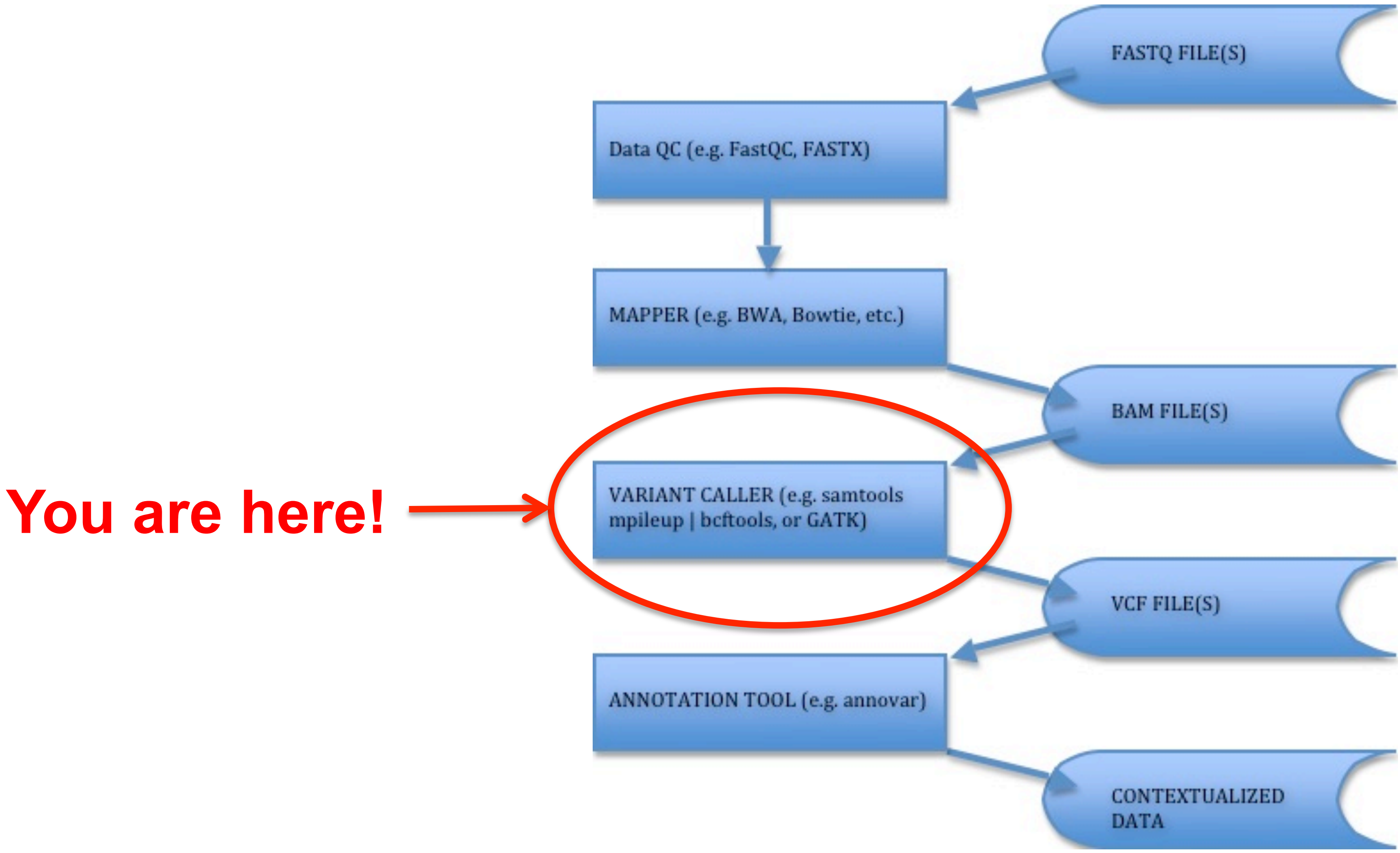
Introduction to Variant Calling



FASTQ FILE(S)

Data QC (e.g. FastQC, FASTX)

MAPPER (e.g. BWA, Bowtie, etc.)

