Hands-on Exercises

CHIPSTER AND FEDERATED CLOUD

Slides and Exercises modified from the CSC presentation (EMBO event)

Outline

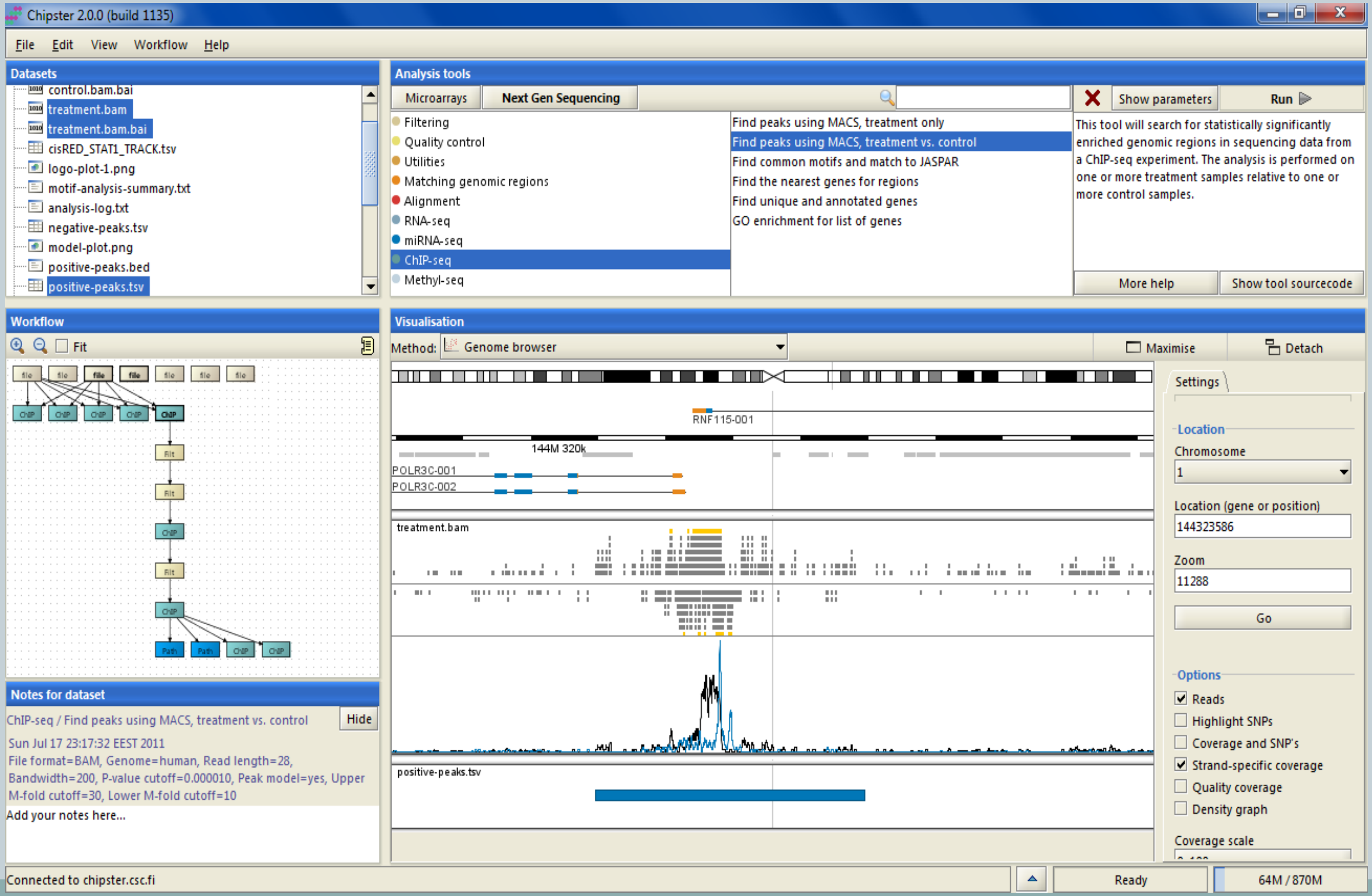
- Introduction to Chipster
- NGS data analysis and visualization
  - Quality control and filtering
  - Alignment
  - Matching sets of genomic regions
  - Visualization of reads and results in their genomic context
  - miRNA-seq: differential expression
- Summary

Why Chipster?

- Goal of Chipster is to enable wet-lab life-science researchers to:
  - ○ Analyse and integrate high-throughput data
  - ○ Visualize results efficiently
  - ○ Save and share automatic workflows

User friendly?

- Interactive visualization and workflow functionality



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.

ChIP-seq / Find peaks using MACS, treatment vs. control
Sun Jul 17 23:17:32 EEST 2011
File format=BAM, Genome=human, Read length=28, Bandwidth=200, P-value cutoff=0.000010, Peak model=yes, Upper M-fold cutoff=30, Lower M-fold cutoff=10
Add your notes here...

Method: Genome browser

Chromosome: 1
Location (gene or position): 144323586
Zoom: 11288

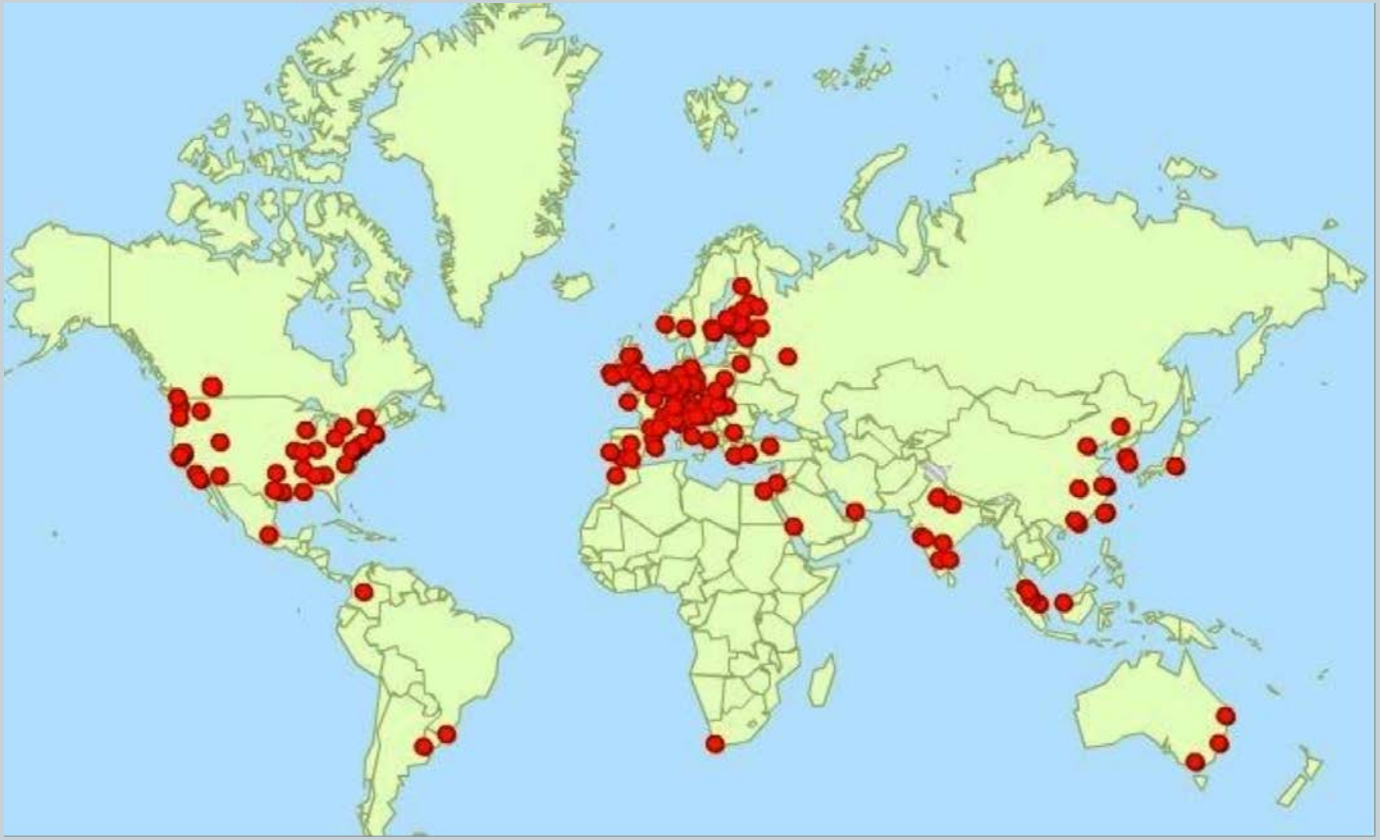
☑ Reads
☐ Highlight SNPs
☐ Coverage and SNP's
☑ Strand-specific coverage
☐ Quality coverage
☐ Density graph

Coverage scale

Connected to chipster.csc.fi

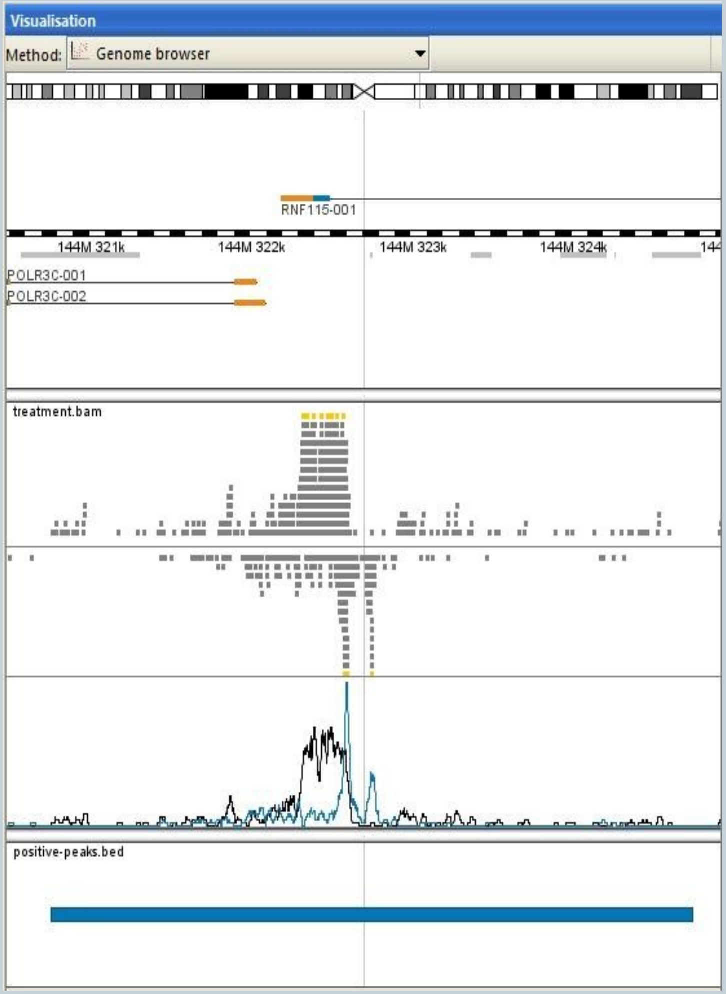
Never heard of it...

