CAMBRIDGE Cancer CENTRE

Cambridge science is changing the way we treat cancer

Calling germline SNVs

Oscar M Rueda

Caldas Lab, CRUK Cambridge Institute. University of Cambridge

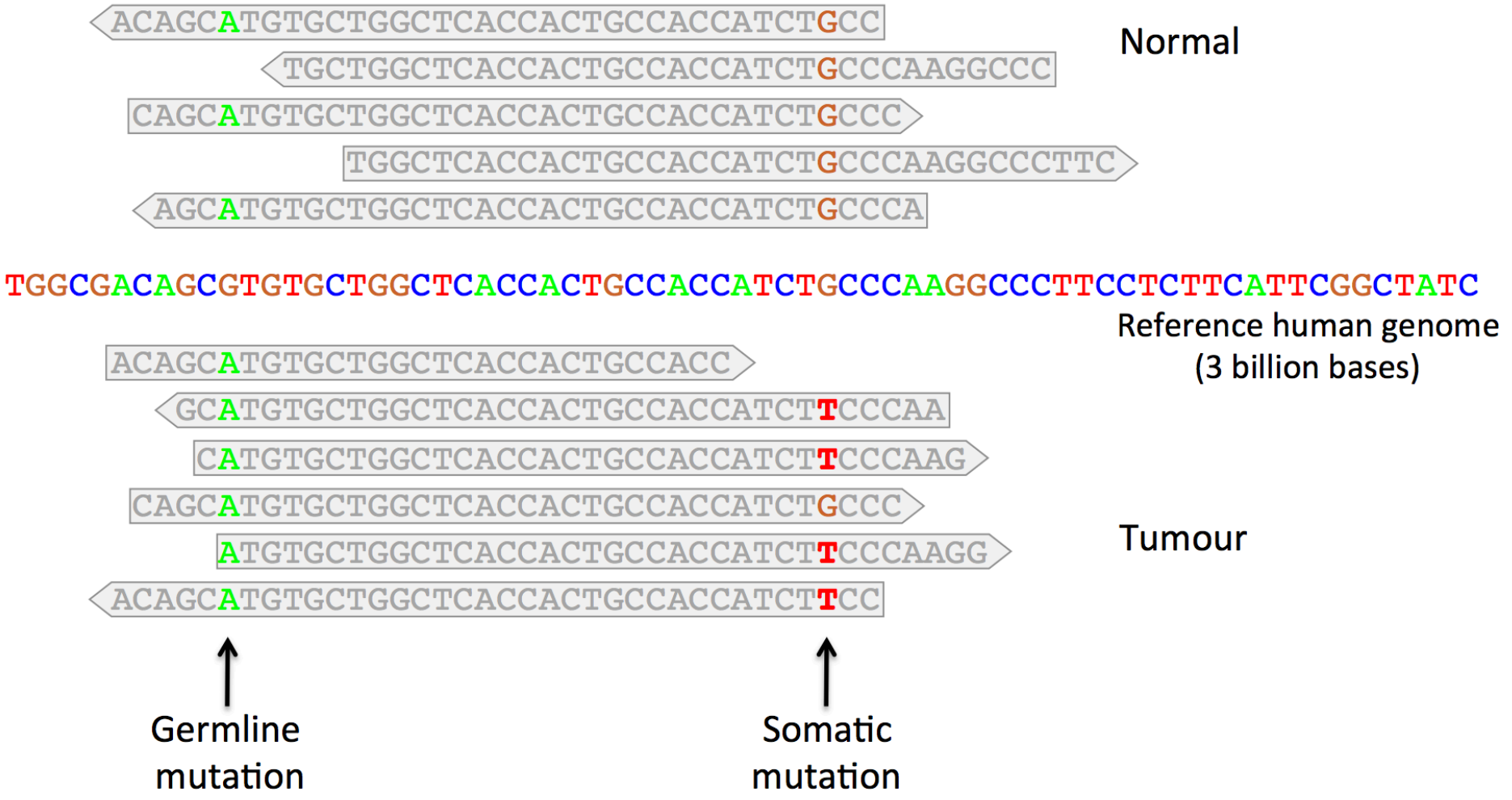
CRUK Summer School, Cambridge 2017.

