bwa vs bowtie2

bwa and bowtie2

defaults settings
bwa vs bowtie2

bwa and bowtie2
defaults settings

bowtie2 tuning
-D 20 -R 3 -N 1 -L 20
Alignment evaluation

The ROC curves

<table>
<thead>
<tr>
<th>Program</th>
<th>Version</th>
<th>Options</th>
<th>100k 100bp SE</th>
<th>100k 2x100bp PE (CPU sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bowtie2</td>
<td>2.0.0-beta4</td>
<td>-X 650; mapQ&gt;1</td>
<td>78.1</td>
<td>154.0 (to be updated)</td>
</tr>
<tr>
<td>bwa</td>
<td>0.5.9-r26-dev</td>
<td>(default); mapQ&gt;0</td>
<td>106.5</td>
<td>230.1</td>
</tr>
<tr>
<td>bwa-sw</td>
<td>0.5.9-r26-dev</td>
<td>(default); mapQ&gt;0</td>
<td>237.4</td>
<td>502.0</td>
</tr>
<tr>
<td>bwa-sw64</td>
<td>0.6.0-r79-dev</td>
<td>(default); mapQ&gt;0</td>
<td>139.4</td>
<td>286.5</td>
</tr>
<tr>
<td>gsnap</td>
<td>2011-10-16</td>
<td>(default); mapQ&gt;3</td>
<td>98.9</td>
<td>538.9</td>
</tr>
<tr>
<td>novoalign</td>
<td>2.05.33</td>
<td>-k14 -s3 -i 500 50; mapQ&gt;3</td>
<td>359.7</td>
<td>349.5</td>
</tr>
<tr>
<td>smalt</td>
<td>~2011-10-17</td>
<td>-k20 -s13 -i 650; mapQ&gt;0</td>
<td>468.8</td>
<td>640.2</td>
</tr>
</tbody>
</table>
• Pileup \rightarrow show all bases at a given index
Sequence data to genotypes

- A common sequencing workflow

Sequencing reads $\rightarrow$ Alignments $\rightarrow$ Variant calls

- FASTQ
- SAM/BAM
- VCF

- a list of short sequences
- a list of short sequences and where they are in the genome
- a list of locations in the genome and what the base is at each
What are variant calls?

- **Naive variant calling**
  - Check all the reads that cover base chr1:291
  - Add up the bases at chr1:291
  - e.g. 10 A's, 2 G's
    - Is this an A/G heterozygous site or two sequencing errors?

- **Actual variant callers**
  - Estimate likelihood of a variant site vs a sequencing error
    - Sequencing error rate
    - Quality scores

Note: it is not always obvious what the underlying assumptions of a snp caller are. Especially when used for genomes other than human/mouse. These are by far the most studied and customized for.
VCF: Variant Call Format

- Represent a list of locations and the variant call at each
  - Simple, right?
- Yes and no.
  - Simple foundation
    - Location and base
  - Complex “bonus features”
    - Indels, structural variants, etc.
    - Multiple samples
    - Haplotype phasing
NAME
vcf – Variant Cell Format

DESCRIPTION
The Variant Call Format (VCF) is a TAB-delimited format with each data line consisting of the following fields:
1  CHROM  CHROMosome name
2  POS    the left-most POSition of the variant
3  ID     unique variant IDentifier
4  REF    the REFerence allele
5  ALT    the ALTernate allele(s) (comma-separated)
6  QUAL   variant/reference QUALITY
7  FILTER FILTERs applied
8  INFO   INFORMATION related to the variant (semicolon-separated)
9  FORMAT FORMAT of the genotype fields (optional; colon-separated)
10+ SAMPLE SAMPLE genotypes and per-sample information (optional)
The following table gives the INFO tags used by samtools and bcftools.

<table>
<thead>
<tr>
<th>TAG</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF1</td>
<td>Max-likelihood estimate of the site allele frequency (AF) of the first ALT allele (double)</td>
</tr>
<tr>
<td>DP</td>
<td>Raw read depth (without quality filtering) (int)</td>
</tr>
<tr>
<td>DP4</td>
<td># high-quality reference forward bases, ref reverse, alternate for and alt rev bases (int[4])</td>
</tr>
<tr>
<td>FQ</td>
<td>Consensus quality. Positive: sample genotypes different; negative: otherwise (int)</td>
</tr>
<tr>
<td>MQ</td>
<td>Root-Mean-Square mapping quality of covering reads (int)</td>
</tr>
<tr>
<td>PC2</td>
<td>Phred probability of AF in group1 samples being larger (smaller) than in group2 (int[2])</td>
</tr>
<tr>
<td>PCHI2</td>
<td>Posterior weighted chi^2 P-value between group1 and group2 samples (double)</td>
</tr>
<tr>
<td>PV4</td>
<td>P-value for strand bias, baseQ bias, mapQ bias and tail distance bias (double[4])</td>
</tr>
<tr>
<td>QCHI2</td>
<td>Phred-scaled PCHI2 (int)</td>
</tr>
<tr>
<td>RP</td>
<td># permutations yielding a smaller PCHI2 (int)</td>
</tr>
<tr>
<td>CLR</td>
<td>Phred log ratio of genotype likelihoods with and without the trio/pair constraint (int)</td>
</tr>
<tr>
<td>UGT</td>
<td>Most probable genotype configuration without the trio constraint (string)</td>
</tr>
<tr>
<td>CGT</td>
<td>Most probable configuration with the trio constraint (string)</td>
</tr>
<tr>
<td>VDB</td>
<td>Tests variant positions within reads. Intended for filtering RNA-seq artifacts around splice sites (float)</td>
</tr>
<tr>
<td>RPB</td>
<td>Mann-Whitney rank-sum test for tail distance bias (float)</td>
</tr>
<tr>
<td>HWE</td>
<td>Hardy-Weinberg equilibrium test (Wigginton et al) (float)</td>
</tr>
</tbody>
</table>
The Variant Call Format (VCF) Version 4.2 Specification

