University of Glasgow

NGS Analysis and Galaxy Part 2

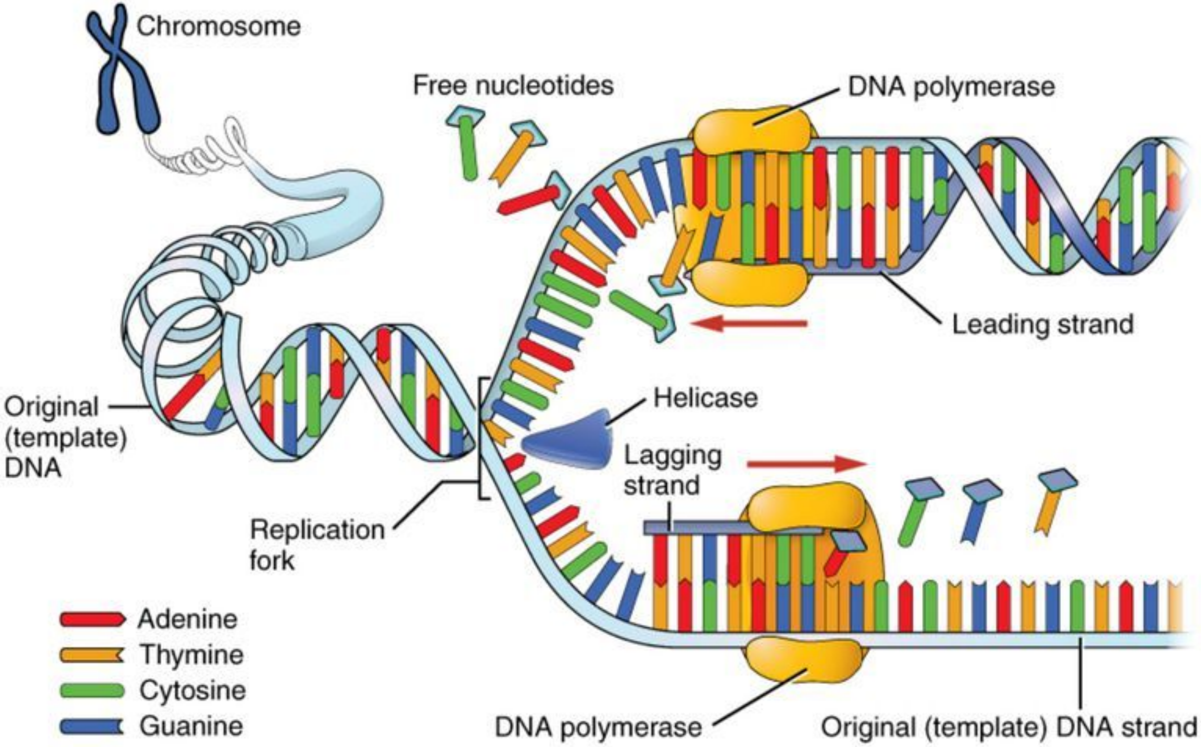
Kathryn Crouch
kathryn.crouch@glasgow.ac.uk

Outline

- Background
  - How does sequencing work?
  - What does the data look like?
- Pre-processing for Analysis
  - QC and trimming
  - Aligning to a reference genome
- Whole-genome sequencing
  - Calling SNPs
  - CNVs

How Does Sequencing Work?

- Carry out replication under controlled conditions
- Artificially slow down the reaction to see the order in which bases are incorporated



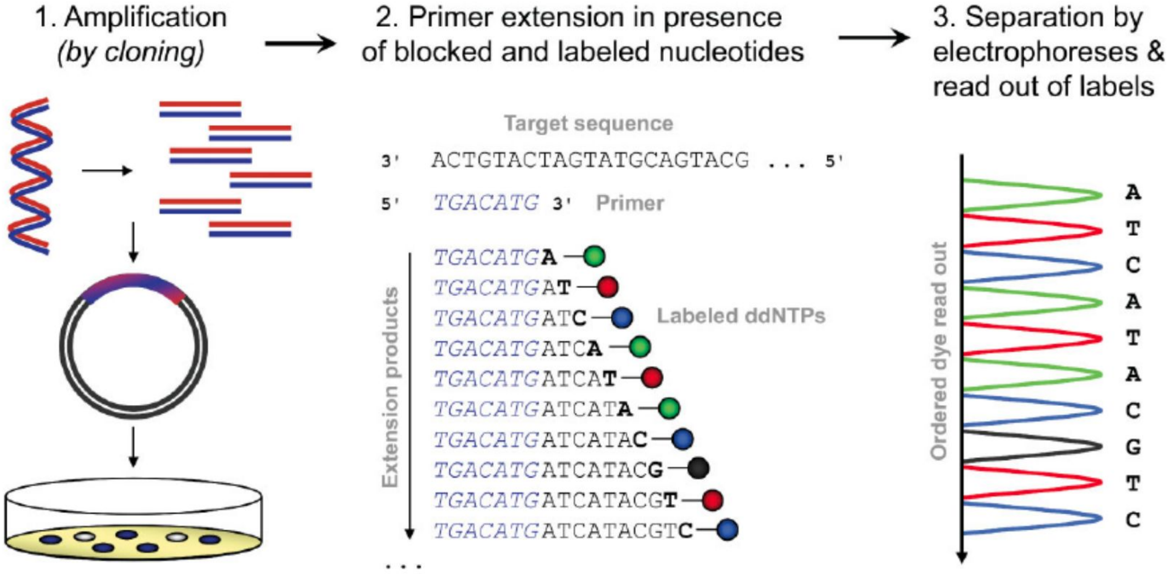
How Does Sequencing Work?

Sanger Sequencing (1975)

Uses modified nucleotides (ddNTPs) that cannot be extended

Each ddNTP is labelled with a different dye so you can see the order in which they are incorporated

Long reads and low error rate, but low throughput



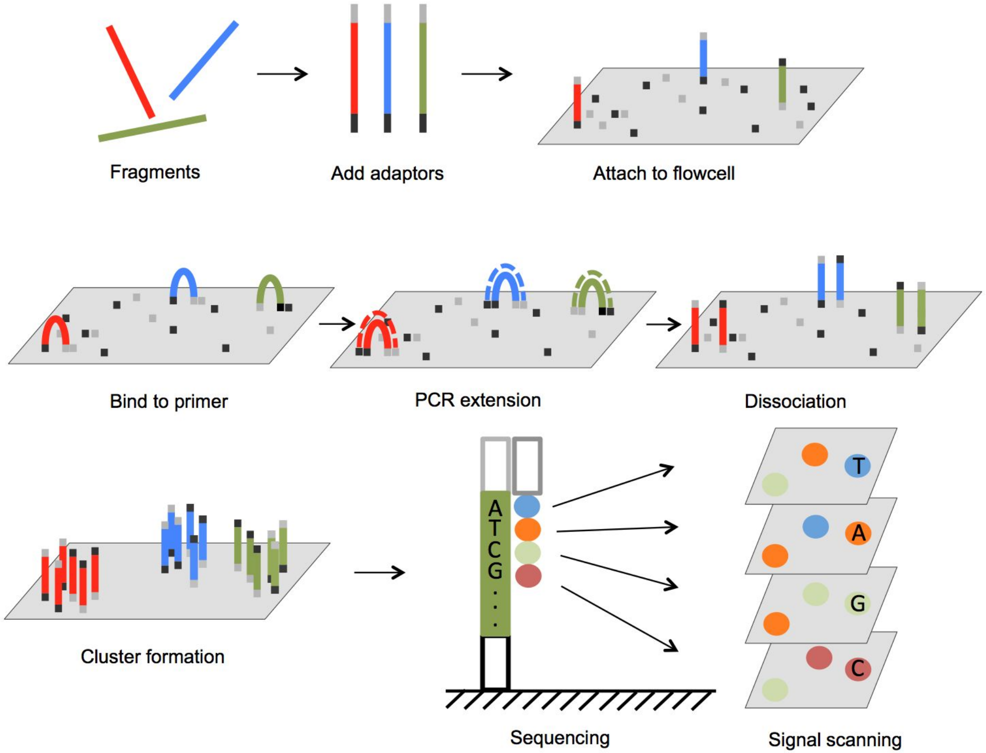
How Does Sequencing Work?

