Mapping Sequence Data

Introduction

Improvements in DNA sequencing technology have lead 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 HiSeq400 can to 12 in 1.5 days… The data from the Illumina machine comes as relatively short stretches of 35-150 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 are much slower and more expensive.

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. Even better are the 3rd generation technologies like Pacific BioScience or Oxford Nano Pore. The results of 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 in at least 50 pieces and for Eukaryotes the assembly is in more than 1000 pieces. Therefore new 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 compounds against severe malaria.

Collaborators generated a new compound that help against severe malaria. Although it is known that the drug is killing parasites, the mechanism of the new compound is not understood. To shed light on the function, a parasite line (PfDd2), which generates quickly resistance, was taken and different clones where 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.

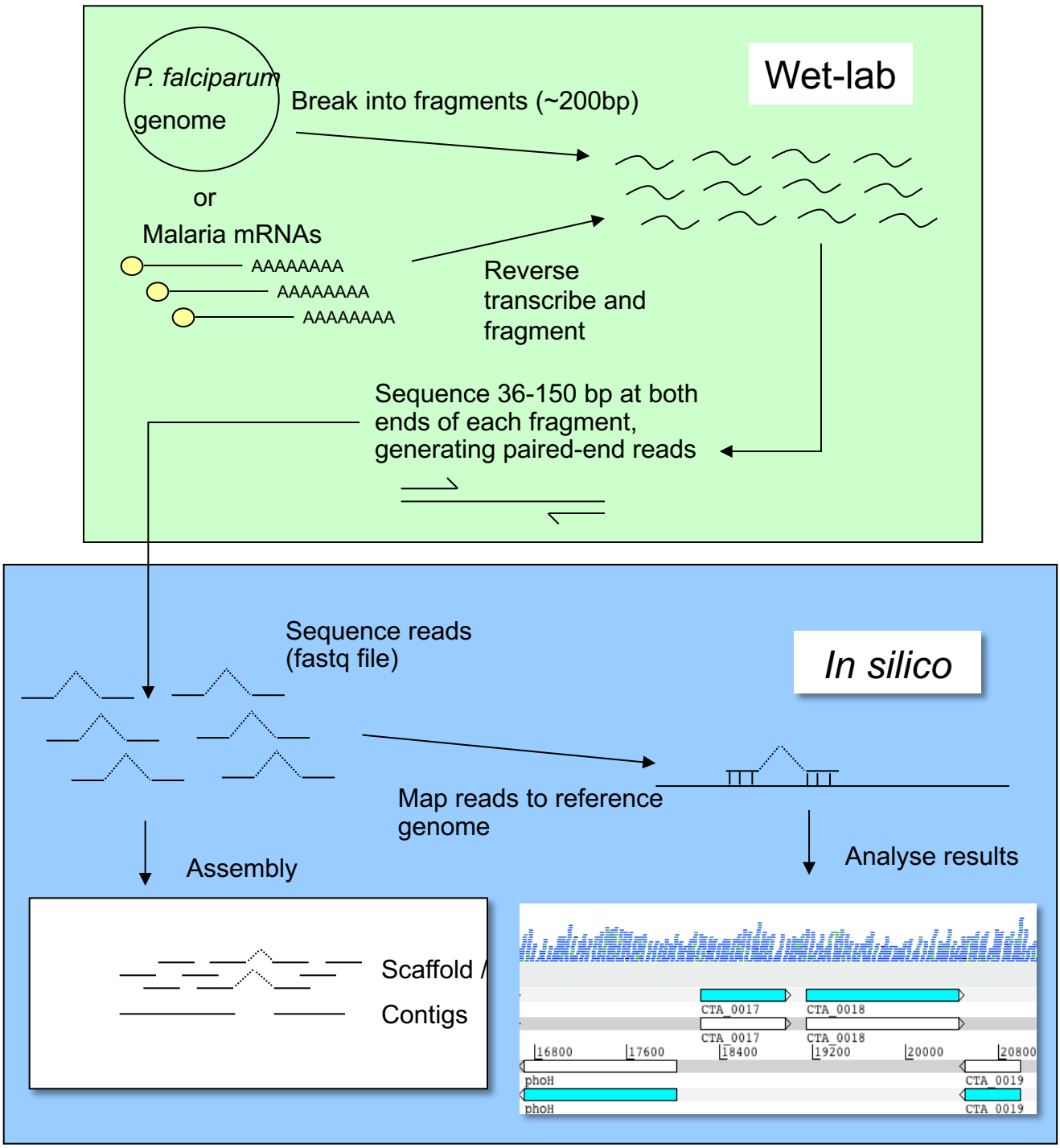
Those reads are then mapped against the *Plasmodium falciparum* reference Pf3D7. The aim is to find differences in the three clones against the WT parasite. If something is unique to the three clones, but does not occur in the parent (WT), then it is likely that this mutation explains the model-of-resistance, and maybe also the mode-of-action of the compound. Important is that this bioinformatics analysis just generates candidates. To proof it, the wet lab is needed again!

The different can be insertion or deletions (indels), point mutations (SNPs) or copy number variation (CNV).

Before we start, some explanations of the methods, the mapping etc are given, see the next pages… if you have any doubts, don’t hesistate 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 what can be done *in silico* (computational) with those reads, *de novo* assembly, or mapping. In this exercise we will focus on the latter.



Resequencing

When resequencing, instead of assembling the reads to produce a new genome sequence and then comparing the two genome sequences, we map the new sequence data to the reference genome. We can then identify **Single Nucleotide Polymorphisms** (SNPs), insertions and deletions (indels) and Copy Number Variants (CNVs) between two similar organisms.

Important is to understand the format of the different data types, see. Below, are examples of four format: fasta, fastq (fasta + quality), sam and vcf. Those files have well defined formats. For example the fastq file contains the sequencing reads with quality values. For each read entry, the first line starts with "@" followed by the read name. The next line(s) is the bases of the sequence. The next line is a "+" indicating that next the line has the quality values of each base. So computer programs will always expect that format, or will crash.

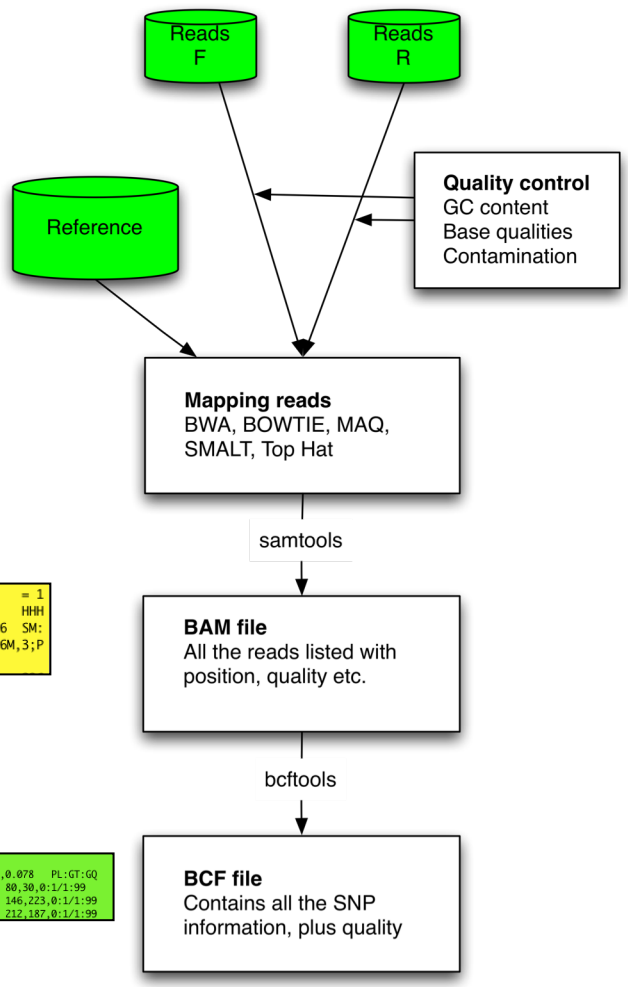
The workflow shows the broad idea of mapping. The paired end reads (in fastq format; F – forward; R - reverse) are mapped against the reference (fasta format). Different tools can be used for that. The results (sam format) can be transformed with samtools to an ordered and index bam file(bam is the binary format of sam). Those bam files can read into programs like Artemis to visualize the alignments of the reads. From the bam file it is possible to call variants. The output format is BCF or VCF. VCF can be loaded easily into excel like tools.

Example of sequence in fastq format

```
@IL39_6014:8:75:16399:10217#12/1
CCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAACCTAAACCCTAAACCCTAAATCCTAAACCCTGAAA
+
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
ctaaacctaaacctaaaccctgaaccctaaaccctaaaccctgaaccctaaaccctgaac
cctgaaccctaaaccctgaaccctaaaccctgaaccctgaaccctaaaccctaaaccctа
aaccctaaaccctaaaccctaaaccctgaaccctaaaccctgaaccctaaaccctaaacc
ctaaaccctaaaccctaaaccctgaaccctaaaccctgaaccctgaaccctaaaacctaa
```



The mapping tools generate sam/bam files

```
IL39_6014:8:45:11459:13097#6    99      Pf3D7_05        1       14      2S9M1D23M1D34M1D8M      = 1
70      246     CCCTAAACCCTAAACCTAAACCCTGAACCCTAAACCTAAACCCTGAACCCTAAACCCTAAACCCTAAACCTAAACC    HHH
HHHHHHHHHHHHHHHHGHHHHHGHHGGGHFHGG=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^CZ4G9^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=20;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=20;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        146,223,0:1/1:99
MAL1    57527   .       C       G       179     .       DP=63;AF1=1;CI95=1,1;DP4=0,0,1,61;MQ=28;FQ=-214 PL:GT:GQ        212,187,0:1/1:99
```

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:

"*Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and BWA-SW.*
The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates."

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 are using a newer version, bwa mem

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*.

The FASTQ sequence format is shown over-page.

Biology

To learn about sequence read mapping and the use of Artemis in conjunction with NGS data we will work with real data the eukaryotic single-celled parasites *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 mutation 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 a the opportunity 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. Important is 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. Illumina reads of the clones and the WT were generated, 100bp.

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 sever 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 it the findings context to other similar experiments.