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Cambridge University Hospitals NHS
NHS Foundation Trust

SNVs, SNPs and somatic mutations



Slide: Matt Eldrigde

Why are we interested in germline variants

Why are we interested in calling germline variants in an **Analysis of Cancer Genomes** course?

- We need to know the inherited variants to filter the somatic mutations
- Some germline variants predispose to cancer (GWAS studies)
- We can build noise models that improve somatic calling methods
- Quality check with multisample sequencing, tumour/normal pairs, DNA/RNA pairs
- Germline variants can refine copy number calling (more on that later today with Geoff)

Naïve genotyping: Maximum likelihood estimator

Suppose we observe 25 alternate reads in a given SNP from 90 total reads. How do we estimate the genotype of the SNP?
We can have three possible genotypes:

- RR
- RA
- AA

We can build the likelihood, that is the probability of our observations given a particular genotype:

- P(25AltReads,65RefReads|RR)
- P(25AltReads,65RefReads|RA)
- P(25AltReads,65RefReads|AA)

Naïve genotyping: Maximum likelihood estimator

In general, the number of alternate reads we observe if our depth is 90 follows a binomial distribution with parameters n=90 and *p,* the probability of sequencing an alternate allele.

$$L(x \mid G) = \binom{90}{25} p^{25}(1-p)^{65}$$

This is a function of p. *Assuming the locus is diploid*, we can only have 3 possible values of p:

- RR -> p=0
- RA -> p=0.5
- AA -> p=1

Naïve genotyping: Maximum likelihood estimator

$$L(x \mid G) = \binom{90}{25} p^{25}(1-p)^{65}$$

We need to consider technical errors in the alignment!

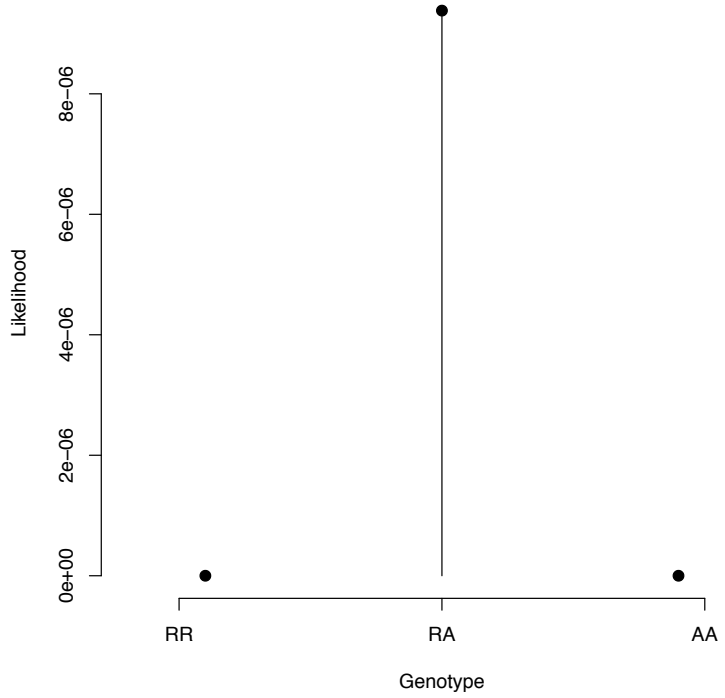
- RR -> p=0.05
- RA -> p=0.5
- AA -> p=0.95

Better error models are available!

Naïve genotyping: Maximum likelihood estimator

$$L(x \mid G) = \binom{90}{25} p^{25}(1-p)^{65}$$

If we plot this function, what is the most likely value of the parameter?



Incorporating additional information in our model

- We know that some genotypes are more common that others in the population
- This information can help us deciding in situations where the likelihood is flat.
- We need **Bayesian Inference** for this.

