1. Introduction
   - Gene regulation
   - Genomics and genome analyses

2. Gene regulation tools and methods
   - Regulatory sequences and motif discovery
   - Transcription factor binding sites
   - Databases

3. Technologies
   - Microarrays
   - Deep sequencing and applications
   - Single cell RNAseq and spatial transcriptomics

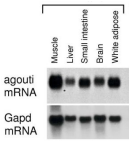
4. Clustering
   - Unsupervized clustering (HCA,K-means, PCA, SOM)
   - Supervized clustering (classification)

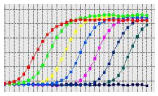
5. Gene ontology, Pathways, Enrichment analysis
   - Databases and tools
   - Gene set enrichment analysis

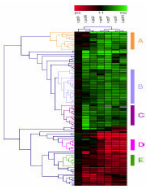
6. Biomolecular networks
   - Small world networks
   - Topology and parameter
   - Network motifs

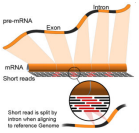
RNA expression profiling

- Northern bloting
  - semi-quantitative
  - few genes
- Real time RT-PCR (qPCR)
  - medium throughput
  - 96/384 per run
- Microarray analysis
  - high throughput
  - 10.000-500.000 elements per chip
- RNA seq
  - high throughput
  - deep sequencing (short reads 100bp)





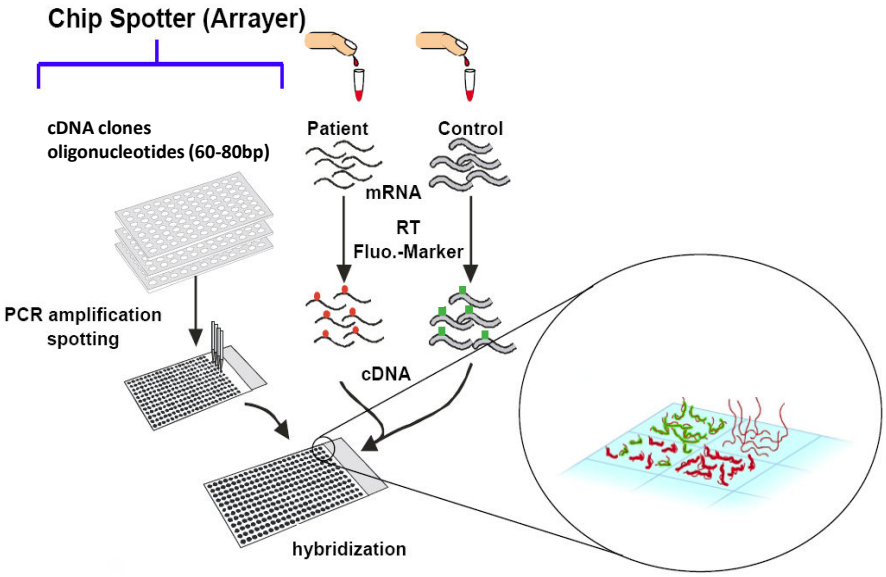




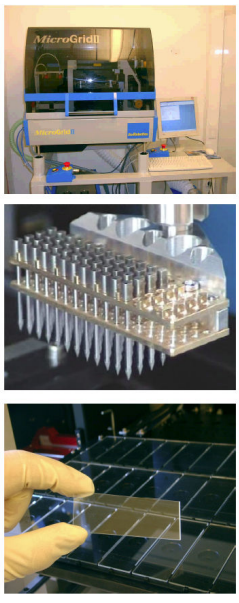
Microarray technology

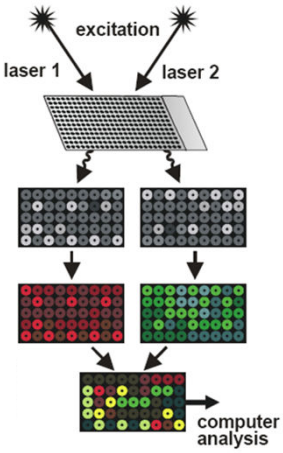
- Two-color microarrays (Custom, Agilent)
- One-color microarrays (Affymetrix)

