RNA-seq data analysis with Chipster

RNA-seq data analysis workshop 7.-10.1.2014

Eija Korpelainen
chipster@csc.fi

CSC

Outline

1. **Introduction to Chipster**
2. **Introduction to RNA-seq**
3. **RNA-seq data analysis, part I**
   - Quality control, preprocessing
   - Alignment to reference
   - Manipulation of alignment files
   - Alignment level quality control
   - Quantitation
   - Visualization of alignments in genome browser
4. **Exercises**
5. **RNA-seq data analysis, part II**
   - Differential expression analysis
6. **More exercises**

CSC

Introduction to Chipster

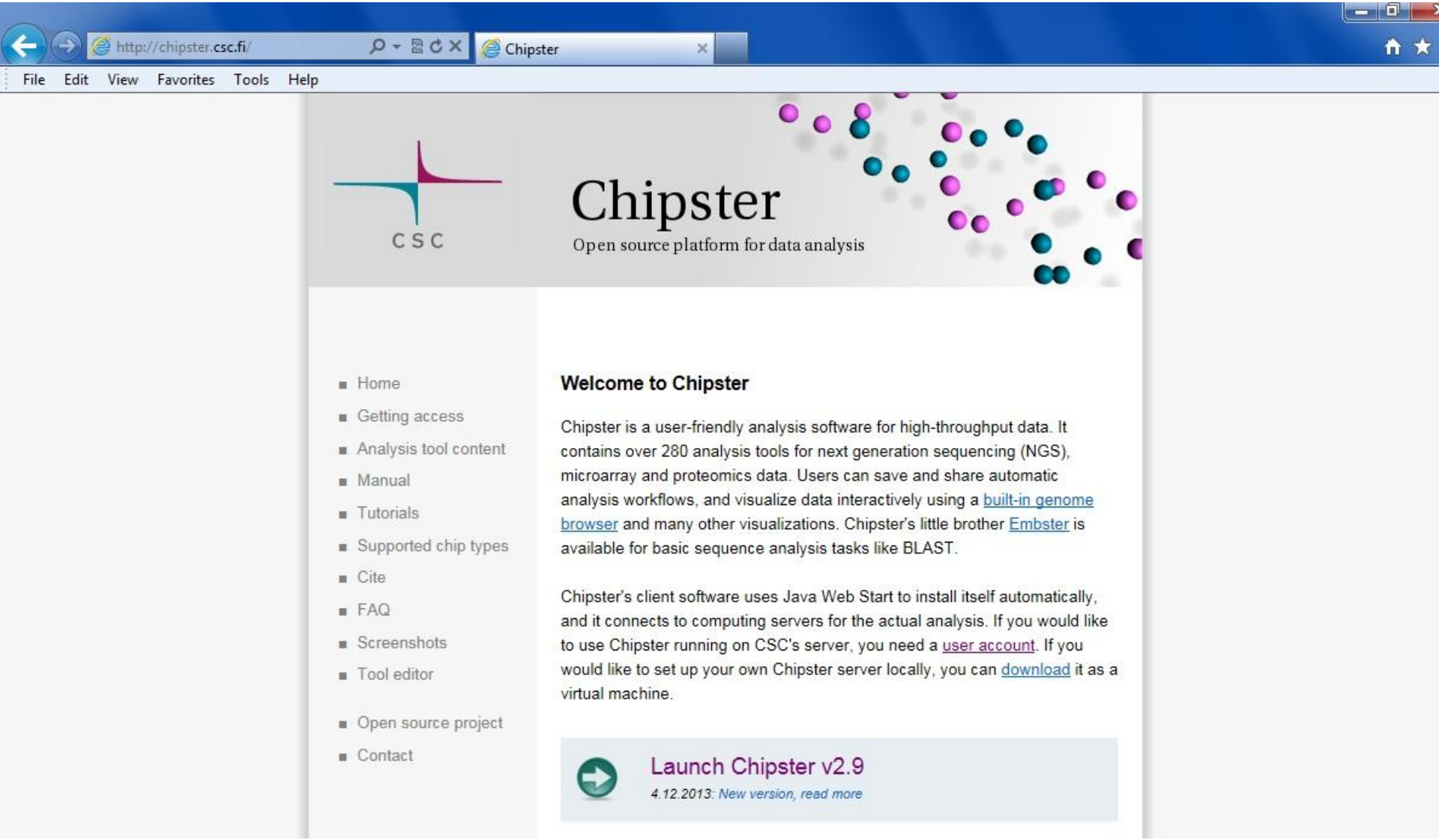
Chipster

- **Provides an easy access to over 280 analysis tools**
  - No programming or command line experience required
- **Free, open source software**

- **What can I do with Chipster?**
  - analyze and integrate high-throughput data
  - visualize data efficiently
  - share analysis sessions
  - save and share automatic workflows

CSC

Chipster start and info page: chipster.csc.fi



Chipster 2.9.0 (build 1418)

File Edit View Workflow Help

Datasets

- Datasets
  - cisRED_STAT1_TRACK.tsv
  - control_chr_1_sorted.bam
  - control_chr_1_sorted.bam.bai
  - treatment_chr_1_sorted.bam
  - treatment_chr_1_sorted.bam.bai
  - analysis-log.txt
  - model-plot.pdf
  - negative-peaks.tsv
  - positive-peaks.bed
  - positive-peaks.tsv
  - filtered-NGS-results.tsv
  - filtered-NGS-results.tsv

Analysis tools

Microarrays | NGS | Show parameters | Run

- Quality control
- Filtering
- Utilities
- Matching sets of genomic regions
- Alignment
- Variants
- RNA-seq
- miRNA-seq
- ChIP-seq and FAIRE-seq
- CNA-seq
- Methyl-seq
- Metagenomics

- Find peaks using MACS, treatment only
- Find peaks using MACS, treatment vs. control
- Find common motifs and match to JASPAR
- Find the nearest genes for regions
- Find unique and annotated genes
- GO enrichment for list of genes
- Find broad peaks using F-seq

This tool will search for statistically significantly enriched genomic regions in sequencing data from a ChIP-seq experiment. The analysis is performed on one or more treatment samples relative to one or more control samples.

More help | Show tool sourcecode

Workflow

Fit

bam, bai, bam, bai, tsv → tsv, txt, tsv, pdf, bed → tsv → tsv → tsv → tsv, pdf, pdf, txt, pdf → tsv, tsv → html, tsv

Notes for dataset

Hide

Import / Import data

Thu Jan 02 15:19:34 EET 2014

Add your notes here...

Visualisation

Method: Genome browser | Help | Maximise | Detach

Annotations

Show all

RNF115-001

144M 319k, 144M 320k, 144M 321k, 144M 322k, 144M 323k, 144M 324k, 144M 325k, 144M 326k, 144M 327k

POLR3C-001

POLR3C-002

control_chr_1_sorted.bam

treatment_chr_1_sorted.bam

filtered-NGS-results.tsv

Settings | Selected | Legend

Genome

Human hg18 (NCBI36.54)

Location

Chromosome: 1

Location (gene or position): 144323304

View size: 9 kb

Go

Options

- Reads
- Highlight SNPs
- Density graph
- Low complexity regions

Coverage type: strand-specific

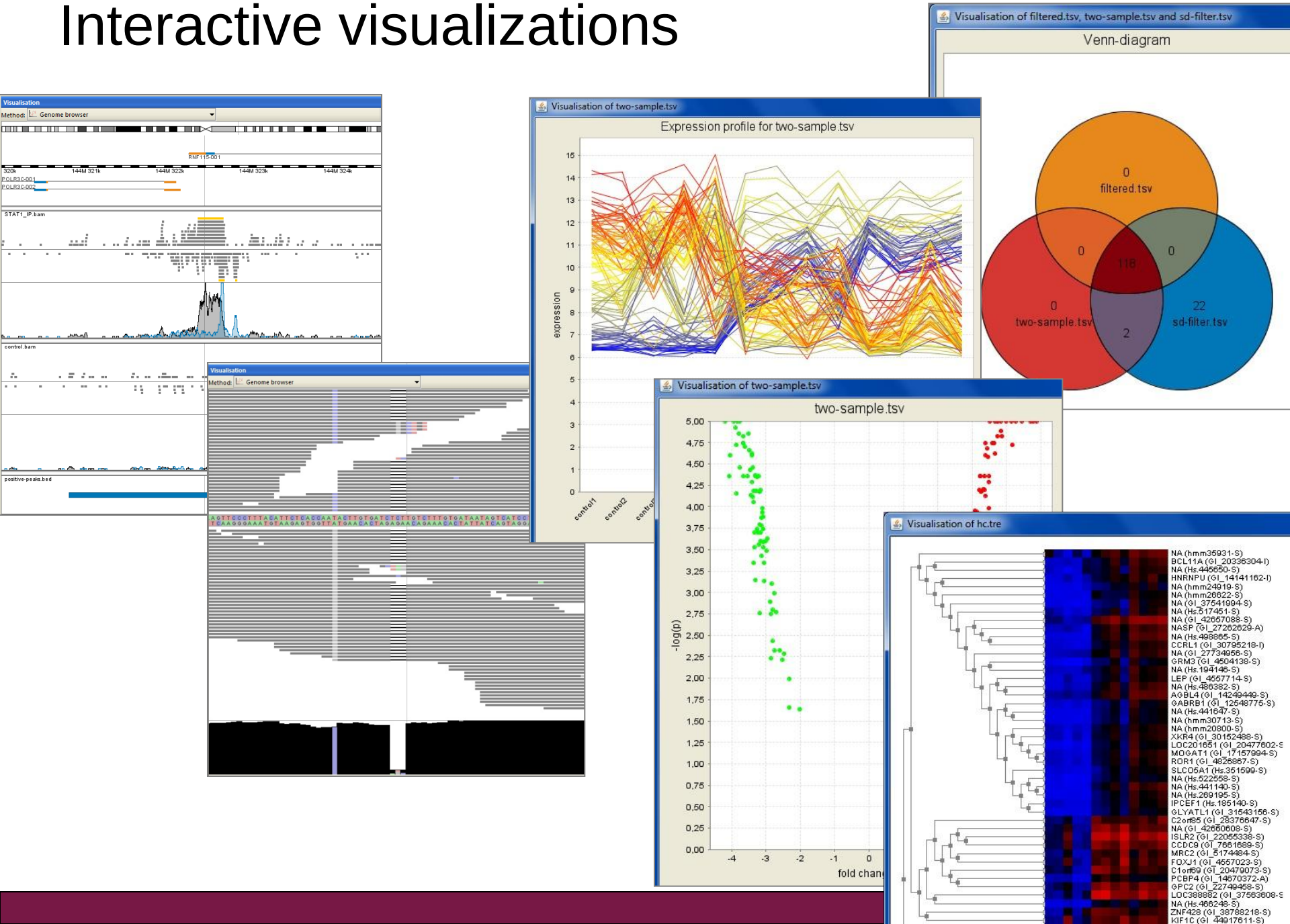
Coverage scale: 50

External links

View this region in Ensembl or UCSC genome browser.

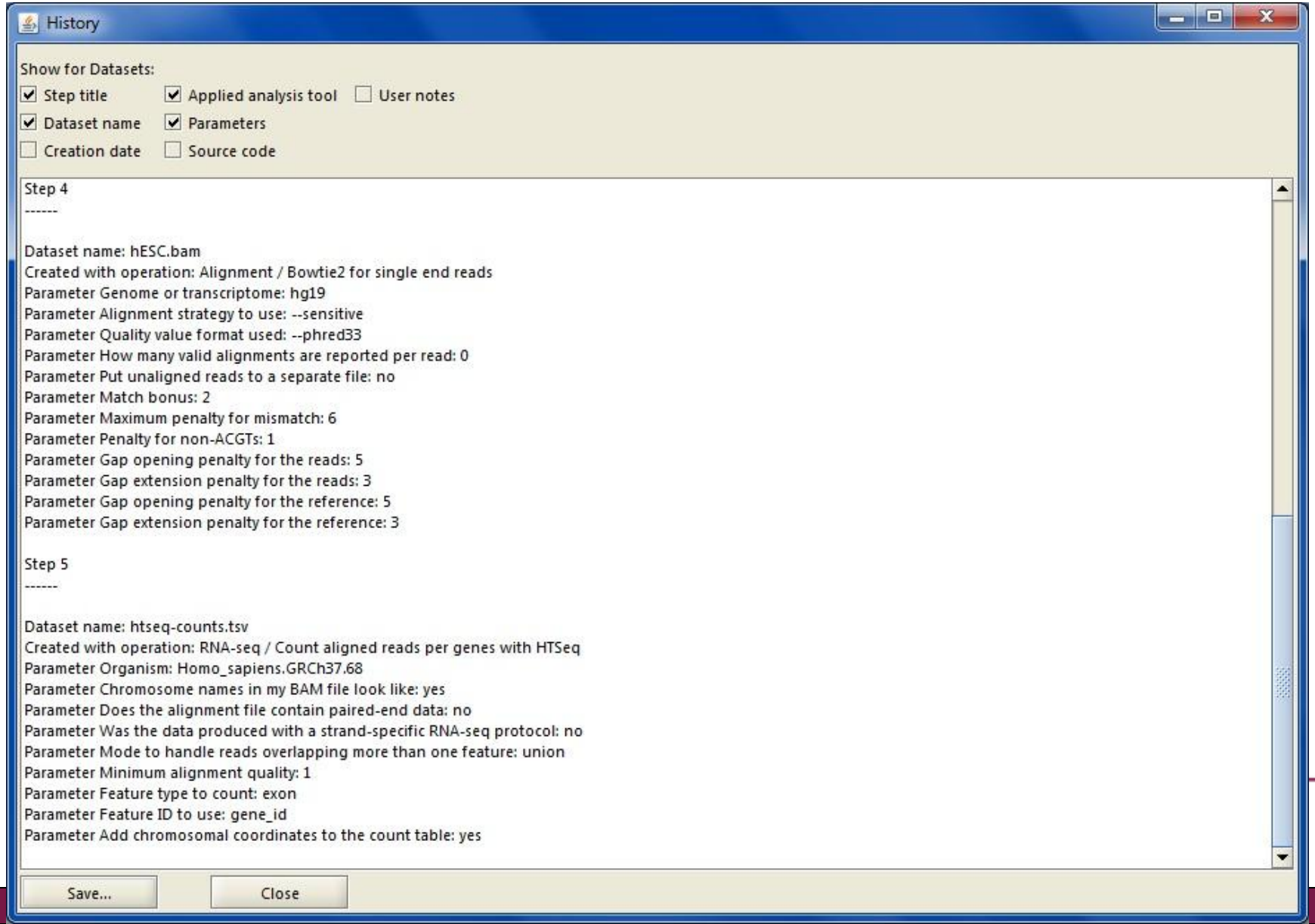
Connected to chipster.csc.fi | Ready | 99M / 870M

Interactive visualizations



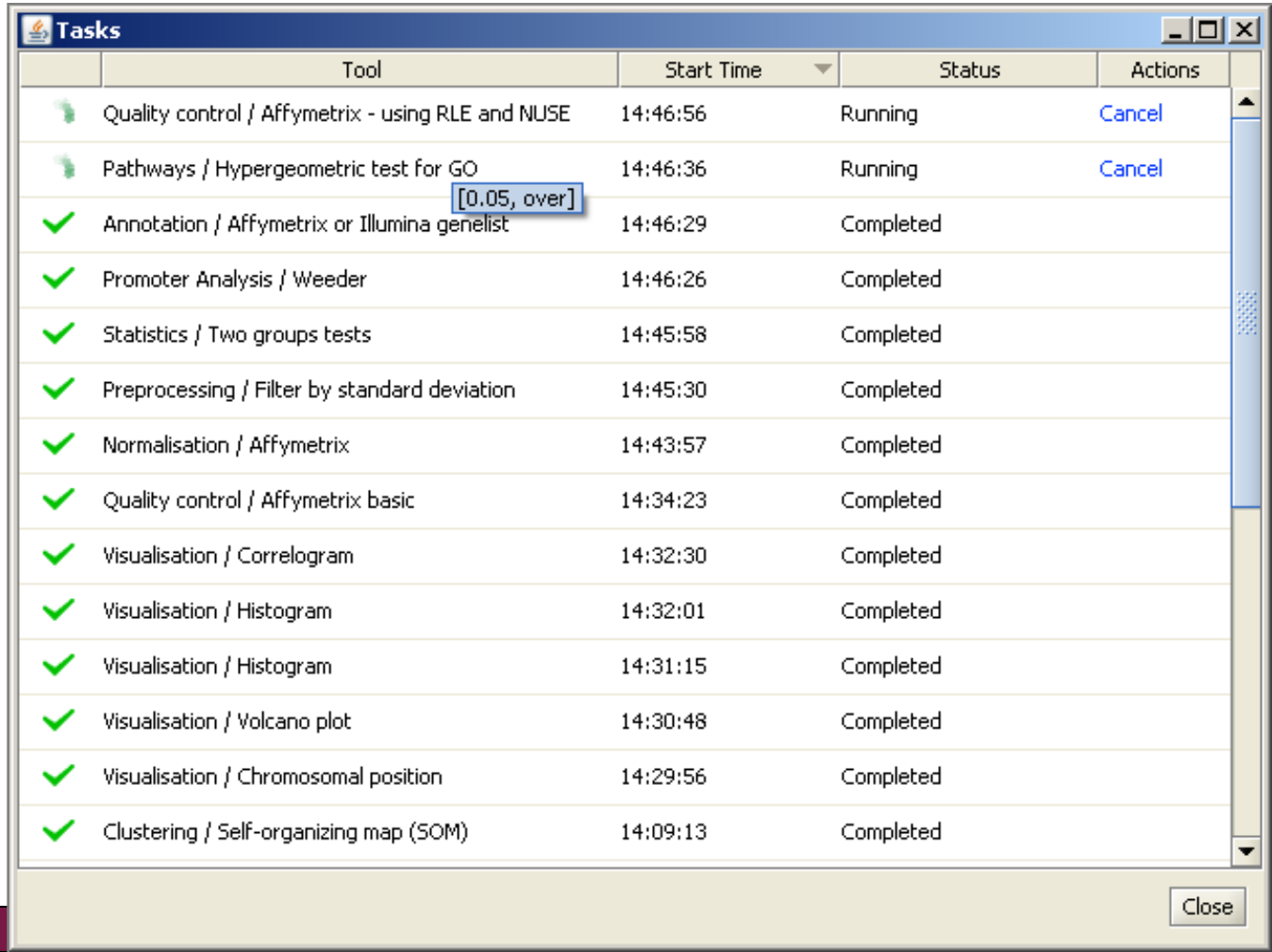
Analysis history is saved automatically

-you can add tool source code to reports if needed



Task manager

- **You can run many analysis jobs at the same time**
- **Use Task manager to**
  - view status
  - cancel jobs
  - view time
  - view parameters