Bayesian Statistics v Frequentist Statistics

- The main problem of Statistical Inference is the estimation of a specific feature (**para**meter) of a population (for example, the genotype of a given individual on a given locus)
- Classic (Frequentist) Statistics assume that the feature is deterministic (fixed at the time of our observation)
- Bayesian Statistics assume that the feature is random
- This (apparently) small technical difference creates a huge philosophical distinction (and a lot of technicalities too)
- Statisticians tend to be very passionate around each approach (although some are pragmatic)

Bayesian Statistics: Principles

- If the feature of the population that we want to estimate is a random variable, it must have a probability distribution. We call this **the prior distribution: P(G)**
- We take a sample from the population, measure our observations, and compute the **likelihood or our data**: **P(R|G)**
- We can then update our knowledge of the distribution of the parameter and obtain its **posterior distribution** using the **Bayes Theorem**:

$$P(G \mid R) = \frac{P(G)P(R \mid G)}{P(R)}$$

- P(R) is difficult to compute and it is the same for all possible genotypes, so in most applications we just need to compute the numerator.
- We obtain a probability distribution! Not a single value! But we can obtain a range of values with high probability (**credible interval**), or summarise the distribution with the most likely value (**maximum a posteriori**) or the mean (**posterior mean**)

Naïve Bayesian genotyping: Posterior probability

$$L(x \mid G) = \binom{90}{25} p^{25}(1-p)^{65}$$

We need to consider Prior probabilities for each genotype:

P(RR)=0.75

P(RA)=0.15

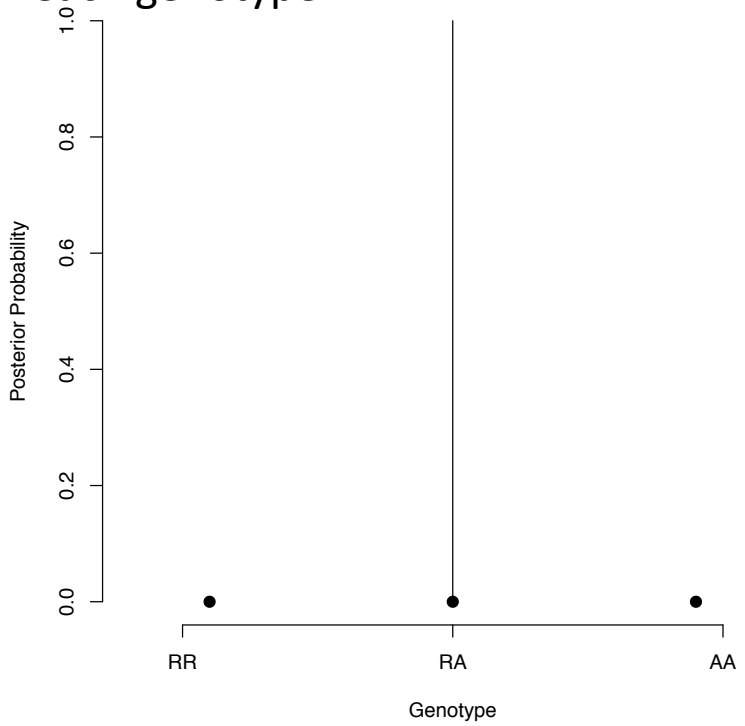
P(AA)=0.10

We can compute posterior probabilities:

P(RR|x)=P(x|RR)P(RR)

