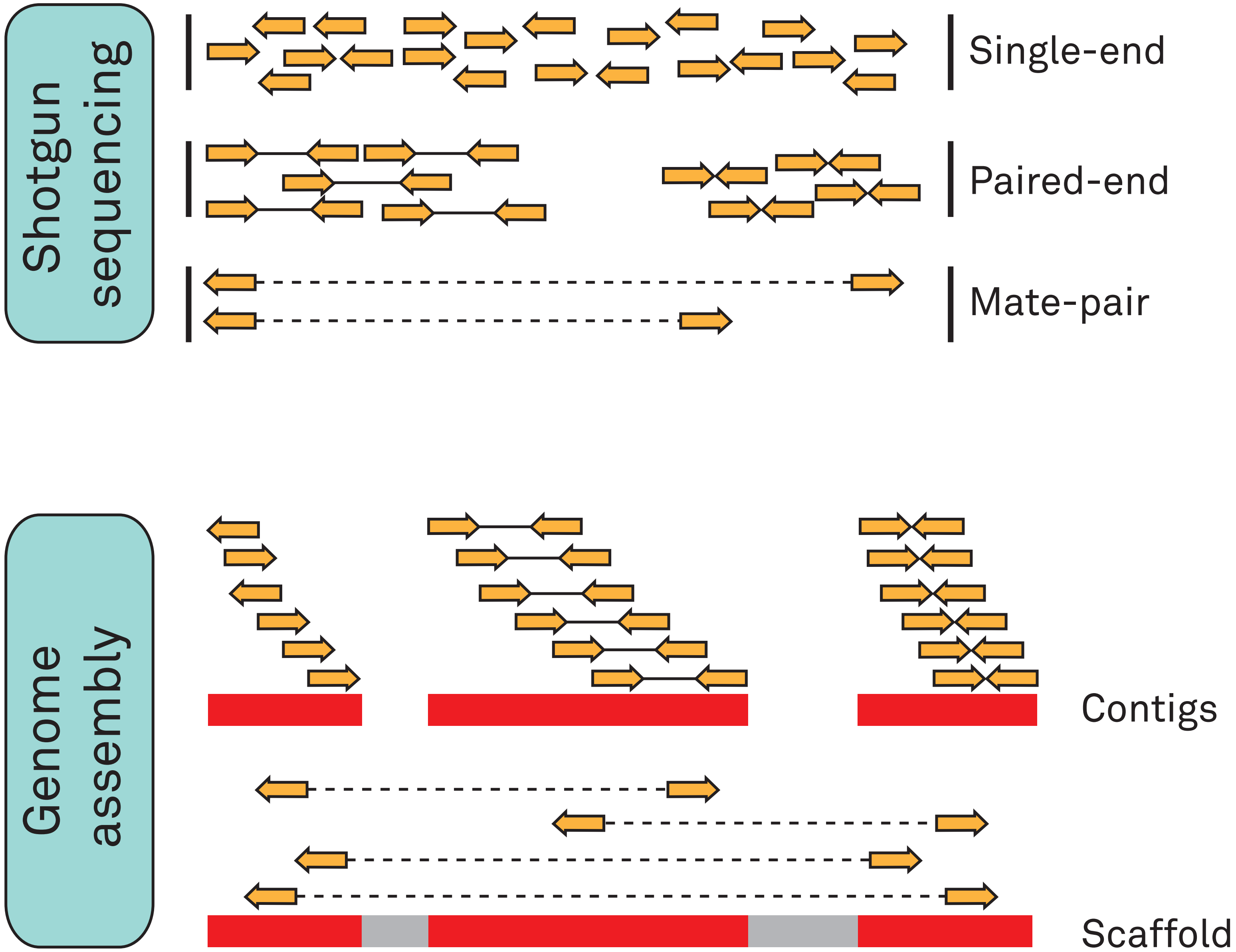
Mapping to a reference genome

What is a reference genome?

