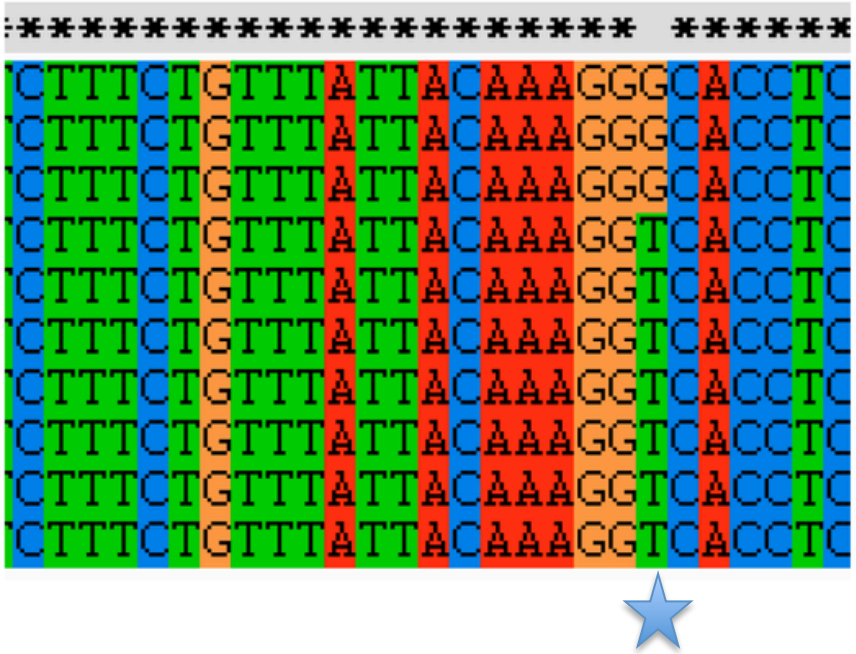
Calling Variants

Variant Calling

- SNP = single nucleotide polymorphism
- SNV = single nucleotide variant
- Indel = insertion/deletion

- Examine the alignments of reads and look for differences between the reference and the individual(s) being sequenced