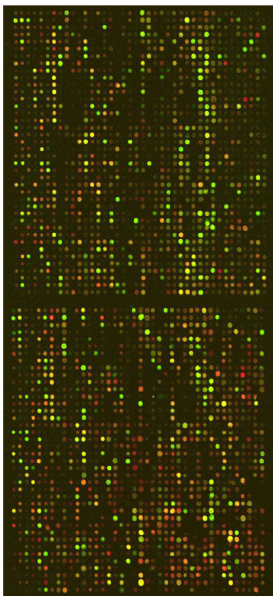
Two-color microarrays



Two-color microarrays







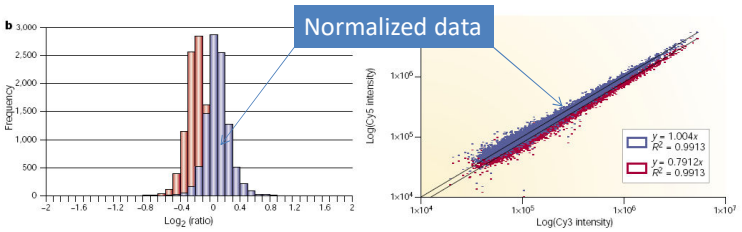
Microarray analysis

- Image analysis (grid alignment, spot vs.background)
- Background correction
- Global normalization
- Log2 transformation (log2-ratio)

Normalization

- Remove bias (e.g. different properties of dyes)
- Basic assumption is that most of the genes are not changing their expression during the studied process
- Same amount of (total) RNA in both channels.

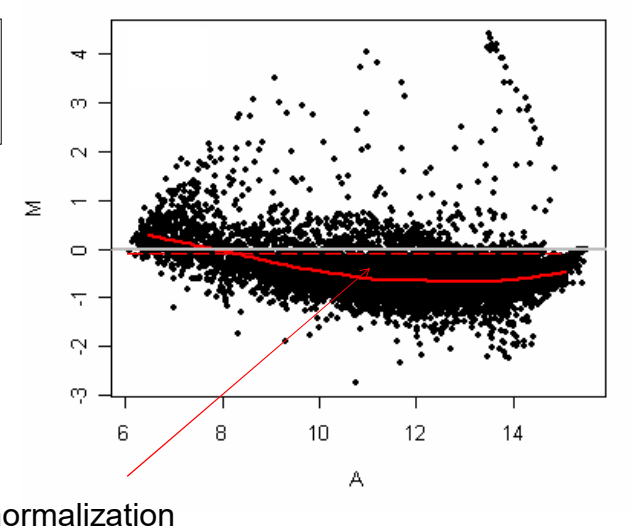
$$N = \frac{\sum_{k=1}^{N_{array}} R_k}{\sum_{k=1}^{N_{array}} G_k} \qquad G_k' = NG_k \text{ and } R_k' = R_k .$$