Module Summary

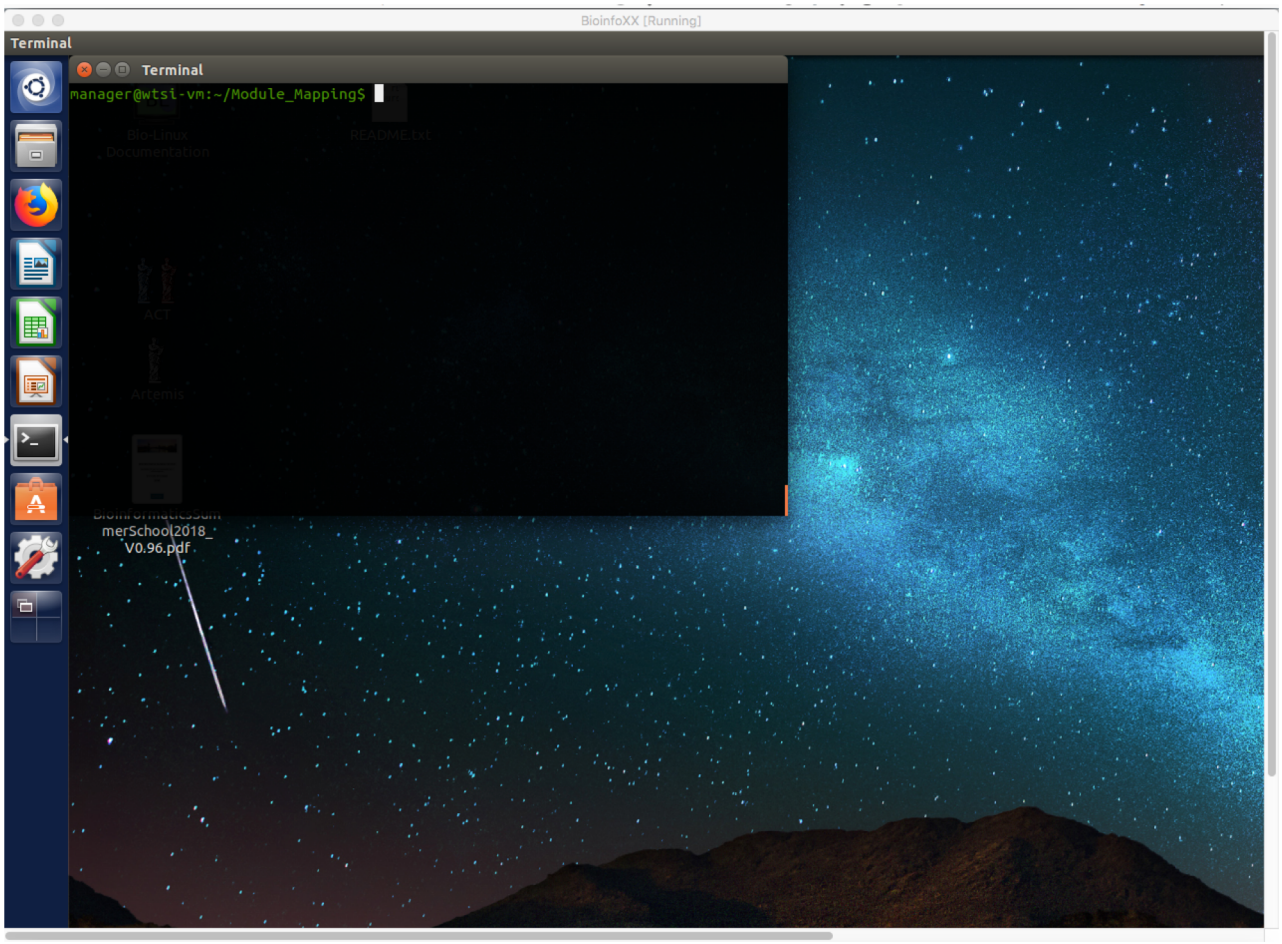
1. Genomic resequencing

- 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 VM machine, 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.

First change to the directory ~/Module_3_Mapping. Double check if you are in the correct directory

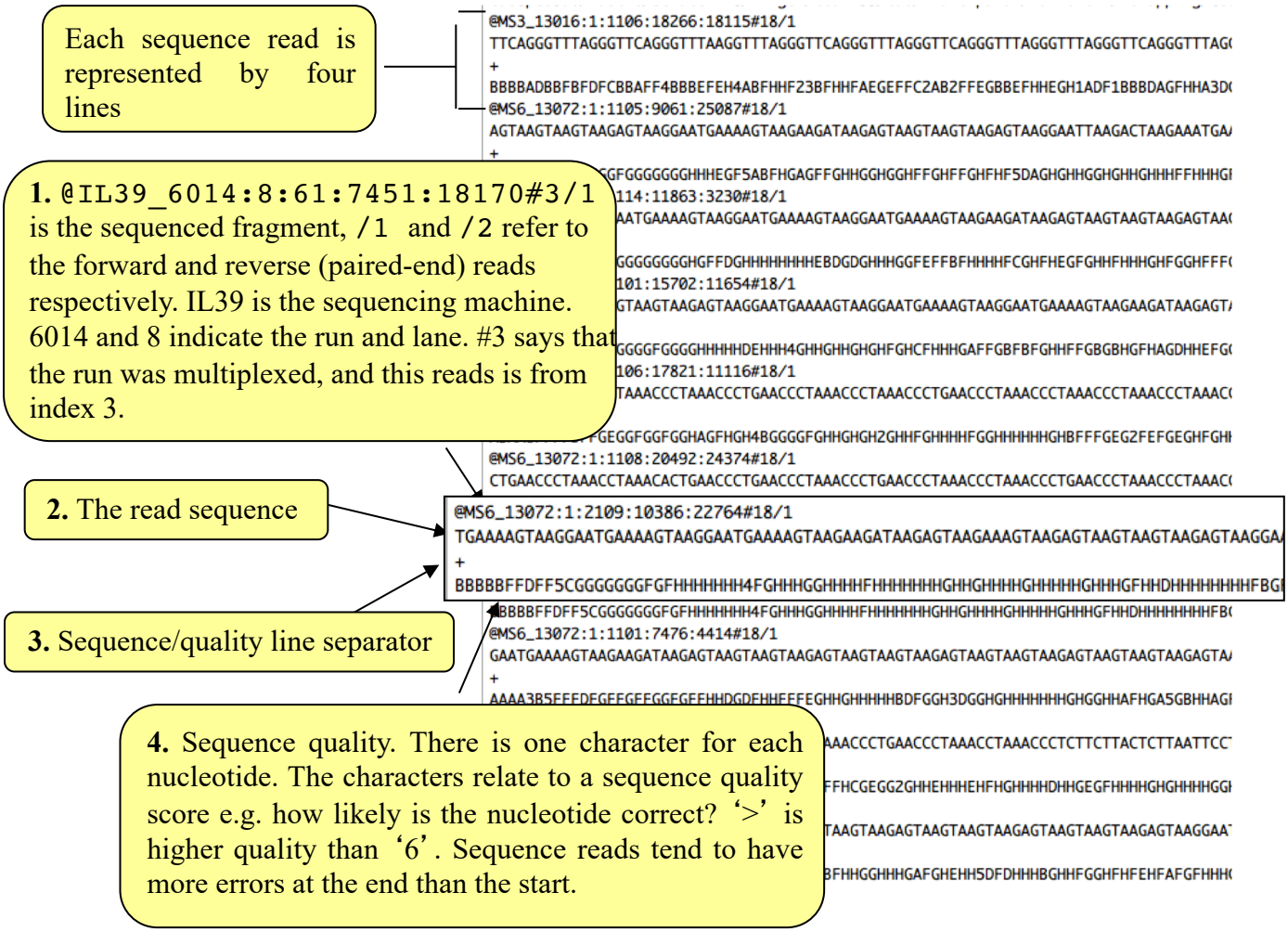
```
$ pwd
```

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
```



Due to time restrictions we will not trimm adapters or low quality regions of the reads. You determine the quality of you reads with the program fastqc. Just type fastqc <READfile> where READfile is your fastq file.

Mapping the reads with BWA mem

Now we will map the IT clone 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`** (the algorithm need to access specific positions on the reference in an efficient way).

```
$ 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 2 reference read_1 read_2 > BWA.sam
```

This should have worked quite fast. **IMPORTANT**, you have specify 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
`$ l`
This will list the files in reverse time order. You should also see the files that the index of bwa generate. If something is missing, adjust your bwa calls.

The details of where each read has been mapped is now stored in the file BWA.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.

To generate the bam file is slightly easier, as BWA already generate a SAM file. The following comment goes over two lines!

```
$ samtools view -Sb BWA.sam | samtools sort - > BWA.18.bam
```
Depending on you samtools version (if < 1.3) you might need to use:
```
$ samtools view -Sb BWA.sam | samtools sort - BWA.18
```

```
$ samtools index BWA.18.bam
```

These commands transform the sam file into a binary format, sort the reads by occurrence against the reference and than indexed the bam file by chromosome/contig.
With the following command you can get the stats of the mapping:

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

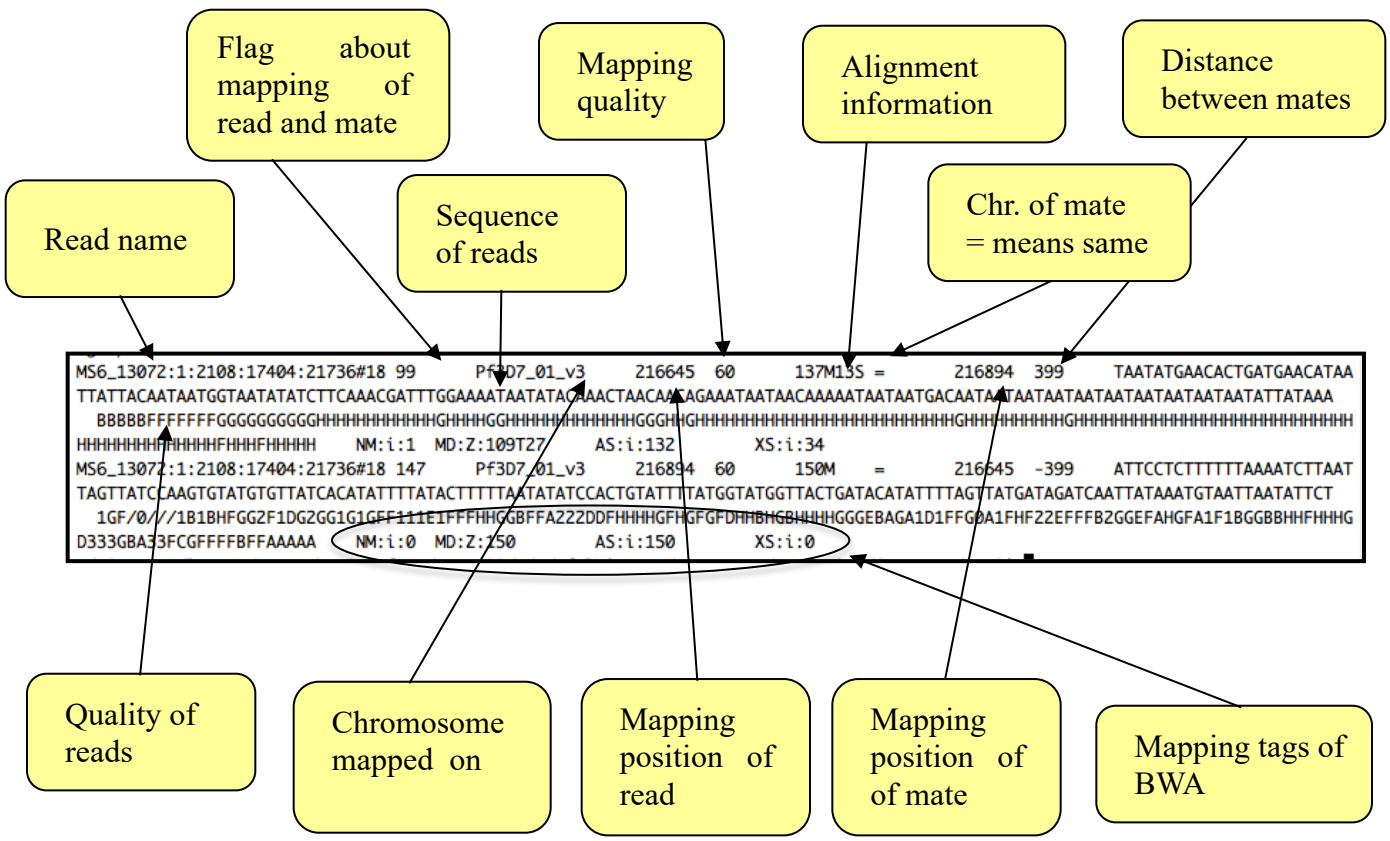
SAMTOOLS format

Before we visualize the alignment of the reads in a Artemis, let's have a look at this sam/bam format. SAMTOOLS was developed to have a standard format to store reads. It contains information about the reference sequence, where a read is mapped, quality of mapping, and where it's mate is mapped. Files ending with .sam, are normally plain text. but as this might take too much space, the file is compressed into a bam file.
All visualization tools will need the bam file. It has to be sorted (by chromosome and position) and indexed. Indexing enables a fast work with the alignments.

Here is an example of a read and its mate in the sam format.