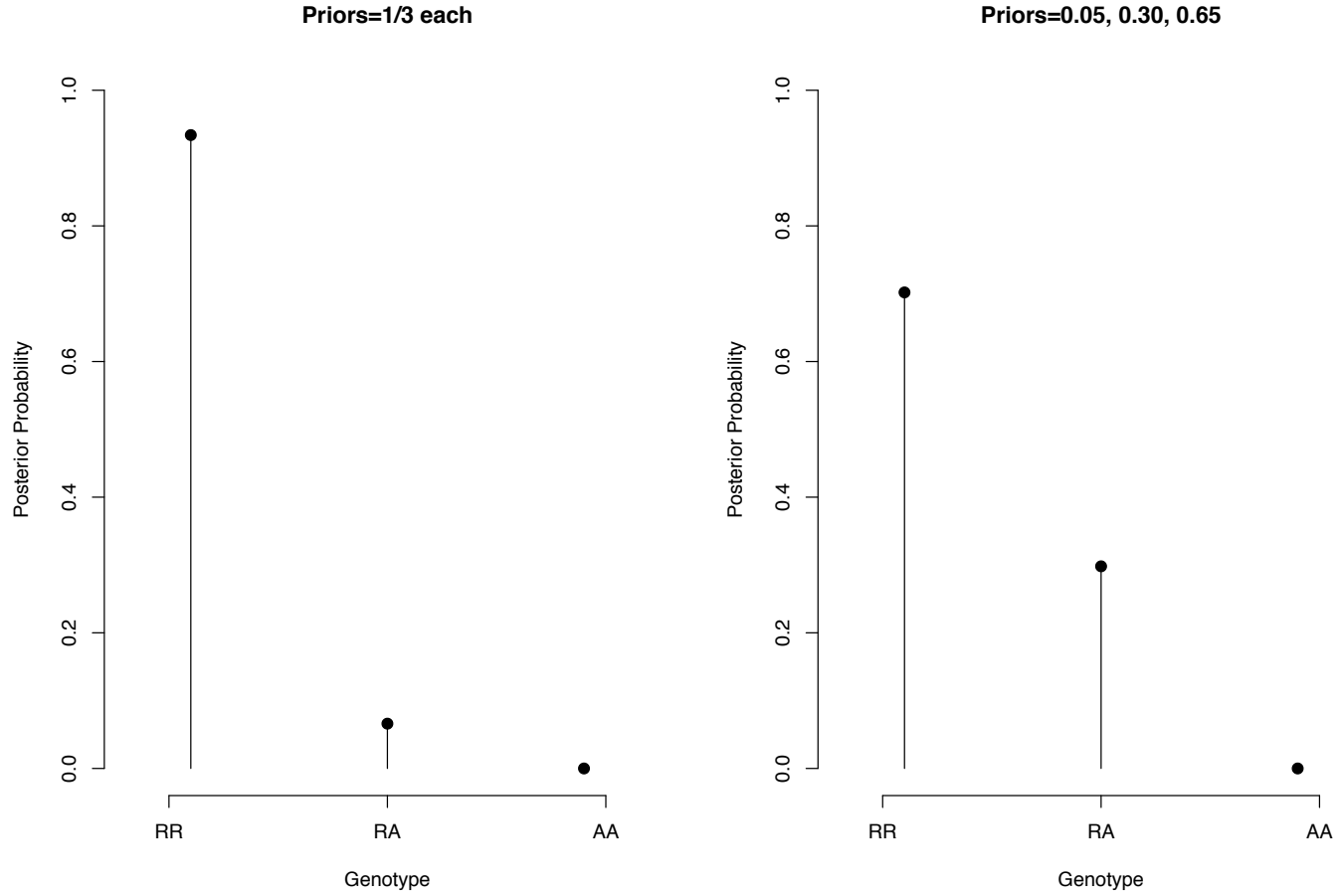
P(RA|x)=P(x|RA)P(RA)

P(AA|x)=P(x|AA)P(AA)



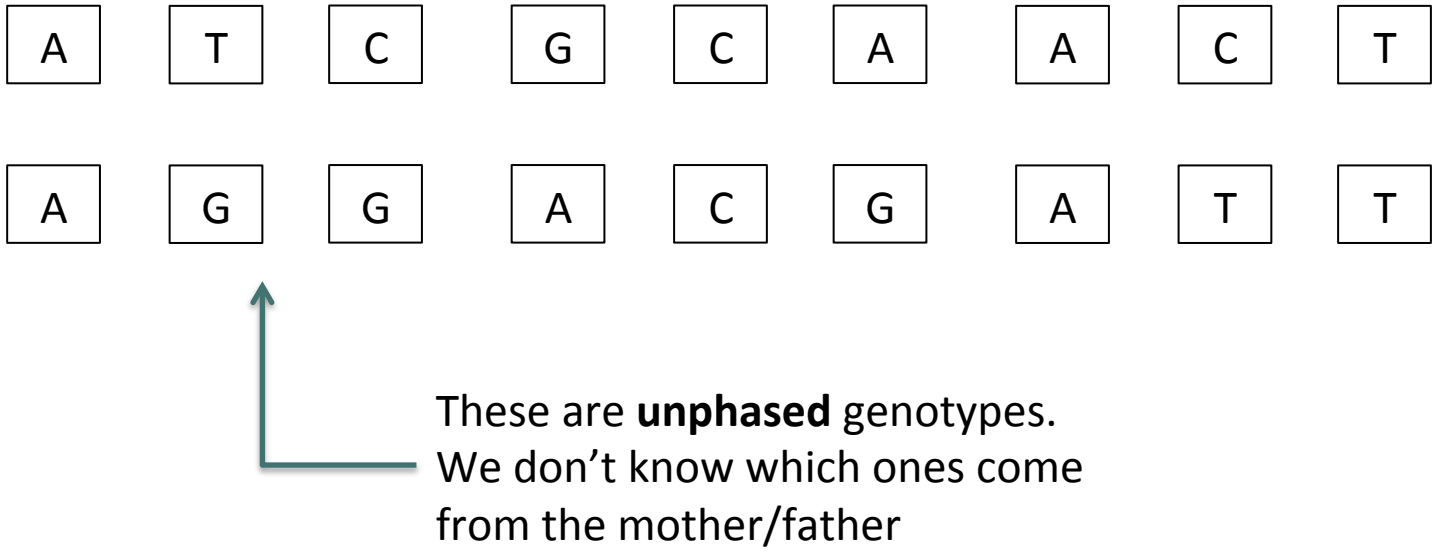
Naïve Bayesian genotyping: Posterior probability

