Part 2: Large-scale gene expression analysis using microarrays and RNA-seq

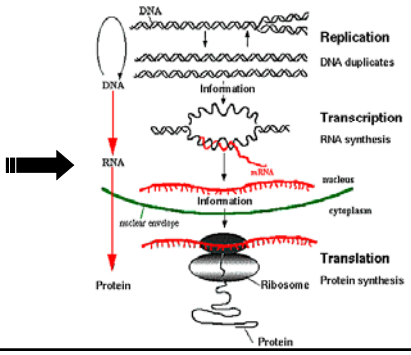
MVE 360 – Bioinformatics, 2012

Erik Kristiansson, erik.kristiansson@chalmers.se

Agenda

- Large-scale mRNA quantification
  - Identification of differentially expressed genes
  - Techniques: microarray and RNA-seq
- *De novo* sequencing of mRNA
  - Identification of the sequence of genes
  - Techniques: RNA-seq

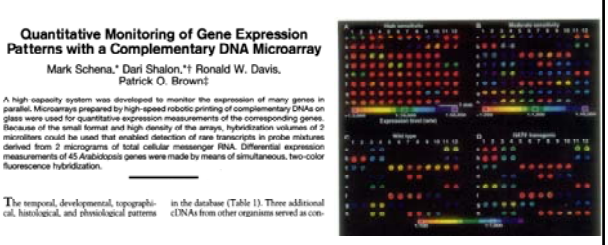
Central Dogma



Large-scale gene expression analysis

- Measurements are done on a genome-wide, *i.e.* for all genes in the genome
- A single experiment results in 10.000-100.000 data points
- Statistical and computational tools are therefore essential for a proper analysis
- Large-scale gene expression analysis is mainly used for explorative research.

The first gene expression microarray!


Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

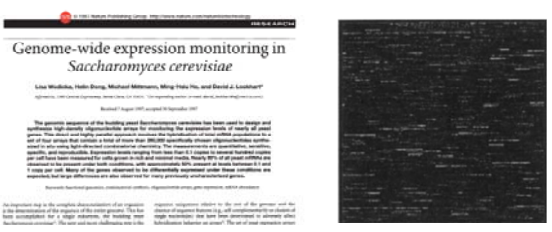
Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

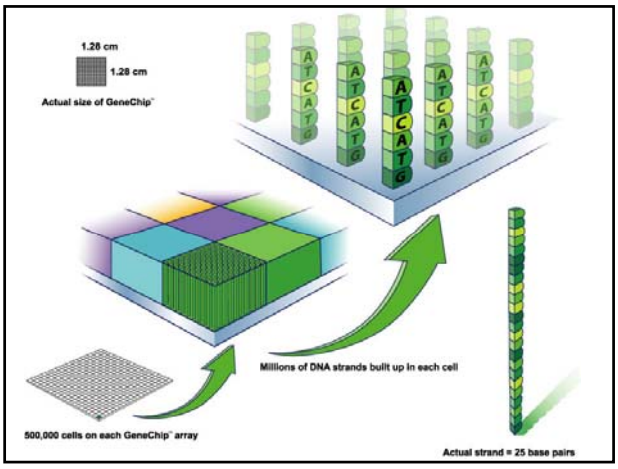
The temporal, developmental, topographical, histological, and physiological patterns in the database (Table 1). Three additional cDNAs from other organisms served as con-

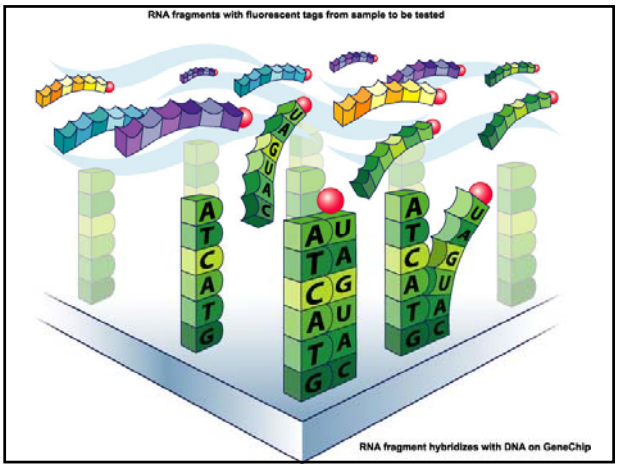
Schena, M., Shalon, D., Davis, R. W., Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* 270, no. 5235, pp 467-470.

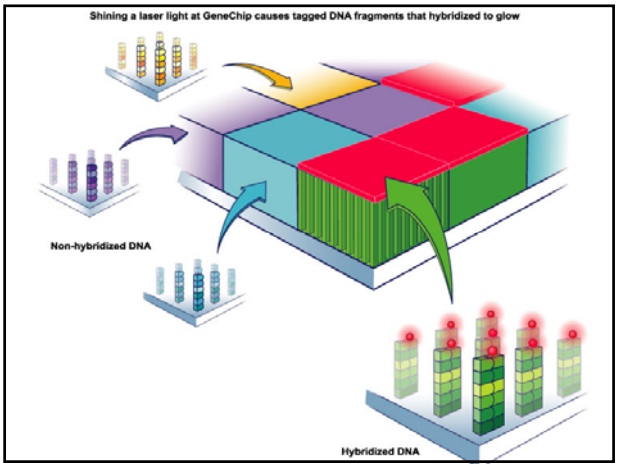
First commercial gene expression microarray!

