A simple SNP calling pipeline

dbolser@ebi.ac.uk
Pipeline overview

1) Align
- Reference sequence
  - e.g. a genome sequence
- NGS reads
  - e.g. RNA-Seq or WGS 're-sequencing' data
- BWA or Bowtie

   SAM format

   BAM format

   Sorted BAM

2) SNP call
- m pileup
- view
- Filter!

   Raw VCF

   VCF for calls

   Final VCF

3) Filter

SAMtools

BCFtools
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
Ambiguity codes

https://wikipedia.org/wiki/Nucleic_acid_notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Bases represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
<td>G</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
<td>T</td>
</tr>
<tr>
<td>U</td>
<td>Uracil</td>
<td>U</td>
</tr>
<tr>
<td>W</td>
<td>Weak</td>
<td>A, T</td>
</tr>
<tr>
<td>S</td>
<td>Strong</td>
<td>C, G</td>
</tr>
<tr>
<td>M</td>
<td>aMino</td>
<td>A, C</td>
</tr>
<tr>
<td>K</td>
<td>Keto</td>
<td>G, T</td>
</tr>
<tr>
<td>R</td>
<td>puRine</td>
<td>A, G</td>
</tr>
<tr>
<td>Y</td>
<td>pYrimidine</td>
<td>C, T</td>
</tr>
<tr>
<td>B</td>
<td>not A (B comes after A)</td>
<td>C, G, T</td>
</tr>
<tr>
<td>D</td>
<td>not C (D comes after C)</td>
<td>A, G, T</td>
</tr>
<tr>
<td>H</td>
<td>not G (H comes after G)</td>
<td>A, C, T</td>
</tr>
<tr>
<td>V</td>
<td>not T (V comes after T and U)</td>
<td>A, C, G</td>
</tr>
<tr>
<td>N or -</td>
<td>aNy base (not a gap)</td>
<td>A, C, G, T</td>
</tr>
</tbody>
</table>
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

Variation data
- VCF / BCF

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.
Example

```plaintext
##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">
##INFO=<ID=GT, Number=1, Type=String, Description="Genotype">
##INFO=<ID=GQ, Number=1, Type=Integer, Description="Genotype Quality (phred score)">
##INFO=<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    1/1:93
1    100 .  <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)

**Phased data** (G and C above are on the same chromosome)