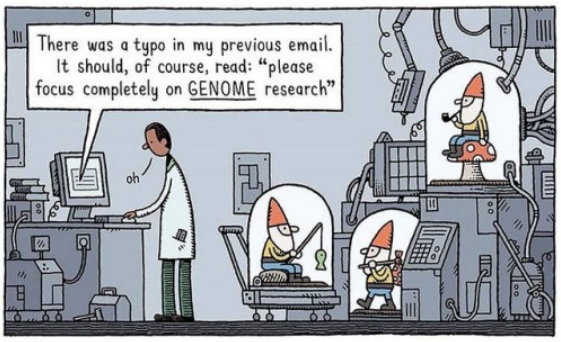
One SNP, two SNPs; many SNPs, few SNPs

How do I filter my data, and when do I stop?

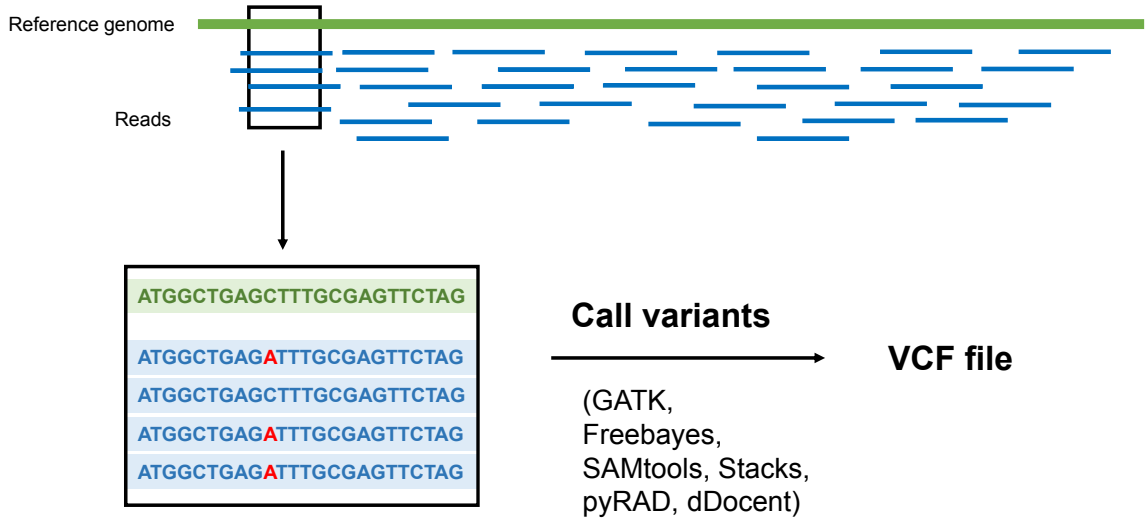
Maddie James
PhD candidate
Ortiz-Barrientos laboratory

Overview

- SNP calling
- What is a VCF file?
- Where do errors come from?
- Data filtering
- Recommendations



SNP calling


Reference genome
Reads
ATGGCTGAGCTTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
ATGGCTGAGCTTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
Call variants
(GATK,
Freebayes,
SAMtools, Stacks,
pyRAD, dDocent)
VCF file

