

2015 Workshop on Genomics

Genomics Laboratory

Instructors:

- Konrad Paszkiewicz k.h.paszkiewicz@exeter.ac.uk

Objectives:

By the end of the lab you will be expected to:

- Understand how short reads are generated.
- Interpret FASTQ quality metrics and remove poor quality reads and adaptor contamination.
- Align reads to a reference sequence to form a SAM file (Sequence Alignment/Map file) using BWA.
- Identify SNPs, Indels, missing or truncated genes with respect to the reference genome.
- Identify and annotate novel genes (with respect to the reference genome).
- Assemble short reads de novo using SPAdes.
- Improve the assembly using longer PacBio reads.
- Assess and compare assemblies using QUAST.
- Compare SNPs searches across several strains.

This document includes a lot of tasks - don't worry if you don't have time to finish it! Don't rush! We hope that most students will be able to finish Part 2 and other parts can be continued during open lab times for those interested.

Part 1: Short read genomics: Introduction

Welcome to the Genomics laboratory. Generating reams of data in Biology is easy these days. In little more than a fortnight we can generate more data than the entire human genome project generated in over a decade of work. Making biological sense out of that data, understanding its limitations and how the analysis algorithms work is now the major challenge for researchers. The aim of this lab is to take you through an example project. On the way you will learn how to evaluate the quality of data as provided by a sequencing facility, how to align the data against a known and annotated reference genome and how to perform a de-novo assembly. In addition you will also learn how to compare results between different samples.

This lab is broken into 5 parts. You should feel free to take as long as you like on each part. It is much more important that you have a thorough understanding of each part, rather than try to race through the entire lab.

The five parts are:

1. Introduction to Illumina sequencing-by-synthesis
2. Remapping a strain of *E.coli* to a reference sequence
3. Assembly of unmapped reads
4. Complete *de-novo* assembly of all reads
5. Repeating parts 3-5 on strains of *Vibrio parahaemolyticus* and comparing them

For this first lab we will assume little background knowledge, save a basic familiarity with the Linux operating system and the Amazon cloud. We will cover the basics of how genomic DNA libraries are generated and sequenced, and the principles behind short read paired-end sequencing. We will look at why data can vary in quality, why adaptor sequences need to be filtered out and how to quality control data. In the second part we will take the plunge and align the filtered reads to a reference genome, call variants and compare them against the published genome to identify missing, truncated or altered genes. This will involve the use of a publicly available set of bacterial *E.coli* Illumina reads and reference genome.

In parts 3 and 4 we will look at how one can identify novel sequences which are not present in the reference genome. In part 5, you will be asked to repeat the steps in parts 1, 2 and 3 on other data sets and to compare the results.

A word on notation. If you see something like this:

```
cd ~/workshop_data/genomics_tutorial/reference_sequence
```

It means, type the highlighted text into your terminal.

Some of the outputs you produce may differ slightly from the screenshots here. DON'T PANIC! This is just due to changes in versions of software between the time of creating the lab exercise and now.

Principles of Illumina-based sequencing:

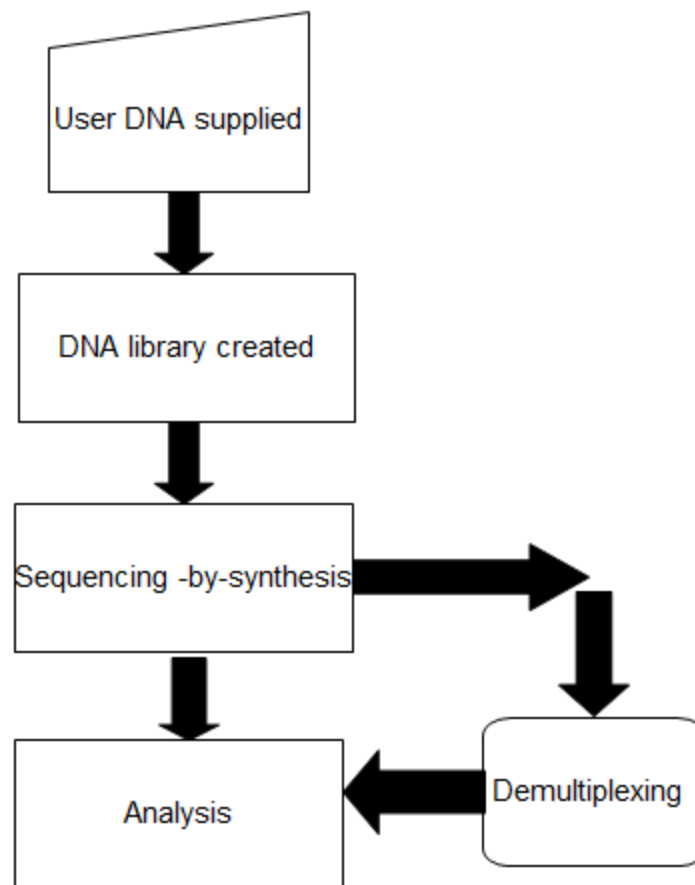
There are several second generation (i.e. non-Sanger) sequencers currently on the market. These include the Life 5500 (formerly known as the ABI SOLiD), the Roche 454 GS FLX and 454 Junior, Ion Torrent, and the Illumina HiSeq and MiSeq systems. All of these systems rely on making hundreds of thousands of clonal copies of a fragment of DNA and sequencing the ensemble of fragments using DNA polymerase or in the case of the SOLiD via ligation. This is simply because the detectors (basically souped-up digital cameras), cannot detect fluorescence (Illumina, SOLiD, 454) or pH changes (Ion Torrent) from a single molecule.

The 'third-generation' Pacific Biosciences SMRT (Single Molecule Real Time) sequencer, is able to detect fluorescence from a single molecule of DNA. However, the machine weighs 2 tons, produces 1/20000th of the data of an Illumina run and has a 10-15% error rate for a single read (not much better than guessing!) with read lengths up to 40kb. Using multiple reads to build up a consensus, it is possible to reduce this error rate to less than 1% (similar to Illumina). This system has become very useful for improving de-novo assemblies, detecting large-scale structural variants, phasing and transcript discovery as well as direct-detection of base modifications.

The Oxford Nanopore, although still in beta-testing, is also capable of sequencing single molecules of DNA by threading the DNA through a nanopore embedded in a membrane and detecting changes in electrical current. The error rates for this sequencer are anywhere between 15-40% in the hands of users (although Oxford Nanopore report error rates below 10%). The read lengths for this platform are between 1-40kb and are primarily dependent on the length of the input DNA. It can prove difficult to prepare long molecules.

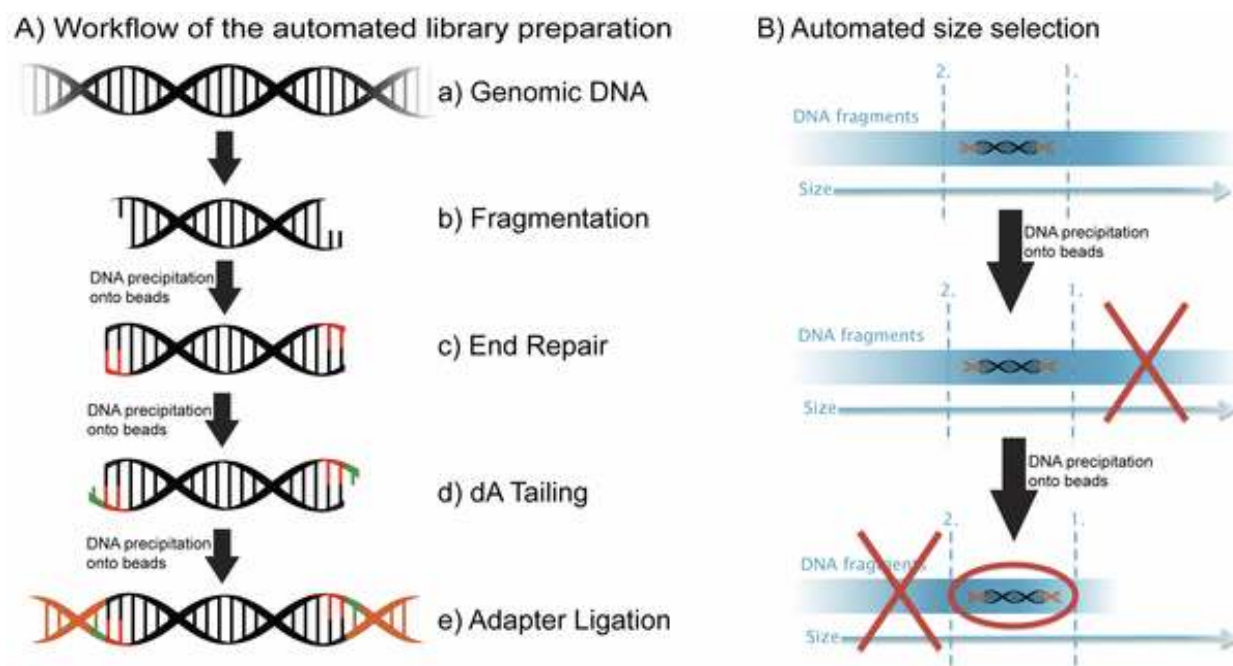
We will primarily look at a pure Illumina sequencing pipeline here, but the basic principles apply to all other sequencers. There are some exercises which also incorporate long read data as an example of the benefits longer reads can provide, such as providing scaffolding that can orient contigs and span repeat regions. If you would like further details on other platforms then I recommend reading *Mardis ER. Next-generation DNA sequencing methods. Annual Reviews Genomics Hum Genet 2008; 9 :387–402*

A typical sequencing run would begin with the user supplying 10ng - 1ug of genomic DNA (depending on the library preparation) to a facility along with quality control information in the form of an Agilent Bioanalyser trace or gel image and quantification information. The following flowchart illustrates the basic workflow.



DNA Library preparation

For most sequencing applications, paired-end libraries are generated. Genomic DNA is sheared into 300-600bp fragments (usually via sonication) and size-selected accordingly. Ends are repaired and an overhanging adenine base is added, after which oligonucleotide adaptors are ligated. In many cases the adaptors contain unique DNA sequences of 6-12 bp which can be used to identify the sample if they are 'multiplexed' together for sequencing. This type of sequencing is used extensively when sequencing small genomes such as those of bacteria because it lowers the overall per-genome cost.



A) Steps a through e explain the main steps in Illumina sample preparation: a) the initial genomic DNA, b) fragmentation of genomic DNA into 500bp fragments, c) end repair, d) addition of A bases to the fragment ends and e) ligation of the adaptors to the fragments.

B) Overview of the automated the size selection protocol: The first precipitation discards fragments larger than the desired interval. The second precipitation selects all fragments larger than the lower boundary of the desired interval.

Borgström E, Lundin S, Lundeberg J, 2011 Large Scale Library Generation for High Throughput Sequencing. *PLoS ONE* 6: e19119. doi:10.1371/journal.pone.0019119

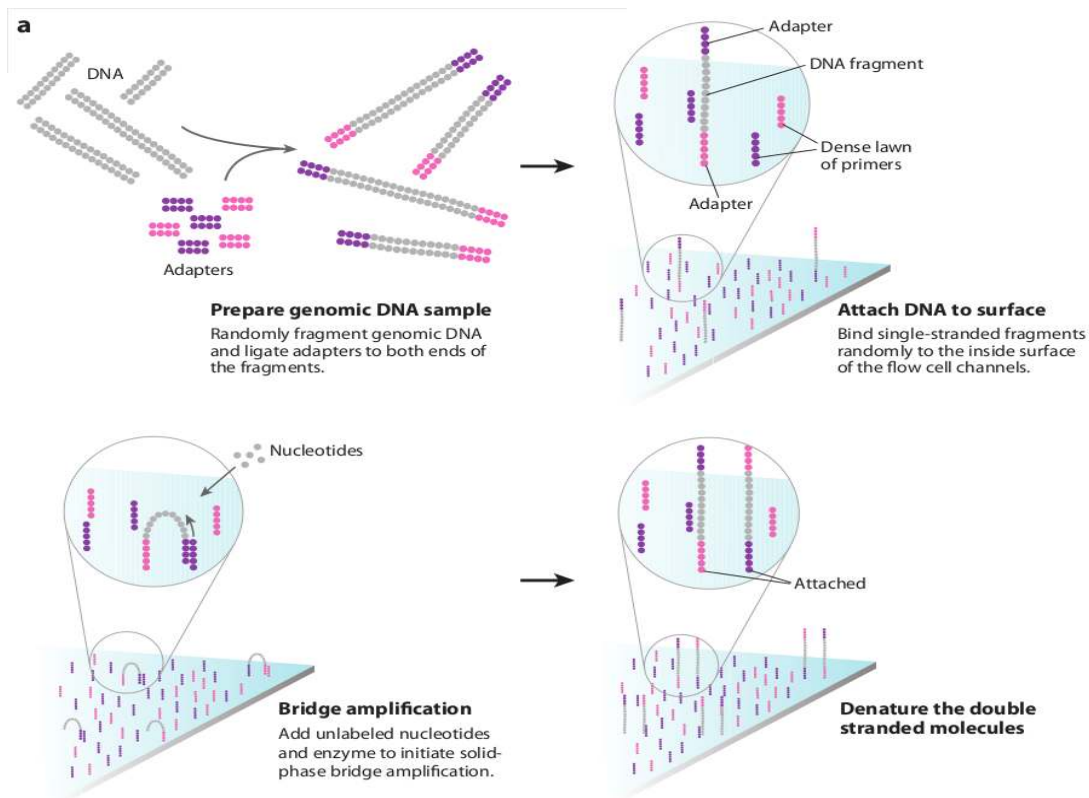
Sequencing

(adapted from Margulis, E.R., reference below)

Once sufficient libraries have been prepared, the task is to amplify single strands of DNA to form monoclonal clusters. The single molecule amplification step for the Illumina HiSeq 2500 starts with an Illumina-specific adapter library and takes place on the oligo-derivatized surface of a flow cell, and is performed by an automated device called a cBot Cluster Station. The flow cell is either a 2 or 8-channel sealed glass microfabricated device that allows bridge amplification of fragments on its surface, and uses DNA polymerase to produce multiple DNA copies, or clusters, that each represent the single molecule that initiated the cluster amplification.

Separate or multiple libraries can be added to each of the eight channels, or the same library can be used in all eight, or combinations thereof. Each cluster contains approximately one million copies of the original fragment, which is sufficient for reporting incorporated bases at the required signal intensity for detection during sequencing. The Illumina system utilizes a sequencing-by-synthesis approach in which all four nucleotides are added simultaneously to the flow cell channels, along with DNA polymerase, for incorporation into the oligo-primed cluster fragments (see figure below for details). Specifically, the nucleotides carry a base-unique fluorescent label and the 3'-OH group is chemically blocked such that each incorporation is a unique event. An imaging step follows each base incorporation step, during which each flow cell lane is imaged in segments by the instrument optics. After each imaging step, the 3' blocking group is chemically removed to prepare each strand for the next incorporation by DNA polymerase. This series of steps continues for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 50–250 bases. A base-calling algorithm assigns sequences and associated quality values to each read and a quality checking pipeline evaluates the Illumina data from each run.

The figure on the following page summarises the process:



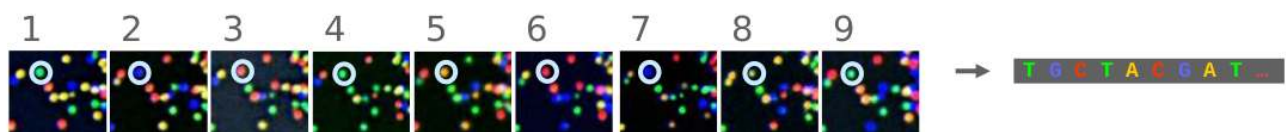
The Illumina sequencing-by-synthesis approach: Cluster strands created by bridge amplification are primed and all four fluorescently labelled, 3'-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3'-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

Next-Generation DNA Sequencing Methods Mardis, E.R. *Annu. Rev. Genomics Hum. Genet.* 2008. 9:387–402

Base-calling:

Base-calling involves evaluating the raw intensity values for each fluorophore and comparing them to determine which base is actually present at a given position during a cycle. To call bases on the Illumina or SOLiD platform, the positions of clusters need to be identified during the first few cycles. This is because they are formed in random positions on the flowcell as the annealing process is stochastic. This is in contrast to the Ion Torrent/454 system where the position of each cluster is defined by steel plate with pico-litre sized holes in which the reaction takes place.

If there are too many clusters, the edges of the clusters will begin to merge and the image analysis algorithms will not be able to distinguish one cluster from another (remember, the software is dealing with upwards of half a million clusters per square millimeter – that's a lot of dots!).



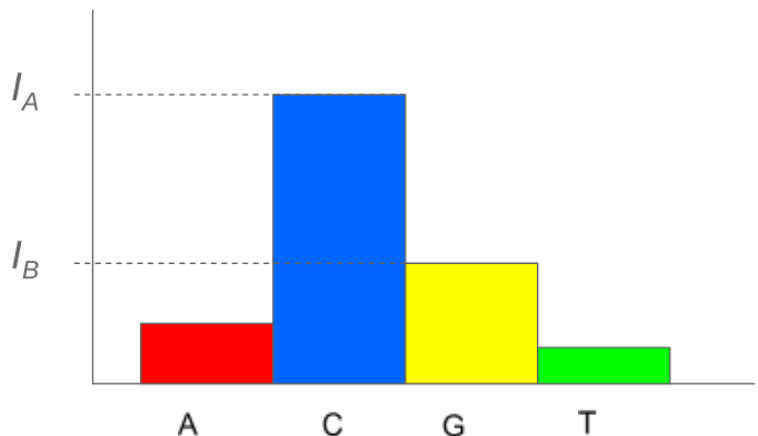
The above figure illustrates the principles of base-calling from cycles 1 to 9. If we focus on the highlighted cluster, one can observe that the colour (wavelength) of light observed at each cycle changes along with the brightness (intensity). This is due to the incorporation of complementary ddNTPs containing fluorophores. So at cycle 1 we have a T base, at 2 a G base and so on. If the colour or intensity is ambiguous the sequencer will mark it as an N. Other clusters are also visible in the images; these will represent different monoclonal clusters with different sequences.

The base calling algorithms turn the raw intensity values into T,G,C,A or N base calls. There are a variety of methods to do this and the one mentioned here is by no means the only one available, but it is often used as the default method on the Illumina systems. Known as the 'Chastity filter' it will only call a base if the intensity divided by the sum of the highest and second highest intensity is less than a given threshold (usually 0.6). Otherwise the base is marked with an N. In addition the standard Illumina pipeline will reject an entire read if two or more of these failures occur in the first 4 bases of a read (it uses these cycles to determine the boundary of a cluster).

Note that these basecalling processes are carried out at the sequencing facility and you will not need to perform any of these tasks under normal circumstances. They are explained here as useful background information.

