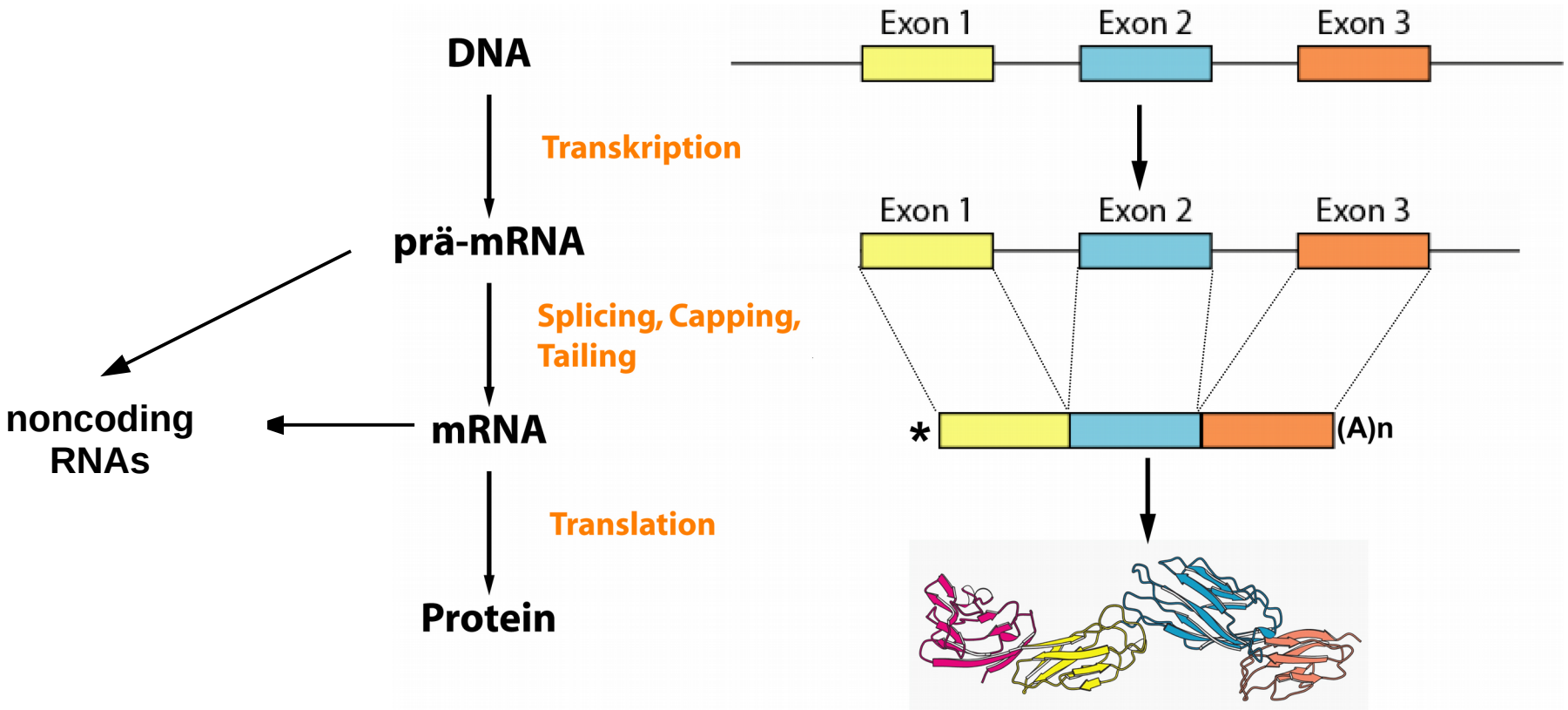
RNA-Seq

Dr. Barbara Hutter
Division of Applied Bioinformatics (B330)
Team Leader Clinical Bioinformatics

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

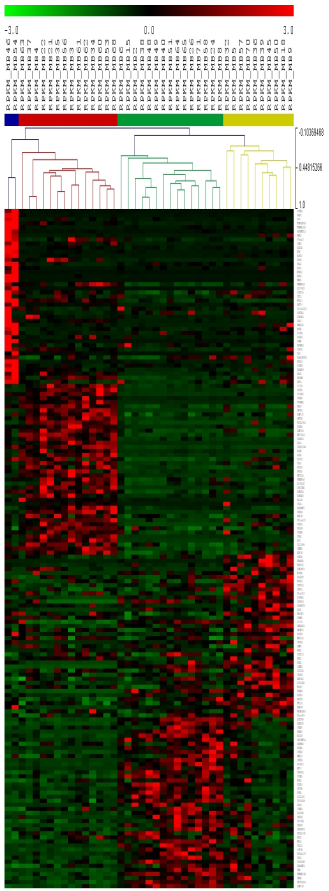
Gene Expression



http://upload.wikimedia.org/wikipedia/commons/2/20/Eukariotische_Genexpression.png

Most Common Use of RNA-Seq

- Quantification of transcripts: “digital gene expression”
  - Classification of transcriptomes for different cell types, developmental stages, conditions
  - Differential expression
    - over- and underexpression
    - clustering and classification
- Fusion genes
- Alternative splicing

