Introduction to variant analysis from sequencing data

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

What will I learn?

- **Basics of exome and genome sequencing**
- **How to operate Chipster software**
- **Detecting genomic variants in sequencing data**
  - Central concepts
  - Analysis steps
  - File formats

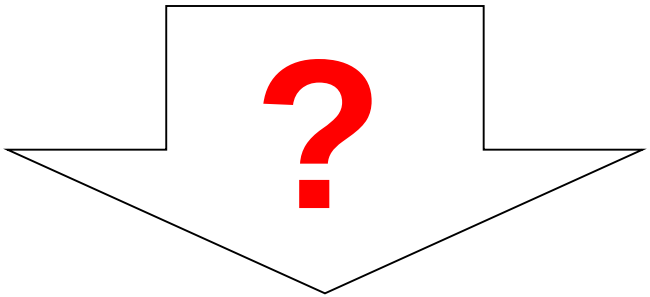
The task

TGGCTGGCTTGGCTGGCTTGGCTGGCTGATTGGTTTGGCT

GGCCATCAGAGAAATGCAAATCAAAACCACAATGAGATAC

CATCTCACACCAGTTAGAATGGCAATCATTAAAAAG

AGAAGGGCGAATGGTGATAGAGAAAATGGAGGTGGC

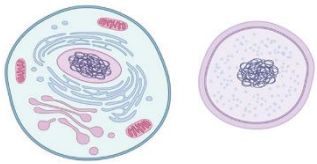


”There is a C → T variant in chr 20 in position 3044461.”
”It changes alanine to valine in the GNRH protein and is probably damaging.”

CSC

Basics of exome and genome sequencing

What happens in the lab?

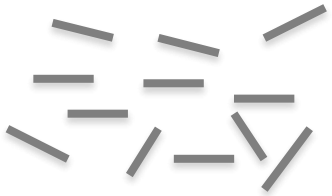


sample

isolate the DNA

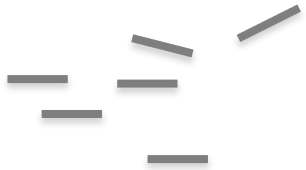
DNA

fragment DNA



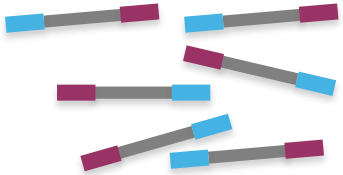
DNA fragments

exome capture



fragments containing exons

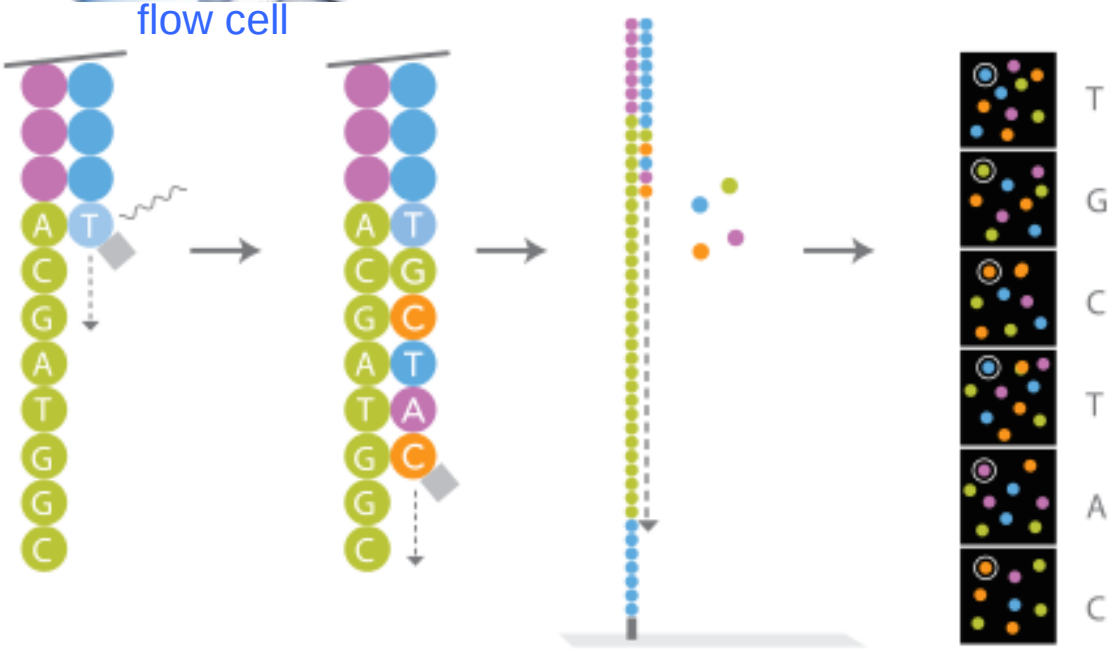
add adapters



fragments with adapters

sequencing

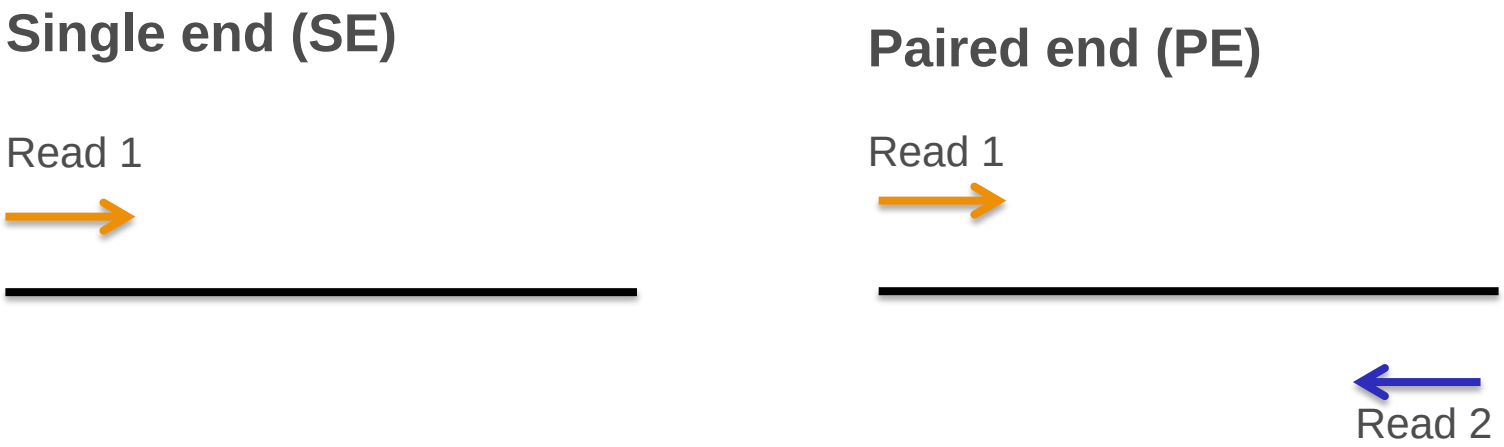
flow cell



Reads (.fastq file)

CSC

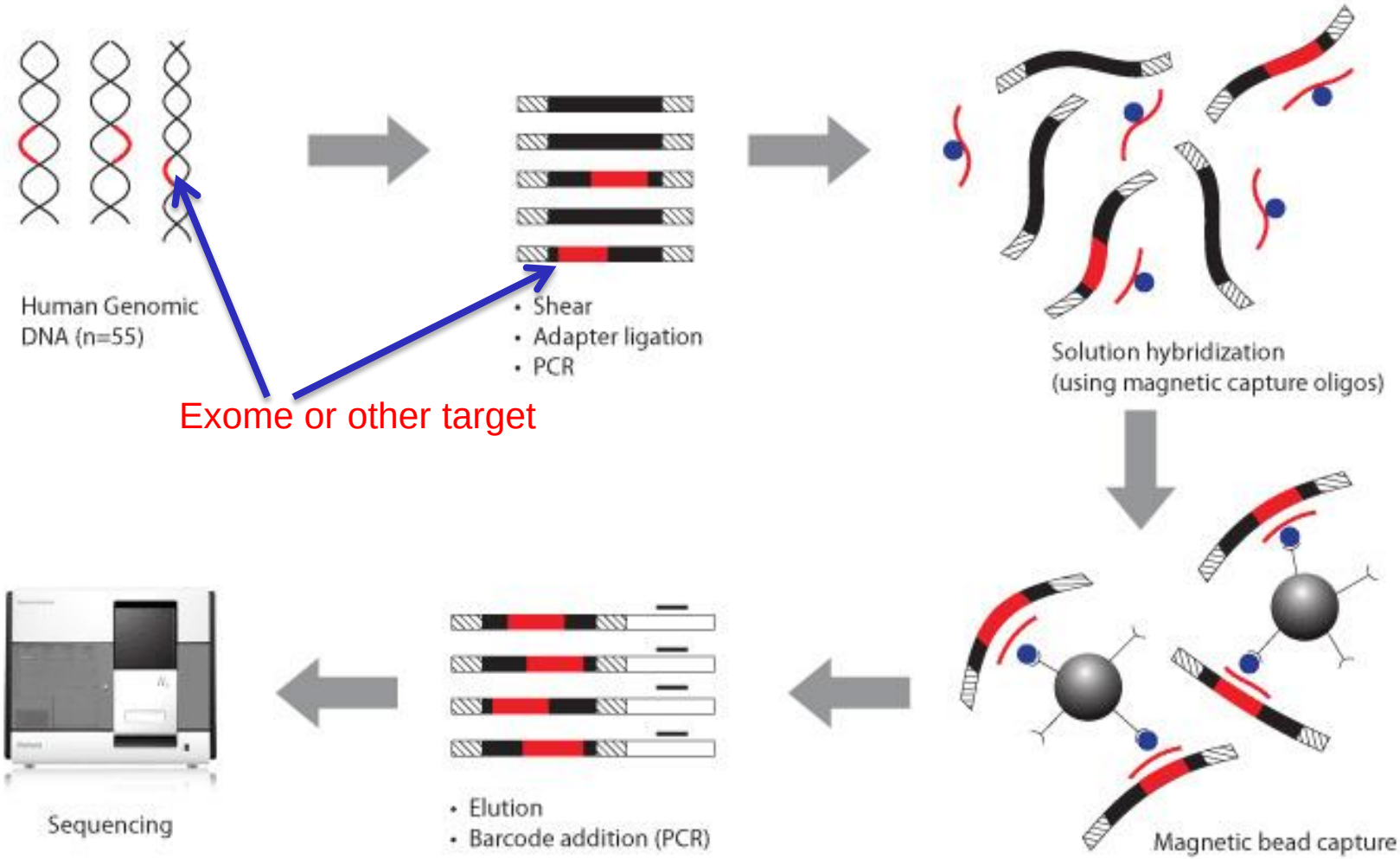
Paired end vs. single end sequencing



http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html

CSC

Exome capture (one example)



https://www.idtdna.com/pages/images/decoded/dcd13_yrht1_092211.jpg?sfvrsn=0

CSC

https://genohub.com/exome-sequencing-library-preparation/

Whole Exome Sequencing (WES)

- **Exome is about 2% of genome**
  - contains 80-90% of known disease related variants
- **Many methods for exome capture**
  - Different targets
    - Exome = ?
    - Non-coding exon flanking regions (UTR, promoter) included…?
  - Different prices, lab times and input requirements (50 ng – 1 ug)
- **On-target rate = enrichment efficiency**
  - varies usually between 0.65-0.75

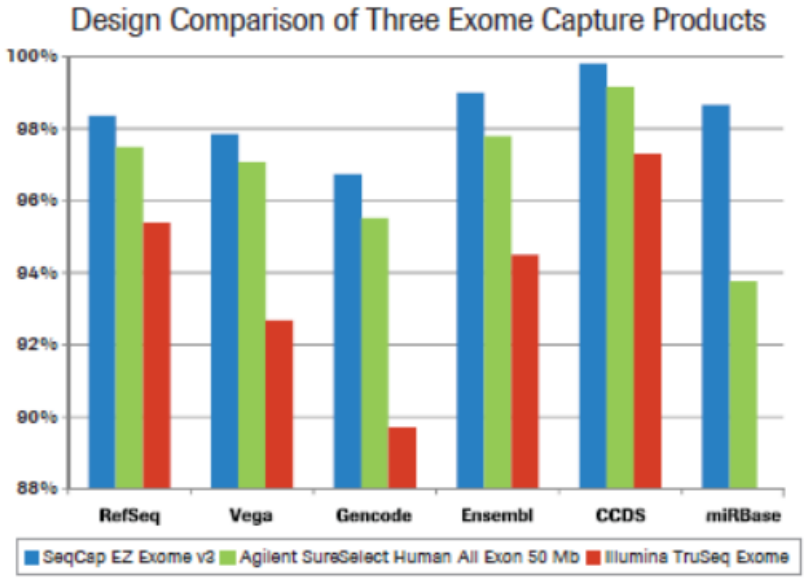


Fig1: Source www.NimbleGen.com

CSC

https://genohub.com/exome-sequencing-library-preparation/

Should I sequence exome or whole genome?

- **Whole Genome Sequencing (WGS)**
  - + Ability to call structural variations & non-coding variants
  - + More reliable and uniform coverage of the exome
  - + PCR amplification not a necessity (no duplicate issues)
  - + No reference bias (since no capturing)
  - + Universal, works for all the species
  - - Expensive (storage, transfer and analysis costs)
  - - Huge amount of data to store and process
  - - Lots of confusing data: how to interpret non-coding area variants?

- **Whole Exome Sequencing (WES)**
  - + Cheaper (although library prep costs)
  - + More reasonable amount of data
  - + More samples & with higher coverage
  - - Problems in capturing: duplicates, biases, old information…
  - - Not for all species
  - - Coverage? What are we missing? On target rate?

What are you looking for and what is required to find it.

CSC

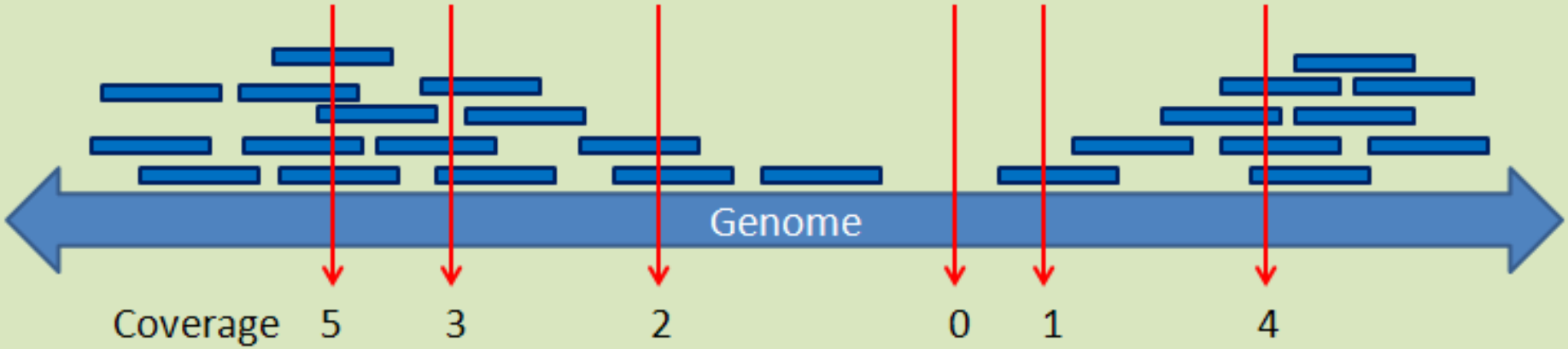
Coverage

https://www.edgebio.com/exome-sequencing-coverage-analysis

- **Read depth, how many reads per each nucleotide on average**
- **Whole genome sequencing**
  - Genotype calls 35x, INDELs 60x, SNVs 30x
- **Exome sequencing**
  - SNPs 100x

$$C = LN / G$$

- C stands for coverage
- G is the haploid genome length
  …or size of target
- L is the read length
- N is the number of reads



Introduction to Chipster

Chipster

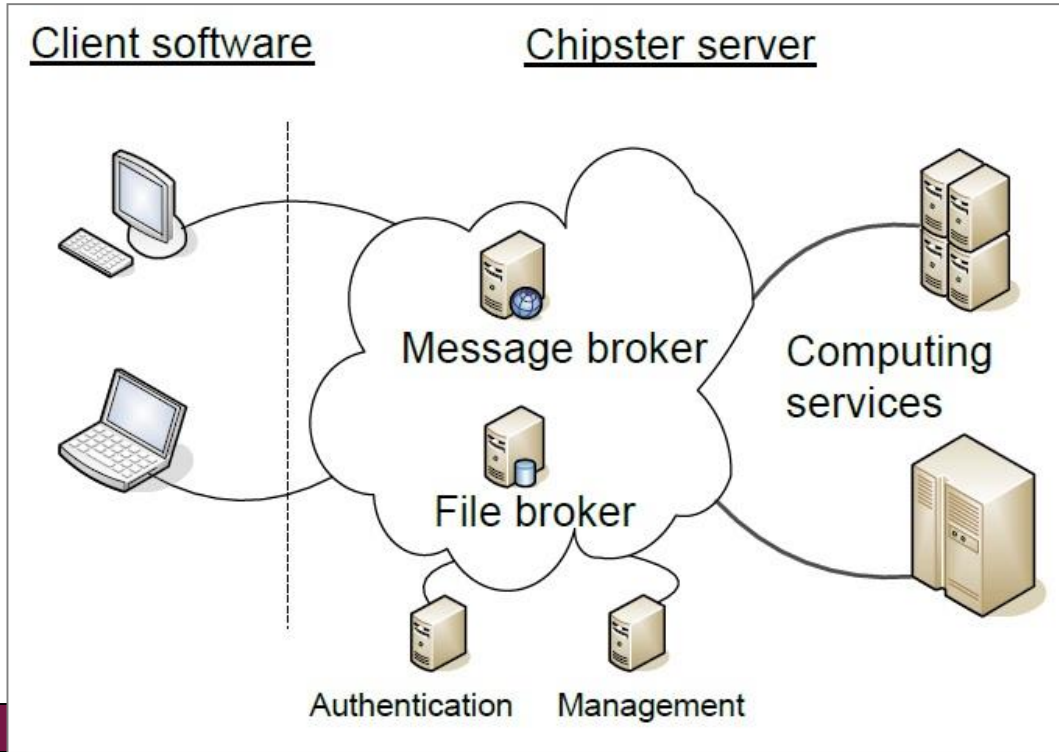
- **Provides an easy access to over 350 analysis tools**
  - Command line tools
  - R/Bioconductor packages
- **Free, open source software**

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

CSC

Technical aspects

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



Chipster

Open source platform for data analysis

- Home
- Getting access
- Analysis tool content
- Screenshots
- Manual
- Tutorial videos
- Cite
- FAQ
- Contact

- For developers:
  - Open source project
  - Tool editor

Welcome to Chipster

Chipster is a user-friendly analysis software for high-throughput data. It contains over 350 analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data. Users can save and share automatic analysis workflows, and visualize data interactively using a built-in genome browser and many other visualizations.

Chipster's client software uses Java Web Start to install itself automatically, and it connects to computing servers for the actual analysis. Chipster is open source and the server environment is available as a virtual machine image free of charge. If you would like to use Chipster running on CSC's server, you need a user account.

Launch Chipster v3.9

...or launch with more memory: 3 GB or 6 GB

If you have trouble launching Chipster, read this

News:

- 25.5.2016 Chipster tutorial on nanotoxicology analysis now available
- 23.5.2016 Version 3.9 released
- 10.7.2015 Chipster tutorial videos now in YouTube
- 19.8.2014 RNA-seq data analysis guidebook with Chipster instructions

Training:

- 10.11.2016 RNA-seq data analysis, Brisbane
- 13.6.2016 Variant analysis, CSC
- 1.-2.6.2016 RNA-seq and ChIP-seq data analysis, Prague
- 28.4.2016 Introduction to CSC services and Chipster, University of Helsinki
- 26.4.2016 RNA-seq data analysis, BTK

Chipster 3.4.0 (build 1441)

File Edit View Workflow Help

Datasets

- two-sample.tsv
- column-value-filter.tsv
- hc.tre
- kmeans.pdf
- kmeans.tsv
- extract.tsv
- seqs.txt.wee
- seqs.html
- annotations.tsv
- annotations.html
- cpdb-pathways.html
- cpdb-pathways.tsv
- cpdb-genes.tsv

Analysis tools

Microarrays | NGS | Misc

Normalisation	One sample tests
Quality control	Two groups tests
Preprocessing	ROTS
Statistics	SAM
Clustering	Several groups tests
Annotation	Linear modelling
Pathways	Linear modelling using user-defined design ma
Promoter analysis	Test proportions
Copy number aberrations	Correlate with phenodata
Visualisation	Correlate miRNA with target expression
Utilities	Time series

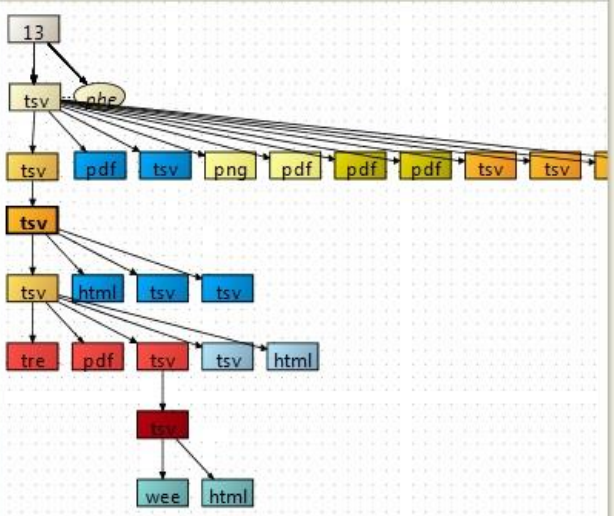
Show parameters | Run

Tests for comparing the mean gene expression of two groups. LPE only works, if the whole normalized data is used, i.e., the data should not be filtered. Other than empiricalBayes might be slow, if run on unfiltered data.

