Bioinformatics Analysis in R

Day 5
Next Generation Sequencing (NGS) Data Analysis and Visualization

Ivan G. Costa, Zhijian Li

Institute for Computational Genomics
RWTH University Hospital
www.costalab.org

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Outline

- Introduction to NGS data analysis pipeline
  - Quality check
  - Alignment
  - Higher level analysis (peak calling)
  - File formats
- Visualization of NGS data using IGV
  - RNA-seq, ChIP-seq, …
  - IGV tools
- Practice

Bioinformatics Analysis in R

Next Generation Sequencing

Institute for
Computational Genomics
RWTH AACHEN
UNIVERSITY

Sequencing

- Read the bases of a DNA/RNA sequence
- Applications
  - Sequence DNA of known or unknown organism
  - Detect variants on patients
  - Sequence the RNA of a cell
  - Detect location of proteins interacting with DNA
- Problem
  - Only short DNA sequences (< 1000 bps) can be read
- Solution
  - Bioinformatics

Information Level vs. NGS

genetic information level

DNA Sequencing
-> detection of genetic variants
-> de-novo reconstructions of genomes

RNA Sequencing
-> quantification of RNA in a cell
-> de-novo identification of RNAs

Detection of Interactions:

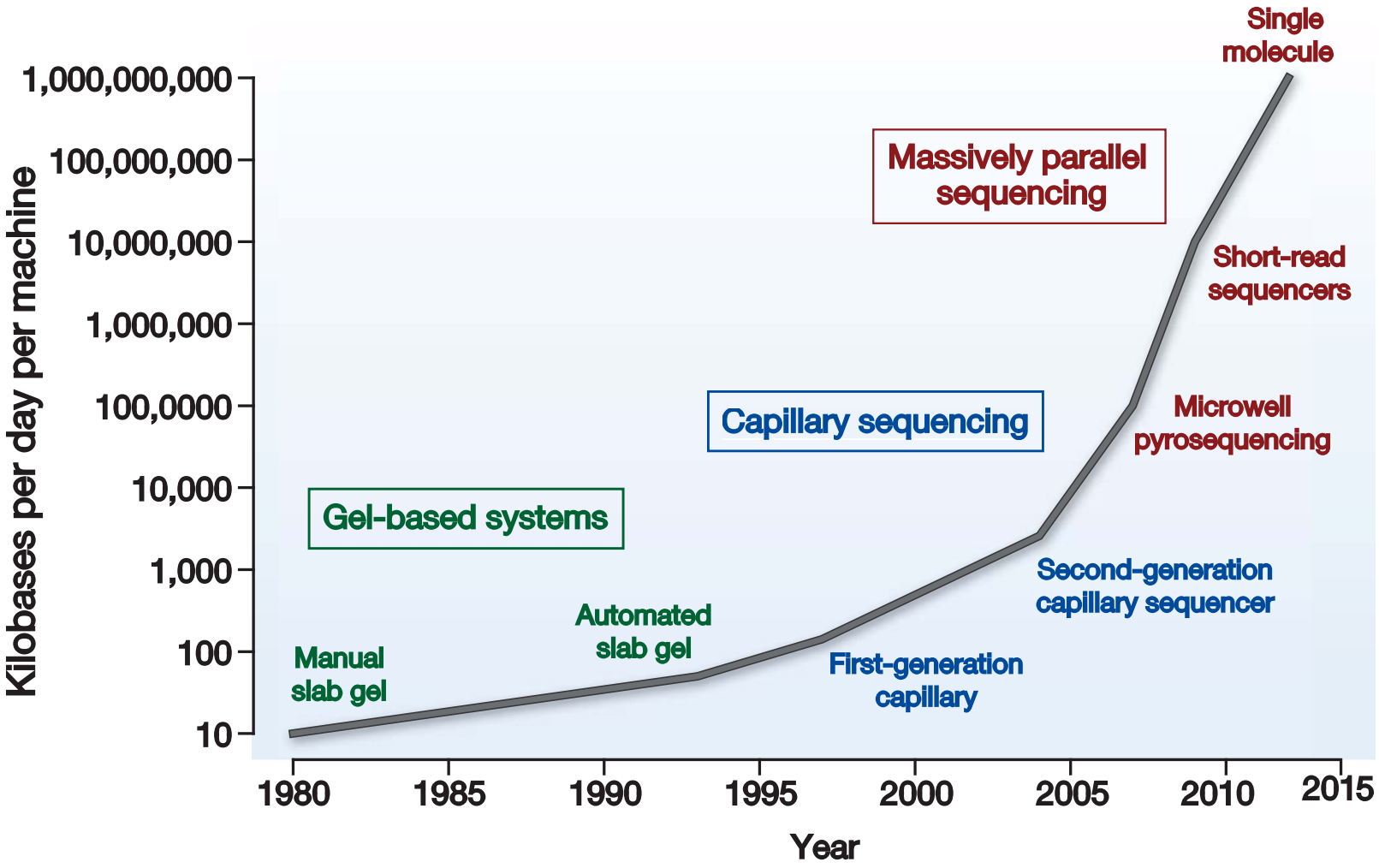
- ChIP Sequencing -> a protein with DNA
- CLIP Sequencing -> a protein with RNA
- ChIRP Sequencing -> a RNA with DNA
- …

See here for a comprehensive list of Seq essays (>50)
https://liorpachter.wordpress.com/seq/

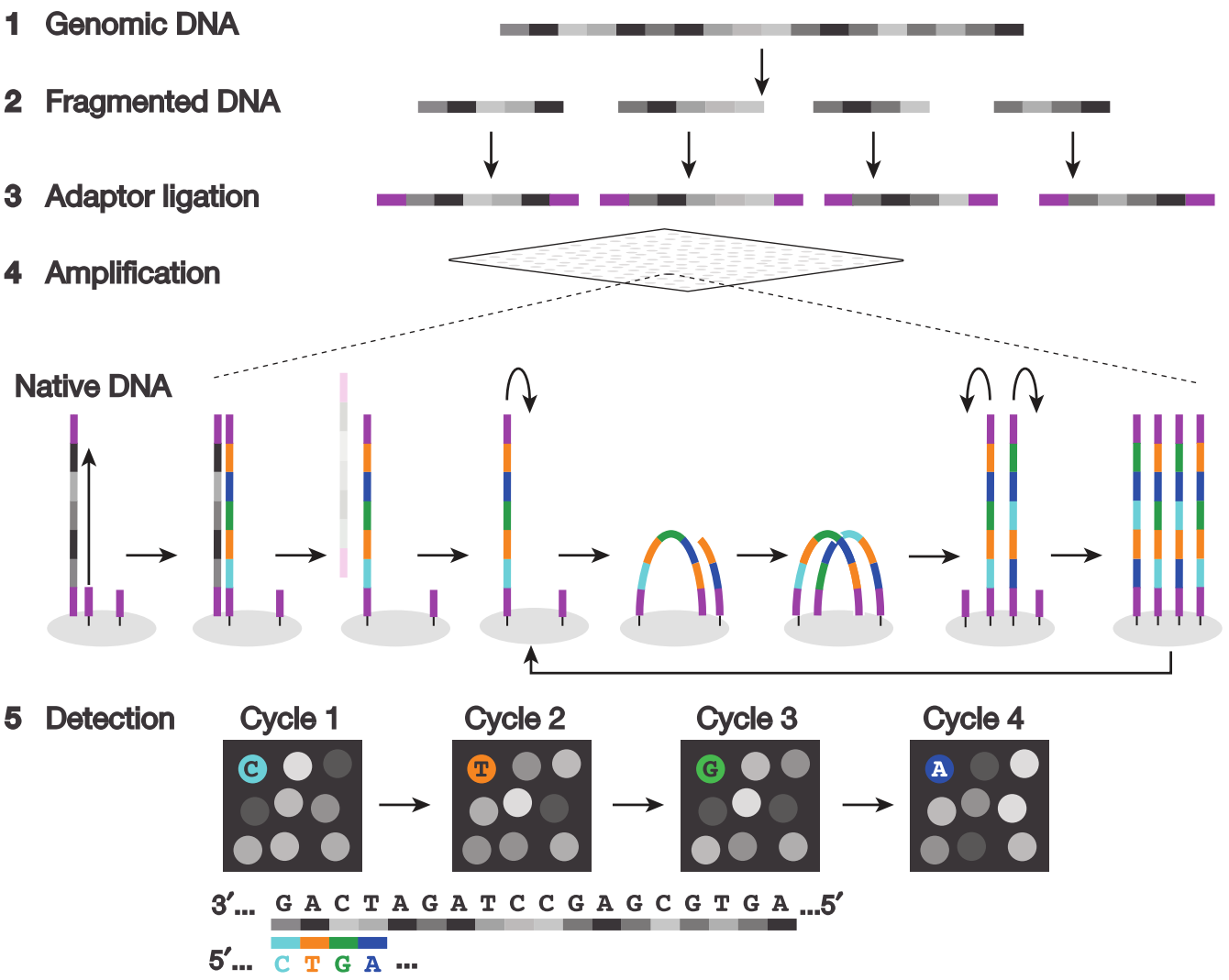
Next Generation Sequencing

- NGS take advantage of parallelization
  - reads millions/billions of reads for a time
  - shorter reads (50-100 bps)
  - higher error rates (0.1-1%)
- commercial products:
  - 454
  - SOLiD
  - Solexa (Illumina)

Next Generation Sequencing



Next Generation Sequencing

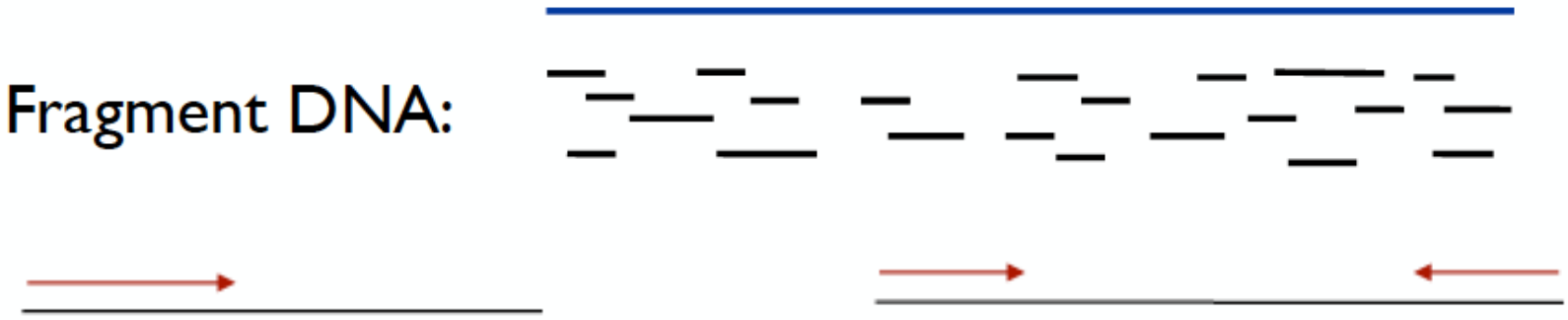


Jay Shendure, *et al (2017), Nature*

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

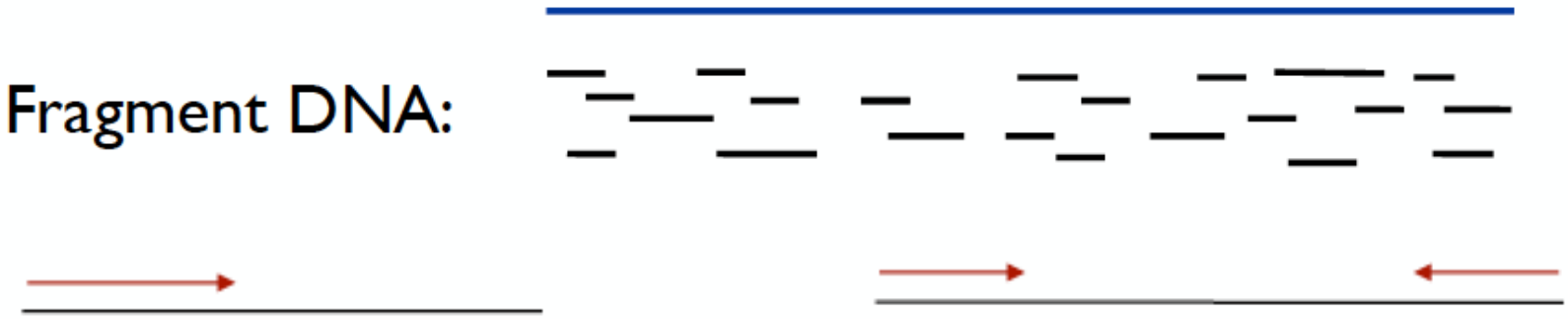
Read Types



Single end

Paired end
Ins: 200-800 bp

Read Types



Single end

Advantages:
- cheaper
-compatible with protocols producing small fragments (Ribo-seq, mirRNA-seq)

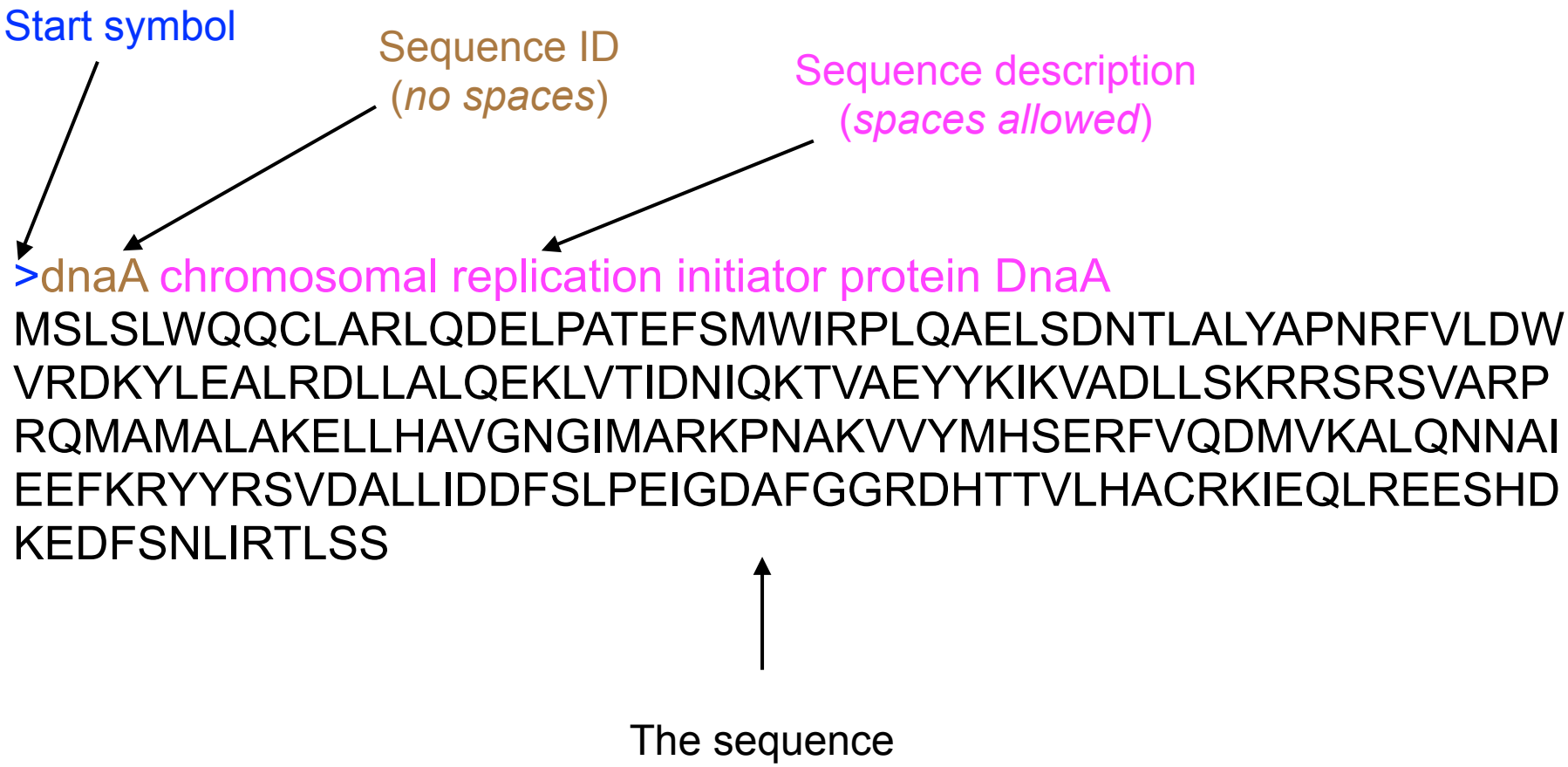
Paired end
Ins: 200-800 bp

Advantages:
- easier to align
-helps detection of variants (DNA), exon pairs (RNA)

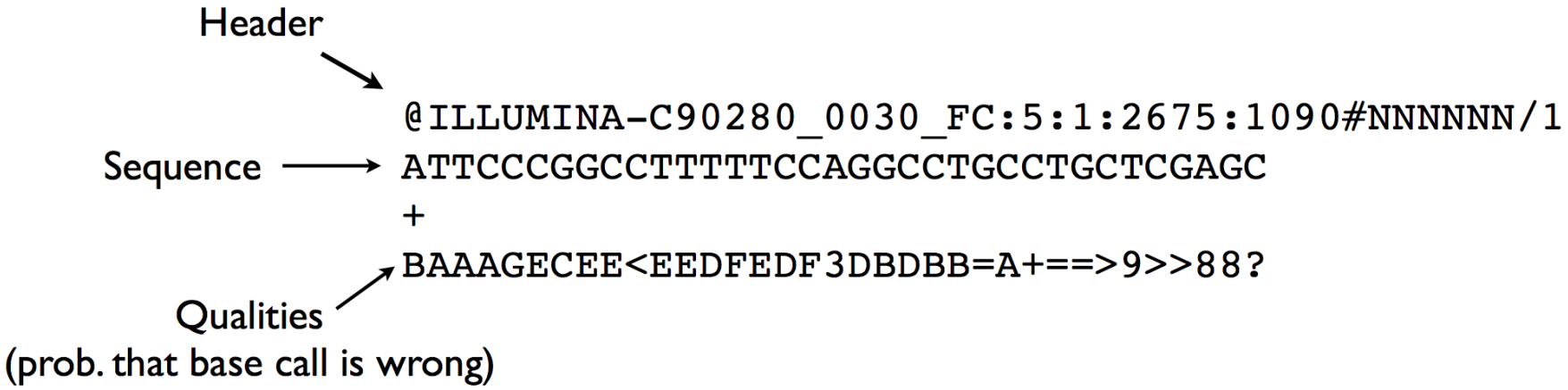
FASTA files

>dnaA chromosomal replication initiator protein DnaA
MSLSLWQQCLARLQDELPATEFSMWIRPLQAELSDNTLALYAPNRFVLDW
VRDKYLEALRDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRRSRSVARP
RQMAMALAKELLHAVGNGIMARKPNAKVVYMHSERFVQDMVKALQNNAI
EEFKRYYRSVDALLIDDFSLPEIGDAFGGRDHTTVLHACRKIEQLREESHD
KEDFSNLIRTLSS

FASTA files



FASTQ files



One character encodes a number using ascii table (0-255)

This number (*Q*) can be converted to *P*

Phred-scale

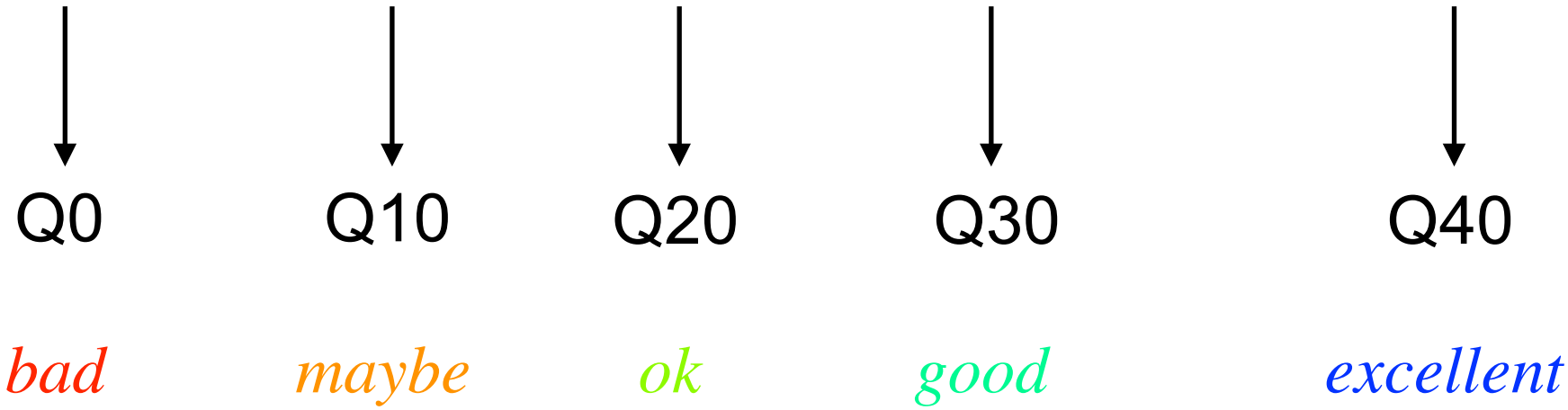
Q = -10 * log10 P

P = 10^(-Q/10)

https://en.wikipedia.org/wiki/Phred_quality_score

FASTQ files

Uses letters/symbols to represent numbers:

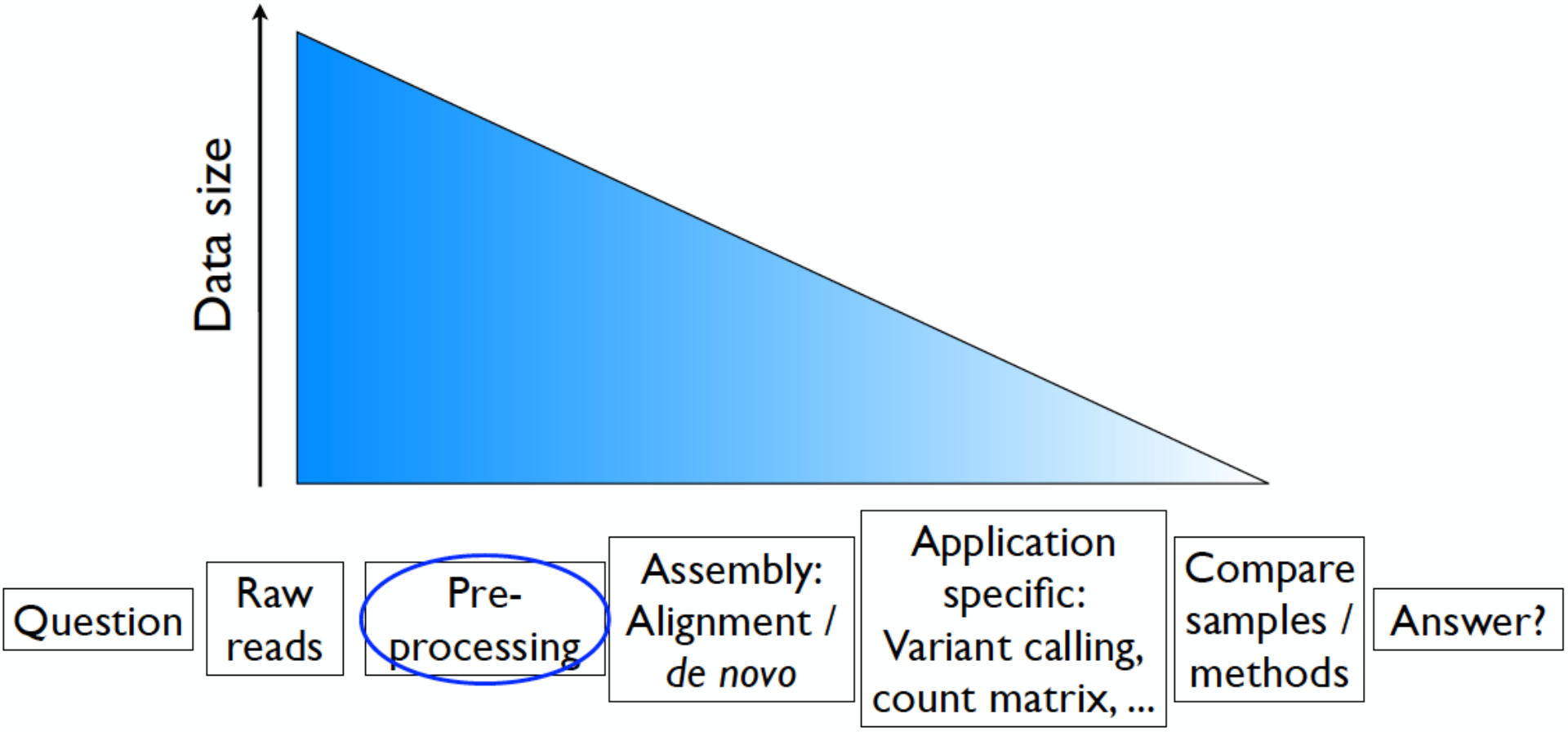


Bioinformatics Analysis in R

Next Generation Sequencing Data Analysis

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Pre-processing



Pre-processing

- Sequencing and sample preparation introduce errors
  - Errors in start/end of reads
  - Bases bias on read positions
  - Presence of adapter sequences
  - Fragment duplication from PCR
  - …
- Tools: FastQC (for checking), Trimmomatic (for trimming), …

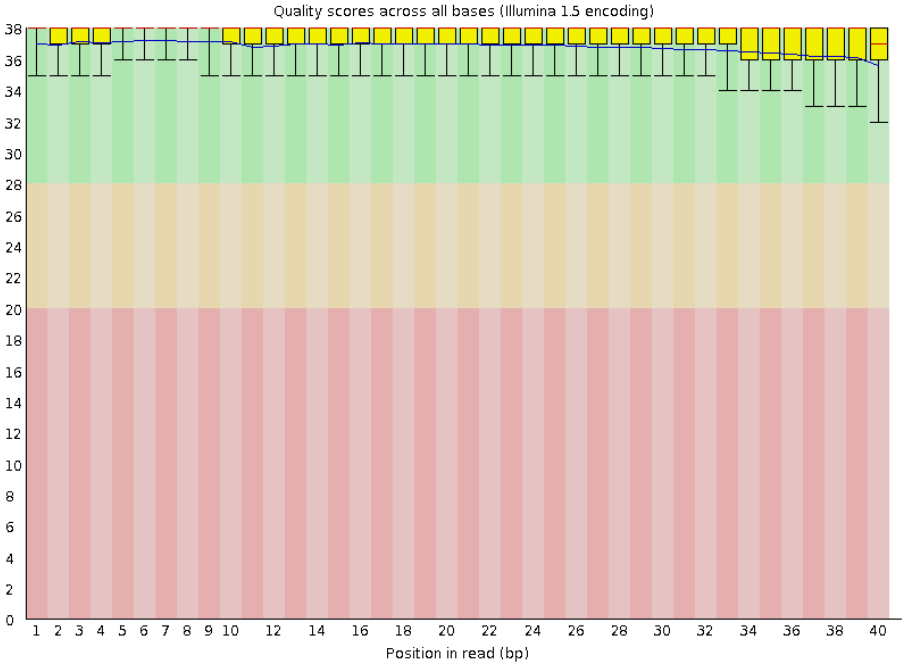
Quality Control

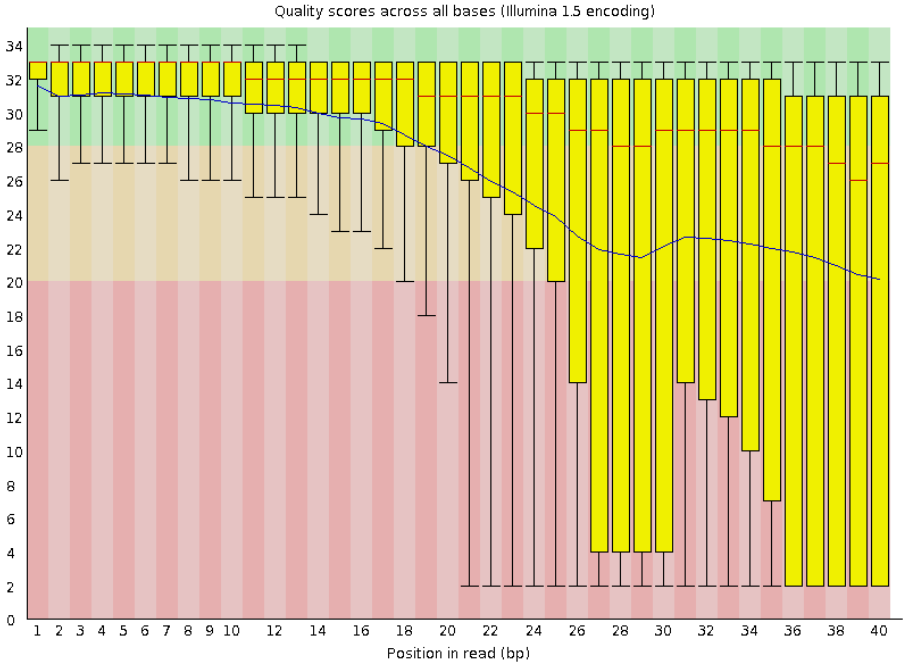
- **FastQC (usually provided by NGS core facilities)**
  - tool to analyse quality of reads from sequencing.
  - indicate problems in library preparation or sequencing steps.
- Example of good sequences:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
- or bad sequences:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Quality Control

Sequencing quality decreases with size.

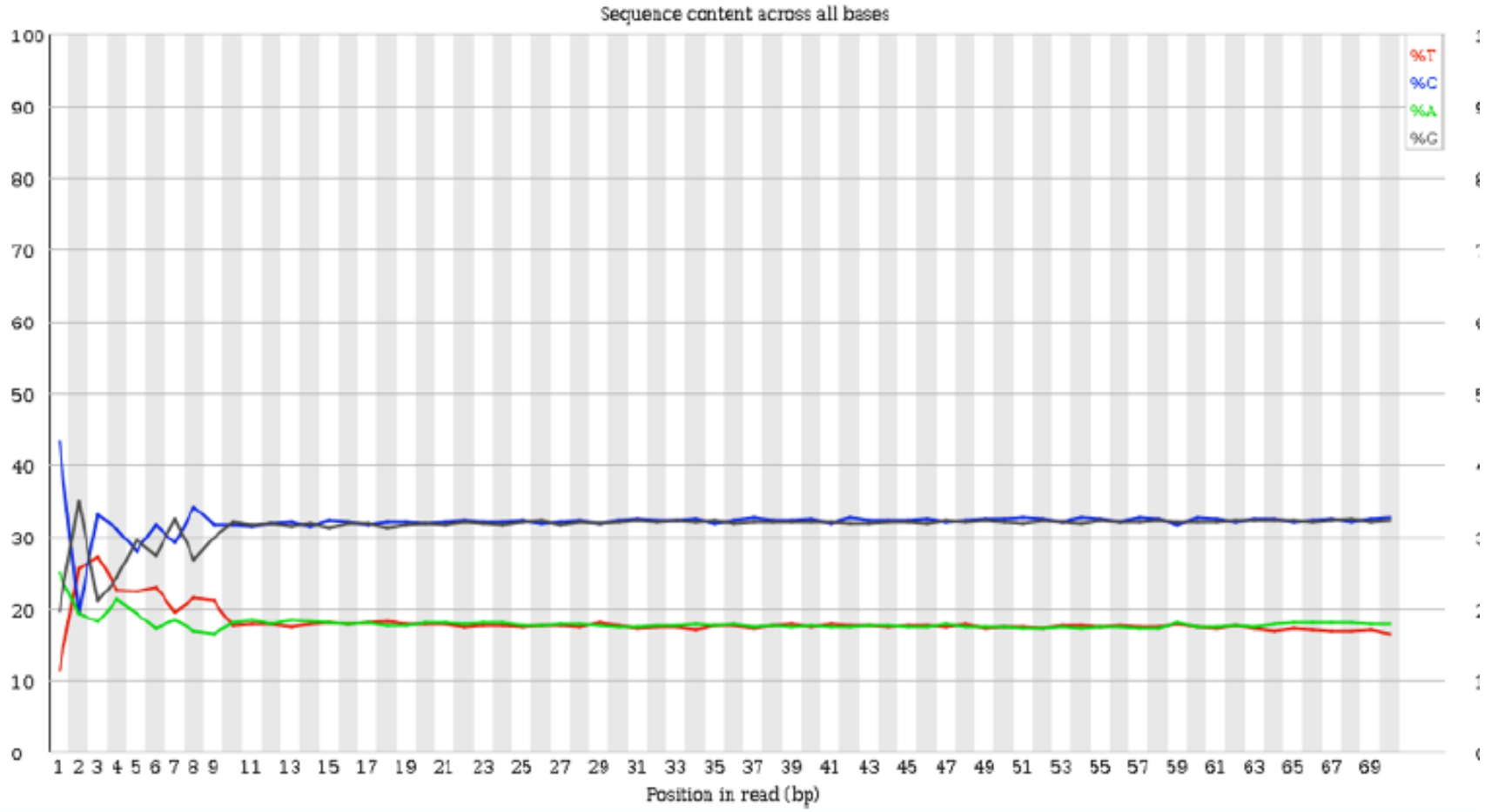




Solution: trim end of reads with low quality

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

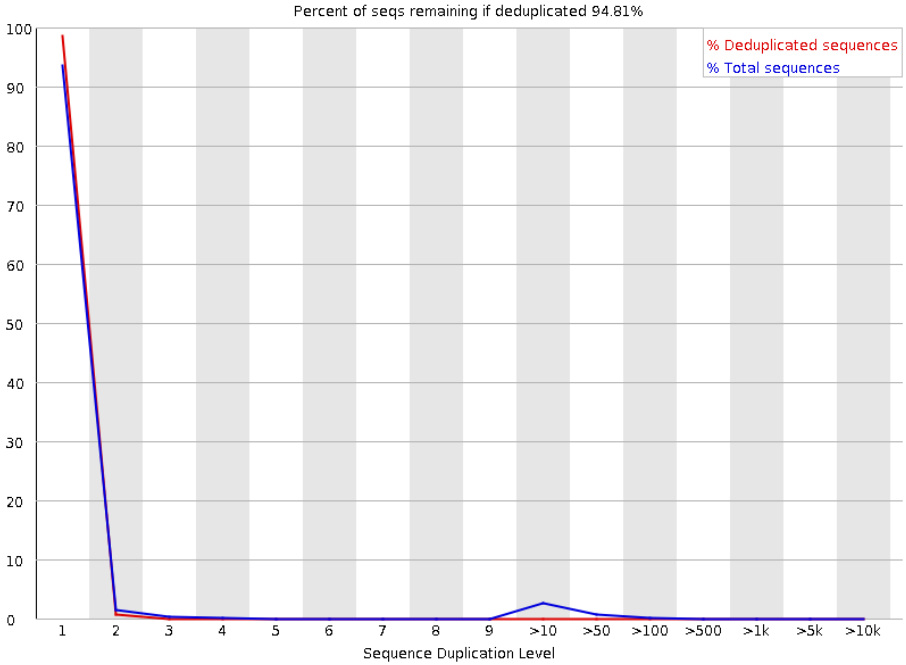
Read position sequence bias

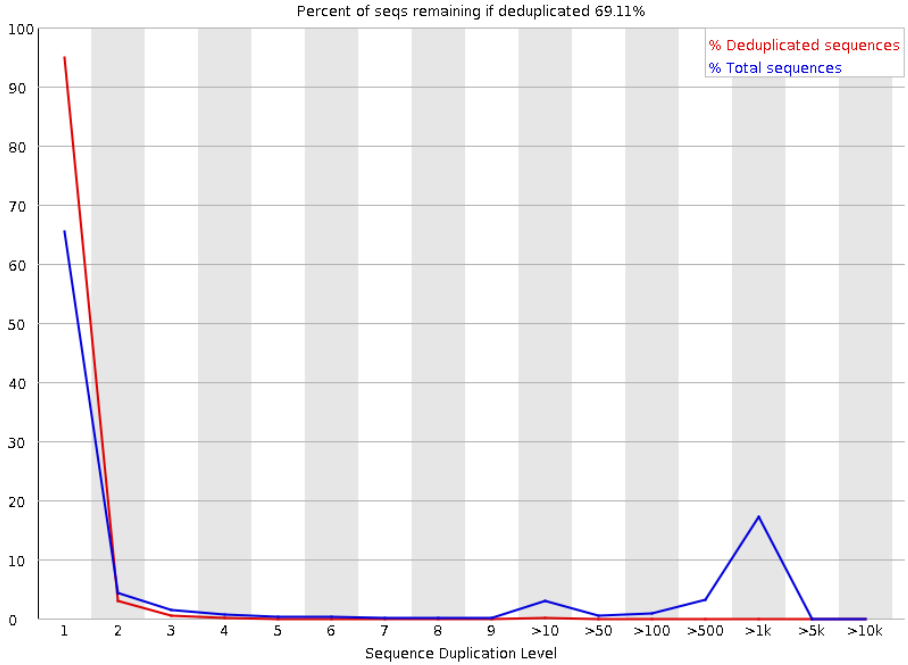


- **Trim read starts**

Quality Control

Sequence duplication levels





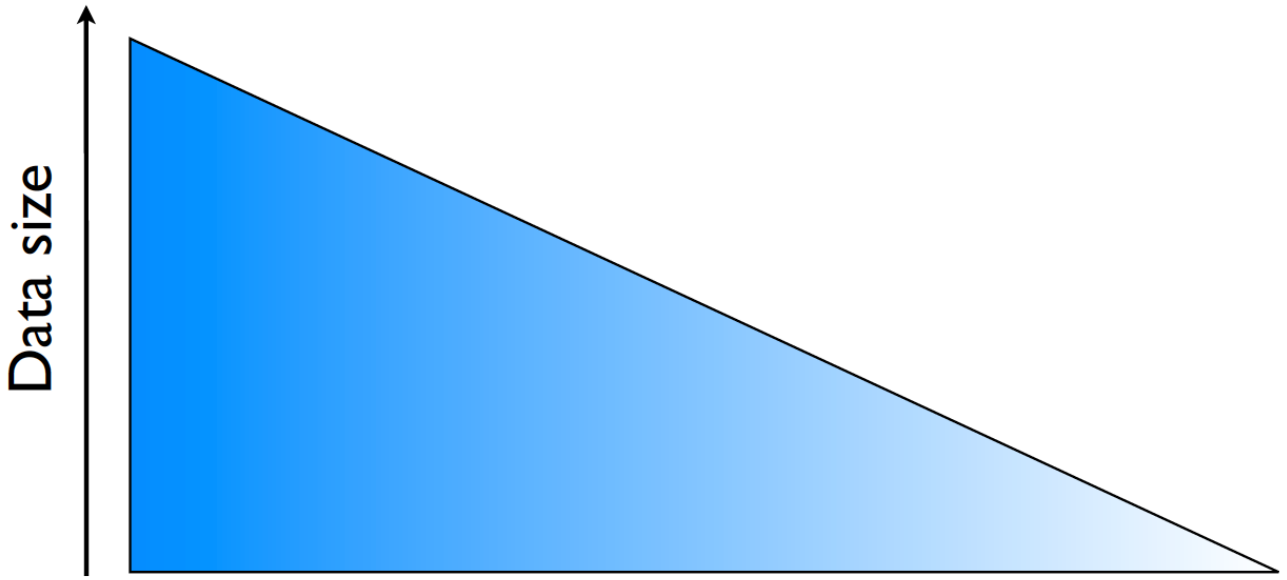
- **Solution: remove duplicates**

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

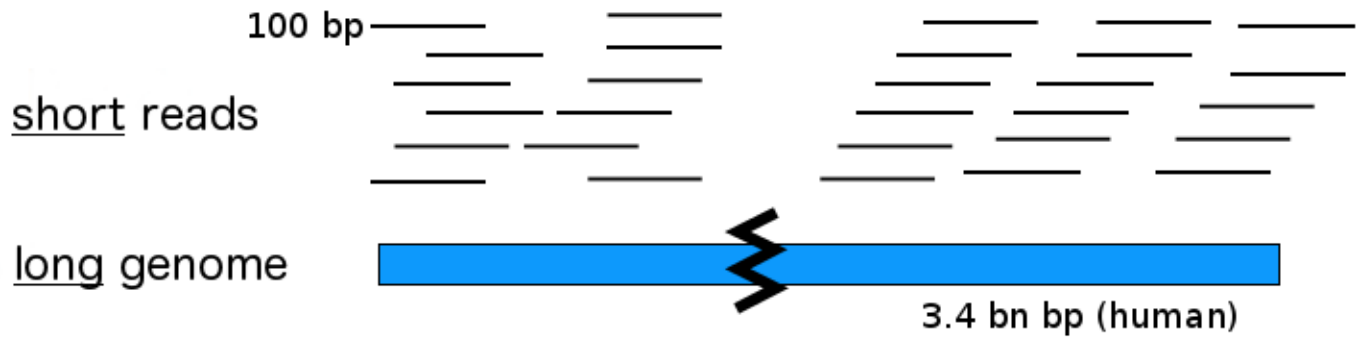
(Short reads) Alignment



Question | Raw reads | Pre-processing | Assembly: Alignment / *de novo* | Application specific: Variant calling, count matrix, ... | Compare samples / methods | Answer?

Short Read Alignment

- Query
  - sequenced reads in FASTQ format
  - huge number of them, 1M ~ 100M
  - short read length, ~100 bp
- Reference
  - human genome in FASTQ format
  - total size ~3 billion bps
- Lots of short vs. a few longs
  - BLAST would take several years to run.



