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:

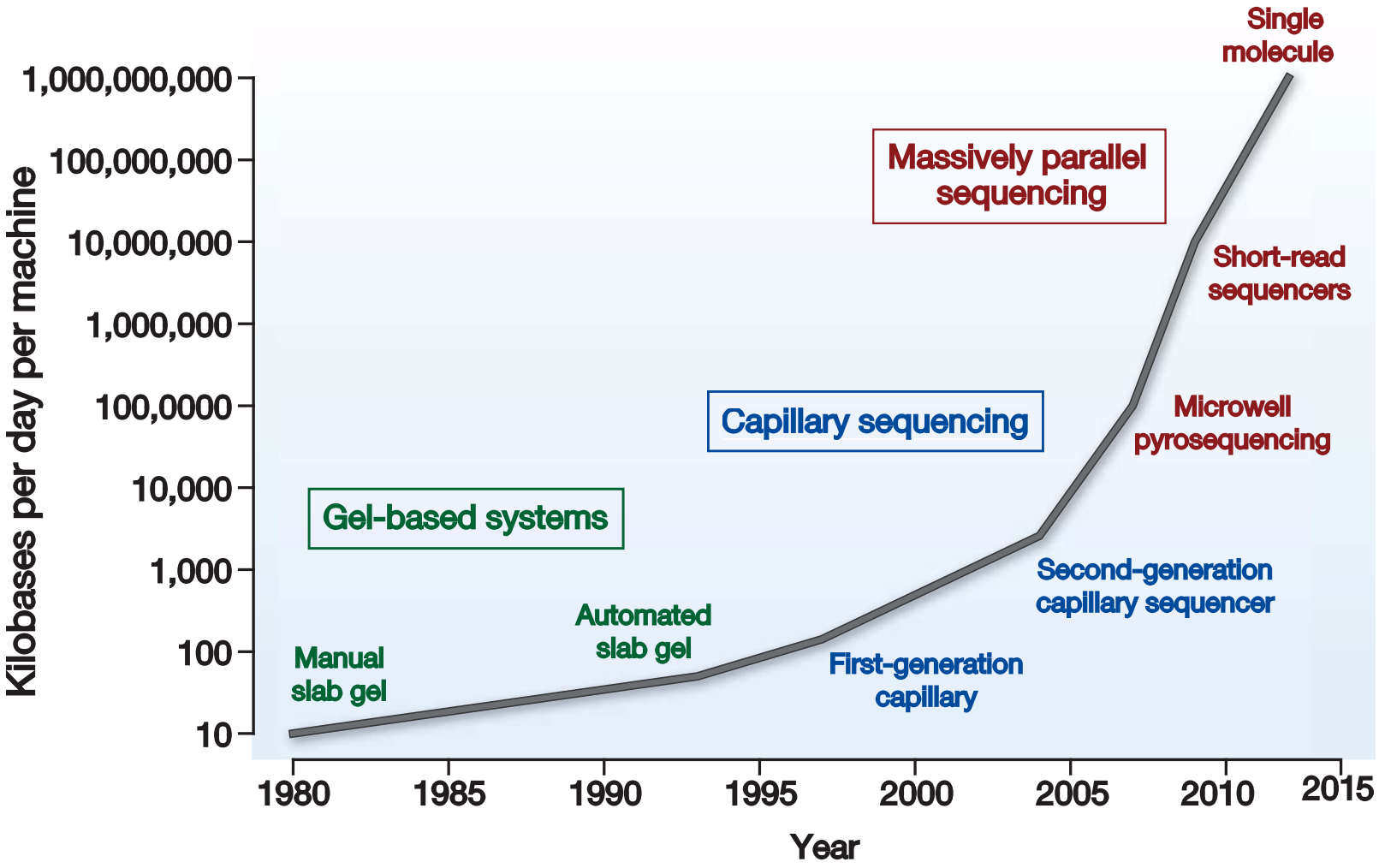
- ChIP Sequencing -> a protein with DNA
- CLIP Sequencing -> a protein with RNA
- ChIRP Sequencing -> a RNA with DNA
- …

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

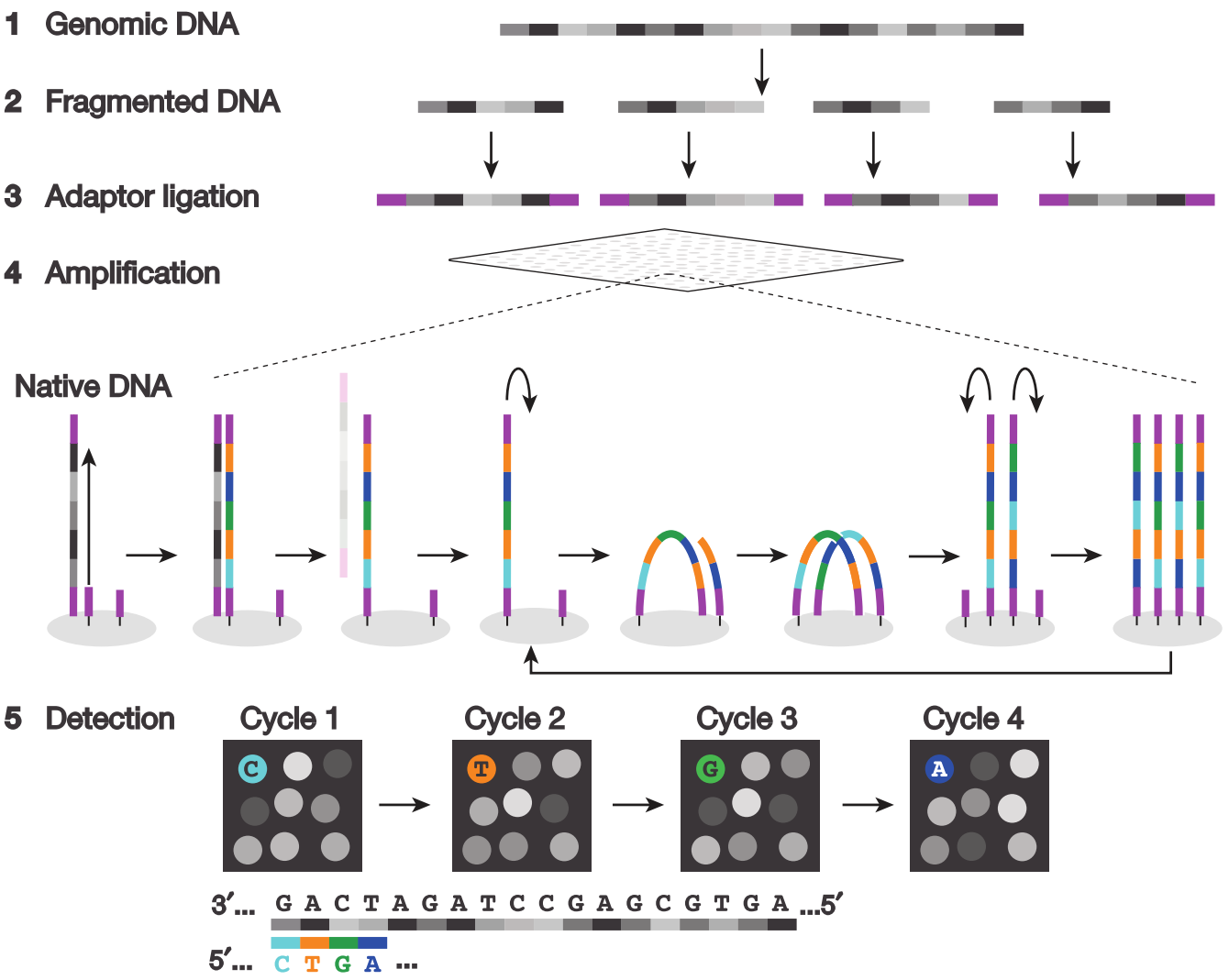
Next Generation Sequencing

- 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

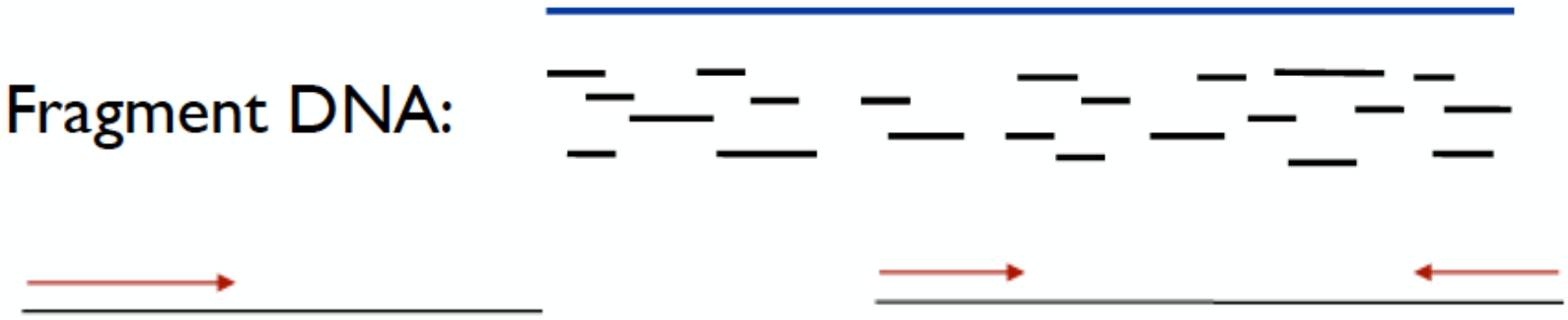


Jay Shendure, *et al (2017), Nature*

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

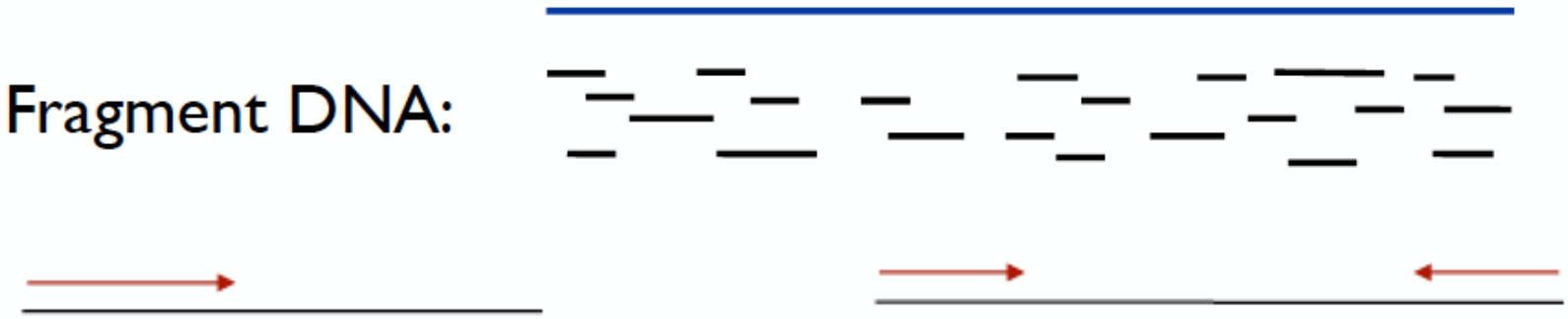
Read Types



Single end

Paired end
Ins: 200-800 bp

Read Types



Single end

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

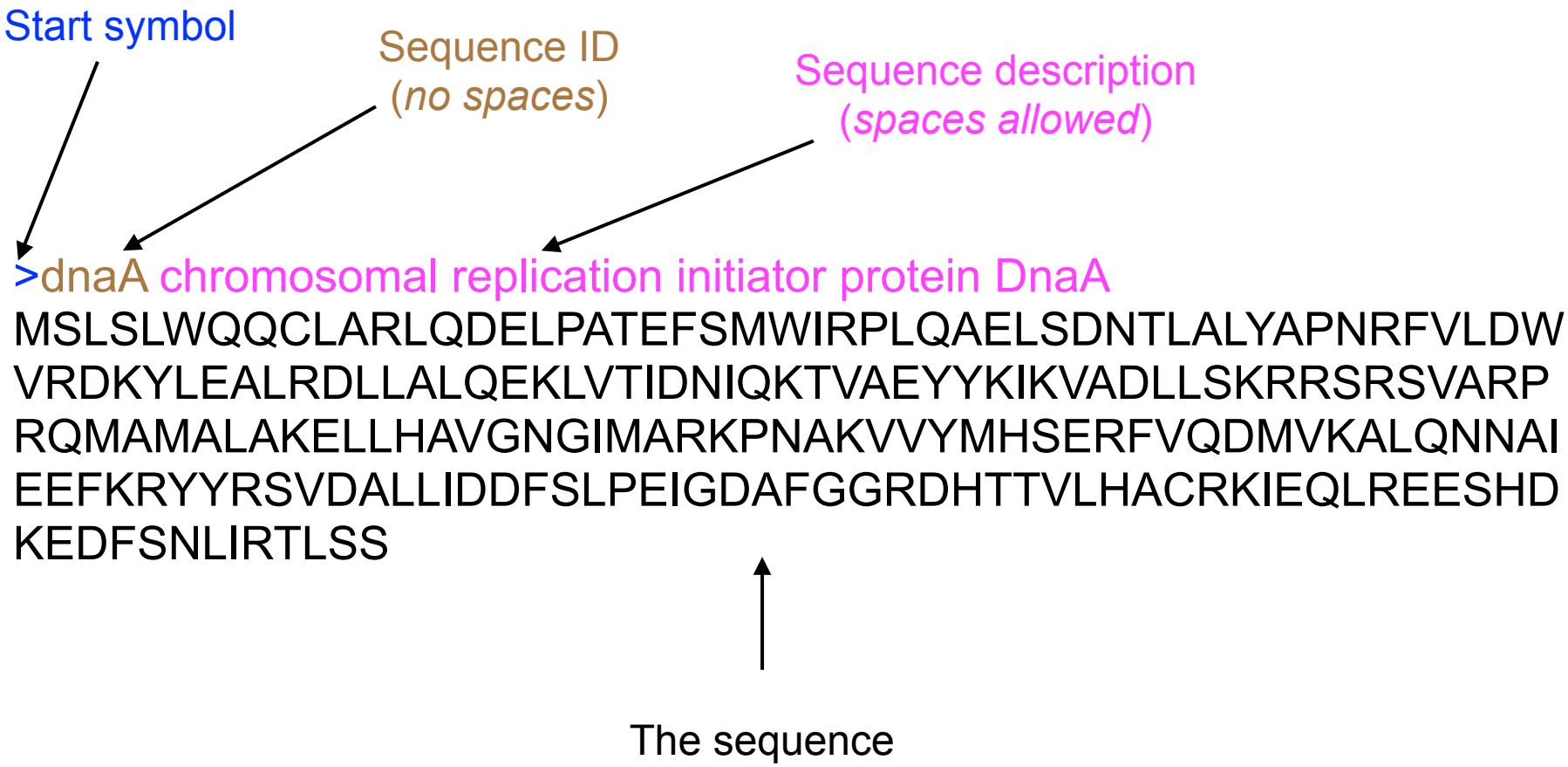
Paired end
Ins: 200-800 bp

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

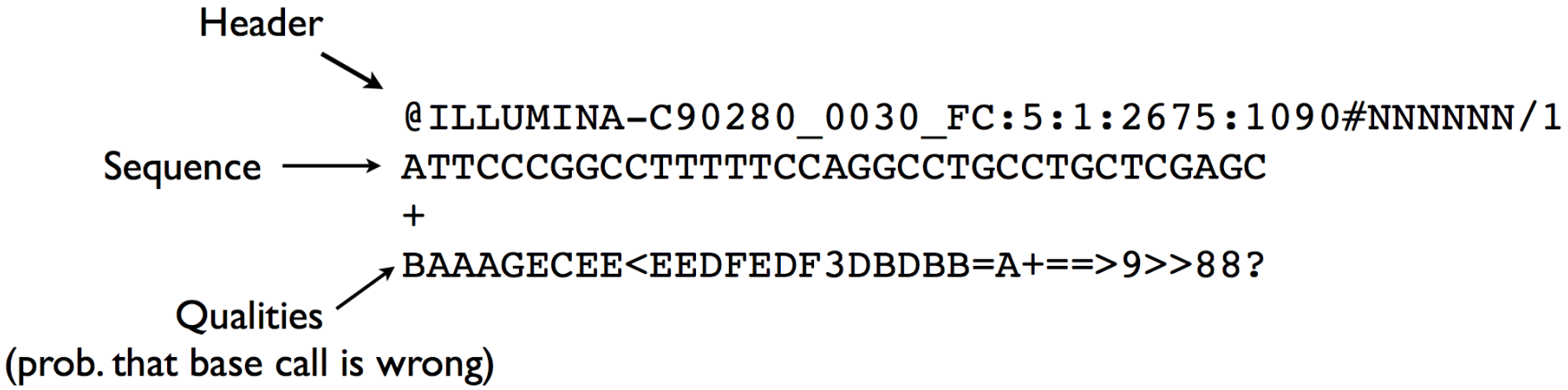
FASTA files

>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

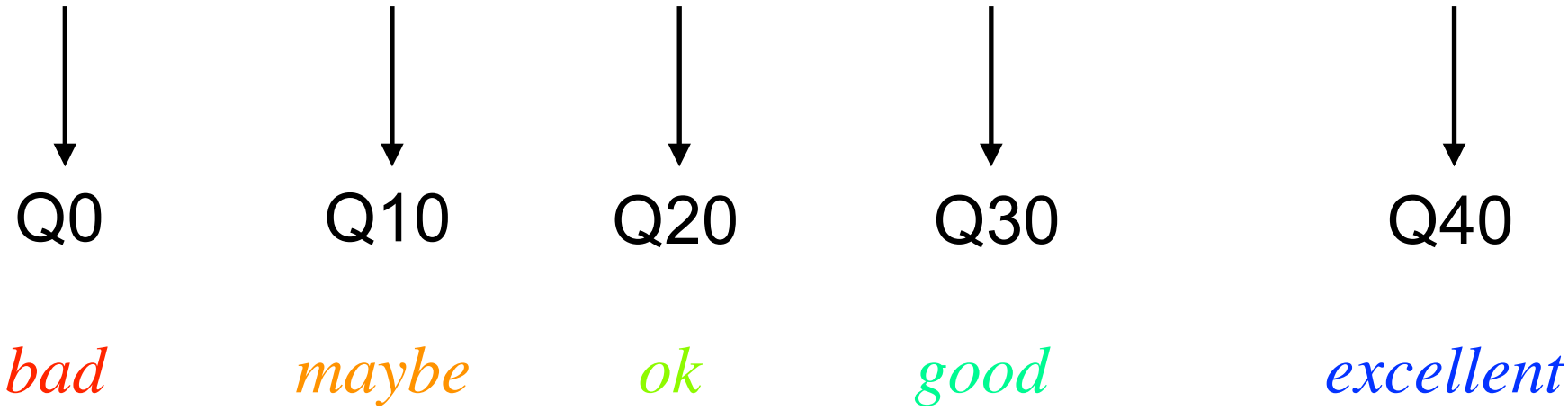
Q = -10 * log10 P

P = 10^(-Q/10)

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

FASTQ files

Uses letters/symbols to represent numbers:

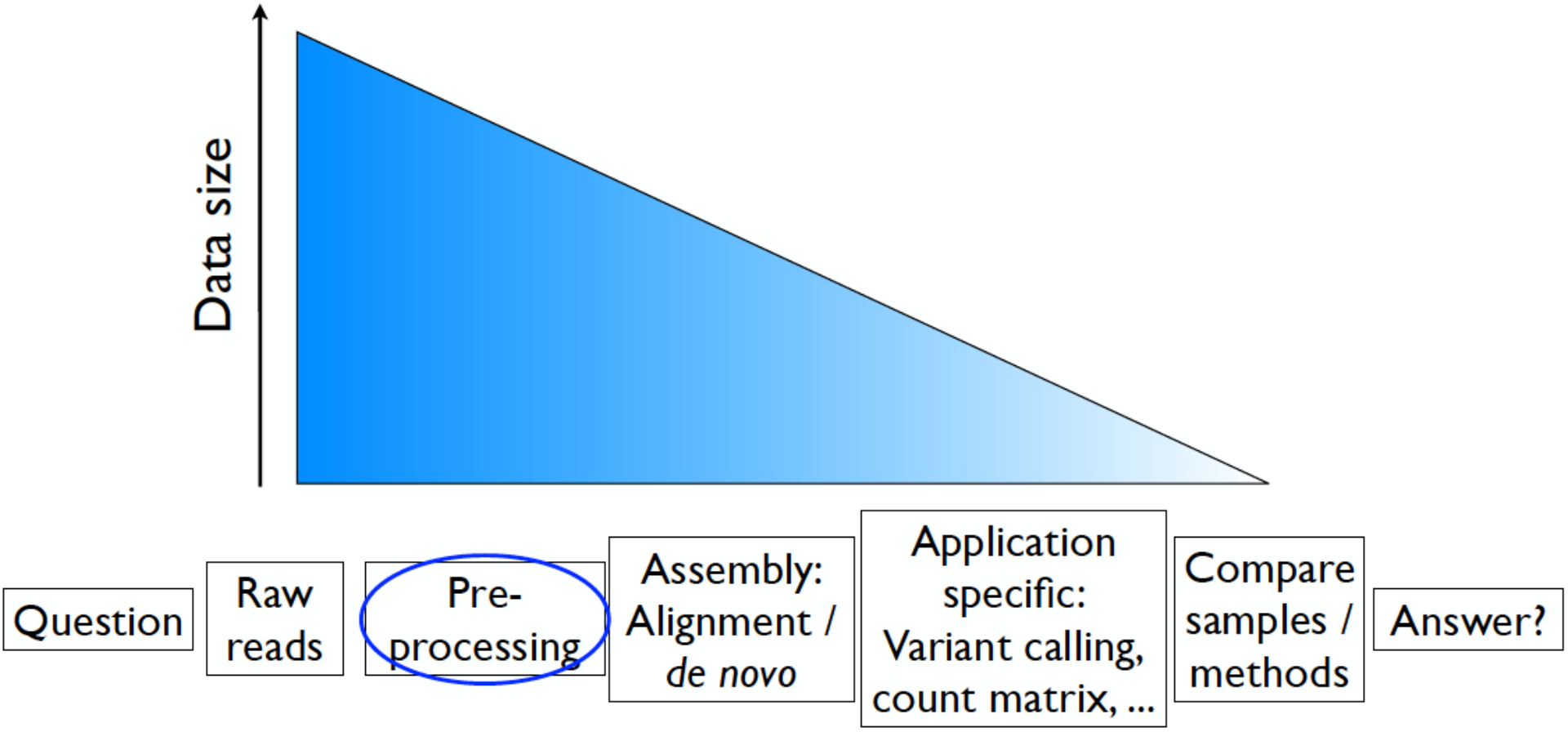


Bioinformatics Analysis in R

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), …

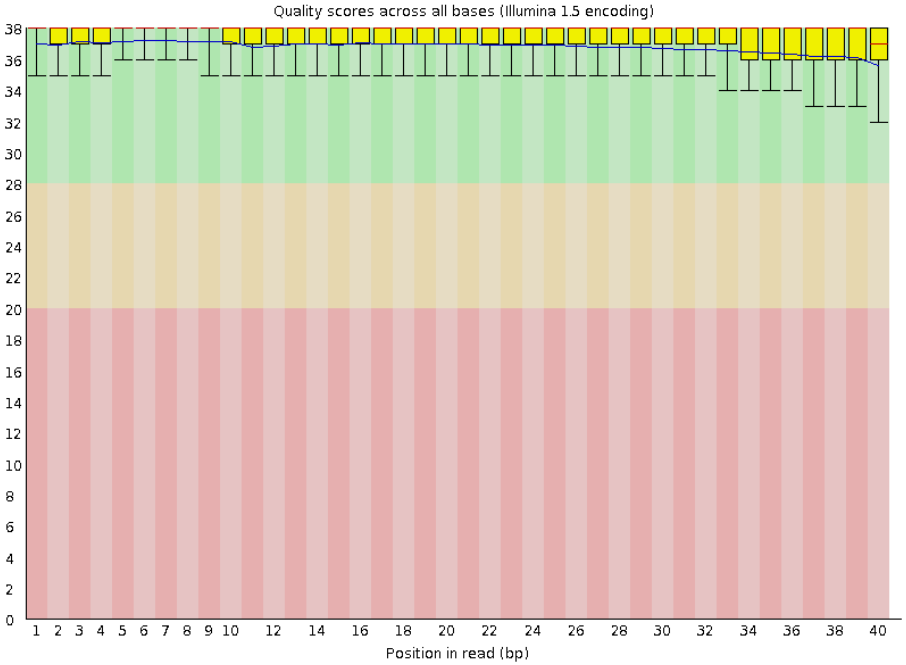
Quality Control

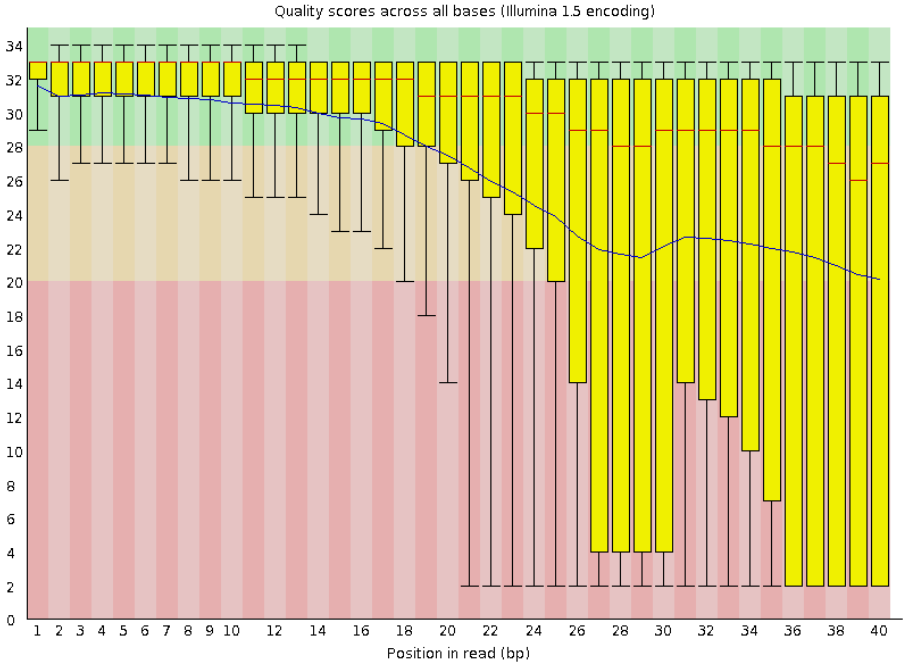
- **FastQC (usually provided by NGS core facilities)**
  - tool to analyse quality of reads from sequencing.
  - indicate problems in library preparation or sequencing steps.
- Example of good sequences:
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
- or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

