Swiss Institute of Bioinformatics

INTRODUCTION

Cancer variant analysis

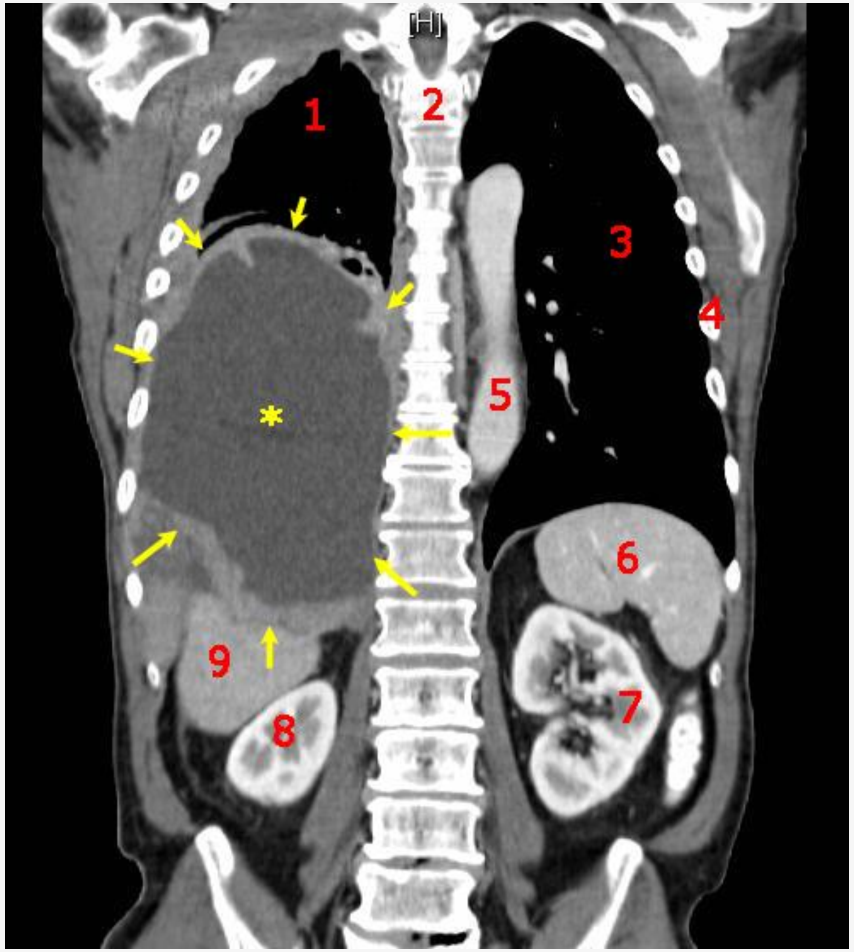
Geert van Geest

13.12.2024

elixir SWITZERLAND

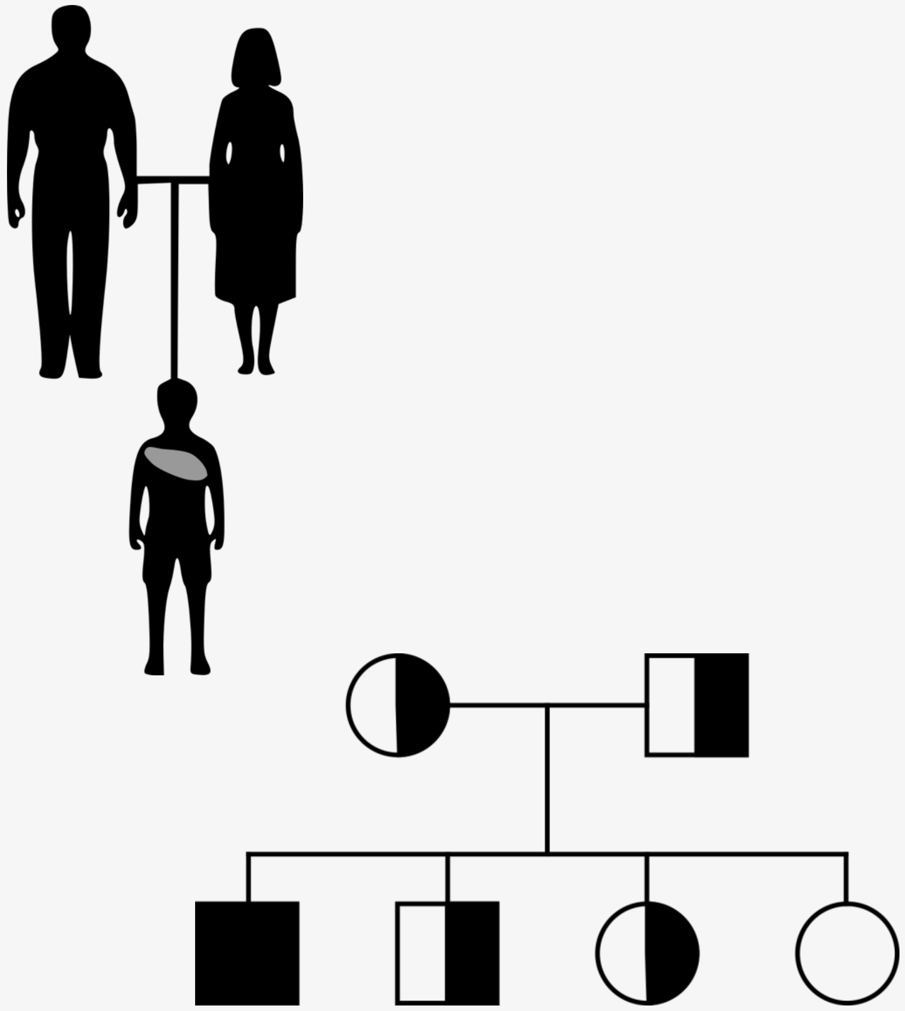
Cancer is a disease of the genome

- Abnormal cell growth
- Potential to invade or spread
- Mostly initiated by **acquired** genomic **mutations** affecting **genes** regulating **cell growth**
- Mutations occur **by chance**, but their rate can be affected by **environmental factors**
- Natural **selection** of **malignant** cells – mutations build up



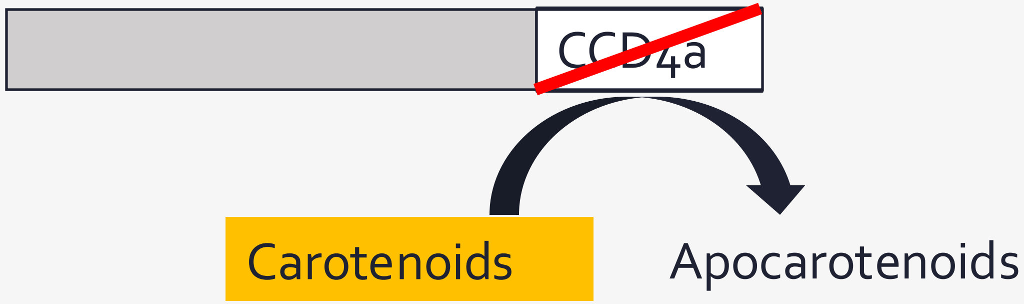
Variants – some definitions

- **Variant**: a difference between DNA
- Caused by a **mutation**: the process of a change in DNA
- **Somatic variant**: occurs only in a part of an organism
- **Germline variant**: can be passed on the next generation
- **Polymorphism**: variant that is common in a population