Pitfals

- (Unknown) divergence of sample and reference genome
- Poor genome reference quality
- Repeats in the genome (larger than read size)
- Recombinations
- Sequencing/read errors

Algorithms - Alignment

Short read alignment is a special problem

- **reference sequence (genome) is large and fixed**
- **query sequence (reads) are short and many**

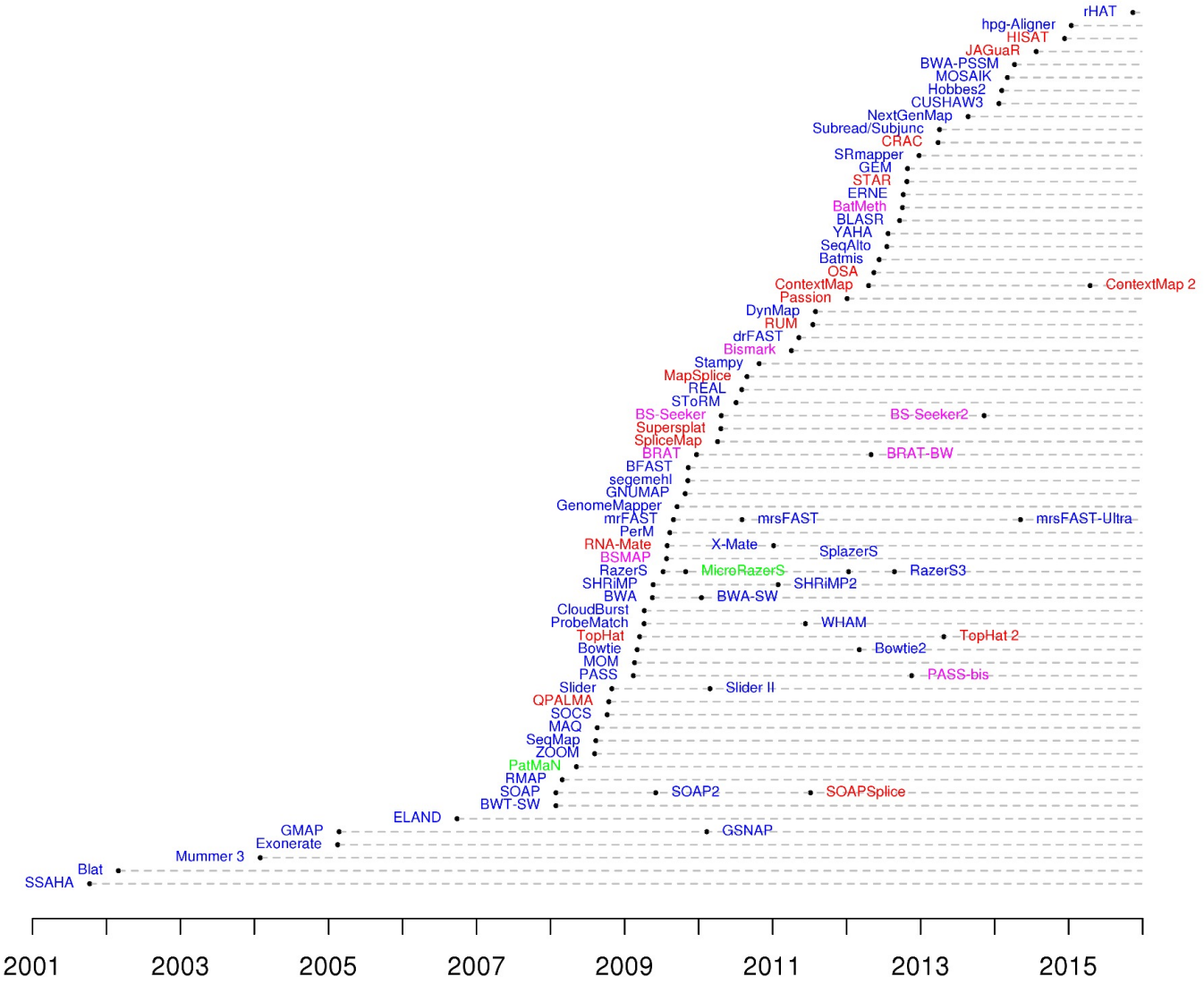
Solution:

1. Pre-process the genome finding all exact alignments for small sequences (>14bps) (index)

- **k-mer hash table (>10GB)**
- **compressed suffix trees (> 4GB)**

2. Break your read in small pieces (>14bps) and extend your alignment on all candidate positions using dynamic programming.

Alignment Tools



https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software

Institute for Computational Genomics

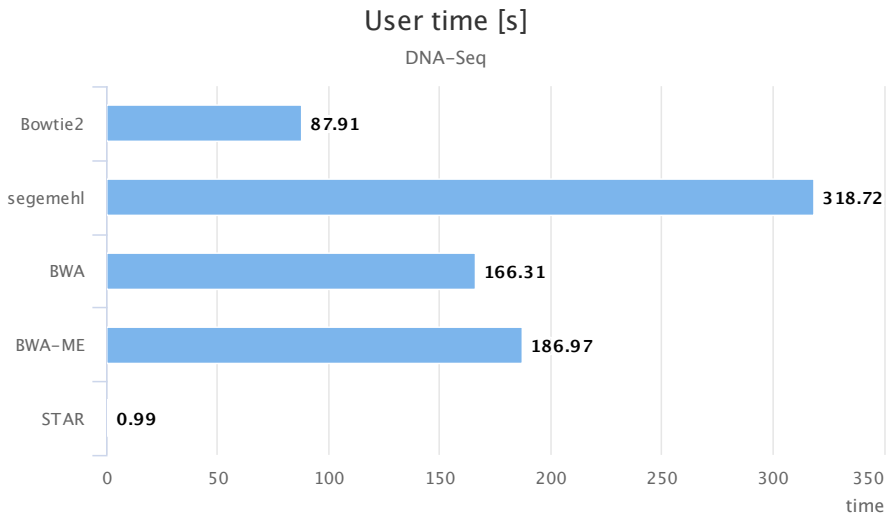
RWTH AACHEN UNIVERSITY

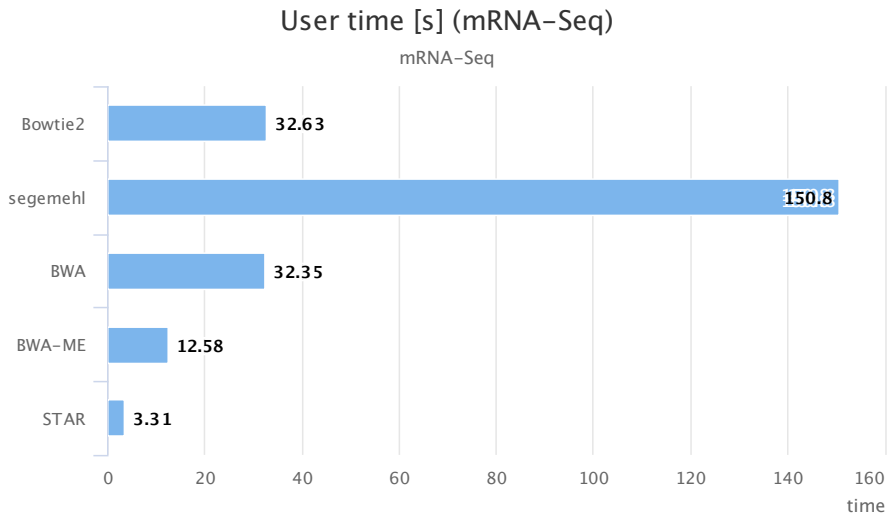
Reference based aligners - Overview

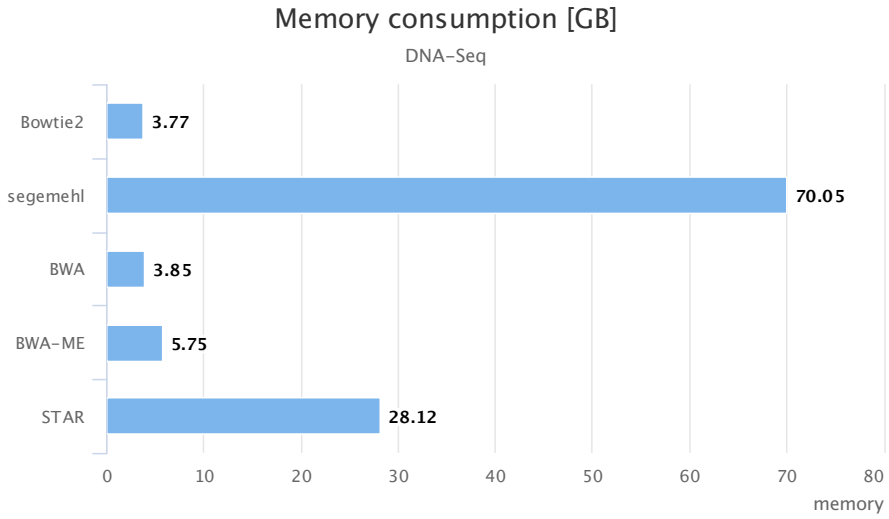
	Time	Precision	Pairs	GAPs	Phred	Memory	Application (Comments)
BOWTIE	+		+	-	-	5GB	General (max. 3 missmatches)
BWA	+		+	+	+	8GB	General (max of 200bps reads)
NOVOALIGN		+	+	+	+	8GB	General (commercial license)
STAR	+		+	-	+	32GB	RNA-Seq (allow split-maps)
BISMARK	+		+	+	+	10GB	Bisulfite/reduced sequencing

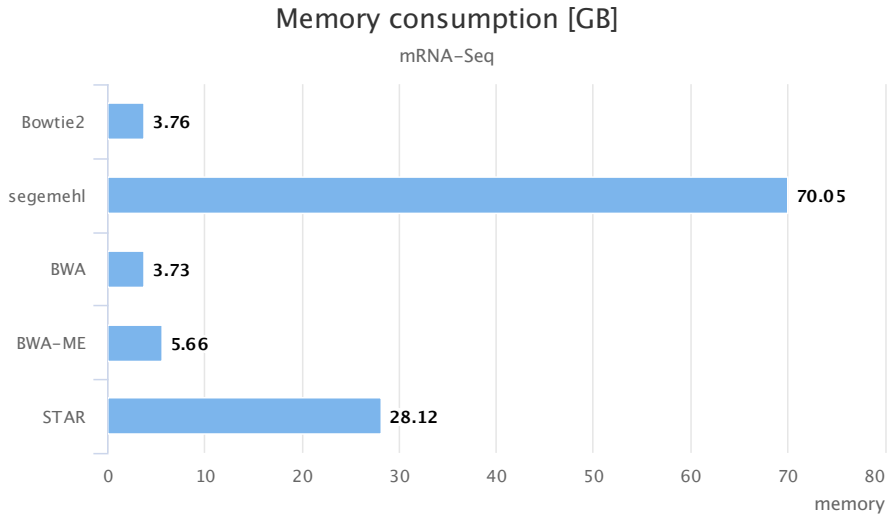
non comprehensive list

Alignment Tools









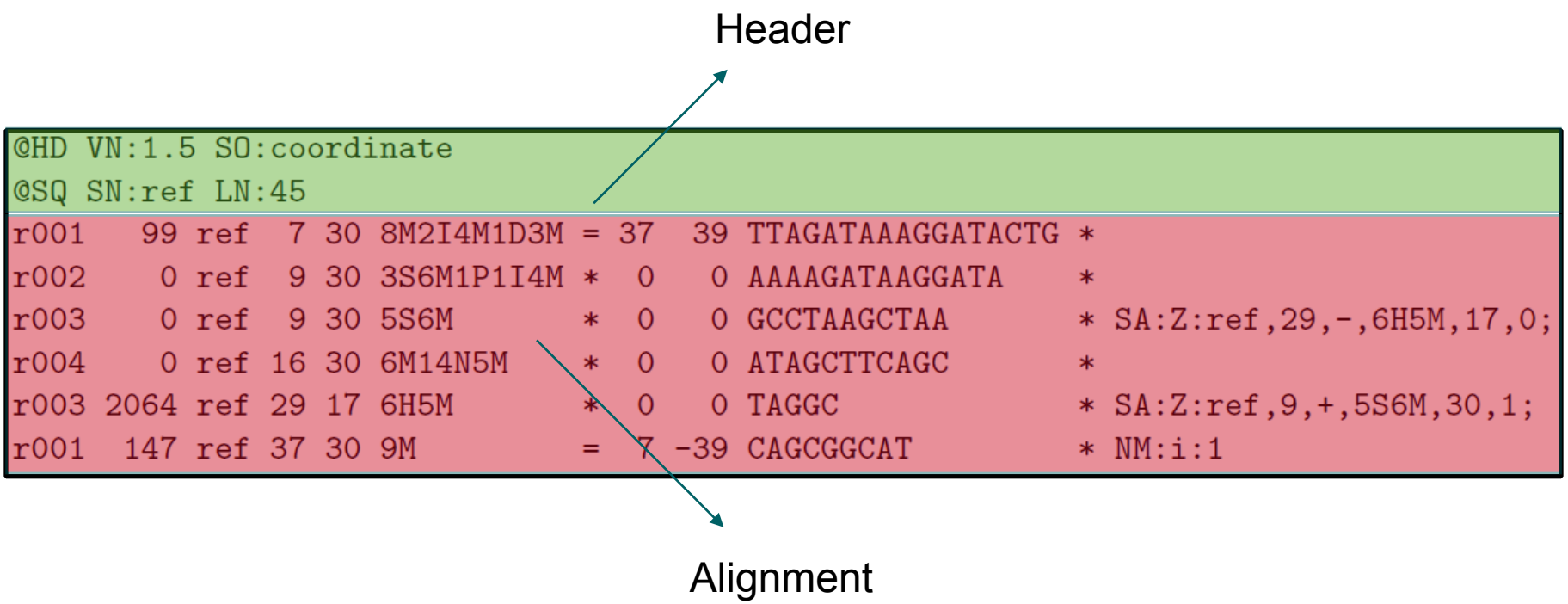
https://www.ecseq.com/support/benchmark

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

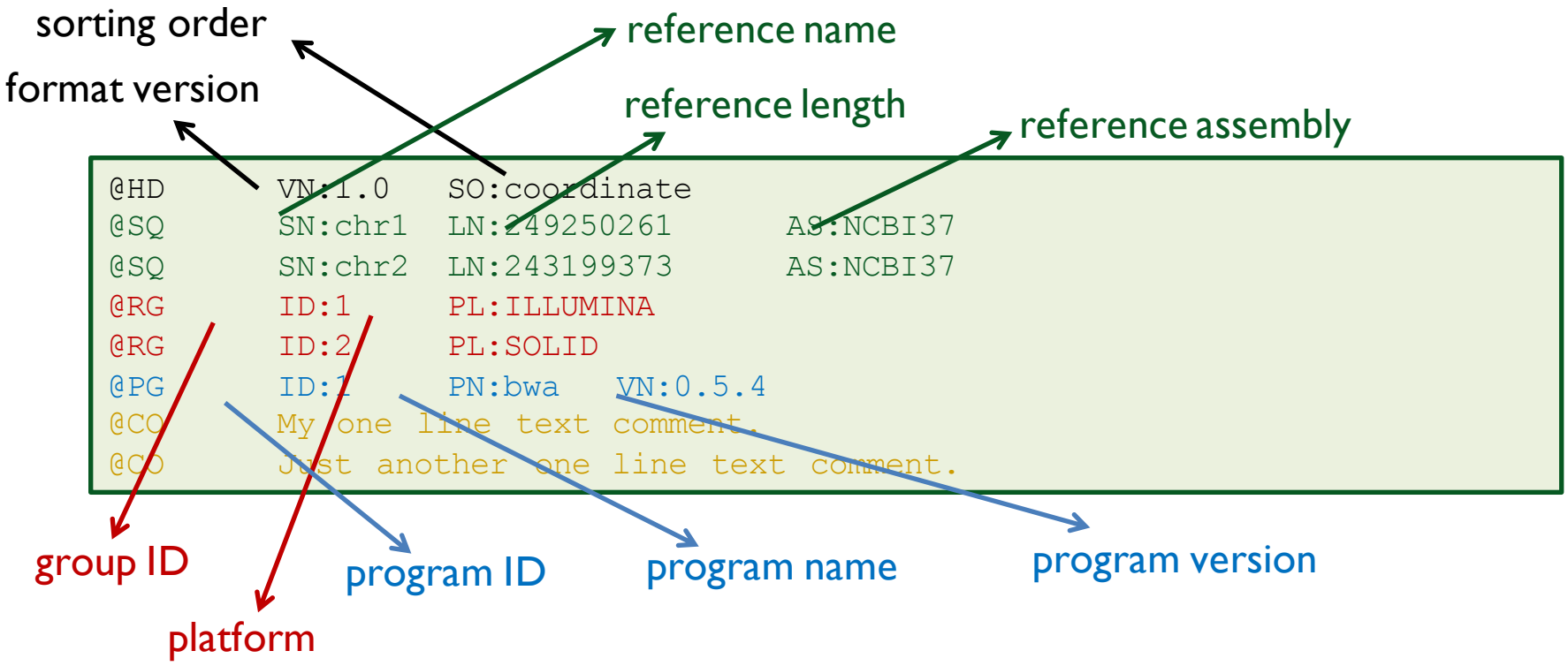
SAM Files

- Store alignment results as text-based file
- Consists of a header and an alignment section

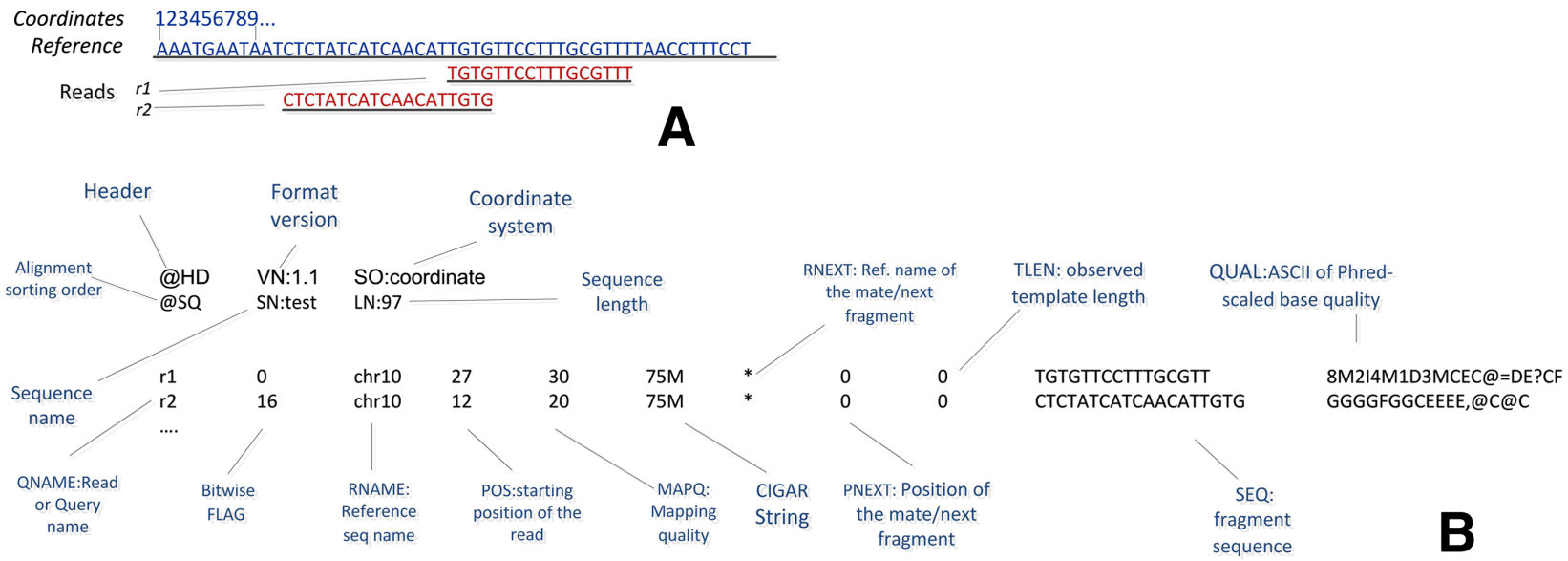


SAM Files - Header

- @HD – Header line.
- @SQ – Reference genome information.
- @RG – Read group information.
- @PG – Program (software) information.
- @CO – Commentary line.