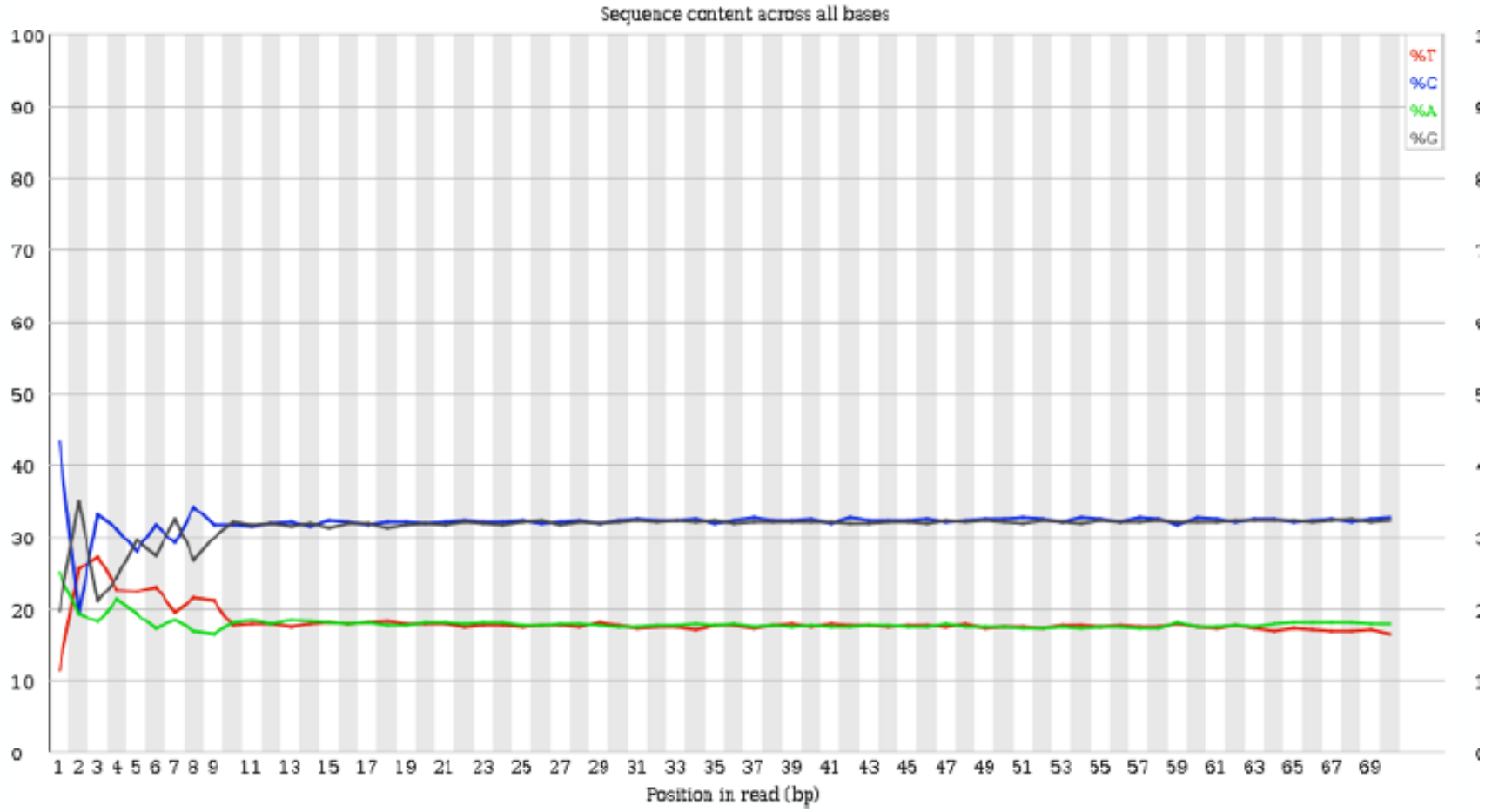




Solution: trim end of reads with low quality

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

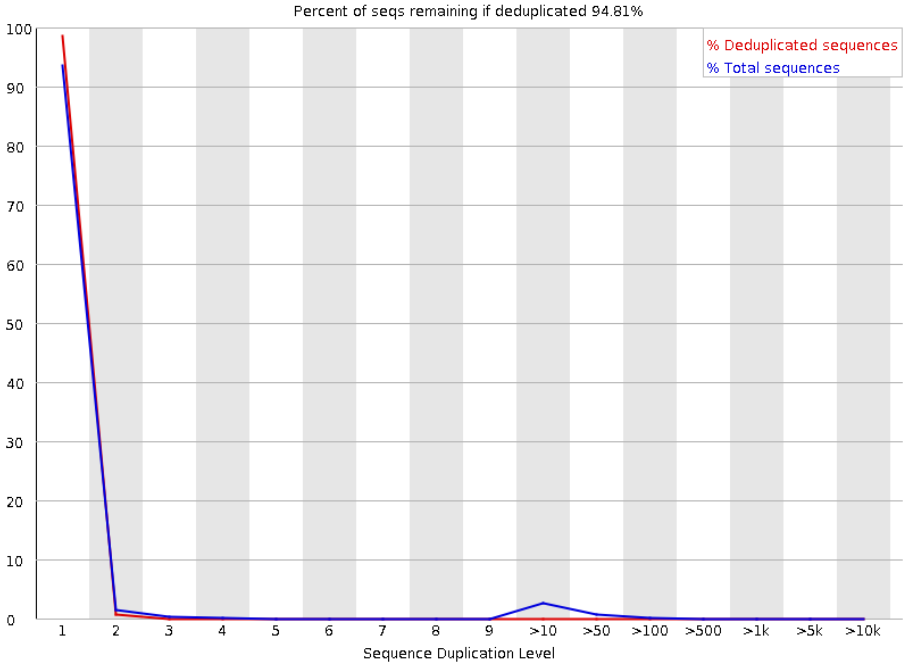
Read position sequence bias

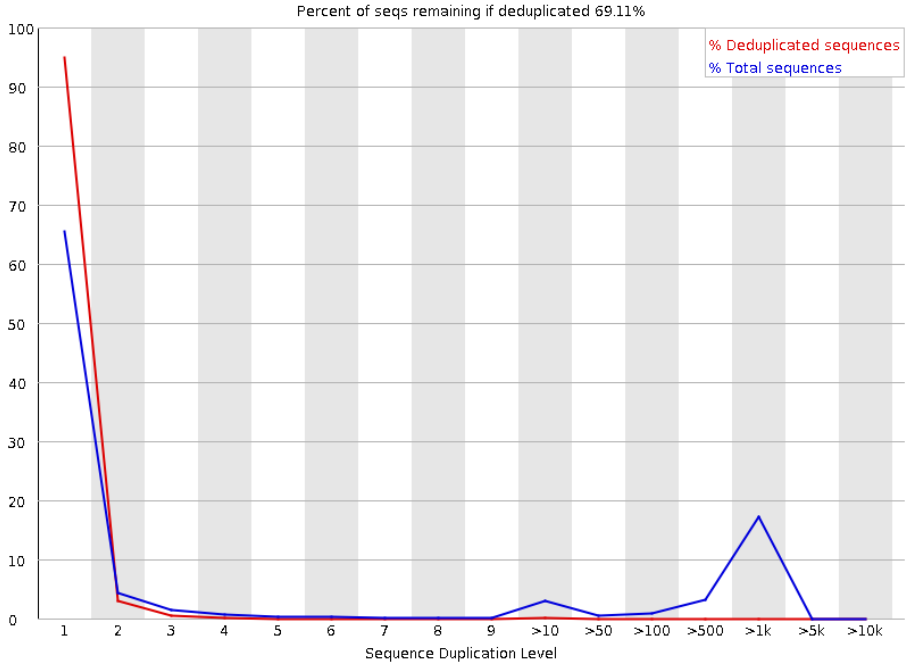


- **Trim read starts**

Quality Control

Sequence duplication levels





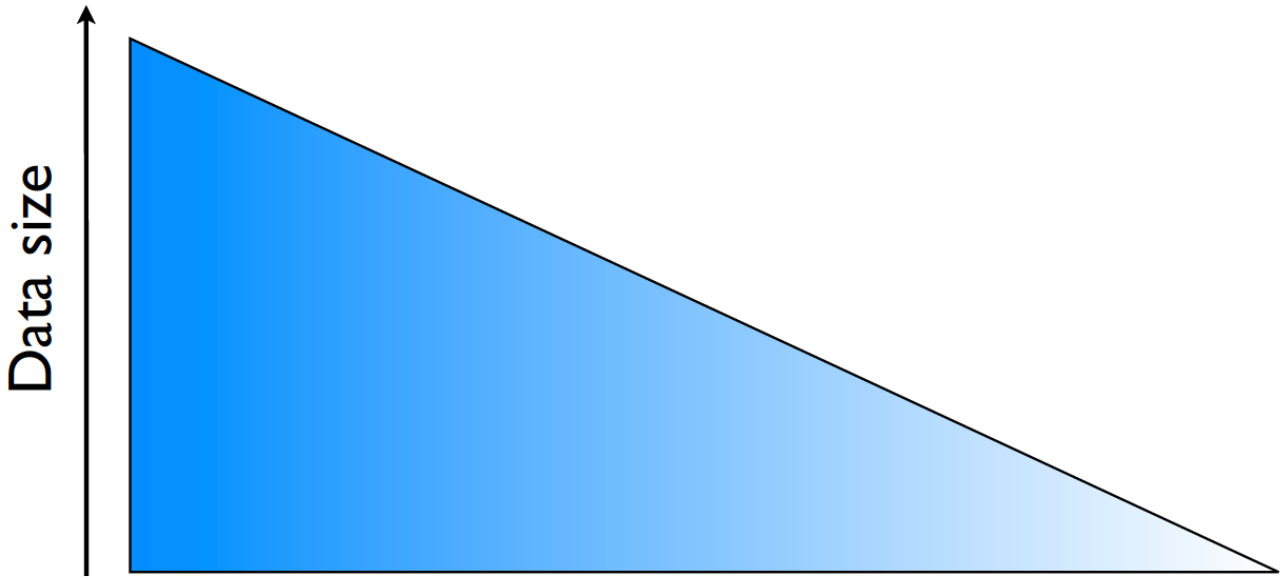
- **Solution: remove duplicates**

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

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

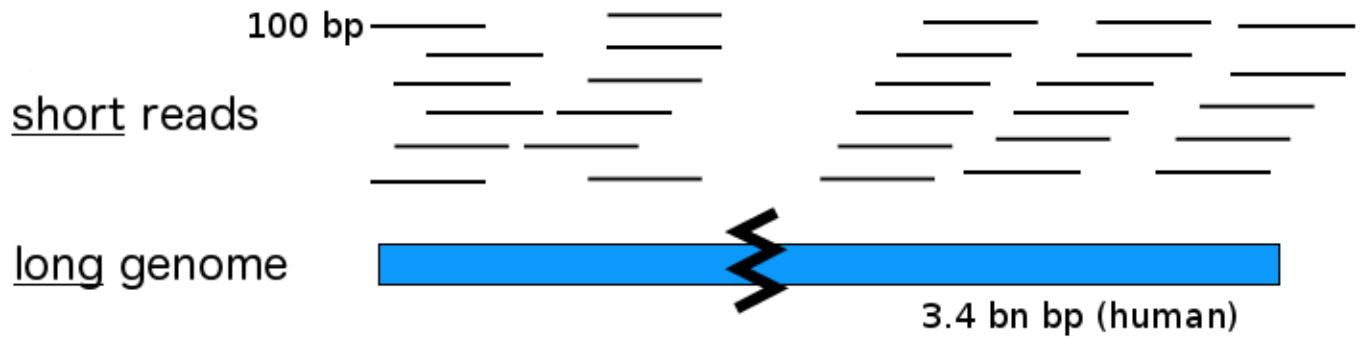
(Short reads) Alignment



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**

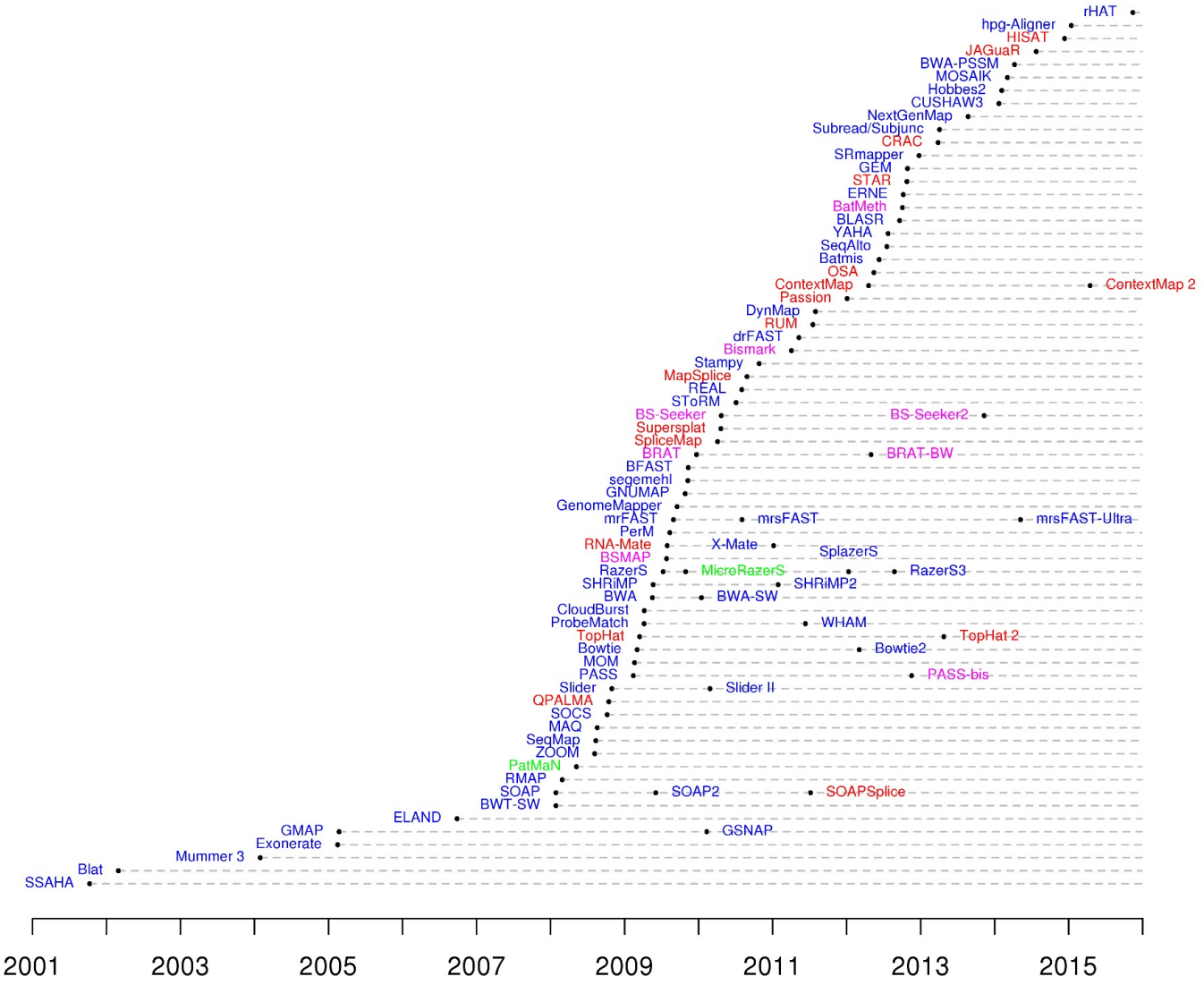
Solution:

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

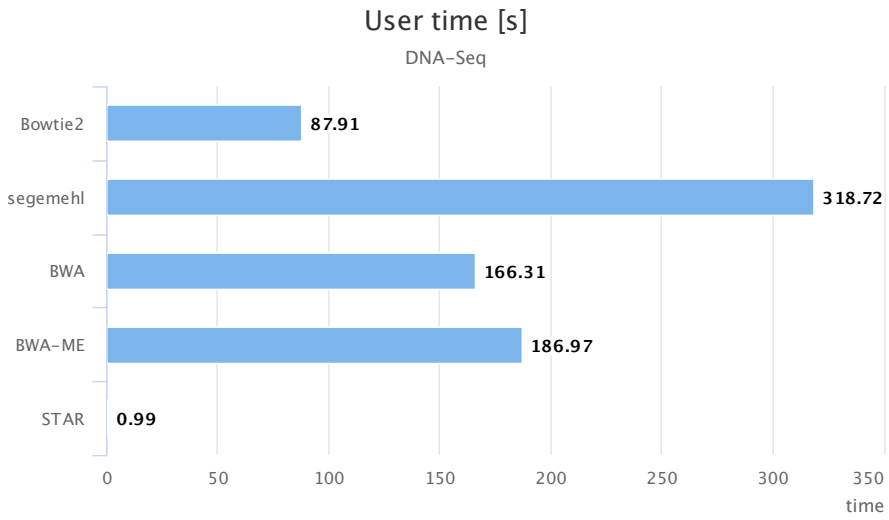
RWTH AACHEN UNIVERSITY

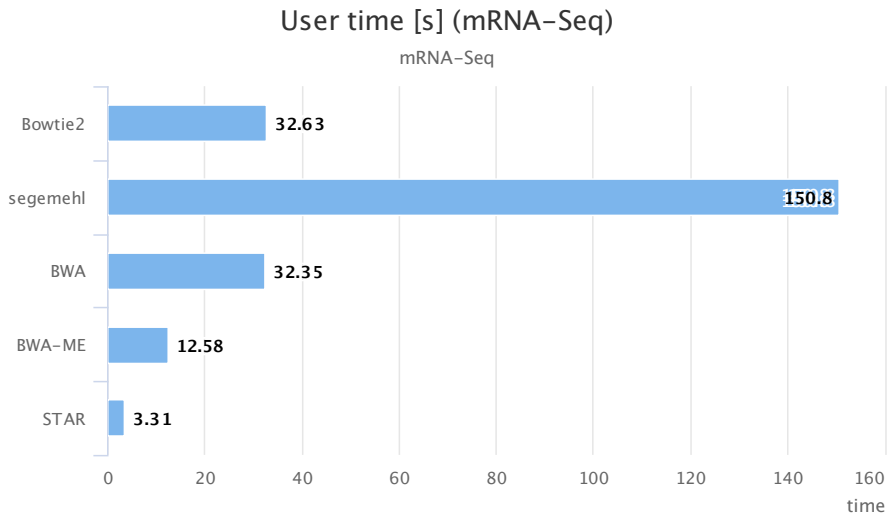
Reference based aligners - Overview

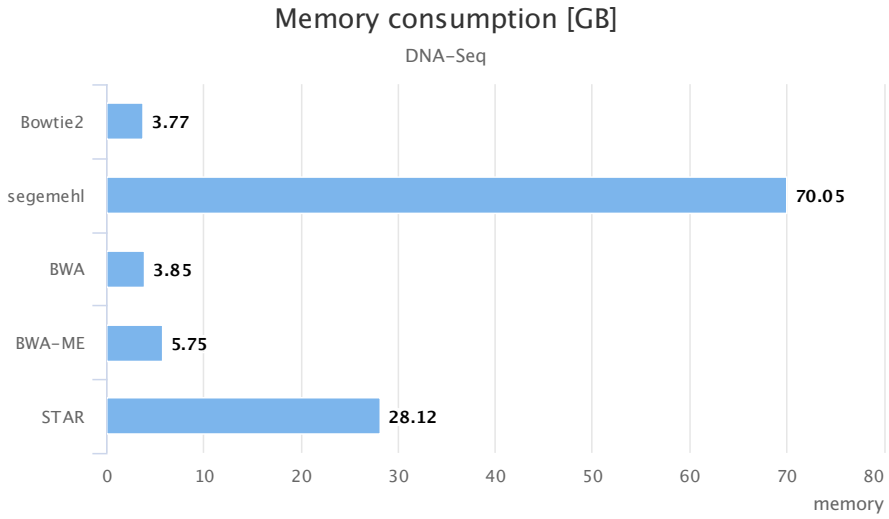
	Time	Precision	Pairs	GAPs	Phred	Memory	Application (Comments)
BOWTIE	+		+	-	-	5GB	General (max. 3 missmatches)
BWA	+		+	+	+	8GB	General (max of 200bps reads)
NOVOALIGN		+	+	+	+	8GB	General (commercial license)
STAR	+		+	-	+	32GB	RNA-Seq (allow split-maps)
BISMARK	+		+	+	+	10GB	Bisulfite/reduced sequencing

