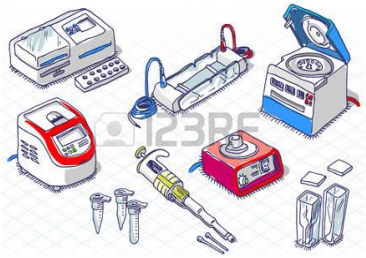
Human Genome Sequencing and Interpretation

Lesson 1 - 20/01/2020
Lesson 2 - 21/01/2020
Lesson 3 - 27/01/2020
(Lesson 4 - 28/01/2020)

Prof. Massimo Delledonne
Functional Genomics lab

Library preparation

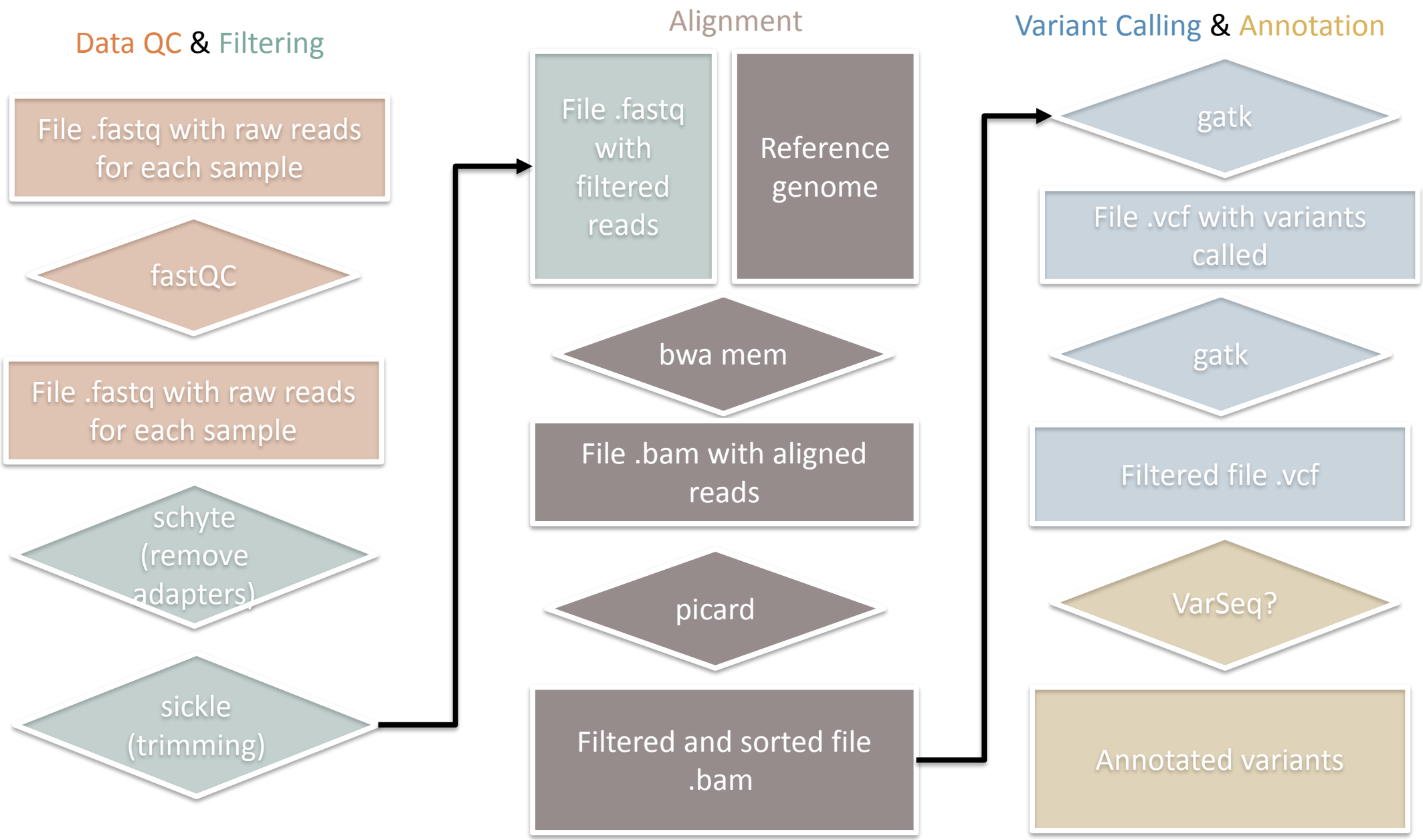
Bioinformatic analysis



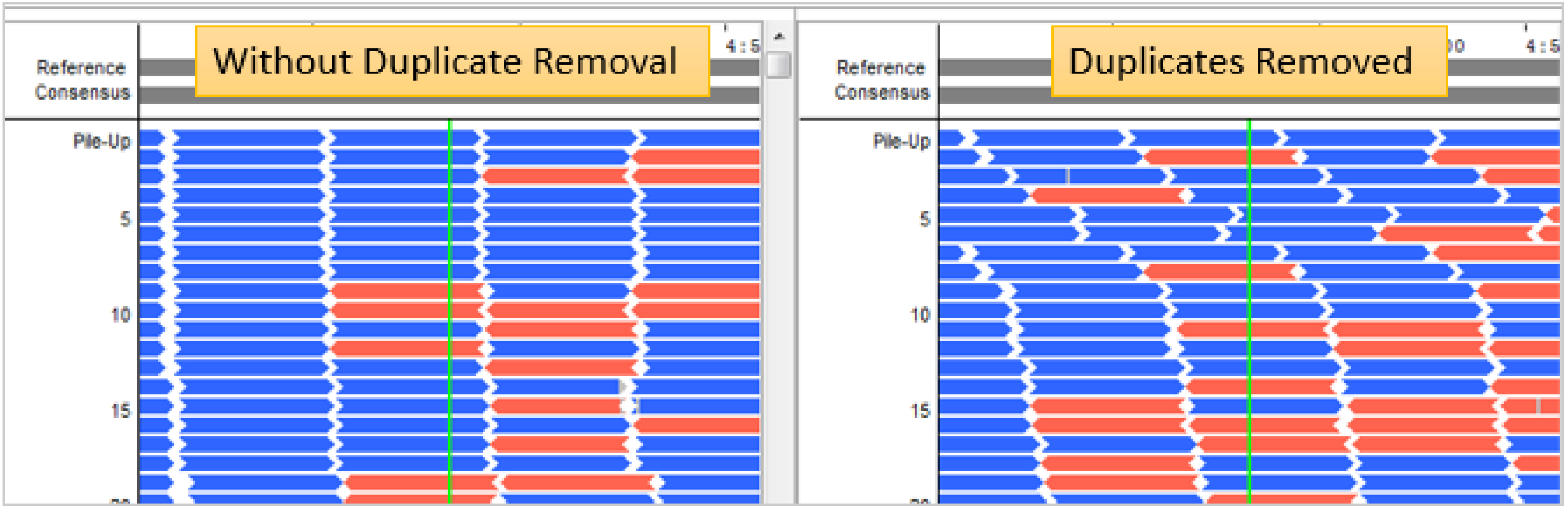
Sequencing

Data QC → Alignment → Variant calling → Variant prioritization

Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
schyte (remove adapters)
sickle (trimming)
Alignment
File .fastq with filtered reads
Reference genome
bwa mem
File .bam with aligned reads
picard
Filtered and sorted file .bam
Variant Calling & Annotation
gatk
File .vcf with variants called
gatk
Filtered file .vcf
VarSeq?
Annotated variants

Picard Mark Duplicates



Connect to the server

1. Enter in the server:
   a. ssh HGSI2020@157.27.26.214
   b. Password: hgsi2020

2. Enter in the folder: cd /attachedvolume/HGSI2020

3. Create your folder: cd your_name

Mark Duplicates

1. Mark Duplicates with picard:

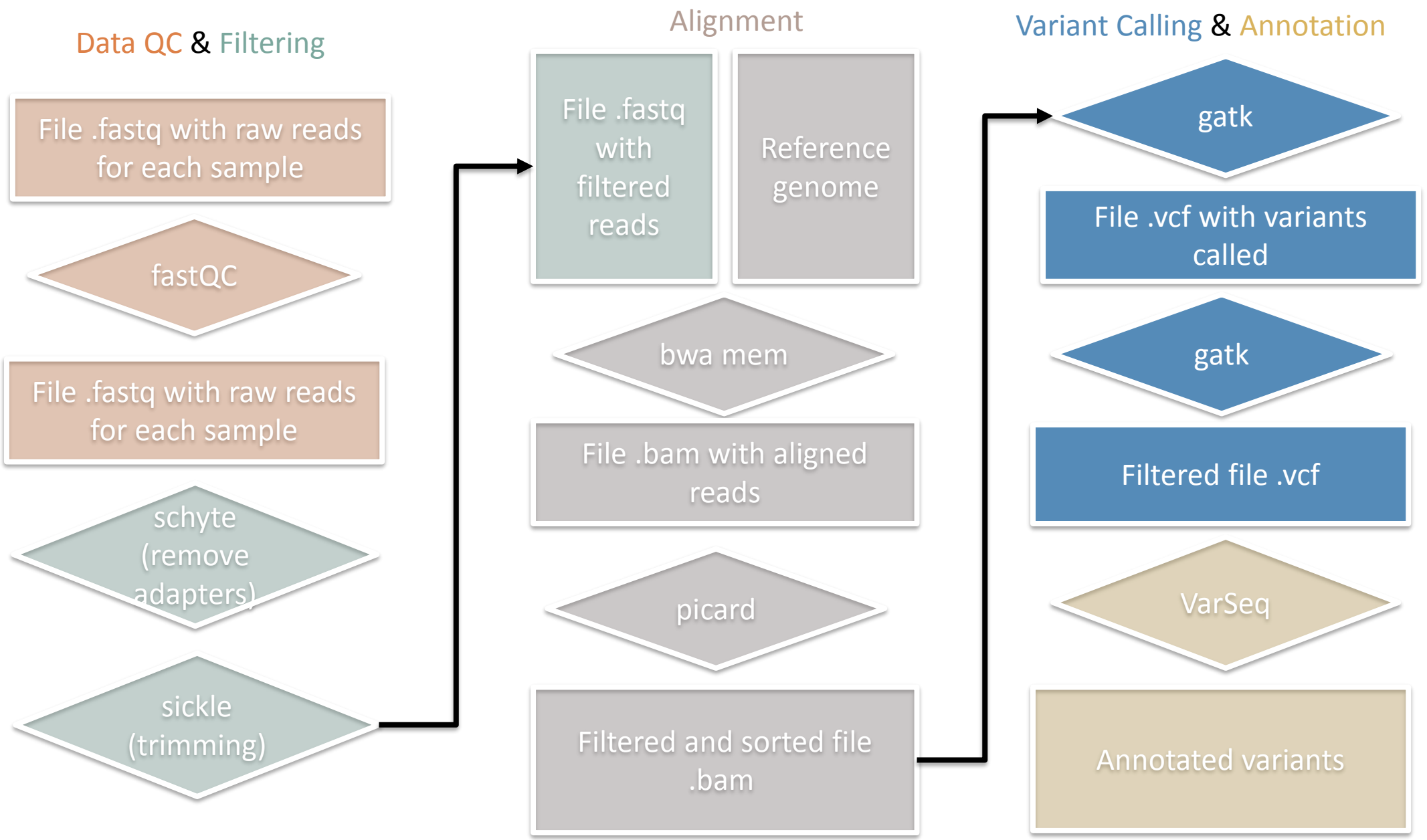
   java -jar /attachedvolume/HGSI2020/example/bin/picard-tools-2.0.1/picard.jar MarkDuplicates INPUT=sample.sorted.bam OUTPUT=sample.sorted.dedup.bam REMOVE_DUPLICATES=false METRICS_FILE=duplicates.txt

2. Open the output file: less –S duplicates.txt

```
## METRICS CLASS        picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED     UNMAPPE_READS  UNPAIRED_READ_DUPLICATES       READ_PAIR_DUPLICATES    READ_
Unknown Library 1       2480    1       0       172     116     0.069341        49106
```

