Meta-Analysis for Omics Datasets

bcf.isb-sib.ch

Pratyaksha "Asa" Wirapati

Bioinformatics Core Facility, Swiss Institute of Bioinformatics

www.isb-sib.ch

Bioinformatics in Genomic and Proteomic Data

November 25–27, 2009, Brno, Czech Republic

Bionformatics Core Facility
Swiss Institute of Bioinformatics

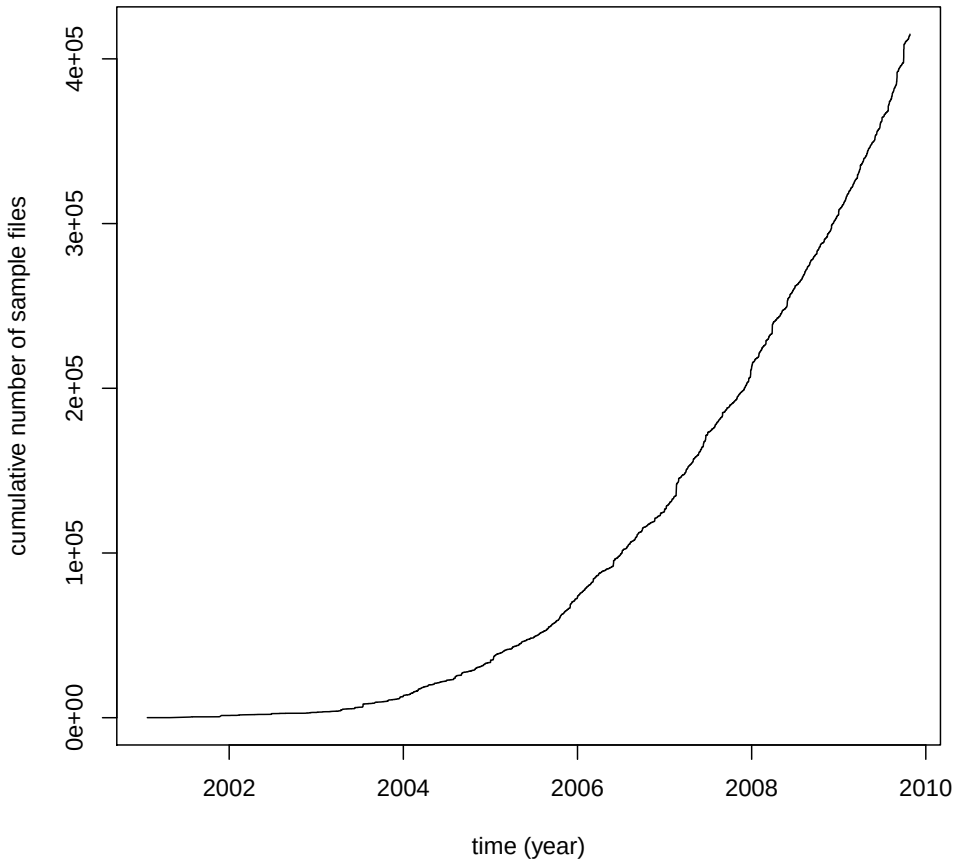
Swiss Alps French Alps Darlene Goldstein Sylvain Pradervand

Lake Geneva Evian Pierre Farmer Fred Schutz Alex Kuhn

Lausanne UNIL EPFL Vlad Popovici Jenny Miggliavacca Thierry Sengstag

Eva Budinská Asa Wirapati **Mauro Delorenzi** M. Faouzi

Growth of Gene Expression Omnibus (GEO) Database



Technology	# samples
in situ oligonucleotide	209391
spotted DNA/cDNA	76911
spotted oligonucleotide	54941
oligonucleotide beads	17013
SAGE	1660
other	1193
high-throughput sequencing	853
RT-PCR	497
spotted protein	390
antibody	337
MPSS	194
mixed spotted oligo/cDNA	109
MS	94
SARST	12

$\Downarrow$

genomics (DNA)
transcriptomics (RNA)
proteomics (protein)

$\Downarrow$

$*$omics (everything else)

Other data sources: ArrayExpress, journal suppl. data, investigator's websites

Omics Biology and Medicine

Data "supertable": studies (rows) $\times$ omics variables (columns)

		DNA			RNA			Protein		Phenotype		Environment	
		SNP	CNV, CGH	UHTS	mRNA	miRNA	SAGE	IHC	proteomics	clinical	Imaging, metabolomics, physiology	drug, therapy	pathogen, toxin
Study design 1 human breast cancer patients, retrospective, clinical outcome, drug	Study 1				■					■			
	Study 2				■	■				■		■	
	Study 3				■					■		■	
	Study 4				■					■			
	Study 5				■								
	Study 6		■		■					■			
	...				■					■			
Study design 2 experimental, time-series, tissue culture	Study a												
	Study b												
Study design 3 cancer cell lines ...	Study x	■	■	■	■	■	■	■	■		■	■	
	Study y						■						
	Study z						■						
	...												

"Horizontal integration": same samples, various omis variables

"Vertical integration": similar variables, multiple studies $\Rightarrow$ our focus

Why re-analyze existing datasets?

- Critical review of the original findings
- Confirmation/validation of results from other studies
- More solid discoveries based on larger sample size
- New discoveries in larger scopes/contexts

Issues in Co-Analysis of Multiple Datasets

I. Dataset curation

- Survey of relevant datasets that are available

  Search literature, public databases, and the web

- Independence of datasets

  Reorganize datasets to ensure non-redundant samples

- Non-uniform variable names and representation

  Rename and recode variables

- Re-mapping probe(set)s and matching across platforms

  Align to a reference sequence database; reduce to single probe per gene

- Quality control of quantitative variables (e.g., gene expression)

  Ensure same unit/transformation; renormalize and rescale if necessary

Issues in Co-Analysis of Multiple Datasets

II. Downstream Analysis

How to do combined analysis of heterogeneous datasets?

- Differences in study designs, populations and sample selection criteria
- Incommensurable quantitative data; systematic measurement artefacts

How to produce the "total" results based on all datasets?

How to assess and incorporate heterogeneity?

How to visualize and present the analysis results?

How to adapt to omics data?

How to adapt to complex analysis, such as hierarchical clustering and prediction?

Outline

- A brief introduction to statistical meta-analysis
- Applications of meta-analysis to omics data
  - An example: breast cancer clinical-expression datasets
  - Differential expression
  - Clustering of genes
  - Clustering of samples
  - Prediction
- Conclusion and future works

Intro to meta-analysis: an example data

UC Berkeley graduate school admission 1973[1]

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Was there a sex bias in the graduate school admission process?

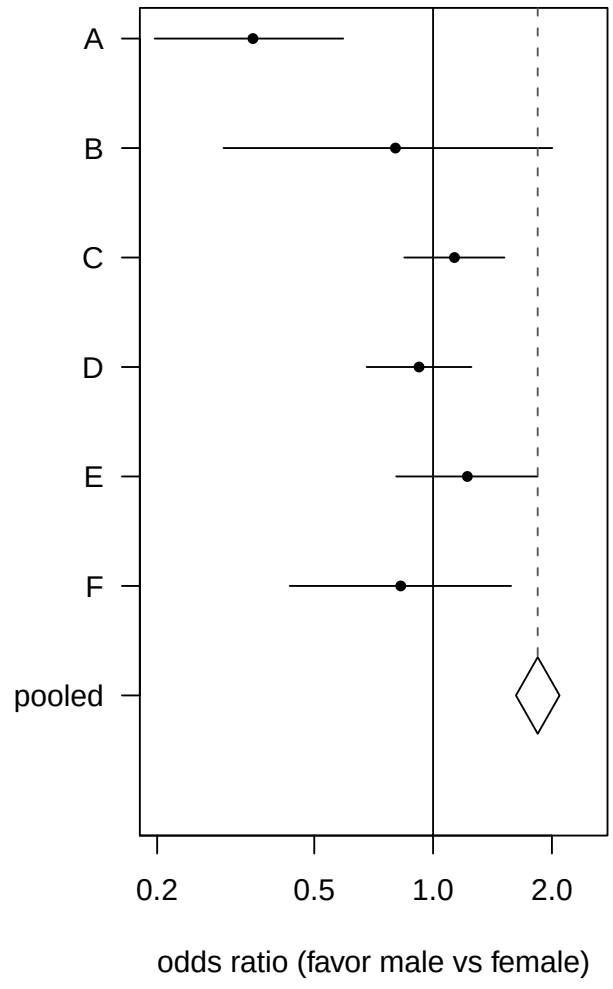
odds ratio: $\frac{1278/557}{1493/1198} = 1.84$, 95% CI: [1.62, 2.09]

p-value: $< 2.2 \times 10^{-16}$

[1] Bickel, Hammel, O'Connell (1975) *Science* **187**:398-403

Stratified Analysis and Forest Plot

Dept.	data	odds ratio	95% C.I.	p-value
A	512 | 89 313 | 19	0.35	[0.20, 0.59]	10^{-5}
B	353 | 17 207 | 8	0.80	[0.30, 0.20]	0.68
C	129 | 202 205 | 391	1.13	[0.84, 1.52]	0.39
D	138 | 131 279 | 244	0.92	[0.68, 1.25]	0.60
E	53 | 94 138 | 299	1.22	[0.80, 1.83]	0.36
F	22 | 24 351 | 317	0.83	[0.43, 1.58]	0.55
pooled	1198 | 557 1439 | 1278	1.84	[1.62, 2.09]	10^{-16}



Simpson's Paradox: "the whole contradicts its parts"
the danger of pooling data $\Rightarrow$ biases due to hidden factors

Meta-Analytical Solution

- Analyze each stratum/study separately
- Average using the inverse variance as weight

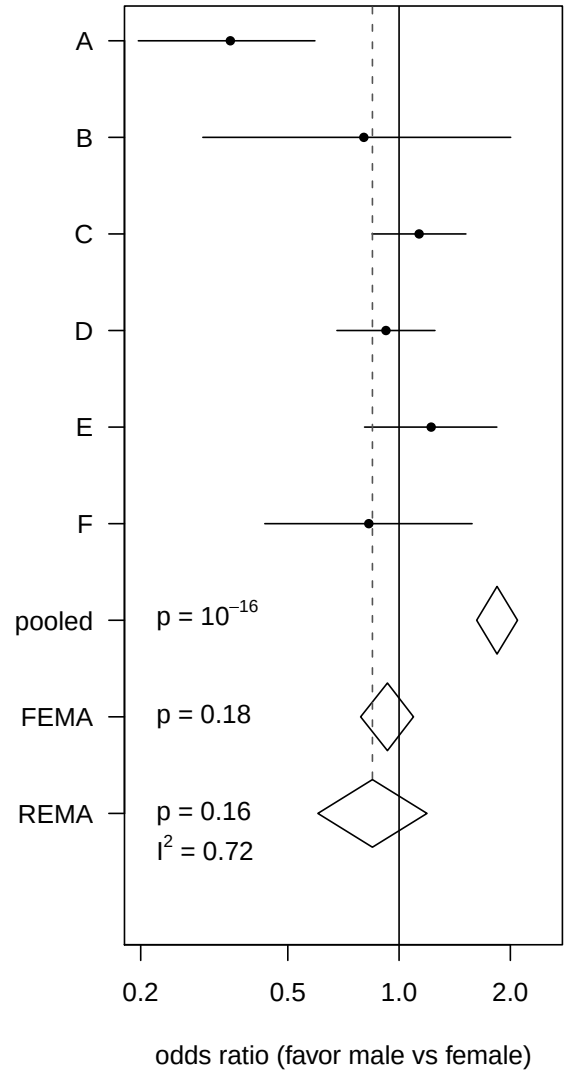
$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} \hat{\beta}_i/(\hat{\sigma}_i^2 + \hat{\tau}^2)}{\sum_{i=1}^{k} 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)}$$

β_i, β_0: effect size (per study and total)