CHASTITY formula:

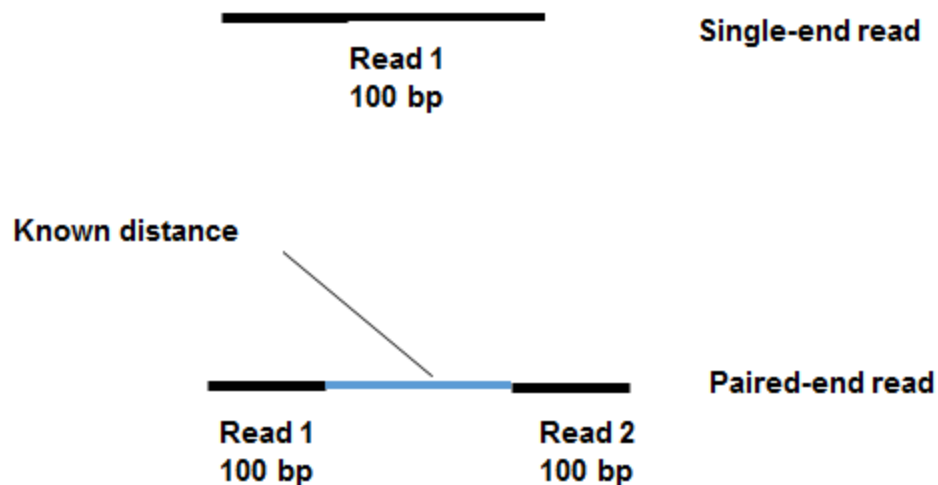
$$C = \frac{I_A}{I_A + I_B}$$



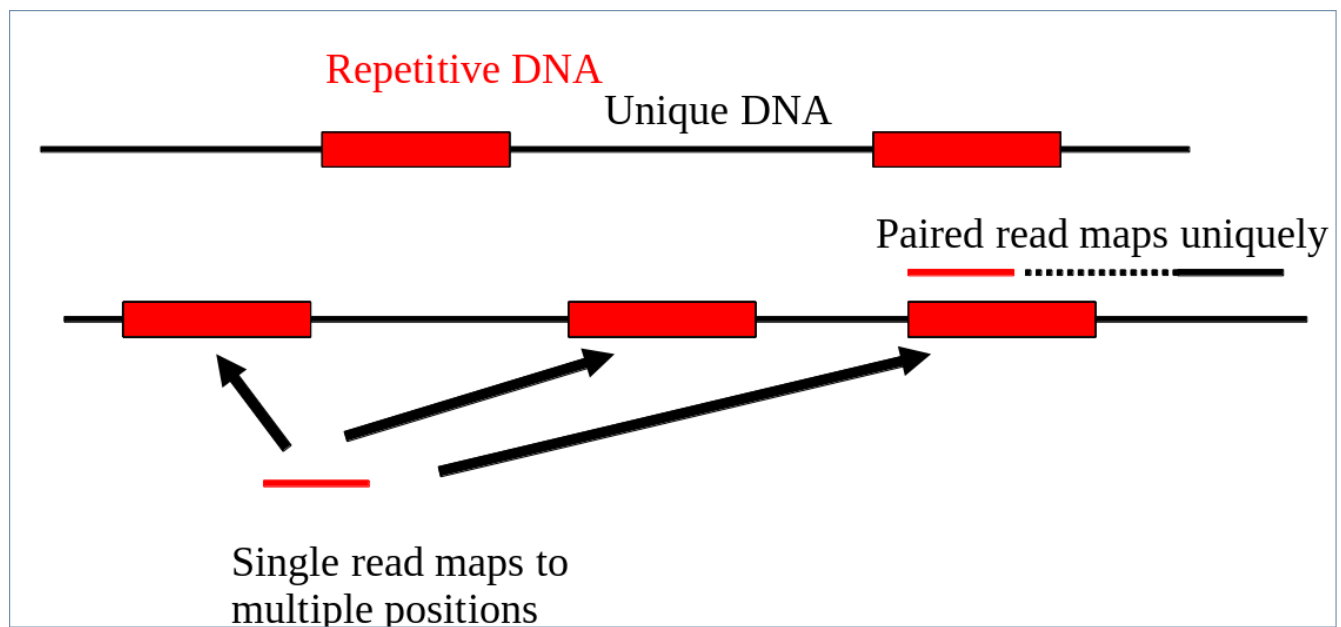
What are paired-end reads and why are they necessary?

Paired-end sequencing is a remarkably simple and powerful modification to the standard sequencing protocol. It is nearly always worth obtaining paired-end reads if performing genomic sequencing. Typically sequencers of any type are only able to sequence a portion of DNA (e.g. 50-300bp in the case of Illumina) before the fidelity of the enzyme and de-phasing of clusters (see later) increase the error rate beyond tolerable levels. As a result, on the Illumina system, a fragment which is 500bp long will may only have the first 100bp sequenced.

If the size selection is tight enough and you know that nearly all the fragments are close to 500bp long, you can repeat the sequencing reaction from the other end of the fragment. This will yield two reads for each DNA fragment separated by a known distance I.e:



The added information gained by knowing the distance between the two reads can be invaluable for spanning repetitive regions. In the figure below, the light coloured regions indicate repetitive sections of DNA. If a read contains only repetitive DNA, an alignment algorithm will be able to align the read to many locations in a reference genome. However, with paired-end reads, there is a greater chance that at least one of the two reads will align to a unique region of DNA. In this way one of the reads can be used to anchor the other read in the pair and help resolve the repetitive region. Paired-end reads are often used when performing de-novo genome sequencing (i.e. when a reference is not available to align against) because they enable contiguous regions of DNA to be ordered, or when characterizing variants such as large insertions or deletions.



Other forms of paired-end sequencing with much larger distances (e.g. 10kb) are possible with so called 'mate-pair' libraries. These are usually used in specific projects to help order contigs in de-novo sequencing projects. We will not cover them here, but the principles behind them are similar.

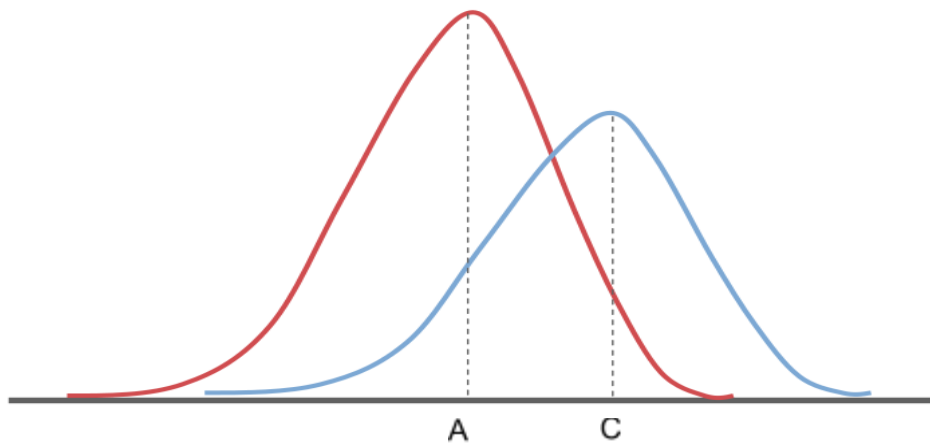
In some cases, such as amplicon sequencing of portions of conserved genes (e.g 16S/18S subunits), it is advantageous to be able to overlap read 1 and read 2. For instance a particular amplicon may have a length of 500bp. If sequenced on the Illumina MiSeq in paired-end 300bp mode, the middle of the amplicon would be sequenced twice reducing the overall likelihood of including erroneous base-calls. Software exists which enables read 1 and read 2 to be merged although no independent benchmark has been performed to evaluate how well the software works.

Inherent sources of error

No measurement is without a certain degree of error. This is true in sequencing. As such there is a finite probability that a base will not be called correctly. There are several possible sources:

Frequency cross-talk and normalisation errors:

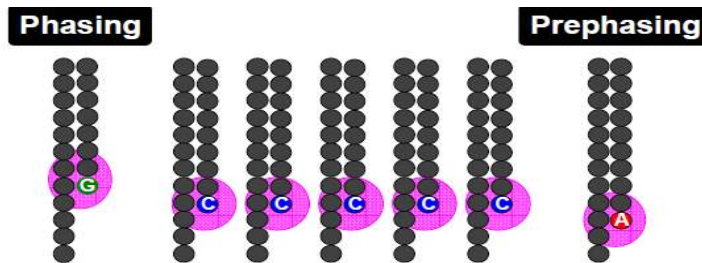
When reading an A base, a small amount of C will also be measured due to frequency overlap and vice-versa. Similarly with G and T bases. Additionally, from the figure below, it should be clear that the extent to which the dyes fluoresce differs. As such it is necessary to normalize the intensities. This normalisation process can also introduce errors.



Frequency response curve for A and C dyes
(Intensity y-axis and frequency on the x-axis)

Phasing/Pre-phasing:

This occurs when a strand of DNA lags or leads the other DNA strands within a cluster. This introduces additional background noise into the signal and reduces the intensity of the true base. In the example below we have a cluster with 7 strands of DNA (very small cluster, but this is just an example). Five strands are on a C-base, whilst 1 is lagging behind (called phasing) on a G base and the remaining strand is running ahead of the pack (confusingly called pre-phasing) on an A base. As such the C signal will be reduced and A and G boosted for the rest of the sequencing run. Too much phasing or pre-phasing (i.e. > 15-20%) usually causes problems for the base calling algorithm and result in clusters being filtered out.



Other issues:

- **Biases introduced by sample preparation** – your sequencing is only as good as your experimental design and DNA extraction. Also, remember that your sample will be put through several cycles of PCR before sequencing. This also introduces a potential source of bias.
- **High AT or GC content sequences** – this reduces the complexity of the sequence and can result in higher error rates. Illumina sequencers used to suffer from this greatly but software improvements post August 2014 have reduced the impact of low complexity sequence.
- **Homopolymeric sequences** – long stretches of a single base can make it difficult to determine phasing and pre-phasing rates. This can introduce errors in determining the precise length of a homopolymeric stretch of sequence. This much more of a problem on the 454 and Ion Torrent than Illumina platforms but still worth bearing in mind. Especially if you encounter indels which have been called in homopolymeric tracts.
- Some motifs can cause loops and other steric clashes

See Nakamura et al, *Sequence-specific error profile of Illumina sequencers* *Nuc. Acid Res.* first published online May 16, 2011 doi:10.1093/nar/gkr344

Quality scores

To account for the possible errors and provide an estimate of confidence in a given base-call, the Illumina sequencing pipeline assigns a quality score to each base called. Most quality scores are calculated using the Phred scale. Each base call has an associated base call quality which estimates the chance that the base call is incorrect.

Q10 = 1 in 10 chance of incorrect base call

Q20 = 1 in 100 chance of incorrect base call

Q30 = 1 in 1000 chance of incorrect base call

Q40 = 1 in 10,000 chance of incorrect base call

For most Ion Torrent, SOLiD and Illumina runs you should see quality scores between Q20 and Q40. Note that these are only estimates of base-quality based on calibration runs performed by the manufacturer against a sample of known sequence with (typically) a GC content of 50%. Extreme GC bias and/or particular motifs or homopolymers can cause the quality scores to become unreliable.

Accurate base qualities are an essential part in ensuring variant calls are correct. For the purposes of this tutorial and as a rough and ready rule we generally assume that with Illumina data anything less than Q20 is not useful data and should be excluded from the analysis.

Reads containing adaptors

Some reads will contain adaptor sequences after sequencing, usually at the end of the read. This is usually because of short sample DNA fragments, which result in the polymerase reading into the adaptor region. Occasionally this can also happen because of mis-priming. It is important to remove or trim sequences containing these sequences as they are not part of the data you are interested in and can prevent reads from aligning or being correctly assembled.

2015 Workshop on Genomics

Part 2: Short read genomics: Remapping

Instructors:

- Konrad Paszkiewicz k.h.paszkiewicz@exeter.ac.uk

Objectives:

By the end of this section you will be expected to:

- Interpret FASTQ quality metrics.
- Remove poor quality data.
- Trim adaptor/contaminant sequences from FASTQ data.
- Count the number of reads before and after trimming and quality control.
- Align reads to a reference sequence to form a SAM file (Sequence AlignMent file) using BWA.
- Convert the SAM file to BAM format (Binary AlignMent format).
- Identify and select high quality SNPs and Indels using SAMtools.
- Identify missing or truncated genes with respect to the reference genome.
- Identify SNPs which overlap with known coding regions.

Introduction

In this section of the lab we will be analysing a strain of *E.coli* which was sequenced at Exeter. It is closely related to the K-12 substrain MG1655 (<http://www.ncbi.nlm.nih.gov/nuccore/U00096>).

We want to obtain a list of single nucleotide polymorphisms (SNPs), insertions/deletions (Indels) and any genes which have been deleted.

Quality control

In this section of the lab we will be learning about evaluating the quality of an Illumina MiSeq sequencing run. The process described here can be used with any FASTQ formatted file from any platform (e.g Illumina, Ion Torrent, PacBio etc).

2nd (and 3rd) generation sequencers produce vast quantities of data. A single Illumina MiSeq lane will produce over 10 Gbases of data. However, the error rates of these platforms are 10-100x higher than Sanger sequencing. They also have very different error profiles. Unlike Sanger sequencing, where the most reliable sequences tend to be in the middle, NGS platforms tend to be most reliable near the beginning of each read.

Quality control usually involves:

- Calculating the number of reads before quality control
- Calculating GC content, identifying over-represented sequences
- Remove or trim reads containing adaptor sequences
- Remove or trim reads containing low quality bases
- Calculating the number of reads after quality control
- Rechecking GC content, identifying over-represented sequences

Quality control is necessary because:

- CPU time required for alignment and assembly is reduced
- Data storage requirements are reduced
- Reduces potential for bias in variant calling and/or de-novo assembly

Quality scores:

Most quality scores are calculated using the Phred scale (*Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194 (1998)*). Each base call has an associated base call quality which estimates chance that the base call is incorrect.

Q10 = 1 in 10 chance of incorrect base call

Q20 = 1 in 100 chance of incorrect base call

Q30 = 1 in 1000 chance of incorrect base call

Q40 = 1 in 10,000 chance of incorrect base call

For most Ion Torrent, SoLID and Illumina runs you should see quality scores between Q20 and Q40.

Note that these are only estimates of base-quality based on calibration runs performed by the manufacturer against a sample of known sequence with (typically) a GC content of 50%. Extreme GC biases and/or particular motifs or homopolymers can cause the quality scores to become unreliable. Accurate base qualities are an essential part in ensuring variant calls are correct. As a rough and ready rule we generally assume that with Illumina data anything less than Q20 is not useful data and should be excluded.

FASTQ format:

A FASTQ entry consists of 4 lines

1. A header line beginning with '@' containing information about the name of the sequencer, and the position at which the originating cluster was located and whether it passed purity filters.
2. The DNA sequence of the read
3. A header line or line beginning with just '+'
4. Quality scores for each base encoded in ASCII format

Typical FASTQ formatted file.

```
@D3P26HQ1:110:d0eh1acxx:8:1101:1116:2122 1:N:0:
AGGTGTCTCCTACAACCAAGCTACAACAGCAATGGGCTATCTGGTGGGATTAAAGGGGTGAAAATGCATCCCCCTTAAATNAAAGTGGTTTT
+
ADDADCFHHHDHGHIII<GIICH4FGCIHIEGFHGHGIIIGDHFDFG?DEHH>FGIG=E@GGADDCCCC@A>ABB>BBC:A>A#,228(4>:??B
```

To reduce storage requirements, the FASTQ quality scores are stored as single characters and converted to numbers by obtaining the ASCII quality score and subtracting either 33 or 64. For example, the above FASTQ file is Sanger formatted and the character '!' has an ASCII value of 33. Therefore the corresponding base would have a Phred quality score of $33-33=Q0$ (i.e. totally unreliable). On the other hand a base with a quality score denoted by '@' which has an ASCII value of 64 would have a Phred quality score of $64-33=Q31$ (i.e. less than 1/1000 chance of being incorrect).

Just to confuse matters, there are several different methods of encoding quality scores in the ASCII format:

```

.....
.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
.....LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [ \ ] ^ _ ` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~
|
33 | 59 | 64 | 73 | 104 | 126
0 .....26...31.....40
-5 .....0.....9.....40
0 .....9.....40
3 .....9.....40
0.2 .....26...31.....41

S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

```

Note that since June 2011 Illumina sequencers have generally provided output in fastq-sanger format rather than Illumina 1.3+. Thus Illumina 1.3+ and other Illumina scoring metrics are unlikely to be encountered if you are using Illumina sequencing data generated after July 2011. If you are uncertain, the presence of a number in the quality scores generally suggests Illumina 1.8+.

Quality control – evaluating the quality of Illumina data

The first task when one receives sequencing data is to evaluate its quality and determine whether all the cash you have handed over was well-spent! To do this we will use the FastQC toolkit (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>). FastQC offers a graphical visualisation of QC metrics, but *does not* have the ability to filter data.

Task 1:

From your home directory change into the
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/ directory and list the directory contents. E.g.:

```
cd ~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/  
ls -l
```

```
drwxrwxr-x  3 ubuntu ubuntu      4096 Jan  9 17:04 blast_precompute  
drwxrwxr-x 11 ubuntu ubuntu      4096 Jan  9 16:47 denovo_assembly  
-rw-r--r--  1 ubuntu ubuntu 426091067 Jan  6 01:44 E_Coli_CGATGT_L001_R1_001.fastq  
-rw-r--r--  1 ubuntu ubuntu 426091067 Jan  6 01:44 E_Coli_CGATGT_L001_R2_001.fastq
```

Note that this is a paired-end run. As such there are two files

- one for read 1 (E_Coli_CGATGT_L001_R1_001.fastq)
- and the other for the reverse read 2 (E_Coli_CGATGT_L001_R2_001.fastq)

Reads from the same pair can be identified because they have the same header. Many programs require that the read 1 and read 2 files have the reads in the same order. To view the first few headers we can use the head and grep commands:

```
head E_Coli_CGATGT_L001_R1_001.fastq | grep MISEQ  
head E_Coli_CGATGT_L001_R2_001.fastq | grep MISEQ
```