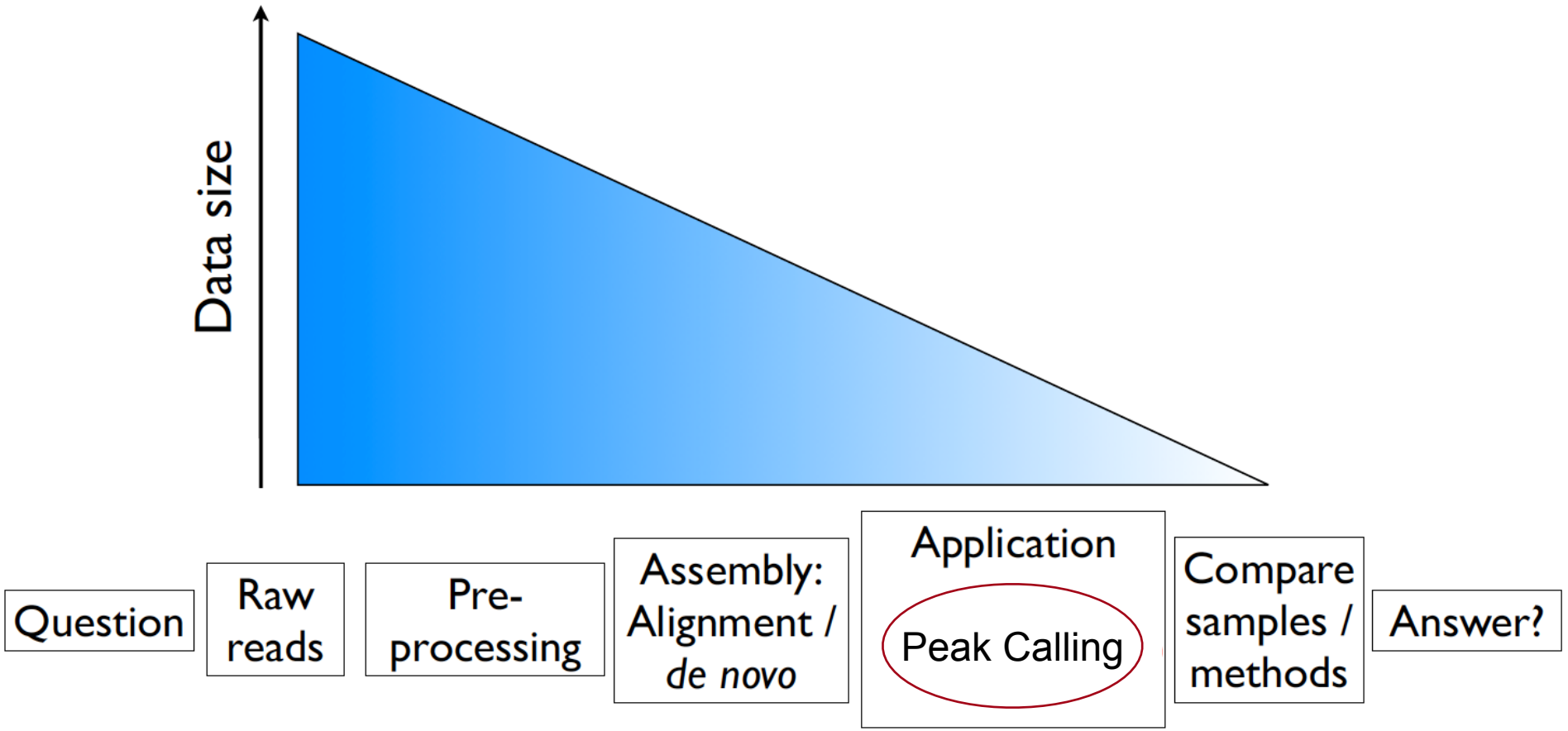


SAM Files- alignment section



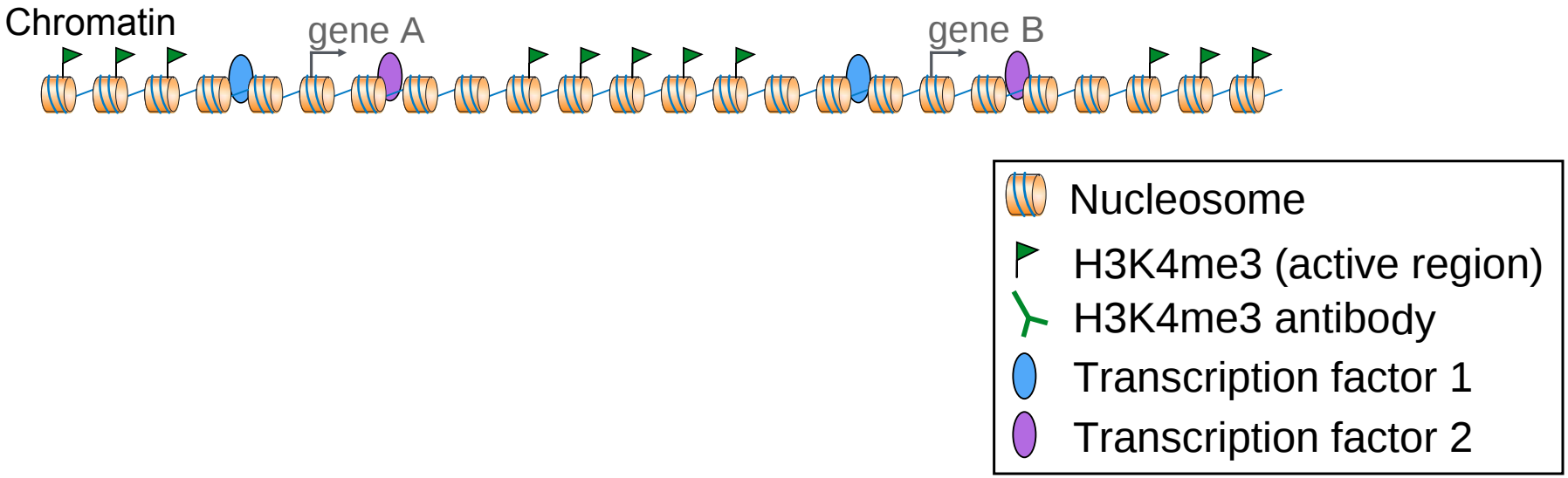
- SAM files will have large size (1-10 Gbs)
  - Usually a experiment has dozens of such files
- BAM files (zipped version of SAM) is more common and reduces the size by 30-50%. This file can be opened in genome browsers if a index file is also given.

Applications - Peak Calling

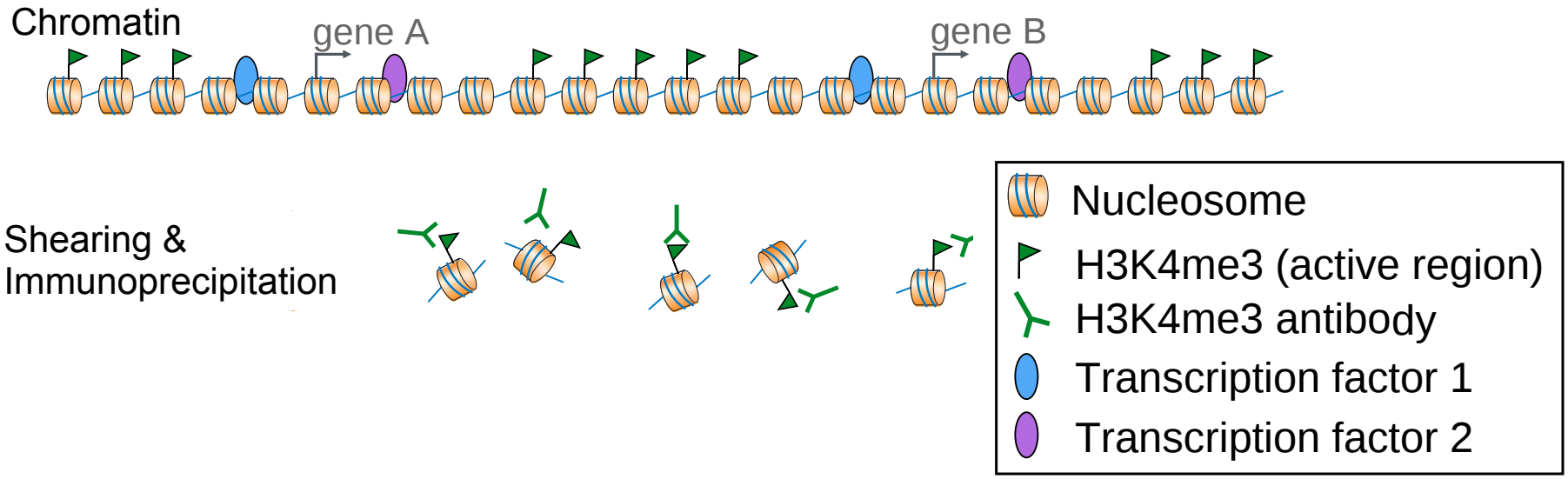


http://en.wikipedia.org/wiki/RNA-Seq

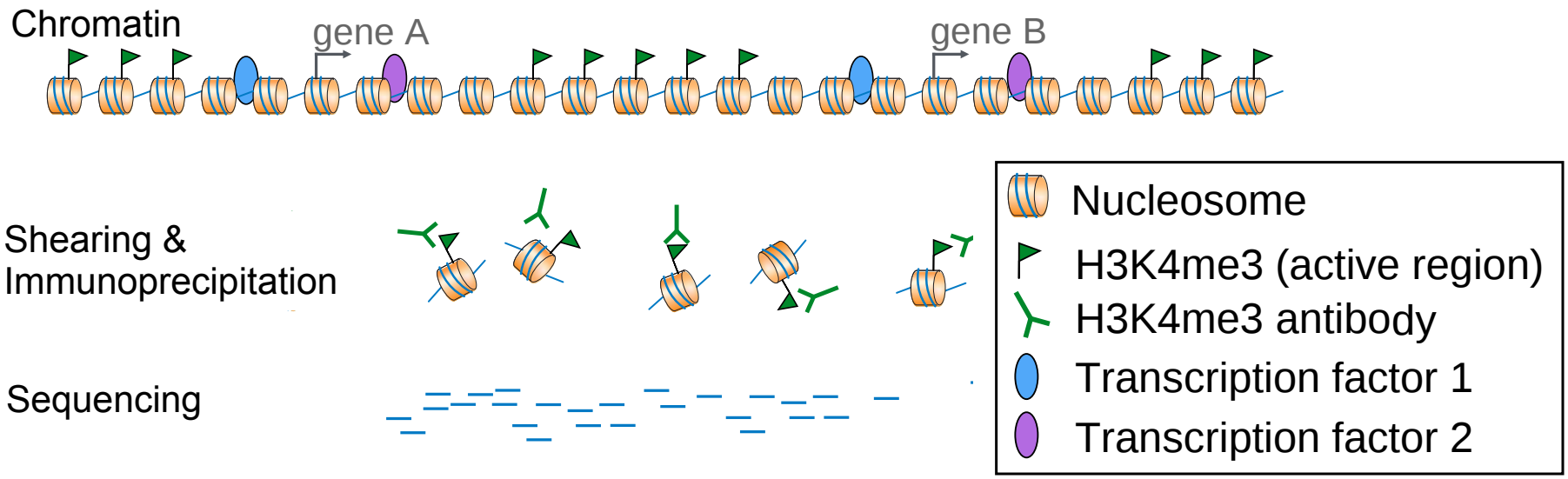
DNA - Protein interactions with ChIP-Seq



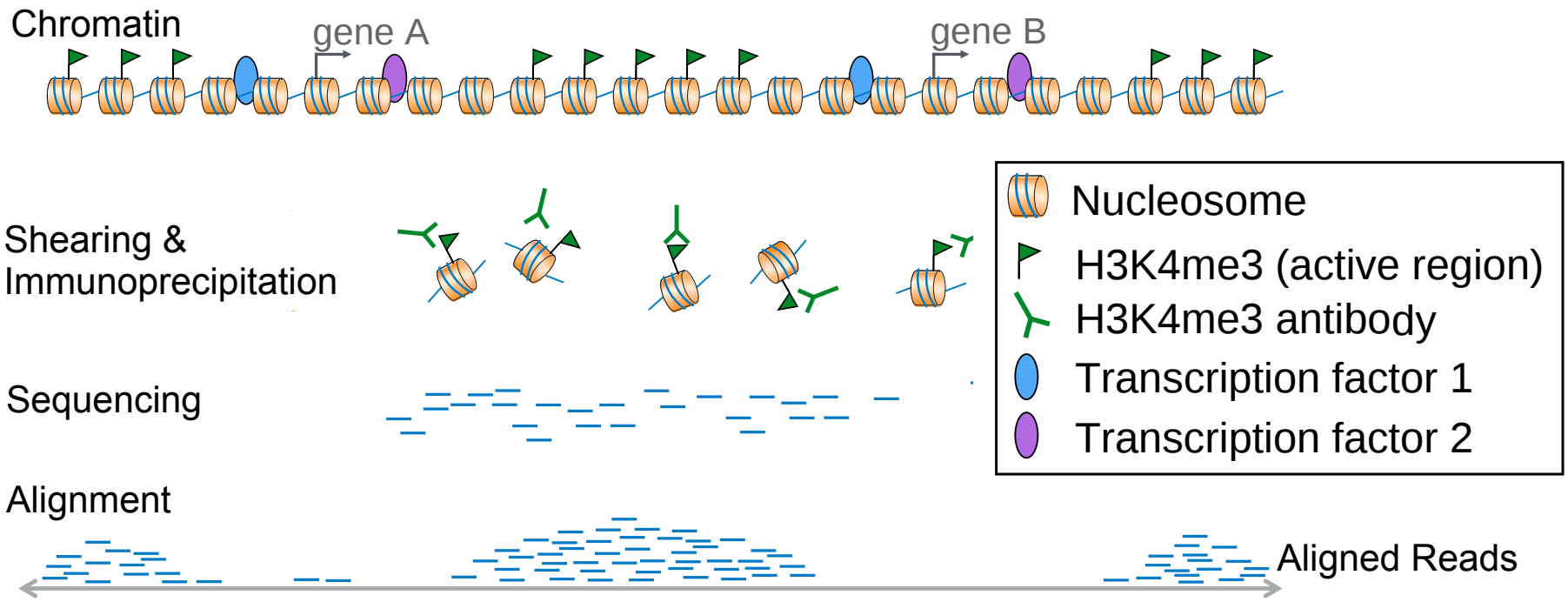
DNA - Protein interactions with ChIP-Seq



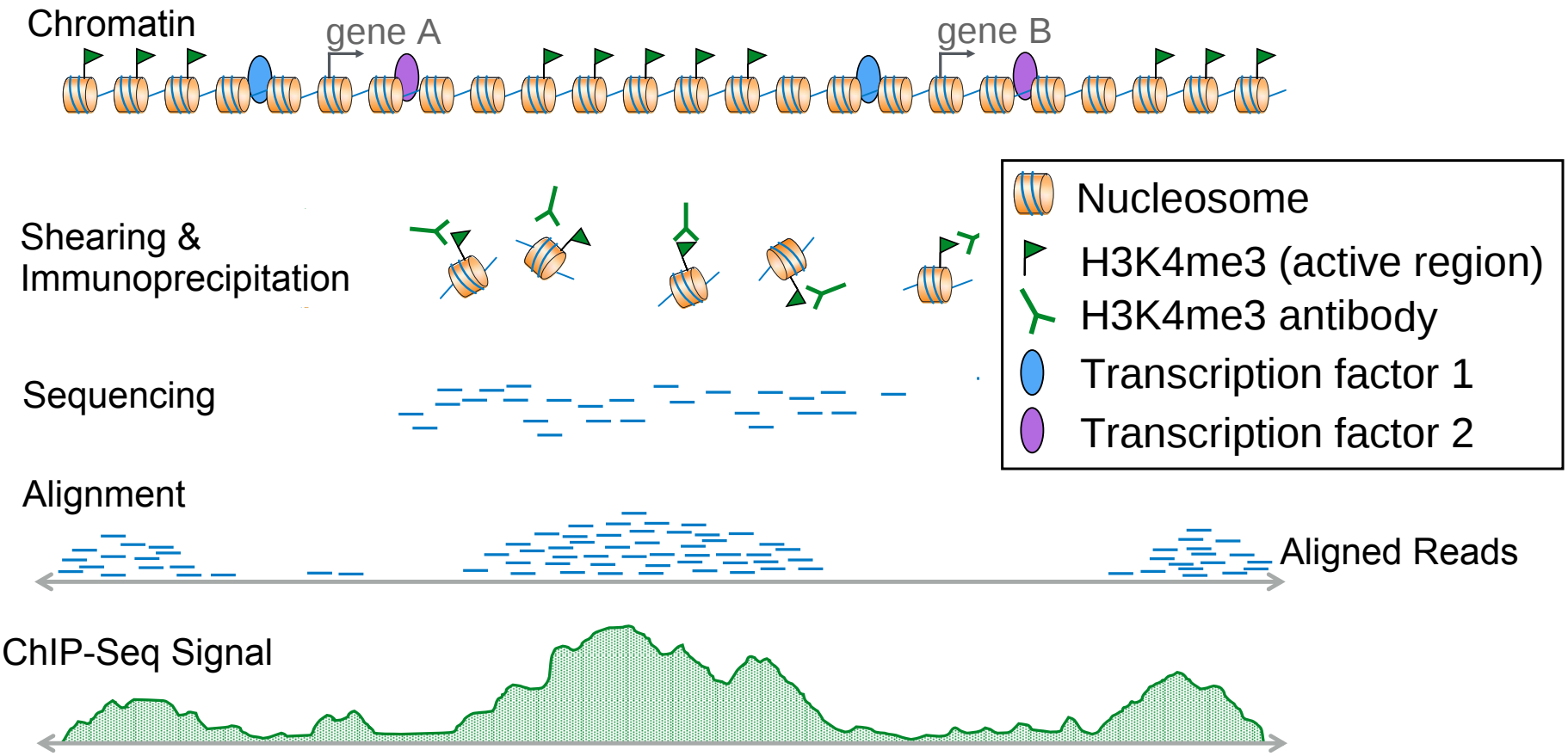
DNA - Protein interactions with ChIP-Seq



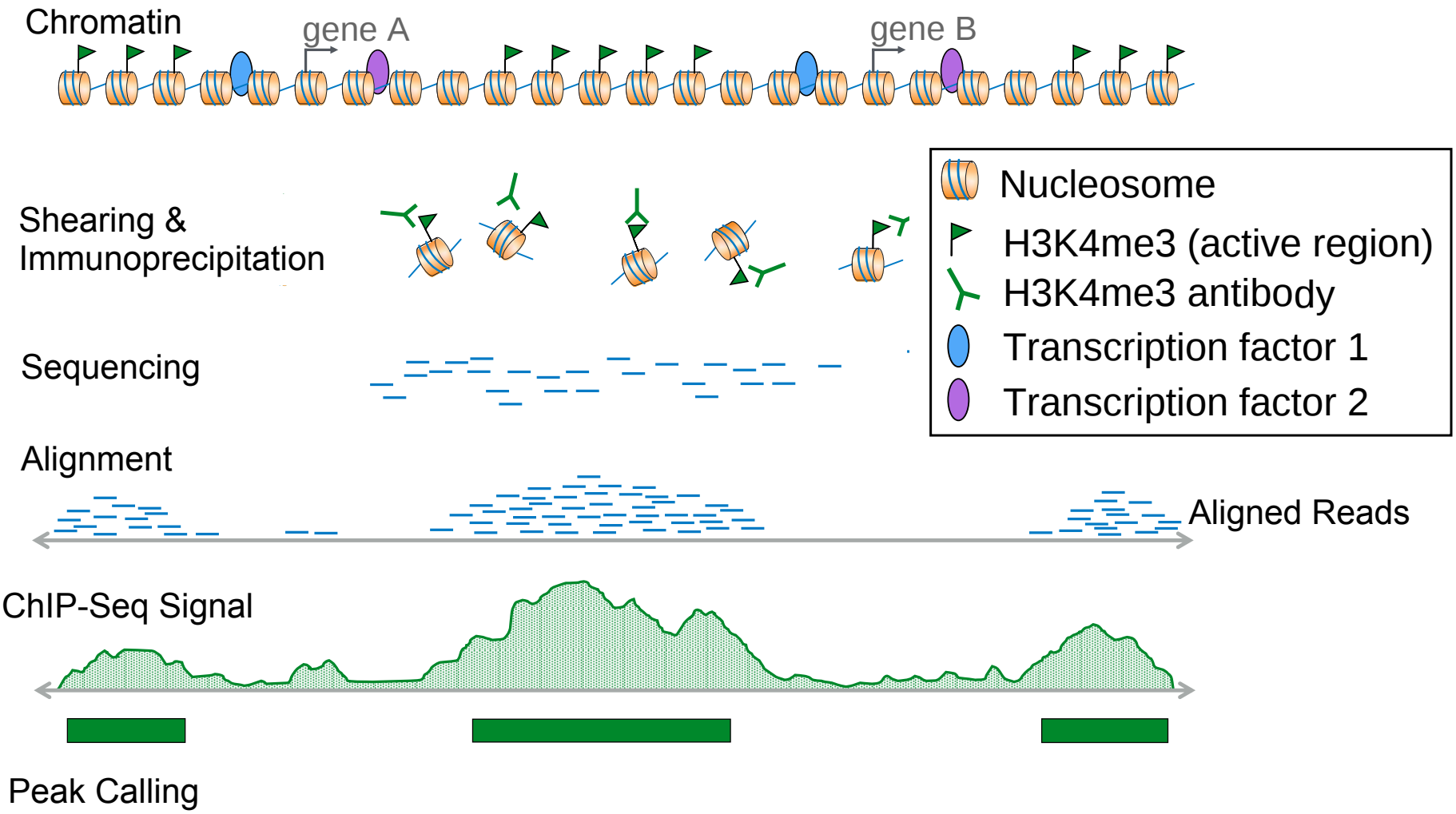
DNA - Protein interactions with ChIP-Seq



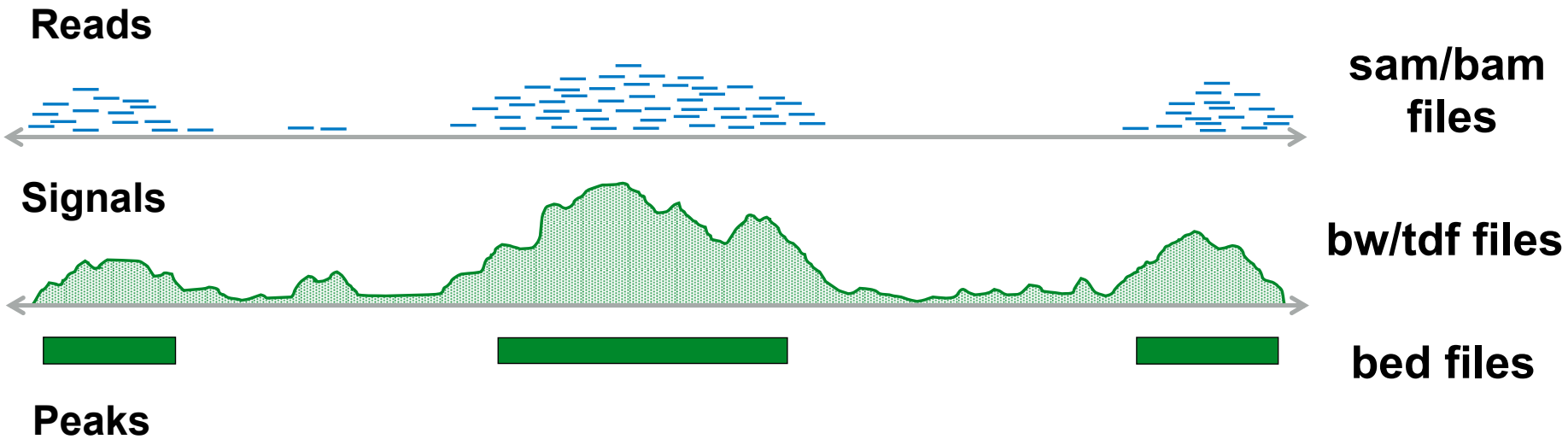
DNA - Protein interactions with ChIP-Seq



DNA - Protein interactions with ChIP-Seq



ChIP-Seq - Data Files



Peaks - Bed files

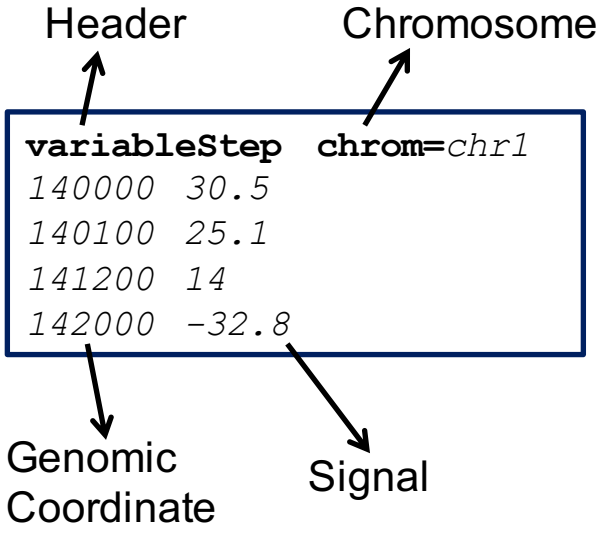
- Peaks / genomic regions are stored in bed files

Chromosome	Start	End	Name	Score	Strand
chr7	*127471196*	*127472363*	*Peak1*	*0*	+
chr7	*127472363*	*127473530*	*Peak2*	*0*	+
chr7	*127473530*	*127474697*	*Peak3*	*0*	+
chr7	*127474697*	*127475864*	*Peak4*	*0*	+
chr7	*127475864*	*127477031*	*Peak5*	*0*	-
chr7	*127477031*	*127478198*	*Peak6*	*0*	-
chr7	*127478198*	*127479365*	*Peak7*	*0*	-
chr7	*127479365*	*127480532*	*Peak8*	*0*	+
chr7	*127480532*	*127481699*	*Peak9*	*0*	-

Genomic Signals - WIG/TDF Files

- Files containing smoothed counts of reads for ChIP-seq, RNA-seq, or similar protocols.

