

Lab session day 2

The power of epigenetic regulation – central dogma

DNA

mRNA

Protein



~23,000 protein coding genes

DNA

mRNA

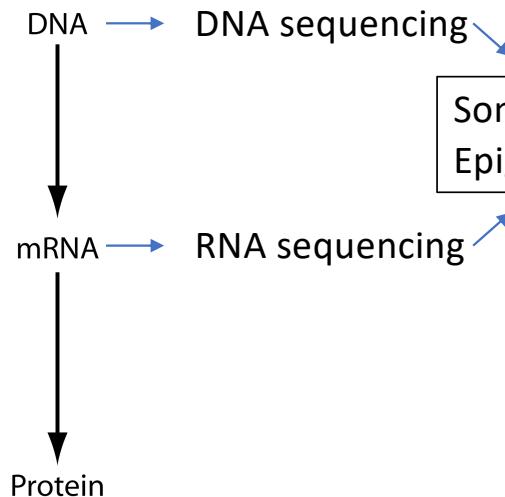
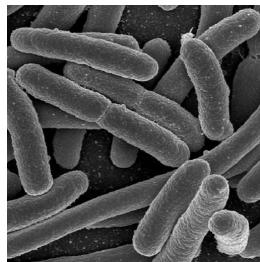
Protein

MASSIVE REGULATION

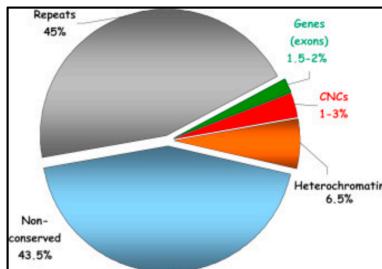


~20,500 protein coding genes

The power of epigenetic regulation – central dogma



Inherent challenges



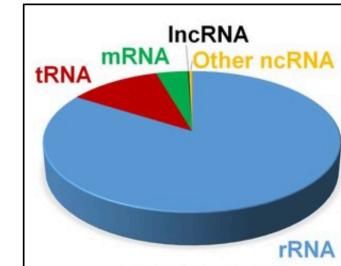
1.5% coding genes

Apply
enrichment
or
WGS

PCR
or
hybridisation

Interrogate protein coding genes

Identify mutations
Structural rearrangements
Microsatellite instability
...



~2% mRNA

Apply
enrichment

polyA-tail
Or
rRNA depletion

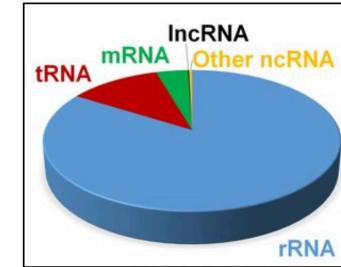
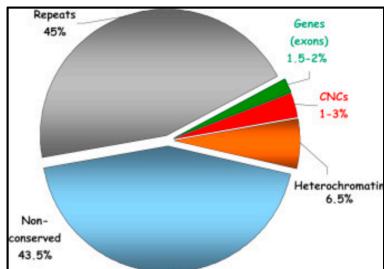
Interrogate protein coding genes

Expressed mutations -> neoantigen -> Immunotherapy
Kinase outlier expression
Gene fusions
Tissue/phenotype specific expression profile
...

Clinical Cancer Genomics – vt 2022

Figures: from googling ...

Inherent challenges



1.5% coding genes

cDNA

~2% mRNA

Library prep

sequencing

Similar but not identical informatics

Interpretation

Figures: from googling ...

Clinical Cancer Genomics – vt 2022

Today's labwork

1. Investigate files needed for processing DNA and RNA data e.g. the human genome reference
2. Run various bioinformatic tools
 1. Processing data
 2. Quality control
3. Investigate processed data in the Integrative Genomics Viewer (IGV).
4. Potential extra task: How to define, order and quality check a targeted sequencing assay.

Human genome reference



- Reference genome: an assembly: an attempt to produce as complete version of the human genome as possible
- Difficulties due to repetitive DNA/non identical DNA stretches etc.
- Pay attention to the version!

Name (link)	Description
1000 Genomes reference	Used in this course. The 1000g reference names chromosomes as follows (chr1 , chr2 , ..., chr22 , chrX , chrY , chrM). This reference includes "decoy" sequences (mostly low complexity sequences) that have been added to the standard genome build sequence. This reduces misalignment of reads that would otherwise get placed somewhere they don't belong. The developer of the BWA aligner documents use of this version of the reference genome. This reference includes the alternative contigs.
Ensembl reference	Ensembl names the chromosomes as follows (1 , 2 , ..., 22 , X , Y , MT). The names of some unplaced contigs also differ. This reference does NOT have the decoy sequences. This reference includes the alternative contigs.
UCSC reference	The UCSC reference names chromosomes as follows (chr1 , chr2 , ..., chr22 , chrX , chrY , chrM). This reference does NOT have the decoy sequences. This reference includes the alternative contigs.
NCBI reference	NCBI names the chromosomes as follows (chr1 , chr2 , ..., chr22 , chrX , chrY , chrMT). This reference does NOT include the decoy sequences. This reference includes the alternative contigs. The major annotation centers such as UCSC and Ensembl start with raw files from NCBI (Various Human Assemblies). Most other people do not use these NCBI files directly but rather get a version of the files from UCSC, Ensembl, etc.
Genomic Data Commons (GDC) reference	The GDC reference names chromosomes as follows (chr1 , chr2 , ..., chr22 , chrX , chrY , chrM). The GDC created their own version of the reference for harmonized analysis of the TCGA and other large cancer sequencing projects. This reference includes "decoy" sequences. This reference does NOT include the alternative contigs. Unique to this reference is the inclusion of several virus sequences for viruses with known or suspected roles in cancer (e.g. HPV, EBV, etc.).

