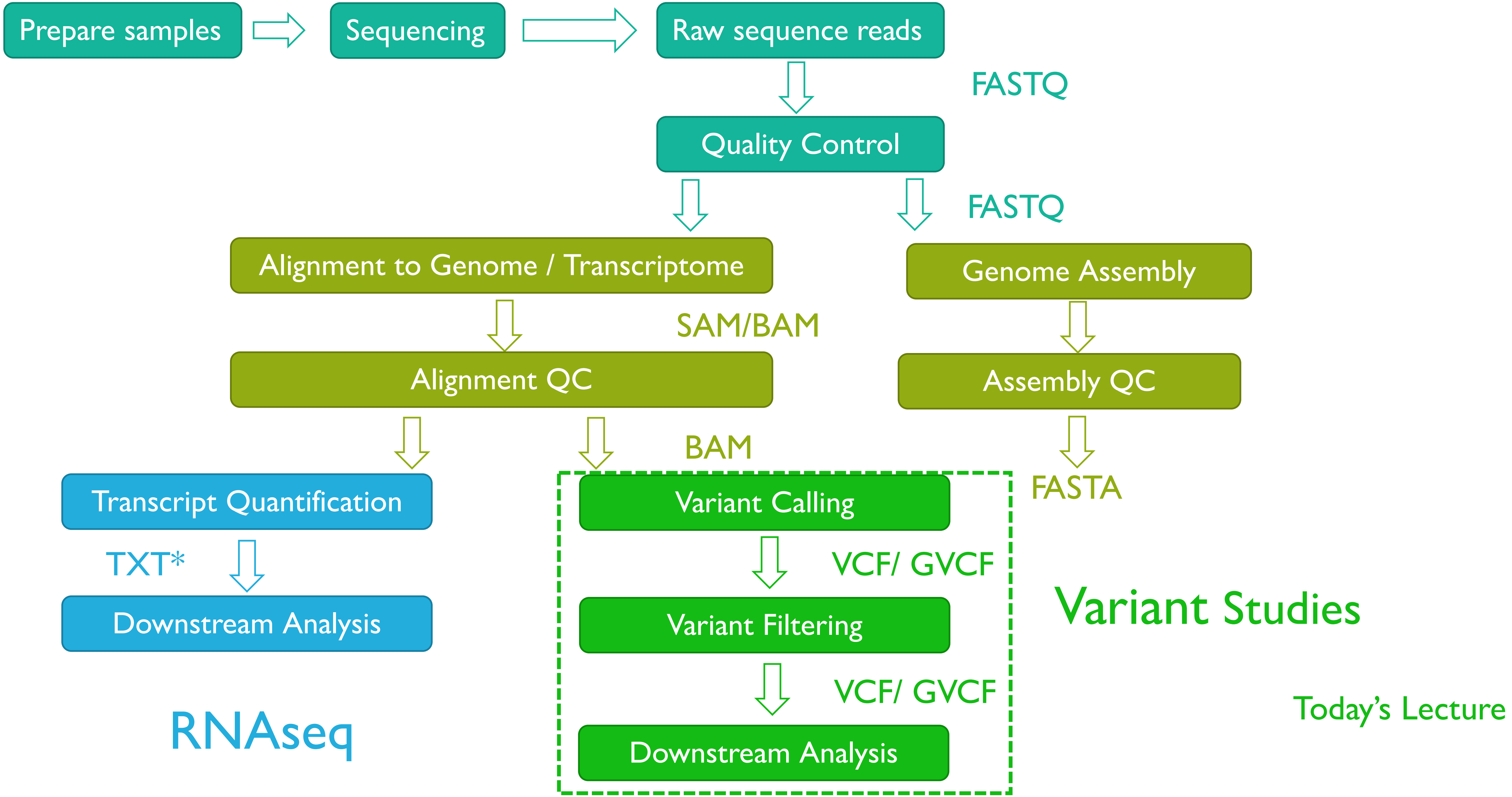
Variants and Filtering

RCN Bioinformatics Mentoring 2021


Prepare samples
Sequencing
Raw sequence reads
FASTQ
Quality Control
FASTQ
Alignment to Genome / Transcriptome
Genome Assembly
SAM/BAM
Alignment QC
Assembly QC
BAM
FASTA
Transcript Quantification
TXT*
Downstream Analysis
RNAseq
Variant Calling
VCF/ GVCF
Variant Filtering
VCF/ GVCF
Downstream Analysis
Variant Studies
Today's Lecture

OUTLINE

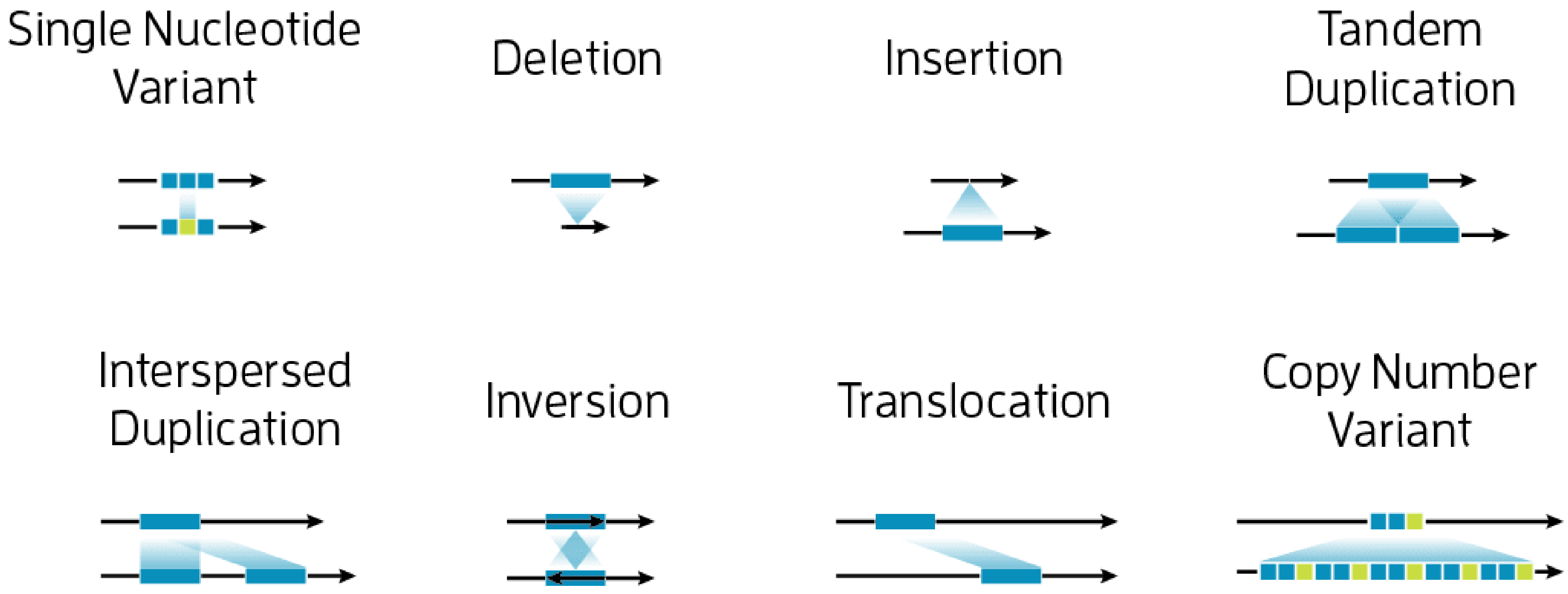
Lecture

- How do we get SNP and genotype calls?
- What is in a VCF file?
- Why and how do we filter?

Tutorial

- SNP calling – review code & Questions
- Filtering – Do it yourself.

GENOMIC VARIATION



Types of Variants

Figure from https://www.pacb.com/applications/whole-genome-sequencing/variant-detection/

