

RNA-seq data analysis: How to find differentially expressed genes?

Eija Korpelainen
CSC – IT Center for Science, Finland
chipster@csc.fi



What will I learn?

- **Introduction to RNA-seq**
- **How to operate the Chipster software used in the exercises**
- **Differential gene expression analysis**
 - Central concepts
 - Analysis steps
 - File formats

Introduction to RNA-seq

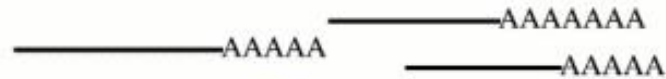


What can I investigate with RNA-seq?

- **Differential gene expression**
- **Isoform switching**
- **New transcripts (and genes)**
- **New transcriptomes**
- **Variants**
- **Allele-specific expression**
- **Etc etc**

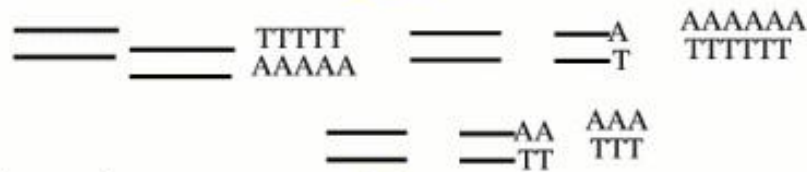
How was your data produced?

extraction of poly-A RNAs



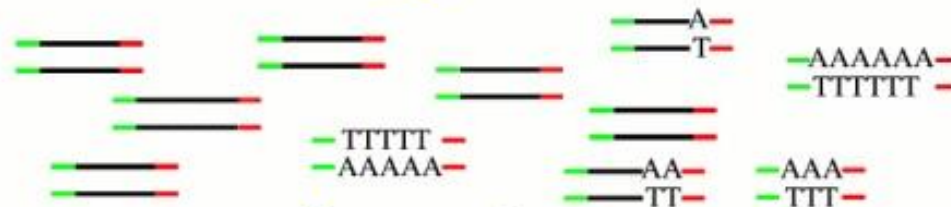
PolyA purification

conversion into ds-cDNA
and shearing



cDNA generation
& fragmentation

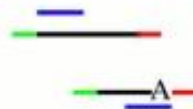
amplification and
adapter ligation



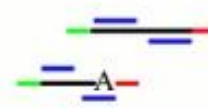
Library construction

sequencing

single end (SET)



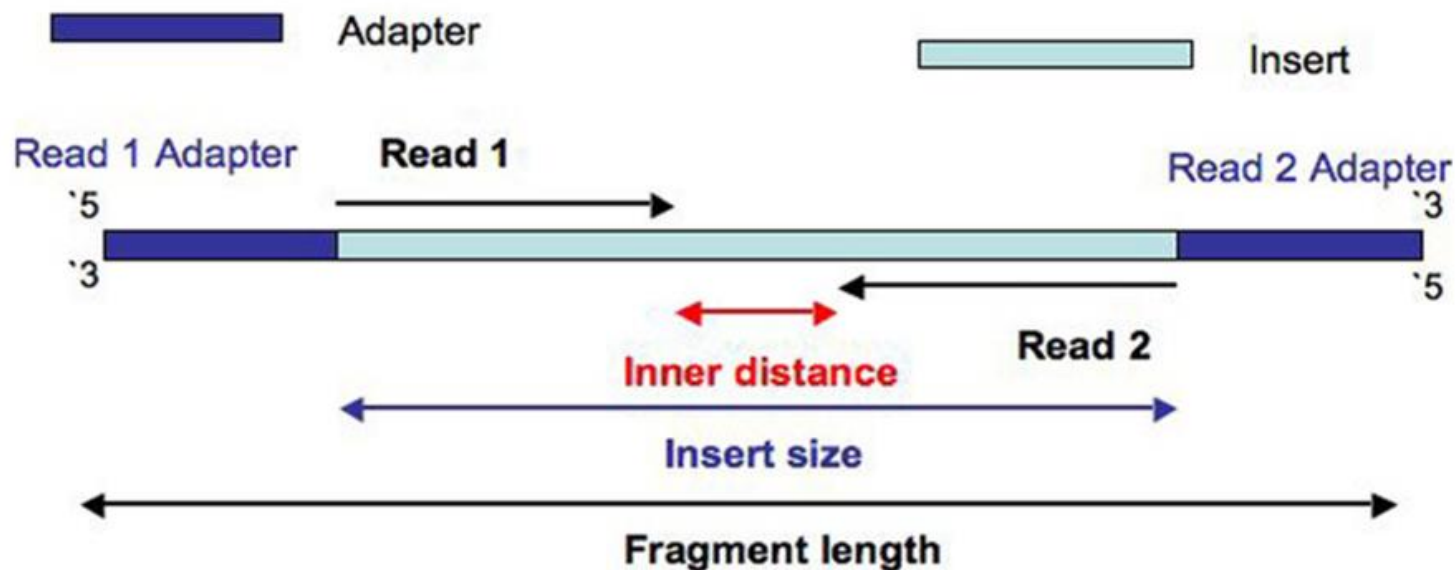
paired-end (PET)



Size selection

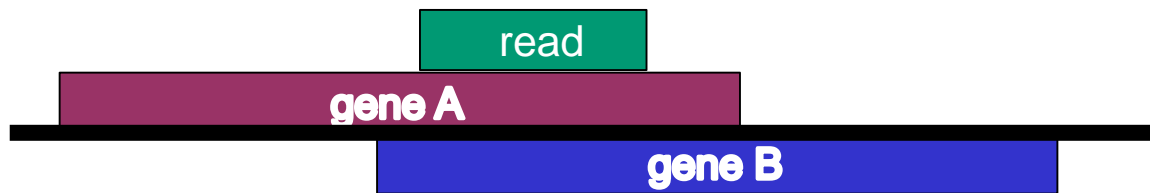
Differently sized fragments & inner distance

- Illumina reads are always of same length
- But the size of the initial RNA fragment (=insert) may vary

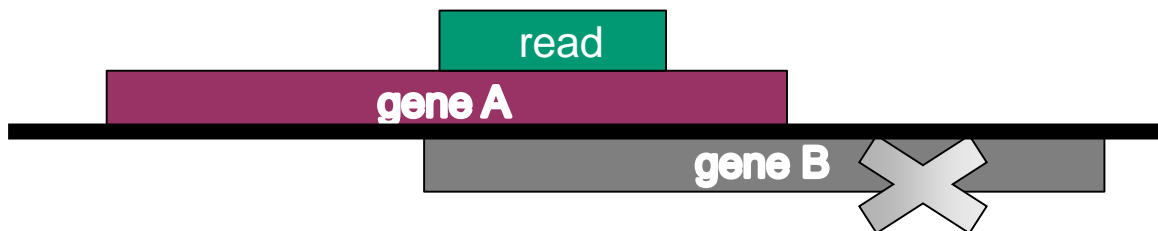


Stranded RNA-seq data

- **Tells if a read maps to the same strand where the parental gene is, or to the opposite strand**
 - Useful information when a read maps to a genomic location where there is a gene on both strands
- **Several lab methods, you need to know which one was used**
 - TruSeq stranded, NEB Ultra Directional, Agilent SureSelect Strand-Specific...



Unstranded data:
Does the read come
from geneA or
geneB?

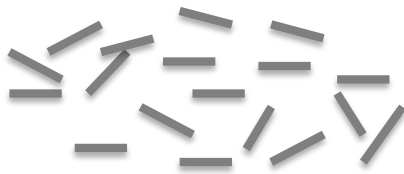


Stranded data
→ the read comes
from geneA

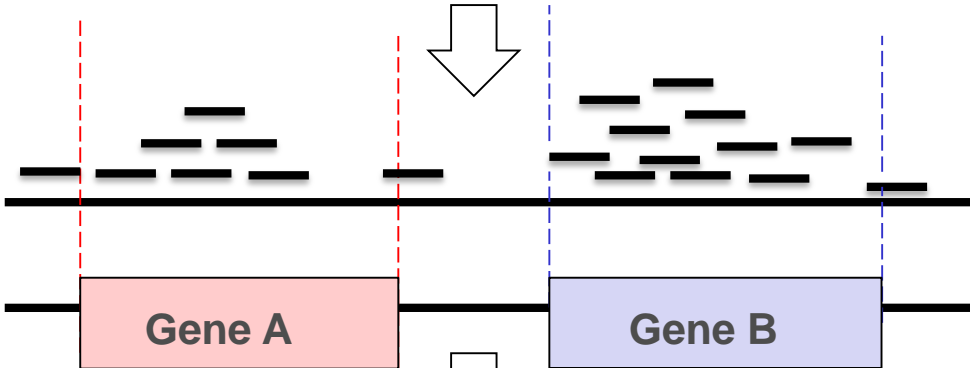


Differential gene expression analysis

DGE analysis: typical steps

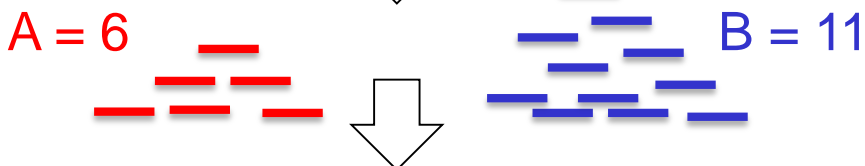


Raw data (reads)



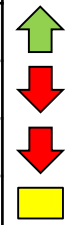
Align reads to reference genome

Match alignment positions with known gene positions



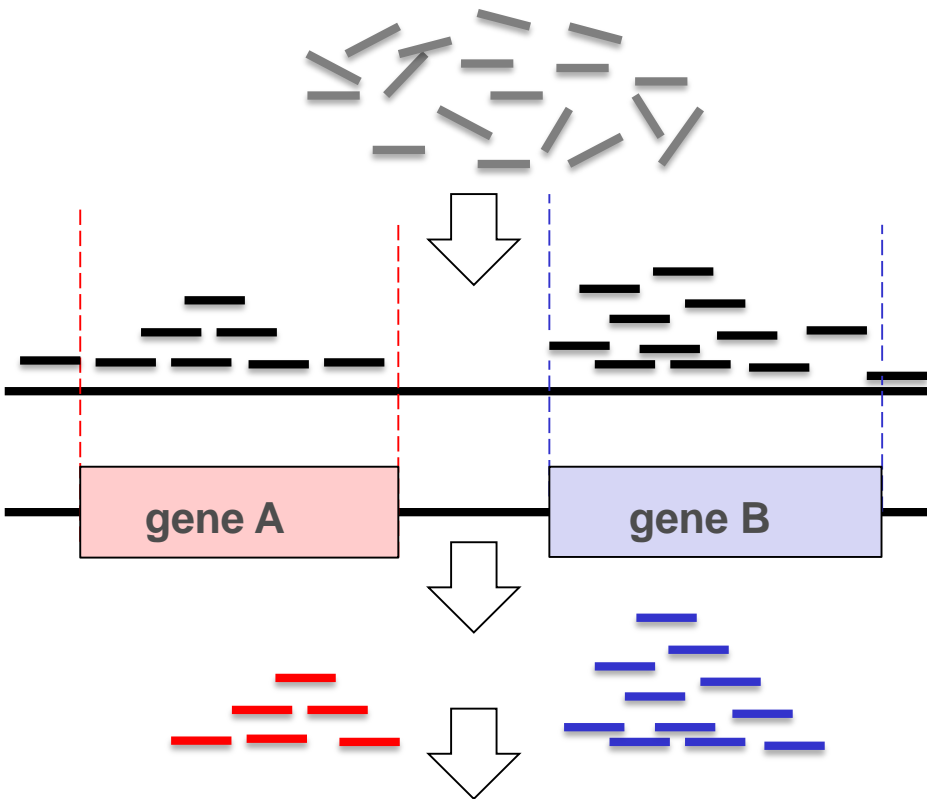
Count how many reads each gene has

	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	170	100	110
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1



Compare sample groups: differential expression analysis

DGE analysis: steps, tools and files



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	170	100	110
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

STEP	TOOL	FILE
Quality control	FastQC	FASTQ
Pre-processing	Trimmo-matic	FASTQ
Alignment	HISAT2	BAM
Quality control	RSeQC	
Quantitation	HTSeq	Read count file (TSV)
Combine count files to table	Define NGS experiment	Read count table (TSV)
Quality control	PCA, clustering	
Differential expression analysis	DESeq2, edgeR	Gene lists (TSV)

CSC

Data analysis workflow

- **Quality control of raw reads**
- **Preprocessing if needed**
- **Alignment to reference genome**
- **Alignment level quality control**
- **Quantitation**
- **Experiment level quality control**
- **Differential expression analysis**
- **Annotation**
- **Pathway analysis**

Data analysis workflow

- **Quality control of raw reads**
- Preprocessing if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

What and why?

➤ **Potential problems**

- low confidence bases, Ns
- sequence specific bias, GC bias
- adapters
- sequence contamination
- ...

Knowing about potential problems in your data allows you to

- **correct for them before you spend a lot of time on analysis**
- **take them into account when interpreting results**

Software packages for quality control

- **FastQC**
- **MultiQC**
- **FastX**
- **TagCleaner**
- ...

Raw reads: FASTQ file format

➤ Four lines per read:

@read name

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+ read name

!"*((((**+))%%%++)(%%%%).1***-+*)"**55CCF>>>>>CCCCCCC65

➤ http://en.wikipedia.org/wiki/FASTQ_format

➤ Attention: Do not unzip FASTQ files

- Chipster's analysis tools can cope with zipped files (.gz)



Base qualities

- If the quality of a base is **20**, the probability that it is wrong is **0.01**.
 - **Phred quality score** $Q = -10 * \log_{10}$ (probability that the base is wrong)

T C A G T A C T C G
40 40 40 40 40 40 40 40 37 35

- **”Sanger” encoding: numbers are shown as ASCII characters**
 - Note that older Illumina data uses different encoding

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy	ASCII coding in FASTQ file
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1,000	99.9%	?
40	1 in 10,000	99.99%	



How to check read quality?

