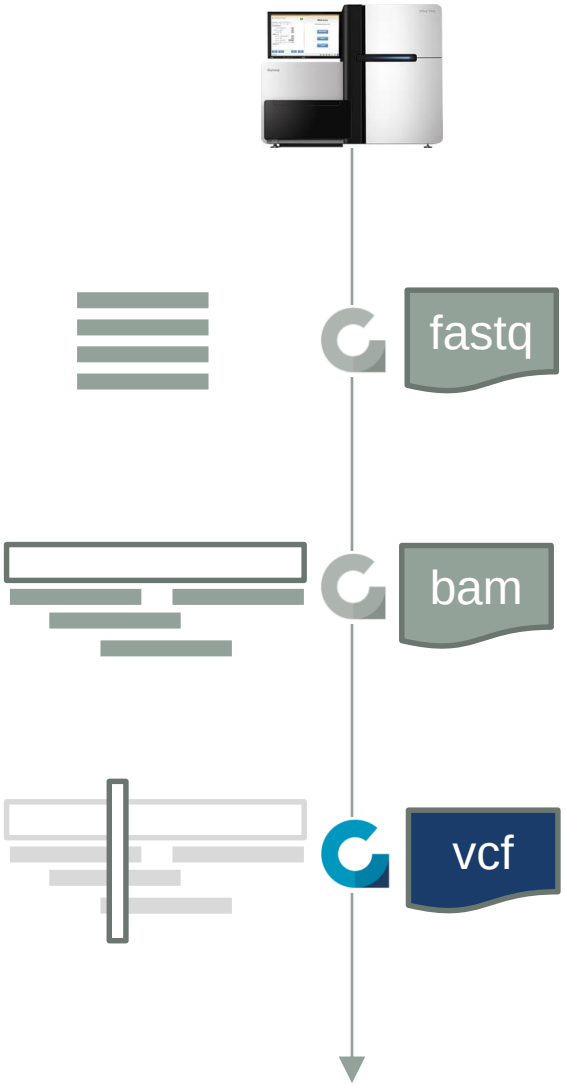
Variant calling:
GATK, quality scores, recalibration, VCF files

GENOMICS CORELEUVEN

Luc Dehaspe
Erika Souche

Overview



Sequencing

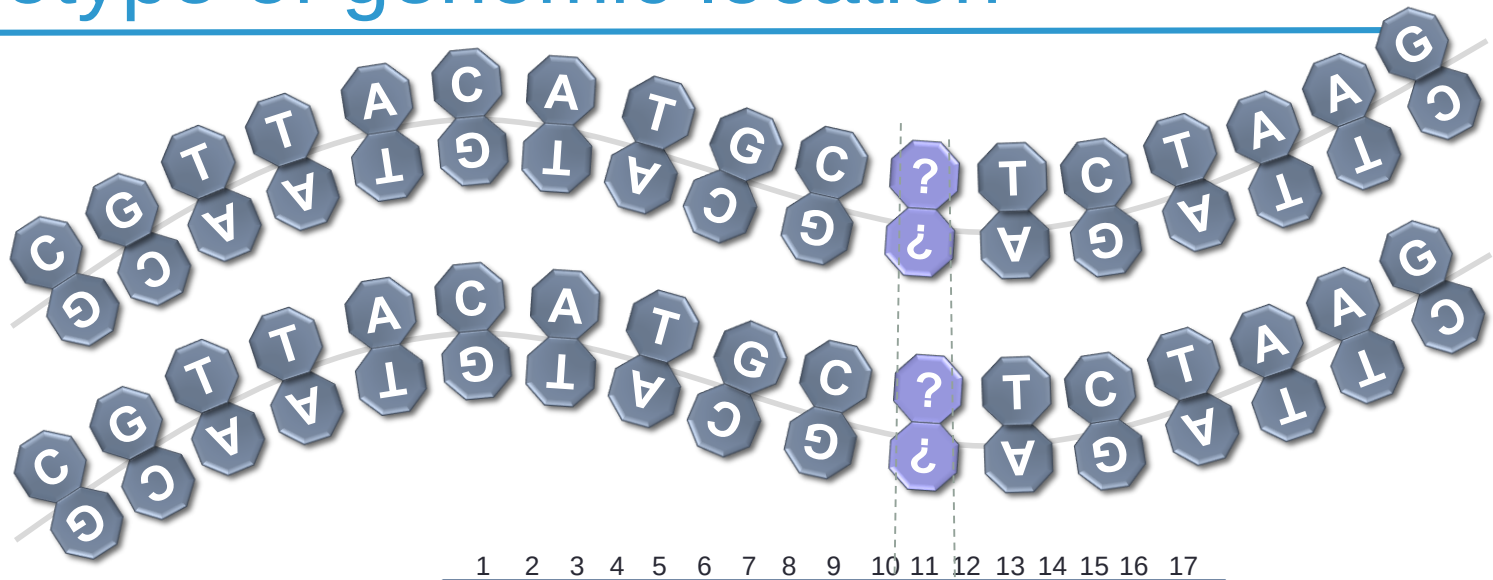
Sequence of nucleotides within a read?

Genomic location of a read?

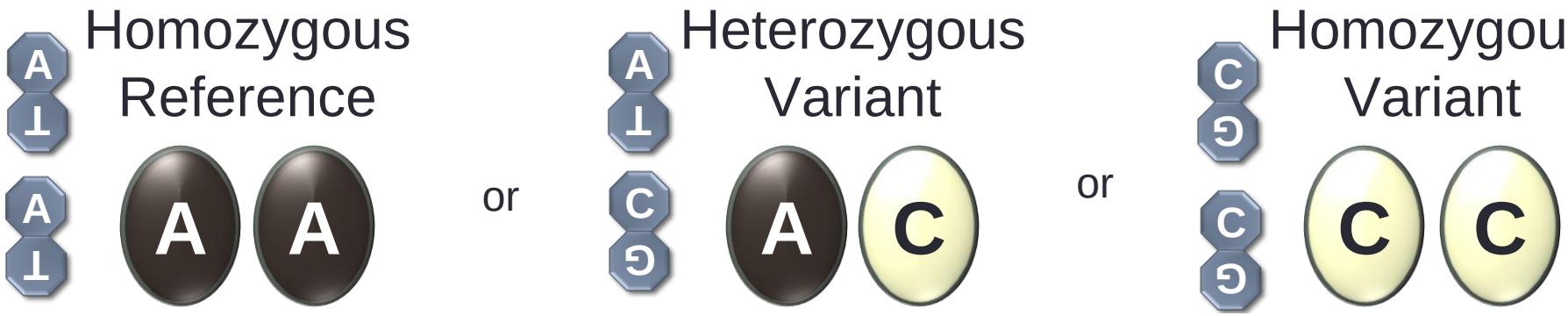
Genotype of a genomic location?

