04. Variant Calling

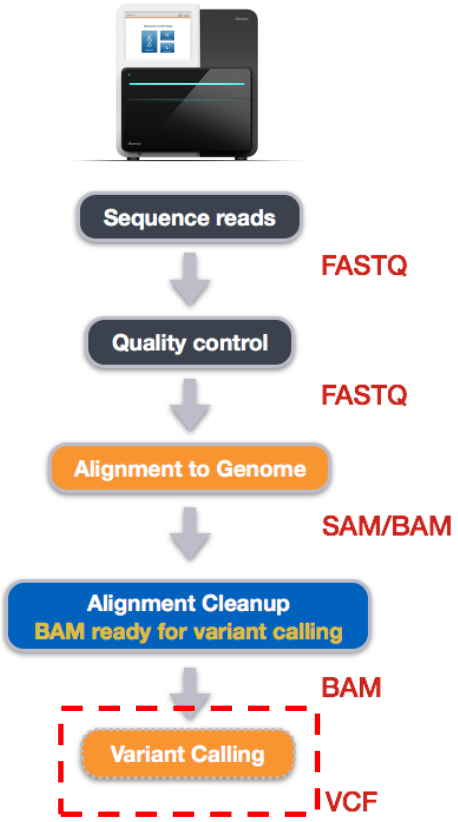
Erik Dassi & Davide Bressan

Genomics Technologies Lab

A Quick Recap

In the last lecture, we aligned the reads on the reference genome, and we obtained a **BAM** file

At this point, we are ready to move on and perform what is called **variant calling**



Variant Calling

Variant calling is a computational process that identifies differences between a sample's DNA sequence and a reference genome. Its applications are broad:

- **Disease** Identification: Determines genetic mutations causing genetic disorders.
- Personalized **Treatment**: Tailors treatments to individual genetic makeups.
- **Cancer** Insights: Identifies cancer-driving somatic mutations.

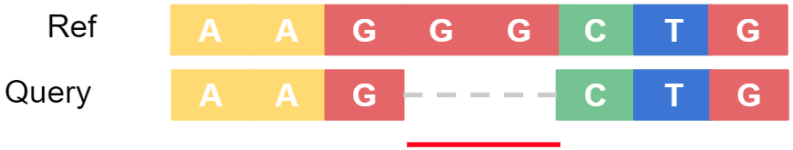
Variant Calling

Genetic variants are often separated into two categories: sequence variants, and structural variants.

Sequence Variants

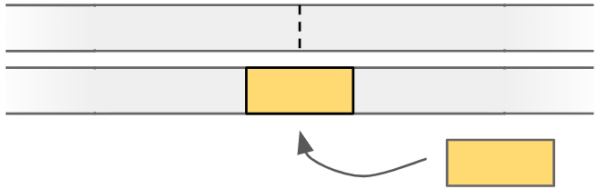
SNV (**S**ingle **N**ucleotide **V**ariant)

INDEL (**In**sertion or **Del**etion)