More help | Show tool sourcecode

Workflow

Fit



Visualisation

Maximise | Detach | Close

two-sample.tsv

472 kB, Wed Sep 03 06:56:07 EEST 2014

(Click here to add your notes)

Analysis history

Statistics / Two groups tests

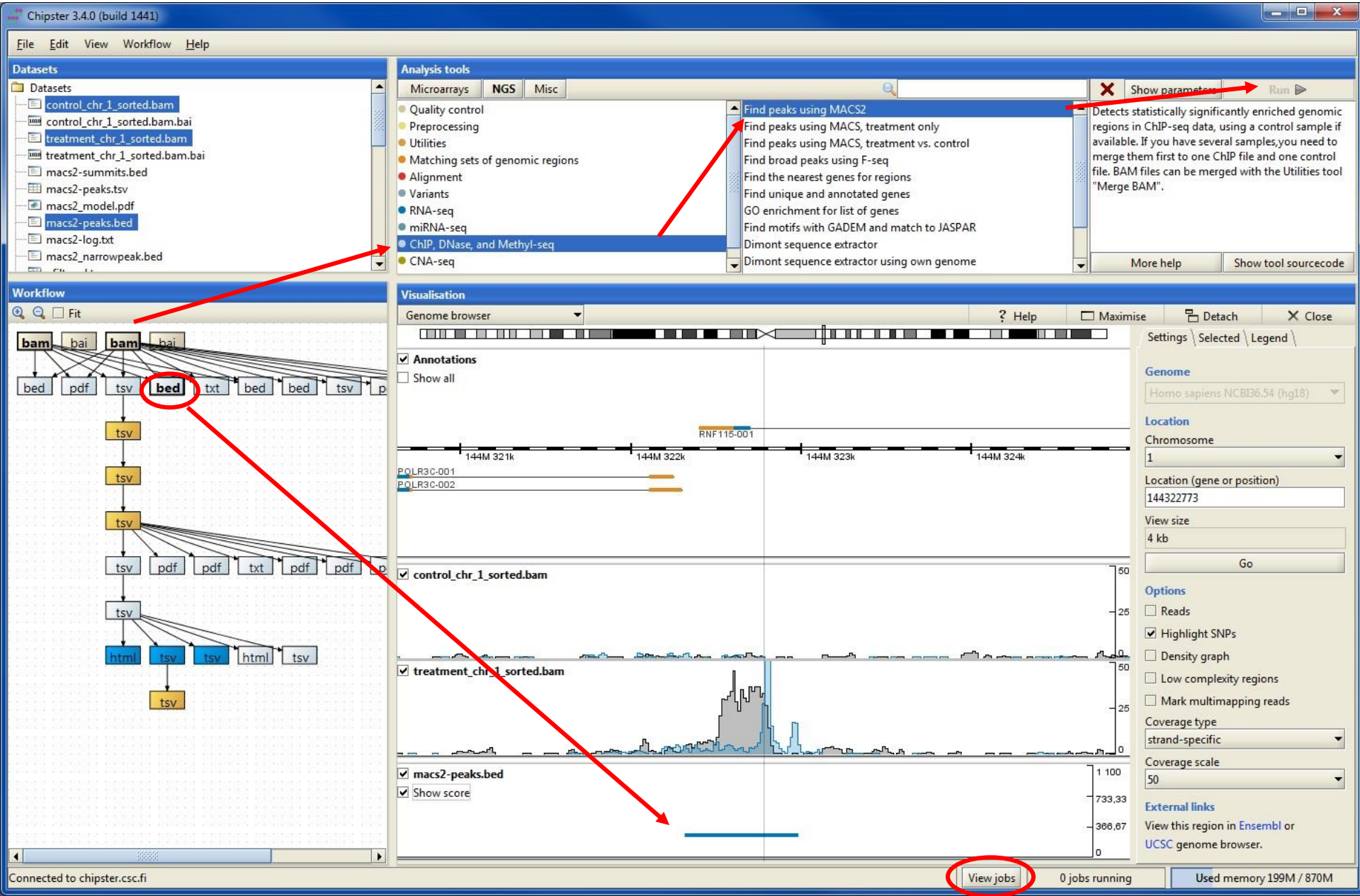
Column	group
Pairing	EMPTY
Test	empirical Bayes
p-value adjustment method	BH
p-value threshold	0.01
Show NA	no

- Spreadsheet
- Heatmap
- Expression profile
- Volcano plot
- Scatterplot
- 3D Scatterplot
- Histogram
- Open in external web browser

Connected to chipster.csc.fi | View jobs | 0 jobs running | Used memory 118M / 870M

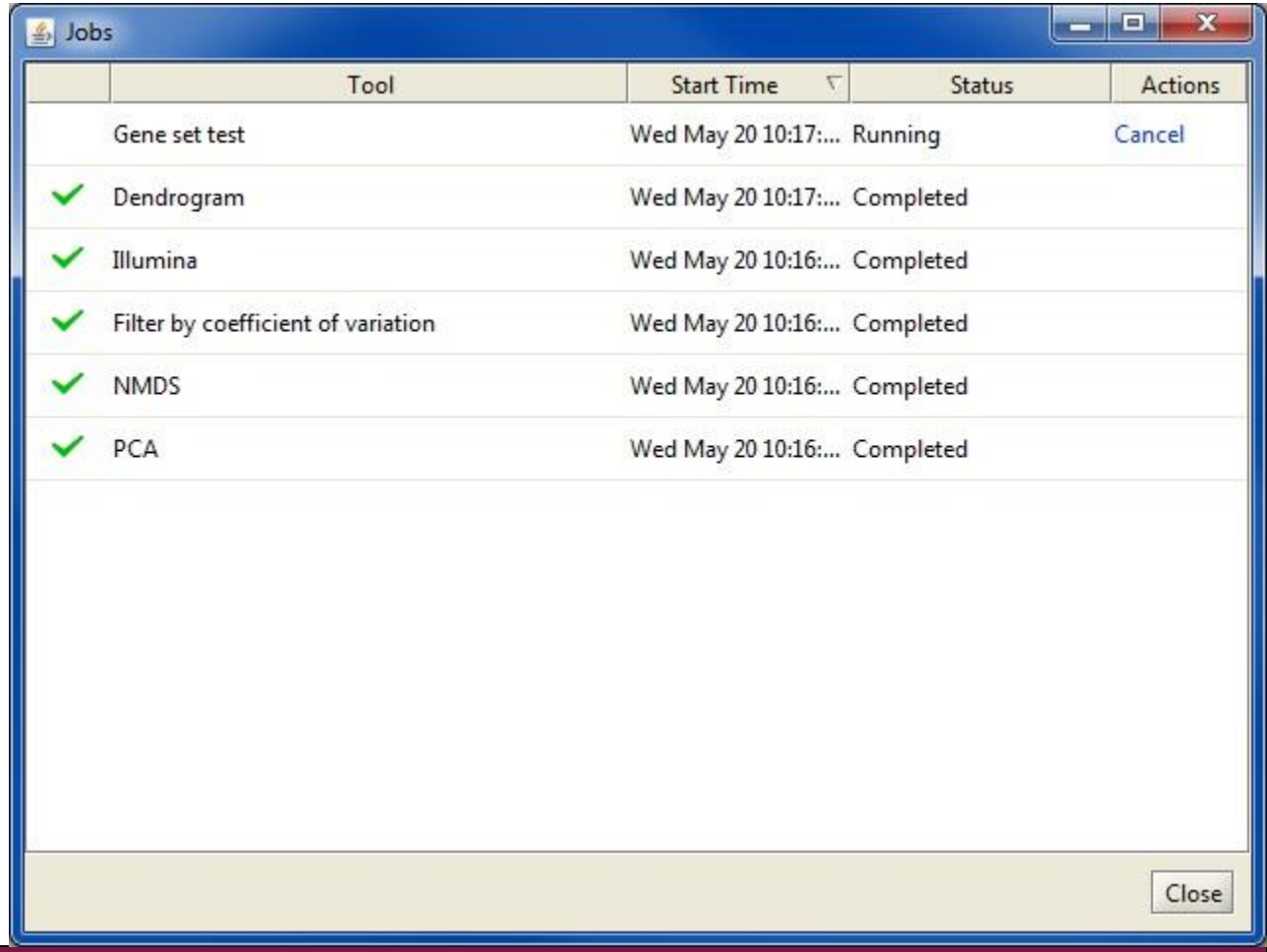
Mode of operation

Select: data → tool category → tool → run → visualize



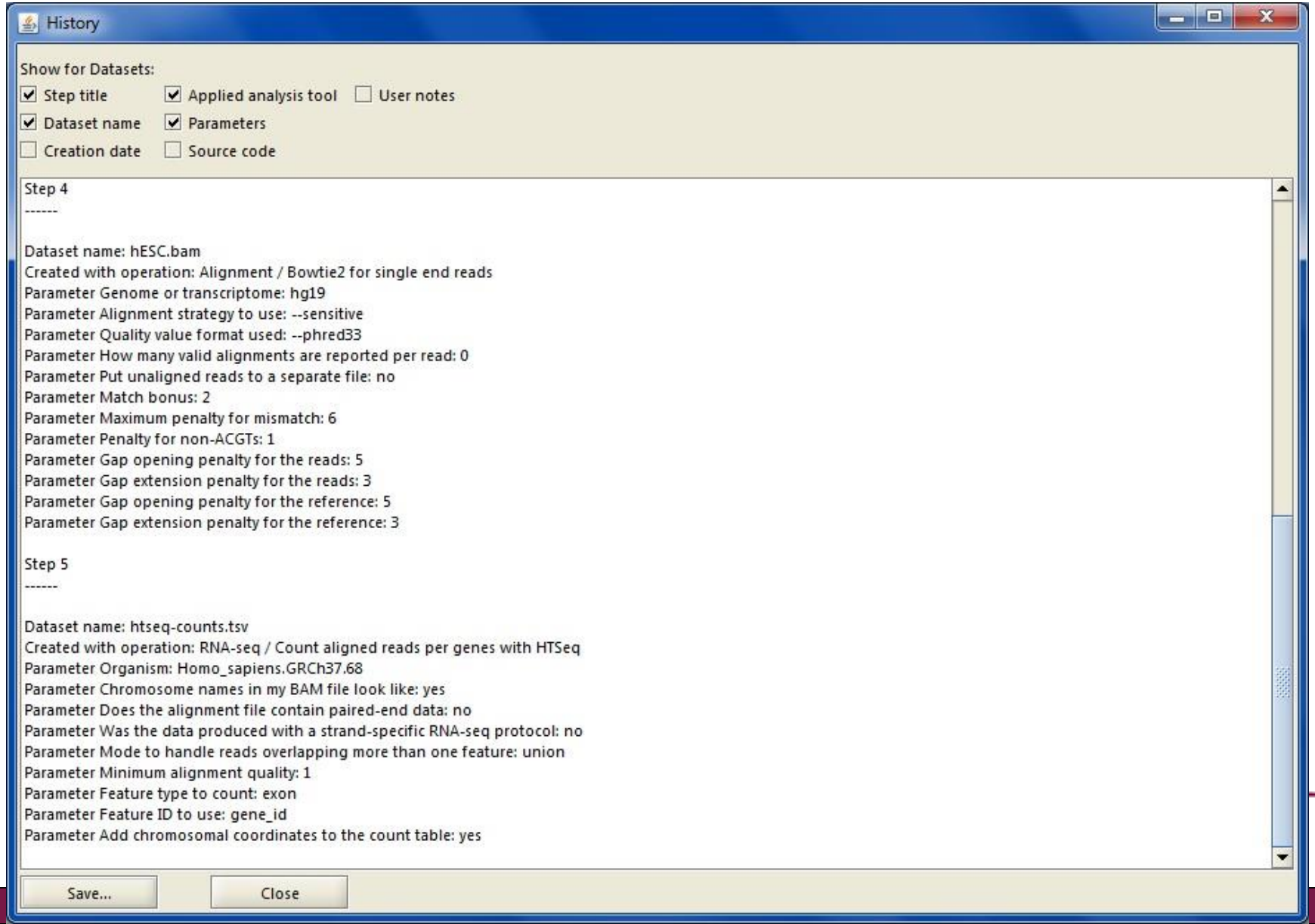
Job manager

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



Analysis history is saved automatically

-you can add tool source code to reports if needed

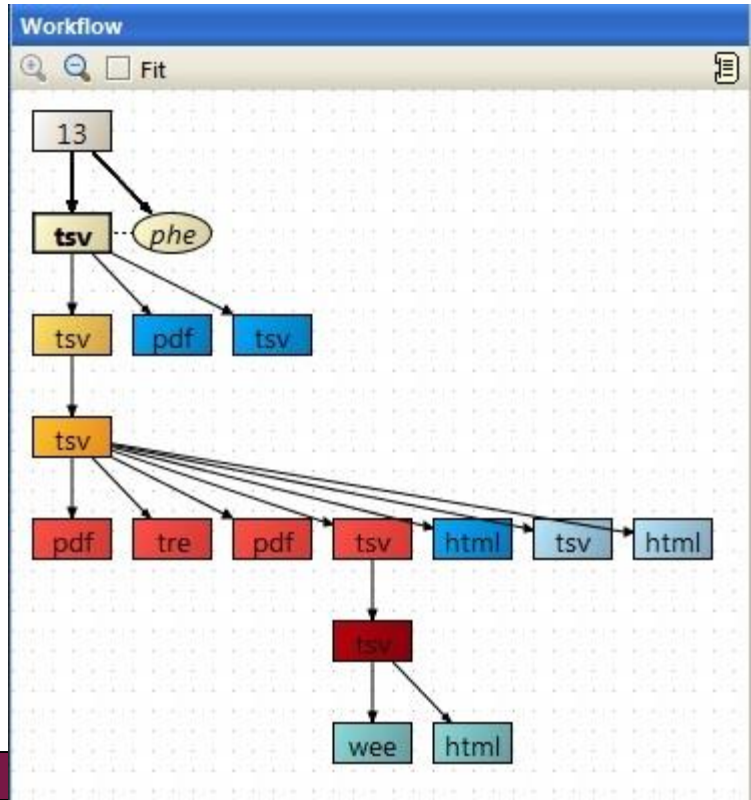


Analysis sessions

- **In order to continue your work later, you have to save the analysis session.**
- **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 or in the cloud.**
- **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

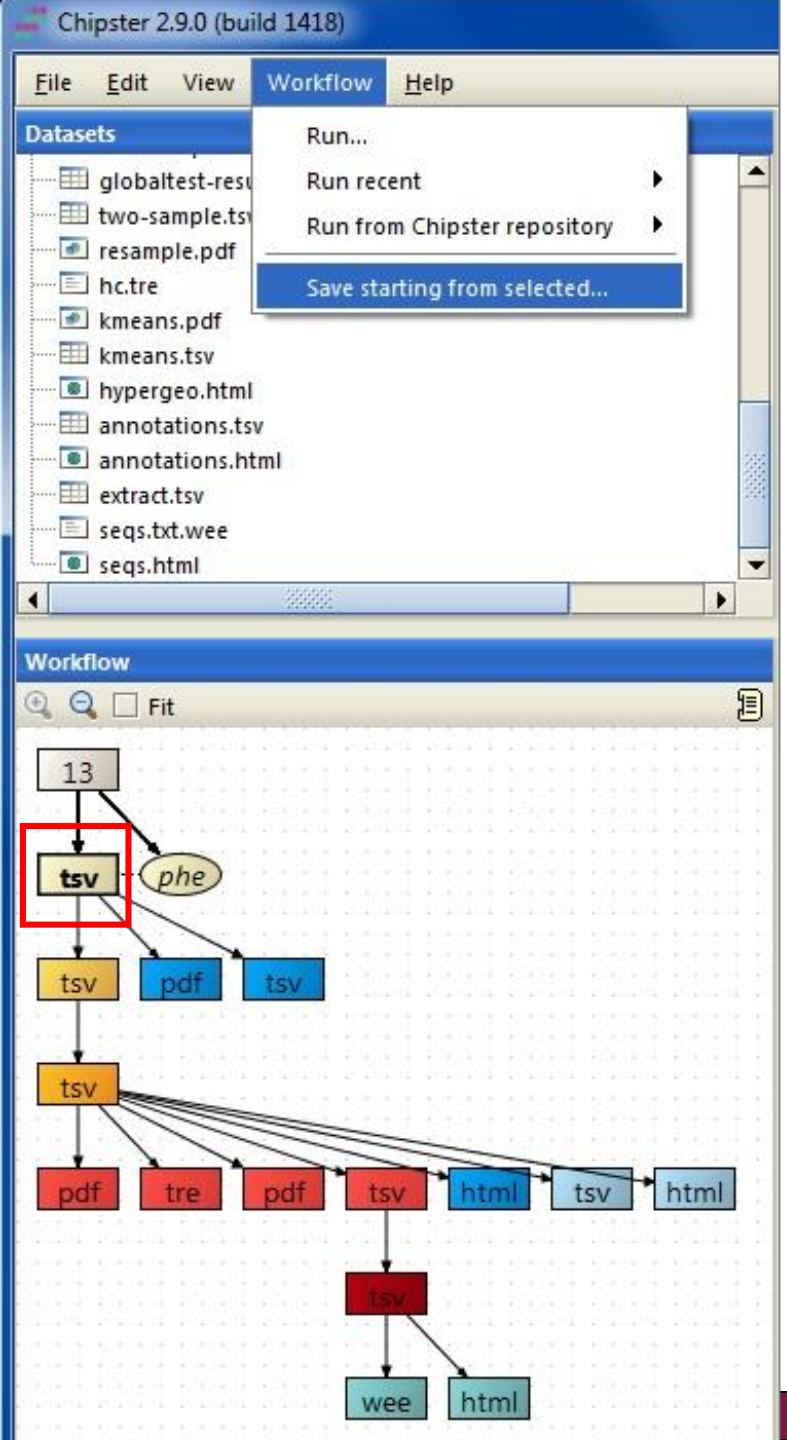


Workflow – reusing and sharing your analysis pipeline

- **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 starting point 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

Analysis tool overview

- **150 NGS tools for**
  - RNA-seq
  - miRNA-seq
  - exome/genome-seq
  - ChIP-seq
  - FAIRE/DNase-seq
  - MeDIP-seq
  - CNA-seq
  - Metagenomics (16S rRNA)
- **140 microarray tools for**
  - gene expression
  - miRNA expression
  - protein expression
  - aCGH
  - SNP
  - integration of different data
- **60 tools for sequence analysis**
  - BLAST, EMBOSS, MAFFT
  - Phylip

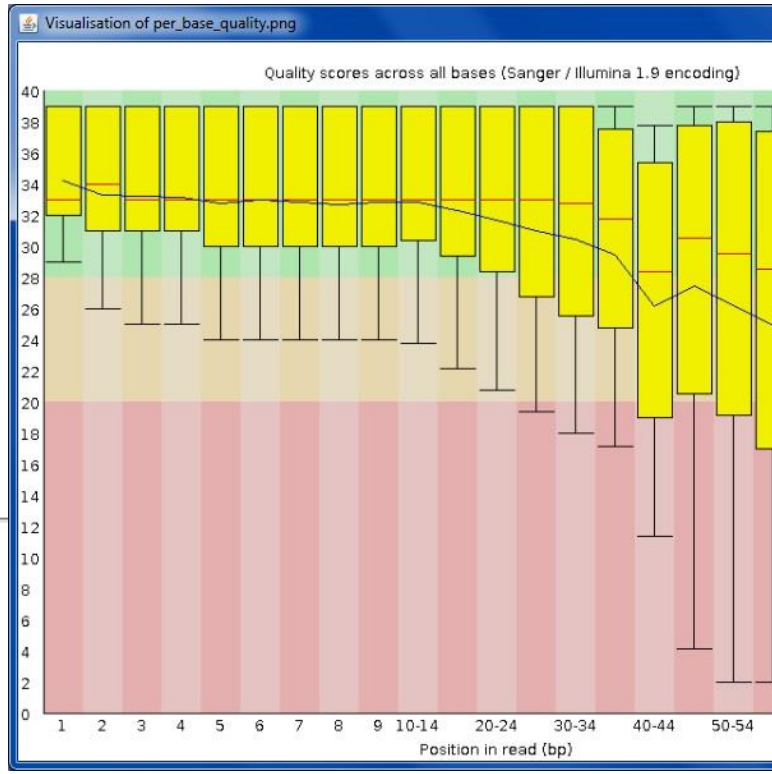
CSC

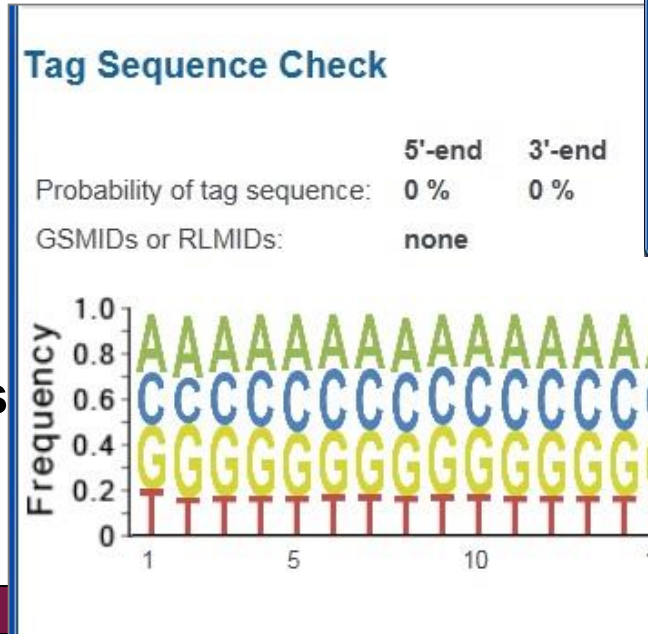
Tools for QC, processing and mapping

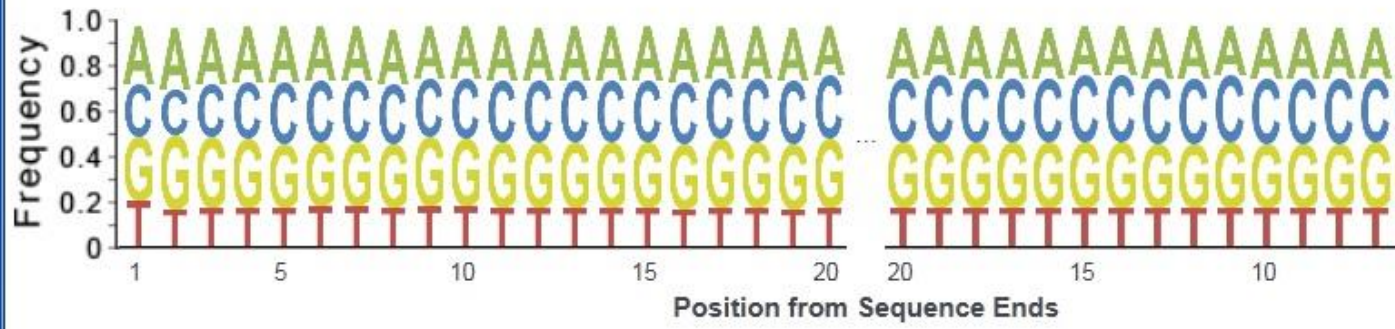
- FastQC
- PRINSEQ
- FastX
- TagCleaner
- Trimmomatic

- Bowtie
- TopHat
- BWA

- Picard
- SAMtools
- BEDTools

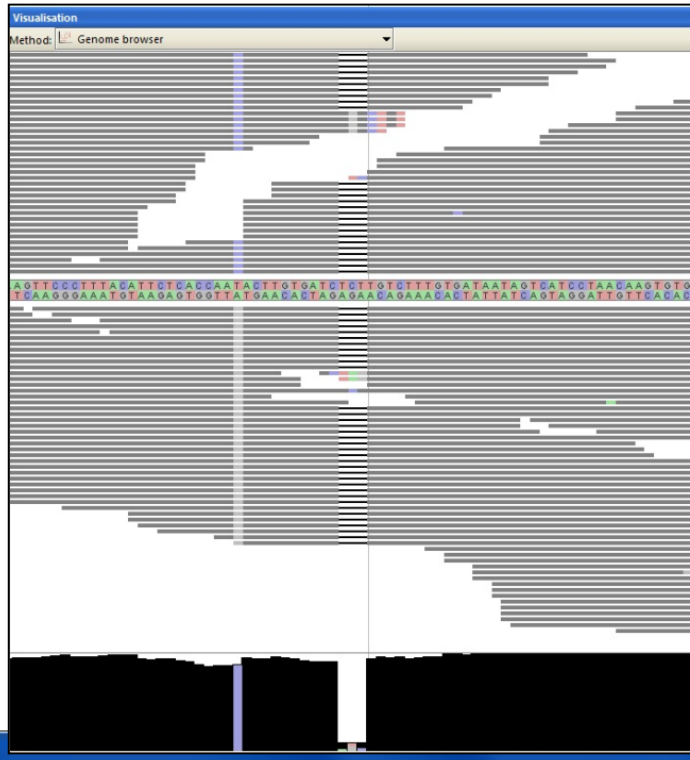






Exome/genome-seq tools

- **Variant calling**
  - Samtools, bcftools
- **Variant filtering and reporting**
  - VCFtools
- **Variant annotation**
  - AnnotateVariant (Bioconductor)
  - Variant Effect Predictor