MA plot

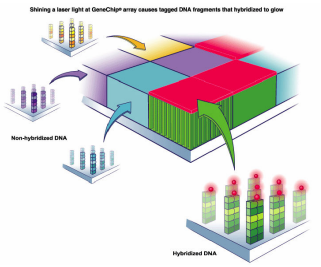
$M = \log_2(R/G)$

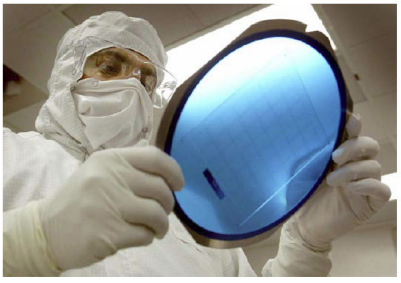
$A = \log_2(R*G)/2$



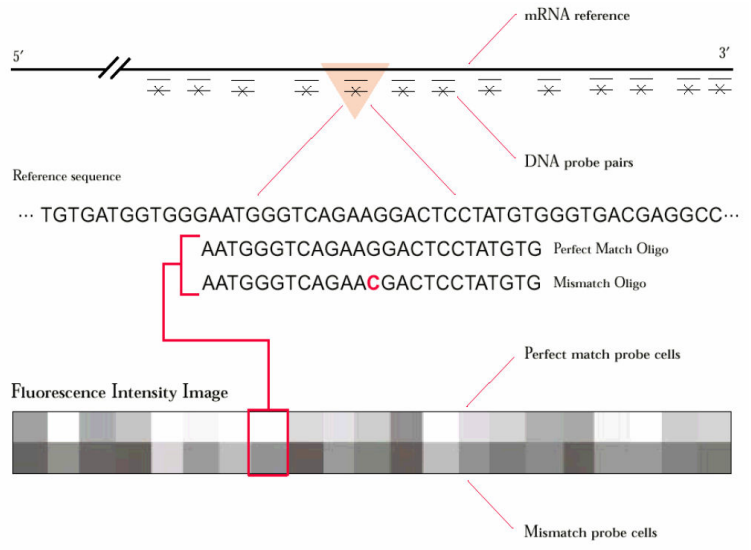
Lowess normalization

One color microarrays (Affymetrix)





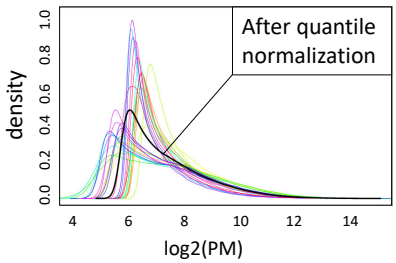
Affymetrix chips



Processing of Affymetrix chips

Robust Microarray Averaging (R/Bioconductor pkg. affy)

- Background modeling (PM vs. MM)
- Quantile normalization across all arrays



- Probe summarization (median polish)
- Log2-transformation (log2-intensities)

Differentially expressed genes

test

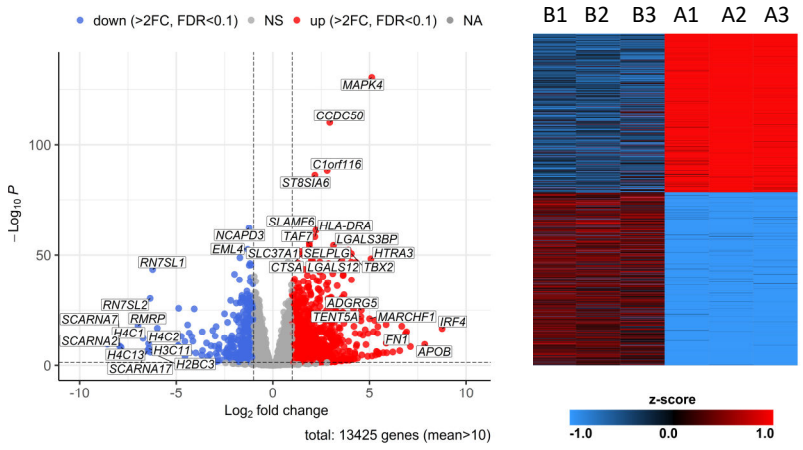
16134 probesets

ID	GENE	KO1	KO2	KO3	WT1	WT2	WT3	logFC	AveExpr	t	P.Value	adj.P.Val
10386473	Srebf1	5.72	5.58	6.06	4.91	4.88	5.09	0.83	5.33	7.66	3.7E-09	4.6E-05
10463355	Scd2	6.63	6.26	6.92	5.13	4.77	5.01	1.64	5.59	7.52	5.6E-09	4.6E-05
10548105	Ccnd2	5.56	5.48	5.49	5.05	5.11	5.02	0.45	5.23	5.21	7.3E-06	3.9E-02
10587284	Elovl5	5.81	5.67	5.97	5.05	5.06	5.35	0.66	5.44	4.87	2.1E-05	8.4E-02
10540122	Slc6a6	7.27	7.16	7.35	6.75	6.81	6.71	0.50	7.04	4.80	2.6E-05	8.5E-02
10605437	Pls3	5.50	5.63	5.41	4.88	4.93	4.87	0.62	5.20	4.63	4.3E-05	9.7E-02
10543791	Podxl	7.30	7.03	7.08	6.31	6.52	6.33	0.75	6.59	4.61	4.6E-05	9.7E-02
10356084	Irs1	8.30	8.76	7.61	6.62	7.33	7.19	1.18	7.60	4.57	5.2E-05	9.7E-02
10346164	Sdpr	5.68	5.37	5.43	5.00	5.03	4.95	0.50	5.17	4.54	5.7E-05	9.7E-02
10387625	Chrnb1	6.31	6.08	6.06	5.73	5.59	5.81	0.44	6.01	4.52	6.0E-05	9.7E-02
10407390	Ptbp1	4.84	5.26	5.07	4.22	3.98	4.64	0.77	4.88	4.43	8.0E-05	1.1E-01
10507539	Elovl1	5.08	4.58	4.89	4.33	4.34	4.55	0.44	4.61	4.40	8.7E-05	1.1E-01
10585988	Myo9a	4.05	4.00	4.01	3.50	3.64	3.79	0.38	3.93	4.39	9.1E-05	1.1E-01
10371959	Elk3	5.94	5.85	5.78	5.28	5.44	5.46	0.47	5.66	4.38	9.3E-05	1.1E-01