Indexing files

1. Many bioinformatic tools are dependent on that the file has been indexed
2. Allows bioinformatic tools to efficiently access only the required information
 1. E.g: a tool wants to look at a specific position on chr 20, not efficient to start from chr 1 and read the file to chr 20.
3. Done for reference genome, annotation files for the human genome etc.
4. Both for DNA- and RNA-seq tools

Course data



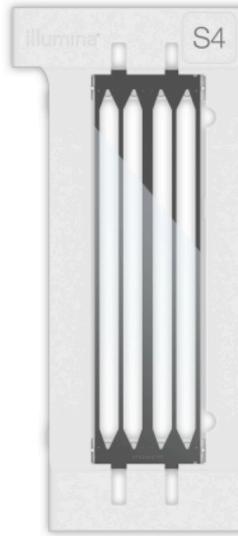
1. Breast cancer cell line (HCC1395) and its matched lymphoblastoid cell line (HCC1395BL).
 1. Cell line is available at the American Type Culture Collection (ATCC) store.
 2. DNA exome-seq, wgs-seq and RNA-seq data.
2. Commercial cfDNA from mCRPC patients (generated at Scilifelab).

Cell line data from pmbio.org

Illumina sequencing

- <https://www.youtube.com/watch?v=fCd6B5HRaZ8>

NovaSeq 6000



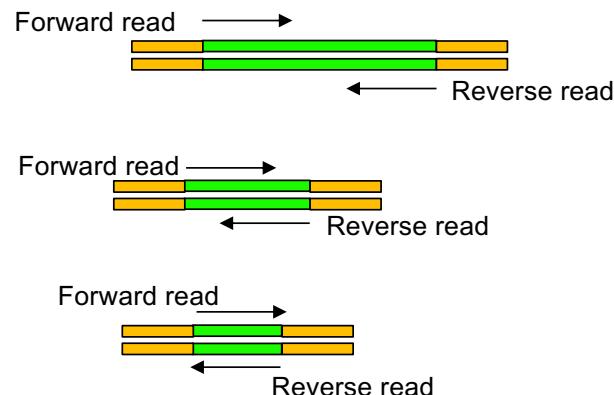
~2000 M read-pairs per lane
2 x 150 bp

iPCM:

tumor DNA, 40M read-pairs
germline DNA, 15M read-pairs

$2000/55 = 36$ Tumor/Normal pairs/lane
3 T/N wgs pairs (90x tumor and 30x gDNA)/lane
166M read pairs/30x genome

Illumina sequencing



Forward read

Reverse read

The sequence

Base qualities

Raw input into all pipelines

```
==> SRR001666_1.fastq <=>
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 → Header
GGGTGATGGCCGCTGGCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 → Header
GTTCAAGGGATACGACGTTTGATTAAAGAACATGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <=>
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 → Header
AAGTTACCCCTAACACTTAAGGGTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIDIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 → Header
-AGCAGAAAGTCGATGATAATACGCGTCGTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
-IIIIIIIIIIIIIIIIIGIT>IIIIII-I)8I
```

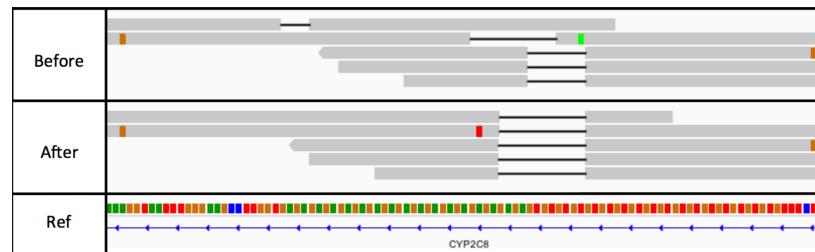
$$Q_{\text{sanger}} = -10 \log_{10} p$$



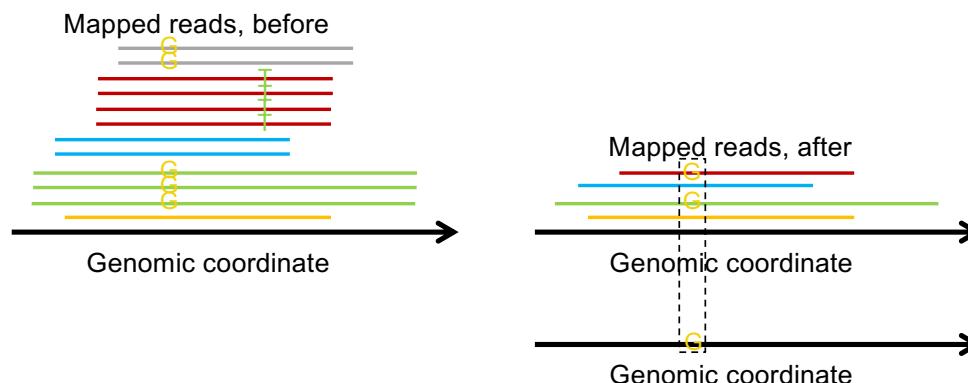
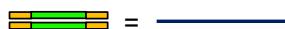
The very basic bioinformatic pipeline