BAM FILE(S)

VARIANT CALLER (e.g. samtools mpileup | bcftools, or GATK)

VCF FILE(S)

ANNOTATION TOOL (e.g. annovar)

CONTEXTUALIZED DATA

You are here!

Some Terminology

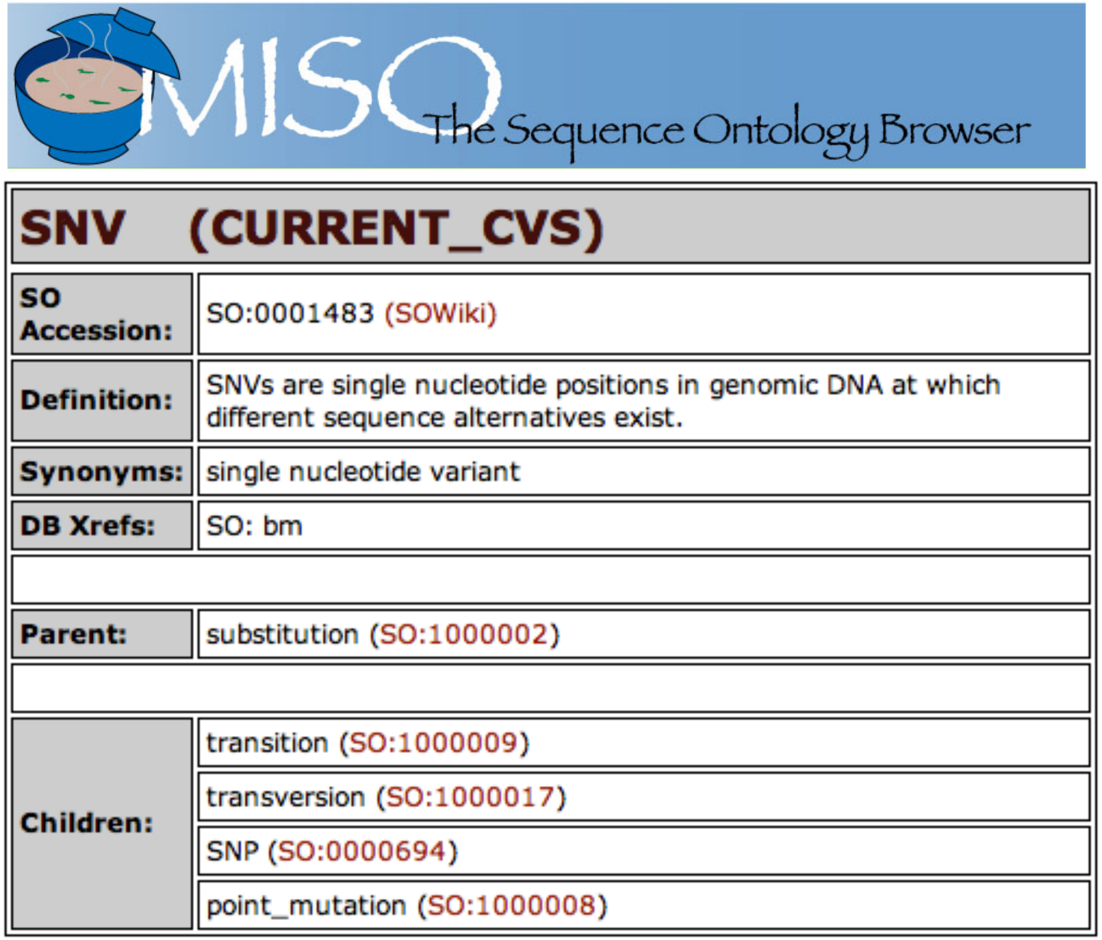
- **Variant** – sequence data difference that exists between individuals in a population.
- **Mutation** – molecular event that created a variant.
- **Allele** – alternative state of a sequence variant.
- **Genotype** – allelic state in a specific individual.
    - AA homozygous or AT heterozygous at specific base
    - Examples: Ara^{+} Lac$^{–}$ *E. coli*, ob/ob mice
    - "20 mice were genotyped for the Klrd1$^{DBA/2J}$ allele."
- **Polymorphism** – sequence variation that is common within a population.
    - "SNP on chromosome 16 associated with obesity"

Types of Genome Sequence Variants

1. Single Nucleotide Variants (SNVs) ∗
   – Single base changes, e.g., A→T.
2. Insertions-Deletions (Indels; DIPs) ∗ ▶
   – Consisting of one or a few bases, e.g., +ATGA, ΔT.
3. Structural Variants (SVs) ▶
   – Everything else: large deletions, insertions, duplications, inversions, translocations, mobile element insertions, horizontal gene transfer

Different sequencing information and different algorithms are used to predict each kind of variant.