condition KO vs. condition WT

Differentially expressed genes

Condition A vs. B



Differentially expressed genes

Moderated t-test (R/Bioconductor package *limma*)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$ => p-value

estimated from all genes

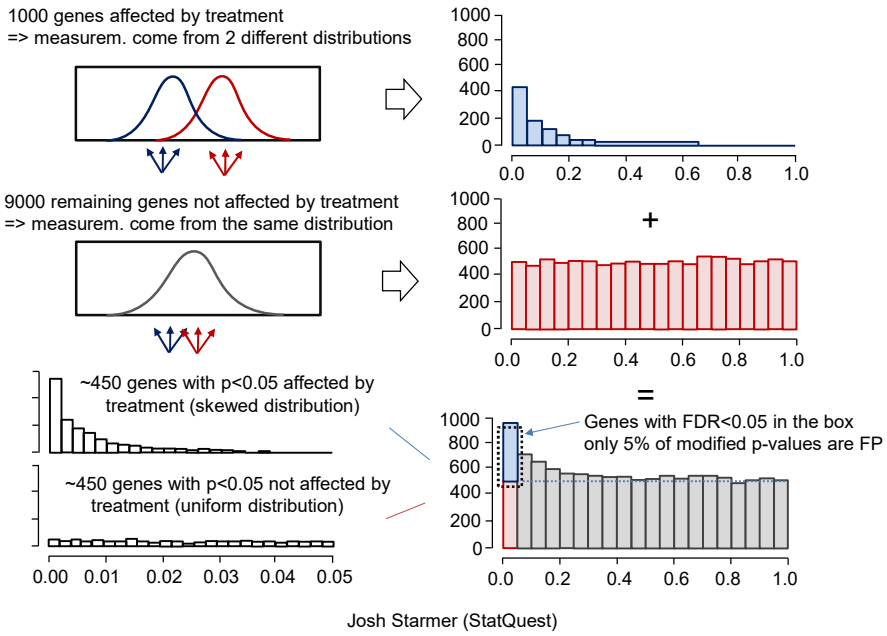
- At a significance level of 0.05 in the case of 10000 tests 500 might be wrong.
- Account for this by correction for multiple hypothesis testing
  - Bonferroni correction (multiply p with number of tests)
  - Benjamini-Hochberg correction (based on the FDR)
- adjusted p-value<0.05 (<0.1) significantly differentially expressed

Methods to correct p-values for multiple testing

	Ranked p	Bonferroni	Benjamini-Hochberg (FDR)	
smallest p →	$p_{(1)}$	$p_{(1)}$ *n	$p_{(1)}$ *n	
	$p_{(2)}$	$p_{(2)}$ *n	$p_{(2)}$ *n/2	
	..	..	..	
	$p_{(i)}$	$p_{(i)}$ *n	$p_{(i)}$ *n/i	..
	..	..	..	
	$p_{(n-1)}$	$p_{(n-1)}$ *n	$p_{(n-1)}$ *n/(n-1)	keep smaller one
largest p →	$p_{(n)}$	$p_{(n)}$ *n	$p_{(n)}$	

$$p_{(i)}^{BH} = \min\{ \min_{j \geq i} \{p_{(j)} * n/j \}, 1 \}$$

P-value distribution



Josh Starmer (StatQuest)