In situations where we don't have much data (lower depth), or there is more noise, the prior can have more impact in the posterior. Let's assume now 10 variant reads out of 50:



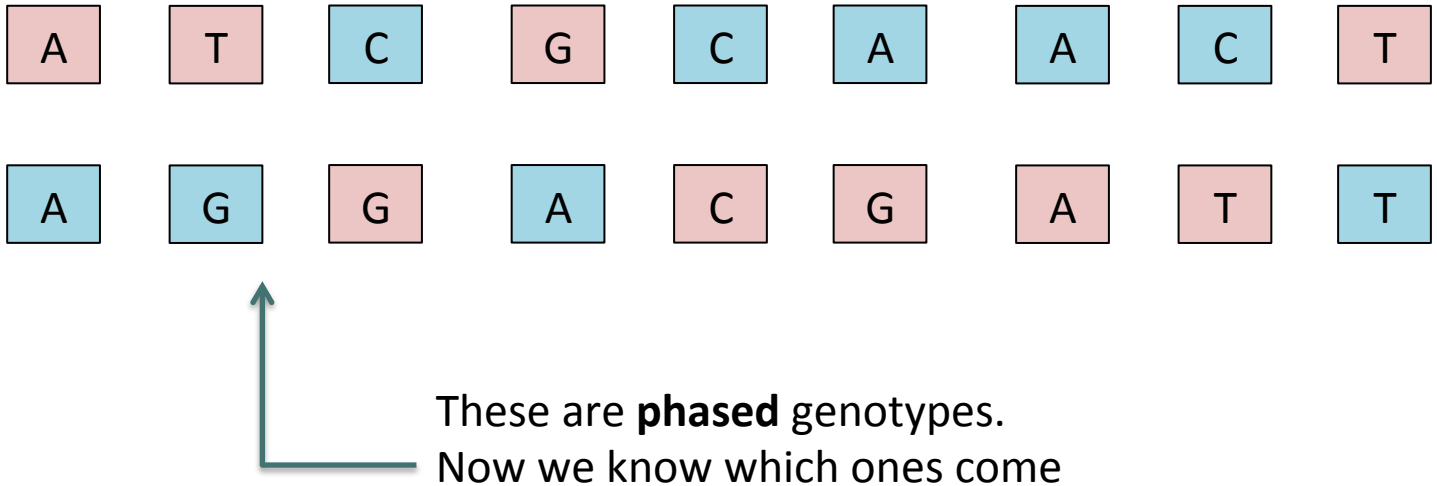
Haplotypes and Computational Phasing

A haplotype is a set of genetic features that *tend to be* inherited together
In our context, we are interested in a set of SNPs that *tend to be* inherited together.
Why? Because it will make variant calling more **robust**



Haplotypes and Computational Phasing

A haplotype is a set of genetic features that *tend to be* inherited together
In our context, we are interested in a set of SNPs that *tend to be* inherited together
Why? Because it will make variant calling more **robust**



Statistical haplotype phasing (unrelated individuals)

Unphased genotypes	Possible phasing A		Possible phasing B		Possible phasing C		Possible phasing D	
A/C	A	C	A	C	A	C	A	C
G/T	G	T	G	T	T	G	T	G
A/T	A	T	T	A	A	T	T	A
Population haplotype frequency	55%	0%	15%	5%	2%	3%	0%	20%
Population frequency of unordered haplotype pair	0%		2 × (15% × 5%) = 1.5%		2 × (2% × 3%) = 0.12%		0%	
Posterior probability of unordered haplotype pair	0%		1.5% / (1.5% + 0.12%) = 93%		0.12% / (1.5% + 0.12%) = 7%		0%	