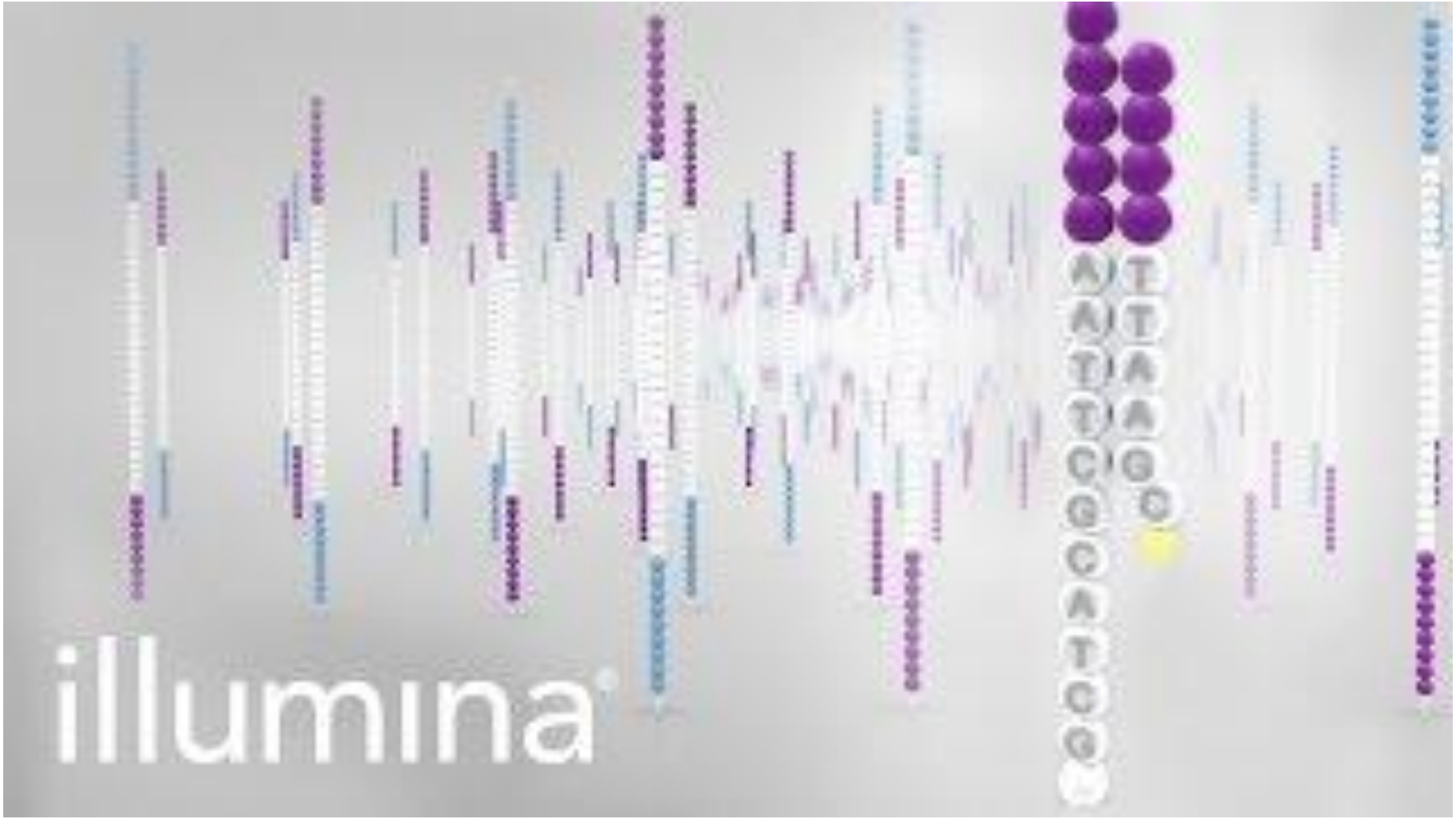
Illumina Sequencing (2005)

DNA fragments are adaptor-ligated and attached to a flow cell

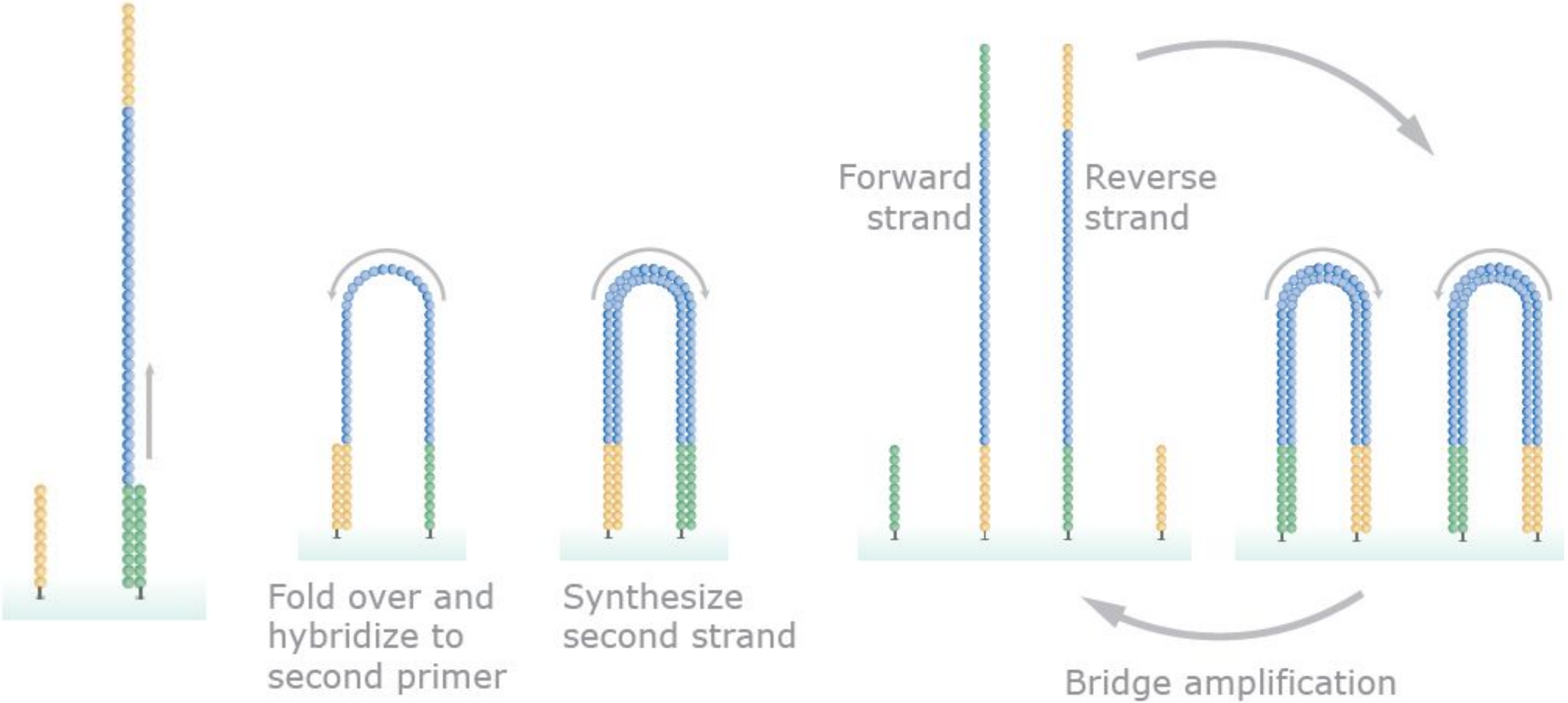
PCR is carried out in situ to form clusters

Sequencing can be carried out on millions of clusters simultaneously

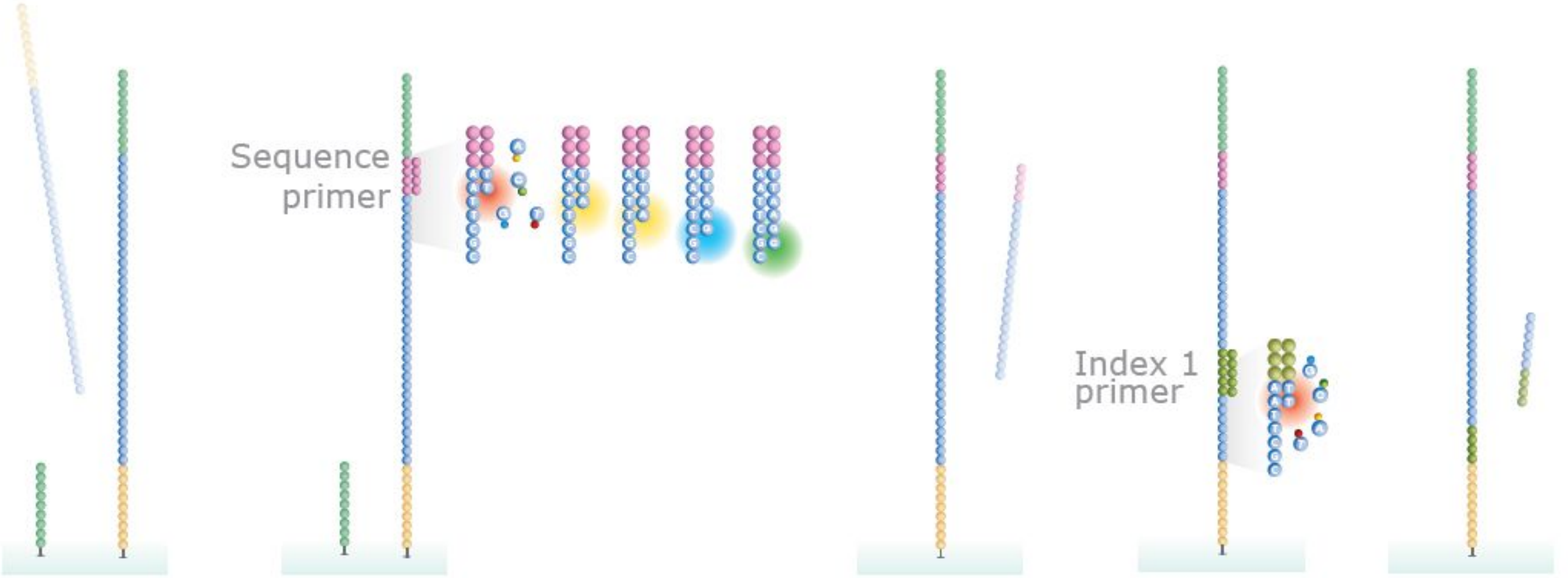



illumina
A A T T C G C A T C G
T T A A G C

How Does Sequencing Work?



How Does Sequencing Work?



The reverse strand is cleaved and washed away

With each cycle, four fluorescently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template.

