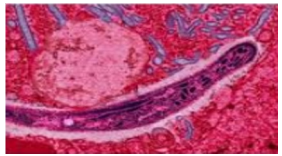
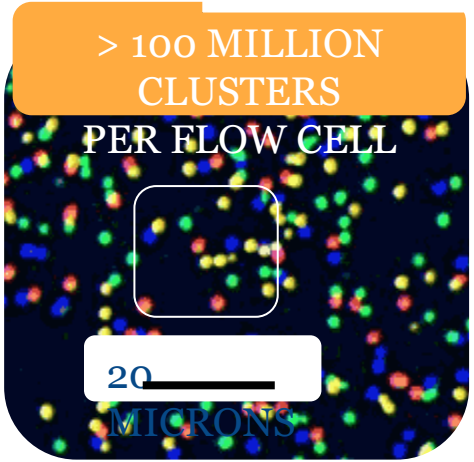


		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

3	6	9	7	10	13
G	T	T	-	A	C
G	T	T	G	A	C



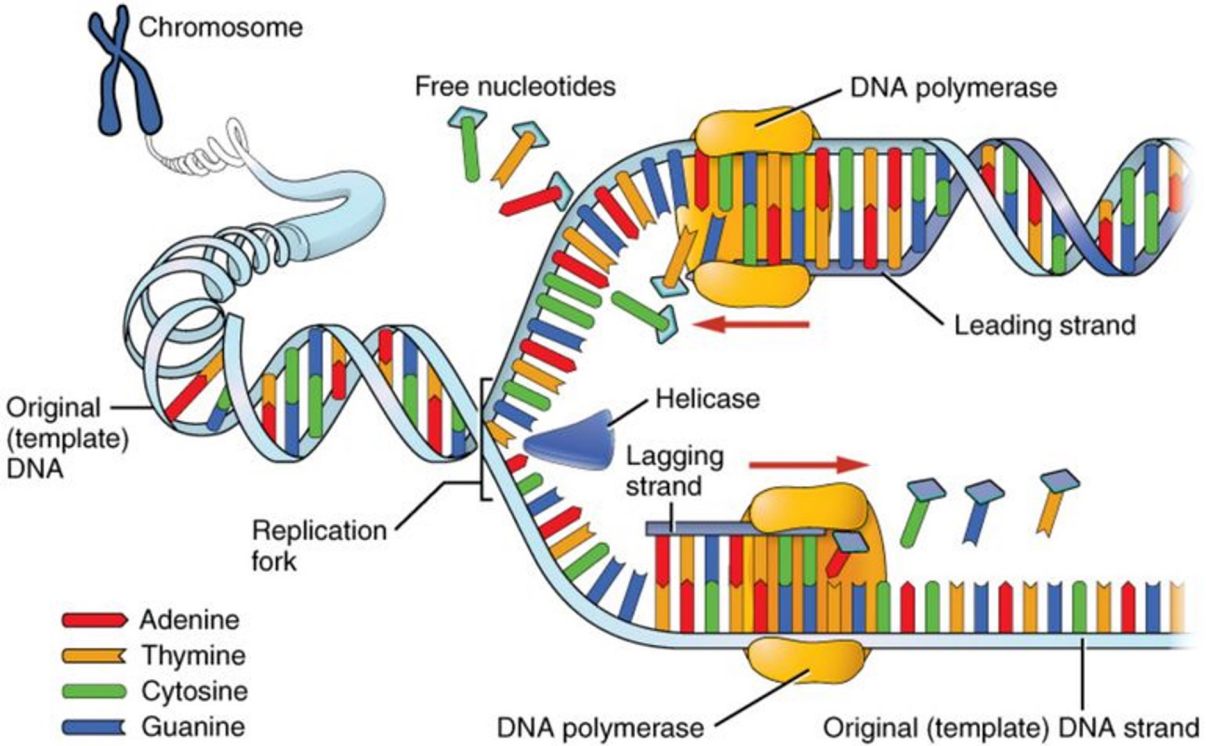
Mapping and variant calling

Outline

- Background
  - How does sequencing work?
  - Why do we do it?
- Analysis - Assembly and Mapping
  - What does the pipeline look like?
  - How do the tools work?
- What can we learn from sequencing
  - Sequence variations
  - Copy number variation

How Does Sequencing Work?

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



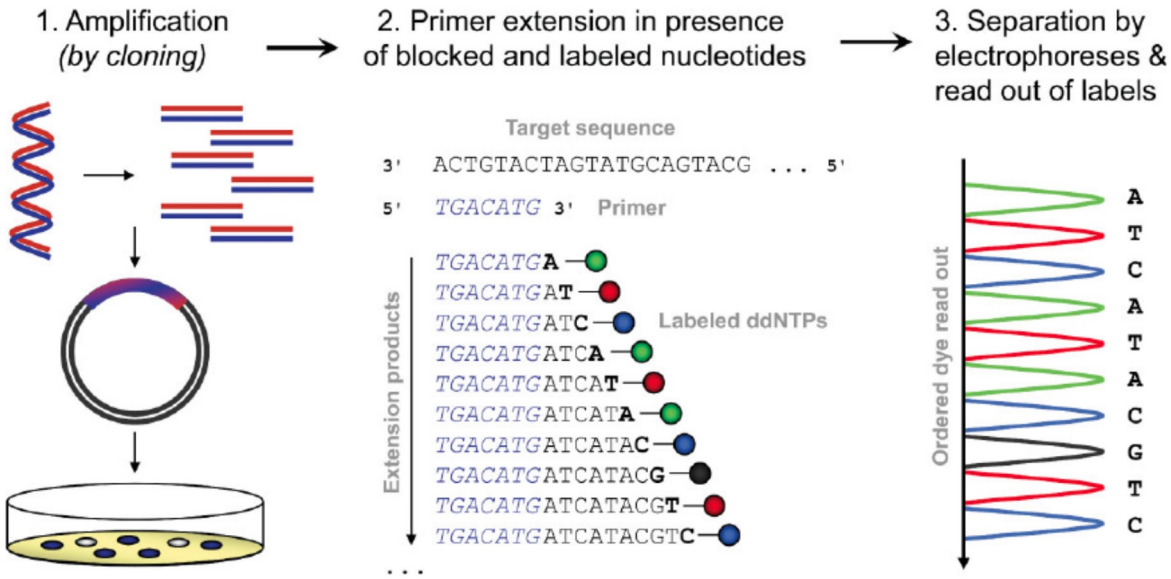
How Does Sequencing Work?

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



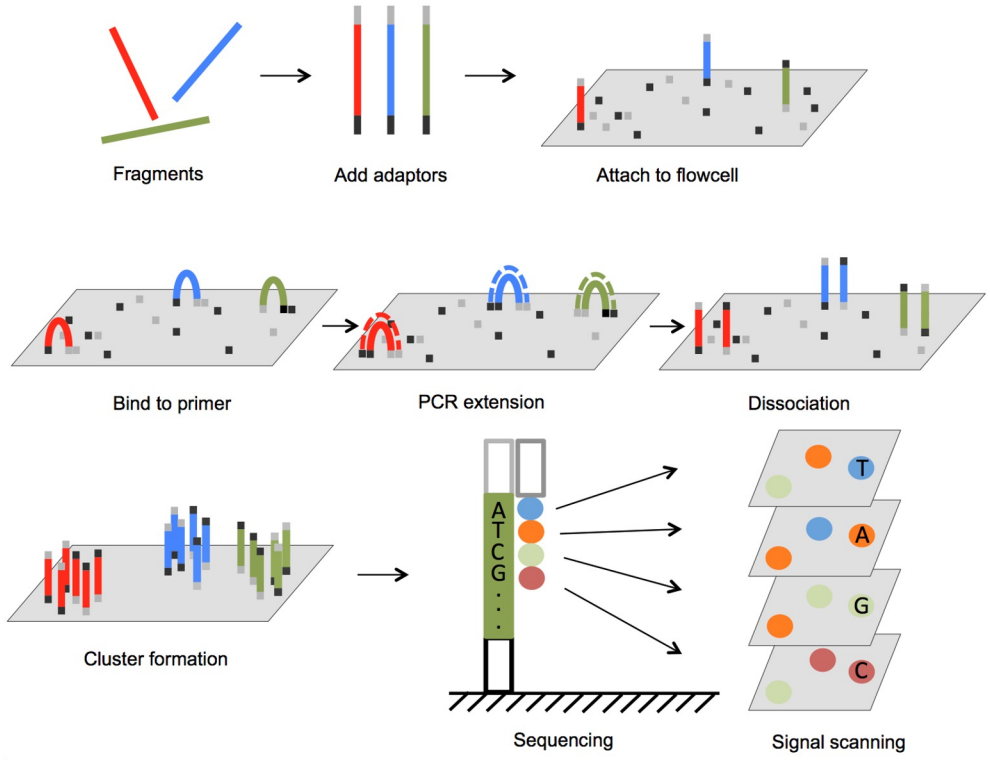
How Does Sequencing Work?

Illumina Sequencing

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



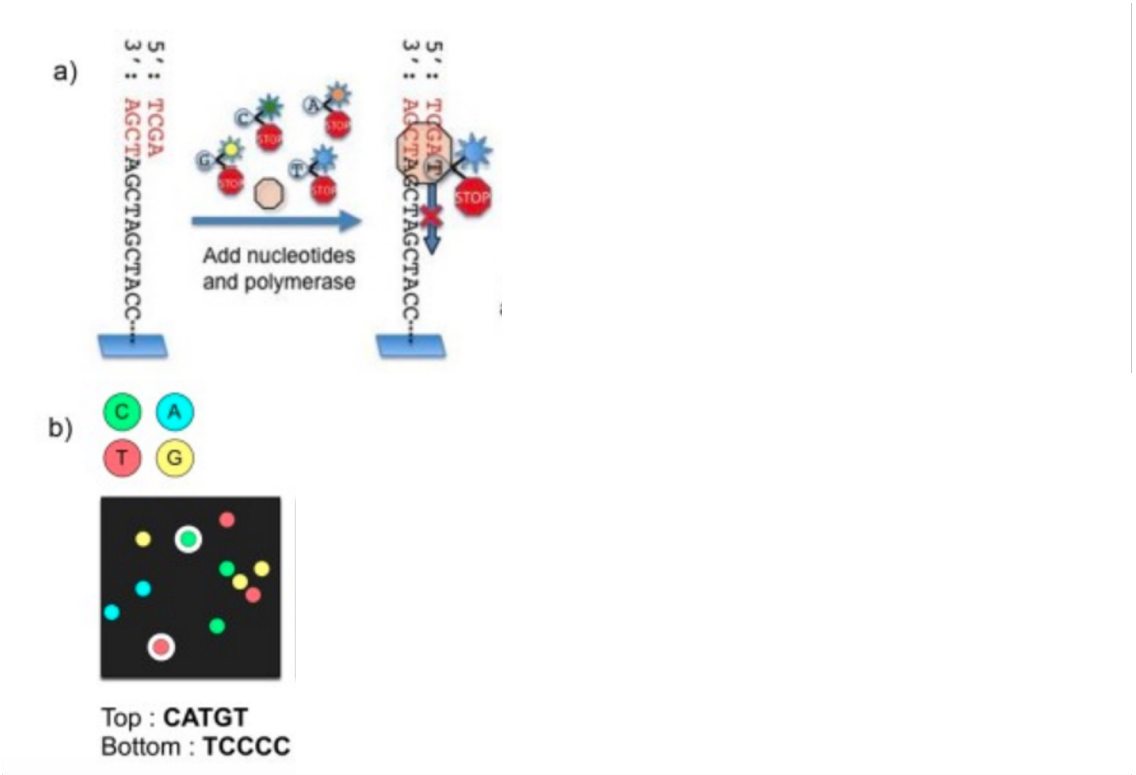
How Does Sequencing Work?

