



## 03. Alignment

Erik Dassi & Davide Bressan

Genomics Technologies Lab

# High-throughput alignment

What we need:

- Reference genome
- Next Generation Sequencing (NGS) reads
- Mapping software

# Reference

## Genome sequence:

- is a digital nucleic acid sequence database, assembled by scientists as a representative example of the set of genes in one idealized individual organism of a species
- Reference genomes are stored in FASTA files

## Genome annotation:

- is the process of attaching biological information to the reference sequences, and particularly in identifying the locations of genes and determining what those genes do
  - Annotations are typically stored in GTF/GFF files
- 
- There are multiple databases from which we can download the genome sequences and annotations: RefSeq, **GENCODE**, Ensembl, UCSC, ...

# Reference

<https://www.encodegenes.org>



Human Mouse How to access data FAQ Documentation About us

## HUMAN

GENCODE 43 (08.02.23)



## MOUSE

GENCODE M32 (08.02.23)



The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.

GENCODE are [updating the annotation](#) of human protein-coding genes linked to SARS-CoV-2 infection and COVID-19 disease.

The [Long-read RNA-seq Genome Annotation Assessment Project](#) (LRGASP) Consortium for systematic evaluation of different methods for transcript computational identification and quantification using long-read sequence data **has launched**.

GENCODE are [supporting the annotation](#) of non-canonical human ORFs predicted by Ribo-seq data.

# Reference

We could also be using a mitochondrial-only reference (not recommended), that could be obtained from:

- <https://www.mitomap.org/foswiki/bin/view/MITOMAP/HumanMitoSeq>
- <https://www.ncbi.nlm.nih.gov/nuccore/251831106>

# Downloading and storing files

- Your user folder maximum size is limited on the lab machines
- But we can use the machine **local** disk to store the files
- Create a folder within the **/var/tmp/** folder (e.g. /var/tmp/myname)
- Store files there when downloading them and do the analyses from there

# Required files

**To perform the alignment, please download:**

- The GENCODE primary assembly file (genome sequence)
- The trimmed sequencing reads files for your group

# Burrows-Wheeler Alignment Tool (BWA)

<https://bio-bwa.sourceforge.net/>

		F	L
mississippi#		#	mississipp i
ississippi#m		i	#mississip p
ssissippi#mi		i	ppi#missis s
sissippi#mis		i	ssippi#mis s
issippi#miss		i	ssissippi# m
ssippi#missi	⇒	m	issippi #
sippi#missis		p	i#mississi p
ippi#mississ		p	pi#mississ i
ppi#mississi		s	ippi#missi s
pi#mississip		s	issippi#mi s
i#mississipp		s	sippi#miss i
#mississippi		s	sissippi#m i



# Building the reference

- To speed-up the alignment and make it **feasible** (time- and memory-wise), we need to preprocess the reference sequence to build an *index*
- BWA provides the *index sub*command:

```
bwa index -p mitoIndex <fasta filename>
```

*-p = prefix of the output index files*

# Aligning the reads

- We can then use the index, together with our QC-passed reads and other parameters to actually run the alignment
- BWA provides the *mem* algorithm to align reads to the reference:

```
bwa mem -t 2 -R "@RG\tID:<sampleName>\tSM:<sampleName>"  
<reference prefix> <reads.fq>
```

-R: read group (fixed value)  
-t: number of parallel threads to use  
-sampleName: name of the sample (e.g.RPE-MITO-NXT)

**Reference prefix:** the value of the -p parameter previously provided to the index command

# Output format

<https://samtools.github.io/hts-specs/SAMv1.pdf>

- SAM – Sequence Alignment/Map format
  - stores alignment information
  - Contains quality information, meta data, alignment information, sequence etc.
- BAM – BGZF compressed SAM format
  - Compressed/binary version of SAM, not human readable
  - Uses a specialized compression algorithm optimized for indexing and record retrieval (bgzip)
  - Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
  - Unsorted, sorted by sequence name, sorted by genome coordinates
  - May be accompanied by an index file (.bai) (only if coordinate sorted)
  - Files are typically very large: ~ 1/5 of SAM, but still very large

# BAM indexing

- We need to index the BAM files to allow for a quick visualization and opening even on PCs with little RAM memory
- Prior to indexing, we need to sort the BAM
- Samtools allows us to do both steps:

```
samtools sort <file.bam> <file_sorted.bam>
```

```
samtools index <file_sorted.bam>
```

- This will produce a .bai file with the same name as the sorted BAM

# Visualizing the alignments

The ***Integrative Genomics Viewer (IGV)*** is a tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata.