Variant – an example

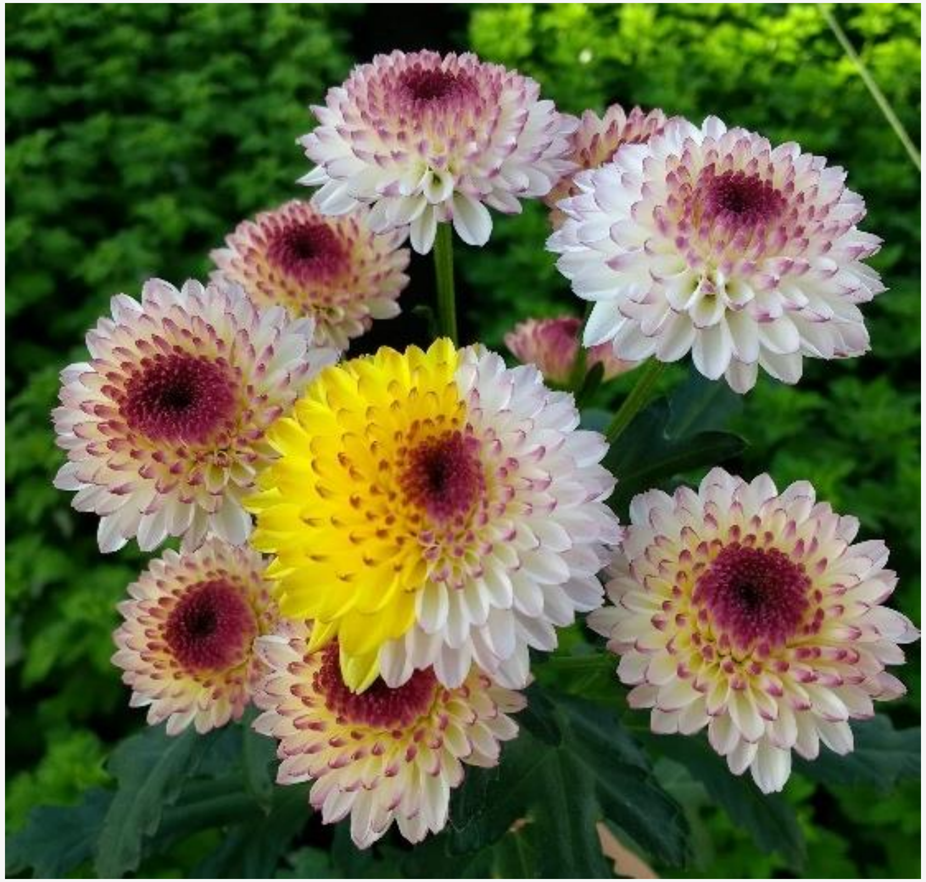
- **Mutation**: the change in DNA that caused the petals to turn yellow
- **Variant**: the difference between the DNA in the yellow petals and the wild type