```
$ grep 2108:17404 BWA.sam
```

You will see the information for two mates.



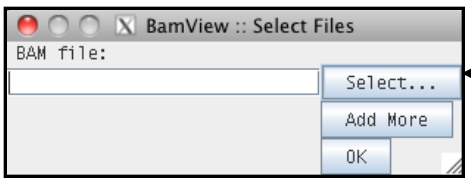
SAMTOOLS is a powerful format. It is unlikely that you will need to work with it directly in the future, although you may use it in a viewer. However, for bioinformaticians this is a perfect format to do further analysis. The specifications are at: http://samtools.sourceforge.net/SAM1.pdf

D. Viewing the mapped reads in Artemis

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

Open Artemis and load `Pf3D7_01_v3.embl` from the `genomic` data directory. This contains exactly 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.

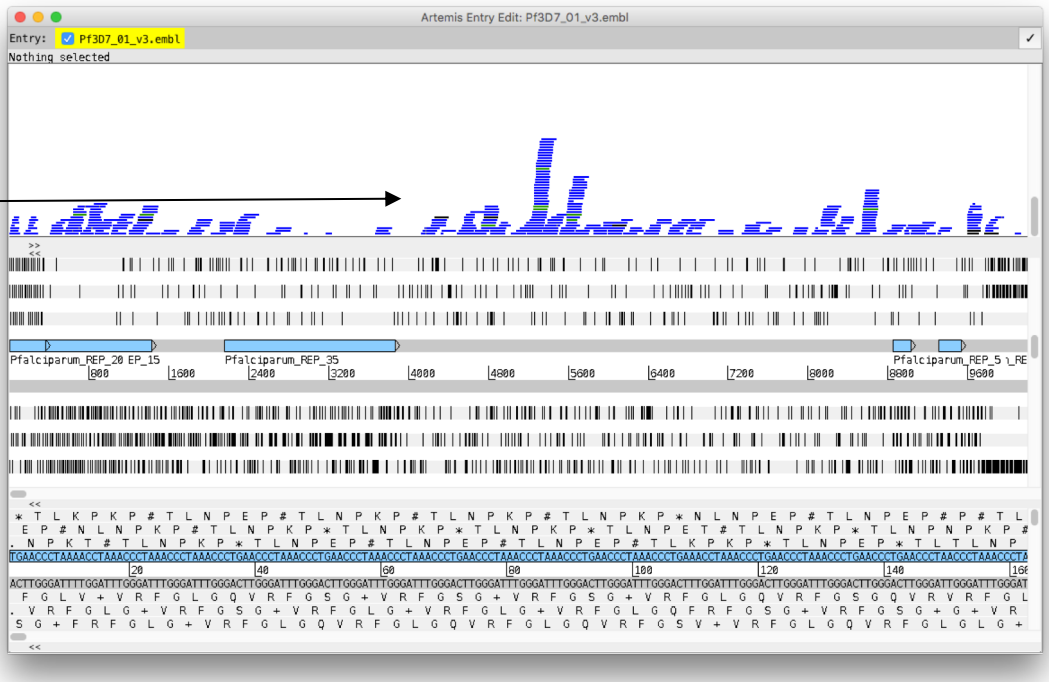


Select BWA.18.bam

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

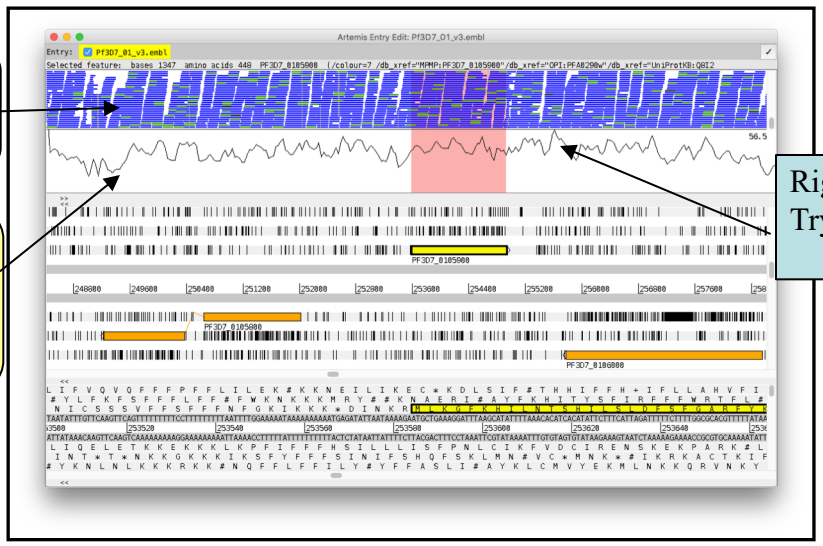


Scroll through the genome. Describe the coverage. Are there regions that are not covered?

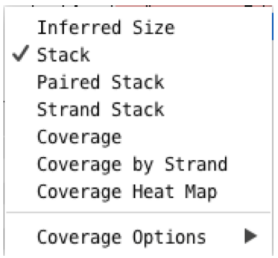
Go to gene PF3D7_0105900 and zoom out once.

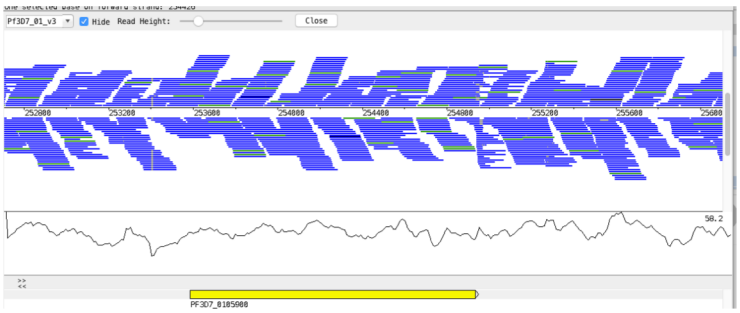
Stack view

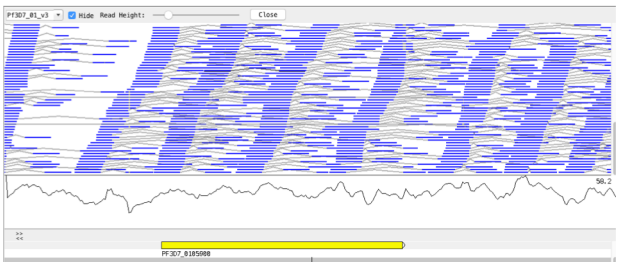
Coverage plot. Right click for more options.

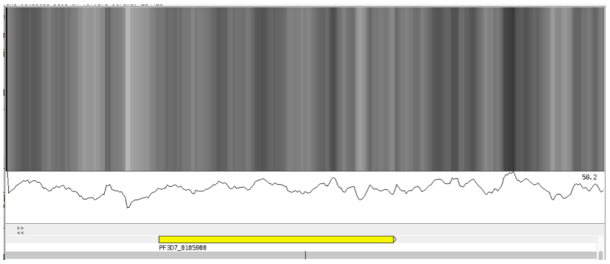


Right click on a read. Try different views.





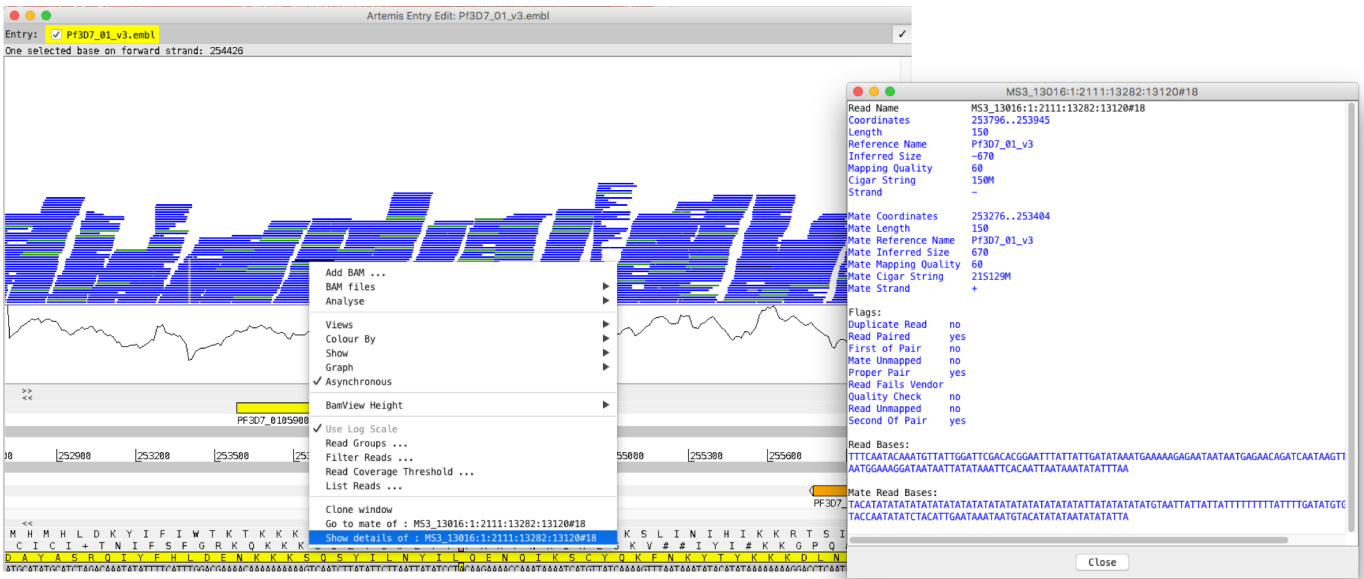




Each view has its advantages:

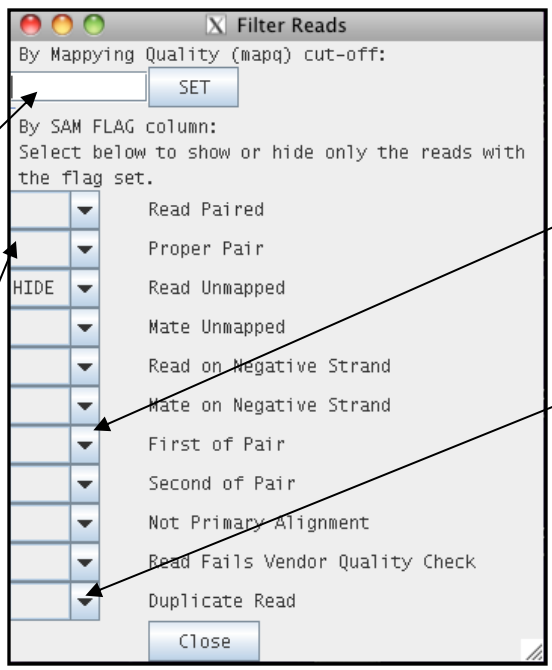
- "Inferred Size" (click also Use log scale) displays the mate pair on the y-axis depending their distance. In this case, some reads map further apart compared to others. Could this be a deletion?
- "Strand Stack": Shows the strand where reads are mapping. Useful for strand specific applications.
- "Paired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

Zoom in and position the mouse over a read until a window pops up. Then right click on the read. Select "`Show details of…`"



1. Those two windows give you information of the mapping of the reads, as it is in the bam file. Notice the "Mapping quality". The maximum value for this is 99. The mapping quality depends on the accuracy of the sequence read and the number of mismatches with the reference. A value of 0 means that the read mapped equally well to at least one other location and therefore is not reliably mapped. The flags describe the read's mate pair mapping. Most of those values can be filtered:

2. Right click on a read. Select "`Filter`". Change some settings and see how the BAMview is changing.



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

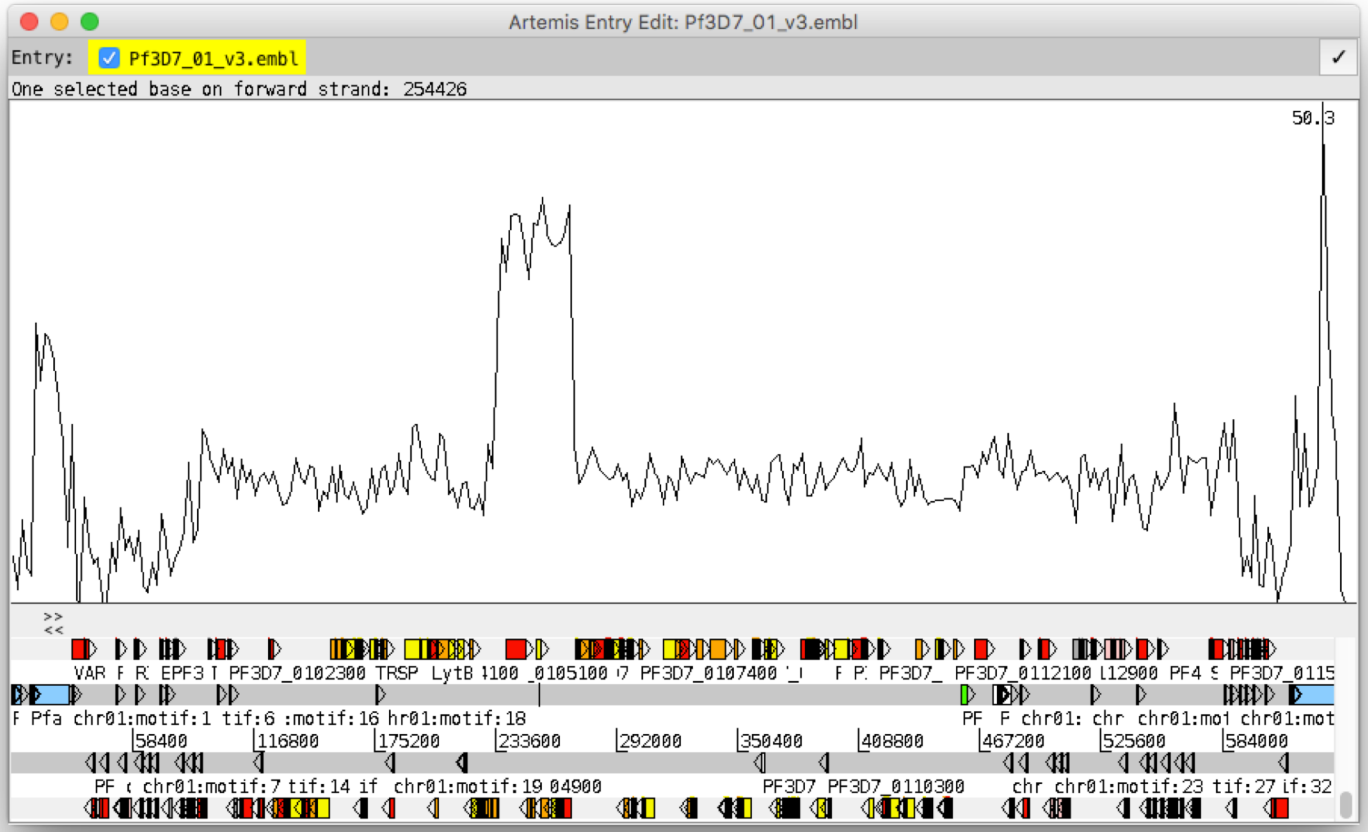
HIDE the proper pairs. What happened?

What is the difference between this and Proper Pair HIDE?

Are there any duplicated reads?

Filtering reads for repetitive regions 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, interesting is this copy number variation (CNV) in the centre of the chromosome. To exclude that this is some special about PfDd2, we have to map the reads of the PfDd2 WT onto the reference and compare the coverage.