17 Dec 2013

The master version of this document can be found at https://github.com/samtools/hts-specs. This printing is version c02ad4c from that repository, last modified on the date shown above.

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.
VCF Poster: great reference

Example

```
##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
1  1   .  ACG  A,AT  .  PASS  .
1  2   rs1  C  T,CT  .  PASS  H2;AA=T
1  5   .  A  G  .  PASS  .
1 100  T  <DEL>  .  PASS  SVTYPE=DEL;END=300

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>SAMPLE1</th>
<th>SAMPLE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT:DP</td>
<td>1/2:13</td>
<td>0/0:29</td>
</tr>
<tr>
<td>GT:GQ</td>
<td>0 1/100</td>
<td>2/2:70</td>
</tr>
<tr>
<td>GT:GQ:DP</td>
<td>1 0:77</td>
<td>1/1:95</td>
</tr>
<tr>
<td>GT:GQ:DP</td>
<td>1 1/12:3</td>
<td>0/0:20</td>
</tr>
</tbody>
</table>
```

- **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)
VCF: The simple part

- location, reference base, your base
  - CHROM/POS, REF, ALT

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_002549</td>
<td>2951</td>
<td>.</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>NC_002549</td>
<td>3939</td>
<td>.</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>NC_002549</td>
<td>14191</td>
<td>.</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>NC_002549</td>
<td>17235</td>
<td>.</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>NC_002549</td>
<td>18307</td>
<td>.</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>NC_002549</td>
<td>18602</td>
<td>.</td>
<td>T</td>
<td>A</td>
</tr>
</tbody>
</table>

- a lot like wgsim's mutations.txt
VCF: the more complicated parts

<table>
<thead>
<tr>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td></td>
<td>DP=42;VDB=GT:PL</td>
<td>1/1:92,117,0</td>
</tr>
<tr>
<td>56</td>
<td></td>
<td>DP=31;VDB=GT:PL</td>
<td>1/1:83,81,0</td>
</tr>
<tr>
<td>4.74549</td>
<td></td>
<td>DP=45;VDB=GT:PL</td>
<td>0/1:36,0,24</td>
</tr>
<tr>
<td>43.8301</td>
<td></td>
<td>DP=50;VDB=GT:PL</td>
<td>1/1:71,111,0</td>
</tr>
<tr>
<td>51</td>
<td></td>
<td>DP=37;VDB=GT:PL</td>
<td>1/1:78,102,0</td>
</tr>
<tr>
<td>18.874</td>
<td></td>
<td>DP=23;VDB=GT:PL</td>
<td>1/1:46,52,0</td>
</tr>
</tbody>
</table>

Small VCF files are easiest to interpret in Excel
VCF: Multiple samples

- VCF can have a variable number of columns!

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>SAMPLE1</th>
<th>SAMPLE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.</td>
<td>ACG</td>
<td>A,AT</td>
<td>PASS</td>
<td></td>
<td></td>
<td>GT:DP</td>
<td>1/2:13</td>
<td>0/0:29</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>rs1</td>
<td>C</td>
<td>T,CT</td>
<td>PASS</td>
<td>H2;AA=T</td>
<td></td>
<td>GT:GQ</td>
<td>0/1:100</td>
<td>2/2:70</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>PASS</td>
<td></td>
<td></td>
<td>GT:GQ</td>
<td>1/0:77</td>
<td>1/1:95</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>T</td>
<td>&lt;DEL&gt;</td>
<td></td>
<td>PASS</td>
<td>SVTYPE=DEL;END=300</td>
<td>GT:GQ:DP</td>
<td>1/1:12:3</td>
<td>0/0:20</td>
<td></td>
</tr>
</tbody>
</table>

- Column headings are the sample names

VCF review

- VCF can represent SNV calls
- and much, much more
  - Indels (G → GC)
  - Multiple variants per site (in ALT column)
  - Multiple samples (SAMPLE columns)
- Check poster for quick overview
- Check full specification for details
Homework 20

- Generate alignments from a mutated genome (or use the prior results).

- Visualize the expected mutations in IGV

- Call SNPS with samtools.

- Overlay your snp calls with those that IGV shows and those that you expect to see.

- Evaluate/discuss the snps that IGV makes visible, those visible in the samtools output versus the real mutations