
Variant calling practical session

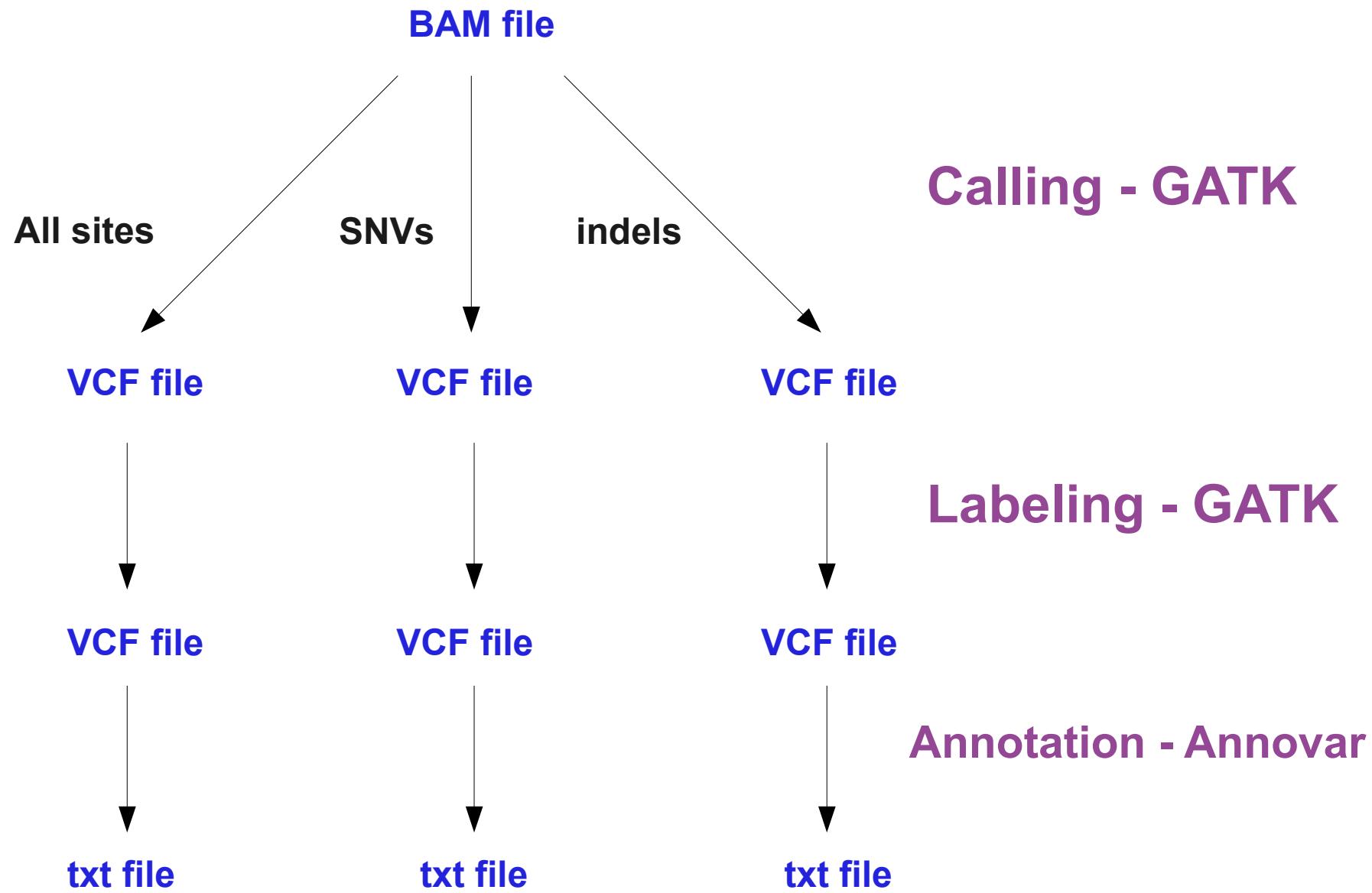
Jorge Jiménez
jjimeneza@cipf.es
BIER - CIBERER
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)