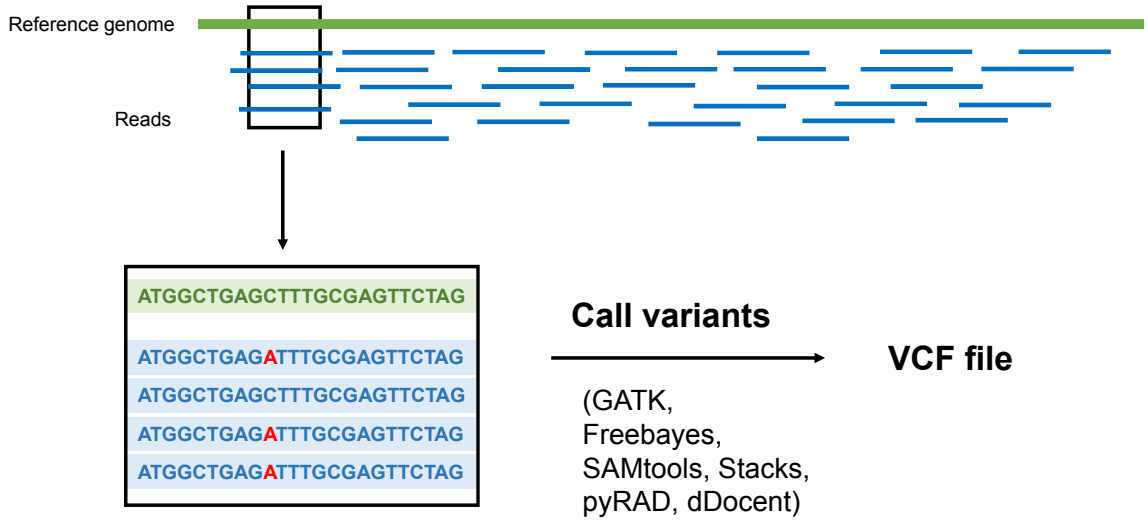
A note on calling SNPs

- Joint calling of SNPs
  - Uses information across all samples to call a SNP (good for samples with low coverage)
  - Assumes all samples are genetically similar
  - Typically, you joint call on “cohorts” of samples (aka populations, species)
  - SNPs from these cohorts are combined into one file
- The default settings of variant callers only output *variant* sites
  - This can be a problem when combining jointly-called cohorts
  - You can’t distinguish between sites that are invariant and sites with missing data
- When calling SNPs on cohorts, output variant and invariant sites (so you know what is missing vs what is invariant)

FreeBayes: --report-monomorphic

- Link: https://gatkforums.broadinstitute.org/gatk/discussion/3686/why-do-joint-calling-rather-than-single-sample-calling-retired

SNP calling


Reference genome
Reads
ATGGCTGAGCTTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
ATGGCTGAGCTTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
ATGGCTGAGATTTGCGAGTTCTAG
Call variants
(GATK,
Freebayes,
SAMtools, Stacks,
pyRAD, dDocent)
VCF file

What is a VCF file and what does it all mean?

```
##fileformat=VCFv4.2
##fileDate=20180911
##source=freeBayes v1.2.0
##reference=/home/uqmjame4/data/PacBio_alignment/Senecio.contigs.fasta
##contig=<ID=tig00000007,length=171107>
##contig=<ID=tig00000008,length=59952>
##contig=<ID=tig00000011,length=37475>
##contig=<ID=tig00000013,length=542510>
##contig=<ID=tig00000014,length=30717>
##contig=<ID=tig00000020,length=61158>
##contig=<ID=tig00000028,length=59932>
##contig=<ID=tig00000034,length=19636>
##contig=<ID=tig00000007,length=171107>
##contig=<ID=tig00000008,length=59952>
##contig=<ID=tig00000011,length=37475>
##contig=<ID=tig00000013,length=542510>
##contig=<ID=tig00000014,length=30717>
##contig=<ID=tig00000020,length=61158>
##contig=<ID=tig00000028,length=59932>
##contig=<ID=tig00000034,length=19636>
##contig=<ID=tig00000041,length=32773>
##contig=<ID=tig00000044,length=14427>
##contig=<ID=tig00000065,length=41833>
##contig=<ID=tig00000075,length=38752>
##contig=<ID=tig00000077,length=13195>
##contig=<ID=tig00000078,length=63063>
##contig=<ID=tig00000082,length=32873>
```

Information about the VCF file

List of contigs

Some extra detail:
https://samtools.github.io/hts-specs/VCFv4.2.pdf
https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference
##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observin
##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation betwee
##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between
##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.
##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to eas
##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
##INFO=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record.">
```

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference
##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observin
##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation betwee
##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between
##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.
##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to eas
##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
##INFO=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record.">
```