1: Map reads to the human genome  Find the location in the human genome

2: Realign reads



3: Remove duplicates



4: Run variant callers

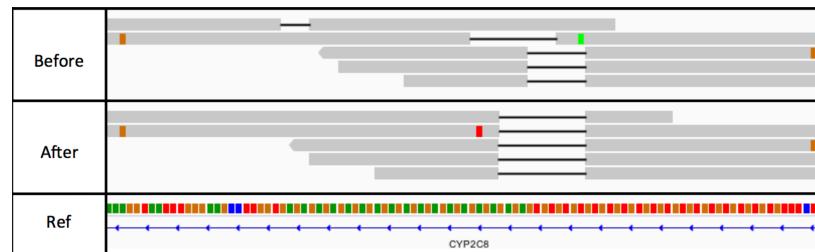
5: Data interpretation for research project



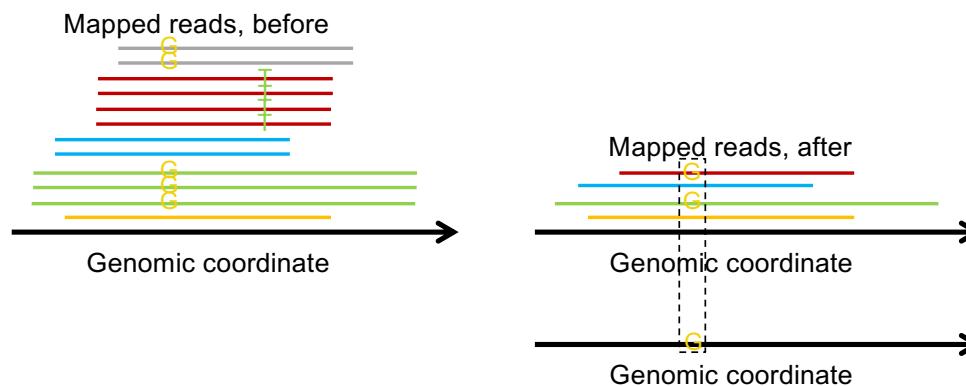
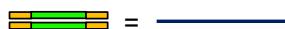
The very basic bioinformatic pipeline

1: Map reads to the human genome  Find the location in the human genome

2: Realign reads



3: Remove duplicates



4: Run variant callers

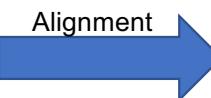
5: Data interpretation for clinical use → Inspect variants manually according to SOP

The very basic bioinformatic pipeline

1: Map reads to the human genome

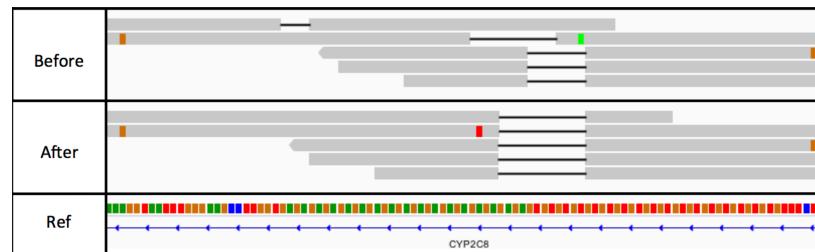


Find the location in the human genome

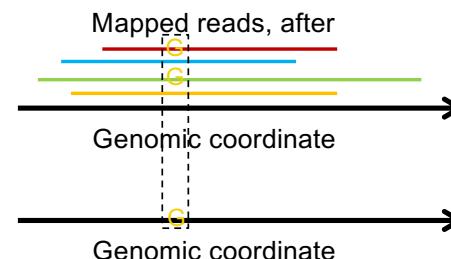
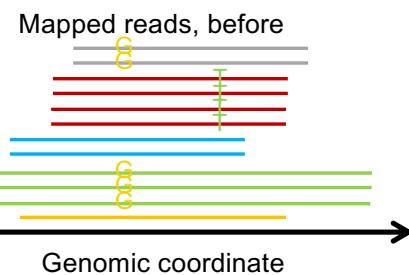
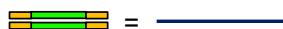


BAM files (Binary Alignment Map).
Compressed version of SAM files,
Sequence Alignment/Map format.
<https://samtools.github.io/hts-specs/>

2: Realign reads



3: Remove duplicates

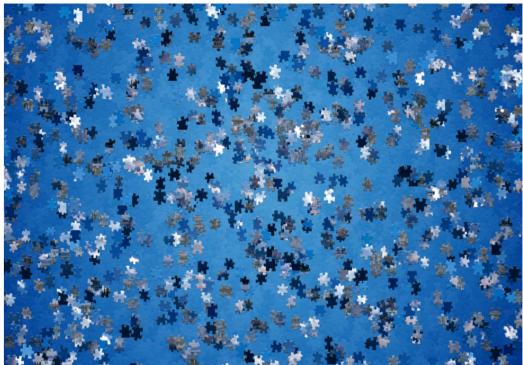


4: Run variant callers

5: Data interpretation for clinical use → Inspect variants manually according to SOP