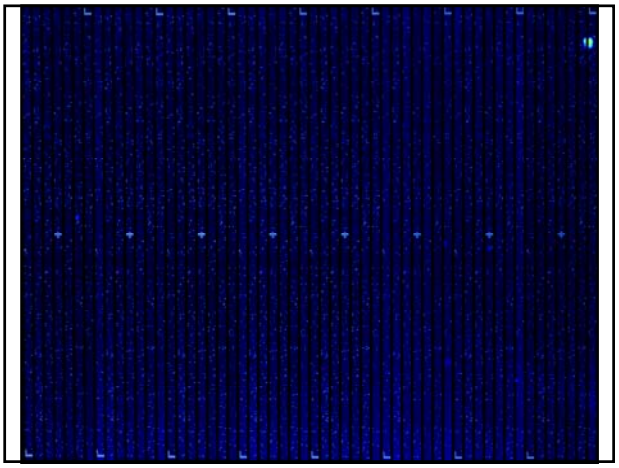

Genome-wide expression monitoring in *Saccharomyces cerevisiae*

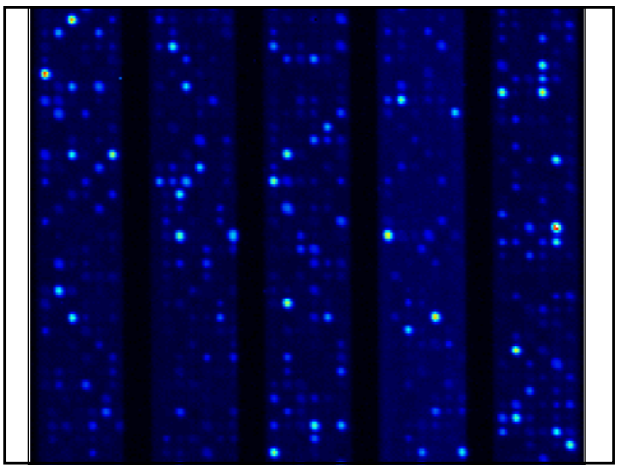
Wodicka, L., Dong, H., Mittmann, M., Ho, MH., Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology **15** pp. 1359-1367.









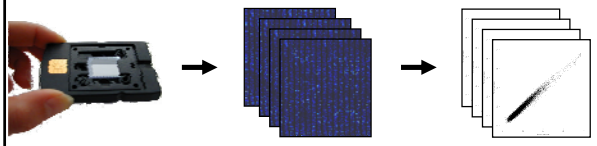


A simple view of a microarray experiment

Experimental Design → Lab work → Pre-processing → Gene Ranking

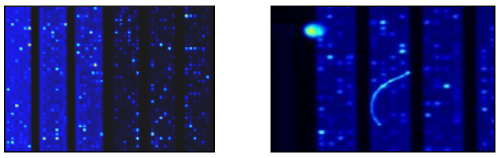
Image Analysis

- Image analysis is then used to transfer the information on image into quantitative data
- For each probe, a foreground and a background intensity is extracted



Background correction

- Background correction is used to remove spatial trends



Background correction

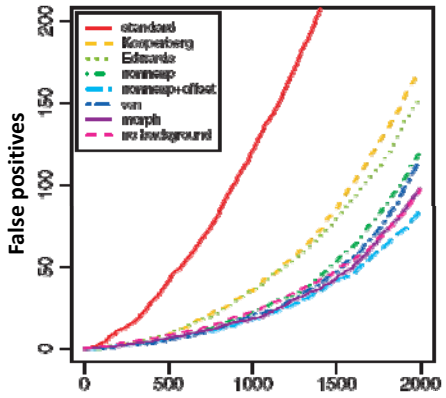
- The most common method is "subtract"

$$\hat{X} = X_f - X_b$$

- Another common method is simply to skip background correction and use

$$\hat{X} = X_f$$

- More complicated procedure include *normexp*, *Kooperberg* and *Edwards*.



Data representation

- Gene expression data from microarrays are transformed to logarithmic scale

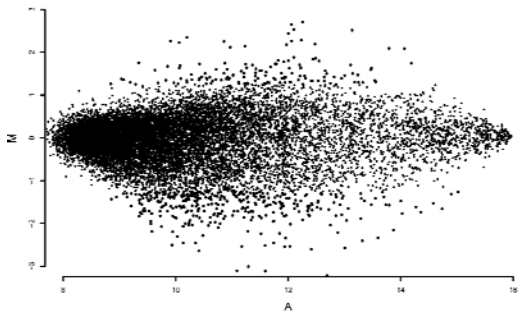
X_{gi} = $\log_2$ (Corrected probe intensity)

	Treatment A			Treatment B		
	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Gene 1	12.34	12.23	11.70	12.78	12.67	11.21
Gene 2	9.30	9.71	9.44	7.65	7.45	7.50
Gene 3	11.45	11.19	11.11	10.58	10.34	10.21
Gene 4	12.45	0.12	0.78	0.12	1.05	0.67
Gene 5	7.41	6.17	7.21	8.67	6.87	7.43
Gene 6	13.24	13.78	12.04	14.12	14.05	13.61
...	...	...	...	...	...	...