Analysis sessions

- **In order to continue your work later, <u>you have to save the analysis session.</u>**

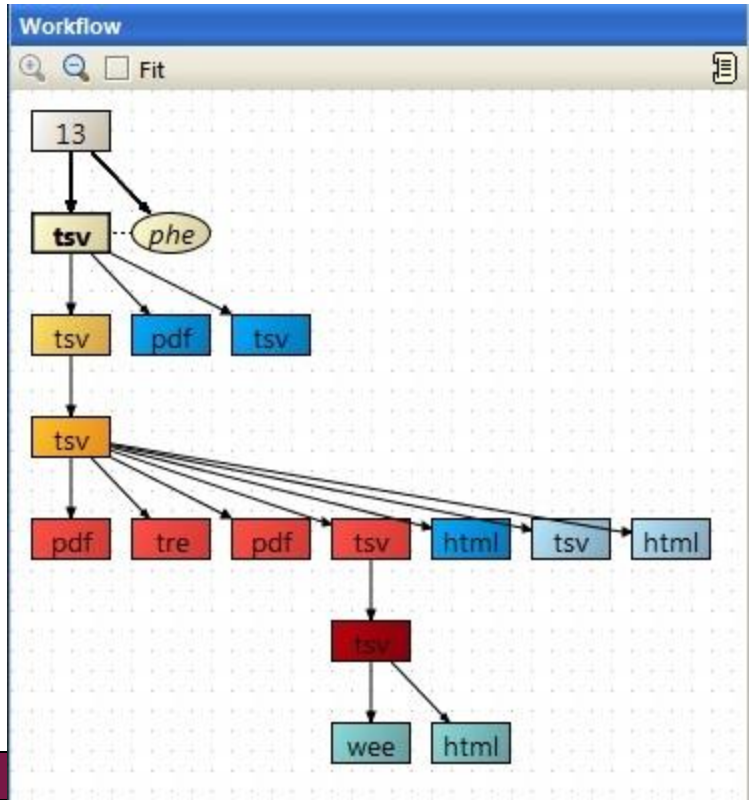
- **Saving the session will save all the files and their relationships. The session is packed into a single .zip file and saved on your computer (in the next Chipster version you can also save it on the server).**

- **Session files allow you to continue the work on another computer, or share it with a colleague.**

- **You can have multiple analysis sessions saved separately, and combine them later if needed.**

Workflow panel

- **Shows the relationships of the files**
- **You can move the boxes around, and zoom in and out.**
- **Several files can be selected by keeping the Ctrl key down**
- **Right clicking on the data file allows you to**
  - Save an individual result file (”Export”)
  - Delete
  - Link to another data file
  - Save workflow

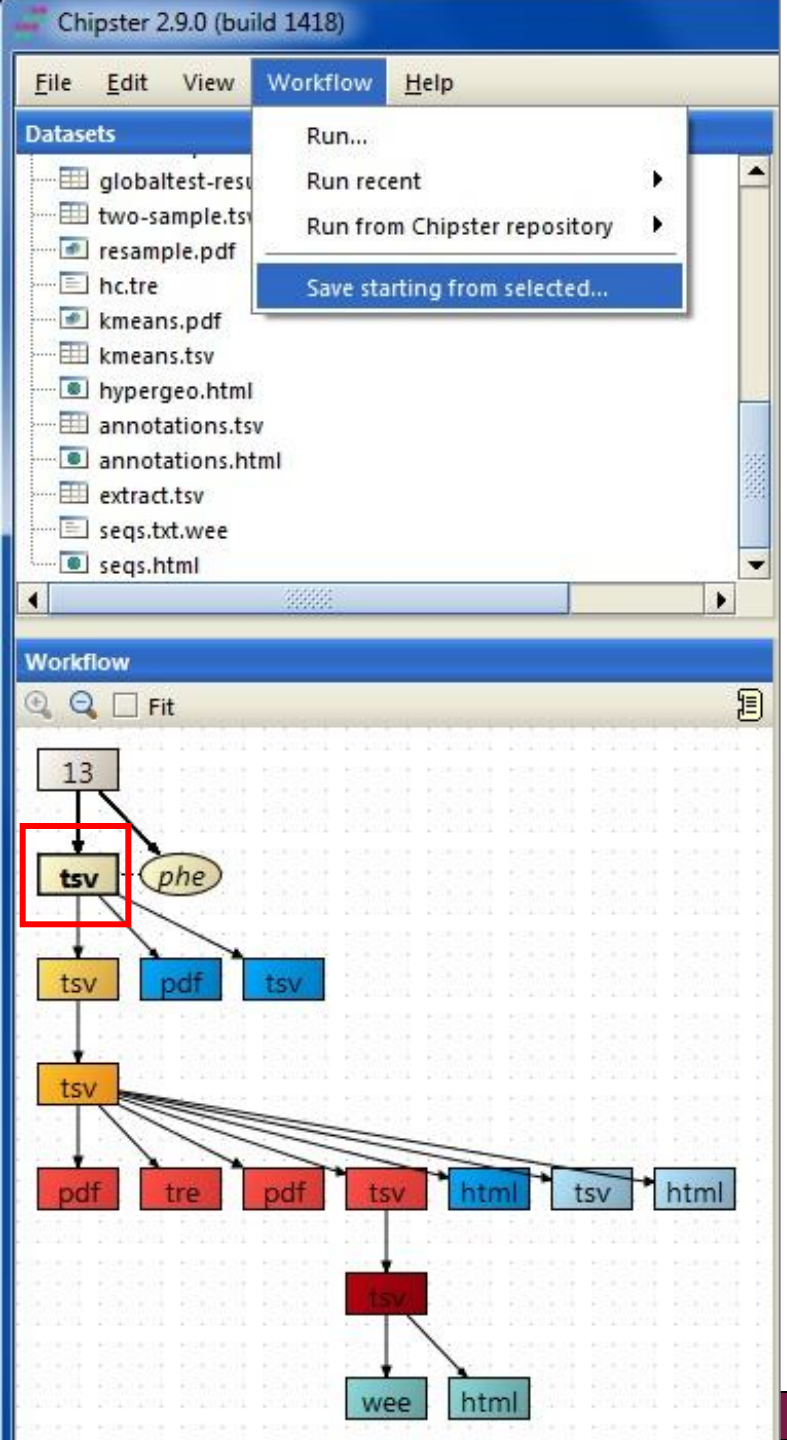


Workflow – reusing and sharing your analysis pipeline

- **You can save your analysis steps as a reusable automatic ”macro”, which you can apply to another dataset**

- **When you save a workflow, all the analysis steps and their parameters are saved as a script file, which you can share with other users**

Saving and using workflows



- **Select the <u>starting point</u> for your workflow**
- **Select ”Workflow/ Save starting from selected”**
- **Save the workflow file on your computer with a meaningful name**
  - Don’t change the ending (.bsh)
- **To run a workflow, select**
  - Workflow->Open and run
  - Workflow->Run recent (if you saved the workflow recently).

CSC

Problems? Send us a support request

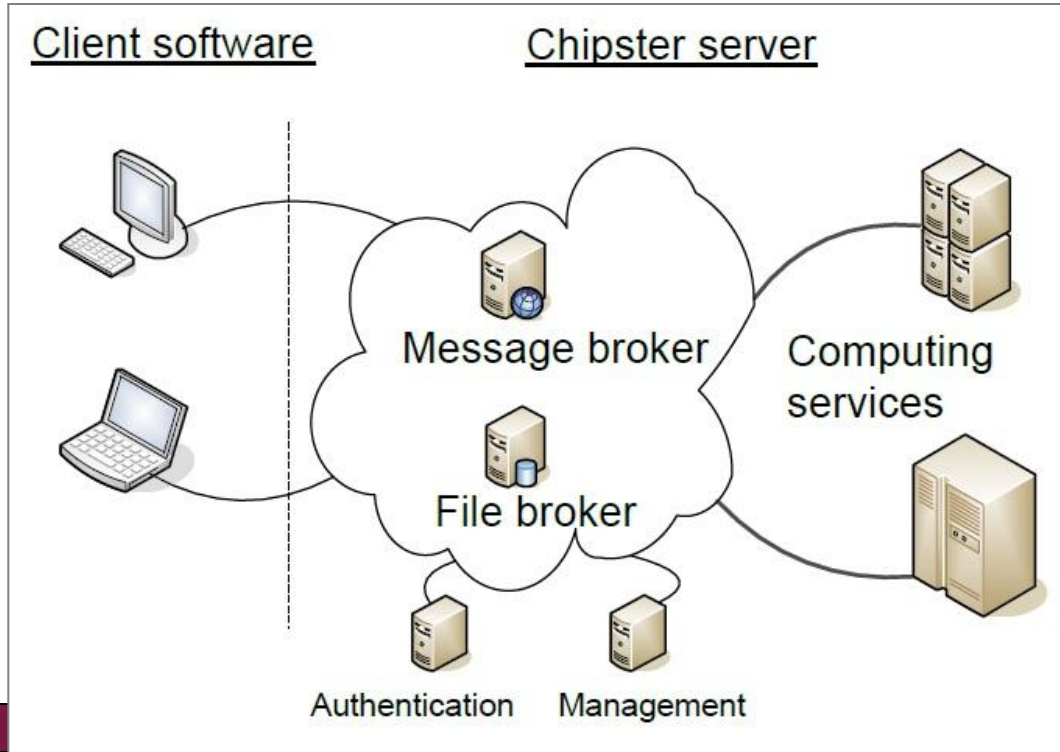
-request includes the error message and link to analysis session (optional)

```
Hi,
I'm trying to normalise my Illumina microarray data (obtained with the Illumina HT-12 v4.0)
For that purpose I have selected the Normalisation option "Illumina - lumi pipeline"
However, the normalisation did not complete successfully.

Any advice to solve this problem ?

Thank you in advance for your precious help.

Best regards

Error message:
in library(chiptype, character.only = T) :
  there is no package called 'Illumina.db'

-----------------------------------------------------------------------

> chipster.common.path = '/opt/chipster/comp/modules/common/R-2.12'
> chipster.module.path = '/opt/chipster/comp/modules/microarray'
> setwd("271661a6-946c-450f-bb21-5d5b5a2837aa")
> probe.identifier <- "Probe_ID"
> transformation <- "log2"
> background.correction <- "none"
> normalize.chips <- "quantile"
> chiptype <- "empty"
> # TOOL norm-illumina-lumi.R: "Illumina - lumi pipeline" (Illumina normalization using
BeadSummaryData files, and using lumi methodology. If you have a BeadSummaryData that reports the
```

CSC

Technical aspects

- **Client-server system**
  - Enough CPU and memory for NGS jobs
  - Centralized maintenance
- **Easy to install**
  - Client uses Java Web Start
  - Server available as a virtual machine



Analysis tool overview

- **140 NGS tools for**
  - RNA-seq
  - miRNA-seq
  - exome/genome-seq
  - ChIP-seq
  - FAIRE-seq
  - MeDIP-seq
  - CNA-seq
  - Metagenomics (16S rRNA)
- **140 microarray tools for**
  - gene expression
  - miRNA expression
  - protein expression
  - aCGH
  - SNP
  - integration of different data

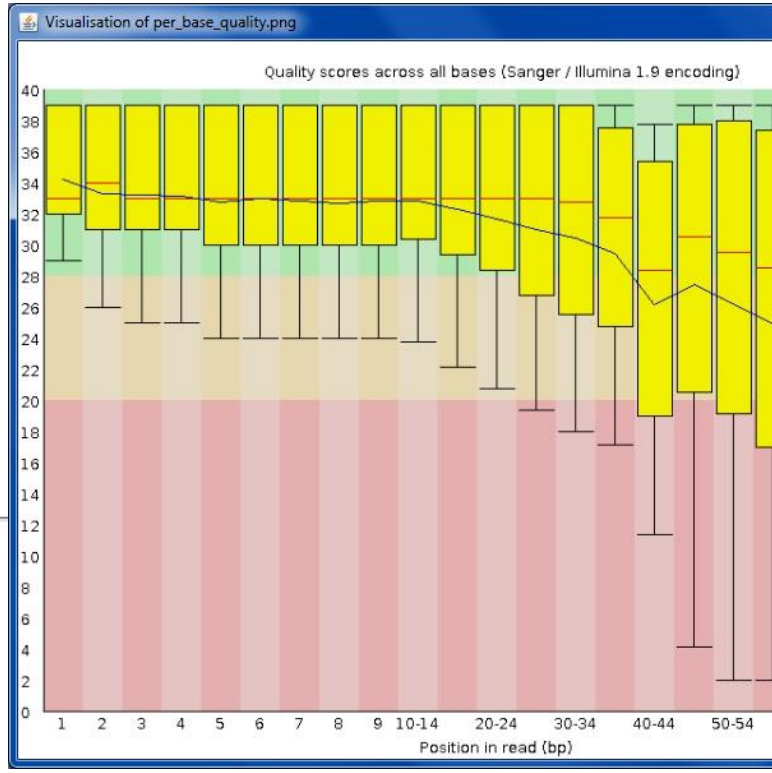
CSC

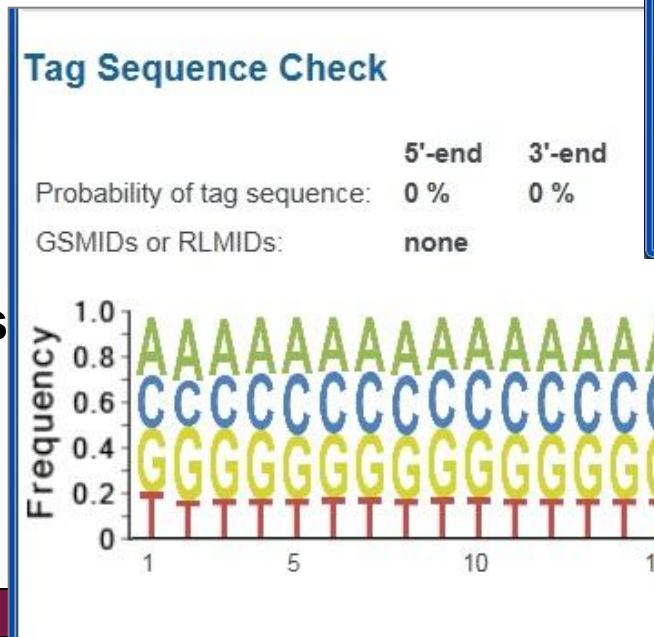
Tools for QC, processing and mapping

- FastQC
- PRINSEQ
- FastX
- TagCleaner

- Bowtie
- TopHat
- BWA

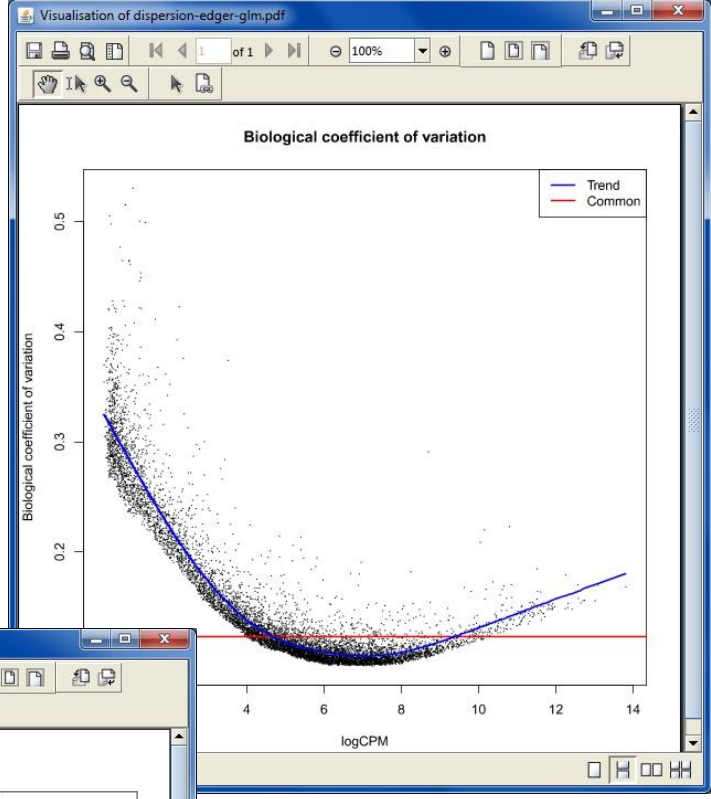
- Picard
- SAMtools
- BEDTools

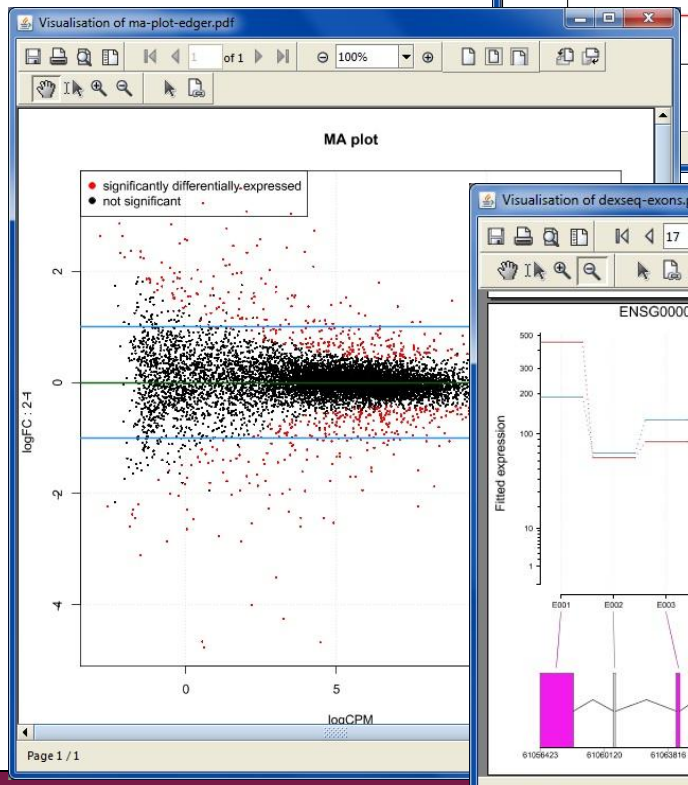


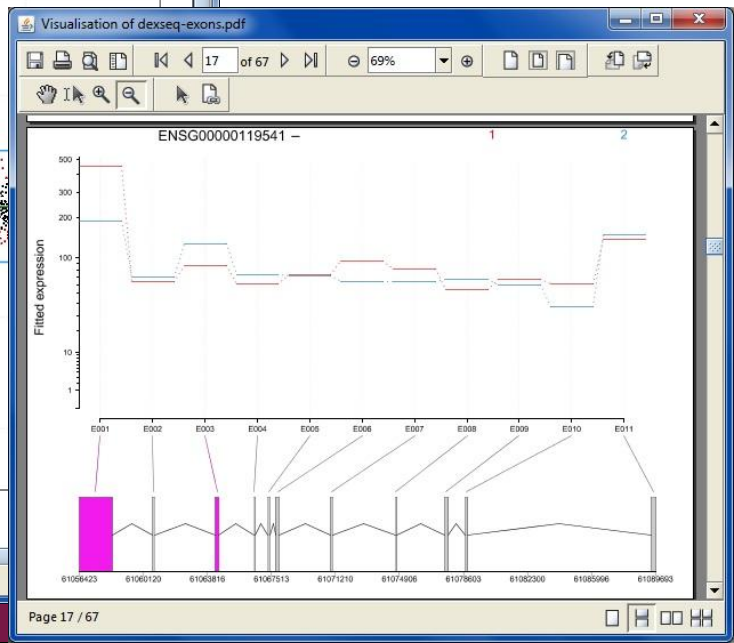


RNA-seq tools

- **Counting**
  - HTSeq
- **Transcript discovery**
  - Cufflinks
- **Differential expression**
  - edgeR
  - DESeq
  - Cuffdiff
  - DEXSeq
- **Pathway analysis**
  - ConsensusPathDB

