Masaryk University
Faculty of Informatics

Computational biomarker discovery: methods and practice

Habilitation Thesis
Collection of articles

Ing. Vlad Popovici, M.Sc., Ph.D.

Brno, 2017

Acknowledgement

The articles collected in this thesis span more than a decade of research activity and are the result of many fruitful collaborations with researchers across the world. My warmest thanks are due to all coauthors of the selected papers. Last but not least, I would like to thank my family for their unshattered patience and especially my wife for her continuous support and encouragements during all these years - without her, this thesis would have never been written.

Abstract

The development of high throughput techniques allows us the exploration of the biological samples at a scale never achieved before. Only two decades ago, the bottleneck of biological discoveries was on the experimental side. Today, it shifted on the analytical side and led to more and more computational disciplines to be drawn into play. The definition of bioinformatics nicely traces these evolutions: in the beginning, it was seen as the application of computer programs to sequence alignment, protein structure prediction and virtual evolution, while nowadays it is more of an umbrella term for a wide spectrum of methods combining computer science, statistics, mathematics and engineering with the goal of analyzing and interpreting biological data.

The present thesis gathers twelve peer-reviewed journal articles in the field of bioinformatics that are related to biomarker discovery and validation. While the methods developed and employed are not specific to any particular pathology, the majority of the results were obtained in the field of oncology, particularly in the case of colon and breast cancers. The articles reproduced here deal with various aspects of biomarker discovery: (i) development of methods for gene expression data normalization with applications (Chapters 7, 14, 15); (ii) classifiers for biomarker design and their applications (Chapters 8, 11, 17); (iii) general methodological aspects for biomarker discovery and validation applied to problems in breast and colon cancers (Chapters 9, 10, 13); and (iv) methods for histopathology image analysis in the context of molecular data for proxy biomarker discovery (Chapters 12, 16, 18). Naturally, this is an over-simplified view since each of these articles is falling under several categories.

The thesis is written as a commentary to a collection of journal articles with estimated personal contribution to each article varying between 5% and 80%, for an average of about 40%.

Contents

PART I

COMMENTARY

1 Introduction

The last years of the XX-*th* century witnessed a true technological revolution in biology: the development of first DNA microarrays. They represented a major step forward from the previous semi-quantitative techniques as, for the first time, it was possible to measure the expression level of hundreds (and later, tens of thousands) of genes. The biology was entering the high-throughput data generation era. The first published results from expression profiling experiments were extremely encouraging so, in the beginning, it was hoped that most of the diseases with high impact (social and economic) would have found a cure within a decade. Yet, twenty years later we still face the same problems in predicting the outcome of a treatment or the likelihood of a cancer to metastasize, despite the tremendous developments during this period. With a few exceptions (e.g. BCR/ABL fusion gene in chronic myelogenous leukemia has now a targeted treatment with very good results), the large majority of cancers are still treated with standard chemotherapy as half a century ago.

So what went wrong? Actually, nothing! As with any new technology of high impact, false hopes and plain naivety fooled us in believing that, finally, the holy grail of modern medicine - individualized treatment - was within grasp. However, these new technologies allowed us to gain insights into a totally new dimension of biology that greatly expanded our knowledge - but also brought numerous challenges in digesting the new types of data.

This thesis is about such challenges of extracting actionable gems of knowledge from large collections of high-throughput genomic data and their transformation into predictive and prognostic models. Additionally, we discuss later development in integrating computational pathology tools both for biomarker discover and for developing a more comprehensive view of the disease of interest. A number of methods for addressing these challenges are presented and discussed and they represent a volume of work in bioinformatics spanning the last decade.

It is clear that, today, the microarrays - the main technological platform used throughout this work - are slowly fading away being replaced by a more versatile technique - the RNASeq. Nevertheless, the work and the results reviewed here remain valid since most of the problems one has in building predictive/prognostic models are the same for RNASeq: normalization, batch effects, validation, model learnability and model interpretability. These aspects are addressed in the various articles reproduced here (and in the corresponding supplemental materials available online from the respective journals) and they represent but a drop in the ocean of all the choices one is presented when challenged to mine genomic data.

The expression of genes represents one facet of the biological reality, many other perspectives could be added by considering the information at protein or epigenetic level, or even at a different scale such that tissue or organism level (Figure 1.1). Ideally, all these data would be taken into account when investigating a pathology but our current ability of managing, mining and interpreting such complex collections of data is still limited.

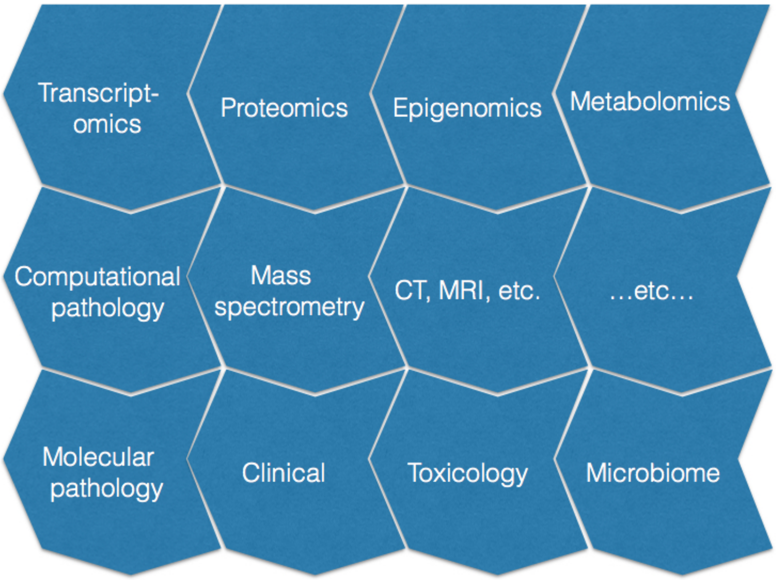


Figure 1.1: Data puzzle in biomarker discovery: a plethora of modalities that each bring a different perspective on the investigated biological phenomenon.

The rest of this first part starts with some background information and a short overview of the technological aspects to facilitate the understanding of the subsequent discussions (Chapters 2,3). Then, the next chapters are dedicated to commenting some aspects of biomarker discovery from both gene expression and histopathology images (Chapters 4,5). The discussion includes some additional results that were not published but which may help enrich the reproduced articles. Finally, some concluding remarks are given in Chpater 6.

2 Of DNA and gene expression

2.1 DNA and genetic information

The *deoxyribonucleic acid (DNA)* is a large molecule that encodes all the biological information needed for the development and reproduction of all living organisms. It is formed of a pair of strands inter-twined in the so called *double helix*. The constitutive unit of this molecule is a *nucleotide* - a monomer consisting of a nucleobase (one of the cytosine (C), guanine (G), adenine (A) or thymine (T)), a sugar (deoxyribose) and a phosphate group. The nucleotides are bound one to another via the covalent links between sugars and phosphates which alternate forming the sugar-phosphate backbone. The nucleobases of one strand bind to the complementary ones from the opposite strand according to base paring rules: C and G, A and T - thus the two strands are said to be antiparallel. The process of base binding is called *hybridization* (or annealing) and the opposite process, of de-coupling the two strands is called *denaturation* (or melting).

The information is coded in the sequences of bases and it relates in part to the production of various proteins or the regulation of various processes. The molecular unit of transmission of hereditary information is the *gene* - a variable-length sequence of bases. In humans, it is estimated that the number of genes is somewhere between 20,000 and 25,000 and they are organized in 23 pairs of *chromosomes*. The *central dogma* of the molecular biology provides a simplified workflow of information transmission within a biological system: "DNA is used to produce ribonucleic acid (RNA) (transcription), RNA is used to produce proteins (translation)". However, there are many other information flows that are not cover by this model, for example the methylation processes which alter the gene expression levels. These main flows of information at molecular level are depicted in Figure 2.2.

The *DNA replication* ensures the transmission of information from parent to progeny and involves the replication of the DNA by a protein complex called replisome, usually in the S-phase of cell cycle. *DNA transcription* is the process of producing messenger RNA (mRNA) from a segment of DNA by RNA polymerases, mainly under the control of various transcription factors. After some post-trasncription mod-

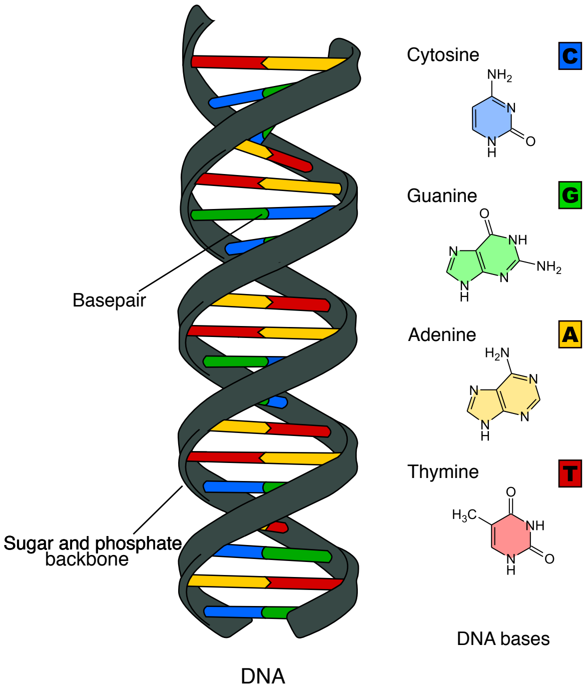


Figure 2.1: Schematic representation of the DNA molecule. Adapted from `https://en.wikipedia.org/wiki/Nucleic_acid`

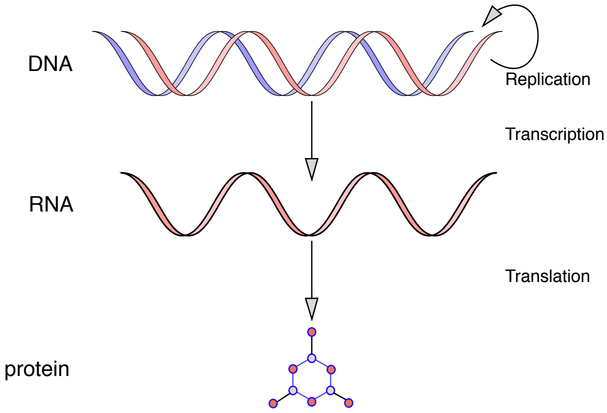


Figure 2.2: A very schematic representation of the main information flow according to the central dogma

ifications (most notably alternate splicing) of precursors of mRNA, the mRNA is externalized from the nucleus and its *translation* takes place in ribosomes leading to the production of polypeptides that, after further processing, will finally result in proteins.

This overly simplified description of the information flow at molecular level provides the basis for understanding the interest of measuring the gene expression levels: in general, it is assumed that the amount of mRNA produced from various genes can be equated to gene expression levels and is proportional with the amount of protein products resulting. Thus, the key of the whole process is the reliable estimation of the amount of specific mRNA sequences.

2.2 DNA microarrays

The *DNA microarray* technology has its roots in electrophoresis methods for the detection of known DNA sequences (Southern blotting) and dates back from the last decades of the twentieth century. It consists in challenging a set of target DNA fragments with a series of known (complementary) DNA sequences and measuring the abundance of the resulting bound molecules, which is usually obtained as the intensity of some electrical signal. In general, the DNA (micro)array is a substrate (nylon membrane, glass or plastic) on which a number of single stranded DNA fragments of known length and sequence are deposited. On this, the properly prepared single stranded target DNA (or RNA) is "washed over" with the intent of having the fragments of target DNA attaching to the the probes (hybridizing) and thus allowing the quantification of their abundance levels (Figure 2.3). While, in theory, this approach should allow a precise quantification of gene expression levels (as abundance of specific sequences), many factors influence the quality of the final measurements, starting from microarray design (including the selection of probes and their distribution over the microarray surface) and sample preparation (quality of the extracted DNA/RNA, chemical agents, amplification protocol, etc), to the data preprocessing methods (normalization, batch effects removal, etc.). An exhaustive presentation of this subject is beyond the scope of the present dissertation, but we will briefly present the data generation process on an Affymetrix (`http:`

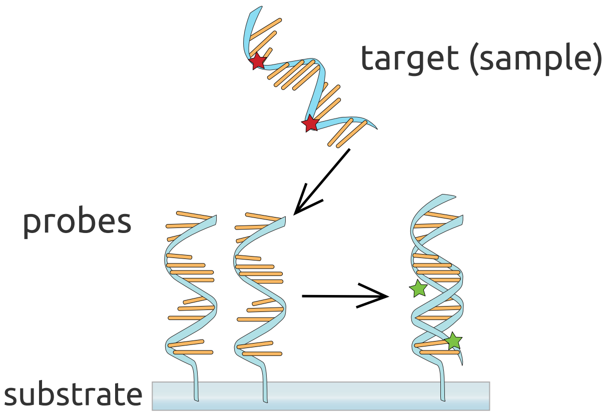


Figure 2.3: The core principle of gene expression level measuring.

//www.affymetrix.com) platform, which became the *de facto* standard for microarray experiments.

We will exemplify the experimental protocol for the Affymetrix Gene Chip HG-133 Plus 2.0 array. The array contains 1,300,000 unique nucleotide probes (features, of length 25 nucleotides) targeting more than 47,000 transcripts and variants representing about 39,000 human genes (and candidate genes). A characteristic of the Affymetrix arrays is the use of probe pairs: for each target sequence there is a pair of probes designed such that one matches perfectly the target (perfect match probe: PM) while the second one has a single nucleotide mismatched (mismatch probe: MM) and is supposed to be used as a negative control to improve the specificity of the measurements. In this array, there are 11 pairs of probes for each sequence, forming a *probeset*. As a side remark, we note that in the latest versions of analytical protocols, these MM probes are no longer used.

A hybridization experiment (Figure 2.4) involves the following main steps[1]:

1. isolation and quantitation of total RNA from the sample

2. reverse transcription to obtain complementary cDNA

[1] see the technical manual at https://assets.thermofisher.com/TFS-Assets/LSG/manuals/expression_analysis_technical_manual.pdf

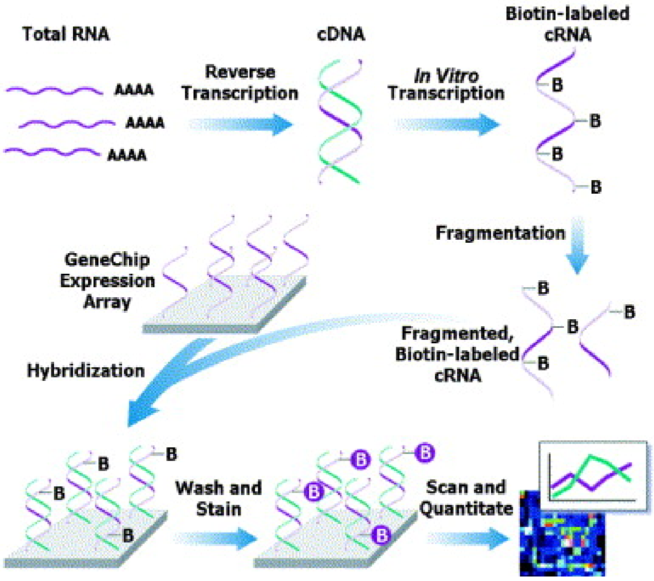


Figure 2.4: Overview of a hibridization experiment on an Affymetrix platform (from [25])

3. transcription and labeling to complementary RNA cRNA, followed by

4. fragmentation (by sonication) to obtain short single stranded RNA segments

5. the RNA segments are hybridized on the array and, after washing it, the raw transcript abundance is obtained as the intensity signal in a scanned image.

Each of these steps has an influence on the final result and a deviation from the protocol may lead to errors that are difficult to detect. As a consequence, aside from following standardized protocols, one has to resort to a number of preprocessing data manipulation and preliminary analyses before data can be considered clean enough for proper analysis. Another consequence is that simply combining data from different protocols is normally not possible due to strong effects introduced in data generation step by individual laboratories (even technicians) and a batch effect removal step is mandatory.

Before concluding this chapter, we note that the quality of the original RNA extracted from the biological sample is equally important. Most notably, the fixation of the biological specimens in formalin and paraffin blocks leads to a degradation of the genetic material. Special protocols need to be devised for such cases, protocols that are accompanied by specific computational methods for data normalization (see Sections 7, 14 and 15 in the present dissertation).

3 Gene expression data preprocessing

Due to the technical variations in gene expression level measurements, their direct analysis is posed to fail. To cope with various artifacts introduced by the experimental conditions and to enhance the signal, various preprocessing steps are needed. It is hoped that after the preprocessing, most of the "true" signal is preserved while the noise is reduced, making the measurements comparable across the samples and, if possible, across experiments.

The preprocessing of microarray experimental data is usually tailored to the platform and many alternative paths are available. The question of choosing the "right" preprocessing pipeline has received considerable attention from the beginning (see, for example [3, 2]) but no definite answer was given. Nevertheless, through experimentation and learning from earlier failures, standard preprocessing workflows emerged for major platforms. Since in almost all the examples discussed in the present dissertation the Affymetrix platform is used (or its derivatives), we will briefly review the main steps for data preprocessing and curation for analysis. One has to bear in mind that while a lot of these preprocessing steps can be (and are) automatized, the detection of abnormalities relies in many cases on manual inspection and *ad hoc* judgement.

3.1 Data acquisition and background correction

As mentioned, for the Affymetrix plaform (as for the majority of microarray platforms), the initial raw data is obtained by scanning the microarrays and quantifying the intensity of the light at each probe location (see Figure 3.1). The amplitude of the signal is given by the quantizer (scanner) and is typically between 0 and $2^{16} - 1$ (for a 16 bit quantizer). From Figure 3.1 it is apparent that the signal is affected by both systematic (while the probes are distributed randomly on the array, the images show a clear stripe pattern of lower intensities which may be attributed to the scanner) and random noise (darker or lighter spots, in the image indicated by white arrows). The background correction has the purpose of removing the systematic noise and performs a locally weighted adaptive background estimation (e.g. 2-percentile

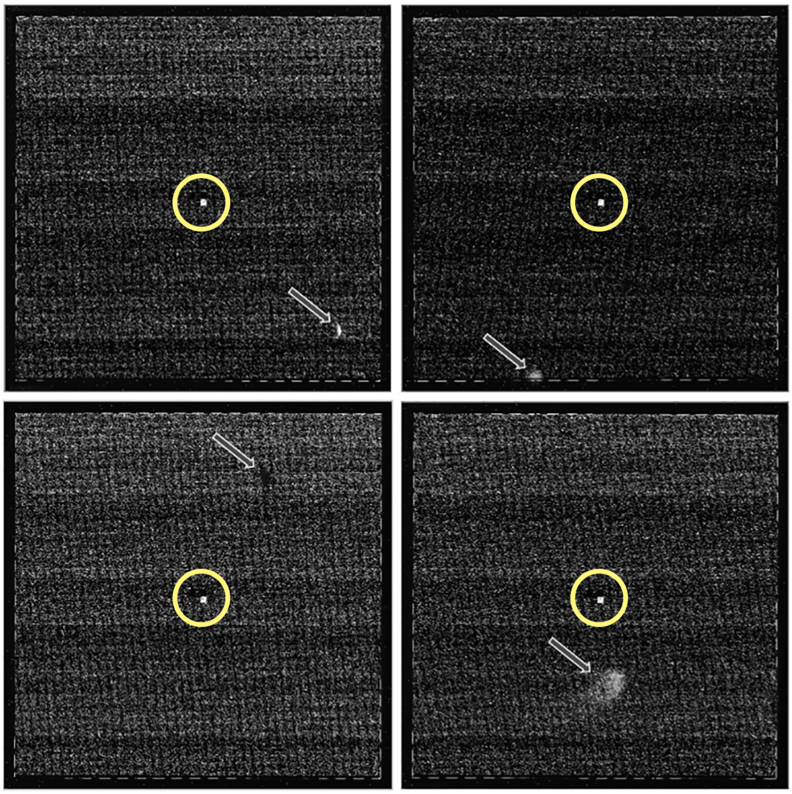


Figure 3.1: An example of scanned Affymetrix microarrays. The yellow circles indicate control spots while the white arrows indicate artifacts. (adapted from [28]).

of the signal intensity in the local neighborhood) and subtraction.

3.2 Signal estimation at probeset level

The goal of this step is to estimate gene (more precisely, probeset)-specific intensity values as a proxy for the amount of transcript in the sample. As mentioned, the Affymetrix microarrays use a pair of probes - perfect match (PM) and mismatch (MM) - in an attempt to improve the specificity of the signal by subtracting an estimate of unspecific hybridization (non targeted molecules that, nevertheless, bind to the probe). The estimates of these quantities are based on robust statistics (like Tukey's biweight estimate). Once they are estimated, for each probeset an average difference is used as the final signal intensity estimate and a scaling step is used to ensure that the signal is comparable across arrays. This signal estimation procedure is the one initially proposed by Affymetrix and implemented in their MAS5 normalization method.

We noted already that the use of MM probes has been discontinued in the later versions of various preprocessing methods, mainly because experience has shown that MM probes are unreliable and often have higher intensities than the PM probes. This observation led to the development of alternative probeset-level signal estimators such as those implemented in RMA (log-scale robust multi-array analysis), which tries to exploit the whole set of arrays in order to obtain better estimates [12]. This method uses median polish iterative procedure for obtaining the estimates of the probeset-level signals.

As a final step, it is customary to apply quantile normalization over the whole batch of arrays in an experiment, in order to align the distribution of the signal across all arrays. The justification of this step comes from the observation that most of the genes are expressed similarly across experimental conditions, with only a small fraction being *differentially expressed*, hence it is reasonable to assume that the overall distribution of signal intensities should not vary much across arrays.

Aside from the MAS5 and RMA, a number of additional methods have been proposed, but they did not reach the popularity of the two mentioned here. For a comparison and discussion, see [11]. In the various articles included in the last chapter of the present dissertation, these two methods are used, with RMA being used in almost all the cases.

One fundamental difference should be emphasized between the MAS5 and RMA (and any other multi-array normalization method): in contrast with MAS5, the multi-array methods estimate their parameters and the probeset-level signal by considering all available arrays (*i.e.* all arrays with enough quality, see next section) and, thus, the final result is influenced, at least theoretically, by each array in an experiment. This observation has implications in evaluating the performance of the predictive models, since any re-sampling method relies on the assumption of independence between training and testing sets. It is clear that, in order to enforce this independence, a proper performance estimation procedure would repeat the normalization of each training set and normalize the testing set using the parameters estimated on the training set alone. This aspect will be later discussed in Chapter 4 (and in Sections 9 and 10).

3.3 Quality control for Affymetrix microarrays

An essential step - actually a suite of steps applied at different stages of data preprocessing - is the control of quality of the samples included in analysis. We have already mentioned the visual inspection of the scanned images (see Figure 3.1) which can help identifying obviously defective arrays. However, as the number of arrays in an experiment increases, this task becomes tedious in addition of being subjective. To help providing a more quantitative measure of array quality, different quality scores and coefficients have been proposed along with guidelines for selecting the quality criteria. Still, depending of the type of experiment, these guidelines have to be adapted. For example, in the case of profiling archival material (formalin fixed, paraffin-embedded (FFPE)), it is expected that the overall intensity of the signal is lower than in the case of fresh frozen material, due to the degradation of the DNA. This aspect is addressed at several occasions in the articles reproduced here – see Sections 11, 14 and 15.

Two main criteria are used to judge the quality of the individual microarrays:

- *percentage of present calls (%PC)*: this is the main quality metrics provided by Affymetrix from the earlier versions of their arrays and is computed using Wilcoxon rank test to test whether significantly more PMs have higher signal than their corresponding MMs and produces a detection call (absent, present or marginal). This method is implemented within the MAS5 normalization procedure. For example, if an array has %PC below 80% than one may choose to call it defective and remove it from further processing.

- *median normalized unscaled standard error (median NUSE)* is the procedure usually used in the context of RMA normalization and relies on the estimation of the residuals from fitting a probe-level model on all arrays in a batch. Briefly, the model assumes that the normalized, background-adjusted, probe-level data is a linear combination (in log-space) of a gene expression in an array, a probe-level effect and a measurement error. Visualizing the distribution of the residuals in an array can help in identifying artifacts but, as mentioned, this is not a feasible approach for large sets. Therefore, a summary statistics (like median) can be computed on a per-array basis and used in deciding whether an array is of sufficient quality (for example, a cut-off value of 1.02 for median NUSE has proven reasonable for fresh frozen samples).

An in depth discussion of this matter is beyond the scope of the present dissertation and has been addressed in several publications (see, for example [13]). In the context of using gene expression data derived from archival material (FFPE) we performed a less formal comparison of the two approaches and found them to be highly concordant. In Figure 3.2 the two criteria are plotted for a set of 240 arrays. The "traditional" cut-off for fresh-frozen samples seemed to be too drastic, therefore a less stringent value - indicated in red - for median NUSE was adopted, since it was expected to have in general a lower signal from this arrays. Still, it is apparent that there is a direct relation between the two quality metrics.

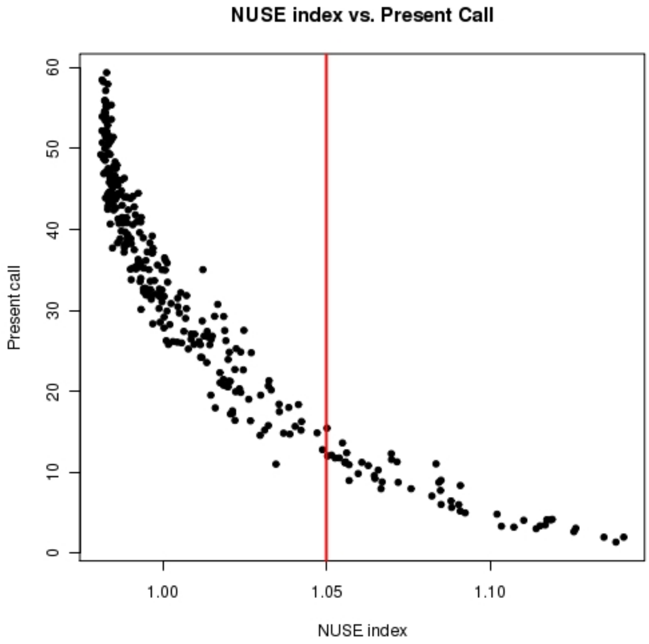


Figure 3.2: Present call *vs.* median NUSE on a set of 240 customized Affymetrix microarrays. This data has been used in articles reproduced in Sections 11, 17.

3.4 A note on normalizing PCR expression data

Polymerase chain reaction (PCR) is a technique of amplifying a DNA region several orders of magnitude to allow the detection and then quantification of the number of copies of that region, which is then converted into an expression level. Without delving into details, we just remark that this technique allows accurate measurement of expression levels of several tens of genes (thus it is not a high-throughput technique as microarrays) and is used as a diagnostic tool due to its low costs and relatively fast processing time. As always, this method requires a proper data normalization before any gene-based score can be computed.

The normalization implies computing a differential expression level with reference to the expression of one or several control genes (housekeeping genes) that are supposed to have stable expression in the given condition. We have shown that some of the traditional control genes may actually vary across a number of cases in breast and colon cancer. Thus, a new set of control genes had to be selected for different pathologies. To avoid the tedious trial-and-error cycle of experimental biology, we developed a computational method that exploits public data from microarray platforms ([23] and Section 7) and proposes a score for estimating the suitability of the candidate gene. This method has been used for selecting control genes in several experiments and as basis for computing genomic prognostic scores in breast cancer (see Sections 14, 15, also [16, 1]).

In Figure 3.3 the variability (standard deviation) as a function of mean expression level for several control genes is shown in a PCR experiment involving 25 fresh frozen breast cancer samples. We note that the proposed procedure allowed the identification of more robust control genes (lower variability, stable low expression - RPLP0, UBB, RPS11) which formed the basis of the reference point for the genomic score.

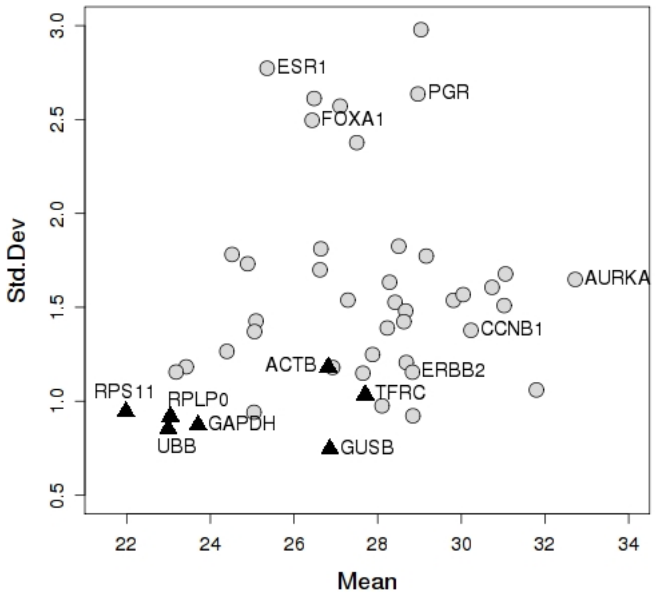


Figure 3.3: Control genes (solid black triangles) and target genes for the development of a genomic score. The newly proposed control genes RPLP0, UBB, RPS11 perform better than the traditional ones GUSB, ACTB and TFRC.

4 Comments on the performance of predictive and prognostic models built on gene expression data

From the beginning, it was clear that one of the main applications of the microarray technology would be in the development of predictive and prognostic models. Here distinguish between two major types of models built on gene expression data: in line with the standard biostatistical nomenclature, we call a model predictive if it is intended to predict whether a patient would respond or not to a treatment, and we call a model prognostic if it is intended to predict whether or not a patient would die from the disease (in a reasonable time frame) or if it predicts the time to an event (such as disease relapse or death).

In general, building predictive models is much more complicated because they require a proper experimental design (as the prognostic ones as well), but one is required to prove that the model is indeed predictive for treatment response, rather than just prognostic (within the specific treatment regimen). Another complication arise from the fact that, at least in theory, for building a predictive model one has to compare a treated group of patients with an untreated group (normally patients would be randomly assigned to the two groups at the enrollment). In the case of severe pathologies like cancer it is nowadays not possible to obtain such cohorts (with, maybe, the exception of very early stages), since denying the treatment to a patient would be unethical. Hence, most of the predictive models refer to predicting the benefit from adding a new compound to the standard of care.

One of the early successes was the identification of the molecular basis of a subtype of the chronic leukemia characterized by the expression of the chimeric BCR/ABL fusion gene [14] and a corresponding targeted therapy [26]. The first prognostic models to reach widespread were intended to predict survival of patients with breast cancers. While they seemed to have a genuine prognostic value, they also stirred a lot of critics quite early on. The main concern was related to the reproducibility of the models: independently developed models led to different gene signatures and were apparently contradictory (see, for example [9]). It took years and a sustained effort to realize that breast cancer was a heterogeneous disease also from a molecular perspective, hence the population sampling would influence dramatically the gene expression signatures learned. Also, since the genes were not independent but rather formed clusters of co-regulated genes (gene modules) some models picked one or other gene from various clusters, the final lists having a small number of common genes. On top of these biology-related causes, the chosen computational modeling approaches were quite different and hence the results differed as well. This early story should serve as a warning that, from a computational/machine learning perspective, there is no single "true" predictive model, but rather there are several views on the same reality. What matters in the end, is the validation and reproducibility of the claimed results, not necessarily the names of the genes in the models.

4.1 General considerations on model learnability

In practice, the usual scenario for developing a new predictive or prognostic model starts with a biological or clinical problem - for example, "build a prognostic model for triple-negatve breast cancer patients". This means that the patient population is (normally) well defined (here, "triple-negative" meaning ER-, PgR- and Her2- breast cancer) as is the end-point (say, time-to-relapse). However, there is no guarantee that the solution to the problem actually exists and, if it exists, whether it can be found in the given feature space (in our case, the gene expression space). Hence, the fundamental question is: *can a model be actually learned for the given problem*? And the usual approach to answer this question is to sistematicaly try solving the problem using a number of different approaches. But when the results are not satisfactory, the question becomes even more difficult to answer, because one can wonder whether the sample size was enough, or whether the methods attempted were appropriate, etc. Ideally, we would like to have a "score" that would indicate how difficult a problem is, independent of the methods. Clearly, as stated, this is an unsolvable problem, but insights into the problem difficulty can be gained by examining the performance of some basic classifiers. In a sense, we would like to find a method of characterizing the prob-

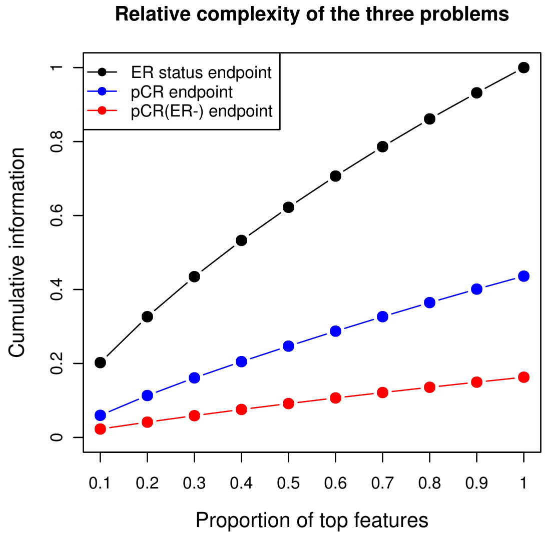


Figure 4.1: Problem complexity as a function of cumulative information: a "simpler" problem would have more informative features and, hence, the curresponding curve would be above the more complex problems (e.g. black curve)

lem difficulty similarly as the classifiers can be characterized by the Vapnik-Chervonenkis dimensionality [31].

This is the context in which we set out to investigate the impact of problem difficulty and sample size, with applications to a classification problem in breast cancer. The MAQC-II project provided the perfect opportunity (and the required data sets) for this investigation. In our investigation (Section 9, [22]) we introduced a new index called *cumulative information* which was used to approximate the problem complexity (Figure 4.1). It is clear that this index is an over-simplification (for example, it does not account for the inter-variable correlations), but it proved its utility in ranking the problems under investigation. This ranking was then confirmed by the classifiers' performance which reproduced the ranking.

Interestingly, the obtained ranking of the biological problems mimicked the clinicians intuition that predicting the oestrogen-receptor (ER) status is much easier that predicting whether a patient will have

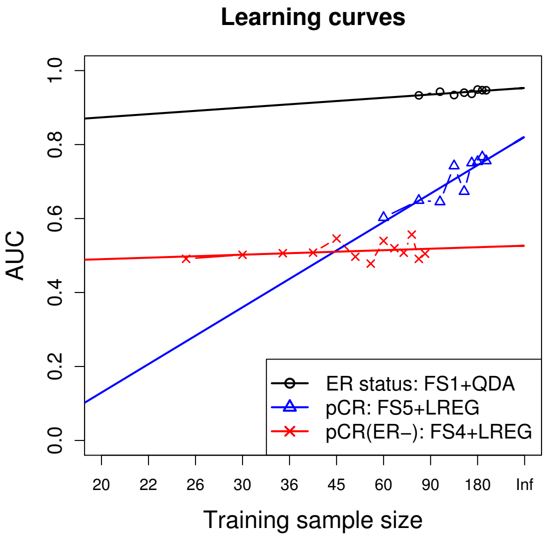


Figure 4.2: Learning curves for three problems in breast cancer. Note the logarithmic scale on the x-axis. See Section 9 and [22] for details.

a complete response to neo-adjuvant chemotherapy (in breast cancer).

We also studied the influence of the sample size on the quality of the predictions. While the required sample size for constructing a classifier for a given endpoint can be estimated only in toy examples (under constraining assumptions regarding the underlying distributions), we found that using *learning curves* for guiding the sample size selection is more appropriate, even though much more computationally intensive. For the same three problems ranked above, the learning curves are shown in Figure 4.2.

The three learning curves suggest different behaviors of the classifiers (here only one representative classifier was chosen per problem): while for the easiest problem, increasing the sample size seem to bring little benefit (as seems to be the case for the most difficult problem as well), for the average difficulty, the learning curves suggest that the model could still be improved - at the cost of doubling (or even tripling) the size of the data set. For the most difficult problem, it seems that there is little hope in gaining anything. Of course, these observations should be taken cautiously, since extrapolating the learning curves may prove delicate, even though the sample size used was in the order of 200 cases.

4.2 A note on model performance estimation

The fast uptake of the microarray technology in biological and clinical studies put under pressure the existing data analysis capabilities of various laboratories and led to a series of sub-optimal or even erroneous analyses. For example, in an early critical review of the gene expression-based prognostic and predictive models in oncology, Dupuy and Simon [8] found that more than 50% of the studies contain at least one of the 3 fundamental errors they considered: (i) incorrect control for multiple-testing in gene filtering; (ii) spurious claims in class discovery studies (usually based on "visual" discovery of classes in cluster analysis); and (iii) incorrect cross-validation procedures resulting in optimistically-biased performance estimation. A decade later, the situation improved dramatically as much more experience has been gained from the many failed trials.

The initial results of Dupuy and Simon [8] were among the causes for setting up the MAQC-II project by US's FDA. The main results are reproduced in Section **??**. Besides them, many other side projects were focussing on more specific aspects. Here we present a different perspective on the published results. Indeed, one key question is whether the estimated performance (at modeling stage) is reproduced by the independent validations. The design of the MAQC-II allowed the investigation of these aspects on a large scale collection of predictive models. The set up of the MAQC-II required that each participating team (in total, there were 38 teams, most of them from US, but also a few from Europe and Asia) submitted a data analysis plan (written before having access to the data) to be applied to each of the 13 predictive problems. Each team had the choice in modeling one, several or all of the problems (endpoints), but the modeling procedure was required to be identical. In the end, more than 30,000 models were submitted for blind validation (some teams had chosen to submit thousands of models, one model per combination of parameters) and by comparing their observed performance with the initial estimates, one can gain some insights into the stability of various an-

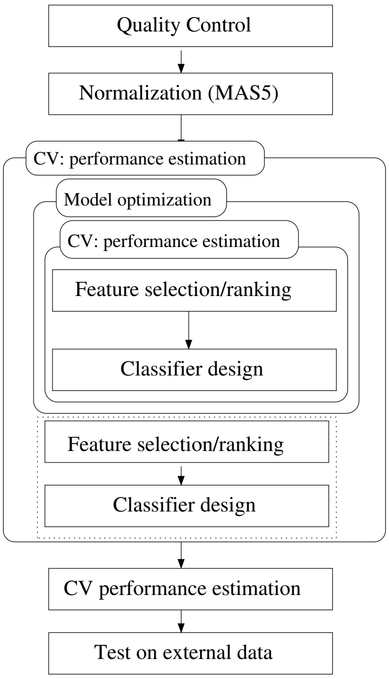

Quality Control
Normalization (MAS5)
CV: performance estimation
Model optimization
CV: performance estimation
Feature selection/ranking
Classifier design
Feature selection/ranking
Classifier design
CV performance estimation
Test on external data

Figure 4.3: Overall design of the analytical pipeline for the MAQC-II project, as put forward by the SIB team

alytical approaches. While the full discussion and main results are presented in Section 10 and [27], here we will briefly discuss some of the results regarding the model performance estimation procedures. At the time of the project, I was with the Swiss Institute of Bioinformatics (SIB) which I represented in the project, hence the "SIB" refers to the results I obtained.

Given the rather constraining nature of the exercise, we have adopted a very conservative approach, with well tested procedures for feature selection and classifier design. The drawback was clear: we might have not profited from tuned-to-the-problem modeling strategies, but the expected benefits were a more robust performance and small bias in its estimation. The overall design of the processing pipeline is given in Figure 4.3.

A first observation is that the overall performance of the system was evaluated by cross-validation (actually a repeated (10×) 5-fold cross validation), corresponding to the outer CV loop in the Figure 4.3.

It was already mentioned elsewhere (Sections 3.2 and 3.3) that the most commonly used quality control and normalization procedures are using batches of microarrays for parameter estimation. In order to avoid repeatedly fitting these models inside the cross-validation, we opted for procedures applicable on individual arrays (%PC for quality control - Section 3.3, and MAS5 for normalization 3.2), thus being able to perform them only once, outside the cross-validation, without violating the performance estimation assumptions (different data for model building and model assessment).

Also, because the training set sample size was relatively small, the feature selection method employed was based on single-variable assessment (ratio of between- to within-group sum of squares - similar to Fisher criterion) and the optimal number of features was estimated within an inner cross-validation loop (Figure 4.3). The same constraints restricted the types of classifiers tested, to those that experience has shown to perform robustly on large number of problems (diagonal LDA, general LDA, logistic and penalized logistic regression, and CART). Again, any meta-parameter those methods had were optimized in the inner cross-validation loop. The details of SIB's data analysis plan were presented during MAQC-II plenary meeting at FDA headquarters in Washington DC (March 2008).

In Figure 4.4 are shown the boxplots corresponding to the performance estimates provided by each participating team. It is already clear that a wide range of performances were expected to be observed on the independent validation sets. More troubling, for the "positive control" endpoints (H and L), which were supposed to be predicted with a performance close to 1.0 (for AUC), some of the models seemed to be far off-target.

Finally, when comparing the estimated/expected performance of the models with the observed performance, the results were even more worrying: in some cases, the AUC bias (in absolute values) was around 0.5 which would be the difference between a perfect model and a totally random one (see left panel in Figure 4.5). It was clear that the performance estimation procedure of some teams was extremely biased. This led to the selection of a set of rules that would guarantee, in principle, an unbiased (or, more likely, low bias) estimate of the performance (but not necessarily a good performance), recommendations that are now part of the FDA's guide for good practices

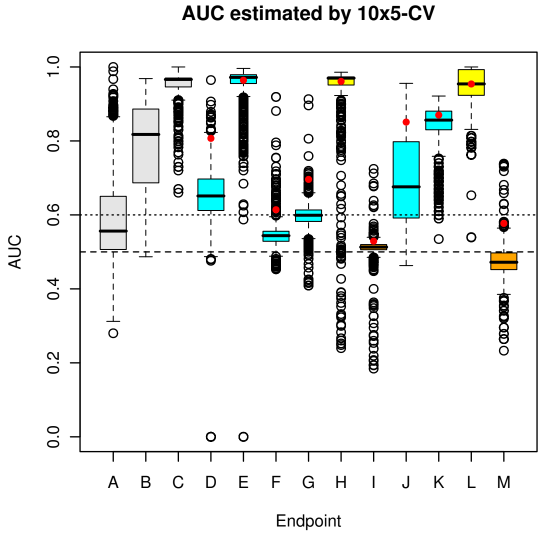


Figure 4.4: Estimated performance of the SIB's models by repeated 5-fold cross validation (red dots) and the estimated performance of all other models submitted to MAQC-II, for the 13 endpoints (A-M). An AUC above 0.6 was considered useful for the prediction of the endpoint. The yellow endpoints (H and L) were later revealed to be "positive controls" - problems easy to predict, while the orange endpoints (I and M) were "negative controls" - randomly assigned labels. The first three endpoints (A-C) were related to toxico-genomics and were not modeled by SIB.

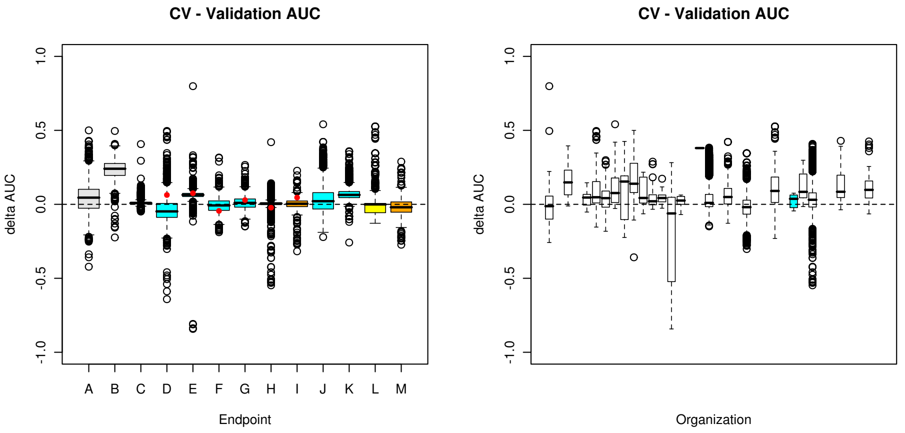


Figure 4.5: Bias of the performance estimation procedures: estimated AUC minus observed AUC. Left panel: estimation bias by endpoint (red dots correspond to SIB), right panel: estimation bias by participating team (cyan colored box corresponds to SIB).

in biomarker discovery.

5 Integration of pathology images: towards a multimodal biomarker discovery

Modern investigation methods in biology and clinical research rely more often than not on multiple sources of information. For example, combining clinical observations, like patient survival or pathologic response, with gene expression data is, nowadays, routinely used for discovering new biomarkers or therapy targets. Similarly, combining gene expression and copy number variation information and/or methylation data, brings a new level of resolution when investigating molecular changes at cellular level. Each of these different modalities provides another perspective on the same underlying biological reality. The current proposal is concerned with the combination of three modalities: histopathology imaging, gene expression and clinical data.

Digital pathology is an active research field which employs methods of image processing and analysis for assisting the interpretation and understanding the histopathology slide images. It has the potential of proposing a more quantitative, thus less subjective, characterization of the slides and of introducing new image descriptors, which can be further mined for diagnostic and prognostic clues [10]. As an example of combining digital pathology and clinical information, the recently proposed immune response score [15], relies on precise counts of all T-cells (TH1) in whole slide images, a task that is clearly too tedious for a human expert to perform for each sample to be assessed. Initial tests show that the score has more prognostic power than even the well-established TNM grading, providing an excellent argument in favor of using digital pathology in clinical practice. However, it relies on special staining for correctly labeling the different types of T-cells.

The histopathology assessment of the samples can be combined with the gene expression and clinical data in a joint model. For example, the tumor grade (a histopathology categorical variable, usually with three levels: "well differentiated", "moderately differentiated" and "undifferentiated") can be combined with expression of ESR1 gene and a genomic proliferation score in order to build a prognostic score for breast tumors (similar to [29]). In this approach the information extracted from the histopathology modality is highly filtered (the human expert extracts only several aspects from the pathology slides, according to the current practice) and extremely summarized (in the above example, only three values are possible), in contrast with the gene expression, which preserves basically all its information. While the process of filtering and summarization greatly improves the signal-to-noise ratio and eases the interpretability of the data, it does this at the expense of discarding some useful information and limiting the descriptive vocabulary of the histopathology images.

In this context we set out to investigate different aspects of exploiting and integrating the whole slide imaging in the biomarker discovery pipeline. Our approach, in contrast with many others, takes a completely data-driven perspective, without benefitting from - nor being biased by - pathologist's expert supervision. However, once the model were built, the pathologists were called for validating them. The advantage is that the resulting models revealed new features, some of them - this being the drawback - without a clear correspondence in pathology practice. The full description is given in Sections 12, 16 and 18.

The computational approach taken was based on extensions to bag-of-visual features method [7]. These extensions aimed at producing more descriptive dictionaries for the histopathology images and investigated the possibility of structuring the visual dictionaries around some semantical terms, allowing an easier interpretation of the results (see Figure 5.1.

Another computational aspect addressed was the optimization of the visual features for the purpose of analyzing pathology images and in-depth analyses were performed using both "classical" features (from Gabor wavelets to local binary patterns) (see [5] for detailed results) and convolutional neural networks features (as used in [20]). Also, in [4] we propose a hierarchical quantification schema for building multiresolution visual dictionaries. Annectodically, this approach led to features that were more appealing to an expert pathologist than the convolutional features, but the overall performance of the system was lower.

The main message of all this investigations is that not only it is possible to combine whole-slide imaging with molecular/gene ex-

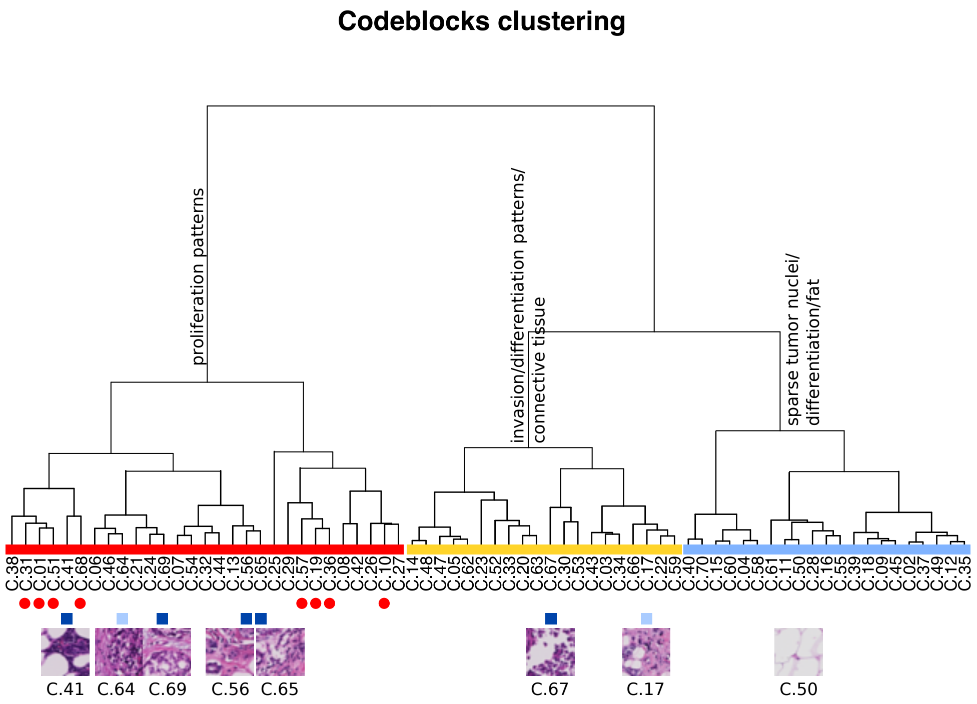


Figure 5.1: The structure of a visual dictionary for breast cancer: three main clusters of features can be observed, each related to a different architectural pattern with clear interpretation [18].

pression data, but this combination reveals new connections between the "genetic program" and the tissue architecture. The results were obtained from breast and colon cancer data but the techniques employed are easily applicable to other pathologies as well. The most important results were:

- Construction of a joint imaging and genomic prognostic score in breast cancer.

  In the case of breast cancer, proliferation of cancer cells is a strong prognostic marker (in addition to, and independent of, ER and Her2 hormonal statuses), and well known to pathologists. Hence it was reassuring to see that a part of the image-based features that were correlated with the outcome reproduced this result. In Figure 5.1 the structure of the visual dictionary is shown along with semantic annotations. The prognostic score built solely on image features was almost as sensitive as its gene expression correspondent. However, a combined image-expression score performed even better [18].

- Recognition of "BRAF-positive" high-risk patients with colorectal cancer.

  In [21] we described a novel gene expression signature identifying a high risk group of patients (positive by the signature, hence called "BRAF-positive"). This group assembled both patients with BRAF V600E mutation (a known risk marker) and other patients not harboring this mutation but subject to the same dismal outcome. In [17] and later in [24] we describe a system for identifying most of these patients based solely on histopathology images. This result shows that typical tumor architecture patterns (including the papillary/serrated phenotype) can be linked to this high risk group allowing its recognition even in the absence of molecular profiling.

- First steps towards a computational imaging characterization of inter-tumoral heterogeneity.

  The fact that solid tumors are heterogeneous is well known and the recent advances in molecular profiling confirmed and expanded the characterization of tumor subtyping. In colorectal cancer, we have proposed a molecular taxonomy based on 5 subtypes, with the observation that not all tumors could be assigned with high confidence to these subtypes (there are probably some lower prevalence subgroups that were not enough present in our data for their characterization) [6]. Using deep learning image features and a hierarchy of support vector machines we were able to construct a decision system capable of predicting the molecular subtypes with high confidence (for four out of the five subtypes). To the best of our knowledge, this is the first image-based predictor of molecular subtypes for any tumor type. The implications of our result go beyond the prediction aspect. Indeed, what we noticed is that part of the tumor correspond to one subtype while other parts may correspond to different subtypes. The decision mostly reflects the dominant subtype but this observation clearly indicates the sensitivity of the results obtained from molecular profiling to the tumor sampling strategy, and impacts the large majority of the results published so far in the field.

6 Concluding remarks

The previous chapters tried to present the context in which the articles reproduced in the second part were written, and also to bring to the attention of the reader some results that were left aside from the publications. It is clear - also from browsing the articles - that the methods evolved along with the technology and the type of problems one is facing in biomarker discovery. The latest directions of research, concerning the joint analysis of histopathology images and gene expression (or other molecular data), clearly show that having the right data and a modern computational infrastructure allows one novel ways of exploring an ever increasingly complex biological reality.

Bibliography

[1] Janine Antonov, Vlad Popovici, Mauro Delorenzi, Pratyaksha Wirapati, Anna Baltzer, Andrea Oberli, Beat Thurlimann, Anita Giobbie-Hurder, Giuseppe Viale, Hans Altermatt, Stefan Aebi, and Rolf Jaggi. Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue. *BMC Cancer*, 10(1):37, 2010.

[2] Reija Autio, Sami Kilpinen, Matti Saarela, Olli Kallioniemi, Sampsa Hautaniemi, and Jaakko Astola. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics*, 10 Suppl 1:S24–S24, December 2008.

[3] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[4] Eva Budinská, Fred Bosman, and Vlad Popovici. Experiments in molecular subtype recognition based on histopathology images. In *International Symposium on Biomedical Imaging*, pages 1168–1172, Prague, CZ, June 2016. IEEE.

[5] Eva Budinská, Lenka Čápková, Daniel Schwarz, Ladislav Dušek, Rolf Jaggi, Josef Feit, and Vlad Popovici. Gene expression-guided selection of histopathology image features. In *IEEE 15th International Conference on Bioinformatics and Bioengineering*, pages 1–6, Belgrade, 2015. IEEE.

[6] Eva Budinská, Vlad Popovici, Mauro Delorenzi, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, Pu Yan, John Graeme Hodgson, Scott Weinrich, Fred Bosman, and Arnaud Roth. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *Journal of Pathology*, 231(1):63–76, July 2013.

[7] G Csurka, C Dance, and L Fan. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[8] A Dupuy and R M Simon. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI Journal of the National Cancer Institute*, 99(2):147–157, January 2007.

[9] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, January 2005.

[10] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and B Yener. Histopathological image analysis: a review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[11] R A Irizarry. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):15e–15, February 2003.

[12] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003.

[13] Jeanette N McClintick and Howard J Edenberg. Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*, 7:49, January 2006.

[14] J V Melo, D E Gordon, NCP Cross, and J M Goldman. The Abl-Bcr Fusion Gene Is Expressed in Chronic Myeloid-Leukemia. *Blood*, 81(1):158–165, 1993.

[15] B Mlecnik, M Tosolini, A Kirilovsky, A Berger, G Bindea, T Meatchi, P Bruneval, Z Trajanoski, W H Fridman, F Pages, and J Galon. Histopathologic-Based Prognostic Factors of Colorectal Cancers Are Associated With the State of the Local Immune Reaction. *Journal of Clinical Oncology*, 29(6):610–618, February 2011.

[16] Andrea Oberli, Vlad Popovici, Mauro Delorenzi, Anna Baltzer, Janine Antonov, Sybille Matthey, Stefan Aebi, Hans Jörg Altermatt, and Rolf Jaggi. Expression profiling with RNA from formalin-fixed, paraffin-embedded material. *BMC Medical Genomics*, 1(1):9, April 2008.

[17] Vlad Popovici. Towards the identification of tissue-based proxy biomarkers. In *AMIA Joint Summits on Translational Science proceedings*, pages 75–83, San Francisco, US, 2016.

[18] Vlad Popovici, Eva Budinská, Lenka Čápková, Daniel Schwarz, Ladislav Dušek, Josef Feit, and Rolf Jaggi. Joint analysis of histopathology image features and gene expression in breast cancer. *BMC Bioinformatics*, 17(1):209, 2016.

[19] Vlad Popovici, Eva Budinská, and Mauro Delorenzi. Rgtsp - a generalized top scoring pairs package for class prediction. *Bioinformatics*, 27(12):1729–1730, 2011.

[20] Vlad Popovici, Eva Budinská, Ladislav Dušek, Michal Kozubek, and Fred Bosman. Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. *Bioinformatics*, January 2017.

[21] Vlad Popovici, Eva Budinská, Sabine Tejpar, Scott Weinrich, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Tao Xie, Fred T Bosman, Arnaud D Roth, and Mauro Delorenzi. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *Journal of Clinical Oncology*, 30(12):1288–1295, April 2012.

[22] Vlad Popovici, Weijie Chen, Brandon G Gallas, Christos Hatzis, Weiwei Shi, Frank W Samuelson, Yuri Nikolsky, Marina Tsyganova, Alex Ishkin, Tatiana Nikolskaya, Kenneth R Hess, Vicente Valero, Daniel Booser, Mauro Delorenzi, Gabriel N Hortobagyi, Leming Shi, W Fraser Symmans, and Lajos Pusztai. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Research*, 12(1):R5–R5, 2010.

[23] Vlad Popovici, Darlene R Goldstein, Janine Antonov, Rolf Jaggi, Mauro Delorenzi, and Pratyaksha Wirapati. Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics*, 10:42–42, December 2008.

[24] Vlad Popovici, Aleš Křenek, and Eva Budinská. Identification of "BRAF-Positive" Cases Based on Whole-Slide Image Analysis. *BioMed Research International*, 2017(24):1–7, 2017.

[25] Margaret M Ryan, Stephen J Huffaker, Maree J Webster, Matt Wayland, Tom Freeman, and Sabine Bahn. Application and optimization of microarray technologies for human postmortem brain studies. *Biological Psychiatry*, 55(4):329–336, February 2004.

[26] D G Savage and K H Antman. Imatinib mesylate - A new oral targeted therapy. *New England Journal of Medicine*, 346(9):683–693, 2002.

[27] Leming Shi, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M Goodsaid, Lajos Pusztai, John D Shaughnessy, André Oberthuer, Russell S Thomas, Richard S Paules, Mark Fielden, Bart Barlogie, Weijie Chen, Pan Du, Matthias Fischer, Cesare Furlanello, Brandon D Gallas, Xijin Ge, Dalila B Megherbi, W Fraser Symmans, May D Wang, John Zhang, Hans Bitter, Benedikt Brors, Pierre R Bushel, Max Bylesjo, Minjun Chen, Jie Cheng, Jing Cheng, Jeff Chou, Timothy S Davison, Mauro Delorenzi, Youping Deng, Viswanath Devanarayan, David J Dix, Joaquin Dopazo, Kevin C Dorff, Fathi Elloumi, Jianqing Fan, Shicai Fan, Xiaohui Fan, Hong Fang, Nina Gonzaludo, Kenneth R Hess, Huixiao Hong, Jun Huan, Rafael A Irizarry, Richard Judson, Dilafruz Juraeva, Samir Lababidi, Christophe G Lambert, Li Li, Yanen Li, Zhen Li, Simon M Lin, Guozhen Liu, Edward K Lobenhofer, Jun Luo, Wen Luo, Matthew N McCall, Yuri Nikolsky, Gene A Pennello, Roger G Perkins, Reena Philip, Vlad Popovici, Nathan D Price, Feng Qian, Andreas Scherer, Tieliu Shi, Weiwei Shi, Jaeyun Sung, Danielle Thierry-Mieg, Jean Thierry-Mieg, Venkata Thodima, Johan Trygg, Lakshmi Vishnuvajjala, Sue Jane Wang, Jianping Wu, Yichao Wu, Qian Xie,

Waleed A Yousef, Liang Zhang, Xuegong Zhang, Sheng Zhong, Yiming Zhou, Sheng Zhu, Dhivya Arasappan, Wenjun Bao, Anne Bergstrom Lucas, Frank Berthold, Richard J Brennan, Andreas Buness, Jennifer G Catalano, Chang Chang, Rong Chen, Yiyu Cheng, Jian Cui, Wendy Czika, Francesca Demichelis, Xutao Deng, Damir Dosymbekov, Roland Eils, Yang Feng, Jennifer Fostel, Stephanie Fulmer-Smentek, James C Fuscoe, Laurent Gatto, Weigong Ge, Darlene R Goldstein, Li Guo, Donald N Halbert, Jing Han, Stephen C Harris, Christos Hatzis, Damir Herman, Jianping Huang, Roderick V Jensen, Rui Jiang, Charles D Johnson, Giuseppe Jurman, Yvonne Kahlert, Sadik A Khuder, Matthias Kohl, Jianying Li, Menglong Li, Quan-Zhen Li, Shao Li, Zhiguang Li, Jie Liu, Ying Liu, Zhichao Liu, Lu Meng, Manuel Madera, Francisco Martinez-Murillo, Ignacio Medina, Joseph Meehan, Kelci Miclaus, Richard A Moffitt, David Montaner, Piali Mukherjee, George J Mulligan, Padraic Neville, Tatiana Nikolskaya, Baitang Ning, Grier P Page, Joel Parker, R Mitchell Parry, Xuejun Peng, Ron L Peterson, John H Phan, Brian Quanz, Yi Ren, Samantha Riccadonna, Alan H Roter, Frank W Samuelson, Martin M Schumacher, Joseph D Shambaugh, Qiang Shi, Richard Shippy, Shengzhu Si, Aaron Smalter, Christos Sotiriou, Mat Soukup, Frank Staedtler, Guido Steiner, Todd H Stokes, Qinglan Sun, Pei-Yi Tan, Rong Tang, Zivana Tezak, Brett Thorn, Marina Tsyganova, Yaron Turpaz, Silvia C Vega, Roberto Visintainer, Juergen von Frese, Charles Wang, Eric Wang, Junwei Wang, Wei Wang, Frank Westermann, James C Willey, Matthew Woods, Shujian Wu, Nianqing Xiao, Joshua Xu, Lei Xu, Lun Yang, Xiao Zeng, Jialu Zhang, Li Zhang, Min Zhang, Chen Zhao, Raj K Puri, Uwe Scherf, Weida Tong, and Russell D Wolfinger. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8):827–838, July 2010.

[28] James P Stewart, Susan Richman, Tim Maughan, Mark Lawler, Philip D Dunne, and Manuel Salto-Tellez. Standardising RNA profiling based biomarker application in cancer—The need for robust control of technical variables. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1868(1):258–272, August 2017.

[29] Carina Strand, Cecilia Ahlin, Pär-Ola Bendahl, Marie-Louise Fjällskog, Ingrid Hedenfalk, Per Malmström, and Mårten Fernö. Combination of the proliferation marker cyclin A, histological grade, and estrogen receptor status in a new variable with high prognostic impact in breast cancer. *Breast Cancer Research and Treatment*, 131(1):33–40, 2012.

[30] Sun Tian, Paul Roepman, Vlad Popovici, Magali Michaut, Ian Majewski, Ramon Salazar, Cristina Santos, Robert Rosenberg, Ulrich Nitsche, Wilma E Mesker, Sjoerd Bruin, Sabine Tejpar, Mauro Delorenzi, Rene Bernards, and Iris Simon. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *Journal of Pathology*, 228(4):586–595, October 2012.

[31] V N Vapnik and A YA Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of probability and its applications*, 16(2):264–280, 1971.

PART II

SELECTED ARTICLES

This second part is dedicated to reproducing a number of articles published over the years in various international journals, dealing with different theoretical and practical aspects of mining, designing, evaluating and validating a number of biomarkers and gene expression signatures. For each article, its current (as of August 2017) bibliometrics information is provided. Most of the articles are accompanied by supplemental materials freely available online at the respective journals web pages.

7 Selecting control genes for RT-QPCR using public microarray data

- BMC Bioinformatics 10(42), 2009
- IF: 2.448
- number of citations: 30
- personal contribution (60%): method design, data processing, experiment implementation (with R code) and statistical analyses, manuscript writing

BMC Bioinformatics

BioMed Central

Methodology article

Selecting control genes for RT-QPCR using public microarray data

Vlad Popovici*[1], Darlene R Goldstein[1,2], Janine Antonov[3], Rolf Jaggi[3], Mauro Delorenzi[1] and Pratyaksha Wirapati[1]

Address: [1]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland, [2]Institut de mathématiques (IMA), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland and [3]Department of Clinical Research, University of Bern, CH-3010 Bern, Switzerland

Email: Vlad Popovici* - vlad.popovici@isb-sib.ch; Darlene R Goldstein - darlene.goldstein@epfl.ch;
Janine Antonov - janine.antonov@dkf.unibe.ch; Rolf Jaggi - rolf.jaggi@dkf.unibe.ch; Mauro Delorenzi - mauro.delorenzi@isb-sib.ch;
Pratyaksha Wirapati - pratyaksha.wirapati@isb-sib.ch

* Corresponding author

Published: 2 February 2009

BMC Bioinformatics 2009, **10**:42 doi:10.1186/1471-2105-10-42

Received: 11 August 2008
Accepted: 2 February 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/42

Abstract

Background: Gene expression analysis has emerged as a major biological research area, with real-time quantitative reverse transcription PCR (RT-QPCR) being one of the most accurate and widely used techniques for expression profiling of selected genes. In order to obtain results that are comparable across assays, a stable normalization strategy is required. In general, the normalization of PCR measurements between different samples uses one to several control genes (e.g. housekeeping genes), from which a baseline reference level is constructed. Thus, the choice of the control genes is of utmost importance, yet there is not a generally accepted standard technique for screening a large number of candidates and identifying the best ones.

Results: We propose a novel approach for scoring and ranking candidate genes for their suitability as control genes. Our approach relies on publicly available microarray data and allows the combination of multiple data sets originating from different platforms and/or representing different pathologies. The use of microarray data allows the screening of tens of thousands of genes, producing very comprehensive lists of candidates. We also provide two lists of candidate control genes: one which is breast cancer-specific and one with more general applicability. Two genes from the breast cancer list which had not been previously used as control genes are identified and validated by RT-QPCR. Open source R functions are available at http://www.isrec.isb-sib.ch/~vpopovic/research/

Conclusion: We proposed a new method for identifying candidate control genes for RT-QPCR which was able to rank thousands of genes according to some predefined suitability criteria and we applied it to the case of breast cancer. We also empirically showed that translating the results from microarray to PCR platform was achievable.

Background

Real-time quantitative reverse transcription PCR (RT-QPCR) has become a method of choice for gene expression profiling in a large number of applications. However, obtaining reliable measurements still depends on the choice of control genes on which the baseline level is constructed. Selecting the control genes remains a critical point in the normalization process. Often, a short list of candidates is produced based on non-systematic and/or often poorly defined biological considerations.

In early studies, normalization was usually based on a single control gene. More recently, the trend is to use several control genes whose average expression level (on a log-scale) is used as baseline [1,2]. Suitable control genes are selected from a short list of 10–15 genes by ranking them according to a criterion that essentially selects those genes having low variation across samples. We describe briefly a few such methods below.

[2] introduces a stability coefficient which is used along with the coefficient of variation for ranking the genes from a predefined list of candidates. Gene stability is defined in terms of average standard deviation of the log-ratios of pairs of candidate genes. Genes are ranked by iteratively removing those most unstable. This approach has the drawback that repeated comparison of pairs of genes is required, which is feasible only when the number of candidates is small. In addition, the method implicitly assumes that there are no co-regulated genes. A model-based approach proposed by [1] aims at estimating the overall variation as well as the between sample variation of each candidate gene. However, with this approach it is cumbersome to integrate different platforms. In an application to plant pathogen profiling, [3] investigates a list of 18 pre-selected candidate housekeeping genes, using the method proposed in [2] and RT-QPCR for measuring the gene expressions. [4] proposes a PCA-based statistical analysis to identify the most suitable control genes among 13 candidates which were selected such that they had independent functions in cellular maintenance.

[5] introduces a strategy which combines the coefficient of variation, maximum fold change and mean expression value in a ranking criterion that is applied to a large number of samples representing a wide variety of tissues. All these samples were hybridized on either Affymetrix HG-U133A or HG-U133 Plus 2.0 arrays and quantile-normalized together prior to ranking. Only probesets common to both arrays were used, with probesets targeting the same gene averaged into a single value.

There are some important differences between the methods described above and our approach (described below). Firstly, in contrast with all the studies based on PCR, we do not require a short list of candidate genes to be produced before assessing their suitability as control genes. Instead, we screen all the genes represented on the microarray chips, giving us the opportunity to assess genes that have not been reported previously. Moreover, we take a meta-analytical approach to the problem, first creating an independent ranking within each data set then aggregating these rankings into a single list. This approach has the advantage of being platform- and normalization-independent. In addition, the approach is not limited to using only genes common between different data sets. Also, by not using the coefficient of variation, we can treat uniformly both single and two-colors arrays. Thus, we are able to exploit data obtained from different platforms without requiring them to be normalized together. Furthermore, the meta-analytical approach allows us to integrate gene lists produced using our ranking system with other ranked gene lists from the literature and we do not require all data to be normalized together. Another key difference is that we introduce a new stability coefficient that combines the mean expression and the standard deviation in a ranking criterion that corresponds to our requirements for candidate control genes for RT-QPCR. In general, these requirements are:

- low variability across different specimens (*e.g.*, subtypes of tumors or normal tissues);

- high and moderate level of expression, such that control genes with expression levels across a larger range may be selected;

- consistency across experiments and platforms.

A key question is whether it is possible to select genes from microarray studies that perform as control genes on PCR platform, given that the two technologies are different. We hypothesize that translating the list of candidate genes from microarray to PCR platform is feasible and we provide empirical evidence in this sense.

Results

Data sets and pre-processing steps

We have collected ten publicly available data sets [6-15], listed in Table 1, from which we derived the quantities of interest: the mean and standard deviations of the log-

Table 1: The ten public microarray data sets used (*n* = number of samples).

Data set ID and reference	*n*	Platform
BWH [6]	47	Affymetrix U133v2
EMC [7]	286	Affymetrix U133A
EXPO [8]	1375	Affymetrix U133Plus2
JRH2 [9]	61	Affymetrix U133A
MGH [10]	60	Agilent
NKI [11]	337	Agilent (custom)
STOCK [12]	159	Affymetrix U133A, B
TGIF1 [13]	49	Affymetrix U133A
UNC [14]	153	Agilent HuA1
UPP [15]	249	Affymetrix U133A, B

intensities (on Affymetrix platforms) or of the log-ratios (on Agilent platforms).

We note here that the original EXPO data set contains a number of different pathologies, but we restrict analysis here to eight different types of cancer (breast, colon, endometrium, kidney, lung, ovary, prostate and uterus) for which a sufficient number of samples existed. EXPO breast cancer samples (n = 328) were used to produce both the breast cancer and general cancer lists of candidate genes.

The Affymetrix data are available as MAS5.0 normalized values. The Agilent data contains log-ratios (base 10) and mean-centered log-intensities. The standard deviations of log-intensities (Affymetrix) and log-ratios (Agilent) were used as measures of variability. The means of log-intensities (both Affymetrix and Agilent) were used as measures of average expression level.

When multiple probesets of the same gene are present, only the most variable one is used. We consider all genes from each platform, the aggregation methods used being able to cope with 'missing' genes (those not represented on the array). Considering only those genes common to all platforms is an unnecessary limiting constraint, as increasing the number of data sets and the heterogeneity of the collection leads to a successively smaller intersection of genes.

Before any further usage of the data, we reduce the variability across platforms by scaling with a factor given by a first order LOESS fit of the data. The effect of this transformation can be seen in Figure 1, where the black line represents the fitted curve. This simple approach seems effective, except for genes with low expression. However, as we are interested in genes with higher mean expression, this deficiency is not problematic.

Ranking the genes

Let us consider that we have M microarray data sets, each containing expression values of a set of genes G_k, k = 1,...,M, and let $G = \cup_k G_k = \{1,...,N\}$ be the set of all genes represented at least once in any of these data sets.

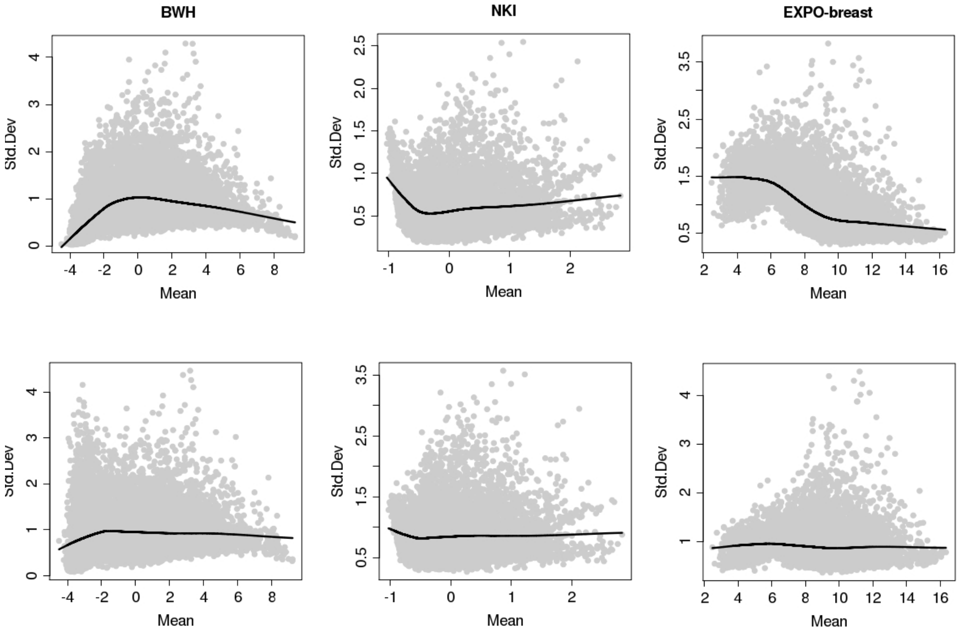


Figure 1
Example of variance stabilization by LOESS correction. LOESS correction applied to three data sets: BWH, NKI and EXPO-breast, respectively. The first row shows the original data with the fitted first order LOESS curve, while the second row shows the variance-stabilized data.

Gene scores

We aim to design a scoring function which ranks the genes such that higher scores correspond to genes that are more suitable to be used as control genes. As mentioned above, the score has to combine each gene's mean expression and standard deviation into a single value such that higher expression levels and lower variances (standard deviations) are favored. Moreover, the score must be independent of the technology used to measure expression levels and the method for normalization.

These requirements lead us to propose a new *stability score* for the gene expression levels. This score for gene *i* in data set *k*, denoted s_{ik}, is defined as

$$s_{ik} = \alpha \log_2(\max\{\mu_{ik} - \beta_k, 0\}) - \sigma_{ik}, \qquad (1)$$

where $\hat{\mu}_{ik}$ and $\hat{\sigma}_{ik}$ are the estimated mean log-intensity and the standard deviation of the gene *i* in data set *k*. The coefficient allows the user to control the trade-off between the mean expression and the standard deviation in gene scoring. Results reported here were obtained with $= 0.25$. The $_k$ parameter allows one to define the level of mean expression below which the genes are not considered for ranking, *i.e.* the score for these genes is $-\infty$. We have set $_k$ to be the 25*th* percentile of the mean expression, for each data set *k*. Genes having a higher score are considered more suitable as control genes. As we see from Eq. 1, high variation in gene expression leads to a lower score when mean expression levels are equal. This is one reason we select the most highly variable probeset from the probesets representing the same gene, in order to encompass the worst-case scenario. Note also that there is no need to normalize the scores to make them comparable across data sets, because they are used solely for ranking the genes within the same data set. Finally, having computed the scores for all the genes within a data set, we order the genes from high to low values of the scores, with ties resolved by ordering by the mean expression (from high to low). From this perspective, the scores can be seen as defining classes of equivalence among genes: all the genes in the same class (having the same score) are equally useful as normalization genes. By using the second ordering criterion, we can select control genes with a desired expression level (examples of classes of equivalence are the equal score levels in Figure 2).

Figure 2 displays the influence of the mean expression level and the standard deviation on the gene score. All genes located on the curves have the same score value (they belong to the same equivalence class). Two consecutive curves are separated by one score unit.

Using this stability score, we ranked the genes from each data set, obtaining the lists that will be later combined. An excerpt from the ten lists for the breast cancer data sets is shown in Table 2 (first ten columns).

Combining results from different data sets

Once genes are ranked according to their scores in each data set (lower ranks correspond to higher scores), the natural next step is to combine these rankings into a global ranked list. We combine the ranks of the genes rather

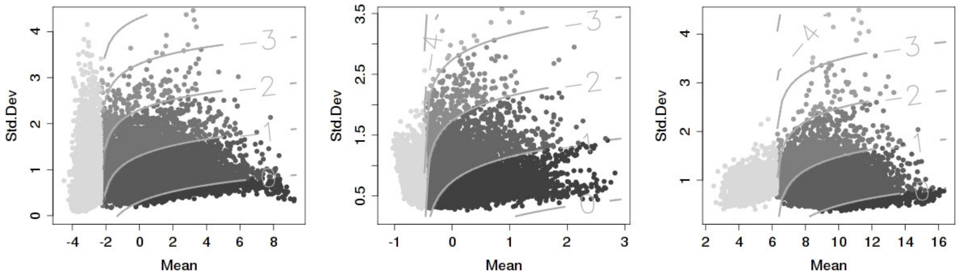


Figure 2

Scatter plots of standard deviation versus mean log-intensity for BWH, NKI and EXPO-breast data sets, respectively. The shading codes the gene stability scores, with darker colors indicating higher scores. These three data sets are from different microarray platforms. The light gray points indicate the discarded genes (those with mean expression level below the value – see Eq. 1). The curves correspond to equal score levels and are one score unit apart.

than their scores to avoid normalizing the scores across different data sets, thereby achieving platform-independence. To this end we use the *rank product score* [16], which is a fast and efficient method for combining ranked lists. It computes, for each gene $i \in G$, a new score

$$R_i = \left(\prod_k \text{rank}_k(i) \right)^{\frac{1}{n_i}}, \qquad (2)$$

where $\text{rank}_k(i)$ is the rank of s_{ik}, the score for gene i in data set k (topmost gene has the rank 1), and n_i is the number of data sets in which the gene i appears. The final list is obtained by sorting the genes in increasing order of R_i. The top 20 genes from the aggregated breast cancer list are given in the 'Meta' (last) column of Table 2.

Validation of the aggregated lists

There is no absolute criterion by which one can judge the quality of the resulting lists. Rather, the aggregated list could be used to select from the top genes (100, for example) those genes that also satisfy further conditions of the specific application.

We can, however, have a subjective impression of the validity of the aggregated list by visualizing the resulting top genes in data sets not used for producing the list. We obtained a list of the top 100 genes by applying the method described above on eight of the ten data sets, leaving NKI and UPP aside as validation sets. The top 100 genes in both validation sets (different microarray platforms) are plotted in Figure 3. As a comparison, we also include the five control genes used in [17] (represented as triangles in the figure). It is seen that the genes are generally concentrated in the lower right part of the plot, corresponding to high mean expression levels and low variance. There is a notable difference between the quality of the results (given by the concentration of the control genes in the lower right corner) on the two platforms, due to the fact that most of the data sets used for gene selection are from Affymetrix platforms. While the top 100 lists contain genes with high stability scores on the Affymetrix platforms (the UPP data set), on the custom Agilent platform (NKI) there are a number of genes that are missed. Nevertheless, those selected still function well as control genes.