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics* 32 496-501.

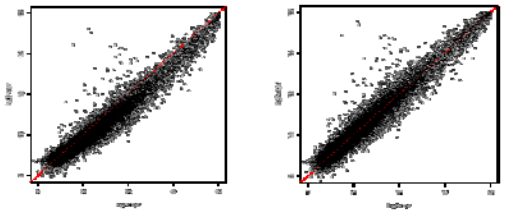
	Treatment A			Treatment B		
	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Gene 1	12.34	12.23	11.70	12.78	12.67	11.21
Gene 2	9.30	9.71	9.44	7.65	7.45	7.50
Gene 3	11.45	11.19	11.11	10.58	10.34	10.21
Gene 4	12.45	0.12	0.78	0.12	1.05	0.67
Gene 5	7.41	6.17	7.21	8.67	6.87	7.43
Gene 6	13.24	13.78	12.04	14.12	14.05	13.61
...	...	...	...	...	...	...

- The M-value is the average difference between treatments.
  - M>0 : more mRNA from treatment A (red)
  - M<0 : more mRNA from treatment B (green)
- The M-value is called the *log fold-change*.
- The A-value is the average *total intensity.*



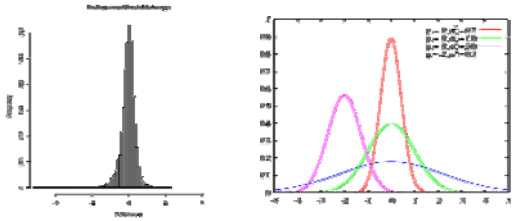
Normalization of microarray data

- Makes microarrays comparable - removes systematic trends and technical artifacts



Statistical modeling of microarray data

- Statistical models are used to describe randomness
- Common assumption: normal distribution

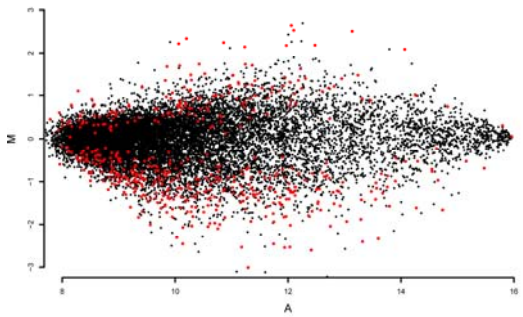


The t-statistic

- The t-statistic is defined as

$$T_g = \frac{M_g}{\sqrt{S_g^2 / n}}$$

- Advantages
  - Takes the variance of the genes into account.
  - We can calculate p-values.
- Disadvantages
  - Model assumptions? Are they correct?
  - Unreliable when few replicates are present!



Moderation of the variance

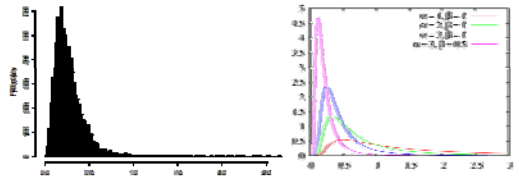
- The t-statistic assumes a normal distribution

$$X_{gi}^{A} \sim \text{Normal}(\mu_g^A, \sigma_g^2), \qquad X_{gi}^{B} \sim \text{Normal}(\mu_g^B, \sigma_g^2).$$

- The parameters μ_g^A, μ_g^B and σ_g^2 are unknown and needs to be estimated
- It is hard to estimate σ_g^2 with few replicates!

Moderation of the variance

- We can add prior information about σ_g^2 to make the estimation more exact!



- Assume that $\sigma_g^2 \sim \Gamma^{-1}(\alpha, \beta)$ (inverse gamma)
- Empirical Bayes model: α and β are estimated from the data.

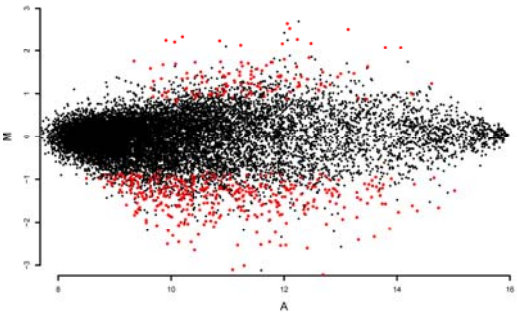
The moderated t-statistic

- The moderated t-statistic is defined as

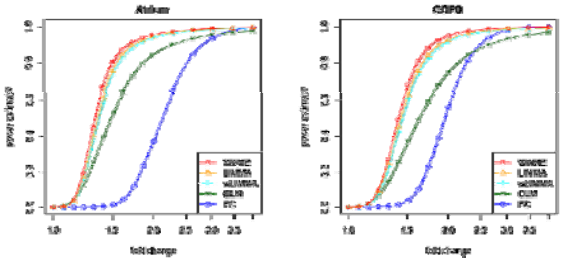
$$T_g^m = \frac{M_g}{\sqrt{S_g^{\,2}/(n-1) + 2\beta}}.$$

Moderation factor

- Robust – works well with few replicates.
- Have *n-1+2α* degrees of freedom. Extra data from the prior assumption!



