

Mapping Sequence Data

Introduction

Improvements in DNA sequencing technology have led to new opportunities for studying organisms at the genomic and transcriptomic levels. Applications include studies of genomic variation within species and gene identification. In this module we will concentrate on data generated using the Illumina Genome Analyzer II, although the techniques you will learn are applicable to other technologies (e.g. Ion Proton or the Illumina HiSeq). A single machine can produce well over 20 Gigabases of sequence data in a week. This is the equivalent of more than 6 human genomes! The data from the Illumina machine comes as relatively short stretches of 35-350 base pairs (bp) of DNA - around 300 million of them. These individual sequences are called **sequencing reads**. The older **capillary sequencing** method produces longer reads of ~500bp, but it is much slower and more expensive to generate the same volume of data.

One of the greatest challenges of sequencing a genome is determining how to arrange sequencing reads into chromosomes. This process of determining how the reads fit together by looking for overlaps between them is called **genome assembly**. Capillary sequencing reads (~500bp) are considered a good length for genome assembly. 3rd generation technologies like Pacific BioScience or Oxford Nanopore are also very good, as they generate even longer reads and have higher throughput. The results for pathogens smaller than 30mb are very good.

Genome assembly using sequence reads of <100bp is more complicated due the high frequency of repeats longer than the read length. Assemblies for bacterial genomes are often in at least 50 pieces and for Eukaryotes the assembly is in more than 1000 pieces. Therefore short-read sequencing technologies are mostly used where a **reference genome** already exists. A reference genome is a well assembled genome from the same or a similar organism that is going to be sequenced. Sequencing a genome with new technology sequencing where a reference genome already exists is called **re-sequencing**.

If you want to do a *de novo* assembly, come and talk to me.

The exercise

In this exercise, we will try to find the gene that is responsible for generating drug resistance to a new compound against severe malaria.

Collaborators generated a new compound to treat severe malaria. Although it is known that the drug kills parasites, the mechanism of the new compound is not understood. To shed light on the function, a parasite line (PfDd2), which quickly generates resistance, was taken and different clones were challenged over half a year with the new compound until they generated resistance.

Then the parent PfDd2, also referred as wild type (WT), and three resistant clones, here called 18, 20 and 23, were sequenced with Illumina, 150bp reads.

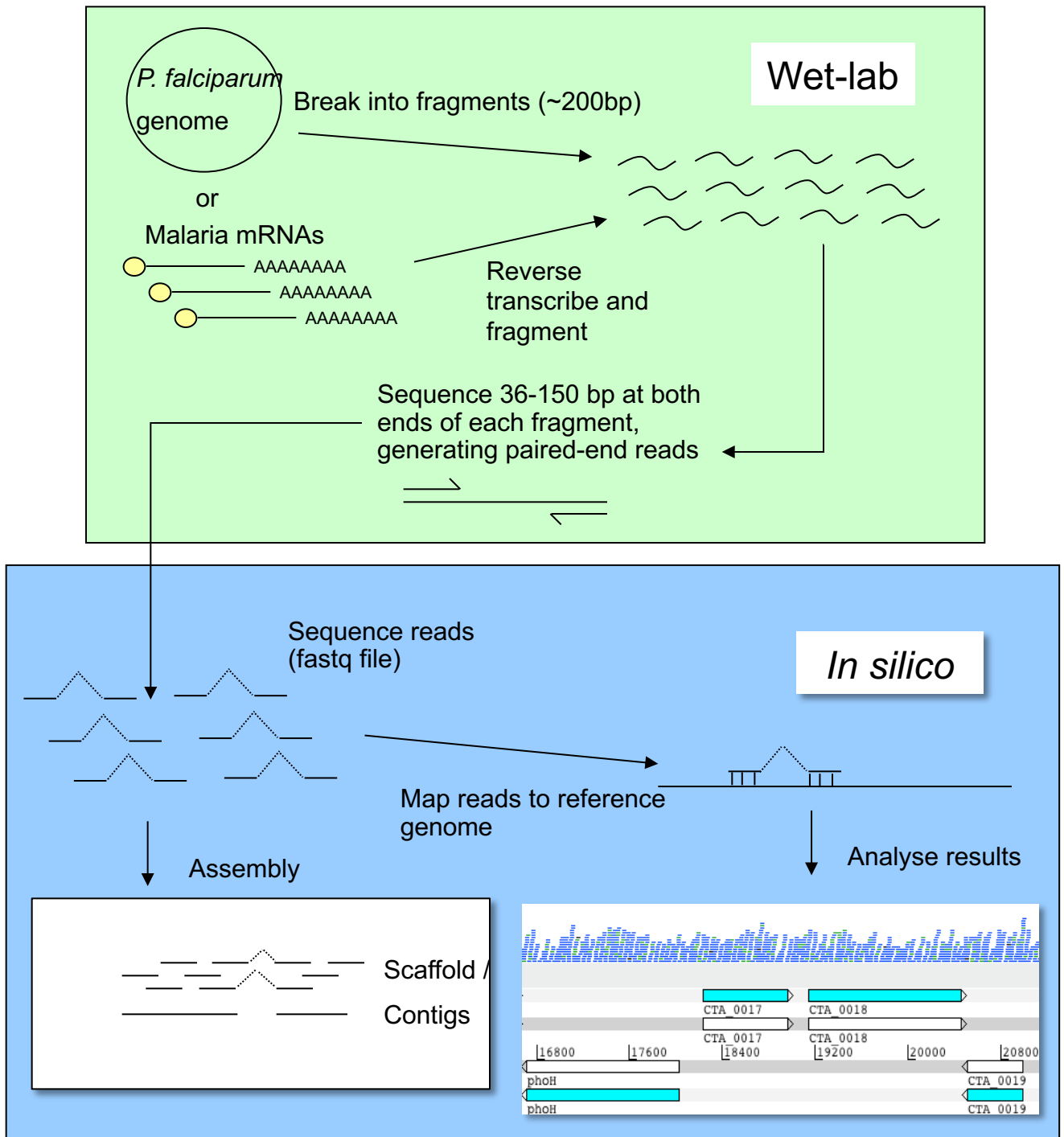
Those reads are then mapped against the *Plasmodium falciparum* reference Pf3D7. The aim is to find differences in the three clones against the WT parasite. If something is unique to the three clones but does not occur in the parent (WT), then it is likely that this mutation explains the model-of-resistance, and maybe also the mode-of-action of the compound. It is important to note that this analysis only generates candidates, which should always be validated in the lab.

Resistance can arise through a number of different types of mutations, including insertions or deletions (indels), point mutations (SNPs) or copy number variation (CNV).

Before we start, some explanations of the methods, the mapping etc are given on the following pages. If you have any doubts, don't hesitate to ask!

Sequencing/Mapping workflow

The diagram below describes the workflows for genomic resequencing and RNA sequencing. For this module, the wet-lab work has been done for you! The blue part gives an overview of the *in silico* (computational) analysis methods, including *de novo* assembly, or mapping. In this exercise we will focus on mapping.



Short-Read Alignment Software

There are multiple short-read alignment programs each with its own strengths, weaknesses, and caveats. Wikipedia has a good list and description of each. Search for “Short-Read Sequence Alignment” if you are interested. We are going to use BWA:

BWA: Burrows-Wheeler Aligner

I quote from <http://bio-bwa.sourceforge.net/> the following:

“BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads..”

Although BWA does not call Single Nucleotide Polymorphisms (SNPs) like some short-read alignment programs, e.g. MAQ, it is thought to be more accurate in what it does do and it outputs alignments in the SAM format which is supported by several generic SNP callers such as SAMtools and GATK.

BWA has a manual that has much more details on the commands we will use. This can be found here: <http://bio-bwa.sourceforge.net/bwa.shtml>. We will use the BWA mem tool described above.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. [PMID: 19451168]

The first thing we are going to do in this Module is to align or map raw sequence read data that is in a standard short-read format (FASTQ) against a reference genome. This will allow us to determine the differences between our sequenced strain and the reference sequence without having to assemble our new sequence data *de novo*.

Biology

To learn about sequence read mapping and the use of Artemis in conjunction with NGS data we will work with real data from the eukaryotic single-celled parasite *Plasmodium* that cause malaria.

Plasmodium falciparum

P. falciparum is the causative agent of **the most dangerous form of malaria in humans**. The reference genome for *P. falciparum* strain 3D7 was determined and published in 2002 (Gardener et al., 2002). Since then the genomes of several other species of *Plasmodium* that infect humans or animals have been elucidated. Malaria is widespread in tropical and subtropical regions, including parts of Asia, Africa, and the Americas. Each year, there are approximately 350–500 million cases of malaria killing more than one million people, the majority of whom are young children in sub-Saharan Africa.

Although several drugs exist to treat malaria, the parasite is acquiring mutations that generate drug resistance, which then subsequently spreads around the world. To overcome this problem many new compounds are tested to kill the parasite. Once a new compound is found, and checked for safety in a mouse model, the mechanism of the drug should be determined. One way to determine that is the generate controlled resistance in several clones from a known drug sensitive background. Once the clones generate mutation that gives them the ability to evade the drug treatment, we sequence the clones and the drug sensitive wild type (WT). This will enable us to find the mode-of-resistance of the parasites, and from that we might be able to determine the mode-of-action of the drug. It is important to have several independent clones to have more statistical power.