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ head E_Coli_CGATGT_L001_R1_001.fastq | grep MISEQ  
@MISEQ:8:000000000-A7VC1:1:1101:14839:1482 1:N:0:CGATGT  
@MISEQ:8:000000000-A7VC1:1:1101:18239:1496 1:N:0:CGATGT  
@MISEQ:8:000000000-A7VC1:1:1101:13371:1512 1:N:0:CGATGT  
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ head E_Coli_CGATGT_L001_R2_001.fastq | grep MISEQ  
@MISEQ:8:000000000-A7VC1:1:1101:14839:1482 2:N:0:CGATGT  
@MISEQ:8:000000000-A7VC1:1:1101:18239:1496 2:N:0:CGATGT  
@MISEQ:8:000000000-A7VC1:1:1101:13371:1512 2:N:0:CGATGT
```

The only difference in the headers for the two reads is the read number. Of course this is no guarantee that all the headers in the file are consistent. To get some more confidence repeat the above commands using 'tail' instead of 'head' to compare reads at the end of the files.

You can also check that there is an identical number of reads in each file using cat, grep and wc -l:

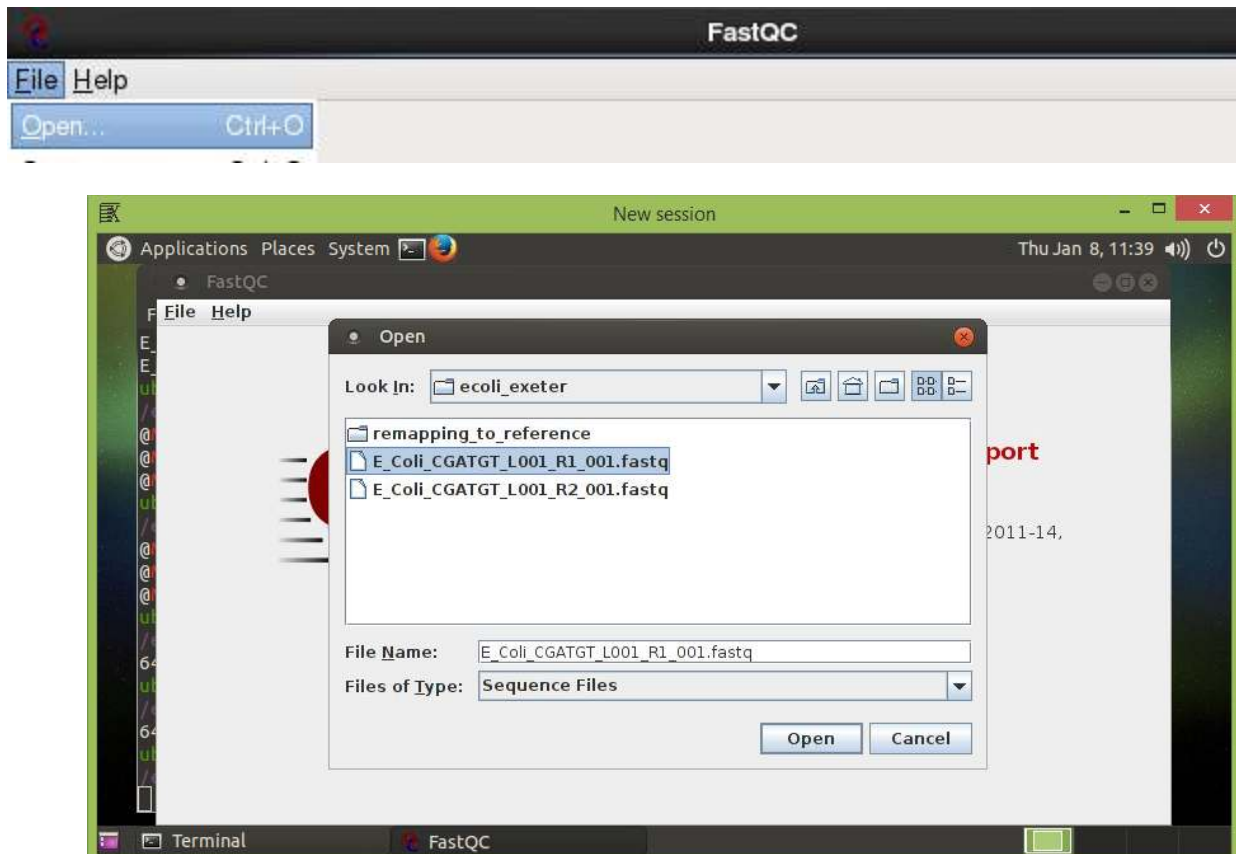
```
cat E_Coli_CGATGT_L001_R1_001.fastq | grep MISEQ | wc -l  
cat E_Coli_CGATGT_L001_R2_001.fastq | grep MISEQ | wc -l
```

Don't worry if this takes a minute or so, remember that the computer is reading and scanning thousands of lines one at a time.

Now, let's start the fastqc program.

`fastqc`

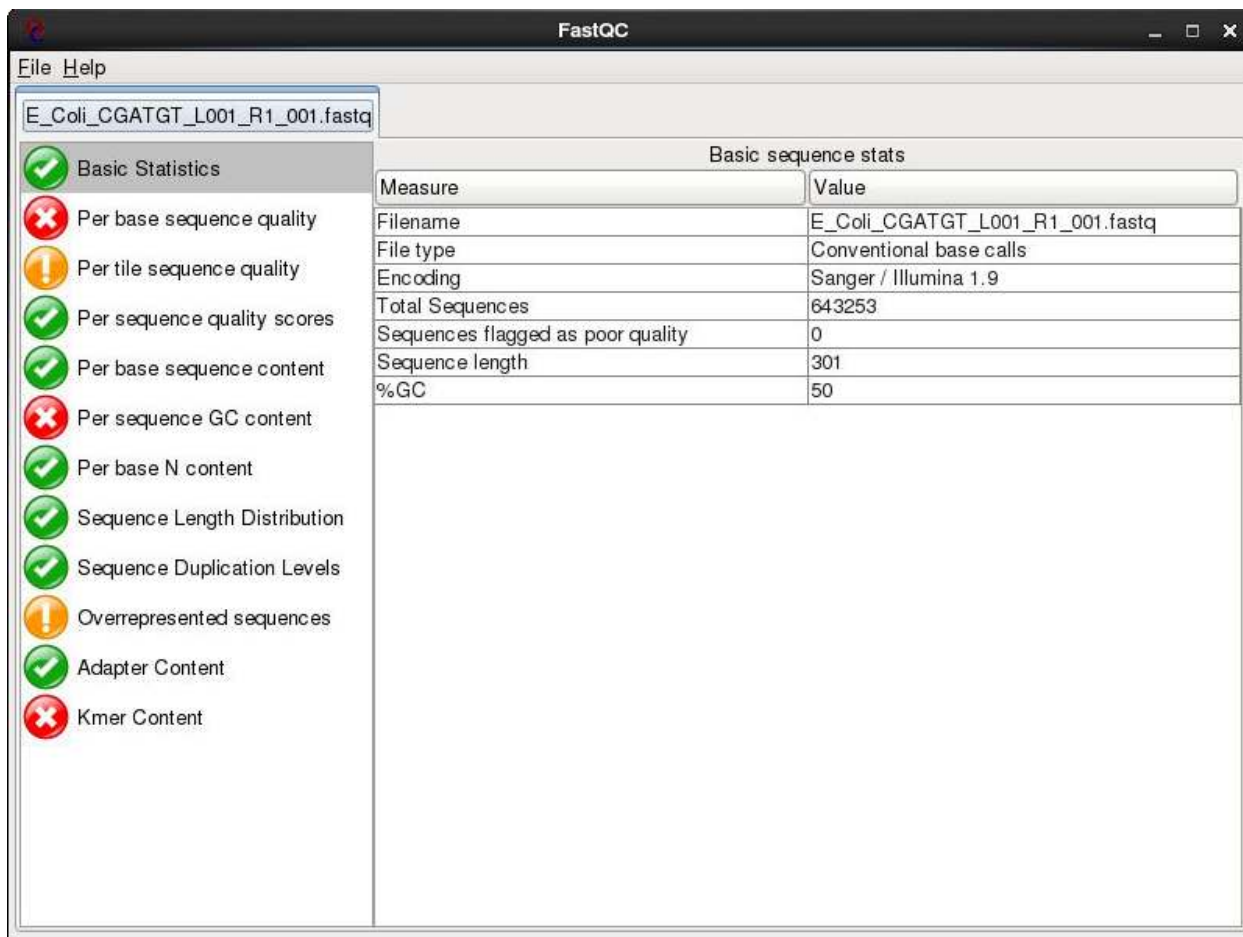
Load the E_Coli_CGATGT_L001_R1_001.fastq file from the
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter directory.



After a few minutes the program should finish analysing the FASTQ file.

The fastqc program performs a number of tests which determines whether a green tick (pass), exclamation mark (warning) or red cross (fail) is displayed. However it is important to realise that fastqc has no knowledge of what your library is or should look like. All of its tests are based on a completely random library with 50% GC content. Therefore if you have a sample which does not match these assumptions, it may 'fail' the library. For example, if you have a high AT or high GC organism it may fail the per sequence GC content. If you have any barcodes or low complexity libraries (e.g. small RNA libraries) they may also fail some of the sequence complexity tests.

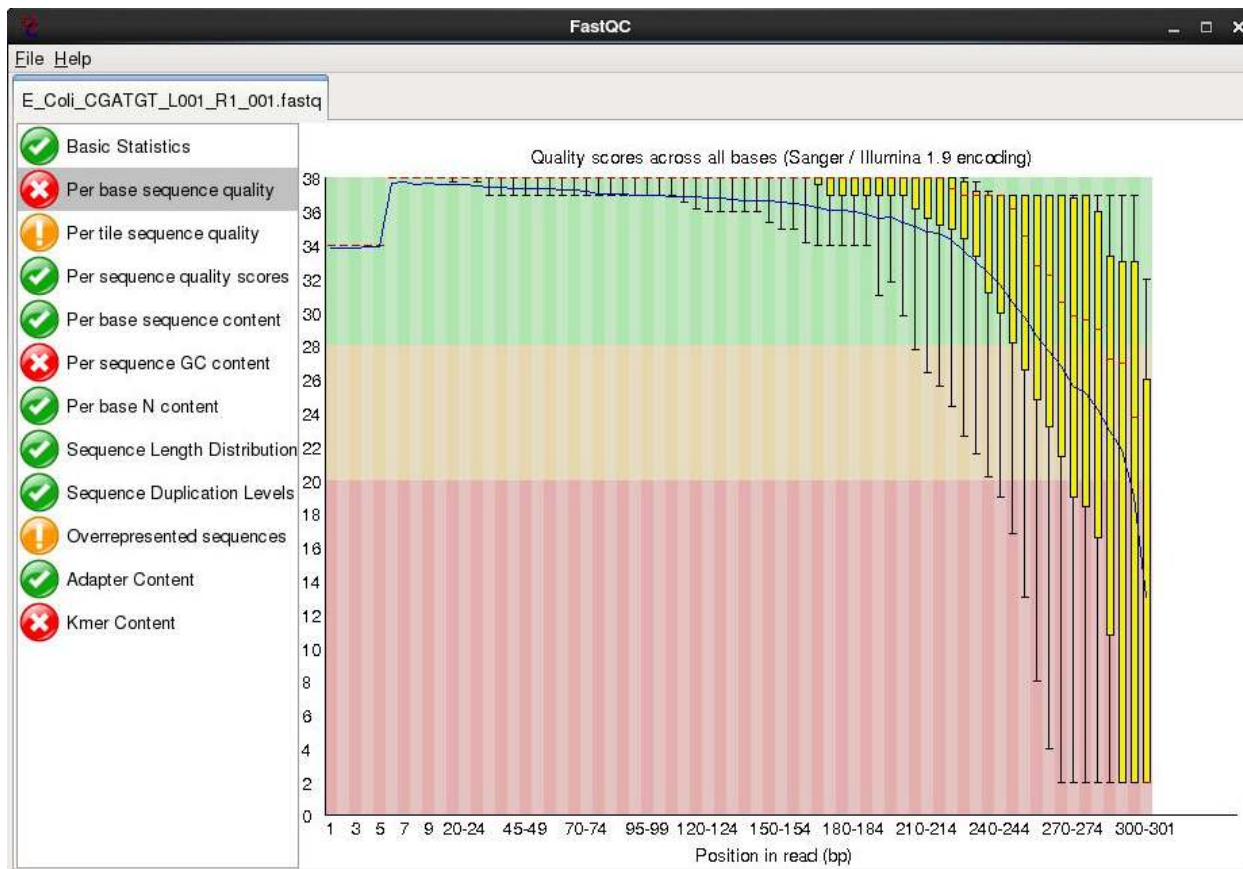
The bottom line is that you need to be aware of what your library is and whether what fastqc is reporting makes sense for that type of library.



In this case we have a number of errors and warnings which at first sight suggest there has been a problem - but don't worry too much yet. Let's go through them in turn.

Quality scores:

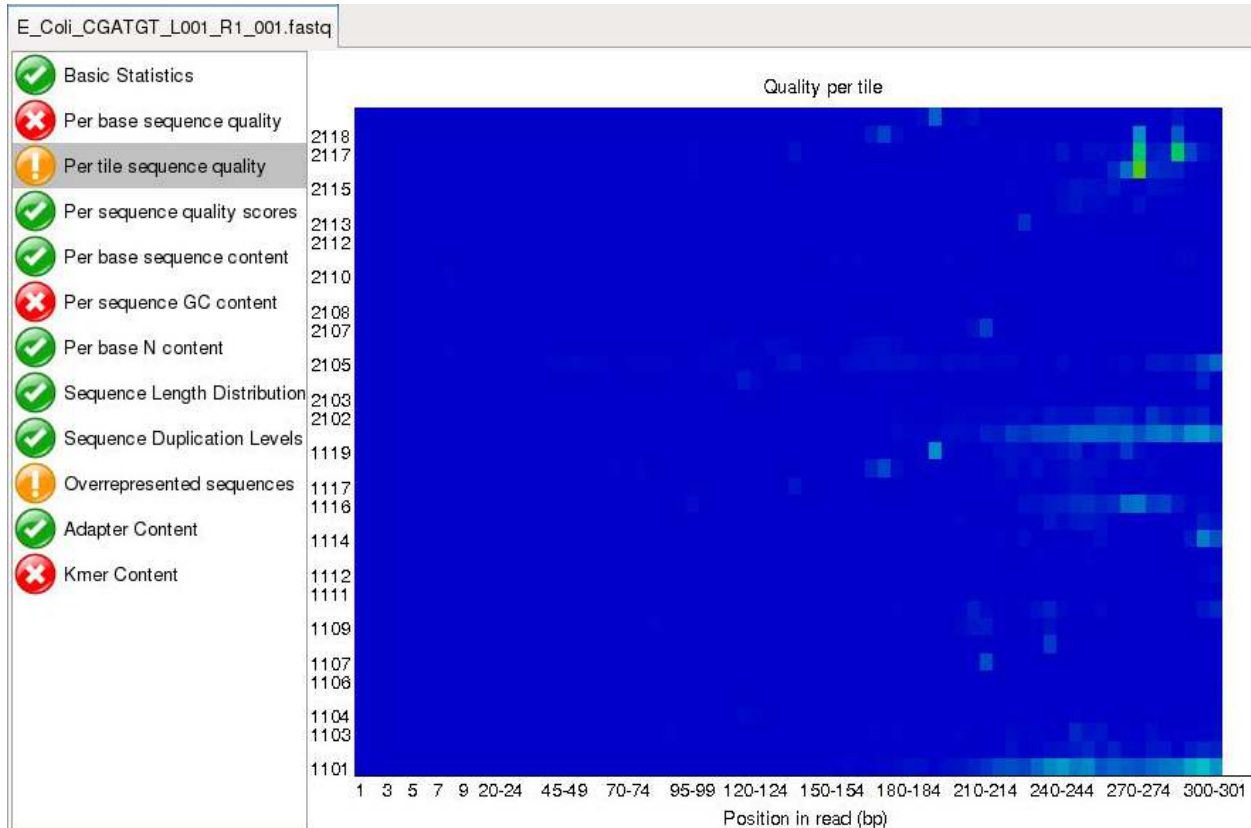
This is one of the most important metrics. If the quality scores are poor, either the wrong FASTQ encoding has been guessed by fastqc (see the title of the chart), or the data itself is poor quality. This view shows an overview of the range of quality values across all bases at each position in the FASTQ file. Generally anything with a median quality score greater than Q20 is regarded as acceptable; anything above Q30 is regarded as 'good'. For more details, see the help documentation in fastqc.



In this case this check is red - and it is true that the quality drops off at the end of the reads. It is normal for read quality to get worse towards the end of the read. You can see that at 250 bases the quality is still very good, we will later trim off the low quality bases so reserve judgment for now..

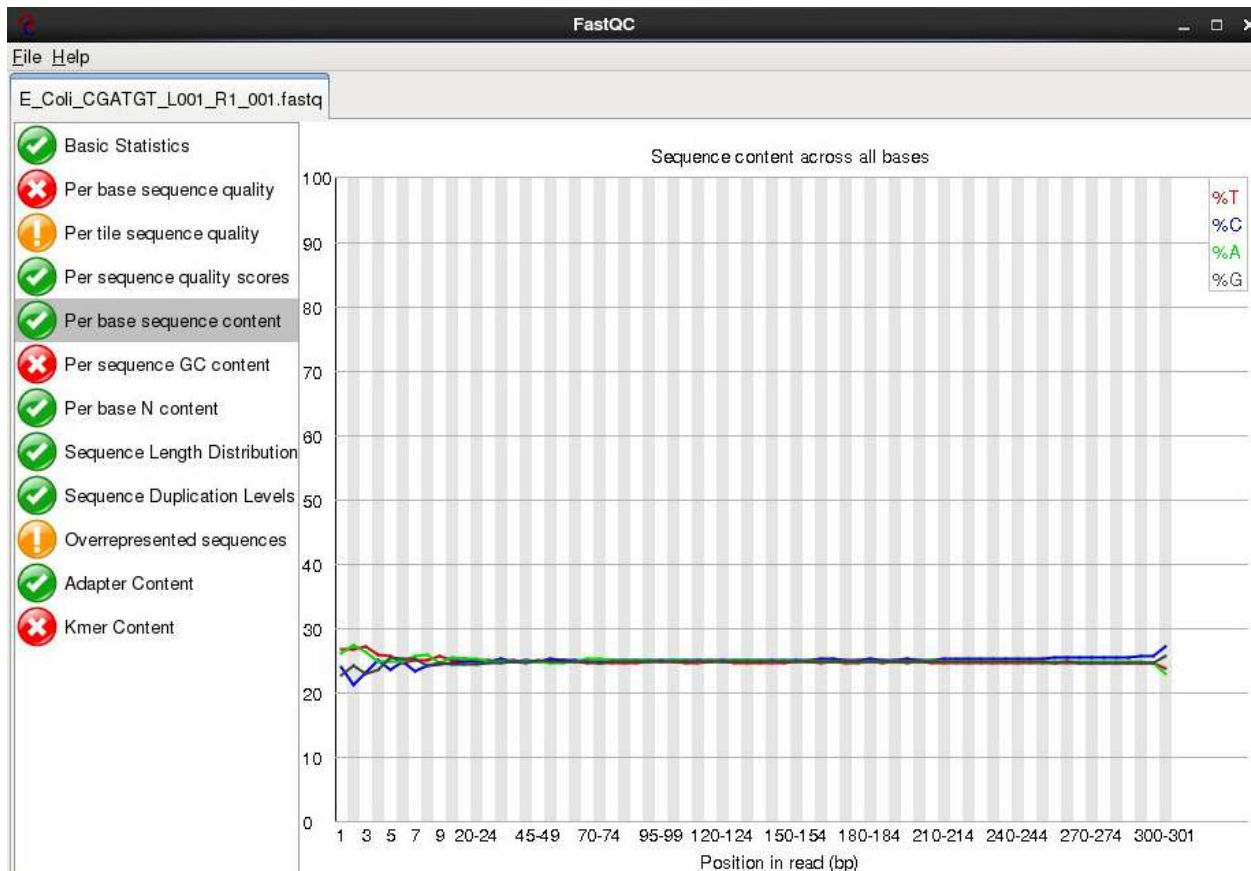
Per tile sequence quality

This is a purely technical view on the sequencing run, it is more important for the team running the sequencer. The sequencing flowcell is divided up into areas called cells. You can see that the read quality drops off in some cells faster than others. This maybe because of the way the sample flowed over the flowcell or a mark or smear on the lens of the optics.



Per base sequence content:

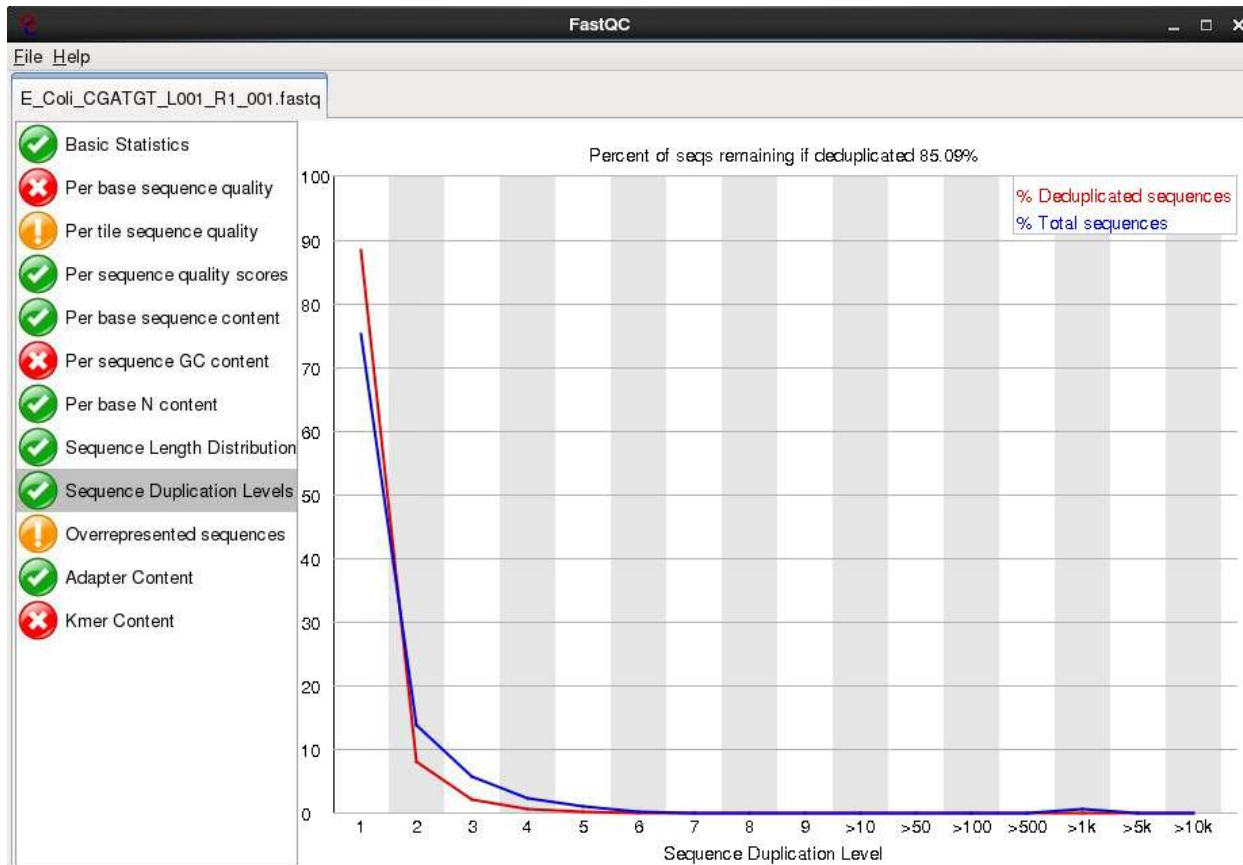
For a completely randomly generated library with a GC content of 50% one expects that at any given position within a read there will be a 25% chance of finding an A,C,T or G base. Here we can see that our library satisfies these criteria, although there appears to be some minor bias at the beginning of the read. This may be due to PCR duplicates during amplification or during library preparation. It is unlikely that one will ever see a perfectly uniform distribution. See http://ess-wiki.exeter.ac.uk/ess/Quality_control for examples of good vs bad runs as well as the fastqc help for more details.



Sequence duplication levels:

In a library that covers a whole genome uniformly most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (e.g. PCR over-amplification).

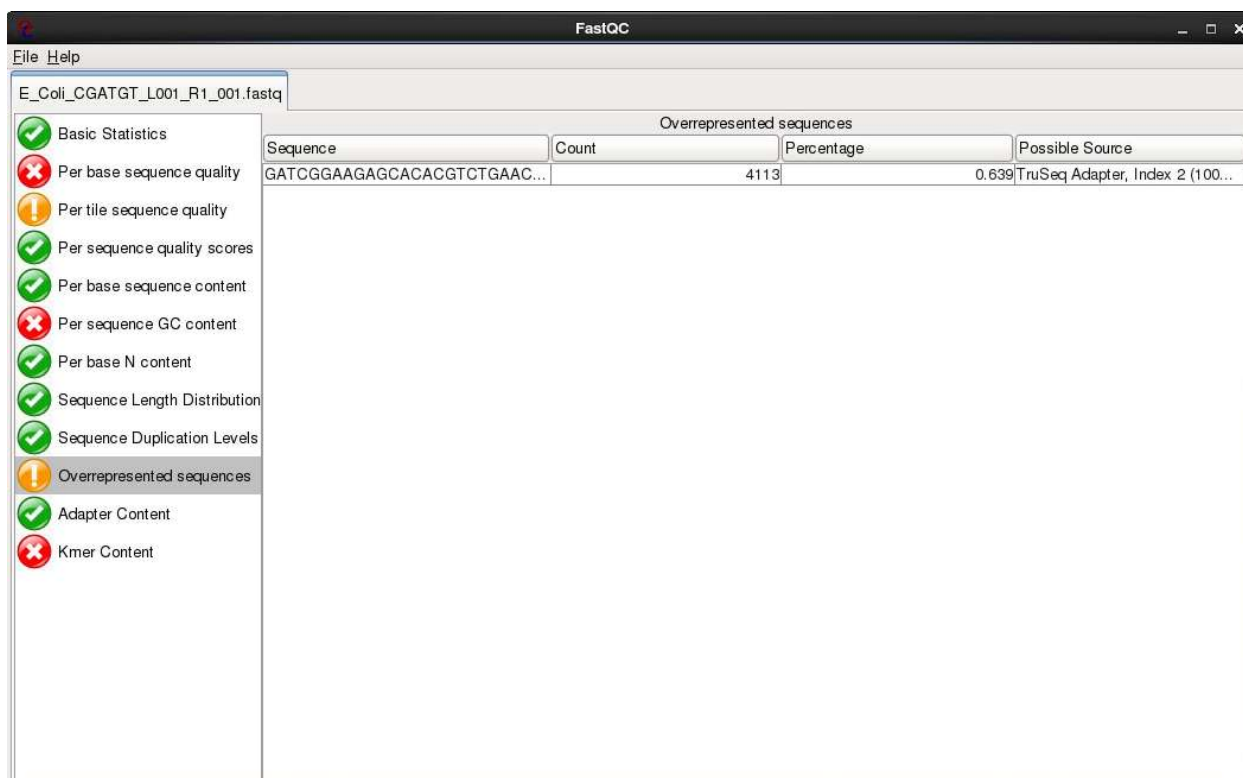
This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.



Overrepresented Sequences

This check for sequences that occur more frequently than expected in your data. It also checks any sequences it finds against a small database of known sequences. In this case it has found that a small number of reads 4000 out of 600000 appear to contain a sequence used in the preparation for the library. (Don't worry that the percent of 0.639 seems incorrect - a subset of reads are analysed to conserve memory). A typical cause is that the original DNA was shorter than the length of the read - so the sequencing overruns the actual DNA and runs into the adaptors used to bind it to the flowcell.

At this level there is nothing to worry about - they will be trimmed in later stages.



There are other reports available:

Have a look at them and at what the author of FastQC has to say.

<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>

Remember the error and warning flags are his (albeit experienced) judgement of what typical data should look like. It is up to you to use some initiative and understand whether what you are seeing is typical for your dataset and how that might affect any analysis you are performing. Bear in mind that some data, such as transcriptome data, is likely to have over represented sequences or some data, such as RAD-seq, will always have the same few base pairs (the digestion site). The important thing is to think about your data!

Task 2:

Do the same for read 2 as we have for read 1. Open fastqc and analyse the read 2 file. Look at the various plots and metrics which are generated. How similar are they?

Note that the number of reads reported in both files is identical. This is because if one read fails to pass the Illumina chastity filter, its partner is automatically excluded too.

Overall, both read 1 and read 2 can be regarded as 'good' data-sets.

Quality control – filtering of Illumina data

In this section we will be filtering the data to ensure any low quality reads are removed and that any sequences containing adaptor sequences are either trimmed or removed altogether. To do this we will use the fastq-mcf program from the ea-utils package (this is installed already on the cloud, but for reference is available at <http://code.google.com/p/ea-utils/>). This package is remarkably fast and ensures that after filtering both read 1 and read 2 files are in the correct order.

Note: Typically when submitting raw Illumina data to NCBI or EBI you would submit unfiltered data, so don't delete your original fastq files!

Make sure you are in the `~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/` directory. We will execute the fastq-mcf program which performs both adaptor sequence trimming and low quality bases (instructions are in the next task). To remove adaptor sequences, we need to supply the adaptor sequences to the program. A list of the most common adaptors used is given in the file `~/workshop_data/genomics_tutorial/data/reference/adaptors/adaptors.fasta` :



