

# **Variant calling and filtering**

**SNPs and indels**

**Dr Johanna Rhodes**

# Learning outcomes

- **Calling single nucleotide polymorphisms (SNPs) and insertions/deletions (indels)**
  - tools
  - considerations
- **Understanding vcf format**
- **Why we need to filter variants**
- **How to filter variants**
  - obtain high-confidence SNPs
  - commands
- **Quality control of filtered SNPs**
- **Mapping SNPs to genes**

# Identifying variation within genomes

- Identifying variants is important for evolution studies, interring resistance mutations and in population genomics studies
  - Reads aligned to a reference genome enables identification of differences between individuals
- SNPs are the most abundant variants, along with insertion/deletions
- Widely used software for variant calling includes GATK and Freebayes
  - Can call variants individually or across multiple samples
  - Determine genotypes e.g. at each position, determine whether variant is homozygous reference, heterozygous or homozygous alternate
- Creates variant call format (vcf) file

# Variant call format (VCF)

## The details

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002
NA000003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51
1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60
0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2
1/1:40:3
```

# Calling high-quality variants

## Commands using GATK

```
gatk HaplotypeCaller -R reference.fa -I isolate_name.post_recal_reads.bam -O  
isolate_name.raw_variants_recal.vcf -ERC GVCF --pcr-indel-model NONE -ploidy 1  
-stand-call-conf 30 -mbq 20 -A QualByDepth
```

```
gatk GenotypeGVCFs -R reference.fa -V isolate_name.raw_variants_recal.vcf -O  
isolate_name.genotyped_variants_recal.vcf
```

# Calling variants

## Considerations

- Repetitive regions
  - Most fungi have 1-25% repetitive DNA content and genome size is correlated with number of transposon families hosted within the genome
  - Difficult to call SNPs in repetitive regions
    - lower quality scores in repetitive regions, so variant calls may be false positives
  - RepeatMasker to mask regions of repeats in the reference genome
    - Include in HaplotypeCaller using the -XL parameter
    - Covered tomorrow

# Variant filtering

# Filtering variants

## What is the need?

- Initial variant calling is generally a very rough approximation, and will incorrectly identify many loci as SNPs or indels
- The INFO and FORMAT fields in the vcf contains a lot of information about each site in the genome, the reads aligned there, and the quality of variant calls.
- The variant filtration algorithm will use this information to filter out uninformative positions.



# Filtering variants

## Commands in GATK

- First, we separate the vcf into SNPs and indels (you can merge them back together after filtering if preferred)

```
gatk SelectVariants -R reference.fa -V  
isolate_name.genotyped_variants_recal.vcf -O isolate_name.raw_snps_recal.vcf --  
select-type-to-include SNP -select 'vc.getGenotype("WGS").getAD().1*1.0 /  
vc.getGenotype("WGS").getDP() > 0.90'
```

```
gatk SelectVariants -R reference.fa -V isolate_name.raw_variants_recal.vcf -O  
isolate_name.raw_indels_recal.vcf --select-type-to-include INDEL
```

# Filtering variants

## Commands in GATK

- To gain a list of high confidence SNPs we filter on mapping quality, depth of coverage, Fisher strand bias and balance of alleles.
- Any SNP that fulfils any one of these criteria is labelled as 'LowConf'. It is not removed
- We also include an additional filter on genotype quality
  - Note that we sort of did a filtering step earlier when selecting SNPs, where we filtered out SNPs that weren't present in at least 90% of mapped reads.

