Statistical analysis of modern sequencing data – quality control, modelling and interpretation

Jörg Rahnenführer

Technische Universität Dortmund, Fakultät Statistik
Email: rahnenfuehrer@statistik.tu-dortmund.de

fdm Seminar
Freiburg Center for Data Analysis and Modeling
Freiburg, 23.01.2015

Research Overview

- Statistical methods in bioinformatics and medicine
  - High-dimensional gene expression data
    - survival analysis models for cancer (prognostic signatures)
    - toxicology (time series and concentration series)
    - distribution shapes, networks, …
  - Proteomics (phylogenetics based on mass spectrometry)
  - Linguistics (analysis of newspaper articles)
  - Efficient algorithms for statistical learning

Model selection

Bias/Variance – „Model fit“ vs. „Model complexity“

1 gene 20.000 genes

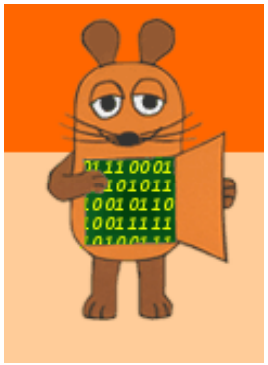
Interdisciplinary character

Computer Science – Bridge between Biology and Statistics

Biology

- Questions
- Experiments
- Interpretation

Statistics



- Model, Design
- Estimates, Tests
- Significance

Computer Science

- Efficient algorithms
- Visualization
- User-friendly tools

Next-generation sequencing (NGS)

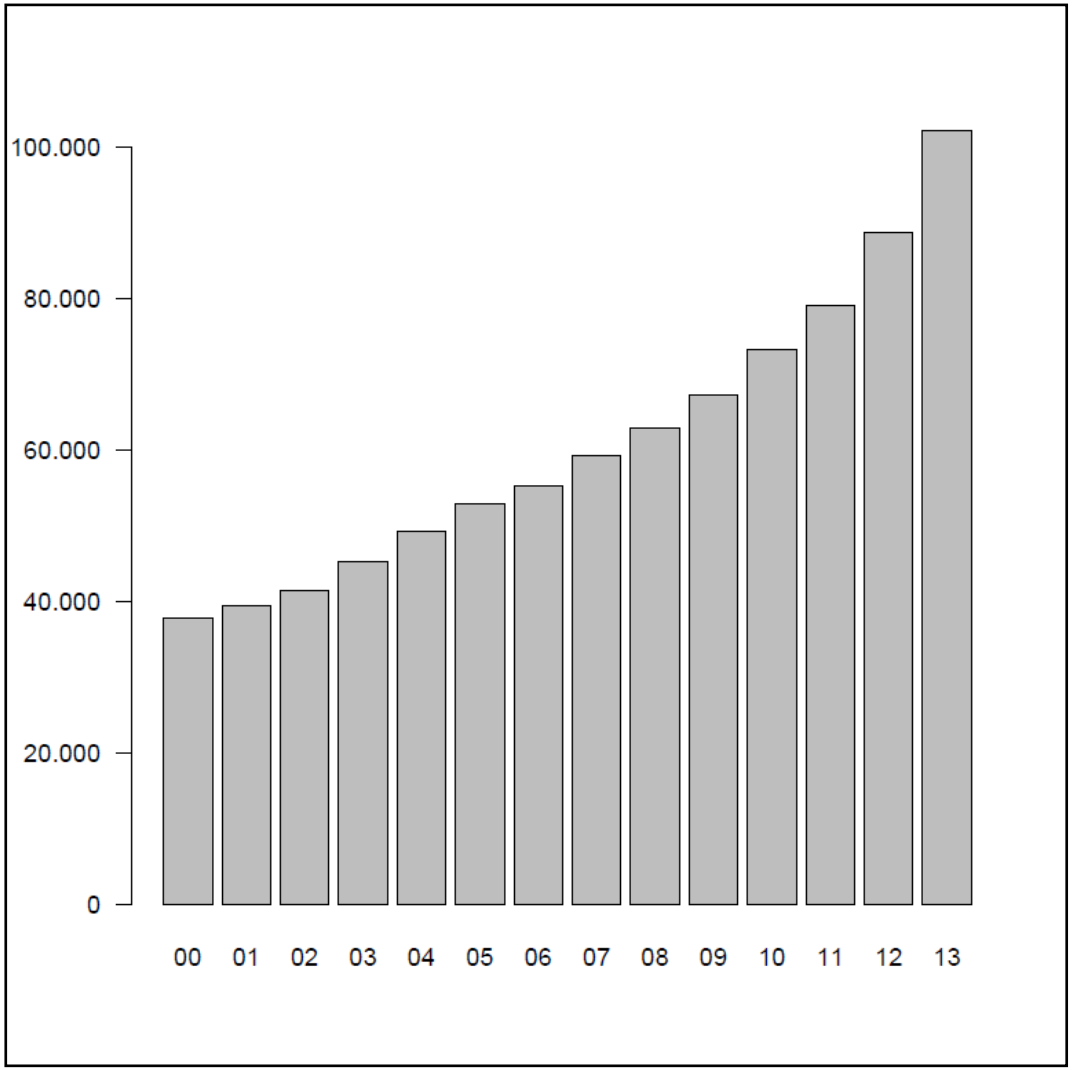
- Sequencing is an analytical key technology for reading bases in DNA or RNA
  - 1977 first sequencing technique by Sanger
  - since 2008 NGS methods
- Compared to Sanger sequencing NGS is much faster and cheaper, but still relatively expensive
- Applications
  - Analysis of genome (parts or whole)
  - Analysis of RNA transcripts for gene expression (RNASeq)
  - Profiling of mRNAs, microRNAs, chromatin structures and DNA methylation patterns
  - Analysis of transcription factors (ChIP-Seq, Protein-DNA-Interaction)

RNAseq technology

- „Whole transcriptome shotgun sequencing“
- Determination of nucleotide series in RNA with *next generation sequencing*
- Translation of RNA in cDNA to apply method of DNA sequencing
- RNAseq provides more information than microarray technology
  - Discrimination between alleles of a gene
  - Posttranscriptional modifications
  - Identification of fusion genes
- Pros and cons
  - Lower background noise
  - Higher resolution & reproducibility for technical and biological replicates
  - Relatively expensive
  - No established analysis pipeline

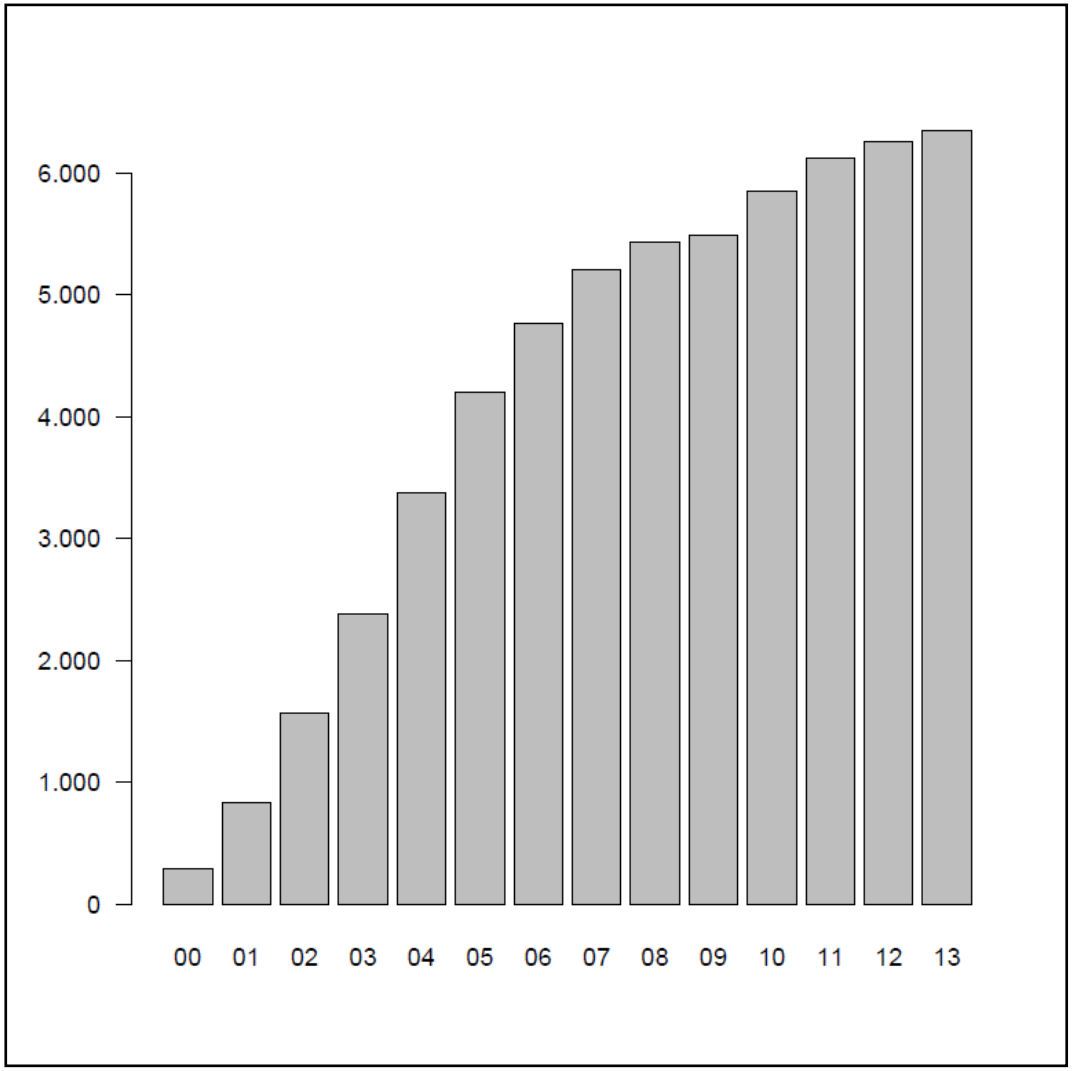
Impact – number of publications

- Absolute number of publications in Pubmed per year (2000-2013)
- Search term **cancer**



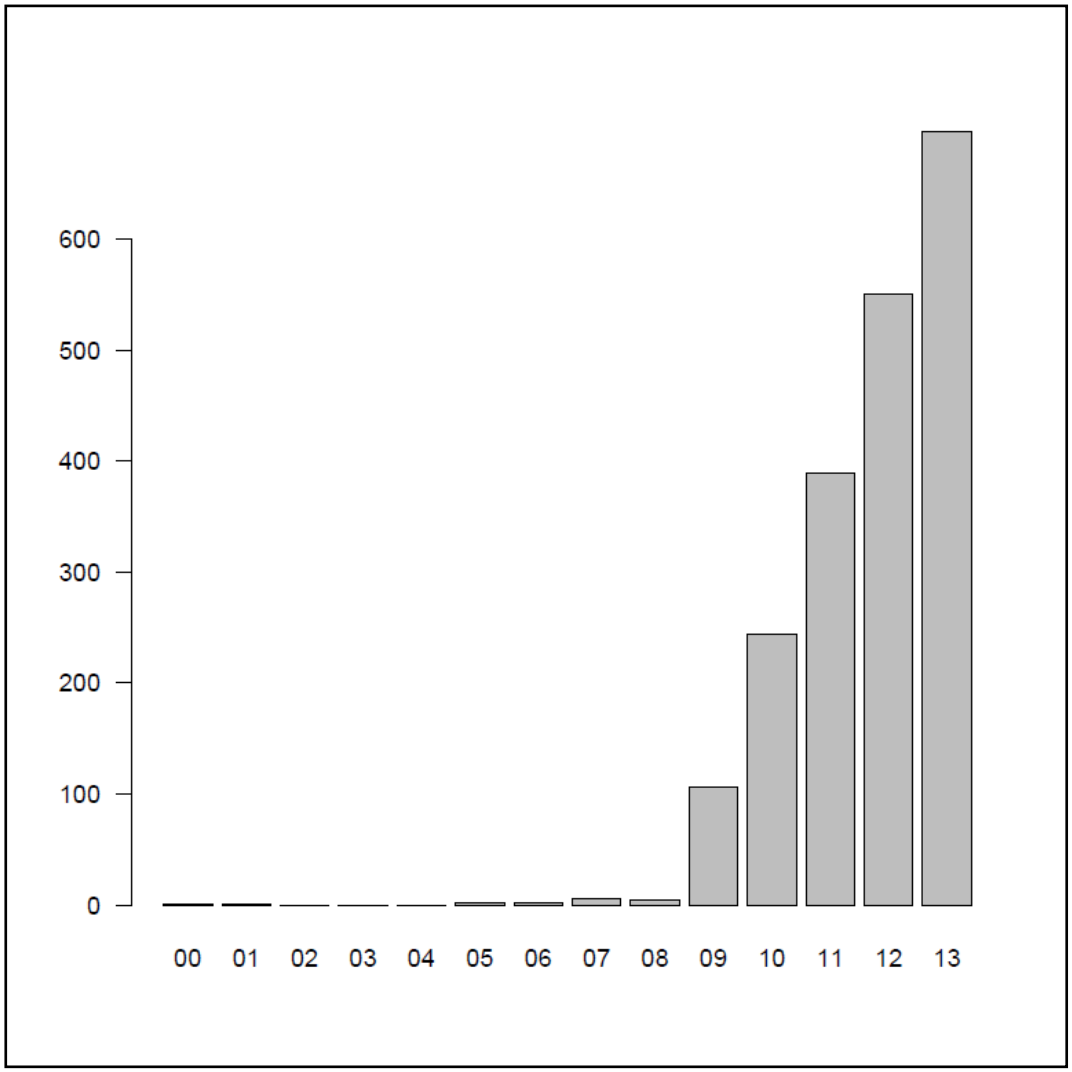
Impact – number of publications

- Absolute number of publications in Pubmed per year (2000-2013)
- Search term **microarray**
- Still large numbers