σ_i^2: within-study variance of β_i, i.e. $[\text{SE}(\beta_i)]^2$

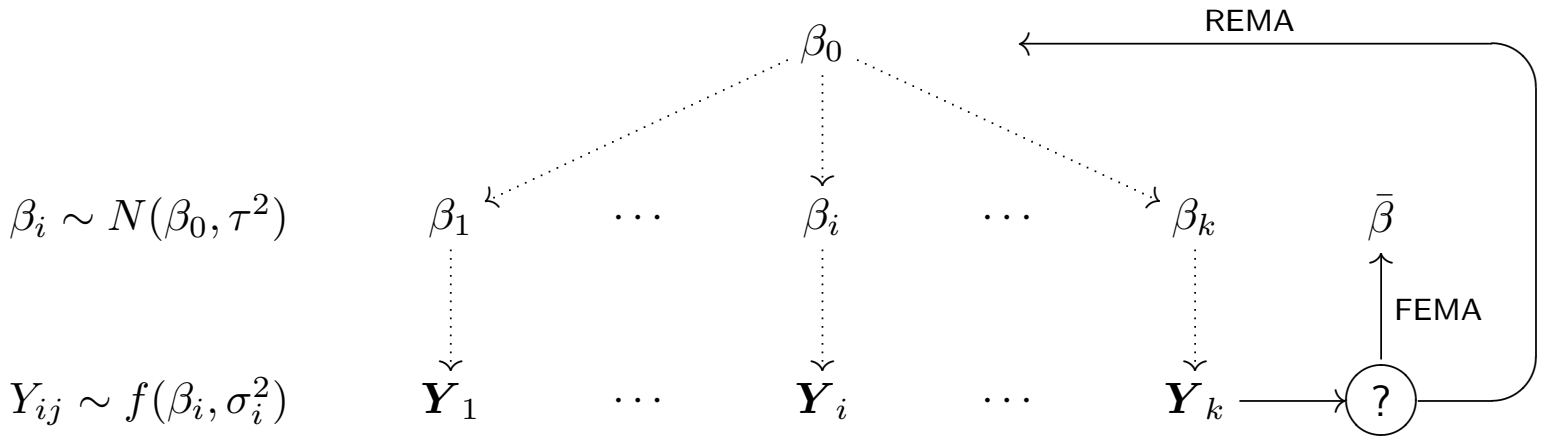
$\hat{\tau}^2$: between-study variance

- If τ^2 is fixed to zero (may not be realistic!)
  $\Rightarrow$ fixed effects meta analysis (FEMA)
- If τ^2 is estimated from the data
  $\Rightarrow$ random effects meta analysis (REMA)
- I^2: proportion of variation due to between study heterogeneity



$I^2 = (Q - (k-1))/Q$, see Higgins & Thompson (2002) *Stat Med*

Hierarchical Sampling Models



Single study:

- Inference about β_i (β_0 + study biases: technical, design, population, ...)

Fixed-effect models

- Inference about $\bar{\beta} = \sum_i \beta_i / k$ (the mean of the specific datasets in hand)
- Confidence interval is not affected by between study variability τ^2

Random-effect/hierarchical models

- Inference about β_0 (the "truth"; expectation of future studies)
- Confidence interval is small if I^2 is small (and vice versa)

Alternative Methods

- (Empirical) Bayes Hierarchical Models

  This is the theoretically "proper" way to hierarchical models

  More flexible than REMA (not limited to normal summaries)

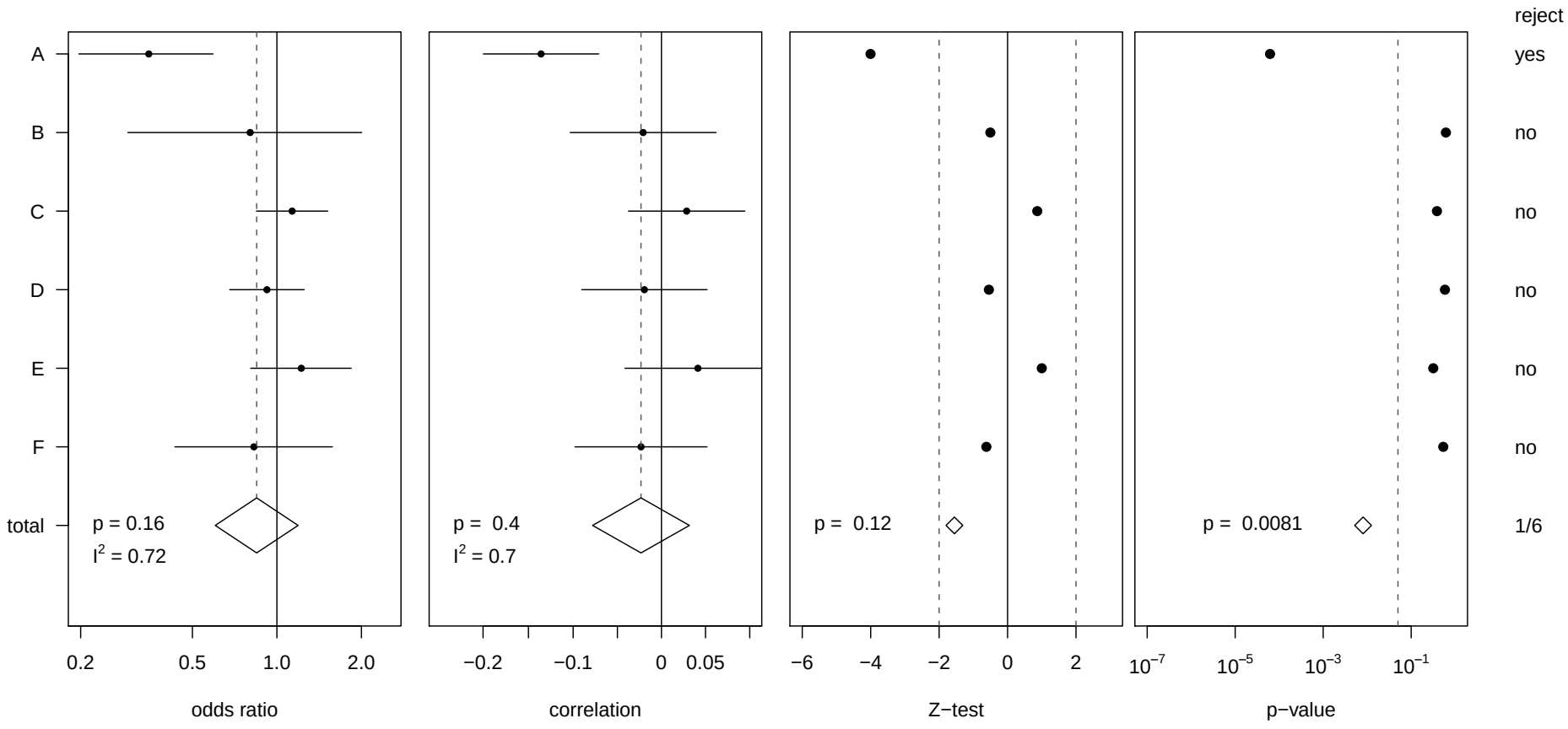
  Simultaneous fitting of model parameters at all levels of hierarchy (while REMA is stage-wise).

  Computationally more expensive (need to maximize marginal likelihood via EM, or MCMC, or quadrature, etc. etc.)

  REMA is an approximate approach to hierarchical models (may even be equivalent in some cases), but easier to calculate. Compromise: maybe less optimal for large number of very small studies.

- For categorical explanatory variables (e.g. ANOVA or contingency tables), the study indicator can be treated as another term, and the heterogeneity is modelled as interaction terms.

Which summary to combine?



odds ratio: regression coefficient (average using REMA)
correlation: measure of dependence or mutual information (average using REMA)
Z-test: significance (signed) $\Rightarrow$ *accumulate* using Stouffer's method: $\sum Z/sqrt(k)$
p-value: significance (unsigned) $\Rightarrow$ *accumulate* using Fisher's method: $-2\sum \log p$
vote counting method: count rejected null hypothesis

Spectrum of possibilities in combining analysis

1. Combine raw data

   (+)easy to apply (–) potential bias, no heterogeneity assessment

2. Combine coefficients (fold change, hazard and odd ratios, . . .)

   (+)physical interpretability (–) affected by measurement unit

3. Combine correlation/dependence (R^2, $\tanh^{-1}(r)$,. . .)

   (+)unit-free (–) affected by sampling/design

4. Combine significance measures (t-test, Z-test, p-value, etc.)

   (–) strong effect + low power = weak effect + high power

5. Combine decisions (reject/accept hypothesis, gene lists)

   (+) easy to apply (–) lacks power

Outline

- A brief introduction to statistical meta-analysis
- Applications of meta-analysis to omics data
  - An example: breast cancer clinical-expression datasets
  - Differential expression
  - Clustering of genes
  - Clustering of samples
  - Prediction
- Conclusion and future works

Breast cancer data collection

Susanne Kunkel

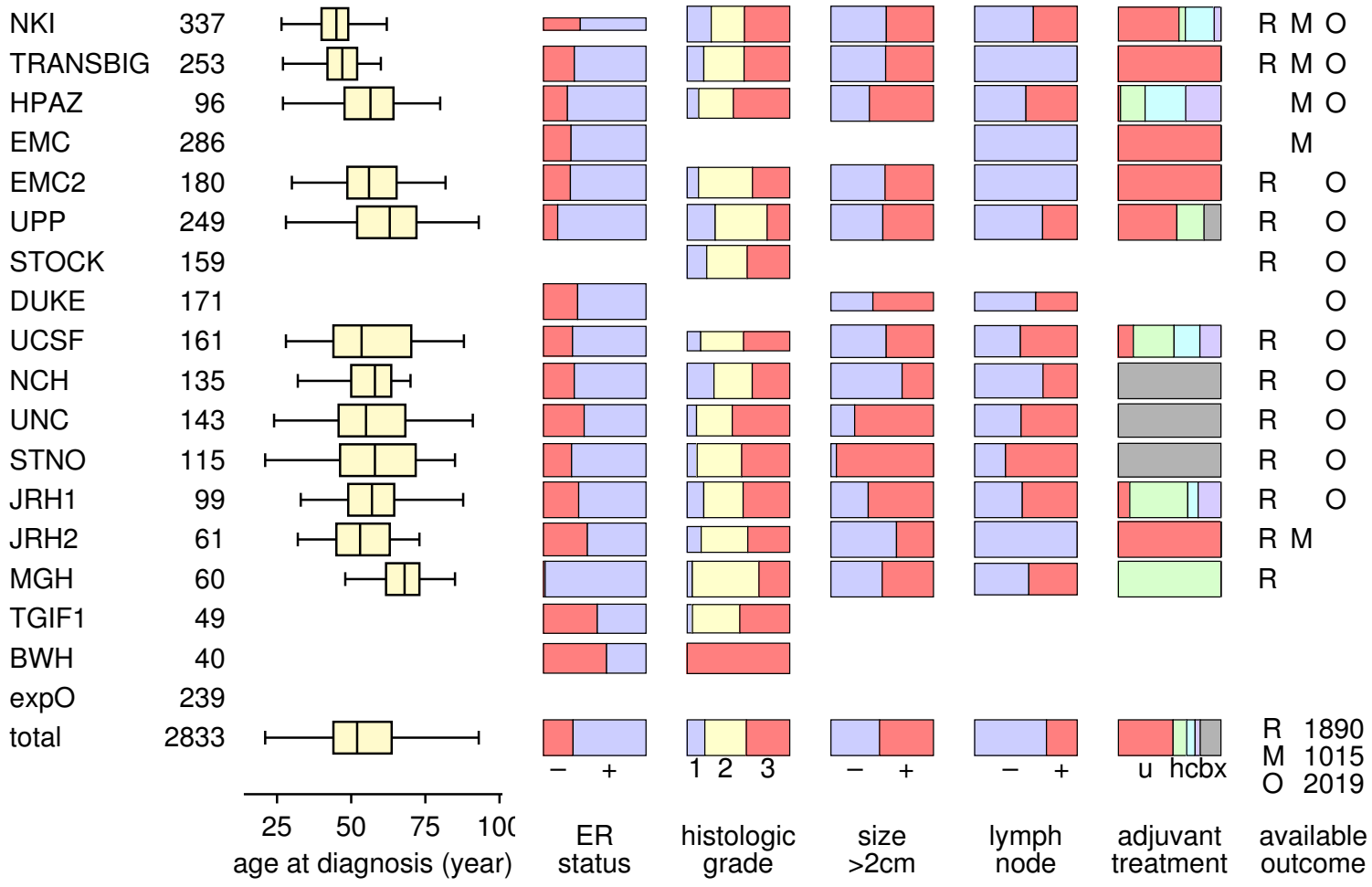
Wirapati *et. al.* 2008 *Breast Cancer Res*