GENOMIC VARIATION

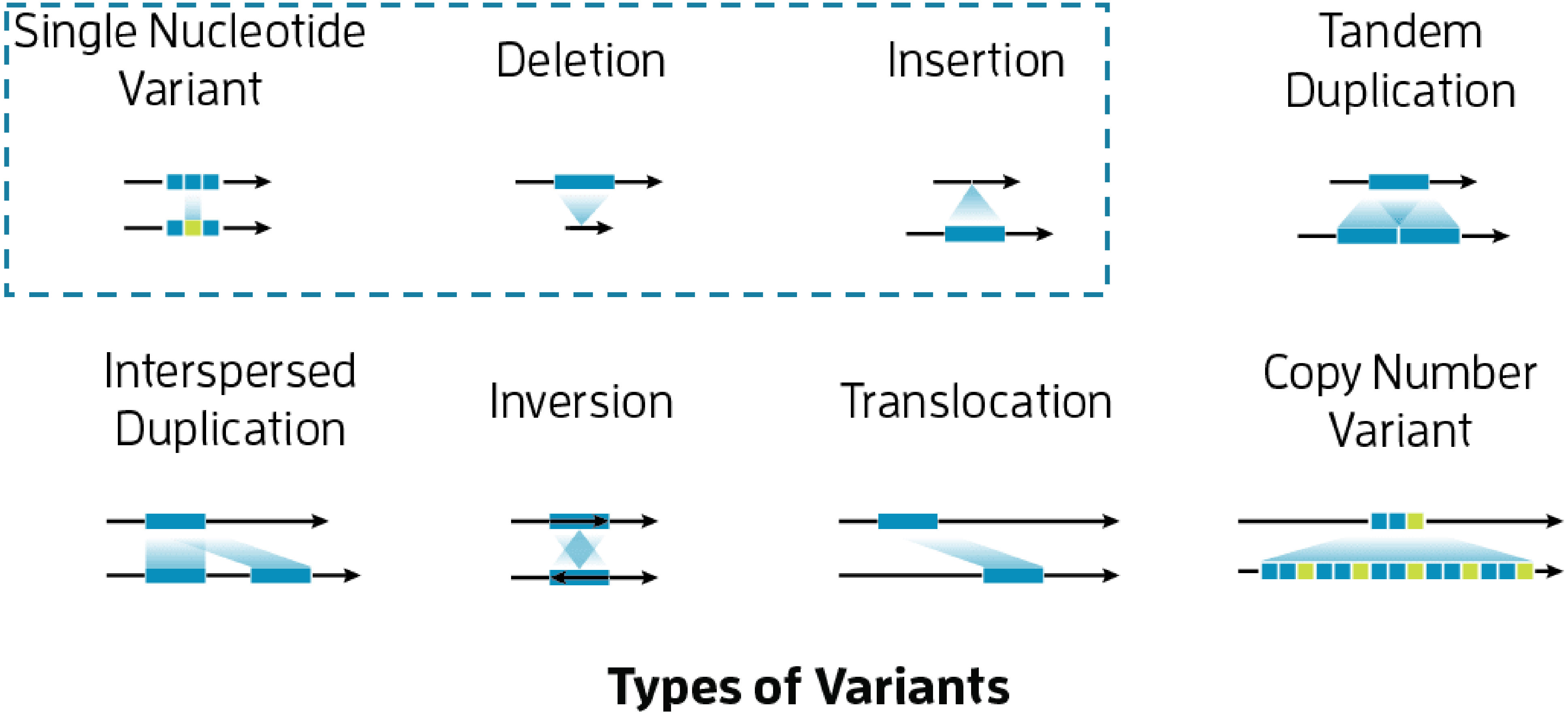


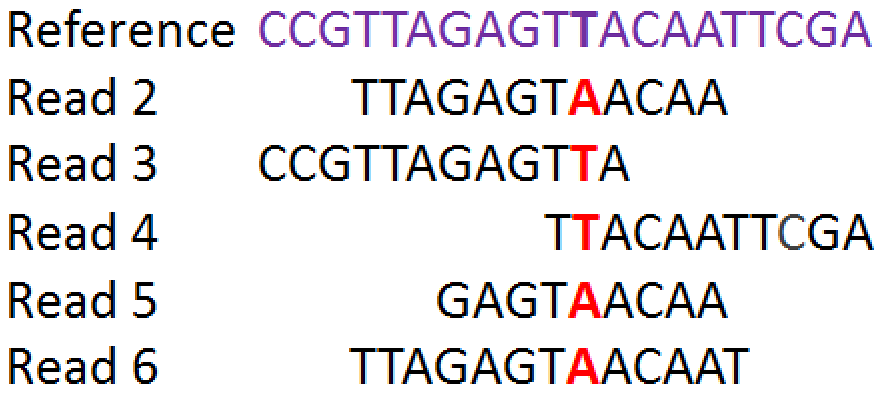
Figure from https://www.pacb.com/applications/whole-genome-sequencing/variant-detection/

SNP AND GENOTYPE CALLING

- *SNP calling* finds the alleles.
- *Genotype calling* finds the genotype of the individual

SNP AND GENOTYPE CALLING

- *SNP calling* finds the alleles.
- *Genotype calling* finds the genotype of the individual



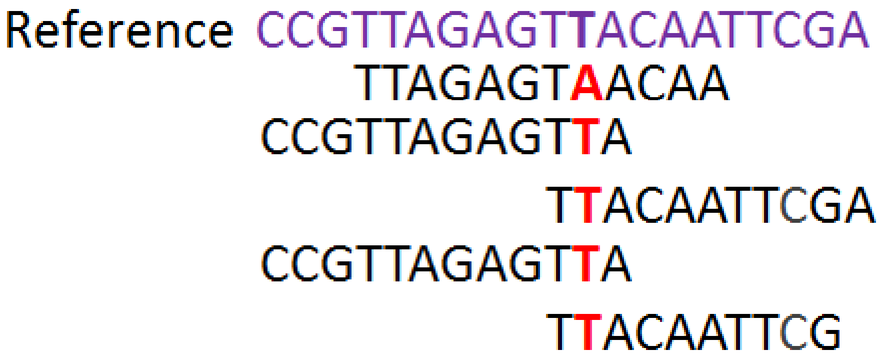
```
Reference  CCGTTAGAGTTACAATTCGA
Read 2        TTAGAGTAACAA
Read 3     CCGTTAGAGTTA
Read 4              TTACAATTCGA
Read 5           GAGTAACAA
Read 6        TTAGAGTAACAAT
```

- Alternate allele A: 3 reads
- Genotype: T/A (2/3)
- Unphased genotype

Figure from https://online.stat.psu.edu/stat555/node/106/

SNP AND GENOTYPE CALLING

- *SNP calling* finds the alleles.
- *Genotype calling* finds the genotype of the individual



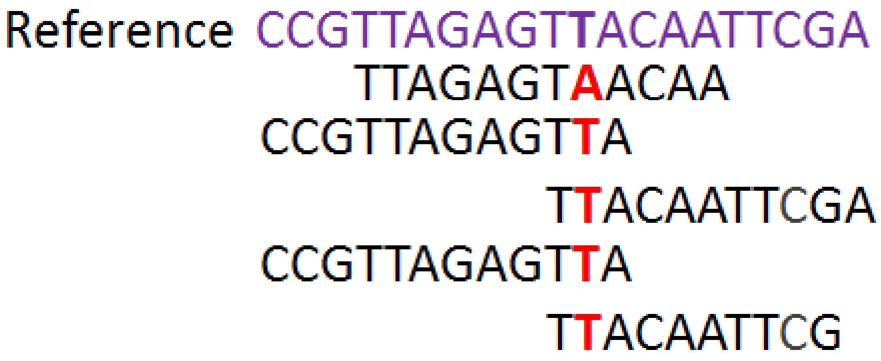
```
Reference CCGTTAGAGTTACAATTCGA
             TTAGAGTAACAA
          CCGTTAGAGTTA
                   TTACAATTCGA
          CCGTTAGAGTTA
                   TTACAATTCG
```

- Alternate allele A: 1 reads
- Genotype: T/A (1/4)

Figure from https://online.stat.psu.edu/stat555/node/106/

SNP AND GENOTYPE CALLING

- *SNP calling* finds the alleles.
- *Genotype calling* finds the genotype of the individual



```
Reference CCGTTAGAGTTACAATTCGA
             TTAGAGTAACAA
          CCGTTAGAGTTA
                   TTACAATTCGA
          CCGTTAGAGTTA
                   TTACAATTCG
```

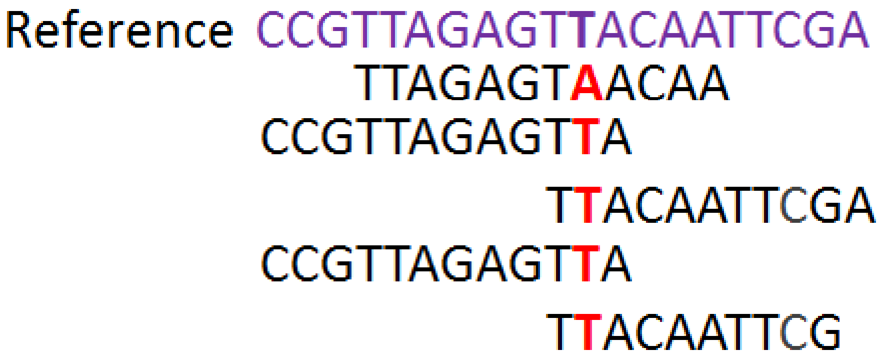
- Alternate allele A: 1 reads
- Genotype: T/A (1/4)
- How confident are you in that A? Do you really think it's a heterozygote?

Figure from https://online.stat.psu.edu/stat555/node/106/

SNP AND GENOTYPE CALLING

- Confidence in alternate allele depends on:
  - Map quality of base in individual reads
  - Number of reads supporting the base

- Confidence in genotype depends on:
  - Map quality of base in individual reads
  - Number of reads supporting each allele in an individual.
  - Ratio of reads supporting 1 allele vs the other.



```
Reference CCGTTAGAGTTACAATTCGA
             TTAGAGTAACAA
          CCGTTAGAGTTA
                   TTACAATTCGA
          CCGTTAGAGTTA
                   TTACAATTCG
```

Figure from https://online.stat.psu.edu/stat555/node/106/

GENOTYPE LIKELIHOODS

- Describes the probability of a genotype given the data

$$L(G=\{A_1,A_2\}|D) \propto Pr(D|G=A_1,A_2), \qquad A_1,A_2 \in \{A,C,G,T\}.$$

  - e.g. $Pr(D|G=\{A_1,A_2\}) = \prod_{i=1}^{M} Pr\,(b_i|G=\{A_1,A_2\}) = \prod_{i=1}^{M}(\frac{1}{2}Pr(b_i|A_1)+\frac{1}{2}Pr(b_i|A_2))$

    Where M is sequencing depth and bi is the base at a given read.

- Different software use different genotype likelihood methods
- Phred scaled -10 log10(Genotype Likelihood) (PL in VCF)

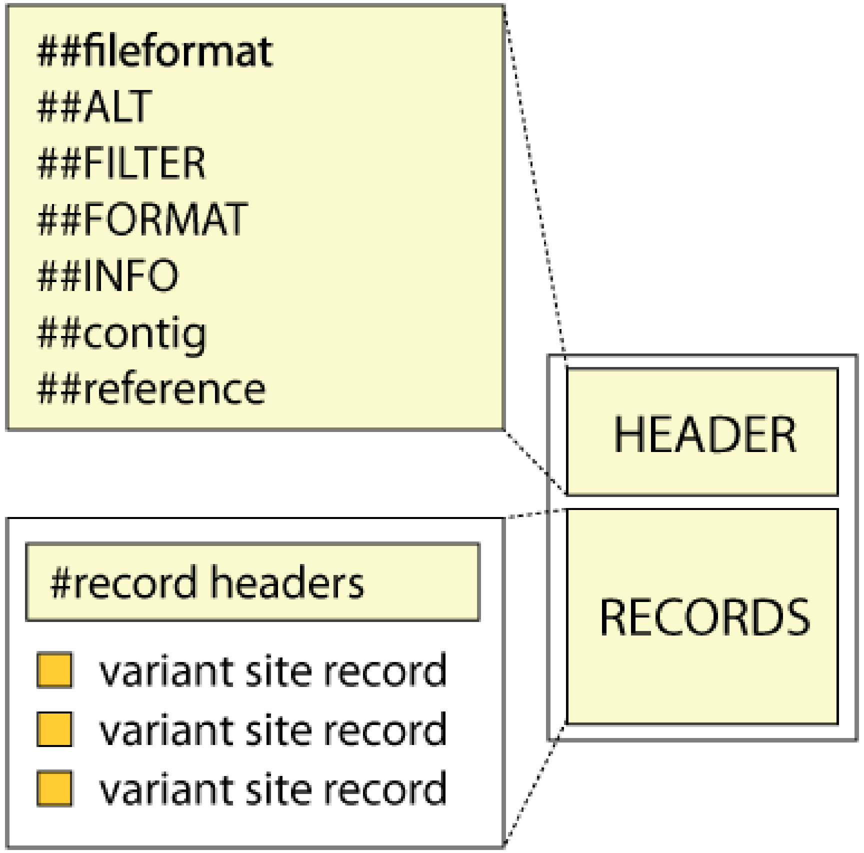
- Take Home message: the quality of your genotypes are going to depend on your sequencing depth!