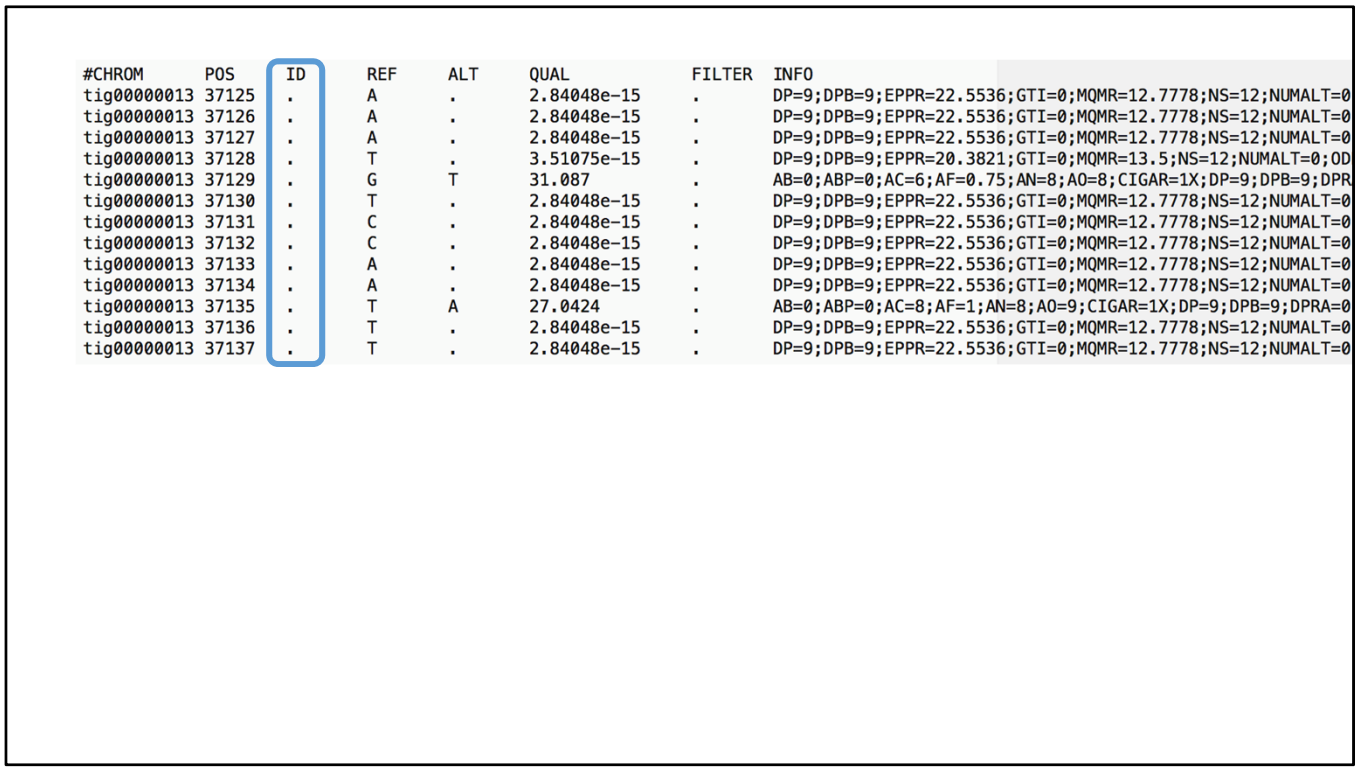
```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference
##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observin
##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation betwee
##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between
##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.
##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to eas
##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
##INFO=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record.">
```