- Quite used across the world as a server / Virtual Machine



Chipster 2.0

- >50 analysis tools for:
  - ChIP-seq
  - RNA-seq
  - miRNA-seq
  - MeDIP-seq
- Integrated genome browser
- 135 microarray analysis tools:
  - Gene expression
  - miRNA expression
  - Protein expression
  - aCGH
  - SNP
  - Integration of different data types

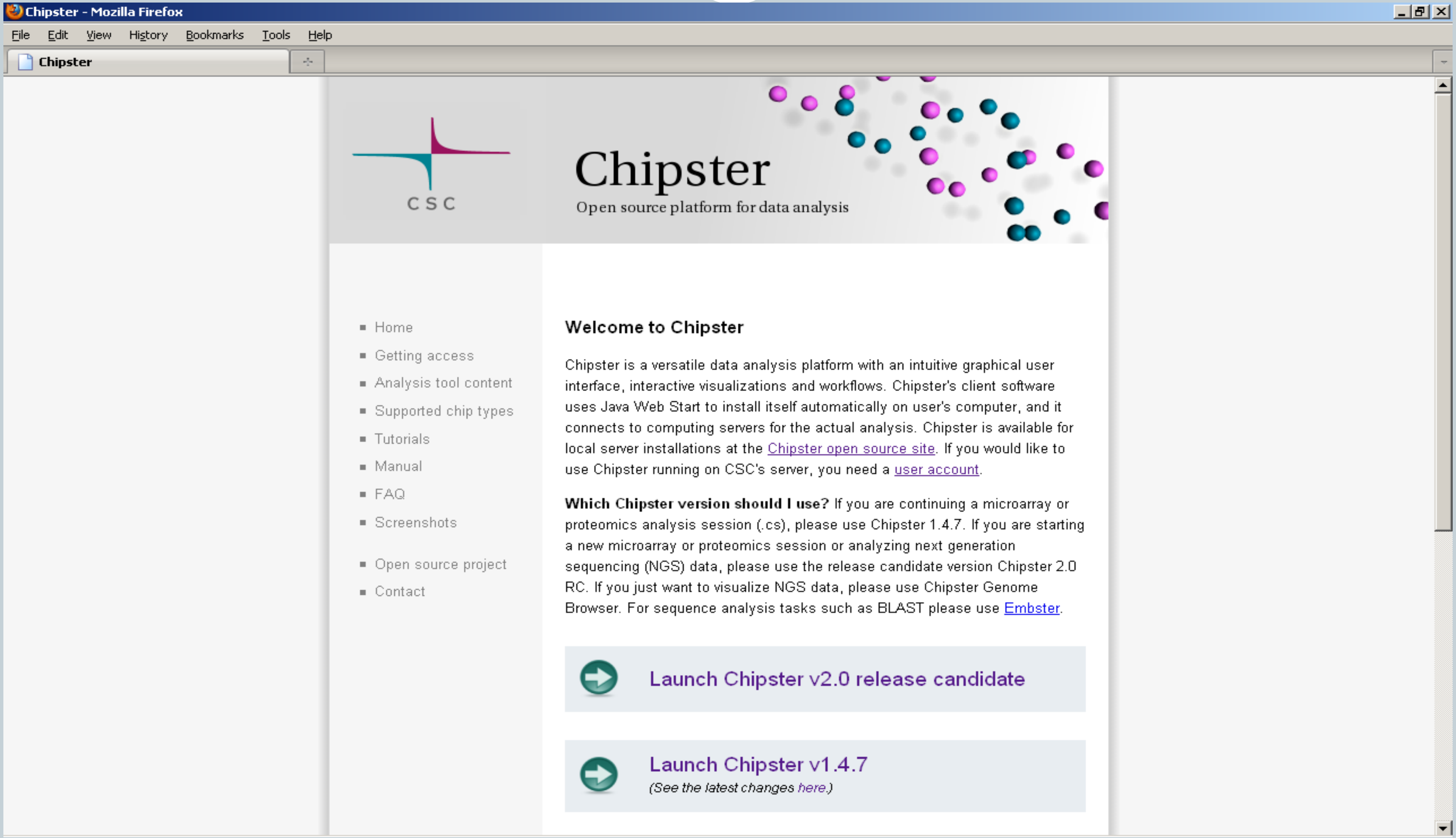

Visualisation

Method: Genome browser

Focus on NGS

- Quality control, filtering, trimming
  - FastX
  - FastQC
- Alignment
  - Bowtie
  - Tophat
- Processing
  - Picard, SAMTools
- Visualization of reads and results in their genomic context
- Genomic region matching
  - In house (Chipster) tools
  - BEDTools
  - HTSeq

Chipster start and info page



Welcome to Chipster

Chipster is a versatile data analysis platform with an intuitive graphical user interface, interactive visualizations and workflows. Chipster's client software uses Java Web Start to install itself automatically on user's computer, and it connects to computing servers for the actual analysis. Chipster is available for local server installations at the Chipster open source site. If you would like to use Chipster running on CSC's server, you need a user account.

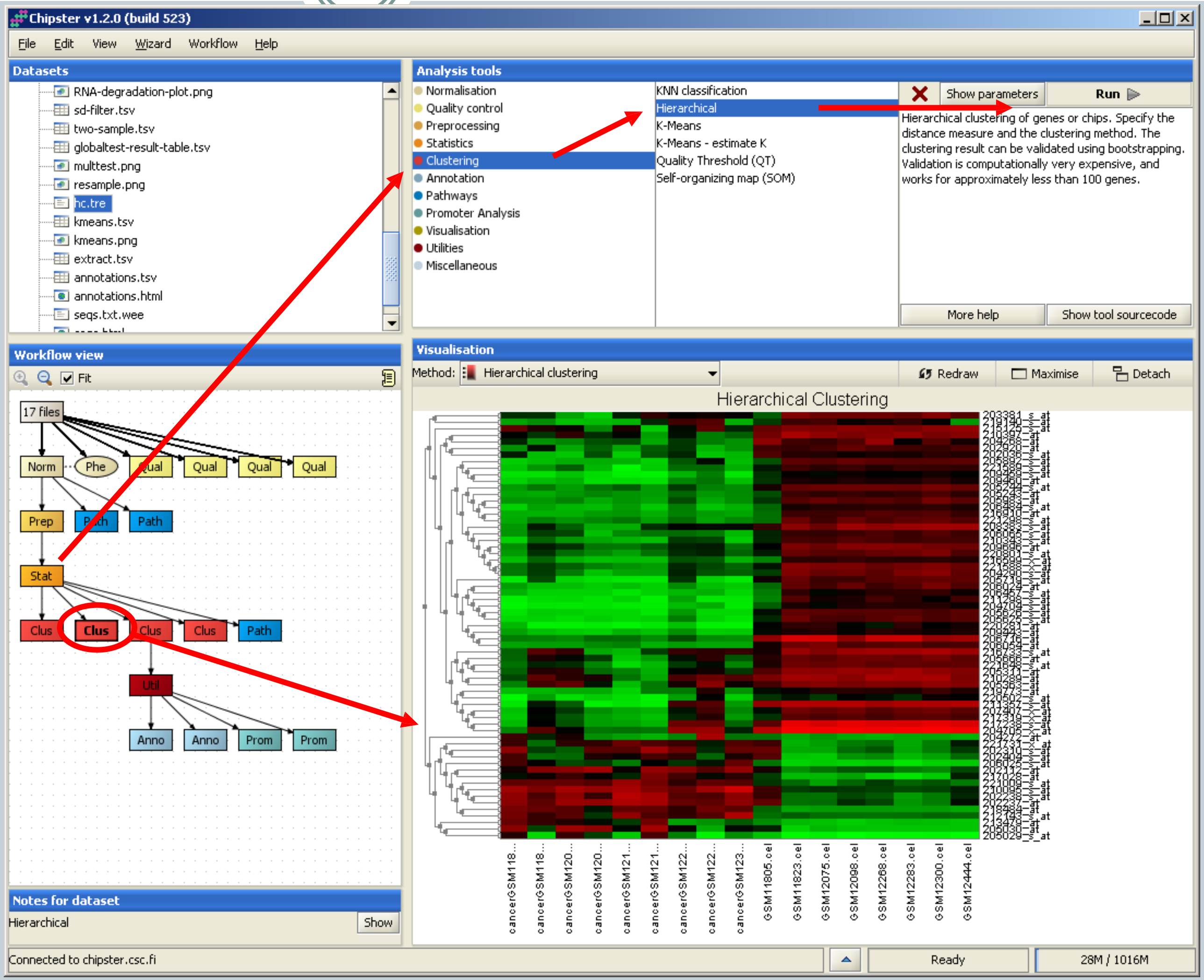
Which Chipster version should I use? If you are continuing a microarray or proteomics analysis session (.cs), please use Chipster 1.4.7. If you are starting a new microarray or proteomics session or analyzing next generation sequencing (NGS) data, please use the release candidate version Chipster 2.0 RC. If you just want to visualize NGS data, please use Chipster Genome Browser. For sequence analysis tasks such as BLAST please use Embster.