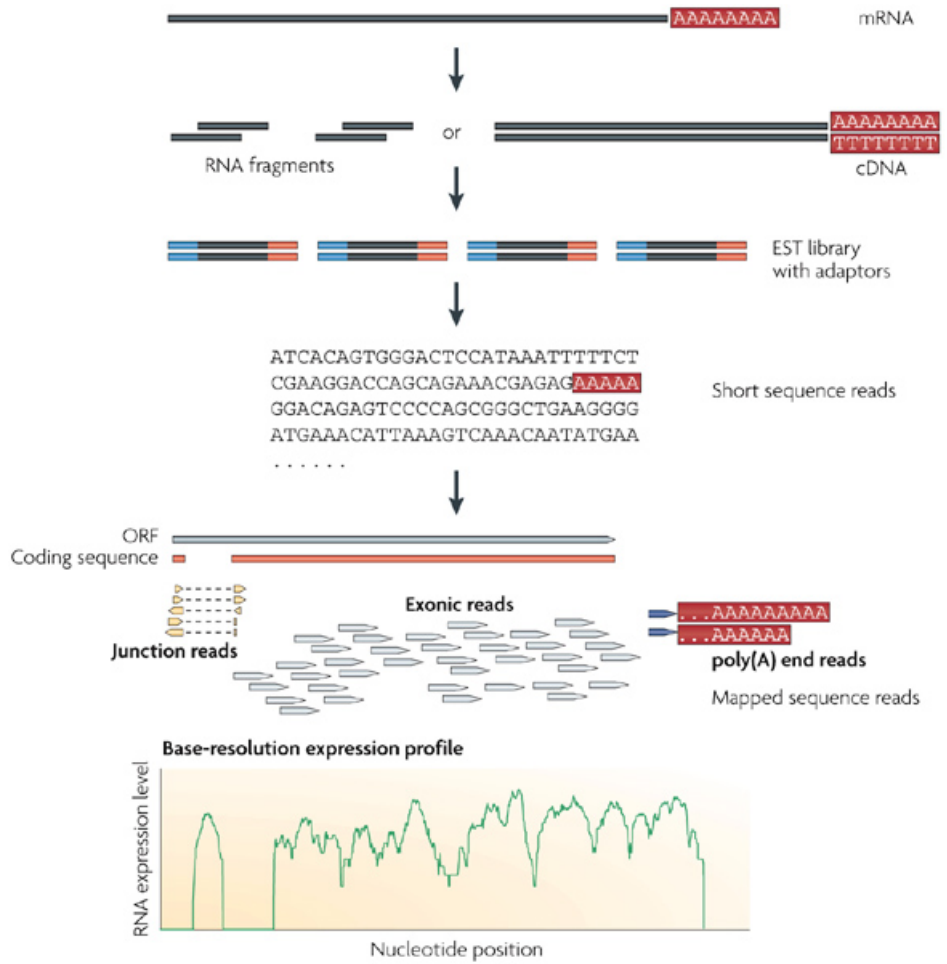


RNA-Seq vs. Microarrays

- Microarrays
  - only known transcripts and isoforms
  - intensities, have to be normalized to be comparable between different experiments
  - limitations known, established protocols
- RNA-Seq
  - good correlation with array data
  - improved identification of lowly expressed genes
  - many proprietary approaches

RNA-Seq

- Library preparation
- Sequencing
- Read mapping
- Read counting



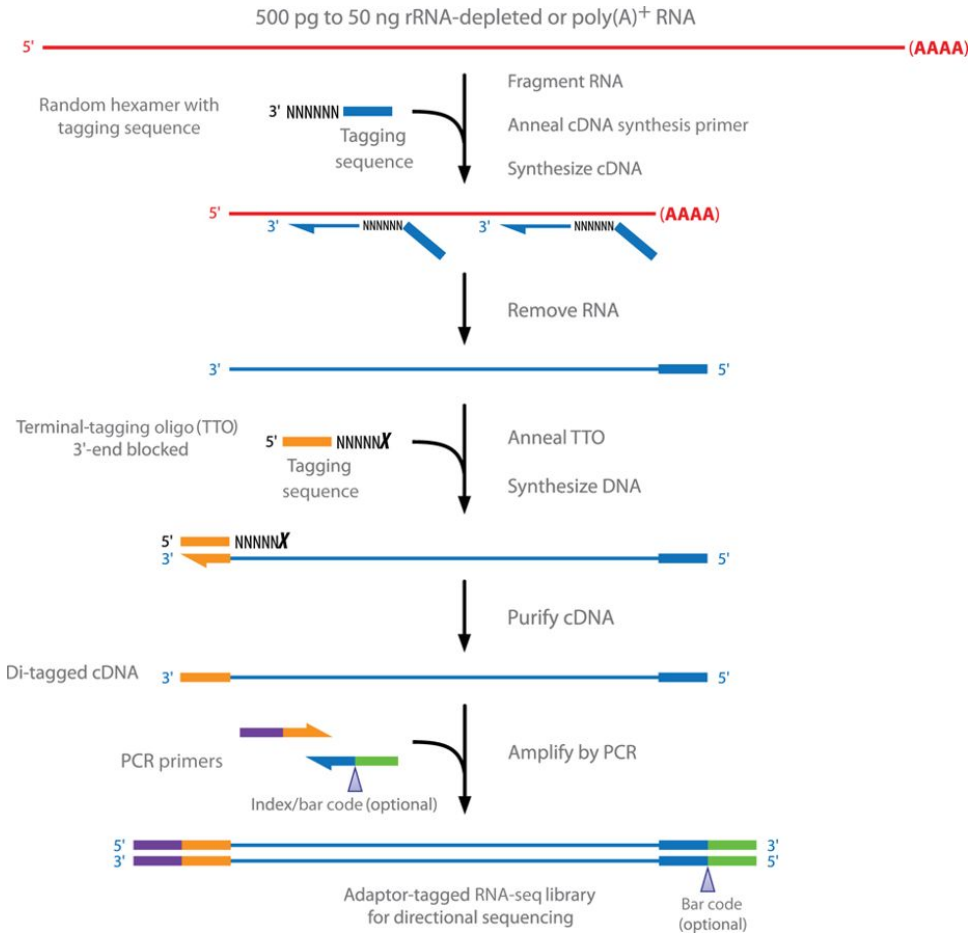
Expression profile in base resolution for a yeast gene with one intron

Z. Wang et al. 2009

Nature Reviews Genetics 10:57-63

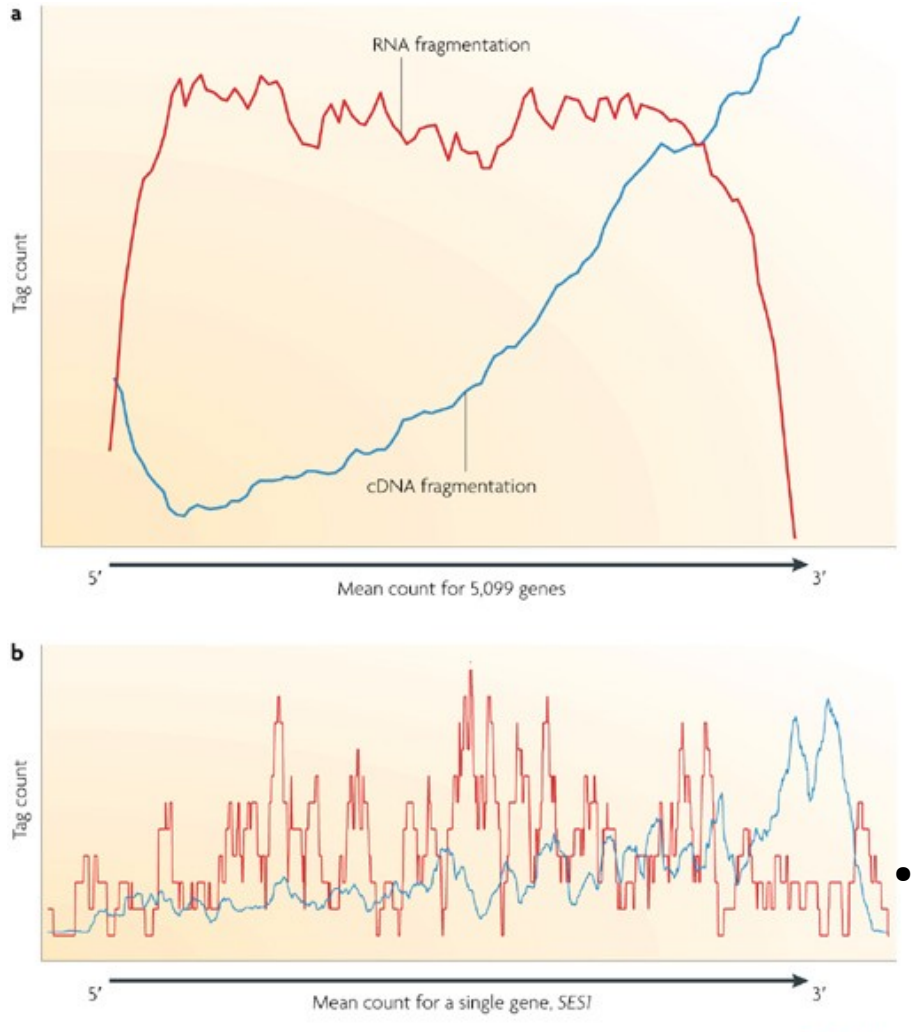
RNA-Seq Library Preparation

- cDNA library
  - poly-A tail of mature mRNA => oligo-T primer for reverse transcription
- RNA fragmentation
  - add DNase
  - get also unspliced RNAs
  - ribosomal RNA (rRNA) depletion
- barcodes for multiplexing
- specialized protocols for strand-specific reads (=> identify antisense transcripts)
- small RNAs (miRNA , ...) different