Control genes lists

We have analyzed ten different data sets which have samples hybridized on different versions of Agilent and Affymetrix platforms. Using our proposed method, we compiled two different lists of candidate control genes: one specific to breast cancer [see Additional file 1] and one resulting from the analysis of eight different types of cancer, thus applicable to cancers in general [see Additional file 2]. From the breast cancer list we selected two new control genes which were validated in an RT-QPCR assay that also included five previously used control genes (ACTB, TFRC, GUSB, RPLP0 and GAPDH – see [17]) and breast cancer-related genes (*e.g.* ESR1, ERBB2, AURKA, *etc.*). The RT-QPCR results confirm the findings from the microarray analysis and show that more stably expressed control genes can be selected by applying the criteria men-

Table 2: Top 20 control genes from the ten breast cancer data sets and top 20 genes from the aggregated list (Meta column)

BWH	EMC	JRH2	MGH	NKI	STOCK	TGIF1	UNC	UPP	EXPO-breast	Meta
RPL37A	PPIA	RPL41	ZNF557	UBC	RPS11	RPL41	RPS10	RPL9	CALM2	RPL37A
RPL41	CALM2	RPL39	CDR1	UBB	RPS24	RPL37A	RPS18	RPL37A	HNRPA1	RPL27A
RPS18	SRP14	RPL23A	PPP1R2	OAZ1	RPL9	EEF1A1	RPLP1	ACTG1	NACA	RPS18
RPL39	RPL37A	RPL37A	TCN2	DYNLL1	RPL37A	RPL30	RPS11	RPL27A	UBA52	RPL30
RPL23A	RPS18	EEF1A1	SSBP1	RAPSN	RPL41	RPL39	RPS23	CFL1	LAPTM4A	RPL41
RPL9	RPL30	RPS23	RPL27A	PCBP1	RPL27A	PPIA	RPL37A	RPS11	RPL27A	CALM2
RPLP1	RPL27A	RPS27	RPS3	KCNH3	RPL39	ACTG1	RPL11	RPS13	RPL30	RPL27
RPS27	RPS11	CALM2	BRCC3	RPL3	RPLP1	CFL1	RPS15	RPL27	RPL9	K-ALPHA-1
RPL27A	RPL39	RPS18	PTMA	RPL8	UBB	RPS23	RPL14	RPL41	RPL31	RPS11
RPL30	RPS15	ACTG1	ABCF2	MYL6	RPS15A	RPL10	NACA	RPS18	RPL37	RPL39
RPS29	RPS24	RPL10	PCDH18	RPL14	CALM2	CALM2	RPL36AL	RPS15	RPS11	RPS13
ACTG1	RPL32	RPS24	LAX1	RPL7A	NACA	RPS11	UBA52	RPL6	RPS29	NACA
CALM2	RPS15A	RPS15A	TPMT	FAU	RPL30	HNRPA1	NEDD8	RPLP1	RPS24	RPL23A
RPS13	RPLP1	RPL32	GALE	ARF1	CFL1	RPL6	PCBP1	RPL32	RPS13	RPS24
HNRPA1	RPL9	RPL27A	MTCH1	CCT3	RPS13	RPL23A	NDUFB2	RPL31	RPS21	HNRPA1
RPS24	UBB	UBB	ATP5G2	PSAP	RPS3A	K-ALPHA-1	HNRPM	RPL39	UBB	RPL9
RPL31	K-ALPHA-1	RPS29	SF3B2	CD81	RPL37	RPS18	HNRPC	UBB	RPS27A	RPLP1
RPL34	RPS13	RPL30	SND1	SQSTM1	RPS18	EEF1G	NDUFB8	RPS24	RPS15	RPL32
RPS15A	RPL27	PPIA	RPL5	K-ALPHA-1	RPL27	TUBA6	ATP5J2	RPS27	RPL32	LAPTM4A
RPS21	FAU	CFL1	SKAP2	CALR	RPL24	RPS3A	TARDBP	DDX5	RPL24	RPS15A

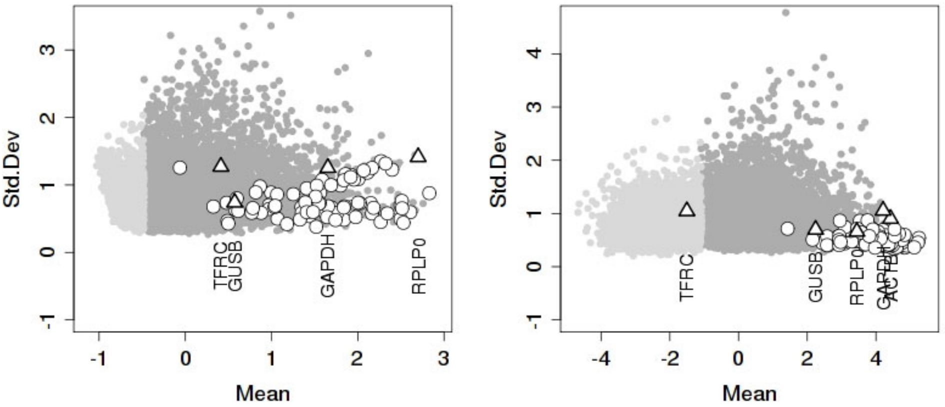


Figure 3
Scatter plots of standard deviation versus mean log-intensity for two validation data sets (from left to right: NKI and UPP). The top 100 breast cancer control genes resulting from aggregating eight data sets are plotted as circles. Triangles correspond to the five control genes used in [17] (NKI does not contain the ACTB gene).

tioned above. Also, they provide empirical evidence supporting the working hypothesis that PCR control genes can be selected from microarray data.

The list of the top 50 control genes obtained from the ten breast cancer data sets is given in Table 3. More comprehensive lists, including one containing the top 2000 candidate breast cancer genes and a similar list compiled from eight different types of cancer, are available [see Additional file 1 and Additional file 2]. In the case of breast cancer control genes, it is interesting to note that some of the "classical" genes (*e.g.* ACTB, GAPDH, TFRC) are not among the top 50.

Evaluation of control genes by RT-QPCR

Motivated by the consistency of the selection process for suitable control genes among different microarray platforms, we performed a small scale RT-QPCR experiment to test the performance of two new control genes along with a number of more commonly used control genes. In this experiment, RNA was isolated from 25 cryo-preserved breast cancer samples and the expression of 47 genes was measured by RT-QPCR [18]. Test genes were selected according to their relatedness to proliferation or estrogen receptor functions. Some of the test genes had been previously identified and used for characterizing primary breast cancers [17]. Two genes, RPS11 and UBB, ranked 9 and 31 in Table 3 respectively, were compared to five additional control genes and to a number of test genes previously measured by [17]. Mean raw expression values of all candidate control and test genes were plotted against standard deviations of each gene (Figure 4). The raw Ct (cycle threshold) value is the number of PCR cycles required for the fluorescence signal to cross the background threshold, so that low Ct values correspond to high expression levels. RPS11 and UBB are clearly among the most stably expressed genes, as their standard deviations are both quite low. Other genes frequently used as control genes are also shown. For comparison, mean expression and standard deviation of several test genes are also indicated. The expression of most test genes is much more variable than UBB and RPS11.

The two new control genes, together with RPLP0, offer the best trade-off between mean expression level and variability, while others like ACTB or TFRC are less stably expressed and therefore seem less suitable for use as normalization genes.

Discussion

We propose a new approach which leverages publicly available microarray data to produce lists of candidate control genes for RT-QPCR. Our method is independent of the microarray platform or normalization methodol-

Table 3: Top 50 control genes as resulting from aggregating the ten breast cancer data sets. Two genes – RPS11 and UBB – were selected as control genes and validated by RT-PCR

Rank	Gene symbol	Gene ID	Description
1	RPL37A	6168	ribosomal protein L37a
2	RPL27A	6157	ribosomal protein L27a
3	RPS18	6222	ribosomal protein S18
4	RPL30	6156	ribosomal protein L30
5	RPL41	6171	ribosomal protein L41
6	CALM2	805	calmodulin 2 (phosphorylase kinase, delta)
7	RPL27	6155	ribosomal protein L27
8	K-ALPHA-1	10376	alpha tubulin
9	**RPS11**	6205	ribosomal protein S11
10	RPL39	6170	ribosomal protein L39
11	RPS13	6207	ribosomal protein S13
12	NACA	4666	nascent-polypeptide-associated complex alpha polypeptide
12	RPL23A	6147	ribosomal protein L23a
14	RPS24	6229	ribosomal protein S24
15	HNRPA1	3178	heterogeneous nuclear ribonucleoprotein A1
16	RPL9	6133	ribosomal protein L9
17	RPLP1	6176	ribosomal protein, large, P1
18	RPL32	6161	ribosomal protein L32
19	LAPTM4A	9741	lysosomal-associated protein transmembrane 4 alpha
20	RPS15A	6210	ribosomal protein S15a
21	DYNLL1	8655	dynein, light chain, LC8-type 1
22	ACTG1	71	actin, gamma 1
23	TUBA6	84790	tubulin, alpha 6
24	SRP14	6727	signal recognition particle 14kDa (homologous Alu RNA binding protein)
25	MYL6	4637	myosin, light chain 6, alkali, smooth muscle and non-muscle
26	RPL24	6152	ribosomal protein L24
27	FAU	2197	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30
28	RPL31	6160	ribosomal protein L31
29	RPS15	6209	ribosomal protein S15
30	MTCH1	23787	mitochondrial carrier homolog 1 (C. elegans)
31	**UBB**	7314	ubiquitin B
32	RPL37	6167	ribosomal protein L37
33	HMGN2	3151	high-mobility group nucleosomal binding domain 2
34	RPS27	6232	ribosomal protein S27 (metallopanstimulin 1)
35	GDF8	2660	growth differentiation factor 8
36	RPL38	6169	ribosomal protein L38
37	RPS29	6235	ribosomal protein S29
38	SULT1C2	27233	sulfotransferase family, cytosolic, 1C, member 2
39	RPL6	6128	ribosomal protein L6
40	UBC	7316	ubiquitin C
41	UBA52	7311	ubiquitin A-52 residue ribosomal protein fusion product 1
42	MRFAP1	93621	Mof4 family associated protein 1
43	HNRPK	3190	heterogeneous nuclear ribonucleoprotein K
44	PARK7	11315	Parkinson disease (autosomal recessive, early onset) 7
45	PSMC1	5700	proteasome (prosome, macropain) 26S subunit, ATPase, 1
46	LOC158572	158572	hypothetical protein LOC158572
47	RPS8	6202	ribosomal protein S8
48	ATP5A1	498	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle
49	EIF4H	7458	eukaryotic translation initiation factor 4H
50	CD63	967	CD63 molecule

ogy, and is able to cope with gene lists that overlap only partially. After screening thousands of genes (generally more than 10,000 genes in each data set), we have produced two separate lists of candidate genes: one specific to breast cancer and one generally applicable to different types of cancer. We do not consider these lists as generally applicable, as the data used do not allow such generalization. Different pathologies may have a different impact on the control genes and some of the control genes we selected may become ineffective in the case of a disease which affects their particular functions. On the other hand, more diverse data should be used if the goal is finding global control genes. The list of the top 50 breast cancer control genes (Table 3) is dominated by ribosomal proteins. This finding is consistent with the fact that ribosomes are a major component of basic physiologic processes in all the cells and not a primary target of changing conditions. Other genes present among the first 50 genes

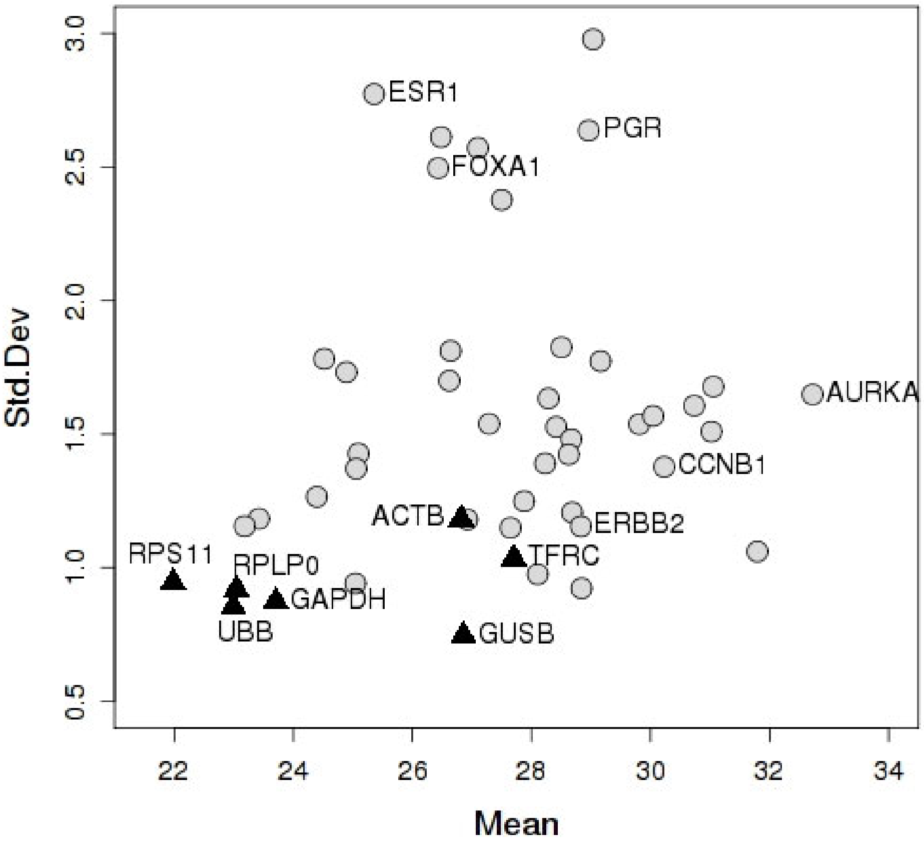


Figure 4
RT-QPCR experiment. Standard deviation as a function of the mean expression level (expressed as raw Ct values) of 47 genes in a RT-QPCR experiment. Higher expression levels correspond to smaller raw Ct values. Control genes are represented by triangles, test genes by circles. The new control genes RPS11 and UBB are in the lower left corner.

code for protein turnover (ubiquitin), tubulin-related proteins or actins, structures which are required in all living cells.

Our results are supported by recent findings of de Jonge *et al.* [5], who used a different ranking method. In addition, the lists of control gene candidates for breast cancer and for diverse types of cancer are similar [see Additional file 1 and Additional file 2], as a large number of the top ranked genes belong to the same functional category (ribosomal genes, protein turnover).

Another important finding is that some of the commonly used control genes in breast cancer (ACTB and TFRC) appear to be less stable than previously assumed. This has an impact on the normalization strategy of the QPCR measurements: indeed, in our more recent experiments we have chosen to use the mean of RPLP0, RPS11 and UBB (on the $\log_2$ scale) for normalizing the expression of test genes.

Finally, we would like to emphasize that these two lists should not be taken in an absolute sense: a gene in top 10 is not necessarily a better choice than a gene in the top 20 to 30. But we do consider it to be definitely a better candidate than a gene not in top 100. Nor do we consider the resulting ranking as providing a solution to the problem of finding normalization genes in all contexts. Rather, the lists produced through this process are meant to guide the choice of control genes while also taking into consideration the specific requirements of any individual analysis. Depending on the planned application, other parameters must be considered. For example, short amplicons or intron-spanning primers must be used when the starting RNA is considerably degraded or when residual DNA contaminations might affect QPCR. The final choice of control genes should be made not by blind adherence to the ranked list, but be imposed by the intended application.

Conclusion

Starting from clearly defined criteria, we have designed a novel method for ranking the candidate genes for their suitability as control genes in RT-QPCR experiments. The genes from a data set were ranked according to their stability score, which represented a trade-off between gene's average expression level and its variance. Finally, the rankings from several data sets were combined into a list of candidate genes, with higher ranked genes being considered to be more suitable as control genes. The proposed approach had the advantage of being platform- and normalization- independent and of not being restricted to only the list of common genes across all data sets.

By applying the proposed method to two particular collections of data sets we were able to produce two lists of candidate genes from which control genes for either breast cancer or more diverse cancer could be easily selected. Two new control genes for breast cancer – UBB and RPS11 – have been identified and validated by RT-QPCR.

Our results support the hypothesis that selecting control genes for QPCR from microarray data is feasible.

Authors' contributions

VP conducted the analysis, devised algorithms and wrote the computer programs. PW collected the datasets and remapped the probes. PW, DRG and MD designed the study and statistical analyses. JA and RJ initiated the biological problems and conducted the RT-PCR validation. All authors have read and approved the final manuscript.

Additional material

Additional file 1

Top 2000 breast cancer candidate control genes. *Excel file containing top 2000 genes as resulted from combining the ten breast cancer data sets.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-42-S1.xls]

Additional file 2

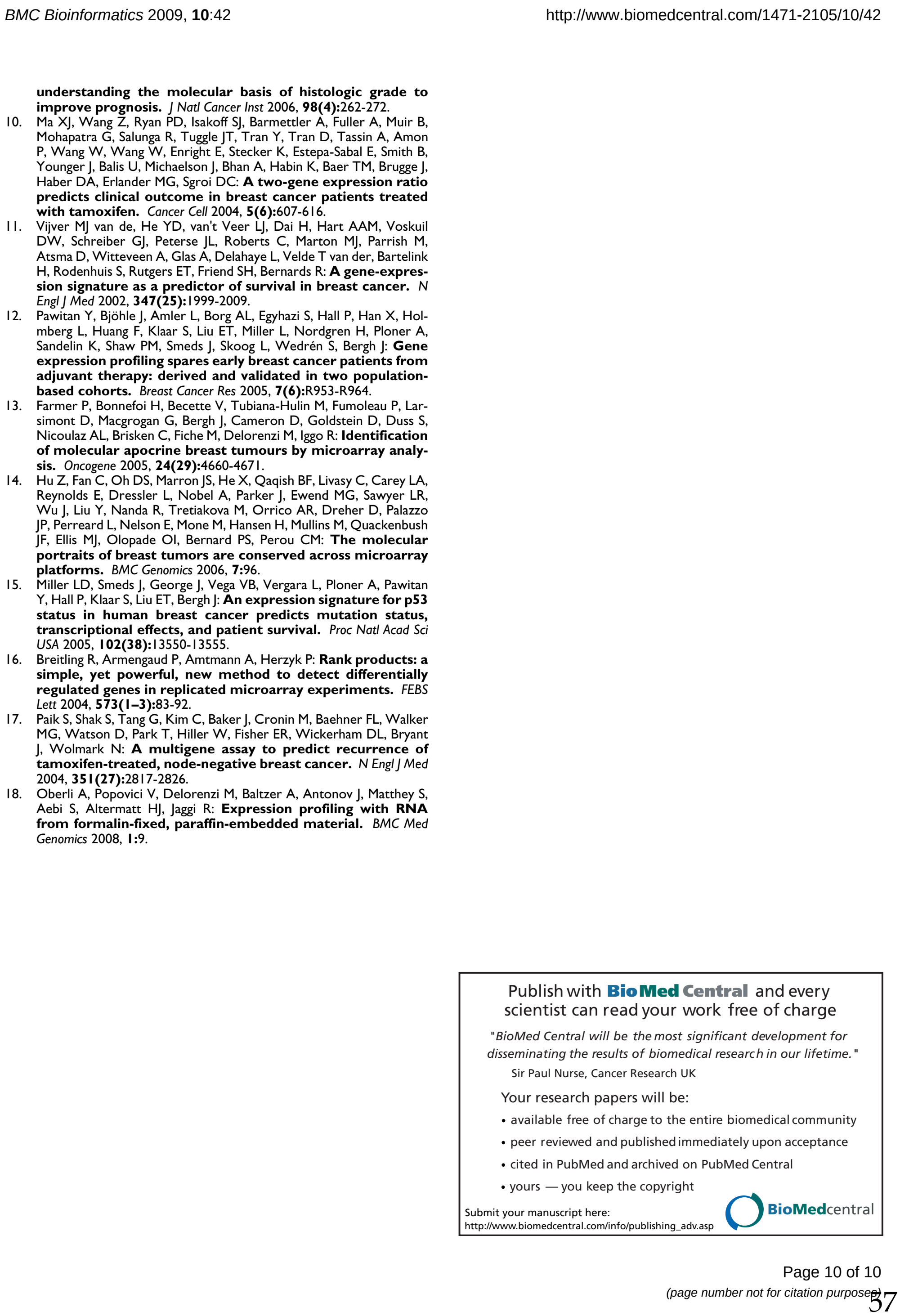
Top 2000 diverse cancer candidate control genes. *Excel file containing top 2000 genes as resulted from combining the eight different types of cancer from the EXPO data set.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-42-S2.xls]

Acknowledgements

This work has been funded by the Swiss National Science Foundation through the National Centre for Competence in Research (Molecular Oncology (MD, VP and PW)) and the National Centre for Competence in Research (Plant Survival (DRG)), and by the Swiss Cancer League (Project OCS 01704-04-2005 (RJ and MD)).

References

1. Andersen CL, Jensen JL, Ørntoft TF: **Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets.** *Cancer Research* 2004, **64(15):**5245-5250.
2. Vandesompele J, Preter KD, Pattyn F, Poppe B, Roy NV, Paepe AD, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biology* 2002, **3(7):**RESEARCH0034.
3. Yan HZ, Liou RF: **Selection of internal control genes for real-time quantitative RT-PCR assays in the oomycete plant pathogen Phytophthora parasitica.** *Fungal Genet Biol* 2006, **43(6):**430-438.
4. de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, Span PN: **Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes.** *Laboratory Investigations* 2005, **85:**154-159.
5. de Jonge HJM, Fehrmann RSN, de Bont ESJM, Hofstra RMW, Gerbens F, Kamps WA, de Vries EGE, Zee AGJ van der, te Meerman GJ, ter Elst A: **Evidence based selection of housekeeping genes.** *PLoS ONE* 2007, **2(9):**e898.
6. Richardson AL, Wang ZC, Nicolo AD, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: **X chromosomal abnormalities in basal-like human breast cancer.** *Cancer Cell* 2006, **9(2):**121-132.
7. Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, van Gelder MEM, Yu J, Jatkoe T, Berns EMJJ, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365(9460):**671-679.
8. IGC: **Expression Project for Oncology.** 2008 [http://www.intgen.org/expo.cfm].
9. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, de Vijver MJV, Bergh J, Piccart M, Delorenzi M: **Gene expression profiling in breast cancer:**



understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006, **98(4):**262-272.

10. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC: **A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.** *Cancer Cell* 2004, **5(6):**607-616.
11. Vijver MJ van de, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Velde T van der, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347(25):**1999-2009.
12. Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Res* 2005, **7(6):**R953-R964.
13. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, Macgrogan G, Bergh J, Cameron D, Goldstein D, Duss S, Nicoulaz AL, Brisken C, Fiche M, Delorenzi M, Iggo R: **Identification of molecular apocrine breast tumours by microarray analysis.** *Oncogene* 2005, **24(29):**4660-4671.
14. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Orrico AR, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7:**96.
15. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.** *Proc Natl Acad Sci USA* 2005, **102(38):**13550-13555.
16. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett* 2004, **573(1–3):**83-92.
17. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351(27):**2817-2826.
18. Oberli A, Popovici V, Delorenzi M, Baltzer A, Antonov J, Matthey S, Aebi S, Altermatt HJ, Jaggi R: **Expression profiling with RNA from formalin-fixed, paraffin-embedded material.** *BMC Med Genomics* 2008, **1:**9.

8 Rgtsp: a generalized top scoring pairs package for class prediction

- Bioinformatics, 27 (12):1729–1730, 2011
- IF: 7.307
- number of citations: 4
- personal contribution (70%): method design, data collection and processing, experimental design and implementation (R package), manuscript writing

Bioinformatics Advance Access published April 19, 2011

Rgtsp: a generalized top scoring pairs package for class prediction

Vlad Popovici [1,2,*], Eva Budinská [1,3] and Mauro Delorenzi [1]

[1]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
[2] Swiss National Center of Competence in Research Molecular Oncology, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3]Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: A top scoring pair (TSP) classifier consists of a pair of variables whose relative ordering can be used for accurately predicting the class label of a sample. This classification rule has the advantage of being easily interpretable and more robust against technical variations in data, as those due to different microarray platforms. Here we describe a parallel implementation of this classifier which significantly reduces the training time, and a number of extensions, including a multi–class approach which have the potential of improving the classification performance.

Availability and Implementation: Full C++ source code and R package Rgtsp are freely available from http://lausanne.isb-sib.ch/∼vpopovic/research/. The implementation relies on existing OpenMP libraries.

Contact: vlad.popovici@isb-sib.ch

1 INTRODUCTION

Top scoring pairs (TSPs) (Geman *et al.* (2004)) are simple two–variables binary classifiers, in which the prediction of the class label is based solely on the relative ranking of the expression levels of the two genes. The rank–based approach to classification ensures a higher degree of robustness to technical variations and makes the rule easily portable across platforms. Also, the direct comparison of the expression level of the genes is easily interpretable in the clinical context, making the TSPs attractive for medical tests.

Let $\mathbf{x} = [x_i]_{i=1,\dots,m} \in \mathbb{R}^m$ be a vector of measurements (*e.g.* gene expression) representing a sample and let the corresponding class label be y, with two classes denoted by 0 and 1. Then, for all pairs of variables i and j, a score is computed,

$$s_{i,j} = P(x_i < x_j|y = 1) - P(x_i < x_j|y = 0), 1 \le i, j \le m \quad (1)$$

where P are conditional probabilities estimated from the data, and the corresponding decision rule is: if $\text{sign}(s_{i,j})x_i < \text{sign}(s_{i,j})x_j$ then predict $y = 1$, otherwise $y = 0$. The pairs are ordered by the absolute values of their scores and the top t pairs ($t \ge 1$) are then considered for the final model (Xu *et al.* (2005); Tan *et al.* (2005); Geman *et al.* (2004)). Remarkably, training a TSP does not require the optimization of any parameter and does not depend on any threshold. Selecting a suitable value for t should be done following the usual machine learning paradigm for optimizing meta–parameters (see, for example, Hastie *et al.* (2001)). Figure 1 shows an example of a TSP predicting the estrogen receptor status. The decision boundary (in grey) is always a line with a slope of 1.

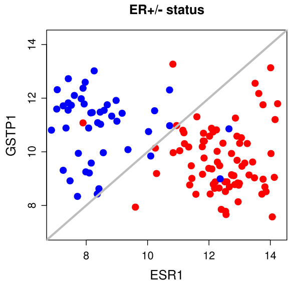


Fig. 1. Predicting estrogen receptor status: if GSTP1 < ESR1, then the sample is considered ER+ (red dots), otherwise ER- (blue dots).

2 IMPLEMENTATION

While the method briefly described above is simple and poses no implementation problems, using it in the context of highly dimensional data requires the evaluation of an extremely large number of pairs of variables making its usage impractical, especially in the context of resampling techniques for performance estimation. However, most if not all of the modern desktop computers are multi–core machines, making parallel programs a feasible alternative to classical serial ones.

Our implementation in C++ exploits the multi-core architecture by using the OpenMP libraries of the system (Chapman *et al.* (2007)), and is wrapped in an R package – Rgtsp. The full source code and the R package are available from http://lausanne.isb-sib.ch/∼vpopovic/research/. As C++ is the main implementation language, the library can easily be extended and integrated with

[*]to whom correspondence should be addressed

other software libraries. Also, the R functions are independent of the domain of application so they could be applied to any kind of data.

3 USAGE EXAMPLES

We present a typical case of using Rgtsp package. These examples represent solely some code snippets and not the full process of developing and assessing the performance of a classifier.

The data used in these examples consists of 130 samples stage I to III breast cancer (Hess *et al.* (2006)) and the goal is to predict the estrogen receptor status (positive or negative coded with ”+1” and ”0”, respectively). For illustration purposes we use only a subset of full data set available from GEO repository under accession number GSE16716.

Before starting R, the user has the option of choosing the number of processing units that will be used, by setting the environment variable OMP_NUM_THREADS. If not set, it defaults to the maximum number of processing units available.

The first steps load the library and the data and build a list of TSPs (note that the matrix X contains the variables as columns):

```
> library(Rgtsp)
> data(mdabr)
> tsp.list = tsp.n(X, y.erpos, 500)
> str(tsp.list)
> print(tsp.list)
```

The function tsp.n() returns at most n TSPs as a list with three components: the first two correspond to the indexes of the selected variables and the third one contains the associated scores. A similar function, tsp.s(), returns all the TSPs that have a score larger than a specified value.

For the $p-$th TSP, the prediction rule can be written as: predict class ”+1” if X[,tsp.list$I[p]] < X[,tsp.list$J[p]] and this forms the core of the predict function. The decision function for $p = 1$ in the above example is shown in Figure 1. Given a list of TSPs one has different choices on how to obtain the final predicted labels. Currently, Rgtsp proposes two means of combining the predictions of individual TSPs: either by majority voting or by weighting the votes with the correspoding scores - giving more weight to the TSPs with better scores. This functionality is available through the predict() generic function:

```
> yp = predict(tsp.list, X, combiner="majority")
> sum(yp != y.erpos) # count the errors
[1] 3
```

By inspecting the list of TSPs, it becomes clear that there are variables that are selected many times as having always either higher or lower value than all its pairing variables. We call such a structure a ***TSP hub*** and we can construct all the hubs larger than a specified size (25 pairs for example) using

```
> h = tsp.hub(tsp.list, min.hub.size=25)
> print(h)
Hub 1: 194 pairs
Center: 953 >
14 25 42 43 44 45 54 105 140 146 149 150 152 202 ...
```

This corresponds to a TSP hub in which the probeset colnames(X)[953] (205225_at, ESR1) has a higher expression than all other probesets in the list tsp.list. The TSP hubs can also be used in predicting the labels, through the same mechanism as above:

```
> yph = predict(h, X, combiner="majority")
> sum(yph != y.erpos) # no. of errors: 6
```

We see that in this particular case the prediction by TSP hubs is slightly less accurate than the combined predictions of the individual TSPs.

The generalization performance of the TSPs classifiers can be estimated by various methods. The Rgtsp package provides a function for k-fold cross–validation of the binary TSP classifiers (either tsp.n() or tsp.s() functions), cv.tsp(), which returns the training and validation performance of the classifier (it defaults to 5–fold cross–validation).

```
> r = cv.tsp(X, y.erpos)
> print(r)
$tr.m
 Error.rate Sensitivity Specificity         AUC
 0.02884615  0.97812500  0.96000000  0.96906250
```

In the case of a multi–class problem, we propose to use classification trees built on top of TSPs predictions. For $C > 2$ classes, one can train TSPs to solve each of the $C(C-1)/2$ pairwise binary classification problems (called one–versus–one (Hsu and Lin (2002)) or round robin (Fürnkranz (2002)) strategy) and then combine the predictions of the TSPs through a classification tree to predict the original classes. For more details the reader is referred to the package web page. This approach is implemented in the function mtsp() and makes use of the ctree() function in the party R package (y4 is an artificial 4–class label vector):

```
> m = mtsp(X, y4)
> yp = predict(m, X)
```

ACKNOWLEDGEMENT

Funding: VP and MD acknowledge the support of the Swiss National Science Foundation NCCR Molecular Oncology. EB acknowledges the support of Fondation Medic.

REFERENCES

Chapman, B., Jost, G., and van der Pas, R. (2007). *Using OpenMP*. The MIT Press.

Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, **2**, 721–747.

Geman, D., d'Avignon, C., Naiman, D. Q., and Winslow, R. L. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, **3**, Article19.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer–Verlag.

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N., and Pusztai, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol*, **24**(26), 4236–4244.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transaction on Neural Networks*, **13**(2), 415–425.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**(20), 3896–3904.

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., and Winslow, R. L. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **21**(20), 3905–3911.

9 Effect of training-sample size and classification difficulty on the accuracy of genomic predictors

- Breast Cancer Research, 12(1):R5
- IF: 6.345
- number of citations: 76
- personal contribution (60%): data processing, experimental design and implementation, statistical analysis of the results, manuscript writing

Popovici *et al. Breast Cancer Research* 2010, **12**:R5
http://breast-cancer-research.com/content/12/1/R5

Breast Cancer RESEARCH

RESEARCH ARTICLE Open Access

Effect of training-sample size and classification difficulty on the accuracy of genomic predictors

Vlad Popovici[1], Weijie Chen[2], Brandon G Gallas[2], Christos Hatzis[3], Weiwei Shi[4], Frank W Samuelson[2], Yuri Nikolsky[4], Marina Tsyganova[5], Alex Ishkin[5], Tatiana Nikolskaya[4,5], Kenneth R Hess[6], Vicente Valero[7], Daniel Booser[7], Mauro Delorenzi[1,8], Gabriel N Hortobagyi[7], Leming Shi[9], W Fraser Symmans[10], Lajos Pusztai[7*]

Abstract

Introduction: As part of the MicroArray Quality Control (MAQC)-II project, this analysis examines how the choice of univariate feature-selection methods and classification algorithms may influence the performance of genomic predictors under varying degrees of prediction difficulty represented by three clinically relevant endpoints.

Methods: We used gene-expression data from 230 breast cancers (grouped into training and independent validation sets), and we examined 40 predictors (five univariate feature-selection methods combined with eight different classifiers) for each of the three endpoints. Their classification performance was estimated on the training set by using two different resampling methods and compared with the accuracy observed in the independent validation set.

Results: A ranking of the three classification problems was obtained, and the performance of 120 models was estimated and assessed on an independent validation set. The bootstrapping estimates were closer to the validation performance than were the cross-validation estimates. The required sample size for each endpoint was estimated, and both gene-level and pathway-level analyses were performed on the obtained models.

Conclusions: We showed that genomic predictor accuracy is determined largely by an interplay between sample size and classification difficulty. Variations on univariate feature-selection methods and choice of classification algorithm have only a modest impact on predictor performance, and several statistically equally good predictors can be developed for any given classification problem.

Introduction

Gene-expression profiling with microarrays represents a novel tissue analytic tool that has been applied successfully to cancer classification, and the first generation of genomic prognostic signatures for breast cancer is already on the market [1-3]. So far, most of the published literature has addressed relatively simple classification problems, including separation of cancer from normal tissue, distinguishing between different types of cancers, or sorting cancers into good or bad prognoses [4]. The transcriptional differences between these conditions or disease states are often large compared with transcriptional variability within the groups, and therefore, reasonably successful classification is possible. The methodologic limitations and performance characteristics of gene expression based classifiers have not been examined systematically when applied to increasingly challenging classification problems in real clinical data sets.

The MicroArray Quality Control (MAQC) (MAQC Consortium project-II: a comprehensive study of common practices for the development and validation of microarray-based predictive models) breast cancer data set (Table 1) offers a unique opportunity to study the performance of genomic classifiers when applied across a range of classification difficulties.

One of the most important discoveries in breast cancer research in recent years has been the realization that estrogen receptor (ER)-positive and -negative breast cancers represent molecularly distinct diseases with large differences in gene-expression patterns [5,6]. Therefore,

* Correspondence: lpusztai@mdanderson.org
[7]Department of Breast Medical Oncology, P.O. Box 301439, Houston, TX 77230-1439, USA

Table 1 Patient characteristics in the training and validation sets

	Training set (*n* = 130)	Validation set (*n* = 100)	*P* value
Median age	51 years (28-79 years)	50 years (26-73 years)	
Race			0.804
Caucasian	85 (65%)	68 (68%)	
African American	13 (10%)	12 (12%)	
Asian	9 (7%)	7 (7%)	
Hispanic	21 (16%)	13 (13%)	
Mixed	2 (2%)	0	
Cancer histology			0.047
Invasive ductal (IDC)	119 (92%)	85 (85%)	
Mixed ductal/lobular (IDC/ILC)	8 (6%)	8 (8%)	
Invasive lobular (ILC)	1 (0.7%)	7 (7%)	
Others	2 (1.3%)	0	
Tumor size			0.643
T0	1 (1%)	2 (2%)	
T1	12 (9%)	8 (8%)	
T2	70 (54%)	62 (62%)	
T3	21 (16%)	13 (13%)	
T4	26 (20%)	15 (15%)	
Lymph node stage			0.935
N0	39 (30%)	27 (27%)	
N1	60 (46%)	47 (47%)	
N2	14 (11%)	13 (13%)	
N3	17 (13%)	13 (13%)	
Nuclear grade (BMN)			0.005
1	2 (2%)	11 (11%)	
2	52 (40%)	42 (42%)	
3	76 (58%)	47 (47%)	
Estrogen receptor			0.813
Estrogen receptor positive	80 (62%)	60 (60%)	
Estrogen receptor negative	50 (38%)	40 (40%)	
HER-2			< 0.001
HER-2 positive	33 (25%)	7 (7%)	
HER-2 negative	96 (74%)	93 (93%)	
Neoadjuvant therapy			0.005
Weekly T × 12 + FAC × 4	112 (86%)	98 (98%)	
3-Weekly T × 4 + FAC × 4	18 (14%)	2 (2%)	
Pathologic complete response (pCR)	33 (25%)	15 (15%)	0.055

Estrogen receptor: cases in which more than 10% of tumor cells stained positive for ER with immunohistochemistry (IHC) were considered positive. HER-2: cases that showed either 3+ IHC staining or had gene copy number greater than 2.0 were considered HER-2 "positive." T = paclitaxel; FAC = 5-fluorouracil, doxorubicin, and cyclophosphamide. The *P* values for the association tests were obtained from a χ^2 test unless the number of cases was fewer than five in any category, in which case, Fisher's Exact test was used.

gene expression-based prediction of ER status represents an easy classification problem.

A somewhat more difficult problem is to predict extreme chemotherapy sensitivity, including all breast cancers in the analysis. This classification problem is facilitated by the association between clinical disease characteristics and chemotherapy sensitivity. For example, ER-negative cancers are more chemotherapy sensitive than are ER-positive tumors [7].

A third, and more difficult, classification problem is to predict disease outcome in clinically and molecularly homogeneous patient populations. Genomic predictors could have the greatest clinical impact here, because traditional clinical variables alone are only weakly discriminatory of outcome in these populations. In the current data set, prediction of chemotherapy sensitivity among the ER-negative cancers represents such a challenge.

The goal of this analysis was to assess how the degree of classification difficulty may affect which elements of prediction methods perform better. We divided the data into a training set (n = 130) and a validation set (n = 100) and developed a series of classifiers to predict (a) ER status, (b) pathologic complete response (pCR) to preoperative chemotherapy for all breast cancers, and (c) pCR for ER-negative breast cancers. A predictor, or classifier, in this article is defined as a set of informative features (generated by a particular feature-selection method) and a trained discrimination rule (produced by applying a particular classification algorithm).

First, we examined whether the success of a predictor was influenced by a feature-selection method. We examined five different univariate feature-selection methods including three variations of a *t* test-based ranking and two methods that order features based on differences in expression values. It has been shown that several different classification algorithms can yield predictors with rather similar performance metrics [8-10]. However, it remains unknown whether the relative performances of different methods may vary depending on the difficulty of the prediction problem. We examined this question for eight different classifiers representing a broad range of algorithms, including linear (LDA), diagonal linear (DLDA), and quadratic discriminant analysis (QDA); logistic regression (LREG); and two versions of support-vector machines (SVM) and k-nearest neighbor (KNN) methods. Altogether, 40 different predictors were developed for each of the three classification problems (five different feature-selection methods × eight different classifiers). We also were interested determine to what extent the cross-validation classification performance is influenced by different data-resampling methods and the difficulty of the classification problem. We estimated the classification performance by using 10-times-repeated fivefold cross validation (10 × 5-CV) and leave-pair-out (LPO) bootstrapping [11] (a method that better accounts for training and testing variability). We calculated performance metrics for each of the 120 predictors (40 predictors × three endpoints) and compared the estimated accuracy in the training set with the observed accuracy in the independent validation set.

Materials and methods

Patients and materials

Gene-expression data from 230 stage I to III breast cancers, without individual patient identifiers, were provided to the MAQC project by the University of Texas M.D. Anderson Cancer Center (MDACC) Breast Cancer Pharmacogenomic Program. Gene-expression results were generated from fine-needle aspiration specimens of newly diagnosed breast cancers before any therapy. The biopsy specimens were collected sequentially during a prospective pharmacogenomic marker discovery study approved by the institutional review board between 2000 and 2008. These specimens represent 70% to 90% pure neoplastic cells with minimal stromal contamination [12]. All patients signed informed consent for genomic analysis of their cancers. Patients received 6 months of preoperative (neoadjuvant) chemotherapy including paclitaxel, 5-fluorouracil, cyclophosphamide, and doxorubicin, followed by surgical resection of the cancer. Response to preoperative chemotherapy was categorized as a pathologic complete response (pCR = no residual invasive cancer in the breast or lymph nodes) or residual invasive cancer (RD). The prognostic value of pCR has been discussed extensively in the medical literature [13]. Genomic analyses of subsets of this sequentially accrued patient population were reported previously [9,14,15]. For each endpoint, we used the first 130 cases as a training set to develop prediction models, and the next 100 cases were set aside as independent validation set. Table 1 and Additional file 1 show patient and sample characteristics in the two data sets.

Gene-expression profiling

Needle-aspiration specimens of the cancer were placed into RNAlater™ solution (Qiagen, Germantown, MD, USA) and stored at -80°C until further analysis. RNA extraction and gene-expression profiling were performed in multiple batches over time, as described previously [16,17] by using Affymetrix U133A (Affymetrix, Santa Clara, CA, USA) microarrays. Gene-expression data have been uploaded to the Gene Expression Omnibus website under the accession number GSE16716. Normalization was performed by using MAS 5.0 software (Affymetrix, Santa Clara, CA, USA) with default settings. Quality-control assessment of the hybridization results were performed with SimpleAffy software by Bioconductor; the percentage present call had to be more than 30%, scaling factor less than 3, and the 3'/5' ratios for β-actin less than 3, and for GAPDH, less than 1.3. These quality-control metrics are presented for each case in Additional file 2.

Ranking of classification problems by informative feature utility score

To assess the relative difficulty of the three classification problems that we selected to study, we adopted an approach similar to that described in [18]. This method defines the utility of a feature *i* as its Fisher score,

$$f_i = \frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{1i}^2 + \sigma_{2i}^2},$$

where μ_{1i} and μ_{2i} are the class means, and σ_{1i} and σ_{2i} are the class standard deviations for the feature *i*, respectively. If features are ordered $f_1 \geq f_2 \geq \ldots$ then, for each endpoint, the cumulative information is defined as

$$F_j = \sum_{i=1}^{j \leq N} f_i,$$

where N is the sample size. This cumulative information score assumes that the features are independent and that their effect on the classification performance is additive. This is rarely the case, as features are often correlated. Nonetheless, this cumulative information score is a simple and straightforward approach to estimate the relative difficulty of a classification problem early in the classifier-development process: an easier problem tends to have larger values for *F* than does a more difficult problem.

Feature-selection methods

No prefiltering of probe sets was done; all probe sets were considered by the feature-ranking methods that included (a) unequal variance *t* test (FS1); (b) unequal variance *t* test with filtering of probe sets that were correlated with one another (Pearson correlation > 0.75) to generate independently informative features (FS2); (c) instead of removing the correlated features, they were combined into meta-features by averaging them (FS3); and (d) we also ranked features according to their ratio of between- to within-group sum of squares (FS4) and (e) according to the absolute differences in the class means (FS5).

Classification algorithms

We examined eight classifiers in combination with the previously mentioned feature-selection methods, including linear discriminant analysis (LDA), diagonal linear discriminant analysis (DLDA), quadratic discriminant analysis (QDA), logistic regression (LREG), two k nearest neighbors classifiers with k = 3 (KNN3) and k = 11 (KNN11), and support vector machines with a radial basis function kernel with two different values for the kernel parameter: $\gamma = 0.5$ (SVM05) and $\gamma = 2.0$ (SVM2), respectively. Overall, 40 models were developed for each of the three prediction problems.

Estimation of predictive performance

Leave-N-out cross-validation and other resampling methods of the training set are often used to select a final predictor for independent validation. Therefore, it is important to understand how resampling-based predictive performance correlates with predictive performance on independent validation cases. To study this question, we used a nested two-level cross-validation scheme, in which the cross-validation in the outer loop had the role of estimating the performance of the whole modeling procedure, whereas the cross-validation in the inner loop was used for selecting the optimal number of features [19].

The procedure in the inner loop is as follows. For each combination of a feature-selection method F and a classification algorithm C, the number of features j (F, C) in the model was considered as a free-parameter (within a predefined set of allowable values) and was optimized. In the inner loop, a repeated (5 times), stratified (to preserve the proportion of the two classes in all training and testing splits), fivefold cross-validation was used to define the number of features that maximized the AUC. A ranking of the features was first obtained by applying *F* on the reduced internal training set (obtained by leaving aside one fold from the current training set). Then the classifier *C* was trained on the same set, but considering only the top j(*F, C*) features. The predictions on the internal testing set (the left-out fold) were recorded, and the procedure was repeated. At the end, an estimation of the AUC was obtained, corresponding to the given combination of *F, C*, and j(*F, C*). The procedure was repeated with different folds, and an average estimate of the AUC was obtained for each *F, C*, and j(*F, C*). The optimal number of features j*(*F, C*) was selected as the value j (*F, C*) yielding the highest average AUC. The number of features allowed for each model was chosen *a priori*, to avoid overfitting of models and to limit the computation time. For the prediction of ER status, the feature size was chosen to contain all values between 2 and 15, whereas for both pCR endpoints, it was {2,5,8,...,41}; 41 being almost half the size of the smallest training set (*n* = 85 ER-negative cancer). For a pseudo-code that details the schema used for cross-validation [see Additional file 3]. To avoid adding variability due to random partitioning the data into folds, all estimates were obtained on the same splits of the data.

We investigated two methods in the outer loop. The first method is a stratified 10-times-repeated fivefold cross-validation (10 × 5-CV). In each of the five cross-validation iterations, 80% of the data were first used as input to the inner loop procedure for feature selection and training the classifier with the selected features, and finally, the remaining 20% of the data were used to test the classifier. The 95% CI for the area under the receiver operating characteristics curve (AUC) was approximated by [AUC - 1.96 SEM, AUC + 1.96 SEM]. The SEM was estimated by averaging the 10 estimates of the standard error of the mean obtained from the five different estimates of the AUC produced by the 5-CV.

The second method in the outer loop is a bootstrap-based method, also known as a smoothed version of cross-validation [20]. Efron and Tibshirani [20] proposed the leave-one-out bootstrap method on the performance metric error rate, and their technique was recently extended by Yousef and colleagues [11] to the performance metric AUC. This method uses a leave-pair-out (LPO) bootstrap approach to estimate the mean AUC (mean over training sets) and a "delta method after bootstrap" to estimate the variability of the estimated mean AUC. We point out that this variability captures both the effect of finite training-set size and the effect of finite testing-set size. In the LPO approach, multiple (n = 5,000) training sets are obtained by stratified bootstrap resampling, and each training set is used as input to the inner-loop procedure for feature selection and training the classifier with the selected features. In testing, any pair of cases (one from the positive class and one from the negative class) is tested on the classifiers trained on the bootstrap samples that do not contain the two held-out cases. The Wilcoxon-Mann-Whitney statistic of the prediction results on pairs of cases is averaged over all bootstrap-training sets and is used to estimate the mean AUC. An advantage of this technique is that it allows estimating the variability of the AUC estimator by using the influence function method [11,20]. By assuming that the estimated AUC is asymptotically normal, the 95% CI of the AUC can be approximated by [AUC - 1.96 SEM; AUC + 1.96 SEM].

The estimated performance and the associated CIs from the training and internal-assessment process are compared with the independent validation performance. The conditional validation performance was obtained by selecting features and training the classifier with the training data set and testing on the validation data set. This performance is conditional on the particular finite training set and may vary when the training set varies. Therefore, we estimated the mean of this conditional performance where the mean is over multiple training sets and obtained by bootstrapping the training set multiple times and averaging the conditional AUCs, as tested on the validation set [21].

We also estimated the variability of the conditional validation performance and decomposed the variance into two components: the variability due to the finite size of the training set and the variability due to the finite size of the test set [21]. The training variability reflects the stability of the classifier performance when the training set varies, and the testing variability reflects the expected performance variation for different test sets.

To compare the ability of the performance estimates of 10 × 5-CV and the LPO bootstrap to predict the performance on the independent set, we used a root mean square error (RMSE) measure, which is defined as

$$RMSE = \sqrt{\frac{1}{40}\sum_{F=1}^{5}\sum_{C=1}^{8}(\bar{A}_{F,C}^{\mathrm{int}ernal} - \bar{A}_{F,C}^{independent})^2},$$

where F and C index feature selection and classifier, respectively, $\bar{A}$ denotes the mean AUC; the superscript "internal" can be "10 × 5-CV" or "LPO bootstrap."

Estimation of predictor learning over increasing training-set size

Predictor learning was evaluated for the models that performed nominally the best in independent validation for each of the three prediction problems. All 230 cases were included in the analysis to fit learning curves to these three models. For the ER-status endpoint, 10 different training-sample sizes, ranging from n = 60 to n = 220 by increments of 20, were used to estimate the dependence of the performance parameters on the sample size. For each sample size, 10 different random samples were drawn from the full set by stratified sampling, and fivefold cross-validation was used to assess the error rate and AUC of the models where all the parameters of the models were recalculated. A similar approach was taken for the pCR (n = 50, 70, ..., 210) and "pCR in ER-negative cancer" predictors (n = 25, 40, ..., 85). By following the work of Fukunaga [22], the following learning-curve model was fit to the resulting AUC: $Y = a + b/TrainingSize$.

Congruence of different predictors at gene and functional pathway level

We were interested in examining the congruence of informative features that were selected by different methods for the same prediction endpoint and also for different endpoints. Both gene-level and pathway-level analyses were performed as described previously [23]. MetaCore protein-function classification was used to group genes into protein functions, and GeneGo Pathway maps were used for functional classification of predictive features. We assessed congruency by using the kappa statistics. The input for kappa involves "learners" that classify a set of objects into categories. We considered each feature-selection method as a learner and each probe set as an object. The probe sets used in this analysis are presented in Additional file 4. Each probe set from the rank-ordered lists is categorized by each feature-selection method either as 1 (that is, selected as informative) or 0 (that is, nonselected). By using such an 0/1 matrix for all probe sets × all feature-selection methods for every prediction endpoint as input, we can calculate Cohen's kappa function for the congruency. For pathway-level analysis, we mapped the probe sets to pathway lists by using hypergeometric enrichment analysis. The pathways are ranked by enrichment *P* values, and the top n pathways (n equals the number of genes in the input list for comparison and consistency between the two levels) were selected for presentation.

All statistical analysis was performed by using R software.

Results

Difficulty of the classification problems

Three distinct classification problems were studied: (a) ER-status prediction, including 80 ER-positive (62%) and 50 ER-negative training cases (38%); (b) pCR prediction, including 33 cases with pCR (25%) and 97 cases with residual cancer (75%) for training; and (c) pCR prediction for ER-negative cancers, including 27 training cases with pCR (54%) and 23 with residual cancer (46%). Figure 1 shows the cumulative information scores for the three endpoints: larger cumulative information is an indicator for a simpler classification problem. The obtained ranking implies that the three endpoints represent different degrees of classification difficulty.

We also assessed the significance of the utility scores by using permutation tests (10,000 permutations) for computing the raw *P* values, followed by Benjamini-Hochberg correction for multiple testing. For the ER-status endpoint, 1,502 features with significant utility scores (*P* value < 0.0001) were used, whereas for the pCR (all cases), 252 significant features and only five features (corresponding to A2M [HGNC:7], RNMT [HGNC:10075], KIAA0460 [HGNC:29039], AHNAK [HGNC:347], and ACSM1 [HGNC:18049] genes) were used for pCR among ER-negative cancers.

Effect of feature-selection methods and classification algorithms on cross-validation performance

Figure 2 illustrates the average cross-validation AUC estimated by 10 × 5-CV for all predictors, stratified by feature-selection method (left column). All feature-selection methods performed similarly in combination with various classification algorithms for a given endpoint. The two non-*t* test-based methods, FS4 and FS5, showed slightly better performances than did *t* test-based feature selection for the most difficult prediction

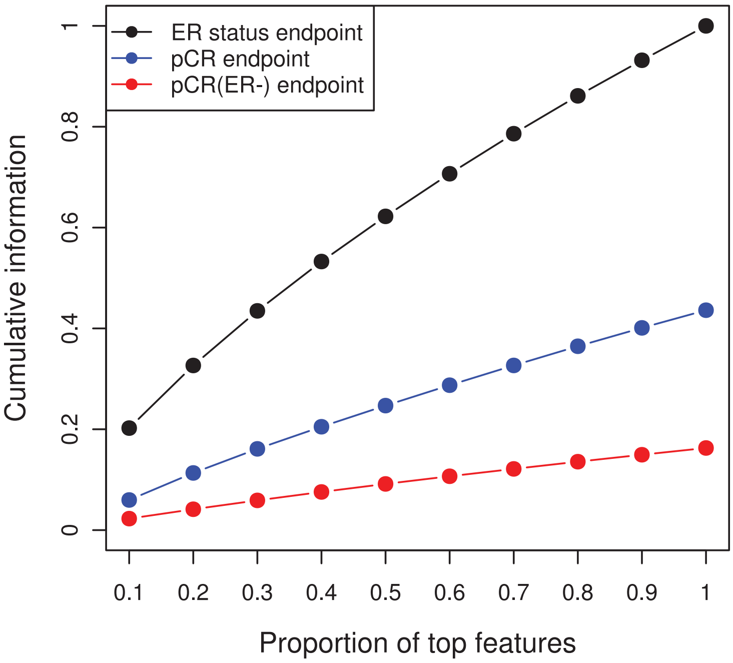


Figure 1 Relative complexity of the three prediction problems. The cumulative information values have been scaled such that the maximum value is 1. To make the curves comparable and to take into account the sample size, the ratio between the number of features used in the cumulative information (F) and the sample size is used on the horizontal axis. Larger values of the cumulative information indicate simpler problems.

endpoint "pCR on ER-negative cancers" in cross validation, but confidence intervals widely overlapped. Additional file 5 shows the average error rates and AUCs generated from 10 × 5-CV for each prediction model applied to all three classification problems, along with the average number of features selected. Interestingly, the number of selected features did not increase as the prediction problem became more difficult. For the most difficult problem, the number of selected features was lower than that for the moderately difficult problem. This is probably because of the lack of informative features: as the classification problem becomes more difficult, fewer features are informative for the outcome (also see Figure 1).

Figure 2 also shows the variability of the classification error rates and AUC estimated through 10 × 5-CV for all predictors, stratified by classification algorithm (right column). All methods performed similarly. The prediction endpoint (that is, classification difficulty) had the greatest effect on the cross-validation AUC. The effects of feature-selection method and choice of classifier algorithm were modest.

Bootstrap and independent-validation results

Figure 3 shows the estimated AUCs obtained with 10 × 5-CV (black square), LPO bootstrap (black circle), and the conditional AUC (blue circle) on the independent validation set and its variability (blue error bar representing ± 2 SD) and mean (red cross). Additional file 5 includes the internal (10 × 5-CV and LPO bootstrap) and independent validation-performance metrics for each predictor. Both internal-estimation methods yielded AUCs that were very close, well within 2 standard deviations of the mean, to the conditional and mean AUCs observed in the independent validation. Internal-performance estimates generated within the training set only slightly overestimated the performance relative to independent validation, indicating both that the modeling approach was correct and that no strong batch effect occurred between training and validation sets. Simpler linear methods, such as LREG, LDA, and DLDA, performed generally well in both internal and independent validation, and these methods were among the top five nominally best-performing models for all prediction endpoints [see Additional file 5]. The non-*t* test-based feature-selection methods (FS4, FS5) that showed good results in cross validation also performed well in independent validation and were included in four of the top five models for each endpoint. However, the 95% CIs of the point estimates overlap broadly for all predictors, and no single strategy emerged as clearly superior for any particular endpoint.

To assess the confidence-interval estimation, we calculated the RMSE for the AUC estimates obtained with 10 × 5-CV and LPO bootstrap for all the three endpoints. Leave-pair-out bootstrap performed better than 10 × 5-CV in terms of the agreement with the mean AUC estimated in the independent-validation set: RMSEs for LPO bootstrap were 0.0484, 0.0491, and 0.357 in comparison with 0.0562, 0.0713, and 0.449 for 10 × 5-CV for the ER status, pCR, and pCR within ER-negative endpoints, respectively.

Figure 3 clearly shows that the variability of the estimated classification performance increases as the level of classification difficulty increases. This implies that, to achieve the same level of statistical precision of the estimated performance, more cases are needed for a more-difficult endpoint. Figure 3 also shows both the conditional (blue circle) and mean validation AUCs (red cross). The larger the difference between the conditional validation AUC and the mean validation AUC, the less stable the predictor is with respect to varying the training sets. A quantitative measure of classifier stability is the training variability, and we have decomposed the variability of the conditional validation AUC shown in Figure 3 into two components (training variability and testing variability) and put the results in Additional file 5.

Predictor-performance and sample-size estimations through learning curves

To estimate the training-set size that is necessary to develop predictors that operate near their respective plateaus, we examined how the performance characteristics of each of the nominally best predictors for each endpoint improved as the training-set size increased. For ER-status prediction, we selected QDA with FS1 (conditional validation AUC = 0.939); for pCR prediction including both the ER-positive and -negative cancers, we selected LREG with FS5 (conditional validation AUC = 0.805); and for pCR in ER-negative cancers, we selected LREG with FS4 (conditional validation AUC = 0.627). Figure 4 shows the observed changes in average AUCs for each of the classifiers as the training-set size increased from 60 to 220 (or from 25 to 85 for pCR prediction in ER-negative cancers) and the projected improvements for assumed larger training sets. The results indicate that for the easiest problem (ER), the predictor seems to perform at its best with a sample size around 80 to 100. For the moderately difficult problem (pCR), the steady increase of the learning curve suggests that the performance of the model can be improved by increasing the sample size, beyond the highest value currently tested (220). For the pCR in ER-negative cancer endpoint, the learning curves manifested a very modest and gradual improvement in performance between training sample sizes of 25 and 85, suggesting that either too few samples were available for a reliable

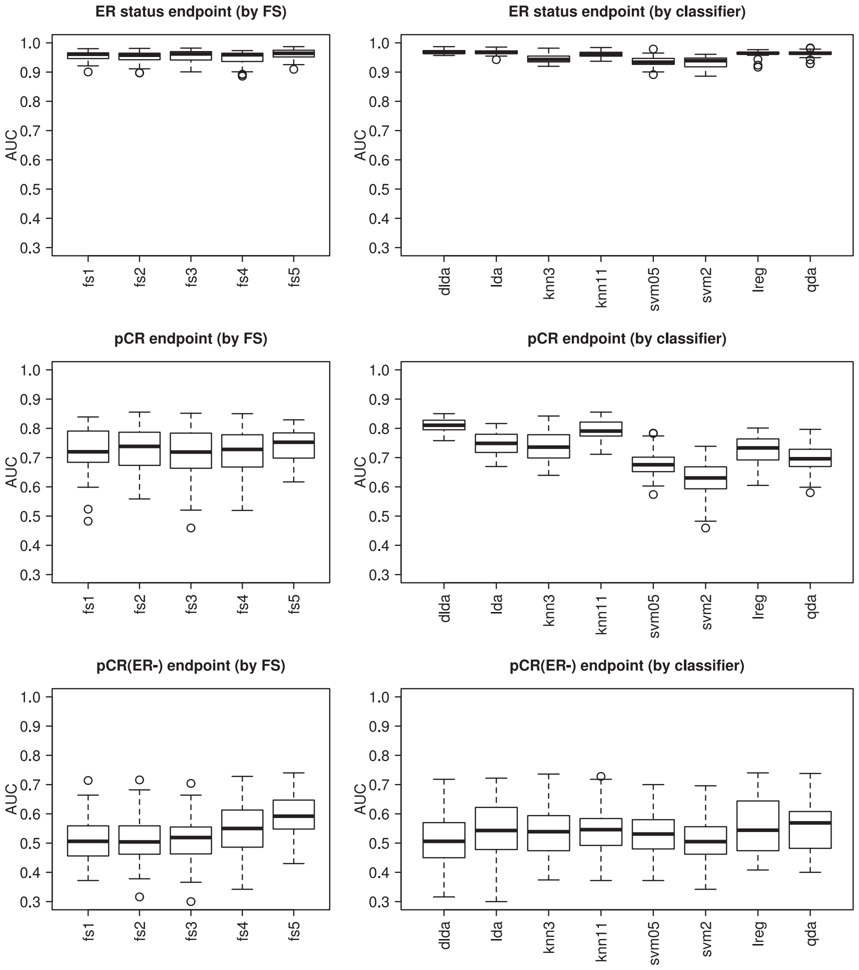


Figure 2 Boxplots of the estimated area under the curve (AUC), stratified by feature-selection and classification methods. The boxplots show the mean AUC in 10 times fivefold cross validation (CV). The left column contains the estimated AUC stratified by the feature-selection method, and the right column contains the estimated AUC stratified by the classification method.

estimation of the learning curve or that limited information in the mRNA space exists to predict this particular outcome with the methods applied in this analysis. The learning curve that had a slope significantly different from 0 was the one for the pCR endpoint (P = 0.001; ER endpoint, P = 0.05; pCR in ER-negative endpoint, P = 0.365).

Functional analysis of predictive features

Our results demonstrate that several different feature sets can yield predictors with statistically similar performances [8-10,24]. This may occur because the various probe sets that represent different genes capture information from the same complex molecular pathways that determine a particular clinical outcome [25]. In other words, different features measure different components of the same informative biologic pathway. To test this concept, we mapped each of the 15 feature sets used in the final validation models to known biologic pathways. The different feature sets selected for a particular prediction endpoint had a high level of congruency at both the gene and the pathway levels across all the

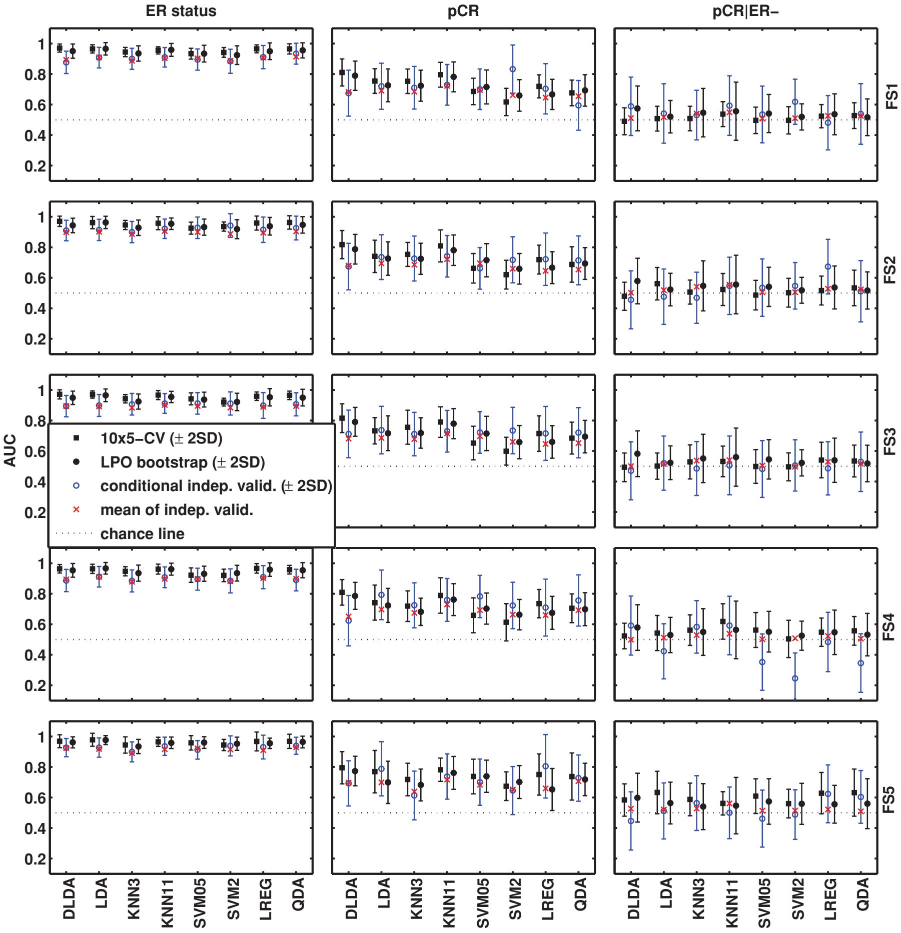


Figure 3 Graphic summaries of the estimated and observed areas under the curve (AUCs) for each of the 120 models. For each combination of feature-selection method and classification algorithm, the AUCs ± 2 standard deviations are plotted. Mean AUCs obtained from 10 × 5-CV (cross-validation; black square), LPO bootstrap (black dot), and the conditional (blue circle) and mean (red cross) validation AUCs are shown.

five different ranking methods (Table 2). The selected gene sets and pathways were also rather similar to each other for the ER and pCR prediction endpoints. However, the genes and pathways predictive of pCR in ER-negative cancers were very different from the other two informative gene sets.

Additional file 6 contains the pathway-enrichment tables for the three endpoints, including pathways with enrichment *P* values < 0.1. Thirty-two pathways contributed to the prediction of ER status; 36, to pCR prediction; and 11, to pCR prediction within ER-negative cancers across the five feature-selection methods. For the ER endpoint, development, cell adhesion, cytoskeleton remodeling, DNA damage, apoptosis, and ER transcription factor activity were the most significant pathway elements common to all informative feature sets. We also noted that most pathways that were involved in pCR prediction (31 of 36) were the same as those involved in ER-status prediction. This is consistent with the known association between pCR rate and ER status [7]. Estrogen receptor-negative cancers had significantly higher pCR rates than ER-positive cancers

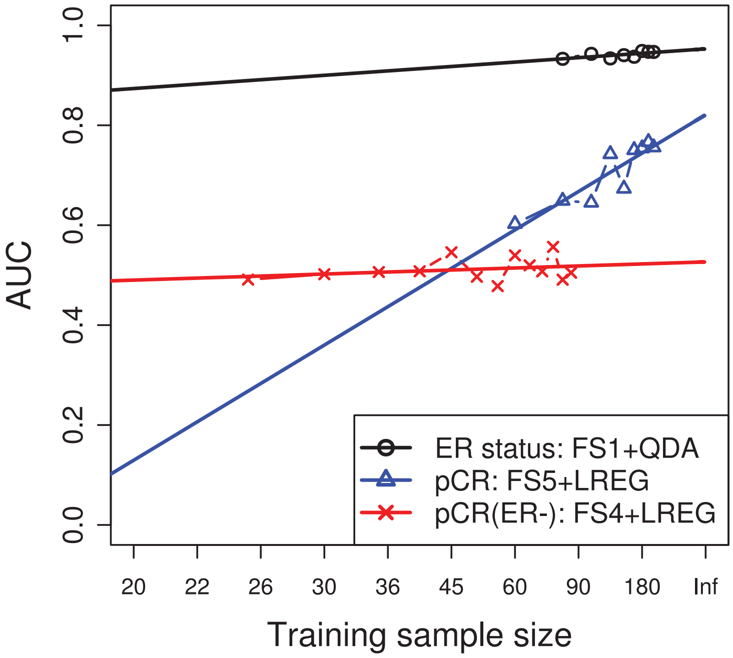


Figure 4 Learning curves for the best predictors for each of the three endpoints. For each endpoint, the learning curve of the best-performing model on the validation set was estimated by fivefold cross-validation for gradually increasing sample sizes. The plot shows both the estimated performance for different sample sizes and the fitted curve. The quadratic discriminant analysis (QDA) classifier required more than 60 samples, so the minimum sample size for it was 80. Note the nonlinear scale of the x-axis.

(54% pCR in ER-negative cancers versus 7.5% pCR in ER-positive cancers; χ^2 test *P* value = 1.068e-08). The pathways that were selected for prediction of pCR in ER-negative cancers were distinct from the pathways that were predictive of pCR in all patients and included immune response-related pathways (IL-2 and T-helper cell activation), opioid-receptor signaling, and endothelial cell-related pathways.

Discussion

The goal of this analysis was to examine how the choice of a univariate feature-selection method and classification algorithm may influence the performance of predictors under varying degrees of classification difficulty. We examined the influence of changing two critical components, feature selection and classification algorithm in the predictor development process, for three different prediction problems that represented three levels of difficulty in a clinically annotated human breast cancer data set. Classification of breast cancer into ER-positive or -negative categories is an easy classification problem; the large number of informative probe sets and high information content of the features allow clear separation of the groups. The AUC values for the 40 different prediction models for this endpoint ranged from 0.875 to 0.939 in the independent validation set. Prediction of pCR across all breast cancers, including both ER-negative and ER-positive cases, represented a slightly more difficult prediction problem with AUCs ranging between 0.61 and 0.80 in the validation set. Prediction of pCR in the molecularly more homogeneous ER-negative breast cancer subpopulaton proved to be the most difficult classification challenge: the validation AUCs ranged from 0.34 to 0.62. No predictor-development strategy emerged as clearly superior for any of the classification problems. The 95% CI of the prediction accuracies overlaped broadly for most of the predictors. However, LDA, DLDA, LREG, and QDA classification algorithms were consistently among the best-performing models for each problem. Interestingly, KNN3 and SVM methods were often among the worst-performing models in independent validation, even though these reached relatively high AUC values in cross validation. It is possible that further fine tuning of parameters for these

Table 2 Congruencies across different endpoints and different feature-selection methods

Same endpoint but different feature selection (FS)		
Endpoint	**Gene-level**	**Level of canonic-pathway maps**
ER status	0.541	0.573
pCR	0.544	0.572
pCR(ER$^-$)	0.593	0.532
Same FS but different endpoints		
FS	Gene-level	Level of canonic-pathway maps
FS1	0.300	0.290
FS2	0.299	0.274
FS3	0.291	0.278
FS4	0.295	0.291
FS5	0.272	0.282

The table shows that kappa statistics (that is, congruency) are high for different feature-selection methods for the same endpoint but are low for the same feature-ranking method for different endpoints. Both gene-level and pathway-level analyses show similar results.

more-complex classifiers (in the sense of an implementable decision boundary) could have improved predictive performance. We examined only the radial function kernel for SVM with two *a priori* set kernel parameters γ = 0.5 and 2.0, and the parameter C (cost of misclassification) was also fixed at 10. Fixing these parameters may have resulted in "less than optimally trained" models that could lead to added variability in the performance of the classifiers. Also, we examined only two versions of KNN with *a priori* set k of 3 and 11, and found that KNN11 outperformed KNN3. Low values of k yield local classifiers with low bias but high variance, whereas higher values led to more-global classifiers with higher bias and lower variance; exploring a broader range of k values could have optimized prediction results. Optimizing the parameters γ or k is not a straightforward task. It should be done within the inner cross-validation process, just as is done with feature selection. Fine tuning different model parameters outside of the two-stage cross-validation process would lead to model-selection bias, or optimization bias [19].

An interesting observation was that simple feature-selection methods that ranked features based on difference in means performed very well in both cross-validation and independent validation relative to the more commonly used t statistic-based ranking. Four of the top five models for each prediction problem used features selected by the non-t test-based methods. However, it is important to recognize that all of the feature-selection methods that we examined represented univariate filtering approaches that rank features individually and independent of the classification method. It is possible that nonparametric or multivariate feature-selection methods could yield different results. Penalized feature-selection methods, which embed feature selection in the classifier fitting step, may also have advantages, because features that might not be discriminatory individually could be jointly predictive in combination with other features. At least one article suggested that multivariate sparse penalized likelihood methods, including lasso and elastic net, might have a slight edge compared with univariate filtering [26]. Other publications that compared several univariate and multivariate feature-selection methods in public cancer data sets by using 10-fold cross-validation estimates found that simple univariate feature-selection methods often outperformed more-complex multivariate approaches [27,28].

Our data demonstrate that many different feature sets and classification methods can yield similarly accurate predictors for a given endpoint. When we mapped the feature sets generated by five different univariate feature-selection methods to biologic pathways, each method tended to identify similar genes and pathways. The biologic pathways that were implicated in ER-status or pCR prediction were distinct from the pathways that were predictive of pCR in ER-negative cancers. This pathway-level analysis is hypothesis generating and will require further laboratory validation to determine the importance of the identified pathways (for example, immune response, endothelial-cell regulation, G-protein signaling) in the biology of chemotherapy response in ER-negative breast cancer.

To estimate potential improvements in predictive performance of the nominally best predictors for each classification problem, we pooled all cases and carried out a series of split-sample training and validation analyses in which the predictors were trained on increasingly larger data sets. For the easy classification problem (ER-status), relatively small sample sizes (80 to 100 samples) were enough for constructing excellent predictors. In contrast, for the moderately difficult problem (pCR prediction), the accuracy of the model steadily improved as the sample size increased. For the most difficult problem, pCR prediction in ER-negative cancer, a minimal improvement was observed over a range of 25 to 85 training cases. It is important to note that the pCR and ER status predictors trained on 80 cases showed good or excellent conditional AUCs (0.65 and 0.94, respectively). This modest performance and limited improvement of the pCR predictor for ER-negative cancer may be due to (a) too small sample size for trainig or (b) the incompleteness of the mRNA expression-based feature space, meaning that this class-separation problem cannot be fully accomplished by using information only from the available probes by using the methods that we applied. However, fitting learning curves to preliminary data sets could assisst investigators in estimating sample-size requirements for a particular prediction problem for any given model.

Conclusions

This analysis confirms that it is possible to build multigene classifiers of clinical outcome that hold up in independent validation. Predictor performance is determined largely by an interplay between training-sample size and classification difficulty. Variations on univariate feature-selection methods and choice of classification algorithm had only a modest impact on predictor performance, and it is clear that within our statistical precision, several equally good predictors can be developed for each of our classification problems. Pathway-level analysis of informative features selected by different methods revealed a high level of congruency. This indicates that similar biologic pathways were identified as informative for a given prediction endpoint by the different univariate feature-selection methods. The independent validation results also showed that internal 10 × 5-CV and LPO bootstrap both yielded reasonably good and only slightly optimistic performance estimates for all the endpoints.

Additional file 1: Supplemental Table S1. Clinical data for all the patients in the training and validation sets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S1.xls]

Additional file 2: Supplemental Table S2. Quality control results.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S2.xls]

Additional file 3: Supplemental Table S3. Pathways mapping for all endpoints.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S3.xls]

Additional file 4: Supplemental methods. Pseudo-code description of the two-level external cross-validation scheme.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S4.pdf]

Additional file 5: Supplemental Table S4. Features (probesets) selected in the 120 models.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S5.xls]

Additional file 6: Supplemental Table S5. Estimated and validation performance of all models.
Click here for file
[http://www.biomedcentral.com/content/supplementary/bcr2468-S6.xls]

Abbreviations

10 × 5-CV: repeated (10 times) fivefold cross validation; AUC: area under the receiver operating characteristic curve; CI: confidence interval; DLDA: diagonal linear discriminant analysis; ER: estrogen receptor; KNN: k nearest-neighbors classifier; LDA: linear discriminant analysis; LPO: leave-pair-out bootstrap; LREG: logistic regression classifier; pCR: pathologic complete response; QDA: quadratic discriminant analysis; RD: residual invasive cancer; RMSE: root mean square error; SD: standard deviation; SEM: standard error of the mean; SVM: support vector machine.

Acknowledgements

This research was supported by grants from the NCI R-01 program (LP), The Breast Cancer Research Foundation (LP and WFS), The MD Anderson Cancer Center Faculty Incentive Funds (WFS), and the Commonwealth Cancer Fundation (LP, WFS). VP and MD acknowledge the support of the Swiss National Science Foundation NCCR Molecular Oncology. Certain commercial materials and equipment are identified to specify experimental procedures adequately. In no case does such identification imply recommendation or endorsement by the FDA, nor does it imply that the items identified are necessarily the best available for the purpose. The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.

Author details

[1]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Génopode Building, Quartier Sorge, Lausanne CH-1015, Switzerland. [2]Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Ave WO62-3124, Silver Springs, MD 20993-0002, USA. [3]Nuvera Biosciences, 400 West Cummings Park, Woburn, MA 01801, USA. [4]GeneGo, Inc., 500 Renaissance Drive, St. Joseph, MI 49085, USA. [5]Department of Systems Biology, Vavilov Institute for General Genetics, Russian Academy of Sciences, Gubkina str. 3 korp. 1, Moscow 119333, Russia. [6]Department of Biostatistics, P.O. Box 301439, Houston, TX 77230-1439, USA. [7]Department of Breast Medical Oncology, P.O. Box 301439, Houston, TX 77230-1439, USA. [8]Swiss NCCR Molecular Oncology, Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [9]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079, USA. [10]Department of Pathology of the University of Texas M. D. Anderson Cancer Center, P.O. Box 301439, Houston, TX 77230-1439, USA.

Authors' contributions

LP, VP, and LS designed the study. VP, WFS, and WC performed the experiments. VP, WC, BG, CH, WS, FS, YN, MT, AI, TN, KH, MD, and LP performed the statistical analyses and interpreted the results. VV, DB, GH, WFS, and LP contributed the clinical, pathologic, and molecular data. All authors contributed to the writing of the manuscript and read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 14 October 2009 Revised: 18 December 2009
Accepted: 11 January 2010 Published: 11 January 2010

References

1. Vijver van de MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Velde van der T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347**:1999-2009.
2. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351**:2817-2826.
3. Ross JS, Hatzis C, Symmans WF, Pusztai L, Hortobágyi GN: **Commercialized multigene predictors of clinical outcome for breast cancer.** *Oncologist* 2008, **13**:477-493.
4. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Statist Assoc* 2002, **97**:77-87.
5. Perou CM, Sørlie T, Eisen MB, Rijn van de M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
6. Pusztai L, Ayers M, Stec J, Clark E, Hess K, Stivers D, Damokosh A, Sneige N, Buchholz TA, Esteva FJ, Arun B, Cristofanilli M, Booser D, Rosales M, Valero V, Adams C, Hortobagyi GN, Symmans WF: **Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences**

between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res* 2003, **9**:2406-2415.