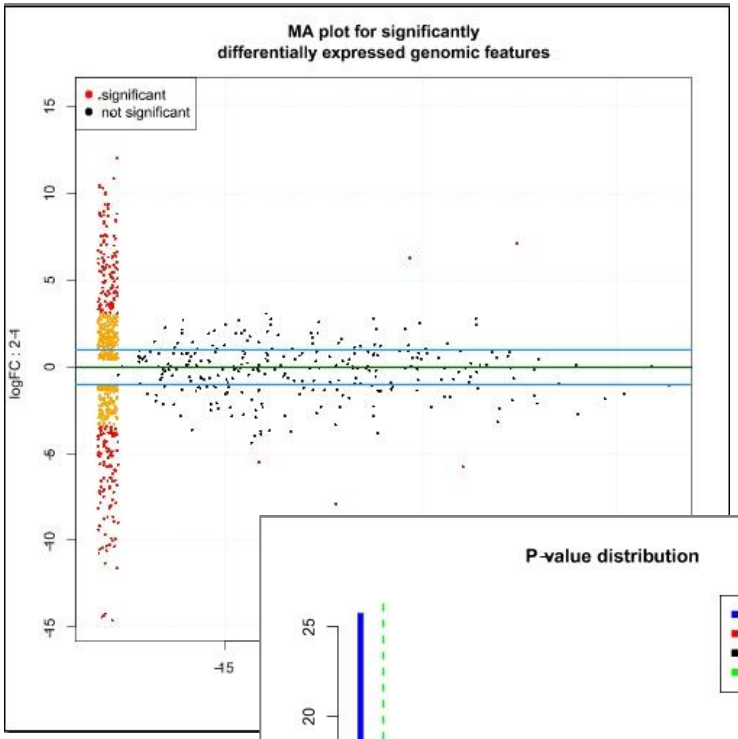


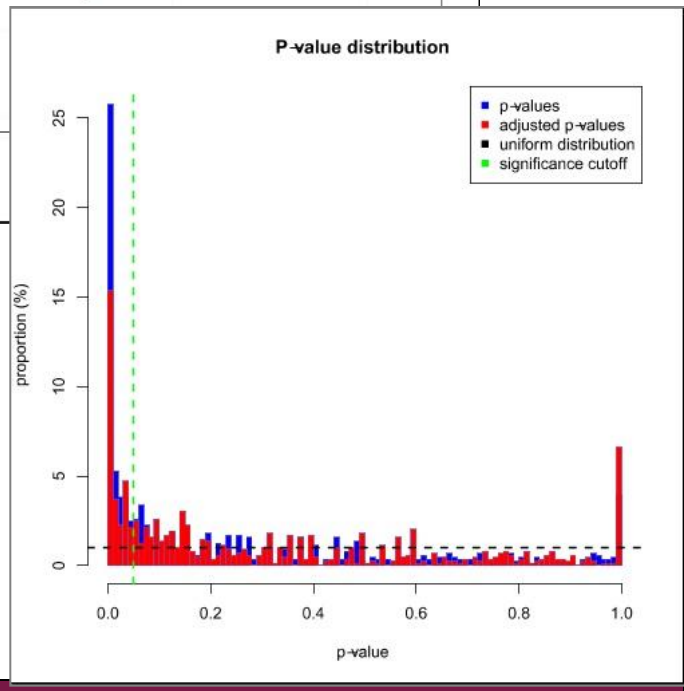




miRNA-seq tools

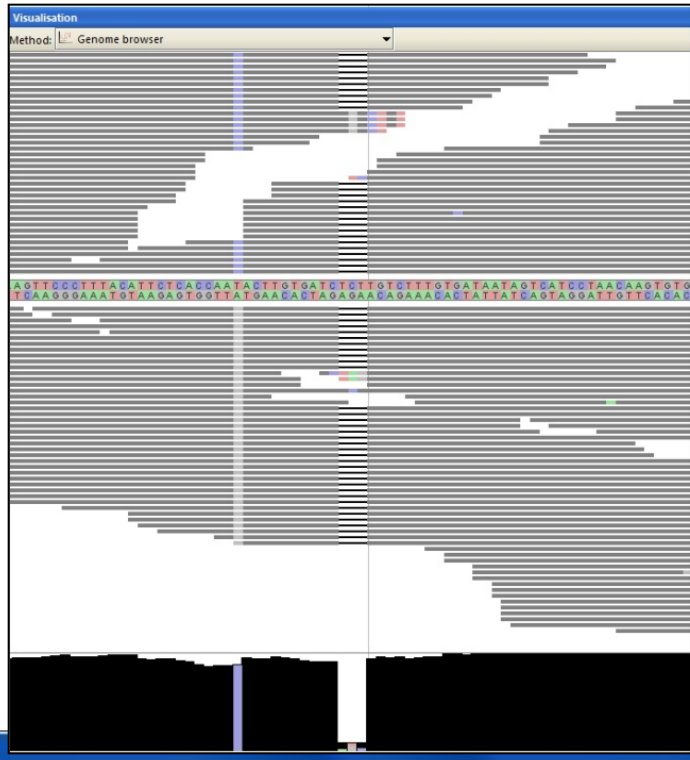
- **Differential expression**
  - edgeR
  - DESeq
- **Retrieve target genes**
  - PicTar
  - miRBase
  - TargetScan
  - miRanda
- **Pathway analysis for targets**
  - GO
  - KEGG
- **Correlate miRNA and target expression**





Exome/genome-seq tools

- **Variant calling**
  - Samtools
- **Variant filtering**
  - VCFtools
- **Variant annotation**
  - AnnotateVariant (Bioconductor)



Visualisation of variants.vcf

Showing 125 rows of 125

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00171	HG00174	NA18486
20	6011756	.	A	G	50.1	.	DP=150;VDB=0.0223;A...	GT:PL:GQ	0/0:0,178,186:99	0/0:0,232,212:99	0/0:0,12,91:17
20	6012323	.	T	C	27.6	.	DP=34;VDB=0.0239;AF...	GT:PL:GQ	0/0:0,42,249:47	0/0:0,42,230:47	0/0:0,3,29:9
20	6014954	.	G	A	999	.	DP=75;VDB=0.0438;AF...	GT:PL:GQ	0/1:69,0,162:80	1/1:206,81,0:86	1/1:154,42,0:47
20	6015419	.	ATGTGT	ATGT	112	.	INDEL;DP=36;VDB=0.0...	GT:PL:GQ	0/0:0,54,255:58	0/0:0,24,255:28	0/0:0,0,0:5
20	6017539	.	C	T	66.6	.	DP=71;VDB=0.0267;AF...	GT:PL:GQ	0/0:0,87,207:89	0/0:0,69,204:71	0/1:74,0,107:72
20	6021948	.	C	T	999	.	DP=106;VDB=0.0392;A...	GT:PL:GQ	0/0:0,15,124:20	0/0:0,24,161:29	0/1:206,0,255:99

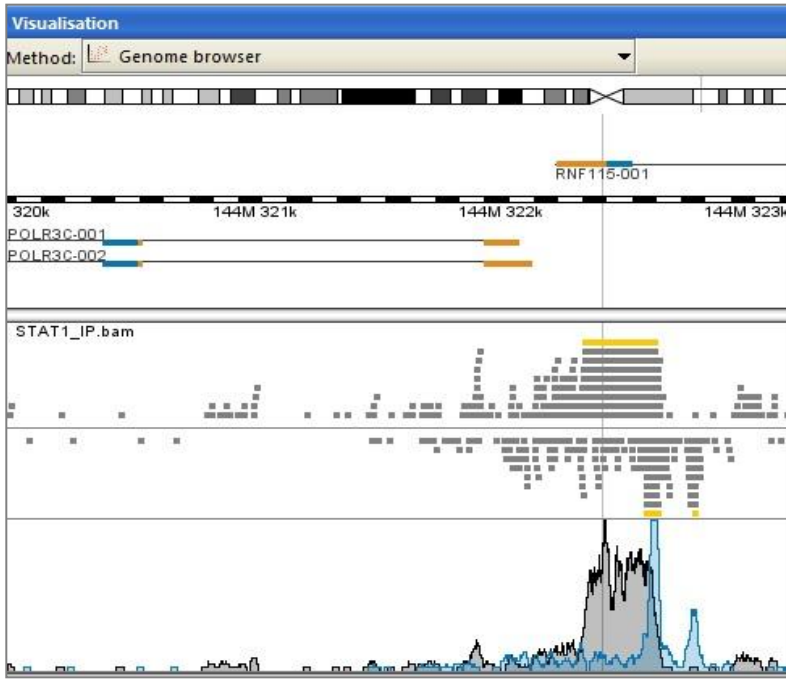
Visualisation of coding-variants.tsv

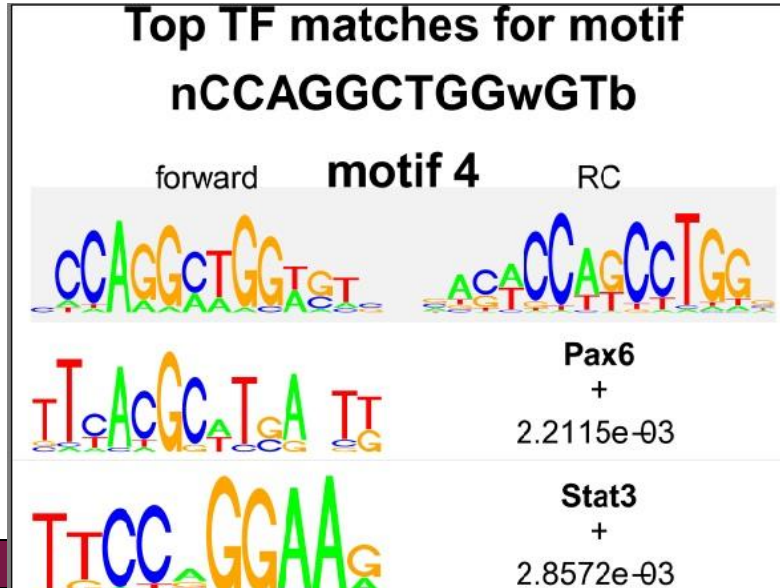
Showing 3 rows of 3

geneID	cdsID	txID	consequence	cdsStart	cdsEnd	width	varAllele	refCodon	varCodon	refAA	varAA	SYMBOL	GENENAME	ENSEMBL
164312	208097	70013	nonsynonymous	421	421	1	G	ACC	GCC	T	A	LRRN4	leucine rich repeat neuronal 4	ENSG00000125872
164312	208097	70014	nonsynonymous	421	421	1	G	ACC	GCC	T	A	LRRN4	leucine rich repeat neuronal 4	ENSG00000125872
650	205075	68975	synonymous	261	261	1	G	TCA	TCG	S	S	BMP2	bone morphogenetic protein 2	ENSG00000125845

ChIP-seq and FAIRE-seq tools

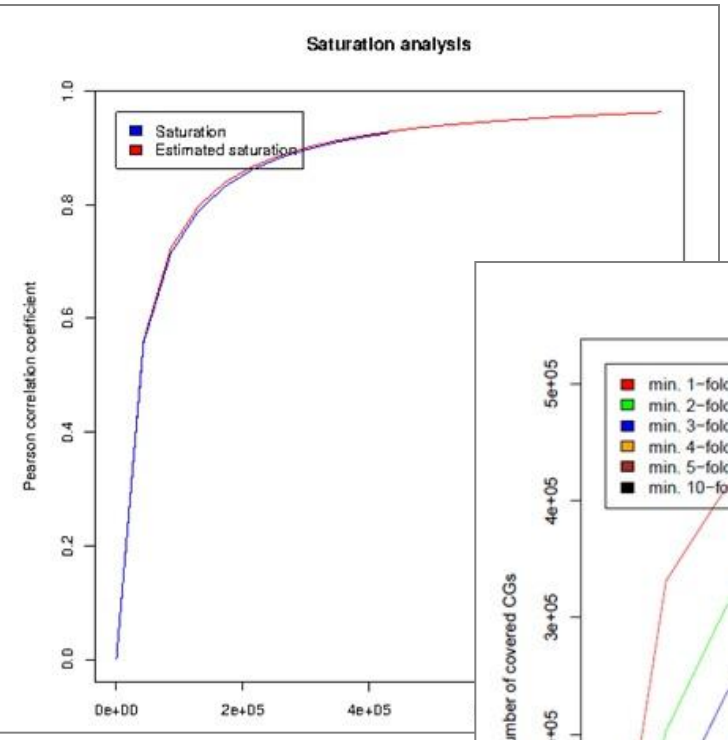
- **Peak detection**
  - MACS
  - F-seq
- **Peak filtering**
  - P-value, no of reads, length
- **Detect motifs, match to JASPAR**
  - MotIV, rGADEM
- **Retrieve nearby genes**
- **Pathway analysis**
  - GO, ConsensusPathDB