Pease & Sooknanan Nature Methods 2012 9, 310

Systematic Errors by Fragmentation Method



Fragmentation into pieces of 200-500 bp

- Oligo-T cDNA
  - with DNAse I or sonication (ultrasound)
  - the more 5', the fewer fragments
- RNA
  - by hydrolysis or nebulization
  - 5' and 3' ends underrepresented
- Tag count: average values for 5099 yeast genes

Yeast gene SES1 (Seryl-tRNA Synthetase)

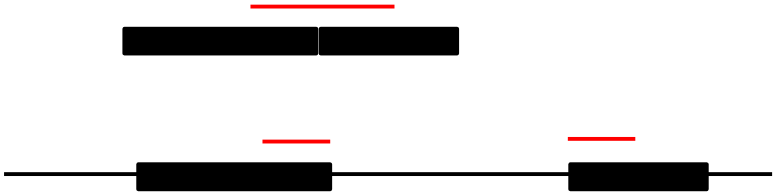
Z. Wang et al. 2009, Nature Reviews Genetics 10:57-63

Mapping RNA-Seq Reads I

- To genome
  - Disadvantages:
    - reads that span exon-exon junctions (splice sites) are not or wrongly aligned
      - the longer the reads, the higher the probability
    - isolated exons shorter than the read size are not covered
    - distances for paired end reads are incorrect if there is an intron in between
  - Advantages:
    - new genes, exons, splice variants, noncoding transcripts (e.g. of repetitive elements) can be detected

Mapping RNA-Seq Reads II

- To transcriptome
  - Advantage:
    - reads that span exon-exon junctions are aligned correctly
  - Disadvantages:
    - restriction to known transcripts => artifacts in mapping
- Solution:

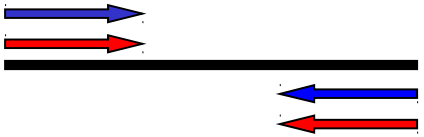


  - split read approach
  - map to transcriptome, genome and collection of splice sites simultaneously or successively
  - include annotations of exons and splice sites for genome mapping

Most Popular Mapping Tools

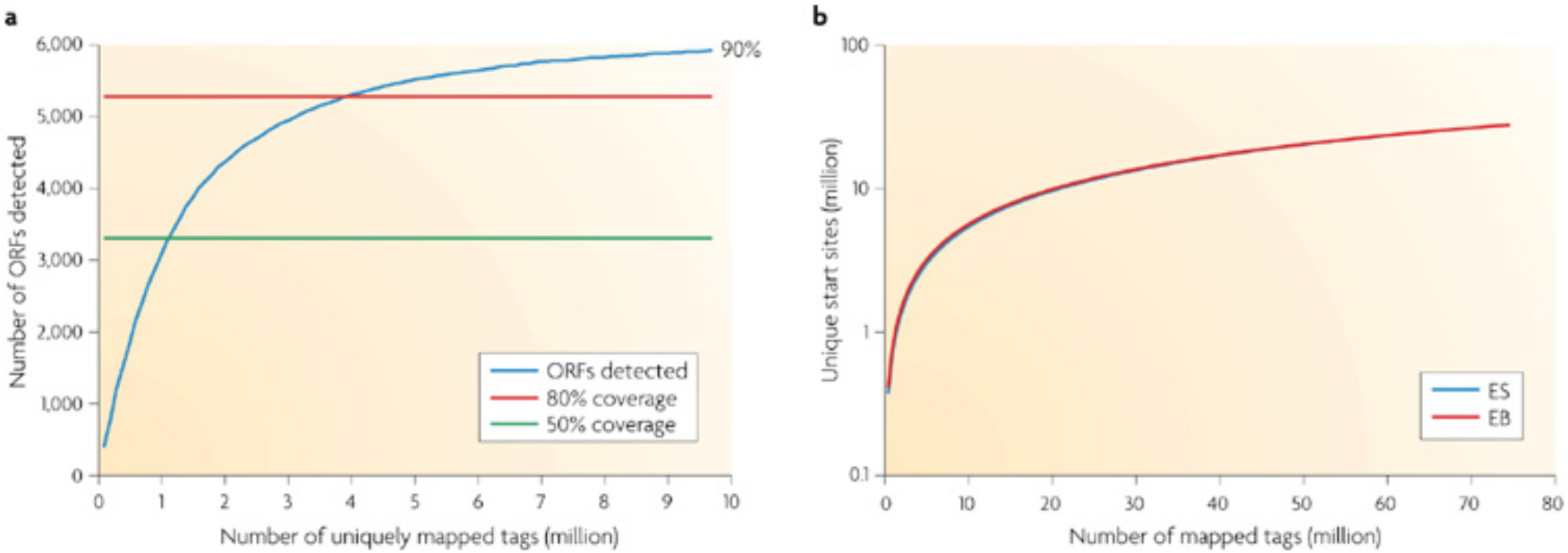
- **TopHat** http://tophat.cbcb.umd.edu/
  - Method has changed twice since publication, now split read mapping (TopHat with Bowtie allows no gaps in alignment)
  - TopHat2 optionally aligns to transcriptome first, remaining reads to genome, uses Bowtie2 => allows gaps in aligment
- **GSNAP** http://research-pub.gene.com/gmap/
  - allows indels, long-distance splicing, and translocations
  - SNP and RNA editing tolerant alignment
  - very slow
- **STAR** http://code.google.com/p/rna-star/
  - extremely fast and memory-consuming (<30 min, 30 GB RAM)
  - very well suited for fusion gene detection (chimeric alignments)

Duplicate Reads in RNA-Seq



- Duplicates have identical start coordinate(s)
  - PCR duplicate: not necessarily same length or same sequence
    - depending on library complexity (initial number of DNA/RNA fragments)
  - Optical duplicate: one cluster on the image is identified as multiple adjacent clusters
- are usually not removed for RNA-Seq:
  - PCR duplicates cannot be distinguished from saturation due to high expression
    - highly expressed gene => large amount of the same mRNA => high probability to map reads at the same position
    - removing duplicates underestimates expression of highly expressed genes
  - low library complexity can be an issue nevertheless
    - estimated library size < 30 Mio reads is potentially problematic, < 20 Mio mostly unusable
    - try to get as high RNA concentration for sequencing as possible
    - biological reasons e.g. in multiple myeloma (Ig genes)

Sequencing Depth and Coverage



Wang et al. 2009

a) 80% of known yeast genes could be found with 4 million uniquely mapped, non-duplicate RNA-Seq reads. Despite increasing sequencing depth, coverage reaches a plateau. Expressed genes: at least 4 independent reads in a 50 bp window at the 3' end.

b) Number of unique transcriptional start sites reaches a plateau at **80 million reads** in 2 mouse transcriptomes: ES, embryonal stem cells; EB, embryonic body.

What “Coverage” is Sufficient ?

