C6215 Advanced Biochemistry and its Methods

Lesson 1

Introduction into Genomics

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a
Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu

Outline

- Definition Of Genomics
- Forward vs Reverse Genetics
- Gene Structure and Identification
- Nucleic Acid Sequencing
- Analysis of Gene Expression

Outline

- Definition Of Genomics

GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
  - Necessary prerequisite: knowledge of the genome (sequence) – work with databases
- *Sensu stricto* (in the narrow sense) – it is interested in FUNCTION of INDIVIDUAL GENES – **FUNCTIONAL GENOMICS**
  - It uses mainly the reverse genetics approaches

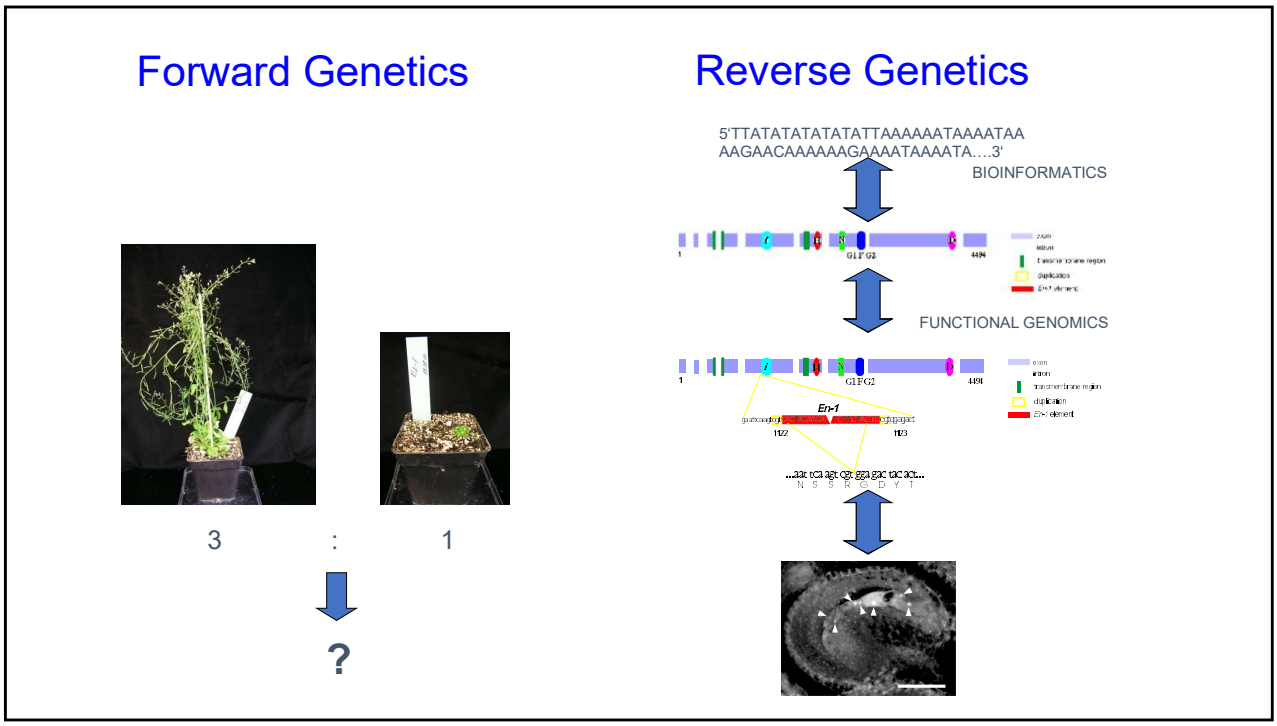
Genomics is a science discipline that is interested in the analysis of genomes. Genome of each organism is a complex of all genes of the respective organism. The genes could be located in cytoplasm (prokaryots) nucleus (in most euckaryotic organisms), mitochondria or chloroplasts (in plants).

The critical prerequisite of genomics is the knowledge of gene sequences.

Functional genomics is interested in function of individual genes.

Outline

- Definition Of Genomics
- Forward vs Reverse Genetics



With the knowledge of gene sequences (or the knowledge of the gene files in the individual organisms, i.e. the knowledge of genomes), **Reverse Genetics** appears that allows study their function.

In comparison to ”classical” or **Forward Genetics**, starting with the phenotype, the reverse genetics starts with the sequence identified as a gene in the sequenced genome. The gene identification using approaches of **Bioinformatics** will be described later (see Lesson 02).

Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

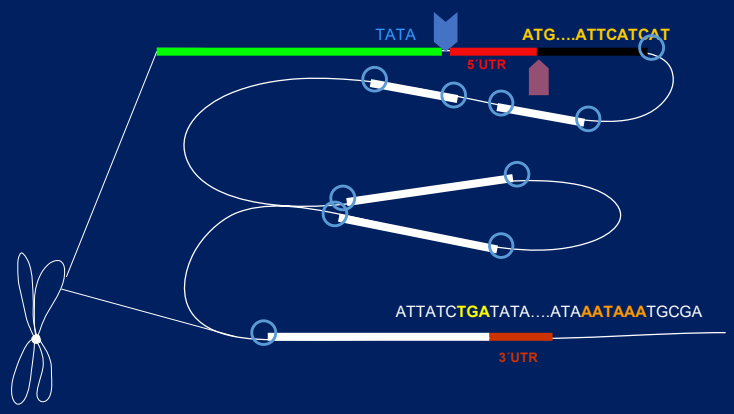
The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype,* but rather as a piece of DNA characterized by the unique *string of nucleotides*. i.e. **physical DNA molecule**.

Outline

- Definition Of Genomics
- Forward vs Reverse Genetics
- Gene Structure and Identification

Gene Structure

- Promoter
- Transcriptional start
- 5´UTR
- Translational start
- Splicing sites
- Stop codon
- 3´UTR
- Polyadenylation signal



Identification of Genes *Ab Initio*

- Omitting 5‘ and 3‘ UTR
- Identification of translation start (ATG) and stop codon (TAG, TAA, TGA)
- Finding donor (typically GT) and acceptor (AG) splicing sites
- Many ORFs are NOT real coding sequences
- Using various statistic models (e.g. Hidden Markov Model – HMM, see recommended literature, Majoros *et al.*, 2003) to evaluate and score the weight of identified donor and acceptor sites

