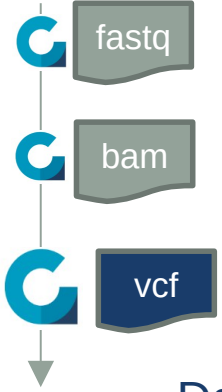
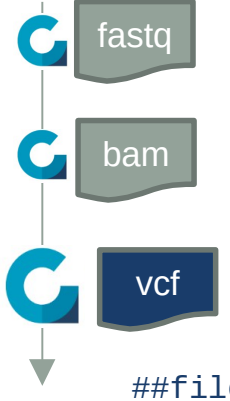


VCF file format

- Data https://software.broadinstitute.org/gatk/blog?id=9044
  2017 Feb workshop presentation slides and tutorial materials
  Germline Data Bundle (Day 2)
- VCF
  - Single individual `data/outputs/filtering/mother.vcf.gz`
  - Trio `data/inputVcfs/trio.vcf.gz`
- gVCF
  - Single individual `data/gvcfs/son.g.vcf`
- Open in text editor

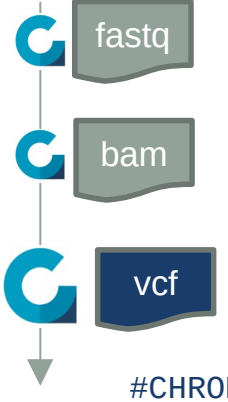
GENOMICS COREL EUVEN



VCF file format

Header

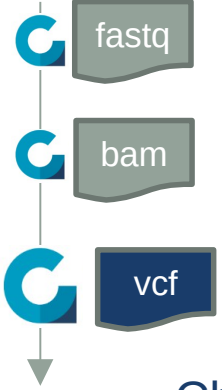
```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate 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">
##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">
#CHROM    POS  ID   REF  ALT  QUAL FILTER    INFO FORMAT    NA12878
```



VCF file format

Variants

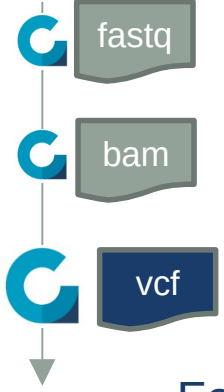
```
#CHROM    POS  ID   REF  ALT  QUAL FILTER    INFO FORMAT    NA12878
20   61098     .    C    T    465.13    .
AC=1;AF=0.500;AN=2;BaseQRankSum=0.516;ClippingRankSum=0.00;DP=44;DP_Orig=124;ExcessHet=3.
0103;FS=0.000;MQ=59.48;MQRankSum=0.803;QD=10.57;ReadPosRankSum=1.54;SOR=0.603
GT:AD:DP:GQ:PL 0/1:28,16:44:99:496,0,938
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=59.45;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=59.81;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
20   63244     .    A    C    923.13    .
AC=1;AF=0.500;AN=2;BaseQRankSum=0.637;ClippingRankSum=0.00;DP=57;DP_Orig=141;ExcessHet=3.
0103;FS=5.470;MQ=59.60;MQRankSum=-1.019e+00;QD=16.20;ReadPosRankSum=0.404;SOR=1.528
GT:AD:DP:GQ:PL 0/1:30,27:57:99:954,0,1064
...
```



VCF file format

Variants

- Chromosome 20
- Position 61098
- 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



