2013 - BMMB 597D: Analyzing Next Generation Sequencing Data

Week 10, Lecture 19

Nick Stoler

The Huck Institutes
of the Life Sciences

Penn State

Sequence data to genotypes

- A common sequencing workflow

Sequencing reads →	Alignments →	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

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

VCF: The simple part

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

ex.vcf

#CHROM	POS	ID	REF	ALT
chrM	2544	.	G	.
chrM	2545	.	C	.
chrM	2546	.	C	G
chrM	2547	.	T	.
chrM	2548	.	G	.
chrM	2550	.	C	A
chrM	2551	.	C	.

  - a lot like wgsim's mutations.txt

VCF: The rest

aln10k.vcf

```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=CLR,Number=1,Type=Integer,Description="Log ratio of genotype likelihoods with and without the constraint">
##INFO=<ID=UGT,Number=1,Type=String,Description="The most probable unconstrained genotype configuration in the trio">
##INFO=<ID=CGT,Number=1,Type=String,Description="The most probable constrained genotype configuration in the trio">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller)
than in group2.">
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2
samples.">
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##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="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  aln10k.bam
chrI    15891   .       A       C       9.49    .       DP=3;VDB=0.0300;AF1=1;AC1=2;DP4=0,0,0,3;MQ=18;FQ=-36    GT:PL:DP:GQ     1/1:41,9,0:3:12
chrI    47991   .       C       T       4.61    .       DP=2;VDB=0.0220;AF1=1;AC1=2;DP4=0,0,1,1;MQ=42;FQ=-33    GT:PL:DP:GQ     1/1:34,6,0:2:5
chrI    50334   .       A       G       15.1    .       DP=3;VDB=0.0124;AF1=1;AC1=2;DP4=0,0,2,1;MQ=41;FQ=-36    GT:PL:DP:GQ     1/1:47,9,0:3:14
chrI    77885   .       C       A       16.1    .       DP=3;VDB=0.0323;AF1=1;AC1=2;DP4=0,0,1,2;MQ=42;FQ=-36    GT:PL:DP:GQ     1/1:48,9,0:3:15
chrI    121354  .       A       T       4.61    .       DP=2;VDB=0.0216;AF1=1;AC1=2;DP4=0,0,1,1;MQ=41;FQ=-33    GT:PL:DP:GQ     1/1:34,6,0:2:5
chrI    134541  .       TTTGT   TT      4.42    .       INDEL;DP=1;AF1=1;AC1=2;DP4=0,0,1,0;MQ=40;FQ=-37.5       GT:PL:DP:GQ     0/1:40,3,0:1:3
chrI    156862  .       G       A       4.61    .       DP=2;VDB=0.0257;AF1=1;AC1=2;DP4=0,0,1,1;MQ=42;FQ=-33    GT:PL:DP:GQ     1/1:34,6,0:2:5
chrI    169815  .       AC      ACCC    4.42    .       INDEL;DP=1;AF1=1;AC1=2;DP4=0,0,0,1;MQ=40;FQ=-37.5       GT:PL:DP:GQ     0/1:40,3,0:1:3
chrI    181090  .       A       C       4.61    .       DP=2;VDB=0.0143;AF1=1;AC1=2;DP4=0,0,1,1;MQ=42;FQ=-33    GT:PL:DP:GQ     1/1:34,6,0:2:5
```

VCF: The full column list

	Col	Field	Description
*	1	CHROM	Chromosome name
*	2	POS	1-based position. For an indel, this is the position preceding the indel.
	3	ID	Variant identifier. Usually the dbSNP rsID.
*	4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
*	5	ALT	Comma delimited list of alternative seuqence(s).
*	6	QUAL	Phred-scaled probability of all samples being homozygous reference.
	7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
	8	INFO	Semicolon delimited list of variant information.
	9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
	10+	Sample(s)	Individual genotype information defined by FORMAT.

- Variant call confidence
  - like Phred score and MAPQ

http://samtools.sourceforge.net/mpileup.shtml

VCF: Multiple variants

- What if your reads have more than 1 base at one location?
  - wgsim's mutations.txt
    - IUPAC notation
- VCF just gives comma-separated lists
  - REF → ALT
  - A → A,C

mutations.txt

chrI	2118	C	T	-
chrI	2206	A	-	-
chrI	2589	C	Y	+
chrI	3888	G	C	-
chrI	4543	T	K	+
chrI	7636	C	S	+
chrI	8832	T	W	+
chrI	9020	T	K	+
chrI	11646	G	K	+
chrI	11807	C	G	-
chrI	11882	C	Y	+
chrI	14995	-	C	-
chrI	15199	A	-	-

http://www.dna.affrc.go.jp/misc/MPsrch/InfoIUPAC.html

VCF: Complex variants

- Can show short indels
  - C → CT (insert T)
  - ACG → A (delete CG)