Genotype of genomic location

2 alleles



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Reference	chrN	C	G	T	T	A	C	A	T	G	C	A	T	C	T	A	A	G
												?						



https://bit.ly/2OwHr5r

Variant calling

- GATK HaplotypeCaller
  - Targeted sequencing
  - WGS

```
java -jar gatk.jar HaplotypeCaller -R genome.fa -I sample.bam --
dbsnp dbsnp_138.hg38.vcf --output sample.HaplotypeCaller.raw.vcf
```

https://software.broadinstitute.org/gatk/documentation/

Variant calling

- Freebayes – Bayesian haplotype-based genetic polymorphism discovery and genotyping
  - Amplicon sequencing
  - Targeted sequencing
  - WGS

```
freebayes -f genome.fa --genotype-qualities sample.bam >
sample.freebayes.vcf
```

https://github.com/ekg/freebayes

Variant Call Format (VCF)

- Header

```
##fileformat=VCFv4.2

##ALT=<ID=NON_REF,Description="Represents any possible alternative
allele at this location">

##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">

##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in
genotypes, for each ALT allele, in the same order as listed">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
```

Variant Call Format (VCF)

- Body

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878

20 61098 . C T 465.13 .
AC=1;AF=0.500;AN=2;BaseQRankSum=0.516;ClippingRankSum=0.00;DP=44;DP_Orig
=124;ExcessHet=3.0103;FS=0.000;MQ=59.48;MQRankSum=0.803;QD=10.57;ReadPos
RankSum=1.54;SOR=0.603 GT:AD:DP:GQ:PL 0/1:28,16:44:99:496,0,938

20 61138 . C CT 155.10 . AC=1;AF=0.500;AN=2;BaseQRankSum=-7.350e-
01;ClippingRankSum=0.00;DP=32;DP_Orig=131;ExcessHet=3.0103;FS=0.000;MQ=5
9.45;MQRankSum=0.790;QD=4.85;ReadPosRankSum=-3.970e-01;SOR=0.591
GT:AD:DP:GQ:PL 0/1:21,11:32:99:195,0,464

20 61795 . G T 2034.16 . AC=1;AF=0.500;AN=2;BaseQRankSum=-6.330e-
01;ClippingRankSum=0.00;DP=60;DP_Orig=164;ExcessHet=3.9794;FS=0.000;MQ=5
9.81;MQRankSum=0.00;QD=17.09;ReadPosRankSum=1.23;SOR=0.723
GT:AD:DP:GQ:PL 0/1:30,30:60:99:1003,0,1027