For this exercise the WT was the Dd2 parasite, of which we generated a high-quality reference with PacBio. 100 bp Illumina reads of the clones and the WT were generated.

Exercise - motivation

Working with the mapped sequence data and Artemis your aim is to find the mode-of-resistance/mode-of-action of a new compound that is very effective against severe malaria. You will need to find differences (genotypes) in several clones that would explain the new phenotype of drug resistance. Once determined, you can perform further analysis set in the findings context to other similar experiments.

Module Summary

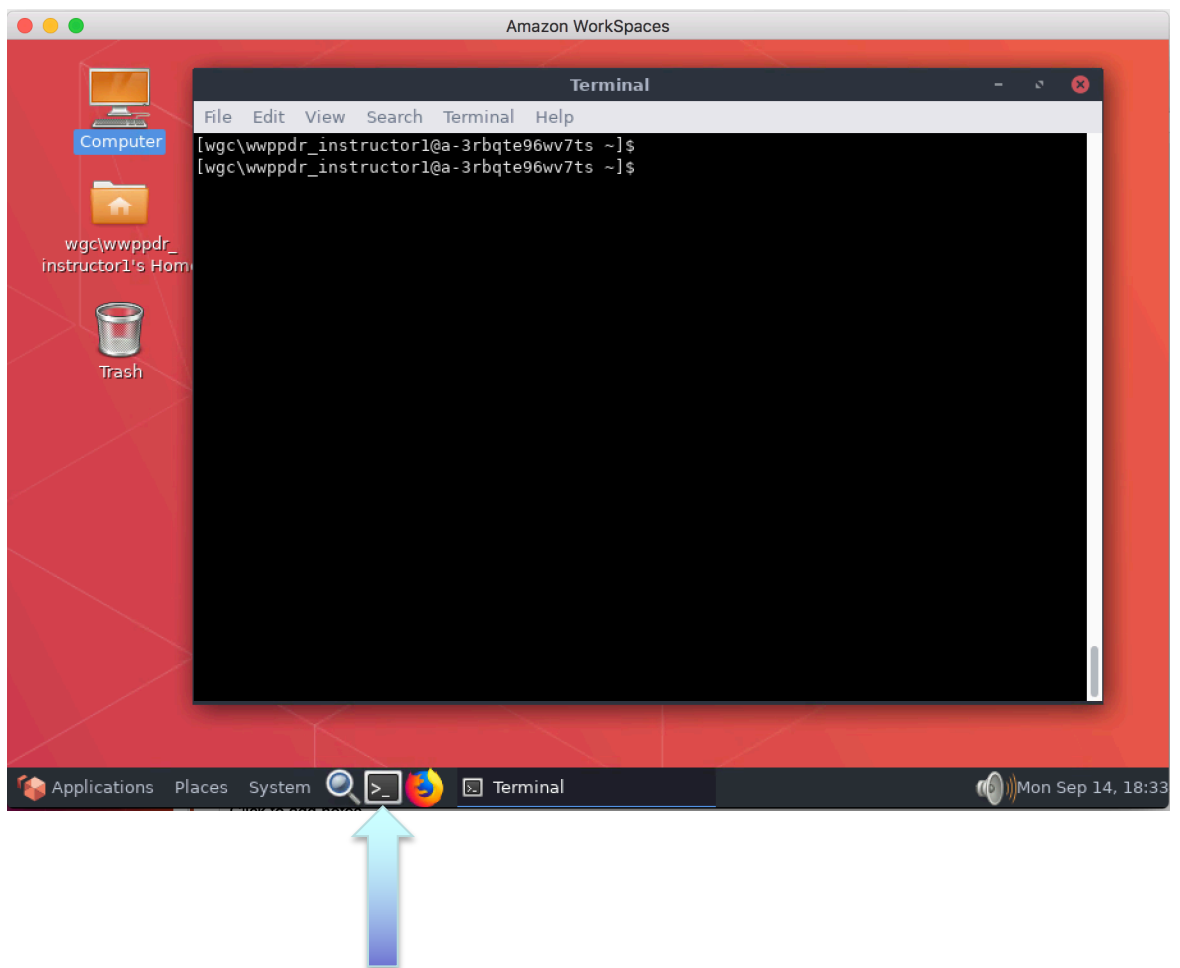
1. Genomic resequencing

- A. File formats
- B. Mapping the data & Converting output to BAM format
- C. Viewing the mapped reads in Artemis
- E. Differences in read coverage between reference and resequenced genomes
- F. Identifying Single Nucleotide Polymorphisms (SNPs) - BCF format

Open a terminal in Linux

Before you will be able to do this exercise, you will need to open a terminal on the Amazon workspace, as described before. Ask for help if needed.

Example of Linux terminal:



1. Genomic resequencing

A. File formats

You have the *P. falciparum* 3D7 clone reference file (Pf3D7_01_v3.fasta). This contains the assembled sequence of the 3D7 genome. You also have two files of sequence reads from the Clone 18 (Reads.18_1.fastq and Reads.18_2.fastq). Look in both the reference file and the read files.

Be sure to be in the directory ~/ Module_3_Mapping. Double check if you are in the correct directory

```
$ pwd
```

Let's first have a look at some file formats... type:

```
$ head -n 5 Pf3D7_01_v3.fasta
```

Compare the format of the reference file above to the format of the fastq files that contain the of sequencing reads:

```
$ head -n 40 Reads.18 1.fastq
```

Each sequence read is represented by four lines

1. @IL39_6014:8:61:7451:18170#3/1
is the sequenced fragment, /1 and /2 refer to
the forward and reverse (paired-end) reads
respectively. IL39 is the sequencing machine.
6014 and 8 indicate the run and lane. #3 says that
the run was multiplexed, and this reads is from
index 3.

2. The read sequence

3. Sequence/quality line separator

4. Sequence quality. There is one character for each nucleotide. The characters relate to a sequence quality score e.g. how likely is the nucleotide correct? ‘>’ is higher quality than ‘6’. Sequence reads tend to have more errors at the end than the start.

Due to time restrictions we will not trim adapters or low quality regions of the reads. You determine the quality of you reads with the program fastqc. Just type fastqc READfile, or see the appendix.

Mapping the reads with BWA mem

Now we will map the clone 18 reads to the 3D7 reference using the short reads mapping program BWA (Li et al, 2009).

First we need to index the reference called **Pf3D7_01_v3.fasta** (this helps BWA to map the reads to the reference more efficiently).

```
$ 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 file is generated. The program is bwa and the program part is “mem”. So type:

```
$ bwa mem
```

You should see the options. The full command with the place holder is

```
$ bwa mem -t 2 reference read_1 read_2 > BWA.18.sam
```

This should have worked quite fast. **IMPORTANT**, you have specified which file the reference, read_1 and read_2 are! -t 2 uses two processors.

Also check if the output file was generated and it is not empty (0 bytes), with

```
$ ls -lta
```

This will list the files in reverse time order. You should also see the files that were generated when you indexed the reference genome. If something is missing, adjust your bwa calls.

The details of where each read has been mapped is now stored in the file BWA.18.sam. We are going to view the mapped reads in Artemis using Artemis BAM view. However the mapping result is not currently in BAM format. To make a BAM file from sam files we need to run a short series of programs.

We can generate a bam file from the sam file with samtools. We also need to sort and index it.

```
$ samtools view -Sb BWA.18.sam | samtools sort - > BWA.18.bam
$ samtools index BWA.18.bam
```

These commands transform the sam file into a binary format, sort the reads by location in the reference and then indexed the bam file by chromosome/contig.

You can also look at some mapping statistics with the following command:

```
$ samtools flagstat BWA.18.bam
```

SAMTOOLS format

D. Viewing the mapped reads in Artemis

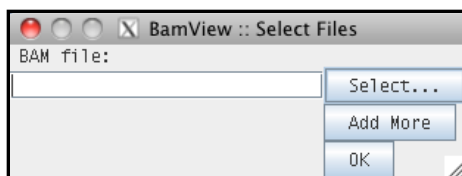
We will now examine the read mapping in Artemis using the BAM view feature.

Open Artemis:

```
$ art &
```

Load the Pf3D7_01_v3.embl file. This contains the same sequence as Pf3D7_01_v3.fasta, but also has genome annotation so we can see the gene models.

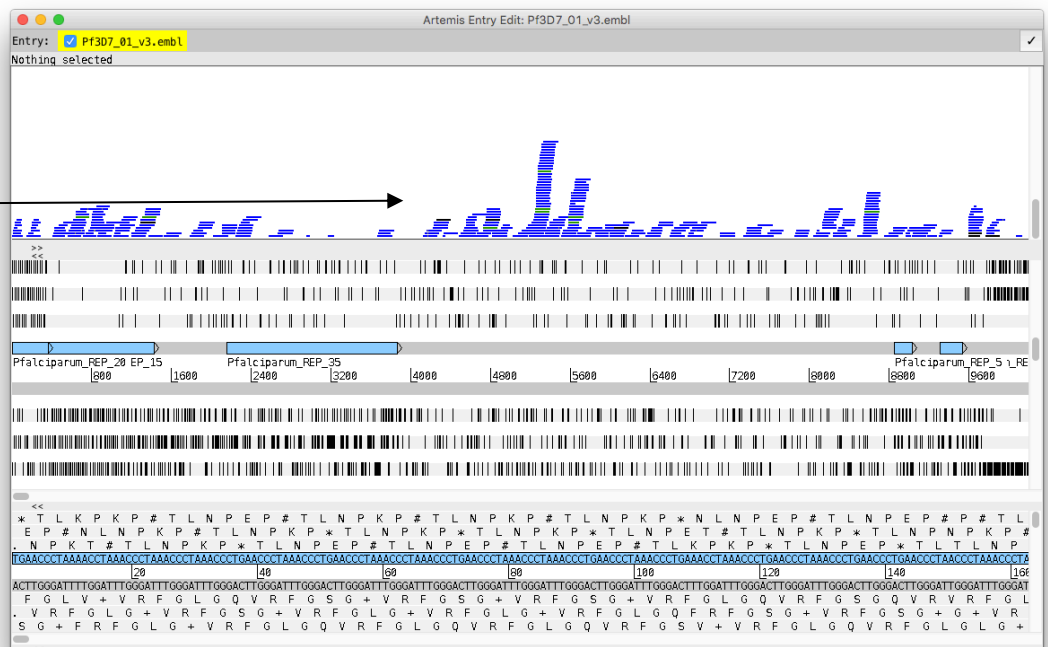
From the Artemis File menu, select 'Read BAM', then locate the file BWA.18.bam from the genomic data directory.



Select BWA.18.bam

You should see the BAM window appear as in the screen shot below. We want to change the view in order to better see how the reads map to the genome.

Right click
here,
Select Graph->
Coverage



Scroll through the genome. Describe the coverage. Are there regions that are not covered?

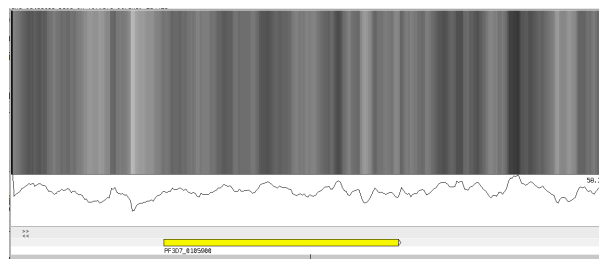
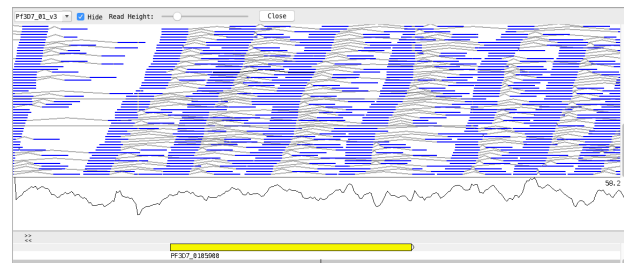
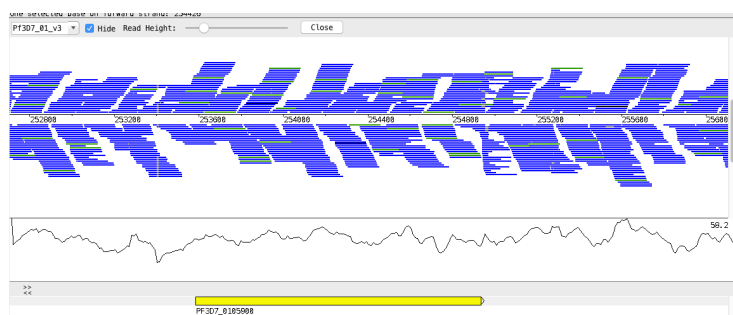
Go to gene PF3D7_0105900 and zoom out once.

Stack view

Coverage plot.
Right click for
more options.

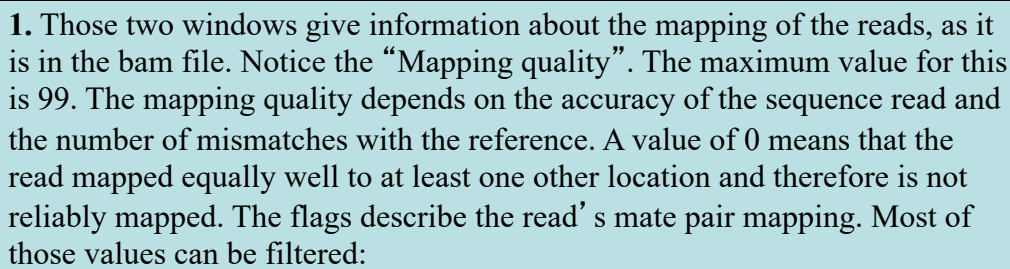
Right click on a read.
Try different views.

Inferred Size
✓ Stack
Paired Stack
Strand Stack
Coverage
Coverage by Strand
Coverage Heat Map
Coverage Options ▶



Each view has its advantages:

- “Inferred Size” (click also Use log scale) displays the mate pair on the y-axis depending their distance. In this case, some reads map further apart compared to others. Could this be a deletion?
- “Strand Stack”: Shows the strand where reads are mapping. Useful for strand specific applications.
- “Paired Stack”: Can be useful to see if two regions are connected.
- “Heatmap”: Can be useful with many different samples!



Reads with less than the mapping quality are not shown. Try 10.

HIDE the proper pairs. What happened?

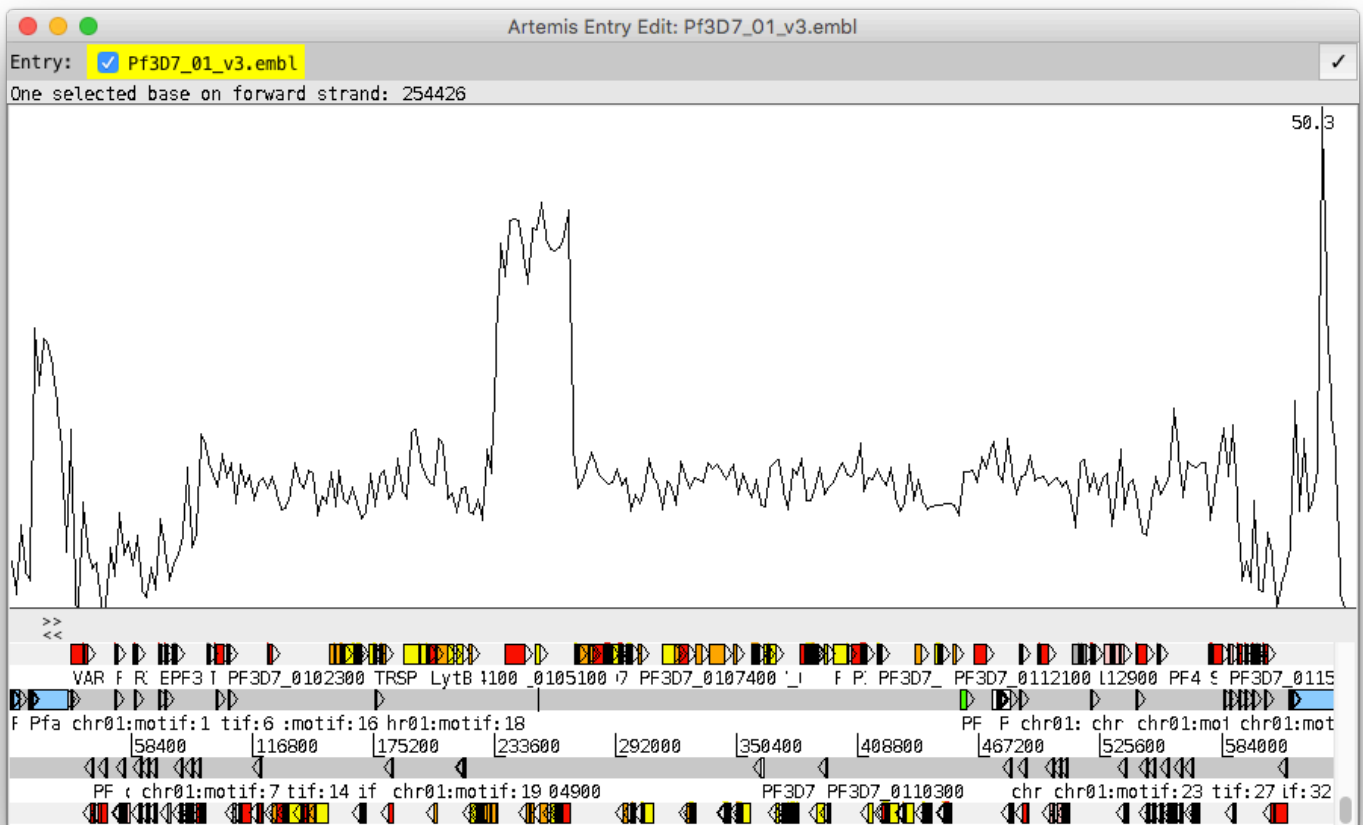
What is the difference between this and Proper Pair HIDE?

Are there any duplicated reads?

Filtering reads for repetitive regions and just see proper paired reads can be a powerful help for analysis.

Now back to biology! The aim of this project is to find the genes responsible for the mode-of-resistance and maybe mode-of-action of a new compound against severe malaria! So far we have looked at one isolate, that turned drug resistant. Did you see anything interesting with this sample?

If not, maybe zoom output and have a look at the coverage plot!



What do you think is interesting here?

And, how can we be sure that this feature is something unique to this clone, and not the PfDd2 isolate, on which the experiments were performed on?

E. Differences in read coverage between reference and resequenced genomes

So yes, this copy number variation (CNV) in the centre of the chromosome is interesting. To exclude that this is something special about PfDd2, we must map the reads of the PfDd2 WT onto the 3D7 reference and compare the coverage.

Use the bwa mem and the samtools command from page 10 and map the reads for the WT/parent against the same reference as before (read_1/ read_2 Reads.WT_1.fastq.gz / Reads.WT_2.fastq.gz) against the same reference as before. Call the output **BWA.WT.sam**.