The very basic bioinformatic pipeline

- Alignment is like solving a puzzle



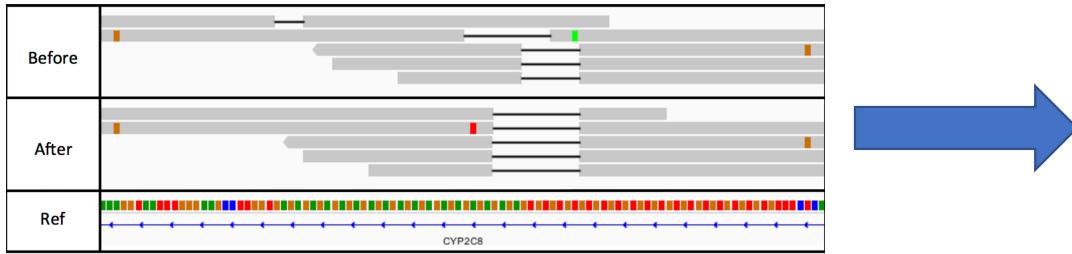
- Use of fast algorithms, mistakes will happen .. (check mapping quality BAM files)
- Different genome builds exist out there – be careful!

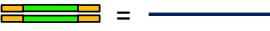
SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived

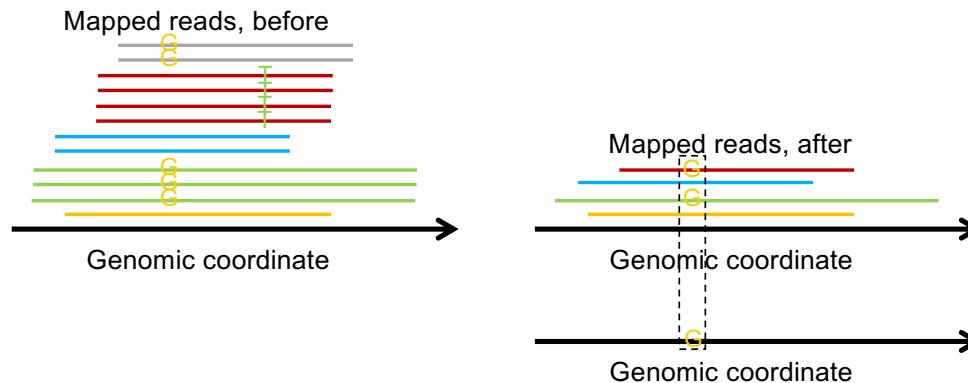
Figs from pmbio.org

The very basic bioinformatic pipeline

1: Map reads to the human genome  Find the location in the human genome

2: Realign reads  A modified bam-file

3: Remove duplicates 



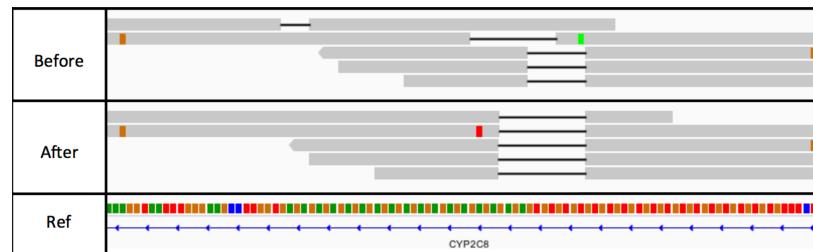
4: Run variant callers

5: Data interpretation for clinical use  Inspect variants manually according to SOP

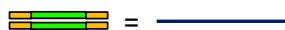
The very basic bioinformatic pipeline

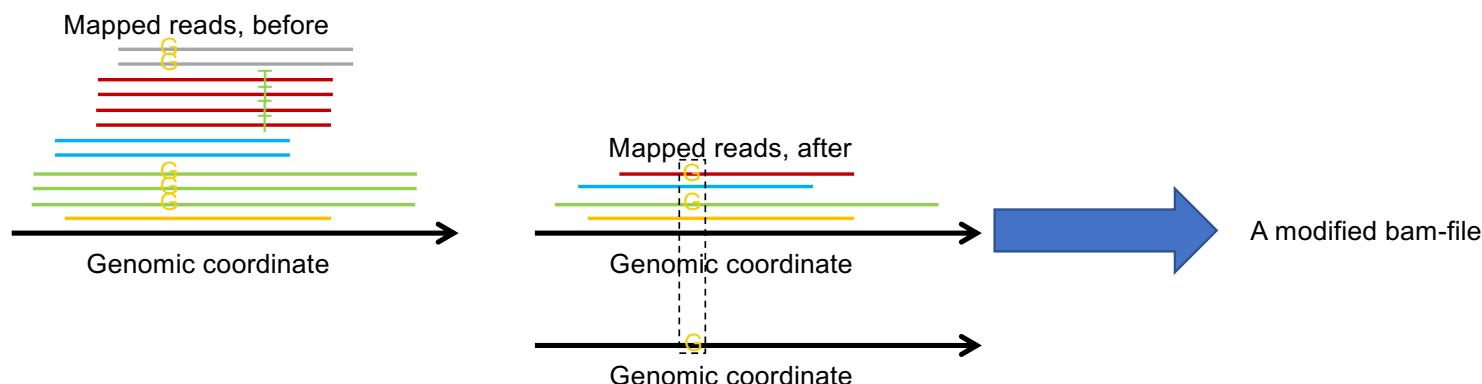
1: Map reads to the human genome  Find the location in the human genome