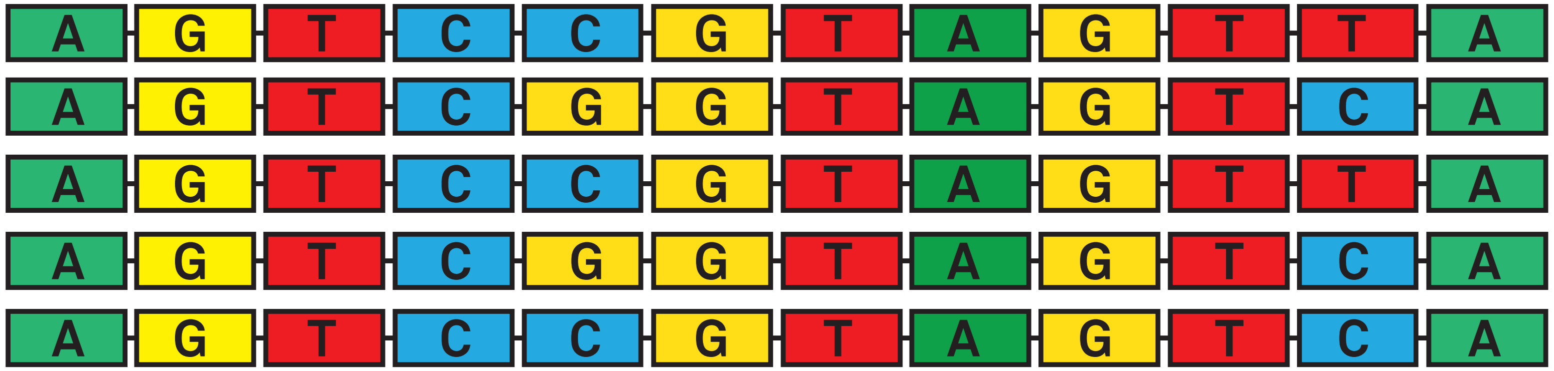
- Human genome >10 years and $500 million to $1 billion
- Now possible with ~$1000 and in a few days
- Sequence representing the actual genome of a species
- Single individual or in some cases, multiple individuals
- Akin to a type specimen used to identify a species
- Not necessarily complete

Assembling a reference genome

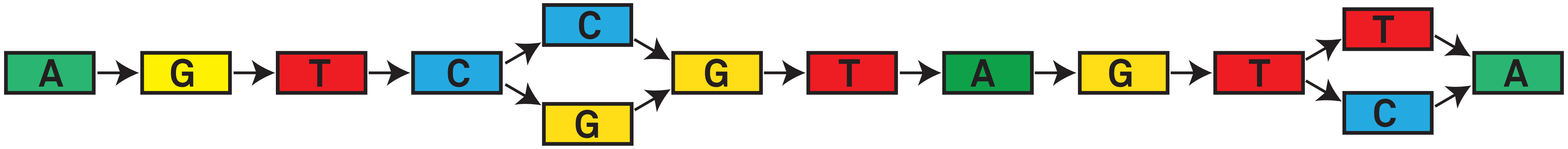


Reference genomes are not perfect

- How well does a single individual represent a species?
- Missing variation in the reference
- Drafts that can be improved over time with new technologies and methods.
- References from multiple individuals? Genome graphs and a shift from a linear perspective on genome structure