```
$ bwa mem -t 2 reference read_1 read_2 > outputFile
```

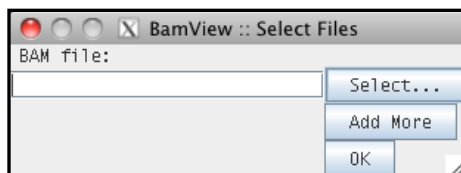
This should have worked quite fast. **IMPORTANT**, you need to specify which files the reference, read_1 and read_2 are!

Now, transform the sam file (outputFile) into a bam file with the samtools view and samtools sort command as before, 6 pages before. Hint: to remember the command, type
\$ h samtools

That will show you all the command from the history that contain the search term samtools ☺

Go now back to Artemis **navigate to the *ATP6* gene locus** using e.g. the Navigator (Goto – Navigator... – Goto Feature With Gene Name). What can you say about the read coverage at this locus? (you may have to zoom out to get a good look).

So far you looked at the **Clone 18**. What about the *atp6* locus in the **Dd2 WT** of the malaria parasite? Right-clicking on the BAMview and choosing ‘Add BAM...’. Select the file BWA.WT.bam



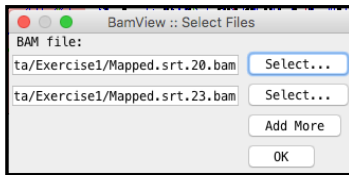
Select BWA.WT.bam

What can you see? Does the drug sensitive parent (wild type - WT) also have this duplication?

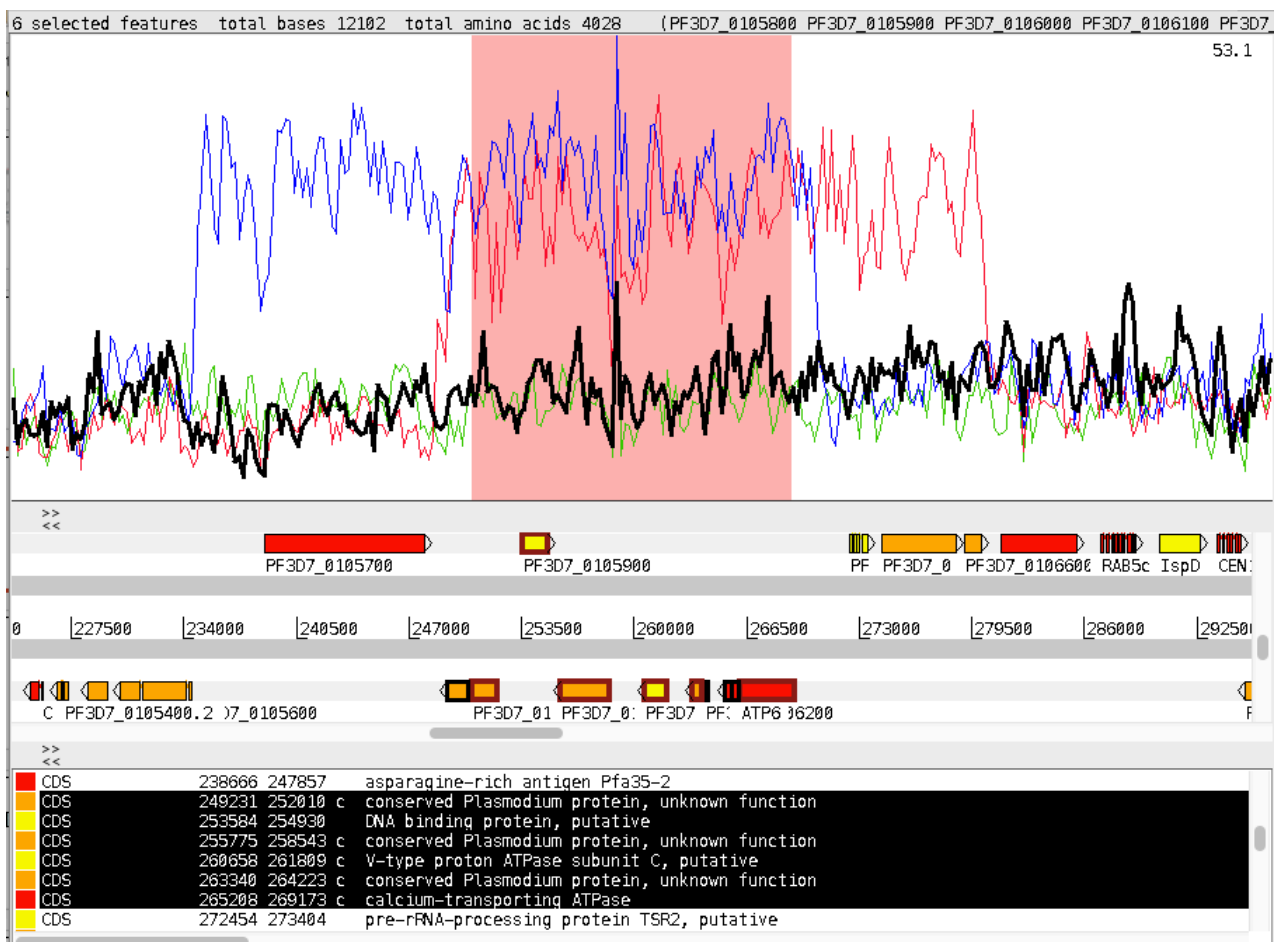
But how big is this duplication (how many genes does it include?)

Does it help us to find the mode-of-action of the new compound?

So let's load in BAM files for the other clones. These are already available in the directory Exercise1/ and they are called Mapped.20.bam and Mapped.23.bam. So as on the page before, do a right-click on the BAMview and choosing 'Add BAM...'. Select those two bam files.



After changing the colours, making the WT black and thicker (right click bamview -> views -> coverage options -> configure lines), filtering for just proper pairs (right click bamview -> Filter reads), this is what you get:

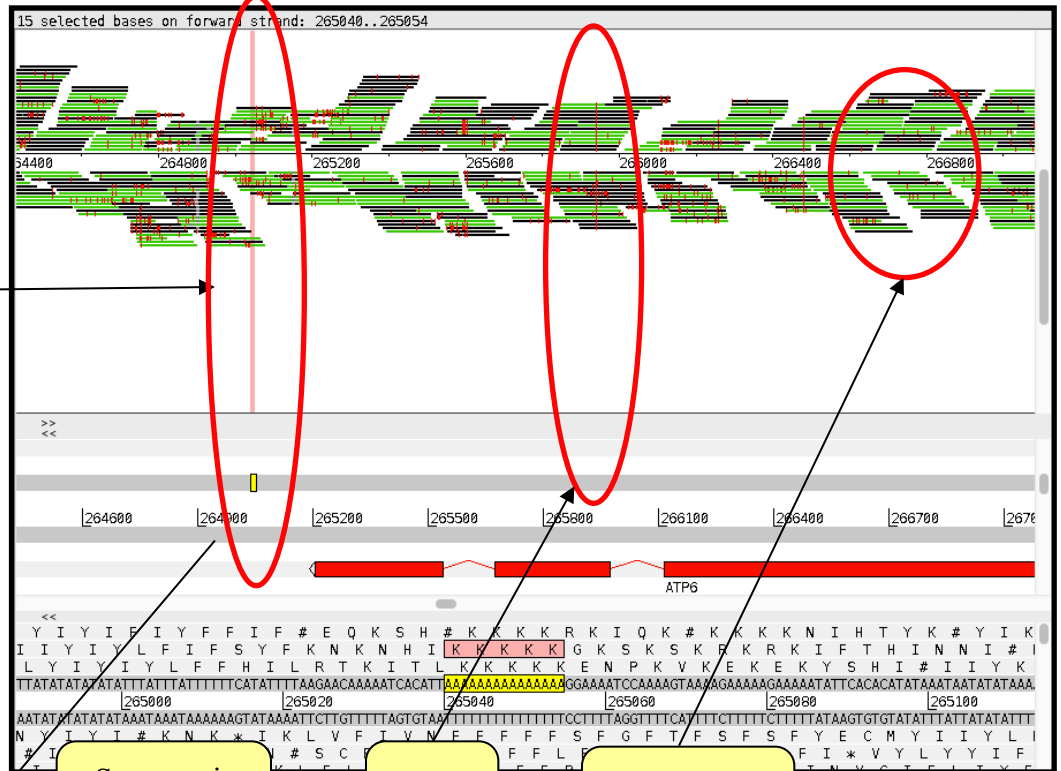


To summarize, we have the WT that is drug sensitive to this new super duper compound (black).

We have three drug resistant clones, blue, red and green, names 18, 20 and 23. Now two seem to have this duplication, over six genes. Any idea which could be an important gene responsible for drug resistance (check the note?).

BUT if the resistance is due to this CNV, what about the green clone? What else could it be?

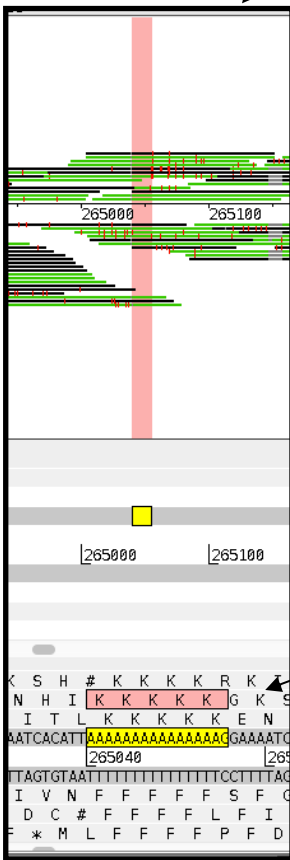
Right click here,
Select Show->
SNPs
&&
Colour by ->
Coverage Plot
Colours..



Systematic error

SNP

Sequencing errors

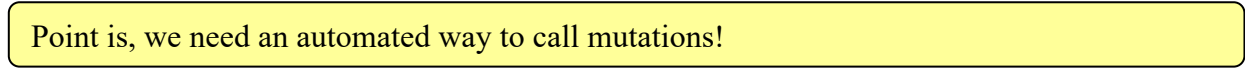


Red marks appear on the stacked reads highlighting every base in a read which does not match the reference. If you zoom in you can distinguish real SNPs as vertical red lines, while the random sequencing errors are more dispersed.

Homopolymer tracts generate systematic errors.

Also GGC motives can generate strand specific errors. Can you find some of those?

Remember, we obtained all the samples from PfDd2. So if a difference is in the WT and the clone, those are just genuine differences between the two isolates Pf3D7 (the reference) and the PfDd2 clone, our parasite we used for the experiment.



Point is, we need an automated way to call mutations!

Calling and analysing SNPs and indels

Obviously, it is not feasible to go through a bam file to look for SNP or indels manually. Fortunately, there are tools to call variants from bam files. The most common is bcftools. Here we are going to explain how to call the variants for the Clone 23 on chromosome 1, starting from the BAM file. There are better ways to call SNPs, like gatk haplotype caller, but due to copyright issues, we cannot present it. But bcftools mpileup will also generate the desired results!

The first step is to generate a pileup of the reads. This is an alignment of the reads over each position, similar to the bam view in Artemis. Now the algorithm knows which bases are covering a given position of the reference, which is given to bcftools call. There, the variant call is performed and stored in a bcf file (*again names in italic are variables and need to be set. Ask if you have doubts!*).