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

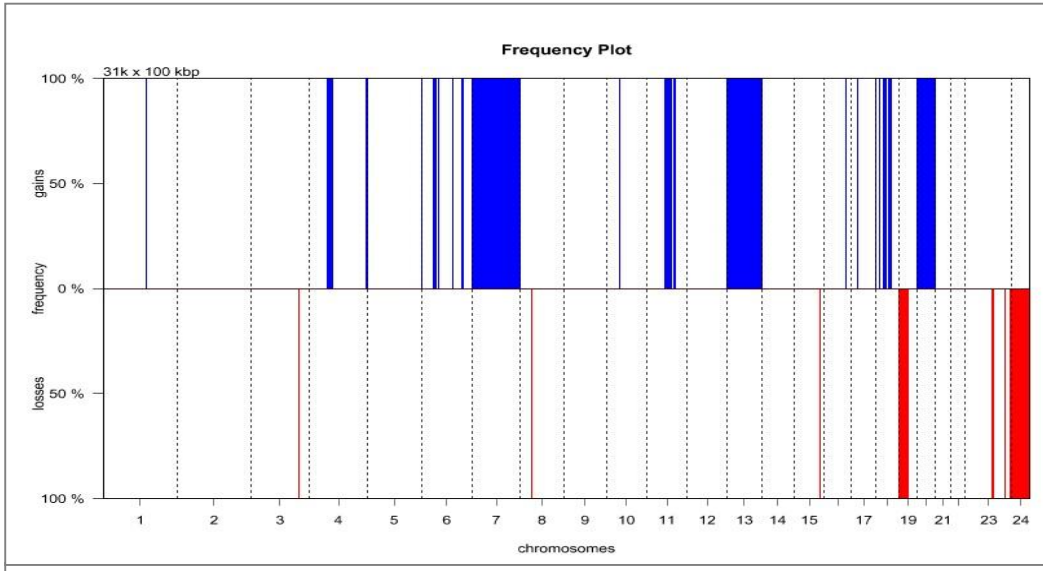
Visualisation of coding-variants.tsv

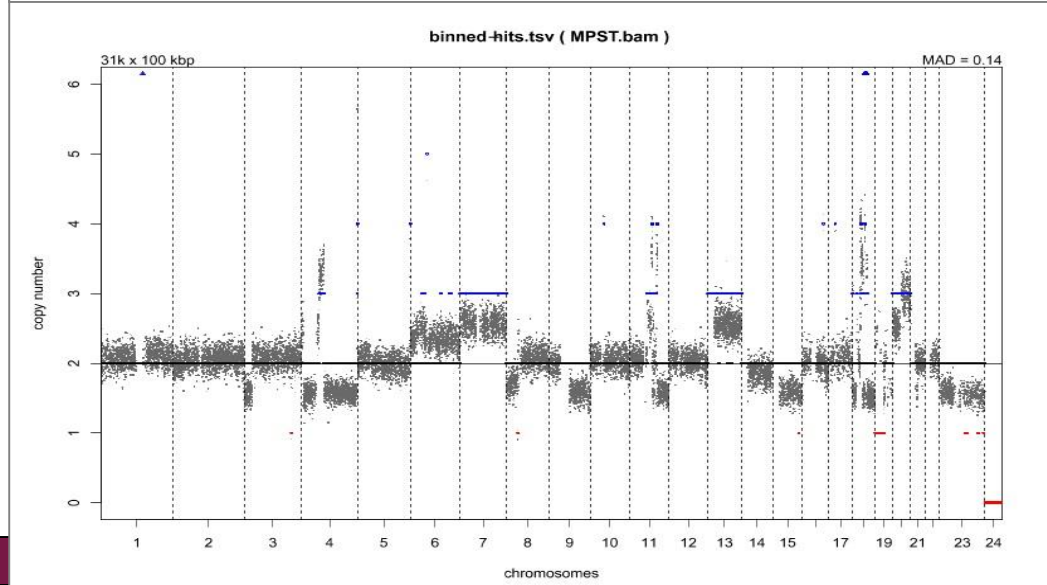
Showing 3 rows of 3

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

CNA-seq tools

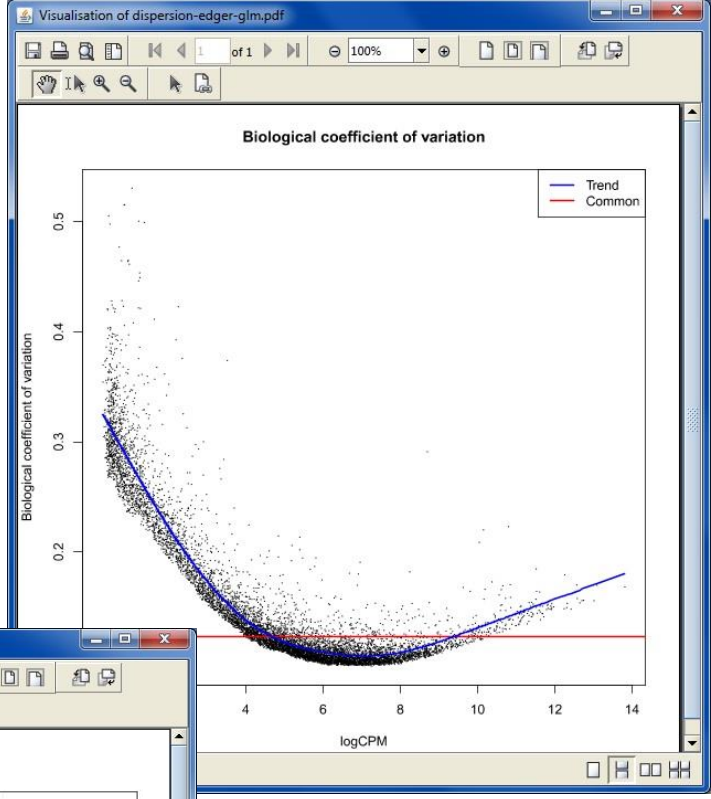
- **Count reads in bins**
  - Correct for GC content
- **Segment and call CNA**
  - Filter for mappability
  - Plot profiles
- **Group comparisons**
- **Clustering**
- **Detect genes in CNA**
- **GO enrichment**
- **Integrate with expression**

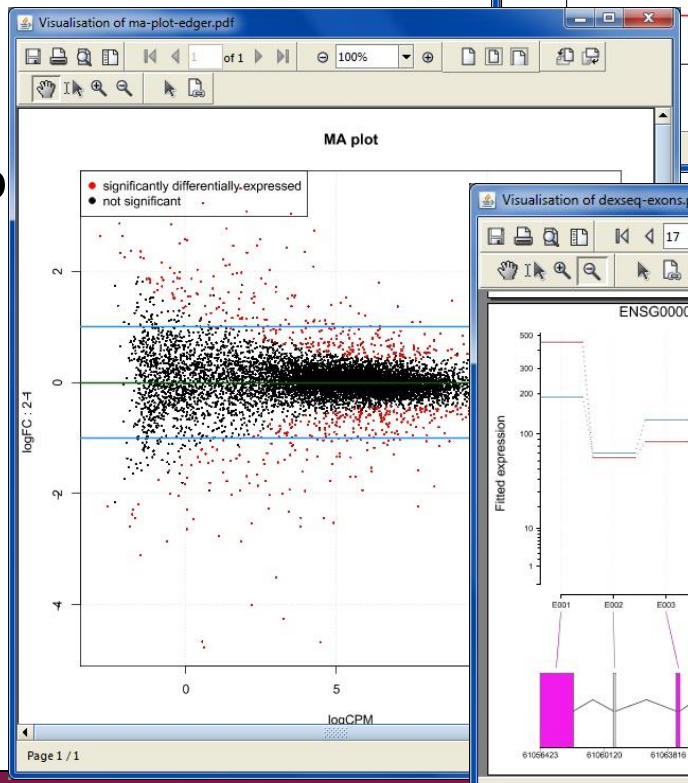


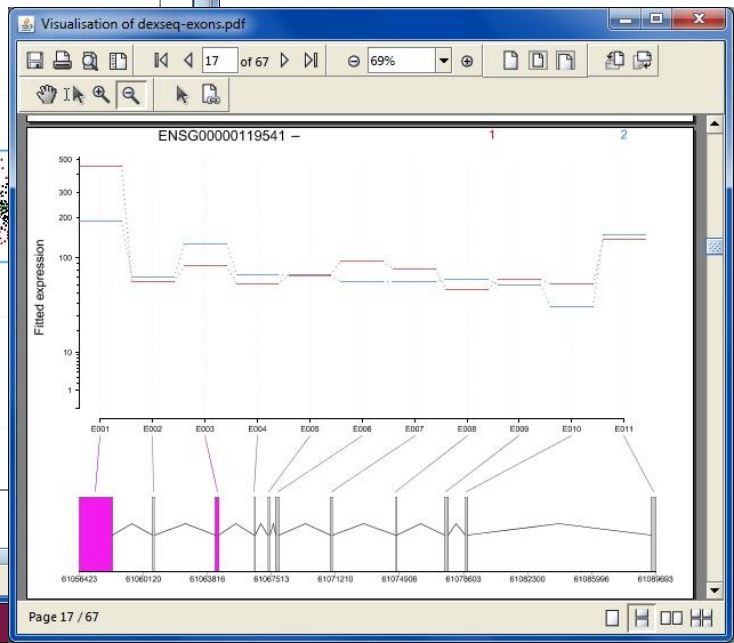


RNA-seq tools

- **Quality control**
  - RseQC
- **Counting**
  - HTSeq
  - eXpress
- **Transcript discovery**
  - Cufflinks
- **Differential expressio**
  - edgeR
  - DESeq
  - Cuffdiff
  - DEXSeq
- **Pathway analysis**
  - ConsensusPathDB

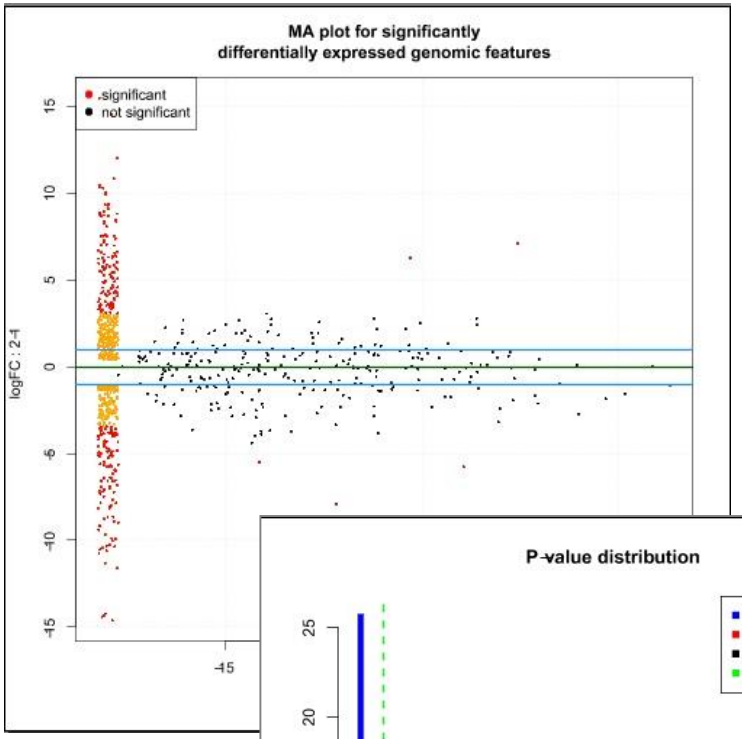


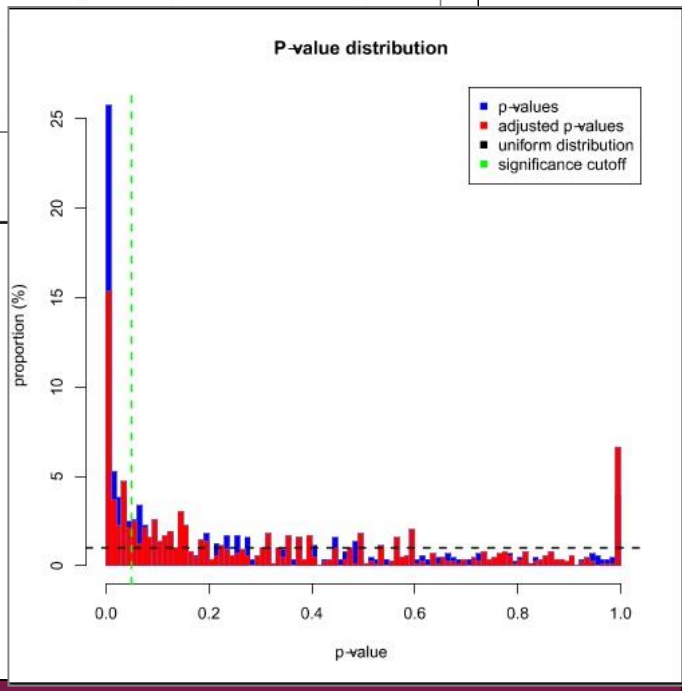




miRNA-seq tools

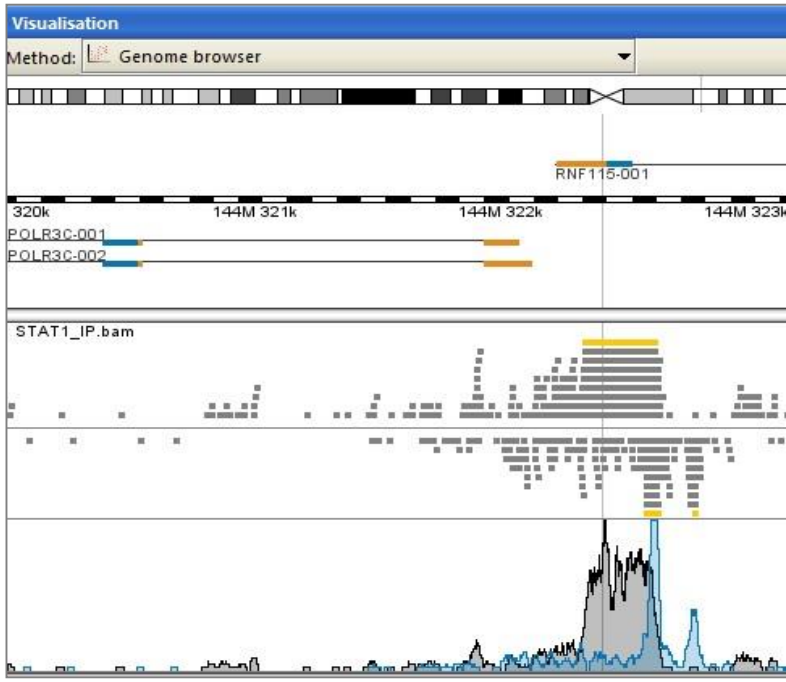
- **Differential expression**
  - edgeR
  - DESeq
- **Retrieve target genes**
  - PicTar
  - miRBase
  - TargetScan
  - miRanda
- **Pathway analysis for targets**
  - GO
  - KEGG
- **Correlate miRNA and target expression**

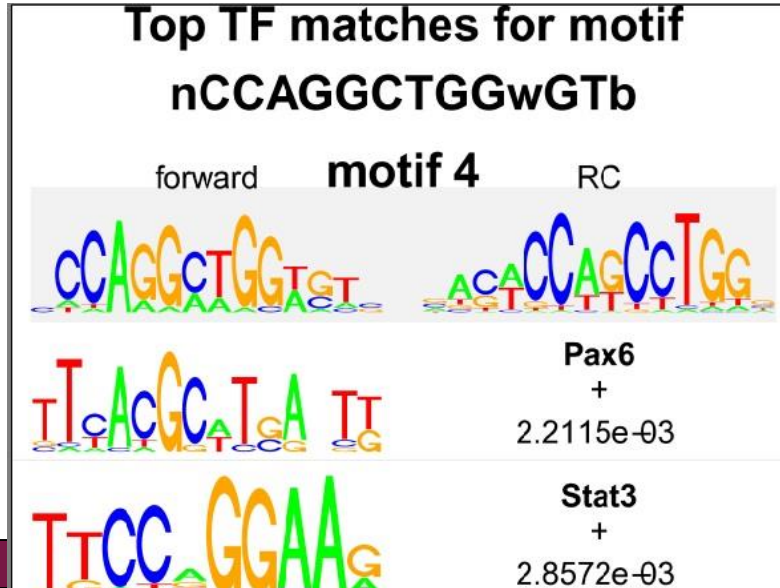




ChIP-seq and DNase-seq tools

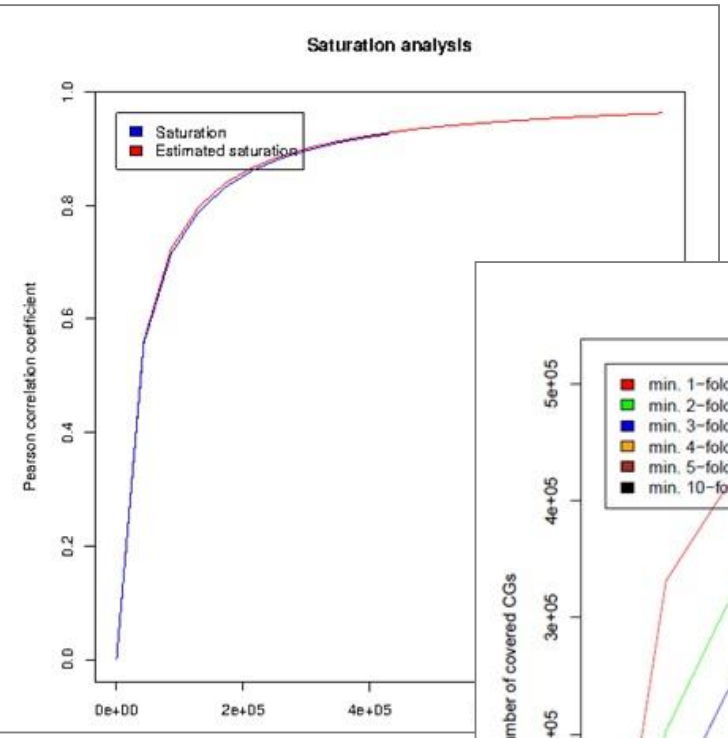
- **Peak detection**
  - MACS
  - F-seq
- **Peak filtering**
  - P-value, no of reads, length
- **Detect motifs, match to JASPAR**
  - MotIV, rGADEM
  - Dimont
- **Retrieve nearby genes**
- **Pathway analysis**
  - GO, ConsensusPathDB

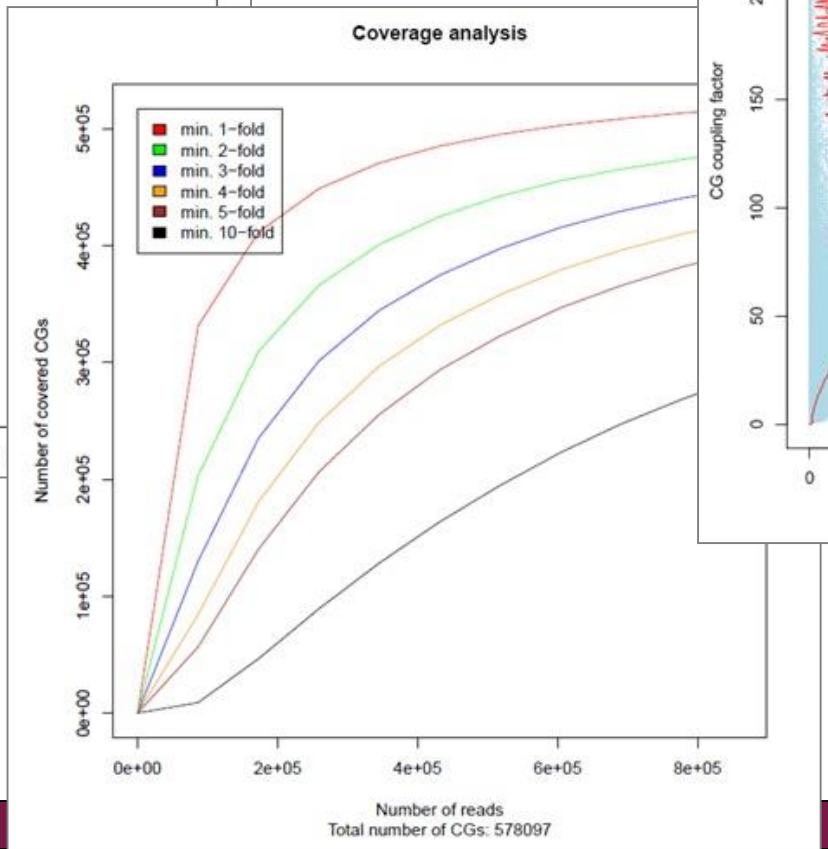


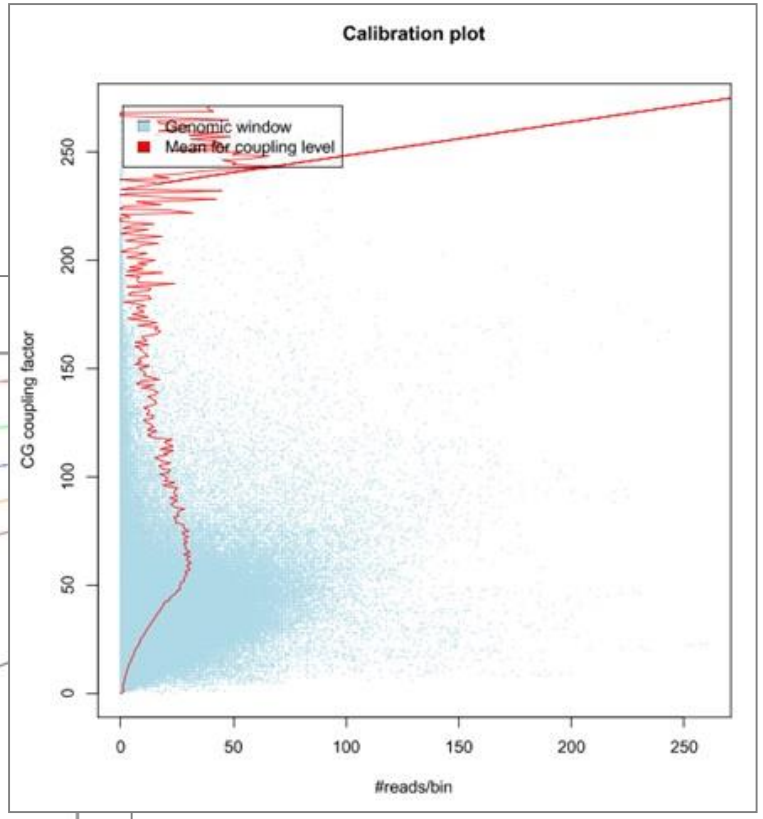


MeDIP-seq tools

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







CSC

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

Visualizing the data

- **Two types of visualizations**

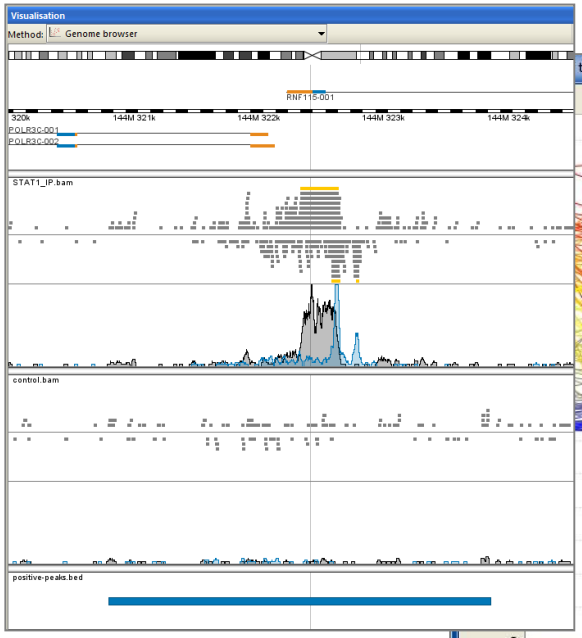
  1. **<u>Interactive visualizations</u> produced by the client program**
     - Select the visualization method from the pulldown menu
     - Save by right clicking on the image

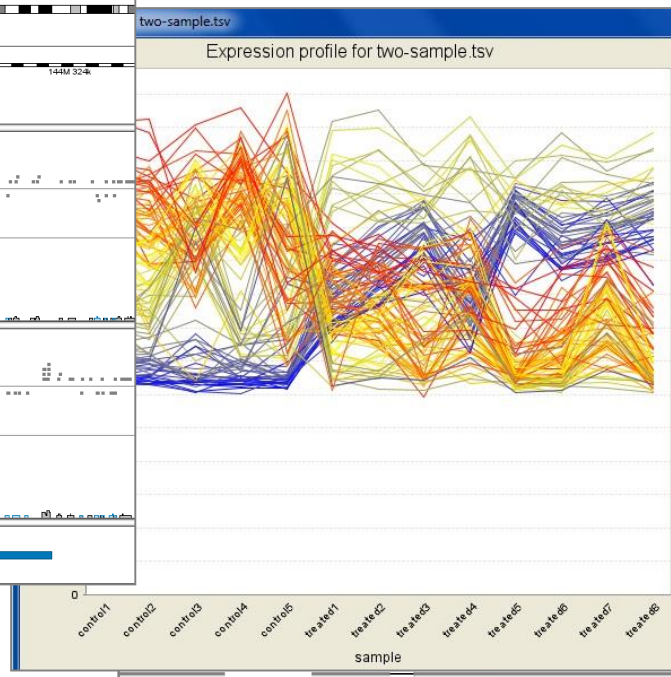
  2. **<u>Static images</u> produced by analysis tools**
     - Select from Analysis tools/ Visualisation
     - View by double clicking on the image file
     - Save by right clicking on the file name and choosing ”Export”

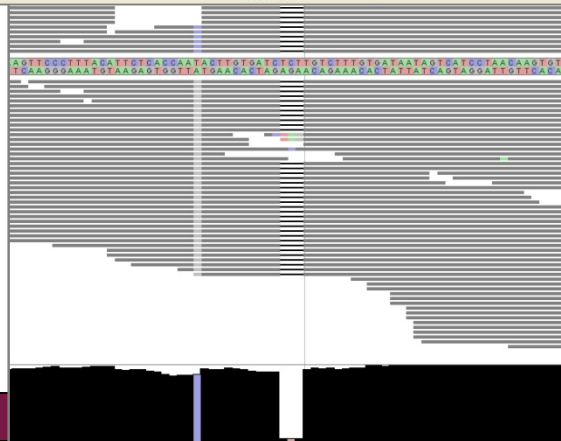
- **Data visualization panel**
  - Maximize and redraw for better viewing
  - Detach = open in a separate window, allows you to view several images at the same time

Interactive visualizations by the client

- **Genome browser**
- **Spreadsheet**
- **Histogram**
- **Venn diagram**
- **Scatterplot**
- **3D scatterplot**
- **Volcano plot**
- **Expression profiles**
- **Clustered profiles**
- **Hierarchical clustering**
- **SOM clustering**