Formulae http://www.popgen.dk/angsd/index.php/Genotype_Likelihoods;
see also https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198575/

VCF FORMAT

- Representation of SNV, and Indels for one or more individuals
- Includes a lot of information on quality, number of reads per site and per individual
- Individual genotype information

Basic structure of a VCF file

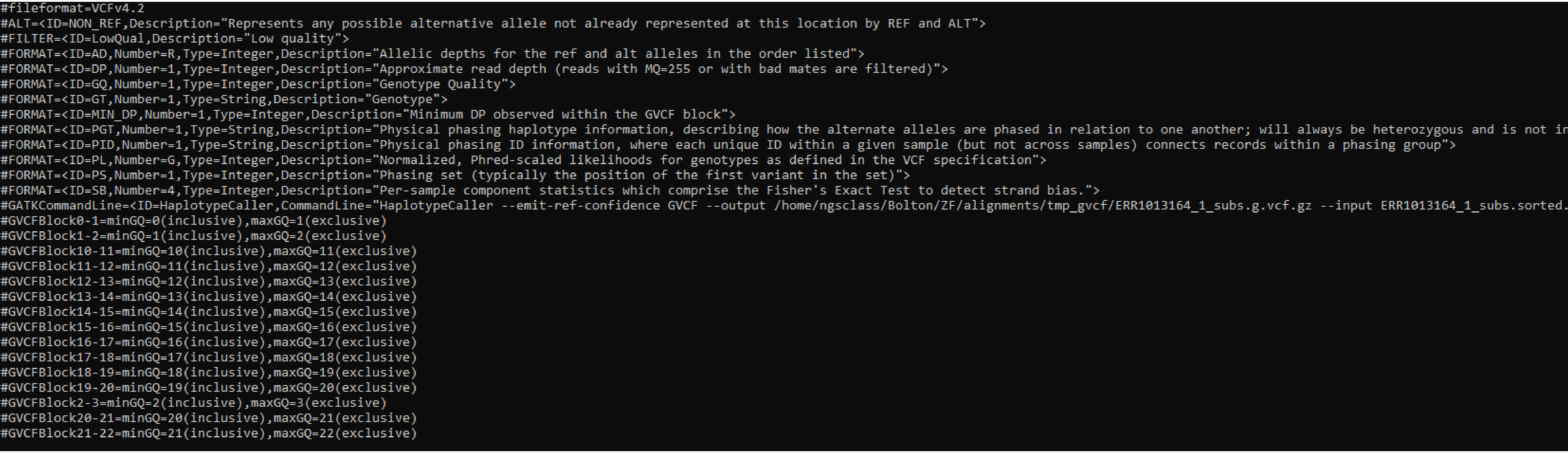


https://gatk.broadinstitute.org/hc/en-us/articles/360035531692

VCF HEADER

- Contains information describing
  - How file was created
  - Record contents metadata
- Commands
- Reference genome and contig names and lengths.
- Preceded by ‘##’

FROM THE TUTORIAL…



```
#fileformat=VCFv4.2
#ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
#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=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
#FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another; will always be heterozygous and is not in
#FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
#FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
#FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
#FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
#GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --emit-ref-confidence GVCF --output /home/ngsclass/Bolton/ZF/alignments/tmp_gvcf/ERR1013164_1_subs.g.vcf.gz --input ERR1013164_1_subs.sorted.
#GVCFBlock0-1=minGQ=0(inclusive),maxGQ=1(exclusive)
#GVCFBlock1-2=minGQ=1(inclusive),maxGQ=2(exclusive)
#GVCFBlock10-11=minGQ=10(inclusive),maxGQ=11(exclusive)
#GVCFBlock11-12=minGQ=11(inclusive),maxGQ=12(exclusive)
#GVCFBlock12-13=minGQ=12(inclusive),maxGQ=13(exclusive)
#GVCFBlock13-14=minGQ=13(inclusive),maxGQ=14(exclusive)
#GVCFBlock14-15=minGQ=14(inclusive),maxGQ=15(exclusive)
#GVCFBlock15-16=minGQ=15(inclusive),maxGQ=16(exclusive)
#GVCFBlock16-17=minGQ=16(inclusive),maxGQ=17(exclusive)
#GVCFBlock17-18=minGQ=17(inclusive),maxGQ=18(exclusive)
#GVCFBlock18-19=minGQ=18(inclusive),maxGQ=19(exclusive)
#GVCFBlock19-20=minGQ=19(inclusive),maxGQ=20(exclusive)
#GVCFBlock2-3=minGQ=2(inclusive),maxGQ=3(exclusive)
#GVCFBlock20-21=minGQ=20(inclusive),maxGQ=21(exclusive)
#GVCFBlock21-22=minGQ=21(inclusive),maxGQ=22(exclusive)
```

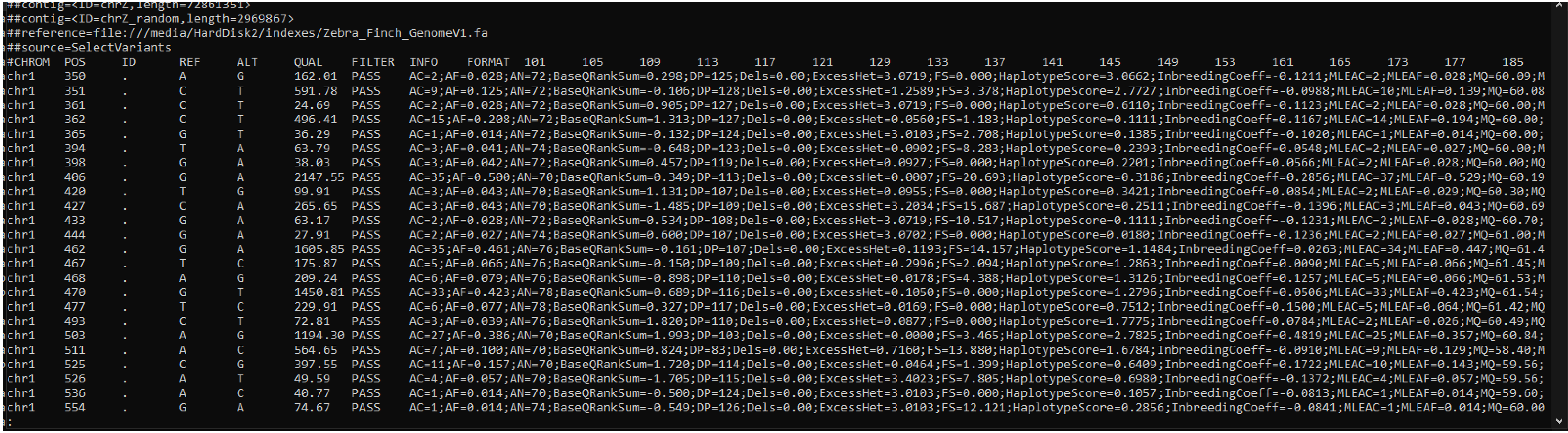
VCF RECORDS

- Each row is a site

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1…
20	10001298	.	T	A	884.77	.	Clipped for brevity	GT:AD:DP:GQ:PL	1/1:0,30:30:89:913,89,0
20	10001436	.	A	AAGGCT	884.77	.	…	GT:AD:DP:GQ:PL	1/1:0,28:28:84:1260,84,0
20	10004769	.	TAAAACTATGC	T	622.73	.	…	GT:AD:DP:GQ:PL	0/1:18,17:35:99:660,0,704

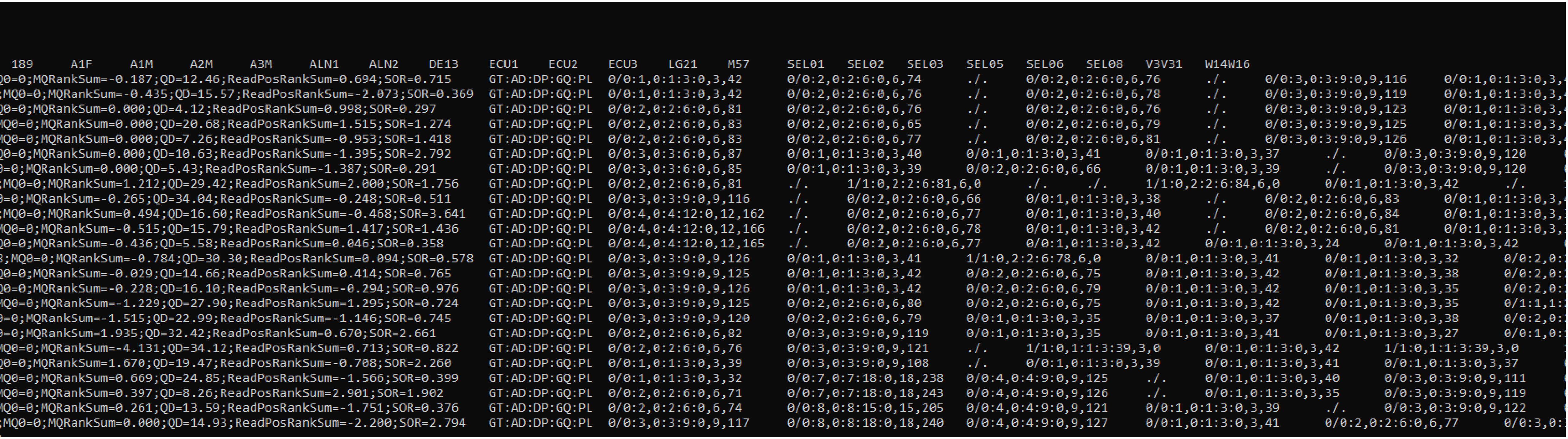
https://gatk.broadinstitute.org/hc/en-us/articles/360035531692

