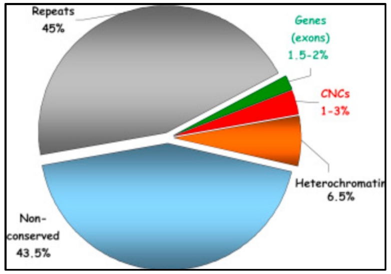
Lab session day 2

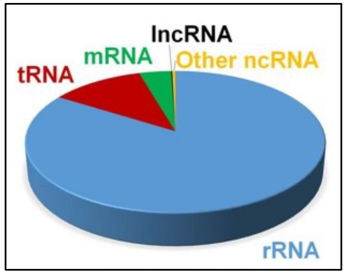
Learning outcomes

- understand the constituents of a bioinformatics pipeline for processing Illumina sequencing data and to run such a pipeline.
- understand the file formats used in high throughput sequencing.
- use the command line and running bioinformatic tools.
- perform quality control on DNA- and RNA sequencing data for cancer sequencing purposes.

Inherent challenges with DNA- and RNA-seq



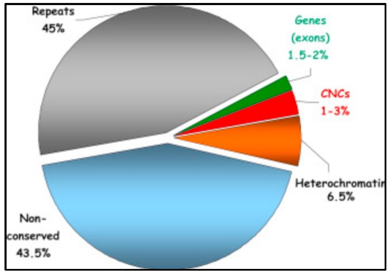
1.5% coding genes



~2% mRNA

Figures: from googling …

Inherent challenges with DNA- and RNA-seq



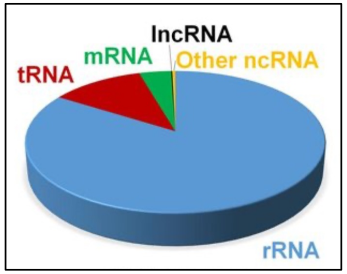
1.5% coding genes

Apply
enrichment
or
WGS

PCR
or
hybridisation

Interrogate protein coding genes

Identify mutations
Structural rearrangements
Microsattelite instability
…



~2% mRNA

Apply
enrichment

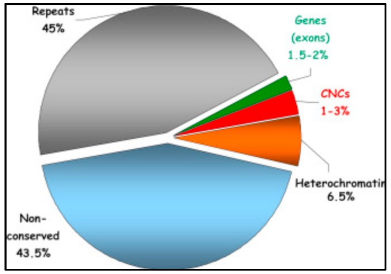
polyA-tail
Or
rRNA depletion

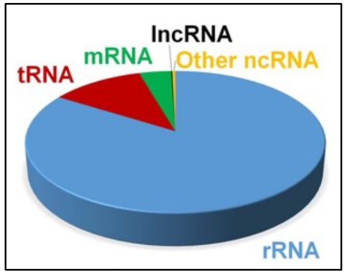
Interrogate protein coding genes

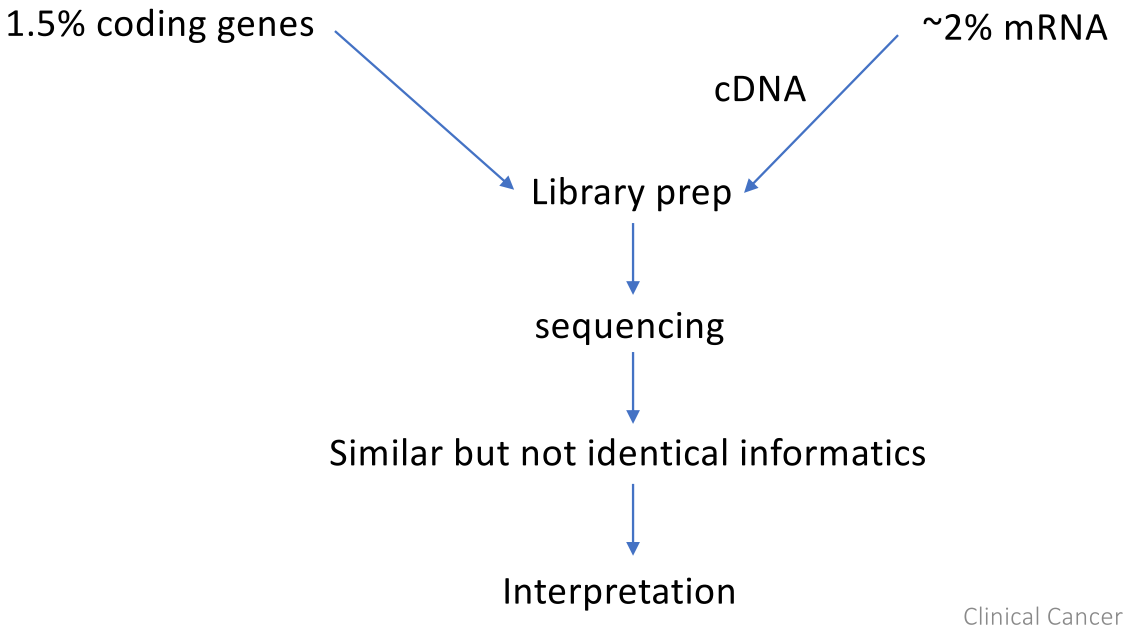
Expressed mutations -> neoantigen -> Immunotherapy
Kinase outlier expression
Gene fusions
Tissue/phenotype specific expression profile
…