```
gatk VariantFiltration -R reference.fa -V isolate_name.raw_snps_recal.vcf -O  
isolate_name.filtered_snps_final.vcf -filter "QD < 2.0" --filter-name "LowConf" -filter  
"FS > 60.0" --filter-name "LowConf" -filter "MQ < 40.0" --filter-name "LowConf" -filter  
"MQRankSum < -12.5" --filter-name "LowConf" -filter "ReadPosRankSum < -8.0" --filter-  
name "LowConf" -filter "SOR > 4.0" --filter-name "LowConf" -filter "DP < 5" --filter-  
name "LowConf" -G-filter "GQ < 50" -G-filter-name "FILTER_GQ-50"
```

# Quality control

- Using the two files, 'isolate\_name.genotyped\_variants\_recal.vcf' and 'isolate\_name.filtered\_snps\_final.vcf', you want to count the number of lines in each, excluding the header
- `grep -v "#" isolate_name.genotyped_variants_recal.vcf | wc -l`
- `grep -v "#" isolate_name.filtered_snps_final.vcf | grep -v "LowConf" | grep -v "FILTER_GQ-50" | wc -l`
- The difference between these two numbers will give you an idea of quality
  - High % filtered out indicates something weird is going on - cross-check with coverage and mapping statistics
    - Hybrid/diploid
    - Not the same species as reference
    - Poor quality sequencing
  - Usually range can be 1-10% filtered out

# Other considerations

## Making pipelines

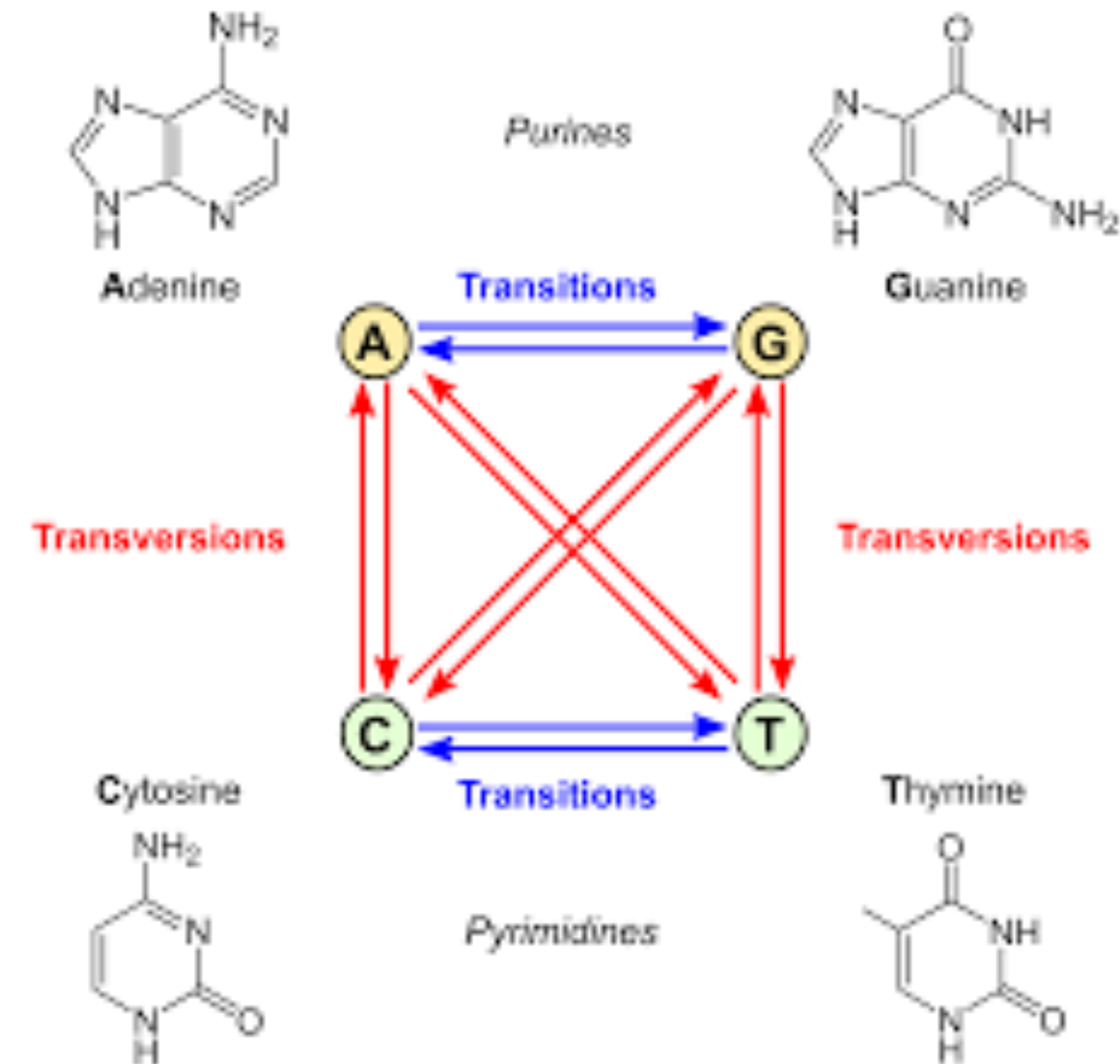
- Running these analyses (from raw read QC to filtering SNPs) one command at a time/one isolate at a time is inefficient
  - Time-consuming
  - Could receive many isolates in one go to analyse
- Automating the process as a 'pipeline'
  - All commands in a script
  - High Performance Computing (HPC)
    - job submission —> running pipeline in parallel
    - may have more processing power available than standard laptop/computer, which is useful for big/highly complex genomes
    - consider how to transfer your data - clients such as Cyberduck provide a user friendly GUI to scp

# Mapping SNPs to genes

# Nucleotide changes

## Transitions and transversions

- Nucleotide substitutions can take two forms:
  - transitions
    - purine-purine or pyrimidine-pyrimidine interchanges
  - transversions
    - purine-pyrimidine interchanges
- Transitions (G>A|C>T or A>G|T>C) occur more frequently in Ascomycota fungi