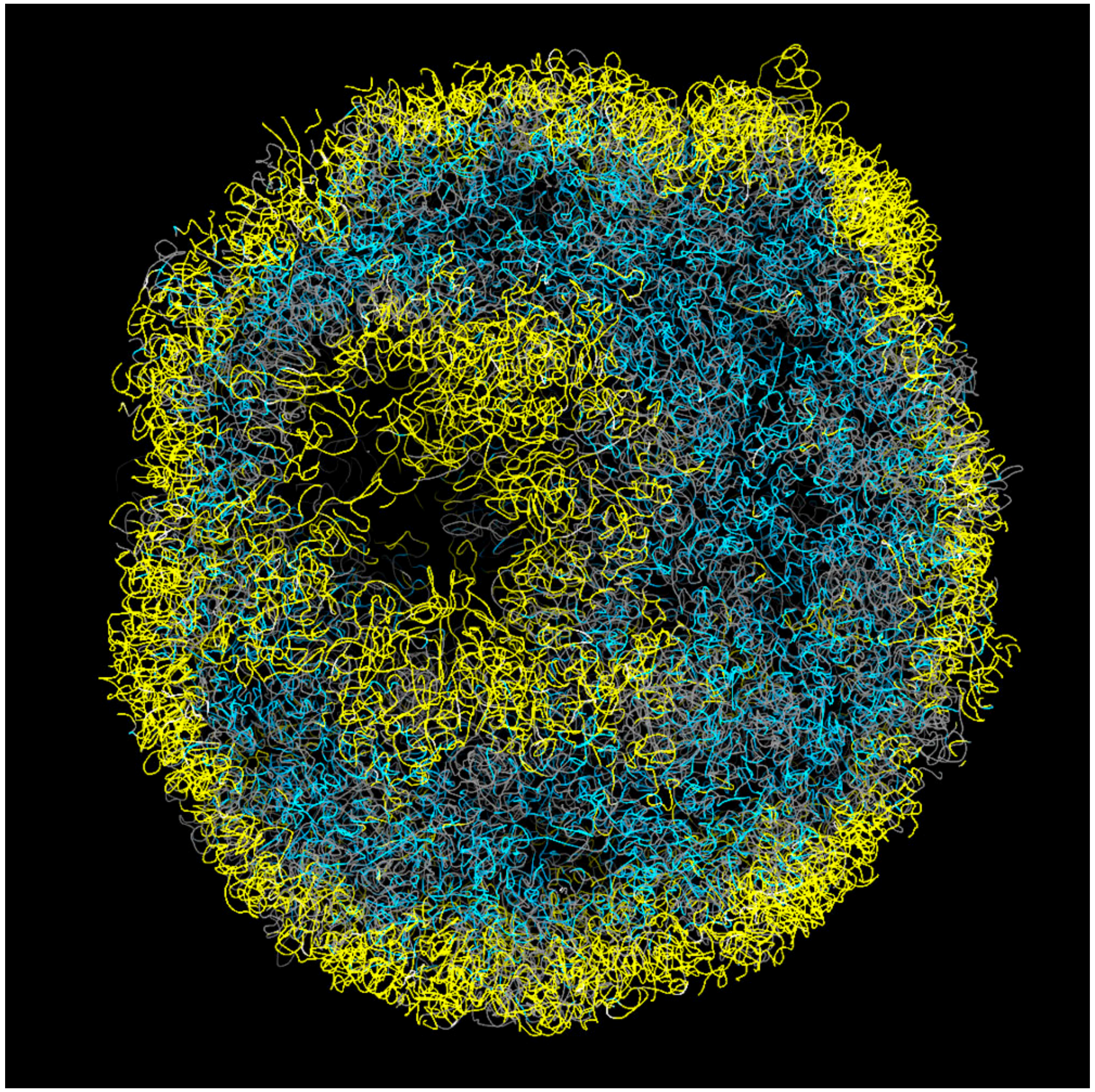


A G T C C G T A G T T A
A G T C G G T A G T C A
A G T C C G T A G T T A
A G T C G G T A G T C A
A G T C C G T A G T C A



Genomes in 3D and 4D?

- Greater appreciation of the 3D structure of genomes - i.e. Hi-C and chromatin binding sequencing
- Even attempts to characterise the genome in 4D - i.e. how structure changes with time