Experimental Gene Identification

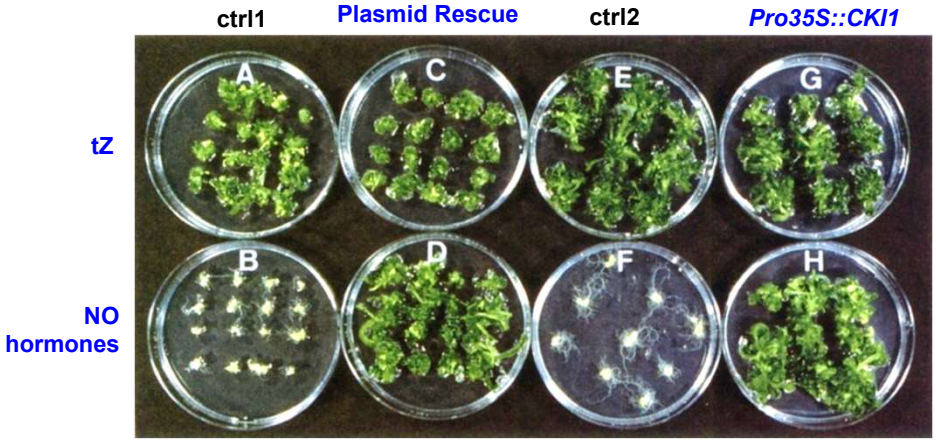
- Principles of experimental identification of genes using forward and revers genetics
  - Alteration of phenotype after mutagenesis
    - **Forward genetics**
  - Identification of sequence-specific mutant and analysis of its phenotype
    - **Reverse genetics**
  - Analysis of expression of a particular gene and its spatiotemporal specifity

Experimental Gene Identification

- Principles of experimental identification of genes using forward and revers genetics
  - Alteration of phenotype after mutagenesis
    - **Forward genetics**

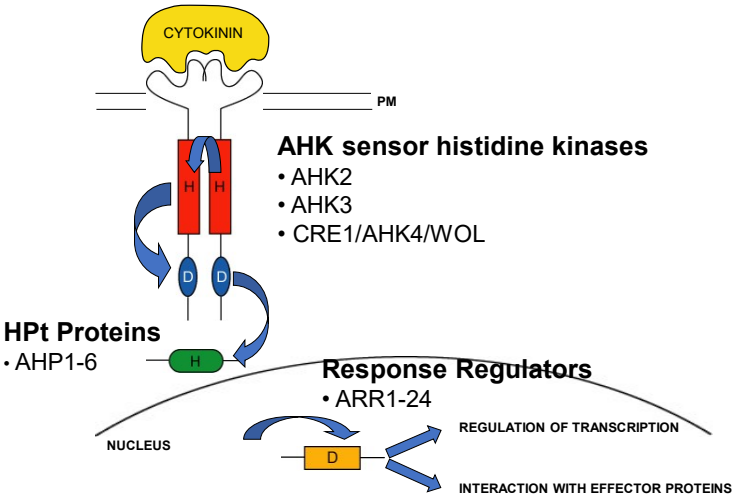
Identification of *CKI1* via Activation Mutagenesis

- *CKI1* overexpression mimics cytokinin response



Kakimoto, *Science*, 1996

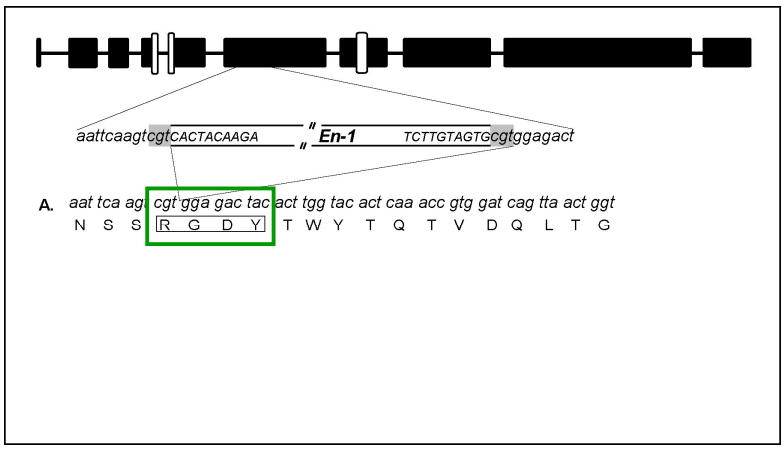
Signal Transduction via MSP



Reverse Genetics

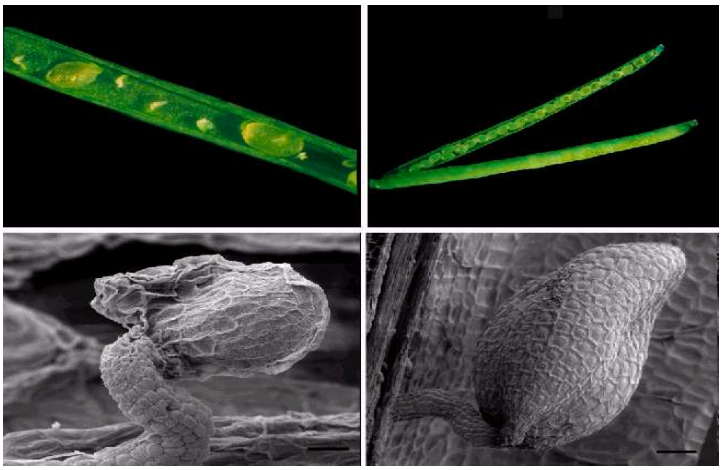
- Principles of experimental identification of genes using forward and revers genetics
  - Alteration of phenotype after mutagenesis
    - **Forward genetics**
  - Identification of sequence-specific mutant and analysis of its phenotype
    - **Reverse genetics**

Identification of insertional *cki1* mutant



CKI1 Regulates Female Gametophyte Development

CKI1/cki1-i ***CKI1/CKI1***



Hejátko et al., *Mol Genet Genomics* (2003)

cki1-i reveals non-Mendelian inheritance

P *CKI1/cki1-i*

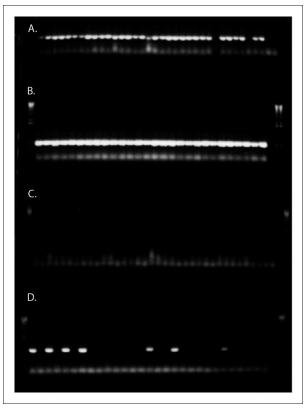
F1 Anticipated: **1** *CKI1* **: 2** *CKI1/cki1-i* **: 1** *cki1-i*