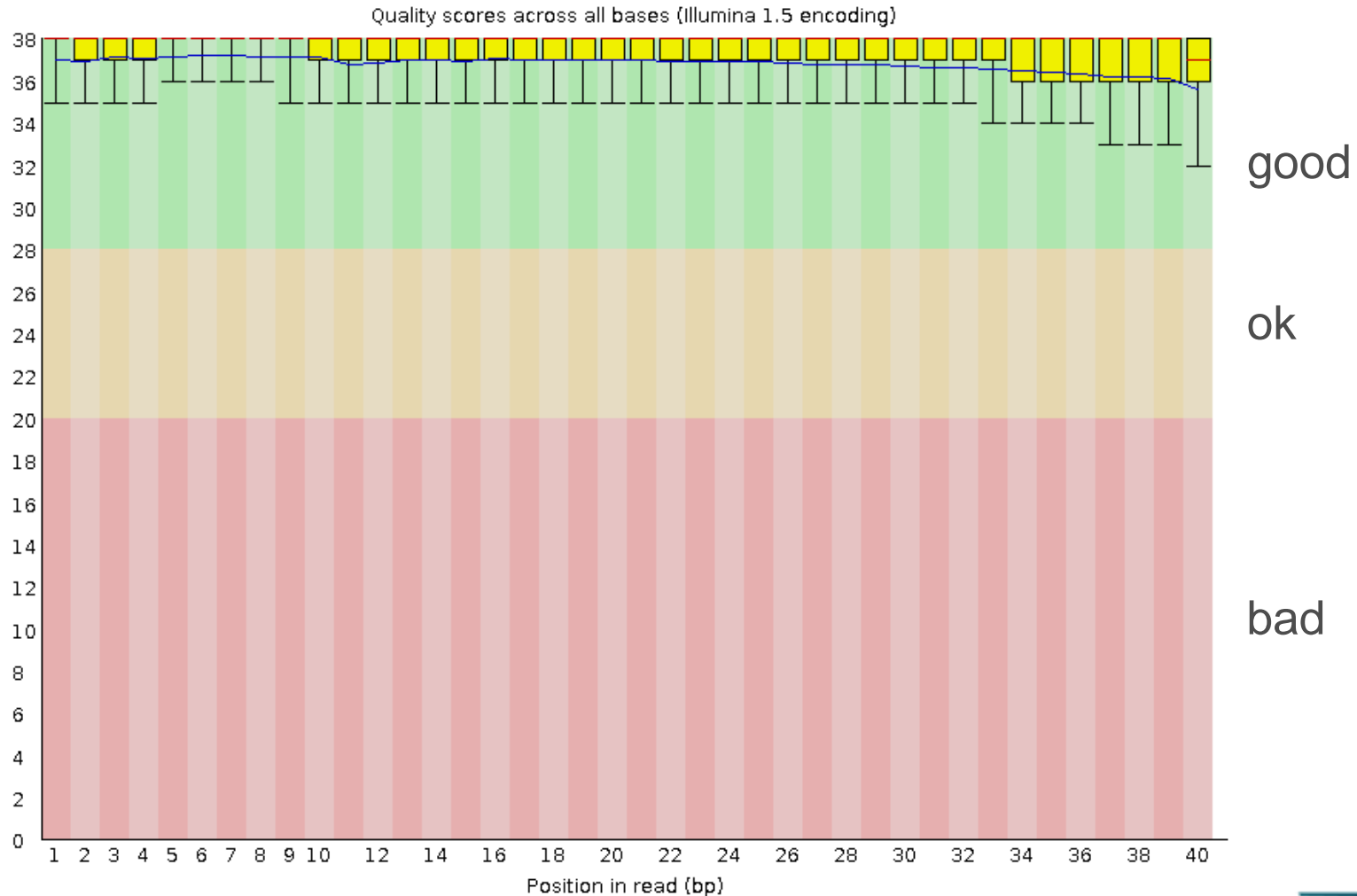
- **You can use FastQC either directly or via MultiQC**
 - If you have many samples, MultiQC is handier
- **Reports many things, including**
 - base quality
 - base composition
 - duplication
 - Ns
 - k-mers
 - adaptors

MultiQC

- **Can combine info from many tools.**
 - In Chipster it uses FastQC
- **Features**
 - Interactive plots
 - Traffic lights (they might not be suitable for your data!)
- **Toolbox (click on the right side panel), allows you to**
 - Highlight samples
 - Show only selected samples
 - Download plots
 - Rename samples
- **Good tutorial video**
 - https://www.youtube.com/watch?v=qPbIIO_KWN0

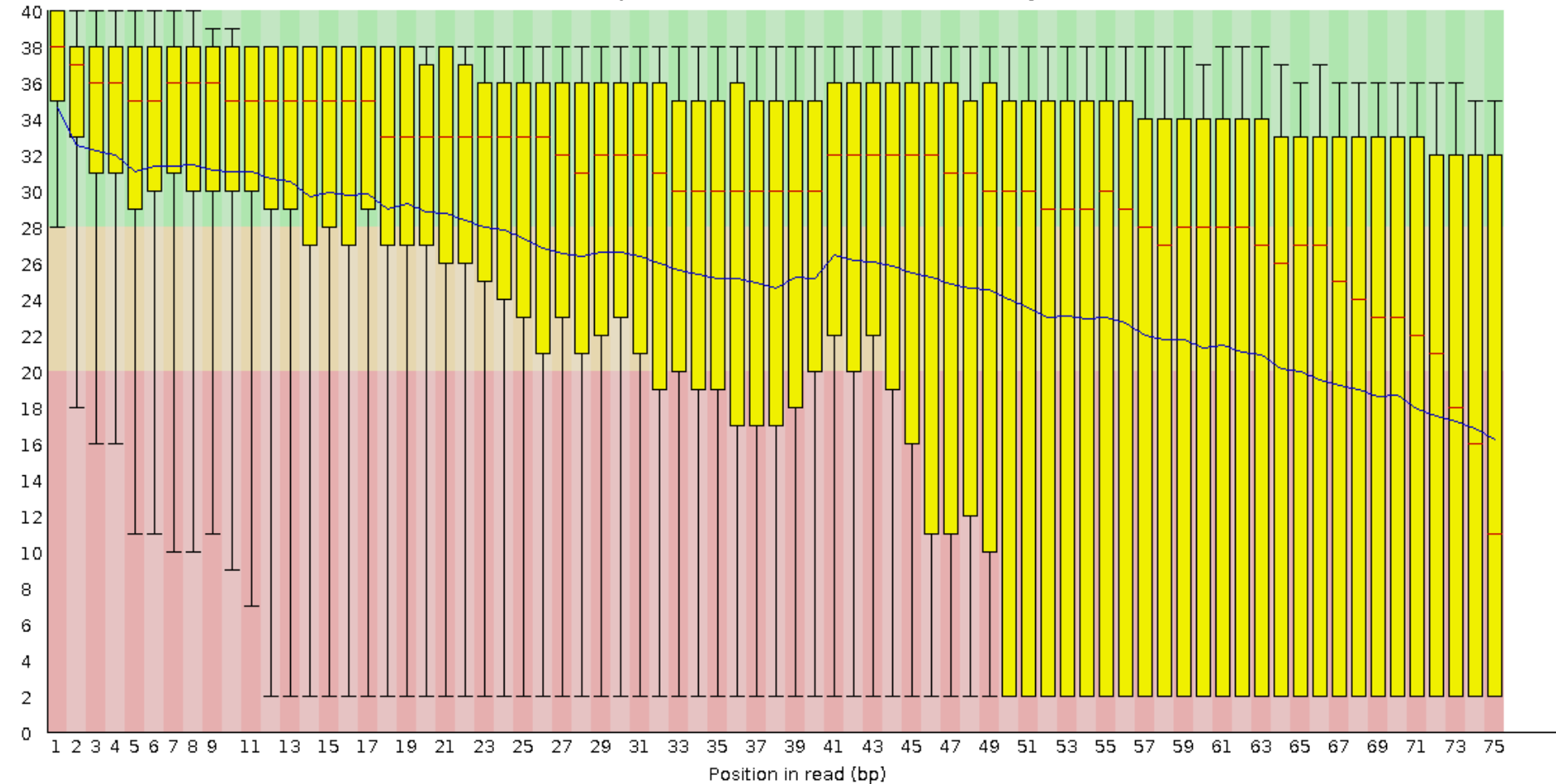


Per position base quality (FastQC)

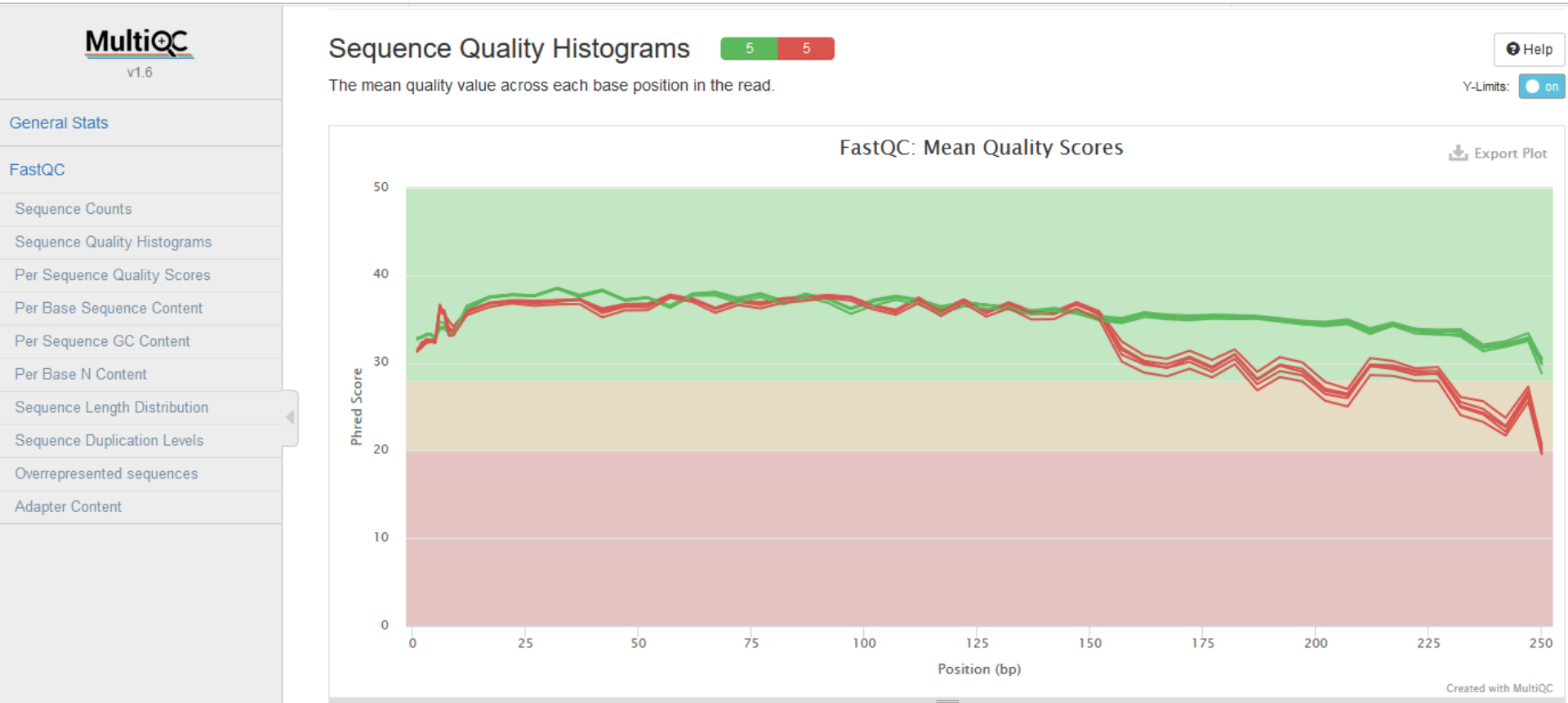


Per position base quality (FastQC)

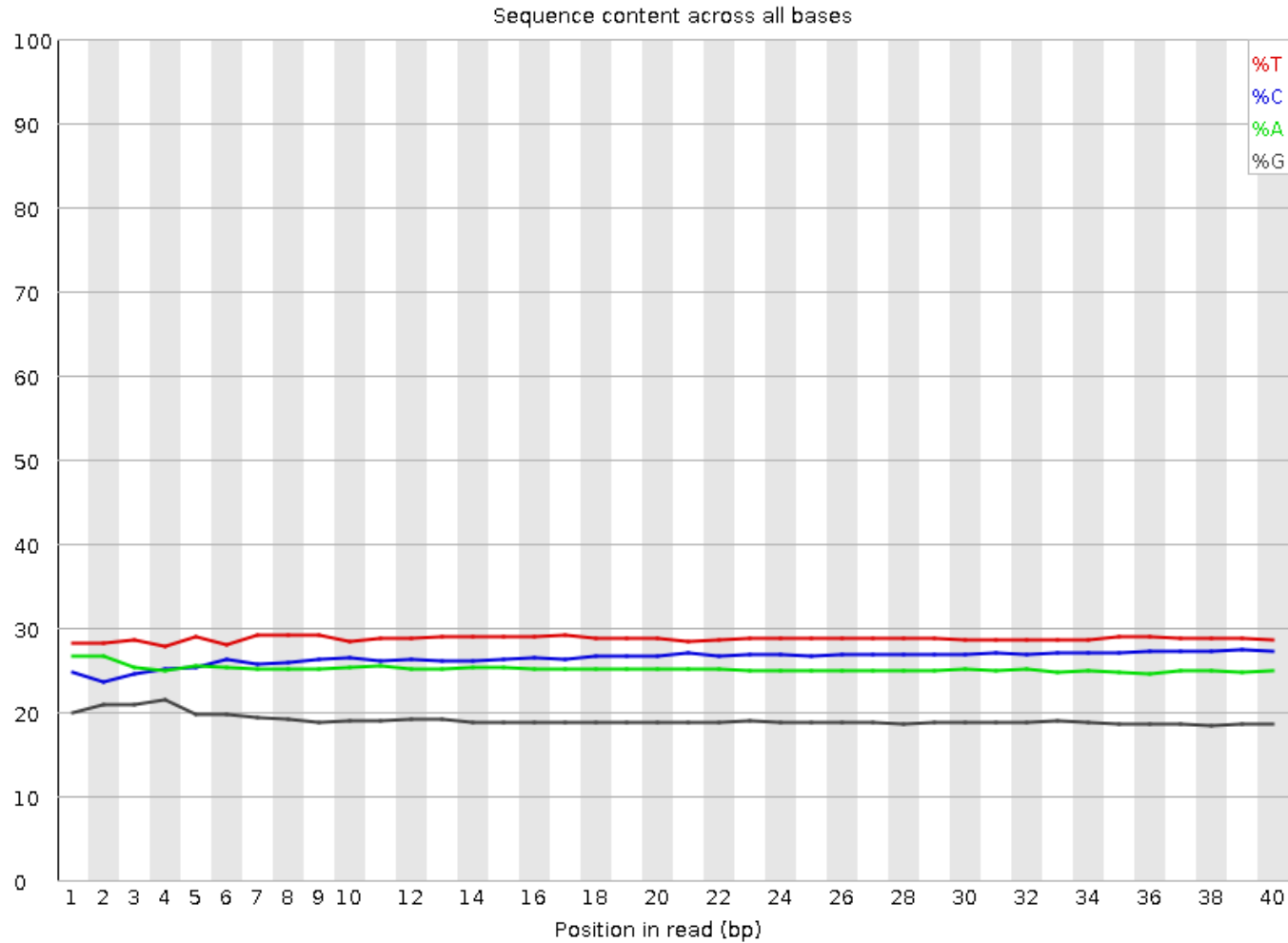
Quality scores across all bases (Illumina 1.5 encoding)