Figures: from googling …

Inherent challenges with DNA- and RNA-seq







1.5% coding genes

~2% mRNA

cDNA

Library prep

sequencing

Similar but not identical informatics

Interpretation

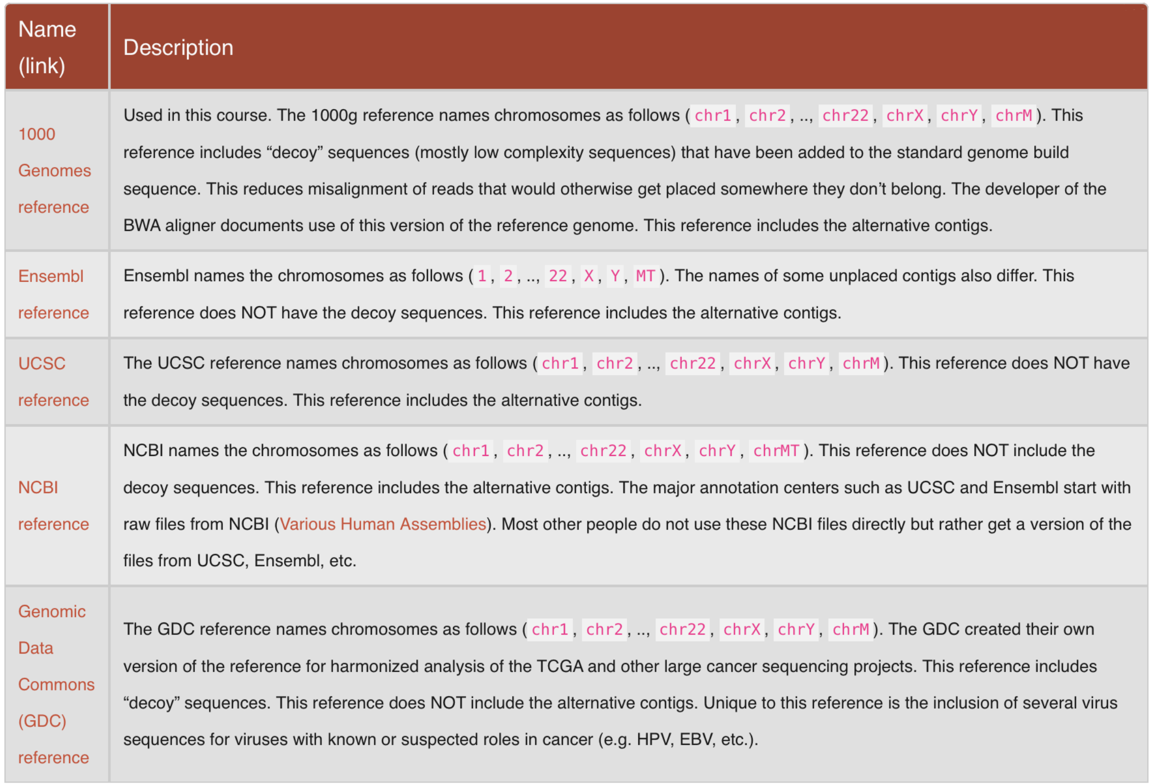
Figures: from googling …

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 comptele 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.).

Table from pmbio.org

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

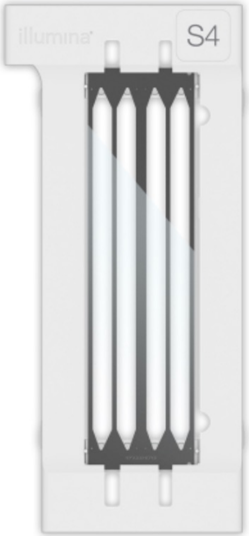
1. Breast cancer cell line (HCC1395) and its matched lymphoblastoid cell line (HCC1395BL).
    1. Cell line is available at the American Type Culture Collection (ATCC) store.
    2. DNA exome-seq, wgs-seq and RNA-seq data.
2. Commercial cfDNA from mCRPC patients (generated at Scilifelab).