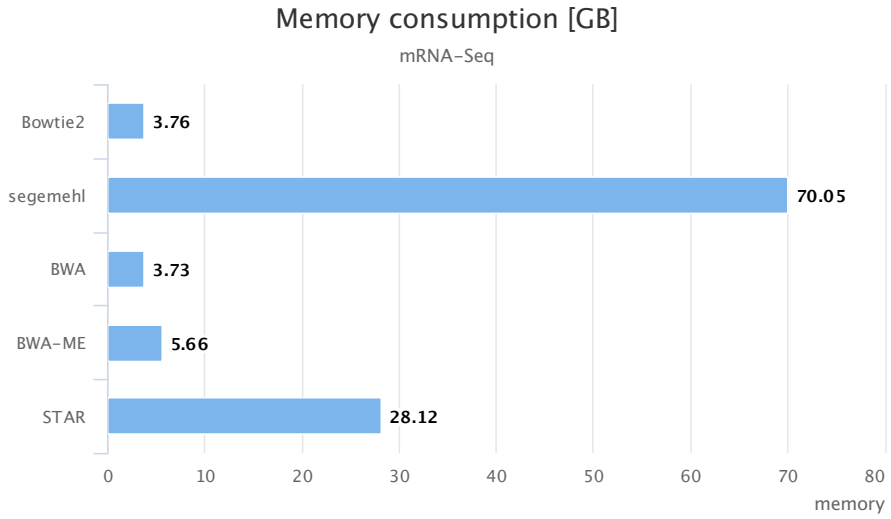
non comprehensive list

Alignment Tools









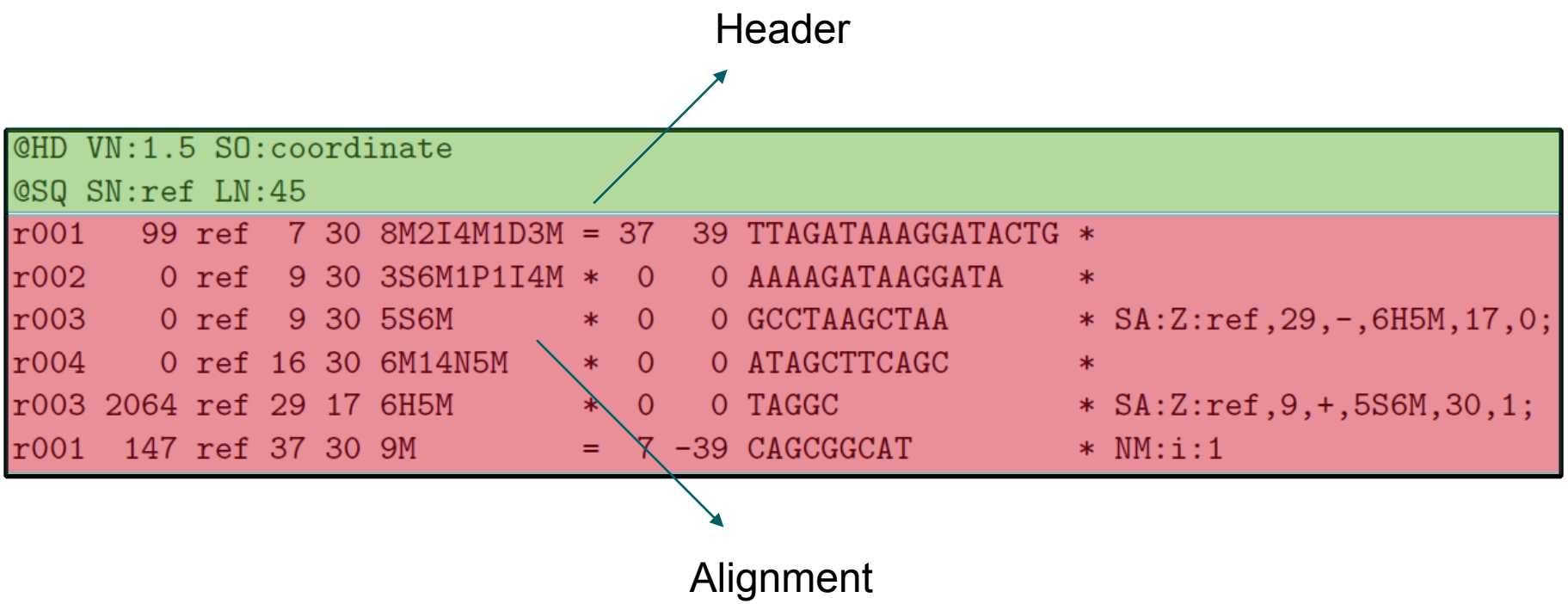
https://www.ecseq.com/support/benchmark

Institute for Computational Genomics

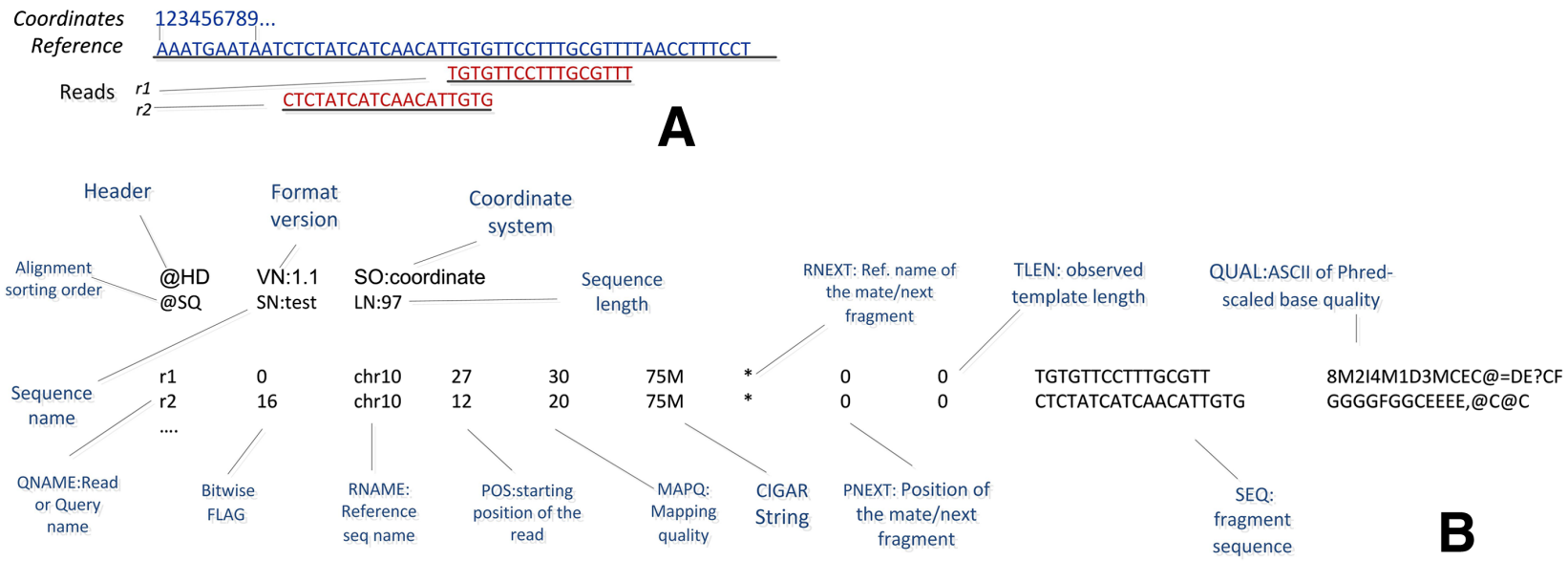
RWTH AACHEN UNIVERSITY

SAM Files

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

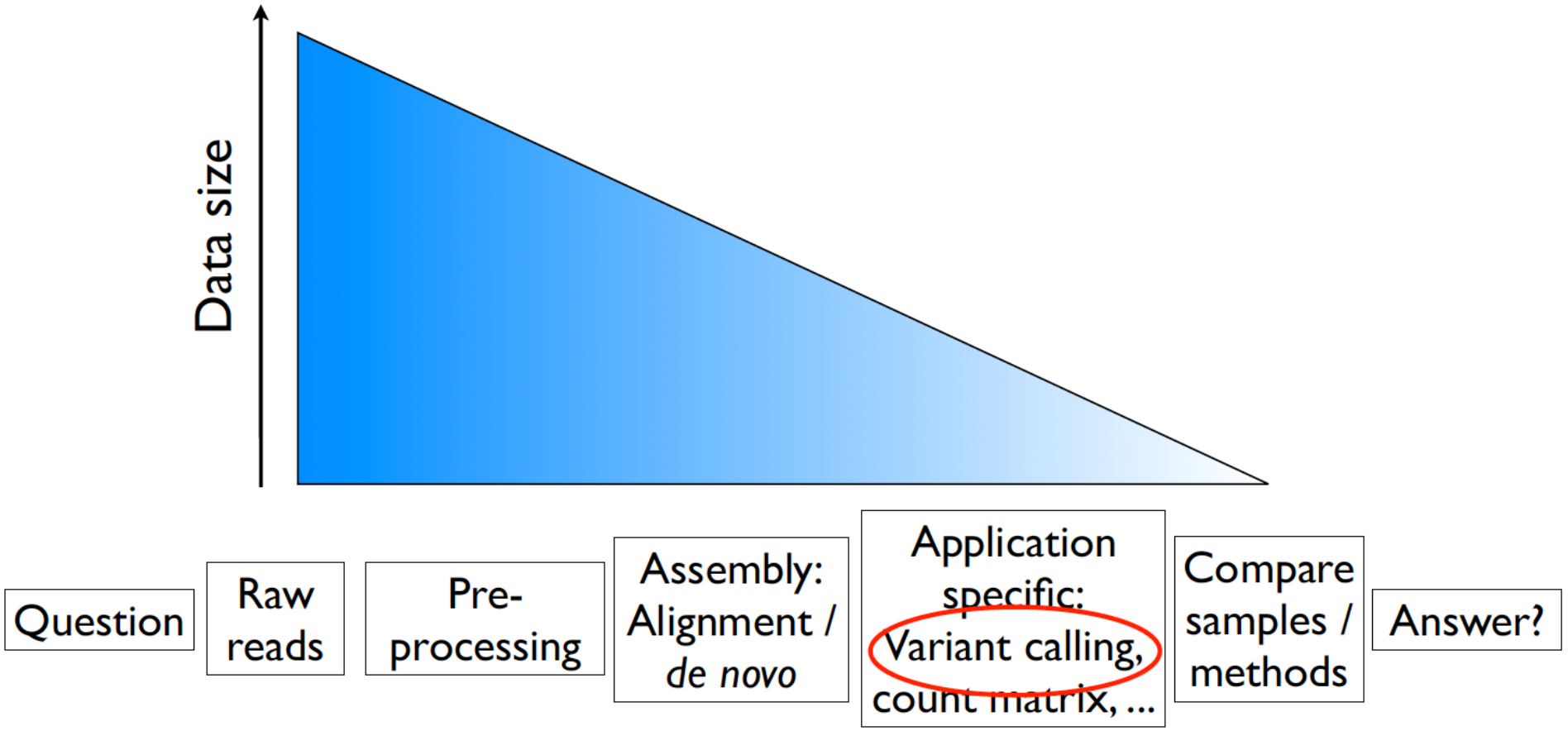


SAM Files- alignment section



- 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 - Variant Calling



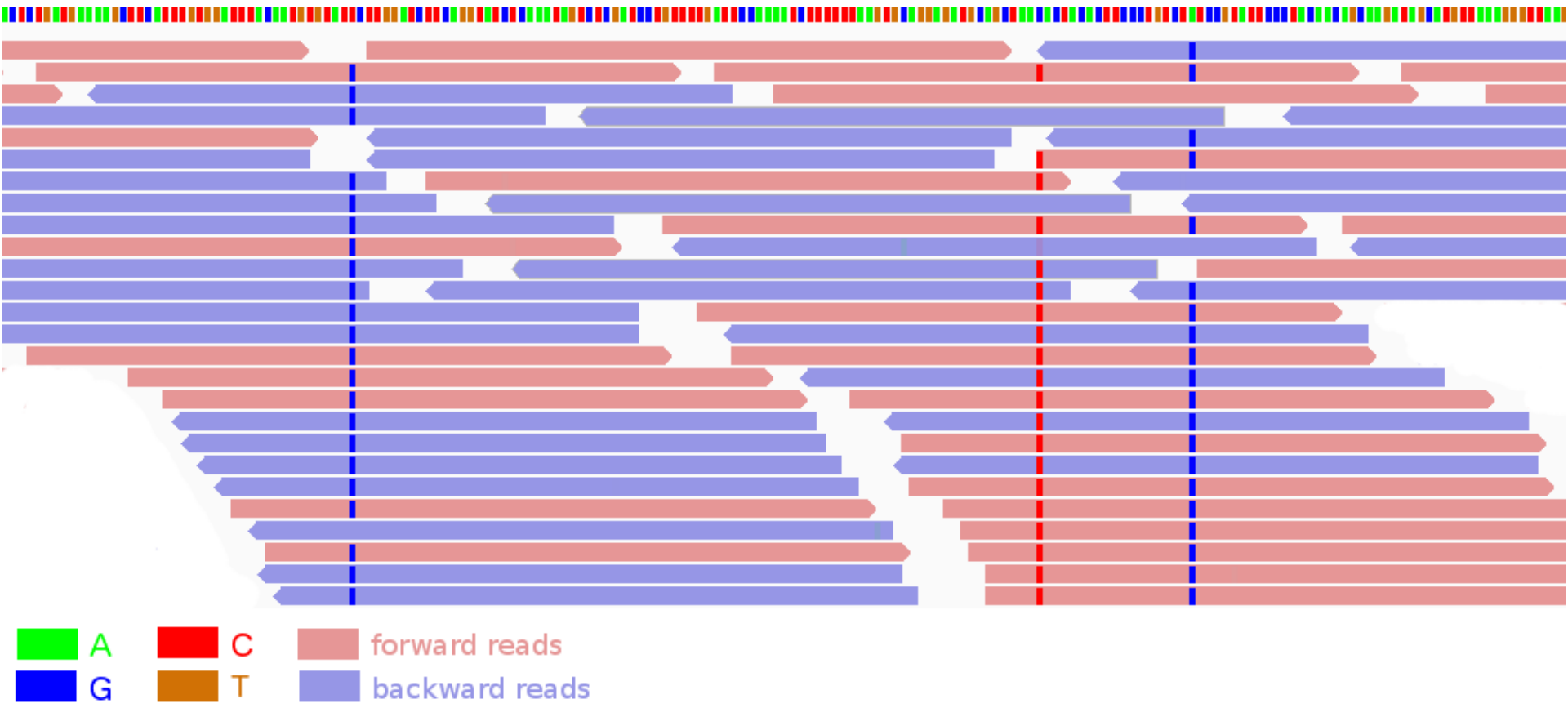
http://en.wikipedia.org/wiki/RNA-Seq

Example Application

SNP Detection - Individual DNA Sequencing

Alignment of short reads on DNA

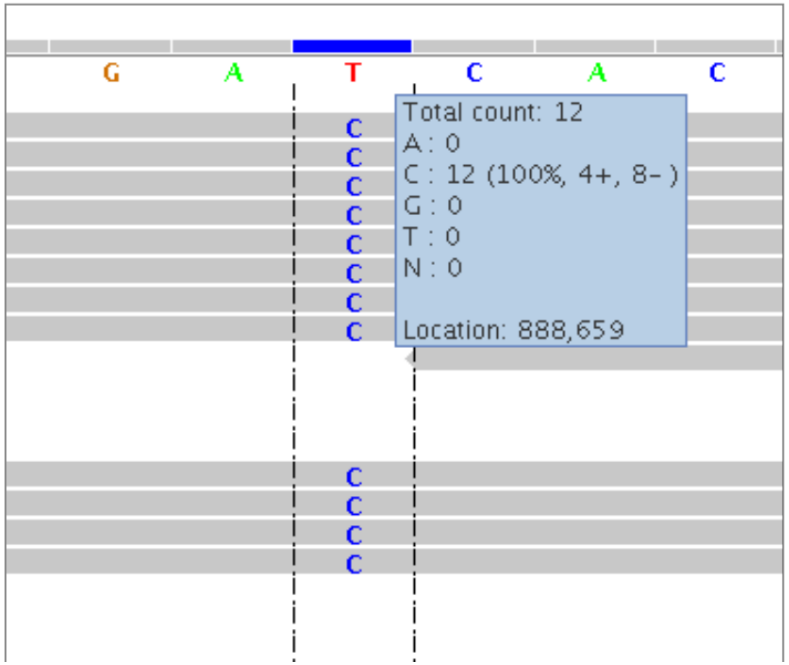
SNP Calling Example - Genome Browser

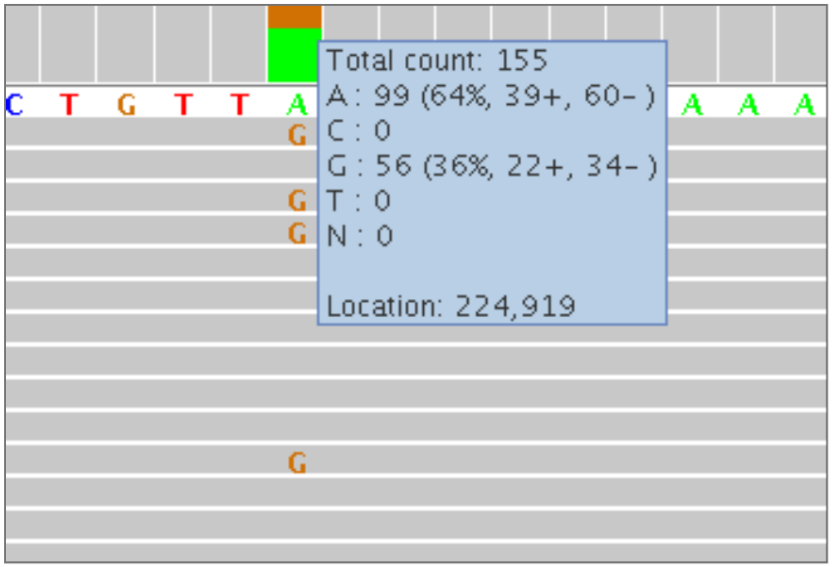


IGV only show miss-matches!

Simple SNP Calling Method

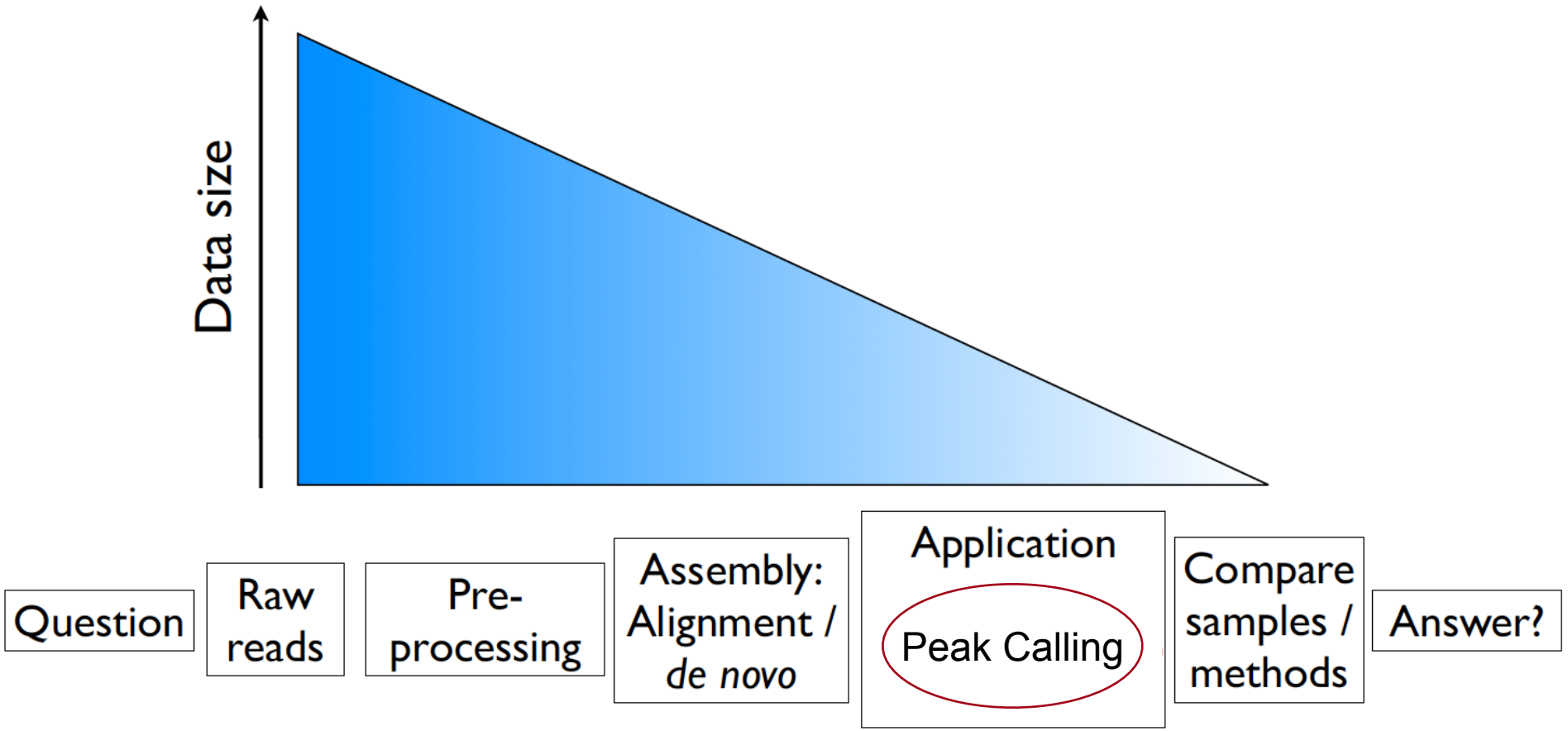
- Aims
  - Variant calling: Identify polymorphic sites => sites that differs from the reference.
  - Genotyping: Determine the genotype for a certain individual at such sites.





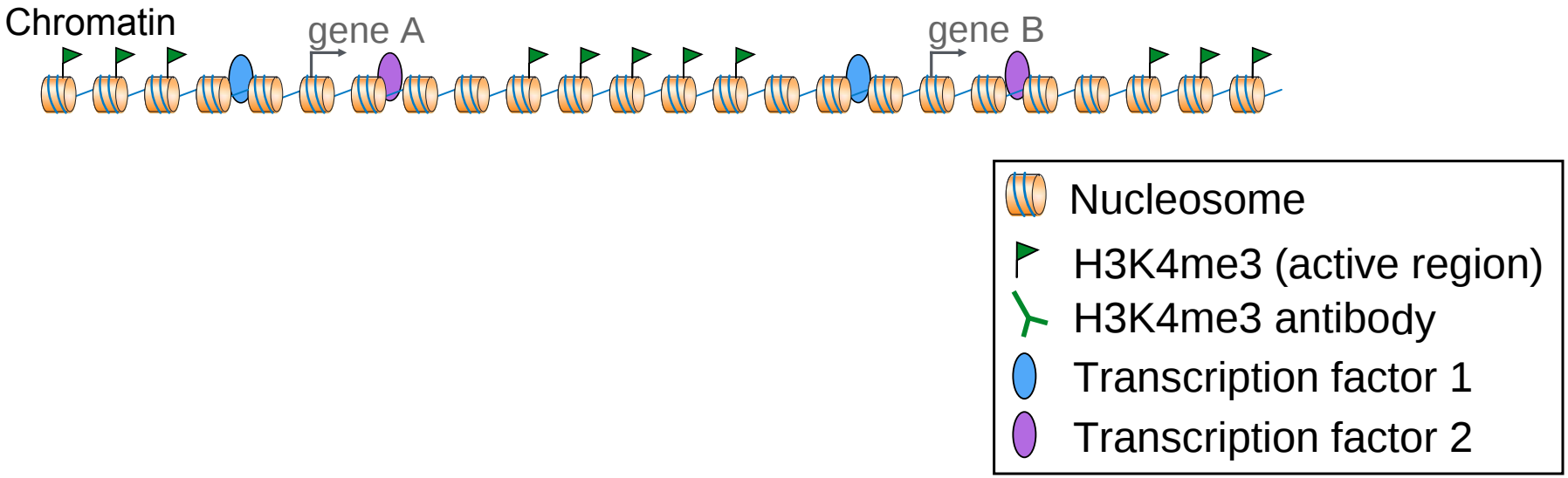
- Early methods
  - Works by simply counting the alleles at each site, and then identifying a variant by use of simple cutoff rules.

Applications - Peak Calling

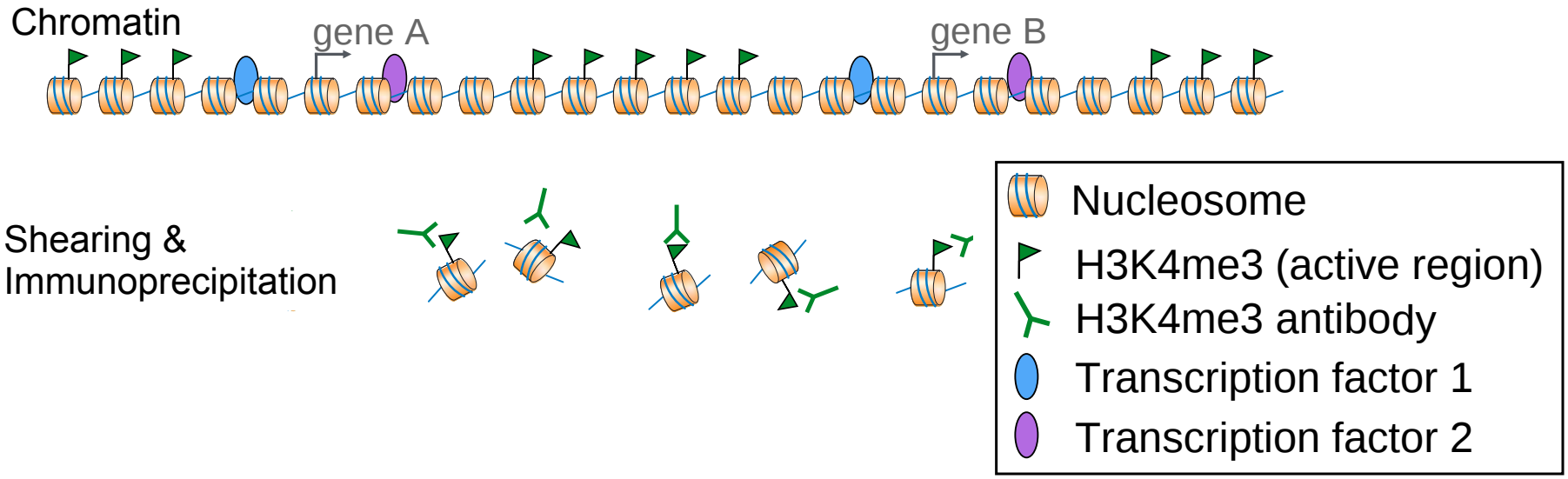


http://en.wikipedia.org/wiki/RNA-Seq

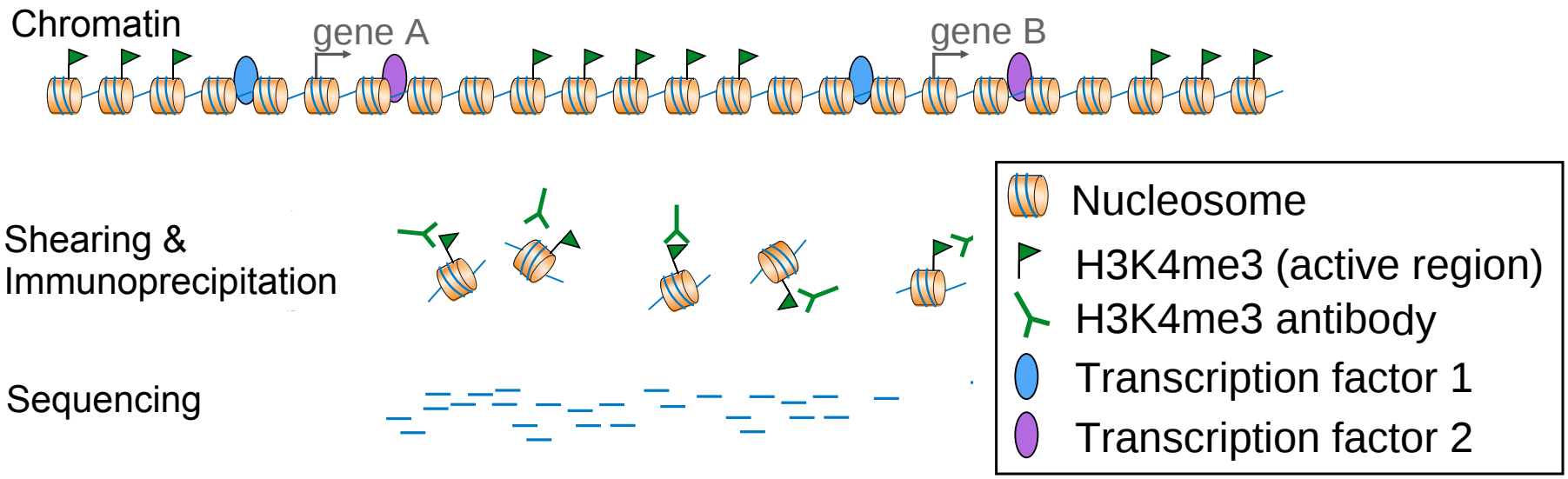
DNA - Protein interactions with ChIP-Seq



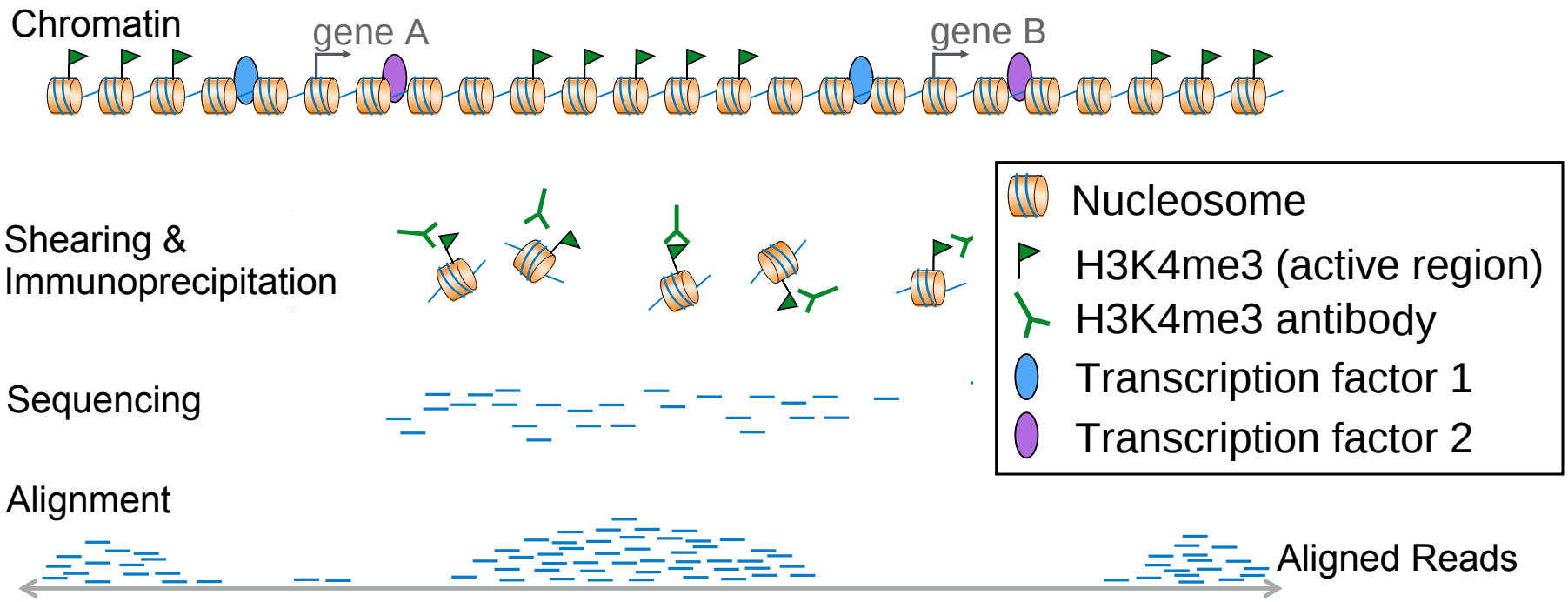
DNA - Protein interactions with ChIP-Seq



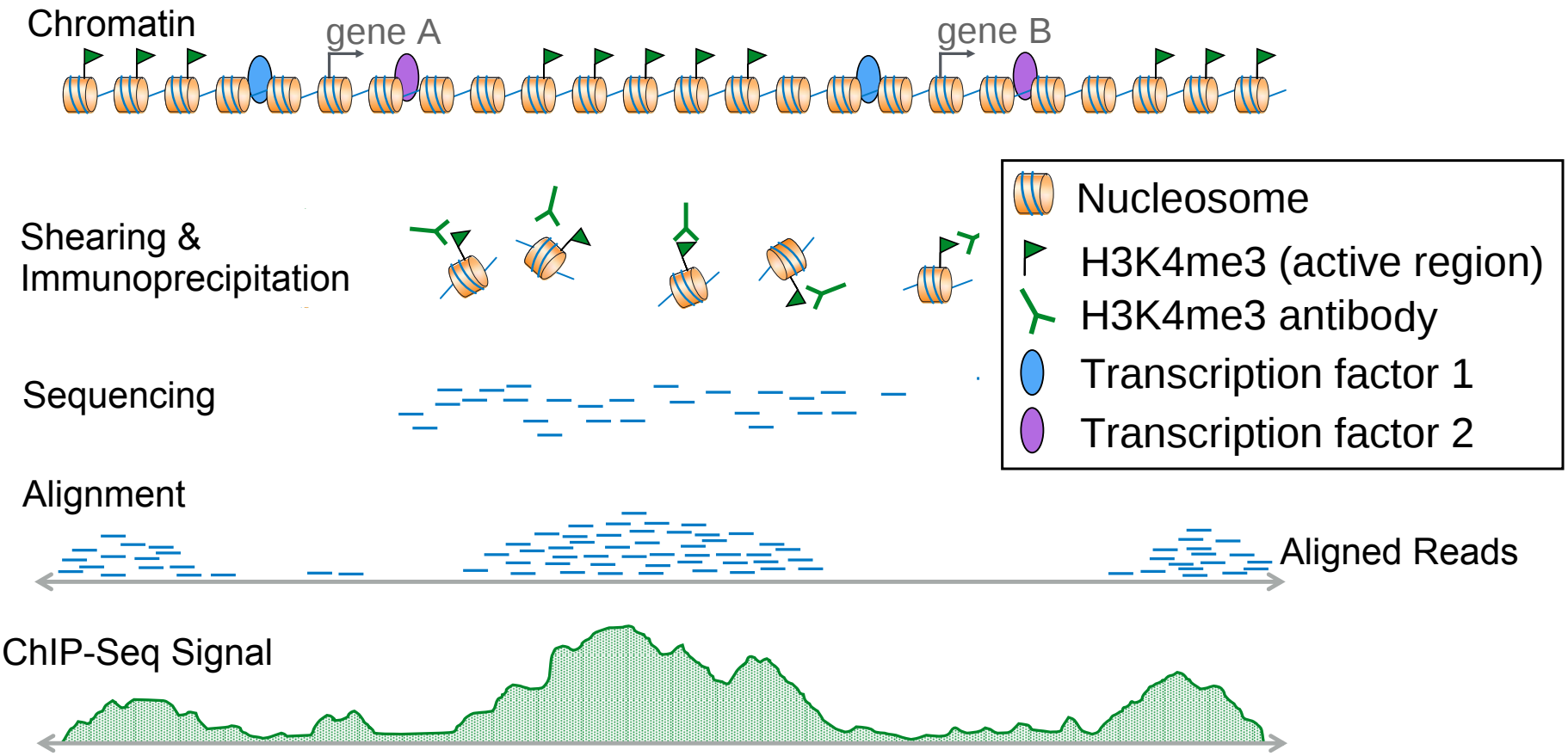
DNA - Protein interactions with ChIP-Seq