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference
##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observin
##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation betwee
##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between
##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.
##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to eas
##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
##INFO=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record.">
```

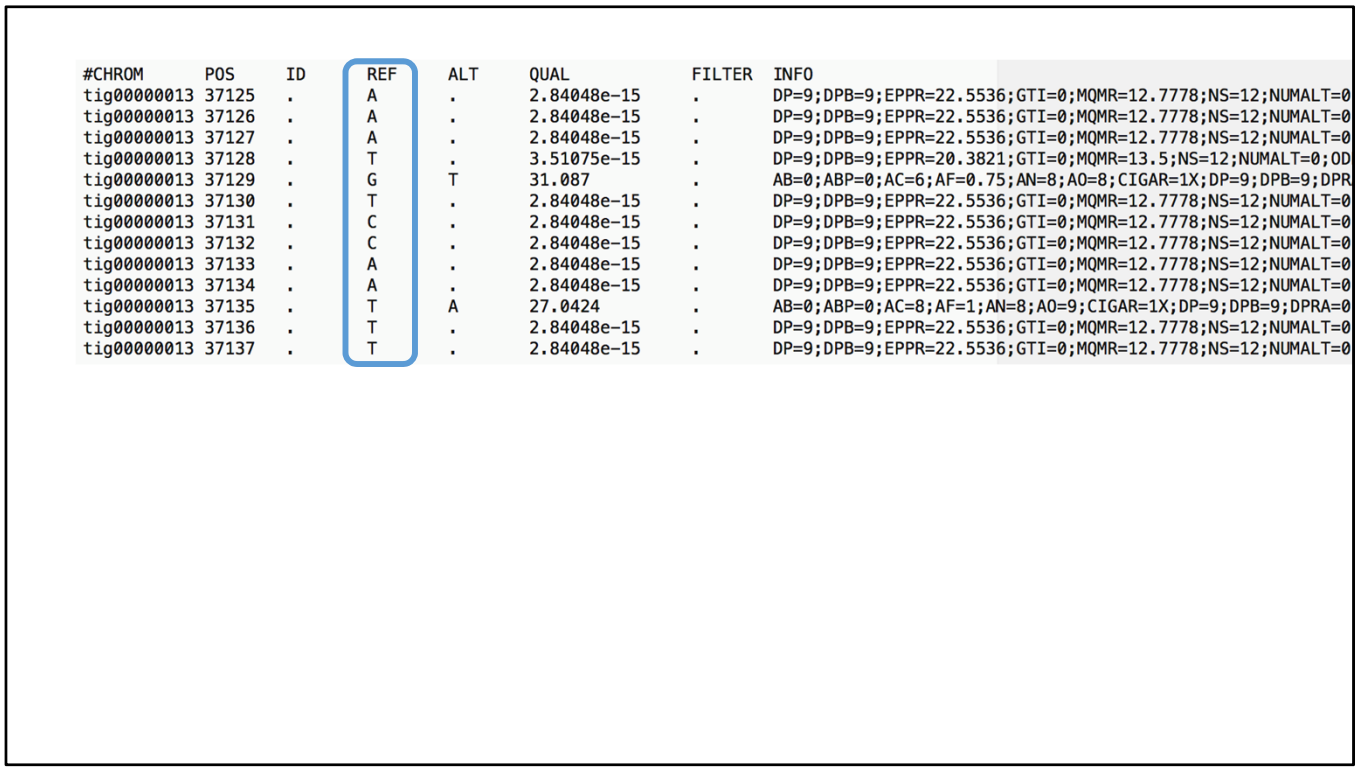
```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of obser
##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference
##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observin
##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation betwee
##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between
##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of obs
##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.
##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to eas
##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
##INFO=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record.">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
tig00000013	37125	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37126	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37127	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37128	.	T	.	3.51075e-15	.	DP=9;DPB=9;EPPR=20.3821;GTI=0;MQMR=13.5;NS=12;NUMALT=0;OD
tig00000013	37129	.	G	T	31.087	.	AB=0;ABP=0;AC=6;AF=0.75;AN=8;AO=8;CIGAR=1X;DP=9;DPB=9;DPR
tig00000013	37130	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37131	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37132	.	C	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37133	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37134	.	A	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37135	.	T	A	27.0424	.	AB=0;ABP=0;AC=8;AF=1;AN=8;AO=9;CIGAR=1X;DP=9;DPB=9;DPRA=0
tig00000013	37136	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0
tig00000013	37137	.	T	.	2.84048e-15	.	DP=9;DPB=9;EPPR=22.5536;GTI=0;MQMR=12.7778;NS=12;NUMALT=0

FORMAT	H15-55	H15-67	H15-60	H15-69	H15-76	H15-58	H15-72
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:68:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:150.924:1:1:1:72:.:.:0	.	0/0:150.924:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	1/1:31.0715:1:0,1:0:0:1:68:-1.3,-0.30103,0	.	1/1:31.0715:2:0,
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	1/1:26.7297:1:0,1:0:0:1:68:-1.3,-0.30103,0	.	1/1:26.8832:2:0,
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:68:.:.:0	.	0/0:151.844:2:2:
GT:GQ:DP:AD:RO:QR:AO:QA:GL	.	.	.	.	0/0:151.844:1:1:1:72:.:.:0	.	0/0:151.844:2:2:

We can use this information in our VCF file to filter our SNPs. We want distinguish between SNPs that are true variants vs SNPs that are errors.