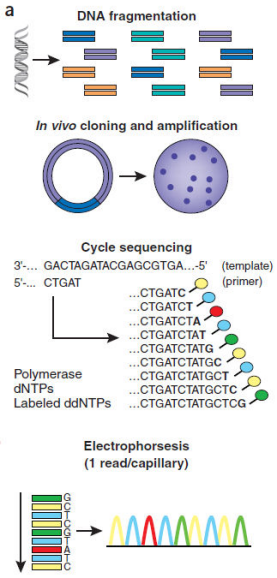
RNAseq vs Microarrays

- Potential for surveying the entire transcriptome, including novel, un-annotated regions.
- Helps to identfy expression and function of regulatory none-coding RNAs (e.g. lincRNA)
- Potential for determining gene structure and isoform level expression using reads mapping to splice junctions.
- Potential for making better presence/absence calls on regions.
- Don‘t need to design probes
- Higher computational and bioinformatics effort
- More expensive than microarrays

Deep (next generation) sequencing technologies

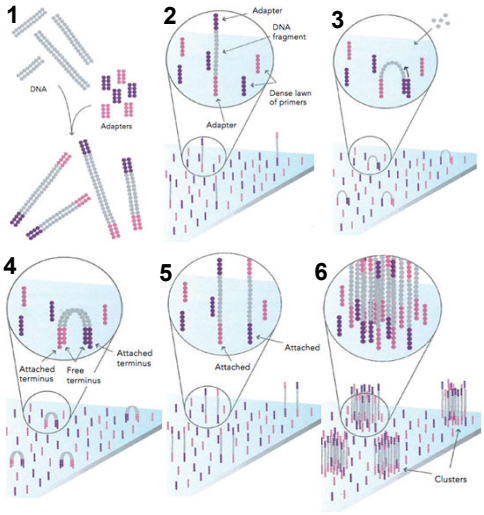
- Sanger (Thermo Fisher Scientific) — 1st gen.
- 454 (Roche)
- Solexa (Illumina)
- Solid (Thermo Fisher Scientific)
- Ion Torrent (Thermo Fisher Scientific) — 2nd gen. (ampl)
- HeliScope (Helicos)
- Pacific Biosciences SMRT
- Oxford Nanopore Sequencing (MinION) — 3rd gen. (no ampl)

Sanger sequencing



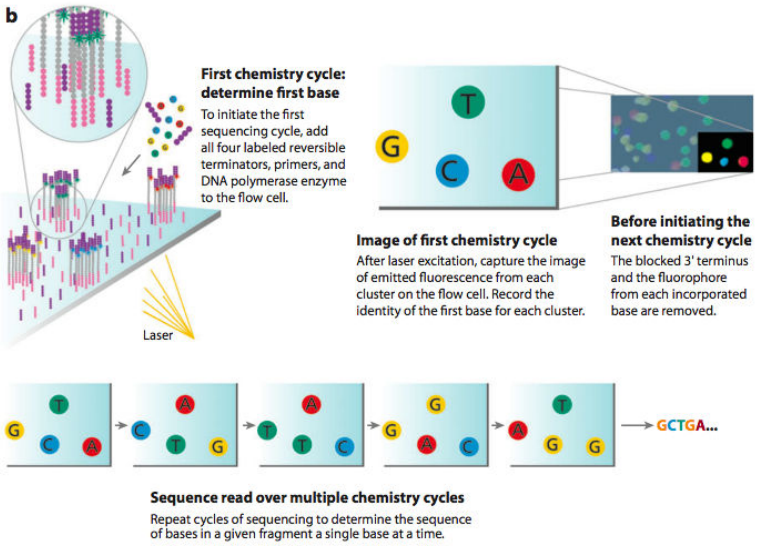
- DNA is fragmented
- Cloned to a plasmid vector, transform bacteria, growth (automated colony picking)
- Using fluorophore labeled dideoxy nucleoside triphosphates (ddNTPs) for chain termination
- Fluorescent readout with capillar electrophoresis (up to 384 capillaries)