```
$ bcftools mpileup -f reference bam-file | bcftools call -cv -Ov
--ploidy 1 -o clone.23.vcf
$ bgzip clone.23.vcf
$ tabix clone.23.vcf.gz
```

As before, the reference is Pf3D7_01_v3.fasta and the bam file is Exercisel/Mapped.srt.23.bam.

The new file contains all the SNP information. We have also indexed it for visualisation.

To look at the output, do

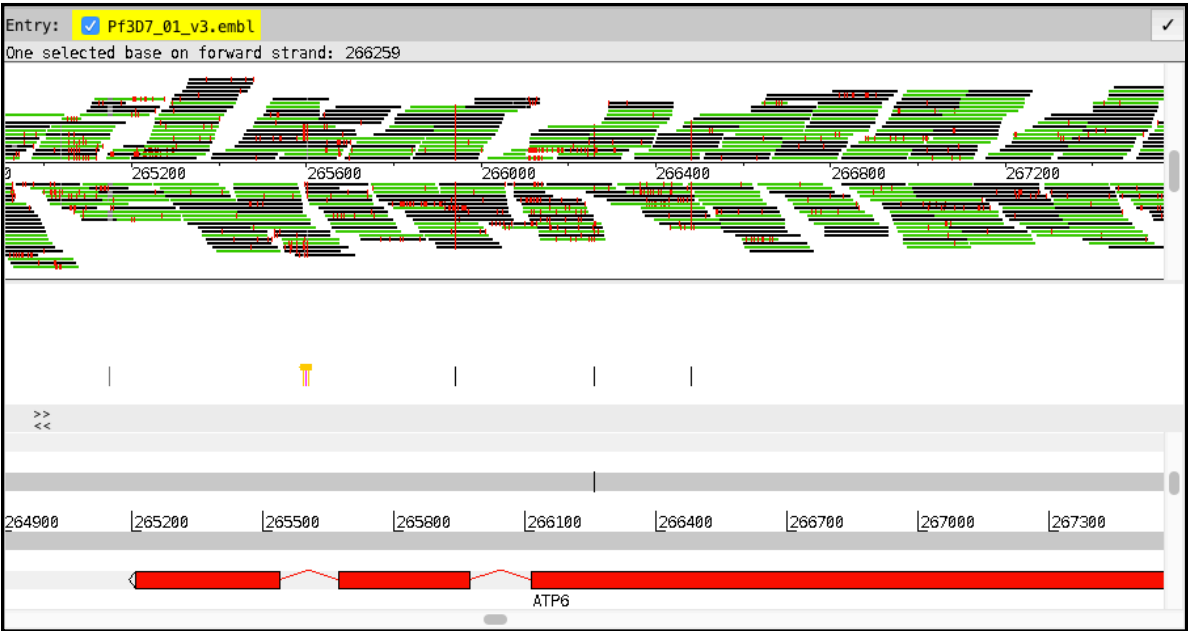
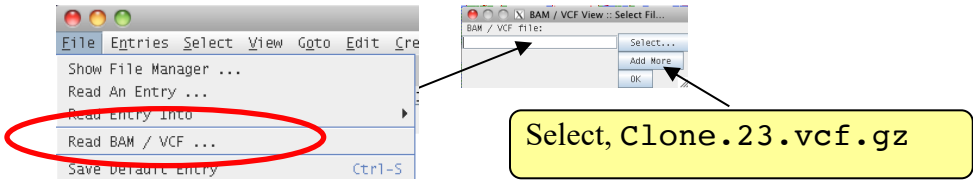
```
$ bcftools view clone.23.vcf.gz | less
```

Exit the less command with “q” for quit.

Position of SNP		Old and new allele		DP - amount of read pair (all quality) DP4 read mapping old allele (forward/reverse), new allele forward /reverse		Quality of SNP		Likelihood which allele is more probable	
Pf3D7_01_v3	266259	A	T	213	DP=10;VDB=1.385735e-01;AF1=1;AC1=2;DP4=0,0,6,4;MQ=60;FQ=-57	GT:PL:GQ	1/1:246,30,0:57		
Pf3D7_01_v3	266480	A	T	198	DP=10;VDB=1.373882e-01;AF1=1;AC1=2;DP4=0,0,4,6;MQ=60;FQ=-57	GT:PL:GQ	1/1:231,30,0:57		

The complete format is described at <http://www.1000genomes.org/node/101>

Go again to the ATP6 gene. Load both BCF files like you included the bam files before. Zoom in and look at all the SNPs and indels in this gene.



Have a look at the different variants in the ATP6 gene. Is it informative at all?

Key to the colours and types of variation shown:

1. Variant (default colour scheme)	
Variant A	Green
Variant G	Blue
Variant T	Black
Variant C	Red
Multiple Alleles	Orange, with circle at top
Insertion	Magenta
Deletion	Grey
Non-variant	Light grey
2. Synonymous/Non-synonymous	
Synonymous SNP	Red
Non-synonymous SNP	Blue
3. Quality Score	
Variants are all on a red colour scale with those with a higher score being darker red.	

Obviously, we need to call SNPs for all the other samples!

This time we are going to use a little script that will generate all the bcf files for us. Type