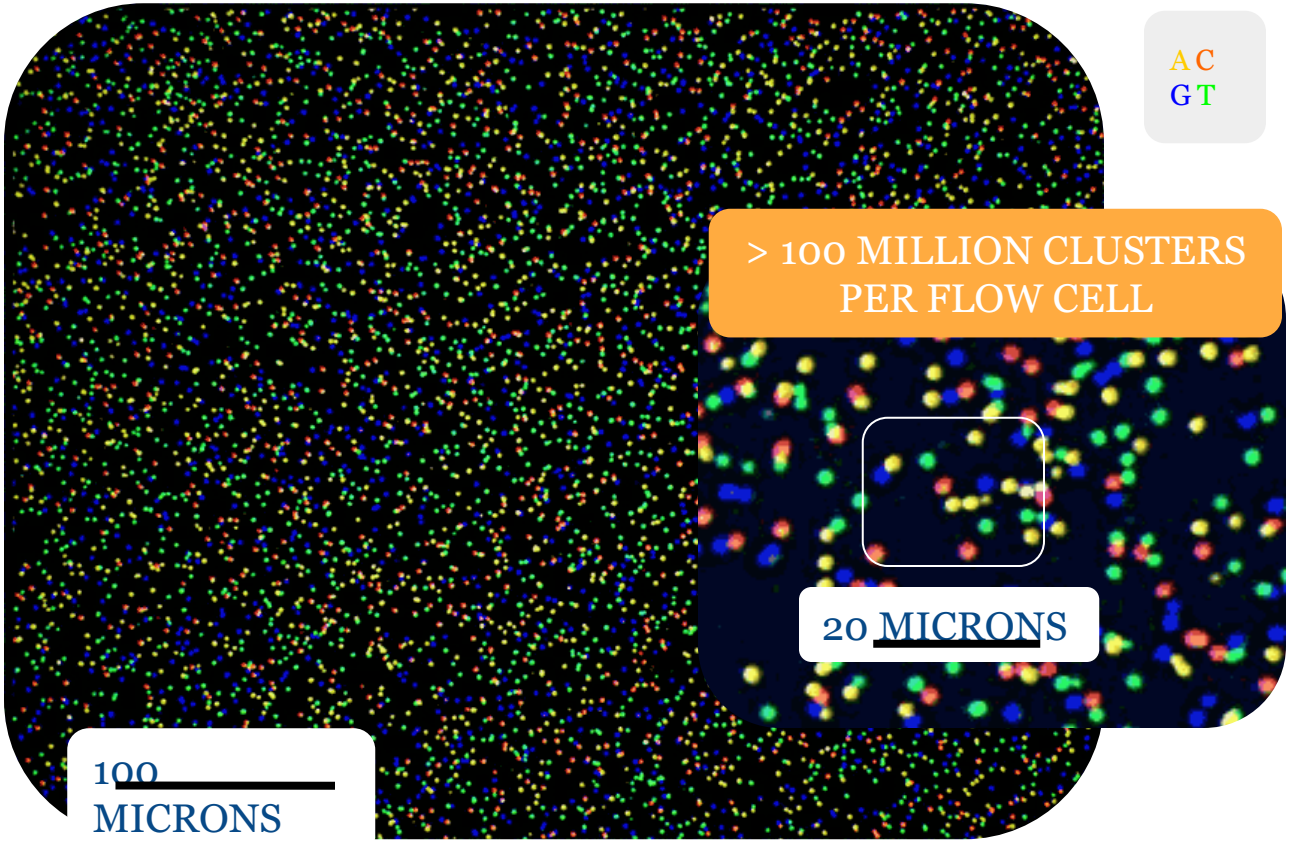
Illumina Sequencing

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 C
G T
> 100 MILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100
MICRONS

Illumina reversible terminator

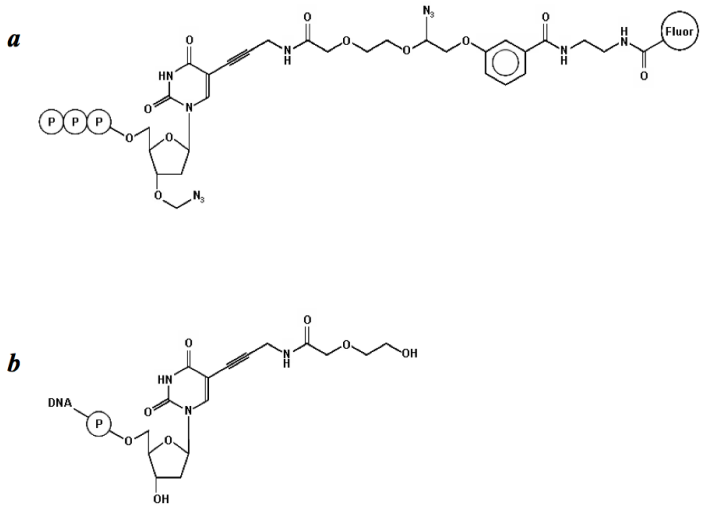


Figure S1. *a*. Structure of the reversible terminator 3'-*O*-azidomethyl 2'-deoxythymine triphosphate (T) labelled with a removable fluorophore. *b*. Structure of the incorporated nucleotide after removal of the fluorophore and terminator group. Each of the four nucleotides have an equivalent structure to the one shown here, except for the different base and a corresponding base-specific fluor.

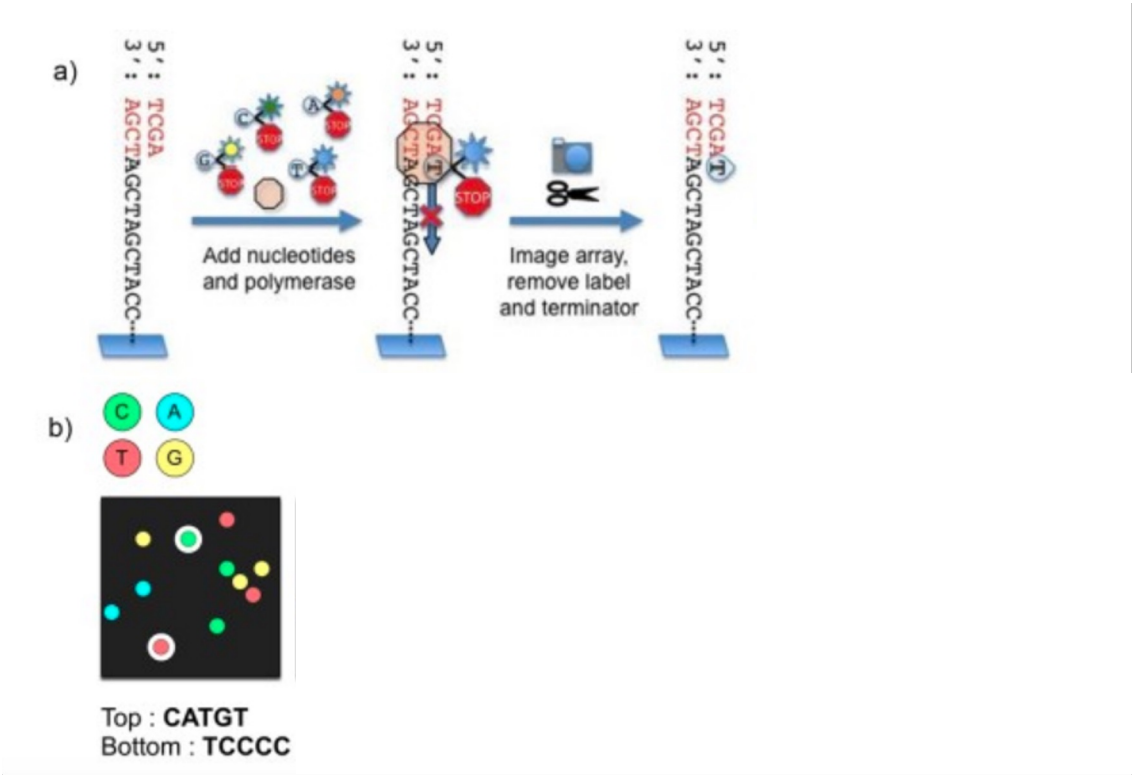
How Does Sequencing Work?

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



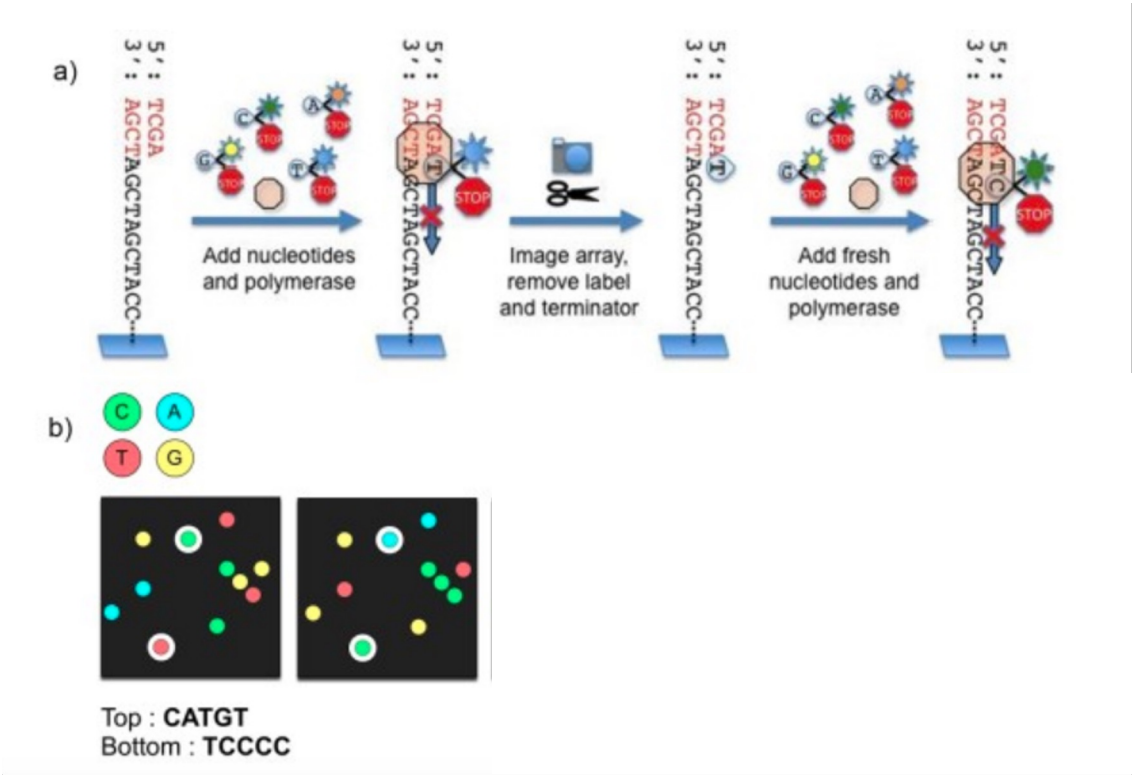
How Does Sequencing Work?

Illumina Sequencing

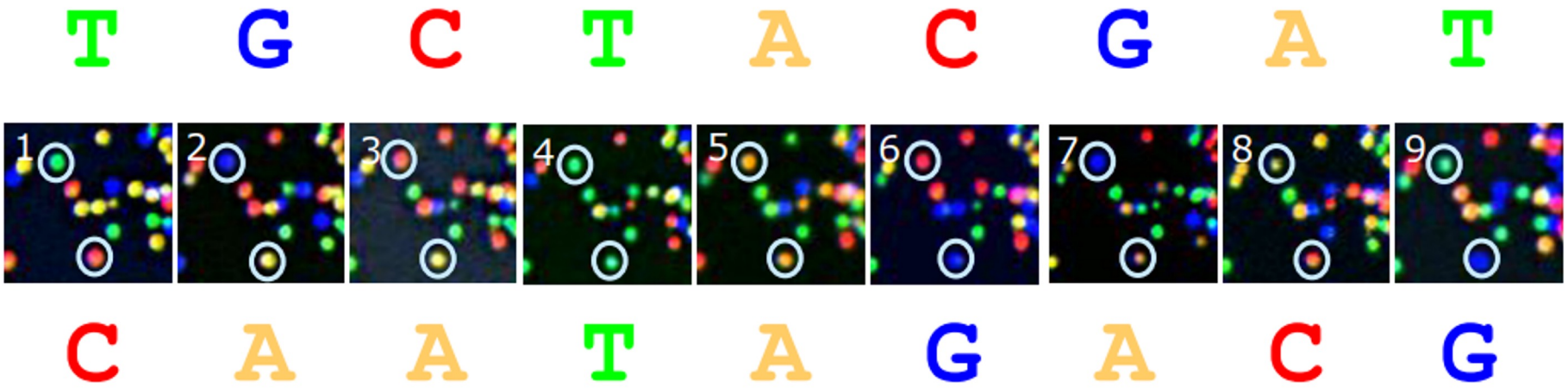
Blocked and labelled nucleotides are added