Launch Chipster v2.0 release candidate

Launch Chipster v1.4.7
(See the latest changes here.)

Chipster mode of operation

- Select data
- Select tool category
- Select tool
- Set parameters
- Click run
- Double-click to view



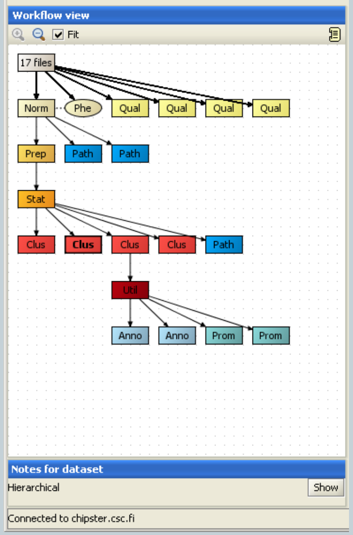
Hierarchical clustering of genes or chips. Specify the distance measure and the clustering method. The clustering result can be validated using bootstrapping. Validation is computationally very expensive, and works for approximately less than 100 genes.

Hierarchical

Connected to chipster.csc.fi

Workflow view

- Shows the relationships of the data sets
- Right-clicking on the data allows you to
  - Save (extract)
  - Delete
  - Visualize
  - Link to another data file
  - View analysis history
  - Save workflow
- Zoom in/out or fit to panel
- View information about the data by clicking on the Show button
- Mousing over a data file shows you the number of data rows (when applicable)
- You can select several datasets (e.g. for a Venn diagram) by keeping the Ctrl key down


Workflow view

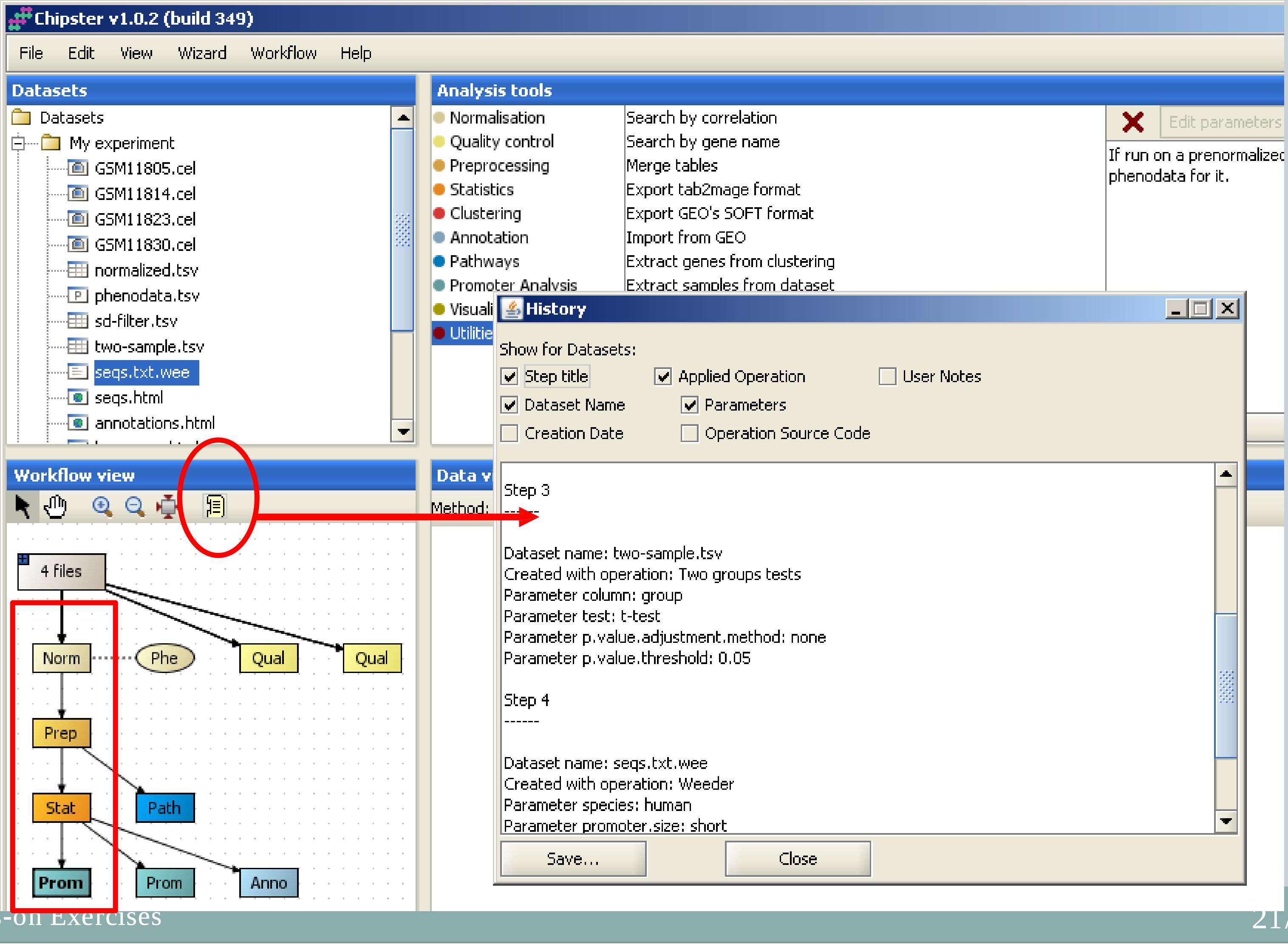
☑ Fit

Notes for dataset

Hierarchical

Connected to chipster.csc.fi

Automatic tracking of analysis history



Show for Datasets:

☑ Step title ☑ Applied Operation □ User Notes
☑ Dataset Name ☑ Parameters
□ Creation Date □ Operation Source Code

```
Step 3
------

Dataset name: two-sample.tsv
Created with operation: Two groups tests
Parameter column: group
Parameter test: t-test
Parameter p.value.adjustment.method: none
Parameter p.value.threshold: 0.05

Step 4
------

Dataset name: seqs.txt.wee
Created with operation: Weeder
Parameter species: human
Parameter promoter.size: short
```

Analysis sessions

- In order to continue your work later on, you have to save the analysis session.
- Saving the session will save all the datasets and their relationships. The session is packed into a single .zip file.
- Session files allow you to continue your work on another computer or share it with a colleague.
- You can have multiple analysis session saved separately, and you can combine them later if needed.

Before everything: we need resources

- We will use resources provided by the training infrastructure of EGI, through the Federated Cloud
- We will launch a number of Chipster servers, one for every “work group”
- Members of the same group will connect to the same server, but each with unique credentials ☺
- The detailed step-by-step instructions can be found here: http://tinyurl.com/pg7avc4

Exercise 0: Start Chipster

- Connect to the UI
- Launch the Chipster VM (unfortunately, 1 in 4 will do this in practice)
- Launch the Chipster client program

Exercise 1: Import data

- Click Import/File and select file:

  `1000readsFromRNAseq.fastq`

- Double-click on the file to see what it looks like

- Select the tab **Next Gen Sequencing (NGS)**

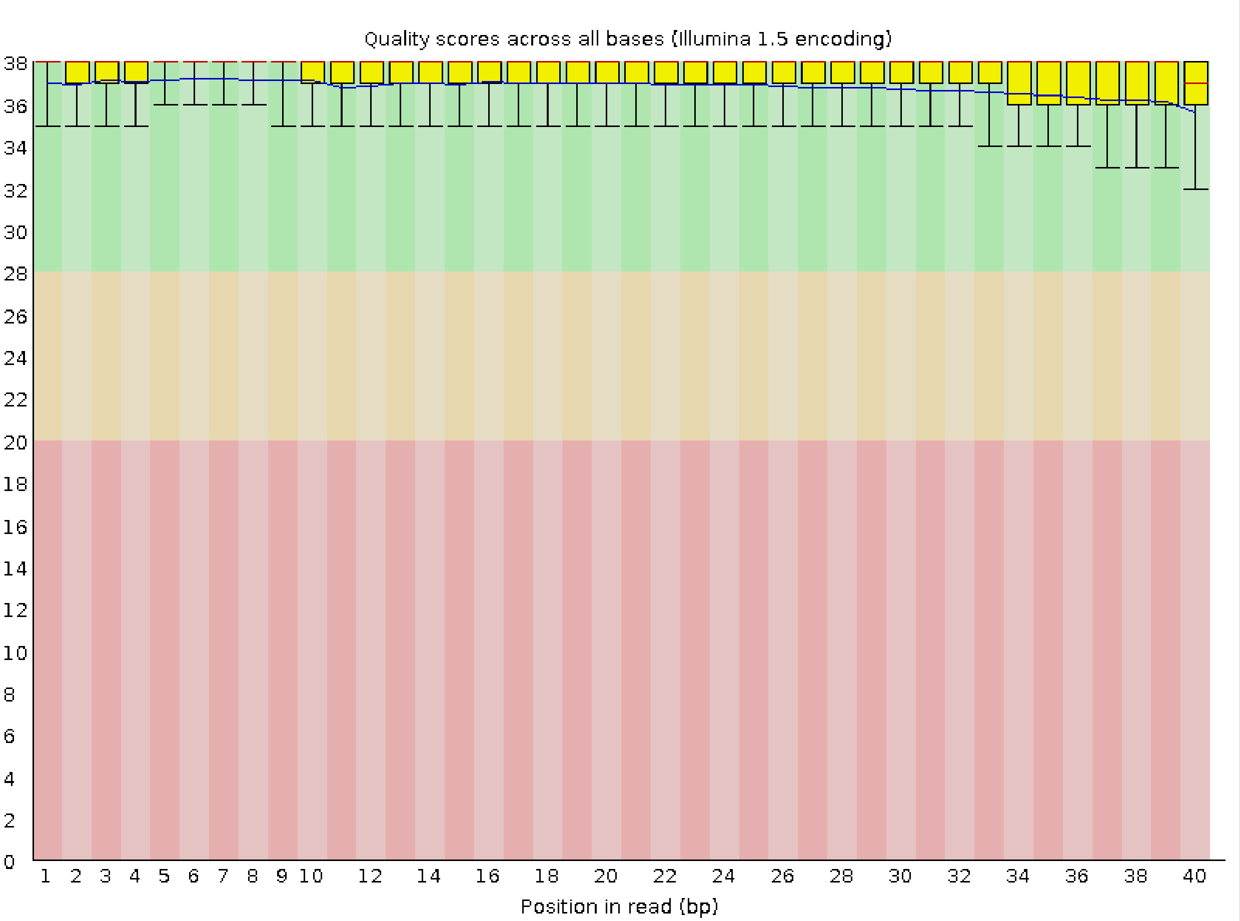
Quality Control

- Why?
- 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

Quality control measurements

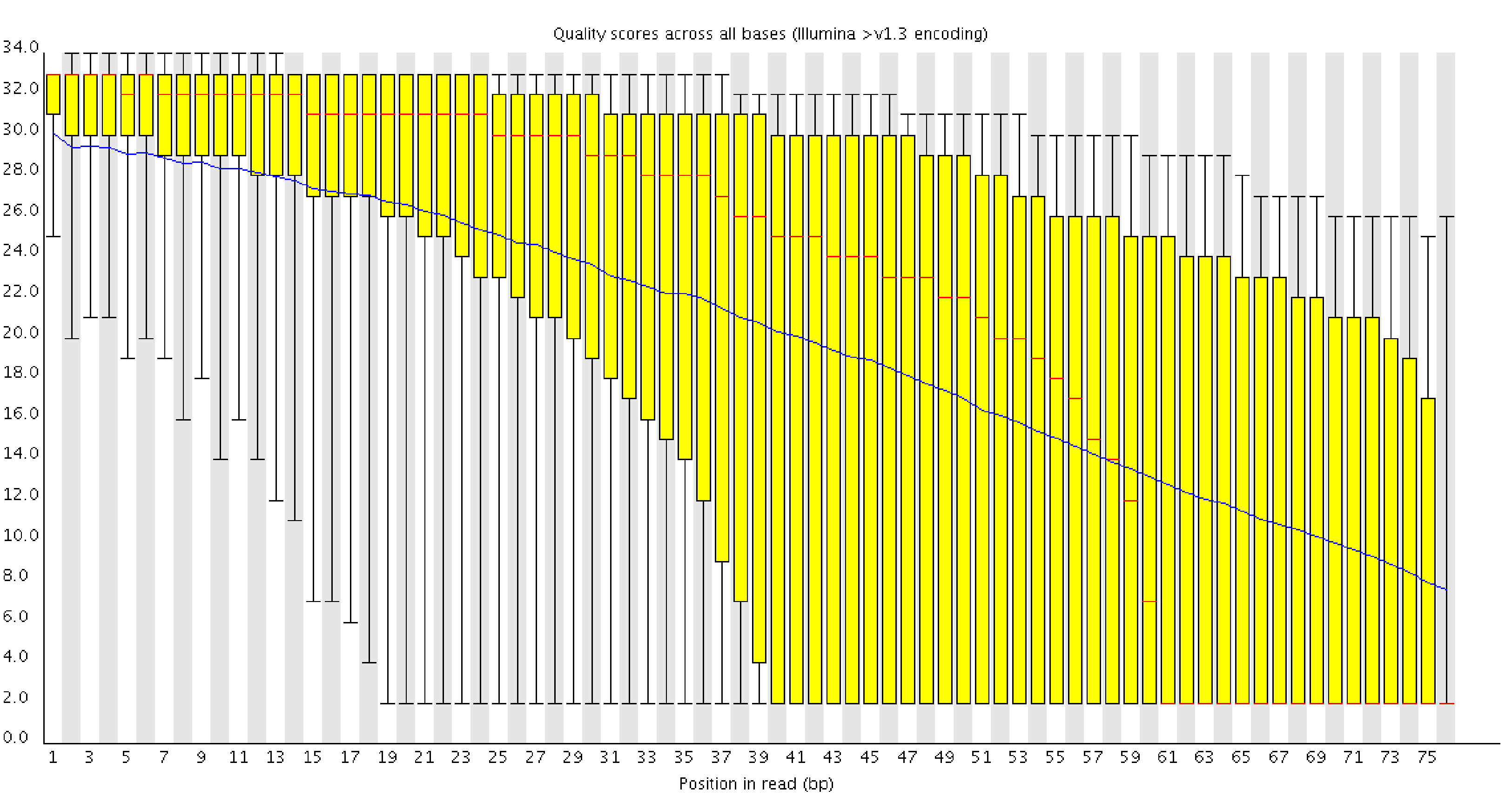
- Quality plots
  - Per base
  - Per sequence
- Composition plots
  - Per base composition
  - GC content and profile
- Contaminant identification
  - Overrepresented sequences and k-mers
  - Duplicate levels

Per base sequence quality



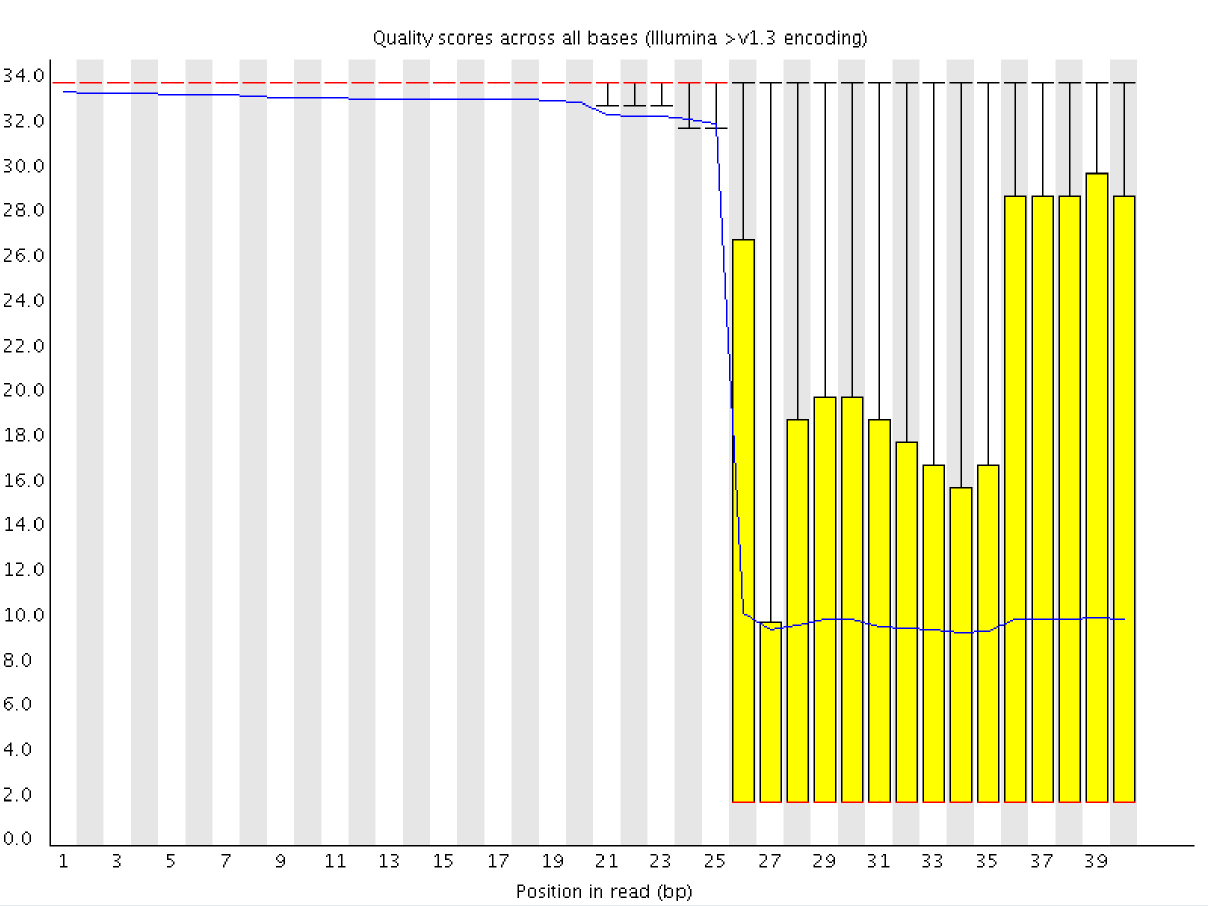
Quality drops gradually

- Typical for longer runs → trim the low-quality ends.