~~CCD4a~~

Carotenoids

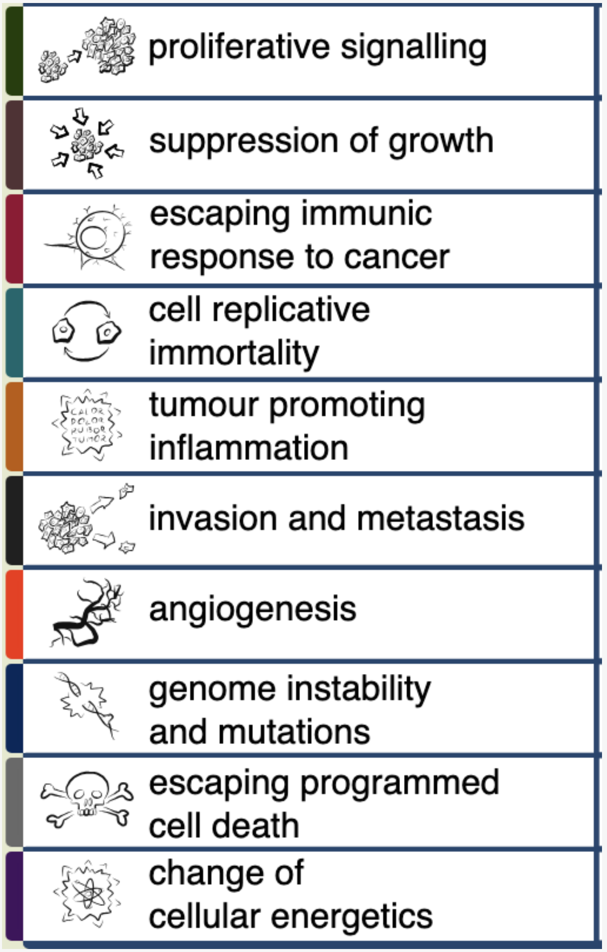
Apocarotenoids



Question

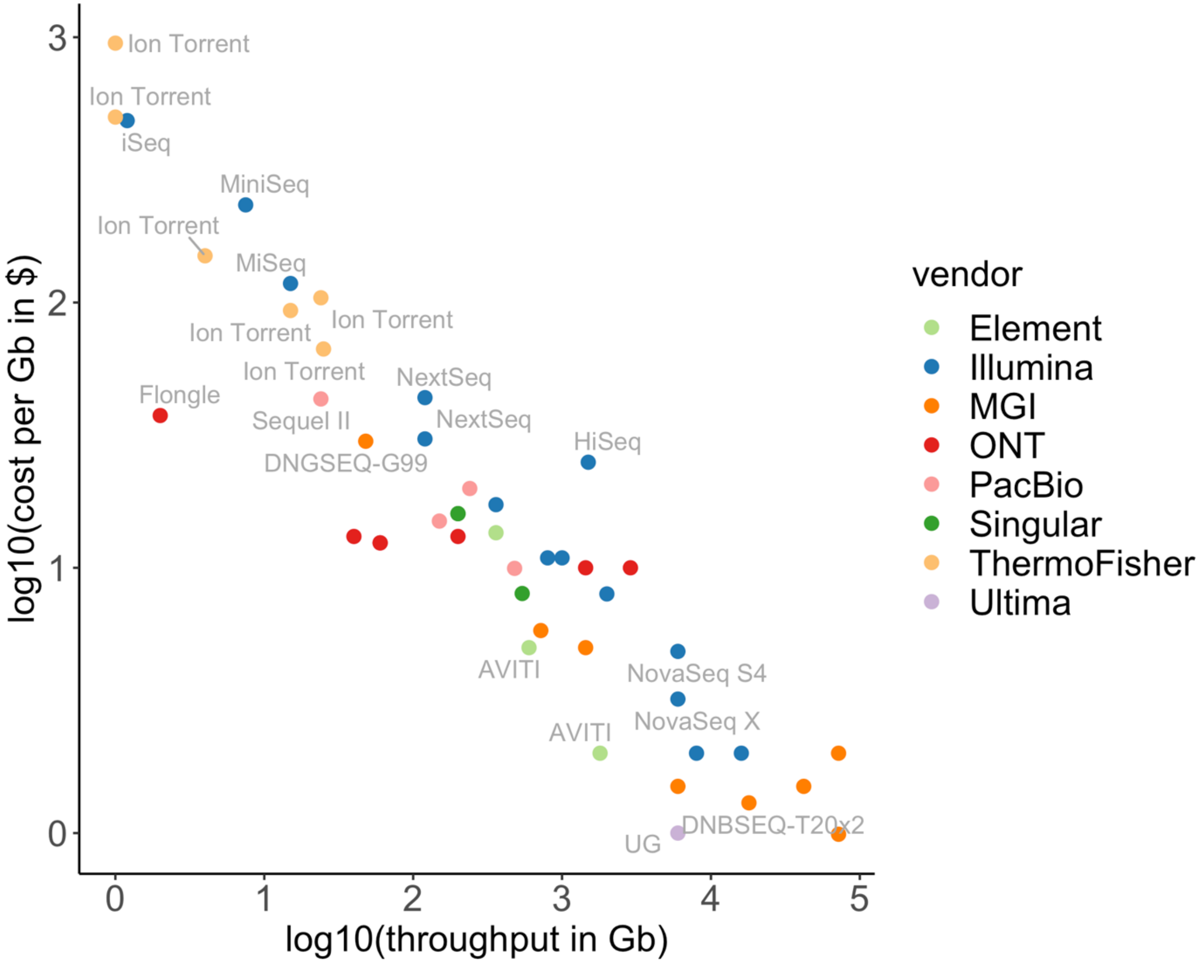
Types of mutations in cancer

- **Small mutations** (SNVs INDELs)
- **Copy Number Variation** (CNV)
  - Loss of heterozygosity (LOH)
- **Structural variation**
  - Fusion transcripts


proliferative signalling
suppression of growth
escaping immunic response to cancer
cell replicative immortality
tumour promoting inflammation
invasion and metastasis
angiogenesis
genome instability and mutations
escaping programmed cell death
change of cellular energetics

Mutation detection - sequencing

- Whole genome sequencing
- Bait capture
  - Whole exome sequencing
  - Custom panels
- Sequencing methods:
  - Short reads (Illumina, MGI, Element)
  - Long reads (PacBio, ONT)

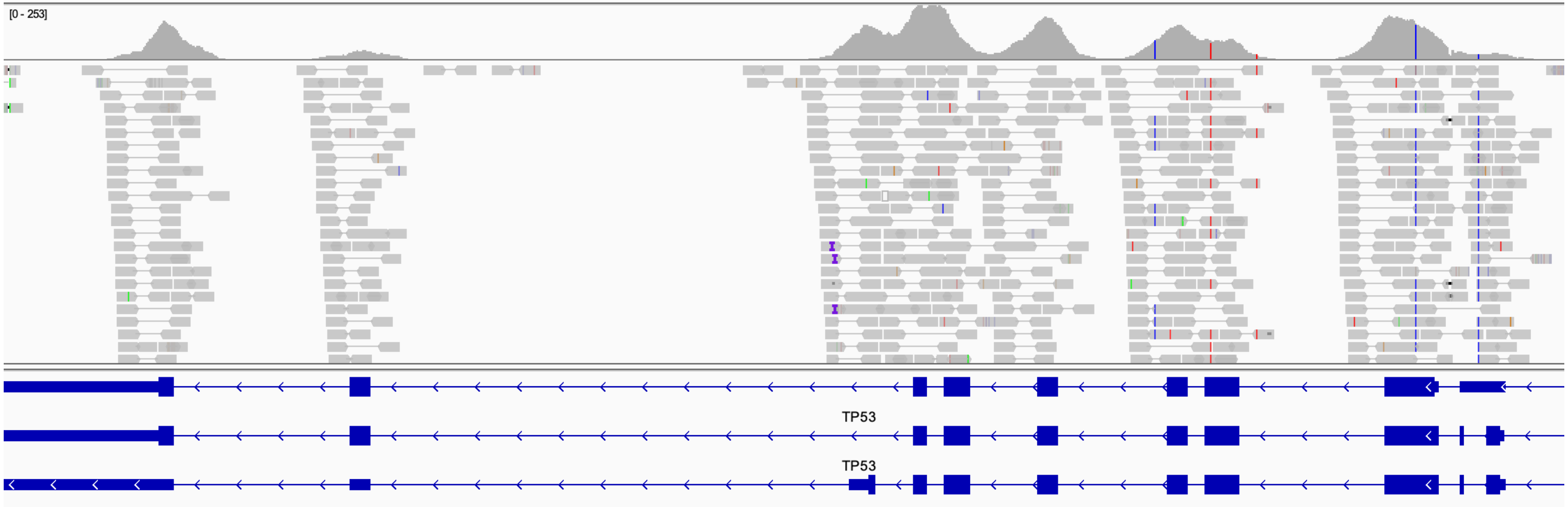


2 questions on NGS

Bait capture

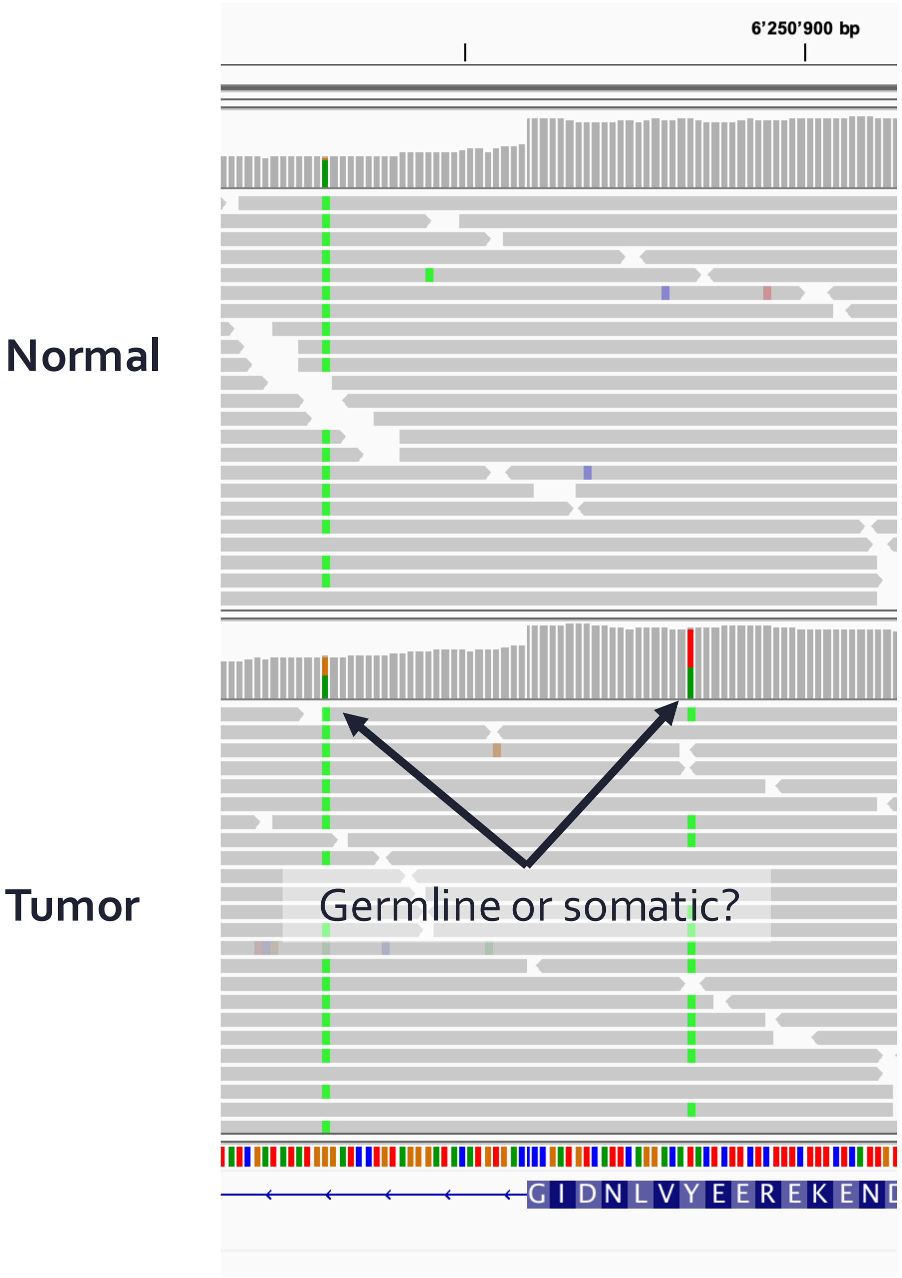
- RNA/DNA baits are designed for regions of interest
- Library prep as usual
- Capture with biotinylated baits