Index

1. Calling SNVs and indels
2. Labeling VCF files
3. Annotating VCF files
4. Visualization of variants

Scheme



Working directory

Working directory

```
cd
```

```
cd mda12
```

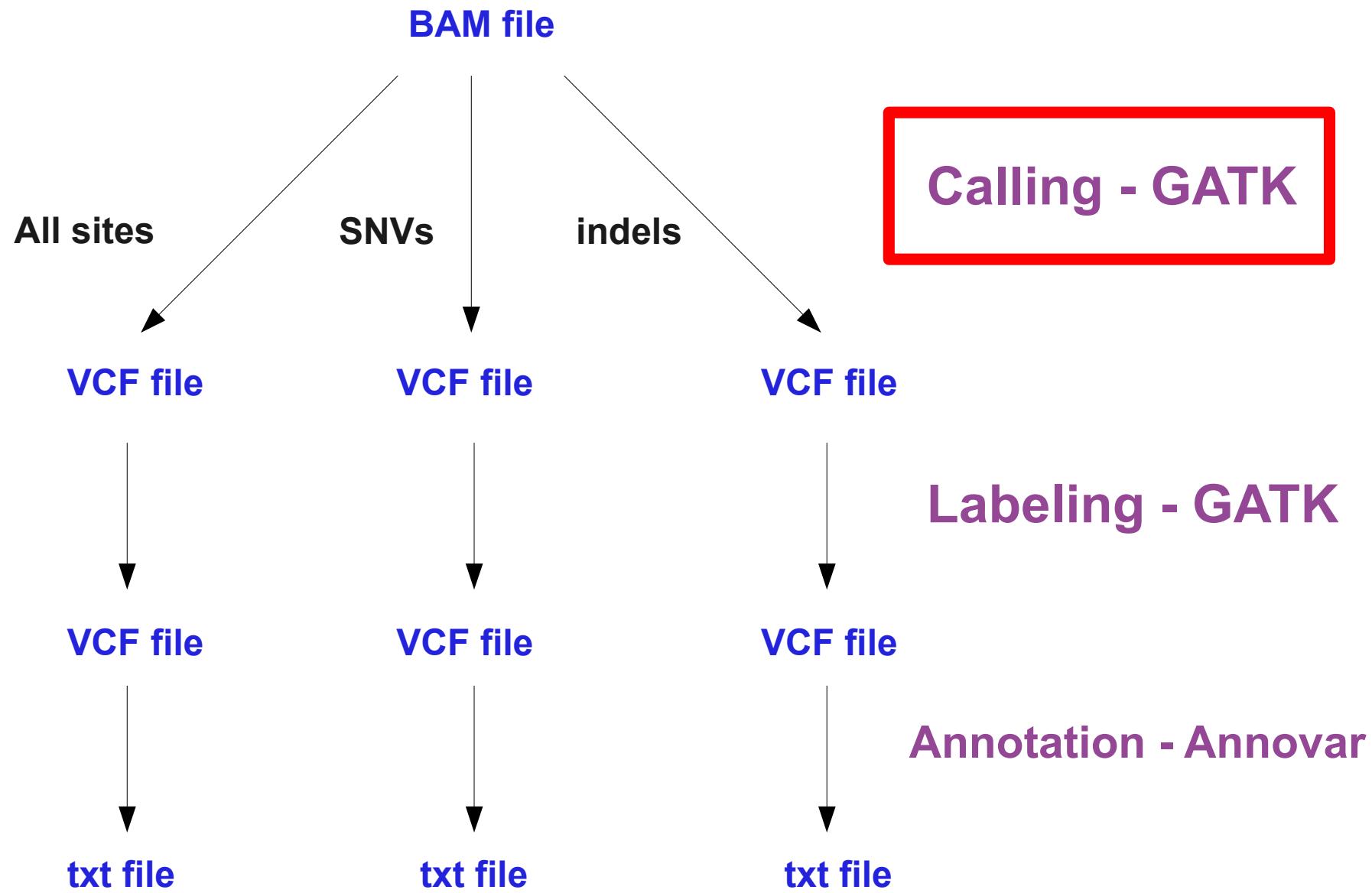
```
ls
```

```
ls mapping
```

```
cd calling
```

```
ls
```