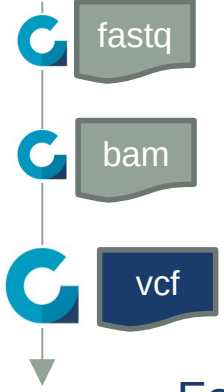
VCF file format

Variants

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

Genotype

0/1 Heterozygous



VCF file format

Variants

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

Genotype

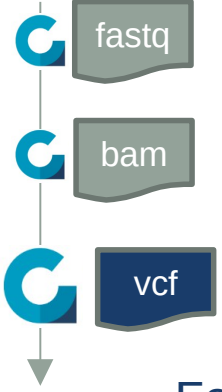
0/1 Heterozygous

Allelic Depth

28 reads supporting reference allele

16 reads supporting alternate allele

GENOMICS COREL EUVEN



VCF file format

Variants

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

Genotype

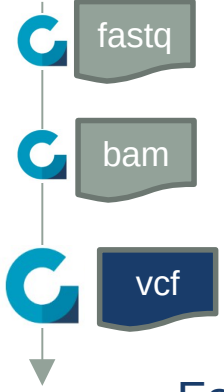
0/1 Heterozygous

Allelic Depth

28 reads supporting reference allele
16 reads supporting alternate allele

Depth

44 reads observed at that position



VCF file format

Variants

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

Genotype

0/1 Heterozygous

Allelic Depth

28 reads supporting reference allele
16 reads supporting alternate allele

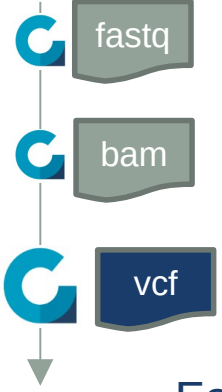
Genotype Quality

Smallest non-zero PL value
Maximum of 99

Depth

44 reads observed at that position

GENOMICS CORELEUVEN



VCF file format

Variants

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

Genotype
0/1 Heterozygous

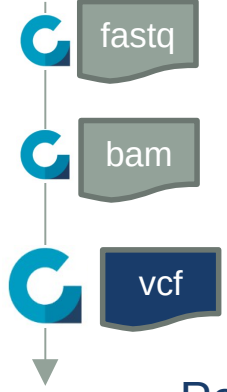
Phred Likelihood
Likelihood 0/0, 0/1, 1/1

Allelic Depth
28 reads supporting reference allele
16 reads supporting alternate allele

Genotype Quality
Smallest non-zero PL value
Maximum of 99

Depth
44 reads observed at that position

GENOMICS CORELEUVEN

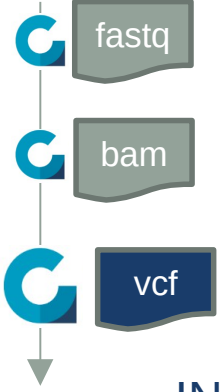


VCF file format

Multiple samples

- Possibility to have more than one sample

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
NA12877 NA12878 NA12882
20      61098   .       C       T       465.13  .       INFO    GT:AD:DP:GQ:PL
0/0:37,0:37:99:0,102,1529   0/1:28,16:44:99:496,0,938    0/0:43,0:43:99:0,99,1496
20      61138   .       C       CT      155.10  .       INFO    GT:AD:DP:GQ:PL
0/0:46,0:46:27:0,27,1425    0/1:21,11:32:99:195,0,464    0/0:43,0:43:99:0,99,1496
20      61795   .       G       T       2034.16 .       INFO    GT:AD:DP:GQ:PL
0/1:29,30:59:99:1063,0,1011 0/1:30,30:60:99:1003,0,1027  0/0:45,0:45:99:0,100,1755
...
```



VCF file format

Multiple samples

- INFO field

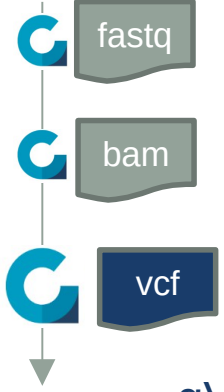
  Summary statistics of all samples included in the VCF file

```
0/0:37,0:37:99:0,102,1529    0/1:28,16:44:99:496,0,938    0/0:43,0:43:99:0,99,1496
```

```
AC=1;AF=0.167;AN=6;BaseQRankSum=0.516;ClippingRankSum=0.00;DP=124;ExcessHet=3.0103
;FS=0.000;MLEAC=1;MLEAF=0.167;MQ=59.48;MQRankSum=0.803;QD=10.57;ReadPosRankSum=1.5
4;SOR=0.603
```

AC=1	Allele Count	1 observed alternate allele
AF=0.167	Allele Frequency	1 alternate out of 6 alleles
DP=124	Depth	Sum of DP of all samples

...



gVCF file format

- gVCF → genome VCF
- Store reference information
  - Nucleotide level
  - Region level