Variant calling - from raw reads to SNPs

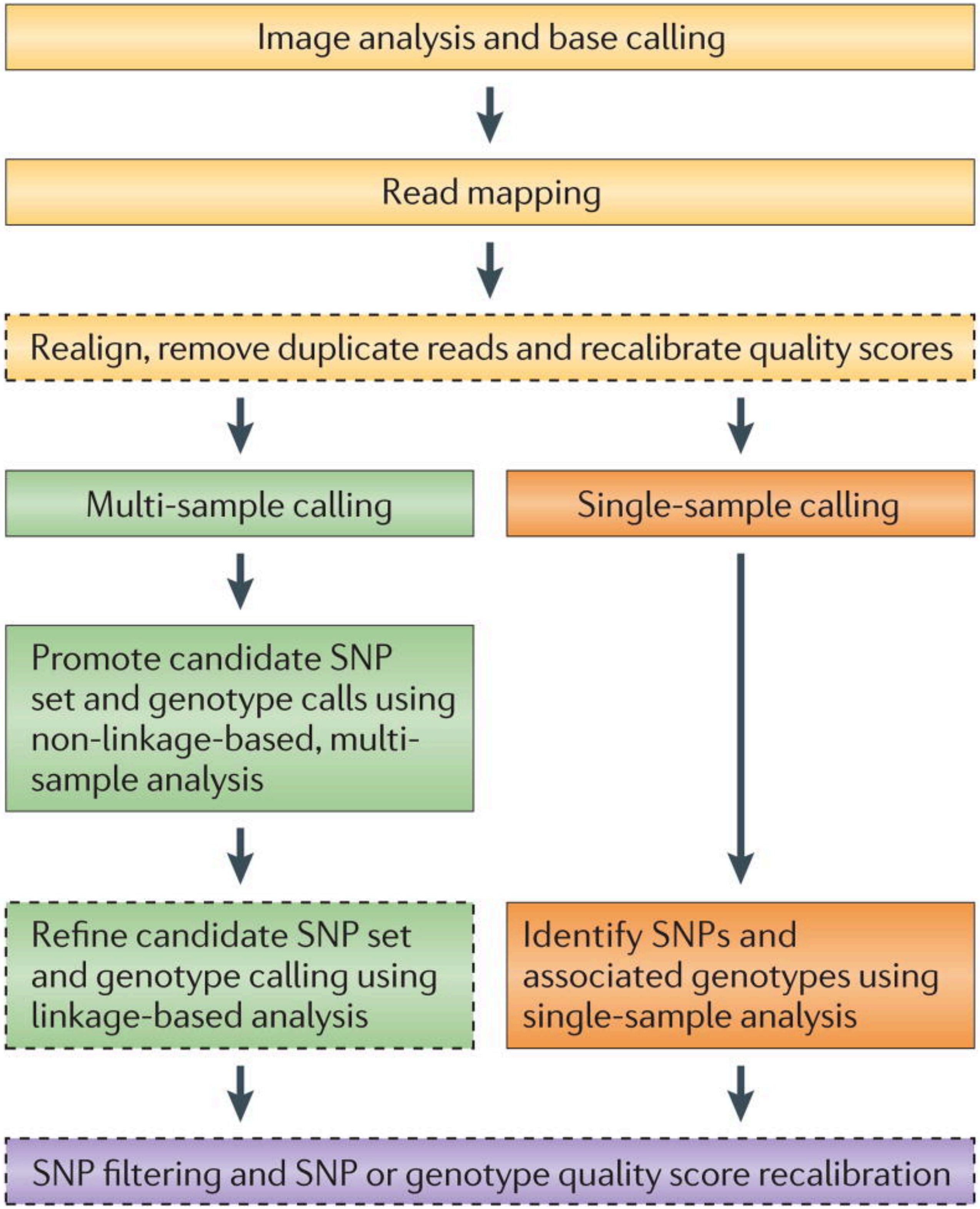


Image analysis and base calling

Read mapping

Realign, remove duplicate reads and recalibrate quality scores

Multi-sample calling

Single-sample calling

Promote candidate SNP set and genotype calls using non-linkage-based, multi-sample analysis

Refine candidate SNP set and genotype calling using linkage-based analysis

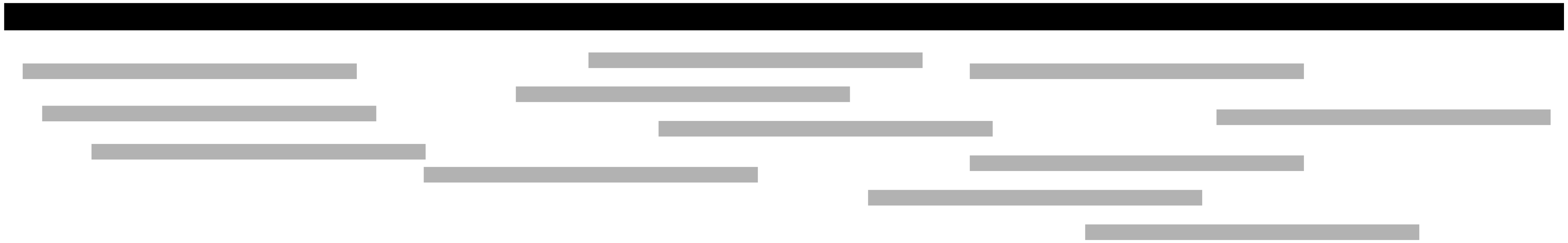
Identify SNPs and associated genotypes using single-sample analysis