```
$ bwa index reference
```

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 reference. Call the output **BWA.WT.sam**. Than index

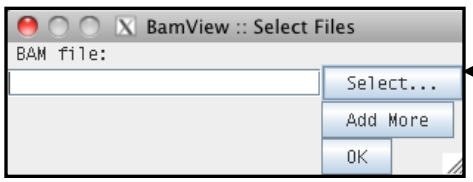
```
$ bwa mem –t 2 reference read_1 read_2 > outputFile
```

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

Then transform the sam file (outputFile) into a bam file with the samtools view and samtools sort command as before.

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).

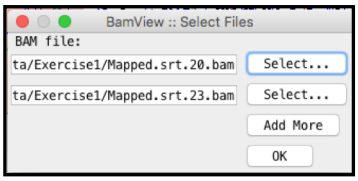
So far you looked at the **Clone 18**. What about the *mdr1* locus in the **Dd2 WT** of the malaria parasite? Right-clicking on the BAMview and choosing 'Add BAM…'. Select the file BWA.WT.bam



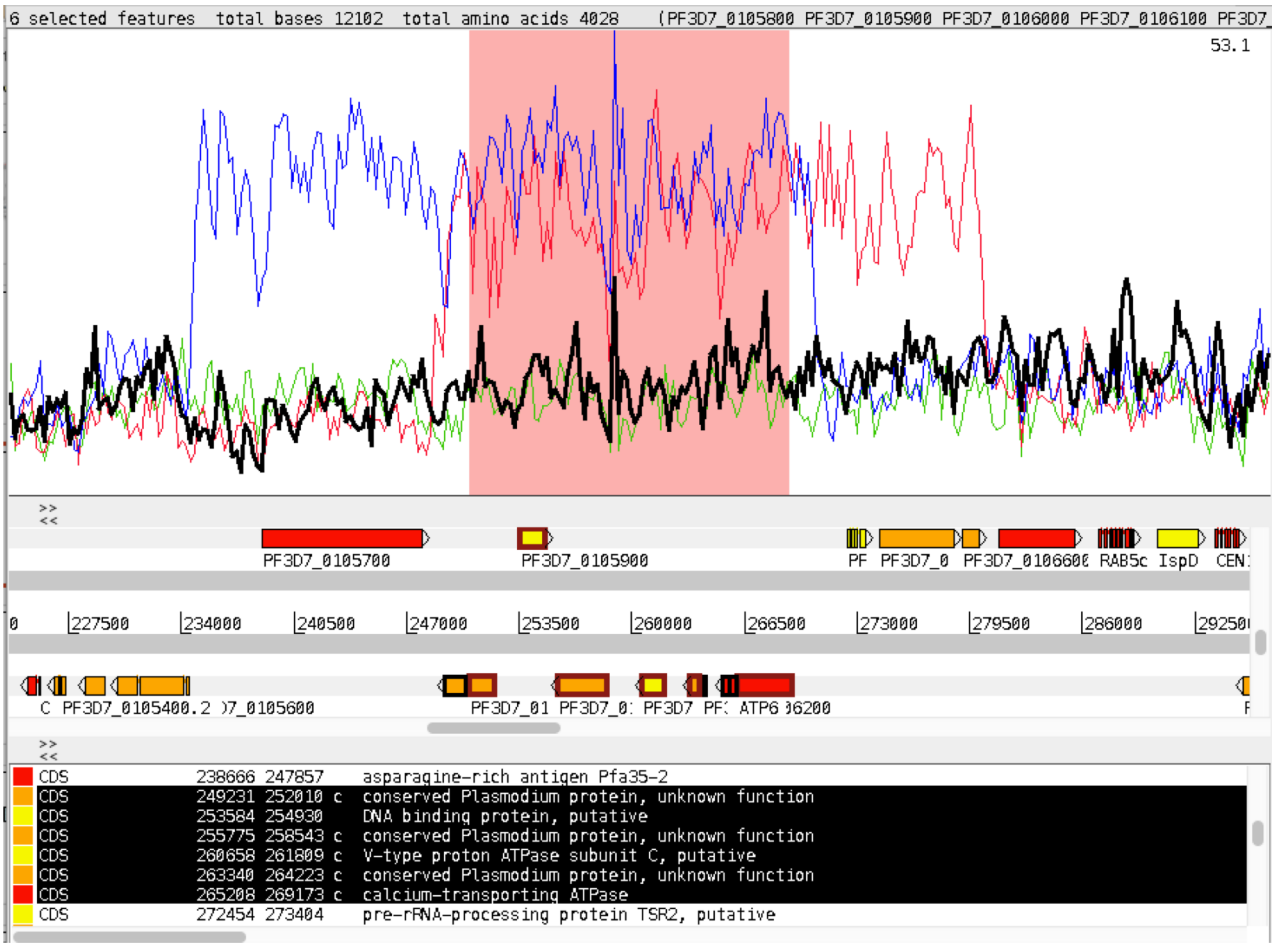
Select BWA.WT.bam

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 further BAM files. You can see them 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?

F. Identifying Single Nucleotide Polymorphisms

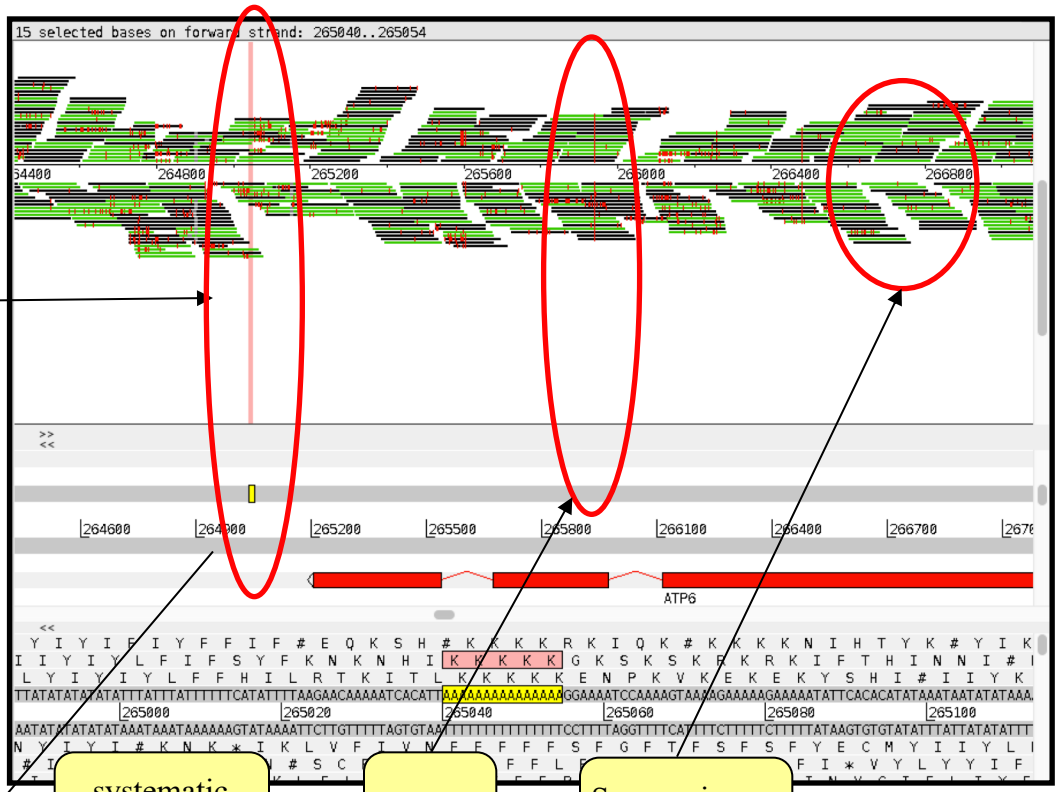
One possibility could be that green (23) clone has mutation, right!!! How could we see those?

Right click here,

Select Show-> SNPs

&&

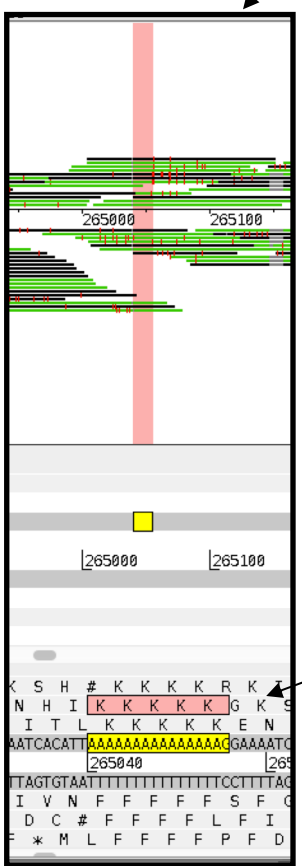
Colour by -> Coverage Plot Colourst..



systematic error

SNP

Sequencing errors



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

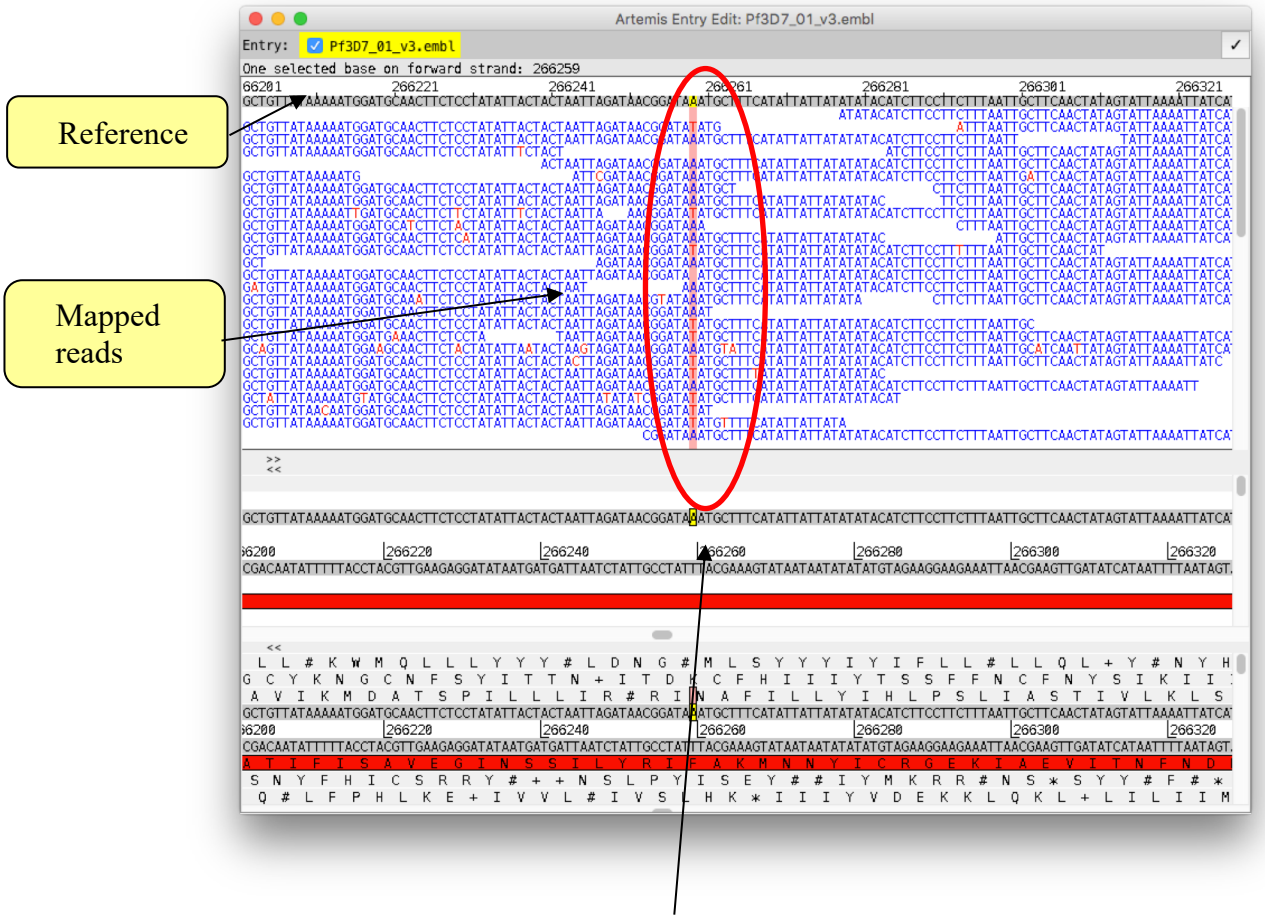
Homopolymer track generate systematic errors.

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

When you zoom in on position 266259 as far as you can go, you can see a SNP: It occurs not in all reads… just the reads of the clone 23...

What is the consequence of this SNP? Can we tell what effect it will have for the clone? Is that the drug mechanism?
Well, but what with the other mutations?

Remember, we obtained all the samples from PfDd2. So if a different 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, fortunately there are tools to call variants from bam files. The most common is samtools. Here we are going to explain how to call the variant for the Clone 23 on chromosome 1, starting from the BAM file. There are better ways to call SNP, like gatk haplotype caller, but due to copyright issues, we cannot present it. But mpileup will also generate the wished 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. 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!***).

samtools version> = 1.1