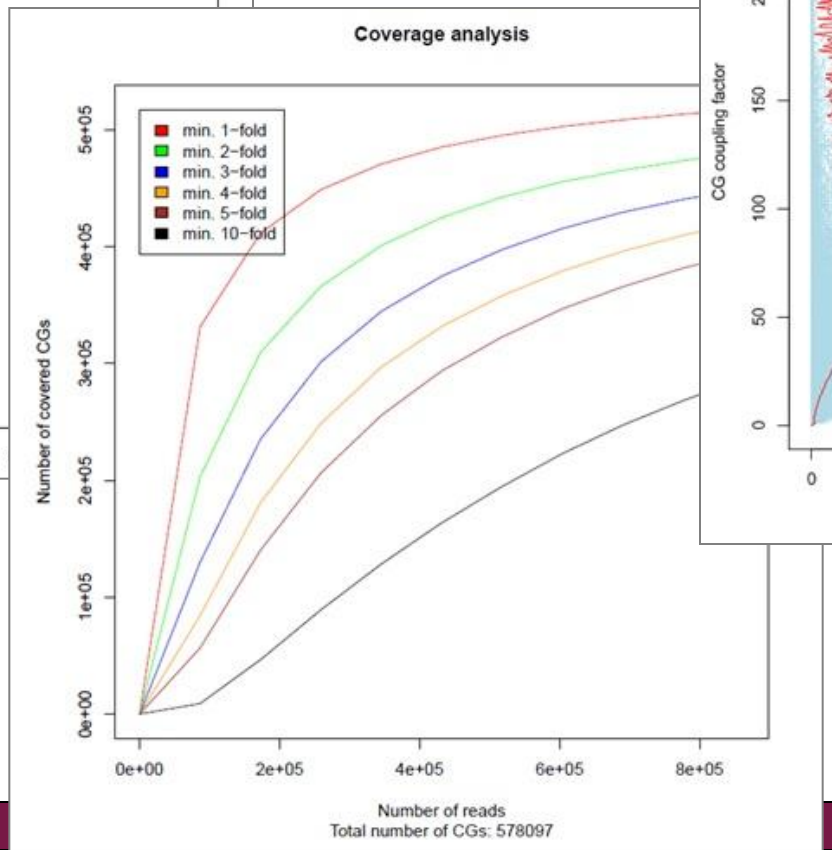


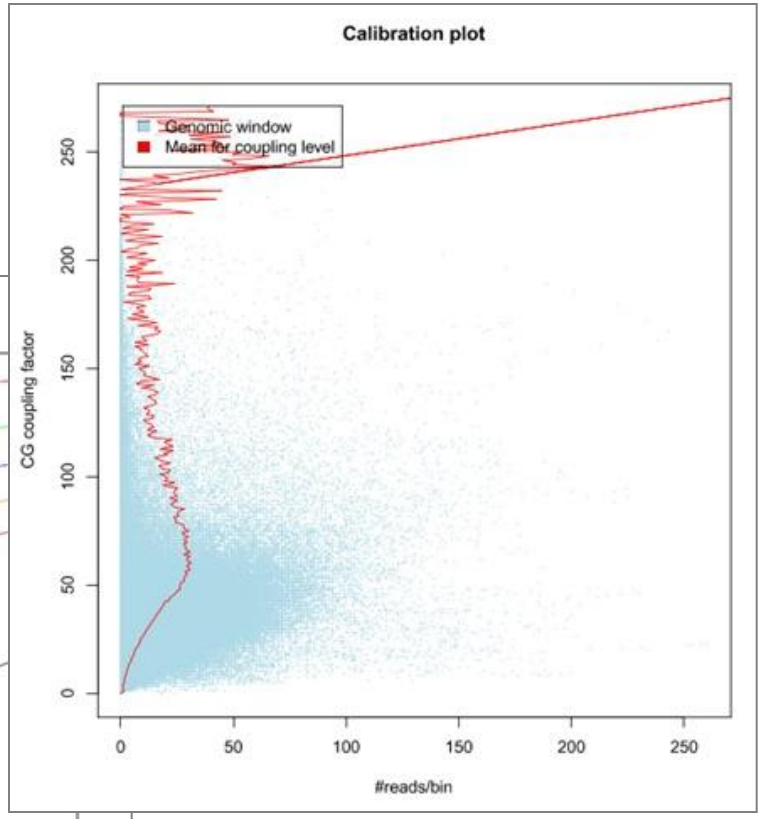


MeDIP-seq tools

- **Detect methylation, compare two conditions**
  - MEDIPS



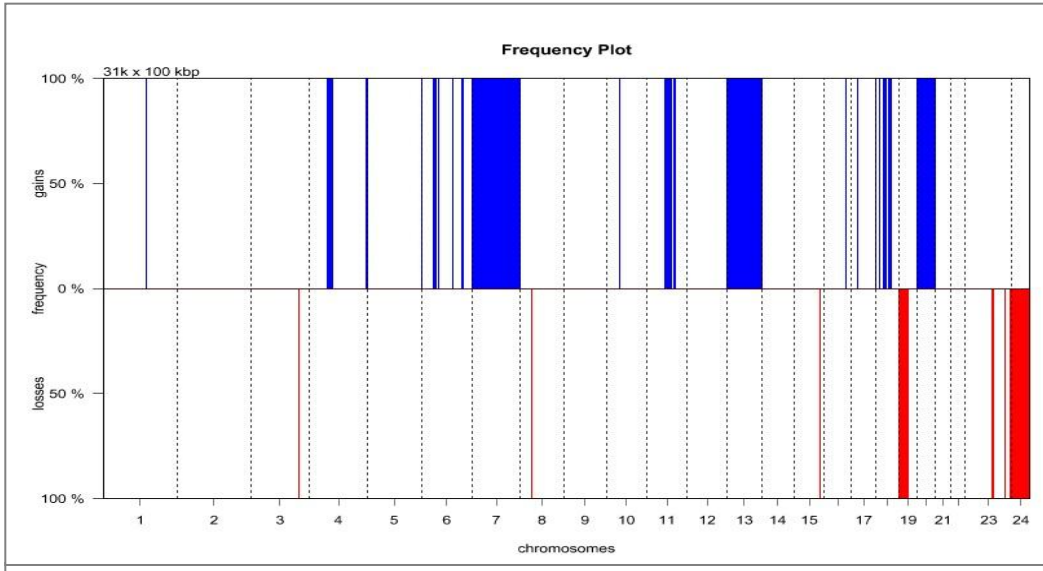


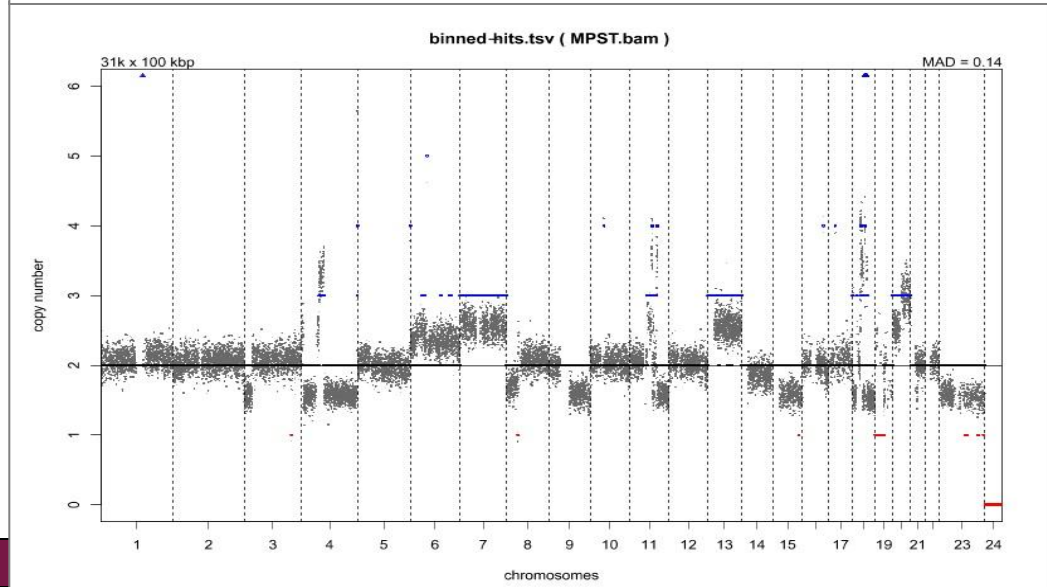


CSC

CNA-seq tools

- **Count reads in bins**
  - Correct for GC content
- **Segment and call CNA**
  - Filter for mappability
  - Plot profiles
- **Group comparisons**
- **Clustering**
- **Detect genes in CNA**
- **GO enrichment**
- **Integrate with expression**



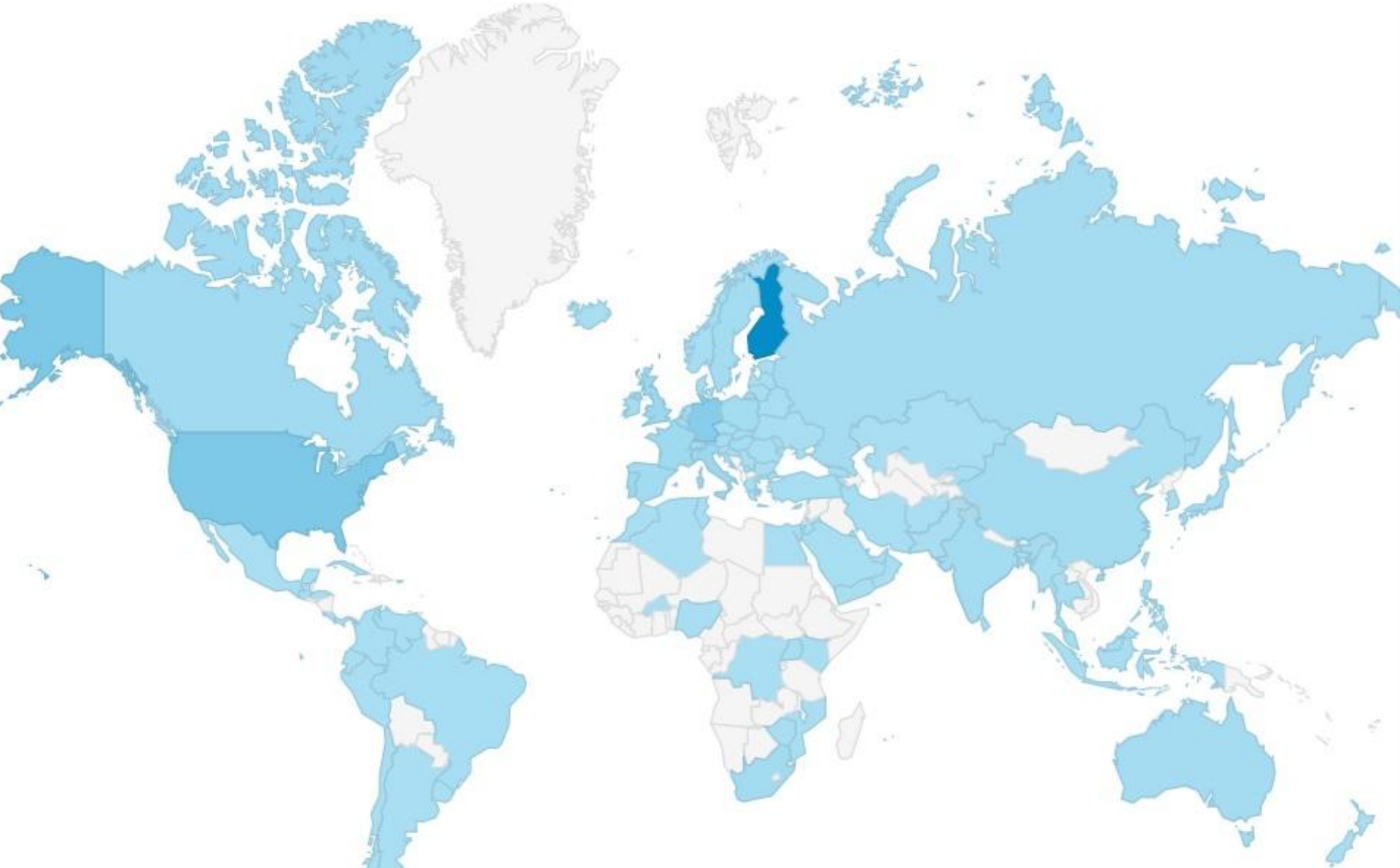


Metagenomics / 16 S rRNA tools

- **Taxonomy assignment with Mothur package**
  - Align reads to 16 S rRNA template
  - Filter alignment for empty columns
  - Keep unique aligned reads
  - Precluster aligned reads
  - Remove chimeric reads
  - Classify reads to taxonomic units
- **Statistical analyses using R**
  - Compare diversity or abundance between groups using several ANOVA-type of analyses

CSC

Acknowledgements to users and contibutors



More info

- **chipster@csc.fi**
- **http://chipster.csc.fi**

GitHub | This repository | Explore | Features | Enterprise | Blog

chipster / **chipster**

Chipster is a user-friendly analysis software for high-throughput data.

5,619 commits | **22** branches | **93** releases | **12** contributors

BMC Genomics

IMPACT FACTOR 4.21

home | journals A-Z | subject areas | advanced search | authors | reviewers | libraries | about | my BioMed Central

Software

Highly accessed | Open Access

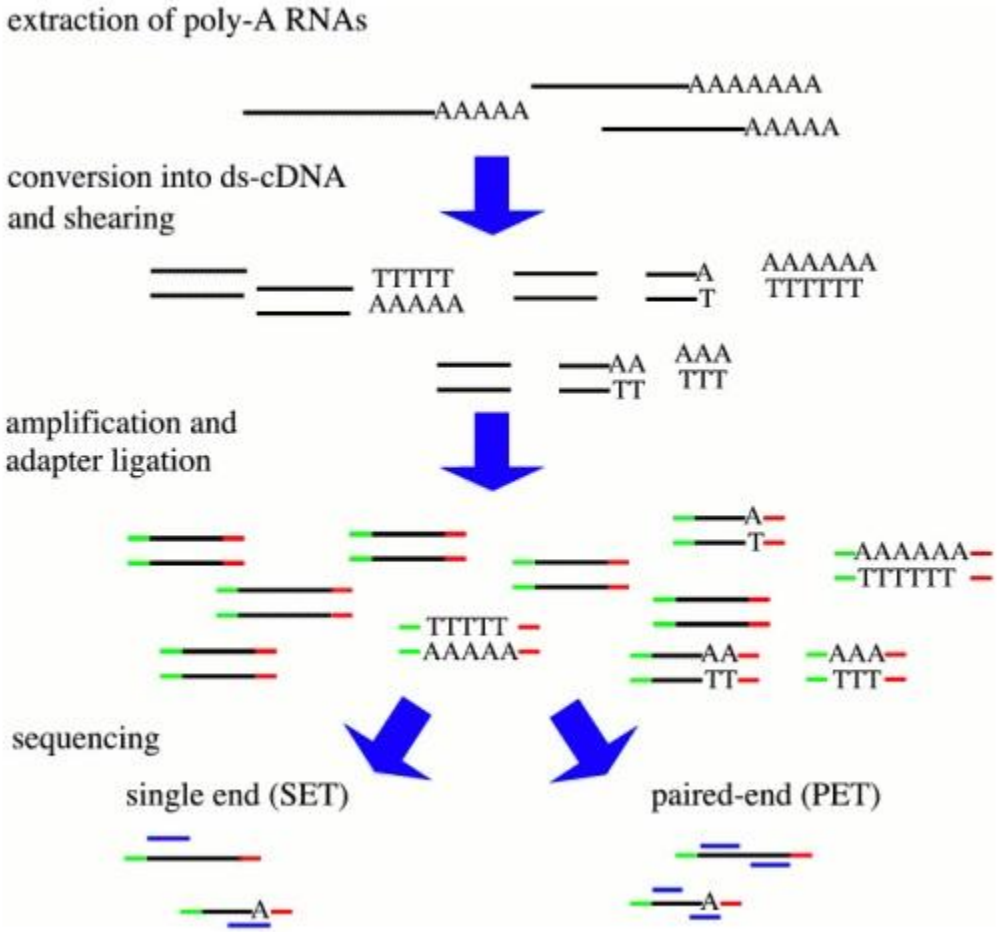
Chipster: user-friendly analysis software for microarray and other high-throughput data

M Aleksi Kallio, Jarno T Tuimala, Taavi Hupponen, Petri Klemela, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Kaki and Eija I Korpelainen

BMC Genomics 2011, **12**:507 doi:10.1186/1471-2164-12-507

Introduction to RNA-seq

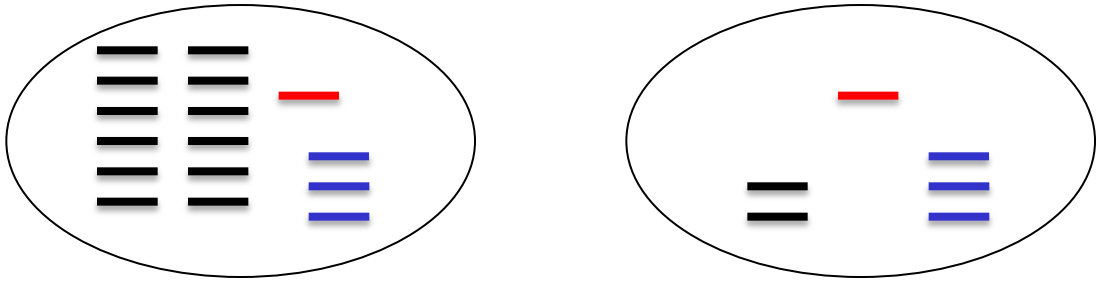
Typical steps in RNA-seq



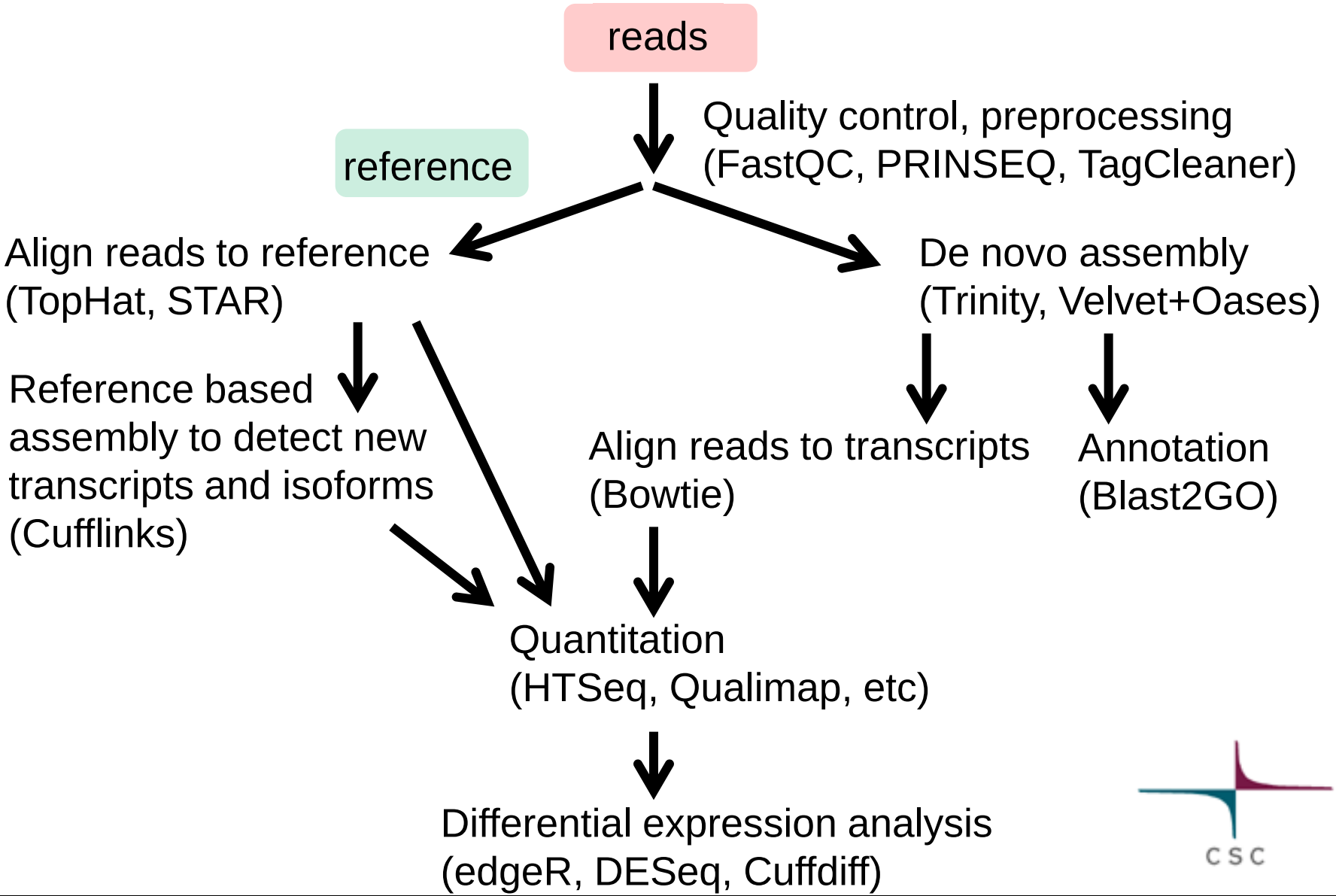
http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/RNA-seq.html

