Human Genomics and Epigenomics

Pratical1 – 18/01/2021
Pratical 2 – 19/01/2021
Pratical 3 – 25/01/2021
Pratical 4 – 26/01/2021

Prof. Massimo Delledonne
Functional Genomics lab

ALIGNMENT AND VARIANT CALLING

1° Day (3h): Pre-processing of raw reads

- The fastq file
- Quality control of fastq files
- Adapter removing and trimming of fastq files
    - Sickle and scythe
    - Trimmomatic
- Reads alignment:
    - The human reference genome (hg19 and hg38, main differences)
    - The BAM file

2° Day (3h): Alignment

- Alignment of trimmed reads to the reference genome
    - BWA-mem
    - Isaac2 pipeline
- Duplicates removal
- Read Clipping
- Visualization of aligned reads on IGV

ALIGNMENT AND VARIANT CALLING

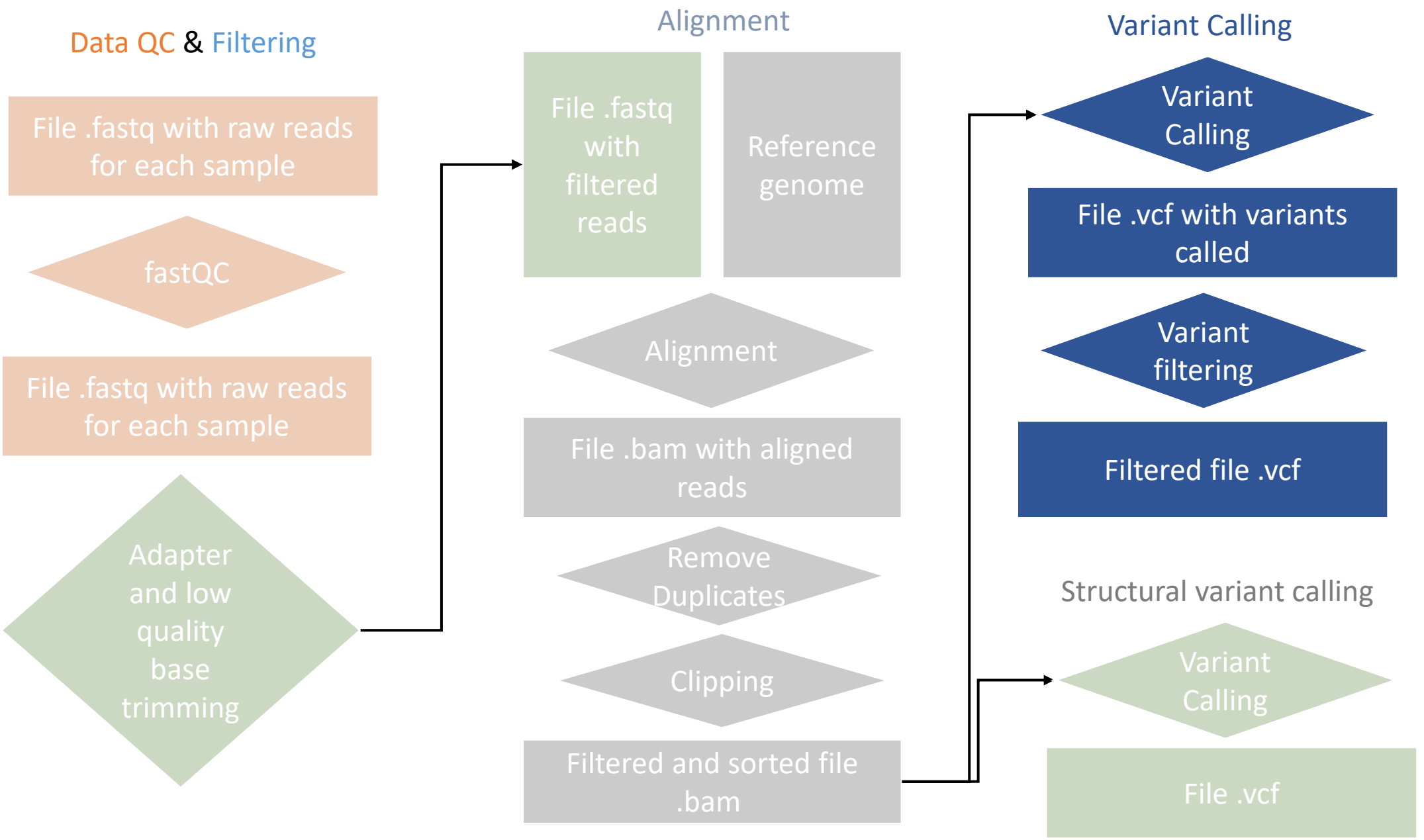
3° Day (3h): Statistics and Variant Calling

- Statistics on reads alignment: main parameters for the evaluation of NGS data
  - Average coverage and uniformity
  - Fold enrichment (on/near/off target)
  - Genotypability (mapping quality besides coverage)
- Variant calling:
  - The VCF and gVCF files
  - Germline variant calling
    - GATK4 Best practice pipeline

4° Day (3h): Variant Calling

- Germline variant calling
  - GATK4 Best practice pipeline
  - Strelka2
- Visualization of genetic variants on IGV
- Structural variant calling

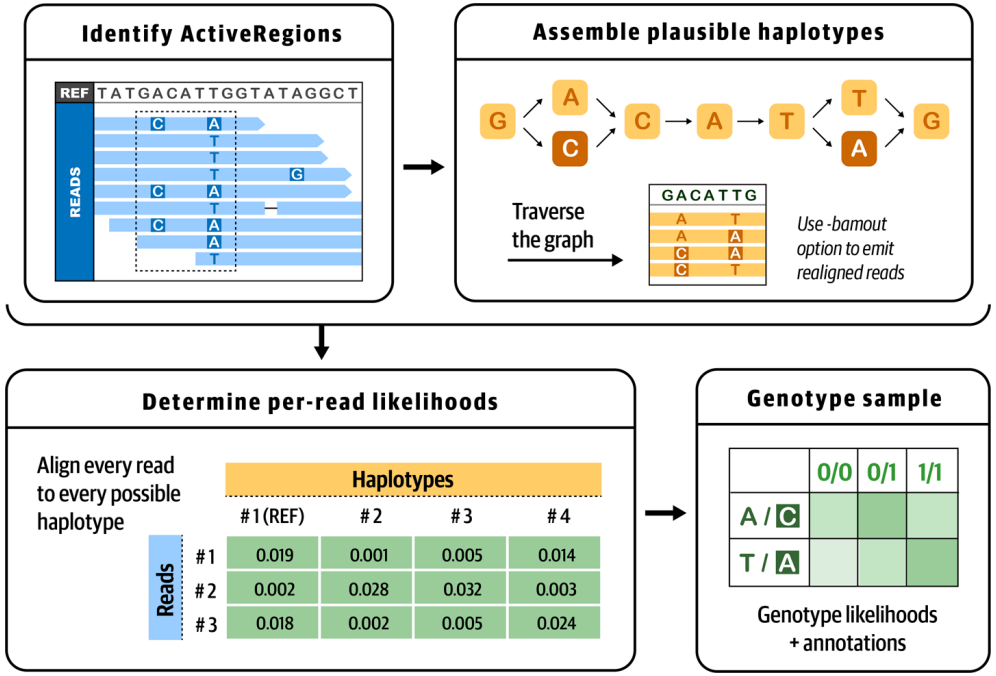
Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
Adapter and low quality base trimming
Alignment
File .fastq with filtered reads
Reference genome
Alignment
File .bam with aligned reads
Remove Duplicates
Clipping
Filtered and sorted file .bam
Variant Calling
Variant Calling
File .vcf with variants called
Variant filtering
Filtered file .vcf
Structural variant calling
Variant Calling
File .vcf

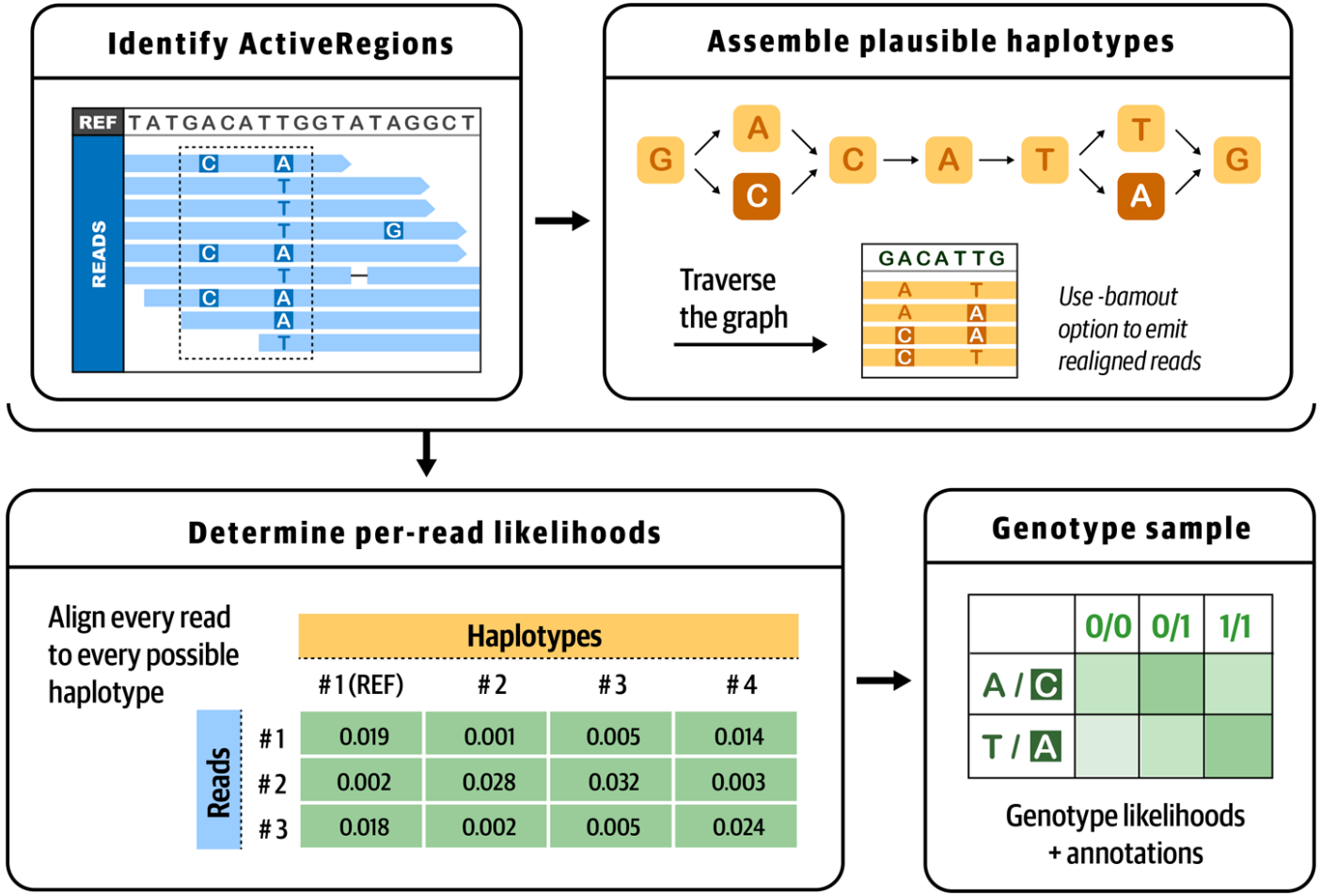
Germline variant calling

Germline variant calling – GATK4

The **HaplotypeCaller** is capable of **calling SNPs and indels simultaneously** via local de-novo assembly of haplotypes in an active region. Whenever the program encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. It creates graphs based on the new assembly and calculate likelihood for each possible haplotype. Outputs the haplotype with a greater score. This allows the HaplotypeCaller to be more accurate when calling regions that are traditionally difficult to call. **It is used for germline variants in target/WES/WGS.**



Germline variant calling – GATK4



1. **Identify ActiveRegions**
   Identify regions were variants are present.
2. **Assemble plausible haplotypes**
   For each region, creates a DeBruijn graph and identifies the possible variation present in the data
3. **Determine per-read likelihoods**
   Each read is aligned to every possible identified haplotype and a score is given.
4. **Genotype sample**
   The likelihood for each genotype is calculated and the most likely genotype is given.

Connect to server

1. Enter in the server:
   a. ssh lessons@157.27.80.26
   b. Password: lez2021
2. Enter in the created folder: cd HGE_2021/your_name

Variant Calling with GATK

1. Enter in the folder: cd HGE_2021/your_name

2. Call variants on the sample:

   java -jar /opt/gatk-3.8/GenomeAnalysisTK.jar -T UnifiedGenotyper -R ../ref/chr6.hg38.fa -I sample.sorted.dedup.clipped.bwa.bamUtils.rg.bam -L ../ref/chr6.hg38.bed -o gatk.raw.vcf

3. Open the file: less -S gatk.raw.vcf

Raw VCF

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##GATKCommandLine.UnifiedGenotyper=<ID=UnifiedGenotyper,Version=3.8-1-0-gf15c1c3ef,Date="Fri Jan 22 17:16:41 CET 2021",Epoch=1611332201504,CommandLineOptions=
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RPA,Number=.,Type=Integer,Description="Number of times tandem repeat unit is repeated, for each allele (including reference)"
##INFO=<ID=RU,Number=1,Type=String,Description="Tandem repeat unit (bases)">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt                    tion bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contin                    strand bias">
##INFO=<ID=STR,Number=0,Type=Flag,Description="Variant is a shor
##contig=<ID=chr6,length=170805979>
##reference=file:///home/lessons/HGE_2021/deni                /ref/chr6.hg38.fa
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  20
chr6    113342  .       T       C       15.65   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    113654  .       T       C       15.65   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    121058  .       C       T       15.65   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    126030  .       G       A       13.72   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    131967  .       A       T       98.03   .       AC=2;AF=1.00;AN=2;DP=6;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    132284  .       T       A       277.78  .       AC=2;AF=1.00;AN=2;DP=8;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    132458  .       T       C       15.65   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
chr6    132572  .       T       C       10.90   .       AC=2;AF=1.00;AN=2;DP=1;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ
gatk.raw.vcf
```

All variants are reported, no filter is applied at the moment.

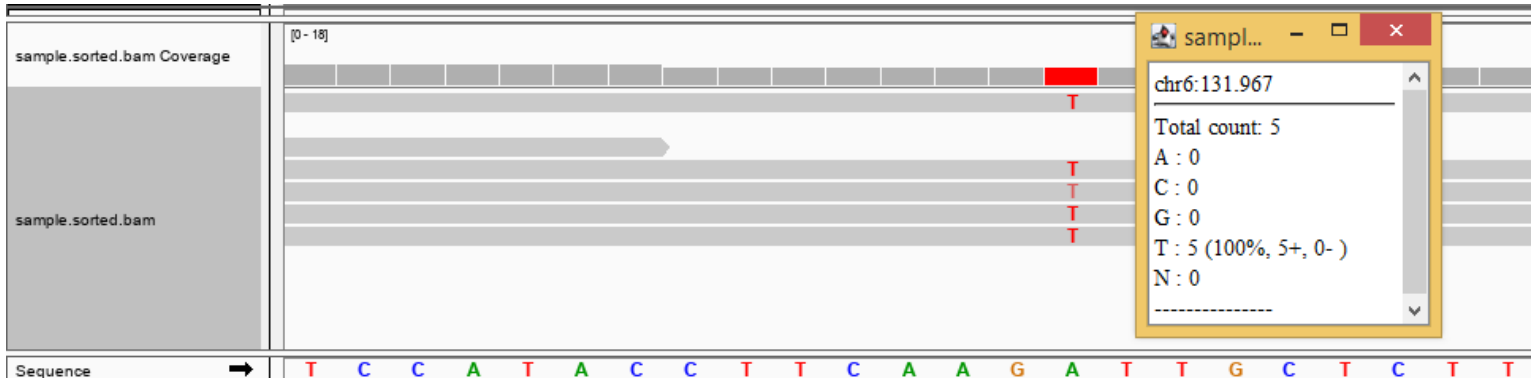
Variant Calling with GATK

1. Filter variants:

```
java -jar /opt/gatk-3.8/GenomeAnalysisTK.jar -T VariantFiltration -R ../ref/chr6.hg38.fa --variant gatk.raw.vcf -o gatk.filtered.vcf --clusterWindowSize 10 --filterExpression 'MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)' --filterName 'HARD_TO_VALIDATE' --filterExpression 'DP < 20' --filterName 'LowCoverage' --filterExpression 'QUAL < 30.0' --filterName 'VeryLowQual' --filterExpression 'QD < 5.0' --filterName 'LowQD' --filterExpression 'FS > 200.0' --filterName 'StrandBias'
```

2. Open the file:

```
less -S gatk.filtered.vcf
```



Filtered VCF

```
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sa
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RPA,Number=.,Type=Integer,Description="Number of times tandem repeat unit is repeated, for each allele (including referenc
##INFO=<ID=RU,Number=1,Type=String,Description="Tandem repeat unit (bases)">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##INFO=<ID=STR,Number=0,Type=Flag,Description="Variant is a short tandem repeat">
##contig=<ID=chr6,length=170805979>
##reference=file:///attachedvolume/HGSI2020/example/reference/chr6.hg38.fa
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  1351S   1352S   135
chr6    131967  .       A       T       122.87  LowCoverage     AC=4;AF=1.00;AN=4;DP=10;D                       ;HaplotypeScore=0.0000;MLEA
chr6    132284  .       T       A       381.61  LowCoverage     AC=6;AF=1.00;AN=6;DP=10;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEA
chr6    140219  .       T       A       2095.16 HARD_TO_VALIDATE        AC=4;AF=0.667;AN=6;BaseQRankSum=-0.059;DP=189;Dels=0.00;FS=0.
chr6    140623  .       G       A       139.93  HARD_TO_VALIDATE;LowCoverage    AC=4;AF=0.667;AN=6;BaseQRankSum=0.742;DP=17;Dels=0.00
chr6    142771  .       C       A       55.59   HARD_TO_VALIDATE;LowQD  AC=2;AF=0.333;AN=6;BaseQRankSum=-0.828;DP=80;Dels=0.00;FS=0.
chr6    142840  .       T       C       1114.16 HARD_TO_VALIDATE        AC=4;AF=0.667;AN=6;BaseQRankSum=-0.565;DP=95;Dels=0.00;FS=9.4
chr6    143085  .       C       G       119.17  HARD_TO_VALIDATE;LowQD  AC=2;AF=0.333;AN=6;BaseQRankSum=1.302;DP=119;Dels=0.00;FS=0.
chr6    144105  .       A       G       98.77   HARD_TO_VALIDATE        AC=3;AF=0.500;AN=6;BaseQRankSum=0.000;DP=23;Dels=0.00;FS=0.00
chr6    144137  .       A       C       301.48  HARD_TO_VALIDATE        AC=6;AF=1.00;AN=6;DP=20;Dels=0.00;FS=0.000;HaplotypeScore=0.
chr6    144967  .       A       C       170.60  LowCoverage     AC=4;AF=0.667;AN=6;BaseQRankSum=0.204;DP=19;Dels=0.00;FS=6.662;Haplo
chr6    147332  .       A       G       2074.90 HARD_TO_VALIDATE        AC=6;AF=1.00;AN=6;BaseQRankSum=-1.400;DP=82;Dels=0.00;FS=0.00
chr6    147363  .       C       T       168.93  LowQD   AC=3;AF=0.500;AN=6;BaseQRankSum=0.093;DP=150;Dels=0.00;FS=8.946;HaplotypeScor
chr6    147404  .       C       T       2409.9  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-0.627;DP=236;Dels=0.00;FS=14.943;HaplotypeSc
```

We set filter if DP<20

Variant passing all the filters we set

Variant Calling with GATK

1. Select Variants passing the filter:

   java -jar /opt/gatk-3.8/GenomeAnalysisTK.jar -T SelectVariants -R ../ref/chr6.hg38.fa --variant gatk.raw.vcf --excludeFiltered -o gatk.selected.variants.vcf

2. Open the file:

   less gatk.selected.variants.vcf

3. Zip and index the file:

   bgzip gatk.selected.variants.vcf

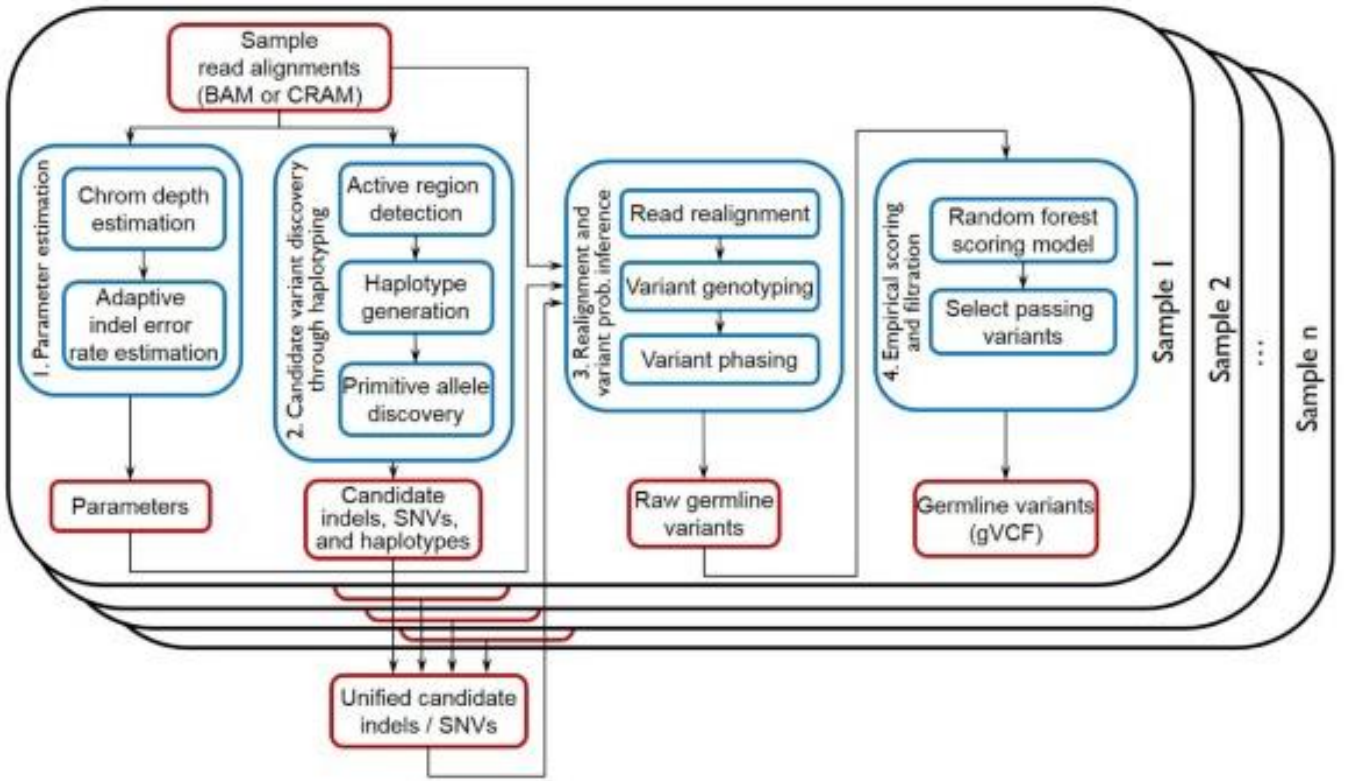
   tabix gatk.selected.variants.vcf.gz

Selected VCF

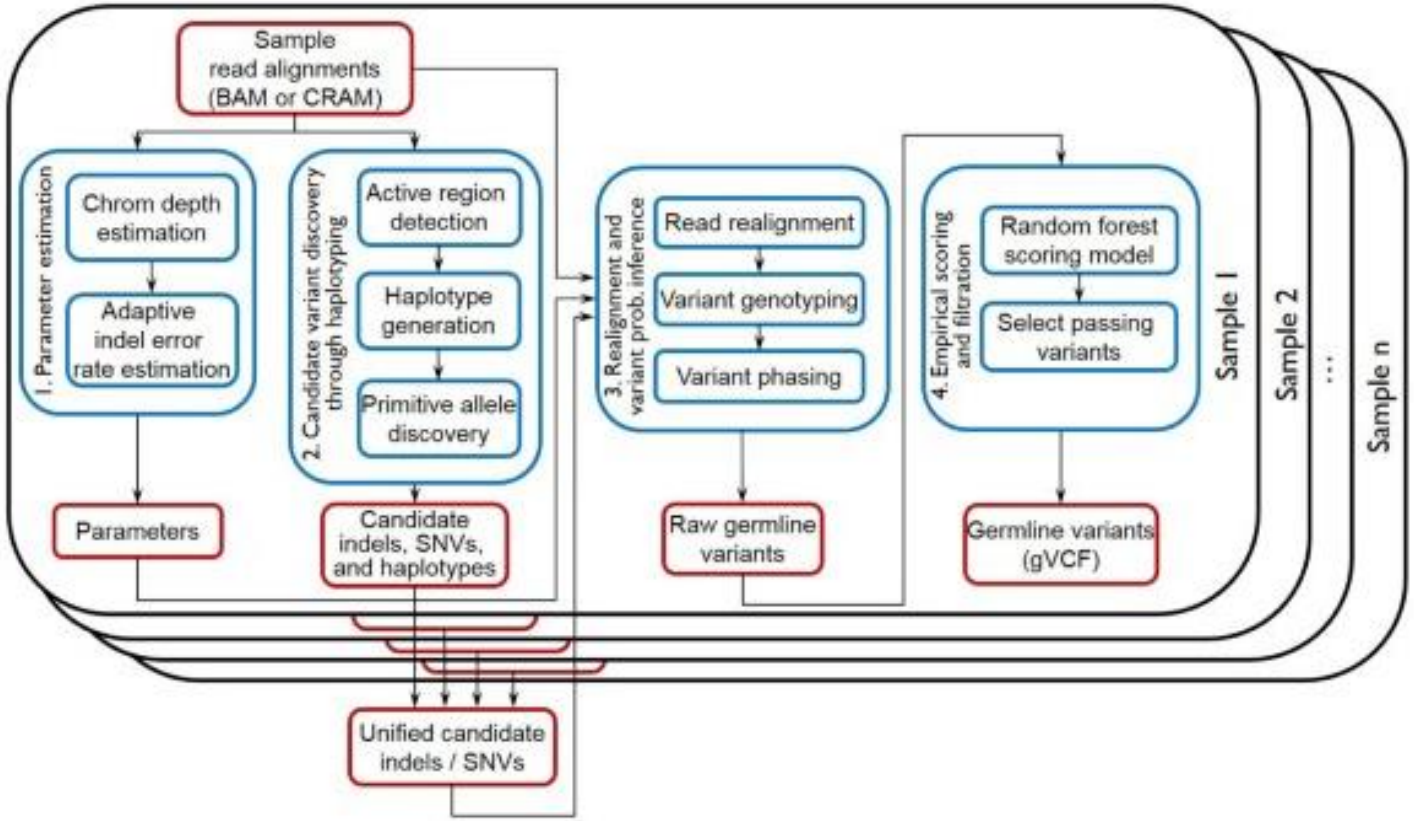
```
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sa
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the s
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RPA,Number=.,Type=Integer,Description="Number of times tandem repeat unit is repeated, for each allele (including referenc
##INFO=<ID=RU,Number=1,Type=String,Description="Tandem repeat unit (bases)">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##INFO=<ID=STR,Number=0,Type=Flag,Description="Variant is a short tandem repeat">
##contig=<ID=chr6,length=170805979>
##reference=file:///attachedvolume/HGSI2020/example/reference/chr6.hg38.fa
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  1351S   1352S   1353S
chr6    147404  .       C       T       2409.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-0.627;DP=236;Dels=0.00;FS=14.943;HaplotypeSc
chr6    147750  .       C       A       478.07  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-1.134;DP=35;Dels=0.00;FS=3.979;HaplotypeScor
chr6    292833  .       G       A       345.92  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=0.185;DP=35;Dels=0.00;FS=0.000;HaplotypeScore
chr6    304890  .       T       A       254.92  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-0.475;DP=31;Dels=0.00;FS=2.783;HaplotypeScor
chr6    325126  .       A       G       22417.90        PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=-1.018;DP=750;Dels=0.00;FS=0.000;Haplo
chr6    325403  .       G       A       6867.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=0.298;DP=729;Dels=0.00;FS=1.750;HaplotypeScor
chr6    325711  .       C       T       6098.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=6.000;DP=750;Dels=0.00;FS=1.816;HaplotypeScor
chr6    325873  .       T       C       19506.90        PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=3.372;DP=731;Dels=0.00;FS=0.000;Haplot
chr6    325961  .       T       C       18061.90        PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=3.651;DP=742;Dels=0.00;FS=6.004;Haplot
chr6    326134  .       G       A       21980.90        PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=1.643;DP=703;Dels=0.00;FS=0.000;Haplot
chr6    334923  .       A       G       798.92  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-1.278;DP=69;Dels=0.00;FS=0.000;HaplotypeScor
chr6    335175  .       A       T       8694.90 PASS    AC=6;AF=1.00;AN=6;DP=257;Dels=0.00;FS=0.000;HaplotypeScore=1.1956;MLEAC=6;MLE
chr6    335251  .       T       C       1447.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=1.147;DP=145;Dels=0.00;FS=3.440;HaplotypeScor
chr6    335253  .       T       C       1428.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=0.090;DP=140;Dels=0.00;FS=3.579;HaplotypeScor
chr6    335268  .       C       T       1310.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-1.548;DP=128;Dels=0.00;FS=2.660;HaplotypeSc
chr6    337804  .       C       T       776.92  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-1.155;DP=40;Dels=0.00;FS=0.000;HaplotypeScor
chr6    337925  .       G       T       6318.90 PASS    AC=6;AF=1.00;AN=6;DP=185;Dels=0.00;FS=0.000;HaplotypeScore=2.5231;MLEAC=6;MLE
chr6    347888  .       A       G       253.93  PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=-0.692;DP=22;Dels=0.00;FS=0.000;HaplotypeScor
chr6    348051  .       A       G       3007.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=0.143;DP=213;Dels=0.00;FS=3.005;HaplotypeScor
chr6    348080  .       A       G       9952.90 PASS    AC=6;AF=1.00;AN=6;DP=294;Dels=0.00;FS=0.000;HaplotypeScore=3.7227;MLEAC=6;MLE
chr6    348906  .       G       A       5354.92 PASS    AC=3;AF=0.500;AN=6;BaseQRankSum=0.519;DP=381;Dels=0.00;FS=8.265;HaplotypeScor
```

Germline variant calling – Strelka2

Strelka2 is a **fast and accurate small variant caller** optimized for analysis of germline variation in small cohorts and somatic variation in tumor/normal sample pairs. For each sample, the germline model estimates parameters and extract candidate variants using haplotype generation. Then, realign the data and calculate a probability for each variant, considering previous results of all the samples (if more than one sample is analyzed). Finally, output results based on the score and filters. The somatic calling model accounts for possible tumor cell contamination in the normal sample. **It is used for germline variants in target/WES/WGS.**



Germline variant calling – Strelka2



1. **Parameter estimation**
   Using mixture model to estimate both indel variant mutation rates and indel noise rates from a set of error processes
2. **Candidate variant discovery**
   The reads at every locus are modeled as depending on the corresponding base call quality strings, the unobserved haplotype that generated the read, and the locus-specific error rates.
3. **Realignment and variant probability inference**
   Reads are realigned using local assembly and the probability of the variant is calculated
4. **Empirical scoring and filtration**
   The empirical variant scoring (EVS) is calculated based on:
   1. the genotype probability,
   2. root-mean-square mapping quality,
   3. strand bias
   4. the fraction of reads consistent with locus haplotype model
   5. the complexity of the reference context

Variant Calling with Strelka2

1. Enter in the folder: cd /home/lessons/HGE_2021/your_name
2. Create the configuration file:

   /opt/strelka-2.9.2/bin/configureStrelkaGermlineWorkflow.py --bam sample.sorted.dedup.clipped.bwa.bamUtils.rg.bam --reference /home/lessons/HGE_2021/ref/chr6.hg38.fa --exome --runDir isaac_results
3. Call the variants on the sample:

   python /home/lessons/HGE_2021/denise/isaac_results/runWorkflow.py -m local
4. Open the file: less -S genome.S1.vcf.gz

GVCF

Open the file: less -S /home/lessons/HGE_2021/denise/isaac_results/results/variants/genome.S1.vcf.gz

Header

Body

```
##FILTER=<ID=LowGQX,Description="Locus GQX is below threshold or not present">
##FILTER=<ID=HighDPFRatio,Description="The fraction of basecalls filtered out at a site is greater than 0.4">
##FILTER=<ID=HighSNVSB,Description="Sample SNV strand bias value (SB) exceeds 10">
##FILTER=<ID=LowDepth,Description="Locus depth is below 3">
##FILTER=<ID=NotGenotyped,Description="Locus contains forcedGT input alleles which could not be genotyped">
##FILTER=<ID=PloidyConflict,Description="Genotype call from variant caller not consistent with chromosome ploidy">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  20
chr6    1       .       N       .       .       LowGQX  END=60064;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    60065   .       A       .       .       LowGQX;LowDepth END=60215;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    60216   .       A       .       .       LowGQX  END=60222;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    60223   .       T       .       .       LowGQX;LowDepth END=60373;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    60374   .       A       .       .       LowGQX  END=61797;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    61798   .       A       .       .       LowGQX;LowDepth END=61860;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61861   .       G       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61862   .       T       .       .       LowGQX;LowDepth .       GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61863   .       A       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61864   .       G       .       .       LowGQX;LowDepth END=61872;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61873   .       C       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61874   .       T       .       .       LowGQX;LowDepth END=61883;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61884   .       C       .       .       LowGQX;HighDPFRatio;LowDepth    END=61885;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61886   .       T       .       .       LowGQX;LowDepth END=61892;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61893   .       G       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61894   .       T       .       .       LowGQX;LowDepth END=61908;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    61909   .       A       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    .:.:0:1:0
chr6    61910   .       T       .       .       LowGQX;LowDepth END=62058;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    62059   .       A       .       .       LowGQX  END=62840;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    62841   .       C       .       .       LowGQX;LowDepth END=62991;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    62992   .       C       .       .       LowGQX  END=63113;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    63114   .       C       .       .       LowGQX;LowDepth END=63264;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    63265   .       G       .       .       LowGQX  END=63465;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    63466   .       A       .       .       LowGQX;LowDepth END=63707;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    63708   .       T       .       .       LowGQX  END=67060;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr6    67061   .       A       .       .       LowGQX;LowDepth END=67159;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr6    67160   .       A       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    0/0:3:1:1:1
chr6    67161   .       G       .       .       LowGQX;LowDepth END=67184;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:5:2:0:2
chr6    67185   .       A       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    0/0:3:1:1:1
chr6    67186   .       A       .       .       LowGQX;LowDepth END=67202;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:5:2:0:2
chr6    67203   .       A       G       6       LowGQX;LowDepth SNVHPOL=4;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:35:6:2:0:1,1:1,0:0,:0.7:LowGQX;Lo
chr6    67204   .       T       .       .       LowGQX;LowDepth END=67211;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP
chr6    67212   .       T       .       .       LowGQX;HighDPFRatio;LowDepth    .       GT:GQX:DP:DPF:MIN_DP    0/0:3:1:1:1
chr6    67213   .       G       .       .       LowGQX;LowDepth END=67252;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:5:2:0:2
```

Reference Position

Variant

VCF

Open the file: less -S /home/lessons/HGE_2021/denise/isaac_results/results/variants/variants.vcf.gz