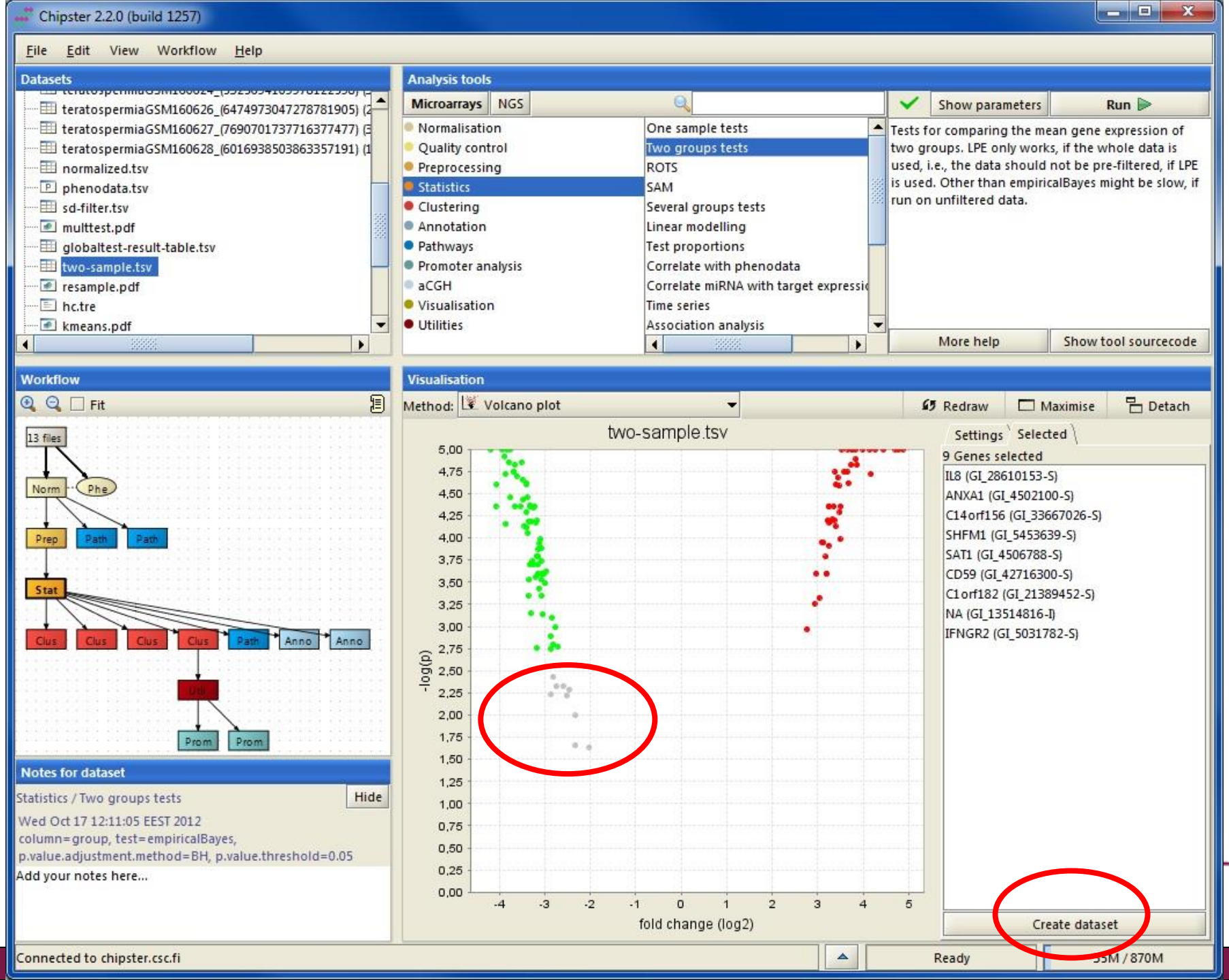






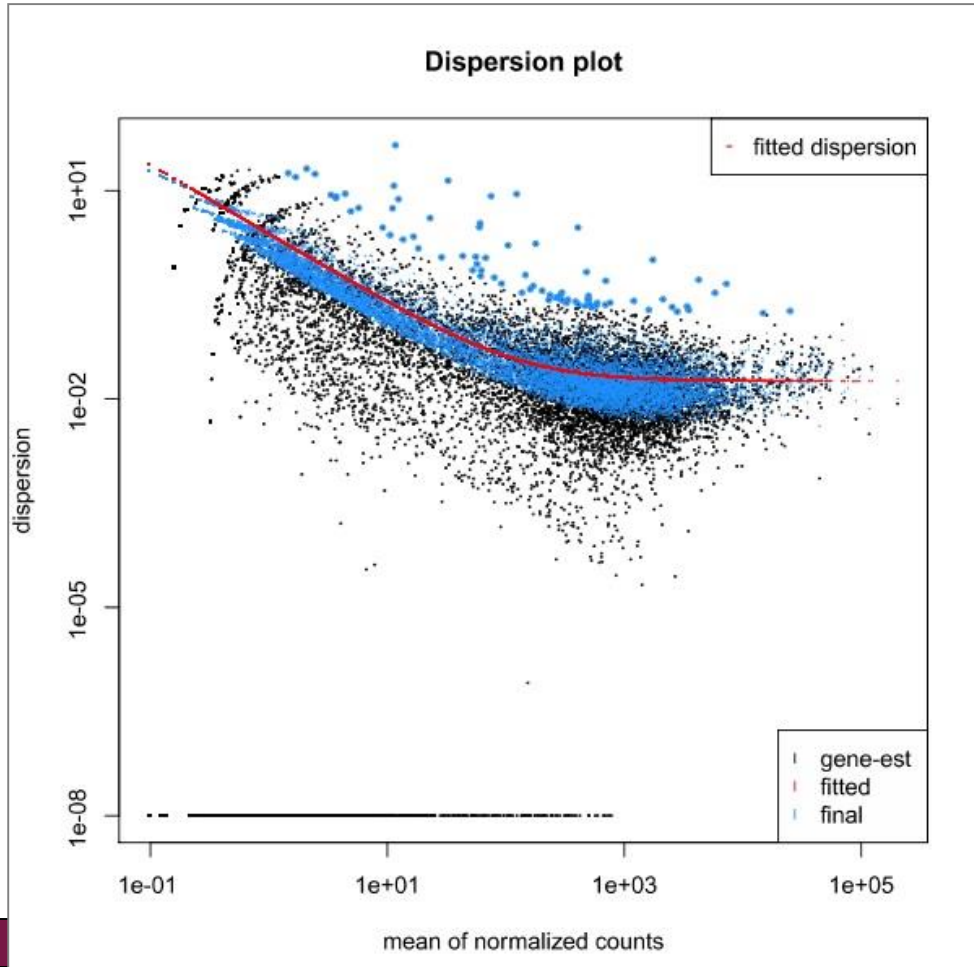
Available actions:

- **Select genes and create a gene list**
- **Change titles, colors etc**
- **Zoom in/out**


Chipster 2.2.0 (build 1257)
File Edit View Workflow Help
Datasets
teratospermiaGSM160626_(6474973047278781905)
teratospermiaGSM160627_(7690701737716377477)
teratospermiaGSM160628_(6016938503863357191)
normalized.tsv
phenodata.tsv
sd-filter.tsv
multtest.pdf
globaltest-result-table.tsv
two-sample.tsv
resample.pdf
hc.tre
kmeans.pdf
Analysis tools
Microarrays NGS
Normalisation
Quality control
Preprocessing
Statistics
Clustering
Annotation
Pathways
Promoter analysis
aCGH
Visualisation
Utilities
One sample tests
Two groups tests
ROTS
SAM
Several groups tests
Linear modelling
Test proportions
Correlate with phenodata
Correlate miRNA with target expressi
Time series
Association analysis
Show parameters
Run
Tests for comparing the mean gene expression of two groups. LPE only works, if the whole data is used, i.e., the data should not be pre-filtered, if LPE is used. Other than empiricalBayes might be slow, if run on unfiltered data.
More help
Show tool sourcecode
Workflow
Fit
13 files
Norm
Phe
Prep
Path
Path
Stat
Clus
Clus
Clus
Clus
Path
Anno
Anno
Util
Prom
Prom
Notes for dataset
Statistics / Two groups tests
Hide
Wed Oct 17 12:11:05 EEST 2012
column=group, test=empiricalBayes,
p.value.adjustment.method=BH, p.value.threshold=0.05
Add your notes here...
Visualisation
Method: Volcano plot
Redraw
Maximise
Detach
two-sample.tsv
-log(p)
fold change (log2)
Settings
Selected
9 Genes selected
IL8 (GI_28610153-S)
ANXA1 (GI_4502100-S)
C14orf156 (GI_33667026-S)
SHFM1 (GI_5453639-S)
SAT1 (GI_4506788-S)
CD59 (GI_42716300-S)
C1orf182 (GI_21389452-S)
NA (GI_13514816-I)
IFNGR2 (GI_5031782-S)
Create dataset
Connected to chipster.csc.fi
Ready

Static images produced by R/Bioconductor

- **Dispersion plot**
- **MA plot**
- **PCA plot**
- **MDS plot**
- **Box plot**
- **Histogram**
- **Heatmap**
- **Idiogram**
- **Chromosomal position**
- **Correlogram**
- **Dendrogram**
- **K-means clustering**
- **etc**



Options for importing data to Chipster

- **Import files/Import folder**
  - Note that you can select several files by keeping the Ctrl key down
  - Compressed files (.gz) are ok
  - FASTQ, BAM, BED, VCF and GTF files are recognized automatically
  - SAM/BAM and BED files can be sorted at the import stage or later
- **Continue with a previous analysis session**
  - Files / Open session
- **Import from URL**
  - Utilities / Download file from URL directly to server
- **Import from SRA database**
  - Utilities / Retrieve FASTQ or BAM files from SRA
- **Import from Ensembl database**
  - Utilities / Retrieve data for a given organism in Ensembl

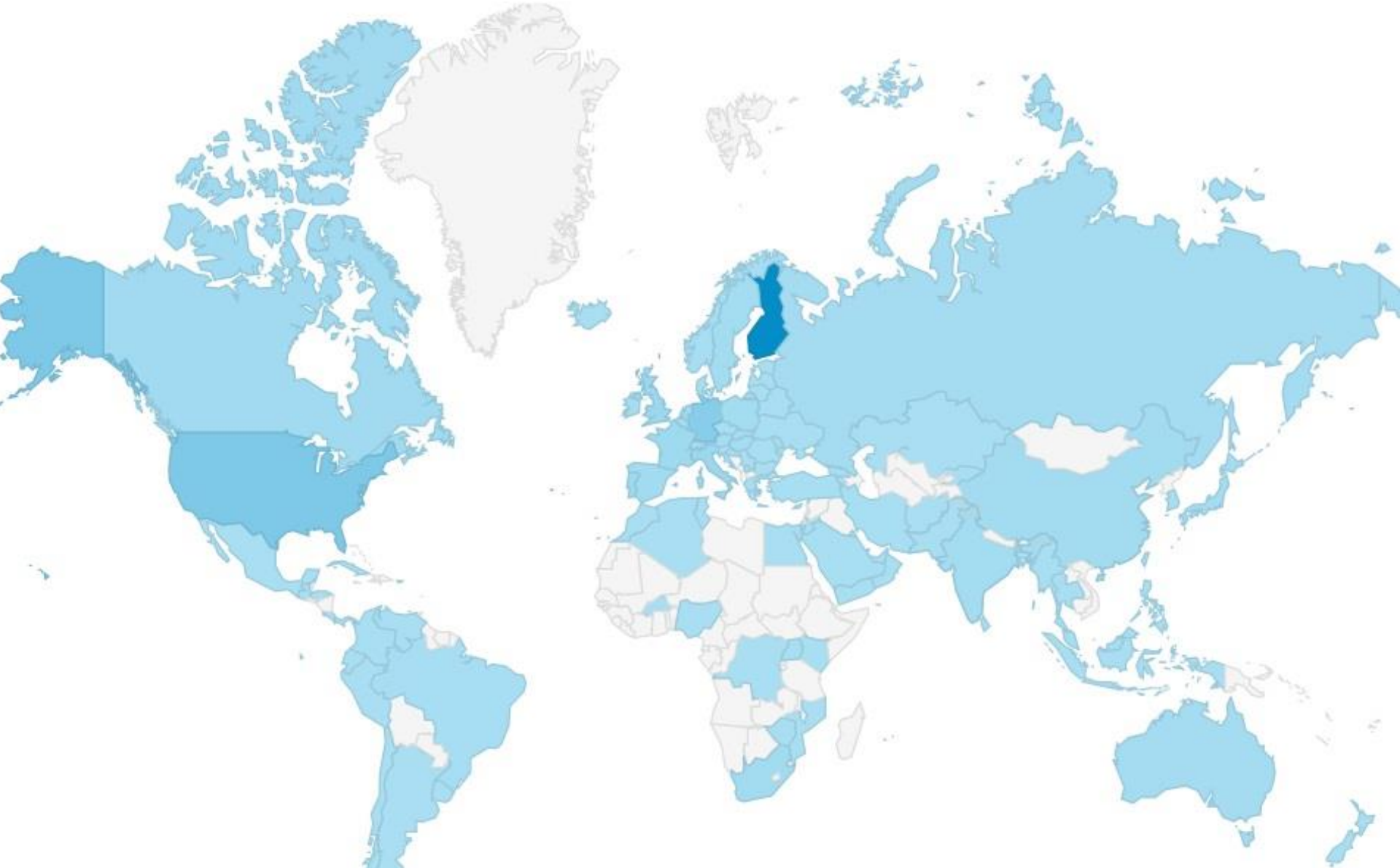
Problems? Send us a support request

-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
```

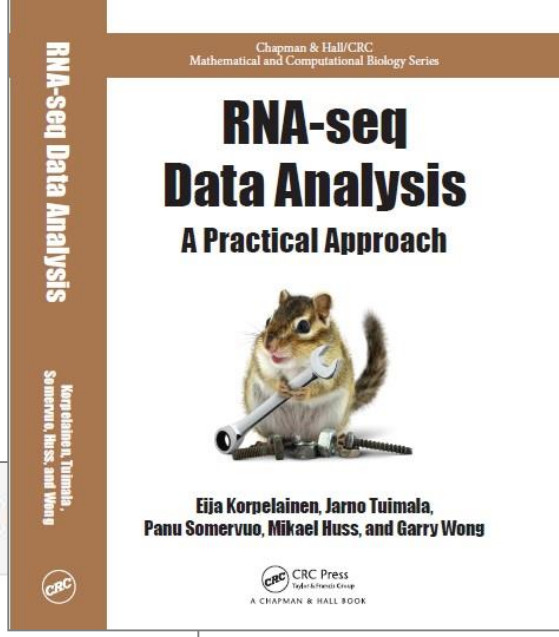
CSC

Acknowledgements to Chipster users and contibutors



More info

- **chipster@csc.fi**
- **http://chipster.csc.fi**
- **Chipster tutorials in YouTube**



GitHub This repository Search Explore Features Enterprise

chipster / chipster

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

7,565 commits 18 branches 123 releases 14 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

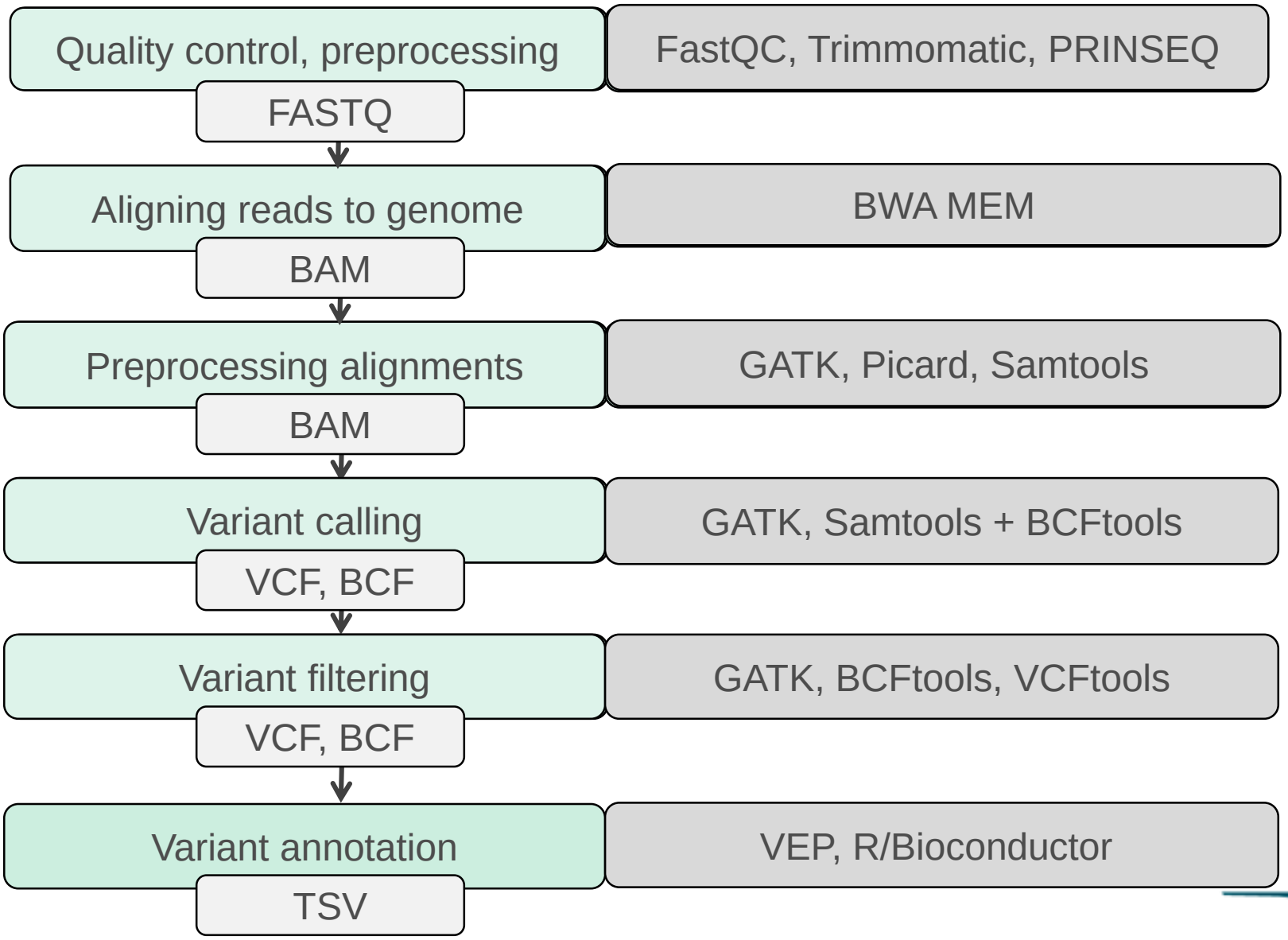
BMC Genomics 2011, **12**:507 doi:10.1186/1471-2164-12-507

Detecting variants in sequencing data

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Match sets of genomic regions**
- **Annotate variants**

Variant analysis workflow



CSC

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

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**
- **FastX**
- **PRINSEQ**
- **TagCleaner**
- **...**

Raw reads: FASTQ file format

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

    @read name

    GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

    + read name

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

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

CSC

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 so that 33 is added to the Phred score**
  - E.g. 39 is encoded as ”H”, the 72nd ASCII character (39+33 = 72)
  - 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.....................................

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

S - Sanger        Phred+33,  raw reads typically (0, 40)
```

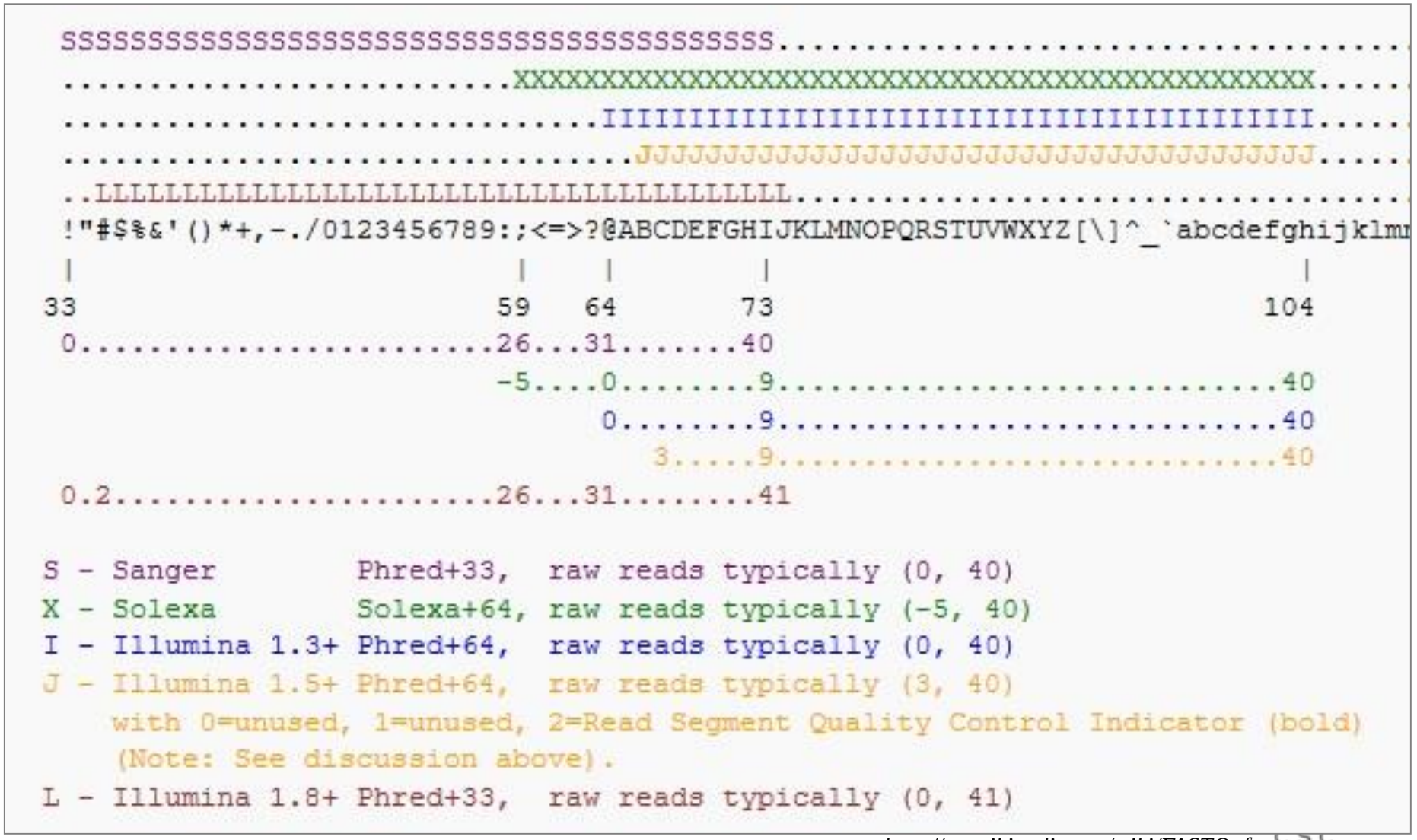
http://en.wikipedia.org/wiki/FASTQ_format

Base quality encoding systems

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
  ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmn
  |                         |    |        |                                    |
 33                        59   64       73                                  104
  0........................26...31.......40

  0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)

