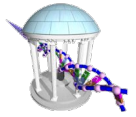


# BCB 716 - Sequence Analysis



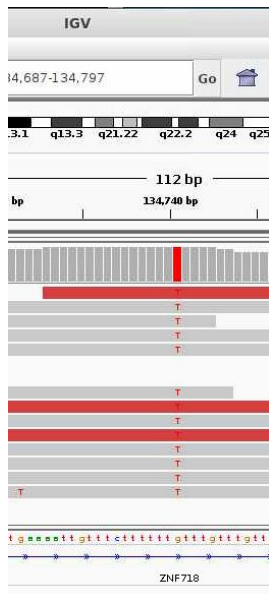
- Lots of problem set issues requires a change in due date for PS#1
- Check the website this weekend

DNA Variant Calling

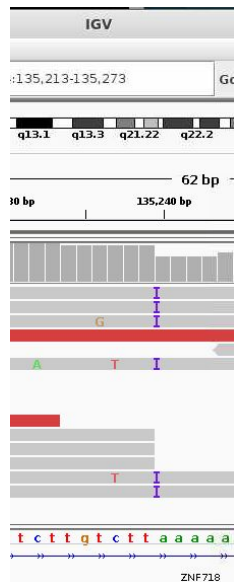
# From last time



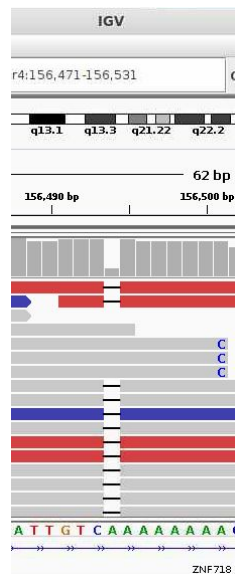
- We noticed using IGV that there are many variants in Genomes



g/T SNP



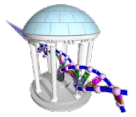
Insertion of  
an A



Deletion of an  
A



Structural variant  
(large 11kb deletion)



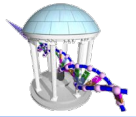
# One of the most popular variant callers

- GATK is a toolkit developed at the Broad Institute for variant calling
- Includes a wide array of tools, many overlapping functions in SAMtools
- Tools for calling variants in a single sample relative to a reference
- Tools for calling variants in a population of multiple sequenced samples
- Tools for calling variants in cases/controls from same sample/organism



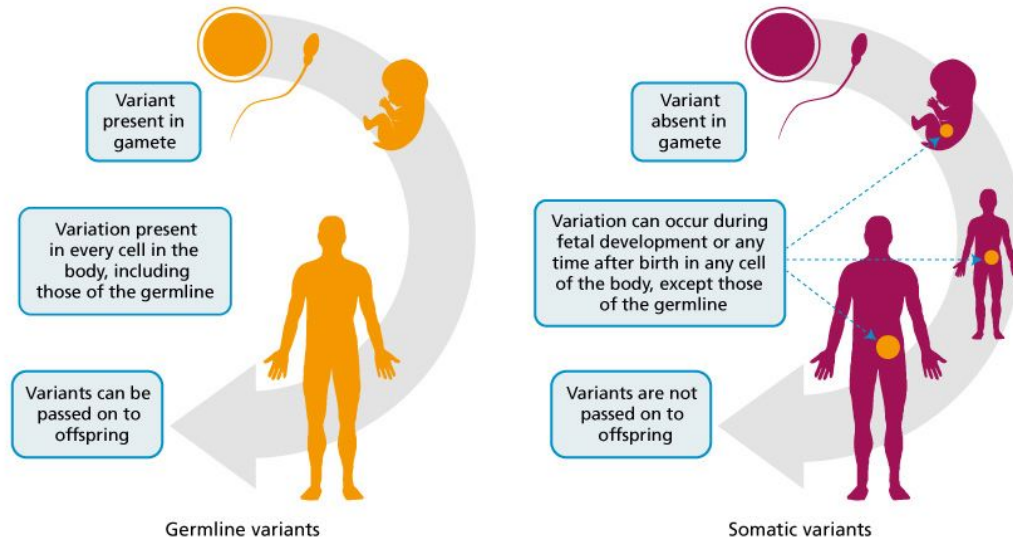
	GERMLINE	SOMATIC
SNPs & INDELS	HaplotypeCaller GVCF	MuTect2
Copy Number	GATK gCNV	GATK CNV + aCNV
Structural Variation	GATK SVDDiscovery (beta)	(planned)