Solexa (Illumina)

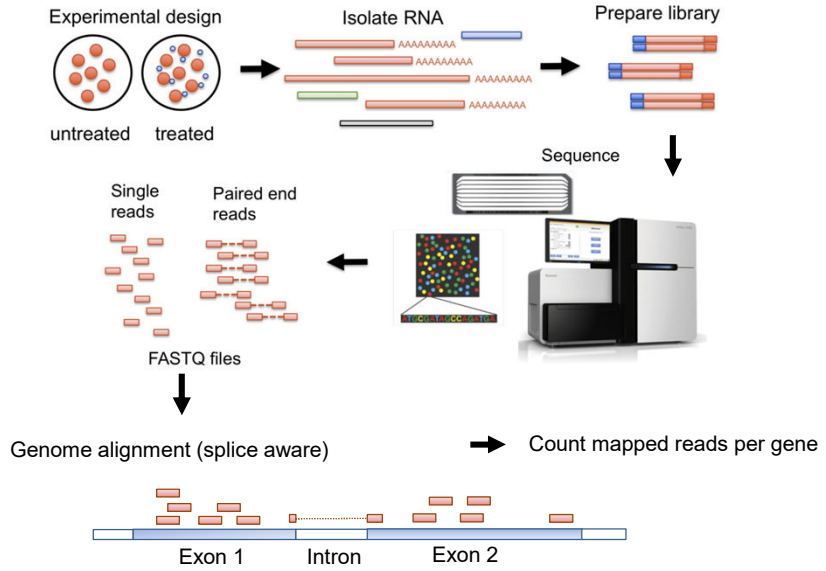


1. Prepare genomic DNA sample
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature double stranded DNA
6. Complete amplification

Solexa (Illumina)



RNAseq



Analysis steps

0. Image analysis and base calling (Phred quality score)

=> FastQ files (sequence and corresponding quality levels)

1. Trimming adaptors and low quality reads
2. Read mapping (Spliced alignment) (STAR)

=> SAM/BAM files

3. Transcriptome reconstruction (reference transcriptome, GTF file)
4. Expression quantitation (transcript isoforms) (featureCounts)
   -count reads
   -normalization
4. Differential expression analysis (negative-binomial test)
   (DESeq2, edgeR)

Phred Quality Score

Q=-10* log P

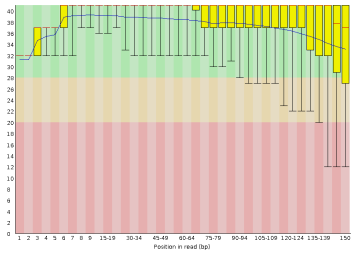
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

fastq format

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

← Q in ASCII

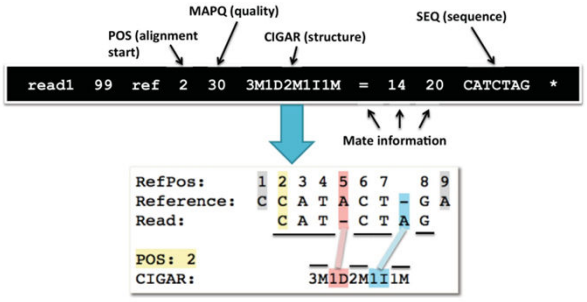
Quality of Sequencing (FASTQC)



SAM/BAM files

Sequence Alignment Map / Binary Alignment Map
SAM.. human readable, BAM.. binary representation, compressed

Header … metadata (sample and read group information)
Record … structured read information (1 line per read record)



Differential gene expression analysis

DESeq2 – generalized linear model (GLM) with NB distribution

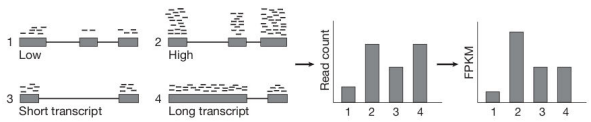
$$\log(\mu_{ij}) = \log(s_j) + x_j\beta_i$$

x_j = 0 if j is control, x_j = 1 if j is treatment sample (design matrix)
μ_{ij} … expected mean count for gene *i* and sample *j*
s_j … size factor for sample *j* (normalization)

Calculate the coefficients β that best fits the data.
Wald test => p-value

Normalization

- Reads per kilobase per million reads (RPKM)
- Fragments per kilobase per million (FPKM) for paired-end seq.