```
##FILTER=<ID=SiteConflict,Description="Site is filtered due to an overlapping indel call filter">
##FILTER=<ID=LowGQX,Description="Locus GQX is below threshold or not present">
##FILTER=<ID=HighDPFRatio,Description="The fraction of basecalls filtered out at a site is greater than 0.4">
##FILTER=<ID=HighSNVSB,Description="Sample SNV strand bias value (SB) exceeds 10">
##FILTER=<ID=LowDepth,Description="Locus depth is below 3">
##FILTER=<ID=NotGenotyped,Description="Locus contains forcedGT input alleles which could not be genotyped">
##FILTER=<ID=PloidyConflict,Description="Genotype call from variant caller not consistent with chromosome ploidy">
##FILTER=<ID=NoPassedVariantGTs,Description="No samples at this locus pass all sample filters and have a variant genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  20
chr6    67203   .       A       G       6       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=4;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:35:6:2:0:
chr6    67278   .       A       T       6       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:34:6:2:0:
chr6    68816   .       T       C       2       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
chr6    69321   .       G       A       6       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:35:6:2:0:
chr6    69405   .       C       T       6       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=4;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:35:6:2:0:
chr6    88580   .       G       T       2       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=27 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
chr6    88590   .       G       C       2       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=5;MQ=27 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
chr6    96530   .       T       C       2       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=12;MQ=27        GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3
chr6    105690  .       A       G       0       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=27 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:19:0:2:0:
chr6    107046  .       G       A       0       LowGQX;NoPassedVariantGTs       SNVHPOL=2;MQ=13 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL
chr6    113342  .       T       C       11      LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL
chr6    113654  .       T       C       10      LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:3:1:0:0