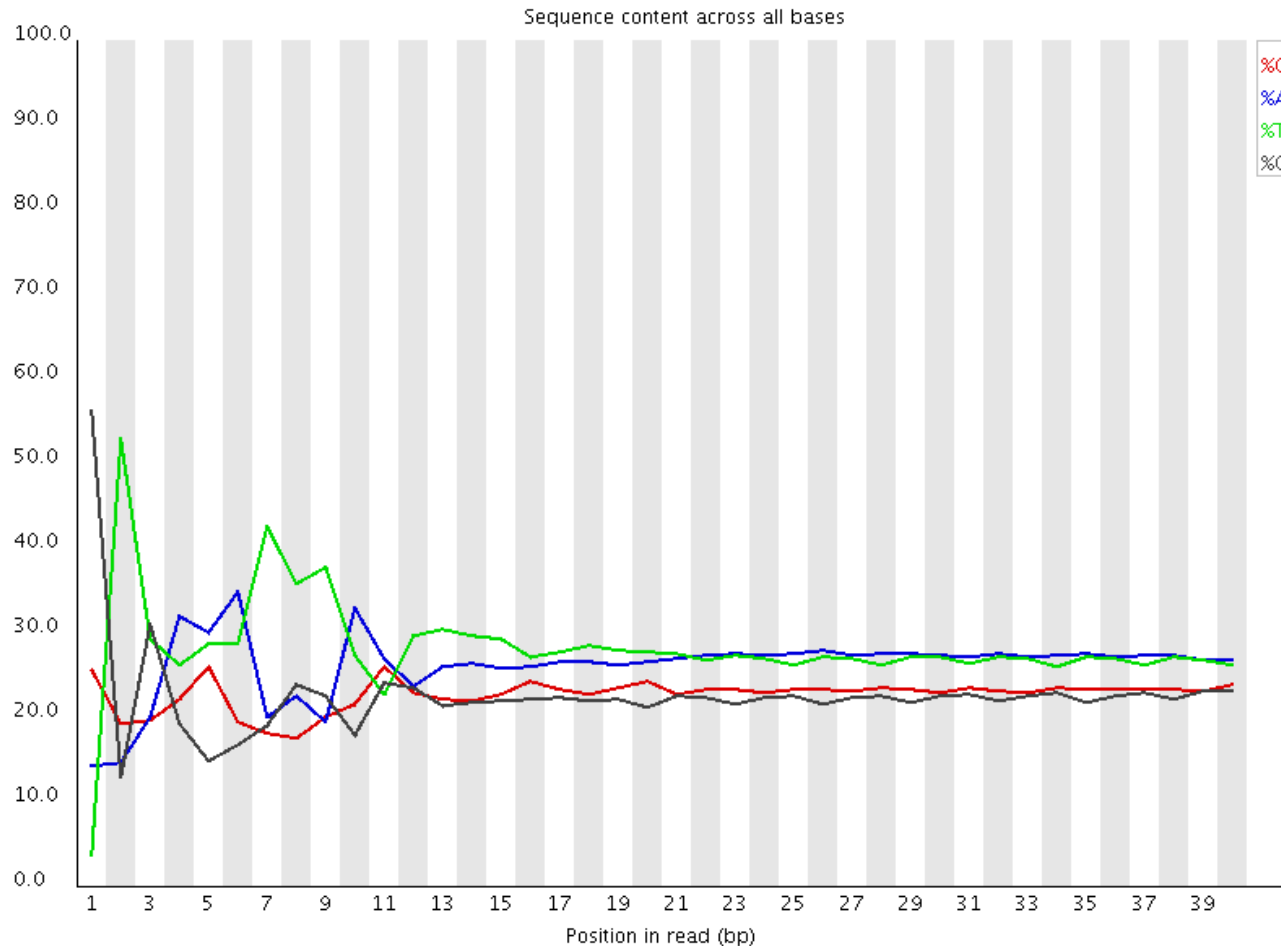
Per position base quality (MultiQC)



Per position sequence content (FastQC)



Per position sequence content (FastQC)

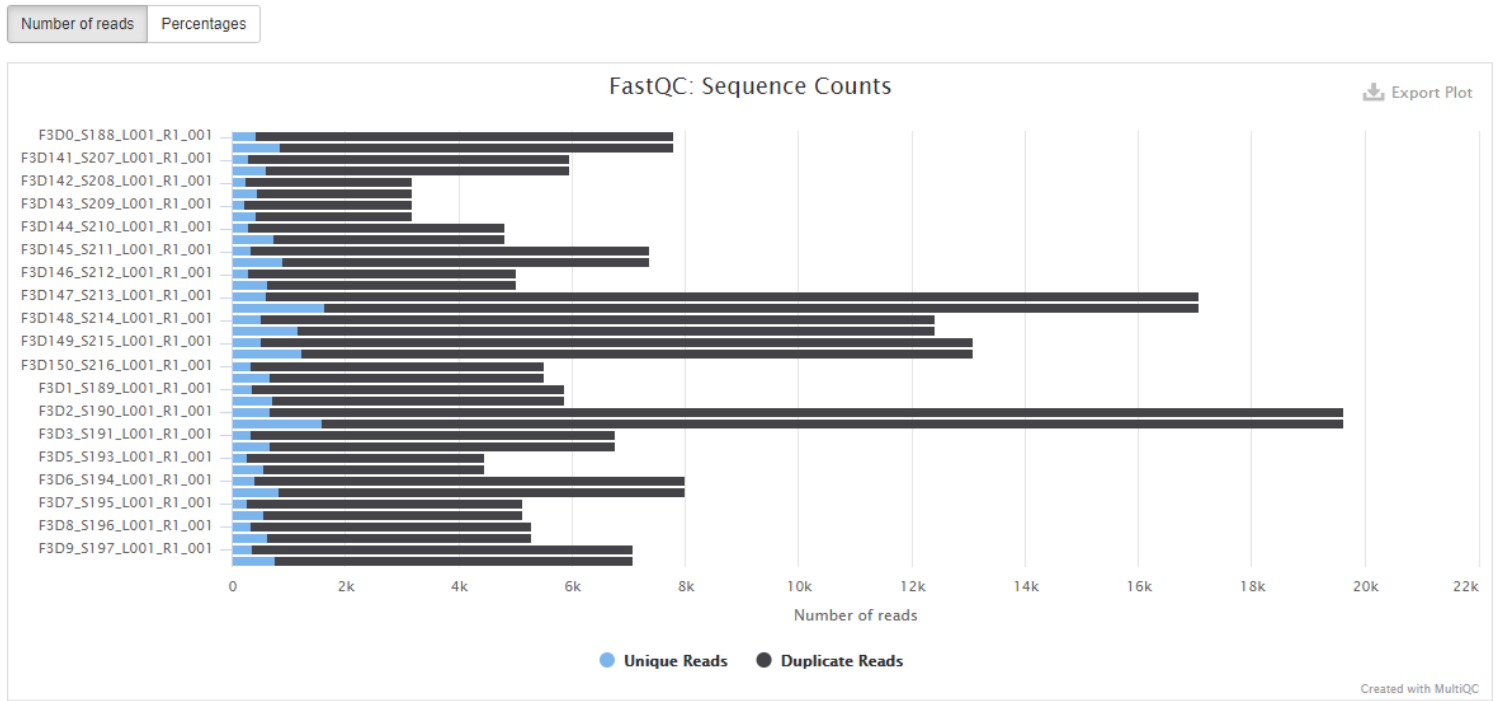


- **Enrichment of k-mers at the 5' end due to use of random hexamers or transposases in the library preparation**
- **Typical for RNA-seq data**
- **Can't be corrected, doesn't usually effect the analysis**

Sequence counts (MultiQC)

MultiQC
v1.7

- General Stats
- FastQC
- Sequence Counts
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content



Was your data made with stranded protocol?

➤ You need to indicate it when

- aligning reads to genome (e.g. HISAT2)
- counting reads per genes (e.g. HTSeq)

➤ If you don't know which stranded sequencing protocol was used, you can check it

- Select your FASTQ file and run the tool Quality control / RNA-seq strandedness inference with RSeQC
- Aligns a subset of the reads to genome and compares the locations to reference annotation

➤ For more info please see the manual

- <http://chipster.csc.fi/manual/library-type-summary.html>



RSeQC strandedness report

experiment_data.txt ...

Text [Details](#)

File size 468.0 bytes.

This is SingleEnd Data

Fraction of reads failed to determine: 0.0433

Fraction of reads explained by "++,--": 0.9498

Fraction of reads explained by "+-, -+": 0.0069

It seems the data is stranded. Read is always on the same strand as the gene.

Corresponding parameters are:

TopHat, Cufflinks and Cuffdiff: library-type fr-secondstrand

HISAT2: RNA-strandness: F

HTSeq: stranded -- yes

RSeQC: ++,--

Input files were assigned as follows:

Read 1 file: hESC.fastq



Data analysis workflow

- Quality control of raw reads
- **Preprocessing (trimming / filtering) if needed**
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

Filtering vs trimming

- **Filtering removes the entire read**
- **Trimming removes only the bad quality bases**
 - It can remove the entire read, if all bases are bad
- **Trimming makes reads shorter**
 - This might not be optimal for some applications
- **Paired end data: the matching order of the reads in the two files has to be preserved**
 - If a read is removed, its pair has to be removed as well

What base quality threshold should be used?

- No consensus
- Trade-off between having good quality reads and having enough sequence
- Start with gentle trimming and check with FastQC

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro¹, Simone Scalabrin², Michele Morgante¹, Federico M. Giorgi^{1,3*}

¹ Institute of Applied Genomics, Udine, Italy, ² IGA Technology Services, Udine, Italy, ³ Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, United States of America

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE

published: 31 January 2014
doi: 10.3389/fgene.2014.00013

On the optimal trimming of high-throughput mRNA sequence data

Matthew D. MacManes^{1,2*}

¹ Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

² Hubbard Center for Genome Studies, Durham, NH, USA

Software packages for preprocessing

- **Trimmomatic**
- **FastX**
- **TagCleaner**
- ...

Trimmomatic options in Chipster

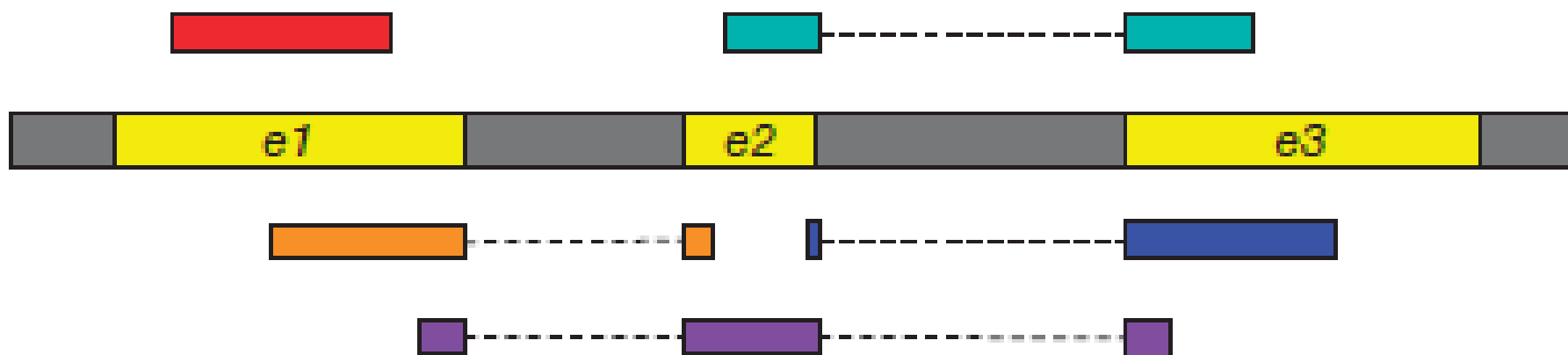
- **Adapters**
- **Minimum quality**
 - Per base, one base at a time or in a sliding window, from 3' or 5' end
 - Per base adaptive quality trimming (balance length and errors)
 - Minimum (mean) base quality
- **Trim x bases from left/ right**
- **Minimum read length after trimming**
- **Copes with paired end data**

Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- **Alignment to reference genome**
- Alignment level quality control
- Quantitation
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

Aligning reads to reference genome

- **The goal is to find the location where the read originated from**
- **Challenges**
 - Reads contain genomic variants and sequencing errors
 - Genomes contain non-unique sequence and introns
- **RNA-seq aligner needs to be able to align splice junction spanning reads to genome non-contiguously**
 - Spliced alignments are difficult because sequence signals at splice sites are limited, and introns can be thousands of bases long



Alignment programs

- **Many aligners have been developed over the years**
 - Convert genome fasta file to a data structure which is faster to search (e.g. BWT index or suffix array)
 - Differ in speed, memory requirements, accuracy and ability to deal with spliced alignments
- **Use splice-aware aligner for mapping RNA-seq reads, for example**
 - STAR (fast and accurate, needs a lot of memory)
 - HISAT2 (fast and accurate, creating the genomic index needs a LOT of memory)
 - TopHat2 (slower, needs less memory)



Splice-aware aligners in Chipster

- **STAR**
 - Human, mouse & rat genomes available
- **HISAT2**
 - Many genomes available
 - You can also supply own reference genome if it is small
- **TopHat2**
 - You can supply own reference genome
- **Output files**
 - BAM = contains the read alignments
 - bai = index file for BAM, required by genome browsers etc
 - log = useful information about the alignment run

HISAT2

- **HISAT = Hierarchical Indexing for Spliced Alignment of Transcripts**
- **Fast spliced aligner with low memory requirement**
- **Reference genome is (BWT FM) indexed for fast searching**
 - Currently Chipster offers human, mouse & rat reference genomes
 - Let us know if you need others!
 - You can provide own (small) reference genome in fasta format
- **Uses two types of indexes**
 - A global index: used to anchor a read in genome (28 bp is enough)
 - Thousands of small local indexes, each covering a genomic region of 56 Kbp: used for rapid extension of alignments (good for spliced reads with short anchors)
- **Uses splice site information found during the alignment of earlier reads in the same run**



HISAT2 parameters

Parameters

Genome Genome or transcriptome that you would like to align your reads against.	Homo_sapiens.GRCh38.95
RNA-strandness Specify strand-specific information. FR means read 1 is always on the same strand as the gene. RF means read 2 is always on the same strand as the gene. The default is unstranded.	unstranded
Base quality encoding used Quality encoding used in the fastq file.	Sanger - Phred+33
How many hits to report per read Instructs HISAT2 to report up to this many alignments to the reference for a given read.	5
Minimum intron length Sets minimum intron length. Default: 20	20
Maximum intron length Sets maximum intron length. Default: 500000	500000
Disallow soft-clipping Is soft-clipping used. By default HISAT2 may soft-clip reads near their 5' and 3' ends.	Use soft-clipping
Require long anchor lengths for subsequent assembly With this option, HISAT2 requires longer anchor lengths for de novo discovery of splice sites. This leads to fewer alignments with short-anchors, which helps transcript assemblers improve significantly in computation and memory usage.	Don't require
Input files	
Reads to align	G1Esubset_R1.fq.gz G1Esubset_R2.fq.gz
List of read 1 files	
List of read 2 files	

- If you have more than 2 FASTQ files per **one** sample, make file name lists for R1 and R2 files first
- Remember to set the strandedness correctly!
- Require long anchors (> 16 bp) if you are going to do transcript assembly
- Soft-clipping = read ends don't need to align, if this maximizes the alignment score

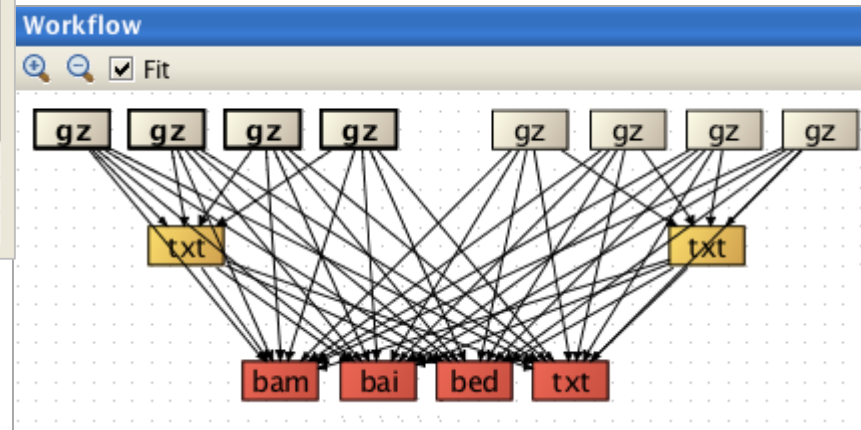


STAR

- **STAR = Spliced Transcripts Alignment to a Reference**
- **Reference genome fasta is converted to a suffix array for fast searching**
- **2-pass alignment process**
 - splice junctions found during the 1st pass are inserted into the genome index, and all reads are re-aligned in the 2nd mapping pass
 - this doesn't increase the number of detected novel junctions, but it allows more spliced reads aligning to novel junctions.
- **Maximum alignments per read -parameter sets the maximum number of loci the read is allowed to map to**
 - Alignments (all of them) will be output only if the read maps to no more loci than this. Otherwise no alignments will be output.
- **Chipster offers an Ensembl GTF file to detect annotated splice junctions**
 - you can also give your own, e.g. GENCODE GTF

If **one** sample has more than two FASTQ files

- **E.g. Illumina NextSeq can produce 8 FASTQ files for each sample**
- **Put all the FASTQ files for the same sample in the same alignment run (one job)**
- **Single end data: Select all the FASTQ files for the sample**
- **Paired end data: Make filename list files first**
 - Select all the read1 files and run the tool "Utilities / Make a list of file names"
 - Repeat with all the read2 files
 - Select all the FASTQ files and both filename list files and run HISAT2/STAR (check that the files have been assigned correctly)



File format for mapped reads: BAM/SAM

- BAM is a compact binary file containing aligned reads. SAM (Sequence Alignment/Map) contains the same information in tab-delimited text.
- You can view BAM as text by running the tool [Create a preview for BAM](#).

lymphnode4a_R1.prev.sam ***

Text Details

File size 66.9 kB.