Percentage of duplicates

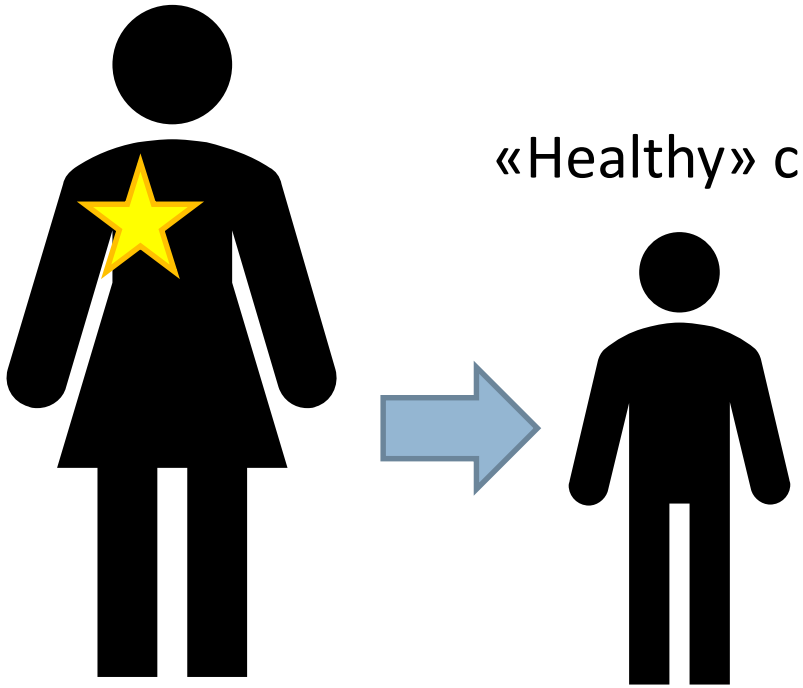
Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
schyte (remove adapters)
sickle (trimming)
Alignment
File .fastq with filtered reads
Reference genome
bwa mem
File .bam with aligned reads
picard
Filtered and sorted file .bam
Variant Calling & Annotation
gatk
File .vcf with variants called
gatk
Filtered file .vcf
VarSeq
Annotated variants

Different type of variants

Somatic Variant

Parent

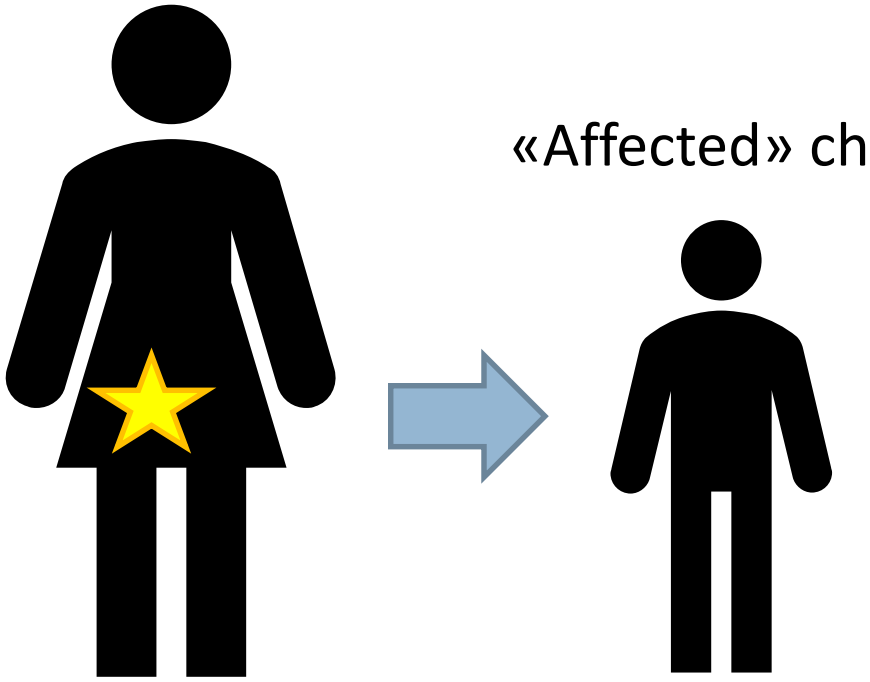


«Healthy» child

- Non germline tissue
- Not trasmitted to child

Germline Variant

Parent



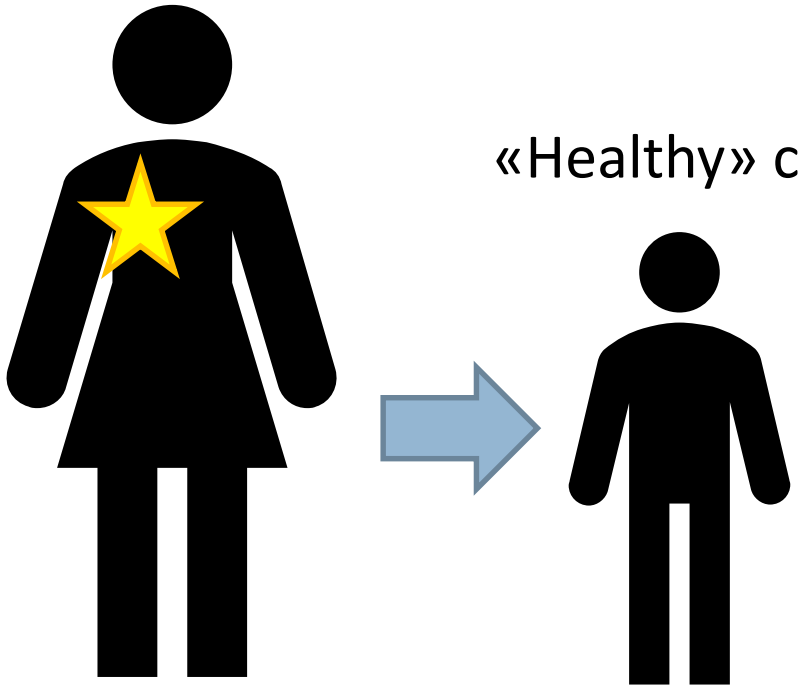
«Affected» child

- Mutation in egg or sperm
- Trasmitted to child

Different type of variants

Somatic Variant

Parent

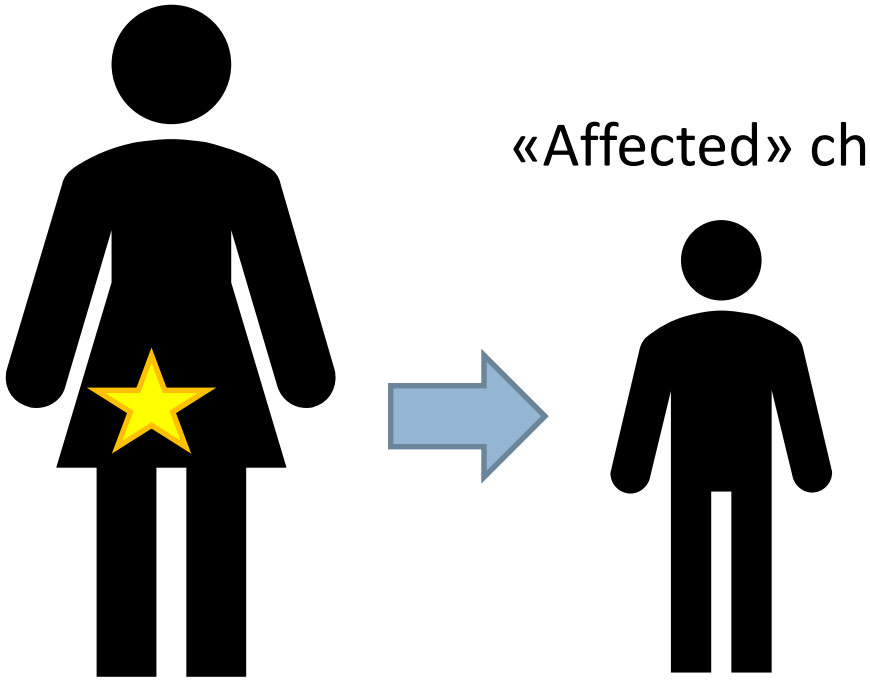


«Healthy» child

- Non germline tissue
- Not trasmitted to child

Germline Variant

Parent



«Affected» child

- Mutation in egg or sperm
- Trasmitted to child

Different type of variants

Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

Interspersed Duplication

Inversion

Translocation

Copy Number Variant

Different type of variants

Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

Interspersed Duplication

Inversion

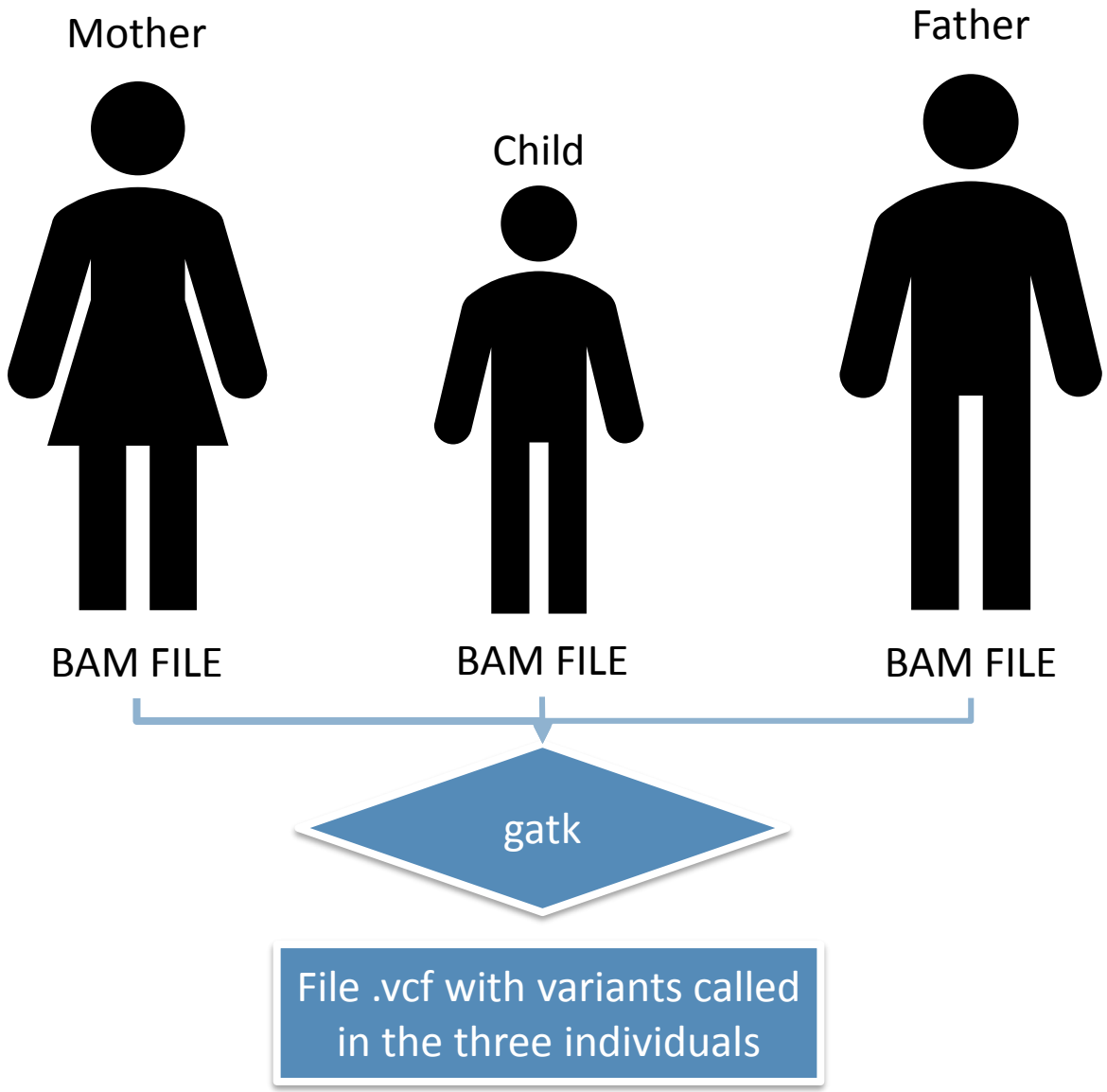
Translocation

Copy Number Variant

Variant Calling

- Starts from alignment data to find differences on the genome
- Decision to call a variant depends on many aspects:
  - Alignment quality
  - Read quality
  - Base coverage
  - …
- Many software are
- available:
  - SOAP2
  - SamTools
  - **GATK**
  - IVC
  - …