Types of variants

SNPs

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
ATGT	2	C	T

Insertions

Alignment	*VCF representation*		
AC-GT	POS	REF	ALT
ACTGT	2	C	CT

Deletions

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
A--T	1	ACG	A

Complex events

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
A-TT	1	ACG	AT

Large structural variants

VCF representation

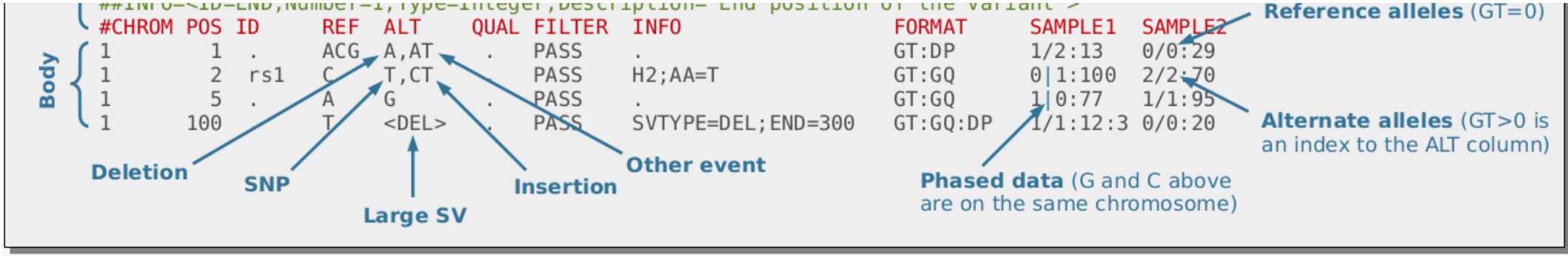
POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

http://vcftools.sourceforge.net/VCF-poster.pdf

VCF: Multiple samples

- VCF can have a variable number of columns!

Col	Field	Description
		...
10+	Sample(s)	Individual genotype information defined by FORMAT.



- Column headings are the sample names

http://vcftools.sourceforge.net/VCF-poster.pdf

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
  - http://vcftools.sourceforge.net/VCF-poster.pdf
- Check full specification for details
  - http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

Samtools can call variants and create a VCF

- Samtools mpileup → BCF
  - BCF is to VCF as BAM is to SAM
    - (roughly)

```
Terminal
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$ samtools mpileup -uDf ../refs/sc.fa sc.bam > sc.bcf
[mpileup] 1 samples in 1 input files
<mpileup> Set max per-file depth to 8000
[Tue Oct 29] me@zen:~/school/bioinformatics/work
```

-u: uncompressed output
-D: include read depth in output
-f: use ../refs/sc.fa as reference

  - The BCF doesn't hold actual calls
    - encodes likelihoods for <u>all</u> variants
- Bcftools view → VCF
  - Performs the actual variant calling

```
Terminal
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$ bcftools view -vcg sc.bcf > sc.vcf
[afs] 0:26910.091 1:634.591 2:608.317
[Tue Oct 29] me@zen:~/school/bioinformatics/work
```

-v: only output non-reference sites
-c: do SNP calling
-g: call genotypes at variant sites

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (2011) 27 (21): 2987-2993.

More mpileup tricks

- Combine multiple BAM files into one BCF

```
Terminal
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$ samtools mpileup -uDf ../refs/sc.fa sc1.bam sc2.bam > sc-total.bcf
[mpileup] 2 samples in 2 input files
<mpileup> Set max per-file depth to 4000
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$
```

- Only include one region

```
Terminal
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$ samtools mpileup -uDf ../refs/sc.fa -r chrI sc.bam > sc-chrI.bcf
[mpileup] 1 samples in 1 input files
<mpileup> Set max per-file depth to 8000
[Tue Oct 29] me@zen:~/school/bioinformatics/work
$
```

Homework 19

- Take your mutations.txt file from wgsim (or create another one) and create a partial VCF file from the first 10 lines (but skip ones with indels)
  - Only the last header line (#CHROM)
  - Only the first 5 columns
  - Refer to IUPAC nucleic acid codes for non-ACGT bases

```
chrI    8832    T       W       +
```

  - means it generated reads with both A and T at this location
- Use samtools/bcftools to create a full VCF file from the alignments you created in the previous homework