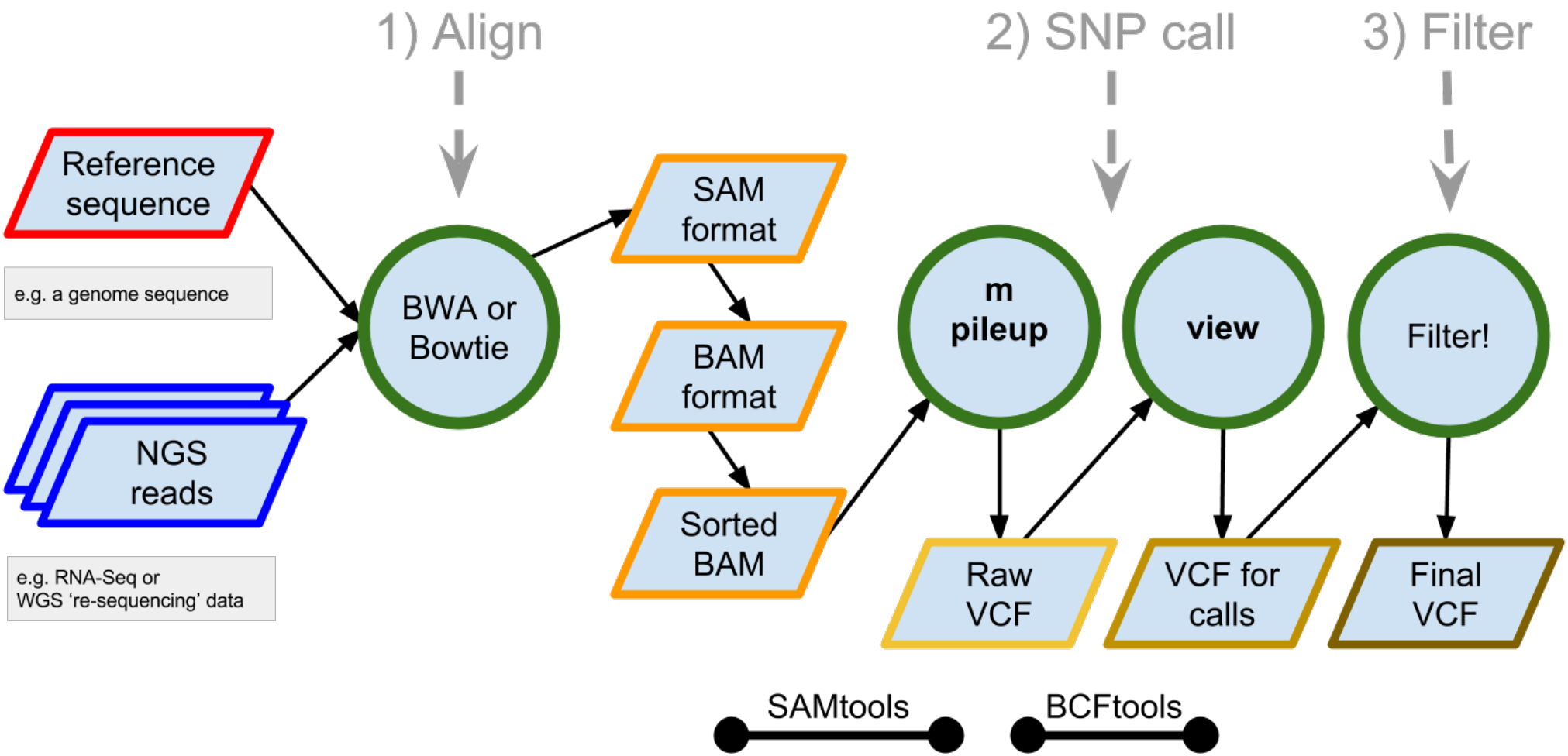
A simple SNP calling pipeline

dbolser@ebi.ac.uk

EMBL-EBI

Pipeline overview

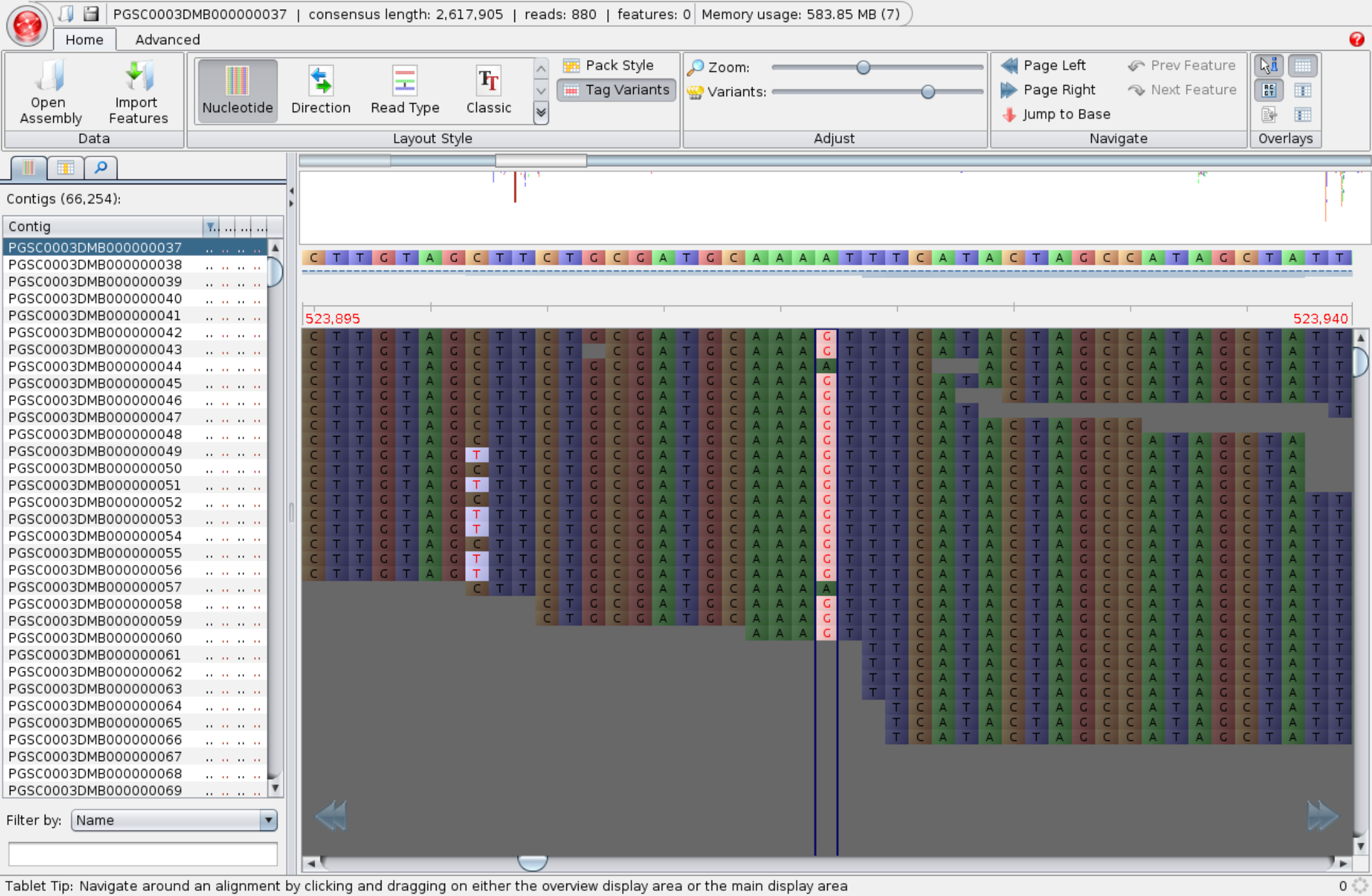


1) Align reads to reference (using BWA)

1. Index the reference (genome) sequence
   - `bwa index my.fasta`
   - `# The various index files are output in the CWD`