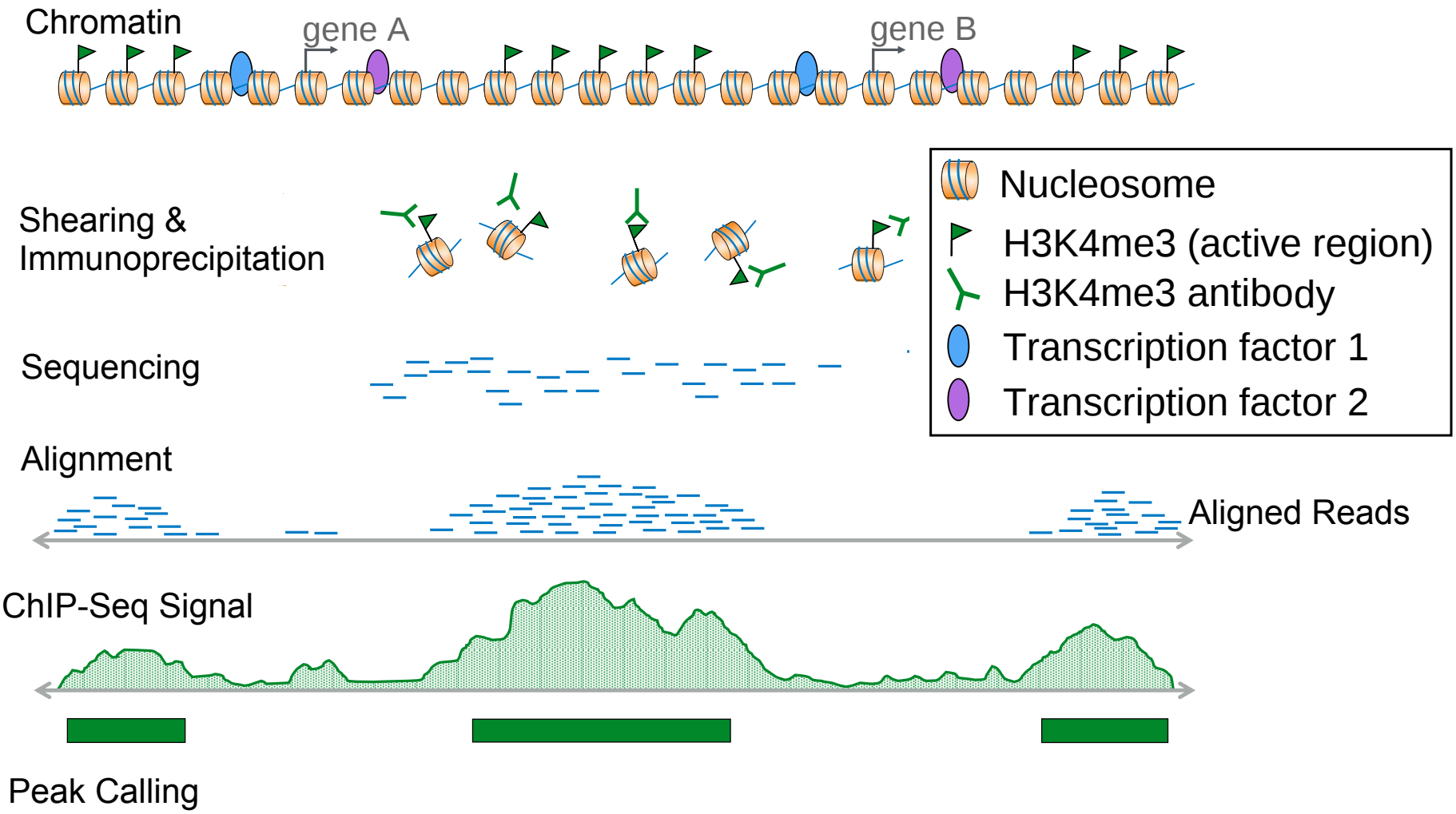
DNA - Protein interactions with ChIP-Seq



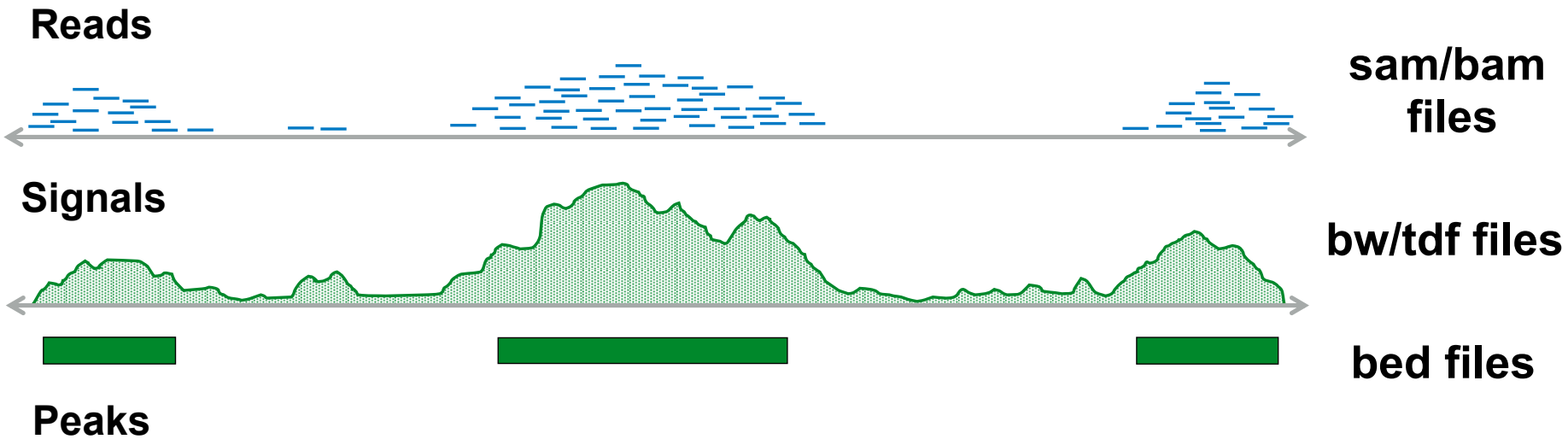
DNA - Protein interactions with ChIP-Seq



DNA - Protein interactions with ChIP-Seq



ChIP-Seq - Data Files



Peaks - Bed files

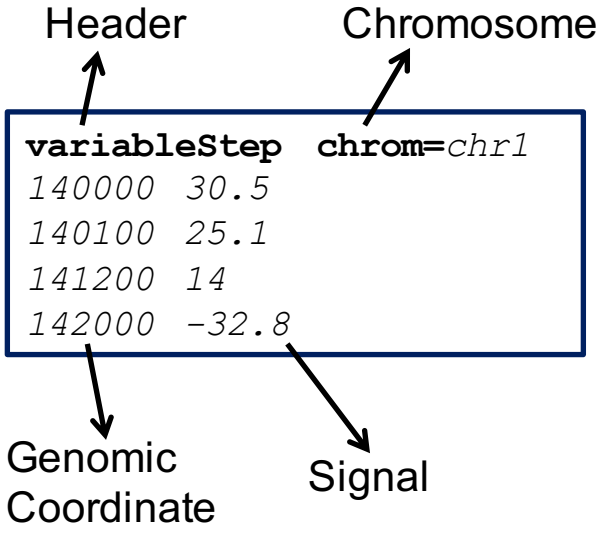
- Peaks / genomic regions are stored in bed files

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

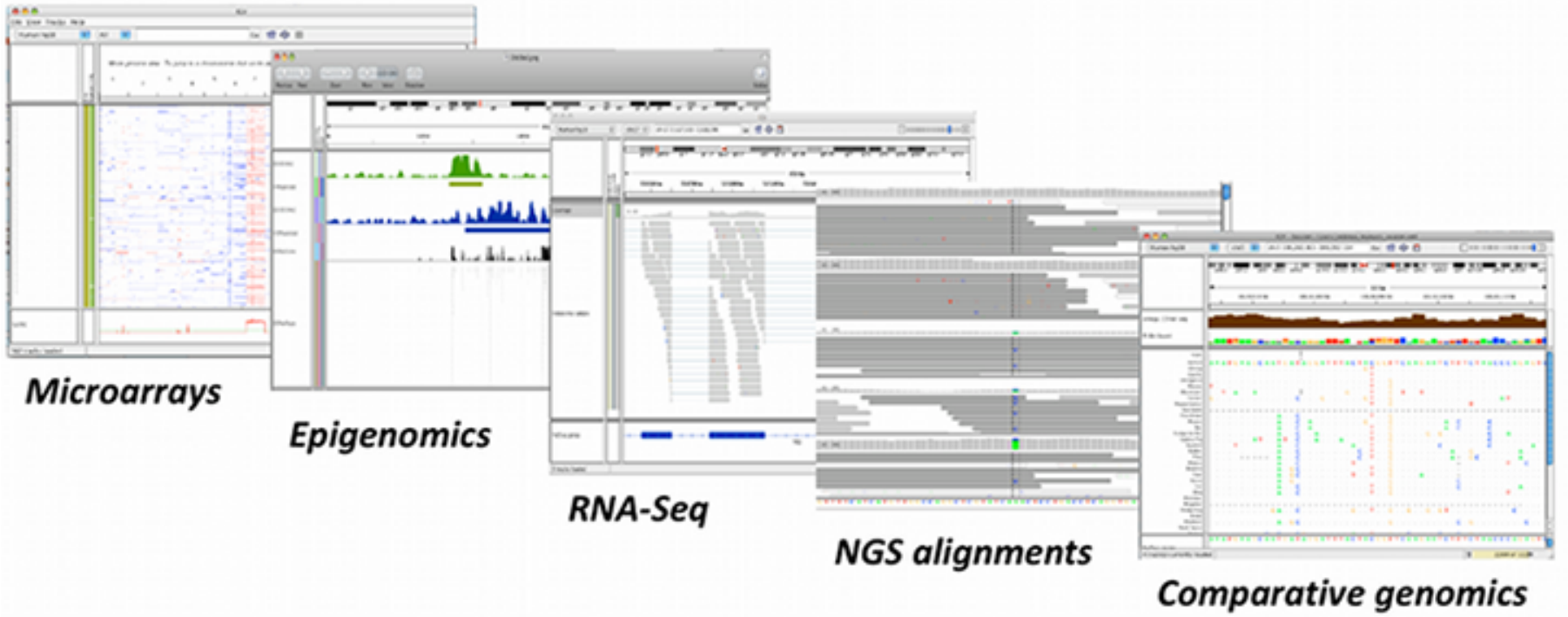
Bioinformatics Analysis in R

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



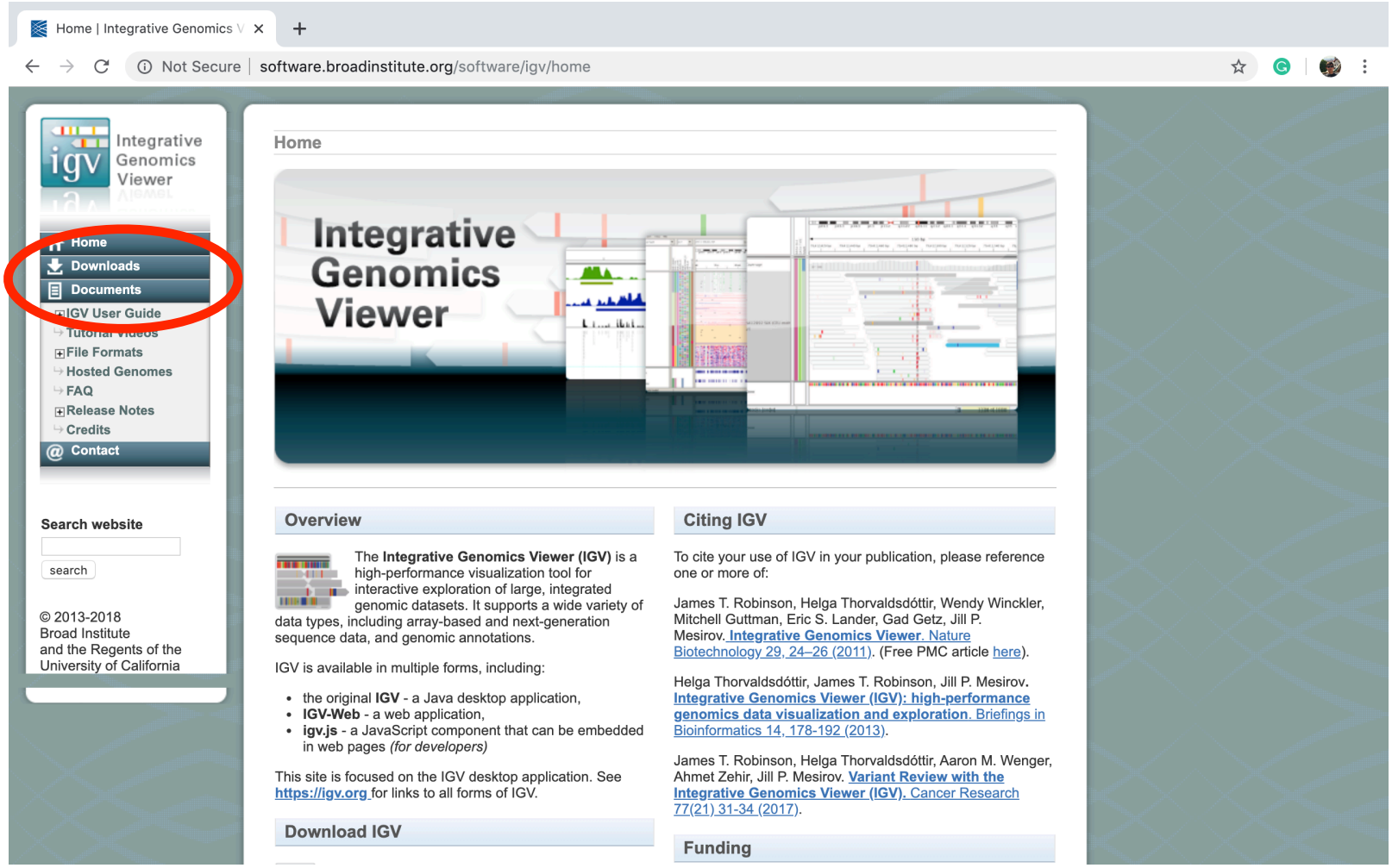
James. T, *et al (2011), Nature Biotechnology*

Advantages

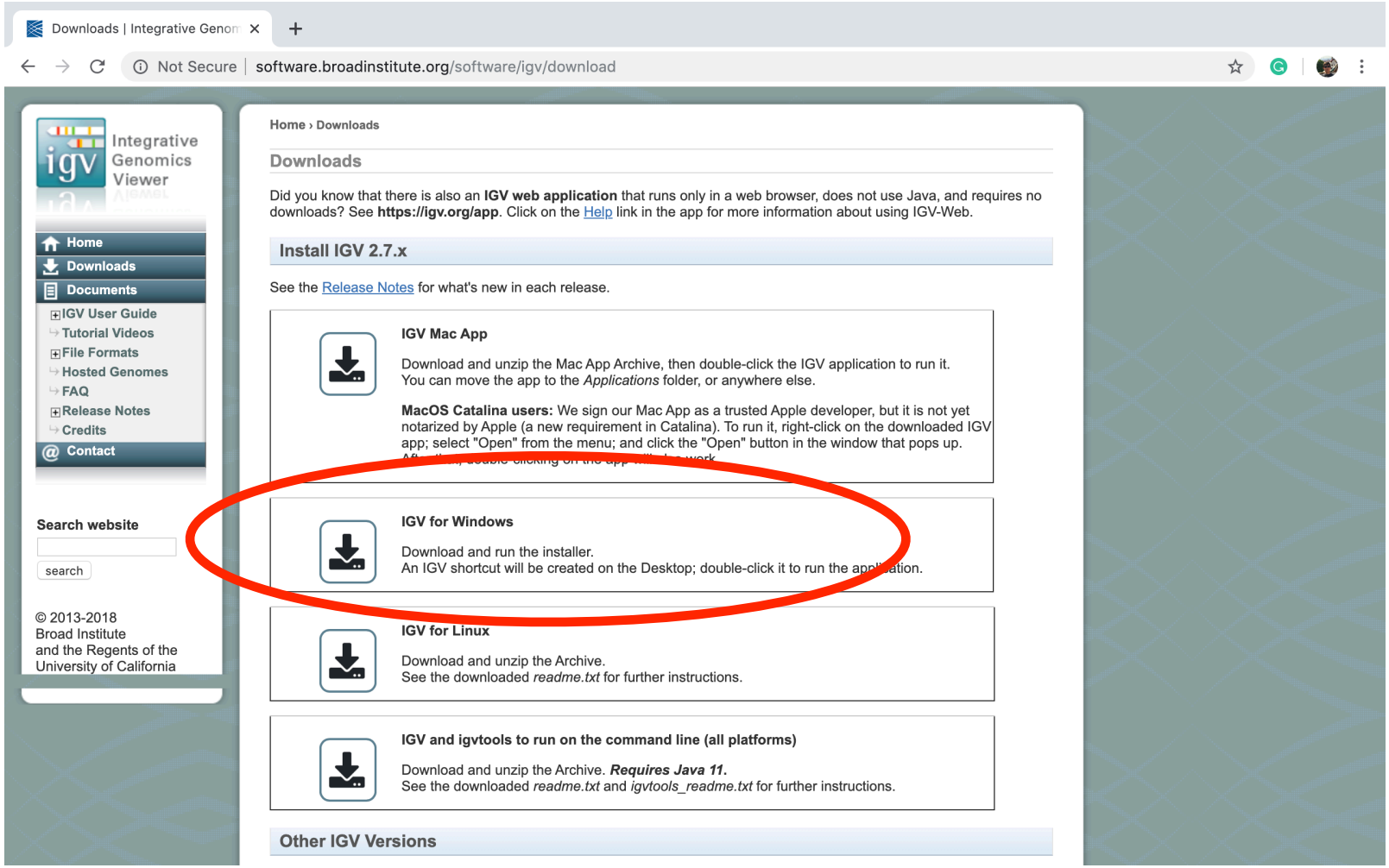
- A high-performance visualization tool
- Allows us interactively explore large, integrated dataset
- Supports a wide variety of data types, including microarray and next-generation sequencing data
- **FREE**

Launch IGV

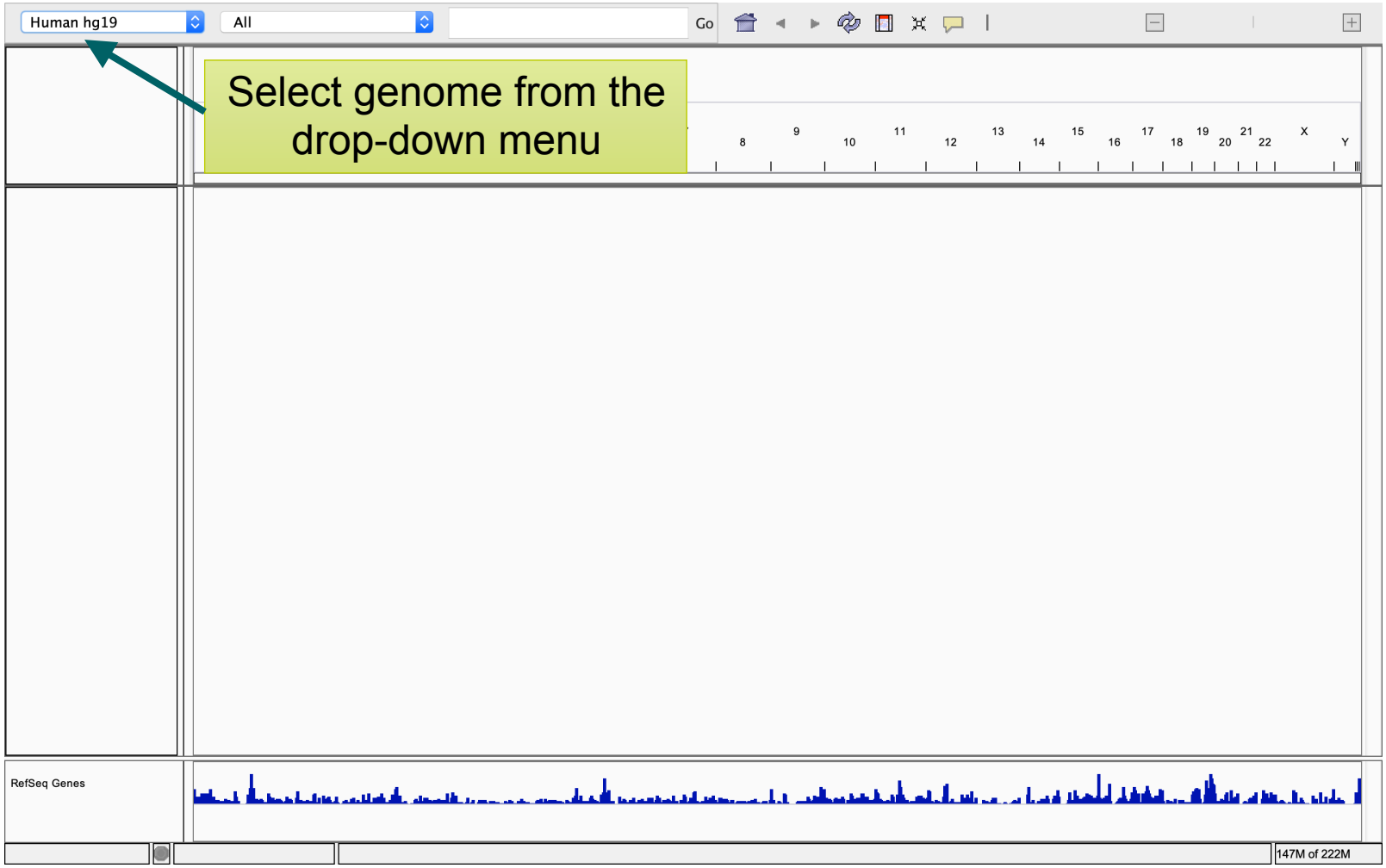
http://software.broadinstitute.org/software/igv/home



Launch IGV


Downloads | Integrative Genom
Not Secure
software.broadinstitute.org/software/igv/download
Integrative Genomics Viewer
Home
Downloads
Documents
IGV User Guide
Tutorial Videos
File Formats
Hosted Genomes
FAQ
Release Notes
Credits
Contact
Search website
search
© 2013-2018
Broad Institute
and the Regents of the
University of California
Home › Downloads
Downloads
Did you know that there is also an IGV web application that runs only in a web browser, does not use Java, and requires no downloads? See https://igv.org/app. Click on the Help link in the app for more information about using IGV-Web.
Install IGV 2.7.x
See the Release Notes for what's new in each release.
IGV Mac App
Download and unzip the Mac App Archive, then double-click the IGV application to run it.
You can move the app to the Applications folder, or anywhere else.
MacOS Catalina users: We sign our Mac App as a trusted Apple developer, but it is not yet notarized by Apple (a new requirement in Catalina). To run it, right-click on the downloaded IGV app; select "Open" from the menu; and click the "Open" button in the window that pops up.
IGV for Windows
Download and run the installer.
An IGV shortcut will be created on the Desktop; double-click it to run the application.
IGV for Linux
Download and unzip the Archive.
See the downloaded readme.txt for further instructions.
IGV and igvtools to run on the command line (all platforms)
Download and unzip the Archive. Requires Java 11.
See the downloaded readme.txt and igvtools_readme.txt for further instructions.
Other IGV Versions

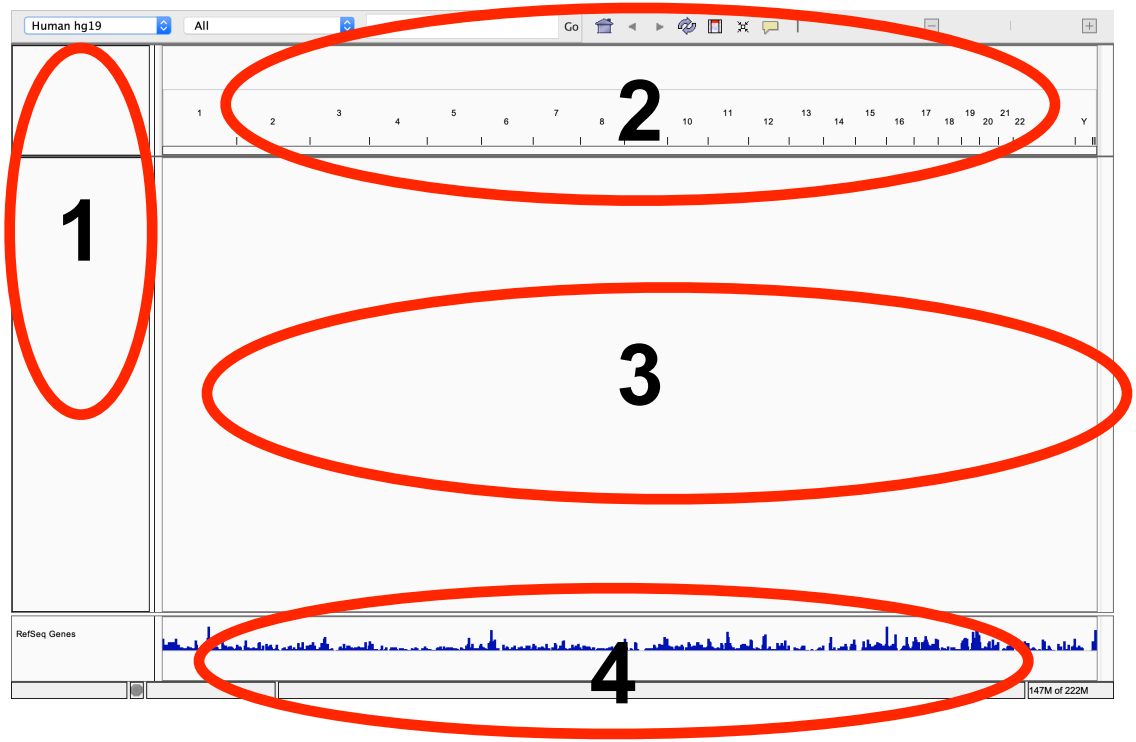
Launch IGV


Human hg19
All
Go
Select genome from the drop-down menu
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
RefSeq Genes
147M of 222M