chr6    121058  .       C       T       11      LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:3:1:0:0
chr6    122701  .       A       G       4       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=44 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:32:4:2:0:
chr6    122814  .       C       T       3       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=44 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:30:0:2:0:
chr6    124046  .       A       ATT     105     PASS    CIGAR=1M2I;RU=T;REFREP=0;IDREP=2;MQ=43  GT:GQ:GQX:DPI:AD:ADF:ADR:FT:PL   1/1:18:18:7:0,7:0,0:0,7:PASS:
chr6    126030  .       G       A       7       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=40 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:2:1:0:0
chr6    131967  .       A       T       41      LowGQX;NoPassedVariantGTs       SNVHPOL=3;MQ=27 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    1/1:10:10:4:1:0,4:0,4
chr6    132019  .       G       C       0       LowGQX;NoPassedVariantGTs       SNVHPOL=2;MQ=37 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:9:0:11:0:10,1:7,1
chr6    132230  .       A       T       0       LowGQX;NoPassedVariantGTs       SNVHPOL=2;MQ=53 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:8:0:11:0:10,1:0,0
chr6    132265  .       C       G       0       LowGQX;NoPassedVariantGTs       SNVHPOL=3;MQ=53 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:6:0:10:0:9,1:0,0:
chr6    132284  .       T       A       113     PASS    SNVHPOL=4;MQ=50 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    1/1:21:21:8:0:0,8:0,0:0,8:0.0:PASS:150,24,0
chr6    132458  .       T       C       11      LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:3:1:0:0
chr6    132572  .       T       C       5       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=4;MQ=33 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
chr6    132577  .       G       A       5       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=6;MQ=33 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
chr6    135415  .       A       C       8       LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=6;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:2:1:0:0
chr6    140219  .       T       A       198     PASS    SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:231:198:73:2:46,27:23,17:23,10:-19.4:PASS
chr6    140622  .       C       T       33      PASS    SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:66:33:8:0:6,2:0,1:6,1:-5.7:PASS:67,0,92