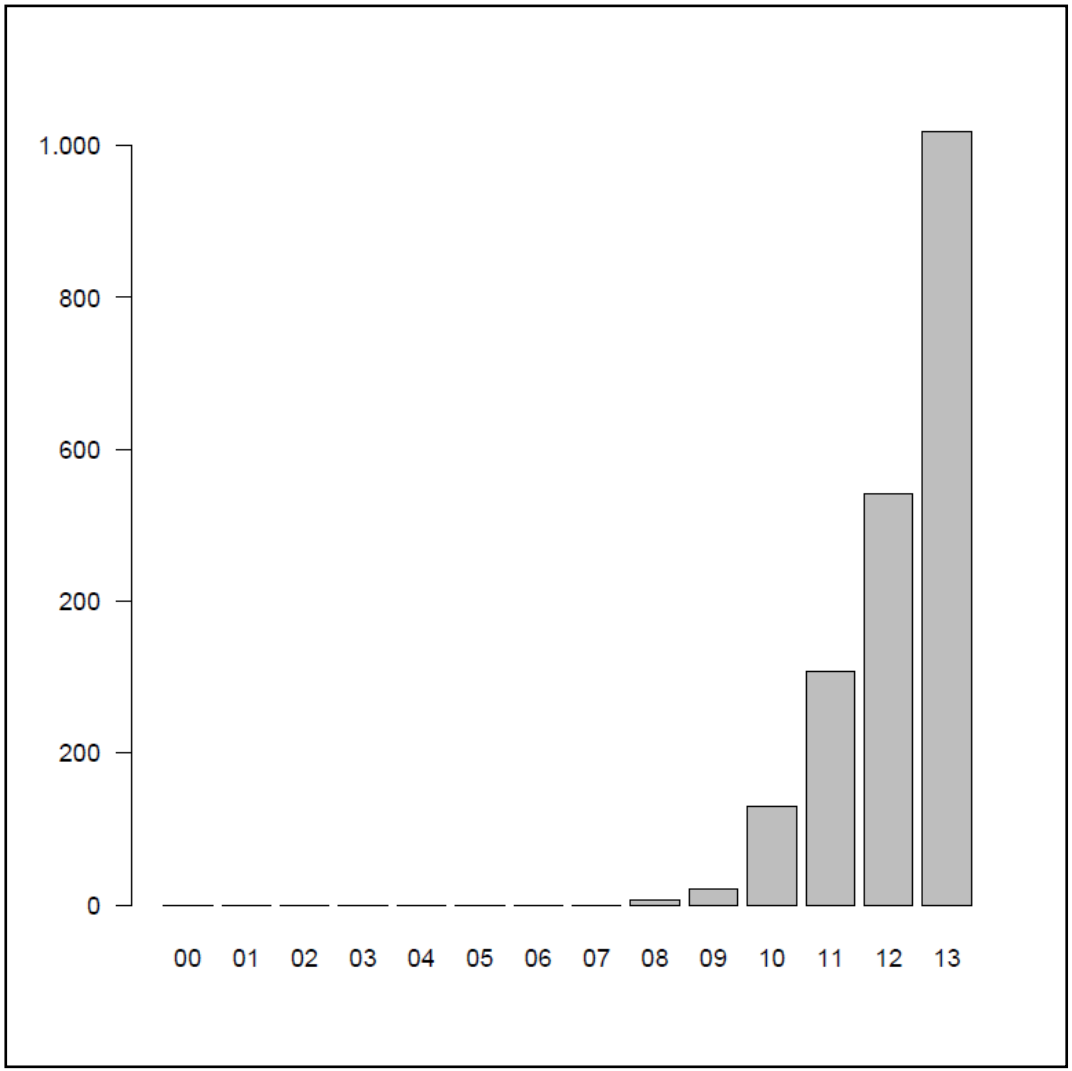
Impact – number of publications

- Absolute number of publications in Pubmed per year (2000-2013)
- Search term **deep sequencing**



Impact – number of publications

- Absolute number of publications in Pubmed per year (2000-2013)
- Search term **RNAseq**
- Little overlap with „deep sequencing“

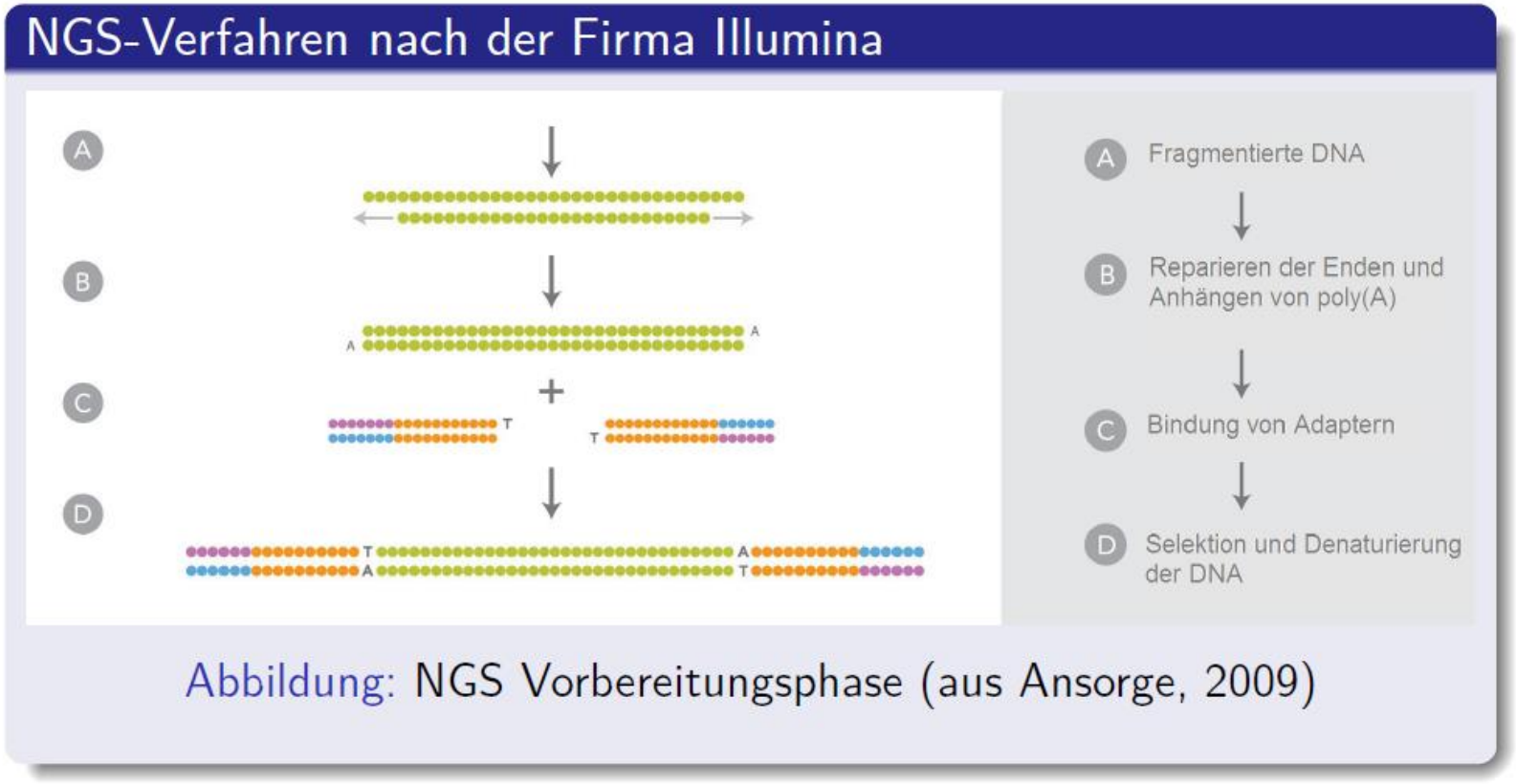


RNAseq technology

- Measuring abundance of mRNA with the help of sequencing techniques
- General procedure
  - Start with mRNA of sample
  - Random fragmentation of mRNAs
  - Backward transcription into cDNA
  - PCR and sequencing of cDNA to obtain reads
  - Assignment of reads to genes (position on genome sequence)
  - Result: Count table of reads assigned to genes

RNAseq technology

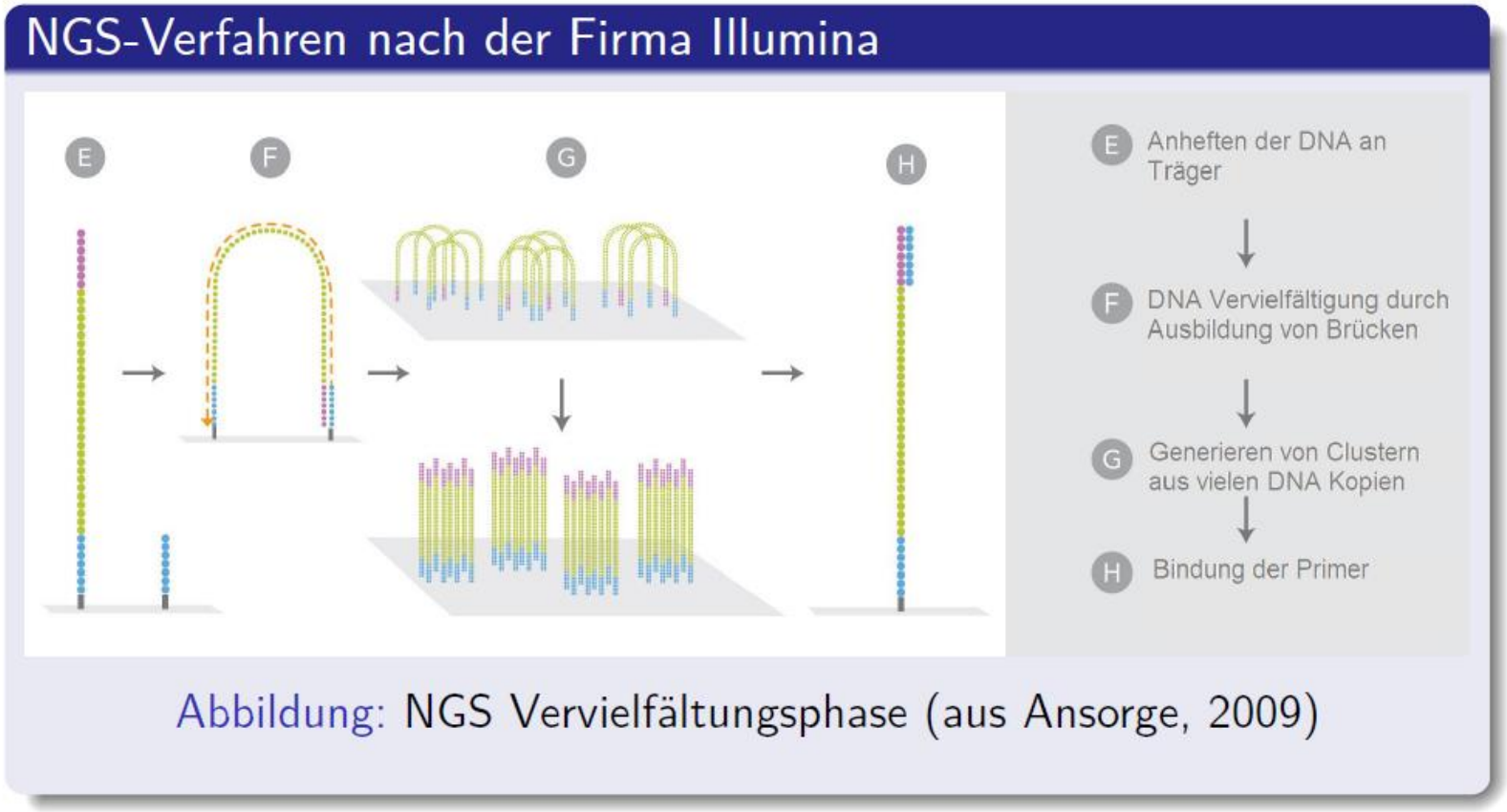
- PCR amplification and sequencing



Ansorge WJ (2009): Next-generation DNA sequencing techniques. *New Biotechnology*, **25**(4): 195-203.

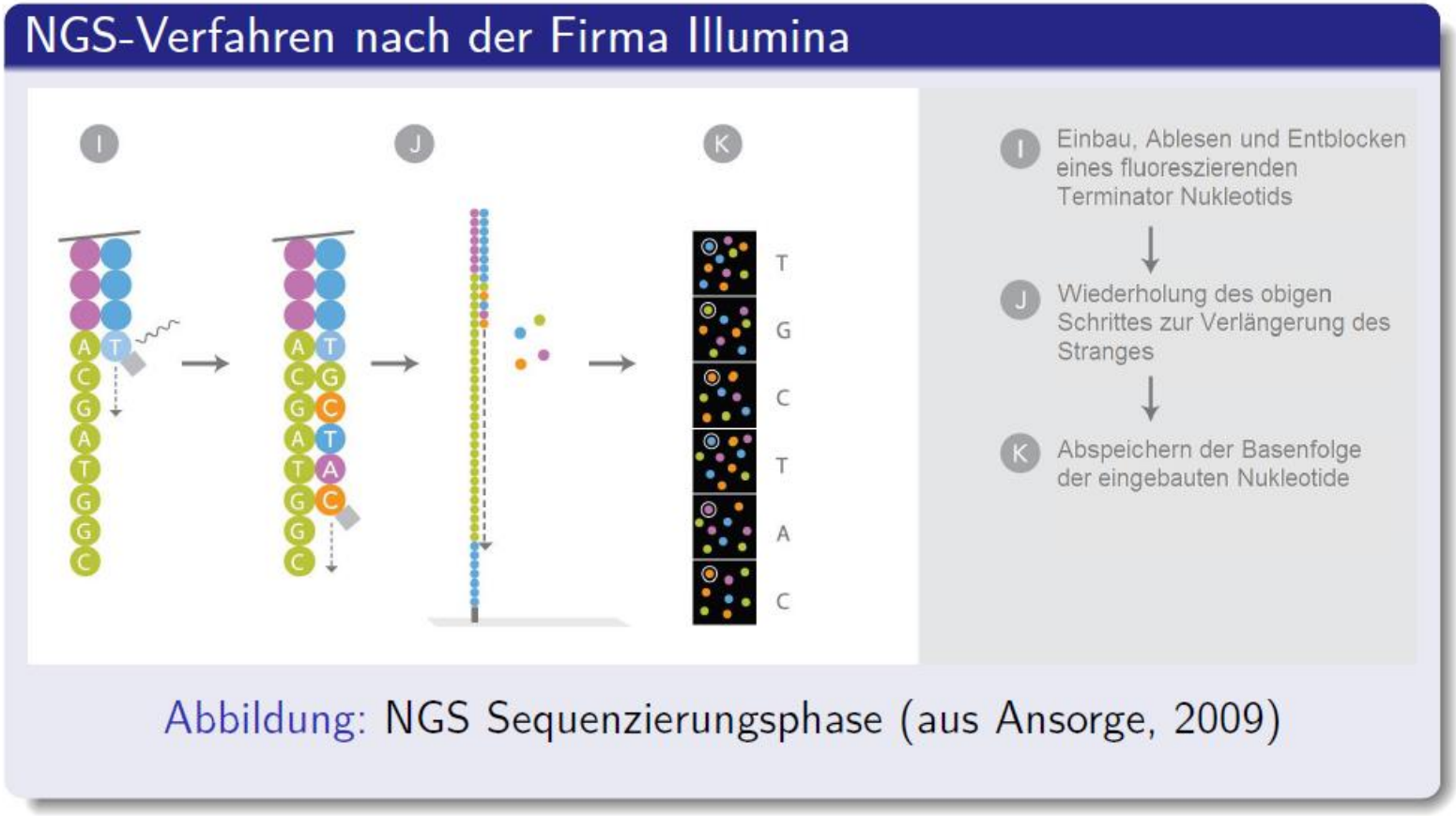
RNAseq technology

- PCR amplification and sequencing



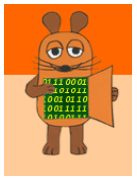
RNAseq technology

- PCR amplification and sequencing

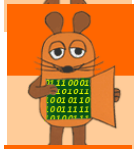


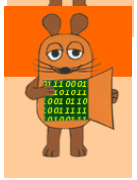
Analysis of RNAseq data

1. Biological question
2. Experimental design
3. Biological experiment
4. Sequence analysis
5. Normalization
6. Statistical analysis
7. Biological verification und interpretation






RNAseq technology

- Jungle of methods for analysis of RNAseq data

http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

Contents

Quantitative analysis and Differential Expression

These tools calculate the abundance of each gene expressed in a RNA-Seq sample (see also Quantification models (http://arxiv.org/abs/1104.3889)). Some software are also designed to study the variability of genetic expression between samples (differential expression). Quantitative and differential studies are largely determined by the quality of reads alignment and accuracy of isoforms reconstruction. See a *comparative study (http://www.amjbot.org/content/99/2/248.long)* of differential expression methods and *Which method should you use for normalization of rna-seq data? (http://www.rna-seqblog.com/data-analysis/which-method-should-you-use-for-normalization-of-rna-seq-data/)*.