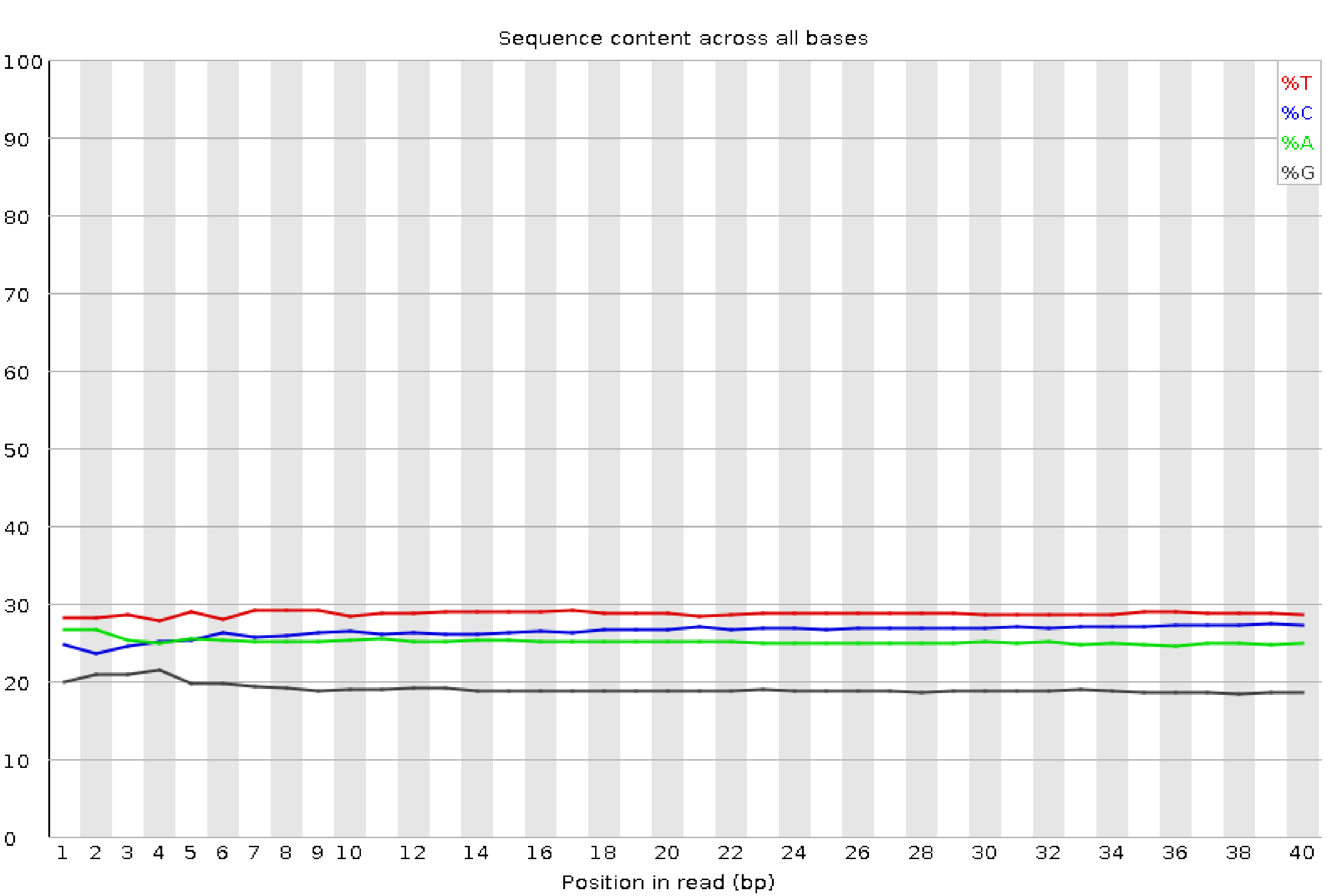


Quality drops suddenly

- Problem in the flow cell → trim the sequences

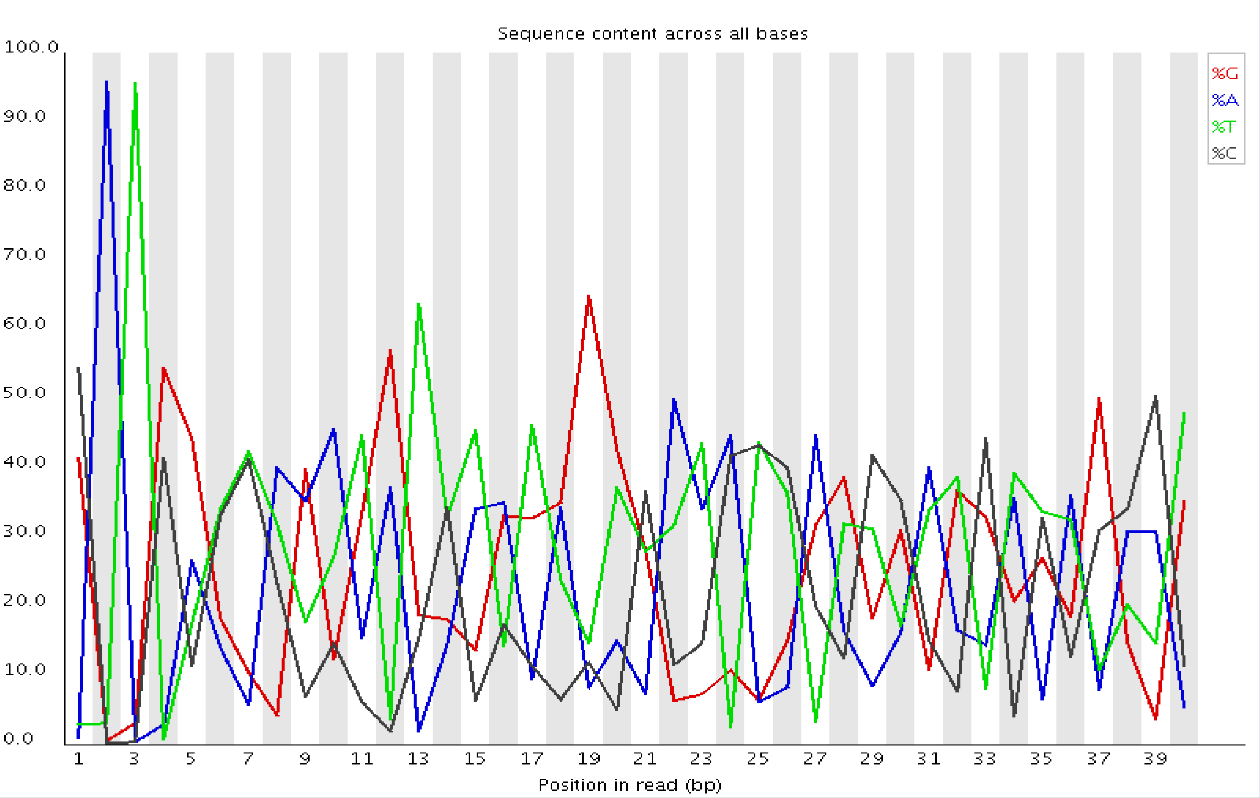


Per base sequence content



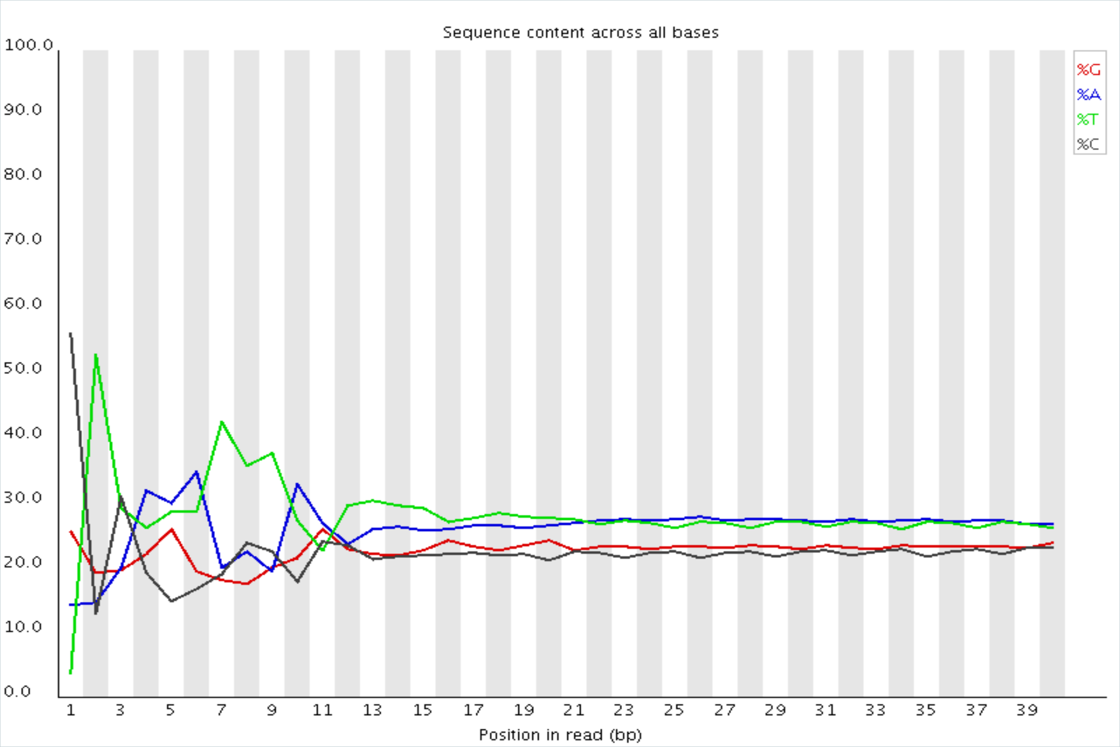
Biased sequence

- Library has a restriction site at the front
- A single sequence makes up of 20% of the library