Scheme



Variant calling - GATK

GATK (http://www.broadinstitute.org/gsa/wiki/index.php/Home_Page)

We need:

- reference
- bed file of regions of capture
- BAM mapping file
- bases to print
- output file

Run program and see options

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar
```

It needs the parameter to do the calling:

UnifiedGenotyper

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T UnifiedGenotyper
```

Reference and bed file

Checking the reference

```
head ~/mida12/resources/ref/human_g1k_v37.chr20.fasta
```

```
head -3000 ~/mida12/resources/ref/human_g1k_v37.chr20.fasta | tail
```

Checking the bed file

```
head ~/mida12/resources/ref/Exon_50mb_hg19_chr20.bed
```

```
20 68319 68439
20 76611 77091
20 123208 123358
20 125995 126237
20 126269 126389
20 138119 138269
20 139359 139719
20 168522 168762
20 170179 170299
20 207898 208018
```

SNV calling of all sites- GATK

SNV Calling of all sites

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T UnifiedGenotyper \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-L ~/mda12/resources/ref/Exon_50mb_hg19_chr20.bed \
-I ~/mda12/resources/mapping/test_final.bam \
-glm SNP \
-out_mode EMIT_ALL_SITES \
-o all_sites.vcf
```

Checking file

```
less all_sites.vcf
```

Counting lines

```
du -hs all_sites.vcf
wc -l all_sites.vcf
```

SNV calling only variants - GATK

Executing SNVs calling of variants

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T UnifiedGenotyper \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-L ~/mda12/resources/ref/Exon_50mb_hg19_chr20.bed \
-I ~/mda12/resources/mapping/test_final.bam \
-glm SNP \
-o snvs.vcf
```