The read product is washed away

Sequence Index1

The read product is washed away

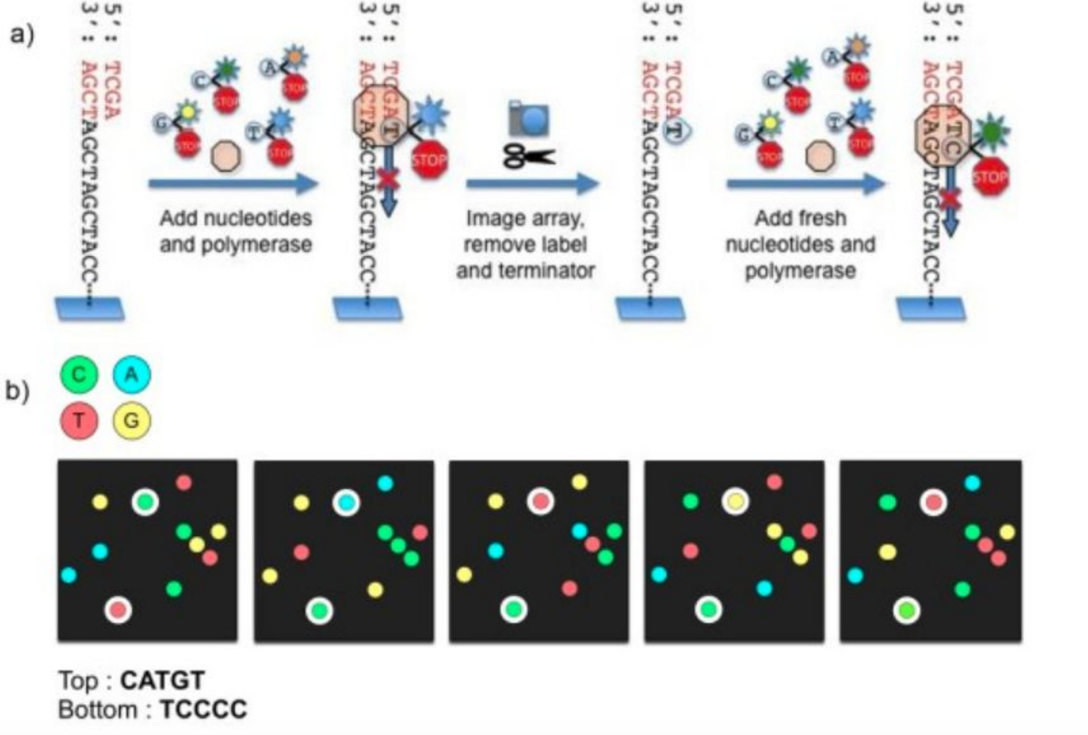
How Does Sequencing Work?

Illumina Sequencing

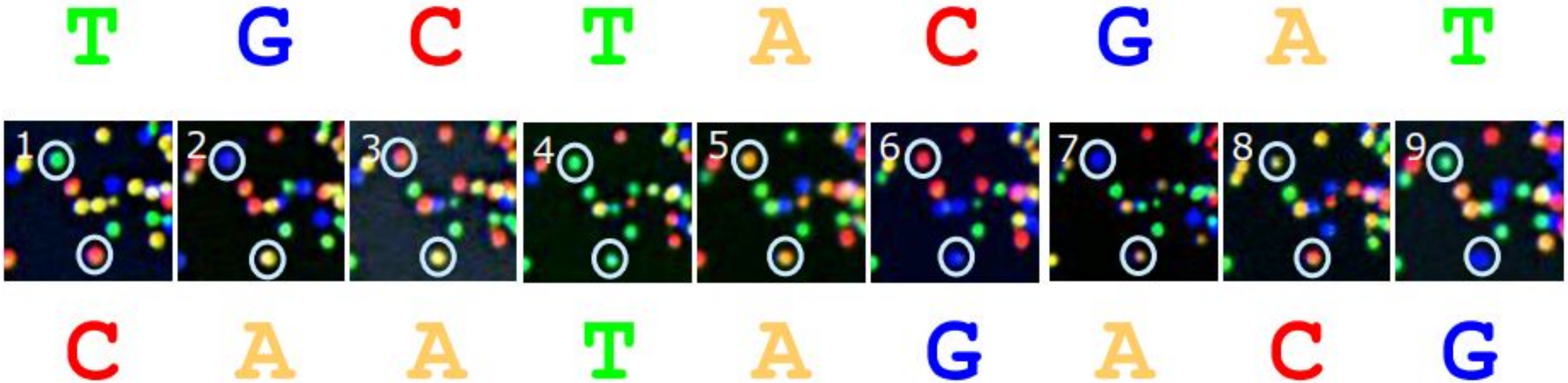
Blocked and labelled nucleotides are added

1 nucleotide is incorporated and an image is taken of the array

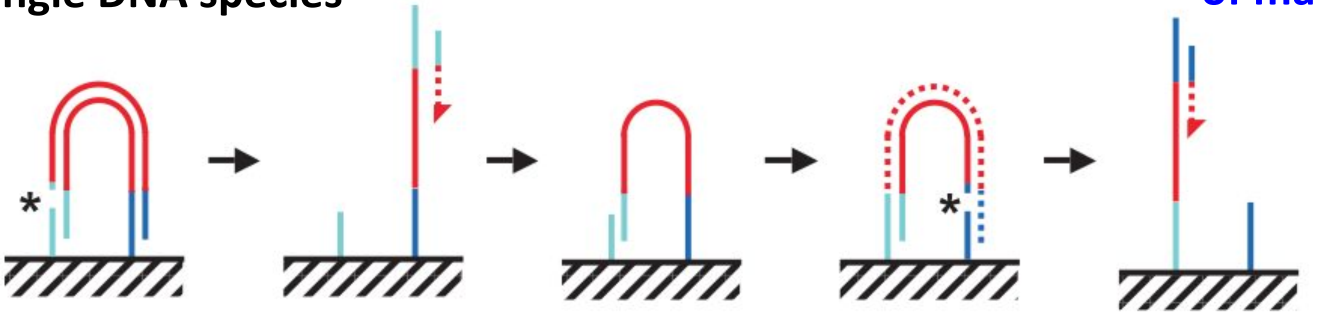
Label and block are removed and cycle repeats


a)
5' : TCGA
3' : AGCTAGCTAGCTACC
Add nucleotides and polymerase
Image array, remove label and terminator
Add fresh nucleotides and polymerase
STOP
b)
C
A
T
G
Top : CATGT
Bottom : TCCCC

How Does Sequencing Work?



How Does Sequencing Work?



Denaturing & washing ➔ single DNA species

Sequencing of mate 1

Sequencing of mate 2

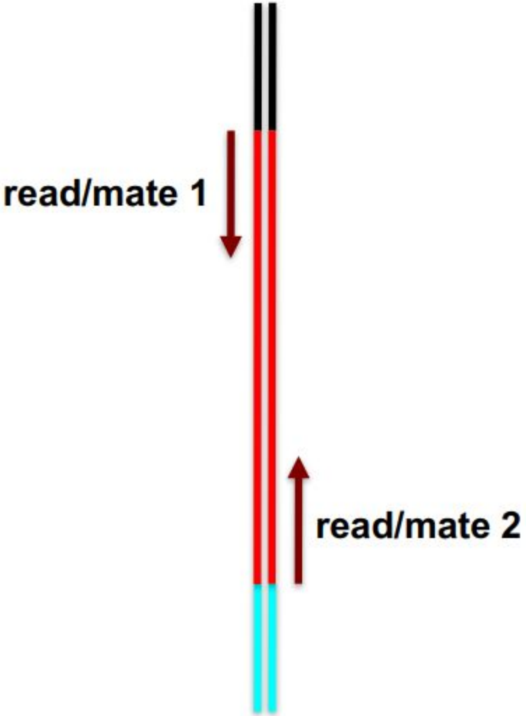
Adaptor-specific DNA-cleavage

Re-synthesis of complementary DNA strand & adaptor-specific DNA-cleavage

Paired End Illumina Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

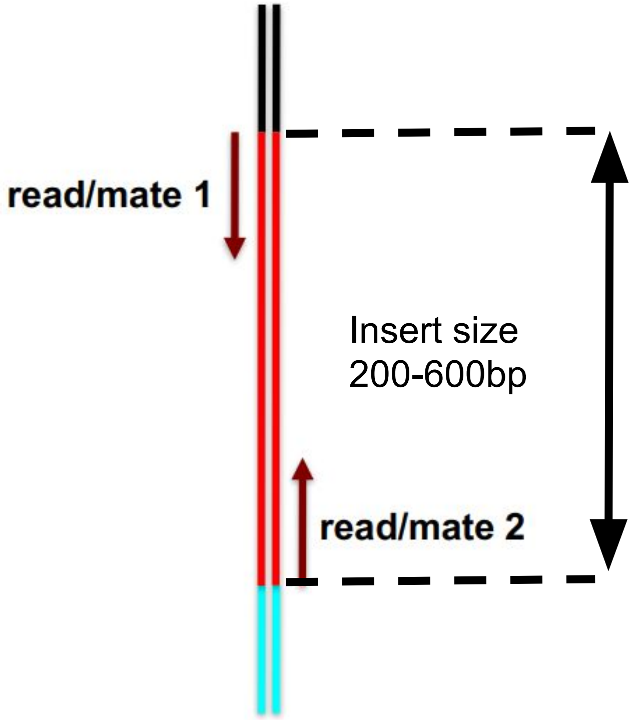


Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors



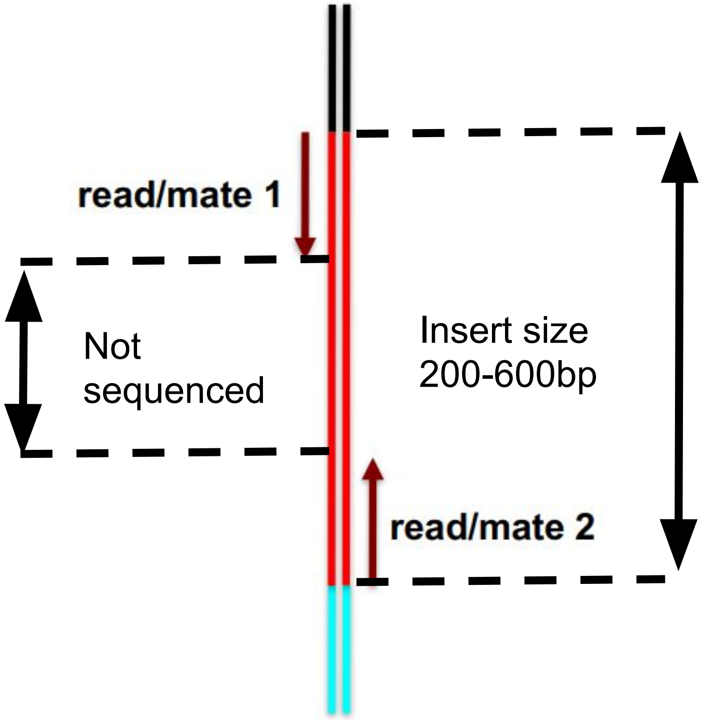
Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors

The region between the reads is not sequenced - it is covered by other fragments



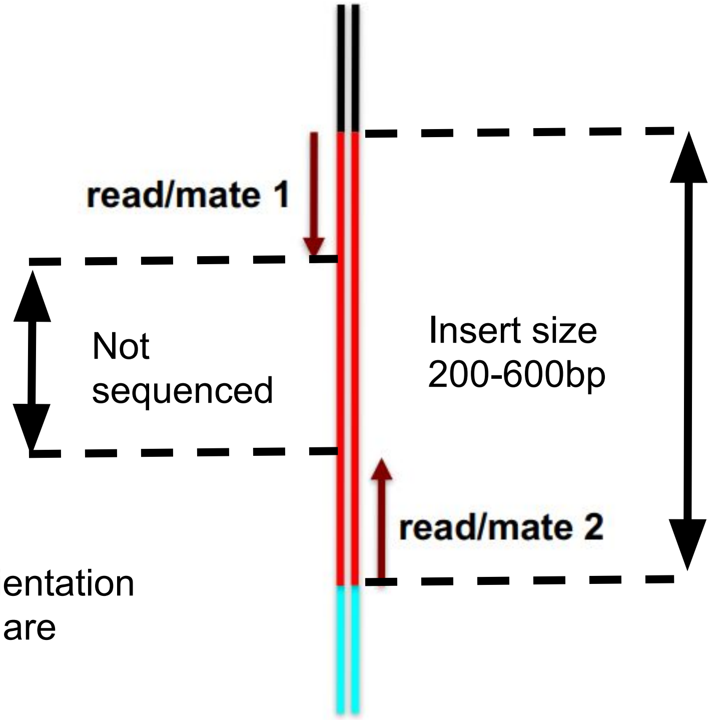
Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors

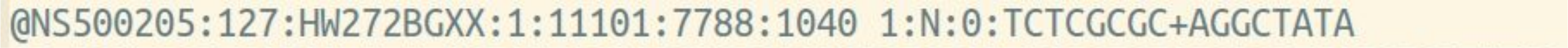
The region between the reads is not sequenced - it is covered by other fragments



Reads that map in the correct orientation and the expected distance apart are “concordant” or “proper pairs”

Concordant alignments are prioritised

File Formats: FASTQ



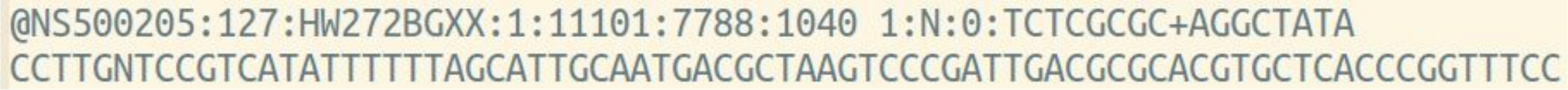
```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
```

For paired-end reads you will have two files

4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)

File Formats: FASTQ



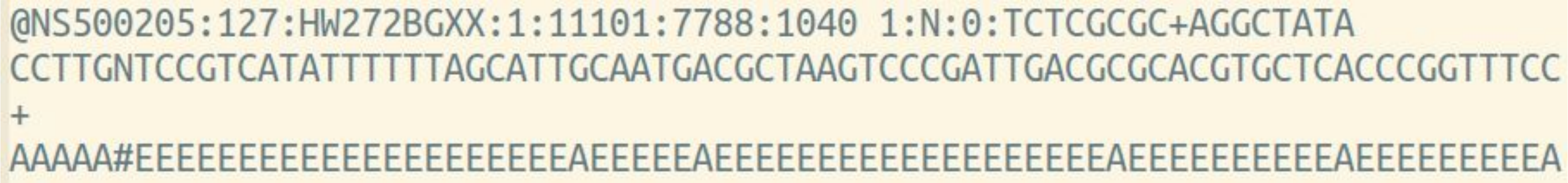
```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
CCTTGNTCCGTCATATTTTTTAGCATTGCAATGACGCTAAGTCCCGATTGACGCGCACGTGCTCACCCGGTTTCC
```

For paired-end reads you will have two files

4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read

File Formats: FASTQ

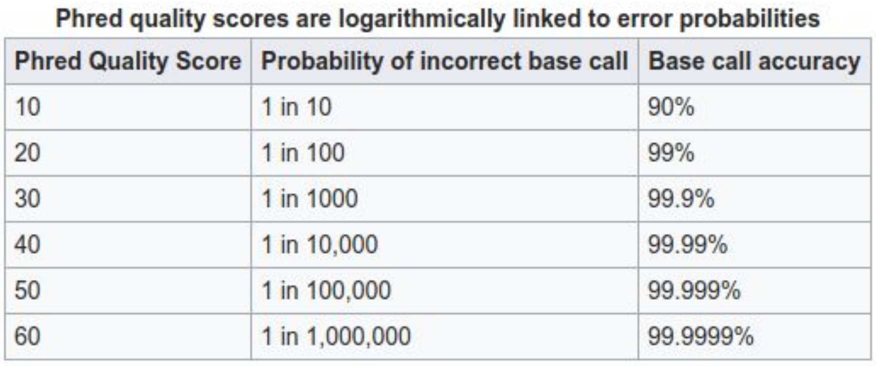


```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
CCTTGNTCCGTCATATTTTTTAGCATTGCAATGACGCTAAGTCCCGATTGACGCGCACGTGCTCACCCGGTTTCC
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEAEEEEEAEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEAEEEEEEEEEA
```

For paired-end reads you will have two files

4 lines per read

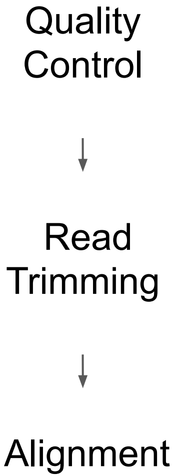
- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read
- Line 4 is the quality for each base
  - Quality is encoded using ASCII
  - http://www.asciitable.com/



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Alignment-based Analysis

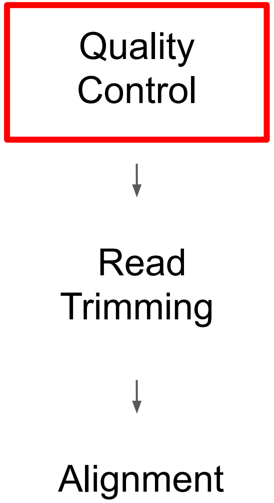


Quality Control

Read Trimming

Alignment

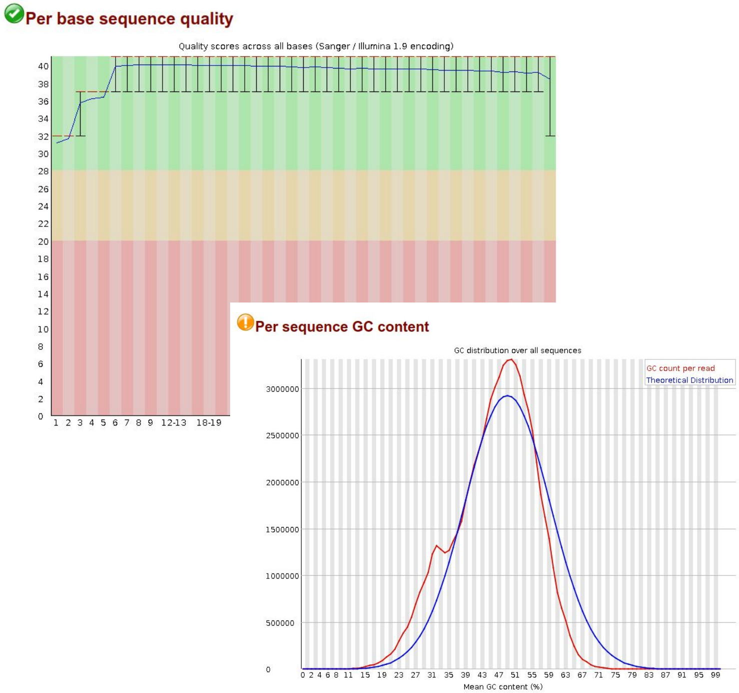
Analysis of NGS Sequencing

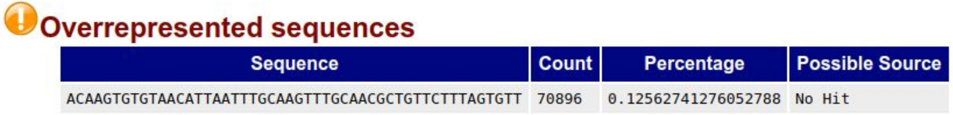

Quality Control

Read Trimming

Alignment

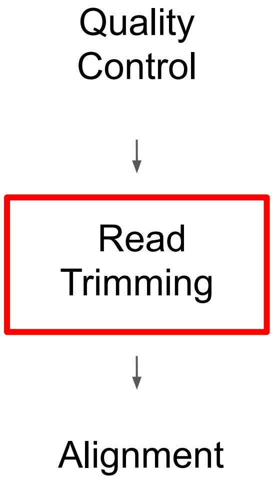
- FASTQC https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
    - Overall sequencing quality
    - GC content
    - N content
    - Read length distribution
    - Over-represented sequences
    - Adaptor content
- Output is an html file that can be opened in a web browser



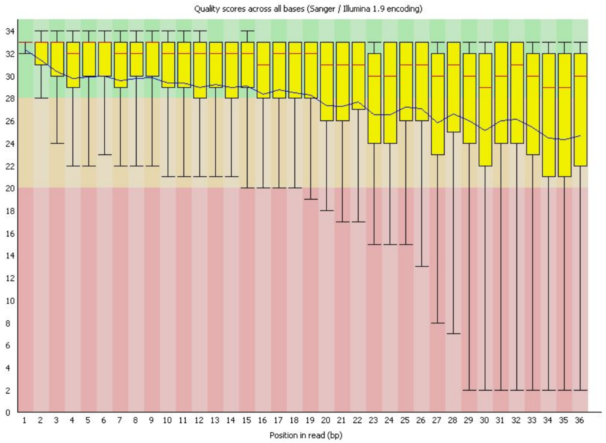

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ACAAGTGTGTAACATTAATTTGCAAGTTTGCAACGCTGTTCTTTAGTGTT	70896	0.12562741276052788	No Hit