- Example of WIG file



- In practice, we use binary version of WIG (BIGWIG) or TDF files

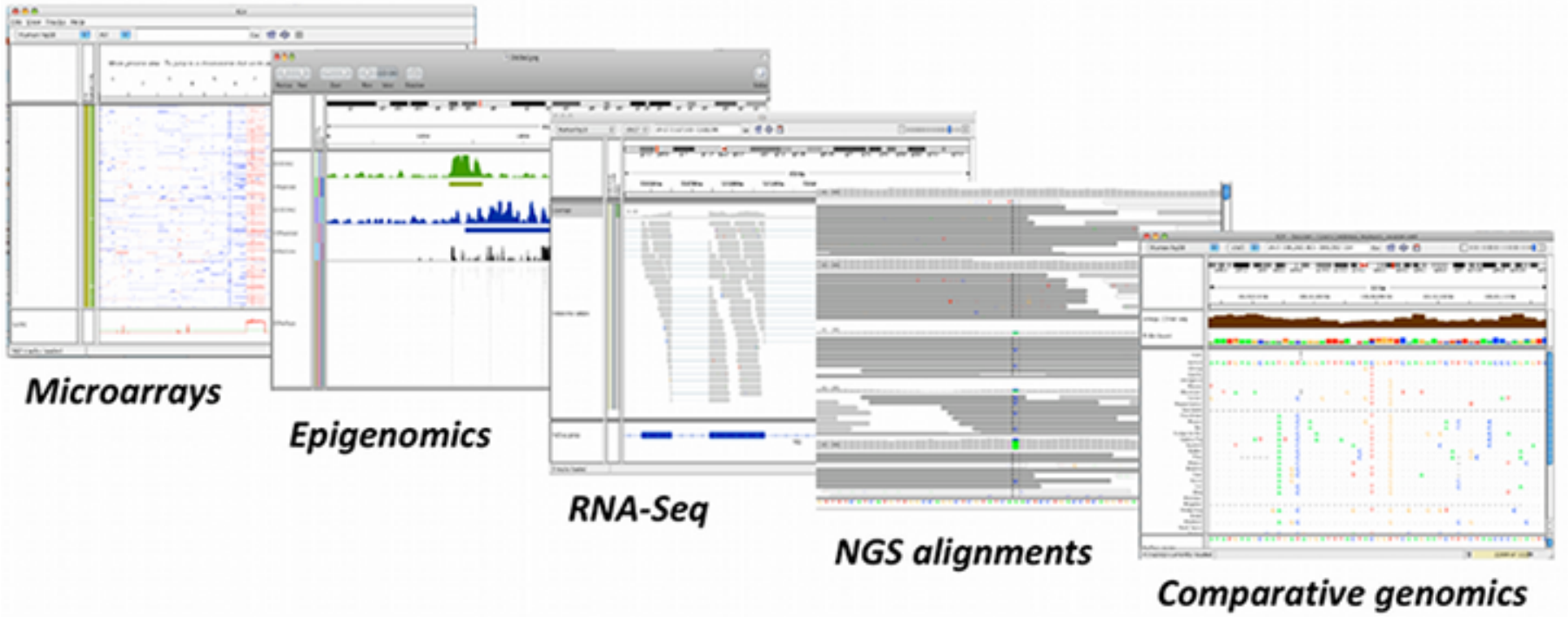
Bioinformatics Analysis in R

Next Generation Sequencing Data Visualization

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

IGV (Integrative Genome Viewer)

- **Desktop** application for the **visual interactive** exploration of **integrated** genomic datasets



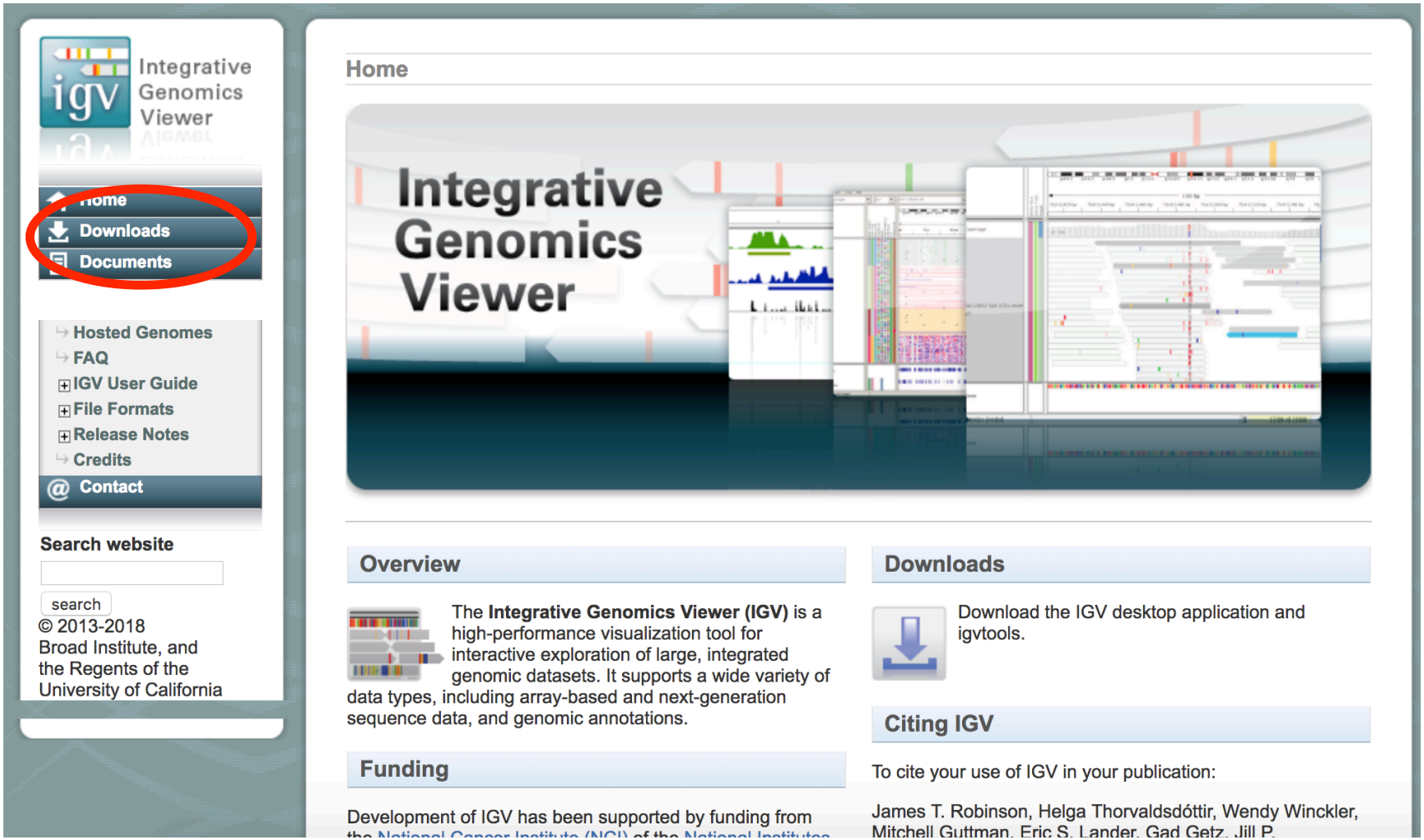
James. T, *et al (2011), Nature Biotechnology*

Advantages

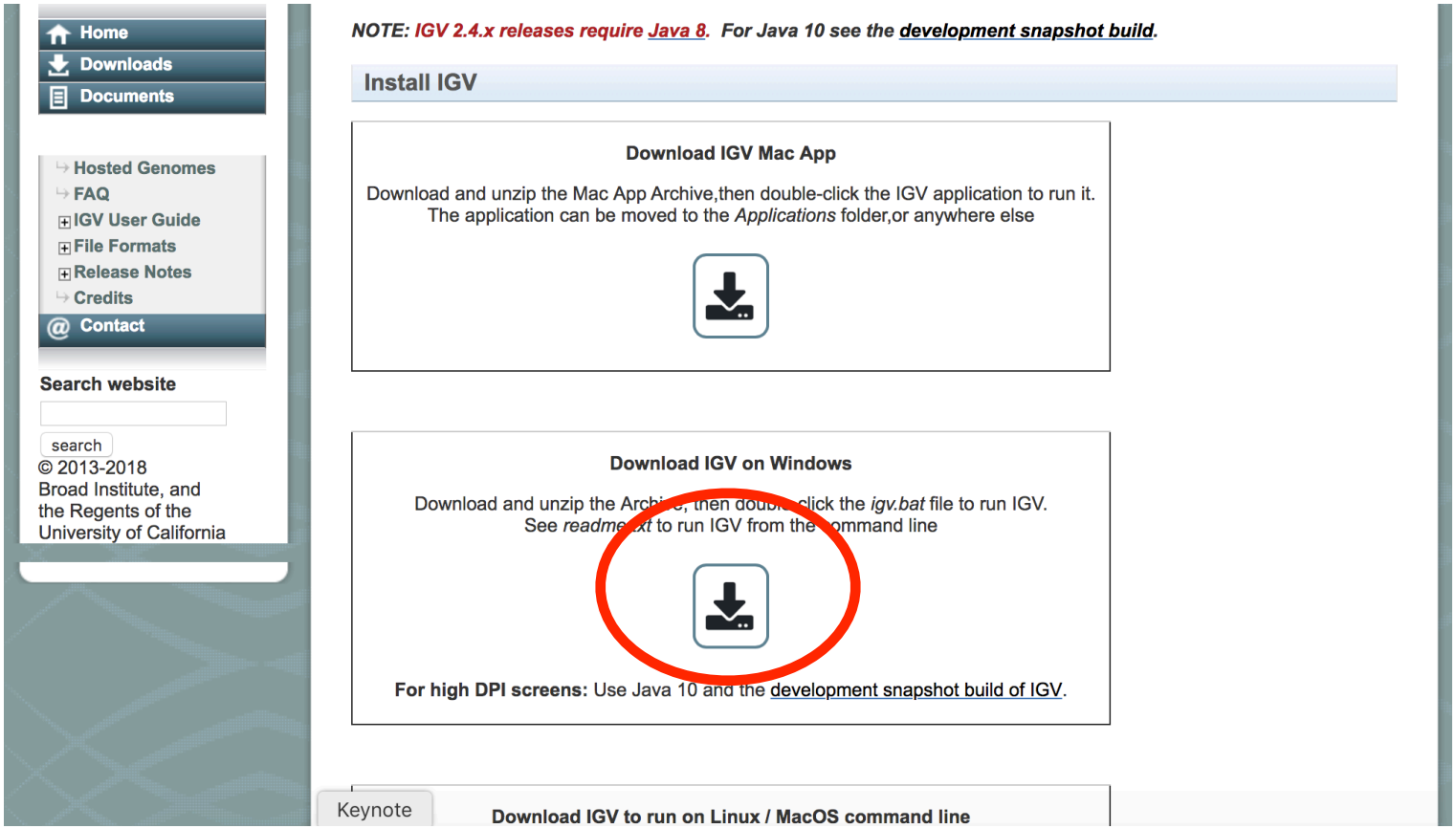
- A high-performance visualization tool
- Allows us interactively explore large, integrated dataset
- Supports a wide variety of data types, including microarray and next-generation sequencing data
- **FREE**

Launch IGV

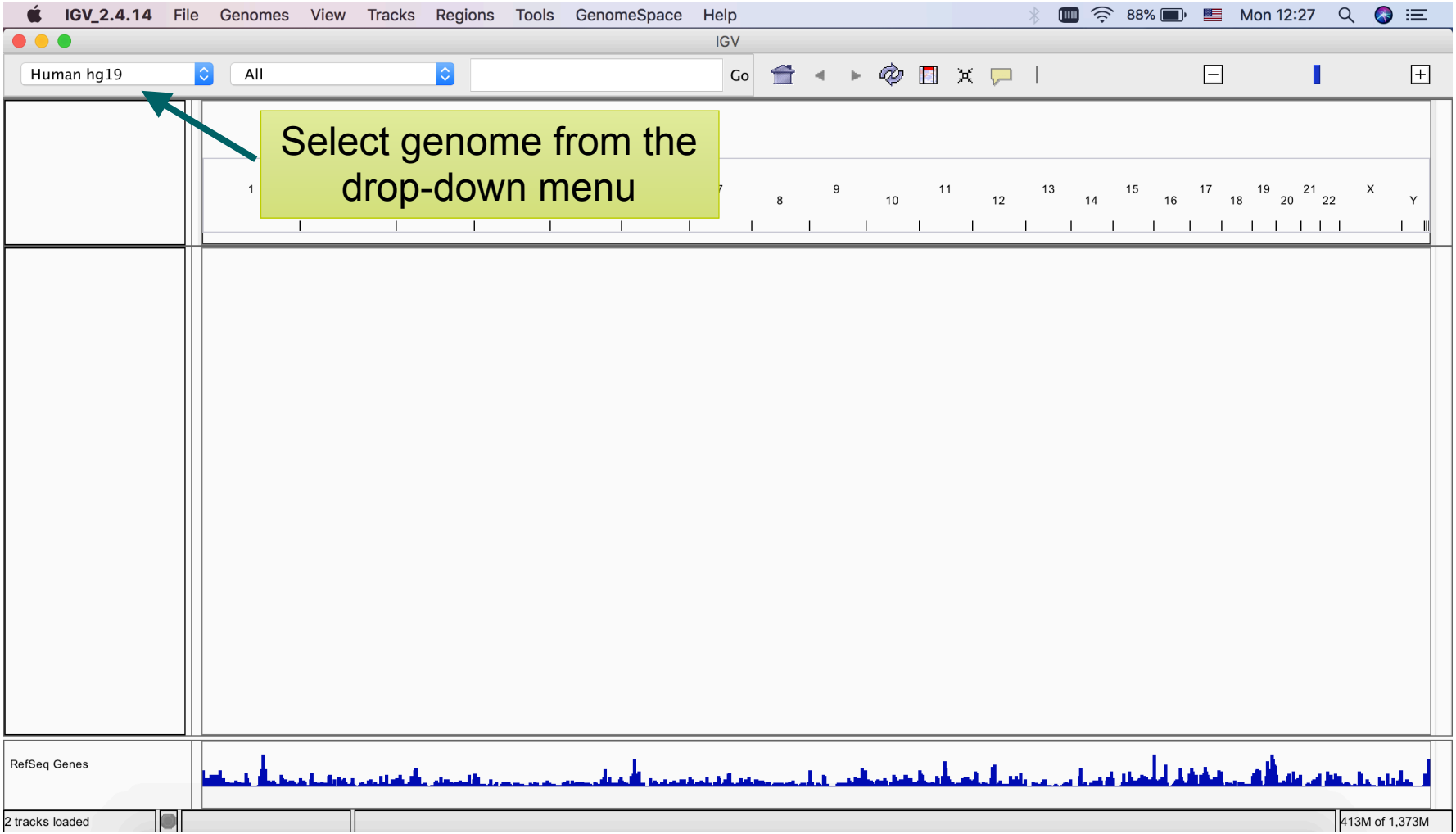
http://software.broadinstitute.org/software/igv/home



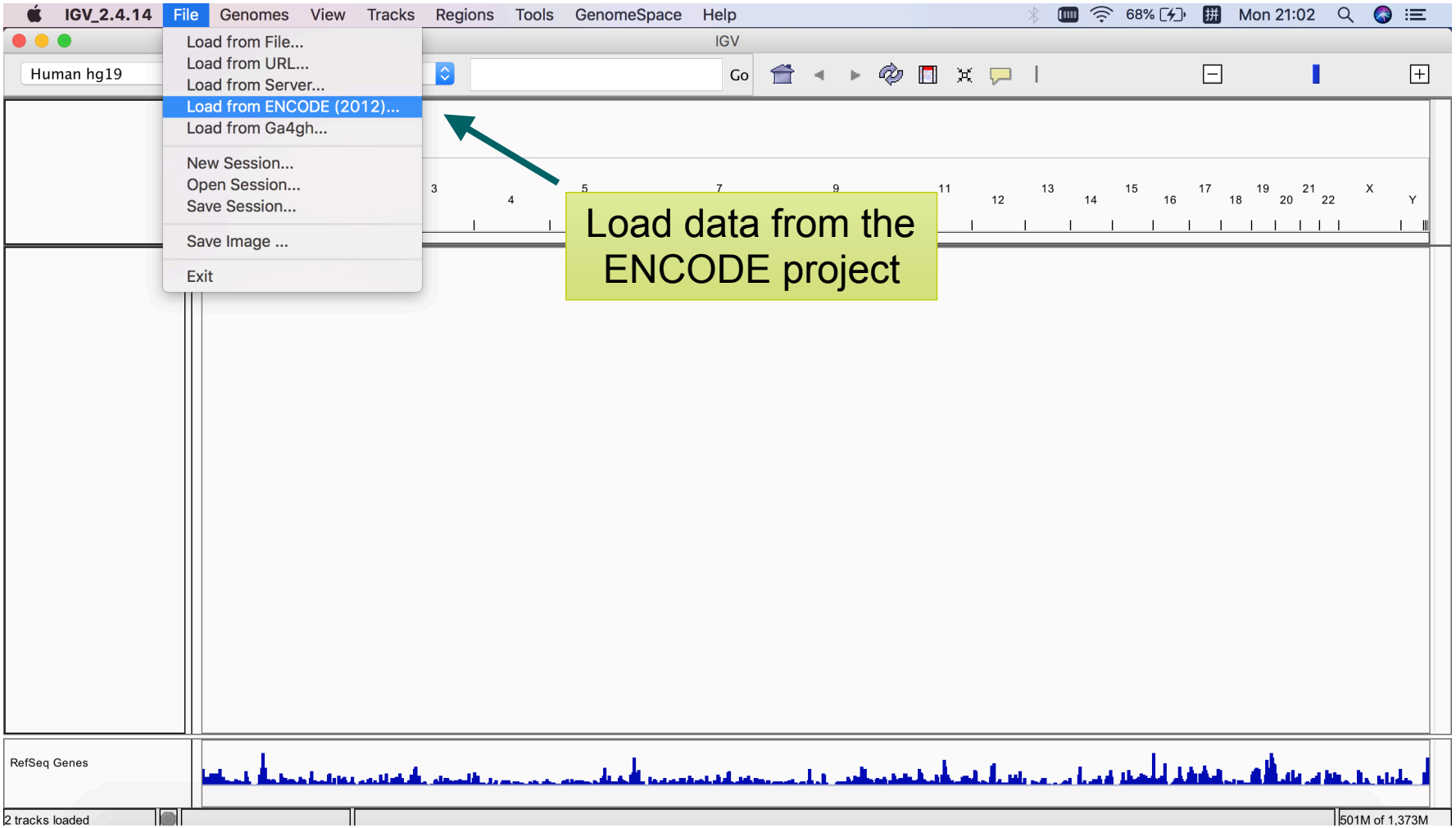
Launch IGV


Home
Downloads
Documents
Hosted Genomes
FAQ
IGV User Guide
File Formats
Release Notes
Credits
Contact
Search website
search
© 2013-2018
Broad Institute, and
the Regents of the
University of California
NOTE: IGV 2.4.x releases require Java 8. For Java 10 see the development snapshot build.
Install IGV
Download IGV Mac App
Download and unzip the Mac App Archive,then double-click the IGV application to run it.
The application can be moved to the Applications folder,or anywhere else
Download IGV on Windows
Download and unzip the Archive, then double-click the igv.bat file to run IGV.
See readme.txt to run IGV from the command line
For high DPI screens: Use Java 10 and the development snapshot build of IGV.
Keynote
Download IGV to run on Linux / MacOS command line