Dataset symbol	No. of arrays	Institution	Reference	Platform	Data source	No. of GeneIDs
Genomic platforms						
NKI	337	Nederlands Kanker Instituut	van't Veer 2002, van de Vijver 2002	Agilent	author's website	13120
EMC	286	Erasmus Medical Center	Wang 2005	Aff. U133A	GEO:GSE2034	11837
UPP	249	Karolinksa Institute (Uppsala)	Miller 2005, Calza 2006	Aff. U133A,B	GEO:GSE4922	15684
STOCK	159	Karolinska Institute (Stockholm)	Pawitan 2005, Calza 2006	Aff. U133A,B	GEO:GSE1456	15684
DUKE	171	Duke University	Huang 2005, Bild 2006	Aff. U95Av2	author's website	8149
UCSF	161+8	UC San Francisco	Korkola 2003	cDNA	author's website	6178
UNC	143+10	University of Carolina	Hu 2006	Agilent HuA1	author's website	13784
NCH	135	Nottingham City Hospital	Naderi 2006	Agilent HuA1	AE:E-UCON-1	13784
STNO	115+7	Stanford Univ./Norwegian Radium Hosp.	Sorlie 2003	cDNA	author's website	5614
JRH1	99	John Radcliffe Hospital	Sotiriou 2003	cDNA	journal's website	4112
JRH2	61	John Radcliffe Hospital	Sotiriou 2006	Aff. U133A	GEO:GSE2990	11837
MGH	60	Massachusetts General Hospital	Ma 2004	Agilent	GEO:GSE1379	11421
expO	239	International Genomic Consortium	http://www.intgen.org	Aff. U133v2	GEO:GSE2109	16634
TGIF1	49	EORTC trial 10994	Farmer 2005	Aff. U133A	GEO:GSE1561	11837
BWH	40+7	Brigham and Women's Hospital	Richardson 2006	Aff. U133v2	GEO:GSE3744	16634
Small diagnostic platforms						
TRANSBIG	253	TRANSBIG Consortium	Buyse 2006	Agilent	AE:E-TABM-77	1052
EMC2	180	Erasmus Medical Center	Foekens 2006	Aff. (custom)	GSE3453	86
HPAZ	96	Hospital La Paz, Madrid	Espinosa 2005	RT-PCR	paper's appendix	61
Total	2865	= 2833 carcinomas			No. of the union of all GeneIDs:	17198
		+ 32 non-malignant breast tissues			No. of GeneIDs common to genomic platforms:	1963

• Abbreviations: No. = number, GEO: = Gene Expression Omnibus accession, AE: = ArrayExpress accession, Aff. = Affymetrix

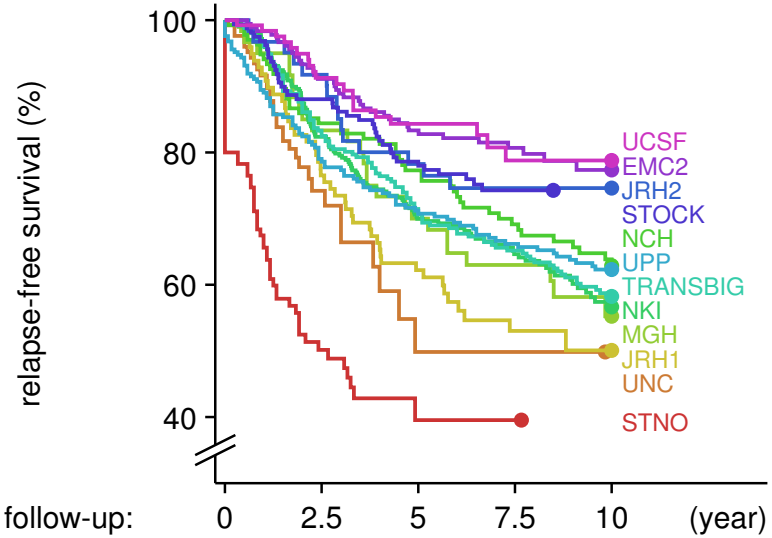
- Reorganize datasets into independent, non-redundant cohorts
- Remap probe(set)s to the same version of RefSeq subset (NM_* only) using BLAT
- Use the most variable probe(set) as the unique representative of a gene

Clinical variable availability and distributions



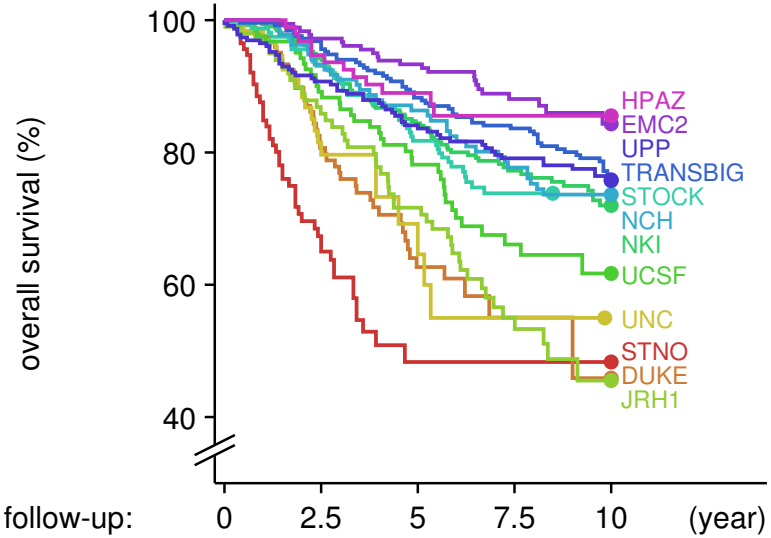
treatment: u untreated, h hormone, c chemo, b both, x unspecified
patient outcome: R relapse-free, M metastasis-free, O overall survival

Heterogeneity in survival data



follow-up: 0 2.5 5 7.5 10 (year)

group:	number at risk:					events
STNO	115	39	10	1		60
UNC	128	33	10	4		32
JRH1	99	75	59	30	1	45
MGH	60	50	42	28	18	25
NCH	135	110	97	81	66	47
NKI	319	260	216	131	75	121
TRANSBIG	253	207	170	147	118	101
UPP	249	185	158	140	107	88
JRH2	61	55	44	38	33	15
STOCK	159	140	124	68		40
EMC2	180	164	149	94	40	37
UCSF	132	97	68	37	11	20
total	1890	1415	1147	799	469	631



follow-up: 0 2.5 5 7.5 10 (year)

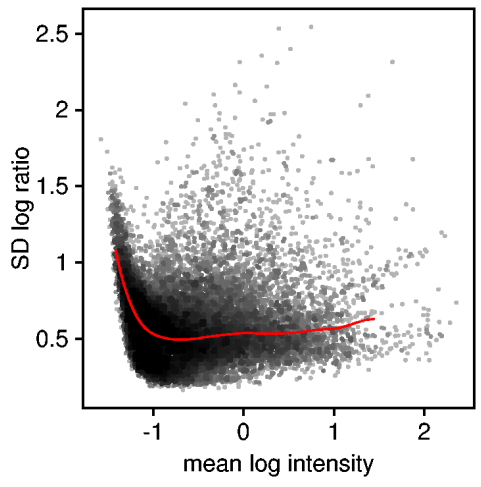
group:	number at risk:					events
STNO	115	54	14	3	1	46
DUKE	170	86	46	12	3	43
UNC	129	39	14	5		22
JRH1	99	85	70	32	1	45
UCSF	132	104	74	43	14	37
NKI	319	290	248	147	89	74
STOCK	159	148	130	64		40
NCH	135	122	111	96	81	34
TRANSBIG	253	240	212	190	154	57
UPP	232	198	173	152	122	51
EMC2	180	175	166	103	44	23
HPAZ	96	87	55	20	3	12
total	2019	1628	1313	867	512	484

Variability between studies greater than that due to natural risk factors or treatments $\Rightarrow$ potential bias in pooled (unstratified) analysis

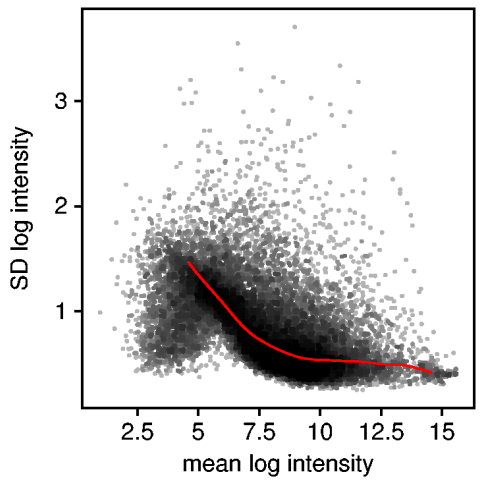
Quality control of original author's normalization

Plot SD-vs-mean of each probe in a dataset
$\Rightarrow$ A characteristic trend for each (platform,normalization) combination

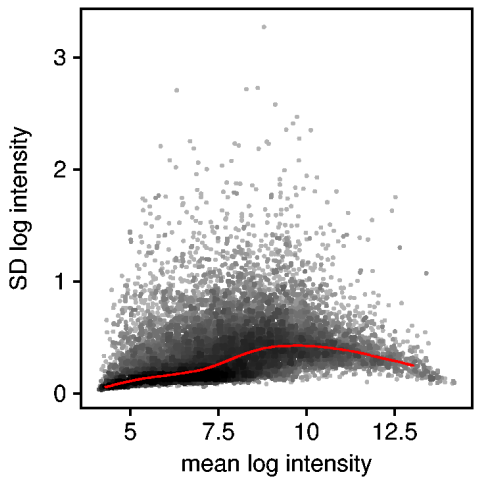
NKI: Agilent (Rosetta) + ?



EMC: Affy U133 + MAS 5.0



TGIF1: Affy U133A + RMA



Raw instrument data (e.g. CEL files) for renormalization from scratch are not always available $\Rightarrow$ possible "post-hoc" corrections:

- Non-parametric variance stabilizing transform
- Global scaling between studies
- Lowess calibration against the mean profile

(In subsequent results in this talk, we used the original without correction)

Differential Expression Analysis

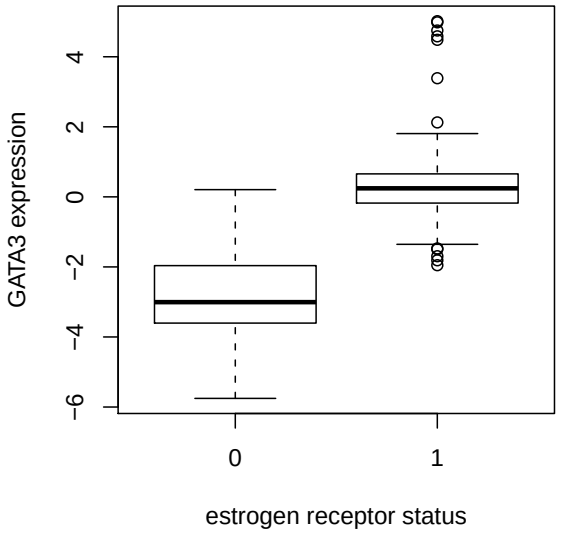
The transcriptome is "scanned" to search for genes whose change in expression is related to changes in other variables (e.g. clinical outcome or experimental conditions)

Adaptation for multiple datasets:

1. Choose the appropriate models that produce an estimate $\pm$ standard error (with normal sampling variation, independent of the location estimate) transformation may be used when appropriate
2. If a gene is missing from a platform, the summary is considered missing value (and simply ignored)
3. Calculate REMA (estimate, SE, heterogeneity)
4. The usual analysis: ranking, multiple testing, etc. on the combined estimates from REMA