```
ec2-user@ip-10-181-81-253:~/genomics_tutorial/data$ more ~/genomics_tutorial/data/reference/adaptors/adaptors.fasta
>Nextera_enrichment
CTGTCTCTTATACACATCT
>TruSeq_Read1
AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
>TruSeq_Read2
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
>Nextera_mate_pair_Read1
CTGTCTCTTATACACATCT
>Nextera_mate_pair_Read2
AGATGTGTATAAGAGACAG
>PolyA
AAAAAAAAAAAAAAAAAAAAAAAAA
>PolyT
TTTTTTTTTTTTTTTTTTTTTTTTT
>PolyC
CCCCCCCCCCCCCCCCCCCCCCCCC
>PolyG
GGGGGGGGGGGGGGGGGGGGGGGGG
>Illumina_Nextera_v2_kmer_701
CGAGATTCGCTTA
>Illumina_Nextera_v2_kmer_702
CGAGATCTAGTACG
>Illumina_Nextera_v2_kmer_703
```

Task 3:

To run the fastq-mcf program, type the following (all on one line):

```
fastq-mcf ../../reference/adaptors/adaptors.fasta E_Coli_CGATGT_L001_R1_001.fastq  
E_Coli_CGATGT_L001_R2_001.fastq -o E_Coli_CGATGT_L001_R1_001.filtered.fastq -o  
E_Coli_CGATGT_L001_R2_001.filtered.fastq -C 1000000 -q 20 -p 10 -u -x 0.01
```

To explain the options:

-o are the output files to be written.

-C specifies the number of reads for subsampling - this takes the results from a smaller number of reads to create a model, and then trims the entire data set based upon this. This setting speeds up the filtering process.

-p specifies the maximum adaptor difference - this is the amount the adaptor can differ by and still be removed.

-u this enables Illumina PF filtering - this is obsolete for all Illumina data past 2011.

-x specifies the bad read percentage causing cycle removal - if the number of reads with a "bad" quality score at that position exceeds this percentage, the position is removed from all reads.

This may take around 20 minutes to finish (go and have a break!) After the filtering should be complete and you should see something similar to:

```
[ec2-user@ip-10-169-87-62 ec2-user]$ fastq-mcf ../../reference/adaptors/adaptors.fasta E_Coli_CGATGT_L001_R1_0  
01.fastq E_Coli_CGATGT_L001_R2_001.fastq -o E_Coli_CGATGT_L001_R1_001.filtered.fastq -o E_Coli_CGATGT_L001_R2_001.  
filtered.fastq -C 1000000 -q 20 -p 10 -u -x 0.01  
Command Line: ../../reference/adaptors/adaptors.fasta E_Coli_CGATGT_L001_R1_001.fastq E_Coli_CGATGT_L001_R2_001.f  
astq -o E_Coli_CGATGT_L001_R1_001.filtered.fastq -o E_Coli_CGATGT_L001_R2_001.filtered.fastq -C 1000000 -q 20 -p 10  
-u -x 0.01  
Scale used: 2.2  
Filtering Illumina reads on purity field  
Phred: 33  
Threshold used: 1609 out of 643253  
Adapter TruSeq_Read1 (AGATCGGAAGAGCACACGTCTGAACTCCAGTCA): counted 8548 at the 'end' of 'E_Coli_CGATGT_L001_R1_001.  
fastq', clip set to 5  
Adapter TruSeq_Read2 (AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT): counted 6204 at the 'end' of 'E_Coli_CGATGT_L001_R2_001.  
fastq', clip set to 5  
Adapter Short Nextera fragment of adaptor (TCGGAAGAGCACACGT): counted 12634 at the 'end' of 'E_Coli_CGATGT_L001_R1  
_001.fastq', clip set to 4  
Adapter Nextera_read_1_external_adapter (ATCGGAAGAGCACACGTCTGAACTCCAGTCAC): counted 12786 at the 'end' of 'E_Coli_  
CGATGT_L001_R1_001.fastq', clip set to 4  
Adapter Nextera_read_2_external_adapter (GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT): counted 6254 at the 'end' of 'E_Coli_C  
GATGT_L001_R2_001.fastq', clip set to 5  
  
Too short after clip: 8895  
Clipped 'end' reads (E_Coli_CGATGT_L001_R1_001.fastq): Count 18042, Mean: 34.11, Sd: 40.08  
Trimmed 505652 reads (E_Coli_CGATGT_L001_R1_001.fastq) by an average of 16.89 bases on quality < 20  
Clipped 'end' reads (E_Coli_CGATGT_L001_R2_001.fastq): Count 9426, Mean: 52.99, Sd: 40.18  
Trimmed 621151 reads (E_Coli_CGATGT_L001_R2_001.fastq) by an average of 60.69 bases on quality < 20
```

You can see that the trimming has been harsher on the R2 reads than on the R1 - this is generally to be expected in Illumina paired end runs.

If we look at the sizes of the files produced:

```
ls -l
```

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ ls -l
total 1568644
-rw-r--r--. 1 ec2-user ec2-user 426091067 Dec  1 10:46 E_Coli_CGATGT_L001_R1_001.fastq
-rw-rw-r--. 1 ec2-user ec2-user 405632367 Dec  1 13:33 E_Coli_CGATGT_L001_R1_001.filtered.fastq
-rw-r--r--. 1 ec2-user ec2-user 426091067 Dec  1 11:21 E_Coli_CGATGT_L001_R2_001.fastq
-rw-rw-r--. 1 ec2-user ec2-user 348453609 Dec  1 13:33 E_Coli_CGATGT_L001_R2_001.filtered.fastq
```

You can see that the original files are exactly the same size, but the R2 filtered file is smaller than R1.

Now count the lines in all the files

```
wc -l *.filtered.fastq
```

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ wc -l *.filtered.fastq
2537432 E_Coli_CGATGT_L001_R1_001.filtered.fastq
2537432 E_Coli_CGATGT_L001_R2_001.filtered.fastq
```

Although the reads have been trimmed differently - the number of reads in the R1 and R2 files are identical. This is required for all the tools we will use to analyse paired end data.

Task 4:

Check the quality scores and sequence distribution in the fastqc program for the two filtered fastq files. You should notice very little change (since comparatively few reads were filtered).

However, you should notice a significant improve in quality and the absence of adaptor sequences.

Task 5:

We can perform a quick check (although this by no means guarantees) that the sequences in read 1 and read 2 are in the same order by checking the ends of the two files and making sure that the headers are the same.

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ head E_Coli_CGATGT_L001_R1_001.filtered.fastq | grep MISEQ
@MISEQ:8:000000000-A7VC1:1:1101:17200:1633 1:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:1101:10456:1673 1:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:1101:16582:1688 1:N:0:CGATGT
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ head E_Coli_CGATGT_L001_R2_001.filtered.fastq | grep MISEQ
@MISEQ:8:000000000-A7VC1:1:1101:17200:1633 2:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:1101:10456:1673 2:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:1101:16582:1688 2:N:0:CGATGT
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ tail E_Coli_CGATGT_L001_R1_001.filtered.fastq | grep MISEQ
@MISEQ:8:000000000-A7VC1:1:2119:19669:25236 1:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:2119:10145:25237 1:N:0:CGATGT
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ tail E_Coli_CGATGT_L001_R2_001.filtered.fastq | grep MISEQ
@MISEQ:8:000000000-A7VC1:1:2119:19669:25236 2:N:0:CGATGT
@MISEQ:8:000000000-A7VC1:1:2119:10145:25237 2:N:0:CGATGT
```

Task 6:

Check the number of reads in each filtered file. They should be the same. To do this use the `grep` command to search for the number of times the header appears. E.g:

```
grep -c MISEQ E_Coli_CGATGT_L001_R1_001.filtered.fastq
```

Do the same for the `strain1_read2.filtered.fastq` file.

Aligning Illumina data to a reference sequence

Now that we have checked the quality of our raw data, we can begin to align the reads against a reference sequence. In this way we can compare how the reference sequence and the strain we have sequenced compare.

To do this we will be using a program called BWA (Burrows Wheeler Aligner *Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.*). This uses an algorithm called (unsurprisingly) Burrows Wheeler to rapidly map reads to the reference genome. BWA also allows for a certain number of mismatches to account for variants which may be present in strain 1 vs the reference genome. BWA allows for insertions or deletions as well (as do most modern short read aligners).

By mapping reads against a reference, what we mean is that we want to go from a FASTQ file listing lots of reads, to another type of file (which we'll describe later) which lists the reads AND where/if it maps against the reference genome. The figure below illustrates what we are trying to achieve here. Along the top in grey is the reference sequence. The coloured sequences below indicate individual sequences and how they map to the reference. If there is a real variant in a bacterial genome we would expect that (nearly) all the reads would contain the variant at the relevant position rather than the same base as the reference genome. Remember that error rates for any single read on second generation platforms tend to be around 0.5-1.5%. Therefore a 300bp read is likely to contain at least a few errors.

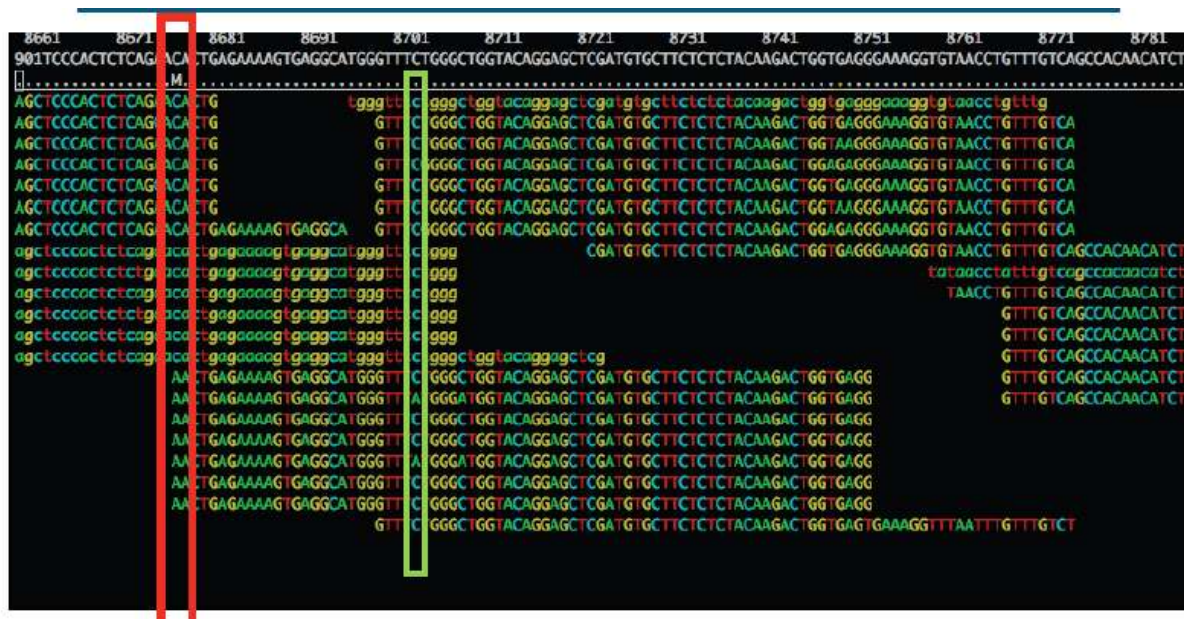
Let's look at 2 potential SNPs which are in fact artefacts.

1. Sequencing error:

The region highlighted in green on the right shows that most reads agree with the reference sequence (i.e. C-base). However, 2 reads near the bottom show an A-base. In this situation we can safely assume that the A-bases are due to a sequencing error rather than a genuine variant.

2. PCR duplication:

The highlighted region red on the left shows where there appears to be a variant (either due to sequencing of a diploid genome or non-clonal samples). A C-base is present in the reference and half the reads, whilst an A-base is present in the other reads.



Is this a genuine difference or a sequencing or sample prep error? What do you think? If this was a real sample, would you expect all the reads containing an A to start at the same location?

The answer is no. This 'SNP' is in fact due to PCR duplication. I.e. an error was made during PCR replication (converting the C to an A) early during the PCR reaction and the same fragment of DNA has been replicated many times more than the average. What we are seeing is many copies of the same erroneous PCR product. We can filter out such reads during alignment to the reference (see later).

A detailed discussion of PCR duplication can be found here

<http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/>

Indexing a reference genome:

Before we can start aligning reads to a reference genome, the genome sequence needs to be indexed. This means sorting the genome into easily searched chunks.

Task 7: Generating an index file from the reference sequence

Change directory to the reference_sequence directory:

```
cd ~/workshop_data/genomics_tutorial/data/reference/U00096/
```

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ cd ~/genomics_tutorial/data/reference/U00096/
[ec2-user@ip-10-169-87-62 U00096]$ ls -l
total 6760
-rw-r--r--. 1 ec2-user ec2-user 4708048 Dec  1 10:43 U00096.fna
-rw-r--r--. 1 ec2-user ec2-user 2208485 Dec  1 10:43 U00096.gff
```

In this directory we have 2 files. U00096.fna is a FASTA file which contains the reference genome sequence. The U00096.gff file contains the annotation for this genome. We will use this later.

First, let's look at the bwa command itself. Type:

```
bwa
```

This should yield something like:

```
Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.7.10-r789
Contact: Heng Li <lh3@sanger.ac.uk>

Usage:  bwa <command> [options]

Command: index          index sequences in the FASTA format
        mem            BWA-MEM algorithm
        fastmap        identify super-maximal exact matches
        pmerge         merge overlapping paired ends (EXPERIMENTAL)
        aln            gapped/ungapped alignment
        samse          generate alignment (single ended)
        sampe          generate alignment (paired ended)
        bwsw           BWA-SW for long queries

        fa2pac         convert FASTA to PAC format
        pac2bwt        generate BWT from PAC
        pac2bwtgen     alternative algorithm for generating BWT
        bwtupdate      update .bwt to the new format
        bwt2sa         generate SA from BWT and Occ

Note: To use BWA, you need to first index the genome with `bwa index'.
      There are three alignment algorithms in BWA: `mem', `bwsw', and
      `aln/samse/sampe'. If you are not sure which to use, try `bwa mem'
      first. Please `man ./bwa.1' for the manual.
```

BWA is actually a suite of programs which all perform different functions. We are only going to use two during this lab, bwa index, bwa mem

If we type:

```
bwa index
```

We can see more options for the bwa index command:

```
[ec2-user@ip-10-169-87-62 U00096]$ bwa index
```

```
Usage:  bwa index [-a bwtsv|is] [-c] <in.fasta>
```

```
Options: -a STR      BWT construction algorithm: bwtsv or is [auto]
          -p STR      prefix of the index [same as fasta name]
          -6           index files named as <in.fasta>.64.* instead of <in.fasta>.*
```

```
Warning: '-a bwtsv' does not work for short genomes, while '-a is' and
          '-a div' do not work not for long genomes. Please choose '-a'
          according to the length of the genome.
```

By default bwa index will use the IS algorithm to produce the index. This works well for most genomes, but for very large ones (e.g. vertebrate) you may need to use bwtsv. For bacterial genomes the default algorithm will work fine.

Now we will create a reference index for the genome using BWA:

```
bwa index U00096.fna
```

```
[ec2-user@ip-10-169-87-62 U00096]$ bwa index U00096.fna
[bwa_index] Pack FASTA... 0.04 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 1.46 seconds elapse.
[bwa_index] Update BWT... 0.04 sec
[bwa_index] Pack forward-only FASTA... 0.02 sec
[bwa_index] Construct SA from BWT and Occ... 0.47 sec
[main] Version: 0.7.10-r789
[main] CMD: bwa index U00096.fna
[main] Real time: 2.398 sec; CPU: 2.041 sec
[ec2-user@ip-10-169-87-62 U00096]$ █
```

If you now list the directory contents using the 'ls' command, you will notice that the BWA index program has created a set of new files. These are the index files BWA needs.