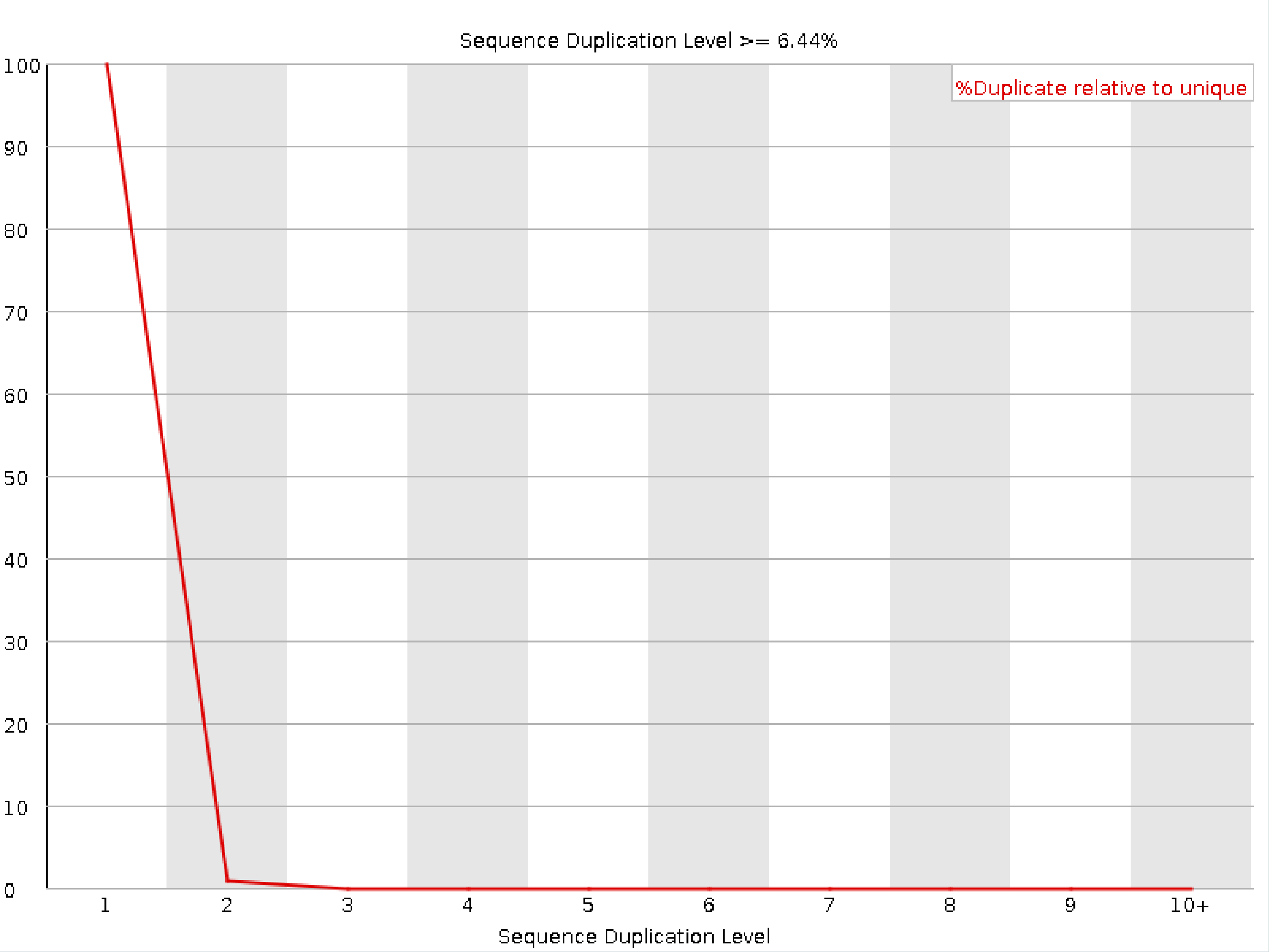


RNA-seq with Illumina

- “Random” primers, enzyme preferences?
- Correct sequence but biases your reads → keep in mind