Full Screen

```
@HD VN:1.0 SO:coordinate
@SQ SN:1 LN:248956422
@SQ SN:2 LN:242193529
@SQ SN:3 LN:198295559
@SQ SN:4 LN:190214555
@SQ SN:5 LN:181538259
@SQ SN:6 LN:170805979
@SQ SN:7 LN:159345973
@SQ SN:8 LN:145138636
@SQ SN:9 LN:138394717
@SQ SN:10 LN:133797422
@SQ SN:11 LN:135086622
@SQ SN:12 LN:133275309
@SQ SN:13 LN:114364328
@SQ SN:14 LN:107043718
@SQ SN:15 LN:101991189
@SQ SN:16 LN:90338345
@SQ SN:17 LN:83257441
@SQ SN:18 LN:80373285
@SQ SN:19 LN:58617616
@SQ SN:20 LN:64444167
@SQ SN:21 LN:46709983
@SQ SN:22 LN:50818468
@SQ SN:X LN:156040895
@SQ SN:Y LN:57227415
@SQ SN:MT LN:16569
@PG ID:hisat2 PN:hisat2 VN:2.1.0 CL:"/opt/chipster/tools/hisat2/hisat2-align-s --wrapper basic-0 --phred33 --min-intronlen 20 --max-intronlen 500000 -x Homo_sapiens.GRCh38.
ERR315371.8524369 409 1 12808 1 101M = 12808 0 GATGCCCTCCACACCTCTTGATCTTCCCTGTGATGTCATCTGGAGCCCTGCTGCTTGCCTGCGGTGGCCTATAAAGCCTCCTGGTCTGGCTCCAAGGCCTGGC BB7
ERR315371.5736647 163 1 14453 1 101M = 14519 167 CTTAAGAACACTGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCAGCTGTGTGGCCTCAAGCCAGCCTTCC BBB
ERR315371.5736647 419 1 14453 1 101M = 185040 150229 CTTAAGAACACTGTGGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCAGACAGAAGTCCCCGCCAGCTGTGTGGCCTCAAGCCAGCCTTCC BBB
ERR315371.5736647 83 1 14519 1 101M = 14453 -167 CCCGCCAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGTGTTCCGTCAACCCCTCCCAAGGAAGTAGGCTGAGCAGCTTGTCTGGCTGTGCCATGTCAGAGCAACG FBB
ERR315371.8571896 355 1 14554 1 101M = 185209 150297 GCTCCTTGAAGCTGGTCTCCACACAGTGTGTTCCGTCAACCCCTCCCAAGGAAGTAGGCTGAGCAGCTTGTCTGGCTGTGCCATGTCAGAGCAACG BBB
ERR315371.2401177 99 1 14701 0 87M14S = 185222 150149 ATCCAGTCGTCCTCGTCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCAGATCGGAAGAGCA BBB
ERR315371.2401177 355 1 14701 0 87M14S = 14701 -115 ATCCAGTCGTCCTCGTCTCCTCTGCCTGTGGCTGCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCAGATCGGAAGAGCA BBB
```

← BAM header

alignment information: one line per read alignment, containing 11 mandatory fields, followed by optional tags

Fields in BAM/SAM files

- **read name** HWI-EAS229_1:2:40:1280:283
- **flag** 272
- **reference name** 1
- **position** 18506
- **mapping quality** 0
- **CIGAR** 49M6183N26M
- **mate name** *
- **mate position** 0
- **insert size** 0
- **sequence**
AGGGCCGATCTTGGTGCCATCCAGGGGGCCTCTACAAGGAT
AATCTGACCTGCTGAAGATGTCTCCAGAGACCTT
- **base qualities**
ECC@EEF@EB:EECFEECCCBEEEE;>5;2FBB@FBFEEFCF@F
FFFCEFFFFEE>FFEFC=@A;@>1@6.+5/5
- **tags** MD:Z:75 NH:i:7 AS:i:-8 XS:A:-



Mapping quality

- **Confidence in read's point of origin**
- **Depends on many things, including**
 - uniqueness of the aligned region in the genome
 - length of alignment
 - number of mismatches and gaps
- **Expressed in Phred scores, like base qualities**
 - $Q = -10 * \log_{10}$ (probability that mapping location is wrong)
- **Values differ in different aligners. E. g. unique mapping is**
 - 60 in HISAT2
 - 255 in STAR
 - 50 in TopHat
 - <https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/>



CIGAR string

- M = match or mismatch
- I = insertion
- D = deletion
- N = intron (in RNA-seq read alignments)
- S = soft clip (ignore these bases)
- H = hard clip (ignore and remove these bases)

- Example:

@HD VN:1.3 SO:coordinate

@SQ SN:ref LN:45

r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *

- The corresponding alignment

```
Ref  AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001          TTAGATAAAGGATA*CTG
```



Flag field in BAM

➤ Read's flag number is a sum of values

- E.g. 4 = unmapped, 1024 = duplicate
- Explained in detail at <http://samtools.github.io/hts-specs/SAMv1.pdf>
- You can interpret them at <http://broadinstitute.github.io/picard/explain-flags.html>

This utility explains SAM flags in plain English.
It also allows switching easily from a read to its mate.

Flag:

Explanation:

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment



How did the alignment go? Check the log file

- **How many reads aligned to the reference genome?**
 - How many of them aligned uniquely?
- **How many read pairs aligned to the reference genome?**
 - How many pairs aligned concordantly?
- **What was the overall alignment rate?**

```
Visualisation
View text
25354832 reads; of these:
  25354832 (100.00%) were paired; of these:
    6098272 (24.05%) aligned concordantly 0 times
    18567284 (73.23%) aligned concordantly exactly 1 time
    689276 (2.72%) aligned concordantly >1 times
----
    6098272 pairs aligned concordantly 0 times; of these:
      724806 (11.89%) aligned discordantly 1 time
----
    5373466 pairs aligned 0 times concordantly or discordantly; of these:
      10746932 mates make up the pairs; of these:
        8812069 (82.00%) aligned 0 times
        1800817 (16.76%) aligned exactly 1 time
        134046 (1.25%) aligned >1 times
82.62% overall alignment rate
```



Log file by STAR

Visualisation	
View text	
Started job on	Feb 17 12:38:11
Started mapping on	Feb 17 12:47:47
Finished on	Feb 17 12:52:32
Mapping speed, Million of reads per hour	320.27
Number of input reads	25354832
Average input read length	202
UNIQUE READS:	
Uniquely mapped reads number	20409554
Uniquely mapped reads %	80.50%
Average mapped length	197.39
Number of splices: Total	12378576
Number of splices: Annotated (sjdb)	12378175
Number of splices: GT/AG	12272618
Number of splices: GC/AG	89423
Number of splices: AT/AC	9589
Number of splices: Non-canonical	6946
Mismatch rate per base, %	0.39%
Deletion rate per base	0.01%
Deletion average length	1.75
Insertion rate per base	0.01%
Insertion average length	1.36
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	970016
% of reads mapped to multiple loci	3.83%
Number of reads mapped to too many loci	11610
% of reads mapped to too many loci	0.05%
UNMAPPED READS:	
% of reads unmapped: too many mismatches	0.00%
% of reads unmapped: too short	15.55%
% of reads unmapped: other	0.08%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%



Other tools for checking BAM files

➤ Count alignments in BAM

- How many alignments does the BAM contain.
- Includes an optional mapping quality filter.

➤ Count alignments per chromosome in BAM

➤ Count alignment statistics for BAM

➤ Collect multiple metrics for BAM

alignment-statistics.txt ***

Text [Details](#)

File size 389.0 bytes.

```
372896 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
372896 + 0 mapped (100.00%:-nan%)
372896 + 0 paired in sequencing
188651 + 0 read1
184245 + 0 read2
335364 + 0 properly paired (89.93%:-nan%)
359009 + 0 with itself and mate mapped
13887 + 0 singletons (3.72%:-nan%)
3129 + 0 with mate mapped to a different chr
1720 + 0 with mate mapped to a different chr (mapQ>=5)
```


Tools for manipulating BAM files

➤ **Make a subset of BAM**

- Retrieve alignments for a given chromosome/region, e.g. chr1:100-1000
- Can filter based on mapping quality

➤ **Index BAM**

➤ **Convert SAM to BAM, sort and index BAM**

- "Preprocessing" when importing SAM/BAM, runs on your computer.
- The tool available in the "Utilities" category runs on the server

➤ **Create a preview for BAM**

- Creates a SAM file containing the BAM header and the first 200 alignments. SAM is a text file so you can view it in Chipster.

Full alignment or lightweight mapping?

- **Aligning reads to reference genome is slow → many quantitation tools offer now lightweight "mapping". Different flavors:**
 - selective alignment (Salmon)
 - quasi-mapping (Sailfish, Salmon)
 - pseudoalignment (kallisto)
- **These tools match reads to transcripts and report transcripts that a read is compatible with (no base-to-base alignments)**
 - Difficult to assign reads to isoforms because they share exons, and technical biases cause non-uniform coverage
 - Need complete transcriptome
- **Srivastava et al 2019: Alignment and mapping methodology influence transcript abundance estimation**
 - Quantification accuracy is better when using traditional alignments



Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- **Alignment level quality control**
- Quantitation
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

Annotation-based quality metrics

➤ **Saturation of sequencing depth**

- Would more sequencing detect more genes and splice junctions?

➤ **Read distribution between different genomic features**

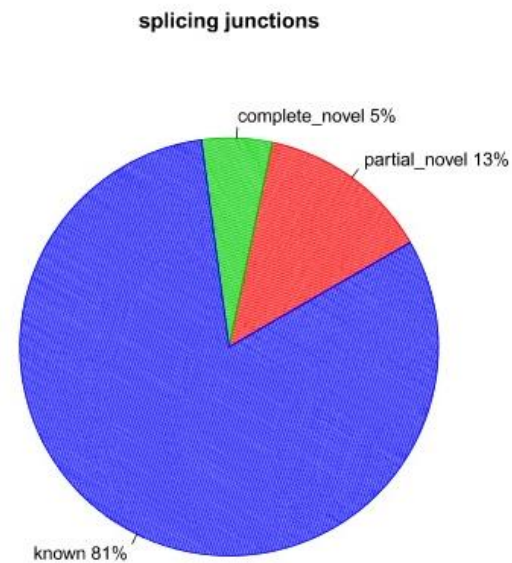
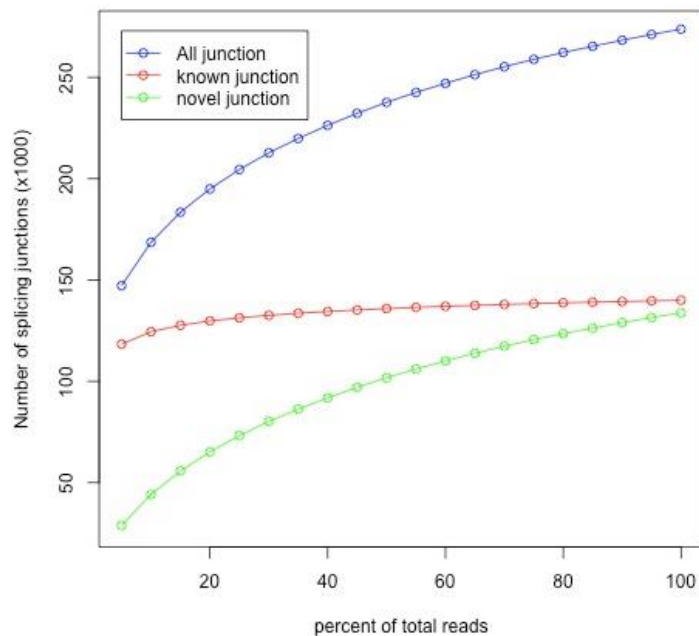
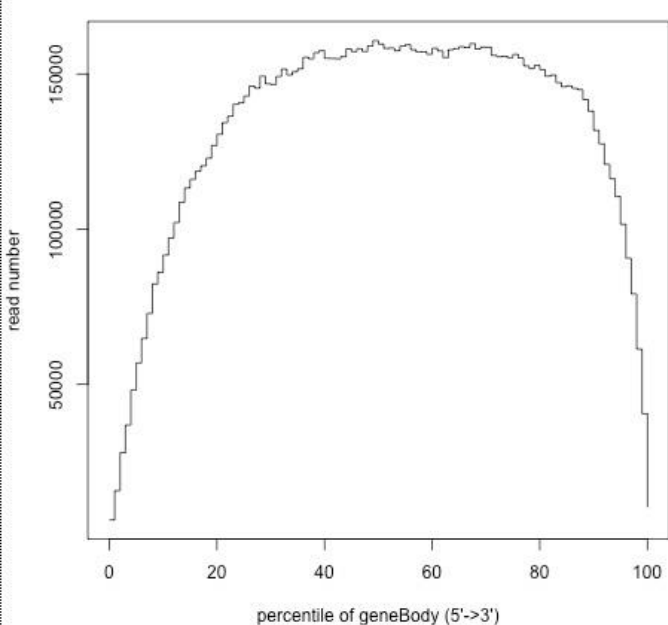
- Exonic, intronic, intergenic regions
- Coding, 3' and 5' UTR exons
- Protein coding genes, pseudogenes, rRNA, miRNA, etc

➤ **Is read coverage uniform along transcripts?**

- Biases introduced in library construction and sequencing
 - polyA capture and polyT priming can cause 3' bias
 - random primers can cause sequence-specific bias
 - GC-rich and GC-poor regions can be under-sampled
- Genomic regions have different mappabilities (uniqueness)

Quality assessment with RSeQC

- Checks coverage uniformity, saturation of sequencing depth, novelty of splice junctions, read distribution between different genomic regions, etc.
- Takes a BAM file and a BED file
 - Chipster has BED files available for several organisms
 - You can also use your own BED if you prefer



BED file format

- **BED (Browser extensible data) file format is used for reporting location of features (e.g. genes and exons) in a genome**
- **5 obligatory columns: chr, start, end, name, score**
- **0-based, like BAM**

column0	column1	column2	column3	column4
chr22	21022480	21024796	JUNC00000001	1
chr19	201609	201783	JUNC00000002	5
chr19	281478	282180	JUNC00000003	3
chr19	282242	282811	JUNC00000004	21
chr19	282751	287541	JUNC00000005	37
chr19	287705	288084	JUNC00000006	6
chr19	288105	291354	JUNC00000007	18
chr19	307484	308600	JUNC00000008	1
chr19	308603	308858	JUNC00000009	2
chr19	308868	311907	JUNC00000010	13
chr19	311872	312256	JUNC00000011	26
chr19	312205	313558	JUNC00000012	22
chr19	313575	325706	JUNC00000013	68

Own BED? Check chromosome names

- **RSeQC needs the same chromosome naming in BAM and BED**
- **Chromosome names in BED files can have the prefix “chr”**
 - e.g. chr1
- **Chipster BAM files are Ensembl-based and don't have the prefix**
 - If you use your own BED (e.g. from UCSC Table browser) you need to remove the prefix (chr1 → 1)
- **Use the tool **Utilities / Modify text** with the following parameters:**
 - Operation = Replace text
 - Search string = chr
 - Input file format = BED