<https://www.igv.org/>

**Open IGV** (*igv* command in the shell)

- Select a reference genome (hg38)
- File > Load from File...
- Select your bam file

IGV User Guide and Tutorials: <http://software.broadinstitute.org/software/igv/>

# Visualizing the alignments

File Genomes View Tracks Regions Tools GenomeSpace Help

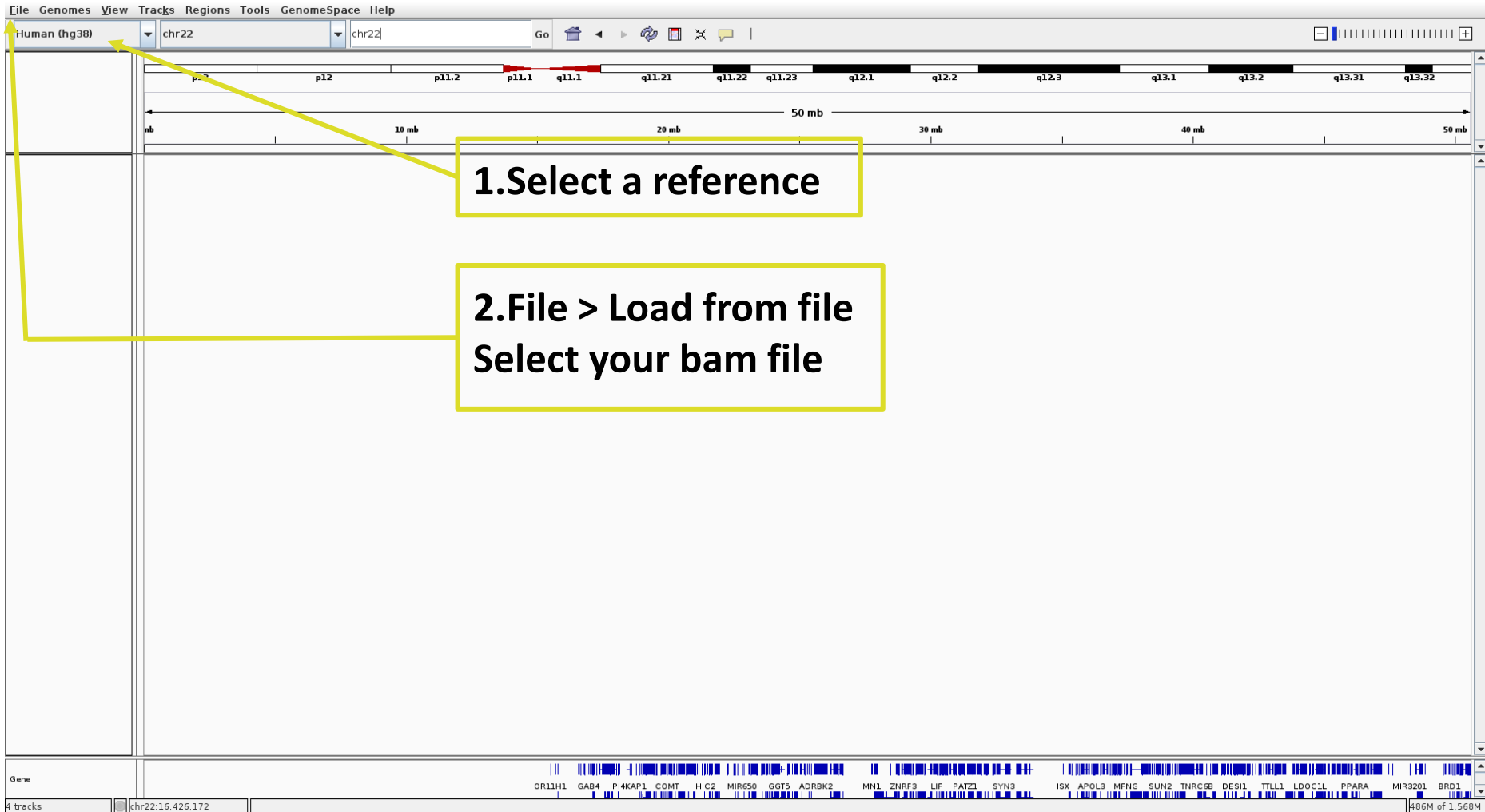
Human (hg38) chr22 chr22 Go

1. Select a reference

2. File > Load from file  
Select your bam file

Gene

4 tracks chr22:16,426,172 H86M of 1,568M



# Visualizing the alignments





***Questions ?***