7. Andre F, Mazouni C, Liedtke C, Kau S-W, Frye D, Green M, Gonzalez-Angulo AM, Symmans WF, Hortobagyi GN, Pusztai L: **HER2 expression and efficacy of preoperative paclitaxel/FAC chemotherapy in breast cancer.** *Breast Cancer Res Treat* 2008, **108**:183-190.
8. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?.** *Bioinformatics* 2005, **21**:171-178.
9. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.** *J Clin Oncol* 2006, **24**:4236-4244.
10. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao M-S, Penn LZ, Jurisica I: **Prognostic gene signatures for non-small-cell lung cancer.** *Proc Natl Acad Sci USA* 2009, **106**:2824-2828.
11. Yousef WA, Wagner RF, Loew MH: **Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier.** *Pattern Recog Lett* 2005, **26**:2600-2610.
12. Symmans WF, Ayers M, Clark EA, Stec J, Hess KR, Sneige N, Buchholz TA, Krishnamurthy S, Ibrahim NK, Buzdar AU, Theriault RL, Rosales MFM, Thomas ES, Gwyn KM, Green MC, Syed AR, Hortobagyi GN, Pusztai L: **Total RNA yield and microarray gene expression profiles from fine-needle aspiration biopsy and core-needle biopsy samples of breast carcinoma.** *Cancer* 2003, **97**:2960-2971.
13. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M, Cristofanilli M, Hortobagyi GN, Pusztai L: **Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.** *J Clin Oncol* 2008, **26**:1275-1281.
14. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, Lecocke M, Metivier J, Booser D, Ibrahim N, Valero V, Royce M, Arun B, Whitman G, Ross J, Sneige N, Hortobagyi GN, Pusztai L: **Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer.** *J Clin Oncol* 2004, **22**:2284-2293.
15. Peintinger F, Anderson K, Mazouni C, Kuerer HM, Hatzis C, Lin F, Hortobagyi GN, Symmans WF, Pusztai L: **Thirty-gene pharmacogenomic test correlates with residual cancer burden after preoperative chemotherapy for breast cancer.** *Clin Cancer Res* 2007, **13**:4078-4082.
16. Stec J, Wang J, Coombes K, Ayers M, Hoersch S, Gold DL, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Symmans WF, Pusztai L: **Comparison of the predictive accuracy of DNA array-based multigene classifiers across cDNA arrays and Affymetrix genechips.** *J Mol Diagn* 2005, **7**:357-367.
17. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN, Pusztai L: **Breast cancer molecular subtypes respond differently to preoperative chemotherapy.** *Clin Cancer Res* 2005, **11**:5678-5685.
18. Ho TK, Basu M: **Complexity measures of supervised classification problems.** *IEEE Trans Patt Anal Mach Intel* 2002, **24**:289-300.
19. Wood IA, Visscher PM, Mengersen KL: **Classification based upon gene expression data: bias and precision of error rates.** *Bioinformatics* 2007, **23**:1363-1370.
20. Efron B, Tibshirani R: **Improvements on cross-validation: the 632+ bootstrap method.** *J Am Statist Assoc* 1997, **92**:548-560.
21. Yousef WA, Wagner RF, Loew MH: **Assessing classifiers from two independent data sets using ROC analysis: a nonparametric approach.** *IEEE Trans Patt Anal Mach Intel* 2006, **28**:1809-1817.
22. Fukunaga K, Hayes RR: **Effects of sample size in classifier design.** *IEEE Trans Patt Anal Mach Intel* 1989, **11**:873-885.
23. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biol* 2007, **8**:R183.
24. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, van't Veer LJ, Perou CM: **Concordance among gene-expression-based predictors for breast cancer.** *N Engl J Med* 2006, **355**:560-569.
25. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schütz F, Goldstein DR, Piccart M, Delorenzi M: **Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.** *Breast Cancer Res* 2008, **10**:R65.
26. Zucknick M, Richardson S, Stronach EA: **Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods.** *Stat Appl Genet Mol Biol* 2008, **7**:Article7.
27. Lai C, Reinders MJT, van't Veer LJ, Wessels LFA: **A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets.** *BMC Bioinformatics* 2006, **7**:235.
28. Lecocke M, Hess KR: **An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data.** *Cancer Inform* 2007, **2**:313-327.

doi:10.1186/bcr2468
Cite this article as: Popovici *et al.*: Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Research* 2010 **12**:R5.

10 The MicroArray Quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

- Nature Biotechnology, 28(8):827–U109
- IF: 41.667
- number of citations: 339
- personal contribution (5%): responsible for SIB participation – experimental design and implementation, manuscript writing

ARTICLES

nature biotechnology

The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

MAQC Consortium*

Gene expression data from microarrays are being applied to predict preclinical and clinical endpoints, but the reliability of these predictions has not been established. In the MAQC-II project, 36 independent teams analyzed six microarray data sets to generate predictive models for classifying a sample with respect to one of 13 endpoints indicative of lung or liver toxicity in rodents, or of breast cancer, multiple myeloma or neuroblastoma in humans. In total, >30,000 models were built using many combinations of analytical methods. The teams generated predictive models without knowing the biological meaning of some of the endpoints and, to mimic clinical reality, tested the models on data that had not been used for training. We found that model performance depended largely on the endpoint and team proficiency and that different approaches generated models of similar performance. The conclusions and recommendations from MAQC-II should be useful for regulatory agencies, study committees and independent investigators that evaluate methods for global gene expression analysis.

As part of the United States Food and Drug Administration's (FDA's) Critical Path Initiative to medical product development (http://www.fda.gov/oc/initiatives/criticalpath/), the MAQC consortium began in February 2005 with the goal of addressing various microarray reliability concerns raised in publications[1–9] pertaining to reproducibility of gene signatures. The first phase of this project (MAQC-I) extensively evaluated the technical performance of microarray platforms in identifying all differentially expressed genes that would potentially constitute biomarkers. The MAQC-I found high intra-platform reproducibility across test sites, as well as inter-platform concordance of differentially expressed gene lists[10–15] and confirmed that microarray technology is able to reliably identify differentially expressed genes between sample classes or populations[16,17]. Importantly, the MAQC-I helped produce companion guidance regarding genomic data submission to the FDA (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm079855.pdf).

Although the MAQC-I focused on the technical aspects of gene expression measurements, robust technology platforms alone are not sufficient to fully realize the promise of this technology. An additional requirement is the development of accurate and reproducible multivariate gene expression–based prediction models, also referred to as classifiers. Such models take gene expression data from a patient as input and as output produce a prediction of a clinically relevant outcome for that patient. Therefore, the second phase of the project (MAQC-II) has focused on these predictive models[18], studying both how they are developed and how they are evaluated. For any given microarray data set, many computational approaches can be followed to develop predictive models and to estimate the future performance of these models. Understanding the strengths and limitations of these various approaches is critical to the formulation of guidelines for safe and effective use of preclinical and clinical genomic data. Although previous studies have compared and benchmarked individual steps in the model development process[19], no prior published work has, to our knowledge, extensively evaluated current community practices on the development and validation of microarray-based predictive models.

Microarray-based gene expression data and prediction models are increasingly being submitted by the regulated industry to the FDA to support medical product development and testing applications[20]. For example, gene expression microarray–based assays that have been approved by the FDA as diagnostic tests include the Agendia MammaPrint microarray to assess prognosis of distant metastasis in breast cancer patients[21,22] and the Pathwork Tissue of Origin Test to assess the degree of similarity of the RNA expression pattern in a patient's tumor to that in a database of tumor samples for which the origin of the tumor is known[23]. Gene expression data have also been the basis for the development of PCR-based diagnostic assays, including the xDx Allomap test for detection of rejection of heart transplants[24].

The possible uses of gene expression data are vast and include diagnosis, early detection (screening), monitoring of disease progression, risk assessment, prognosis, complex medical product characterization and prediction of response to treatment (with regard to safety or efficacy) with a drug or device labeling intent. The ability to generate models in a reproducible fashion is an important consideration in predictive model development.

A lack of consistency in generating classifiers from publicly available data is problematic and may be due to any number of factors including insufficient annotation, incomplete clinical identifiers, coding errors and/or inappropriate use of methodology[25,26]. There

*A full list of authors and affiliations appears at the end of the paper. Correspondence should be addressed to L.S. (leming.shi@fda.hhs.gov or leming.shi@gmail.com).

Received 2 March; accepted 30 June; published online 30 July 2010; doi:10.1038/nbt.1665

are also examples in the literature of classifiers whose performance cannot be reproduced on independent data sets because of poor study design[27], poor data quality and/or insufficient cross-validation of all model development steps[28,29]. Each of these factors may contribute to a certain level of skepticism about claims of performance levels achieved by microarray-based classifiers.

Previous evaluations of the reproducibility of microarray-based classifiers, with only very few exceptions[30,31], have been limited to simulation studies or reanalysis of previously published results. Frequently, published benchmarking studies have split data sets at random, and used one part for training and the other for validation. This design assumes that the training and validation sets are produced by unbiased sampling of a large, homogeneous population of samples. However, specimens in clinical studies are usually accrued over years and there may be a shift in the participating patient population and also in the methods used to assign disease status owing to changing practice standards. There may also be batch effects owing to time variations in tissue analysis or due to distinct methods of sample collection and handling at different medical centers. As a result, samples derived from sequentially accrued patient populations, as was done in MAQC-II to mimic clinical reality, where the first cohort is used for developing predictive models and subsequent patients are included in validation, may differ from each other in many ways that could influence the prediction performance.

The MAQC-II project was designed to evaluate these sources of bias in study design by constructing training and validation sets at different times, swapping the test and training sets and also using data from diverse preclinical and clinical scenarios. The goals of MAQC-II were to survey approaches in genomic model development in an attempt to understand sources of variability in prediction performance and to assess the influences of endpoint signal strength in data. By providing the same data sets to many organizations for analysis, but not restricting their data analysis protocols, the project has made it possible to evaluate to what extent, if any, results depend on the team that performs the analysis. This contrasts with previous benchmarking studies that have typically been conducted by single laboratories. Enrolling a large number of organizations has also made it feasible to test many more approaches than would be practical for any single team. MAQC-II also strives to develop good modeling practice guidelines, drawing on a large international collaboration of experts and the lessons learned in the perhaps unprecedented effort of developing and evaluating >30,000 genomic classifiers to predict a variety of endpoints from diverse data sets.

MAQC-II is a collaborative research project that includes participants from the FDA, other government agencies, industry and academia. This paper describes the MAQC-II structure and experimental design and summarizes the main findings and key results of the consortium, whose members have learned a great deal during the process. The resulting guidelines are general and should not be construed as specific recommendations by the FDA for regulatory submissions.

RESULTS

Generating a unique compendium of >30,000 prediction models

The MAQC-II consortium was conceived with the primary goal of examining model development practices for generating binary classifiers in two types of data sets, preclinical and clinical (**Supplementary Tables 1** and **2**). To accomplish this, the project leader distributed six data sets containing 13 preclinical and clinical endpoints coded A through M (**Table 1**) to 36 voluntary participating data analysis teams representing academia, industry and government institutions (**Supplementary Table 3**). Endpoints were coded so as to hide the identities of two negative-control endpoints (endpoints I and M, for which class labels were randomly assigned and are not predictable by the microarray data) and two positive-control endpoints (endpoints H and L, representing the sex of patients, which is highly predictable by the microarray data). Endpoints A, B and C tested teams' ability to predict the toxicity of chemical agents in rodent lung and liver models. The remaining endpoints were predicted from microarray data sets from human patients diagnosed with breast cancer (D and E), multiple myeloma (F and G) or neuroblastoma (J and K). For the multiple myeloma and neuroblastoma data sets, the endpoints represented event free survival (abbreviated EFS), meaning a lack of malignancy or disease recurrence, and overall survival (abbreviated OS) after 730 days (for multiple myeloma) or 900 days (for neuroblastoma) post treatment or diagnosis. For breast cancer, the endpoints represented estrogen receptor status, a common diagnostic marker of this cancer type (abbreviated 'erpos'), and the success of treatment involving chemotherapy followed by surgical resection of a tumor (abbreviated 'pCR'). The biological meaning of the control endpoints was known only to the project leader and not revealed to the project participants until all model development and external validation processes had been completed.

To evaluate the reproducibility of the models developed by a data analysis team for a given data set, we asked teams to submit models from two stages of analyses. In the first stage (hereafter referred to as the 'original' experiment), each team built prediction models for up to 13 different coded endpoints using six training data sets. Models were 'frozen' against further modification, submitted to the consortium and then tested on a blinded validation data set that was not available to the analysis teams during training. In the second stage (referred to as the 'swap' experiment), teams repeated the model building and validation process by training models on the original validation set and validating them using the original training set.

To simulate the potential decision-making process for evaluating a microarray-based classifier, we established a process for each group to receive training data with coded endpoints, propose a data analysis protocol (DAP) based on exploratory analysis, receive feedback on the protocol and then perform the analysis and validation (**Fig. 1**). Analysis protocols were reviewed internally by other MAQC-II participants (at least two reviewers per protocol) and by members of the MAQC-II Regulatory Biostatistics Working Group (RBWG), a team from the FDA and industry comprising biostatisticians and others with extensive model building expertise. Teams were encouraged to revise their protocols to incorporate feedback from reviewers, but each team was eventually considered responsible for its own analysis protocol and incorporating reviewers' feedback was not mandatory (see Online Methods for more details).

We assembled two large tables from the original and swap experiments (**Supplementary Tables 1** and **2**, respectively) containing summary information about the algorithms and analytic steps, or 'modeling factors', used to construct each model and the 'internal' and 'external' performance of each model. Internal performance measures the ability of the model to classify the training samples, based on cross-validation exercises. External performance measures the ability of the model to classify the blinded independent validation data. We considered several performance metrics, including Matthews Correlation Coefficient (MCC), accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC) and root mean squared error (r.m.s.e.). These two tables contain data on >30,000 models. Here we report performance based on MCC because it is informative when the distribution of the two classes in a data set is highly skewed and because it is simple to calculate and was available for all models. MCC values range from +1 to –1, with +1 indicating perfect prediction (that is, all samples classified correctly and none incorrectly), 0 indicates random prediction and –1 indicating perfect inverse prediction.

The 36 analysis teams applied many different options under each modeling factor for developing models (**Supplementary Table 4**) including 17 summary and normalization methods, nine batch-effect removal methods, 33 feature selection methods (between 1 and >1,000 features), 24 classification algorithms and six internal validation methods. Such diversity suggests the community's common practices are

Table 1 Microarray data sets used for model development and validation in the MAQC-II project

				Training set[a]				Validation set[a]				
Date set code	Endpoint code	Endpoint description	Microarray platform	Number of samples	Positives (P)	Negatives (N)	P/N ratio	Number of samples	Positives (P)	Negatives (N)	P/N ratio	Comments and references
Hamner	A	Lung tumorigen vs. non-tumorigen (mouse)	Affymetrix Mouse 430 2.0	70	26	44	0.59	88	28	60	0.47	The training set was first published in 2007 (ref. 50) and the validation set was generated for MAQC-II
Iconix	B	Non-genotoxic liver carcinogens vs. non-carcinogens (rat)	Amersham Uniset Rat 1 Bioarray	216	73	143	0.51	201	57	144	0.40	The data set was first published in 2007 (ref. 51). Raw microarray intensity data, instead of ratio data, were provided for MAQC-II data analysis
NIEHS	C	Liver toxicants vs. non-toxicants based on overall necrosis score (rat)	Affymetrix Rat 230 2.0	214	79	135	0.58	204	78	126	0.62	Exploratory visualization of the data set was reported in 2008 (ref. 53). However, the phenotype classification problem was formulated specifically for MAQC-II. A large amount of additional microarray and phenotype data were provided to MAQC-II for cross-platform and cross-tissue comparisons
Breast cancer (BR)	D	Pre-operative treatment response (pCR, pathologic complete response)	Affymetrix Human U133A	130	33	97	0.34	100	15	85	0.18	The training set was first published in 2006 (ref. 56) and the validation set was specifically generated for MAQC-II. In addition, two distinct endpoints (D and E) were analyzed in MAQC-II
	E	Estrogen receptor status (erpos)		130	80	50	1.6	100	61	39	1.56	
Multiple myeloma (MM)	F	Overall survival milestone outcome (OS, 730-d cutoff)	Affymetrix Human U133Plus 2.0	340	51	289	0.18	214	27	187	0.14	The data set was first published in 2006 (ref. 57) and 2007 (ref. 58). However, patient survival data were updated and the raw microarray data (CEL files) were provided specifically for MAQC-II data analysis. In addition, endpoints H and I were designed and analyzed specifically in MAQC-II
	G	Event-free survival milestone outcome (EFS, 730-d cutoff)		340	84	256	0.33	214	34	180	0.19	
	H	Clinical parameter S1 (CPS1). The actual class label is the sex of the patient. Used as a "positive" control endpoint		340	194	146	1.33	214	140	74	1.89	
	I	Clinical parameter R1 (CPR1). The actual class label is randomly assigned. Used as a "negative" control endpoint		340	200	140	1.43	214	122	92	1.33	
Neuroblastoma (NB)	J	Overall survival milestone outcome (OS, 900-d cutoff)	Different versions of Agilent human microarrays	238	22	216	0.10	177	39	138	0.28	The training data set was first published in 2006 (ref. 63). The validation set (two-color Agilent platform) was generated specifically for MAQC-II. In addition, one-color Agilent platform data were also generated for most samples used in the training and validation sets specifically for MAQC-II to compare the prediction performance of two-color versus one-color platforms. Patient survival data were also updated. In addition, endpoints L and M were designed and analyzed specifically in MAQC-II
	K	Event-free survival milestone outcome (EFS, 900-d cutoff)		239	49	190	0.26	193	83	110	0.75	
	L	Newly established parameter S (NEP_S). The actual class label is the sex of the patient. Used as a "positive" control endpoint		246	145	101	1.44	231	133	98	1.36	
	M	Newly established parameter R (NEP_R). The actual class label is randomly assigned. Used as a "negative" control endpoint		246	145	101	1.44	253	143	110	1.30	

The first three data sets (Hamner, Iconix and NIEHS) are from preclinical toxicogenomics studies, whereas the other three data sets are from clinical studies. Endpoints H and L are positive controls (sex of patient) and endpoints I and M are negative controls (randomly assigned class labels). The nature of H, I, L and M was unknown to MAQC-II participants except for the project leader until all calculations were completed.

[a] Numbers shown are the actual number of samples used for model development or validation.

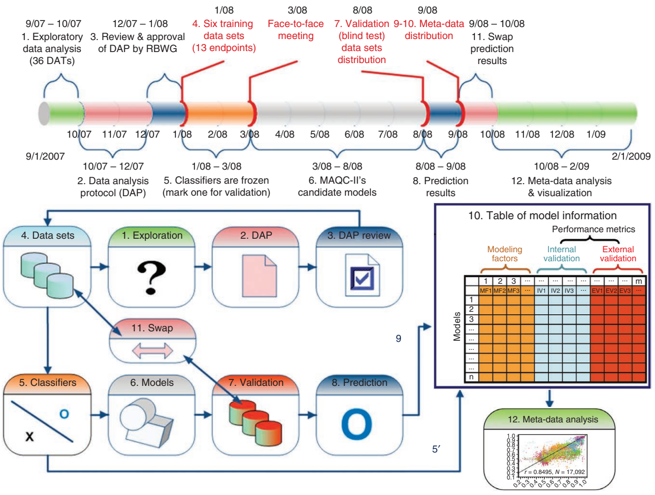

9/07 – 10/07
1. Exploratory data analysis (36 DATs)
12/07 – 1/08
3. Review & approval of DAP by RBWG
1/08
4. Six training data sets (13 endpoints)
3/08
Face-to-face meeting
8/08
7. Validation (blind test) data sets distribution
9/08
9-10. Meta-data distribution
9/08 – 10/08
11. Swap prediction results
10/07
11/07
12/07
1/08
2/08
3/08
4/08
5/08
6/08
7/08
8/08
9/08
10/08
11/08
12/08
1/09
9/1/2007
2/1/2009
10/07 – 12/07
2. Data analysis protocol (DAP)
1/08 – 3/08
5. Classifiers are frozen (mark one for validation)
3/08 – 8/08
6. MAQC-II's candidate models
8/08 – 9/08
8. Prediction results
10/08 – 2/09
12. Meta-data analysis & visualization
4. Data sets
1. Exploration
?
2. DAP
3. DAP review
11. Swap
5. Classifiers
o
x
6. Models
7. Validation
8. Prediction
O
9
5′
10. Table of model information
Performance metrics
Modeling factors
Internal validation
External validation
1 2 3 ... ... m
MF1 MF2 MF3 ... IV1 IV2 IV3 ... EV1 EV2 EV3 ...
Models
1
2
3
...
n
12. Meta-data analysis
r = 0.8495, N = 17,092

Figure 1 Experimental design and timeline of the MAQC-II project. Numbers (1–11) order the steps of analysis. Step 11 indicates when the original training and validation data sets were swapped to repeat steps 4–10. See main text for description of each step. Every effort was made to ensure the complete independence of the validation data sets from the training sets. Each model is characterized by several modeling factors and seven internal and external validation performance metrics (**Supplementary Tables 1** and **2**). The modeling factors include: (i) organization code; (ii) data set code; (iii) endpoint code; (iv) summary and normalization; (v) feature selection method; (vi) number of features used; (vii) classification algorithm; (viii) batch-effect removal method; (ix) type of internal validation; and (x) number of iterations of internal validation. The seven performance metrics for internal validation and external validation are: (i) MCC; (ii) accuracy; (iii) sensitivity; (iv) specificity; (v) AUC; (vi) mean of sensitivity and specificity; and (vii) r.m.s.e. s.d. of metrics are also provided for internal validation results.

well represented. For each of the models nominated by a team as being the best model for a particular endpoint, we compiled the list of features used for both the original and swap experiments (see the MAQC Web site at http://edkb.fda.gov/MAQC/). These comprehensive tables represent a unique resource. The results that follow describe data mining efforts to determine the potential and limitations of current practices for developing and validating gene expression–based prediction models.

Performance depends on endpoint and can be estimated during training

Unlike many previous efforts, the study design of MAQC-II provided the opportunity to assess the performance of many different modeling approaches on a clinically realistic blinded external validation data set. This is especially important in light of the intended clinical or preclinical uses of classifiers that are constructed using initial data sets and validated for regulatory approval and then are expected to accurately predict samples collected under diverse conditions perhaps months or years later. To assess the reliability of performance estimates derived during model training, we compared the performance on the internal training data set with performance on the external validation data set for of each of the 18,060 models in the original experiment (**Fig. 2a**). Models without complete metadata were not included in the analysis.

We selected 13 'candidate models', representing the best model for each endpoint, before external validation was performed. We required that each analysis team nominate one model for each endpoint they analyzed and we then selected one candidate from these nominations for each endpoint. We observed a higher correlation between internal and external performance estimates in terms

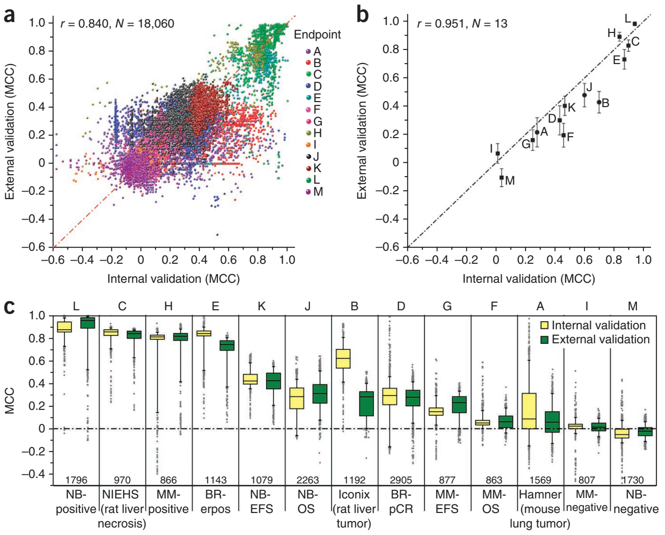
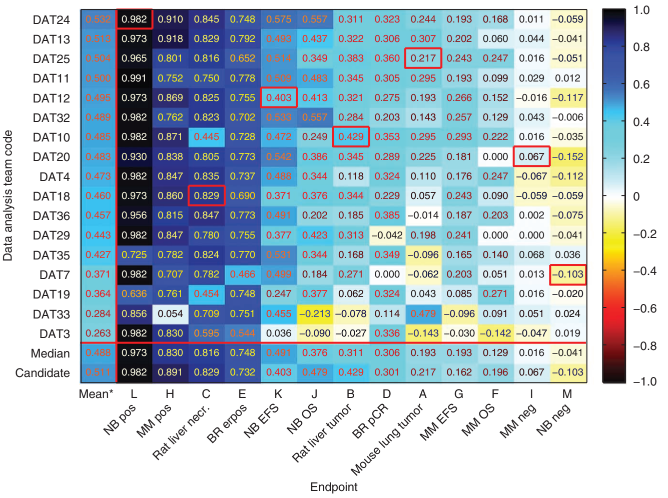


Figure 2 Model performance on internal validation compared with external validation. (**a**) Performance of 18,060 models that were validated with blinded validation data. (**b**) Performance of 13 candidate models. *r*, Pearson correlation coefficient; *N*, number of models. Candidate models with binary and continuous prediction values are marked as circles and squares, respectively, and the standard error estimate was obtained using 500-times resampling with bagging of the prediction results from each model. (**c**) Distribution of MCC values of all models for each endpoint in internal (left, yellow) and external (right, green) validation performance. Endpoints H and L (sex of the patients) are included as positive controls and endpoints I and M (randomly assigned sample class labels) as negative controls. Boxes indicate the 25% and 75% percentiles, and whiskers indicate the 5% and 95% percentiles.



Data analysis team code	Mean*	L	H	C	E	K	J	B	D	A	G	F	I	M
DAT24	0.532	0.982	0.910	0.845	0.748	0.575	0.557	0.311	0.323	0.244	0.193	0.168	0.011	−0.059
DAT13	0.513	0.973	0.918	0.829	0.792	0.493	0.437	0.322	0.306	0.307	0.202	0.060	0.044	−0.041
DAT25	0.504	0.965	0.801	0.816	0.652	0.514	0.349	0.383	0.360	0.217	0.243	0.247	0.016	−0.051
DAT11	0.500	0.991	0.752	0.750	0.778	0.509	0.483	0.345	0.305	0.295	0.193	0.099	0.029	0.012
DAT12	0.495	0.973	0.869	0.825	0.755	0.403	0.413	0.321	0.275	0.193	0.266	0.152	−0.016	−0.117
DAT32	0.489	0.982	0.762	0.823	0.702	0.533	0.557	0.284	0.203	0.143	0.257	0.129	0.043	−0.006
DAT10	0.485	0.982	0.871	0.445	0.728	0.472	0.249	0.429	0.353	0.295	0.293	0.222	0.016	−0.035
DAT20	0.483	0.930	0.838	0.805	0.773	0.542	0.386	0.345	0.289	0.225	0.181	0.000	0.067	−0.152
DAT4	0.473	0.982	0.847	0.835	0.737	0.488	0.344	0.118	0.324	0.110	0.176	0.247	−0.067	−0.112
DAT18	0.460	0.973	0.860	0.829	0.690	0.371	0.376	0.344	0.229	0.057	0.243	0.090	−0.059	−0.059
DAT36	0.457	0.956	0.815	0.847	0.773	0.491	0.202	0.185	0.385	−0.014	0.187	0.203	0.002	−0.075
DAT29	0.443	0.982	0.847	0.780	0.755	0.377	0.423	0.313	−0.042	0.198	0.241	0.000	0.000	−0.041
DAT35	0.427	0.725	0.782	0.824	0.770	0.531	0.344	0.168	0.349	−0.096	0.165	0.140	0.068	0.036
DAT7	0.371	0.982	0.707	0.782	0.466	0.499	0.184	0.271	0.000	−0.062	0.203	0.051	0.013	−0.103
DAT19	0.364	0.636	0.761	0.454	0.748	0.247	0.377	0.062	0.324	0.043	0.085	0.271	0.016	−0.020
DAT33	0.284	0.856	0.054	0.709	0.751	0.455	−0.213	−0.078	0.114	0.479	−0.096	0.091	0.051	0.024
DAT3	0.263	0.982	0.830	0.595	0.544	0.036	−0.090	−0.027	0.336	−0.143	−0.030	−0.142	−0.047	0.019
Median	0.488	0.973	0.830	0.816	0.748	0.491	0.376	0.311	0.306	0.193	0.193	0.129	0.016	−0.041
Candidate	0.511	0.982	0.891	0.829	0.732	0.403	0.479	0.429	0.301	0.217	0.162	0.196	0.067	−0.103
Endpoint		NB pos	MM pos	Rat liver necr.	BR erpos	NB EFS	NB OS	Rat liver tumor	BR pCR	Mouse lung tumor	MM EFS	MM OS	MM neg	NB neg

Figure 3 Performance, measured using MCC, of the best models nominated by the 17 data analysis teams (DATs) that analyzed all 13 endpoints in the original training-validation experiment. The median MCC value for an endpoint, representative of the level of predicability of the endpoint, was calculated based on values from the 17 data analysis teams. The mean MCC value for a data analysis team, representative of the team's proficiency in developing predictive models, was calculated based on values from the 11 non-random endpoints (excluding negative controls I and M). Red boxes highlight candidate models. Lack of a red box in an endpoint indicates that the candidate model was developed by a data analysis team that did not analyze all 13 endpoints.

of MCC for the selected candidate models ($r = 0.951$, $n = 13$, **Fig. 2b**) than for the overall set of models ($r = 0.840$, $n = 18{,}060$, **Fig. 2a**), suggesting that extensive peer review of analysis protocols was able to avoid selecting models that could result in less reliable predictions in external validation. Yet, even for the hand-selected candidate models, there is noticeable bias in the performance estimated from internal validation. That is, the internal validation performance is higher than the external validation performance for most endpoints (**Fig. 2b**). However, for some endpoints and for some model building methods or teams, internal and external performance correlations were more modest as described in the following sections.

To evaluate whether some endpoints might be more predictable than others and to calibrate performance against the positive- and negative-control endpoints, we assessed all models generated for each endpoint (**Fig. 2c**). We observed a clear dependence of prediction performance on endpoint. For example, endpoints C (liver necrosis score of rats treated with hepatotoxicants), E (estrogen receptor status of breast cancer patients), and H and L (sex of the multiple myeloma and neuroblastoma patients, respectively) were the easiest to predict (mean MCC > 0.7). Toxicological endpoints A and B and disease progression endpoints D, F, G, J and K were more difficult to predict (mean MCC ~0.1–0.4). Negative-control endpoints I and M were totally unpredictable (mean MCC ~0), as expected. For 11 endpoints (excluding the negative controls), a large proportion of the submitted models predicted the endpoint significantly better than chance (MCC > 0) and for a given endpoint many models performed similarly well on both internal and external validation (see the distribution of MCC in **Fig. 2c**). On the other hand, not all the submitted models performed equally well for any given endpoint. Some models performed no better than chance, even for some of the easy-to-predict endpoints, suggesting that additional factors were responsible for differences in model performance.

Data analysis teams show different proficiency

Next, we summarized the external validation performance of the models nominated by the 17 teams that analyzed all 13 endpoints (**Fig. 3**). Nominated models represent a team's best assessment of its model-building effort. The mean external validation MCC per team over 11 endpoints, excluding negative controls I and M, varied from 0.532 for data analysis team (DAT)24 to 0.263 for DAT3, indicating appreciable differences in performance of the models developed by different teams for the same data. Similar trends were observed when AUC was used as the performance metric (**Supplementary Table 5**) or when the original training and validation sets were swapped (**Supplementary Tables 6** and 7). **Table 2** summarizes the modeling approaches that were used by two or more MAQC-II data analysis teams.

Many factors may have played a role in the difference of external validation performance between teams. For instance, teams used different modeling factors, criteria for selecting the nominated models, and software packages and code. Moreover, some teams may have been more proficient at microarray data modeling and better at guarding against clerical errors. We noticed substantial variations in performance among the many *K*-nearest neighbor algorithm (KNN)-based models developed by four analysis teams (**Supplementary Fig. 1**). Follow-up investigations identified a few possible causes leading to the discrepancies in performance[32]. For example, DAT20 fixed the parameter 'number of neighbors' $K = 3$ in its data analysis protocol for all endpoints, whereas DAT18 varied K from 3 to 15 with a step size of 2. This investigation also revealed that even a detailed but standardized description of model building requested from all groups failed to capture many important tuning variables in the process. The subtle modeling differences not captured may have contributed to the differing performance levels achieved by the data analysis teams. The differences in performance for the models developed by various data analysis teams can also be observed from the changing patterns of internal and external validation performance across the 13 endpoints (**Fig. 3**, **Supplementary Tables 5**–7 and **Supplementary Figs. 2**–**4**). Our observations highlight the importance of good modeling practice in developing and validating microarray-based predictive models including reporting of computational details for results to be replicated[26]. In light of the MAQC-II experience, recording structured information about the steps and parameters of an analysis process seems highly desirable to facilitate peer review and reanalysis of results.

Swap and original analyses lead to consistent results

To evaluate the reproducibility of the models generated by each team, we correlated the performance of each team's models on the original training data set to performance on the validation data set and repeated this calculation for the swap experiment (**Fig. 4**). The correlation varied from 0.698–0.966 on the original experiment and from

Table 2 Modeling factor options frequently adopted by MAQC-II data analysis teams

		Original analysis (training => validation)		
Modeling factor	Option	Number of teams	Number of endpoints	Number of models
Summary and normalization	Loess	12	3	2,563
	RMA	3	7	46
	MAS5	11	7	4,947
Batch-effect removal	None	10	11	2,281
	Mean shift	3	11	7,279
Feature selection	SAM	4	11	3,771
	FC+P	8	11	4,711
	T-Test	5	11	400
	RFE	2	11	647
Number of features	0~9	10	11	393
	10~99	13	11	4,445
	≥1,000	3	11	474
	100~999	10	11	4,298
Classification algorithm	DA	4	11	103
	Tree	5	11	358
	NB	4	11	924
	KNN	8	11	6,904
	SVM	9	11	986

Analytic options used by two or more of the 14 teams that submitted models for all endpoints in both the original and swap experiments. RMA, robust multichip analysis; SAM, significance analysis of microarrays; FC, fold change; RFE, recursive feature elimination; DA, discriminant analysis; Tree, decision tree; NB, naive Bayes; KNN, K-nearest neighbors; SVM, support vector machine.

0.443–0.954 on the swap experiment. For all but three teams (DAT3, DAT10 and DAT11) the original and swap correlations were within ±0.2, and all but three others (DAT4, DAT13 and DAT36) were within ±0.1, suggesting that the model building process was relatively robust, at least with respect to generating models with similar performance. For some data analysis teams the internal validation performance drastically overestimated the performance of the same model in predicting the validation data. Examination of some of those models revealed several reasons, including bias in the feature selection and cross-validation process[28], findings consistent with what was observed from a recent literature survey[33].

Previously, reanalysis of a widely cited single study[34] found that the results in the original publication were very fragile—that is, not reproducible if the training and validation sets were swapped[35]. Our observations, except for DAT3, DAT11 and DAT36 with correlation <0.6, mainly resulting from failure of accurately predicting the positive-control endpoint H in the swap analysis (likely owing to operator errors), do not substantiate such fragility in the currently examined data sets. It is important to emphasize that we repeated the entire model building and evaluation processes during the swap analysis and, therefore, stability applies to the model building process for each data analysis team and not to a particular model or approach. **Supplementary Figure 5** provides a more detailed look at the correlation of internal and external validation for each data analysis team and each endpoint for both the original (**Supplementary Fig. 5a**) and swap (**Supplementary Fig. 5d**) analyses.

As expected, individual feature lists differed from analysis group to analysis group and between models developed from the original and the swapped data. However, when feature lists were mapped to biological processes, a greater degree of convergence and concordance was observed. This has been proposed previously but has never been demonstrated in a comprehensive manner over many data sets and thousands of models as was done in MAQC-II[36].

The effect of modeling factors is modest

To rigorously identify potential sources of variance that explain the variability in external-validation performance (**Fig. 2c**), we applied random effect modeling (**Fig. 5a**). We observed that the endpoint itself is by far the dominant source of variability, explaining >65% of the variability in the external validation performance. All other factors explain <8% of the total variance, and the residual variance is ~6%. Among the factors tested, those involving interactions with endpoint have a relatively large effect, in particular the interaction between endpoint with organization and classification algorithm, highlighting variations in proficiency between analysis teams.

To further investigate the impact of individual levels within each modeling factor, we estimated the empirical best linear unbiased predictors (BLUPs)[37]. **Figure 5b** shows the plots of BLUPs of the corresponding factors in **Figure 5a** with proportion of variation >1%. The BLUPs reveal the effect of each level of the factor to the corresponding MCC value. The BLUPs of the main endpoint effect show that rat liver necrosis, breast cancer estrogen receptor status and the sex of the patient (endpoints C, E, H and L) are relatively easier to be predicted with ~0.2–0.4 advantage contributed on the corresponding MCC values. The rest of the endpoints are relatively harder to be predicted with about −0.1 to −0.2 disadvantage contributed to the corresponding MCC values. The main factors of normalization, classification algorithm, the number of selected features and the feature selection method have an impact of −0.1 to 0.1 on the corresponding MCC values. Loess normalization was applied to the endpoints (J, K and L) for the neuroblastoma data set with the two-color Agilent platform and has 0.1 advantage to MCC values. Among the Microarray Analysis Suite version 5 (MAS5), Robust Multichip Analysis (RMA) and dChip normalization methods that were applied to all endpoints (A, C, D, E, F, G and H) for Affymetrix data, the dChip method has a lower BLUP than the others. Because normalization methods are partially confounded with endpoints, it may not be suitable to compare methods between different confounded groups. Among classification methods, discriminant analysis has the largest positive impact of 0.056 on the MCC values. Regarding the number of selected features, larger bin number has better impact on the average across endpoints. The bin number is assigned by applying the ceiling function to the log base 10 of the number of selected features. All the feature selection methods have a slight impact of −0.025 to 0.025

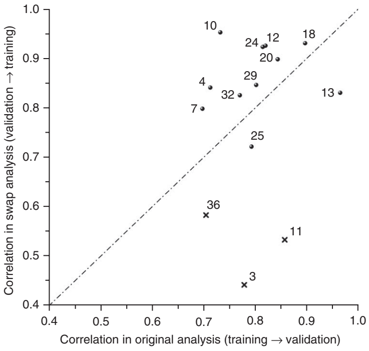


Figure 4 Correlation between internal and external validation is dependent on data analysis team. Pearson correlation coefficients between internal and external validation performance in terms of MCC are displayed for the 14 teams that submitted models for all 13 endpoints in both the original (*x* axis) and swap (*y* axis) analyses. The unusually low correlation in the swap analysis for DAT3, DAT11 and DAT36 is a result of their failure to accurately predict the positive endpoint H, likely due to operator errors (**Supplementary Table 6**).

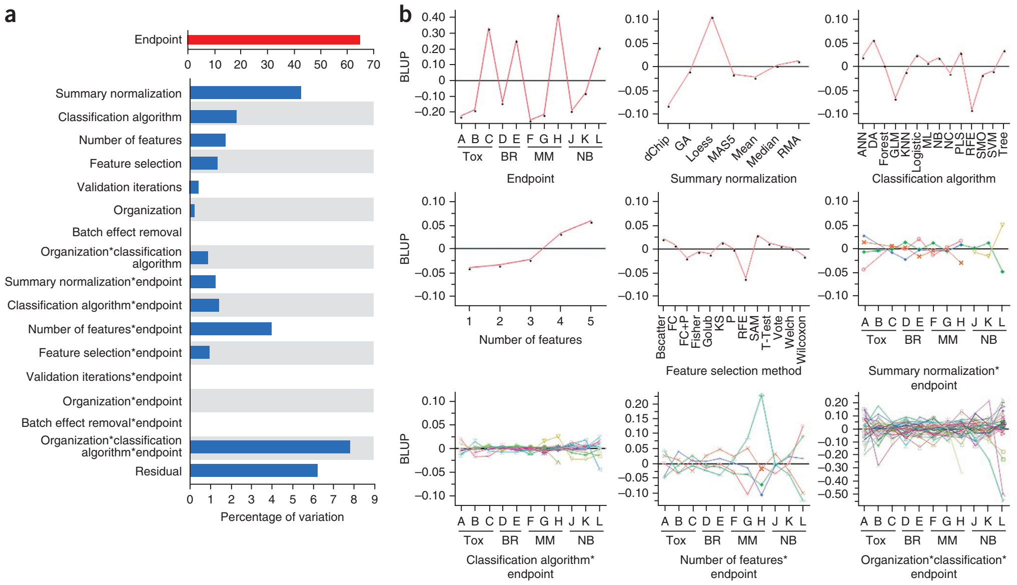


Figure 5 Effect of modeling factors on estimates of model performance. (**a**) Random-effect models of external validation performance (MCC) were developed to estimate a distinct variance component for each modeling factor and several selected interactions. The estimated variance components were then divided by their total in order to compare the proportion of variability explained by each modeling factor. The endpoint code contributes the most to the variability in external validation performance. (**b**) The BLUP plots of the corresponding modeling factors having proportion of variation larger than 1% in **a**. Endpoint abbreviations (Tox., preclinical toxicity; BR, breast cancer; MM, multiple myeloma; NB, neuroblastoma). Endpoints H and L are the sex of the patient. Summary normalization abbreviations (GA, genetic algorithm; RMA, robust multichip analysis). Classification algorithm abbreviations (ANN, artificial neural network; DA, discriminant analysis; Forest, random forest; GLM, generalized linear model; KNN, K-nearest neighbors; Logistic, logistic regression; ML, maximum likelihood; NB, Naïve Bayes; NC, nearest centroid; PLS, partial least squares; RFE, recursive feature elimination; SMO, sequential minimal optimization; SVM, support vector machine; Tree, decision tree). Feature selection method abbreviations (Bscatter, between-class scatter; FC, fold change; KS, Kolmogorov-Smirnov algorithm; SAM, significance analysis of microarrays).

on MCC values except for recursive feature elimination (RFE) that has an impact of –0.006. In the plots of the four selected interactions, the estimated BLUPs vary across endpoints. The large variation across endpoints implies the impact of the corresponding modeling factor on different endpoints can be very different. Among the four interaction plots (see **Supplementary Fig. 6** for a clear labeling of each interaction term), the corresponding BLUPs of the three-way interaction of organization, classification algorithm and endpoint show the highest variation. This may be due to different tuning parameters applied to individual algorithms for different organizations, as was the case for KNN[32].

We also analyzed the relative importance of modeling factors on external-validation prediction performance using a decision tree model[38]. The analysis results revealed observations (**Supplementary Fig. 7**) largely consistent with those above. First, the endpoint code was the most influential modeling factor. Second, feature selection method, normalization and summarization method, classification method and organization code also contributed to prediction performance, but their contribution was relatively small.

Feature list stability is correlated with endpoint predictability

Prediction performance is the most important criterion for evaluating the performance of a predictive model and its modeling process. However, the robustness and mechanistic relevance of the model and the corresponding gene signature is also important (**Supplementary Fig. 8**). That is, given comparable prediction performance between two modeling processes, the one yielding a more robust and reproducible gene signature across similar data sets (e.g., by swapping the training and validation sets), which is therefore less susceptible to sporadic fluctuations in the data, or the one that provides new insights to the underlying biology is preferable. Reproducibility or stability of feature sets is best studied by running the same model selection protocol on two distinct collections of samples, a scenario only possible, in this case, after the blind validation data were distributed to the data analysis teams that were asked to perform their analysis after swapping their original training and test sets. **Supplementary Figures 9** and **10** show that, although the feature space is extremely large for microarray data, different teams and protocols were able to consistently select the best-performing features. Analysis of the lists of features indicated that for endpoints relatively easy to predict, various data analysis teams arrived at models that used more common features and the overlap of the lists from the original and swap analyses is greater than those for more difficult endpoints (**Supplementary Figs. 9–11**). Therefore, the level of stability of feature lists can be associated to the level of difficulty of the prediction problem (**Supplementary Fig. 11**), although multiple models with different feature lists and comparable performance can be found from the same data set[39]. Functional analysis of the most frequently selected genes by all data analysis protocols shows that many of these genes represent biological processes that are highly relevant to the clinical outcome that is being predicted[36]. The sex-based endpoints have the best overlap, whereas more difficult survival endpoints (in which disease processes are confounded by many other factors) have only marginally better overlap with biological processes relevant to the disease than that expected by random chance.

Summary of MAQC-II observations and recommendations

The MAQC-II data analysis teams comprised a diverse group, some of whom were experienced microarray analysts whereas others were graduate students with little experience. In aggregate, the group's composition likely mimicked the broad scientific community engaged in building and publishing models derived from microarray data. The more than 30,000 models developed by 36 data analysis teams for 13 endpoints from six diverse clinical and preclinical data sets are a rich source from which to highlight several important observations.

First, model prediction performance was largely endpoint (biology) dependent (**Figs. 2c** and **3**). The incorporation of multiple data sets and endpoints (including positive and negative controls) in the MAQC-II study design made this observation possible. Some endpoints are highly predictive based on the nature of the data, which makes it possible to build good models, provided that sound modeling procedures are used. Other endpoints are inherently difficult to predict regardless of the model development protocol.

Second, there are clear differences in proficiency between data analysis teams (organizations) and such differences are correlated with the level of experience of the team. For example, the top-performing teams shown in **Figure 3** were mainly industrial participants with many years of experience in microarray data analysis, whereas bottom-performing teams were mainly less-experienced graduate students or researchers. Based on results from the positive and negative endpoints, we noticed that simple errors were sometimes made, suggesting rushed efforts due to lack of time or unnoticed implementation flaws. This observation strongly suggests that mechanisms are needed to ensure the reliability of results presented to the regulatory agencies, journal editors and the research community. By examining the practices of teams whose models did not perform well, future studies might be able to identify pitfalls to be avoided. Likewise, practices adopted by top-performing teams can provide the basis for developing good modeling practices.

Third, the internal validation performance from well-implemented, unbiased cross-validation shows a high degree of concordance with the external validation performance in a strict blinding process (**Fig. 2**). This observation was not possible from previously published studies owing to the small number of available endpoints tested in them.

Fourth, many models with similar performance can be developed from a given data set (**Fig. 2**). Similar prediction performance is attainable when using different modeling algorithms and parameters, and simple data analysis methods often perform as well as more complicated approaches[32,40]. Although it is not essential to include the same features in these models to achieve comparable prediction performance, endpoints that were easier to predict generally yielded models with more common features, when analyzed by different teams (**Supplementary Fig. 11**).

Finally, applying good modeling practices appeared to be more important than the actual choice of a particular algorithm over the others within the same step in the modeling process. This can be seen in the diverse choices of the modeling factors used by teams that produced models that performed well in the blinded validation (**Table 2**) where modeling factors did not universally contribute to variations in model performance among good performing teams (**Fig. 5**).

Summarized below are the model building steps recommended to the MAQC-II data analysis teams. These may be applicable to model building practitioners in the general scientific community.

Step one (design). There is no exclusive set of steps and procedures, in the form of a checklist, to be followed by any practitioner for all problems. However, normal good practice on the study design and the ratio of sample size to classifier complexity should be followed. The frequently used options for normalization, feature selection and classification are good starting points (**Table 2**).

Step two (pilot study or internal validation). This can be accomplished by bootstrap or cross-validation such as the ten repeats of a fivefold cross-validation procedure adopted by most MAQC-II teams. The samples from the pilot study are not replaced for the pivotal study; rather they are augmented to achieve 'appropriate' target size.

Step three (pivotal study or external validation). Many investigators assume that the most conservative approach to a pivotal study is to simply obtain a test set completely independent of the training set(s). However, it is good to keep in mind the exchange[34,35] regarding the fragility of results when the training and validation sets are swapped. Results from further resampling (including simple swapping as in MAQC-II) across the training and validation sets can provide important information about the reliability of the models and the modeling procedures, but the complete separation of the training and validation sets should be maintained[41].

Finally, a perennial issue concerns reuse of the independent validation set after modifications to an originally designed and validated data analysis algorithm or protocol. Such a process turns the validation set into part of the design or training set[42]. Ground rules must be developed for avoiding this approach and penalizing it when it occurs; and practitioners should guard against using it before such ground rules are well established.

DISCUSSION

MAQC-II conducted a broad observational study of the current community landscape of gene-expression profile–based predictive model development. Microarray gene expression profiling is among the most commonly used analytical tools in biomedical research. Analysis of the high-dimensional data generated by these experiments involves multiple steps and several critical decision points that can profoundly influence the soundness of the results[43]. An important requirement of a sound internal validation is that it must include feature selection and parameter optimization within each iteration to avoid overly optimistic estimations of prediction performance[28,29,44]. To what extent this information has been disseminated and followed by the scientific community in current microarray analysis remains unknown[33]. Concerns have been raised that results published by one group of investigators often cannot be confirmed by others even if the same data set is used[26]. An inability to confirm results may stem from any of several reasons: (i) insufficient information is provided about the methodology that describes which analysis has actually been done; (ii) data preprocessing (normalization, gene filtering and feature selection) is too complicated and insufficiently documented to be reproduced; or (iii) incorrect or biased complex analytical methods[26] are performed. A distinct but related concern is that genomic data may yield prediction models that, even if reproducible on the discovery data set, cannot be extrapolated well in independent validation. The MAQC-II project provided a unique opportunity to address some of these concerns.

Notably, we did not place restrictions on the model building methods used by the data analysis teams. Accordingly, they adopted numerous different modeling approaches (**Table 2** and **Supplementary Table 4**).

For example, feature selection methods varied widely, from statistical significance tests, to machine learning algorithms, to those more reliant on differences in expression amplitude, to those employing knowledge of putative biological mechanisms associated with the endpoint. Prediction algorithms also varied widely. To make internal validation performance results comparable across teams for different models, we recommended that a model's internal performance was estimated using a ten times repeated fivefold cross-validation, but this recommendation was not strictly followed by all teams, which also allows us to survey internal validation approaches. The diversity of analysis protocols used by the teams is likely to closely resemble that of current research going forward, and in this context mimics reality. In terms of the space of modeling factors explored, MAQC-II is a survey of current practices rather than a randomized, controlled experiment; therefore, care should be taken in interpreting the results. For example, some teams did not analyze all endpoints, causing missing data (models) that may be confounded with other modeling factors.

Overall, the procedure followed to nominate MAQC-II candidate models was quite effective in selecting models that performed reasonably well during validation using independent data sets, although generally the selected models did not do as well in validation as in training. The drop in performance associated with the validation highlights the importance of not relying solely on internal validation performance, and points to the need to subject every classifier to at least one external validation. The selection of the 13 candidate models from many nominated models was achieved through a peer-review collaborative effort of many experts and could be described as slow, tedious and sometimes subjective (e.g., a data analysis team could only contribute one of the 13 candidate models). Even though they were still subject to over-optimism, the internal and external performance estimates of the candidate models were more concordant than those of the overall set of models. Thus the review was productive in identifying characteristics of reliable models.

An important lesson learned through MAQC-II is that it is almost impossible to retrospectively retrieve and document decisions that were made at every step during the feature selection and model development stage. This lack of complete description of the model building process is likely to be a common reason for the inability of different data analysis teams to fully reproduce each other's results[32]. Therefore, although meticulously documenting the classifier building procedure can be cumbersome, we recommend that all genomic publications include supplementary materials describing the model building and evaluation process in an electronic format. MAQC-II is making available six data sets with 13 endpoints that can be used in the future as a benchmark to verify that software used to implement new approaches performs as expected. Subjecting new software to benchmarks against these data sets could reassure potential users that the software is mature enough to be used for the development of predictive models in new data sets. It would seem advantageous to develop alternative ways to help determine whether specific implementations of modeling approaches and performance evaluation procedures are sound, and to identify procedures to capture this information in public databases.

The findings of the MAQC-II project suggest that when the same data sets are provided to a large number of data analysis teams, many groups can generate similar results even when different model building approaches are followed. This is concordant with studies[29,33] that found that given good quality data and an adequate number of informative features, most classification methods, if properly used, will yield similar predictive performance. This also confirms reports[6,7,39] on small data sets by individual groups that have suggested that several different feature selection methods and prediction algorithms can yield many models that are distinct, but have statistically similar performance. Taken together, these results provide perspective on the large number of publications in the bioinformatics literature that have examined the various steps of the multivariate prediction model building process and identified elements that are critical for achieving reliable results.

An important and previously underappreciated observation from MAQC-II is that different clinical endpoints represent very different levels of classification difficulty. For some endpoints the currently available data are sufficient to generate robust models, whereas for other endpoints currently available data do not seem to be sufficient to yield highly predictive models. An analysis done as part of the MAQC-II project and that focused on the breast cancer data demonstrates these points in more detail[40]. It is also important to point out that for some clinically meaningful endpoints studied in the MAQC-II project, gene expression data did not seem to significantly outperform models based on clinical covariates alone, highlighting the challenges in predicting the outcome of patients in a heterogeneous population and the potential need to combine gene expression data with clinical covariates (unpublished data).

The accuracy of the clinical sample annotation information may also play a role in the difficulty to obtain accurate prediction results on validation samples. For example, some samples were misclassified by almost all models (**Supplementary Fig. 12**). It is true even for some samples within the positive control endpoints H and L, as shown in **Supplementary Table 8**. Clinical information of neuroblastoma patients for whom the positive control endpoint L was uniformly misclassified were rechecked and the sex of three out of eight cases (NB412, NB504 and NB522) was found to be incorrectly annotated.

The companion MAQC-II papers published elsewhere give more in-depth analyses of specific issues such as the clinical benefits of genomic classifiers (unpublished data), the impact of different modeling factors on prediction performance[45], the objective assessment of microarray cross-platform prediction[46], cross-tissue prediction[47], one-color versus two-color prediction comparison[48], functional analysis of gene signatures[36] and recommendation of a simple yet robust data analysis protocol based on the KNN[32]. For example, we systematically compared the classification performance resulting from one- and two-color gene-expression profiles of 478 neuroblastoma samples and found that analyses based on either platform yielded similar classification performance[48]. This newly generated one-color data set has been used to evaluate the applicability of the KNN-based simple data analysis protocol to future data sets[32]. In addition, the MAQC-II Genome-Wide Association Working Group assessed the variabilities in genotype calling due to experimental or algorithmic factors[49].

In summary, MAQC-II has demonstrated that current methods commonly used to develop and assess multivariate gene-expression based predictors of clinical outcome were used appropriately by most of the analysis teams in this consortium. However, differences in proficiency emerged and this underscores the importance of proper implementation of otherwise robust analytical methods. Observations based on analysis of the MAQC-II data sets may be applicable to other diseases. The MAQC-II data sets are publicly available and are expected to be used by the scientific community as benchmarks to ensure proper modeling practices. The experience with the MAQC-II clinical data sets also reinforces the notion that clinical classification problems represent several different degrees of prediction difficulty that are likely to be associated with whether mRNA abundances measured in a specific data set are informative for the specific prediction problem. We anticipate that including other types of biological data at the DNA, microRNA, protein or metabolite levels will enhance our capability to more accurately predict the clinically relevant endpoints. The good modeling practice guidelines established by MAQC-II and lessons learned from this unprecedented collaboration provide a solid foundation from which other high-dimensional biological data could be more reliably used for the purpose of predictive and personalized medicine.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. All MAQC-II data sets are available through GEO (series accession number: GSE16716), the MAQC Web site (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/), ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/) or CEBS (http://cebs.niehs.nih.gov/) accession number: 009-00002-0010-000-3.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The MAQC-II project was funded in part by the FDA's Office of Critical Path Programs (to L.S.). Participants from the National Institutes of Health (NIH) were supported by the Intramural Research Program of NIH, Bethesda, Maryland or the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, North Carolina. J.F. was supported by the Division of Intramural Research of the NIEHS under contract HHSN273200700046U. Participants from the Johns Hopkins University were supported by grants from the NIH (1R01GM083084-01 and 1R01RR021967-01A2 to R.A.I. and T32GM074906 to M.M.). Participants from the Weill Medical College of Cornell University were partially supported by the Biomedical Informatics Core of the Institutional Clinical and Translational Science Award RFA-RM-07-002. F.C. acknowledges resources from The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine and from the David A. Cofrin Center for Biomedical Information at Weill Cornell. The data set from The Hamner Institutes for Health Sciences was supported by a grant from the American Chemistry Council's Long Range Research Initiative. The breast cancer data set was generated with support of grants from NIH (R-01 to L.P.), The Breast Cancer Research Foundation (to L.P. and W.F.S.) and the Faculty Incentive Funds of the University of Texas MD Anderson Cancer Center (to W.F.S.). The data set from the University of Arkansas for Medical Sciences was supported by National Cancer Institute (NCI) PO1 grant CA55819-01A1, NCI R33 Grant CA97513-01, Donna D. and Donald M. Lambert Lebow Fund to Cure Myeloma and Nancy and Steven Grand Foundation. We are grateful to the individuals whose gene expression data were used in this study. All MAQC-II participants freely donated their time and reagents for the completion and analyses of the MAQC-II project. The MAQC-II consortium also thanks R. O'Neill for his encouragement and coordination among FDA Centers on the formation of the RBWG. The MAQC-II consortium gratefully dedicates this work in memory of R.F. Wagner who enthusiastically worked on the MAQC-II project and inspired many of us until he unexpectedly passed away in June 2008.

DISCLAIMER

This work includes contributions from, and was reviewed by, individuals at the FDA, the Environmental Protection Agency (EPA) and the NIH. This work has been approved for publication by these agencies, but it does not necessarily reflect official agency policy. Certain commercial materials and equipment are identified in order to adequately specify experimental procedures. In no case does such identification imply recommendation or endorsement by the FDA, the EPA or the NIH, nor does it imply that the items identified are necessarily the best available for the purpose.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630–631 (2004).
2. Frantz, S. An array of problems. *Nat. Rev. Drug Discov.* **4**, 362–363 (2005).
3. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–492 (2005).
4. Ntzani, E.E. & Ioannidis, J.P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
5. Ioannidis, J.P. Microarrays and molecular research: noise discovery? *Lancet* **365**, 454–455 (2005).
6. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178 (2005).
7. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* **103**, 5923–5928 (2006).
8. Shi, L. *et al.* QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev. Mol. Diagn.* **4**, 761–777 (2004).
9. Shi, L. *et al.* Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6** Suppl 2, S12 (2005).
10. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
11. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169 (2006).
12. Canales, R.D. *et al.* Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24**, 1115–1122 (2006).
13. Patterson, T.A. *et al.* Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol.* **24**, 1140–1150 (2006).
14. Shippy, R. *et al.* Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.* **24**, 1123–1131 (2006).
15. Tong, W. *et al.* Evaluation of external RNA controls for the assessment of microarray performance. *Nat. Biotechnol.* **24**, 1132–1139 (2006).
16. Irizarry, R.A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
17. Strauss, E. Arrays of hope. *Cell* **127**, 657–659 (2006).
18. Shi, L., Perkins, R.G., Fang, H. & Tong, W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr. Opin. Biotechnol.* **19**, 10–18 (2008).
19. Dudoit, S., Fridlyand, J. & Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–87 (2002).
20. Goodsaid, F.M. *et al.* Voluntary exploratory data submissions to the US FDA and the EMA: experience and impact. *Nat. Rev. Drug Discov.* **9**, 435–445 (2010).
21. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
22. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* **98**, 1183–1192 (2006).
23. Dumur, C.I. *et al.* Interlaboratory performance of a microarray-based gene expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers. *J. Mol. Diagn.* **10**, 67–77 (2008).
24. Deng, M.C. *et al.* Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. *Am. J. Transplant.* **6**, 150–160 (2006).
25. Coombes, K.R., Wang, J. & Baggerly, K.A. Microarrays: retracing steps. *Nat. Med.* **13**, 1276–1277, author reply 1277–1278 (2007).
26. Ioannidis, J.P.A. *et al.* Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155 (2009).
27. Baggerly, K.A., Edmonson, S.R., Morris, J.S. & Coombes, K.R. High-resolution serum proteomic patterns for ovarian cancer detection. *Endocr. Relat. Cancer* **11**, 583–584, author reply 585–587 (2004).
28. Ambroise, C. & McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **99**, 6562–6566 (2002).
29. Simon, R. Using DNA microarrays for diagnostic and prognostic prediction. *Expert Rev. Mol. Diagn.* **3**, 587–595 (2003).
30. Dobbin, K.K. *et al.* Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res.* **11**, 565–572 (2005).
31. Shedden, K. *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* **14**, 822–827 (2008).
32. Parry, R.M. *et al.* K-nearest neighbors (KNN) models for microarray gene-expression analysis and reliable clinical outcome prediction. *Pharmacogenomics J.* **10**, 292–309 (2010).
33. Dupuy, A. & Simon, R.M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.* **99**, 147–157 (2007).
34. Dave, S.S. *et al.* Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351**, 2159–2169 (2004).
35. Tibshirani, R. Immune signatures in follicular lymphoma. *N. Engl. J. Med.* **352**, 1496–1497, author reply 1496–1497 (2005).

36. Shi, W. *et al.* Functional analysis of multiple genomic signatures demonstrates that classification algorithms choose phenotype-related genes. *Pharmacogenomics J.* **10**, 310–323 (2010).
37. Robinson, G.K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32 (1991).
38. Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Statist.* **15**, 651–674 (2006).
39. Boutros, P.C. *et al.* Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl. Acad. Sci. USA* **106**, 2824–2828 (2009).
40. Popovici, V. *et al.* Effect of training sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* **12**, R5 (2010).
41. Yousef, W.A., Wagner, R.F. & Loew, M.H. Assessing classifiers from two independent data sets using ROC analysis: a nonparametric approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1809–1817 (2006).
42. Gur, D., Wagner, R.F. & Chan, H.P. On the repeated use of databases for testing incremental improvement of computer-aided detection schemes. *Acad. Radiol.* **11**, 103–105 (2004).
43. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7**, 55–65 (2006).
44. Wood, I.A., Visscher, P.M. & Mengersen, K.L. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics* **23**, 1363–1370 (2007).
45. Luo, J. *et al.* A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* **10**, 278–291 (2010).
46. Fan, X. *et al.* Consistency of predictive signature genes and classifiers generated using different microarray platforms. *Pharmacogenomics J.* **10**, 247–257 (2010).
47. Huang, J. *et al.* Genomic indicators in the blood predict drug-induced liver injury. *Pharmacogenomics J.* **10**, 267–277 (2010).
48. Oberthuer, A. *et al.* Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. *Pharmacogenomics J.* **10**, 258–266 (2010).
49. Hong, H. *et al.* Assessing sources of inconsistencies in genotypes and their effects on genome-wide association studies with HapMap samples. *Pharmacogenomics J.* **10**, 364–374 (2010).

Leming Shi[1], Gregory Campbell[2], Wendell D Jones[3], Fabien Campagne[4], Zhining Wen[1], Stephen J Walker[5], Zhenqiang Su[6], Tzu-Ming Chu[7], Federico M Goodsaid[8], Lajos Pusztai[9], John D Shaughnessy Jr[10], André Oberthuer[11], Russell S Thomas[12], Richard S Paules[13], Mark Fielden[14], Bart Barlogie[10], Weijie Chen[2], Pan Du[15], Matthias Fischer[11], Cesare Furlanello[16], Brandon D Gallas[2], Xijin Ge[17], Dalila B Megherbi[18], W Fraser Symmans[19], May D Wang[20], John Zhang[21], Hans Bitter[22], Benedikt Brors[23], Pierre R Bushel[13], Max Bylesjo[24], Minjun Chen[1], Jie Cheng[25], Jing Cheng[26], Jeff Chou[13], Timothy S Davison[27], Mauro Delorenzi[28], Youping Deng[29], Viswanath Devanarayan[30], David J Dix[31], Joaquin Dopazo[32], Kevin C Dorff[33], Fathi Elloumi[31], Jianqing Fan[34], Shicai Fan[35], Xiaohui Fan[36], Hong Fang[6], Nina Gonzaludo[37], Kenneth R Hess[38], Huixiao Hong[1], Jun Huan[39], Rafael A Irizarry[40], Richard Judson[31], Dilafruz Juraeva[23], Samir Lababidi[41], Christophe G Lambert[42], Li Li[7], Yanen Li[43], Zhen Li[31], Simon M Lin[15], Guozhen Liu[44], Edward K Lobenhofer[45], Jun Luo[21], Wen Luo[46], Matthew N McCall[40], Yuri Nikolsky[47], Gene A Pennello[2], Roger G Perkins[1], Reena Philip[2], Vlad Popovici[28], Nathan D Price[48], Feng Qian[6], Andreas Scherer[49], Tieliu Shi[50], Weiwei Shi[47], Jaeyun Sung[48], Danielle Thierry-Mieg[51], Jean Thierry-Mieg[51], Venkata Thodima[52], Johan Trygg[24], Lakshmi Vishnuvajjala[2], Sue Jane Wang[8], Jianping Wu[53], Yichao Wu[54], Qian Xie[55], Waleed A Yousef[56], Liang Zhang[53], Xuegong Zhang[35], Sheng Zhong[57], Yiming Zhou[10], Sheng Zhu[53], Dhivya Arasappan[6], Wenjun Bao[7], Anne Bergstrom Lucas[58], Frank Berthold[11], Richard J Brennan[47], Andreas Buness[59], Jennifer G Catalano[41], Chang Chang[50], Rong Chen[60], Yiyu Cheng[36], Jian Cui[50], Wendy Czika[7], Francesca Demichelis[61], Xutao Deng[62], Damir Dosymbekov[63], Roland Eils[23], Yang Feng[34], Jennifer Fostel[13], Stephanie Fulmer-Smentek[58], James C Fuscoe[1], Laurent Gatto[64], Weigong Ge[1], Darlene R Goldstein[65], Li Guo[66], Donald N Halbert[67], Jing Han[41], Stephen C Harris[1], Christos Hatzis[68], Damir Herman[69], Jianping Huang[36], Roderick V Jensen[70], Rui Jiang[35], Charles D Johnson[71], Giuseppe Jurman[16], Yvonne Kahlert[11], Sadik A Khuder[72], Matthias Kohl[73], Jianying Li[74], Li Li[75], Menglong Li[76], Quan-Zhen Li[77], Shao Li[36], Zhiguang Li[1], Jie Liu[1], Ying Liu[35], Zhichao Liu[1], Lu Meng[35], Manuel Madera[18], Francisco Martinez-Murillo[2], Ignacio Medina[78], Joseph Meehan[6], Kelci Miclaus[7], Richard A Moffitt[20], David Montaner[78], Piali Mukherjee[33], George J Mulligan[79], Padraic Neville[7], Tatiana Nikolskaya[47], Baitang Ning[1], Grier P Page[80], Joel Parker[3], R Mitchell Parry[20], Xuejun Peng[81], Ron L Peterson[82], John H Phan[20], Brian Quanz[39], Yi Ren[83], Samantha Riccadonna[16], Alan H Roter[84], Frank W Samuelson[2], Martin M Schumacher[85], Joseph D Shambaugh[86], Qiang Shi[1], Richard Shippy[87], Shengzhu Si[88], Aaron Smalter[39], Christos Sotiriou[89], Mat Soukup[8], Frank Staedtler[85], Guido Steiner[90], Todd H Stokes[20], Qinglan Sun[53], Pei-Yi Tan[7], Rong Tang[2], Zivana Tezak[2], Brett Thorn[1], Marina Tsyganova[63], Yaron Turpaz[91], Silvia C Vega[92], Roberto Visintainer[16], Juergen von Frese[93], Charles Wang[62], Eric Wang[21], Junwei Wang[50], Wei Wang[94], Frank Westermann[23], James C Willey[95], Matthew Woods[21], Shujian Wu[96], Nianqing Xiao[97], Joshua Xu[6], Lei Xu[1], Lun Yang[1], Xiao Zeng[44], Jialu Zhang[8], Li Zhang[8], Min Zhang[1], Chen Zhao[50], Raj K Puri[41], Uwe Scherf[2], Weida Tong[1] & Russell D Wolfinger[7]

[1]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [2]Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, Maryland, USA. [3]Expression Analysis Inc., Durham, North Carolina, USA. [4]Department of Physiology and Biophysics and HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York, USA. [5]Wake Forest Institute for Regenerative Medicine, Wake Forest University, Winston-Salem, North Carolina, USA. [6]Z-Tech, an ICF International Company at NCTR/FDA, Jefferson, Arkansas, USA. [7]SAS Institute Inc., Cary, North Carolina, USA. [8]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA. [9]Breast Medical Oncology Department, University of Texas (UT) M.D. Anderson Cancer Center, Houston, Texas, USA. [10]Myeloma Institute for Research

and Therapy, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA. [11]Department of Pediatric Oncology and Hematology and Center for Molecular Medicine (CMMC), University of Cologne, Cologne, Germany. [12]The Hamner Institutes for Health Sciences, Research Triangle Park, North Carolina, USA. [13]National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina, USA. [14]Roche Palo Alto LLC, South San Francisco, California, USA. [15]Biomedical Informatics Center, Northwestern University, Chicago, Illinois, USA. [16]Fondazione Bruno Kessler, Povo-Trento, Italy. [17]Department of Mathematics & Statistics, South Dakota State University, Brookings, South Dakota, USA. [18]CMINDS Research Center, Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, Massachusetts, USA. [19]Department of Pathology, UT M.D. Anderson Cancer Center, Houston, Texas, USA. [20]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA. [21]Systems Analytics Inc., Waltham, Massachusetts, USA. [22]Hoffmann-LaRoche, Nutley, New Jersey, USA. [23]Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [24]Computational Life Science Cluster (CLiC), Chemical Biology Center (KBC), Umeå University, Umeå, Sweden. [25]GlaxoSmithKline, Collegeville, Pennsylvania, USA. [26]Medical Systems Biology Research Center, School of Medicine, Tsinghua University, Beijing, China. [27]Almac Diagnostics Ltd., Craigavon, UK. [28]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [29]Department of Biological Sciences, University of Southern Mississippi, Hattiesburg, Mississippi, USA. [30]Global Pharmaceutical R&D, Abbott Laboratories, Souderton, Pennsylvania, USA. [31]National Center for Computational Toxicology, US Environmental Protection Agency, Research Triangle Park, North Carolina, USA. [32]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain. [33]HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York, USA. [34]Department of Operation Research and Financial Engineering, Princeton University, Princeton, New Jersey, USA. [35]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing, China. [36]Institute of Pharmaceutical Informatics, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China. [37]Roche Palo Alto LLC, Palo Alto, California, USA. [38]Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston, Texas, USA. [39]Department of Electrical Engineering & Computer Science, University of Kansas, Lawrence, Kansas, USA. [40]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [41]Center for Biologics Evaluation and Research, US Food and Drug Administration, Bethesda, Maryland, USA. [42]Golden Helix Inc., Bozeman, Montana, USA. [43]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. [44]SABiosciences Corp., a Qiagen Company, Frederick, Maryland, USA. [45]Cogenics, a Division of Clinical Data Inc., Morrisville, North Carolina, USA. [46]Ligand Pharmaceuticals Inc., La Jolla, California, USA. [47]GeneGo Inc., Encinitas, California, USA. [48]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. [49]Spheromics, Kontiolahti, Finland. [50]The Center for Bioinformatics and The Institute of Biomedical Sciences, School of Life Science, East China Normal University, Shanghai, China. [51]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA. [52]Rockefeller Research Laboratories, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [53]CapitalBio Corporation, Beijing, China. [54]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA. [55]SRA International (EMMES), Rockville, Maryland, USA. [56]Helwan University, Helwan, Egypt. [57]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. [58]Agilent Technologies Inc., Santa Clara, California, USA. [59]F. Hoffmann-La Roche Ltd., Basel, Switzerland. [60]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA. [61]Department of Pathology and Laboratory Medicine and HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York, USA. [62]Cedars-Sinai Medical Center, UCLA David Geffen School of Medicine, Los Angeles, California, USA. [63]Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow, Russia. [64]DNAVision SA, Gosselies, Belgium. [65]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [66]State Key Laboratory of Multi-phase Complex Systems, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China. [67]Abbott Laboratories, Abbott Park, Illinois, USA. [68]Nuvera Biosciences Inc., Woburn, Massachusetts, USA. [69]Winthrop P. Rockefeller Cancer Institute, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA. [70]VirginiaTech, Blacksburg, Virgina, USA. [71]BioMath Solutions, LLC, Austin, Texas, USA. [72]Bioinformatic Program, University of Toledo, Toledo, Ohio, USA. [73]Department of Mathematics, University of Bayreuth, Bayreuth, Germany. [74]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. [75]Pediatric Department, Stanford University, Stanford, California, USA. [76]College of Chemistry, Sichuan University, Chengdu, Sichuan, China. [77]University of Texas Southwestern Medical Center (UTSW), Dallas, Texas, USA. [78]Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain. [79]Millennium Pharmaceuticals Inc., Cambridge, Massachusetts, USA. [80]RTI International, Atlanta, Georgia, USA. [81]Takeda Global R & D Center, Inc., Deerfield, Illinois, USA. [82]Novartis Institutes of Biomedical Research, Cambridge, Massachusetts, USA. [83]W.M. Keck Center for Collaborative Neuroscience, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. [84]Entelos Inc., Foster City, California, USA. [85]Biomarker Development, Novartis Institutes of BioMedical Research, Novartis Pharma AG, Basel, Switzerland. [86]Genedata Inc., Lexington, Massachusetts, USA. [87]Affymetrix Inc., Santa Clara, California, USA. [88]Department of Chemistry and Chemical Engineering, Hefei Teachers College, Hefei, Anhui, China. [89]Institut Jules Bordet, Brussels, Belgium. [90]Biostatistics, F. Hoffmann-La Roche Ltd., Basel, Switzerland. [91]Lilly Singapore Centre for Drug Discovery, Immunos, Singapore. [92]Microsoft Corporation, US Health Solutions Group, Redmond, Washington, USA. [93]Data Analysis Solutions DA-SOL GmbH, Greifenberg, Germany. [94]Cornell University, Ithaca, New York, USA. [95]Division of Pulmonary and Critical Care Medicine, Department of Medicine, University of Toledo Health Sciences Campus, Toledo, Ohio, USA. [96]Bristol-Myers Squibb, Pennington, New Jersey, USA. [97]OpGen Inc., Gaithersburg, Maryland, USA.

ONLINE METHODS

MAQC-II participants. MAQC-II participants can be grouped into several categories. Data providers are the participants who provided data sets to the consortium. The MAQC-II Regulatory Biostatistics Working Group, whose members included a number of biostatisticians, provided guidance and standard operating procedures for model development and performance estimation. One or more data analysis teams were formed at each organization. Each data analysis team actively analyzed the data sets and produced prediction models. Other participants also contributed to discussion and execution of the project. The 36 data analysis teams listed in **Supplementary Table 3** developed data analysis protocols and predictive models for one or more of the 13 endpoints. The teams included more than 100 scientists and engineers with diverse backgrounds in machine learning, statistics, biology, medicine and chemistry, among others. They volunteered tremendous time and effort to conduct the data analysis tasks.

Six data sets including 13 prediction endpoints. To increase the chance that MAQC-II would reach generalized conclusions, consortium members strongly believed that they needed to study several data sets, each of high quality and sufficient size, which would collectively represent a diverse set of prediction tasks. Accordingly, significant early effort went toward the selection of appropriate data sets. Over ten nominated data sets were reviewed for quality of sample collection and processing consistency, and quality of microarray and clinical data. Six data sets with 13 endpoints were ultimately selected among those nominated during a face-to-face project meeting with extensive deliberations among many participants (**Table 1**). Importantly, three preclinical (toxicogenomics) and three clinical data sets were selected to test whether baseline practice conclusions could be generalized across these rather disparate experimental types. An important criterion for data set selection was the anticipated support of MAQC-II by the data provider and the commitment to continue experimentation to provide a large external validation test set of comparable size to the training set. The three toxicogenomics data sets would allow the development of predictive models that predict toxicity of compounds in animal models, a prediction task of interest to the pharmaceutical industry, which could use such models to speed up the evaluation of toxicity for new drug candidates. The three clinical data sets were for endpoints associated with three diseases, breast cancer (BR), multiple myeloma (MM) and neuroblastoma (NB). Each clinical data set had more than one endpoint, and together incorporated several types of clinical applications, including treatment outcome and disease prognosis. The MAQC-II predictive modeling was limited to binary classification problems; therefore, continuous endpoint values such as overall survival (OS) and event-free survival (EFS) times were dichotomized using a 'milestone' cutoff of censor data. Prediction endpoints were chosen to span a wide range of prediction difficulty. Two endpoints, H (CPS1) and L (NEP_S), representing the sex of the patients, were used as positive control endpoints, as they are easily predictable by microarrays. Two other endpoints, I (CPR1) and M (NEP_R), representing randomly assigned class labels, were designed to serve as negative control endpoints, as they are not supposed to be predictable. Data analysis teams were not aware of the characteristics of endpoints H, I, L and M until their swap prediction results had been submitted. If a data analysis protocol did not yield models to accurately predict endpoints H and L, or if a data analysis protocol claims to be able to yield models to accurately predict endpoints I and M, something must have gone wrong.

The Hamner data set (endpoint A) was provided by The Hamner Institutes for Health Sciences. The study objective was to apply microarray gene expression data from the lung of female B6C3F1 mice exposed to a 13-week treatment of chemicals to predict increased lung tumor incidence in the 2-year rodent cancer bioassays of the National Toxicology Program[50]. If successful, the results may form the basis of a more efficient and economical approach for evaluating the carcinogenic activity of chemicals. Microarray analysis was performed using Affymetrix Mouse Genome 430 2.0 arrays on three to four mice per treatment group, and a total of 70 mice were analyzed and used as MAQC-II's training set. Additional data from another set of 88 mice were collected later and provided as MAQC-II's external validation set.

The Iconix data set (endpoint B) was provided by Iconix Biosciences. The study objective was to assess, upon short-term exposure, hepatic tumor induction by nongenotoxic chemicals[51], as there are currently no accurate and well-validated short-term tests to identify nongenotoxic hepatic tumorigens, thus necessitating an expensive 2-year rodent bioassay before a risk assessment can begin. The training set consists of hepatic gene expression data from 216 male Sprague-Dawley rats treated for 5 d with one of 76 structurally and mechanistically diverse nongenotoxic hepatocarcinogens and nonhepatocarcinogens. The validation set consists of 201 male Sprague-Dawley rats treated for 5 d with one of 68 structurally and mechanistically diverse nongenotoxic hepatocarcinogens and nonhepatocarcinogens. Gene expression data were generated using the Amersham Codelink Uniset Rat 1 Bioarray (GE HealthCare)[52]. The separation of the training set and validation set was based on the time when the microarray data were collected; that is, microarrays processed earlier in the study were used as training and those processed later were used as validation.

The NIEHS data set (endpoint C) was provided by the National Institute of Environmental Health Sciences (NIEHS) of the US National Institutes of Health. The study objective was to use microarray gene expression data acquired from the liver of rats exposed to hepatotoxicants to build classifiers for prediction of liver necrosis. The gene expression 'compendium' data set was collected from 418 rats exposed to one of eight compounds (1,2-dichlorobenzene, 1,4-dichlorobenzene, bromobenzene, monocrotaline, *N*-nitrosomorpholine, thioacetamide, galactosamine and diquat dibromide). All eight compounds were studied using standardized procedures, that is, a common array platform (Affymetrix Rat 230 2.0 microarray), experimental procedures and data retrieving and analysis processes. For details of the experimental design see ref. 53. Briefly, for each compound, four to six male, 12-week-old F344 rats were exposed to a low dose, mid dose(s) and a high dose of the toxicant and sacrificed 6, 24 and 48 h later. At necropsy, liver was harvested for RNA extraction, histopathology and clinical chemistry assessments.