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

Label and block are removed and cycle repeats



Basecalling



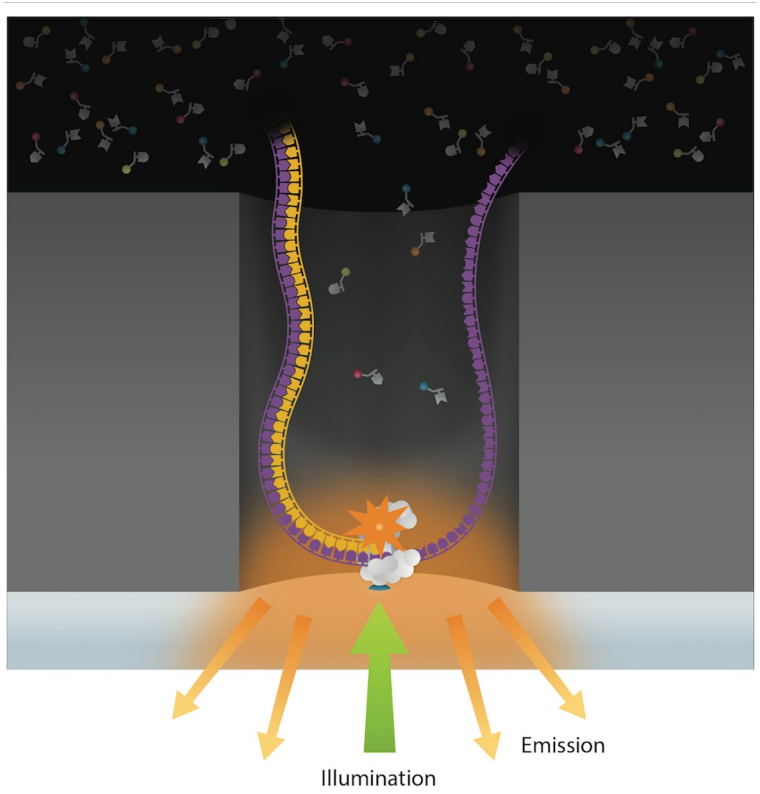
How Does Sequencing Work?

Third Generation
PacBio

A polymerase trapped in a well on a plate synthesises DNA

Labelled phosphonucleotides enable the sequence to be read from emission spectra

Long reads



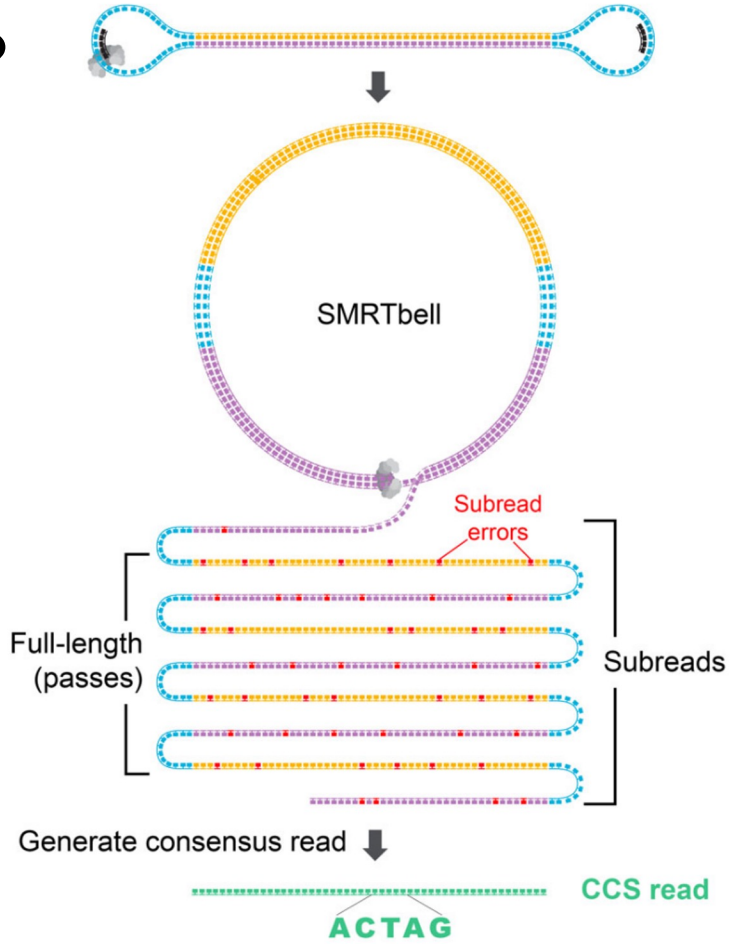
How Does Sequencing Work?

Third Generation
PacBio

A polymerase trapped in a well on a plate synthesises DNA

High stochastic error rate can be mitigated by circular consensus sequencing

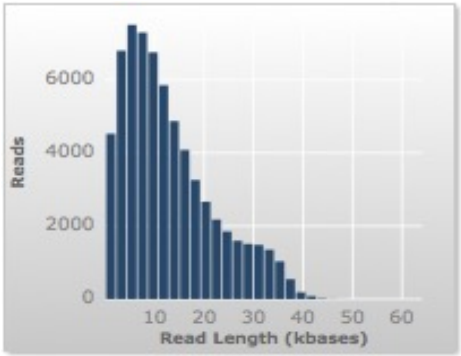
This reduces the read length



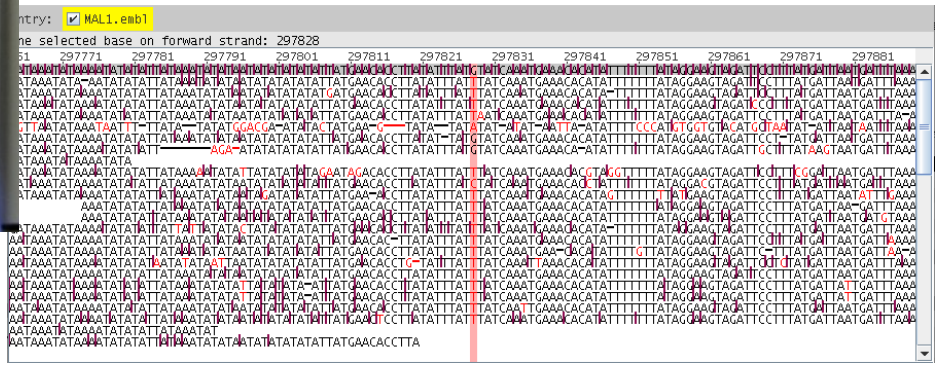
Mapping & Co.

PacBio RS

raw reads



reads of insert



Vertical lines are insertions in read

ABN map

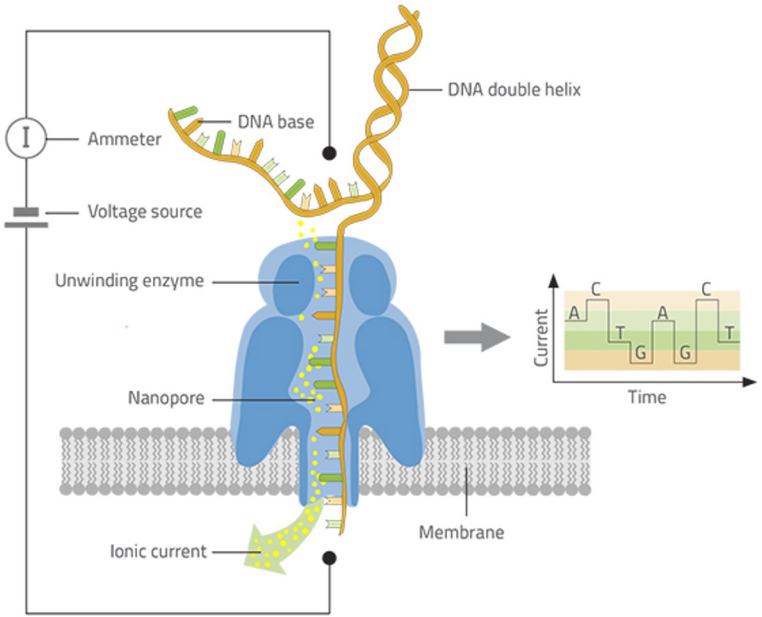
How Does Sequencing Work?

Third Generation
Oxford Nanopore

Protein (alpha haemolysin) pores are created in a flow cell

Applying a current across the flow cell allows DNA to pass through the pores

Sequence is read as disturbances in the electrical current



How Does Sequencing Work?

Third Generation
Oxford Nanopore

A processive enzyme (e.g., an exonuclease) may be coupled to the pore to slow down the process

Read length restricted only by DNA extraction techniques

Cheap and highly portable

High error rate with non-stochastic profile

Which Technology to Use?

Sanger	Illumina	PacBio	Nanopore
Cheap	Mid	Expensive	Cheap
1000 bp	30 - 350 bp	10,000 bp	100,000 bp (?)
99% accuracy	90% accuracy	85% accuracy	70% accuracy
Low throughput			Portable
Sequence a mutant or construct	Genome wide SNV analysis, differential expression	Generating a new reference assembly	Reference assembly, field monitoring

Why Do We Sequence Genomes?

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

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

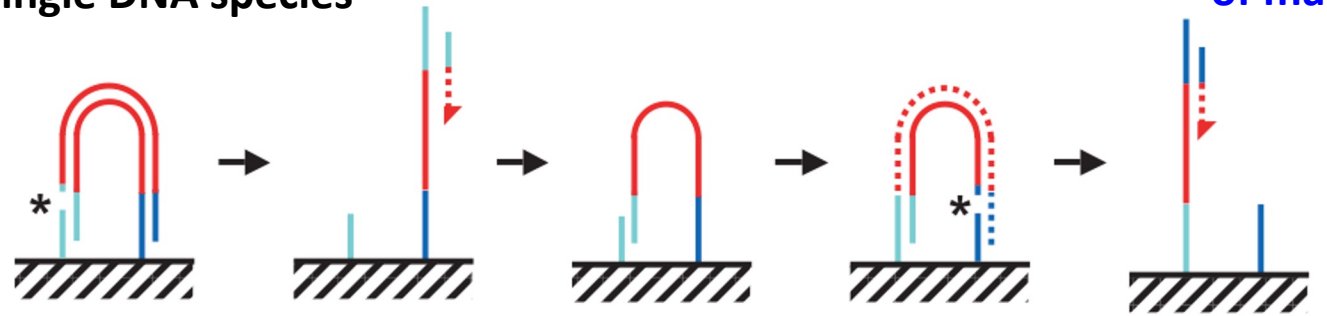
Which Technology to Use?

Sanger	Illumina	PacBio	Nanopore
Cheap	Mid	Expensive	Cheap
1000 bp	30 - 350 bp	10,000 bp	100,000 bp (?)
99% accuracy	90% accuracy	85% accuracy	70% accuracy
Low throughput			Portable
Sequence a mutant or construct	Genome wide SNV analysis, differential expression	Generating a new reference assembly	Reference assembly, field monitoring

What Does Illumina Sequencing Look Like?

