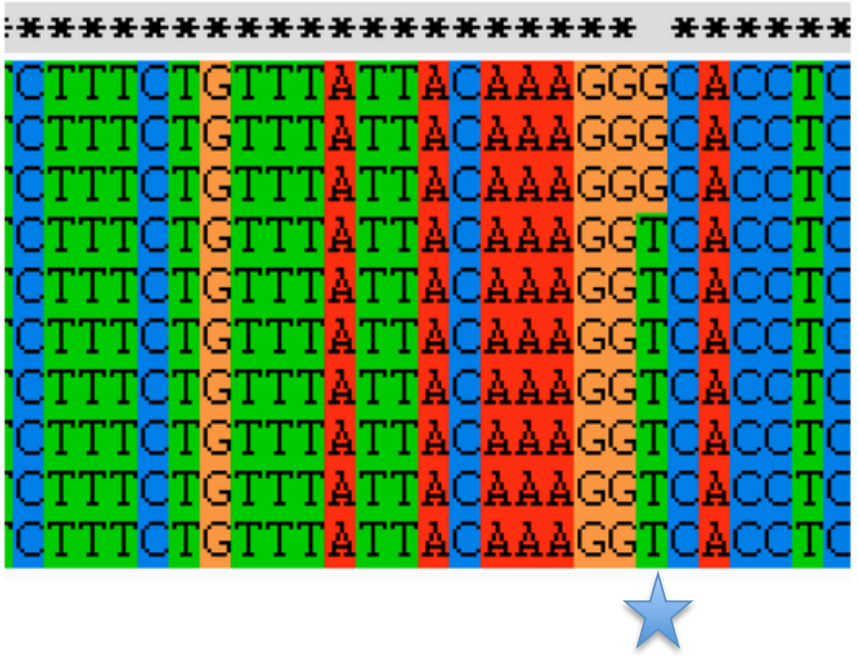
Calling Variants

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.)