Family Analysis



Variant Calling with GATK

1. Enter in the folder: cd /attachedvolume/HGSI2020/ your_name
2. Remove old files:

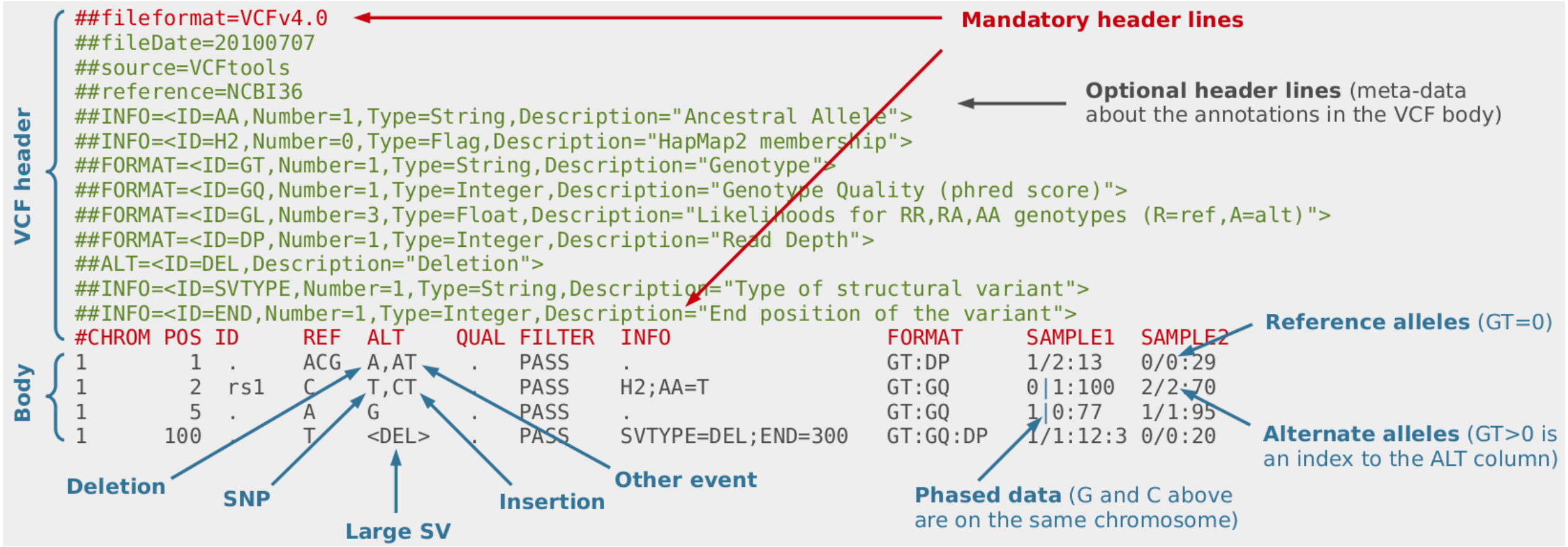
   rm trimmed* sample* duplicates.txt
3. Call variants on the family:

   java -jar /attachedvolume/HGSI2020/example/bin/GenomeAnalysisTK-3.3-0.jar -T UnifiedGenotyper
   -R /attachedvolume/HGSI2020/example/reference/chr6.hg38.fa
   -o snp.raw.vcf
   -I /attachedvolume/HGSI2020/example/samples/1351S/1351S.hg38.chr6.bam
   -I /attachedvolume/HGSI2020/example/samples/1352S/1352S.hg38.chr6.bam
   -I /attachedvolume/HGSI2020/example/samples/1353S/1353S.hg38.chr6.bam
   -L /attachedvolume/HGSI2020/example/reference/chr6.hg38.bed

   bash /attachedvolume/HGSI2020/example/scripts/step3.varCal.sh

4. Open the file: less -S snp.raw.vcf

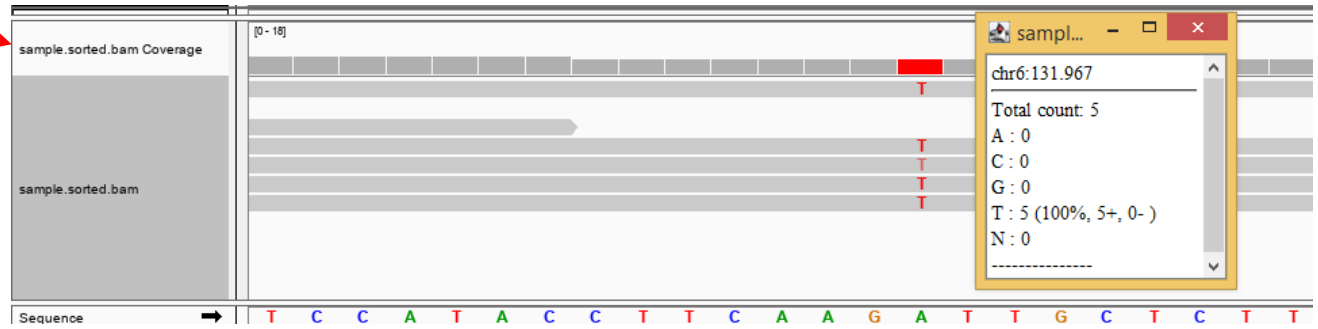
VCF


VCF header
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
Body
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Variant Calling with GATK

1. Filter variants:

```
java -jar /attachedvolume/HGSI2020/example/bin/GenomeAnalysisTK-3.3-0.jar
-T VariantFiltration
-R /attachedvolume/HGSI2020/example/reference/chr6.hg38.fa --variant snp.raw.vcf
-o snp.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'
```

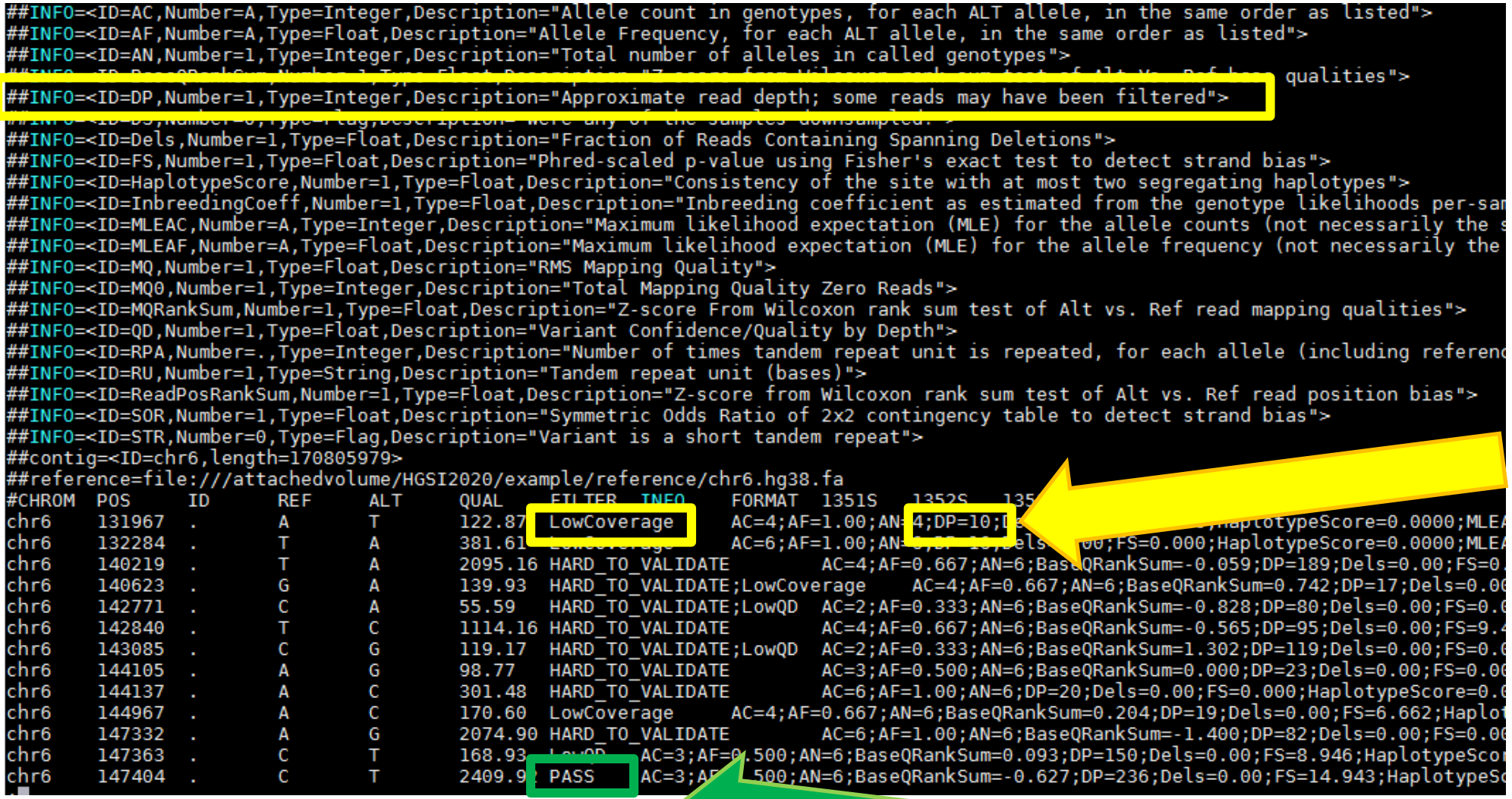



```
bash /attachedvolume/HGSI2020/example/scripts/step4.filterVariants.sh
```

2. Open the file:

```
less -S snp.filtered.vcf
```

Filtered VCF



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 /attachedvolume/HGSI2020/example/bin/GenomeAnalysisTK-3.3-0.jar
-T SelectVariants
-R /attachedvolume/HGSI2020/example/reference/chr6.hg38.fa
--variant snp.filtered.vcf
--excludeFiltered
-o selected.variants.vcf
```

```
bash /attachedvolume/HGSI2020/example/scripts/step4.varCal.sh
```

2. Open the file:

```
zless –S selected.variants.vcf
```

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       22417.9 0       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       19506.9 0       PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=3.372;DP=731;Dels=0.00;FS=0.000;Haplot
chr6    325961  .       T       C       18061.9 0       PASS    AC=6;AF=1.00;AN=6;BaseQRankSum=3.651;DP=742;Dels=0.00;FS=6.004;Haplot
chr6    326134  .       G       A       21980.9 0       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
```

GVCF

- Open the file: zless /attachedvolume/HGSI2020/example/samples/1351S/1351S.genome.vcf.gz