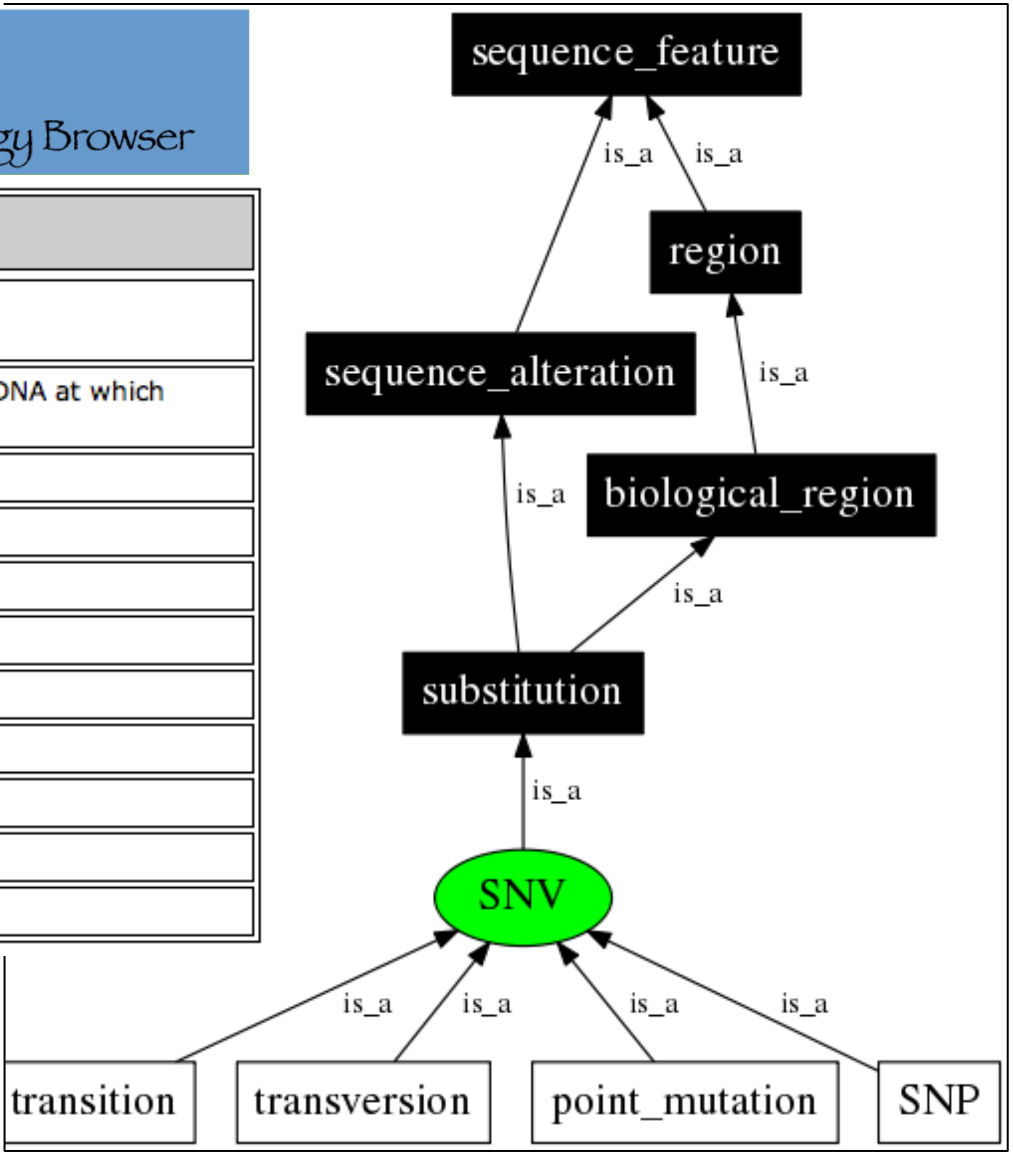
Sequence Ontology



MISO The Sequence Ontology Browser

SNV (CURRENT_CVS)