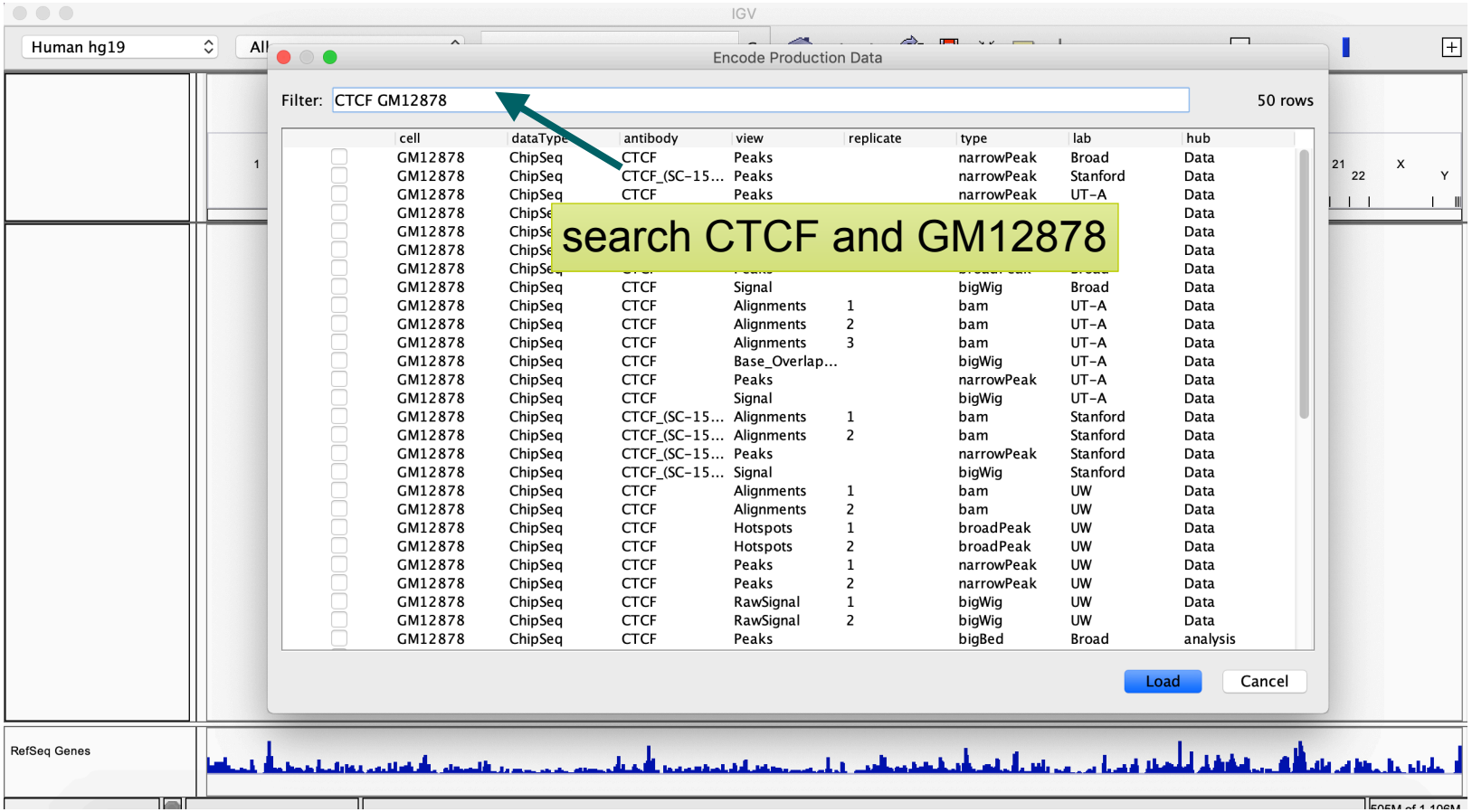
Launch IGV


IGV_2.4.14 File Genomes View Tracks Regions Tools GenomeSpace Help
88% Mon 12:27
IGV
Human hg19
All
Go
Select genome from the drop-down menu
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
RefSeq Genes
2 tracks loaded
413M of 1,373M

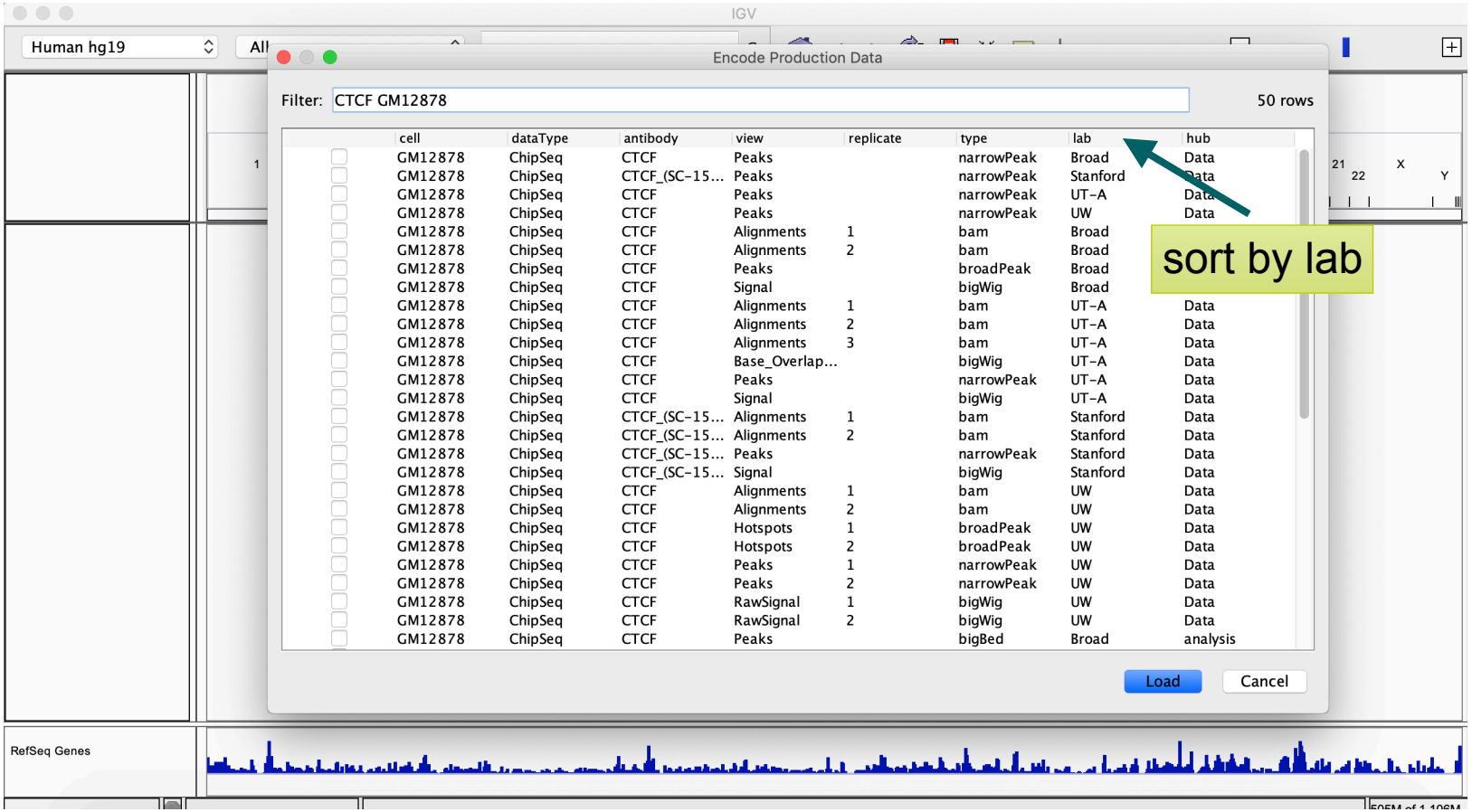
Launch IGV


IGV_2.4.14
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Mon 21:02
IGV
Human hg19
Go
Load from File...
Load from URL...
Load from Server...
Load from ENCODE (2012)...
Load from Ga4gh...
New Session...
Open Session...
Save Session...
Save Image ...
Exit
Load data from the ENCODE project
RefSeq Genes
2 tracks loaded
501M of 1,373M

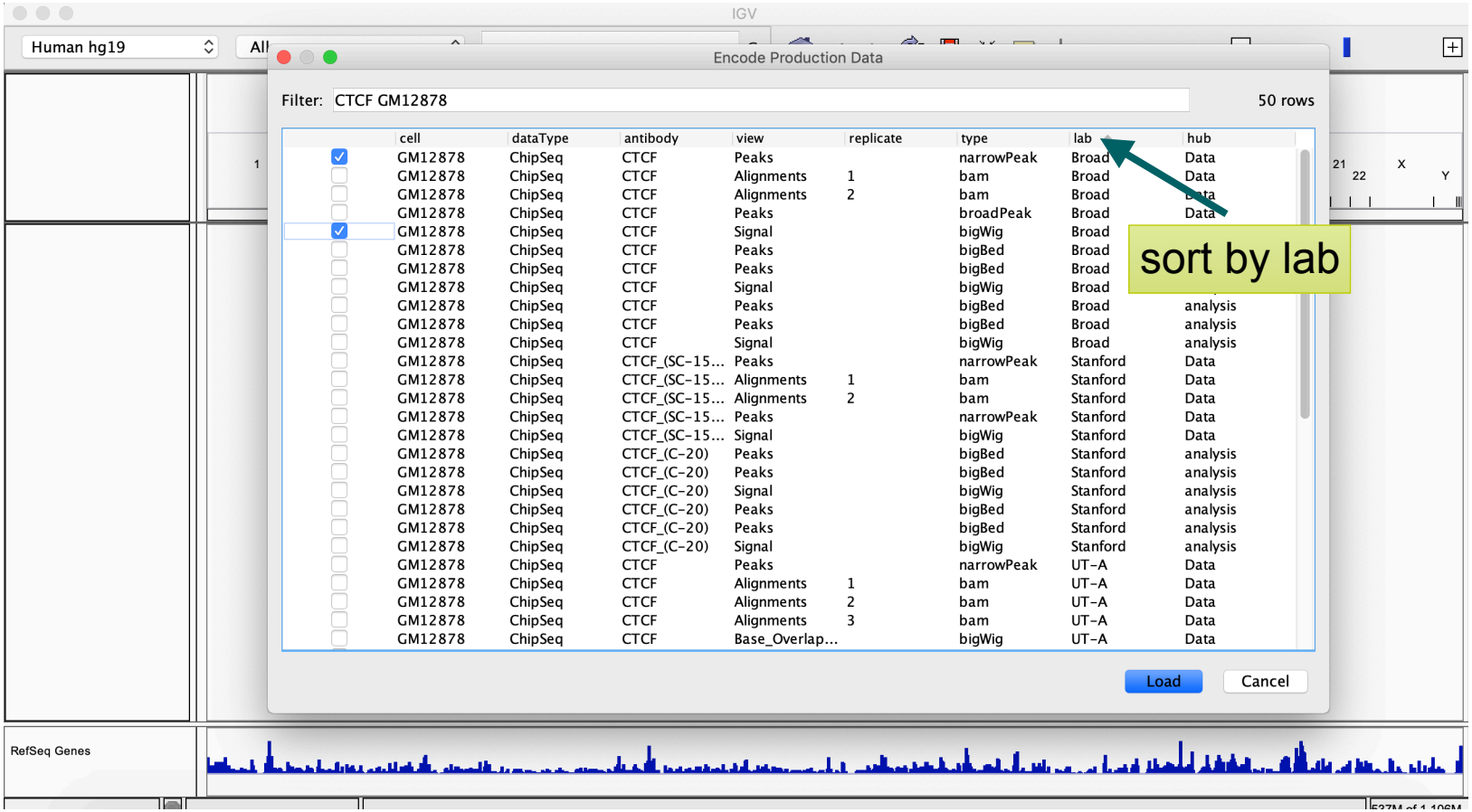
Launch IGV



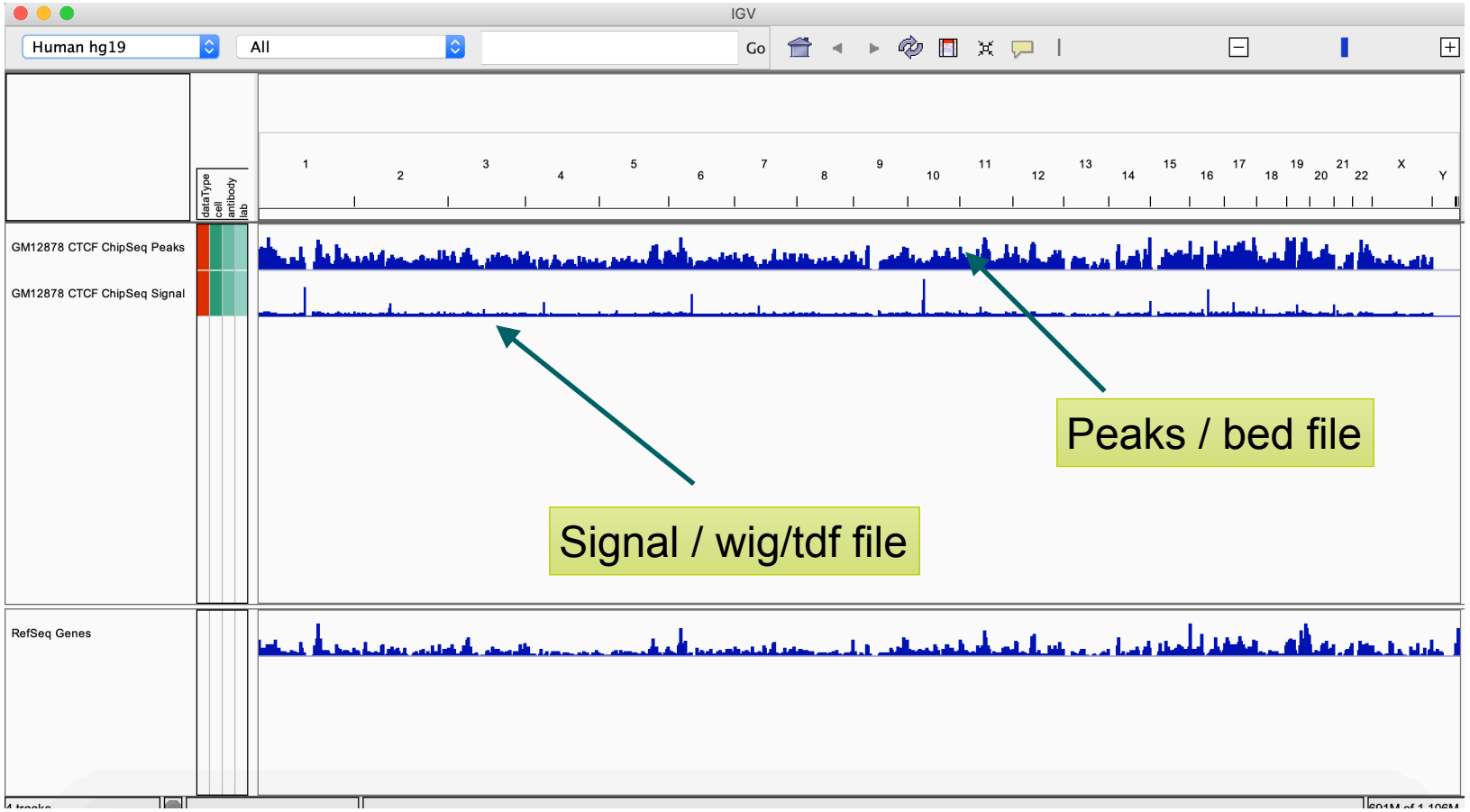
Launch IGV


IGV
Human hg19
Encode Production Data
Filter: CTCF GM12878
50 rows
cell dataType antibody view replicate type lab hub
GM12878 ChipSeq CTCF Peaks narrowPeak Broad Data
GM12878 ChipSeq CTCF_(SC-15... Peaks narrowPeak Stanford Data
GM12878 ChipSeq CTCF Peaks narrowPeak UT-A Data
GM12878 ChipSeq CTCF Peaks narrowPeak UW Data
GM12878 ChipSeq CTCF Alignments 1 bam Broad
GM12878 ChipSeq CTCF Alignments 2 bam Broad
GM12878 ChipSeq CTCF Peaks broadPeak Broad
GM12878 ChipSeq CTCF Signal bigWig Broad
GM12878 ChipSeq CTCF Alignments 1 bam UT-A Data
GM12878 ChipSeq CTCF Alignments 2 bam UT-A Data
GM12878 ChipSeq CTCF Alignments 3 bam UT-A Data
GM12878 ChipSeq CTCF Base_Overlap... bigWig UT-A Data
GM12878 ChipSeq CTCF Peaks narrowPeak UT-A Data
GM12878 ChipSeq CTCF Signal bigWig UT-A Data
GM12878 ChipSeq CTCF_(SC-15... Alignments 1 bam Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Alignments 2 bam Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Peaks narrowPeak Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Signal bigWig Stanford Data
GM12878 ChipSeq CTCF Alignments 1 bam UW Data
GM12878 ChipSeq CTCF Alignments 2 bam UW Data
GM12878 ChipSeq CTCF Hotspots 1 broadPeak UW Data
GM12878 ChipSeq CTCF Hotspots 2 broadPeak UW Data
GM12878 ChipSeq CTCF Peaks 1 narrowPeak UW Data
GM12878 ChipSeq CTCF Peaks 2 narrowPeak UW Data
GM12878 ChipSeq CTCF RawSignal 1 bigWig UW Data
GM12878 ChipSeq CTCF RawSignal 2 bigWig UW Data
GM12878 ChipSeq CTCF Peaks bigBed Broad analysis
sort by lab
Load
Cancel
RefSeq Genes

Launch IGV


IGV
Human hg19
Encode Production Data
Filter: CTCF GM12878
50 rows
cell dataType antibody view replicate type lab hub
GM12878 ChipSeq CTCF Peaks narrowPeak Broad Data
GM12878 ChipSeq CTCF Alignments 1 bam Broad Data
GM12878 ChipSeq CTCF Alignments 2 bam Broad Data
GM12878 ChipSeq CTCF Peaks broadPeak Broad Data
GM12878 ChipSeq CTCF Signal bigWig Broad Data
GM12878 ChipSeq CTCF Peaks bigBed Broad Data
GM12878 ChipSeq CTCF Peaks bigBed Broad Data
GM12878 ChipSeq CTCF Signal bigWig Broad Data
GM12878 ChipSeq CTCF Peaks bigBed Broad analysis
GM12878 ChipSeq CTCF Peaks bigBed Broad analysis
GM12878 ChipSeq CTCF Signal bigWig Broad analysis
GM12878 ChipSeq CTCF_(SC-15... Peaks narrowPeak Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Alignments 1 bam Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Alignments 2 bam Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Peaks narrowPeak Stanford Data
GM12878 ChipSeq CTCF_(SC-15... Signal bigWig Stanford Data
GM12878 ChipSeq CTCF_(C-20) Peaks bigBed Stanford analysis
GM12878 ChipSeq CTCF_(C-20) Peaks bigBed Stanford analysis
GM12878 ChipSeq CTCF_(C-20) Signal bigWig Stanford analysis
GM12878 ChipSeq CTCF_(C-20) Peaks bigBed Stanford analysis
GM12878 ChipSeq CTCF_(C-20) Peaks bigBed Stanford analysis
GM12878 ChipSeq CTCF_(C-20) Signal bigWig Stanford analysis
GM12878 ChipSeq CTCF Peaks narrowPeak UT-A Data
GM12878 ChipSeq CTCF Alignments 1 bam UT-A Data
GM12878 ChipSeq CTCF Alignments 2 bam UT-A Data
GM12878 ChipSeq CTCF Alignments 3 bam UT-A Data
GM12878 ChipSeq CTCF Base_Overlap... bigWig UT-A Data
sort by lab
Load
Cancel
RefSeq Genes