From Cluster to Sequence

Denaturing
& washing
➔ single DNA species



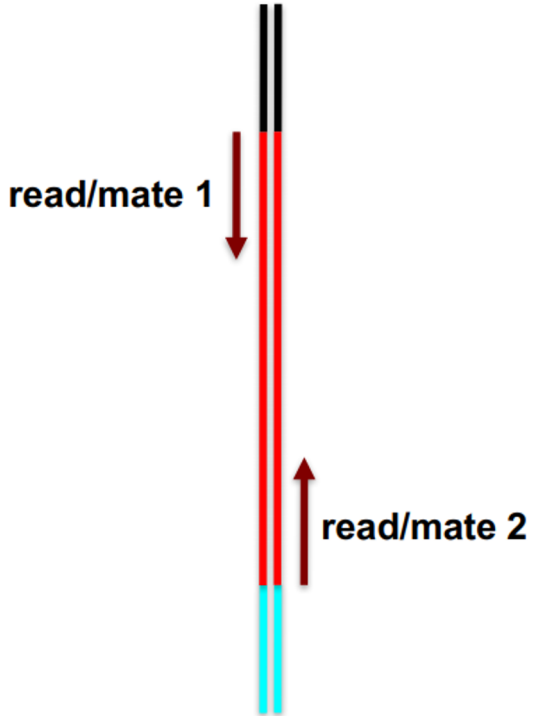
Adaptor-specific DNA-cleavage

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

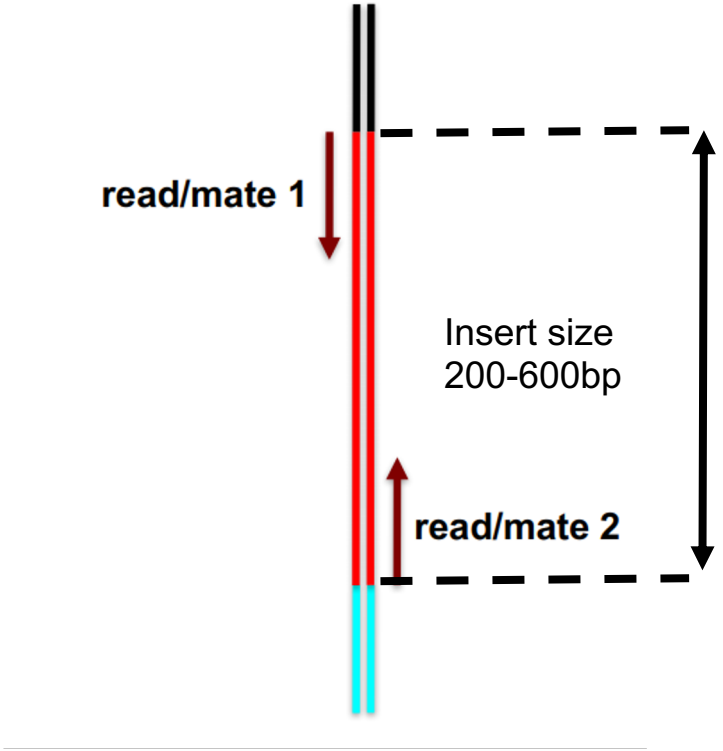


Paired End Sequencing

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



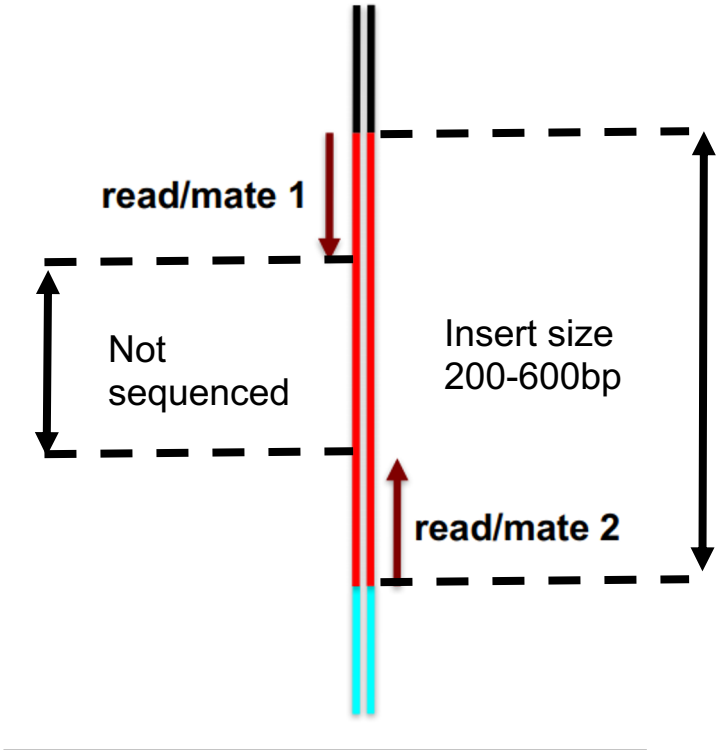
Paired End Sequencing

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



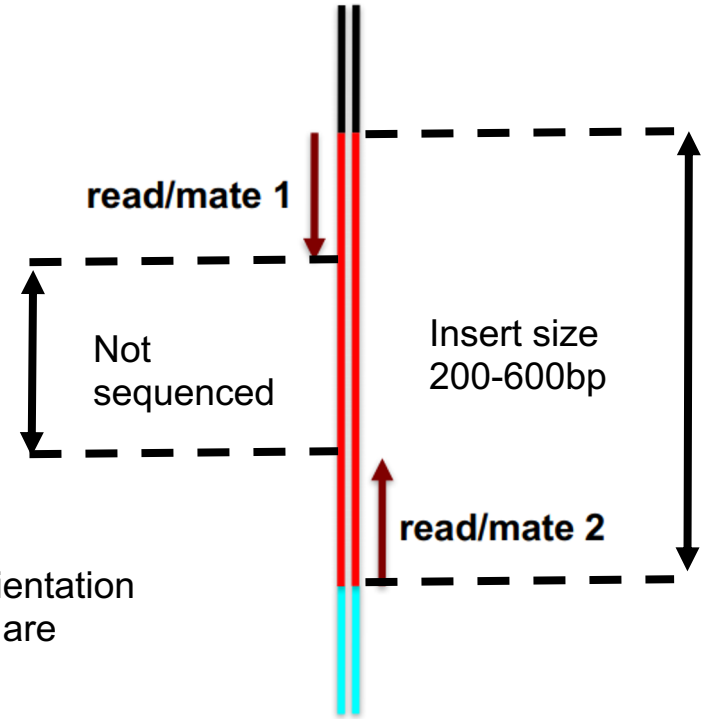
Paired End Sequencing

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

Illumina Overview

- **Highly parallel:**100 million clusters sequenced at the same time
- **Short sequence reads:** 35 bp - 350 bp
- **Paired end sequencing:** sequencing both ends of the fragment improves mapping accuracy
- **Multiplexing:** Sequence multiple samples together
- **Error rate:** 0.1%

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#EEEEEEEEEEEEEEEEEEEEAEEEEEAEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEAEEEEEEEEEA
```

For paired-end reads you will have two files

4 lines per read

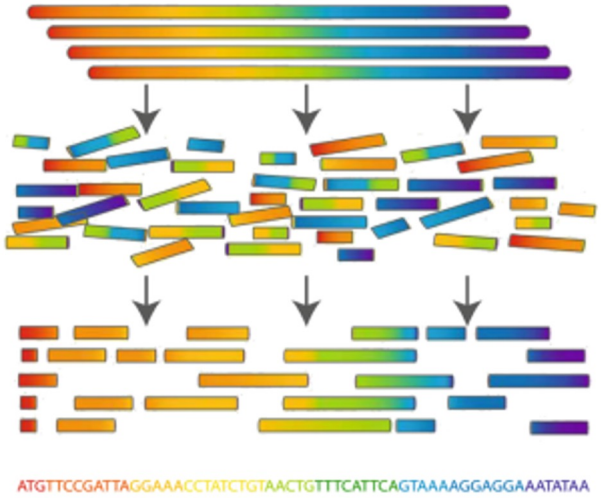
- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read
- Line 4 is the quality for each base
    - Quality is encoded using ASCII
    - http://www.asciitable.com/

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%

De Novo Assembly

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



Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

↓

Alignment

↓

SNP Calling

Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

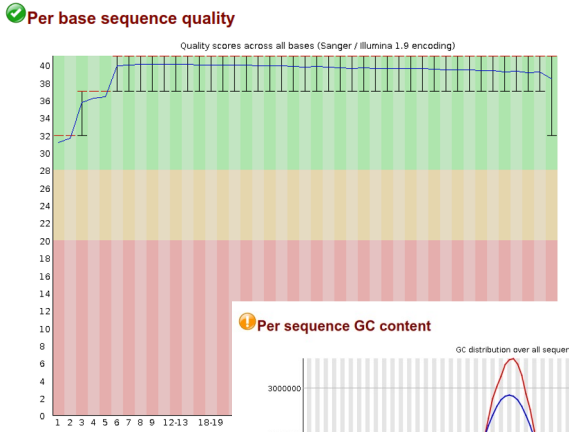
↓

Alignment

↓

SNP Calling

- 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

Quality Control

↓

Read Trimming

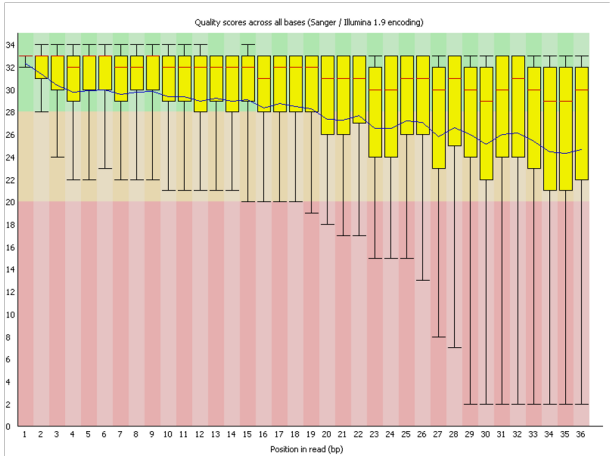
↓

Alignment

↓

SNP Calling

- 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



Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

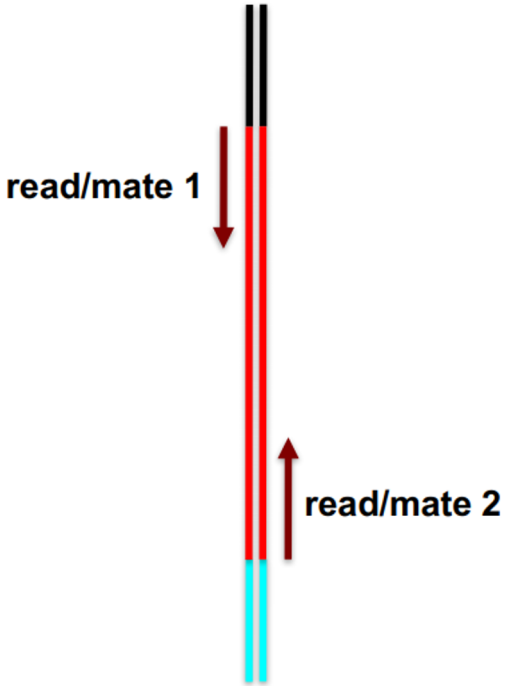
↓

Alignment

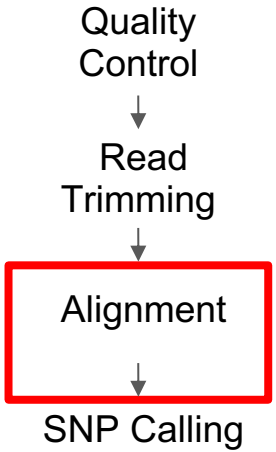
↓

SNP Calling

- 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



Analysis of NGS Sequencing



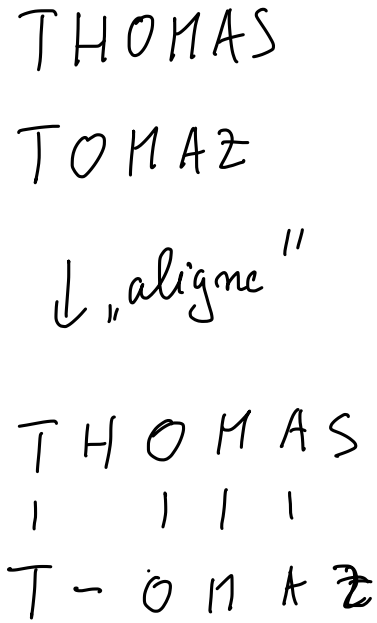
What is an alignment?

Two sequences:

ATTGAAAGCTA
GAAATGAAAAGG

How would you align one to the other?

- THOMAS
- T-OMAZ



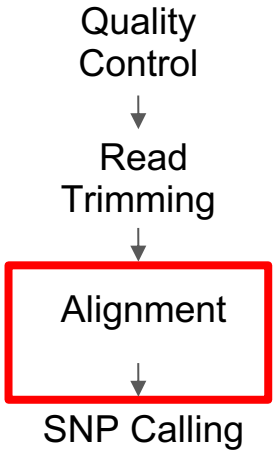
What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

Analysis of NGS Sequencing



What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

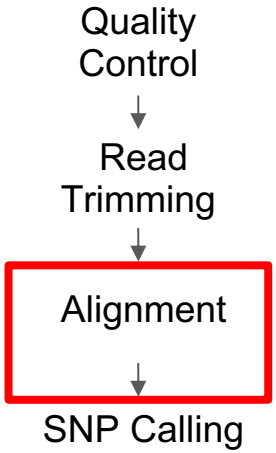
How would you align one to the other?