Observed: **1** *CKI1* : **1** *CKI1/cki1-i*

♀ \ ♂	*CKI1*	*cki1-i*
CKI1	*CKI1/CKI1*	*CKI1/cki1-i*
cki1-i	*CKI1/cki1-i*	

CKI1 and Megagametogenesis

- *cki1-i* is not transmitted through the female gametophyte



A. ♂ **wt x** ♀ ***CKI1/cki1-i***

CKI1 **specific primers (PCR positive control)**

B. ♂ ***CKI1/cki1-i*** **x** ♀ **wt**

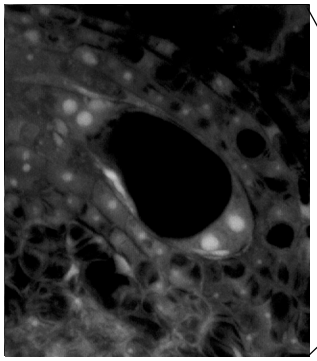
C. ♂ **wt x** ♀ ***CKI1/cki1-i***

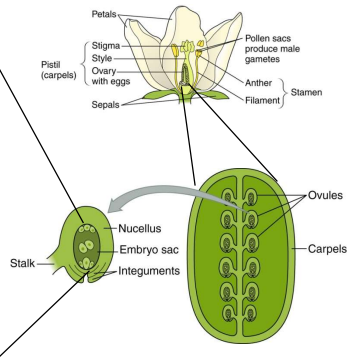
cki1-i **specific primers**

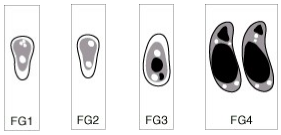
D. ♂ ***CKI1/cki1-i*** **x** ♀ **wt**

CKI1 and Megagametogenesis

FG 4

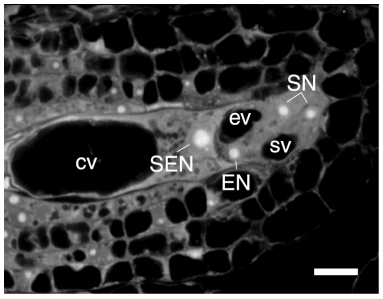






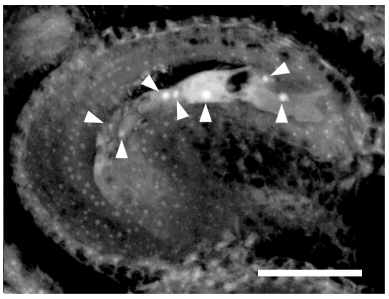
CKI1 and Megagametogenesis

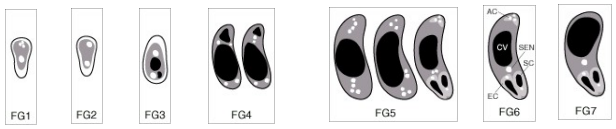
CKI1 — Late FG5



Hejátko et al., *Mol Genet Genomics* (2003)

cki1-i — 48 HAE

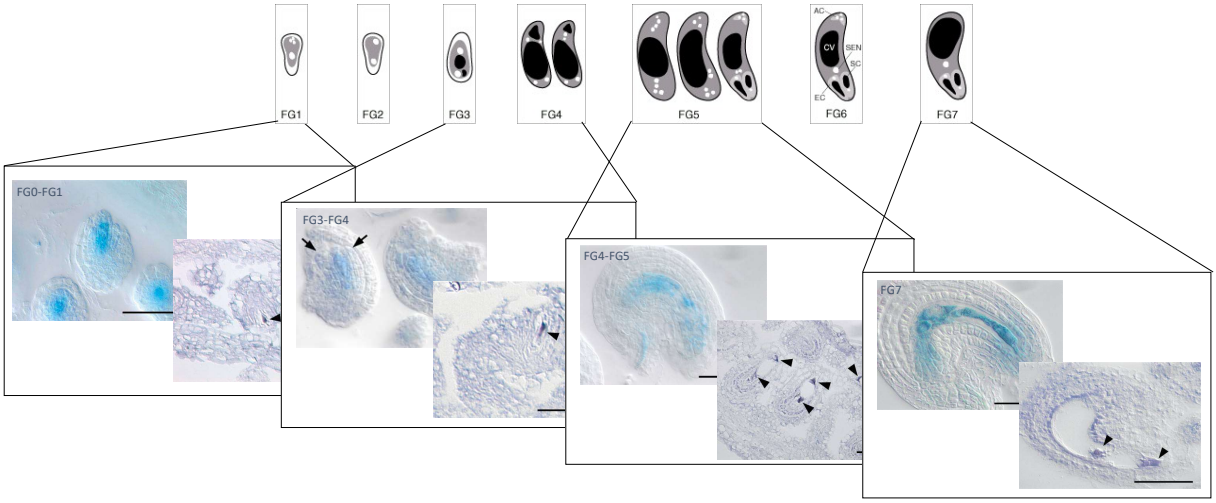




Experimental Gene Identification

- Principles of experimental identification of genes using forward and revers genetics
  - Alteration of phenotype after mutagenesis
    - **Forward genetics**
  - Identification of sequence-specific mutant and analysis of its phenotype
    - **Reverse genetics**
  - Analysis of expression of a particular gene and its spatiotemporal specifity

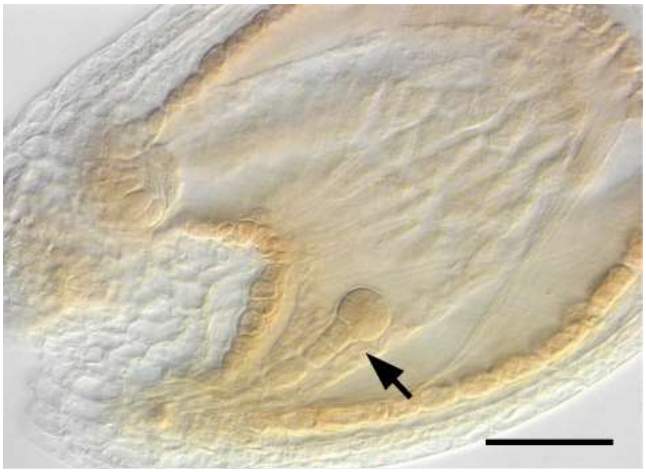
CKI1 and Megagametogenesis



Paternal *CKI1* is Expressed Early after Fertilization

♀ wt x ♂ Pro*CKI1:GUS*

22 HAP
(hours after pollination)



Hejátko et al., *Mol Genet Genomics* (2003)

Outline

- Definition Of Genomics
- Forward vs Reverse Genetics
- Genes Structure and Identification
- Nucleic Acid Sequencing

Sanger Sequencing

Frederick Sanger

1958 – Nobel prize – insulin structure

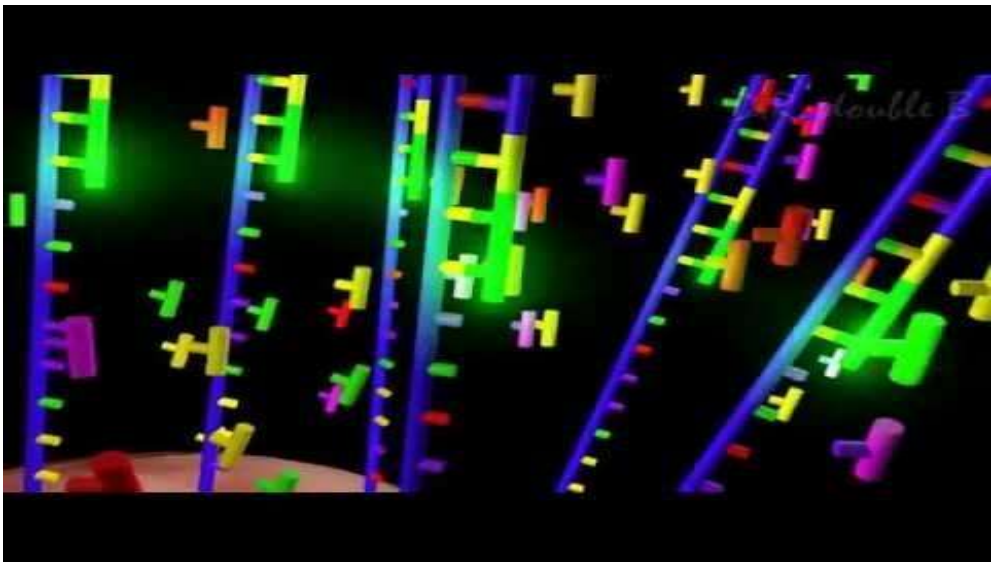
1975 - Dideoxy sequencing method

1980 – second Nobel prize for NA sequencing

Sanger Sequencing

Original video @ https://www.youtube.com/watch?v=K0RThNB-HE

NGS Sequencing



Original video at https://www.youtube.com/watch?v=-7GK1HXwCtE.