Animal use in the studies was approved by the respective Institutional Animal Use and Care Committees of the data providers and was conducted in accordance with the National Institutes of Health (NIH) guidelines for the care and use of laboratory animals. Animals were housed in fully accredited American Association for Accreditation of Laboratory Animal Care facilities.

The human breast cancer (BR) data set (endpoints D and E) was contributed by the University of Texas M.D. Anderson Cancer Center. Gene expression data from 230 stage I–III breast cancers were generated from fine needle aspiration specimens of newly diagnosed breast cancers before any therapy. The biopsy specimens were collected sequentially during a prospective pharmacogenomic marker discovery study between 2000 and 2008. These specimens represent 70–90% pure neoplastic cells with minimal stromal contamination[54]. Patients received 6 months of preoperative (neoadjuvant) chemotherapy including paclitaxel (Taxol), 5-fluorouracil, cyclophosphamide and doxorubicin (Adriamycin) followed by surgical resection of the cancer. Response to preoperative chemotherapy was categorized as a pathological complete response (pCR = no residual invasive cancer in the breast or lymph nodes) or residual invasive cancer (RD), and used as endpoint D for prediction. Endpoint E is the clinical estrogen-receptor status as established by immunohistochemistry[55]. RNA extraction and gene expression profiling were performed in multiple batches over time using Affymetrix U133A microarrays. Genomic analysis of a subset of this sequentially accrued patient population were reported previously[56]. For each endpoint, the first 130 cases were used as a training set and the next 100 cases were used as an independent validation set.

The multiple myeloma (MM) data set (endpoints F, G, H and I) was contributed by the Myeloma Institute for Research and Therapy at the University of Arkansas for Medical Sciences. Gene expression profiling of highly purified bone marrow plasma cells was performed in newly diagnosed patients with MM[57–59]. The training set consisted of 340 cases enrolled in total therapy 2 (TT2) and the validation set comprised 214 patients enrolled in total therapy 3 (TT3)[59]. Plasma cells were enriched by anti-CD138 immunomagnetic bead selection of mononuclear cell fractions of bone marrow aspirates in a central laboratory. All samples applied to the microarray contained >85% plasma cells as determined by two-color flow cytometry ($CD38^+$ and $CD45^-$/dim) performed after selection. Dichotomized overall survival (OS) and event-free survival (EFS) were determined based on a 2-year milestone cutoff. A gene expression model of high-risk multiple myeloma was developed and validated by the data provider[58] and later on validated in three additional independent data sets[60–62].

doi:10.1038/nbt.1665

The neuroblastoma (NB) data set (endpoints J, K, L and M) was contributed by the Children's Hospital of the University of Cologne, Germany. Tumor samples were checked by a pathologist before RNA isolation; only samples with ≥60% tumor content were used and total RNA was isolated from ~50 mg of snap-frozen neuroblastoma tissue obtained before chemotherapeutic treatment. First, 502 preexisting 11 K Agilent dye-flipped, dual-color replicate profiles for 251 patients were provided[63]. Of these, profiles of 246 neuroblastoma samples passed an independent MAQC-II quality assessment by majority decision and formed the MAQC-II training data set. Subsequently, 514 dye-flipped dual-color 11 K replicate profiles for 256 independent neuroblastoma tumor samples were generated and profiles for 253 samples were selected to form the MAQC-II validation set. Of note, for one patient of the validation set, two different tumor samples were analyzed using both versions of the 2 × 11K microarray (see below). All dual-color gene-expression of the MAQC-II training set were generated using a customized 2 × 11K neuroblastoma-related microarray[63]. Furthermore, 20 patients of the MAQC-II validation set were also profiled using this microarray. Dual-color profiles of the remaining patients of the MAQC-II validation set were performed using a slightly revised version of the 2 × 11K microarray. This version V2.0 of the array comprised 200 novel oligonucleotide probes whereas 100 oligonucleotide probes of the original design were removed due to consistent low expression values (near background) observed in the training set profiles. These minor modifications of the microarray design resulted in a total of 9,986 probes present on both versions of the 2 × 11K microarray. The experimental protocol did not differ between both sets and gene-expression profiles were performed as described[63]. Furthermore, single-color gene-expression profiles were generated for 478/499 neuroblastoma samples of the MAQC-II dual-color training and validation sets (training set 244/246; validation set 234/253). For the remaining 21 samples no single-color data were available, due to either shortage of tumor material of these patients (n = 15), poor experimental quality of the generated single-color profiles (n = 5), or correlation of one single-color profile to two different dual-color profiles for the one patient profiled with both versions of the 2 × 11K microarrays (n = 1). Single-color gene-expression profiles were generated using customized 4 × 44K oligonucleotide microarrays produced by Agilent Technologies. These 4 × 44K microarrays included all probes represented by Agilent's Whole Human Genome Oligo Microarray and all probes of the version V2.0 of the 2 × 11K customized microarray that were not present in the former probe set. Labeling and hybridization was performed following the manufacturer's protocol as described[48].

Sample annotation information along with clinical co-variates of the patient cohorts is available at the MAQC web site (http://edkb.fda.gov/MAQC/). The institutional review boards of the respective providers of the clinical microarray data sets had approved the research studies, and all subjects had provided written informed consent to both treatment protocols and sample procurement, in accordance with the Declaration of Helsinki.

MAQC-II effort and data analysis procedure. This section provides details about some of the analysis steps presented in **Figure 1**. Steps 2–4 in a first round of analysis was conducted where each data analysis team analyzed MAQC-II data sets to generate predictive models and associated performance estimates. After this first round of analysis, most participants attended a consortium meeting where approaches were presented and discussed. The meeting helped members decide on a common performance evaluation protocol, which most data analysis teams agreed to follow to render performance statistics comparable across the consortium. It should be noted that some data analysis teams decided not to follow the recommendations for performance evaluation protocol and used instead an approach of their choosing, resulting in various internal validation approaches in the final results. Data analysis teams were given 2 months to implement the revised analysis protocol (the group recommended using fivefold stratified cross-validation with ten repeats across all endpoints for the internal validation strategy) and submit their final models. The amount of metadata to collect for characterizing the modeling approach used to derive each model was also discussed at the meeting.

For each endpoint, each team was also required to select one of its submitted models as its nominated model. No specific guideline was given and groups could select nominated models according to any objective or subjective criteria. Because the consortium lacked an agreed upon reference performance measure (**Supplementary Fig. 13**), it was not clear how the nominated models would be evaluated, and data analysis teams ranked models by different measures or combinations of measures. Data analysis teams were encouraged to report a common set of performance measures for each model so that models could be reranked consistently a posteriori. Models trained with the training set were frozen (step 6). MAQC-II selected for each endpoint one model from the up-to 36 nominations as the MAQC-II candidate for validation (step 6).

External validation sets lacking class labels for all endpoints were distributed to the data analysis teams. Each data analysis team used its previously frozen models to make class predictions on the validation data set (step 7). The sample-by-sample prediction results were submitted to MAQC-II by each data analysis team (step 8). Results were used to calculate the external validation performance metrics for each model. Calculations were carried out by three independent groups not involved in developing models, which were provided with validation class labels. Data analysis teams that still had no access to the validation class labels were given an opportunity to correct apparent clerical mistakes in prediction submissions (e.g., inversion of class labels). Class labels were then distributed to enable data analysis teams to check prediction performance metrics and perform in depth analysis of results. A table of performance metrics was assembled from information collected in steps 5 and 8 (step 10, **Supplementary Table 1**).

To check the consistency of modeling approaches, the original validation and training sets were swapped and steps 4–10 were repeated (step 11). Briefly, each team used the validation class labels and the validation data sets as a training set. Prediction models and evaluation performance were collected by internal and external validation (considering the original training set as a validation set). Data analysis teams were asked to apply the same data analysis protocols that they used for the original 'Blind' Training → Validation analysis. Swap analysis results are provided in **Supplementary Table 2**. It should be noted that during the swap experiment, the data analysis teams inevitably already had access to the class label information for samples in the swap validation set, that is, the original training set.

Model summary information tables. To enable a systematic comparison of models for each endpoint, a table of information was constructed containing a row for each model from each data analysis team, with columns containing three categories of information: (i) modeling factors that describe the model development process; (ii) performance metrics from internal validation; and (iii) performance metrics from external validation (**Fig. 1**; step 10).

Each data analysis team was requested to report several modeling factors for each model they generated. These modeling factors are organization code, data set code, endpoint code, summary or normalization method, feature selection method, number of features used in final model, classification algorithm, internal validation protocol, validation iterations (number of repeats of cross-validation or bootstrap sampling) and batch-effect-removal method. A set of valid entries for each modeling factor was distributed to all data analysis teams in advance of model submission, to help consolidate a common vocabulary that would support analysis of the completed information table. It should be noted that since modeling factors are self-reported, two models that share a given modeling factor may still differ in their implementation of the modeling approach described by the modeling factor.

The seven performance metrics for internal validation and external validation are MCC (Matthews Correlation Coefficient), accuracy, sensitivity, specificity, AUC (area under the receiver operating characteristic curve), binary AUC (that is, mean of sensitivity and specificity) and r.m.s.e. For internal validation, s.d. for each performance metric is also included in the table. Missing entries indicate that the data analysis team has not submitted the requested information.

In addition, the lists of features used in the data analysis team's nominated models are recorded as part of the model submission for functional analysis and reproducibility assessment of the feature lists (see the MAQC Web site at http://edkb.fda.gov/MAQC/).

Selection of nominated models by each data analysis team and selection of MAQC-II candidate and backup models by RBWG and the steering committee. In addition to providing results to generate the model information table, each team nominated a single model for each endpoint as its preferred model for validation, resulting in a total of 323 nominated models, 318 of which were applied to the prediction of the validation sets. These nominated models were peer reviewed, debated and ranked for each endpoint by the RBWG before validation set predictions. The rankings were given to the MAQC-II steering committee, and those members not directly involved in developing models selected a single model for each endpoint, forming the 13 MAQC-II candidate models. If there was sufficient evidence through documentation to establish that the data analysis team had followed the guidelines of good classifier principles for model development outlined in the standard operating procedure (**Supplementary Data**), then their nominated models were considered as potential candidate models. The nomination and selection of candidate models occurred before the validation data were released. Selection of one candidate model for each endpoint across MAQC-II was performed to reduce multiple selection concerns. This selection process turned out to be highly interesting, time consuming, but worthy, as participants had different viewpoints and criteria in ranking the data analysis protocols and selecting the candidate model for an endpoint. One additional criterion was to select the 13 candidate models in such a way that only one of the 13 models would be selected from the same data analysis team to ensure that a variety of approaches to model development were considered. For each endpoint, a backup model was also selected under the same selection process and criteria as for the candidate models. The 13 candidate models selected by MAQC-II indeed performed well in the validation prediction (**Figs. 2c** and **3**).

50. Thomas, R.S., Pluta, L., Yang, L. & Halsey, T.A. Application of genomic biomarkers to predict increased lung tumor incidence in 2-year rodent cancer bioassays. *Toxicol. Sci.* **97**, 55–64 (2007).
51. Fielden, M.R., Brennan, R. & Gollub, J. A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicol. Sci.* **99**, 90–100 (2007).
52. Ganter, B. *et al.* Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* **119**, 219–244 (2005).
53. Lobenhofer, E.K. *et al.* Gene expression response in target organ and whole blood varies as a function of target organ injury phenotype. *Genome Biol.* **9**, R100 (2008).
54. Symmans, W.F. *et al.* Total RNA yield and microarray gene expression profiles from fine-needle aspiration biopsy and core-needle biopsy samples of breast carcinoma. *Cancer* **97**, 2960–2971 (2003).
55. Gong, Y. *et al.* Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. *Lancet Oncol.* **8**, 203–211 (2007).
56. Hess, K.R. *et al.* Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.* **24**, 4236–4244 (2006).
57. Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020–2028 (2006).
58. Shaughnessy, J.D. Jr. *et al.* A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276–2284 (2007).
59. Barlogie, B. *et al.* Thalidomide and hematopoietic-cell transplantation for multiple myeloma. *N. Engl. J. Med.* **354**, 1021–1030 (2006).
60. Zhan, F., Barlogie, B., Mulligan, G., Shaughnessy, J.D. Jr. & Bryant, B. High-risk myeloma: a gene expression based risk-stratification model for newly diagnosed multiple myeloma treated with high-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone. *Blood* **111**, 968–969 (2008).
61. Chng, W.J., Kuehl, W.M., Bergsagel, P.L. & Fonseca, R. Translocation t(4;14) retains prognostic significance even in the setting of high-risk molecular signature. *Leukemia* **22**, 459–461 (2008).
62. Decaux, O. *et al.* Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myelome. *J. Clin. Oncol.* **26**, 4798–4805 (2008).
63. Oberthuer, A. *et al.* Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J. Clin. Oncol.* **24**, 5070–5078 (2006).

doi:10.1038/nbt.1665

11 Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer

- Journal of Clinical Oncology, 30(12):1288–1295
- IF: 24.008
- number of citations: 79
- personal contribution (60%): method design, data preprocessing, experimental design and implementation, biomarker discovery, results analysis, manuscript writing

VOLUME 30 · NUMBER 12 · APRIL 20 2012

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Identification of a Poor-Prognosis *BRAF*-Mutant–Like Population of Patients With Colon Cancer

Vlad Popovici, Eva Budinska, Sabine Tejpar, Scott Weinrich, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Tao Xie, Fred T. Bosman, Arnaud D. Roth, and Mauro Delorenzi

See accompanying editorial on page 1255; listen to the podcast by Dr Meyerhardt at www.jco.org/podcasts

Vlad Popovici, Eva Budinska, and Mauro Delorenzi, Swiss Institute of Bioinformatics; Fred T. Bosman and Mauro Delorenzi, Lausanne University Medical Center, Lausanne; Arnaud D. Roth, Geneva University Hospital, Geneva; Arnaud D. Roth, The Swiss Group for Clinical Cancer Research, Bern, Switzerland; Sabine Tejpar and Eric Van Cutsem, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium; and Scott Weinrich, Heather Estrella, Graeme Hodgson, and Tao Xie, Pfizer, La Jolla, CA.

Submitted September 23, 2011; accepted December 20, 2011; published online ahead of print at www.jco.org on March 5, 2012.

Supported by the Swiss National Center for Competence in Research Molecular Biology (V.P. and M.D.) and Fondation Medic (E.B.); S.T. is a senior clinical investigator of the Fund for Scientific Research Flanders and has received research grants from the Belgian Federation Against Cancer and from the Belgian National Cancer Plan.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Vlad Popovici, PhD, Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Quartier Sorge, Genopode, CH-1015 Lausanne, Switzerland; e-mail: vlad.popovici@isb-sib.ch.

0732-183X/12/3012-1288/$20.00

DOI: 10.1200/JCO.2011.39.5814

ABSTRACT

Purpose
Our purpose was development and assessment of a *BRAF*-mutant gene expression signature for colon cancer (CC) and the study of its prognostic implications.

Materials and Methods
A set of 668 stage II and III CC samples from the PETACC-3 (Pan-European Trails in Alimentary Tract Cancers) clinical trial were used to assess differential gene expression between c.1799T>A (p.V600E) *BRAF* mutant and non-*BRAF*, non-*KRAS* mutant cancers (double wild type) and to construct a gene expression–based classifier for detecting *BRAF* mutant samples with high sensitivity. The classifier was validated in independent data sets, and survival rates were compared between classifier positive and negative tumors.

Results
A 64 gene-based classifier was developed with 96% sensitivity and 86% specificity for detecting *BRAF* mutant tumors in PETACC-3 and independent samples. A subpopulation of *BRAF* wild-type patients (30% of *KRAS* mutants, 13% of double wild type) showed a gene expression pattern and had poor overall survival and survival after relapse, similar to those observed in *BRAF*-mutant patients. Thus they form a distinct prognostic subgroup within their mutation class.

Conclusion
A characteristic pattern of gene expression is associated with and accurately predicts *BRAF* mutation status and, in addition, identifies a population of *BRAF* mutated-like *KRAS* mutants and double wild-type patients with similarly poor prognosis. This suggests a common biology between these tumors and provides a novel classification tool for cancers, adding prognostic and biologic information that is not captured by the mutation status alone. These results may guide therapeutic strategies for this patient segment and may help in population stratification for clinical trials.

J Clin Oncol 30:1288-1295. © 2012 by American Society of Clinical Oncology

INTRODUCTION

Activation of the *KRAS/BRAF/MEK/ERK* cascade is believed to occur frequently in colorectal (CRC) cancer on the basis of the observed 40% incidence of *KRAS* mutations and 10% to 15% incidence of *BRAF* mutations.[1-4] *KRAS* and *BRAF* mutations occur in a mutually exclusive pattern in CRC, which has long been interpreted as a sign of functional redundancy. However, these mutations occur in different histopathologic subtypes of CRC,[5,6] and we recently showed[7] that the prognosis of patients with *KRAS* and *BRAF* mutant metastatic CRC is quite different, with a clearly worse prognosis for *BRAF*-mutant disease. It has been suggested this could be due to higher levels of mitogen-activated protein kinase activation in *BRAF*-mutant (BRAFm) colon cancer.[8,9] Unlike the majority of *KRAS*-mutant (KRASm) CRCs, BRAFm metastatic CRCs do not respond to any current chemotherapy, and the outcome of patients with BRAFm CRC is similar to that of untreated patients.

Our main objective was to better understand the underlying biology of BRAFm CRCs as captured by gene expression. We developed a BRAFm gene signature that allowed an accurate identification of BRAFm samples, and which, when applied to *BRAF* wild-type samples, identified additional colon cancer (CC) samples that manifested a similar gene expression pattern. Although a substantial amount of work has been dedicated to the development of BRAFm gene expression signatures in melanoma,[10-12] to the best of our knowledge, there is no such published work in the CC context. Taking advantage of a large series of tumors with gene expression and mutation data from the PETACC-3 (Pan-European Trails in Alimentary Tract Cancers) clinical trial,[13] we studied the genes differentially expressed between c.1799T>A (p.V600E) BRAFm and double-wild-type (WT2) tumors, defined as non-*BRAF* mutant, non-*KRAS* mutant. We purposely excluded the KRASm tumors from this comparison because it was unclear whether KRASm carcinomas had overlapping biology with BRAFm. Next, we built a classifier able to recognize with high sensitivity BRAFm CCs in our own and external data sets.

When the *BRAF* classifier was applied to the whole population, it identified a *BRAF* wild-type subpopulation, with similar gene expression and prognostic characteristics. Approximately 62% of these BRAFm-like tumors were KRASm (30% of all KRASm were BRAFm-like), with the rest being WT2 (13% of all WT2). In our data, the BRAFm-like population represented 18% of CCs. This intriguing finding suggests a common biology between these tumors, not predicted by the mutation status. The results obtained show that our current classifications of tumors as *KRAS*- or *BRAF*-mutant or mitogen-activated protein kinase–active versus nonactive are inadequate to capture the whole underlying biology and clinical behavior.

MATERIALS AND METHODS

Tumor Samples and Data Preparation

Within the PETACC-3 clinical trial,[13] formalin-fixed paraffin-embedded tissue blocks were collected after cancer diagnosis and independently of future research plans, and DNA was extracted from 1,404 microdissected tissue sections. The analysis of *KRAS* exon 2 and *BRAF* exon 15 was performed by allele-specific real-time polymerase chain reaction.[7] The mutation status has been confirmed for all samples by a second analysis, using Sequenom.[14] RNA of sufficient quantity and quality was extracted from 895 samples, and gene expressions were measured on the ALMAC Colorectal Cancer DSA platform (Craigavon, Northern Ireland)—a customized Affymetrix chip with 61,528 probe sets mapping to 15,920 unique Entrez Gene IDs—in two phases (phase 1: n = 322, phase 2: n = 573). In total, 688 unique samples passed the final quality control (phase 1: n = 265 [82.3%], phase 2: n = 423 [73.8%]) and were used in subsequent analysis (Data Supplement). Of this series of CCs, 257 (37.4%) were *KRAS* mutated, whereas *BRAF* mutation was detected in 47 (6.8%) of the cases (Data Supplement).

The stage III subset included all samples for which profile data could be obtained and is thus representative of the clinical population of the trial. The stage II subset included all patients with relapse for whom profile data could be obtained and is thus also representative of this group, whereas from the nonrelapsing patients, a randomly selected population was profiled.

Three additional independent data sets[15-17] were used for validation of the signature, whereas a fourth data set,[18] with available survival information, was used for validating the prognostic value of the signature.

Statistical Analysis

PETACC-3 gene expression data were retrospectively analyzed to derive the *BRAF* gene signature discriminating between c.1799T>A (p.V600E) BRAFm and double-wild-type (WT2; *BRAF* and *KRAS* wild-type) tumors. Samples with missing mutation information (n = 39) were discarded from the gene signature development, but were included later in the survival analysis.

Gene expression data were normalized using RMA (Robust Microchip Average)[19] and summarized at the gene level by choosing the probe set with the highest standard deviation as a representative of each gene, in each data set individually.

Differentially expressed genes were obtained by fitting multivariate linear models (using LIMMA[20] package) to probe set–level data to fully exploit the potential of the platform. To account for known association between microsatellite instability-high (MSI-H), BRAFm, and right-sided tumors,[7] the linear model for the whole population included factors for *BRAF* mutation, MSI status, and tumor site (all binary variables). For the microsatellite stable (MSS) subpopulation, the model included only the *BRAF* mutation status and tumor site. The false discovery rate was controlled by Benjamini-Hochberg procedure[21] and required to be at most 1%, whereas the minimum absolute log-fold change was 0.585 (= log2 1.5). As the MSI-H subpopulation was small and consisted only of right-sided samples, the differentially expressed genes were derived by comparing BRAFm and WT2 only in the right colon, with a false discovery rate less than 25% and no constraint on the fold change.

For signature generation, an adapted version of the top scoring pairs algorithm[22] (multiple top scoring pairs [mTSP]; Data Supplement) was used, resulting in gene pairs deemed as the most informative in the process of classifier construction. The final classification model consisted of two groups of genes (G1 and G2), and the prediction was made comparing the averages of these groups: If, for a given sample, the average of G1 was smaller than the average of G2, then the sample was predicted to be BRAFm, otherwise WT2.

We also defined a *BRAF* score (BS) as the difference between the average expression of G2 genes and the average expression of G1 genes (from the mTSP model) and used it to analyze the stratification for different threshold values (a threshold of 0 leading to the original decision rule). An alternative threshold for the *BRAF* score was obtained as the value that maximized Matthews correlation coefficient[23] on the PETACC-3 data set.

The performance of the classifier was estimated by repeated (10 times) stratified five-fold cross-validation, following the MAQC-II guidelines,[24] and measured in terms of sensitivity, specificity, and error rate. The final *BRAF* classifier was built from all BRAFm and WT2 samples in the PETACC-3 data set and then applied to the full PETACC-3 data set (including KRASm) and independent validation sets for the analysis of stratification of the population (Data Supplement). Because the stage II subgroup of PETACC-3 is smaller and not fully representative, the analysis of the prognostic value of the signature is focused on stage III subgroup. However, results for both stages are given (Data Supplement).

The association between predicted class and survival outcomes was tested using Cox proportional hazard models (log-likelihood test) and log-rank test for dichotomous variables. Three survival outcomes have been considered: overall survival, relapse-free survival and survival after relapse. Fisher's exact test was used for testing differences in proportions in contingency tables.

RESULTS

BRAFm: Characteristic Genes and Classifier

In the PETACC-3 data set, we identified 314 differentially expressed probe sets between BRAFm and WT2 (see Materials and Methods for details), mapping to 223 unique EntrezGene IDs. Top 50 differentially expressed probe sets are given in Table 1, with the full table given in the Data Supplement. We also derived lists of differentially expressed genes for the MSI-H and MSS tumors separately (Data Supplement).

Using the technique of mTSP, a 32-gene pair BRAFm signature (Table 2) was obtained by training on the c.1799T>A (p.V600E) BRAFm and WT2 samples, considering all genes, whether or not they were previously identified to be differentially expressed. Its performance was estimated at a sensitivity of 95.8% and a specificity of 86.5% (Table 3). Fifty of the 64 genes of the signature were among the 223 differentially expressed genes (Data Supplement).

BRAFm-Like Tumors

To make the distinction between the true and classifier-predicted mutation status, we prefix the predictions by "pred-": pred-BRAFm denotes the samples predicted to be BRAFm, whereas pred-BRAFwt

Table 1. Top 50 Differentially Expressed Probe Sets Between c.1799T>A (p.V600E) BRAFm and WT2

Probe Set ID	Gene Symbol	Entrez GeneID	LFC	Official Full Name
ADXCRPD.7995.C1_x_at	AQP5	362	−2.91	Aquaporin 5
ADXCRIH.384.C1_s_at	REG4	83998	−2.80	Regenerating islet-derived family, member 4
ADXCRAG_BC014461_x_at	CDX2	1045	2.02	Caudal type homeobox 2
ADXCRAG_BC014461_at	CDX2	1045	1.97	Caudal type homeobox 2
ADXCRPD.10572.C1_at	HSF5	124535	1.70	Heat shock transcription factor family member 5
ADXCRAG_AK024491_s_at	SOX8	30812	−1.95	SRY (sex determining region Y)-box 8
ADXCRSS.Hs#S2988180_at	HSF5	124535	2.02	Heat shock transcription factor family member 5
ADXCRPD.7687.C1_at	TM4SF4	7104	−1.70	Transmembrane 4 L six family member 4
ADXCRAG_M14335_s_at	F5	2153	−1.18	Coagulation factor V (proaccelerin, labile factor)
ADXCRAG_AJ250717_s_at	CTSE	1510	−2.62	Cathepsin E
ADXCRAG_AJ132099_s_at	VNN1	8876	−0.93	Vanin 1
ADXCRAD_NM_025113_s_at	C13orf18	80183	1.77	Chromosome 13 open reading frame 18
ADXCRAG_NM_182510_s_at	LOC146336	146336	−1.33	Hypothetical LOC146336
ADXCRAG_BC028581_s_at	PIWIL1	9271	−0.72	Piwi-like 1 (*Drosophila*)
ADXCRAD_BX094012_s_at	SOX13	9580	−0.72	SRY (sex determining region Y)-box 13
ADXCRPDRC.4289.C1_at	RNF43	54894	1.38	Ring finger protein 43
ADXCRPD.10016.C1_at	SATB2	23314	1.82	SATB homeobox 2
ADXCRPDRC.8321.C1_s_at	TFCP2L1	29842	1.26	Transcription factor CP2-like 1
ADXCRIH.1549.C1_at	ELOVL5	60481	0.94	ELOVL family member 5, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast)
ADXCRAG_BC028581_x_at	PIWIL1	9271	−1.72	Piwi-like 1 (*Drosophila*)
ADXCRIH.1305.C1_s_at	LYZ	4069	−1.61	Lysozyme
ADXCRSS.Hs#S1405714_at	RNF43	54894	1.27	Ring finger protein 43
ADXCRSS.Hs#S3740849_at	HSF5	124535	1.21	Heat shock transcription factor family member 5
ADXCRSS.Hs#S3012761_at	HSF5	124535	1.20	Heat shock transcription factor family member 5
ADXCRAD_BM825250_s_at	TM4SF4	7104	−0.99	Transmembrane 4 L six family member 4
ADXCRPD.7300.C1_s_at	LOC388199	388199	−1.28	Proline rich 25
ADXCRIH.4080.C1_s_at	SPINK1	6690	2.09	Serine peptidase inhibitor, Kazal type 1
ADXCRAD_NM_006113_s_at	VAV3	10451	1.38	Vav 3 guanine nucleotide exchange factor
ADXCRIH.546.C1_at	GGH	8836	1.49	γ-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)
ADXCRAD_AJ709424_s_at	ABLIM3	22885	−0.65	Actin binding LIM protein family, member 3
ADXCRPDRC.1943.C1_at	AXIN2	8313	1.32	Axin 2
ADXCRAD_BG470190_s_at	CDX2	1045	0.77	Caudal type homeobox 2
ADXCRAG_XM_371238_at	TRNP1	388610	−1.03	TMF1-regulated nuclear protein 1
ADXCRAD_BU664688_s_at	SLC14A1	6563	−0.82	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)
ADXCRPD.12823.C1_s_at	SYT13	57586	−0.77	Synaptotagmin XIII
ADXCRAD_CK823169_at	ANXA10	11199	−0.80	Annexin A10
ADXCRPD.8346.C1_at	HSF5	124535	1.34	Heat shock transcription factor family member 5
ADXCRPD.15182.C1_at	MIR142	406934	0.95	MicroRNA 142
ADXCRIH.31.C9_at	LYZ	4069	−1.61	Lysozyme
ADXCRAD_BP299698_s_at	VNN1	8876	−0.96	Vanin 1
ADXCRPD.14261.C1_at	ANO1	55107	−1.12	Anoctamin 1, calcium activated chloride channel
ADXCRAG_NM_002526_at	NT5E	4907	−1.27	5′-nucleotidase, ecto (CD73)
ADXCRAD_CN404528_s_at	DCBLD2	131566	−0.76	Discoidin, CUB and LCCL domain containing 2
ADXCRAD_BM852899_at	DUSP4	1846	−0.98	Dual specificity phosphatase 4
ADXCRAD_BP376354_at	AXIN2	8313	1.27	Axin 2
ADXCRAG_U04313_s_at	SERPINB5	5268	−0.89	Serpin peptidase inhibitor, clade B (ovalbumin), member 5
ADXCRIH.482.C1_at	KLK6	5653	−0.76	Kallikrein-related peptidase 6
ADXCRAD_BM718216_s_at	TRNP1	388610	−1.16	TMF1-regulated nuclear protein 1
ADXCRAG_XM_031357_s_at	KIAA0802	23255	−0.82	KIAA0802
ADXCRPD.1115.C1_s_at	MLPH	79083	−1.32	Melanophilin

NOTE. Positive LFC indicates higher expression in WT2.
Abbreviations: LFC, log fold change; WT2, double wild type.

denotes those predicted to be *BRAF* wild type. The pred-BRAFm samples consist of true *BRAF* mutants and the subset of WT2 and KRASm samples that are positive for the signature. These tumors share a common gene expression pattern, as can be seen in Appendix Figure A1 (online only). We call the subset of *BRAF* wild-type samples that are positive for the signature BRAFm-like to distinguish them from the true BRAFm.

Having identified a population of BRAFm-like samples, we proceeded to its characterization: In the population stratification analysis of PETACC-3, approximately 30% (76 of 257) of KRASm and 13%

Table 2. 32 Pairs of Genes Defining the *BRAF* Signature

Pair	Gene 1 (G1)	Gene 2 (G2)	Pair	Gene 1 (G1)	Gene 2 (G2)
1	C13orf18	CTSE	17	VAV3	OSBP2
2	DDC	AQP5	18	CFTR	KLK10
3	PPP1R14D	REG4	19	PHYH	DUSP4
4	HSF5	RSBN1L	20	PLCB4	HOXD3
5	SATB2	RASSF6	21	ZNF141	C11orf9
6	TNNC2	CRIP1	22	PPP1R14C	CD55
7	GGH	PPPDE2	23	FLJ32063	TRNP1
8	SPINK1	PLK2	24	APCDD1	FSCN1
9	PTPRO	TM4SF4	25	ACOX1	KIAA0802
10	ZSWIM1	MLPH	26	C10orf99	PLLP
11	RNF43	RBM8A	27	MIR142	IRX3
12	CELP	SOX8	28	ARID3A	SLC25A37
13	CBFA2T2	PIWIL1	29	C20orf111	PIK3AP1
14	PTPRD	LOC388199	30	AMACR	TPK1
15	CDX2	S100A16	31	AIFM3	ZIC2
16	TSPAN6	RBBP8	32	CTTNBP2	SERPINB5

NOTE. A sample is predicted to be *BRAF* mutant if the average expression of the genes in the Gene 1 (G1) columns is lower than the average expression of genes in Gene 2 (G2) columns.

(46 of 345) of WT2 samples were BRAFm-like. The BRAFm-like samples were significantly enriched in right-sided tumors in comparison with non–BRAF-like overall and also separately for KRASm (51% were right-sided) and WT2 (63% were right-sided). There was no association with a particular *KRAS* mutation subtype. Approximately 29% of the BRAFm-like samples were MSI-H (whereas 41% of the BRAFm were MSI-H). On the other hand, 50% of the MSI-H samples were BRAFm-like, with an additional 27% being BRAFm (Data Supplement). Separate hierarchical clustering of the KRASm and WT2 subpopulations, based on the genes from the signature, showed a split between BRAFm-like and the rest of the samples (Data Supplement). The identified BRAFm-like subpopulation was further described in terms of clinicopathologic features (Data Supplement), survival rates (Table 4 and Data Supplement), and differentially expressed genes between BRAFm-like and BRAFm samples (Data Supplement). The two groups of patients were similar with respect to their clinical and pathologic parameters, with the only exceptions being age (BRAFm-like comprise more patients older than 60 years) and tumor site (56% of BRAFm-like were right-sided, whereas 77% of BRAFm are right-sided; Data Supplement).

Table 3. Performance Metrics for the *BRAF* Signature

Data Set	Sensitivity	Specificity	Error Rate
PETACC-3[13]			
%	95.78	86.52	12.41
Standard deviation	4.04	0.18	0.14
Kim,[16] n = 20			
%	100.00	54.55	25.00
No.	9/9	6/11	5/20
Koinuma,[15] n = 20			
%	100.00	72.73	15.00
No.	9/9	8/11	3/20
Cetuximab,[17] n = 94			
%	85.71	91.95	8.51
No.	6/7	80/87	8/94
Aggregated, on validation sets, n = 134			
%	96.00	86.24	11.94
No.	24/25	94/109	16/134

NOTE. PETACC-3: cross-validation estimated performance. For the other data sets, the values indicate the observed performance.
Abbreviation: PETACC-3, Pan-European Trials in Alimentary Tract Cancers.

Prognostic Value of the Classifier

The prognostic value of the *BRAF* signature was assessed in the combined stage II and III population and in the stage III only subpopulation for three end points—overall survival (OS), relapse-free survival (RFS), and survival after relapse (SAR)—within the whole population, WT2 only, and KRASm only subpopulations, respectively. To account for the known prognostic effect of the MSI status (mainly for RFS) and its association with the *BRAF* mutation, the survival analysis was also performed within the MSS population only. The small number of MSI-H samples prevented a similar analysis of the signature predictions within MSI-H. In whole population and in MSS, the BRAFm and BRAFm-like patients have shorter survival times (OS and SAR), as can be seen in Figure 1 and the Data Supplement for different stratifications. The BRAFm-likeness showed the strongest prognostic effect for SAR, for both KRASm and WT2 (in all and MSS-only samples; see Figs 1F and 1H). The corresponding hazard ratios and their 95% CIs as well as the corresponding log-rank test *P* values for each of these comparisons are summarized in Table 4.

No statistically significant difference in survival was found between the BRAFm and BRAFm-like subpopulations, even though a tendency was observed for the patients with a BRAFm-like tumor to have a slightly better prognosis than those with a BRAFm tumor.

To identify potential drivers of the prognostic effect, we assessed the prognostic value of each of the 64 genes in the signature by fitting univariate Cox regression models in the whole PETACC-3 population and in the subset of *BRAF* wild-type samples (KRASm and WT2). Most of these genes were found to be significantly associated with the SAR end point, and, for 25 of them, the association was found also in the *BRAF* wild-type subgroup. These results reveal multiple interesting genes for future studies (Data Supplement).

External Validation

The *BRAF* signature was validated on three external data sets: Koinuma,[15] Kim,[16] and an internal series of patients with cetuximab-treated stage IV disease with gene expression data from primary tumors.[17] When genes from the signature were not represented on a platform, only the complete pairs of genes were considered. The aggregated observed sensitivity was 96.0% (24 of 25 BRAFm correctly identified) and the specificity was 86.24% (94 of 109 WT2 and KRASm correctly predicted; Table 3). This confirmed the highly sensitive recognition of tumors with a BRAFm and their distinction from majority non-BRAFm tumors, whereas approximately 14% of the latter were also wrongly classified as BRAFm. The reported specificity refers to KRASm and WT2 samples that should have been labeled as BRAF wild type by the classifier. The existence of a BRAFm-like group of patients is thus confirmed in these data sets.

The prognostic value of the *BRAF* signature has been validated in all and in the stage II and III only samples from the Moffitt data set[18] for OS and SAR (RFS being only marginally significant in stage II and III). No information on *BRAF* or *KRAS* mutational status was available,

Table 4. Survival Analyses Results

Data Set	OS			RFS			SAR		
	P	HR	95% CI	*P*	HR	95% CI	*P*	HR	95% CI
PETACC-3, all									
pred-BRAFm/*pred-BRAFwt*	**.0005**	**1.67**	**1.25 to 2.25**	.2447	1.17	0.90 to 1.53	**< .001**	**2.85**	**2.06 to 3.95**
BRAFm/*BRAFwt*	**.0021**	**2.01**	**1.28 to 3.17**	.1602	1.37	0.88 to 2.12	**< .001**	**3.68**	**2.20 to 6.16**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.5196	1.16	0.74 to 1.83	.4724	1.17	0.76 to 1.78	**.0021**	**2.13**	**1.30 to 3.48**
Within WT2: BRAFm-like/*pred-BRAFwt*	.1312	1.58	0.87 to 2.87	.4866	1.20	0.72 to 2.01	**.0011**	**2.72**	**1.46 to 5.06**
PETACC-3, stage III									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**1.93**	**1.41 to 1.79**	.0455	1.34	1.00 to 1.79	**< .0001**	**3.04**	**2.15 to 4.29**
BRAFm/*BRAFwt*	**.0024**	**2.14**	**1.29 to 3.55**	.1685	1.41	0.86 to 2.32	**< .0001**	**4.53**	**2.54 to 8.07**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.1916	1.37	0.85 to 2.21	.8203	1.05	0.68 to 1.64	**.0038**	**2.09**	**1.26 to 3.46**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0365	1.90	1.03 to 3.50	.2154	1.40	0.82 to 2.40	**.0012**	**2.75**	**1.45 to 5.19**
PETACC-3, MSS									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**2.19**	**1.57 to 3.07**	.0159	1.46	1.07 to 1.99	**< .0001**	**3.16**	**2.17 to 4.59**
BRAFm/*BRAFwt*	**< .0001**	**2.91**	**1.74 to 4.88**	.0228	1.79	1.08 to 2.98	**< .0001**	**4.67**	**2.57 to 8.45**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.0511	1.59	0.99 to 2.53	.4690	1.17	0.76 to 1.82	**.0043**	**2.07**	**1.24 to 3.43**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0642	1.98	0.95 to 4.16	.3464	1.37	0.71 to 2.63	**.0001**	**4.24**	**1.89 to 9.47**
PETACC-3, MSS/stage III									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**2.27**	**1.58 to 3.25**	.0105	1.54	1.10 to 2.15	**< .0001**	**2.97**	**2.01 to 4.40**
BRAFm/*BRAFwt*	**.0024**	**2.43**	**1.35 to 4.40**	.1149	1.59	0.89 to 2.86	**< .0001**	**3.88**	**1.99 to 7.56**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.0216	1.77	1.08 to 2.89	.1765	1.37	0.87 to 2.16	**.0089**	**1.98**	**1.18 to 3.34**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0220	2.35	1.11 to 4.98	.2789	1.46	0.73 to 2.93	**< .0001**	**4.67**	**2.05 to 10.63**
Moffitt[18]									
pred-BRAFm/*pred-BRAFwt*	.0376	1.67	1.02 to 2.73	.0956	1.77	0.90 to 3.50	**.0014**	**3.78**	**1.58 to 9.04**
pred-BRAFm/*pred-BRAFwt* (stages II,III)	**.0003**	**3.22**	**1.66 to 6.26**	.0498	2.02	0.99 to 4.15	**.0017**	**3.97**	**1.58 to 9.99**
pred-BRAFm/*pred-BRAFwt* (stage III)	**.0002**	**4.26**	**1.87 to 9.69**	.0204	2.79	1.13 to 6.87	**.0028**	**4.95**	**1.58 to 15.44**
	OS			PFS			SAR		
Cetuximab,[17] MSS	*P*	HR	95% CI	*P*	HR	95% CI	*P*	HR	95% CI
pred-BRAFm/*pred-BRAFwt*				**< .0001**	**4.49**	**2.40 to 8.38**	**< .0001**	**4.58**	**2.45 to 8.56**
BRAFm/*BRAFwt*				**.0018**	**3.24**	**1.46 to 7.19**	**< .0001**	**5.72**	**2.49 to 13.12**
Within BRAFwt: BRAFm-like/*pred-BRAFwt*				**.0017**	**3.45**	**1.56 to 7.63**	**< .0001**	**3.26**	**1.47 to 7.22**

NOTE. Highly significant results ($P < .01$) are set in bold. For the Cetuximab data set, only two end points could be considered: SAR and PFS. This data set contained also only stage IV MSS patients. When the predictions are considered within KRASm or WT2 subpopulations, those samples positive for the signature are called BRAFm-like (see the Results section). The comparison is given in the first column, with the reference category in italic font.

Abbreviations: BRAFm, true *BRAF* mutant; BRAFwt, true *BRAF* wild type; HR, hazard ratio; MSS, microsatellite stable; OS, overall survival; PETACC-3, Pan-European Trails in Alimentary Tract Cancers; PFS, progression-free survival; pred-BRAFm, classifier-predicted *BRAF* mutant; pred-BRAFwt, classifier-predicted *BRAF* wild type; SAR, survival after relapse.

making it impossible to draw any conclusions on the prognostic value of the signature within the KRASm or WT2 subpopulations. The signature was confirmed to be prognostic for SAR and progression-free survival (PFS) in the cetuximab[17] data set as well (OS information was not available for this data set). The survival analysis results and the corresponding Kaplan-Meier curves are given in Table 4 and in the Data Supplement.

DISCUSSION

Our results show that for c.1799T>A (p.V600E) BRAFm tumors, a characteristic gene expression signature of high sensitivity can be identified, and this signature extends to a population of *BRAF* wild-type subgroup of colon carcinomas (BRAFm-like) sharing similar clinicopathologic and gene expression features of potential prognostic importance. The *BRAF* mutation status has been previously shown to have prognostic value in CRC,[7,25-27] both in MSS and MSI-H tumors, and this feature is also shared by our signature in the case of MSS tumors. Because of the limited number of MSI-H tumors, we could not assess its prognostic value in those samples. The BRAFm-like tumors, either KRASm or double wild type, show a similar poor prognostic in all and MSS-only samples. This effect was also independent of tumor stage.

Globally, the group of BRAFm-like tumors discovered studying the gene expression data shows clinicopathologic features more similar to the BRAFm tumors (Data Supplement) than to pred-BRAFwt. As previously described,[13,28] BRAFm tumors are found with higher frequencies in right (proximal) colon, are enriched for the MSI-H phenotype, and are of higher grade. In our study, the frequencies of high-grade were 30% in BRAFm, 20% in BRAFm-like, and 5% in pred-BRAFwt; of MSI-H, 30%, 30%, and 3%, respectively; of right-side, 75%, 55%, and 30%, respectively. The mucinous tumors are most frequently BRAFm-like (45%) and are less often BRAFm (30% *v* only 10% in pred-BRAFwt). The exception is age, for which the frequency of young patients is highest in BRAFm-like (55%) and lowest in BRAFm (35%).

From a biologic perspective, this finding supports the notion that the poor outcome of tumors with BRAFm is shared with some non–BRAF-mutated tumors, suggesting that they have common biology that drives poor survival after relapse. For the genes in the signature, the c.1799T>A (p.V600E) BRAFm tumors display a homogeneous

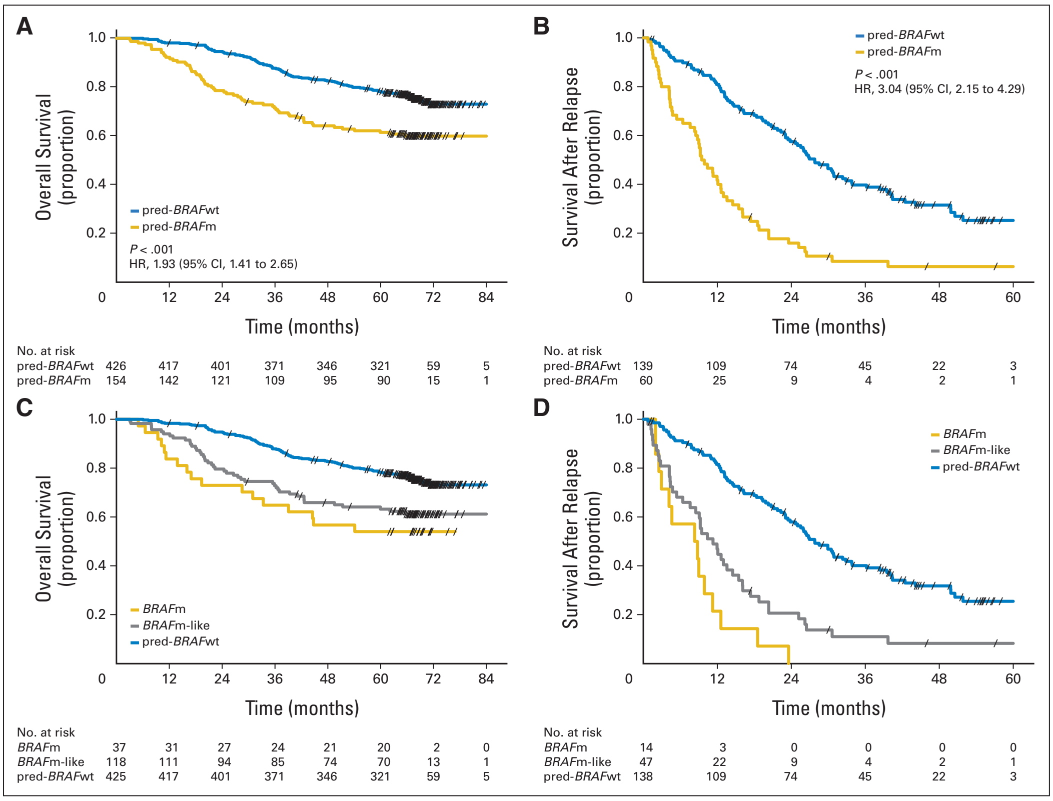


Fig 1. Kaplan-Meier curves for different stratifications of the stage III subpopulation and different end points. Columns correspond to overall survival and survival after relapse end points, respectively. Panels A-D correspond to stratifications into samples predicted to be *BRAF* mutant (pred-BRAFm)/predicted to be *BRAF* wild type (pred-BRAFwt; A, B) and *BRAF* mutant (BRAFm)/*BRAF* mutant like (BRAFm-like)/pred-BRAFwt (C, D) in the whole stage III subpopulation. Panels E-H correspond to stratifications BRAFm-like/pred-BRAFwt within *KRAS* mutant (E, F) and double wild type (WT2; G, H) subpopulations, in microsatellite stable. For the cases when only two populations are compared, the log-rank test *P* values and the hazard ratios (HRs; with 95% CIs) are given.

gene expression pattern, which is also found in some KRASm and WT2 samples (approximately 30% and 13% in our data, respectively; Appendix Fig A1). It is interesting to note that *BRAF* mutations have been strongly associated with the serrated adenoma pathway,[29,30] and thus the clear differences in gene expression between BRAFm and other colon tumors may be related to a different adenoma-carcinoma progression sequence. The existence of several subgroups of CCs, defined by their DNA methylation and mutation status, was first discovered in a population-based study[31] and was then subsequently confirmed.[32,33] A recent study[34] similarly presented evidence validating the existence of a cluster that included all BRAFm samples and a fraction of KRASm (18% of all KRASm) and WT2 samples and that was enriched for CIMP-positive, MLH1 hypermethylated, and right-sided tumors. For the moment, we can only speculate about the relation between our BRAFm-like concept and this cluster. In any case, it also supports the idea that c.1799T>A (p.V600E) BRAFm tumors form a homogeneous group with respect to the genes in the signature and that a sizeable set of other tumors show similar characteristics. The underlying driver biology of this BRAFm-like group remains unknown, although it is clearly associated with clinicopathologic features, such as MSI-H, right-sidedness, and mucinous histology.

The identification of a BRAFm-like subpopulation of CC that includes KRASm and WT2 samples and that manifests a coherent clinical behavior suggests that a new definition of CC subgroups is needed. To the best of our knowledge, this is the first reported split based on gene expression data of the KRASm tumors (see also Data Supplement), which were considered until now as a compact group, based solely on their mutation status.

The genes associated with the *BRAF* c.1799T>A (p.V600E) mutation in CC and in melanoma are dissimilar, indicating tissue-specific biology that needs to be understood and targeted differently. It is therefore not surprising that *BRAF*-specific inhibitors, such as PLX4032 or GSK2118436, although very successful in BRAFm melanoma, have failed in BRAFm colorectal cancer treatment.[35,36]

In summary, our results show that for c.1799T>A (p.V600E) BRAFm tumors, a high-sensitivity gene expression signature can be

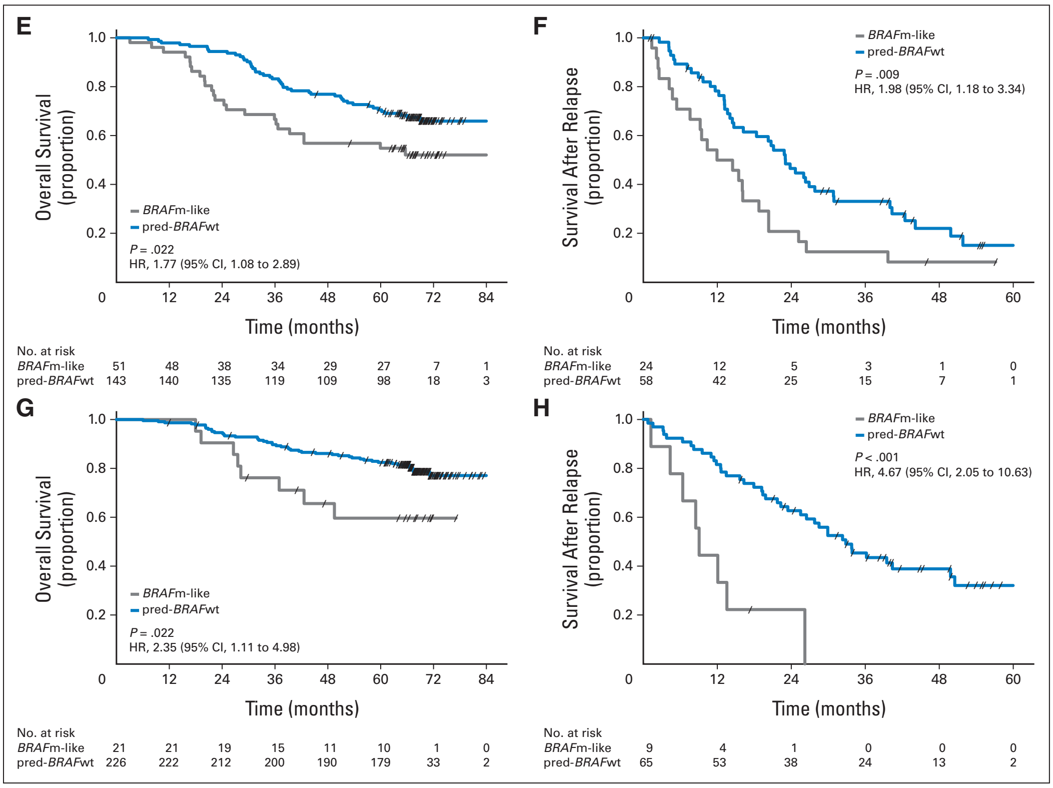


Fig 1. (continued).

derived and that this signature identifies also a subgroup of BRAFm-like tumors sharing similar clinicopathologic features of potential prognostic importance. They also indicate histologic and prognostic heterogeneity within the KRASm and thus challenge the current assumption that these tumors can all be considered alike. This stratification may be of interest in randomized clinical trials and in drug development studies and can easily be obtained by applying the proposed classifier.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: Scott Weinrich, Pfizer (C); Heather Estrella, Pfizer (C); Graeme Hodgson, Pfizer (C); Tao Xie, Pfizer (C) **Consultant or Advisory Role:** Sabine Tejpar, Pfizer (C); Fred T. Bosman, Pfizer (C); Arnaud D. Roth, Pfizer (C) **Stock Ownership:** Scott Weinrich, Pfizer; Heather Estrella, Pfizer; Graeme Hodgson, Pfizer **Honoraria:** Arnaud D. Roth, Pfizer **Research Funding:** Sabine Tejpar, Pfizer, Merck; Eric Van Cutsem, Pfizer; Fred T. Bosman, Pfizer **Expert Testimony:** None **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Vlad Popovici, Eva Budinska, Sabine Tejpar, Arnaud D. Roth, Mauro Delorenzi
Provision of study materials or patients: Eric Van Cutsem
Collection and assembly of data: Vlad Popovici, Eva Budinska, Sabine Tejpar, Scott Weinrich, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Tao Xie, Fred T. Bosman, Arnaud D. Roth
Data analysis and interpretation: Vlad Popovici, Eva Budinska, Sabine Tejpar, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Fred T. Bosman, Mauro Delorenzi
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Samowitz WS, Albertsen H, Herrick J, et al: Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology 129:837-845, 2005

2. Nosho K, Irahara N, Shima K, et al: Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. PLoS One 3:e3698, 2008

3. Brink M, de Goeij AF, Weijenberg MP, et al: K-ras oncogene mutations in sporadic colorectal cancer in The Netherlands Cohort Study. Carcinogenesis 24:703-710, 2003

4. English DR, Young JP, Simpson JA, et al: Ethnicity and risk for colorectal cancers showing somatic BRAF V600E mutation or CpG island methylator phenotype. Cancer Epidemiol Biomarkers Prev 17:1774-1780, 2008

5. Rosenberg DW, Yang S, Pleau DC, et al: Mutations in BRAF and KRAS differentially distinguish serrated versus non-serrated hyperplastic aberrant crypt foci in humans. Cancer Res 67:3551-3554, 2007

6. Velho S, Moutinho C, Cirnes L, et al: BRAF, KRAS and PIK3CA mutations in colorectal serrated polyps and cancer: Primary or secondary genetic events in colorectal carcinogenesis? BMC Cancer 8:255, 2008

7. Roth AD, Tejpar S, Delorenzi M, et al: Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: Results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. J Clin Oncol 28:466-474, 2010

8. Pratilas CA, Xing F, Solit DB: Targeting oncogenic BRAF in human cancer. Curr Top Microbiol Immunol [epub ahead of print on August 5, 2011]

9. Pratilas CA, Taylor BS, Ye Q, et al: (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. Proc Natl Acad Sci U S A 106:4519-4524, 2009

10. Dry JR, Pavey S, Pratilas CA, et al: Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). Cancer Res 70:2264-2273, 2010

11. Pavey S, Johansson P, Packer L, et al: Microarray expression profiling in melanoma reveals a BRAF mutation signature. Oncogene 23:4060-4067, 2004

12. Kannengiesser C, Spatz A, Michiels S, et al: Gene expression signature associated with BRAF mutations in human primary cutaneous melanomas. Mol Oncol 1:425-430, 2008

13. Van Cutsem E, Labianca R, Bodoky G, et al: Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. J Clin Oncol 27:3117-3125, 2009

14. De Roock W, Claes B, Bernasconi D, et al: Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: A retrospective consortium analysis. Lancet Oncol 11:753-762, 2010

15. Koinuma K, Yamashita Y, Liu W, et al: Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. Oncogene 25:139-146, 2006

16. Kim IJ, Kang HC, Jang SG, et al: Oligonucleotide microarray analysis of distinct gene expression patterns in colorectal cancer tissues harboring BRAF and K-ras mutations. Carcinogenesis 27:392-404, 2006

17. Budinska E, Delorenzi M, De Roock W, et al: New insights to gene expression signatures from primary FFPE tumors for the prediction of response to cetuximab in KRAS and BRAF wild-type colorectal cancer (CRC). J Clin Oncol 28, 243s, 2010 (suppl; abstr 3588)

18. Smith JJ, Deane NG, Wu F, et al: Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138:958-968, 2010

19. Irizarry RA, Bolstad BM, Collin F, et al: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31:e15, 2003

20. Smyth GK: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3, 2004

21. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57:289-300, 1995

22. Tan AC, Naiman DQ, Xu L, et al: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21:3896-3904, 2005

23. Matthews BW: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442-451, 1975

24. Shi L, Campbell G, Jones WD, et al: The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827-838, 2010

25. Samowitz WS, Sweeney C, Herrick J, et al: Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. Cancer Res 65:6063-6069, 2005

26. Ogino S, Nosho K, Kirkner GJ, et al: CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. Gut 58:90-96, 2009

27. French AJ, Sargent DJ, Burgart LJ, et al: Prognostic significance of defective mismatch repair and BRAF V600E in patients with colon cancer. Clin Cancer Res 14:3408-3415, 2008

28. Li WQ, Kawakami K, Ruszkiewicz A, et al: BRAF mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status. Mol Cancer 5:2, 2006

29. Snover DC: Update on the serrated pathway to colorectal carcinoma. Hum Pathol 42:1-10, 2011

30. Leggett B, Whitehall V: Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology 138:2088-2100, 2010

31. Ogino S, Kawasaki T, Kirkner GJ, et al: CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: Possible associations with male sex and KRAS mutations. J Mol Diagn 8:582-588, 2006

32. Yagi K, Akagi K, Hayashi H, et al: Three DNA methylation epigenotypes in human colorectal cancer. Clin Cancer Res 16:21-33, 2010

33. Dahlin AM, Palmqvist R, Henriksson ML, et al: The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. Clin Cancer Res 16:1845-1855, 2010

34. Hinoue T, Weisenberger DJ, Lange CP, et al: Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res [epub ahead of print on June 9, 2011]

35. Kopetz S, Desai J, Chan E, et al: PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. J Clin Oncol 28:269s, 2010 (suppl; abstr 3534)

36. Kefford R, Arkenau H, Brown MP, et al: Phase I/II study of GSK2118436, a selective inhibitor of oncogenic mutant BRAF kinase, in patients with metastatic melanoma and other solid tumors. J Clin Oncol 28:611s, 2010 (suppl; abstr 8503)

12 Identification of "BRAF-Positive" Cases Based on Whole-Slide Image Analysis

- Biomed Research International, art. no.:3926498, 2017
- IF: 2.476
- number of citations: 0
- personal contribution (80%): image analysis method design, data collection and processing, experimental design and implementation, manuscript writing

Hindawi
BioMed Research International
Volume 2017, Article ID 3926498, 7 pages
https://doi.org/10.1155/2017/3926498

Research Article

Identification of “BRAF-Positive” Cases Based on Whole-Slide Image Analysis

Vlad Popovici,[1] Aleš Křenek,[2] and Eva Budinská[3]

[1] *Institute of Biostatistics and Analyses, Faculty of Medicine and Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masarykova Univerzita, Kamenice 5, 625 00 Brno, Czech Republic*
[2] *Institute of Computer Science, Masarykova Univerzita, Šumavská 15, 602 00 Brno, Czech Republic*
[3] *Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masarykova Univerzita, Kamenice 5, 625 00 Brno, Czech Republic*

Correspondence should be addressed to Vlad Popovici; popovici@iba.muni.cz

Received 11 November 2016; Accepted 20 March 2017; Published 24 April 2017

Academic Editor: Xudong Huang

A key requirement for precision medicine is the accurate identification of patients that would respond to a specific treatment or those that represent a high-risk group, and a plethora of molecular biomarkers have been proposed for this purpose during the last decade. Their application in clinical settings, however, is not always straightforward due to relatively high costs of some tests, limited availability of the biological material and time, and procedural constraints. Hence, there is an increasing interest in constructing tissue-based surrogate biomarkers that could be applied with minimal overhead directly to histopathology images and which could be used for guiding the selection of eventual further molecular tests. In the context of colorectal cancer, we present a method for constructing a surrogate biomarker that is able to predict with high accuracy whether a sample belongs to the “BRAF-positive” group, a high-risk group comprising V600E BRAF mutants and BRAF-mutant-like tumors. Our model is trained to mimic the predictions of a 64-gene signature, the current definition of BRAF-positive group, thus effectively identifying histopathology image features that can be linked to a molecular score. Since the only required input is the routine histopathology image, the model can easily be integrated in the diagnostic workflow.

1. Introduction

The pathologic assessment of the tumor specimen provides the essential information for patient management, outcome estimation, and treatment decision. In the case of colorectal cancer (CRC), the main parameters of the pathologic assessment include the TNM stage, histologic grade, tumor type, vascular infiltration, and status of the resection margins [1]. Aside from these classical parameters, the discovery of molecular drivers and markers for resistance led to refined prognostic and predictive models [2]. For example, it has been shown that *KRAS*-mutated tumors are resistant to anti-*EGFR* treatment [3, 4]. In parallel several molecular taxonomies partially explaining intertumoral heterogeneity have been proposed for CRC [5–7]. Of interest for the current study is the identification of a high-risk group of CRC patients consisting of *V600E BRAF* mutants and a sizeable *BRAF*-wild type subset of tumors which display a similar pattern of gene activation, the so-called *BRAF*-mutant-like tumors [8]. This group is collectively called *BRAF*-positive, as the defining 64-gene signature has positive values for these cases [8]. These are only a few of the plethora of gene expression signatures proposed for CRC (in other types of cancer, the situation being similar) and they all have in common the requirement for profiling a rather large panel of genes and the limited usage in clinical practice. Among the reasons for their slow adoption are the associated costs for tests and limited availability of biological material. On the other hand, if one could robustly predict the outcome of some of these molecular tests directly from the data available for the pathologic assessment, significant speed-ups and cost cuts would be achieved. This is one of the main justifications of the present study, in which we propose an image analysis model for recognizing the "*BRAF*-positive" cases of CRC, that is, to predict the (dichotomized) outcome of the *BRAF* signature [8]. A second and broader in scope justification is the interest in identifying and understanding the connections between tumor architecture and gene activity as captured by transcriptomics.

Such connections between phenotypical appearance of the tumor and gene activity have been established before. For example, in the case of breast cancer the lobular phenotype is associated with deletions in the *CDH1* gene (encoding E-cadherin) [9] and the mesenchymal/metaplastic features are predictive in the case of *AR*-positive triple negative breast cancers [10]. In the case of colorectal cancer (CRC) the association of mucinous/serrated carcinomas with *BRAF* mutations is well known and we have shown that such association can be extended to the group of "*BRAF*-mutated-like" tumors, characterized by a specific genomic signature [8]. Similarly, connections between nuclear morphometry and molecular data have been identified in glioblastoma [11] and exploited in a multimodal prognostic signature in breast cancer [12]. When deriving molecular subtypes for colorectal cancer, we have also identified tumor architecture patterns preferentially enriched in those subtypes [5]. These observations all support the idea that genomic and phenotypic traits can be put in correspondence and, by consequence, that some phenotypic features could potentially be used as proxies for genomic markers.

In the present work, we propose an approach at building a histology image-based classifier able to predict the "*BRAF*-positive" status, as defined by the genomic signature. The gene expression data for the signature is supposed to be obtained from the same (or adjacent) tumor section as the histopathology whole-slide image. The key point of our approach resides in a convenient summarization of the imaging data into a code vector used for building the classification model. Apart from our own earlier results [13], there were no other studies to guide our selection of image features useful for this task. Hence, we took a data-driven approach in which the implicit hypothesis was that local tumor appearance contained enough information to build a predictor for the genomic "*BRAF*-positive" status. Thus, our approach was prior-free, in the sense that we did not restrict ourselves to a set of predefined (by an expert pathologist) measurements, with the potential drawback of limiting interpretability of the results.

Having a tissue-based surrogate biomarker for a genomic test allows an immediate integration in the routine diagnostic workflow and may provide the pathologist with hints for further genomic testing. This integration is supported by the increased adoption of digital pathology solutions. Additionally, such models can be applied to pathology image archives for the selection of cases for retrospective studies.

2. Materials and Methods

2.1. Data. The data collection used consisted of $n = 291$ samples for which both histopathology whole-slide images and clinical data (including *BRAF* and *KRAS* mutation status) were available, along with gene expression necessary for computing the *BRAF* score [8]. These samples were a subset of the data collected in the PETACC-3 clinical trial [14] and were selected based on the image quality and availability of the mutation information. A summary of the data is presented in Table 1 detailing the following clinical and molecular parameters, in this order: tumor stage; microsatellite stability status (high microsatellite instability (MSI-M) versus low microsatellite instability (MSI-L) or microsatellite stable (MSS)); mutation status of *BRAF* (V600E mutation) and *KRAS* (in codons 12 and 13) oncogenes; BRAF score (from the genomic signature) and the mucinous histology status of the tumor.

Table 1: Summary of main clinical parameters.

Parameter	*N*	Proportion (%)
Stage		
Stage II	55	18.9
Stage III	236	81.1
MSI		
MSI-H	12	4.1
MSI-L & MSS	279	95.6
V600E BRAF status		
Mutated	16	5.5
Wild type	275	94.5
KRAS (codons 12 and 13) status		
Mutated	113	38.8
Wild type	178	61.2
BRAF score		
Positive	59	20.3
Negative	232	79.7
Mucinous		
Yes	33	11.3
No	258	88.7

For each sample, a whole-slide image of haematoxylin-eosin (H&E) stained tumor sections was acquired at 20x magnification, using Hamamatsu NanoZoomer C9600 scanner. The resulting images were compressed by the image acquisition software using JPEG standard (at 80% quality) and stored in the proprietary NDPI format. The resolution of the images was 455 nm/pixel (equivalent to 55824 DPI) for a typical size of 100,000 × 50,000 pixels (varying with the size of the tissue section). The images were exported in standard TIFF format using OpenSlide software library [15].

2.2. Image Preprocessing. The whole-slide images were downscaled to an equivalent 5x magnification and only tumoral regions were retained from each sample (manually cut following the pathologist's annotations), the pixels outside the tumors being set to zero. To obtain the intensity signal corresponding to the haematoxylin and eosin dyes, the color deconvolution method from [16] was used, resulting in two single channel (intensity) images (H- and E-images).

2.3. Feature Extraction and Image Summarization. Our main assumption for image data modeling was that local appearance of the tissue section (local texture) contains enough information to yield discriminative features. However, the representation of an image in terms of a set of local descriptors still does not allow a direct comparison of two images (required for building a classifier); hence further summarization and standardization of the representation are needed. A suitable framework is represented by the image-retrieval applications based on Bag-of-Visual-Words methods [17]. In this framework, the local descriptors are used to construct a codebook for image representation (the information in the image is highly compressed) and the image is recoded in terms of frequencies of elements (visual codewords) from the codebook. We adapted this general approach to the problem at hand, as follows.

We decided to use a two-level approach to image representation with the first level (L1) being generic for all images and the second one (L2) specific to each class. The main reason behind this approach was that the first coding level was designed to capture the appearance of small structures (several cells, patches of stroma, parts of the colon crypts, etc.), while the second level was intended to capture larger arrangements of basic structures, which might be specific to each class. Additionally, since the classification problem was highly imbalanced, such separation would allow structures of both classes to be equally represented. Such multilevel approach has been already used in natural scene categorization [18]; however in our method we used the class label in generating the second level representation.

The first level (L1) of coding considered local patches of size 32 × 32 pixels as the basic processing unit. For such patches, we used the Gabor descriptors computed on both H- and E-images for each sample. These descriptors were based on the real component of the Gabor filter [19]:

$$G(x, y; \nu, \theta, \sigma) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \times \exp\left(2\pi\nu j\left(x\cos\theta + y\sin\theta\right)\right), \quad (1)$$

where $j = \sqrt{-1}$ and ν was the frequency, θ the orientation, and σ the bandwidth of the Gaussian kernel, respectively. The parameters were fixed throughout all experiments: $\sigma \in \{1, 2\sqrt{2}\}$, $\theta \in \{k(\pi/4) \mid k = 0, \ldots, 3\}$, and $\nu \in \{3/4, 3/8, 3/16\}$. In total, there were 24 Gabor filters that led to a 48-valued descriptor vector for each H- and E-image, with the first 24 values representing the mean response and the last 24 values representing the variance of the filter responses, over the considered 32 × 32 pixels' patch. Thus, to each local patch from the original images corresponded 96-value descriptor vectors obtained by concatenating the Gabor descriptors of the H- and E-images.

From each image in the training set (which will be generated within the cross-validation loop, see Classifier Design), 1,000 random patch descriptors were selected for building the L1 codebook using the standard k-means clustering, with $K_1 = 128$ clusters. Then, all the patches were assigned a code $1, \ldots, K_1$ based on the closest cluster (codeword) from the L1 codebook.

The second level of coding (L2) considered neighborhoods of 15 × 15 L1 patches (i.e., 480 × 480 pixels). For each such neighborhood, the descriptor computed was the vector of frequencies of the L1 codes (a vector with K_1 values). Similarly to L1 coding, a new codebook was constructed by clustering L2 descriptors (500 random L2 descriptors selected from each image) with $K_2 = 128$ clusters. Two such codebooks were constructed, one of each class (*BRAF*-positive and *BRAF*-negative), and then both used for coding each image, leading to a representation with codes $1, \ldots, 2K_2$.

The process described above led to a recoding of each image in terms of a histogram with $2K_2$ bins, each corresponding to an L2 code. We note that, in all the steps for image coding, the patches containing more than 50% of background pixels were excluded.

2.4. Classifier Design. After the image recoding step, to each image corresponded a $2K_2$-value vector which constituted the input data for the classifier design. The classifier design included the following main steps:

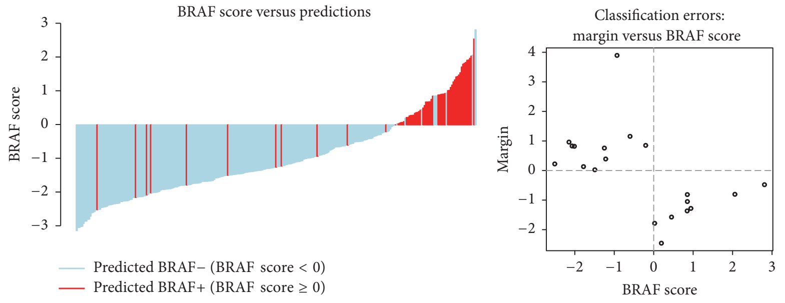
(1) Classifier feature selection: features (elements of the input vectors) were ordered based on recursive feature elimination (RFE) method [20] and subsets of features of sizes $f = 30, 50, \ldots, 130$ (approximately half of total number of features) were considered for Step (2).

(2) For each subset of features, a Support Vector Machine (SVM) [21] with Radial Basis Function (RBF) kernel was trained and its metaparameters were optimized in an inner cross-validation loop. Its performance was estimated by cross-validation and the estimated area under the ROC curve (AUC) recorded.

(3) The number of features yielding the maximum AUC was deemed optimal and the final SVM was trained on that number of features.

To estimate the performance of the system, the image recoding procedure followed by Steps (1)–(3) above was embedded into an external 10-fold stratified cross-validation loop, thus ensuring an unbiased estimation. The vector of predicted labels within this outer cross-validation was taken to represent typical predictions of the model and used in statistical analyses to avoid overly optimistic conclusions that would have been obtained from the predictions made by the model trained on the full data set.

2.5. Statistical Analyses. The main performance parameter for the classifier was AUC, but sensitivity and specificity were equally measured. For sensitivity and specificity 95% confidence intervals were computed using Agresti-Coull approximation [22] while for AUC they were obtained by bootstrap [23]. To test the association between individual image features and the class label, univariable logistic regression models were fit and the sign of the resulting coefficient was used to determine the sense of the association. To test for the association between clinical variables and classifier predictions we used χ^2-test on 2 × 2 contingency tables. Survival analysis was performed using survival package (version 2.39-4) from R statistical computing environment (version 3.3.1,



(a)

(b)

Figure 1: Analysis of the classifier's predictions. (a) Waterfall plot of the BRAF scores and the corresponding predictions (color-coded). (b) The relationship between the genomic score (x-axis) and the prediction margin (y-axis) for the misclassified samples.

Table 2: Confusion matrix for classifier predictions. The ground truth is given by the genomic signature.

	Predicted BRAF-negative	Predicted BRAF-positive
Genomic BRAF-negative	221	11
Genomic BRAF-positive	9	50

http://www.r-project.org). The estimation of hazard ratios was obtained from Cox proportional hazards regression in the absence of any other covariates, while the comparison of survival experiences of different subgroups was assessed by log-rank test (Mantel-Haenszel test). Statistical significance level was chosen to be $p = 0.05$ and no adjustment for multiple hypotheses testing was performed.

3. Results and Discussion

3.1. Image-Based Predictor. The estimated performance of the classifier was AUC = 0.938, 95% CI = (0.903–0.972), with a default operating point yielding a sensitivity Se = 0.848, 95% CI = (0.733–0.920), and a specificity Sp = 0.926, 95% CI = (0.917–0.974), corresponding to an accuracy Acc = 0.931, 95% CI = (0.896–0.956). The optimal number of features varied throughout the cross-validation iterations between 70 and 110. In Table 2, the confusion matrix from the cross-validation predictions is shown.

The relationship between the image-based classifier predictions (from cross-validation) and the genomic score can be seen in Figure 1. The misclassified samples are covering the whole range of genomic scores (Figure 1(a)). For the SVMs, the margin of a sample can be viewed as a confidence in the prediction; hence we were interested in studying the classification errors in the context of their corresponding margins. In Figure 1(b), the margins are shown as a function of genomic score. It appears that smaller margin corresponds to larger (in absolute value) *BRAF* scores indicating that the confidence in those (erroneous) predictions is rather low.

A different trade-off between sensitivity and specificity could be obtained by adapting the classifier's threshold: for example, an operating point yielding Se = 0.915, 95% CI = (0.812–0.967), and Sp = 0.776, 95% CI = (0.718–0.825), would favor the detection of *BRAF*-positives.

3.2. Relationship with Clinical Parameters. Further investigation of the classifier's errors showed that most of the false negatives were *KRAS* mutants (6 out of 9) while the majority of the false positives were double wild type (*BRAF* and *KRAS* wild type). We also note that the classifier labeled two cases (out of 16) of *BRAF* mutant tumors as "BRAF-negative"; however, one of them had also a negative genomic score. The predictions were also associated with the mucinous status of the tumors (χ^2 test p value $= 0.0066$), the microsatellite instability status (χ^2 test p value < 0.0001), and the grade (χ^2 test p value $= 0.0006$) as expected [8] but not with other clinical parameters including *KRAS* mutation status and tumor stage.

The *BRAF* genomic signature was shown to have a strong prognostic value for overall survival (OS) and survival after relapse (SAR) and limited value for relapse-free survival (RFS) [8]. In the subset of samples considered, the genomic signature maintained its prognostic value and the classifier predictions inherited, to some degree, this property: the predictions were prognostic for OS ($p = 0.007$, HR = 1.81, 95% CI = (1.17–2.81)) and SAR ($p = 0.010$, HR = 1.89, 95% CI = (1.16–3.10)) but not for RFS ($p = 0.072$, HR = 1.44, 95% CI = (0.97–2.13)).

3.3. The Predictive Image Features. We investigated the structure of the final model generated using the complete data set, on which both image recoding and the classifier design steps were applied as described above. For this model, 90 features (corresponding to codewords from the L2 codebook) were selected as the optimal set and using the logistic regression coefficient (from single-variable models) they were divided into "positive features" (preferentially present in *BRAF*-positive cases, 58 features in total, see Figure 2) and "negative

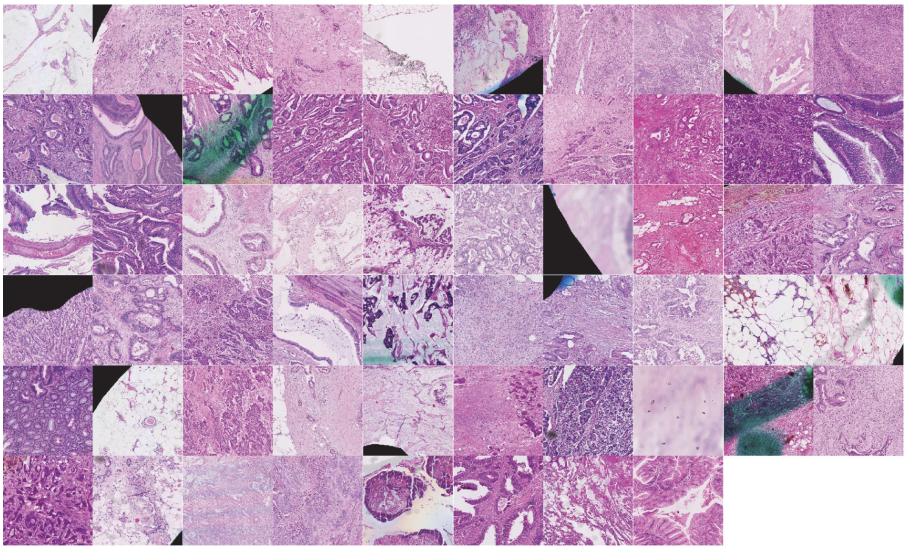


Figure 2: “Positive features”: image patterns associated with BRAF-positive class. Each feature is a 480 × 480 image patch and corresponds to an L2 codeword. Higher resolution image is available at DOI: 10.5281/zenodo.376999.

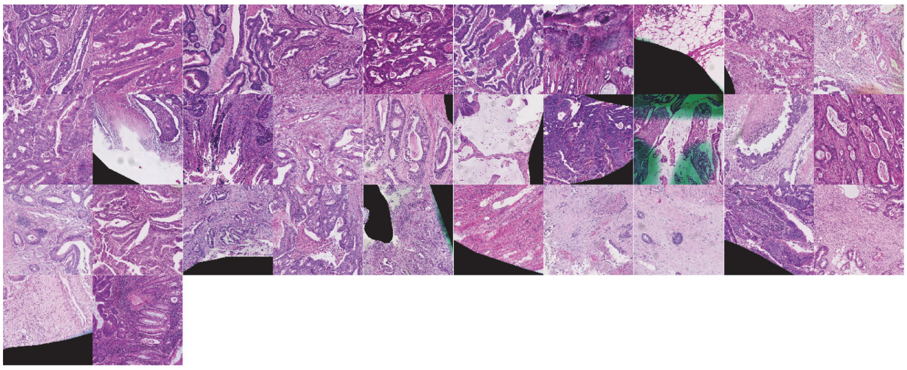


Figure 3: “Negative features”: image patterns associated with BRAF-negative class. Each feature is a 480 × 480 image patch and corresponds to an L2 codeword. Higher resolution image is available at DOI: 10.5281/zenodo.376999.

features” (preferentially present in *BRAF*-negative cases, 32 features in total, see Figure 3). We note that a number of features were dedicated to representing the border of the tumors and that some were partially affected by the markings present on the slides. It appears that the color deconvolution used in combination with Gabor descriptors made the representation robust to this type of noise. A second observation was that there were, roughly, twice as many image features representing the positive class compared to the negative one. This was to some degree not unexpected: indeed, in general, the *BRAF*-mutated and MSI-H CRC tumors show more intratumoral heterogeneity than the rest; however our results may suggest that this characteristic is common to a larger group of tumors.