Figure 1 | **Statistical phasing of unrelated individuals using haplotype frequencies.** Consider one individual with a heterozygous genotype at each of three SNPs in a region. There are four possible haplotype configurations that are consistent with the genotype data (possible phasing patterns A–D). Suppose that haplotype frequencies are available from other individuals in the population at these sites (provided below each phasing pattern). These frequencies may have been estimated from population data without additional modelling (with the *a priori* assumption that all haplotype frequency configurations are equally likely) or from a model that accounts for the biological processes of recombination and mutation (such as the Li and Stephens model[30]). The population frequency of a haplotype pair is obtained using the Hardy–Weinberg principle (independence of the two haplotypes within an individual); the factor of two in the frequency of the haplotype pairs accounts for both possible assignments of maternal and paternal origin to the two haplotypes. The posterior probabilities of the phased data are obtained from the population frequencies of the possible haplotype pairs. In this example, the posterior probability of phasing B (93%) is much greater than that of phasing C (7%).

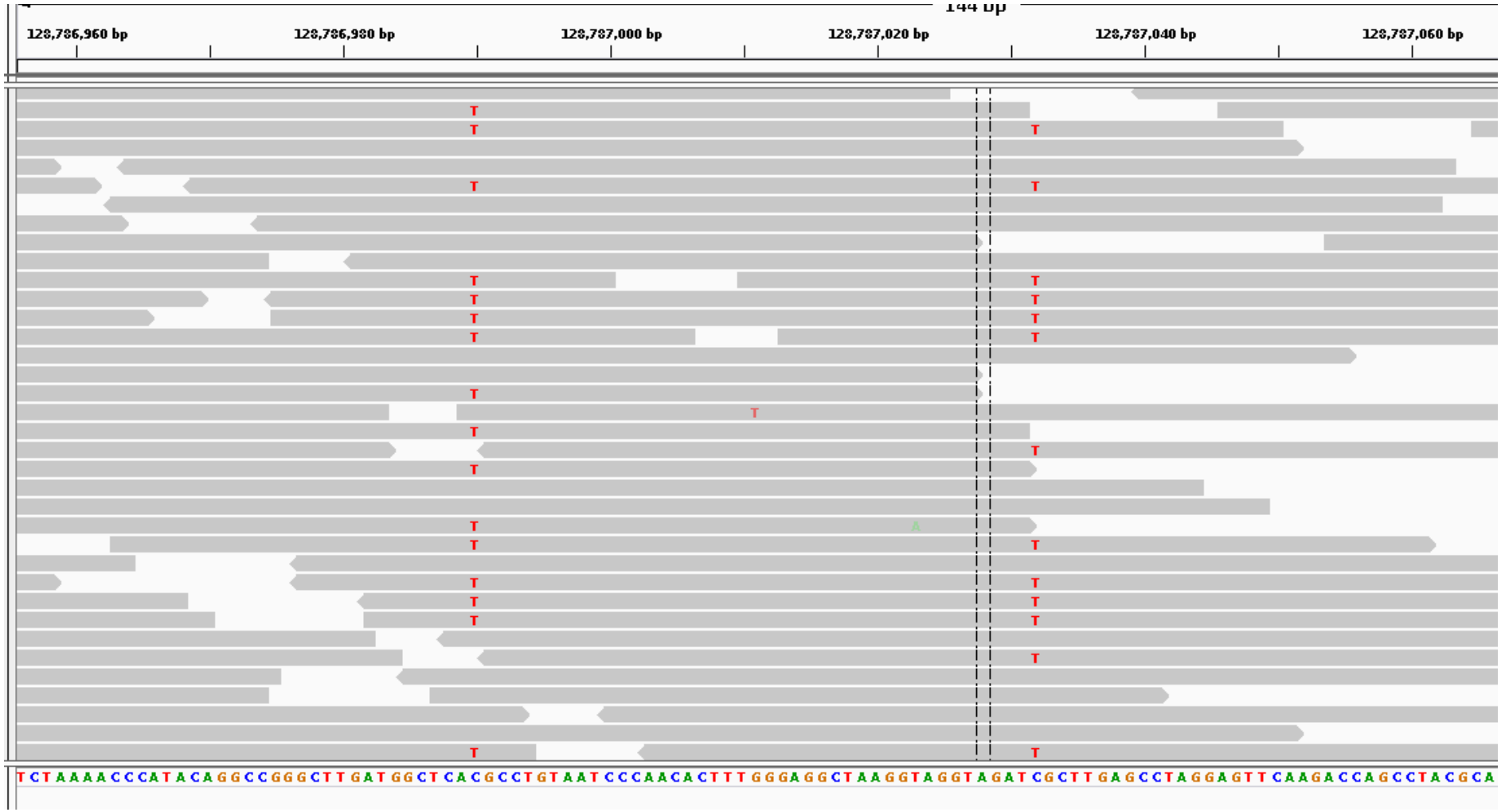
Source: Browning SR and Browning BL. Haplotype phasing: existing methods and new developments (2011). Nature Reviews Genetics, 12