Things to take into account

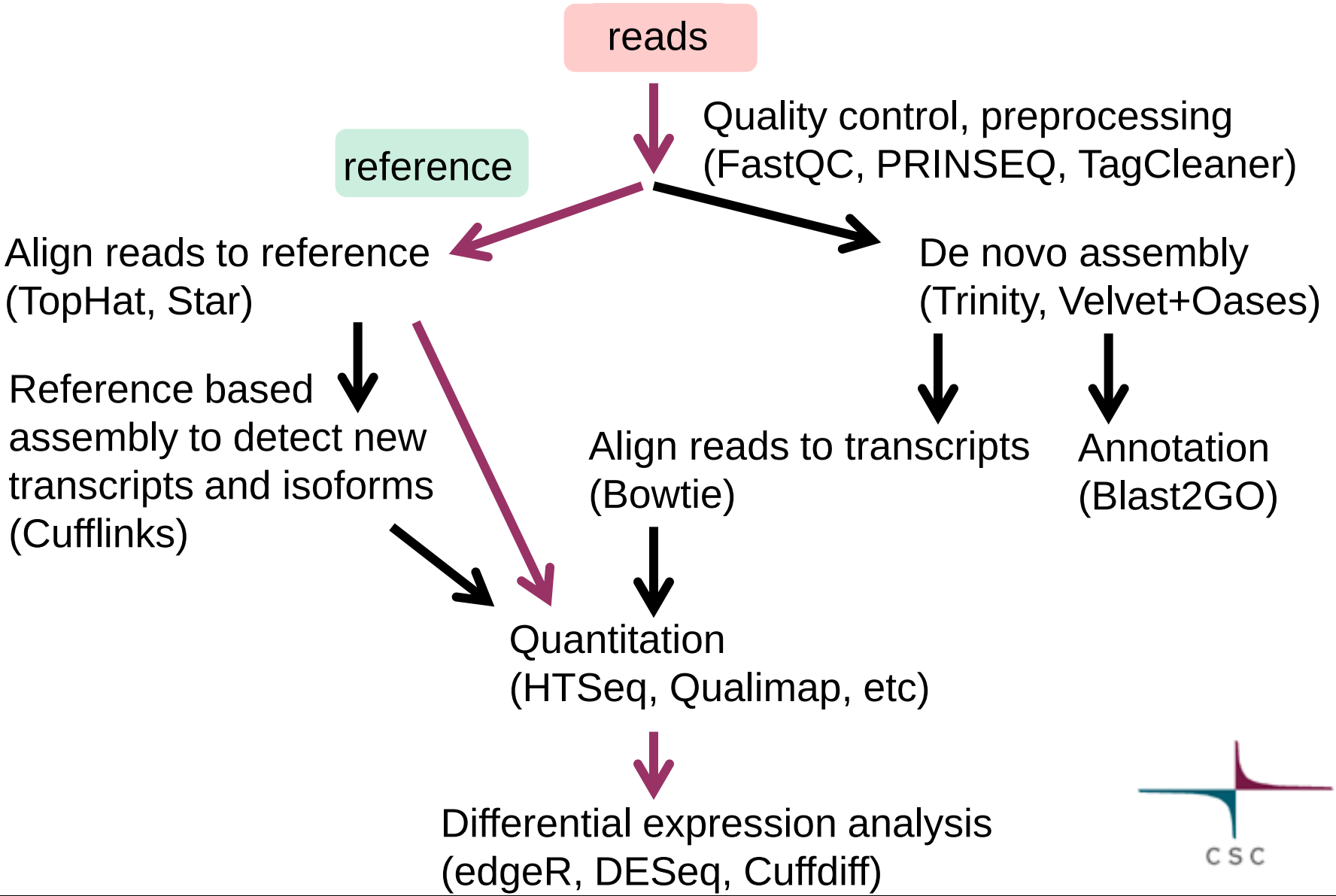
- **Non-uniform coverage along transcripts**
  - Biases introduced in library construction and sequencing
    - polyA capture and polyT priming can cause 3' bias
    - random primers have different binding affinities
    - GC-rich and GC-poor regions can be under-sampled
  - Regions have different mappabilities (uniqueness)
- **Longer transcripts give more counts**
- **RNA composition effect due to sampling:**



RNA-seq data analysis workflow



RNA-seq data analysis today



Quality control of raw reads

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**

CSC

Software packages for quality control

- **FastQC**
- **FastX**
- **PRINSEQ**
- **TagCleaner**
- **Qualimap**
- **...**

Raw reads: FASTQ file format

- **Four lines per read:**
  - Line 1 begins with a '@' character and is followed by a sequence identifier.
  - Line 2 is the sequence.
  - Line 3 begins with a '+' character and can be followed by the sequence identifier.
  - Line 4 encodes the quality values for the sequence, encoded with a single ASCII character for brevity.
  - Example:

    @SEQ_ID
    GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
    +
    !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