L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

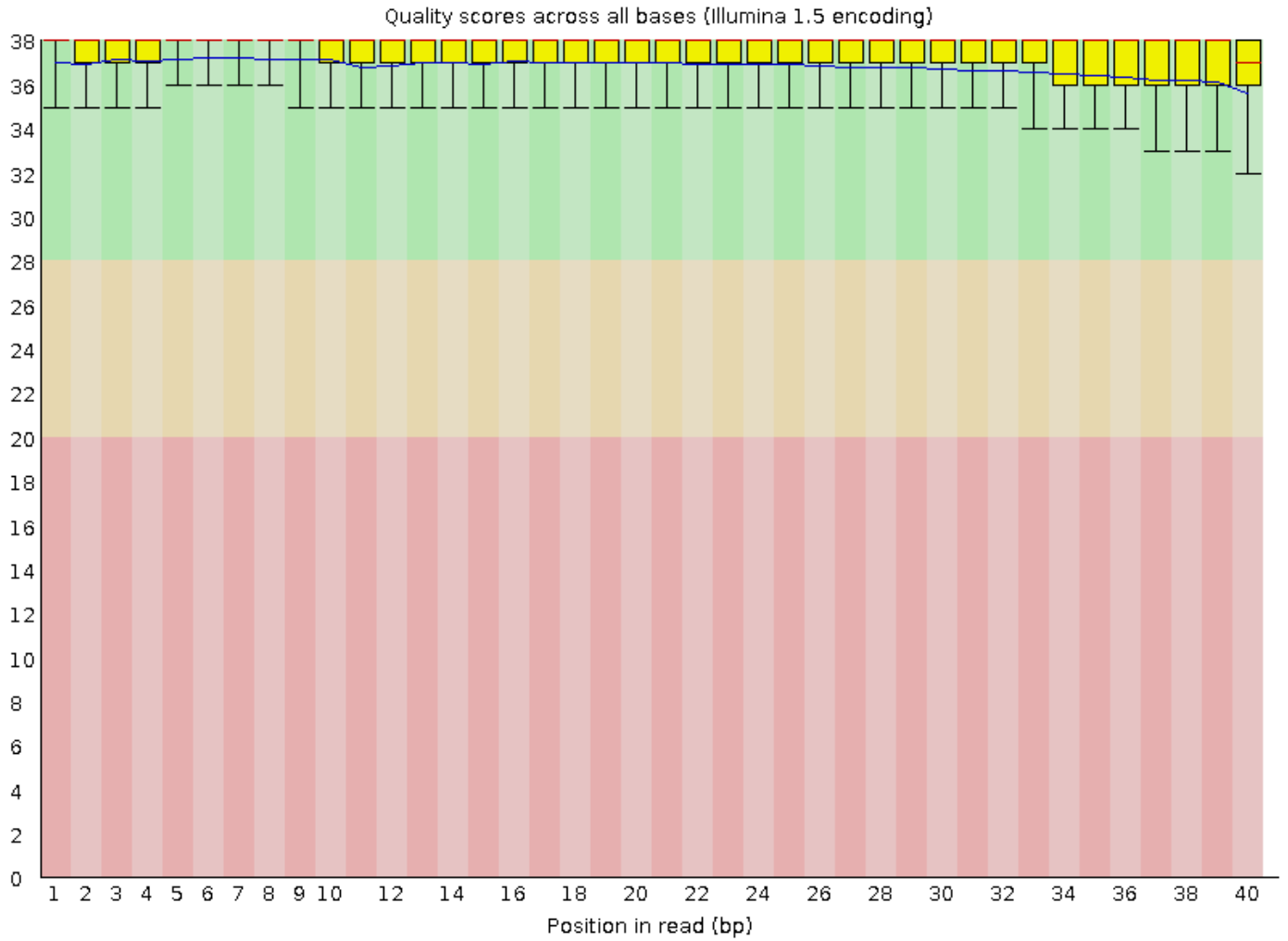
http://en.wikipedia.org/wiki/FASTQ_format

Base quality encoding systems



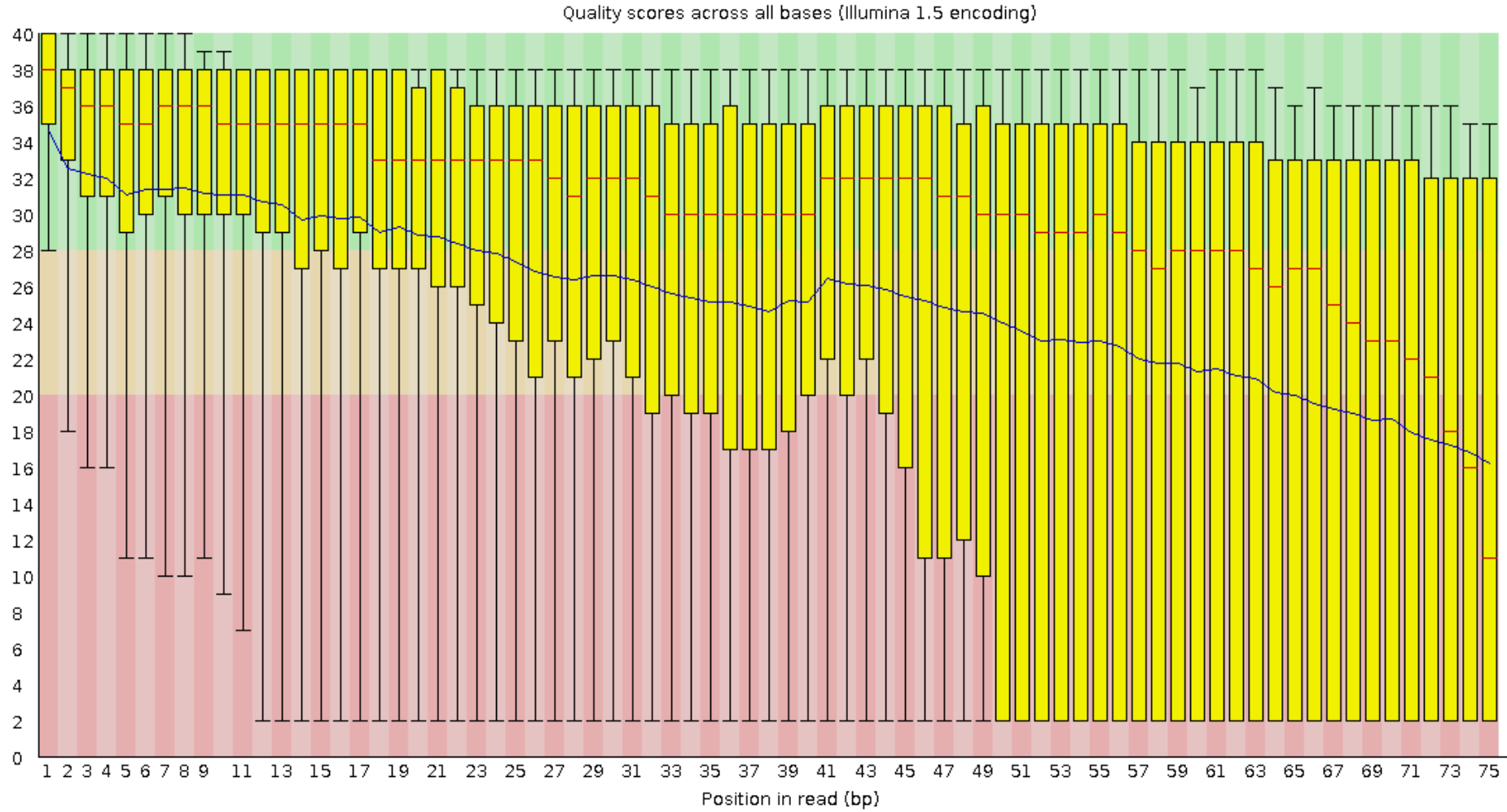
http://en.wikipedia.org/wiki/FASTQ_format

Per position base quality (FastQC)

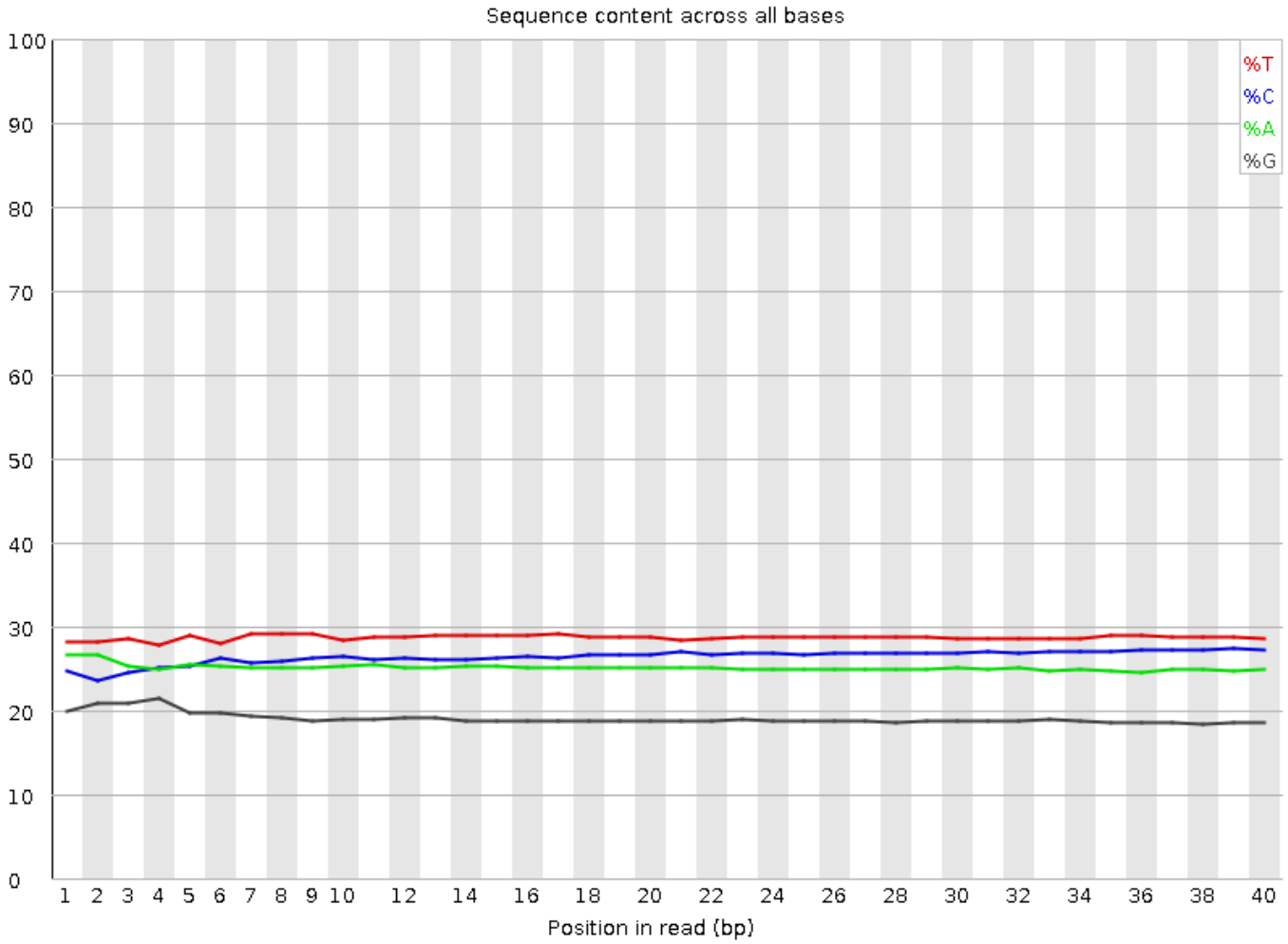


CSC

Per position base quality (FastQC)

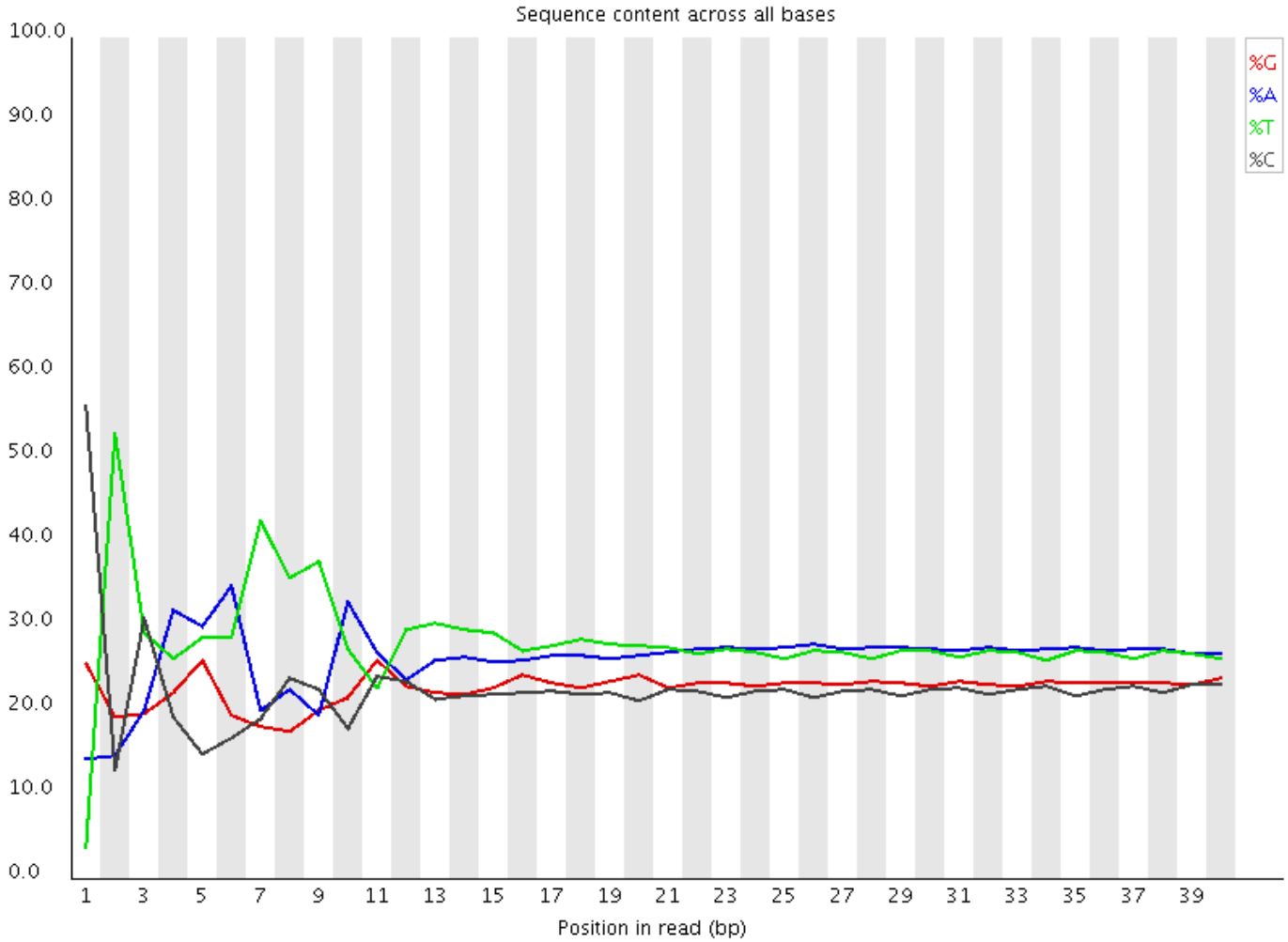


Per position sequence content (FastQC)



CSC

Per position sequence content (FastQC)



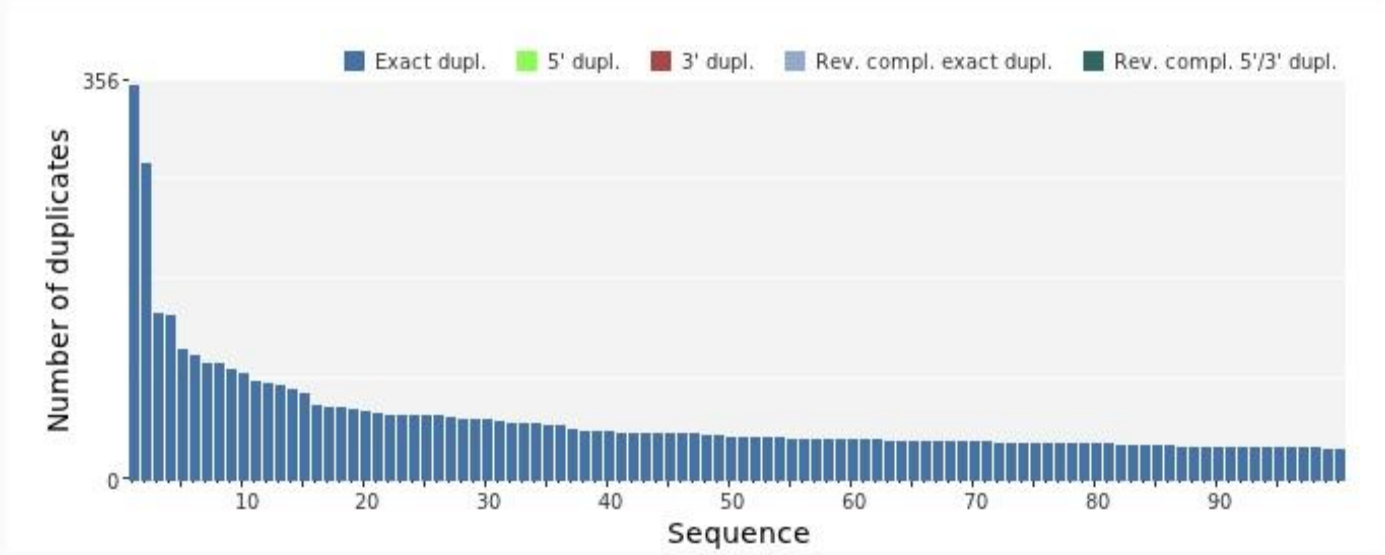
- **Sequence specific bias: Correct sequence but biased location, typical for Illumina RNA-seq data**

CSC

Duplicate reads plot (PRINSEQ)

SEQUENCE DUPLICATION

	# Sequences	Max duplicates
Exact duplicates	29478 (29.74%)	353
5' duplicates	0	0
3' duplicates	0	0
Exact duplicates with reverse complements	2 (0.00%)	1
5'/3' duplicates with reverse complements	0	0
Total	29480 (29.74%)	-



CSC

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

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**
- **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 removed as well

What base quality threshold should be used?

- **No consensus yet**
- **Trade-off between having good quality reads and having enough sequence**

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

Cristian Del Fabbro[1☯], Simone Scalabrin[2☯], Michele Morgante[1], 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

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

Trimmomatic options in Chipster

- **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) read quality
- **Trim x bases from left/ right**
- **Adapters**
- **Minimum read length after trimming**
- **Copes with paired end data**

CSC

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
- **Copes with paired end data**

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 x bases from left/ right**
- **Trim to length x**
- **Trim polyA/T tails**
  - Minimum number of A/Ts
  - From left or right
- **Copes with paired end data**

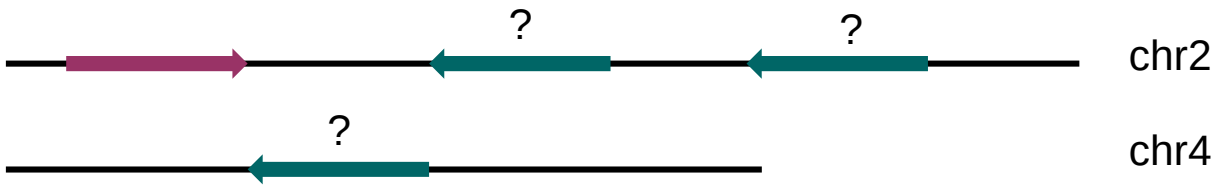
CSC

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Manipulate alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Alignment to reference genome

- **The goal is to find out where a read originated from**
- **Challenges**
  - Variants
  - Sequencing errors
  - Repeated sequence in genome
  - Low complexity sequences
  - Reference errors and gaps
- **Paired end reads can help**



CSC

Tens of aligners are available

- **http://wwwdev.ebi.ac.uk/fg/hts_mappers/**

BIOINFORMATICS **REVIEW**

Vol. 28 no. 24 2012, pages 3169–3177
doi:10.1093/bioinformatics/bts605

Sequence analysis

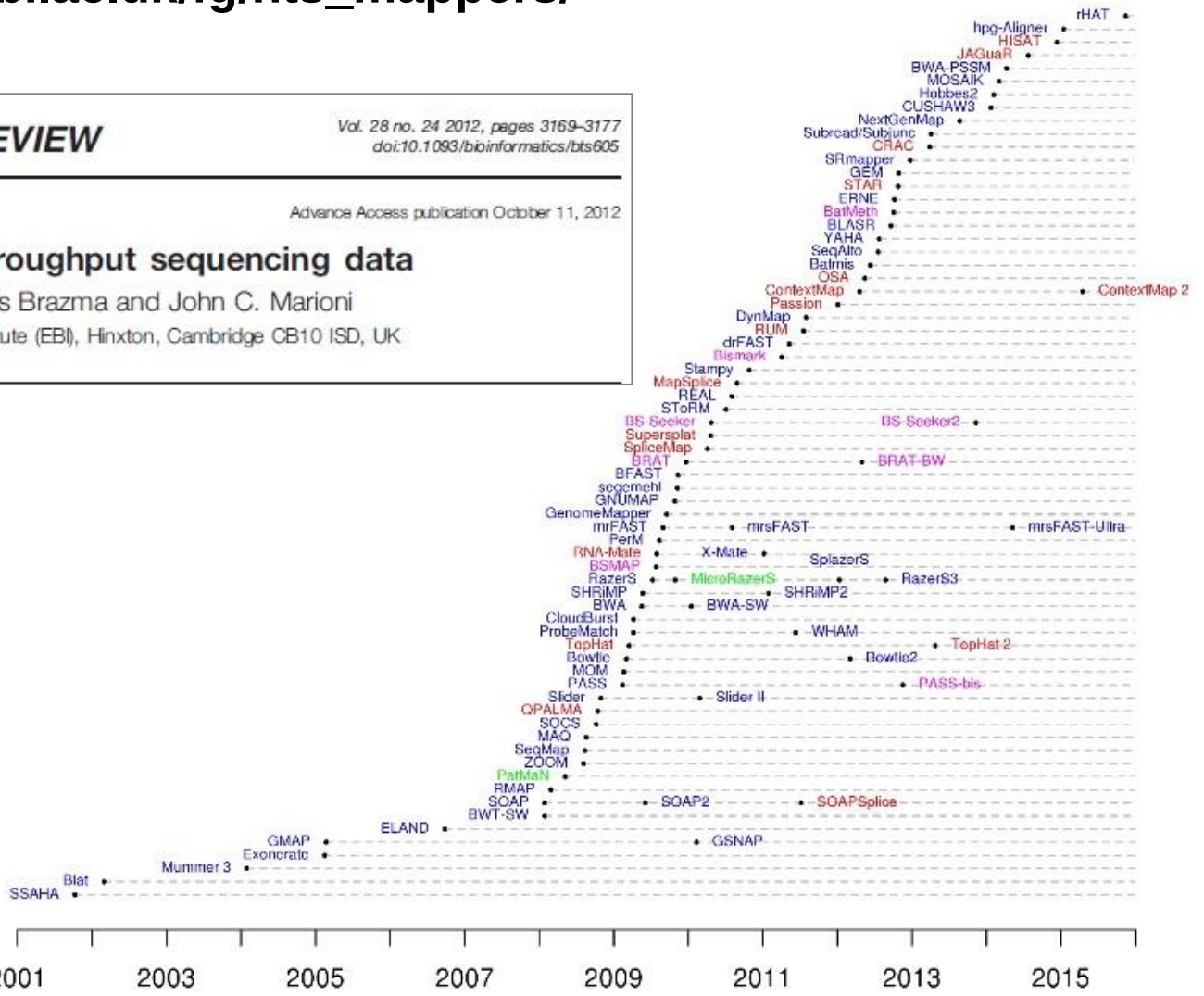
Advance Access publication October 11, 2012

Tools for mapping high-throughput sequencing data

Nuno A. Fonseca*, Johan Rung, Alvis Brazma and John C. Marioni

EMBL Outstation, European Bioinformatics Institute (EBI), Hinxton, Cambridge CB10 ISD, UK

Associate Editor: Jonathan Wren



BWA MEM

- **Burrows-Wheeler Aligner Maximal Exact Match**
- **Works well with long reads (from 70 bp to a few megabases)**
- **Robust to sequencing errors**
- **Reference genome is indexed**
  - Allows fast searching
- **Chooses automatically between local and end-to-end alignment**
  - Soft clips read ends which don't match

BWA MEM steps

1. **Looks for maximal exact matches (MEM), “seeds”**
   - exact match that cannot be extended further to either direction
   - you can set the minimum length of MEM, by default it is 19
2. **Chains nearby seeds together**
3. **Extends seeds by allowing mismatches and gaps**
   - Smith-Waterman alignment, “SW”

Seed extension: clipping penalty

```
CGATG--GCTAGCATAGCTAGAGTTC
  |||  ||||||||||||||||
  ATGATGCTAGCATAGCTAGACAC
```

Under the BWA-MEM scoring, the best local hit is the green region with both ends clipped off. True variants may be clipped.

BWA-MEM gives a bonus to an extension reaching the end. It may prefer to reach the left end.