A ZF VCF FILE



```
##contig=<ID=chrZ,length=72861351>
##contig=<ID=chrZ_random,length=2969867>
##reference=file:///media/HardDisk2/indexes/Zebra_Finch_GenomeV1.fa
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  101     105     109     113     117     121     129     133     137     141     145     149     153     161     165     173     177     185
chr1    350     .       A       G       162.01  PASS    AC=2;AF=0.028;AN=72;BaseQRankSum=0.298;DP=125;Dels=0.00;ExcessHet=3.0719;FS=0.000;HaplotypeScore=3.0662;InbreedingCoeff=-0.1211;MLEAC=2;MLEAF=0.028;MQ=60.09;M
chr1    351     .       C       T       591.78  PASS    AC=9;AF=0.125;AN=72;BaseQRankSum=-0.106;DP=128;Dels=0.00;ExcessHet=1.2589;FS=3.378;HaplotypeScore=2.7727;InbreedingCoeff=-0.0988;MLEAC=10;MLEAF=0.139;MQ=60.08
chr1    361     .       C       T       24.69   PASS    AC=2;AF=0.028;AN=72;BaseQRankSum=0.905;DP=127;Dels=0.00;ExcessHet=3.0719;FS=0.000;HaplotypeScore=0.6110;InbreedingCoeff=-0.1123;MLEAC=2;MLEAF=0.028;MQ=60.00;M
chr1    362     .       C       T       496.41  PASS    AC=15;AF=0.208;AN=72;BaseQRankSum=1.313;DP=127;Dels=0.00;ExcessHet=0.0560;FS=1.183;HaplotypeScore=0.1111;InbreedingCoeff=0.1167;MLEAC=14;MLEAF=0.194;MQ=60.00;
chr1    365     .       G       T       36.29   PASS    AC=1;AF=0.014;AN=72;BaseQRankSum=-0.132;DP=124;Dels=0.00;ExcessHet=3.0103;FS=2.708;HaplotypeScore=0.1385;InbreedingCoeff=-0.1020;MLEAC=1;MLEAF=0.014;MQ=60.00;
chr1    394     .       T       A       63.79   PASS    AC=3;AF=0.041;AN=74;BaseQRankSum=-0.648;DP=123;Dels=0.00;ExcessHet=0.0902;FS=8.283;HaplotypeScore=0.2393;InbreedingCoeff=0.0548;MLEAC=2;MLEAF=0.027;MQ=60.00;M
chr1    398     .       G       A       38.03   PASS    AC=3;AF=0.042;AN=72;BaseQRankSum=0.457;DP=119;Dels=0.00;ExcessHet=0.0927;FS=0.000;HaplotypeScore=0.2201;InbreedingCoeff=0.0566;MLEAC=2;MLEAF=0.028;MQ=60.00;MQ
chr1    406     .       G       A       2147.55 PASS    AC=35;AF=0.500;AN=70;BaseQRankSum=0.349;DP=113;Dels=0.00;ExcessHet=0.0007;FS=20.693;HaplotypeScore=0.3186;InbreedingCoeff=0.2856;MLEAC=37;MLEAF=0.529;MQ=60.19
chr1    420     .       T       G       99.91   PASS    AC=3;AF=0.043;AN=70;BaseQRankSum=1.131;DP=107;Dels=0.00;ExcessHet=0.0955;FS=0.000;HaplotypeScore=0.3421;InbreedingCoeff=0.0854;MLEAC=2;MLEAF=0.029;MQ=60.30;MQ
chr1    427     .       C       A       265.65  PASS    AC=3;AF=0.043;AN=70;BaseQRankSum=-1.485;DP=109;Dels=0.00;ExcessHet=3.2034;FS=15.687;HaplotypeScore=0.2511;InbreedingCoeff=-0.1396;MLEAC=3;MLEAF=0.043;MQ=60.69
chr1    433     .       G       A       63.17   PASS    AC=2;AF=0.028;AN=72;BaseQRankSum=0.534;DP=108;Dels=0.00;ExcessHet=3.0719;FS=10.517;HaplotypeScore=0.1111;InbreedingCoeff=-0.1231;MLEAC=2;MLEAF=0.028;MQ=60.70;
chr1    444     .       G       A       27.91   PASS    AC=2;AF=0.027;AN=74;BaseQRankSum=0.600;DP=107;Dels=0.00;ExcessHet=3.0702;FS=0.000;HaplotypeScore=0.0180;InbreedingCoeff=-0.1236;MLEAC=2;MLEAF=0.027;MQ=61.00;M
chr1    462     .       G       A       1605.85 PASS    AC=35;AF=0.461;AN=76;BaseQRankSum=-0.161;DP=107;Dels=0.00;ExcessHet=0.1193;FS=14.157;HaplotypeScore=1.1484;InbreedingCoeff=0.0263;MLEAC=34;MLEAF=0.447;MQ=61.4
chr1    467     .       T       C       175.87  PASS    AC=5;AF=0.066;AN=76;BaseQRankSum=-0.150;DP=109;Dels=0.00;ExcessHet=0.2996;FS=2.094;HaplotypeScore=1.2863;InbreedingCoeff=0.0090;MLEAC=5;MLEAF=0.066;MQ=61.45;M
chr1    468     .       A       G       209.24  PASS    AC=6;AF=0.079;AN=76;BaseQRankSum=-0.898;DP=110;Dels=0.00;ExcessHet=0.0178;FS=4.388;HaplotypeScore=1.3126;InbreedingCoeff=0.1257;MLEAC=5;MLEAF=0.066;MQ=61.53;M
chr1    470     .       G       T       1450.81 PASS    AC=33;AF=0.423;AN=78;BaseQRankSum=0.689;DP=116;Dels=0.00;ExcessHet=0.1050;FS=0.000;HaplotypeScore=1.2796;InbreedingCoeff=0.0506;MLEAC=33;MLEAF=0.423;MQ=61.54;
chr1    477     .       T       C       229.91  PASS    AC=6;AF=0.077;AN=78;BaseQRankSum=0.327;DP=117;Dels=0.00;ExcessHet=0.0169;FS=0.000;HaplotypeScore=0.7512;InbreedingCoeff=0.1500;MLEAC=5;MLEAF=0.064;MQ=61.42;MQ
chr1    493     .       C       T       72.81   PASS    AC=3;AF=0.039;AN=76;BaseQRankSum=1.820;DP=110;Dels=0.00;ExcessHet=0.0877;FS=0.000;HaplotypeScore=1.7775;InbreedingCoeff=0.0784;MLEAC=2;MLEAF=0.026;MQ=60.49;MQ
chr1    503     .       A       G       1194.30 PASS    AC=27;AF=0.386;AN=70;BaseQRankSum=1.993;DP=103;Dels=0.00;ExcessHet=0.0000;FS=3.465;HaplotypeScore=2.7825;InbreedingCoeff=0.4819;MLEAC=25;MLEAF=0.357;MQ=60.84;
chr1    511     .       A       C       564.65  PASS    AC=7;AF=0.100;AN=70;BaseQRankSum=0.824;DP=83;Dels=0.00;ExcessHet=0.7160;FS=13.880;HaplotypeScore=1.6784;InbreedingCoeff=-0.0910;MLEAC=9;MLEAF=0.129;MQ=58.40;M
chr1    525     .       C       G       397.55  PASS    AC=11;AF=0.157;AN=70;BaseQRankSum=1.720;DP=114;Dels=0.00;ExcessHet=0.0464;FS=1.399;HaplotypeScore=0.6409;InbreedingCoeff=0.1722;MLEAC=10;MLEAF=0.143;MQ=59.56;
chr1    526     .       A       T       49.59   PASS    AC=4;AF=0.057;AN=70;BaseQRankSum=-1.705;DP=115;Dels=0.00;ExcessHet=3.4023;FS=7.805;HaplotypeScore=0.6980;InbreedingCoeff=-0.1372;MLEAC=4;MLEAF=0.057;MQ=59.56;
chr1    536     .       A       C       40.77   PASS    AC=1;AF=0.014;AN=70;BaseQRankSum=-0.500;DP=124;Dels=0.00;ExcessHet=3.0103;FS=0.000;HaplotypeScore=0.1057;InbreedingCoeff=-0.0813;MLEAC=1;MLEAF=0.014;MQ=59.60;
chr1    554     .       G       A       74.67   PASS    AC=1;AF=0.014;AN=74;BaseQRankSum=-0.549;DP=126;Dels=0.00;ExcessHet=3.0103;FS=12.121;HaplotypeScore=0.2856;InbreedingCoeff=-0.0841;MLEAC=1;MLEAF=0.014;MQ=60.00
:
```