Statistical haplotype phasing (related individuals)

SNP index	Unphased genotypes		Shared haplotype	IBD-phased genotypes				Possible phasing A				Possible phasing B			
	Individual 1	Individual 2		Individual 1		Individual 2		Individual 1		Individual 2		Individual 1		Individual 2	
1	A/C	A/C	?	?	?	?	?	A	C	A	C	C	A	C	A
2	C/T	C/C	C	C	T	C	C	C	T	C	C	C	T	C	C
3	T/T	T/G	T	T	T	T	G	T	T	T	G	T	T	T	G
4	G/G	A/G	G	G	G	G	A	G	G	G	A	G	G	G	A
5	C/C	C/C	C	C	C	C	C	C	C	C	C	C	C	C	C
	Population frequency of haplotype (second instance of shared haplotype in parentheses)							5%	6%	(5%)	0.1%	0.2%	0.3%	(0.2%)	3%
	Population frequency of ordered trio of haplotypes							$5\% \times 6\% \times 0.1\% = 3.0 \times 10^{-4}\%$				$0.2\% \times 0.3\% \times 3\% = 1.8 \times 10^{-5}\%$			
	Posterior probability of phasing (normalized population frequency of trio of haplotypes)							$3 \times 10^{-4}\% / (3 \times 10^{-4}\% + 1.8 \times 10^{-5}\%) = 94\%$				$1.8 \times 10^{-5}\% / (3 \times 10^{-4}\% + 1.8 \times 10^{-5}\%) = 6\%$			

Figure 3 | **Use of identity-by-descent to determine haplotype phase.** First, we discuss how to determine phase using identity-by-descent (IBD) alone (main columns 1–4).When two individuals are known to be identical-by-descent (for example, if they are a parent–offspring pair), the individuals share an allele at each marker, and this allele is determined by the genotype data when one or both individuals are homozygous. In this example, the two individuals with unphased genotypes shown in main column 2 are identical-by-descent. SNP 1 is heterozygous in both individuals and thus cannot be phased using IBD but may be able to be phased using population haplotype frequencies (see below). SNP 2 is homozygous in individual 2, and so the shared haplotype must have the C allele. Analogously, SNPs 3 and 4 are homozygous in individual 1, so the shared alleles are T and G, respectively. SNP 5 is homozygous in both individuals, so phasing is trivial. The inferred shared haplotype is shaded green. Use of IBD phasing alone gives the phasing shown in the IBD-phased haplotype columns, in which the phasing of SNP 1 is unknown. Second, we discuss how to determine phase using IBD and haplotype frequencies. Consider the same two identical-by-descent individuals as above. The phase is determined by IBD at SNPs 2–5 (main column 3) but is not determined at SNP 1, which is heterozygous in both individuals. Only haplotype phasings that satisfy the IBD-phasing constraints need be considered. Here the two identical-by-descent individuals are phased jointly, so the joint phase at SNP 1 must be consistent with the IBD, and the identical-by-descent haplotype is only included once in the probability of the haplotype configuration. The inferred identical-by-descent haplotype is shaded in main columns 5 and 6. Haplotype phasing pattern A is much more probable (94%) than phasing pattern B (6%).