2. Perform the alignment
   - `bwa aln [opts] my.fasta my.fastq > my.sai`

3. Output results in SAM format (single end)
   - `bwa samse my.fasta my.sai my.fastq > my.sam`


PGSC0003DMB000000037 | consensus length: 2,617,905 | reads: 880 | features: 0 | Memory usage: 583.85 MB (7)
Home
Advanced
Open Assembly
Import Features
Data
Nucleotide
Direction
Read Type
Classic
Pack Style
Tag Variants
Layout Style
Zoom:
Variants:
Adjust
Page Left
Page Right
Jump to Base
Prev Feature
Next Feature
Navigate
Overlays
Contigs (66,254):
Contig
PGSC0003DMB000000037
PGSC0003DMB000000038
PGSC0003DMB000000039
PGSC0003DMB000000040
PGSC0003DMB000000041
PGSC0003DMB000000042
PGSC0003DMB000000043
PGSC0003DMB000000044
PGSC0003DMB000000045
PGSC0003DMB000000046
PGSC0003DMB000000047
PGSC0003DMB000000048
PGSC0003DMB000000049
PGSC0003DMB000000050
PGSC0003DMB000000051
PGSC0003DMB000000052
PGSC0003DMB000000053
PGSC0003DMB000000054
PGSC0003DMB000000055
PGSC0003DMB000000056
PGSC0003DMB000000057
PGSC0003DMB000000058
PGSC0003DMB000000059
PGSC0003DMB000000060
PGSC0003DMB000000061
PGSC0003DMB000000062
PGSC0003DMB000000063
PGSC0003DMB000000064
PGSC0003DMB000000065
PGSC0003DMB000000066
PGSC0003DMB000000067
PGSC0003DMB000000068
PGSC0003DMB000000069
Filter by:
Name
523,895
523,940
Tablet Tip: Navigate around an alignment by clicking and dragging on either the overview display area or the main display area
0