chr6    140623  .       G       A       58      PASS    SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:89:58:8:0:3,5:1,0:2,5:0.7:PASS:92,0,91
chr6    140847  .       T       A       2       LowGQX;LowDepth;NoPassedVariantGTs       SNVHPOL=15;MQ=60        GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3
chr6    140854  .       C       G       11      LowGQX;LowDepth;NoPassedVariantGTs      SNVHPOL=3;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:3:1:0:0
chr6    140909  .       A       C       1       LowGQX;HighDPFRatio;LowDepth;NoPassedVariantGTs  SNVHPOL=2;MQ=60 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3
```

Filtered variants

Variants passing all the filters

Visualization of genetic variants on IGV

Download the vcf file

- Download the vcf file and the index file on your pc:
- Open new terminal:

cd Desktop/HGE_2021

rsync -auv lessons@157.27.80.26:/home/lessons/HGE_2021/denise/gatk.selected.variants.vcf.gz* .

rsync -auv lessons@157.27.80.26:/home/lessons/HGE_2021/denise/isaac_results/results/variants/variants.vcf.gz* .

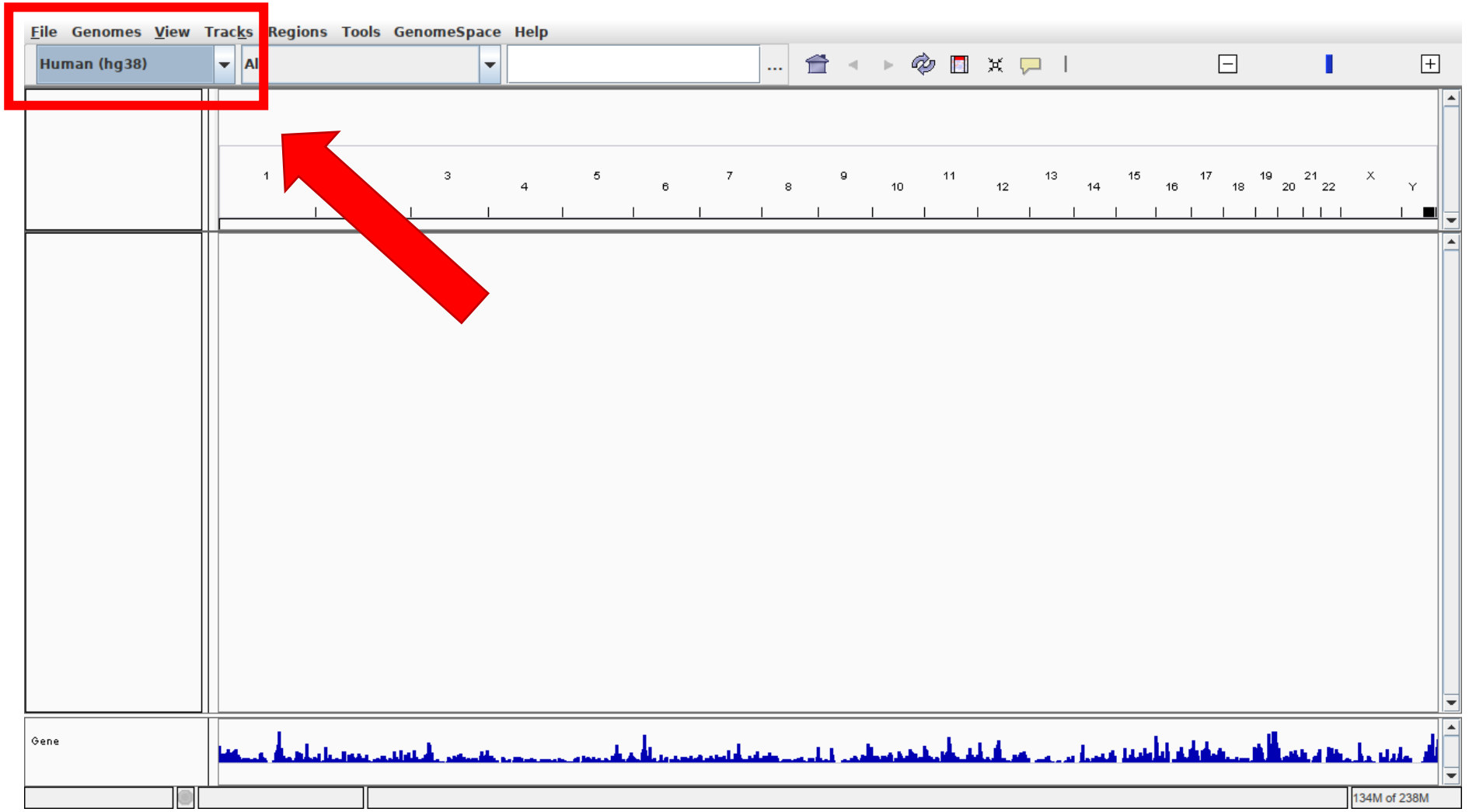
Password: lez2021

- Check if you have downloaded: ls

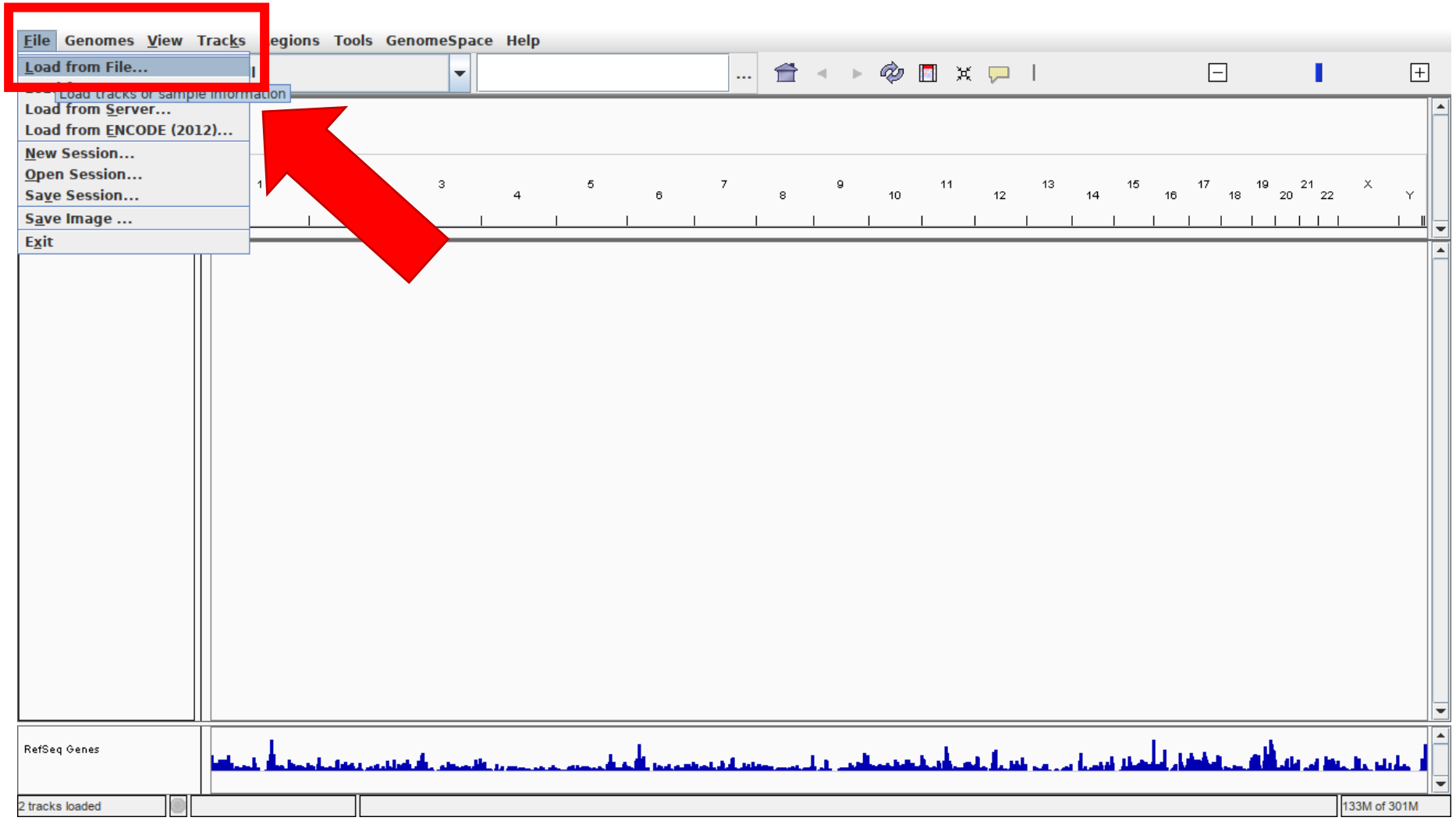
- Open IGV

./igv.sh for Ubuntu

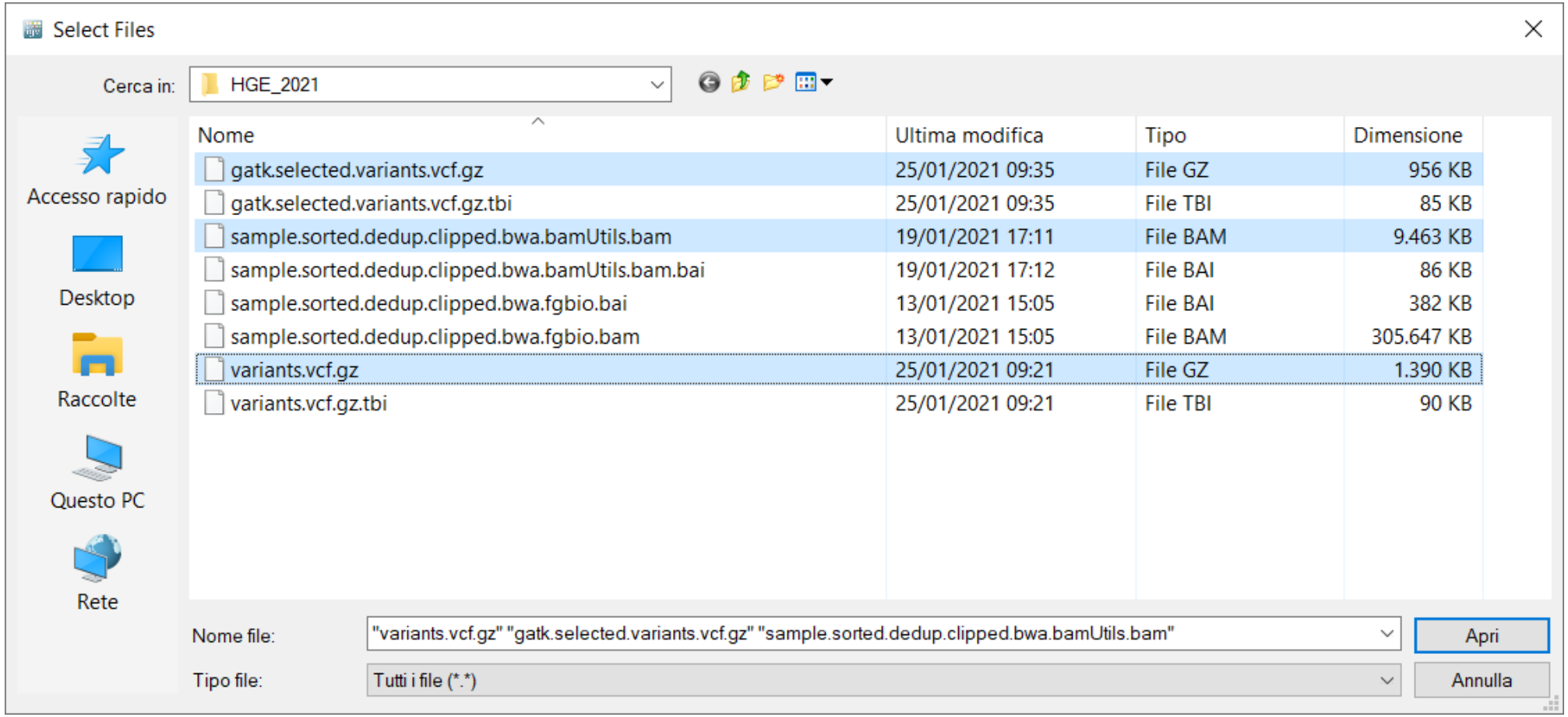
Download the vcf file



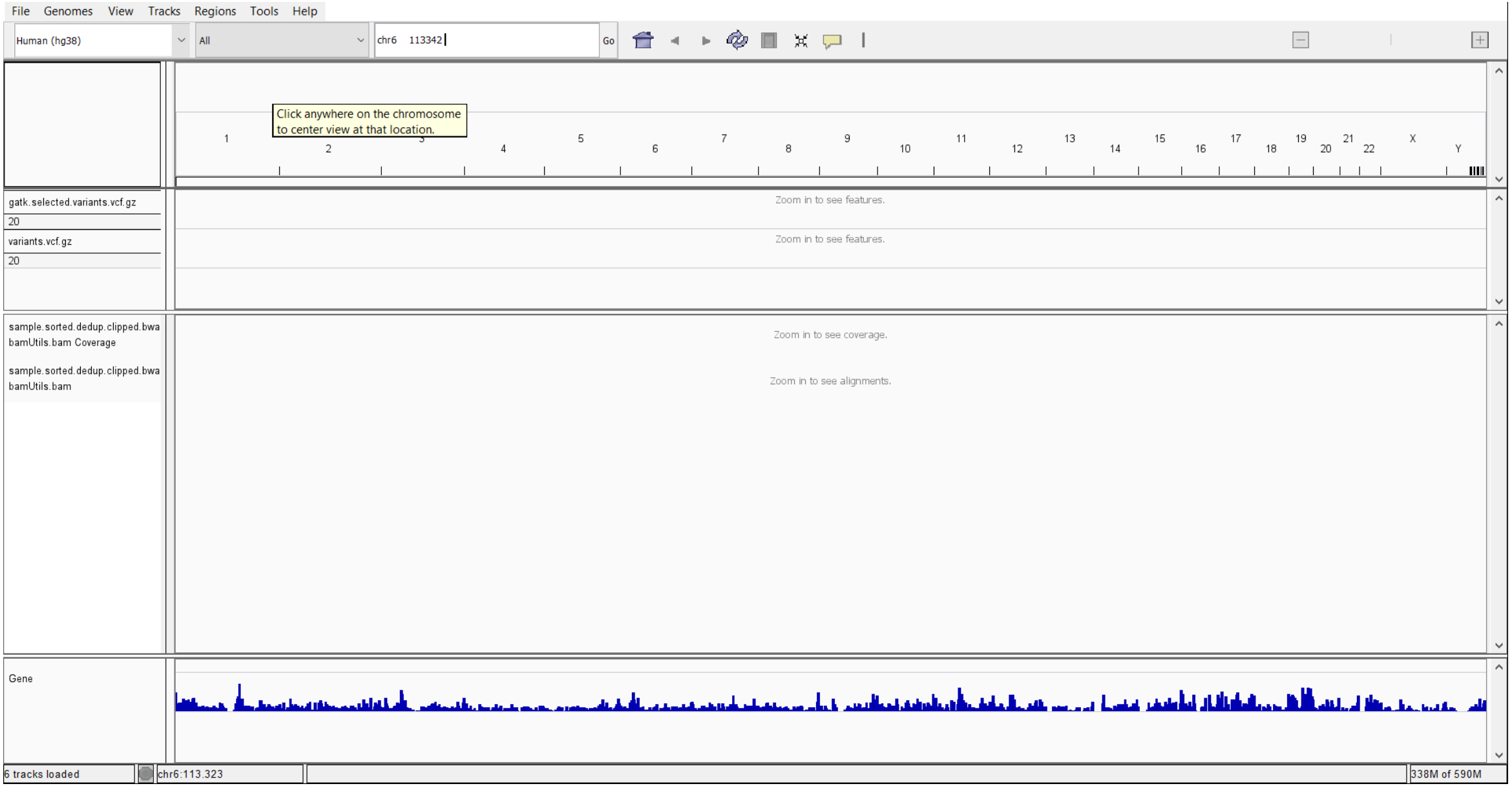
Download the vcf file



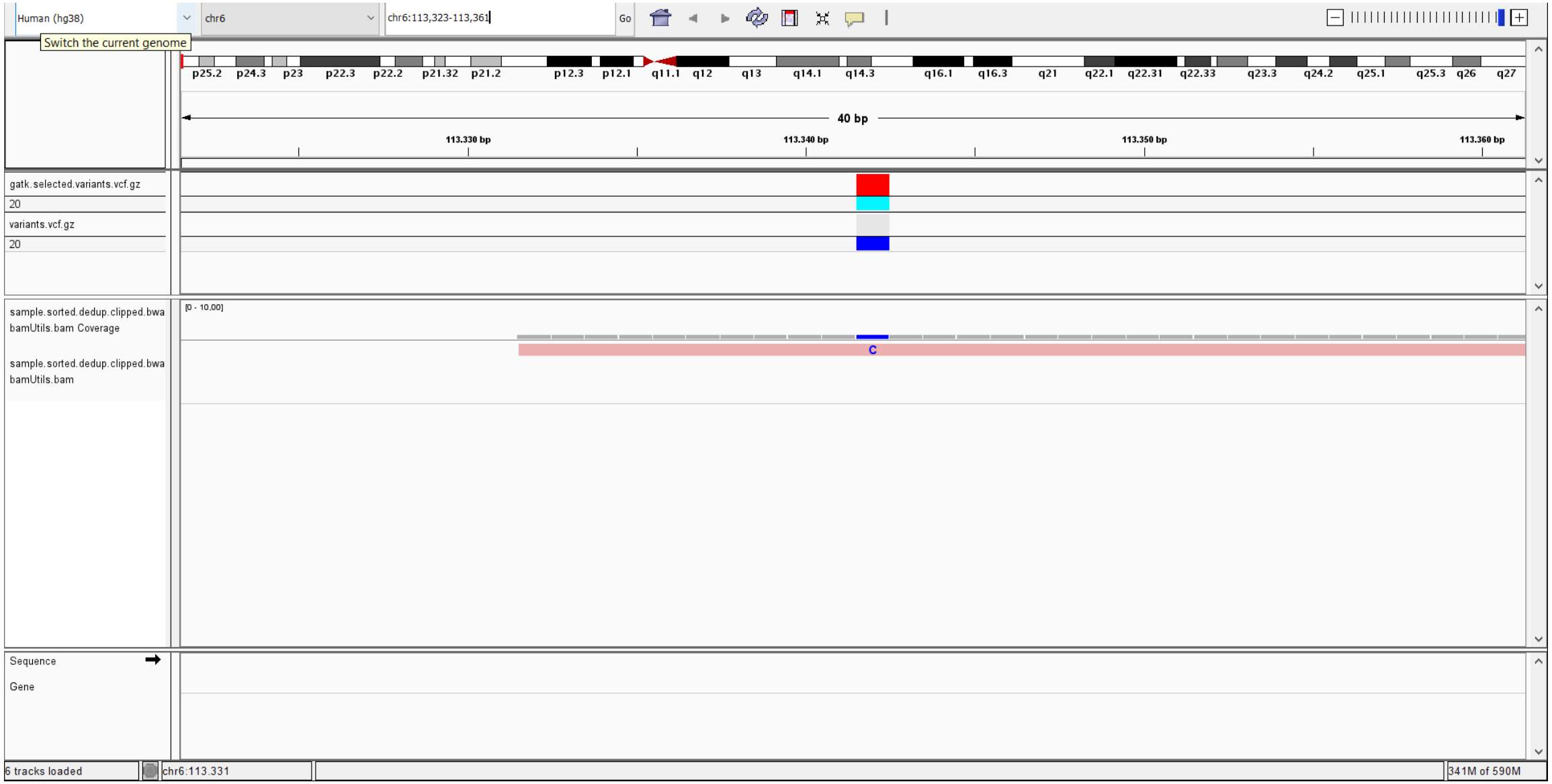
Select the files



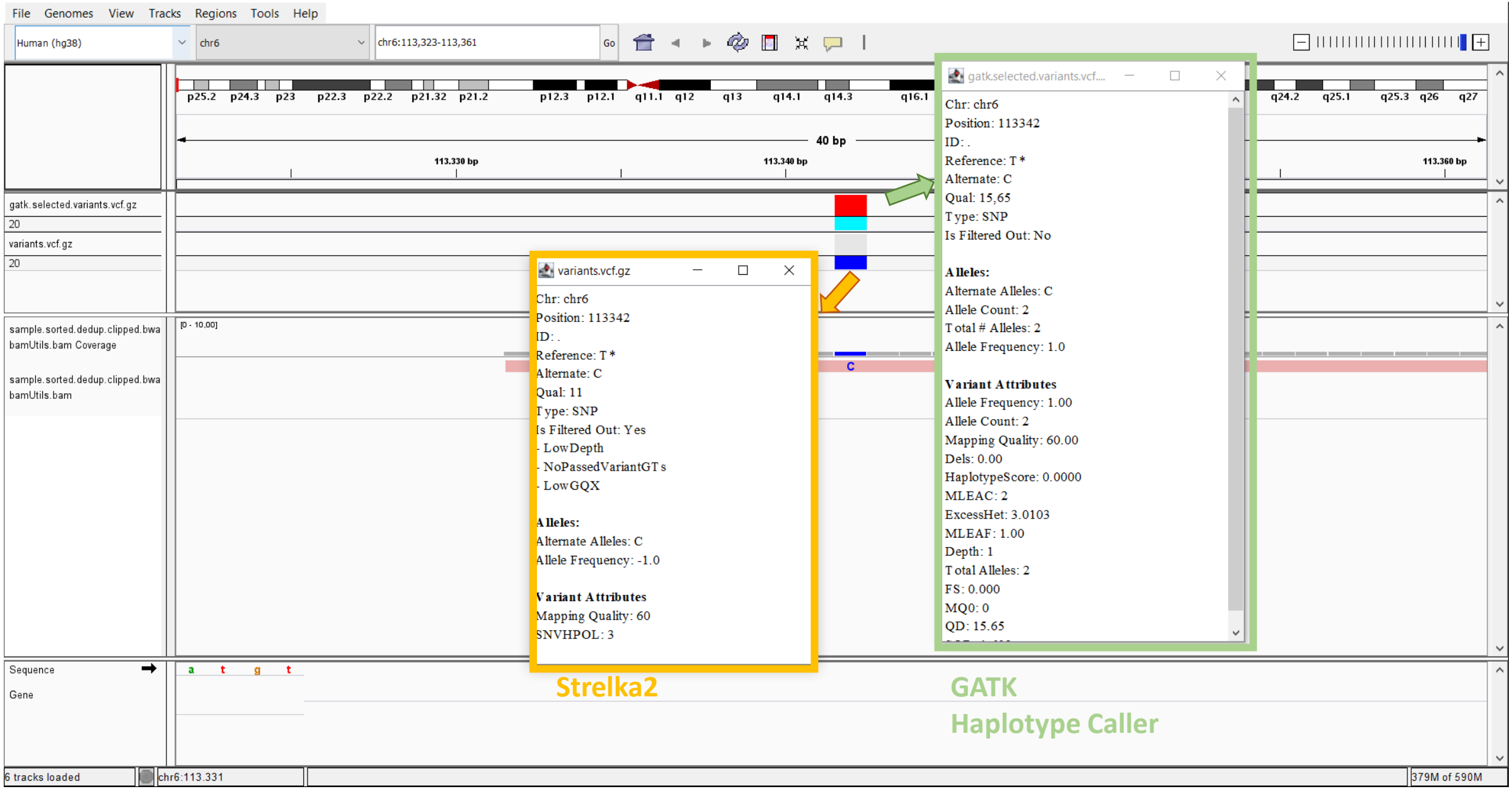
Search for a position



Visualization of the variant

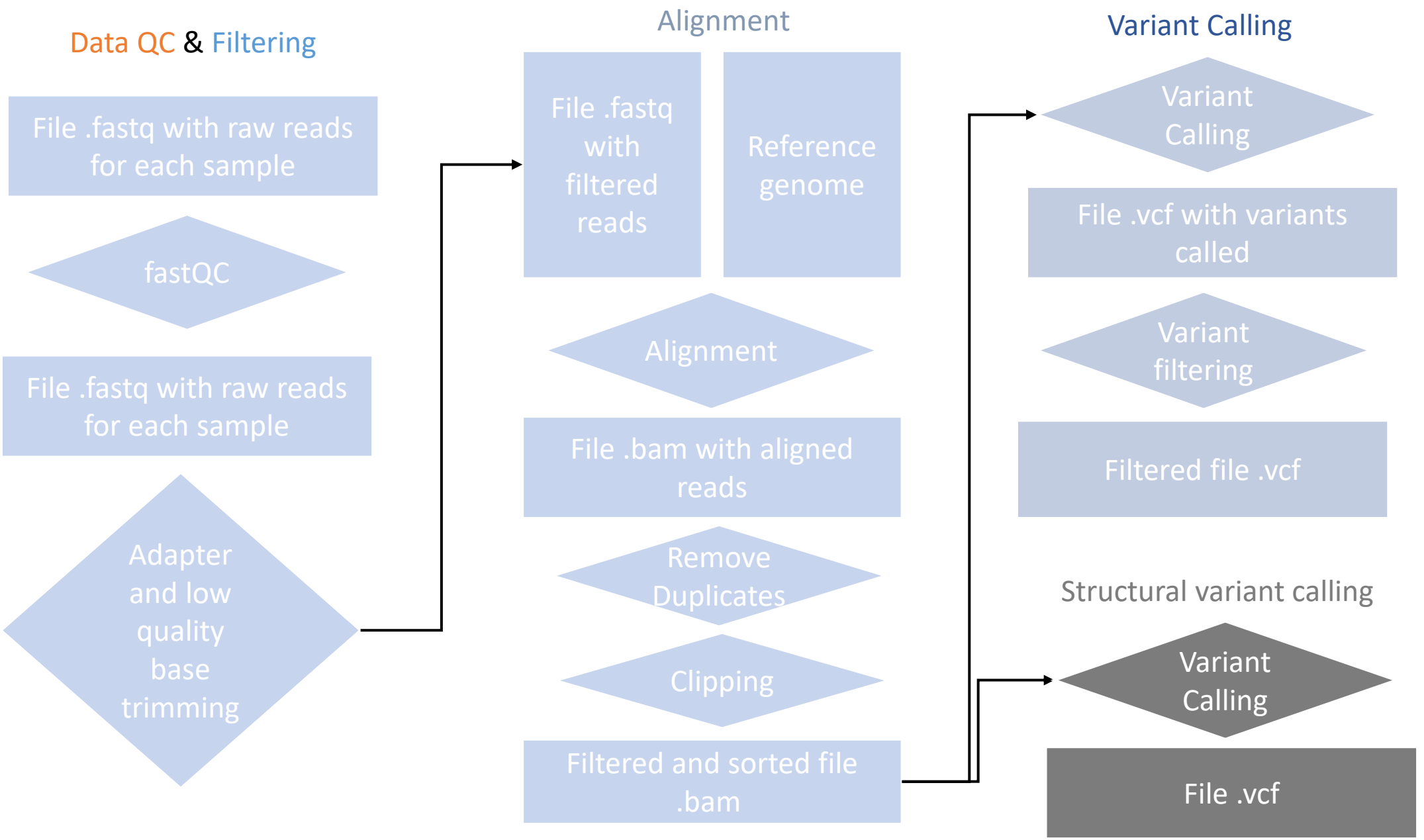


VCF files from the two software



Structural variant calling

Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
Adapter and low quality base trimming
Alignment
File .fastq with filtered reads
Reference genome
Alignment
File .bam with aligned reads
Remove Duplicates
Clipping
Filtered and sorted file .bam
Variant Calling
Variant Calling
File .vcf with variants called
Variant filtering
Filtered file .vcf
Structural variant calling
Variant Calling
File .vcf

Structural Variants identification with NGS platforms

Outline

- Signatures of different SV types
- Approaches and software for SV detection
- SV visual inspection

Signatures of different SV types

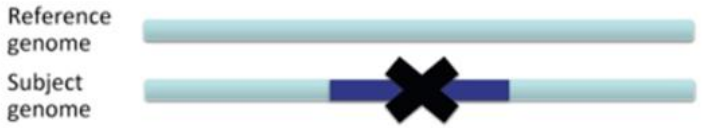
Different features may suggest the occurrence of a Structural Variant (SV)

- Insert size
  - Greater than expected -> Deletion
  - Smaller than expected -> Insertion
- Coverage depth