...
```

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value
Chromosome	20
Position	61098
dbSNP ID	.
Reference allele	C
Alternate allele	T
Quality	465.13
Filter	.
Info	AC=1;AF=0.500;AN=2;BaseQRankSum=0.516; ClippingRankSum=0.00;DP=44;DP_Orig=124; ExcessHet=3.0103;FS=0.000;MQ=59.48; MQRankSum=0.803;QD=10.57; ReadPosRankSum=1.54;SOR=0.603
Format	GT:AD:DP:GQ:PL
Sample	0/1:28,16:44:99:496,0,938

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value				
Format	GT:	AD:	DP:	GQ:	PL
Sample	0/1:	28,16:	44:	99:	496,0,938

Genotype
0/1 = heterozygous

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value				
Format	GT:	AD:	DP:	GQ:	PL
Sample	0/1:	28,16:	44:	99:	496,0,938

Genotype
0/1 = heterozygous

Allelic Depth
28 reference alleles
16 alternate alleles

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value				
Format	GT:	AD:	DP:	GQ:	PL
Sample	0/1:	28,16:	44:	99:	496,0,938

Genotype
0/1 = heterozygous

Allelic Depth
28 reference alleles
16 alternate alleles

Depth
44 reads at this position

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value				
Format	GT:	AD:	DP:	GQ:	PL
Sample	0/1:	28,16:	44:	99:	496,0,938

Genotype
0/1 = heterozygous

Allelic Depth
28 reference alleles
16 alternate alleles

Genotype quality
Smallest non-zero PL value
Maximum of 99

Depth
44 reads at this position

Variant Call Format (VCF)

- Looking at one position (GATK VCF)

Field	Value				
Format	GT:	AD:	DP:	GQ:	PL
Sample	0/1:	28,16:	44:	99:	496,0,938

Genotype
0/1 = heterozygous

Phred Likelihood
Likelihood 0/0, 0/1, 1/1

Allelic Depth
28 reference alleles
16 alternate alleles

Genotype quality
Smallest non-zero PL value
Maximum of 99

Depth
44 reads at this position

Checking variants in IGV

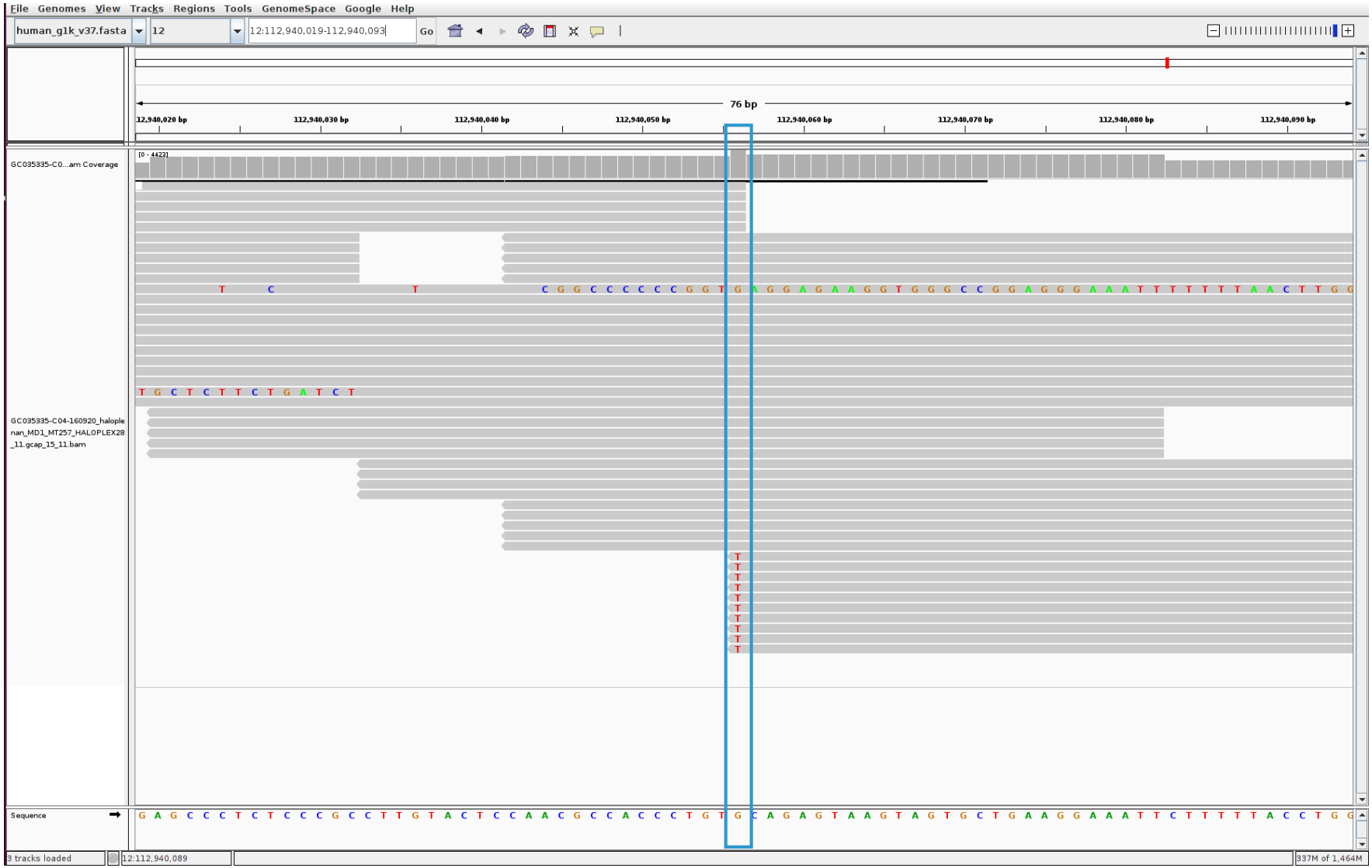
(1) variant detected with unbalance allelic depth

Amplicon based assay

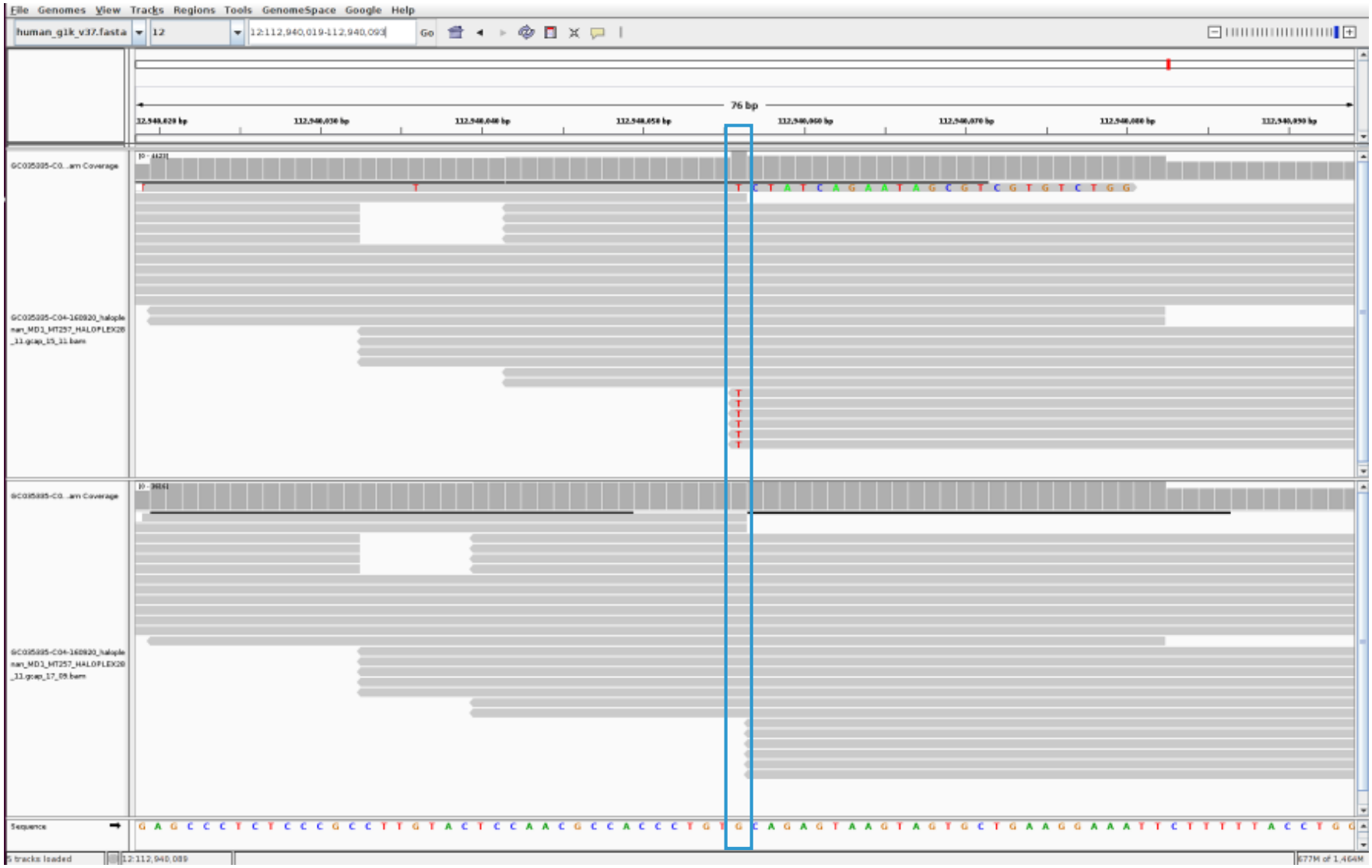
Heterozygous variant G/T

Format	Value
GT	0/1
AB	0.770
AD	**3791,1108**
DP	4899
GQ	99
PL	19373,0,137338

(1) unbalance AD



(1) unbalance AD



Checking variants in IGV

(2) variants on same allele ?

Targeted assay

Recessive disorder

Two heterozygous variants: GACT/G and G/A

Format	Variant 1	Variant 2
GT	0/1	0/1
AB	-	0.490
AD	50,39	42,44
DP	89	86
GQ	99	99
PL	1460,0,1947	1328,0,1128

(2) Variants on same allele ?

Checking variants in IGV

(3) real *de novo* variant ?

Targeted assay

Trio *i.e.* child & parents

Heterozygous deletion TA/T in child

Format	Value
GT	0/1
AD	43,51
DP	94
GQ	99
PL	1742,0,1448

(3) real *de novo* ?

Checking variants in IGV

(3) real *de novo* variant ?

Targeted assay

Trio *i.e.* child & parents

Heterozygous deletion TA/T in child

Format	Index	Mother	Father
GT	0/1	0/0	0/0
AD	43,51	103,0	81,0
DP	94	103	81
GQ	99	-	-
PL	1742,0,1448	-	-

Joint calling/genotyping

- Make one VCF file with many samples
- Benefits
  - Easy comparison (reference calls included)
  - Inheritance inference
  - More accurate genotypes

Joint calling/genotyping