```
$ 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
```

Here again depending on your samtools version (< 1.1)

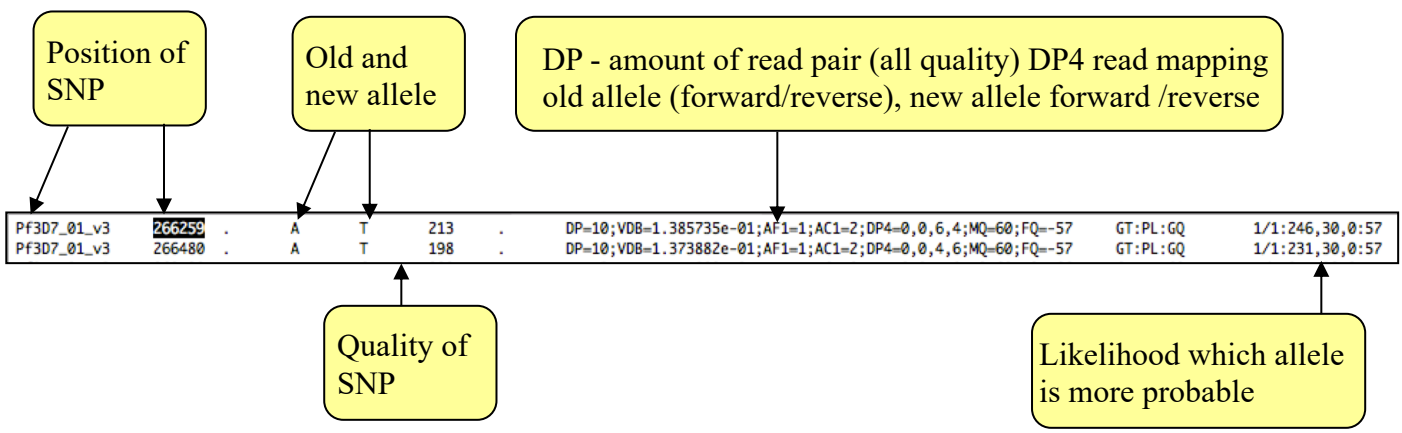
```
$ samtools mpileup -ugf reference bam-file |
bcftools view -bcvg - > Clone.23.bcf
$ bcftools index Clone.23.bcf
```

Alternative,
As before, the reference is `Pf3D7_01_v3.fasta` and the bam file is `Exercise1/Mapped.srt.23.bam`.

The new file contains all the SNP information. For visualization, it now just needs to be indexed:
To look at the output, do

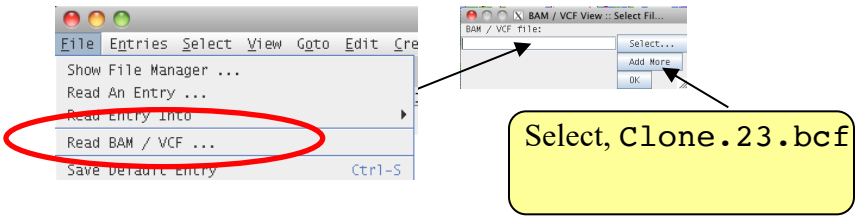
```
$ bcftools view clone.23.vcf.gz | less
```

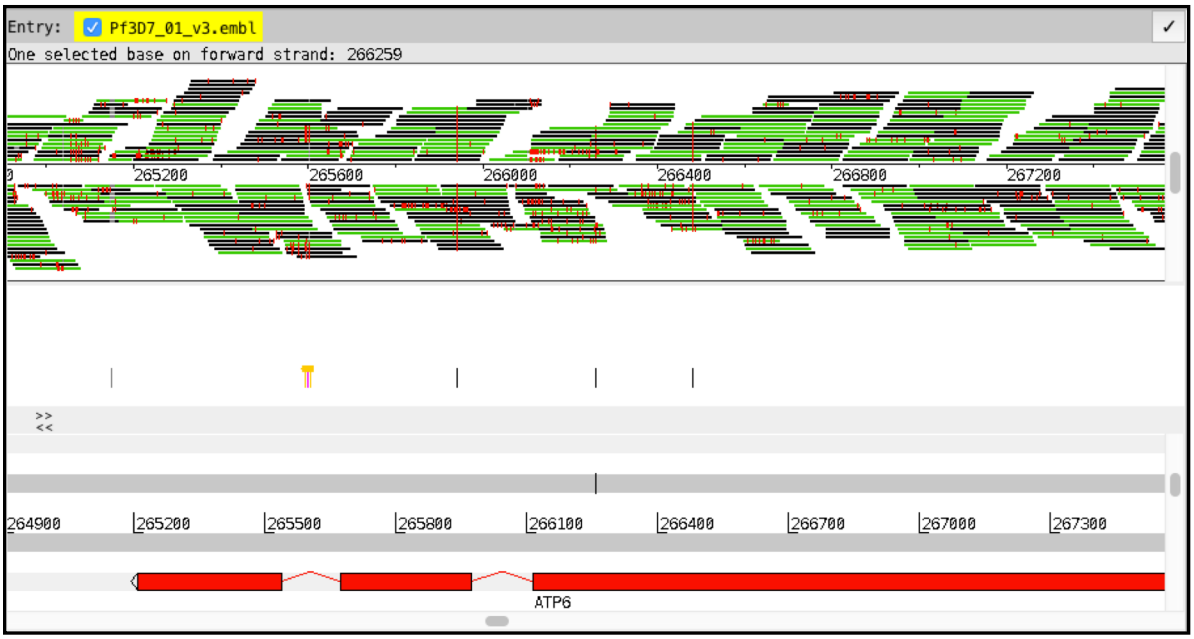
Exit the less command with “q” for quit.



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 gene.





Have a look at the different variants there are in 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 do the calls 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

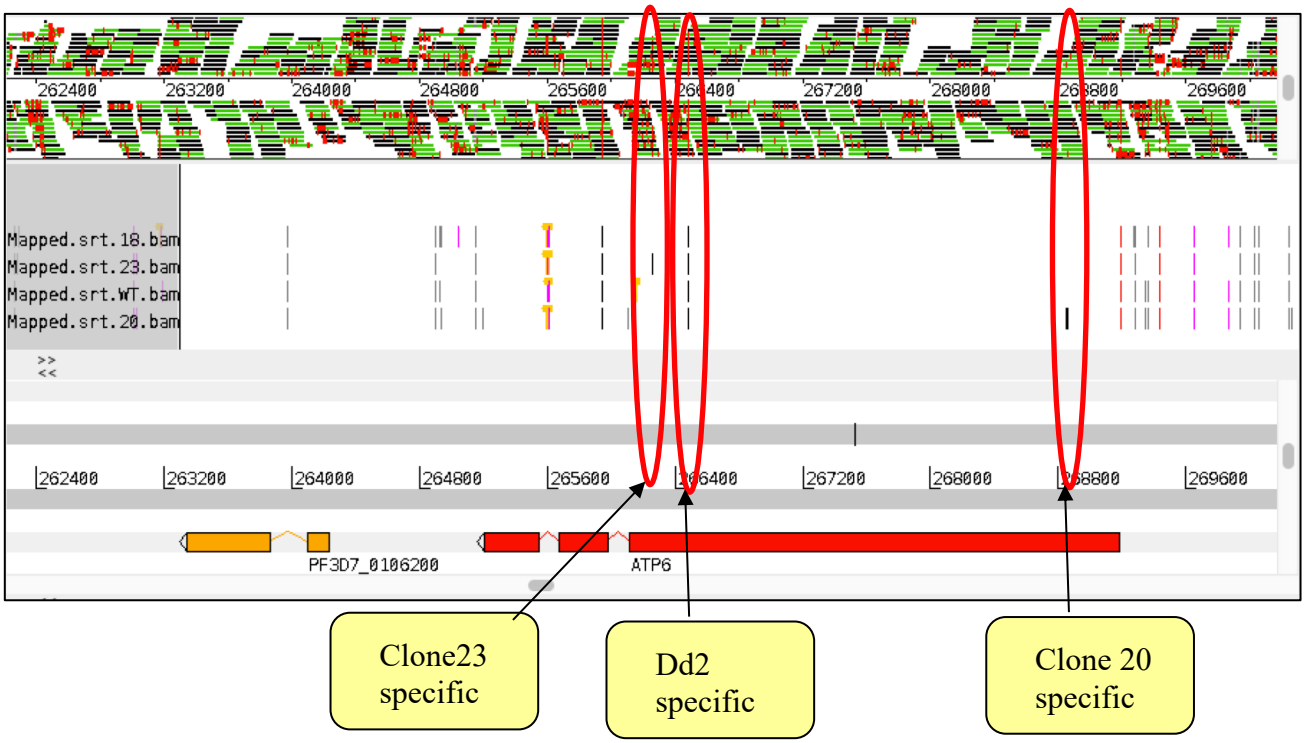
```
$ bash ./do.SNP.sh &> out.txt &
```

Maybe have a look at the script:

```
$ cat do.SNP.sh
for x in 18 20 WT ; do
 bcftools mpileup -f Pf3D7_01_v3.fasta Exercise1/Mapped.srt.$x.bam | bcftools
call -cv -Ov --ploidy 1 -o clone.$x.vcf;
bgzip clone.$x.vcf;
tabix clone.$x.vcf.gz ;
done
```

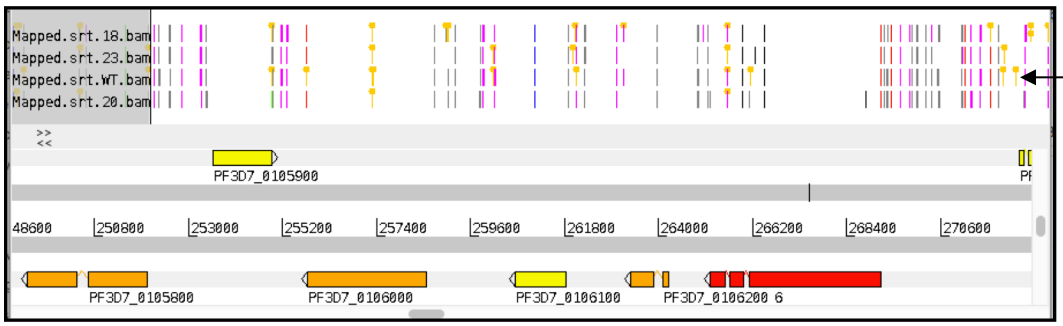
It basically iterates through the four samples: So the variable x will hold the name of the sample (18. 20, 23 or WT), and then the same command is done four times.
It takes a little bit, but you have free time… well, not for long!

Load the four files into artemis (right click on the SNP view -> add). They are called clone.18.vcf.gz, clone.WT.vcf.gz etc...



So we can already see mutation specific for PfDd2 (not really interesting, right?) And then SNP 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



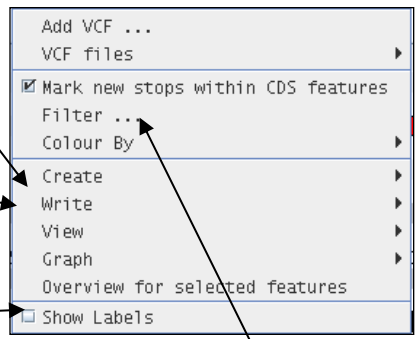
Right click, and a window opens

In this menu are some useful options:

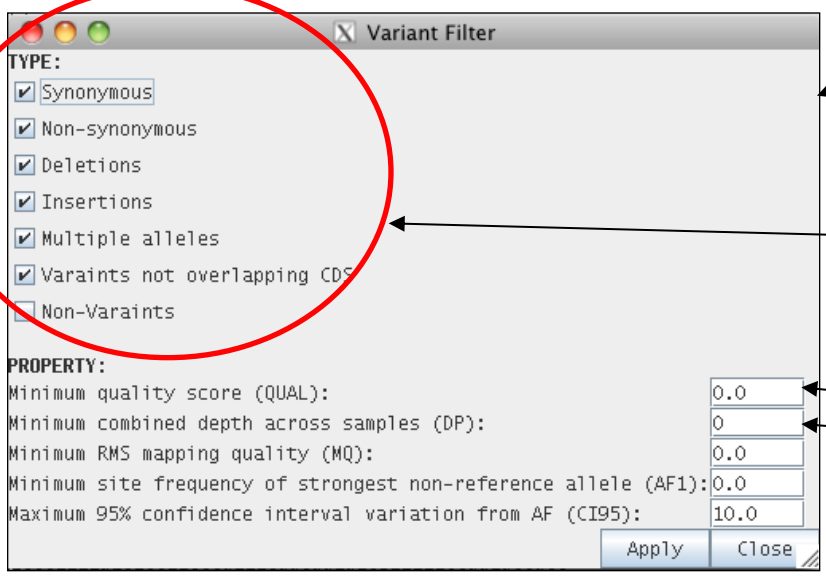
Write filtered vcf, or fasta of the genes with the new alleles

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

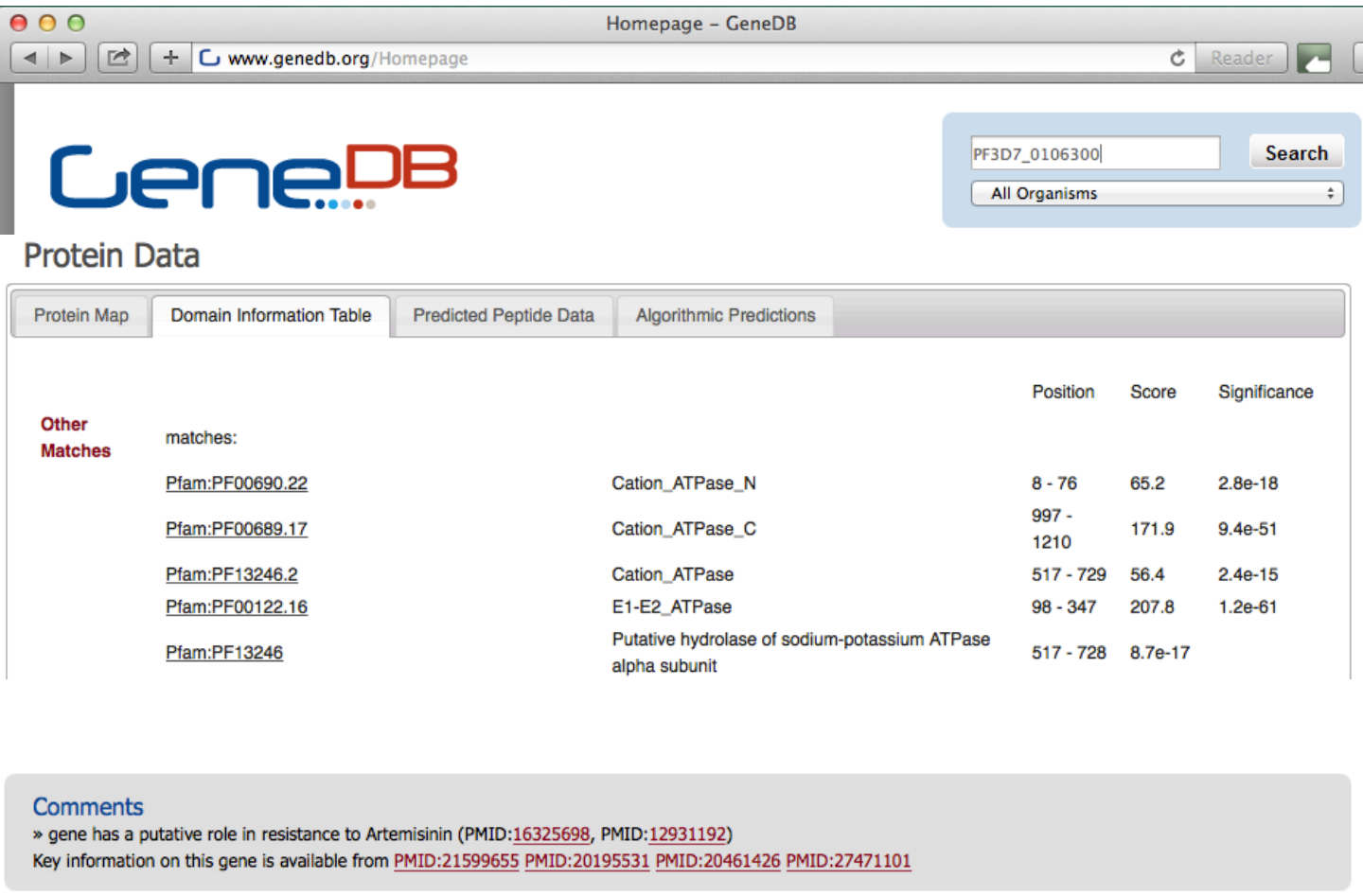
So it seems that ATP6 is the one interesting target. Get the new sequence of it gene (right click on BCF view -> view -> FASTA of selected feature). Click ok on the next window and three windows with fasta sequences will open. What happens if you blast it on plasmoDB?