```
--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

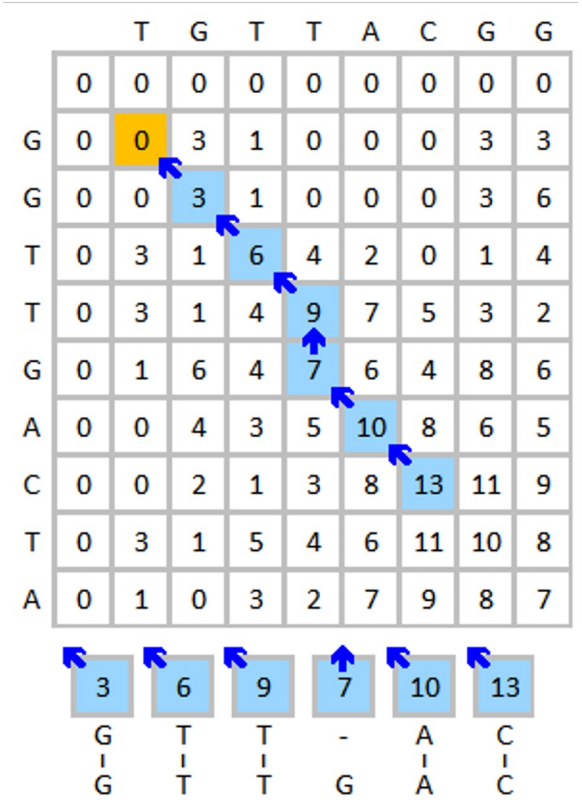
Alignment Algorithms - Smith Waterman

Algorithm for pairwise local alignments

Exhaustive - will find the best alignment

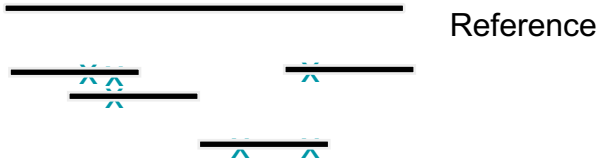
Slow and memory intensive

Too slow for high-throughput sequencing (hundreds of millions of reads)

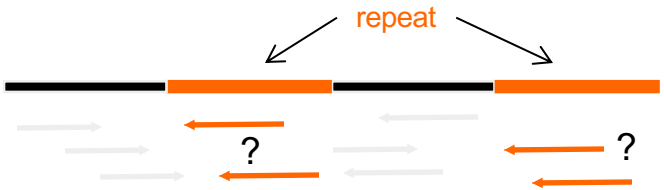


Assembly

Mapping



- Mapping is easier than assembly
- Longer reads are easier to analyze



Read pairs help to position the reads

- Read pairs help to resolve repeats
- More coverage is better
- Accuracy

Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

↓

Alignment

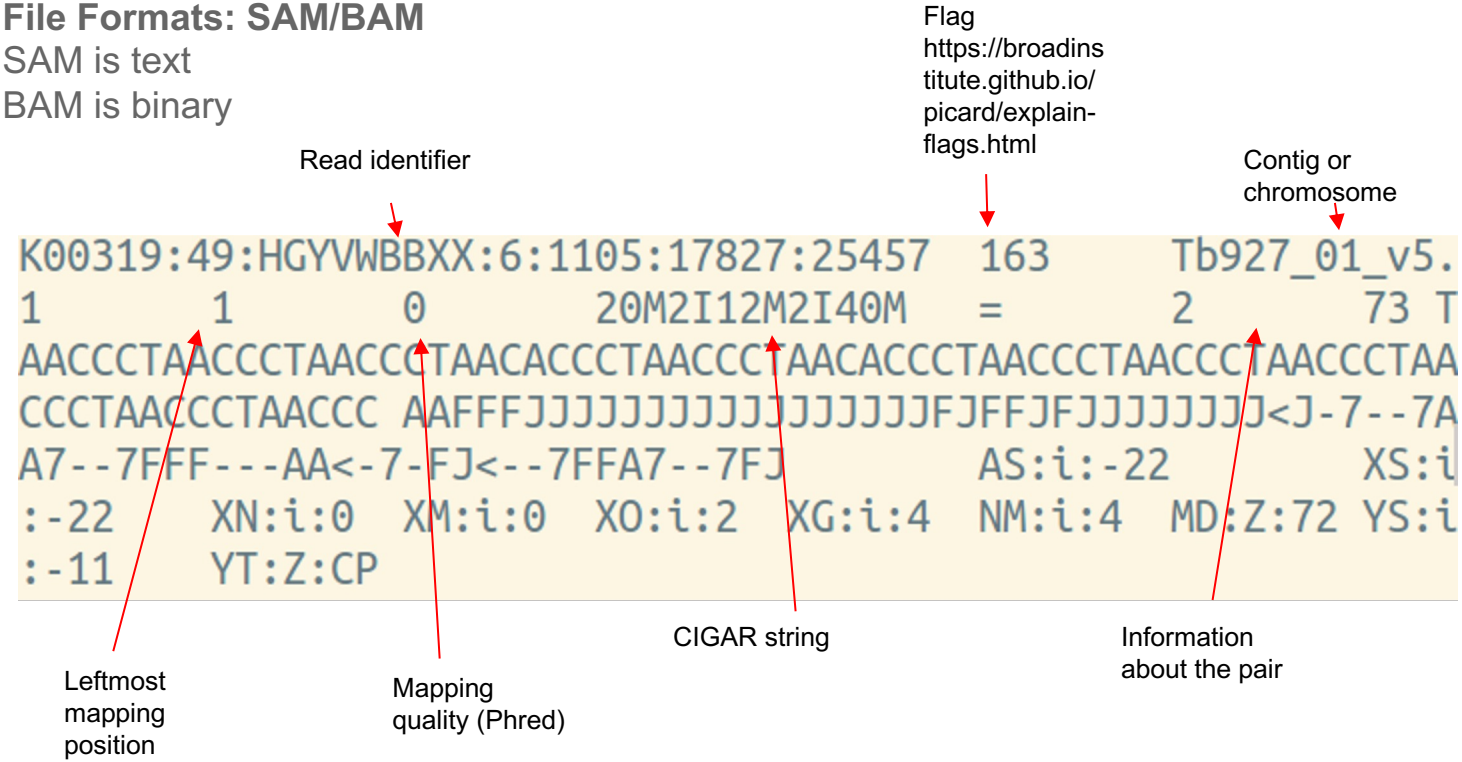
↓

SNP Calling

File Formats: SAM/BAM

SAM is text

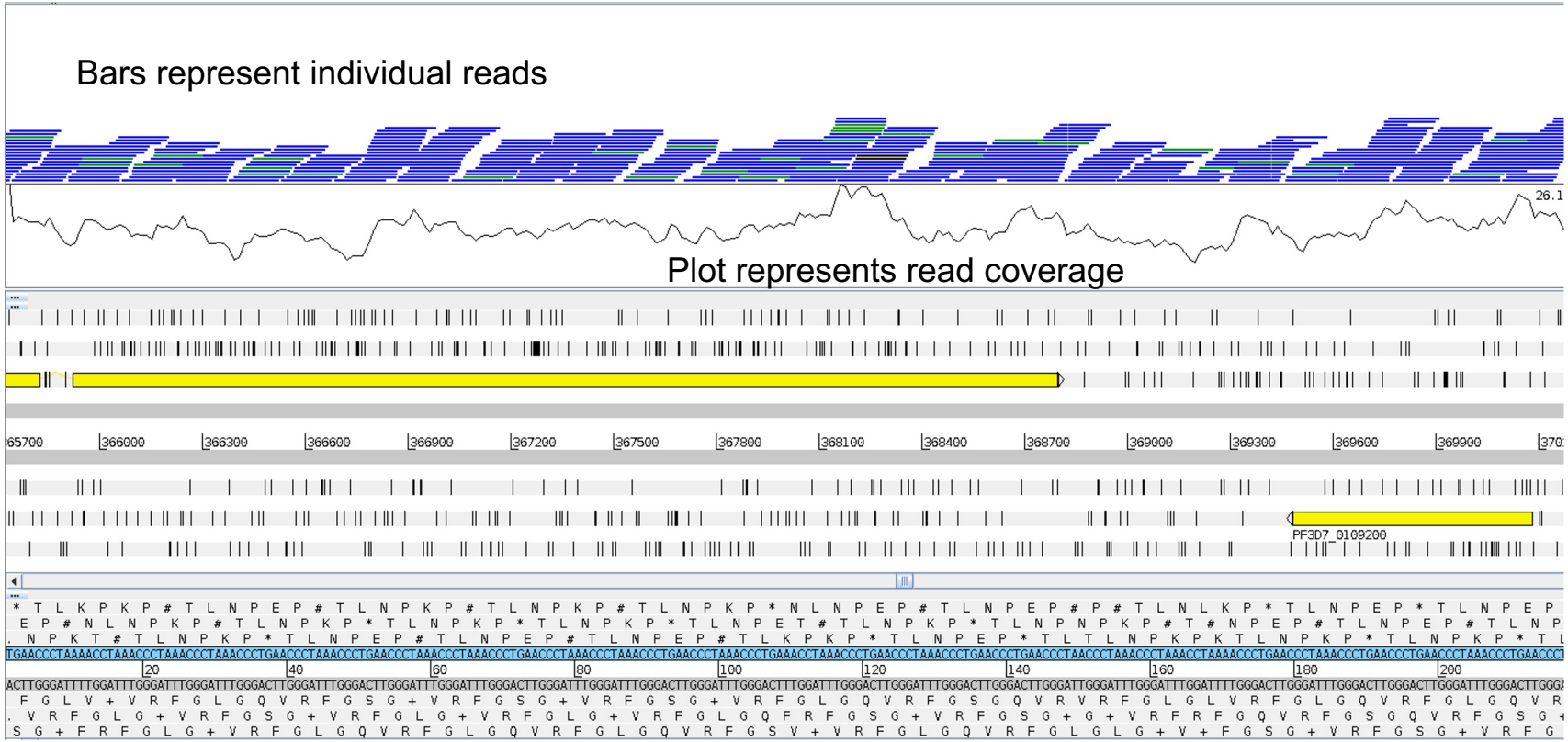
BAM is binary



Samtools

- Samtools is a set of software tools for manipulating SAM and BAM files
  - Conversion between formats
  - Filtering
  - Sorting
  - Quickly view BAM files
- In the exercises we will use samtools to convert SAM to BAM, sort BAM files and index BAM files for visualisation and downstream analysis
- http://www.htslib.org/doc/samtools.html

Visualising Alignments in Artemis



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

Analysis of NGS Sequencing

Quality Control

↓

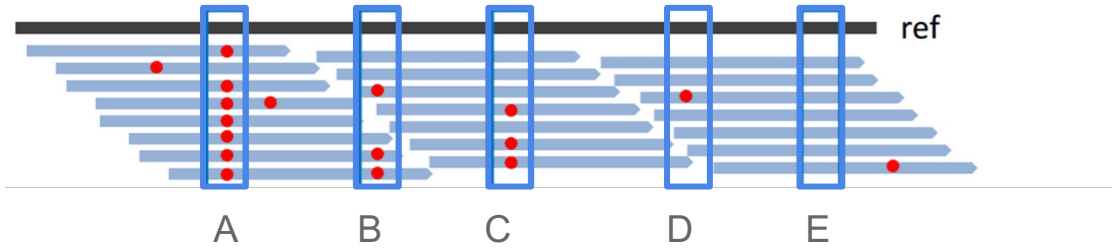
Read Trimming

↓

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.

What could die different blue boxes represent?