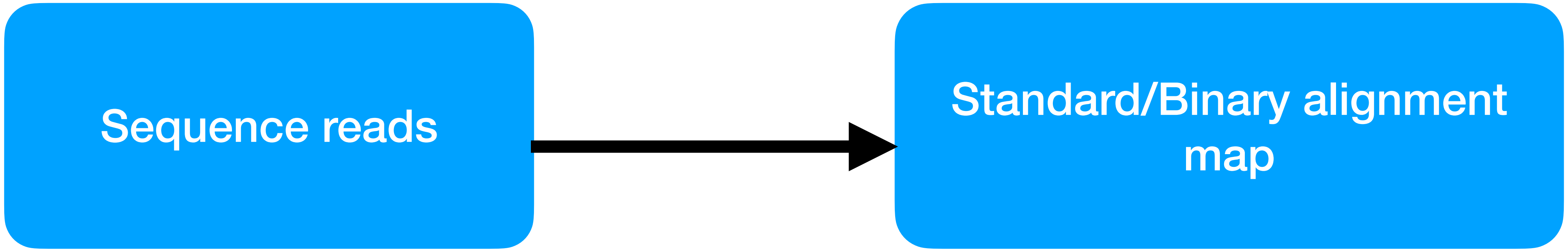
SNP filtering and SNP or genotype quality score recalibration

Nielsen 2011 **Nature Review Genetics**

Aligning to a reference genome

- Align reads to a reference genome
  - using short read aligner





Sequence reads

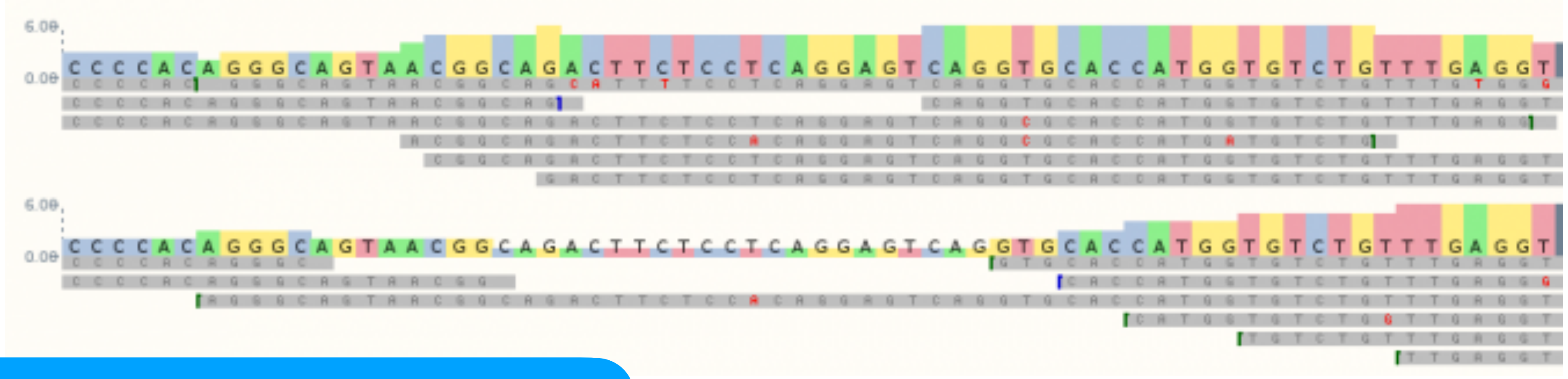
Standard/Binary alignment map

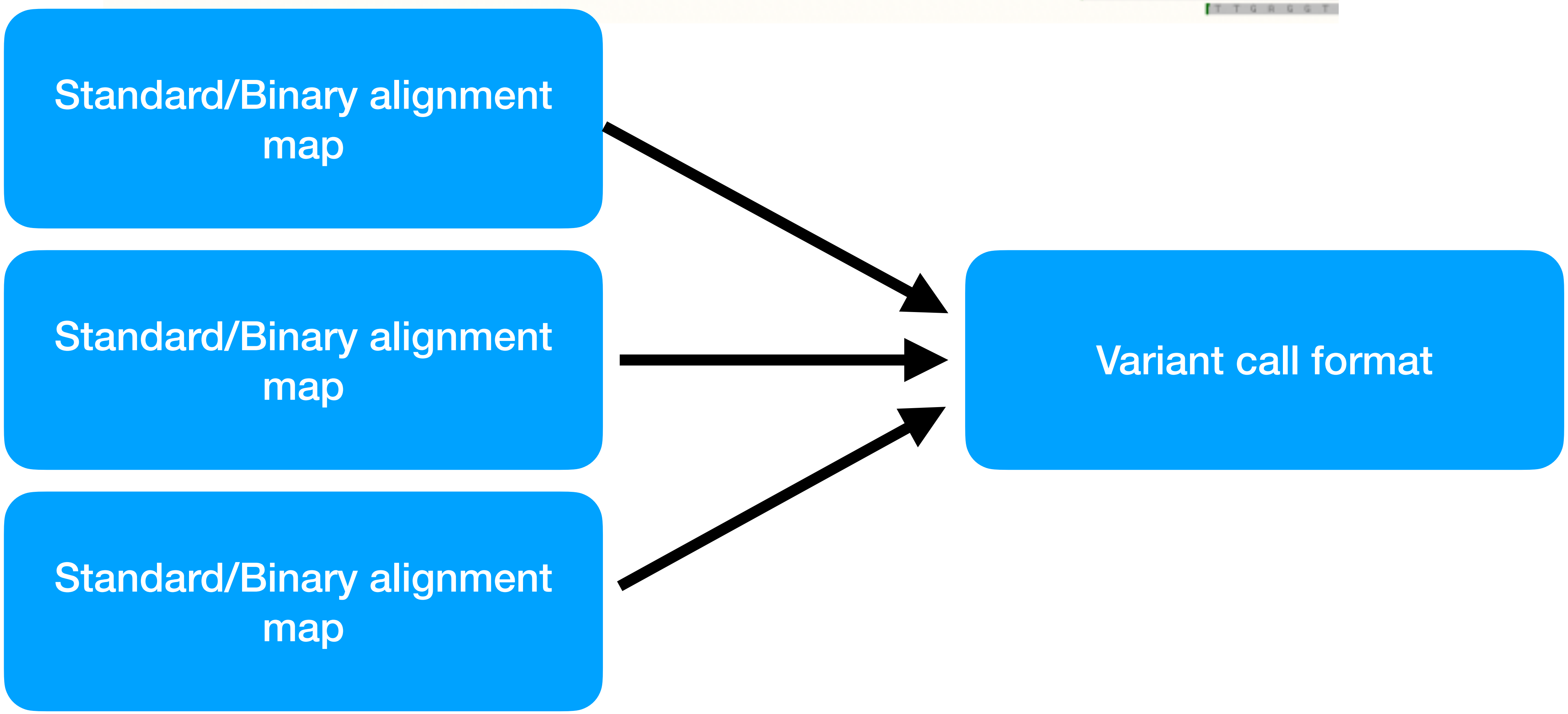
Per individual

How does alignment work?

- Index the genome - make a ‘hash’
- Match reads to genome locations based on partial matches
- Continually refine matching until ‘best hit’ location found
- Many short read aligners available - over 90!

Variant calling across multiple individuals





Standard/Binary alignment map

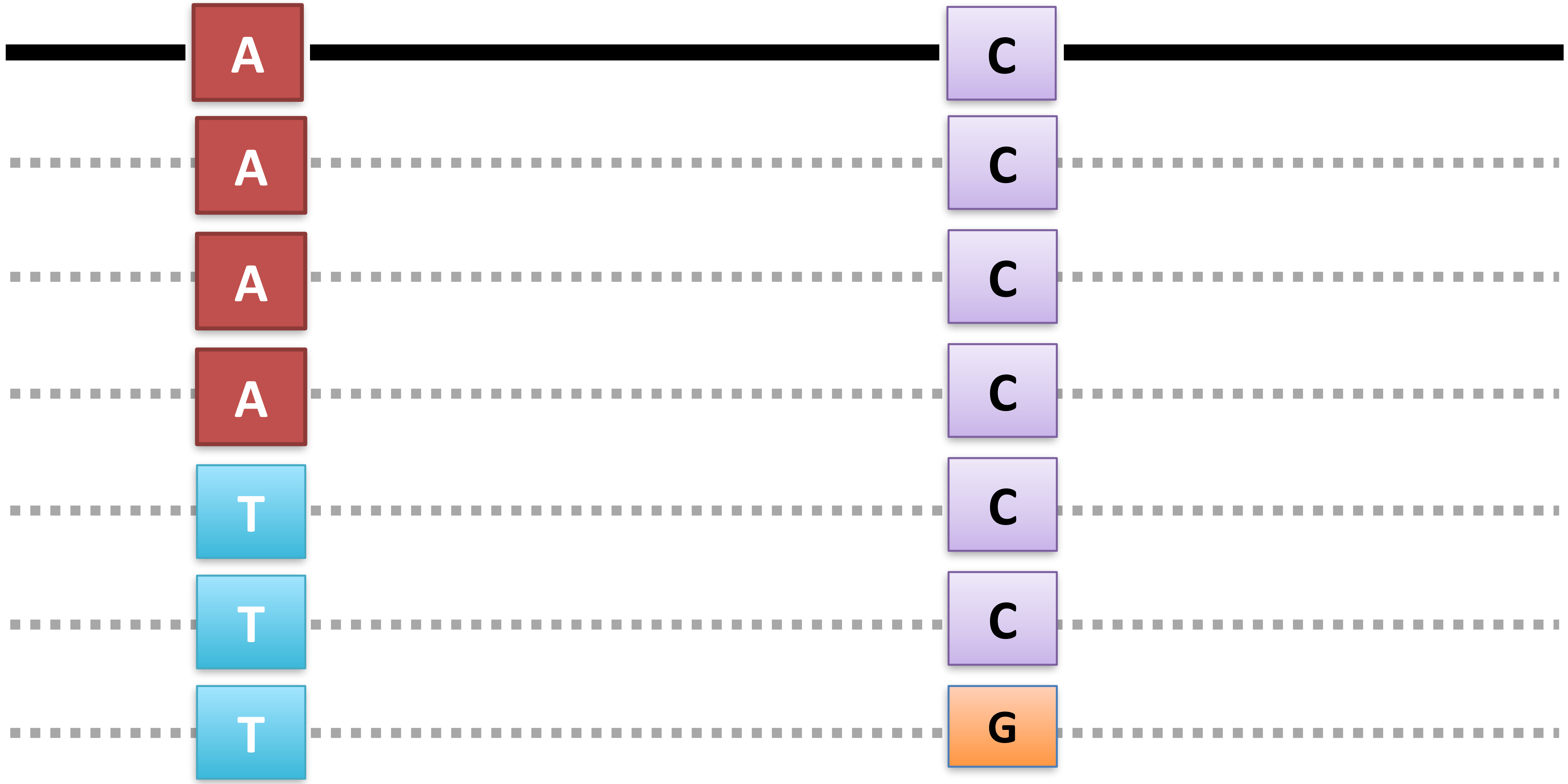
Standard/Binary alignment map

Standard/Binary alignment map

Variant call format

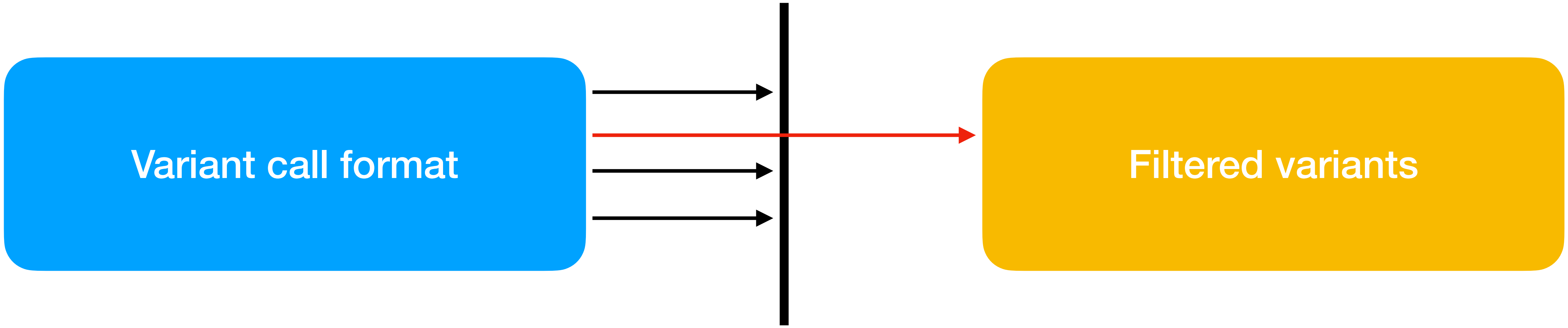
Variant calling from aligned reads

- variant calling tools such as bcftools, GATK, FreeBayes, Stacks, ANGSD
- Probabilistic models - Bayesian, Maximum likelihood



Filtering VCF to reduce error

- Error from:
  - library preparation
  - sequencing
  - alignment



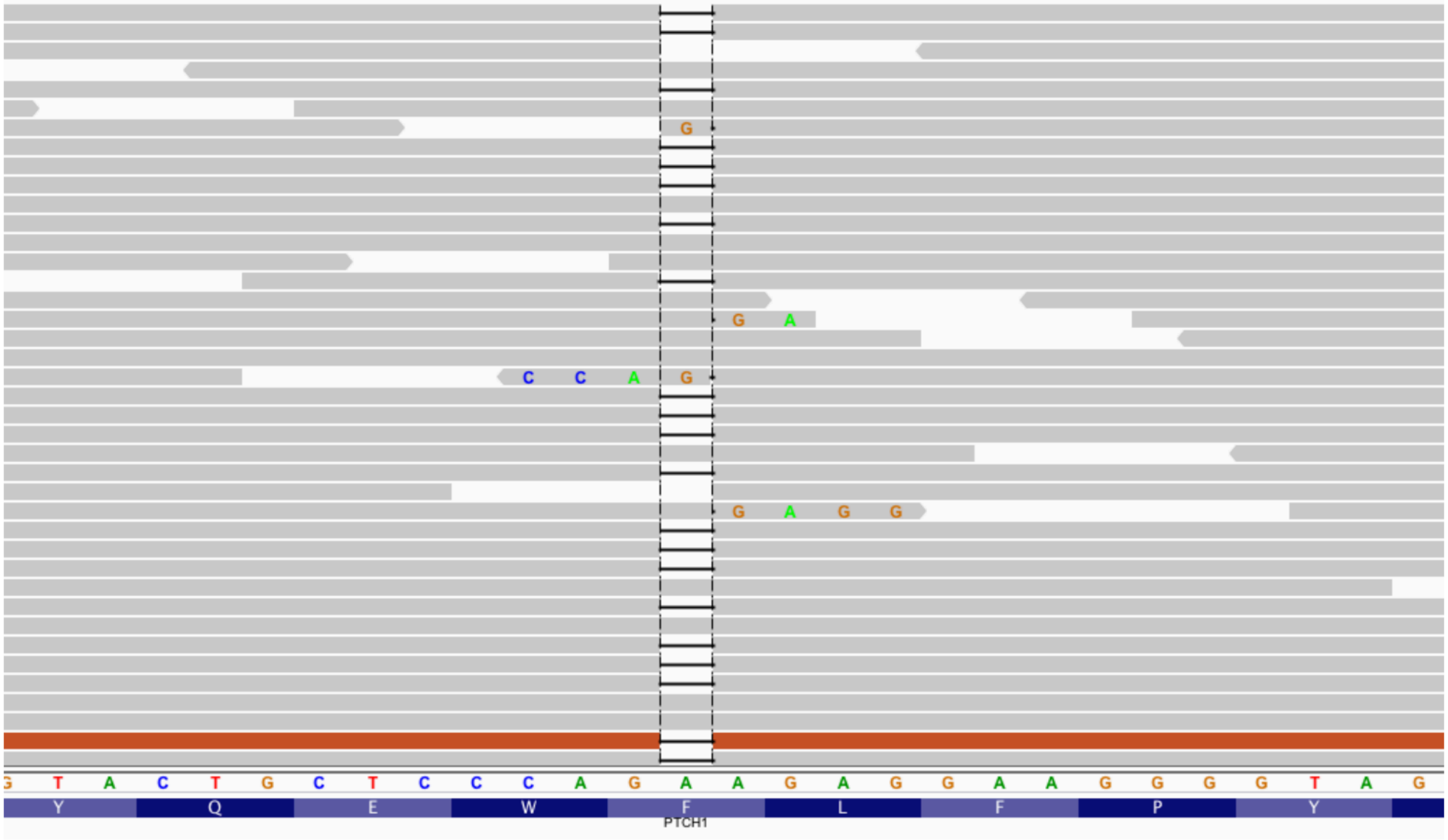
Variant call format

Filtered variants

- Apply stringent filters to remove them.
- Get used to throwing out some data!

Alignment error - indels

- indels hard to deal with and can cause errors around them because of poor alignment



What we will do today

- Map reads to a reference genome
- Call variants across individuals
- Filter these variants and learn about the Variant Call Format