Checking file

```
less snvs.vcf
```

Counting lines

```
du -hs snvs.vcf
wc -l snvs.vcf
```

indel calling - GATK

Executing indels calling of variants

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T UnifiedGenotyper \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-L ~/mda12/resources/ref/Exon_50mb_hg19_chr20.bed \
-I ~/mda12/resources/mapping/test_final.bam \
-glm INDEL \
-o indels.vcf
```

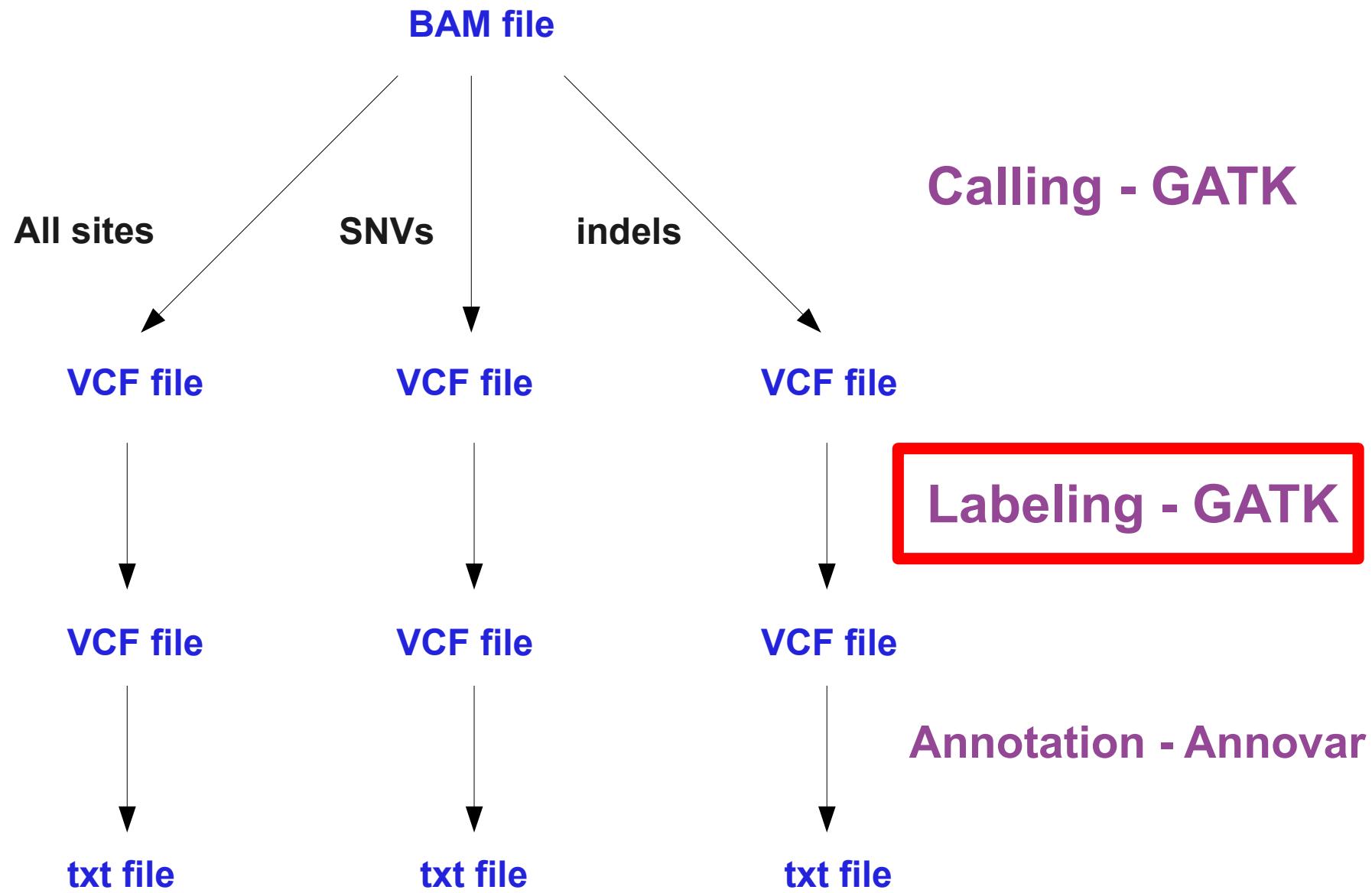
Checking file

```
less indels.vcf
```

Counting lines

```
du -hs indels.vcf
wc -l indels.vcf
```

Scheme



Labeling VCF files - GATK

Options for VariantFiltration:

- filter
- filter name
- reference
- input VCF file
- output VCF file

GATK parameter:

VariantFiltration

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T VariantFiltration
```

Labeling all sites VCF file - GATK

Labeling all sites VCF file

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T VariantFiltration \
-filter "QD < 2.0 || MQ < 40.0 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
-filterName "STD_FILTER" \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-V all_sites.vcf \
-o all_sites_labeled.vcf
```

Checking files

```
wc -l all_sites_labeled.vcf
wc -l all_sites.vcf

grep PASS all_sites_labeled.vcf | wc -l
```

Labeling SNVs VCF file - GATK

Labeling SNVs VCF file

```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T VariantFiltration \
-filter "QD < 2.0 || MQ < 40.0 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
-filterName "STD_FILTER" \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-V snvs.vcf \
-o snvs_labeled.vcf
```

Checking files