```
java -jar gatk.jar HaplotypeCaller -R genome.fa -I sample.bam --
dbsnp dbsnp_138.hg38.vcf --output-mode EMIT_ALL_SITES --all-site-
pls --output sample.HaplotypeCaller.raw.g.vcf
```

https://software.broadinstitute.org/gatk/documentation/

genomeVCF (gVCF)

- Store reference and candidate variant information
  - Nucleotide level

```
#CHROM POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT MOTHER
14     29237220     .      C      <NON_REF>    .      .      END= 29237220
GT:DP:GQ:MIN_DP:PL 0/0:103:99:103:0,120,1800
14     29237221     .      T      <NON_REF>    .      .      END= 29237221
GT:DP:GQ:MIN_DP:PL 0/0:103:99:103:0,120,1800
14     29237222     .      A      <NON_REF>    .      .      END= 29237222
GT:DP:GQ:MIN_DP:PL 0/0:103:99:103:0,120,1800
14     29237223     .      C      <NON_REF>    .      .      END= 29237223
GT:DP:GQ:MIN_DP:PL 0/0:103:99:103:0,120,1800
...
```

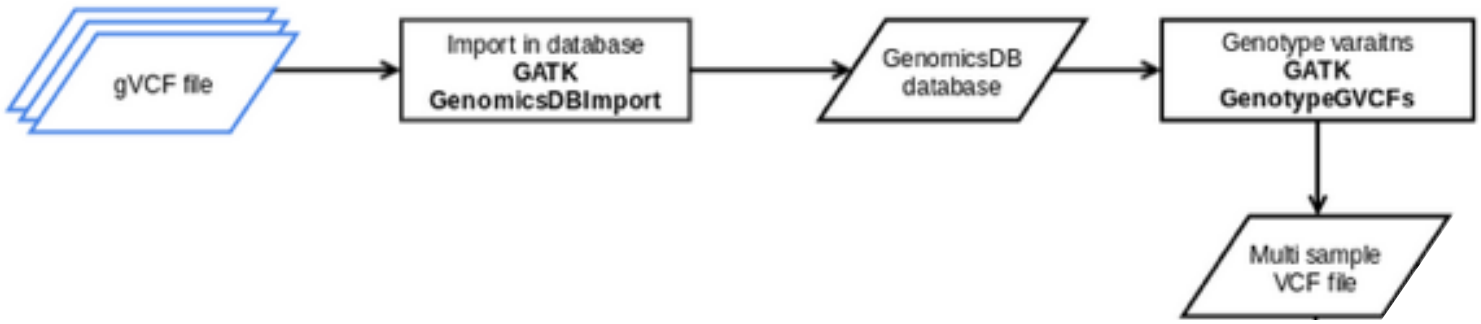
genomeVCF (gVCF)

- Store reference and candidate variant information
  - Nucleotide level
  - Region level

```
#CHROM POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT MOTHER
14     29237220     .      C      <NON_REF>    .      .      END= 29237223
GT:DP:GQ:MIN_DP:PL 0/0:103:99:103:0,120,1800
14     29237224     .      C      <NON_REF>    .      .      END= 29237242
GT:DP:GQ:MIN_DP:PL 0/0:47:88:103:0,88,1800
...
```

Joint calling/genotyping

```
java -jar gatk.jar HaplotypeCaller -R genome.fa -I sample.bam --dbsnp dbsnp_138.hg38.vcf --output-mode EMIT_ALL_SITES --all-site-pls --output sample.HaplotypeCaller.raw.g.vcf
```




```
java -jar gatk.jar GenomicsDBImport -V mother.g.vcf.gz -V father.g.vcf.gz -V son.g.vcf.gz --genomicsdb-workspace-path database
```

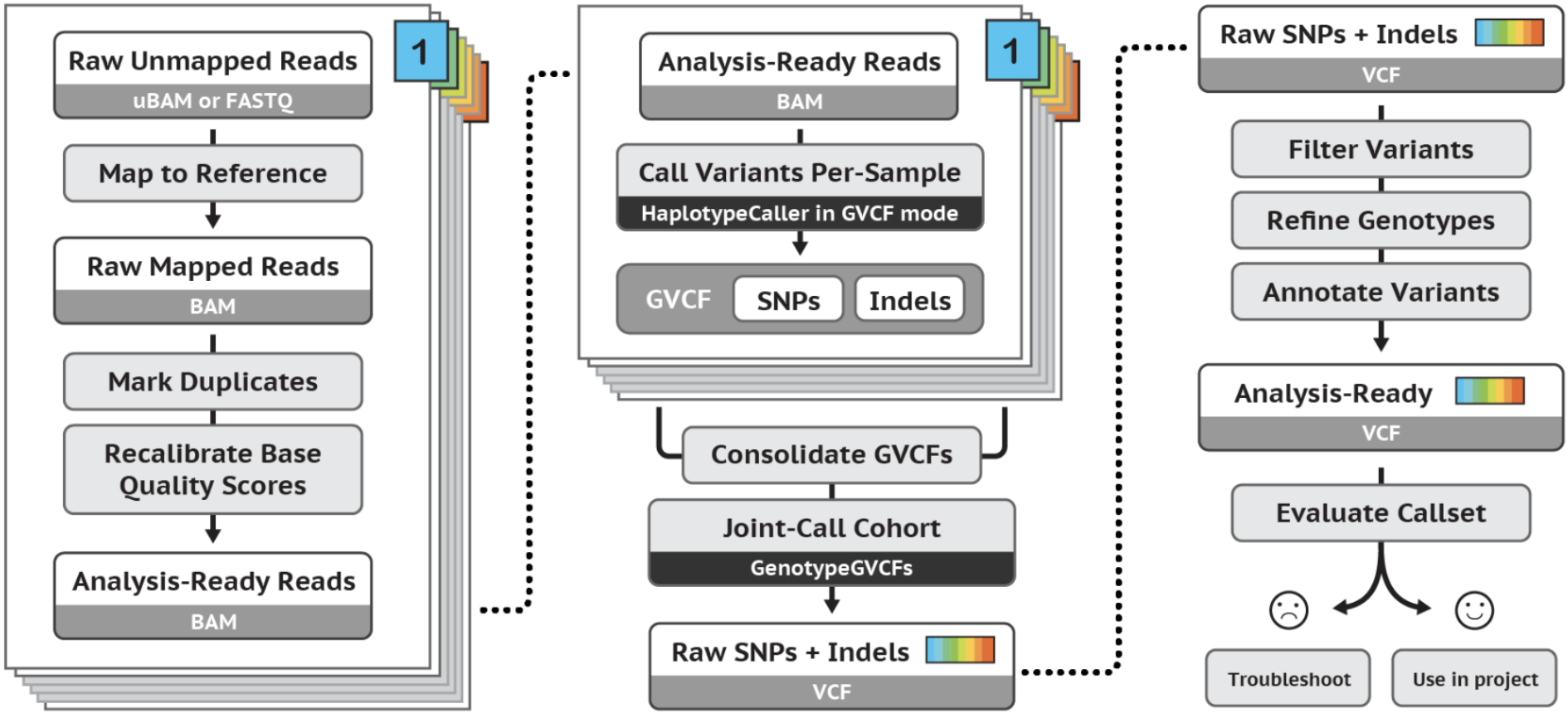
```
java -jar gatk.jar GenotypeGVCFs -R genome.fa -V gendb://database -O trio.vcf.gz
```

https://software.broadinstitute.org/gatk/documentation/

Multi-sample VCF

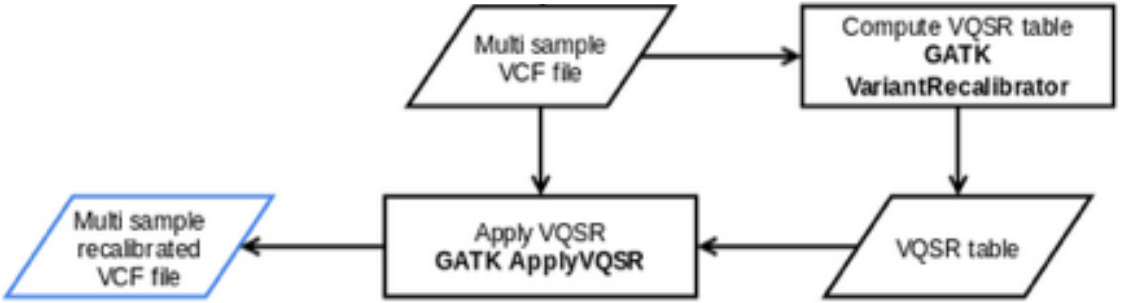
```
#CHROM POS     ID      REF     ALT     QUAL    FILTER INFO    FORMAT INDEX  FATHER
       MOTHER