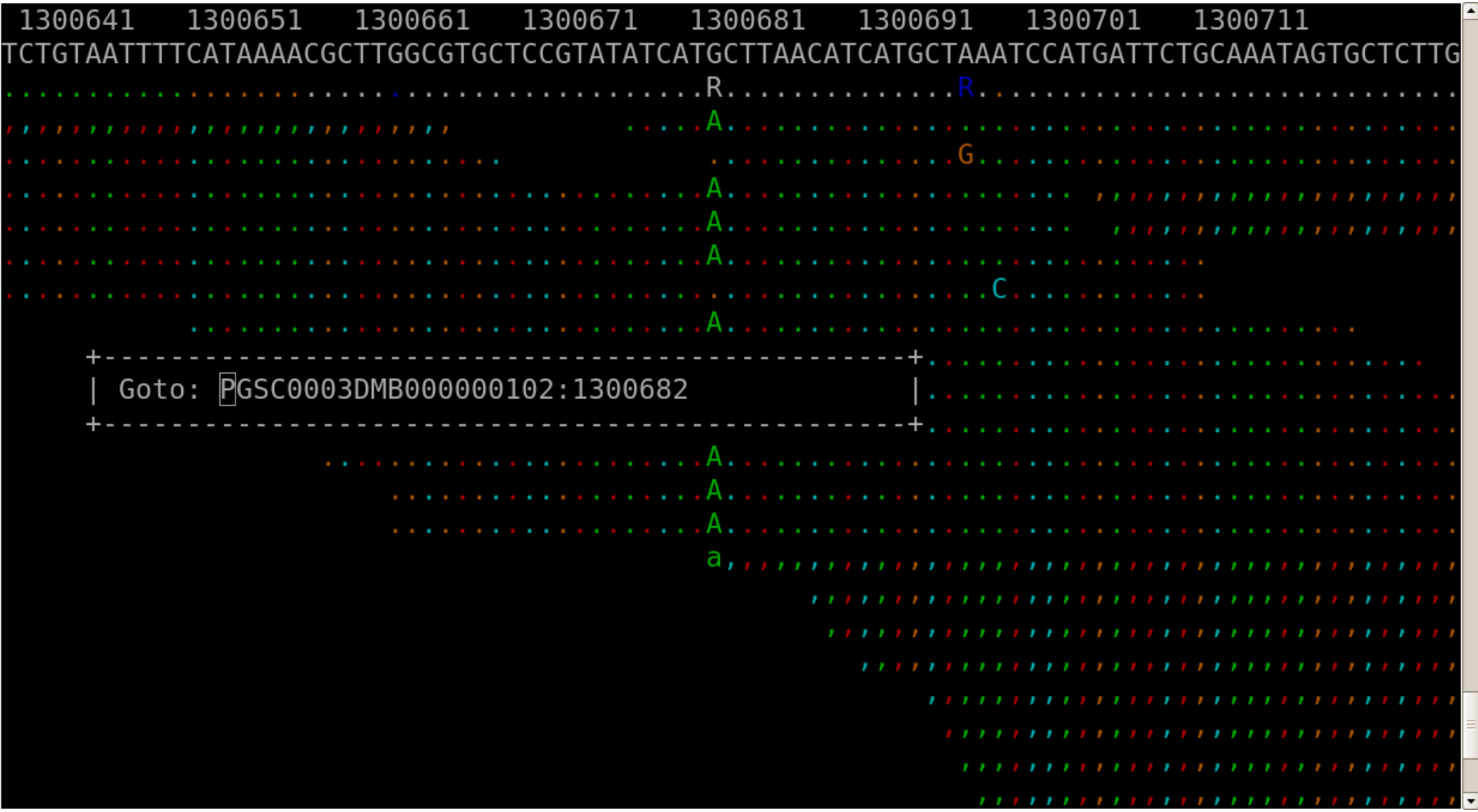
Ambiguity codes

https://wikipedia.org/wiki/Nucleic_acid_notation

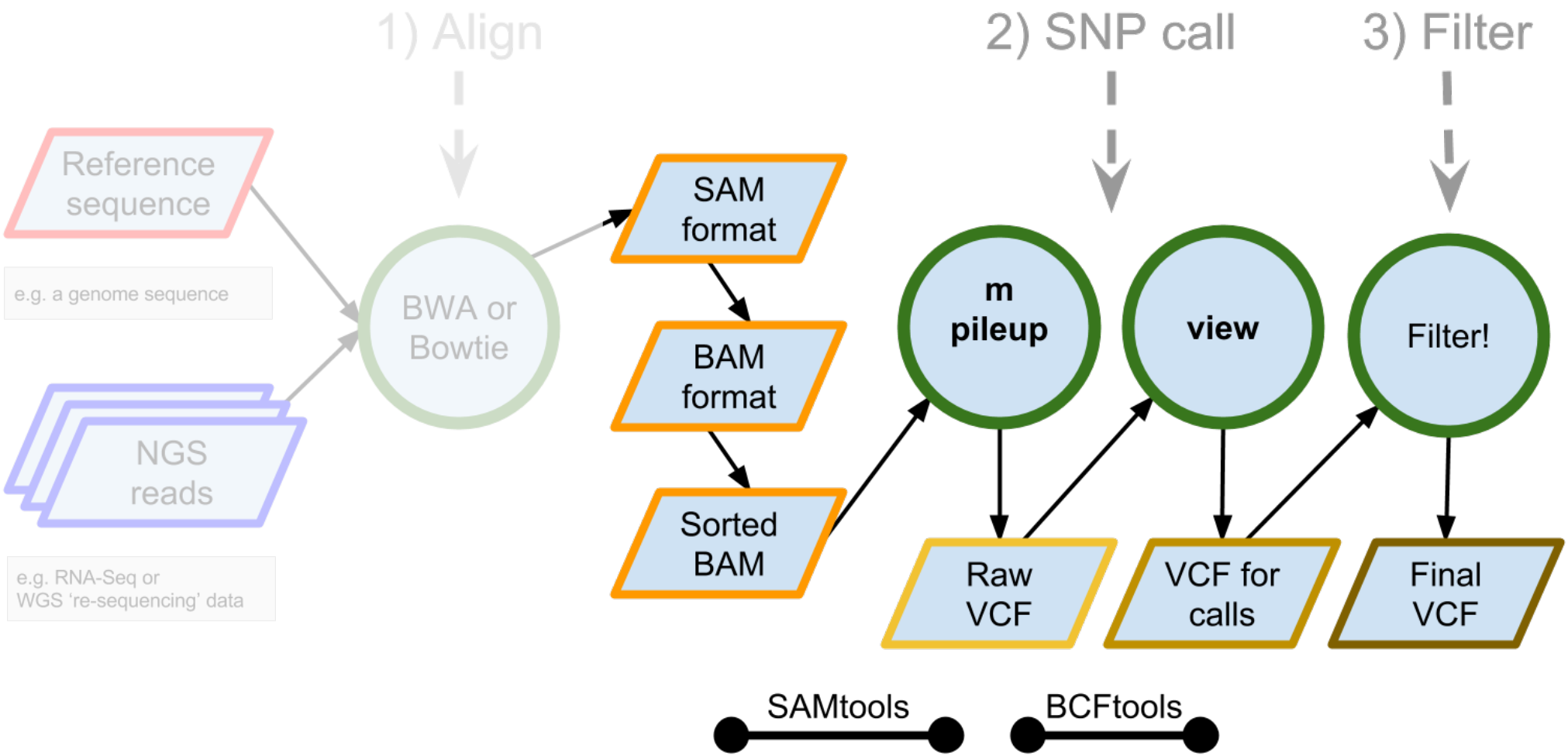
Nomenclature Committee of the International Union of Biochemistry (NC-IUB) (1984). "Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences". Retrieved 2008-02-04.

