
# BIO306: Bioinformatics

## Lecture 7

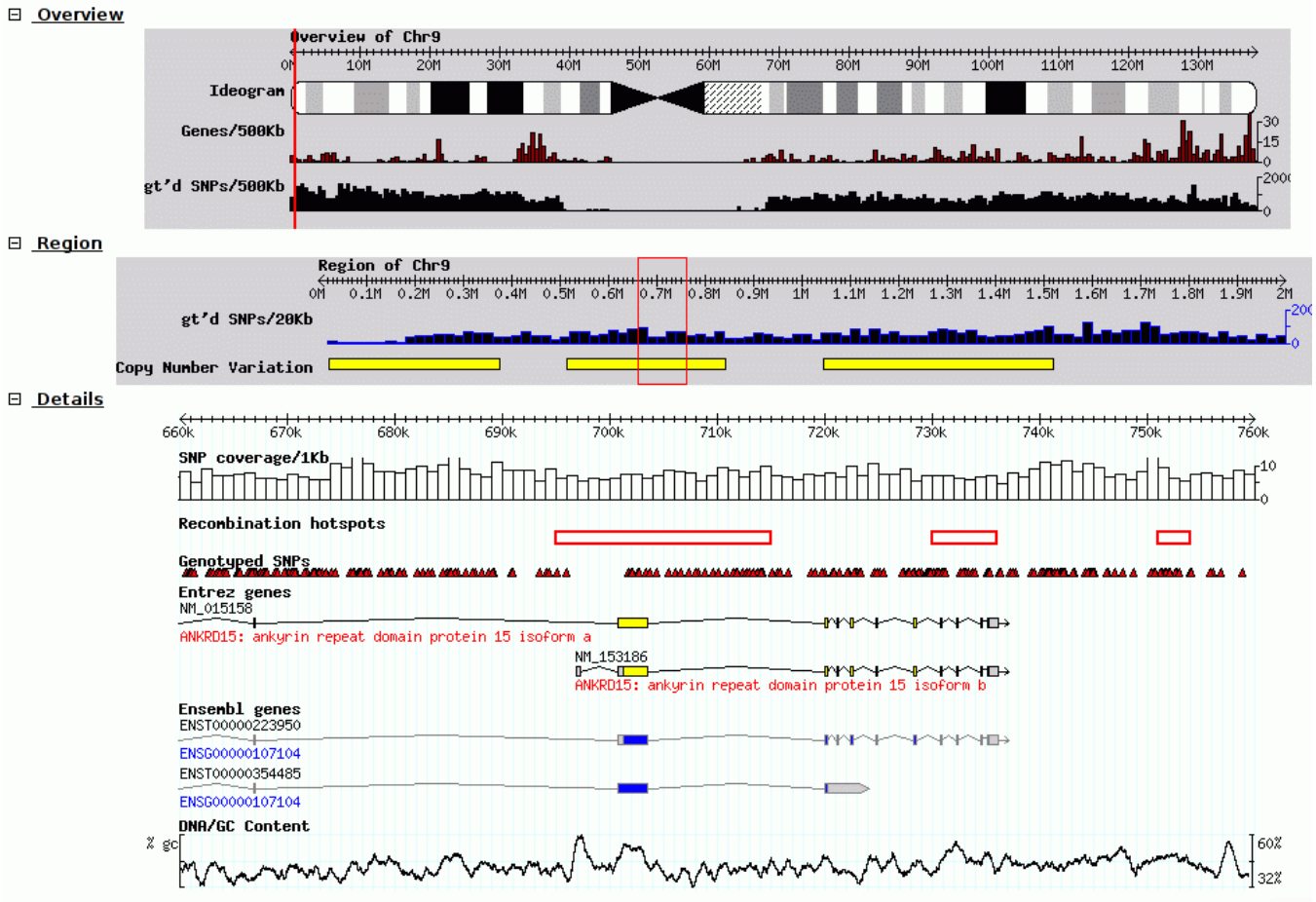
## Variant calling and output format

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

# Genomics data visualization

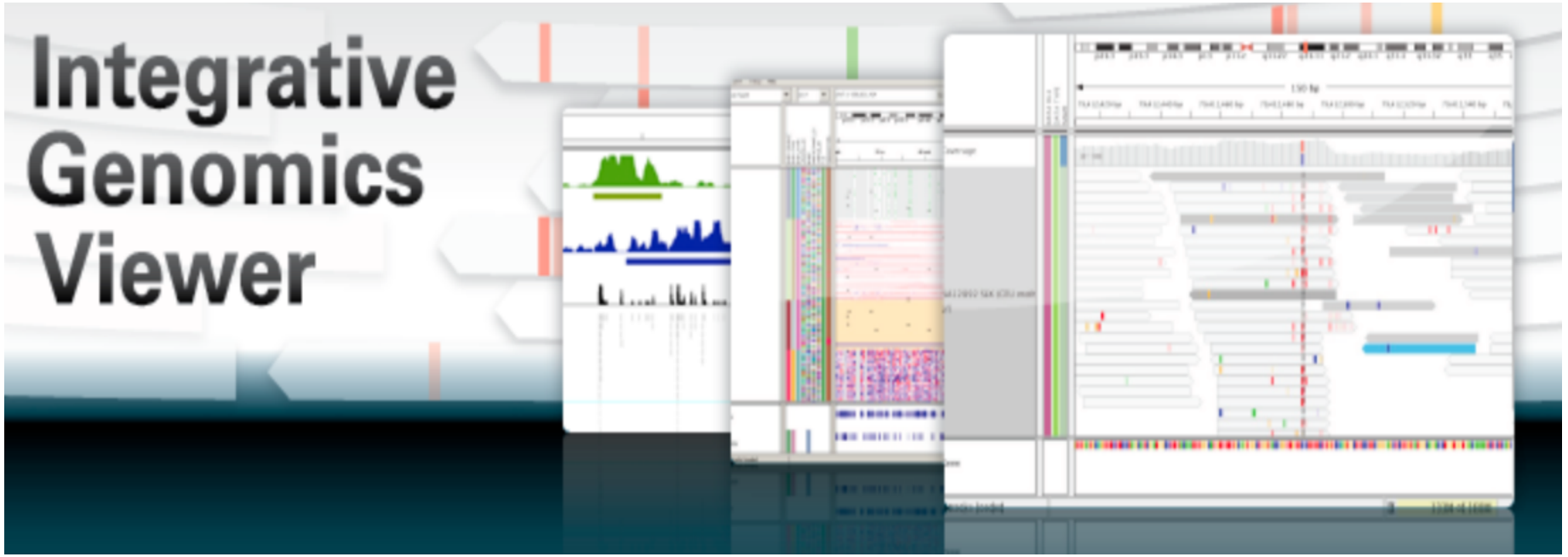
- Online websites data are also repositories
  - UCSC, Ensembl, Gbrowse
  - WashU EpiGenome Browser
  - ...
- Desktop software with graphical user interface
  - IGV
  - BamView
  - ...