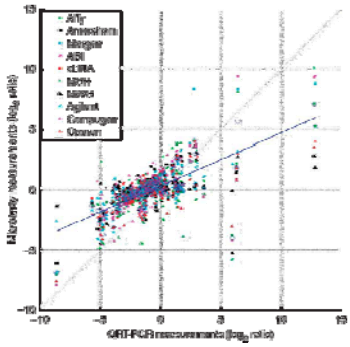
Gene ranking



Microarrays – how good are they?

- Early microarray studies resulted in a high error rate
- Recently, Kuo *et al.* evaluated the performance of 10 microarray platforms.
  - Reference material consisting of mRNA from cortex and retina in mouse.
  - PCR results from 150 genes were used as a "golden standard". The expression from the different platforms were compared against these genes.

General performance of microarrays



Kuo W.P. et al. (2006). A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature Biotechnology* 24 (7).

Why is there a difference?

- We might be measuring the same gene, but are we measuring the same piece of mRNA?

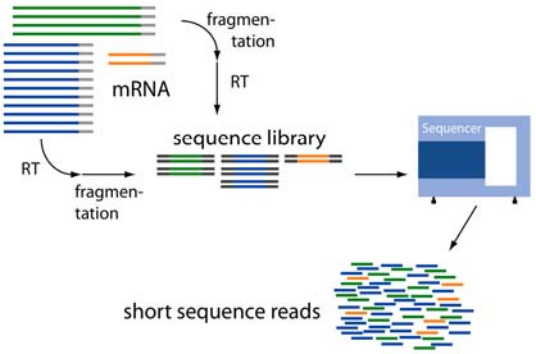
Why is there a difference?

- Disadvantages of microarrays
  - The properties of the probes is vital.
  - Cross-hybridization. Off-target binding causes non-trivial correlations in the data.
  - Alternative splicing makes probe position important.
  - Many complicated steps. The performance depends on the optimization of the protocol (e.g. cDNA-synthesis, fragmentation, hybridization temperature, washing, etc.) .

Gene expression analysis using RNA-seq

- RNA-seq is based on next generation DNA sequencing
- Modern alternative to microarrays
- Illumina and SOLiD are the most used sequencing technologies in RNA-seq

Gene expression analysis using RNA-seq



Data from RNA-seq data

- Data from RNA-seq comes as reads per gene

X_{gi} = Number of reads matching gene g

	Treatment A			Treatment B		
	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Gene 1	66489	29192	18643	21721	84669	80540
Gene 2	11288	2899	1062	6130	9581	17251
Gene 3	44979	12906	14604	10378	85043	39478
Gene 4	7133	4772	1124	319	6863	7286
Gene 5	34282	14379	13748	6133	12648	7620
Gene 6	6531	7184	1962	651	1334	13125
...	...	...	...	...	...	...
Total	17070232	5913427	9103289	4735558	15326223	12020031

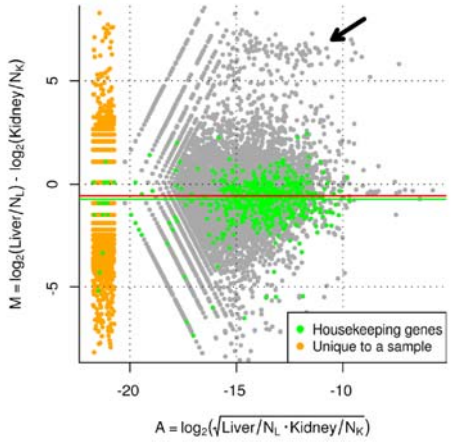
X_{gi} N_i

Normalization of RNA-seq data

- Normalization of RNA-seq data is necessary
- Naïve: Calculate the relative abundance

$$R_{gi} = \frac{X_{gi}}{N_i}$$

- Not good! High-expressed genes will affect the global expression level.



Normalization of RNA-seq data

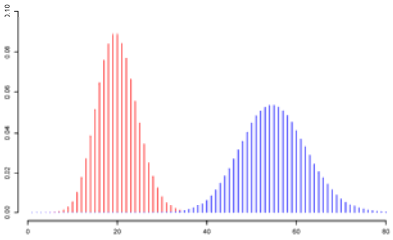
- Robust scaling
  - TTM – trimmed mean of M-values (Robinson & Oshlack 2010)
  - Robust scaling (Anders and Huber 2010)

$$R_{gi} = \frac{X_{gi}}{\hat{N}_i} \qquad \hat{N}_i = \underset{g}{\text{median}} \frac{X_{gi}}{\left(\prod_{i=1}^{m} X_{gi}\right)^{1/m}}$$

Statistical analysis of RNA-seq data

- Data from RNA-seq is discrete

$$X_{gi} \sim \text{Poisson}(\lambda_g) \qquad \text{E}(X_{gi}) = \lambda_g \qquad \text{Var}(X_{gi}) = \lambda_g$$