Launch IGV

Genome navigation

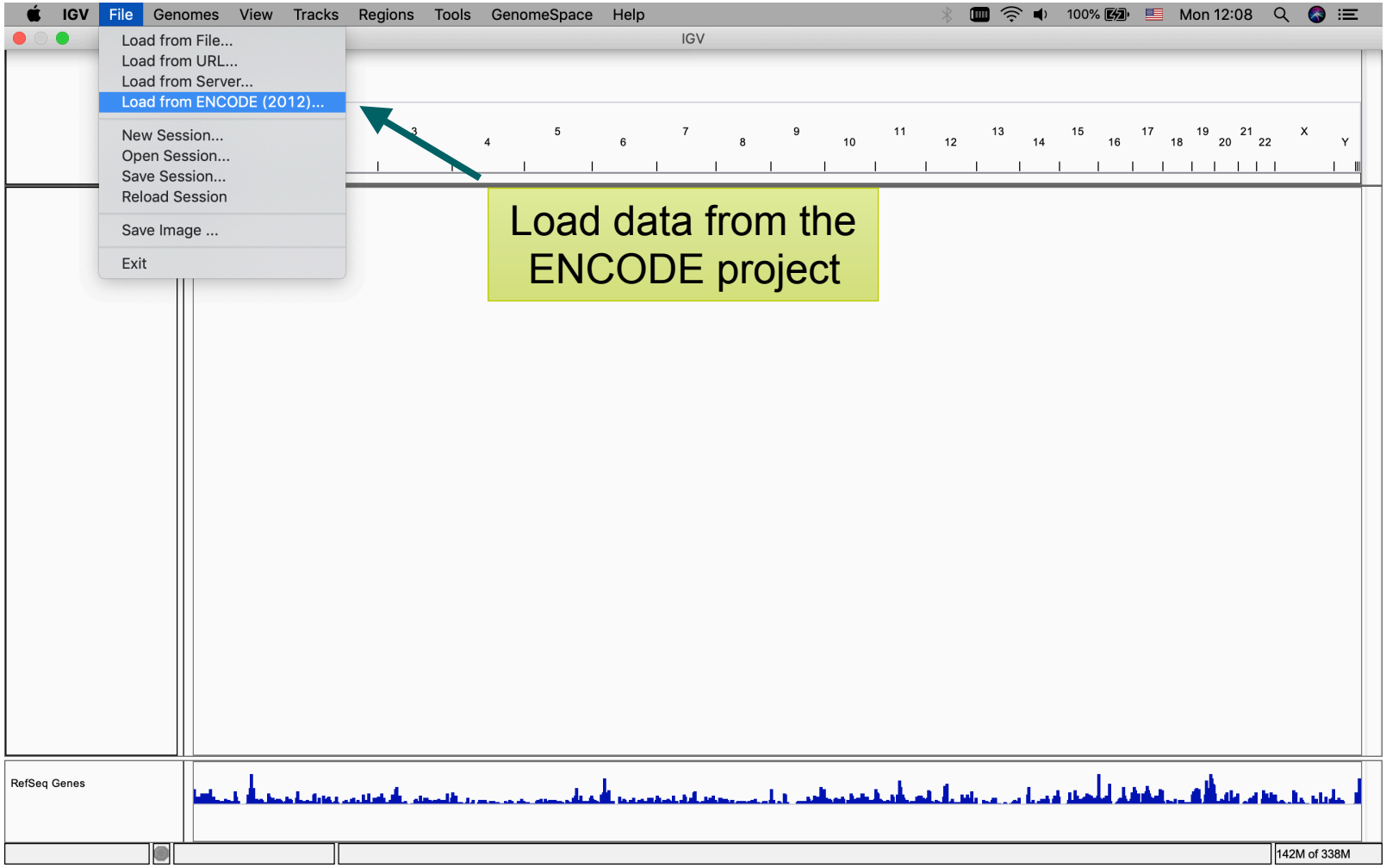
Sample information



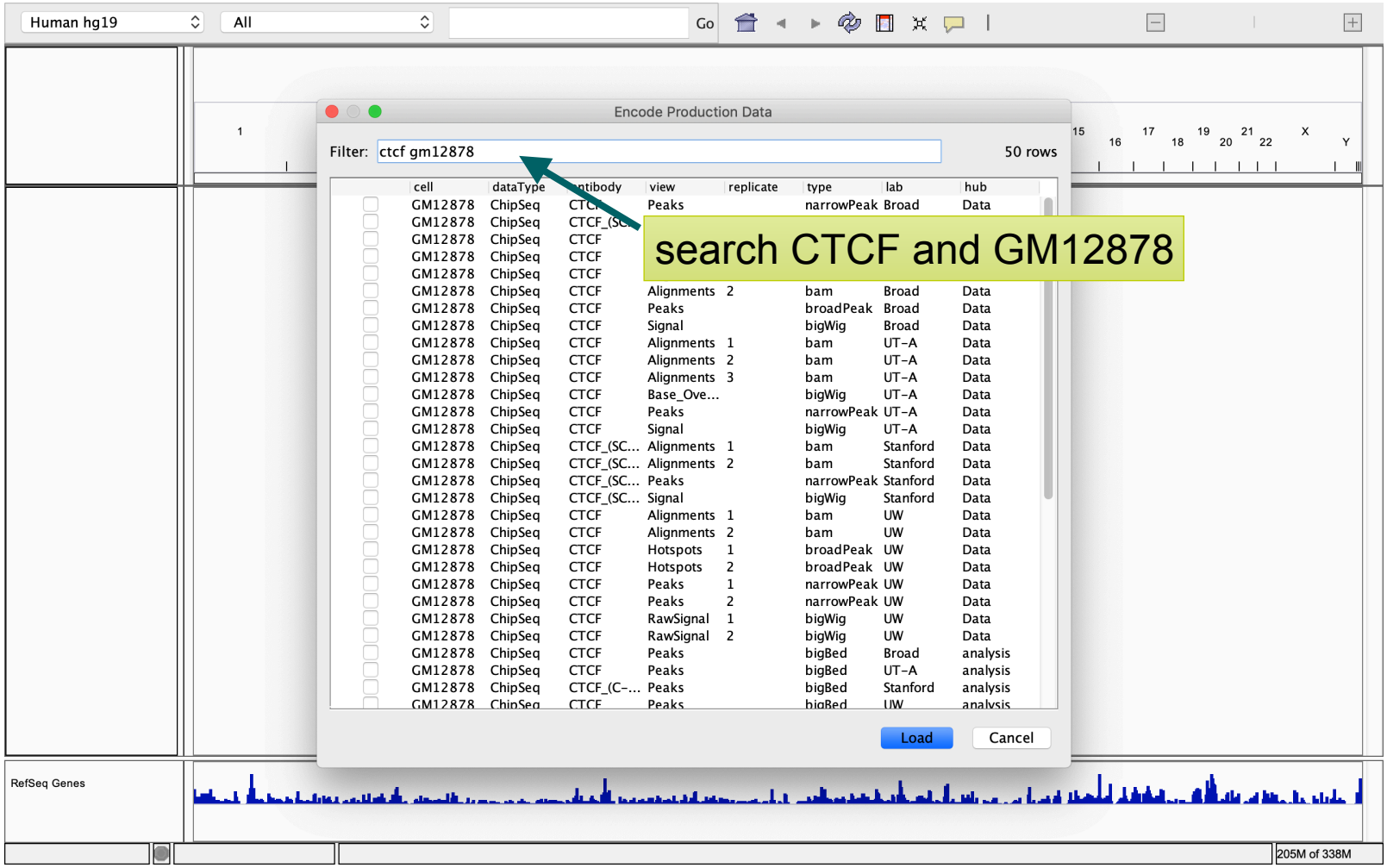
Data

Genome locations and sequence

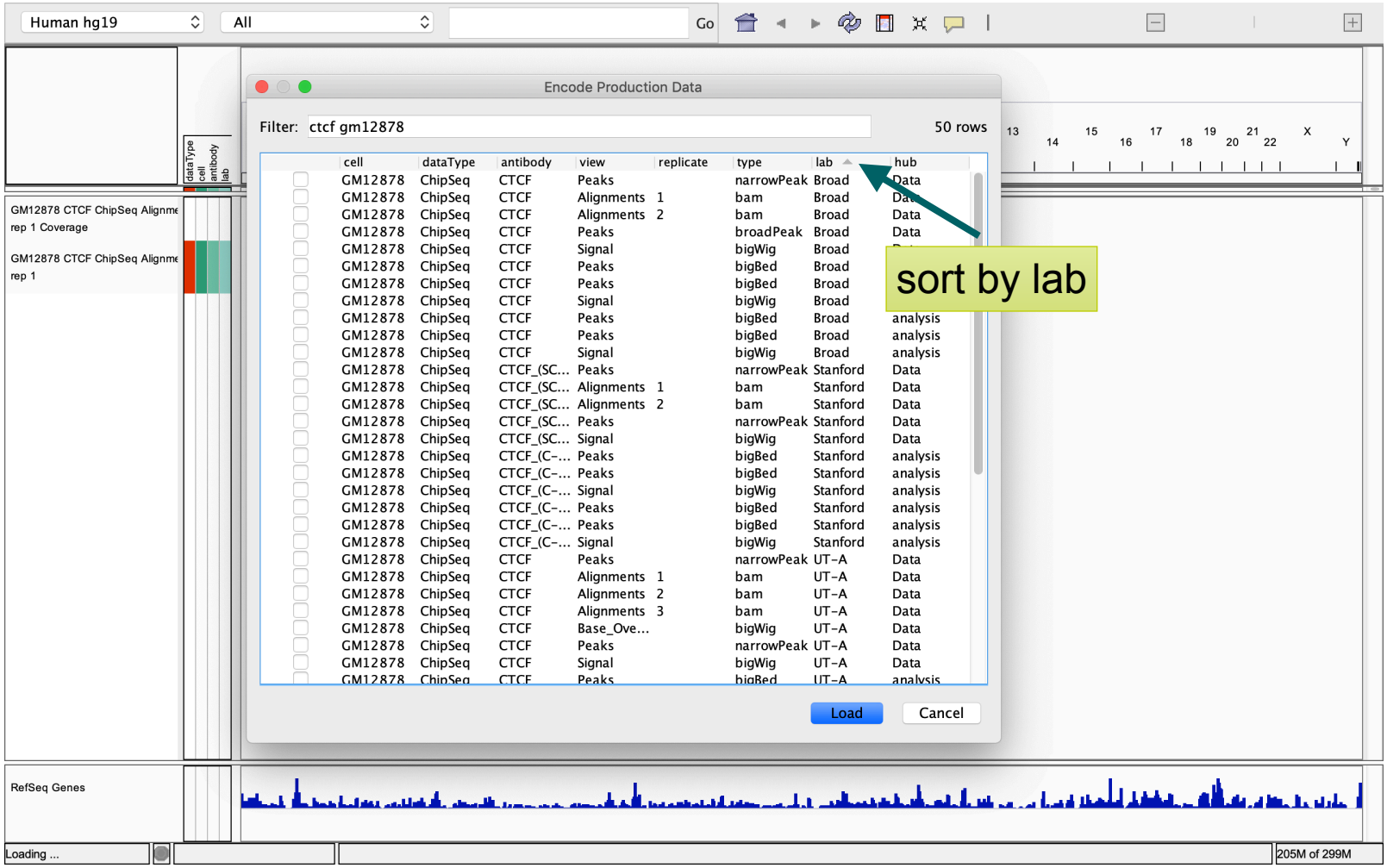
Viewing NGS Data



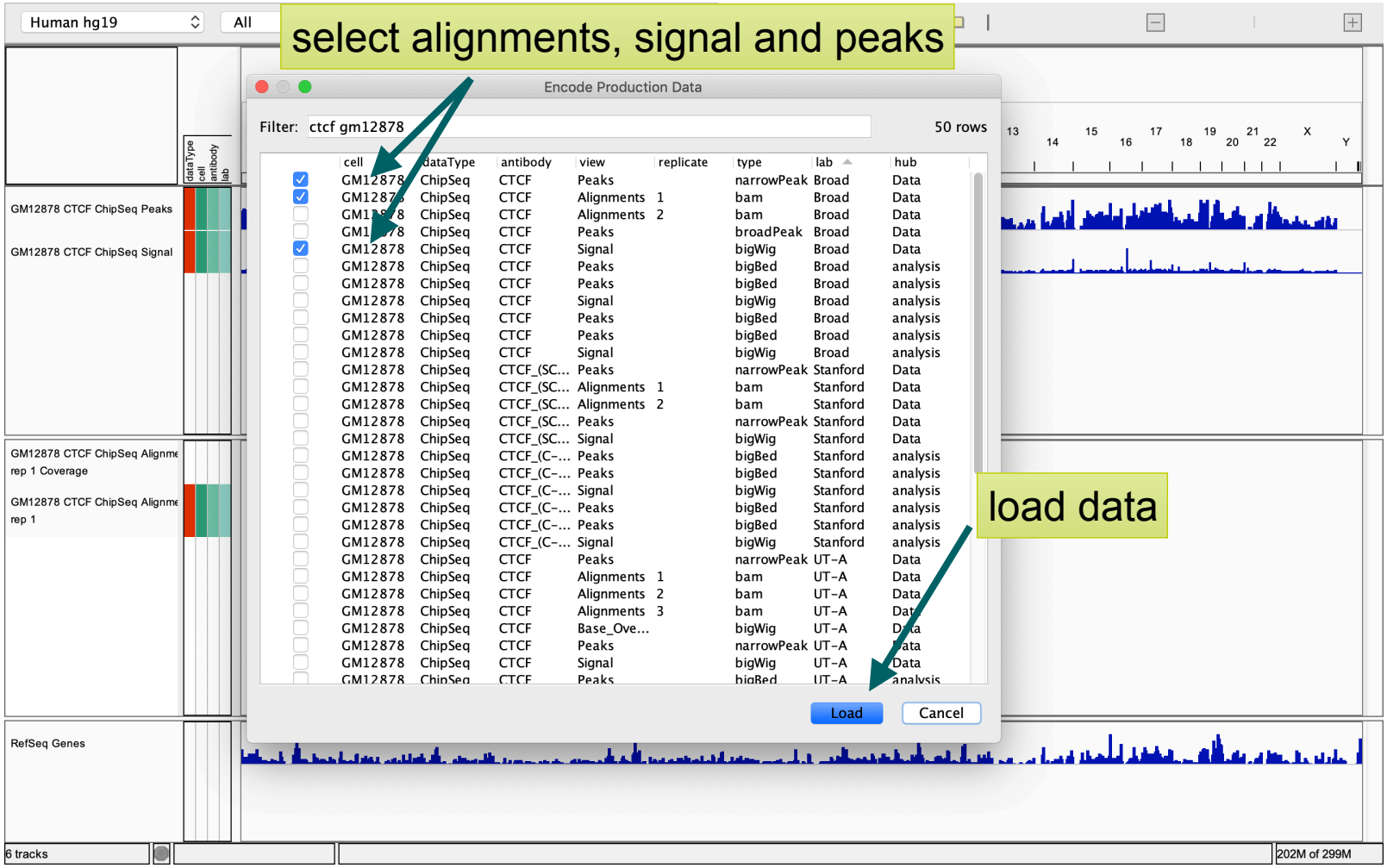
Viewing NGS Data



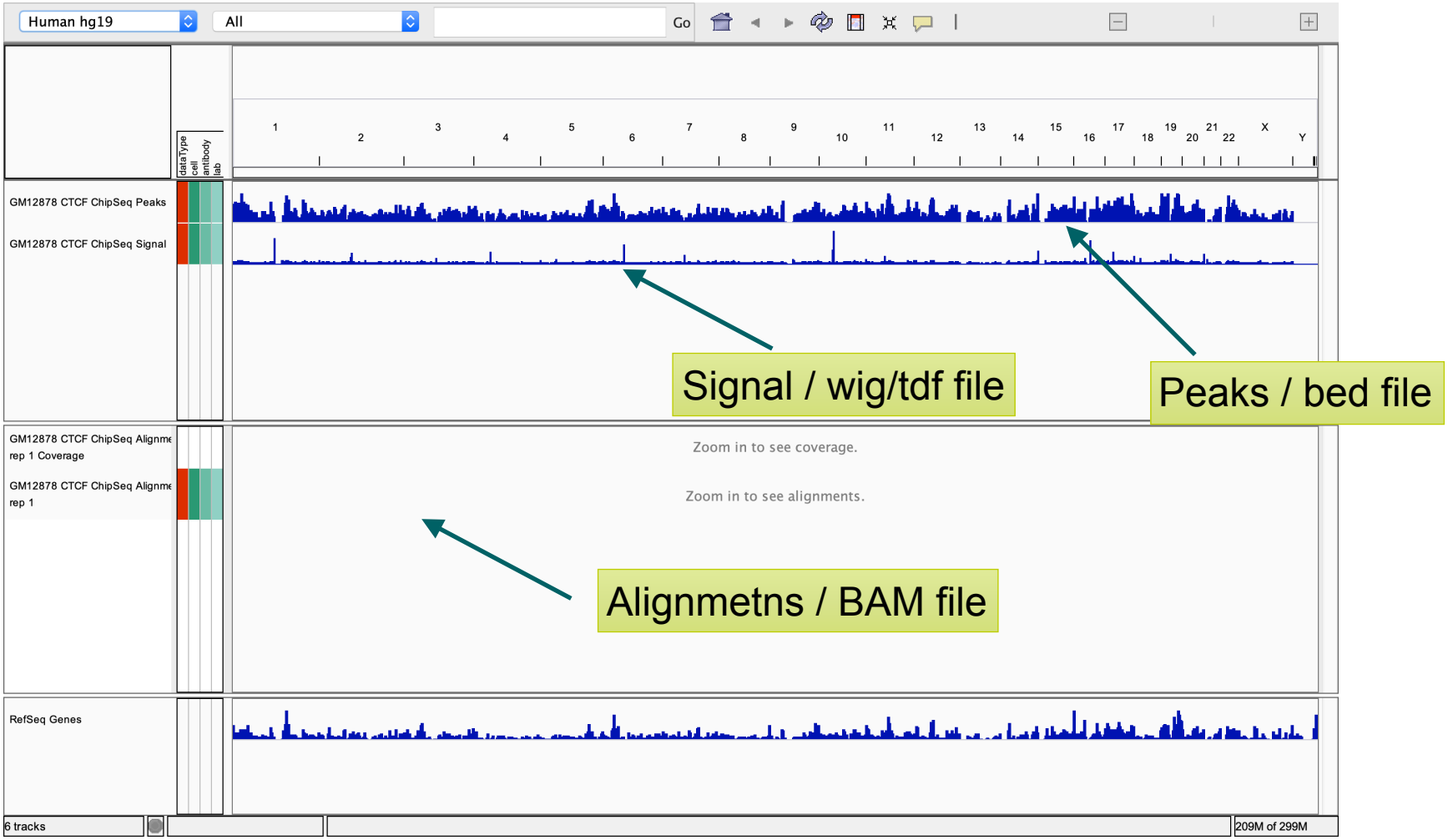
Viewing NGS Data



Viewing NGS Data



Viewing NGS Data



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

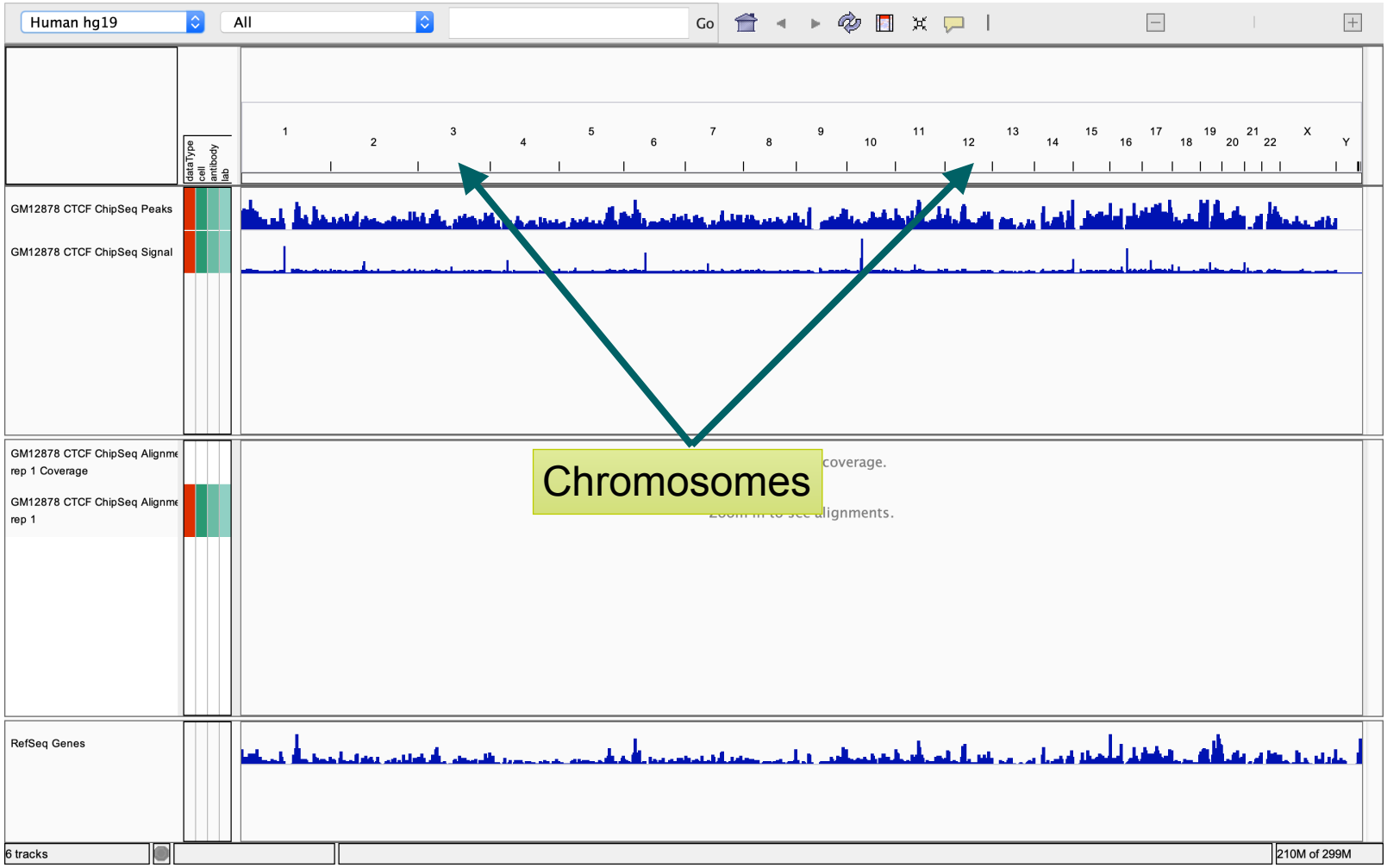
File Formats

Note: for large files use indexed formats

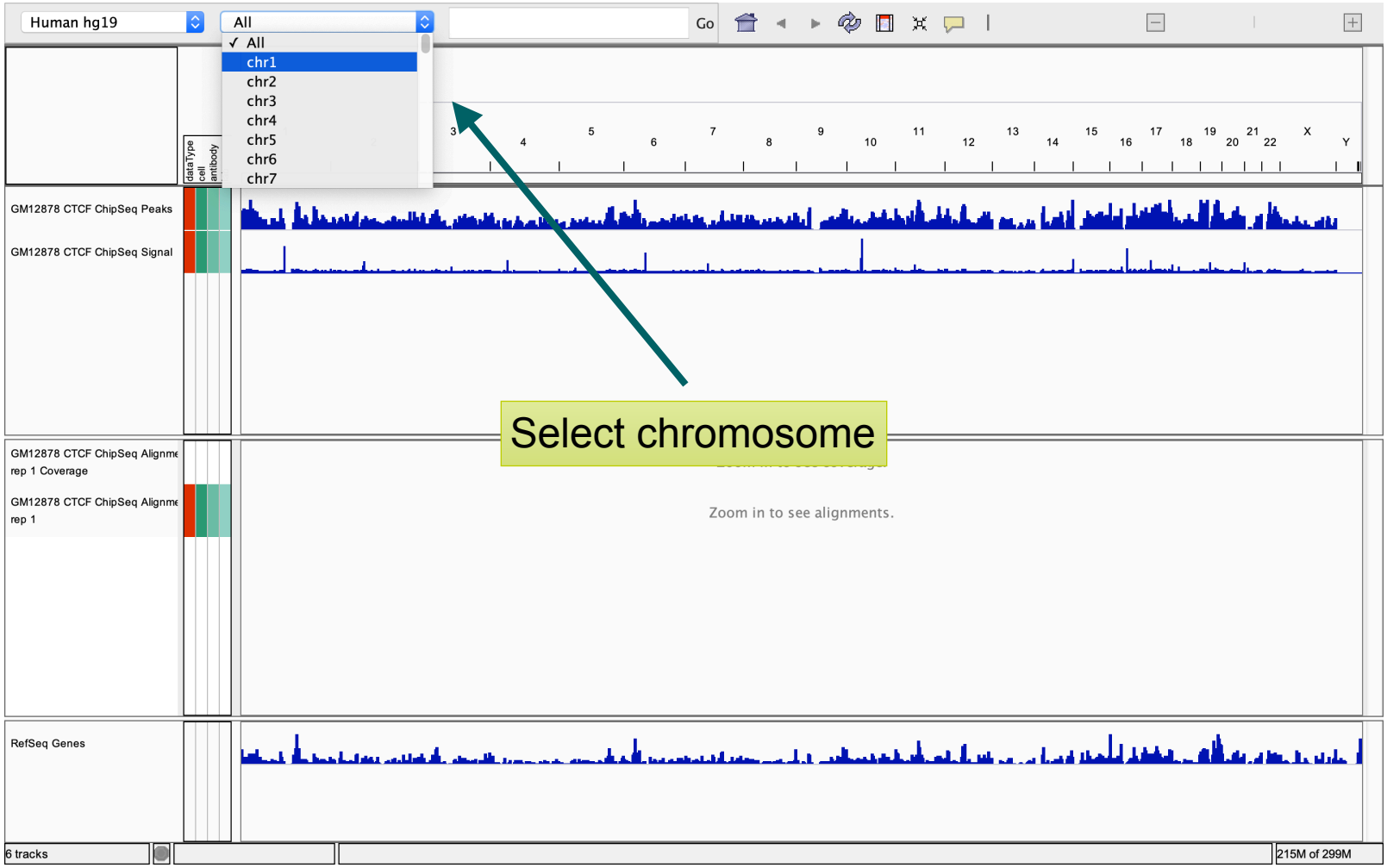
Hands on

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

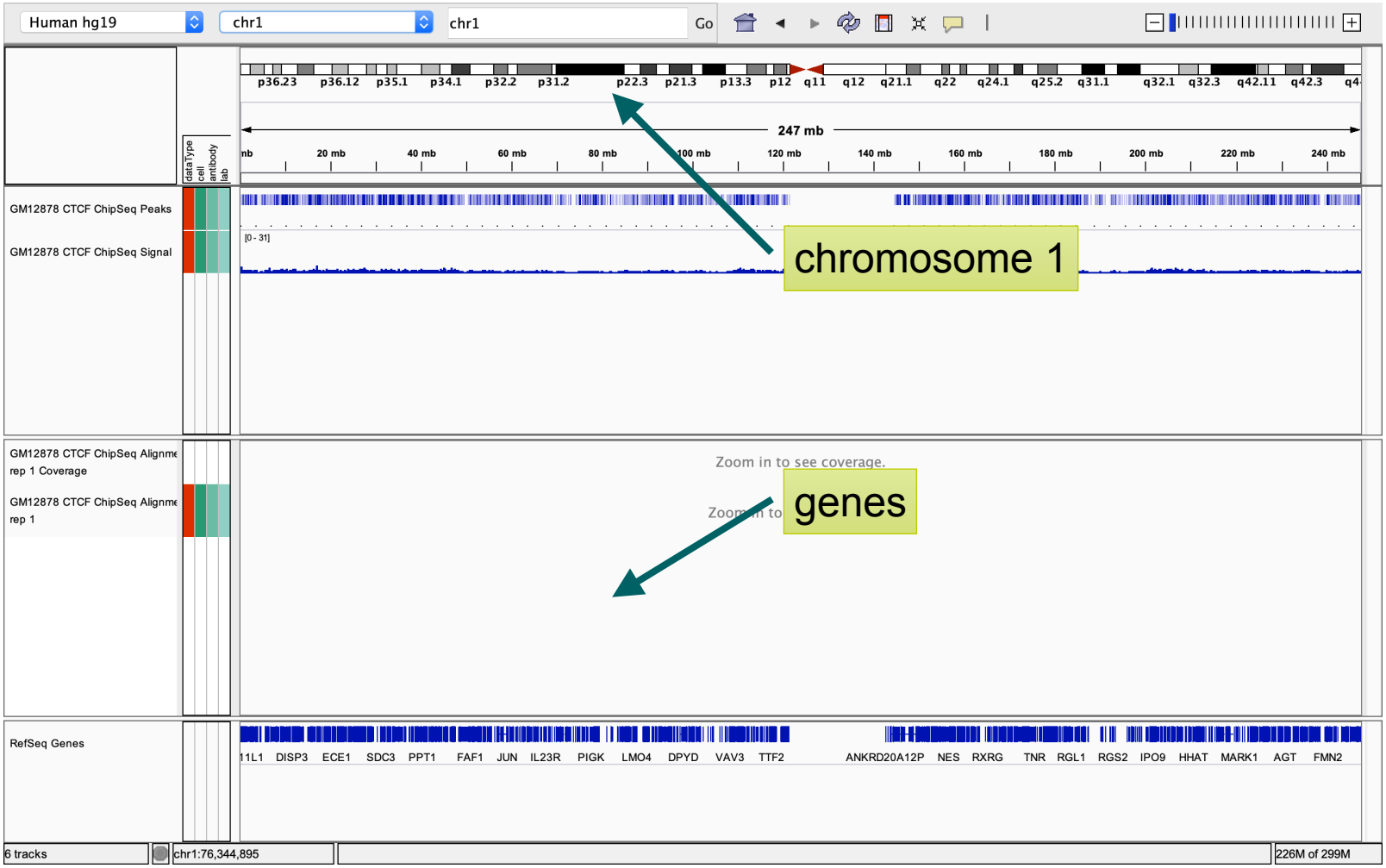
Navigation


Human hg19
All
Go
dataType
cell
antibody
lab
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
GM12878 CTCF ChipSeq Alignme
rep 1 Coverage
GM12878 CTCF ChipSeq Alignme
rep 1
Chromosomes
coverage.
lignments.
RefSeq Genes
6 tracks
210M of 299M