```
#CHROM  POS             ID      REF     ALT             QUAL    FILTER
    INFO    FORMAT                          NA12882
20      10000000        .       T       <NON_REF>       .       .
    END=10000001    GT:DP:GQ:MIN_DP:PL      0/0:20:57:20:0,57,855
20      10000002        .       G       <NON_REF>       .       .
    END=10000002    GT:DP:GQ:MIN_DP:PL      0/0:20:48:20:0,48,555
20      10000003        .       T       <NON_REF>       .       .
    END=10000003    GT:DP:GQ:MIN_DP:PL      0/0:19:22:19:0,22,513
20      10000004        .       T       <NON_REF>       .       .
    END=10000004    GT:DP:GQ:MIN_DP:PL      0/0:20:31:20:0,31,553
...
```

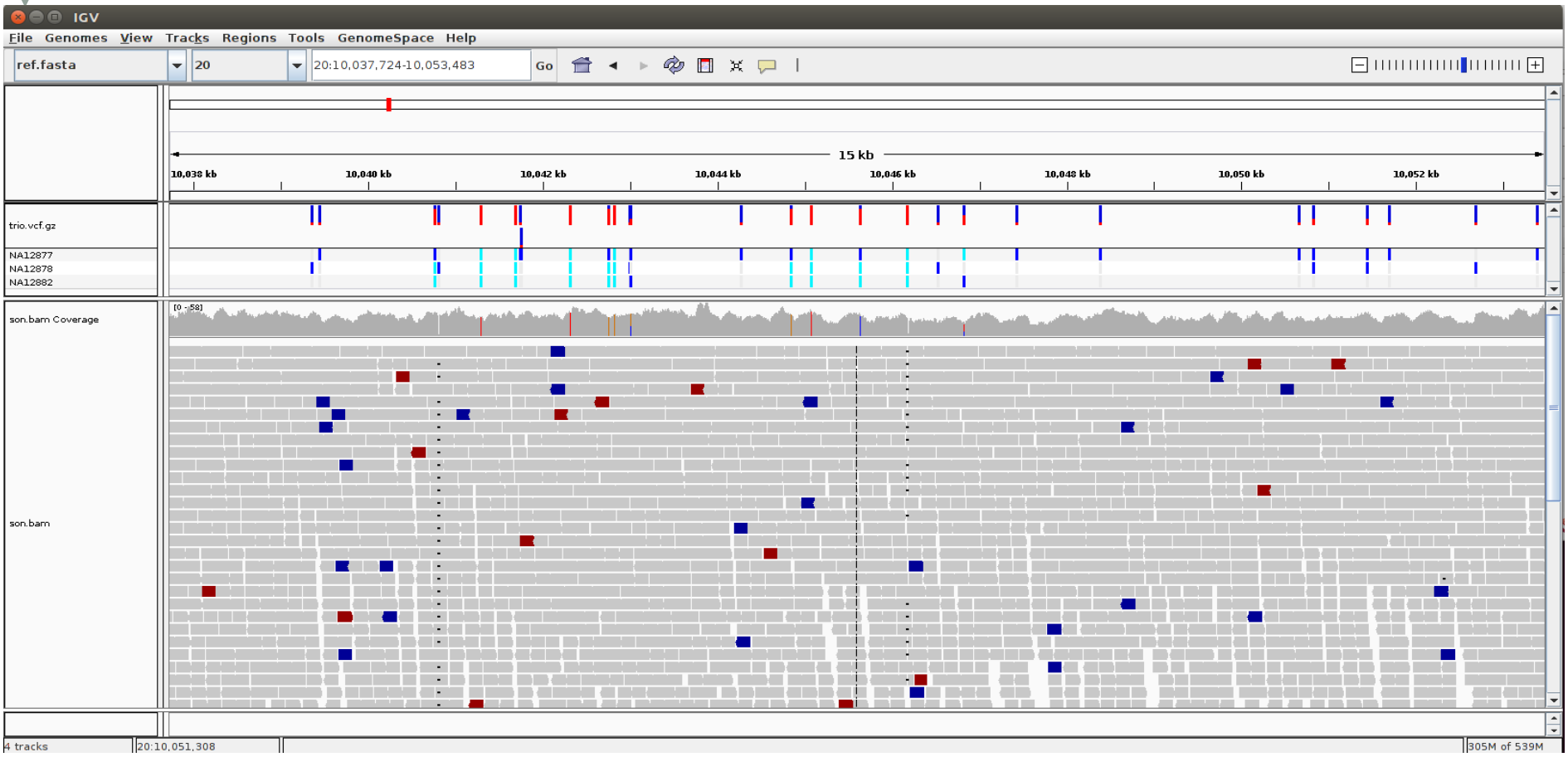
fastq

bam

vcf

Viewing variants

IGV



GENOMICS CORELEUVEN

fastq

bam

vcf

Viewing variants