The exact contribution of each feature to the final decision is less obvious as their involvement in the classifier’s prediction is through the RBF kernel and since the support vectors (actually a number of images from the training set) are

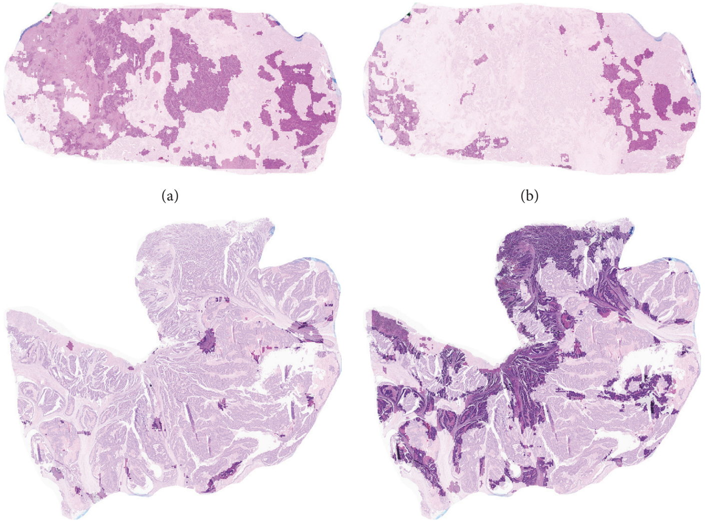


Figure 4: Spatial distribution of (positive and negative) features in two correctly classified images. The regions with low contrast were not involved in the classification process. (a-b) A *BRAF*-positive tumor: (a) positive image features; (b) negative image features. (c-d) A *BRAF*-negative tumor: (c) positive image features; (d) negative image features. Higher resolution images are available at DOI: 10.5281/zenodo.376999.

defining the separation boundary between classes. However, a visualization of their spatial distribution in images may help in qualitatively understanding the model: in Figure 4 two examples of correctly classified tumors are shown. It appears that the features identified as "positive features" cover a relatively larger region in the *BRAF*-positive tumors than the "negative features." The inverse relationship holds for the *BRAF*-negative tumors.

We also investigated whether the codebooks (for both levels of coding, L1 and L2) are biased towards one or a small group of images. We recall that the codebooks have been generated using an equal number of image patches randomly selected from the images. None of the clusters of the codebooks was dominated by a particular image, indicating that the codebooks capture general features.

4. Conclusions

We presented an image-based classifier that was able to predict with high accuracy the outcome of a genomic score. The input images were scans of H&E pathology slides making the system suitable for integration in the routine diagnostic procedures. Since the predictions of the classifier (as those of the corresponding genomic score) were not correlated with the TNM staging, they brought an independent indication of high-risk tumors (in the case of positive predictions). The system could also be applied for the retrospective selection of cases from tumor archives, reducing the volume of cases that an expert would need to evaluate.

Another important outcome is the observation that some gene expression based signatures may be translated into an image-based surrogate biomarker. Such tissue-based biomarkers may be used as a filtering step before the genomic tests.

Disclosure

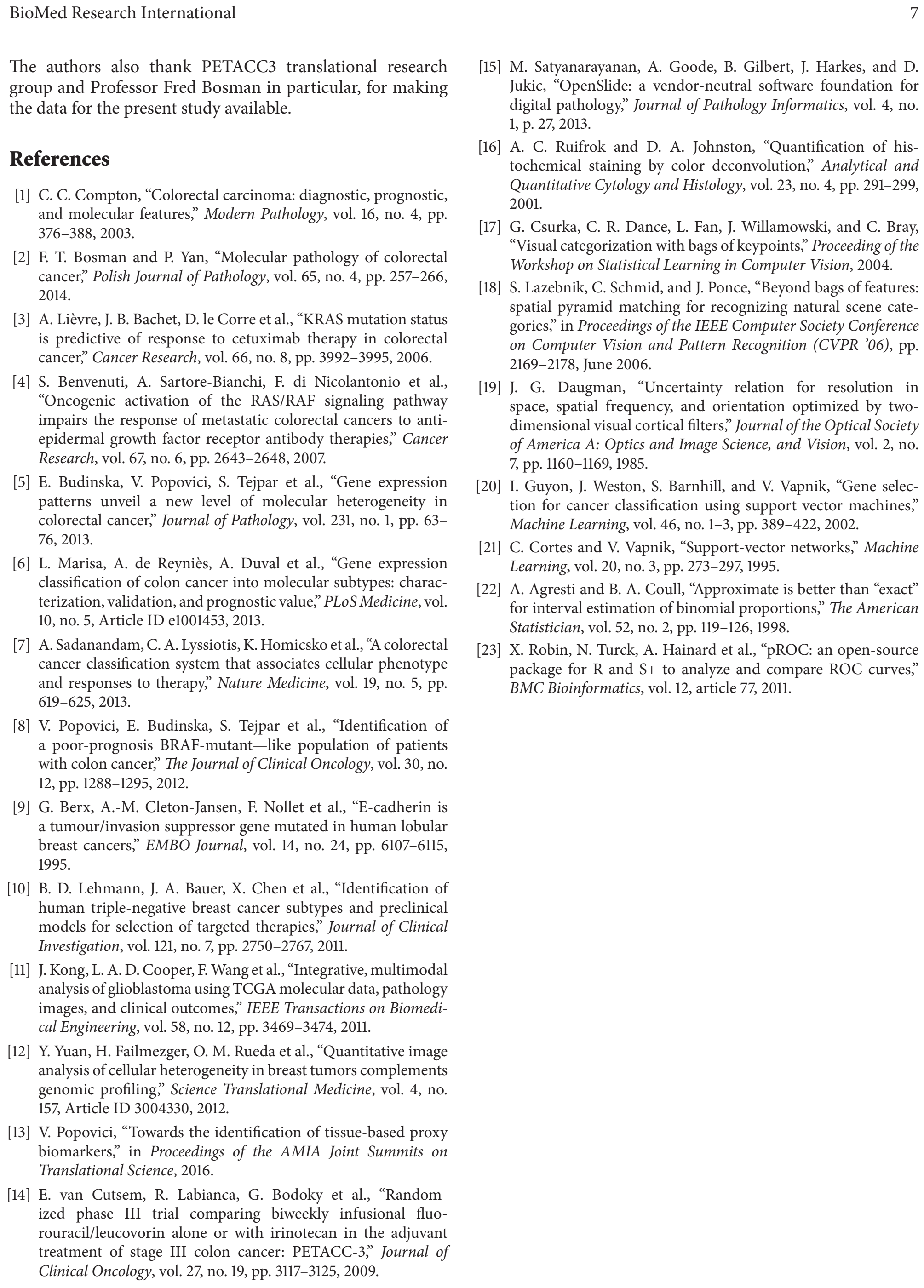
This article reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project is financed by the SoMoPro II programme. The research leading to this result has acquired a financial grant from the People Programme (Marie Curie action) of the Seventh Framework Programme of EU according to the REA Grant Agreement no. 291782. The research is further cofinanced by the South Moravian Region. The necessary computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085 projects under the programme "Projects of Projects of Large Research, Development, and Innovations Infrastructures."



The authors also thank PETACC3 translational research group and Professor Fred Bosman in particular, for making the data for the present study available.

References

[1] C. C. Compton, "Colorectal carcinoma: diagnostic, prognostic, and molecular features," *Modern Pathology*, vol. 16, no. 4, pp. 376–388, 2003.

[2] F. T. Bosman and P. Yan, "Molecular pathology of colorectal cancer," *Polish Journal of Pathology*, vol. 65, no. 4, pp. 257–266, 2014.

[3] A. Lièvre, J. B. Bachet, D. le Corre et al., "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer," *Cancer Research*, vol. 66, no. 8, pp. 3992–3995, 2006.

[4] S. Benvenuti, A. Sartore-Bianchi, F. di Nicolantonio et al., "Oncogenic activation of the RAS/RAF signaling pathway impairs the response of metastatic colorectal cancers to anti-epidermal growth factor receptor antibody therapies," *Cancer Research*, vol. 67, no. 6, pp. 2643–2648, 2007.

[5] E. Budinska, V. Popovici, S. Tejpar et al., "Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer," *Journal of Pathology*, vol. 231, no. 1, pp. 63–76, 2013.

[6] L. Marisa, A. de Reyniès, A. Duval et al., "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value," *PLoS Medicine*, vol. 10, no. 5, Article ID e1001453, 2013.

[7] A. Sadanandam, C. A. Lyssiotis, K. Homicsko et al., "A colorectal cancer classification system that associates cellular phenotype and responses to therapy," *Nature Medicine*, vol. 19, no. 5, pp. 619–625, 2013.

[8] V. Popovici, E. Budinska, S. Tejpar et al., "Identification of a poor-prognosis BRAF-mutant—like population of patients with colon cancer," *The Journal of Clinical Oncology*, vol. 30, no. 12, pp. 1288–1295, 2012.

[9] G. Berx, A.-M. Cleton-Jansen, F. Nollet et al., "E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers," *EMBO Journal*, vol. 14, no. 24, pp. 6107–6115, 1995.

[10] B. D. Lehmann, J. A. Bauer, X. Chen et al., "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, 2011.

[11] J. Kong, L. A. D. Cooper, F. Wang et al., "Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12, pp. 3469–3474, 2011.

[12] Y. Yuan, H. Failmezger, O. M. Rueda et al., "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Science Translational Medicine*, vol. 4, no. 157, Article ID 3004330, 2012.

[13] V. Popovici, "Towards the identification of tissue-based proxy biomarkers," in *Proceedings of the AMIA Joint Summits on Translational Science*, 2016.

[14] E. van Cutsem, R. Labianca, G. Bodoky et al., "Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3," *Journal of Clinical Oncology*, vol. 27, no. 19, pp. 3117–3125, 2009.

[15] M. Satyanarayanan, A. Goode, B. Gilbert, J. Harkes, and D. Jukic, "OpenSlide: a vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 27, 2013.

[16] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[17] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *Proceeding of the Workshop on Statistical Learning in Computer Vision*, 2004.

[18] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.

[19] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 2, no. 7, pp. 1160–1169, 1985.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.

[23] X. Robin, N. Turck, A. Hainard et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, article 77, 2011.

13 A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency

- Journal of Pathology, 228(4):586-595, 2012
- IF: 6.894
- number of citations: 16
- personal contribution (10%): data preprocessing, experimental design, data analysis and results interpretation, manuscript writing

Journal of Pathology
J Pathol 2012
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.4092

ORIGINAL PAPER

A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency#

Sun Tian,[1] Paul Roepman,[1] Vlad Popovici,[2] Magali Michaut,[3] Ian Majewski,[3] Ramon Salazar,[4] Cristina Santos,[4] Robert Rosenberg,[5] Ulrich Nitsche,[5] Wilma E Mesker,[6] Sjoerd Bruin,[3] Sabine Tejpar,[7] Mauro Delorenzi,[2,8,9] Rene Bernards[1,3] and Iris Simon[1]*

[1] *Agendia NV, Amsterdam, The Netherlands; and Agendia Inc., Irvine, CA, USA*
[2] *Swiss Institute for Bioinformatics, Lausanne, Switzerland*
[3] *Netherlands Cancer Institute, Amsterdam, The Netherlands*
[4] *IDIBELL, Institut Catala d'Oncologia, L'Hospitalet de Llobregat, Barcelona, Spain*
[5] *Klinikum Rechts der Isar, Technische Universität München, Germany*
[6] *Leiden University Medical Centre, Leiden, The Netherlands*
[7] *University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Belgium*
[8] *Département de Formation et Recherche, Centre Hospitalier Universitaire Vaudois, France*
[9] *University of Lausanne, Switzerland*

**Correspondence to: I Simon, Agendia NV, Amsterdam, The Netherlands; OR Agendia Inc. Irvine, CA, USA. e-mail: iris.simon@agendia.com*

#Array data have been deposited on: http://research.agendia.com/

Abstract

Microsatellite instability (MSI) occurs in 10–20% of colorectal tumours and is associated with good prognosis. Here we describe the development and validation of a genomic signature that identifies colorectal cancer patients with MSI caused by DNA mismatch repair deficiency with high accuracy. Microsatellite status for 276 stage II and III colorectal tumours has been determined. Full-genome expression data was used to identify genes that correlate with MSI status. A subset of these samples ($n = 73$) had sequencing data for 615 genes available. An MSI gene signature of 64 genes was developed and validated in two independent validation sets: the first consisting of frozen samples from 132 stage II patients; and the second consisting of FFPE samples from the PETACC-3 trial ($n = 625$). The 64-gene MSI signature identified MSI patients in the first validation set with a sensitivity of 90.3% and an overall accuracy of 84.8%, with an AUC of 0.942 (95% CI, 0.888–0.975). In the second validation, the signature also showed excellent performance, with a sensitivity 94.3% and an overall accuracy of 90.6%, with an AUC of 0.965 (95% CI, 0.943–0.988). Besides correct identification of MSI patients, the gene signature identified a group of MSI-like patients that were MSS by standard assessment but MSI by signature assessment. The MSI-signature could be linked to a deficient MMR phenotype, as both MSI and MSI-like patients showed a high mutation frequency (8.2% and 6.4% of 615 genes assayed, respectively) as compared to patients classified as MSS (1.6% mutation frequency). The MSI signature showed prognostic power in stage II patients ($n = 215$) with a hazard ratio of 0.252 ($p = 0.0145$). Patients with an MSI-like phenotype had also an improved survival when compared to MSS patients. The MSI signature was translated to a diagnostic microarray and technically and clinically validated in FFPE and frozen samples.

Keywords: Colorectal cancer; microsatellite instability; deficient mismatch repair system; gene expression; mutation frequency; genomic signature; prognosis

Received 25 June 2012; Revised 14 August 2012; Accepted 15 August 2012

Sun Tian, Paul Roepman, Rene Bernards and Iris Simon are employees of Agendia (the company that developed ColoPrint and the MSI signature).

Introduction

There are at least two recognized pathways of colorectal carcinogenesis [1]. The most common pathway is a progressive model that involves stepwise accumulation of genetic alterations in several key oncogenes and tumour suppressor genes, such as KRAS, BRAF, TP53 and, importantly, the adenomatous polyposis coli (APC) gene [2,3]. These tumours account for approximately 85% of all sporadic disease and commonly display a chromosomal instability (CIN) phenotype that is associated with widespread structural alterations. A second class of colon tumours manifests a microsatellite instability (MSI) phenotype;

these tumours typically display various insertions or deletions, most commonly in short tandem repeats, the so-called microsatellites [4]. MSI is the molecular fingerprint of a deficient mismatch repair system. Approximately 15% of colorectal cancers (CRCs) display MSI, owing either to epigenetic silencing of MLH1 or to somatic or germline mutations in one of the mismatch repair genes MLH1, MLH3, MSH2, MSH6 or PMS2 [5]. Consequently, the MSI phenotype is also referred to as the deficient MMR (dMMR) phenotype. MSI rates vary with tumour stage and, in the adjuvant setting, MSI patients have been associated with longer survival than patients with microsatellite-stable (MSS) tumours [6,7]. The deficiencies in MMR genes lead to loss of function of tumour suppressor genes and are associated with activating mutations in oncogenes such as BRAF [8].

Patients with MSI cancers might have different responses to chemotherapy compared to MSS patients [1,9]. The MMR involves the recognition and repair of incorrectly paired nucleotides during DNA replication. 5-Fluorouracil (5-FU)-based chemotherapy is the standard treatment for stage II and III CRCs after surgery, and the survival advantage associated with this treatment is about 10% [10]. Data from patients with MSI and from cell lines with dMMR indicate that MSI promotes resistance to 5-FU treatment [1]. However, results from clinical studies are conflicting. It seems that MSI patients with stage II cancer have no benefit from 5-FU treatment [11,12], while stage III MSI patients might benefit from treatment, but this is predominantly seen in patients that have a germline predisposition [13]. Evidence supporting the preferential efficacy of irinotecan in MSI tumours continues to emerge, but are still considered preliminary [14]. Other studies have shown that MSI colorectal cancer might be specifically sensitive to compounds inhibiting the phosphatidylinositol 3-kinase (PI3K)–AKT–mammalian target of rapamycin (mTOR) pathway [15].

Considering the different prognosis and treatment response of MSI patients when compared to MSS patients, an accurate diagnosis is needed to facilitate appropriate treatment decisions. Today, several methods for the detection of MSI status are used. MSI can be detected by PCR amplification of specific microsatellite repeats. The presence of instability is determined by comparing the length of nucleotide repeats in tumour cells and normal cells. A consensus conference established a panel of microsatellite markers with appropriate sensitivity and specificity to diagnose MSI [16]. This reference panel, known as the Bethesda panel, included five microsatellite loci: two mononucleotides (Bat25 and Bat26) and three dinucleotides (D5s346, D2s123 and D17s250) [17]. Immunohistochemical analysis of MMR proteins is an alternative method to detect MSI in the clinical setting and complements the genetic testing of Lynch syndrome [18]. Lack of expression of one or more of the MMR proteins is indicative of deficient MMR, and can help to determine which gene harbours a germline mutation or has been inactivated by another mechanism. However, traditional methods for determining MSI status might not identify all patients with a deficient mismatch repair system and other methods might be required for a more comprehensive detection [19].

As demonstrated by others [15,20] and in this paper, patients with MSI have a very distinct gene expression pattern that allows the development of strong gene expression signatures. Pairwise comparisons between studies showed that 94–98% of genes have consistent changes in expression, even though samples were analysed on different platforms and in different studies [20]. Here we describe the development and validation of a robust gene expression signature that identifies patients with MSI status, determined by standard methods (PCR, IHC) with high accuracy, and additionally identifies a group of MSS patients with a MSI-like phenotype. The signature was translated into a diagnostic test that can be used in fresh or FFPE material and can be performed in combination with other gene expression signatures [21,22] for further classification of early-stage colon cancer patients.

Methods

Patients and samples

In this study, microsatellite instability was assessed in three patient cohorts that have been described previously: a development cohort (A) [22], a first independent validation cohort (B) [23] and a second independent cohort in the subset of the PETACC-3 gene expression dataset with complete MSI status information (cohort D) [24–26]. The prognostic value of the developed MSI signature was assessed on cohort B combined with an additional set of samples with patient follow-up data but without hospital-based MSI assessment (cohort C). Patient and sample characteristics are shown in Table 1. All tissue samples were collected from patients with appropriate informed consent. The study was carried out in accordance with the ethical standards of the Helsinki Declaration and was approved by the medical ethical boards of the participating medical centres and hospitals.

Hospital-based assessment for microsatellite instability (MSI)

For the development cohort (cohort A), fresh-frozen tumour samples from patients with colorectal cancer were collected (n = 276; Table 1). For 90 patients, 5 µm slides were immunohistochemically stained for the markers MLH1 and PMS2. For the remaining 186 patients and for all patients in validation cohort B (n = 132; Table 1) the MSI/MSS status was assessed by PCR amplification, following the standard protocol of the hospital and described in [21,22,26] and in Supplementary methods (see Supplementary material).

Table 1. Patient characteristics

Cohorts	A	B	C	D	Total
	Development	Validation	Validation (prognosis)	Validation	
Patients (*n*)	276	132	131	625	1164
Tissue type	Fresh	Fresh	Fresh	FFPE	
Age					
< 70	157	84	60	529	830
≥ 70	119	48	71	96	334
Stage					
I	40	–	–	–	40
II	157	132	131	104	524
III	78	–	–	521	599
IV	1	–	–	–	1
Gender					
Male	165	74	66	382	687
Female	111	58	65	243	477
Location					
Left colon	143	76	56	391	666
Right colon	96	56	57	234	443
Rectum	37	–	10	–	47
Not available	–	–	8	–	8
Grade					
1	83	1	21	–	105
2	172	90	87	567*	916*
3	20	41	21	55*	137*
Not available	1	–	2	3	6
BRAF					
Activating mutation	24	18	13	46	101
Wild-type/unknown mutation	248	86	92	577	1003
Not available	4	28	26	2	60
Microsatellite stability					
MSI	29	31	–	70	130
MSS	247	101	–	555	903
Not available	–	–	131	–	131

* The PETACC3 dataset dichotomized the grade information by grouping stages 1 and 2, and 3 and 4, respectively.

Patients who had at least two microsatellite unstable markers were defined as MSI. A tumour with only normal markers was defined as microsatellite-stable (MSS). MSI assessment of the PETACC-3 samples (cohort D) was performed as described previously, using a standard panel of 10 mononucleotide and dinucleotide microsatellite loci by PCR amplification of normal/tumour DNA pairs [26]. Irregularity in one marker (two in the PETACC-3 study) was defined as low-grade microsatellite instability (MSI-L); irregularity in more markers was defined as high-grade microsatellite instability (MSI) [27]. Patients with MSI-L were classified as MSS for all analysis.

Development and validation of a 64-gene signature associated with MSI status

RNA extraction, T7-based linear amplification, Cy-dye labelling and hybridization to Agilent arrays was performed as described previously [22]. All tumour samples contained > 30% tumour cells. Samples were analysed against a common reference that was generated using a pool of 44 CRC samples. Gene expression measurements were normalized (Lowess normalization) and log-ratios were used for identification of genes that were associated with the MSI status of the tumours (based on two-sided Student's t-test). We used a 10-fold cross-validation (CV10) procedure that has been described previously [22,28]. The CV10 procedure was applied on the development cohort ($n = 276$) and repeated 1000 times to determine classification performance and for robust gene selection. During each CV10 round, genes were ranked by p value. The 64 genes (see Supplementary material, Table S1) with the highest frequency of appearance within the top-ranking genes in each of the 1000 CV loops were selected as the final set with the strongest MSI association (http://research.agendia.com/).

The 64 gene set was used to construct a nearest centroid-based classification method (cosine correlation); a MSI gene signature index for the individual samples was defined as the difference of the two correlations. Samples were classified within the MSI group if their index exceeded a predefined optimized threshold. This threshold was determined to reach a maximal overall accuracy (sum of sensitivity and specificity).

The 64-gene signature was validated on 132 independent CRC samples analysed in the same way as the development cohort, using the same microarray platform and threshold (cohort B, Table 1). Samples were classified as MSI if their index (the difference of the two correlations) exceeded the predefined optimized threshold. A second validation was performed on data from the PETACC-3 study comprising 625 colon tumour FFPE samples with known MSI status,

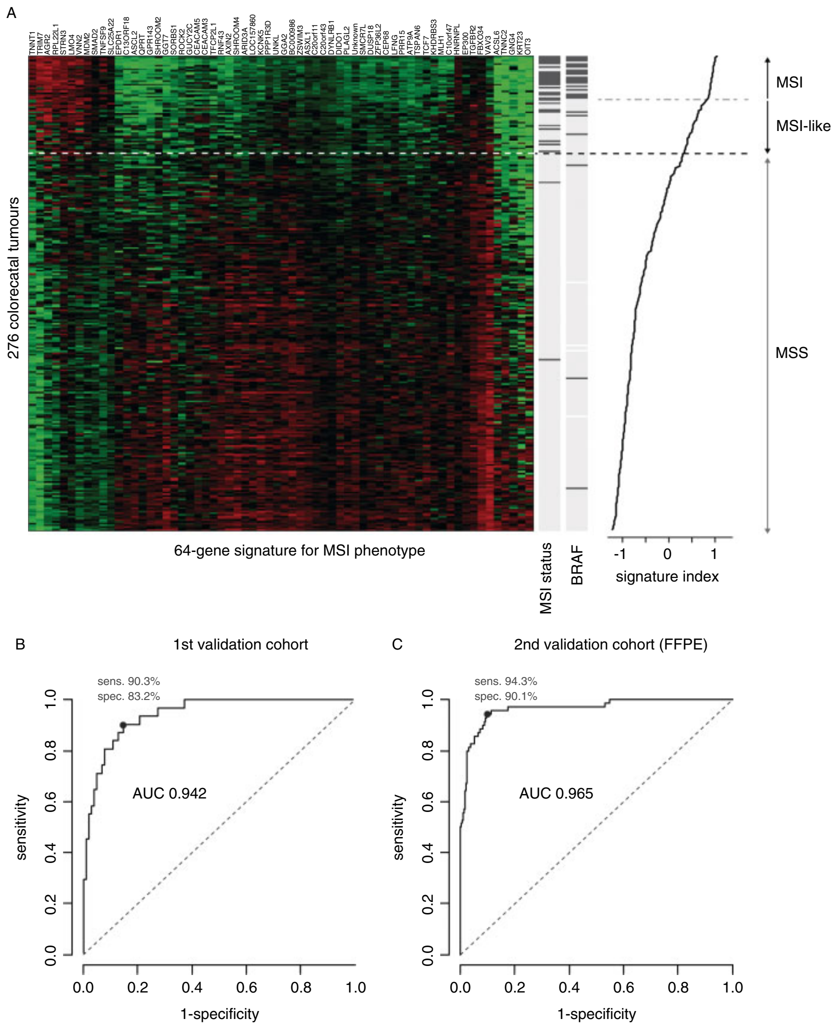


Figure 1. (A) A 64-gene expression signature for identification of colorectal cancer samples with MSI, MSI-like and MSS phenotypes. The MSI signature read-outs (index) are shown for 276 tumour samples (cohort A): red, relative up-regulation; green, down-regulation. Standard hospital-based MSI assessment is indicated in the middle bars, together with the *BRAF* V600E mutation status: light grey, MSS or *BRAF* wild-type, dark grey, MSI or *BRAF* mutation. (B) ROC curve and AUC of the signature read-out on validation cohort B. (C) ROC curve and AUC of the MSI signature on validation cohort D (PETACC-3 study). The optimal sensitivity and specificity (with a sensitivity of at least 0.9 and sum of sensitivity and specificity is maximal) is indicated in grey.

of which 70 (11.2%) were MSI (cohort D, Table 1). As described previously [25], these 625 samples had been hybridized to a custom Affymetrix platform optimized for analysis of degraded RNA in FFPE samples. We could identify 58 of the 64 MSI signature genes. Read-out of the MSI gene signature index on the Affymetrix data was done in a similar fashion as for the first validation cohort. A receiver operating characteristic (ROC) curve was plotted and the area under the ROC curve (AUC) was calculated. Sensitivity and specificity were calculated based on the optimal overall accuracy, with a sensitivity of at least 90%.

Besides the main binary classification of MSS and MSI samples, a secondary threshold was determined to subclassify MSI-like samples that were positive by MSI gene expression signature but typically classified as MSS by hospital assessment. Both thresholds for MSI and MSI-like classification were determined using the development cohort A only and are indicated in Figure 1A.

Functional analysis of 64-gene signature

Functional analysis of the genes in the signature was performed by using the Database for Annotation, Visualization and Integrated Discovery (DAVID) software, v 6.7 [29]. The enriched functional annotation clusters were calculated by DAVID through grouping enriched functional terms. The parameter set used had a similarity threshold of 0.4, multiple linkage threshold of 0.3 and an EASE parameter of 0.5. Only clusters larger than three functional terms were used.

Investigation of mutation frequency

DNA fragment libraries were prepared using the TruSeq DNA Sample Preparation Kit (Illumina) and were hybridized to the SureSelect Human Kinome bait library according to the manufacturer's protocol (Agilent). Captured DNA samples were sequenced on a HiSeq 2000 (Illumina), using a 55 bp paired-end protocol. Sequence reads were aligned to the human genome (GRCh37/hg19) and unique pairs were used for variant calling. Candidate variants were identified using SAMtools and the following inclusion criteria were applied: minimum coverage 10; minimum variant count 5; a variant must be detected on both strands. Variants were assessed using the Ensembl variant effect predictor (v 62) to define those that were likely to impact protein coding sequences and to filter out germline polymorphisms. Matched germline DNA was sequenced for 19 of the 73 tumour samples and an additional 56 normal samples were used to improve the removal of germline SNPs and sequencing errors. In this paper we focus on mutation load; a full analysis of the sequence alterations is the subject of another study.

Statistical and survival analysis

All analyses and statistical tests were performed in Matlab (MathWorks) or R (v 2.14.1; www.r-project.org). All tests were two-sided and the significance level of p values was set to be 0.05. Survival analysis was performed on cohorts B and C combined, using Cox proportional hazard models with 10-year distant metastasis-free survival (dmfs) as end point.

Results

Development of an MSI signature

A cohort of 276 colorectal tumour samples (cohort A, Table 1) was analysed for their microsatellite status [microsatellite instability (MSI) or stability (MSS)] according to the local standard methodology at the originating hospital (see Methods for details); 11% (n = 29) of the tumours were identified as MSI (Table 1). This cohort was used for identification of genes with expression strongly associated with MSI status. Using a 10-fold cross-validation procedure, we identified a set of 64 genes (see Supplementary material, Table S1) that formed the basis of a single sample-based classifier to accurately identify MSI tumours (Figure 1A). Optimal accuracy was reached upon classification of 57 samples as MSI by the signature and 219 samples as MSS, corresponding to a sensitivity of 93.1% and a specificity of 87.9% (Table 2).

The 64-gene signature was validated in an independent cohort of 132 stage II colon cancer samples (validation cohort B, Table 1) that was analysed using the single sample predictor (SSP), as established in the development cohort. Performance in the validation samples showed an area under the ROC curve (AUC) of 0.942 (95% CI, 0.888–0.975) with a sensitivity of 90.3% and a specificity of 83.2% when applying the established threshold for MSS and MSI classification (Figure 1B, Table 2).

A second independent validation of gene signature was performed on a prospective cohort of FFPE tissue samples from the randomized PETACC-3 study (cohort D, Table 1) [24]. Signature read-out in the PETACC-3 samples showed a very high concordance with hospital-based MSI assessment, with an ROC of 0.965 (95% CI, 0.943–0.988), which has an optimal sensitivity of 94.3% and specificity of 90.1% (Figure 1C, Table 2). Besides validating the signature in an independent prospective study, this result showed that the developed 64-gene signature can be successfully translated to a different microarray platform and can likely be used for MSI assessment on FFPE samples.

MSI signature and mutation frequency

In all patient cohorts, the MSI signature was able to correctly identify nearly all MSI patients (sensitivity above 92%) but they were classified as MSI by the gene signature (Figure 1A). We hypothesized that, although these MSI-like tumours were assessed as MSS by standard methods, they do have a deficient mismatch repair (dMMR)-related phenotype. As such, the developed gene signature might be able to identify MSI samples but also MSS samples that harbour a dMMR phenotype (MSI-likes).

To test this hypothesis, we have deep-sequenced 73 tumour samples for their 'cancer kinome' (615 genes in total). The sequencing results confirmed that samples identified as MSI by the gene signature have a significantly higher mutation frequency (on average, candidate mutations were identified in 7.4% of the analysed genes) compared to MSS samples (on average, candidate mutations were identified in 1.6% of the genes) (Student's t-test, p = 3.15e-12). Importantly, further classification into MSI and MSI-like samples indicated that the mutation frequency of the MSI-like tumours (average 6.4%) is also significantly higher than that of MSS samples (Student's t-test, p = 6.26e-6) and comparable to the mutation frequency in MSI samples

Table 2. Performance of MSI gene signature: performance of MSI and MSS classification by the 64-gene signature compared to standard local hospital methodology

	Tissue	*n*	Sensitivity	Specificity	Overall accuracy
Development cohort A	Fresh	276	93.1	87.9	88.4
Validation cohort B	Fresh	132	90.3	83.2	84.8
Validation cohort D (PETTAC-3)	FFPE	625	94.3	90.1	90.6

(8.2%) (Figure 2). This result suggests that MSI-like tumours also harbour a dMMR phenotype, resulting in a higher mutation rate.

It is important to note that the MSI-like patients, as identified by the signature, were not patients with a low-grade MSI (MSI-low) assessment by the hospital (data not shown), confirming that the MSI-likes might be a subclass that cannot be identified by standard MSI assessment.

Investigation of activating mutations in *BRAF* showed that 64.3% of all samples classified as MSI by the gene signature harboured an activating *BRAF* mutation (36 of 56 samples with a known *BRAF* mutation status). In the MSI-like class, 17.4% of the samples had an activating *BRAF* mutation, while the MSS classified samples were almost exclusively (98.0%) *BRAF* wild-type (342 of 349 samples).

Functional annotation

The association between the MSI gene signature and a dMMR phenotype was further supported by functional analysis. The results indicated that four functional annotation clusters were significantly enriched in the 64 signature genes (see Methods; see also Supplementary material, Tables S1 and S2). Annotation cluster 1 indicated that the encoded proteins of the signature are enriched with zinc-finger domain proteins, which are often found as part of transcription, translation, DNA replication and repair machineries [30]. Together with the enrichment in functional terms related to DNA binding and the nucleic acid metabolic processes (annotation cluster 2), these results are in agreement with the nature of DNA mismatch repair proteins as DNA interacting/metabolism partners that often form large complexes in the nucleus (annotation cluster 4) [31]. In addition, annotation cluster 3 indicated that the signature genes are also involved in apoptosis.

MSI-signature and prognosis

The prognostic value of the 64-gene MSI signature was tested on 263 mostly (80%) untreated stage II tumours: 132 samples from validation cohort B, plus an additional set of 131 stage II colon tumours with no available hospital-based MSI assessment (validation cohort C, Table 1). Patients with samples classified as MSI by the gene signature showed a significantly better distant metastasis-free survival (DMFS) compared to patients with MSS tumours, with a hazard ratio (HR) of 0.252 (95% CI, 0.076–0.83, p = 0.0145) (Figure 3A). After further subclassification into MSI, MSI-like and MSS, the MSI-like group also showed a significantly better survival compared to MSS samples. Interestingly, the MSI group with concordant MSI classification by signature and hospital method showed a 100% survival rate (Figure 3B). In contrast to stage II, investigation in stage III samples (n = 201) showed no prognostic value of MSI/MSS classification (p = 0.29) (data not shown).

It has been postulated that MSI patients might be resistant to 5-FU treatment and that this resistance is associated with thymidylate synthase (TYMS) activity. We therefore investigated the expression of *TYMS* in the tumours. Samples classified as MSI showed a significant higher expression of *TYMS* compared to samples classified as MSS (cohort A, p < 1e-18). Samples classified as MSI-like showed also a significantly higher expression of *TYMS* compared to MSS (p = 3.9e-13) (Figure 4).

Technical validation of the MSI gene signature

The reproducibility of the MSI signature was investigated by replicate hybridization and analysis of 53 samples. MSI gene signature results were highly reproducible, with an R^2 value of 0.992 (Figure 5A) and, importantly, all samples resulting in the same classification (100% concordance). Matching samples from the same patients (n = 60) that were either preserved as formalin-fixed and paraffin-embedded (FFPE) or preserved fresh in RNA-retain were analysed to address tumour heterogeneity and technical differences between FFPE and fresh preservation. The read-outs of MSI signature score from these two biopsies were highly correlated (R = 0.93) and the binary results (MSS versus MSI) were 98.4% concordant. In addition, a repeated assessment was performed for three samples over 20 consecutive days by five different technicians. Signature read-out was stable across the 20 consecutive days, with an average standard deviation of well below 5% of the total dynamic range (Figure 5B). Of the 60 measurements, only two read-outs resulted in a change in classification outcome (96.7% concordance). Finally, validation of the signature on the PETACC-3 study (Figure 1C) indicated that the gene signature, which has been developed and validated on fresh-frozen tissue samples, can be used for assessment of FFPE samples as well as fresh tissue.

Discussion

In this report we describe the development of a 64-gene expression signature that identifies patients

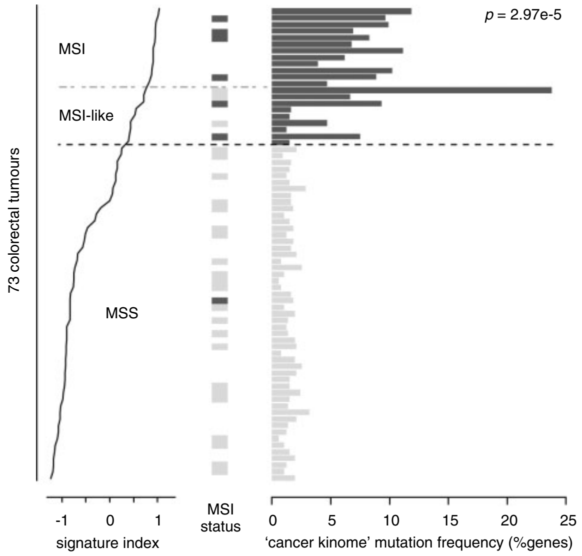


Figure 2. MSI and MSI-like samples classified by the 64-gene signature show an increased mutation frequency. Seventy-three colorectal tumour samples were sorted according to their MSI-signature index; the middle bar shows standard hospital-based MSI assessment when available (light grey, MSS; dark grey, MSI) and the right barplot show the mutation frequency (% of genes mutated) of each sample in the 'cancer kinome' (615 genes).

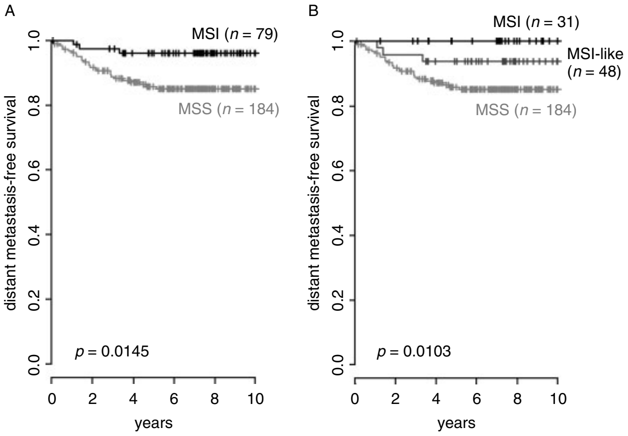


Figure 3. Prognostic value of the 64-gene MSI signature in 263 stage II colorectal cancer (cohorts B and C combined). (A) Kaplan–Meier (KM) survival curves for samples classified as MSI (MSI and MSI-like combined) and MSS by the gene signature; (B) KM curves for samples classified as MSI, MSI-like and MSS by the gene signature. *p* values are based on log-rank test.

with DNA mismatch repair deficiency resulting in a MSI phenotype. The signature was developed and independently validated in large sets of samples and showed high reproducibility in technical validation experiments. To our knowledge, this is the first report to describe a genomic MSI-signature directly linked to mutation frequency, which was translated into a robust diagnostic test.

The MSI-signature identifies patients with MSI status with high accuracy (85% and 91% accuracy in validation sets B and D, respectively) but also identifies a group of MSI-like patients who are not recognized by traditional methods as MSI but have features similar to MSI patients, eg high mutation frequency, frequent *BRAF* mutations, high *TYMS* expression and better prognosis. This observation is in good agreement with

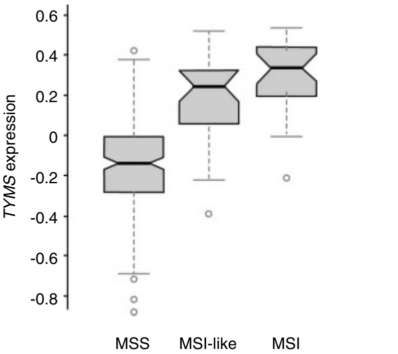


Figure 4. Relative gene expression levels ($\log_{10}$ scale) of thymidylate synthase (*TYMS*) in samples classified as MSS, MSI and MSI-like by the 64-gene signature. Samples classified as MSI showed a significant higher expression of *TYMS* compared to samples classified as MSS (p < 1e-18, Student's t-test) Samples classified as MSI-like also showed a significantly higher expression of *TYMS* compared to MSS (p = 3.9e-13, Student's t-test).

a recently published study from the Cancer Genome Atlas (TCGA) Network that also identified a group of patients with MSI-like features (high mutation frequency) who were classified as MSS by traditional methods [19]. This is clinically relevant because these patients might be better served without adjuvant chemotherapy if they are stage II. Additionally, these result indicate that that microsatellite instability is not necessarily a good surrogate for dMMR in all patients.

Interestingly, in our study, the sample with the highest mutation frequency (23.8%) is MSI-like by gene signature but was classified MSS by standard PCR assessment. This is again in good agreement with observations from TCGA that found that six patients with highest mutation rates were classified as MSS by standard methods. On the other hand, the single sample that was MSI by standard methods but with a strong MSS gene expression pattern in our set did not have an increased mutation rate, suggesting that this sample was incorrectly classified by standard MSI assessment (Figure 2).

The more comprehensive identification of MSI and MSI-like patients might be explained by the fact that the read-out of gene expression is a measurement of cellular consequences of DNA MMR deficiency in colorectal tumour, and is therefore independent of knowing the cause of the defect. At this moment, not all components of the MMR pathway in human cells are known, eg the human counterparts of *Escherichia coli MutH* and *UvrD* are not yet identified [31]. Although the epigenetic silencing of *MLH1* is often observed as the main factor, other factors are known to play a role. MMR defects can be caused by any genetic or epigenetic alteration of the genes in the DNA MMR pathway. Knock-out mouse models of *Msh2*, *Msh3*, *Msh6*, *Mlh1*, *Mlh3*, *Pms1*, *Pms2* and *Exo1* all confer a MSI phenotype [32,33]. It is therefore difficult to comprehensively measure all possible sources causing MMR deficiency. Moreover, although somatic mutations in known mismatch-repair genes might be detectable, the mutations do not always result in microsatellite instability, at least not in those microsatellites that are traditionally assessed [19]. However, it is possible to summarize the cellular consequence of DNA MMR deficiency with a dominant gene expression pattern, as with the 64-gene signature, that measures the downstream effect. The functional annotation of the 64 genes further supports the theory that the signature measures an activation that is caused by MMR deficiency, rather than the deficiency itself. Proteins with classical conserved zinc-finger domains, DNA binding domains and associated to the nucleic acid metabolic processes were enriched in the signature. The expression signature is indicative of active DNA damage signalling, which in turn leads to cell cycle arrest and apoptosis (see Supplementary material, Table S2).

The 64-gene signature summarizes the gene expression pattern displayed by colorectal tumours with DNA MMR deficiency, regardless of the diverse causes of this defect, and therefore might have advantages when compared to IHC or PCR methods [9]. Using a gene expression signature for MSI assessment might also have technical advantages: it does not require a comparison of DNA microsatellite regions from paired normal and tumour tissues; in addition, the nature of a molecular signature builds upon the read-out of a relatively large set of genes, resulting in robust and reproducible measurements; additionally, the MSI signature can be read out from the same tissue biopsy and in the same assay as other diagnostic signatures [20,21], minimizing sample requirements and systematic errors.

It has been well established that stage II MSI patients have better prognosis compared to patients with functional mismatch repair [34]. Consistent with this knowledge, we report here that tumours predicted by the 64-gene signature as MSI showed better distant metastasis-free survival. While the good prognosis of MSI tumours is well documented, the value of MSI to predict response to adjuvant chemotherapy is still under investigation. Cell line models support the idea that CRCs require a functional MMR system to induce apoptosis in response to 5-FU treatment [35]. In addition, meta-analysis of seven independent clinical studies indicated that MSI patients do not benefit from adjuvant chemotherapy with 5-FU [12]. The mechanism of action of 5-FU is through its metabolite dUMP, which competes for the binding site of thymidylate synthase (*TYMS*), an enzyme catalysing conversion of dUMP to dTMP during DNA synthesis. The non-responsiveness to 5-FU therapy in MSI patients might be related to higher expression of *TYMS* in these tumours [36]. In our dataset, we have confirmed this association, as MSI patients identified by the signature have high expression of *TYMS*. MSI-like patients might present an additional population of CRC patients that are unlikely to respond to treatment with 5-FU.

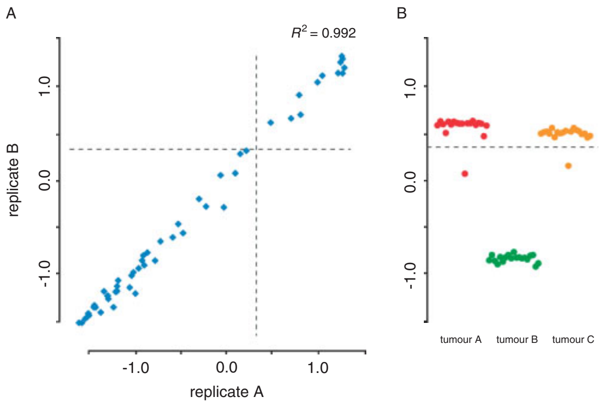


Figure 5. Reproducibility and precision of the 64-gene signature. (A) Replicate analysis of 53 tumour samples shows a very high correlation in signature index. (B) Stability of the MSI signature read-out for three representative diagnostic samples across a time period of 20 consecutive days. In both panels, the classification threshold (MSI vs MSS) is indicated by the dashed line.

To conclude, we have developed a 64-gene signature characterizing DNA MMR deficiency in colorectal tumours. This signature is technically robust and can be used as an alternative diagnostic tool to assess MSI status. It was implemented on a diagnostic array and validated in both fresh-frozen and FFPE tumour samples. The results from this test provide information on the prognosis of colorectal cancer patients and aid decision making for the selection of appropriate chemotherapeutic agents.

Acknowledgements

The authors thank Mireille Snel, Frank Nieboer, Jeroen Heijmans and Astrid Bosma for sample and microarray processing, Christa Dreezen for support with clinical annotation and Jeroen van den Akker and Annuska Glas for technical support and discussion. This work was funded in part by the European Commission 7th Framework research programme COLTHERES (Modelling and Predicting Sensitivity to Targeted Therapies in Colorectal Cancers; Grant No. 259015). ST is a senior clinical investigator of the Fund for Scientific Research Flanders and has received research grants from the Belgian Federation Against Cancer and Belgian National Cancer Plan.

Author contributions

All authors were involved in writing the manuscript and in reviewing the final draft; ST, PR, RB and IS conceived experiments and study design; ST, PR, VP, MM, IA and MD performed data analysis; RS, CS, RR, UN, WM, SB and SaT were involved in sample collection, updating patient information and/or generating MSI-data; and PR, VP, IA, RS, UN, SaT, MD, RB and IS were involved in data interpretation.

Abbreviations

5-FU, 5-fluorouracil; CRC, colorectal cancer; MMR, mismatch repair; MSI, microsatellite instability; MSS, microsatellite stability.

References

1. Warusavitarne J, Schnitzler M. The role of chemotherapy in microsatellite unstable (MSI-H) colorectal cancer. *Int J Colorect Dis* 2007; **22**(7): 739–748.
2. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell* 1996; **87**(2): 159–170.
3. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**(5): 759–767.
4. Ionov Y, Peinado MA, Malkhosyan S, *et al*. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* 1993; **363**(6429): 558–561.
5. Kane MF, Loda M, Gaida GM, *et al*. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Res* 1997; **57**(5): 808–811.
6. Tejpar S, Bertagnolli M, Bosman F, *et al*. Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist* 2010; **15**(4): 390–404.
7. Koopman M, Kortman GA, Mekenkamp L, *et al*. Deficient mismatch repair system in patients with sporadic advanced colorectal cancer. *Br J Cancer* 2009; **100**(2): 266–273.
8. Miquel C, Jacob S, Grandjouan S, *et al*. Frequent alteration of DNA damage signalling and repair pathways in human colorectal cancers with microsatellite instability. *Oncogene* 2007; **26**(40): 5919–5926.
9. Vilar E, Gruber SB. Microsatellite instability in colorectal cancer – the stable evidence. *Nat Rev Clin Oncol* 2010; **7**(3): 153–162.
10. Sargent D, Sobrero A, Grothey A, *et al*. Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20 898 patients on 18 randomized trials. *J Clin Oncol* 2009; **27**(6): 872–877.
11. Ribic CM, Sargent DJ, Moore MJ, *et al*. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based

adjuvant chemotherapy for colon cancer. *N Engl J Med* 2003; **349**(3): 247–257.

12. Des GG, Schischmanoff O, Nicolas P, *et al*. Does microsatellite instability predict the efficacy of adjuvant chemotherapy in colorectal cancer? A systematic review with meta-analysis. *Eur J Cancer* 2009; **45**(10): 1890–1896.
13. Sinicrope FA, Foster NR, Thibodeau SN, *et al*. DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. *J Natl Cancer Inst* 2011; **103**(11): 863–875.
14. Bertagnolli MM, Niedzwiecki D, Compton CC, *et al*. Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: Cancer and Leukemia Group B Protocol 89803. *J Clin Oncol* 2009; **27**(11): 1814–1821.
15. Vilar E, Mukherjee B, Kuick R, *et al*. Gene expression patterns in mismatch repair-deficient colorectal cancers highlight the potential therapeutic role of inhibitors of the phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin pathway. *Clin Cancer Res* 2009; **15**(8): 2829–2839.
16. Gonzalez-Garcia I, Moreno V, Navarro M, *et al*. Standardized approach for microsatellite instability detection in colorectal carcinomas. *J Natl Cancer Inst* 2000; **92**(7): 544–549.
17. Umar A, Boland CR, Terdiman JP, *et al*. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome): and microsatellite instability. *J Natl Cancer Inst* 2004; **96**(4): 261–268.
18. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology* 2010; **56**(2): 167–179.
19. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**(7407): 330–337.
20. Jorissen RN, Lipton L, Gibbs P, *et al*. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* 2008; **14**(24): 8061–8069.
21. Sun T, Simon I, Moreno V, *et al*. A combined oncogenic pathway signature of *BRAF*, *KRAS* and *PI3KCA* mutation improves colorectal cancer classification and Cetuximab treatment prediction. *Gut* 2012 Jul 14. [Epub ahead of print] PMID: 22798500
22. Salazar R, Roepman P, Capella G, *et al*. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**(1): 17–24.
23. Maak M, Simon I, Nitsche U, *et al*. Independent validation of a prognostic genomic signature (ColoPrint): for stage II colon cancer patients. *Ann Surg* 2012 (in press).
24. Van CE, Labianca R, Bodoky G, *et al*. Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J Clin Oncol* 2009; **27**(19): 3117–3125.
25. Popovici V, Budinska E, Tejpar S, *et al*. Identification of a poor-prognosis *BRAF*-mutant-like population of patients with colon cancer. *J Clin Oncol* 2012; **30**(12): 1288–1295.
26. Roth AD, Tejpar S, Delorenzi M, *et al*. Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60–00 trial. *J Clin Oncol* 2010; **28**(3): 466–474.
27. Nardon E, Glavac D, Benhattar J, *et al*. A multicenter study to validate the reproducibility of MSI testing with a panel of five quasimonomorphic mononucleotide repeats. *Diagn Mol Pathol* 2010; **19**(4): 236–242.
28. Roepman P, Jassem J, Smit EF, *et al*. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin Cancer Res* 2009; **15**(1): 284–290.
29. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**(1): 44–57.
30. Krishna SS, Majumdar I, Grishin NV. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res* 2003; **31**(2): 532–550.
31. Li GM. Mechanisms and functions of DNA mismatch repair. *Cell Res* 2008; **18**(1): 85–98.
32. Wei K, Kucherlapati R, Edelmann W. Mouse models for human DNA mismatch-repair gene defects. *Trends Mol Med* 2002; **8**(7): 346–353.
33. Prolla TA, Baker SM, Harris AC, *et al*. Tumour susceptibility and spontaneous mutation in mice deficient in *Mlh1*, *Pms1* and *Pms2* DNA mismatch repair. *Nat Genet* 1998; **18**(3): 276–279.
34. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol* 2005; **23**(3): 609–618.
35. Arnold CN, Goel A, Boland CR. Role of hMLH1 promoter hypermethylation in drug resistance to 5-fluorouracil in colorectal cancer cell lines. *Int J Cancer* 2003; **106**(1): 66–73.
36. Ricciardiello L, Ceccarelli C, Angiolini G, *et al*. High thymidylate synthase expression in colorectal cancer with microsatellite instability: implications for chemotherapeutic strategies. *Clin Cancer Res* 2005; **11**(11): 4234–4240.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

Supplementary methods

Table S1. Genes in the MSI-64 gene signature

Table S2. Functional annotation: functional category analysis of the 64 genes by DAVID software

14 Expression profiling with RNA from formalin-fixed, paraffin-embedded material

- BMC Medical Genomics, 1(9), 2008
- IF: 2.848
- number of citations: 36
- personal contribution (30%): design of genomic signatures and scores, statistical analyses, manuscript writing

BMC Medical Genomics

BioMed Central

Technical advance

Expression profiling with RNA from formalin-fixed, paraffin-embedded material

Andrea Oberli†1, Vlad Popovici†2, Mauro Delorenzi2,3, Anna Baltzer1, Janine Antonov1, Sybille Matthey1, Stefan Aebi1, Hans Jörg Altermatt4 and Rolf Jaggi*1

Address: 1Department of Clinical Research, University of Bern, Murtenstrasse 35 CH-3010 Bern, Switzerland, 2Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland, 3National Center of Competence in Research (NCCR) Molecular Oncology, Swiss Institute for Experimental Cancer Research (ISREC), Epalinges, Switzerland and 4Pathology Länggasse, Forstweg 56, CH-3012 Bern, Switzerland

Email: Andrea Oberli - andrea.oberli@dkf.unibe.ch; Vlad Popovici - vlad.popovici@isb-sib.ch;
Mauro Delorenzi - mauro.delorenzi@isrec.unil.ch; Anna Baltzer - anna.baltzer@dkf.unibe.ch; Janine Antonov - janine.antonov@dkf.unibe.ch;
Sybille Matthey - sybille.matthey@dkf.unibe.ch; Stefan Aebi - stefan.aebi@insel.ch; Hans Jörg Altermatt - altermatt@patholaenggasse.ch;
Rolf Jaggi* - rolf.jaggi@dkf.unibe.ch

* Corresponding author   †Equal contributors

Published: 19 April 2008

BMC Medical Genomics 2008, **1**:9   doi:10.1186/1755-8794-1-9

Received: 17 October 2007
Accepted: 19 April 2008

This article is available from: http://www.biomedcentral.com/1755-8794/1/9

Abstract

Background: Molecular characterization of breast and other cancers by gene expression profiling has corroborated existing classifications and revealed novel subtypes. Most profiling studies are based on fresh frozen (FF) tumor material which is available only for a limited number of samples while thousands of tumor samples exist as formalin-fixed, paraffin-embedded (FFPE) blocks. Unfortunately, RNA derived of FFPE material is fragmented and chemically modified impairing expression measurements by standard procedures. Robust protocols for isolation of RNA from FFPE material suitable for stable and reproducible measurement of gene expression (e.g. by quantitative reverse transcriptase PCR, QPCR) remain a major challenge.

Results: We present a simple procedure for RNA isolation from FFPE material of diagnostic samples. The RNA is suitable for expression measurement by QPCR when used in combination with an optimized cDNA synthesis protocol and TaqMan assays specific for short amplicons. The FFPE derived RNA was compared to intact RNA isolated from the same tumors. Preliminary scores were computed from genes related to the ER response, HER2 signaling and proliferation. Correlation coefficients between intact and partially fragmented RNA from FFPE material were 0.83 to 0.97.

Conclusion: We developed a simple and robust method for isolating RNA from FFPE material. The RNA can be used for gene expression profiling. Expression measurements from several genes can be combined to robust scores representing the hormonal or the proliferation status of the tumor.

Background

Breast cancer has been widely studied in the past and molecular characterization has increased the understanding of biological pathways that are altered during neoplastic transformation of cells [1-4]. However, the findings based on molecular profiling have not yet altered diagnosis, and decisions about treatment still rely mostly on histopathological and immunohistochemical techniques which are at best semi-quantitative [5,6]. Currently, many patients with primary, non-metastatic breast cancer with positive estrogen receptor (ER) status undergo several cycles of chemotherapy, although a substantial proportion of them does not benefit from it. Presently, no conventional parameters exist for many patients which allow to identify individuals who will benefit from chemotherapy. Personalized diagnosis on the basis of highly specific molecular analyses has the potential to improve the situation of many patients by optimizing medication, and at the same time, sparing others from unnecessary treatment regimens.

DNA chip studies are based on measuring gene expression for many genes in parallel [1,4,7,8]. Most protocols for gene expression analysis on the basis of DNA chips are sensitive to RNA degradation and RNA must be isolated from freshly prepared or FF tumor material. As a consequence, material is fairly limited and often originates from convenience samples of heterogeneous patients. Many of these studies including meta-analyses have revealed genes and biological functions of their products which are relevant for classification and prognosis [9,10]. However, many samples were derived from patients who did not participate in clinical studies and their treatment regimens were not standardized. Therefore, follow up data must still be interpreted with caution.

Obviously, procedures based on formalin-fixed, paraffin-embedded (FFPE) material would greatly facilitate and speed up research in this area as large amounts of highly valuable material and clinical data have already been collected. In many cases, FFPE blocks are still available and they could be used for a molecular analysis. Especially material from clinical trials would allow investigating distinct clinical questions with existing material rather than material from newly designed studies.

Many efforts are currently made to individualize diagnosis of breast cancer by including molecular parameters into diagnosis. Fresh frozen material would obviously be ideal for a molecular analysis by gene expression measurements but it may be difficult to implement novel procedures which complicate current workflows of daily routine. Procedures based on FFPE material would be more feasible as they do not interfere with current protocols and they do not affect routine diagnosis as material for molecular analysis could be collected after standard diagnosis has been terminated. Only relatively few molecular approaches have been described which are based on FFPE material. For example, Paik and co-workers have established a recurrence score (RS, Oncotype DX), it allows to quantify the likelihood of distant recurrence and to predict the magnitude of chemotherapy benefit [11,12].

It is generally accepted that molecular profiles which reflect primarily biological characteristics of tumor cells, may complement clinical and histopathological diagnosis, resulting in a more detailed characterization of individual tumors, a pre-requisite for better treatment decisions. In this study we present the development of a novel procedure for RNA isolation from FFPE material and an optimized workflow for expression measurements by QPCR.

Methods

Human breast cancer samples

Human breast cancer specimens were divided into two aliquots, one of which was processed for histological diagnosis by fixation with formalin and embedding in paraffin. FFPE material was obtained from the Institute of Pathology (University of Bern) and the Pathology Länggasse, Bern. Tissue (3–5 mm thick slices of tumor) was fixed over night in buffered formalin and processed for paraffin embedding in a Tissue Processing Center TPC 15 (Medite Medizintechnik, Germany). The second aliquot was frozen on dry ice and stored at -80°C. Fresh frozen material was obtained from the Tumorbank Bern. Both, FF and FFPE samples were checked by hematoxylin and eosin staining and only samples with more than 50% tumor cells were used for this study. An informed consent to use the material for research was obtained from all the patients.

RNA Extraction

Intact RNA was isolated from four 25 μm thick kryo-sections of approximately 0.5 cm^2. The tissue was homogenized in 420 μl lysis buffer (4 M guanidinium thiocyanate, 30 mM Tris pH 8.0, 1% Triton-X-100), 8.0,1 using a TissueLyser (Mixer Mill, Retsch GmbH, Haan, Germany) at 15 Hz for 3 min. Total RNA was bound to silica-based columns (Epoch Biolabs, Huston Texas), treated with DNase I (30 Kunitz units for 20 min. at room temperature; Qiagen, Hilden, Germany), washed once with lysis buffer (containing 30% ethanol) and once with 20 mM NaCl (containing 20% ethanol) and eluted in 50 μl 10 mM Tris pH 7.4, 0.1 mM EDTA and stored at -20°C. RNA quantity was measured on an ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and quality assessed by capillary electrophoresis with an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA) using Agilent RNA 6000 Series Nano kits.

RNA was isolated from ten 10 µm thick FFPE sections according to the RNeasy FFPE protocol of Qiagen (Fig. 1, lanes B), the ncLysis protocol of Applied Biosystems (lanes C) and the protocol developed in our laboratory (lanes D). Paraffin sections were de-paraffinized with xylene, washed with ethanol and dried in a speed vac. For our own protocol, 200 µl lysis buffer (4 M guanidinium thiocyanate, 30 mM Tris, pH 8.0, 1% Triton-X-100) was added to the dried sections and immediately homogenized in a Mixer Mill at 20 Hz for 4 min. Proteinase K (Roche Diagnostics, Mannheim, Germany) was added (1 mg/ml final concentration) and tissue was digested for 1 hour at 55°C. One milliliter dilution buffer (30 mM Tris, pH 8.0, 1% Triton-X-100) was added to each lysate and digestion continued for 1 hr after adding fresh proteinase K (final concentration 1 mg/ml). RNA was de-modified by adding 318 µl of de-modification solution (5 M NH_4Cl) and incubating at 94°C for 20 min or as described in the text. RNA was bound to silica-based columns and digested with DNase I as described for fresh-frozen tissue samples. The reproducibility of our own procedure was tested by isolating several independent RNAs from consecutive sections of the same tissue block. About 10 µg of total RNA could be isolated from 5 to 10 FFPE sections (0.5–1 cm^2/section). RNA was isolated from closely matched sections using the RNeasy FFPE kit (Qiagen) or the ncLysis system (Applied Biosystems) according to the protocols included with the kits. In both cases, the RNA was purified on silica-based columns. 22 samples were available. In 14 cases sufficient RNA was obtained from all 4 parallel isolations. In 2 cases of FF material (samples 4 and 11) and in 6 cases of FFPE material (samples 1, 5, 7, 9, 12 and 21) less than 1.5 µg RNA could be isolated with the ncLysis protocol. These samples were excluded from further analysis.

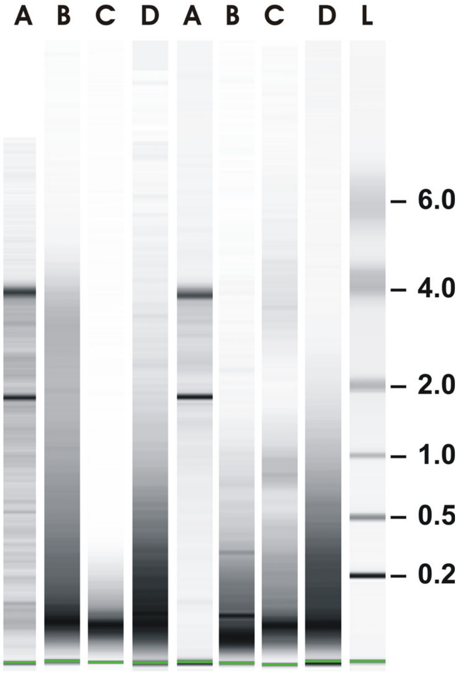


Figure 1
RNA isolation and characterization. Total RNA was isolated from kryo-sections (lanes A) and from paraffin sections according to the RNeasy FFPE protocol of Qiagen (lanes B), the ncLysis protocol of Applied Biosystems (lanes C) or according to our own protocol (lanes D). Aliquots of each RNA were separated by capillary electrophoresis (Agilent Bioanalyzer) on Nano chips along with RNA ladder (L; Ambion). Shown are RNAs from two representative tumors (Tu#10 and #18).

cDNA synthesis and QPCR

Aliquots of 100 to 500 ng of total RNA were reverse transcribed using MultiScribe™ MuLV reverse transcriptase (High-Capacity cDNA Archive Kit; Applied Biosystems, Foster City, CA, USA) and random or gene-specific primers. Reverse primers were kindly provided by Applied Biosystems, they were used at 1 µM each, cDNAs were made in the presence of 3, 10 or 22 reverse primers as 3-plex, 10-plex or 22-plex, respectively. Regular Assays on Demand (Applied Biosystems) were used for QPCR (Table 1). Manually designed assays coding for short, medium-size and long amplicons of the insulin growth factor-binding protein 5 (IGBP5) were selected with Primer Express (Version 3, Applied Biosystems). Forward primer and probe were kept constant for all assays while reverse primers were selected such that amplicons of different sizes were generated [13]. QPCR reactions were carried out in triplicates in a final volume of 10 µl in 1× FAST Master mix (Applied Biosystems) and cDNA corresponding to 4 ng total RNA. QPCR was performed on an ABI 7500 FAST instrument (2 min at 95°C, followed by 45 cycles of 95°C for 3 sec and 60°C for 30 sec). The quality of the assays and the absence of contaminating DNA were assessed with water and RNA instead of cDNA, respectively (data not shown). Three positive controls containing cDNA derived of ZR-7-51 cells were included on each 96-well plate. Cycle threshold values (Ct) were determined using the SDS software of the 7500 FAST System (Version 1.3.1). Constant threshold values were set for each gene throughout the study.

Table 1: QPCR assays. QPCR assays (Assays on Demand) were from Applied Biosystems (Palo Alto, CA). Reverse primers from each assay were used for the synthesis of gene-specific cDNAs. They were provided separately by Applied Biosystems. Three assays (IGBP5_short, IGBP5_medium, IGBP5_long) were designed manually.

AoD	Assay	Acc_Nr	AmpliconSize	Module
Hs00608023_m1	BCL2	NM_000633	81	Estrogen
Hs00221277_m1	CEGP1	NM_020974	64	Estrogen
Hs00174860_m1	ESR1	NM_000125	62	Estrogen
Hs00172183_m1	PGR	NM_000926	118	Estrogen
Hs00180450_m1	GRB7	NM_005310	70	Her2
Hs01001598_g1	HER2	NM_004448	55	Her2
Hs00952036_m1	CTSL2	NM_001333	72	Invasion
Hs00171829_m1	STMY3	NM_005940	66	Invasion
Hs01030097_m1	CCNB1	NM_031966	66	Proliferation
Hs01032443_m1	MKI67	NM_002417	66	Proliferation
Hs00231158_m1	MYBL2	NM_002466	81	Proliferation
Hs00269212_m1	STK15	NM_003600	85	Proliferation
Hs00153353_m1	SURV	NM_001168	93	Proliferation
Hs99999903_m1	ACTB	NM_001101	171	Reference
Hs00266705_g1	GAPDH	NM_002046	74	Reference
Hs99999908_m1	GUSB	NM_000181	81	Reference
Hs99999902_m1	RPLP0	NM_001002	105	Reference
Hs00174609_m1	TFRC	NM_003234	79	Reference
Hs00430290_m1	UBB	NM_018955	120	Reference
Hs01630490_s1	RPL7A	BX641050	84	Reference
Hs00817975_g1	RPS11	NM_001015	168	Reference
Hs01922548_s1	RPS23	NM_001025	90	Reference
Hs00185390_m1	BAG1	NM_004323	58	
Hs00154355_m1	CD68	NM_001251	68	
Hs01383449_s1	GSTM1	AY532925	65	
(own design)	IGBP5_short	NM_000599	60	Test
(own design)	IGBP5_medium	NM_000599	109	Test
(own design)	IGBP5_long	NM_000599	147	Test

Data processing and determination of breast cancer classification scores

All the measured cycle threshold (Ct) values represent $\log_2$ expression levels. These values need to be normalized such that they are comparable across samples and suitable for generating scores. For a gene, a large Ct value corresponds to a low expression level, so the first processing step needed was to reverse the sense of this relationship by letting

$$Ct' = \max(cut\_off - Ct, 0)$$

be the new value for each measured gene. The cut off value was set empirically to 35.0 as any higher raw Ct value was deemed unreliable. This cut off was fixed a priori and kept constant throughout all the experiments reported here. Then, the final value of each target gene was taken to be

$$\Delta Ct = max\_val*(Ct' - R + cut\_off)/(2*cut\_off),$$

where R represents the reference value and was taken as the mean of Ct' values of 5 selected reference genes (GAPDH, GUSB, RPLP0, TFRC, UBB, see Results section for details). The approach guarantees that all ΔCt values are positive and upper bounded by max_val (set to 33 for all the results reported here).

We used the scores associated with three of the gene groups listed in Table 1: the ER, HER2 and Proliferation group. While for the HER2 and Proliferation groups the scores were taken as the average ΔCt value of the genes in the group, for the ER group more weight was given to the ESR1 gene:

$$ER\_score = 0.55*ESR1 + 0.15*(BCL2 + CEGP1 + PGR)$$

where the gene symbols stand for the corresponding ΔCt values.

Finally, for each tumor a Total score was computed as

$$Total\_score = (Proliferation\_score + HER2\_score - ER\_score + max\_val)/3$$

The Total score, together with the group scores as computed above, are used in all subsequent discussions.

Results

Isolation of RNA from FFPE material

Total RNA was isolated from FF human breast cancer specimen which resulted in intact RNA in all samples (Fig. 1, lanes A, shown are RNAs from two representative tumors from a series of 14 tumors). RNA from FF tissue was used as reference for partially fragmented RNA isolated from FFPE material of the same tumors. RNA was assessed by capillary electrophoresis. The size distribution of RNA isolated according to our own protocol was in the range of 200 to 1000 nucleotides (Fig. 1, panel D) while the majority of RNA fragments was in the range of 100 nucleotides when RNA was isolated according to RNeasy FFPE (panel B) or the ncLysis system (panel C). Gene expression was measured by QPCR using 25 commercially available and three own TaqMan assays [13] (Tab. 1). The cycle threshold values (Ct values) were determined from RNAs isolated according to one of the three protocols for FFPE material and compared to Cts obtained with intact RNA of the same tumors. Fig. 2 shows correlation coefficients between intact RNA (A) and FFPE-derived RNAs isolated according to the RNeasy FFPE protocol, (A vs B); the ncLysis system (A vs C); or our own protocol (A vs D) for all 14 tumors using the expression levels of 5 genes (GAPDH, GUSB, RPLP0, TFRC, UBB; see below). The cDNAs were made in the presence of random (white boxes) or gene-specific primers (gray boxes). Clearly, correlation coefficients between intact and partially fragmented RNA were higher with gene-specific primers than random primers and RNA isolated according to our own protocol resulted in cDNA which performed better in QPCR than cDNA made from RNA isolated according to RNeasy FFPE and ncLysis protocols.

Parameters affecting the RNA quality and QPCR

Several parameters were systematically optimized to improve the protocol for RNA isolation from FFPE-derived sections. For example, QPCR made in the presence of primers specific for large amplicons (Fig. 3, dashed line) is very sensitive to RNA fragmentation and modification resulting in higher Ct values than primers specific for medium-size amplicons (dotted line) or short amplicons (non-interrupted line). In addition, the effect of de-modification of FFPE-derived RNA is apparent: the Ct determined from de-modified RNA is 3 or more units lower than the Ct measured from the same RNA but without de-modification. The effect was consistently observed with several tumors and also when expression was measured

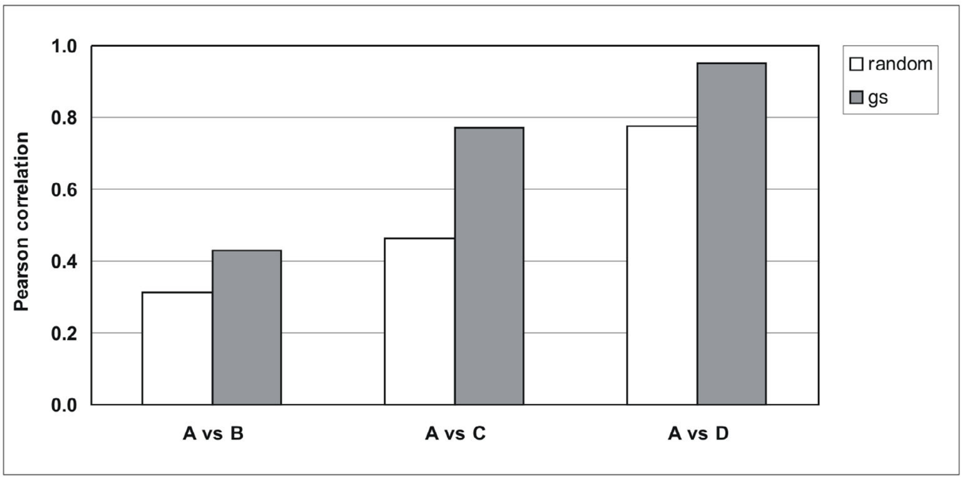


Figure 2

Comparison of RNAs isolated according to different protocols. RNA was reverse transcribed in the presence of random primers (white boxes) or gene-specific primers (hatched boxes). Gene expression was measured from an equivalent of 4 ng of RNA by QPCR for five reference genes (GAPDH, GUSB, RPLP0, TFRC and UBB). Pearson correlations were computed between matched Cts for the five reference genes and each tumor RNA isolated from FF (A) and FFPE material. Shown are correlations between intact RNA and RNA isolated from FFPE material according to the RNeasy FFPE protocol (A versus B), intact RNA and RNA isolated from FFPE material according to the ncLysis system (A versus C) and intact RNA and RNA isolated according to our own protocol (A versus D).

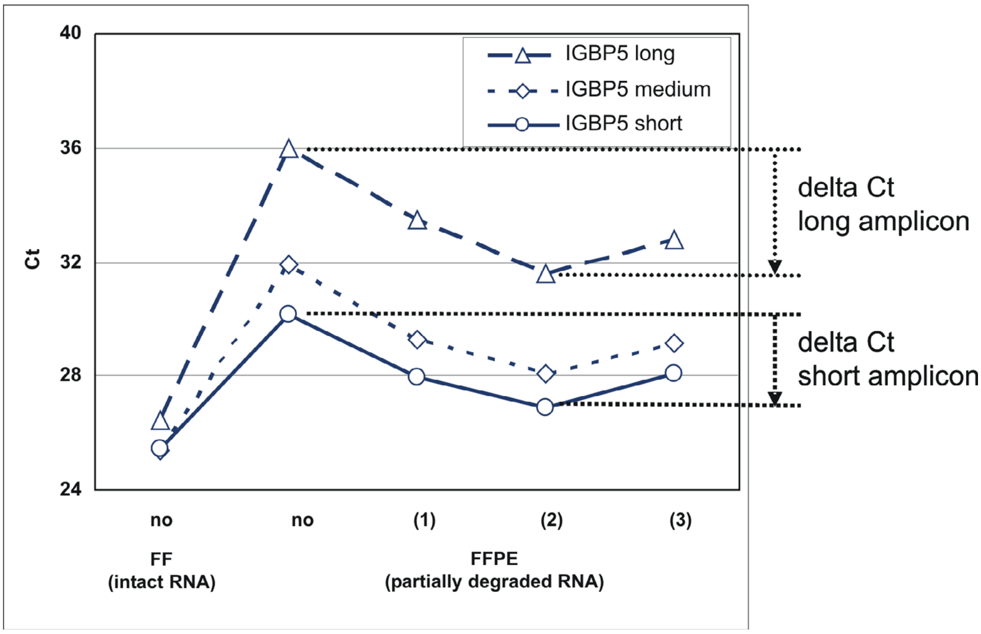


Figure 3

De-modification of RNA results in higher efficiency during subsequent QPCR. RNA was isolated from FFPE material according to our own protocol and compared to intact RNA derived of FF tissue. RNA samples were reverse transcribed without previous de-modification (labeled "no") or after de-modification at room temperature (1), 94°C and pH 8.0 (2) or 94°C and pH 5.0 (3). Each RNA was tested by QPCR using three amplicons for IGBP5. Primers used code for short (60 bp, □), medium-size (109 bp, ◈) or long amplicons (147 bp, △). Shown are raw Ct values from intact RNAs from FF material and from RNAs derived of FFPE material of the same tumors. The benefit of de-modification is visualized as delta Ct values. They are indicated for short and long amplicons (dotted lines).

with TaqMan assays from Applied Biosystems (data not shown). The optimum time of demodification was 20 min, longer times led to higher Ct values (not shown).

The different protocols of RNA isolation from FFPE material were further compared by measuring expression levels of reference genes in the 14 tumors and by comparing the results to Cts generated from corresponding intact RNAs (Fig. 4). Experimental variation was reduced by comparing mean Ct values from 5 reference genes (GAPDH, GUSB, RPLP0, TFRC, UBB) instead of their single values. Mean Cts of the five reference genes were plotted for each tumor and each protocol (panel A) and their distribution summarized (panel B). As expected, the Ct values generated with intact RNA resulted in the lowest and most stable Cts (diamonds). RNA prepared from FFPE tissue according to our own protocol (circles) resulted in higher but fairly constant Ct values (compare diamonds and circles). RNA isolated according to the RNeasy FFPE protocol (squares) and the ncLysis protocol (triangles) resulted in Ct values that were not only much higher than with intact RNA, they also exhibited large variations among different isolates when compared to corresponding Cts based on intact RNA. This result suggests a generally poorer and more variable quality of RNA isolated according to the two commercial protocols than our own protocol, leading to relatively large variations of Cts for the 5 reference genes among the different tumors. The Ct values generated from RNA isolated according to our own protocol were on average 2.9 units higher than Cts from intact RNA. RNA isolated according to RNeasy FFPE and ncLysis were 7.6 and 5.8 units higher than Cts from intact RNA of the same tumors, respectively (Fig. 4B). Standard deviations of Cts for the 14 tumors were 0.45 for intact RNA,

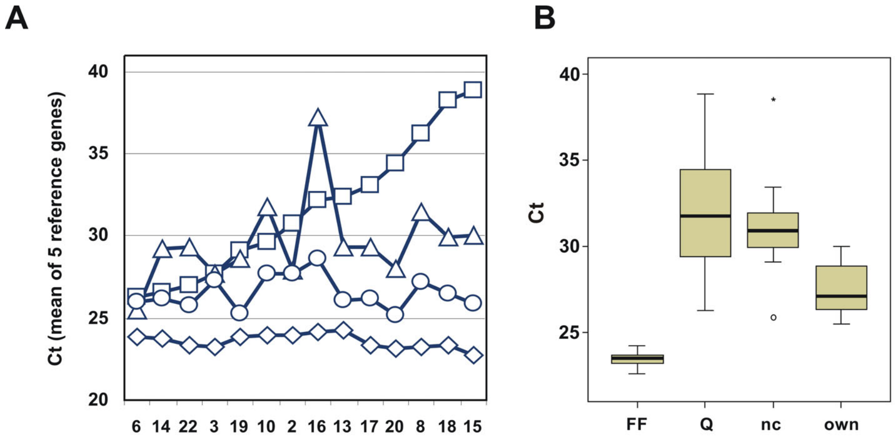


Figure 4

Comparison of RNA isolation methods. Shown are the means of raw Cts of five reference genes (GAPDH, GUSB, RPLP0, TFRC, UBB) for intact RNA (◈, FF) and for RNA isolated from matched FFPE material according to the protocols of Qiagen (□, Q), Applied Biosystems (△, AB) and our own (, own). Individual mean Cts of the 14 tumors and summarized box plots of Cts are shown in panel A and panel B, respectively. Tumors are aligned according to increasing Ct in FFPE-derived RNA (Qiagen protocol).

and 4.21, 2.69 and 1.01 for FFPE-derived RNA isolated according to the RNeasy FFPE, ncLysis and our own protocol, respectively.

An important aspect when working with RNA from FFPE material relates to the reproducibility of the RNA isolation procedure. This was directly tested for our own protocol by isolating independent samples of RNA from closely matched FFPE sections of the same tissue block and measuring gene expression by QPCR from both RNAs (Fig. 5A and 5B showing two representative examples). RNAs were also isolated from two independent tumors from the same patient, resulting in a third panel of data sets (C). Data points are shown as polygonal diagrams of raw Cts for each gene measured. Horizontal, parallel lines indicate closely similar expression, crossing lines indicate discrepancies between two measurements in matched samples. The Pearson correlation of raw Cts between matched samples was 0.99 for replicates shown in panels A and B and 0.74 for results shown in panel C.