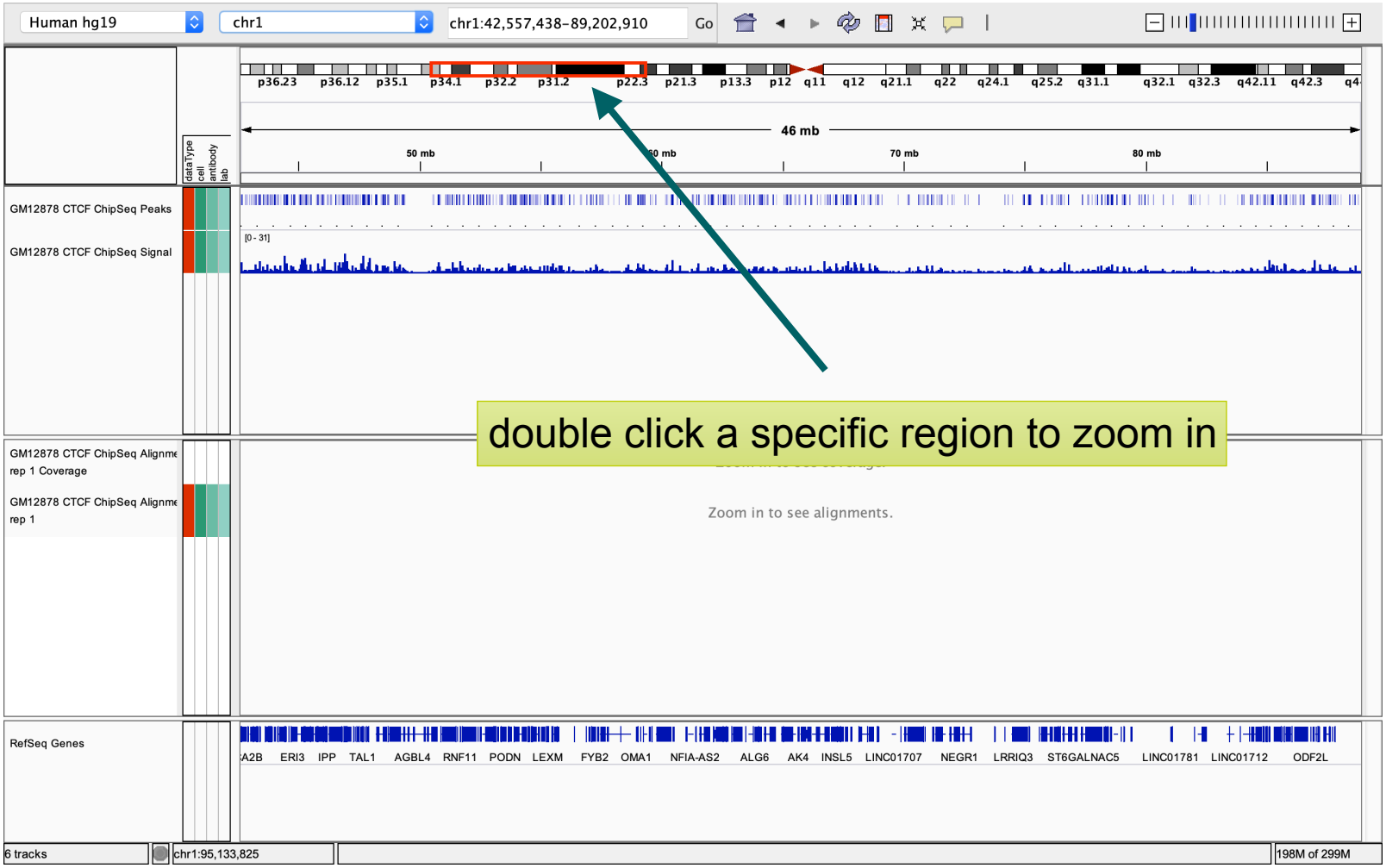
Navigation


Human hg19
All
Go
✓ All
chr1
chr2
chr3
chr4
chr5
chr6
chr7
dataType
cell
antibody
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
GM12878 CTCF ChipSeq Alignme
rep 1 Coverage
GM12878 CTCF ChipSeq Alignme
rep 1
Zoom in to see alignments.
RefSeq Genes
6 tracks
215M of 299M
Select chromosome

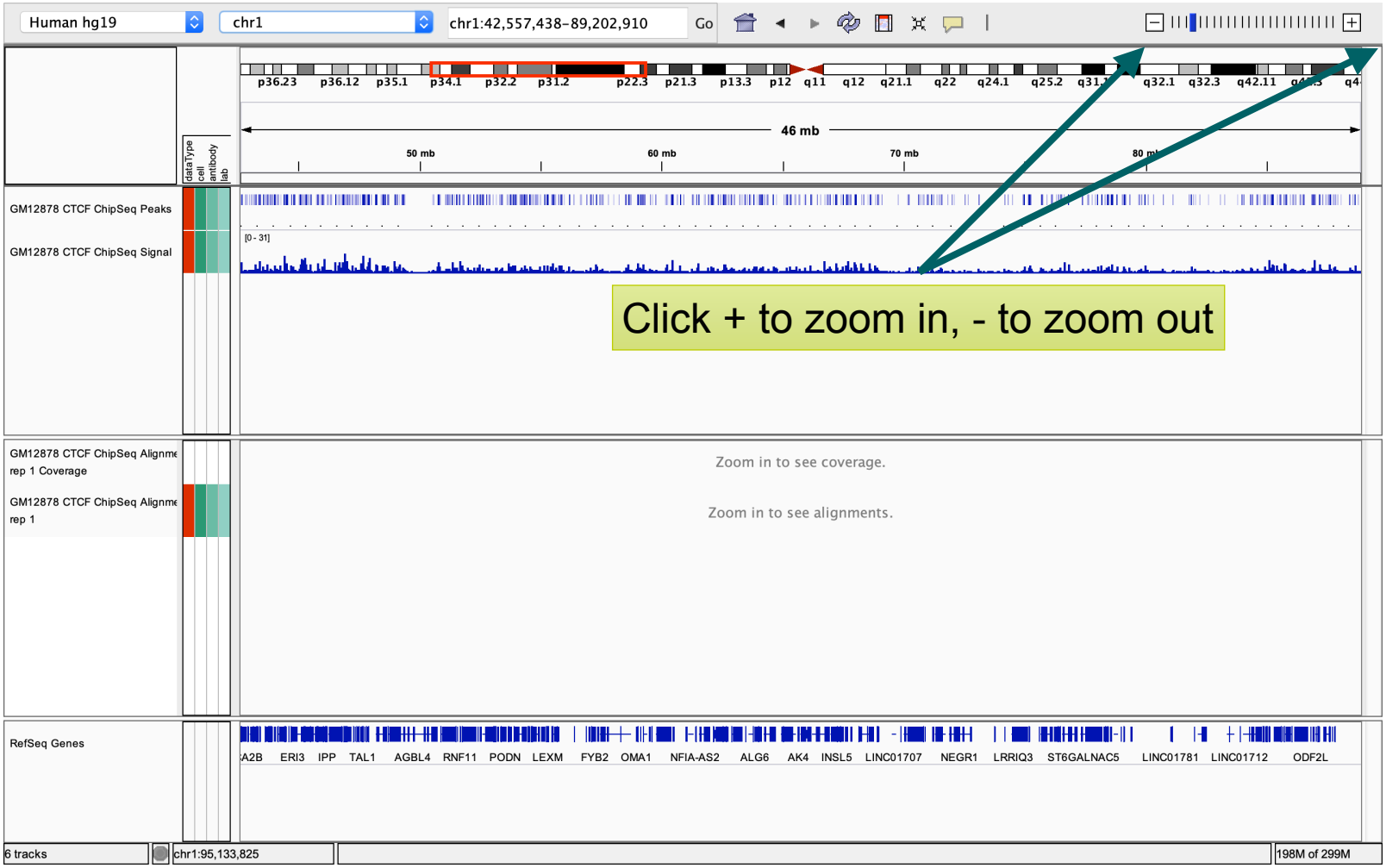
Navigation


Human hg19
chr1
chr1
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3
247 mb
20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb 160 mb 180 mb 200 mb 220 mb 240 mb
dataType cell antibody lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 31]
chromosome 1
GM12878 CTCF ChipSeq Alignme rep 1 Coverage
GM12878 CTCF ChipSeq Alignme rep 1
Zoom in to see coverage.
genes
RefSeq Genes
11L1 DISP3 ECE1 SDC3 PPT1 FAF1 JUN IL23R PIGK LMO4 DPYD VAV3 TTF2 ANKRD20A12P NES RXRG TNR RGL1 RGS2 IPO9 HHAT MARK1 AGT FMN2
6 tracks
chr1:76,344,895
226M of 299M

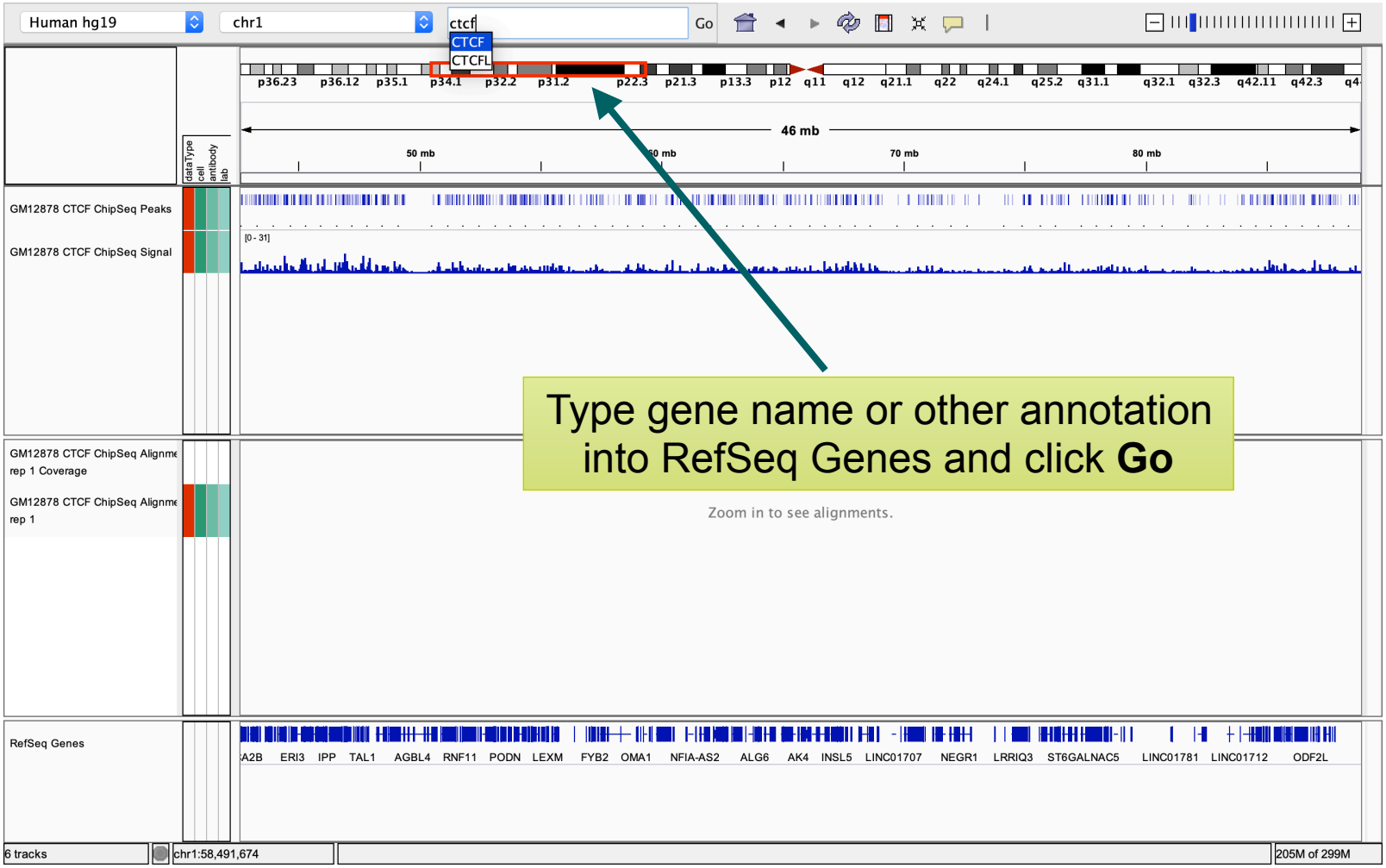
Navigation


Human hg19
chr1
chr1:42,557,438-89,202,910
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3
46 mb
50 mb
60 mb
70 mb
80 mb
dataType
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 31]
GM12878 CTCF ChipSeq Alignme
rep 1 Coverage
GM12878 CTCF ChipSeq Alignme
rep 1
Zoom in to see alignments.
RefSeq Genes
A2B ERI3 IPP TAL1 AGBL4 RNF11 PODN LEXM FYB2 OMA1 NFIA-AS2 ALG6 AK4 INSL5 LINC01707 NEGR1 LRRIQ3 ST6GALNAC5 LINC01781 LINC01712 ODF2L
6 tracks
chr1:95,133,825
198M of 299M
double click a specific region to zoom in

Navigation

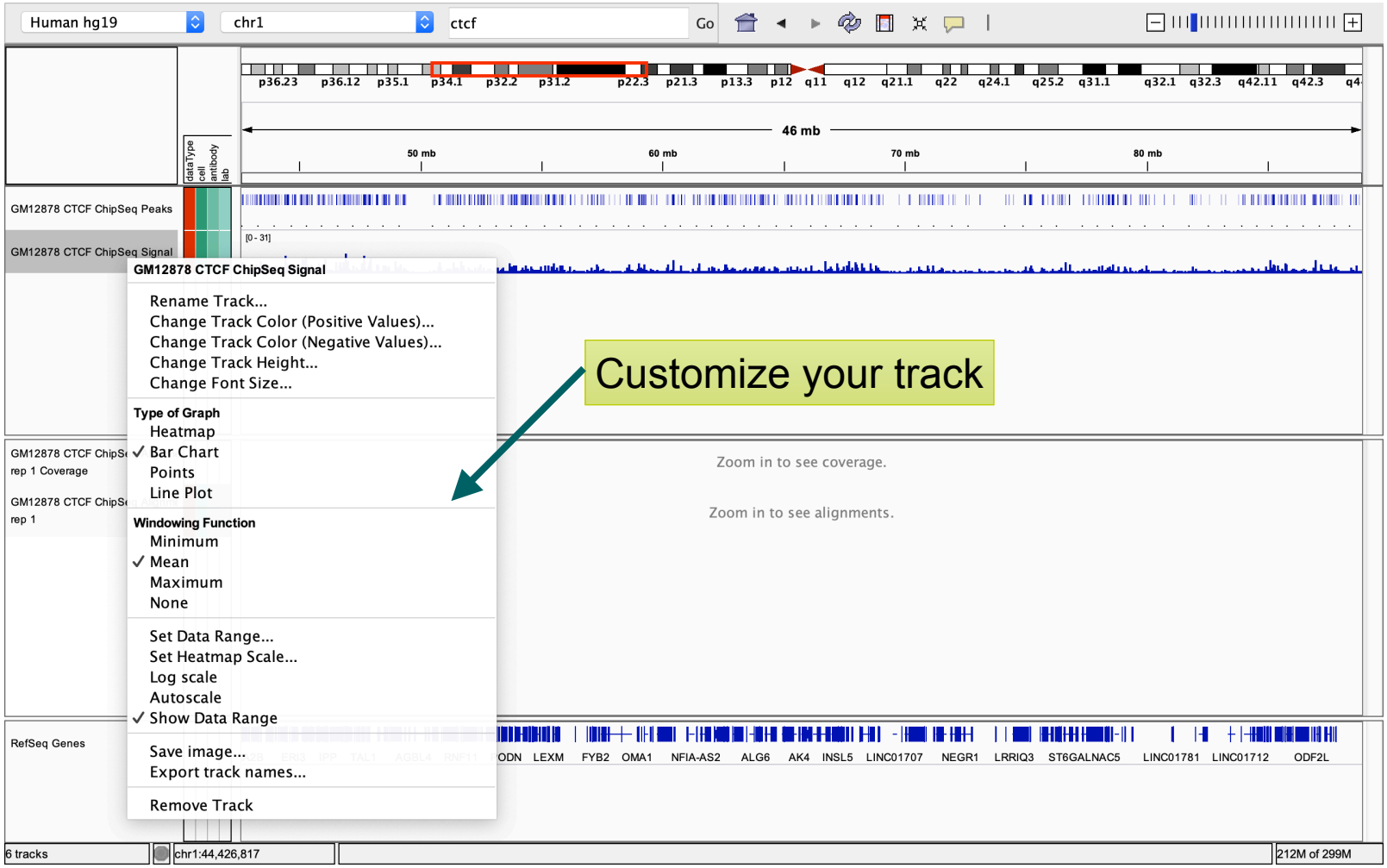

Human hg19
chr1
chr1:42,557,438-89,202,910
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q22 q24.1 q25.2 q32.1 q32.3 q42.11
46 mb
50 mb
60 mb
70 mb
80 mb
dataType
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 31]
Click + to zoom in, - to zoom out
GM12878 CTCF ChipSeq Alignme rep 1 Coverage
GM12878 CTCF ChipSeq Alignme rep 1
Zoom in to see coverage.
Zoom in to see alignments.
RefSeq Genes
A2B ERI3 IPP TAL1 AGBL4 RNF11 PODN LEXM FYB2 OMA1 NFIA-AS2 ALG6 AK4 INSL5 LINC01707 NEGR1 LRRIQ3 ST6GALNAC5 LINC01781 LINC01712 ODF2L
6 tracks
chr1:95,133,825
198M of 299M

Navigation

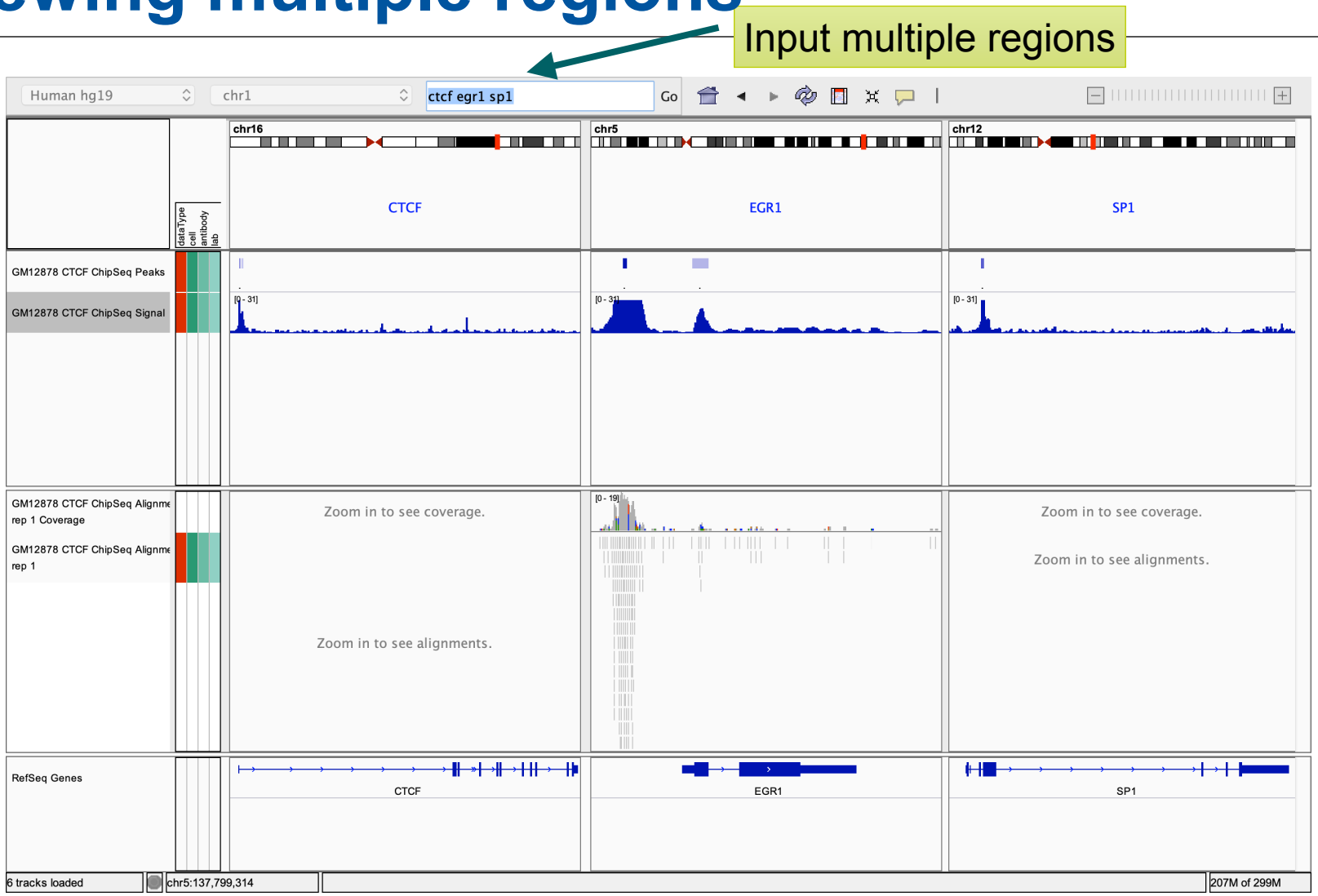


Type gene name or other annotation into RefSeq Genes and click **Go**

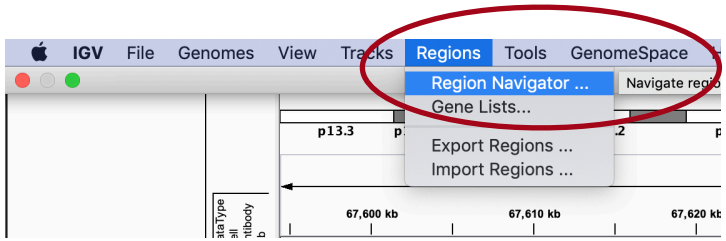
Navigation


Human hg19
chr1
ctcf
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3
46 mb
50 mb
60 mb
70 mb
80 mb
dataType
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 31]
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
Customize your track
GM12878 CTCF ChipSeq rep 1 Coverage
GM12878 CTCF ChipSeq rep 1
Zoom in to see coverage.
Zoom in to see alignments.
RefSeq Genes
ODN LEXM FYB2 OMA1 NFIA-AS2 ALG6 AK4 INSL5 LINC01707 NEGR1 LRRIQ3 ST6GALNAC5 LINC01781 LINC01712 ODF2L
6 tracks
chr1:44,426,817
212M of 299M

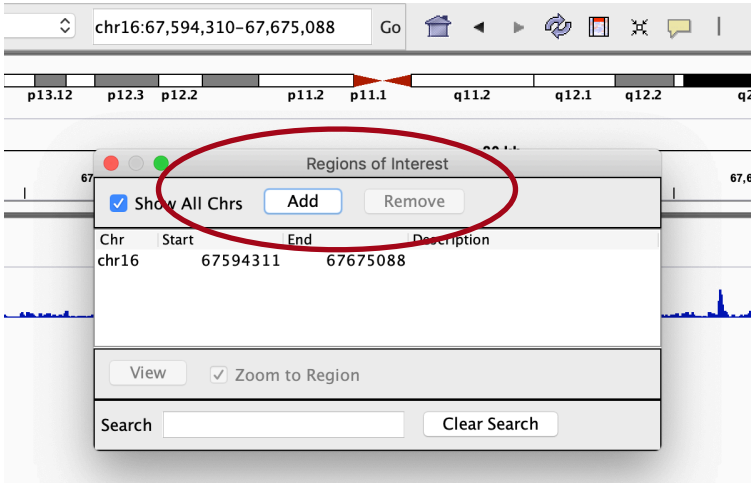
Viewing multiple regions



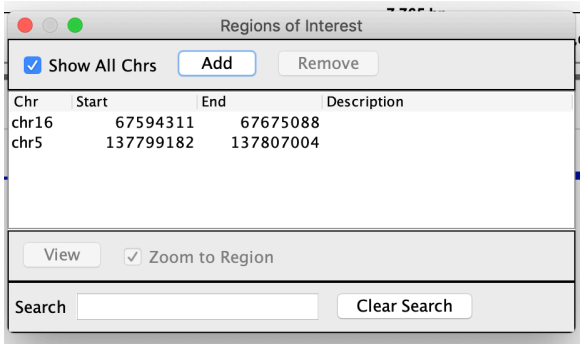
Viewing multiple regions



Click **Regions >Region Navigator**

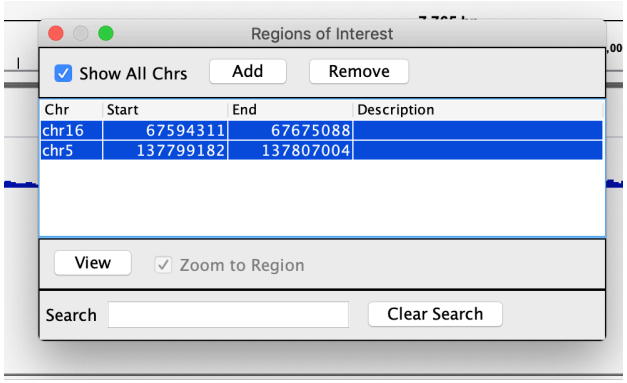


Click **add**



Repeat above step to add more regions

Viewing multiple regions



Select two or more regions
click **View**

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

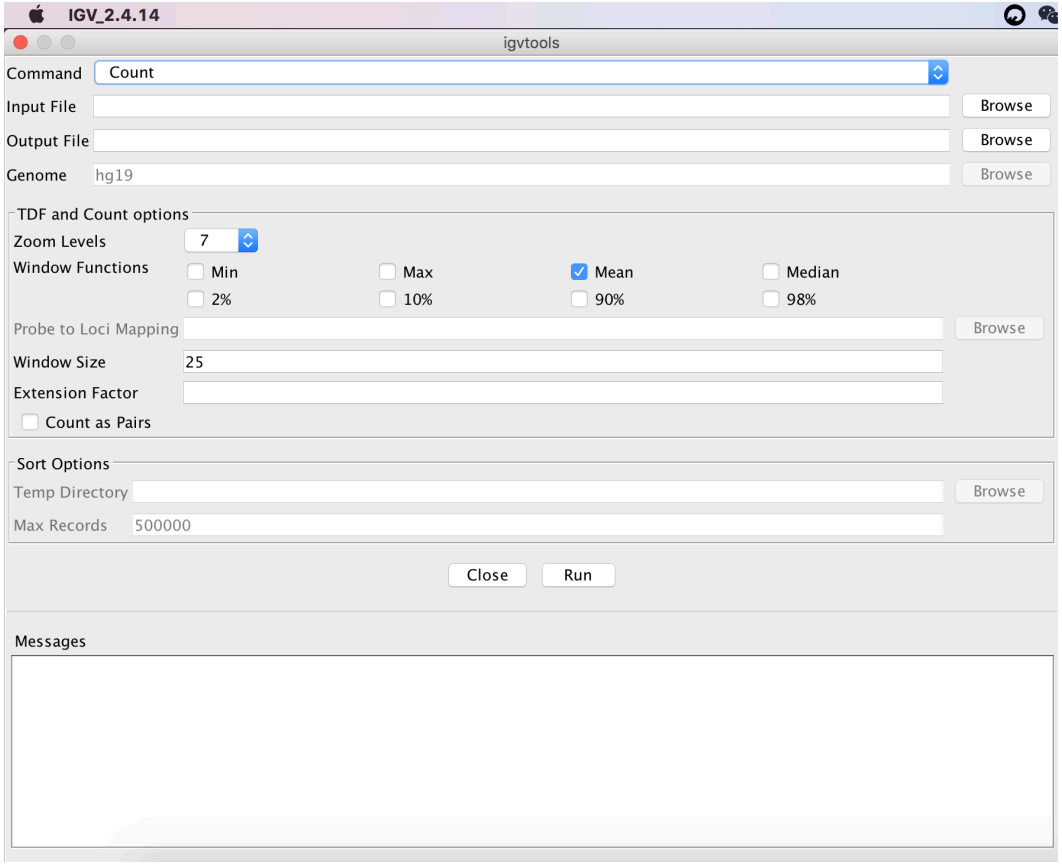
A set of utilities for preparing files for efficient display

- 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

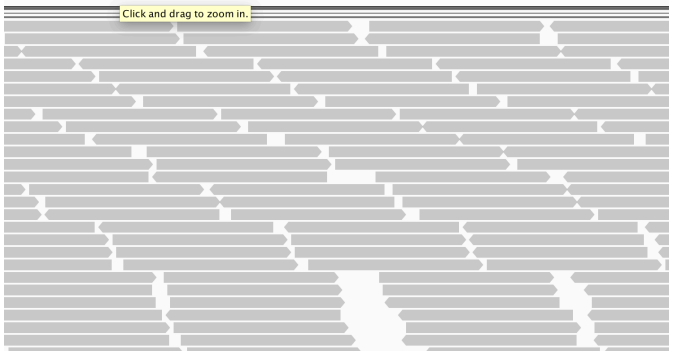
The **toTDF** utility converts large data files into tiled data format (.tdf) files

TDF files have the following advantages:

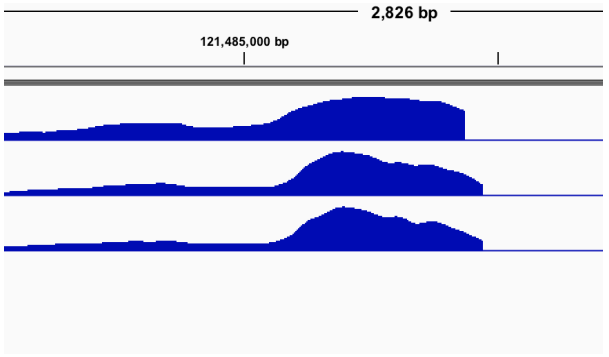
- Data is indexed for efficient retrieval
- Data is preprocessed for zoomed out views
- TDF files are web friendly - large data can be shared over the web.