Statistical analysis of RNA-seq data

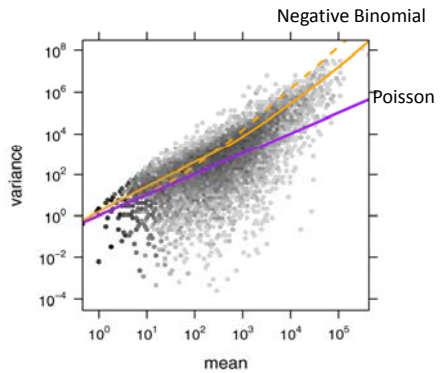
- Data from RNA-seq is overdispersed

$$\text{Var}(X_{gi}) > \text{E}(X_{gi})$$

- Data is there often modeled using a negative binomial distribution

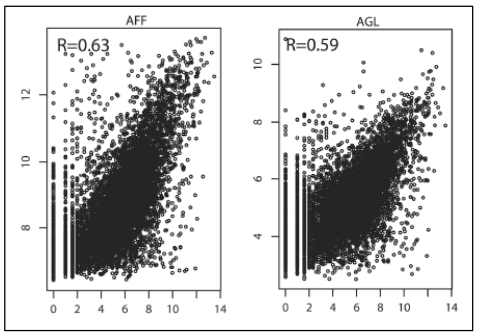
$$X_{gi} \sim \text{NegBin}(\mu_g, \phi)$$
$$\text{E}(X_{gi}) = \mu_g$$
$$\text{Var}(X_{gi}) = \mu_g + \phi\mu_g^2$$

Statistical analysis of RNA-seq data

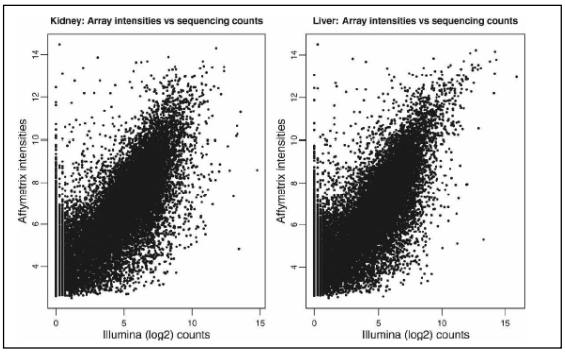


RNA-seq – better than microarrays?

- In general yes!
  - No probes!
  - No cross-hybridization!
  - Lower technical noise!
- However,
  - Problems with GC/AT-rich regions
  - A high sequencing depth is needed to accurately quantify low-expressed genes
  - Still a slightly higher cost
  - The statistics is currently more complicated



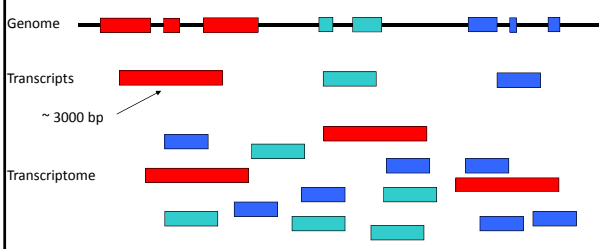
t'Hoen et al. 2008
Correlation is ~60%



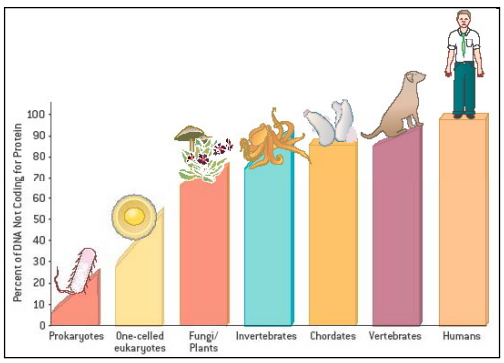
Marioni et al. 2008
Correlation is ~70%

Large-scale gene expression analysis

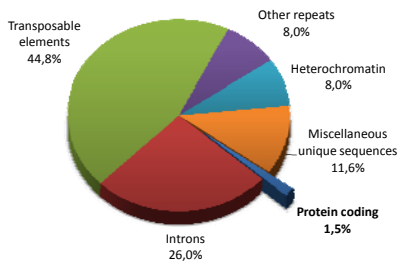
- mRNA quantification
  - Identification of differentially expressed genes
  - Techniques: microarray and RNA-seq
- *De novo* sequencing of mRNA
  - Identification of the sequence of genes
  - Techniques: RNA-seq

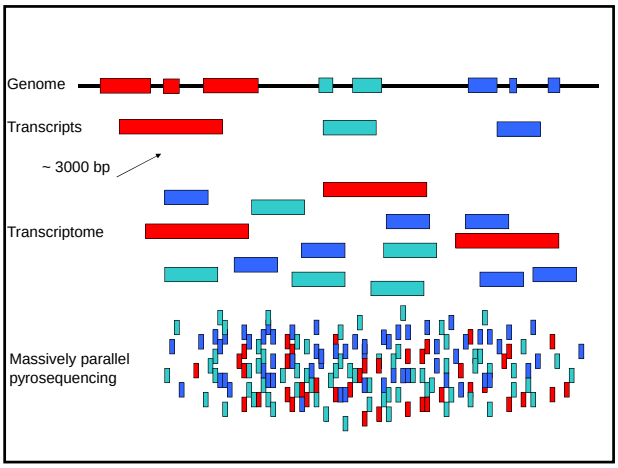


- Up to 50% of the human genome is transcribed.
- 5%-10% of the human genome is believed to be functional.
- < 2% of the human genome is coding and thus translated to a protein.