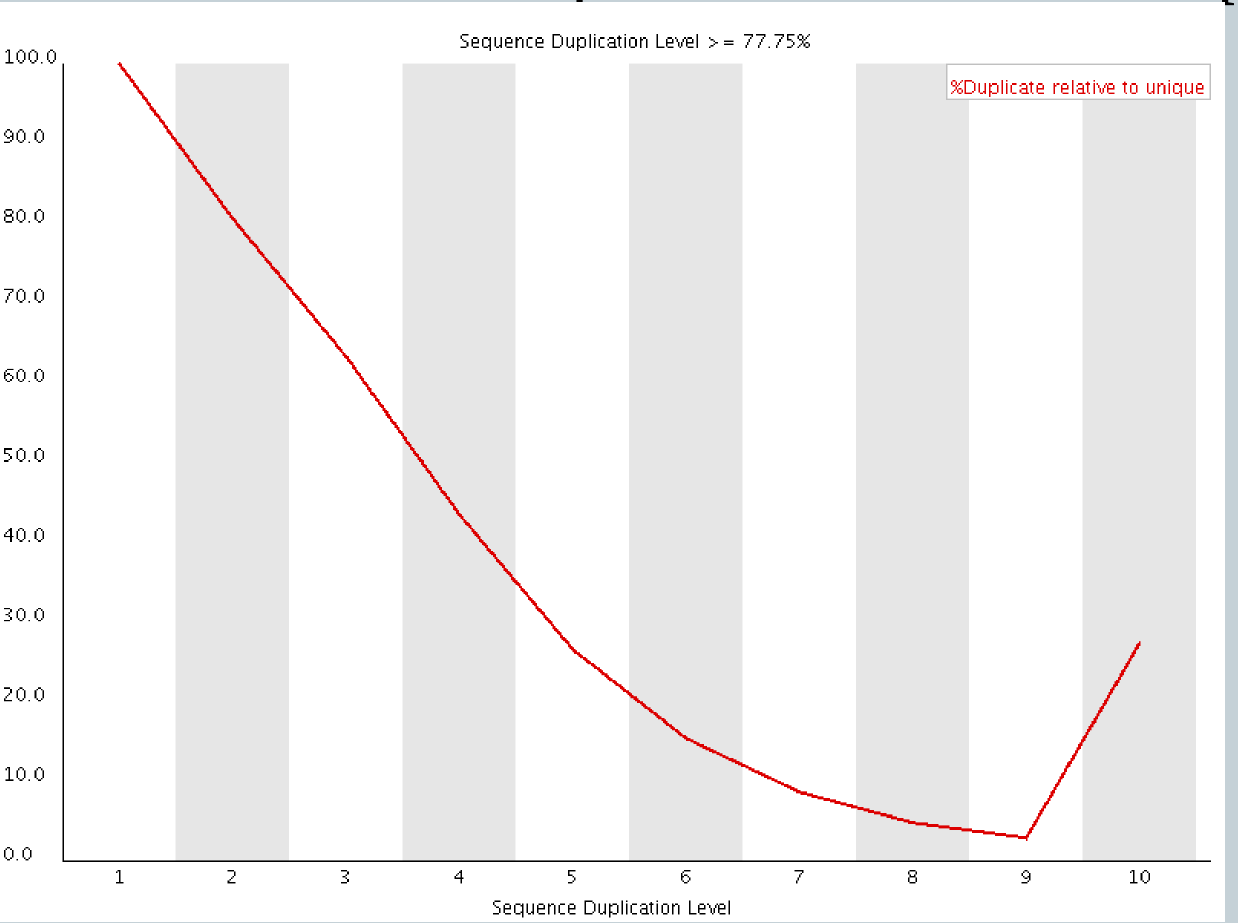


Sequence duplication level



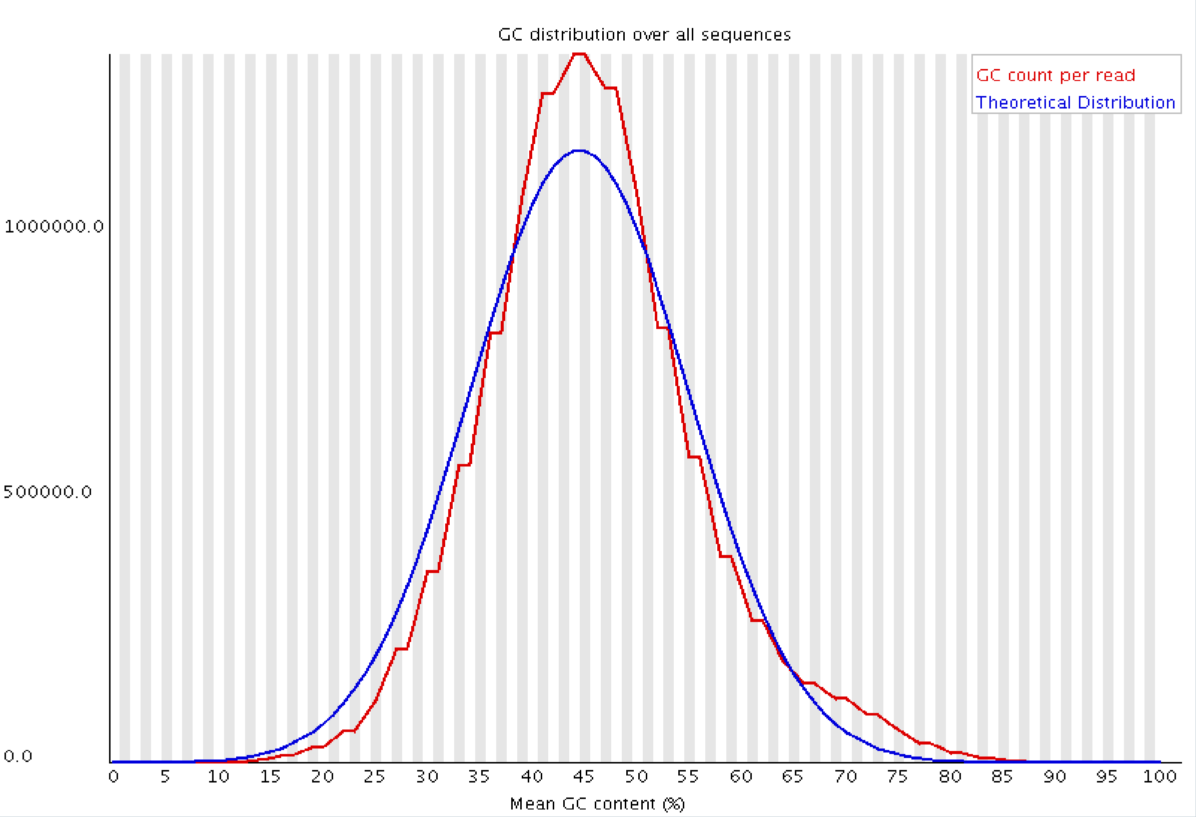
Duplicated reads

- Library has been over-amplified → remove duplicate reads

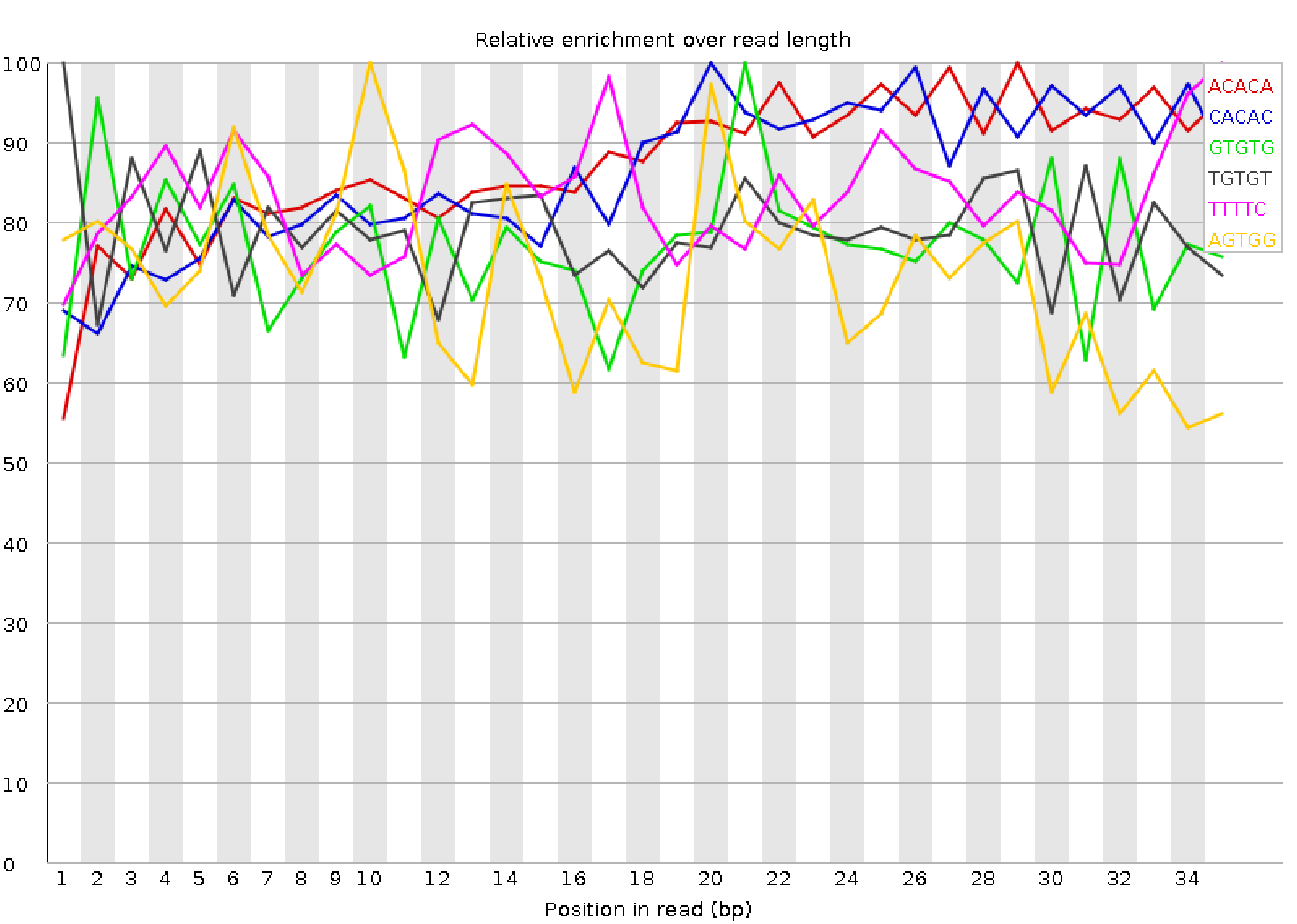


Per sequence GC content

- Median GC content is 45% instead of 42% → bacterial sequences in a human library

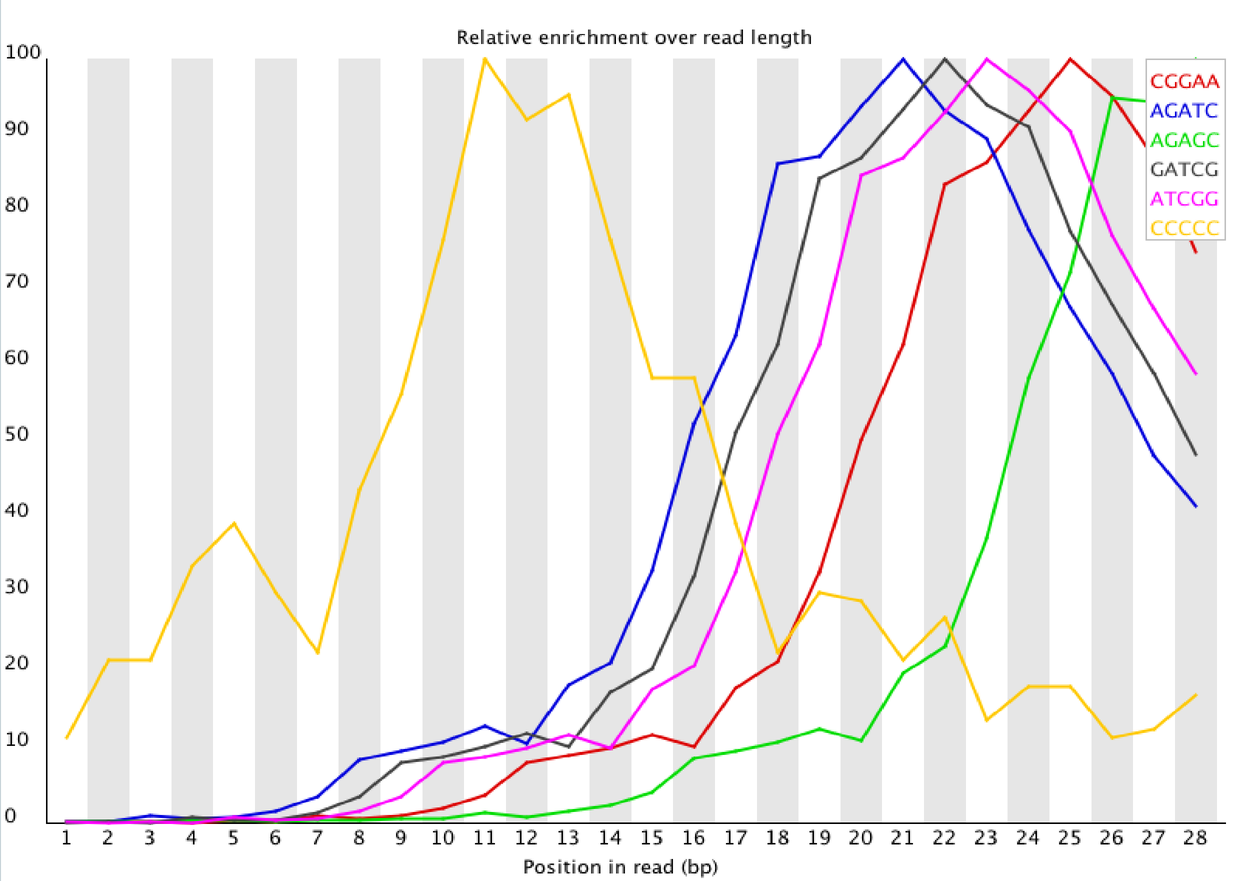


k-mer profile



k-mer enrichment rises towards the end

- Read contain partial Illumina adapter sequences → trim



Exercise 2: Quality control plots

- Go to the quality control category
- Select the tool “Read quality with FastQC” and click run
  - How long are the reads?
  - Up to what length is the quality acceptable?
  - Is the base content uniform all the way? If not, why?

Filter and trim low quality sequences: FastX

- Filter sequences based on quality
  - What is the minimum allowed quality
  - What percentage of bases in a read are required to have this quality or higher
- Trim all reads to a give n length

- Note that some aligners (like BowTie) give you the option to align only a part of the read

Exercise 3: Filter and trim reads

- Select the tool “Preprocessing / Filter reads for several criteria with PRINSEQ”, set the Quality cut-off value to 30 and run
  - How many reads were filtered out?
- Run again the tool “Read quality with FastQC”
  - Does the per base quality now look acceptable?
- Select the tool “Preprocessing / Trim reads with FastX”, set the last base to keep to 80 and run.
- Run again the tool “Read quality with FastQC”
- Which approach would you use to get rid of low quality sequence: trimming or filtering based on qualities? Why?

Exercise 4: Convert FASTq to FASTA

- Select the tools “Utilities / Convert FASTQ to FASTA” and run
- Open the result file. What happened to the qualities? What could you use this file for?