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



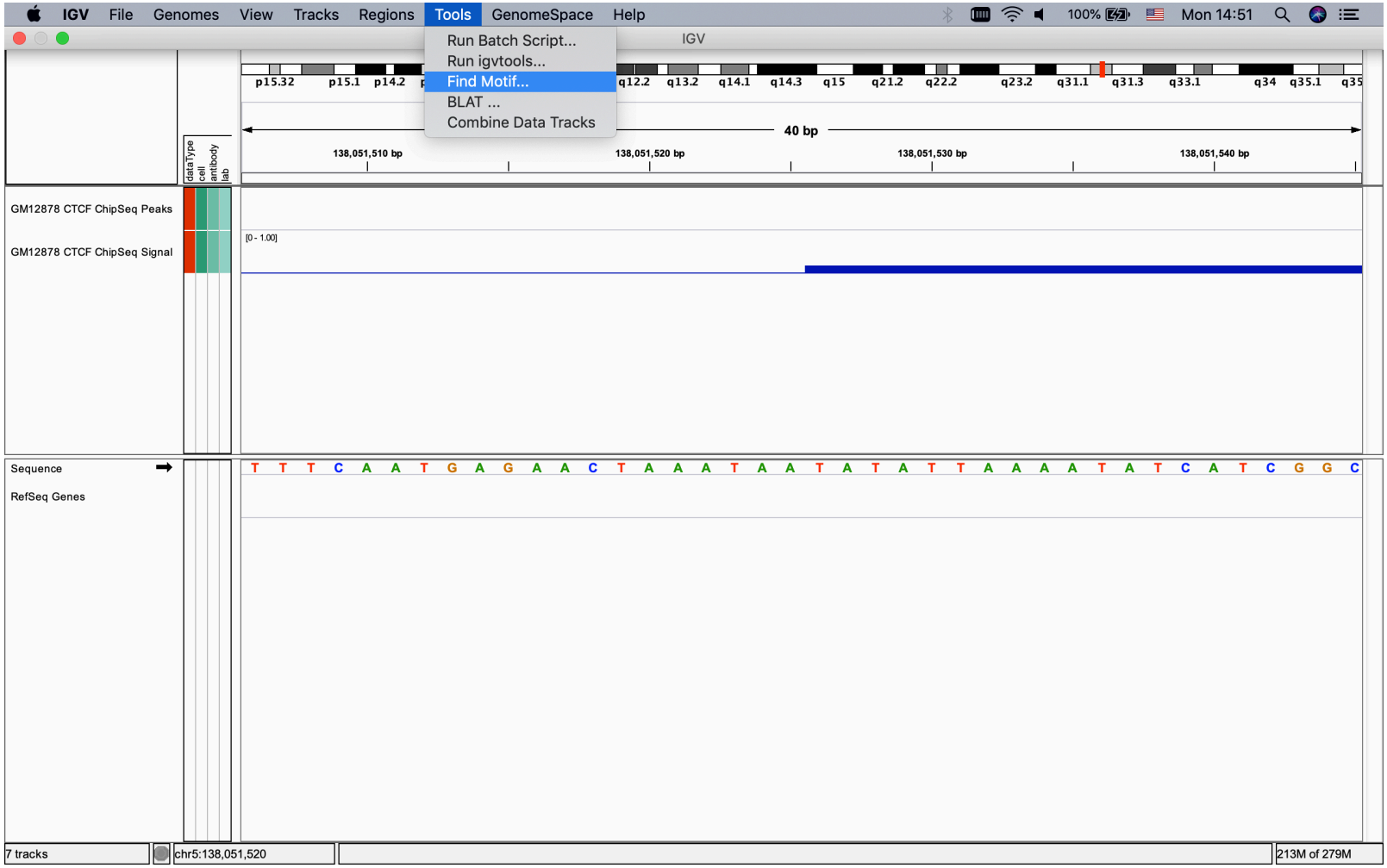
count



Alignment

Read Density

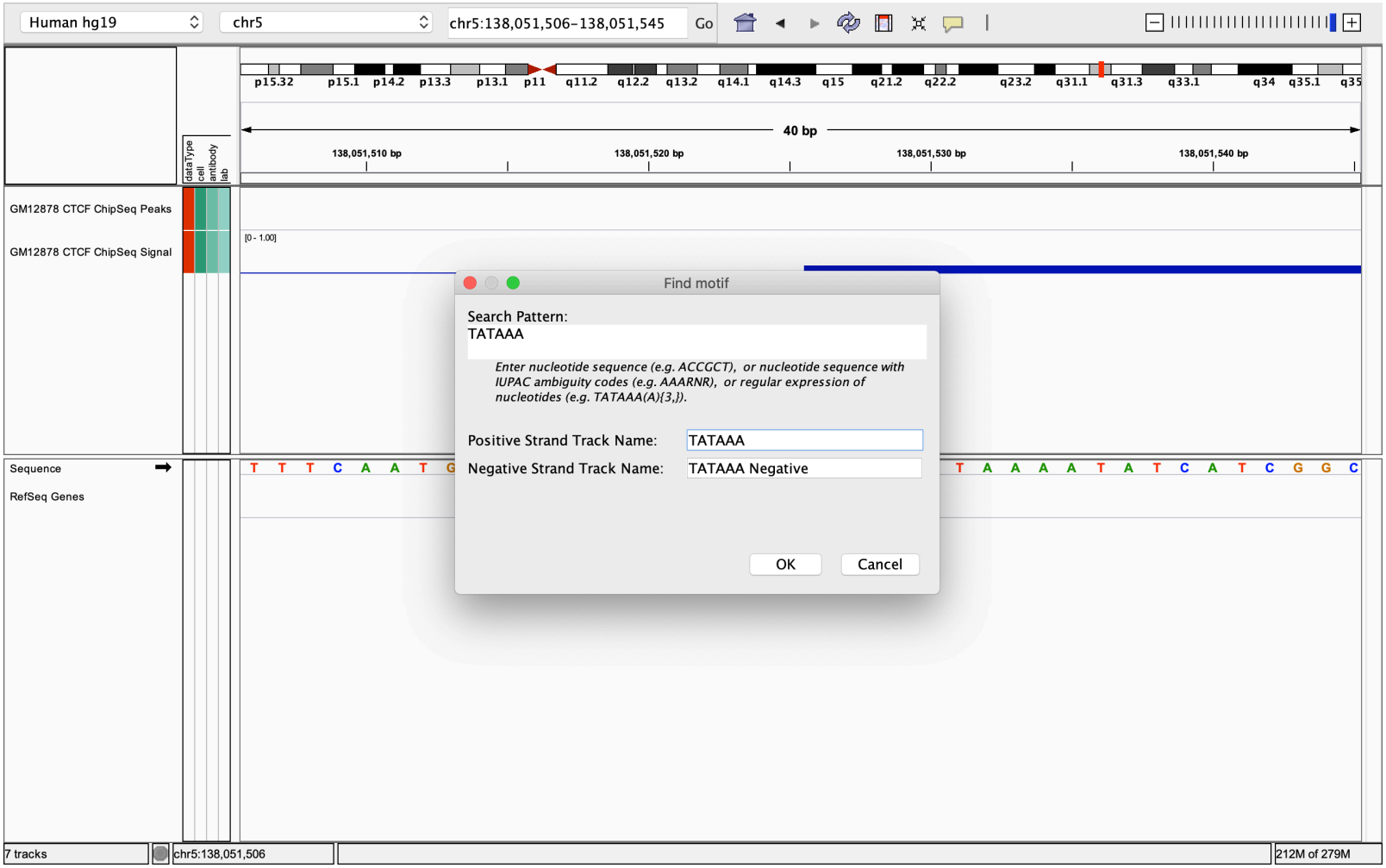
Find Motif



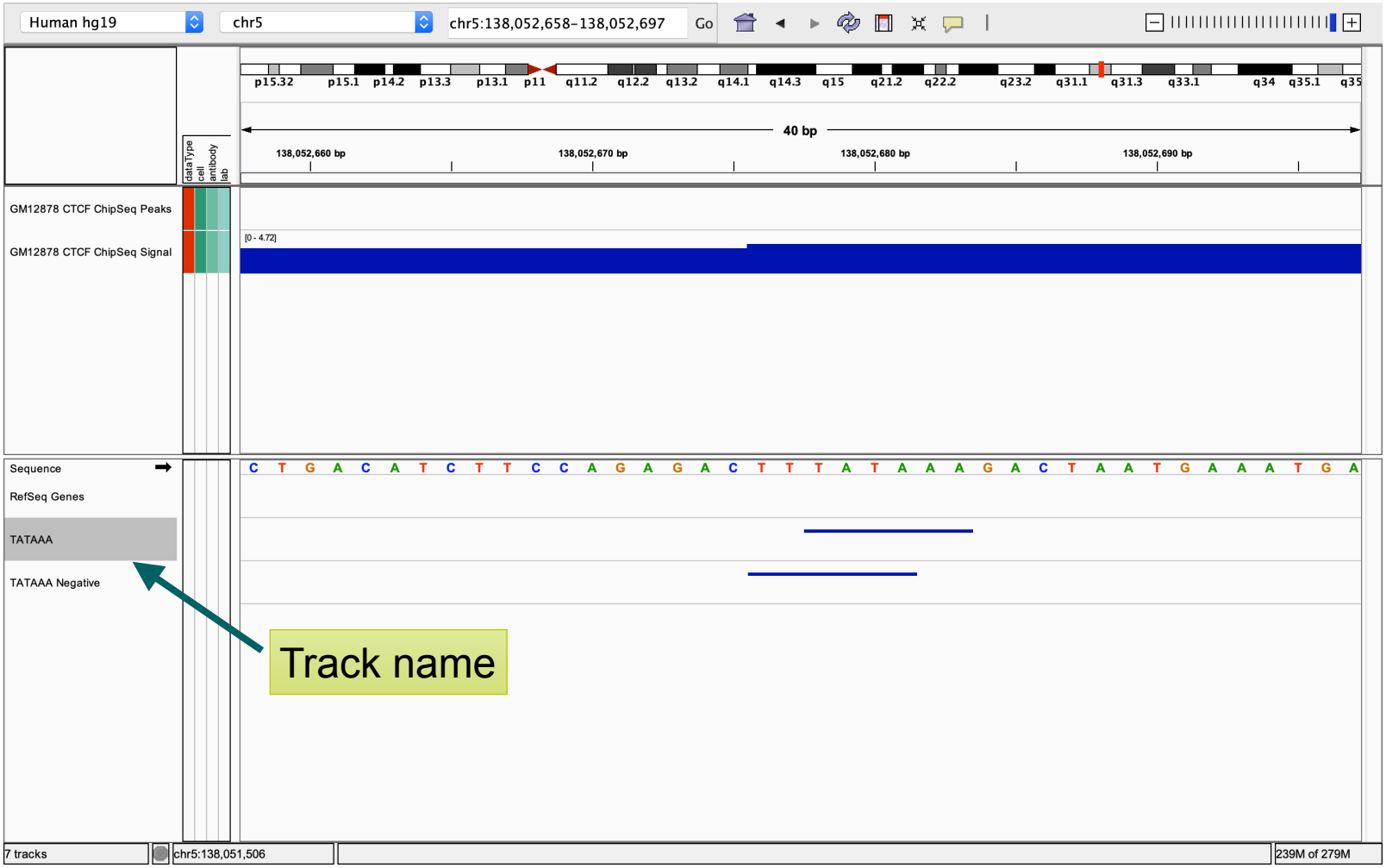
Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Find Motif


Human hg19
chr5
chr5:138,051,506-138,051,545
Go
p15.32 p15.1 p14.2 p13.3 p13.1 p11 q11.2 q12.2 q13.2 q14.1 q14.3 q15 q21.2 q22.2 q23.2 q31.1 q31.3 q33.1 q34 q35.1 q35
40 bp
138,051,510 bp
138,051,520 bp
138,051,530 bp
138,051,540 bp
dataType
cell
antibody
lab
GM12878 CTCF ChipSeq Peaks
GM12878 CTCF ChipSeq Signal
[0 - 1.00]
Find motif
Search Pattern:
TATAAA
Enter nucleotide sequence (e.g. ACCGCT), or nucleotide sequence with IUPAC ambiguity codes (e.g. AAARNR), or regular expression of nucleotides (e.g. TATAAA(A){3,}).
Positive Strand Track Name:
TATAAA
Negative Strand Track Name:
TATAAA Negative
OK
Cancel
Sequence
RefSeq Genes
7 tracks
chr5:138,051,506
212M of 279M

Find Motif

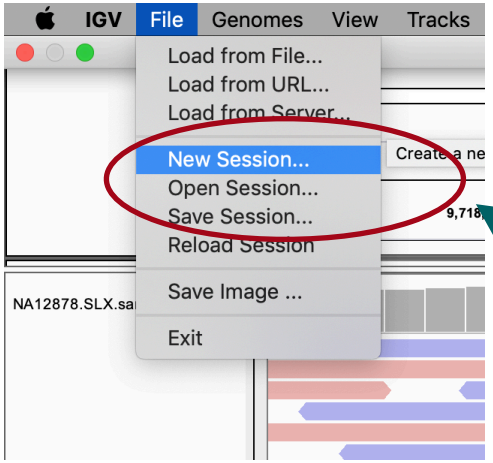


Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Viewing SNPs

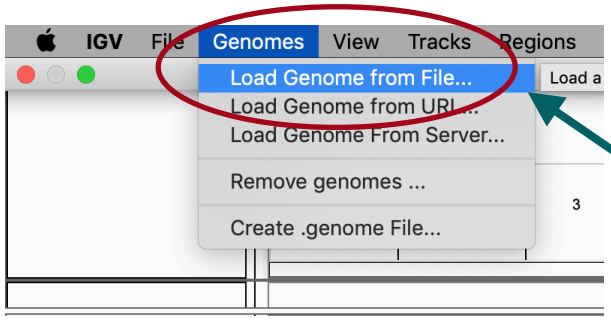
- Download data:
  - https://costalab.ukaachen.de/open_data/Bioinformatics_Analysis_in_R_2019/BIAR_D5/igvData.tar.gz
- Create a new session



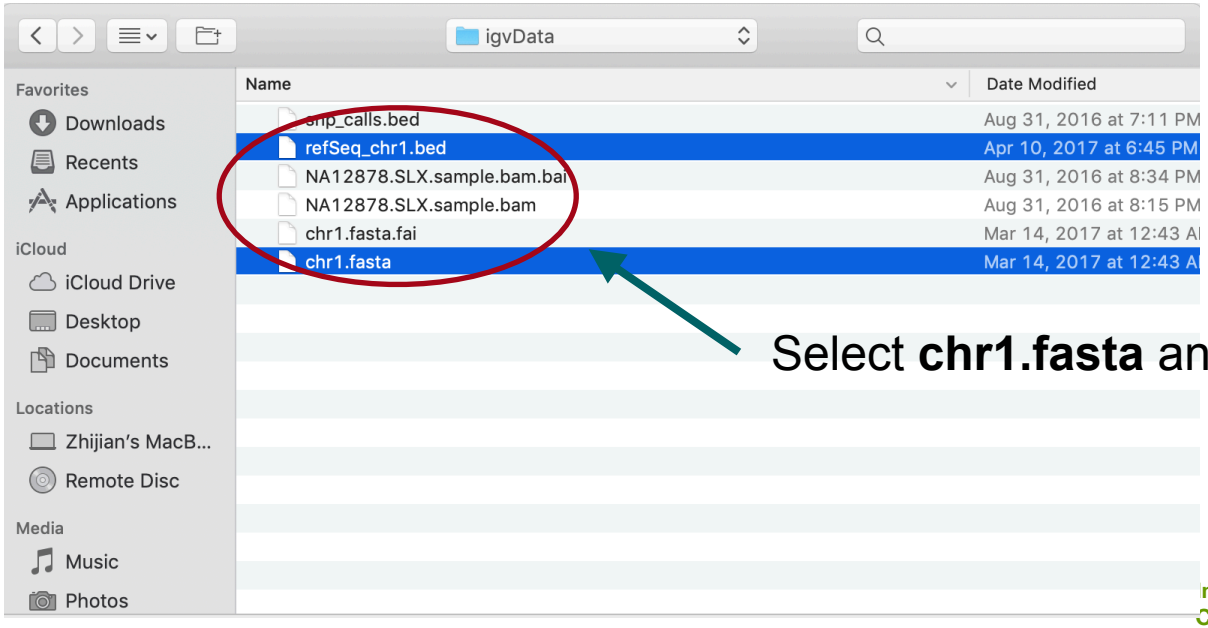
Click **New Session**

Viewing SNPs

- Load reference genome:



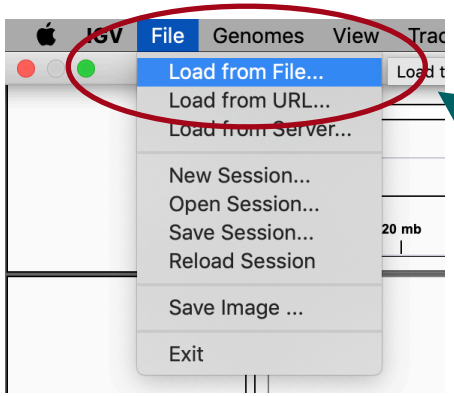
Click **Genomes > Load Genome from File**



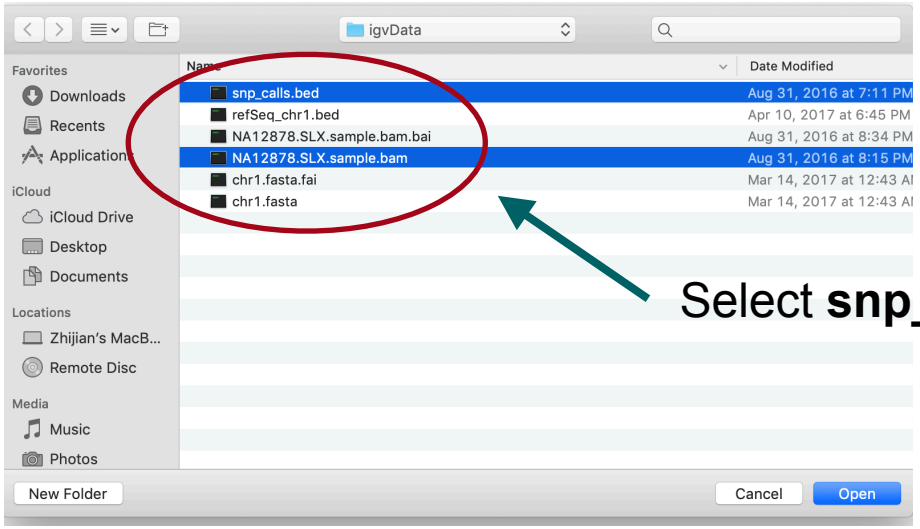
Select **chr1.fasta** and **refSeq_chr1.bed**

Viewing SNPs

- Load sequencing data:

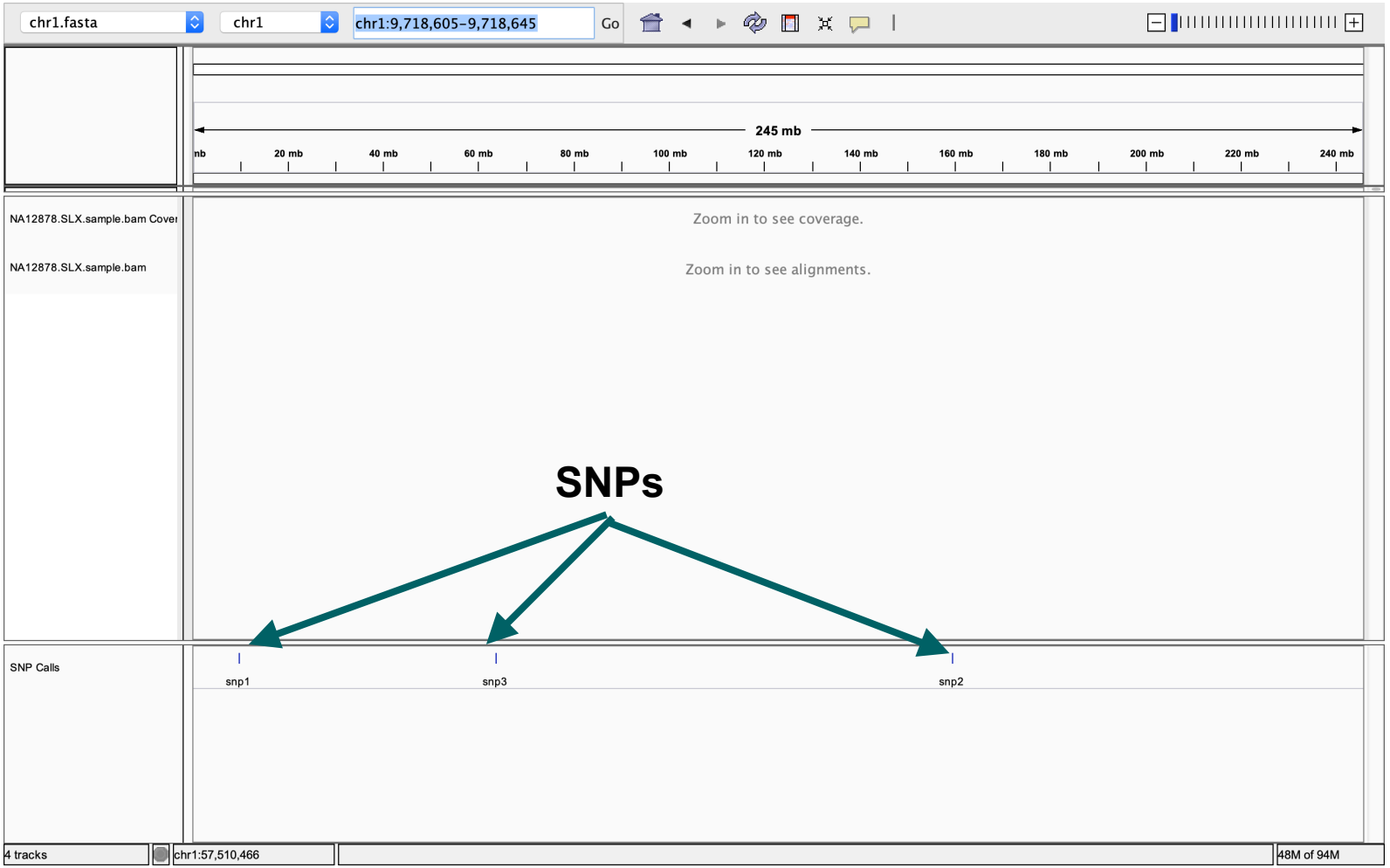


Click **File > Load Genome from File**

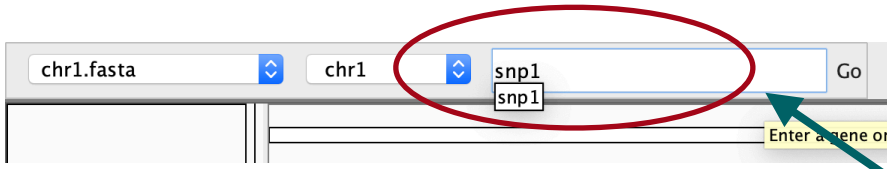


Select **snp_call,bed** and**NA12878.SLX.sample.bam**

Viewing SNPs



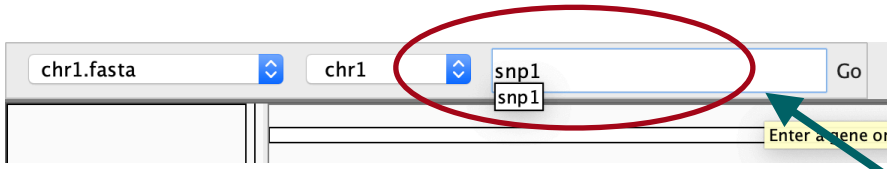
Viewing SNPs



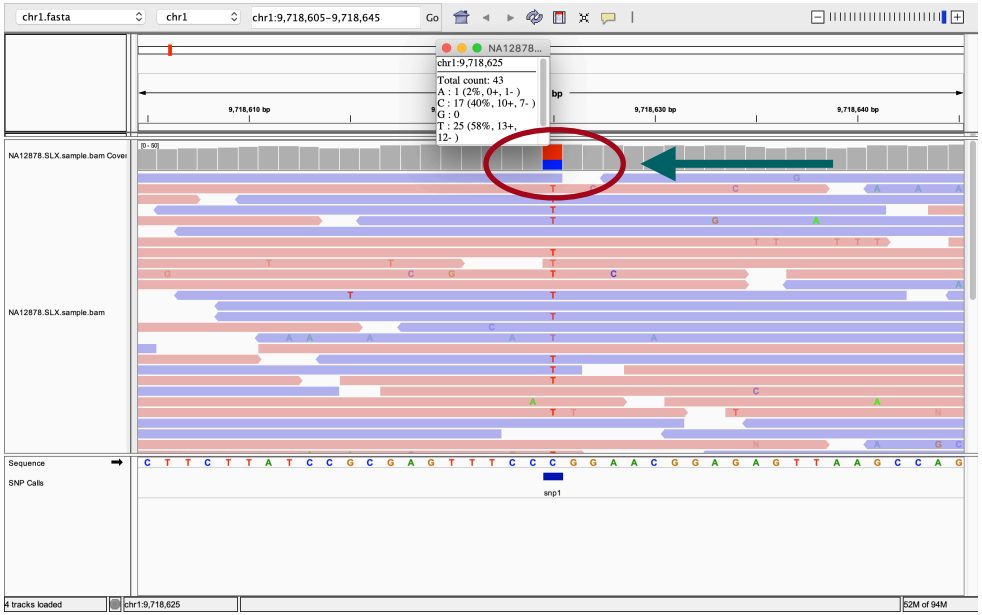
Input snp1 and Click **Go**

a SNP with **C -> T**

Viewing SNPs



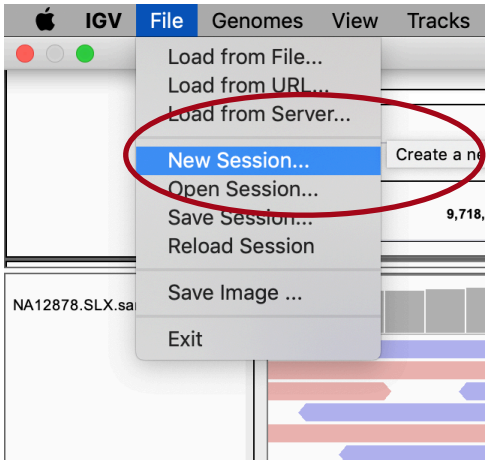
Input snp1 and Click **Go**



Click here to see statistics

Viewing RNA-seq data

- Create a new session



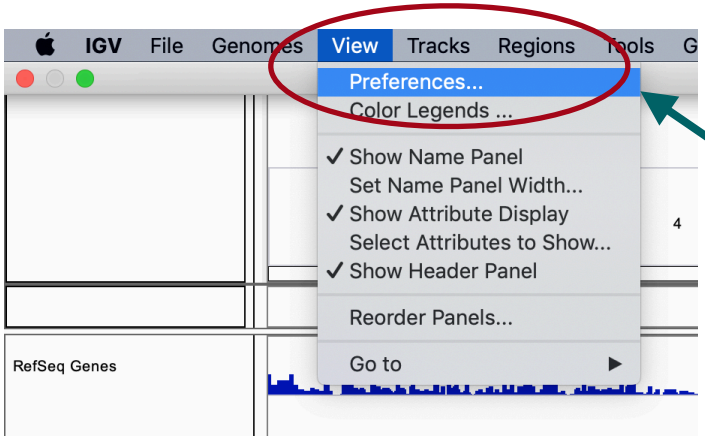
Click **New Session**

- Choose hg19 as reference genome

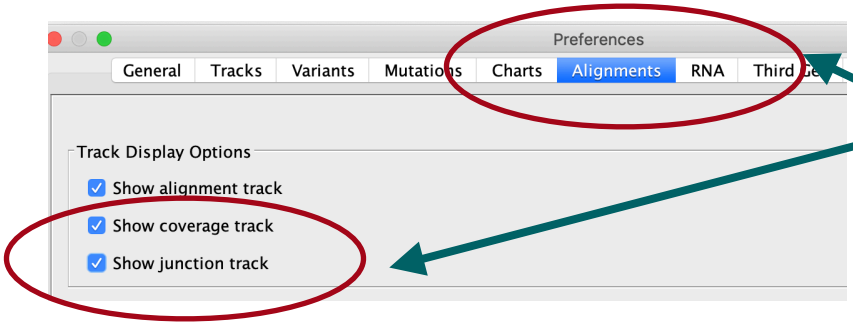
Click **drop-down list** and select **hg19**