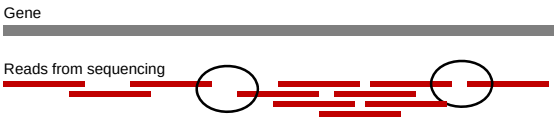


The human genome





How much data do we get?



- The reads are joined into contigs.
- Reads that can not joined with anything are called singlets.

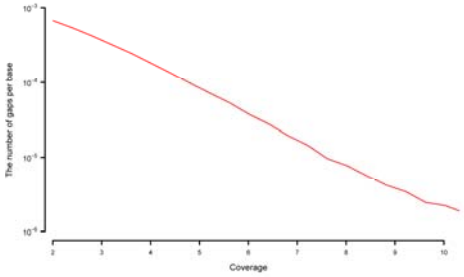
How much data do we get?

- The result from one run on a Genome Sequencer FLX using Titanium chemistry

1.000.000 reads
× 500 bases
500.000.000 bases

- The transcriptome of a higher eukaryote is up to around 50 million bases
- We can, theoretically, cover a transcriptome 10 times

How much data do we get?



Transcriptome sequencing

1. Pre-processing
Removes:
- Redundant 454-reads (ghost reads)
- 5'/3' sequencing vectors

2. Cleaning
Removes:
- PolyA-tails
- Simple repeats
- Bacterial contamination

3. Masking
Masks/removes:
- Low-complexity regions (Transposons, etc.)
- Contamination (rRNA, etc)

4. Assembly
- Contig assembly using megablast and CAP3.

Sequence data processing

>E6LSDQW02HPHGI
TGACTAAGATCCATCACATCAGGCCAGGTAGGAGTCTCTTATATTAGGTATCAATACCTTCCGGGT
GGATACCTTTGAGGCATAAGCTGGACAGGCACAGAACCTCGAGGCAGAACTTCCCGACTGCTTGAT
GTGTATCAAGGTCAATCAATCTGAAAATCAGCTGCCTAAGCACCAGTTCNNNNNNNNNNNNNNNNN
TTTGCTCAACTCCTCTTAGTAGCTGAGCGGGCTGGCAAGGC

>E5R7OVD09FMUGM
TCACTAACTGTAGNCACACACACACATCACACACACACACACACACACACACACACACACACACACAC
TAGNNNCACACACACACACACACACACACACACACACACACACACACACACACACACTACTATAATAAATAAAGAAGA
AGAAGTAGTTAGTTAGTACTTAACGTTAACGGTACGGTACGTAGGTACGGTAACCGGTAACCGGTA
ACCCGGAACCGTACGGTACGGTCGGTACGGTACGGTACGGTACGTACGTAACCGTTAAAAACCGGT
TTAAAAAGGTAAAAAGGGTAAAAAGGGTTAAAACGGGTTAACGGGTAACGTAGTAGNA

>E6LSDQW02GGDBU
TGACTAACAAATTTTAATTACACTTAAGGTGTATATTTTCTATGCAACCCATCAATTCAAGAGGTG
TAATGTGCTGATGACTATTTGTAATCGTTATACATTCTGACCCGAAGTCAGAAAGTATTTCTCTGT
CTGTGTGTTCACAGGCAGTGTGGTTGATTACATGAAATTCAGTACATTTGCAGTCTCGTTGCCCTT
CTCACCTGCCTTTCGTCATTACCGACGGTATTGAATTTCGTTTTCCCCGTTGGGGTTCTCCGGACA
AGGAG

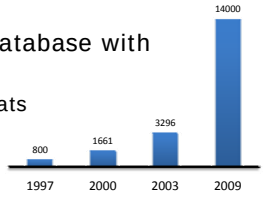
```
Sequences producing significant alignments:                      (bits) Value

Ecoligenome                                                        519   e-148

>Ecoligenome
          Length = 4639675

 Score =  519 bits (262), Expect = e-148
 Identities = 262/262 (100%)
 Strand = Plus / Minus


Query: 8       caaatttttaattacacttaaggtgtatattttctatgcaacccatcaattcaagaggtgt 67
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4566813 caaatttttaattacacttaaggtgtatattttctatgcaacccatcaattcaagaggtgt 4566754


Query: 68      aatgtgctgatgactatttgtaatcgttatacattctgacccgaagtcagaaagtatttc 127
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4566753 aatgtgctgatgactatttgtaatcgttatacattctgacccgaagtcagaaagtatttc 4566694


Query: 128     tctgtctgtgtgttcacaggcagtgtggttgattacatgaaattcagtacatttgcagtc 187
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4566693 tctgtctgtgtgttcacaggcagtgtggttgattacatgaaattcagtacatttgcagtc 4566634


Query: 188     tcgttgccctttctcacctgcctttcgtcattaccgacggtattgaatttcgttttccccg 247
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4566633 tcgttgccctttctcacctgcctttcgtcattaccgacggtattgaatttcgttttccccg 4566574


Query: 248     ttggggttctccggacaaggag 269
                |||||||||||||||||||||
Sbjct: 4566573 ttggggttctccggacaaggag 4566552
```