- (Much) lower sequencing and processing costs :
  - WES 100x: 25 M 2 x 100 bp
  - WGS 30x: 450 M 2 x 100 bp
- Not sequenced regions can contain valuable information, e.g. structural variation



Experimental design

- Tumor tissue is often not purely ‘tumor’. It contains e.g. immune cells.
- How to discriminate between germline and somatic variation?
- Per patient two samples:
  - Tumor
  - Normal (e.g. blood)



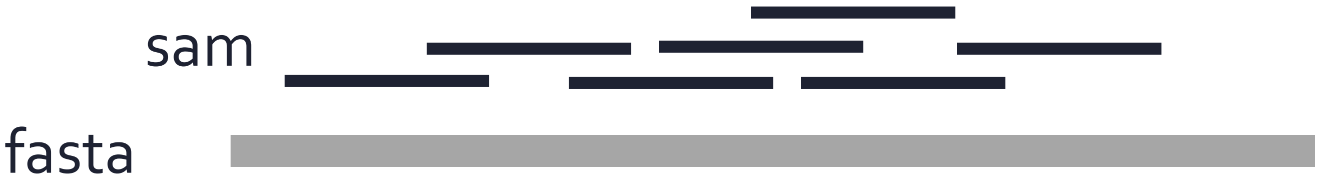
Downstream processing

- **Adapter removal**
  - e.g. fastp
- **Alignment**
  - bwa-mem or dragen
- **Adding read groups**
  - during alignment or e.g. samtools addreplacerg

- **Base quality score recalibration**
  - GATK framework
- **Deduplication**
  - gatk MarkDuplicates
- **Coverage/enrichment evaluation**
  - e.g. mosdepth or gatk CollectHsMetrics

The sam/bam file format

- Usage: store alignments
- bam = compressed sam
- Important format to understand!

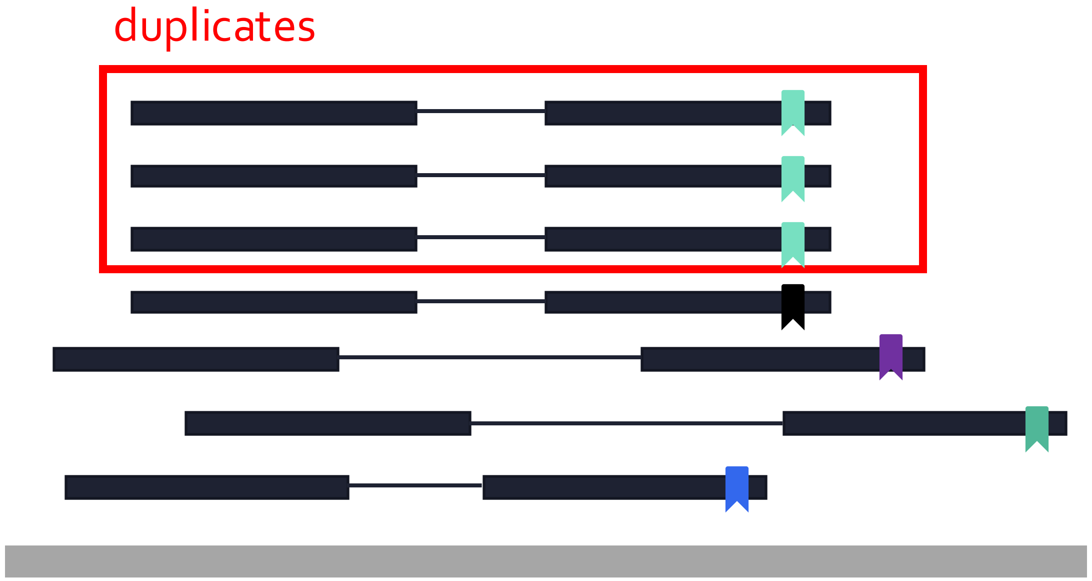


```
@HD    VN:1.6  SO:coordinate
@SQ    SN:chr6 LN:170805979
@SQ    SN:chr17       LN:83257441
@PG    ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem ref_genome.fa normal_R1.fastq.gz normal_R2.fastq.gz
@PG    ID:samtools    PN:samtools    PP:bwa  VN:1.21 CL:samtools sort

HWI-ST466:135068617:C1TD1ACXX:7:1114:9733:82689 163    chr6    60001   60     100M    =     60106   205    GATCTTATATAACTGTGAGATTAATCTCAGATAATGACACAA
CCCFFFFFHHHHHJJIJHIJJJEIJIJJJJJIJJIJJJJIJJ    NM:i:0  MD:Z:100 MC:Z:100M AS:i:100 XS:i:0

HWI-ST466:135068617:C1TD1ACXX:7:1303:2021:90688 99    chr6    60001   60     100M    =     60104   203    GATCTTATATAACTGTGAGATTAATCTCAGATAATGACACAA
@CCFFFFFFFGGHHIIFHGIEGHJICEHIJJIIJIJEGHIIJJ    NM:i:0  MD:Z:100 MC:Z:100M AS:i:100 XS:i:0

HWI-ST466:135068617:C1TD1ACXX:7:2304:7514:30978 113    chr6    60001   60     2S98M   =     61252   1194    TAGATCTTATATAACTGTGAGATTAATCTCAGATAATGACAC
DEDDCEEFEEEEFFFFFFFHHHHHHIJJJHIGJJJJIHHFGGJ    NM:i:0  MD:Z:98 MC:Z:60S40M AS:i:98 XS:i:0
```

More about the bam file format at the course NGS – Quality Control, Alignment, Visualization

Marking duplicates

- Variant caller assumes each fragment is **unique representation** of the genome
- PCR/optical **duplicates** are **not**
- **Marking** PCR/optical duplicates **improves** variant calling **accuracy**
- Adding Unique Molecular Identifiers (**UMI**) strongly improves **duplicate marking accuracy**