- ENCODE recommendations

http://genome.ucsc.edu/ENCODE/protocols/dataStandards/RNA_standards_v1_2011_May.pdf

  - for transcript quantification, 10-30 Mio reads
  - for transcript reconstruction, up to 200 Mio reads

- own experience:
  - 50 Mio 36 bp single end reads are sufficient for differential expression
  - at least 100 Mio reads (1/3 HiSeq2500 lane) paired end 100 bp needed for detection of fusion genes

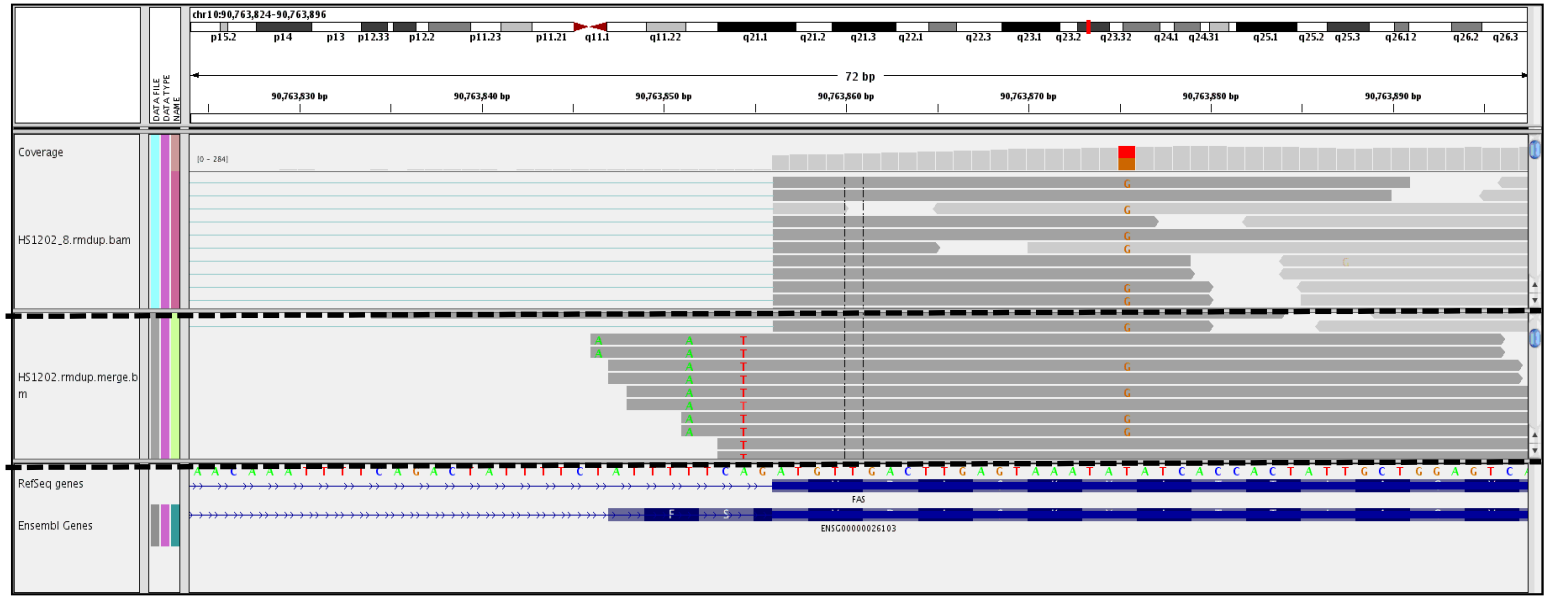
Quality Control and Issues I

- QC: base quality, number of reads, mapping rate (>80%), percent mapped (uniquely) to known exons (>90% for polyT); coverage of housekeeping genes
- RNASeqQC https://github.com/SamuelHLewis/RNASeqQC
  - estimated library size (> 30 Mio), duplication rate (< 80%), genes detected (> 22.000), intergenic rate (<5%), rRNA (ribosomal RNA) (<2%)
- Problems on library level:
  - RNA degradation
    - RIN value does not correlate with usability of RNA-Seq data
  - ribosomal RNA, DNA contamination
  - adapters: fragments are shorter than the read length; trim adapters

Quality Control and Issues II

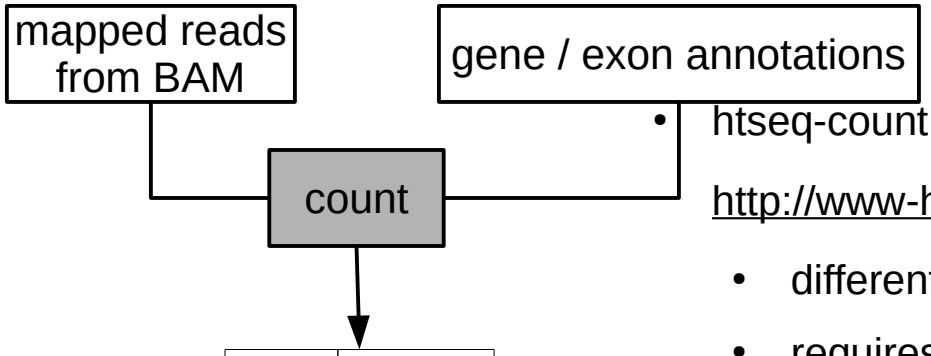
- Problems on bioinformatics level:
  - repetitive regions, paralogs => some exons not mappable even with paired end reads and splice awareness
  - long exons get more reads even (even when read counts are normalized by total exon length)
  - mapping artifacts
  - for statistical evaluation: batch effects

Alignment Artifacts



- Upper panel: correct alignment considering spliced reads reverals truncating mutation
- Middle panel: alignment to genome introduces false positive SNVs and misses the truncation
- Figure by courtesy of Marc Zapatka, DKFZ

Read Counts



A	1359
B	223
C	237
D	25
E	4

- htseq-count from the HTSeq Python package

  http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html
  - different ways of handling overlaps
  - requires BAM to be name sorted for paired end reads
- featureCounts from the R package subread (Liao et al. 2014)

  http://subread.sourceforge.net/
  - much faster, internal sorting
- coverageBed from the BEDtools package (Quinlan et al. 2010)

  https://github.com/arq5x/bedtools2
- Kallisto (Bray et al. 2016) https://pachterlab.github.io/kallisto/about
  - pseudoalignment of reads to reference transcriptome
  - very memory demanding, very fast

Fold Change

Matrix of read counts

gene	s1 (basis)	s2	s3	s4	s5	s6
A	1359	1433	3509	660	1410	3229
B	223	566	3496	273	3222	3207
C	237	241	1184	152	764	1295
D	25	265	2266	50	1599	2379
E	4	13	119	3	166	140

Fold Change

Activation / repression of genes judged by fold change

-2fold	0	2fold	5fold	10fold	50fold

gene	s1 (basis)	s2	s3	s4	s5	s6
A	1359	1433	3509	660	1410	3229
B	223	566	3496	273	3222	3207
C	237	241	1184	152	764	1295
D	25	265	2266	50	1599	2379
E	4	13	119	3	166	140

Fold Change

- fold change often given as logarithm: log fold change
  - makes distribution of expression values more symmetrical