Analysis of NGS Sequencing

Quality Control

↓

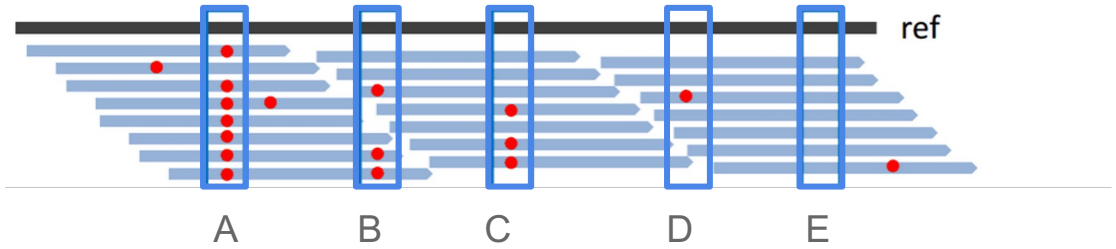
Read Trimming

↓

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

↓

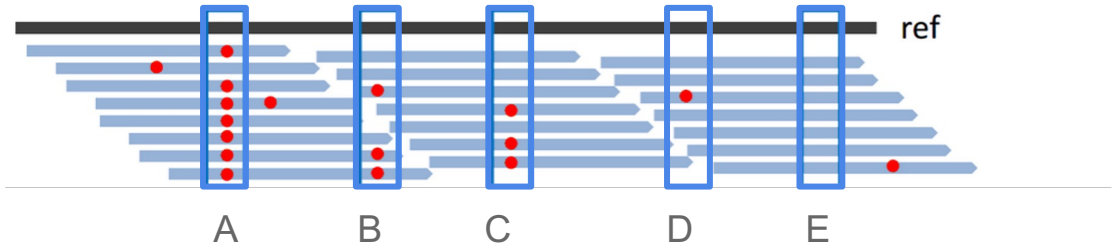
Read Trimming

↓

Alignment

↓

SNP Calling



Things to think about:

- 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
  - Too much depth can introduce noise

Analysis of NGS Sequencing

Quality Control

↓

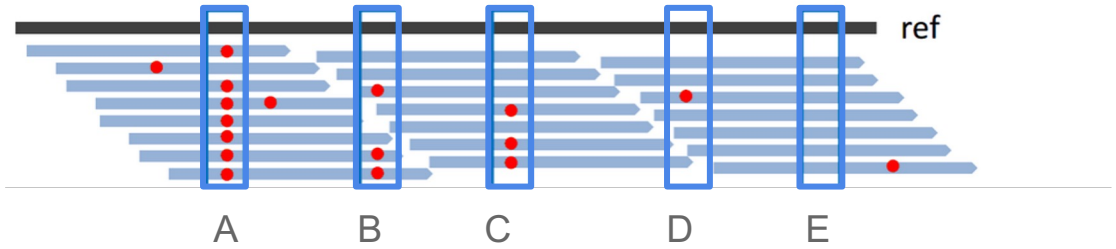
Read Trimming

↓

Alignment

↓

SNP Calling



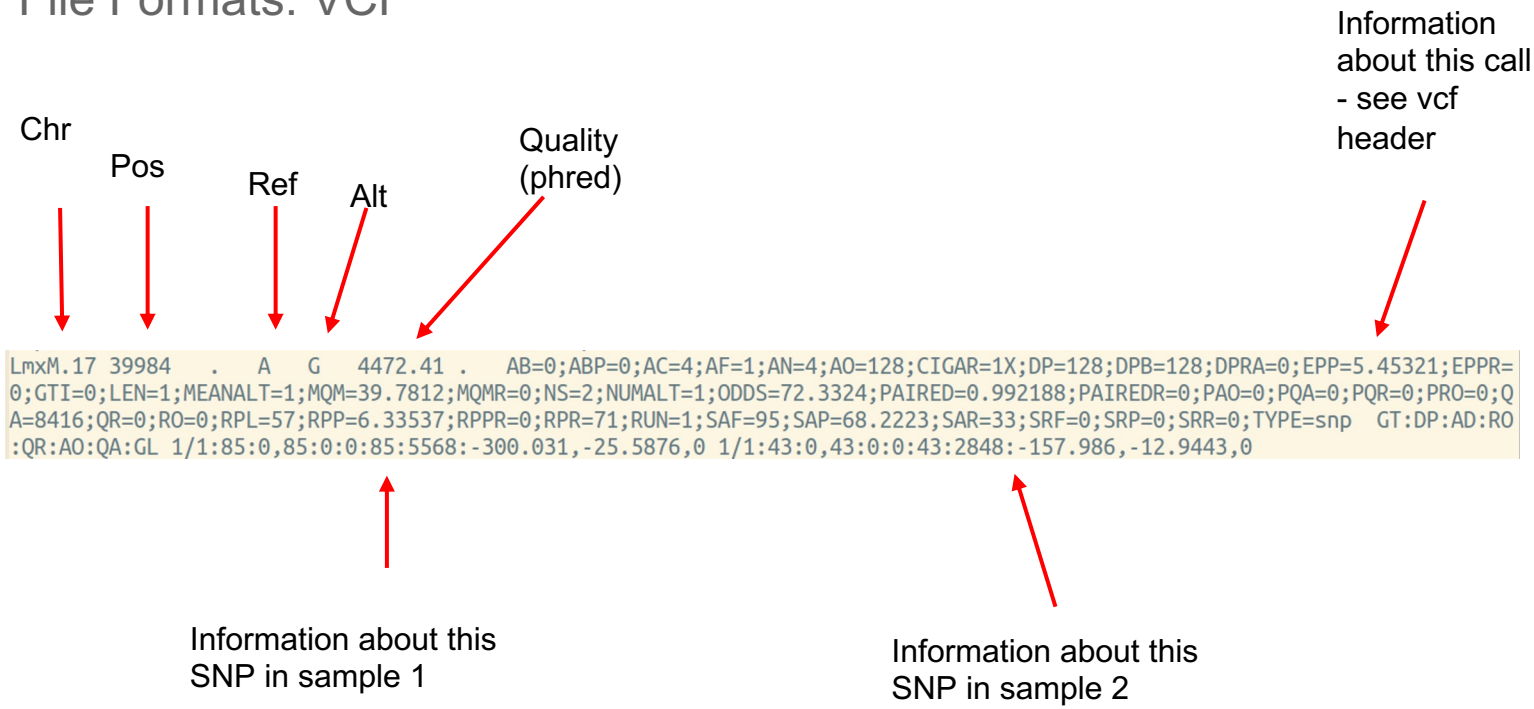
BCF tools MPileup https://samtools.github.io/bcftools/howtos/variant-calling.html

Automated tool to call SNPS

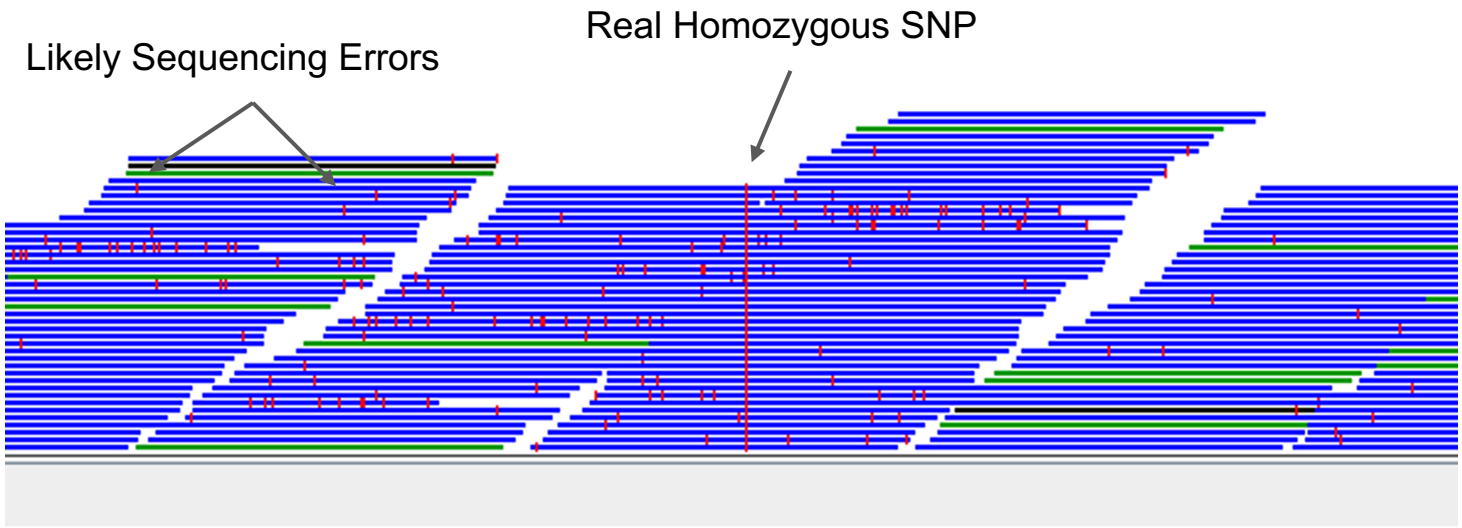
You may also come across other tools including GATK and Freebayes.

Calling SNPs

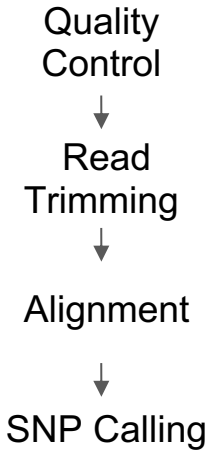
File Formats: VCF



Visualising SNPs in Artemis



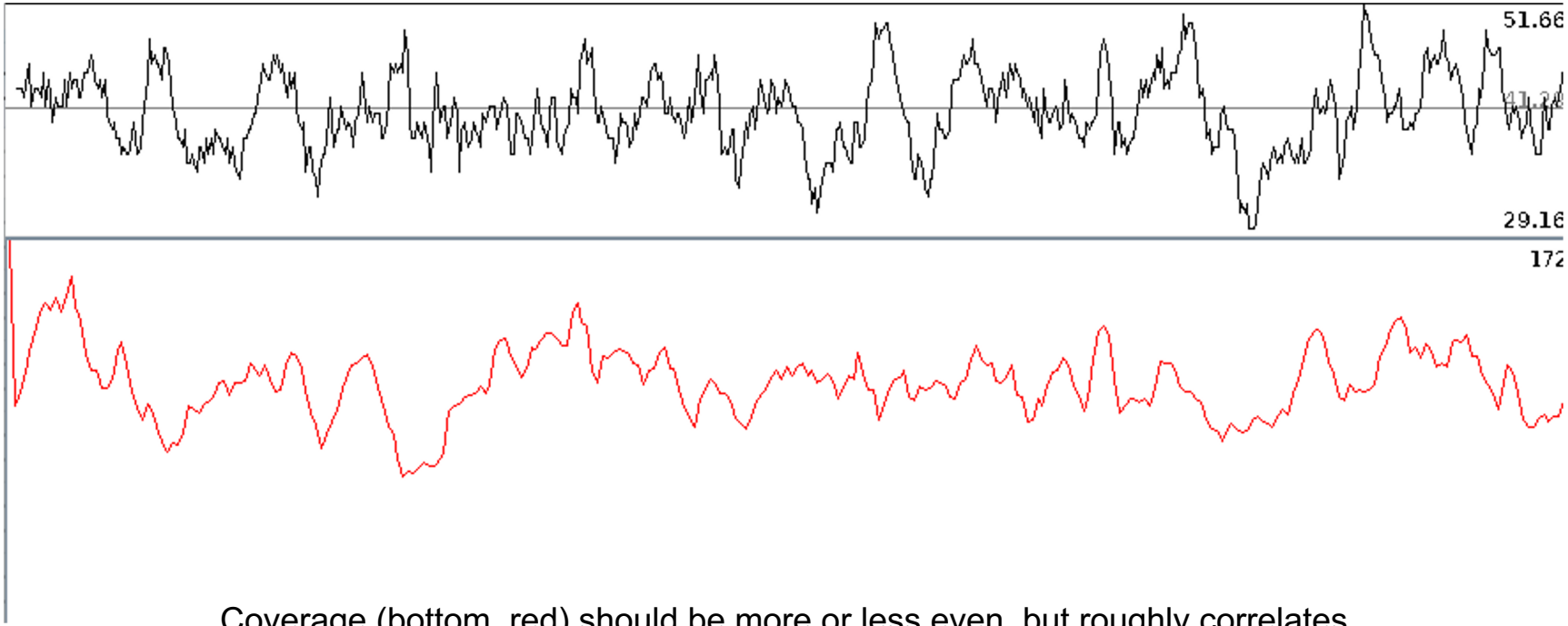
What Else Can We Find Out?