# We will look for Germline Variants

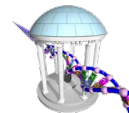


## Inherited genome

(basis genome of the germ cells from which an organism is derived)

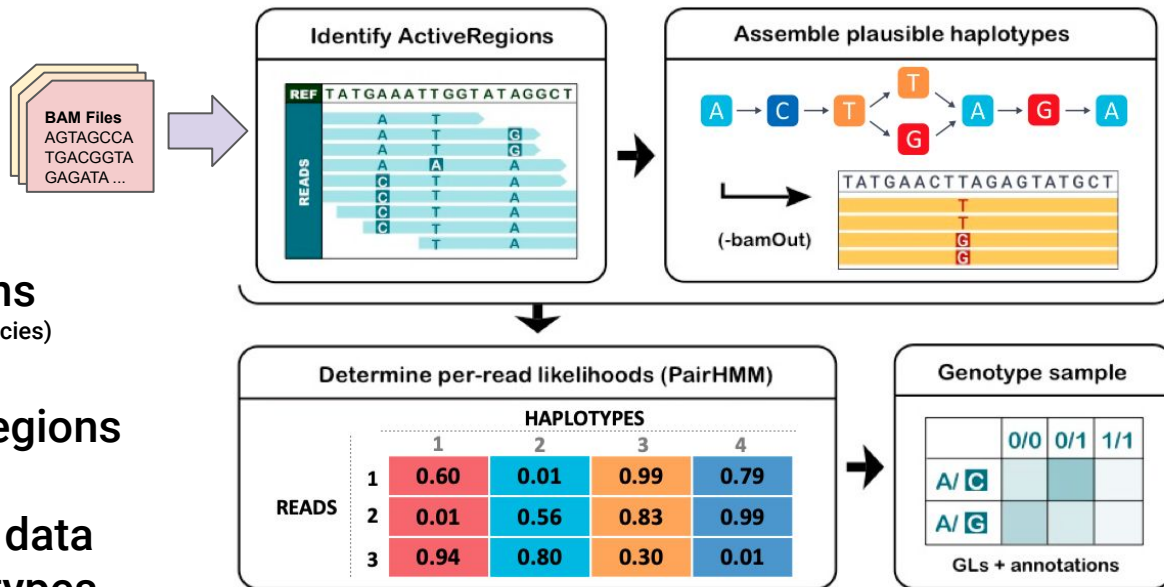


# HaplotypeCaller

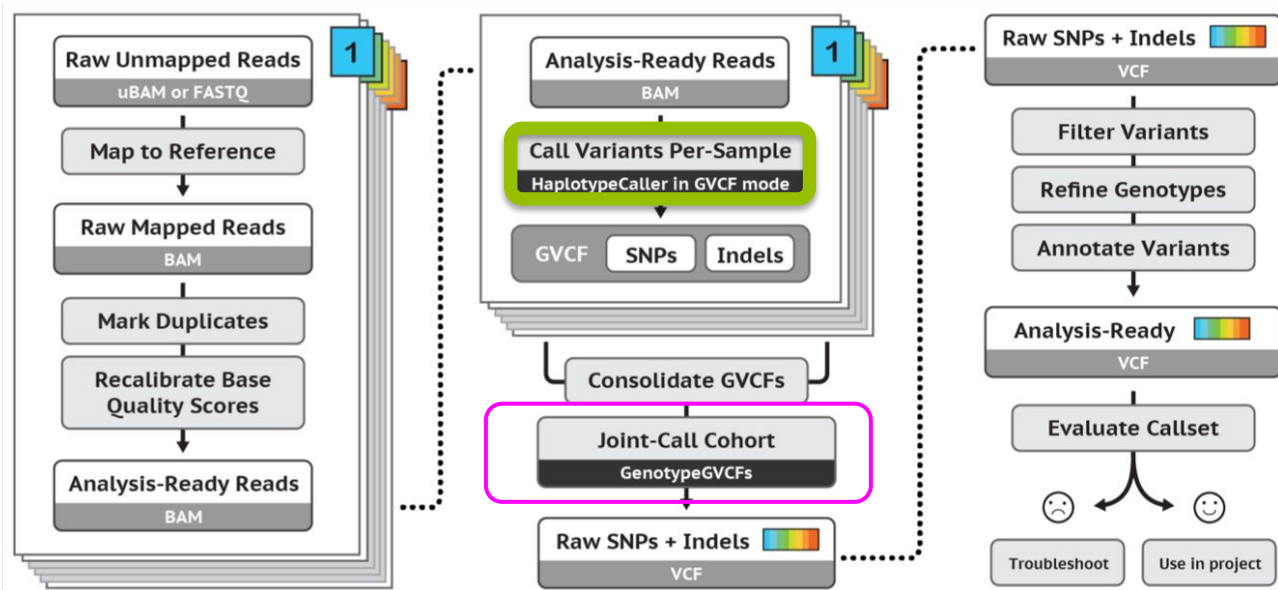
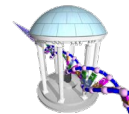


## Four Steps:

1. Identifies active regions  