Cell line data from pmbio.org

Illumina sequencing

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

NovaSeq 6000

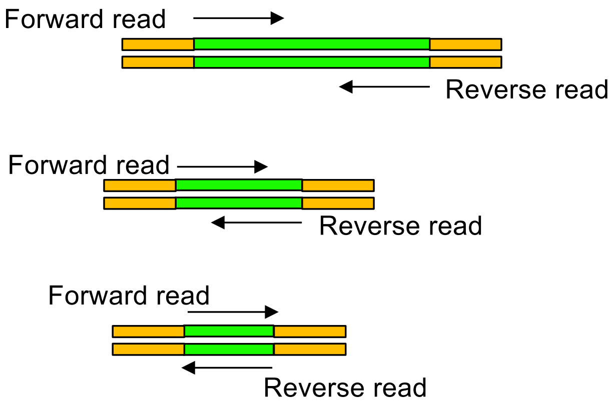


~2000 M read-pairs per lane
2 x 150 bp

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



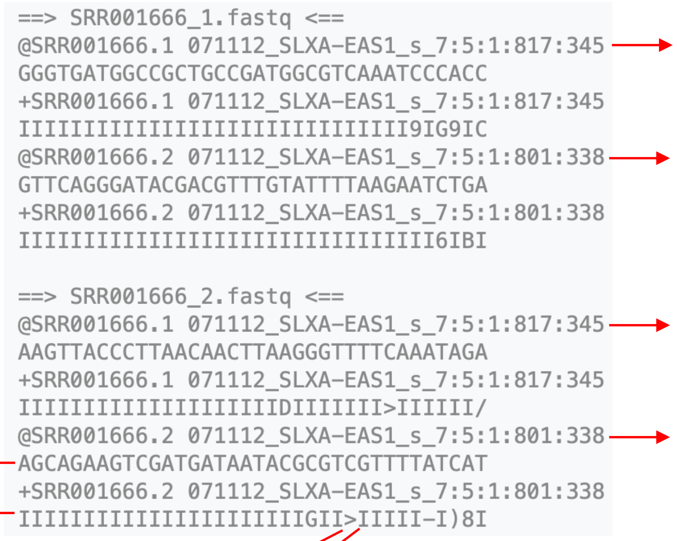
Raw input into all pipelines

Forward read

Reverse read

The sequence

Base qualities



```
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

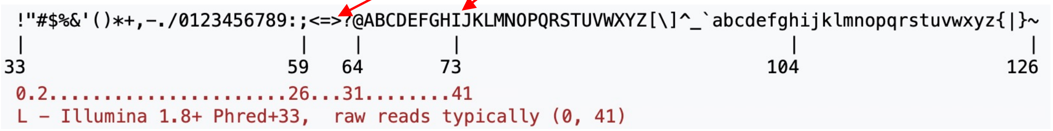
Header

Header

Header

Header

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



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                    |
33                        59   64       73                             104                  126
0.2......................26...31........41
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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

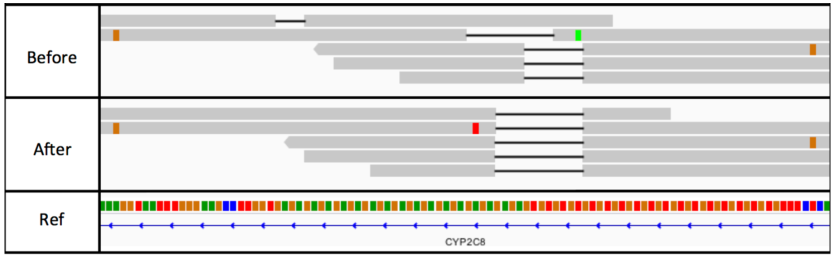
The very basic bioinformatic pipeline

Karolinska Institutet

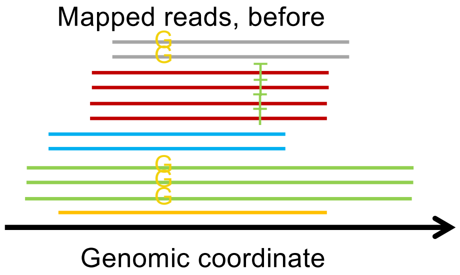
1: Map reads to the human genome

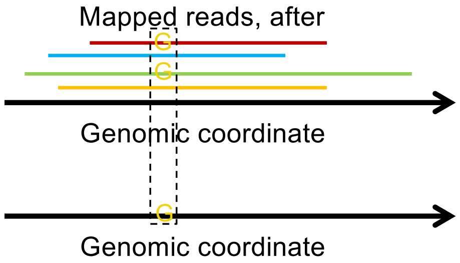