Generalized Linear Models

normal

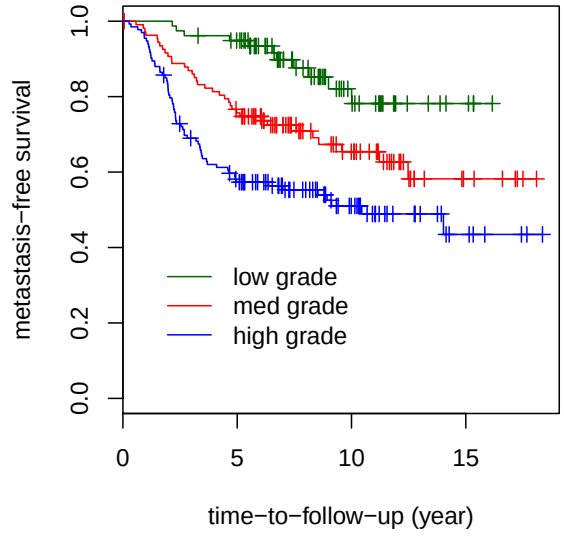


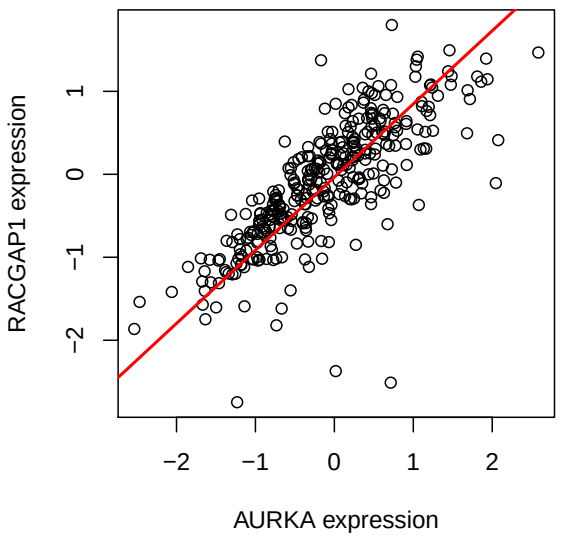
logistic

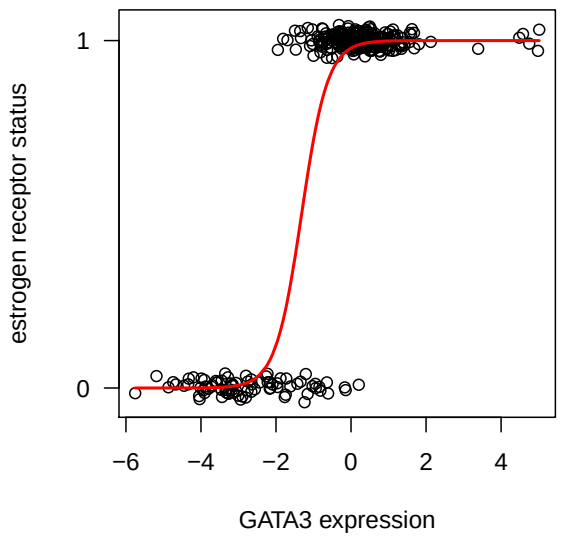
histologic grade

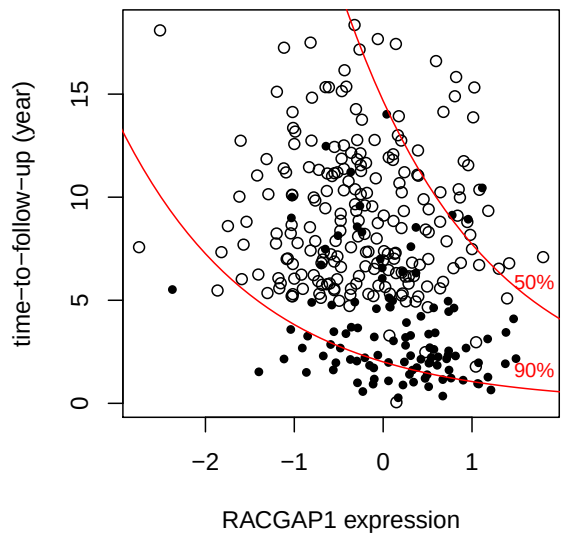
ER		low	med	high
	0	4	11	66
	1	75	98	83

survival



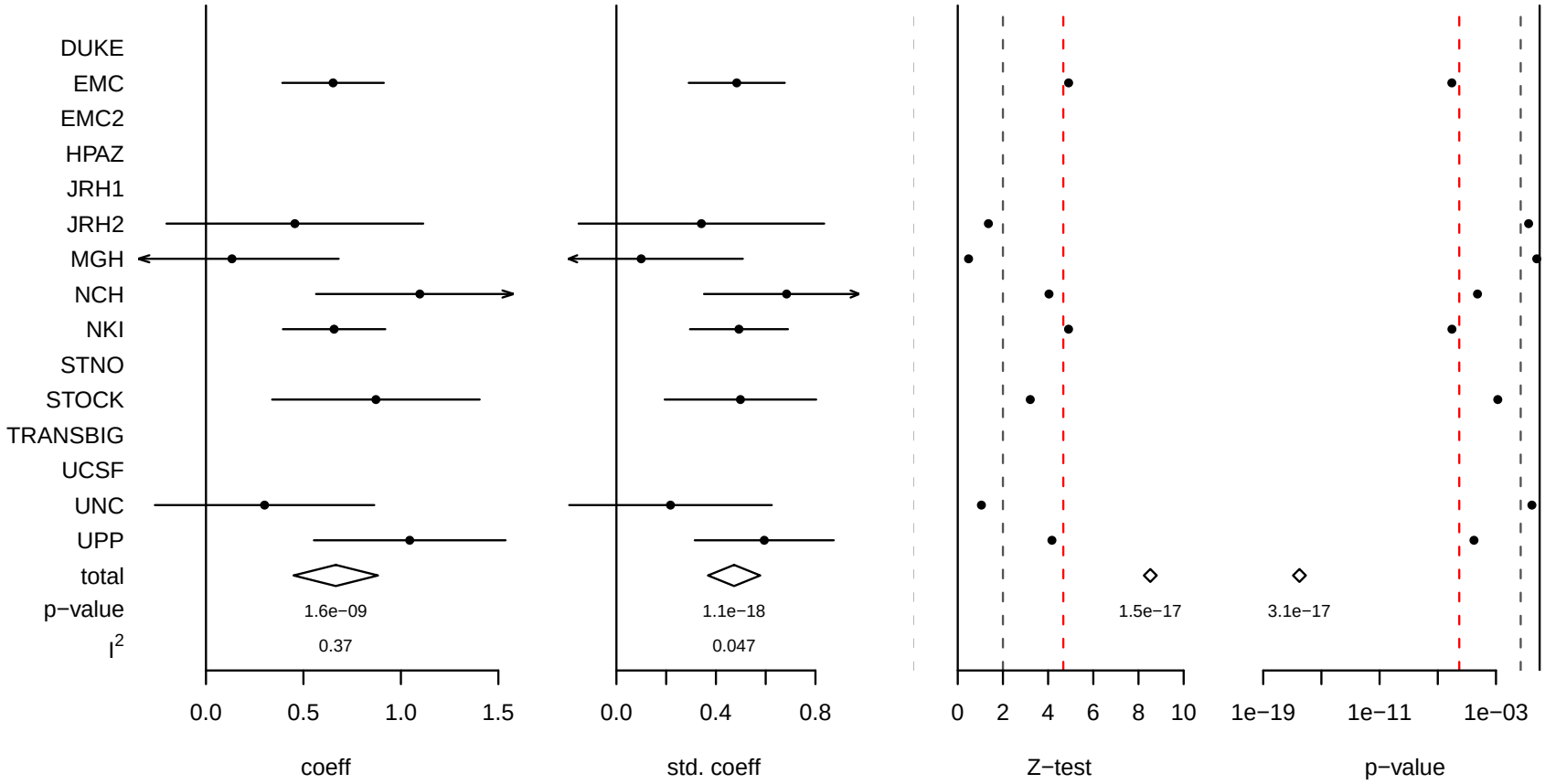






An example: prognostic genes in breast cancer

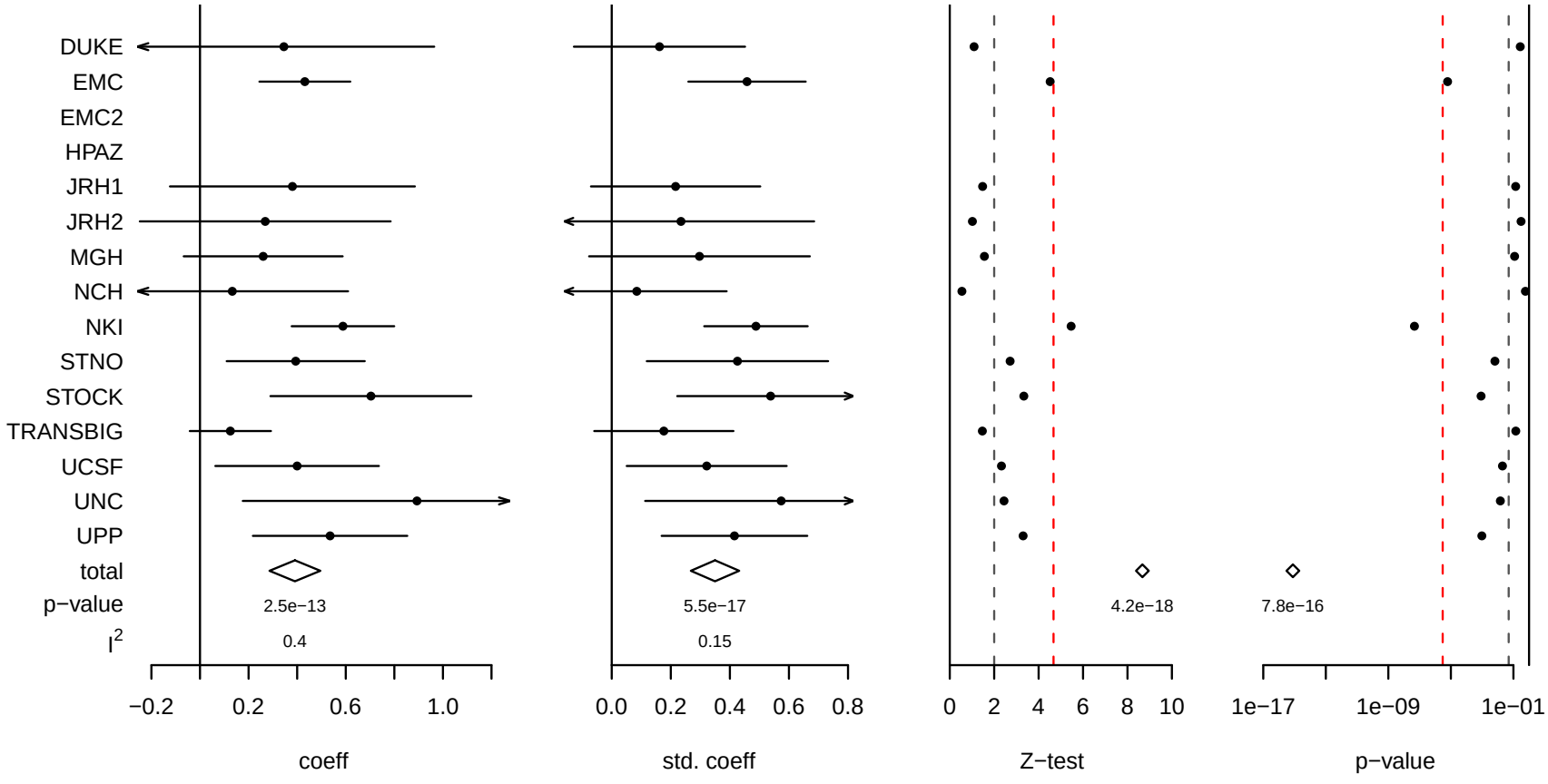
Gene: RACGAP1; Model: Cox proportional hazard
Response variable: metastasis-free survival; explanatory variable: $\log_2$ expression



coeff: $\log_e$(hazard change)/$\log_2$(fold change) $\Rightarrow$ effect size with physical interpretation
std. coeff: measure of correlation (mutual information), equivalent to (pseudo) R^2
Z-test: significance, equivalent to p-value, but with direction of effect $(-/+)$
Only significant (after multiple testing) in two studies