14  25103662  .       T       C       1340.16       .
       AC=4;AF=0.667;AN=6;BaseQRankSum=2.86;ClippingRankSum=0.318;DP=54;
   FS=0.000;MLEAC=4;MLEAF=0.667;MQ=60.00;MQ0=0;MQRankSum=-6.350e-
   01;QD=24.82;ReadPosRankSum=0.421;SOR=0.811       GT:AB:AD:DP:GQ:PL:TP
       0|1:0.500:7,7:14:99:238,0,217:711|1:.:0,24:24:71:903,71,0:71
       0|1:0.560:9,7:16:99:231,0,280:71

14  29237221  .       TA      T       1711.13       .
       AC=1;AF=0.167;AN=6;BaseQRankSum=-3.000e-
   02;ClippingRankSum=0.448;DP=279;FS=8.696;MLEAC=1;MLEAF=0.167;MQ=60.11
   ;MQ0=0;MQRankSum=0.918;QD=18.20;ReadPosRankSum=-7.290e-01;SOR=1.911
       GT:AD:DP:GQ:PL:TP   0/1:43,51:94:99:1742,0,1448:36
       0/0:81,0:81:99:0,120,1800:36     0/0:103,0:103:99:0,120,1800:36

14      31344406      .       T       G       3458.90       .
       AC=6;AF=1.00;AN=6;DP=92;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=60.00;MQ0=
0;QD=32.40;SOR=5.762       GT:AD:DP:GQ:PL:TP
       1|1:0,34:34:99:1317,102,0:801|1:0,31:31:92:1163,92,0:80
       1|1:0,27:27:81:1005,81,0:80
```

GATK best practices



Variant recalibration

- Variant Quality Score Recalibration (VQSR)
  - Compute new quality score
  - Distinguish good from bad variants
  - Based on validated variant resources



Variant recalibration

```
java -jar gatk.jar VariantRecalibrator -R genome.fasta -V
multi_sample.vcf.gz --resource
hapmap,known=false,training=true,truth=true,prior=15.0:hapmap.s
ites.vcf.gz --resource
omni,known=false,training=true,truth=false,prior=12.0:1000G_omn
i2.5.sites.vcf.gz --resource
1000G,known=false,training=true,truth=false,prior=10.0:1000G_ph
ase1.snps.high_confidence.vcf.gz --resource
dbsnp,known=true,training=false,truth=false,prior=2.0:Homo_sapi
ens.dbsnp138.vcf.gz -an QD -an MQ -an MQRankSum -an
ReadPosRankSum -an FS -an SOR -mode SNP -O output.recal --
tranches-file vqsr.tranches --rscript-file vqsr.plots.R
```