# Gbrowse: Generic Genome Browser

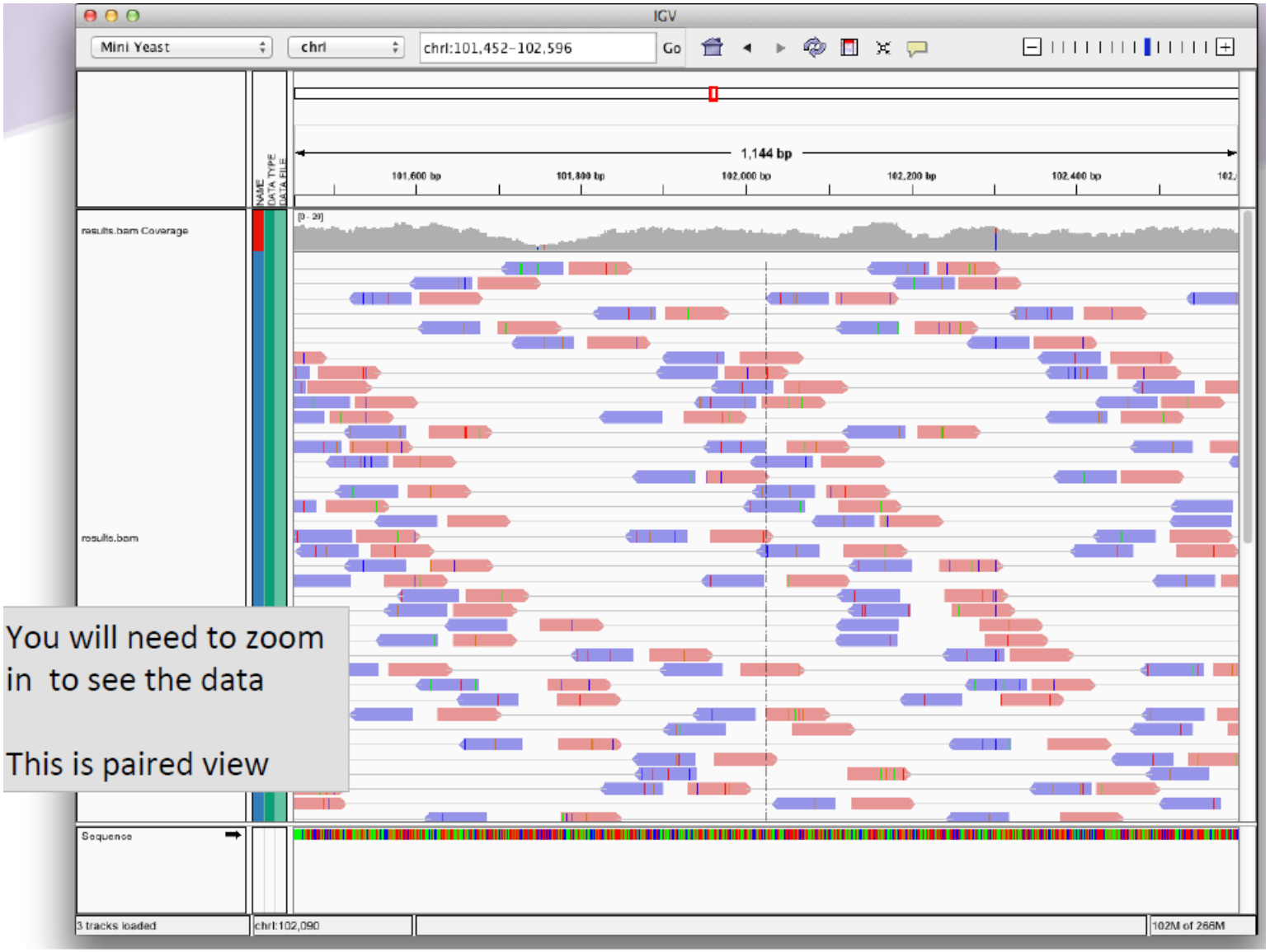


# WashU EpiGenome Browser

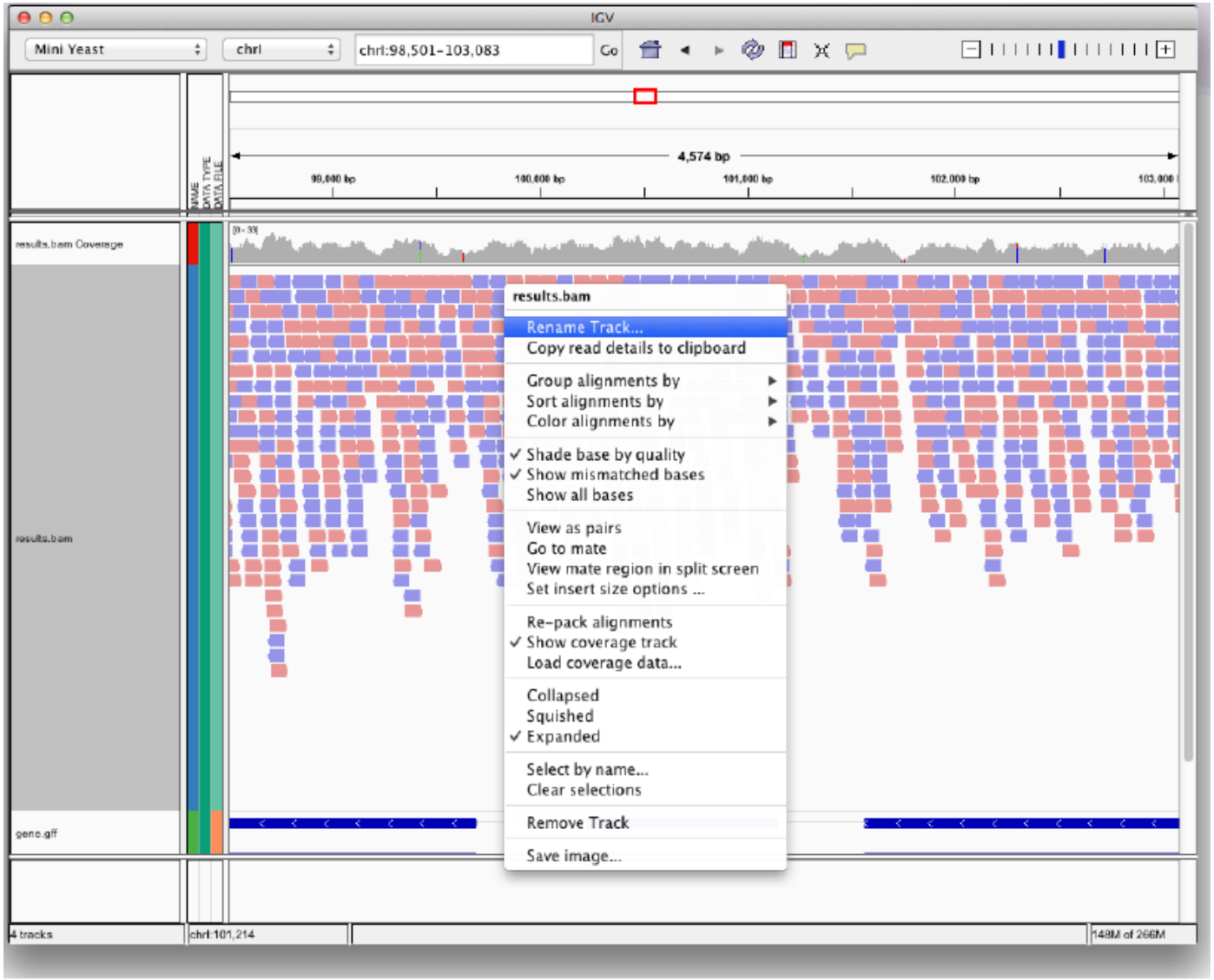
# IGV: integrative genomics viewer

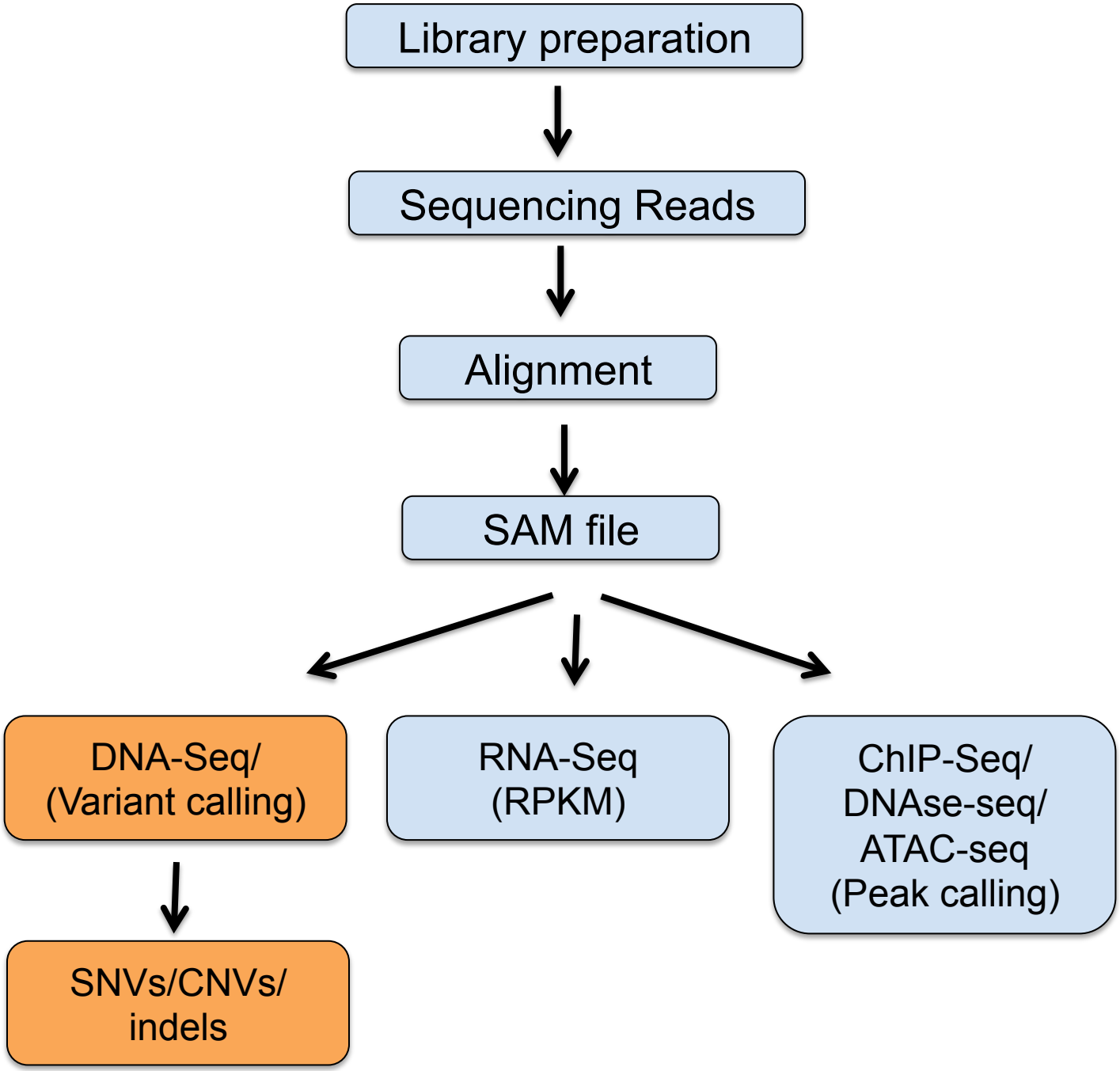


# Visualizing BAM file



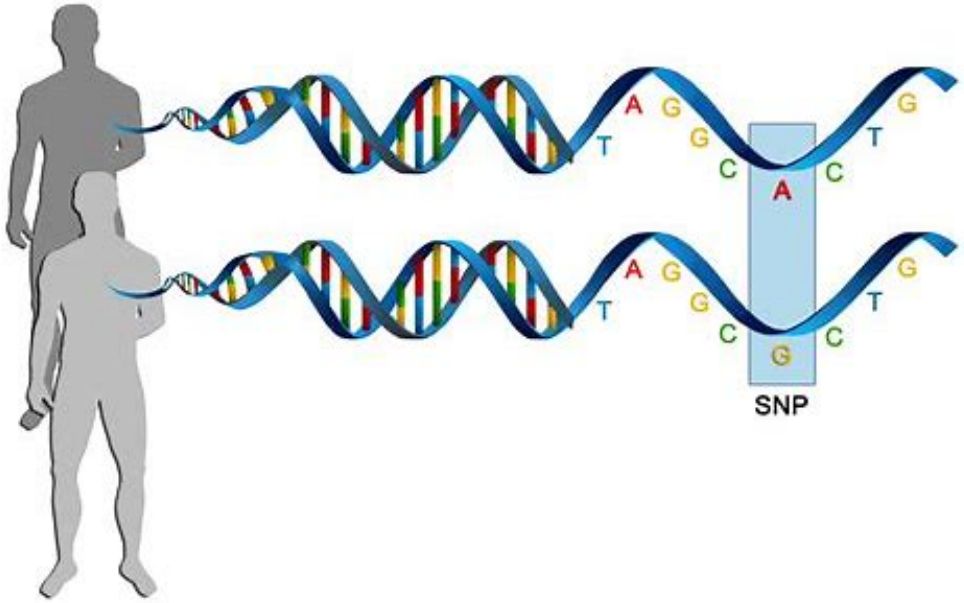
# Right click on the view to set options




Library preparation
Sequencing Reads
Alignment
SAM file
DNA-Seq/
(Variant calling)
RNA-Seq
(RPKM)
ChIP-Seq/
DNAse-seq/
ATAC-seq
(Peak calling)