- **ALDEx2** *ALDEx2 (https://github.com/ggloor/ALDEx2)* is a tool for comparative analysis of high-throughput sequencing data. ALDEx2 uses compositional data analysis and can be applied to RNAseq, 16S rRNA gene sequencing, metagenomic sequencing, and selective growth experiments.
- **Alexa-Seq** *Alexa-Seq (http://www.alexaplatform.org/alexa_seq/downloads.htm)* is a pipeline that makes possible to perform gene expression analysis, transcript specific expression analysis, exon junction expression and quantitative alternative analysis. Allows wide alternative expression visualization, statistics and graphs. See also *seqanswers/Alexa-Seq (http://seqanswers.com/wiki/ALEXA-Seq)*.
- **ASC** *ASC (http://www.stat.brown.edu/ZWU/research.aspx)*. See also *seqanswers/ASC (http://seqanswers.com/wiki/ASC)*.
- **Ballgown** *Ballgown (https://github.com/alyssafrazee/ballgown)*.
- **BaySeq** *BaySeq (http://www.bioconductor.org/packages/release/bioc/html/baySeq.html)* is a Bioconductor package to identify differential expression using next-generation sequencing data, via empirical Bayesian methods. There is an option of using the "snow" package for parallelisation of computer data processing, recommended when dealing with large data sets. See also *seqanswers/BaySeq (http://seqanswers.com/wiki/BaySeq)*.
- **BBSeq** *BBSeq (http://www.bios.unc.edu/research/genomic_software/BBSeq/)*. See also *seqanswers/BBSeq (http://seqanswers.com/wiki/BBSeq)*.
- **BitSeq** *BitSeq (http://code.google.com/p/bitseq/)*.
- **CEDER** *CEDER (http://www-ref.usc.edu/~fsun/Programs/CEDER/CEDERmain.html)*.
- **CPTRA** *CPTRA (http://people.tamu.edu/~zyuan/cptra/cptra.html)*.
- **casper** *casper (http://bioconductor.org/packages/release/bioc/html/casper.html)* is a Bioconductor package to quantify expression at the isoform level. It combines using informative data summaries, flexible estimation of experimental biases and statistical precision considerations which (reportedly) provide substantial reductions in estimation error.
- **Cufflinks/Cuffdiff** *Cufflinks (http://cufflinks.cbcb.umd.edu)* is appropriate to measure global *de novo* transcript isoform expression. It performs assembly of transcripts, estimation of abundances and determines differential expression (Cuffdiff) and regulation in RNA-Seq samples. See also *seqanswers/Cufflinks (http://seqanswers.com/wiki/Cufflinks)*. [4]
- **DESeq** *DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html)* is a Bioconductor package to perform differential gene expression analysis based on negative binomial distribution. See also *seqanswers/DESeq (http://seqanswers.com/wiki/DESeq)*.
- **DEGSeq** *DEGSeq (http://www.bioconductor.org/packages/2.6/bioc/html/DEGseq.html)*. See also *seqanswers/DEGSeq (http://seqanswers.com/wiki/DEGseq)*.
- **DEXSeq** *DEXSeq (http://bioconductor.org/packages/release/bioc/html/DEXSeq.html)* is Bioconductor package that finds differential differential exon usage based on RNA-Seq exon counts between samples. DEXSeq employs negative binomial distribution, provides options to visualization and exploration of the results.
- **DEXUS** *dexus (http://www.bioconductor.org/packages/2.13/bioc/html/dexus.html)* is a Bioconductor package that identifies differentially expressed genes in RNA-Seq data under all possible study designs such as studies without replicates, without sample groups, and with unknown conditions.[5] In contrast to other methods, DEXUS does not need replicates to detect differentially expressed transcripts, since the replicates (or conditions) are estimated by the EM method for each transcript.
- **DiffSplice** *DiffSplice (http://www.netlab.uky.edu/p/bioinfo/DiffSplice)* is a method for differential expression detection and visualization, not dependent on gene annotations. This method is supported on identification of alternative splicing modules (ASMs) that diverge in the different isoforms. A non-parametric test is applied to each ASM to identify significant differential transcription with a measured false discovery rate.
- **EBSeq** *EBSeq (http://www.biostat.wisc.edu/~kendzior/EBSEQ/)* is a Bioconductor package for identifying genes and isoforms differentially expressed (DE) across two or more biological conditions in an RNA-seq experiment. It also can be used to identify DE contigs after performing de novo transcriptome assembly. While performing DE analysis on isoforms or contigs, different isoform/contig groups have varying estimation uncertainties. EBSeq models the varying uncertainties using an empirical Bayes model with different priors.
- **EdgeR** *EdgeR (http://www.bioconductor.org/packages/release/bioc/html/edgeR.html)* is a R package for analysis of differential expression of data from DNA sequencing methods, like RNA-Seq, SAGE or ChIP-Seq data. edgeR employs statistical methods supported on negative binomial distribution as a model for count variability. See also *seqanswers/EdgeR (http://seqanswers.com/wiki/EdgeR)*.
- **ESAT** *ESAT (http://garberlab.umassmed.edu/software/esat/index.html)* The End Sequence Analysis Toolkit (ESAT) is specially designing to be applied for quantification of annotation of specialized RNA-Seq gene libraries that target the 5' or 3' ends of transcripts.
- **eXpress** *eXpress (http://bio.math.berkeley.edu/eXpress/index.html)* performance includes transcript-level RNA-Seq quantification, allele-specific and haplotype analysis and can estimate transcript abundances of the multiple isoforms present in a gene. Although could be coupled directly with aligners (like Bowtie), eXpress can also be used with de novo assemblers and thus is not needed a reference genome to perform alignment. It runs on Linux, Mac and Windows.
- **ERANGE** *ERANGE (http://woldlab.caltech.edu/rnaseq)* performs alignment, normalization and quantification of expressed genes. See also *seqanswers/ERANGE (http://seqanswers.com/wiki/ERANGE)*.
- **featureCounts** *featureCounts (http://bioinf.wehi.edu.au/featureCounts/)* an efficient general-purpose read quantifier. It is part of the *SourceForge Subread package (http://subread.sourceforge.net)* and *Bioconductor Rsubread package (http://bioconductor.org/packages/release/bioc/html/Rsubread.html)*.
- **FDM** *FDM (http://cebio-linux001.cs.unc.edu/nextgen/software/FDM/)*
- **GPSeq** *GPSeq (http://www-ref.usc.edu/%7Eliangche/software.html)*
- **MATS** *MATS (http://rnaseq-mats.sourceforge.net/)*.

- **MMSEQ** *MMSEQ (http://bgx.org.uk/software/mmseq.html)* is a pipeline for estimating isoform expression and allelic imbalance in diploid organisms based on RNA-Seq. The pipeline employs tools like Bowtie, TopHat, ArrayExpressHTS and SAMtools. Also, edgeR or DESeq to perform differential expression. See also *seqanswers/MMSEQ (http://seqanswers.com/wiki/MMSEQ)*.
- **Myrna** *Myrna (http://bowtie-bio.sourceforge.net/myrna/index.shtml)* is a pipeline tool that runs in a cloud environment (*Elastic MapReduce (http://aws.amazon.com/elasticmapreduce/)*) or in a unique computer for estimating differential gene expression in RNA-Seq datasets. Bowtie is employed for short read alignment and R algorithms for interval calculations, normalization, and statistical processing. See also *seqanswers/Myrna (http://seqanswers.com/wiki/Myrna)*.
- **NEUMA** *NEUMA (http://neuma.kobic.re.kr/)* is a tool to estimate RNA abundances using length normalization, based on uniquely aligned reads and mRNA isoform models. NEUMA uses known transcriptome data available in databases like RefSeq.
- **NOISeq** *NOISeq (http://bioinfo.cipf.es/noiseq/doku.php?id=start)*. See also *seqanswers/NOISeq (http://seqanswers.com/wiki/NOISeq)*.
- **NPEBseq** *NPEBseq (http://bioinformatics.wistar.upenn.edu/NPEBseq)* is a nonparametric empirical bayesian- based method for differential expression analysis.
- **NSMAP** *NSMAP (http://sites.google.com/site/nsmapformaseq/)* allows inference of isoforms as well estimation of expression levels, without annotated information. The exons are aligned and splice junctions are identified using TopHat. All the possible isoforms are computed by combination of the detected exons.
- **Qlucore** Easy to use for analysis and visualization. One button import of BAM files. Qlucore.
- **RNAeXpress** *RNAeXpress (http://rnaexpress.org/)* Can be run with Java GUI or command line on Mac, Windows and Linux. Can be configured to perform read counting, feature detection or GTF comparison on mapped maseq data.

- **RNA-Skim** *RNA-Skim (http://www.csbio.unc.edu/rs/)*
- **rDiff** *rDiff (http://bioweb.me/rdiff/)* is a tool that can detect differential RNA processing (e.g. alternative splicing, polyadenylation or ribosome occupancy).
- **rSeq** *rSeq (http://www-personal.umich.edu/~jianghui/rseq/)*
- **RSEM** *RSEM (http://deweylab.biostat.wisc.edu/rsem/)*. See also *seqanswers/RSEM (http://seqanswers.com/wiki/RSEM)*.
- **rQuant** *rQuant (http://www.raetschlab.org/suppl/rquant)* is a web service (Galaxy (computational biology) installation) that determines abundances of transcripts per gene locus, based on quadratic programming. rQuant is able to evaluate biases introduced by experimental conditions. A combination of tools is employed: PALMapper (reads alignment), mTiM and mGene (inference of new transcripts).
- **Salmon** *Salmon (https://github.com/kingsfordgroup/sailfish/tree/develop)* is an software tool for computing transcript abundance from RNA-seq data using either an alignment-free (based directly on the raw reads) or an alignment-based (based on pre-computed alignments) approach. It uses an online stochastic optimization approach to maximize the likelihood of the transcript abundances under the observed data. The software itself is capable of making use of many threads to produce accurate quantification estimates quickly. It is part of the *Sailfish (http://www.cs.cmu.edu/~ckingsf/software/sailfish/)* suite of software, and is the successor to the Sailfish tool.
- **SAJR** *SAJR (http://storage.bioinf.fbb.msu.ru/~mazin/)* is a java-written read counter and R-package for differential splicing analysis. It uses junction reads to estimate exon exclusion and reads mapped within exon to estimate its inclusion. SAJR models it by GLM with quasibinomial distribution and uses log likelihood test to assess significance.
- **Scotty** *Scotty (http://euler.bc.edu/marthlab/scotty/scotty.php/)* Performs power analysis to estimate the number of replicates and depth of sequencing required to call differential expression.
- **SplicingCompass** *SplicingCompass (http://www.ichip.de/software/SplicingCompass.html)*.

RNAseq – analysis guidelines

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

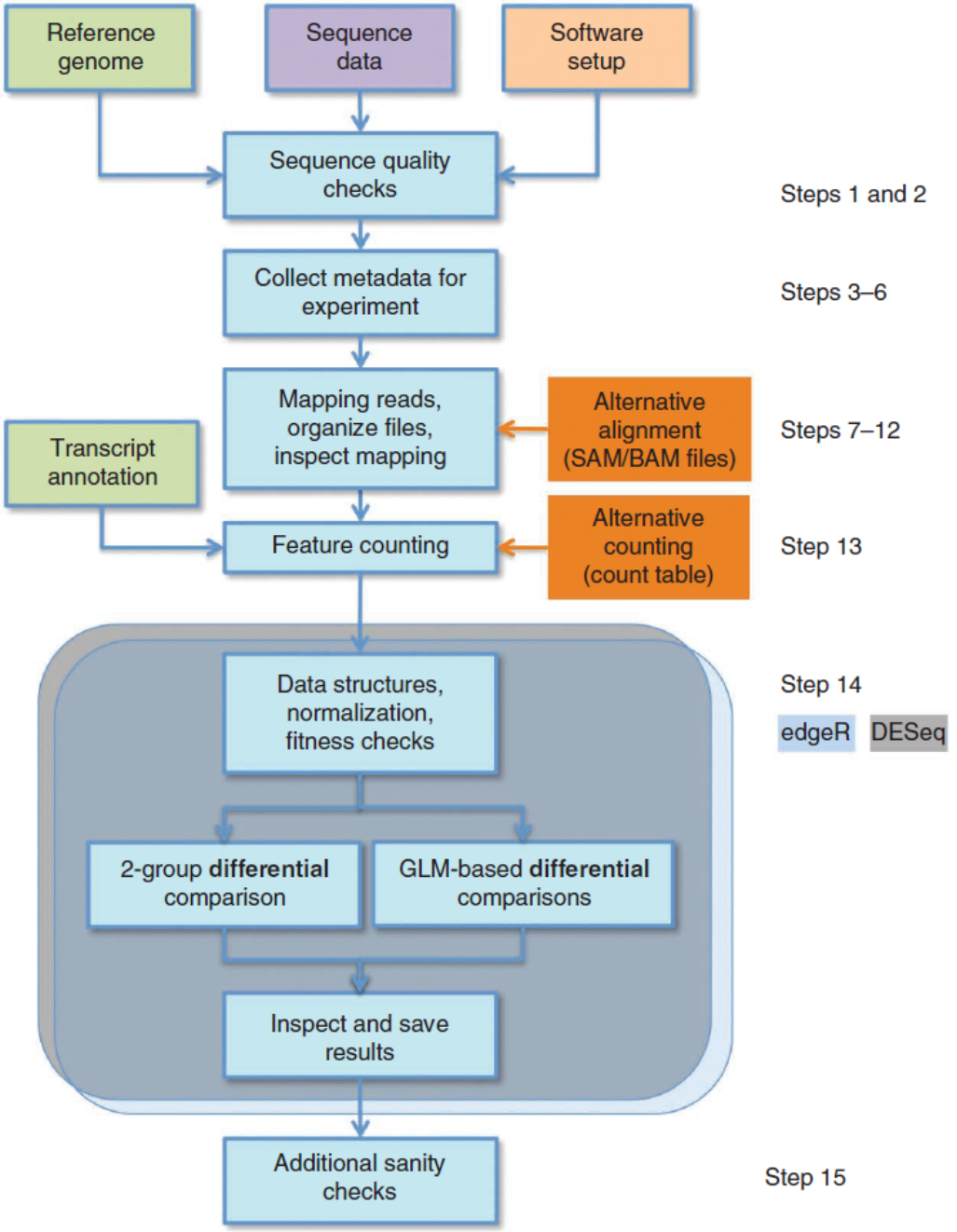
[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Functional Genomics Center UNI ETH, Zurich, Switzerland. [7]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.D.R. (mark.robinson@imls.uzh.ch) or W.H. (whuber@embl.de).

Published online 22 August 2013; doi:10.1038/nprot.2013.099

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations) while optionally adjusting for other systematic factors that affect the data-collection process. There are a number of subtle yet crucial aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a state-of-the-art computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and, in particular, on two widely used tools, DESeq and edgeR. Hands-on time for typical small experiments (e.g., 4–10 samples) can be <1 h, with computation time <1 d using a standard desktop PC.