(places with lots of alignment inconsistencies)
2. Finds haplotypes by reassembling active regions
3. Assesses haplotype likelihoods given read data
4. Assigns sample genotypes



# Workflow for germline short variant discovery



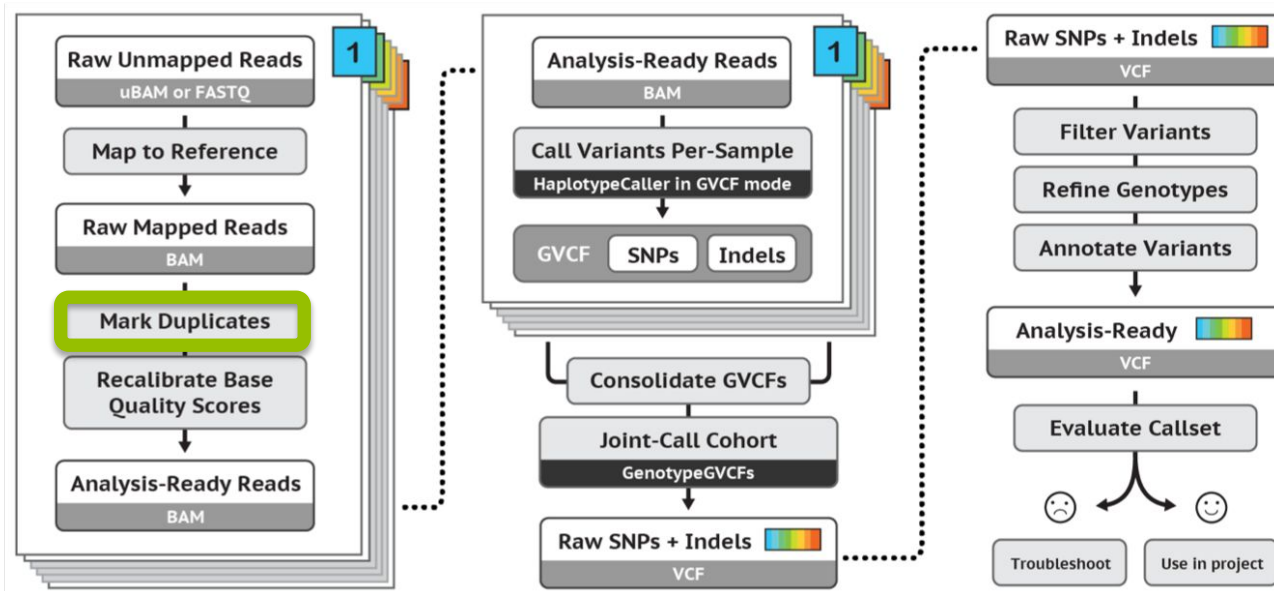
But you don't just run GATK on your BAM file!