SNVs/CNVs/
indels

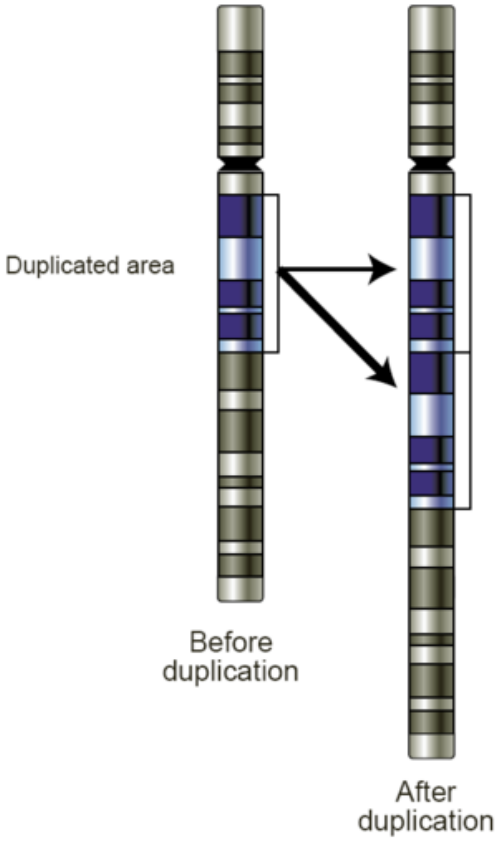
# Single nucleotide variant (SNV)

- SNV: Single-nucleotide variant: A variant called in an individual sequence
- SNP: Single-nucleotide polymorphism. A variant that is polymorphic within a population
- SNV and SNP are used interchangeably



# Structural variation

- Variation in structure of an organism's chromosome
- Including deletions, duplications, copy-number variants, insertions, inversions and translocations.