  - Higher than expected -> Duplication/Copy Number Variation
  - Lower than expected -> Deletion
- Read pair orientation
  - Same direction (LL and RR) -> Inversion
  - Pointing towards the outside (RL) -> Tandem duplication/Intra-chromosomal translocation
- Read pairs mapping to different chromosomes
  - Inter-chromosomal translocation

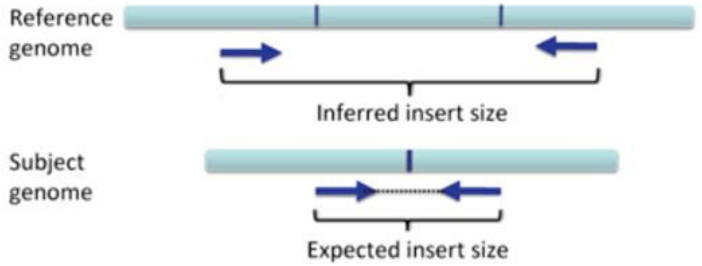
Greater insert size: Deletions

Deletions

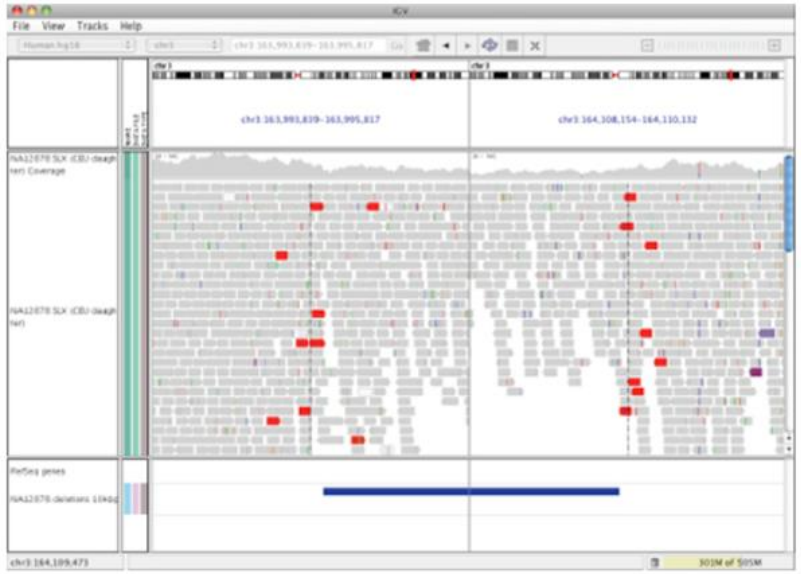
In a deletion a section of DNA is absent in the subject genome compared to the reference genome.



When pairs from a section of DNA spanning the deletion are aligned to the genome the inferred insert size will be larger than expected. This is due to the deleted section of the genome, not present in the subject. Schematically this can be visualized as follows:



So in the case of a deletion, the inferred insert size is GREATER THAN the expected insert size. In IGV such an event might look like the following.

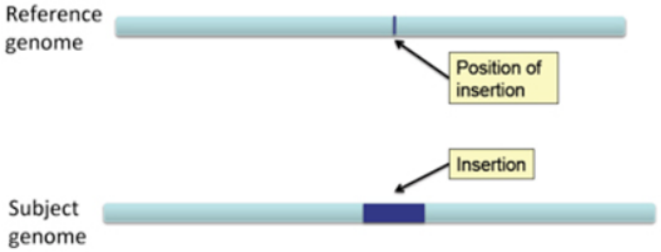


Reads that are colored red have larger than expected inferred sizes, and therefore indicate possible deletions.

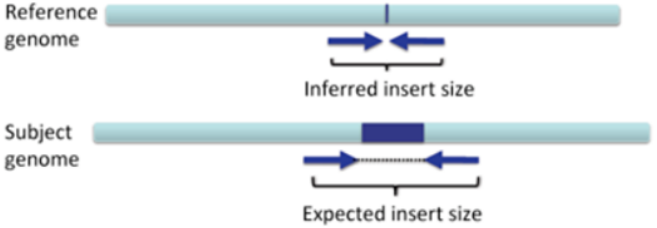
Smaller insert size: Insertions

Insertions

In the case of an insertion, a section of DNA is present in the subject genome that is not represented in the reference genome.



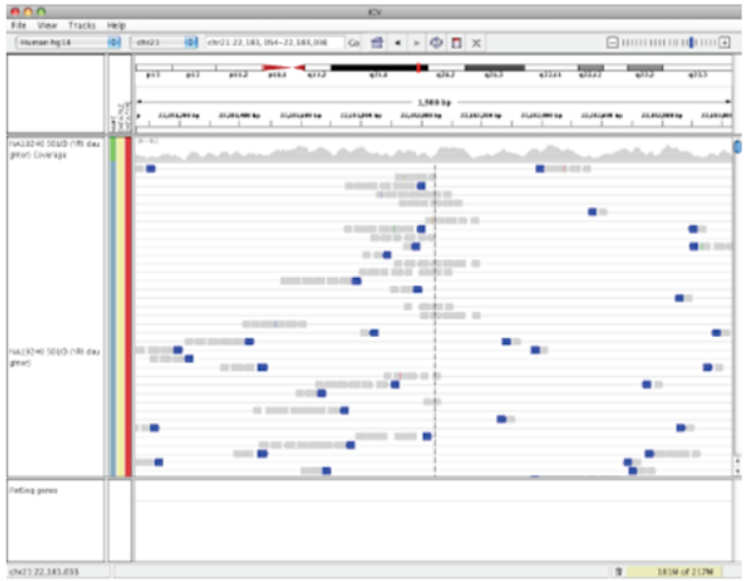
The effect on distance between aligned pairs is opposite in the case of a deletion; the "inferred insert size" is smaller than expected.



The maximum size of an insertion detectable by insert size anomaly is limited by the size of the fragments. They must be long enough to span the insertion and include sequences on both ends that are mapped to the reference. The maximum detectable size is approximately equal to:

fragment length - (2x read length)

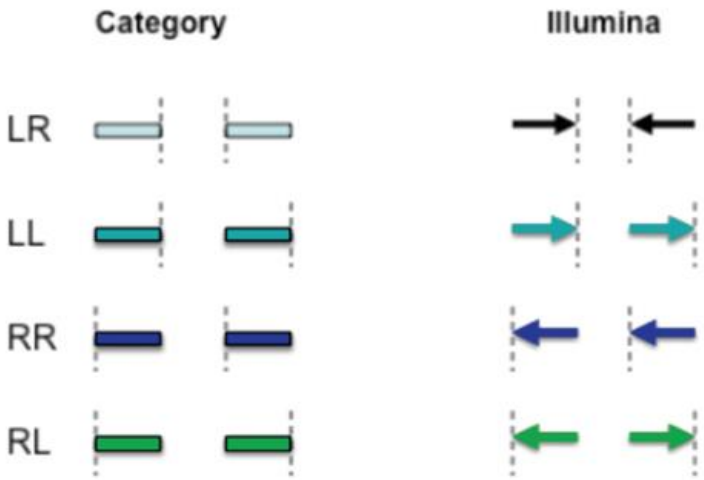
Detection of this event is therefore more likely with larger fragment libraries, such as Illumina mate-pair (not paired-end) and SOLID.



In the example above reads that are colored blue have smaller than expected inferred sizes, and therefore indicate insertions.

Read pair orientation

Interpretation of read pair orientations

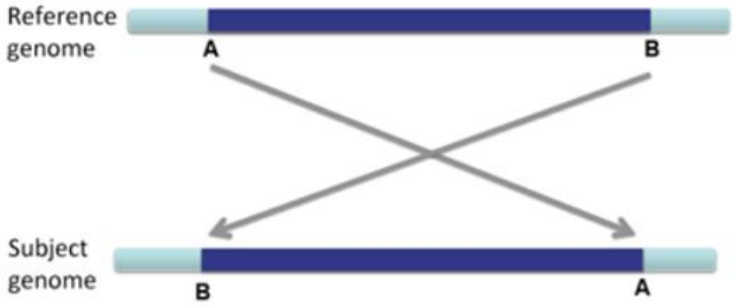


LR	Normal reads. The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
LL,RR	Implies inversion in sequenced DNA with respect to reference.
RL	Implies duplication or translocation with respect to reference.

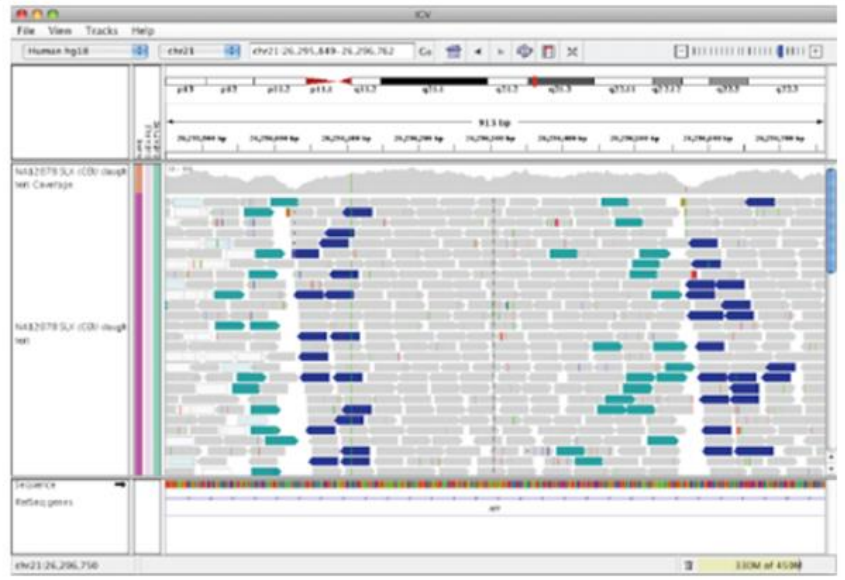