```
$ bash ./do.SNP.sh &> out.txt &
```

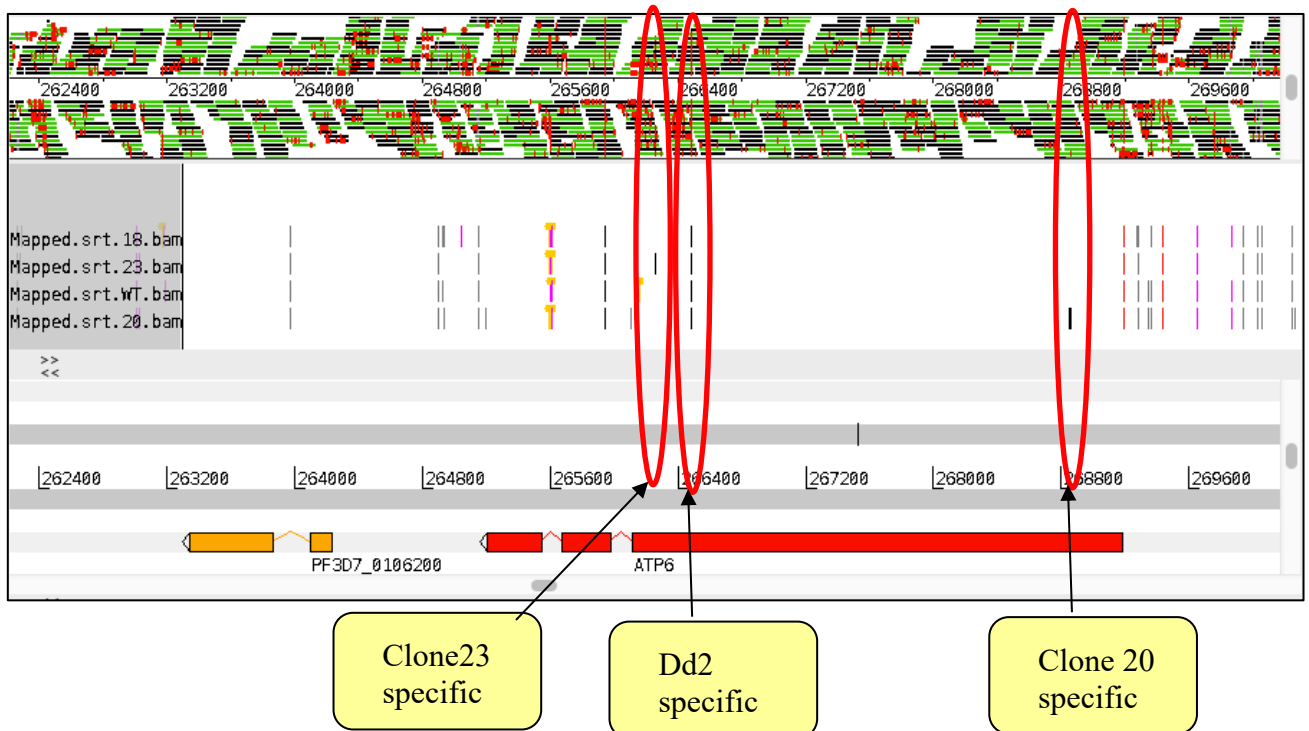
If you are interested, you have a look at the script:

```
$ cat do.SNP.sh
```

This script iterates through the four samples: So the variable x will hold the name of the sample (18, 20, 23 or WT), and then the same command is done four times.

It will take a little bit of time to run.

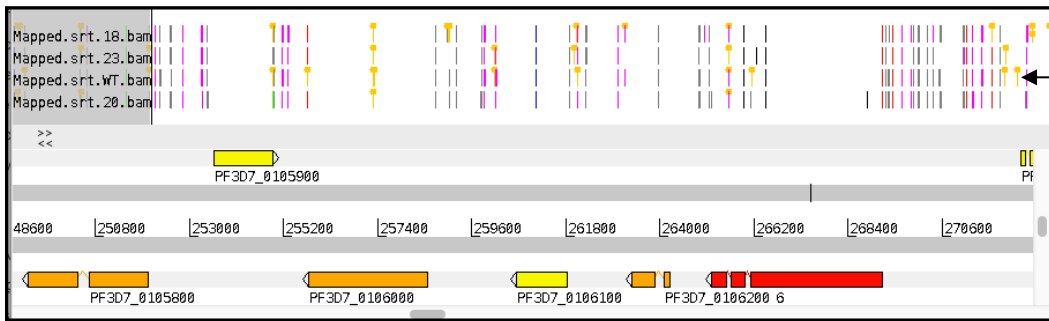
Load the four files into artemis (right click on the SNP view -> add). They are called SNP.*.vcf.gz



So we can already see mutation specific for PfDd2 (not really interesting, right?) And then SNPs specific for the different clones.

Do you find genes that are duplicated in the clones, with private mutations?

But let's do some filtering first



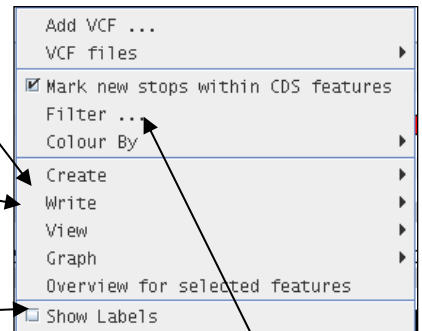
Right click,
and a window
opens

In this menu
are some
useful options:

Write filtered vcf, or fasta
of the genes with the new
alleles

View the sequence, with
SNPs substituted in.

Shows the name of the
variant files



Option to
change the
filters

Able or disable different
types of variants.

Show filter are the same
fields as explained the
page before. Good values
are DP of 10 and Qual of
at least 30

So it seems that ATP6 is the one interesting target. Get the new sequence of the gene (right click on BCF view -> view -> FASTA of selected feature). Click ok on the next window and three windows with fasta sequences will open. What happens if you blast it on PlasmDB?

```
> PF3D7_0106300.1 | gene=PF3D7_0106300 | organism=Plasmodium falciparum_3D7
| gene_product=calcium-transporting ATPase | transcript_product=calcium-transporting
ATPase | location=PF3D7_01_v3:265208-269173(-)
| length=3687 | sequence_SO=chromosome
SO=protein_coding
Length=3687

Score = 5613 bits (6224), Expect = 0.0
Identities = 3117/3120 (99%), Gaps = 0/3120 (0%)
Strand=Plus/Plus

Query 1   ATGGAAGAGGTTATTAAGAATGCTCATACATACGATGTTGAGGATGTACTAAAAATTTTG 60
Sbjct 1   ATGGAAGAGGTTATTAAGAATGCTCATACATACGATGTTGAGGATGTACTAAAAATTTTG 60