SO Accession:	SO:0001483 (SOWiki)
Definition:	SNVs are single nucleotide positions in genomic DNA at which different sequence alternatives exist.
Synonyms:	single nucleotide variant
DB Xrefs:	SO: bm
Parent:	substitution (SO:1000002)
Children:	transition (SO:1000009)
	transversion (SO:1000017)
	SNP (SO:0000694)
	point_mutation (SO:1000008)



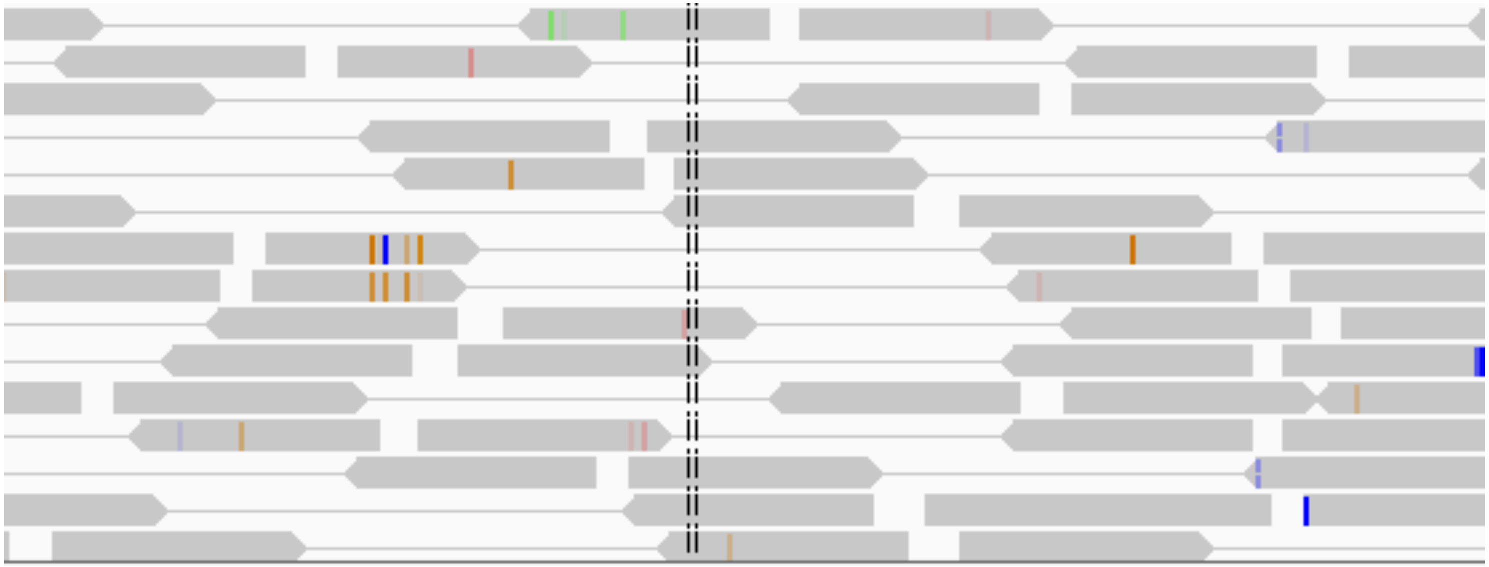
http://www.sequenceontology.org/browser/

Hierarchical controlled vocabulary like the Gene Ontology (GO terms).