Normalization

Results generated in the presence of partially fragmented RNA cannot be directly aligned with results produced from intact RNA and a suitable normalization is required to eliminate or reduce the effects of fragmentation and residual modification in RNA from FFPE material. Nine putative reference genes were selected from the literature [14] and from microarray results [15]. Expression was measured from intact and FFPE-derived RNA and raw Cts from all the 14 tumors are plotted for each putative reference gene (Fig. 6). Analyses based on intact RNAs revealed that 8 of the 9 tested genes performed similarly well (panel A). RPS23 which was hardly measurable (mean Ct in intact RNA > 37) was characterized by a large variation between the different tumors. A slightly higher variation was observed when expression levels were compared for FFPE-derived RNAs (panel B): GAPDH, GUSB, RPLP0, TFRC, RPL7A and UBB showed a similar performance and small variations between the 14 tumors as was seen with intact RNA. In contrast, the Ct values with RNA from FFPE material revealed larger variations for ACTB and RPS11 and therefore, the two genes were excluded as reference genes. The ACTB and RPS11 amplicons are larger than amplicons for the other reference genes and also for the test genes (Tab. 1, see also Fig. 3). Five genes were used as reference genes: GAPDH, GUSB, RPLP0, TFRC and UBB. For comparison, raw Ct values are shown for 4 genes related to the ER response (BCL2, CEPG1, ESR1, PGR) (Fig. 6, left). As expected, a high variation was observed for these genes between the 14 tumors. Protocols B and C did not yield enough usable data, precluding the data from further analysis. For example, protocol B did not have data for all the reference genes and for protocol C

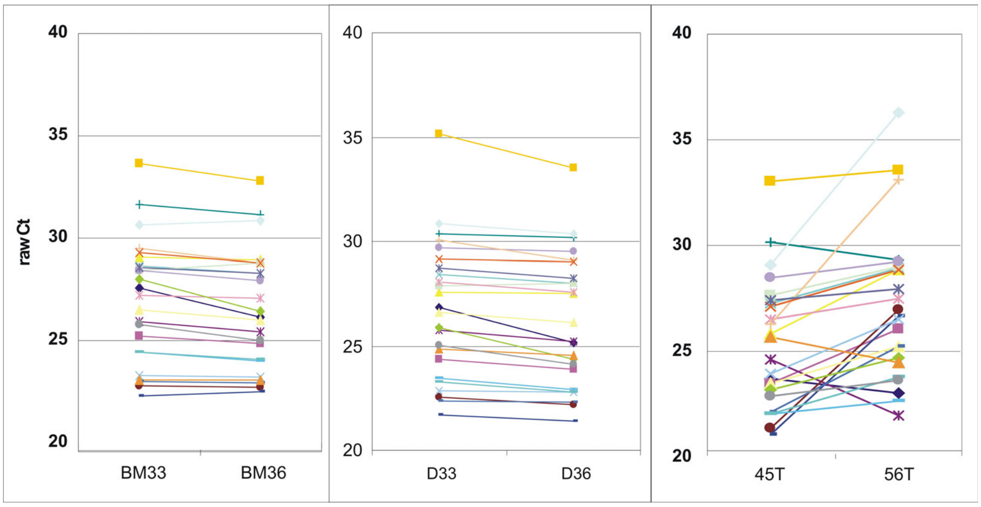


Figure 5
Reproducibility of RNA isolation from FFPE material. The RNAs were isolated from paraffin blocks according to our own protocol. BM33 and BM36 (panel A) are two separate RNAs isolated from tissue block "BM", D33 and D36 are RNAs isolated from block "D" (panel B). For comparison, 45T and 56T originate from two distinct tumors isolated from one patient (panel C). Gene expression was measured by QPCR for 24 genes and raw Ct values are shown for each gene measured from the two matching RNAs.

several test genes could not be measured reliably (e.g. BCL2, PGR of the ER group).

RNAs isolated from FFPE material according to our own protocol were also compared to RNA derived of kryo-preserved material of the same tumors in a different way. The arithmetic mean of the five reference genes (GAPDH, GUSB, RPLP0, TFRC and UBB) was used for normalizing expression values of all the genes in each RNA. Normalized expression values were compared between intact and FFPE-derived RNA for each gene and each tumor [see Additional File 1]. Good conservation of inter-tumor differences were observed between kryo-preserved and FFPE samples for most genes.

Module scores

Normalized expression values were also used to compute scores representing ER-related genes (ESR1, PGR, BCL2, CEPG1), HER2-related genes (HER2 and GRB7), genes related to proliferation (STK15/AURKA, CCNB1, MYBL2, MKI67, BIRC5/SURV) and a Total score representing all the genes of the three scores (for details see Methods). The computation of biologically meaningful scores with multiple genes instead of relying on just one has the scope to reduce noise variation. Module scores and Total scores were computed separately from normalized expression values of intact RNAs (circles) and of RNAs isolated according to our own protocol (triangles) and Total scores are depicted separately for each tumor (Fig. 7). The figure demonstrates that similar values are obtained for each tumor irrespective of whether they are computed from intact RNA or from RNA derived of FFPE material. This suggests that scores can be computed with RNA from FF samples as well as with RNA from FFPE samples. ER and HER2 scores were visualized in scatter plots, where the ER and HER2 scores were represented on the x- and y-axis, respectively (Fig. 8A and 8B). It was apparent that the three immunohistochemically ER-negative tumors have low ER scores (#15, #18, #20) and the only immunohistochemically HER2 positive tumor (#6) among the 14 tested tumors had a high HER2 score and an intermediate ER score (see also Table 2). The remaining tumors were all ER positive as assessed by immunohistochemistry (IHC) and they had relatively high ER scores. ER-negative and HER2-positive tumors all had high Proliferation scores (visualized by the red color of the dots). A larger spectrum of Proliferation scores (from blue to red) was found for ER positive tumors. Similar distributions were found when

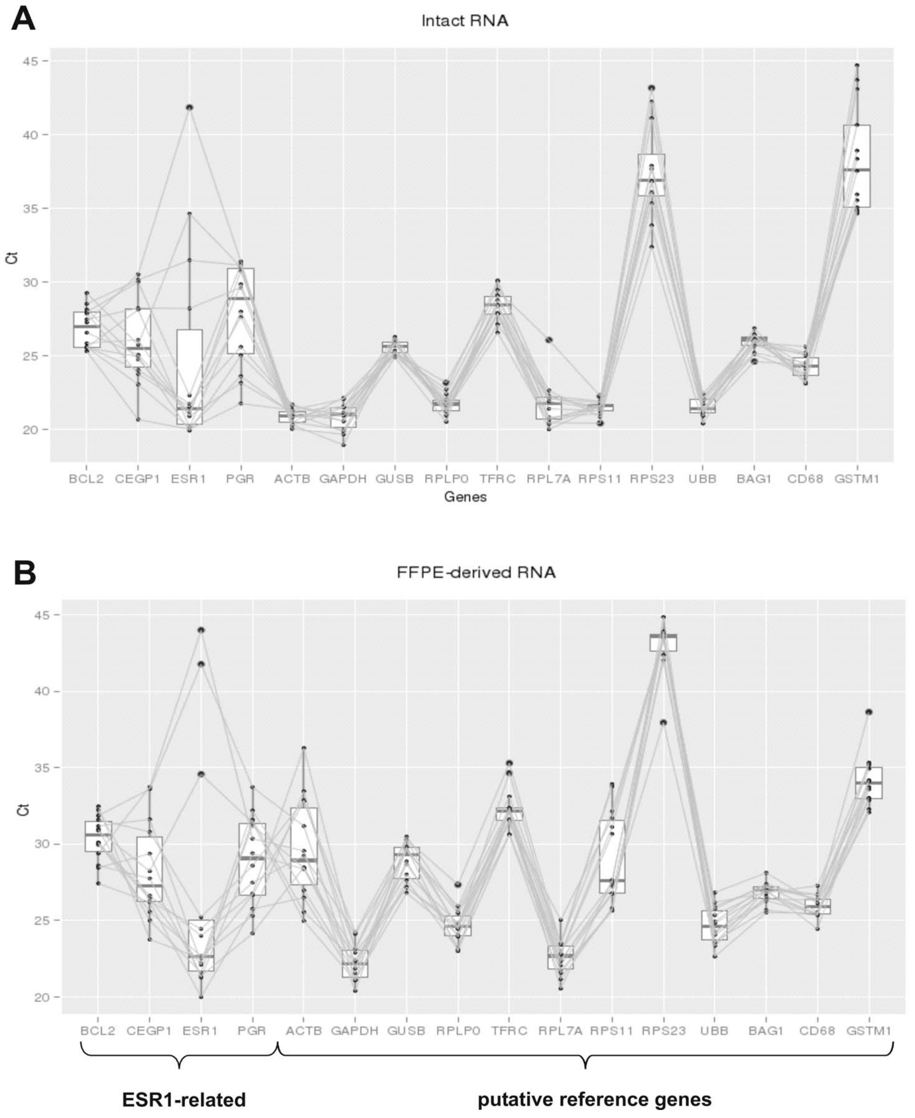


Figure 6

Stability of reference gene expression in RNA isolated from FF and FFPE material. Raw Cts are shown for 9 putative reference genes (ACTB, GAPDH, GUSB, RPLP0, TFRC, RPL7A, RPS11, RPS23 and UBB). Results based on intact RNA derived of FF material (A) and based on RNA isolated according to our own protocol from FFPE material (B) are depicted for all the 14 tumors. The Ct values for 4 ER-related genes (BCL2, CEPG1, ESR1 and PGR) are shown for comparison (left).

Table 2: Clinical and molecular parameters of breast cancers. Clinical and molecular parameters are given for each breast cancer used in this study. Module scores for each tumor were calculated from the results based on intact RNA (FF material) and based on RNA isolated from FFPE material according to our own method. N.A., data not available.

	Clinical classification			Immunohistochemistry			Module Score (FF/FFPE)			
Tu#	**T**	**N**	**Grade**	**ER**	**PR**	**ErbB2**	**ER**	**HER2**	**Prolif.**	**Histological type**
2	2	0	3	70% pos.	neg.	1+	16.6/17.1	15.8/16.3	14.2/14.5	invasive ductal
3	2	1a	2	70% pos.	pos.	1+	17.2/17.7	16.4/16.6	14.3/14.0	mixed (duct./lob)
6	1c	3a	3	>90% pos.	pos.	3+	15.7/16.2	17.2/18.1	14.5/15.4	invasive ductal
8	2	2a	3	>90% pos.	pos.	2+	16.5/17.2	15.8/16.3	14.7/14.7	invasive ductal
10	1c	N.A.	2	>90% pos.	pos.	2+	14.5/16.6	15.2/16.2	13.5/14.4	invasive ductal
13	2	3a	3	>90% pos.	neg.	1+	16.6/17.0	15.7/15.6	14.6/14.5	invasive ductal
14	1c	N.A.	2	>90% pos.	neg.	1+	16.5/16.9	16.2/16.4	13.8/13.5	invasive ductal
15		N.A.	3	neg.	neg.	0	11.8/13.0	14.9/15.5	14.9/15.5	invasive ductal
16	2	N.A.	1	65% pos.	pos.	0	17.9/18.3	16.4/16.6	13.8/14.1	invasive ductal/cribriform
17	2	0	3	>90% pos.	pos.	2+	16.5/17.2	16.3/16.9	14.9/15.6	invasive ductal
18	2	N.A.	3	neg.	neg.	0	13.0/12.9	15.1/15.9	15.2/16.0	invasive ductal
19	2	N.A.	2	>90% pos.	pos.	0	17.2/17.6	15.9/16.0	13.8/14.2	invasive ductal
20	1c	N.A.	2	neg.	neg.	0	12.4/13.0	15.7/15.9	14.6/15.2	invasive ductal
22	2	0	2	N.A.	N.A.	N.A.	16.8/17.4	15.6/16.1	13.3/13.9	invasive ductal

scores were computed from intact RNA (Fig. 8A) and FFPE-derived RNA that was isolated according to our own protocol (B). A different presentation of scores is shown where ER, HER2, Proliferation and Total Scores are plotted separately for each tumor [see Additional file 2]. The scores determined from the 14 FF and FFPE-derived samples are in the same range and only few tumors were classified in a different order between intact and FFPE-derived RNAs (leading to crossing lines).

The similarity between the results generated from intact and partially fragmented RNA was also assessed by calculating Pearson correlation coefficients between the scores of both RNAs. Correlation coefficients (and corresponding p-values and 95% confidence intervals) were 0.966 (p = 2.071*10-8, CI = 0.893; 0.989), 0.856 (p = 9.32*10-5, CI = 0.597; 0.954) and 0.833 (p = 2.177*10-4, CI = 0.541; 0.946) for ER, HER2 and Proliferation scores, respectively. The corresponding Spearman correlations were 0.938 (p < 2.2*10-16), 0.851 (p = 1.167*10-4) and 0.867 (p = 2.048*10-5), respectively.

Discussion

Methods and protocols for RNA isolation from formalin-fixed tissues have been published since almost 20 years [16-32].

RNA was quantified by dot blot hybridization [23], semi-quantitative PCR [19] and more recently, by QPCR [24,18,26,13,17,33,32] and other methods [28-30]. RNA derived of FFPE material is not only partially hydrolyzed but also chemically modified: formalin reacts with nucleotides leading to the formation of methylol groups in nucleobases. These groups tend to further react and form intra- and inter-molecular methylene bridges in RNA, DNA [34,35,31] and protein [36]. As a result, reverse transcription is impaired and threshold cycle values (Ct values) increase during subsequent QPCR.

The protocol for RNA isolation described here was complemented by adding a separate demodification step which involves incubation at elevated temperature in a buffer containing ammonium chloride which favors the reversion of methylol groups to amino groups in nucleobases. It does not only improve the efficiency of downstream applications (mainly reverse transcription), it also improves the recovery of RNA from FFPE sections. RNA yield and quality can be further improved by extensive digestion of FFPE material with protease in a buffer containing guanidinium thiocyanate. Reverse transcription in the presence of gene-specific primers prevents the initiation of cDNA synthesis inside amplicons and therefore, cDNA made in the presence of gene-specific primers is a better template for QPCR than cDNA made from random primers (Fig. 2). Several papers have demonstrated that QPCR with primers coding for short amplicons are more efficient than primers coding for long amplicons [17,20,24,13,32].

Finally, normalization of raw data is used to eliminate or at least reduce the effect of poorer quality of starting RNA. Various approaches of normalization were proposed in the literature [37,14,38,32]. They are based on calculating relative expression values: expression levels of genes of interest are expressed relative to the expression of one or a panel of several suitable reference genes. An ideal refer-

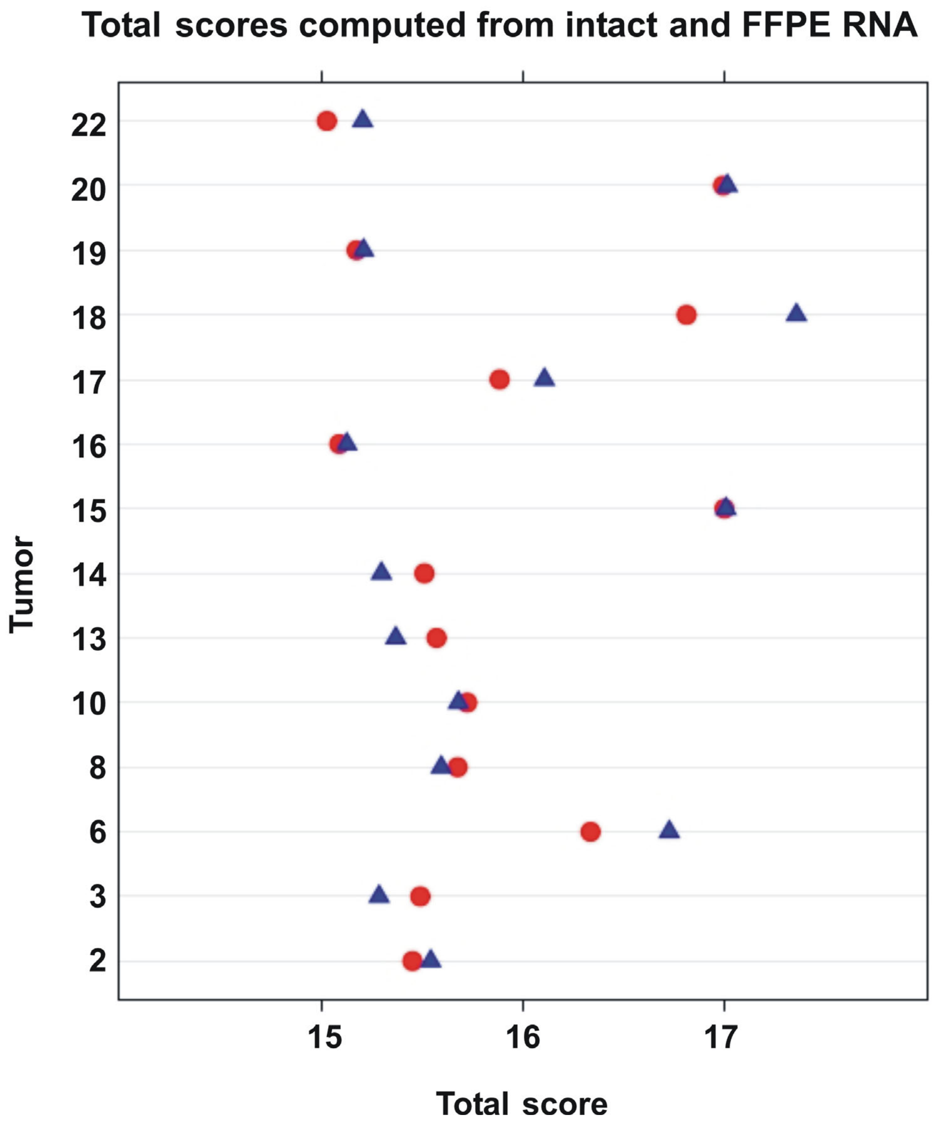


Figure 7

Comparison of Total scores computed from intact and FFPE-derived RNA. Total scores were computed from normalized expression values based on the results of intact RNA ( ) and FFPE-derived RNA (△, own protocol) as described in the Methods section. They are shown separately for each of the 14 tumors.

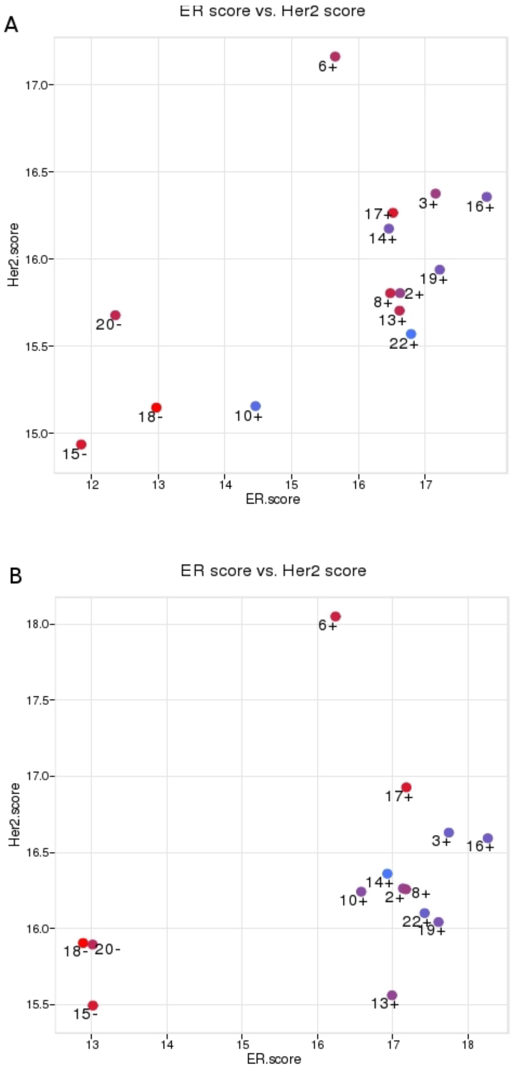


Figure 8
Module scores. ER, HER2 and proliferation scores were computed from expression values of 14 breast cancers and visualized in a scatter plot. The ER score was determined from four genes, the HER2 score from 2 and the proliferation score from 5 genes (see Methods). Tumors are positioned according to their ER score (x-axis) and HER2 score (y-axis). Proliferation scores are color coded. The histological ER status is indicated by a "-" or "+" sign next to the tumor numbers in the plot. The results were computed from intact RNA derived of FF material (A) and RNA isolated from FFPE material according to our own protocol (B). Individual scores for each tumor are given in Table 2.

ence gene has a stable expression level in all the samples under investigation. As such "ideal" reference gene normally does not exist, the mean or median expression level of several suitably chosen reference genes is used as a relatively stable reference Ct value. We used a formalized approach to characterize all candidate reference genes. Candidate reference genes were ranked according to their standard deviations of raw Ct values in RNA from FF and FFPE material. The final rank of each candidate reference gene was taken as the mean of the two ranks obtained with RNA from intact and FFPE material. Genes with higher ranks were excluded as reference genes.

We also applied GeNorm [14] to characterize candidate reference genes: ACTB and RPS11 had poorest stability measure M [14] for FFPE-derived RNA and RPL7A had a poor stability measure when RNA from FF material was tested (data not shown). For these reasons GAPDH, GUSB, RPLP0, TFRC and UBB were used as reference genes in this study.

Our own RNA isolation protocol was compared to RNA that was isolated from the same material but according to commercial protocols and products (Qiagen RNeasy FFPE and ncLysis system of Applied Biosystems). Additional products for FFPE material from commercial providers (e.g. Stratagene, Ambion) were tested and the results obtained with our own protocol were superior to all tested commercial products (data not shown).

We determined module scores for each of the 14 tumors in this study. The limited number of samples does not allow statements about the clinical significance of module scores but they can be used to compare scores computed from intact RNA from FF material and RNA isolated from FFPE according to our own protocol. Pearson correlations between these RNAs in the 14 tumors were 0.966, 0.856 and 0.833 for ER, HER2 and Proliferation scores, respectively. As kryo-preserved RNA and RNA from FFPE material always originated from different portions of the same tumor, a certain variation of gene expression cannot be excluded and, as a consequence, part of the observed variability between kryo and FFPE material may be attributed to biological heterogeneity in the tumors. The three module scores were combined to a Total score. The Total score is similar to the recurrence score described by Paik [11], with high expression of genes related to proliferation and HER2 and low expression of ER-related genes indicating higher risk.

The data generated from FF and FFPE material were also compared to ER and HER2 levels assessed by IHC results from the same tumors. Three tumors (#15, #18, #20) were ER-negative and one was strongly HER2-positive (#6) (Tab. 2). The same tumors had low ER scores when assessed by QPCR (Fig. 8). Tumor #6 had a high HER2 score and an intermediate ER score. These results are in good agreement with the expected distribution of the three scores [15,39]. By comparing QPCR based data with well known tumor subtypes allowed to validate the protocols developed here, even if no new biological findings are provided. The primary issue of this work was to document that stable and robust expression values can be determined from FFPE-derived RNA which are close to the values computed from intact RNA of the same tumors. The optimization and validation of the scoring procedure remains an important issue but obviously, the available number of samples is not sufficient to deal with this aspect and it will be addressed separately and on a larger collection of samples.

While IHC results are at most semi-quantitative, QPCR-based results reflect more accurately the expression level of genes in question. The module scores proposed here integrate quantitative gene expression data from several genes, this makes the resulting scores more robust than measurements based on single genes. QPCR is not only quantitative, it is also very sensitive over a large dynamic range. The number of genes which can be measured by QPCR is not limited and additional genes and module scores can be included in the analysis if this will be required.

Importantly, certain predictive parameters still cannot be determined with current technologies. For example, breast cancers are classified into histological grade 1, 2 or 3. This grading most likely reflects the proliferative state of tumor cells [40]. Grading may be especially important as high grade tumors seem to respond more favorably to chemotherapy than low grade tumors. Unfortunately, many tumors are histological grade 2 and for those tumors the benefit is not clear. Paik and co-workers documented that their recurrence score (RS) was also predictive for a response to chemotherapy [41]. The RS defined by Paik and coworkers is composed of 16 test genes mainly representing ER response genes, proliferation-associated genes, HER2-related genes and invasion genes and 5 genes for normalization [11,41].

The genes selected for this study (Tab. 1) were selected from published DNA chip studies with breast cancer samples [15]. They mostly coincide with the genes used by Paik ad co-workers.

Conclusion

The results presented in this study reveal that RNA isolated from FFPE material according to the protocol developed in our laboratory can be used for expression measurements by QPCR although the RNA is partially degraded. The optimized isolation and de-modification procedures combined with a normalization procedure results in stable and robust gene expression data. Robustness of results was further increased by computing scores from several genes representing the hormonal and the proliferation status of the tumor. Molecular profiling from FFPE material may be of interest for routine diagnostics in the near future as FFPE material is always available [42]. Similarly, molecular profiling from FFPE material may be of great interest in the context of existing and newly planed clinical trials for which only formalin-fixed samples exist.

Abbreviations

ER, estrogen receptor; FF, fresh frozen tissue; FFPE, formalin-fixed, paraffin-embedded tissue; IHC, immunohistochemistry; PGR, progesterone receptor; QPCR, quantitative polymerase chain reaction.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

AO and AB developed the procedure, performed validation studies and were involved in all the experiments presented here. HJA and SA contributed clinical and pathological information. JA and SM participated in the design of the study and VP and MD performed the statistical analysis and developed the scoring system. RJ conceived the study, participated in its design and coordination and drafted the manuscript. All the authors read and approved the final manuscript.

Additional material

Additional file 1

Comparison of normalized expression for each gene in FF and FFPE material*. Expression was determined by QPCR from RNA derived of FF and FFPE material (own protocol). Normalized expression levels (see Methods for details) are shown for each gene and the 14 tumors as polygonal plots.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1755-8794-1-9-S1.jpeg]

Additional file 2

Polygonal representation of ER, HER2, Proliferation and Total scores*. Gene expression was measured from RNA derived of FF and FFPE material (own protocol) and ER, HER2, proliferation and Total scores were computed for each RNA of the 14 tumors and results are shown as polygonal plots. Parallel lines indicate good correlations and crossing lines are indicative for discrepancies between scores computed from FF and FFPE-derived RNA*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1755-8794-1-9-S2.jpeg]

Acknowledgements

The authors thank Dr. R. Haener, University of Bern for helpful suggestions on formalin-induced modification of RNA, Drs. M. Müller, C. Meier and A. Günthert for providing tumors for the tumorbank, Drs. A. Fleischmann and H. Burger for providing FFPE tumor material, and Dr. M. Schobesberger for critical reading of the manuscript. This work was supported by the Swiss Cancer League (to RJ and MD), the NCCR "Molecular Oncology" (to MD), the Bernese Cancer League and Applied Biosystems, Rotkreuz Switzerland (to RJ). Kryo-preserved material was provided by the Tumorbank Bern. The Tumorbank Bern is sponsored by the Department of Clinical Research, the Institute of Pathology, University of Bern, and the Bernese Cancer League. Informed consent was provided from the patients for all the samples used in this study.

References

1. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci U S A* 2001, **98(19):**10869-10874.
2. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci U S A* 2003, **100(14):**8418-8423.
3. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: **Breast cancer classification and prognosis based on gene expression profiles from a population-based study.** *Proc Natl Acad Sci U S A* 2003, **100(18):**10393-10398.
4. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406(6797):**747-752.
5. Rampaul RS, Pinder SE, Elston CW, Ellis IO: **Prognostic and predictive factors in primary breast cancer and their role in patient management: The Nottingham Breast Team.** *Eur J Surg Oncol* 2001, **27(3):**229-238.
6. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Thurlimann B, Senn HJ: **Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005.** *Ann Oncol* 2005, **16(10):**1569-1583.
7. Sotiriou C, Powles TJ, Dowsett M, Jazaeri AA, Feldman AL, Assersohn L, Gadisetti C, Libutti SK, Liu ET: **Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer.** *Breast Cancer Res* 2002, **4(3):**R3.
8. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871):**530-536.
9. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, al. : **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol* 2006, **24(9):**1151-1161.
10. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, Klijn JG, Larsimont D, Buyse M, Bontempi G, Delorenzi M, Piccart MJ, Sotiriou C: **Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.** *J Clin Oncol* 2007, **25(10):**1239-1246.
11. Paik S: **Molecular profiling of breast cancer.** *Curr Opin Obstet Gynecol* 2006, **18(1):**59-63.
12. Cronin M, Pho M, Dutta D, Stephans JC, Shak S, Kiefer MC, Esteban JM, Baker JB: **Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay.** *Am J Pathol* 2004, **164(1):**35-42.

13. Antonov J, Goldstein DR, Oberli A, Baltzer A, Pirotta M, Fleischmann A, Altermatt HJ, Jaggi R: **Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization.** *Lab Invest* 2005, **85(8):**1040-1050.
14. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3(7):**RESEARCH0034.
15. Wirapati P, Kunkel S, Goldstein DG, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Sengstag T, Schütz F, Piccart M, Sotiriou C, Delorenzi M: **Integrative analysis of gene-expression profiles: toward a unified understanding of breast cancer subtyping and prognosis signatures.** 2007 in press.
16. Shibutani M, Uneyama C, Miyazaki K, Toyoda K, Hirose M: **Methacarn fixation: a novel tool for analysis of gene expressions in paraffin-embedded tissue specimens.** *Lab Invest* 2000, **80(2):**199-208.
17. Abrahamsen HN, Steiniche T, Nexo E, Hamilton-Dutoit SJ, Sorensen BS: **Towards quantitative mRNA analysis in paraffin-embedded tissues using real-time reverse transcriptase-polymerase chain reaction: a methodological study on lymph nodes from melanoma patients.** *J Mol Diagn* 2003, **5(1):**34-41.
18. Godfrey TE, Kim SH, Chavira M, Ruff DW, Warren RS, Gray JW, Jensen RH: **Quantitative mRNA expression analysis from formalin-fixed, paraffin-embedded tissues using 5' nuclease quantitative reverse transcription-polymerase chain reaction.** *J Mol Diagn* 2000, **2(2):**84-91.
19. Houze TA, Gustavsson B: **Sonification as a means of enhancing the detection of gene expression levels from formalin-fixed, paraffin-embedded biopsies.** *Biotechniques* 1996, **21(6):**1074-8, 1080, 1082.
20. Koopmans M, Monroe SS, Coffield LM, Zaki SR: **Optimization of extraction and PCR amplification of RNA extracts from paraffin-embedded tissue in different fixatives.** *J Virol Methods* 1993, **43(2):**189-204.
21. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P: **Unlocking the archive--gene expression in paraffin-embedded tissue.** *J Pathol* 2001, **195(1):**66-71.
22. Reichmuth C, Markus MA, Hillemanns M, Atkinson MJ, Unni KK, Saretzki G, Hofler H: **The diagnostic potential of the chromosome translocation t(2;13) in rhabdomyosarcoma: a Pcr study of fresh-frozen and paraffin-embedded tumour samples.** *J Pathol* 1996, **180(1):**50-57.
23. Rupp GM, Locker J: **Purification and analysis of RNA from paraffin-embedded tissues.** *Biotechniques* 1988, **6(1):**56-60.
24. Specht K, Richter T, Muller U, Walch A, Werner M, Hofler H: **Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue.** *Am J Pathol* 2001, **158(2):**419-429.
25. Stanta G, Bonin S: **RNA quantitative analysis from fixed and paraffin-embedded tissues: membrane hybridization and capillary electrophoresis.** *Biotechniques* 1998, **24(2):**271-276.
26. Thomazy VA, Luthra R, Uthman MO, Davies PJ, Medeiros LJ: **Determination of cyclin D1 and CD20 mRNA levels by real-time quantitative RT-PCR from archival tissue sections of mantle cell lymphoma and other non-Hodgkin's lymphomas.** *J Mol Diagn* 2002, **4(4):**201-208.
27. Mies C: **A simple, rapid method for isolating RNA from paraffin-embedded tissues for reverse transcription-polymerase chain reaction (RT-PCR).** *J Histochem Cytochem* 1994, **42(6):**811-813.
28. Bibikova M, Talantov D, Chudin E, Yeakley JM, Chen J, Doucet D, Wickham E, Atkins D, Barker D, Chee M, Wang Y, Fan JB: **Quantitative gene expression profiling in formalin-fixed, paraffin-embedded tissues using universal bead arrays.** *Am J Pathol* 2004, **165(5):**1799-1807.
29. Bibikova M, Chudin E, Arsanjani A, Zhou L, Garcia EW, Modder J, Kostelec M, Barker D, Downs T, Fan JB, Wang-Rodriguez J: **Expression signatures that correlated with Gleason score and relapse in prostate cancer.** *Genomics* 2007, **89(6):**666-672.
30. Haller AC, Kanakapalli D, Walter R, Alhasan S, Eliason JF, Everson RB: **Transcriptional profiling of degraded RNA in cryopreserved and fixed tissue samples obtained at autopsy.** *BMC Clin Pathol* 2006, **6:**9.
31. Rait VK, Zhang Q, Fabris D, Mason JT, O'Leary TJ: **Conversions of formaldehyde-modified 2'-deoxyadenosine 5'-monophosphate in conditions modeling formalin-fixed tissue dehydration.** *J Histochem Cytochem* 2006, **54(3):**301-310.
32. Koch I, Slotta-Huspenina J, Hollweck R, Anastasov N, Hofler H, Quintanilla-Martinez L, Fend F: **Real-time quantitative RT-PCR shows variable, assay-dependent sensitivity to formalin fixation: implications for direct comparison of transcript levels in paraffin-embedded tissues.** *Diagn Mol Pathol* 2006, **15(3):**149-156.
33. Hamatani K, Eguchi H, Takahashi K, Koyama K, Mukai M, Ito R, Taga M, Yasui W, Nakachi K: **Improved RT-PCR amplification for molecular analyses with long-term preserved formalin-fixed, paraffin-embedded tissue specimens.** *J Histochem Cytochem* 2006, **54(7):**773-780.
34. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K: **Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples.** *Nucleic Acids Res* 1999, **27(22):**4436-4443.
35. Chaw YF, Crane LE, Lange P, Shapiro R: **Isolation and identification of cross-links from formaldehyde-treated nucleic acids.** *Biochemistry* 1980, **19(24):**5525-5531.
36. Orlando V, Strutt H, Paro R: **Analysis of chromatin structure by in vivo formaldehyde cross-linking.** *Methods* 1997, **11(2):**205-214.
37. Fleige S, Walf V, Huch S, Prgomet C, Sehm J, Pfaffl MW: **Comparison of relative mRNA quantification models and the impact of RNA integrity in quantitative real-time RT-PCR.** *Biotechnol Lett* 2006, **28(19):**1601-1613.
38. Andersen CL, Jensen JL, Orntoft TF: **Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets.** *Cancer Res* 2004, **64(15):**5245-5250.
39. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM: **Concordance among gene-expression-based predictors for breast cancer.** *N Engl J Med* 2006, **355(6):**560-569.
40. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M: **Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.** *J Natl Cancer Inst* 2006, **98(4):**262-272.
41. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, Costantino JP, Geyer CE Jr., Wickerham DL, Wolmark N: **Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer.** *J Clin Oncol* 2006, **24(23):**3726-3734.
42. Sun Y, Goodison S, Li J, Liu L, Farmerie W: **Improved breast cancer prognosis through the combination of clinical and genetic markers.** *Bioinformatics* 2007, **23(1):**30-37.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1755-8794/1/9/prepub

15 Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue

- BMC Cancer, 10(37), 2010
- IF: 3.288
- number of citations: 6
- personal contribution (30%): design of genomic signatures and scores, statistical analyses, manuscript writing

Antonov *et al. BMC Cancer* 2010, **10**:37
http://www.biomedcentral.com/1471-2407/10/37

BMC Cancer

RESEARCH ARTICLE Open Access

Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue

Janine Antonov[1†], Vlad Popovici[2†], Mauro Delorenzi[2], Pratyaksha Wirapati[2], Anna Baltzer[1], Andrea Oberli[1], Beat Thürlimann[3,8], Anita Giobbie-Hurder[4], Giuseppe Viale[5], Hans Jörg Altermatt[6], Stefan Aebi[1,7,8], Rolf Jaggi[1*]

Abstract

Background: The purpose of the work reported here is to test reliable molecular profiles using routinely processed formalin-fixed paraffin-embedded (FFPE) tissues from participants of the clinical trial BIG 1-98 with a median follow-up of 60 months.

Methods: RNA from fresh frozen (FF) and FFPE tumor samples of 82 patients were used for quality control, and independent FFPE tissues of 342 postmenopausal participants of BIG 1-98 with ER-positive cancer were analyzed by measuring prospectively selected genes and computing scores representing the functions of the estrogen receptor (eight genes, ER_8), the progesterone receptor (five genes, PGR_5), Her2 (two genes, HER2_2), and proliferation (ten genes, PRO_10) by quantitative reverse transcription PCR (qRT-PCR) on TaqMan Low Density Arrays. Molecular scores were computed for each category and ER_8, PGR_5, HER2_2, and PRO_10 scores were combined into a RISK_25 score.

Results: Pearson correlation coefficients between FF- and FFPE-derived scores were at least 0.94 and high concordance was observed between molecular scores and immunohistochemical data. The HER2_2, PGR_5, PRO_10 and RISK_25 scores were significant predictors of disease free-survival (DFS) in univariate Cox proportional hazard regression. PRO_10 and RISK_25 scores predicted DFS in patients with histological grade II breast cancer and in lymph node positive disease. The PRO_10 and PGR_5 scores were independent predictors of DFS in multivariate Cox regression models incorporating clinical risk indicators; PRO_10 outperformed Ki-67 labeling index in multivariate Cox proportional hazard analyses.

Conclusions: Scores representing the endocrine responsiveness and proliferation status of breast cancers were developed from gene expression analyses based on RNA derived from FFPE tissues. The validation of the molecular scores with tumor samples of participants of the BIG 1-98 trial demonstrates that such scores can serve as independent prognostic factors to estimate disease free survival (DFS) in postmenopausal patients with estrogen receptor positive breast cancer.

Trial Registration: Current Controlled Trials: NCT00004205

Background

Clinical and histopathological factors such as lymph node status, tumor size, histological grade, age, and expression of estrogen receptor (ER) and Her2 have traditionally guided treatment decisions of patients with operable breast cancer [1,2]. Various prognostic models are based on these factors, for example the Nottingham Prognostic Index (NPI) [3,4], Adjuvant!Online [5,6] and others [7]. Despite providing excellent estimates of the average risk of recurrence, there remains substantial variation in outcome which may be explained by molecular differences among these tumors [8,9].

DNA-chip based expression analyses have confirmed the heterogeneity of breast cancer and allowed the development of clinically relevant gene "signatures" or

* Correspondence: rolf.jaggi@dkf.unibe.ch
† Contributed equally
[1] Department of Clinical Research, University of Bern, Bern, Switzerland

"profiles" [10-20]. Such profiles are being implemented widely in routine patient care even though many signatures were developed and validated on heterogeneous patient cohorts with respect to stage of disease and therapy. The utility of gene signatures as part of the decision making process is being validated in ongoing studies (TAILORx [21] and MINDACT [22]). Most profiling studies are based on fresh-frozen (FF) or RNAlater conserved tissue. Such material must be collected and processed separately after surgery, complicating the implementation of molecular analyses into the clinical workflow. Procedures based on formalin-fixed, paraffin-embedded (FFPE) material simplify the acquisition of tumor material and can easily be established as part of the routine pathological procedures. In addition, FFPE tissues collected in the framework of clinical trials could be a valuable resource for future research.

We prospectively selected genes from publicly available microarray data and developed molecular scores representing the ER, progesterone receptor (PgR), Her2 and proliferation (PRO) status, and the overall risk of recurrence (RISK). The reproducibility and robustness of the molecular scores was validated by comparing expression data with RNA from FF and FFPE material of 82 tumors. Molecular scores were determined from 342 ER positive tumor samples of the BIG 1-98 clinical trial. Multivariate Cox proportional hazard models revealed that molecular scores are independent prognostic factors to estimate disease free survival (DFS).

Methods

To assess the quality of expression profiling from FFPE material, matched FF and FFPE samples from 82 human breast cancers were used. Histopathological information was irreversibly anonymized according to Swiss law. Independent FFPE blocks and corresponding clinical data of 437 Swiss participants of the trial BIG 1-98 were provided by the International Breast Cancer Study Group. The ethics committees and required health authorities of each participating institution approved the study protocol, and all patients gave written informed consent (ClinicalTrials.gov number, NCT00004205) [23]. Retrospective tissue collection was carried out in accordance with institutional guidelines and national laws. The patient and tumor characteristics of these patients were similar to the entire BIG 1-98 population (Table 1). BIG 1-98 is a randomized controlled clinical trial of adjuvant hormonal therapy for postmenopausal patients with endocrine-responsive breast cancer comparing 4 arms: 5 years of tamoxifen, 5 years of letrozole, two years of tamoxifen followed by 3 years of letrozole, or vice versa [24-26]. All the patients from the BIG 1-98 were treated by mastectomy or breast conserving surgery [24-26]. The available paraffin blocks contained material derived from representative tumor regions.

Tissue samples and data processing

The RNA was isolated from 4 sections (25 μm) of FF material and from 10 paraffin sections (10 μm thick) as described previously [27]. After demodification, the RNA was bound to silica-based columns, DNase I digested and eluted with water. The protocols and reagents for RNA isolation from FF and FFPE tissues were recently incorporated in commercial protocols (RNAready and FFPE RNAready, AmpTec, Hamburg, Germany). RNA qualities were assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA). RNA prepared from FF material had a RIN>6 (RNA integrity number), the RIN of RNA from FFPE was 2-3. The percentage of tumor cells in each FFPE block was evaluated on stained tissue sections. From 437 available FFPE samples 43 samples (9.8%) with less than ~30% tumor cells, 10 ER-negative tumor samples and 7 samples (1.6%) with less than 1.5 μg total RNA recovery were excluded from further analysis. Approximately 30% of the sections contained 30-50% tumor cells, and about 60% contained 50-100% tumor cells. Each of the remaining RNAs was tested by quantitative reverse transcription PCR (qRT-PCR) with 3 control genes (GUSB, RPLP0 and UBB). The mean of the three raw Cts (cycle thresholds) was determined. In 35 samples (8%) the mean Ct was >31, indicating poor quality of the RNA. These RNAs were excluded from further analyses. For the remaining 342 RNAs (78.3%), the expression of 34 genes (see Table 1) was measured by qRT-PCR on TaqMan Low Density Arrays (TLDAs) (Applied Biosystems, Foster City, CA, USA) using a one step protocol (Invitrogen, Basel, Switzerland) on an Applied Biosystems 7900HT instrument. Technical replicates were performed for several intact and several partially degraded RNAs from FF and FFPE material, respectively. They revealed Pearson correlation coefficients higher than 0.95 for all 34 assays.

Genes with high correlation to the expression of ER, PgR, Her2 and proliferation related genes were prospectively selected from publicly available microarray data [28]. A complete list of microarray data sets used in the meta-analysis is available at ".http://breast-cancer-research.com/content/10/4/R65/table/T1[28] (Additional File 1, Table S1). The scores were defined by giving equal weight to each gene in the four groups (proliferation, estrogen response, progesterone response, Her2 response). Thus, a training set was not used as the scores were based on in silico gene selection.

Raw Ct values were normalized against the mean expression of GUSB, RPLP0 and UBB. Scores for ER (ER_8), PgR (PGR_5), Her2 (HER2_2) and proliferation

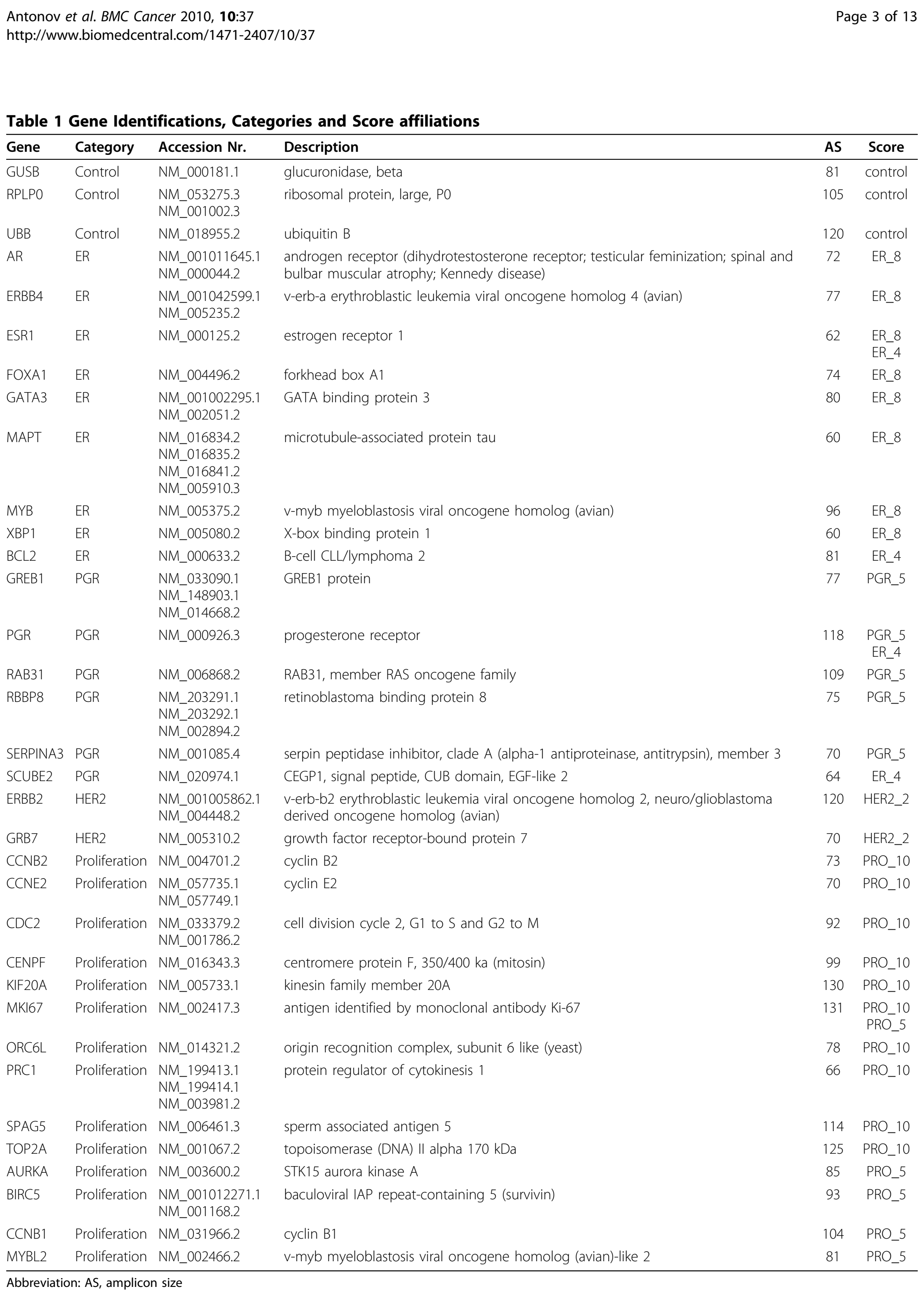


Table 1 Gene Identifications, Categories and Score affiliations

Gene	Category	Accession Nr.	Description	AS	Score
GUSB	Control	NM_000181.1	glucuronidase, beta	81	control
RPLP0	Control	NM_053275.3 NM_001002.3	ribosomal protein, large, P0	105	control
UBB	Control	NM_018955.2	ubiquitin B	120	control
AR	ER	NM_001011645.1 NM_000044.2	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)	72	ER_8
ERBB4	ER	NM_001042599.1 NM_005235.2	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	77	ER_8
ESR1	ER	NM_000125.2	estrogen receptor 1	62	ER_8 ER_4
FOXA1	ER	NM_004496.2	forkhead box A1	74	ER_8
GATA3	ER	NM_001002295.1 NM_002051.2	GATA binding protein 3	80	ER_8
MAPT	ER	NM_016834.2 NM_016835.2 NM_016841.2 NM_005910.3	microtubule-associated protein tau	60	ER_8
MYB	ER	NM_005375.2	v-myb myeloblastosis viral oncogene homolog (avian)	96	ER_8
XBP1	ER	NM_005080.2	X-box binding protein 1	60	ER_8
BCL2	ER	NM_000633.2	B-cell CLL/lymphoma 2	81	ER_4
GREB1	PGR	NM_033090.1 NM_148903.1 NM_014668.2	GREB1 protein	77	PGR_5
PGR	PGR	NM_000926.3	progesterone receptor	118	PGR_5 ER_4
RAB31	PGR	NM_006868.2	RAB31, member RAS oncogene family	109	PGR_5
RBBP8	PGR	NM_203291.1 NM_203292.1 NM_002894.2	retinoblastoma binding protein 8	75	PGR_5
SERPINA3	PGR	NM_001085.4	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	70	PGR_5
SCUBE2	PGR	NM_020974.1	CEGP1, signal peptide, CUB domain, EGF-like 2	64	ER_4
ERBB2	HER2	NM_001005862.1 NM_004448.2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	120	HER2_2
GRB7	HER2	NM_005310.2	growth factor receptor-bound protein 7	70	HER2_2
CCNB2	Proliferation	NM_004701.2	cyclin B2	73	PRO_10
CCNE2	Proliferation	NM_057735.1 NM_057749.1	cyclin E2	70	PRO_10
CDC2	Proliferation	NM_033379.2 NM_001786.2	cell division cycle 2, G1 to S and G2 to M	92	PRO_10
CENPF	Proliferation	NM_016343.3	centromere protein F, 350/400 ka (mitosin)	99	PRO_10
KIF20A	Proliferation	NM_005733.1	kinesin family member 20A	130	PRO_10
MKI67	Proliferation	NM_002417.3	antigen identified by monoclonal antibody Ki-67	131	PRO_10 PRO_5
ORC6L	Proliferation	NM_014321.2	origin recognition complex, subunit 6 like (yeast)	78	PRO_10
PRC1	Proliferation	NM_199413.1 NM_199414.1 NM_003981.2	protein regulator of cytokinesis 1	66	PRO_10
SPAG5	Proliferation	NM_006461.3	sperm associated antigen 5	114	PRO_10
TOP2A	Proliferation	NM_001067.2	topoisomerase (DNA) II alpha 170 kDa	125	PRO_10
AURKA	Proliferation	NM_003600.2	STK15 aurora kinase A	85	PRO_5
BIRC5	Proliferation	NM_001012271.1 NM_001168.2	baculoviral IAP repeat-containing 5 (survivin)	93	PRO_5
CCNB1	Proliferation	NM_031966.2	cyclin B1	104	PRO_5
MYBL2	Proliferation	NM_002466.2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	81	PRO_5

Abbreviation: AS, amplicon size

(PRO_10) were defined as mean expression of all genes in each category (Table 1). A RISK score comprising 25 genes was calculated as follows: RISK_25 = PRO_10 +HER2_2-(8 × ER_8+5 × PGR_5)/13. For comparison, ER_4 and PRO_5 scores were calculated based on 4 and 5 genes described previously [27]. The genes corresponding to ER_4 and PRO_5 scores corresponded to the genes used for calculating the recurrence score (RS) [29].

Concordance of molecular scores and pathological parameters

Histopathological data of BIG 1-98 samples were derived from a central review, with the exception of the grade which was locally assessed. The ER and PgR status were dichotomized into positive (≥ 10% immunoreactive cells) or negative (<10%) [30]. Her2 was measured by fluorescence in-situ hybridization or immunohistochemistry (IHC) and tumors were classified according to Rasmussen et al. [31]. The Ki-67 labeling index (LI) was centrally assessed by IHC as described and classified into low or high using the median LI (11%) as cut-off [32]. The same assays and cut-offs were used for the 82 matched samples with the exception of Her2 which was measured using the CB11 monoclonal antibody and using a cut-off of ≥ 50% [33]. Continuous molecular scores were compared to binary IHC parameters using the area under the curve (AUC). The 95% confidence intervals (CI) were estimated by a bootstrap method (100 bootstraps). Two-sided Mann-Whitney tests were used to assess the association between clinicopathological factors and scores.

Statistical analyses

Primary endpoint of survival analyses was DFS as defined previously [25]. Forty-five events were observed in 342 patients with a median follow-up time (estimated by reverse Kaplan-Meier [34]) of 60 months. DFS was estimated by Kaplan Meier analysis. Patients were classified into low and high PRO or RISK scores using the corresponding median score as cut-off. The differences in survival experience between the two resulting groups were assessed with log rank tests. Univariate and multivariate Cox proportional hazard models were used [35] and hazard ratios (HR), CIs and p-values were obtained. The multivariate models were assessed using the log-likelihood and the deviance of residuals. Likelihood ratio tests (LRT) were used to compare different nested multivariate models. No adjustments were made for multiple testing. Univariate Cox proportional hazard models were applied to estimate the rate of events and to produce corresponding plots.

Results

Reliable expression profiling from FFPE tumor tissue

Gene expression was measured from 34 genes using TLDAs with RNA isolated from FF and FFPE material of 82 breast cancers. These data were used solely for the assessment of the expression profiling from FFPE material. Pearson correlation coefficients between FF and FFPE expression values for each tumor and all assays ranged from 0.91 to 0.98. The mean increase of raw Ct values derived of FFPE compared to matched FF tissues was 1.30 units. This Ct shift was mostly compensated by normalization (Additional File 2, Figure S1. and Additional File 3, Figure S2).

Unsupervised hierarchical clustering demonstrated the stability of gene clusters and revealed an excellent agreement between FF- and FFPE-based expression profiles (Additional File 4, Figure S3). Molecular scores were determined for ER, PGR, HER2 and PRO. A linear relationship of scores was found for RNA from FF and RNA from FFPE material (Figure 1). Pearson correlation coefficients for the four scores were 0.968, 0.974, 0.942 and 0.944, respectively. The distributions of ER_8, PGR_5 and HER2_2 scores are shown as histograms together with the fitted mixture of two Gaussian distributions (Additional File 1, Figure S4) used for discriminating the subtypes.

The agreement between molecular scores and corresponding binary IHC variables was assessed by receiver operating characteristic (ROC) curves and AUC. AUCs and 95% CI were calculated for ER_8 (FF = 0.940 (0.835-1.00), FFPE = 0.931 (0.804-1.00)), PGR_5 (FF = 0.919 (0.828-0.986), FFPE = 0.916 (0.806-0.987) and HER2_2 (FF = 0.961 (0.895-1.00), FFPE = 0.963 (0.915-0.993)). PRO_10 was compared with IHC data for Ki-67 using a cut-off of 11% and the resulting AUCs were 0.798 (0.609-0.900) for FF and 0.810 (0.660-0.907) for FFPE, respectively. In conclusion, the agreement of the IHC with FFPE samples was as good as with FF samples.

Concordance between pathological parameters and molecular scores for tumors of the BIG 1-98 clinical trial

Molecular scoring was applied to an independent set of tissue samples from Swiss patients participating in the BIG 1-98 randomized clinical trial and scores were compared to centrally assessed histopathological data by ROC curves. From a total of 437 provided tumor samples 342 ER-positive tumors (78.3%) were suitable for analysis. The AUC was 0.974 (95% CI = 0.946-0.995) for HER2_2 and 0.847 (95% CI = 0.794-0.902) for PGR_5. PRO_10 scores positively correlated with Ki-67 LI (Pearson correlation coefficient 0.51); the AUC was 0.815 (95% CI = 0.768-0.864) for Ki-67 binarized at 11% [32].

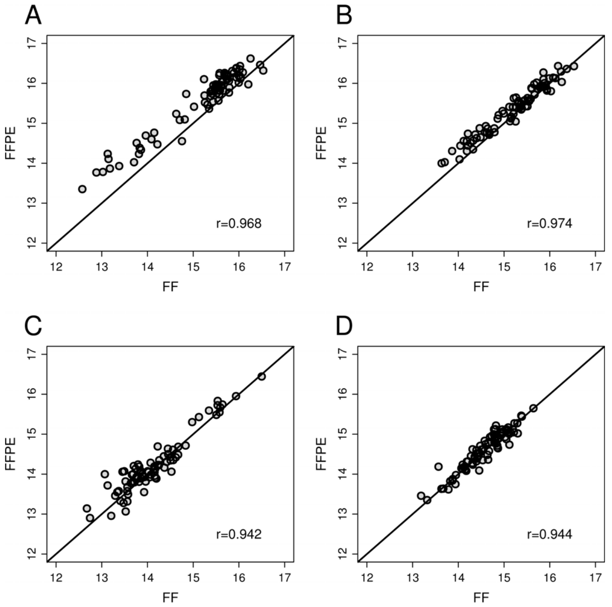


Figure 1 Comparison of scores computed from intact RNA and partially degraded RNA from FFPE material. Scores were determined for RNA from FF material and RNA from corresponding FFPE tumor material of 82 patients. Scatter plots are shown between scores from FF and FFPE tissues representing ER_8 (A), PGR_5 (B), HER2_2 (C) and PRO_10 (D) for each tumor. Pearson correlations are indicated.

The PRO_10 score correlates with histological grade and other clinical factors

The histological grade was assessed according to Elston and Ellis [36]. The PRO_10 score positively correlated with Elston and Ellis scores and with grade (Pearson correlation coefficient 0.453 and 0.409, respectively) (Figure 2). Furthermore, PRO_10 scores were significantly higher in Her2 positive tumors, in tumors larger than 2 cm and in tumors with axillary lymph node metastasis as compared to Her2 negative tumors, T1 tumors and N0 tumors ($p \leq 0.0015$, Mann-Whitney tests), respectively (data not shown).

PRO and RISK scores predict disease free survival in lymph node positive patients and patients with grade II breast cancer

The prognostic values of PRO_10 and RISK_25 scores were assessed by their ability to assign patients to low and high risk groups. Patients were stratified according to histological grade and low or high PRO_10 and RISK_25 scores using the corresponding medians as cut-offs (Figure 3). As expected, patients with grade III tumors had poorer DFS than patients with grade I or grade II tumors (p = 0.0019, panel A). High PRO_10 scores correlated with poorer DFS compared to low scores in all (p = 0.0043, panel B) and in histological grade II tumors (p = 0.0024, panel C). Similarly, RISK_25 discriminated between favorable and poor DFS in all (p = 0.0005, panel D) and in node positive tumors (p = 0.0009, panel E). Univariate Cox proportional hazards regression analysis confirmed these observations.

The PGR_5, PRO_10 and RISK_25 scores were all significant predictors of DFS ($p < 0.05$) as were histological grade, tumor size, number of positive lymph nodes and Ki-67 LI (Table 2). The PRO_5 score was also a significant predictor of DFS but PRO_10 score was numerically better than PRO_5 in terms of log-likelihood (L) and deviance of residuals (D) (PRO_10: L = -223.35, D = 225.83; PRO_5: L = -224.16, D = 227.57).

Figure 4 shows the estimated rate of recurrence as a function of PRO_10, PGR_5 and RISK_25 scores. The PRO_5, PRO_10 and the RISK_25 scores remained significant predictors of DFS when applied to patients with grade II breast cancer.

PRO_10 and PGR_5 scores are independent risk factors in multivariate analyses

The impact of the molecular scores PRO_10 and PGR_5 was further documented in multivariate models

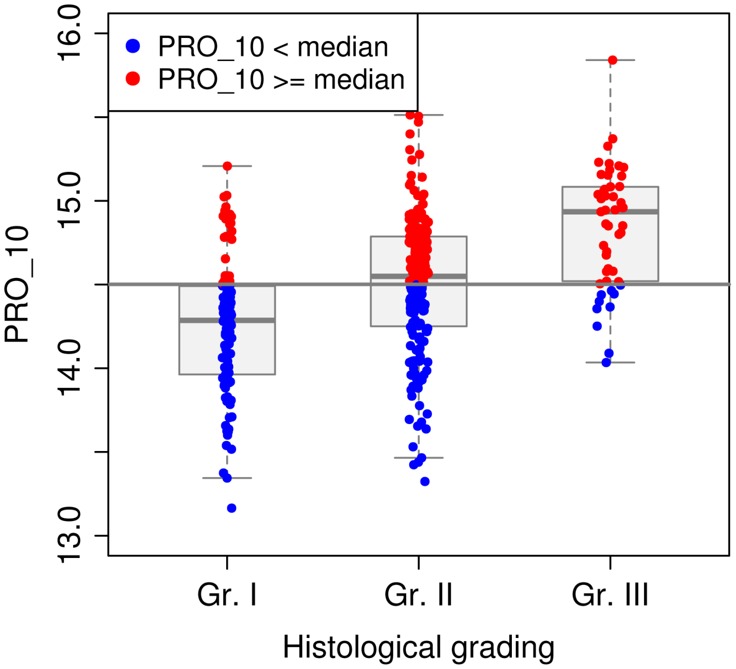


Figure 2 Comparison of scores and immunohistochemical analysis. Correlation of histological grading and PRO_10 score. The 342 tumors were classified according to histological grading. The data are shown as boxplots with median (solid line), interquartile ranges (boxes) and minimum and maximum non-outlier values (whiskers). The PRO_10 scores higher and lower than the median are indicated as red and blue dots, respectively for each grade.

comprising clinicopathologic predictors and molecular scores that were significant in univariate analyses.

Multivariate analyses revealed that PRO_10 is a predictor of DFS independent of tumor size (T), number of positive lymph nodes (N), grade (G) and Ki-67 LI. PRO_10 represents proliferation-related genes and it was of interest to compare it to Ki-67. Table 2 shows the results of multivariate analyses including T, N, G and either Ki-67 (model 1) or PRO_10 (model 3) in comparison with a model containing both markers (model 2). The full model (model 2) was significantly better than model 1 (LRT p = 0.0071). No significant difference was found for PRO_10 between models 2 and 3 (LRT p = 0.8075). Thus, adding PRO_10 to T, N, G and Ki-67 significantly improved the model. In contrast, adding Ki-67 to T, N, G and PRO_10 did not bring additional information.

The same procedure was used to evaluate whether PGR_5 further improved model 6 containing T, N, G and PRO_10 (Table 2). The full model including all 5 variables (model 5) performed better than model 4 (T, N, G, PGR_5; LRT p = 0.0089) and model 6 (T, N, G, PRO_10; LRT p = 0.0339). Both, PGR_5 and PRO_10 remained significant in model 5 suggesting that the two scores contain independent information with respect to prognosis and outcome.

Discussion

Gene expression profilings define clinically relevant gene signatures [15,17,37,38]. For the present work, we selected genes correlating with the ER, PgR, Her2 and proliferative status using a meta-analysis of gene expression profiles [28]. The prognostic power of resulting gene expression scores for ER, PgR, proliferation and overall risk of recurrence was validated using tissues and clinical data from a representative subset of participants of trial BIG 1-98 confirming the correlation structure of these genes and their association with clinical and outcome variables.

Multiple genes representing each score were quantified by qRT-PCR. RNA from 82 matched FF and FFPE tissues were compared by qRT-PCR on TLDAs. The mean increase of raw Ct values between RNA from FF and FFPE tissues was 1.3 units. This is similar to the

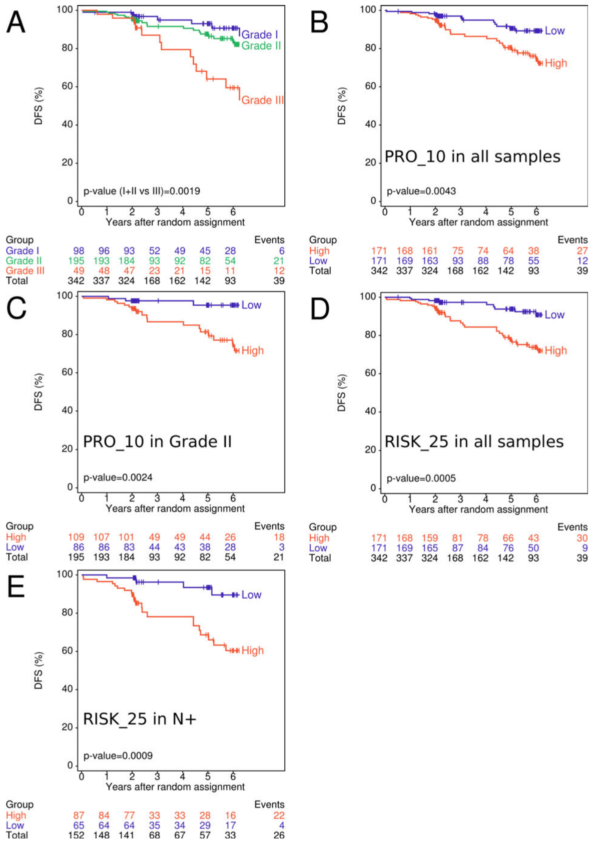


Figure 3 Survival data based on molecular scores. Kaplan-Meier plots for DFS. Patients were stratified into grade I (blue), II (green) and III (red line) (A), into low (blue) and high (red) PRO_10 scores in all samples (B) and in Grade II samples (C). The RISK_25 score is shown for all samples (D) and for tumors of patients with lymph node positive (N+) cancer (E). Median values of the scores were used as cut-offs. The p-values correspond to Log-rank test.

findings of Cronin and co-workers (+2.0 units) in a comparable setting [39]. Duration of formalin fixation, storage time and conditions influence the quality of RNA derived of FFPE tissues with direct effects on the sensitivity of subsequent PCR reactions [40]. However, normalization effectively compensated for this shift of Ct values (Additional File 2, Fig S1 and Additional File 3, Figure S2).

The mean expression of eight genes related to ER and five genes related to PgR were used to calculate the ER_8 and PGR_5 scores. Scores representing different functional categories were combined in RISK_25 score. The molecular scores determined from 82 paired samples of FF and FFPE tumors were highly concordant, as were molecular scores and immunohistochemically assessed parameters demonstrating the reliability of the procedure.

Molecular scores were validated in an independent set of tumor tissues from 342 participants of trial BIG 1-98. In contrast to histological analyses which can also be performed from tissue sections that contain considerable normal, stromal or fat components the architecture of the tissue is completely lost during work up for molecular analyses and therefore, it was important to exclude samples with inadequate tumor content. A histological section was taken from the immediate vicinity of each sample that was used for molecular analyses. Each section was assessed by an experienced pathologist (H.J.A.) and molecular analyses were restricted to samples containing at least 30% tumor cells. For comparison, RNA

Table 2 Baseline characteristics.

Characteristic	Patients with FFPE profiles from Swiss participants used in the study (N = 342)	Provided material of Swiss participants (N = 437)	Patients of the BIG 1-98 population not used in the study (N = 7573)	Overall BIG 1-98 population (N = 8010)
Menopausal category - N (%)				
Postmen. before chemo	321 (93.9)	413 (94.5)	7279 (96.1)	7692 (96.0)
Postmen. after chemo	10 (2.9)	11 (2.5)	181 (2.4)	192 (2.4)
Premenopausal (ineligible)	0 (0.0)	2 (0.5)	21 (0.3)	23 (0.3)
Uncertain status	10 (2.9)	10 (2.3)	92 (1.2)	102 (1.3)
Unknown/ missing	1 (0.3)	1 (0.2)	0	1 (<0.1)
Age at randomization - years				
Median	62	62	61	61
Range	41-86	41-86	38-90	38-90
Tumor size - N (%)				
≤ 2 cm	195 (57.0)	251 (57.4)	4706 (62.1)	4957 (61.9)
> 2 cm	144 (42.1)	179 (41.0)	2794 (36.9)	2973 (37.1)
Unknown/ missing	3 (0.9)	7 (1.6)	73 (1.0)	80 (1.0)
Tumor grade - N (%)				
Grade 1	94 (27.5)	124 (28.4)	2007 (26.5)	2131 (26.6)
Grade 2	196 (57.3)	251 (57.4)	3649 (48.2)	3900 (38.7)
Grade 3	49 (14.3)	59 (13.5)	1166 (15.4)	1225 (15.3)
Unknown/ missing	3 (0.9)	3 (0.7)	751 (9.9)	754 (9.4)
Nodal status - N (%)				
Negative (including Nx)	186 (54.4)	245 (56.1)	4342 (57.3)	4587 (57.3)
Positive	152 (44.4)	188 (43.0)	3123 (41.2)	3311 (41.3)
Unknown/ missing	4 (1.2)	4 (1.0)	108 (1.4)	112 (1.4)
ER and PgR status - N (%)				
ER pos and PgR pos.	268 (78.4)	340 (77.8)	4715 (62.3)	5055 (63.1)
ER pos and PgR neg.	66 (19.3)	87 (19.9)	1544 (20.4)	1631 (20.4)
ER pos and PgR unknown	1 (0.3)	1 (0.2)	1153 (15.2)	1154 (14.4)
ER neg and PgR pos.	5 (1.5)	7 (1.6)	136 (1.8)	143 (1.8)
ER unknown, PGR pos.	0	0	7 (0.1)	7 (0.1)
Other	2 (0.6)	2 (0.5)	18 (0.3)	20 (0.2)
Local therapy - N (%)				
BCS and RT	236 (69.0)	310 (70.9)	3987 (52.7)	4297 (53.7)
BCS and no RT	13 (3.8)	16 (3.7)	228 (3.0)	244 (3.0)

Table 2: Baseline characteristics. *(Continued)*

Mastectomy and RT	24 (7.0)	25 (5.7)	1415 (18.7)	1440 (18.0)
Mastectomy and no RT.	68 (19.9)	85 (19.5)	1926 (25.4)	2011 (25.1)
Other	1 (0.3)	1 (0.2)	17 (0.2)	18 (0.2)
Adjuvant or neoadjuvant				
chemo (or both) - N (%)				
Yes	133 (38.9)	159 (36.4)	1865 (24.6)	2024 (25.3)
No	209 (61.1)	278 (63.6)	5708 (75.4)	5986 (74.7)

Abbreviations: BCS, breast conserving surgery; Nx, nodal status unknown; postmen., postmenopausal; RT, radiotherapy; PgR, progesterone receptor; pos., positive; neg., negative

was also isolated from tumor-surrounding cells which led to rather poor RNA recoveries from comparable tissue areas (data not shown). However, this does not exclude that tumor-surrounding cells may have a limited impact on molecular scores in such analyses. Contamination by non-tumor cells may be reduced by macrodissecting tumors before RNA isolation and molecular assessment. The same procedure would also make tumors accessible to molecular analysis when sections contain less than 30% tumor cells.

Classification of patients by low and high PRO_10 and RISK_25 scores corresponded to low and high risk of recurrence. PRO, RISK and PGR scores were prognostic for DFS not only in the entire patient population but also in a subpopulation of patients with node positive disease (Figure 3D and 3E). We provide evidence independent of Genomic Health™ that a RISK score based on similar biological processes as the recurrence score (RS), but with other genes selected through a different procedure, can predict DFS [29,41,42]. In contrast to the RS which was validated with tamoxifen-treated patients, PRO_10, RISK_25 and PGR_5 scores were validated with patients treated with tamoxifen, letrozole or a sequence of both drugs; therefore, they may apply to patients who received either of these drugs.

Histological grading is an important factor in estimating the risk of recurrence of patients with breast cancer [2,43]. Recently, Sortiriou and colleagues have developed the gene expression grade index (GGI) based on the expression of 97 genes related to proliferation. They demonstrated that grade II cancers are comprised of tumors which are similar to genomic grade I or grade III with corresponding clinical outcomes [16,44]. Our findings agree with these observations as grade II tumors could be further classified into low and high risk of recurrence by 10 genes (PRO_10) (Figure 3C) or even by 5 genes (PRO_5 score) (data not shown). Seven of the PRO_10 and three of the PRO_5 genes are also part of GGI. The PRO_5 genes (Table 1) corresponded to the proliferation-related genes of the RS [29]. The assessment of gene signatures related to proliferation such as GGI or PRO scores is of special interest in ER positive, grade II breast cancer for whom therapeutic decisions are often difficult. Both, GGI and RS were shown to be associated with response to chemotherapy [45,46]. In contrast to GGI which requires FF tumor material, PRO scores or RS can be determined from a few microtome slices or cores such as used for tissue microarrays [47]. Material for molecular analysis can be taken from the same FFPE tissue block used for histological and immunohistochemical analyses without interfering with clinicopathological workflow.

The prognostic value of Ki-67 in early breast cancer was recently confirmed [48]. However, Ki-67 is not used uniformly in clinical practice [49,50] as it appears to be difficult to agree on cut-off values separating high and low proliferation tumors or on its value in assisting the choice of adjuvant therapy [50,51]. Therefore, instead of dichotomizing Ki-67 it may be more feasible to use Ki-67 as continuous variable [52]. Here, we made a comparison between centrally assessed Ki-67 LI and a qRT-PCR based proliferation signature. The PRO_10 score correlated with Ki-67 LI, and both were significant predictors of DFS in univariate Cox analyses. In multivariate models however, PRO_10 offered superior prognostic value and outperformed Ki-67 LI (Table 3). Moreover, the PRO_10 score added independent prognostic information to anatomical staging.

PgR, as measured by immunohistochemistry [30] or microarray analysis [53], was shown to positively correlate with prognosis. Here we show that the molecular PGR_5 score was also positively associated with DFS (Figure 4) and added independent prognostic information to anatomical staging and PRO_10 (Table 3). Thus, PGR_5 and PRO_10 scores independently predict prognosis in the BIG 1-98 population.

Compared to immunohistochemically assessed parameters, qRT-PCR based scores are quantitative,

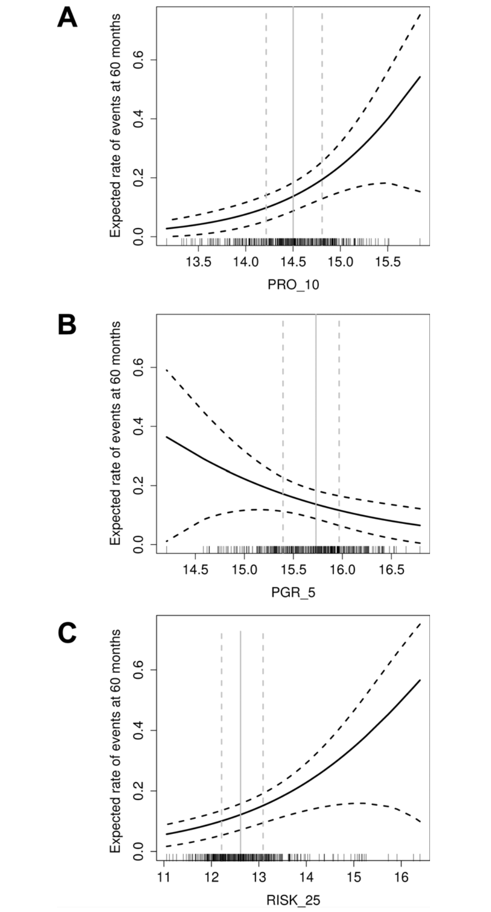


Figure 4 Expected rate of disease-free survival (DFS). The expected rate of events at 60 months (solid line) is shown as a function of PRO_10 (A), PGR_5 (B) and RISK_25 scores (C). The 95% confidence intervals are indicated (dashed lines). Vertical lines represent the median of all scores (solid line) and 25% and 75% quantiles (dashed lines).

relatively independent on operator expertise and less affected by inter-observer variability. The procedure is simple, economical and can be standardized easily with good control genes, reference samples and quality control procedures.

The results of this study are based on a limited number of patients and follow-up time (60 months). Similar

Table 3 Cox Proportional Hazard Analyses.

Covariate	P-value	HR (95% CI)
Univariate Analyses*		
Clinicopathological Variables		
HER2	0.7816	1.18 (0.36 - 3.84)
PgR	0.5147	0.78 (0.36 - 1.66)
Histological grade	0.0032	1.99 (1.26 - 3.14)
Ki-67 LI	0.0226	1.02 (1.00 - 1.04)
Tumor size	0.0047	1.22 (1.06 - 1.39)
Number of positive nodes	<0.0001	1.13 (1.08 - 1.18)
Treatment (4 categories)	0.1540	-
Molecular scores		
HER2_2	0.1080	1.20 (0.96 - 1.51)
PGR_5	0.0344	0.66 (0.44 - 0.97)
PRO_5	0.0003	2.14 (1.42 - 3.22)
PRO_10	<0.0001	2.09 (1.45 - 3.00)
RISK_25	0.0001	1.54 (1.24 - 1.91)
Multivariate Analyses: Comparison of PRO_10 and Ki-67 LI**		
Model 1: log-likelihood = -179.38, Deviance = 188.11		
Number of positive nodes	<0.0001	1.19 (1.12 - 1.27)
Tumor size	0.0370	1.19 (1.01 - 1.39)
Grade	0.4200	1.25 (0.72 - 2.17)
Ki-67 LI	0.1300	1.02 (1.00 - 1.04)
Model 2: log-likelihood = -175.75, Deviance = 180.71		
Number of positive nodes	<0.0001	1.19 (1.12 - 1.27)
Tumor size	0.1300	1.14 (0.96 - 1.34)
Grade	0.9600	0.99 (0.55 - 1.76)
PRO_10	0.0092	2.12 (1.20 - 3.72)
Ki-67 LI	0.8100	1.00 (0.97 - 1.03)
Model 3: log-likelihood = -175.78, Deviance = 180.77		
Number of positive nodes	<0.0001	1.19 (1.12 - 1.27)
Tumor size	0.1200	1.14 (0.97 - 1.34)
Grade	0.9400	0.98 (0.55 - 1.74)
PRO_10	0.0026	2.03 (1.28 - 3.23)
Multivariate Analyses: Role of PGR_5***		
Model 4: log-likelihood = -215.27, Deviance = 214.30		
Number of positive nodes	<0.0001	1.12 (1.07 - 1.16)
Tumor size	0.2000	1.11 (0.95 - 1.30)
Grade	0.0170	1.78 (1.11 - 2.87)
PGR_5	0.0570	0.68 (0.45 - 1.01)
Model 5: log-likelihood = -211.85, Deviance = 208.03		
Number of positive nodes	<0.0001	1.06 (1.06 - 1.16)
Tumor size	0.4300	1.07 (0.91 - 1.26)
Grade	0.3000	1.32 (0.78 - 2.23)
PRO_10	0.0092	1.73 (1.15 - 2.62)
PGR_5	0.0360	0.65 (0.43 - 0.97)
Model 6: log-likelihood = -214.10, Deviance = 211.25		
Number of positive nodes	<0.0001	1.11 (1.06 - 1.16)
Tumor size	0.1700	1.13 (0.95 - 1.34)

Table 3: Cox Proportional Hazard Analyses. *(Continued)*

Grade	0.2100	1.40 (0.83 - 2.37)
PRO_10	0.0150	1.71 (1.11 - 2.62)

*Histological grading was analyzed according to three categories (histological grade I, II or III). Number of lymph node metastases and tumor size were continuous variables. PgR and Her2 were centrally assessed and binary IHC data were included in the analyses [30,31]. Centrally assessed Ki-67 labeling index and molecular scores were included as continuous variables.