Read groups

Have multiple groups of reads in a bam file

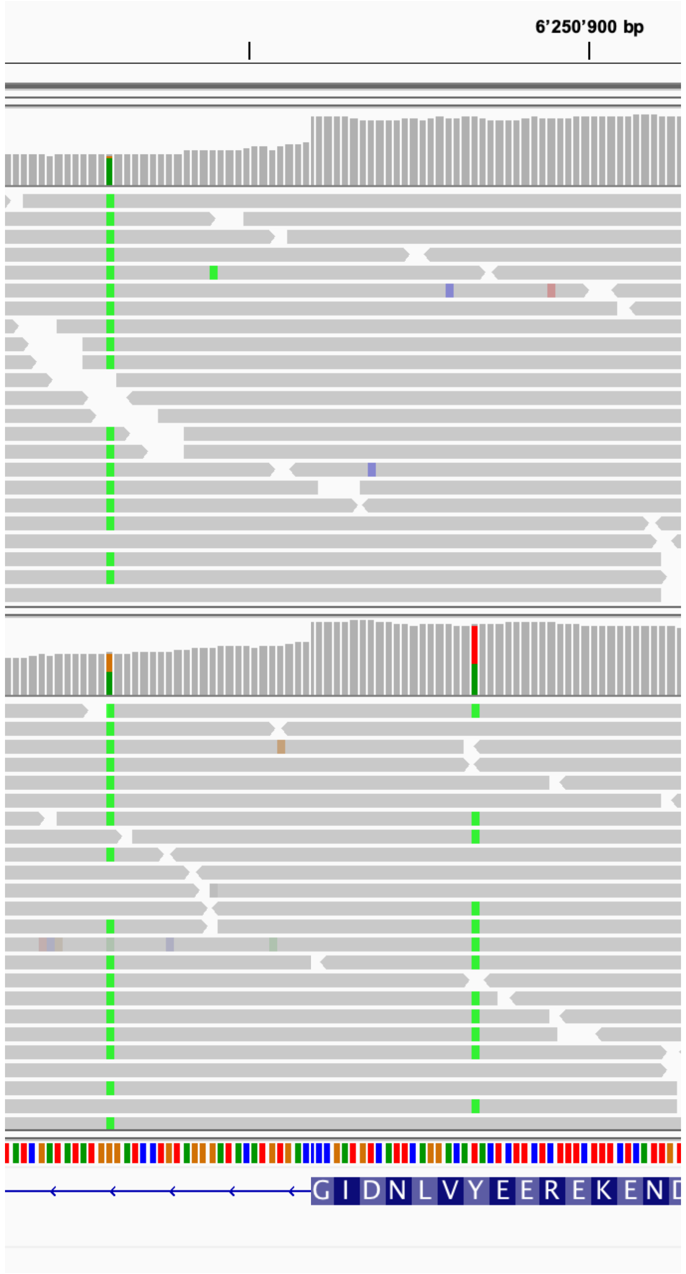
Add metadata to alignments:

- Samples
- Libraries
- Lanes
- ..

```
@RG ID:rg1 LB:lib1 SM:sampleA
@RG ID:rg2 LB:lib2 SM:sampleA
read1    456345   chr20    RG:Z:rg1
read2    456348   chr20    RG:Z:rg2
read3    456357   chr20    RG:Z:rg2
read4    456359   chr20    RG:Z:rg1
```

Small variants – different assumptions

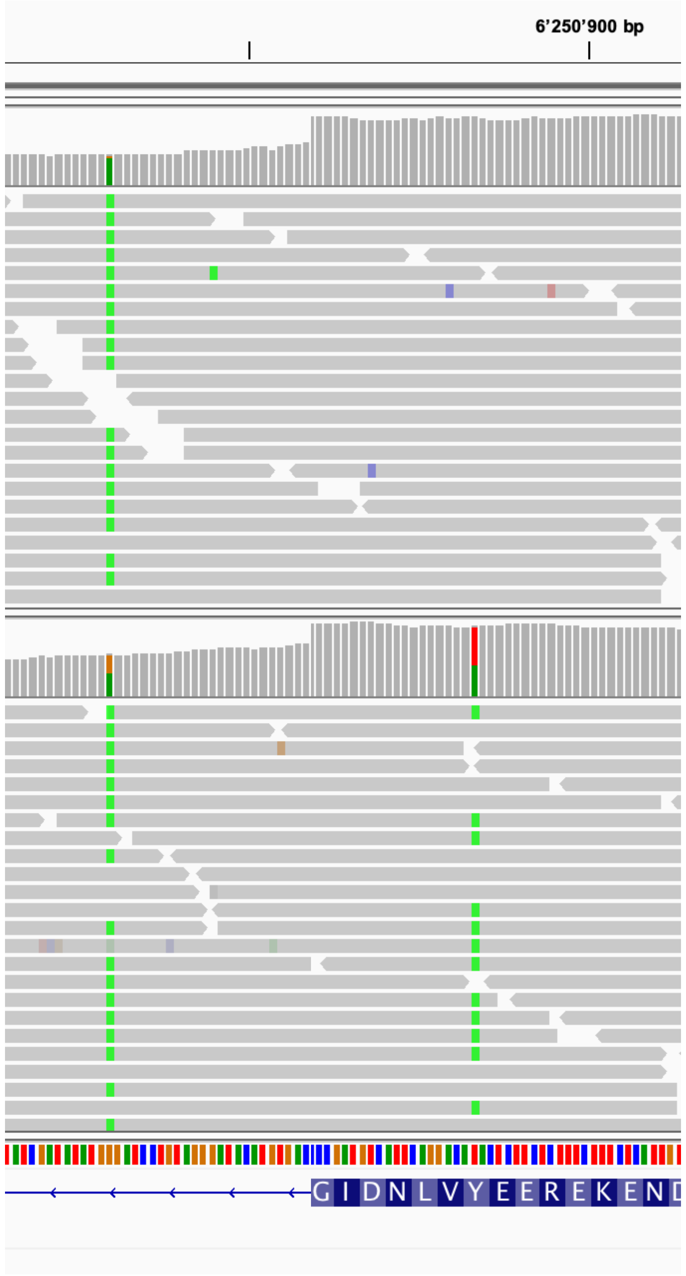
- Small variants: SNVs and INDELs
- Germline variants have **assumed variant allele frequencies**:
  - 100% if homozogous alternative
  - 50% if heterozygous
- These do **not hold** for somatic variants – heavily relies on tumor purity and heterogeneity
- **Biased error** strongly influences accuracy