Symbol[2]	Description	Bases represented				
A	**A**denine	A				1
C	**C**ytosine		C			
G	**G**uanine			G		
T	**T**hymine				T	
U	**U**racil				U	
W	**W**eak	A			T	2
S	**S**trong		C	G		
M	a**M**ino	A	C			
K	**K**eto			G	T	
R	pu**R**ine	A		G		
Y	p**Y**rimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	a**N**y base (not a gap)	A	C	G	T	4


235661 235671 235681 235691 235701 235711 235721
CAGCTGAGGAAATACCAGCTTTCCATTTTCACTGGCTGAGAGTTACCCTTATTTATTTTTGCTTGTTGAGAATATACATATGGTTAA
Goto: PGSC0003DMB000000454:235697


1300641 1300651 1300661 1300671 1300681 1300691 1300701 1300711
TCTGTAATTTTCATAAAACGCTTGGCGTGCTCCGTATATCATGCTTAACATCATGCTAAATCCATGATTCTGCAAATAGTGCTCTTG
Goto: PGSC0003DMB000000102:1300682

Alignment is done! Next, SNP calling!!



First... convert alignments (using SAMtools)

1. Convert SAM to BAM for sorting
   - `samtools view -S -b my.sam > my.bam`

2. Sort BAM for SNP calling
   - `samtools sort my.bam my-sorted`

Alignments are both:

- compressed for long term storage and
- sorted for variant discovery.

2) Call SNPs (using SAMtools)

1. Index the genome assembly (again!)
   - `samtools faidx my.fasta`

2. Run 'mpileup' to generate VCF format
   - `samtools mpileup -g -f my.fasta my-sorted-1.bam my-sorted-2.bam my-sorted-n.bam > my-raw.bcf`

NB: All we did so far (roughly) is to perform a format conversion from BAM to VCF!

2) Call SNPs (using bcftools)

3. Call SNPs...
   - `bcftools view -bvcg my-raw.bcf > my-var.bcf`

Again...

- `samtools mpileup`
  - Collects summary information in the input BAMs, computes the likelihood of data given each possible genotype and stores the likelihoods in the BCF format.

- `bcftools view`
  - Applies the prior and does the actual calling.

3) Filter SNPs

1. Filter SNPs
   - `bcftools view my.var.bcf |`
     `vcfutils.pl varFilter - > my.var-final.vcf`

Now...

Your turn!

Options for BWA

For details, see http://bio-bwa.sourceforge.net/bwa.shtml

Options for SAMtools

For details, see http://samtools.sourceforge.net/samtools.shtml

Options for vcfutils.pl varFilter

```
Usage:   vcfutils.pl varFilter [options] <in.vcf>

Options: -Q INT    minimum RMS mapping quality for SNPs [10]
         -d INT    minimum read depth [2]
         -D INT    maximum read depth [10000000]
         -a INT    minimum number of alternate bases [2]
         -w INT    SNP within INT bp around a gap to be filtered [3]
         -W INT    window size for filtering adjacent gaps [10]
         -1 FLOAT  min P-value for strand bias (given PV4) [0.0001]
         -2 FLOAT  min P-value for baseQ bias [1e-100]
         -3 FLOAT  min P-value for mapQ bias [0]
         -4 FLOAT  min P-value for end distance bias [0.0001]
         -e FLOAT  min P-value for HWE (plus F<0) [0.0001]
         -p        print filtered variants

Note: Some of the filters rely on annotations generated by SAMtools/BCFtools.
```

Glossary of file formats

Sequence data formats:

- FASTA:
  Simple format for DNA or peptide sequences.