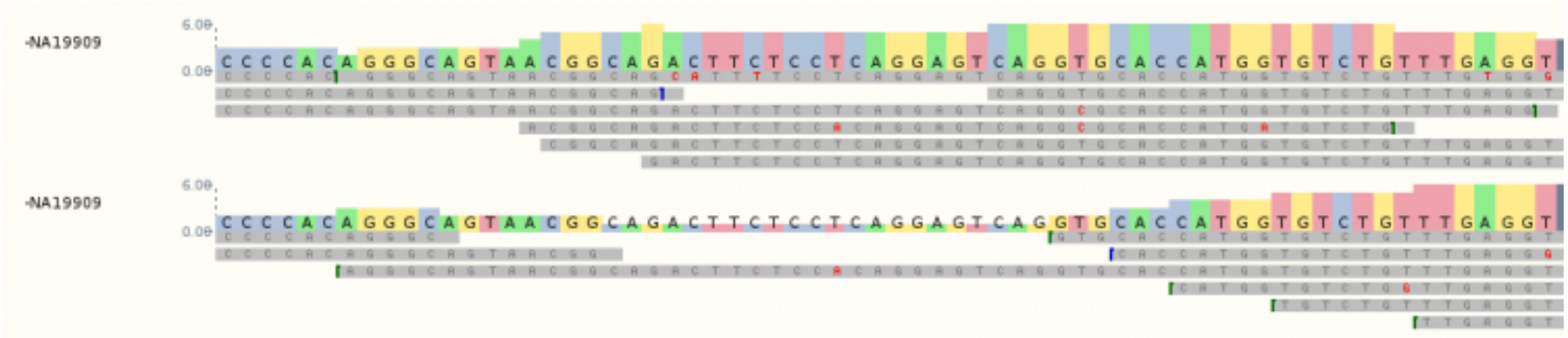


# Mutations nomenclature

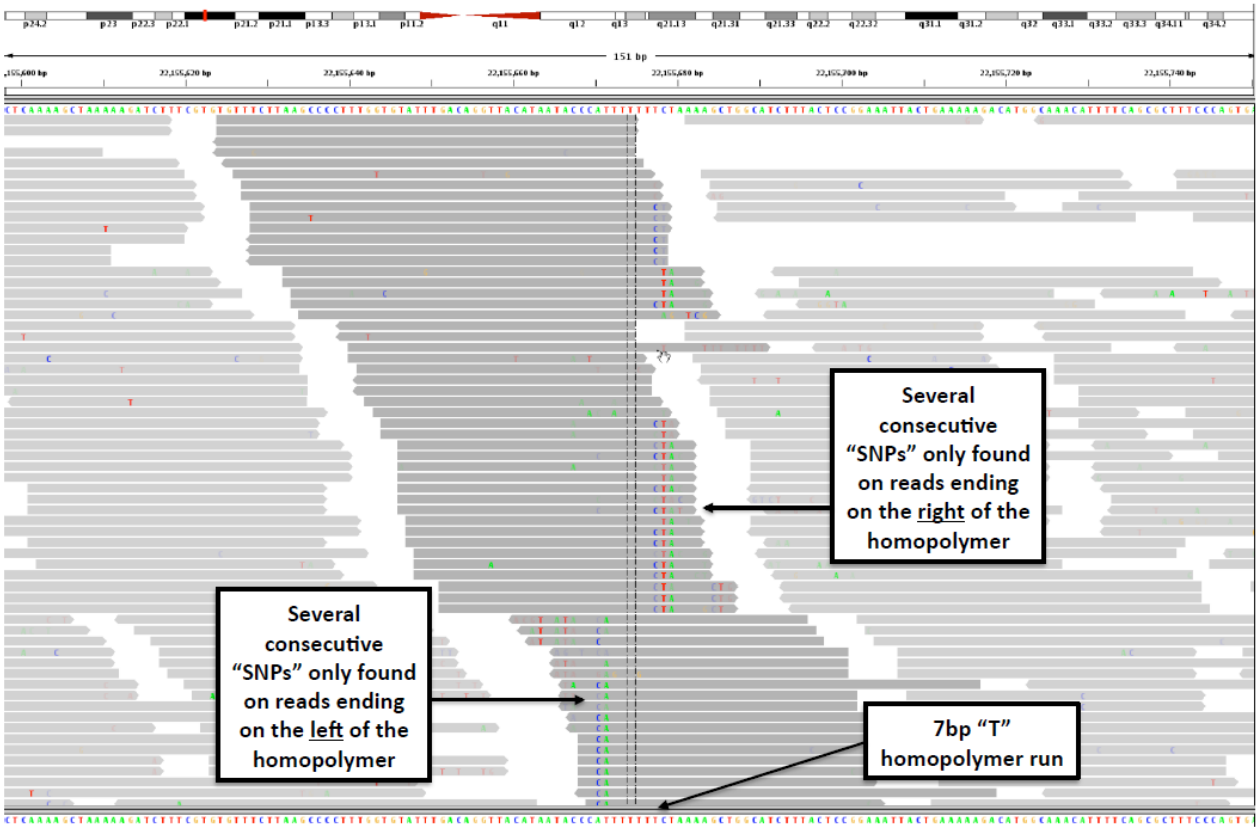
- Major classes of mutation are: missense, nonsense, insertion, deletion, splice-site mutation
- DNA: A,C,G,T
  - Simple subsitution c.123A>G
  - Deletion c.123delA
  - Duplication c.123dupA
  - Insertion c.123 124insC
- Protein: 1- or 3-letter code
  - p.A212P, Ala212Pro

# Introduction of variant calling

- Variant calling is the process by which we identify variants from sequence data.



# Sorted bam


Several consecutive "SNPs" only found on reads ending on the right of the homopolymer
Several consecutive "SNPs" only found on reads ending on the left of the homopolymer
7bp "T" homopolymer run

# Pileup （1）

- Pileup format facilitates SNP/indel calling and brief alignment viewing by eyes.
- Each line consists of chromosome, 1-based coordinate, reference base, the number of reads covering the site, read bases and base qualities.

```
...
21 31587791 T 24 .,..,,.,,.....,..,..,,^], ?EFGDDEEEFEFFEEFDD?EE=>;
21 31587792 G 25 .,..,,.,,.....,..,..,,^], BCHI9H89IJ7IF78G8I:9I::::
21 31587793 A 26 .,..g,Gg,.G..GG,G.gG.,,g,^], 8G=B6F56GC4BC45I5B76BA?8AA
21 31587794 G 27 .$,$..,,.,,.....,..,..,,,,^], <D?F9G89GH7HC78F8H:9EC@>BBB
21 31587795 T 26 ..,,.,,.....,..,..,,,,,^]. ;CEECDEAB@A@AD=@FBC@@>>=>B
...
```

- a dot: match to the reference base on the forward strand
- a comma: match on the reverse strand
- ‘ACGTN’: for a mismatch on the forward strand
- ‘acgtn’: for a mismatch on the reverse strand.