Small variants - likelihood

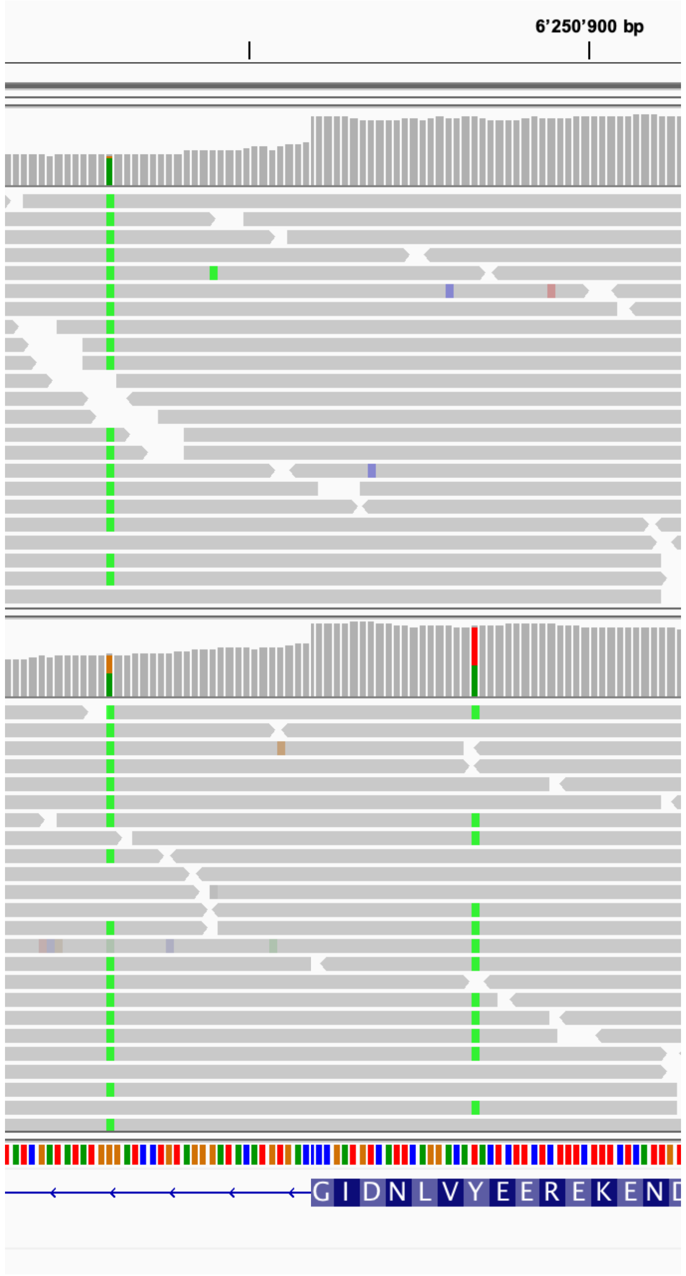
Estimating likelihood and filtering can be based on:

- **Sequencing error**: base qualities and variant allele frequency
- **Technical artifacts**: 'panel of normals' – information about expected non-random errors/artifacts
- Possibility of being a **germline variant**:
  - Germline call of normal sample
  - Databases of known germline variants, e.g. gnomAD



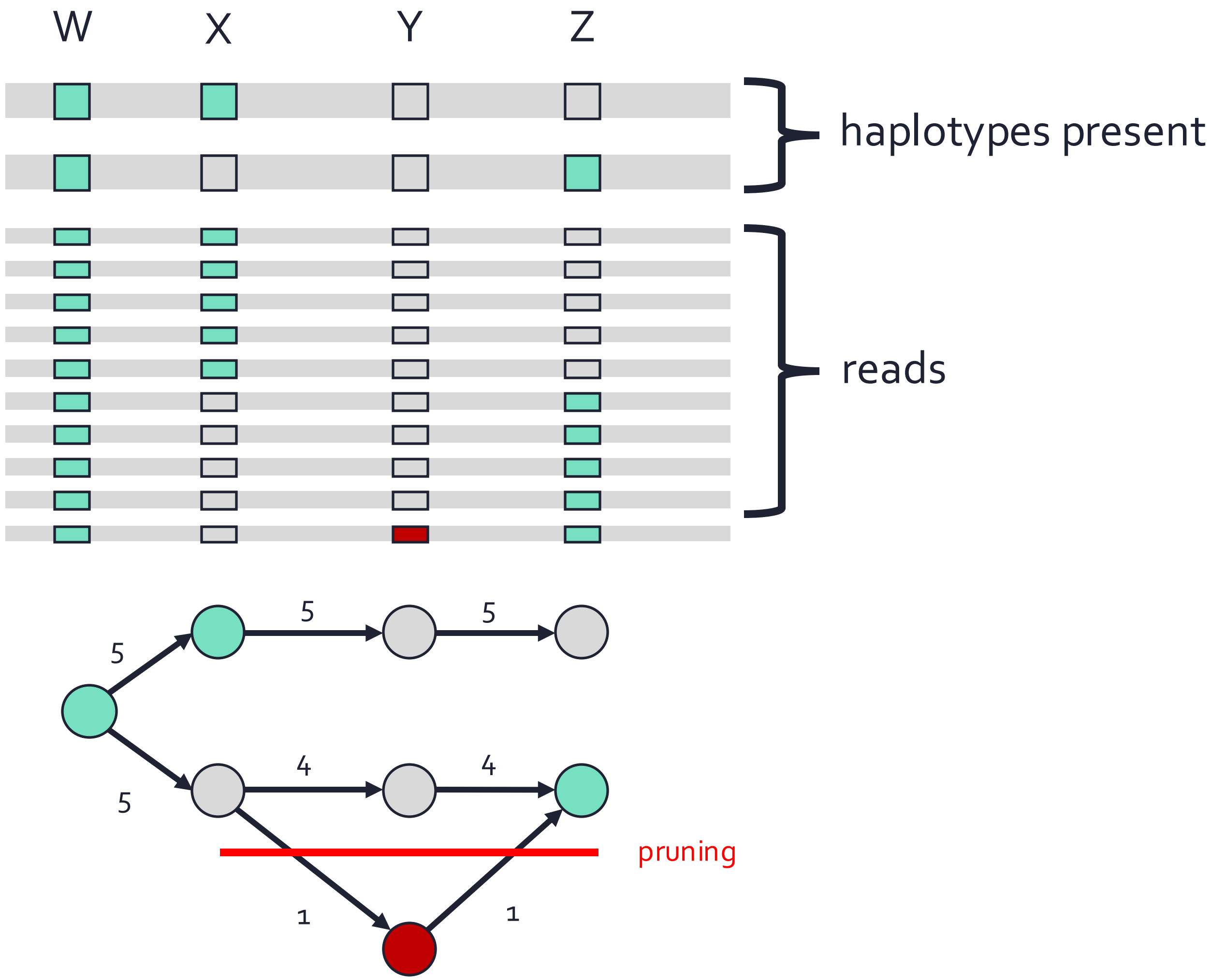
Small variants – GATK workflow

- Mutect2 – haplotype aware variant calling
- FilterMutectCalls – filtering variants based on:
  - Sequencing error
  - Technical artifacts
  - Germline variants



Small variants - haplotypes

- **Haplotype**: alleles in phase on the same chromosome
- Helps to detect sequencing error
- **Germline**: maximum 2 haplotypes
- **Tumor**: can have more than 2 hapltoypes