- FASTQ:
  Stores sequences and sequence **quality** information together.

Alignment data formats

- SAM / BAM
  Sequence Alignment/Map

Variation data

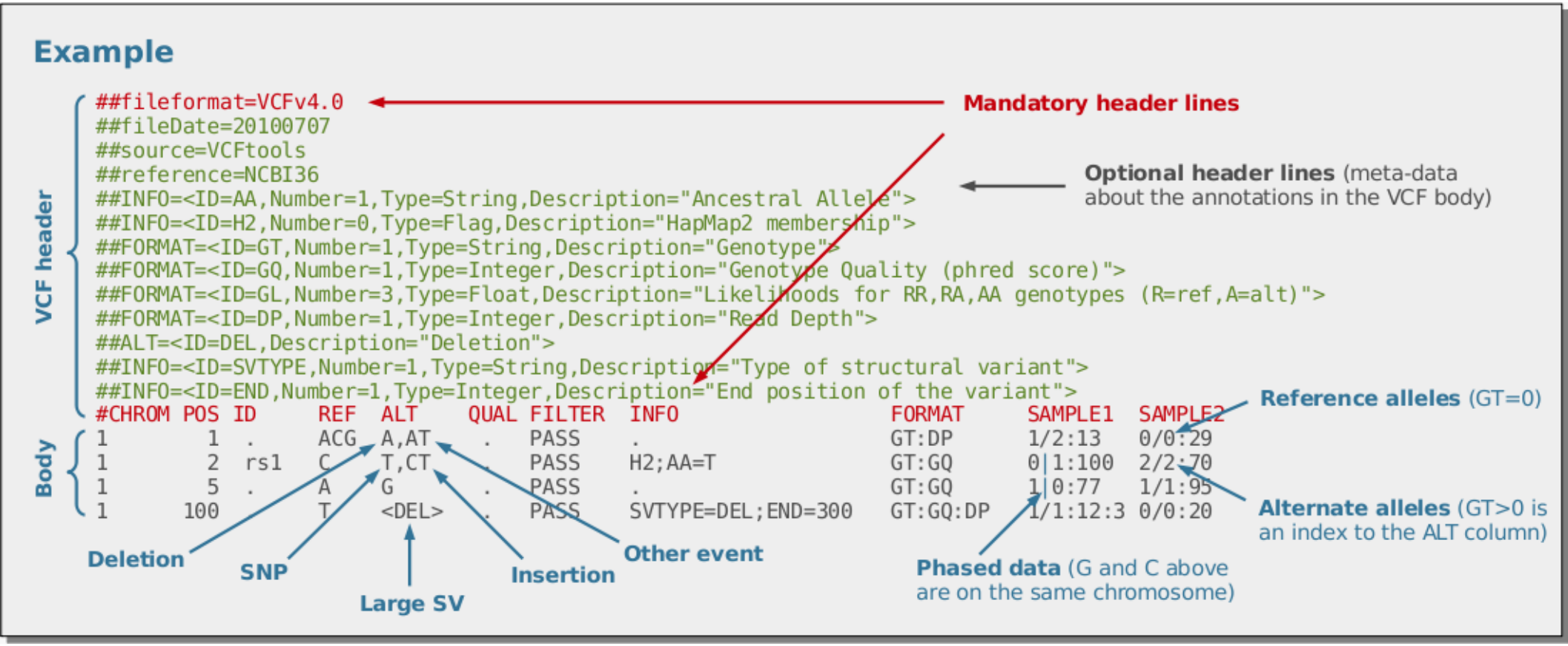
- VCF / BCF
  Variant Call Format

http://www.ebi.ac.uk/ena/about/read-file-formats

VCF format

- A standard format for sequence variation: SNPs, indels and structural variants.
- Compressed and indexed.
- Developed for the 1000 Genomes Project.
- VCFtools for VCF like SAMtools for SAM.
- Specification and tools available from http://vcftools.sourceforge.net

Example


VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)