# Pileup（2）

- A pattern `\+[0-9]+[ACGTNacgtn]+' indicates there is an insertion between this reference position and the next reference position. The length of the insertion is given by the integer in the pattern, followed by the inserted sequence. Here is an example of 2bp insertions on three reads:

  seq2 156 A 11 .$.....+2AG.+2AG.+2AGGG <975;:<<<<<

- Similarly, a pattern `-[0-9]+[ACGTNacgtn]+' represents a deletion from the reference. Here is an example of a 4bp deletions from the reference, supported by two reads:

  seq3 200 A 20 ,,,,,..,.-4CACC.-4CACC....,.,,.^~. ==<<<<<<<<<<<::<;2<<

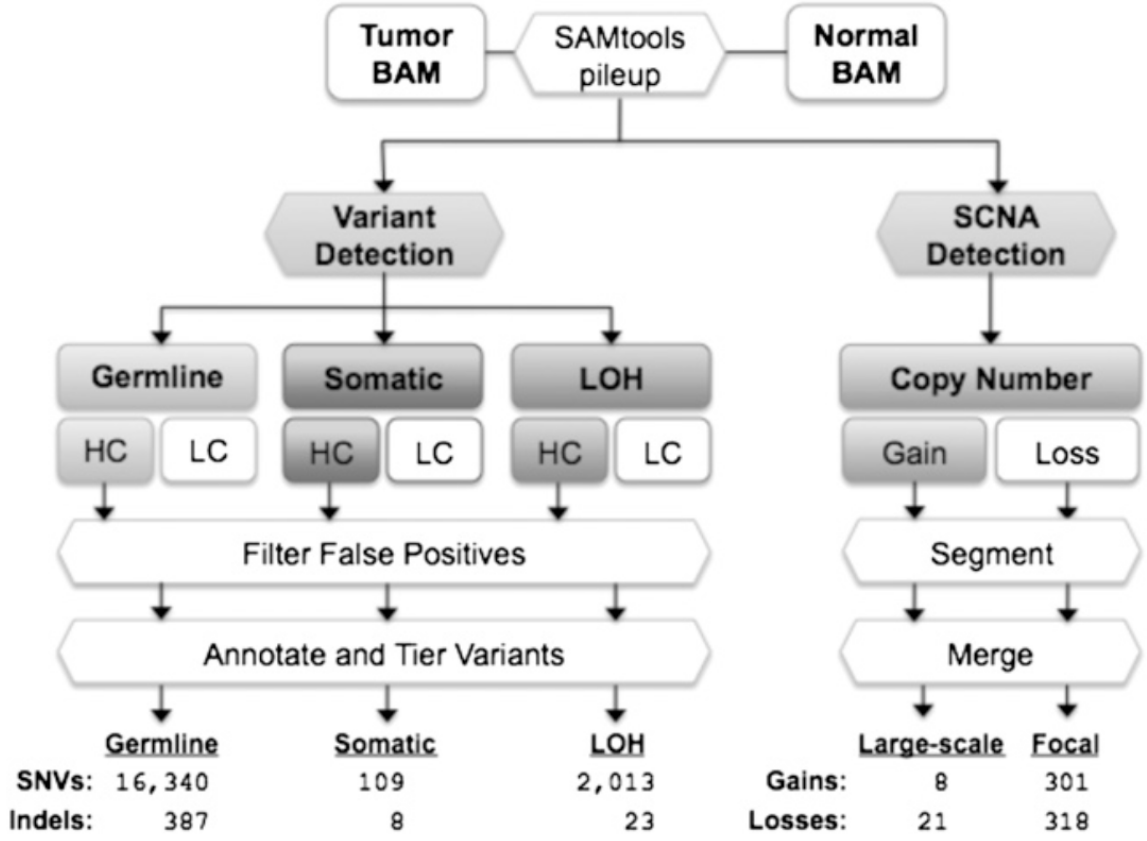
# Variant calling by viewing

- Let us now consider position 31587793

21 31587793 A 26 .,..g,Gg,.G..GG,G.gG.,,g,^], 8G=B6F56GC4BC45I5B76BA?8AA

- This position is covered by a total of 26 reads. 16 reads favor the reference A, and 10 reads favor an alternate base, G
- Intuitively, this seems likely to be a heterozygous variant

# Simple approaches to variant calling (VarScan 2) schematic




Koboldt DC et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568-76.

# Simple approaches to variant calling: VarScan 2

- The following steps are performed for each position of genome in parallel for the tumor sample and the matched normal sample
  - Determine if both samples meet the minimum coverage requirement (by default, three reads with base quality 20)
  - Determine a genotype for each sample individually based upon the read bases observed. By default, a variant allele must be supported by at least two independent reads and at least 8% of all reads.
  - Variants are called homozygous if supported by 75% or more of all reads at a position; otherwise they are called heterozygous.

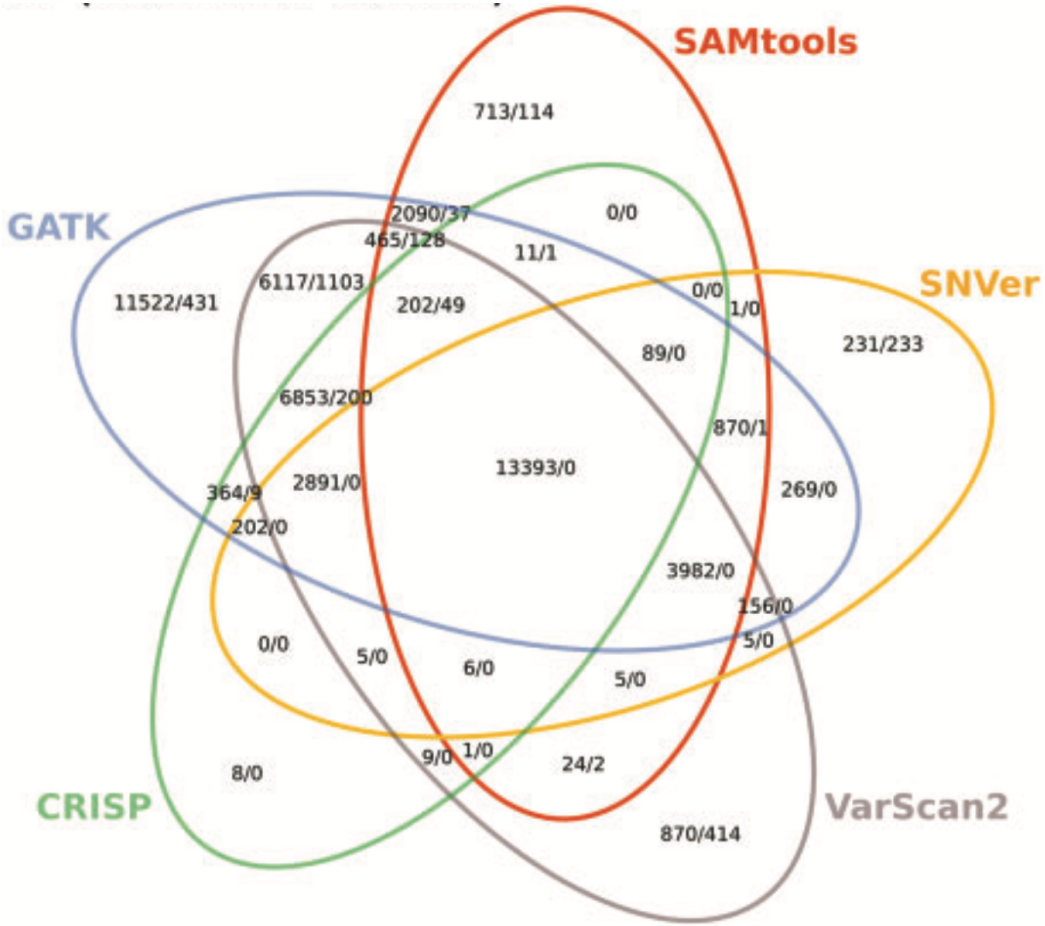
# Simple approaches to variant calling: Varscan