Coverage

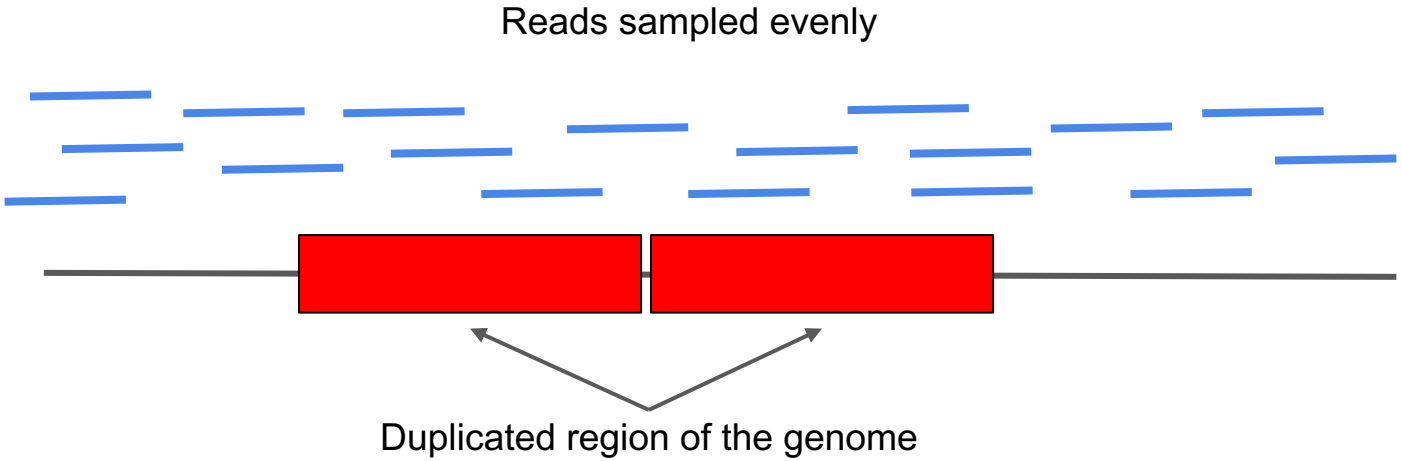
- Expect coverage to be even across the genome
- In reality, we see variation associated with:
  - GC content
  - Repetitive or highly variable regions
  - Large scale insertions and deletions
- Note that doing alignments and examining coverage is the basis of RNAseq analysis too!

Coverage and Copy Number Variations

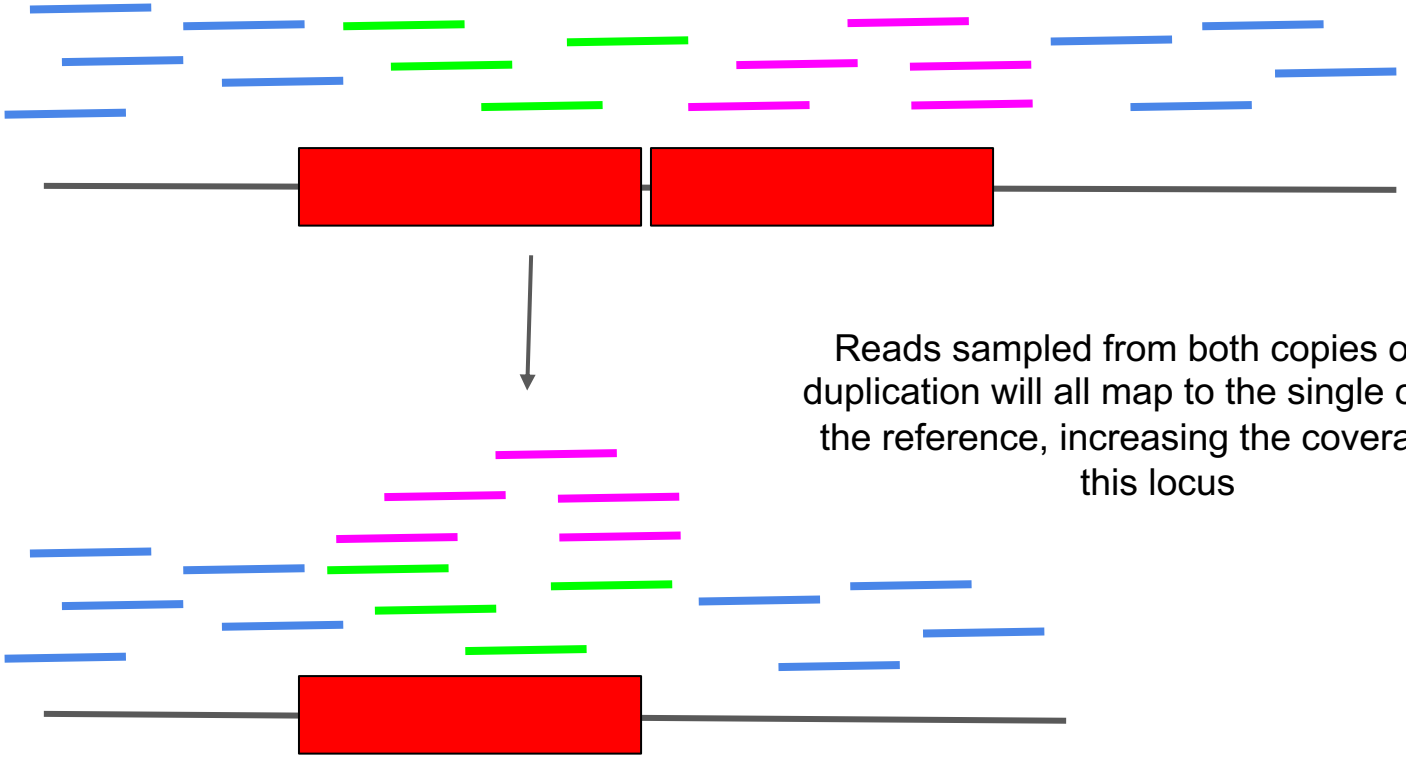


Coverage (bottom, red) should be more or less even, but roughly correlates with GC content (upper black).
This can be mitigated by using amplification-free library prep methods.

Coverage and Copy Number Variations



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

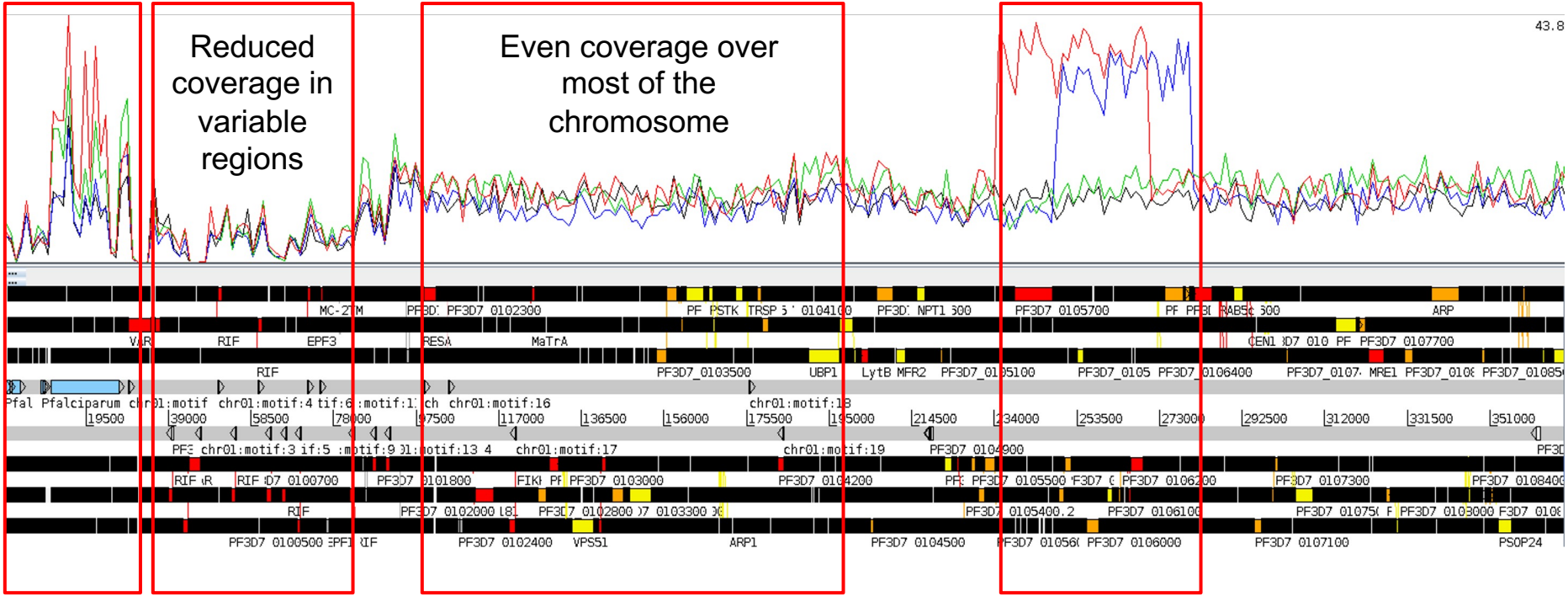
Global Coverage in Artemis

Variable coverage in repetitive regions

Reduced coverage in variable regions

Even coverage over most of the chromosome

Duplications in two samples



Example of sequence in fastq format

```
@IL39_6014:8:75:16399:10217#12/1
CCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAACCTAAACCTAAACCCTAAATCCTAAACCCTGAAA
+
HGGHHHHGHHGGGGGGHHB9GGHHGGGGGGHGCAEHEG;CB=DAE;:=A/A5667)=>2937&6:&055.5621/3&
@IL39_6014:8:13:15323:7428#4/1
```

Reference are given in the fasta format

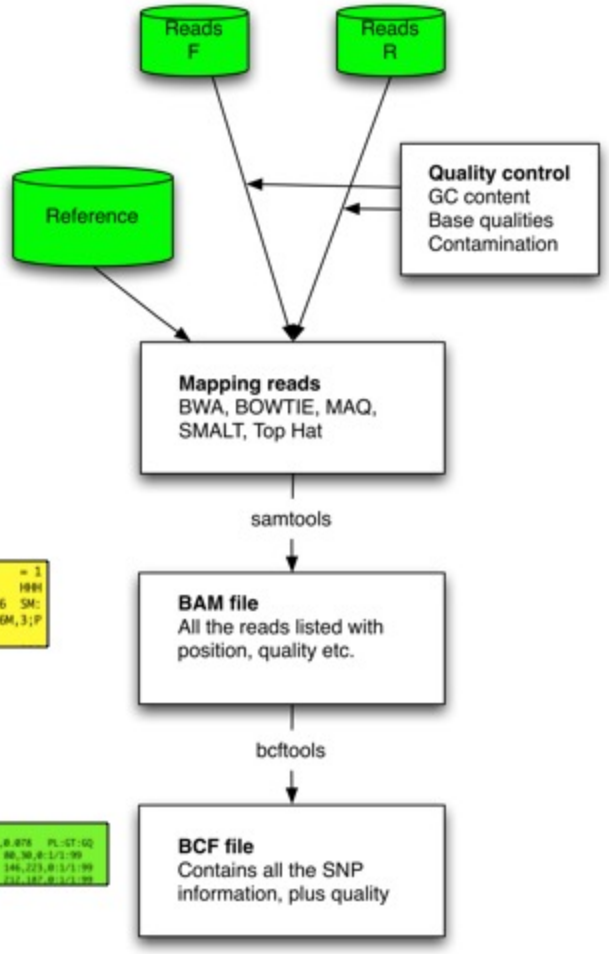
```
>MAL1
ctaaaccctaaaccctaaaccctgaaccctaaaccctaaaccctgaaccctaaaccctgaac
cctgaaccctaaaccctgaaccctaaaccctgaaccctgaaccctaaaccctaaaccctaa
aaccctaaaccctaaaccctaaaccctgaaccctaaaccctgaaccctaaaccctaaaccc
ctaaaccctaaaccctaaaccctgaaccctaaaccctgaaccctgaaccctaaaaccctaa
```