-2fold	0	2fold	5fold	10fold	50fold

gene	s1 (basis)	s2	s3	s4	s5	s6
A	1359	1433	3509	660	1410	3229
B	223	566	3496	273	3222	3207
C	237	241	1184	152	764	1295
D	25	265	2266	50	1599	2379
E	4	13	119	3	166	140

  - Expression changes of one gene
  - Correlations and anticorrelations of expression levels across genes

Significance of Expression Changes

RNA-Seq	mapped reads	gene A	gene B
sample 1	16 000 000	2000 (0.01%)	2000 (0.01%)
sample 2	17 000 000	2100 (0.01%)	3000 (0.02%)

contingency table (cross tabulation)

gene	sample 1	sample 2
mapped to gene	2000	2100
mapped elsewhere	15 998 000	16 997 900

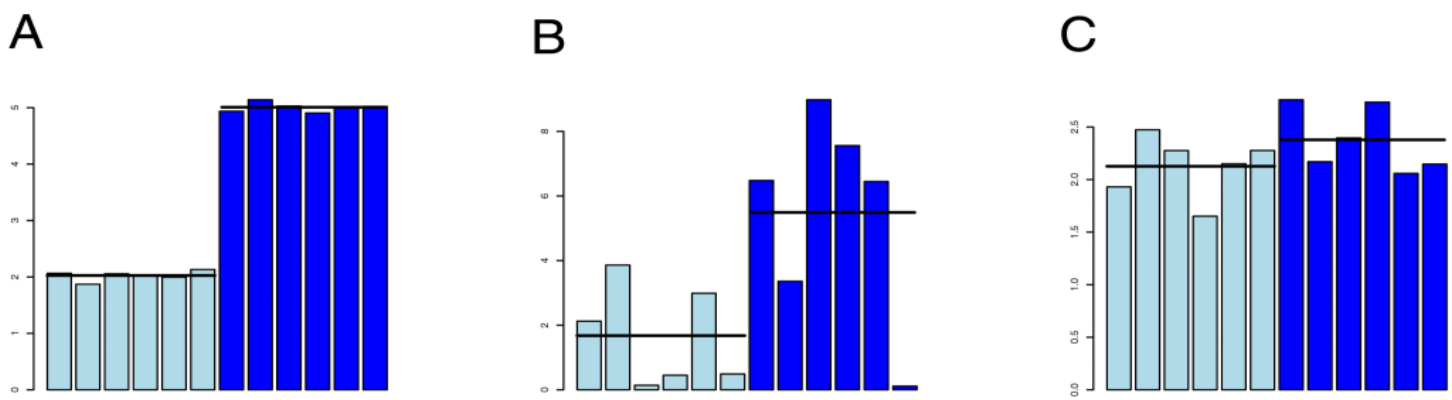
sample 2
3000
16 997 000

Fisher's exact test or Chi square test => p value

- Gene A: p = 0.71 => no significant difference between samples
- Gene B: p = $2.2e^{-16}$ (= $2.2*10^{-16}$) => expression is significantly higher in sample 2 than in sample 1

Differential Expression

- A simple test for differential expression would be simply the **fold change**, i.e. avg($r_{g,A}$) / avg($r_{g,B}$) if comparing conditions *A* and *B*
- This doesn‘t account for **variance**, i.e. the scatter around the expected value
- This is what **statistical tests** have been made for



- Slide by courtesy of Benedikt Brors, DKFZ

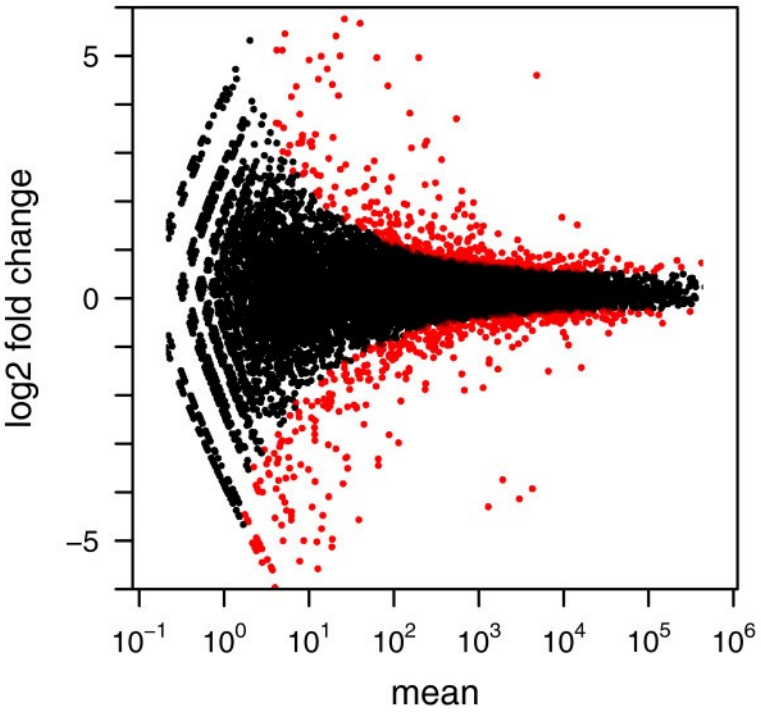
Analysis of Read Counts

- Read count 0 => fold change ∞
  - add a pseudocount of 1
- Type 1 error
  - too many false positives

=> need to be conservative

- But: type II error
  - false negatives

=> Statistical modelling of read count distribution



Anders & Huber 2010,
Genome Biology 11:R106

Tools for Differential Expression from RNA-Seq

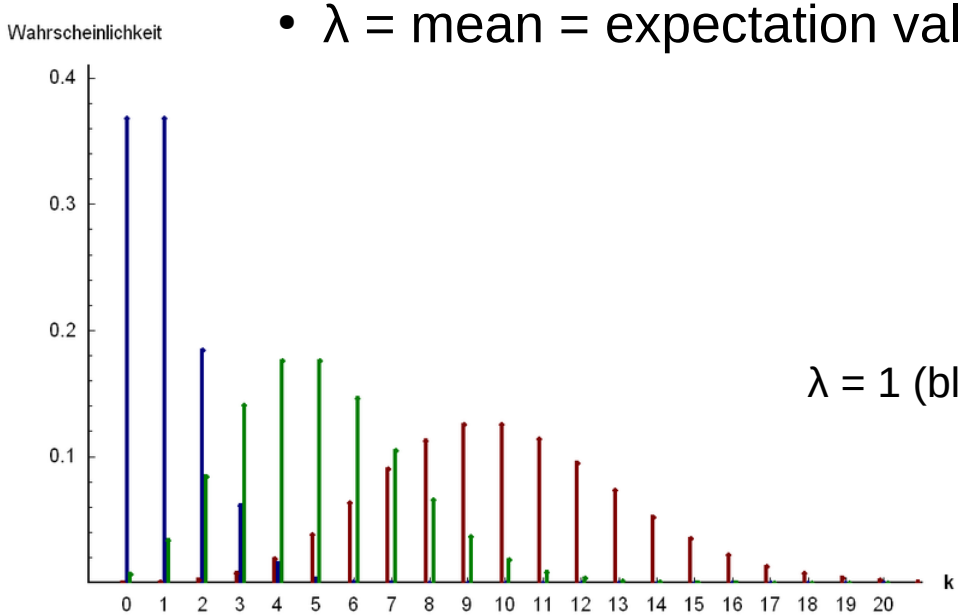
- **cuffdiff / cufflinks** http://cufflinks.cbcb.umd.edu/index.html
  - Fisher's exact test
- **DESeq** http://www-huber.embl.de/users/anders/DESeq/
  - **D**ifferential **E**xpression analysis for **Seq**uence count data
  - package in R Bioconductor
  - also suited for ChIP-Seq analysis
- **DESeq2** (Love at el. 2014)
  - improved statistical models
- **EdgeR** http://bioconductor.org/packages/release/bioc/html/edgeR.html
  - **E**mpirical analysis of **d**igital **g**ene **e**xpression data in **R**
  - similar to DESeq