Mapped Reads to Variants

BAM/SAM databases make it easy to iterate over mapped data in two ways:

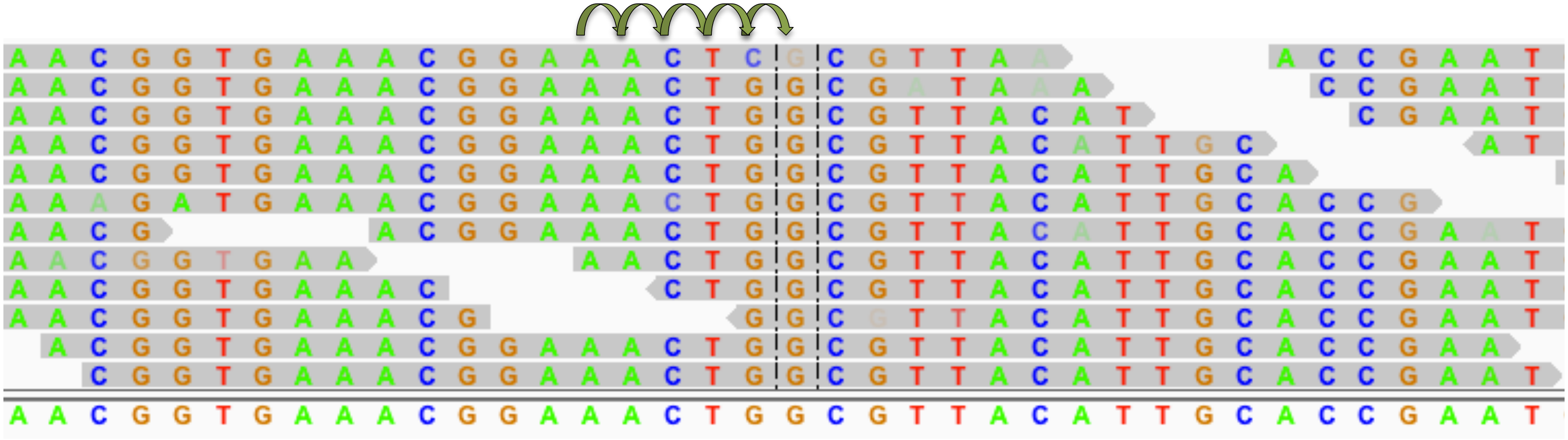
1. Information about each read or read pair



Mapped Reads to Variants

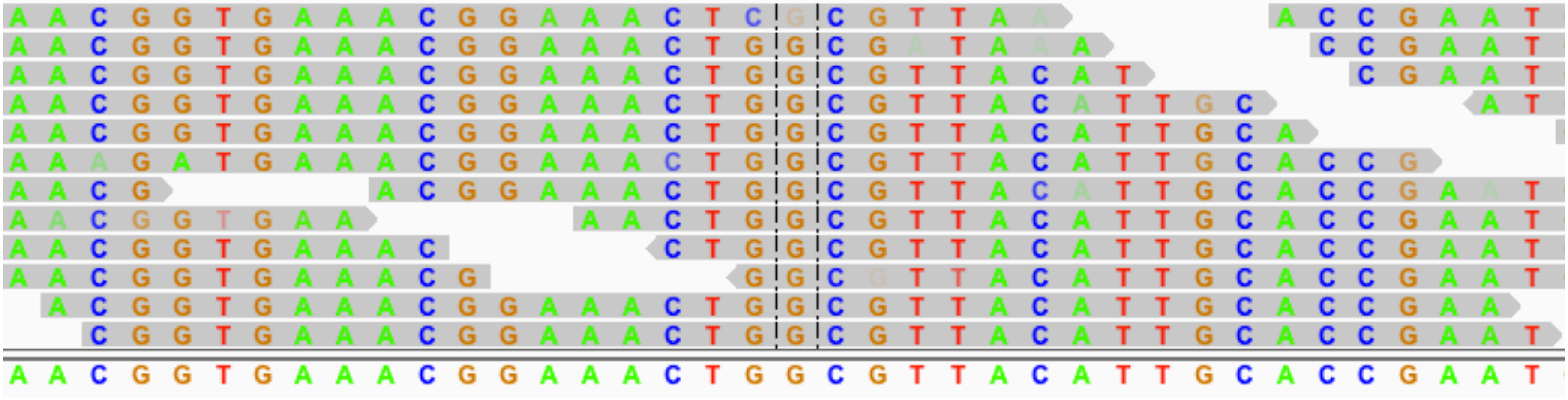
BAM/SAM databases make it easy to iterate over mapped data in two ways:

1. Information about each mapped read
2. Bases mapped to each reference column


A A C G G T G A A A C G G A A A C T G G C G T T A C A T T G C A C C G A A T

SAMtools mpileup

- Processes one or multiple samples aligned to the same reference simultaneously to tabulate all information about how reads are aligned to each reference column.