Phasing with short reads



freeBayes variant caller

Features of the method:

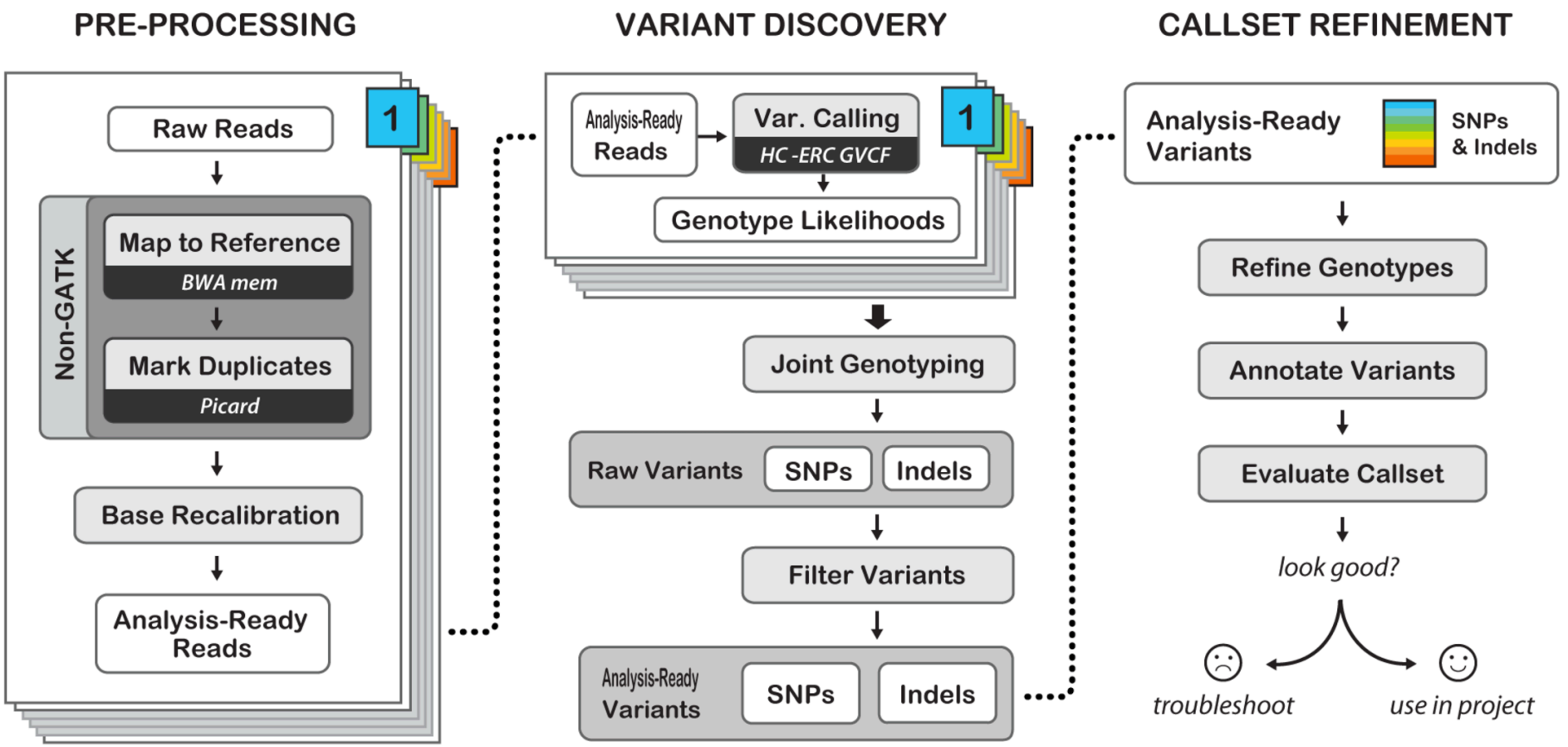
- It is a Bayesian method
- It incorporates information from multiple samples
- It uses haplotype blocks

Components of the model:

- Modeling of the number of reads given a certain genotype as a multinomial distribution
- It incorporates the probability of errors in the reads (as a function of the quality scores)

Garrison, E and Marth, G. Haplotype-based variant detection from short-read sequencing

GATK Pipeline



GATK Pipeline

How HaplotypeCaller works

1. Define active regions
2. Determine haplotypes by assembly of the active region
3. Determine likelihoods of the haplotypes given the read data
4. Assign sample genotypes