Another example

gene: AURKA

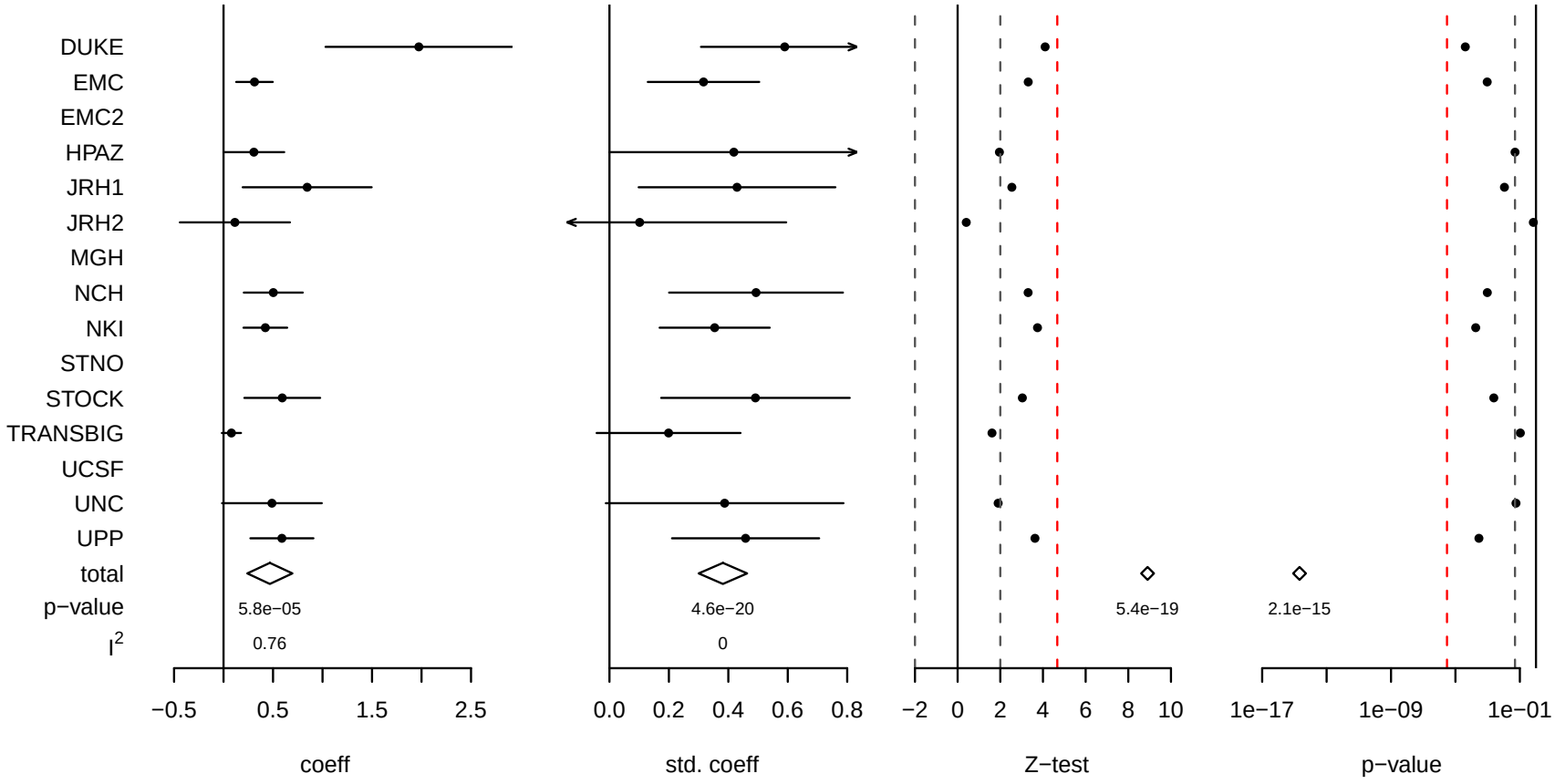


Coefficients are less heterogeneous than in RACGAP1

Present in all genome-wide platforms

Another example

gene: MELK

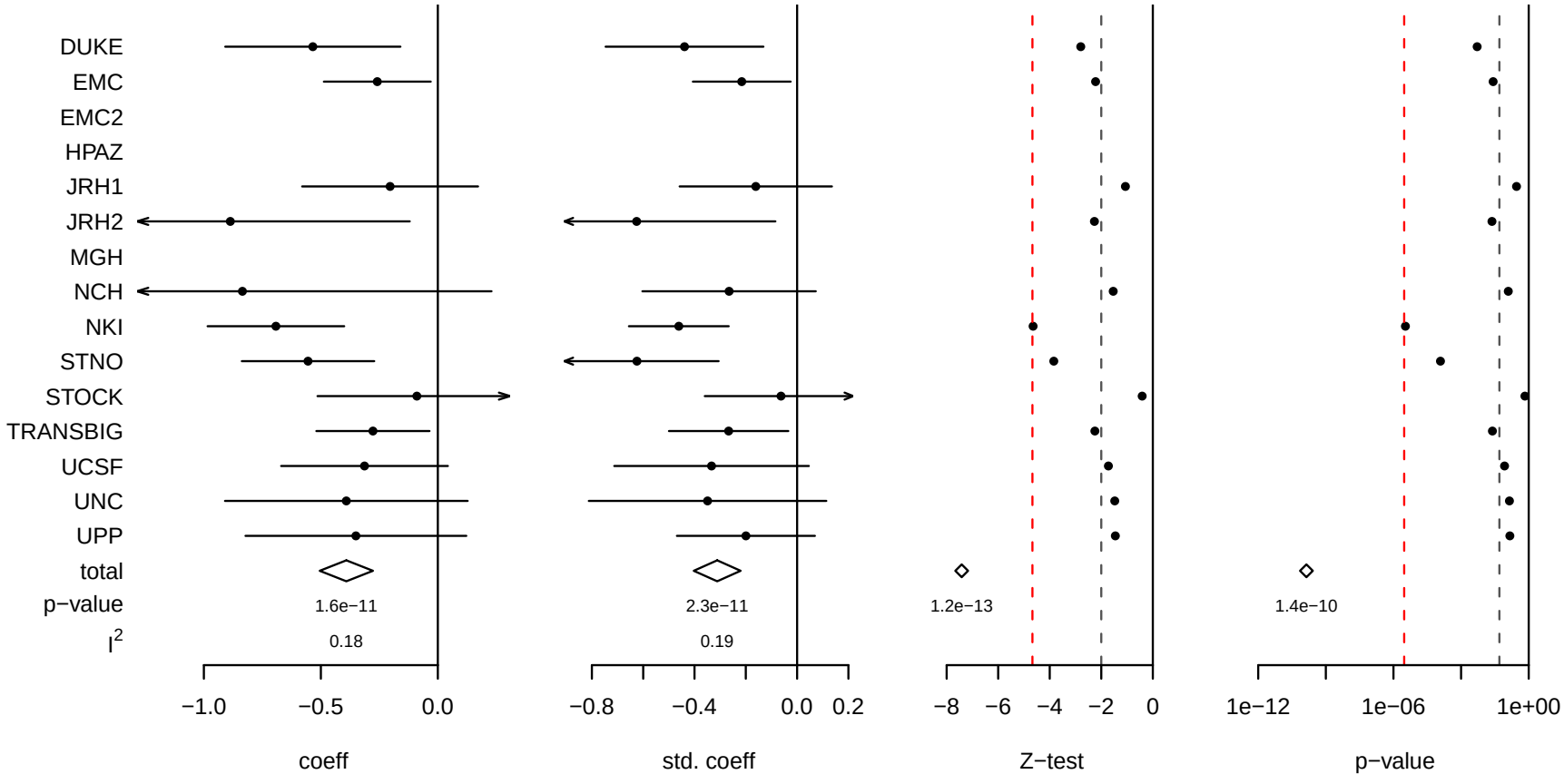


Coefficients are heterogeneous; correlation (std. coeff) is homogeneous
$\Rightarrow$ normalization issue? or the $\log_2$ scale is less consistent in general?

Not significant in individual studies

Another example

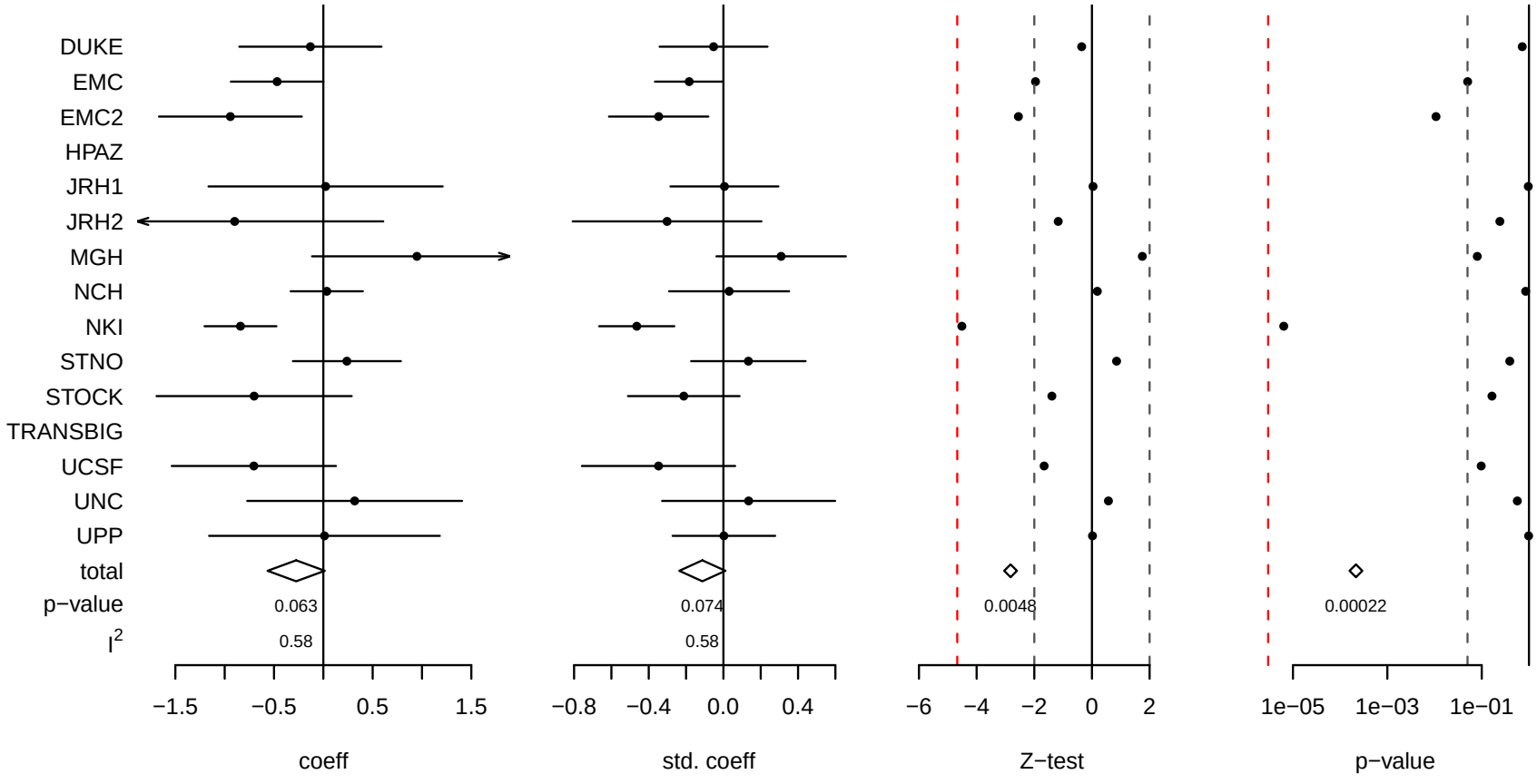
gene: BTG2



Negative effects (over-expression is protective)