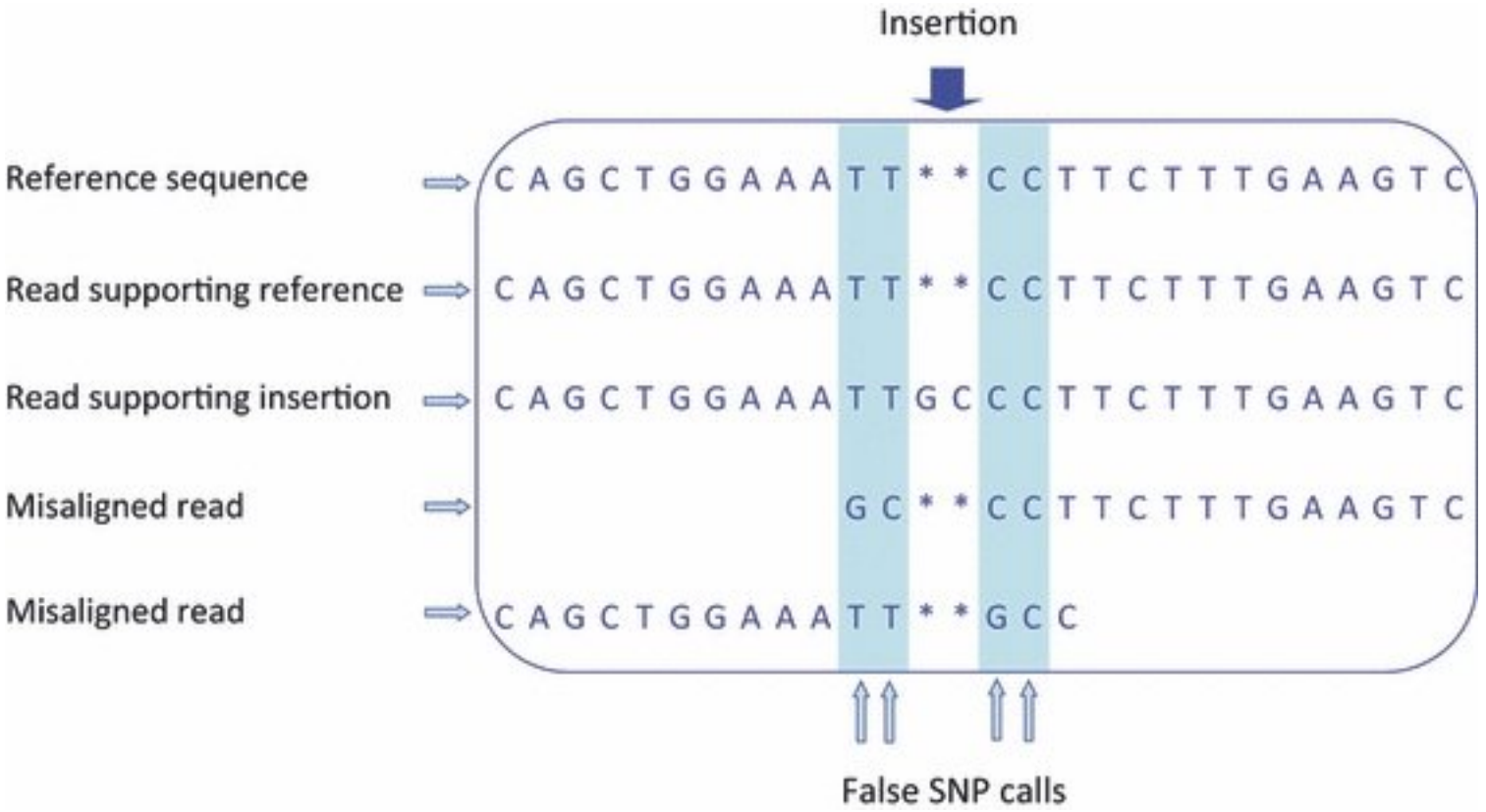
Variant Calling Difficulties

- 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.

Example of misalignment



The effect of next-generation sequencing technology on complex trait research. Day-Williams and Zeggini, 2010

Another example from the Li paper

Indels are far more problematic to call than SNPs.

- Genome Analysis Toolkit
  - Open source
  - Originally published in 2010
  - Continues to expand and improve
  - Complex but worthwhile
  - January 2018 − GATK4
    - Runs on the cloud!
    - New features
    - covers all major variant classes (SNPs, indels, copy number, and structural variation)
    - both germline and somatic mutations (ie. cancer)
    - for genomes and targeted sequencing assays

- https://software.broadinstitute.org/gatk/gatk4

User Guide

Very complete. Much documents. So quality. WOW.