Replicates

- To distinguish noise (random variance) from real (biological) differences
- Technical replicates
  - different libraries from same sample
  - does not mean running the same library on different lanes or just using different barcodes for multiplexing!
- Biological replicates
  - different samples
  - recommended: at least 6 samples per condition
- No biological replicates:
  - only for exploration and hypothesis generation
  - overestimating variance
  - only a fraction of the hits obtained with replicates is recovered

Poisson Distribution

- Assumption: read counts follow a multinomial distribution
- Poisson distribution = “distribution of rare events”
  - $P_\lambda (X=k) = \lambda^k/k! * e^{-\lambda}$
  - λ = mean = expectation value = variance

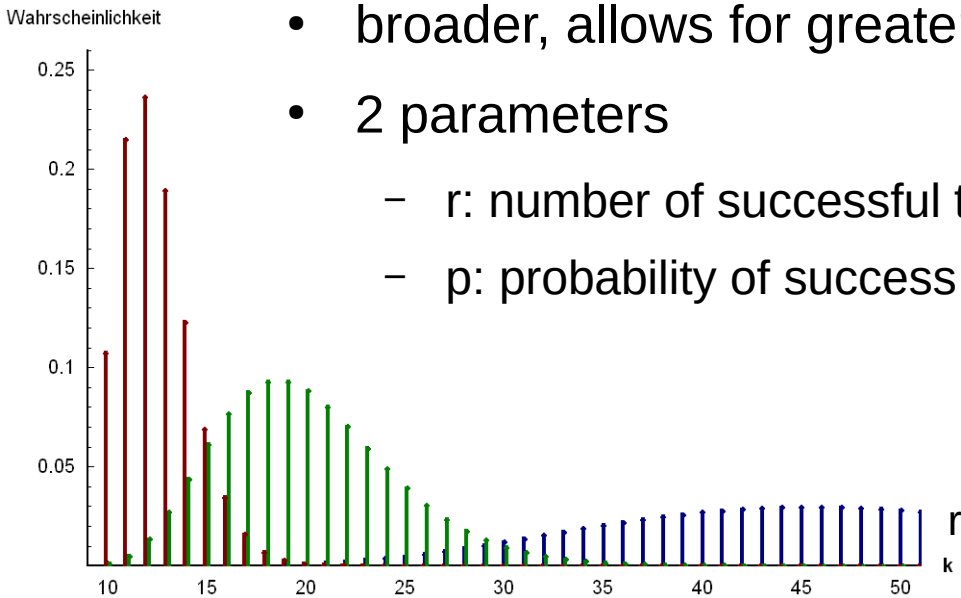


λ = 1 (blue), λ = 5 (green), λ = 10 (red)

http://de.wikipedia.org/w/index.php?title=Datei:Poisson-Verteilung.PNG

DESeq

- Poisson distribution fits for shot noise between technical replicates
- But there is extra (biological) variation in biological replicates: overdispersion => negative binomial distribution
  - variance is not identical to mean but larger
  - broader, allows for greater variance
  - 2 parameters
    - r: number of successful trials
    - p: probability of success in a single trial



r = 10; p = 0.2 (blue), p = 0.5 (green), p = 0.8 (red)

http://upload.wikimedia.org/wikipedia/commons/2/2c/Negativ_Binomial_Distribution.PNG

False Negatives

- Noise between replicates must be lower than that between different conditions
- Genes with low counts
  - shot noise “overwhelms” real differences
  - need to sequence deeper, or leave these genes out of the analysis
- For high counts, false negatives due to conservativeness
  - more replicates needed
- Big fold changes may mask smaller ones
  - do a second DEG analysis after removing genes identified as differentially expressed in first round
  - DEseq2 has improved models

Drawbacks of Read Count Data

- High read counts != highly expressed
  - long genes get more reads
- Number of reads and those that are aligned can be very variable depending on experiment or sequencing method
  - sophisticated scaling by DESeq(2)
- Only pairwise comparisons between conditions
- No time courses
- Different genes cannot be compared to each other

RPKM and FPKM

- Normalisization of read counts by length of the transcript (or single exon) => compare expression levels of different genes
- RPKM: ***R****eads* ***P****er* ***K****ilobase of exon per* ***M****illion mapped reads*
- Often refered to as “mRNA copies per cell”
- RPKM = ($ReadsA_{sX}$/ ExonsumA / 1000) / ($allReads_{sX}$ / 1 000 000)
- FPKM: ***F****ragments* ***P****er* ***K****ilobase of exon per* ***M****illion mapped reads*

Transcripts per Million

$$\mathrm{TPM}_g = \frac{r_g \cdot l_r \cdot 10^6}{L_g \cdot T} \qquad T = \sum_{g \in G} \frac{r_g l_r}{L_g}$$

- r_g: number of reads for gene g
- l_r: read length
- L_g : length of gene/transcript/exon
- T: number of transcripts

- Proportional to RPKM, but with a sample-specific scaling factor

Slide by courtesy of Benedikt Brors, DKFZ

Expression judged by RPKM

- Rule of thumb:

< 1 very low	< 10 low	10 - 30 moderate	30 - 70 quite high	70 - 100 high	>> 100 over-expressed