Header

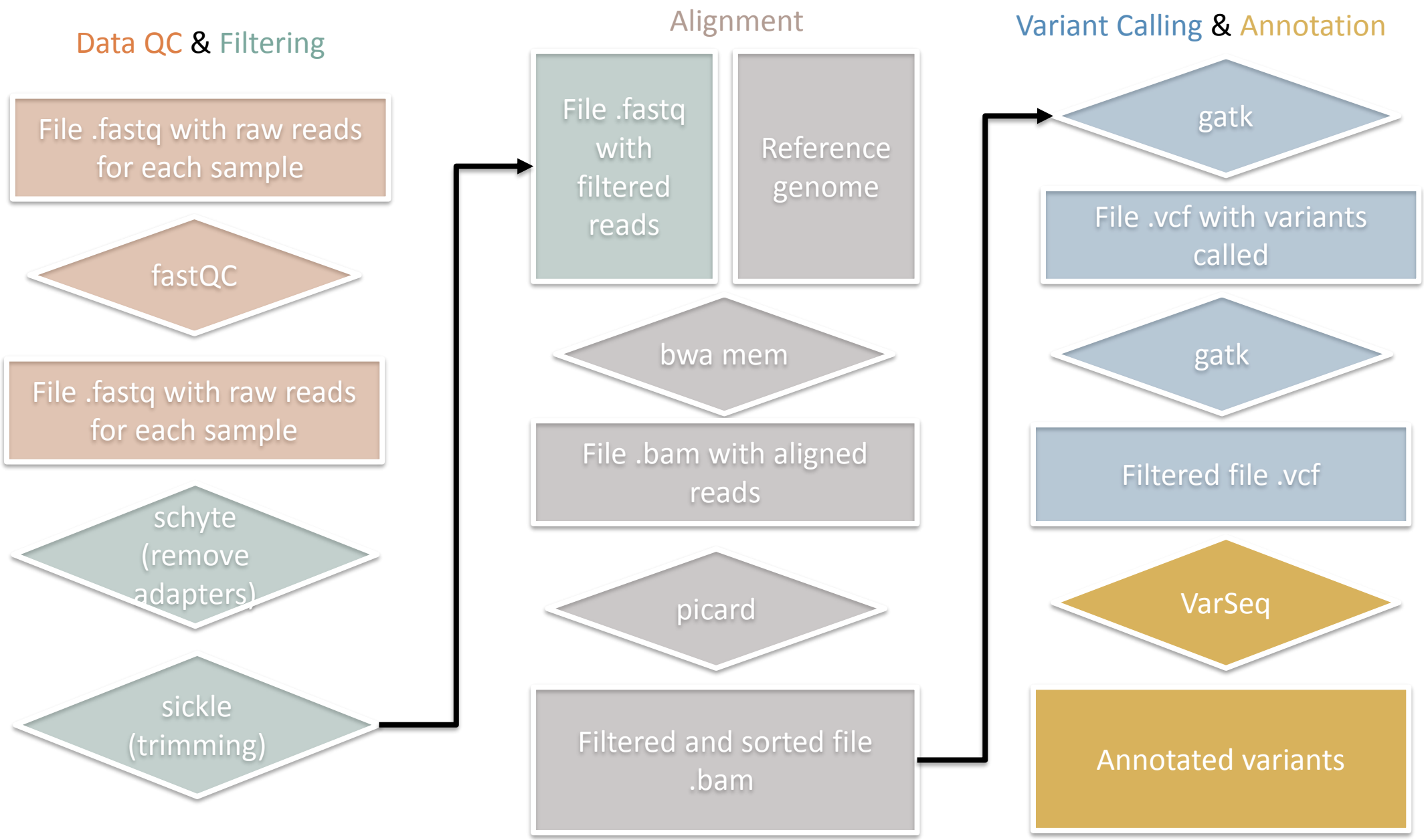
Body

```
##FORMAT=<ID=SB,Number=1,Type=Float,Description="Sample site strand bias">
##FILTER=<ID=IndelConflict,Description="Indel genotypes from two or more loci conflict in at least one sample">
##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">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  1351S
chr1    65565   .       A       .       .       LowGQX  END=69461;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr1    69462   .       C       .       .       LowGQX;LowDepth END=69510;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr1    69511   .       A       G       1       LowGQX;LowDepth SNVHPOL=3;MQ=25 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:3:0:1:0:0
,1:0,1:0,0:0.0:LowGQX;LowDepth:30,3,0
chr1    69512   .       C       .       .       LowGQX;LowDepth END=69831;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr1    69832   .       G       .       .       LowGQX  END=70000;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr1    70001   .       A       .       .       LowGQX;LowDepth END=70008;BLOCKAVG_min30p3a     GT:GQX:DP:DPF:MIN_DP    0/0:3:1:0:1
chr1    182709  .       A       .       .       PASS    END=182746;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:237:90:1:80
chr1    183114  .       G       .       .       PASS    END=183240;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:107:70:1:45
chr1    183922  .       A       .       .       PASS    END=183932;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP
chr1    183933  .       C       .       .       PASS    END=184158;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:210:161:1:143
chr1    450740  .       T       .       .       LowGQX  END=451678;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr1    586839  .       T       .       .       LowGQX  END=586955;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr1    601398  .       C       .       .       PASS    END=601427;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:21:9:0:8
chr1    601428  .       A       .       .       LowGQX  .       GT:GQX:DP:DPF:MIN_DP    0/0:4:10:0:10
chr1    601429  .       G       .       .       PASS    END=601459;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:15:11:0:8
chr1    601460  .       T       .       .       PASS    END=601514;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP
chr1    601515  .       T       C       12      LowGQX  SNVHPOL=4;MQ=18 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:44:12:29:2:21,8:
6,5:5,3:-4.2:LowGQX:46,0,181
chr1    601516  .       G       .       .       PASS    END=601543;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:68:31:2:29
chr1    601544  .       G       A       26      PASS    SNVHPOL=5;MQ=20 GT:GQ:GQX:DP:DPF:AD:ADF:ADR:SB:FT:PL    0/1:59:26:32:1:22,10
14,5:8,5:-5.3:PASS:60,0,190
chr1    601545  .       T       .       .       PASS    END=601577;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:78:32:1:30
chr1    607955  .       C       .       .       LowGQX  END=609082;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    .:.:0:0:0
chr1    609083  .       C       .       .       PASS    END=609142;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:87:42:0:30
chr1    609143  .       T       .       .       PASS    .       GT:GQX:DP:DPF:MIN_DP    0/0:46:52:1:52
chr1    609144  .       C       .       .       PASS    END=609246;BLOCKAVG_min30p3a    GT:GQX:DP:DPF:MIN_DP    0/0:156:91:1:53
```

Reference Position

Variant

Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
schyte (remove adapters)
sickle (trimming)
Alignment
File .fastq with filtered reads
Reference genome
bwa mem
File .bam with aligned reads
picard
Filtered and sorted file .bam
Variant Calling & Annotation
gatk
File .vcf with variants called
gatk
Filtered file .vcf
VarSeq
Annotated variants

Download the file on your PC

- Open new terminal and enter in your folder:

  cd Desktop/HGSI2020/

- Download gvcf file:

  rsync -auv HGSI2020@157.27.26.214:/attachedvolume/HGSI2020/example/samples/*/*.genome.vcf.gz* .

- Download VCF file:

  rsync -auv HGSI2020@157.27.26.214:/attachedvolume/HGSI2020/example/samples/selected.variants.chr6.vcf.gz* .

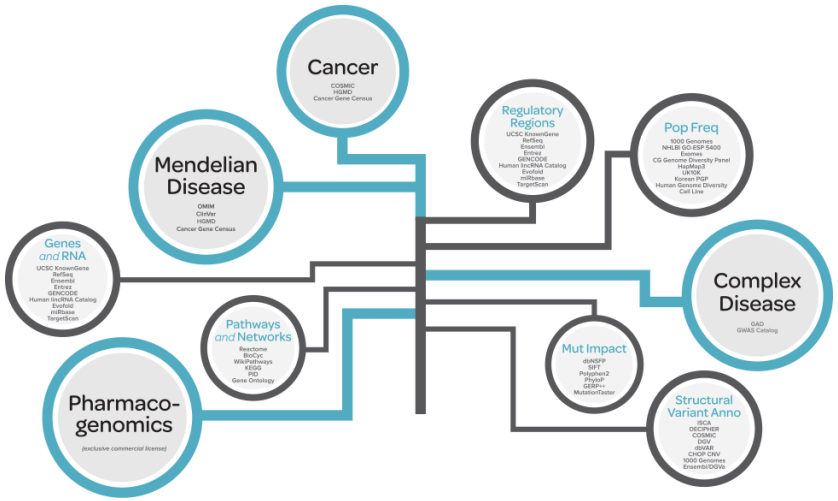
- Download excel file:

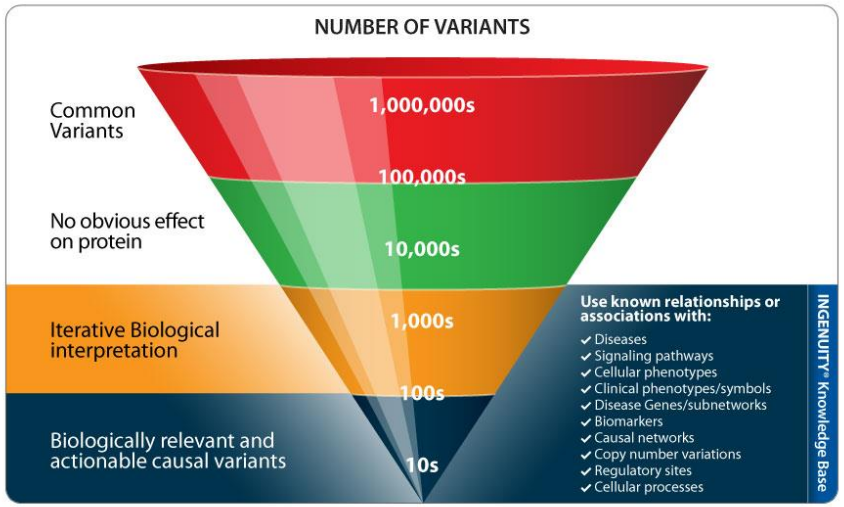
  rsync -auv HGSI2020@157.27.26.214:/attachedvolume/HGSI2020/example/samples/Lezioni2020_trio_1351S_chr6.xlsx .

Variant Annotation

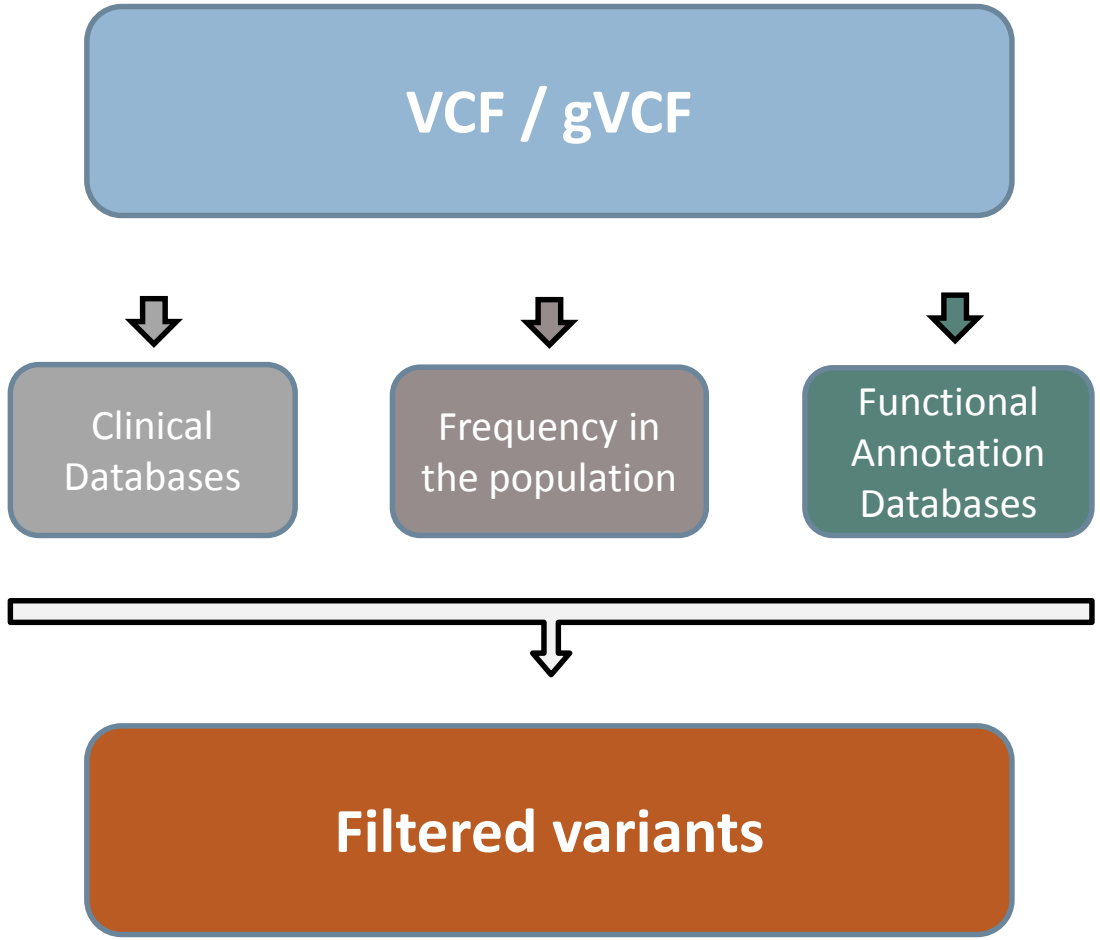
→ The process of assigning functional information to DNA variants to aid the identification of disease-causing mutations