2: Realign reads



3: Remove duplicates

 = 



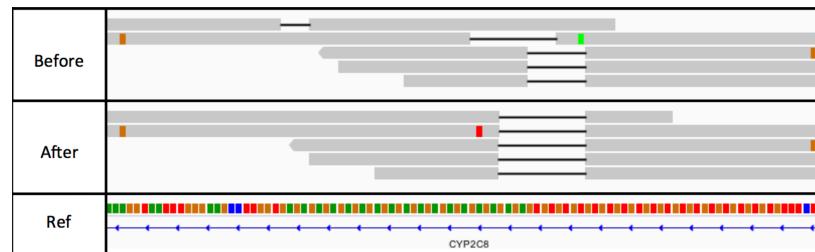
4: Run variant callers

5: Data interpretation for clinical use  Inspect variants manually according to SOP

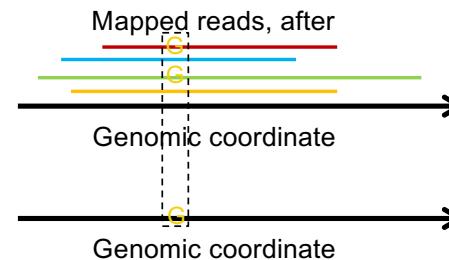
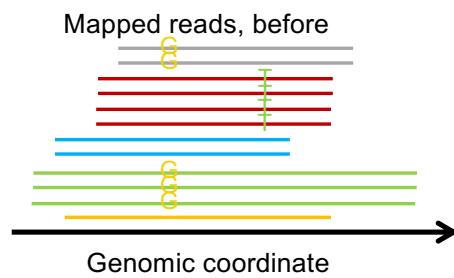
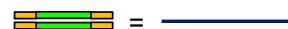
The very basic bioinformatic pipeline

1: Map reads to the human genome  Find the location in the human genome

2: Realign reads



3: Remove duplicates



4: Run variant callers

Variant Call Format (VCF)
Mutation annotation format (MAF)
Other tab delimited formats...

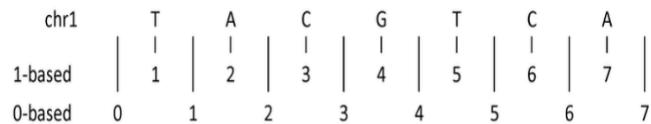
5: Data interpretation for clinical use → Inspect variants manually according to SOP

The very basic bioinformatic pipeline

Watch out for booby traps!

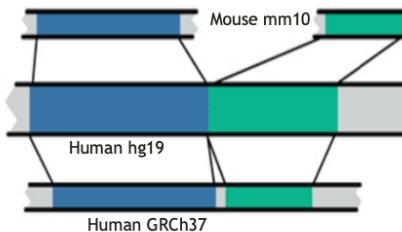
Genomic coordinate systems

- 1-based vs. 0-based



Genome builds

- And annotation builds
- “Liftover” tools



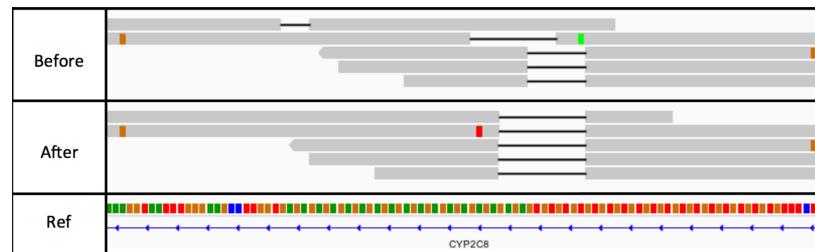
Left-shifted vs right-shifted

REF	GGCA	2	GGCA	GG	Not left trimmed
ALT	GG	1			
REF	GCA	3	GCA	G	Normalized (left aligned & parsimonious)
ALT	G	1			

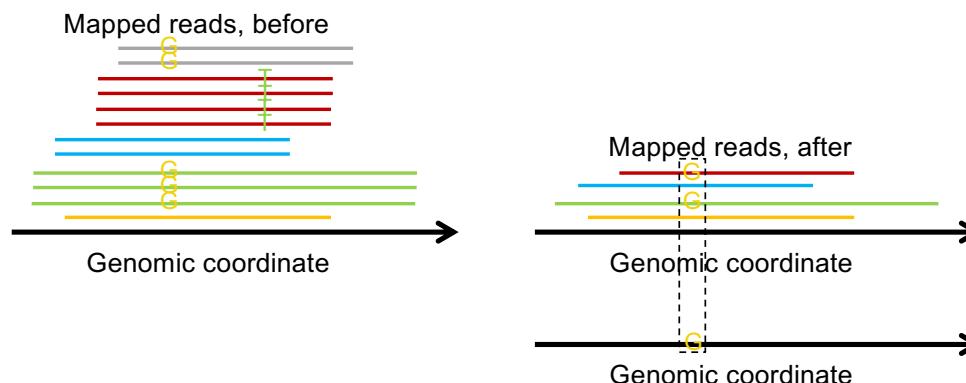
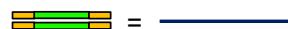
The very basic bioinformatic pipeline