QUICK START GUIDE
Take a brief orientation tour and get started today

WHAT'S IN THE BOX

Tool Documentation
Usage documentation for each tool

Methods and Algorithms
Analysis details and recommendations

BEFORE YOU ASK

Frequently Asked Questions
Questions that many people have asked

Solutions to Problems
Tips for solving common problems and errors

HIGH-LEVEL VIEW

Best Practices
Workflows for variant discovery analysis

Presentations
Materials from workshops and conferences

NEWBIE ZONE

Tutorials
Step-by-step instructions targeted by use case

Dictionary
Definitions of terms used in the docs

CHANGE YOU CAN BELIEVE IN

Version History
Historical record of changes by version

Bugs & Feature Requests
Known issues and requested enhancements

RUNNING GATK AT SCALE

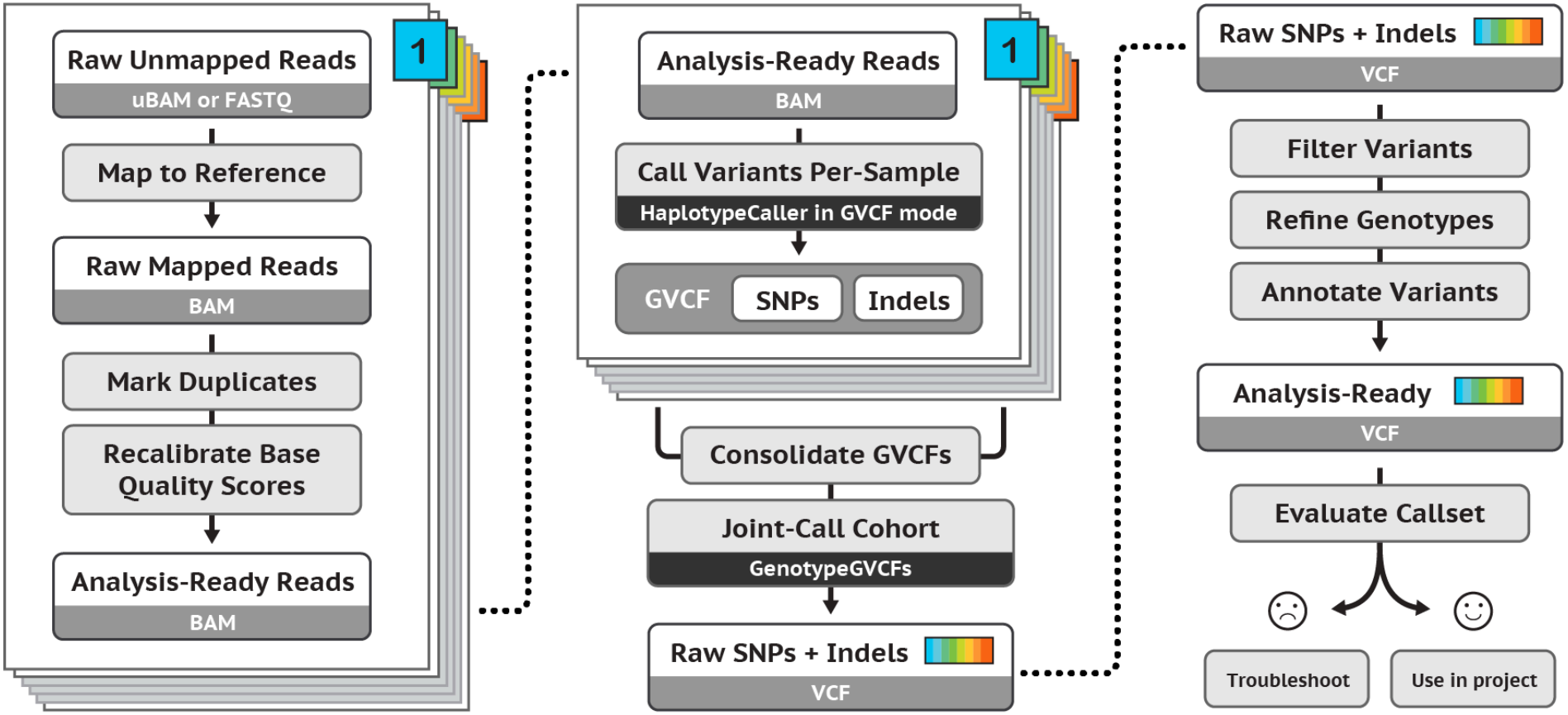
Pipelining Options
Tools and scripts for analysis pipelines

GATK on FireCloud
A secure and open cloud-based analysis portal

Note that the information in this documentation guide is targeted at end-users. For developers, the source code and related resources are available on GitHub.

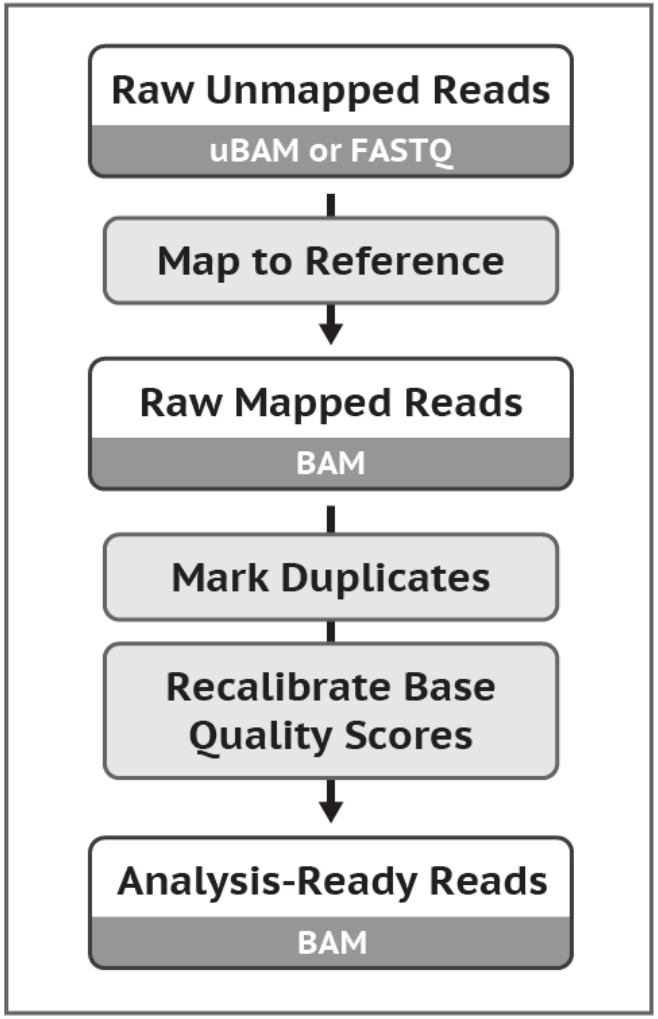
☑ Germline short variant discovery (SNPs + Indels)