- TPM (transcripts per million) (better as totals are equal between samples)
- Quantile normalization (upper quantile normalization)
- TMM (trimmed mean of M values) (edgeR)
- Relative log expression (RLE) (DESeq2)

RPKM (FPKM)

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1
Tens(Mio)	3.5	4.5	10.6

1. Divide by millions of reads

RPM

A (2kb)	2.86	2.61	2.83
B (4kb)	5.71	5.43	5.66
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.09

2. Divide by gene length in kb

RPKM

A (2kb)	1.43	1.30	1.42
B (3kb)	1.43	1.36	1.42
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.01

TPM

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

1. Divide by gene length in kb

A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Tens(Mio)	1.5	2.025	4.51

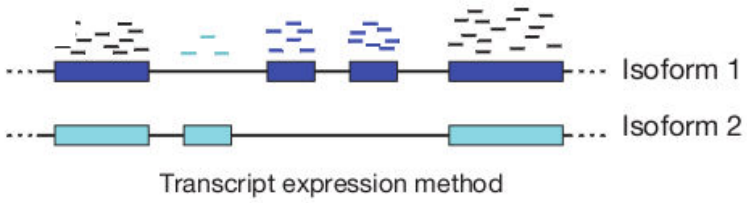
RPK

2. Divide by millions of RPK

A (2kb)	3.33	2.96	3.326
B (3kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

TPM

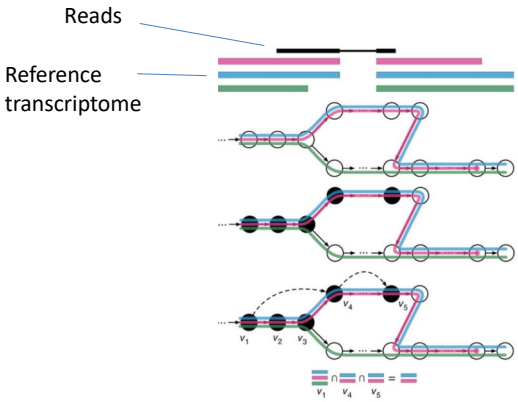
Isoform quantification



Transcript expression method

- Uncertainy in assigning reads to isoforms
- Paired-end sequencing
- Spliced alignment
- Alternative splicing (statistical significant?)

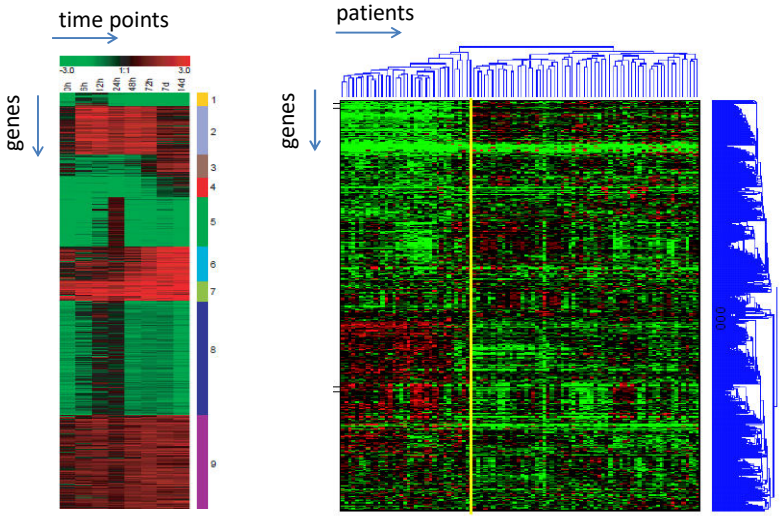
RNA seq quantification using pseudoalignment (kallisto)



Transcriptome de Bruijn Graph (T-DBG) where nodes (v1, v2, v3, ...) are *k*-mers

Bray et al. Nature Biotechnology 2016

Gene expression profiling



cell development

cancer