*files.figshare.com/1421612/**mem**_poster.pdf* by Heng Li

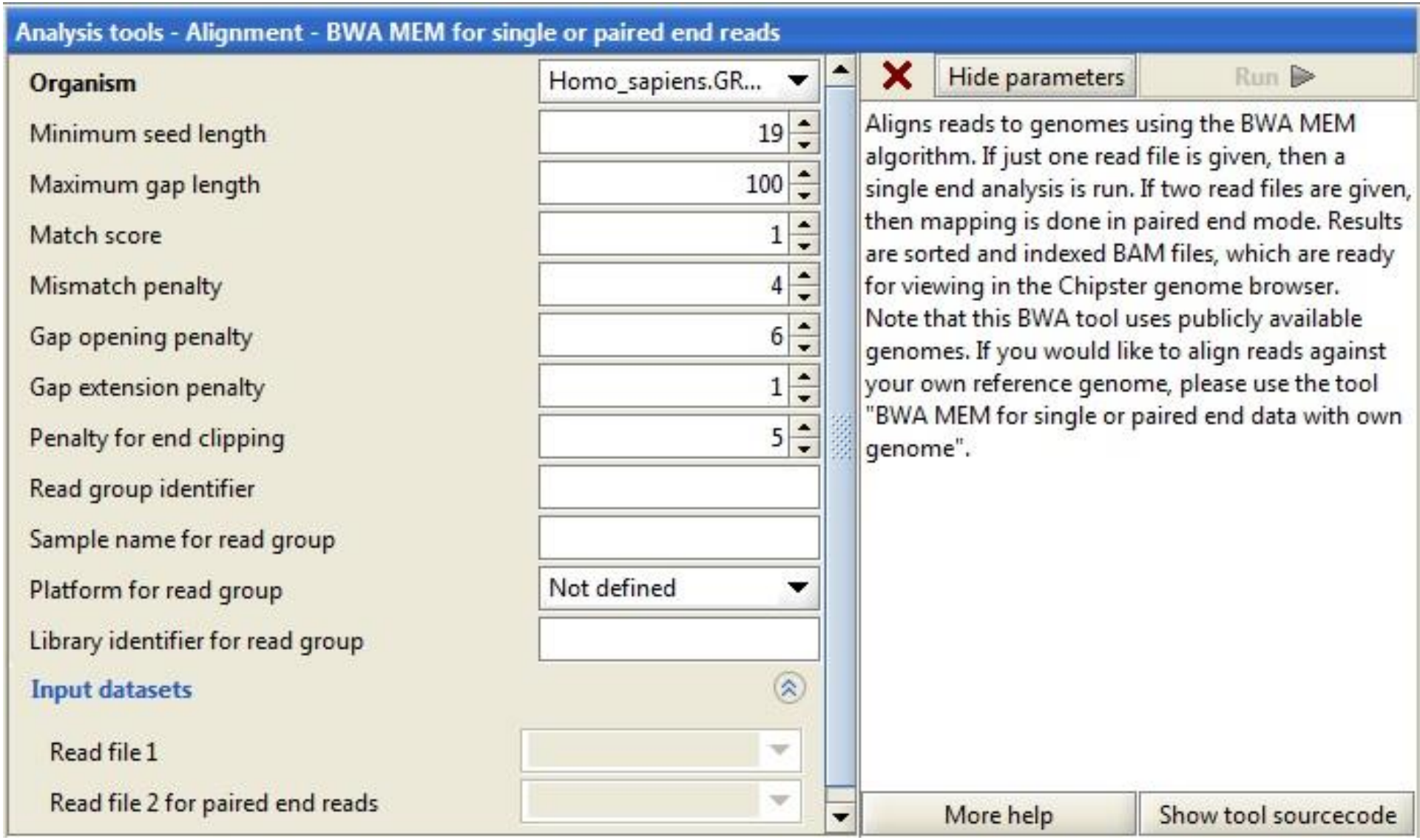
CSC

BWA MEM and paired end reads

- **Reads are first aligned separately**
- **Aligned reads are paired**
  - Considers alignment scores, insert size and the possibility of chimera when pairing.
- **If one of the reads in a pair did not align during the single end alignment step, BWA tries to rescue it now using a more thorough SW alignment**

BWA MEM in Chipster

- **Reference genome indexes available for a number of genomes**
  - Let us know if other reference genomes need to be added
- **You can use your own reference too, but indexing may take some hours**
  - Supply reference genome in FASTA format
- **The following parameters are available**
  - Match score, penalties for mismatch, gap open and extension
  - Clipping penalty
  - Minimum exact match
  - Maximum gap length
  - Read group information (needed for variant calling if one sample has reads in several files)
- **The output is a coordinate-sorted BAM file which is indexed**


Analysis tools - Alignment - BWA MEM for single or paired end reads
Organism
Homo_sapiens.GR...
Minimum seed length
19
Maximum gap length
100
Match score
1
Mismatch penalty
4
Gap opening penalty
6
Gap extension penalty
1
Penalty for end clipping
5
Read group identifier
Sample name for read group
Platform for read group
Not defined
Library identifier for read group
Input datasets
Read file 1
Read file 2 for paired end reads
Hide parameters
Run
Aligns reads to genomes using the BWA MEM algorithm. If just one read file is given, then a single end analysis is run. If two read files are given, then mapping is done in paired end mode. Results are sorted and indexed BAM files, which are ready for viewing in the Chipster genome browser.
Note that this BWA tool uses publicly available genomes. If you would like to align reads against your own reference genome, please use the tool "BWA MEM for single or paired end data with own genome".
More help
Show tool sourcecode

CSC

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)

- **Note that even if a read is mapped to its correct genomic position and it has good mapping quality, individual bases can still be misaligned, leading to false SNPs**

```
ref     aggttttataaaac----aattaagtctacagagcaacta
sample  aggttttataaaacAAATaattaagtctacagagcaacta
read1   aggttttataaaac****aaAtaa
```

File format for mapped reads: BAM/SAM

- SAM (Sequence Alignment/Map) is a tab-delimited text file containing read alignment data. BAM is a binary form of SAM.

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   566152541681648866(6(6(6261?8==(B=513);(/BB=141=>6>?=<=?B>9B?>BA<66>BA>BB
CC:Z:chr15MD:Z:40C34XG: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
```

optional header section

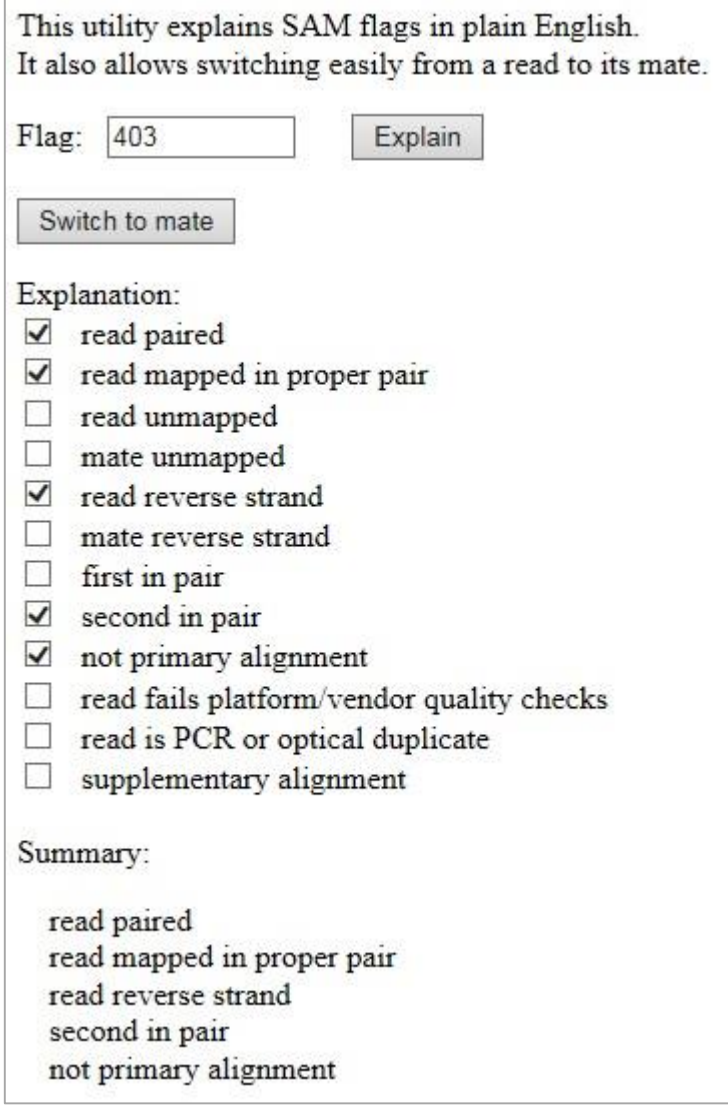
alignment section: one line per read,
containing 11 mandatory fields, followed by optional tags

Fields in BAM/SAM files

- **read name** 206B4ABXX100825:7:8:4978:179024
- **flag** 161
- **reference name (chr)** 20
- **position** 3020970
- **mapping quality** 60
- **CIGAR** 76M
- **reference name of mate** =
- **mate position** 3021227
- **insert size** 326
- **sequence**
  TAGAGCATGACCTGAAATGTTGTGAAAGTGCAGATTTTGTAGGTAA
  GATGGCTTCAGATTCGCCCTGAACCAAGGG
- **base qualities**
  AAGBDDCBJDHI9KEGGDJDEJBGEHGH>LJEKFECE:JC=H6@C8G=
  CGCGFD9/;-;<><DB->A:6?HD0H4=
- **tags** MD:Z:76 NM:i:0 AS:i:76 XS:i:0

Flag field in BAM

- **Information whether the read is**
  - Mapped
  - Paired
  - Mapped in proper pair
  - Located in reverse strand
  - Duplicate
  - Primary alignment
  - From file 1 or file 2
- **Read's flag number is a sum of values**
  - E.g. 4 = unmapped, 1024 = duplicate



- http://samtools.github.io/hts-specs/SAMv1.pdf
- http://broadinstitute.github.io/picard/explain-flags.html

CSC

CIGAR string

- M = match or mismatch
- I = insertion
- D = deletion
- S = soft clip (“these bases don’t map in the reference, ignore them”)
- H = hard clip (“these bases map somewhere else in the reference, ignore and remove them from this alignment”)
- N = intron (in RNA-seq read alignments)

- 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
```

CSC

BAM 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).**

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

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**
- **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**
- **Make a subset of BAM**
  - Retrieves alignments for a given chromosome/region, e.g. chr1:100-1000.
  - Includes an optional mapping quality filter.

Count alignment statistics for BAM

- **Based on Samtools flagstat command**
- **Example output:**

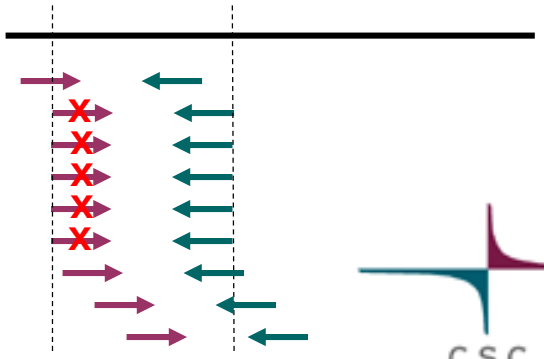
52841623 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
52841623 + 0 mapped (100.00%:-nan%)
52841623 + 0 paired in sequencing
28919461 + 0 read1
23922162 + 0 read2
42664064 + 0 properly paired (80.74%:-nan%)
44904884 + 0 with itself and mate mapped
7936739 + 0 singletons (15.02%:-nan%)
999152 + 0 with mate mapped to a different chr
357082 + 0 with mate mapped to a different chr (mapQ>=5)

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Marking duplicates

- **Duplicate reads arise when the same DNA fragment is sequenced several times**
  - PCR duplicates and optical duplicates
- **Duplicates are defined here as <u>paired reads which have the same outer coordinates</u>**
  - Note that the read sequence or length doesn't need to be identical
- **Duplicate reads should not be used in variant calling**
  - Not independent measurements
  - Errors get propagated to all duplicates → false SNPs
- **Marking duplicates with a flag in BAM allows them to be ignored**
- **Tools in Chipster**
  - Mark duplicates in BAM (Picard)
  - Remove duplicates in BAM (Samtools)



Identify duplicates using orientation + “unclipped” 5’ position

Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	–	C	CAG		A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

Blue maps to forward strand
Orange maps to reverse strand
Grey bases are clipped

Underlined is the expected 5’ start of the read, given the mapping

So…what are the duplicate sets?

https://www.broadinstitute.org/gatk/events/slides/1503/GATKwh6-BP-1A-Mapping_and_Dedupping.pdf

CSC

Misaligned bases near INDELs

- **Even if a read is mapped to its correct genomic position and it has good mapping quality, individual bases can still be misaligned, leading to false SNPs**

- **Typical when there is an INDEL near the ends of reads, because creating a mismatch in an alignment is ”cheaper” than creating a gap**
  - E.g. in BWA MEM the default mismatch penalty is -4, gap opening penalty is -6 and gap extension penalty is -1.

Misalignment around INDELs → false SNPs

```
ref     aggttttataaaac----aattaagtctacagagcaacta
sample  aggttttataaaacAAATaattaagtctacagagcaacta
read1   aggttttataaaac****aaAtaa
read2    ggttttataaaac****aaAtaaTt
read3        ttataaaacAAATaattaagtctaca
read4            CaaaT****aattaagtctacagagcaac
read5              aaT****aattaagtctacagagcaact
read6                T****aattaagtctacagagcaacta
```

Local realignment

- **Realigning a short stretch around an INDEL can solve the problem**
  - Original alignment is an alignment between two sequences
  - Realignment makes use of all the reads covering that position (multiple sequence alignment, can consider base qualities as well)
- **Two approaches**
  - Samtools marks bases with base alignment quality (BAC)
  - GATK can perform (computationally heavy) realignment
  - You can do GATK realignment even if you are using Samtools for variant calling

After realignment

```
ref     aggttttataaaac----aattaagtctacagagcaacta
sample  aggttttataaaacaaataattaagtctacagagcaacta
read1   aggttttataaaacaaataa
read2    ggttttataaaacaaataatt
read3        ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                    taattaagtctacagagcaacta
```

GATK realignment has two options

1. **Realign at known INDELs only**
   - Fast
   - Low coverage is ok
   - Need known INDELs
2. **Use mismatching bases to determine if a site should be realigned. Then realign all those sites.**
   - Slow
   - Needs sufficient coverage
   - Can also include known INDELs in this

➢ **Note that realignment is not necessary if you use an assembly based variant caller (like GATK's HaplotypeCaller)**

Base quality recalibration

- **Base quality values might be over- or under-estimated.**
  - This is a problem because they are used extensively in variant calling.
- **Quality values can be re-calibrated with GATK**
  - Easier if you have a set of known variants (e.g. dbSNP)

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Detecting variants (SNPs, INDELs,…)

- **Compare reads to the reference genome and look for differences**

- **BUT differences can be caused also by sequencing and mapping errors, so we need to consider**
  - Sequencing depth (coverage)
  - Base quality of each base which support the variant
  - Mapping qualities of the reads
  - Proximity to INDELs
  - Proximity to homopolymer runs (> 10 for Illumina)
  - Individual vs multi-sample calling

- **Need to balance**
  - sensitivity (should not miss real variants)
  - specificity (should not report sequencing and mapping errors)

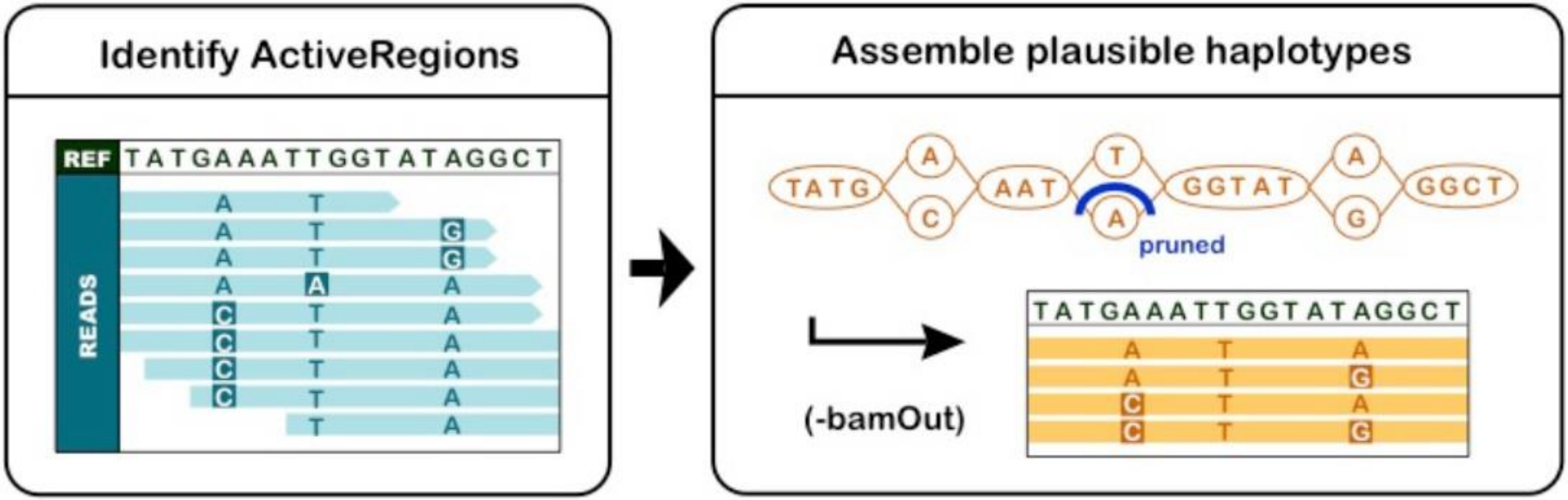
Variant calling tools

- **Samtools + bcftools**
  - Position based calling
  - Integrated in Chipster
- **GATK**
  - HaplotypeCaller (assembly based calling)
  - UnifiedGenotyper (position based calling)
  - MuTect (for somatic variant discovery)
  - Licensing limits GATK integration to Chipster
- **FreeBayes**

- **For intersting comparisons, see the Blue Collar Bioinformatics blog by Brad Chapman**
  - https://bcbio.wordpress.com/

Variant calling with GATK HaplotypeCaller

- **Per-sample variant calling followed by joint genotyping across samples: Scalable and incremental**
- **Assembly-based variant calling with HaplotypeCaller**
  - More accurate than position-based callers
  - When it encounters a region of variation, it discards the existing alignment information and completely reassembles the reads
  - Does not need local realignment around INDELs (which is still needed for base quality recalibration though)
  - Runs per-sample to generate gVCF, which is then used by the GenotypeGVCFs tool for joint genotyping of multiple samples



Variant calling with Samtools and bcftools

1. **Samtools mpileup command**
   - calculates genotype likelihoods for each mismatching position and stores the likelihoods in the BCF format. It doesn't call variants
2. **Bcftools call command**
   - calls variants using the genotype likelihoods calculated by Samtools

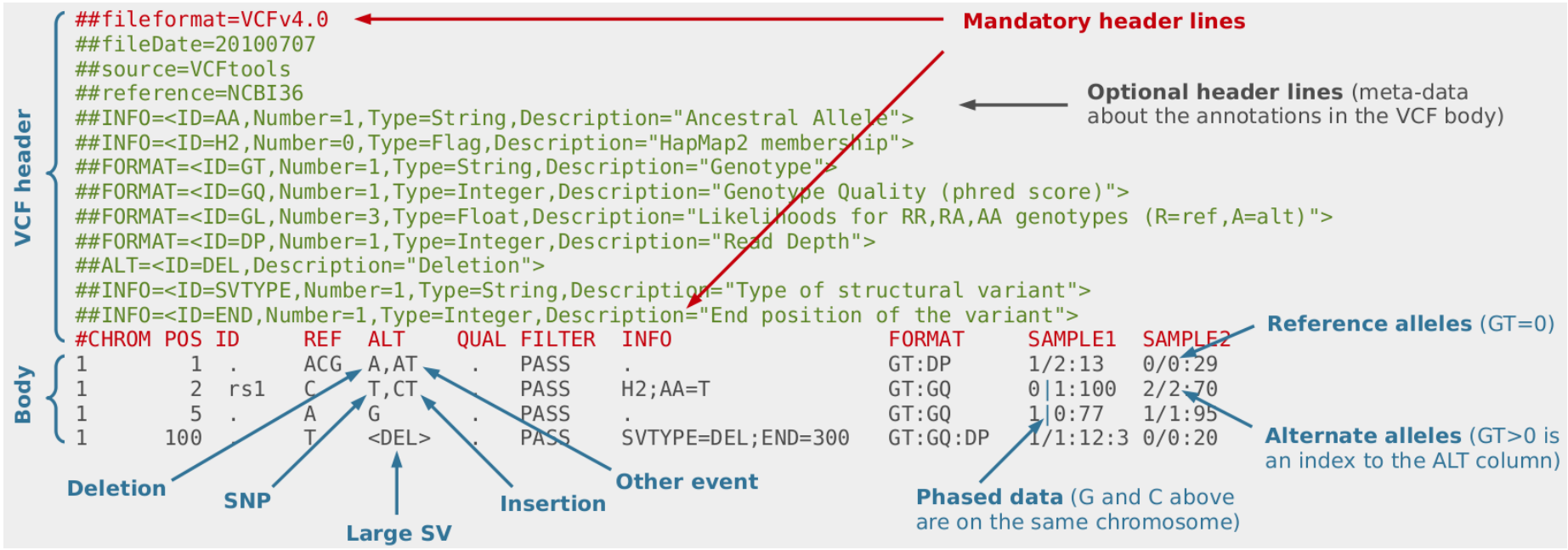
- **Incorporates different types of information, e.g.**
  - number of reads that share a mismatch
  - base quality data
  - expected sequencing error rates
- **Combined in the Chipster tool "Call SNPs and short INDELs"**
- **http://samtools.sourceforge.net/mpileup.shtml**

Call SNPs and short INDELs: parameters

- **Reference sequence**
- **Format of chromosome names (chr1 or 1?)**
- **Call only SNPs, INDELs not considered**
- **Minimum and maximum read depth**
  - Maximum should be about twice the average read depth
  - Note that data from individuals is pooled
- **Minimum mapping quality for an alignment to be used**
- **Minimum base quality for a base to be considered**
- **Disable probabilistic realignment for computation of BAC**
- **Downgrading coefficient for mapping quality**
  - Reduce the effect of reads with excessive mismatches
  - 0: turn the functionality off; 50: suggested value for BWA
- **Call variants for a certain region only**
- **Output per sample read depth and number of high-quality non-reference bases**
- **Output all INFO fields**

CSC

Variant call file format (VCF)



- **BCF is a binary form of VCF (like BAM is a binary form of SAM)**
  - Same content but more compact

CSC

Variant call format (VCF) files

- **Header**
  - VCF format version
  - Program and reference genome used
  - definitions of variant call description fields (INFO, FORMAT)
- **Variant calls, one line with several fields for each variant:**
  - CHROM chromosome name
  - POS leftmost position of the variant
  - ID identifier of the variant
  - REF reference allele
  - ALT variant allele
  - QUAL variant quality
  - FILTER filters applied
  - INFO information about the variant, separated by semi-colon
  - FORMAT format of the genotype fields, separated by semi-colon
  - SAMPLE 1-n sample columns containing genotypes etc

- **See https://samtools.github.io/hts-specs/VCFv4.2.pdf**

VCF header: tool and reference data used for producing the file

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##samtoolsVersion=1.2+htslib-1.2.1
##samtoolsCommand=samtools mpileup -u -C 0 -q 50 -Q 20 -t DP -t DV -f /opt/chipster/tools/genomes/fasta/
   Homo_sapiens.GRCh38.fa alignment001.bam alignment002.bam alignment003.bam
##reference=file:///opt/chipster/tools/genomes/fasta/Homo_sapiens.GRCh38.fa
##contig=<ID=1,length=248956422>
##contig=<ID=22,length=50818468>
##contig=<ID=X,length=156040895>
##contig=<ID=Y,length=57227415>
##contig=<ID=MT,length=16569>
##ALT=<ID=X,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Number of high-quality bases">
##FORMAT=<ID=DV,Number=1,Type=Integer,Description="Number of high-quality non-reference bases">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as l
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and al
##bcftools_callVersion=1.2+htslib-1.2.1
##bcftools_callCommand=call -vmO z –
#CHROM  POS ID REF ALT QUAL FILTER INFO FORMAT alignment001.bam alignment002.bam alignment003.ba
```

VCF header: INFO about all samples, FORMAT for each sample

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##samtoolsVersion=1.2+htslib-1.2.1
##samtoolsCommand=samtools mpileup -u -C 0 -q 50 -Q 20 -t DP -t DV -f /opt/chipster/tools/genomes/fasta/
  Homo_sapiens.GRCh38.fa alignment001.bam alignment002.bam alignment003.bam
##reference=file:///opt/chipster/tools/genomes/fasta/Homo_sapiens.GRCh38.fa
##contig=<ID=1,length=248956422>
##contig=<ID=22,length=50818468>
##contig=<ID=X,length=156040895>
##contig=<ID=Y,length=57227415>
##contig=<ID=MT,length=16569>
##ALT=<ID=X,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Number of high-quality bases">
##FORMAT=<ID=DV,Number=1,Type=Integer,Description="Number of high-quality non-reference bases">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as l
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and al
##bcftools_callVersion=1.2+htslib-1.2.1
##bcftools_callCommand=call -vmO z –
#CHROM  POS ID REF ALT QUAL FILTER INFO FORMAT alignment001.bam alignment002.bam alignment003.ba
```

VCF variant site lines

- **QUAL**
  - Phred-scaled estimate of confidence that there is a variation at a given site in at least one sample.
  - QUAL = -10 * log (probability that a variant is called incorrectly). E.g. 20 means that there is 1:100 chance that there is no variant.
  - can be large when a large amount of data is used for variant calling
- **FORMAT explains sample columns**
  - GT:PL:DP:DV
  - 1/1:255,255,0:161:160