For more detailed description see e.g. https://www.youtube.com/watch?v=fCd6B5HRaZ8.

Outline

- Definition Of Genomics
- Forward vs Reverse Genetics
- Genes Structure and Identification
- Nucleic Acid Sequencing
- Analysis of Gene Expression

Gene Expression Assays

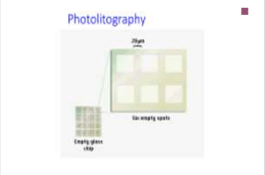
- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips
    - Next generation transcriptional profiling
  - Qualitative analysis of gene expression
    - Preparation of transcriptional fusion of promoter of analysed gene with a reporter gene
    - Preparation of translational fusion of the coding region of the analysed gene with reporter gene
    - Use of the data available in public databases
    - Tissue- and cell-specific gene expression analysis

Expression Assays

- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips

DNA Chips

- DNA čipy
  - metoda umožňující rychlé porovnání velkého množství genů/proteinů mezi testovaným vzorkem a kontrolou
  - nejčastěji jsou používané oligo DNA čipy
    - k dispozici komerčně dostupné sady pro celý genom
    - firma Operon (Qiagen), 29.110 70-mer oligonulkleotidů reprezentujících 26.173 genů kódujících proteiny, 28.964 transkriptů a 87 microRNA genů *Arabidopsis thaliana*
    - možnost používat pro přípravu čipů fotolitografické techniky-usnadnění syntézy oligonukleotidů např. pro celý genom člověka (cca 3,1 x 10^9 bp) je touto technikou možno připravit 25-mery v pouze 100 krocích)



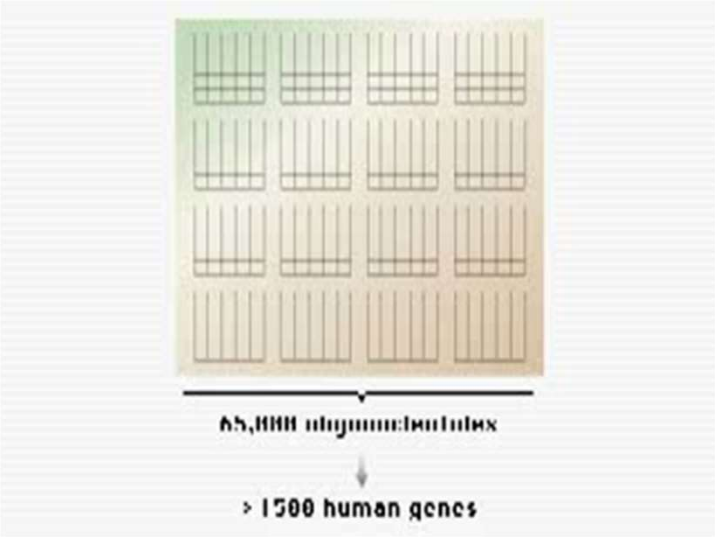
- čipy nejen pro analýzu exprese, ale např. i genotypování (SNPs – jednonukleotidové polymorfizmy, sekvenování pomocí čipů, …)

Affymetrix ATH1 *Arabidopsis* genome array

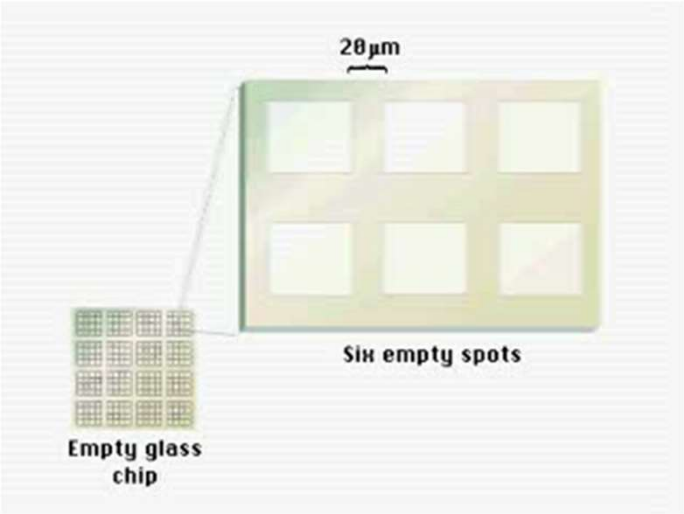
Critical Specifications	
Number of arrays	One
Number of sequence represented	>24,000 gene sequences
Feature size	18 µm
Oligonucleotide probe length	25-mer
Probe pairs/sequence	11
Control sequences	*E. coli* genes *bioB, bioC, bioD*. *B. subtilis* gene *lysA*. Phage P1 *cre* gene. Arabidopsis maintenance genes GAPDH, Ubiquitin, and Actin
Detection sensitivity	1:100,000*

*As measured by detection in comparative analysis between a complex target containing spiked control transcriptions and a complex target with no spikes.

DNA Chips

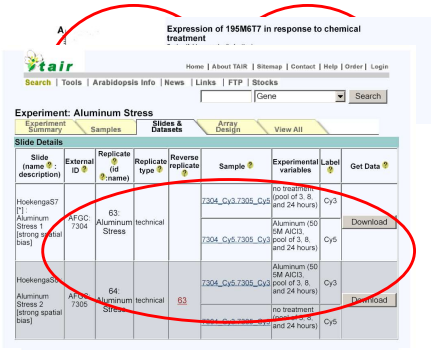


Photolitography



DNA Chips

- For the correct interpretation of the results, good knowledge of advanced statistical methods is required
- It is necessary to include a sufficient number of controls and repeats
  - Control of **accuracy** of the measurement (repeated measurements on several chips with the same sample, comparing the same samples analysed on different chips with each other)
  - Control of **reproducibility** of measurements (repeated measurements with different samples isolated under the same conditions on the same chip – comparing with each other)
  - Identification of **reliable measurement treshold**
  - Finally comparing the experiment with the control or comparing different conditions with each other -> the result
  - Currently there's been a great number of results of various experiments in publicly accessible databases