- If the genotypes do not match, then their read counts are evaluated by one-tailed Fisher's exact test in a two-by-two table

| | Reference | alternate |
|---|---|---|
| **Tumor reads** | Tumor reads 1 | Tumor reads 2 |
| **Normal reads** | Normal reads 1 | Normal reads 2 |

- Fisher exact test is performed. If the P value is significant then
  - if the normal sample is called reference and the tumor sample is called alternate, then the variant is called somatic

# Intersection of variants called by different approaches




Pabinger S et al., A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2013

# Challenges of variant calling

Reasons for seeing a mismatch in a pileup include..

- True variant
- Error from library prep (PCR artifacts)
- Machine sequencing errors
- Misalignment (mapping error)
- Error in reference sequence
- Contamination

# Estimating Parameters from Data

- In many situations in bioinformatics, we want to estimate "optimal" parameters from data. In the examples we have seen in the lectures on variant calling, these parameters might be:
  - Error rate for reads
  - Proportion of a certain genotype
  - Proportion of nonreference bases
  - ...
- However, the hello world example for this sort of thing is the coin toss, so we will start with that.

# Coin toss

Let's say we have two coins that are each tossed 10 times

- Coin 1: H,T,T,H,H,H,T,H,T,T
- Coin 2: T,T,T,H,T,T,T,H,T,T

Intuitively, we might guess that coin one is a fair coin, i.e., $P(X = H) = 0.5$, and that coin 2 is biased, i.e.,$P(X = H) \neq 0.5$

# Discrete Random Variable

Let us begin to formalize this. We model the coin toss process as follows.

- The outcome of a single coin toss is a random variable $X$ that can take on values in a set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$
- In our example, of course, $n = 2$, and the values are $x_1 = 0$ (tails) and $x_2 = 1$ (heads)
- We then have a probability mass function $p : \mathcal{X} \longrightarrow [0, 1]$; the law of total probability states that $\sum_{x \in \mathcal{X}} p(x_i) = 1$
- This is a Bernoulli distribution with parameter $\mu$:

$$p(X = 1; \mu) = \mu \tag{1}$$

# Probability of sequence of events

In general, for a sequence of two events $X_1$ and $X_2$, the joint probability is

$$P(X_1, X_2) = p(X_2|X_1)p(X_1) \tag{2}$$

Since we assume that the sequence is iid (identically and independently distributed), by definition $p(X_2|X_1) = P(X_2)$. Thus, for a sequence of $n$ events (coin tosses), we have

$$p(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} p(x_i; \mu) \tag{3}$$

if the probability of heads is 30%, the the probability of the sequence for coin 2 can be calculated as

$$p(T, T, T, H, T, T, T, H, T, T; \mu) = \mu^2(1-\mu)^8 = \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8 \tag{4}$$

# Probability of sequence of events

Thus far, we have considered $p(x; \mu)$ as a function of $x$, parametrized by $\mu$. If we view $p(x; \mu)$ as a function of $\mu$, then it is called the **likelihood function**.

Maximum likelihood estimation basically chooses a value of $\mu$ that maximizes the likelihood function given the observed data.

# Maximum likelihood for Bernoulli

The likelihood for a sequence of i.i.d. Bernoulli random variables $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ with $x_i \in \{0, 1\}$ is then

$$p(\mathbf{X}; \mu) = \prod_{i=1}^{n} p(x_i; \mu) = \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i} \tag{5}$$

We usually maximize the log likelihood function rather than the original function

- Often easier to take the derivative
- the log function is monotonically increasing, thus, the maximum (argmax) is the same
- Avoid numerical problems involved with multiplying lots of small numbers

# Log likelihood

Thus, instead of maximizing this

$$p(\mathbf{X}; \mu) = \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i} \tag{6}$$

we maximize this

$$\begin{aligned}
\log p(\mathbf{X}; \mu) &= \log \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i} \\
&= \sum_{i=1}^{n} \log \left\{ \mu^{x_i}(1-\mu)^{1-x_i} \right\} \\
&= \sum_{i=1}^{n} \left[ \log \mu^{x_i} + \log(1-\mu)^{1-x_i} \right] \\
&= \sum_{i=1}^{n} \left[ x_i \log \mu + (1-x_i) \log(1-\mu) \right]
\end{aligned}$$

# Log likelihood

Note that one often denotes the log likelihood function with the symbol $\mathcal{L} = \log p(\mathbf{X}; \mu)$.

A function f defined on a subset of the real numbers with real values is called monotonic (also monotonically increasing, increasing or non-decreasing), if for all $x$ and $y$ such that $x \leq y$ one has $f(x) \leq f(y)$

Thus, the monotonicity of the log function guarantees that

$$\underset{\mu}{\operatorname{argmax}}\, p(\mathbf{X}; \mu) = \underset{\mu}{\operatorname{argmax}}\, \log p(\mathbf{X}; \mu) \tag{7}$$

# ML estimate

The ML estimate of the parameter $\mu$ is then

$$\underset{\mu}{\text{argmax}} \sum_{i=1}^{n} [x_i \log \mu + (1 - x_i) \log(1 - \mu)] \tag{8}$$

We can calculate the argmax by setting the first derivative equal to zero and solving for $\mu$

# ML estimate

Thus

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log p(\mathbf{X}; \mu) &= \sum_{i=1}^{n} \frac{\partial}{\partial \mu} \left[ x_i \log \mu + (1 - x_i) \log(1 - \mu) \right] \\
&= \sum_{i=1}^{n} x_i \frac{\partial}{\partial \mu} \log \mu + \sum_{i=1}^{n} (1 - x_i) \frac{\partial}{\partial \mu} \log(1 - \mu) \\
&= \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{1}{1 - \mu} \sum_{i=1}^{n} (1 - x_i)
\end{aligned}$$

# ML estimate

and finally, to find the maximum we set $\frac{\partial}{\partial \mu} \log p(\mathbf{X}; \mu) = 0$:

$$
\begin{aligned}
0 &= \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{1}{1-\mu} \sum_{i=1}^{n} (1 - x_i) \\
\frac{1-\mu}{\mu} &= \frac{\sum_{i=1}^{n} (1 - x_i)}{\sum_{i=1}^{n} x_i} \\
\frac{1}{\mu} - 1 &= \frac{\sum_{i=1}^{n} 1}{\sum_{i=1}^{n} x_i} - 1 \\
\frac{1}{\mu} &= \frac{n}{\sum_{i=1}^{n} x_i} \\
\hat{\mu}_{ML} &= \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}
$$

Reassuringly, the maximum likelihood estimate is just the proportion of flips that came out heads.