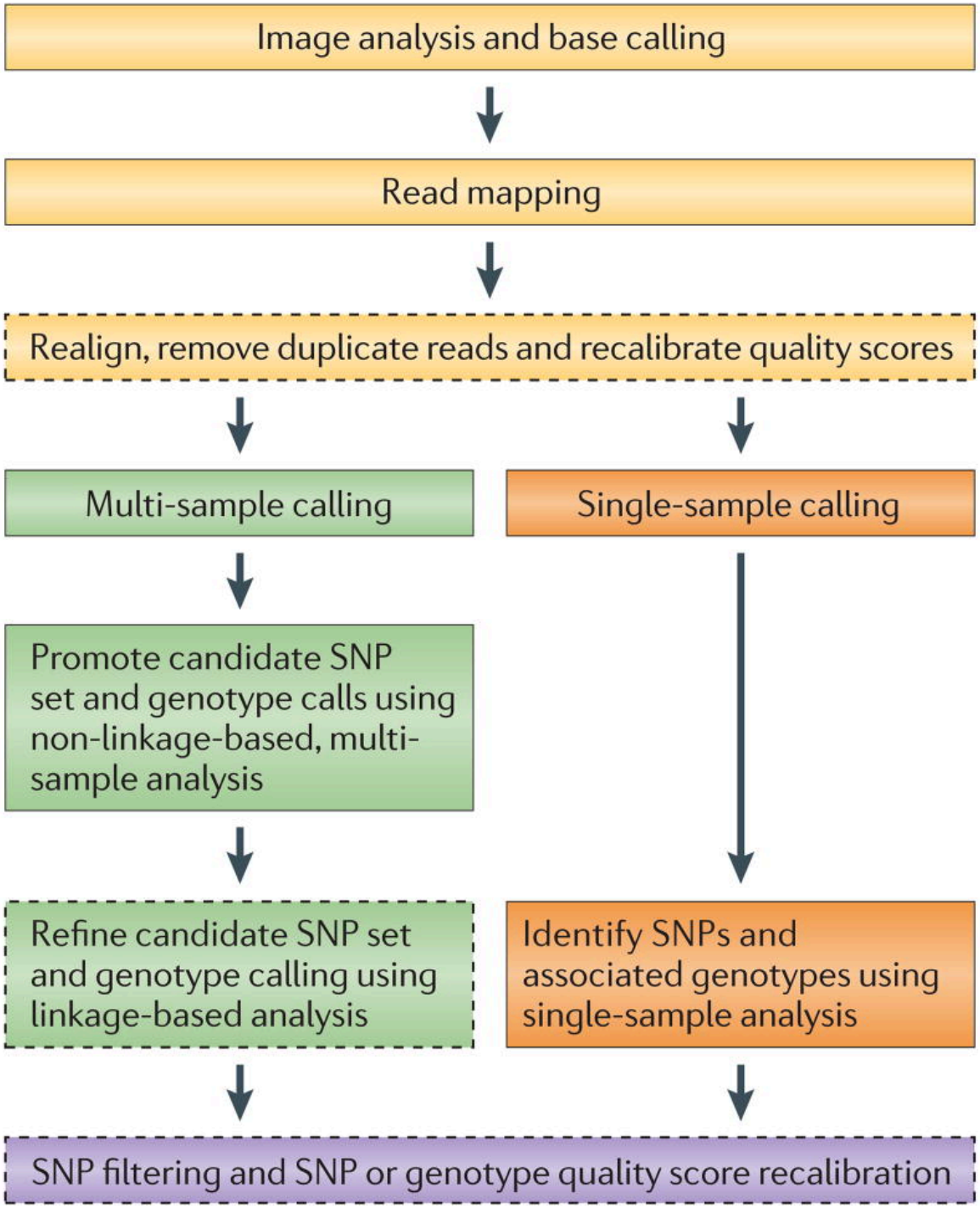


Feng et al 2014. Novel Single Nucleotide Polymorphisms of the Insulin-Like Growth Factor-I Gene and Their Associations with Growth Traits in Common Carp (Cyprinus carpio L.)

May or may not be worth preprocessing: https://bcbio.wordpress.com/2013/10/21/updated-comparison-of-variant-detection-methods-ensemble-freebayes-and-minimal-bam-preparation-pipelines/

Workflow

(Optional steps with dashed lines)



Nielson et al., 2013

Variant Calling Difficulties

- Difficulties:
  - Cloning process (PCR) artifacts
  - Errors in the sequencing reads
  - Incorrect mapping
  - Errors in the reference genome

- Heng Li, developer of BWA, looked at major sources of errors in variant calls*:
  - erroneous realignment in low-complexity regions
  - the incomplete reference genome with respect to the sample

* Li 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics.

Indel

```
coor      12345678901234    567890123456789012345 6
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac****aaAtaa
read2      ggttttataaaac****aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT****aattaagtctacagagcaac
read5                aaT****aattaagtctacagagcaact
read6                  T****aattaagtctacagagcaacta
```

Can be difficult to decide where the best alignment actually is.

Indels are far more problematic to call than SNPs.

*https://bioinf.comav.upv.es/courses/sequence_analysis/snp_calling.html

Step 1. Quality Score Recalibration

- Implemented in Genome Analysis Toolkit (GATK) and SOAPsnp
- Quality scores output by the sequencing machine have systematic error
- Fix by re-estimating the quality score using an empirical analysis of the current dataset
- Uses (supposedly) non-polymorphic sites to calculate the number of mismatches and look for patterns

- Example from GATK documentation:
  - "we can identify that, for a given run, whenever we called two A nucleotides in a row, the next base we called had a 1% higher rate of error. So any base call that comes after AA in a read should have its quality score reduced by 1%"
  - Also tries to examine machine cycles with higher error rates than other cycles