Image sources:
https://www.yourgenome.org/facts/what-is-pcr-polymerase-chain-reaction
http://www.well.ox.ac.uk/ogc/sequencing-quality-monitoring-run/

See **dDocent** filtering: http://ddocent.com/filtering/

Link:
https://software.broadinstitute.org/gatk/documentation/article.php?id=4860

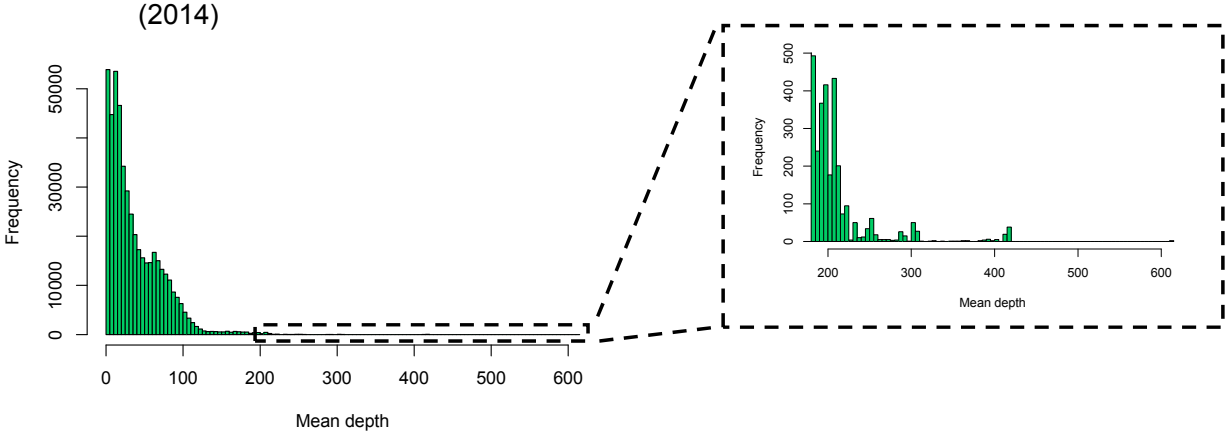
Minimum depth: Depends if you are joint calling or not. If you are calling SNPs for each sample by itself, you may want 10 reads as a minimum. If you are jointly calling (which uses information across all samples), you might be ok with 3 reads. But if you have really high coverage you might want to have 20 reads as a minimum.

Missing data per site: If you filter stringently here, and you have a bunch of low quality individuals (i.e. they are the ones contributing the most to the missing data), you will remove many sites. If we first have a relaxed missing data filter, and then remove low quality individuals, you will end up with more SNPs at the end.

Iterative filtering:
See O'Leary et al. (2018) "These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists"

Maximum mean depth

- We want to remove paralogues
  - Generally, the mean read depth per locus should be approximately normally distributed
  - 90th quantile; two times the mode (Willis et al. 2017); d+3*sqrt(d), d=mean depth; Li (2014)

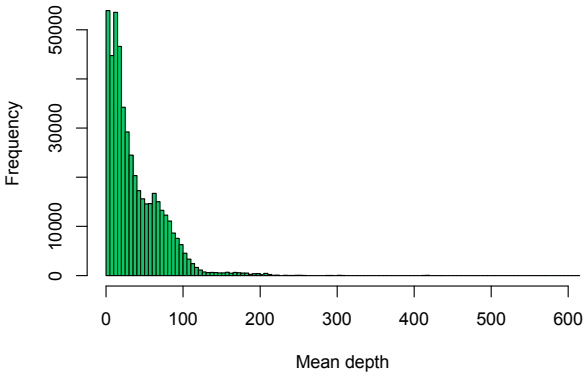


Willis et al. (2017) "Haplotyping RAD loci: An efficient method to filter paralogs and account for physical linkage"

Li (2014) "Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples"

Maximum mean depth

- We want to remove paralogues
  - Generally, the mean read depth per locus should be approximately normally distributed
  - 90th quantile; two times the mode (Willis et al. 2017); d+3*sqrt(d), d=mean depth; Li (2014)