RNAseq

- Analysis pipeline
- From raw data to the gene list
- Schematic procedure from Nature Protocols
  - Count-based differential expression analysis of RNA sequencing data using R and Bioconductor
  - Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, Mark D Robinson



RNAseq – modelling

- RNAseq data are count data
  - Discrete, skewed, large span
  - Normal distribution not adequate
  - Rank and permutation methods mostly also not adequate (often small number of samples, many genes)
- Modelling with suitable distributions
  - Poisson distribution
  - Negative binomial distribution
- Typical pipeline
  - Normalization of data
  - Estimation of model parameters (expected value and variance)
  - Often: Determination of differentially expressed genes

RNAseq – modelling

- Poisson distribution
  - Distribution for rare events
  - Limit distribution for binomial distribution with large number of tries and small probability ($\lambda = n \cdot p$)

$$P(X = x) = \frac{\lambda^x}{x!}\exp(-\lambda), \qquad E(X) = \lambda,\ Var(X) = \lambda$$

- Negative binomial distribution
  - Number of failures until *r*-th success in Bernoulli process with success probability *p*

$$P(X = x) = \binom{x + r - 1}{x} p^r (1-p)^x, \qquad E(X) = r\frac{(1-p)}{p},\ Var(X) = r\frac{(1-p)}{p^2}$$

  - Alternative parameterization: $E(X) = \mu,\ Var(X) = \mu + \mu^2\phi$
  - Advantage: Additional parameter for estimation of individual variance

Data example

- Dataset from Mainz
  - Institut für Molekulare Medizin, Universitätsmedizin der Johannes Gutenberg-Universität Mainz
- Comparison of two groups of mice
  - 3 healthy mice
  - 3 genetically manipulated mice (overexpression of Bcl3 in T cells) with inflammation of the colon and problems with the development of different subgroups of T cells
  - Analysis of genetic material from two types of T cells: CD4 cells and CD8 cells
  - Very small sample size, thus stable analysis required

Data example

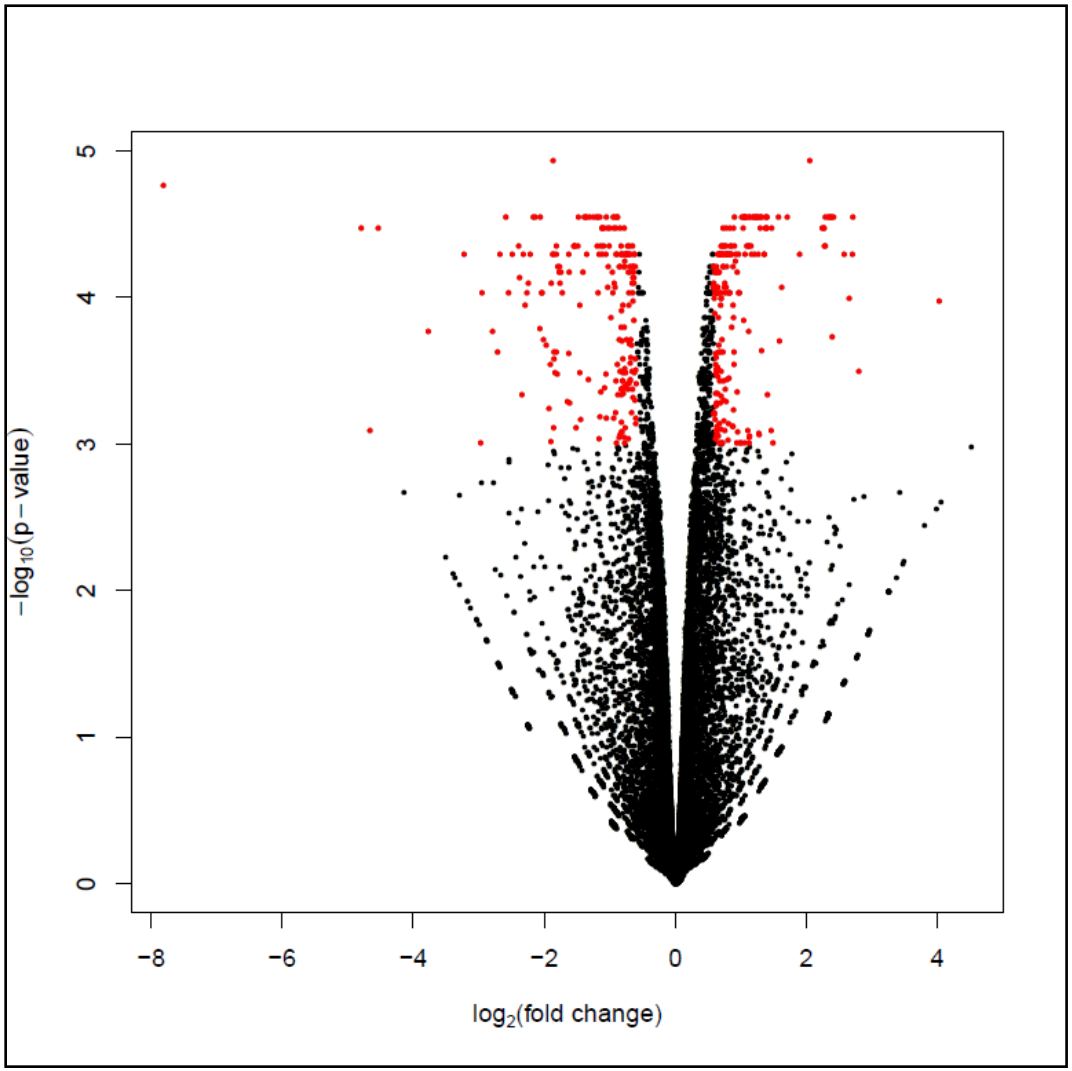
- Analysis of RNAseq data: 3 early R packages
- DESeq
  - Simon Anders, Wolfgang Huber
    Differential expression analysis for sequence count data
    Genome Biology 11, R106 (2010)
- edgeR
  - Robinson MD, McCarthy DJ, Smyth GK:
    edgeR: a Bioconductor package for differential expression analysis of digital gene expression data
    Bioinformatics 26(1):139-40 (2010)
- PoissonSeq
  - Li J, Witten DM, Johnstone IM, Tibshirani R
    Normalization, testing, and false discovery rate estimation for RNA-sequencing data
    Biostatistics 13(3):523-538 (2012)

Volcano-Plots: PoissonSeq

- X axis: effect ($\log_2$ fold change)
- Y axis: p-value ($\log_{10}$ p)

Red points

- Effect: FC > 3/2 or FC < 2/3
- P-value: p < 0.001

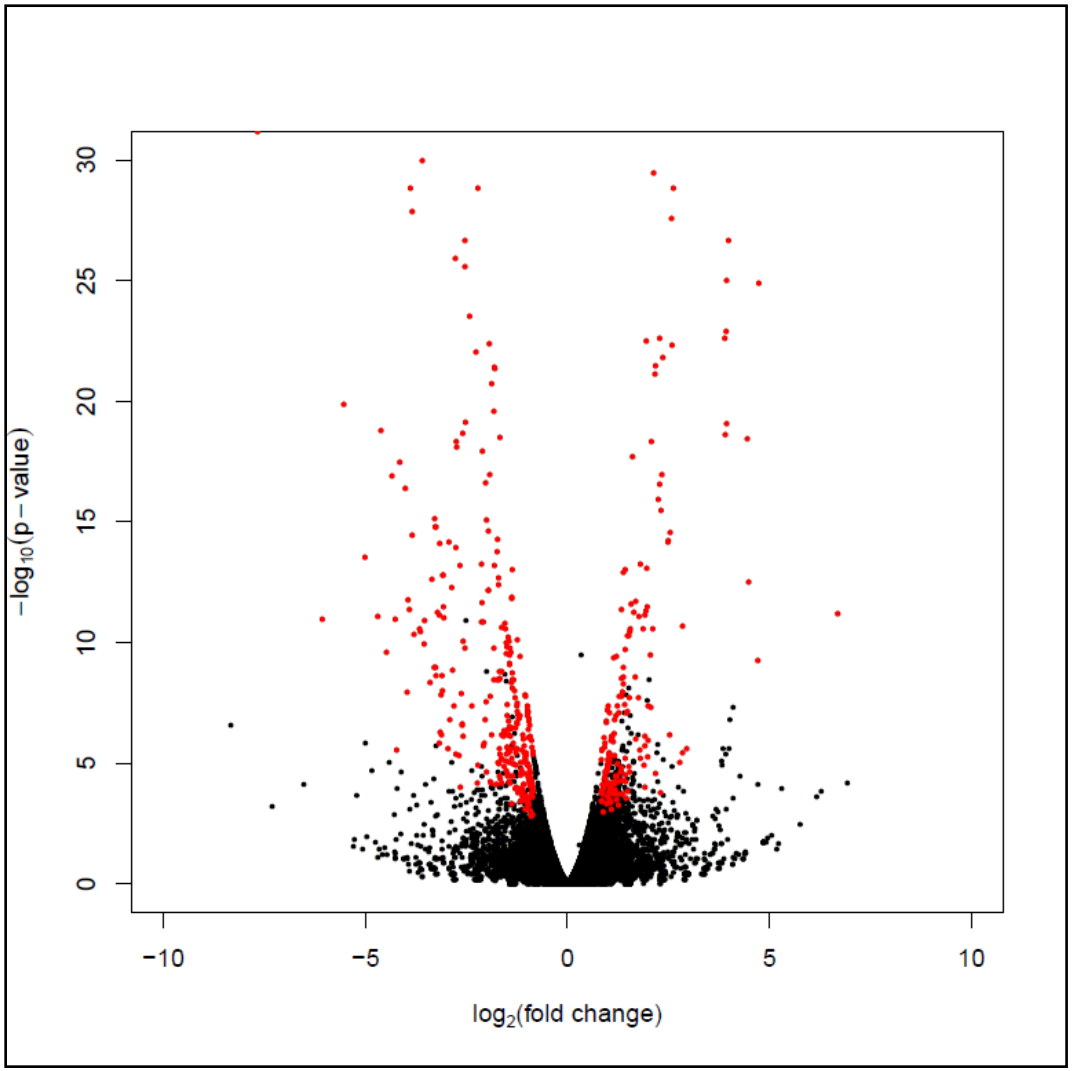


Volcano-Plots: DESeq

- X axis: effect ($\log_2$ fold change)
- Y axis: p-value ($\log_{10}$ p)

Red points

- Effect:
  FC > 3/2 or FC < 2/3
- P-value:
  p < 0.001

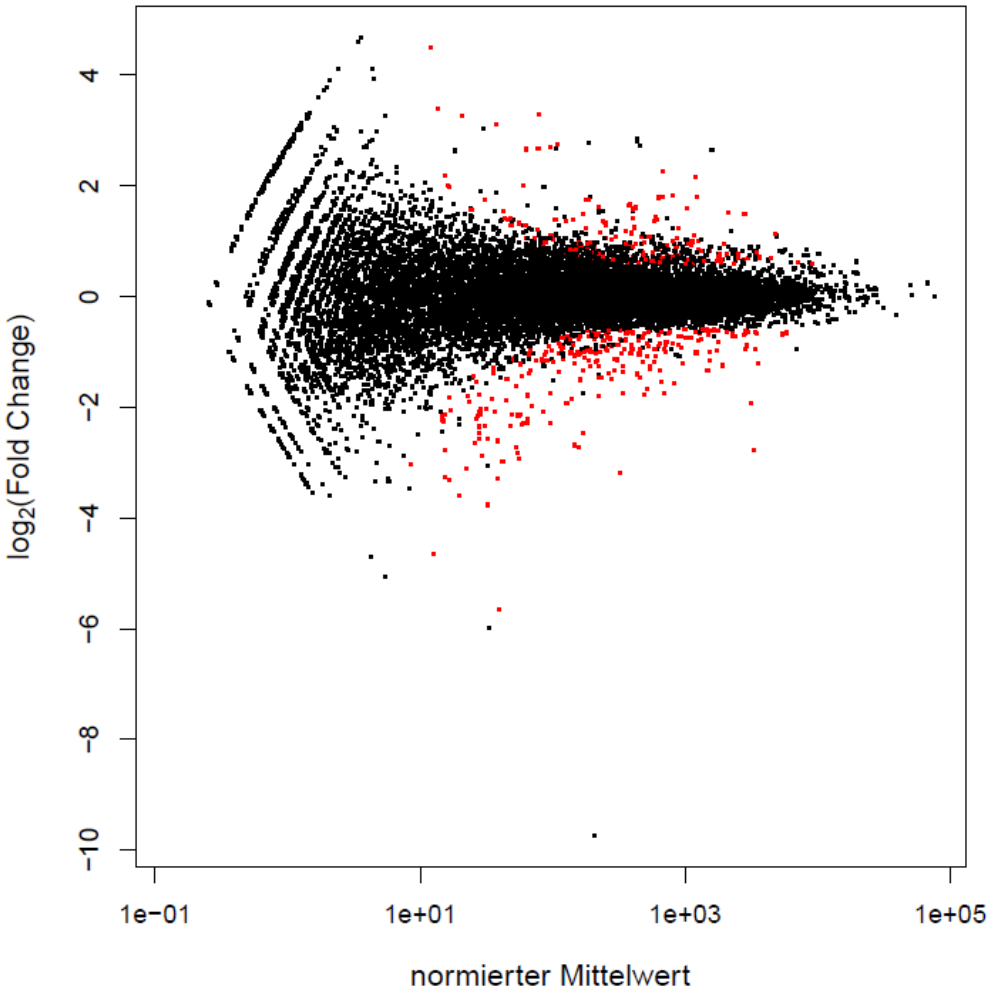


MA-Plot: DESeq

- X axis: intensity (mean of exp. & ctrl.)
- Y axis: effect (log_2 fold change)

Red points

- Effect: FC > 3/2 or FC < 2/3
- P-value: p < 0.001



Normalization

- Total-count normalization
  - Sum of all counts is set to the same value
    (see next slide, not sufficient)
- RPKM (Mortazavi et al., 2008)
  - Reads Per Kilobase per Million mapped reads
  - Normalization w.r.t length of gene and total number of counts
  - Necessary but not sufficient (Wagner, Theory Biosci, 2012)
  - Must be careful when modelling with “counts”
- Quantile normalization
  - „All“ quantiles set to the same value (across experiments)
- „Median-Reference-Normalization“
  - Scaling factor d_j per sample *j*
  - Per gene *i* artificial reference sample $\hat{d}_j = \underset{i}{\text{median}} \frac{Y_{ij}}{(\prod_{h=1}^{n} Y_{ih})^{1/n}}$

RNAseq – Normalization

- Total-count normalization (TCN, same sum) often not suitable
- Example demonstrates false positive results

Original data

Gene	Group 1	Group 2
1	80	80
2	100	100
3	30	30
4	200	200
5	240	240
6	310	310
7	0	300
8	20	520
9	5	100
10	15	120
Sum	1000	2000

Data after TCN

Gene	Group 1	Group 2
1	80	40
2	100	50
3	30	15
4	200	100
5	240	120
6	310	155
7	0	150
8	20	260
9	5	50
10	15	60
Sum	1000	1000

RNAseq – Normalization

- Total-count normalization (TCN, same sum) often not suitable
- Example demonstrates false positive results

Original data

Gene	Group 1	Group 2
1	80	80
2	100	100
3	30	30
4	200	200
5	240	240
6	310	310
7	0	300
8	20	520
9	5	100
10	15	120
Sum	1000	2000

Data after TCN

Gene	Group 1	Group 2
1	80	40
2	100	50
3	30	15
4	200	100
5	240	120
6	310	155
7	0	150
8	20	260
9	5	50
10	15	60
Sum	1000	1000

RNAseq – Normalization

- Scaling factor d_j per sample j : $\hat{d}_j = \underset{i}{\text{median}} \frac{Y_{ij}}{(\prod_{h=1}^{n} Y_{ih})^{1/n}}$
- Data: Median of transformed data always 1, this means no change

Original data

Gene	Group 1	Group 2
1	80	80
2	100	100
3	30	30
4	200	200
5	240	240
6	310	310
7	0	300
8	20	520
9	5	100
10	15	120
Sum	1000	2000

Transformed data

Gene	Group 1	Group 2
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	„0/0“	„Inf“
8	0.196	5.10
9	0.224	4.47
10	0.354	2.83
Median	$d_1 = 1$	$d_2 = 1$

RNAseq: DESeq

- Estimation of model parameters for DESeq
- Y_{ij} is number of counts for gene *i* in sample *j* (in group *k*)

$$Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$$

- q_{ik} corresponds to the expected normalized count value for gene *i* in group *k*
- d_j is a scaling factor for sample *j*
- Variance decomposition in technical and biological part
- $v_{ik} = f_k(q_{ik}), \quad f_k : \mathbb{R}^+ \to \mathbb{R}^+$ smooth function

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$
- Estimator for expected value: $\hat{\mu}_{ij} = \hat{d}_j \hat{q}_{ik}$
- Estimator for q_{ik} : $\hat{q}_{ik} = \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{y_{ij}}{\hat{d}_j}$
- Here n_k is the number of samples in group k.