- Biggest problem – you need to already know variants!
  - This is great for human, mouse, etc, but far less helpful if you are the first person to ever call variants in an organism.

http://gatkforums.broadinstitute.org/wdl/discussion/44/base-quality-score-recalibration-bqsr

Step 2. SNP/Variant Calling Software

Many available, most common options:

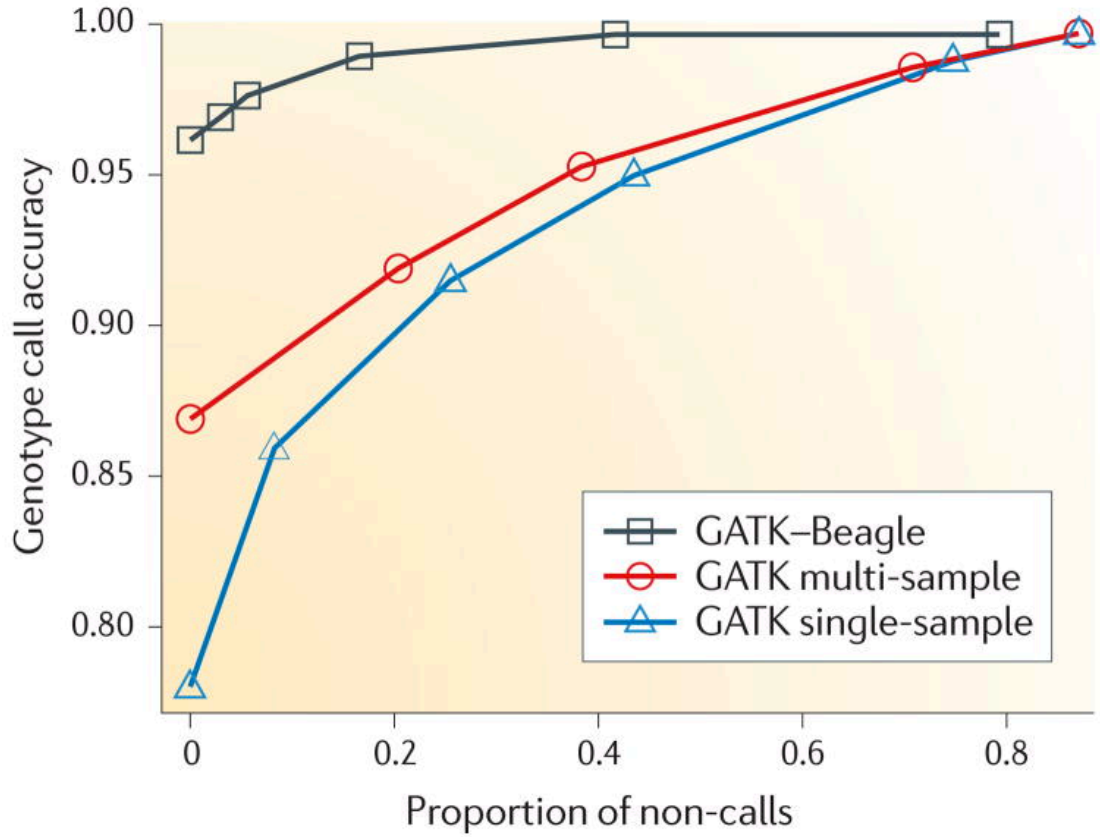
- Samtool's mpileup -> bcftools
- GATK (Genome Analysis Tooklit)
- FreeBayes

- All of these handle SNPs, indels, and other small variants
- All handle data from multiple individuals
- All use Bayesian probabilistic methods
  - Take into account the quality value of the individual base and the quality value for the alignment of the read
  - Can utilize information about previously called/confirmed SNPs
  - Provide measures of statistical uncertainty

Genotype Likelihoods

- Calculates the probability of the observed data given each genotype
- Include the prior probability for that variant (is it a known SNP?)
- Return the most likely genotype
- Probabilistic framework (dependent on software)
  - Errors from base calling
  - Errors from alignment
  - Substitution type
  - Prior information – allele frequences
  - Prior information – LD
- Quality of calls is increased by multiple samples
  - HWE