Best Practices Workflows | Created 2018-01-07 | Last updated 2018-07-26



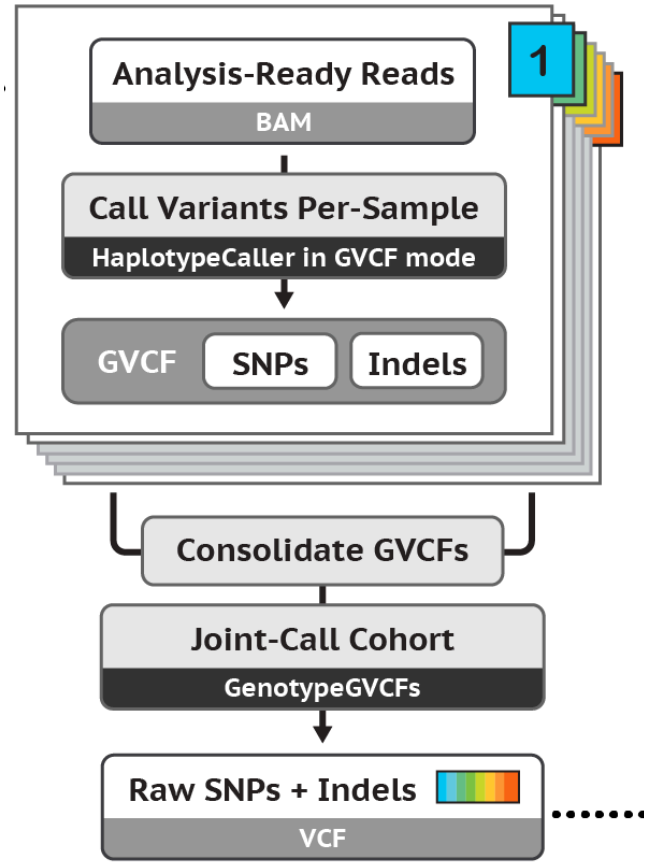
https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145

Data Preprocessing



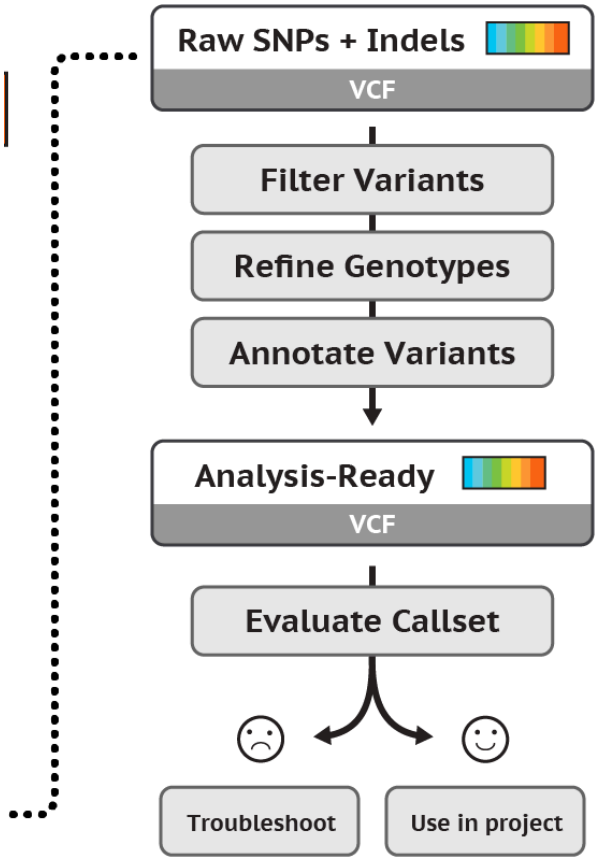
- BWA listed online as their suggested mapper
- MarkDuplicates – identify reads from artifacts
- Quality Recalibration - applying machine learning to detect and correct for patterns of systematic errors