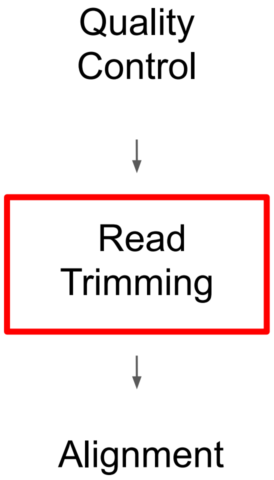
Analysis of NGS Sequencing



- Trimmomatic https://github.com/usadellab/Trimmomatic
- 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



Analysis of NGS Sequencing

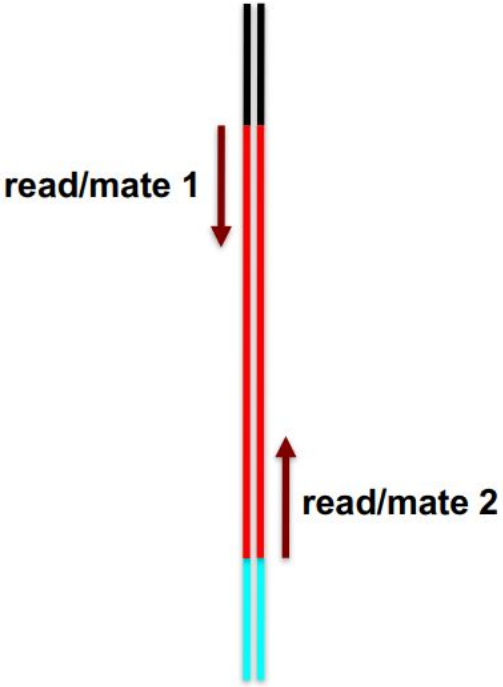


Quality Control

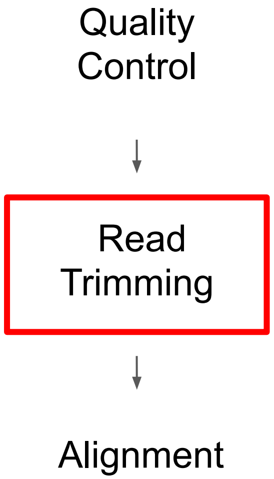
Read Trimming

Alignment

- Trimmomatic https://github.com/usadellab/Trimmomatic
- 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



Analysis of NGS Sequencing



Quality Control

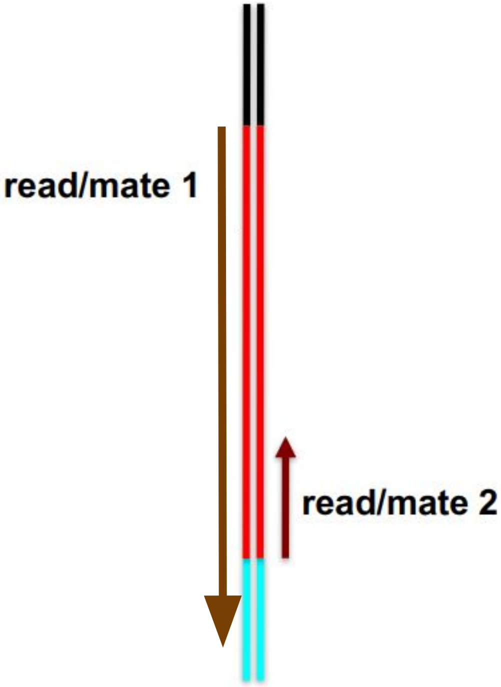
↓

Read Trimming

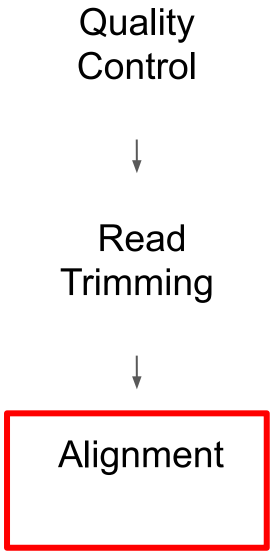
↓

Alignment

- Trimmomatic https://github.com/usadellab/Trimmomatic
- 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



Analysis of NGS Sequencing



Quality Control

↓

Read Trimming

↓

Alignment

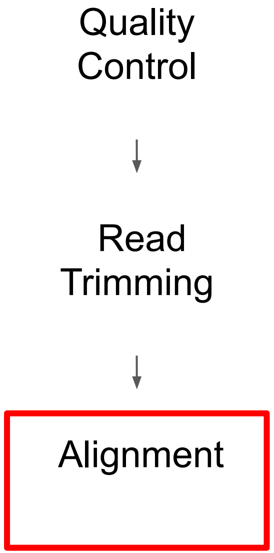
What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

Analysis of NGS Sequencing



What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

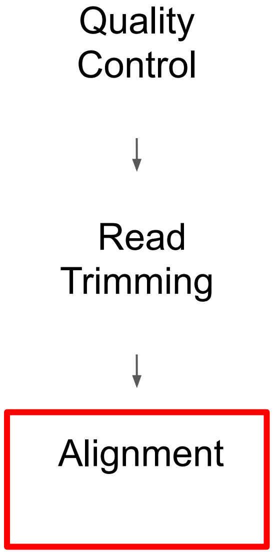
How would you align one to the other?

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

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

Which one is better??

Analysis of NGS Sequencing



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

Alignment algorithms are designed to align data in a reasonable time on a standard computer

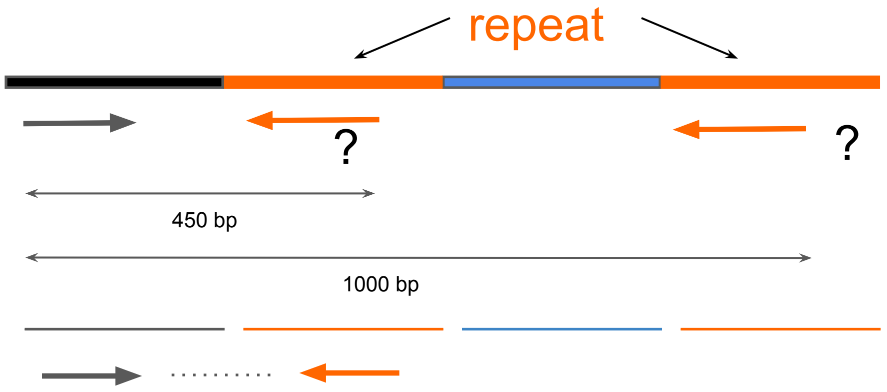
Heuristic - not exhaustive, but “good enough”

Improving Alignments with Paired End Reads

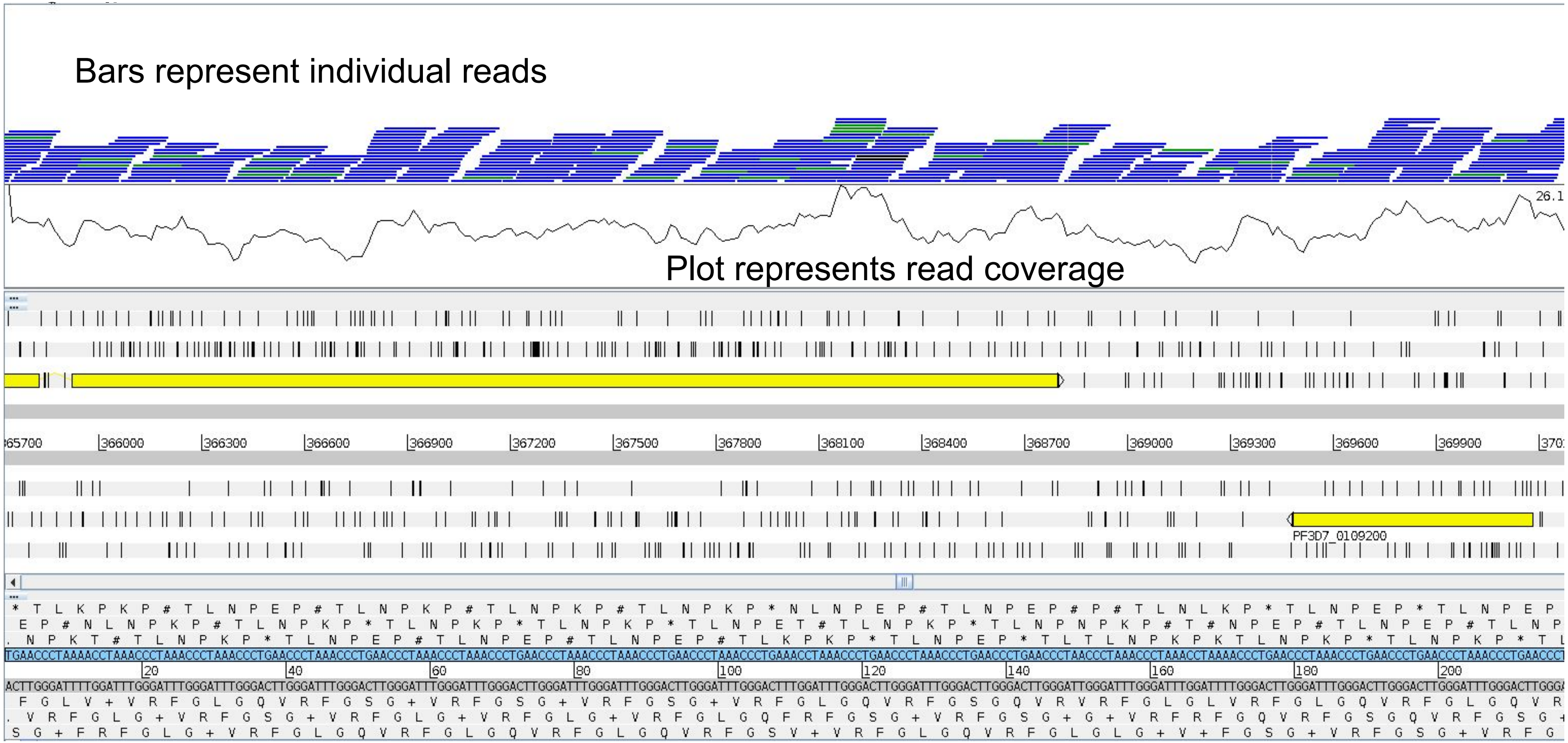
Paired end reads can resolve alignments in repetitive regions

A read that could map in multiple locations due to repetitions in the genome can be located accurately by inference from the position of its pair

In this case, we know that the insert size is ~400 bp, so we can infer that the first alignment is more likely to be correct.