Find the location in the human genome

2: Realign reads



3: Remove duplicates





4: Run variant callers

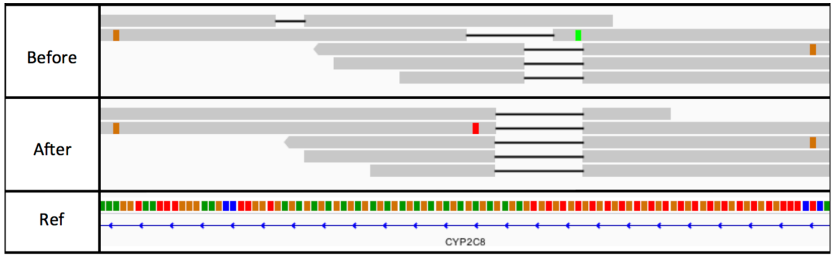
5: Data interpretation for research project ∞

The very basic bioinformatic pipeline

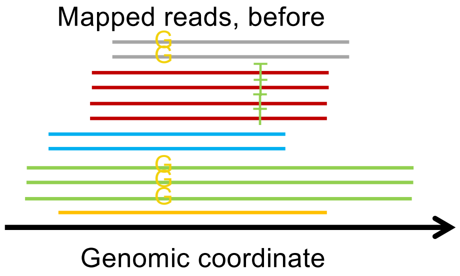
1: Map reads to the human genome

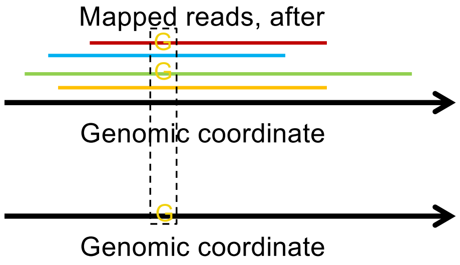
Find the location in the human genome

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

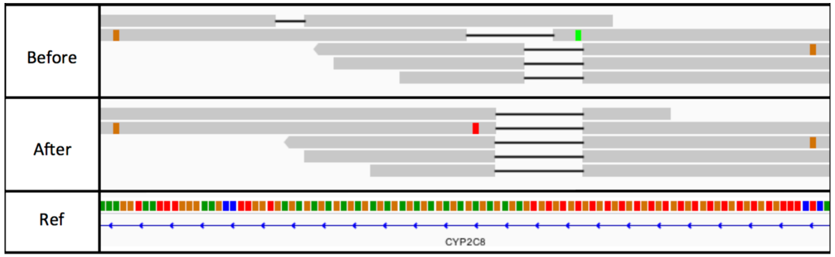
1: Map reads to the human genome

Find the location in the human genome

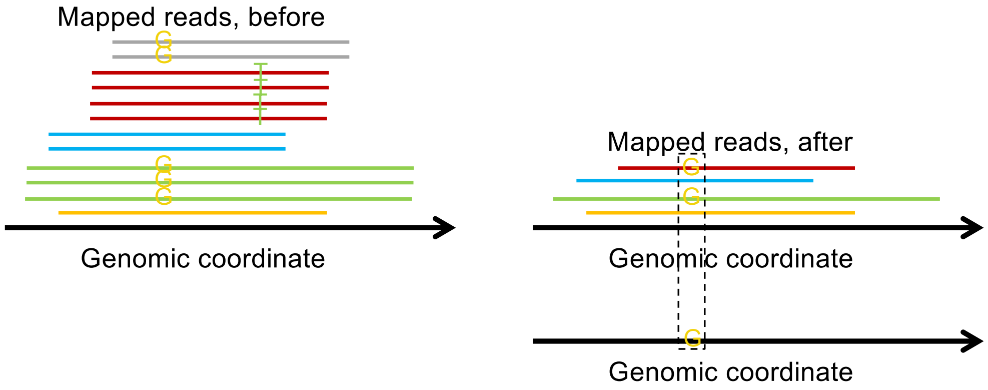
Alignment →

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



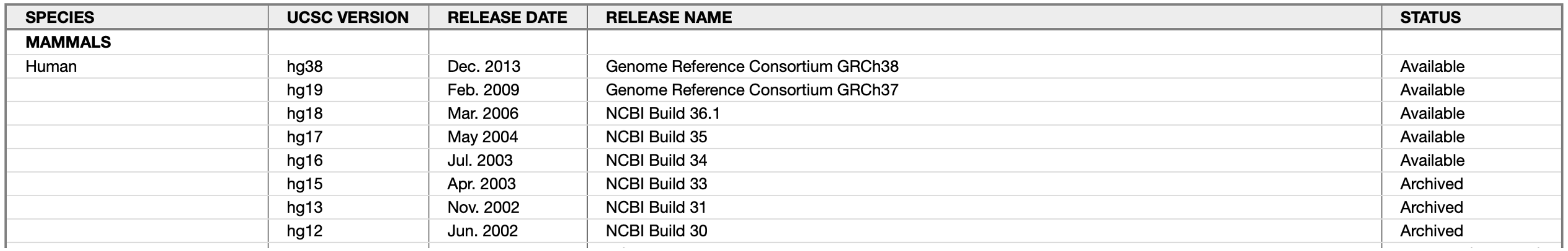
4: Run variant callers

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

The very basic bioinformatic pipeline

- Alignment is like solving a pussle

- 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

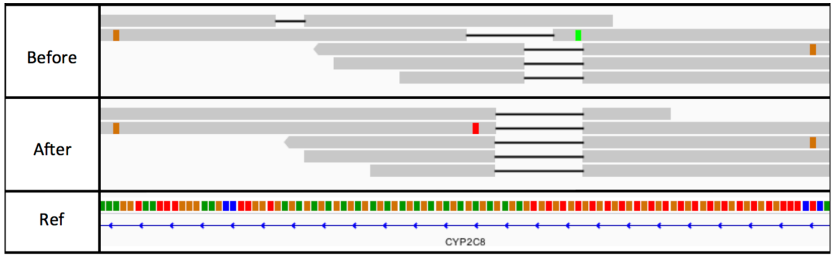
The very basic bioinformatic pipeline

Karolinska Institutet

1: Map reads to the human genome

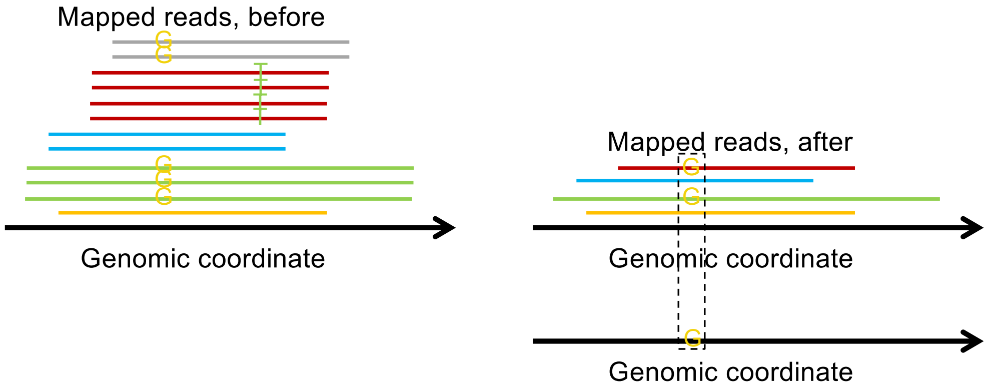
Find the location in the human genome