```
wc -l snvs_labeled.vcf
wc -l snvs.vcf

grep PASS snvs_labeled.vcf | wc -l
```

Labeling indels VCF file - GATK

Labeling indels VCF file

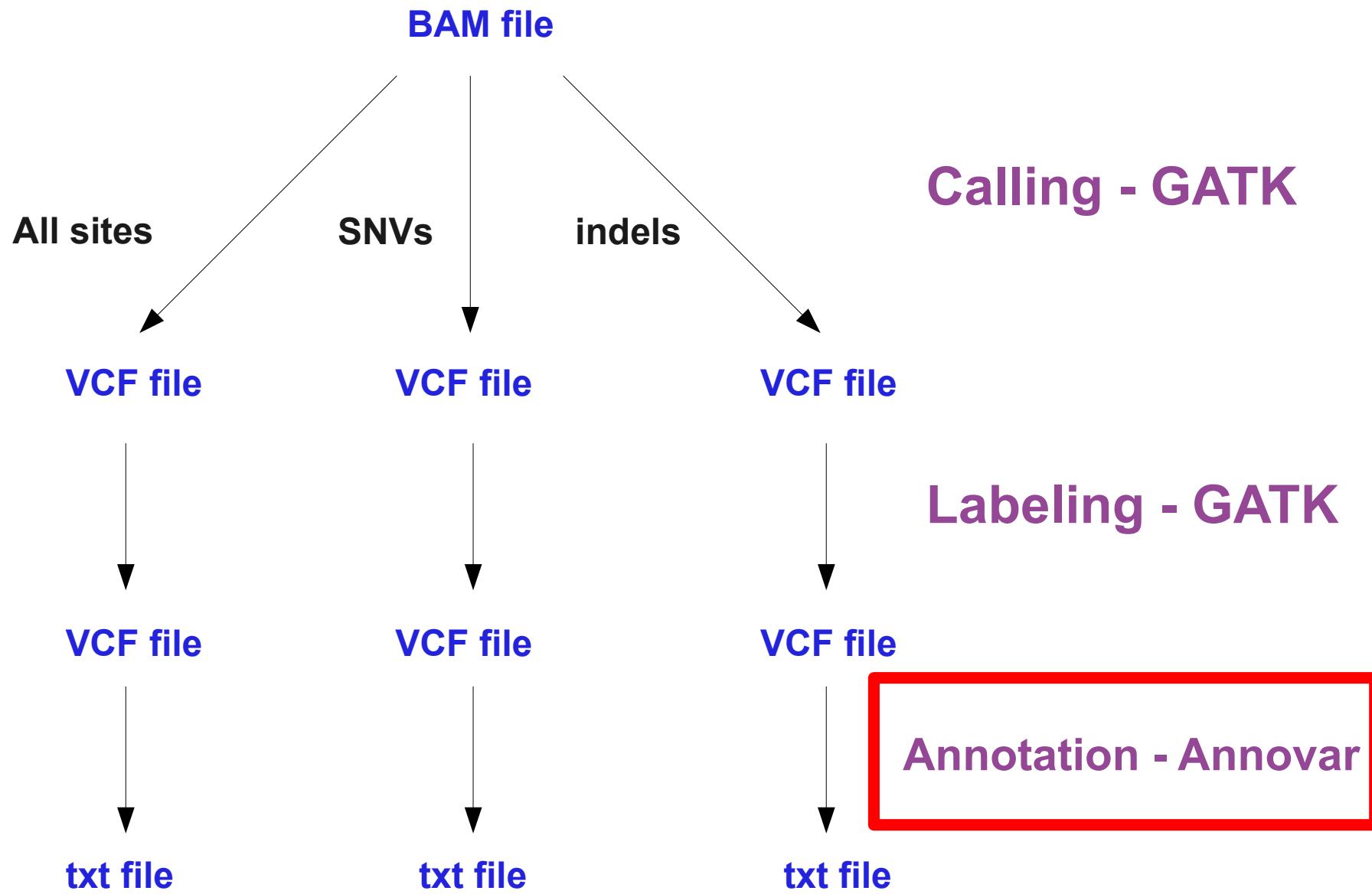
```
~/mda12/calling/software/GenomeAnalysisTK-1.4-15-gcd43f01/GenomeAnalysisTK.jar \
-T VariantFiltration \
-filter "QD < 2.0 || ReadPosRankSum < -20.0 || FS > 200.0" \
-filterName "STD_FILTER" \
-R ~/mda12/resources/ref/human_g1k_v37.chr20.fasta \
-V indels.vcf \
-o indels_labeled.vcf
```

Checking files