A ZF VCF FILE



```
  189     A1F     A1M     A2M     A3M     ALN1    ALN2    DE13    ECU1    ECU2    ECU3    LG21    M57     SEL01   SEL02   SEL03   SEL05   SEL06   SEL08   V3V31   W14W16
Q0=0;MQRankSum=-0.187;QD=12.46;ReadPosRankSum=0.694;SOR=0.715     GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,42      0/0:2,0:2:6:0,6,74      ./.     0/0:2,0:2:6:0,6,76      ./.     0/0:3,0:3:9:0,9,116     0/0:1,0:1:3:0,3,4
;MQ0=0;MQRankSum=-0.435;QD=15.57;ReadPosRankSum=-2.073;SOR=0.369  GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,42      0/0:2,0:2:6:0,6,76      ./.     0/0:2,0:2:6:0,6,78      ./.     0/0:3,0:3:9:0,9,119     0/0:1,0:1:3:0,3,4
Q0=0;MQRankSum=0.000;QD=4.12;ReadPosRankSum=0.998;SOR=0.297       GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,81      0/0:2,0:2:6:0,6,76      ./.     0/0:2,0:2:6:0,6,76      ./.     0/0:3,0:3:9:0,9,123     0/0:1,0:1:3:0,3,4
MQ0=0;MQRankSum=0.000;QD=20.68;ReadPosRankSum=1.515;SOR=1.274     GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,83      0/0:2,0:2:6:0,6,65      ./.     0/0:2,0:2:6:0,6,79      ./.     0/0:3,0:3:9:0,9,125     0/0:1,0:1:3:0,3,4
MQ0=0;MQRankSum=0.000;QD=7.26;ReadPosRankSum=-0.953;SOR=1.418     GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,83      0/0:2,0:2:6:0,6,77      ./.     0/0:2,0:2:6:0,6,81      ./.     0/0:3,0:3:9:0,9,126     0/0:1,0:1:3:0,3,4
Q0=0;MQRankSum=0.000;QD=10.63;ReadPosRankSum=-1.395;SOR=2.792     GT:AD:DP:GQ:PL  0/0:3,0:3:6:0,6,87      0/0:1,0:1:3:0,3,40      0/0:1,0:1:3:0,3,41      0/0:1,0:1:3:0,3,37      ./.     0/0:3,0:3:9:0,9,120
0=0;MQRankSum=0.000;QD=5.43;ReadPosRankSum=-1.387;SOR=0.291       GT:AD:DP:GQ:PL  0/0:3,0:3:6:0,6,85      0/0:1,0:1:3:0,3,39      0/0:2,0:2:6:0,6,66      0/0:1,0:1:3:0,3,39      ./.     0/0:3,0:3:9:0,9,120
;MQ0=0;MQRankSum=1.212;QD=29.42;ReadPosRankSum=2.000;SOR=1.756    GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,81      ./.     1/1:0,2:2:6:81,6,0      ./.     ./.     1/1:0,2:2:6:84,6,0      0/0:1,0:1:3:0,3,42      ./.
0=0;MQRankSum=-0.265;QD=34.04;ReadPosRankSum=-0.248;SOR=0.511     GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,116     ./.     0/0:2,0:2:6:0,6,66      0/0:1,0:1:3:0,3,38      ./.     0/0:2,0:2:6:0,6,83      0/0:1,0:1:3:0,3,
;MQ0=0;MQRankSum=0.494;QD=16.60;ReadPosRankSum=-0.468;SOR=3.641   GT:AD:DP:GQ:PL  0/0:4,0:4:12:0,12,162   ./.     0/0:2,0:2:6:0,6,77      0/0:1,0:1:3:0,3,40      ./.     0/0:2,0:2:6:0,6,84      0/0:1,0:1:3:0,3,
MQ0=0;MQRankSum=-0.515;QD=15.79;ReadPosRankSum=1.417;SOR=1.436    GT:AD:DP:GQ:PL  0/0:4,0:4:12:0,12,166   ./.     0/0:2,0:2:6:0,6,78      0/0:1,0:1:3:0,3,42      ./.     0/0:2,0:2:6:0,6,81      0/0:1,0:1:3:0,3,
Q0=0;MQRankSum=-0.436;QD=5.58;ReadPosRankSum=0.046;SOR=0.358      GT:AD:DP:GQ:PL  0/0:4,0:4:12:0,12,165   ./.     0/0:2,0:2:6:0,6,77      0/0:1,0:1:3:0,3,42      0/0:1,0:1:3:0,3,24      0/0:1,0:1:3:0,3,42
3;MQ0=0;MQRankSum=-0.784;QD=30.30;ReadPosRankSum=0.094;SOR=0.578  GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,126     0/0:1,0:1:3:0,3,41      1/1:0,2:2:6:78,6,0      0/0:1,0:1:3:0,3,41      0/0:1,0:1:3:0,3,32      0/0:2,0:
Q0=0;MQRankSum=-0.029;QD=14.66;ReadPosRankSum=0.414;SOR=0.765     GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,125     0/0:1,0:1:3:0,3,42      0/0:2,0:2:6:0,6,75      0/0:1,0:1:3:0,3,42      0/0:1,0:1:3:0,3,38      0/0:2,0:
Q0=0;MQRankSum=-0.228;QD=16.10;ReadPosRankSum=-0.294;SOR=0.976    GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,126     0/0:1,0:1:3:0,3,42      0/0:2,0:2:6:0,6,79      0/0:1,0:1:3:0,3,42      0/0:1,0:1:3:0,3,35      0/0:2,0:
MQ0=0;MQRankSum=-1.229;QD=27.90;ReadPosRankSum=1.295;SOR=0.724    GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,125     0/0:2,0:2:6:0,6,80      0/0:2,0:2:6:0,6,75      0/0:1,0:1:3:0,3,42      0/0:1,0:1:3:0,3,35      0/1:1,1:
0=0;MQRankSum=-1.515;QD=22.99;ReadPosRankSum=-1.146;SOR=0.745     GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,120     0/0:2,0:2:6:0,6,79      0/0:1,0:1:3:0,3,35      0/0:1,0:1:3:0,3,37      0/0:1,0:1:3:0,3,38      0/0:2,0:
0=0;MQRankSum=1.935;QD=32.42;ReadPosRankSum=0.670;SOR=2.661       GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,82      0/0:3,0:3:9:0,9,119     0/0:1,0:1:3:0,3,35      0/0:1,0:1:3:0,3,41      0/0:1,0:1:3:0,3,27      0/0:1,0:
MQ0=0;MQRankSum=-4.131;QD=34.12;ReadPosRankSum=0.713;SOR=0.822    GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,76      0/0:3,0:3:9:0,9,121     ./.     1/1:0,1:1:3:39,3,0      0/0:1,0:1:3:0,3,42      1/1:0,1:1:3:39,3,0
Q0=0;MQRankSum=1.670;QD=19.47;ReadPosRankSum=-0.708;SOR=2.260     GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,39      0/0:3,0:3:9:0,9,108     ./.     0/0:1,0:1:3:0,3,39      0/0:1,0:1:3:0,3,41      0/0:1,0:1:3:0,3,37
MQ0=0;MQRankSum=0.669;QD=24.85;ReadPosRankSum=-1.566;SOR=0.399    GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,32      0/0:7,0:7:18:0,18,238   0/0:4,0:4:9:0,9,125     ./.     0/0:1,0:1:3:0,3,40      0/0:3,0:3:9:0,9,111
MQ0=0;MQRankSum=0.397;QD=8.26;ReadPosRankSum=2.901;SOR=1.902      GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,71      0/0:7,0:7:18:0,18,243   0/0:4,0:4:9:0,9,126     ./.     0/0:1,0:1:3:0,3,35      0/0:3,0:3:9:0,9,119
MQ0=0;MQRankSum=0.261;QD=13.59;ReadPosRankSum=-1.751;SOR=0.376    GT:AD:DP:GQ:PL  0/0:2,0:2:6:0,6,74      0/0:8,0:8:15:0,15,205   0/0:4,0:4:9:0,9,121     0/0:1,0:1:3:0,3,39      ./.     0/0:3,0:3:9:0,9,122
;MQ0=0;MQRankSum=0.000;QD=14.93;ReadPosRankSum=-2.200;SOR=2.794   GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,117     0/0:8,0:8:18:0,18,240   0/0:4,0:4:9:0,9,127     0/0:1,0:1:3:0,3,41      0/0:2,0:2:6:0,6,77      0/0:3,0:
```

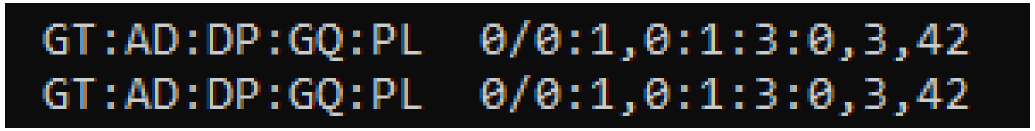
RECORD INFO

- Exact definitions included in the header!
- QUAL = Phred scaled (-10*log10(p)) probability that there is a polymorphism at this site.
- FILTER= names of any filters applied
- INFO= site level annotations, semi-colon separated. See header for more info.
- FORMAT= names and order of presentation of sample level variant information

https://gatk.broadinstitute.org/hc/en-us/articles/360035531692

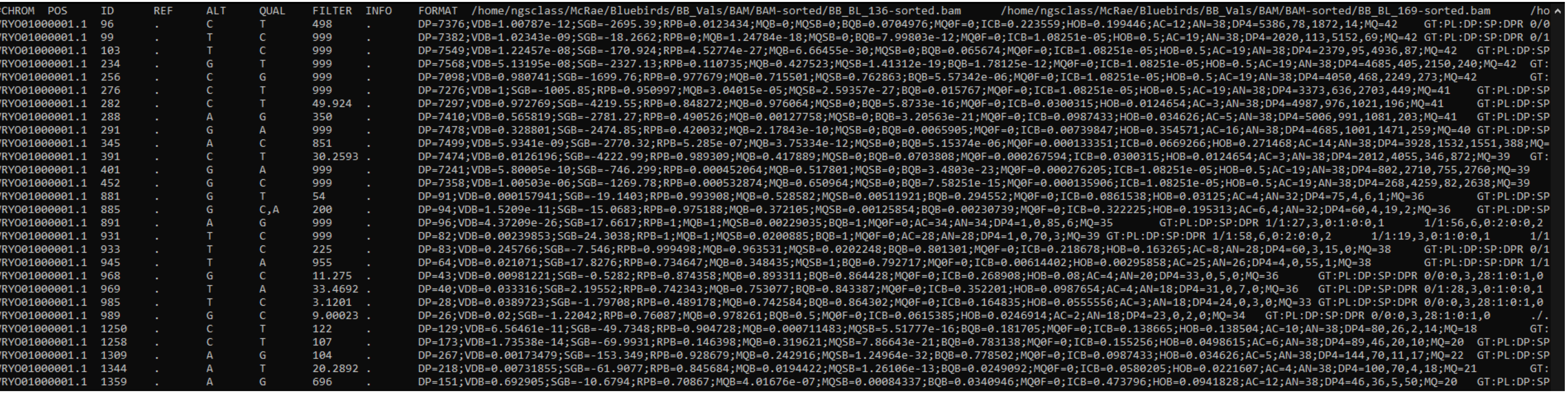
RECORD INFO

- FORMAT: GT:AD:DP:GQ:PL
- GT = genotype,
  - 0/0 homozygous for reference allele
  - 0/1 heterozygous reference/alternate
  - ... | indicates a phased genotype
- AD=Depth per allele
- DP= Depth
- GQ=Genotype Quality
- PL=Phred-scaled genotype likelihood of each genotype (0/0, 0/1, 1/1)



```
GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,42
GT:AD:DP:GQ:PL  0/0:1,0:1:3:0,3,42
```