Create a list of mean depth per site:
vcftools --vcf input_file.vcf --site-mean-depth --out mean_depth

Results are stored in:
mean_depth.ldepth.mean

Filter for maximum mean depth:
vcftools --vcf input_file.vcf --max-meanDP 120 --recode --recode-INFO-all --out output_file.vcf

Minimum mean depth

- We want to be confident with out SNP calls
  - Commonly 20-30

```
vcftools --vcf input_file.vcf --min-meanDP 20 --recode --recode-INFO-all --out output_file
```

Depth and quality scores

See: http://ddocent.com/filtering/

- High coverage can lead to inflated quality scores
  - Removal of variants with high quality scores and high depth
  - Typical to remove these sites before the mean depth filtering

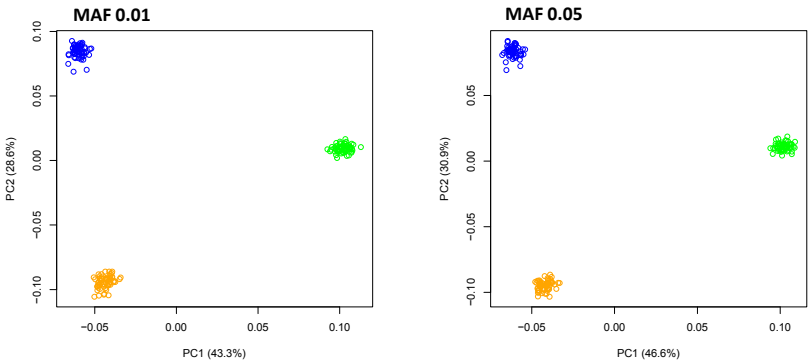
Mapping quality

- We want to filter out sites with reads that haven't mapped well to the reference genome
  - Typically 20-30

```
vcffilter -f "MQ > 30" input_file.vcf > output_file.vcf
(vcffilter is within vcflib)
```

Minor allele frequency

- Typically minor allele frequency of 0.01 or 0.05, or minor allele count of 1
  - Depends on what statistic you are measuring
  - Are you interested in rare variants? Singletons might be errors
  - Do some PCAs - does the structure change depending on the minor allele frequency?



~10,000 SNPs, three closely related populations, neutral loci

Minor allele frequency:
vcftools --vcf input_file.vcf --maf 0.01--recode --recode-INFO-all --out output_file

Minor allele count:
vcftools --vcf input_file.vcf --mac 1--recode --recode-INFO-all --out output_file

Are your results robust to differences in minor allele frequencies, or do you detect different structure depending on the minor allele frequency?

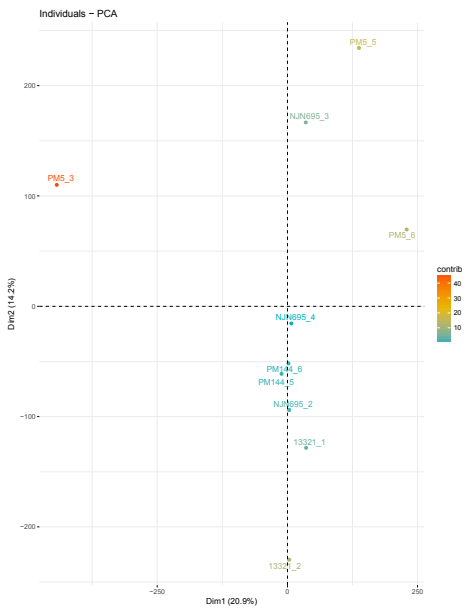
See Linck & Battey “Minor allele frequency thresholds strongly affect population structure inference with genomic datasets”

Missing data

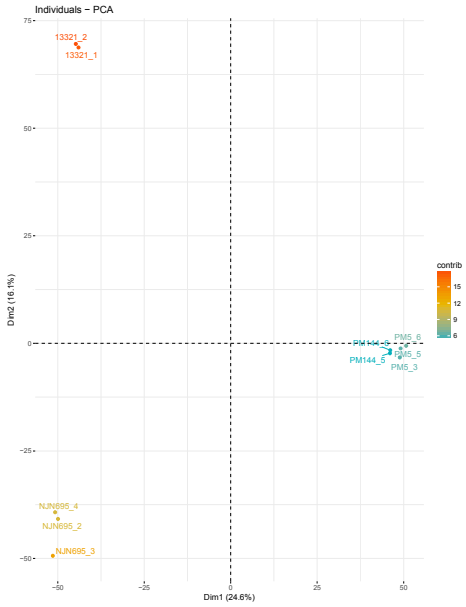
- Overall % of missing data per locus
  - Typically 0-20%
- Population specific missing data
  - Do you need each SNP to be sequenced in every population?
- Consider how each program deals with missing data!