Nielson et al, 2010



See the math: https://software.broadinstitute.org/gatk/guide/article?id=4442

Step 3. Filtering

- Genotype Likelihood calculation should render this unnecessary, but alas, real data sets often still benefit from additional filtering
- Hard cut off on depth
  - How many reads do you need to sample to confidently call a SNP? (For a diploid?)
  - > 20X = very good
  - 5-20X = okay
  - < 5X = missing many heterozygous calls
- High coverage – can indicate a duplicated region in the genome
- Highly variable region – can also indicate a duplicated region (take into account HWE)
- Low complexity regions

Your mileage may vary

- Different decisions about how to align reads and identify variants can yield very different results

Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, Zhi Wei, Kai Wang and Gholson J Lyon

Genome Medicine 2013 5:28 | DOI: 10.1186/gm432

- 5 pipelines
- "SNV concordance between five Illumina pipelines across all 15 exomes was 57.4%, while 0.5 to 5.1% of variants were called as unique to each pipeline. Indel concordance was only 26.8% between three indel-calling pipelines"

Last step (?): Imputation

If one site has low coverage but is tightly linked to other sites with high coverage, the information can be "imputed"

Rescue missing data!

- Utilize LD across loci (i.e. known haplotypes)
- Depends on haplotype estimation (phasing)
- Many software options
  - BEAGLE
  - Impute2
  - MaCH

Phasing

Heterozygous genotypes at 3 sites

AC TG AT

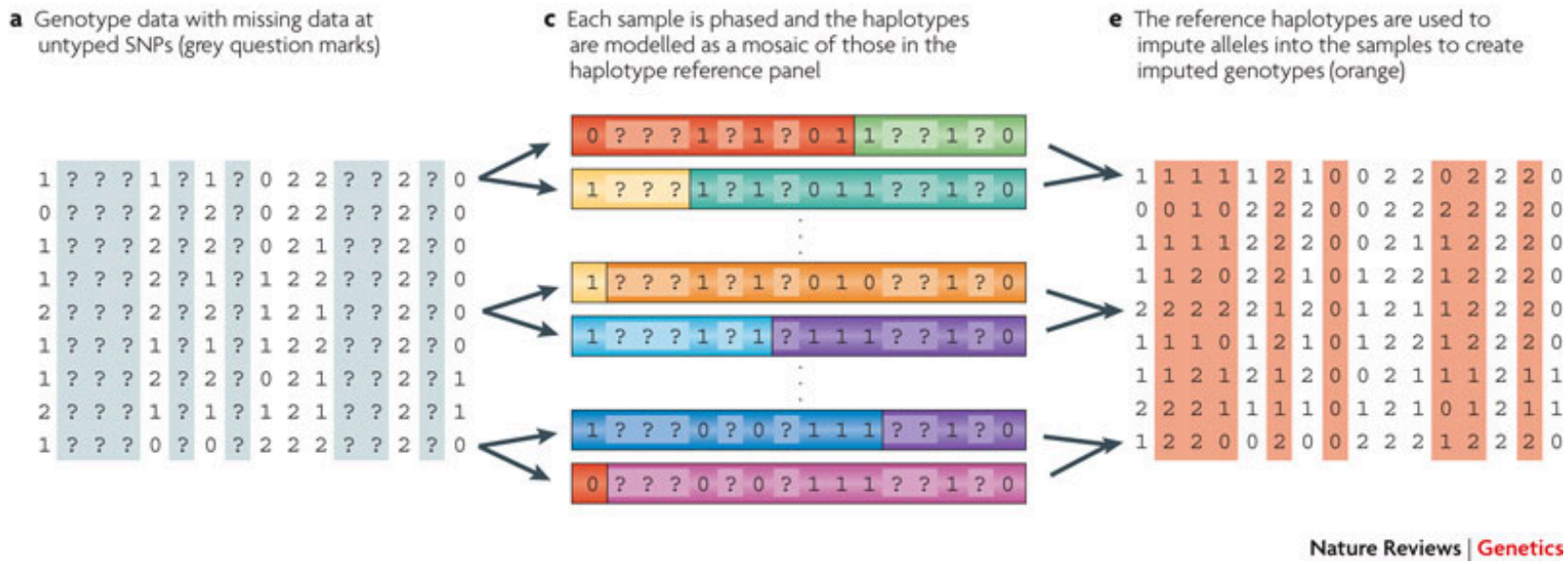
The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Reference Haplotypes

Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0



Review article: Genotype imputation for genome-wide association studies. Marchini and Howie 2010.

Novel SNPs/Indels

- What is the effect of this variant?
- Is the variant inside a gene?
  - Does it change an amino acid?
  - Does it create a stop codon?
  - Does it shift the open reading frame?

- Software:
  - SnpEff/SnpSift
  - Annovar
  - Variant Effect Predictor

Choice of Transcripts and Software has a large effect on variant annotation
McCarthy et al., 2014

"When comparing results from Annovar and VEP using Ensembl transcripts, matching annotations were seen for only 65% of loss-of-function variants and 87% of all exonic variants, with splicing variants revealed as the category with the greatest discrepancy"

HTSLIB/SAMTOOLS/BCFTOOLS

SAMtools, BCFtools, HTSLib

- http://www.htslib.org/
- Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

1. Samtools

   Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
2. BCFtools

   Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarizing SNP and short indel sequence variants
3. HTSlib

   A C library for reading/writing high-throughput sequencing data

- Example workflow:
- http://www.htslib.org/workflow/#mapping_to_variant

samtools

- View – print alignments to your screen or convert between formats. Can reduce files to a particular region only
- Tview - text alignment viewer, nifty for quick viewing of files
- Mpileup – generates a special mpileup formatted file needed for calling variants
- Sort – sort the alignments (by default, sorts by coordinate). Sorting is needed for most downstream applications.
- Merge – concatenate bam files together, while maintaining sorting order
- Index - index a bam or cram file, needed for most downstream applications
- Idxstats – get some stats about your bam file
- Faidx - index a fasta file, need for most downstream applications using a bam file
- Bam2fq – convert a bam file to a fastq file
- More…

Always the format

Samtools subcommand –flags –moreflags

Mpileup format

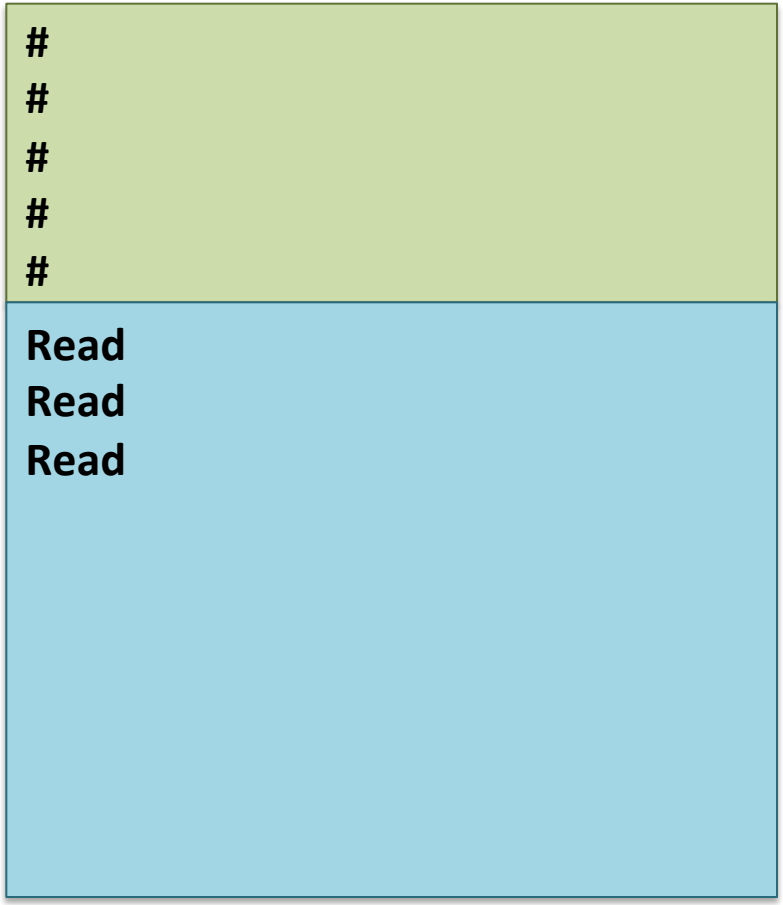
- Mpileup format
- For each base in the reference
  - reference base
  - the number of reads covering the site
  - read bases
  - base qualities
  - alignment mapping qualities
- You will rarely ever use this format, just need to generate it and pass it straight to the SNP caller

