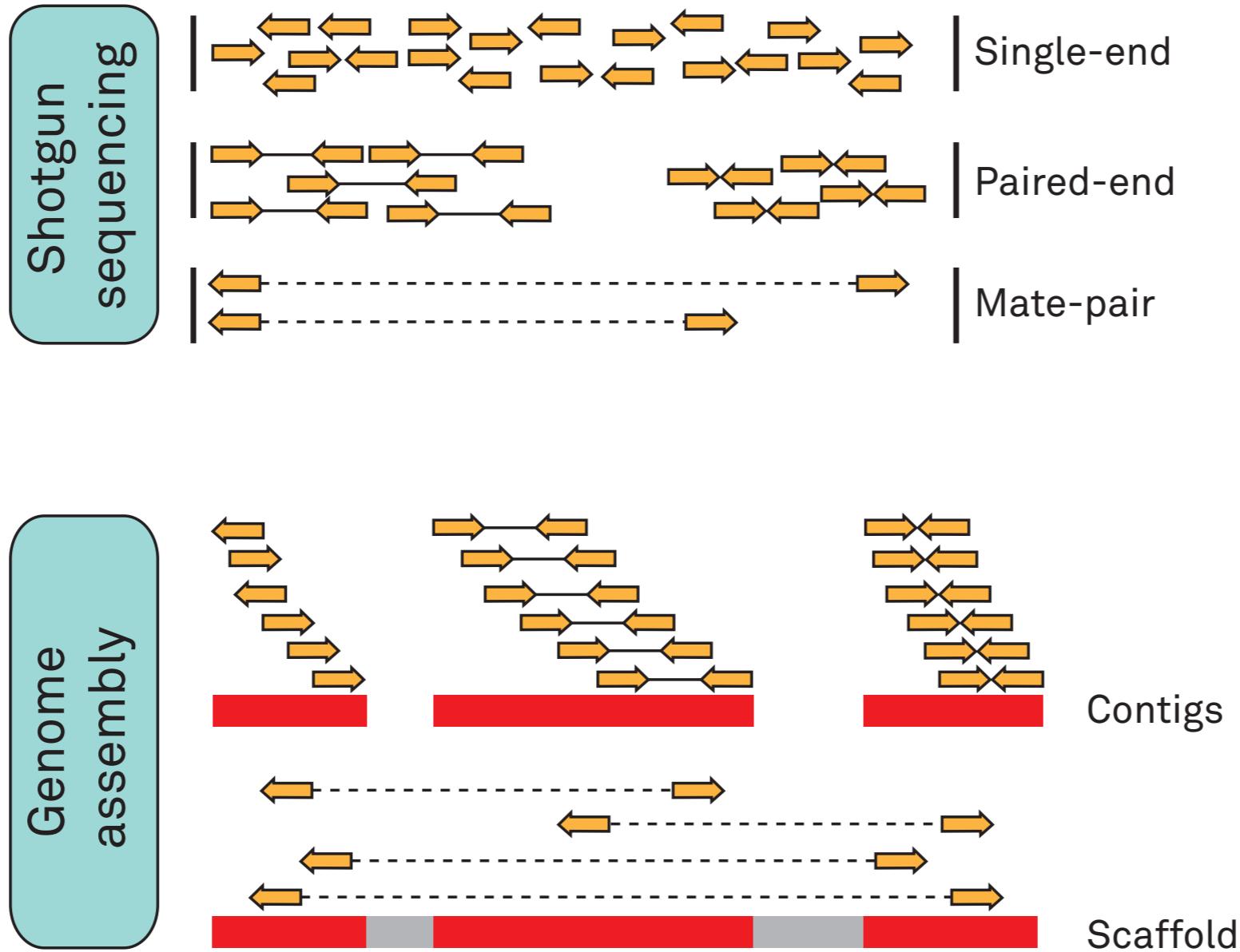


# Mapping to a reference genome

# What is a reference genome?

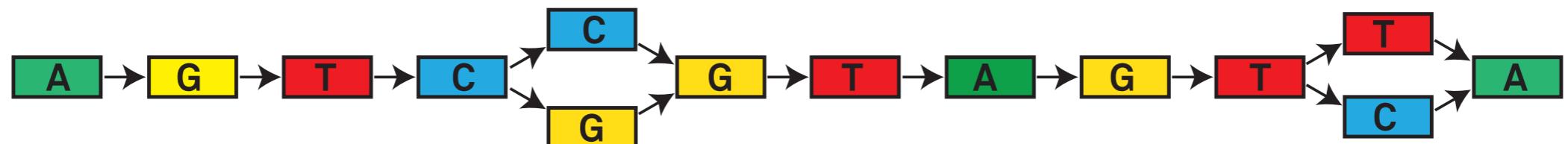
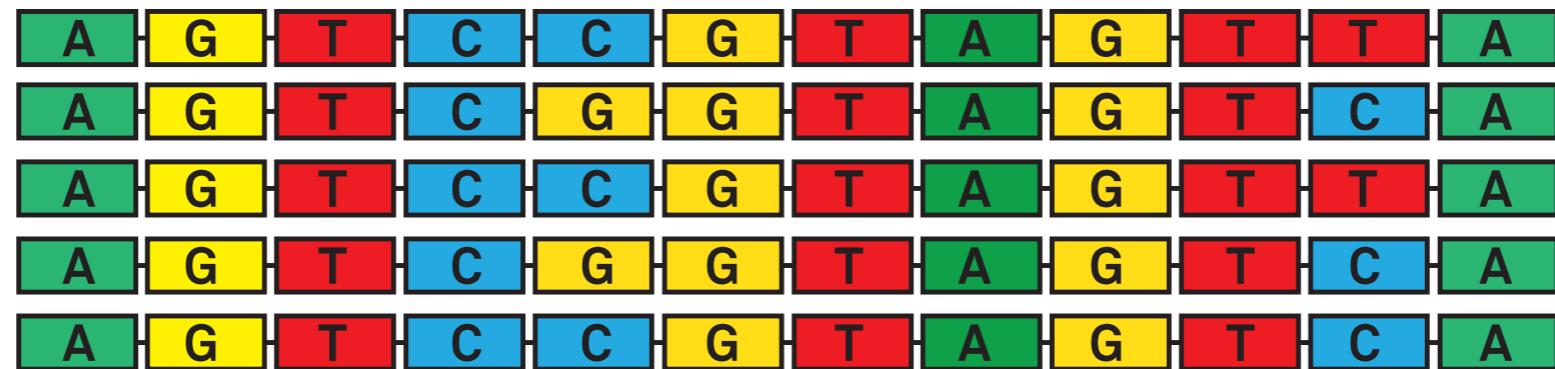
- Human genome >10 years and \$500 million to \$1 billion
- Now possible with ~\$1000 and in a few days
- Sequence representing the actual genome of a species
- Single individual or in some cases, multiple individuals
- Akin to a type specimen used to identify a species
- Not necessarily complete

# Assembling a reference genome

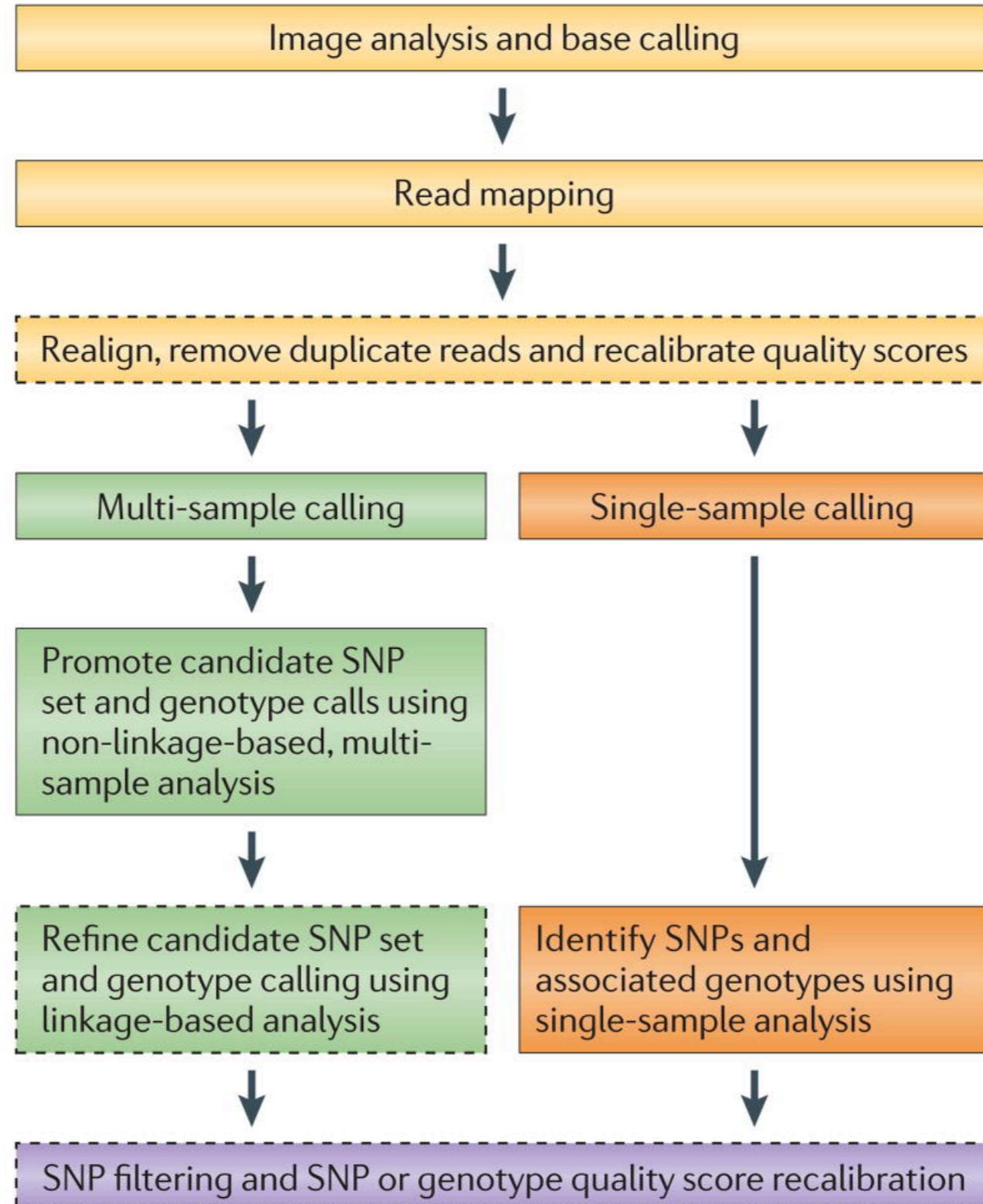


# Reference genomes are not perfect

- 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

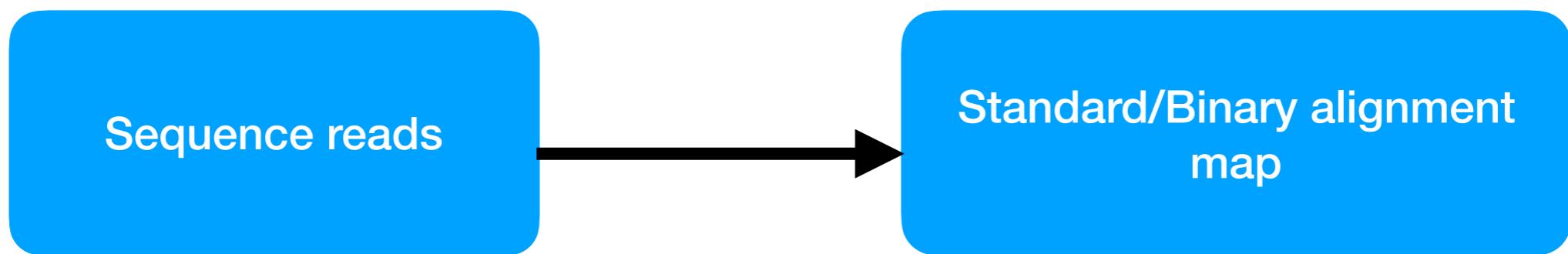


# Variant calling - from raw reads to SNPs



# Aligning to a reference genome

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



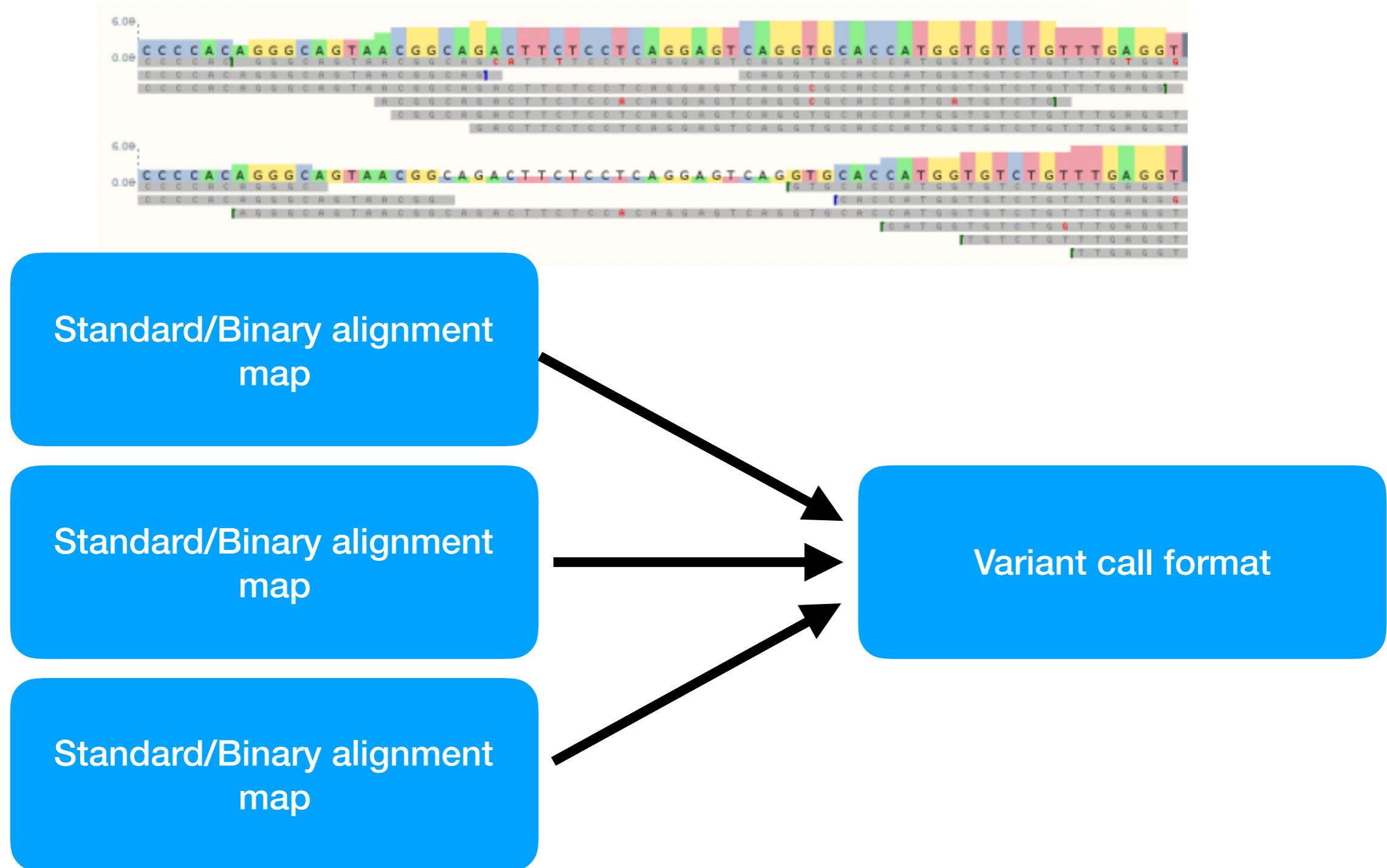
**Per individual**

# 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!

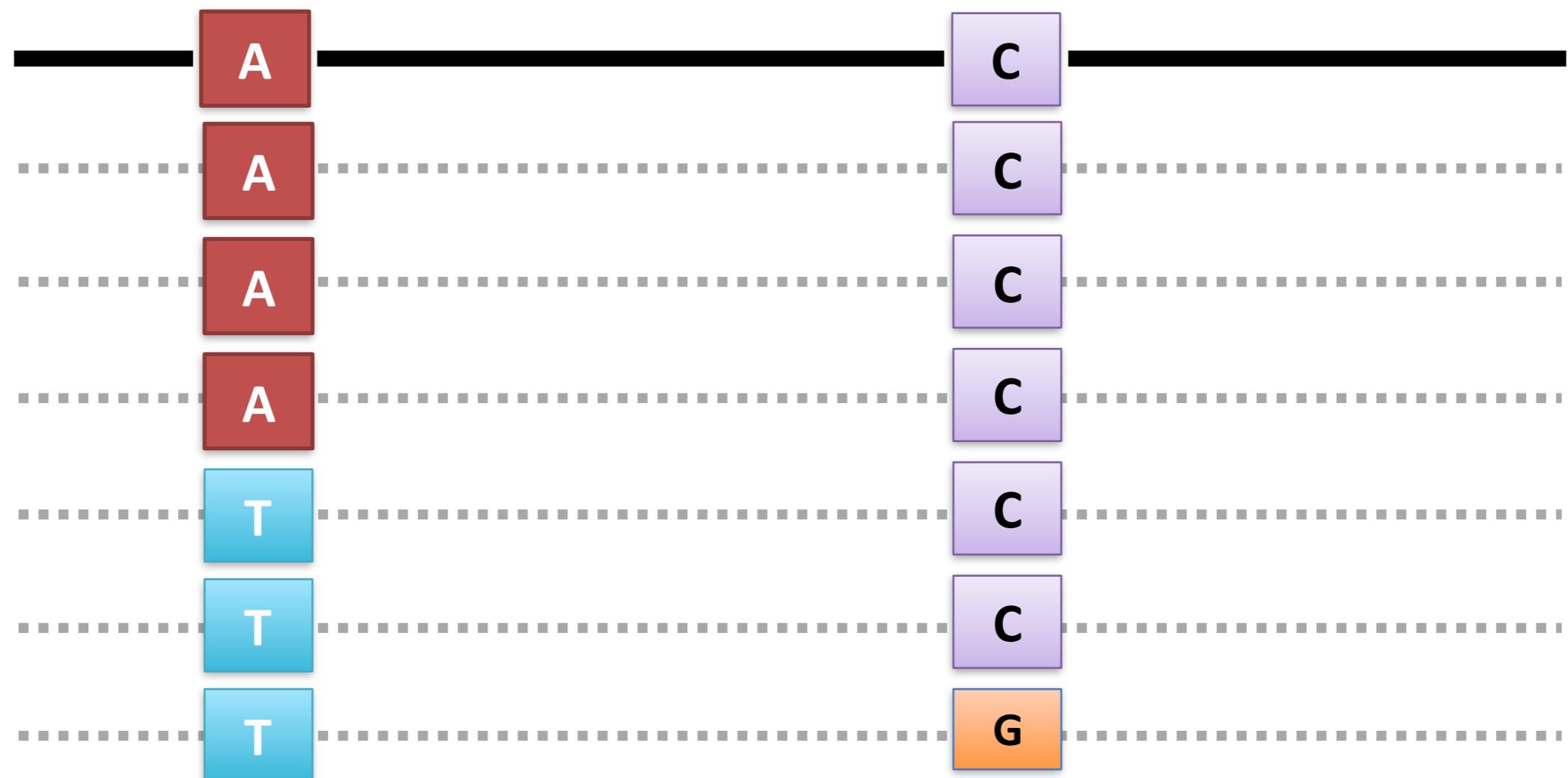


# Variant calling across multiple individuals

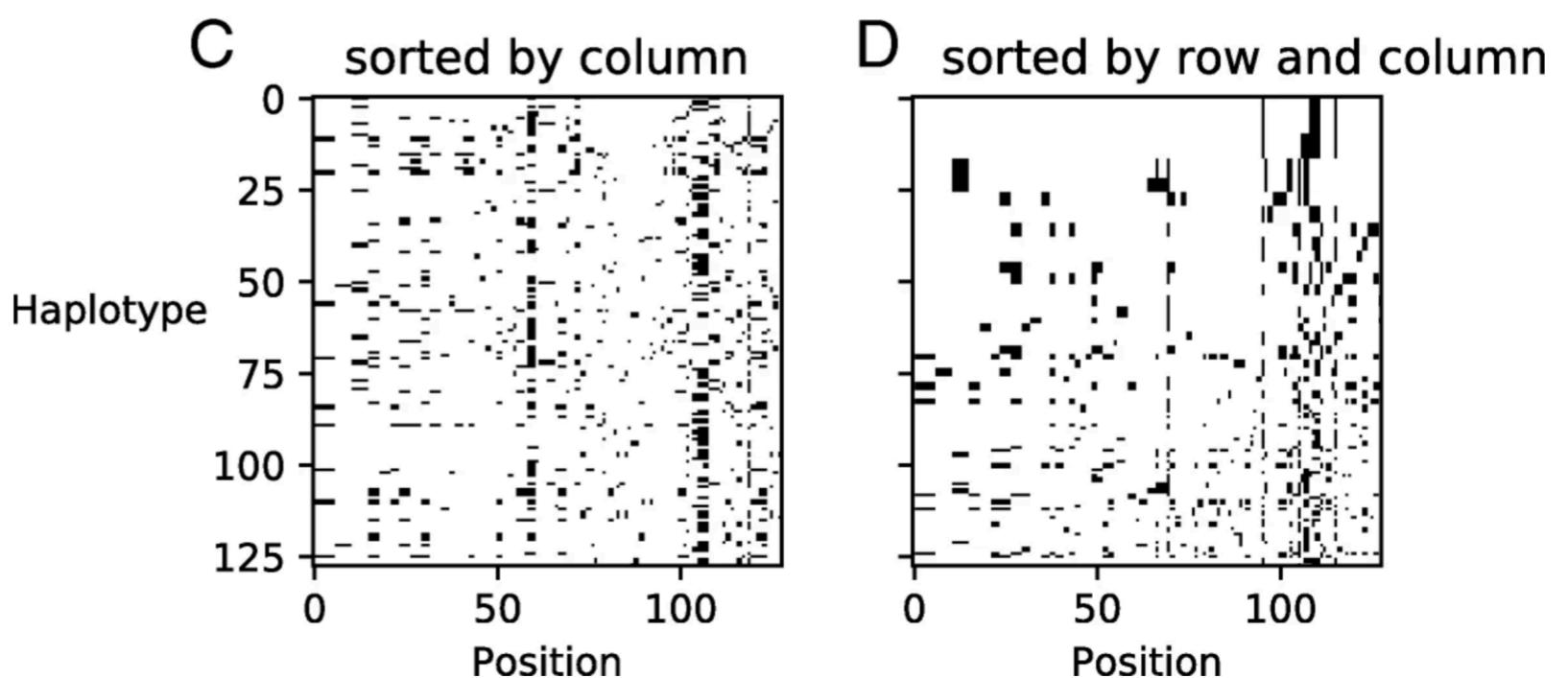
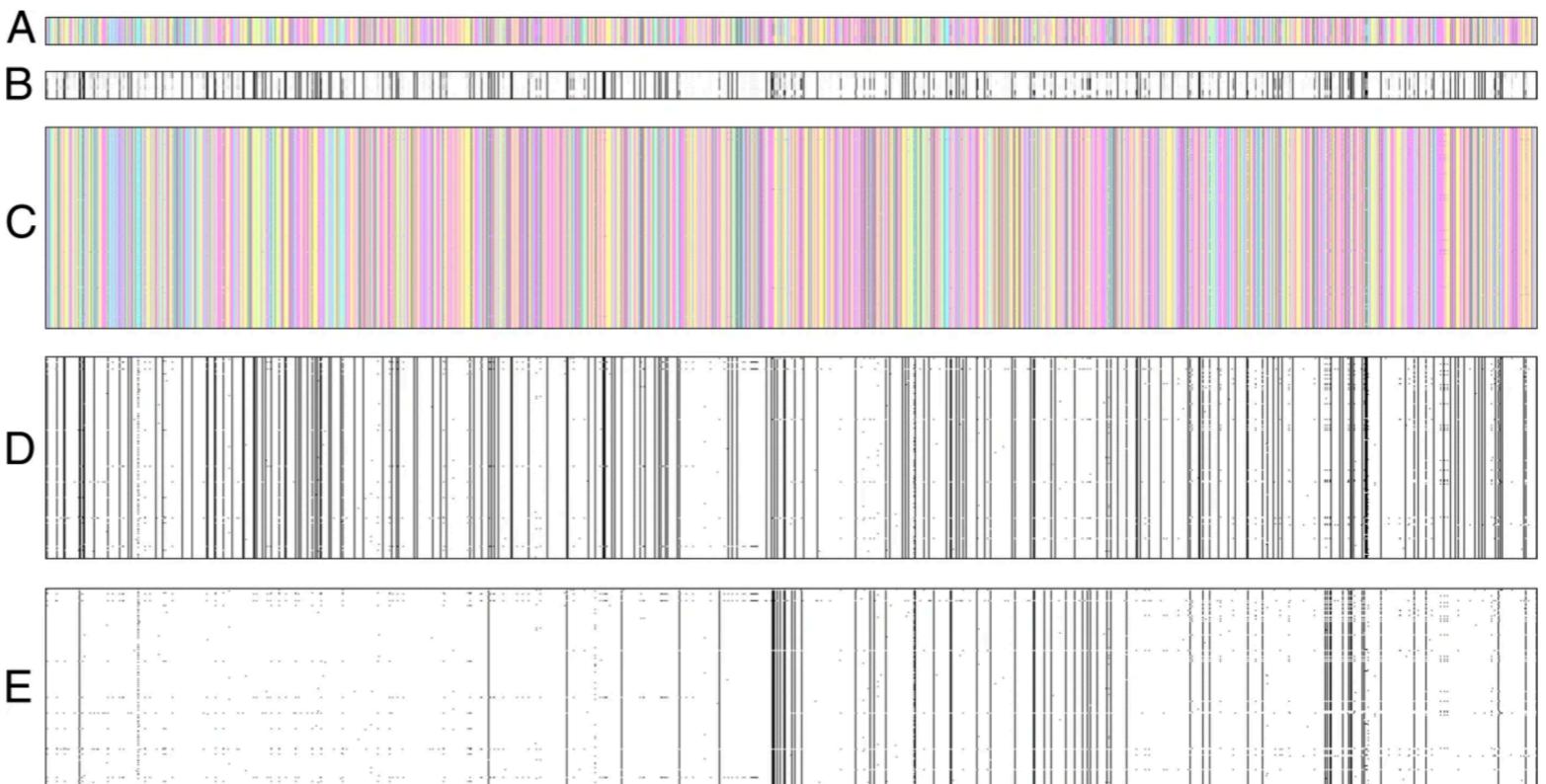
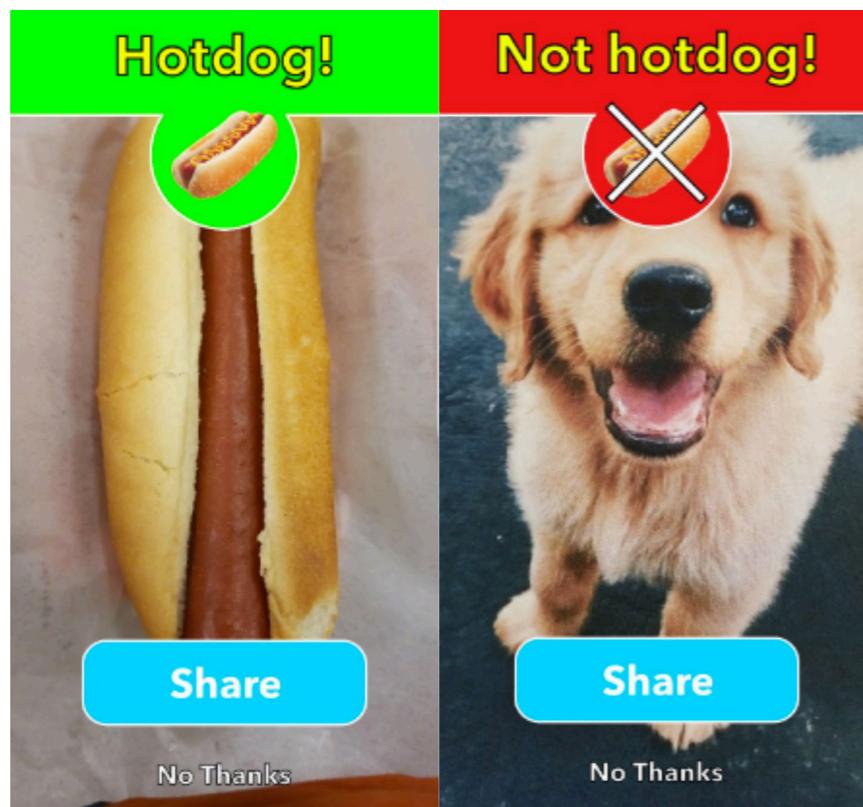


# Variant calling from aligned reads

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

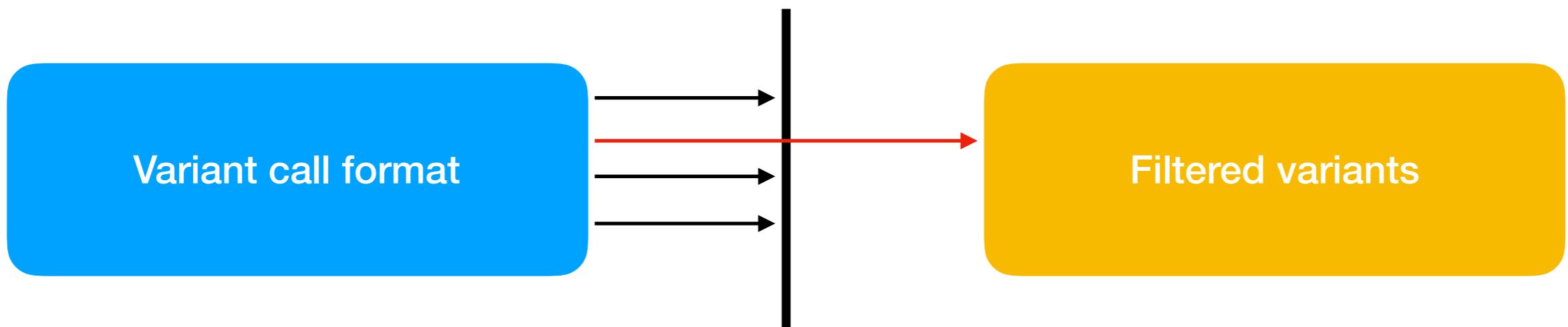


# Variant calling- the future?



# Filtering VCF to reduce error

- Error from:
  - library preparation
  - sequencing
  - alignment



- 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