Sequence cleaning

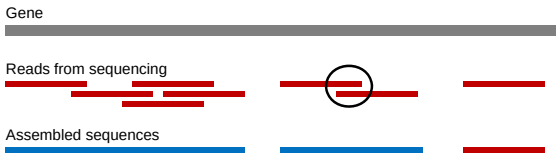
- Contamination
  - mRNA from other types of species
  - rRNA or other unwanted types of RNA
- Repetitive elements
  - polyA-tails
  - Simple Sequence Repeats (SSR)
  - More complex repeats like SINEs, LINEs and transposons
- Repetitive elements are typically located outside coding regions

Sequence cleaning

- RepeatMasker is a tool for identification of repetitive elements
  - *ab initio* prediction of repeats
  - database matching
- Repbase Update is a database with
  - Transposable elements
  - Simple Sequence Repeats
  - Pseudogenes

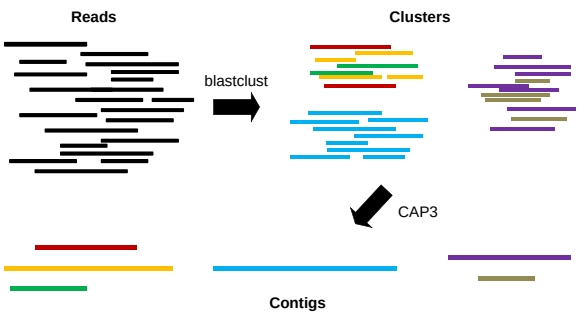


Assembly



- Similarity threshold
  - Less strict setting results in longer contigs with more errors
  - More strict setting results in shorter contigs with fewer errors

The Gene Indices Clustering Tools



A few existing projects and their results

	Sequencing			Assembly		
Project	System	Reads	Length	Algorithm	Contigs	Length
Barrel clover	GS20	300,000	110	Custom	34,000	?
Glanville fritillary butterfly	GS20	600,000	110	Custom	48,000	197
Largemouth bass	GS20	550,000	105	Newbler	33,000	?
Eucalyptus	GS20+FLX	630,000+ 400,000	105+ 210	Newbler	71,000	247
Coral larva	FLX	630,000	233	Custom+ Newbler	44,000	440
Flesh fly	FLX	210,000	241	Custom+ Newbler	21,000	332
Viviparous eelpout	FLX	400,000	237	Custom	36,000	395
Bank vole	FLX Titanium	1,000,000	305	Custom	64,000	481

Annotation

- Functional similarity from sequence similarity
- Assign information to the assembled transcripts
  - Gene description
  - Functional annotation (e.g. pathways)

GenBank UniProt ensembl

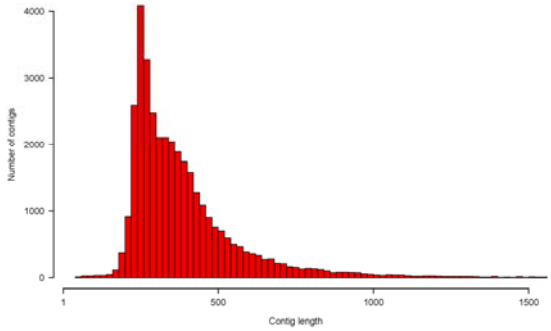
Case study: Sequencing of the transcriptome of *Zoarces viviparus*

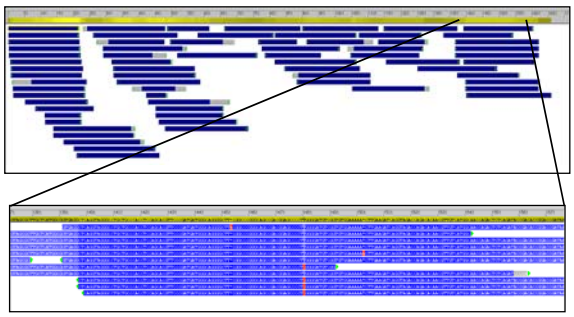
- The BALCOFISH project
- No suitable model species
- *Zoarces viviparus* (eelpout)
  - Stationary
  - Gives birth to live young
- Large-scale gene expression assays in eelpout
  - Sequencing of the liver transcriptome
  - Design of an eelpout microarray