Yet Another Example

gene: RPL11



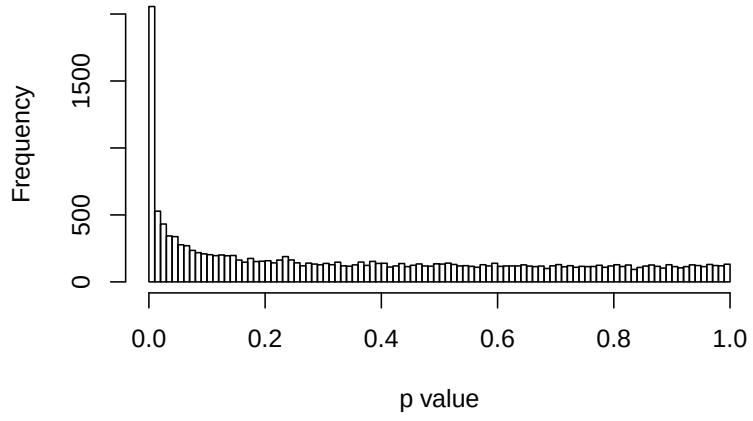
A gene that doesn't work. (It's a housekeeping gene)

The Usual Analysis and Visualization

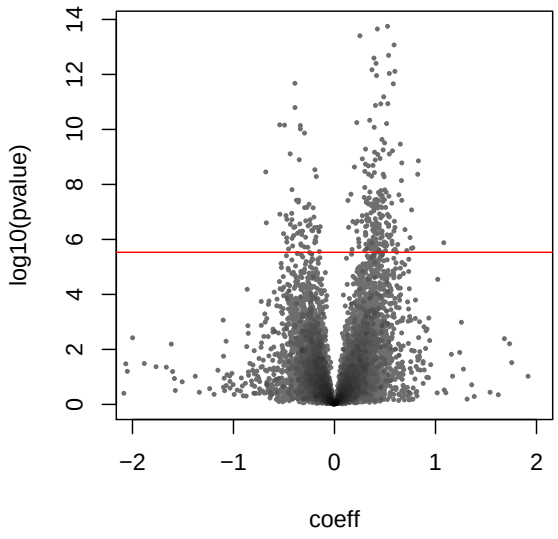
Gene rank table

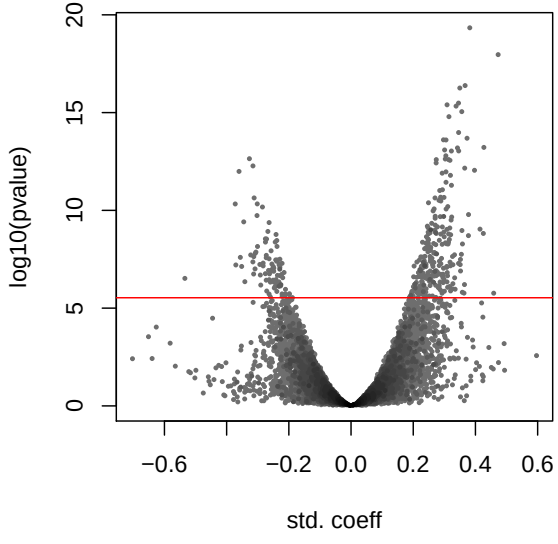
	est	se	Z	pval	p.bonf
SEC61G	0.5252896	0.06852138	7.666069	1.773481e-14	2.963486e-10
CEP55	0.4241852	0.05554382	7.636946	2.224340e-14	3.716872e-10
BIRC5	0.2513234	0.03322773	7.563666	3.918662e-14	6.548084e-10
PSMA7	0.5896986	0.07901168	7.463436	8.429511e-14	1.408571e-09
NP	0.5357213	0.07291376	7.347327	2.022091e-13	3.378915e-09
AURKA	0.3907769	0.05340849	7.316757	2.540361e-13	4.244944e-09
NEK2	0.4112018	0.05666095	7.257236	3.950808e-13	6.601800e-09
UBE2S	0.3708391	0.05161736	7.184387	6.750934e-13	1.128081e-08
PSMD2	0.5975764	0.08338927	7.166107	7.716040e-13	1.289350e-08
TCEB1	0.5424997	0.07595975	7.141937	9.202507e-13	1.537739e-08
SPAG5	0.4161139	0.05846667	7.117114	1.102106e-12	1.841618e-08
P4HA2	0.5822613	0.08292219	7.021779	2.190609e-12	3.660507e-08
GARS	0.4871429	0.07092937	6.867999	6.510866e-12	1.087966e-07
TXNRD1	0.5284003	0.07786935	6.785729	1.155019e-11	1.930036e-07
MYBL2	0.4579217	0.06750750	6.783271	1.174851e-11	1.963175e-07
GINS2	0.4053210	0.05991814	6.764579	1.336972e-11	2.234081e-07
ADFP	0.3487663	0.05298368	6.582524	4.625270e-11	7.728826e-07
NDRG1	0.2208146	0.03369460	6.553412	5.623725e-11	9.397245e-07
RAD51	0.5155052	0.07881440	6.540749	6.121145e-11	1.022843e-06
SHCBP1	0.3931051	0.06053550	6.493795	8.370070e-11	1.398639e-06
CDK2AP1	0.4698637	0.07412179	6.339076	2.311474e-10	3.862472e-06
C20orf24	0.4956172	0.07873671	6.294614	3.081649e-10	5.149436e-06
DDX39	0.6519741	0.10384654	6.278245	3.424157e-10	5.721766e-06
TGFBI	0.3072691	0.04945128	6.213572	5.179349e-10	8.654693e-06
ZWINT	0.4816099	0.07764377	6.202815	5.546219e-10	9.267732e-06

p-value histogram



Volcano plots





Many significant genes even after the stringent Bonferroni multiple testing correction for $>$17,000 genes (red lines, p.bonff $= 0.05$)

Standardized coefficients yield more significant genes (≈ 400 vs ≈ 300)

Hierarchical Clustering of Genes

1. Calculate Pearson correlation r_{ijk} for each pair of gene (i, j) in each study k

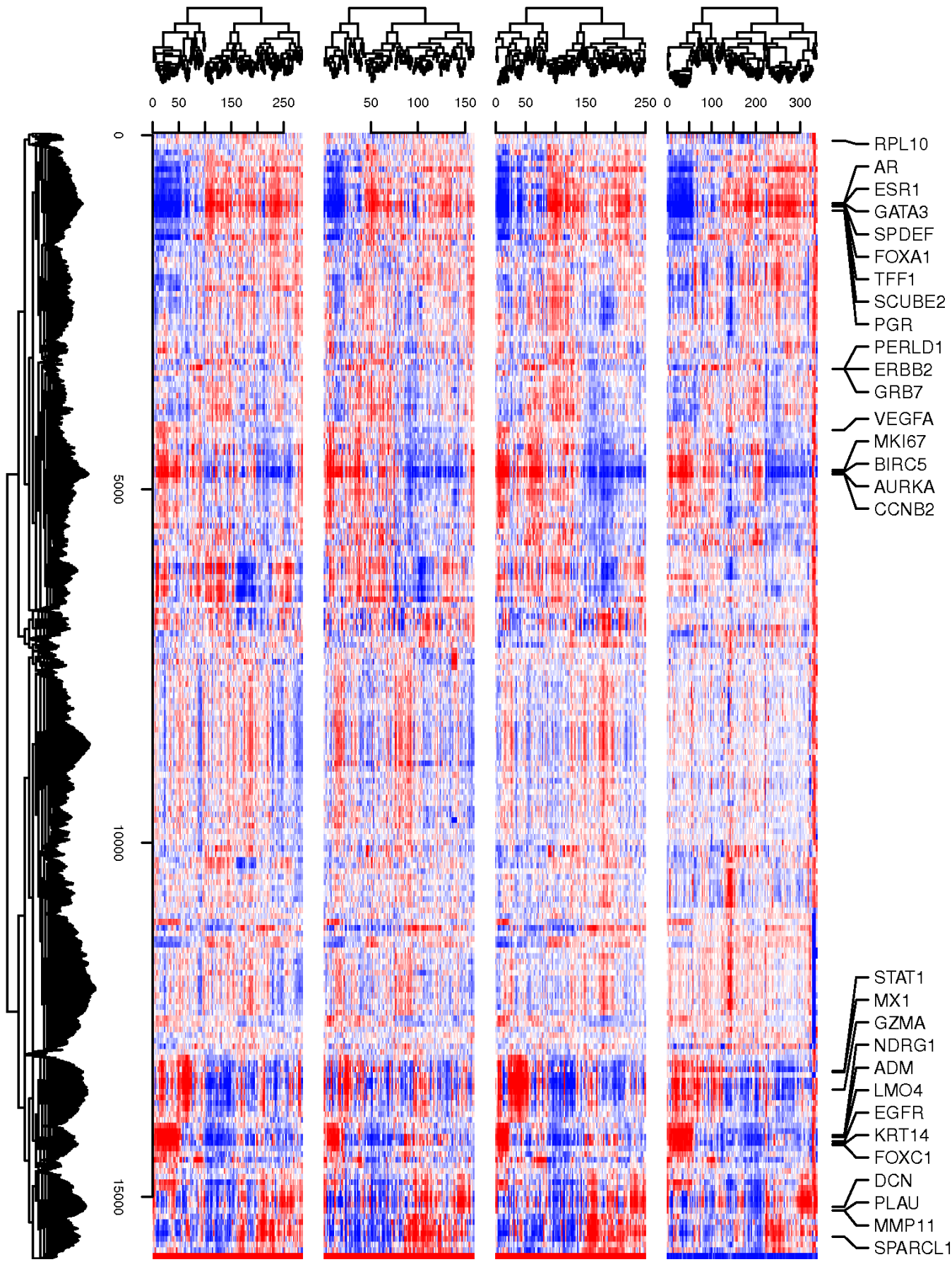
2. r isn't normal (bounded by $[-1, 1]$, asymmetric variance)
   $\Rightarrow$ transform using (yet another) Fisher's method:

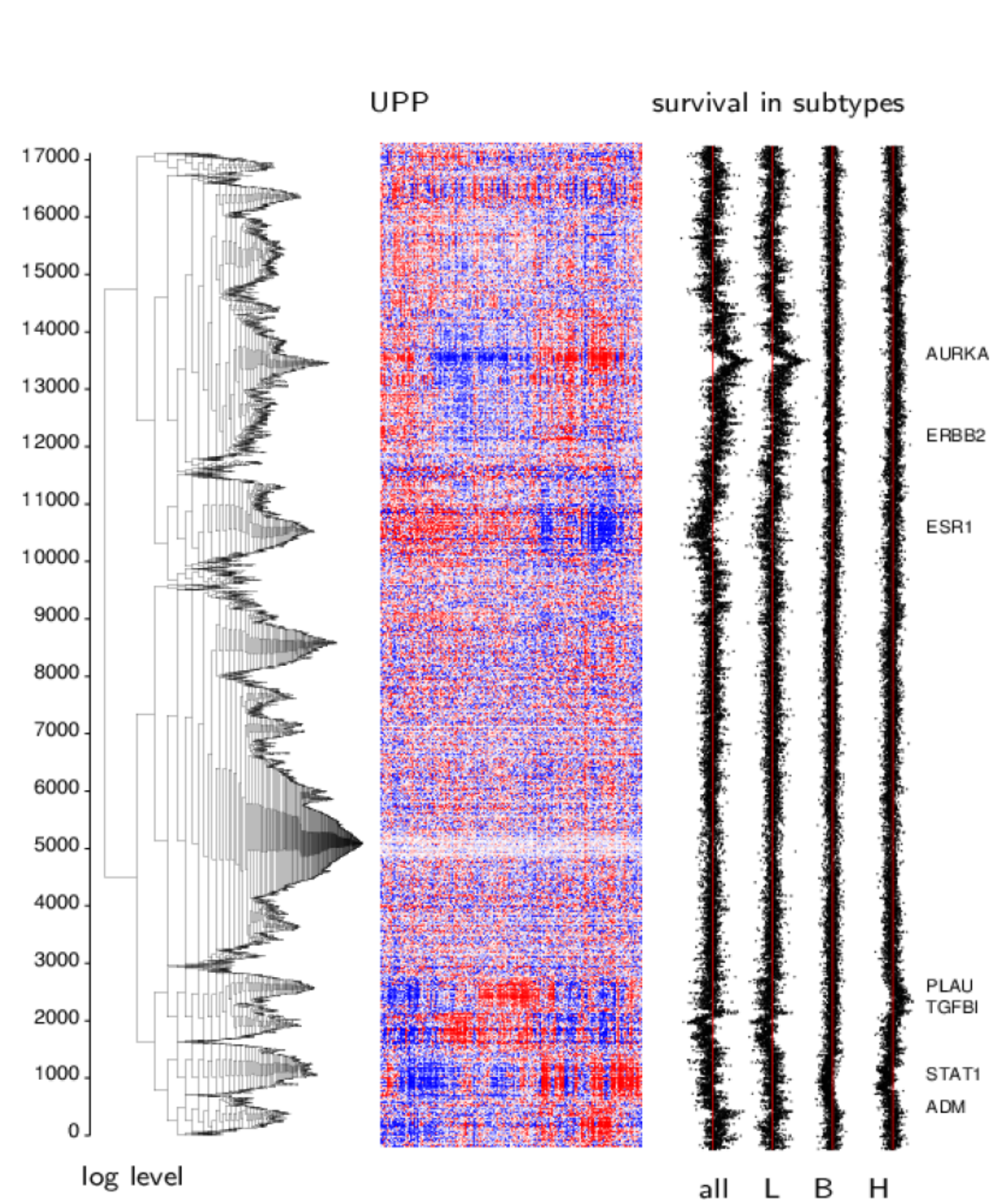
$$z_{ijk} = \tanh^{-1}(r_{ijk}), \qquad \text{Var}(z_{ijk}) = 1/(n-3)$$