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$
- Estimator for variance:

$$w_{ik} = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} \left(\frac{y_{ij}}{\hat{d}_j} - \hat{q}_{ik} \right)^2 \text{ and } s_{ik} = \frac{\hat{q}_{ik}}{n_k} \sum_{j=1}^{n_k} \frac{1}{\hat{d}_j}$$

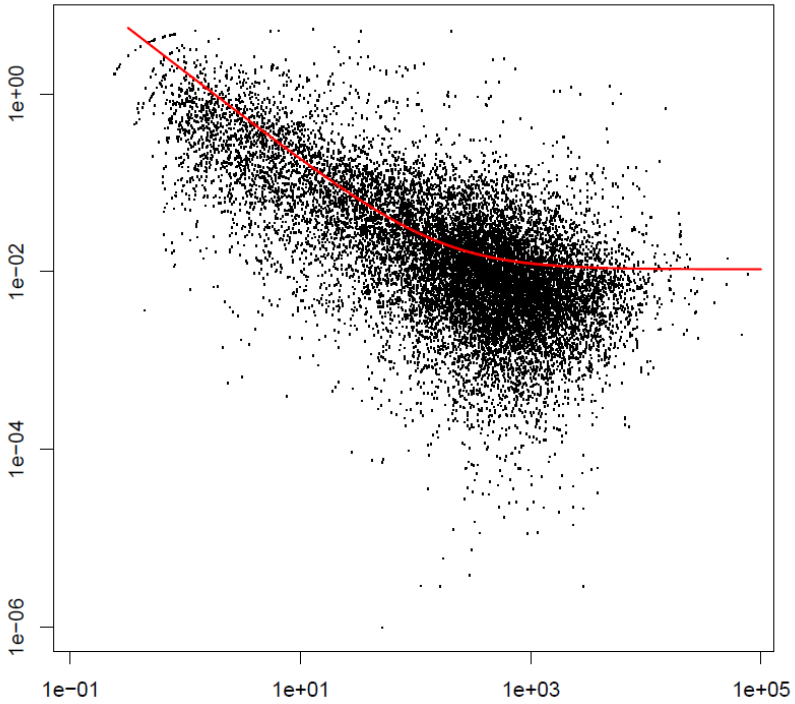
w_{ik} - s_{ik} is an unbiased estimator for v_{ik}

- But w_{ik} has high variability for small n_k
- Thus estimate directly functional dependence between expected value and variance

$$\hat{f}_k(\hat{q}_{ik}) = \omega_k(\hat{q}_{ik}) - s_{ik}$$

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$

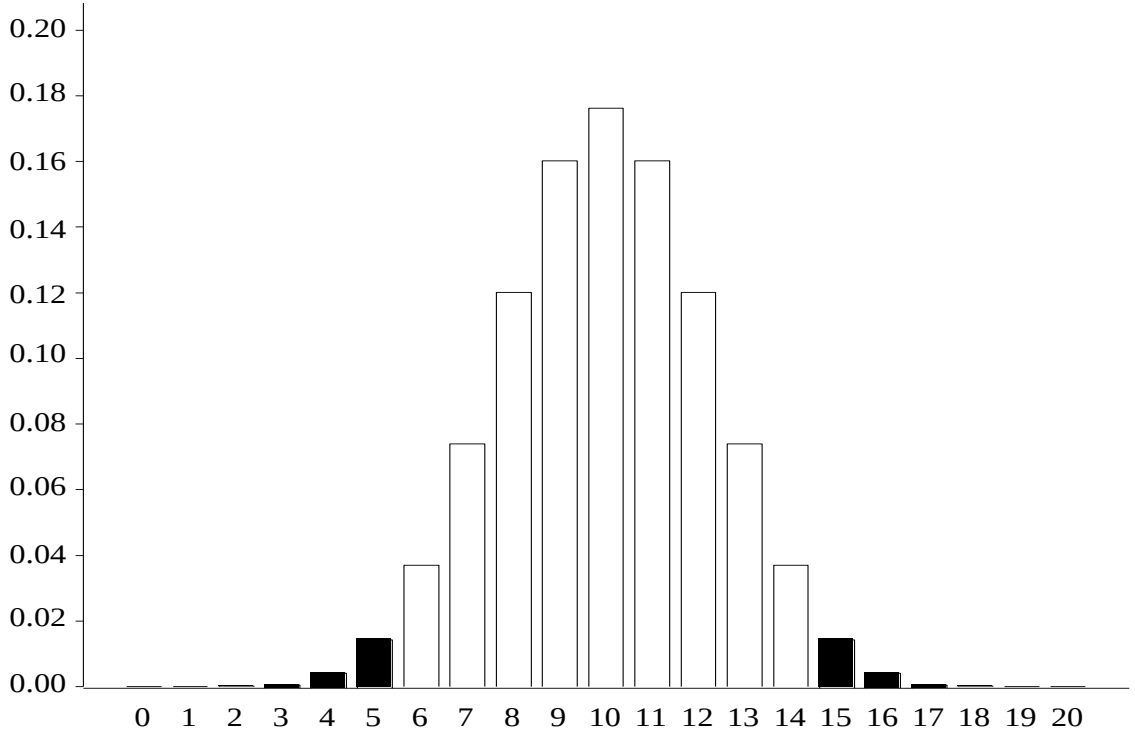


$$\hat{f}_k(\hat{q}_{ik}) = \omega_k(\hat{q}_{ik}) - s_{ik}$$

- First approach was: Use red regression line (local linear fit) as estimates
- Now: Use maximum of red regression line and empirical value

Significance values for discrete distributions

- Reminder: Determination of significance for binomial distribution (two-sided)



Rejection region: 0-5, 15-20

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$
- Differential gene expression
  - Test for difference between 2 groups
  - Test for gene *i* of null hypothesis $H_0 : q_{i1} = q_{i2}, \quad i = 1..m$
  - Define

$$Z_{ik} := \sum_{j=1}^{n_k} Y_{ij} \qquad Z_{i\cdot} := Z_{i1} + Z_{i2}$$

  - Assumption: Z_{ik} is approximately *NB* – distributed

		Number of counts for gene *i*
Group	1	Z_{i1}
	2	Z_{i2}
	$\sum$	$Z_{i\cdot}$

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$

$$Z_{ik} := \sum_{j=1}^{n_k} Y_{ij}$$

$$Z_{i\cdot} := Z_{i1} + Z_{i2}$$

		Number of counts for gene *i*
Group	1	Z_{i1}
	2	Z_{i2}
	$\sum$	$Z_{i\cdot}$

- Define $p(a,b) := P(Z_{i1} = a, Z_{i2} = b)$
- Then p-value for a pair *(a,b)* of observed *count* values is

$$p_i = \frac{\sum\limits_{a+b=z_{i\cdot}} p(a,b)}{\sum\limits_{a+b=z_{i\cdot}} p(a,b)} \quad , \quad a,b \in \{0,\ldots,z_{i\cdot}\}$$
$$p(a,b) \le p(z_{i1},z_{i2})$$

Alternative:
Sum over all pairs with more extreme fold change

RNAseq: DESeq

- Assumptions: $Y_{ij} = NB(\mu_{ij}, \sigma_{ij}^2), \quad \mu_{ij} = d_j q_{ik}, \quad \sigma_{ij}^2 = \mu_{ij} + d_j^2 v_{ik}^2$

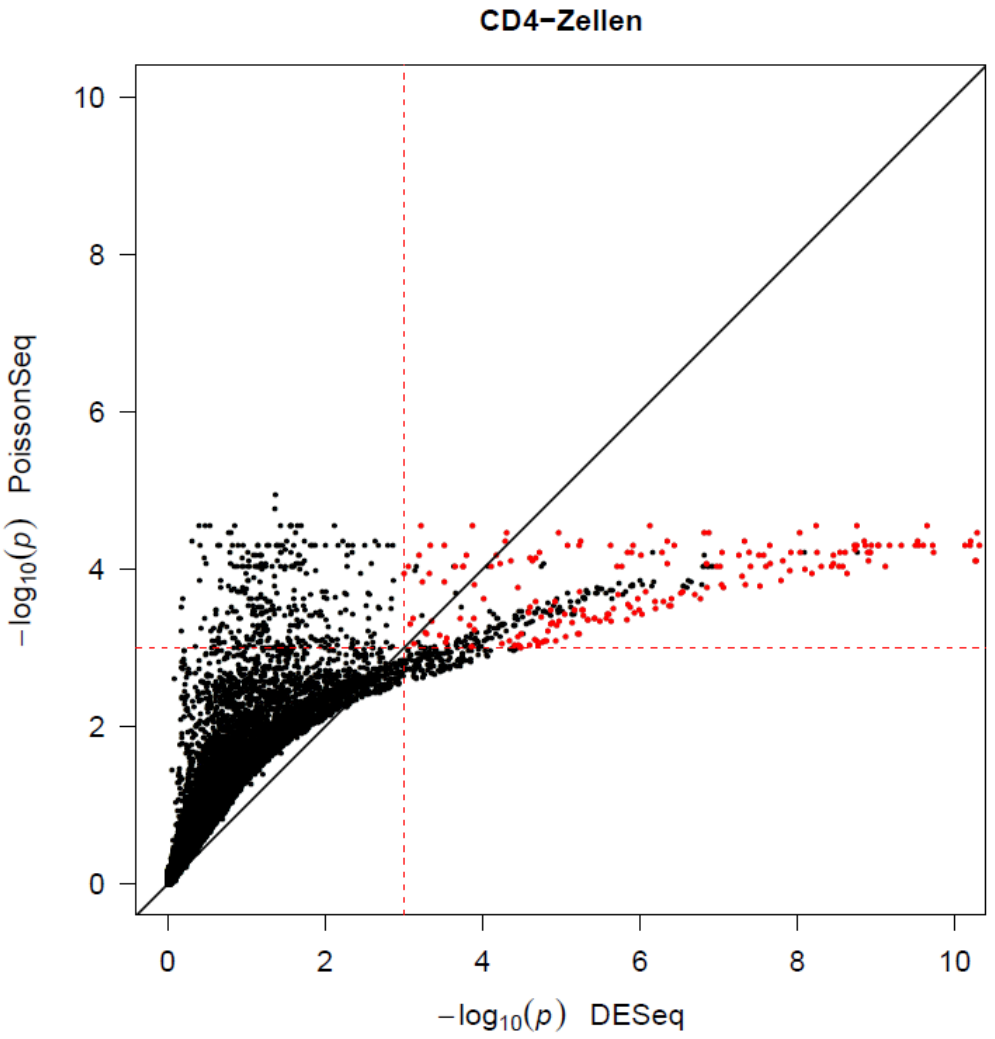
  $Z_{ik} := \sum_{j=1}^{n_k} Y_{ij} \qquad Z_{i\cdot} := Z_{i1} + Z_{i2}$

- Necessary to estimate parameters of the negative binomial distribution for both groups (under H_0)
- Pooled mean of normalizes values: $\hat{q}_{i0} = \sum_{j=1}^{n} \frac{y_{ij}}{\hat{d}_j}$
  (n total number of samples)
- Estimates for mean and variance in group k :

$$\hat{\mu}_{ik} = \sum_{j=1}^{n_k} \hat{d}_j \hat{q}_{i0} \quad \text{and} \quad \hat{\sigma}_{ik}^2 = \sum_{j=1}^{n_k} \hat{d}_j \hat{q}_{i0} + \hat{d}_j^2 \hat{f}_k(\hat{q}_{i0})$$

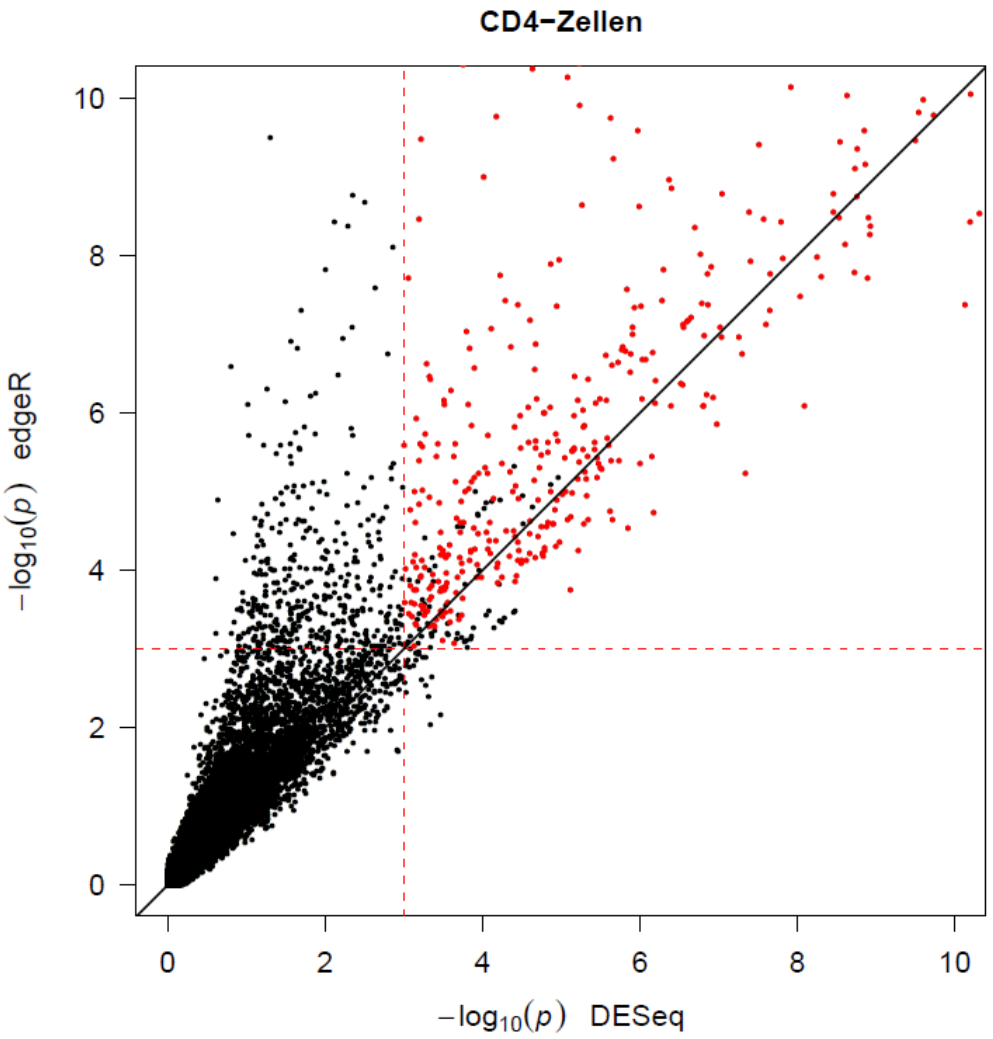
Results – comparison

- PoissonSeq identifies more differential genes than DESeq
- Extremely different distributions