VCF FROM SAMTOOLS MPILEUP



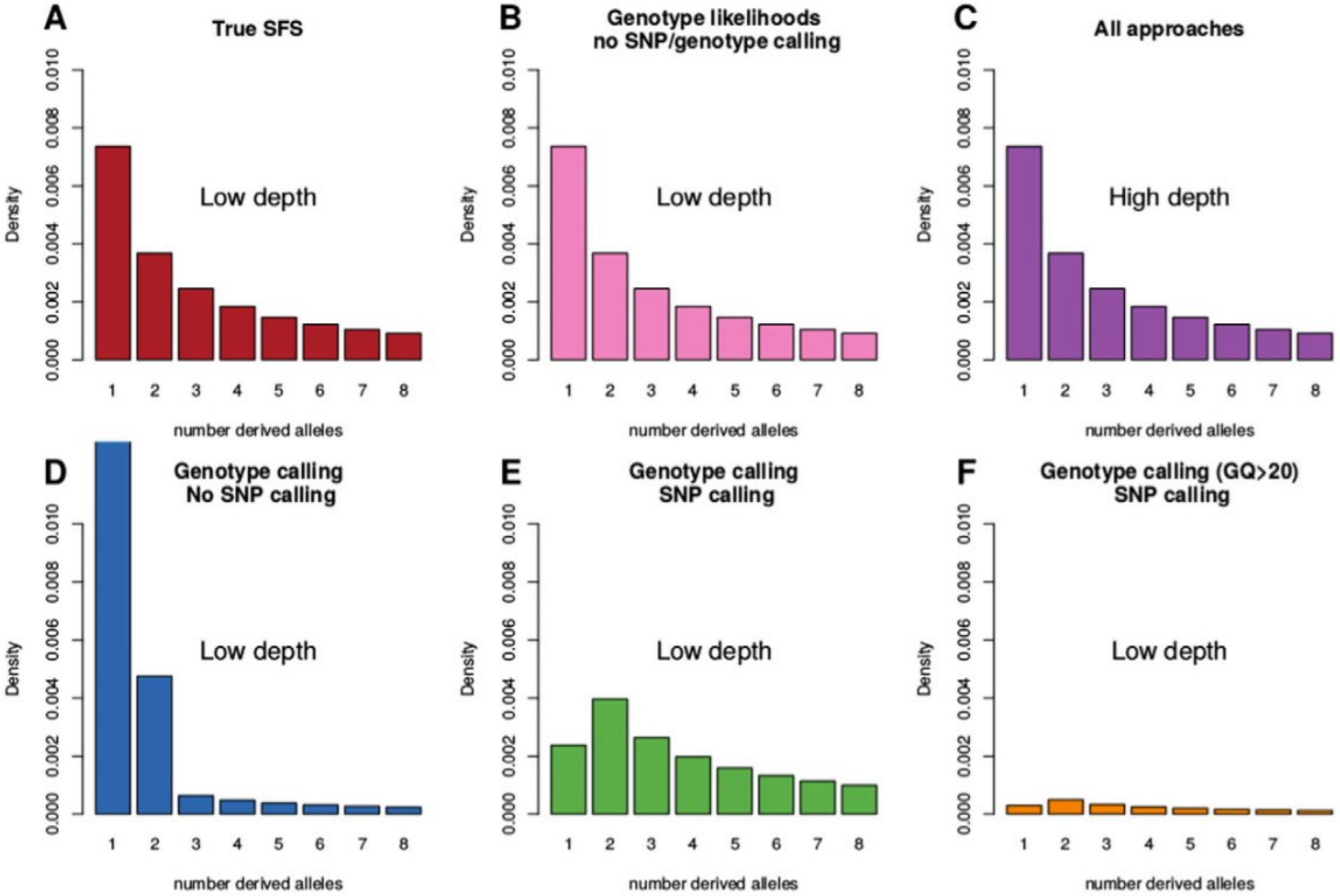
```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  /home/ngsclass/McRae/Bluebirds/BB_Vals/BAM/BAM-sorted/BB_BL_136-sorted.bam      /home/ngsclass/McRae/Bluebirds/BB_Vals/BAM/BAM-sorted/BB_BL_169-sorted.bam      /ho
VRYO01000001.1  96      .       C       T       498     .       DP=7376;VDB=1.00787e-12;SGB=-2695.39;RPB=0.0123434;MQB=0;MQSB=0;BQB=0.0704976;MQ0F=0;ICB=0.223559;HOB=0.199446;AC=12;AN=38;DP4=5386,78,1872,14;MQ=42    GT:PL:DP:SP:DPR 0/0
VRYO01000001.1  99      .       T       C       999     .       DP=7382;VDB=1.02343e-09;SGB=-18.2662;RPB=0;MQB=1.24784e-18;MQSB=0;BQB=7.99803e-12;MQ0F=0;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=2020,113,5152,69;MQ=42 GT:PL:DP:SP:DPR 0/1
VRYO01000001.1  103     .       T       C       999     .       DP=7549;VDB=1.22457e-08;SGB=-170.924;RPB=4.52774e-27;MQB=6.66455e-30;MQSB=0;BQB=0.065674;MQ0F=0;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=2379,95,4936,87;MQ=42   GT:PL:DP:SP
VRYO01000001.1  234     .       G       T       999     .       DP=7568;VDB=5.13195e-08;SGB=-2327.13;RPB=0.110735;MQB=0.427523;MQSB=1.41312e-19;BQB=1.78125e-12;MQ0F=0;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=4685,405,2150,240;MQ=42  GT:
VRYO01000001.1  256     .       C       G       999     .       DP=7098;VDB=0.980741;SGB=-1699.76;RPB=0.977679;MQB=0.715501;MQSB=0.762863;BQB=5.57342e-06;MQ0F=0;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=4050,468,2249,273;MQ=42        GT:
VRYO01000001.1  276     .       C       T       999     .       DP=7276;VDB=1;SGB=-1005.85;RPB=0.950997;MQB=3.04015e-05;MQSB=2.59357e-27;BQB=0.015767;MQ0F=0;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=3373,636,2703,449;MQ=41     GT:PL:DP:SP
VRYO01000001.1  282     .       C       T       49.924  .       DP=7297;VDB=0.972769;SGB=-4219.55;RPB=0.848272;MQB=0.976064;MQSB=0;BQB=5.8733e-16;MQ0F=0;ICB=0.0300315;HOB=0.0124654;AC=3;AN=38;DP4=4987,976,1021,196;MQ=41      GT:PL:DP:SP
VRYO01000001.1  288     .       A       G       350     .       DP=7410;VDB=0.565819;SGB=-2781.27;RPB=0.490526;MQB=0.00127758;MQSB=0;BQB=3.20563e-21;MQ0F=0;ICB=0.0987433;HOB=0.034626;AC=5;AN=38;DP4=5006,991,1081,203;MQ=41   GT:PL:DP:SP
VRYO01000001.1  291     .       G       A       999     .       DP=7478;VDB=0.328801;SGB=-2474.85;RPB=0.420032;MQB=2.17843e-10;MQSB=0;BQB=0.0065905;MQ0F=0;ICB=0.00739847;HOB=0.354571;AC=16;AN=38;DP4=4685,1001,1471,259;MQ=40 GT:PL:DP:SP
VRYO01000001.1  345     .       A       C       851     .       DP=7499;VDB=5.9341e-09;SGB=-2770.32;RPB=5.285e-07;MQB=3.75334e-12;MQSB=0;BQB=5.15374e-06;MQ0F=0.000133351;ICB=0.0669266;HOB=0.271468;AC=14;AN=38;DP4=3928,1532,1551,388;MQ=
VRYO01000001.1  391     .       C       T       30.2593 .       DP=7474;VDB=0.0126196;SGB=-4222.99;RPB=0.989309;MQB=0.417889;MQSB=0;BQB=0.0703808;MQ0F=0.000267594;ICB=0.0300315;HOB=0.0124654;AC=3;AN=38;DP4=2012,4055,346,872;MQ=39   GT:
VRYO01000001.1  401     .       G       A       999     .       DP=7241;VDB=5.80005e-10;SGB=-746.299;RPB=0.000452064;MQB=0.517801;MQSB=0;BQB=3.4803e-23;MQ0F=0.000276205;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=802,2710,755,2760;MQ=39
VRYO01000001.1  452     .       G       C       999     .       DP=7358;VDB=1.00503e-06;SGB=-1269.78;RPB=0.000532874;MQB=0.650964;MQSB=0;BQB=7.58251e-15;MQ0F=0.000135906;ICB=1.08251e-05;HOB=0.5;AC=19;AN=38;DP4=268,4259,82,2638;MQ=39
VRYO01000001.1  881     .       G       T       54      .       DP=91;VDB=0.000157941;SGB=-19.1403;RPB=0.993908;MQB=0.528582;MQSB=0.00511921;BQB=0.294552;MQ0F=0;ICB=0.0861538;HOB=0.03125;AC=4;AN=32;DP4=75,4,6,1;MQ=36        GT:PL:DP:SP
VRYO01000001.1  885     .       G       C,A     200     .       DP=94;VDB=1.5209e-11;SGB=-15.0683;RPB=0.975188;MQB=0.372105;MQSB=0.00125854;BQB=0.00230739;MQ0F=0;ICB=0.322225;HOB=0.195313;AC=6,4;AN=32;DP4=60,4,19,2;MQ=36    GT:PL:DP:SP
VRYO01000001.1  891     .       A       G       999     .       DP=96;VDB=4.37209e-26;SGB=17.6617;RPB=1;MQB=1;MQSB=0.00229035;BQB=1;MQ0F=0;AC=34;AN=34;DP4=1,0,85,6;MQ=35       GT:PL:DP:SP:DPR 1/1:27,3,0:1:0:0,1      1/1:56,6,0:2:0:0,2
VRYO01000001.1  931     .       T       C       999     .       DP=82;VDB=0.00239853;SGB=24.3038;RPB=1;MQB=1;MQSB=0.0200885;BQB=1;MQ0F=0;AC=28;AN=28;DP4=1,0,70,3;MQ=39 GT:PL:DP:SP:DPR 1/1:58,6,0:2:0:0,2      1/1:19,3,0:1:0:0,1      1/1
VRYO01000001.1  933     .       T       C       225     .       DP=83;VDB=0.245766;SGB=-7.546;RPB=0.999498;MQB=0.963531;MQSB=0.0202248;BQB=0.801301;MQ0F=0;ICB=0.218678;HOB=0.163265;AC=8;AN=28;DP4=60,3,15,0;MQ=38     GT:PL:DP:SP:DPR 0/1
VRYO01000001.1  945     .       T       A       955     .       DP=64;VDB=0.021071;SGB=17.8276;RPB=0.734647;MQB=0.348435;MQSB=1;BQB=0.792717;MQ0F=0;ICB=0.00614402;HOB=0.00295858;AC=25;AN=26;DP4=4,0,55,1;MQ=38     GT:PL:DP:SP:DPR 1/1
VRYO01000001.1  968     .       G       C       11.275  .       DP=43;VDB=0.00981221;SGB=-0.5282;RPB=0.874358;MQB=0.893311;BQB=0.864428;MQ0F=0;ICB=0.268908;HOB=0.08;AC=4;AN=20;DP4=33,0,5,0;MQ=36      GT:PL:DP:SP:DPR 0/0:0,3,28:1:0:1,0
VRYO01000001.1  969     .       T       A       33.4692 .       DP=40;VDB=0.033316;SGB=2.19552;RPB=0.742343;MQB=0.753077;BQB=0.843387;MQ0F=0;ICB=0.352201;HOB=0.0987654;AC=4;AN=18;DP4=31,0,7,0;MQ=36   GT:PL:DP:SP:DPR 0/1:28,3,0:1:0:0,1
VRYO01000001.1  985     .       T       C       3.1201  .       DP=28;VDB=0.0389723;SGB=-1.79708;RPB=0.489178;MQB=0.742584;BQB=0.864302;MQ0F=0;ICB=0.164835;HOB=0.0555556;AC=3;AN=18;DP4=24,0,3,0;MQ=33 GT:PL:DP:SP:DPR 0/0:0,3,28:1:0:1,0
VRYO01000001.1  989     .       G       C       9.00023 .       DP=26;VDB=0.02;SGB=-1.22042;RPB=0.76087;MQB=0.978261;BQB=0.5;MQ0F=0;ICB=0.0615385;HOB=0.0246914;AC=2;AN=18;DP4=23,0,2,0;MQ=34   GT:PL:DP:SP:DPR 0/0:0,3,28:1:0:1,0      ./.
VRYO01000001.1  1250    .       C       T       122     .       DP=129;VDB=6.56461e-11;SGB=-49.7348;RPB=0.904728;MQB=0.000711483;MQSB=5.51777e-16;BQB=0.181705;MQ0F=0;ICB=0.138665;HOB=0.138504;AC=10;AN=38;DP4=80,26,2,14;MQ=18        GT:
VRYO01000001.1  1258    .       C       T       107     .       DP=173;VDB=1.73538e-14;SGB=-69.9931;RPB=0.146398;MQB=0.319621;MQSB=7.86643e-21;BQB=0.783138;MQ0F=0;ICB=0.155256;HOB=0.0498615;AC=6;AN=38;DP4=89,46,20,10;MQ=20  GT:PL:DP:SP
VRYO01000001.1  1309    .       A       G       104     .       DP=267;VDB=0.00173479;SGB=-153.349;RPB=0.928679;MQB=0.242916;MQSB=1.24964e-32;BQB=0.778502;MQ0F=0;ICB=0.0987433;HOB=0.034626;AC=5;AN=38;DP4=144,70,11,17;MQ=22  GT:PL:DP:SP
VRYO01000001.1  1344    .       A       T       20.2892 .       DP=218;VDB=0.00731855;SGB=-61.9077;RPB=0.845684;MQB=0.0194422;MQSB=1.26106e-13;BQB=0.0249092;MQ0F=0;ICB=0.0580205;HOB=0.0221607;AC=4;AN=38;DP4=100,70,4,18;MQ=21        GT:
VRYO01000001.1  1359    .       A       G       696     .       DP=151;VDB=0.692905;SGB=-10.6794;RPB=0.70867;MQB=4.01676e-07;MQSB=0.00084337;BQB=0.0340946;MQ0F=0;ICB=0.473796;HOB=0.0941828;AC=12;AN=38;DP4=46,36,5,50;MQ=20   GT:PL:DP:SP
```