Query 61  GATGTAACAAAGATAATGGTTTAAAGAATGAGGAATTGGATGATAGAAGATTAAATAT 120
Sbjct 61  GATGTAACAAAGATAATGGTTTAAAGAATGAGGAATTGGATGATAGAAGATTAAATAT 120
```

Functional information

By now we determined the amount of mutations in field isolates and know that our candidate gene is a calcium transporter. But do we know when it is expressed? Do we know if it is already a drug target? Does it interact with other genes? Where do the mutations sit?

Here we want you to explore two webpages, www.genedb.org and www.plasmodb.org.

In GeneDB put the gene ID in the search box. This will open the gene page. Exploring the page we can find publications associated to this gene and also check if the mutation falls into known PFAM domains or transmembrane proteins. The latter would give you an idea of the impact of the mutation versus the function.

The screenshot shows the GeneDB homepage with the search bar containing 'PF3D7_0106300'. The 'Protein Data' section is active, displaying a table of matches for the gene. The table includes columns for 'Other Matches', 'Position', 'Score', and 'Significance'. The matches listed are Pfam:PF00690.22, Pfam:PF00689.17, Pfam:PF13246.2, Pfam:PF00122.16, and Pfam:PF13246. The matches are associated with Cation_ATPase_N, Cation_ATPase_C, Cation_ATPase, E1-E2_ATPase, and Putative hydrolase of sodium-potassium ATPase alpha subunit.

Protein Data

Protein Map | Domain Information Table | Predicted Peptide Data | Algorithmic Predictions

Other Matches	Position	Score	Significance	
matches:				
Pfam:PF00690.22	Cation_ATPase_N	8 - 76	65.2	2.8e-18
Pfam:PF00689.17	Cation_ATPase_C	997 - 1210	171.9	9.4e-51
Pfam:PF13246.2	Cation_ATPase	517 - 729	56.4	2.4e-15
Pfam:PF00122.16	E1-E2_ATPase	98 - 347	207.8	1.2e-61
Pfam:PF13246	Putative hydrolase of sodium-potassium ATPase alpha subunit	517 - 728	8.7e-17	

Comments

» gene has a putative role in resistance to Artemisinin (PMID:16325698, PMID:12931192)

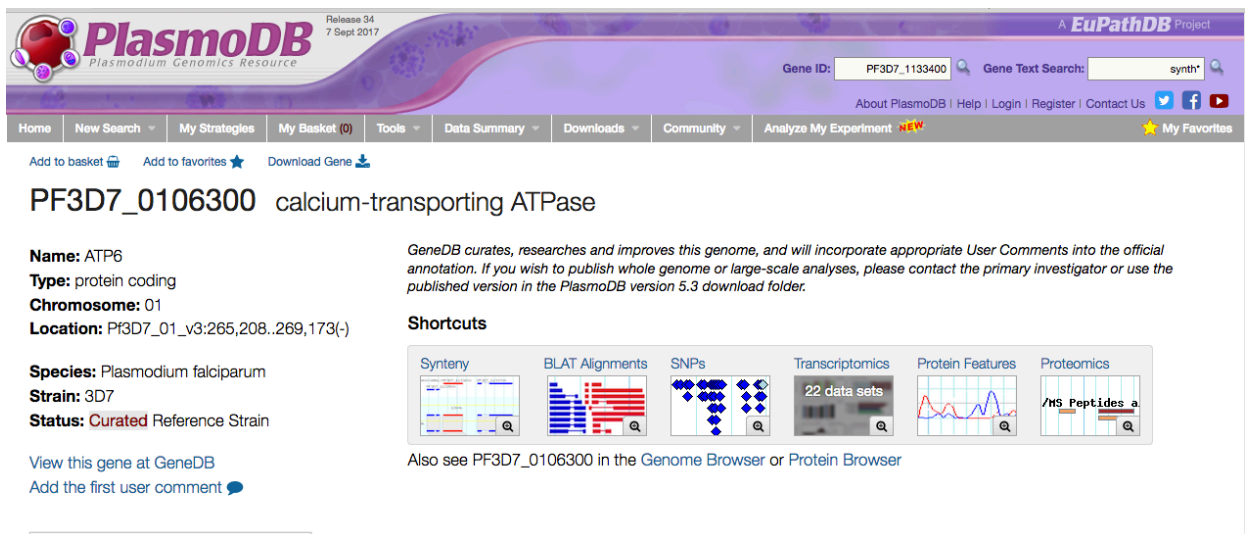
Key information on this gene is available from [PMID:21599655](#) [PMID:20195531](#) [PMID:20461426](#) [PMID:27471101](#)

Where are the mutation of the Clone 20 and Clone 23? Are they in a specific domain?

Functional information

Open the gene page in PlasmoDB. Can you learn anything more? When is the gene expressed? Does it make sense in terms of resistance? Are there known SNPs?

Do you think that this gene is also involved in drug resistance for Artemisinin.



PlasmoDB Plasmodium Genomics Resource Release 34
7 Sept 2017 A **EuPathDB** Project

Gene ID: Gene Text Search:

[Home](#) [New Search](#) [My Strategies](#) [My Basket \(0\)](#) [Tools](#) [Data Summary](#) [Downloads](#) [Community](#) [Analyze My Experiment](#) [About PlasmoDB](#) [Help](#) [Login](#) [Register](#) [Contact Us](#) [Twitter](#) [Facebook](#) [YouTube](#) [My Favorites](#)

[Add to basket](#) [Add to favorites](#) [Download Gene](#)

PF3D7_0106300 calcium-transporting ATPase

Name: ATP6
Type: protein coding
Chromosome: 01
Location: PF3D7_01_v3:265,208..269,173(-)

Species: Plasmodium falciparum
Strain: 3D7
Status: Curated Reference Strain

[View this gene at GeneDB](#)
[Add the first user comment](#)

GeneDB curates, researches and improves this genome, and will incorporate appropriate User Comments into the official annotation. If you wish to publish whole genome or large-scale analyses, please contact the primary investigator or use the published version in the PlasmoDB version 5.3 download folder.

Shortcuts

[Synteny](#)
[BLAT Alignments](#)
[SNPs](#)
[Transcriptomics](#)
[Protein Features](#)
[Proteomics](#)

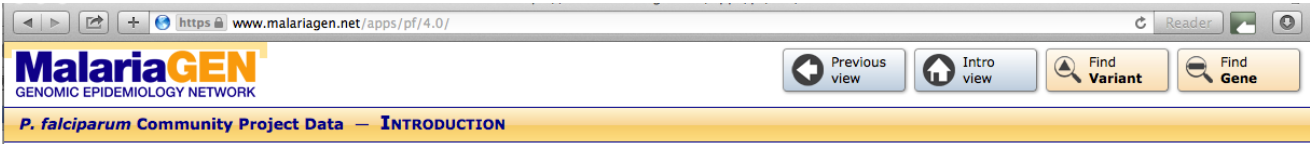
Also see PF3D7_0106300 in the [Genome Browser](#) or [Protein Browser](#)

Where are the mutations from Clone 20 and Clone 23? Are they in a specific domain?

Panoptes

Now we found mutations in the clones that might explain the drug resistance. But would it be a powerful new compound if those mutations that generate resistance already exist in *Plasmodium* field isolates? We are now going to search in the Panoptes database that contains over 5000 *Plasmodium* field isolates, if that gene has many mutations and is the exact mutations we found in our clones, were already found.

Open a web browser and go to <https://www.malariagen.net/apps/pf/4.0/>.



Gene PF3D7_0106300

Names: 124505761;1351996;23510637;301599275;301599277;301599279;301599281;301599283;301599285;301599289;301599293;301599295;301599297;301599299;301599301;301599303;301599305;301599307;301599309;301599311;301

Description: calcium-transporting ATPase (ATP6)

Position: Pf3D7_01_v3:265208-269173

Show list of variants

Show position on genome

Find in **GeneDb** Find in **PlasmoDB**

Click on find a gene and insert the ID of the calcium transporter. Show the variants...

Position	NRAFWAF	NRAFCAF	NRAFEAF	NRAFSAS	NRAFWSEA	NRAFESEA	NRAFOCE	NRAFSAM	MAFGlobal	Fst	Amino acid	Type
1:267878	0	0.003	0	0	0.000	0	0	0	0.000	0.003	E432D	Non-syn
1:267881	0.001	0	0.002	0	0	0	0	0	0.000	0.001	431E	Syn
1:267882	0	0	0	0.007	0	0	0	0	0.001	0.007	E431G	Non-syn
1:267883	0.118	0.134	0.245	0.107	0.008	0.018	0.007	0	0.080	0.081	E431K	Non-syn
1:267884	0	0.002	0	0	0	0	0	0	0.000	0.002	430G	Syn
1:267894	0.000	0.001	0	0	0	0.000	0	0	0.000	0.000	T427K	Non-syn

Are those two mutations new in the field or are they already reported?

Is this database useful for your research?

Summary

Uff, that was a lot of work!

But you did the analysis that was central to find the mode-of- resistance of a new compound!

At the same time you learnt how to map reads, call mutations, start a script and analysed the data. You looked at the data in Artemis and finally did some functional analysis!

You are very close to being a bioinformacian! If you want to learn more about the compound and target, this work is part of following publication:

SC83288 is a clinical development candidate for the treatment of severe malaria. Nature Communitation -

<https://www.ncbi.nlm.nih.gov/pubmed/28139658>

The pdf is in the Module directory, called paper.pdf – have fun reading it and remember, you replicated the bioinformatics analysis!

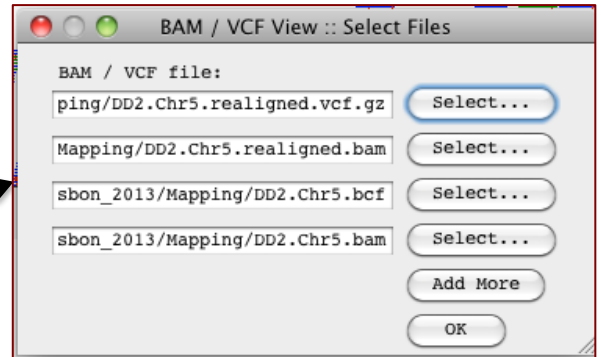
H. Realignment

When analysing the BAM files and the variant calls, you might have noticed heterozygous SNPs. This is very unlikely in a haploid genome. This could indicate a collapsed repeat, CNV or improperly aligned reads.

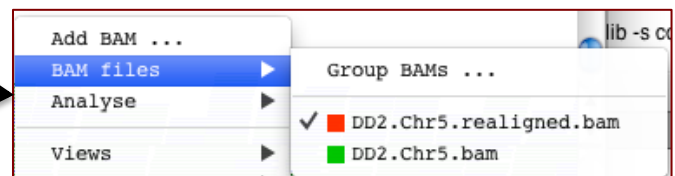
To improve the alignment of reads, we are going to compare the DD2 bam file with a realigned bam file (generated with GATK) to understand the "realignment" process.