A A C G G T G A A A C G G A A A C T G G C G T T A C A T T G C A C C G A A T

- Output is in variant call format (VCF).

Variant Call Format (VCF)

- Like SAM, VCF is "human readable" with a fixed number of columns followed by optional key=data fields.
- VCF files can contain information about variation across a number of samples (not just one).

Fixed fields in VCF format:

CHROM chromosome **POS** 1-indexed position **ID** unique identifiers

REF reference base(s): Each base must be one of A,C,G,T,N. Bases should be in uppercase. Multiple bases are permitted.

ALT comma separated list of alternate non-reference alleles called on at least one of the samples.

QUAL Phred-scaled quality score for the assertion made in ALT

FILTER filter: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated list of codes for filters that fail. e.g. “q10;s50”

INFO additional fields encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data].

Working with VCF files

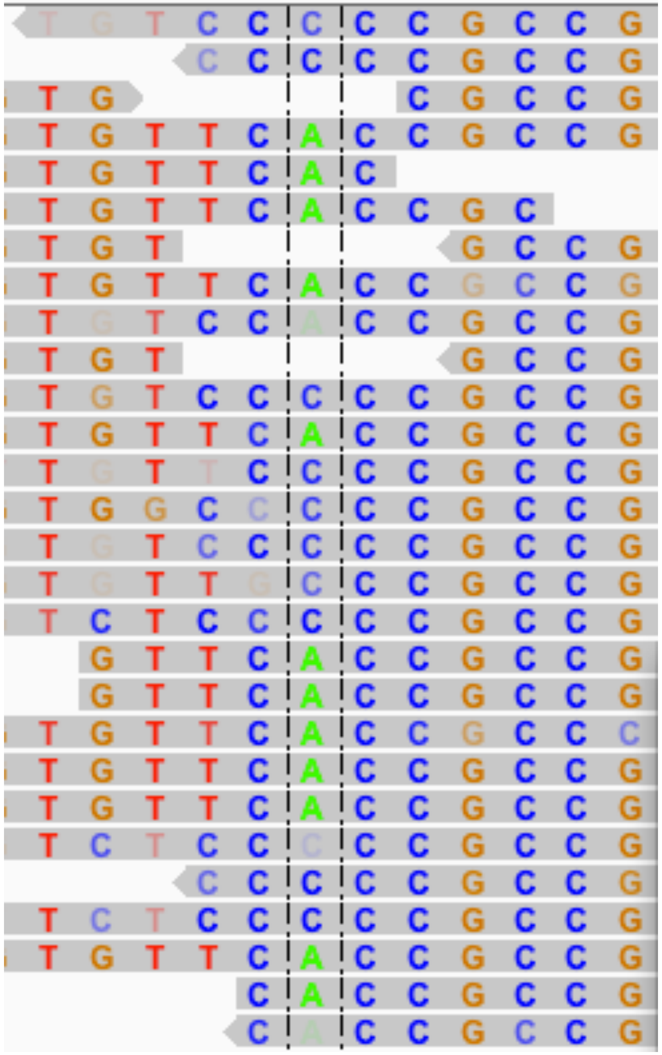
You guessed it: SAM is to BAM as VCF is to ______.

bcftools

- Part of SAMtools, with similar command structure.
- Can convert, filter, and index BCF and VCF files.
- Used to call variants from raw "data-dump" output from `samtools mpileup`.

```
#CHROM       POS      ID  REF ALT QUAL    FILTER
     INFO   FORMAT  samtools_bwa/SRR030257.sorted.bam
NC_012967   161041  .   T   G,X 0   .
    DP=60;I16=0,0,24,32,0,0,1595,46069,0,0,1120,22400,0,0,590,7460;
    VDB=0.0839 PL  212,169,0,212,169,212
NC_012967   161041  .   T   G   222 .
    DP=62;VDB=0.0626;AF1=1;AC1=2;DP4=0,0,24,34;MQ=52;FQ=-202
    GT:PL:GQ   1/1:255,175,0:99
```

Bayesian variant calling?