- **http://en.wikipedia.org/wiki/FASTQ_format**

CSC

Base qualities

- **If the quality of a base is 30, the probability that it is wrong is 0.001.**
  - Phred quality score Q = -10 * log10 (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
```

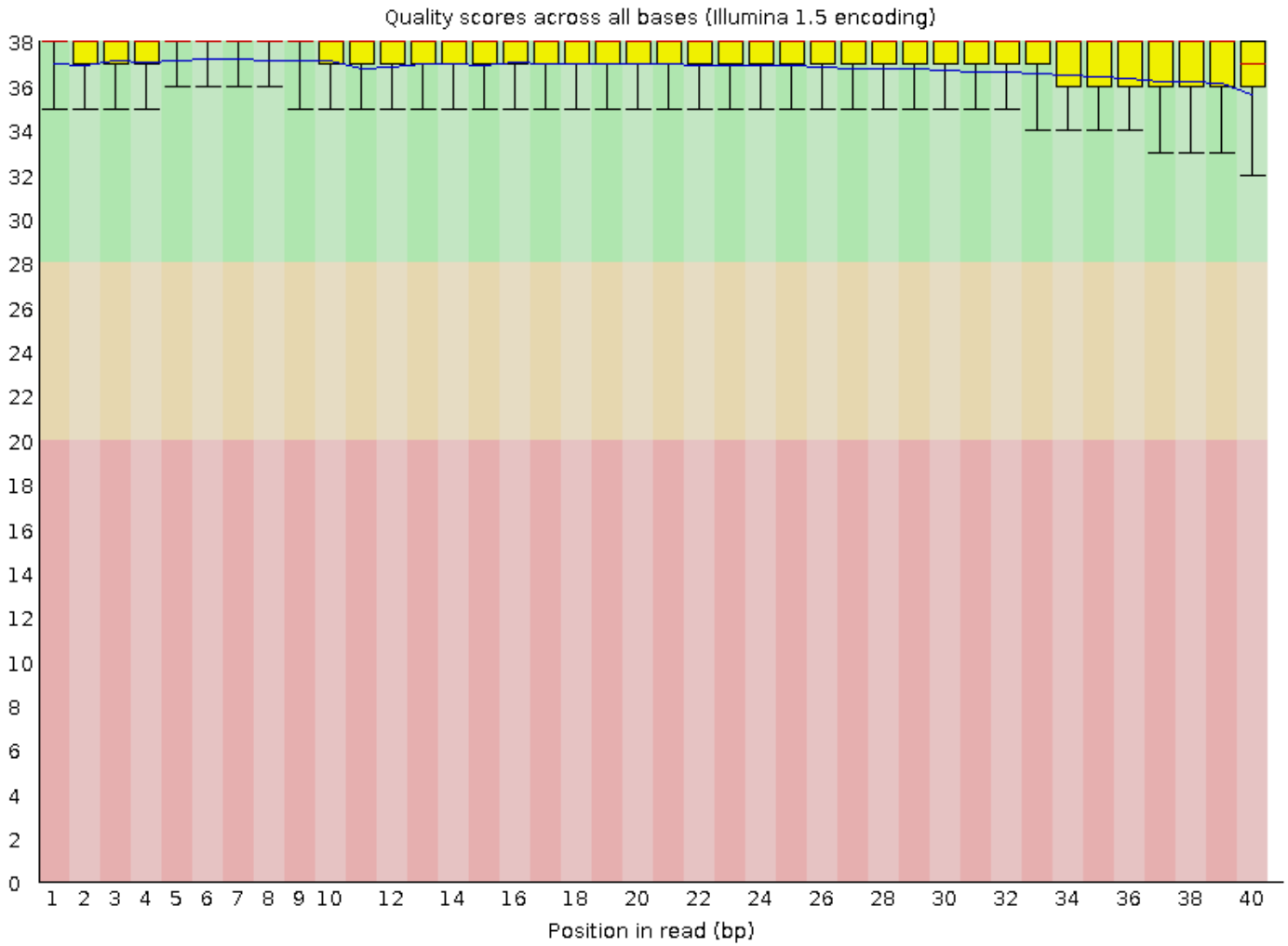
- **Encoded as ASCII characters so that 33 is added to the Phred score**
  - This ”Sanger” encoding is used by Illumina 1.8+, 454 and SOLiD
  - Note that older Illumina data uses different encoding
    - Illumina1.3: add 64 to Phred
    - Illumina 1.5-1.7: add 64 to Phred, ASCII 66 ”B” means that the whole read segment has low quality

Base quality encoding systems

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.......................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmn
 |                         |    |        |                              |
33                        59   64       73                            104
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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

Per position base quality (FastQC)

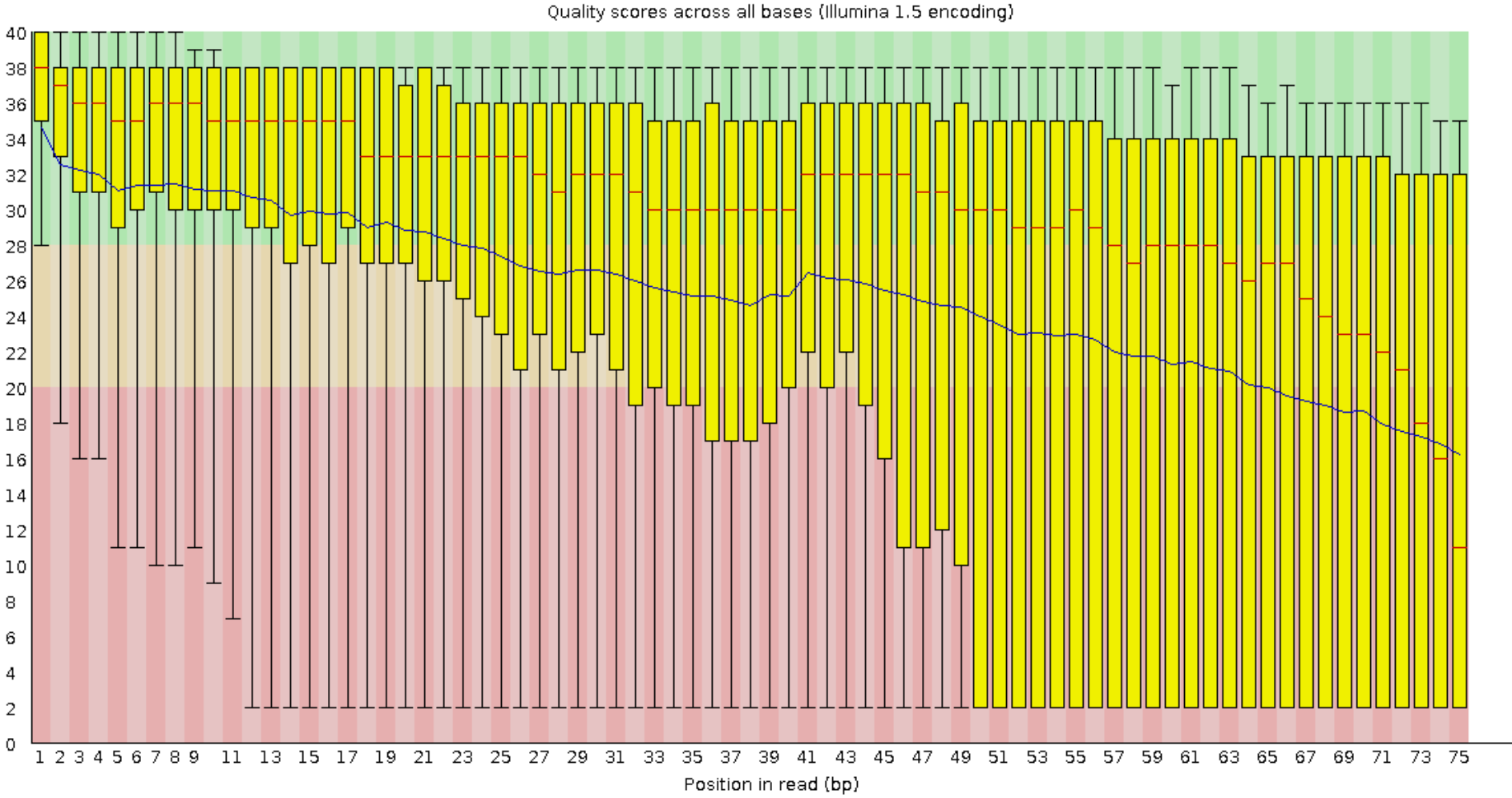


good

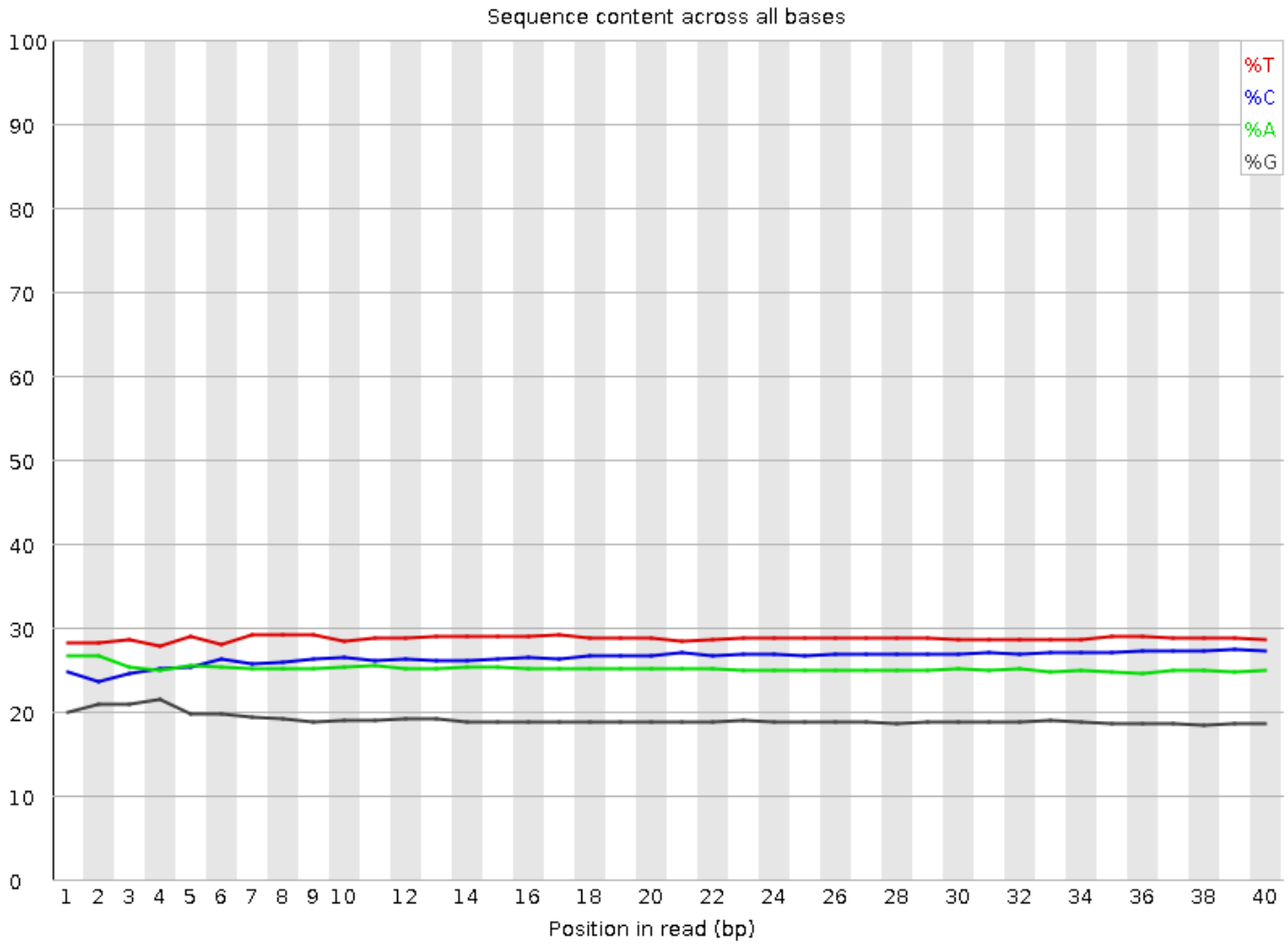
ok

bad

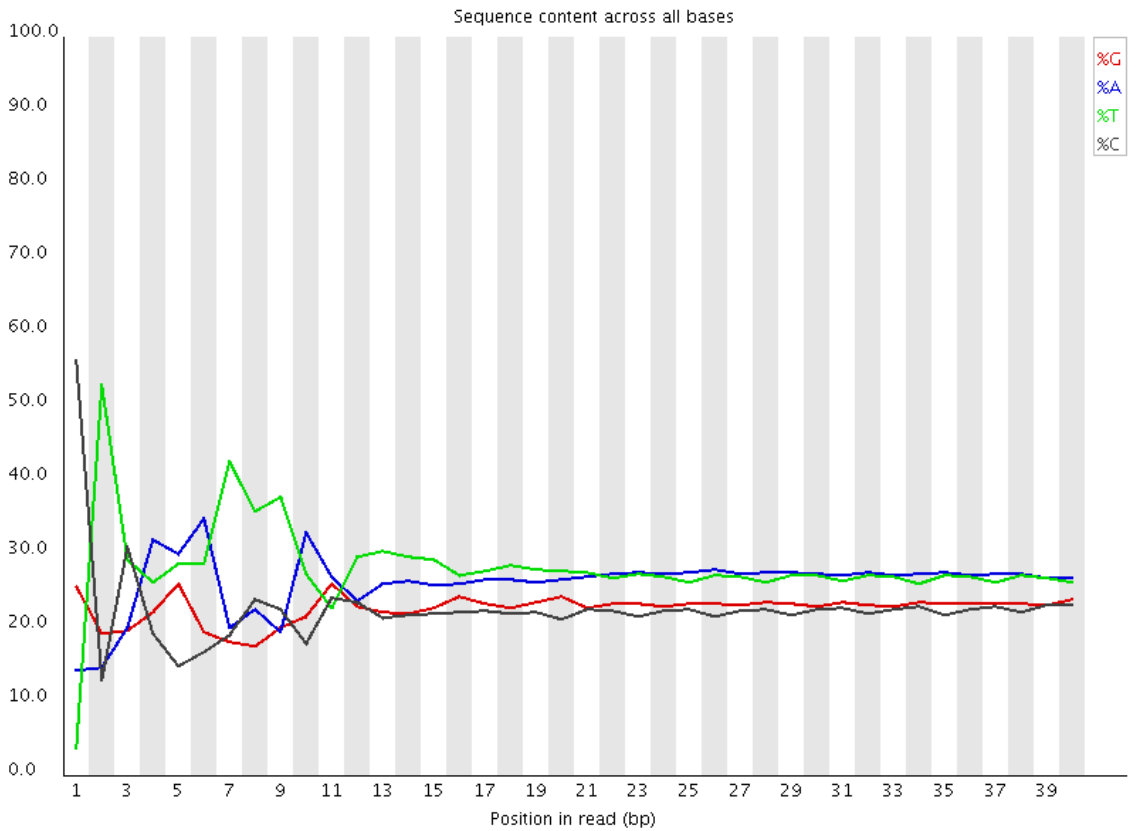
Per position base quality (FastQC)



Per position sequence content (FastQC)



Per position sequence content (FastQC)



- **Sequence specific bias: Correct sequence but biased location, typical for Illumina RNA-seq data**

Preprocessing: Filtering and trimming low quality reads

Software packages for preprocessing

- **FastX**
- **PRINSEQ**
- **TagCleaner**
- **Trimmomatic**
- **Cutadapt**
- **TrimGalore!**
- **...**

PRINSEQ filtering possibilities in Chipster

- **Base quality scores**
  - Minimum quality score per base
  - Mean read quality
- **Ambiguous bases**
  - Maximum count/ percentage of Ns that a read is allowed to have
- **Low complexity**
  - DUST (score > 7), entropy (score < 70)
- **Length**
  - Minimum length of a read
- **Duplicates**
  - Exact, reverse complement, or 5'/3' duplicates
- **Tool "Filter for several criteria"**
  - Combines all above and <u>copes with paired end data</u>

CSC

PRINSEQ trimming possibilities in Chipster

- **Trim based on quality scores**
  - Minimum quality, look one base at a time
  - Minimum (mean) quality in a sliding window
  - From 3’ or 5’ end
- **Trim polyA/T tails**
  - Minimum number of A/Ts
  - From left or right
- **Trim based on several criteria**
  - Trim x bases from left/ right
  - Trim to length x
  - All above and <u>copes with paired end data</u>

CSC

Data

- **Human data for 2 cell lines (h1-hESC and GM12878) from the ENCODE project**
  - 76 b single-end reads, no replicates

Aligning (=mapping) reads to reference

Alignment to reference genome/transcriptome

- **Goal is to find out where a read originated from**
  - Challenge: variants, sequencing errors, repetitive sequence
- **Mapping to**
  - transcriptome allows you to count hits to known transcripts
  - genome allows you to find new genes and transcripts
- **Many organisms have introns, so RNA-seq reads map to genome non-contiguously → spliced alignments needed**
  - Difficult because sequence signals at splice sites are limited and introns can be thousands of bases long

Splice-aware aligners

- **TopHat (uses Bowtie)**
- **STAR**
- **GSNAP**
- **RUM**
- **MapSplice**
- **...**

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström[1,13], Tamara Steijger[1], Botond Sipos[1], Gregory R Grant[2,3], André Kahles[4,5], The RGASP Consortium[6], Gunnar Rätsch[4,5], Nick Goldman[1], Tim J Hubbard[7], Jennifer Harrow[7], Roderic Guigó[8,9] & Paul Bertone[1,10–12]

Nature methods 2013 (10:1185)

CSC

Mapping quality

- **Confidence in read's point of origin**
- **Depends on many things, including**
  - length of alignment
  - number of mismatches and gaps
  - uniqueness of the aligned region in the genome
- **Expressed in Phred scores, like base qualities**
  - Q = -10 * log10 (probability that read was mapped to a wrong location)

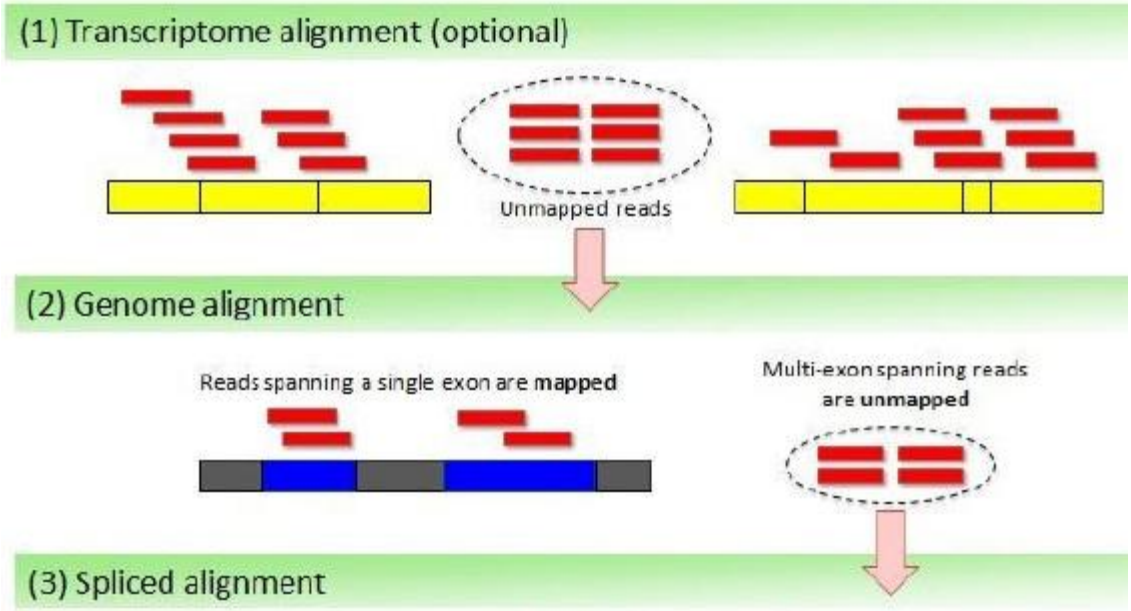
Bowtie2

- **Fast and memory efficient aligner**
- **Can make gapped alignments (= can handle indels)**
- **Cannot make spliced alignments but is used by TopHat2 which can**
- **Two alignment modes:**
  - End-to-end
    - Read is aligned over its entire length
    - Maximum alignment score = 0, deduct penalty for each mismatch (less for low quality base), N, gap opening and gap extension
  - Local
    - Read ends don’t need to align, if this maximizes the alignment score
    - Add bonus to alignment score for each match
- **Reference (genome) is indexed to speed up the alignment process**

CSC

TopHat2

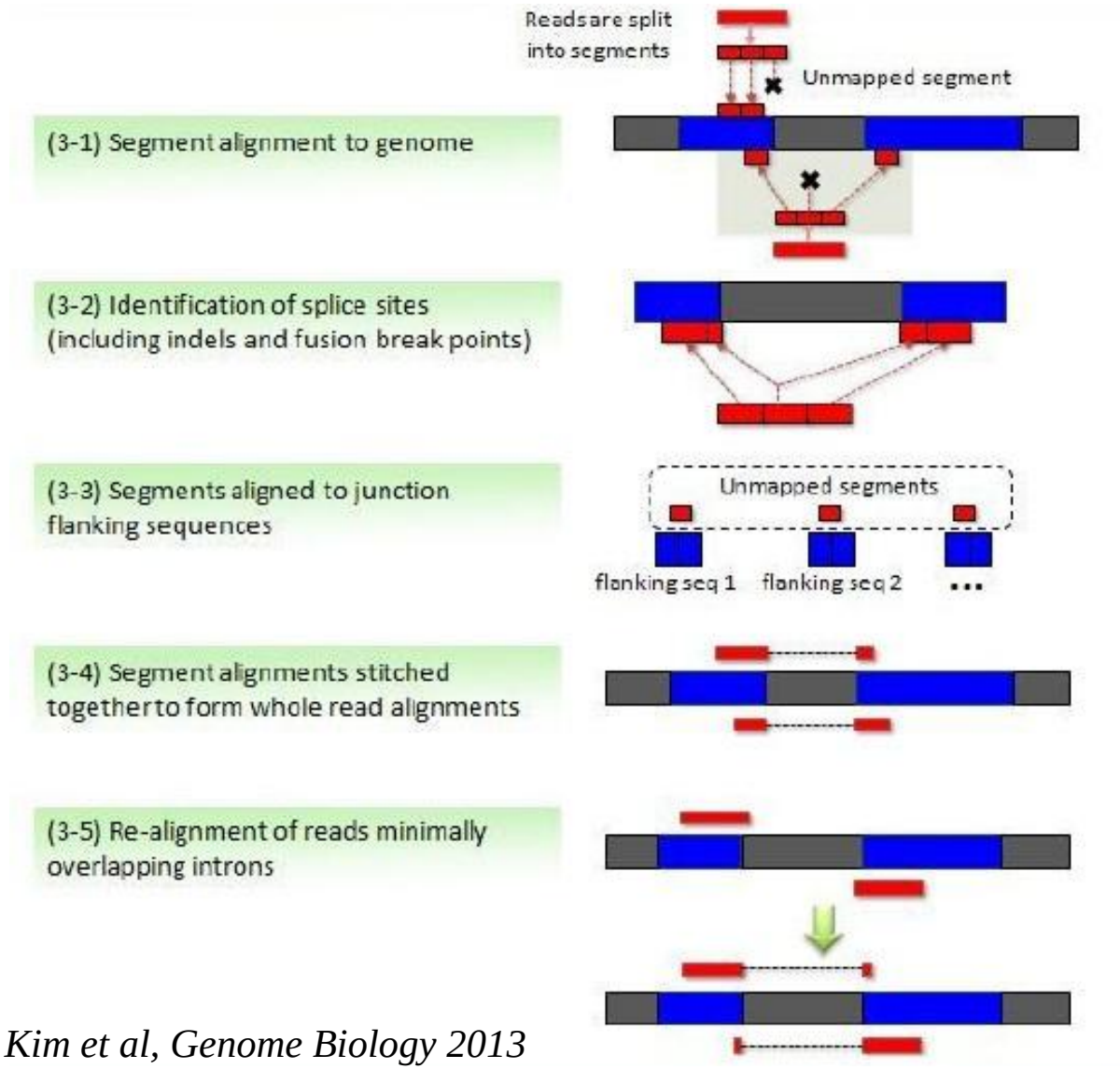
- **Relatively fast and memory efficient spliced aligner**
- **Performs several alignment steps**
- **Uses Bowtie2 end-to-end mode for aligning**
  - Low tolerance for mismatches
- **If annotation (GTF file) is available, builds a virtual transcriptome and aligns reads to that first**



Kim et al, Genome Biology 2013

CSC

TopHat2 spliced alignment steps



Kim et al, Genome Biology 2013

CSC

Analysis tools - Alignment - TopHat2 for paired end reads

Genome	Human genome (hg...
Use annotation GTF	yes
When GTF file is used, ignore novel junctions	yes
Base quality encoding used	Sanger - Phred+33
Expected inner distance between mate pairs	200
Standard deviation for the inner distances between mate pairs	20
How many hits is a read allowed to have	20
Number of mismatches allowed in final alignment	2
Minimum anchor length	8
Maximum number of mismatches allowed in the anchor	0
Minimum intron length	70
Maximum intron length	500000

CSC

File format for aligned reads: BAM/SAM

- SAM (Sequence Alignment/Map) is a tab-delimited text file. BAM is a binary form of SAM.
- Optional header (lines starting with @)
- One line for each alignment, with 11 mandatory fields:
  - read name, flag, reference name, position, mapping quality, CIGAR, mate name, mate position, fragment length, sequence, base qualities
  - CIGAR reports match (M), insertion (I), deletion (D), intron (N), etc
- 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
```

Visualisation

Method: BAM viewer Redraw Restore

```
@HD	VN:1.4	SO:coordinate
@SQ	SN:chr1	LN:249250621
@SQ	SN:chr10	LN:135534747
@SQ	SN:chr11	LN:135006516
@SQ	SN:chr12	LN:133851895
@SQ	SN:chr13	LN:115169878
@SQ	SN:chr14	LN:107349540
@SQ	SN:chr15	LN:102531392
@SQ	SN:chr16	LN:90354753
@SQ	SN:chr17	LN:81195210
@SQ	SN:chr18	LN:78077248
@SQ	SN:chr19	LN:59128983
@SQ	SN:chr2	LN:243199373
@SQ	SN:chr20	LN:63025520
@SQ	SN:chr21	LN:48129895
@SQ	SN:chr22	LN:51304566
@SQ	SN:chr3	LN:198022430
@SQ	SN:chr4	LN:191154276
@SQ	SN:chr5	LN:180915260
@SQ	SN:chr6	LN:171115067
@SQ	SN:chr7	LN:159138663
@SQ	SN:chr8	LN:146364022
@SQ	SN:chr9	LN:141213431
@SQ	SN:chrM	LN:16571
@SQ	SN:chrX	LN:155270560
@SQ	SN:chrY	LN:59373566
@PG	ID:TopHat	VN:2.0.9	CL:/opt/chipster/tools/tophat2/tophat -p 2 --read-mismatches 2 -a 8 -m 0 -i 70 -I 500000 -g 20 --library-type fr-unstranded
--transcriptome-index=/opt/chipster/tools/bowtie2/indexes/hg19.ti --no-novel-juncs /opt/chipster/tools/bowtie2/indexes/hg19 reads1.fq
HWI-EAS229_1:4:82:1371:1147	272	chr1	18378	1	2M6358N73M	*	0	0
TCCTGCTGAAGATGTCTCCAGAGACCTTCTGCAGGTACTGAAGGGCATCCGCCATCTGCTGGACGGCCTCCTCTC	5661525416816488666(6(6(6261?8==(B=513);(/BB=141=>6>?=<=?B>9B?>BA<66>BA>BBB
CC:Z:chr15	MD:Z:40C34	XG:i:0	NH:i:3	HI:i:0	NM:i:1	XM:i:1	XN:i:0	XO:i:0	CP:i:102506354	AS:i:0	XS:A:-	YT:Z:UU
```

CSC

BAM file (.bam) and index file (.bai)

- **BAM files can be sorted by chromosomal coordinates and indexed for efficient retrieval of reads for a given region.**
- **The index file must have a matching name. (e.g. reads.bam and reads.bam.bai)**
- **Genome browser requires both BAM and the index file.**
- **The alignment tools in Chipster automatically produce sorted and indexed BAMs.**
- **When you import BAM files, Chipster asks if you would like to preproces them (convert SAM to BAM, sort and index BAM).**

CSC

Manipulating BAM files (SAMtools, Picard)

- **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.
- **Index BAM**
- **Statistics for BAM**
  - How many reads align to the different chromosomes.
- **Count alignments in BAM**
  - How many alignments does the BAM contain.
  - Includes an optional mapping quality filter.
- **Retrieve alignments for a given chromosome/region**
  - Makes a subset of BAM, e.g. chr1:100-1000, inc quality filter.
- **Create consensus sequence from BAM**

CSC

Region file formats: BED

- **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
chr19	325637	326573	JUNC00000014	55

CSC

Region file formats: GFF/GTF

- **9 obligatory columns: chr, source, name, start, end, score, strand, frame, attribute**
- **1-based**

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";

CSC

Quality control of aligned reads

Quality metrics for aligned reads

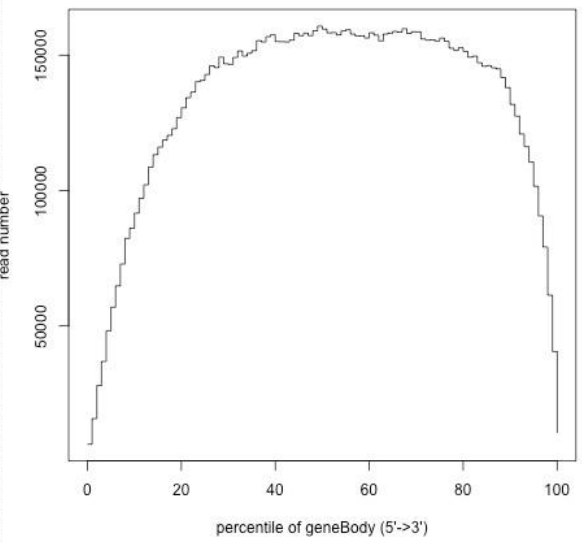
- **How many reads mapped and how many mapped uniquely?**
- **How many pairs mapped, how many mapped concordantly, and what proportion of pairs map to identical location?**
- **Mapping quality distribution?**