You should do a series of preprocessing steps called "best-practices"

<https://software.broadinstitute.org/gatk/best-practices/>

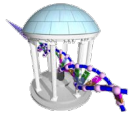


# Mark Duplicates



A considerable number of reads are "identical" technical artifacts

Both mated reads have the same exact sequence

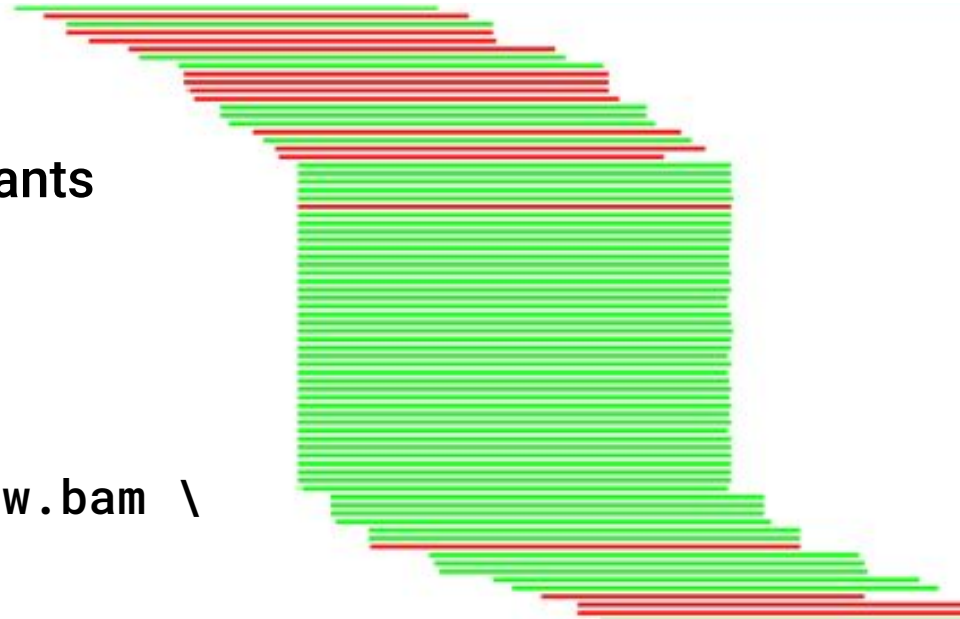


# Duplicate reads

- PCR duplicates - library preparation
- Optical duplicates - sequencing
  - Multiple optical clusters in random flow cells
- Don't add unique information
- Gives false allelic ratios of variants
- Should be removed/marked

Command line:

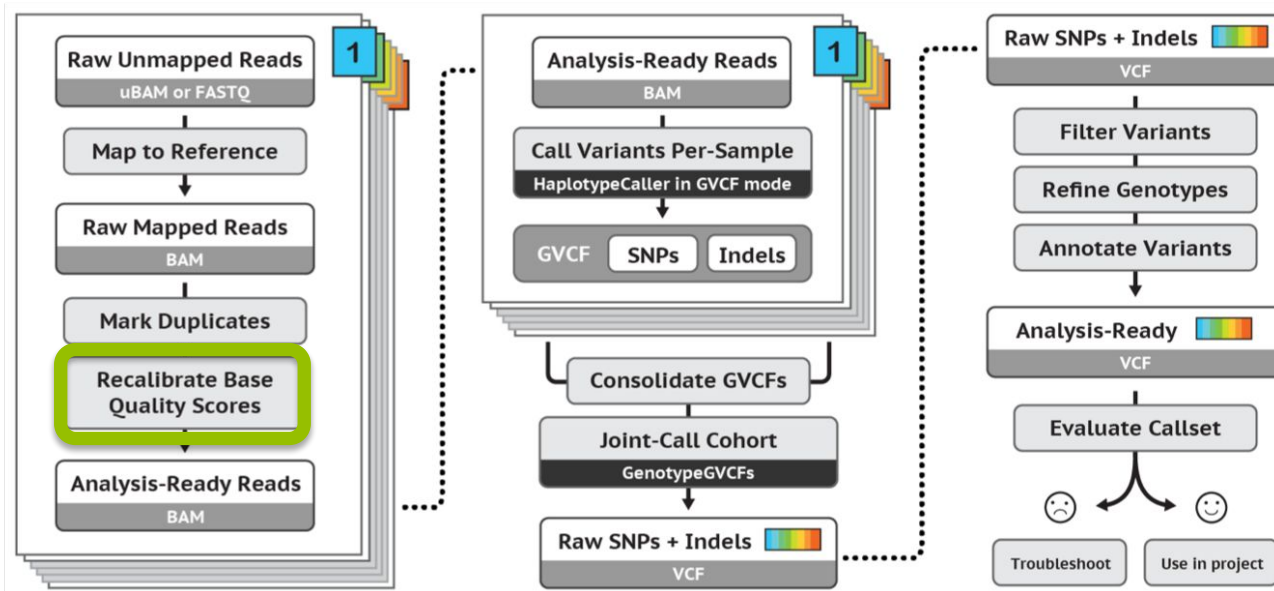
```
$ picard MarkDuplicates -I raw.bam \  
    -O marked.bam \  
    -M markedMetrics.txt
```







# Base Quality Score Recalibration (BQSR)



The base confidence (quality string) of an individual read does not account for other reads.

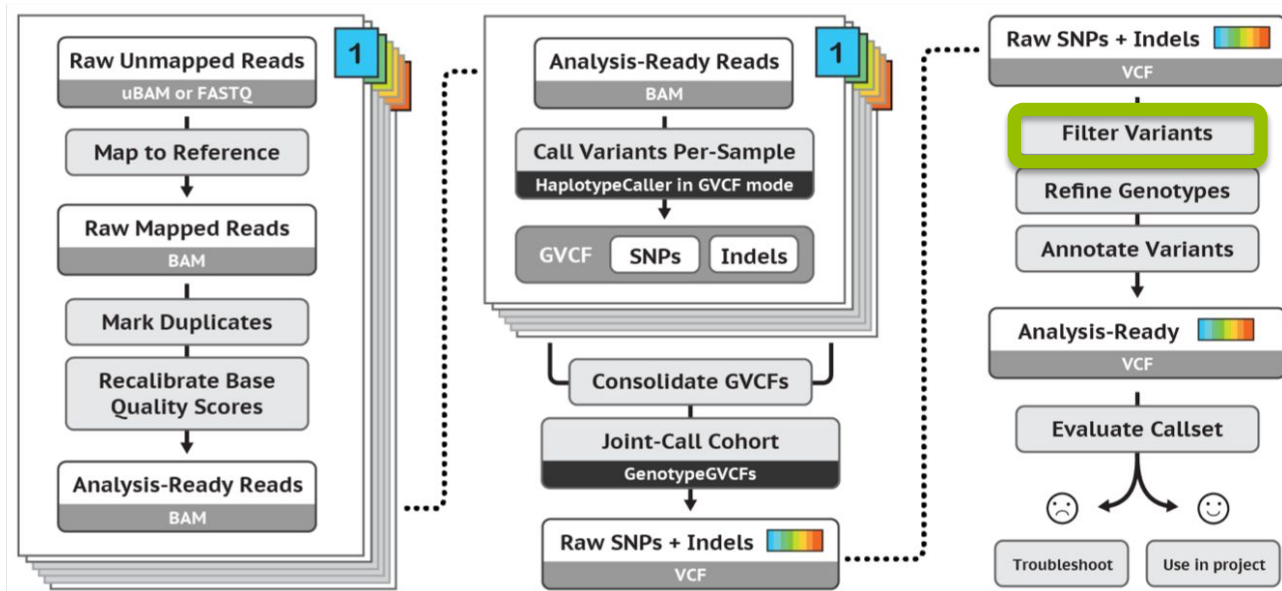
Post-alignment we can revisit the "optical" quality in the context of supporting evidence