QC tables by RSeQC

```

=====
#All numbers are READ count (alignment, actually...)
=====
Total records:                103284

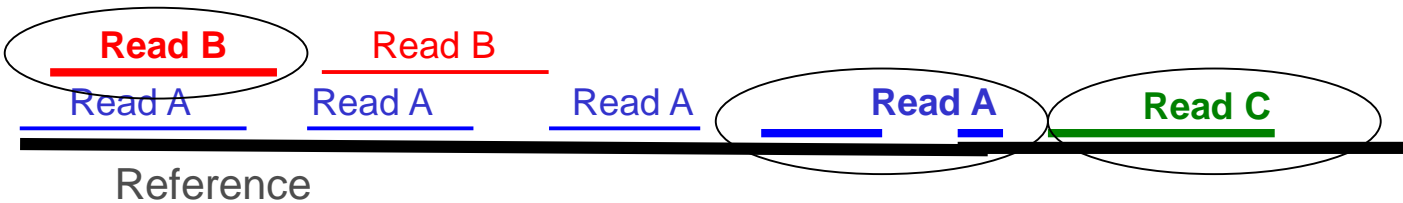
QC failed:                    0
Optical/PCR duplicate:        0
Non primary hits              18476
Unmapped reads:               0
mapq < mapq_cut (non-unique): 4208
                               Default=30
mapq >= mapq_cut (unique):    80600
Read-1:                       0
Read-2:                       0
Reads map to '+':             48292
Reads map to '-':             32308
Non-splice reads:             50919
Splice reads:                 29681
Reads mapped in proper pairs: 0
Proper-paired reads map to different chrom:0
    
```

```

read_distribution:
Total Reads                84808
Total Tags                 116738
Total Assigned Tags          111352
=====
Group           Total_bases   Tag_count   Tags/Kb
CDS_Exons       2211343    90961      41.13
5'UTR_Exons     529860    1662       3.14
3'UTR_Exons     1415234   12423      8.78
Introns         25801210  5349       0.21
TSS_up_1kb     1295771   31         0.02
TSS_up_5kb     5332522   321        0.06
TSS_up_10kb    8804879   584        0.07
TES_down_1kb   1292506   217        0.17
TES_down_5kb   5108821   344        0.07
TES_down_10kb  8282641   373        0.05
=====
    
```

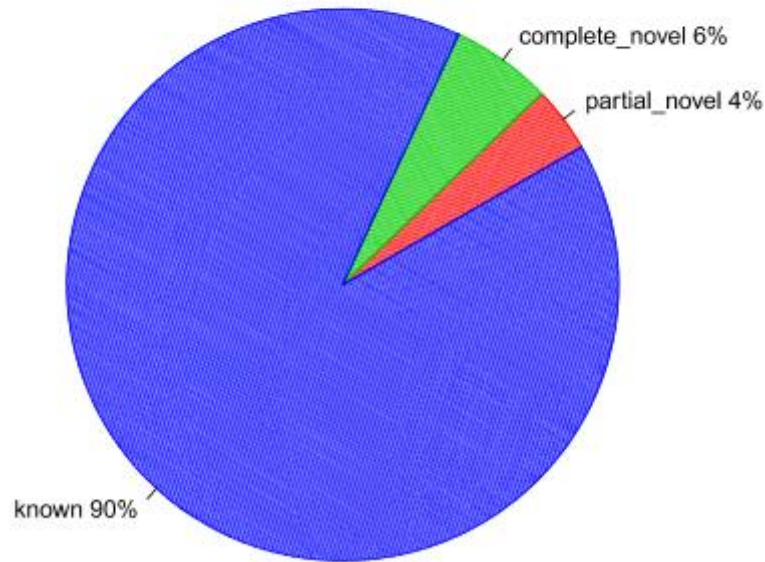
```

Total records:      7
Non primary hits:  4
Total reads:        3
Total tags:         8
    
```

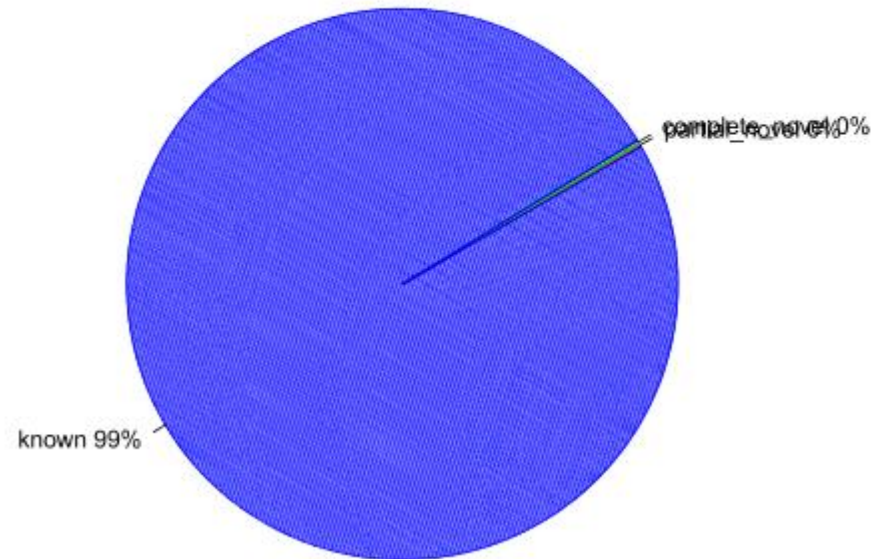


Splicing graphs by RSeQC

splicing junctions



splicing events



- **Splicing junction = exon-exon junction covered by one or more reads**
- **Splicing event = a read is split across a splice junction**

Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- **Quantitation**
- Describing the experiment with phenodata
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

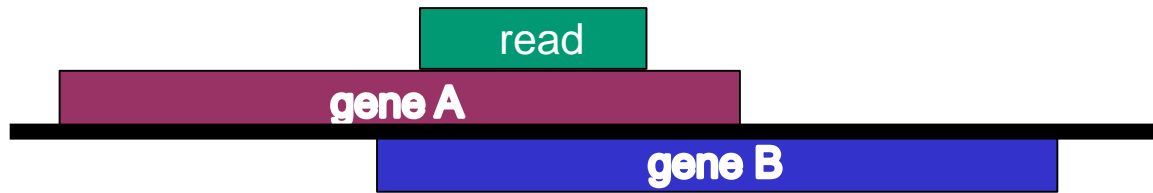
Software for counting reads per genes or transcripts

- **HTSeq**
- **StringTie**
- **Cufflinks**
- **Salmon**
- **Kallisto**

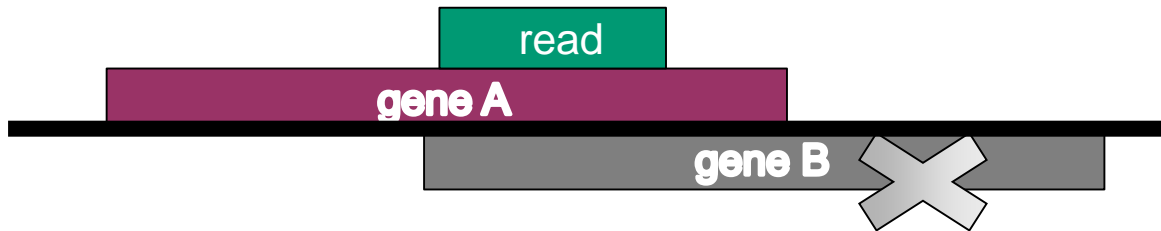
Counting reads per genes with HTSeq

- **Given a BAM file and a GTF file with gene locations, counts how many reads map to each gene.**
 - A gene is considered as the union of all its exons.
 - Reads can be counted also per exons.
- **Chipster provides Ensembl GTF files, but you can give your own**
 - Note that GTF and BAM must use the same chromosome naming
 - All exons of a gene must have the same gene_id (avoid UCSC GTFs)
- **Multimapping reads and ambiguous reads are not counted**
- **3 modes to handle reads which overlap several genes**
 - Union (default), Intersection-strict, Intersection-nonempty
- **Attention: was your data made with stranded protocol?**
 - You need to select the right counting mode!

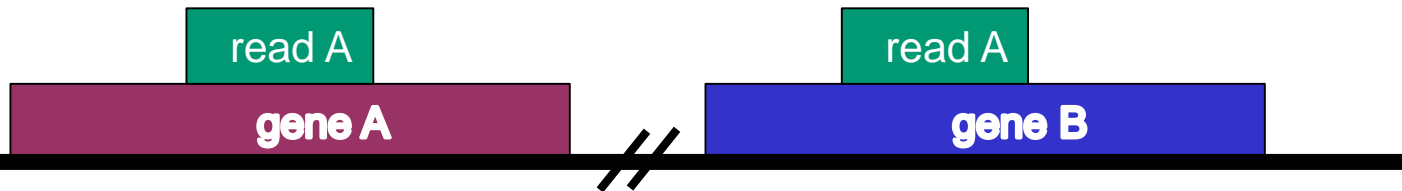
Not unique or ambiguous?



Ambiguous



Stranded data
→ Not ambiguous



Multimapping
(not unique)



HTSeq count modes

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous



HTSeq result files: summary info and counts

htseq-count-info.txt ...

Text [Details](#)

File size 150.0 bytes.

```
__no_feature      149620
__ambiguous       28344
__too_low_aQual   0
__not_aligned     0
__alignment_not_unique 55693

not_counted      233657
counted          401694
total            635351
```

hESC2_chr18.tsv

Showing all 63677 rows.

id	chr	start	end	len	strand	count
ENSG00000152234	chr18	43664109	43684300	20191	-	28474
ENSG00000235552	chr18	6462142	6463014	872	-	15356
ENSG00000074657	chr18	56529831	56653712	123881	+	9285
ENSG00000133313	chr18	72163050	72188366	25316	+	8325
ENSG00000046604	chr18	29078005	29128971	50966	+	7373
ENSG00000134440	chr18	55267887	55289445	21558	-	7196
ENSG00000175886	chr18	36914835	36915639	804	-	6102
ENSG00000235297	chr18	72057118	72057532	414	-	6009
ENSG00000101680	chr18	6941742	7117813	176071	-	5647
ENSG00000177426	chr18	3411605	3458409	46804	+	5024
ENSG00000265273	chr18	29542140	29543581	1441	+	4900
ENSG00000118680	chr18	3261906	3278282	16376	+	4603
ENSG00000176014	chr18	12307667	12344319	36652	+	4538
ENSG00000134759	chr18	33709406	33757909	48503	+	4202
ENSG00000134779	chr18	34359986	34409158	49172	-	4067
ENSG00000081913	chr18	60382671	60647666	264995	+	3997
ENSG00000101608	chr18	3247478	3256234	8756	+	3962
ENSG00000141401	chr18	11981023	12030876	49853	+	3816
ENSG00000101544	chr18	77866914	77905406	38492	+	3687
ENSG00000167088	chr18	19192227	19210417	18190	+	3392
ENSG00000167315	chr18	47309868	47340330	30462	-	3290
ENSG00000141425	chr18	33564349	33647539	83190	-	3165

GTF file format

- **9 obligatory columns: chr, source, name, start, end, score, strand, frame, attribute**
- **1-based**
- **For HTSeq to work, all exons of a gene must have the same gene_id**
 - Use GTFs from Ensembl, avoid UCSC

chr1	unknown	exon	14362	14829	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	14970	15038	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	15796	15947	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	16607	16765	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	16858	17055	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	17233	17368	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	17606	17742	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	17915	18061	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	18268	18366	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	24738	24891	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";
chr1	unknown	exon	29321	29370	.	-	.	gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "NR_024540"; tss_id "TSS7245";



Isoform switching can confound DGE analysis

- **The number of reads obtained from an expressed gene depends on the transcript length**
 - Longer transcripts produce more fragments and hence more reads
- **If a gene switches from a long transcript isoform to a short one, this can confound DGE analysis**



Control sample



Cancer sample



Expression level of gene A is the same in both samples, but cancer cells express the shorter isoform

Isoform switching can confound DGE analysis

- **The number of reads obtained from an expressed gene depends on the transcript length**
 - Longer transcripts produce more fragments and hence more reads
- **If a gene switches from one transcript isoform to another one, this can confound DGE analysis**



Control sample



Cancer sample



We get twice as many reads from the control sample
→ is gene A down-regulated in cancer?

Is isoform switching a major problem?

- **The magnitude of the effect depends on**
 - the extent of differential transcript usage (DTU)
 - the difference in length between the differentially expressed isoforms.
 - If the longer isoform is < 34% longer, false positives are controlled ok
 - Among all human transcript pairs in which both transcripts belong to the same gene, the median length ratio is 1.85
 - For one third of such pairs the longer isoform is < 38% longer
- **Many human genes express mainly one, dominant isoform**
 - → the global impact of isoform switching is relatively small in many real datasets (as opposed to simulated ones)

Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- **Describing the experiment with phenodata**
- Experiment level quality control
- Differential expression analysis
- Annotation
- Pathway analysis

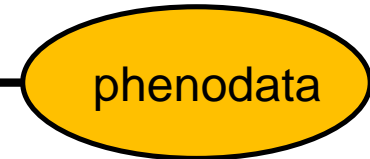
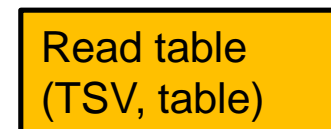
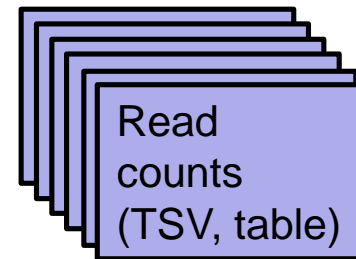
Combine individual count files into a count table

- Select all the count files and run “Utilities / Define NGS experiment”
- This creates a table of counts and a phenodata file, where you can describe experimental groups

						Control 1
Gene	Gene	Gene	Gene	Gene A	Gene A	6
Gene	Gene	Gene	Gene	Gene B	Gene B	11
Gene	Gene	Gene	Gene	Gene C	Gene C	200
Gene	Gene	Gene	Gene	Gene D	Gene D	0



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1



Phenodata file: describe the experiment

- **Describe experimental groups, time, pairing etc with numbers**
 - e.g. 1 = control, 2 = cancer
- **Define sample names for visualizations in the Description column**



sample	original_name	description	patient	group	treatment	time	hours
ngs001.tsv	SRR479052	1_C_24	1	1	Control	1	24h
ngs002.tsv	SRR479053	1_C_48	1	1	Control	2	48h
ngs003.tsv	SRR479054	1_DP_24	1	2	DPN	1	24h
ngs004.tsv	SRR479055	1_DP_48	1	2	DPN	2	48h
ngs007.tsv	SRR479058	2_C_24	2	1	Control	1	24h
ngs008.tsv	SRR479059	2_C_48	2	1	Control	2	48h
ngs009.tsv	SRR479060	2_DP_24	2	2	DPN	1	24h
ngs011.tsv	SRR479062	2_DP_48	2	2	DPN	2	48h
ngs015.tsv	SRR479066	3_C_24	3	1	Control	1	24h
ngs016.tsv	SRR479067	3_C_48	3	1	Control	2	48h
ngs017.tsv	SRR479068	3_DP_24	3	2	DPN	1	24h
ngs018.tsv	SRR479069	3_DP_48	3	2	DPN	2	48h