```
[ec2-user@ip-10-169-74-140 U00096]$ ls
U00096.fna  U00096.fna.amb  U00096.fna.ann  U00096.fna.bwt  U00096.fna.pac  U00096.fna.sa  U00096.gff
[ec2-user@ip-10-169-74-140 U00096]$
```

Task 8: Aligning reads to the indexed reference sequence:

Now we can begin to align read 1 and read 2 to the reference genome. First of all change back into the ~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/ directory and create a subdirectory to contain our remapping results.

```
cd ~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/
mkdir remapping_to_reference
cd remapping_to_reference
```

```
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ cd ~/genomics_tutorial/data/sequencing/ecoli_exeter/
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ mkdir remapping_to_reference
[ec2-user@ip-10-169-87-62 ecoli_exeter]$ cd remapping_to_reference
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ pwd
/home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/remapping_to_reference
```

Let's explore the alignment options BWA MEM has to offer. Type:

```
bwa mem
```

Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]

Algorithm options:

```
-t INT          number of threads [1]
-k INT          minimum seed length [19]
-w INT          band width for banded alignment [100]
-d INT          off-diagonal X-dropoff [100]
-r FLOAT        look for internal seeds inside a seed longer than {-k} * FLOAT [1.5]
-c INT          skip seeds with more than INT occurrences [500]
-D FLOAT        drop chains shorter than FLOAT fraction of the longest overlapping
chain [0.50]
-W INT          discard a chain if seeded bases shorter than INT [0]
-m INT          perform at most INT rounds of mate rescues for each read [50]
-S             skip mate rescue
-P             skip pairing; mate rescue performed unless -S also in use
-e            discard full-length exact matches
-A INT         score for a sequence match, which scales options -TdBOELU unless
overridden [1]
-B INT         penalty for a mismatch [4]
-O INT[,INT]   gap open penalties for deletions and insertions [6,6]
-E INT[,INT]   gap extension penalty; a gap of size k cost '{-O} + {-E}*k' [1,1]
-L INT[,INT]   penalty for 5'- and 3'-end clipping [5,5]
-U INT         penalty for an unpaired read pair [17]
```

```

-x STR      read type. Setting -x changes multiple parameters unless overridden
[null]
pacbio: -k17 -W40 -r10 -A2 -B5 -O2 -E1 -L0
pbread: -k13 -W40 -c1000 -r10 -A2 -B5 -O2 -E1 -N25 -FeaD.001

```

Input/output options:

```

-p          first query file consists of interleaved paired-end sequences
-R STR      read group header line such as '@RG\tID:foo\tSM:bar' [null]

-v INT      verbose level: 1=error, 2=warning, 3=message, 4+=debugging [3]
-T INT      minimum score to output [30]
-h INT      if there are <INT hits with score >80% of the max score, output all in
XA [5]
-a          output all alignments for SE or unpaired PE
-C          append FASTA/FASTQ comment to SAM output
-Y          use soft clipping for supplementary alignments
-M          mark shorter split hits as secondary

-I FLOAT[,FLOAT[,INT[,INT]]]
            specify the mean, standard deviation (10% of the mean if absent), max
            (4 sigma from the mean if absent) and min of the insert size
distribution.
            FR orientation only. [inferred]

```

Note: Please read the man page for detailed description of the command line and options.

The basis format of the command is

```
Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]
```

We can see that we need to provide BWA with a FASTQ files containing the raw reads (denoted by <in.fq> and <in2.fq>) to align to a reference file (unhelpfully this is listed as <idxbase>). There are also a number of options. The most important are the maximum number of differences in the seed (-k i.e. the first 32 bp of the sequence vs the reference), the number of processors the program should use (-t – our machine has 2 processors).

Our reference sequence is in

```
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna
```

Our filtered reads in

```
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R1_001.
filtered.fastq
```

```
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R2_001.
filtered.fastq
```

So to align our paired reads using processors and output to file
E_Coli_CGATGT_L001_filtered.sam:

Type, all on one line:

```
bwa mem -t 2 ~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna  
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R1_001.  
filtered.fastq  
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R2_001.  
filtered.fastq > E_Coli_CGATGT_L001_filtered.sam
```

There will be quite a lot of output but the end should look like:

```
[main] Version: 0.7.10-r789  
[main] CMD: bwa mem -t 4 /home/ec2-user/genomics_tutorial/data/reference/U00096/U00096.fna /home/ec2-user/genomics  
_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R1_001.filtered.fastq /home/ec2-user/genomics_tutorial/d  
ata/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R2_001.filtered.fastq  
[main] Real time: 102.319 sec; CPU: 193.716 sec
```

Viewing the alignment

Once the alignment is complete, list the directory contents and check that the alignment file is present.

```
ls -lh
```

```
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lh  
total 795M  
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
```

Note: ls -lh outputs the size of the file in human readable format (795Mb in this case)

The raw alignment is stored in what is called SAM format (Simple AlignMent format). It is in plain text format and you can view it if you wish using the 'less' command. Do not try to open the whole file in a text editor as you will likely run out of memory!

```
less E_Coli_CGATGT_L001_filtered.sam
```

```

@SQ      SN:gi|545778205|gb|U00096.3|      LN:4641652
@PG      ID:bwa  PN:bwa  VN:0.7.10-r789  CL:bwa mem -t 4 /home/ec2-user/genomics_tutorial/data/reference/U00096/U00
096.fna /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R1_001.filtered.fastq /ho
me/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/E_Coli_CGATGT_L001_R2_001.filtered.fastq
MISEQ:8:000000000-A7VC1:1:1101:17200:1633      83      gi|545778205|gb|U00096.3|      881006 60      137M      =
      880711 -432      GGTAAAGATGCCGGGGCGACGGGAAGCCGGGAACGGCGTGGTTCATCGGTAATGTTCCGCAACCGGGCGATCAGGTTTCGGTGGCAGACT
TGAACAAAGGTGTGATTATCCAGTCCGGTAATGACGCCCTGTATTGC      @9,@D>@8+>>@>+?A+AE?A86+++B:+8++>+B,B:,,,8,,EA,AC,8++++,,B
@8++C@@,C:,CC8++++EC,CC,C@<,,,CCC,CCC@C,C,CC,9E8CCFEEDGGGGGGGGGGGGGGGGGGCCCC9      NM:i:5 MD:Z:12T13T4C2T23T78
      AS:i:112      XS:i:0
MISEQ:8:000000000-A7VC1:1:1101:17200:1633      163      gi|545778205|gb|U00096.3|      880711 60      84M      =
      881006 432      TACTCGGGTGGCCTTTCTCCCGCACTACTCCTCTCTCCTCGTCTCTCCAGCGGGTCTGCATTTTCTCCTTTTTCCTTTTCCCC      8,A
6C,,+;+,,;CC,<,,,+8++7,,6CC<C@C<CECCC,,,9CCC,<,,,+++88BC,<,<99@B,9:BEB@@@=+:??A      NM:i:9 MD:Z:12A3A1G0A4A19G4G
19A9G4      AS:i:39 XS:i:0
MISEQ:8:000000000-A7VC1:1:1101:10456:1673      83      gi|545778205|gb|U00096.3|      1864278 60      42M      =
      1863862 -458      GGGTAAACTTGTGAAATCGATCTTGAATCACATGGCGAATT      CC;,@C<,,,9EAFFFC7GGGGFGGGGGGGGGGGGGGGGGGCCCC9
      NM:i:0 MD:Z:42 AS:i:42 XS:i:0

```

Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and a variable number of optional fields for flexible or aligner specific information. For further details as to what each field means see <http://samtools.sourceforge.net/SAM1.pdf>

Task 9: Convert SAM to BAM file

Before we can visualise the alignment however, we need to convert the SAM file to a BAM (Binary AlignMent format) which can be read by most software analysis packages. To do this we will use another suite of programs called samtools. Type:

samtools view

```
Usage:  samtools view [options] <in.bam>|<in.sam> [region1 [...]]
```

```
Options: -b          output BAM
        -h          print header for the SAM output
        -H          print header only (no alignments)
        -S          input is SAM
        -u          uncompressed BAM output (force -b)
        -l          fast compression (force -b)
        -x          output FLAG in HEX (samtools-C specific)
        -X          output FLAG in string (samtools-C specific)
        -c          print only the count of matching records
        -B          collapse the backward CIGAR operation
        -@ INT      number of BAM compression threads [0]
        -L FILE     output alignments overlapping the input BED FILE [null]
        -t FILE     list of reference names and lengths (force -S) [null]
        -T FILE     reference sequence file (force -S) [null]
        -o FILE     output file name [stdout]
        -R FILE     list of read groups to be outputted [null]
        -f INT      required flag, 0 for unset [0]
        -F INT      filtering flag, 0 for unset [0]
        -q INT      minimum mapping quality [0]
        -l STR      only output reads in library STR [null]
        -r STR      only output reads in read group STR [null]
        -s FLOAT    fraction of templates to subsample; integer part as seed [-1]
        -?          longer help

```

We can see that we need to provide samtools view with a reference genome in FASTA format file (-T), the -b and -S flags to say that the output should be in BAM format and the input in SAM, plus the alignment file.

Remember our reference sequence is in

~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna

Type (all on one line):

```
samtools view -bS -T
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna
E_Coli_CGATGT_L001_filtered.sam > E_Coli_CGATGT_L001_filtered.bam
```

This should take around 2 minutes.

```
ls -lh
```

It's always good to check that your files have produced correctly if something goes wrong it's better to catch it immediately.

```
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ samtools view -bS -T ~/genomics_tutorial/data/reference/U00096/
U00096.fna E_Coli_CGATGT_L001_filtered.sam > E_Coli_CGATGT_L001_filtered.bam
[samopen] SAM header is present: 1 sequences.
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lh
total 1.1G
-rw-rw-r--. 1 ec2-user ec2-user 254M Dec  1 14:45 E_Coli_CGATGT_L001_filtered.bam
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
```

Note that the bam file is smaller than the sam file - this is to be expected as the binary format is more efficient.

Task 10: Sort BAM file

Once this is complete we then need to sort the BAM file so that the reads are stored in the order they appear along the chromosomes (don't ask me why this isn't done automatically....). We can do this using the samtools sort command.

```
samtools sort E_Coli_CGATGT_L001_filtered.bam E_Coli_CGATGT_L001_filtered.sorted
```

This will take another minute or so.

Note that the output file is called .sorted, but actually the program appends .bam to the end of the file (see below). Just to add confusion.

```
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ samtools sort E_Coli_CGATGT_L001_filtered.bam E_Coli_CGATGT_L001_filtered.sorted
[bam_sort_core] merging from 2 files...
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lh
total 1.3G
-rw-rw-r--. 1 ec2-user ec2-user 254M Dec  1 14:45 E_Coli_CGATGT_L001_filtered.bam
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:52 E_Coli_CGATGT_L001_filtered.sorted.bam
[ec2-user@ip-10-169-87-62 remapping_to_reference]$
```

Task 11: Fill the 'MD' field using samtools

This mysterious task is an obscure term whose origins are lost in the mists of time. However, all it means is that we want samtools to look at the BAM file and annotate where it thinks SNPs should be.

On the command-line type:

```
samtools fillmd -b E_Coli_CGATGT_L001_filtered.sorted.bam
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna >
E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
```

Again, this should take around 1 minute.

As always check the results.

```
-rw-rw-r--. 1 ec2-user ec2-user 254M Dec  1 14:45 E_Coli_CGATGT_L001_filtered.bam
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:52 E_Coli_CGATGT_L001_filtered.sorted.bam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:56 E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
```

Task 12: Remove suspected PCR duplicates

Especially when using paired-end reads, samtools can do a reasonably good job of removing potential PCR duplicates (see the first part of this lab if you are unsure what this means).

Again, samtools has a great little command to do this called rmdup.

On the command-line type:

```
samtools rmdup E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
```

```
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ samtools rmdup E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam E_
Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
[bam_rmdup_core] processing reference gi|545778205|gb|U00096.3|...
[bam_rmdup_core] inconsistent BAM file for pair 'MISEQ:8:000000000-A7VC1:1:2117:24462:9451'. Continue anyway.
[bam_rmdup_core] inconsistent BAM file for pair 'MISEQ:8:000000000-A7VC1:1:2117:25993:22308'. Continue anyway.
[bam_rmdup_core] inconsistent BAM file for pair 'MISEQ:8:000000000-A7VC1:1:1113:11936:9295'. Continue anyway.
[bam_rmdup_core] 31 unmatched pairs
[bam_rmdup_core] 9680 / 458452 = 0.0211 in library ' '
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lh
total 1.6G
-rw-rw-r--. 1 ec2-user ec2-user 254M Dec  1 14:45 E_Coli_CGATGT_L001_filtered.bam
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:52 E_Coli_CGATGT_L001_filtered.sorted.bam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:56 E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
-rw-rw-r--. 1 ec2-user ec2-user 183M Dec  1 14:59 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
```

You will notice some warnings about inconsistent BAM file for pair - this is just a warning that a pair of reads does not align together on the genome within the expected tolerance - it is normal to expect some of these, and you can ignore.

Task 13: Index the BAM file

Most programs used to view BAM formatted data require an index file to locate the reads mapping to a particular location quickly. We'll use the samtools index command to do this.

Type:

```
samtools index E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
```

```
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ samtools index E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.
bam
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lh
total 1.6G
-rw-rw-r--. 1 ec2-user ec2-user 254M Dec  1 14:45 E_Coli_CGATGT_L001_filtered.bam
-rw-rw-r--. 1 ec2-user ec2-user 795M Dec  1 14:21 E_Coli_CGATGT_L001_filtered.sam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:52 E_Coli_CGATGT_L001_filtered.sorted.bam
-rw-rw-r--. 1 ec2-user ec2-user 185M Dec  1 14:56 E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
-rw-rw-r--. 1 ec2-user ec2-user 183M Dec  1 14:59 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
-rw-rw-r--. 1 ec2-user ec2-user  15K Dec  1 15:00 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam.bai
```

We should obtain a .bai file (known as a BAM-index file).

Task 14: Obtain mapping statistics

Finally we can obtain some summary statistics.

```
samtools flagstat E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam >
mappingstats.txt
```

This should only take a few seconds. Once complete view the mappingstats.txt file using a text-editor (e.g. pluma or nano) or the 'more' command.

```

quencing/ecoli_exeter/remapping_to_reference]$ cat mappingstats.txt
1250552 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
898917 + 0 mapped (71.88%:-nan%)
1250552 + 0 paired in sequencing
625340 + 0 read1
625212 + 0 read2
894492 + 0 properly paired (71.53%:-nan%)
897495 + 0 with itself and mate mapped
1422 + 0 singletons (0.11%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

```

So here we can see we have 1250552 reads in total, none of which failed QC. 71.88% of reads mapped to the reference genome and 71.53% mapped with the expected 500-600bp distance between them. 1422 reads could not have their read-pair mapped (i.e read 1 mapped but read 2 did not or vice-versa).

0 reads have mapped to a different chromosome than their pair (0 has a mapping quality > 5 – this is a Phred scaled quality score much as we say in the FASTQ files). If there were any such reads they would likely be due to repetitive sequences (e.g phage insertion sites) or an insertion of plasmid or phage DNA into the main chromosome.

Task 15: Cleanup

We have a number of leftover intermediate files which we can now remove to save space.

Type (all on one line):

```

rm E_Coli_CGATGT_L001_filtered.sam E_Coli_CGATGT_L001_filtered.bam
E_Coli_CGATGT_L001_filtered.sorted.bam
E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam

```

You should now be left with the processed alignment file, the index file and the mapping stats.

```

[ec2-user@ip-10-169-87-62 remapping_to_reference]$ rm E_Coli_CGATGT_L001_filtered.sam E_Coli_CGATGT_L001_filtered.
bam E_Coli_CGATGT_L001_filtered.sorted.bam E_Coli_CGATGT_L001_filtered.sorted.fillmd.bam
[ec2-user@ip-10-169-87-62 remapping_to_reference]$ ls -lg
total 186704
-rw-rw-r--. 1 ec2-user 191163911 Dec  1 14:59 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
-rw-rw-r--. 1 ec2-user    14624 Dec  1 15:01 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam.bai
-rw-rw-r--. 1 ec2-user      383 Dec  1 15:02 mappingstats.txt

```

Well done! You have now mapped, filtered and sorted your first whole genome data-set!
Let's take a look at it!

Task 16: QualiMap

Qualimap (<http://qualimap.bioinfo.cipf.es/>) is a program that summarises the alignment in much more detail than the mapping stats file we produced. It's a technical tool which allows you to assess the sequencing for any problems and biases in the sequencing and the alignment rather than a tool to deduce biological features.

There are a few options to the program, We want to run bamqc. Type:

```
qualimap bamqc
```

to get some help on this command.

To get the report, first make sure you are in the directory:

```
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/remapping_to_reference
```

then run the command:

```
qualimap bamqc -outdir bamqc -bam  
E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam -gff  
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.gff
```

this creates a subfolder called bamqc

```
[ec2-user@ip-10-165-124-92 remapping_to_reference]$ ls -l  
total 186692  
drwxrwxr-x. 5 ec2-user ec2-user      4096 Dec  2 10:00 bamqc  
-rw-rw-r--. 1 ec2-user ec2-user 191144523 Dec  1 17:15 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam  
-rw-rw-r--. 1 ec2-user ec2-user   14608 Dec  1 17:16 E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam.bai  
-rw-rw-r--. 1 ec2-user ec2-user     383 Dec  1 17:16 mappingstats.txt
```

cd to this directory and run