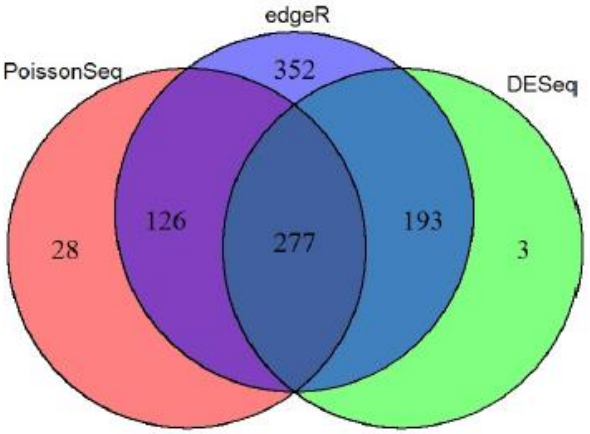
Results – comparison

- edgeR identifies more differential genes than DESeq
- Similar distributions



Results – comparison

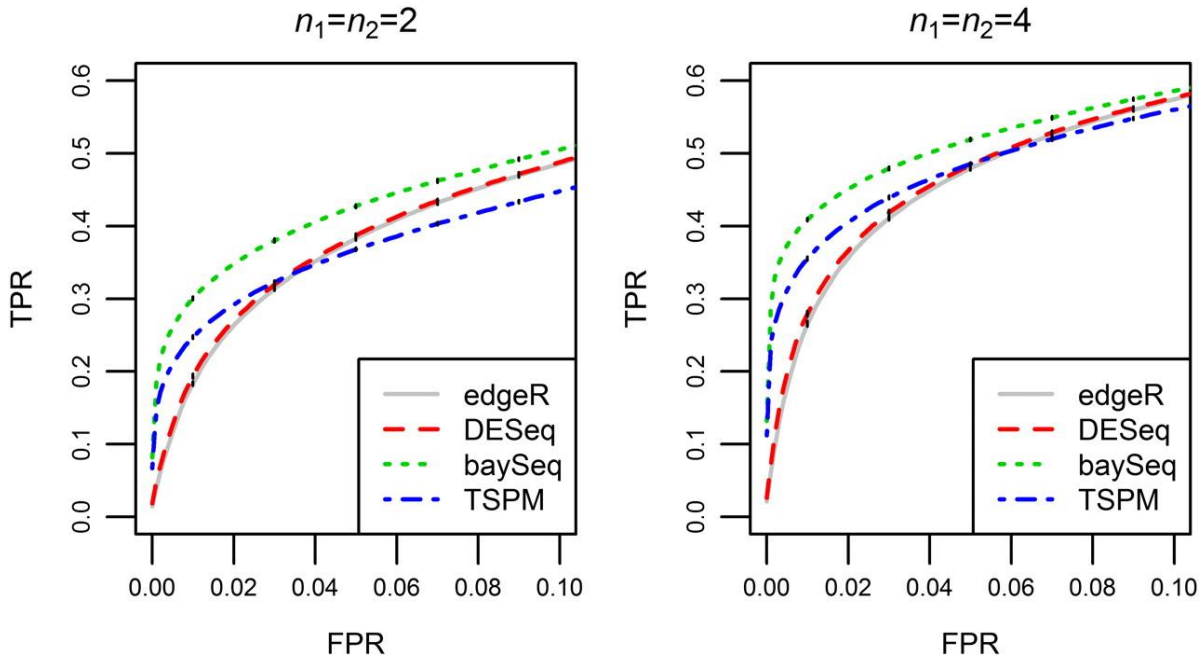
- HOW TO DECIDE WHICH METHOD IS BETTER?
- Comparison of results on data: Probably not…
  - DESeq hat less power, but edgeR is more sensitive regarding outliers
  - Agreement between methods could increase confidence, but …
  - Scientific method: Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Regina Nuzzo, Nature 505, 150-152, 2014



- Simulation studies
  - Simulate data distributions as close as possible to real world situation
- Validation
  - Analyze different cohorts, validate hypothesis on new data, meta analysis…

Results – comparison

- HOW TO DECIDE WHICH METHOD IS BETTER?
- Simulation study: ROC curves comparing the performance to detect differential (half of the genes follow Poisson distributions and the other half follow overdispersed Poisson distributions)



Kvam V.M. et al.: **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data,** *Am. J. Bot.* 99:248-256, 2012

Validation

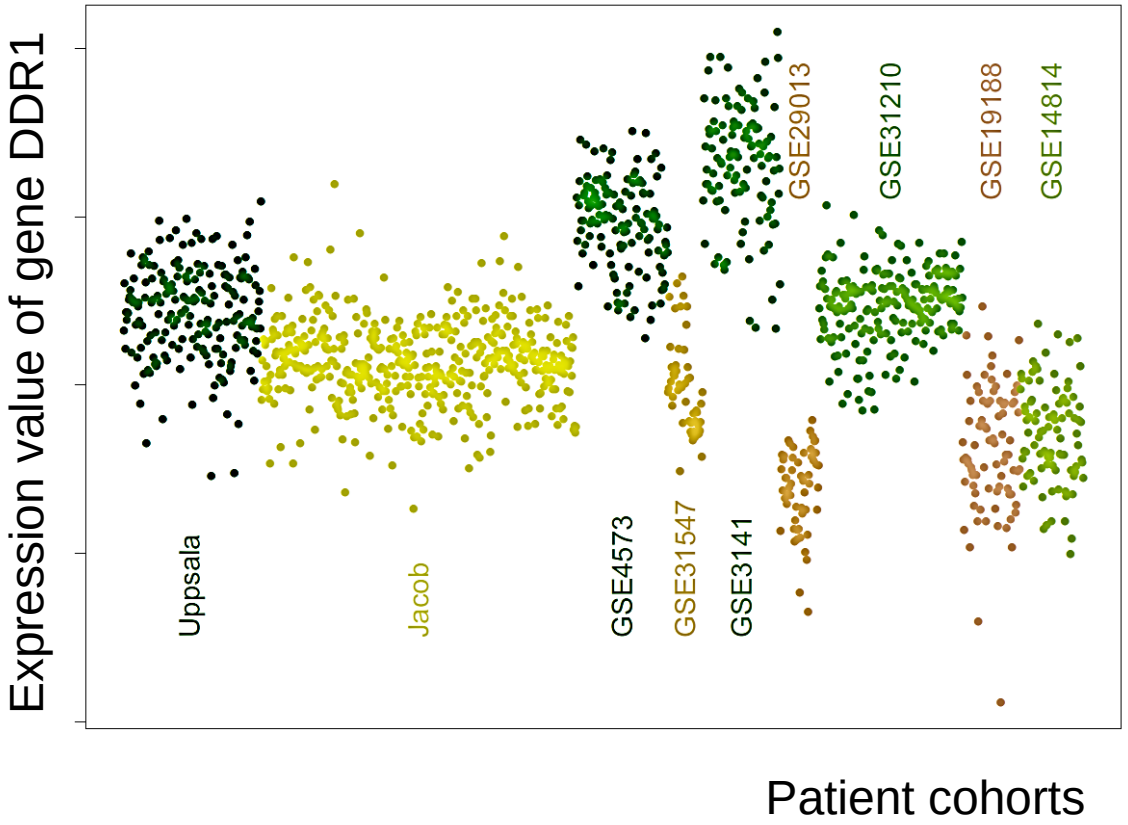
- Analysis of the same research question using „different“ patient cohorts

Problem

- Cohorts show batch effects

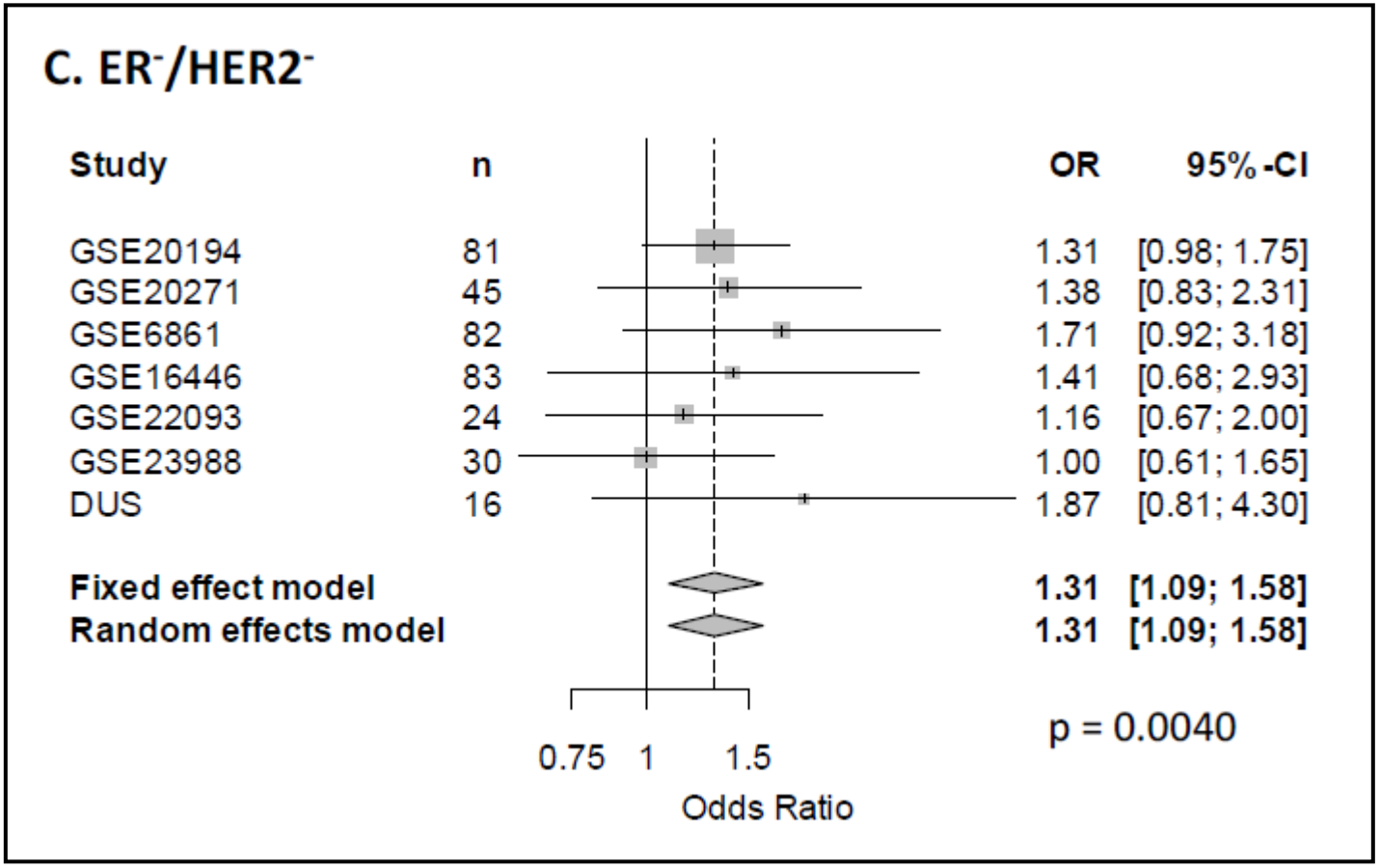
Solutions

- Gene wise standardization of expression values
- Iterative validation procedures
- Meta analysis!



Validation – meta analysis

- IGKC is associated with better response (pCR) for breast cancer patients treated with anthracycline-based neoadjuvant chemotherapy



Multiple testing

- Multiple testing: Many hypotheses tested simultaneously
  - Higher risk of “false positive” decisions
    - For 20.000 genes and p-value cutoff 0.01 even under the null hypothesis (no effect at all) approx. 20.000·0.01 = 200 genes will be called “differential”
- Solutions
  - Control of FWER (family-wise error rate)
    - FWER: Probability to have at least one false positive result
  - Control of FDR (false discovery rate)
    - FDR: Expected proportion of false positive results among the rejected null hypotheses
    - *FDR = E(Q)* with $Q = \begin{cases} V/\mathrm{R} & \text{für } R > 0 \\ 0 & \text{für } R = 0 \end{cases}$
  - No control
    - Sometimes useful in enrichment analyses

Gene group tests

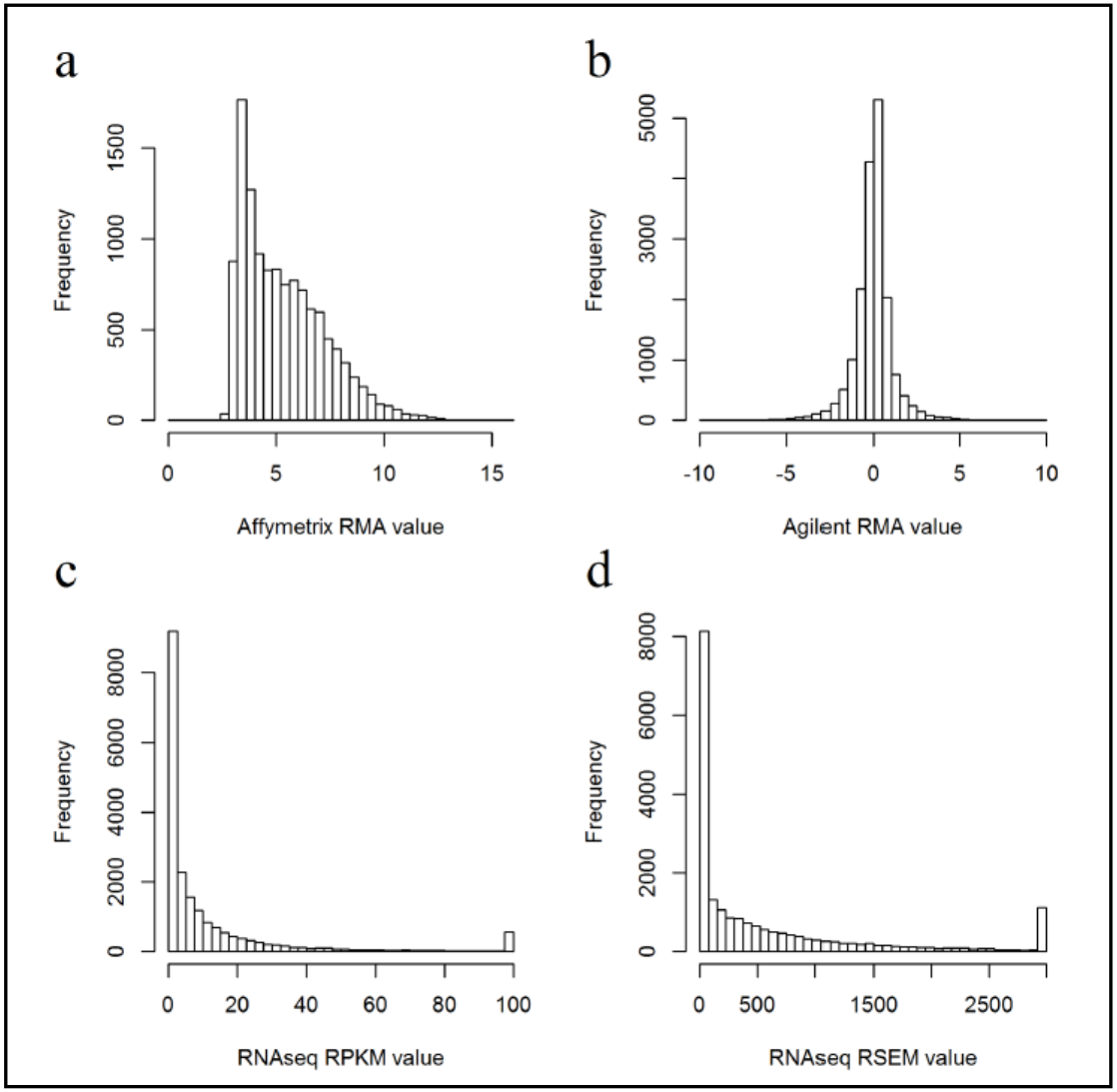
- General idea
  - Summarize genes to gene groups, that are in a joint predefined functional context
- Typical gene groups
  - Gene Ontology groups
  - Gene groups defined by transcription factors
  - Genes that belong to the same pathway
    - Regulatory pathways
    - Metabolic pathways
    - Signaling pathways
  - Gene groups obtained from previous experiments
    - Differential genes between samples / conditions
    - Genes relevant in specific diseases

Comparison microarray – RNAseq

- Guo Y, Sheng Q, Li J, Ye F, Samuels DC, et al. (2013) Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data, PLoS ONE 8(8): e71462.
  - Technologies (microarray and RNAseq) are very different, and RNAseq is rapidly replacing microarrays
  - Here: Large comparison study using TCGA (The Cancer Genome Atlas)
  - High correlation between Affymetrix and RNAseq (Korr~0.8, but worse for small counts)
  - Low correlation between Agilent and RNAseq (Korr~0.2)
  - In general good comparability for comparison tumor vs. normal
- Positive conclusion
  - Huge amounts of available microarray data can still be used

Comparison microarray – RNAseq

- Guo et al, 2013
- Comparison of distributions of gene expression based on 258 tumor samples
- RNAseq:
  Values greater 100 are set to 100



Why male-female classifier?