Calling Variants



- Haplotype Caller – super smart variant caller, does local haplotype assemblies
- (basically anywhere there are signs of variation, it throws away the alignments and starts over)
- We'll learn about VCFs and GVCFs in a minute (they're file formats)
- GenotypeGVCFs – use information from all samples to increase sensitivity

Filtering Variants



- VariantRecalibrator – use machine learning to identify variants that are likely to be real, and assigns a more reliable quality score
- Requires a lot of data AND database of known variants
- Won't work on organisms without a lot of high quality known variants, on small datasets or on some targeted/RNASeq datasets

- There are other ways to filter (we'll see bcftools filtering during lab)

GATK3 Lab

- Still an ok pipeline (but you should upgrade)
- We'll do indel realignment, which is no long necessary if you use a special variant caller that performs a haplotype assembly step
- IE follow the newest GATK best practices and your variant analysis will be great

Genotype Likelihoods

- Calculates the probability of the observed data given each genotype
- Usually (but not always) phred-scaled (PL)
- In the case of GATK, this is a Bayesian calculation
- Take into account:
  - Mapped reads
  - Quality value of bases
  - prior probability for that variant (is it a known SNP?)
- Return the most likely genotype
- Quality of calls is increased by multiple samples

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

Step 3. Filtering

- 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)

More options

- Freebayes
- Samtools/Bcftools
- SNVer
- Platypus
- VarScan
- VarDict

Specific for reduced representation data (GBS/RAD)