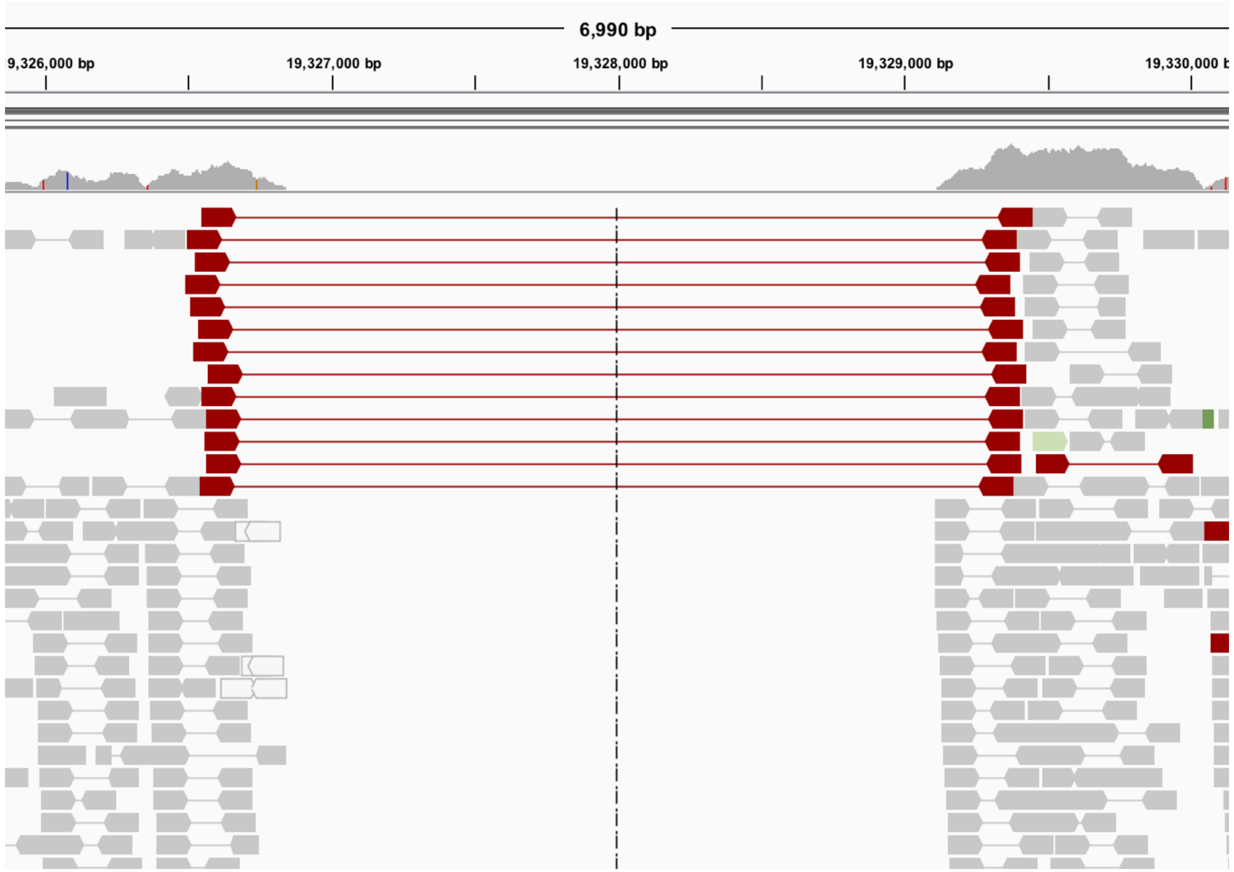


The VCF file format

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,>
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">