**Data of 299 patients with available Ki-67 LI were included in model 1, 2 and 3, respectively.

***Data of all 342 patients were included in model 4, 5 and 6, respectively.

Abbreviations: HR, hazard ratio; CI, confidence interval; LRT, likelihood ratio test; Ki-67 LI, Ki-67 labeling index.

Models 3 and 6 should not be compared directly as they were fitted on different sample sizes, due to missing data in Ki-67 LI.

analyses with independent, larger sample sizes and more mature follow-up data are planned to further consolidate the prognostic and possibly predictive value of the proposed scores in each treatment arm separately.

Gene expression profiling has improved the understanding of molecular subtypes of breast cancer. FFPE material is not widely used although it may facilitate and speed up the development and validation of novel gene signatures due to the availability of well-characterized tissues from numerous clinical trials [54,55]. The same material can be used for molecular diagnostics. The investigation of gene signatures may become more important in the future as an increasing proportion of agents under development for breast cancer treatment have defined molecular targets. Early integration of biomarker analysis in the drug development process has the potential to improve the specificity and efficiency of novel therapeutics. This opens the possibility to further individualize therapy of patients with breast cancer.

Conclusions

We define four molecular scores based on quantitative measurement of gene expression with RNA derived of FFPE tissues. The genes for each score were selected from a large meta-analysis of microarrays. The genes do not coincide with genes used for other molecular scores like the RS (except genes that were previously used as immunohistochemical markers such as ER, PgR or Her2). Two of the described scores are shown to be independent predictors of disease-free survival of postmenopausal patients with operable, estrogen receptor positive breast cancer. The proliferation-associated score outperforms the Ki-67 labeling index measured by immunohistochemistry.

List of abbreviations

AUC: area under the (ROC) curve; CI: confidence interval; DFS: disease-free survival; ER: estrogen receptor; FF: fresh frozen; FFPE: formalin-fixed, paraffin embedded; HR: hazard ratio; IHC: immunohistochemistry; GGI: gene expression grade index; LI: labeling index; LRT: likelihood ratio tests; PCR: polymerase chain reaction; RIN: RNA integrity number; PgR: progesterone receptor; ROC: receiver operating characteristic; RS: recurrence score; TLDA: TaqMan Low Density Arrays.

Additional file 1: Publicly available gene expression data from breast cancer studies.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-10-37-S1.PDF]

Additional file 2: Effect of normalization. Mean expression of 34 assays determined for 82 RNAs isolated from FFPE and from corresponding FF tissue. Shown are the differences between FFPE and FF before (Raw) and after normalization against the mean of three control genes (UBB, RPLP0 and GUSB) (Normalized).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-10-37-S2.PDF]

Additional file 3: Unsupervised hierarchical clustering of data from FF- and FFPE-derived RNA. Shown are heat maps based on normalized expression from RNA of FF (A) and FFPE tissues (B). Proliferation (red box), Her2 (blue box) and ER or PgR related genes (green box) are indicated. The hormone receptor status of each tumor was also assessed by IHC. ER negative (closed circles) and Her2 positive tumors (open circles) are indicated.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-10-37-S3.PDF]

Additional file 4: Distribution of molecular scores. Shown are histograms of ER, PGR and HER2 scores and fitted mixtures of Gaussian distributions. Results of 82 matched samples are shown for ER_8 (A, B), PGR_5 (C, D) and HER2_2 (E, F) scores derived from FF (A, C, E) and FFPE tissues (B, D, F).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-10-37-S4.PDF]

Acknowledgements
We thank all patients, physicians and pathologists who participated in the BIG 1-98 trial. We thank the IBCSG for providing material and corresponding clinicopathological data. We thank Dr. M. Schobesberger and K. Price for helpful discussions and critical reading of the manuscript and R. Kammler for preparing FFPE material. This work was supported by Cancer Research Switzerland/Oncosuisse, the NCCR "Molecular Oncology" and the Bernese Cancer League. The IBCSG Statistical Center receives funding from the US NCI (CA-75362). The BIG 1-98 trial was financed by Novartis and coordinated by the International Breast Cancer Study Group (funded by the Swedish Cancer Society, The Cancer Council Australia, Australian New Zealand Breast Cancer Trials Group, Frontier Science and Technology Research Foundation, Swiss Group for Clinical Cancer Research (SAKK), Cancer Research Switzerland/Oncosuisse and the Foundation for Clinical Cancer Research of Eastern Switzerland (OSKK).

BIG 1-98 Steering Committee: B. Thürlimann (Chair), S. Aebi, L. Blacher, M. Castiglione, A. S. Coates, T. Cufer, P. Dinh, J. F. Forbes, R. D. Gelber, A. Giobbie-Hurder, A. Goldhirsch, A. Hiltbrunner, S. B. Holmberg, R. Maibach, A. Martoni, L. Mauriac, G. McGrogan, H. T. Mouridsen, R. Paridaens, K. N. Price, M. Rabaglio, B.B. Rasmussen, M.M. Regan, A. Santoro, I. E. Smith, A. Wardley, G. Viale. Novartis: H. A. Chaudri-Ross, R. Dias, D. B. Evans, C. Sguotti, U. Trostmann

Swiss Group for Clinical Cancer Research (SAKK): A. Goldhirsch (up to January 2004), R. Herrmann (from June 2004): **Kantonsspital Aarau, Zentrum f. Onkologie, Aarau**: A. Schönenberger, W. Mingrone, Ch. Honegger, E. Bärtschi, M. Neter, M. Rederer, G. Schär; **University Hospital Basel**, Basel: C. Rochlitz, R. Herrmann, D. Oertli, E. Wight, H. Moch; **Institute of Oncology of Southern Switzerland: Ospedale San Giovanni,**

Bellinzona: J. Bernier, L. Bronz, F. Cavalli, E. Gallerani, A. Richetti, A. Franzetti; **Ospedale Regionale di Lugano (Civico & Italiano), Lugano**: M. Conti-Beltraminelli, M. Ghielmini, T. Gyr, S. Mauri, P. C. Saletti; **Ospedale Regionale Beata Vergine, Mendrisio**: A. Goldhirsch, O. Pagani, R. Graffeo, M. Locatelli, S. Longhi, P.C. Rey, M. Ruggeri; **Ospedale Regionale La Carità, Locarno**: E. Zucca, D. Wyss; **Istituto Cantonale di Patologia, Locarno**: L. Mazzucchelli, E. Pedrinis, T. Rusca; **Inselspital, Berne**: S. Aebi, M. F. Fey, M. Castiglione, M. Rabaglio; **Kantonsspital Olten**, Olten: S. Aebi, M. F. Fey, M. Zuber, G. Beck; **Bürgerspital, Solothurn**: S. Aebi, M. F. Fey, R. Schönenberger;**Spital Thun-Simmental AG Thun**: J.M. Lüthi, D. Rauch; **Hôpital Cantonal Universitaire HCUG**, Geneva: H. Bonnefoi; **Rätisches Kantons- und Regionalspital,** Chur: F. Egli, R. Steiner, P. Fehr; **Centre Pluridisciplinaire d'Oncologie, Lausanne**: L. Perey, P. de Grandi, W. Jeanneret, S. Leyvraz, J.-F. Delaloye; **Kantonsspital St. Gallen**, St. Gallen: B. Thürlimann, D. Köberle, F. Weisser, S., Mattmann, A. Müller, T. Cerny, B. Späti, M. Höfliger, G. Fürstenberger, B. Bolliger, C. Öhlschlegel, U. Lorenz, M. Bamert, J. Kehl-Blank, E. Vogel; **Kantonales Spital Herisau**, Herisau: B. Thürlimann, D. Hess, I. Senn, D. Köberle, A. Ehrsam, C. Nauer, C. Öhlschlegel, J. Kehl-Blank, E. Vogel; **Stadtspital Triemli, Zürich**: L. Widmer, M. Häfner; **Universitätsspital Zürich**, Zürich: B. C. Pestalozzi, M. Fehr, R. Caduff, Z. Varga, R. Trüb, D. Fink.

Swiss Private MDs: Private Praxis, Zürich: B. A. Bättig; Sonnenhof-Klinik Engeried, Berne: K. Buser; Frauenklinik Limmattalspital, Schlieren: N. Bürki; Private Praxis, Birsfelden: A. Dieterle; Private Praxis, Biel: L. Hasler; Private Praxis, Baar: M. Mannhart-Harms; Brust-Zentrum, Zürich: C. Rageth; Private Praxis, Berne: J. Richner; Private Praxis, Bellinzona: V. Spataro; Private Praxis, Winterthur: M. Umbricht.

Author details

[1]Department of Clinical Research, University of Bern, Bern, Switzerland. [2]National Center of Competence in Research (NCCR) Molecular Oncology, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland. [3]Senology Center of Eastern Switzerland, Kantonsspital, St. Gallen, Switzerland. [4]International Breast Cancer Study Group Statistical Center, Dana-Farber Cancer Institute, Boston, MA, USA. [5]Division of Pathology and Laboratory Medicine, European Institute of Oncology, University of Milan, Milan, Italy. [6]Pathology Länggasse, Bern, Switzerland. [7]Medical Oncology, University Hospital Bern, Bern, Switzerland. [8]Swiss Group of Clinical Cancer Research (SAKK), Bern, Switzerland.

Authors' contributions

JA, SA and RJ organized the study, planned the experiments and wrote the manuscript. SA and BT organized samples from the International Breast Cancer Study Group. AO and AB carried out RNA isolations, quality controls and gene expression measurements. VP, PW, MD and AGH carried out the statistical analyses. HJA and GV were responsible for histological assessment of stained sections. All authors contributed to the manuscript, they read and approved the final manuscript.

Competing interests

JA, VP, MD, PW, AB, AO, AGH, GV, HJA, SA and RJ declare that they have no competing interest. B.T. holds stocks from Novartis (Ciba Geigy) since 1990.

Received: 10 September 2009
Accepted: 9 February 2010 Published: 9 February 2010

References

1. Carlson RW, Jahanzeb M, Kiel K, Marks LB, Mc Cromick B, Pierce LJ, Ward JH, Topham NS: NCCN Clinical Practice Guidelines in Oncology V.2.2008. *Book NCCN Clinical Practice Guidelines in Oncology V.2* 2008http://www.nccn.org.
2. Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thurlimann B, Senn HJ: Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. *Ann Oncol* 2007, **18**:1133-1144.
3. Blamey RW, Pinder SE, Ball GR, Ellis IO, Elston CW, Mitchell MJ, Haybittle JL: Reading the prognosis of the individual with breast cancer. *Eur J Cancer* 2007, **43**:1545-1547.
4. Galea MH, Blamey RW, Elston CE, Ellis IO: The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992, **22**:207-219.
5. Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, Davis GJ, Chia SK, Gelmon KA: Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *J Clin Oncol* 2005, **23**:2716-2725.
6. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, Parker HL: Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol* 2001, **19**:980-991.
7. Ellis MJ, Tao Y, Luo J, A'Hern R, Evans DB, Bhatnagar AS, Chaudri Ross HA, von Kameke A, Miller WR, Smith I, *et al*: Outcome prediction for estrogen receptor-positive breast cancer based on postneoadjuvant endocrine therapy tumor characteristics. *J Natl Cancer Inst* 2008, **100**:1380-1388.
8. Andre F, Pusztai L: Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy. *Nat Clin Pract Oncol* 2006, **3**:621-632.
9. Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF: Molecular classification of breast cancer: limitations and potential. *Oncologist* 2006, **11**:868-877.
10. Brenton JD, Carey LA, Ahmed AA, Caldas C: Molecular classification and molecular forecasting of breast cancer: ready for clinical application?. *J Clin Oncol* 2005, **23**:7350-7360.
11. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assignies MS, Bergh J, Lidereau R, Ellis P, *et al*: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006, **98**:1183-1192.
12. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS, *et al*: Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007, **13**:3207-3214.
13. Perou CM, Sorlie T, Eisen MB, Rijn van de M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al*: Molecular portraits of human breast tumours. *Nature* 2000, **406**:747-752.
14. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, *et al*: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003, **100**:8418-8423.
15. Sotiriou C, Piccart MJ: Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?. *Nat Rev Cancer* 2007, **7**:545-553.
16. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, *et al*: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006, **98**:262-272.
17. Stadler ZK, Come SE: Review of gene-expression profiling and its clinical use in breast cancer. *Crit Rev Oncol Hematol* 2008, 1-11.
18. Vijver van de MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, *et al*: A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, **347**:1999-2009.
19. van 't Veer LJ, Dai H, Vijver van de MJ, He YD, Hart AA, Mao M, Peterse HL, Kooy van der K, Marton MJ, Witteveen AT, *et al*: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, **415**:530-536.
20. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, *et al*: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005, **365**:671-679.
21. Sparano JA, Paik S: Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol* 2008, **26**:721-728.
22. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ: Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol* 2008, **26**:729-735.
23. Viale G, Giobbie-Hurder A, Regan MM, Coates AS, Mastropasqua MG, Dell'Orto P, Maiorano E, MacGrogan G, Braye SG, Ohlschlegel C, *et al*: Prognostic and predictive value of centrally reviewed Ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from Breast International Group Trial 1-98 comparing adjuvant tamoxifen with letrozole. *J Clin Oncol* 2008, **26**:5569-5575.
24. Coates AS, Keshaviah A, Thurlimann B, Mouridsen H, Mauriac L, Forbes JF, Paridaens R, Castiglione-Gertsch M, Gelber RD, Colleoni M, *et al*: Five years of letrozole compared with tamoxifen as initial adjuvant therapy for postmenopausal women with endocrine-responsive early breast cancer: update of study BIG 1-98. *J Clin Oncol* 2007, **25**:486-492.

25. Thurlimann B, Keshaviah A, Coates AS, Mouridsen H, Mauriac L, Forbes JF, Paridaens R, Castiglione-Gertsch M, Gelber RD, Rabaglio M, *et al*: A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. *N Engl J Med* 2005, **353**:2747-2757.
26. Mouridsen H, Giobbie-Hurder A, Goldhirsch A, Thurlimann B, Paridaens R, Smith I, Mauriac L, Forbes JF, Price KN, Regan MM, *et al*: Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. *N Engl J Med* 2009, **361**:766-776.
27. Oberli A, Popovici V, Delorenzi M, Baltzer A, Antonov J, Matthey S, Aebi S, Altermatt HJ, Jaggi R: Expression profiling with RNA from formalin-fixed, paraffin-embedded material. *BMC Med Genomics* 2008, **1**:1-9.
28. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, *et al*: Meta-analysis of gene-expression profiles in breast cancer: toward a unified understanding of breast cancer sub-typing and prognosis signatures. *Breast Cancer Res* 2008, **10**:R65.
29. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, *et al*: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004, **351**:2817-2826.
30. Viale G, Regan MM, Maiorano E, Mastropasqua MG, Dell'Orto P, Rasmussen BB, Raffoul J, Neven P, Orosz Z, Braye S, *et al*: Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. *J Clin Oncol* 2007, **25**:3846-3852.
31. Rasmussen BB, Regan MM, Lykkesfeldt AE, Dell'Orto P, Del Curto B, Henriksen KL, Mastropasqua MG, Price KN, Mery E, Lacroix-Triki M, *et al*: Adjuvant letrozole versus tamoxifen according to centrally-assessed ERBB2 status for postmenopausal women with endocrine-responsive early breast cancer: supplementary results from the BIG 1-98 randomised trial. *Lancet Oncol* 2008, **9**:23-28.
32. Viale G, Regan MM, Mastropasqua MG, Maffini F, Maiorano E, Colleoni M, Price KN, Golouh R, Perin T, Brown RW, *et al*: Predictive value of tumor Ki-67 expression in two randomized trials of adjuvant chemoendocrine therapy for node-negative breast cancer. *J Natl Cancer Inst* 2008, **100**:207-212.
33. Hayes DF, Thor AD, Dressler LG, Weaver D, Edgerton S, Cowan D, Broadwater G, Goldstein LJ, Martino S, Ingle JN, *et al*: HER2 and response to paclitaxel in node-positive breast cancer. *N Engl J Med* 2007, **357**:1496-1506.
34. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. *Control Clin Trials* 1996, **17**:343-346.
35. Cox DR: Regression models and life-tables. *J R Stat Soc B* 1972, **34**:187-220.
36. Elston CW, Ellis IO: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991, **19**:403-410.
37. Ma XJ, Hilsenbeck SG, Wang W, Ding L, Sgroi DC, Bender RA, Osborne CK, Allred DC, Erlander MG: The HOXB13:IL17BR expression index is a prognostic factor in early-stage breast cancer. *J Clin Oncol* 2006, **24**:4611-4619.
38. Ma XJ, Salunga R, Dahiya S, Wang W, Carney E, Durbecq V, Harris A, Goss P, Sotiriou C, Erlander M, Sgroi D: A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer. *Clin Cancer Res* 2008, **14**:2601-2608.
39. Cronin M, Pho M, Dutta D, Stephans JC, Shak S, Kiefer MC, Esteban JM, Baker JB: Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. *Am J Pathol* 2004, **164**:35-42.
40. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M: Determinants of RNA quality from FFPE samples. *PLoS ONE* 2007, **2**:e1261.
41. Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE, *et al*: Prognostic Utility of the 21-Gene Assay in Hormone Receptor-Positive Operable Breast Cancer Compared With Classical Clinicopathologic Features. *J Clin Oncol* 2008, **26**:4063-4071.
42. Paik S: Methods for gene expression profiling in clinical trials of adjuvant breast cancer therapy. *Clin Cancer Res* 2006, **12**:1019s-1023s.
43. Carlson RW, Allred DC, Anderson BO, Burstein HJ, Carter WB, Edge SB, Erban JK, Farrar WB, Goldstein LJ, Gradishar WJ, *et al*: NCCN Practice Guidelines in Oncology: Breast Cancer., v.1 2009.
44. Desmedt C, Giobbie-Hurder A, Neven P, Paridaens R, Christiaens MR, Smeets A, Lallemand F, Haibe-Kains B, Viale G, Gelber RD, *et al*: The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1-98 trial. *BMC Med Genomics* 2009, **2**:40.
45. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, *et al*: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2006, **24**:3726-3734.
46. Liedtke C, Hatzis C, Symmans WF, Desmedt C, Haibe-Kains B, Valero V, Kuerer H, Hortobagyi GN, Piccart-Gebhart M, Sotiriou C, Pusztai L: Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *J Clin Oncol* 2009, **27**:3185-3191.
47. Schobesberger M, Baltzer A, Oberli A, Kappeler A, Gugger M, Burger H, Jaggi R: Gene expression variation between distinct areas of breast cancer measured from paraffin-embedded tissue cores. *BMC Cancer* 2008, **8**:343.
48. Viale G, Giobbie-Hurder A, BIG 1-98 Collaborative Group and International Breast Cancer Study Group (IBCSG): Value of centrally-assessed Ki-67 labeling index as a marker of prognosis and predictor of response to adjuvant endocrine therapy in the BIG 1-98 trial of postmenopausal women with estrogen receptor-positive breast cancer. *Breast Cancer Res Treat* 2007, **106(Supplement 1)**:S17, Abstract 64.
49. Carlson RW, Allred DC, Anderson BO, Burstein HJ, Carter WB, Edge SB, Erban JK, Farrar WB, Goldstein LJ, Gradishar WJ, *et al*: Breast cancer. Clinical practice guidelines in oncology. *J Natl Compr Canc Netw* 2009, **7**:122-192.
50. de Azambuja E, Cardoso F, de Castro G Jr, Colozza M, Mano MS, Durbecq V, Sotiriou C, Larsimont D, Piccart-Gebhart MJ, Paesmans M: Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. *Br J Cancer* 2007, **96**:1504-1513.
51. Whitfield ML, George LK, Grant GD, Perou CM: Common markers of proliferation. *Nat Rev Cancer* 2006, **6**:99-106.
52. Urruticoechea A, Smith IE, Dowsett M: Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol* 2005, **23**:7212-7220.
53. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, *et al*: Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 2007, **25**:1239-1246.
54. Chang JC, Makris A, Gutierrez MC, Hilsenbeck SG, Hackett JR, Jeong J, Liu ML, Baker J, Clark-Langone K, Baehner FL, *et al*: Gene expression patterns in formalin-fixed, paraffin-embedded core biopsies predict docetaxel chemosensitivity in breast cancer patients. *Breast Cancer Res Treat* 2008, **108**:233-240.
55. Paik S: Molecular assays to predict prognosis of breast cancer. *Clin Adv Hematol Oncol* 2007, **5**:681-682.

Pre-publication history

The pre-publication history for this paper can be accessed here:http://www.biomedcentral.com/1471-2407/10/37/prepub

doi:10.1186/1471-2407-10-37

Cite this article as: Antonov *et al*.: Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue. *BMC Cancer* 2010 **10**:37.

16 Joint analysis of histopathology image features and gene expression in breast cancer

- BMC Bioinformatics, 17:209, 2016
- IF: 2.448
- number of citations: 1
- personal contribution (80%): image analysis method design, data collection and processing, experimental design and implementation, statistical analyses and results interpretation, manuscript writing

Popovici *et al. BMC Bioinformatics* (2016) 17:209
DOI 10.1186/s12859-016-1072-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Joint analysis of histopathology image features and gene expression in breast cancer

Vlad Popovici[1]*, Eva Budinská[1,2], Lenka Čápková[1], Daniel Schwarz[1], Ladislav Dušek[1], Josef Feit[1] and Rolf Jaggi[3]

Abstract

Background: Genomics and proteomics are nowadays the dominant techniques for novel biomarker discovery. However, histopathology images contain a wealth of information related to the tumor histology, morphology and tumor-host interactions that is not accessible through these techniques. Thus, integrating the histopathology images in the biomarker discovery workflow could potentially lead to the identification of new image-based biomarkers and the refinement or even replacement of the existing genomic and proteomic signatures. However, extracting meaningful and robust image features to be mined jointly with genomic (and clinical, etc.) data represents a real challenge due to the complexity of the images.

Results: We developed a framework for integrating the histopathology images in the biomarker discovery workflow based on the bag-of-features approach – a method that has the advantage of being assumption-free and data-driven. The images were reduced to a set of salient patterns and additional measurements of their spatial distribution, with the resulting features being directly used in a standard biomarker discovery application. We demonstrated this framework in a search for prognostic biomarkers in breast cancer which resulted in the identification of several prognostic image features and a promising multimodal (imaging and genomic) prognostic signature. The source code for the image analysis procedures is freely available.

Conclusions: The framework proposed allows for a joint analysis of images and gene expression data. Its application to a set of breast cancer cases resulted in image-based and combined (image and genomic) prognostic scores for relapse-free survival.

Keywords: Histopathology images, Image analysis, Biomarker discovery, Gene expression, Multimodal data mining

Background

The recent technological progress made scanning the whole pathology slides affordable and its integration in the routine pathology workflow feasible. This resulted in a revived interest in developing new computational methods for nuclear morphometry and tissue architecture characterization, as well as for developing new tissue-based biomarkers [1]. In the last decade, genomic and proteomic techniques have been the methods of choice for novel biomarker discovery. When applied to the same sample, the pathology imaging and *omics technologies allow the investigation of the underlying biology from different perspectives, increasing the chances for identifying effective biomarkers. Ideally, these perspectives could be integrated in a common data analytical framework, to enable a joint (or multimodal) data mining and decision [2].

Traditionally, the methods for analyzing pathology images focused on extracting quantitative measures for a set of predefined morphological parameters (e.g. counting, classifying and characterizing the nuclei) and on reproducing the expert's decision in diagnostic applications (for a review see Gurcan et al. [3]). More recently, a number of applications of pathology image analysis combined image-based quantitative features with genomic

*Correspondence: popovici@iba.muni.cz
[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic
Full list of author information is available at the end of the article

information. For example, Yuan et al. [4] showed that nuclear morphometry is an independent prognostic factor that can improve a genomic signature. A similar approach is discussed by Kong et al. [5] in the case of glioblastoma where they show how nuclear and cytoplasmic features can be linked to genomic profiles and survival outcome. More advanced techniques combine several image-derived characteristics, such as co-localization of tumor nuclei and lymphocyte infiltration [6]. In all these cases however, the imaging features were predefined and based on previous known associations between histopathology and diagnostic/prognostic.

Our interest is in developing a more general computational framework that would allow the integration of the standard histopathology images in the biomarker discovery workflow and in which the image features would be learned in a data-driven fashion, enabling a prior-free data mining. The main challenge when analyzing the pathology images stems from their high complexity and size, and seeming incompatibility with *omics data. In the present work we propose to use the *bag-of-features* approach [7] for reducing the dimensionality of the images and extracting salient features. This approach has already been used in histopathology image classification applications [8, 9] and has the main advantage of allowing an unsupervised learning of image representation. The features extracted describe mostly the textural appearance of small neighborhoods and may be combined with other types of features (e.g. nuclear morphometry) in later stages of image analysis, but these approaches will not be discussed here. As an alternative to bag-of-features, one could use deep learning methods for learning image features as proposed by Cireşan et al. [10] or Cruz-Roa et al. [11]. However, these methods require a larger sample size and were applied in a supervised learning context.

We propose a novel representation of histopathology images which extends the standard bag-of-features with a number of derived measurements aimed at capturing more global characteristics of the tissue sample. In addition, we introduce an objective criterion for optimizing the image representation. The new computational framework is demonstrated in a biomarker discovery scenario, where prognostic features (both imaging and gene expression) for relapse-free survival in breast cancer are sought. We see the application of this approach as a succession of two independent steps, not necessarily performed on the same data corpus. In the first step, a histopathology image representation is learned from a collection of images representative for the pathology under investigation. In the second step, the images of interest are recoded based on the constructed representation and the resulting image features are jointly analyzed with the molecular and clinical data.

Methods

Data

The data used in this study is a subset of the data from Moor et al. [12], selected solely based on the availability of the material for analysis. Overall there were $n = 196$ standard pathology (haematoxylin-eosin-stained) slides with breast tissue sections, not all containing a tumoral component and not necessarily from different cases. All images were obtained by whole-slide scanning of the pathology slides at 40× magnification, resulting in color images of about $150,000 \times 100,000$ pixels.

These data were partitioned into an image model learning set ($n = 131$) and a biomarker discovery/data mining set ($n = 65$). In the biomarker discovery set we kept unique cases for which the slides contained $> 70\,\%$ tumor component and the clinical, survival and gene expression data were all available. The expression profiles of 47 target genes (including 5 control genes) were obtained by quantitative real-time PCR (qRT-PCR). A full description of the data set is available in Moor et al. [12] and the major characteristics of the biomarker discovery set used here are given in Additional file 1.

We computed the genomic prognostic signature (PRO_10) as described in Antonov et al. [13] for all the cases with full genomic profiles.

Image processing

Preprocessing

All images were downscaled to an equivalent of 2.5× magnification by subsampling the Gaussian-filtered higher resolution images (the 4-th level in a Gaussian pyramid). In the resulting images a mask corresponding to the tissue regions was obtained by adaptive thresholding in the green channel. The mask was subsequently refined by morphological operations: erosion with a circular structuring element with radius 13 followed by gap filling and removal of small objects.

For each image we estimated the intensity of haematoxylin (H) staining by deconvolving the RGB-images as described by Ruifrok et al. [14]. The intensity levels of the haematoxylin image (H-image) were adjusted by adaptive histogram equalization. Finally, the background pixels were masked out using the tissue region mask computed as above. In all subsequent image processing steps, only the H-images were used.

Learning the image representation

The bag-of-features [7] approach has two main stages: (i) learning an appropriate *codebook* for representing the images of interest and (ii) re-coding the images based on the frequencies of each *codeblock* (codeword from the codebook). Thus, the resulting representation of the image is a histogram of the codeblocks. For the current application, we extended this representation to include several derived features. We point out that once an appropriate image representation is learned, it can be applied unchanged to other similar image collections thus this step does not need to be repeated on each new data set.

Codebook learning The codebook is a collection of representative local descriptors $\{C_1, \ldots, C_K\}$ obtained as centers of K clusters resulting from k-means clustering of a number of image local descriptors (i.e. a vector quantization procedure). For this, the images are decomposed in a set of local neighborhoods for which descriptor vectors are computed. The local descriptors range from pixels intensities to responses to filter banks or other textural descriptor. For the histopathology images, the Gabor wavelets provide a good set of descriptors, so they were adopted in the present work. Each local neighborhood of size $w \times w$ was convolved with a bank of 24 Gabor filters [15],

$$G(x, y; \nu, \theta, \sigma) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \times \exp\left(2\pi \nu j(x\cos\theta + y\sin\theta)\right)$$

where $j = \sqrt{-1}$, ν was the frequency, θ the orientation and σ the bandwidth of the Gaussian kernel. These parameters were set to $\sigma \in \{1, 2\sqrt{2}\}, \theta \in \{k\frac{\pi}{4} | k = 0, \ldots, 3\}$ and $\nu \in \{3/4, 3/8, 3/16\}$, respectively. They were kept fixed throughout all the experiments. For each filter response, its mean and standard deviations were recorded, thus each local neighborhood $w \times w$ was represented by 48 values (24 means and 24 standard deviations). A comparison of Gabor wavelets with other local descriptors, in the context of histopathology image analysis, is given by Budinská et al. [9].

The size of the codebook (i.e. the number of clusters in k-means clustering), K, is a free parameter that has to be chosen/guessed at the moment of codebook construction [8]. It can also be optimized for the problem at hand [9] using, for example, the Gap statistic [16]. Here we took advantage of having available a number of examples for different tissue components (fat, fat foamy macrophages, comedo necrosis, connective tissue and carcinoma infiltrating fat – for examples see Additional file 1) which we used as reference categories. The goal was to choose the size of the dictionary K in such a way that the representations of these categories are sparse and have a minimal overlap. For each image i, let $y_i = \{j \mid \text{if codeblock } C_j \text{ is used in coding the sample } i\}$, be the set of codeblocks used in its coding. Then we define the following quantities (where $|\cdot|$ denotes the cardinality of a set):

- total Jaccard index,

$$J(K) = 0.5 \sum \frac{|y_i \cap y_j|}{|y_i \cup y_j|},$$

where the sum is taken over all pairs (i, j) of images from different reference categories;

- total sum of within-cluster distances,

$$D(K) = \sum_{k=1}^{K} \sum_{i \in \text{cluster } k} \|\mathbf{x}_i - C_k\|^2,$$

where $\mathbf{x}_i$ are the descriptor vectors.

With these quantities, we defined an (empirical) objective function:

$$\Psi(K) = \log \frac{n_c(n_c - 1)}{2} - \log J(K) - \log \sqrt{D(K)} - 0.75 \log K,$$

where n_c is the number of reference categories (in our case $n_c = 5$). The overall goal of our image recoding step is to find a low dimensional (sparse) representation which still bears enough information for discriminating major tissue components. For this, we minimize $J(K)$, i.e. the overlap between the representations of the reference categories. At the same time, we require tight clusters (small within-cluster total distances $D(K)$) and sparse representation (small K). Hence, the desired value for K is the one that maximizes $\Psi(K)$, where we note that the first term is constant (included to bring the values closer to 0) and that the scaling factor 0.75 is used to reduce the influence of K.

Image recoding Once a suitable K is found and a codebook is constructed by k-means clustering, the standard bag-of-feature approach represents the images as codeblock histograms. However, in this coding, all spatial information about the distribution of the codeblocks is lost. Consider the situation in Fig. 1a: all four images have the same number of patches assigned to the same codeblock, but the spatial arrangement is very different. In order to characterize these spatial differences, we extend the image representation with a number of statistics on the distribution of the codeblocks. For a given image and for each codeblock $k \in \{1, \ldots, K\}$, we construct a binary image in which 1s represent regions assigned to the codeblock and 0s everything else. In these binary images, the connected components (4-neighbor connectivity) define individual objects and for each of them we compute the area (in pixels) and the compactness index (ratio of the squared perimeter to the area of the object). Finally, for each image and each codeblock, we compute (i) the median area, (ii) the maximum area, (iii) the ratio of the maximum area to the total area of the objects, (iv) the skewness of the distribution of the area values and (v) the mean compactness. Thus, for each codeblock in an image, aside from its frequency, we add five new values aimed at characterizing the distribution of the codeblock in the image. We will refer to these additional quantities as the "extended set of features". The final representation of an

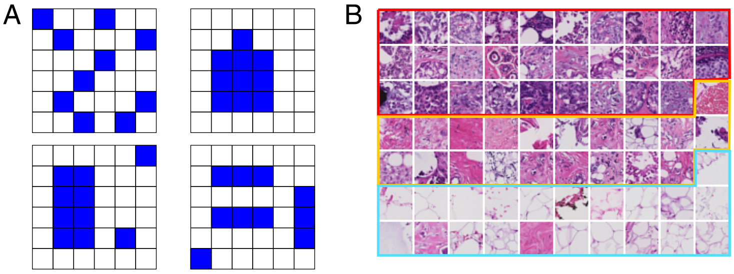


Fig. 1 Codeblocks and codebook. **a** An example of four different hypothetical distributions of the codeblocks leading to identical frequencies. To cope with such situations, the distribution of codeblocks is also taken into account through extended image features. **b** A visual representation of the obtained codebook. The 70 image patches are the closest to the codeblocks obtained after *k*-means clustering. The three groups of codeblocks (with 29, 20 and 21 elements, respectively) correspond to the major clusters in Fig. 2 and the ordering of the image patches is the same as in the clustering

image has a length of $6K$: K values for the codeblock histogram (the standard representation) and $5K$ values of the extended representation.

Joint data mining

The new representation of the images allows for direct application of standard data mining techniques. In the case of multi-modality data mining, the choice of a proper similarity metric/measure is of crucial importance. Two main strategies may be attempted for defining a proper similarity: combination of single, modality-specific, metrics or building/learning a fully multi-modality metric. The first approach has the advantage of using established metrics usually resulting in easily interpretable models and facilitating the comparison with known results. The second approach promises to build a similarity metric that better exploits the multi-modality nature of the data. These ideas can be implemented, for example, in the context of kernel machines (such as Support Vector Machines) where composite kernels (based on closure properties – see [17] p.75) would represent a possible implementation of the first approach and multiple kernel learning [18] an implementation of the latter.

In the present work and in order to demonstrate the general analytical framework, we make use of standard statistical tools. We aim at identifying image features that could be linked to expression levels of the genes of interest (genotype-phenotype association) and potential image biomarkers that alone or in combination with gene expression can be used for defining a prognostic signature. Besides the gene expression, we also used a proliferation gene signature PRO_10 [12, 13], which was shown to be prognostic in various cohorts of patients with breast cancer.

To test the association between image features and tumor size (T) and grade (G) we dichotomized the clinical variables (T: {T1, T2} vs {T3, T4}, and G: {G1,G2} vs. G3, respectively) and used two-sided t-test, with 0.05 significance level. The association of image features with gene expression was assessed based on correlation test (Pearson) with significance level 0.05 and the condition that the correlation coefficient was at least 0.5 (in absolute value). We also used canonical correlation analysis (CCA) to study the associations between image features and molecular data with significance level of 0.05 for Wilks' test. The association between image features and survival outcome (relapse-free survival – RFS) was tested using Cox proportional hazard models (log-likelihood test), with significance level of 0.05. The hazard ratios were estimated from interquartile range-standardized variables (both image and genomic variables). To test if an image feature improves the prognostic value of the gene signature, we tested the difference between the models with and without the variable of interest using likelihood ratio tests. To assess the difference in survival between two groups we used log-rank tests. We binarized the variables by their median value, into high- and low- expressions or values. Since the work reported here is purely exploratory and the sample size is rather small, no adjustment for multiple hypotheses testing was performed. We used hierarchical clustering (Ward method) with Euclidean distance between samples to cluster the codeblocks.

All statistical analyses were performed in `R` package for statistical computing (http://www.r-project.org) version 3.2.2.

Results

Codebook

The image analysis methods described above were implemented in a `Python` package (available at https://github.com/vladpopovici/WSItk), using the `scikit-image` [19] and `Mahotas` [20] libraries.

For the codebook construction we used only the modeling set of images, none of the image used in the data mining phase being used for learning the codebook. From each image, a set of 3000 random patches of size 32 × 32 was extracted and the corresponding Gabor descriptors computed (vectors of 48 elements). These descriptor vectors were clustered using the k-means algorithm to build the codebooks. We estimated the optimal (in the sense of the Ψ objective function, described above) codebook size by evaluating $\Psi(k)$ for $k = 10, 20, \ldots, 1000$. The optimal value was found to be $K = 70$ (see Additional file 1 for a plot of $\Psi(k)$) leading to 420 feature vectors for each image. Since the codeblocks are centers of the clusters (the means of descriptor vectors assigned to the respective cluster), they might not necessarily correspond to observed image regions. Thus we selected the closest regions to the codeblocks (the corresponding descriptor vectors were the closest to the codeblocks) to provide an approximate visual representation of the codebook - Fig. 1b. In the following, to designate a specific codeblock from the codebook, we will use the notation *C.xy*. We have extensively investigated the stability of the learned codebooks and the resulting image representations and we found the process to be stable – see Additional file 1.

The hierarchical clustering of the codeblocks (Fig. 2) revealed a rather structured content: three major groups of codeblocks could be identified. We tentatively labeled them as “proliferation patterns”, “invasion/differentiation patterns/connective tissue” and “sparse tumor nuclei/differentiation/fat” to indicate the major components in the clusters - without claiming a precise histopathological characterization.

A number of codeblocks were found to be associated with tumor size (C.10, C.18, C.29, C.38, C.41, and C.42) and grade (C.09, C.34, C.43, C.45, C.48, and C.62).

Correlations between image features and gene expression

The association analysis between image features and gene expression identified a number of significant ($p < 0.05$ and $\rho > 0.5$) pairwise correlations (all in the range $0.50 - 0.60$). In all, eight different codeblocks were associated with different genes, most of them with *CCNE1* and *CCNB2*. The codeblock C.31 was associated with most genes (*CCNE1, CCNB2, BIRC5, PRC1, SPAG5*) either by its frequency of appearance in the image or by the skewness of its distribution. By summing the frequencies corresponding to image features that are highly correlated (e.g. C.38, C.31, C.01, C.51, C.41, C.68) the correlations coefficients were improved to $0.65 - 0.70$. CCA confirmed the association between these image features and gene expression data (Wilks’ test $p = 0.026$). The image features C.10, C.19, C.57, and C.68 and the genes *CCNE1*, *CCNB2*, and *SPAG5* had the strongest impact on the canonical dimensions. These were also the most stable image features-gene expression correlations in the image representation stability experiments – see Additional file 1.

Despite the fact that the PRO_10 gene signature is an average of proliferation genes which were found to be

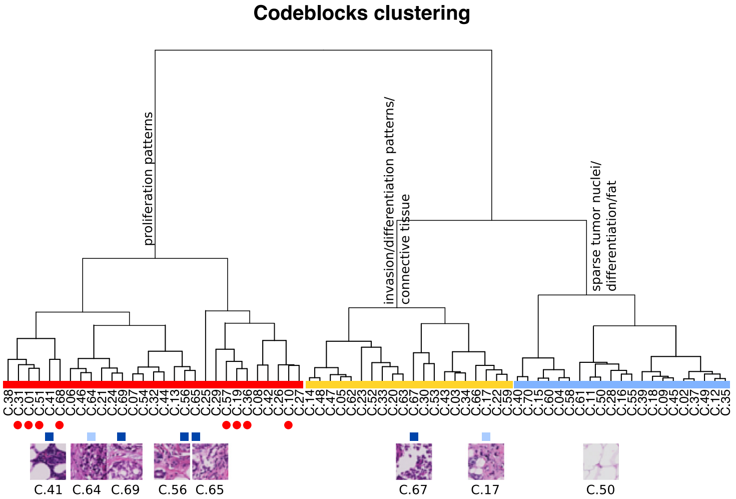


Fig. 2 Hierarchical clustering of the codebook. Clustering the codeblocks led to identification of three major clusters, to which generic terms have been assigned. The codeblocks correlated with gene expression are marked with *red dots*. The codeblocks with potential prognostic value (in univariate analysis) are marked with blue squares (*dark blue* for p-value < 0.01, *light blue* for $0.01 \leq p$-value ≤ 0.05

correlated with image features, the correlations between image features and PRO_10 did not reach the required significance level in all but one case: the skewness of codeblock C.31.

Survival analyses

The goal of the analyses performed was to assess the utility of image-based variables for predicting relapse-free survival independently, or combined with the PRO_10 signature. In the set of samples analyzed, the genomic score is a strong prognostic marker (Cox regression: $p = 0.001, \text{HR} = 2.12, 95\ \%\ \text{CI} = (1.29, 3.51)$).

Univariate Cox proportional hazards models were fit for each of the 420 image features resulting in the identification of several significant associations with relapse-free survival endpoint. The most prognostic image features were C.41, C.56, C.65, C.67, C.69, with $p < 0.01$ and HR between 1.16 and 1.70. From the extended set of features, the median area of the regions assigned to clusters C.15 and C.26 were significantly associated with RFS ($p <$ 0.05). The strongest predictor among the image features was C.69 ($p = 0.0018, \text{HR} = 1.7, 95\ \%\ \text{CI} = (1.22, 2.37)$).

In combined models (image feature and genomic score) a number of image features led to improved models (likelihood ratio test $p < 0.05$), most of them from the extended set of features. From all these image features, C.69 remained significant in the multivariate model (with PRO_10) and had no significant interaction with the genomic signature.

We defined an image score variable by averaging C.41, C.56, C.65, C.67, C.69 which resulted in a stronger prognostic factor (Cox regression: $p = 0.0003$ and HR $=$ $1.76, 95\ \%\ \text{CI} = (1.30, 2.40)$ - see also Figure 3). In a regression model including both the genomic and the image scores, both remained independent significant variables (PRO_10: $p = 0.05$, image score: $p = 0.007$, no significant interaction) and the model was signficantly better than the corresponding univariate models ($p = 0.013$). In Fig. 4 the Kaplan-Meier curves for binarized (by median value) scores are shown, together with corresponding p-values (log-rank tests) and hazard ratios. Another visualization of the prognostic scores is given in Fig. 5 where the expected survival at 4 years is shown as a function of the genomic, image-based, and combined scores, respectively. Two examples of high risk cases, according to the image-based score, are given in Additional files 2 and 3.

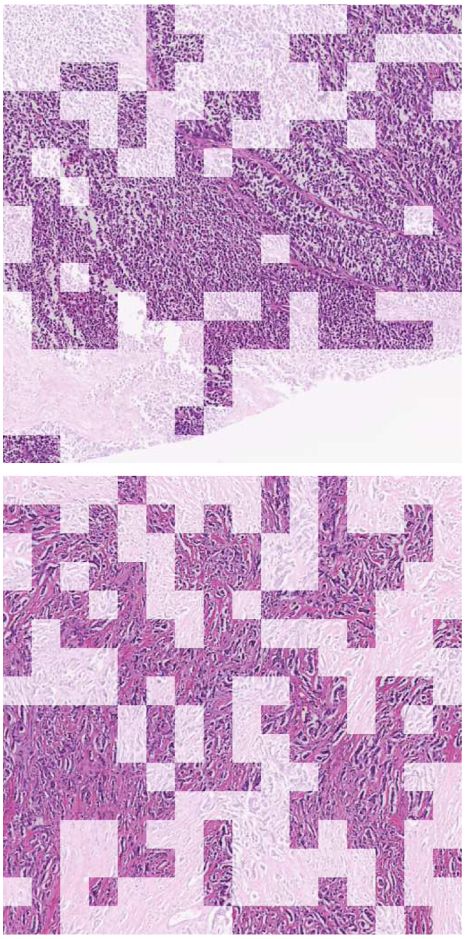


Fig. 3 Regions assigned to the most prognostic codeblocks. 512 × 512 regions from two different samples with high image score (high risk of relapse), at 2.5× magnification. The image patches represented in full color were assigned to one of the C.41, C.56, C.65, C.67 or C.69 codeblocks. In Additional files 2 and 3, the corresponding whole slide images are provided

Discussion

The main challenge in introducing the histopathology images in the general data mining biomarker discovery framework stems from their high complexity and low level of information representation. Thus, while the images contain a huge amount of data (in the order of 10^{10} pixels) the extraction of information implies a considerable effort. Traditionally, this effort is performed by the expert pathologists or, more recently, by using quantitative methods for measuring a set of predefined morphological aspects to complement the pathology report. In this work, we took a third approach, in which the image data is reduced to a number of essential patterns (the codeblocks) whose frequency and spatial distribution in the image is used for data mining. The codeblocks are learned independent of any prior knowledge about the images, potentially

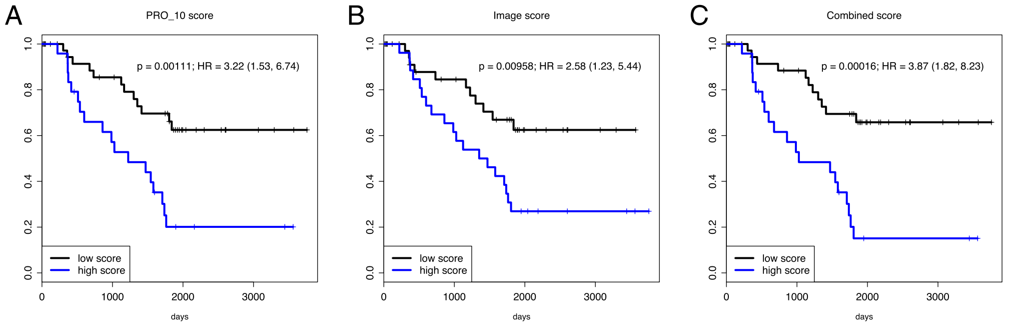


Fig. 4 Kaplan-Meier curves for binarized scores. The genomic (**a**), image-based (**b**) and combined scores (**c**) were binarized by the respective median values into "low score" (low risk) and "high score" (high risk) categories. The combined score slightly improves on the genomic score

enabling the discovery of new image features not necessarily assessed during the pathology review of the cases. The obvious drawback is the difficulty of interpreting some of the patterns and the possibility of having also artifacts in the model. The adopted representation of local neighborhoods in the image (responses to a bank of Gabor filters) encouraged the identification of codeblocks with distinctive textural appearance (Fig. 1). This local appearance may be later on combined with a nuclei detector and classifier (as in Yuan et al. [4]), for example, to obtain a more comprehensive characterization of the image.

By examining the similarities between codeblocks, we identified three major aspects of the images that are captured: proliferation, invasion/differentiation (within connective tissue) and isolated tumor nuclei (within regions predominantly with fat component) (Fig. 2). This result combined with the observation that the whole third cluster did not contribute to the prognostic models, suggests a possible refinement of the current method, in which these regions with high fat content are discarded in an initial preprocessing stage and a more detailed model is used to characterize the remaining regions.

We demonstrated the integration of the image features in a standard biomarker discovery scenario, in which both image-genes correlations (precursors to genotype-phenotype associations) as well as various survival prognostic models were tested. Since the main purpose of this exercise was to demonstrate the integration of image features with genomic information and the sample size was relatively modest, we did not adjust for multiple hypotheses testing and restricted ourselves to an exploratory analysis. Thus the associations found, while hypothesis-generating, have to be taken with caution and more validation is needed.

Most of the genes in the panel were related to proliferation processes, thus it is not surprising that the correlations with image features involved almost exclusively these genes. The strongest associations were found

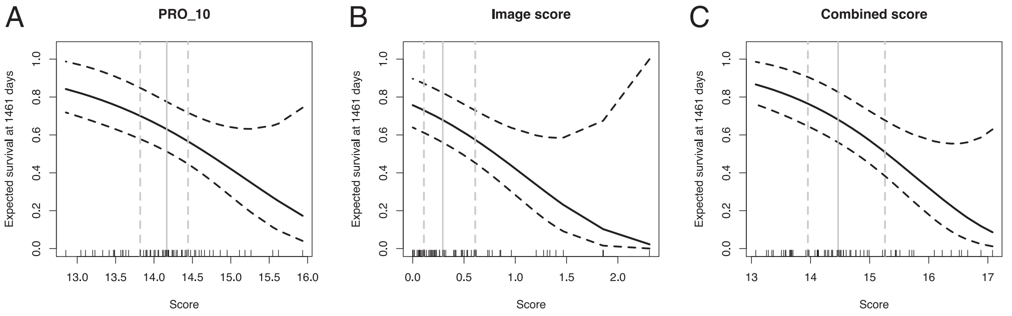


Fig. 5 Prognostic scores at 4 years. Predicting the likelihood of an event (relapse) at 4 years, based on genomic signature (PRO_10 - panel **a**), the image-based score (panel **b**) and the combined score (panel **c**)

with *CCNE1* and *CCNB2*. Somehow surprising, no significant correlation was found with *MKI67* gene, a common marker (with Ki-67 specific staining) for proliferation.

A number of image features were found to be prognostic for RFS and we proposed a simple image-based prognostic score which averages five basic image features. The new score is strongly prognostic and is not correlated with the genomic score considered (PRO_10). When combining the two scores in a multivariable Cox regression, the two remained significant (with a marginal significance for the genomic score) and independent predictors (no significant interaction) leading to an improved model. Thus, the image-based score can be used either alone - as a first line predictor - or in combination with the genomic predictor. These results also demonstrate the complementarity of the two modalities - histopathology imaging and genomics - and suggest that refined predictors can be built by a combination thereof.

It must be noted that the sample size and the number of events did not allow for more variables in the regression models. Further analysis of the scores (either image-based or combined) in the context of usual clinical predictors (TNM-staging, hormonal status, etc.) is required before a definite conclusion about its clinical utility can be drawn. Nevertheless, the image-based score can already be used in applications like searching or indexing in histopathology image archives.

Conclusions

We proposed a general framework for integrating the histopathology images in the routine genomic data analysis pipeline. The image features used are based on the responses of Gabor filters applied to single channel images. The approach can easily be extended to exploit the full color information and to include other types of features.

When applying our method to a data collection of breast cancer samples, we were able to identify a number of associations between image features and gene expression levels. More importantly, several prognostic image features were identified, some of them complementary to the genomic score. Thus, we could build an image-based and a combined survival score, improving on the performance of the genomic score. These results must be validated in larger data sets.

The code implementing the methods described is made freely available and continues to be under active development.

Availability of data and materials

The source code for the image analysis methods described in the paper is available from the `GitHub` repository https://github.com/vladpopovici/WSItk.

The data used to demonstrate the methods described is not publicly available.

Ethics approval and consent to participate

The data used to demonstrate the methods in this study has been graciously provided by the Department of Medical Oncology, Inselspital Bern, Switzerland. All patients gave a general consent for the use of their tissue samples in research.

Additional files

Additional file 1: Codebook construction details [PDF file]. The codebook was optimized based on a objective function and a set of reference categories. This file contains the plot of the objective function and example images for the selected categories. (PDF 12390 kb)

Additional file 2: High risk carcinoma according to image-based score (Example 1). [JPG file]. Whole-slide image of a tumor labeled as high risk by the image score, with the regions used in scoring highlighted. (JPG 9758 kb)

Additional file 3: High risk carcinoma according to image-based score (Example 2). [JPG file]. Whole-slide image of a tumor labeled as high risk by the image score, with the regions used in scoring highlighted. (JPG 12800 kb)

Abbreviations

HR: hazard ratio; qRT-PCR: quantitative real-time polymerase chain reaction; RFS: relapse-free survival.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VP initiated and designed the study, performed image and statistical data analyses. EB and LB performed bioinformatics and biostatistics analyses. JF performed image analyses and contributed pathology interpretation of the images. RJ performed the gene expression profiling, provided clinical and biological interpretation of the results. DS and LD performed image and statistical analyses. All authors contributed to the manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the support of Ministry of Health of the Czech Republic through grant NT/14134. The results presented here reflect only the authors' views.

Funding

This study was funded by the Ministry of Health of the Czech Republic through grant NT/14134.

Author details

[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic. [2] RECETOX, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic. [3] Department of Clinical Research, Faculty of Medicine, University of Bern, Bern, Switzerland.

Received: 9 December 2015 Accepted: 4 May 2016
Published online: 11 May 2016

References

1. Hamilton PW, Bankhead P, Wang Y, Hutchinson R, Kieran D, McArt DG, James J, Salto-Tellez M. Digital pathology and image analysis in tissue biomarker research. Methods. 2014;70(1):59–73.
2. Colen R, Foster I, Gatenby R, Giger ME, Gillies R, Gutman D, Heller M, Jain R, Madabhushi A, Madhavan S, Napel S, Rao A, Saltz J, Tatum J, Verhaak R, Whitman G. NCI Workshop Report: Clinical and Computational

Requirements for Correlating Imaging Phenotypes with Genomics Signatures. Transl Oncol. 2014;7(5):556–69.

3. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng. 2009;2: 147–71.
4. Yuan Y, Failmezger H, Rueda OM, Ali HR, Graf S, Chin SF, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, Johnson N, Doyle S, Turashvili G, Provenzano E, Aparicio S, Caldas C, Markowetz F. Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. Sci Transl Med. 2012;4(157):143.
5. Kong J, Cooper LAD, Wang F, Gutman DA, Gao J, Chisolm C, Sharma A, Pan T, Van Meir EG, Kurc TM, Moreno CS, Saltz JH, Brat DJ. Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. IEEE Trans Biomed Eng. 2011;58(12):3469–74.
6. Nawaz S, Heindl A, Koelble K, Yuan Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. Mod Pathol. 2015;28(6):766–77.
7. Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. Work Stat Learn Comput Vision ECCV. 200459–74.
8. Caicedo JC, Cruz A, Gonzalez FA. Histopathology Image Classification Using Bag of Features and Kernel Functions In: Combi C, Shahar Y, Abu-Hanna A, editors. 12th Conference on Artificial Intelligence in Medicine. Berlin Heidelberg: Springer; 2009. p. 126–35.
9. Budinská E, Čápková L, Schwarz D, Dušek L, Jaggi R, Feit J, Popovici V. Gene expression-guided selection of histopathology image features. In: 15th International Conference on Bioinformatics and Bioengineering. Belgrade: IEEE; 2015. p. 1–6.
10. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin Heidelberg: Springer; 2013. p. 411–8.
11. Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks In: Gurcan MN, Madabhushi A, editors. SPIE Medical Imaging. San Diego, USA: SPIE; 2014. p. 904103.
12. Moor AE, Guevara C, Altermatt HJ, Warth R, Jaggi R, Aebi S. PRO_10 – A new tissue-based prognostic multigene marker in patients with early estrogen receptor-positive breast cancer. Pathobiology. 2011;78(3):140–8.
13. Antonov J, Popovici V, Delorenzi M, Wirapati P, Baltzer A, Oberli A, Thurlimann B, Giobbie-Hurder A, Viale G, Altermatt H, Aebi S, Jaggi R. Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue. BMC Cancer. 2010;10(1):37.
14. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001;23(4):291–9.
15. Daugman JG. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A. 1985.
16. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001.
17. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press; 2004.
18. McFee B, Lanckriet GRG. Learning Multi-modal Similarity. J Mach Learn Res. 2011;12:491–523.
19. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, Scikit-image contributors. scikit-image: image processing in Python. PeerJ. 2014;2:e453.
20. Coelho LP. Mahotas: Open source software for scriptable computer vision. J Open Res Softw. 2013;1(1):e3.

17 Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer

- Journal of Pathology, 231(1):63-76, 2013
- IF: 6.894
- number of citations: 121
- personal contribution (20%): statistical analyses, results interpretation, manuscript writing

Journal of Pathology
J Pathol 2013; **231:** 63–76
Published online 8 July 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.4212

ORIGINAL PAPER

Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer

Eva Budinska,[1,2]* Vlad Popovici,[1,2] Sabine Tejpar,[3] Giovanni D'Ario,[1] Nicolas Lapique,[1] Katarzyna Otylia Sikora,[1] Antonio Fabio Di Narzo,[1] Pu Yan,[4] John Graeme Hodgson,[5] Scott Weinrich,[5] Fred Bosman,[5] Arnaud Roth[6,7] and Mauro Delorenzi[1,8]

[1] *Bioinformatics Core Facility, Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland*
[2] *Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic*
[3] *Department of Oncology, University Hospital Gasthuisberg, Katholik Universiteit Leuven, Belgium*
[4] *University Institute of Pathology, Lausanne University Medical Centre, Switzerland*
[5] *Pfizer Inc., Worldwide Research and Development, Oncology Research Unit, La Jolla, CA, USA*
[6] *Oncosurgery, Geneva University Hospital, Switzerland*
[7] *Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland*
[8] *Département de Formation et Recherche, Lausanne University Medical Centre, Switzerland*

**Correspondence to: Eva Budinska, Institute of Biostatistics and Analyses, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic e-mail: budinska@iba.muni.cz*

Abstract

The recognition that colorectal cancer (CRC) is a heterogeneous disease in terms of clinical behaviour and response to therapy translates into an urgent need for robust molecular disease subclassifiers that can explain this heterogeneity beyond current parameters (MSI, *KRAS*, *BRAF*). Attempts to fill this gap are emerging. The Cancer Genome Atlas (TGCA) reported two main CRC groups, based on the incidence and spectrum of mutated genes, and another paper reported an EMT expression signature defined subgroup. We performed a prior free analysis of CRC heterogeneity on 1113 CRC gene expression profiles and confronted our findings to established molecular determinants and clinical, histopathological and survival data. Unsupervised clustering based on gene modules allowed us to distinguish at least five different gene expression CRC subtypes, which we call surface crypt-like, lower crypt-like, CIMP-H-like, mesenchymal and mixed. A gene set enrichment analysis combined with literature search of gene module members identified distinct biological motifs in different subtypes. The subtypes, which were not derived based on outcome, nonetheless showed differences in prognosis. Known gene copy number variations and mutations in key cancer-associated genes differed between subtypes, but the subtypes provided molecular information beyond that contained in these variables. Morphological features significantly differed between subtypes. The objective existence of the subtypes and their clinical and molecular characteristics were validated in an independent set of 720 CRC expression profiles. Our subtypes provide a novel perspective on the heterogeneity of CRC. The proposed subtypes should be further explored retrospectively on existing clinical trial datasets and, when sufficiently robust, be prospectively assessed for clinical relevance in terms of prognosis and treatment response predictive capacity. Original microarray data were uploaded to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under Accession Nos E-MTAB-990 and E-MTAB-1026.

Keywords: colorectal cancer; histopathology; gene expression; molecular heterogeneity

Received 3 February 2013; Revised 10 May 2013; Accepted 14 May 2013

Conflict of interest: ST and FTB have received research funds from Pfizer; and AR is a compensated consultant for Pfizer.

Introduction

Current classifications of sporadic colorectal cancer take into consideration stage, histological type and grade [1]. Colorectal cancer (CRC) is a highly heterogeneous disease, with clinicopathologically similar tumours differing strikingly in treatment response and patient survival. These differences are only partly explained by current concepts regarding the molecular events leading to CRC. In recent years, microsatellite instability (MSI) emerged as an important classifier with significant prognostic impact and potential for patient stratification for therapy [2,3]. Some molecular markers, as well as the mutation status of *BRAF* or *KRAS* genes (predictive for anti-EGFR [4]), are in use for treatment decisions and patient stratification. However, patient groups defined by these molecular markers still differ remarkably in behaviour and therapy response [5,6]. Several approaches to further subtype CRC have been proposed, based on combinations of clinical, histopathological, gene expression, CNV, epigenetic and single gene parameters [7–13]. Each of these different modalities provides its own perspective on the same underlying biological reality. The CpG island methylator phenotype (CIMP) status is emerging as important molecular determinant of CRC heterogeneity [11]. The cancer genome atlas (TCGA) analysis identified a hypermutant group not entirely captured by MSI status [13]. Several studies have addressed CRC subtyping using genome-wide gene expression profiling of relatively large patient cohorts [12,14]. One study used unsupervised clustering of stage II and III CRCs to identify three stage-independent subtypes, with *BRAF* mutation and MSI status dominating one of the subtypes [14]. A study of stage I–IV CRC samples segregated CRC into two prognostic subtypes with epithelial–mesenchymal transition (EMT) as a main determinant [12]. Another study on 88 stage I–IV samples identified four subtypes, one correlated with MSI, *BRAF* mutation and mucinous histology, two with stromal component and one with high nuclear β-catenin expression [15].

We recently reported CRC expressing a *BRAF*-mutated signature [6], which strongly overlaps with the methylation-based group of Hinoue [11], and a MSI-like gene expression group that captures the hypermutant tumours of TCGA [13], indicating the potential for identification of robust biological subgroups. We now describe CRC subtypes based upon unsupervised clustering of genome-wide expression patterns. We characterized these subtypes in terms of biological motifs, common clinical variables, association with known CRC molecular markers and morphological patterns. A key element in our approach was the use of a system of unsupervised gene modules—groups of genes with correlated expression. They are more resistant to noise and have a higher chance of having at least a few members represented on various platforms. In addition, as each gene module is represented by its median expression, the modules with fewer genes contribute equally to the subtype definition. We and others have successfully used similar strategies previously [16–18]. We validated the existence of the subtypes and their respective clinical and molecular marker characteristics in an independent dataset. Ultimately, it will be mandatory to integrate the various sources of information on CRC heterogeneity into an integrative, robust and reproducible subclassifier that can become a tool for clinical use.

Materials and methods

A detailed description of all the datasets and analysis procedures is given in Supplementary methods and results (see Supplementary material).

Data acquisition and processing

We have built two non-overlapping data collections: a discovery collection, comprising four publicly available (425 samples) and two previously unpublished datasets (688 samples with 10 year follow-up in a clinical trial setting and 64 normal samples) with known stage status, and a validation collection of eight publicly available datasets (720 CRC samples) (see Supplementary material, Supplementary methods and results). Observations derived from the analysis of 64 normal samples were further validated on five publicly available datasets, with both carcinoma and normal samples available in one batch (totalling 205 normal/adenoma/carcinoma samples). Copy number data was available for 154 of the PETACC3, as in [19]. Our analysis included a total of 2102 samples.

The discovery collection contained the previously unpublished 688 CRC formalin-fixed, paraffin-embedded (FFPE) samples of PETACC3 [6] and 64 FFPE normal colon tissue samples from Centre Hospitalier Universitaire Vaudois's Biobank, which were uploaded to ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), under Accession Nos E-MTAB-990 and E-MTAB-1026, respectively. Gene expression data were processed by standard tools to obtain normalized, probeset-level expression data. For each EntrezID in the datasets, the probeset with the highest variability was selected as representative and the number of EntrezIDs entering the analysis was reduced to 3025 by applying non-specific filtering. For PETACC3 and normal colon samples, patients signed an informed consent form in which the use of tissue specimens was included, and all marker study proposals were subjected to the approval of the trial steering committee.

Subtype definition and validation

For model development (gene modules and subtype definition, classifier training, identification of subtype-specific genes) only the 1113 CRC samples of the discovery set were used, no sample in the validation collection being used for any model tuning. Hierarchical clustering (complete linkage, Pearson correlation similarity measure) and dynamic cut tree [20] were used to produce *gene modules* (groups of genes with correlated expression), from which non-robust modules (see Supplementary material, Supplementary methods and results) and a gender-related module were discarded. Each expression profile was then reduced to a vector of *meta-gene*s by taking the median of the values of genes in each gene module. The meta-genes were then further grouped into clusters using hierarchical clustering.

The subtypes were defined in terms of *core samples*—those samples from the discovery collection that were assigned to clusters by hierarchical clustering, using a consensus distance [21] followed by pruning of the dendrogram (see Supplementary material, Supplementary methods and results). The clusters to which the core samples were assigned were called *subtypes*. The rest of the samples from the discovery collection, not assigned to subtypes by this procedure, were called *non-core samples*. This approach allowed the reduction of noise in subtype-defining samples, and thus a higher consistency of the resulting subtypes defining the ground truth for downstream analyses. The stability of the obtained clusters was assessed under different perturbations of the processing pipeline (different parameters and clustering methods) to ensure that the results were not simple artefacts (see Supplementary material, Supplementary methods and results). A multiclass linear discriminant (LDA) [22] was trained on core samples with meta-genes as variables to assign new samples to one of the subtypes. Minimal gene sets characteristic to each subtype were identified using ElasticNet [23] on gene-level data.

In order to validate the existence of subtypes (and their independence on data selection) and the modelling choices in subtype discovery, we applied the same subtyping procedure (including parameters) to the validation collection. The clusters identified in the validation collection were put in correspondence with the subtypes in the training set by LDA predictions and correlations of subtype-specific moderated *t* statistic [24] values, corresponding to the gene-wise comparison of the respective subtype with the other subtypes (one-versus-all comparison). A simple classifier application would have led the validation samples to be classified as one of the subtypes, but it would have not informed us of possible over-fitting of the data in the discovery procedure.

Subtype characterization

If not specified differently, all the reported *p* values were adjusted for multiple hypothesis testing, using the Benjamini–Hochberg procedure. Significance level was set at 0.1. Pathway analysis for each set of gene modules was carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [25]. Gene set enrichment analysis of gene signatures was performed using the mygsea2 tool, in each subtype and normal samples, on average expression-ordered median-centred lists of genes. Differential expression analysis was performed using limma [24] and sign test using BSDA [26]. The Cox proportional hazards model was used to analyse the prognostic value of interquartile range (IQR)-standardized values of meta-genes, for overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR), stratified by dataset. The Wald test was used to assess the global significance of the models. Pairwise differences in survival were assessed using the log-rank test. For subtype comparison, the survival was truncated at 7 years. Subtype enrichment for clinical or molecular markers was assessed by the Fisher test to the baseline, defined as the proportion of the marker in the whole dataset. Morphological pattern differences were assessed pairwise by Fisher test.

Histology

The identified subtypes were characterized histologically in terms of six different architectural patterns: complex tubular; solid/trabecular; mucinous; papillary; desmoplastic; and serrated (Figure 4A), which were called dominant or secondary depending on their presence in the histology slides (for details on immunohistochemistry, see Supplementary material, Supplementary methods and results).

Results

Gene modules and subtype definition

We identified 54 gene modules, reproducible across all datasets in the discovery collection, comprising 658 genes from an initial list of 3025 identified as the most variable. The assignment of genes to gene modules and gene module clusters is listed in Table S1 (see Supplementary material); meta-gene expression profiles for the discovery set are shown in Figure 1A; and between meta-gene correlations in Figure S1C (see Supplementary material). Based on gene modules, we identified five major subtypes: surface crypt-like (A), lower crypt-like (B), CIMP-H-like (C), mesenchymal (D) and mixed (E), totalling 765 samples (69% of discovery data; see Supplementary material, Supplementary methods and results).

Subtype reproducibility in an independent validation set

In the validation set of 720 CRC samples we identified a set of subtypes comprising 602 samples (83.6% of the validation set) and associated them with our discovery subtypes using the subtype classifier (see Supplementary material, Table S2) and correlations of subtype-specific patterns based on moderated *t* statistic (see Supplementary material, Table S3). All five major subtypes reappeared in the validation set, confirming the robustness of our approach. Figure S2 (see Supplementary material) presents gene expression profiles of both discovery and validation sets. Two notable differences were observed: (i) subtype B in the validation set was split into two subgroups (B1, B2), as observed in the discovery set too, but only at lower pruning height; (ii) another cluster passed the minimal size criteria, corresponding to the small subtype (F) which, in the discovery set, was not considered for further characterization because of small sample size. Validation of other subtype characteristics (to the extent of available information) is described in each of the respective sections.

Subtypes are characterized by distinct biological components

We set out to assign biological labels to gene modules that define the subtypes (Table 1; see also Supplementary material, Table S1). Of the 54 meta-genes,

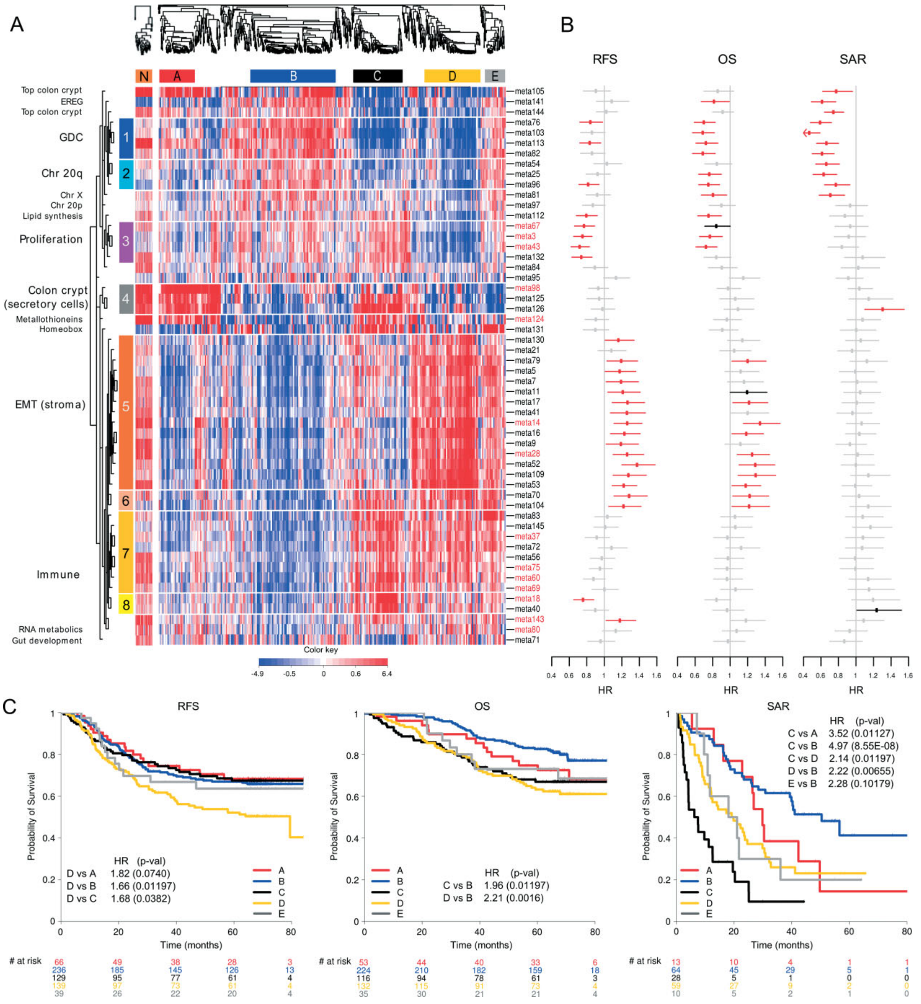


Figure 1. Meta-gene expression pattern in subtypes, connected with prognostic effect of subtypes and meta-genes, in the discovery set. (A) Two heat maps clustering normal (left) and CRC (right) samples (columns) and meta-genes (rows). Colours represent decreased (blue) or increased (red) meta-gene expression relative to their medians. Normal samples were clustered independently on meta-genes centred to CRC meta-gene medians. For comparative purposes, ordering of meta-genes in normal samples is imposed to correspond to that of CRC samples. White horizontal lines denote eight unsupervised clusters of meta-genes, each assigned a colour bar on the left; meta-genes not belonging to a cluster have no colour bar. Names of the meta-genes corresponding to gene modules with gene–gene correlations in normal samples comparable to those in cancer samples are marked red (see Supplementary material, Figure S1D). (B) Effect of inter-quartile range (IQR) standardized expression of meta-genes on RFS, OS and SAR. Points represent estimated hazard ratio (HR), bars represent 95% CI. Bold lines represent effects significant at 5% without adjustment for multiple hypothesis testing; red lines represent effects significant at FDR $< 10\%$; details are provided in Table S6 (see Supplementary material). (C) Kaplan–Meier plots for RFS, OS and SAR, with HR for significant pairwise comparisons (p values adjusted for FDR). Numbers below x axes represent number of patients at risk at selected time points.

41 could be further grouped into eight gene module clusters; 13 meta-genes remained ungrouped, each possibly representing a distinct biological motif. Pathway analysis characterized five of eight gene module clusters by the following biological motifs: chromosome 20q (cluster 2), proliferation (cluster 3), EMT/stroma (cluster 5) and immune response (clusters 7 and 8). Literature searching identified biological motifs associated with other gene modules. We labelled cluster 1 as GDC (genes differentially expressed in CRC), as

Table 1. Biological identification of gene modules

Cluster name	Number of genes	Pathway analysis result (number of overlapping genes, p value) OR description based on literature search	Selected genes
1. GDC	27	Genes involved in differentiation of colon crypt and/or whose expression was reported to be affected in colorectal cancer and/or with prognostic effect in CRC	Intestinal differentiation genes: *CDX2*[45], *IHH*[46], *VAV3*[47], *ASCL2*[35], *PLAGL2*[48] Genes reported altered in colorectal cancer with prognostic effect: *PITX2*[49], *DDC*[50], *PRLR*[51], *SPINK1*[52] Other genes connected to CRC: *GGH* – connected to CIMP^{+} phenotype [53] *NR1I2* – connected to chemoresistance [54]
2. Chromosome 20q genes	33	Chromosome 20 (26 genes, 9.2E-34)	Other, non-20q genes: *TP53RK*, *ANO9*, *NEU1*, *CLDN3*, *PRSS8*
3. Proliferation	83	Cell cycle (36 genes, 3.0E-33) Mitosis (26 genes, 1.4E-29) Chromosome (26 genes, 2.5E-17) DNA metabolic process (20 genes, 4.9E-10) Lipid synthesis (4 genes, 5.0E-2)	Mitotic checkpoint kinases: *BUB1*, *BUB1B* Cyclins: *CCNA2*, *CCNB2* Centromere proteins: *CENPA*, *CENPE*, *CENPN* Kinesins: *KIF11*, *KIF23*, *KIF4A* Topoisomerase II (*TOP2A*) Cell division cycle 2 *CDC2*
4. Colon crypt markers (secretory cells)	16		*AGR2*[55], *AGR3*, *MUC2*, *SPINK4*[56], *RETNLB*[57], *REG4*[58]
5. EMT/stroma	310	Extracellular region part (90 genes) 2.7E-36 Cell adhesion (57 genes) 1.2E-17 Extracellular matrix (44 genes) 5.3E-30 Collagen (16 genes) 1.2E-15 EGF-like domain (26 genes) 1.6E-12 Cell motion (33 genes) 7.2E-8 Blood vessel development (25 genes) 1.1E-8 Growth factor binding (6 genes) 6.0E-5 Frizzled related (5 genes) 6.7E-3 Cell junction organization (7 genes) 1.8E-2 WNT receptor signalling pathway (8 genes) 1.4E-1	Inhibitors of β-catenin-dependent canonical WNT: *SFRP1*, *SFRP2*, *SFRP4*, *DKK3*, *FZD1,7*, *PRICKLE1*, *NXN* Mesenchymal markers: N-cadherin, OB cadherin, *SPARC*, *DDR2* EMT inducers(TFs): *SNAI2*, *ZEB1*, *ZEB2*, *TWIST1*, *CDH11* ECM remodelling and invasion: *MMP14*, *VIM* ECM proteins: fibronectin 1, collagens Angiogenesis: *PLAT*, *PLAU*, *NRP1*, *NRP2*, *THBS1*, *THBS2*, *THBS4* TGFs, their receptors and binding proteins: *IGF1*, *IGFBP5*, *IGFBP7*,*TGFB*, *LTBP1*, *LTBP2*, *PDGFRA*, *PDGFRB*
6. Unidentified	14		*DUSP1*, *EGR2*, *SERPINE1*
7 and 8. Immune response	103	Immune response (42 genes) 2.0E-28 Positive regulation of immune system process (16 genes) 4.0E-9 Antigen processing and presentation via MHC class II (6 genes) 7.5E-5 Defence response (31 genes) 3.3E-17 Chemokine signalling pathway (9 genes) 2.2E-3 Lymphocyte activation (11 genes) 2.1E-5 Regulation of programmed cell death (14 genes) 2.1E-2	Cytokines: *CCL3*, *CXCL5*, *CXCL9*,*CXCL10*, *CXCL11*, *SPP1*, *LTB* MHC class II: *HLA-DMB*, *HLA-DPA1*, *HLA-DRA*, *CD74* MHC class I: *HLA-F*, *TAP1*, *TAP2* Anti-apoptotic: *BCL2A1*, *CD74*, *BIRC3*, *IFI6*, *TNFAIP3*, *TNFAIP3* Apoptotic: *STAT1*, *XAF1* Interferon-induced proteins: *IFI30*, *IFI16*, *IFI44*, *IFI16*, *IFIH1*, *IFIT3*
Cluster-unassigned meta-genes with colon crypt cell markers (enterocytes/top of the crypt)			
Meta-gene 105	6	Top of the crypt genes	*FAM55A*, *FAM55D*, *MUC12* and *CEACAM7*[59], *SLC26A2*[59], *SLC26A3*[59]
Meta-gene 144	5	Enterocytes, goblet cells markers	*LOC644844*, *NGEF*, *HEPH*, *KRT20*[59], *MUC20*[59]
Cluster-unassigned meta-genes associated with chromosomal location 0			
Meta-gene 81	7	Chromosome X (7 genes) 1.1E-8	*CXorf15*, *EIF1AX*, *HDHD1A*, *MED14*, *PNPLA4*, *SCML1*, *SMC1A*
Meta-gene 97	6	Chromosome 20p (5 genes) 5.0E-11	*CDC25B*, *CSNK2A1*, *MRPS26*, *PTPRA*, *RP5-1022P6.2*, *SNRPB*
Meta-gene 84	7	Chromosome 8 (7 genes) 5.4E-9	*AGPAT5*, *FDFT1*, *GTF2E2*, *LONRF1*, *MTUS1*, *VPS37A*, *ZNF395*
Other cluster-unassigned meta-genes			
Meta-gene 141	5	EREG	*AK3L1*, *ARID3A*, *EREG*, *LDLRAD3*, *ZBTB10*
Meta-gene 112	6	Lipid synthesis (4 genes) 5.0E-2	*DHCR7*, *FASN*, *FGFBP1*, *HMGCS1*, *IDI1*, *PCSK9*
Meta-gene 95	6	Homeobox genes	*HOXA10*, *HOXA11*, *HOXA13*, *HOXA5*, *HOXA7*, *HOXA9*
Meta-gene 124	5	Metallothioneins	*MT1E*, *MT1F*, *MT1G*, *MT1M*, *MT1X*
Meta-gene 131	5	Disulphide bonds (5 genes) 1.7E-02	*CXCL5*, *IL6*, *MMP1*, *MMP3*, *PTGS2*
Meta-gene 143	5	Unidentified	*DUSP5*, *ERRFI1*, *KLF6*, *MXD1*, *PLAUR*
Meta-gene 80	7	Regulation of RNA metabolic process (6 genes) 4.9E-2	*ATF3*, *C8orf4*, *FOS*, *JUNB*, *NR4A1*, *SIK1*, *ZFP36*
Meta-gene 71	8	Gut development (3 genes) 3.5E-2	*CCL11*, *CH25H*, *EDNRB*, *F2RL2*, *FOXF1*, *FOXF2*, *PCDH18*, *WNT5A*

Table 2. Subtype-specific minimal gene set as identified by Elastic net

	Minimal gene sets specifying a subtype	
Subtype	Up-regulated from population mean	Down-regulated from population mean
A. Surface crypt-like	*ADTRP, B3GNT7, CLCA1, MUC2, NR3C2, PADI2, RETNLB, STYK1*	*CHI3L1, FNDC1, TIMP3, SULF1*
B. Lower crypt-like	*CCDC113, CDHR1, FARP1, GPSM2, GRM8, HNF4A, IHH, KCNK5, KIAA0226L, MYRIP, PLAGL2, PRR15, QPRT, RNF43, RPS6KA3, SLC5A6, TP53RK, TSPAN6, VAV3, YAE1D1*	*ALOX5, BASP1, CREB3L1, CXCR4, EPB41L3, FSCN1, GFPT2, GPX8, ITPRIP, KCNMA1, KCTD12,MT1E, RARRES3, RNASE1, SGK1, SOCS3*
C. CIMP-H-like	*ANP32E, EGLN3, IDO1, PLK2, RAB27B, RARRES3, RPL22L1, TFAP2A*	*ATP9A, C10orf99, CXCL14, KIAA0226L*
D. Mesenchymal	*ANK2, BOC, C7, CRYAB, DCHS1, DDR2, GEM, PRICKLE1, TAGLN*	*HOOK1, RBM47*
E. Mixed	*CEACAM6, CXCL5, HSD11B1, IL1B, IL6, MRPS31, PI15, RAP2A, UQCC*	*AGR3, RAB27B, REG4*

it consisted of a number of genes significantly associated with CRC. The analysis of pairwise intra-gene module correlations in normal samples of both discovery and validation set identified as cancer-specific gene modules of chromosome 20q, several immune response, EMT/stroma and GDC gene modules, homeobox genes and gut development (see Supplementary material, Figure S1D). The relationship between subtypes and meta-genes is illustrated by the heat map (Figure 1A), in which the major molecular motifs and their role in subtype definition stand out. Table S4 (see Supplementary material) contains median subtype values per meta-gene and the results of differential meta-gene expression testing between subtypes. Subtypes are not determined by individual biological components but each of them contributes to the molecular identity of the subtypes. The EMT/stroma cluster stands out in subtypes A + B (low expression) and D + E (high expression), while subtype C notably contained a high expression of immunity-associated cluster. High expression of meta-genes representing upper colon crypt cells in subtypes A and B, correlated with serrated and papillary (A) and complex tubular (B) morphological patterns (see below). Given the enterocyte-like morphology and retained polarity of the neoplastic cells in these patterns, they are considered as well differentiated. Subtype C is associated with the mucinous phenotype. Interestingly, subtypes A and C show high expression of metallothioneins, subtypes C and E show high expression of the homeobox gene module, while subtypes E and B strongly express a gene module containing the *EREG* gene (Table 1). The high expression of chromosome 20q cluster in subtype B was correlated with a significantly higher copy number gain/amplification of all of 20q in this subtype (see Supplementary material, Figure S8). The low expression of lipid synthesis genes is striking for subtype D and low expression of the gut development gene module for subtype C. A refined picture of differences is given by a quantitative comparison of (meta-)gene expression between subtype pairs (see Supplementary material, Tables S4 and S5, Figure S4). For each subtype we also identified a minimum set of characteristic genes (Table 2; for more details, see Supplementary material, Supplementary methods and results).