```
> 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     ATGGAAGAGGTTATTAAGAATGCTCATACATACGATGTTGAGGATGTACTAAAATTTTTG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGGAAGAGGTTATTAAGAATGCTCATACATACGATGTTGAGGATGTACTAAAATTTTTG  60

Query  61    GATGTAAACAAAGATAATGGTTTAAAGAATGAGGAATTGGATGATAGAAGATTAAAATAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GATGTAAACAAAGATAATGGTTTAAAGAATGAGGAATTGGATGATAGAAGATTAAAATAT  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 geneID 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 fall into known PFAM domains or transmembrane proteins. Latter would give you an idea of the impact of the mutation versus the function.

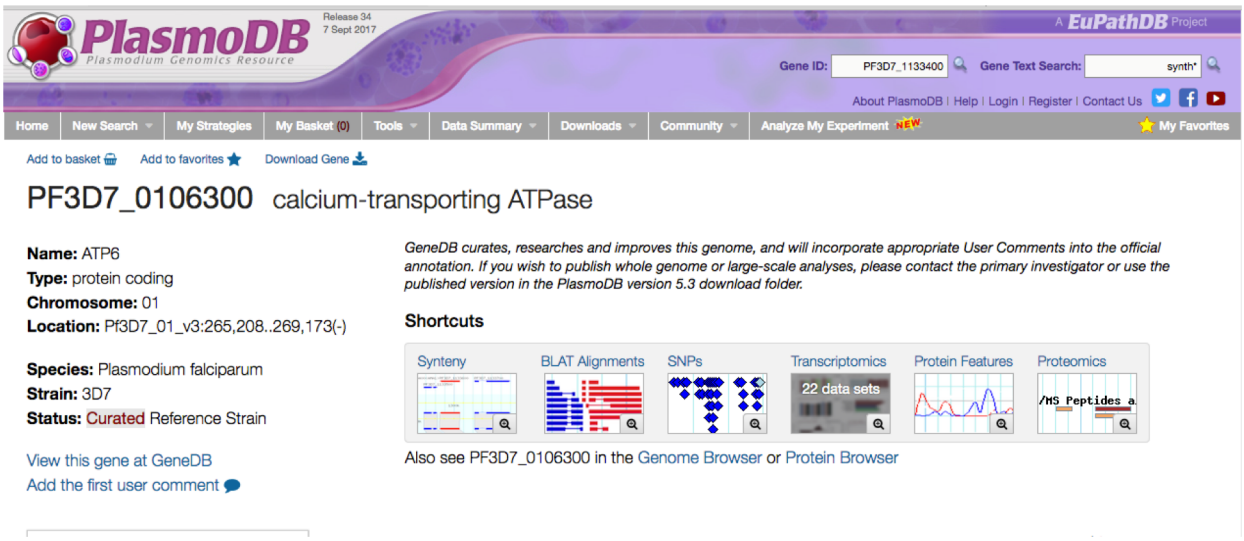


Where are the mutation of the Clone 20 and Clone 23? Are they in a specific domain?

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.

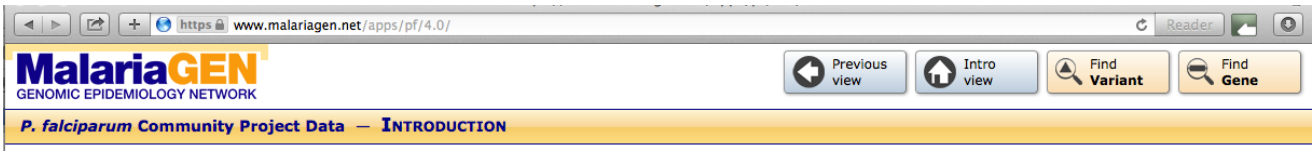


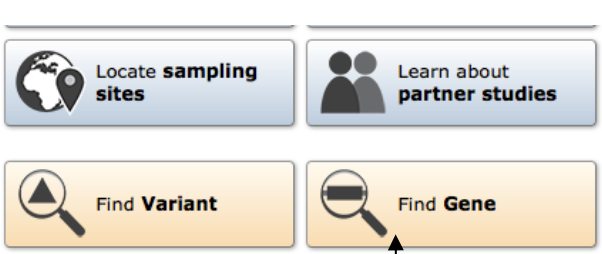
Where are the mutation of the Clone 20 and Clone 23? Are they in a specific domain?

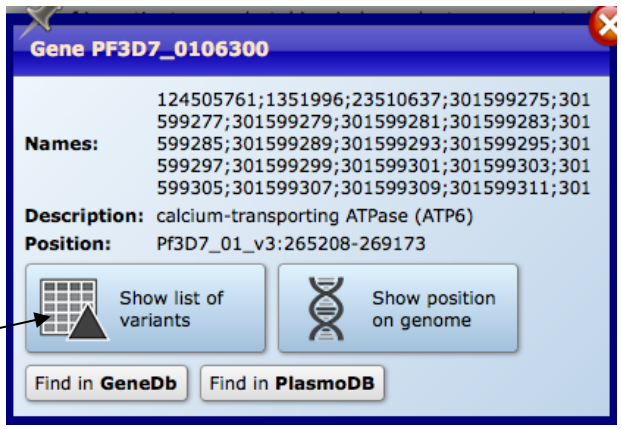
Panoptes

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/.







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

Position	NRAF WAF	NRAF CAF	NRAF EAF	NRAF SAS	NRAF WSEA	NRAF ESEA	NRAF OCE	NRAF SAM	MAF Global	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 exported?

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, started a script and analysed the data. You looked at the data in Artemis and finally did some functional analysis!

You are very close to be 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

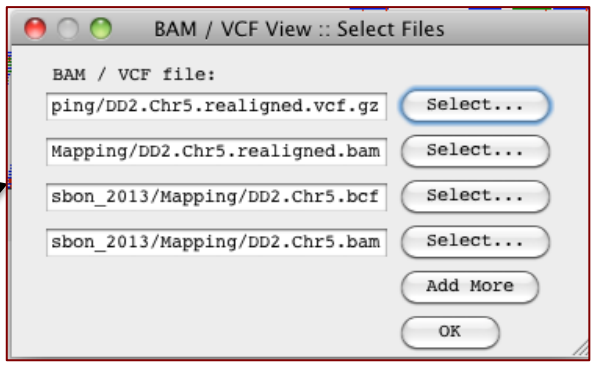
The pdf is in the Module directory, called paper.pdf – have fun reading it and remember, you replicated the bioinformatics analysis!