Launch IGV - Example on CTCF ChIP-seq



File Formats

- The file format defines the track type
- The track type determines the display options
- IGV supports many file formats
  - BAM
  - BED
  - BedGraph
  - bigBed
  - bigWig
  - Birdsuite Files
  - broadPeaks
  - CBS
  - CN
  - Cufflinks Files
  - FASTA
  - GCT
  - genePred
  - GFF
  - GISTIC
  - Goby
  - GWAS
  - IGV
  - LOH
  - MAF
  - MUT
  - narrowPeaks
  - PSL
  - RES
  - SAM
  - SEG
  - SNP
  - TAB
  - TDF
  - TrackLine
  - TypeLine
  - VCF
  - WIG

http://software.broadinstitute.org/software/igv/FileFormats

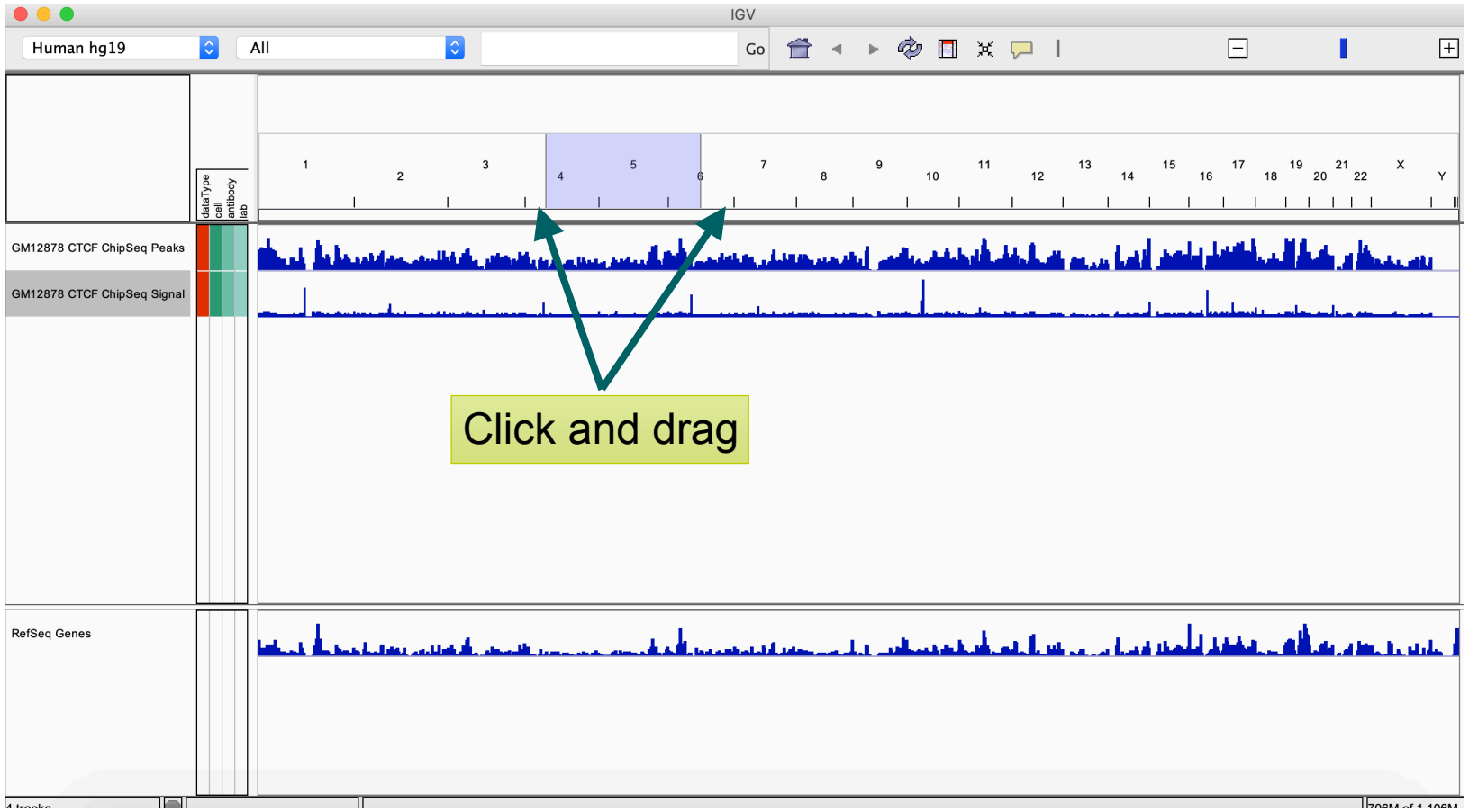
File Formats

Note: for large files use indexed formats

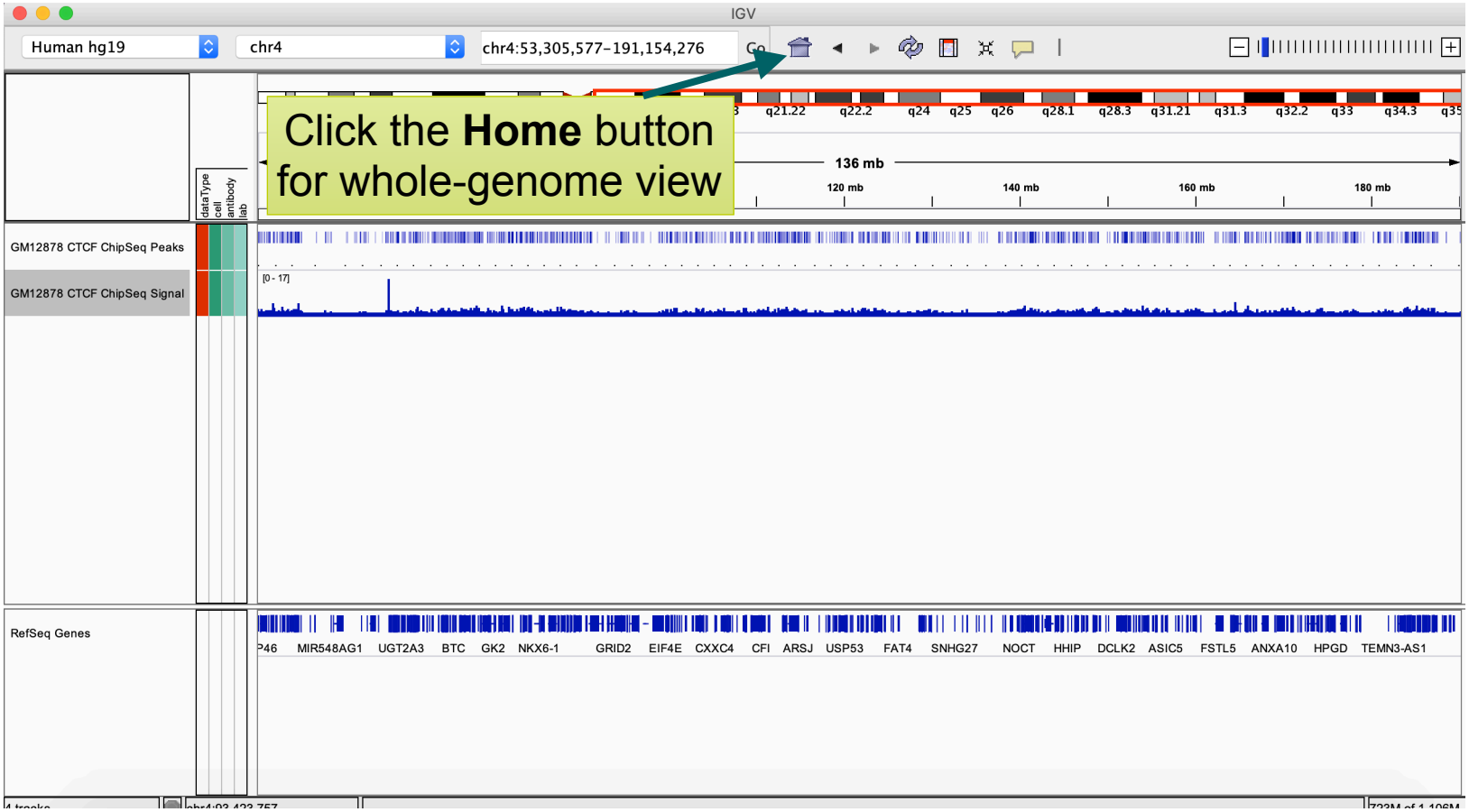
Hands on

- Lunch IGV on your computer
- Choose human genome hg19
- Load data from ENCODE project
  - ChIP-seq of factor CTCF for GM12878 cell type
  - Peaks and signal

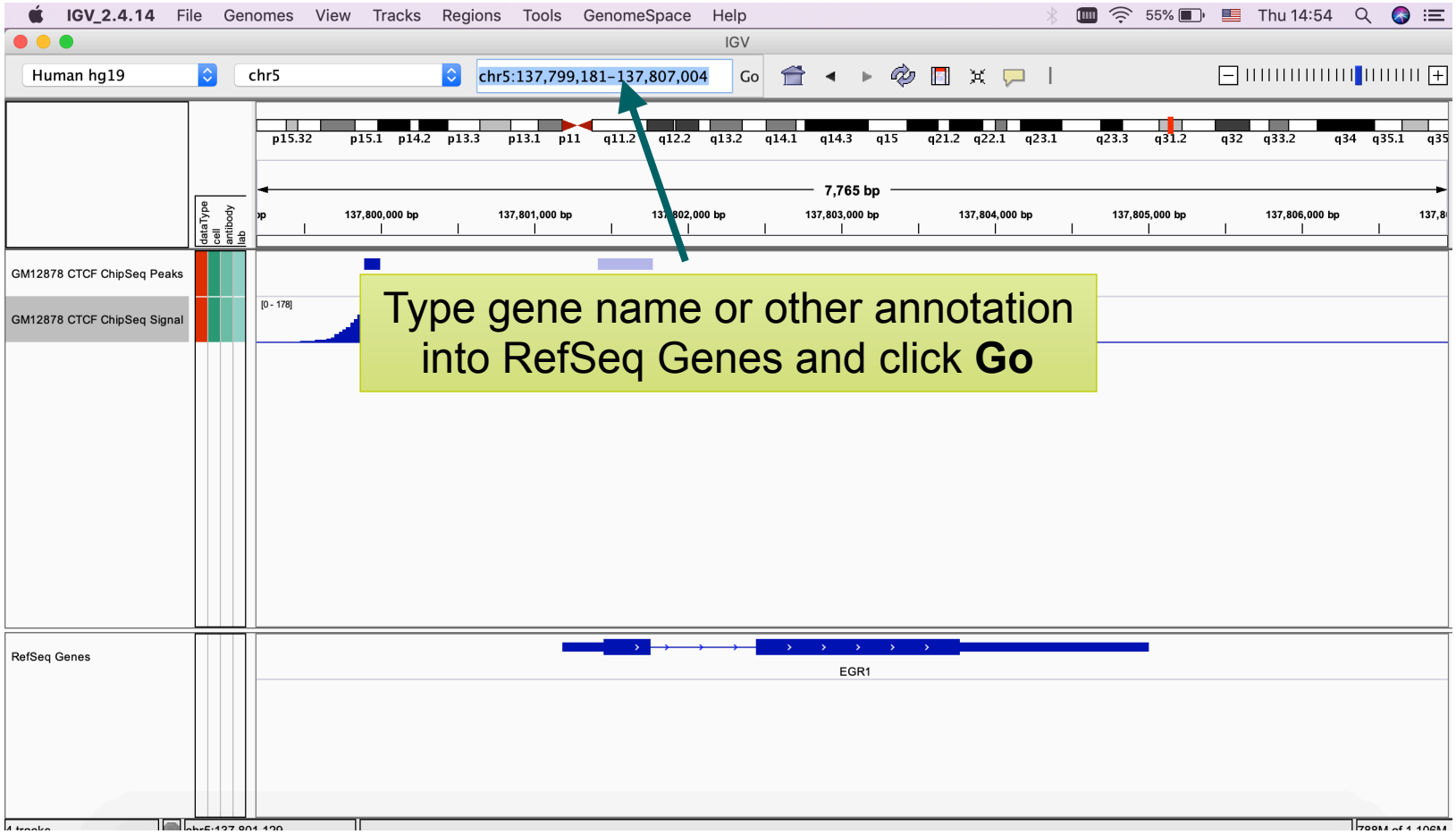
Navigate


IGV
Human hg19
All
Go
dataType
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
RefSeq Genes
Click and drag

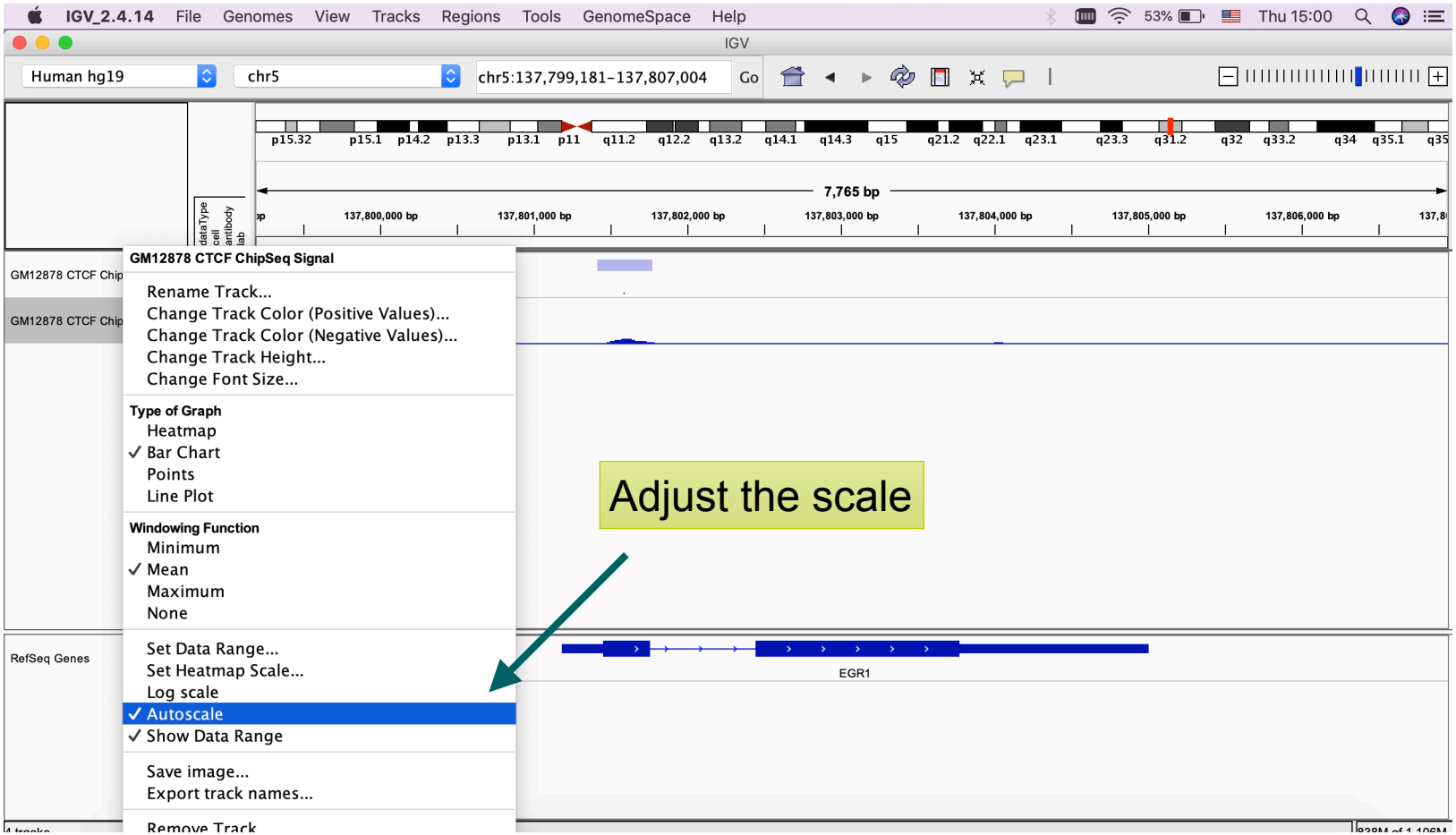
Navigate


IGV
Human hg19
chr4
chr4:53,305,577-191,154,276
Go
Click the **Home** button for whole-genome view
136 mb
120 mb
140 mb
160 mb
180 mb
data Type
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 17]
RefSeq Genes
MIR548AG1
UGT2A3
BTC
GK2
NKX6-1
GRID2
EIF4E
CXXC4
CFI
ARSJ
USP53
FAT4
SNHG27
NOCT
HHIP
DCLK2
ASIC5
FSTL5
ANXA10
HPGD
TEMN3-AS1

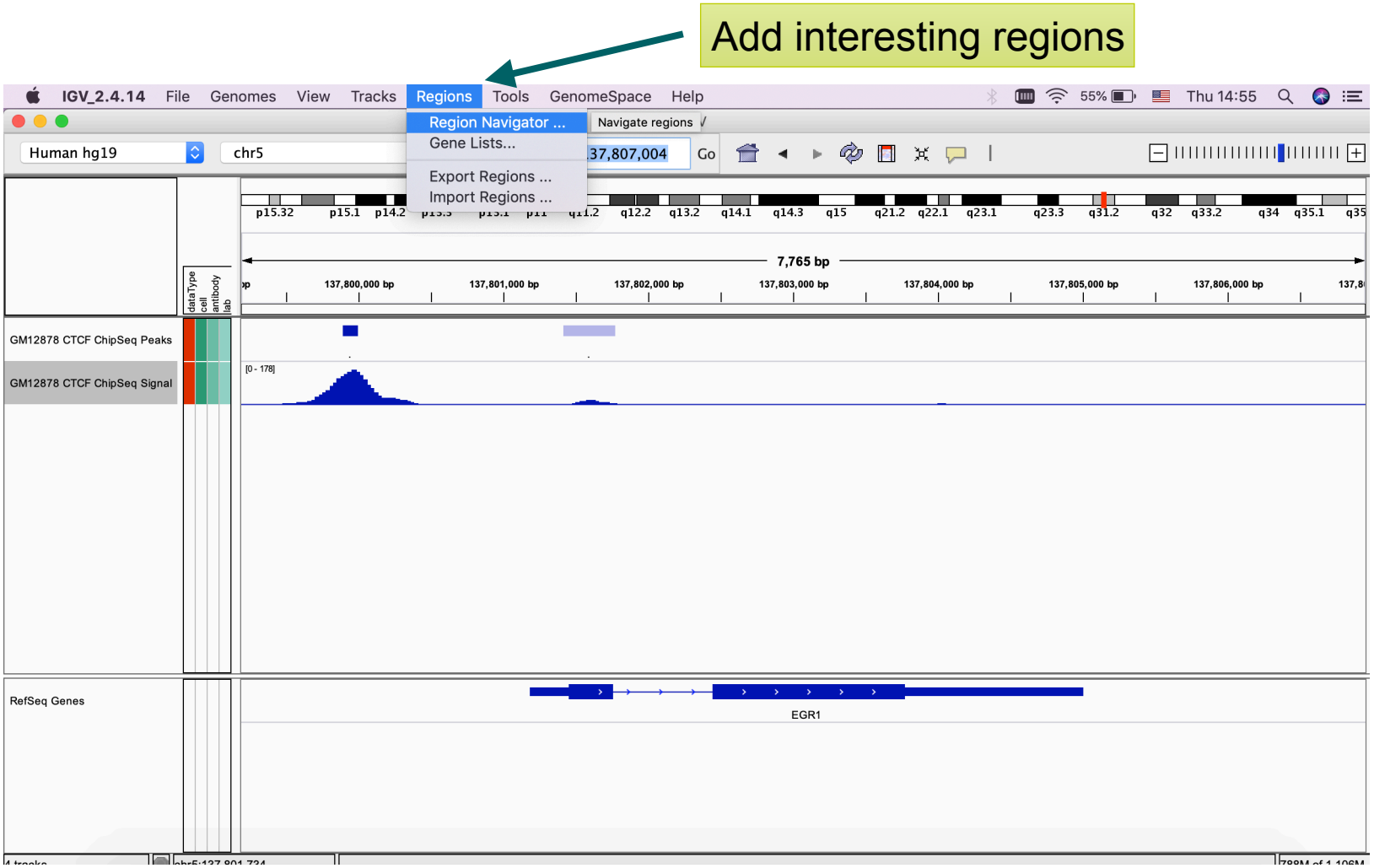
Navigate


IGV_2.4.14 File Genomes View Tracks Regions Tools GenomeSpace Help
IGV
Human hg19
chr5
chr5:137,799,181-137,807,004
Go
7,765 bp
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
RefSeq Genes
EGR1
Type gene name or other annotation into RefSeq Genes and click **Go**

Navigate


IGV_2.4.14 File Genomes View Tracks Regions Tools GenomeSpace Help
IGV
Human hg19
chr5
chr5:137,799,181-137,807,004
Go
7,765 bp
GM12878 CTCF ChipSeq Signal
Rename Track...
Change Track Color (Positive Values)...
Change Track Color (Negative Values)...
Change Track Height...
Change Font Size...
Type of Graph
Heatmap
Bar Chart
Points
Line Plot
Windowing Function
Minimum
Mean
Maximum
None
Set Data Range...
Set Heatmap Scale...
Log scale
Autoscale
Show Data Range
Save image...
Export track names...
Remove Track
Adjust the scale
RefSeq Genes
EGR1