Filter for overall 20% missing data:
vcftools --vcf input_file.vcf --max-missing 0.8 --recode --recode-INFO-all --out output_file

You might consider imputing missing data.



Missing data

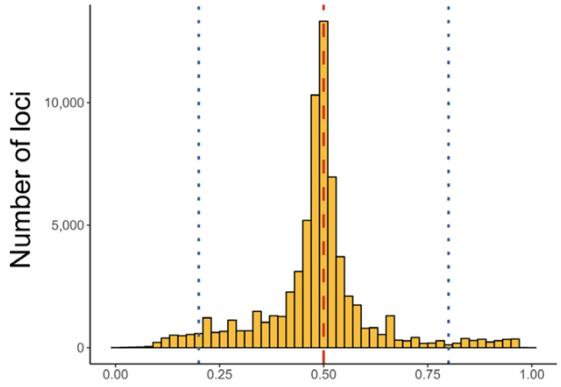


No missing data

PCAs by Alicia Toon

Allele balance

- For heterozygous sites, the ratio of the number of reads for the reference allele compared to the number of reads for the alternative allele

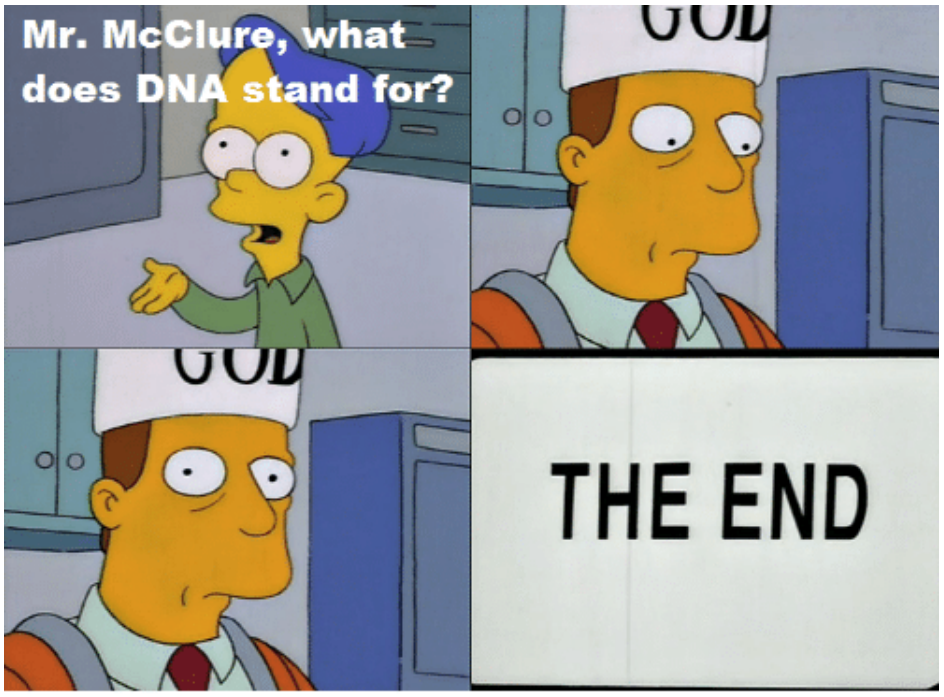


Source: O'Leary et al. (2018)

- We expect close to 0.5
- Typically remove sited with allele balance less than 0.25 and greater than 0.75

```
vcffilter -s -f "AB > 0.25 & AB < 0.75 | AB < 0.01" input_file.vcf > output_file.vcf
```

Allele balance not close to 0.5 could indicate false homozygotes (i.e. errors!)


Mr. McClure, what does DNA stand for?
GOD
GOD
THE END