SNP CALLING SOFTWARE

- samtools mpileup
- FreeBayes
- Genome Analysis Toolkit (GATK)
- Angsd (based on Genotype Likelihoods)

…

FILTERING AND DEPTH MATTERS



From da Fonseca et al (2016)

FILTERING

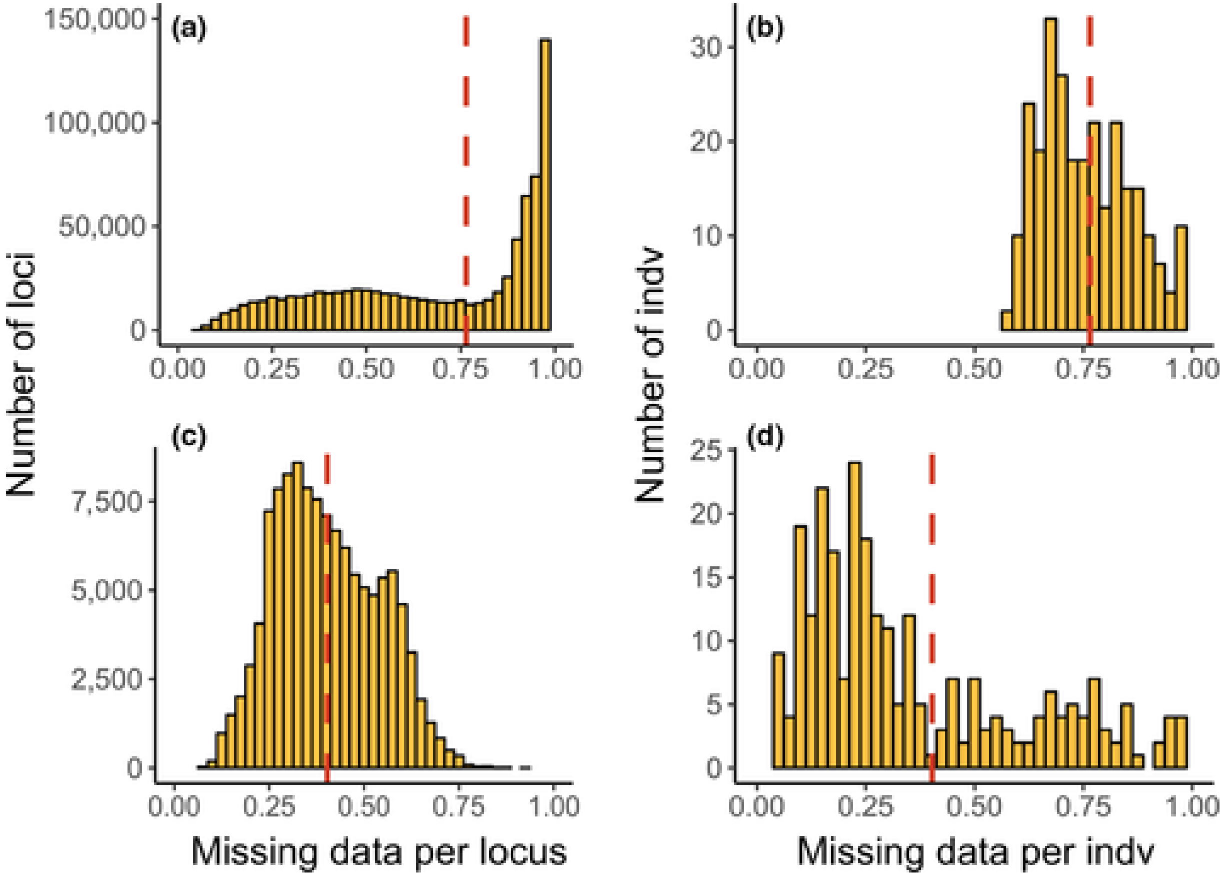
- Will depend on your sampling design and your questions
- In general:
  - High quality genotypes – depth, allele balance etc
  - Low level of missing data