- **Exercise**
  - Import 1000readsFromRNAseq_2.fastq
  - Run quality control and try to salvage some good quality reads

- Save session with name qc.zip
- Select “New session”

Alignment to Reference

- Most NGS applications (apart from de novo assembly) require mapping the reads to a genome or transcriptome
  - RNA-seq
  - Re-sequencing, variant detection
  - ChIP-seq
  - Assembly by mapping
  - Methyl-seq
  - …

Software packages for alignment

- Bowtie, Bowtie 2 (available in Chipster)
- TopHat2 (available in Chipster)
- BWA (available in Chipster)
- MAQ
- SHRiMP
- …

- Differences in speed, memory consumption, handling indels and spliced reads

Bowtie

- Fast and memory efficient (Burrows-Wheeler index)
- Does not support gapped alignments
- Two modes
  - (n) Limit mismatched only in a user-specified seed region.
  - (v) Limit mismatches across the whole read
- Careful: the default parameters are dangerous:
  - Use “-best” to get the best alignment if there are several
  - Use “strata” to get only alignments of the best class

Exercise 5: Align reads to genome

- Import the files:
  - `e_coli_1000.fq`
  - `NC_008253.fna`
  - Select both files by keeping the Ctrl key down
- Select “Alignment / Bowtie2 for single end reads and own genome”
  - In the parameters, check that read and genome files are correctly assigned. Click run
  - How many reads were aligned?
  - Play with the parameter settings (number of mismatches, allowed number of hits). Do you get more alignments?
- Save the session with name ecoli.zip

Visualization

- Why?
  - Nothing beats the human eye in detecting potentially interesting patterns in the data
- Software packages for visualization
  - Chipster genome browser ☺
  - IGV
  - GenomeView
  - UCSC Genome Browser
  - …
- Differences in memory consumption, interactivity, ability to edit, annotation, contig view,….

Chipster Genome Browser

- Integrated with Chipster analysis envirnoment
- Automatic sorting and indexing of BAM and BED
- Automatic coverage calculation
- Zoom in to nucleotide level
- Highlight SNPs
- Support for spliced reads
- Jump to locations using a BED file
- Several views (reads, coverage profile, density graph)
- Low memory requirements

Exercise 6

- Open session `ChIP-seq_STAT1.zip`
- Open the file `positive-peaks.bed`, detach it, and put it down
- Select 5 files:
  - `treatment.bam and treatment.bam.bai`
  - `control.bam and control.bam.bai`
  - `positive-peaks.bed`
- In the visualization panel, select “genome browser”
- Select genome hg18, set the scale to 100, type gene “RNF115” in the location field and click go

Exercise 7: Use Chipster genome browser

- Zoom in to nucleotide level, select “highlight SNPs”
- Look at all the reads by selecting “Show full height”. Then unselect this.
- Zoom out a little and select strand-specific coverage to see the shape of the peaks. Move sideways.
- Bring the detached bed file up. Sort it by the last column, and navigate through the most significant peaks by clicking at the start position.
- Close the session.

Exercise 8: Count reads per miRNAs

- Import session miRNA-seq.zip
- Select files bowtie.bam and miRBase16-preprocessed.bed
- Select tool “”, check that the input files are correctly assigned, and run.
- Open the output file to see what columns it has.

Exercise 9: Look at edgeR result files

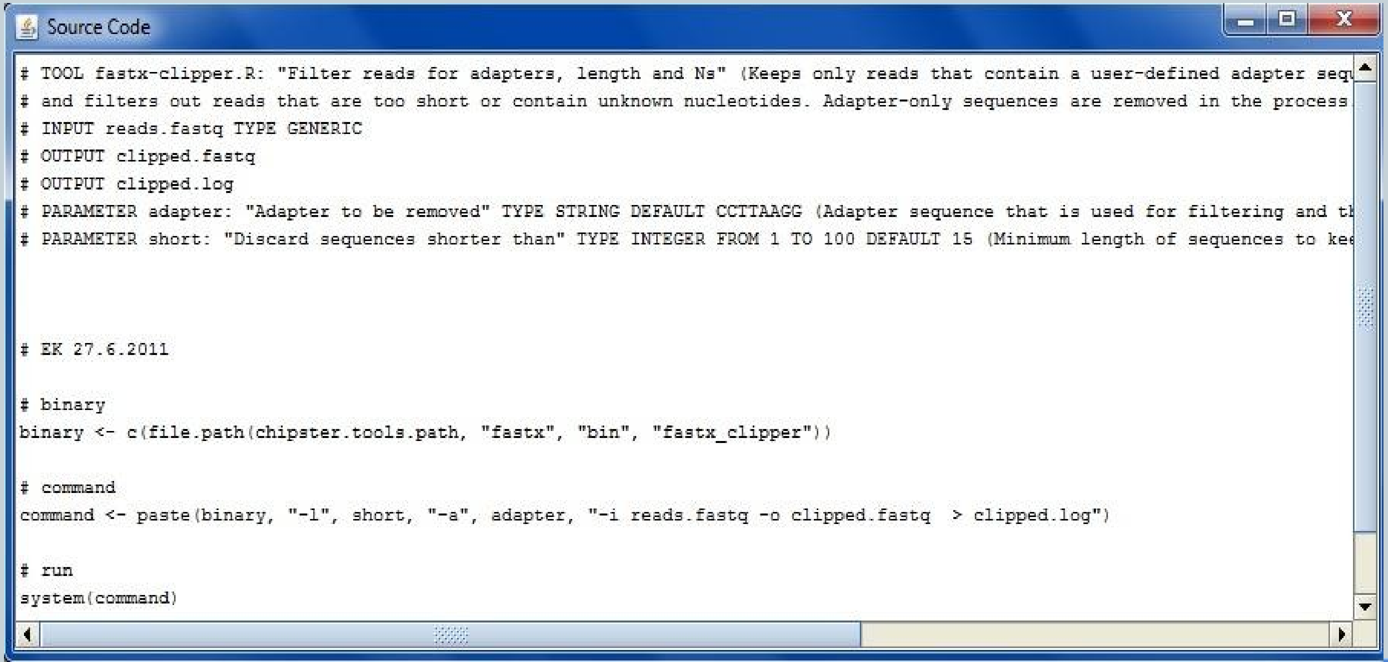
- Your current session miRNA-seq.zip contains an analysis of differentially expressed miRNAs. Open the edgeR result files to study how they look like.
- Import, open and detach the file miRNA-seq.bed
- Use the genome browser to visualize the genomic alignment and miRNA-seq.bed. Use the previously detached bed file to go to mir-370.

Summary

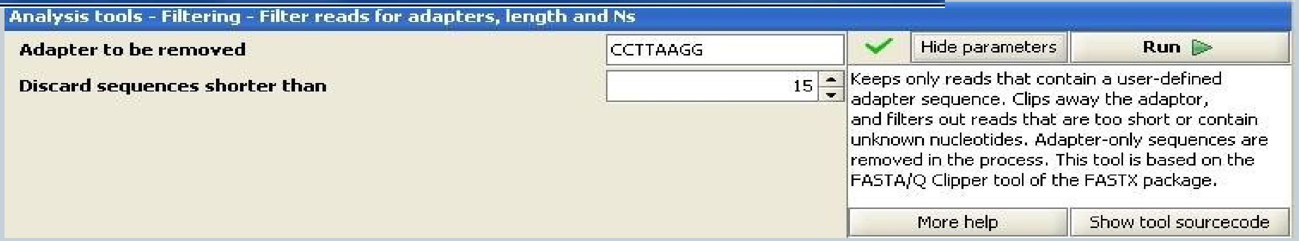
- What can I do with Chipster?
  - Wet-lab scientist
    - Analyze, visualize and integrate your data
    - Share workflows and analysis sessions with colleagues
  - Bioinformatician
    - Offload routine tasks to wet-lab researchers
    - Prepare workflows for them
    - Customize Chipster for your users by adding new tools
  - Analysis method developer
    - Easy way to provide a GUI for your tool,thereby enlarging the user community.

Easy to add analysis tools

- Command line, R-based, web-services



```
# TOOL fastx-clipper.R: "Filter reads for adapters, length and Ns" (Keeps only reads that contain a user-defined adapter sequ
# and filters out reads that are too short or contain unknown nucleotides. Adapter-only sequences are removed in the process
# INPUT reads.fastq TYPE GENERIC
# OUTPUT clipped.fastq
# OUTPUT clipped.log
# PARAMETER adapter: "Adapter to be removed" TYPE STRING DEFAULT CCTTAAGG (Adapter sequence that is used for filtering and th
# PARAMETER short: "Discard sequences shorter than" TYPE INTEGER FROM 1 TO 100 DEFAULT 15 (Minimum length of sequences to kee



# EK 27.6.2011

# binary
binary <- c(file.path(chipster.tools.path, "fastx", "bin", "fastx_clipper"))

# command
command <- paste(binary, "-l", short, "-a", adapter, "-i reads.fastq -o clipped.fastq  > clipped.log")

# run
system(command)
```


Analysis tools - Filtering - Filter reads for adapters, length and Ns

Adapter to be removed: CCTTAAGG

Discard sequences shorter than: 15

Keeps only reads that contain a user-defined adapter sequence. Clips away the adaptor, and filters out reads that are too short or contain unknown nucleotides. Adapter-only sequences are removed in the process. This tool is based on the FASTA/Q Clipper tool of the FASTX package.

Acknowledgments

- Kimmo Mattila
  Application specialist, CSC
- Diego Scardaci
  Technical Outreach Expert, EGI.eu
- EGI FedCloud Resources
  GRNET, CESNET
- All the people at CERTH/INAB and AUTH/IPL that made this workshop happen! ☺

Thank you for your patience!