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

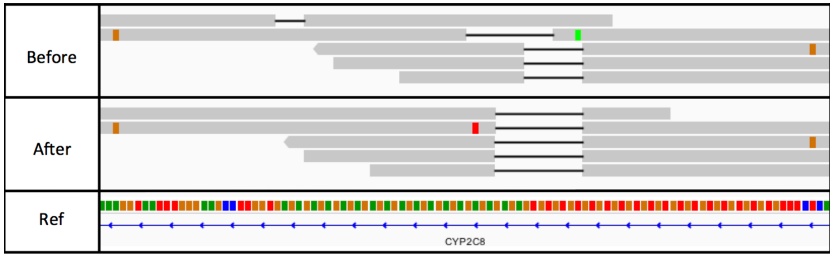
The very basic bioinformatic pipeline

Karolinska Institutet

1: Map reads to the human genome

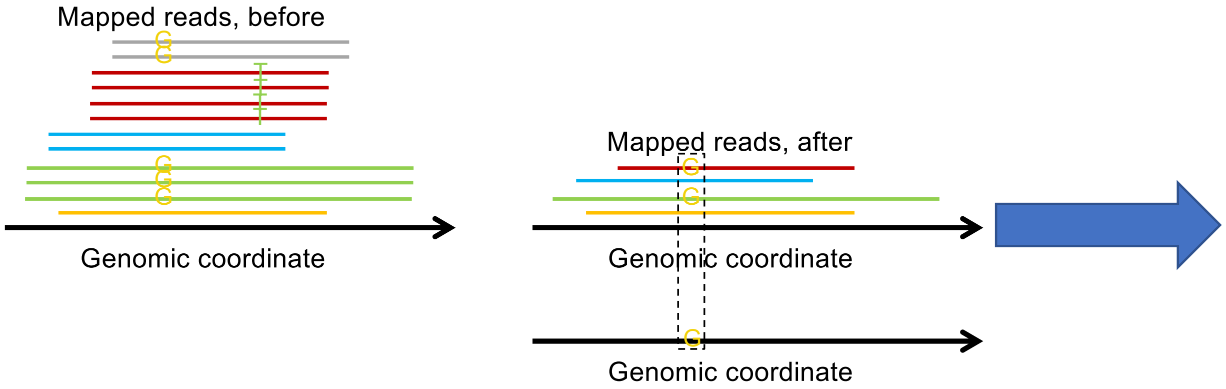
Find the location in the human genome

2: Realign reads



3: Remove duplicates

=



Mapped reads, before

Genomic coordinate

Mapped reads, after

Genomic coordinate

A modified bam-file

4: Run variant callers

Genomic coordinate

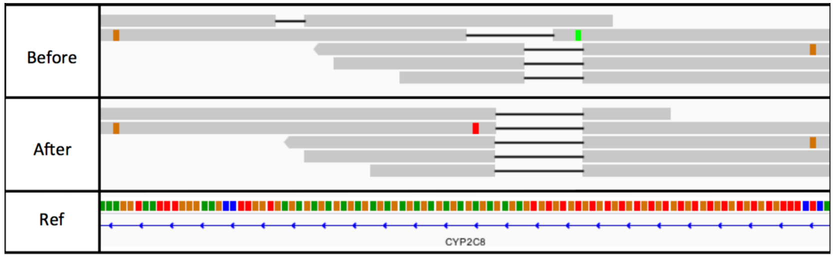
5: Data interpretation for clinical use → Inspect variants manually according to SOP

The very basic bioinformatic pipeline

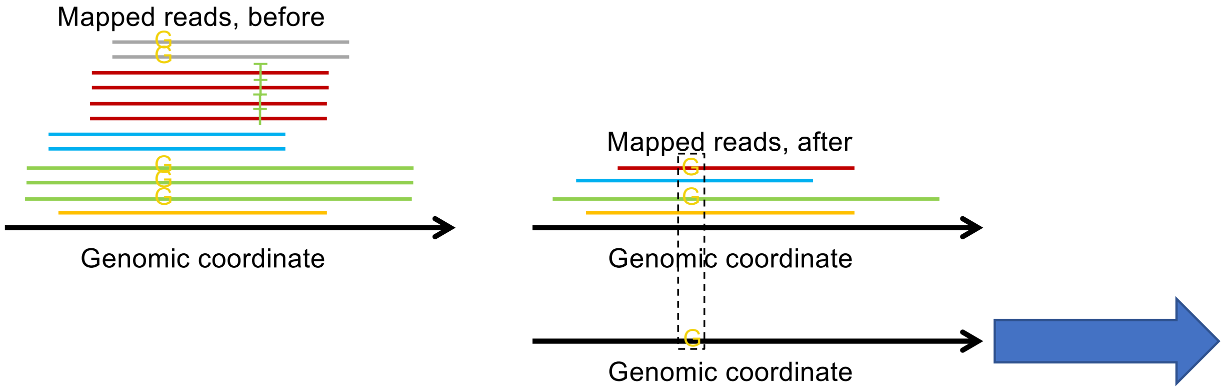
1: Map reads to the human genome

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…

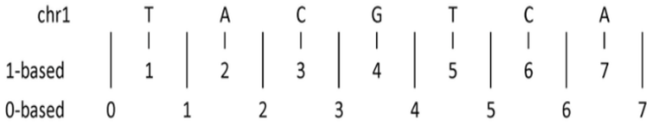
5: Data interpretation for clinical use → Inspect variants manually according to SOP

The very basic bioinformatic pipeline

Watch out for booby traps!

Genomic coordinate systems

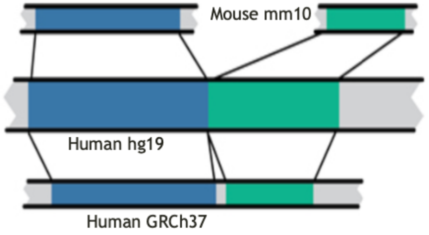
- 1-based vs. 0-based



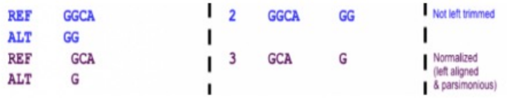
```
chr1        T     A     C     G     T     C     A