```
java -jar gatk.jar ApplyVQSR -R Homo_sapiens_assembly38.fasta -
V input.vcf.gz -O output.vcf.gz --truth-sensitivity-filter-
level 99.0 --tranches-file vqsr.tranches --recal-file
output.recal -mode SNP
```

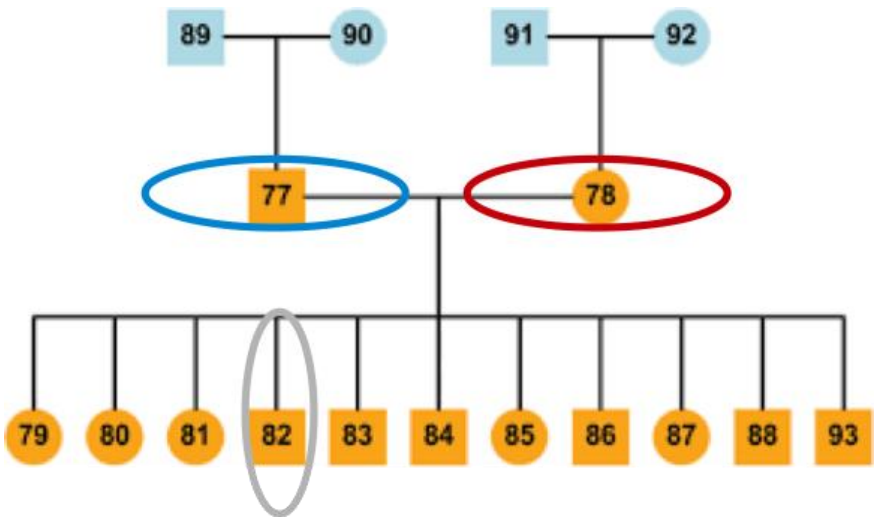
https://software.broadinstitute.org/gatk/documentation/

Variant calling pipelines available at GC

- Validated for diagnostic use (human)
  - Amplicon
    - Multiplicom
    - Multiplex PCR
  - Targeted
    - Custom assays
    - Whole Exome Sequencing (WES)
  - WGS
- Research
  - Molecular Inversion Probes (MIPs)
  - Non-human species

Validation of SNV pipeline

- Use of cell lines
  - Father NA12877
  - Mother NA12878
  - Son NA12882



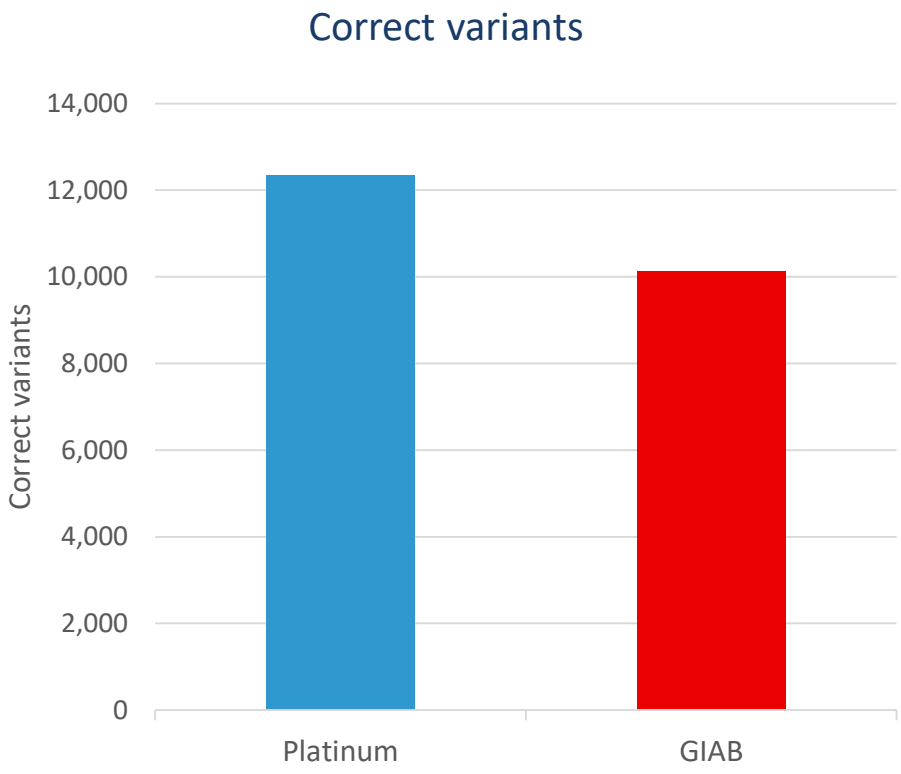
Cell line with Illumina Platinum calls

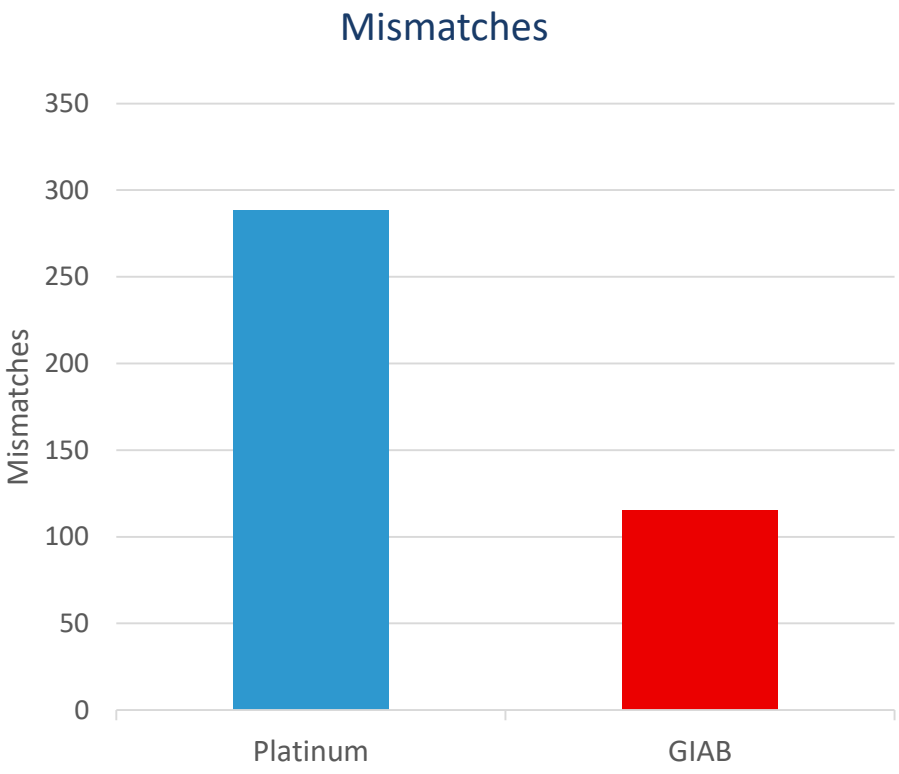
Cell line with Illumina Platinum calls + Genome in a bottle

Cell line