#CHROM  POS      ID         REF  ALT     QUAL  FILTER  INFO                    FORMAT       NA00001         NA00002
20      14370    rs6054257  G    A       29    PASS    NS=2;DP=9;AF=0.25       GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51
20      17330    .          T    A       3     q10     NS=2;DP=8;AF=0.25       GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3
20      1110696  rs6040355  A    G,T     67    PASS    NS=2;DP=6;AF=0.5,0.5    GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2
20      1230237  .          T    .       47    PASS    NS=2;DP=11              GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51
20      1234567  microsat1  GTC  G,GTCT  50    PASS    NS=2;DP=6               GT:GQ:DP     0/1:35:4        0/2:17:2
```

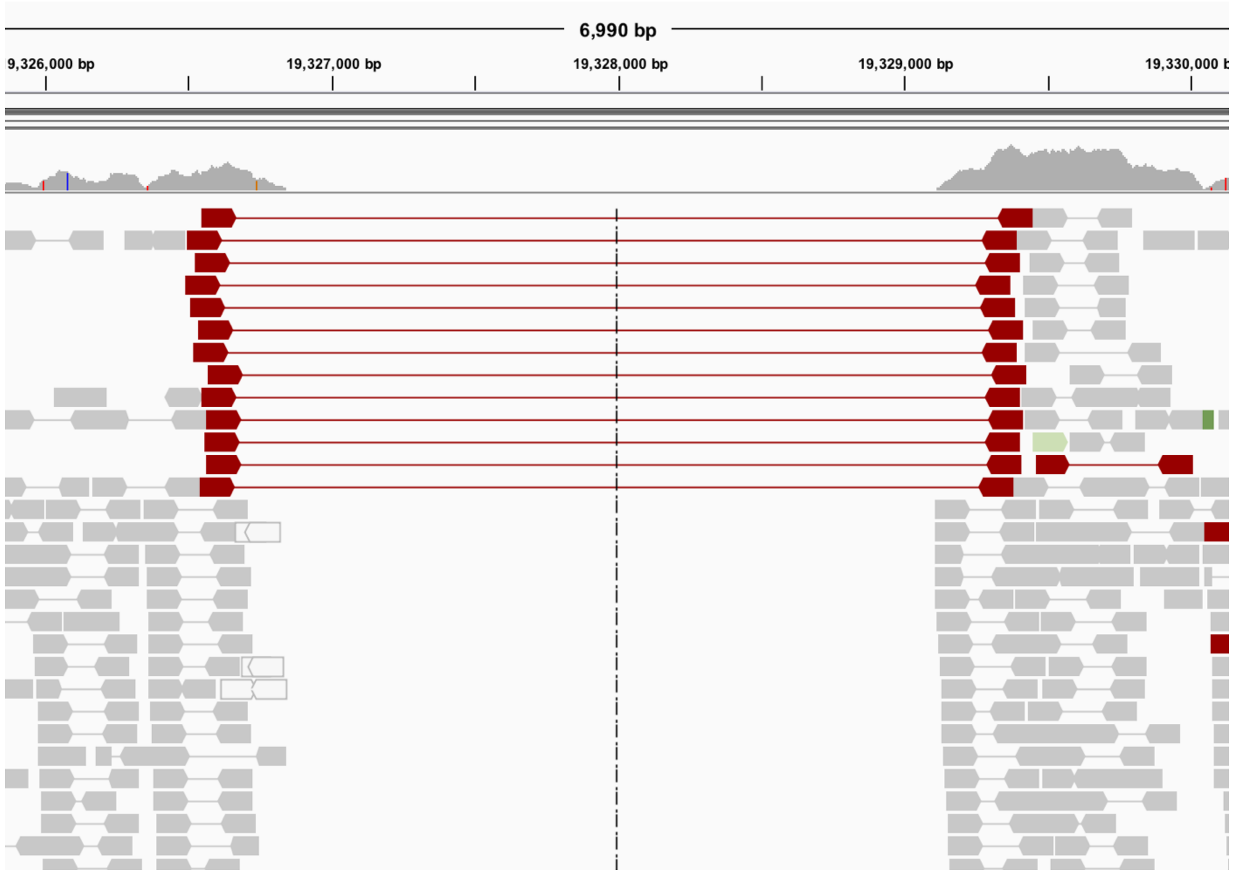
Question

Question


6,990 bp
9,326,000 bp
19,327,000 bp
19,328,000 bp
19,329,000 bp
19,330,000 b

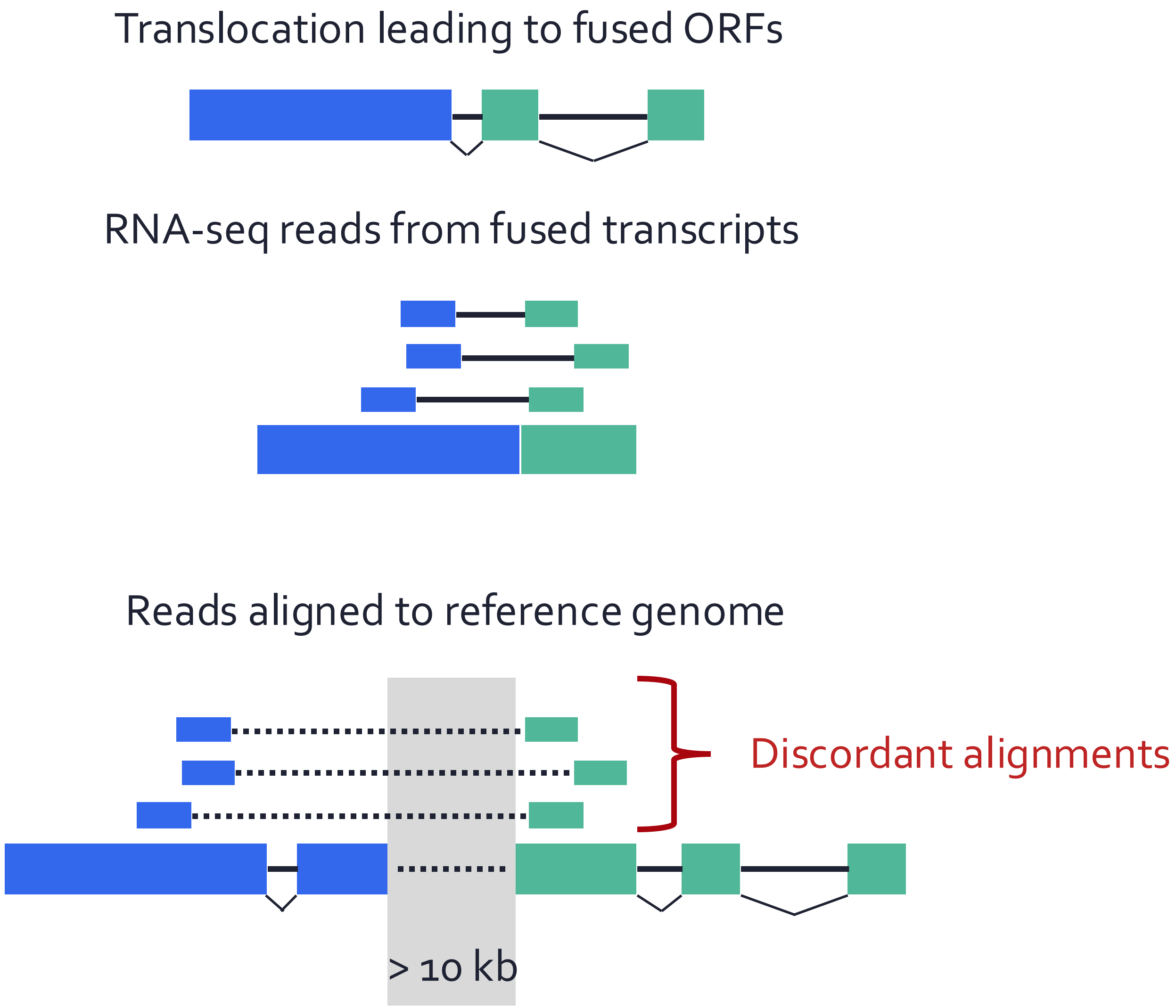
Structural variation

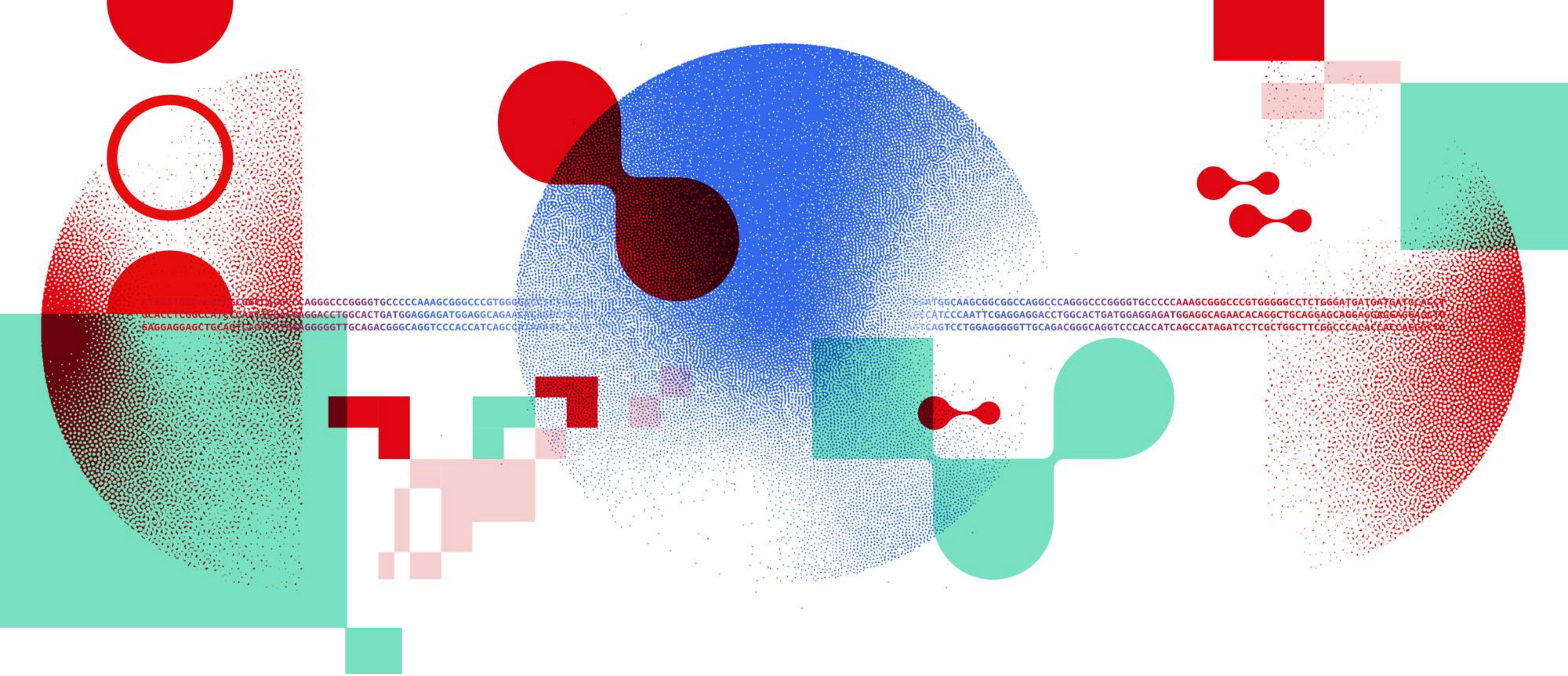
- Large INDELs
- Translocations
- Inversions
- Tools: manta, tiddit
- If this causes a loss or multiplication: **copy number variation** (CNV)
- Translocations/inversions can lead to **gene fusions**



Gene fusion

- Fusion of ORF at the level of the genome
- Results in fused transcripts
- Can be estimated based on:
  - WGS
  - WES
  - RNA-seq
- Detection by discordant alignments


Translocation leading to fused ORFs
RNA-seq reads from fused transcripts
Reads aligned to reference genome
Discordant alignments
> 10 kb



Time for exercises!

DATA SCIENTISTS FOR LIFE

sib.swiss