```
wc -l indels.vcf
wc -l indels_labeled.vcf

grep PASS indels_labeled.vcf | wc -l
```

Scheme



Annotation

Software:

Annovar: <http://www.openbioinformatics.org/annovar/>

Steps:

1. Convert VCF file to annovar format file
2. Annotate variants: SNVs and indels

Output:

- only exonic variants.
- all variants.

Annotation of SNVs – Annovar (1)

Converting SNV VCF file to annovar format file

```
~/mda12/calling/software/annovar/convert2annovar.pl \
-format vcf4 \
-filter PASS snvs_labeled.vcf > \
snvs_labeled.vcf.annovar
```

Checking files

```
wc -l snvs_labeled.vcf.annovar
grep PASS snvs_labeled.vcf | wc -l
```

```
head snvs_labeled.vcf.annovar
```

20	76962	76962	T	C	hom	8096.43	244	56.21	33.18
20	139362	139362	G	A	hom	2129.32	61	59.24	34.91
20	168728	168728	T	A	hom	6509173	58.52	37.62	
20	209932	209932	G	T	het	308.17	72	51.77	4.28
20	210061	210061	G	A	het	1856.36	128	57.24	14.50
20	238507	238507	A	C	het	135.67	24	57.00	5.65
20	239688	239688	G	A	het	353.20	43	54.98	8.21
20	239697	239697	G	C	het	452.21	46	54.22	9.83
20	256573	256573	A	G	het	1113.27	81	57.09	13.74
20	256727	256727	T	A	het	533.82	45	58.97	11.86

Annotation of SNVs – Annovar (2)

Annotating

```
~/mda12/calling/software/annoVar/annotate_variation.pl \  
--geneanno \  
--buildver hg19 \  
--dbtype gene \  
snvs_labeled.vcf.annoVar \  
~/mda12/calling/software/annoVar/humandb/
```

```
ls -latr
```

Output

```
head snvs_labeled.vcf.annoVar.exonic_variant_function
```

```
head snvs_labeled.vcf.annoVar.variant_function
```

```
head snvs_labeled.vcf.annoVar.log
```

Annotation of indels – Annovar (1)

Converting indels VCF file to annovar format file

```
~/mda12/calling/software/annovar/convert2annoar.pl \
-format vcf4 \
-filter PASS indels_labeled.vcf > \
indels_labeled.vcf.annoar
```

Checking files

```
wc -l indels_labeled.vcf.annoar
grep PASS indels_labeled.vcf | wc -l
```

```
head indels_labeled.vcf.annoar
```

20	126156	126159	CAAA-	het	1926.74	113	55.75	17.05	
20	126311	126312	CC -	het	866.91	102	52.73	8.50	
20	138179	138179	C -	hom	1281.43	35	59.30	36.61	
20	238436	238441	TGGTCT -	het	402.53	20	54.78	20.13	
20	746424	746432	TATCTGCC	-	het	482.13	19	47.89	25.38
20	2618033	2618033	-	AAAA	het	944.09	45	49.31	20.98
20	3740621	3740622	CA -	hom	763.01	26	53.64	29.35	
20	4880133	4880148	GCTCAATGCCTTCTGC	-	hom	9459.12	139	45.11	68.05
20	10622081	10622081	A -	het	1189.25	148	59.54	8.04	
20	11790885	11790885	- TT	hom	5752.58	110	60.45	52.30	

Annotation of indels – Annovar (2)

Annotating