→ Based on different databases and literature





Variant Annotation

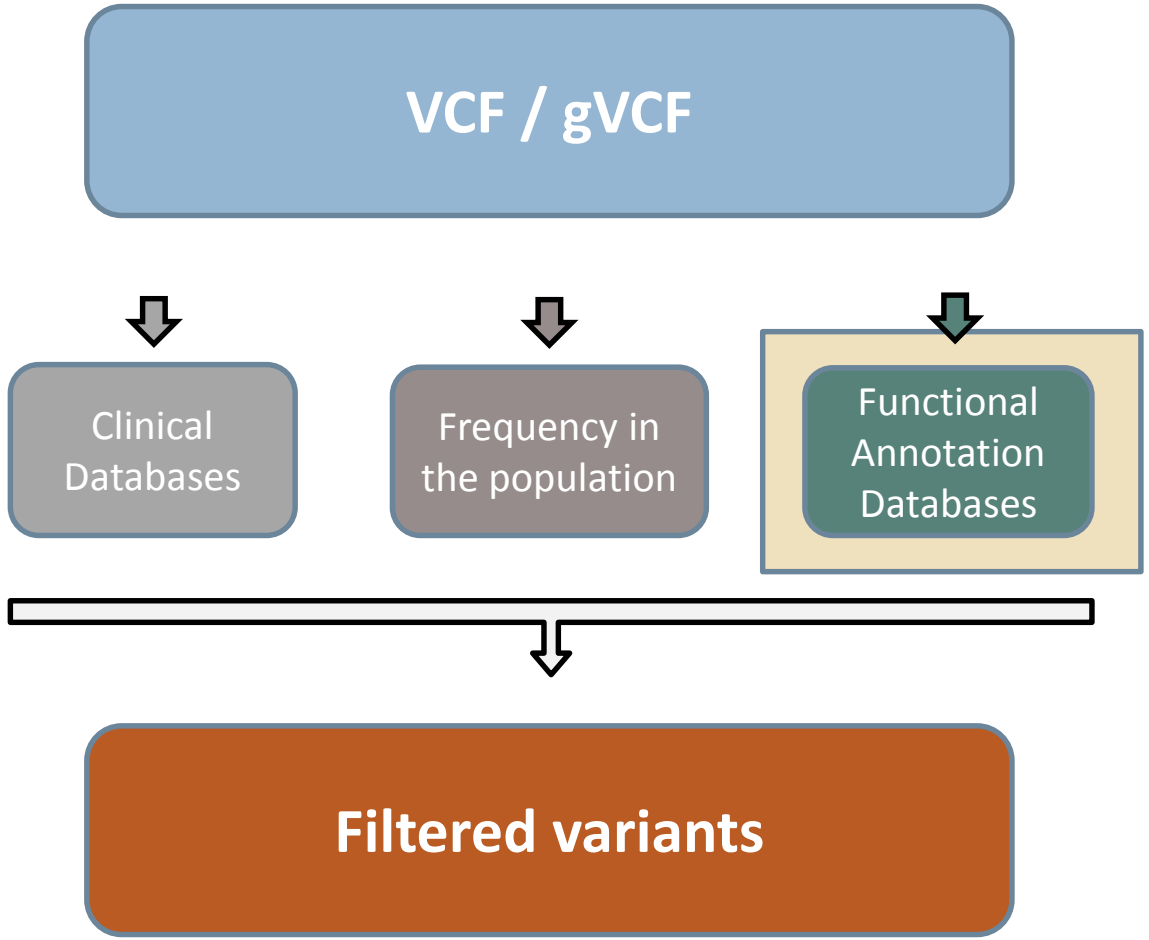


Example of VarSeq annotation output

Open the excel file: Lezioni2020_trio_1351S_chr6.xlsx

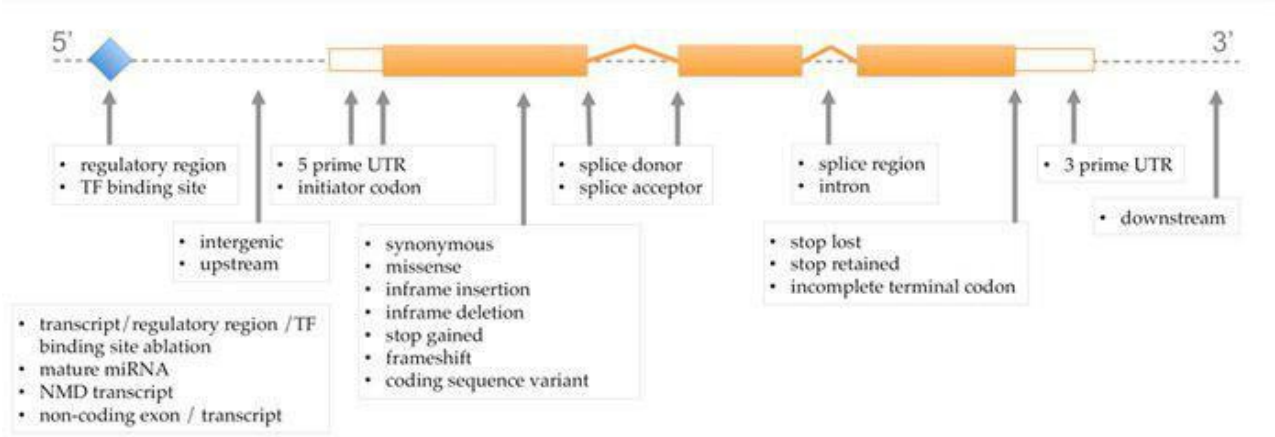
Variant Info		Flags	0/1 Genotypes (GT)			Read Depths (DP)			Allelic Depths (AD)			Zygosity			Alt Allele Freq			lendel Err	ound Het Variants for	ariant Typ	bSNP Com	dbSNP Common 151, NCBI						
Chr:Pos	Ref/Alt	Rare Varia	Proband (	Mother (1	Father (13	Proband (	Mother (1	Father (13	Proband (	Mother (1	Father (13	Proband (	Mother (1	Father (13	Proband (	Mother (1	Father (13	Proband (	Compoun	Inherited	Variant Ty	In dbSNP(	Ref/Alt	Identifier	dbSNPBui	Flags	Variation	1kG Varia	Ge
6:147404	C/T	False	0/1	0/1	0/1	76	85	75	45,31	61,24	27,48	Heterozyg	Heterozyg	Heterozyg	0.407895	0.282353	0.64	Transmitt	False	NA	SNP	False							
6:147750	C/A	False	0/1	0/1	0/1	10	16	9	4,6	8,8	1,8	Heterozyg	Heterozyg	Heterozyg	0.6	0.5	0.888889	Transmitt	False	NA	SNP	False							
6:396923	C/T	False	0/1	0/1	0/1	6	12	7	3,3	7,5	3,4	Heterozyg	Heterozyg	Heterozyg	0.5	0.416667	0.571429	Transmitt	False	NA	SNP	False							IR
6:489202	C/T	False	0/0	0/0	0/1	19	19	14	19,0	19,0	7,7	Reference	Reference	Heterozyg	0	0	0.5	Untransm	False	NA	SNP	True	C/T	rs7895186	132	In Intron,I	SNV	0.987,0.01	EX
6:908983	A/T	False	1/1	1/1	0/1	19	15	25	0,19	0,15	10,15	Homozygc	Homozygc	Heterozyg	1	1	0.6	Transmitt	False	NA	SNP	True	A/T	rs2756306	100	RS orienta	SNV	0.7863,0.213	
6:909005	A/G	False	1/1	1/1	0/1	29	25	32	0,29	1,24	14,18	Homozygc	Homozygc	Heterozyg	1	0.96	0.5625	Transmitt	False	NA	SNP	True	A/G	rs2756305	100	RS orienta	SNV	0.7861,0.213	
6:909193	C/G	False	0/1	0/1	0/1	175	187	242	86,89	89,97	111,131	Heterozyg	Heterozyg	Heterozyg	0.508571	0.521505	0.541322	Transmitt	False	NA	SNP	False							
6:909214	A/T	False	0/1	0/1	0/1	181	205	248	93,88	110,95	97,150	Heterozyg	Heterozyg	Heterozyg	0.486188	0.463415	0.607287	Transmitt	False	NA	SNP	False							
6:909475	T/C	False	0/0	0/1	0/0	48	203	116	48,0	135,68	112,4	Reference	Heterozyg	Reference	0	0.334975	0.0344828	Untransm	False	NA	SNP	False							
6:909483	T/C	False	1/1	0/1	0/0	40	196	105	1,39	94,102	81,24	Homozygc	Heterozyg	Reference	0.975	0.520408	0.228571	MIE	False	NA	SNP	False							
6:909490	G/C	False	1/1	0/0	0/0	38	202	101	0,38	170,31	79,20	Homozygc	Reference	Reference	1	0.154229	0.20202	MIE	False	NA	SNP	False							
6:910777	A/G	False	0/1	1/1	0/0	123	167	120	52,71	6,161	119,1	Heterozyg	Homozygc	Reference	0.577236	0.964072	0.0083333	Transmitt	False	NA	SNP	False							
6:910781	A/T	False	1/1	1/1	1/1	122	161	122	1,121	2,158	1,121	Homozygc	Homozygc	Homozygc	0.991803	0.9875	0.991803	Transmitt	False	NA	SNP	True	A/T	rs2756302	100	RS orienta	SNV	0.5587,0.441	
6:922566	G/T	False	1/1	0/1	0/1	18	4	12	0,18	1,3	8,4	Homozygc	Heterozyg	Heterozyg	1	0.75	0.333333	Transmitt	False	NA	SNP	True	G/T	rs2145914	96	RS orienta	SNV	0.3321,0.667	
6:922572	C/T	False	1/1	0/1	0/1	16	3	12	0,16	1,2	8,4	Homozygc	Heterozyg	Heterozyg	1	0.666667	0.333333	Transmitt	False	NA	SNP	True	C/T	rs2145913	96	RS orienta	SNV	0.3327,0.667	
6:1312020	C/T	False	0/1	0/0	0/1	155	137	123	80,75	137,0	64,59	Heterozyg	Reference	Heterozyg	0.483871	0	0.479675	Transmitt	False	NA	SNP	True	C/T	rs5483195	142	Has Subm	SNV	0.9952,0.004	
6:1624401	C/T	False	0/1	0/1	0/0	78	63	82	40,38	26,37	82,0	Heterozyg	Heterozyg	Reference	0.487179	0.587302	0	Transmitt	False	NA	SNP	True	C/T	rs1511614	134	Has Subm	SNV	0.9832,0.0	GM
6:2245587	C/A	False	0/0	0/0	0/1	11	15	17	11,0	15,0	7,10	Reference	Reference	Heterozyg	0	0	0.588235	Untransm	False	NA	SNP	False							GM
6:2870270	G/C	False	0/1	0/0	0/1	11	7	6	6,5	7,0	2,4	Heterozyg	Reference	Heterozyg	0.454545	0	0.666667	Transmitt	False	NA	SNP	False							
6:2870294	A/G	False	0/1	0/0	0/1	11	7	9	6,5	7,0	3,6	Heterozyg	Reference	Heterozyg	0.454545	0	0.666667	Transmitt	False	NA	SNP	True	A/G	rs1814447	135	Has Subm	SNV	0.9505,0.049	
6:3157295	G/A	False	0/0	0/1	0/0	86	93	87	86,0	47,46	85,2	Reference	Heterozyg	Reference	0	0.494624	0.0229885	Untransm	False	NA	SNP	False							TU
6:3273220	G/A	False	0/0	0/1	0/0	79	97	75	79,0	50,47	75,0	Reference	Heterozyg	Reference	0	0.484536	0	Untransm	False	NA	SNP	False							SL
6:3356072	C/G	False	0/1	1/1	1/1	7	7	7	1,6	0,7	0,7	Heterozyg	Homozygc	Homozygc	0.857143	1	1	MIE	False	NA	SNP	False							SL
6:3356078	A/G	False	0/1	0/1	0/1	7	7	8	3,3	3,4	6,2	Heterozyg	Heterozyg	Heterozyg	0.5	0.571429	0.25	Transmitt	False	NA	SNP	False							SL
6:3457117	C/T	False	0/1	0/1	0/1	11	6	12	8,3	4,2	9,3	Heterozyg	Heterozyg	Heterozyg	0.272727	0.333333	0.25	Transmitt	False	NA	SNP	False							
6:3723887	C/A	False	0/0	0/1	0/0	23	22	16	23,0	6,16	16,0	Reference	Heterozyg	Reference	0	0.727273	0	Untransm	False	NA	SNP	True	C/A	rs226963	79	RS orienta	SNV	0.9149,0.0	PX
6:3979240	C/T	False	0/1	0/0	0/1	23	20	36	13,10	20,0	21,15	Heterozyg	Reference	Heterozyg	0.434783	0	0.416667	Transmitt	False	NA	SNP	False							
6:4031690	A/G	False	0/1	0/0	0/1	111	65	109	47,64	65,0	60,49	Heterozyg	Reference	Heterozyg	0.576577	0	0.449541	Transmitt	False	NA	SNP	False							PR
6:4135723	T/G	False	0/1	0/1	0/1	14	19	20	8,6	10,9	10,10	Heterozyg	Heterozyg	Heterozyg	0.428571	0.473684	0.5	Transmitt	False	NA	SNP	False							
6:4943770	T/A	False	0/0	0/1	0/0	60	74	51	60,0	44,30	50,1	Reference	Heterozyg	Reference	0	0.405405	0.0196078	Untransm	False	NA	SNP	False							CD
6:4998602	G/A	False	0/0	0/0	0/1	36	25	39	36,0	25,0	25,14	Reference	Reference	Heterozyg	0	0	0.358974	Untransm	False	NA	SNP	True	G/A	rs3536567	126	Has Subm	SNV	0.9892,0.0	RP
6:4998696	A/G	False	0/0	0/1	0/0	71	41	71	71,0	17,24	71,0	Reference	Heterozyg	Reference	0	0.585366	0	Untransm	False	NA	SNP	True	A/G	rs1475758	134	Has Subm	SNV	0.9978,0.0	RP
6:5771091	C/T	False	0/0	0/0	0/1	8	11	13	8,0	11,0	5,8	Reference	Reference	Heterozyg	0	0	0.615385	Untransm	False	NA	SNP	True	C/T	rs7282192	130	Has Subm	SNV	0.9804,0.0	FA