# Problems with ML estimation

Does it really make sense that

- H,T,H,T $\rightarrow \hat{\mu} = 0.5$
- H,T,T,T $\rightarrow \hat{\mu} = 0.25$
- T,T,T,T$\rightarrow \hat{\mu} = 0.0$

ML estimation does not incorporate any prior knowledge and does not generate an estimate of the certainty of its results.

# Maximum a posteriori Estimation

Bayesian approaches try to reflect our belief about $\mu$. In this case, we will consider $\mu$ to be a random variable.

$$p(\mu|\mathbf{X}) = \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \tag{9}$$

Thus, Bayes' law converts our prior belief about the parameter $\mu$ (before seeing data) into a posterior probability, $p(\mu|\mathbf{X})$, by using the likelihood function $p(\mathbf{X}|\mu)$. The maximum a-posteriori (MAP) estimate is defined as

$$\hat{\mu}_{MAP} = \underset{\mu}{\text{argmax}}\, p(\mu|\mathbf{X}) \tag{10}$$

# Maximum a posteriori Estimation

Note that because $p(\mathbf{X})$ does not depend on $\mu$, we have

$$\begin{aligned}
\hat{\mu}_{MAP} &= \underset{\mu}{\text{argmax}}\, p(\mu|\mathbf{X}) \\
&= \underset{\mu}{\text{argmax}}\, \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \\
&= \underset{\mu}{\text{argmax}}\, p(\mathbf{X}|\mu)p(\mu)
\end{aligned}$$

This is essentially the basic idea of the MAP equation used by SNVMix for variant calling

# Beta distribution: Background

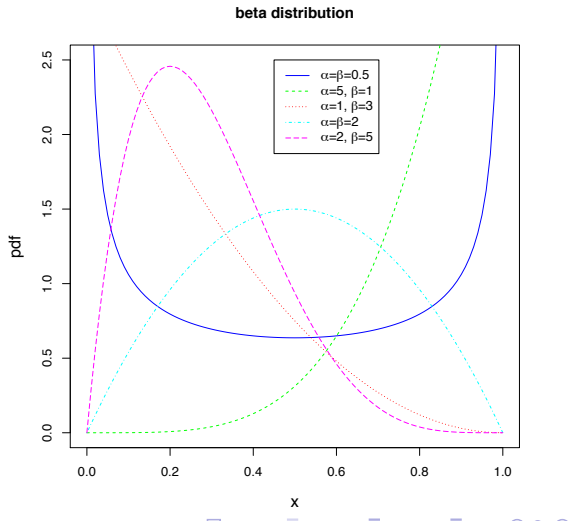
The Beta distribution is appropriate to express prior belief about a Bernoulli distribution. The Beta distribution is a family of continuous distributions defined on $[0, 1]$ and parametrized by two positive shape parameters, $\alpha$ and $\beta$

$$p(\mu) = \frac{1}{B(\alpha, \beta)} \cdot \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

here, $\mu \in [0, 1]$, and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)}$$

where $\Gamma$ is the Gamma function (extension of factorial).



# Maximum a posteriori (MAP) Estimation

Let's go back now to our problem of predicting the results of the next election. Essentially, we plug in the equations for the distributions of the likelihood (a Bernoulli distribution) and the prior (A Beta distribution).

$$\Pr(\mu|\mathbf{X}) \propto \Pr(x_i|\mu) \cdot \Pr(\mu)$$

- posterior
- Likelihood (Bernoulli)
- prior (Beta)

# Maximum a posteriori (MAP) Estimation

We thus have that

- $\Pr(x_i|\mu) = \text{Bernoulli}(x_i|\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$
- $\Pr(\mu) = \text{Beta}(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \cdot \mu^{\alpha-1}(1-\mu)^{\beta-1}$

thus

$$\Pr(\mu|\mathbf{X}) \propto \Pr(\mathbf{X}|\mu)\Pr(\mu)$$

is equivalent to

$$\Pr(\mu|\mathbf{X}) \propto \left\{ \prod_i \text{Bernoulli}(x_i|\mu) \right\} \cdot \text{Beta}(\mu|\alpha,\beta) \qquad (15)$$

# Maximum a posteriori (MAP) Estimation

Furthermore

$$\begin{aligned}
\mathcal{L} &= \log \Pr(\mu|\mathbf{X}) \\
&= \log \left\{ \prod_i \text{Bernoulli}(x_i|\mu) \right\} \cdot \text{Beta}(\mu|\alpha, \beta) \\
&= \sum_i \log \text{Bernoulli}(x_i|\mu) + \log \text{Beta}(\mu|\alpha, \beta)
\end{aligned}$$

We solve for $\hat{\mu}_{MAP} = \text{argmax}_\mu \mathcal{L}$ as follows

$$\underset{\mu}{\text{argmax}} \sum_i \log \text{Bernoulli}(x_i|\mu) + \log \text{Beta}(\mu|\alpha, \beta)$$

Note that this is almost the same as the ML estimate except that we now have an additional term resulting from the prior

# Maximum a posteriori (MAP) Estimation

Again, we find the maximum value of $\mu$ by setting the first derivative of $\mathcal{L}$ equal to zero and solving for $\mu$

$$\frac{\partial}{\partial \mu}\mathcal{L} = \sum_i \frac{\partial}{\partial \mu} \log \mathrm{Bernoulli}(x_i|\mu) + \frac{\partial}{\partial \mu} \log \mathrm{Beta}(\mu|\alpha, \beta)$$

The first term is the same as for ML[3], i.e.

$$\sum_i \frac{\partial}{\partial \mu} \log \mathrm{Bernoulli}(x_i|\mu) = \frac{1}{\mu}\sum_{i=1}^{n} x_i - \frac{1}{1-\mu}\sum_{i=1}^{n}(1-x_i) \quad (16)$$

# Maximum a posteriori (MAP) Estimation

To find the second term, we note

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log \mathrm{Beta}(\mu|\alpha,\beta) &= \frac{\partial}{\partial \mu} \log \left\{ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \mu^{\alpha-1}(1-\mu)^{\beta-1} \right\} \\
&= \frac{\partial}{\partial \mu} \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} + \frac{\partial}{\partial \mu} \log \mu^{\alpha-1}(1-\mu)^{\beta-1} \\
&= 0 + \frac{\partial}{\partial \mu} \log \mu^{\alpha-1}(1-\mu)^{\beta-1} \\
&= (\alpha-1)\frac{\partial}{\partial \mu} \log \mu + (\beta-1)\frac{\partial}{\partial \mu} \log (1-\mu) \\
&= \frac{\alpha-1}{\mu} - \frac{\beta-1}{1-\mu}
\end{aligned}
$$

# The use of MAP in MAQ

Recall from the lecture that we call the posterior probabilities of the three genotypes given the data $D$, that is a column with $n$ aligned nucleotides and quality scores of which $k$ correspond to the reference $a$ and $n-k$ to a variant nucleotide $b$.

$$
\begin{aligned}
p(G = \langle a,a \rangle | D) &\propto p(D|G = \langle a,a \rangle) p(G = \langle a,a \rangle) \\
&\propto \alpha_{n,k} \cdot (1-r)/2 \\
p(G = \langle b,b \rangle | D) &\propto p(D|G = \langle b,b \rangle) p(G = \langle b,b \rangle) \\
&\propto \alpha_{n,n-k} \cdot (1-r)/2 \\
p(G = \langle a,b \rangle | D) &\propto p(D|G = \langle a,b \rangle) p(G = \langle a,b \rangle) \\
&\propto \binom{n}{k} \frac{1}{2^n} \cdot r
\end{aligned}
$$

# MAQ: Consensus Genotype Calling

Note that MAQ does not attempt to learn the parameters, rather it uses user-supplied parameter $r$ which roughly corresponds to $\mu$ in the election.