- Prior probabilities
  Beliefs sans evidence
  - Equal probability for each base (25% for all bases)
  - Only a small number of changes from reference genome (99.99% ref base)
- Update **beliefs** given evidence (aligned bases) according to Bayes' rule:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Example of Updating Priors

- We initially believe that a given base is **C** with 97% probability:

$P_0(\text{ref } \mathbf{A}) = 0.01 \quad P_0(\mathbf{T}) = 0.01 \quad P_0(\mathbf{C}) = 0.97 \quad P_0(\mathbf{G}) = 0.01$

- If we observe that a base (B) with Q=30 in a read mapped to the position is an **A**, what should we believe now?

$P(\text{read B}|\text{ref } \mathbf{A}) = (1–10^{–Q/10}) = 0.999$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$P(B|\mathbf{T}) = P(B|\mathbf{C}) = P(B|\mathbf{G}) = ⅓10^{–Q/10} = 0.00033$

$P(B) = P(B|\mathbf{A})P(\mathbf{A}) + P(B|\mathbf{T})P(\mathbf{T}) + P(B|\mathbf{C})P(\mathbf{C}) + P(B|\mathbf{G})P(\mathbf{G})$
$= 0.0103167$

- Updated probabilities

$P_1(\text{ref } \mathbf{A}) = P(\mathbf{A}|B) = (0.999)\ (0.01)\ /\ 0.0103169 = 0.968$

$P_1(\mathbf{T}) = 0.00032 \qquad P_1(\mathbf{C}) = 0.031 \qquad P_1(\mathbf{G}) = 0.00032$

SAMtools mpileup-bcftools

- Real variant callers consider mapping quality and better error probabilities in these calculations.
- Can process multiple samples aligned to the same reference simultaneously (integrating information about the error model to call variants in all of them).
- Perform Bayesian SNV and indel variant calling and a slew of other calculations to spot systematic errors.

  –read stand bias –base quality bias –mapping quality bias
  –base alignment quality (BAQ) –coverage cutoff

- Favor sensitivity (recovery of true positives) over specificity, typically leaving many false-positives for you to filter.

Variant Call Format (VCF)

Example of a simple VCF file after bcftools (lines are wrapped):