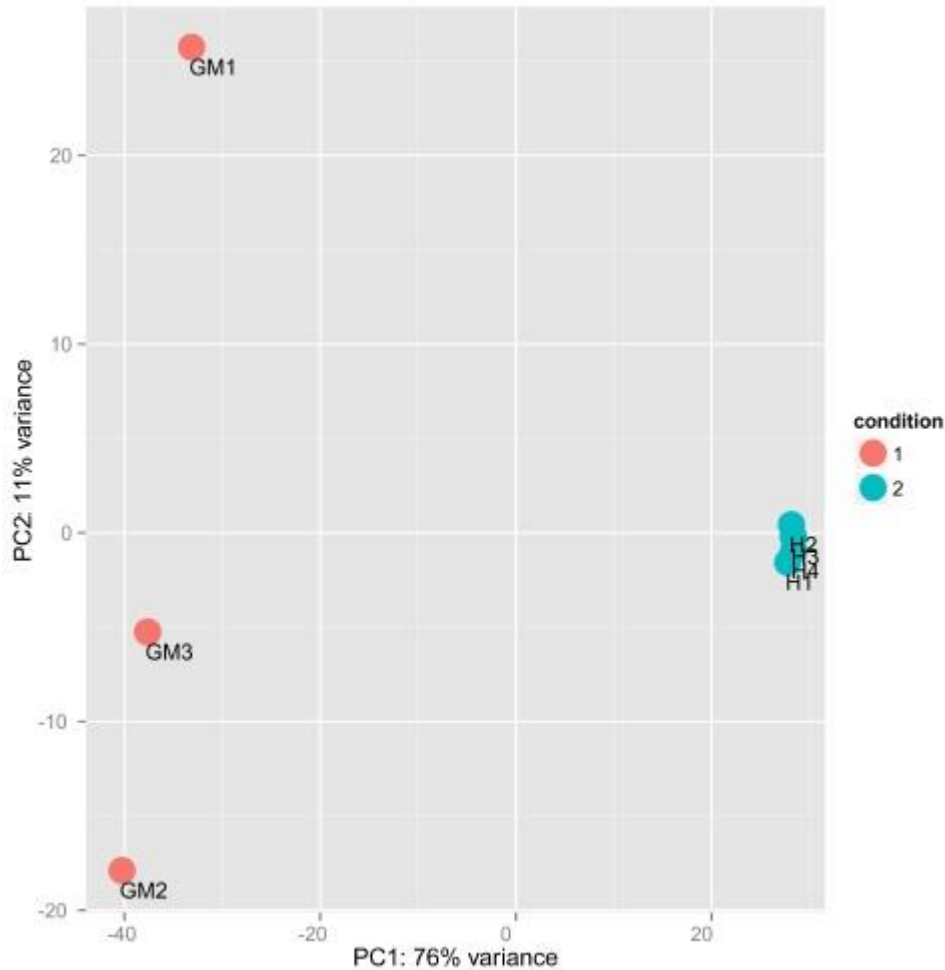
Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- **Experiment level quality control**
- Differential expression analysis
- Annotation
- Pathway analysis

Experiment level quality control

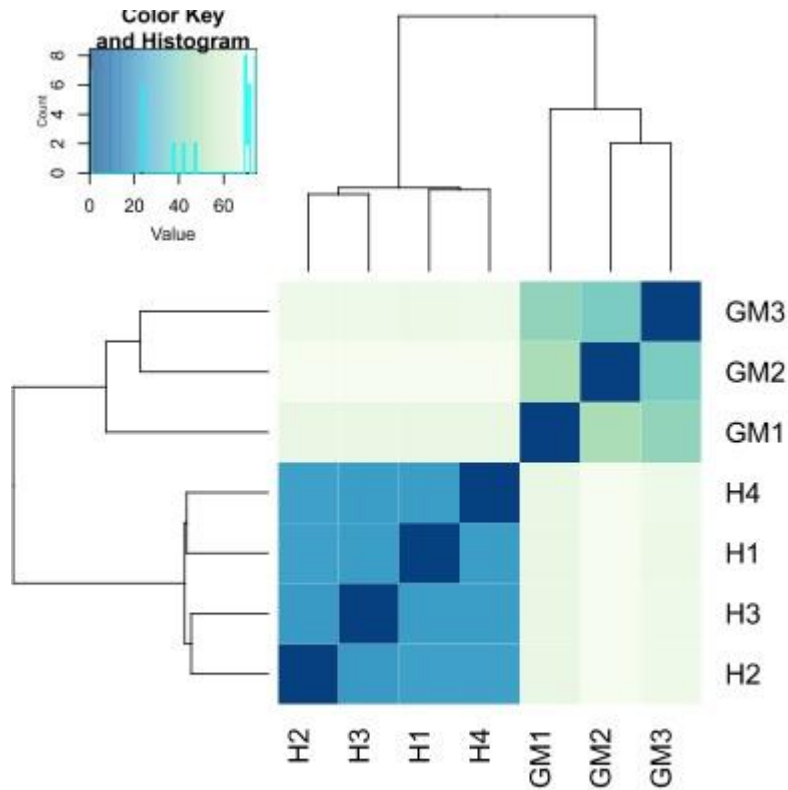
- **Getting an overview of similarities and dissimilarities between samples allows you to check**
 - Do the experimental groups separate from each other?
 - Is there a confounding factor (e.g. batch effect) that should be taken into account in the statistical analysis?
 - Are there sample outliers that should be removed?
- **Several methods available**
 - MDS (multidimensional scaling)
 - PCA (principal component analysis)
 - Clustering

PCA plot by DESeq2



- **The first two principal components, calculated after variance stabilizing transformation**
- **Indicates the proportion of variance explained by each component**
 - If PC2 explains only a small percentage of variance, it can be ignored

Sample heatmap by DESeq2



- **Euclidean distances between the samples, calculated after variance stabilizing transformation**

Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- Experiment level quality control
- **Differential expression analysis**
- Annotation
- Pathway analysis

Software packages for DE analysis

- **edgeR**
- **DESeq2**
- **Sleuth**
- **DRIMSeq**
- **DEXSeq**
- **Cuffdiff, Ballgown**
- **Limma + voom, limma + vst**
- **...**

Differential gene expression analysis

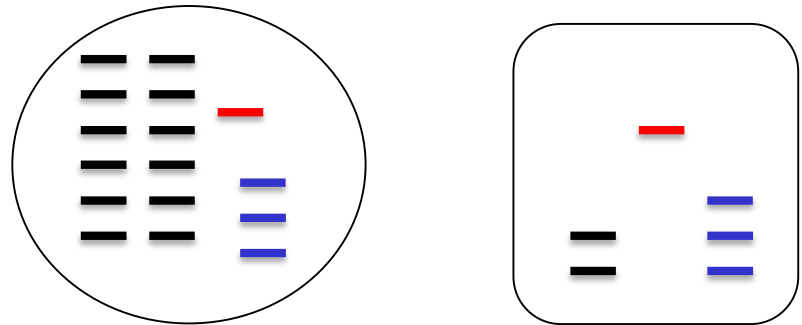
- **Normalization**
- **Dispersion estimation**
- **Log fold change estimation**
- **Statistical testing**
- **Filtering**
- **Multiple testing correction**

Differential expression analysis: Normalization

Normalization

➤ For comparing gene expression between (groups of) samples, normalize for

- Library size (number of reads obtained)
- RNA composition effect



➤ The number of reads for a gene is also affected by transcript length and GC content

- When studying differential gene expression we *assume that they stay the same*

Normalization by edgeR and DESeq

- **Aim to make normalized counts for non-differentially expressed genes similar between samples**
 - Do not aim to adjust count distributions between samples
- **Assume that**
 - Most genes are not differentially expressed
 - Differentially expressed genes are divided equally between up- and down-regulation
- **Do not transform data, but use normalization factors within statistical testing**

Normalization by edgeR and DESeq – how?

➤ DESeq(2)

- Take geometric mean of gene's counts across all samples
- Divide gene's counts in a sample by the geometric mean
- Take median of these ratios → sample's normalization factor (applied to read counts)

➤ edgeR

- Select as reference the sample whose upper quartile is closest to the mean upper quartile
- Log ratio of gene's counts in sample vs reference → M value
- Take weighted trimmed mean of M-values (TMM) → normalization factor (applied to library sizes)
 - Trim: Exclude genes with high counts or large differences in expression
 - Weights are from the delta method on binomial data

edgeR and DESeq2 expect raw read counts

- **Raw counts are needed to assess the quantification uncertainty**
- **Uncertainty information is lost if counts are transformed to FPKM**
 - FPKM = fragments per kilobase per million mapped reads.
 - Normalizes for gene length and library size. Example:
 - 20 kb transcript has 400 counts, library size is 20 million reads: $FPKM = (400/20) / 20$
 - 0.5 kb transcript has 10 counts, library size is 20 million reads: $FPKM = (10/0.5) / 20$
→ in both cases $FPKM = 1$, but it is less likely to get 400 reads just by chance
- **The negative binomial assumption of edgeR and DESeq2 is flexible enough to deal with gene-level counts summarized from Salmon's transcript-level abundance estimates**

Differential expression analysis: Dispersion estimation



Dispersion

- **When comparing gene's expression levels between groups, it is important to know also its within-group variability**
- **Dispersion = (BCV)²**
 - BCV = gene's biological coefficient of variation
 - E.g. if gene's expression typically differs from replicate to replicate by 20% (so BCV = 0.2), then this gene's dispersion is $0.2^2 = 0.04$
- **Note that the variability seen in counts is a sum of 2 things:**
 - Sample-to-sample variation (dispersion)
 - Uncertainty in measuring expression by counting reads

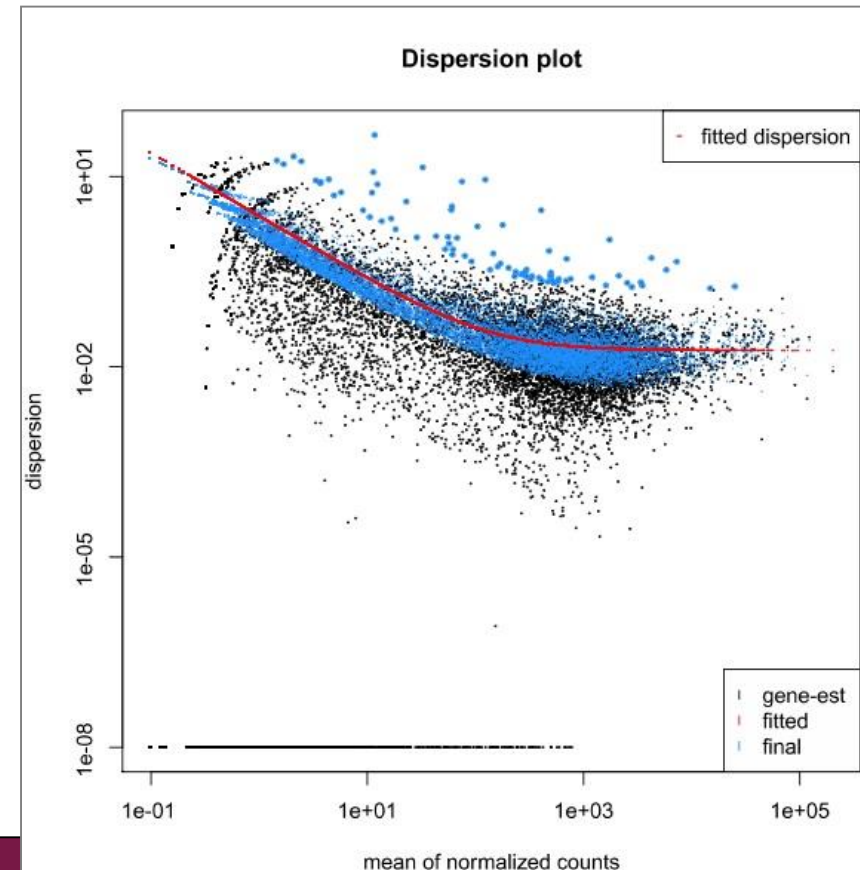


How to estimate dispersion reliably?

- **We cannot typically afford tens or hundreds of biological replicates**
 - it is difficult to estimate within-group variability
- **Solution: pool information across genes which are expressed at similar level**
 - assumes that genes of similar average expression strength have similar dispersion
- **Different approaches**
 - edgeR
 - DESeq2

Dispersion estimation by DESeq2

- Estimates genewise dispersions using maximum likelihood
- Fits a **curve** to capture the dependence of these estimates on the average expression strength
- Shrinks **genewise values towards the curve** using an empirical Bayes approach
 - The amount of shrinkage depends on several things including sample size
 - Genes with high gene-wise dispersion estimates are dispersion outliers (blue circles above the cloud) and they are not shrunk



Differential expression analysis: Statistical testing

Generalized linear models

- **Model the expression of each gene as a linear combination of explanatory factors (eg. group, time, patient)**

- $y = a + (b \cdot \text{group}) + (c \cdot \text{time}) + (d \cdot \text{patient}) + e$

y = gene's expression

a , b , c and d = parameters estimated from the data

a = intercept (expression when factors are at reference level)

e = error term

- **Generalized linear model (GLM) allows the expression value distribution to be different from normal distribution**

- Negative binomial distribution used for count data



Statistical testing

➤ DESeq2

- Generalized linear model, Wald test for significance
 - Log fold change is divided by its standard error and the resulting z statistic is compared to a standard normal distribution

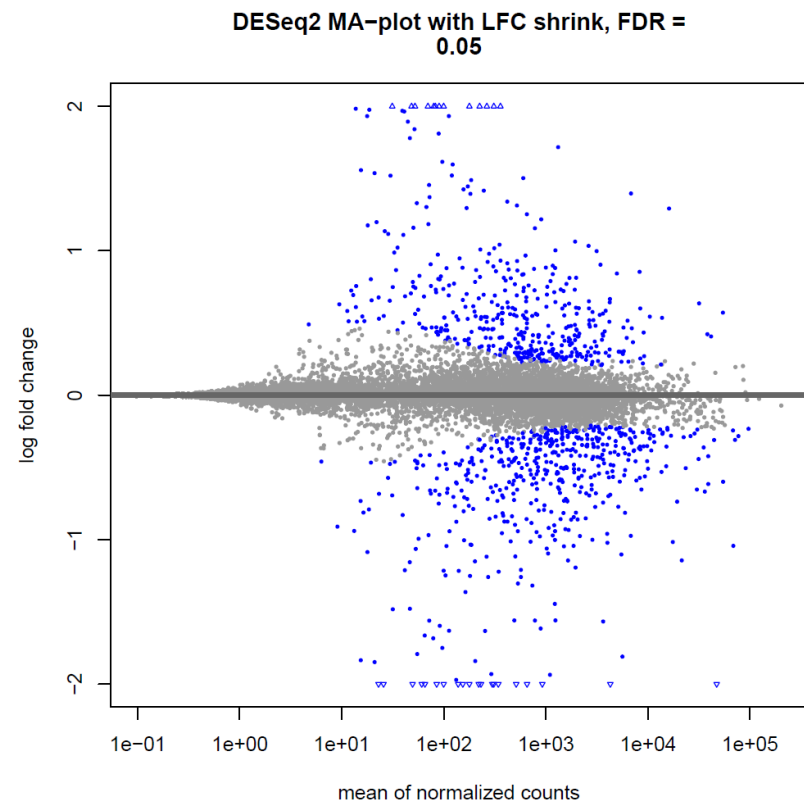
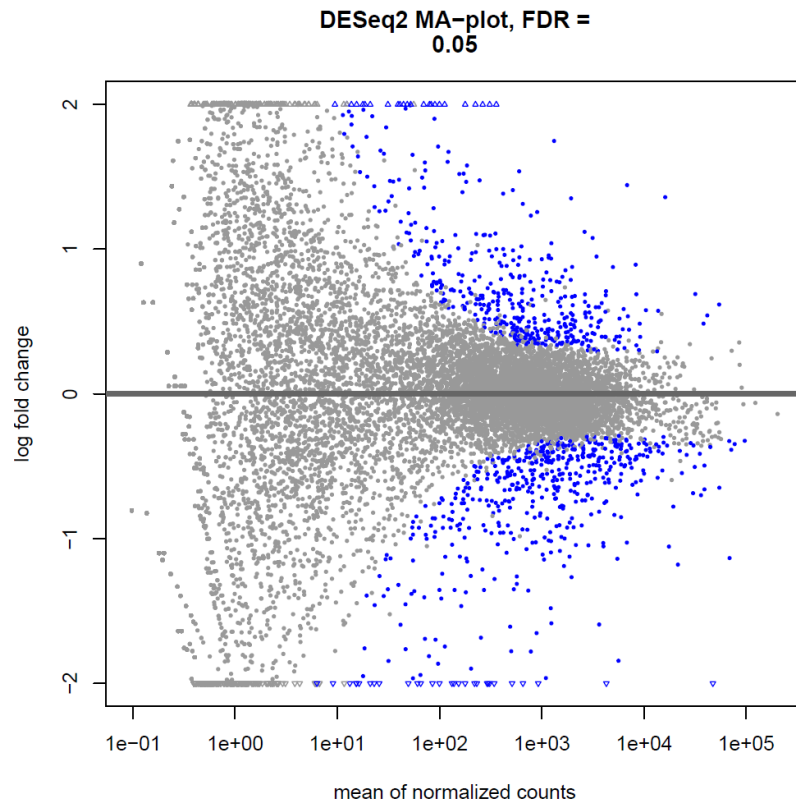
➤ edgeR