LL and RR read pair orientation: Inversions

Inversions

An inversion is a large section of DNA that is reversed in the subject genome compared to the reference genome.



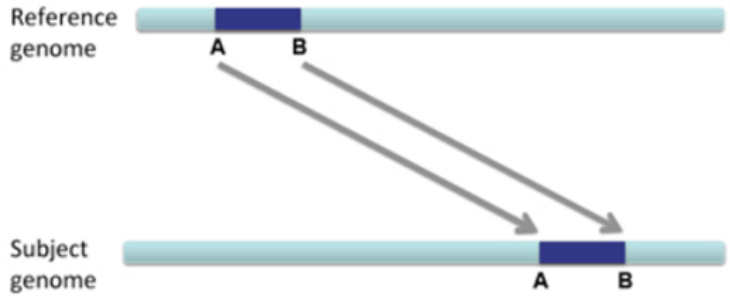
When an inversion shows up in paired-end reads, the reads are distinctively variant from the reference genome.



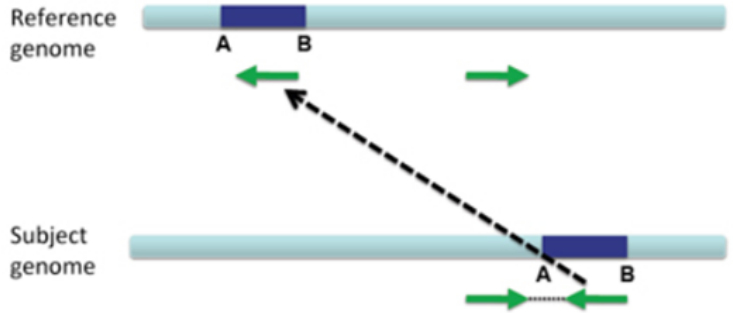
RL read pair orientation: Intra-chromosomal translocation

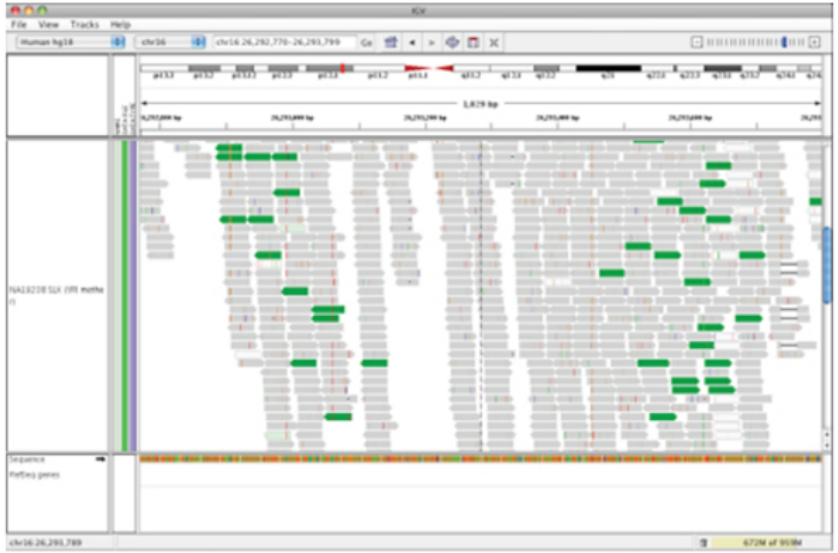
Translocation on the Same Chromosome

When a large section of DNA is removed from one location and inserted elsewhere, that is a translocation.

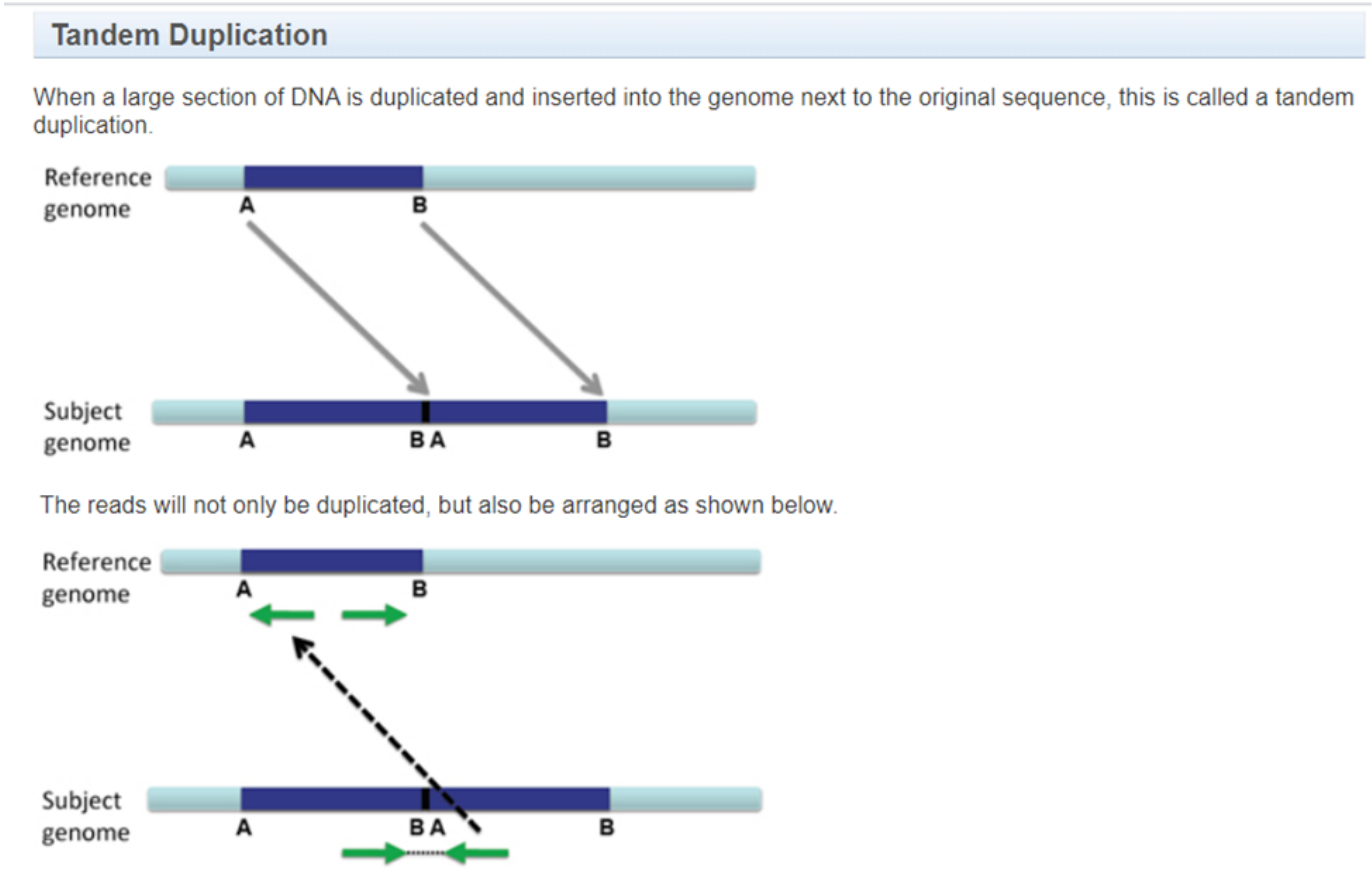


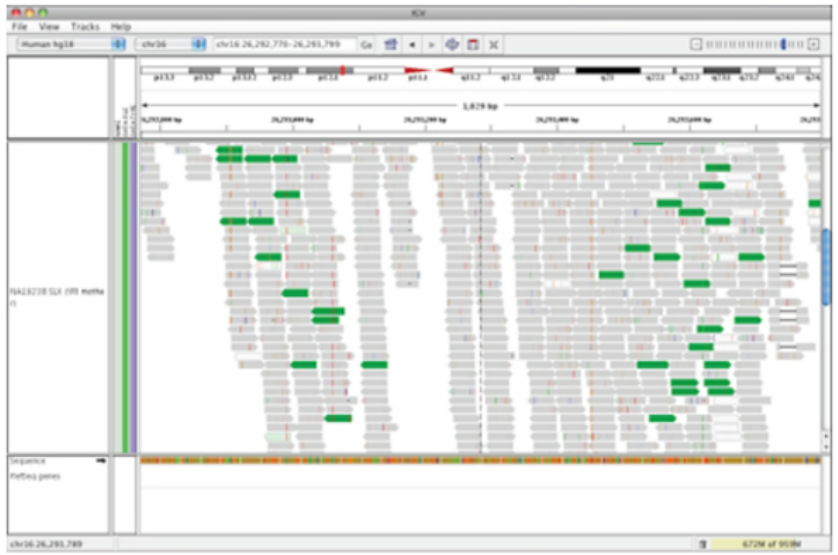
Translocations on the same chromosome can be detected by color-coding for pair orientation, whereas translocations between two chromosomes can be detected by coloring by insert size.





RL read pair orientation: Tandem duplications



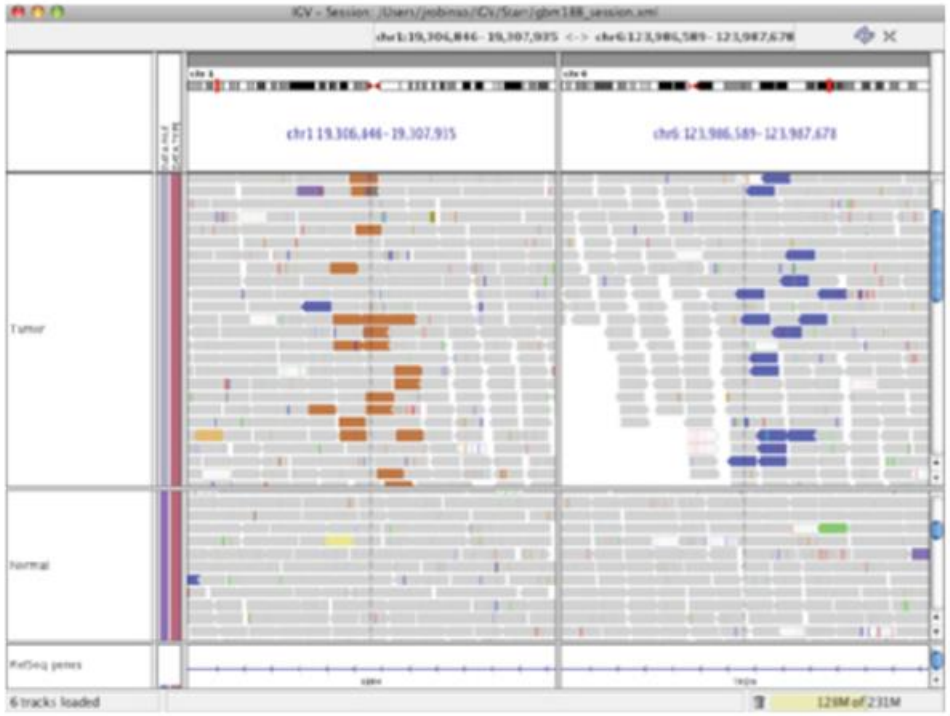


Tandem duplications can be distinguished from intra-chromosomal translocation by looking at coverage depth

Inter-chromosomal translocation

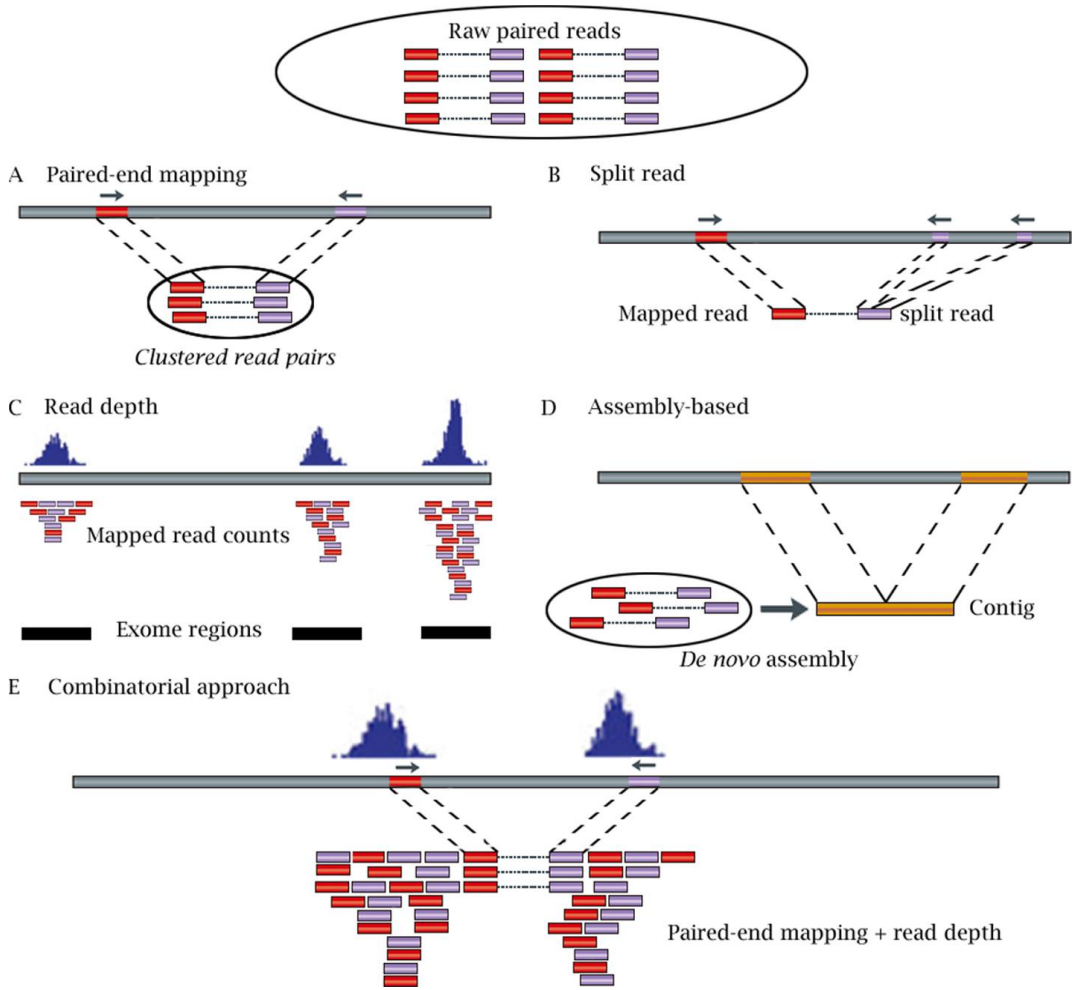
Inter-chromosomal Rearrangement

IGV codes inserts for inter-chromosomal rearrangements. For instance, in this case, one end is on chromosome 1 and the other is on chromosome 6.

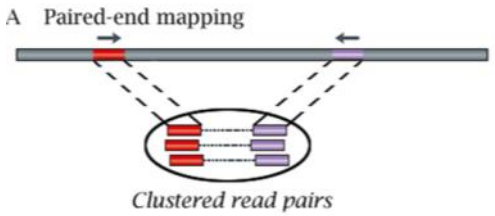


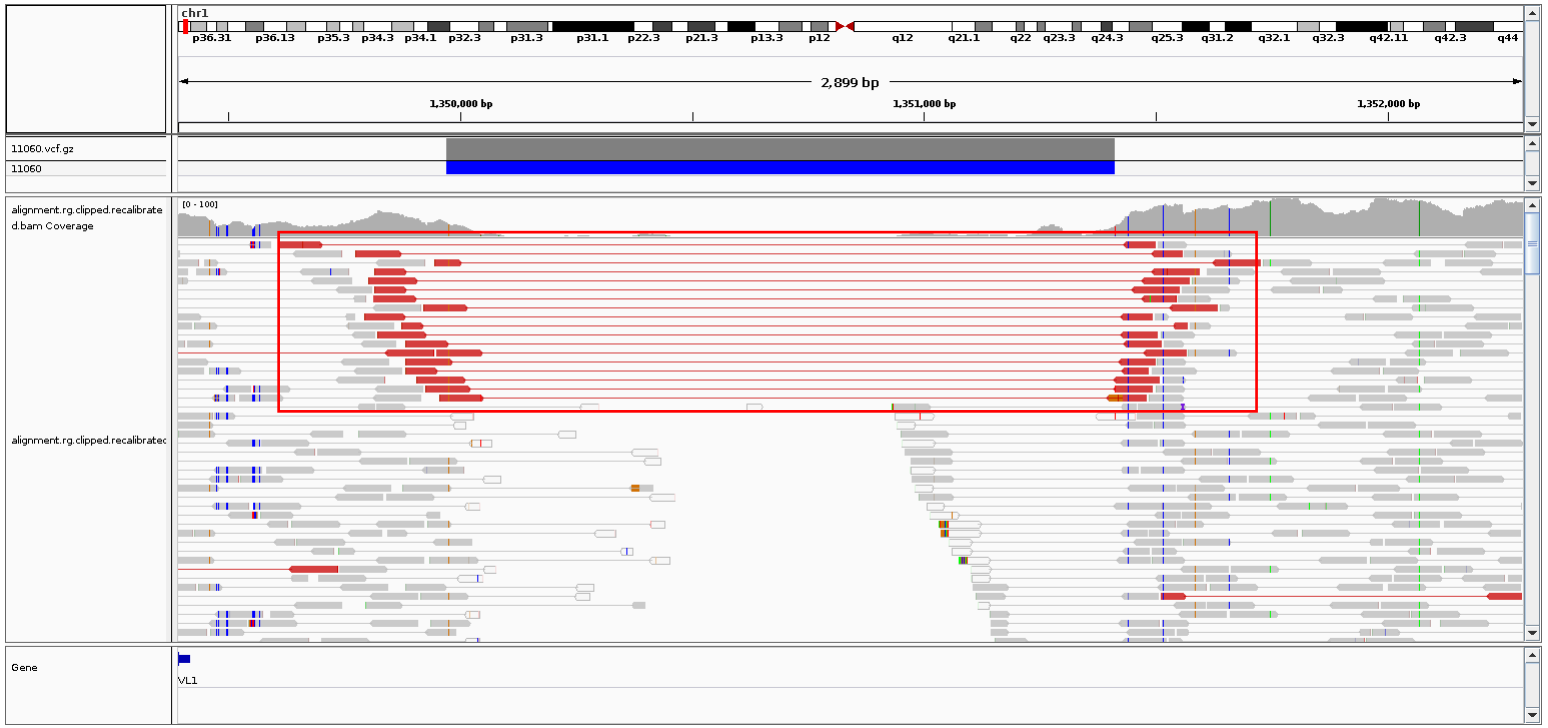
Approaches and software for SV detection

Many approaches are exploited by software for SVs identification



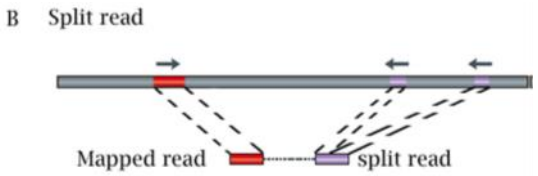
Paired-end mapping approach

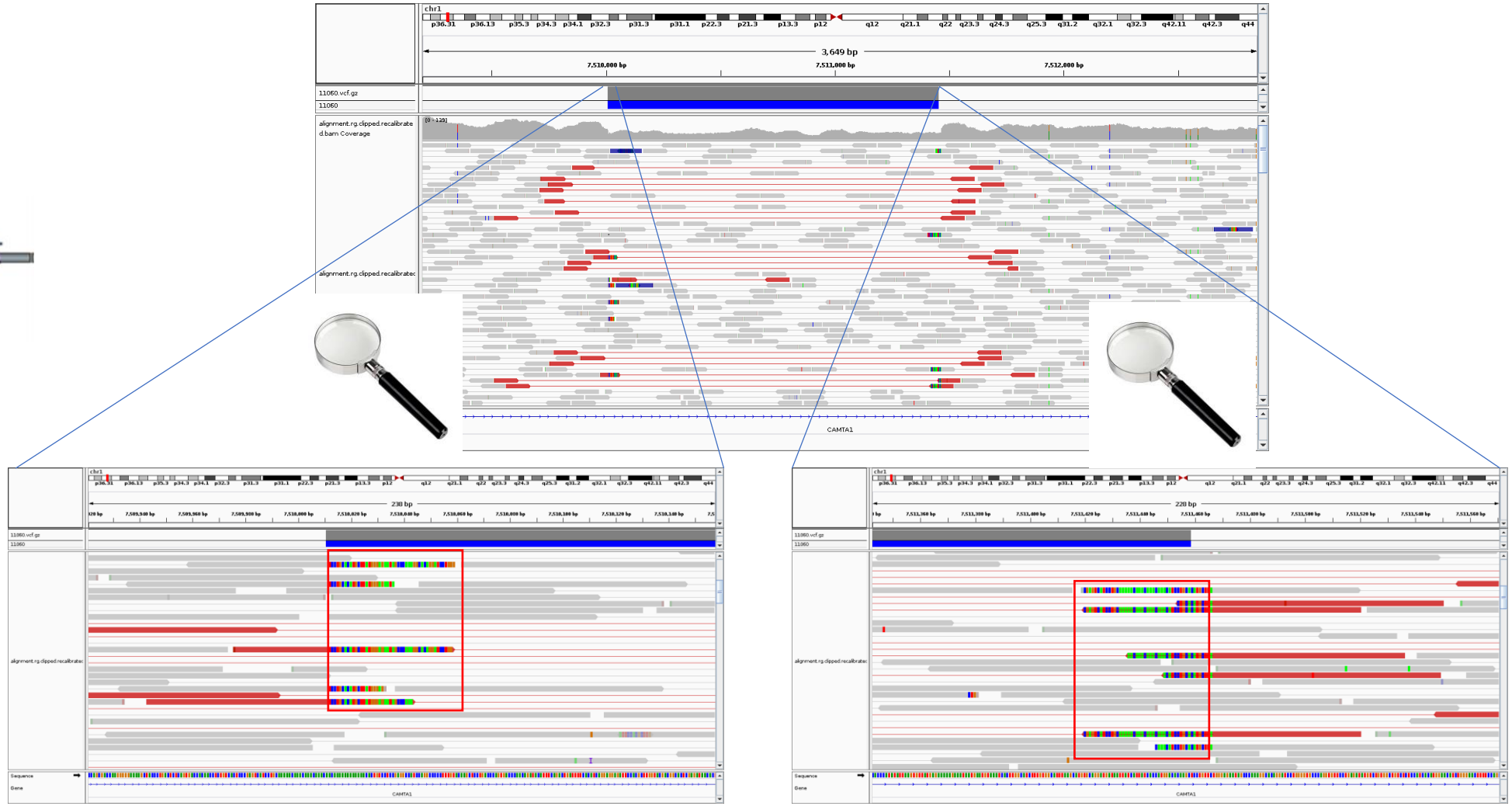




Paired-end reads with greater insert size suggest a deletion

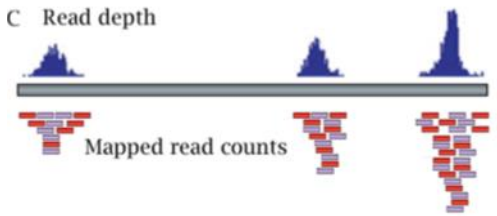
Split-read approach



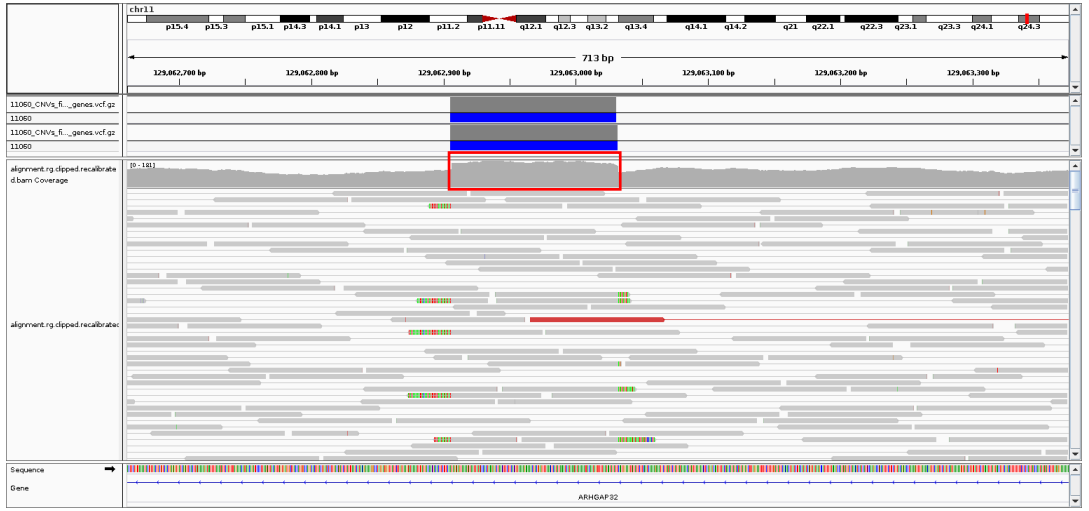


Soft-clipped portions of reads are split and mapped at the SV other breakpoint

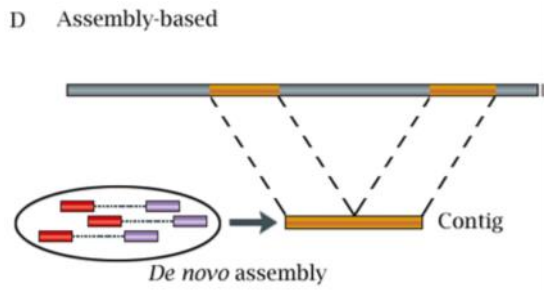
Read-depth approach

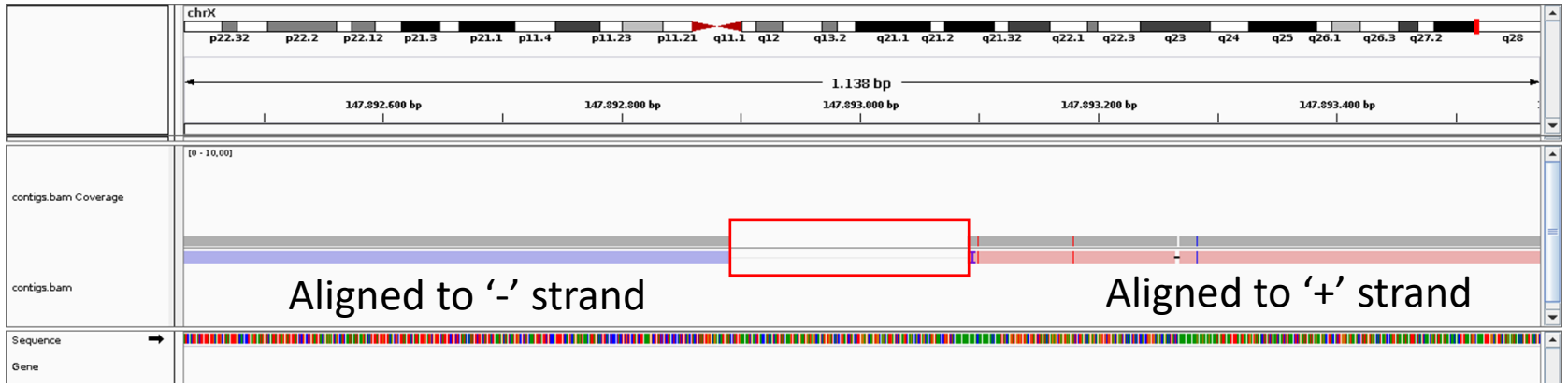


Decrease or increase in coverage depth suggest the presence of deletions or duplications



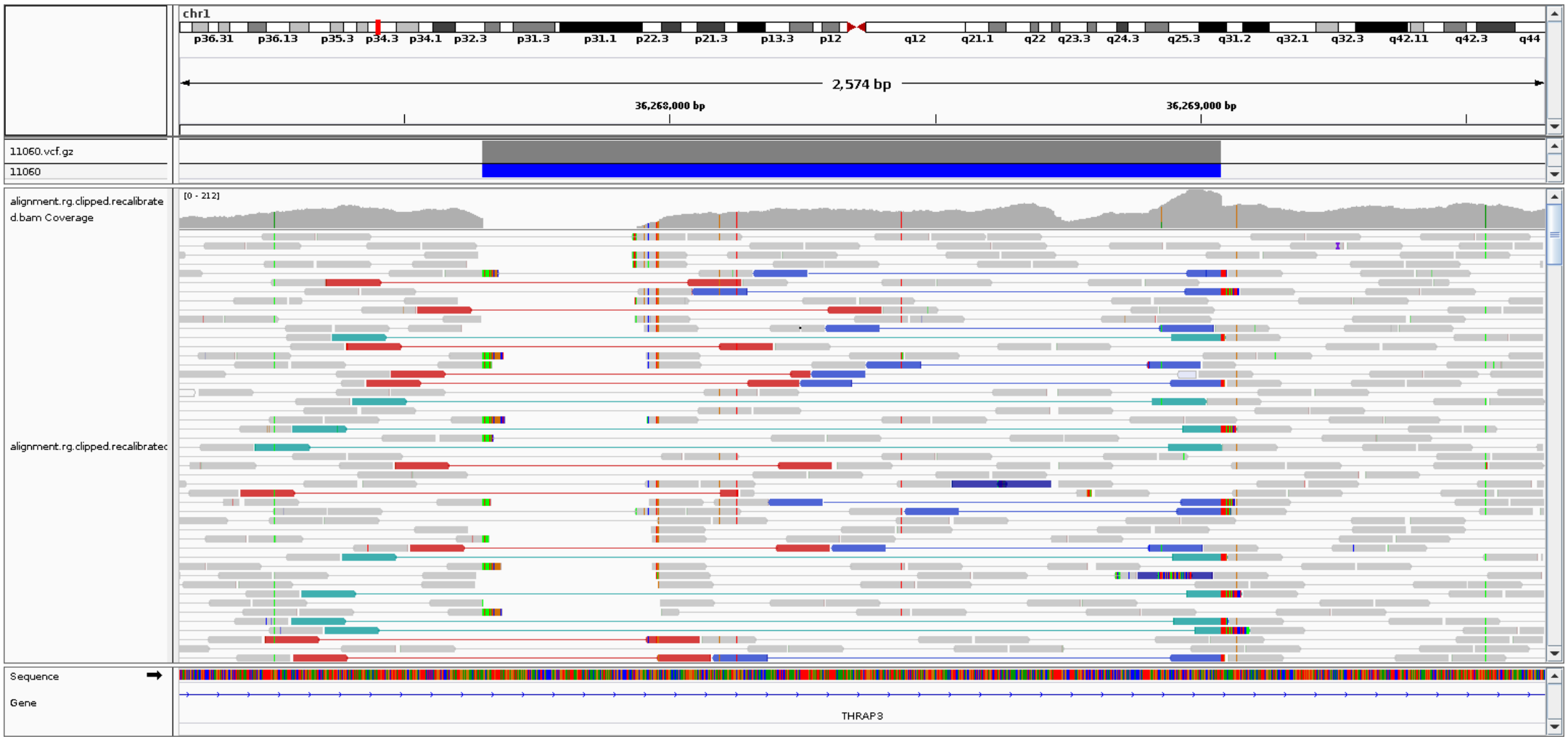
Assembly-based approach



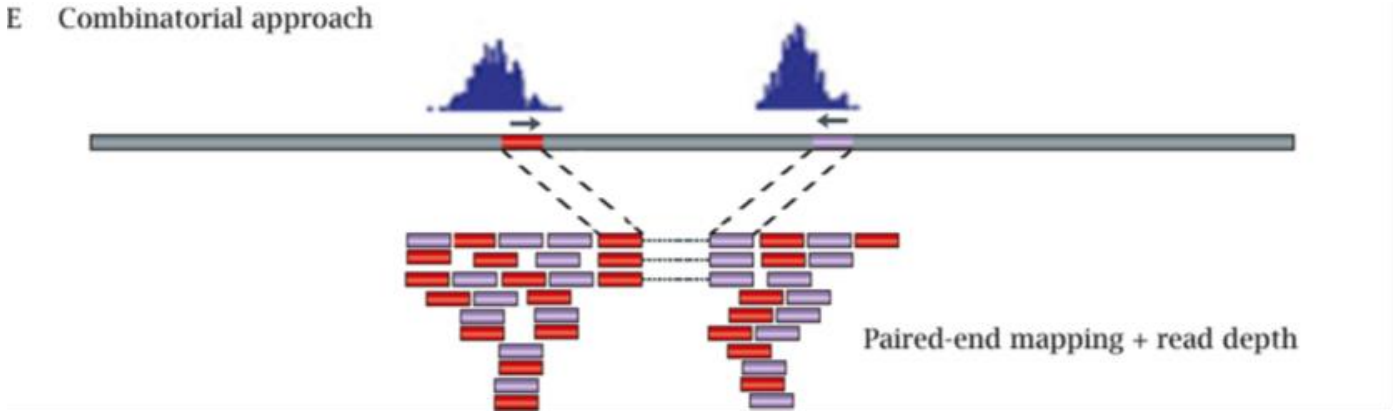


Two portions of the same contig aligned to different strands and a region with 0 coverage depth suggest the presence of an inversion and a deletion

Test: which SV signatures and SV types can you spot?



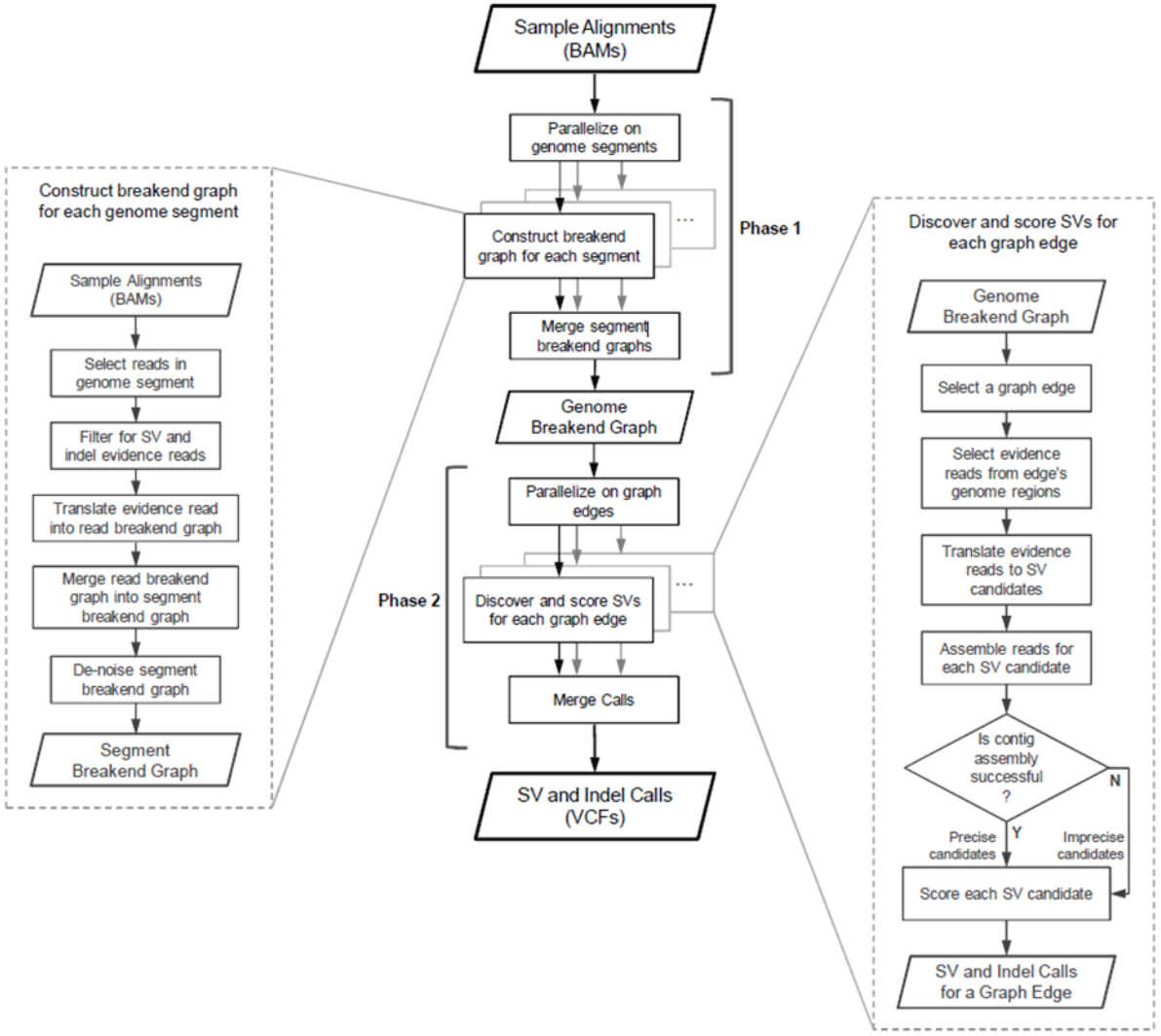
Best performing SV callers are based on a combinatorial approach exploiting multiple sources of evidence