Visualising Alignments (Artemis)

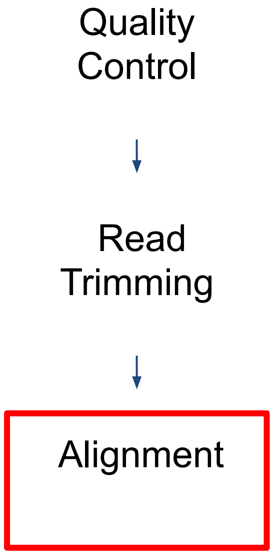


Bars represent individual reads

Plot represents read coverage

```
 * T L K P K P # T L N P E P # T L N P K P # T L N P K P # T L N P K P * N L N P E P # T L N P E P # P # T L N L K P * T L N P E P * T L N P E P
  E P # N L N P K P # T L N P K P * T L N P K P * T L N P K P * T L N P E T # T L N P K P * T L N P N P K P # T # N P E P # T L N P E P # T L N P
. N P K T # T L N P K P * T L N P E P # T L N P E P # T L N P E P # T L K P K P * T L N P E P * T L T L N P K P K T L N P K P * T L N P K P * T L
TGAACCCTAAAACCTAAACCCTAAACCCTAAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAACCCTAAACCCTGAAACCTAAACCCTGAACCCTAAACCCTGAACCCTGAACCCTAACCCTAAACCCTAAACCTAAAACCCTGAACCCTAAACCCTGAACCCTGAACCCTAAACCCTGAACCCT
                    20                  40                  60                  80                  100                 120                 140                 160                 180                 200
ACTTGGGATTTTGGATTTGGGATTTGGGATTTGGGACTTGGGATTTGGGACTTGGGATTTGGGATTTGGGACTTGGGATTTGGGACTTGGGATTTGGGACTTTGGATTTGGGACTTGGGATTTGGGACTTGGGACTTGGGATTGGGATTTGGGATTTGGATTTTGGGACTTGGGATTTGGGACTTGGGACTTGGGATTTGGGACTTGGGA
  F G L V + V R F G L G Q V R F G S G + V R F G S G + V R F G S G + V R F G L G Q V R F G S G Q V R V R F G L G L V R F G L G Q V R F G L G Q V R
. V R F G L G + V R F G S G + V R F G L G + V R F G L G + V R F G L G Q F R F G S G + V R F G S G + G + V R F R F G Q V R F G S G Q V R F G S G +
 S G + F R F G L G + V R F G L G Q V R F G L G Q V R F G L G Q V R F G S V + V R F G L G Q V R F G L G L G + V + F G S G + V R F G S G + V R F G
```

Alignment Tools



Quality Control

Read Trimming

Alignment

Mapping Tools for DNA-seq data

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

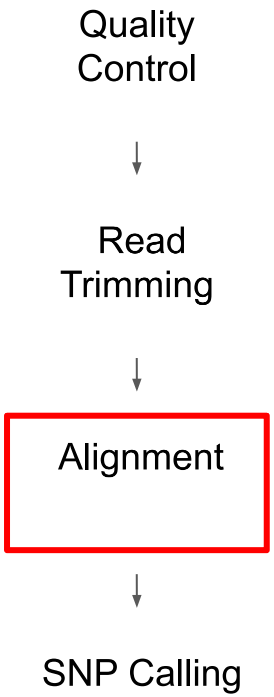
Whole Genome Sequencing (DNA-seq, WGS)

What Can We Discover From Aligned Reads?

- Where and how is our sample different from the reference?
  - Discovery of SNVs and Indels

- Coverage
  - Discovery of copy number variations
  - Discovery of regions of high variability

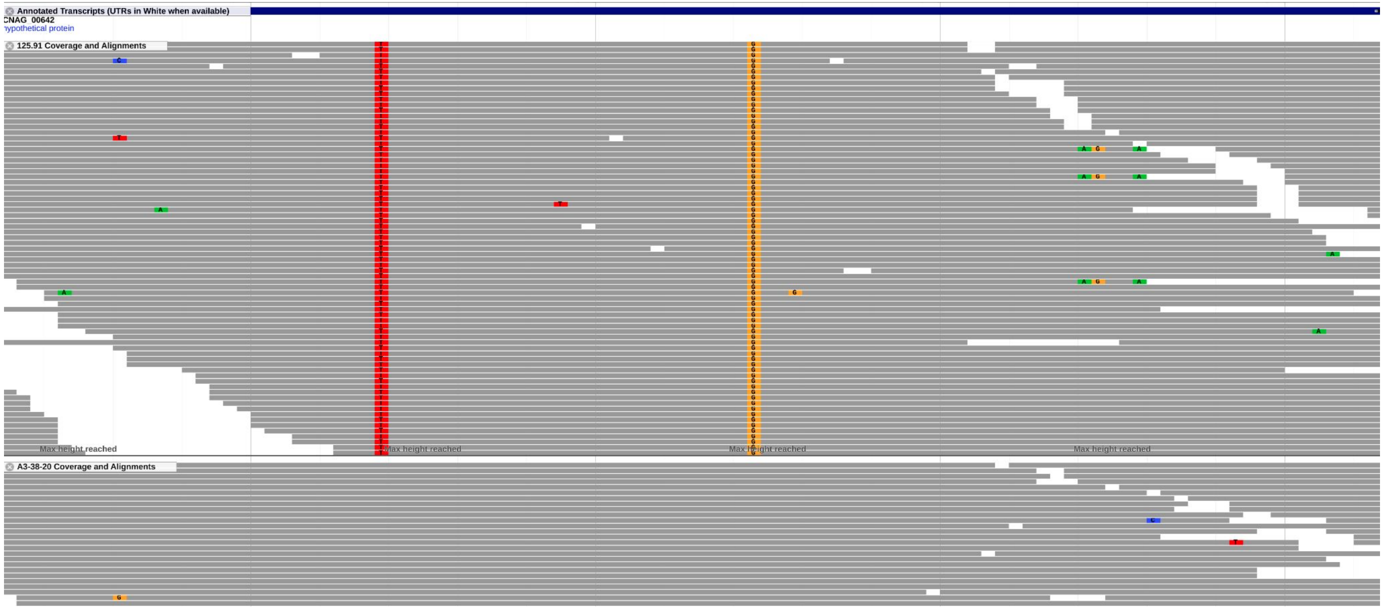
Finding SNVs


Quality Control

Read Trimming

Alignment

SNP Calling


Annotated Transcripts (UTRs in White when available)
CNAG_00642
ypothetical protein
125.91 Coverage and Alignments
Max height reached
Max height reached
Max height reached
Max height reached
A3-38-20 Coverage and Alignments

Finding Variants
If we load alignments into a genome viewer, we can see variants

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

Analysis of NGS Sequencing

Quality Control

↓

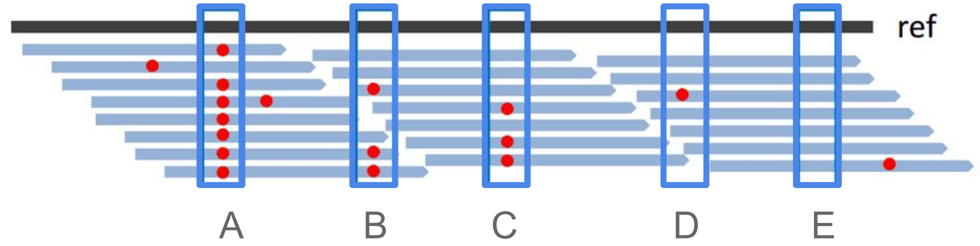
Read Trimming

↓

Alignment

↓

SNP Calling



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

Analysis of NGS Sequencing

Quality Control

↓

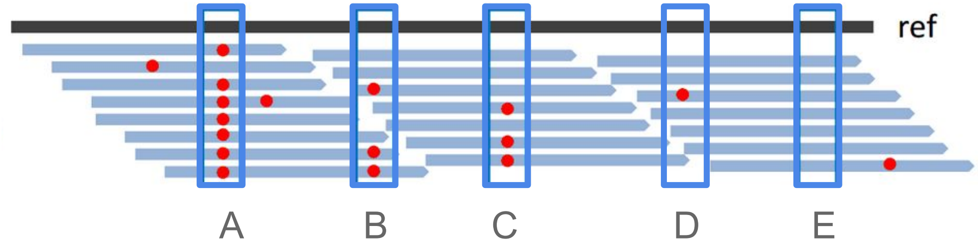
Read Trimming

↓

Alignment

↓

SNP Calling



Things to think about:

- What happens if your sample is not a clone?
- What happens if your sequencing depth is low?