- Fast-GBS
- TASSEL-GBS
- IGST

Specific for reduced representation data (GBS/RAD) *without a reference genome*

- UNEAK
- Stacks

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

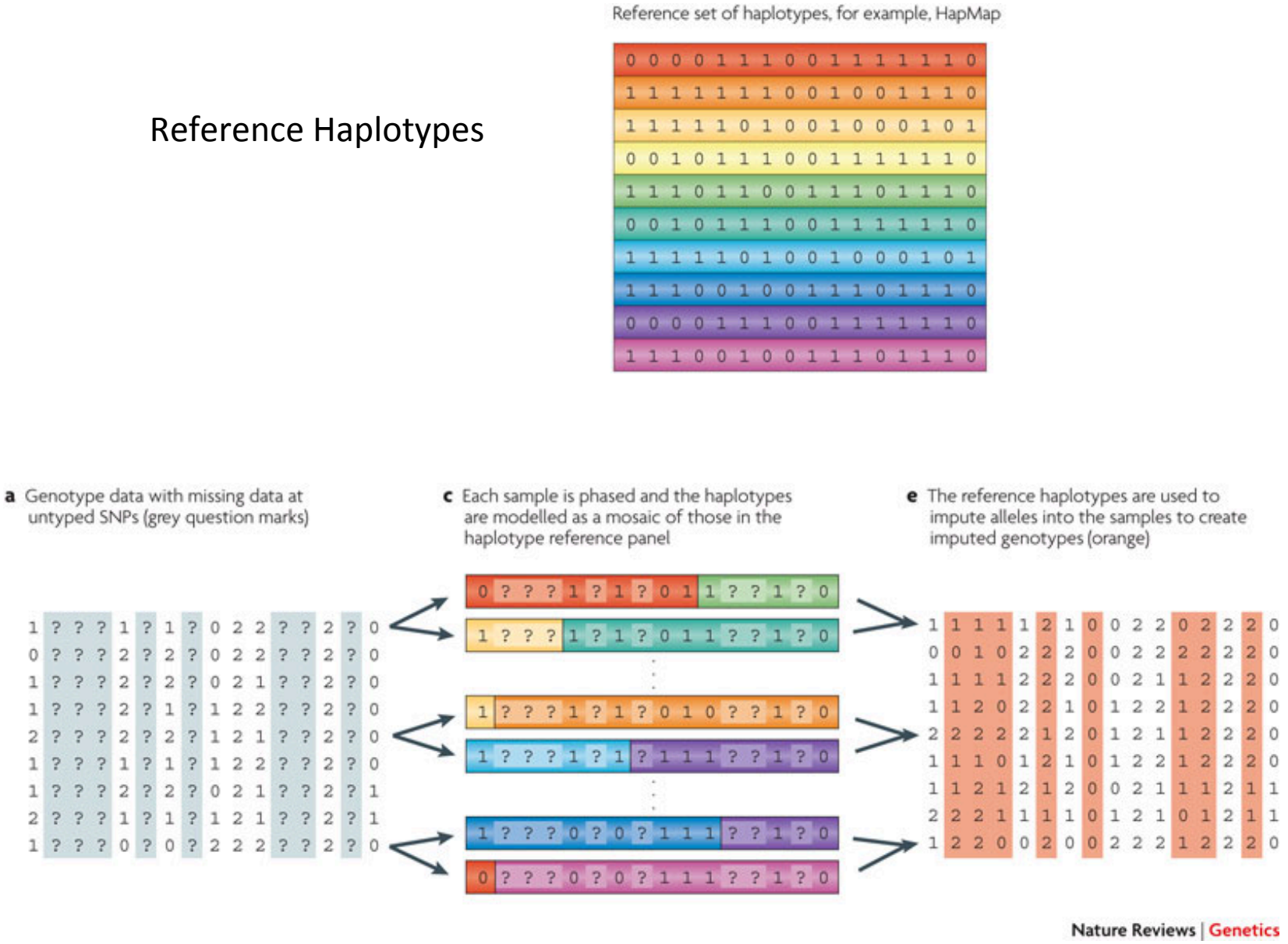
Heterozygous genotypes at 3 sites

AC TG AT

The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Reference Haplotypes



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

SNPs/Indels – what are they doing? Anything interesting?

- 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

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 –anotherflag paraemter -yetanotherflag

Mpileup format

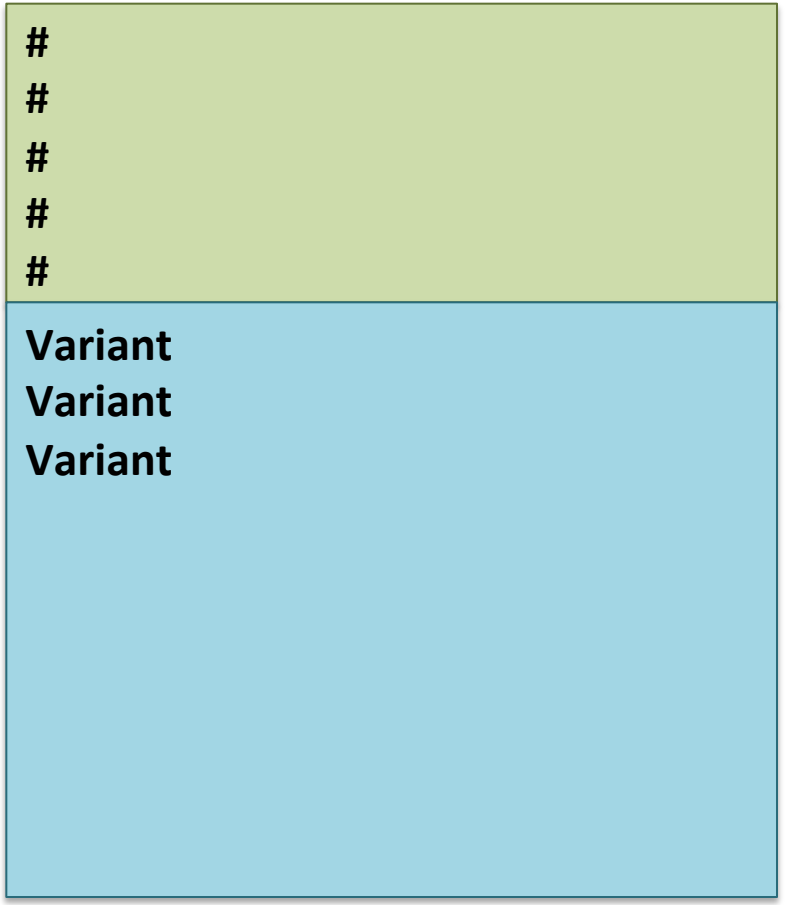
- Mpileup format
- For each base in the reference
  - reference base
  - the number of reads covering the site
  - read bases
  - base qualities
  - alignment mapping qualities
- You will rarely ever use this format, just need to generate it and pass it straight to the SNP caller

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/VCFv4.3.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

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

bcftools

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

- Call – SNP/indel calling
- 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/