1: Map reads to the human genome  Find the location in the human genome

2: Realign reads



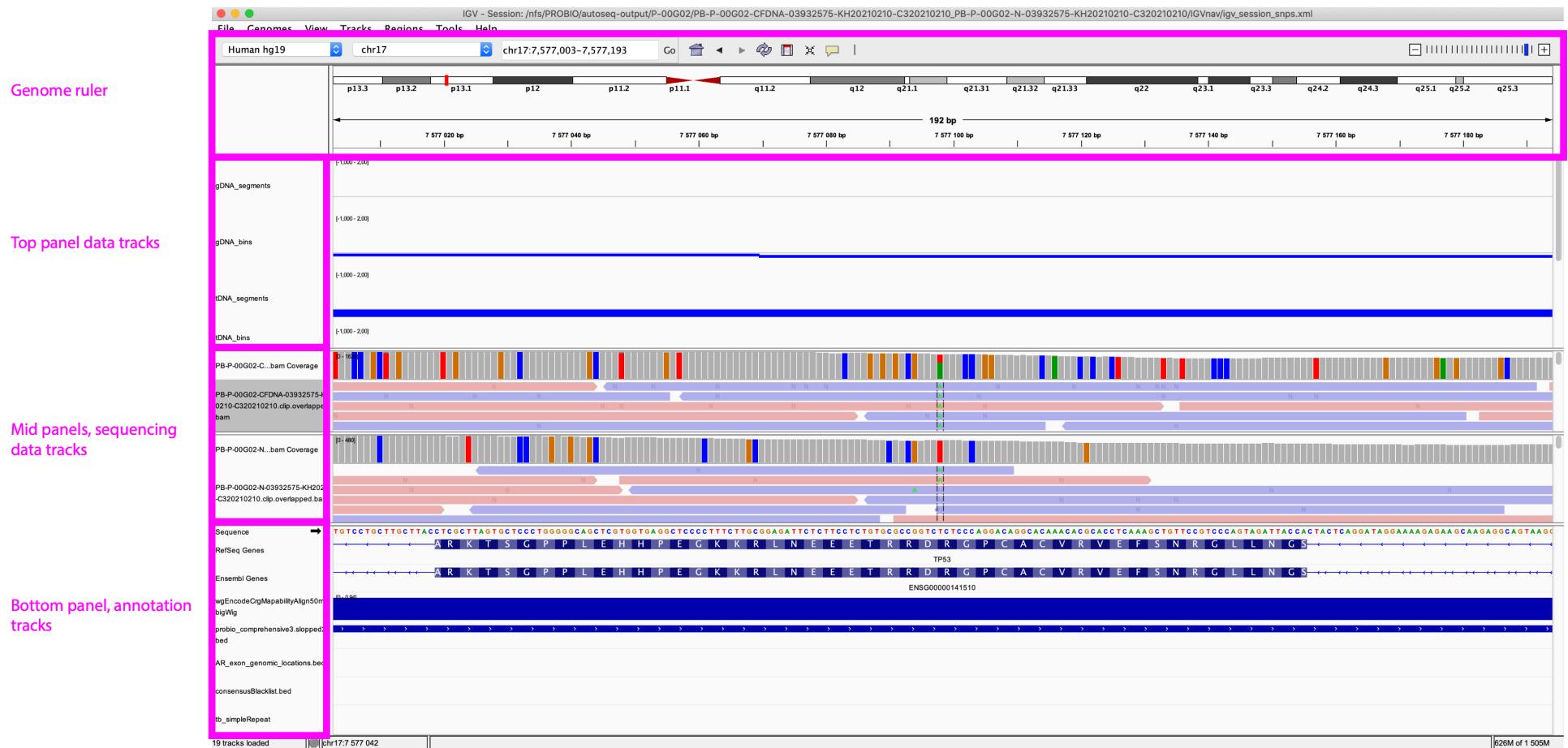
3: Remove duplicates



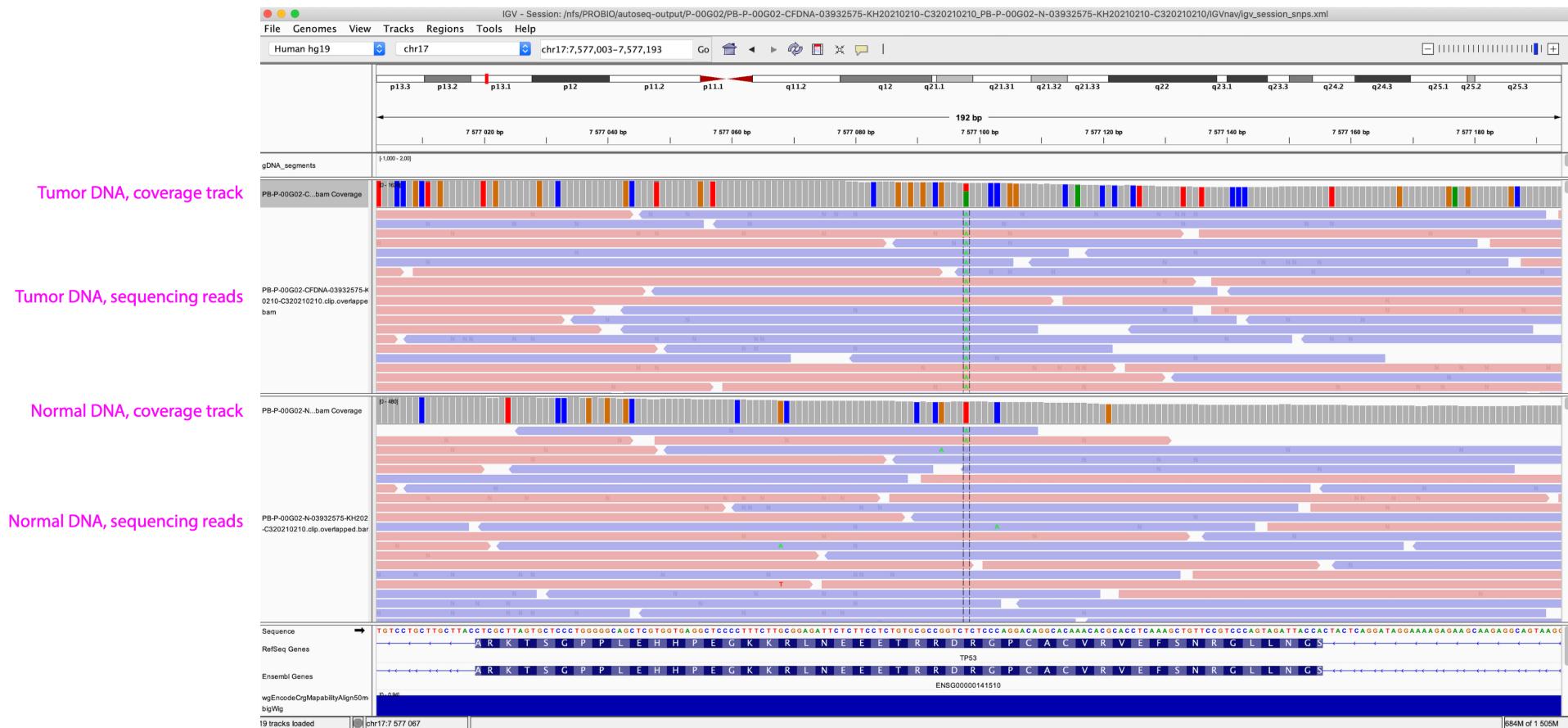