```
firefox qualimapReport.html
```

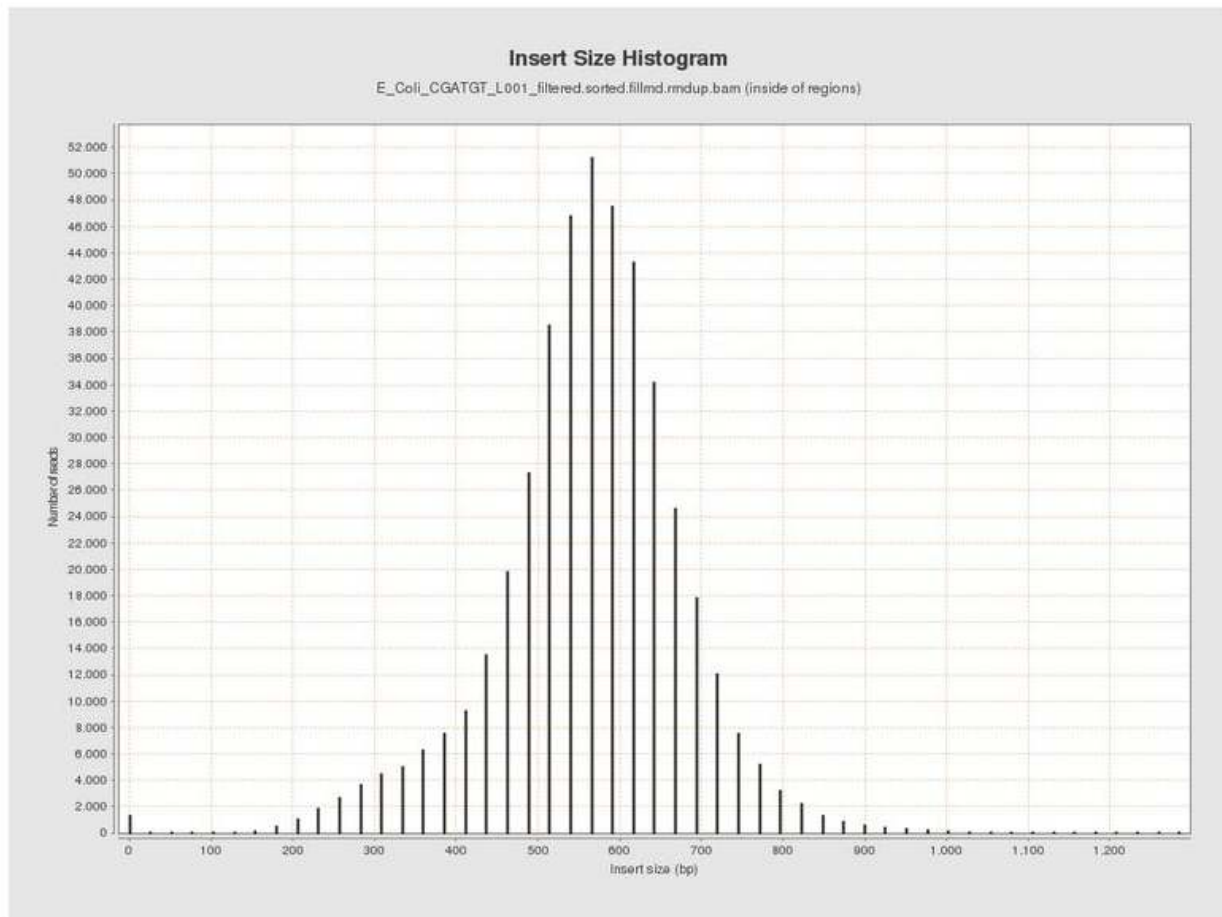
There is a lot in the report so just a few highlights:

Coverage across reference



This shows the number of reads that 'cover' each section of the genome. The red line shows a rolling average around 50x - this means that on average every part of the genome was sequenced 50X. It is important to have sufficient depth of coverage in order to be confident that any features you find in your data are real and not a result of sequencing errors.

Insert Size Histogram



The Insert Size Histogram displays the range of sizes of the DNA fragments. It shows how well your DNA was size selected before sequencing. Note that the 'insert' refers to the DNA that was inserted between the sequencing adaptors, so equates to the size range of the DNA that was used. In this case we have 300 paired end reads and our insert size varies around 600 bases - so there should only be a small gap between the reads that was not sequenced.

Have a look at some of the other graphs produced.

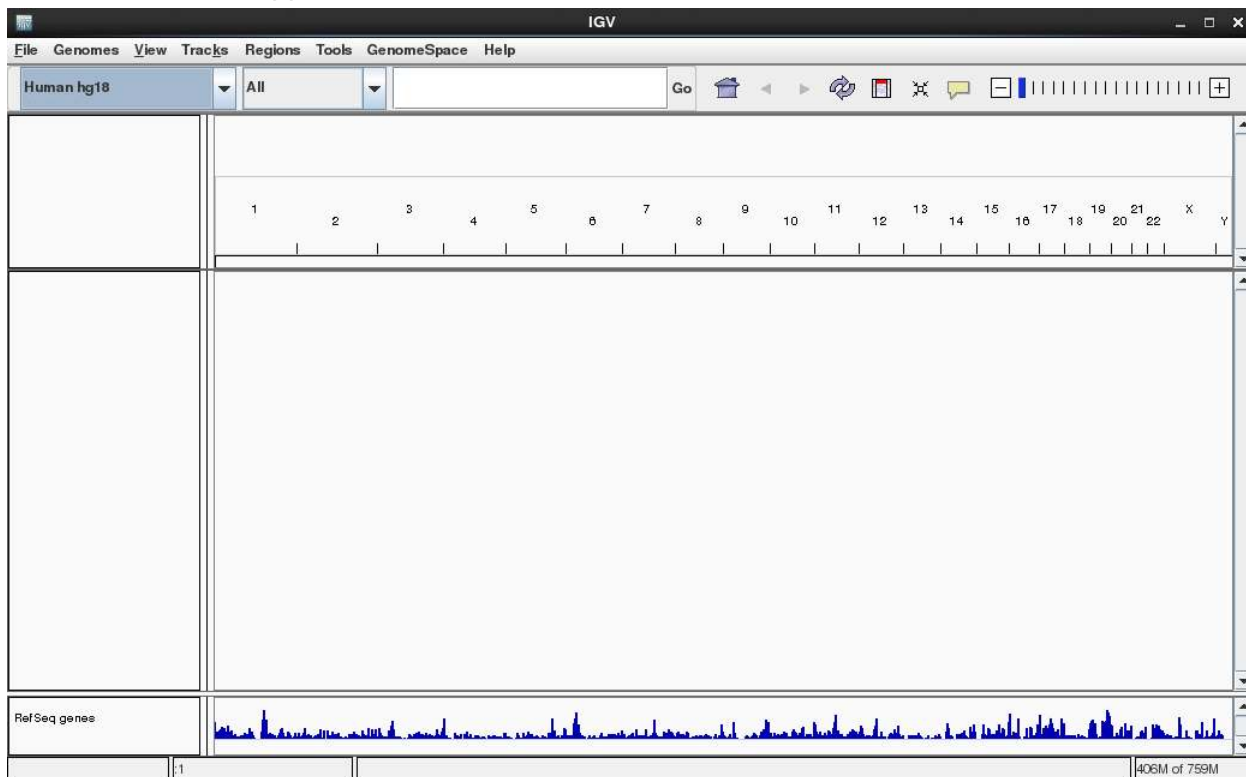
Task 17: Load the Integrative Genomics Viewer

The Integrative Genome Viewer (IGV) is a tool developed by the Broad Institute for browsing interactively the alignment data you produced. It has a wealth of features and we can only cover some basics to get you started. Go to <http://www.broadinstitute.org/igv/> to get more information.

In your terminal, type