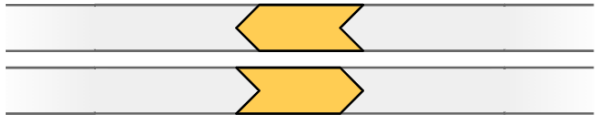


Structural Variants

Insertion



Inversion

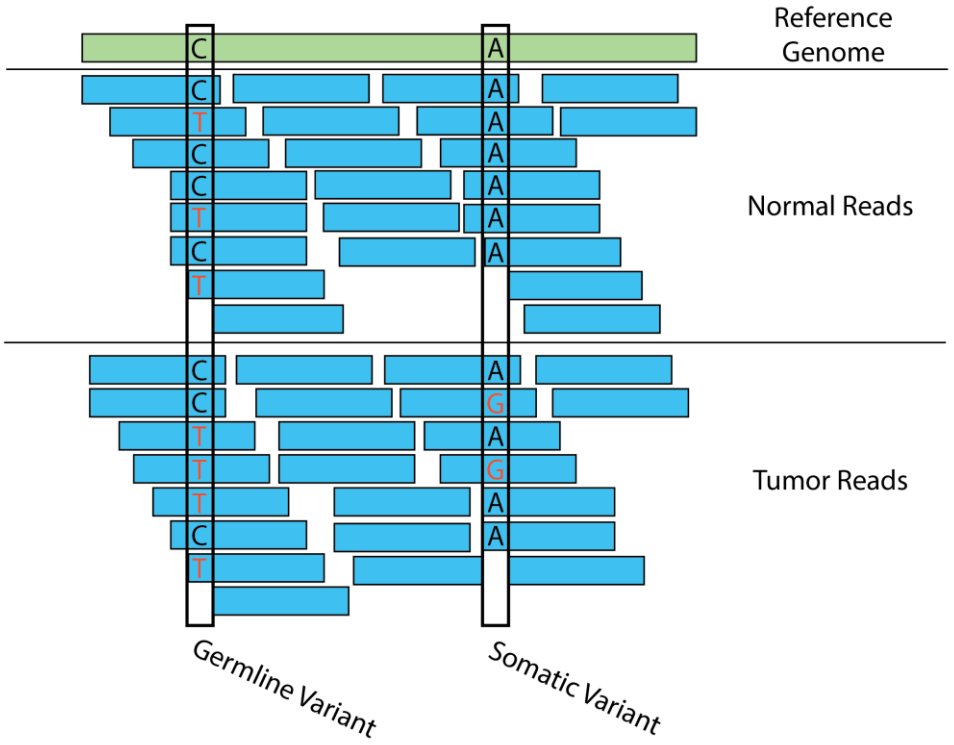


We will mainly focus on SNVs

Germline vs Somatic Variants

- A **germline variant** is a genetic change present in the reproductive cells, which means it can be inherited by offspring and potentially passed down through generations.

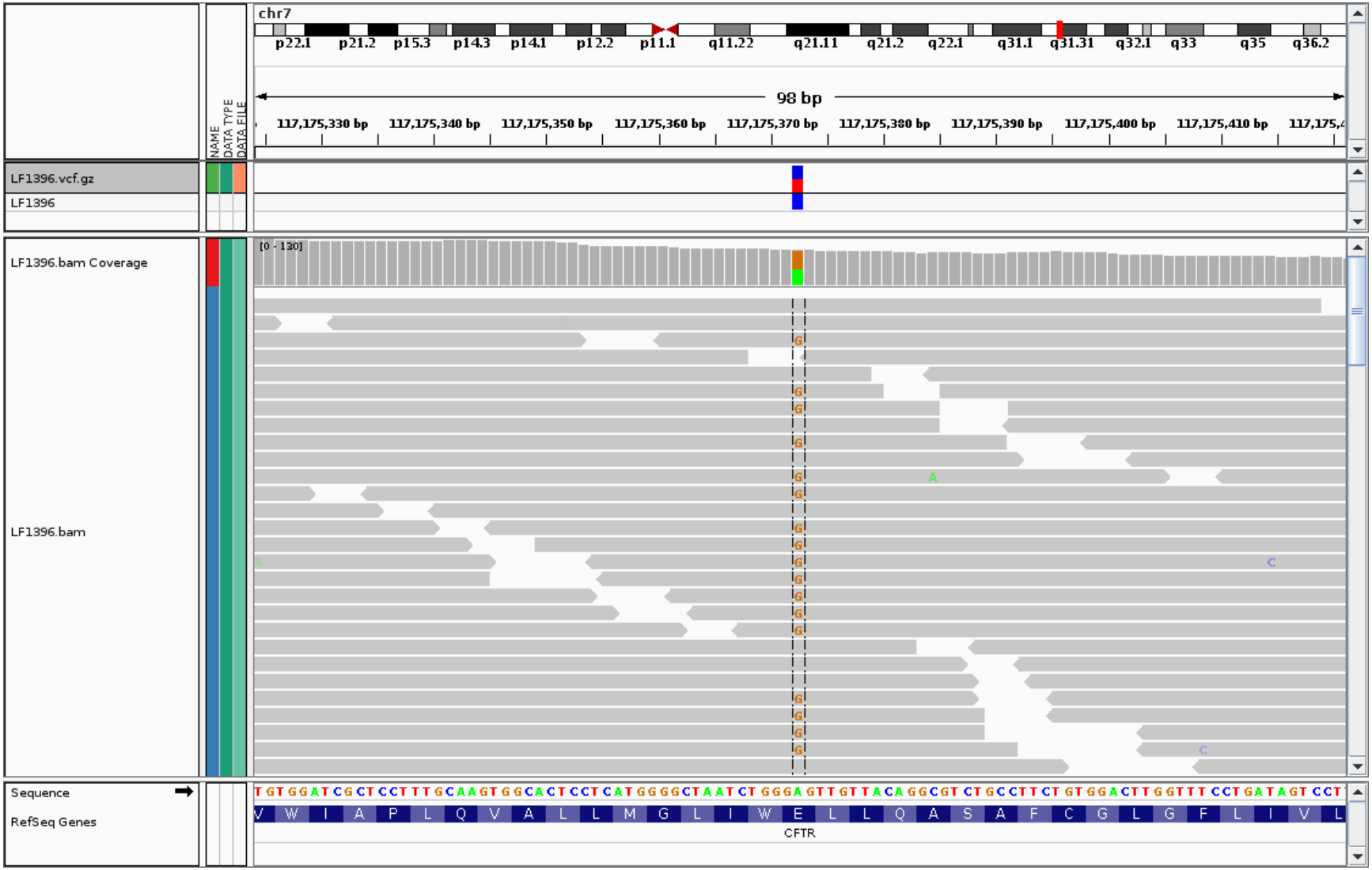
- **Somatic variants** are the most common cause of cancer; some of these genetic alterations produce noticeable traits or physical changes (phenotypes), while others do not. These variants are not hereditary.