bcftools

- BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF.
- Ack, more formats!!!

VCF

- Variant Call Format
- Official spec: http://samtools.github.io/hts-specs/VCFv4.2.pdf
- Header lines starting with # signs
- Lines with variants afterward



VCF (cont)

- Tab delimited fields
  - Chromosome
  - Location
  - ID (if this is a named variant)
  - Reference sequence
  - Alternate sequence
  - Quality score
  - Filter (true/false – whether or not it passed filtering)
  - Info – lots of additional info such as CIGAR string, depth across different samples, etc.
  - Columns follow for each genotype if available
- BCF is the compressed binary format
  - SAM <-> BAM
  - VCF <-> BCF

Quite variable depending on software used to call SNPs

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

Standard

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

Info field gives general information about this position across all samples. The codes are defined in the header of the file, can vary.

NS = Number of samples with data

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

DP = combined depth across samples

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

AF = allele frequence for alternate allele

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

DB = dbSNP membership

H2 = HapMap2 membership

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

Format field

Explains the format used for information about each sample.

Variable by SNP caller.

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

GT = genotype
0/0 0/1 1/1 1/2

The / is replaced with a | if the alleles are phased

0|0 0|1 1|1

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

GQ = Genotype Quality

Phred-scaled confidence in genotype call

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

DP = Read Depth

# of reads from this location for this individual

VCF Example

```
#CHROM      20
POS         14370
ID          rs6054257
REF         G
ALT         A
QUAL        29
FILTER      PASS
INFO        NS=3;DP=14;AF=0.5;DB;H2
FORMAT      GT:GQ:DP:HQ
NA00001     0|0:48:1:51,51
NA00002     1|0:48:8:51,51
NA00003     1/1:43:5:.,.
```

HQ = Haplotype Quality

Only for phased loci, added by phasing software

Flexible info fields

- SNPEff has standardized the addition of variant effect information
- Additional tag ANN in the info field

Chromosome 1411926 . G C 228.0 PASS DP=97;VDB=1.42407e-36;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,45,41;MQ=60;ANN=C|missense_variant|MODERATE|ttcA|b1344|transcript|AAC74426.1...

bcftools

- Okay, now that we know what VCF and BCF are, what does bcftools do?
- Will call SNPs!

- Call – SNP/indel calling
- CNV – copy number variation caller
- Concat – merge VCF files together
- Consensus –resequenced an individual and generate the reference sequence for that individual
- Filter – filter the variants by quality
- Stats - statistics
- Convert – convert between formats

Overview

Samtools	• Works with SAM/BAM files • Produces mpileup	Alignment Data
Bcftools	• Call SNPs from mpileup • Works with VCF/BCF files	Variant Data

IGV

- high-performance visualization tool for interactive exploration of large, integrated genomic datasets
- Run on local computer

- Visualizes lots of data types
  - NGS read alignments
  - Gene annotation
  - Variants
  - Etc.

http://www.broadinstitute.org/igv/