IGV

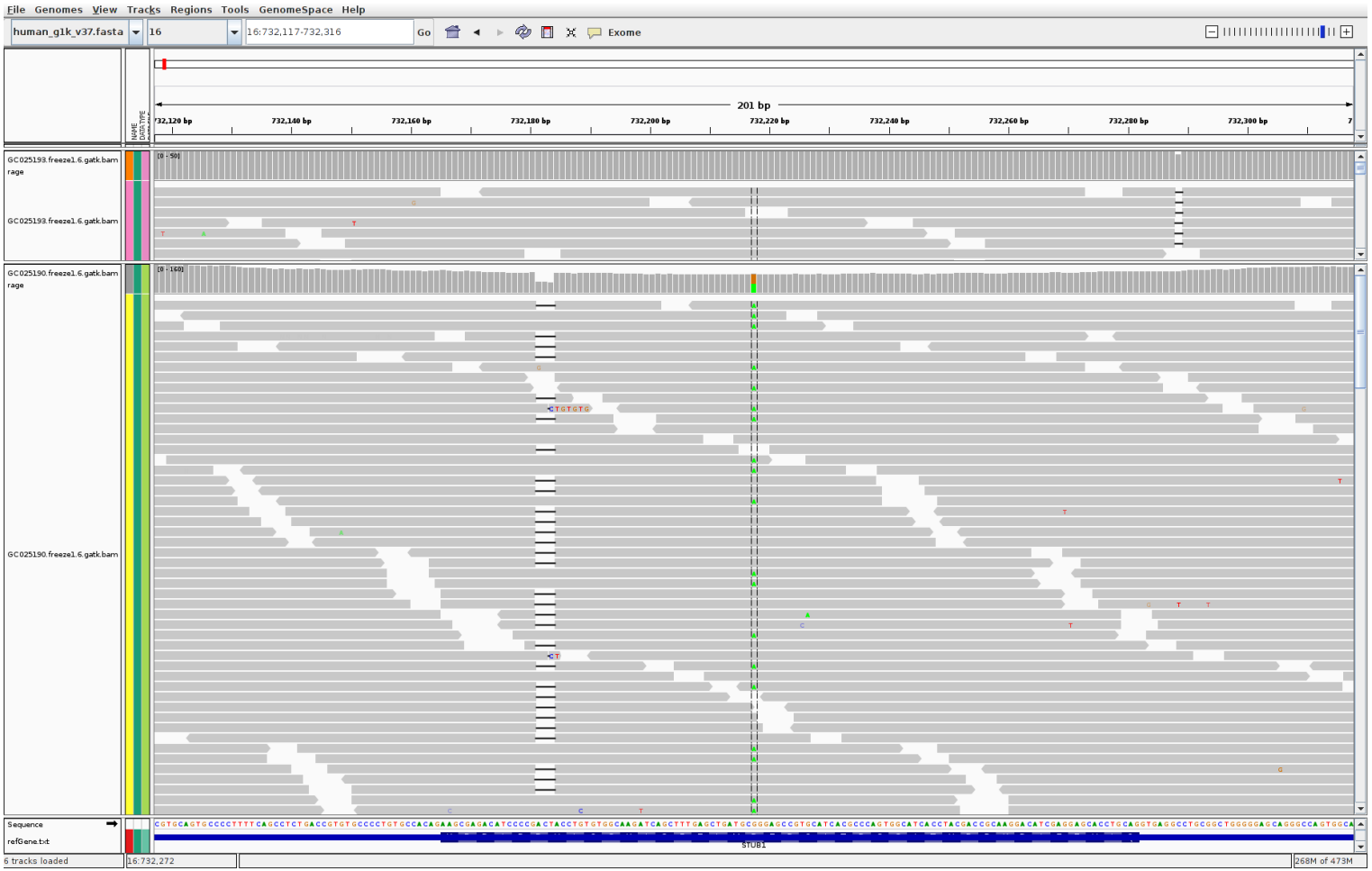
GENOMICS CORELEUVEN

fastq

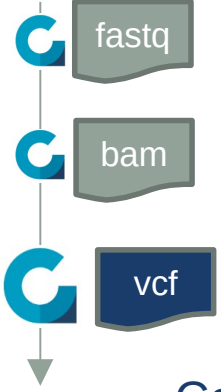
bam

vcf

Viewing variants
IGV



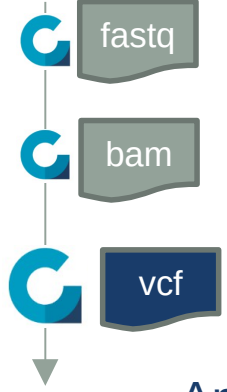
GENOMICS CORELEUVEN



VCF quality control

- Count and monitor
  - Number of variants per sample
    - SNPs
    - Indels
  - Transition/Transversion ratio
  - Insertion to deletion ratio
  - Number of heterozygous variants
  - ...

- GATK VariantEval
  http://gatkforums.broadinstitute.org/gatk/discussion/6211/howto-evaluate-a-callset-with-varianteval



VCF quality control

- Annotated VCF files
  - Number of new variants (i.e. not in dbSNP)
  - Number of synonymous vs non-synonymous variants
  - Transition/Transversion ratio in coding regions
  - ...

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