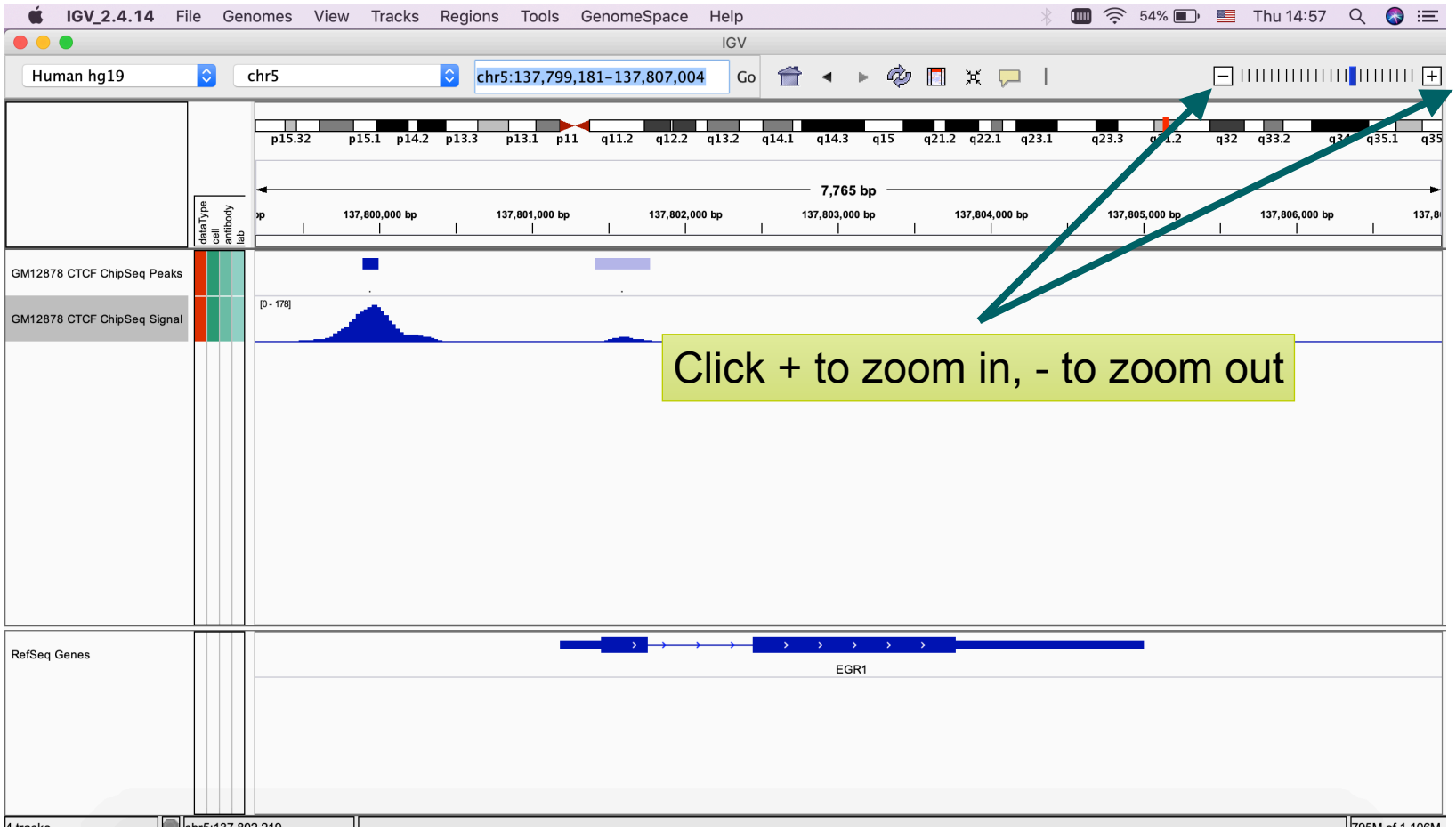
Navigate



Institute for Computational Genomics

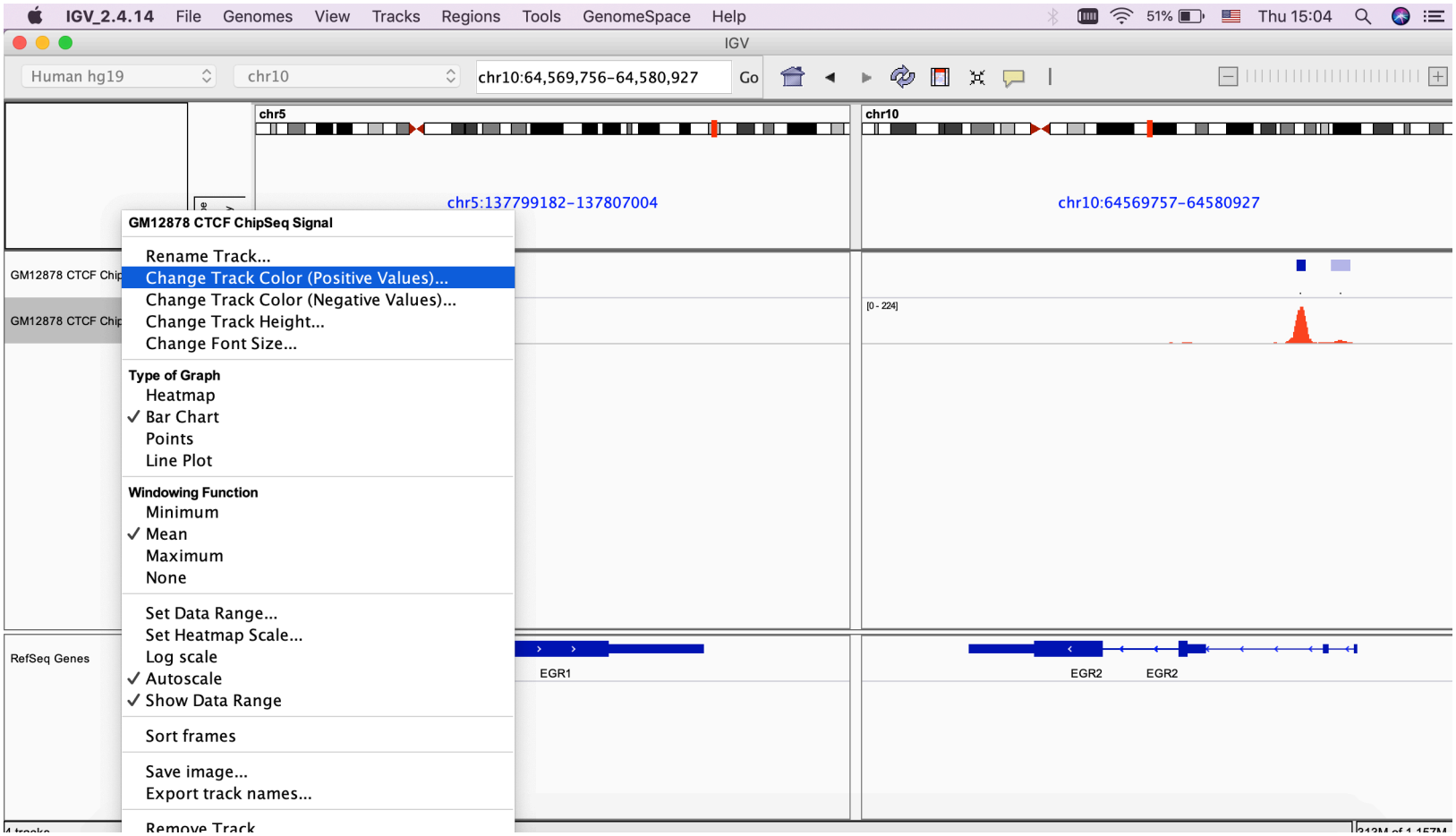
RWTH AACHEN UNIVERSITY

Navigate


IGV_2.4.14 File Genomes View Tracks Regions Tools GenomeSpace Help
IGV
Human hg19
chr5
chr5:137,799,181-137,807,004
Go
7,765 bp
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
RefSeq Genes
EGR1
Click + to zoom in, - to zoom out

Viewing multiple regions

Change the track color



Hands on

- Change the color of CTCF track
- Find more than two interesting regions
- View the multiple regions
- Load more dataset from ENCODE project
  - H3K4me1 of GM12878
  - H3K4me3 of GM12878
- Use different color for these three tracks

IGV Tools

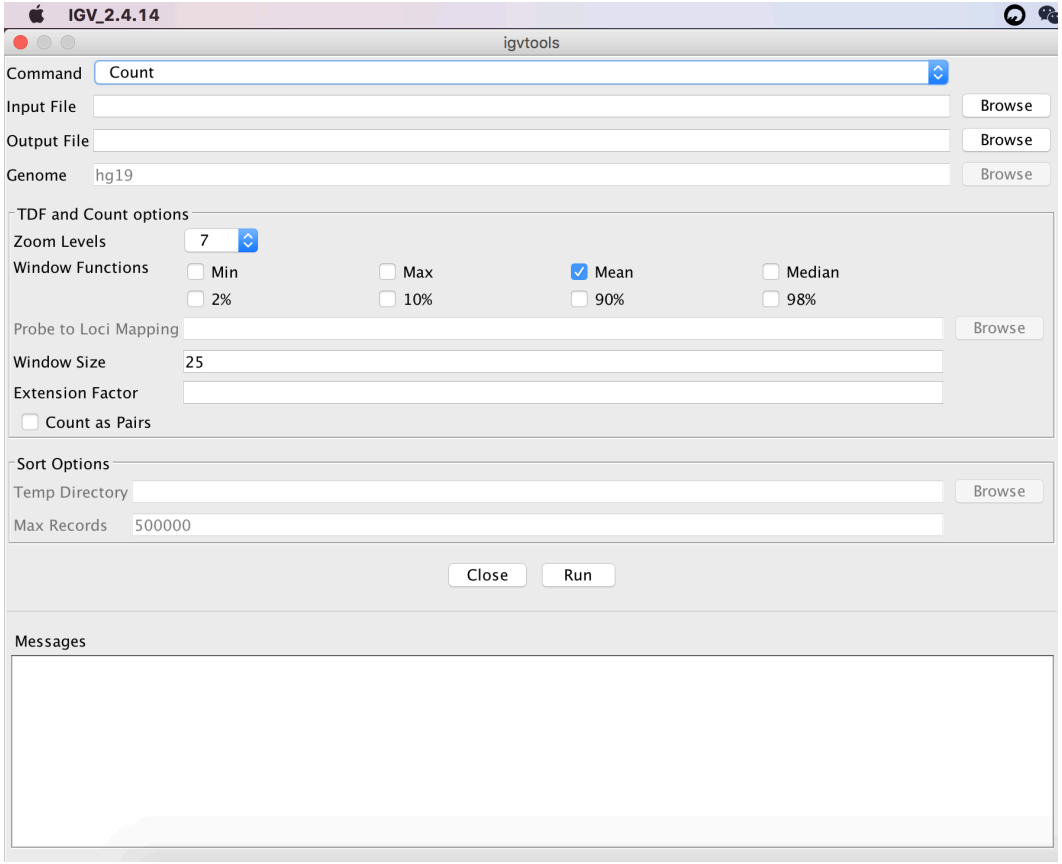
A set of utilities for preparing files for efficient display

- toTDF
  - Converts sorted data file to binary file (TDF).
- counts
  - Computes average alignment or feature over a window size across the genome
- sort
  - Sorts file by genomic position
- index
  - Creates an index file for alignment or feature file

IGV Tools

Can be lunched from the IGV user interface

Tools > Run igvtools…



toTDF

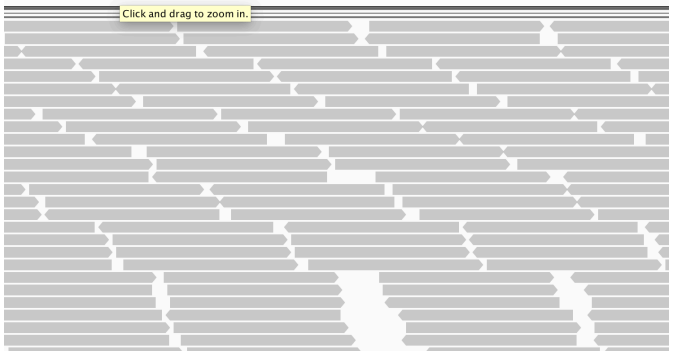
The **toTDF** utility converts large data files into tiled data format (.tdf) files

TDF files have the following advantages:

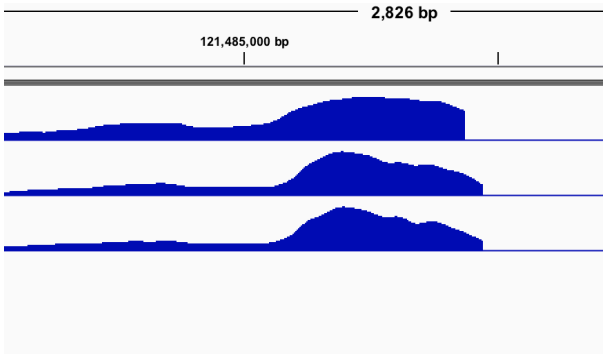
- Data is indexed for efficient retrieval
- Data is preprocessed for zoomed out views
- TDF files are web friendly - large data can be shared over the web.

count

The **count command** is used to transform alignment files to read density TDF files, e.g. for ChIP-seq, RNA-seq and similar alignment counting experiments



count



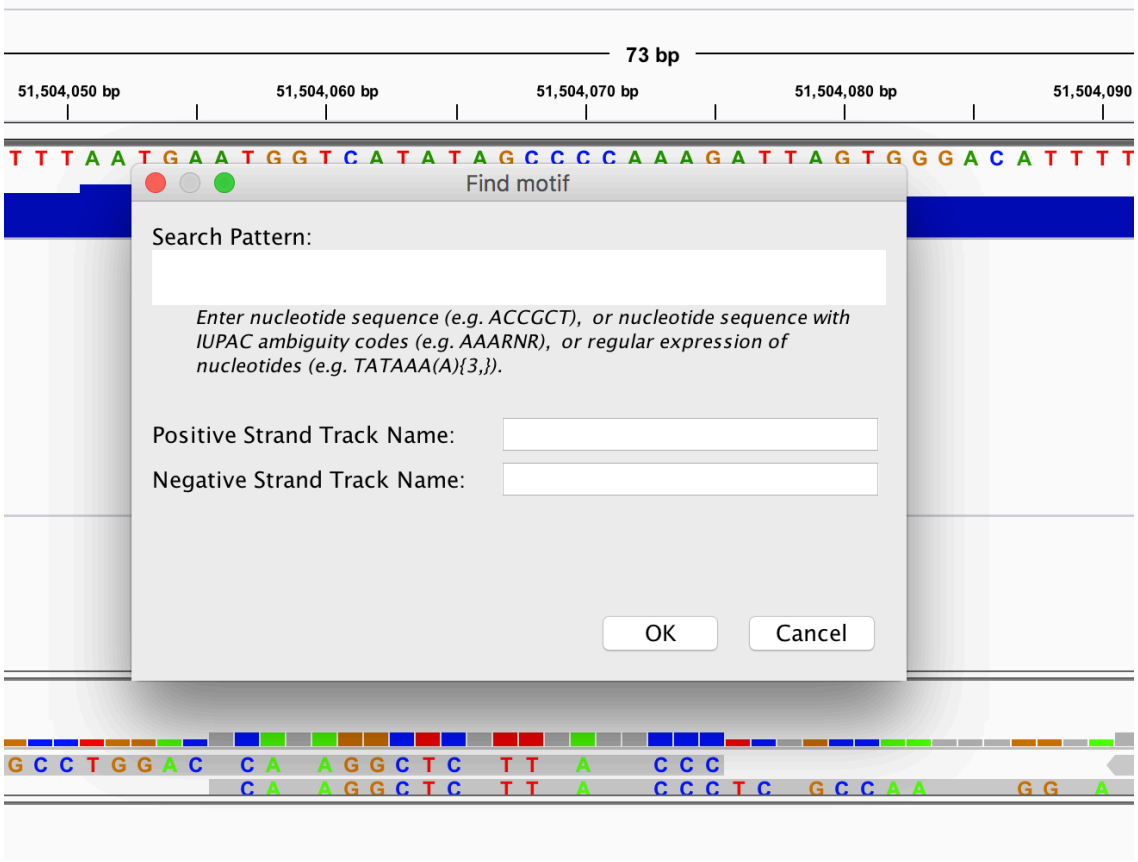
Alignment

Read Density

Find Motif

Lunched from the IGV user interface

Tools > *Find Motif*



hands on

download the BAM file from:

https://costalab.ukaachen.de/open_data/Bioinformatic_Analysis_in_R_2018/BIAR_D5/practice/igv_data.zip

Files included in the zip file:

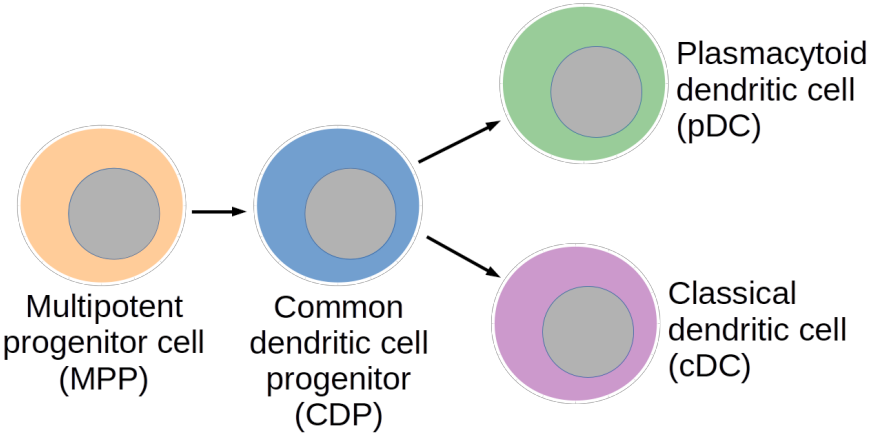
- cDC_H3K4me1_chr8.bam
- cDC_H3K4me1_chr8_peak.bed
- cDC_PU1_chr8.bam
- cDC_PU1_chr8_peaks.bed
- script.zsh (only for reference)

Convert BAM into TDF using *count*

Explore the data

Example: Epigenetic Changes in Cell Differentiation

***invitro* system for mimicking dendritc cell differentiation**



ChIP-Seq

PU.1

H3K4me3

H3K4me1

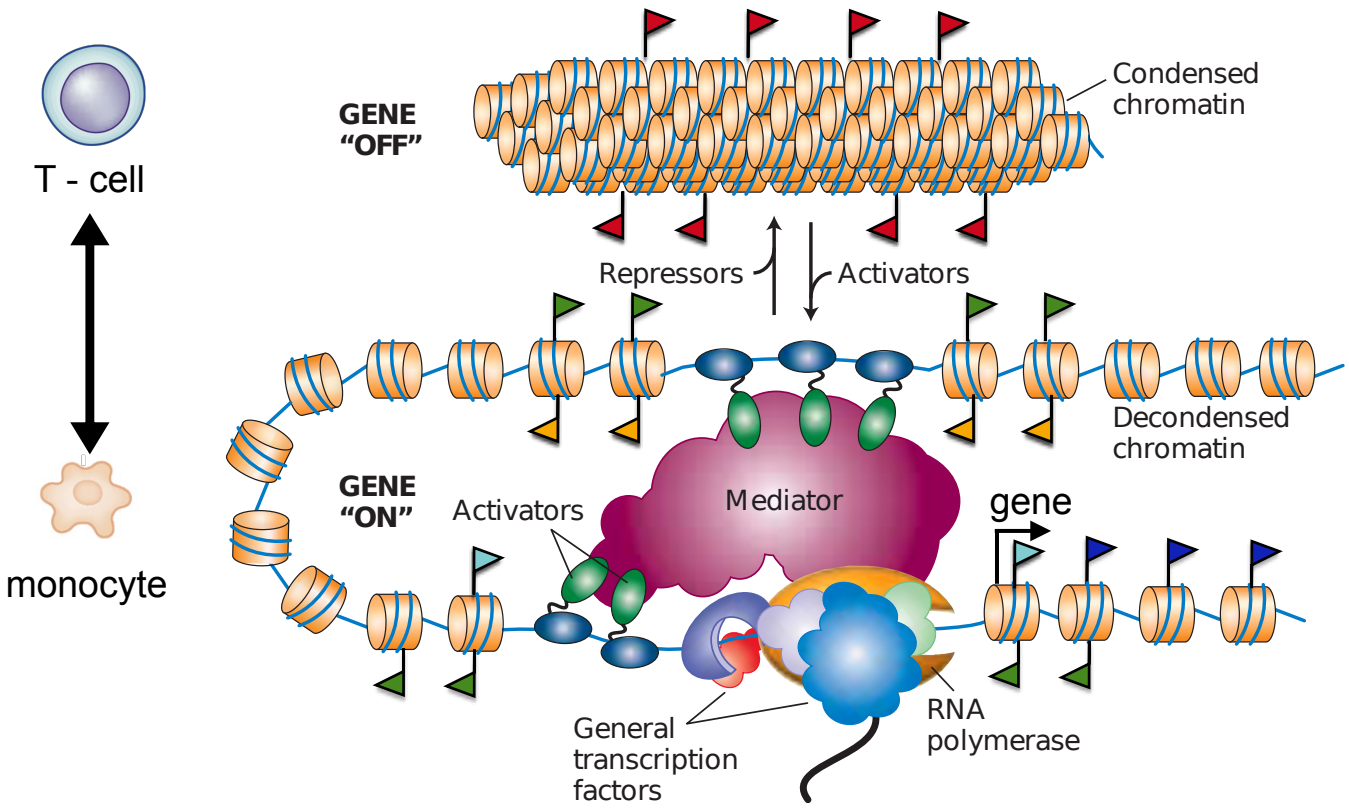
H3K27me3

Find binding sites of the transcription factor PU.1

PU.1 Motif

Find histone modifications associated to PU.1 binding sites

Chromatin, Histone Code and TF binding



Histone Code

Transcription

H3K79me2, H3k36me3

Active Regions

H3K27ac, H3K9ac

Active Promoters

H3K4me3

Active Enhancers

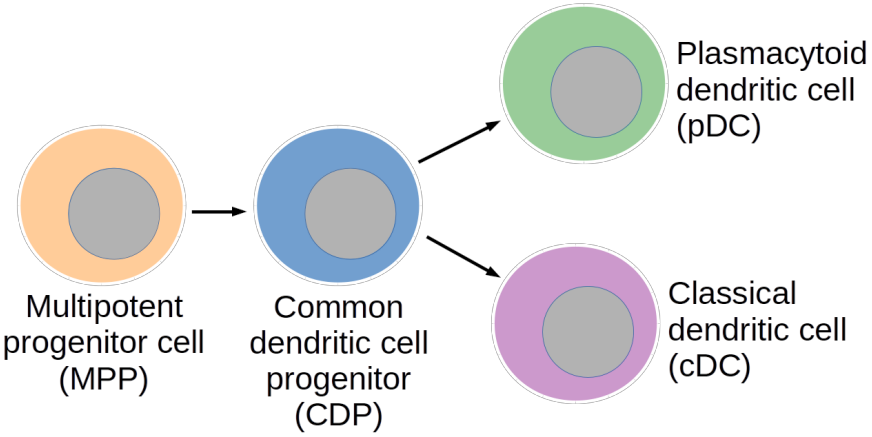
H3K4me1

Repressed regions

H3K27me3, H3K9me3

Example: Epigenetic Changes in Cell Differentiation

***invitro* system for mimicking dendritc cell differentiation**



ChIP-Seq

PU.1

Find binding sites of the transcription factor PU.1

PU.1 Motif

H3K4me3

H3K4me1

H3K27me3

Active Promoters

Active Enhancers

Repressed regions

difications
.1 binding sites

www.costalab.org

Institute for
Computational Genomics
01011011010
10100100101

RWTH AACHEN
UNIVERSITY