H. Realignment

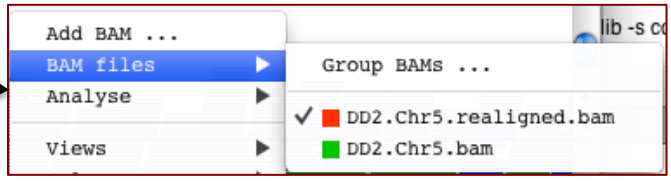
When analysing the BAM files and the variant calls, you might have noticed heterozygous SNPs. This is very unlikely in a haploid genome. This could indicate a collapsed repeat, CNV or inproperly aligned reads.
To improve the alignment of reads, we are going to compare the DD2 bam file with a realigned 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.

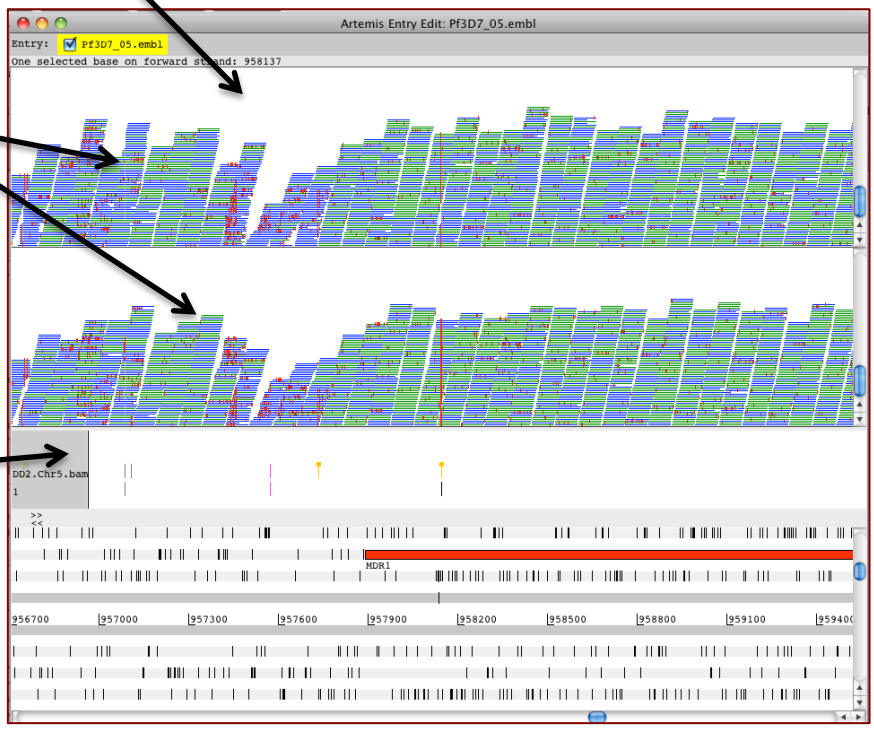


3. Clone the BAMview, so that on top is the DD2.Chr5.bam, and below is the file re-aligned BAMfile.



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

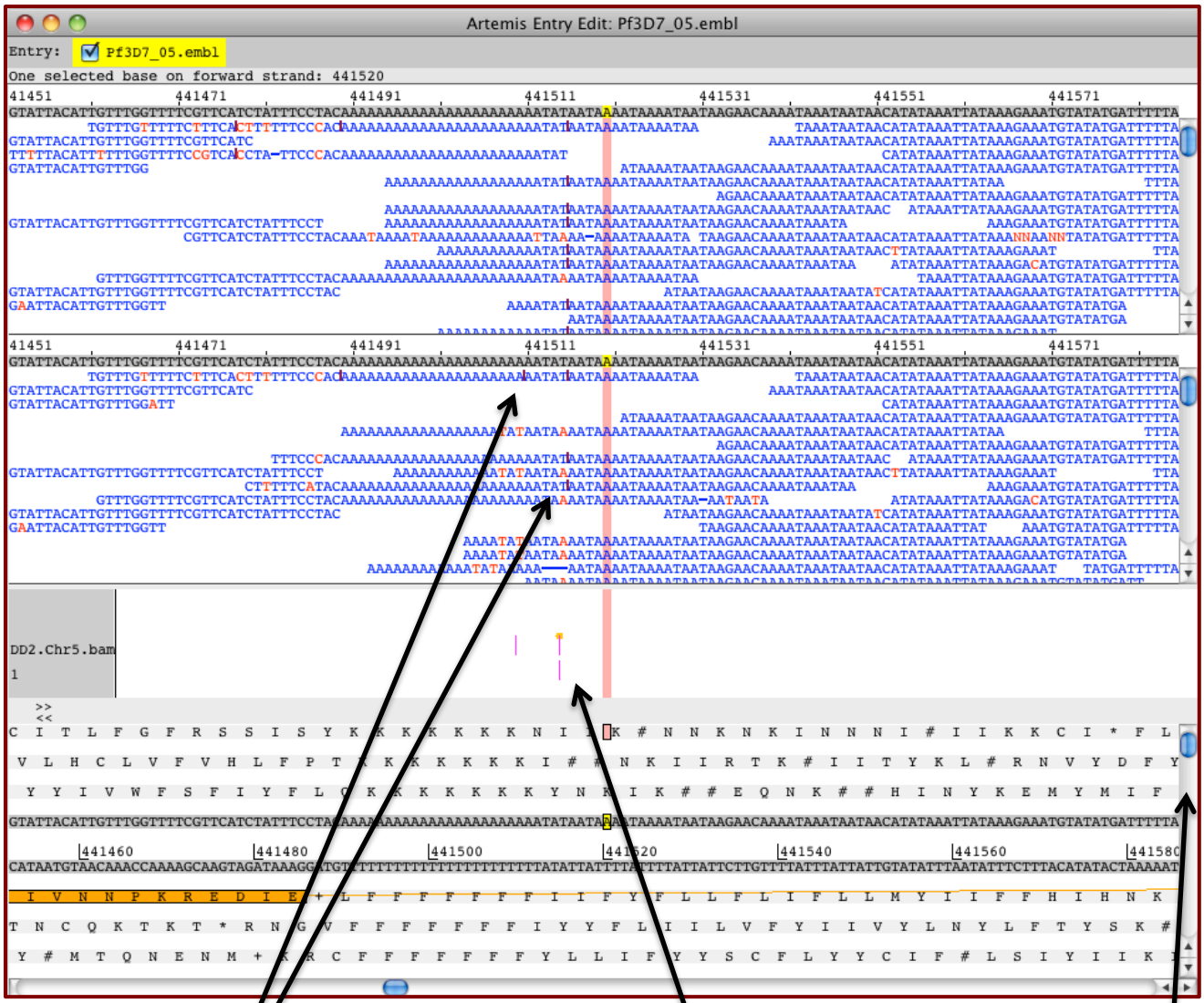
5. Show the labels in the BCF view (right click -> show labels).



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

Go to position 4415200 and zoom until you see the bases of the reads.

1. Can you see where the alignments in the bam files are different? (find insertions - vertical bars)
2. Can you explain the different variant calls?
3. Which calls seem to be more accurate?



Look at the insertions. The read indicate an insertion, that BWA aligner places once left, mostly right. In the re-aligned BAM file, the insert is always right.

The variant calls are different.

Zoom in here.

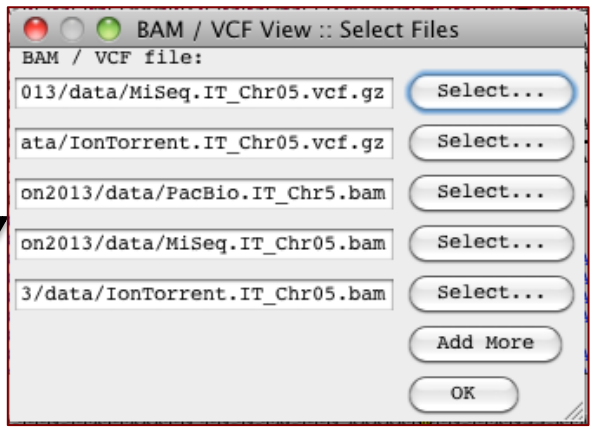
G. Comparing different sequencing technologies

So far we just 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:

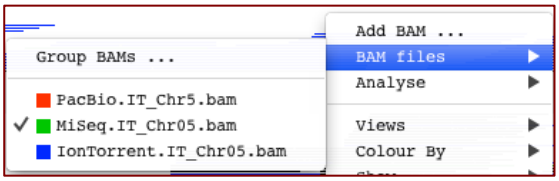
1. Look at coverage variation in each technology. Is there a correlation with the GC content?
2. Which technology has the longest reads, and which has the most accurate reads?
3. Which technology is better for SNP calling?

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).

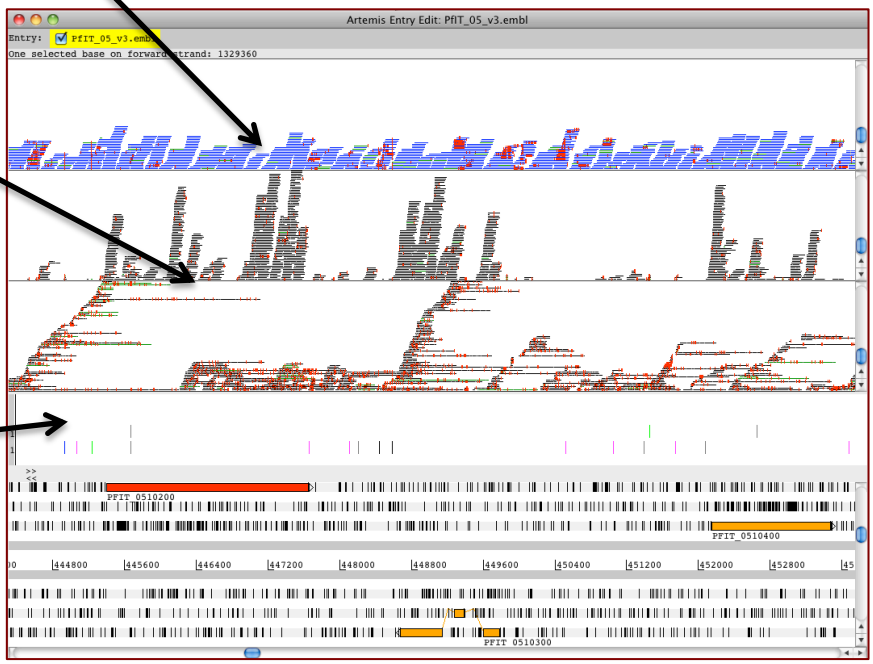


3. Clone the BAMview three times as done before.



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. Maq will randomly report only one of the mapping locations and give the read a score of 0.

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. Maq) 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, TopHat (Trapnell et al., 2009) 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 options. 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 (but the maximum score you will usually see is 40). 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.

.