    0/0 0/1 1/1 (0/0 = hom-ref, 0/1 = het-alt, 1/1 = hom-alt)
  - Genotype likelihoods (PL) are normalized so that the most likely genotype has the value of 0. Genotype quality (GQ) is equal to the likelihood of the second most likely genotype.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	alignment001.bam	alignment002.bam	
20	3022243	.	A	C	469	.	DP=716;AC=3;DP4=25,194,21,212	GT:PL:DP:DV	1/1:255,255,0:161:160	0/0:0,255,255:147:1	0/1:255,0,255:144:72
20	3026582	.	G	A	289	.	DP=39;AC=3;DP4=14,1,19,0	GT:PL:DP:DV	1/1:192,30,0:10:10	0/0:0,36,250:12:0	0/1:138,0,91:12:9
20	3026846	.	T	C	394	.	DP=152;AC=3;DP4=2,65,2,65	GT:PL:DP:DV	1/1:255,141,0:47:47	0/0:0,123,248:41:0	0/1:180,0,205:46:20
20	3027048	.	AGGTGG	AGGTGGTGG	469	.	INDEL=1;IDV=41;DP=160;AC=3;DP4=63,1,73,0	GT:PL:DP:DV	1/1:255,129,0:43:43	0/0:0,107,255:54:1	0/1:255,0,215:40:29
20	3027593	.	C	T	338	.	DP=87;AC=3;DP4=29,2,36,1	GT:PL:DP:DV	1/1:203,63,0:21:21	0/0:0,69,255:23:0	0/1:176,0,157:24:16

BCF and gVCF

- **BCF is a binary form of VCF**
  - Same content but more compact, like BAM is a binary form of SAM
- **gVCF = genomic VCF**
  - similar to a normal VCF, but contains extra information
  - has records for all sites, whether there is a variant call or not (used for joint analysis of a cohort)
  - produced by GATK Haplotype Caller
  - non-variant sites can be grouped into intervals to save space

Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Software packages for visualization

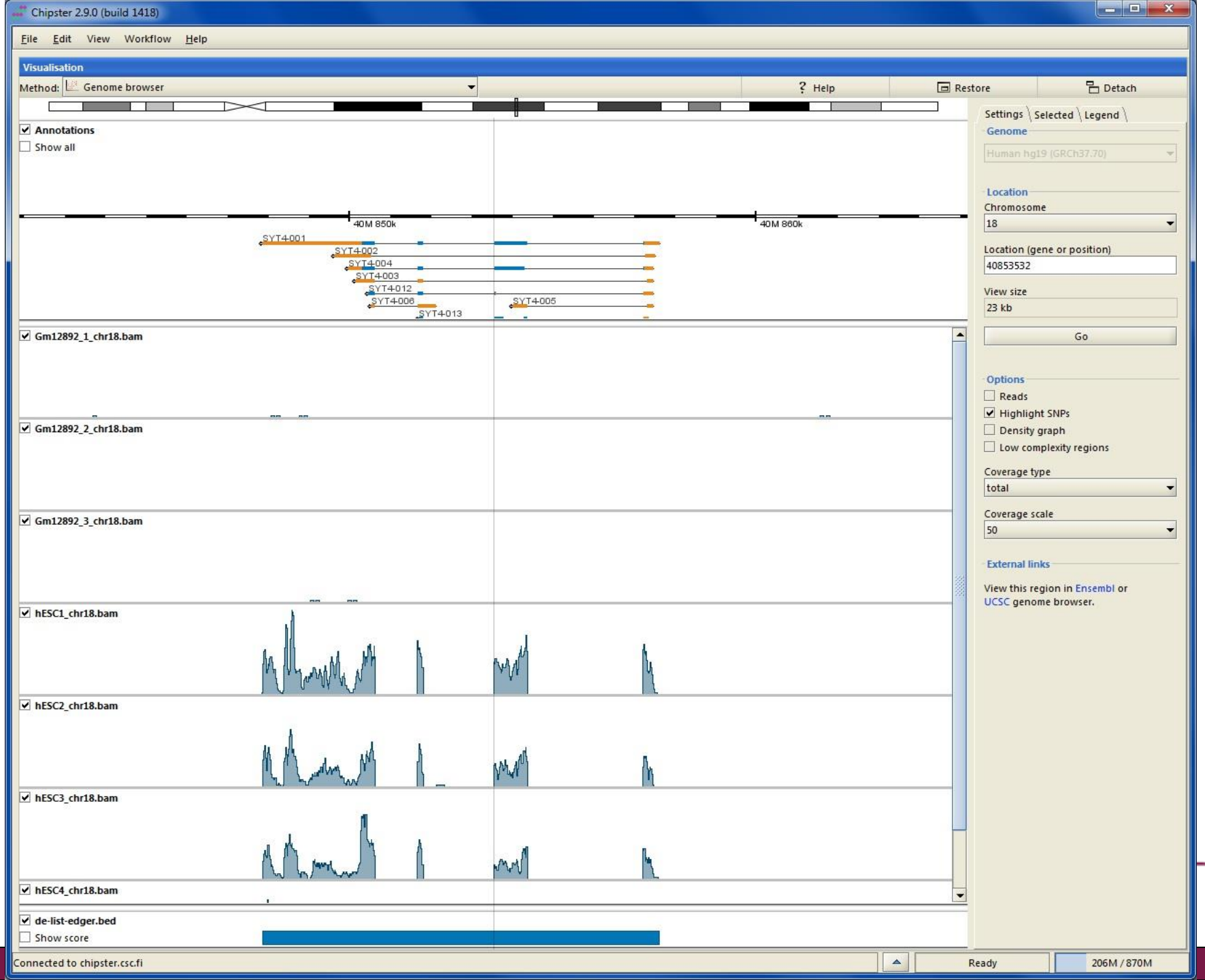
- **Chipster genome browser**
- **IGV**
- **UCSC genome browser**
- **....**

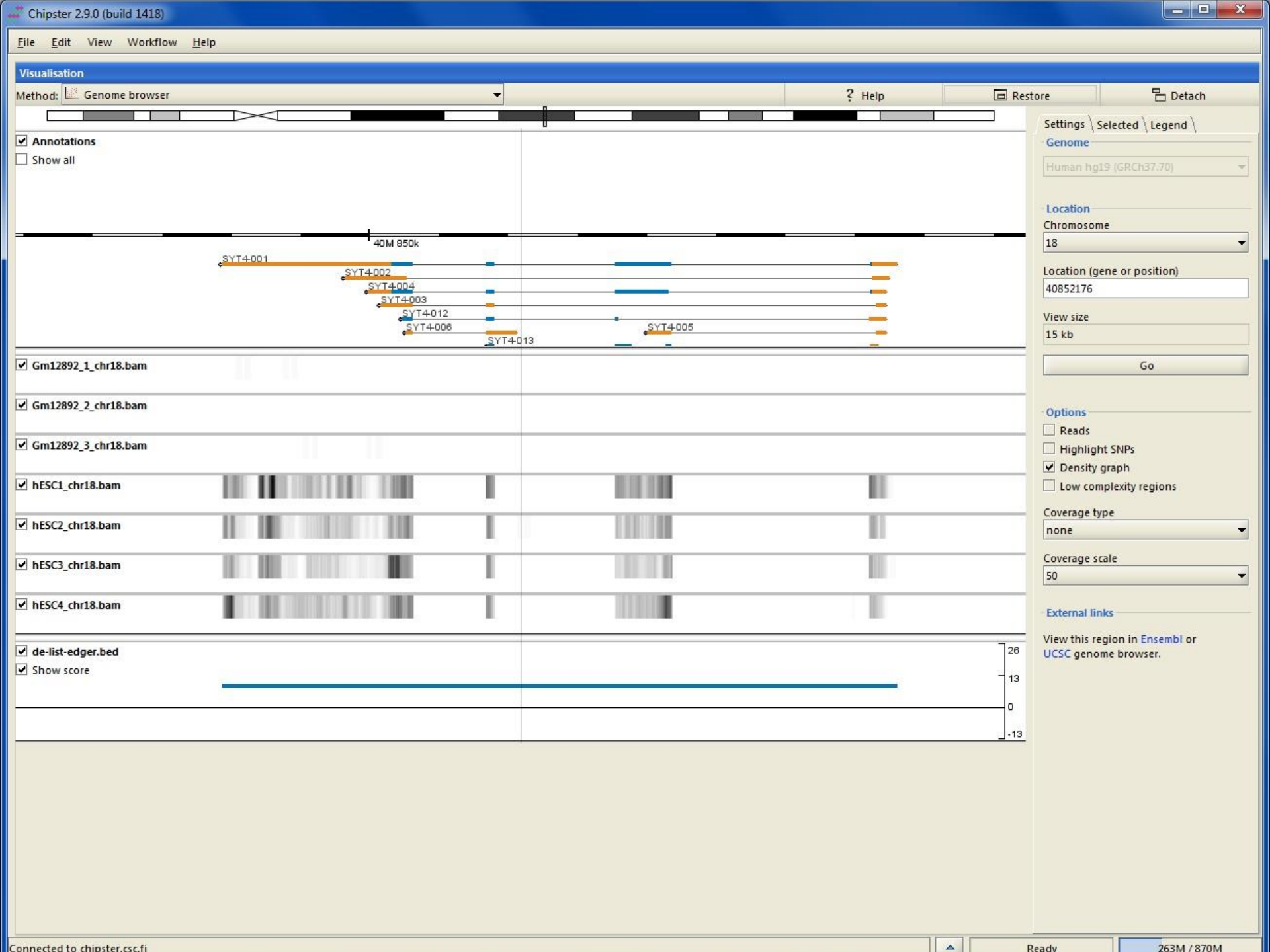
- **Differences in memory consumption, interactivity, annotations, navigation,...**

Chipster Genome Browser

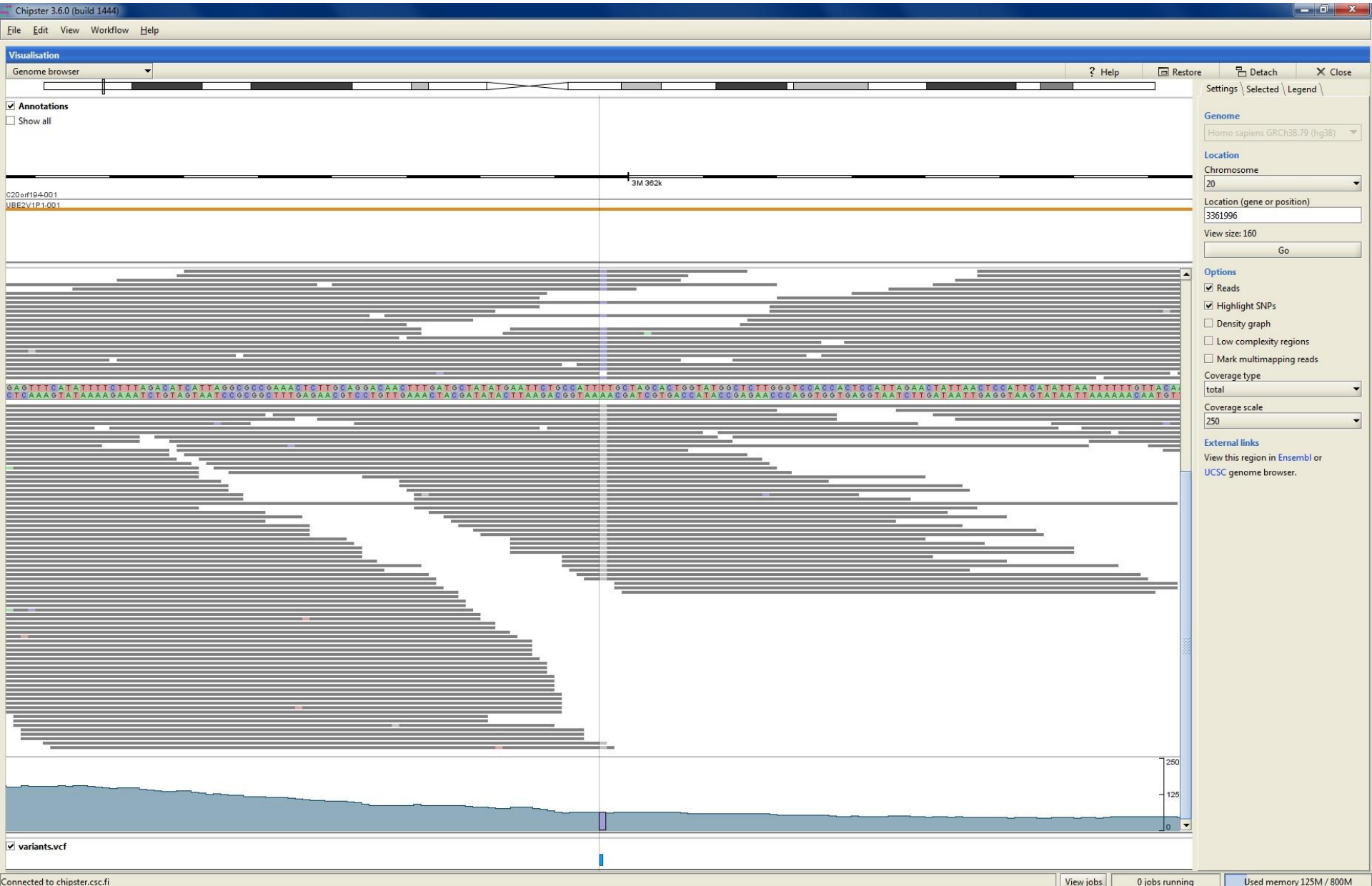
- **Integrated with Chipster analysis environment**
- **Automatic sorting and indexing of BAM, BED and GTF files**
- **Automatic coverage calculation (total and strand-specific)**
- **Zoom in to nucleotide level**
- **Highlight variants**
- **Jump to locations using BED, GTF, VCF and tsv files**
- **View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

Highlight variants at nucleotide level



Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Filtering variants

- **QUAL = variant quality score**
  - Not always reliable: variants in regions with deep coverage can have artificially inflated QUAL scores.
  - GATK VariantRecalibrator uses machine learning to assign more reliable VQSLOD scores. Needs high-quality sets of known variants and a lot of data (min 30 exomes)
  - GATK HaplotypeCaller provides a normalized value QD (QualityByDepth) which is QUAL by AD (allele depth)
- **Read depth (DP)**
- **Strand bias (FS < 60)**
- **Mapping quality (MQ > 40)**
- **Different thresholds for SNPs and INDELs, filter separately**

- **Filter based on locations and effects**
  - E.g. exome capture target areas, functional annotations

Chipster tool Calculate statistics on VCF file

- **Based on VCFtools**
- **Produces several reports**
  - per-site allele count and frequency
  - p-value for HWE
  - Linkage disequilibrium statistics
  - SNP density in bins of a given size
- **We can add more reports, just let us know what is needed**
  - See https://vcftools.github.io/man_latest.html#OUTPUT OPTIONS

Variant analysis workflow

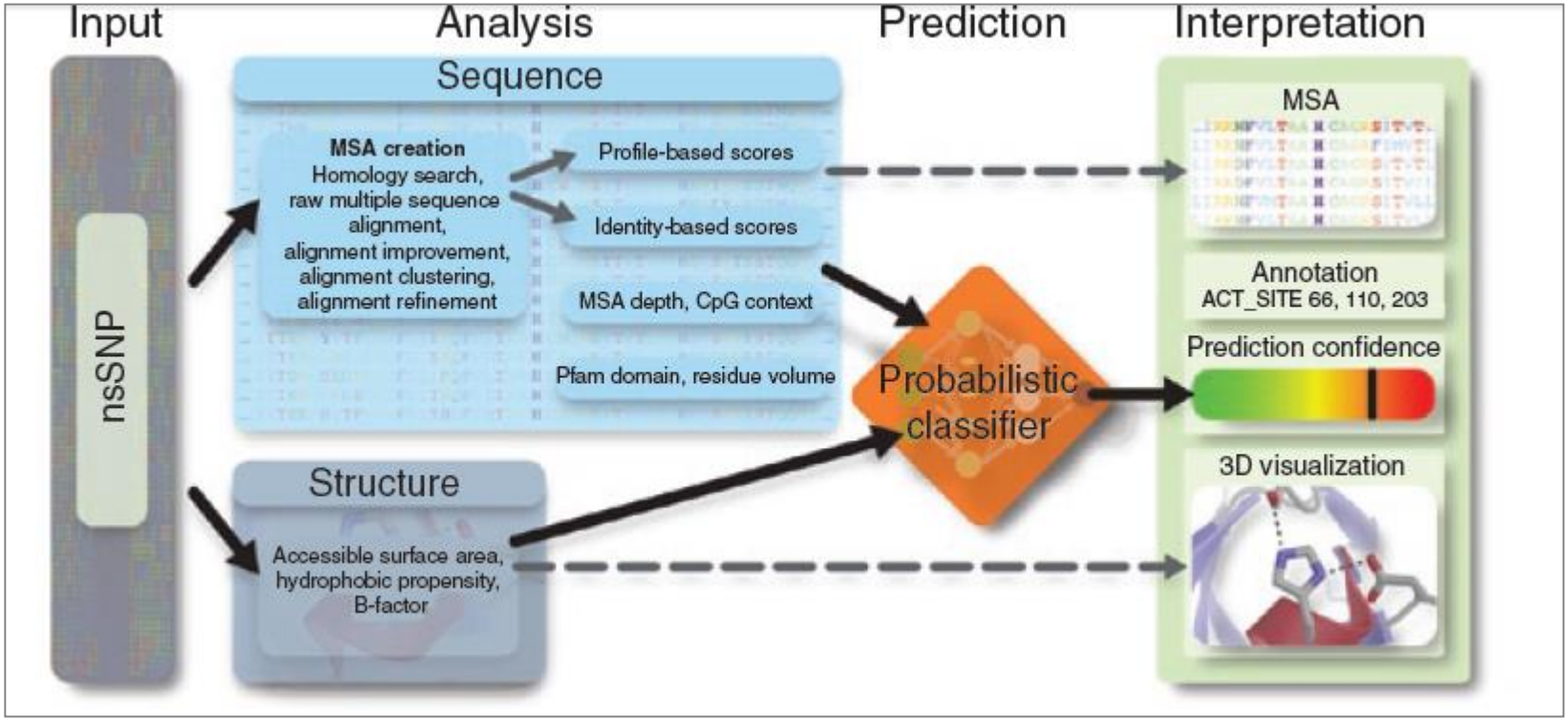
- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Annotating variants

- **Is it a known variant (from, e.g., dbSNP)**
- **Location (chromosome, base pair)**
- **Gene, upstream, downstream,…**
- **Consequence type (synonymous, etc.)**
- **SIFT prediction of impact on protein function**
  - based on sequence homology and the physical properties of amino acids
- **PolyPHEN prediction of impact on protein structure and function**
  - Based on physical and evolutionary comparative considerations
- **Associated disease(s)**
- **Frequency from the1000 genomes etc.**

CSC

PolyPHEN



- **http://genetics.bwh.harvard.edu/pph2/dokuwiki/_media/nmeth0410-248.pdf**

CSC

Annotate variants tool in Chipster

- **Uses Bioconductor package VariantAnnotation**
- **Coding, 3’ and 5’ UTR, intronic, splice site, and promoter variants**
- **dbSNP identifiers and PolyPhen information**
- **Genomes available: hg19 and hg38**

Annotate variants output files

all-variants.tsv:

GENEID	SEQNAMES	START	END	WIDTH	STRAND	LOCATI...	LOCSTART	LOCEND	QUERYID	TXID	CDSID	PRE...	FO...	SYMBOL	GENENAME	ENSEMBL
113730	chr22	50987558	50987558	1	+	coding	963	963	10149	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987601	50987601	1	+	coding	1006	1006	10150	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50986647	50986647	1	+	coding	52	52	10144	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987175	50987175	1	+	coding	580	580	10145	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987273	50987273	1	+	coding	678	678	10146	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987287	50987287	1	+	coding	692	692	10147	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987353	50987353	1	+	coding	758	758	10148	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988141	50988141	1	+	coding	1546	1546	10160	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987825	50987825	1	+	coding	1230	1230	10152	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987843	50987843	1	+	coding	1248	1248	10153	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987970	50987970	1	+	coding	1375	1375	10154	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987990	50987990	1	+	coding	1395	1395	10155	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988016	50988016	1	+	coding	1421	1421	10156	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988038	50988038	1	+	coding	1443	1443	10157	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988062	50988062	1	+	coding	1467	1467	10158	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988105	50988105	1	+	coding	1510	1510	10159	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988183	50988183	1	+	coding	1588	1588	10161	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988193	50988193	1	+	coding	1598	1598	10162	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988317	50988317	1	+	coding	1722	1722	10163	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988352	50988352	1	+	coding	1757	1757	10164	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50988376	50988376	1	+	coding	1781	1781	10165	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	chr22	50987690	50987690	1	+	coding	1095	1095	10151	74399	216445	char...	cha...	KLHDC7B	kelch domain containing 7B	ENSG00000130487
1890	chr22	50965102	50965102	1	-	coding	831	831	9823	75301	218800	char...	cha...	TYMP	thymidine phosphorylase	ENSG00000025708
1890	chr22	50965624	50965624	1	-	coding	735	735	9826	75297	218801	char...	cha...	TYMP	thymidine phosphorylase	ENSG00000025708
1890	chr22	50965624	50965624	1	-	coding	735	735	9826	75298	218801	char...	cha...	TYMP	thymidine phosphorylase	ENSG00000025708

coding-variants.tsv:

geneID	rsID	cdsID	txID	consequence	cdsStart	cdsEnd	width	varAllele	refCodon	varCodon	refAA	varAA	SYMBOL	GENENAME	ENSEMBL
113730	rs62239508	216445	74399	nonsynonymous	52	52	1	A	GTC	ATC	V	I	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs116500907	216445	74399	nonsynonymous	580	580	1	T	GTG	TTG	V	L	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	22:50987273_G/T	216445	74399	synonymous	678	678	1	T	GTG	GTT	V	V	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs5770886	216445	74399	nonsynonymous	692	692	1	G	CAG	CGG	Q	R	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs117231091	216445	74399	synonymous	1230	1230	1	T	ATC	ATT	I	I	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs61746062	216445	74399	synonymous	1248	1248	1	T	TAC	TAT	Y	Y	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	22:50987970_C/A	216445	74399	nonsynonymous	1375	1375	1	A	CTC	ATC	L	I	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs77297390	216445	74399	synonymous	1395	1395	1	A	AGG	AGA	R	R	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	22:50988016_A/G	216445	74399	nonsynonymous	1421	1421	1	G	GAC	GGC	D	G	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs115466691	216445	74399	synonymous	1443	1443	1	T	AGC	AGT	S	S	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs140519	216445	74399	synonymous	1467	1467	1	T	GTG	GTT	V	V	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs36062310	216445	74399	nonsynonymous	1510	1510	1	A	GTG	ATG	V	M	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	rs115170030	216445	74399	nonsynonymous	1546	1546	1	A	GGC	AGC	G	S	KLHDC7B	kelch domain containing 7B	ENSG00000130487
113730	22:50987353_T/C	216445	74399	nonsynonymous	758	758	1	C	GTG	GCG	V	A	KLHDC7B	kelch domain containing 7B	ENSG00000130487

Ensembl Variant Effect Predictor (VEP)

- **Genes and transcripts affected**
- **Relative location (coding, non-coding, upstream…)**
- **Impact**
  - High: protein truncation, loss of function, triggers decay
  - Moderate: might change protein effectiveness
  - Low: mostly harmless
  - Modifier: usually non-coding variant
- **Consequence (stop gained, missense, frameshift…)**
  - http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences
- **Known variants that match yours**
- **SIFT & PolyPhen scores**
- **VEP in Chipster uses EBI's VEP service**
- **Genomes available: hg38, more can be added**
- **http://www.ensembl.org/info/docs/tools/vep/**