```
~/mda12/calling/software/annoVar/annotate_variation.pl \  
--geneanno \  
--buildver hg19 \  
--dbtype gene \  
indels_labeled.vcf.annoVar \  
~/mda12/calling/software/annoVar/humandb/
```

```
ls -latr
```

Output

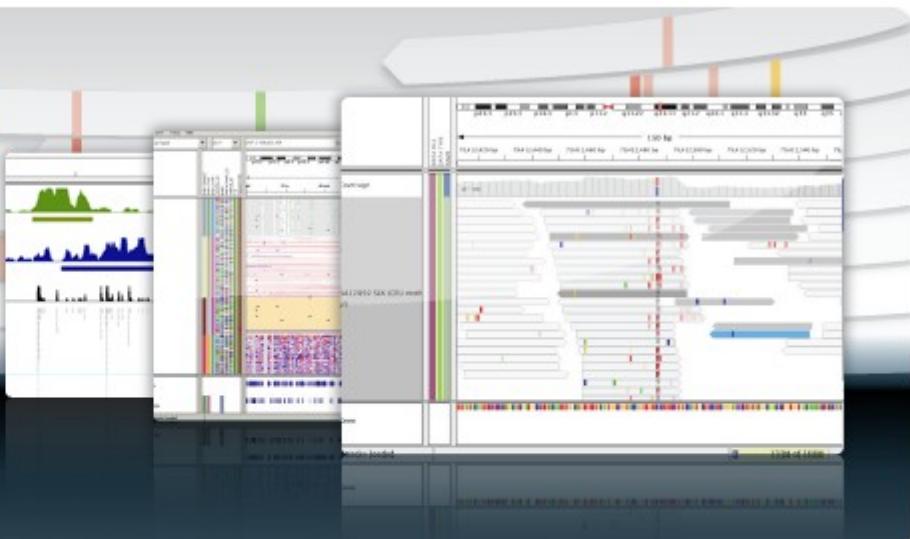
```
head indels_labeled.vcf.annoVar.exonic_variant_function
```

```
head indels_labeled.vcf.annoVar.variant_function
```

```
head indels_labeled.vcf.annoVar.log
```

Visualization - IGV

IGV: <http://www.broadinstitute.org/igv/>



The screenshot shows the IGV software interface. On the left, there is a navigation sidebar with the IGV logo, a search bar, and links to Home, Downloads, Documents, Hosted Genomes, FAQ, User Guide, File Formats, Release Notes, Credits, and Contact. Below this is a "Search website" input field and a "Broad Home" link. At the bottom, there is a copyright notice for the Broad Institute. The main area displays the IGV home page with the title "Integrative Genomics Viewer" and a large image of the software's user interface.

Home

Integrative Genomics Viewer

IGV is a genome browser for visualizing genomic data. It integrates multiple data sources and allows users to explore their own data alongside public datasets. IGV is designed to be easy to use and provides a powerful set of tools for analyzing genomic data.

IGV is available for download from the Broad Institute website. It is a Java application and requires Java 6 or later to run. IGV supports a wide range of data formats, including BED, GFF, VCF, and BigWig. It also supports RNA-seq, ChIP-seq, and other high-throughput sequencing data.

IGV is used by researchers around the world to analyze genomic data. It has been cited in numerous scientific publications, including the Nature Biotechnology paper "IGV: a tool for exploring genomic data" (Robinson et al., 2011).

What's New

July 3, 2012. Soybean (*Glycine max*) and Rat (rn5) genomes have been updated.

April 20, 2012. IGV 2.1 has been released. See the [release notes](#) for more details.

April 19, 2012. See our new [IGV paper](#) in *Briefings in Bioinformatics*.

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011), or

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov

Questions?