Variant Annotation



Variants effect on genes

→ Variants are annotated using gene names and classified based on the effects that the variants cause on genes

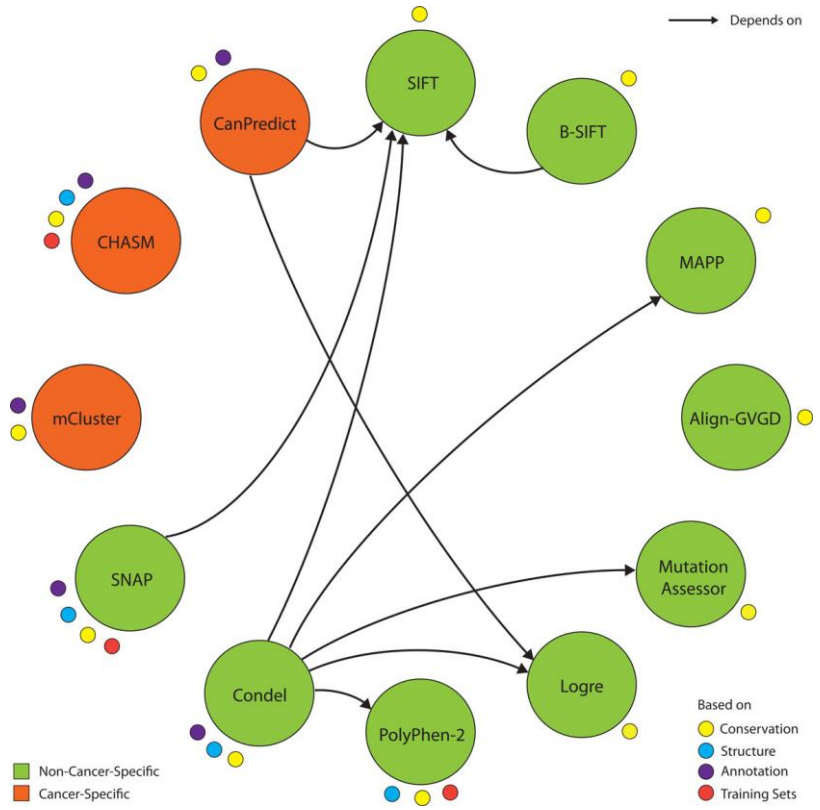


Loss of Function	The variant is likely to cause the transcript's product to lose function. The ontologies included in this category are: transcript ablation, exon loss variant, stop lost, stop gained, initiator codon variant, frameshift variant, splice acceptor variant, splice donor variant
Missense	The variant will cause at least one amino acid to change or cause a premature start codon in the UTR5. The ontologies included in this category are: disruptive inframe deletion, disruptive inframe insertion, inframe deletion, inframe insertion, 5 prime UTR premature start codon gain variant, missense variant
Other	The variant is likely to have a low or unknown effect on the transcript's functional product. These changes do not change the amino acid sequence of the protein. The ontologies included in this category are: synonymous variant, stop retained variant, splice region variant, 3 prime UTR variant, 5 prime UTR variant, intron variant, non-coding exon variant, intergenic variant

Variants effect prediction

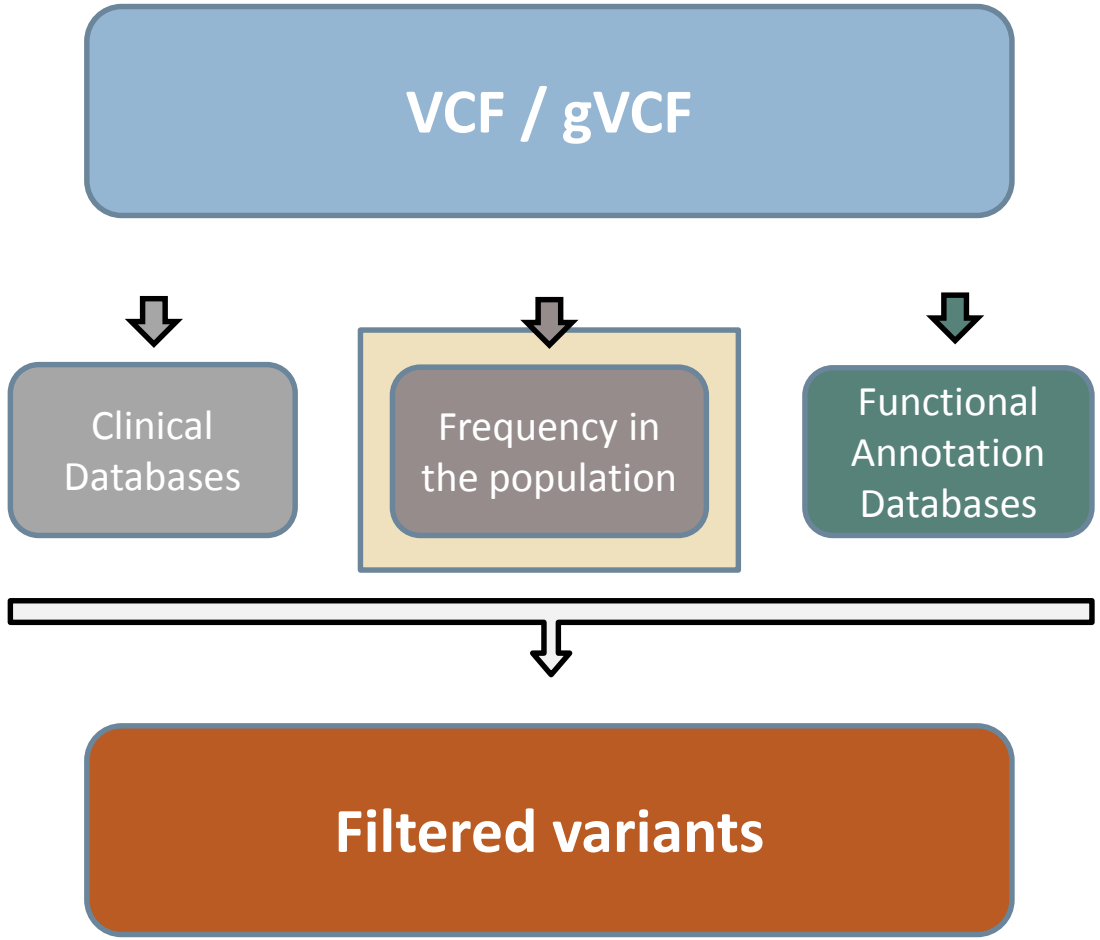
→ Several tools predict the effects of single nucleotide variations

What is the functional prediction scores? ⟶ SIFT, PolyPhen, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR scores, …