The mapping tools generate sam/bam files

```
IL39_6014:8:45:11459:13097#6    99      Pf3D7_05        1       14      2S9M1D23M1D34M1D8M      = 1
70      246     CCCTAAACCCTAAACCTAAACCCTGAACCCTAAACCCTAAACCCTGAACCCTAAACCCTAAACCCTAAACCTAAACC  HHH
HHHHHHHHHHHHHHH2HHHHGHHGGGHFHGG=EFFCEGGBCD>ECBDBEE@EA;BB<A=@;>?9A>:<=@;E       XT:A:M  NM:i:6  SM:
i:14    AM:i:14 XM:i:3  XO:i:3  XG:i:3  MD:Z:9^GZC20^C24G9^C3G4 XA:Z:Pf3D7_05,-1343398,40M1D36M,3;P
f3D7_05,-1343268,11M1D65M,3;
```

From bam files variation is called and stored in vcf/bvf format

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Res.5853.s_6.bam
MAL1	57405	.	A	C	36.2	.	DP=28;AF1=1;CI95=0.5,1;DP4=1,0,0,4;MQ=26;FQ=-35;PV4=0.2,0.018,1,0.078	PL:GT:GQ	
MAL1	57450	.	C	T	47	.	DP=57;AF1=1;CI95=1,1;DP4=0,0,0,10;MQ=28;FQ=-57	PL:GT:GQ	80,30,0:1/1:99
MAL1	57484	.	C	T	113	.	DP=76;AF1=1;CI95=1,1;DP4=0,0,1,73;MQ=24;FQ=-250	PL:GT:GQ	146,225,0:1/1:99
MAL1	57527	.	C	G	179	.	DP=63;AF1=1;CI95=1,1;DP4=0,0,1,61;MQ=28;FQ=-214	PL:GT:GQ	232,187,0:1/1:99



Working at the Command Line

```
$ bwa index reference
```

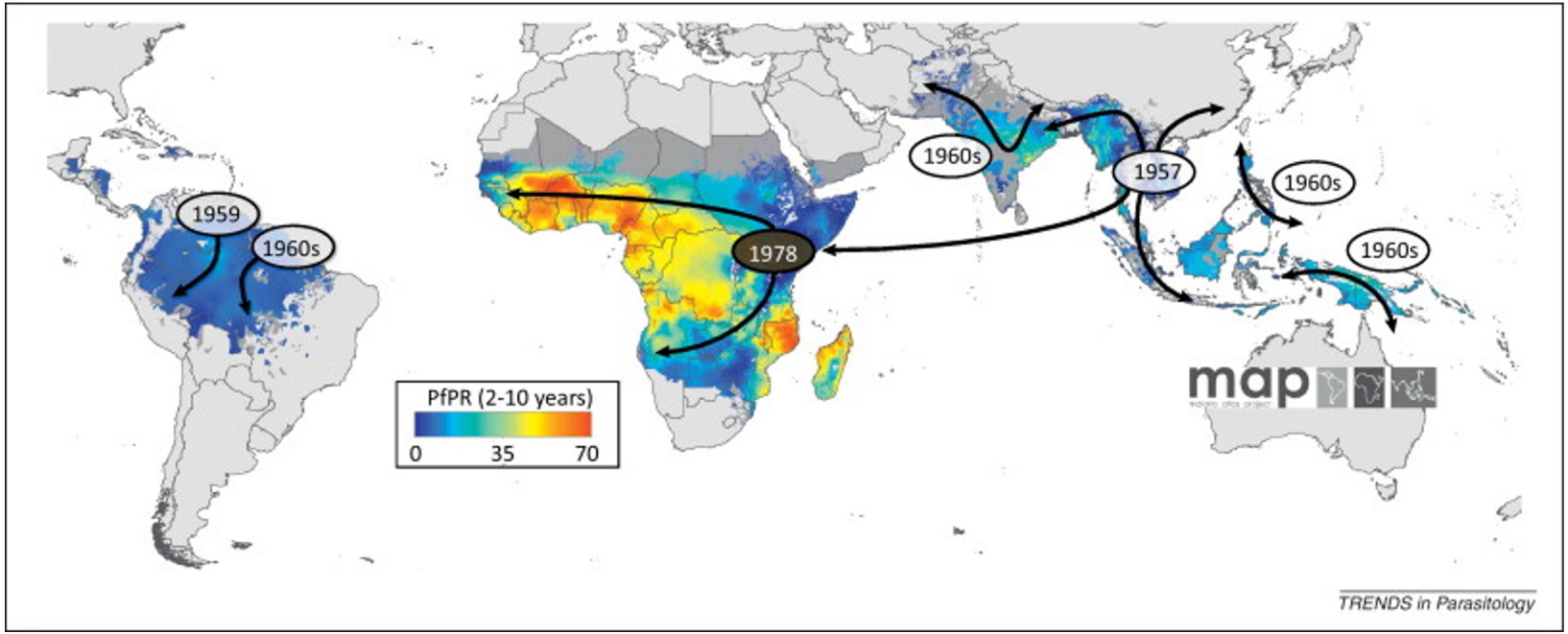
Next, read files (read_1: `Reads.18_1.fastq` read_2: `Reads.18_2.fastq`) are mapped against the reference and a sam files is generated. The program is bwa and the program part is “mem”. So type to see the options.

```
$ bwa mem
```

To see the options. The full command with the place holder is

```
$ bwa mem -t 8 reference read_1 read_2 > BWA.sam
```

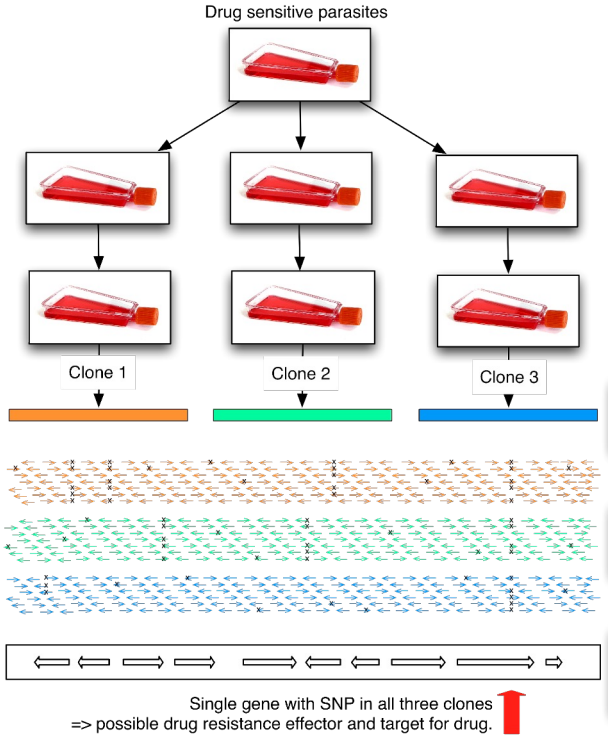
Exercise - Drug Resistance in *Plasmodium*



Exercise- Drug Resistance in *Plasmodium*

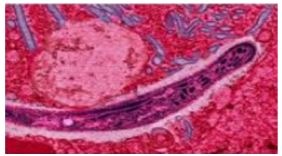
- *Plasmodium* is a protozoan parasite that causes malaria
- Drug resistance is a problem - development of new drugs is important
- For licensing, it is important to know the mode of action of a drug
- Grow parasites in increasing concentrations of the drug until they become resistant - look at what has changed
- We will align three drug resistant lines to the reference and look at what has changed to determine the drug target
- Exercise 3 will repeat this in another parasite, *Leishmania*, but with less guidance

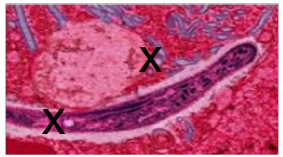
Exercise - Drug Resistance in *Plasmodium*

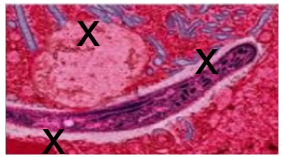


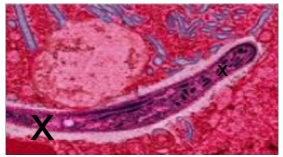
Spot the difference

parent









Detecting genes responsible for drug resistance

New compounds: Vaidya et al Nature communication 2014

Baragaña et al Nature 2015

Understanding gametocyte control Sinha et al Nature 2014

Questions?

exercise

- download the manual pdf from git-hub
- open a terminal in the VM
- starting reading the manual and do as suggested!

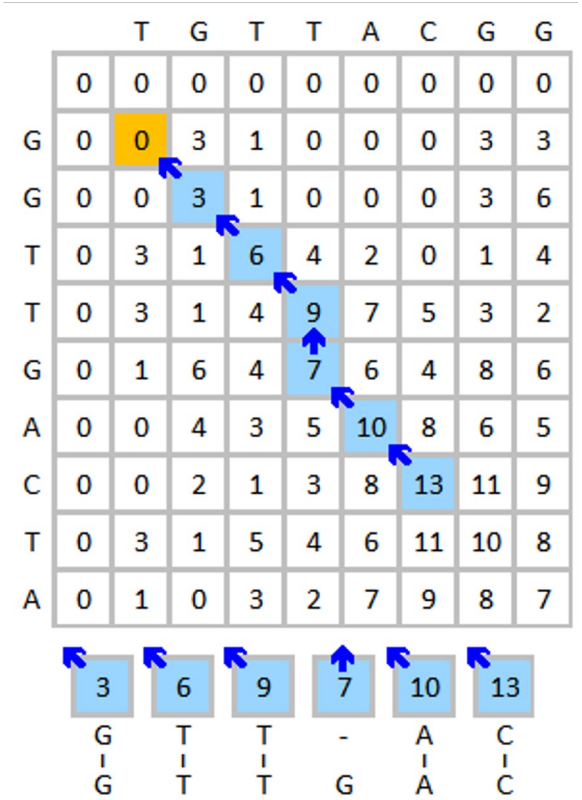
Alignment Algorithms - Smith Waterman

Algorithm for pairwise local alignments

Exhaustive - will find the best alignment

Slow and memory intensive

Too slow for high-throughput sequencing (hundreds of millions of reads)



Alignment Algorithms - Burrows Wheeler

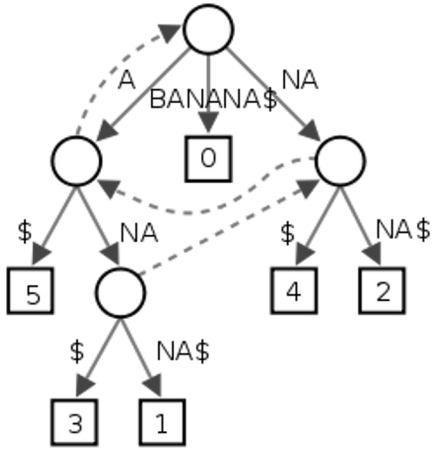
Algorithm most commonly used for aligning NGS data to a reference

Uses suffix tree for fast text searching combined with Burrows-Wheeler transform to reduce memory use

Can align NGS data on most computers

Heuristic not exhaustive, but “good enough”

Suffix tree of the word BANANA



Transformation				
1. Input	2. All rotations	3. Sort into lexical order	4. Take the last column	5. Output
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	**A**NANA|^B **AN**A|^BAN **A**|^BANAN **B**ANANA|^ **N**ANA|^BA **NA**|^BANA **^**BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|^ NANA|^**BA** NA|^BANA ^BANANA| |^BANAN**A**	BNN^AA|A

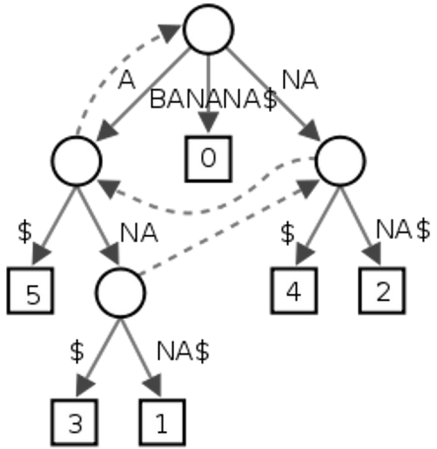
Alignment Algorithms - Burrows Wheeler

In the exercises, we will use BWA https://github.com/lh3/bwa

BWA is designed to efficiently map many short reads against a reference

Also takes into account pair concordance

Bowtie2 is a similar tool that you may come across



Transformation				
1. Input	2. All rotations	3. Sort into lexical order	4. Take the last column	5. Output
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	**A**NANA|^B **AN**A|^BAN **A**|^BANAN **B**ANANA|^ **N**ANA|^BA **NA**|^BANA **^**BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|^ NANA|^**BA** NA|^BANA ^BANANA| |^BANAN**A**	BNN^AA|A