1-based     1     2     3     4     5     6     7
0-based  0     1     2     3     4     5     6     7
```

Genome builds

- And annotation builds
- “Liftover” tools



Left-shifted vs right-shifted



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

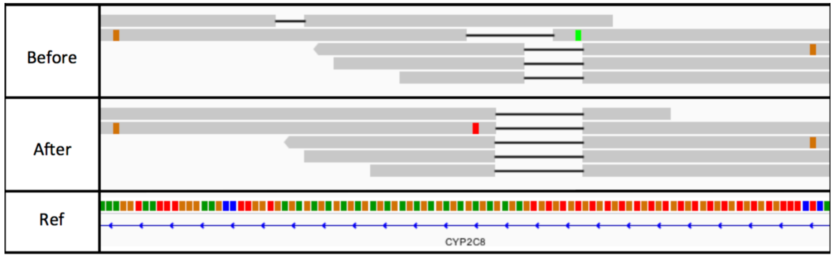
Figs from pmbio.org

The very basic bioinformatic pipeline

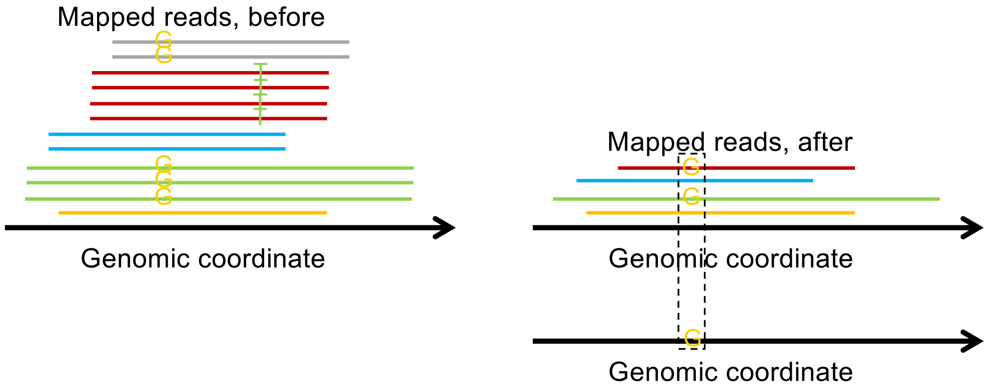
1: Map reads to the human genome

Find the location in the human genome

2: Realign reads



3: Remove duplicates

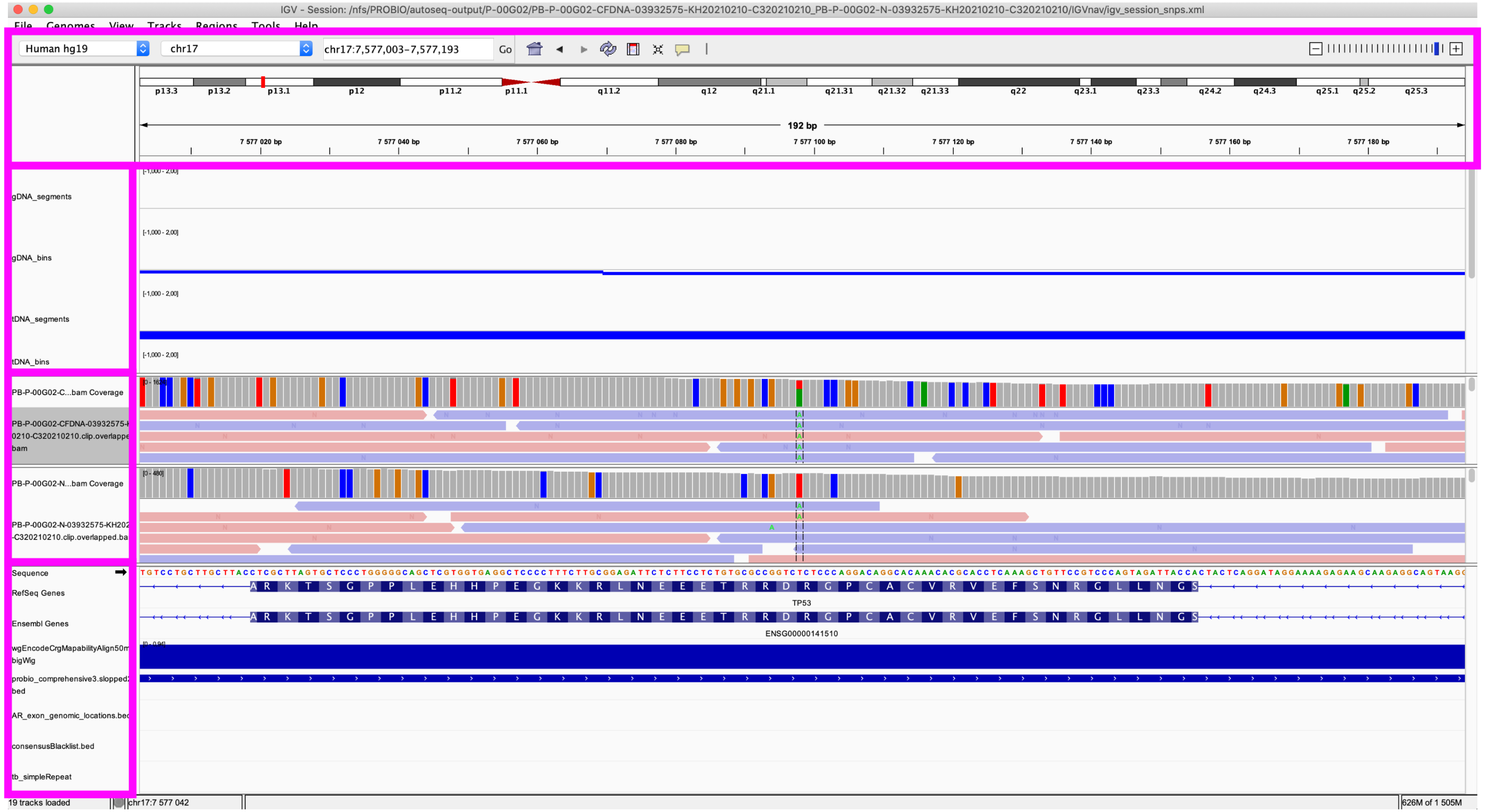


4: Run variant callers

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

IGV galore!

Integrative Genomics Viewer (IGV)


IGV - Session: /nfs/PROBIO/autoseq-output/P-00G02/PB-P-00G02-CFDNA-03932575-KH20210210-C320210210_PB-P-00G02-N-03932575-KH20210210-C320210210/IGVnav/igv_session_snps.xml
File Genomes View Tracks Regions Tools Help
Human hg19
chr17
chr17:7,577,003–7,577,193
Go
p13.3 p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.32 q21.33 q22 q23.1 q23.3 q24.2 q24.3 q25.1 q25.2 q25.3
192 bp
7 577 020 bp
7 577 040 bp
7 577 060 bp
7 577 080 bp