Best performing SV callers:

- Manta
- Lumpy
- Others (Delly, GRIDSS, …)

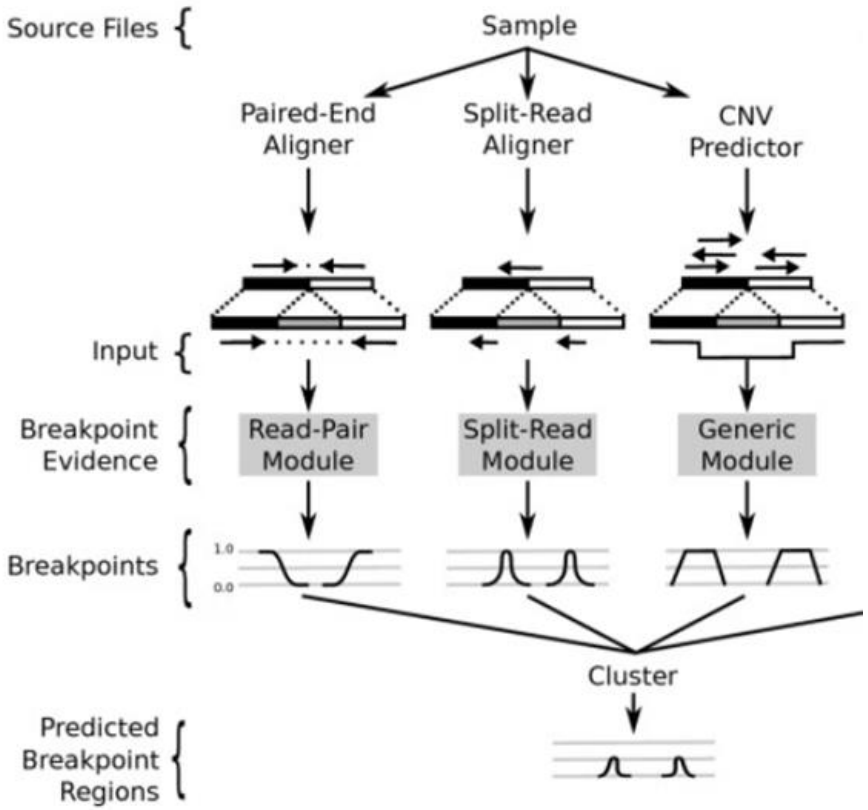
Manta workflow



- Scan bam file with reads aligned to reference
- Identify reads supporting SVs based on **paired-end mapping** and **split-read** approaches
- Construct a graph to represent candidate breakpoints and filter SV candidates
- Perform **local assembly** to refine breakpoints
- Report SVs in vcf file

Chen X et al. “Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications”. Bioinformatics. 2016 Apr 15;32(8):1220-2.
https://github.com/Illumina/manta

Lumpy workflow



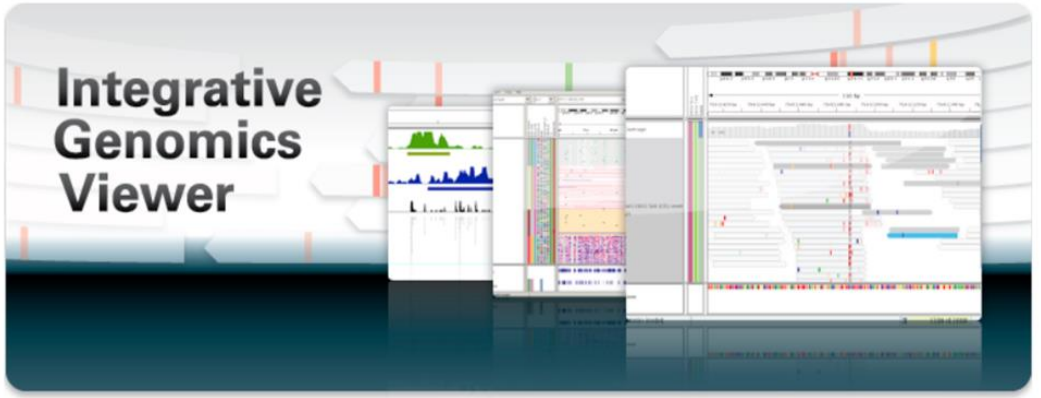
- Scan bam file with reads aligned to reference
- Identify reads supporting SVs based on **paired-end mapping**, **split-read** and **read-depth** approaches
- Cluster breakpoints identified with different approaches
- Report SVs in vcf file

Layer RM et al. “LUMPY: a probabilistic framework for structural variant discovery”. Genome Biol. 2014 Jun 26;15(6)
https://github.com/arq5x/lumpy-sv

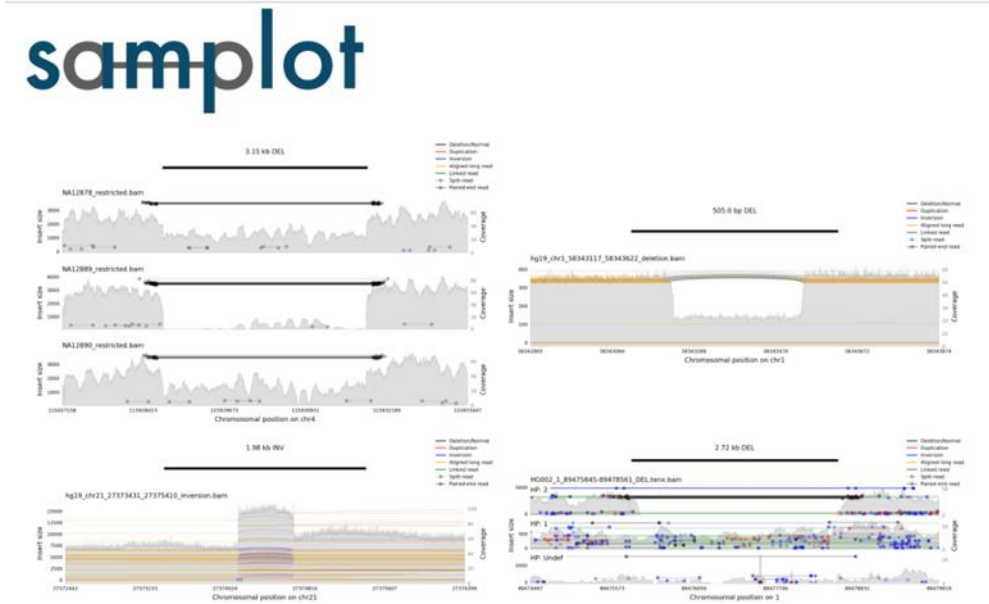
SV visual inspection

SV visual inspection is useful for spotting spurious calls

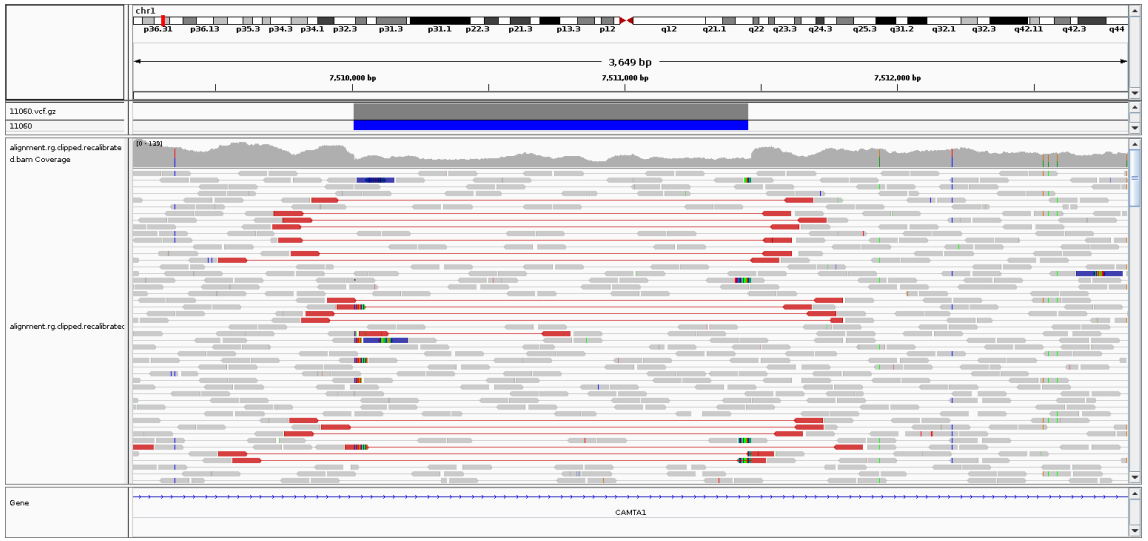
- Integrative Genomics Viewer (IGV)
  - General purpose genome browser
  - A graphical interface allows browsing the genome
  - Not suitable to inspect a huge amount of regions
  - http://software.broadinstitute.org/software/igv/



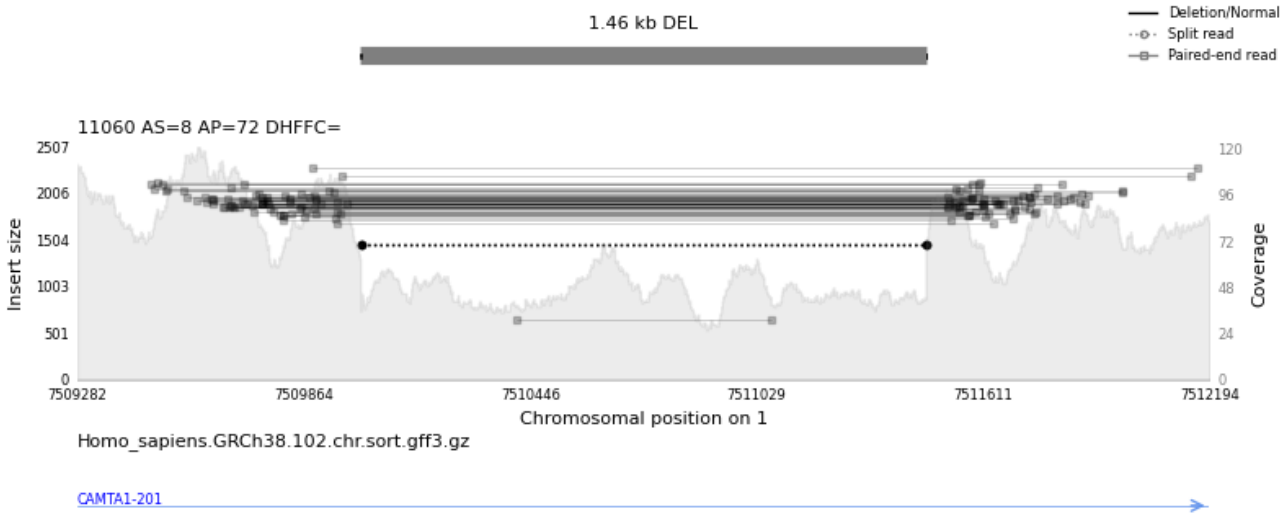
- Samplot
  - Command line tool
  - Specific for SV visual inspection
  - It highlights alignment and depth signals supporting the SV
  - https://github.com/ryanlayer/samplot



SV visualization: heterozygous deletion



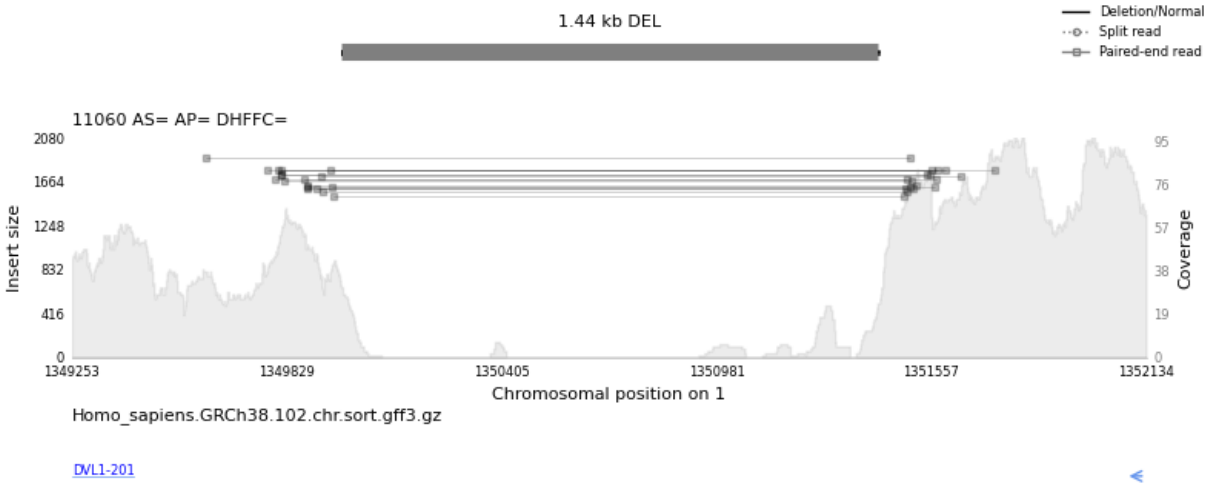
IGV



Samplot

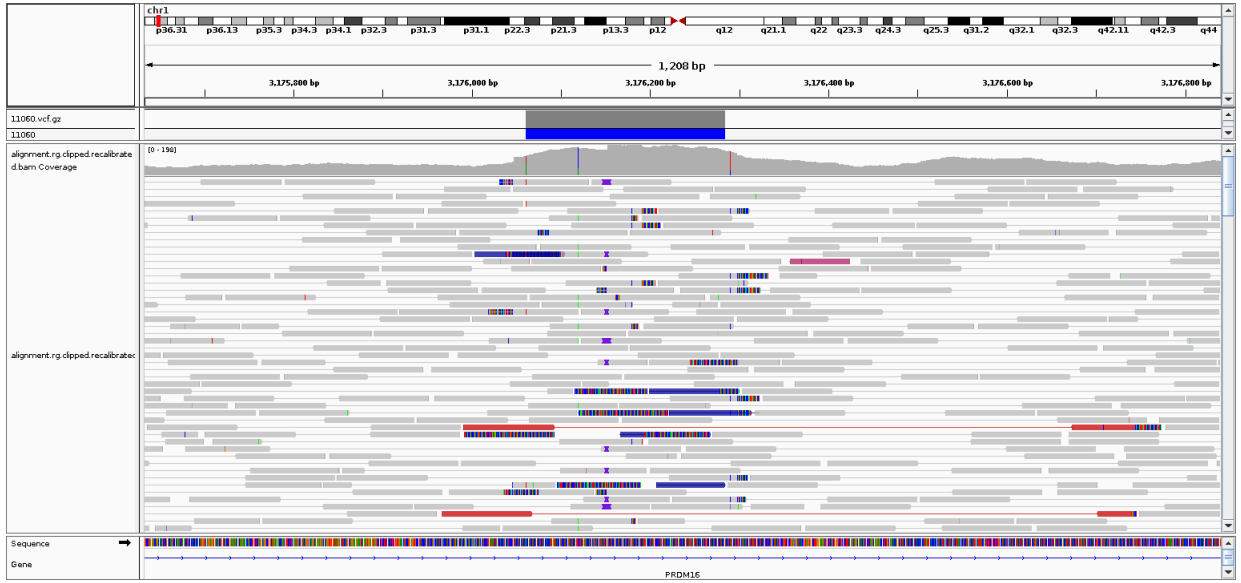
SV visualization: homozygous deletion

IGV

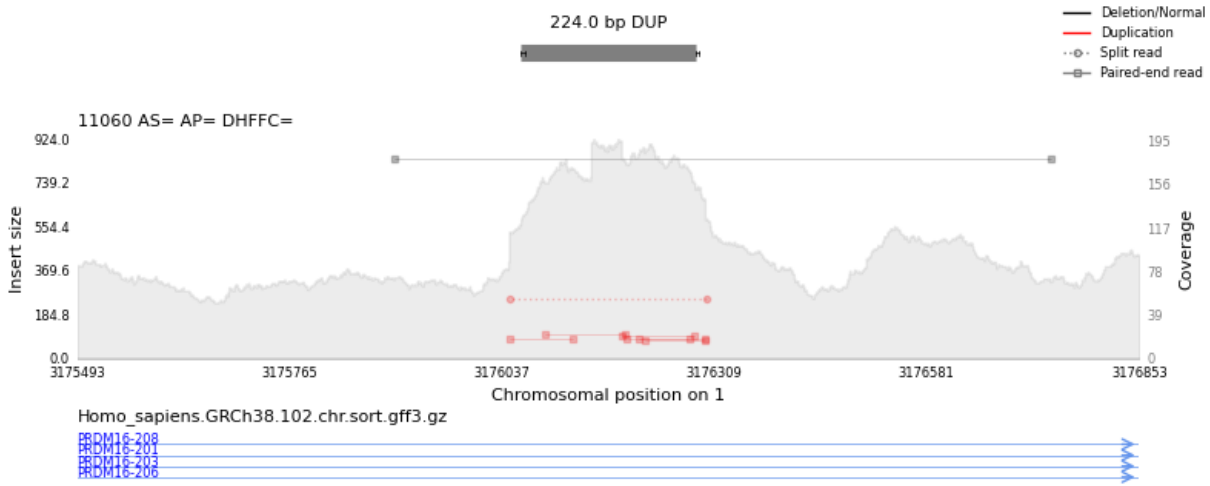


Samplot

SV visualization: duplication (repeat expansion)

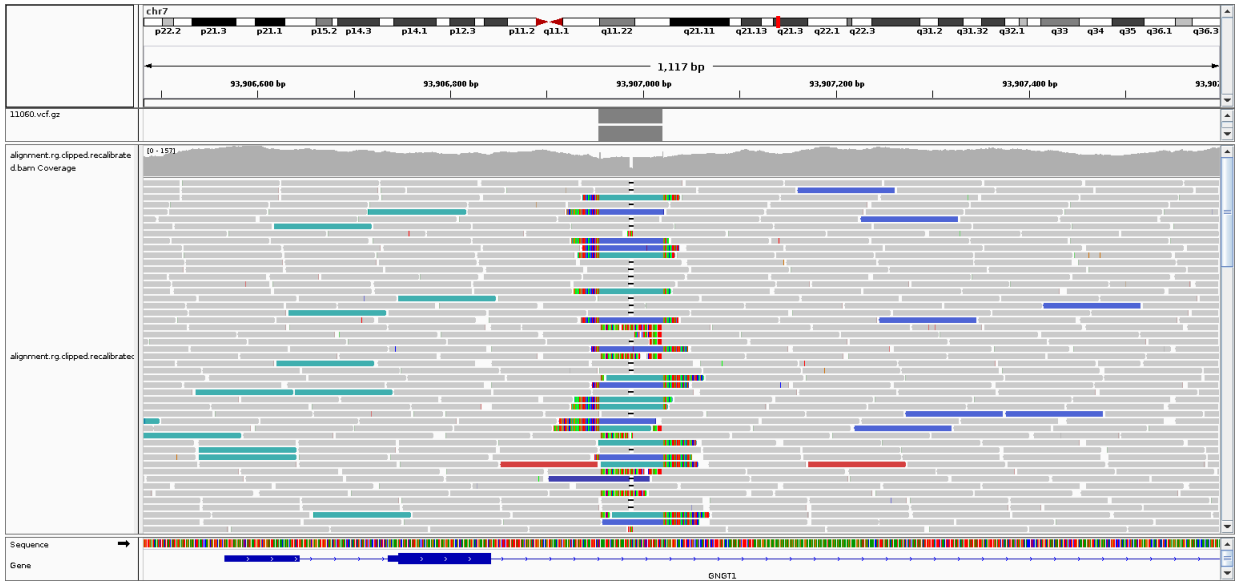


IGV

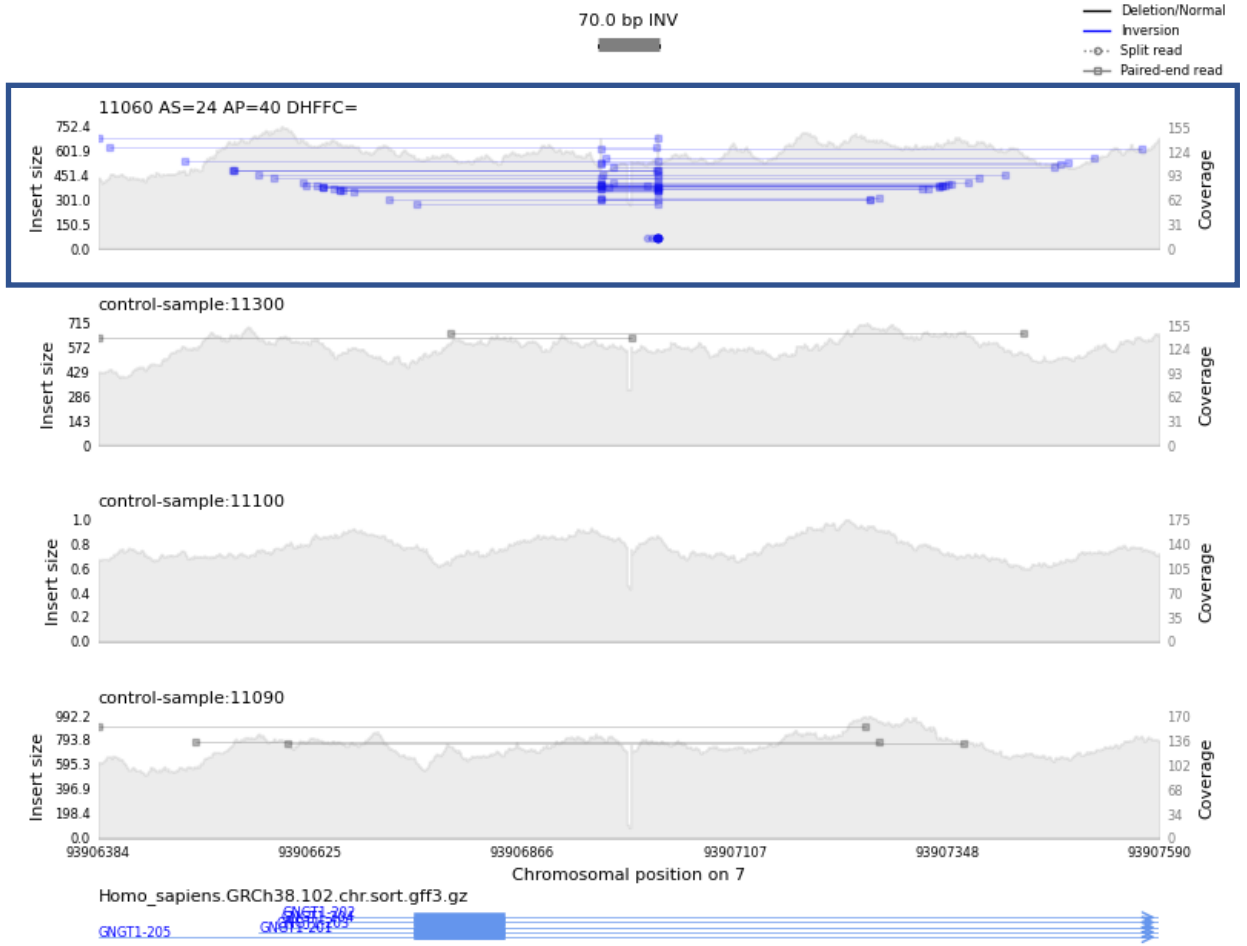


Samplot

SV visualization: inversion



IGV



Samplot

Inspecting additional «control» samples may help identifying spurious SV calls

Summary

- Exploiting many signatures and combining multiple approaches is necessary for accurate SV identification
- Best performing SV callers as Manta and Lumpy are based on a combinatorial approach exploiting multiple sources of evidence
- SV visual inspection with IGV or Samplot is useful for spotting spurious calls