Analysis of NGS Sequencing

Quality Control

↓

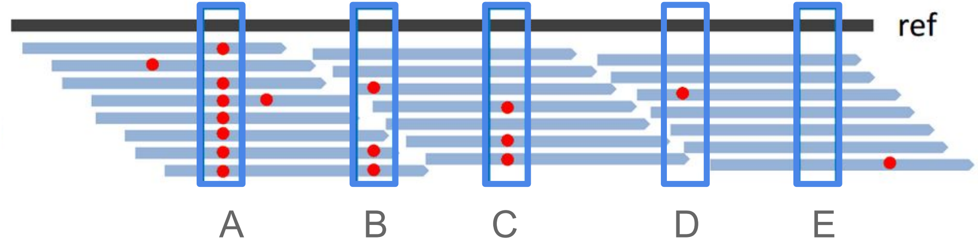
Read Trimming

↓

Alignment

↓

SNP Calling

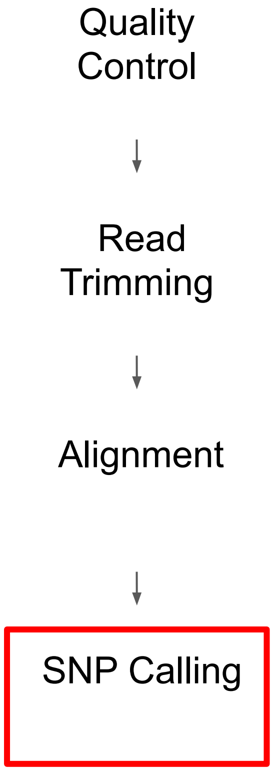


Freebayes https://github.com/freebayes/freebayes

Automated tool to call SNPs

You may also come across other tools including GATK and BCFTools.

What Else Can We Find Out?



Quality Control

Read Trimming

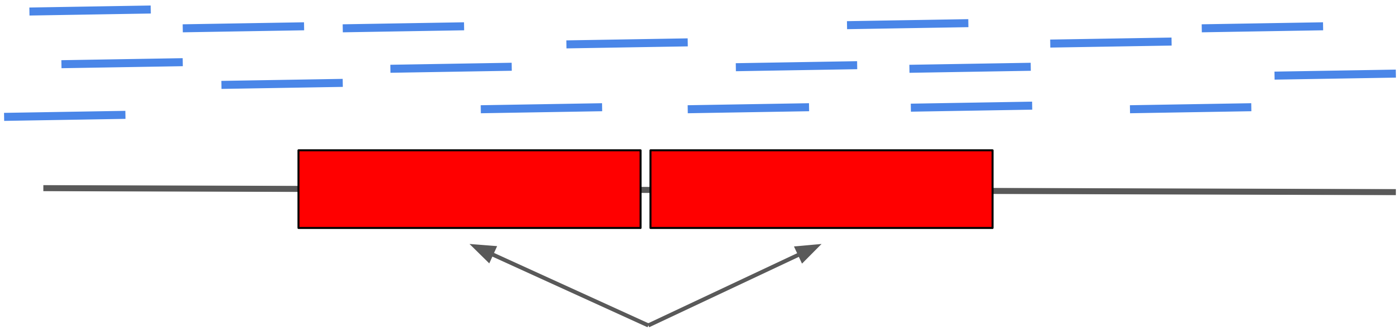
Alignment

SNP Calling

Coverage

- Expect coverage to be even across the genome
- In reality, we see variation associated with:
    - GC content
    - Repetitive or highly variable regions
    - Large scale insertions and deletions
- Note that doing alignments and examining coverage is the basis of RNAseq and ChIPseq analysis too!