7 577 100 bp
7 577 120 bp
7 577 140 bp
7 577 160 bp
7 577 180 bp
[-1,000 - 2,00]
gDNA_segments
gDNA_bins
tDNA_segments
tDNA_bins
[0 - 1624]
PB-P-00G02-C...bam Coverage
PB-P-00G02-CFDNA-03932575-K
0210-C320210210.clip.overlappe
bam
[0 - 480]
PB-P-00G02-N...bam Coverage
PB-P-00G02-N-03932575-KH202
-C320210210.clip.overlapped.ba
Sequence
RefSeq Genes
TP53
Ensembl Genes
ENSG00000141510
[0 - 0.94]
wgEncodeCrgMapabilityAlign50m
bigWig
probio_comprehensive3.slopped2
bed
AR_exon_genomic_locations.bed
consensusBlacklist.bed
tb_simpleRepeat
19 tracks loaded
chr17:7 577 042
626M of 1 505M

Genome ruler

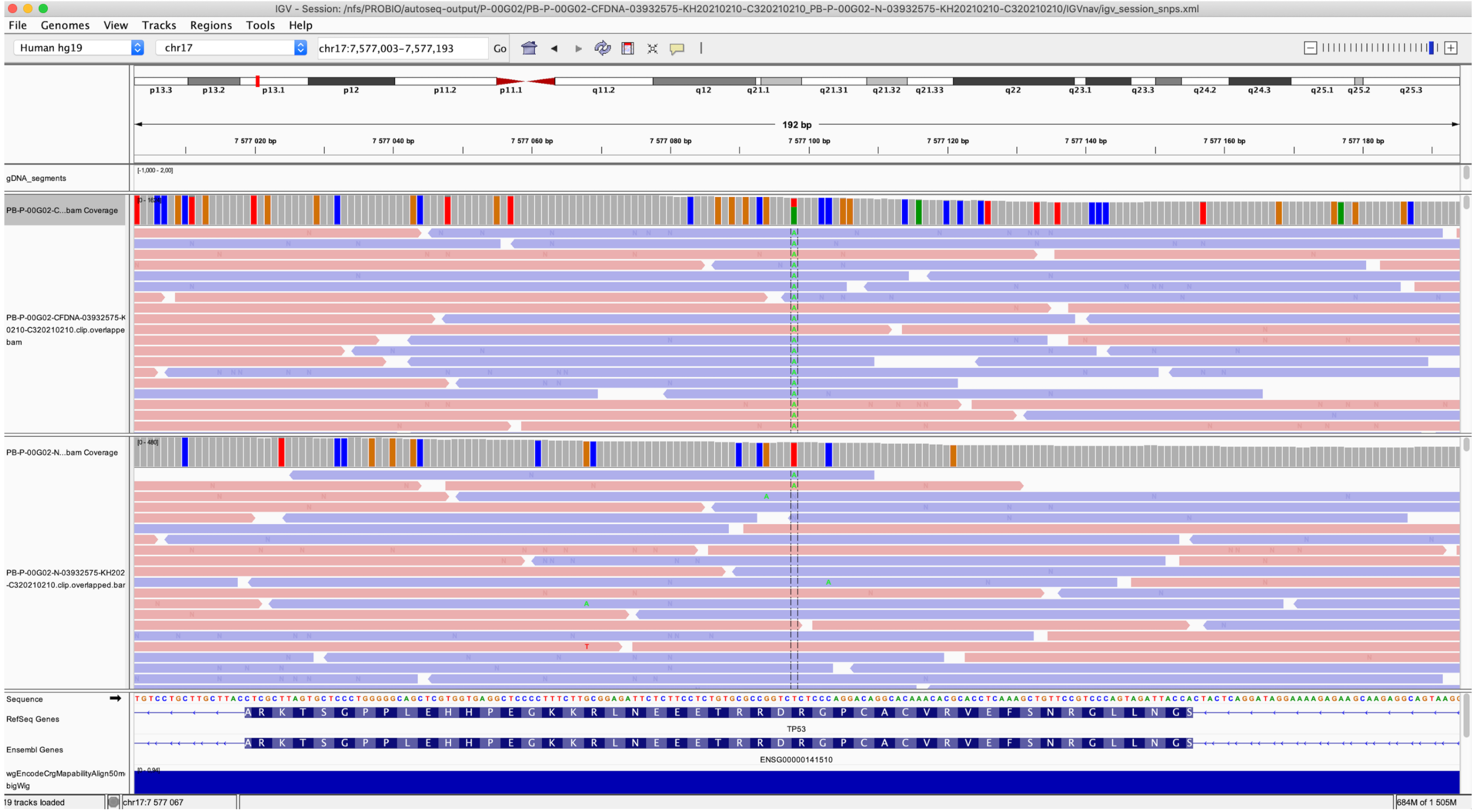
Top panel data tracks

Mid panels, sequencing data tracks

Bottom panel, annotation tracks

Integrative Genomics Viewer (IGV)

Karolinska Institutet



Tumor DNA, coverage track

Tumor DNA, sequencing reads

Normal DNA, coverage track

Normal DNA, sequencing reads

IGV - Session: /nfs/PROBIO/autoseq-output/P-00G02/PB-P-00G02-CFDNA-03932575-KH20210210-C320210210_PB-P-00G02-N-03932575-KH20210210-C320210210/IGVnav/igv_session_snps.xml

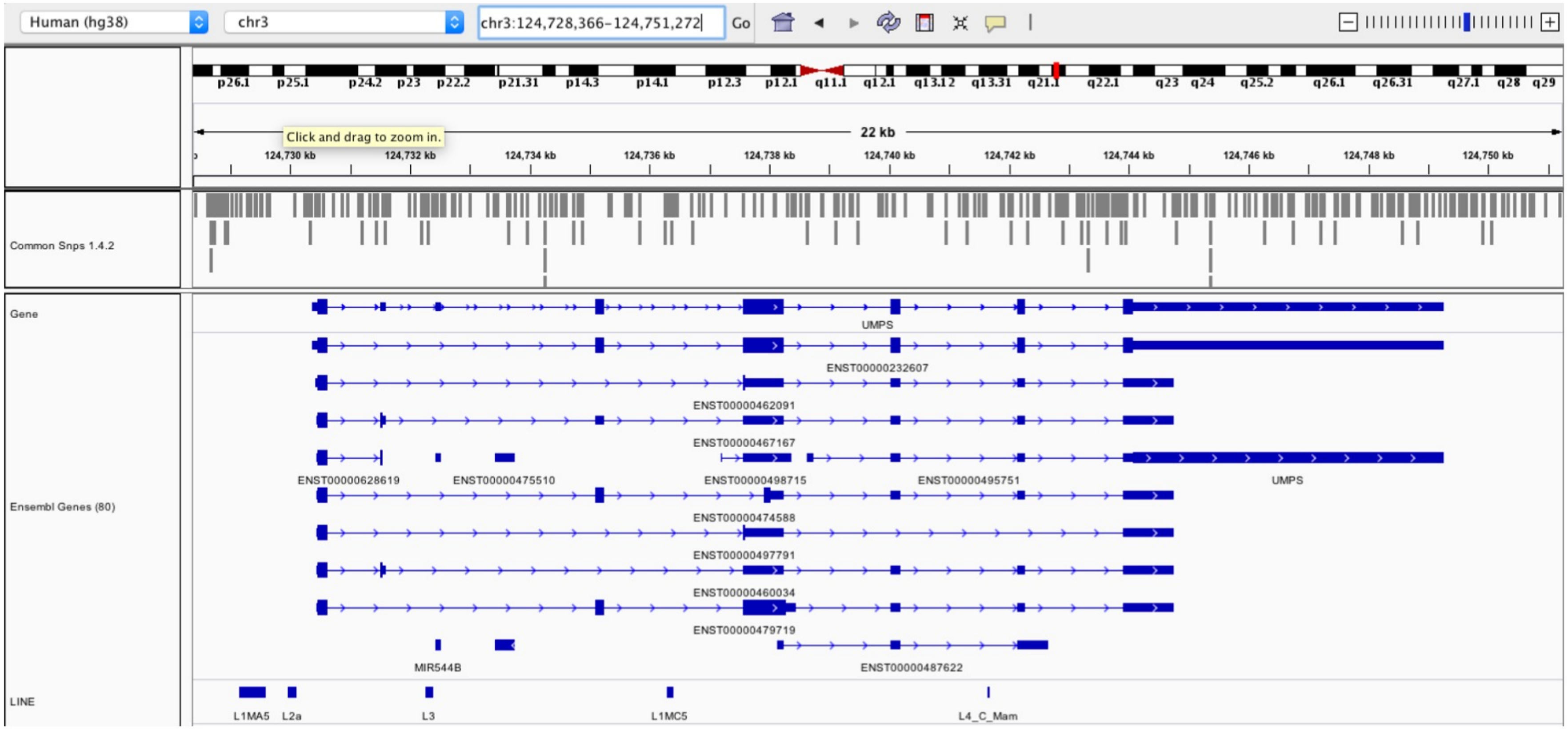
Human hg19