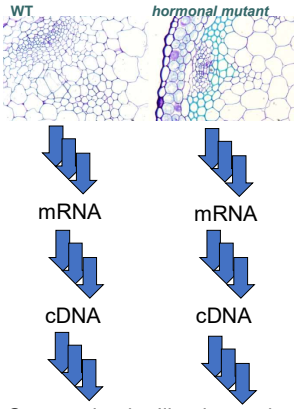
Che et al., 2002

Gene Expression Assays

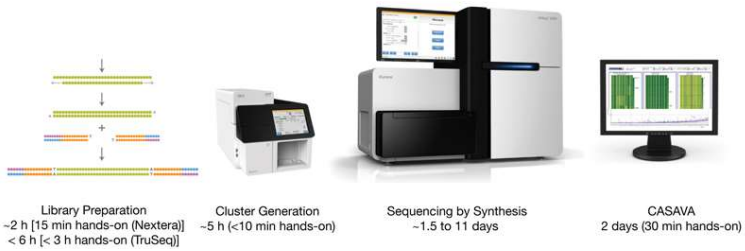
- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips
    - Next generation transcriptional profiling

Next Gen Transcriptional Profiling

- ***Transcriptional profiling*** via ***RNA sequencing***



Sequencing by Illumina and
number of transcripts determination



Results of –omics Studies vs Biologically Relevant Conclusions

- Transcriptional profiling yielded more then **7K differentially regulated genes**…

gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
AT1G07795	1:2414285-2414967	WT	MT	OK	0	1,1804	1.79769e+308	1.79769e+308	6.88885e-05	0,00391801	yes
HRS1	1:4556891-4558708	WT	MT	OK	0	0,696583	1.79769e+308	1.79769e+308	6.61994e-06	4.67708e-05	yes
ATMLO14	1:9227472-9232296	WT	MT	OK	0	0,514609	1.79769e+308	1.79769e+308	9.74219e-05	0,00053055	yes
NRT1.6	1:9400663-9403789	WT	MT	OK	0	0,877865	1.79769e+308	1.79769e+308	3.2692e-08	3.50131e-07	yes
AT1G27570	1:9575425-9582376	WT	MT	OK	0	2,0829	1.79769e+308	1.79769e+308	9.76039e-06	6.647e-05	yes
AT1G60095	1:22159735-22162419	WT	MT	OK	0	0,688588	1.79769e+308	1.79769e+308	9.95901e-08	9.84992e-07	yes
AT1G03020	1:698206-698515	WT	MT	OK	0	1,78859	1.79769e+308	1.79769e+308	0,00913915	0,0277958	yes
AT1G13609	1:4662720-4663471	WT	MT	OK	0	3,55814	1.79769e+308	1.79769e+308	0,00021683	0,00108079	yes
AT1G21550	1:7553100-7553876	WT	MT	OK	0	0,562868	1.79769e+308	1.79769e+308	0,00115582	0,00471497	yes
AT1G22120	1:7806308-7809632	WT	MT	OK	0	0,617354	1.79769e+308	1.79769e+308	2.48392e-06	1.91089e-05	yes
AT1G31370	1:11238297-11239363	WT	MT	OK	0	1,46254	1.79769e+308	1.79769e+308	4.83523e-05	0,000285143	yes
APUM10	1:13253397-13255570	WT	MT	OK	0	0,581031	1.79769e+308	1.79769e+308	7.87855e-06	5.46603e-05	yes
AT1G48700	1:18010728-18012871	WT	MT	OK	0	0,556525	1.79769e+308	1.79769e+308	6.53917e-05	0,000374736	yes
AT1G59077	1:21746209-21833195	WT	MT	OK	0	138,886	1.79769e+308	1.79769e+308	0,00122789	0,00496816	yes
AT1G60050	1:22121549-22123702	WT	MT	OK	0	0,370087	1.79769e+308	1.79769e+308	0,00117953	0,0048001	yes
AT4G15242	**4:8705786-8706997**	**WT**	**MT**	**OK**	**0,00930712**	**17,9056**	**10,9098**	**-4,40523**	**1.05673e-05**	**7.13983e-05**	**yes**
ATSG33251	**5:12499071-12500433**	**WT**	**MT**	**OK**	**0,0498375**	**52,2837**	**10,0349**	**-9,8119**	**0**	**0**	**yes**
AT4G12520	**4:7421055-7421738**	**WT**	**MT**	**OK**	**0,0195111**	**15,8516**	**9,66612**	**-3,90043**	**9.60217e-05**	**0,000528904**	**yes**
AT1G60020	**1:22100651-22105276**	**WT**	**MT**	**OK**	**0,0118377**	**7,18823**	**9,24611**	**-7,50382**	**6.19504e-14**	**1.4988e-12**	**yes**
ATSG15360	**5:4987235-4989182**	**WT**	**MT**	**OK**	**0,0988273**	**56,4834**	**9,1587**	**-10,4392**	**0**	**0**	**yes**

Ddii et al., *unpublished*

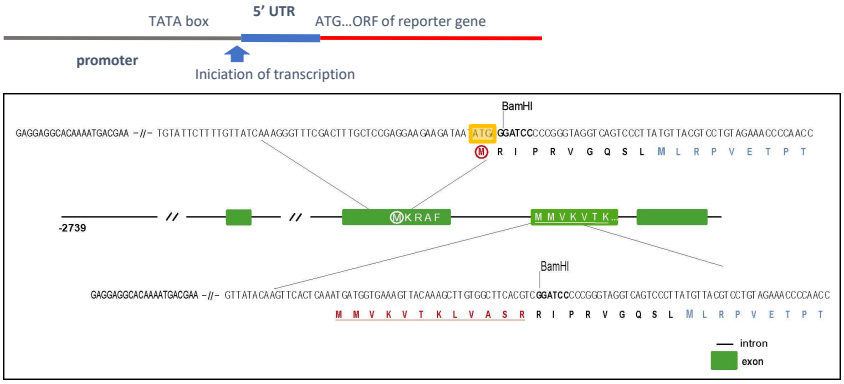
Excample of an output of transcriptional profiling study using Illumina sequencing performed in our lab. Shown is just a tiny fragment of the complete list, copmprising about 7K genes revealing differential expression in the studied mutant.

Gene Expression Assays

- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips
    - Next generation transcriptional profiling
  - Qualitative analysis of gene expression
    - Preparation of transcriptional fusion of promoter of analysed gene with a reporter gene

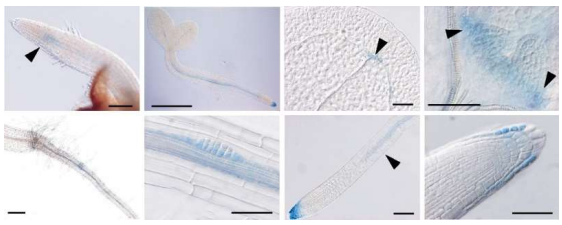
Transcriptional Fusion

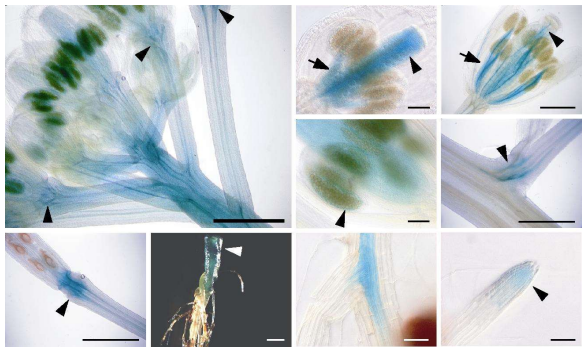
- Identification and cloning of the promoter region of the gene
- Preparation of recombinant DNA carrying the promoter and the reporter gene (uidA, GFP)



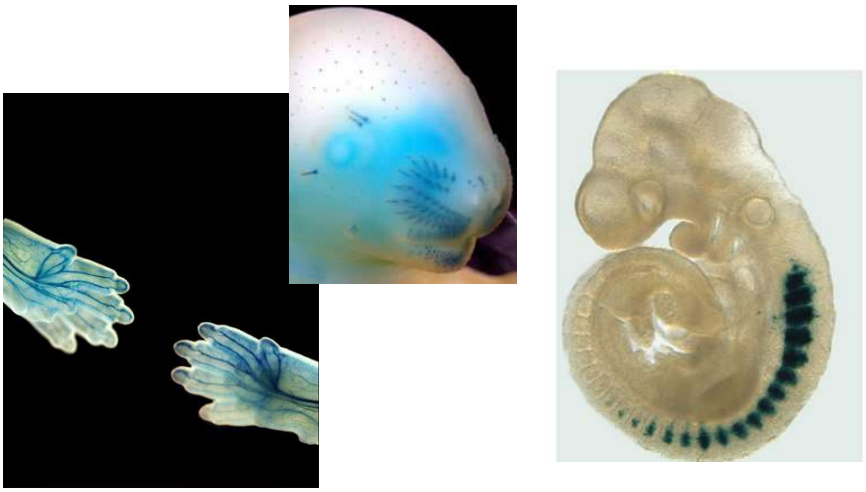
Transcriptional Fusion

- Identification and cloning of the promoter region of the gene
- Preparation of recombinant DNA carrying the promoter and the reporter gene (uidA, GFP)
- Preparation of transgenic organisms carrying this recombinant DNA and their histological analysis





LacZ Reporter in Mouse Embryos

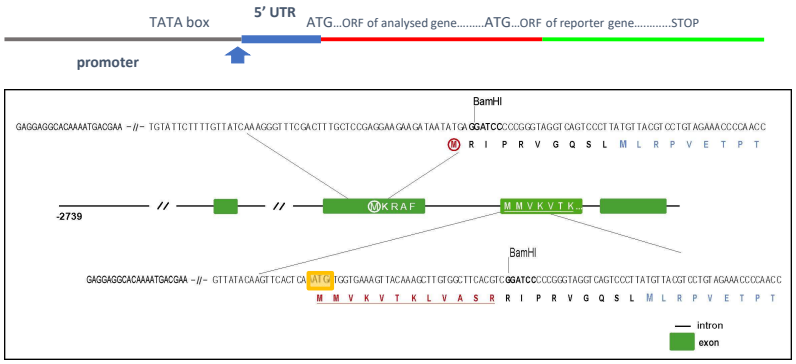


Gene Expression Assays

- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips
    - Next generation transcriptional profiling
  - Qualitative analysis of gene expression
    - Preparation of transcriptional fusion of promoter of analysed gene with a reporter gene
    - Preparation of translational fusion of the coding region of the analysed gene with reporter gene

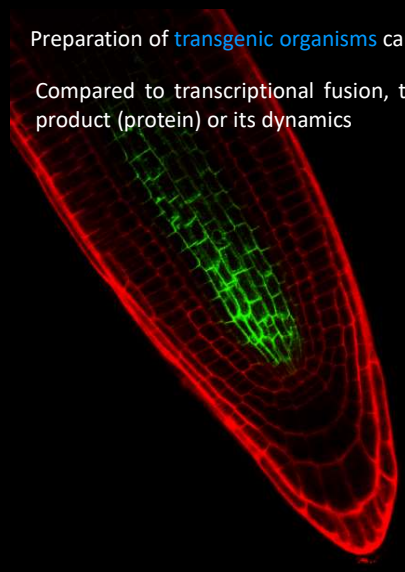
Translational Fusion

- Identification and cloning of the promoter and coding region of the analyzed gene
- Preparation of a recombinant DNA carrying the promoter and the coding sequence of the studied gene in a fusion with the reporter gene (uidA, GFP)

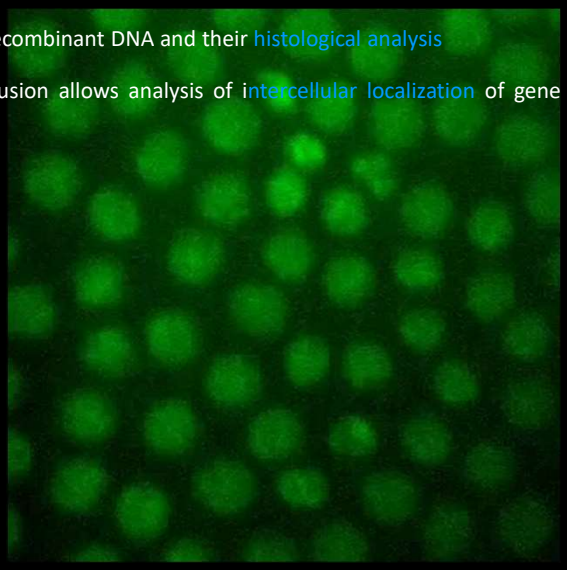


Translational Fusion

- Preparation of transgenic organisms carrying the recombinant DNA and their histological analysis
- Compared to transcriptional fusion, translation fusion allows analysis of intercellular localization of gene product (protein) or its dynamics