1. Reopen chromosome 5 of *P. falciparum* (Pf3D7_05_v3.embl) in Artemis

2. Load the following two BAM and BCF files (file -> Read BAM/ VCF).
Directory ~/Module_3_Mapping/Exercise2.

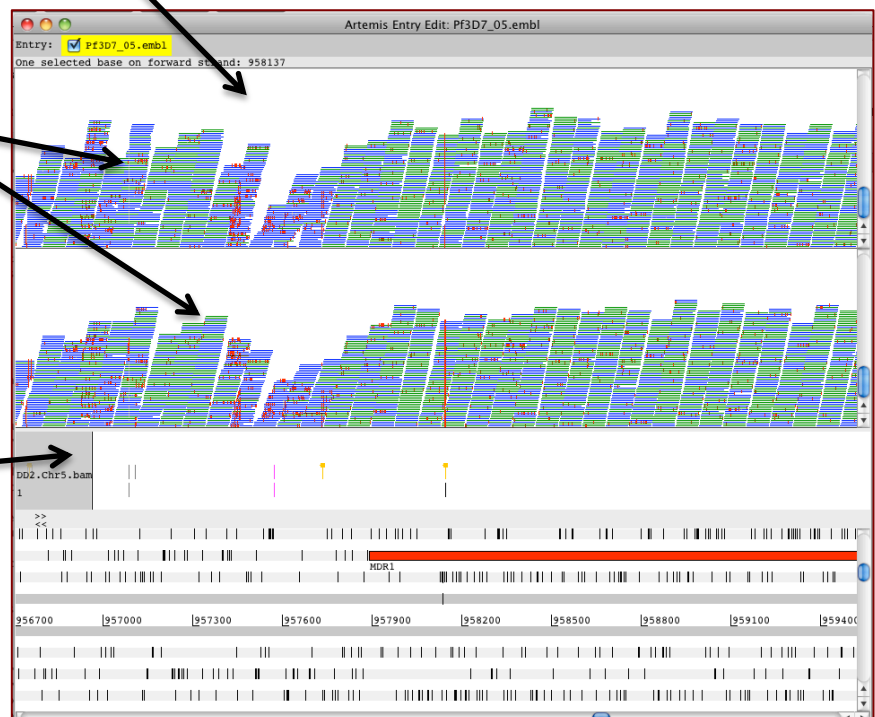


3. Clone the BAMview, so that on top is the DD2.Chr5.bam, and below is the file re-aligned BAMfile.



4. Show the SNPs (show -> SNP mark) in BAMview

5. Show the labels in the BCF view (right click -> show labels).



Can you find differences between the realigned reads and the SNP call between the two versions?

Zoom in here.

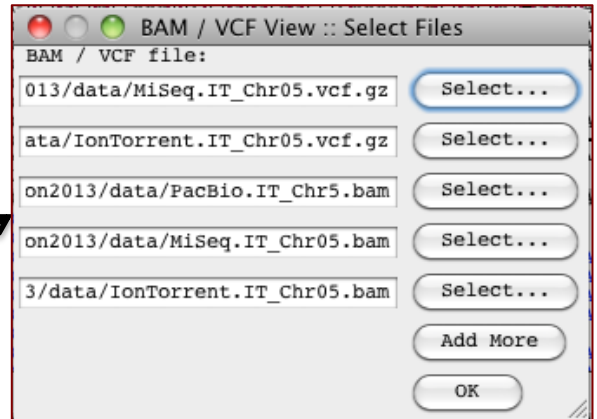
G. Comparing different sequencing technologies

So far we only used data from the Illumina platform. Here we include reads from the Ion Torrent and PacBio platform. The reads are from the 3D7 genome and were mapped against the chromosome 5 of the IT clone. The data are in ~/Module_3_Mapping/Exercise3. Try to:

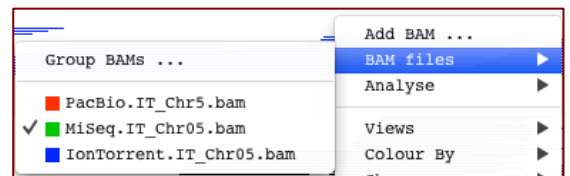
1. Look at coverage variation in each technology. Is there a correlation with the GC content?
2. Which technology has the longest reads, and which has the most accurate reads?
3. Which technology is better for SNP calling?

1. Open the chromosome 5 of *P. falciparum* IT clone (PfIT_05_v3.embl) from the Exercise3 directory in Artemis.

2. Load the following BAM and BCF files (file -> Read BAM/VCF).

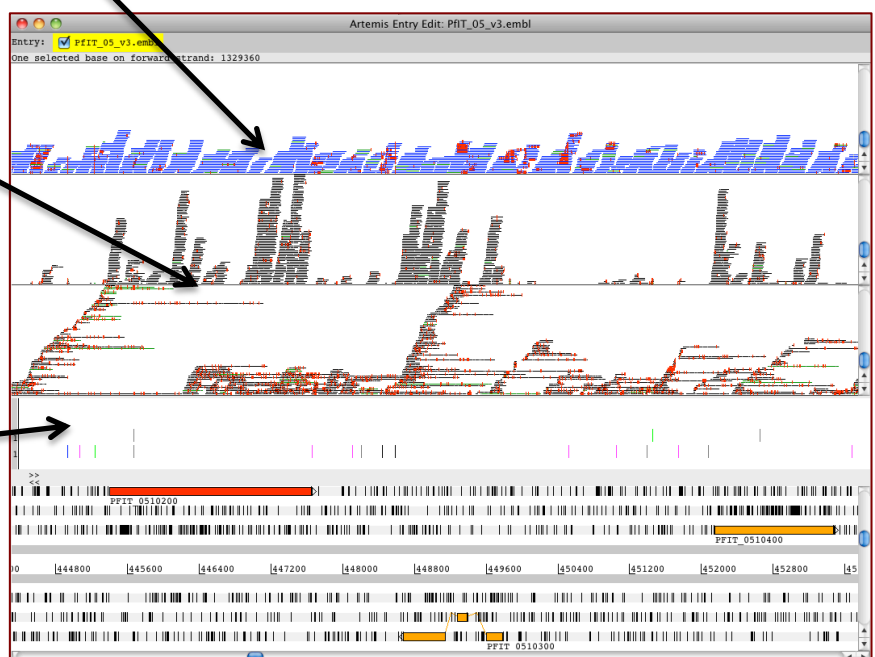


3. Clone the BAMview three times as done before.



4. Show the SNPs (show -> SNP mark) in BAMview

5. Show the labels in the BCF view. Can you guess which lane contains variant calls from the Ion Torrent platform?



Important aspects of the mapping procedure

Non-unique/repeat regions

A sequence read may map equally well to multiple locations in the reference genome. In such cases it is unclear where the read should be placed. Different mapping algorithms have different strategies for this problem.

GC content

Some organisms have genomes with extreme GC content. The *Plasmodium* genome, for instance, is 19% GC, meaning 81% of bases are A or T. The result of this is that reads are more likely to map by chance to multiple locations in the genome than in a genome with neutral GC content (e.g. 40-60% GC).

Insert size

When mapping paired reads, the mapping algorithm (e.g. BWA) takes the expected insert (e.g. sequenced DNA fragment) size into account. If the fragments are expected to be, on average, 200bp and the sequence reads are 50bp, then the paired reads should be ~100bp apart. If the paired reads are significantly further apart then we can say that the reads do not map reliably and discard them. This information can help to produce a more reliable mapping.

Tips

- It is always a good idea to try different programs for any particular problem in computational biology. If they all produce the same answer you can be more certain it is correct.
- Alternative short read mappers include SOAP (Li et al., 2008b), Ssaha (Ning et al., 2001), MAQ (Li et al., 2008) and Bowtie2 (Langmead et al., 2009). As seen, Hisat2 (Kim et al., 2019) is particularly useful for RNAseq mapping as it supports spliced mapping.
- New tools for mapping sequence reads are continually being developed. This reflects improvements in mapping technology but it is also due to changes in the sequence data to be mapped. The sequencing machines we are using now (e.g. Illumina Genome Analyzer II, 454 GS FLX etc) will not be the ones we are using in a few years time and the data the new machines produce will not be best mapped with current tools.
- For SNP calling, especially with many samples, GATK is a very good option. This includes merging BAM files with different read tags, doing the re-alignment etc, as shown during the introduction talk and Section H.

Optional: Understanding base quality and the ASCII code

Each base in the FASTQ file has an associated quality score Q , which reflects the probability p that the base is incorrect. The formula to get Q is $Q = -10 \log_{10} p$. The values of Q can range from 0 to 93. To save space in a FASTQ file, each quality score is transformed into a single ASCII character.

To understand this a bit better, let's transform the quality "I" into the probability that the base is wrong. First search the internet for "ASCII table" and get the decimal value for the ASCII character I.

It is 73. The convention is to subtract 33, which makes 40. Is this a good quality? We change the formal $Q = -10 \log_{10} p$ to $p = 10^{(Q/-10)}$.

With $Q = 40$, p is 0.0001. So there is a chance of 0.01 % that the base is wrong – pretty good!

Could you please transform following base qualities:

Quality in fastq	Q in decimal	p
I	40	0.0001
		0.1
	23	
:		
#		

For more information, have a look at http://en.wikipedia.org/wiki/FASTQ_format.

There should be also spend some attention to quality 2, which has following meaning (Illumina manual page 30): *If a read ends with a segment of mostly low quality (Q15 or below), then all of the quality values in the segment are replaced with a value of 2 ... This Q2 indicator does not predict a specific error rate, but rather indicates that a specific final portion of the read should not be used in further analyses.*

For those who would like to look into PERL, you can use following command on the command line to get the value of "#":

```
$ perl -e 'print (ord("#")-33)'
```

```
$ perl -e '$x=40;print (10**($x/-10))'
```

Will give you the p value for Q 40.

Run fastqc

After this nice introduction, let's run fastqc to check the quality of the reads. First you will need to change to the correct directory with the cd command.

```
$ cd ~/Module_3_Mapping
$ fastqc Reads.18_2.fastq
```

Use firefox to look at the data.

```
$ firefox Reads.18_2_fastqc.html
```

Use firefox to look at the data. That looks pretty good, all green!

If the quality is not so good, you can trim poor-quality bases from the 3' end with sickle. In this case there is no need to do this, but the command is below for information.

```
$ sickle se -t sanger -f Reads.18_2.fastq -o Reads.18_2_trim.fastq -q 20
```

For this to work you need to install the program with

```
$ sudo apt install sickle
```

And type your password (Glasgow2020) You could rerun fastqc to see differences

```
$ fastqc Reads.18_2_trim.fastq
$ firefox Reads.18_2_trim_fastqc.html
```

Can you see that the overall quality in the boxplot went up?

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

✓ Per base sequence quality