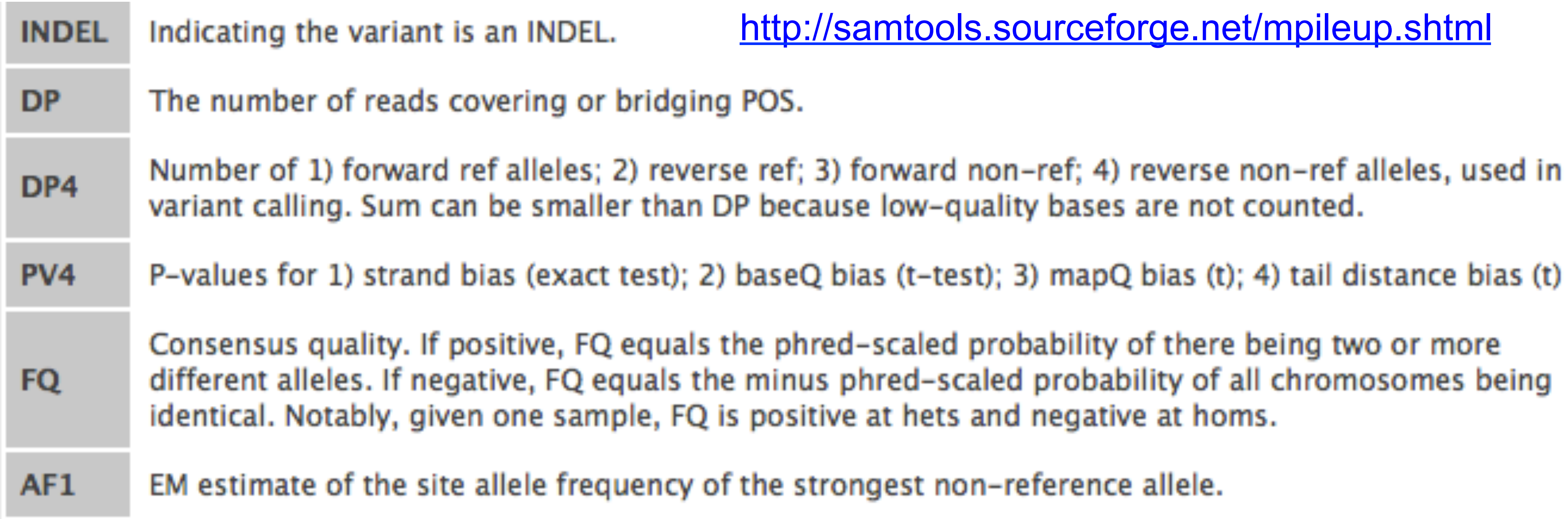
```
#CHROM    POS       ID   REF  ALT  QUAL FILTER
     INFO FORMAT    samtools_bwa/SRR030257.sorted.bam
NC_012967 33801     .    T    G    5.46 .
     DP=47;VDB=0.0423;AF1=0.4999;AC1=1;DP4=6,16,6,1;MQ=53;FQ=7.8;PV4=0.011,0.00019,4.1e-07,1
     GT:PL:G0/1:34,0,227:34
NC_012967 90953     .    T    G    13.2 .
     DP=65;VDB=0.1016;AF1=0.5;AC1=1;DP4=8,29,18,0;MQ=50;FQ=16.1;PV4=1.1e-08,1e-08,2.9e-05,1
     GT:PL:G0/1:43,0,236:46
NC_012967 92359     .    G    T    4.77 .
     DP=48;VDB=0.0258;AF1=0.4999;AC1=1;DP4=4,23,9,0;MQ=54;FQ=6.99;PV4=7.6e-06,6.7e-06,0.012,1
     GT:PL:GQ  0/1:33,0,205:33
NC_012967 139812    .    G    T    21   .
     DP=56;VDB=0.0071;AF1=0.5;AC1=1;DP4=4,27,15,0;MQ=53;FQ=24;PV4=7.6e-09,1.4e-05,1.8e-11,1
     GT:PL:G0/1:51,0,201:54
NC_012967 161041    .    T    G    222  .
     DP=62;VDB=0.0626;AF1=1;AC1=2;DP4=0,0,24,34;MQ=52;FQ=-202    GT:PL:GQ  1/1:255,175,0:99
NC_012967 165565    .    A    C    16.1 .
     DP=35;VDB=0.0423;AF1=0.5;AC1=1;DP4=8,0,1,5;MQ=51;FQ=19.1;PV4=0.003,4.1e-05,3.9e-05,0.43
     GT:PL:G0/1:46,0,107:49
```

Tons of information specific to the variant caller is jammed into the INFO fields. This is useful for filtering.

mpileup-bcftools specific INFO

```
#CHROM       POS      ID  REF ALT QUAL    FILTER
     INFO   FORMAT  samtools_bwa/SRR030257.sorted.bam
NC_012967   90953    .   T   G   13.2    .
    DP=65;VDB=0.1016;AF1=0.5;AC1=1;DP4=8,29,18,0;MQ=50;FQ=16.1;
    PV4=1.1e-08,1e-08,2.9e-05,1   GT:PL:G     0/1:43,0,236:46
NC_012967   161041  .   T   G   222 .
    DP=62;VDB=0.0626;AF1=1;AC1=2;DP4=0,0,24,34;MQ=52;FQ=-202
    GT:PL:GQ    1/1:255,175,0:99
```

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



INDEL	Indicating the variant is an INDEL.
DP	The number of reads covering or bridging POS.
DP4	Number of 1) forward ref alleles; 2) reverse ref; 3) forward non-ref; 4) reverse non-ref alleles, used in variant calling. Sum can be smaller than DP because low-quality bases are not counted.
PV4	P-values for 1) strand bias (exact test); 2) baseQ bias (t-test); 3) mapQ bias (t); 4) tail distance bias (t)
FQ	Consensus quality. If positive, FQ equals the phred-scaled probability of there being two or more different alleles. If negative, FQ equals the minus phred-scaled probability of all chromosomes being identical. Notably, given one sample, FQ is positive at hets and negative at homs.
AF1	EM estimate of the site allele frequency of the strongest non-reference allele.