- Transition:transversion ratio indicative of certain biological processes



# Mutations

## In coding regions

- Synonymous (sSNP)
  - A SNP is described as synonymous if there is no change in the AA encoded
    - ‘silent’/evolutionary neutral
- Non-synonymous (nsSNP)
  - Change in the AA encoded
    - sometimes called ‘missense’
- Stop mutation (STP)
  - AA change encodes an AA that is a stop codon, resulting in premature truncation of protein
    - sometimes called ‘nonsense’
- Read-through mutation (RTH)
  - AA change results in stop codon changing to other AA and protein extended

# Mutations

## In non-coding regions

- Mutations in UTR regions and intergenic regions



# Mapping SNPs to genes

## snpEff

- Cingolani *et al.* 2012 PMID: 22728672
- Genomic variant annotations and functions effect prediction
- Has ~38k genomes pre-loaded, or can make custom database for new species/genomes
- Annotate the final vcf
- `java -jar snpEff.jar genome_ref isolate_name.final_SNPs.vcf > isolate_name.snps.annotated`
- Produces an annotated vcf with an additional 'INFO' column, and a html report

# Mapping SNPs to genes

## snpEff

- html results report

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	7,115	34.968%
NONSENSE	46	0.226%
SILENT	13,186	64.806%

Missense / Silent ratio: 0.5396

Number of effects by type and region

Type			Region		
Type (alphabetical order)	Count	Percent	Type (alphabetical order)	Count	Percent
downstream_gene_variant	103,013	38.927%	DOWNSTREAM	103,013	38.934%
intergenic_region	27,313	10.321%	EXON	20,325	7.682%
intron_variant	247	0.093%	INTERGENIC	27,313	10.323%
missense_variant	7,105	2.685%	INTRON	231	0.087%
non_coding_transcript_exon_variant	1	0%	SPLICE_SITE_ACCEPTOR	1	0%
splice_acceptor_variant	1	0%	SPLICE_SITE_DONOR	1	0%
splice_donor_variant	1	0%	SPLICE_SITE_REGION	37	0.014%
splice_region_variant	47	0.018%	UPSTREAM	113,663	42.959%
start_lost	2	0.001%			
stop_gained	46	0.017%			
stop_lost	8	0.003%			
stop_retained_variant	21	0.008%			
synonymous_variant	13,165	4.975%			
upstream_gene_variant	113,663	42.951%			



# Mapping SNPs to genes

## Annotated vcf

- Lots of information
- Perhaps searching for AA that results in drug resistance?
  - Search for gene of interest on FungiDB/NCBI and use ID to search:

```
Jo Rhodes@MacBook-Pro-4 Annotated % grep AFUA_4G06890 C101.snps.annotated | grep "missense"
NC_007197.1 1780473 . T C 615.04 PASS AC=1;AF=1.00;AN=1;DP=20;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=32.37;SOR=1.292;ANN=C|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|2/2|c.1279A>G|p.Lys427Glu|1279/1548|1279/1548|427/515||,Clupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-3022T>C|||||3022| GT:AD:DP:GQ:PL 1:0,19:19:99:625,0
NC_007197.1 1780987 . C G 632.04 PASS AC=1;AF=1.00;AN=1;DP=23;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=27.48;SOR=1.179;ANN=G|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|2/2|c.765G>C|p.Glu255Asp|765/1548|765/1548|255/515||,Glupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-2508C>G|||||2508| GT:AD:DP:GQ:PL 1:0,23:23:99:642,0
NC_007197.1 1781009 . G T 675.04 PASS AC=1;AF=1.00;AN=1;DP=23;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=29.35;SOR=0.963;ANN=T|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|2/2|c.743C>A|p.Thr248Asn|743/1548|743/1548|248/515||,Tlupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-2486G>T|||||2486| GT:AD:DP:GQ:PL 1:0,23:23:99:685,0
NC_007197.1 1781238 . C T 689.04 PASS AC=1;AF=1.00;AN=1;DP=26;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=29.96;SOR=0.963;ANN=T|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|2/2|c.514G>A|p.Val172Met|514/1548|514/1548|172/515||,Tlupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-2257C>T|||||2257| GT:AD:DP:GQ:PL 1:0,23:23:99:699,0
NC_007197.1 1781459 . A T 966.04 PASS AC=1;AF=1.00;AN=1;DP=31;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=32.20;SOR=0.693;ANN=T|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|2/2|c.293T>A|p.Leu98His|293/1548|293/1548|98/515||,Tlupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-2036A>T|||||2036| GT:AD:DP:GQ:PL 1:0,30:30:99:976,0
NC_007197.1 1781686 . T A 884.04 PASS AC=1;AF=1.00;AN=1;DP=28;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=31.57;SOR=0.836;ANN=A|missense_variant|MODERATE|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protei
n_coding|1/2|c.137A>T|p.Tyr46Phe|137/1548|137/1548|46/515||,Alupstream_gene_variant|MODIFIER|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protein_coding||c.-1809T>A|||||1809|,Alupstream_gene_variant|MODIFIER|AFUA_4G06910|AFUA_4G06910|transcript|XM_747042.1|protein_coding||c.-4872T>A|||||4872| GT:AD:DP:GQ:PL 1:0,28:28:99:894,0
NC_007197.1 1784350 . T C 553.04 PASS AC=1;AF=1.00;AN=1;DP=18;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=32.53;SOR=0.804;ANN=C|missense_variant|MODERATE|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protei
n_coding|6/7|c.566T>C|p.Ile189Thr|566/1920|566/1920|189/639||,Clupstream_gene_variant|MODIFIER|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protein_coding||c.-2528A>G|||||2528|,Clupstream_gene_variant|MODIFIER|AFUA_4G06910|AFUA_4G06910|transcript|XM_747042.1|protein_coding||c.-2208T>C|||||2208|,Cldownstream_gene_variant|MODIFIER|AFUA_4G06920|AFUA_4G06920|transcript|XM_747041.1|protein_coding||c.*4075A>G|||||4075| GT:AD:DP:GQ:PL 1:0,17:17:99:563,0
NC_007197.1 1784694 . A G 945.04 PASS AC=1;AF=1.00;AN=1;DP=28;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=33.75;SOR=1.382;ANN=G|missense_variant|MODERATE|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protei
n_coding|6/7|c.910A>G|p.Thr304Ala|910/1920|910/1920|304/639||,Glupstream_gene_variant|MODIFIER|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protein_coding||c.-2872T>C|||||2872|,Glupstream_gene_variant|MODIFIER|AFUA_4G06910|AFUA_4G06910|transcript|XM_747042.1|protein_coding||c.-1864A>G|||||1864|,Gldownstream_gene_variant|MODIFIER|AFUA_4G06920|AFUA_4G06920|transcript|XM_747041.1|protein_coding||c.*3731T>C|||||3731| GT:AD:DP:GQ:PL 1:0,28:28:99:955,0
NC_007197.1 1785669 . T C 1385.04 PASS AC=1;AF=1.00;AN=1;DP=31;FS=0.000;MLEAC=1;MLEAF=1.00;MQ=60.00;QD=29.60;SOR=0.756;ANN=C|missense_variant|MODERATE|AFUA_4G06900|AFUA_4G06900|transcript|XM_747043.1|protei
n_coding|7/7|c.1826T>C|p.Leu609Pro|1826/1920|1826/1920|609/639||,Clupstream_gene_variant|MODIFIER|AFUA_4G06890|AFUA_4G06890|transcript|XM_747044.1|protein_coding||c.-3847A>G|||||3847|,Clupstream_gene_variant|MODIFIER|AFUA_4G06910|AFUA_4G06910|transcript|XM_747042.1|protein_coding||c.-889T>C|||||889|,Clupstream_gene_variant|MODIFIER|AFUA_4G06930|AFUA_4G06930|transcript|XM_747040.1|protein_coding||c.-4457T>C|||||4457|,Cldownstream_gene_variant|MODIFIER|AFUA_4G06920|AFUA_4G06920|transcript|XM_747041.1|protein_coding||c.*2756A>G|||||2756| GT:AD:DP:GQ:PL 1:0,31:31:99:1395,0
Jo Rhodes@MacBook-Pro-4 Annotated %
```