4: Run variant callers

5: Data interpretation for clinical use → Inspect variants manually according to SOP → IGV galore!

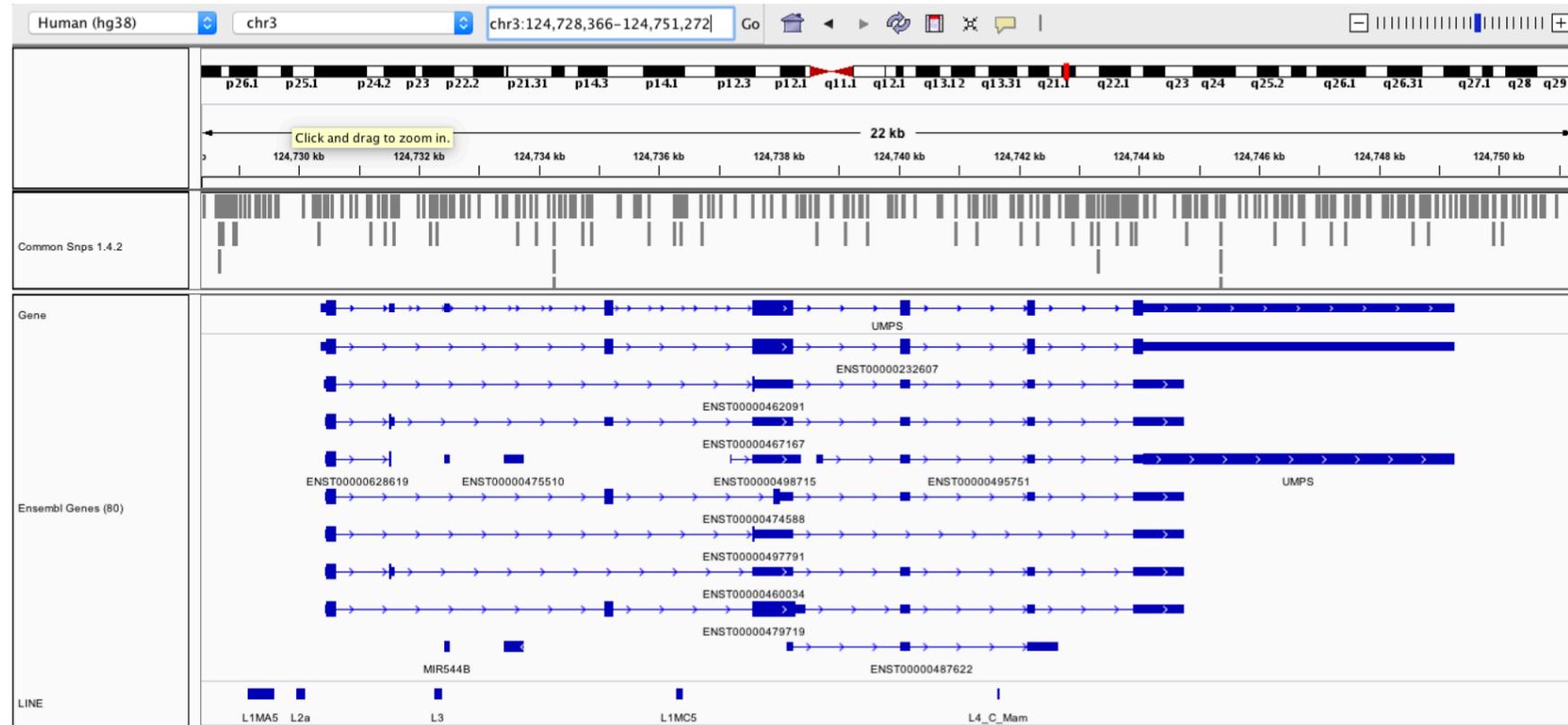
Integrative Genomics Viewer (IGV)