# Base Quality Score Recalibration (BQSR)

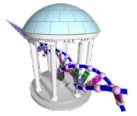


- During base calling, the sequencer estimates a quality score for each base. This is the quality scores present in the fastq files.
- Systematic (non-random) errors in the base quality score estimation can occur.
  - due to the physics or chemistry of the sequencing reaction
  - manufacturing flaws in the equipment
  - etc
- Can cause bias in variant calling
- Base Quality Score Recalibration helps to calibrate the scores so that they correspond to the real per-base sequencing error rate (phred scores)

# Filter variants

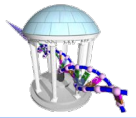


This is a post-variant calling step that selects takes a VCF file, which we will discuss shortly and identifies that high-confidence calls



# Filtering

- **Remove low quality variants**
- **Variant quality score recalibration (VQSR):**
  - For large data sets ( >1 WGS or >30WES samples)
  - GATK has a machine learning algorithm that can be trained to recognise "likely false" variants
  - Use VQSR when possible!
- **Hard filters:**
  - For smaller data sets
  - Hard filters on information in the VCF file
  - For example: Flag variants with "QD < 2" and "MQ < 40.0"
  - GATK recommendations on hard filters:  
<https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set>

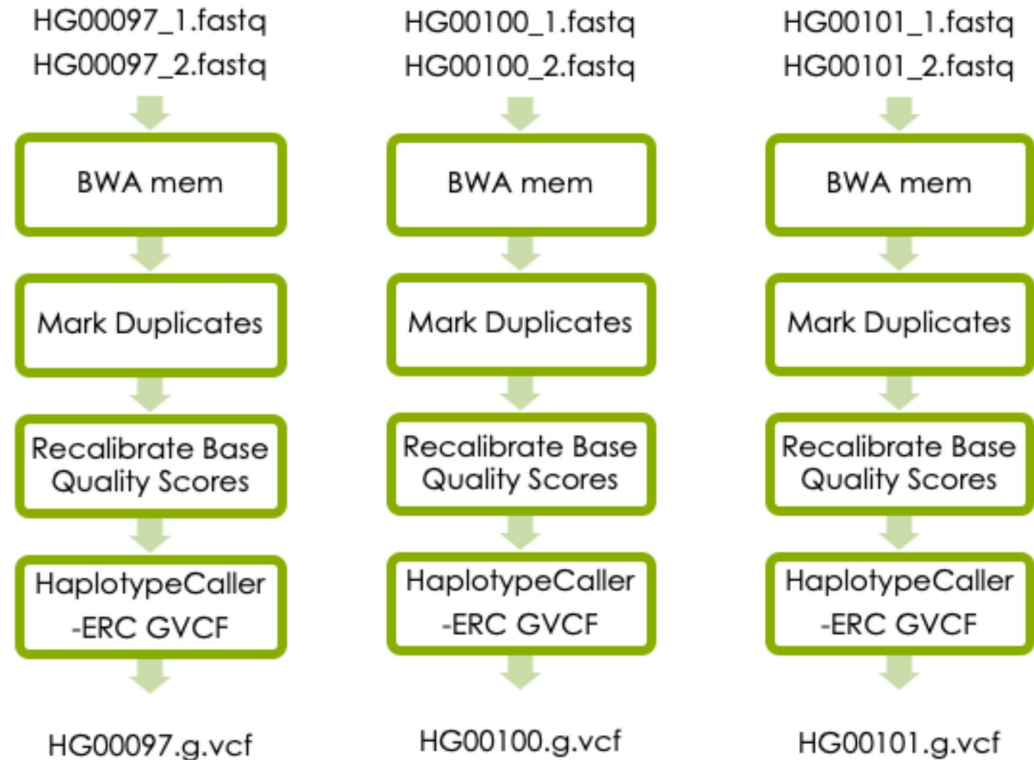


# For One Sample

Here is a typical GATK pipeline for variant calling

1. Align
2. Mark duplicates
3. Recalibrate Quality strings
4. Call Variants

Repeat for multiple samples

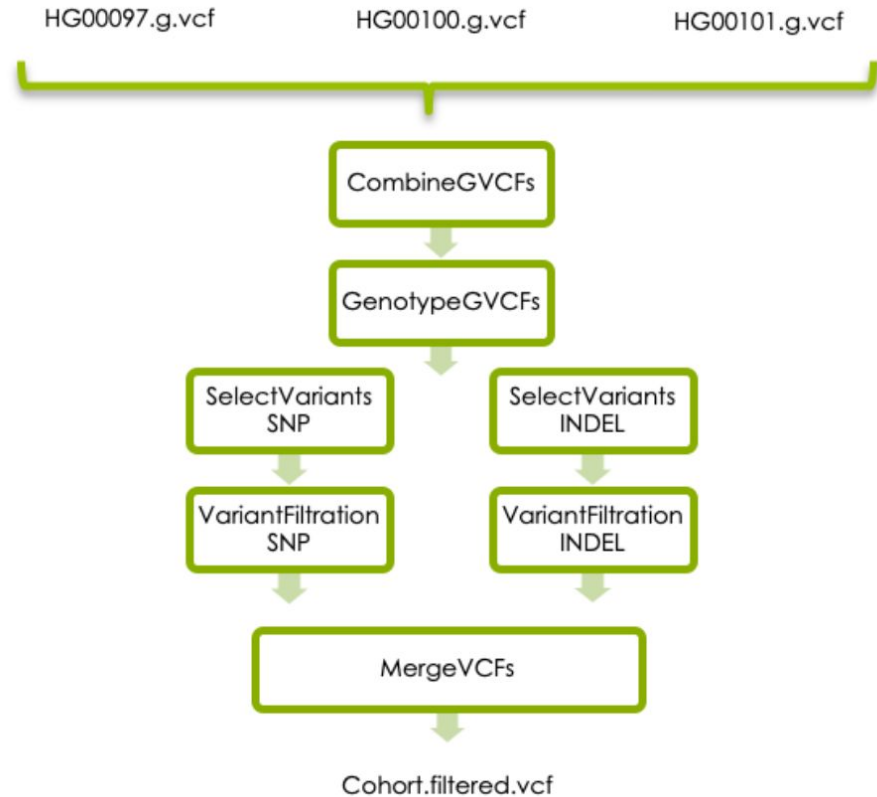




# Multiple Samples

There are additional steps and tools involved to create a Variant Calling File (VCF) from multiple samples.

These are used in populations or tissue specific (ex: tumor vs normal) variant calling

