Week 11, Lecture 21

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State
## Sequence data to genotypes

- A common sequencing workflow

<table>
<thead>
<tr>
<th>Sequencing reads</th>
<th>Alignments</th>
<th>Variant calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASTQ</td>
<td>SAM/BAM</td>
<td>VCF</td>
</tr>
</tbody>
</table>

- 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
Pileups

- Pileup → show all bases at a given index

```
ialbert@grit ~/edu/lec18
$ samtools mpileup bwa.bam | head -3
[mpileup] 1 samples in 1 input files
<mPILEUP> Set max per-file depth to 8000
NC_002549  1   N   1   ^}C   5
NC_002549  2   N   1   G    5
NC_002549  3   N   2   G^}G   55
(biostar)
ialbert@grit ~/edu/lec18
$ samtools mpileup -f ~/refs/852/NC.fa bwa.bam | head -3
[mpileup] 1 samples in 1 input files
<mPILEUP> Set max per-file depth to 8000
NC_002549  1   c   1   ^].   5
NC_002549  2   g   1   .    5
NC_002549  3   g   2   .^].   55
(biostar)
ialbert@grit ~/edu/lec18
$`
```
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 Call 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

```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=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"> 
##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
1      1    .    ACG   A,AT  .      PASS   .  
1      2    rs1  C    T,CT  .      PASS   H2;AA=T 
1      5    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>
</tbody>
</table>
```

### Mandatory header lines
- `##fileformat=VCFv4.0`
- `##fileDate=20100707`
- `##source=VCFtools`
- `##reference=NCBI36`

### Optional header lines (meta-data about the annotations in the VCF body)
- `##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=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">`
- `##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">`

### Reference alleles (GT=0)
- `GT:DP 1/2:13 0/0:29`
- `GT:GQ 0/1:100 2/2:70`
- `GT:GQ:DP 1/0:77 1/1:95`

### Alternate alleles (GT>0 is an index to the ALT column)
- `GT:DP 1/1:12:3 0/0:20`

### Important events
- **Deletion**
- **SNP**
- **Large SV**
- **Insertion**
- **Other event**
- **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
Small VCF files are easiest to interpret in Excel

<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>
VCF: Multiple samples

- VCF can have a variable number of columns!

- 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
VCF and BCF

• The same relationship as SAM and BAM formats

• BCF – binary, compressed VCF – much smaller but need to be operated on with **bcftools**
DIY snp caller

```python
# DIY snp caller

import sys
from collections import defaultdict

# This code was written in about 20 minutes to demonstrate
# a really simple snp caller. As it is it calls only
# homozygous mutation defined as over 50% of calls.

print "##fileformat=VCFv4.2"
print '##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">'
print '##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype"'
print "\t".join("##CHROM POS ID REF ALT QUAL FILTER INFO FORMAT bwa.bam".split())

for line in sys.stdin:
    # Strip the newline then split by tabs.
    column = line.strip().split("\t")

    # The fifth base contains the basecalls
    # (zero based counting).
    bases = column[4]

    # The quals column will tell us how many bases are called.
    quals = column[5]

    # Upper case everything.
```
SNP calling comparison
Homework 21

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

- Visualize the expected mutations in IGV

- Call SNPS with a snp caller of your choice.

- 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