Integrative Genomics Viewer (IGV)

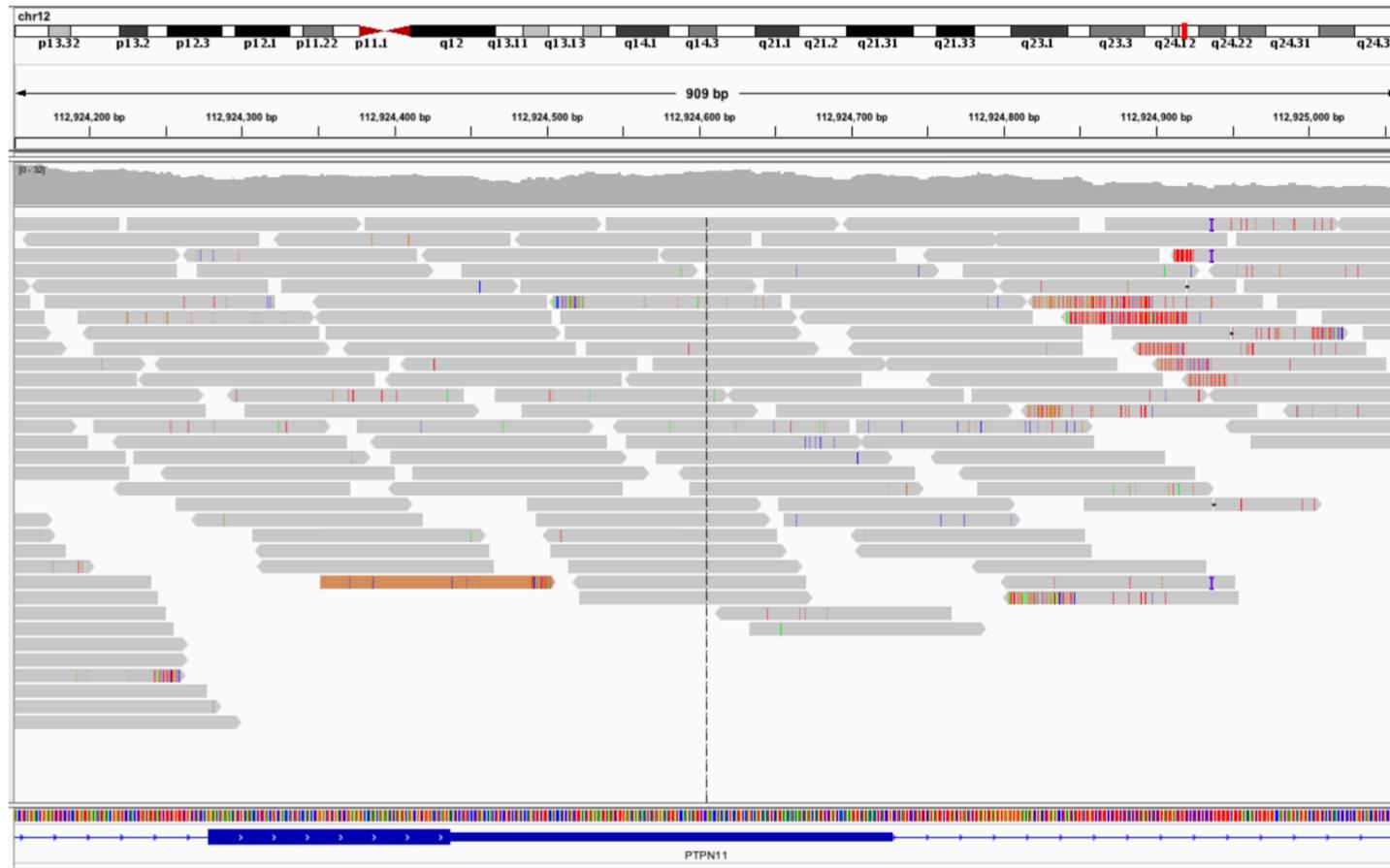


Integrative Genomics Viewer (IGV)



Figs from pmbio.org

Integrative Genomics Viewer (IGV)



Figs from pmbio.org

Lets get started!