Normal colon mucosa in the context of subtypes

When applied to the 64 normal samples, the LDA classifier assigned them all to subtype A, with posterior probability > 0.99, supporting the observation that A is well differentiated and closest to normal colonic epithelium in terms of gene expression pattern. For validation, we analysed five public datasets comprising 205 profiles of normal/adenoma/carcinoma samples. Most of the normal and adenoma samples were classified by LDA as subtype A (74.5% of 51 and 69.0% of 71, respectively) or subtype B (28.2% and 21.6%, respectively), confirming subtype A as the most normal-like. The 80 carcinoma samples were distributed over all subtypes (26.2% A, 30.0% B, 11.3% C, 18.7% D and 13.8% E).

Subtypes and patient survival

We assessed whether subtypes differ in survival, as a general read-out of biological significance, and then tested the association of each meta-gene with prognosis, using the complete discovery set of 1113 patients (Figure 1B-C see also Supplementary material, Table S6). Kaplan–Meier curves for RFS, OS, SAR, hazard ratios (HRs) and *p* values of pairwise differences between subtypes are shown in Figure 1C. The results indicate that subtypes C and D are associated with poor OS. For subtype D, this is primarily due to early relapse correlated with high expression of EMT genes and low expression of proliferation-associated genes. For subtype C it is the result of short SAR, correlated with low expression of GDC, top colon crypt, EREG and Chr 20q genes and high expression of meta-gene 126 (see Supplementary material, Table S1). For subtype E the trend towards poorer OS and RFS was not statistically significant, although borderline significant poorer SAR was found relative to subtype B. Subtypes A and B had better prognosis than D for all three endpoints, although for OS in subtype A this was not significant.

The analysis of clinical and molecular markers (below) showed that subtype C is enriched for MSI tumours and *BRAF* mutant tumours, the latter present also in subtype D. The literature indicates that MSI is associated with better RFS, while *BRAF* mutation is an indicator of worse SAR [27]. To analyse how these two contradictory components affect survival in

Table 3. Result of additive multivariate Cox proportional hazards model, with subtype, *BRAF* mutation, MSI and stage[a]

Variable	RFS HR	*p*	OS HR	*p*	SAR HR	*p*
A	0.906	0.760	1.381	0.390	1.726	0.180
C	0.940	0.850	1.560	0.220	3.675	0.0022*
D	1.688	0.0055*	2.161	0.0011*	1.906	0.014*
E	1.506	0.210	2.201	0.035*	2.046	0.075
BRAFm	1.633	0.085	2.472	0.0034*	3.361	0.00072*
MSI	0.478	0.044*	0.275	0.004*	0.356	0.036*
Stage 3	0.770	0.190	0.943	0.820	1.780	0.062*

[a] Baseline is subtype B, MSS, *BRAF* wt and Stage 2.

*Variables significant in the model.

Hazard ratios (HR) for relapse-free survival (RFS), overall survival (OS) and survival after relapse (SAR).

subtypes, we built a multivariate Cox proportional hazard model with subtype, stage, *BRAF* and MSI (Table 3; see also Supplementary material, Table S6). Subtype C remained significantly associated with poor SAR, even after the adjustment for *BRAF*, MSI and stage, but not with RFS. Subtypes B and D remained significantly prognostic for RFS, OS and SAR. No equivalent survival data were available for the datasets in the validation series, hence these observations could not be validated.

Colorectal stem cell and Wnt signatures within subtypes

We investigated the association of subtypes with Wnt [28–32], putative colon cancer stem cell (CSC) [33–35] signatures, and two signatures specific for upper and lower colon crypt compartments [36], using gene set enrichment analysis (Figure 2; see also Supplementary material, Table S7). Subtypes B and E highly expressed canonical Wnt signalling target signatures. Subtypes A and D and also normal samples, however, showed low expression of these signatures. This was in concordance with the differences in β-catenin nuclear immunoreactivity at the invasion front (IF; see Supplementary material, Figure S9 and Supplementary methods and results). Subtypes B and E showed the highest percentages, while subtypes A and D showed significantly lower percentages of the β-catenin-positive nuclei. Subtype C exhibited almost no β-catenin nuclear immunoreactivity at the IF. We analysed CSC signatures derived from low colon crypt compartment cells that had been identified either by a Wnt reporter construct TOP GFP or by high surface expression of *EphB2*. Subtypes D and E expressed both TOP GFP and *EphB2*-derived CSC signatures, while subtype B mainly expressed only the TOP GFP signature (Figure 2).

Subtypes complement clinical and molecular markers

An important goal of this study was to assess how our molecular subtypes complement known clinical variables and molecular markers. We found that MSI, *BRAF* mutation status, site, mucinous histology and expression of p53 were significantly associated with various subtypes (Figure 3), but not tumour stage, age, gender, *SMAD4* or *PIK3CA* mutations (see Supplementary material, Figure S5A). Subtype D was not significantly enriched for any of the tested variables except for the *BRAF* mutated signature and possibly represents a mixture of tumours that have the EMT/stroma signature in common. *KRAS* mutants occurred in all subtypes (see Supplementary material, Figure S5C), supporting the emerging notion that *KRAS*-mutated CRC are substantially heterogeneous [5,6,37], the oncogenic role of *KRAS* varying per specific mutation and the molecular background of the tumour in which it occurs [38]. Subtype C expressed the *BRAF* mutant signature we identified earlier [6] (87.0%), a CIMP-H signature ([11], Figure 2), and its characteristics (enrichment for MSI, right side and mucinous histology) corresponded with those of the previously reported CIMP-H phenotype [9,11,39,40] and hypermutated tumours [13]. Regarding the latter, subtype C had a similar low frequency of copy number variations (see Supplementary material, Figure S7). The distribution of MSI status, stage, age, gender, grade and site over the subtypes in the validation set followed the same patterns established in the discovery set [cf Figures 3 and S5B (see Supplementary material)]. A classification tree, trained with a combination of available clinical and molecular markers, did not identify our subtypes (see Supplementary material, Figure S5D), indicating that gene expression patterns reveal a layer of heterogeneity that goes beyond conventional CRC classification approaches.

Histological characteristics of subtypes

To study whether or not our molecular subtypes are associated with histological patterns, we examined haematoxylin and eosin (H&E)-stained paraffin sections of a randomly selected subset of each subtype (23, 31, 31, 29 and 19 cases for subtypes A, B, C, D and E, respectively). In attempting to match histological morphotypes to molecular subtypes, architectural patterns were used, as illustrated in Figure 4A, rather than the recognized WHO classification of CRCs [1]. Not surprisingly, given intratumour heterogeneity, none of the tumours had a single pattern. However, the prevalent patterns showed appreciable differences between the subgroups (Figure 4B, C; see also Supplementary material, Figure S6). In subtype A, the serrated pattern was most frequent, followed by the papillary pattern; in

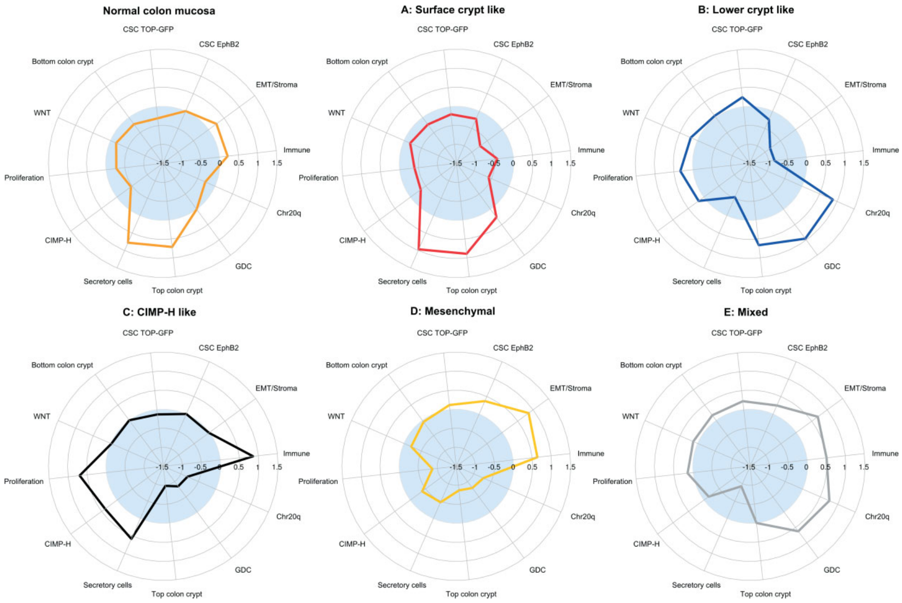


Figure 2. Subtypes and biological motifs. Subtype-specific fingerprints of biological motifs, represented either as mean values of gene set enrichment scores of gene sets from corresponding gene modules (EMT/stroma, immune, secretory cells, proliferation, GDC, chromosome 20q, top of the crypt—meta105 and meta144) or composed gene set enrichment scores of particular signatures (canonical Wnt targets, CSC-TopGFP, CSC-EphB2, colon crypt bottom and CIMP-H). The gene set enrichment scores represent whether the genes from the gene set show statistically significant enrichment between the down-regulated (negative scores, light blue area) or up regulated (positive scores) genes of a given subtype; details of score calculation can be found in the Supplementary material (Supplementary methods and results and Table S7.).

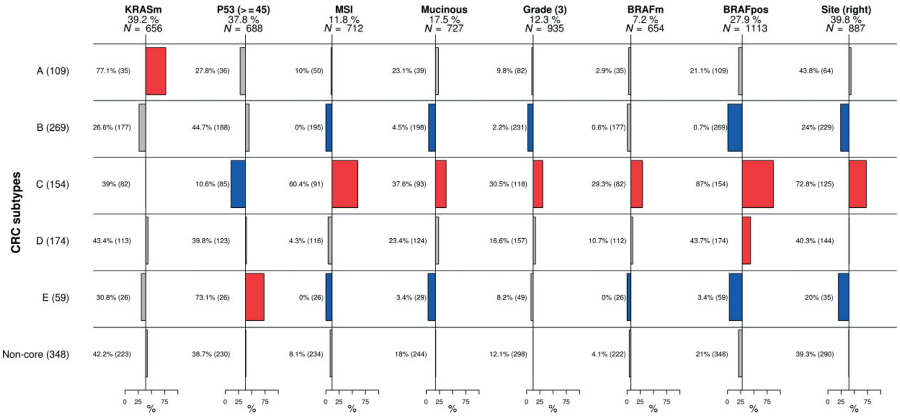


Figure 3. Clinical and mutational characterization of subtypes. Columns represent variables and rows subtypes. Horizontal bar plots represent proportions of the corresponding variable in each of the subtypes and non-core samples. Non-core samples were tested as one group to ensure that they did not share a common characteristic that would set them apart. Numbers in brackets adjacent to subtype name represent overall number of samples in the subtype. Under the title of each variable we denote the percentage representing baseline proportion in the population, with available information, and *N* denotes the number of patients for which the information on the respective feature was available. Bars in red represent significant enrichment and bars in blue significant depletion of a feature in the subtype in comparison to baseline, at the 5% significance level. Adjacent to each bar is the percentage of samples in the subtype with the specific feature and in brackets the overall number of samples in the subtype with the information available. We can read that, for instance, subtype C, comprising 154 samples, is enriched for microsatellite-unstable (MSI) tumours, where 60.4% of 91 samples with available information are MSI.

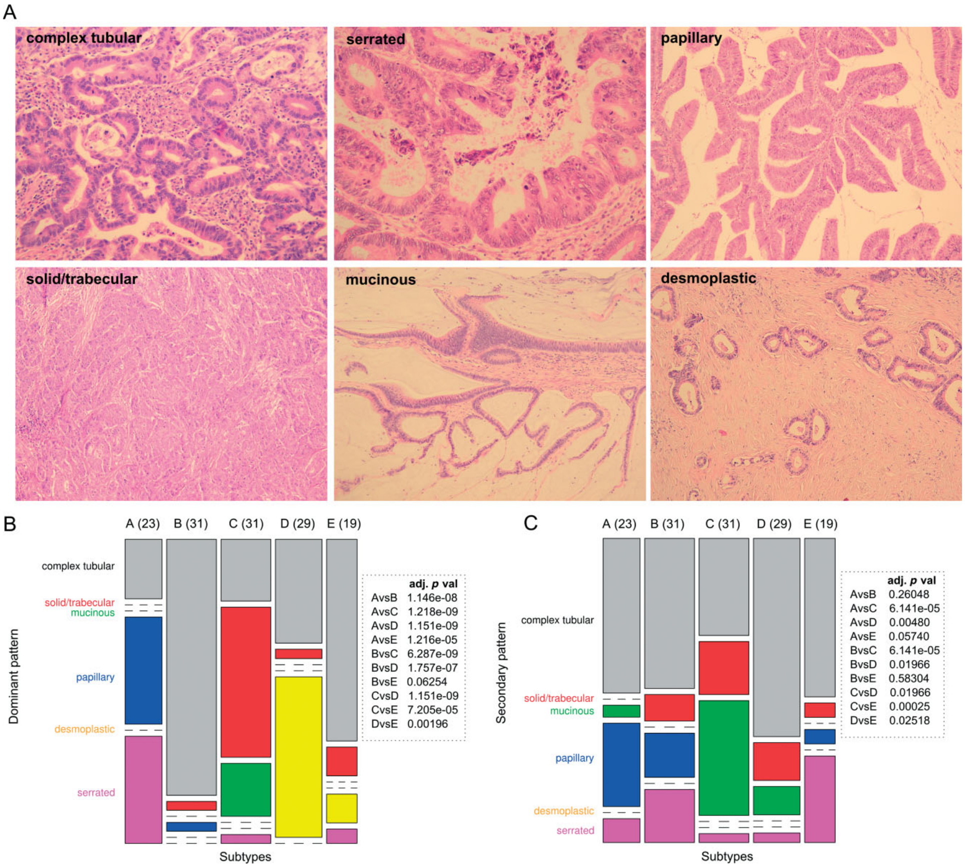


Figure 4. Morphological CRC patterns. (A) morphological CRC patterns scored in subtypes. (B, C) Distribution of dominant (B) and secondary (C) histological patterns in subtypes. Columns represent subtypes and widths are proportional to subtype frequency (numbers of samples in each subtype); rows represent dominant (B) or secondary (C) patterns and heights are proportional to pattern frequency. Boxes show adjusted *p* values of pairwise statistical testing of morphological pattern distribution between subtypes.

subtypes B and E, complex tubular dominated; in subtype C the solid pattern dominated, with mucinous as the second; most striking was the presence of a strong stromal reaction in subtype D.

Discussion

Our approach, using gene modules on a large panel of samples, allowed us to identify five main CRC gene expression subtypes (Table 4). It is relevant to note that subtyping can be performed on FFPE tissues, an important prerequisite for wide clinical applications. An example is the hypermutated group identified in the TCGA study by whole exome sequencing [13], but according to our data also by gene expression profiling on routinely processed tissues (CIMP-H-like subtype).

The combination of gene expression, clinical, mutational, survival and morphological data contributes new insight into the heterogeneity of CRC. While the validation confirmed the robustness of our findings across different platforms (ALMAC versus Affymetrix), sample preparation methods (FFPE versus fresh-frozen) and dataset collections, larger datasets are necessary to assess and characterize the relevance of lower frequency subtypes (eg F, or further segregation of B into B1 and B2). Our data indicate that several major biological processes are key determinants of a complex subtype structure of CRC. Therefore our subtypes defined by gene expression do not substitute but complement groups defined by current clinicopathological variables and molecular markers. Notably, morphological subclassification of CRC has clearly reached its limits, given the often striking intratumour

Table 4. Summary of subtype characteristics

	CRC markers and mutations				Histopathology	IHC	Median survival (months)			Clinical		Gene expression	
Subtype	MSI	*BRAF*	*KRAS*	*P53*	Dominant	Nuclear β-catenin at IF	OS	RFS	SAR	Site	Grade	Up-regulated	Down-regulated
A: Surface crypt-like		−	+		Papillary or serrated	−	NA	NA	28.9			Top colon crypt, secretory cell, metallothioneins	EMT/stroma, Wnt, CSC, Chr20q, proliferation
B: Lower crypt-like	−	−			Complex tubular	+	NA	NA	50.4	Left	2	Top colon crypt, proliferation, Wnt	EMT/stroma, immune, secretory cell
C: CIMP-H-like	+	+		−	Solid/trabecular or mucinous	−	NA	NA	6.9	Right	3	Proliferation, immune, metallothioneins	GDC, top colon crypt, Chr20q
D: Mesenchymal					Desmoplastic	−	NA	79.5	19.8			EMT/stroma, CSC, immune	Proliferation, secretory cell, top colon crypt, GDC, Wnt, Chr20q
E: Mixed	−	−		+	Complex tubular	+	NA	NA	19.6	Left		EMT/stroma, immune, top colon crypt, Chr20q, GDC, CSC	Secretory cell

+, significantly enriched; −, significantly depleted; IF, invasion front; NA, not attained; no value, no significant enrichment in comparison to population baseline.

heterogeneity, which made us use a (primary and secondary) architectural pattern approach rather than the canonized histological subtypes (WHO). Profiling of microdissected patterns within a single tumour might reveal molecular mechanisms responsible for these morphotypes. This additional heterogeneity within the subtypes may reflect tumour polyclonality, similar to breast cancer [41]. Ultimately, aggregating clinical, pathological and further detailed molecular characteristics (including CNV, miRNA and methylation) will contribute to a more detailed perception of CRC heterogeneity and it is likely that more subtypes will emerge. This, however, would need more detailed molecular annotation of larger clinically well documented CRCs.

A striking association was found between the stromal subtype D and the EMT signature. The previously discovered EMT [12] also emerged from our analysis as the largest cluster of meta-genes associated with poor RFS (subtype D). Our histological assessment suggests that the EMT signature is the reflection of a strong mesenchymal stromal reaction, and this histological characteristic deserves to be tested for its capacity to predict resistance to therapy, in view of its strong association with poor survival. Studies requiring high tumour cell content as sample inclusion criteria (eg [13]) could miss this poor prognosis subtype. Identification of this subtype in cell lines or xenograft models is less straightforward and would benefit from the analysis of gene expression patterns between microdissected tumour and stromal cells.

EMT, however important, only partly explains CRC heterogeneity, as even subtypes with similar expression of EMT-associated genes (A–C or D–E) differ in survival, mutational, clinical and gene expression characteristics. Additional biological components, such as differentiation, immune response, proliferation, chromosome 20q or cluster of genes deregulated in CRCs, are important co-determinants that underpin a need for further subdivision of CRCs. The findings from the analysis of CSC and WNT signatures support the recently suggested hypothesis that the colon stem cell signature under the condition of silenced canonical WNT targets is associated with higher risk of recurrence (subtype D) [33]. This is consistent with subtype D showing a significantly lower percentage of β-catenin-positive nuclei than subtype B, with its Wnt-associated gene expression and better survival.

MSI tumours represent a subclass in most unsupervised analyses and can be recognized at the gene expression level [42]. The more recent gene expression studies [14,15] suggest that MSI and *BRAF* share distinct gene expression patterns. Subtype C was enriched for both MSI and *BRAF* mutants and had one of the best outcomes for RFS, but the worse outcome in SAR, in concordance with previously reported results [43]. Subtype C retained its poor SAR prognostic value, even in the population of MSS and *BRAF* wild-type patients. Our data suggest that subtype C represents tumours with a common biology and a gene expression pattern that might best characterize a group of tumours resistant to chemotherapy, once metastatic. In this sense, our work not only agrees with the current known markers (*BRAF* mutation status and MSI) but clearly adds new insight, putting together these previously unrelated clusters into one biologically meaningful group. This observation is in line with recently published work [6].

Our observations show that gene expression profiling contributes substantially to our insight into CRC heterogeneity in confirming and complementing data from sequencing, CNV and promoter methylation analysis. Our subtypes can be further functionally interrogated for driving oncogenes/events by *in vitro* functional screens. High-risk subtypes D and C might contribute to therapeutic decision making in either adjuvant or metastatic settings. Retrospective analysis of clinical trial series may identify drug sensitivity associated with particular subtypes, and might open new treatment optimization strategies to be tested in clinical trials with stratified cohorts, similar to the I-SPY2 trial for breast cancer [44].

In conclusion, our unsupervised approach using gene modules resulted in the identification of distinct molecularly defined CRC subtypes, which adds a new layer of complexity to CRC heterogeneity and opens new opportunities for understanding the disease. The challenge is now to assimilate conventional and these new molecular approaches into a comprehensive consensus classification, which might then be used in further clinical studies for patient stratification and experimental studies to further elucidate mechanisms involved in the development and progression of CRC.

Acknowledgements

We thank all the clinicians who enrolled patients and participated in the PETACC-3 trial (see Appendix at: http://jco.ascopubs.org/content/27/19/3117.long), in particular the coordinators D Cunningham, R Labianca and E Van Cutsem. We thank F Schütz, T Sengstag and A Wirapati for providing analytical tools for clustering and gene set enrichment analysis. This work was supported in part by the Swiss National Science Foundation (SNF; http://www.snf.ch/E/; Grant No. 320030_135421, to AR and MD); Krebsforschung Schweiz (KFS; http://www.krebsforschung.ch/; Grant No. 02697-08-2010, to AR and MD); Fondation Medic (to EB and MD); and the Swiss National Centre for Competence in Research Molecular Biology (http://www.nccr-oncology.ch/; to VP and MD). ST is a Senior Clinical Investigator of the Fund for Scientific Research Flanders (FWO-Vlaanderen) and has received research grants from the Belgian National Cancer Plan (Nationaal Kankerplan). We thank the TCGA research network for the availability of data.

Author contributions

EB and MD designed the study; YP, FTB, ST, JGH and SW conceived and carried out microarray experiments; YP and FTB performed histopathological experiments and β-catenin scoring; EB, VP, GD, NL and AFN analysed the data, EB, VP, ST, FTB, KOS, NL, JGH, SW, MD and AR performed data interpretation; EB, NL and KOS performed the literature search; and EB generated figures and conceived the first manuscript draft. All authors were involved in writing the paper and had final approval of the submitted and published versions.

References

1. Bosman FT, World Health Organization, International Agency for Research on Cancer. *WHO Classification of Tumours of the Digestive System*, 4th edn. International Agency for Research on Cancer (IARC): Lyons, 2010.
2. Tejpar S, Saridaki Z, Delorenzi M, *et al*. Microsatellite instability, prognosis and drug sensitivity of stage II and III colorectal cancer: more complexity to the puzzle. *J Natl Cancer Inst* 2011; **103**: 841–844.
3. Sinicrope FA, Sargent DJ. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin Cancer Res* 2012; **18**: 1506–1512.
4. Vecchione L, Jacobs B, Normanno N, *et al*. EGFR-targeted therapy. *Exp Cell Res* 2011; **317**: 2765–2771.
5. Martini M, Vecchione L, Siena S, *et al*. Targeted therapies: how personal should we go? *Nat Rev Clin Oncol* 2011; **9**: 87–97.
6. Popovici V, Budinska E, Tejpar S, *et al*. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J Clin Oncol* 2012; **30**: 1288–1295.
7. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* 2007; **50**: 113–130.
8. Shen L, Toyota M, Kondo Y, *et al*. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 2007; **104**: 18654–18659.
9. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* 2008; **10**: 13–27.
10. Furlan D, Carnevali IW, Bernasconi B, *et al*. Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Mod Pathol* 2011; **24**: 126–137.
11. Hinoue T, Weisenberger DJ, Lange CP, *et al*. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012; **22**: 271–282.
12. Loboda A, Nebozhyn MV, Watters JW, *et al*. EMT is the dominant program in human colon cancer. *BMC Med Genom* 2011; **4**: 9.
13. TCGA CGAN. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330–337.
14. Salazar R, Roepman P, Capella G, *et al*. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**: 17–24.
15. Perez Villamil B, Romera Lopez A, Hernandez Prieto S, *et al*. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* 2012; **12**: 260.
16. Wirapati P, Sotiriou C, Kunkel S, *et al*. Meta-analysis of gene expression profiles in breast cancer: toward a unified

understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008; **10**: R65.

17. Farmer P, Bonnefoi H, Becette V, *et al*. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005; **24**: 4660–4671.
18. Shedden K, Taylor JM, Enkemann SA, *et al*. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; **14**: 822–827.
19. Xie T, G DA, Lamb JR, *et al*. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS One* 2012; **7**: e42001.
20. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008; **24**: 719–720.
21. Monti S, Tamayo P, Mesirov J, *et al*. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003; **52**: 91–118.
22. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer: New York, 2009.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; **67**: 301–320.
24. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**: Article 3.
25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
26. Arnholt AT. BSDA: Basic statistics and data analysis. R package v 1.01, 2012; http://cran.r-project.org/web/packages/BSDA/index.html
27. Tejpar S, Bertagnolli M, Bosman F, *et al*. Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist* 2010; **15**: 390–404.
28. Mokry M, Hatzis P, de Bruijn E, *et al*. Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One* 2010; **5**: e15092.
29. Hatzis P, van der Flier LG, van Driel MA, *et al*. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* 2008; **28**: 2732–2744.
30. Van der Flier LG, Sabates-Bellver J, Oving I, *et al*. The intestinal Wnt/TCF signature. *Gastroenterology* 2007; **132**: 628–632.
31. Sansom OJ, Reed KR, Hayes AJ, *et al*. Loss of APC *in vivo* immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev* 2004; **18**: 1385–1390.
32. Fevr T, Robine S, Louvard D, *et al*. Wnt/β-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol Cell Biol* 2007; **27**: 7551–7559.
33. de Sousa EMF, Colak S, Buikhuisen J, *et al*. Methylation of cancer stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell* 2011; **9**: 476–485.
34. Merlos-Suarez A, Barriga FM, Jung P, *et al*. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 2011; **8**: 511–524.
35. van der Flier LG, van Gijn ME, Hatzis P, *et al*. Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* 2009; **136**: 903–912.
36. Kosinski C, Stange DE, Xu C, *et al*. Indian hedgehog regulates intestinal stem cell fate through epithelial–mesenchymal interactions during development. *Gastroenterology* 2010; **139**: 893–903.
37. Faris JE, Ryan DP. Trees, forests, and other implications of a *BRAF* mutant gene signature in patients with *BRAF* wild-type disease. *J Clin Oncol* 2012; **30**: 1255–1257.
38. Singh A, Sweeney MF, Yu M, *et al*. TAK1 inhibition promotes apoptosis in *KRAS*-dependent colon cancers. *Cell* 2012; **148**: 639–650.
39. Tanaka H, Deng G, Matsuzaki K, *et al*. *BRAF* mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. *Int J Cancer* 2006; **118**: 2765–2771.
40. Hawkins N, Norrie M, Cheong K, *et al*. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology* 2002; **122**: 1376–1387.
41. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.
42. Tian S, Roepman P, Popovici V, *et al*. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol* 2012; **228**: 586–595.
43. Dahlin AM, Palmqvist R, Henriksson ML, *et al*. The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. *Clin Cancer Res* 2010; **16**: 1845–1855.
44. Barker AD, Sigman CC, Kelloff GJ, *et al*. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther* 2009; **86**: 97–100.
45. Moskaluk CA, Zhang H, Powell SM, *et al*. Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Mod Pathol* 2003; **16**: 913–919.
46. van den Brink GR, Bleuming SA, Hardwick JC, *et al*. Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation. *Nat Genet* 2004; **36**: 277–282.
47. Liu JY, Seno H, Miletic AV, *et al*. Vav proteins are necessary for correct differentiation of mouse cecal and colonic enterocytes. *J Cell Sci* 2009; **122**: 324–334.
48. Zheng H, Ying H, Wiedemeyer R, *et al*. PLAGL2 regulates Wnt signaling to impede differentiation in neural stem cells and gliomas. *Cancer Cell* 2010; **17**: 497–509.
49. Hirose H, Ishii H, Mimori K, *et al*. The significance of PITX2 overexpression in human colorectal cancer. *Ann Surg Oncol* 2011; **18**: 3005–3012.
50. Kontos CK, Papadopoulos IN, Fragoulis EG, *et al*. Quantitative expression analysis and prognostic significance of L-DOPA decarboxylase in colorectal adenocarcinoma. *Br J Cancer* 2010; **102**: 1384–1390.
51. Bhatavdekar J, Patel D, Ghosh N, *et al*. Interrelationship of prolactin and its receptor in carcinoma of colon and rectum: a preliminary report. *J Surg Oncol* 1994; **55**: 246–249.
52. Gaber A, Johansson M, Stenman UH, *et al*. High expression of tumour-associated trypsin inhibitor correlates with liver metastasis and poor prognosis in colorectal cancer. *Br J Cancer* 2009; **100**: 1540–1548.
53. Kawakami K, Ooyama A, Ruszkiewicz A, *et al*. Low expression of gamma-glutamyl hydrolase mRNA in primary colorectal cancer with the CpG island methylator phenotype. *Br J Cancer* 2008; **98**: 1555–1561.
54. Chen Y, Tang Y, Guo C, *et al*. Nuclear receptors in the multidrug resistance through the regulation of drug-metabolizing enzymes and drug transporters. *Biochem Pharmacol* 2012; **83**: 1112–1126.
55. Park SW, Zhen G, Verhaeghe C, *et al*. The protein disulfide isomerase AGR2 is essential for production of intestinal mucus. *Proc Natl Acad Sci USA* 2009; **106**: 6950–6955.
56. Noah TK, Kazanjian A, Whitsett J, *et al*. SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells. *Exp Cell Res* **316**: 452–465.

57. Steppan CM, Brown EJ, Wright CM, *et al*. A family of tissue-specific resistin-like molecules. *Proc Natl Acad Sci USA* 2001; **98**: 502–506.

58. Heiskala K, Giles-Komar J, Heiskala M, *et al*. High expression of RELP (Reg IV) in neoplastic goblet cells of appendiceal mucinous cystadenoma and pseudomyxoma peritonei. *Virchows Arch* 2006; **448**: 295–300.

59. Dalerba P, Kalisky T, Sahoo D, *et al*. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; **29**: 1120–1127.

60. *R Development Core Team. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.

61. *Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**: R80.

62. *Therenau T. *A Package for Survival Analysis in S*. R package version 2.36–14, 2012.

63. *Bolstad BM, Collin F, Simpson KM, *et al*. Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol* 2004; **60**: 25–58.

64. *Venables WNR, Ripley BD. Modern Applied Statistics with S, 4th edn. Springer: New York, 2002.

65. *Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Statist* 2006; **15**: 651–674.

66. *Van Cutsem E, Labianca R, Bodoky G, *et al*. Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J Clin Oncol* 2009; **27**: 3117–3125.

67. *Jorissen RN, Gibbs P, Christie M, *et al*. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin Cancer Res* 2009; **15**: 7642–7651.

68. *IGC. Expression Project for Oncology, 2008 [cited; available from: http://www.intgen.org/expo/]

69. *Smith JJ, Deane NG, Wu F, *et al*. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010; **138**: 958–968.

70. *Skrzypczak M, Goryca K, Rubel T, *et al*. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One* 2010; **5**: e13091.

71. *Hong Y, Ho KS, Eu KW, *et al*. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007; **13**: 1107–1114.

72. *Gyorffy B, Molnar B, Lage H, *et al*. Evaluation of microarray preprocessing algorithms based on concordance with RT–PCR in clinical samples. *PLoS One* 2009; **4**: e5645.

73. *Galamb O, Sipos F, Solymosi N, *et al*. Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 2835–2845.

74. *Galamb O, Spisak S, Sipos F, *et al*. Reversal of gene expression changes in the colorectal normal–adenoma pathway by NS398 selective COX2 inhibitor. *Br J Cancer* 2010; **102**: 765–773.

75. *Koinuma K, Yamashita Y, Liu W, *et al*. Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene* 2006; **25**: 139–146.

76. *Jorissen RN, Lipton L, Gibbs P, *et al*. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* 2008; **14**: 8061–8069.

77. *Grone J, Lenze D, Jurinovic V, *et al*. Molecular profiles and clinical outcome of stage UICC II colon cancer patients. *Int J Colorectal Dis* 2011; **26**: 847–858.

78. *Birnbaum DJ, Laibe S, Ferrari A, *et al*. Expression profiles in stage II colon cancer according to APC gene status. *Transl Oncol* 2012; **5**: 72–76.

79. *Giancarlo R, Scaturro D, Utro F. Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, *Gap Statistics and Model Explorer. BMC Bioinformat* 2008; **9**: 462.

80. *Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: x1–13.

*Cited only in the Supplementary material.

SUPPLEMENTARY MATERIAL ON THE INTERNET

The following supplementary material may be found in the online version of this article:

Supplementary methods and results (contains a further table and two further figures)

Figure S1. (A) Consensus clustering and similarity dendrogram of samples. (B) Subtype projection in the four-dimensional space of LDA axes. (C) Heat map matrix of pairwise meta-gene Fisher Z-transformed Pearson pairwise correlations. (D) Box plots of intra gene module pairwise gene–gene Pearson correlations in normal samples in both discovery and validation sets

Figure S2. Validation of meta-gene expression pattern of subtypes represented by heat maps

Figure S3. (A) Heat map representing validation of gene expression patterns of subtypes. (B) Pairwise Fisher Z-transformed correlations of meta-genes in validation set. (C) Box plots representing medians of pairwise gene–gene Pearson correlations in the validation datasets

Figure S4. Expression of top five down- and top five u*p* regulated genes from all pairwise comparisons between subtypes

Figure S5. (A) Other clinical and mutational markers tested and found non-significant between subtypes. (B) Clinical variables tested in the clusters of the validation test. (C) Distribution of significant clinical and mutational markers across subtypes. (D) Classification tree trained on clinical variables

Figure S6. Graphs of joined distribution of dominant vsersus secondary patterns in each of the subtypes

Figure S7. Heat map of CNV profiles of 154 samples from the discovery set, randomly ordered inside each of the subtypes

Figure S8. Result of hypothesis testing of median log-scale copy number estimates of chromosome 20 of subtype B versus all other subtypes

Figure S9. Distribution of β-catenin immunoreactivity of the invasion front counts between subtypes

Table S1. Detailed description of gene module members and detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

Table S2. Multiclass linear discriminant (LDA) subtype assignment of samples from validation set

Table S3. Correlations of subtype-specific gene expression profiles (1 versus all moderated *t* test statistics) when accounting for subtype F in the training set

Table S4. Detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

Table S5. Detailed results of pairwise comparisons of differentially expressed gene between subtypes

Table S6. Detailed results of Cox proportional hazards models for RFS, OS and SAR for subtype, stage, MSI and *BRAF* and for meta-genes

Table S7. Results of GSEA comparison of enrichment tested signatures in individual subtypes and normal samples

18 Image-based surrogate biomarkers for molecular subtypes of colorectal cancer

- Bioinformatics, 33(13):2002–2009, 2017
- IF: 7.307
- number of citations: 0
- personal contribution (80%): image analysis method design, data collection and processing, experimental design and implementation, manuscript writing

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Image-based surrogate biomarkers for molecular subtypes of colorectal cancer

Vlad Popovici [1,*], Eva Budinská [2], Ladislav Dušek [1], Michal Kozubek [3] and Fred Bosman [4]

[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic
[2] Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masaryk University, Brno, Czech Republic
[3] Faculty of Informatics, Masaryk University, Brno, Czech Republic and
[4] University Institute of Pathology, University of Lausanne, Switzerland.

* To whom correspondence should be addressed.

Associate Editor: Prof. Robert Murphy

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Whole genome expression profiling of large cohorts of different types of cancer led to the identification of distinct molecular subcategories (subtypes) that may partially explain the observed inter-tumoral heterogeneity. This is also the case of colorectal cancer where several such categorizations have been proposed. Despite recent developments, the problem of subtype definition and recognition remains open, one of the causes being the intrinsic heterogeneity of each tumor, which is difficult to estimate from gene expression profiles. However, one of the observations of these studies indicates that there may be links between the dominant tumor morphology characteristics and the molecular subtypes. Benefiting from a large collection of colorectal cancer samples, comprising both gene expression and histopathology images, we investigated the possibility of building image–based classifiers able to predict the molecular subtypes. We employed deep convolutional neural networks for extracting local descriptors which were then used for constructing a dictionary–based representation of each tumor sample. A set of support vector machine classifiers were trained to solve different binary decision problems, their combined outputs being used to predict one of the five molecular subtypes.

Results: A hierarchical decomposition of the multi-class problem was obtained with an overall accuracy of 0.84 (95%CI=(0.79-0.88)). The predictions from the image-based classifier showed significant prognostic value similar to their molecular counterparts.

Availability: Source code used for the image analysis is freely available from `https://github.com/higex/qpath`

Contact: popovici@iba.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

The last two decades witnessed fundamental changes in the way we investigate the biology of living organisms, with technological developments fueling major breakthroughs in our understanding of various pathologies and paving the road towards a personalized medicine. Currently, the researchers are armed with a battery of techniques for interrogating the same biological reality at various scales (from sub-cellular to whole population) and from very diverse perspectives (clinical, imaging, genomic, proteomic, etc) generating high throughput multimodal data. The bottleneck is now represented by our limited ability to interpret such data in an integrated way (Li *et al.* (2016)) and the need for a more inter-disciplinary approach is epitomized by large scale projects such as The Cancer Genome Atlas (TCGA). In cancer research, one of the main goals it to identify homogeneous groups of patients - *i.e.* to stratify the patient population - in the hope of finding the common causes and tailored treatments. Traditional stratification of cancer patients is based on histologic and morphologic assessment of the tumor sample and it still defines the golden standard. Lately, various molecular biomarkers have been proposed for the same purpose. The two perspectives are partly overlapping and partly orthogonal, making their integration more challenging. Our present work focusses on translating a gene expression-based cancer patient population stratification into an image-based biomarker, thus trying to bring transcriptomics data into a histopathologic context.

Colorectal cancer (CRC) is the third most frequent cancer worldwide and the second leading cause of cancer mortality in Europe, with metastatic disease accounting for 40% to 50% of newly diagnosed patients. At the same time, it is a highly heterogeneous disease in terms of prognosis and its response to therapy. Using whole-genome profiling of large data collections, several systems for sub-categorization of CRC have been proposed recently (Budinská *et al.*, 2013; Marisa *et al.*, 2013; Sadanandam *et al.*, 2013; Roepman *et al.*, 2013; De Sousa E Melo *et al.*, 2013). In general, they relied on clustering the CRC tumors in order to identify patterns of co-regulation of genes that could be indicative of common oncogenic pathways and coherent treatment responses of these tumors. Our own analysis (Budinská *et al.*, 2013) identified five stable tumor clusters (labeled as subtypes A, B,..., E), but also showed that a relatively high proportion of cases remained unaccounted for by this system. A recent effort (Guinney *et al.*, 2015) to harmonize all these discoveries confirmed the presence of four distinct and reproducible subtypes across all studies, labeled CMS1,..., CMS4, which match closely our subtypes A,..., D (Guinney *et al.*, 2015). The current golden standard for the identification of the molecular subtype of a given tumor requires the interrogation of a large panel of genes and the application of a genomic classifier. In the analyses reported here, we will use the subtypes as defined in Budinská *et al.*, 2013. There are several reasons for this choice: Firstly, since they were derived from the same gene expression data that accompany the images we use, it is hoped that the subtype assignment is less noisy. Secondly, in Budinská *et al.*, 2013 it is noted that an expert pathologist, when presented with the molecular categorization for a set of cases, was able to identify a number of morphological features that were preferentially enriched in one or a few of the subtypes hence, showing preliminary evidence that such connections exist. And thirdly, we are interested in identifying the imaging support for the five previously identified subtypes.

The problem of recognizing the tumor subtype based on imaging data is not new and probably the most studied is the case of breast cancer. For these cancers, five molecular subtypes are currently considered - Luminal A, Luminal B, basal, Her2-enriched and normal-like (Perou *et al.*, 2000) - and surrogate immunohistochemical stains are available (corresponding to hormonal status of ER, PR and Her2 and the invasion marker Ki-67, respectively). Consequently, automatic stain quantification is the strategy of choice for molecular subtype recognition from image data and it was shown to outperform the human expert (Stålhammar *et al.*, 2016). A systematic review of the connections between histological and molecular subtypes in breast cancer is given in Weigelt *et al.*, 2010. Other efforts concentrated on the recognition of the high risk group of triple negative breast cancers on various imaging platforms (Agner *et al.*, 2014; Dogan and Turnbull, 2012). The quantitative image analysis of pathology slides can also serve as a main means for subtype definition. For example, Chang *et al.*, 2011 found five subtypes of glioblastoma, one of which being predictive value and correlated with the expression of several genes. Similarly, Lan *et al.*, 2015 propose an alternative subtyping of ovarian cancer based on quantitative analysis of tumor microenvironment. A general approach to the identification of disease subtype based on morphologic analysis of pathology slides is described in Cooper *et al.*, 2012.

In the case of CRC, Budinská *et al.*, 2013 showed that subtype A had either serrated or papillary architecture, subtype B represented typical colorectal adenoma with complex tubular architecture, subtype C was mucinous or solid trabecular, subtype D was a mixture of desmoplastic and complex tubular architecture, and subtype E was mixed (see Budinská *et al.*, 2013 for example images). However, these annotations did not lead to a strong classifier.

This observation - that associations can be found between the molecular subtypes and morphological traits of the tumors - constitutes the starting point of our investigations reported here. Our interest is to construct a histopathology image-based classifier able to predict the molecular subtype of a given tumor section without resorting to any other staining but the standard haematoxylin-eosine. This classifier may be seen as a surrogate image biomarker (actually, as we will see, a combination of several biomarkers) for the molecular subtypes and, to the best of our knowledge, it is the first such biomarker to be proposed. This constitutes the main contribution of our work reported here and it represents a largely improved result from our earlier explorations (Budinská *et al.*, 2016). Equally important, our approach does not rely on predefined morpho-pathological features: the feature selection is guided by the prediction task. This would allow identifying potentially unknown (or overlooked) image features but may also make the interpretation of the models less obvious.

There are many potential application of such a system once established and well tested. First, since it does not require any special laboratory work, it could be easily integrated in the diagnostic workflow to provide hints about the molecular subtype, with no extra costs. It could also be used for sample stratification and selection for retrospective studies, where large collections of samples could easily be filtered for the subtypes of interest without the need of the much more expensive molecular profiling.

Currently, the molecular subtype is established by profiling the expression of a set of genes from the DNA/RNA extracted from the tumoral region of a tissue section and combining their values through a genomic classifier. The whole process involves a number of parameters (from defining the characteristics of the region to be profiled - tumor content, presence/absence of stroma, etc - to the cut-offs of the classifiers) that are yet to be formalized, thus being error-prone and leading to noisy labels. While we consider the molecular subtypes as the ground truth our image-based classifier is measured against, one has to keep in mind the somehow fuzzy nature of the class definition. These specific settings of our problem make it even more challenging than the more classical applications in the field of digital/computational pathology.

The rest of the paper is structured as follows: the data and the methods used are described in Section 1, followed by the discussion of the results in Section 2 and conclusions in Section 3.

1 Methods

1.1 Data

The present work is based on the data from a subset of the PETACC3 clinical trial (Van Cutsem *et al.*, 2009) samples. The trial compared two treatment regimens (fluorouracil/leucovorin alone or in combination with irinotecan) in colorectal cancer and found no differences between the two. The gene expression data for a set of $n = 688$ samples was used (along with other data sets) in the derivation of the molecular subtypes of CRC (Budinská *et al.*, 2013) and is publicly available from ArrayExpress under accession number E-MTAB-990. In (Budinská *et al.*, 2013) the molecular subtypes (denoted A-E) were assigned to a number of $n = 458$ cases, the rest being considered ambiguous (or representing other low-prevalence subtypes) and were labeled as "outliers". From those 458 samples, $n = 300$ cases were selected for this study based purely on technical considerations (availability of histopathology tumor section, acceptable whole slide image quality, tissue sample not too fragmented,

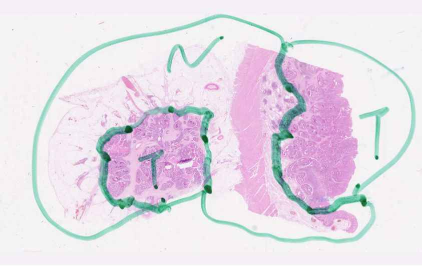


Fig. 1. Typical whole slide image from the data collection. At $10\times$ magnification, this image is 39936×22528 pixels in size. The regions marked with a "T" correspond to tumoral component, while the "N" annotation indicate normal tissue.

etc.). The "outlier" (from a molecular subtype perspective) cases were not considered in the present study.

All molecular subtypes were represented in this collection with the following frequencies: A: 21, B:140, C:37, D: 81, and E: 21, respectively. The slides were annotated by an expert pathologist and these annotations were present in the digital versions - a typical example is given in Figure 1 (note the annotations delineating the loosely the tumoral and normal tissue components).

From the whole collection of 300 images a subset of 100 images was selected by stratified random sampling to form the *development set*. This development set was used for selecting the image representation model and for designing the classification approach. We did not use the whole available data in order to reduce the likelihood of obtaining a model too adapted to our particular collection of samples (overfitted). For the same reason we also preferred limiting the number of experiments, comparing only several modeling approaches.The remaining 200 images were added at a later stage when the multi-class classifier performance was estimated by cross-validation. Other strategies of selecting a development set (eventually larger, equal number of cases per class, etc.) could have been attempted, with their own advantages and drawbacks, but we found the chosen approach to provide a reasonable trade-off.

1.2 Image acquisition and preprocessing

All whole slide images of haematoxylin-eosin stained tumor sections were acquired at $20\times$ magnification, using a Hamamatsu NanoZoomer C9600 scanner. The resulting images were compressed by the image acquisition software using JPEG standard (at 80% quality) and stored in the proprietary NDPI format. The resolution of the images was 455nm/pixel (equivalent of 55824 DPI) for a typical size of $100,000 \times 50,000$ pixels (depending on the size of the tissue section). The images were exported in standard TIFF format using OpenSlide software library (Satyanarayanan *et al.*, 2013).

The images were down-scaled to an equivalent $10\times$ magnification and only tumoral regions were retained from each sample (manually cut following the pathologist's annotations) - the pixels outside the tumors being set to zero. For example, the image in Figure 1 contains two tumoral regions (marked with "T"). No further preprocessing was applied to the images.

1.3 Local descriptors

We based our sample description on the aggregation of local information over the tumor regions in the image. The choice of image features plays a major role in the performance of image recognition/classification system. Traditionally, most of such features are handcrafted, consisting of some dense sampling of local patches, like in wavelet decomposition, Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), Local Binary Patterns (LBP) (Ojala *et al.*, 1996), etc. These local descriptors are later pooled into a global representations by means of methods such as Bag-of-Visual-Words (BoVW) (Csurka *et al.*, 2004), Fisher Vector (FV) (Perronnin and Dance, 2007), or Vector of Locally Aggregated Descriptors (VLAD) (Jégou *et al.*, 2010, 2012).

More recently, Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1989, 2015) gained momentum due to the superior performance of the systems employing them and to the increasing availability of dedicated software (and hardware) systems facilitating their use. While the CNNs also require a number of design decisions (such as their structure), they also have a large number of parameters that are learned from data, leading to adapted image descriptions. Cimpoi *et al.*, 2016 provide a detailed comparison of deep image features and some standard ones in the general context of texture classification. In biomedical imaging, there are a number of successful recognition systems based on various CNNs architectures, such as U-Net (Ronneberger *et al.*, 2015). In general, training CNN-based recognition systems requires a large number of labeled image examples, the deeper the architecture more images being needed. For example, the well-known image recognition systems like ImageNet (Krizhevsky *et al.*, 2012) or GoogleNet (Szegedy *et al.*, 2015) were trained on millions of images. Such large data collections are usually not available in biomedical field, thus the interest in transferring general pre-trained CNN models to the medical applications. For example, van Ginneken *et al.* (2015) and Kawahara *et al.* (2016) describe such successful systems that are based on pre-trained CNN features.

An alternate route for obtaining local descriptors is represented by the autoencoding methods, where an identity function is learned under the constraint of a lower dimensional (or sparse) internal representation. The parameters of the function are obtained through an optimization process, where the distance (usually L_2) between the original and reconstructed image is minimized, eventually with some additional constraints over the parameters. Examples of such methods are represented predictive sparse decomposition methods (as used in Chang *et al.* (2015) for example) and deep autoencoding networks. We do not explore further this direction on the present work.

For the problem addressed here, we chose to use a very deep CNN trained on *ImageNet* data collection – *imagenet-vgg-f* (Chatfield *et al.*, 2014) – as implemented in the MatConvNet library (Vedaldi and Lenc, 2015)[1]. The network is trained to predict the probability of an input color image of size 224×224 to belong to one of the $1,000$ categories. By using the output of the next to last layer (*relu7*, before the classification layers), a $4,096$ element description vector can be obtained. Since we will use Gaussian Mixture Models (GMMs - see Section 1.4) for building the coding dictionary, such a high dimensional space would require a prohibitively large number of samples for a good fit of the models, so we choose to perform PCA to further reduce the dimension of the local descriptor vectors by retaining the first $d = 128$ coordinates (chosen to be fixed, non-trainable). Thus, a local RGB patch of 224×224 pixels was reduced to a set of 128 values corresponding to the projection of the $4,096-$value ImageNet vector onto the first 128 principal axes.

As a side note, we remark that the CNN-based descriptor vector is itself the result of a combination of a number of filters applied to even smaller neighborhoods. However, in this work we consider the basic neighborhood to be the 224×224 patch on which the CNN is applied.

[1] for the architecture see `http://www.vlfeat.org/matconvnet/models/imagenet-vgg-f.svg`

1.4 Aggregating local descriptors

Once a set of local descriptors is obtained from an image, they are pooled into a summarizing feature vector supposed to capture the global aspects of the image. The first step of the process involves the re-coding of the image in terms of elements of a *visual dictionary (codebook)*, the same for all classes, which is followed by the computation of the image representation.

For the construction of the codebook, $k-$means clustering and Gaussian Mixture Models (GMMs) are the most common choices, and are typically used with either the standard *Bag-of-Visual-Words* (Csurka *et al.*, 2004) or other aggregators. Jégou *et al.*, 2012 give a comprehensive comparison of various design choices. Here we shortly remind the main differences between BoVW, FV and VLAD:

- *Bag-of-Visual-Words* typically uses $k-$means clustering for obtaining a codebook, with the K centroids from the clustering being the codewords (visual words). Then the representation of an image is simply the histogram of the number of local descriptors assigned to each codeword, thus an image is reduced to a $K-$dimensional vector. This histogram can be further normalized using Manhattan or Euclidean normalization Jégou *et al.*, 2012. One can also use a soft-coding scheme in which the patches are assigned, for example, a code based on the distance to the centroids Sivic and Zisserman, 2003.
- *Fisher Vector* represents a generalization of BoVW as it encodes higher order statistics of the distribution of the codewords. In this case, the codebook is usually obtained as a GMM with K components fitted via expectation maximization on the training data. The FV encodes the gradient of a given sample's likelihood with respect to parameters of the fitted GMM, thus it indicates the direction in the parameter space in which the learned GMM has to be modified to accommodate the observed data Jégou *et al.*, 2012. For a full FV that accounts for differences both in mean and variance between the model and observed data, the resulting representation vector has $2Kd$ elements (d being the size of the local descriptor vector).
- *VLAD* can be seen as a non-probabilistic version of FV Jégou *et al.*, 2012 and was designed to provide a low dimensional representation of the image Jégou *et al.*, 2010 that would allow the indexing of very large image databases in memory. It tries to combine the simplicity of BoVW with some ideas of FV: the codebook is learned via $k-$means clustering and each patch is assigned the closest codeword as in BoVW, but the feature vector accumulates the differences between each patch and its corresponding codeword, similar to FV. See Arandjelovic and Zisserman, 2013 for a detailed discussion and further extensions.

In the present work, we decided to use a common method for constructing the visual codebook, namely the Gaussian Mixture Models. This allowed us to test a soft-coding scheme as well, in which codes were based on the posterior probabilities of being generated by a particular component of the GMM.

1.5 Classifier training and performance estimation

Training the system could be summarized by the following steps:

1. for each image, extract the local descriptors (based on ImageNet) for all non-overlapping regions corresponding to tumoral component(s);
2. construct a visual codebook by:
   a. performing PCA and retain the first 128 components (the PCA model is saved for later application on validation set)
   b. fitting a $K = 128$-component GMM on PCA-transformed local descriptors (the visual codebook is saved for later usage on validation set)
3. train the binary classifiers (save the models for validation). Each such binary classifier was a support vector machine with a radial basis function kernel. Two parameters were tuned in an inner cross-validation loop: the γ parameter of the kernel and the C parameter for the misclassification penalty. The final prediction of the subtype label is made according to the decision tree in Figure 2. This particular decomposition of the multi-class problem was the result of the analysis of misclassified samples in the development set which suggested that firstly subtypes A, B should be separated from the rest (see Sec. 2.1).

Table 1. Confusion matrix for BoVW. Empty cells correspond to null values.

	Predicted						
	A	B	C	D	E	Precision	Recall
A	3	4				0.75	0.43
B	1	41		5		0.76	0.87
C		3	7	2		0.44	0.58
D		4	8	13	2	0.59	0.48
E	1	2	1	2	1	0.33	0.14

Since the ImageNet is an external model independent of the data analyzed, it does not need to be included in the cross-validation loop, this being an additional reason for preferring a pre-built CNN model. The other steps, however, were repeated at each cross-validation iteration on the corresponding training data.

1.6 Statistical analyses

For the identification of image features enriched/depleted in a subtype with respect to the other subtypes, we used Wilcoxon rank-sum tests since the measurements were not normally distributed. For hierarchical clustering we used the Ward method with an Euclidean distance between feature vectors. Survival analysis was performed using `survival` package (version 2.39-4) from `R` statistical computing environment (version 3.3.1, `www.r-project.org`). The estimation of hazard ratios was obtained from Cox proportional hazards regression in the absence of any other covariates, while the comparison of survival experience of different subgroups was assessed by log-rank test (Mantel-Haenszel test). Statistical significance level was chosen to be $p = 0.01$ and all tests yielding a $p-$value $0.01 \leq p \leq 0.05$ were considered marginally significant. Finally, the 95% confidence intervals (95%CI) for binomial random variables (such as accuracy) were estimated using the (Agresti and Coull, 1998) method.

2 Results and discussion

The results discussed here are complemented by larger images on the project's website: `http://bias.cerit-sc.cz/somopro-subtypes.html`.

2.1 Initial experiments

As mentioned, in an attempt to avoid overfitting the available data, a development set has been used to guide the design decisions and to set a number of meta-parameters. We tested dictionaries with $K_1 = 64$ and $K_2 = 128$ codewords and compared the performance of BoVW, FV and VLAD representations when predicting the five molecular subtypes. We performed this comparison under two standard decompositions of the multi-class classification problem, namely *1-vs-all* and *1-vs-1*.

These tests showed that BoVW with GMM-based quantization performed as good as the more involved representation by FV and VLAD

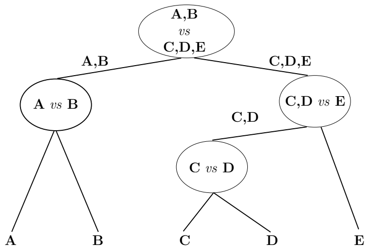


Fig. 2. Decomposition of the multi-class classification problem. For each non-terminal node a binary classifier was trained to split the respective groupings of molecular subtypes.

Table 2. 10-fold cross-validation confusion matrix for the multi-class classifier and corresponding per-class performance metrics. Empty cells correspond to null values.

	Predicted						
	A	B	C	D	E	Precision	Recall
A	21					0.95	1.00
B	1	119		13	7	0.91	0.85
C		2	29	6		0.91	0.78
D		8	1	71	1	0.75	0.88
E		2	2	5	12	0.60	0.57

see Supplement materials - Sec. 1. The small sample size definitely influences this observation, since both FV and VLAD have much higher dimensionality and would require more data for a better training. Table 1 shows the results for BoVW method with 1-vs-all decomposition of the multi-class problem, on the development set (obtained by stratified 4-fold cross-validation) - for the other approaches the results were similar, so they are not detailed here.

Another important observation was that the 1-vs-1 and 1-vs-all decompositions of the multi-class classification problem might not be the best suited for the present case. By analyzing the confusion matrix and taking into account the performance indexes (precision and recall) it appeared that a first split would have been more advantageous between classes A,B on one side and C,D,E, on the other. This observation is also supported by the results in Budinská *et al.*, 2013 where it is noted that subtypes A, B, on one hand, and C, D, E, on the other hand, share dominant and secondary dominant morphological features as well as similar survival expectancy. So, the final design for the multi-class classifier was chosen to be as depicted in Figure 2.

2.2 Prediction of molecular subtypes

Once the final decisions for the classification system were taken based on the initial experiments described above, the performance of the system was assessed using $10-$fold cross validation, on the whole set of 300 samples.

The estimated overall accuracy of the multi-class classifier was $\mathrm{Acc} = 0.84, 95\%\mathrm{CI} = (0.79-0.88)$ for a weighted average recall and precision of $\mathrm{R} = 0.85$, $95\%\mathrm{CI} = (0.80 - 0.89)$ and $\mathrm{P} = 0.84$, $95\%\mathrm{CI} = (0.80 - 0.88)$, respectively. Table 2 details the performance metrics of the classifier. We note the good performance of the first decision level ({A,B} *vs* {C,D,E}) ($\mathrm{Acc} = 0.89, 95\%\mathrm{CI} = (0.85-0.92)$) but also the poor recognition of the subtype E.

We repeated the same experiments on the 200 samples not used in the development set and the results were in line with those above (thus not repeated here), only with subtype A being slightly worse separated from subtype B (see Supplemental materials - Sec. 2). This indicates that the current sample size may still be too small for some cases and some improvements may be expected by enlarging the training set.

2.3 Associations between predictions and clinical data

The study Budinská *et al.*, 2013 indicated that some associations could be found between molecular subtypes and clinical variables and molecular markers. Hence, we were interested in testing whether such associations are transferrable to the predictions made by the image-based classifier. To avoid overly-optimistic discoveries, we use the predictions (A-E labels) produced during the cross-validation estimation of the system. There is also one caveat: as explained the selection of the cases was governed by technical constraints and thus it does not represent the true population-based statistics for various clinical variables and the results reported here should not be compared directly with those in Budinská *et al.*, 2013. Nevertheless, we investigate these associations and compare them with those found between gene expression-based subtypes and the clinical variables, on the same set of cases.

We first tested whether the predicted subtypes were associated with relapse free survival (RFS). In Budinská *et al.*, 2013, subtypes A and B have a lower risk of relapse than subtypes C, D, and E. The same can be observed in the set of 300 samples used here ($p = 0.0014$, $\mathrm{HR} = 1.75, 95\%\mathrm{CI} = (1.24 - 2.49)$, Figure 3(a)). The image-based subtype predictions also produce a statistically significant stratification of the population ($p = 0.012$, $\mathrm{HR} = 1.56, 95\%\mathrm{CI} = (1.10 - 2.21)$, Figure 3(b)).

We also found associations between microsatellite stability, BRAF and KRAS mutations, and mucinous histology and various subtypes - both image-based and gene expression-based. In the case of image-based predictions, subtypes A and C were enriched in mucinous histology compared to the sample average, while subtype E was almost depleted of it. BRAF mutated cases (5.8% of all cases) were mostly found in subtype C (20% of cases predicted), and rarely in subtype B (2.4%), while KRAS mutation (38.4% of all cases) represented 77% of cases predicted as subtype A and only 29% and 22% of cases predicted as subtypes B and E, respectively. Finally, high microsatellite instability (MSI) was almost exclusively found in subtype C (10 out of 13 cases). The same trends were found in gene-expression subtypes, with some variations below statistical significance.

A related question was whether the misclassified samples were enriched in any particular type of tumors. The only significant association was between the misclassified subtype B samples, which were enriched in higher T-stage and N-stage tumors. This observation may provide hints about further refinement of the classifier for subtype B. Detailed results are given in Supplemental materials - Sec. 3.

2.4 Visual codebook

We explored the structure of the visual codebook as obtained by training the model on the full data set. A visual depiction of the extracted codewords (centers of the Gaussian components) is shown in Figure 4 and a higher resolution image is given in Supplemental materials - Sec. 4. Note that the visual codewords are the centers of the Gaussians in the GMM, hence the means of feature vectors obtained by projecting the ImageNet features in the PCA space. The patches shown are just the closest image neighborhoods to these centers, thus they are an approximation of the true centers (whose visual appearance would require inverting the CNN function). We use this simplification only for visualization purposes and to get a qualitative assessment of the results.

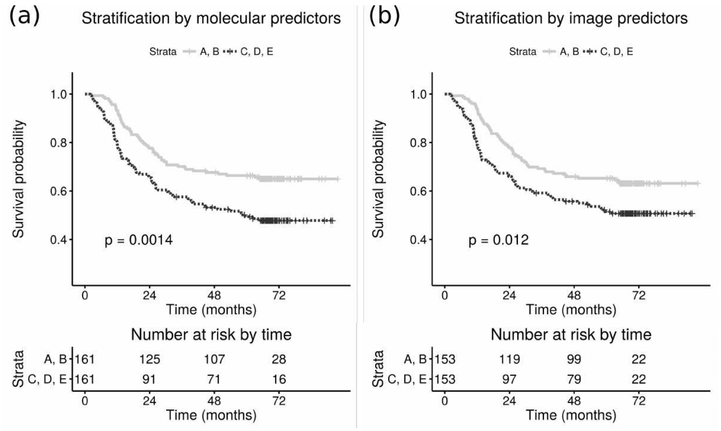


Fig. 3. Survival analysis: risk of relapse stratified by (a) molecular subtypes and (b) image-based classifier, respectively. Subtypes A and B represent a lower risk group, while subtypes C, D and E a higher risk, respectively.

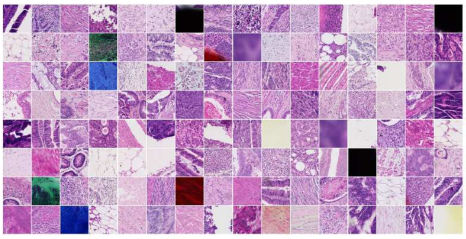


Fig. 4. Visual dictionary for colorectal cancer. While most of the selected visual words correspond to various tissue architectures, some are clearly linked to artifacts still present in the images, or regions partially covered by the annotations. The ordering of the image patches is given by the index in the GMM, with indexes from 0 to 127 (by rows).

As one can see most of the codewords could be associated with distinct tissue architectures (from various parts of the glands, papillary or tubular structures, to necrotic and fat regions). On the other hand, it is apparent that some of the codewords were affected - to different degrees - by the markings on the slides. Finally, a few codewords clearly corresponded to artifacts (either due to out-of-focus regions or markings). However, none of these artifact-related codewords were found to be associated with the subtypes, indicating that the approached use can cope, to some extent, with the noise inherent in such images.

Some of the codebooks had a much higher incidence in a particular subtype than in all the others (Wilcoxon rank-sum test). In Figure 5 the top four visual codewords resulted from this analysis are shown along with the corresponding p-values (no adjustment for multiple testing was performed, since this is purely exploratory). For all the subtypes but E the associations were statistically significant ($p \leq 0.01$). The subtype E seemed to not have a strong preference for any of the codewords, the few found associations being weakly statistically significant ($0.01 \leq p \leq 0.05$). It appears that subtype A is associated with well differentiated morphology (Figure 5 (a-d)), with subtype B being less well differentiated (Figure 5 (e-h)). For subtypes C, D and E, the top codewords could be associated with either necrotic tissue (Figure 5 (j),(l)), stromal reaction (Figure 5 (m-p)) or poorly differentiated morphology (Figure 5 (q)). It is important to stress that the classifiers were built based on non-linear support vector machines, so the results from this analysis cannot be directly extrapolated to understanding the classification models.

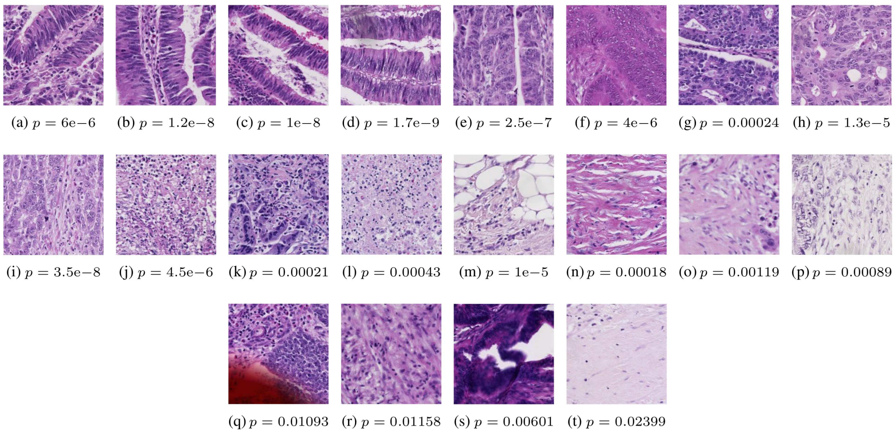
We performed a hierarchical clustering (Ward method) of all the codewords using Euclidean distance and the result showed a rather structured codebook (see Supplemental materials - Sec. 5). By corroborating the clustering results with those above, one can see that there are two major clusters - one corresponding mostly to features that are enriched in subtypes A and B (and depleted in C, D, E) and one corresponding to features enriched in subtypes C, D, E. This post-hoc analysis supports our decision of having a first decision level separating subtypes A, B from subtypes C, D, E.

3 Conclusion

We presented an approach at recognizing the colorectal cancer molecular subtypes from the routine histology images. The results indicate that an automated system could be built to identify with high confidence at least four of the five subtypes - subtype E apparently being much more challenging to recognize. The predictions made by the classifier were found to be also prognostic for relapse free survival and associated with other clinical parameters, as their molecular counterparts.

The models used for predicting the subtypes are based on support vector machine classifiers with radial basis functions kernels, making the direct interpretation of the models rather intricate. Nevertheless, we qualitatively evaluated the image features by testing their associations with various subtypes and inspecting their distribution in the whole image. To obtain better insights, we plan to also build simplified models - even at the expense on degraded performance - that would better lend themselves to a biological interpretation, a mandatory condition for the acceptance of the system.

In the current work, we concentrated on recognizing the five molecular subtypes from pre-segmented tumoral regions. This simplification will be addressed in future work where we plan to use an automatic segmentation of the tumor region as a preprocessing step for the subtype recognition.



(a) $p = 6\text{e}{-6}$ (b) $p = 1.2\text{e}{-8}$ (c) $p = 1\text{e}{-8}$ (d) $p = 1.7\text{e}{-9}$ (e) $p = 2.5\text{e}{-7}$ (f) $p = 4\text{e}{-6}$ (g) $p = 0.00024$ (h) $p = 1.3\text{e}{-5}$

(i) $p = 3.5\text{e}{-8}$ (j) $p = 4.5\text{e}{-6}$ (k) $p = 0.00021$ (l) $p = 0.00043$ (m) $p = 1\text{e}{-5}$ (n) $p = 0.00018$ (o) $p = 0.00119$ (p) $p = 0.00089$

(q) $p = 0.01093$ (r) $p = 0.01158$ (s) $p = 0.00601$ (t) $p = 0.02399$

Fig. 5. Top four prototypes associated with each subtype: (a-d) subtype A; (e-h) subtype B; (i-l) subtype C; (m-p) subtype D; and (q-t) subtype E. Under each image the corresponding p-value from Wilcoxon rank-sum test is shown.

Another question we will address in the future pertains the classification of the so-called "outliers": tumors for which no molecular subtype was assigned. It would be interesting to see how the subtypes predicted by the current image-based classifier correlate with the similarity between their expression profiles and those of well assigned tumors.

One has to bear in mind that, despite recent efforts to consolidate the molecular taxonomy of CRC, the sub-categorization of CRC is still not definitive. Indeed, depending on the size of the cohort and parameters chosen for cut-offs, more or less molecular subtypes can be observed, thus this categorization is still fluid. Nevertheless, in the present work it has been considered the golden standard to which the image-based models were compared against. We believe that actually combining the observations from the two modalities may led to an even more refined subtyping of the CRC. However, this would probably involved a more supervised (by expert pathologists) construction of the image-based models.

As they stand now, our results are clearly supporting the possibility of translating some molecular observations into image-based models, as it is the case of molecular subtypes. These results are reinforced by similar observations made by an expert pathologist (Budinská *et al.*, 2013), where several tissue architectural patterns could be linked, in a supervised analysis, to the molecular subtypes. It is interesting to note that some of the the regions/patterns found representative in our data-driven analysis are also visually similar to those hand-picked by an expert (see example images in Budinská *et al.*, 2013). On the other hand, the intra-tumoral heterogeneity and pathology sampling region clearly influence sample's assignment to a molecular subtype (Dunne *et al.*, 2016). In the light of the results presented here, it can be imagined an image-analysis approach to the delineation of the tissue sampling regions to improve the stability of the subtype assignment.

While it is too early for considering any clinical application of the models described here, they could, however, be used for indexing/annotating or for retrieval of samples of interest from archives. Consider the situation in which one would like to test for some biomarker which is hypothesized to work in one or several subtypes on a retrospective collection of samples. Since determining the molecular subtypes relies on profiling hundreds of genes, it makes more sense to use a classifier such the one proposed here, to select the most promising samples. And this can be implemented without significant effort since more and more of the pathology departments are adopting the digital pathology workflows, thus the images being readily available.

Acknowledgements

The necessary computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085 projects under the programme "Projects of Projects of Large Research, Development, and Innovations Infrastructures".

Funding

This project is financed from the SoMoPro II programme. The research leading to this result has acquired a financial grant from the People Programme (Marie Curie action) of the Seventh Framework Programme of EU according to the REA Grant Agreement No.291782. The research is further co-financed by the South Moravian Region. This article reflects only the authors views and the Union is not liable for any use that may be made of the information contained therein.

References

Agner, S. C., Rosen, M. A., Englander, S., Tomaszewski, J. E., Feldman, M. D., Zhang, P., Mies, C., Schnall, M. D., and Madabhushi, A. (2014). Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. *Radiology*, **272**(1), 91–99.

Agresti, A. and Coull, B. A. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, **52**(2), 119–126.

Arandjelovic, R. and Zisserman, A. (2013). All About VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585. IEEE.

Budinská, E., Popovici, V., Delorenzi, M., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Hodgson, J. G., Weinrich, S., Bosman, F., and Roth, A. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *Journal of Pathology*, **231**(1), 63–76.

Budinská, E., Bosman, F., and Popovici, V. (2016). Experiments in molecular subtype recognition based on histopathology images. In *International Symposium on Biomedical Imaging*, pages 1168–1172. Masaryk University, Brno, Czech Republic, IEEE.

Chang, H., Fontenay, G. V., Han, J., Cong, G., Baehner, F. L., Gray, J. W., Spellman, P. T., and Parvin, B. (2011). Morphometic analysis of TCGA glioblastoma multiforme. *BMC Bioinformatics*, **12**(1), 484.

Chang, H., Zhou, Y., Borowsky, A., Barner, K., Spellman, P., and Parvin, B. (2015). Stacked Predictive Sparse Decomposition for Classification of Histology Sections. *International Journal of Computer Vision*, **113**(1), 3–18.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference*.

Cimpoi, M., Maji, S., Kokkinos, I., and Vedaldi, A. (2016). Deep Filter Banks for Texture Recognition, Description, and Segmentation. *International Journal of Computer Vision*, **118**(1), 65–94.

Cooper, L. A. D., Kong, J., Gutman, D. A., Wang, F., Gao, J., Appin, C., Cholleti, S., Pan, T., Sharma, A., Scarpace, L., Mikkelsen, T., Kurc, T., Moreno, C. S., Brat, D. J., and Saltz, J. H. (2012). Integrated morphologic analysis for the identification and characterization of disease subtypes. *Journal of the American Medical Informatics Association*, **19**(2), 317–323.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 59–74.

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P. M. H., de Jong, J. H., de Boer, O. J., van Leersum, R., Bijlsma, M. F., Rodermond, H., van der Heijden, M., van Noesel, C. J. M., Tuynman, J. B., Dekker, E., Markowetz, F., Medema, J. P., and Vermeulen, L. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, pages 1–8.

Dogan, B. E. and Turnbull, L. W. (2012). Imaging of triple-negative breast cancer. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, **23 Suppl 6**, vi23–9.

Dunne, P. D., McArt, D. G., Bradley, C. A., O'Reilly, P. G., Barrett, H. L., Cummins, R., O'Grady, T., Arthur, K., Loughrey, M. B., Allen, W. L., McDade, S. S., Waugh, D. J., Hamilton, P. W., Longley, D. B., Kay, E. W., Johnston, P. G., Lawler, M., Salto-Tellez, M., and Van Schaeybroeck, S. (2016). Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. *Clinical Cancer Research*, **22**(16), 4095–4104.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, **21**(11), 1350–1356.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3304–3311. INRIA, Le Chesnay, France, IEEE.

Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., and Schmid, C. (2012). Aggregating Local Image Descriptors into Compact Codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(9), 1704–1716.

Kawahara, J., BenTaieb, A., and Hamarneh, G. (2016). Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging*, pages 1397–1400. Simon Fraser University, Burnaby, Canada, IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, pages 1106–1114.

Lan, C., Heindl, A., Huang, X., Xi, S., Banerjee, S., Liu, J., and Yuan, Y. (2015). Quantitative histology analysis of the ovarian tumour microenvironment. *Scientific Reports*, **5**, 16317–16317.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural computation*, **1**(4), 541–551.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436–444.

Li, G., Bankhead, P., Dunne, P. D., O'Reilly, P. G., James, J. A., Salto-Tellez, M., Hamilton, P. W., and McArt, D. G. (2016). Embracing an integromic approach to tissue biomarker research in cancer: Perspectives and lessons learned. *Briefings in Bioinformatics*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1150–1157. The University of British Columbia, Vancouver, Canada.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., and Boige, V. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine*, **10**(5), e1001453.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, **29**(1), 51–59.

Perou, C. M., Sorlie, T., Eisen, M. B., and van de Rijn, M. (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–752.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Xerox Research Centre Europe, Meulan, France, IEEE.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F., Bernards, R., and Simon, I. M. (2013). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer*, **134**(3), 552–562.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, Cham. Springer International Publishing.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W., and Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, pages 1–8.

Satyanarayanan, M., Goode, A., Gilbert, B., Harkes, J., and Jukic, D. (2013). OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, **4**(1), 27.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1470–1477. University of Oxford, Oxford, United Kingdom.

Stålhammar, G., Martinez, N. F., Lippert, M., Tobin, N. P., Mølholm, I., Kis, L., Rosin, G., Rantalainen, M., Pedersen, L., Bergh, J., Grunkin, M., and Hartman, J. (2016). Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathology*, **29**(4), 318–329.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. IEEE.

Van Cutsem, E., Labianca, R., Bodoky, G., Barone, C., Aranda, E., Nordlinger, B., Topham, C., Tabernero, J., Andre, T., Sobrero, A. F., Mini, E., Greil, R., Di Costanzo, F., Collette, L., Cisar, L., Zhang, X., Khayat, D., Bokemeyer, C., Roth, A. D., and Cunningham, D. (2009). Randomized Phase III Trial Comparing Biweekly Infusional Fluorouracil/Leucovorin Alone or With Irinotecan in the Adjuvant Treatment of Stage III Colon Cancer: PETACC-3. *Journal of Clinical Oncology*, **27**(19), 3117–3125.

van Ginneken, B., Setio, A. A. A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI*, pages 286–289. Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands, IEEE.

Vedaldi, A. and Lenc, K. (2015). MatConvNet – Convolutional Neural Networks for MATLAB. In *ACM International Conference on Multimedia*, pages 1–55.

Weigelt, B., Geyer, F. C., and Reis-Filho, J. S. (2010). Histological types of breast cancer: how special are they? *Molecular Oncology*, **4**(3), 192–208.