- Popular approach
  - Comprehensive transcriptomic analysis of clinically annotated human tissue
- Validation important
  - Use of multiple datasets for the same question
  - Public accessibility of many raw data set together with information on clinico-pathological parameters
  - Gain of statistical power
- Results and conclusions depend on the reliability of the available information!
- Reasons for sample mix-up are manifold

Identification of sample misannotations

- We propose gene expression based methods for identifying sample misannotations in public omics datasets

Two statistical methods

- a likelihood-based classifier that is able to differentiate between samples from male and female patients
- expression correlation analysis that identifies multiple measurements of the same tissue

Data sets

- 45 publicly available sample collections
  - accessible gene expression data measured on Affymetrix HG-U133A or HG-U133 plus 2.0 arrays
  - sex information
- Cancer datasets
  - 10 NSCLC (non-small cell lung cancer) datasets
  - 7 colon cancer datasets
  - 5 other cancer datasets
  - 7 non-cancer related datasets
  - 8 breast cancer datasets
  - 4 ovarian cancer datasets
  - 4 prostate cancer datasets
- In total 4913 samples (3034 females, 1879 males)

Male-female classifier

- Selection of features (probe sets)

For every probe set of the overlap of the Affmetrix© HG-U133A and HG-U133 plus 2.0 array

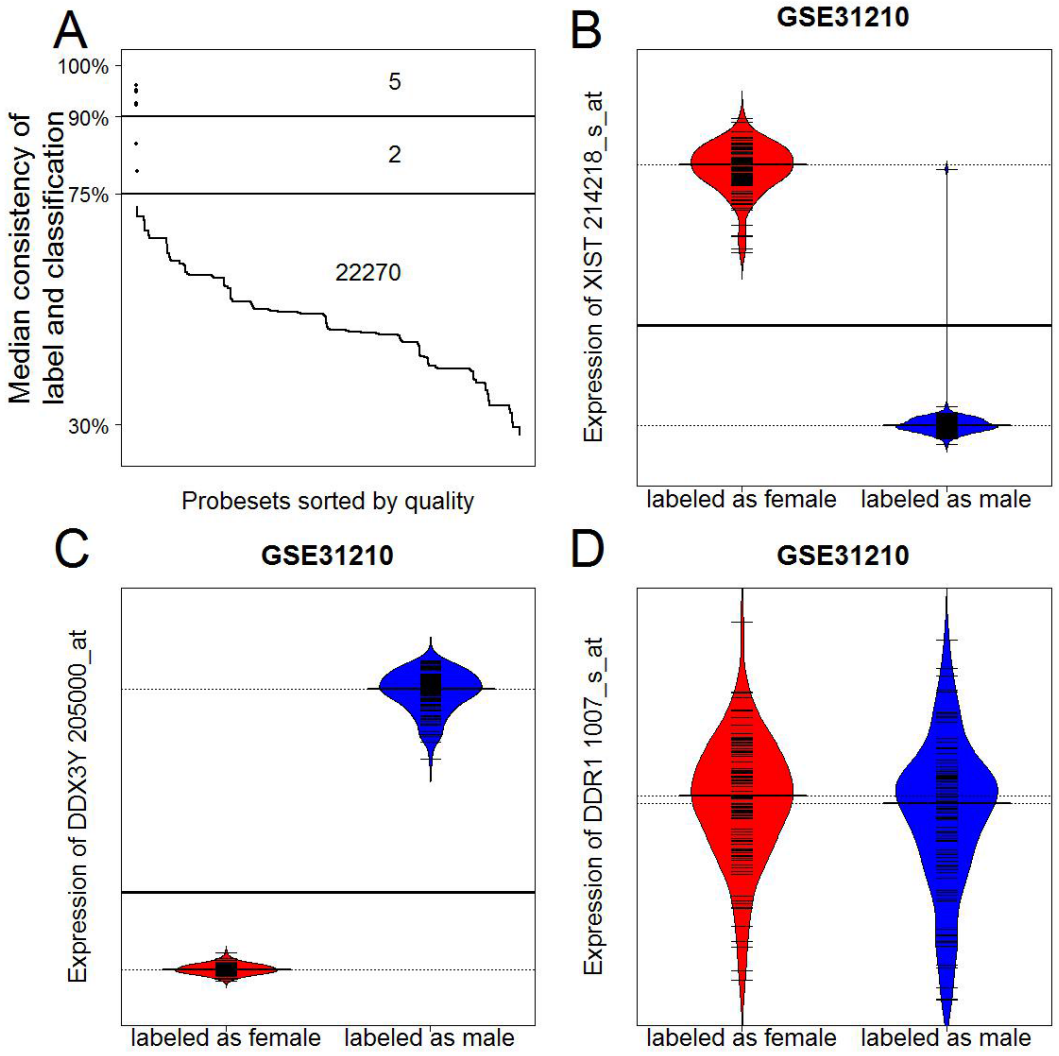
1. Estimate location and scale with *median* m_0 and *Rousseeuw-Croux estimator* $Q_{n,0}$ of the **lower** expression values (males *or* females);
2. assign a sex-specific Gaussian distribution f_0 to the low expression values $N\left(m_0, (2.22 \cdot Q_{n,0})^2\right)$;
3. calculate the 99.9% quantile of the distribution f_0 $q_{0.999}$ and classify a sample to the group with larger values if $x > q_{0.999}$;
4. compute the median of correctly classified samples across all training (NSCLC) datasets
   $\rightarrow$ *median classification accuracy (MCA).*

Male-female classifier

Selection of suitable features for a sex classifier

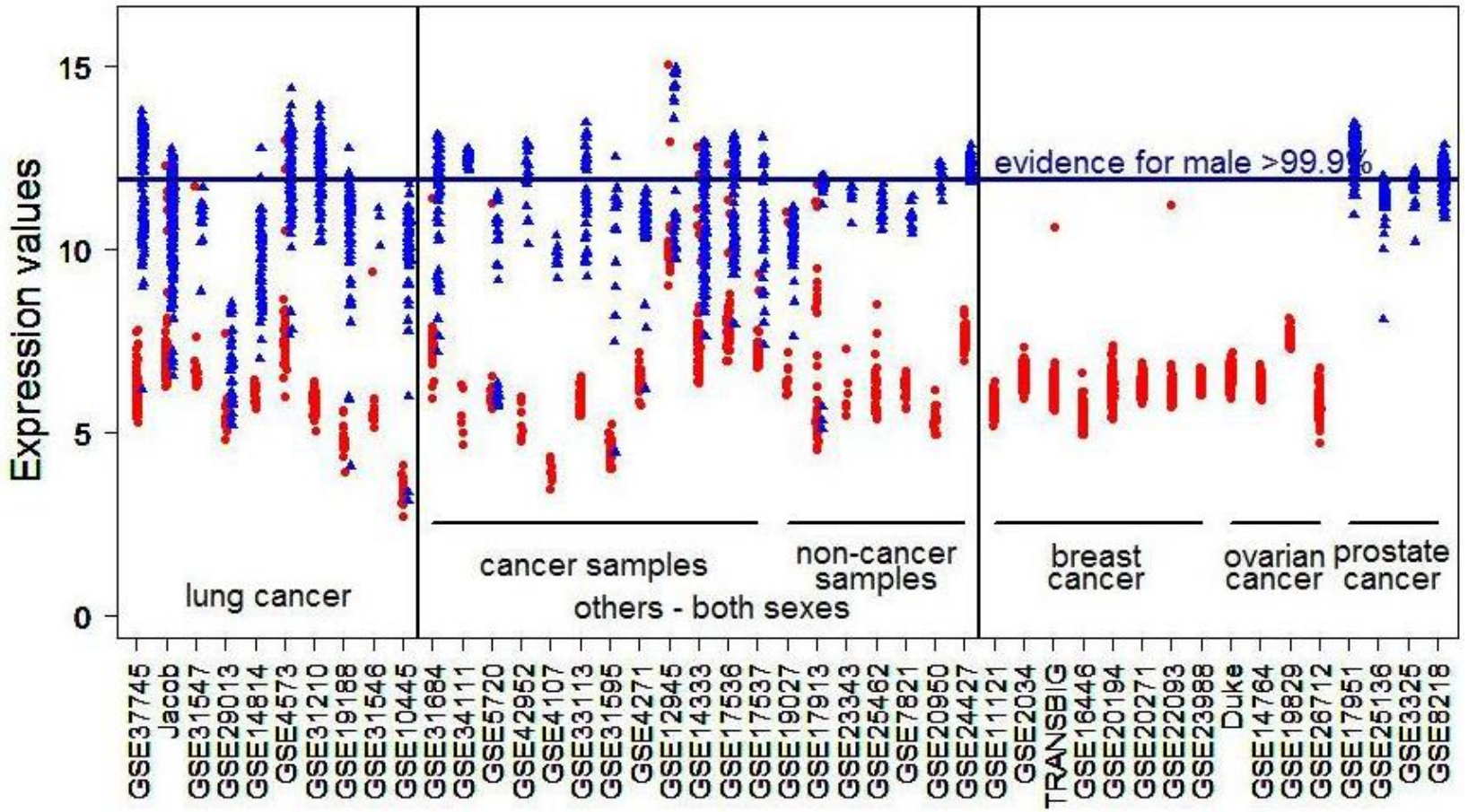
- Only five of 22 277 probe sets with MCA > 90%!
- The top four probe sets were included in our classifier
- Two map to the XIST gene (221728_x_at and 214218_s_at), located on the X chromosome and highly expressed in females
- Two map to RPS4Y1 (201909_at) and DDX3Y (205000_at), respectively, both located on the Y chromosome and highly expressed in males

Male-female classifier



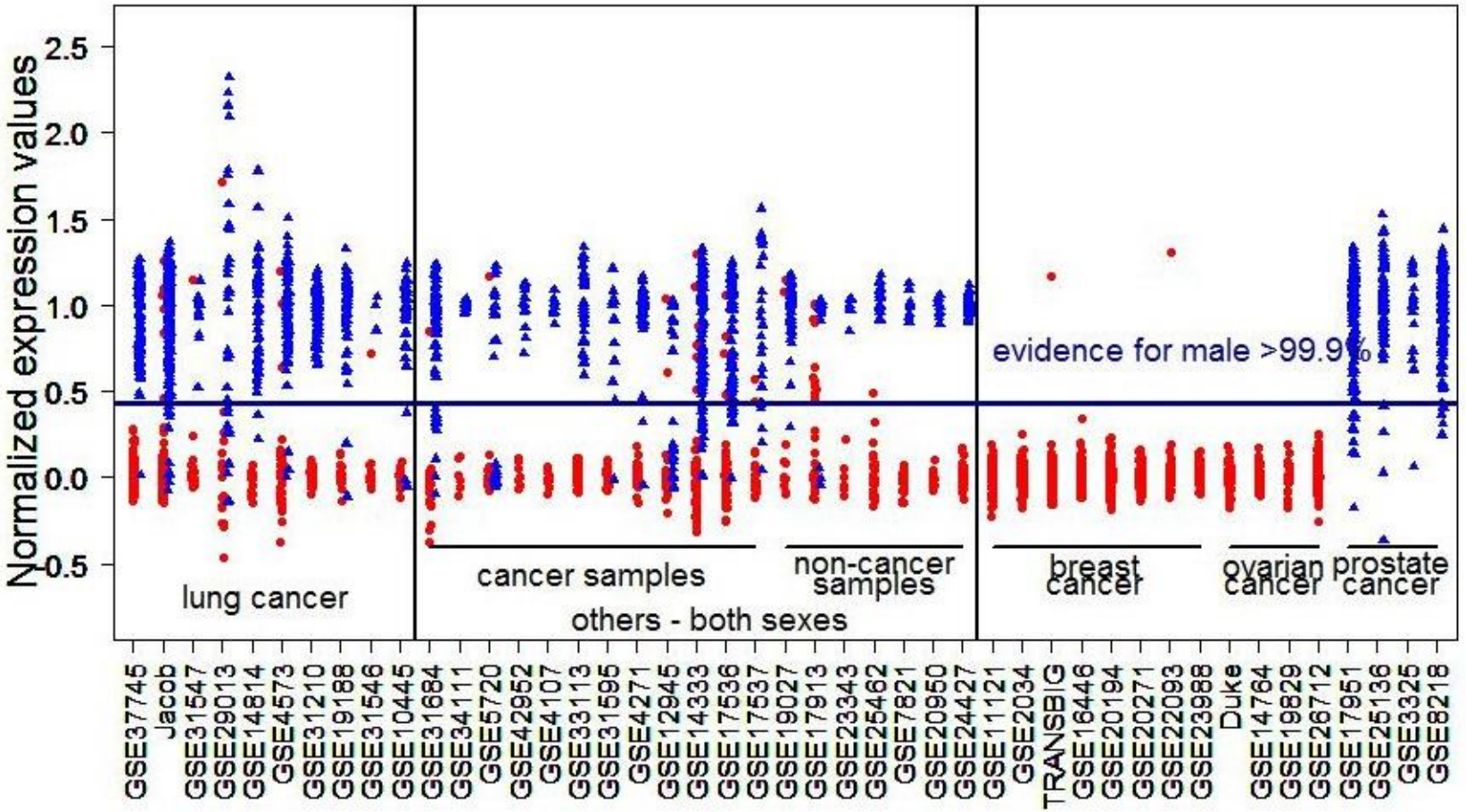
Normalization

- Raw expression values of probe set 201909_at (RPS4Y1)



Normalization

- After normalization (group medians shifted to 0 and 1)



Normalization

- For datasets containing both sexes: Cluster samples in two groups and shift group medians to 0 and 1, respectively.

$$\tilde{x} = \frac{x - m_0}{m_1 - m_0}.$$

- For single sex datasets: Use pooled median and Rousseeuw-Croux estimator of normalized training datasets to normalize the data