3. Combine z using REMA

4. Treat the combined correlations as similarity measures in hierarchical agglomerative clustering. No need to back transform z_{ij0} to r_{ij0} (irrelevant for single- and complete link, maybe even better for average link)

5. Display the heatmaps in stratified manner


0 50 150 250
50 100 150
0 50 150 250
0 100 200 300
0
5000
10000
15000
RPL10
AR
ESR1
GATA3
SPDEF
FOXA1
TFF1
SCUBE2
PGR
PERLD1
ERBB2
GRB7
VEGFA
MKI67
BIRC5
AURKA
CCNB2
STAT1
MX1
GZMA
NDRG1
ADM
LMO4
EGFR
KRT14
FOXC1
DCN
PLAU
MMP11
SPARCL1


UPP
survival in subtypes
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
log level
AURKA
ERBB2
ESR1
PLAU
TGFBI
STAT1
ADM
all
L
B
H

Hierarchical Clustering of Samples

This doesn't fit the framework of REMA.

(Dis)similarity measures are not summary statistic from a regression model, rather it is a kind of a distance.

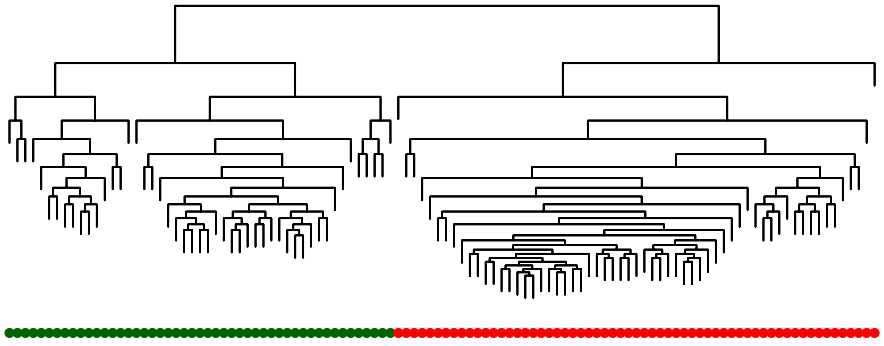
We need to have separate clustering tree for each study, but we need to know the correspondence across studies.

Pooling the data is inevitable. Expression profiles will be compared between and within studies.

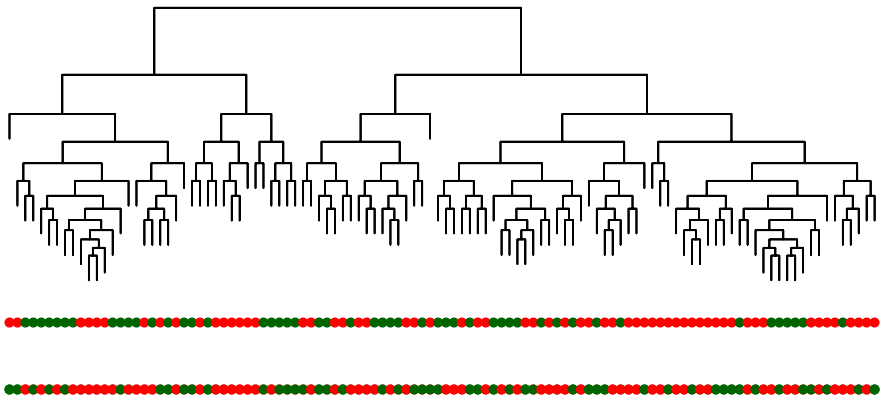
The problem: how to ensure the similarity measures are biological (rather than technical, e.g. due to batch effect), which will results clustering by the data of origin.

Simplest solution: mean center each gene for each dataset before clustering

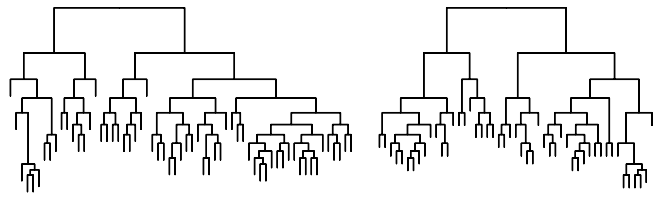
without mean centering



with mean centering



stratify by splitting the tree

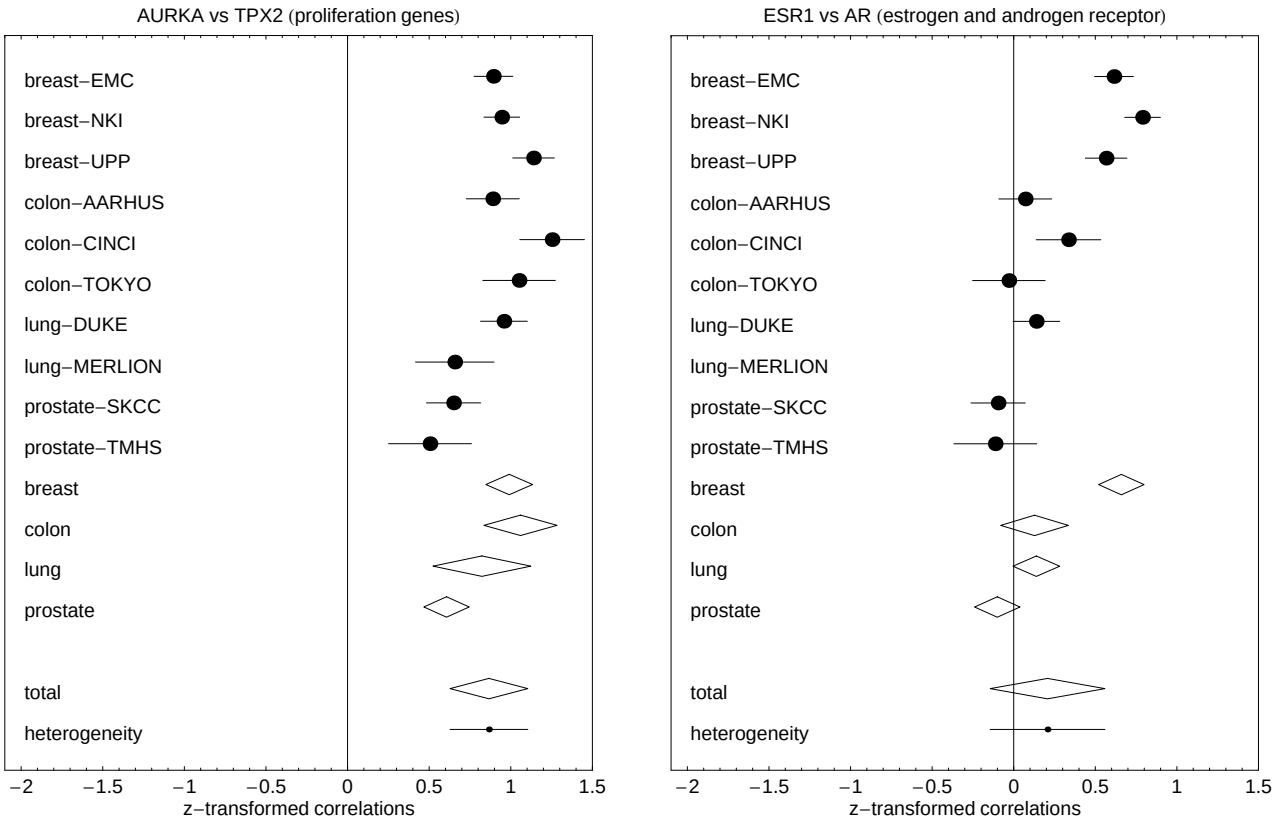


Extension to Multilevel Gene Clustering

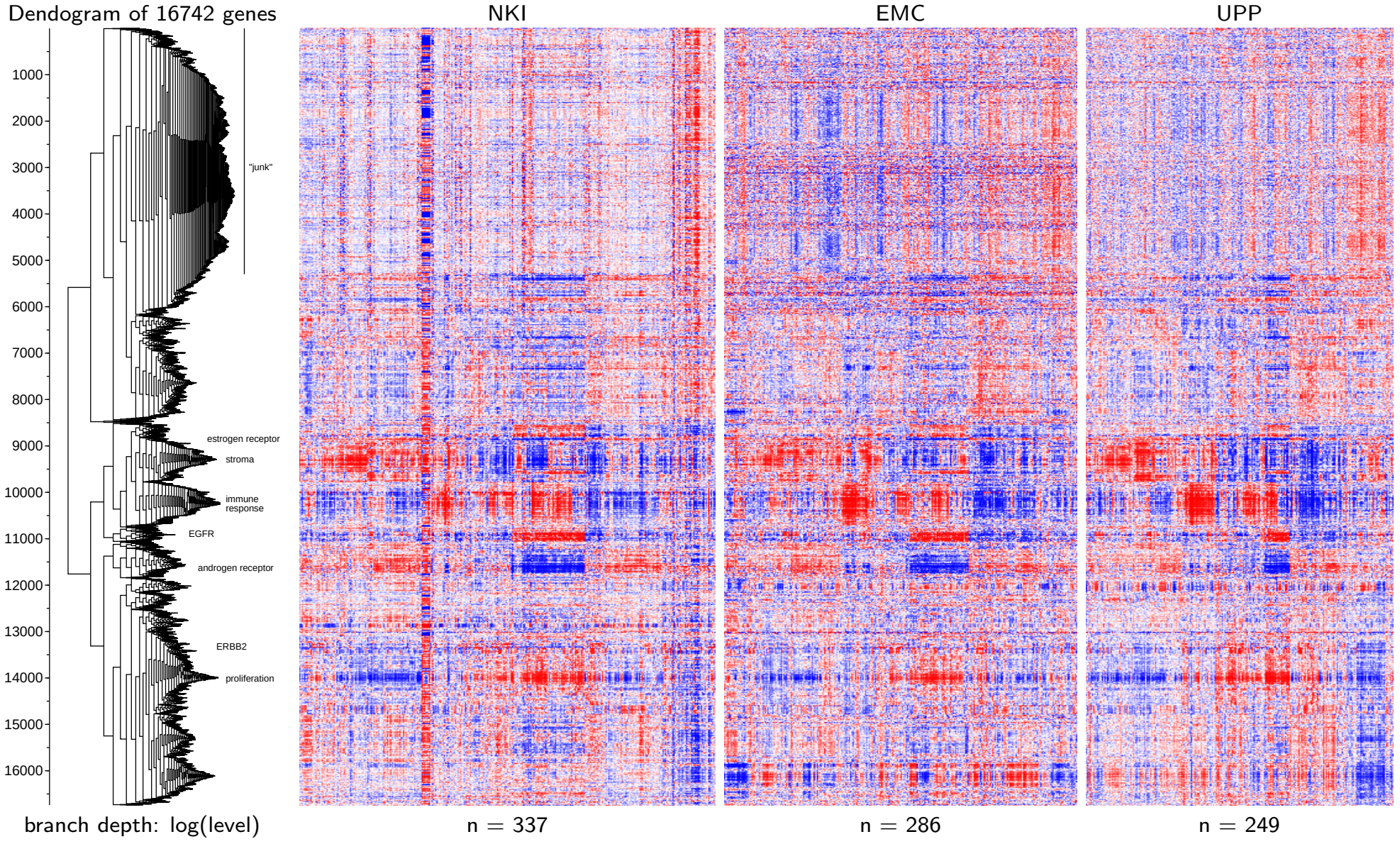
Multi-stage random-effects meta-analysis can be use to both combine the correlations and assess *differential co-expression* using the between-strata variance.