- Two group comparisons
 - Exact test for negative binomial distribution
- Multifactor experiments
 - Generalized linear model, likelihood ratio test

DESeq2 shrinks log2 fold changes estimates

➤ uses the **ashr (Adaptive SHRinkage)** method

- the amount of shrinkage is determined from the data
- measurements with high standard error will undergo more shrinkage
- <https://github.com/stephens999/ashr>



Multiple testing correction

- **We test thousands of genes, so it is possible that some genes get good p-values just by chance**
 - This problem is much bigger, if you test transcripts (DTE)
- **To control this problem of false positives, p-values need to be corrected for multiple testing**
- **Several methods are available, the most popular one is the Benjamini-Hochberg correction (BH)**
 - largest p-value is not corrected
 - second largest $p = (p * n) / (n-1)$
 - third largest $p = (p * n) / (n-2)$
 - ...
 - smallest $p = (p * n) / (n - n + 1) = p * n$
- **The adjusted p-value is FDR (false discovery rate)**

Filtering

- **Reduces the severity of multiple testing correction by removing some genes (makes n smaller)**
- **Filter out genes which have little chance of showing evidence for significant differential expression**
 - genes which are not expressed
 - genes which are expressed at very low level (low counts are unreliable)
- **Should be independent**
 - do not use information on what group the sample belongs to
- **DESeq2 selects filtering threshold automatically**



edgeR result table

- **logFC = log2 fold change**
- **logCPM = average log2 counts per million**
- **Pvalue = raw p-value**
- **FDR = false discovery rate (Benjamini-Hochberg adjusted p-value)**

de-list-edger.tsv

Showing all 757 rows.

	chip.treat	chip.trea	chip.trea	chip.unt	chip.un	chip.un	chip.unt	logFC	logCPM	PValue	FDR
FBgn0039155	56	62	74	1756	2238	1081	1229	-4.69899422081019	6.03560818295738	7.20308463504266e-135	6.22058389082284e-131
FBgn0029167	1649	1624	1175	6810	9182	4877	4919	-2.23433823802528	8.24793917660261	1.99034107006072e-67	8.59429274052218e-64
FBgn0034736	35	44	29	358	593	295	283	-3.49985798252498	4.04448825313873	6.11382510741391e-60	1.75996645425422e-56
FBgn0035085	179	212	172	1075	1292	749	861	-2.52597752387173	5.53857558363739	4.25001852624925e-54	9.17578999817213e-51
FBgn0000071	775	822	645	78	151	86	96	2.74080535171334	4.6808323300972	2.25722664890845e-48	3.89868186799468e-45
FBgn0029896	192	156	123	704	1274	595	611	-2.43819206791615	5.18811150378514	1.46605619431662e-45	2.11014354901973e-42
FBgn0039827	27	20	46	588	656	471	501	-4.27016967435994	4.60377905047843	3.03506500413406e-42	3.74440305367167e-39
FBgn0033764	196	184	155	16	21	12	10	3.48503218393543	2.5488391163408	2.526432524815e-41	2.7272839105378e-38
FBgn0034434	17	13	20	226	299	219	237	-4.01870183550384	3.44909319532712	2.01715887090541e-39	1.93557600101545e-36
FBgn0051092	349	384	331	64	91	49	44	2.4232155517234	3.68036599989021	4.54650866609765e-37	3.92636488404193e-34
FBgn0011260	650	455	466	84	139	81	75	2.35185375445073	4.24174827367645	7.37076473305667e-35	5.78672038497067e-32

DESeq2 result table

- baseMean = mean of counts (divided by size factors) taken over all samples
- **log2FoldChange = log2 of the ratio meanB/meanA**
- lfcSE = standard error of log2 fold change
- stat = Wald statistic
- pvalue = raw p-value
- **padj = Benjamini-Hochberg adjusted p-value**

de-list-deseq2-rt.tsv

Showing all 973 rows.

	chip.treat	chip.treat	chip.treat	chip.unt	chip.unt	chip.unt	chip.unt	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FBgn0039155	56	62	74	1756	2238	1081	1229	924.27	-4.17	0.15	-28.66	1.246e-180	9.941e-177
FBgn0026562	13037	17789	14377	61534	87864	55454	72147	47282.42	-2.32	0.09	-24.84	3.333e-136	1.330e-132
FBgn0029167	1649	1624	1175	6810	9182	4877	4919	4287.44	-2.13	0.1	-21.79	3.121e-105	8.300e-102
FBgn0039827	27	20	46	588	656	471	501	342.77	-3.49	0.18	-19.48	1.660e-84	3.311e-81
FBgn0035085	179	212	172	1075	1292	749	861	654.94	-2.36	0.12	-19.01	1.332e-80	2.126e-77
FBgn0034736	35	44	29	358	593	295	283	231.7	-2.93	0.17	-17.25	1.119e-66	1.488e-63
FBgn0000071	775	822	645	78	151	86	96	359.53	2.42	0.14	16.76	4.495e-63	5.123e-60
FBgn0034434	17	13	20	226	299	219	237	153.84	-3.11	0.19	-16.25	2.071e-59	2.065e-56

Interactive Venn diagram

- You can compare result files, e.g. were the same genes found
 - Select 2-3 tsv files and click **Draw**
- Make a new gene list
 - Click on the image to select an area (e.g. the intersection) and click **Create file** → new gene list appears
 - If the files have columns with the same name (e.g. padj), the values to the new gene list are taken from the input file that you selected first

The screenshot displays the Chipster software interface, which is used for bioinformatics workflows. The interface is split into two main panels.

Left Panel: Workflow Editor

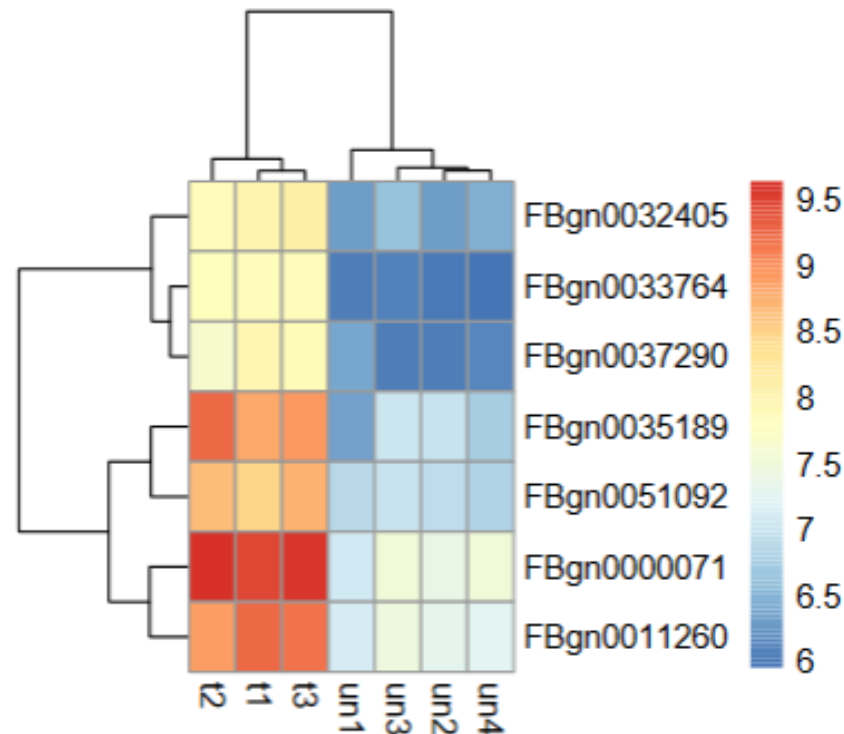
- At the top, there are navigation tabs: "Chipster", "Analyze", "Sessions", "Manual", and "Contact".
- Below the tabs, there are "Files" and "Workflow" sections. The "Workflow" section shows a tree diagram of a workflow with nodes labeled "pdf", "txt", and "tsv".
- A search bar labeled "Find file" is present.
- An "Add file" button is visible.

Right Panel: Venn Diagram

- The title is "de-list-edger.tsv de-list-deseq2.tsv de-list-deseq2-rt.tsv".
- The diagram is a three-set Venn diagram with the following counts:
 - de-list-edger.tsv only: 53
 - de-list-deseq2-rt.tsv only: 247
 - de-list-deseq2.tsv only: 0
 - Intersection of de-list-edger.tsv and de-list-deseq2-rt.tsv: 2
 - Intersection of de-list-edger.tsv and de-list-deseq2.tsv: 3
 - Intersection of de-list-deseq2-rt.tsv and de-list-deseq2.tsv: 18
 - Intersection of all three: 699
- On the right side of the Venn diagram, there is a "Common denominator" dropdown menu set to "identifier".
- Below that, it says "Selected: 699" and a green "Create file" button.
- A list of gene identifiers is shown, including:
 - FBgn0039155
 - FBgn0029167
 - FBgn0034736
 - FBgn0035085
 - FBgn0000071
 - FBgn0029896
 - FBgn0039827
 - FBgn0033764
 - FBgn0034434
 - FBgn0051092
 - FBgn0011260
 - FBgn0038832
 - FBgn0024288
 - FBgn0052407
 - FBgn0003501
 - FBgn0034897
 - FBgn0040091
 - FBgn0026562
 - FBgn0035189

Heatmap of differentially expressed genes

- Use the tool **Heatmap for RNA-seq results**
- **Counts are transformed using variance stabilization transformation**
 - calculated using the experiment-wide trend of variance over mean
 - You need to give 2 input files: the original count table and the list of differentially expressed genes. Check that they are correctly assigned!



What if I have several experimental factors?

➤ The tool **Differential expression using edgeR for multivariate experiments** can cope with 3 main effects and pairing

➤ Main effects can be treated as

- Linear = is there a trend towards higher numbers?
- Factor = are there differences between the levels?

If the main effect has only two levels (e.g. control and cancer), selecting linear or factor gives the same result

➤ Note that the result table contains all the genes, so in order to get the differentially expressed genes you have to filter it

- Use the tool **Utilities / Filter using a column value**
- Select the FDR column that corresponds to the comparison of your interest



PValue-as.factor(group)2	FDR-as.factor(group)2	logFC-as.factor(time)2	logCPM-as.factor(time)2	LR-as.factor(time)2	PValue-as.factor(time)2	FDR-as.factor(time)2	logFC-as.factor(patient)2	
4.4e-05	0.01647	-1.583315	5.782999	308.077737	0	0	-2.002844	
0.000102	0.027761	1.473723	-0.397191	12.823652	0.000342	0.003208	-0.696252	
0	0.000276	0.037768	6.612959	0.328104	0.566777	0.746462	0.287009	
0.000182	0.037215	0.56783	7.896608	84.211177	0	0	0.667624	
0.000245	0.044683	0.319444	7.146574	27.072628	0	5e-06	-0.029676	
8e-06	0.004906	-0.087083	9.060264	0.78067	0.376936	0.592923	0.216045	
0.000285	0.049417	-0.073242	7.146943	1.895127	0.168625	0.35584	-0.445068	

Analyzing differential gene expression: things to take into account

- **Biological replicates are important!**
- **Normalization is required in order to compare expression between samples**
 - Different library sizes
 - RNA composition bias caused by sampling approach
- **Raw counts are needed to assess measurement precision**
 - Counts are the "the units of evidence" for expression
 - Gene-level counts summarized from Salmon's transcript-level estimates seem to be ok
- **Multiple testing problem**

Data analysis workflow

- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- Experiment level quality control
- Differential expression analysis
- **Annotation of gene identifiers**
- Pathway analysis

Add gene symbols and descriptions to data

➤ Tool Utilities / Annotate Ensembl identifiers

- Ensembl IDs can be
 - in the first column, with or without a title
 - in the middle of the file if the column title is `ensembl_id`
- Fetches annotations from the EBI
 - Max100 000 Ensembl IDs can be annotated in one job

annotated.tsv

×

Showing all 3600 rows.

	symbol	description	chr	start	end	length	sequence	chip.sample001.t
ENSG00000064042	LIMCH1	LIM and calponin homology domains 1 [Source:HGNC Symbol;Acc:HGNC:29191]	4	41359606	41700044	340438	NA	1948
ENSG00000185499	MUC1	mucin 1, cell surface associated [Source:HGNC Symbol;Acc:HGNC:7508]	1	155185823	155192916	7093	NA	630
ENSG00000198722	UNC13B	unc-13 homolog B [Source:HGNC Symbol;Acc:HGNC:12566]	9	35161991	35405338	243347	NA	851
ENSG00000013588	GPRC5A	G protein-coupled receptor class C group 5 member A [Source:HGNC Symbol;Acc:HGNC:9836]	12	12890781	12917937	27156	NA	3009
ENSG00000131400	NAPSA	napsin A aspartic peptidase [Source:HGNC Symbol;Acc:HGNC:13395]	19	50358476	50365830	7354	NA	3532
ENSG00000112782	CLIC5	chloride intracellular channel 5 [Source:HGNC Symbol;Acc:HGNC:13517]	6	45898450	46080395	181945	NA	1436
ENSG00000170017	ALCAM	activated leukocyte cell adhesion molecule [Source:HGNC Symbol;Acc:HGNC:400]	3	105366908	105576900	209992	NA	1019
ENSG00000111319	SCNN1A	sodium channel epithelial 1 subunit alpha [Source:HGNC Symbol;Acc:HGNC:10599]	12	6346842	6377730	30888	NA	494
ENSG00000122852	SFTPA1	surfactant protein A1 [Source:HGNC Symbol;Acc:HGNC:10798]	10	79610938	79615455	4517	NA	28321

Summary of DGE analysis steps and files

- **Quality control / Read quality with MultiQC** → html report
- (Preprocessing / Trim reads with Trimmomatic → FASTQ)
- (Utilities / Make a list of file names → txt)
- **Alignment / HISAT2 or STAR** → BAM
- **Quality control / RNA-seq quality metrics with RSeQC** → pdf
- **RNA-seq / Count aligned reads per genes with HTSeq** → tsv
- **Utilities / Define NGS experiment** → tsv
- **Quality control / PCA and heatmap of samples with DESeq2** → pdf
- **RNA-seq / Differential expression using DESeq2** → tsv
- **Utilities / Annotate Ensembl identifiers** → tsv