Validation of SNV pipeline

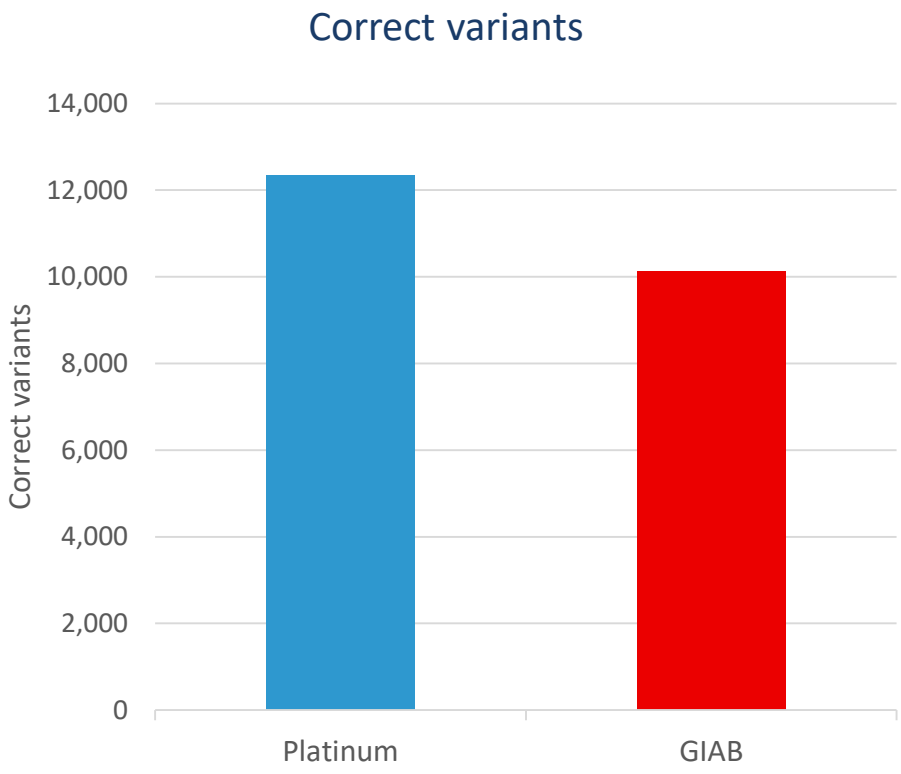
- Clinical exome (160X) - NA12878

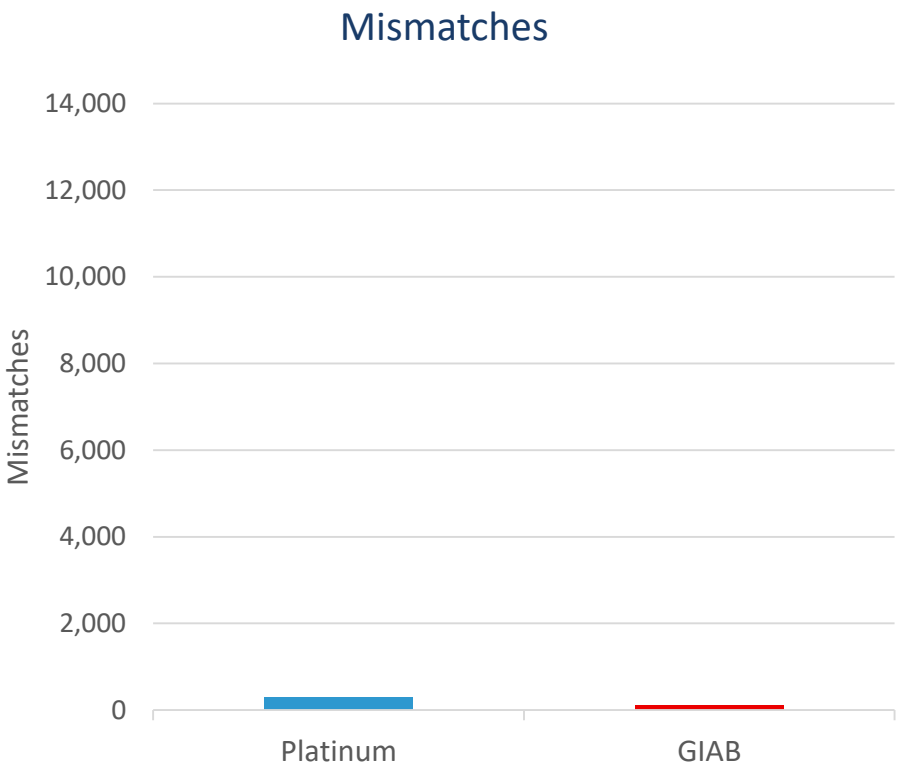




Validation of SNV pipeline

- Clinical exome (160X) - NA12878





Validation of SNV pipeline

- Clinical exome (160X) - NA12878
  - Correct variants
    - True Positives (TP)
  - Correct reference calls
    - True Negatives (TN)
  - Mismatches
    - False Positives (FP)
    - False Negatives (FN)

Sensitivity
Specificity

Use optimal variant filtering strategy to
- Maximise sensitivity & specificity
- Avoid the need to check variants in IGV

VCF quality control

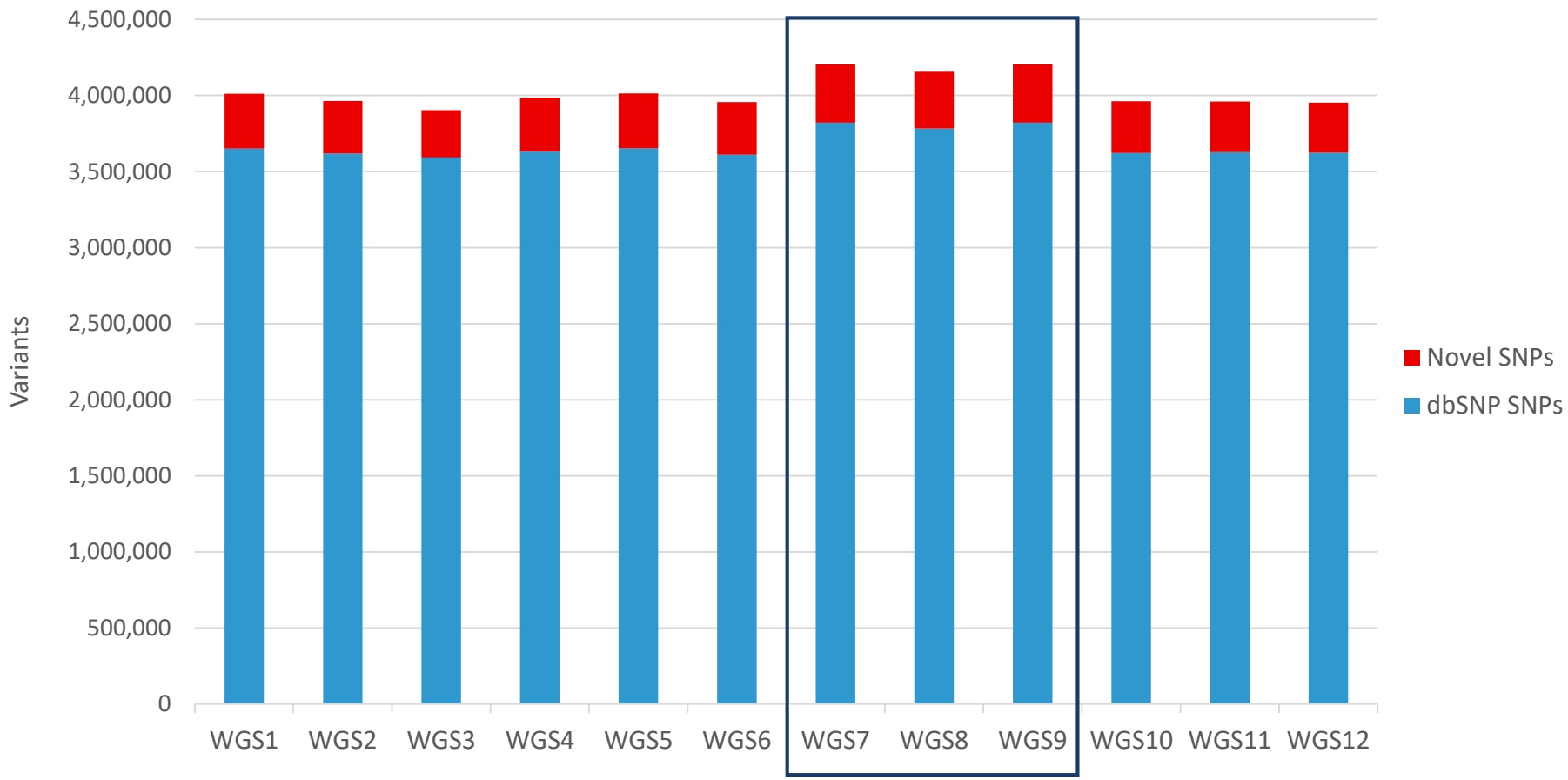
- Count and monitor
  - Number of variants per sample
    - SNPs
    - Indels
  - Transition/Transversion ratio
  - Insertion to deletion ratio
  - Number of heterozygous variants
  - Number of new variants (*i.e.* not in dbSNP)