- RPKM > 100: housekeeping genes such as Actin; oncogenes in amplifications

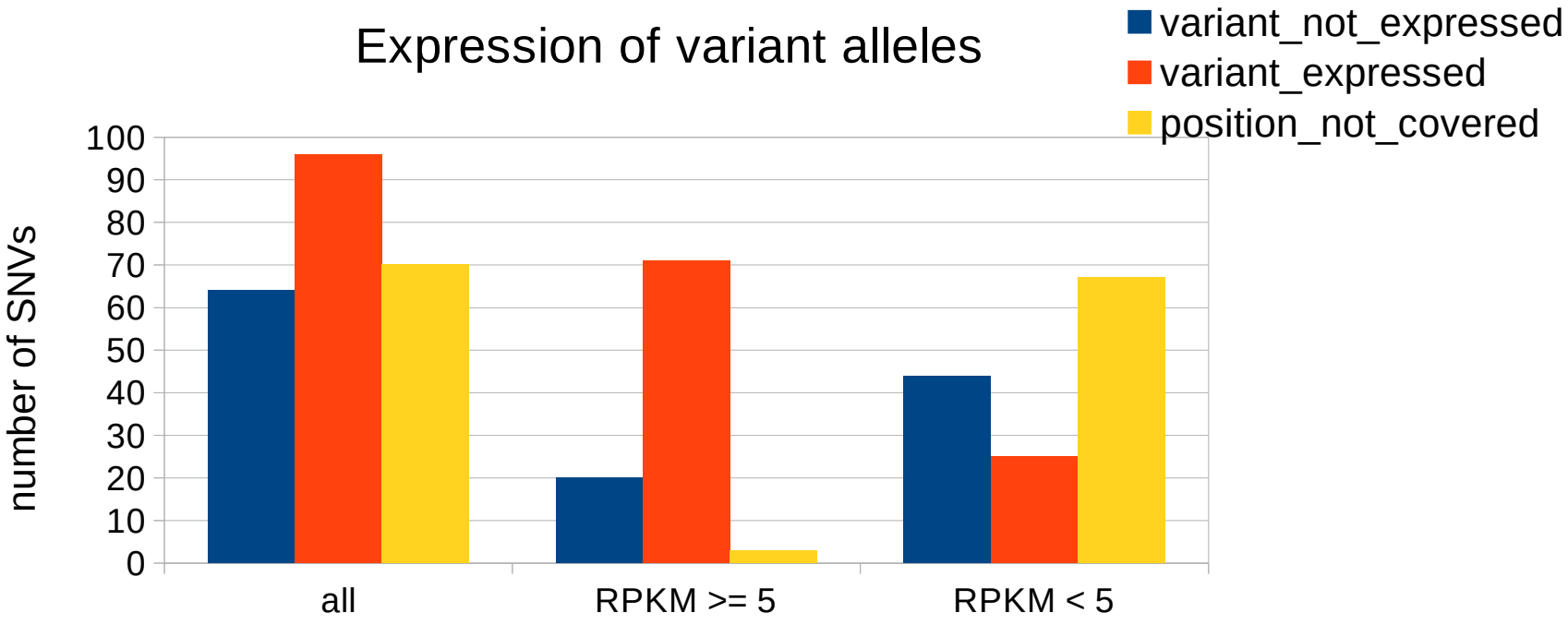
gene	s1	s2	s3	s4	s5	s6
A	61.26	61.52	150.78	29.8	59.85	131.61
B	6.58	15.33	99.17	7.33	88.45	85.51
C	5.61	5.60	26.67	3.44	16.94	27.72
D	0.52	6.79	63.95	1.19	41.73	62.82
E	0.23	0.89	7.51	0.21	10.29	8.41

Other Applications of RNA-Seq

- Identification of novel transcripts and genes
  - new exons, isoforms, and alternative transcription start sites
  - novel protein-coding and noncoding mRNA
- Variant calling
- Allele-specific gene expression
- RNA editing
- Reflexion of DNA changes in RNA
  - are mutated alleles expressed at all?
  - do SNVs change splicing?
  - do mutations in promoter regions influence transcription?

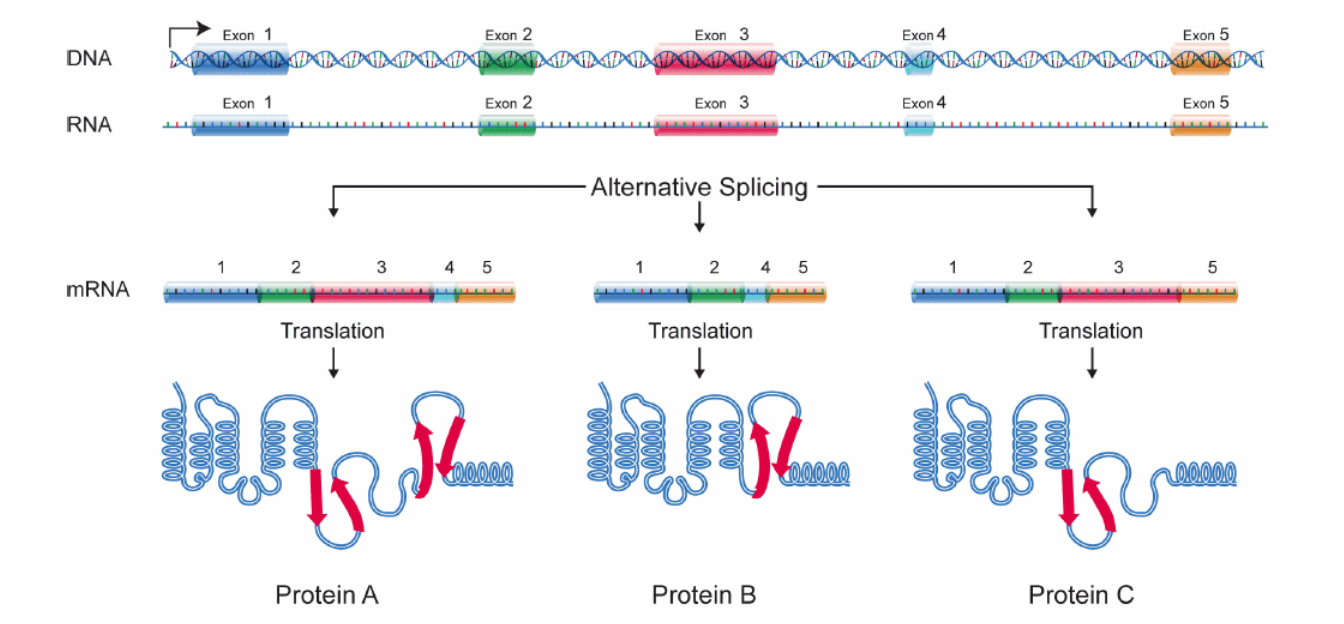
Reflexion of SNVs in RNA

- samtools mpileup at positions of SNVs detected in genome
- RPKM threshold to distinguish between low / absent expression of the gene and lack of coverage at the SNV position



Isoforms and Alternative Splicing

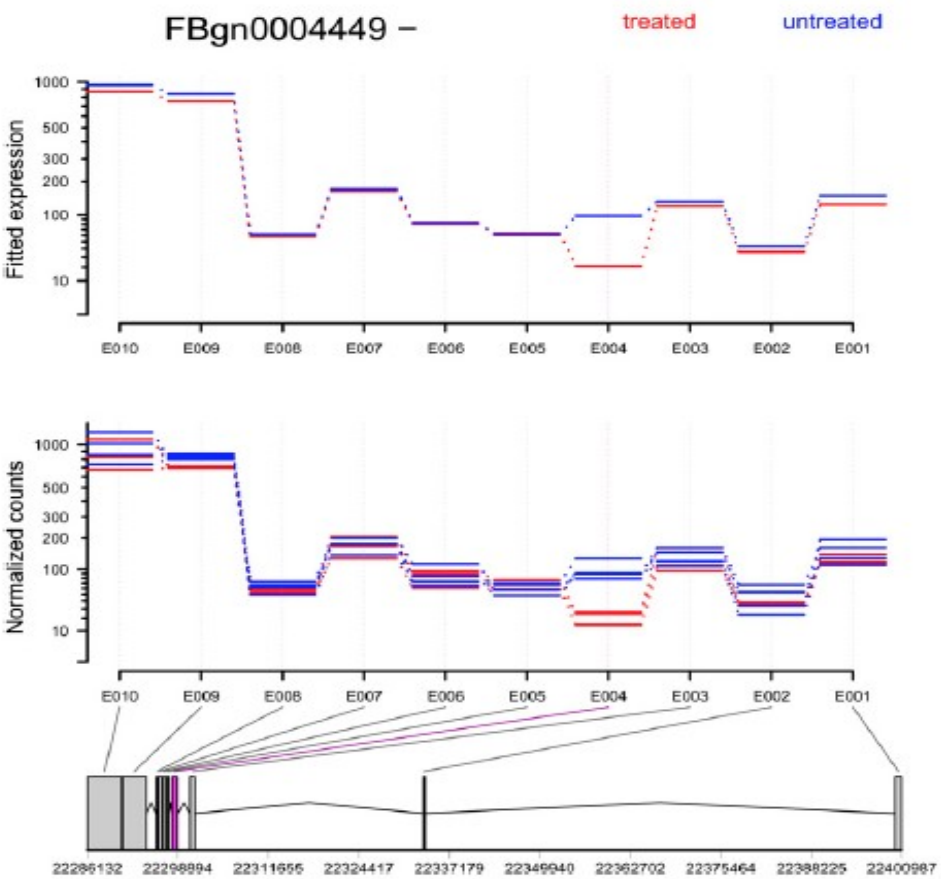
- Tissue-specific usage of alternative isoforms and promotors
- noncoding RNAs