chr17

chr17:7,577,003–7,577,193

File
Genomes
View
Tracks
Regions
Tools
Help
Go
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
192 bp
7 577 020 bp
7 577 040 bp
7 577 060 bp
7 577 080 bp
7 577 100 bp
7 577 120 bp
7 577 140 bp
7 577 160 bp
7 577 180 bp
gDNA_segments
[-1,000 - 2,00]
PB-P-00G02-C...bam Coverage
[0 - 1624]
PB-P-00G02-CFDNA-03932575-K 0210-C320210210.clip.overlappe bam
PB-P-00G02-N...bam Coverage
[0 - 480]
PB-P-00G02-N-03932575-KH202 -C320210210.clip.overlapped.bar
Sequence
RefSeq Genes
Ensembl Genes
wgEncodeCrgMapabilityAlign50m bigWig
[0 - 0.94]
TGTCCTGCTTGCTTACCTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCTCCCCTTTCTTGCGGAGATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAAAGAGAAGCAAGAGGCAGTAAG
A R K T S G P P L E H H P E G K K R L N E E E T R R D R G P C A C V R V E F S N R G L L N G S
TP53
ENSG00000141510
19 tracks loaded
chr17:7 577 067
684M of 1 505M

Integrative Genomics Viewer (IGV)



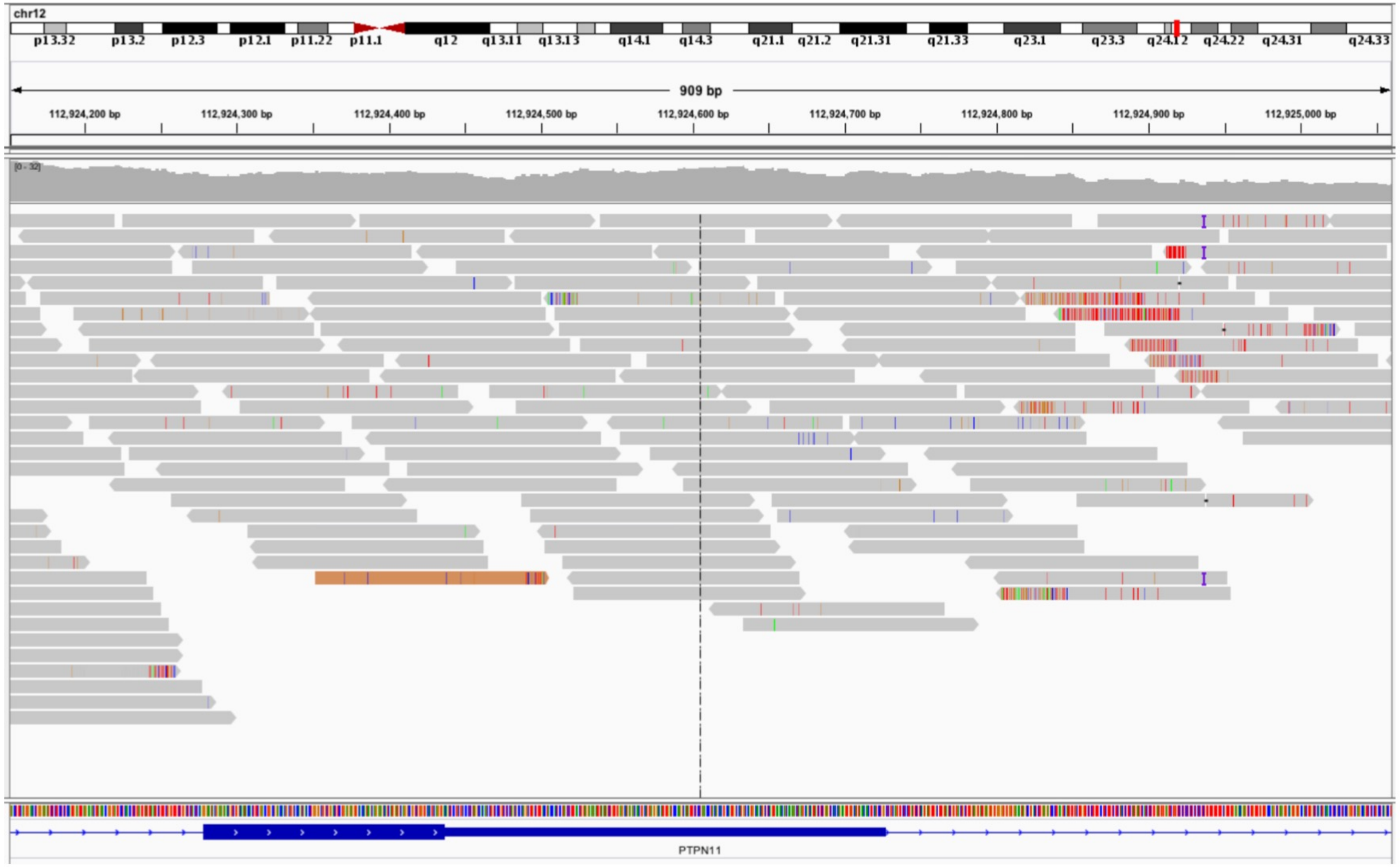
Human (hg38)

chr3

chr3:124,728,366-124,751,272

Figs from pmbio.org

Integrative Genomics Viewer (IGV)



Figs from pmbio.org

Lets get started!