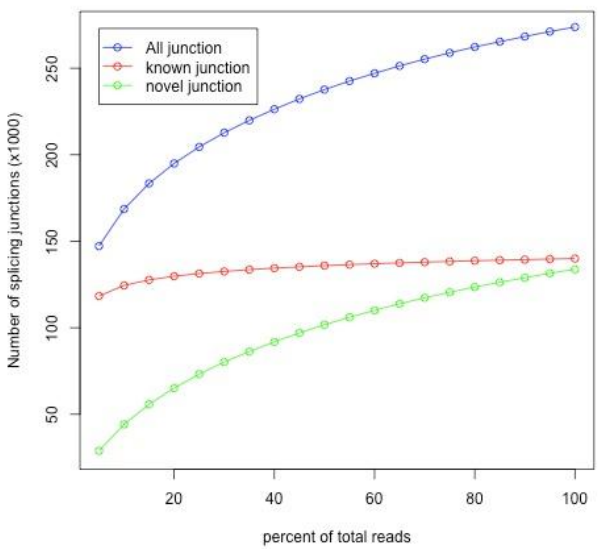
- **Saturation of sequencing depth**
  - Would more sequencing detect more genes and splice junctions?
- **Read distribution between different genomic features**
  - Exonic, intronic, intergenic regions
  - Coding, 3’ and 5’ UTR exons
  - Protein coding genes, pseudogenes, rRNA, miRNA, etc
- **Coverage uniformity along transcripts**

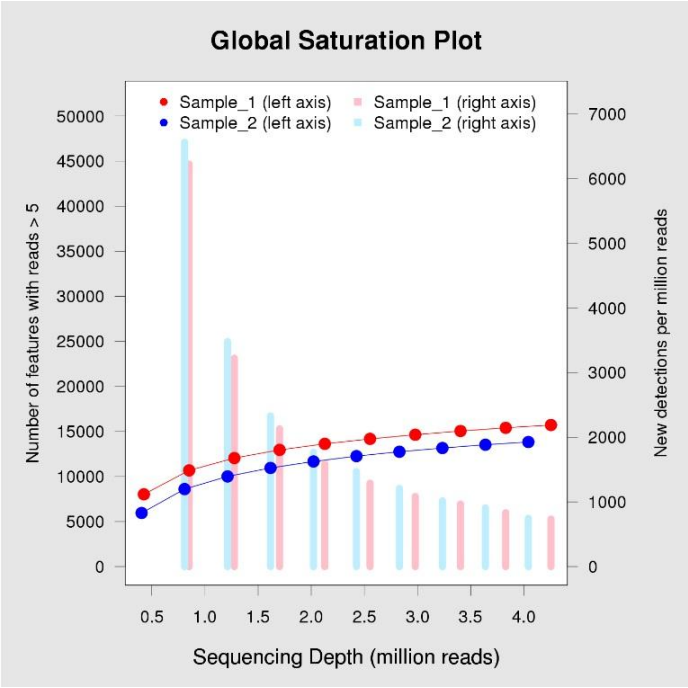
CSC

Quality control programs for aligned reads

- **RseQC (soon in Chipster)**
- **RNA-seqQC**
- **Qualimap**
- **Picards's CollectRnaSeqMetrics**







Visualization of reads and results in genomic context

Software packages for visualization

- **Chipster genome browser**
- **IGV**
- **UCSC genome browser**
- **....**

- **Differences in memory consumption, interactivity, annotations, navigation,...**

CSC

Chipster Genome Browser

- **Integrated with Chipster analysis environment**
- **Automatic sorting and indexing of BAM and BED files**
- **Automatic coverage calculation (total and strand-specific)**
- **Zoom in to nucleotide level**
- **Highlight variants**
- **Jump to locations using BED and tsv files**
- **View details of selected BED features**
- **Several views (reads, coverage profile, density graph)**

CSC

Chipster 2.9.0 (build 1418)

File Edit View Workflow Help

Visualisation

Method: Genome browser

Help Restore Detach

Annotations

Show all

40M 850k 40M 860k

SYT4-001
SYT4-002
SYT4-004
SYT4-003
SYT4-012
SYT4-006
SYT4-005
SYT4-013

Gm12892_1_chr18.bam

Gm12892_2_chr18.bam

Gm12892_3_chr18.bam

hESC1_chr18.bam

hESC2_chr18.bam

hESC3_chr18.bam

hESC4_chr18.bam

de-list-edger.bed

Show score

Settings Selected Legend

Genome

Human hg19 (GRCh37.70)

Location

Chromosome

18

Location (gene or position)

40853532

View size

23 kb

Go

Options

Reads

Highlight SNPs

Density graph

Low complexity regions

Coverage type

total

Coverage scale

50

External links

View this region in Ensembl or UCSC genome browser.

Connected to chipster.csc.fi

Ready

206M / 870M

Chipster 2.9.0 (build 1418)

File Edit View Workflow Help

Visualisation

Method: Genome browser

Help Restore Detach

Annotations

Show all

40M 850k

SYT4-001
SYT4-002
SYT4-004
SYT4-003
SYT4-012
SYT4-006
SYT4-013
SYT4-005

Gm12892_1_chr18.bam

Gm12892_2_chr18.bam

Gm12892_3_chr18.bam

hESC1_chr18.bam

hESC2_chr18.bam

hESC3_chr18.bam

hESC4_chr18.bam

de-list-edger.bed

Show score

26
13
0
-13

Settings \ Selected \ Legend \

Genome

Human hg19 (GRCh37.70)

Location

Chromosome

18

Location (gene or position)

40852176

View size

15 kb

Go

Options

Reads

Highlight SNPs

Density graph

Low complexity regions

Coverage type

none

Coverage scale

50

External links

View this region in Ensembl or UCSC genome browser.

Connected to chipster.csc.fi

Ready

263M / 870M

Chipster 2.9.0 (build 1418)

File Edit View Workflow Help

Visualisation

Method: Genome browser

Help Restore Detach

Annotations

Show all

40M 847k 900

SYT4-001

Gm12892_1_chr18.bam

Gm12892_2_chr18.bam

Gm12892_3_chr18.bam

hESC1_chr18.bam

hESC2_chr18.bam

hESC3_chr18.bam

hESC4_chr18.bam

Show all

GAAACGGGTTGGGCAAAAGCAGCAATTTATTCATGTCTTACAAACATGGTCTTACAAACATCATACAGAGCTGGCACAAAAGAACAGACACTGTCTCACAC

CTTTGCCCAACCCGTTTTCGTCGTTAAATAAGTACAGAATGTTTGTACCAGAATGTTTGTAGTATGTCTCGACCGTGTTTTCTTGTCTGTGACAGAGTGTG

de-list-edger.bed

Settings Selected Legend

Genome

Human hg19 (GRCh37.70)

Location

Chromosome

18

Location (gene or position)

40847900

View size

104

Go

Options

Reads

Highlight SNPs

Density graph

Low complexity regions

Coverage type

total

Coverage scale

50

External links

View this region in Ensembl or UCSC genome browser.

Connected to chipster.csc.fi

Ready

276M / 870M

Quantitation

Software for counting aligned reads per genomic features (genes/exons/transcripts)

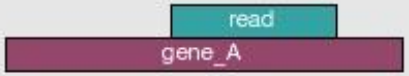
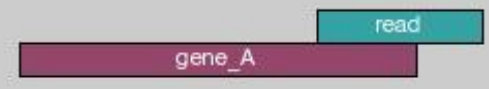
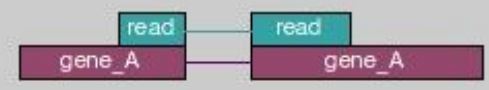
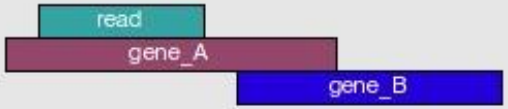
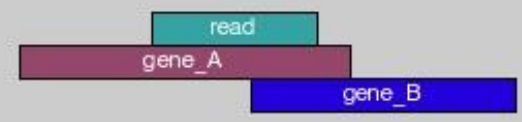
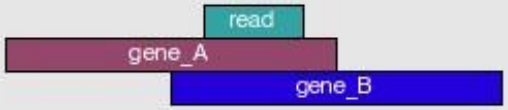
- **HTSeq**
- **Cuffdiff**
- **BEDTools**
- **Qualimap**
- **...**

HTSeq count

- **Given a BAM file and a list of genomic features, counts how many reads map to each feature.**
  - For RNA-seq the features are typically genes, where each gene is considered as the union of all its exons.
  - Also exons can be considered as features, e.g., in order to check for alternative splicing.
- **Features need to be supplied in GTF file**
  - Note that GTF and BAM must use the same chromosome naming
- **3 modes to handle reads which overlap several genes**
  - Union (default)
  - Intersection-strict
  - Intersection-nonempty

CSC

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

CSC

Differential expression analysis

Things to take into account

- **Normalization is required in order to compare expression between samples**
  - Different library sizes
  - RNA composition bias caused by sampling approach
- **Model has to account for overdispersion in biological replicates → negative binomial distribution**
- **Raw counts are needed to assess measurement precision**
  - Units of evidence for expression
- **Multiple testing problem**

CSC

Software packages for DE analysis

- **edgeR**
- **DESeq**
- **DEXSeq**
- **Cuffdiff**
- **BaySeq**
- **SAMseq**
- **NOIseq**
- **Limma + voom, limma + vst**
- **...**

Briefings in Bioinformatics Advance Access published December 2, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 12 doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Comprehensive evaluation of differential expression analysis methods for RNA-seq data

Franck Rapaport [1], Raya Khanin [1], Yupu Liang [1], Azra Krek [1], Paul Zumbo [2,4], Christopher E. Mason [2,4], Nicholas D. Socci [1], Doron Betel [3,4]

[1]Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York
[2]Department of Physiology and Biophysics, Weill Cornell Medical College, New York
[3] Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, New York
[4] Institute for Computational Biomedicine, Weill Cornell Medical College, New York

January 24, 2013

A comparison of methods for differential expression analysis of RNA-seq data

BMC Bioinformatics 2013, **14**:91 doi:10.1186/1471-2105-14-91

Charlotte Soneson (Charlotte.Soneson@isb-sib.ch)
Mauro Delorenzi (Mauro.Delorenzi@unil.ch)

CSC

Comments from comparisons

- **”Methods based on negative binomial modeling have improved specificity and sensitivities as well as good control of false positive errors”**
- **”Cuffdiff performance has reduced sensitivity and specificity. We postulate that the source of this is related to the normalization procedure that attempts to account for both alternative isoform expression and length of transcripts”**

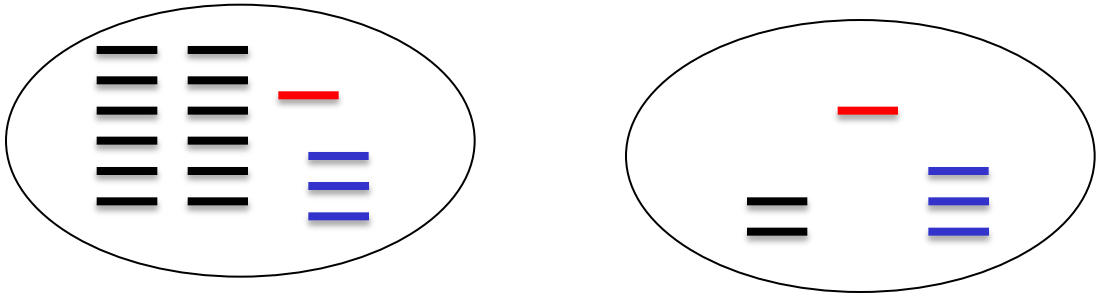
CSC

Differential expression analysis:

Normalization

Normalization

- **For comparing gene expression <u>within sample</u>, normalize for**
  - Gene length
  - Gene GC content
- **For comparing gene expression <u>between samples</u>, normalize for**
  - Library size (number of reads obtained)
  - RNA composition effect



Briefings in Bioinformatics Advance Access published September 17, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

- **"FPKM and TC are ineffective and should be definitely abandoned in the context of differential analysis"**
- **"In the presence of high count genes, only DESeq and TMM (edgeR) are able to maintain a reasonable false positive rate without any loss of power"**

CSC

RPKM and FPKM

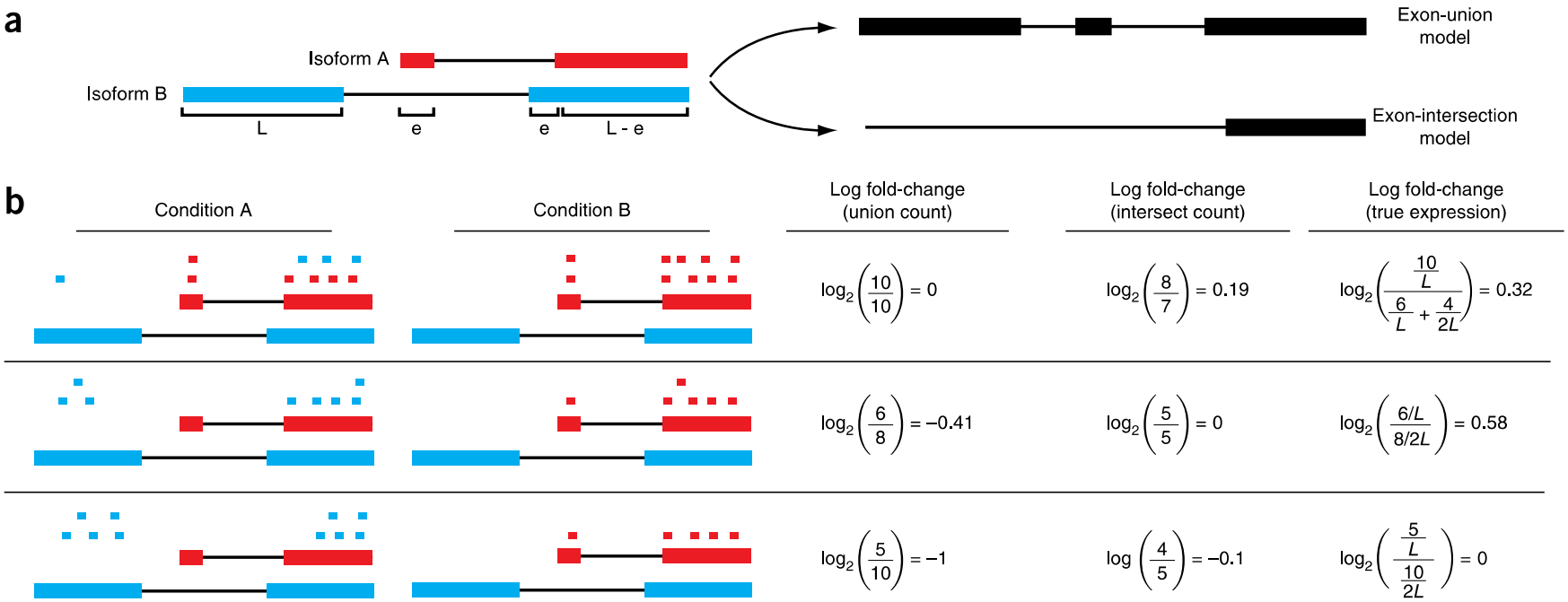
- **Reads/fragments per kilobase per million mapped reads. Examples:**
  - 20 kb transcript has 400 counts, library size is 20 million reads

  → RPKM = (400/20) / 20 = 1
  - 0.5 kb transcript has 10 counts, library size is 20 million reads

  → RPKM = (10/0.5) / 20 = 1
- **Normalizes for gene length and library size**
- **Can be used only for reporting expression values, not for testing differential expression**
  - Raw counts are needed to assess the measurement precision correctly

CSC

Estimating gene expression
-isoform switching problem



Trapnell et al. Nature Biotechnology 2013

CSC

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
- **Assumes 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**
  - 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

CSC

Filtering

Filtering

- **Filter out genes which have little chance of showing significant evidence for differential expression**
  - genes which are not expressed
  - genes which are expressed at very low level

- **Reduces the severity of multiple testing adjustment**

- **Should be independent**
  - do not use information on what group the sample belongs to

Differential expression analysis:

Dispersion estimation

Dispersion

- **Dispersion = $(BCV)^2$**
  - BCV = gene's biological coefficient of variation
  - E.g. if gene's expression typically differs from replicate to replicate by 20%, this gene's dispersion is $0.2^2 = 0.04$