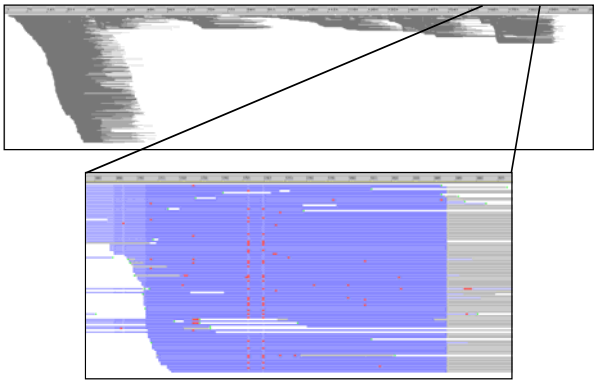
Assembly results and statistics

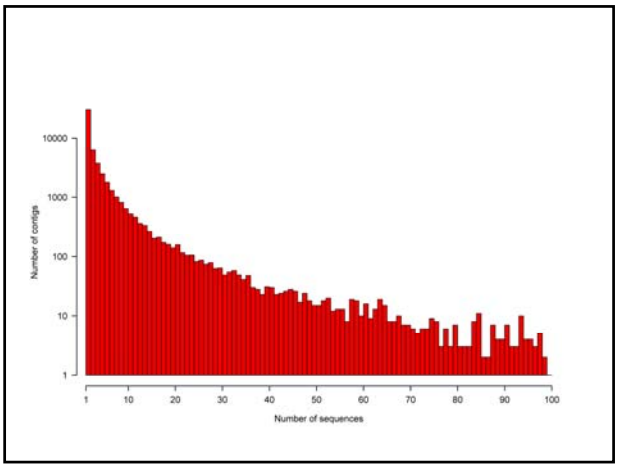
- Massively parallel pyrosequencing
  - 400.000 reads with an average length of 237 bases
  - 90 million bases in total

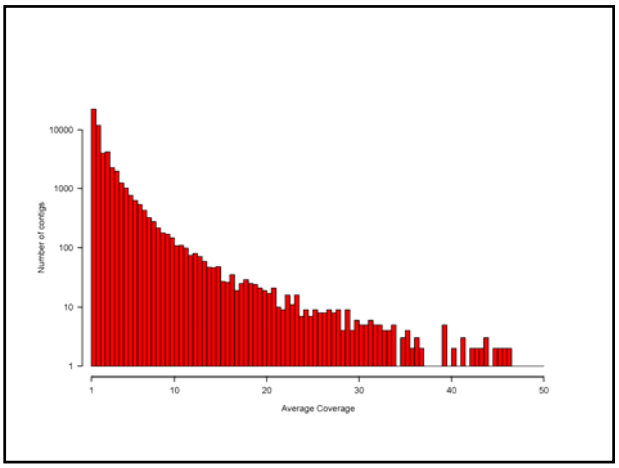
	Contigs	Singlets	Total
Number of sequences	36,110	17,347	53,457
Number of bases	14,250,156	4,050,061	18,300,217
Average length	395	233	342
Average coverage	3.46	1	2.67
Annotated	89.2%	87.3%	88.6%





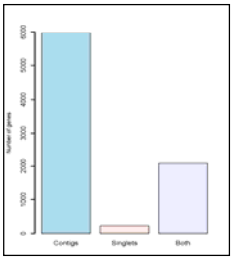






Depth and coverage

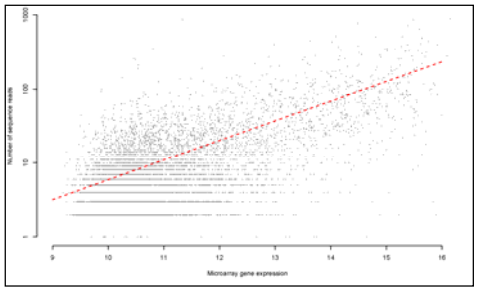
- The size of the stickleback transcriptome is ~30 million bases.
- The 18 million bases covers ~40% of the eelpout transcriptome.
- The eelpout sequencing is deep
  - Matches ~8,000 stickleback genes (15,000 genes in total).
  - Few eelpout stickle back genes are represented by singlets.



Gene	Pyrosequencing		Genbank	
	Accession	Length	Accession	Length
Vitellogenin	ZOVI0010766	1,826	AJ416326	1,229
Zona Pelucida 2	ZOVI0014264	1,100	-	-
Zona Pelucida 3	ZOVI0034606	989	-	-
Estrogen receptor	ZOVI0044876	852	AY223902	3,256
Metallothionein	ZOVI0049137	363	X97270	312
Heat-shock protein 70	ZOVI0038668	1,460	-	-
Heat-shock protein 90	ZOVI0020982	938	-	-
Cytochrome P450 1A	ZOVI0005392	1,652	-	-
Superoxide dismutase	ZOVI0007529	747	-	-
Glutathione peroxidase	ZOVI0037346	1,208	-	-

Gene expression analysis using high-throughput sequencing

- The correlation for the eelpout microarray was ~60%.



Kristiansson *et al.* 2009 Characterization of the Zoarces viviparus transcriptome using massively parallel pyrosequencing.

Conclusions

- Massively parallel pyrosequencing provides means for fast and cost-efficient *de novo* transcriptome sequencing.
- One full run on a 454 sequencer is enough to cover a substantial part of the transcriptome from a higher eukaryote.
- Bioinformatics competence and computational resources are needed to assemble the generated data into transcripts.