MAQ calls the the genotype with the highest posterior probability:

$$\hat{g} = \underset{g \in (\langle a,a \rangle, \langle a,b \rangle, \langle b,b \rangle)}{\text{argmax}} p(g|D)$$

The probability of this genotype is used as a measure of confidence in the call.

# MAQ: Consensus Genotype Calling

A major problem in SNV calling is false positive heterozygous variants. It seems less probable to observe a heterozygous call at a position with a common SNP in the population. For this reason, MAQ uses a different prior ($r$) for previously known SNPs ($r = 0.2$) and "new" SNPs ($r = 0.001$).

Let us examine the effect of these two priors on variant calling. In R, we can write

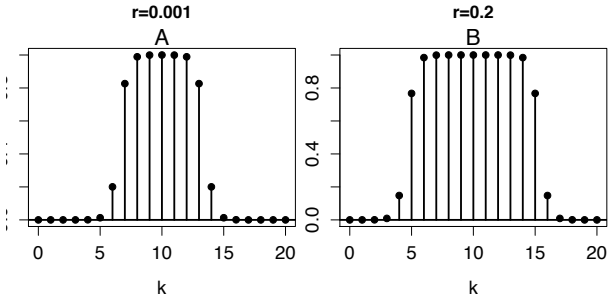
$$p(G = \langle a, b \rangle | D) \propto \binom{n}{k} \frac{1}{2^n} \cdot r$$

as

```
> dbinom(k,n,0.5) * r
```

where $k$ is the number of non-reference bases, $n$ is the total number of bases, and 0.5 corresponds to the probability of seeing a non-ref base given that the true genotype is heterozygous.

# MAQ: Consensus Genotype Calling




- The figures show the posterior for the heterozygous genotype according to the simplified MAQ algorithm discussed in the previous lecture
- The prior $r = 0.0001$ means than positions with 5 or less ALT bases do not get called as heterozygous, whereas the prior with $r = 0.2$ means that positions with 5 bases do get a het call

# Software for variant calling

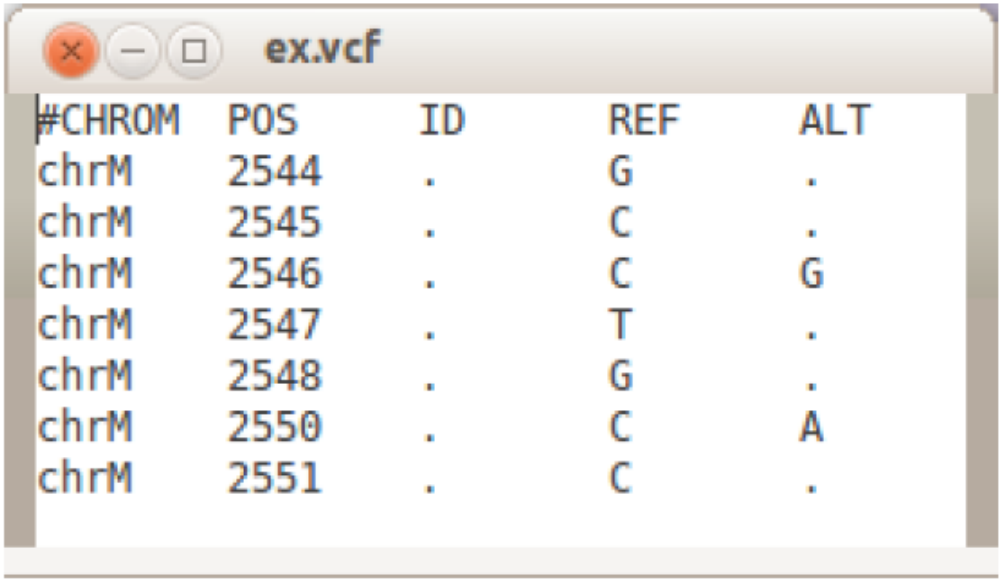
- SAMTools
- GATK
- CRISP
- varScan 2
- SNVer
- SeattleSeq
- Oncotator
- Annovar
- snv-mix

Format for saving the called variants

# VCF: Variant Call Format

# VCF: The simple part

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



# 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.

## 1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

# VCF: 8 mandatory columns + optional

| | 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

# VCF: Complex variants

## Types of variants

### SNPs

*Alignment*

```
ACGT
ATGT
```

*VCF representation*

```
POS REF ALT
2   C   T
```

### Insertions

*Alignment*

```
AC-GT
ACTGT
```

*VCF representation*

```
POS REF ALT
2   C   CT
```

### Deletions

*Alignment*

```
ACGT
A--T
```

*VCF representation*

```
POS REF ALT
1   ACG A
```

### Complex events

*Alignment*

```
ACGT
A-TT
```

*VCF representation*

```
POS REF ALT
1   ACG AT
```

### Large structural variants

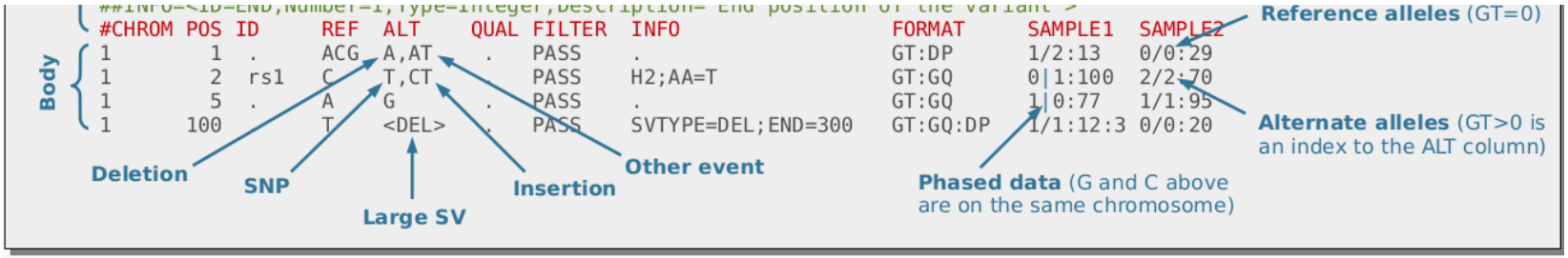
*VCF representation*

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

# 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

Thank you for your attention