http://de.wikipedia.org/w/index.php?title=Datei:DNA_alternative_splicing.gif

DEXSeq

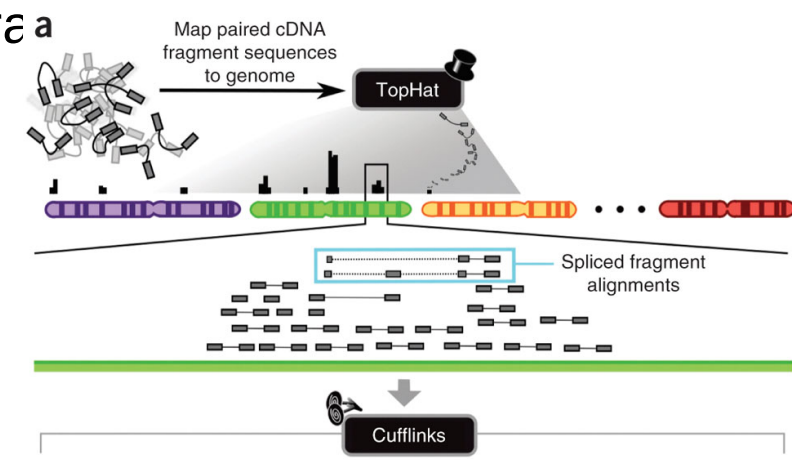
- Similar to DESeq
- Comparison of all exons over a gene (based on fixed gene model)
- Exon skipping
- Alternative exon usage
- Alternative splice sites

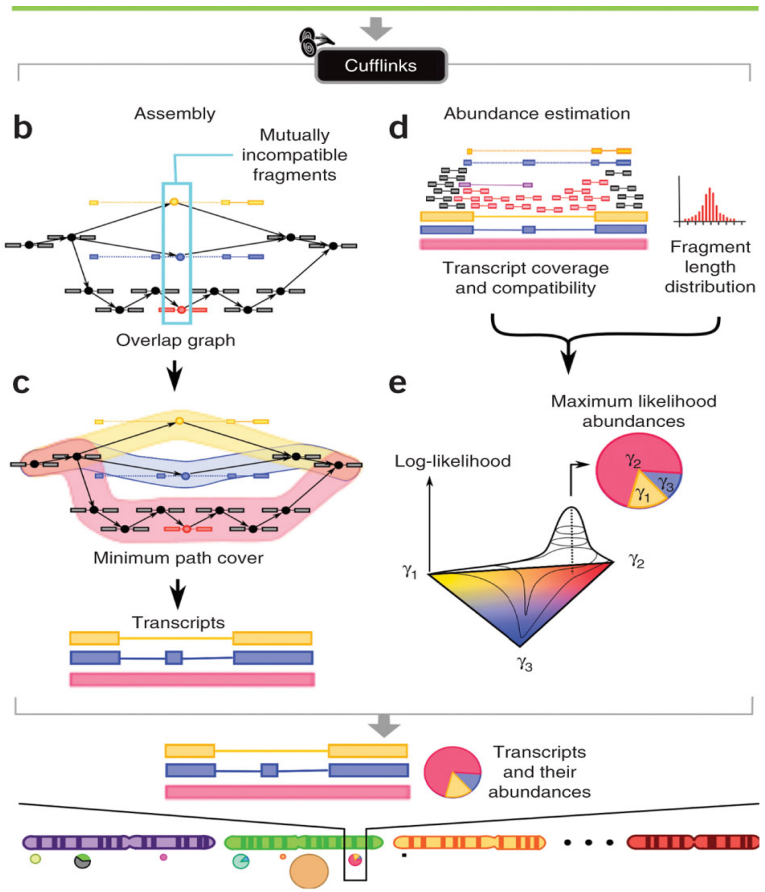


Anders S, Reyes A & Huber W (2012)

Cufflinks

- Assemby of reads into transcripts
- At the same time calculate FPKM
- Estimate probability how many reads belong to which transcript
- But: transcripts detected by cufflinks often do not fit well with known tra





Trapnell et al. 2010

Fusion Gene Detection

- SOAPfuse (Jia et al. 2013) https://sourceforge.net/projects/soapfuse/
- deFuse (McPherson et al. 2011) https://sourceforge.net/projects/defuse/
- STAR-Fusion https://github.com/STAR-Fusion/STAR-Fusion
- arriba https://github.com/suhrig/arriba
  - in-house tool based on STAR chimeric aligments
  - short runtime and high sensitivity for clinical applications
  - can detect breakpoints in introns and intergenic regions

Other Applications

- Expression during time course
- Clustering of RPKM values analogous to microarray expression data
- New alternative promoters
  - high read coverage 5' of known promoters
  - CAGE / TSS (RNA-Seq of 5‘ ends, mapping to transcriptional start sites)
- Small-RNA sequencing (miRNA etc.)
  - need trimming of adapters
  - usually mapped to sequence database of known RNAs
  - if mapping to genome, high preference for certain classes (e.g. snoRNAs)
- single cell RNA-Seq
- long read sequencing

RNA-Seq Workflow

- offered by the DKFZ Omics IT and Data Management Core Facility (ODCF) for human and mouse
- aligment with STAR
- featureCounts for read counts
- RNASeqQC for quality control
  - automated blocking of data that does not reach the standard thresholds
- fusion genes from arriba (only for human)

References I

- Anders S & Huber W (2010) Differential expression analysis for sequence count data. Genome Biol.11(10):R106 (DESeq)
- Anders S, Reyes A & Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res. 22(10):2008 (DEXSeq)
- Bray et al. (2016) Near-optimal probabilistic RNA-seq quantification. Nature Biotech. 34:525 (Kallisto)
- Dobin A et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15
- Jia W. et al (2103) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol.14:R12
- Kim D et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol.14(4):R36
- Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923

References II

- Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550
- McPherson A et al. (2011) deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. PLOS Comp.Biol. 7(5): e1001138.
- Pease J & Sooknanan R. (2012) A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. Nature Methods 9:310
- Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841
- Trapnell C et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28:511 (Cufflinks)
- Wang Z, Gerstein M & Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57-63
- Wu TD & Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873 (GSNAP)