Data analysis workflow

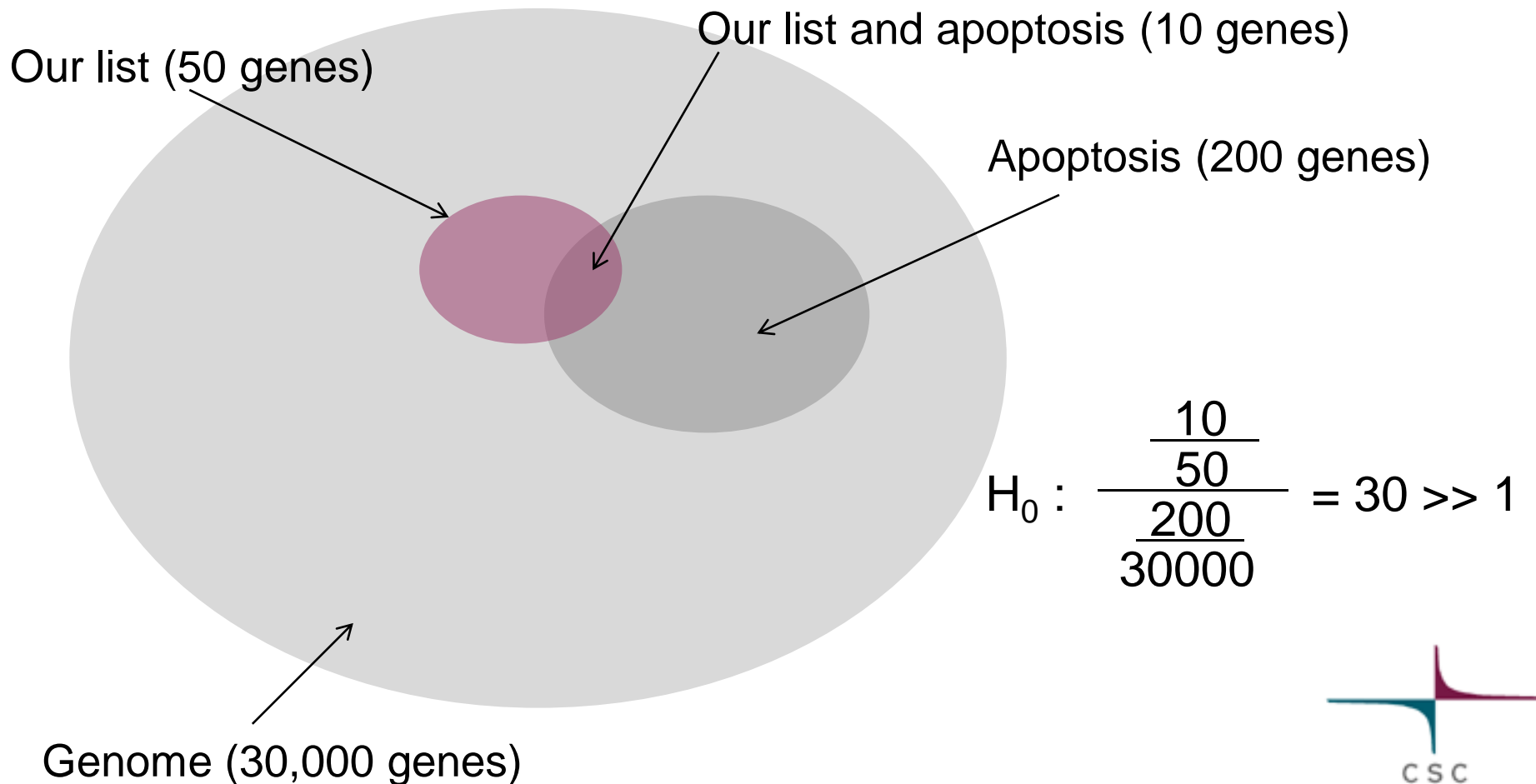
- Quality control of raw reads
- Preprocessing (trimming / filtering) if needed
- Alignment to reference genome
- Alignment level quality control
- Quantitation
- Experiment level quality control
- Differential expression analysis
- Annotation
- **Pathway analysis**

Pathway analysis – why?

- **Statistical tests can yield thousands of differentially expressed genes**
- **It is difficult to make "biological" sense out of the result list**
- **Looking at the bigger picture can be helpful, e.g. which pathways are differentially expressed between the experimental groups**
- **Databases such as KEGG, GO, Reactome and ConsensusPathDB provide grouping of genes to pathways, biological processes, molecular functions, etc**

Gene set enrichment analysis

1. Perform a statistical test to find differentially expressed genes
2. Check if the list of differentially expressed genes is "enriched" for some pathways



ConsensusPathDB

- **One-stop shop: Integrates pathway information from 32 databases covering**
 - biochemical pathways
 - protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions
- **Developed by Ralf Herwig's group at the Max-Planck Institute in Berlin**
- **ConsensusPathDB over-representation analysis tool is integrated in Chipster**
 - runs on the MPI server in Berlin



Design of experiments

When planning an experiment, consider

- **The number of biological replicates needed. Depends on**
 - Biological variability and technical noise
 - Expression level, fold change and sequencing depth
- **Sample pairing**
- **Sequencing decisions**
 - Number of reads per sample
 - Read length (longer is better)
 - Paired end or single end (PE is better)
 - Stranded or unstranded (stranded is better)
 - Batch effects

How many biological replicates?

- **Publication quality data needs at least 3 biological replicates per sample group**
 - This can be sufficient for cell-cultures and/or test animals
- **More reasonable numbers:**
 - Cell cultures / test animals: 3 is minimum, 4-5 OK, >7 excellent
 - Patients: 3 is minimum, 10-20 OK, >50 good
 - Power analysis can be used to estimate sample sizes



How many reads per sample do I need?

- **Depends on the transcriptome and the analysis goal**
 - Differential expression 10-25 M reads
 - Allele specific expression 50-100 M
 - Alternative splicing 50-100 M
 - *De novo* assembly >100 M

<https://genohub.com/recommended-sequencing-coverage-by-application/>



More reads or more replicate samples?

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

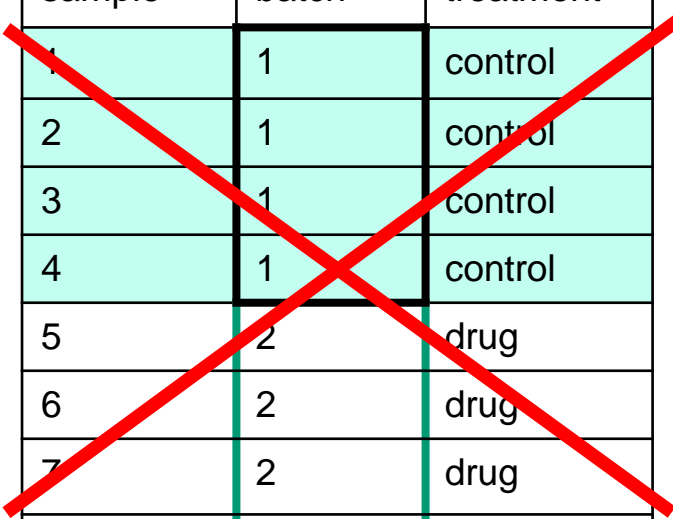
Balance sample groups across batches

- You can't account for a batch effect if all your control samples were run in one batch and the drug samples in the other
 - DESeq2 would give an error: "*The model matrix is not full rank*"
- **Balance sample groups cross batches**

Case: You have 8 samples, 4 controls and 4 treated samples. There will be 2 batches in the sequencing run. How do you form the batches?

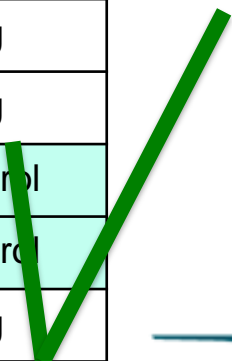
Option A:

sample	batch	treatment
1	1	control
2	1	control
3	1	control
4	1	control
5	2	drug
6	2	drug
7	2	drug
8	2	drug



Option B:

sample	batch	treatment
1	1	control
2	1	control
5	1	drug
6	1	drug
3	2	control
4	2	control
7	2	drug
8	2	drug

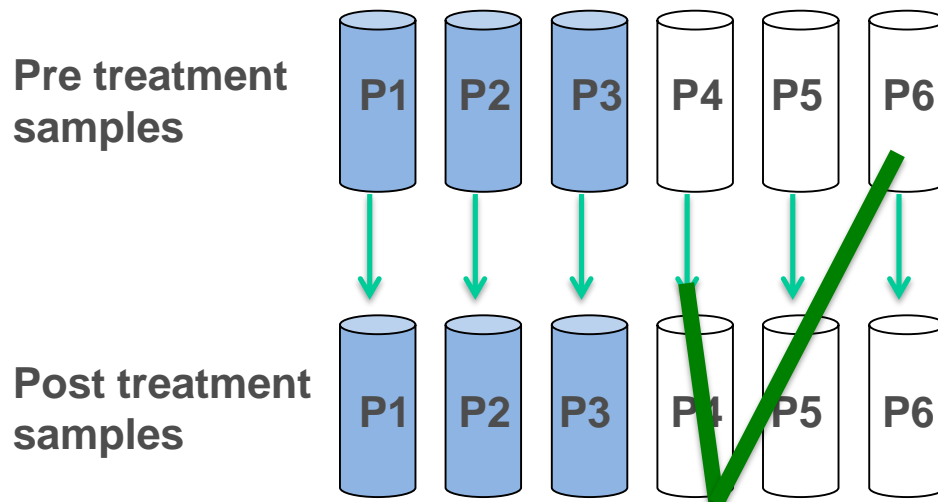


Paired samples

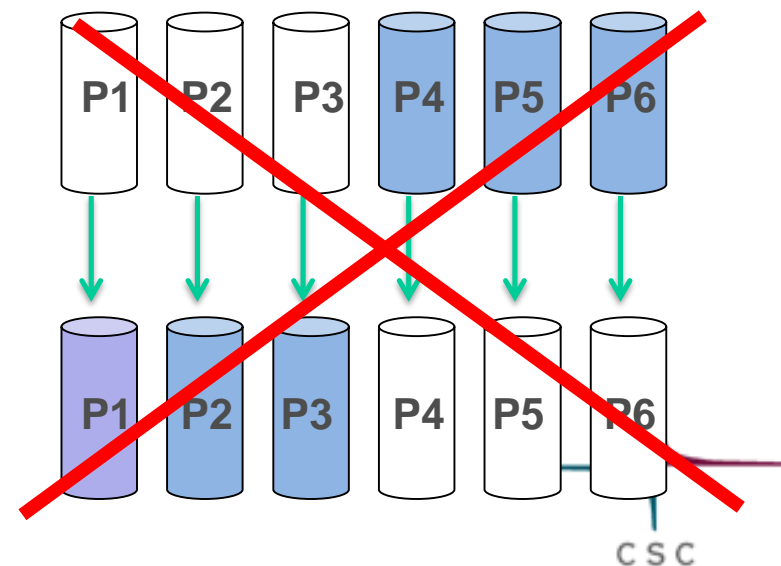
- **Individual variation can be tackled using a matched control**
 - Before and after treatment samples from the same patient
 - Tumor vs. normal samples from the same patient

Case : 6 patients, 2 samples from each. You can afford to sequence only 6 samples. Which option do you choose?

Option A:

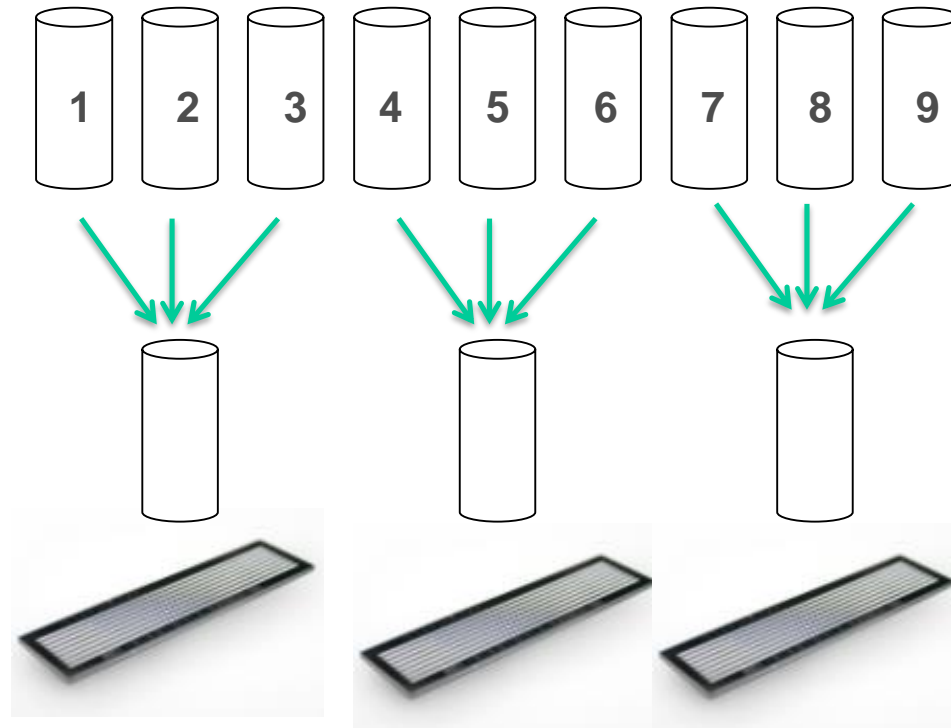


Option B:



Pooling

- **When possible, measure each sample on its own**
 - Pool only if there is not enough material to run the samples individually)
- **If one of your samples is an outlier or has a contamination, the whole pool is unusable**

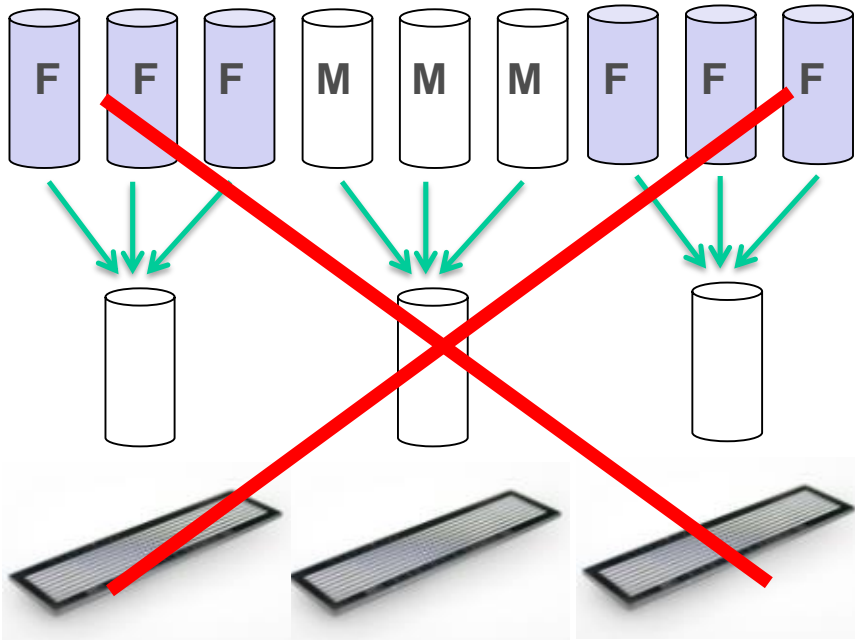


Pooling

- **Make pools as similar as possible**
- **Avoid pooling of samples of similar type into one pool**

Case: We have 9 control samples, 6 samples from females and 3 from males. We need to pool 3 samples together. How do we do it?

Option A:



Option B:

