DNA sequencing and variants
VEuPathDB Workshop 2021

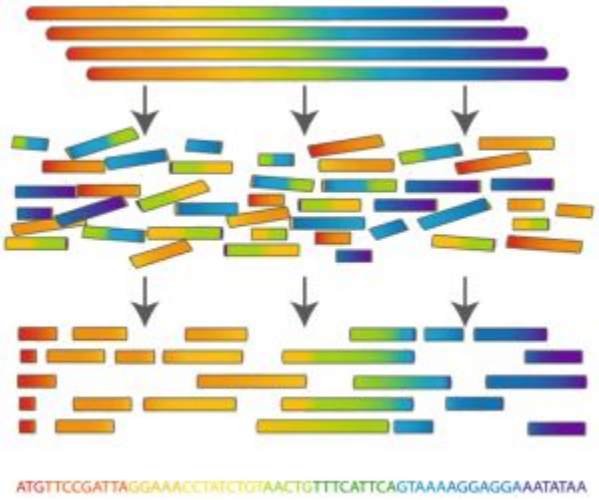
Kathryn Crouch
kathryn.crouch@glasgow.ac.uk

Why Do We Sequence Genomes?

- To create a reference
- To do comparative studies
  - Compare a free living organism vs a parasite
  - Compare virulent vs avirulent strains
- To understand how a species responds to pressure
  - Changes under drug pressure
  - Changes under metabolic pressure
  - Change under environmental pressure
- To understand how species diversify
  - Population dynamics
- To understand how species are related
  - Phylogenetics

De Novo Assembly

- All sequencing technologies fragment DNA to some extent
- *De novo* assembly aims to reconstruct a genome from the fragments
- Easier to do with a sequencing technology that generates longer reads (PacBio or Nanopore)
- Applications:
  - To generate a genome for a completely new organism
  - To assess regions that vary highly between organisms (surface antigens, immunoglobulins)



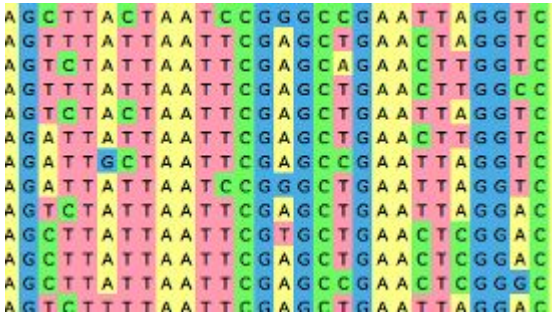
Why Do We Sequence Genomes?

- To create a reference
- To do comparative studies
  - Compare a free living organism vs a parasite
  - Compare virulent vs avirulent strains
- To understand how a species responds to pressure
  - Changes under drug pressure
  - Changes under metabolic pressure
  - Change under environmental pressure
- To understand how species diversify
  - Population dynamics
- To understand how species are related
  - Phylogenetics

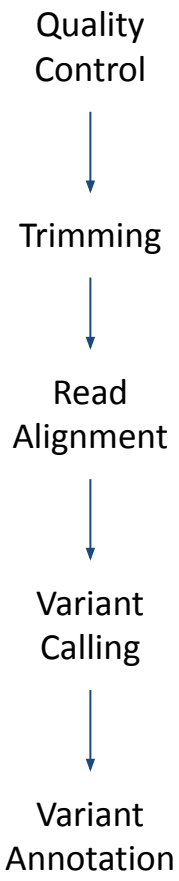
Exploring Sequence Variation

Exploring Sequence Diversity

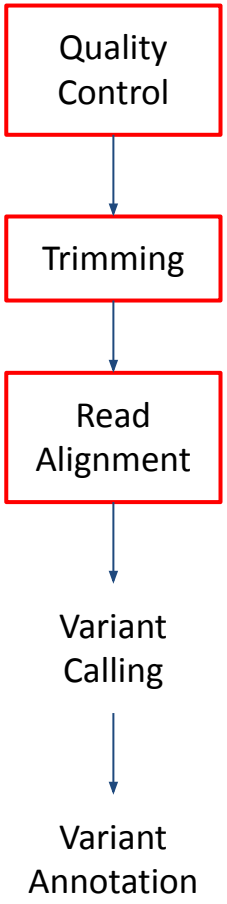
- Applications
  - Population biology
  - Phylogeny
  - Comparative studies
- Low error rates are important to explore sequence variation
- Illumina is a commonly used technology
- Analysis uses alignment to compare sequences



DNA Sequencing Analysis



DNA Sequencing Analysis



Does this look familiar?!

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

Read Alignment

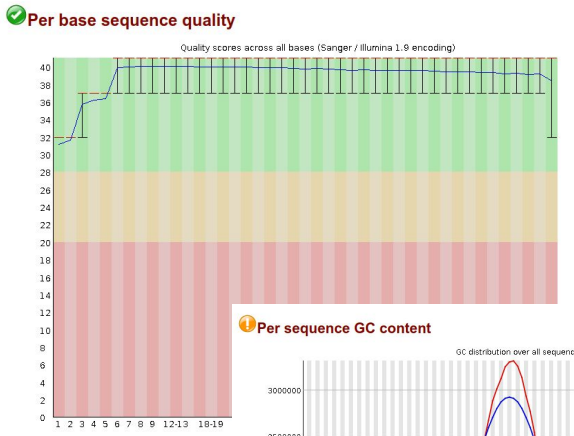
↓

Variant Calling

↓

Variant Annotation

- FASTQC https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - Overall sequencing quality
  - GC content
  - N content
  - Read length distribution
  - Over-represented sequences
  - Adaptor content
- Output is an html file that can be opened in a web browser



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ACAAGTGTGTAACATTAATTTGCAAGTTTGCAACGCTGTTCTTTAGTGTT	70896	0.12562741276052788	No Hit

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

Read Alignment

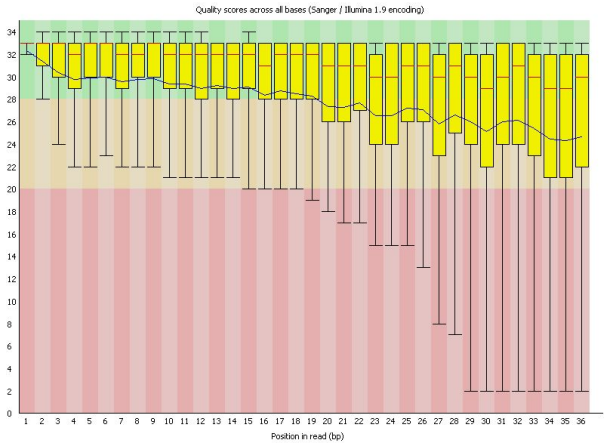
↓

Variant Calling

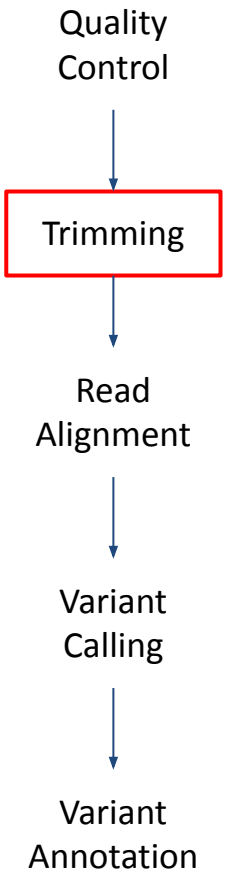
↓

Variant Annotation

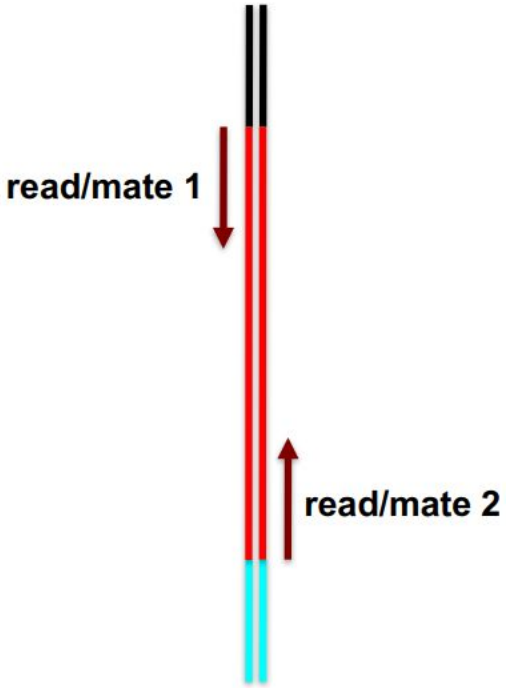
- Sickle https://github.com/najoshi/sickle
- Sequence quality tends to decrease towards the 3' end of reads - these can affect mapping
- Sickle will:
  - Remove poor quality reads from the 3' end of each read
  - Check for reads that are too short and discard them
  - Check that all reads still have a pair and discard those that don't
- Some trimming tools can also remove adaptors



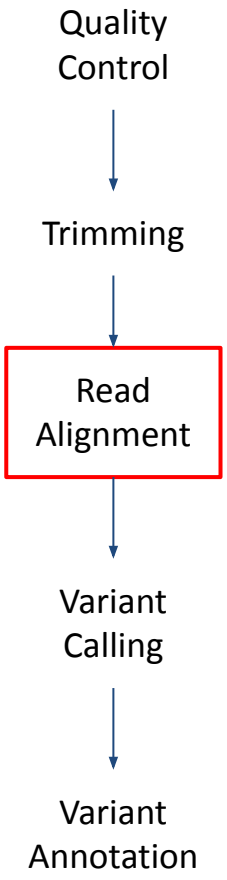
DNA Sequencing Analysis



- Sickle https://github.com/najoshi/sickle
- Sequence quality tends to decrease towards the 3' end of reads - these can affect mapping
- Sickle will:
  - Remove poor quality reads from the 3' end of each read
  - Check for reads that are too short and discard them
  - Check that all reads still have a pair and discard those that don't
- Some trimming tools can also remove adaptors



DNA Sequencing Analysis



What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

```
--ATTGAAA-GCTA
  | ||||| |
GAAATGAAAAGG--

ATTGAAA-GCTA---
   |||| |   |
---GAAATGAAAAGG
```

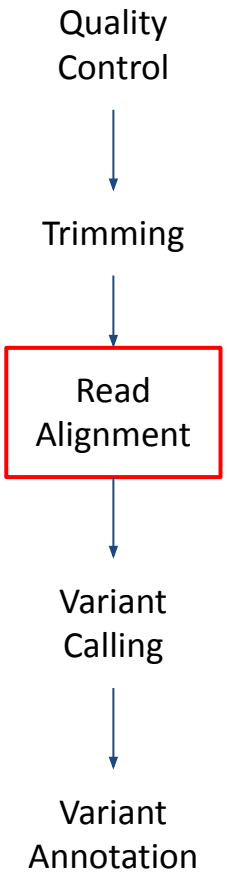
Which one is better??

Alignment scoring:

- 1 for a match
- -1 for a mismatch
- -2 for a gap

Now alignment 1 scores -4 and alignment 2 scores -10 so we would choose alignment 1

DNA Sequencing Analysis



What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

```
--ATTGAAA-GCTA
  | ||||| |
GAAATGAAAAGG--
```

```
ATTGAAA-GCTA---
   |||| |   |
---GAAATGAAAAGG
```

Which one is better??

Alignment scoring:

- 1 for a match
- -1 for a mismatch
- -2 for a gap

Now alignment 1 scores -4 and alignment 2 scores -10 so we would choose alignment 1

When aligning genomic sequences to each other we do not need to worry about splicing

- BWA: https://github.com/lh3/bwa
- Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

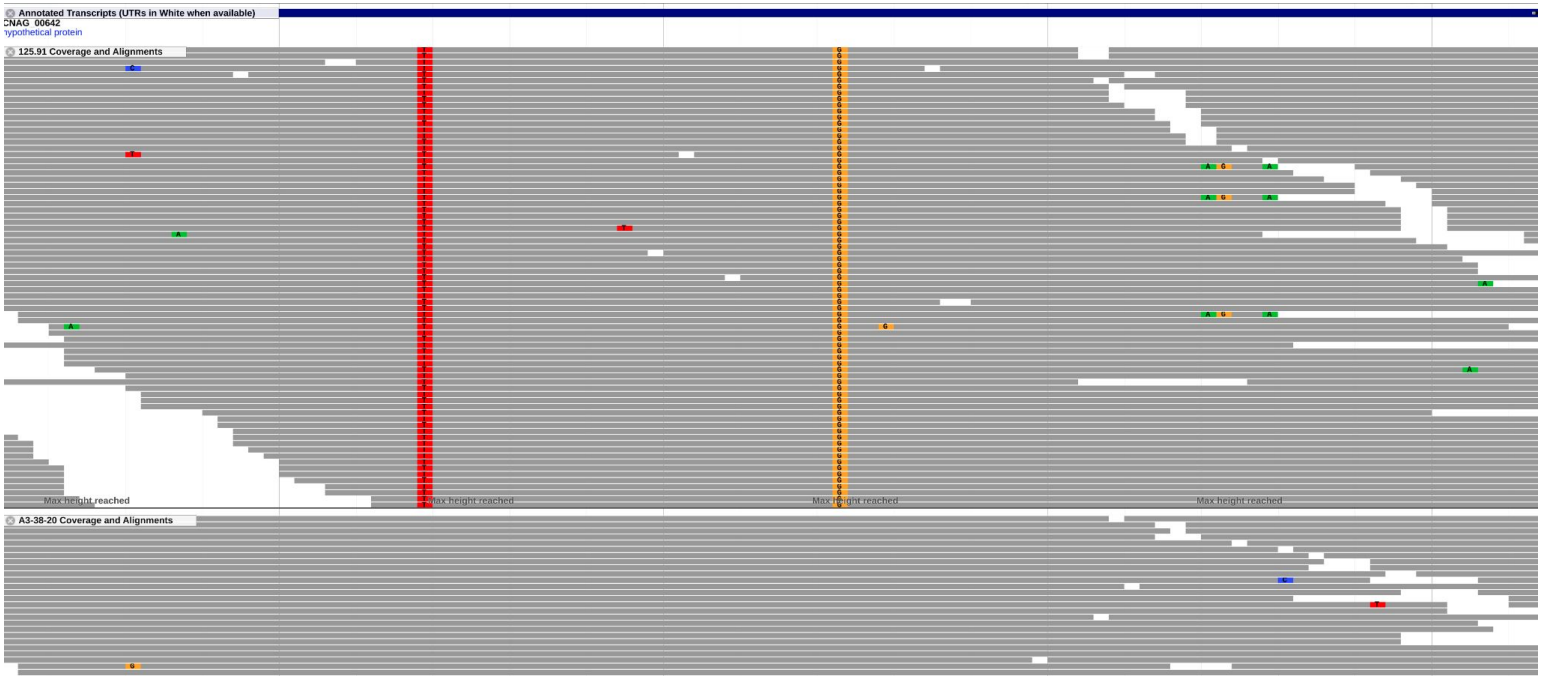
Read Alignment

↓

Variant Calling

↓

Variant Annotation



Finding Variants

At this point, we can load our alignment into a genome browser and see variants

How do we find them globally? How do we assess them?

https://tinyurl.com/2de356ju

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

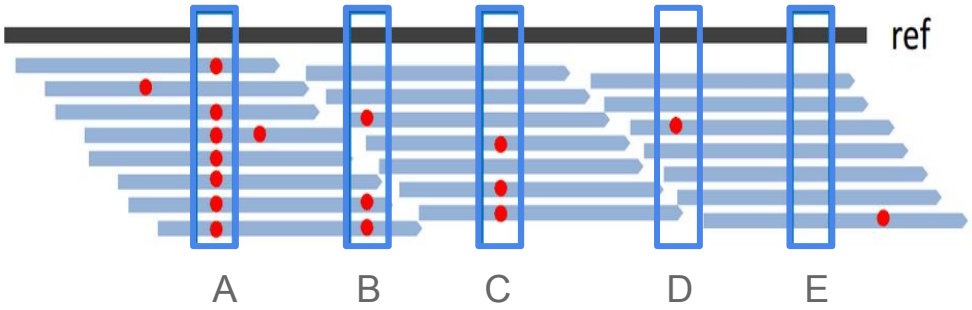
Read Alignment

↓

Variant Calling

↓

Variant Annotation



Blue lines are reads aligned against a reference (black). Red dots indicate individual bases where a base in a read differs from the reference.

A: Most reads differ from the reference -> homozygous SNP
B and C: Roughly 50% of reads differ from the reference -> potential heterozygous SNP
D: Only one base differs from the reference -> probably a sequencing error
E: All bases the same as the reference

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

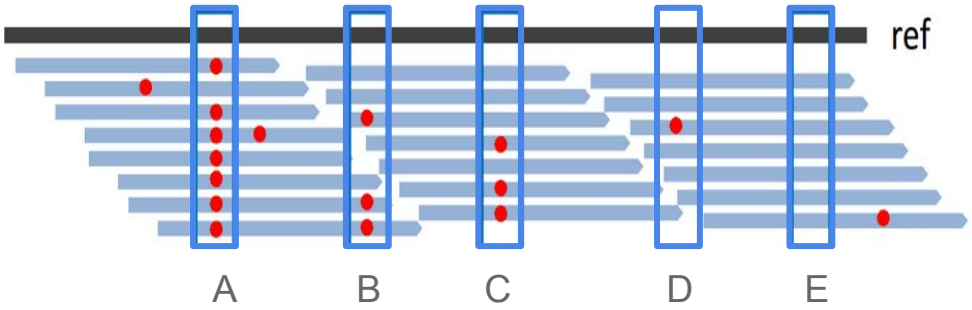
Read Alignment

↓

Variant Calling

↓

Variant Annotation



Things to think about:

- Allelic ratios assume a clone (culture or sample from an individual). In a population, these will not hold up
- Illumina has an accuracy of 90%
  - Deeper sequencing can help distinguish real variants from sequencing errors - but only to a point
  - Too much depth can introduce noise

DNA Sequencing Analysis

Quality Control

↓

Trimming

↓

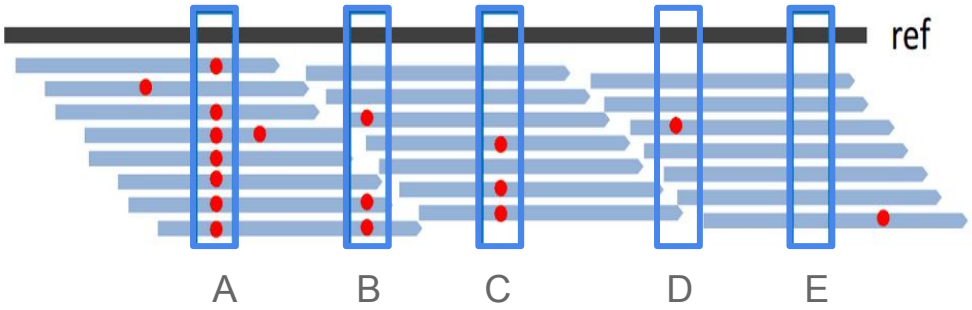
Read Alignment

↓

Variant Calling

↓

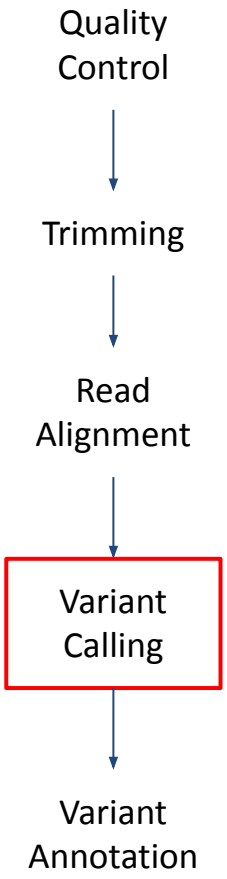
Variant Annotation



Automated tool to call SNPS

BCFtools http://samtools.github.io/bcftools/bcftools.html
GATK https://gatk.broadinstitute.org/hc/en-us
FreeBayes https://github.com/ekg/freebayes

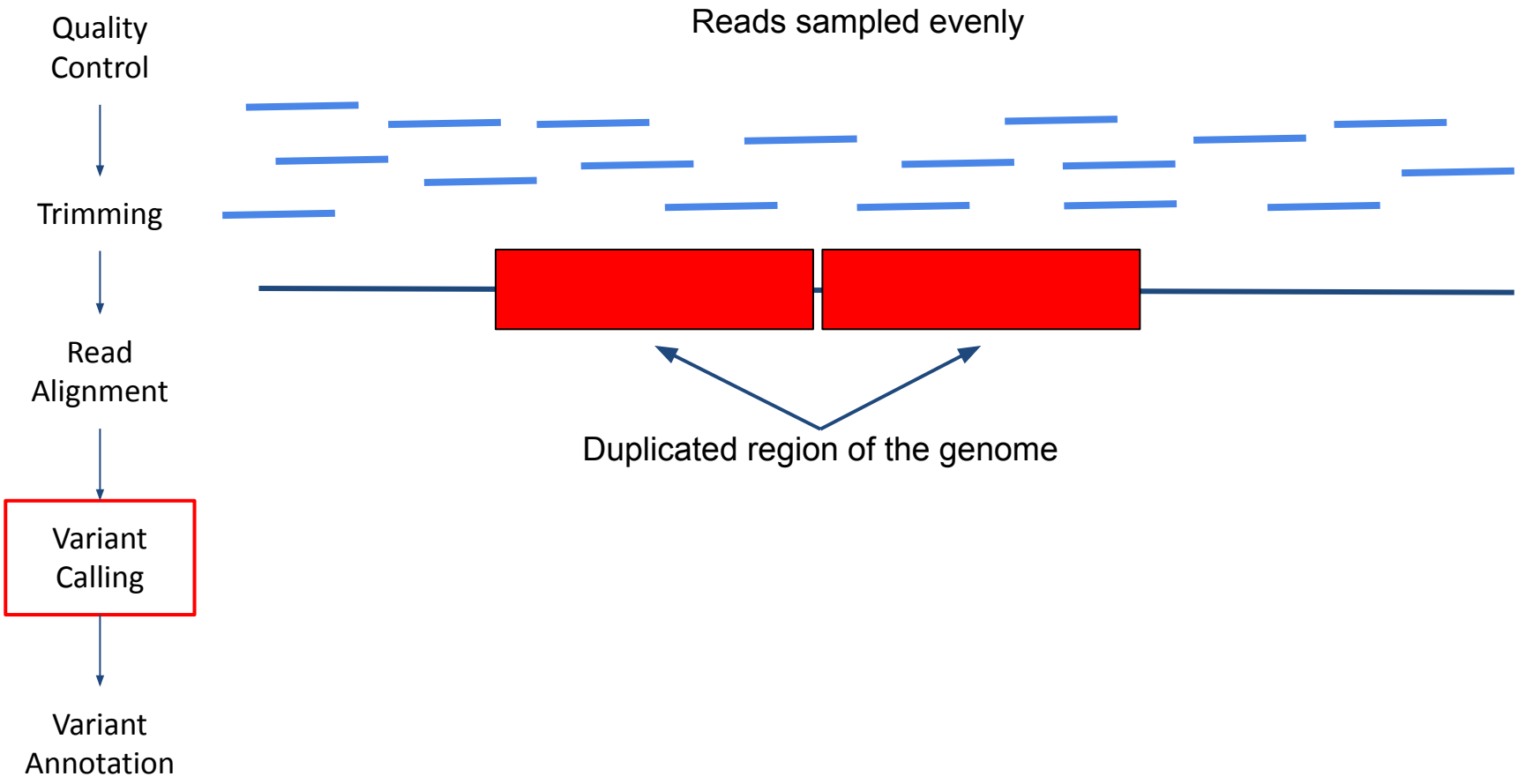
DNA Sequencing Analysis



Copy Number Variation

- Expect coverage to be even across the genome
- In reality, we see local variation associated with:
  - GC content
  - Repetitive or highly variable regions
- Changes in coverage can also tell us about copy number variations

DNA Sequencing Analysis



DNA Sequencing Analysis

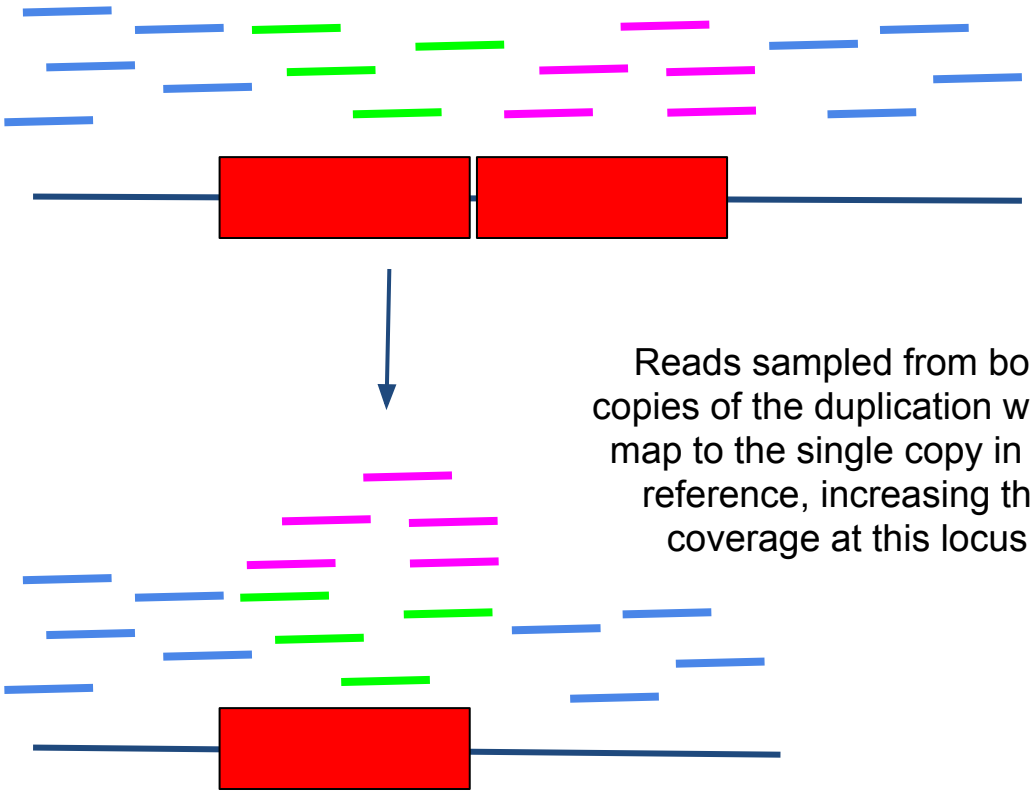
Quality Control

Trimming

Read Alignment

Variant Calling

Variant Annotation



Reads sampled from both copies of the duplication will all map to the single copy in the reference, increasing the coverage at this locus

Reference genome without duplication

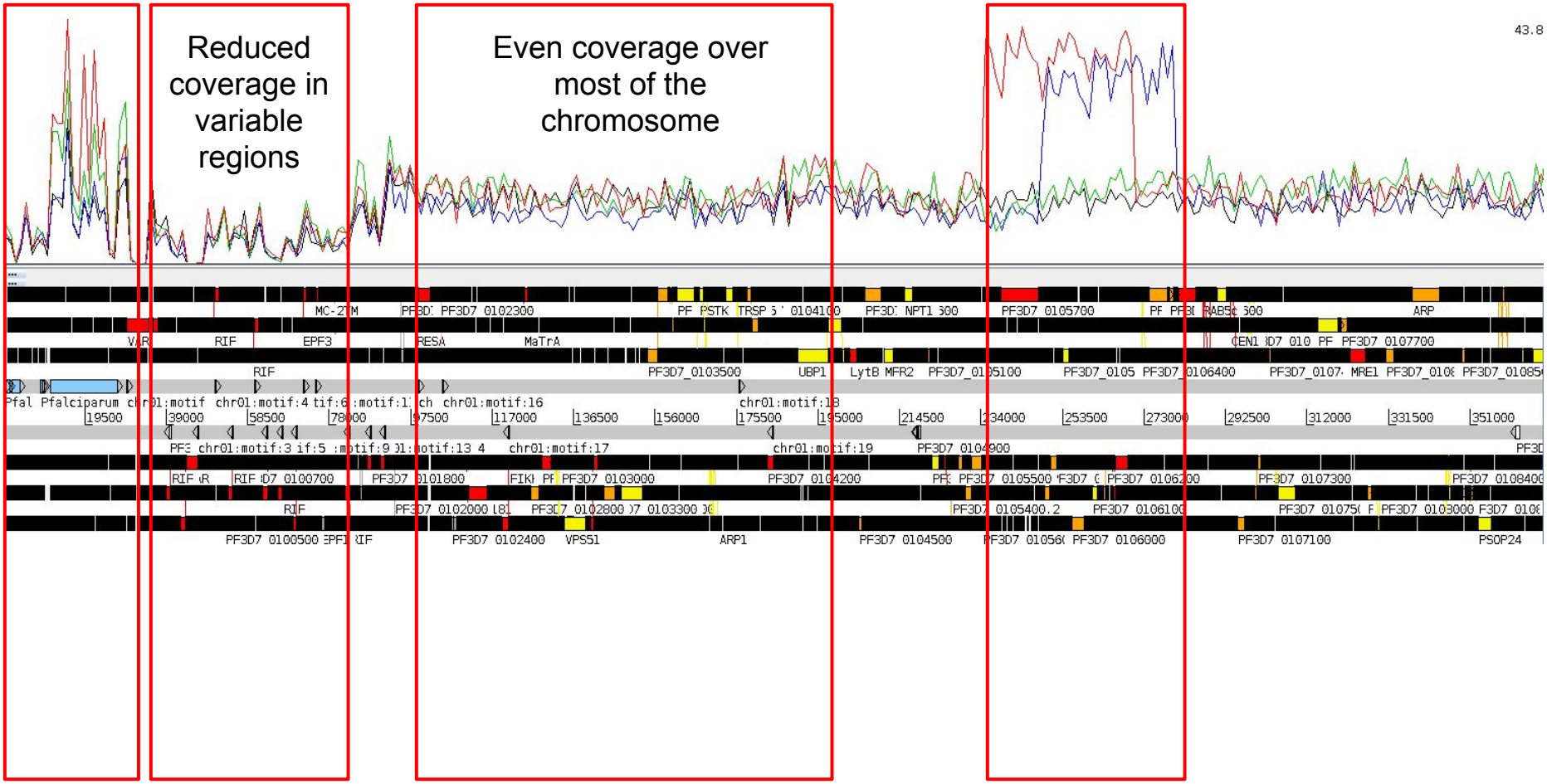
Global Coverage and CNVs

Variable coverage in repetitive regions

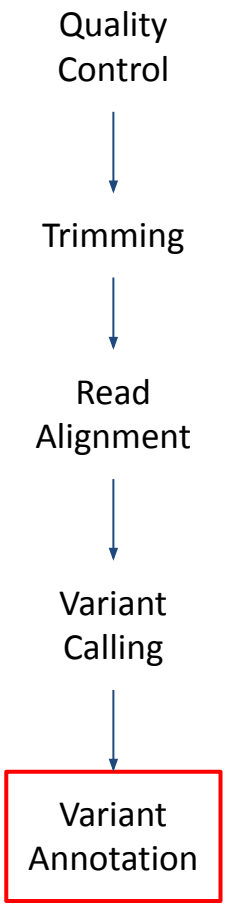
Reduced coverage in variable regions

Even coverage over most of the chromosome

Duplications in two samples



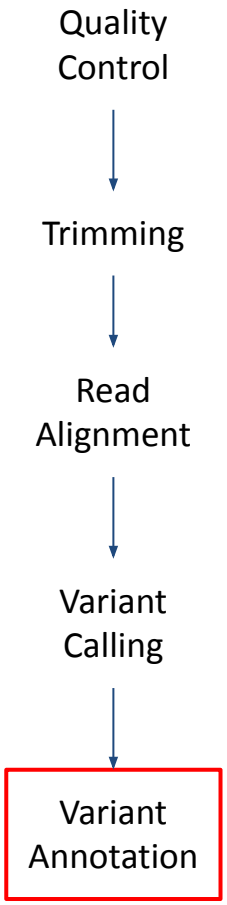
DNA Sequencing Analysis



Variant Annotation

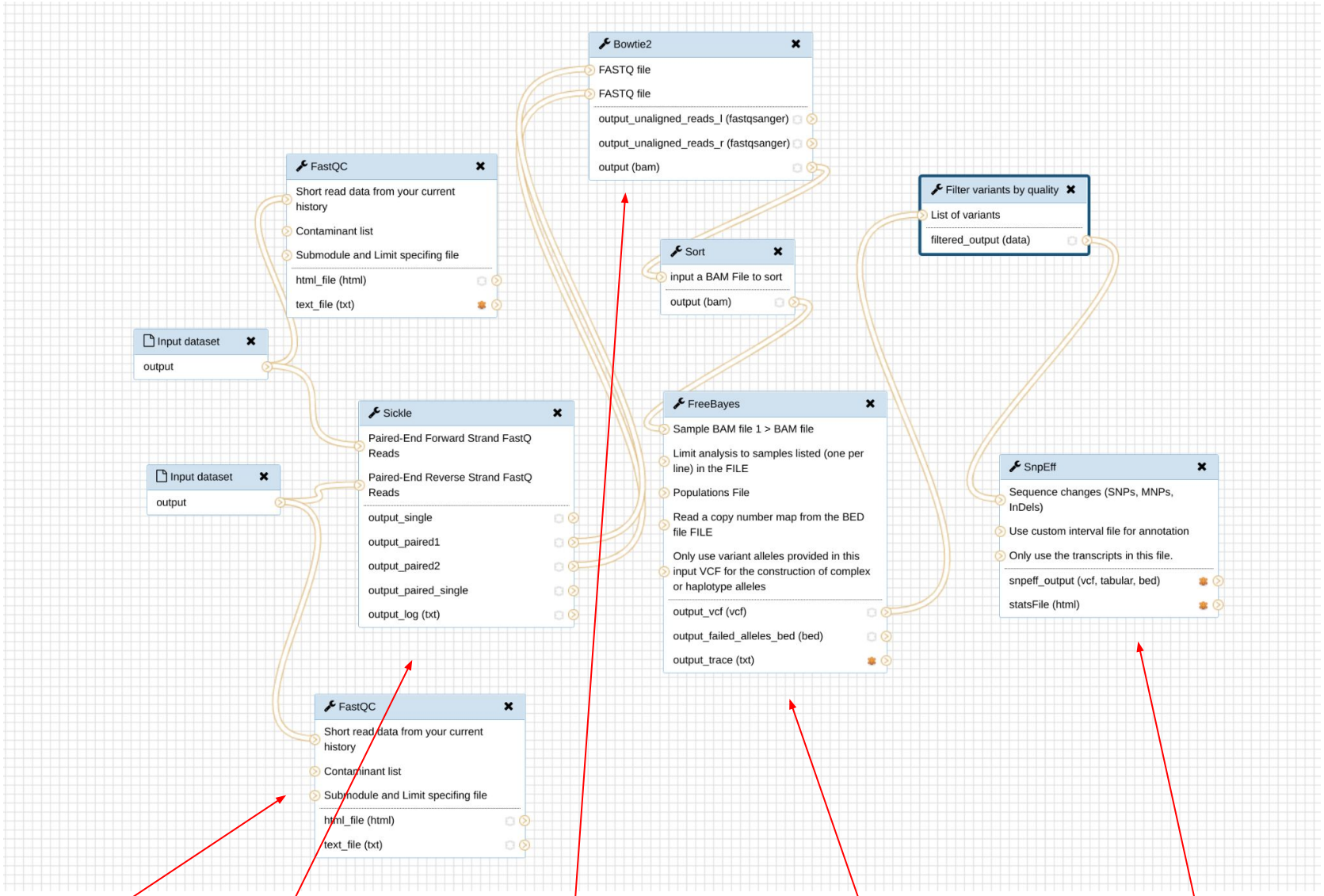
- Up to this point we have been solely concerned with sequence
- Variant annotation is concerned with determining the effects of variants
  - Coding or non-coding?
  - Synonymous or non-synonymous?
  - Missense or nonsense?
- Quality of variant annotation depends on the quality of the genome annotation
- Annotation can be *de novo* or we can use databases of known SNPs

DNA Sequencing Analysis



Variant Annotation

- Variant Effect Predictor: https://www.ensembl.org/info/docs/tools/vep/index.html
- SnpEff: http://pcingola.github.io/SnpEff/
- Annovar: https://annovar.openbioinformatics.org/en/latest/



QC → Trimming → Alignment → Variant Calling → Variant Annotation