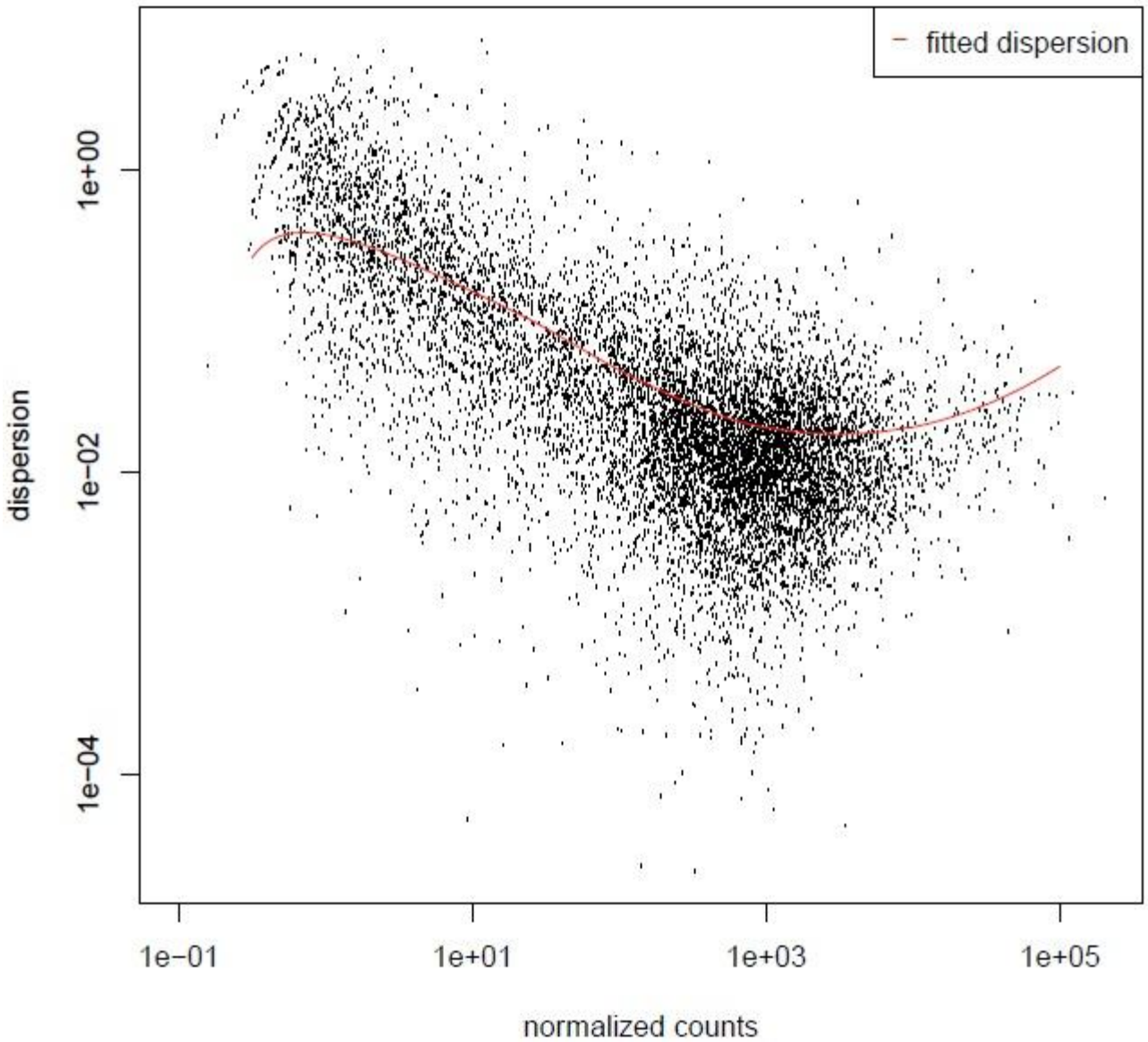
- **Note that the variance seen in counts is a sum of 2 things:**
  - Sample-to-sample variation (dispersion)
  - Uncertainty in measuring expression by counting reads

Dispersion estimation by edgeR and DESeq

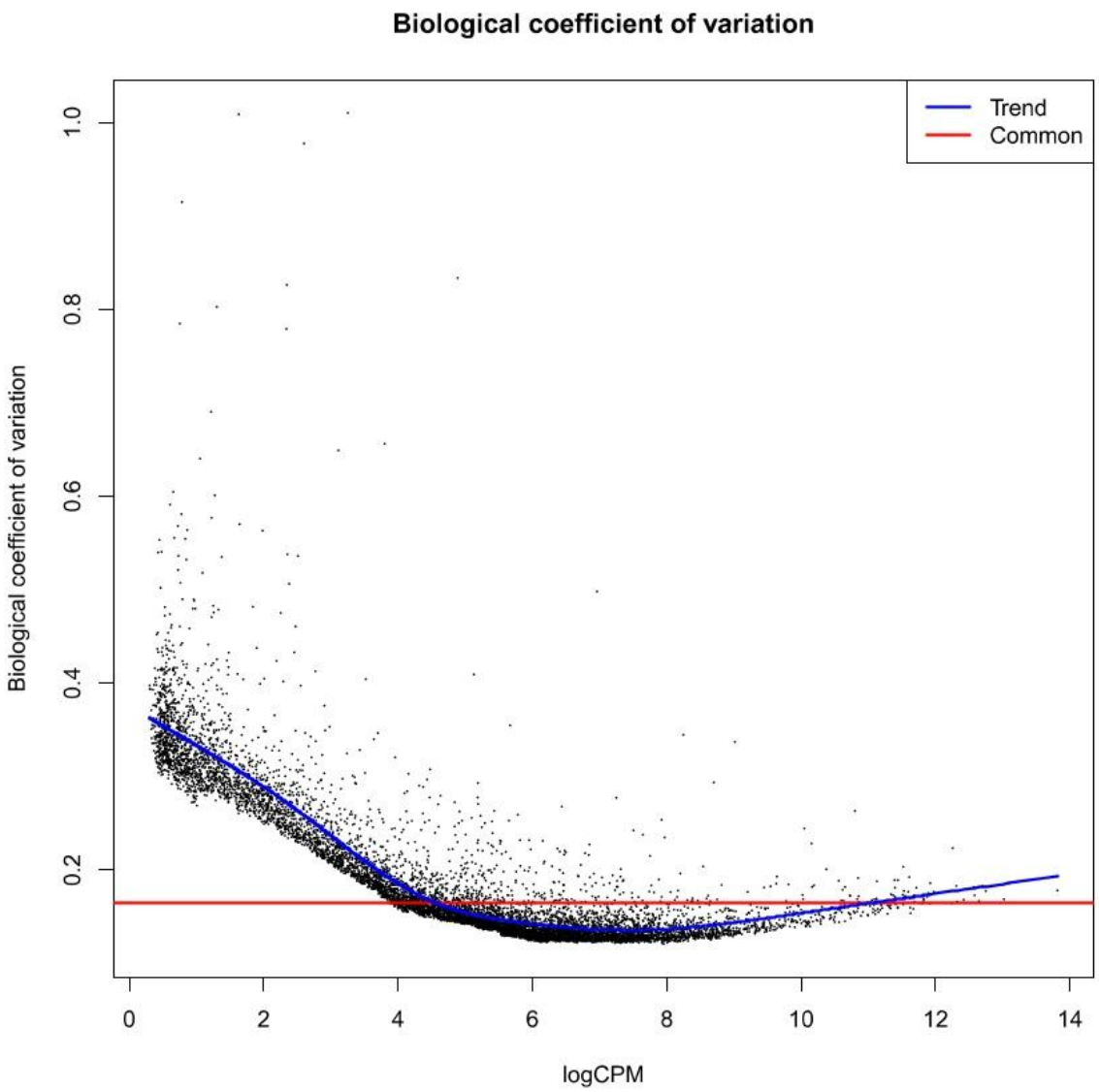
- **DESeq**
  - Models the observed mean-variance relationship for the genes using either parametric or local regression
  - User can choose to use the fitted values always, or only when they are higher than the genewise value

- **edgeR**
  - Estimates <u>common</u> dispersion for all genes using a conditional maximum likelyhood approach
  - <u>Trended</u> dispersion: takes binned common dispersion and abundance, and fits a curve though these binned values
  - <u>Tagwise</u> dispersion: uses empirical Bayes strategy to shrink gene-wise dispersions towards the common/trended one using a weighted likelyhood approach → genes that are consistent between replicates are ranked more highly

CSC


fitted dispersion
dispersion
1e+00
1e−02
1e−04
1e−01
1e+01
1e+03
1e+05
normalized counts

CSC


Biological coefficient of variation
Trend
Common
Biological coefficient of variation
1.0
0.8
0.6
0.4
0.2
0
2
4
6
8
10
12
14
logCPM

CSC

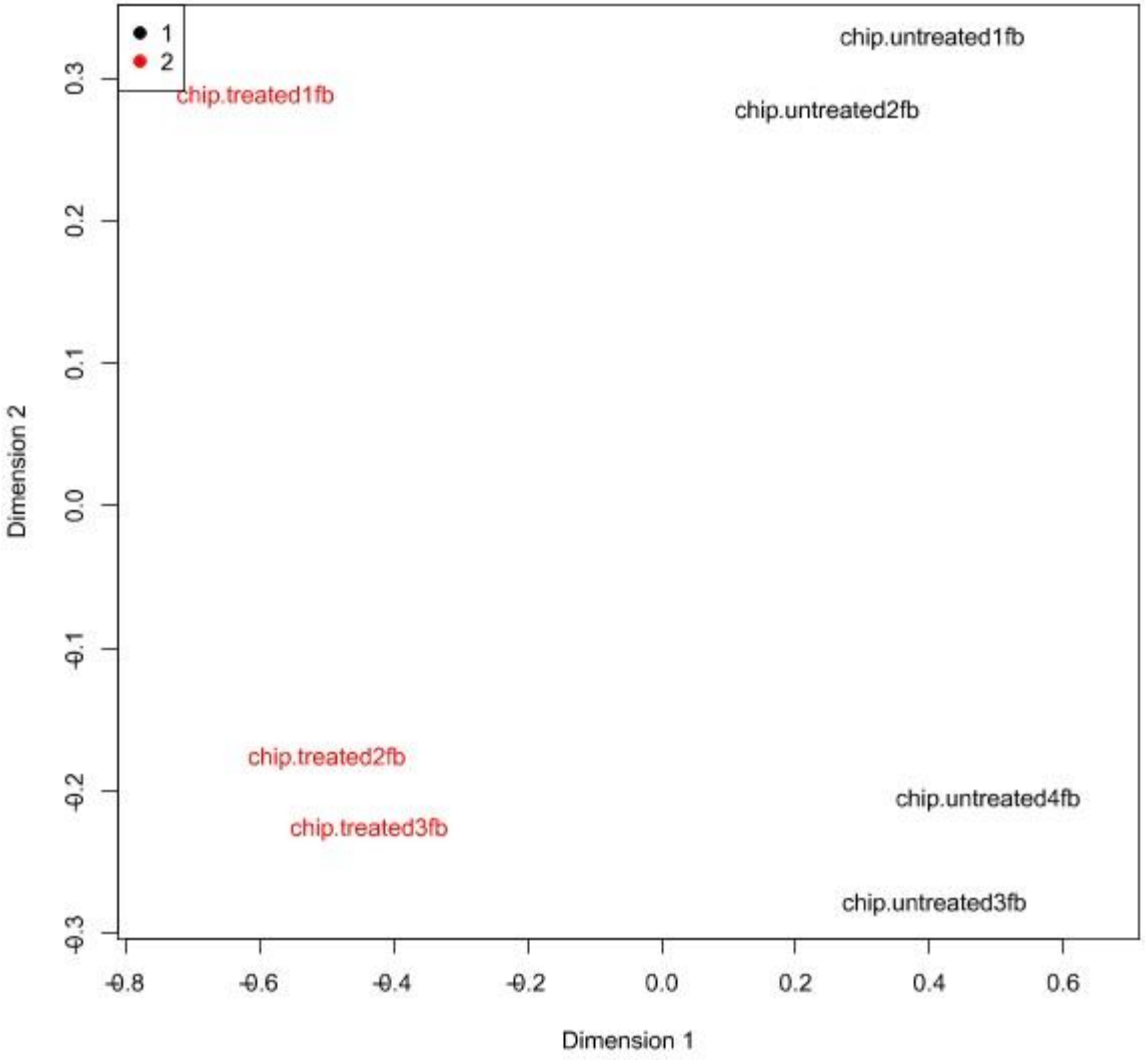
Data exploration using MDS plot

- **edgeR outputs multidimensional scaling (MDS) plot which shows the relative similarities between samples**

- **Allows you to see if replicates are consistent and if you can expect to find differentially expressed genes**

- **Distances correspond to the biological coefficient of variation between each pair of samples**
  - Calculated using 500 most heterogenous genes (that have largest tagwise dispersion treating all libraries as one group)

MDS plot by edgeR



Differential expression analysis:

Statistical testing

Statistical testing by DESeq and edgeR

- **Two group comparisons**
  - Exact test for negative binomial distribution

- **Multifactor experiments**
  - Generalized linear model (GLM) likelyhood ratio test
    - GLM = extension of linear models to non-normally distributed response data

Data

- **Drosophila data from RNAi knock-down of pasilla gene**
  - 4 untreated samples
    - 2 sequenced single end
    - 2 sequenced paired end
  - 3 samples treated with RNAi
    - 1 sequenced single end
    - 2 sequenced paired end

Extra: Technical slides about Chipster

Adding analysis tools is easy
- simple tool description syntax

Analysis tools - RNA-seq - Differential expression using edgeR

Column describing groups	group
Apply TMM normalization	yes
Dispersion method	tagwise
Dispersion value used if no replicates are available	0.1
Analyze only genes which have counts in at least this many samples	0
P-value cutoff	0.05
Multiple testing correction	BH
Plot width	600
Plot height	600

Hide parameters | Run

Differential gene expression analysis using the exact statistical methods of the edgeR Bioconductor package. You can create the input count table and phenodata file using the tool "Utilities - Define NGS experiment". Please note that this tool is suitable only for two group comparisons. For multifactor experiments please use the tool "Differential expression using edgeR for multivariate experiments".

More help | Show tool sourcecode

Source Code

```
# TOOL ngs-dea-edger-RNA.R: "Differential expression using edgeR" (Differential gene expression analysis using the exa
# INPUT data.tsv TYPE GENERIC
# INPUT phenodata.tsv TYPE GENERIC
# OUTPUT OPTIONAL de-list-edger.tsv
# OUTPUT OPTIONAL de-list-edger.bed
# OUTPUT OPTIONAL ma-plot-edger.pdf
# OUTPUT OPTIONAL mds-plot-edger.pdf
# OUTPUT OPTIONAL edger-log.txt
# OUTPUT OPTIONAL p-value-plot-edger.pdf
# OUTPUT OPTIONAL dispersion-edger.pdf
# PARAMETER column: "Column describing groups" TYPE METACOLUMN_SEL DEFAULT group (Phenodata column describing the grou
# PARAMETER OPTIONAL normalization: "Apply TMM normalization" TYPE [yes, no] DEFAULT yes (Should normalization based o
# PARAMETER OPTIONAL dispersion_method: "Dispersion method" TYPE [common, tagwise] DEFAULT tagwise (The dispersion of
# PARAMETER OPTIONAL dispersion_estimate:"Dispersion value used if no replicates are available" TYPE DECIMAL FROM 0 TO
# PARAMETER OPTIONAL filter: "Analyze only genes which have counts in at least this many samples" TYPE INTEGER FROM 0
# PARAMETER OPTIONAL p_value_threshold: "P-value cutoff" TYPE DECIMAL FROM 0 TO 1 DEFAULT 0.05 (The cutoff for adjuste
# PARAMETER OPTIONAL p_value_adjustment_method: "Multiple testing correction" TYPE [none, Bonferroni, Holm, Hochberg,
# PARAMETER OPTIONAL w: "Plot width" TYPE INTEGER FROM 200 TO 3200 DEFAULT 600 (Width of the plotted image)
# PARAMETER OPTIONAL h: "Plot height" TYPE INTEGER FROM 200 TO 3200 DEFAULT 600 (Height of the plotted image)
```

To make it even easier:

Tool editor GUI for writing tool descriptions

localhost:8080/tool-editor

localhost:8080/tool-editor/

Seurattavat | Seurattavat 2 | Luetuksi | Vipunen | Chipster | Trello | BN | Takki

Tool Editor

	ACTIONS
▼Differential expression using DESeq (dea-deseq.R)	
▶Inputs	+
▶Outputs	+
▼Parameters	+
Parameter Column describing groups METACOLUMN_SEL	✕
Parameter Apply normalization ENUM OPTIONAL	✕
Parameter Dispersion estimation method ENUM OPTIONAL	✕
Parameter Use fitted dispersion values ENUM OPTIONAL	✕
Parameter Multiple testing correction ENUM OPTIONAL	✕
Parameter P-value cutoff DECIMAL OPTIONAL	✕
Parameter Plot width INTEGER OPTIONAL	✕
Parameter Plot height INTEGER OPTIONAL	✕

Parameter **Dispersion estimation method ENUM OPTIONAL**

User defined parameter: dispersion_estimate

Display name: Dispersion estimation method

Type: ENUM

ID	NAME	ACTION
parametric	parametric	✕
local	local	✕

New Row

Maximum value:

Minimum value:

Default: local

Optional: ☑

Description: Dispersion can be estimated using a local fit or a two-coefficient parametric model. Local fit is suitable in most cases, including when there are no biological replicates. The parametric model may be preferable under certain circumstances.

✕

Clear All

```
# TOOL dea-deseq.R: "Differential expression using DESeq" (Differential expression analysis using the DESeq Bioconductor package. You can create the input count table and phenodata file using the tool \"Utilities - Define NGS experiment\".)
# INPUT data.tsv TYPE GENERIC
# INPUT phenodata.tsv TYPE GENERIC
# OUTPUT OPTIONAL de-list-deseq.tsv
# OUTPUT OPTIONAL de-list-deseq.bed
# OUTPUT OPTIONAL ma-plot-deseq.pdf
# OUTPUT OPTIONAL dispersion-plot-deseq.pdf
# OUTPUT OPTIONAL p-value-plot-deseq.pdf
# PARAMETER column: "Column describing groups" TYPE METACOLUMN_SEL DEFAULT group (Phenodata column describing the groups to test.)
# PARAMETER OPTIONAL normalization: "Apply normalization" TYPE [yes, no] DEFAULT yes (Should effective library size be estimated. This corrects for RNA composition bias. Note that if you have supplied library size in phenodata, size factors are ca
based on the library size total, and composition bias is not corrected.)
# PARAMETER OPTIONAL dispersion_estimate:"Dispersion estimation method" TYPE [parametric: "parametric", local: "local"] DEFAULT local (Dispersion can be estimated using a local fit or a two-coefficient parametric model. Local fit is suitable in mo
```

Server is easy to install and update

- **Virtual machine image**
  - for KVM, VirtualBox, VMware platforms
  - contains all analysis tools and related data
    - easy for the admin
    - large size → we will make species-specific bundles

- **Update script**
  - no need to download the whole thing when updating to new Chipster version
  - updates everything (tools, databases, client, server)

- **Compute service can be also deployed to queue system, but a cloud-like cluster is a better match**
  - responsiveness, efficient resource usage

CSC

Upcoming in Chipster v3.0

- **Data handling improvement**
  - Permanent server side sessions
  - Data can come to the server directly from a url

- **Admin GUI to monitor and manage**
  - Disk space usage per user
  - Running compute services and connected clients
  - Jobs and statistics

- **Improvements to client GUI**
  - More space for viewing dataset's metadata
  - Shortcuts to visualization options

CSC

Admin GUI

- keep track of disk space usage, server instances, jobs