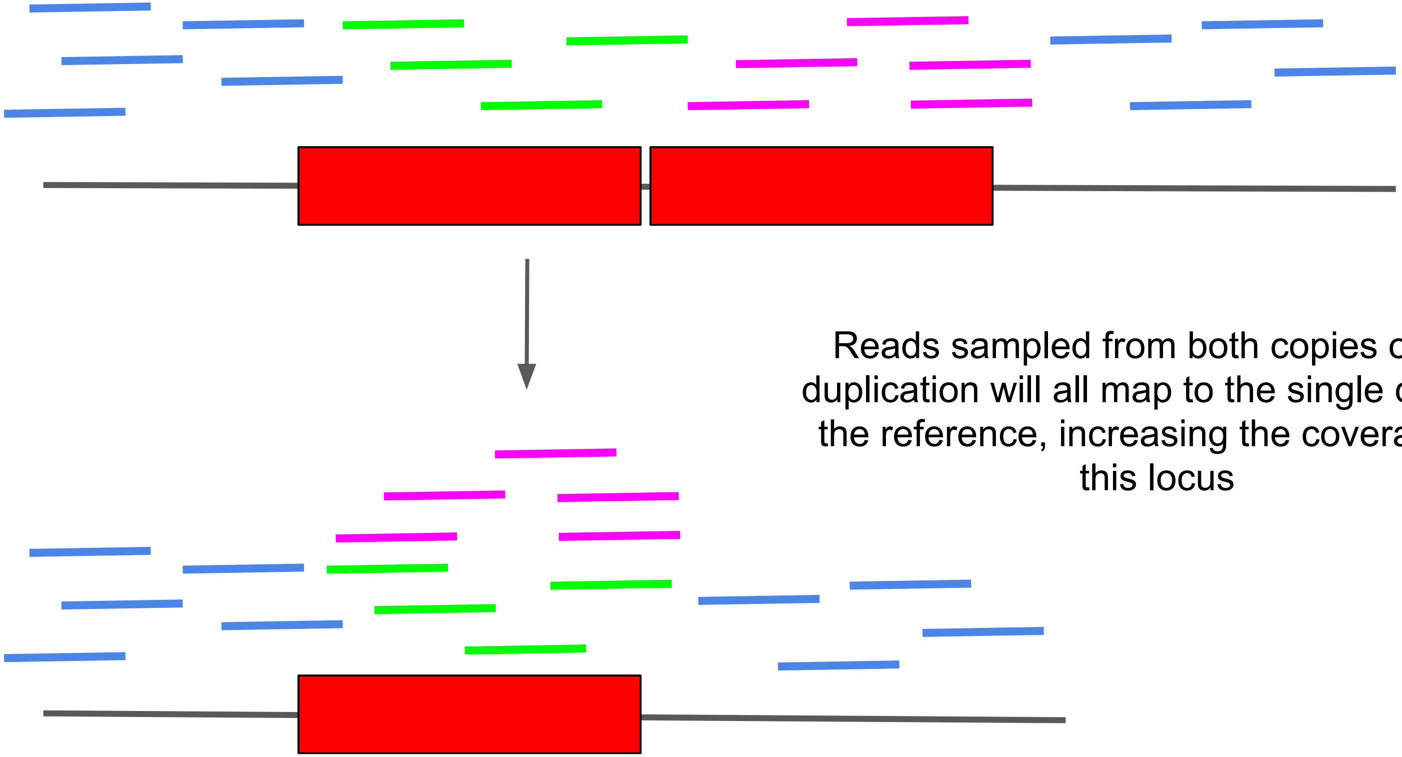
Coverage and Copy Number Variations



Reads sampled evenly

Duplicated region of the genome

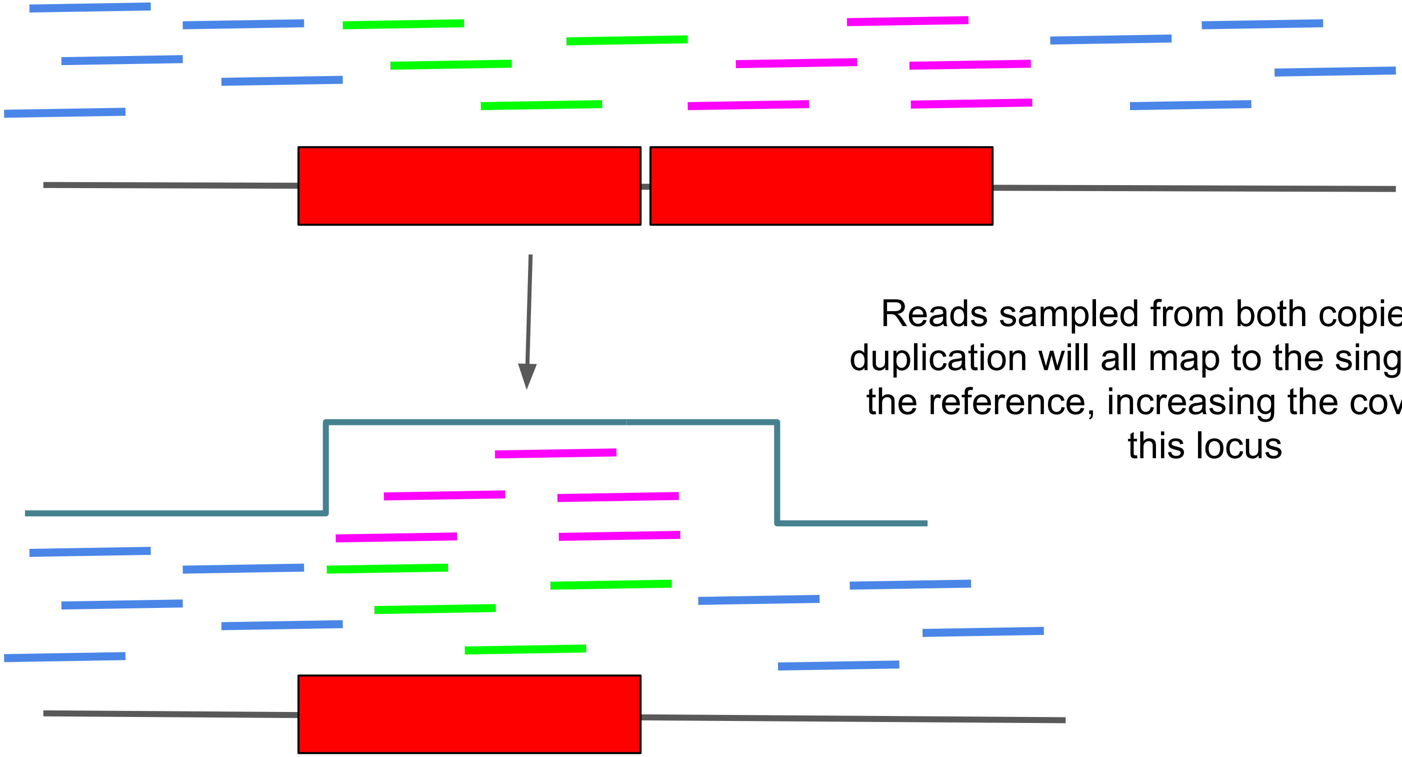
Copy Number Variations



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

Copy Number Variations



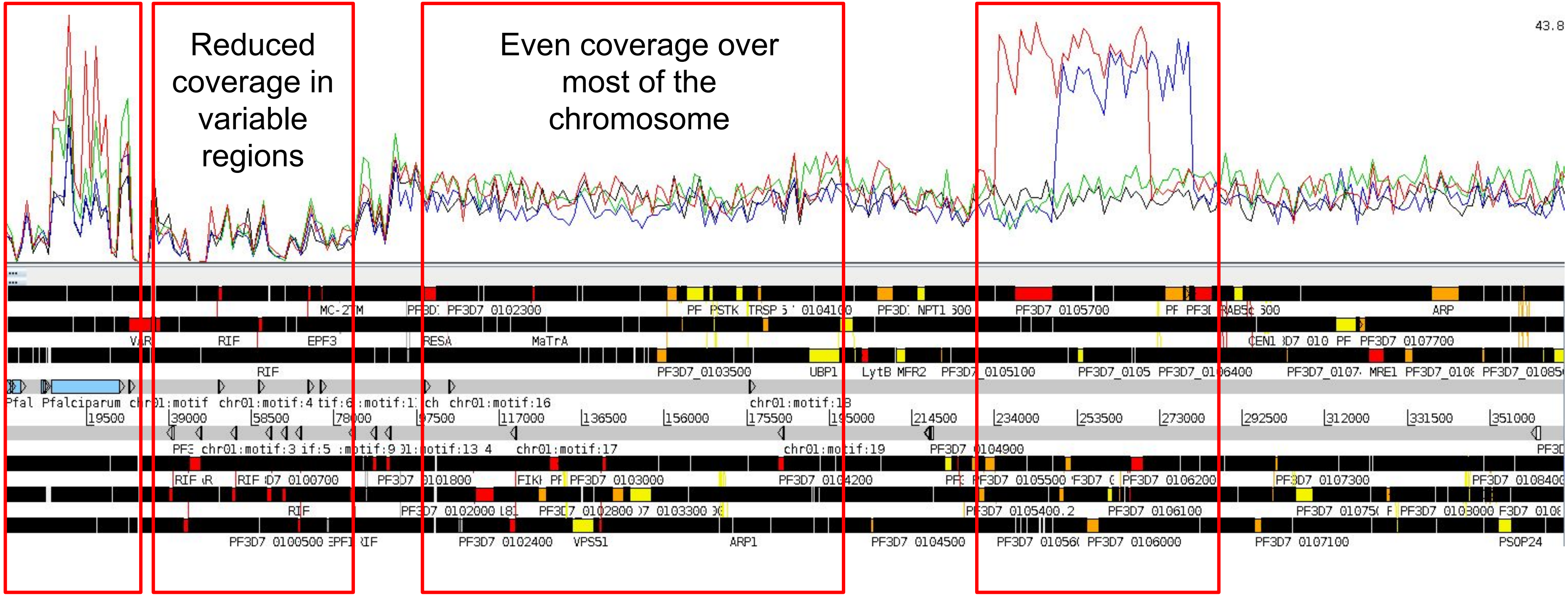
Reads sampled from both copies of the duplication will all map to the single copy in the reference, increasing the coverage at this locus

Reference genome without duplication

Global Coverage

Variable coverage in repetitive regions

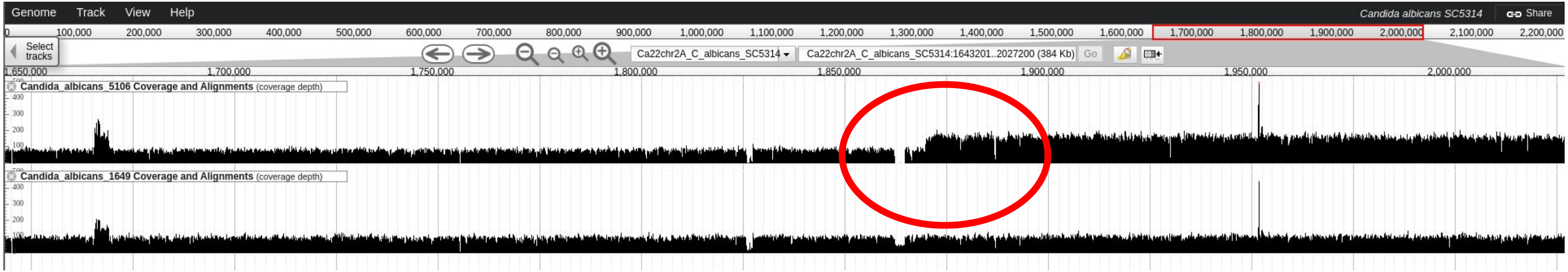
CNV (triplication) in two samples



Reduced coverage in variable regions

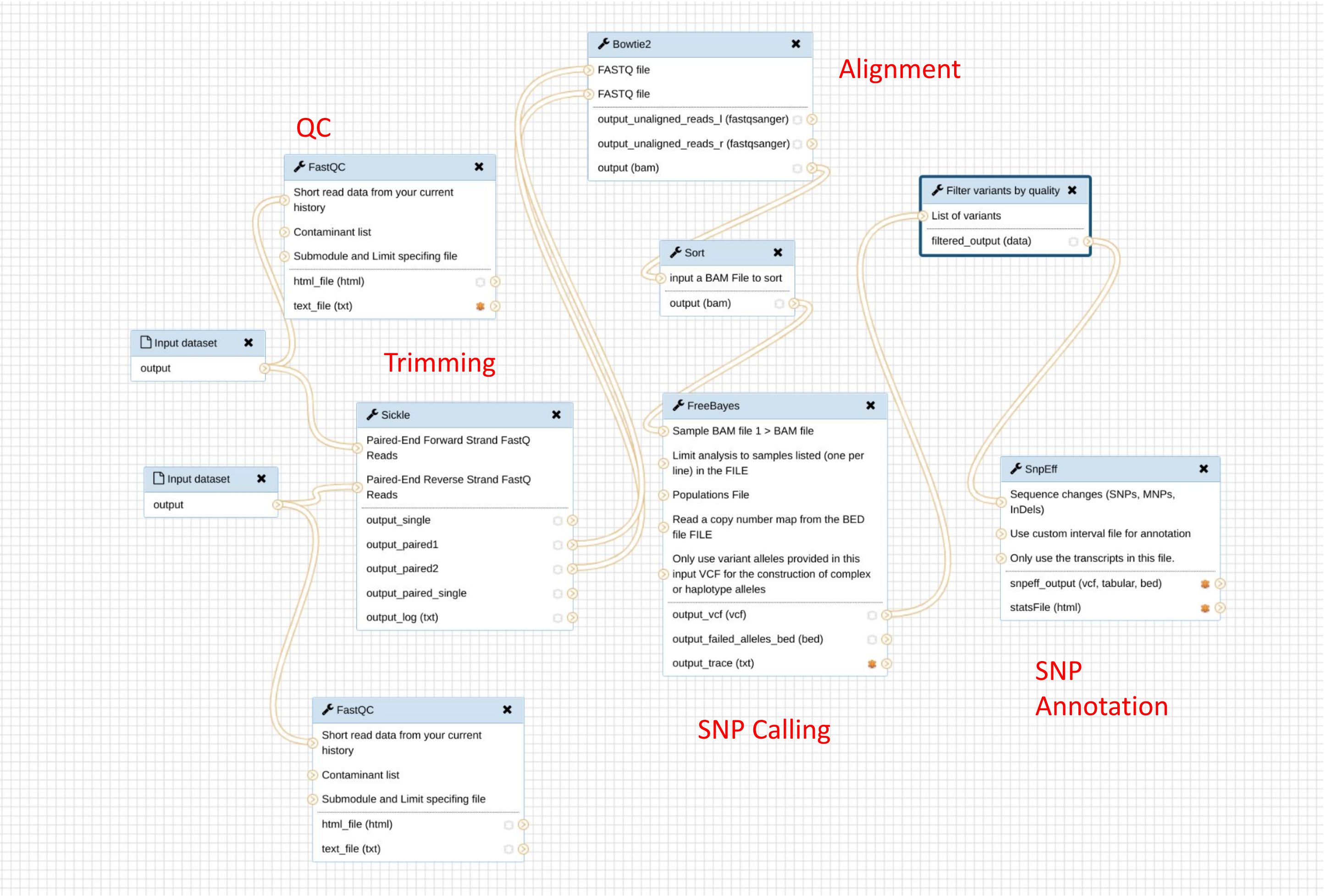
Even coverage over most of the chromosome

Segmental Duplication in *C. albicans*



Segmental tetraploidy on the right arm of chromosome 2 in a clinical sample

Galaxy


QC
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Input dataset
output
Trimming
Sickle
Paired-End Forward Strand FastQ Reads
Paired-End Reverse Strand FastQ Reads
output_single
output_paired1
output_paired2
output_paired_single
output_log (txt)
Input dataset
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Bowtie2
FASTQ file
FASTQ file
output_unaligned_reads_l (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
Alignment
Sort
input a BAM File to sort
output (bam)
FreeBayes
Sample BAM file 1 > BAM file
Limit analysis to samples listed (one per line) in the FILE
Populations File
Read a copy number map from the BED file FILE
Only use variant alleles provided in this input VCF for the construction of complex or haplotype alleles
output_vcf (vcf)
output_failed_alleles_bed (bed)
output_trace (txt)
SNP Calling
Filter variants by quality
List of variants
filtered_output (data)
SnpEff
Sequence changes (SNPs, MNPs, InDels)
Use custom interval file for annotation
Only use the transcripts in this file.
snpeff_output (vcf, tabular, bed)
statsFile (html)
SNP Annotation

Accessing Data

- Workflows failed because we ran of disc space - you didn't do anything wrong!
- We will explore the output of workflows that Eve pre-ran
- We will do this in the live Galaxy site NOT the workshop Galaxy site!