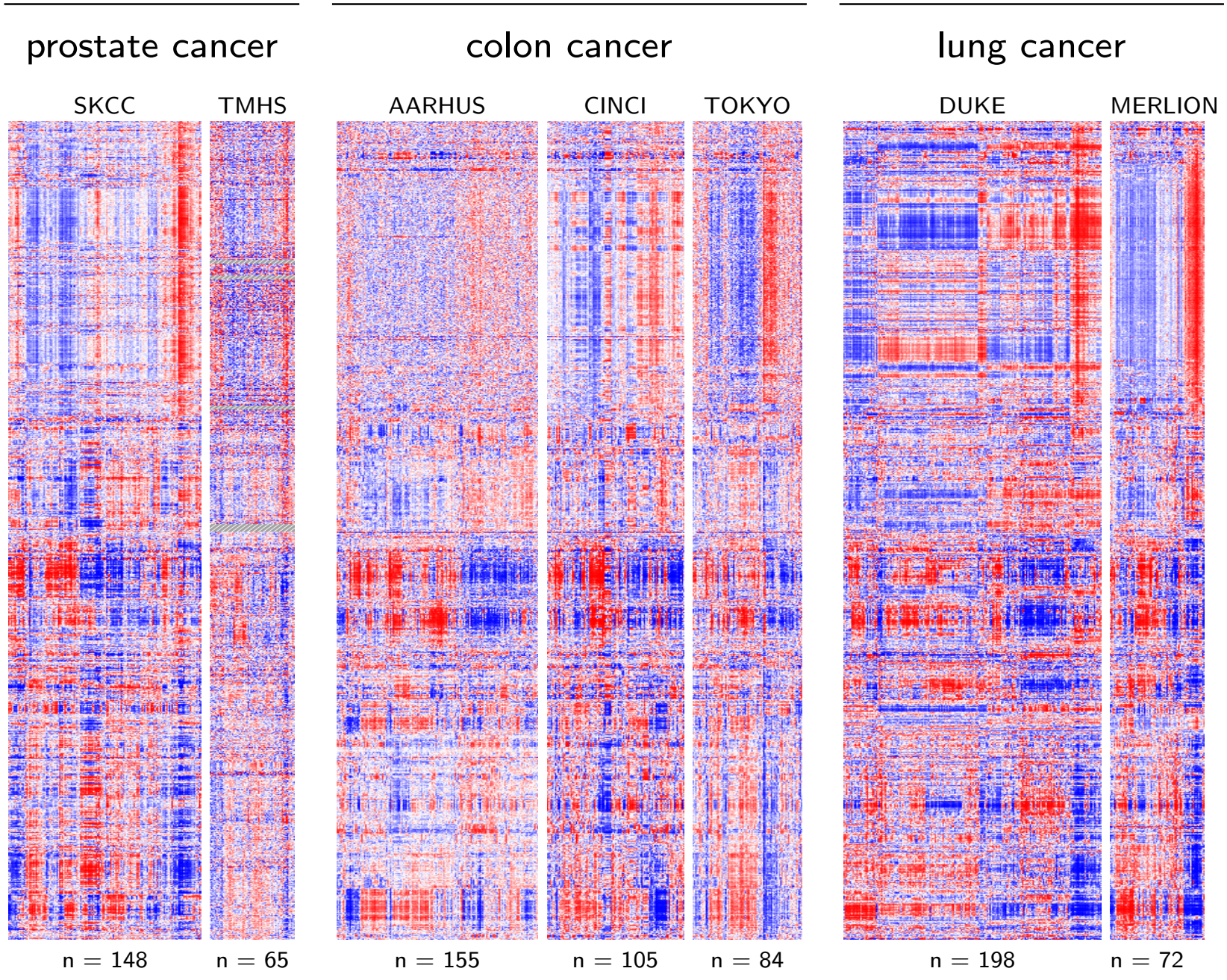
Example: cluster genes in multiple types of cancer, each having multiple studies

Examples of consistently correlated pairs (left) and breast-cancer-only pairs (right)



breast cancer




prostate cancer
colon cancer
lung cancer
SKCC
TMHS
AARHUS
CINCI
TOKYO
DUKE
MERLION
n = 148
n = 65
n = 155
n = 105
n = 84
n = 198
n = 72

Prediction

Components of classifiers:

- Gene list ("signature"): identified by feature selection step
- Model parameters (e.g. coefficients, neural network weights, etc.): identified by model fitting.
- Cutoff

  Very difficult to calibrate. Sensitive to changes in the distribution of both predictor variables and outcome.

  e.g. disease prevalence (or baseline hazard in survival data) in the target populations may be different from those in retrospective study datasets

Naïve/Idiot Bayes predictors

Assume conditional independence amongst predictor variables (conditioned on the response).

DLDA, Tukey's compound covariate, etc. are based on this principle. Penalized regression is similar, if the penalty is large.

- Fit gene-by-gene models
- Select top genes
- Use the gene-by-gene coefficients or significance (t-stat or Z-score, or simply the $\pm$ signs) as weights in linear predictor: $\sum w_i x_i$; the cutoff is to be calibrated from the training set

Still one of the best for microarray data.

$\Rightarrow$ Most amenable to cross-platform applications, because it's insensitive to the exact weights or missing genes.

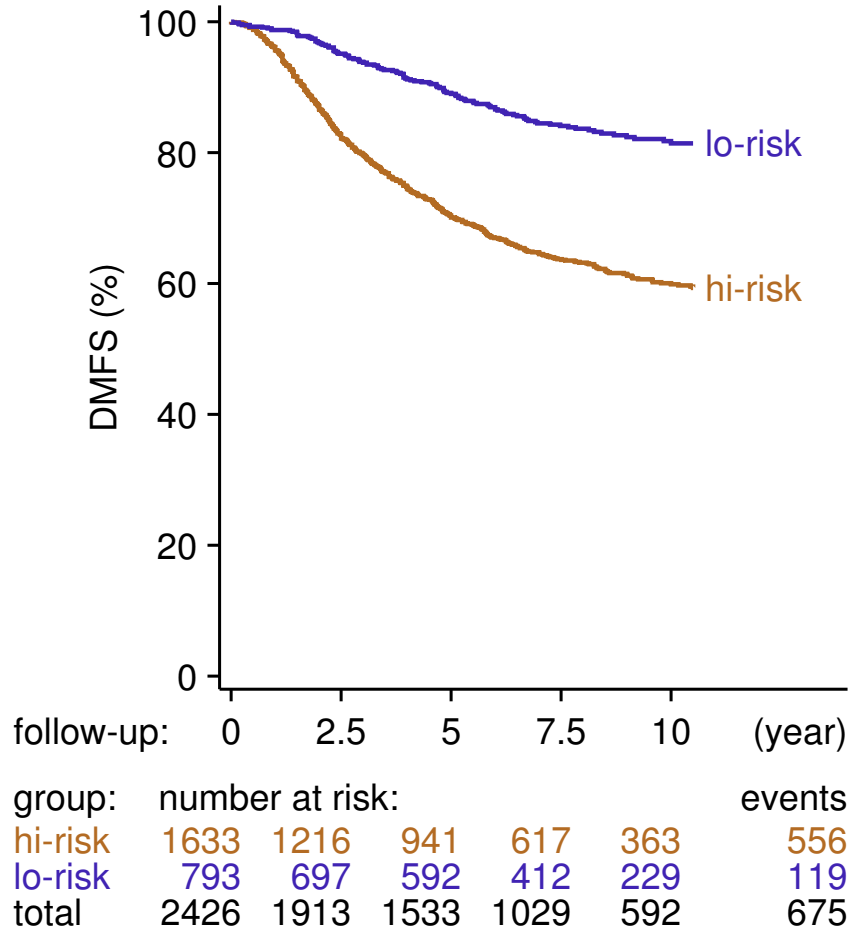
Cross validation schemes

1. Within dataset
    - Split each dataset into learning and test parts
    - Select top genes (ranking based on REMA summaries)
    - In each dataset, apply the model with dataset-specific parameters to the test part
    - Combine performance
2. Cross-dataset
    - Split datasets into learning and test datasets
    - Fit model in the test datasets
    - Apply to test datasets: global weights, local cutoff (need its own CV)

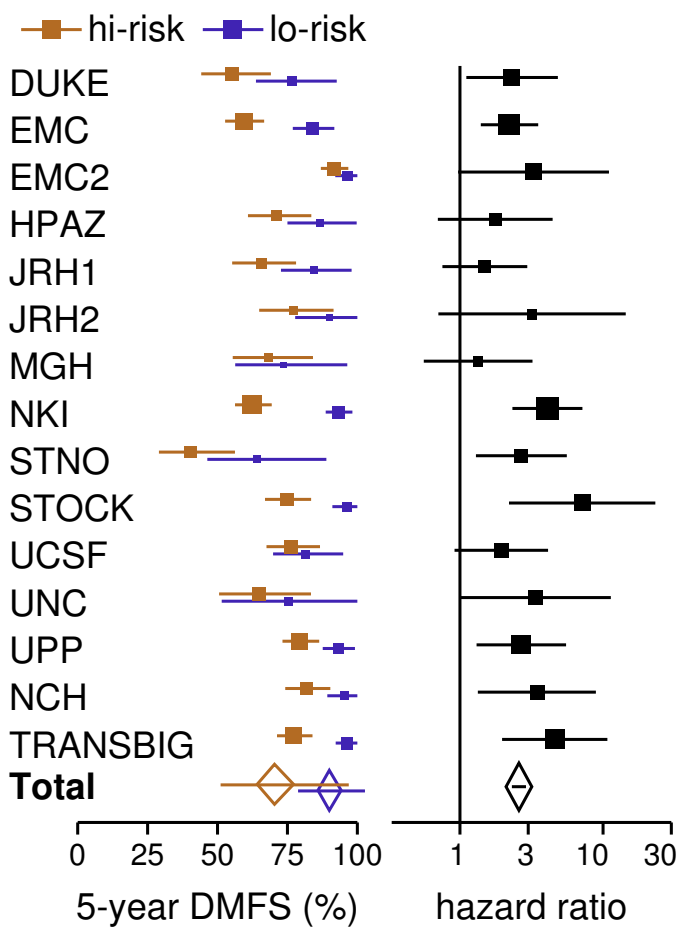
$\Rightarrow$ "Leave-one-dataset out CV" is particularly simple

Example of LODOCV: Breast cancer datasets

Cutoff is 30% low-risk



group:	number at risk:					events
hi-risk	1633	1216	941	617	363	556
lo-risk	793	697	592	412	229	119
total	2426	1913	1533	1029	592	675



Summary

Multiple omics datasets can be co-analyzed under the framework of "standard" statistical methods (e.g. generalized linear models, meta-analysis, hierarchical sampling models).

Extension to complex analysis (e.g., prediction, cluster analysis) is possible, by incorporating REMA for combining summaries, at the appropriate stage of analysis.

Future Work

Release (hopefully soon) R packages for:

- Fast, meta-analytical scanning of GLM (normal, logit, survival).
- Fast multilevel meta-analytical hierarchical clustering

A system for data clean-up and curation (this is the most time consuming part):

- text mining of clinical data and mapping to ontologies
- QC and renormalization/retransformation of expression data