Viewing RNA-seq data

- Set preference for viewing RNA-seq data



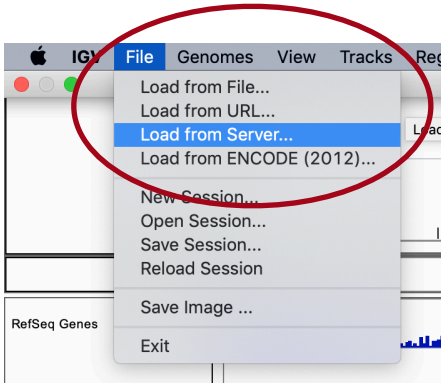
Click **View > Preference**



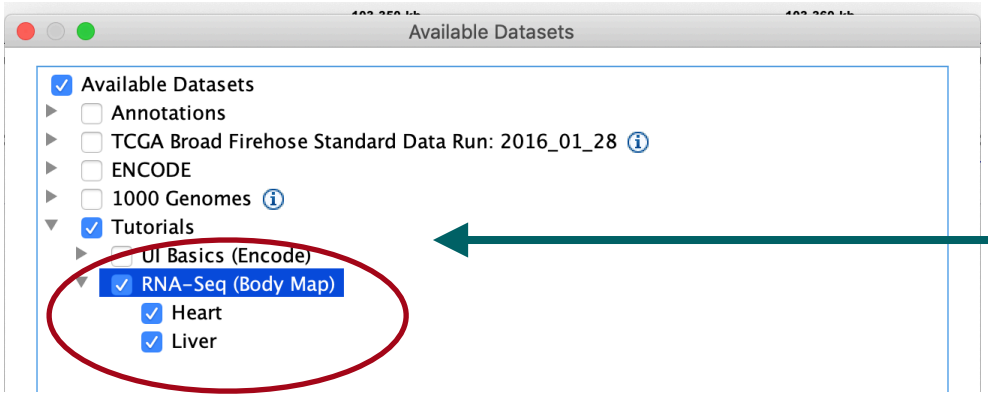
Select **Alignments** tab
Check **junction track**

Viewing RNA-seq data

- Load data



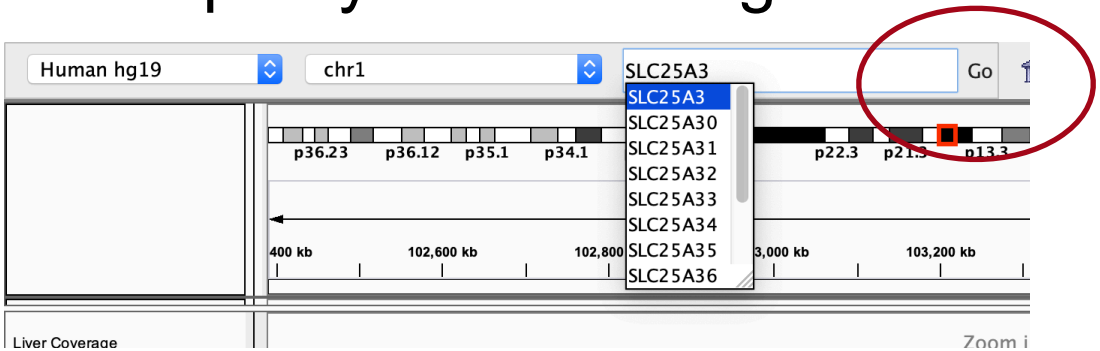
Click **File > Load from Server**



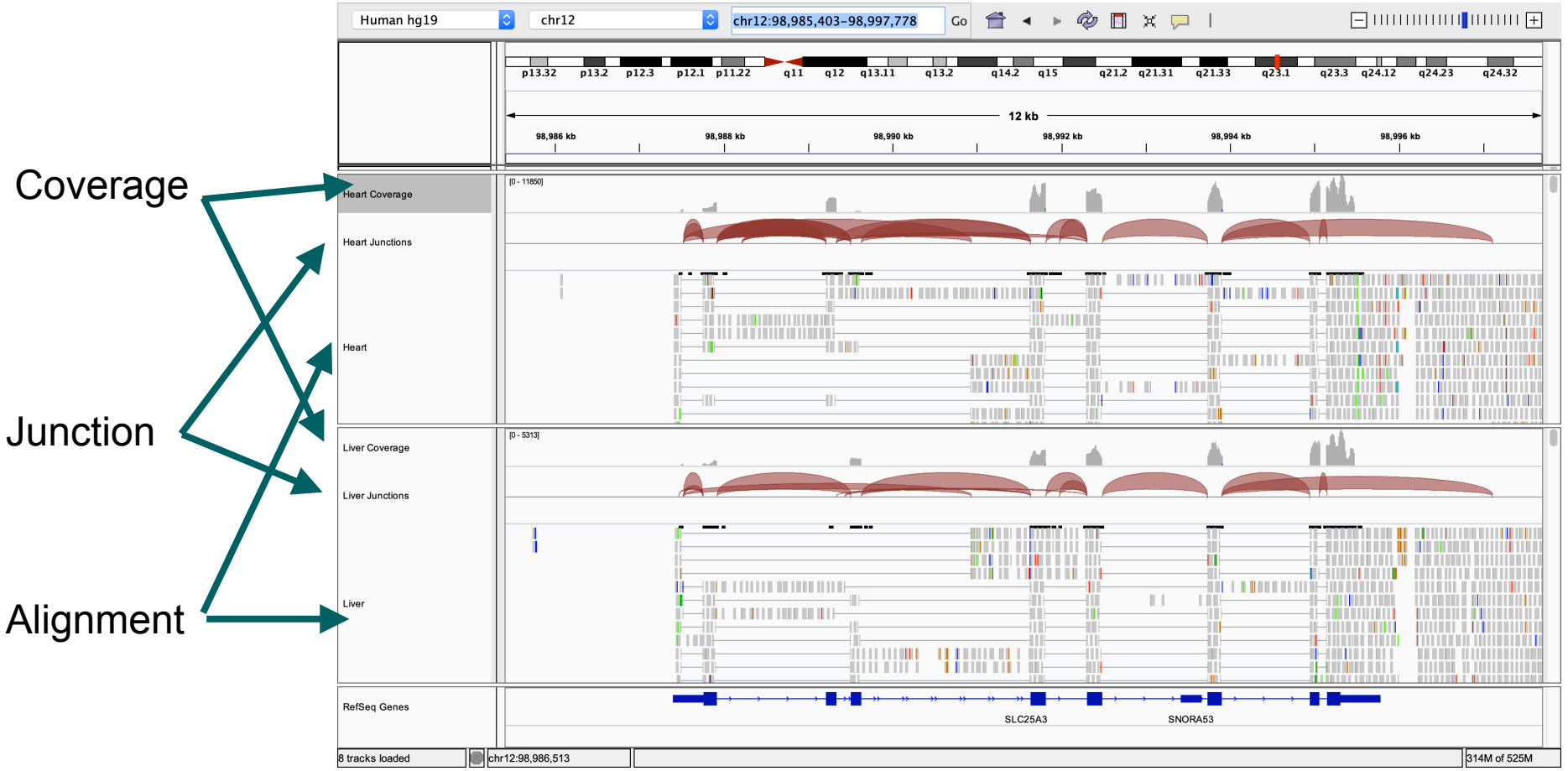
Open **Tutorial**
Click on **RNA-seq > OK**

Viewing RNA-seq data

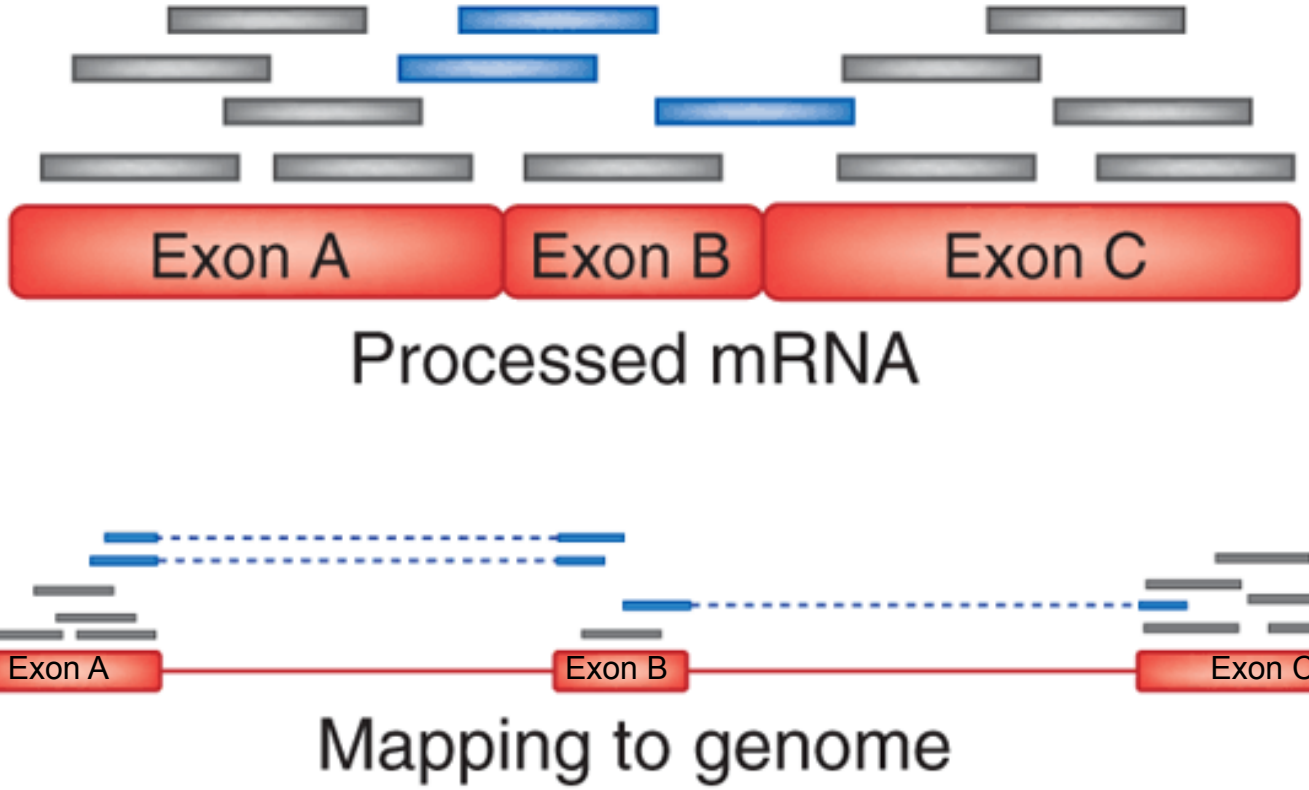
- Jump to your favorite gene



Viewing RNA-seq data



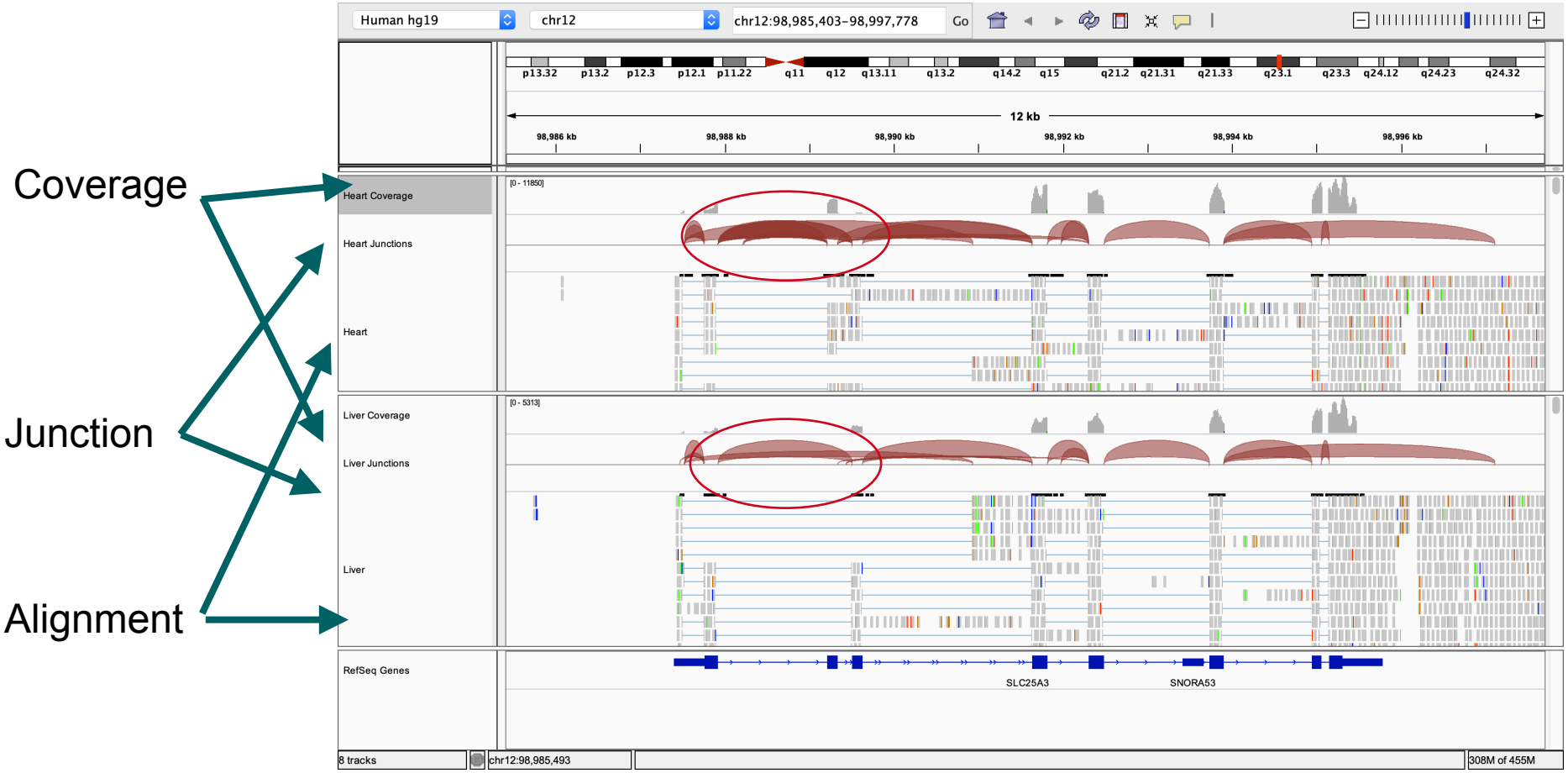
Split Read Mapping (RNA-Seq)



- **Reads spanning distinct exons indicate junctions**

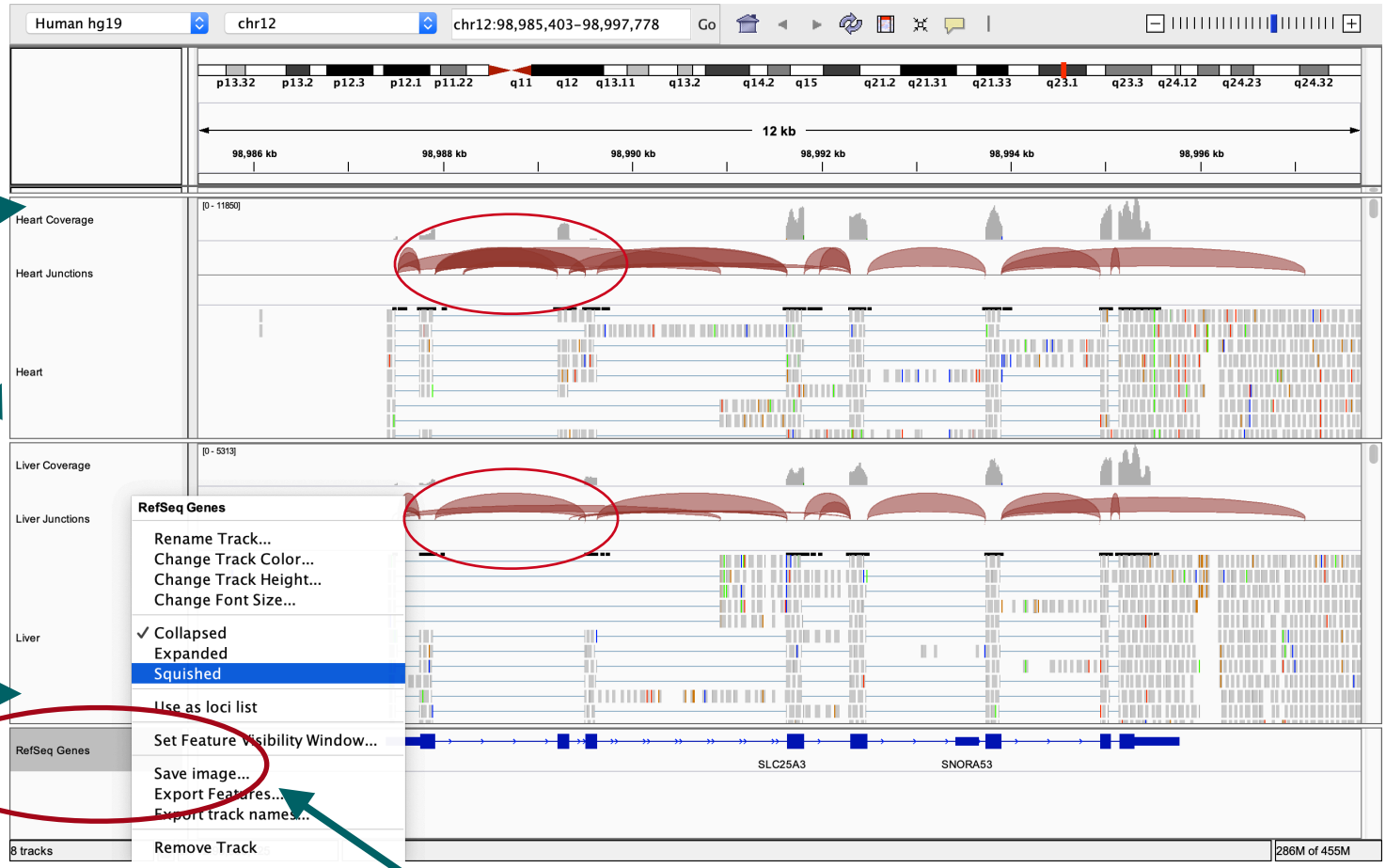
Viewing RNA-seq data

- Example of alternative splicing



Viewing RNA-seq data

- Example of alternative splicing



Coverage

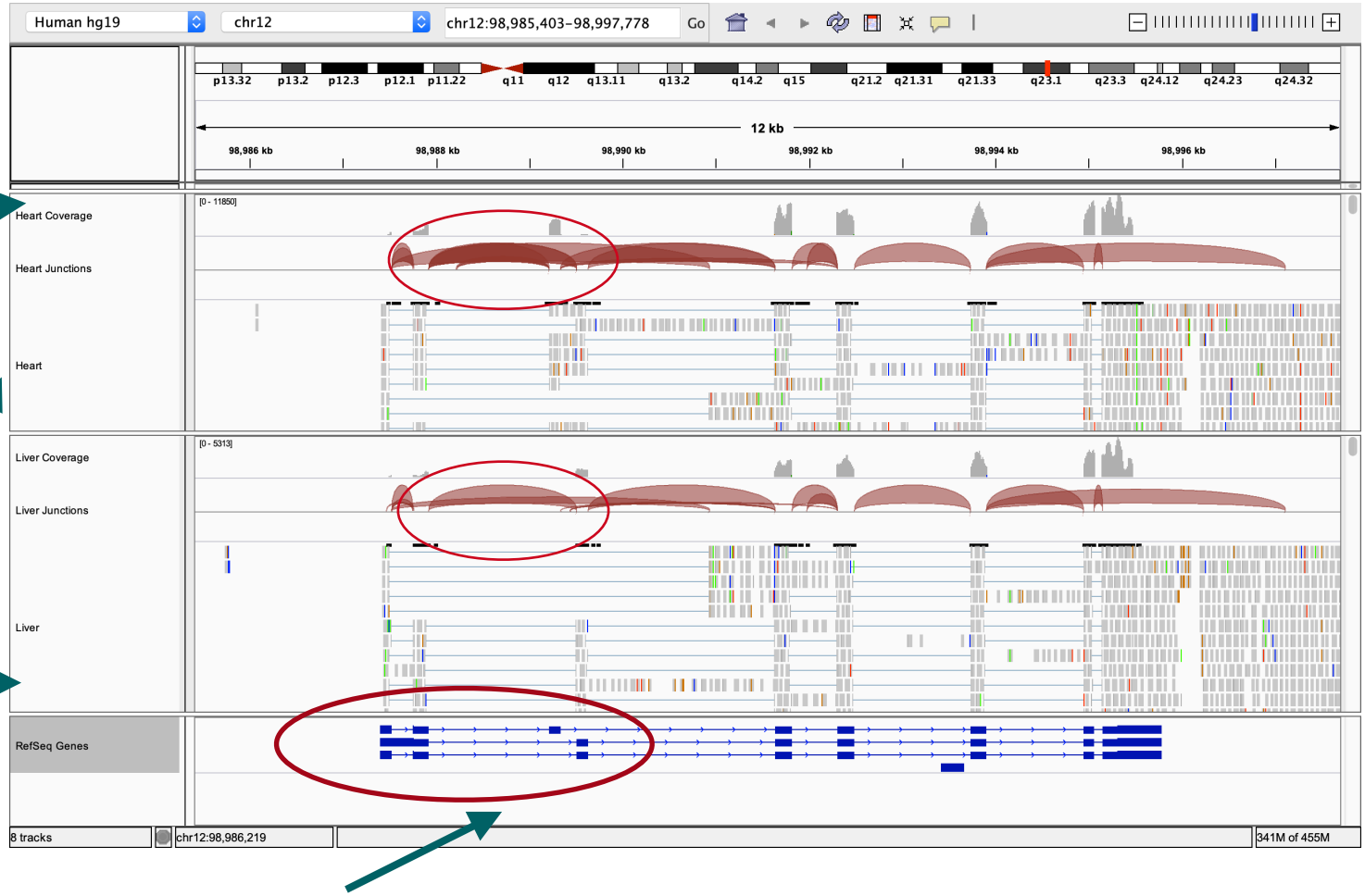
Junction

Alignment

Right click and select **Squished**

Viewing RNA-seq data

- Example of alternative splicing



Coverage

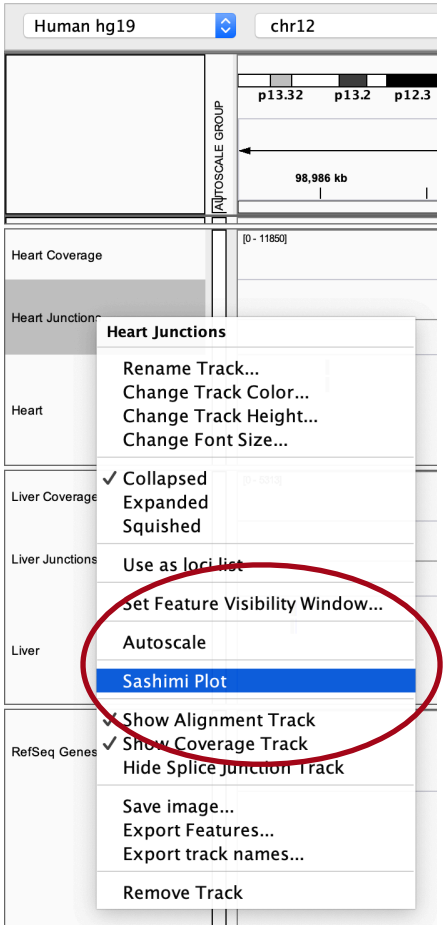
Junction

Alignment

Alternative splicing

Viewing RNA-seq data

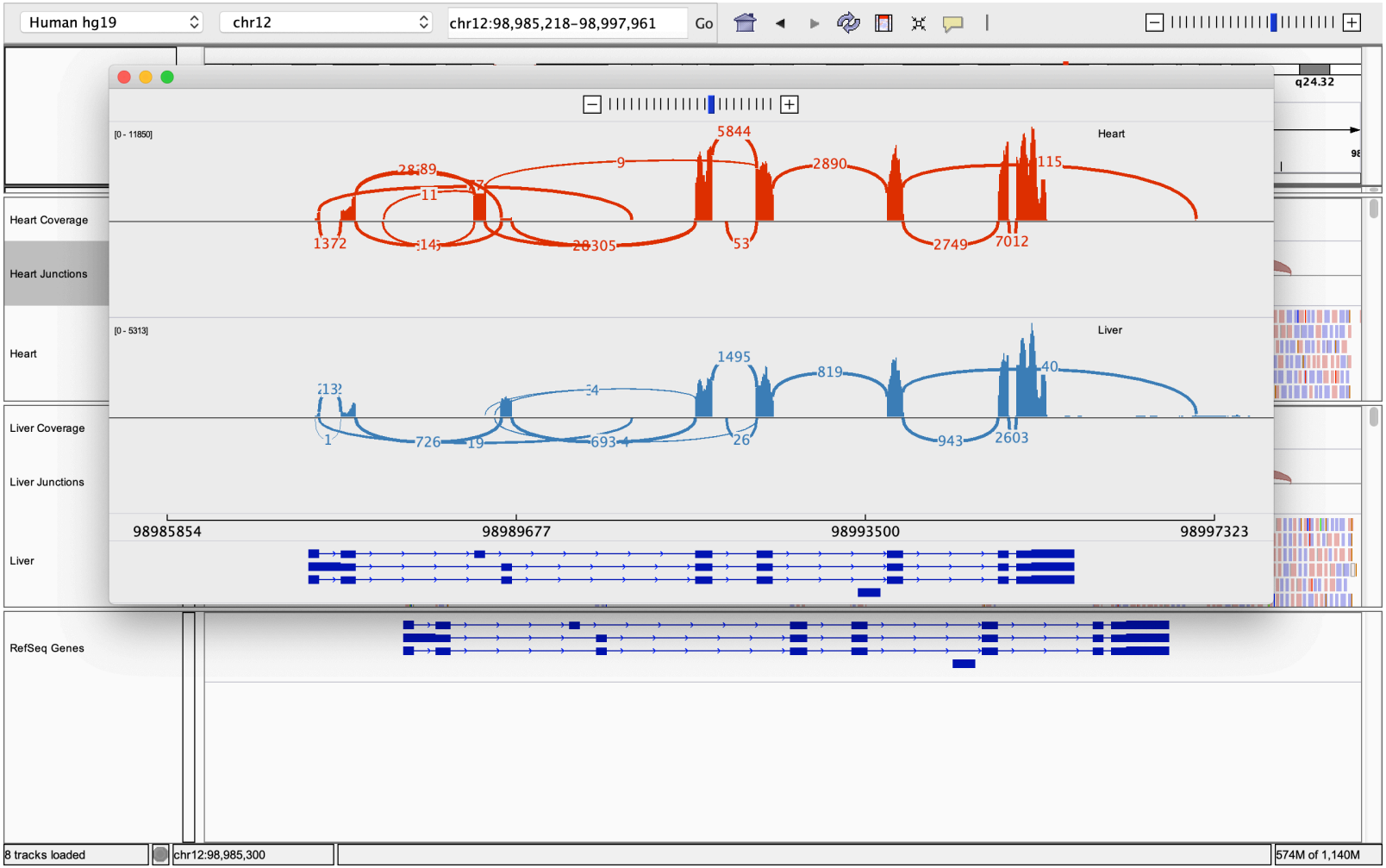
- Example of alternative splicing



Right click **Junction > Sashimi plot**

Viewing RNA-seq data

- Example of alternative splicing



www.costalab.org

Institute for
Computational Genomics
01011011010
10100100101

RWTH AACHEN
UNIVERSITY