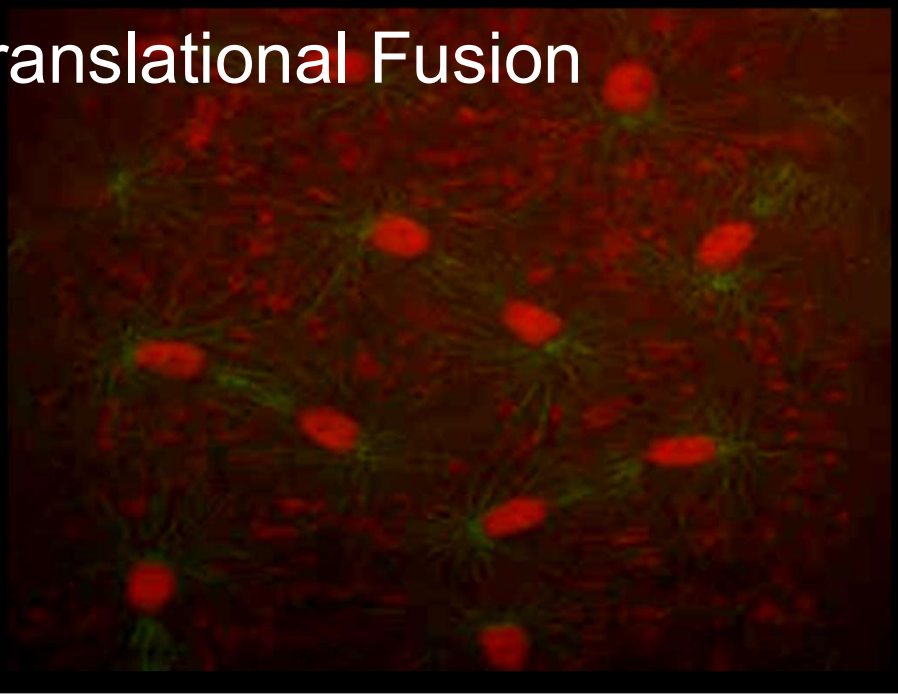


PIN1-GFP in *Arabidopsis*



Histone 2A-GFP in *Drosophila* embryo by PAM

Translational Fusion

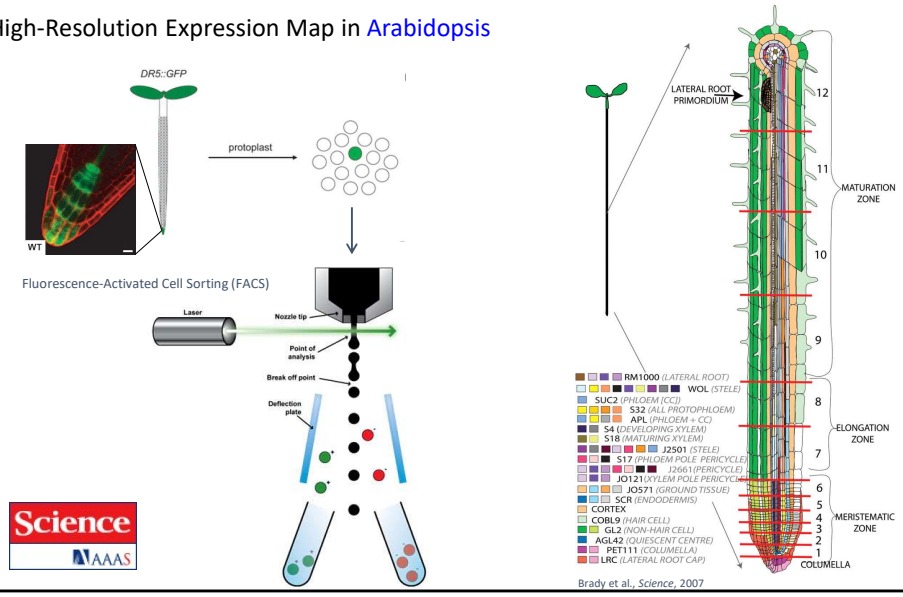


Gene Expression Assays

- Methods of gene expression analysis
  - Quantitative analysis of gene expression
    - DNA chips
    - Next generation transcriptional profiling
  - Qualitative analysis of gene expression
    - Preparation of transcriptional fusion of promoter of analysed gene with a reporter gene
    - Preparation of translational fusion of the coding region of the analysed gene with reporter gene
    - Use of the data available in public databases
    - Tissue- and cell-specific gene expression analysis

Gene Expression Assays

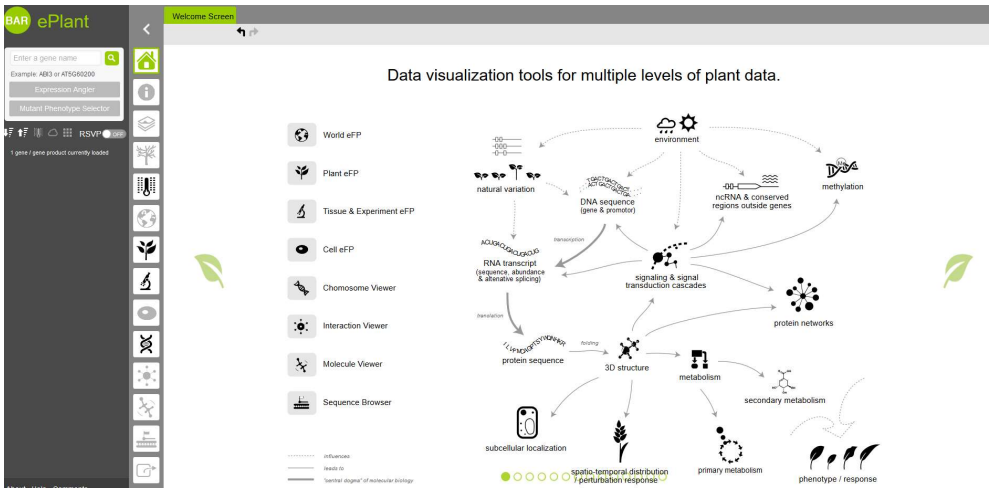
- High-Resolution Expression Map in Arabidopsis



Microarray expression profiles of 19 fluorescently sorted GFP-marked lines were analyzed (*3*–*9*, *23*, *24*). The colors associated with each marker line reflect the developmental stage and cell types sampled. Thirteen transverse sections were sampled along the root's longitudinal axis (red lines) (*10*). CC, companion cells.

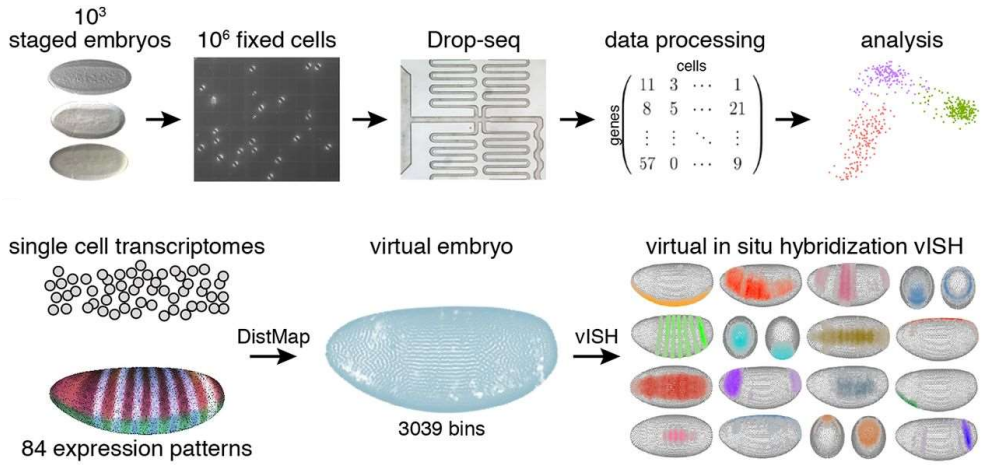
BAR ePlant

https://bar.utoronto.ca/eplant/



Expression Maps - RNA

- High-Resolution Expression Map in Drosophilla



Nikos Karaiskos et al. Science 2017;science.aan3235

Deconstructing and reconstructing the embryo by single-cell transcriptomics combined with spatial mapping.

(A) Single-cell sequencing of the Drosophila embryo: ~1000 handpicked stage 6 fly embryos are dissociated per Drop-seq replicate, cells are fixed and counted, single cells are combined with barcoded capture beads, and libraries are prepared and sequenced. Finally, single-cell transcriptomes are deconvolved, resulting in a digital gene expression matrix for further analysis.