Gnad *et al. BMC Genomics* 2013 **14**:S7

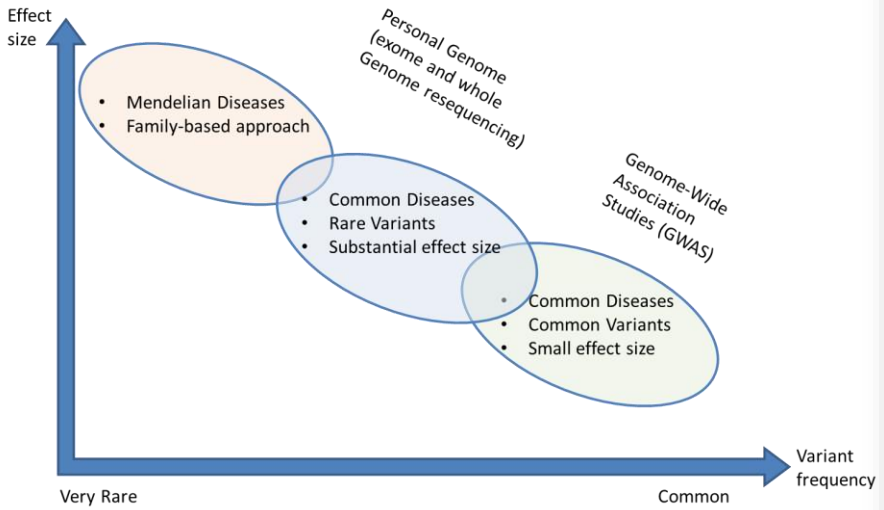
Variant Annotation

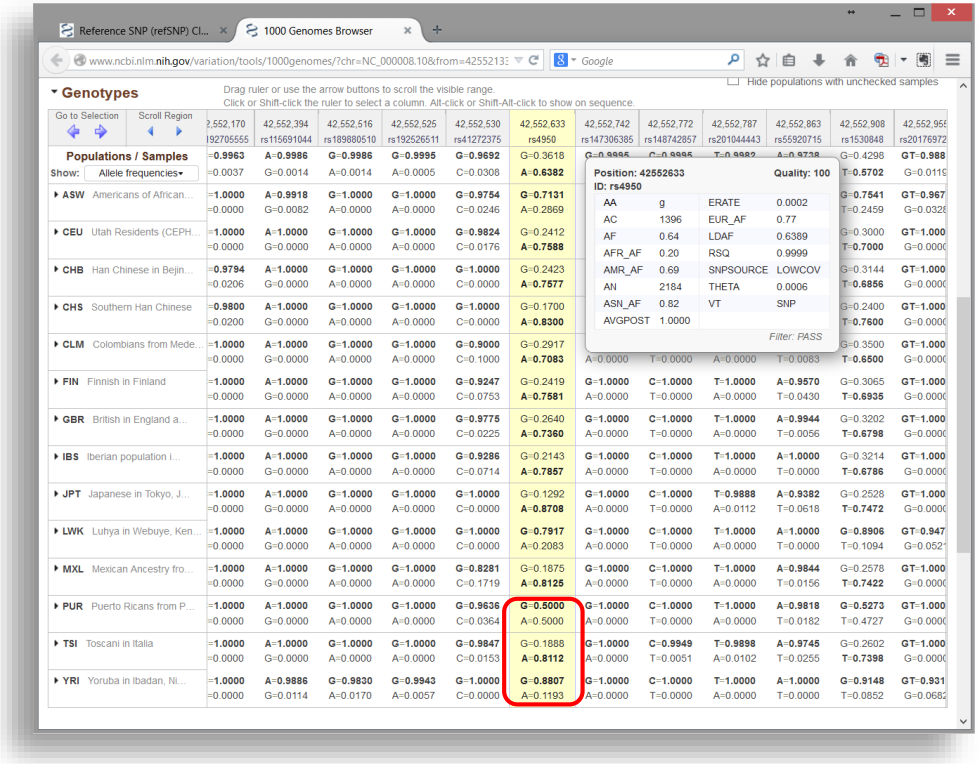


Variants frequency

→ Rare single nucleotide variants are much more likely to be population specific

What is the allele frequency? ⟶ 1000Genomes, ESP6500, ExAC,…

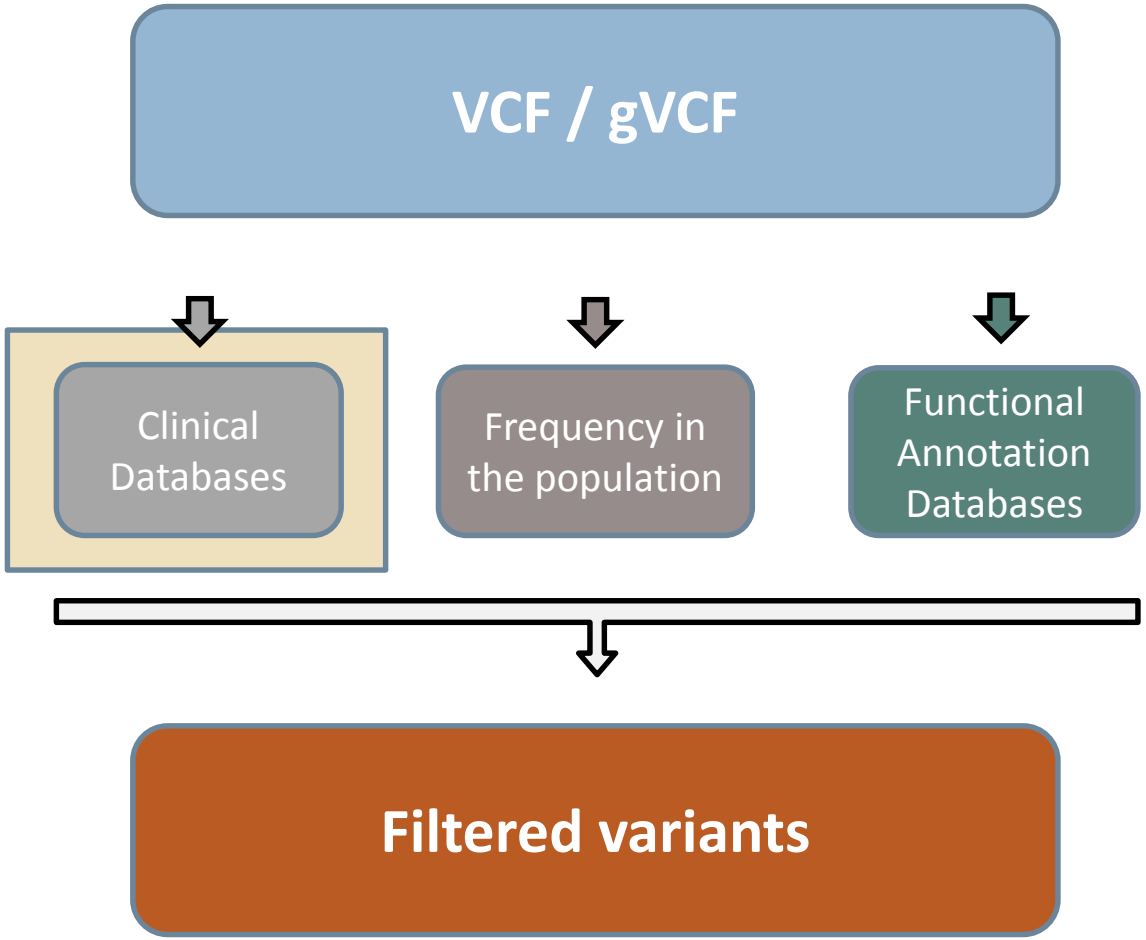




Example of variant frequency DB

NHLBI ESP6500SI-V2-SSA137 Exomes Variant Frequencies 0.0.30, GHI			ExAC Variant Frequencies 0.3 v2, BROAD			ExAC European AF	Exomes Variant Frequencies 2.0.1		gnomAD Exomes European AF	nomAD Genomes Variant Frequencies 2.0.1 v2, BROAI		gnomAD Genomes Europea
Ref/Alt	All AAF	European American AAF	Ref/Alt	Alt Alle	Adjusted Allele Frequency (AF_Adj)	ExAC European AF	Ref/Alt	Alt Allele Freq (AF)	gnomAD Exomes European AF	Ref/Alt	Alt Allele Freq (AF)	gnomAD Genomes European
										C/T	0.00421687	0.00193548
										C/A	0.0433898	0.0113636
										G/A	0.495425	0.425206
										T/A	0.499419	0.499576
										A/G	0.983638	0.990883
										G/A	0.445146	0.460159
										C/T	0.20497	0.23763
										T/C	0.995626	0.998647
										T/C	0.976526	0.983463
										G/A	0.99557	0.99923
										A/G	0.357182	0.417695
A/T	0.998001	0.999884	A/T	0.999	0.999348	0.999955	A/T	0.99948	0.999982	A/T	0.998032	1
										T/C	0.480021	0.476034
										T/C	0.480098	0.47604
										C/T	0.491512	0.478901
										C/T	0.364637	0.428104
										G/T	0.997193	0.999933
										A/G	0.479987	0.476053
A/G	0.435473	0.423908								A/G	0.480061	0.476171
A/G	0.997078	0.999884	A/G	0.999	0.998837	0.99985	A/G	0.999086	0.999901	A/G	0.99758	0.999933
			G/A	0.474	0.488234	0.486466	G/A	0.481123	0.480105	G/A	0.479977	0.476046
										C/T	0.468758	0.475875
										T/C	0.102443	0.0756101
										C/T	0.0331695	0.0414102
										T/A	0.687136	0.873232
										A/G	0.717898	0.785695
										A/G	0.717898	0.785666
										T/C	0.71142	0.788128
										T/A	0.597832	0.756943
										C/T	0.0240524	0.0371901
G/C	0.36091	0.484302	G/C	0.43	0.430548	0.459771	G/C	0.443986	0.469396	G/C	0.398413	0.489514
										T/C	0.471519	0.601764
										G/A	0.452138	0.42964

Variant Annotation



Clinical Significance

→ Once filtered, search for the associations with the disease of the case of study

- HGMD
- ClinVar
- OMIM
- …

→ Aim is to find the list of genes correlated with the disease and reported on medical and curated databases

HGMD

The Human Gene Mutation Database (**HGMD®**) represents an attempt to collate known (published) gene lesions responsible for human inherited disease.

It includes DNA sequence variants that are either disease-associated and of likely functional significance, or of clear functional significance even though no associated clinical phenotype may have been identified to date.

Effects of mutation:

- **DM:** disease-causing (pathological) mutation
- **DM?:** likely disease-causing (likely pathological) mutation
- **DP:** disease-associated polymorphism
- **DFP:** disease-associated polymorphism with additional supporting functional evidence

http://www.hgmd.cf.ac.uk/ac/index.php

HGMD

hgmd_pro_2018

Citatio	Ref/Alt	Identifier	Refere	Alterna	Quality	Filter	EFFECT	MUT	GENE	STRAN	DNA	PROT	In dbSNP? (DB)	PHEN	RANKS
,11423500	C/G	CM960827	C	G			DM	ALT	HFE	+	NM_00041	NP_00040	rs1799945	"Haemochromatosis"	0.1
	C/T	CM170949	C	T			DM?	ALT	HSPA1L	-	NM_00552	NP_00551	rs2075800	"Inflammatory_bowel_disease"	0.1
3038528,1	G/C	CM994664	G	C			DM?	ALT	CYP21A2	+	NM_0005(	NP_00049	rs6472	"Adrenal_hyperplasia"	0.1
,	T/G	CS114776	T	G			DM?	ALT	MUT	-	NM_000255.3:c.1676		rs9381786	"Methylmalonic_aciduria"	
,,,,	C/T	CM102732	C	T			DM?	ALT	EYS	-	NM_00114	NP_00113	rs111991705	"Retinitis_pigmentosa"	0.12
,,,	G/A	CM1719220	G	A			DM	ALT	MYO6	+	NM_00499	NP_004990.3:p.D1223N		"Hearing_loss_non-syndromic"	0.13
4	C/G	CR181226	C	G			DM?	ALT	ELOVL4	-			rs62407622	"Stargardt_disease"	
,,,,	G/A	CM030232	G	A			DM?	ALT	LAMA2	+	NM_00042	NP_00041	rs117781224	"Laminin_alpha_2_chain_deficiency_partial"	0.1
	G/A	CM1616147	G	A			DM?	REF	MAP3K4	+	NM_00592	NP_00591	rs4559074	"Premature_ovarian_failure"	0.11
,	G/A	CM056983	G	A			DM?	ALT	PRKN	-	NM_00456	NP_00455	rs55830907	"Parkinson_disease_early-onset"	0.14

The Human Gene Mutation Database

at the Institute of Medical Genetics in Cardiff

Home Search help Statistics New genes What is new Background Publications Contact Register Login LSDBs Other links Edit details Logout

Gene symbol Go! Symbol: Missense/nonsense Go!

Gene Symbol	Chromosomal location	Gene name	cDNA sequence	Extended cDNA	Mutation viewer
HFE (Aliases: available to subscribers)	6p21.3	homeostatic iron regulator (Aliases: available to subscribers)	NM_000410.3	Not available	Available to subscribers

Mutation type	Number of mutations	Mutation data by type (register or log in)
Missense/nonsense	34	Get mutations
Splicing	4	Get mutations
Regulatory	3	Get mutations
Small deletions	7	Get mutations
Small insertions	1	Get mutations
Small indels	0	No mutations
Gross deletions	2	Get mutations
Gross insertions/duplications	0	No mutations
Complex rearrangements	1	Get mutations
Repeat variations	0	No mutations
Get all mutations by type	Available to subscribers	
Public total (HGMD Professional 2019.4 total)	**52** (61)	

Disease/phenotype	Number of mutations	Mutation data by disease/phenotype
Haemochromatosis	36	Available to subscribers
Haemochromatosis ?	5	Available to subscribers
Iron overload	4	Available to subscribers

ClinVar

ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.

ClinVar differs from NCBI's variation archives, namely dbSNP and dbVar, which have the responsibility of maintaining information about the types and locations of all sequence variation.

Clinical significance value	Guidance for use in ClinVar SCV records
Benign	As recommended by ACMG for variants interpreted for Mendelian disorders.
Likely benign	As recommended by ACMG for variants interpreted for Mendelian disorders.
Uncertain significance	As recommended by ACMG for variants interpreted for Mendelian disorders.
Likely pathogenic	As recommended by ACMG for variants interpreted for Mendelian disorders.
Pathogenic	As recommended by ACMG for variants interpreted for Mendelian disorders.
drug response	A general term for a variant that affects a drug response, not a disease. We anticipate adding more specific drug response terms based on a recommendation by CPIC.
association	For variants identified in a GWAS study and further interpreted for their clinical significance.
risk factor	For variants that are interpreted not to cause a disorder but to increase the risk.
protective	For variants that decrease the risk of a disorder, including infections.
Affects	For variants that cause a non-disease phenotype, such as lactose intolerance.
conflicting data from submitters	Only for submissions from a consortium, where groups within the consortium have conflicting intepretations of a variant but provide a single submission to ClinVar.
other	If ClinVar does not have the appropriate term for your submission, we ask that you submit "other" as clinical significance and contact us to discuss if there are other terms we should add.
not provided	For submissions without an interpretation of clinical significance. The primary goal of ClinVar is to archive reports of clinical significance of variants. Therefore submissions with a clinical significance of "not provided" should be limited to: • "literature only" submissions that report a publication about the variant, without interpreting the clinical significance • "research" submissions that provide functional significance (e.g. undetectable protein level) but no interpretation of clinical significance • "clinical testing" submissions from clinics or physicians that provide additional information about individuals with the variant, such as observed phenotypes, but do not interpret the clinical significance

NM_002834.3(PTPN11):c.1510A>G (p.Met504Val) AND Noonan's syndrome

Clinical significance: Pathogenic (Last evaluated: Oct 30, 2013)
Review status: ☆ ☆ ☆ ☆ (0/4)

NM_000492.3(CFTR):c.2392C>T (p.Pro798Ser) AND Cystic fibrosis

Clinical significance: Uncertain significance (Last evaluated: Aug 18, 2011)
Review status: ★ ☆ ☆ ☆

NM_002834.3(PTPN11):c.1510A>G (p.Met504Val) AND Rasopathy

Clinical significance: Pathogenic (Last evaluated: Mar 5, 2015)
Review status: ★ ★ ☆ ☆

NM_000492.3(CFTR):c.2453delT (p.Leu818Trpfs) AND Cystic fibrosis

Clinical significance: Pathogenic (Last evaluated: Mar 28, 2013)
Review status: ★ ★ ★ ☆

NM_000492.3(CFTR):c.2052delA (p.Lys684Asnfs) AND Cystic fibrosis

Clinical significance: Pathogenic (Last evaluated: Mar 3, 2004)
Review status: ★ ★ ★ ★

https://www.ncbi.nlm.nih.gov/clinvar/

ClinVar

	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	
1	ClinVar 2019-11-01, NCBI															
2	Ref/Alt	Variant	Classification	Clinical Significance	Aggreg	Review	RSID	Gene N	Gene II	HGVS g	HGVS g	HGVS c	HGVS p	Per Cor	Conditions	Clinical Sig
1535	C/G	10	Pathogenic	Pathogenic, Other	Pathogeni	(2 Stars) C	rs1799945	HFE,LOC1(	3077,1087	NC_00000	NC_00000	NM_0004	NP_00040	RCV00000	Hemochromatosis type 1,Microvascular complications of diabetes 7,not provided,Hereditary hemochromatosis,Alzheimer's disease;Familial porphyria cutanea tarda;Hemochromatosis	Pathogenic
11582																
11583																
11584																
11585																

NM_000410.3(HFE):c.187C>G (p.His63Asp) AND Hemochromatosis type 1

Clinical significance: Pathogenic (Last evaluated: May 28, 2019)
Review status: ★ ★ ☆ ☆

Based on: 7 submissions [Details]
Record status: current
Accession: RCV000000026.18

Allele description [Variation Report for NM_000410.3(HFE):c.187C>G (p.His63Asp)]

NM_000410.3(HFE):c.187C>G (p.His63Asp)

Genes: LOC108783645:HFE antisense RNA [Gene]
HFE:homeostatic iron regulator [Gene - OMIM - HGNC]

Variant type: single nucleotide variant

Cytogenetic location: 6p22.2

Genomic location: Chr6: 26090951 (on Assembly GRCh38)
Chr6: 26091179 (on Assembly GRCh37)

Preferred name: NM_000410.3(HFE):c.187C>G (p.His63Asp)

HGVS: NC_000006.12:g.26090951C>G
NG_008720.2:g.8671C>G
NM_000410.3:c.187C>G

...more

Protein change: H40D; His63Asp

Links: Genetic Testing Registry (GTR): GTR000021464; Genetic Testing Registry (GTR): GTR000509340; UniProtKB: Q30201#VAR_004396; OMIM: 613609.0002; dbSNP: rs1799945

NCBI 1000 Genomes Browser: rs1799945

OMIM

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily.