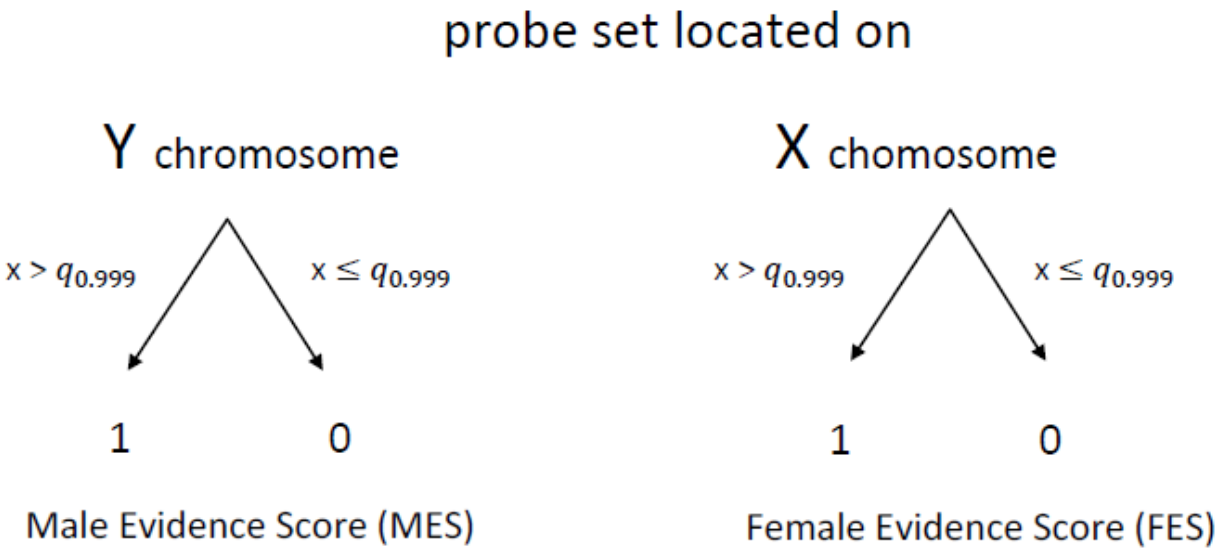
$$\tilde{x} = \frac{x - m^{\mathsf{s}}_{\text{training}}}{Q^{\mathsf{s}}_{n,\text{training}}} \cdot Q^{\mathsf{s}}_{n,\text{training}},$$

s = male/female, respectively, and add 1 if s is the group with larger median values.

Classification

Sex classification

For each of the four selected probe sets assign a group-specific Gaussian distribution $N\left(m_0, (2.22 \cdot Q_{n,0})^2\right)$ (m_0, $Q_{n,0}$ estimated from training data) with 0.999 quantile $q_{0.999}$.



Classification

Sex classification

Define:

$$\text{MaxFES} = \max_i (\text{FES}_i) \quad \wedge \quad \text{MaxMES} = \max_i (\text{MES}_i)$$

Sample labeled as female

MaxFES \ MaxMES	0	1
0	unconfident	misclassified
1	correctly classified	unconfident

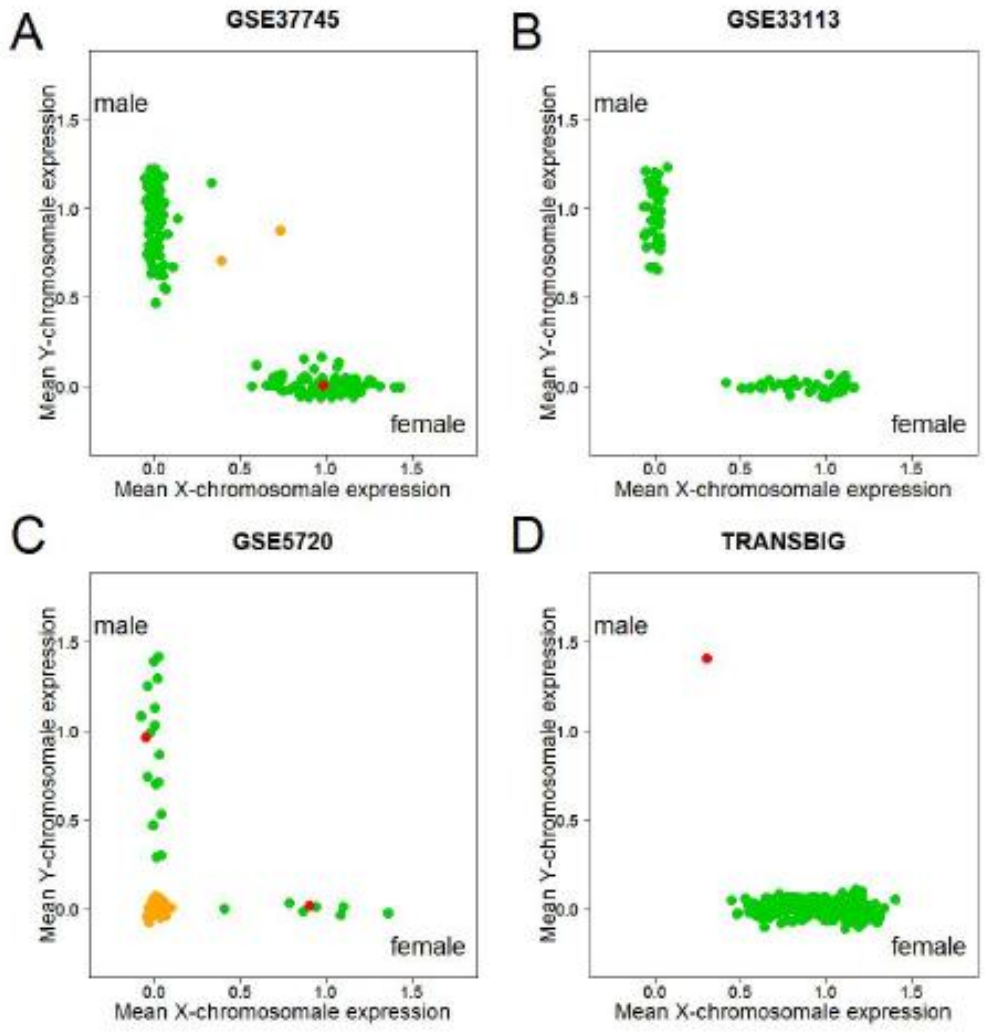
Sample labeled as male

MaxFES \ MaxMES	0	1
0	unconfident	correctly classified
1	misclassified	unconfident

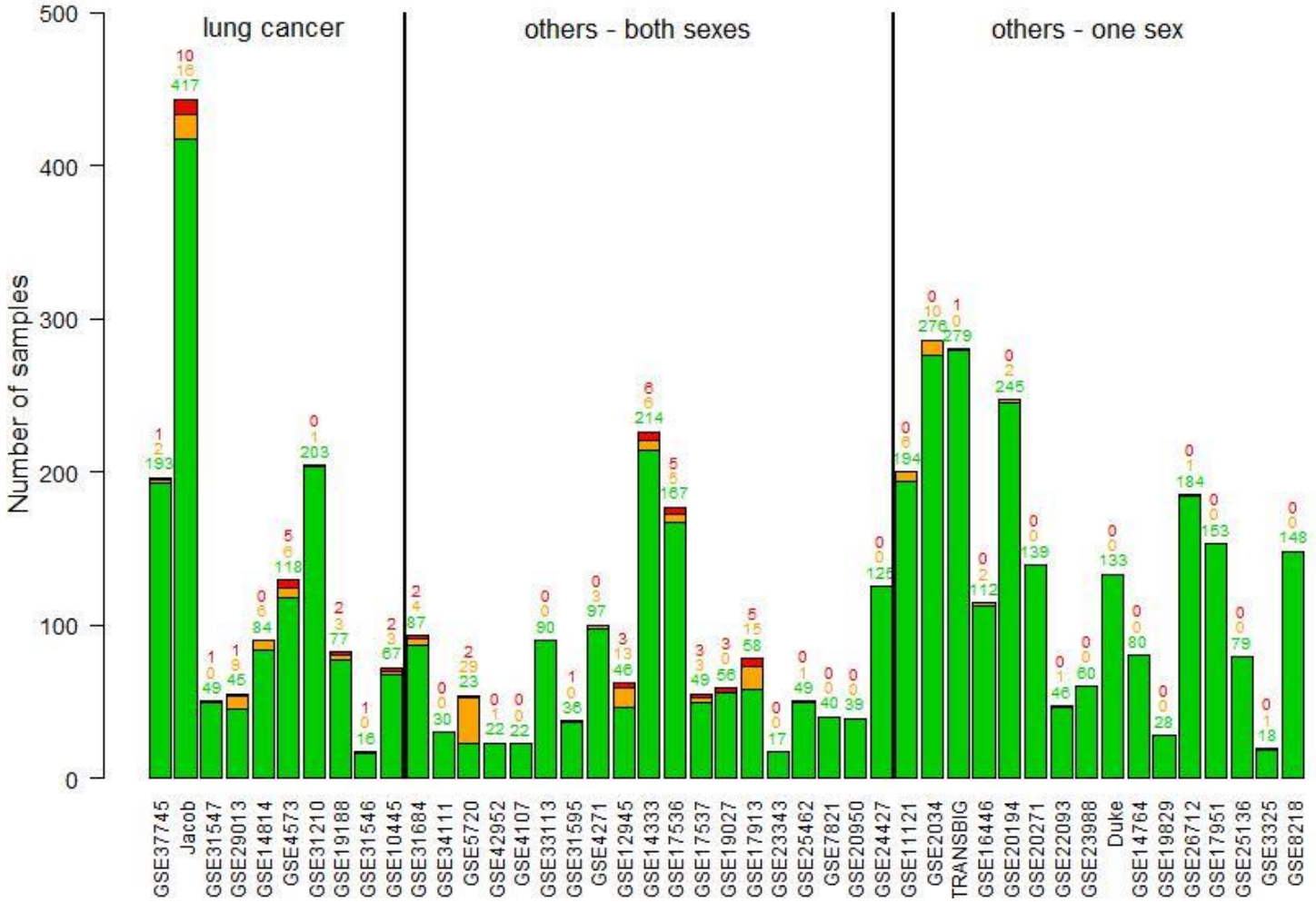
- Only if one sex evidence score yields clear evidence and the other does not, we classify to the respective sex

Male-female classifier – Results

- Results for four data sets
- Green: Predicted and true sex label agree
- Red: Predicted and true sex label disagree



Male-female classifier – Results



Male-female classifier – Results

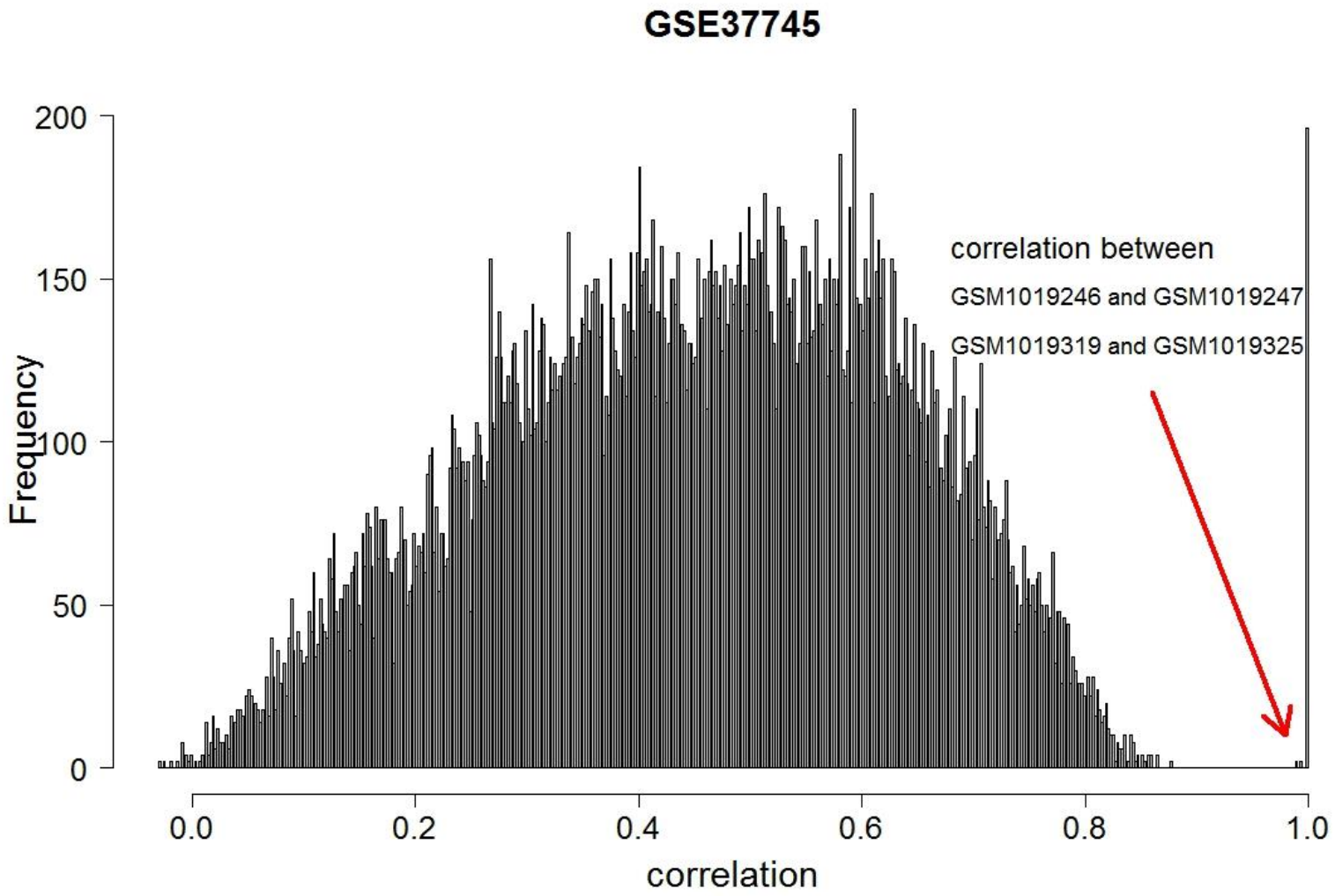
- The male-female classifier was applied to all 45 cohorts, including 4913 patients (3034 females, 1879 males).
- In total, 54 patients (1.1%) were clearly „misclassified“
- 149 (3.0%) were labelled „unconfident“
- For 15 out of the 45 cohorts (33.3%), all samples were „correctly classified“
- In 18 of the 45 cohorts (40%), at least one clearly „misclassified“ sample was detected

Duplicated measurements

Heuristic procedure

- Separately for each cohort, select the 1000 probe sets with highest variance
- Calculate Pearson correlation coefficient between all pairs of samples in each cohort
- Cutpoint to discriminate between pairs of measurements from different samples and duplicated measurements: The largest distance between all ordered correlations.

Duplicated measurements



Duplicated measurements

- Duplications in 15 of the 45 cohorts (33.3%).
- In total 32 duplicates were detected.
- 9 of the 54 sex „misclassified“ assignments (16.7%) could be explained by duplicated measurements

- Validation analysis for GSE37745
  - DNA and RNA samples of misclassified sample (GSM1019247) and five additional control samples were prepared from the original biobanked tissue
  - Correlation between old and new sample of GSM1019247 was only 0.46, otherwise >0.9
  - Clear evidence for duplicated measurements

Acknowledgements

Biostatistics

Jan Hengstler

Eugen Rempel

Katrin Madjar

Marianna Grinberg