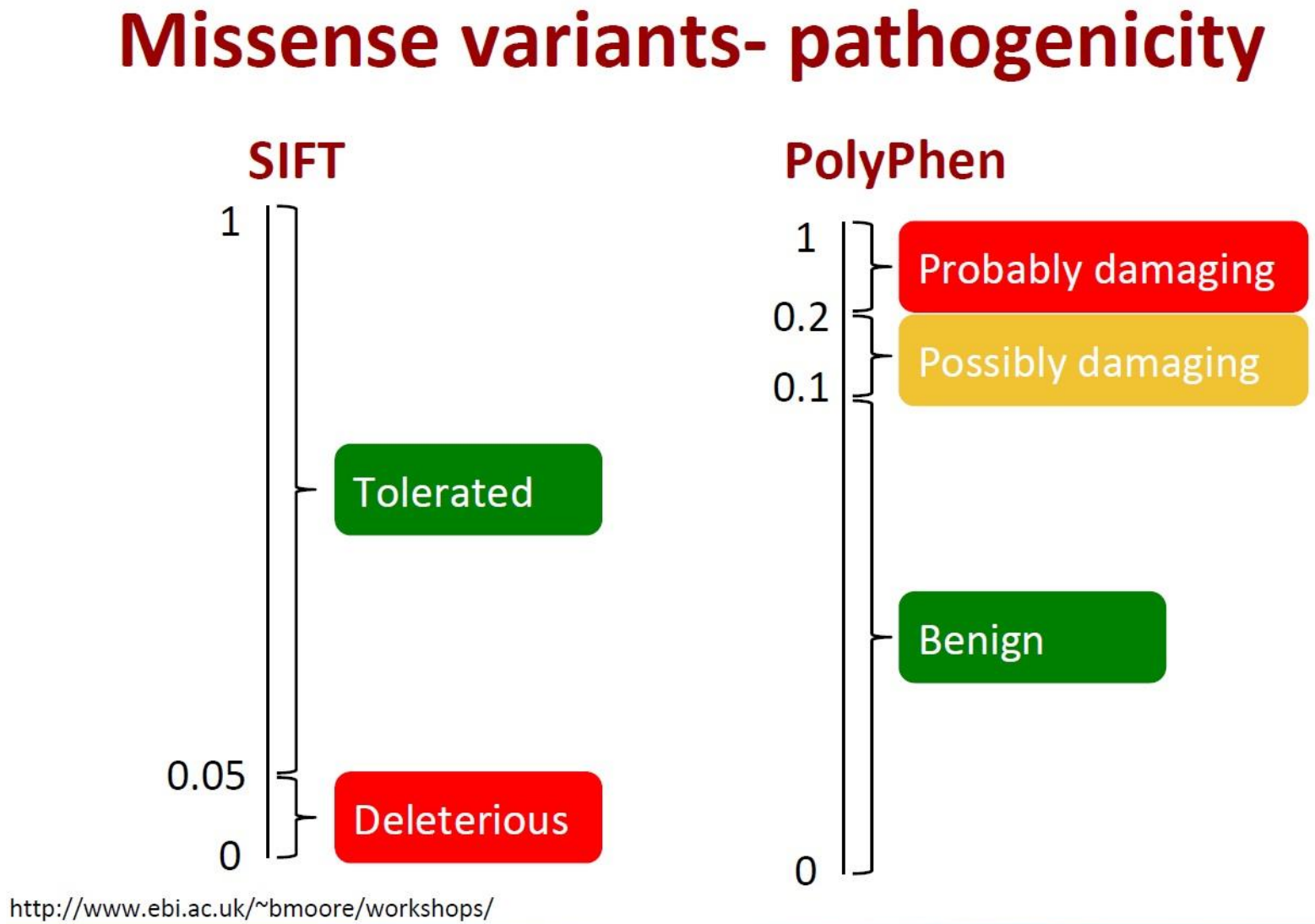
CSC

VEP output file

#U...	Location	Allele	Consequence	IMPACT	SYMBOL	Gene	Feature_ty...	Feature	BIOTYPE	EXON	INTRON	cDNA_position	CDS_position	Protein_position	Amin...	Codons
.	20:3803463-3803463	T	synonymous_variant	LOW	CDC25B	ENSG00000101224	Transcript	ENST00000344256	protein_coding	14/16	-	1584	1224	408	I	atC/atT
.	20:3803463-3803463	T	synonymous_variant	LOW	CDC25B	ENSG00000101224	Transcript	ENST00000245960	protein_coding	14/16	-	2113	1416	472	I	atC/atT
.	20:3857794-3857794	G	missense_variant	MODERATE	MAVS	ENSG00000088888	Transcript	ENST00000428216	protein_coding	3/7	-	405	277	93	Q/E	Cag/Gag
.	20:3190691-3190691	T	missense_variant	MODERATE	DDRGK1	ENSG00000198171	Transcript	ENST00000354488	protein_coding	9/9	-	965	907	303	A/T	Gcc/Acc
.	20:3229655-3229655	A	missense_variant,NMD_tr...	MODERATE	SLC4A11	ENSG00000088836	Transcript	ENST00000474451	nonsense_med...	14/20	-	1709	1484	495	T/M	aCg/aTg
.	20:3304493-3304493	C	missense_variant	MODERATE	C20orf194	ENSG00000088854	Transcript	ENST00000252032	protein_coding	21/37	-	1797	1729	577	R/G	Aga/Gga
.	20:3669558-3669558	A	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000356518	protein_coding	20/22	-	2562	2320	774	P/S	Ccc/Tcc
.	20:3669558-3669558	A	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000379861	protein_coding	20/22	-	2562	2320	774	P/S	Ccc/Tcc
.	20:3669558-3669558	A	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000619289	protein_coding	18/20	-	2087	1960	654	P/S	Ccc/Tcc
.	20:3669558-3669558	A	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000350009	protein_coding	19/21	-	2369	2242	748	P/S	Ccc/Tcc
.	20:3669587-3669587	G	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000356518	protein_coding	20/22	-	2533	2291	764	M/T	aTg/aCg
.	20:3669587-3669587	G	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000379861	protein_coding	20/22	-	2533	2291	764	M/T	aTg/aCg
.	20:3669587-3669587	G	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000619289	protein_coding	18/20	-	2058	1931	644	M/T	aTg/aCg
.	20:3669587-3669587	G	missense_variant	MODERATE	ADAM33	ENSG00000149451	Transcript	ENST00000350009	protein_coding	19/21	-	2340	2213	738	M/T	aTg/aCg
.	20:3044461-3044461	T	missense_variant	MODERATE	GNRH2	ENSG00000125787	Transcript	ENST00000380346	protein_coding	1/3	-	50	47	16	A/V	gCc/gTc
.	20:3044461-3044461	T	missense_variant	MODERATE	GNRH2	ENSG00000125787	Transcript	ENST00000380347	protein_coding	1/3	-	452	47	16	A/V	gCc/gTc
.	20:3044461-3044461	T	missense_variant	MODERATE	GNRH2	ENSG00000125787	Transcript	ENST00000359100	protein_coding	2/4	-	98	47	16	A/V	gCc/gTc
.	20:3044461-3044461	T	missense_variant	MODERATE	GNRH2	ENSG00000125787	Transcript	ENST00000245983	protein_coding	2/4	-	98	47	16	A/V	gCc/gTc
.	20:3694686-3694686	A	missense_variant	MODERATE	SIGLEC1	ENSG00000088827	Transcript	ENST00000344754	protein_coding	11/21	-	2921	2921	974	A/V	gCa/gTa
.	20:3694851-3694851	G	missense_variant	MODERATE	SIGLEC1	ENSG00000088827	Transcript	ENST00000344754	protein_coding	11/21	-	2756	2756	919	H/P	cAc/cCc
.	20:3701479-3701479	T	missense_variant	MODERATE	SIGLEC1	ENSG00000088827	Transcript	ENST00000344754	protein_coding	6/21	-	1391	1391	464	R/H	cGc/cAc
.	20:3704082-3704082	C	missense_variant	MODERATE	SIGLEC1	ENSG00000088827	Transcript	ENST00000344754	protein_coding	4/21	-	716	716	239	K/R	aAg/aGg
.	20:3705789-3705789	T	missense_variant	MODERATE	SIGLEC1	ENSG00000088827	Transcript	ENST00000344754	protein_coding	3/21	-	661	661	221	V/M	Gtg/Atg
.	20:3022243-3022243	C	intron_variant	MODIFIER	PTPRA	ENSG00000132670	Transcript	ENST00000216877	protein_coding	-	14/22	-	-	-	-	-
.	20:3022243-3022243	C	intron_variant	MODIFIER	PTPRA	ENSG00000132670	Transcript	ENST00000356147	protein_coding	-	14/22	-	-	-	-	-

Existing_variation	Extra
rs1056720,COSM3758565,COSM375...	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:1726;SOMATIC=1,1;PHENO=1,1
rs1056720,COSM3758565,COSM375...	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:1726;SOMATIC=1,1;PHENO=1,1
rs17857295,COSM4134535	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:29233;SIFT=tolerated(0.11);PolyPhen=benign(0.116);SOMATIC=1;PHENO=1
rs11591	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:16110;SIFT=tolerated(0.36);PolyPhen=benign(0.008)
rs41281860,COSM3773985,COSM37...	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:16438;SIFT=deleterious_low_confidence(0.02);PolyPhen=benign(0.063);SOMATIC=1,1;...
rs2422864	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:17721;SIFT=tolerated(1);PolyPhen=benign(0)
rs2280090,CM099912,COSM3758549	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.1);PolyPhen=benign(0.04);SOMATIC=1;PHENO=1,1
rs2280090,CM099912,COSM3758549	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.1);PolyPhen=benign(0.018);SOMATIC=1;PHENO=1,1
rs2280090,CM099912,COSM3758549	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.08);PolyPhen=benign(0.016);SOMATIC=1;PHENO=1,1
rs2280090,CM099912,COSM3758549	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.1);PolyPhen=benign(0.041);SOMATIC=1;PHENO=1,1
rs2280091,CM099913,COSM3758550	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.28);PolyPhen=benign(0.007);SOMATIC=1;PHENO=1,1
rs2280091,CM099913,COSM3758550	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.23);PolyPhen=benign(0.007);SOMATIC=1;PHENO=1,1
rs2280091,CM099913,COSM3758550	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.6);PolyPhen=benign(0.011);SOMATIC=1;PHENO=1,1
rs2280091,CM099913,COSM3758550	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:15478;SIFT=tolerated(0.27);PolyPhen=benign(0.015);SOMATIC=1;PHENO=1,1
rs6051545,COSM4001869	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:4420;SIFT=tolerated(0.5);PolyPhen=probably_damaging(0.971);SOMATIC=1;PHENO=1
rs6051545,COSM4001869	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:4420;SIFT=tolerated(0.45);PolyPhen=probably_damaging(0.971);SOMATIC=1;PHENO=1
rs6051545,COSM4001869	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:4420;SIFT=tolerated(0.45);PolyPhen=probably_damaging(0.971);SOMATIC=1;PHENO=1
rs6051545,COSM4001869	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:4420;SIFT=tolerated(0.5);PolyPhen=probably_damaging(0.935);SOMATIC=1;PHENO=1
rs3746638	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:11127;SIFT=tolerated(0.14);PolyPhen=benign(0.098)
rs709012	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:11127;SIFT=tolerated(1);PolyPhen=benign(0)
rs34924243	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:11127;SIFT=deleterious(0);PolyPhen=probably_damaging(0.996)
rs625372	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:11127;SIFT=tolerated(0.29);PolyPhen=benign(0.115)
rs6037651,COSM1411675,COSM413...	STRAND=-1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:11127;SIFT=tolerated(0.11);PolyPhen=benign(0.21);SOMATIC=1,1;PHENO=1,1
rs779664797,rs544090	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:9664
rs779664797,rs544090	STRAND=1;SYMBOL_SOURCE=HGNC;HGNC_ID=HGNC:9664

SIFT and PolyPHEN scores in VEP



Variant analysis workflow

- **Check the quality of reads**
- **Remove bad quality data if needed**
- **Align/map reads to reference genome**
- **Process alignment files**
  - Sort, index, QC
  - Mark duplicates, realign indels, recalibrate base quality
- **Call variants**
- **Visualize reads and variants in genomic context**
- **Filter variants**
- **Annotate variants**
- **Match sets of genomic regions**

Intersect genomic regions: examples

- **Give me only the reads that map to exons**
- **Give me only the variants that are located in exons**
- **Does my SNP list contain known SNPs?**
- **Give me only the reads which do not match known genes / SNPs**
- **Compare lists of variants**
- **….**

BED file format

- **5 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

Intersect BED (BEDTools)

- **Looks for overlapping regions between two BED/GFF/VCF files**
  - One of the files can also be BAM
  - Option for strand-awareness
- **Reporting options**
  - Only the overlapping region
  - Original region in file A or B
  - Region in A so that the overlapping part is removed
  - Remove the portion of a region that is overlapped by another region
- **The B file is loaded to memory**
  - So the smaller one should be B (e.g. BAM = A, exons = B)

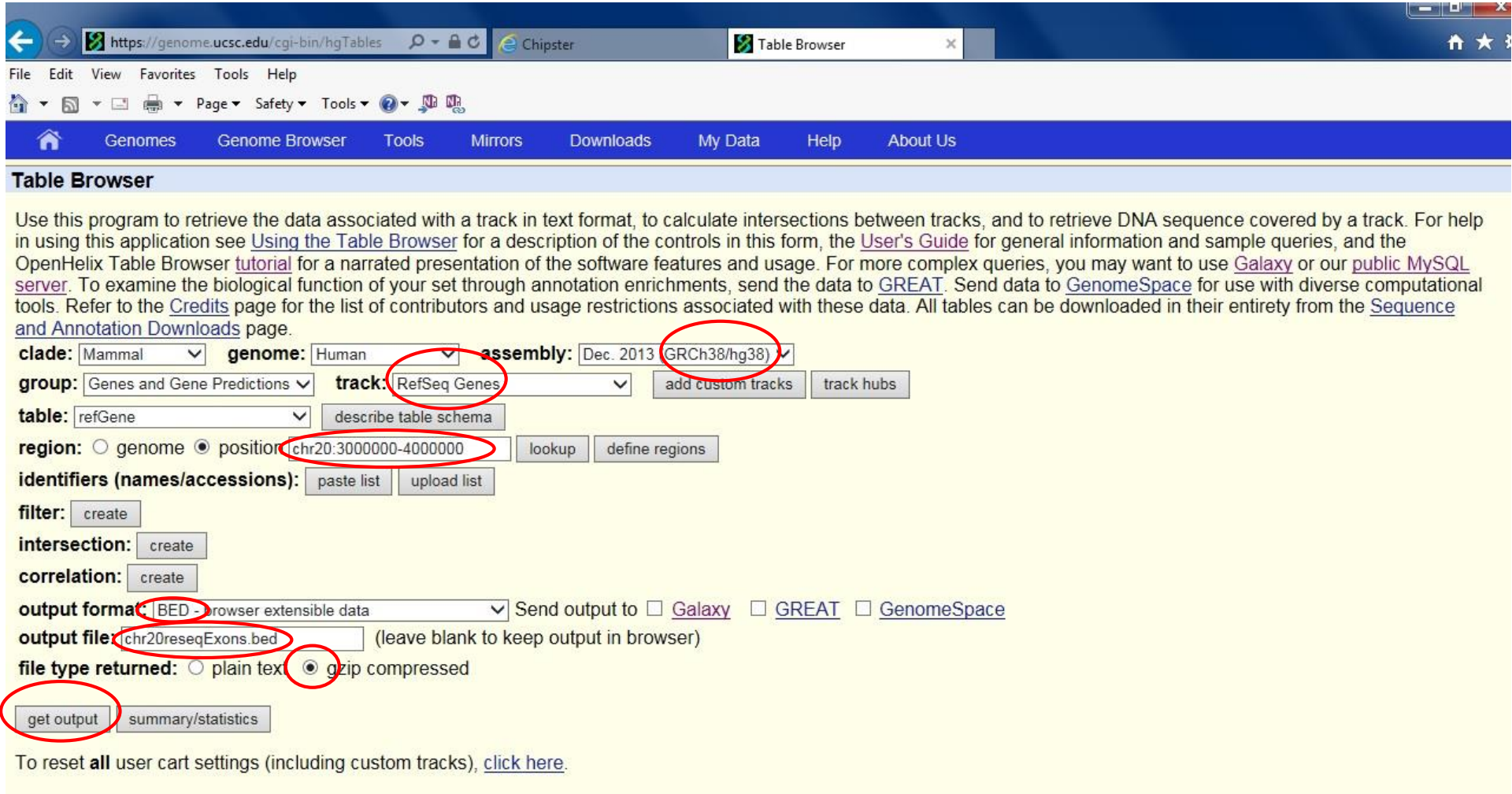
Closest BED (BEDTools)

- **Looks for overlapping regions between two BED/GFF/VCF files and if no overlap is found, the closest region is reported.**
- **Reports a region in A followed by its closest region in B.**
- **Option for strand-awareness**
- **E.g. What is the nearest gene to this SNP?**

Window BED (BEDTools)

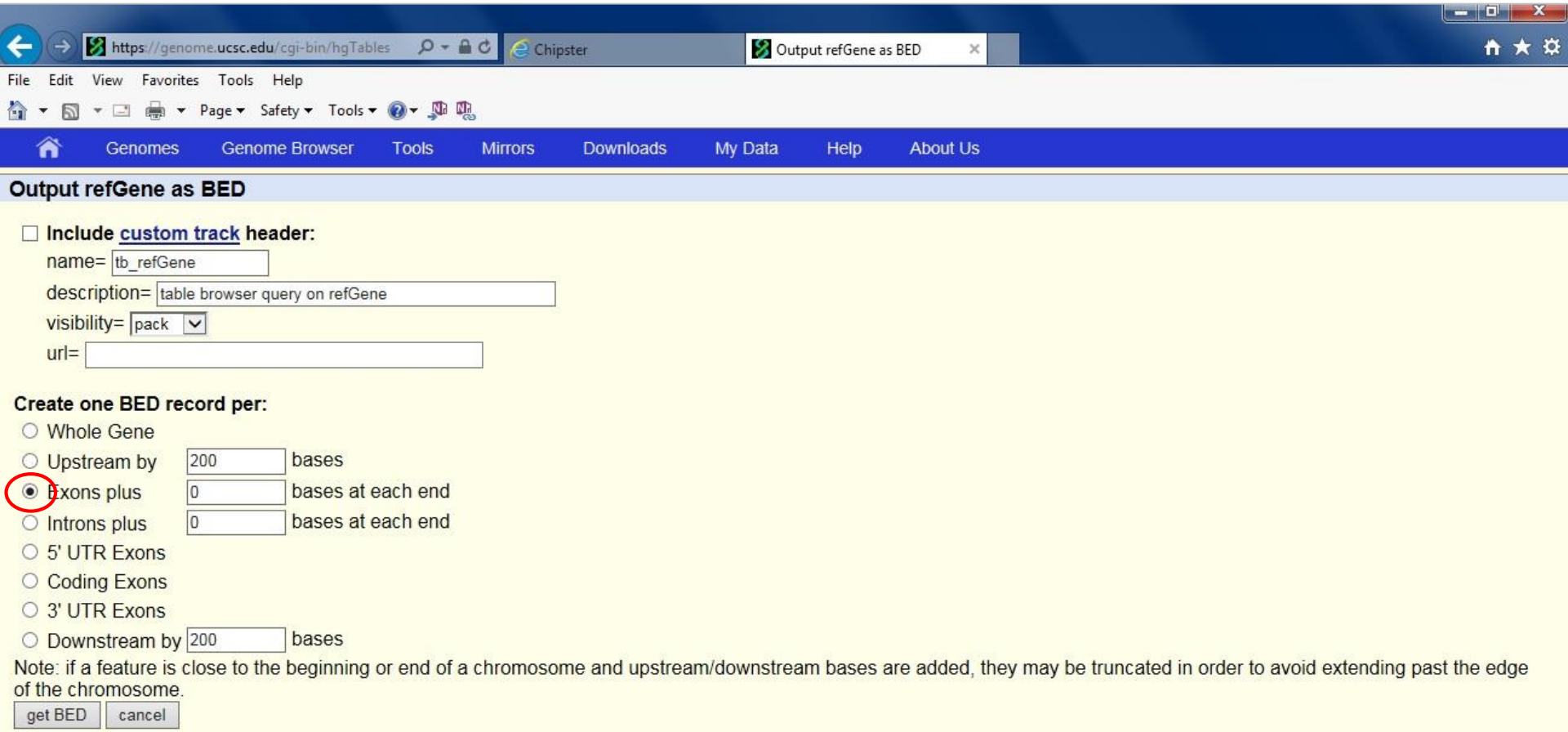
- **Looks for overlapping regions between two BED/GFF/VCF files after adding a given number of bases upstream and downstream of regions in A.**
  - One of the files can be BAM
- **Reports the regions in A which overlap with regions in B.**
- **Option for strand-awareness**

Obtain BED from UCSC Table Browser



CSC

UCSC Table Browser, part II



CSC

Preprocessing UCSC BED in Chipster

Change chromosome names to match the reference (chr1 → 1)

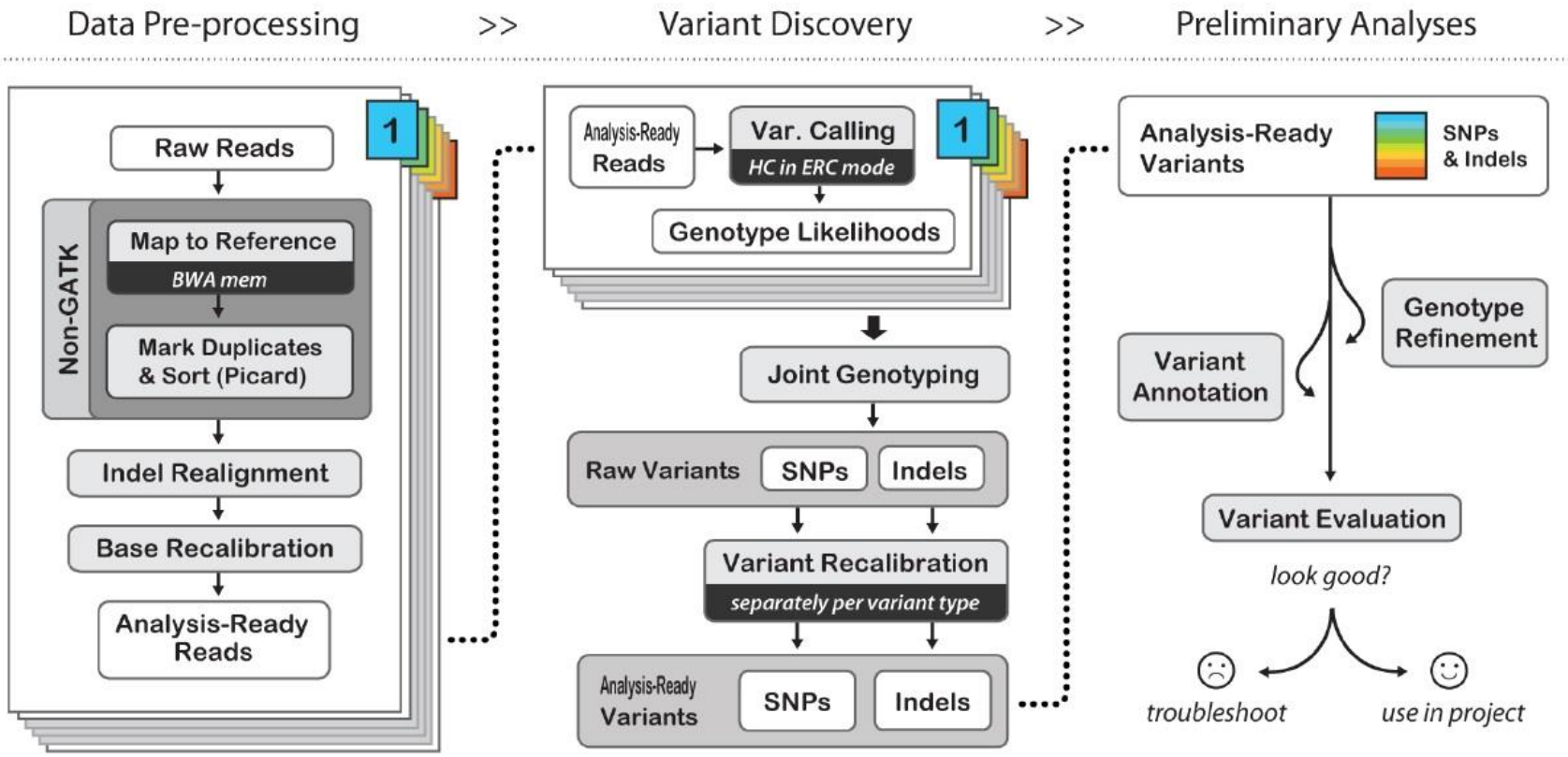
- Select **Utilities / Modify text**
- In the parameters change the following:
  - Operation = Replace text
  - Search string = chr
  - Replacement string = leave empty
  - Input file format = BED

Summary

Summary

- **Variant calling is a complicated, multi-step process**
- **The big challenge is to separate real variants from sequencing and mapping errors.**
  - Use good raw data
  - Use accurate aligner
  - Mark duplicates
  - Deal with misalignments around INDELs (use BAC or realign)
  - Recalibrate base quality values if possible
  - Do multi-sample calling
  - Recalibrate variant quality values if possible
  - Filter variants
  - Annotate variants

GATK workflow



CSC

More info

- **Manuals of individual analysis tools**
  - Varying quality…
- **GATK best practises documents**
  - https://www.broadinstitute.org/gatk/guide/best-practices.php
- **Biostars**
  - https://www.biostars.org/
- **SEQanswers**
  - http://seqanswers.com/