PLOTTING THE DATA

- Before

- Apply Filter

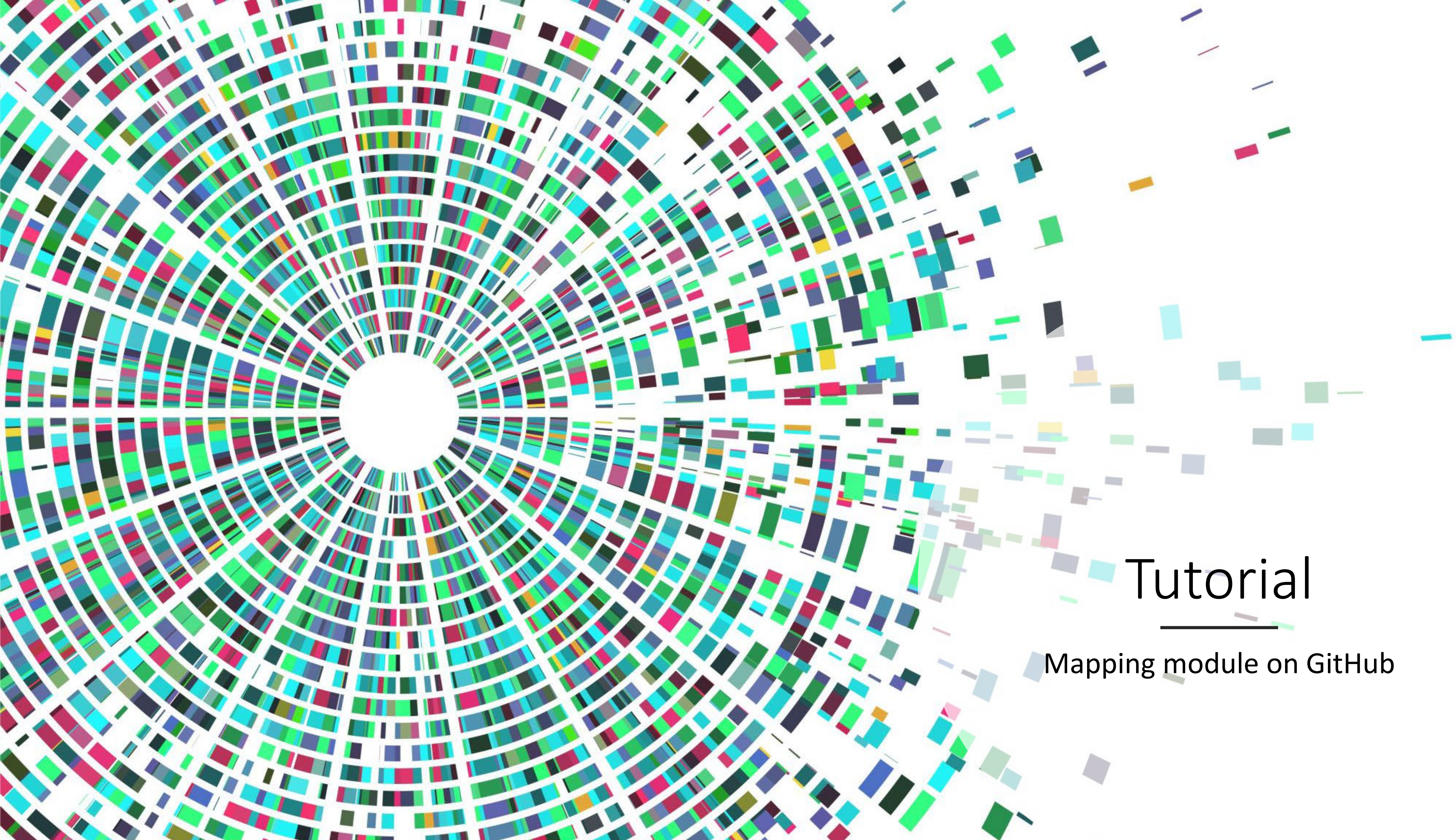
- After



O’Leary et al 2018 Mol Ecol https://doi.org/10.1111/mec.14792

FURTHER READING & REFERENCES

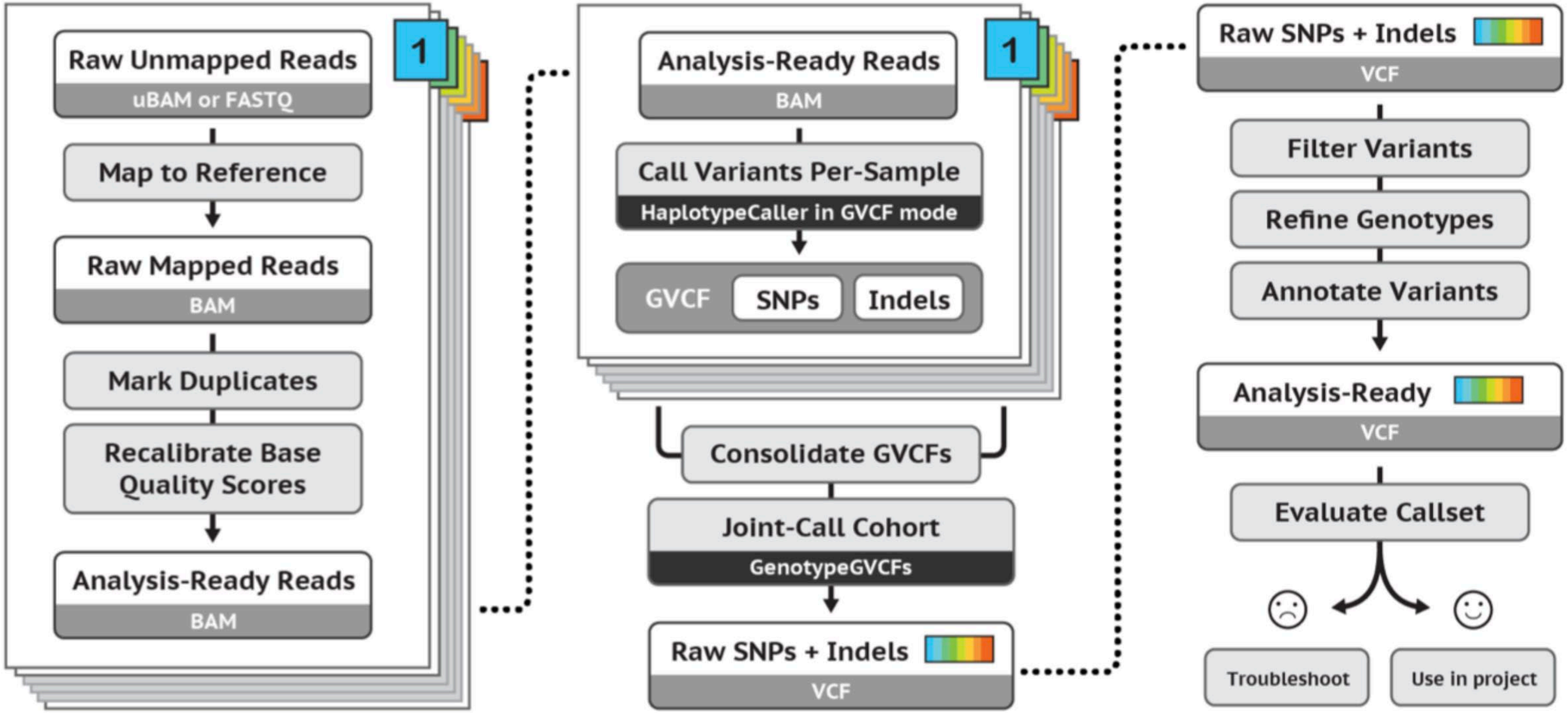
- da Fonseca et al (2016) Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics.*
- Korneliussen et al (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics.*
- Nielsen et al (2012) SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*
- Nielsen et al (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics.*
- O’Leary et al (2018). These aren’t the loci you’e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology.*
- Bhatia et al. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Research.*
- Linck and Battey (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *Moleuclar Ecology Resources.*
- Broad Institute (2021) GATK Best Practices Workflow: Germline short variant discovery. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-
- Salter & Faircloth (2020) Running GATK in Parallel. http://protocols.faircloth-lab.org/en/latest/protocols-computer/analysis/analysis-gatk-parallel.html



Tutorial

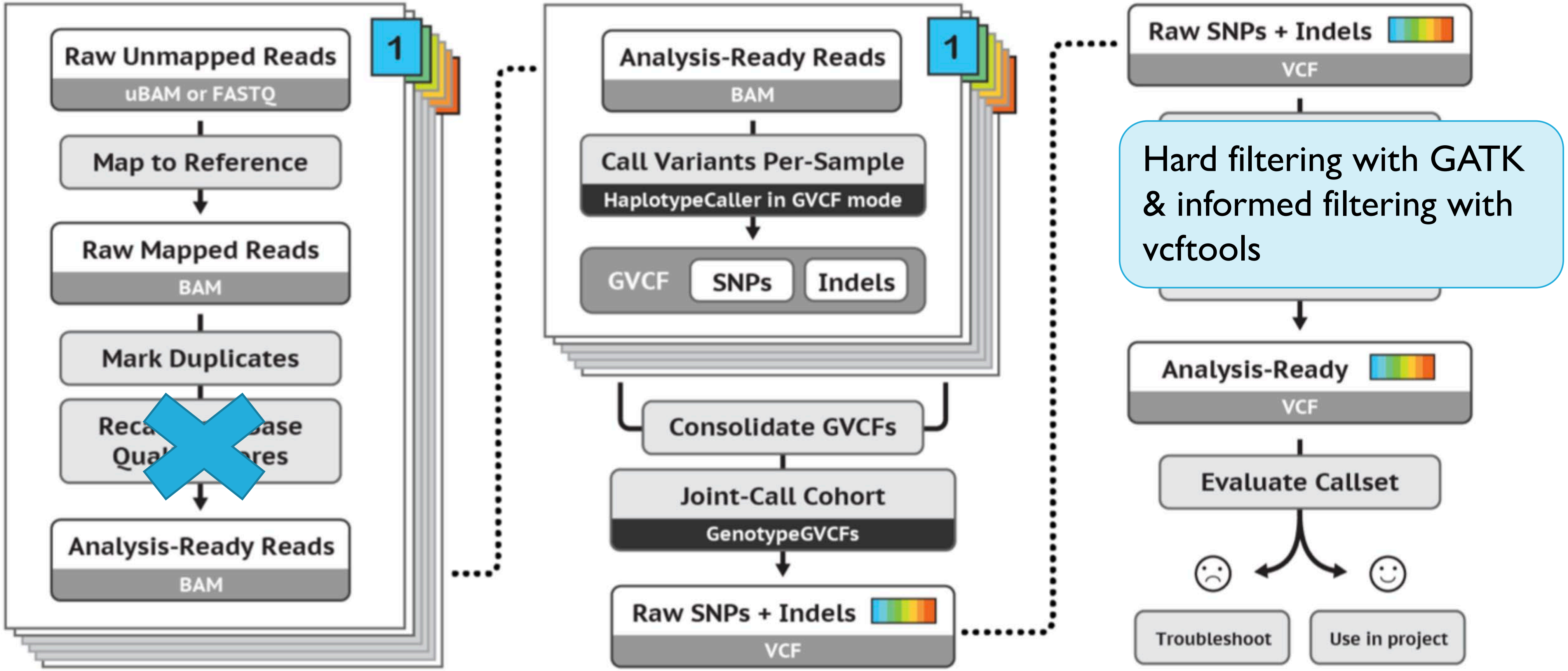
Mapping module on GitHub

GATK GERMLINE BEST PRACTICES



https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

GATK WITH NON-MODEL ORGANISM



Hard filtering with GATK & informed filtering with vcftools

https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

A NOTE ABOUT READ GROUPS

- Set at mapping step or afterwards using PicardTools
- Fastq files can be split by lane
- Enables multiple libraries, flowcells, and lanes to be run per “individual” – SM.
- GATK can take into account some of the batch effects.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890671-Read-groups