UNIVERSITÀ DI TRENTO

CIBIO

In IGV



How do we find sequence variants?

The **GATK suite** is a group of software to perform variant calling and all the tasks needed to obtain high quality SNVs.

Mutect v2

Mutect2 employs a Bayesian statistical framework based on the **Allelic Fraction**

$$\boldsymbol{Allelic\ Fraction\ (AF)} = \frac{\boldsymbol{Number\ of\ variant\ reads}}{\boldsymbol{Total\ reads\ at\ loci}}$$

The allelic fraction is used as a **filter** criterion. SNVs with a very low allelic fraction may be likely due to sequencing errors or very rare subclonal mutations

Must do!

Before using GATK, we need to run this command to configure it:

```
source /usr/local/GenomicsTechnologies/setup_env.sh
```

How to run Mutect 2

- Before actually callling the variants, we need to create a sequence dictionary, which is simply an index that Mutect2 needs to run.

- You can generate this dictionary by typing on the shell:

```
gatk CreateSequenceDictionary -R <fasta filename>
```

<fasta filename>: is the fasta file with the reference genome that you used for the alignment index generation

How to run Mutect 2

- At this point, you can run Mutect2 by typing:

```
gatk Mutect2 --reference <fasta filename> --input <sample_name.sorted.bam> --output <sample_name.unfiltered.vcf.gz>
```

- Note that the input is the sorted bam filed that you obtained with samtools

The VCF file

https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF (Variant Call Format) is a text file format.
It contains:

- meta-information lines
- header line
- data lines each containing information about a position in the genome.

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##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">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Description

Metadata

Data

Filter the VCF

The output from Mutect2 is a raw variant calling output and the calls need to be **filtered** to ensure against errors such as:

- Technical artifacts
- Germline mutations
- Sequencing Errors

FilterMutectCalls will annotate the FILTER field in the VCF file with whether the variant is passing with **PASS** or the reasons why it failed filtering.

```
gatk FilterMutectCalls –reference <fasta filename> --variant
<sample_name.unfiltered.vcf.gz> --output <sample_name.filtered.vcf.gz>
```

Filter the VCF

FilterMutectCalls contains a set of filters, divided into three categories: technical artifacts, non-somatic, and sequencing error.

Filter	Threshold	Explanation
clustered_events	max-events-in-region	mutations sharing an assembly region
duplicate_evidence	unique-alt-read-count	unique insert start/end pairs of alt reads
multiallelic	max-alt-alleles-count	passing alt alleles at a site
base_qual	min-median-base-quality	median base quality of alt reads
map_qual	min-median-mapping-quality	median mapping quality of alt reads
fragment	max-median-fragment-length-difference	difference of alt and ref reads' median fragment lengths
position	min-median-read-position	median distance of alt mutations from end of read
panel_of_normals	panel-of-normals	presence in panel of normals

For more details on the filtering steps see the second chapter of Mutect manual:

https://github.com/broadinstitute/gatk/blob/master/docs/mutect/mutect.pdf

Questions ?