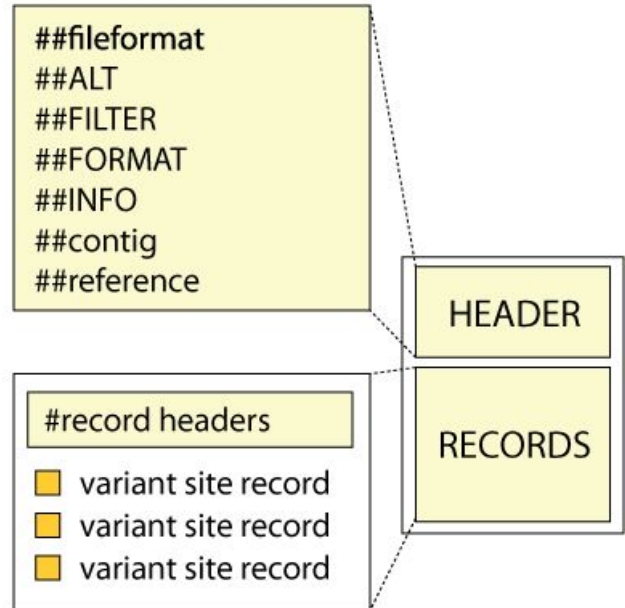




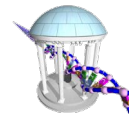
# Variant Calling Files (VCF)

- **Multiple Headers ("##")**
  - Information about dataset
  - Reference used
  - Annotations of options and filters
  - Parameter settings
  - Command line used
  - Header column formats
  - Variant header starts with single "#"
- **Variant Calls**
  - One line for each variant
  - Annotates supporting evidence

## Basic structure of a VCF file



# VCF header column descriptions

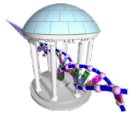


- Part of header that describes each column of a variant call

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles i
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or
##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 bloc
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, desc
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for ger
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which compri
##GATKCommandLine.HaplotypeCaller=<ID=HaplotypeCaller,Version=3.7-0-gcfedb67,Date="Fri Jan 20 1
##GATKCommandLine=<ID=GenotypeGVCFs,CommandLine="[command-line goes here]",Version=4.beta.6-117
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the sa
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of
##INFO=<ID=ClippingRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have be
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=END,Number=1,Type=Integer,Description="Stop position of the interval">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of ex
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test t
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimate
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for th
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Al
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RAW_MQ,Number=1,Type=Float,Description="Raw data for RMS Mapping Quality">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table t
```

← All of these header elements are describing valid options for the "INFO" column





# VCF variant calls

- Tab separated columns
- Semicolons separate multivalued column entries

```
[HEADER LINES]
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10001019 . T G 364.77 . AC=1;AF=0.500;AN=2;BaseQRankSum=0.699;ClippingRankSum=C
20 10001298 . T A 884.77 . AC=2;AF=1.00;AN=2;DP=30;ExcessHet=3.0103;FS=0.000;MLEAC
20 10001436 . A AAGGCT 1222.73 . AC=2;AF=1.00;AN=2;DP=29;ExcessHet=3.0103;FS=0.000;M
20 10001474 . C T 843.77 . AC=2;AF=1.00;AN=2;DP=27;ExcessHet=3.0103;FS=0.000;MLEAC
20 10001617 . C A 493.77 . AC=1;AF=0.500;AN=2;BaseQRankSum=1.63;ClippingRankSum=0.
```

# VCF call supporting evidence



## Showing the FORMAT and NA12878 columns

**GT:** 0/0 : homozygous reference; 0/1 : heterozygous; 1/1 : the sample is homozygous alternate

**AD:** Allele depth data, reads with ref allele, reads with alt allele

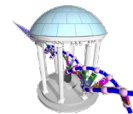
**DP:** "Filtered" depth of coverage at position

**GQ:** Genotype Quality represents the Phred-scaled confidence

**PL:** "Normalized" Phred-scaled likelihoods of the possible genotypes,  
one number for each genotype 1/1,0/1,0/0

```
20 10001298 . T A 884.77 . [CLIPPED] GT:AD:DP:GQ:PL 1/1:0,30:30:89:913,89,0
20 10001436 . A AAGGCT 1222.73 . [CLIPPED] GT:AD:DP:GQ:PL 1/1:0,28:28:84:1260,84,
20 10004769 . TAAACTATGC T 622.73 . [CLIPPED] GT:AD:DP:GQ:PL 0/1:18,17:35:99:660
```

# Next Time



## We move to RNA sequence analysis