→ Contain information on **all known mendelian disorders** and over 15,000 genes

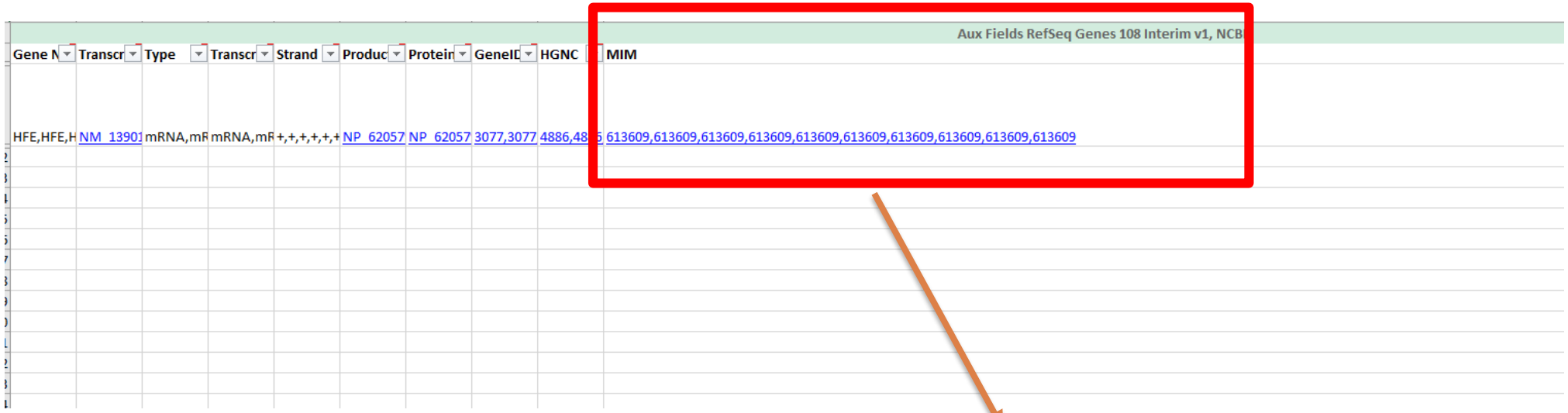
→ Focused on the relationship between phenotype and genotype

→ Based on the published peer-reviewed biomedical literature and curated by experts

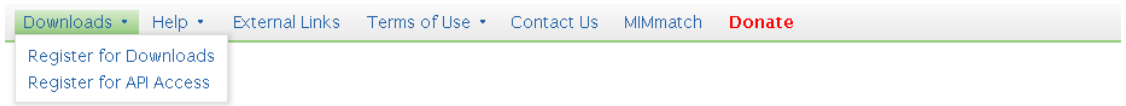
The gene-phenotype relationship is established considering the existence of multiple, **unrelated** individuals with pathogenic variants in the same gene (or the variants occur **de novo** in a statistically significant number of individuals)

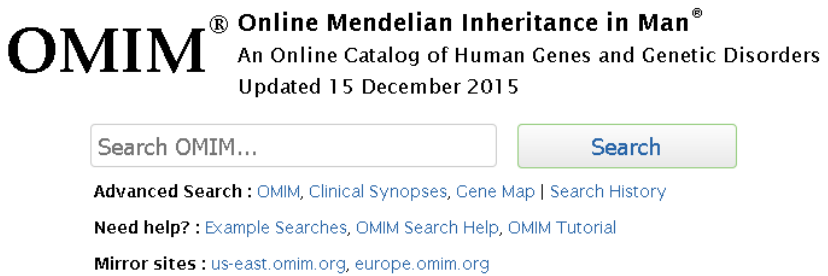
https://omim.org/

OMIM



Aux Fields RefSeq Genes 108 Interim v1, NCB									
Gene N	Transcr	Type	Transcr	Strand	Produc	Protein	GeneID	HGNC	MIM
HFE,HFE,H	NM_13901	mRNA,mR	mRNA,mR	+,+,+,+,+,+	NP_62057	NP_62057	3077,3077	4886,48	613609,613609,613609,613609,613609,613609,613609,613609,613609,613609





OMIM

* 613609

HOMEOSTATIC IRON REGULATOR; HFE

Alternative titles; symbols

HFE GENE
HLAH

HGNC Approved Gene Symbol: HFE

Cytogenetic location: 6p22.2 ***Genomic coordinates (GRCh38): 6:26,087,280-26,098,342*** (from NCBI)

Gene-Phenotype Relationships

View clinical synopses as a table

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
6p22.2	[Transferrin serum level QTL2]	614193		3
	{Alzheimer disease, susceptibility to}	104300	AD	3
	{Microvascular complications of diabetes 7}	612635		3
	{Porphyria cutanea tarda, susceptibility to}	176100	AD, AR	3
	[illegible]	176200	AD	3
	Hemochromatosis	235200	AR	3

PheneGene Graphics

External Links

- Genome
- DNA
- Protein
- Gene Info
- Clinical Resources
- Variation
  - 1000 Genome
  - ClinVar
  - ExAC
  - gnomAD
  - GWAS Catalog
  - GWAS Central
  - HGMD
  - HGVS
  - NHLBI EVS
  - PharmGKB
- Animal Models
- Cellular Pathways

TEXT

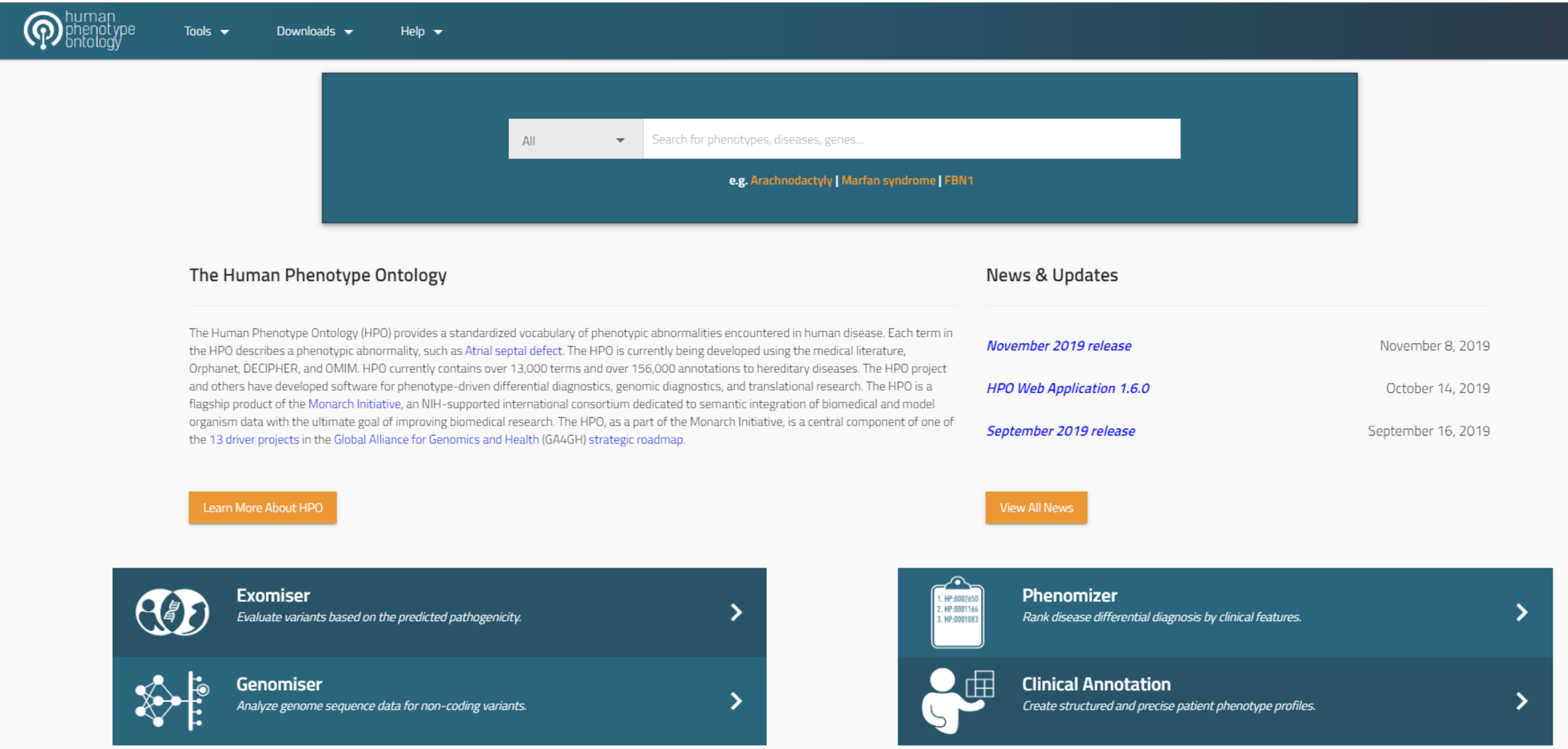
A number sign (#) is used with this entry because hemochromatosis type 1 (HFE1) is caused by homozygous or compound heterozygous mutation in the HFE gene (613609) on chromosome 6p22.

Description

Hereditary hemochromatosis is an autosomal recessive disorder of iron metabolism wherein the body accumulates excess iron (summary by Feder et al., 1996). Excess iron is deposited in a variety of organs leading to their failure, and resulting in serious illnesses including cirrhosis, hepatomas, diabetes, cardiomyopathy, arthritis, and hypogonadotropic hypogonadism. Severe effects of the disease usually do not appear until after decades of progressive iron loading. Removal of excess iron by therapeutic phlebotomy decreases morbidity and mortality if instituted early in the course of the disease. Classic hemochromatosis (HFE) is most often caused by mutation in a gene designated HFE on chromosome 6p21.3.

Adams and Barton (2007) reviewed the clinical features, pathophysiology, and management of hemochromatosis.

HPO



https://hpo.jax.org/app/

HPO

- The HPO provides computational resources that allow large-scale computational analysis of the human phenome.
- Each term in the HPO describes a clinical abnormality. These may be general terms or very specific ones.
- Each term is assigned to one of the five subontologies (see Table below).
- The terms have unique IDs.

Subontology	Description
Phenotypic abnormality	This is the main ontology of the HPO and contains descriptions of clinical abnormalities. The level 1 children of this class are formed by terms such as *Abnormality of the skeletal system* and *Abnormality of blood and blood-forming tissues*.
Mode of Inheritance	This relatively small ontology is intended to describe the mode of inheritance and contains terms such as *Autosomal dominant inheritance*.
Clinical modifier	This ontology contains classes that describe typical modifiers of clinical symptoms. For example the speed of progression, triggering factors, location or severity. It contains terms such as *Episodic*, *Bilateral*, or *Triggered by exertion*.
Clinical course	This sub-ontology describes the course a disease typically takes from its onset, progression in time, and eventual resolution or death of the affected individual.
Frequency	Frequency with patients that show a particular clinical feature. Examples are *Obligate*, *Frequent*, and *Occasional*. These terms are were defined in collaboration with Orphanet.

HPO

Hemochromatosis, Type 1 OMIM:235200

Hemochromatosis type 1 (classic) is the most common form of hereditary hemochromatosis (HH), a group of diseases characterized by excessive tissue iron deposition. Due to its incidence (1/200-1/1000), it is not considered as a rare disease, unlike the other subforms of the disease

Export Associations

HPO Associations | Gene Associations

Inheritance [1 annotation]

Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0000007	Autosomal recessive inheritance	-	-	

Metabolism/Laboratory abnormality [3 annotations]

Term Identifier	Term Name	Onset	Frequency	Source(s)
HP:0003452	Increased serum iron	-	-	
HP:0001952	Glucose intolerance	-	-	
HP:0003281	Increased serum ferritin	-	-	

IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK
Repeating Elements by RepeatMasker, UCSC						ity Region	gions and I	Complexit	ACMG li			Increased Serum Iron PhoRank
repNan	repClas	repFam	repStar	repEnd	repLeft	Matche	Type	Matche	In ACM	Gene Rank	Gene Score	Path
						False		False	False	1	3.93969	HFE / HP:0003452 (Increased serum iron)