(B) Mapping cells back to the embryo: Single-cell transcriptomes are correlated with high-resolution gene expression patterns across 84 marker genes, cells are mapped to positions within a virtual embryo, and expression patterns are computed by combining the mapping probabilities with the expression levels (virtual in situ hybridization).

Drosophila Virtual Expression eXplorer

https://shiny.mdc-berlin.de/DVEX/

DVEX | t-SNE | vISH | vISHs | vISH - D. vir. | Gradients | Archetypes | Download | About

DVEX.org

MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION
BIMSB THE BERLIN INSTITUTE FOR MEDICAL SYSTEMS BIOLOGY

Loading data ... (DVEX is currently better supported on Linux / Mac OS X)

Data loaded for 8924 genes.

Drosophila Virtual Expression eXplorer

DVEX is an online resource tool which offers an easy way to explore the transcriptome of the stage 6 *Drosophila* embryo at the single cell level. It is part of the collaboration between the Rajewsky and Zinzen labs in the Berlin Institute of Medical Systems Biology of the Max Delbrueck Center in Berlin. DVEX accompanies the following publication:

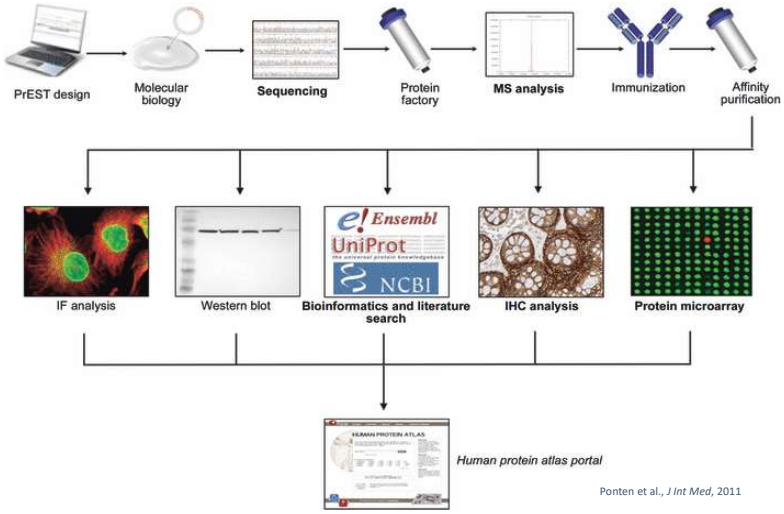
> Science 358 (6360), 194-199
> The Drosophila Embryo at Single Cell Transcriptome Resolution
>
> Nikos Karaiskos[a], Philipp Wahle[b], Jonathan Alles[a], Anastasiya Boltengagen[a], Salah Ayoub[a], Claudia Kipar[b], Christine Kocks[a], Nikolaus Rajewsky[a], Robert Zinzen[b],
>
> [a]Systems biology of gene regulatory elements, BIMSB, MDC
> [b]Systems biology of neural tissue differentiation, BIMSB, MDC

Correspondence regarding the publication: Nikolaus Rajewsky and Robert P. Zinzen.

DVEX is created and maintained by Nikos Karaiskos. Contact the author for questions, or troubleshooting.

Expression Maps - Proteins

- Human Protein Atlas (http://www.proteinatlas.org/)

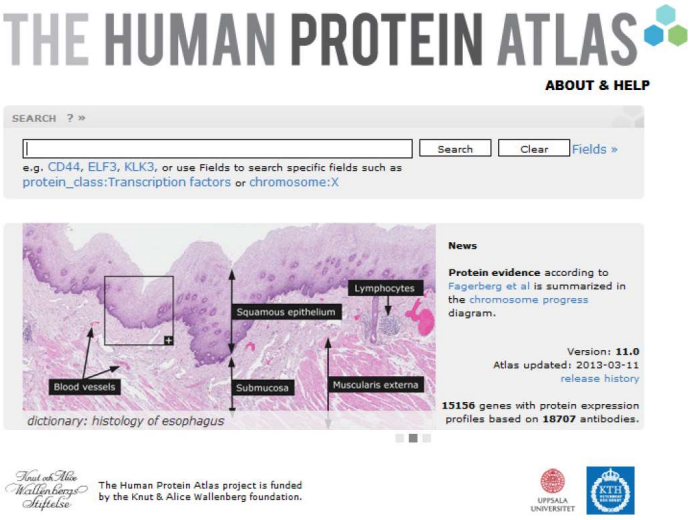


Ponten et al., *J Int Med*, 2011

Schematic flowchart of the Human Protein Atlas. For each gene, a signature sequence (PrEST) is defined from the human genome sequence, and following RT-PCR, cloning and production of recombinant protein fragments, subsequent immunization and affinity purification of antisera results inmunospecific antibodies. The produced antibodies are tested and validated in various immunoassays. Approved antibodies are used for protein profiling in cells (immunofluorescence) and tissues (immunohistochemistry) to generate the images and protein expression data that are presented in the Human Protein Atlas (Ponten et al., *J Int Med*, 2011).

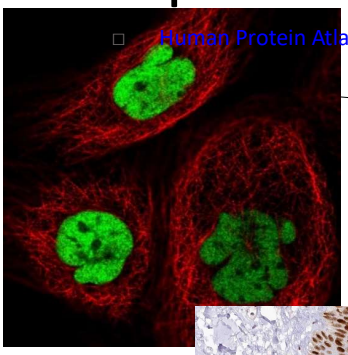
Expression Maps - Proteins

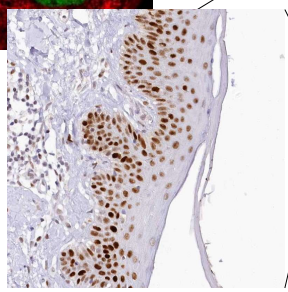
- Human Protein Atlas (http://www.proteinatlas.org/)

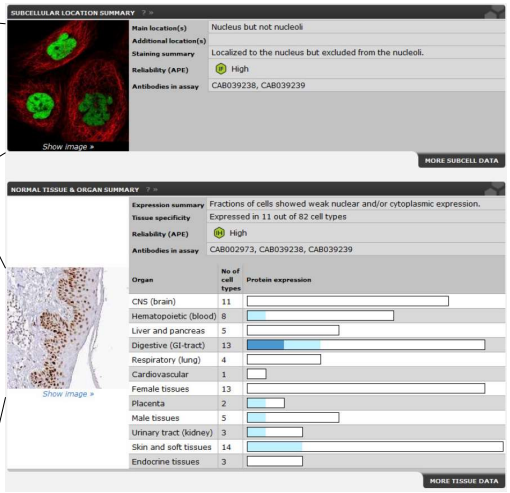


Expression Maps - Proteins

- Human Protein Atlas (http://www.proteinatlas.org/)







Summary

- Definition Of Genomics
- Forward vs Reverse Genetics
- Genes Structure and Identification
- Nucleic Acid Sequencing
- Analysis of Gene Expression

Discussion