```
java -jar CollectVariantCallingMetrics INPUT=sample.vcf.gz
OUTPUT=sample.vcf.metrics DBSNP=dbsnp138.vcf
```

https://broadinstitute.github.io/picard/

Variant metrics

- Known vs new variants (WGS)



Annotation

- *de novo* variant: TA/T

Format	Index	Mother	Father
GT	0/1	0/0	0/0
AD	43,51	103,0	81,0
DP	94	103	81
GQ	99	99	99
PL	1742,0,1448	0,120,1800	0,120,1800

  - FOXG1, associated to Rett syndrome (autosomal dominant)
  - Frameshift
  - Not found in public databases

Annovar

- Gene-based
  - Genic vs intergenic (RefSeq, UCSC, or ensembl)
  - Synonymous, non-synonymous, splicing, frameshift, intronic
- Region-based
  - Segmental duplications
  - Conserved regions
- Filter-based
  - Allele frequencies (gnomad, …)
  - Pathogenicity prediction scores (SIFT, PolyPhen, CADD, …)
  - Clinical interpretation (clinvar_20180603 & intervar_20180118)

```
table_annovar.pl sample.vcf database/ -buildver hg38 -out
sample_annotatef -remove -protocol refseq,cadd –operation
```

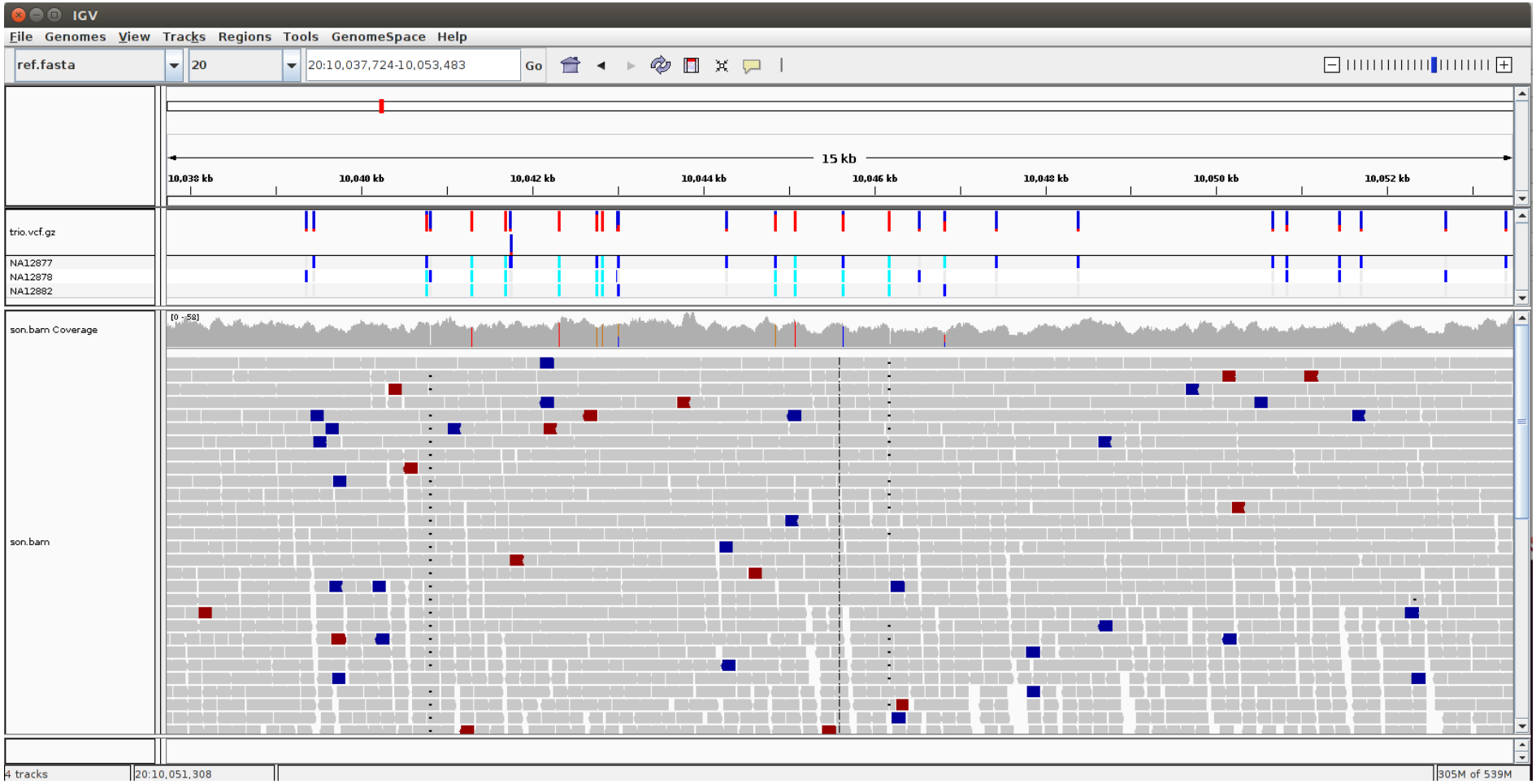
http://annovar.openbioinformatics.org/en/latest/

Questions?

GENOMICS CORELEUVEN

Luc Dehaspe
Erika Souche

Viewing variants in IGV



Genotype refinement workflow

- Extra step proposed by GATK