```
igv.sh
```

IGV viewer should appear:

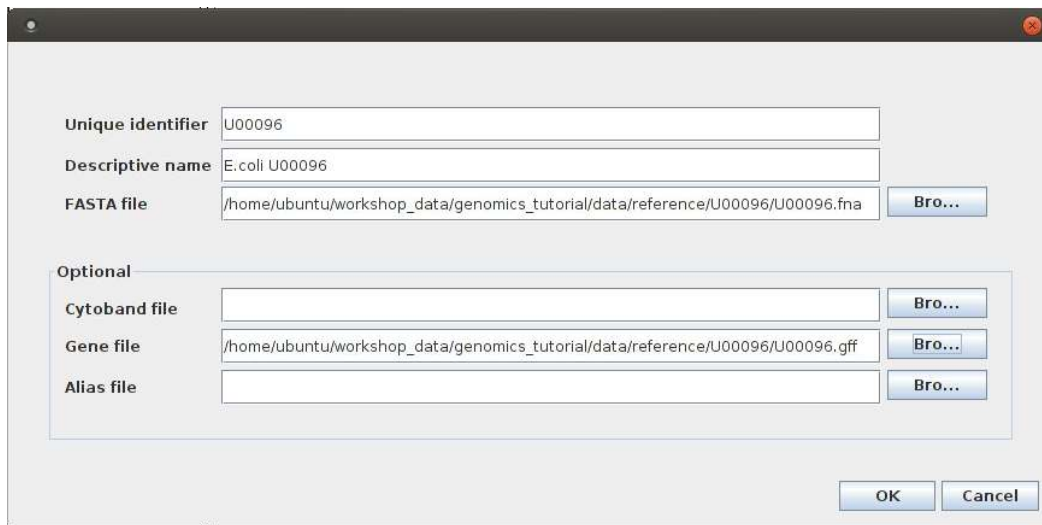


Notice that by default a human genome has been loaded.

Task 17a: Import the *E.coli* U0009 reference genome to IGV

By default IGV does not contain our reference genome. We'll need to import it.

Click on 'Genomes ->Create .genome file...'

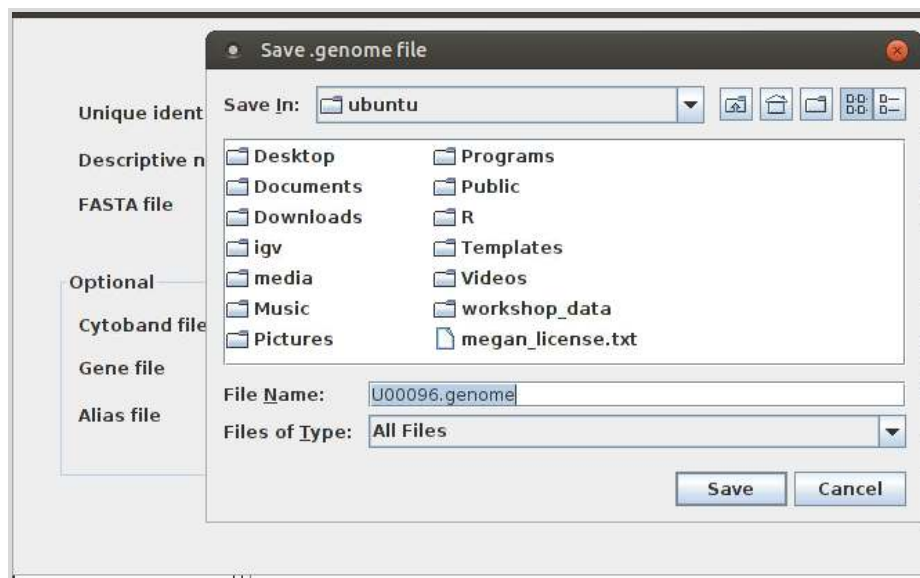


The screenshot shows a dialog box titled 'Create .genome file'. It contains the following fields and buttons:

- Unique identifier:** U00096
- Descriptive name:** E.coli U00096
- FASTA file:** /home/ubuntu/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna (with a 'Bro...' button)
- Optional section:**
 - Cytoband file:** (empty) (with a 'Bro...' button)
 - Gene file:** /home/ubuntu/workshop_data/genomics_tutorial/data/reference/U00096/U00096.gff (with a 'Bro...' button)
 - Alias file:** (empty) (with a 'Bro...' button)
- Buttons:** OK, Cancel

Enter the information above and click on 'OK'.

IGV will ask where it can save the genome file. Your home directory will be fine.

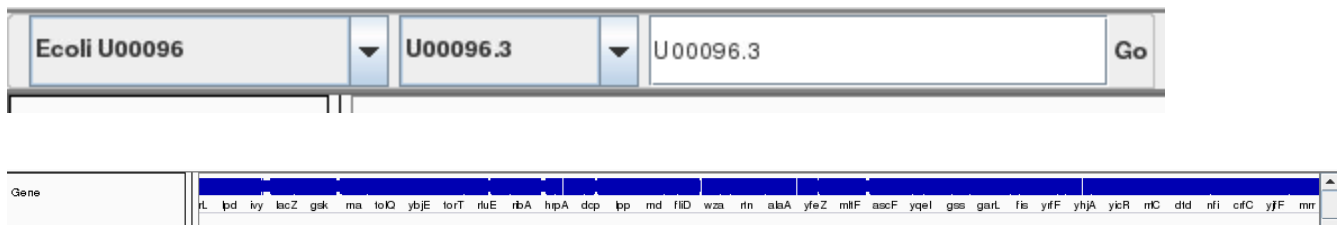


The screenshot shows a 'Save .genome file' dialog box. It contains the following fields and buttons:

- Save In:** ubuntu (with a dropdown arrow)
- File list:** Desktop, Documents, Downloads, igv, media, Music, Pictures, Programs, Public, R, Templates, Videos, workshop_data, megan_license.txt
- File Name:** U00096.genome
- Files of Type:** All Files
- Buttons:** Save, Cancel

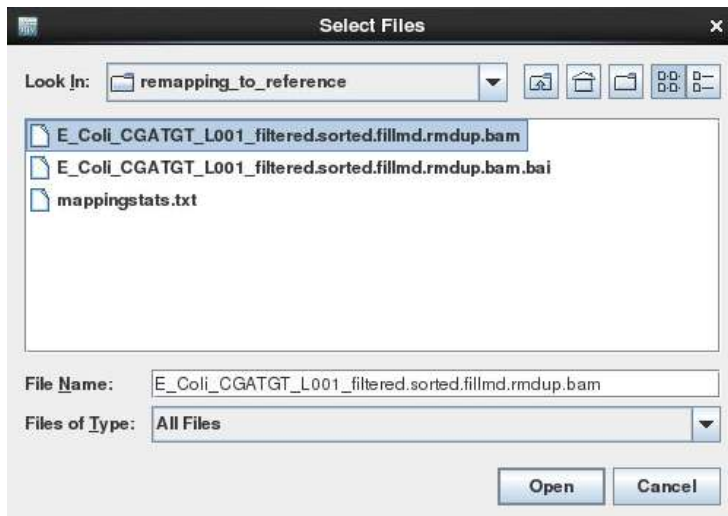
Click 'Save' again.

Note that the genome and the annotation have now been imported.



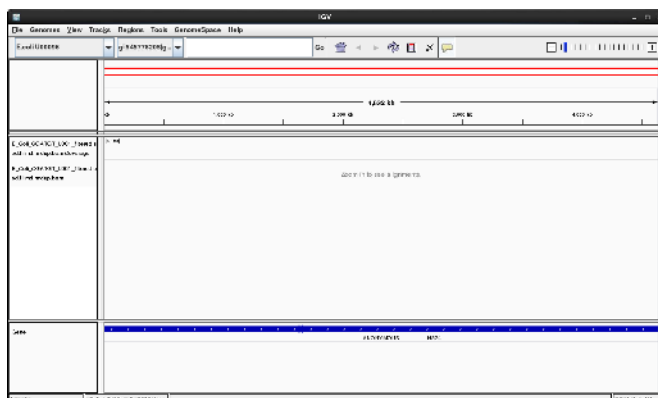
Task 17b: Load the BAM file

Load the alignment file. Note that IGV requires the .bai index file to also be in the same directory.
Select File... and Load From File



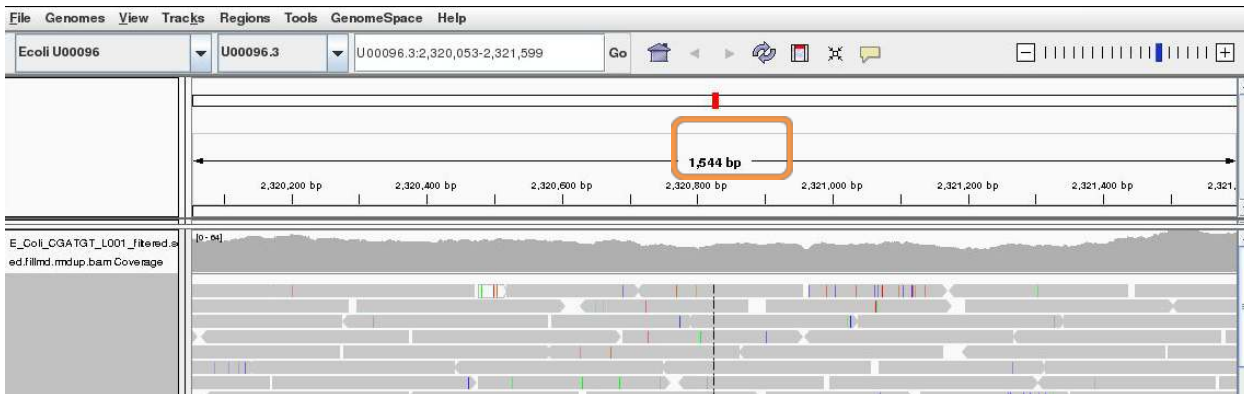
Select the bam file and click open

Once loaded your screen should look similar to the following. Note that you can load more BAM files if you wish to compare different samples or the results of different mapping programs.

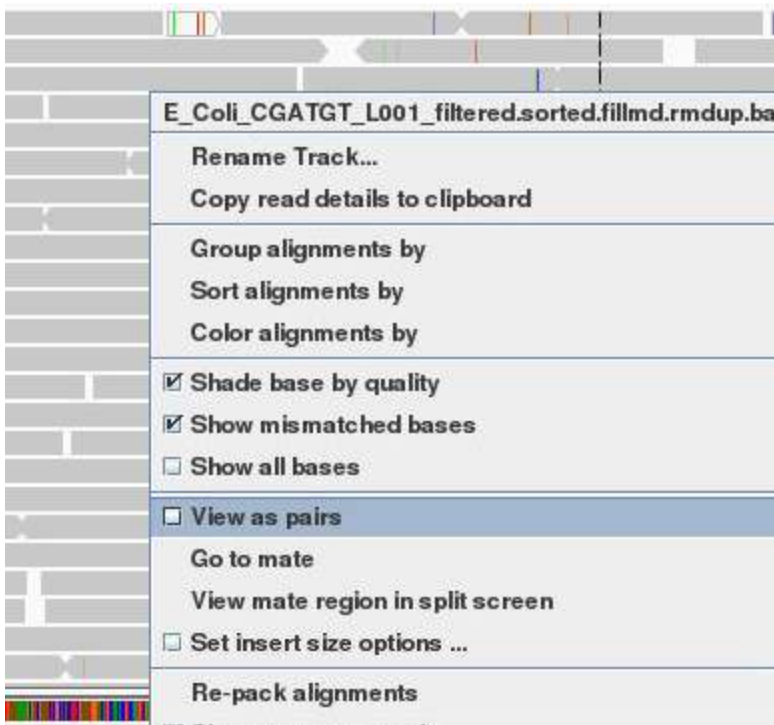




Select the chromosome U00096.3 if it is not already selected



Use the +/- keys to zoom in or use the zoom bar at the top right of the screen to zoom into about 1-2kbases as above



Right click on the main area and select view as pairs

The gray graph at the top of the figure indicates the coverage of the genome:



The more reads mapping to a certain location, the higher the peak on the graph. You'll see a coloured line of blue, green or red in this coverage plot if there are any SNPs (single-nucleotide polymorphisms) present (there are none in the plot). If there are any regions in the genome which are not covered by the reads, you will see these as gaps in the coverage graph. Sometimes these gaps are caused by repetitive regions; others are caused by genuine insertions/deletions in your new strain with respect to the reference.

Below the coverage graph is a representation of each read pair as it is mapped to the genome. One pair is highlighted.



This pair consists of 2 reads with a gap (there may be no gap if the reads overlap) Any areas of mismatch either due to inconsistent distances between paired-end reads or due to differences between the reference and the read and are highlighted by a colour. The brighter the colour, the higher the base-calling quality is estimated to be. Differences in a single read are likely to be sequencing errors. Differences consistent in all reads are likely to be mutations.

Hover over a read to get detailed information about the reads' alignment:

Left alignment	Right alignment
Read name = MISEQ:8:000000000-A7VC1:1:2112:3986:8017	Read name = MISEQ:8:000000000-A7VC1:1:2112:3986:8017
Location = U00096.3:2,319,925	Location = U00096.3:2,319,925
Alignment start = 2,319,293 (+)	Alignment start = 2,319,859 (-)
Cigar = 270M	Cigar = 187M
Mapped = yes	Mapped = yes
Mapping quality = 60	Mapping quality = 60
Secondary = no	Secondary = no
Supplementary = no	Supplementary = no
Duplicate = no	Duplicate = no
Failed QC = no	Failed QC = no
Mate is mapped = yes	Base = C
Mate start = U00096.3:2319858 (-)	Base phred quality = 25
Insert size = 753	Mate is mapped = yes
First in pair	Mate start = U00096.3:2319292 (+)
Pair orientation = F1R2	Insert size = -753
MD = 221A10A37	Second in pair
NM = 2	Pair orientation = F1R2
AS = 260	MD = 12T27T146
XS = 0	NM = 2
	AS = 177

You don't need to understand every value, but compare this to the SAM format to get an idea of what is there.

SNPs and Indels

The following 3 tasks are open-ended. Please take your time with these. Read the examples on the following page if you get stuck.

Task 18: Read about the alignment display format

Visit <http://www.broadinstitute.org/software/igv/AlignmentData>

Task 19a: Manually identify a region without any reads mapping.

It can be quite difficult to find even with a very small genome. Zoom out as far as you can and still see the reads. Use the coverage plot from QualiMap to try to find it. Are there genes associated?

Task 19b: Manually identify a region containing repetitive sequences.

Again try to use the QualiMap report to give you an idea. What is this region? Is there a gene close-by? What do you think this is? (Think about repetitive sequences, what does BWA do if a region in the genome has been duplicated)

Task 20: Identify SNPs and Indels manually

Can you find any SNPs? Which genes (if any) are they in? How reliable do they look? (Hint – look at the number of reads mapping, their orientation - which strand they are on and how bright the base-calls are).

Optional:

Zoom in to maximum magnification at the site of the SNP. Can you determine whether a SNP results in a synonymous (i.e. silent) or non-synonymous change in the amino acid? Can you use PDB (<http://www.rcsb.org/pdb/home/home.do>) or other resources to determine whether or not this occurs in a catalytic site or other functionally crucial region? (Note this may not always be possible).

What effect do you think this would have on the cell?

Example: Identifying Variants manually

Here are some regions where there are differences in the organism sequenced and the reference:
Can you interpret what has happened to the genome of our strain? Try to work out what is going on yourself before looking at the comment

Paste each of the genomic locations in this box and click go

Ecoli U00096	▼	U00096.3	▼	U00096.3:2,108,392-2,133,153	Go
--------------	---	----------	---	------------------------------	----

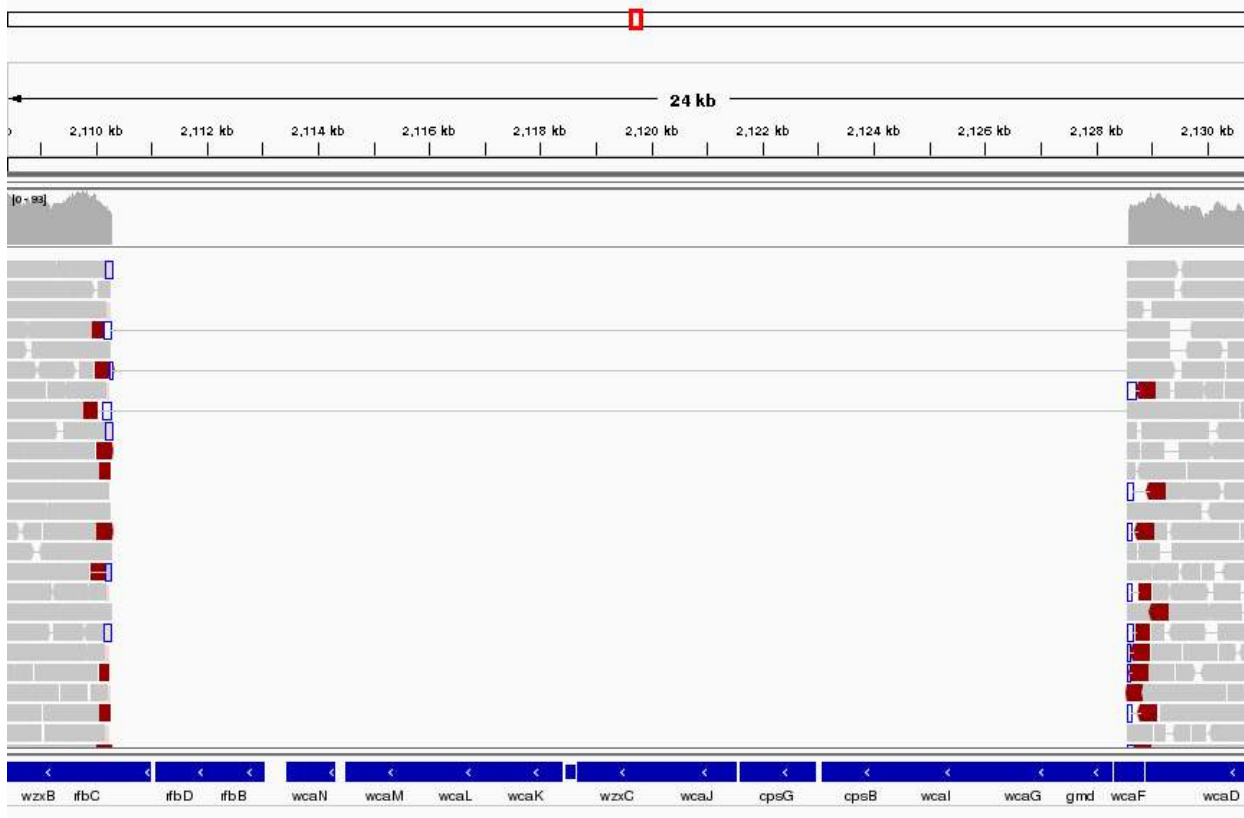
U00096.3:2,108,392-2,133,153

U00096.3:3,662,049-3,663,291

U00096.3:4,296,249-4,296,510

U00096.3:565,965-566,489

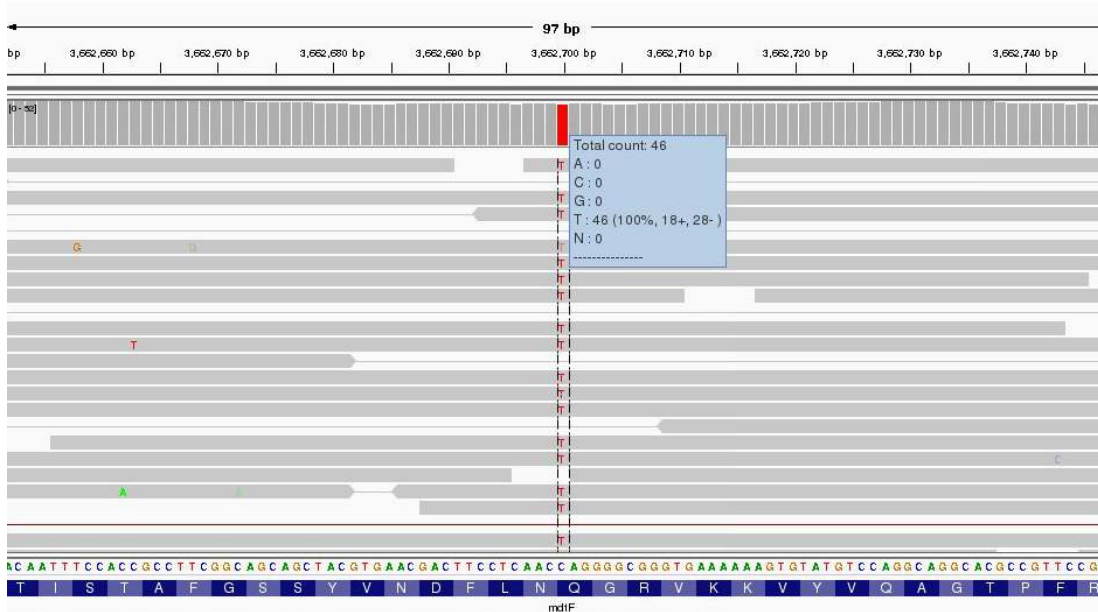
Region U00096.3:2,108,392-2,133,153



This area corresponds to the drop in coverage identified by Qualimap. It looks like a fairly large region of about 17 kbases which was present in the reference and is missing from our sequenced genome. It looks like about 12 genes from the reference strain are not present in our strain - is this real or an artefact?

The evidence we have shows that we have coverage of about 60X either side of the deletion and no reads at all mapping within the region. There are nice clean edges to the start and end of the suspected deletion. We also have paired reads which should be 300-400bp apart which span the suspected deletion. This is exactly what you would expect if the two regions either side of the suspected deletion were actually joined together.

Region U00096.3:3,662,049-3,663,291

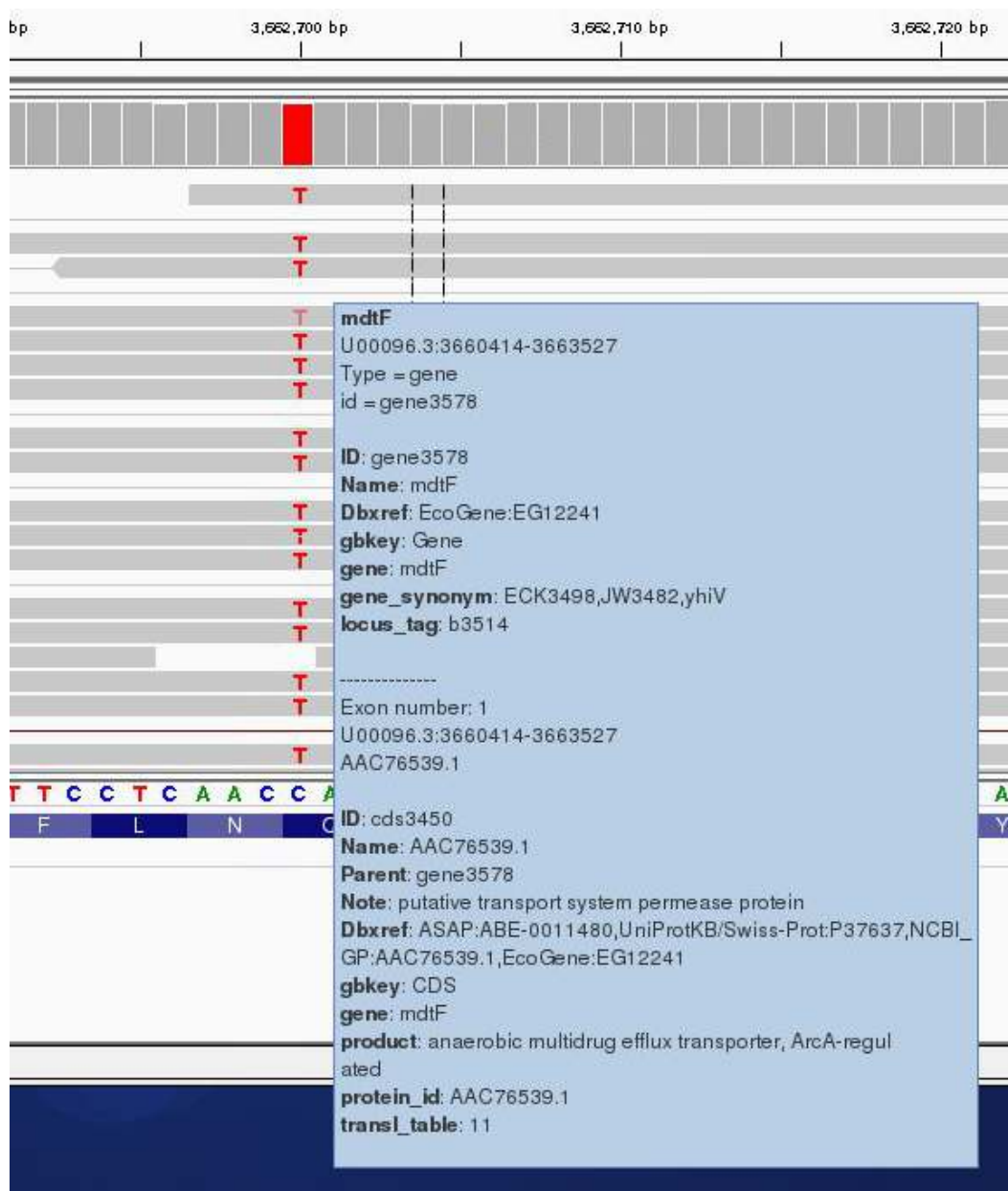


Zoom right in until you can see the reference sequence and protein sequence at the bottom of the display.

The first thing to note is that only discrepancies with respect to the reference are shown. If a read is entirely the same as its reference, it will appear entirely grey. Blue and red blocks indicate the presence of an 'abnormal' distance between paired-end reads. Note that unless this is consistent across most of the reads at a given position, it is not significant.

Here we have a C->T SNP. This changes the codon from CAG->TAG (remember to check what strand the gene is on this one is on the forward strand, if it was on the reverse strand you would have to take the reverse complement of the codon to interpret the amino acid it codes for.) and results in a Gln->Stop mutation in the final protein product which is very likely to change the effect of the protein product.

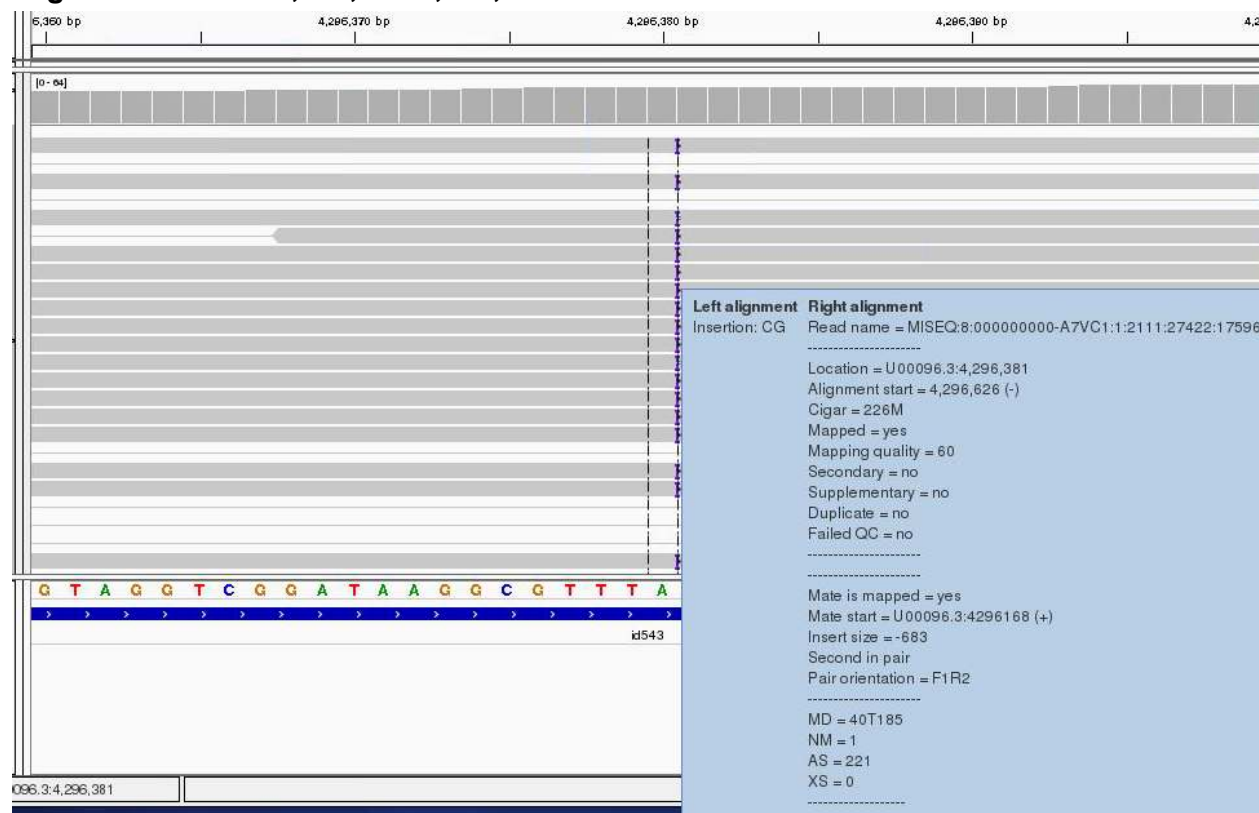
Hover over the gene to get some more information from the annotation... Since it is a drug resistance protein it could be very significant.



One additional check is that the SNPs also occur when reading the forward strand. We can check this by looking at the direction of the grey reads, or by hovering over the coverage graph - see previous diagram. We can see that approximately half of the bases reporting the C->T mutation occur in read 1 (forward arrow), and half in read 2 (reverse arrow). This adds confidence to the base-call as it reduces the likelihood of this SNP being the result of a PCR duplication error.

Note that sequencing errors in Illumina data are quite common (look at the odd bases showing up in the screen above). We rely on depth of sequencing to average out these errors. This is why people often mention that a minimum median coverage of 20-30x across the genome is required for accurate SNP-calling with Illumina data on a diploid organism. This is not necessarily true for simple organisms such as prokaryotes, but for diploid and polyploid organisms it becomes important because each position may have one, two or many alleles changed.

Regions U00096.3:4,296,249-4,296,510



Much the same guidelines apply for indels as they do for SNPs. Here we have an insertion of two bases CG in our sample compared to the reference. Again, we can see how much confidence we have that the insertion is real by checking that the indel is found on both read 1 and read 2 and on both strands.

The insertion is signified by the presence of a purple bar. Hover your mouse over it to get more details as above. We can hover our mouse over the reference sequence to get details of the gene. We can see that it occurs in a repeat region and is unlikely to have very significant effects.

One can research the effect that a SNP or Indel may have by finding the relevant gene at <http://www.uniprot.org> (or google 'mdtF uniprot' in the previous case).

It should be clear from this quick exercise that trying to work out where SNPs and Indels are manually is a fairly tedious process if there are many mutations. As such, the next section will look at how to obtain spread-sheet friendly summary details of these.

Region U00096.3:565,965-566,489

This last region is more complex try to understand what genomic mutation could account for this pattern - discuss with a colleague or an instructor.

Recap: SNP/Indel identification

1. Only changes from the reference sequence are displayed in IGV
2. Genuine SNPs/Indels should be present on both read 1 and read 2
3. Genuine SNPs/Indels should be present on both strands
4. Genuine SNPs/Indels should be supported by a good (i.e. 20-30x) depth of coverage
4. Very important mutations (i.e. ones relied upon in a paper) should be confirmed via PCR/Sanger sequencing.

Automated analyses

Viewing alignments is useful when convincing yourself or others that a particular mutation is real rather than an artefact and for getting a feel for short read sequencing datasets. However, if we want to quickly and easily find variants we need to be able to generate lists of variants, in which gene they occur (if any) and what effect they have. We also need to know which (if any) genes are missing (i.e. have zero coverage).

Automated variant calling

To call variants we can use a number of packages (e.g. VarScan, GATK). However here, we will show you how to use the bcftools package which comes with samtools. First we need to generate a pileup file which contains only locations with the variants and pass this to bcftools.

Task 21: Identify SNPs and Indels using automated variant callers

Make sure you are in the directory.

~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/remapping_to_reference

Type the following:

```
samtools mpileup
```

You should see a screen similar to the following

```
Usage: samtools mpileup [options] in1.bam [in2.bam [...]]
```

Input options:

```
-6          assume the quality is in the Illumina-1.3+ encoding
-A          count anomalous read pairs
```

```

-B          disable BAQ computation
-b FILE     list of input BAM filenames, one per line [null]
-C INT      parameter for adjusting mapQ; 0 to disable [0]
-d INT      max per-BAM depth to avoid excessive memory usage [250]
-E          recalculate extended BAQ on the fly thus ignoring existing BQs
-f FILE     faidx indexed reference sequence file [null]
-G FILE     exclude read groups listed in FILE [null]
-l FILE     list of positions (chr pos) or regions (BED) [null]
-M INT      cap mapping quality at INT [60]
-r STR      region in which pileup is generated [null]
-R          ignore RG tags
-q INT      skip alignments with mapQ smaller than INT [0]
-Q INT      skip bases with baseQ/BAQ smaller than INT [13]
--rf INT     required flags: skip reads with mask bits unset []
--ff INT     filter flags: skip reads with mask bits set []

```

Output options:

```

-D          output per-sample DP in BCF (require -g/-u)
-g          generate BCF output (genotype likelihoods)
-O          output base positions on reads (disabled by -g/-u)
-s          output mapping quality (disabled by -g/-u)
-S          output per-sample strand bias P-value in BCF (require -g/-u)
-u          generate uncompressed BCF output

```

SNP/INDEL genotype likelihoods options (effective with '-g' or '-u'):

```

-e INT      Phred-scaled gap extension seq error probability [20]
-F FLOAT    minimum fraction of gapped reads for candidates [0.002]
-h INT      coefficient for homopolymer errors [100]
-I          do not perform indel calling
-L INT      max per-sample depth for INDEL calling [250]
-m INT      minimum gapped reads for indel candidates [1]
-o INT      Phred-scaled gap open sequencing error probability [40]
-p          apply -m and -F per-sample to increase sensitivity
-P STR      comma separated list of platforms for indels [all]

```

Notes: Assuming diploid individuals.

If you are running this on your own datasets, please make sure you set the -d parameter if you have high coverage (i.e. > 100x coverage) per sample.

As the samtools mpileup command outputs an unfriendly output, we will pass it directly to the bcftools view command using the linux pipe ('|'). Type the following:

```

samtools mpileup -uf
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna
E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam | bcftools view -bvcg -> var.raw.bcf

```

This may take 15 minutes or so and will generate a BCF (Binary Call Format) file containing the raw unfiltered variant calls in a binary format.

This is not readable by humans, so let's use the bcftools view command and use the linux pipe (|) with the vcftutils.pl varFilter command. We can see what the options are for this program by typing in:

vcftutils.pl varFilter

Usage: vcftutils.pl varFilter [options] <in.vcf>

Options: -Q INT minimum RMS mapping quality for SNPs [10]
 -d INT minimum read depth [2]
 -D INT maximum read depth [10000000]
 -a INT minimum number of alternate bases [2]
 -w INT SNP within INT bp around a gap to be filtered [3]
 -W INT window size for filtering adjacent gaps [10]
 -1 FLOAT min P-value for strand bias (given PV4) [0.0001]
 -2 FLOAT min P-value for baseQ bias [1e-100]
 -3 FLOAT min P-value for mapQ bias [0]
 -4 FLOAT min P-value for end distance bias [0.0001]
 -e FLOAT min P-value for HWE (plus F<0) [0.0001]
 -p print filtered variants

Note: Some of the filters rely on annotations generated by SAMtools/BCFtools.

We will use the -d option to limit variant calls to those positions where there are at least 10 reads. Type:

```
bcftools view var.raw.bcf | vcftutils.pl varFilter -d 10 > var.flt.vcf
```

Once complete, view the file using the 'more' command. You should see something similar to: (lines beginning with # are just comment lines explaining the output)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
000096.3	378700	.	A	C	222	.	DP=47;VDB=3.492280e-01;AF1=1;AC1=2;DP4=0,0,20,26;MQ=60;FQ=-165
000096.3	566173	.	C	G	140	.	DP=74;VDB=1.335471e-01;RBP=-1.366788e+00;AF1=0.5;AC1=1;DP4=22,35,7,9;MQ=60;FQ=143;PV4=0.78,0.051,1,1
000096.3	566205	.	T	C	152	.	DP=70;VDB=3.660676e-02;RBP=-2.810193e-01;AF1=0.5;AC1=1;DP4=22,31,6,9;MQ=60;FQ=155;PV4=1,1,1,1
000096.3	566245	.	G	A	133	.	DP=67;VDB=1.726489e-02;RBP=7.739471e-01;AF1=0.5;AC1=1;DP4=22,29,5,9;MQ=60;FQ=136;PV4=0.76,1,1,0.35
000096.3	566277	.	C	T	55	.	DP=63;VDB=3.921215e-03;RBP=2.597793e-01;AF1=0.5;AC1=1;DP4=25,28,3,6;MQ=60;FQ=58;PV4=0.49,1,1,1
000096.3	566323	.	C	T	71	.	DP=58;VDB=6.304791e-03;RBP=2.418227e+00;AF1=0.5;AC1=1;DP4=25,23,3,6;MQ=60;FQ=74;PV4=0.47,1,1,1
000096.3	566326	.	T	C	57	.	DP=57;VDB=5.476300e-03;RBP=2.654789e+00;AF1=0.5;AC1=1;DP4=24,23,3,6;MQ=60;FQ=60;PV4=0.47,1,1,1
000096.3	566332	.	T	G	26	.	DP=57;VDB=3.998488e-03;RBP=2.444295e+00;AF1=0.5;AC1=1;DP4=25,22,3,7;MQ=60;FQ=29;PV4=0.3,0.32,1,1
000096.3	566356	.	T	C	71	.	DP=60;VDB=3.343644e-02;RBP=3.626135e+00;AF1=0.5;AC1=1;DP4=25,21,3,7;MQ=60;FQ=74;PV4=0.3,0.49,1,0.13

You can see the chromosome, position, reference and alternate allele as well as a quality score for the SNP. This is a VCF file (Variant Call File). This is a standard developed for the 1000 genomes project. The full specification is given at <http://samtools.github.io/hts-specs/VCFv4.2.pdf>

The lines starting DP and INDEL contain various details concerning the variants. For haploid organisms, most of these details are not necessary.

Variant qualities:

Typically one should only accept variant calls over a certain quality threshold. Typically a threshold of 60 is used (i.e. a 1 in 1000000 chance of a mis-called variant). Here you can see that all these

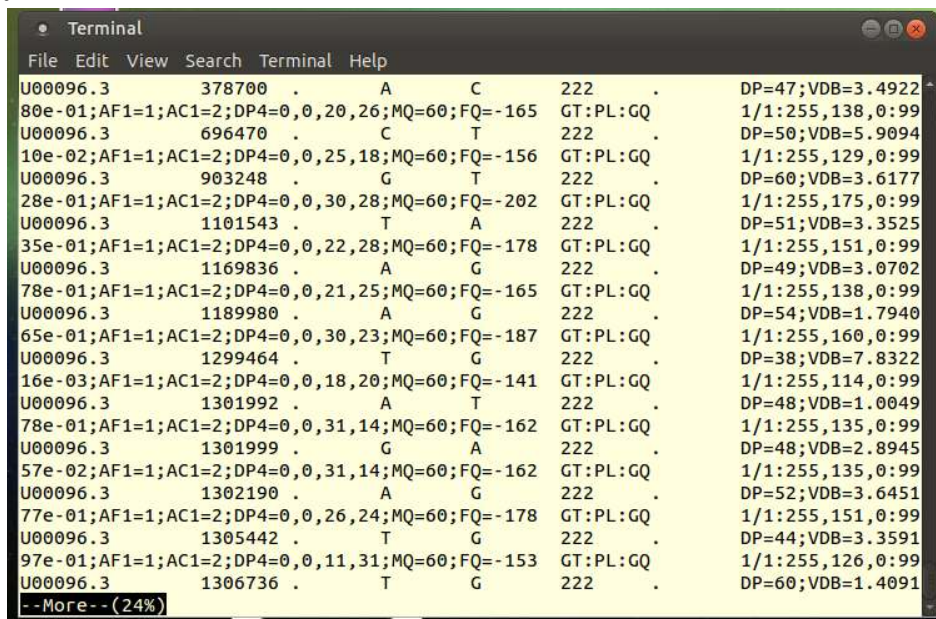
variants would pass these thresholds. However, for future reference, we can use the Linux 'awk' command to filter the data on the quality column (i.e. column 6):

```
awk '($6>=60)' var.flt.vcf > out.snps.vcf4.temp
```

Unfortunately samtools assumes the organism is diploid so we also want to remove any heterozygous calls:

```
awk '($10~/1V1/)' out.snps.vcf4.temp > out.snps.vcf4
```

Again viewing the final output file out.snps.vcf4 using a text-editor or the 'more' command should yield:



```
U00096.3 378700 . A C 222 . DP=47;VDB=3.4922
80e-01;AF1=1;AC1=2;DP4=0,0,20,26;MQ=60;FQ=-165 GT:PL:GQ 1/1:255,138,0:99
U00096.3 696470 . C T 222 . DP=50;VDB=5.9094
10e-02;AF1=1;AC1=2;DP4=0,0,25,18;MQ=60;FQ=-156 GT:PL:GQ 1/1:255,129,0:99
U00096.3 903248 . G T 222 . DP=60;VDB=3.6177
28e-01;AF1=1;AC1=2;DP4=0,0,30,28;MQ=60;FQ=-202 GT:PL:GQ 1/1:255,175,0:99
U00096.3 1101543 . T A 222 . DP=51;VDB=3.3525
35e-01;AF1=1;AC1=2;DP4=0,0,22,28;MQ=60;FQ=-178 GT:PL:GQ 1/1:255,151,0:99
U00096.3 1169836 . A G 222 . DP=49;VDB=3.0702
78e-01;AF1=1;AC1=2;DP4=0,0,21,25;MQ=60;FQ=-165 GT:PL:GQ 1/1:255,138,0:99
U00096.3 1189980 . A G 222 . DP=54;VDB=1.7940
65e-01;AF1=1;AC1=2;DP4=0,0,30,23;MQ=60;FQ=-187 GT:PL:GQ 1/1:255,160,0:99
U00096.3 1299464 . T G 222 . DP=38;VDB=7.8322
16e-03;AF1=1;AC1=2;DP4=0,0,18,20;MQ=60;FQ=-141 GT:PL:GQ 1/1:255,114,0:99
U00096.3 1301992 . A T 222 . DP=48;VDB=1.0049
78e-01;AF1=1;AC1=2;DP4=0,0,31,14;MQ=60;FQ=-162 GT:PL:GQ 1/1:255,135,0:99
U00096.3 1301999 . G A 222 . DP=48;VDB=2.8945
57e-02;AF1=1;AC1=2;DP4=0,0,31,14;MQ=60;FQ=-162 GT:PL:GQ 1/1:255,135,0:99
U00096.3 1302190 . A G 222 . DP=52;VDB=3.6451
77e-01;AF1=1;AC1=2;DP4=0,0,26,24;MQ=60;FQ=-178 GT:PL:GQ 1/1:255,151,0:99
U00096.3 1305442 . T G 222 . DP=44;VDB=3.3591
97e-01;AF1=1;AC1=2;DP4=0,0,11,31;MQ=60;FQ=-153 GT:PL:GQ 1/1:255,126,0:99
U00096.3 1306736 . T G 222 . DP=60;VDB=1.4091
---More--- (24%)
```

This forms our definitive list of variants for this sample.

Take a look at some of the look at some of the variants we just excluded, was it justified. Remember there is no filter that can keep all the correct variants and remove all the dubious!

Task 22: Compare the variants found using this method to those you found in the manual section

Can you see any variants which may have been missed? Often variants within a few bp of indels are filtered out as they could be spurious SNPs thrown up by a poor alignment. This is especially the case if you use non-gapped aligners such as Bowtie.

Quickly locating genes which are missing compared to the reference

We can use a command from the BEDTools package (<http://bedtools.readthedocs.org/en/latest/>) to identify annotated genes which are not covered by reads across their full length.

Type the following on one line:

```
coverageBed -abam E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam -b  
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.gff > gene_coverage.txt
```

This should only take a minute or so. The output contains one row per annotated gene - the 13th column contains the proportion of the gene that is covered but reads from our sequencing. 1.00 means the gene is 100% covered and 0.00 means no coverage.

If we sort by this column we can see which genes are missing

```
sort -t '$\t' -g -k 13 gene_coverage.txt | more
```

There is another region of about 10kb which is absent from our sequences - can you find it in IGV?

Evaluating the impact of variants

So far we have found a number of genes missing from this strain of *E.coli* which obviously could have a phenotypic effect. Let's now take a closer look at the variants. We'd like to obtain a list of genes in which these variants occur and whether they result in amino acid changes.

To do this we'll use a custom perl script developed by David Studholme and Konrad Paszkiewicz.

We'll just need the reference annotation, sequence and the VCF file containing the SNPs.

Task 23: Determine the effect of variants

Type (all on one line):

```
snp_comparator.pl 10  
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna  
~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.gff out.snps.vcf4 >  
snp_report.txt
```

You will see lots of warnings about 'Use of uninitialized value \$gene_name - you can ignore these.

This program takes the information from the reference sequence and annotation, and the VCF SNP files and determines whether the variant occurs within a gene, and if so the effect of each mutation.

Once complete, view the snp_report.txt file using the more command:

```
## Table of SNP and Indel occurrences between these samples. Note that any comma-separated values (e.g. A,C indicate potential heterozygosity)
Chrom  Pos    Ref    out.snps.vcf4  Gene description      Status
U00096.3  1101543 T      A      curli production assembly%2Ftransport outer membrane lipoprotein  ,non-silent aaa -> Taa;
U00096.3  1169836 A      G      L%2CD-transpeptidase linking Lpp to murein  ,non-silent ctg -> cCg;
U00096.3  1189980 A      G      response regulator in two-component regulatory system with PhoQ  ,silent act -> acC;
U00096.3  1299464 T      G
U00096.3  1301992 A      T      oligopeptide transporter subunit  ,non-silent aat -> Tat;
U00096.3  1301999 G      A      oligopeptide transporter subunit  ,non-silent agc -> aAc;
U00096.3  1302190 A      G      oligopeptide transporter subunit  ,non-silent aac -> Gac;
U00096.3  1305442 T      G      oligopeptide transporter subunit  ,non-silent gtc -> gGc;
```

In later sections we will see how we can use this program to compare results between different strains.

Task 24: Check each variant in IGV

N.B. If a variant doesn't seem to match what the snp_report file says, check the reverse reading frames.

That concludes the first part of the course. You have successfully, QC'd, filtered, remapped and analysed a whole bacterial genome! Well done!

In the next installment we will be looking at how to extract and assemble unmapped reads. This will enable us to look at material which may be present in the strain of interest but not in the reference sequence.

2015 Workshop on Genomics

Part 3:

Short read genomics:

Assembly of unmapped reads

Instructors:

- Konrad Paszkiewicz k.h.paszkiewicz@exeter.ac.uk

Objectives:

By the end of this section you will be expected to be able to:

- Extract reads which do not map to the reference sequence.
- Assemble these reads de novo using SPAdes.
- Generate summary statistics for the assembly.
- Identify potential genes within the assembly.
- Search for matches within the ncbi database via BLAST.
- Visualize the taxonomic distribution of BLAST hits.
- Perform gene prediction and annotation using RAST.

3.1 Introduction

In this section of the lab we will continue the analysis of a strain of *E.coli*. In the previous section we cleaned our data, checked QC metrics, mapped our data and obtained a list of variants and an overview of any missing regions.

Now, we will examine those reads which did not map to the reference genome. We want to know what these sequences represent. Are they novel genes, plasmids or just contamination?

To do this we will extract unmapped reads, evaluate their quality, prepare them for de novo assembly, assemble them using SPAdes, generate assembly statistics and then produce some annotation via BLAST and RAST.

3.2 Extraction and QC of unmapped reads

Task 1: Extract the unmapped reads

First of all make sure you are in the
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter directory (hint: use the cd command). Then create a directory called unmapped_assembly in which we will do our de novo assembly and analysis.

```
cd unmapped_assembly/
```

Now we will use the bam2fastq program
(<http://gsl.hudsonalpha.org/information/software/bam2fastq>) to extract from the BAM file just those reads which did NOT map to the reference genome. The bam2fastq program has a number of options, most of which are self-explanatory. Type (all on one line):

```
bam2fastq --no-aligned -o unaligned#.fastq  
../remapping_to_reference/E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
```

The --no-aligned option means only extract reads which did not align. The -o unaligned\# means dump read 1 into a file called unaligned_1.fastq and read 2 into a file unaligned_2.fastq. Below we can see that the program has successfully created the two files.

```
[ec2-user@ip-10-171-67-183 unmapped_assembly]$ bam2fastq --no-aligned -o unaligned/#.fastq ../remapping_to_reference/E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
This looks like paired data from lane 8.
Output will be in unaligned/_1.fastq and unaligned/_2.fastq
1250574 sequences in the BAM file
351638 sequences exported
WARNING: 1414 reads could not be matched to a mate and were not exported
[ec2-user@ip-10-171-67-183 unmapped_assembly]$ bam2fastq --no-aligned -o unaligned#.fastq ../remapping_to_reference/E_Coli_CGATGT_L001_filtered.sorted.fillmd.rmdup.bam
This looks like paired data from lane 8.
Output will be in unaligned_1.fastq and unaligned_2.fastq
1250574 sequences in the BAM file
351638 sequences exported
WARNING: 1414 reads could not be matched to a mate and were not exported
```

Note that some reads were singletons (i.e. one read mapped to the reference, but the other did not). These will not be included in this analysis.

Task 2: Check that the number of entries in both fastq files is the same. Also check that the last few entries in the read 1 and read 2 files have the same header (i.e. that they have been correctly paired).

Task 3: Evaluate QC of unmapped reads

Use the fastqc program to look at the statistics and QC for the unaligned_1.fastq and unaligned_2.fastq files.

Do these look reasonably good? Remember, some reads will fail to map to the reference because they are poor quality, so the average scores will be lower than the initial fastqc report we did in the remapping section. The aim here is to see if it looks as though there are reads of reasonable quality which did not map.

Assuming these reads look ok, we will proceed with preparing them for de novo assembly.

De novo assembly

de novo is a Latin expression meaning "from the beginning," "afresh," "anew," "beginning again." when we perform a de novo assembly we try to reconstruct a genome or part of the genome from our reads without making any prior assumptions (in contrast to remapping where we compare our reads to what we think is a close reference sequence).

The advantage is that is that de novo assembly can reveal completely novel results, identify horizontal gene transfer events for example. This disadvantage is that it is difficult to get a good assembly from short reads and it can be prone to misleading results due to contamination and mis-assembly.

Task 4: Learn more about de novo assemblers

To understand more about de-novo assemblers, read the technical note at:

http://res.illumina.com/documents/products/technotes/technote_denovo_assembly_ecoli.pdf

N.B. You will also learn more in the next section so don't worry if it doesn't all make sense immediately. You should however understand the idea of the k-mer and broadly how the assembly is built up from them.

Task 5: Generate the assembly.

We will be using an assembler called SPAdes (<http://bioinf.spbau.ru/spades>). It generally performs pretty well with a variety of genomes. It can also incorporate longer reads produced from PacBio sequencers that we will use later in the course.

One big advantage is that it is not just a pure assembler - it is a suite of programs that prepare the reads you have, assembles them and then refines the assembly.

SPAdes runs the modules that are required for a particular dataset and it produces the assembly with a minimum of preparation and parameter selection - making it very straightforward to produce a decent assembly. As with everything in bio-informatics you should try to assess the results critically and understand the implications for further analysis.

Let's start the assembler because it takes about 20 minutes to run:

```
spades.py -k 21,33,55,77,99,127 --careful -o spades_assembly -1 unaligned_1.fastq -2 unaligned_2.fastq
```

We are telling it to run the SPAdes assembly pipeline with a range of k-mer sizes (-k); specifying --careful tells it to run a mismatch correction algorithm to reduce the number of errors; put the output in the spades_assembly directory and the reads to assemble.

Just because SPAdes does a lot for you does not mean you should not try to understand the process.

Have a read of this:

<http://thegenomefactory.blogspot.co.uk/2013/08/how-spades-differs-from-velvet.html>

It is a discussion of how SPAdes differs from Velvet another widely used assembler, it explains the overall process nicely:

"

1. Read error correction based on k-mer frequencies using [BayesHammer](#)
2. De Bruijn graph assembly at *multiple* k-mer sizes, not just a single fixed one.
3. Merging of different k-mer assemblies (good for varying coverage)
4. Scaffolding of contigs from paired end/mate pair reads
5. Repeat resolution from paired end/mate pair data using [rectangle graphs](#)
6. Contig error correction based on aligning the original reads with [BWA](#) back to contigs

"

Try to understand the steps in the context of the whole picture:

Can you explain why error correction of reads becomes more important as k-mer length increases?

When the assembly is complete:

```
===== Mismatch correction finished.
```

```
* Corrected reads are in /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/unmapped_assembly/spades_assembly2/corrected/
```

```
* Assembled contigs are in /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/unmapped_assembly/spades_assembly2/contigs.fasta (contigs.fastg)
```

```
* Assembled scaffolds are in /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/unmapped_assembly/spades_assembly2/scaffolds.fasta (scaffolds.fastg)
```

```
===== SPAdes pipeline finished.
```

```
SPAdes log can be found here: /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/unmapped_assembly/spades_assembly2/spades.log
```

```
Thank you for using SPAdes!
```

—

Change to the spades_assembly directory (use cd) and look at the output.

```
[ec2-user@ip-10-113-191-224 spades_assembly]$ ls -latr
total 2356
-rw-rw-r--. 1 ec2-user ec2-user 266 Dec 4 10:23 input_dataset.yaml
-rw-rw-r--. 1 ec2-user ec2-user 1440 Dec 4 10:23 params.txt
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec 4 10:27 corrected
-rw-rw-r--. 1 ec2-user ec2-user 145 Dec 4 10:27 dataset.info
drwxrwxr-x. 4 ec2-user ec2-user 4096 Dec 4 10:29 K21
drwxrwxr-x. 4 ec2-user ec2-user 4096 Dec 4 10:30 K33
drwxrwxr-x. 4 ec2-user ec2-user 4096 Dec 4 10:32 K55
drwxrwxr-x. 4 ec2-user ec2-user 4096 Dec 4 10:33 K77
drwxrwxr-x. 4 ec2-user ec2-user 4096 Dec 4 10:34 K99
drwxrwxr-x. 5 ec2-user ec2-user 4096 Dec 4 10:37 K127
-rw-rw-r--. 1 ec2-user ec2-user 361823 Dec 4 10:37 before_rr.fasta
-rw-rw-r--. 1 ec2-user ec2-user 363295 Dec 4 10:37 before_rr.fastg
drwxrwxr-x. 2 ec2-user ec2-user 4096 Dec 4 10:37 misc
-rw-rw-r--. 1 ec2-user ec2-user 360030 Dec 4 10:41 contigs.fasta
-rw-rw-r--. 1 ec2-user ec2-user 375980 Dec 4 10:41 contigs.fastg
-rw-rw-r--. 1 ec2-user ec2-user 360030 Dec 4 10:46 scaffolds.fasta
drwxrwxr-x. 2 ec2-user ec2-user 4096 Dec 4 10:46 tmp
-rw-rw-r--. 1 ec2-user ec2-user 375980 Dec 4 10:46 scaffolds.fastg
-rw-rw-r--. 1 ec2-user ec2-user 144114 Dec 4 10:46 spades.log
```

Let's take a look at some of the more important content.

params.txt

This contains a summary of the parameters used for assembly - it is useful so you can repeat the exact analysis performed, or can remember you setting when you want to publish the genome.

contigs.fasta

This contains the final results of the assembly in fasta format.

scaffolds.fasta

This contains the final results after scaffolding (which means using paired end information to join contigs together with gaps). In this case the files are identical, probably because the sum of the lengths of our paired reads is not much smaller than our insert size (there are very few large gaps between reads).

contigs/scaffolds.fastg

These contain the same results in fastg format - this is a slightly different format that contains more information than fasta - for example it can contain alternative alleles in diploid assemblies. We don't need it here, but see http://fastg.sourceforge.net/FASTG_Spec_v1.00.pdf if you might be working with diploid organisms.

Task 6: Assessment of the assembly

We will use QUAST (<http://bioinf.spbau.ru/quast>) to generate some statistics on the assembly.

```
quast.py --output-dir quast contigs.fasta
```

This will create a directory called quast and create some statistics on the assembly you produced.

```
cat quast/report.txt
```

```
ubuntu@ip-10-37-155-238:16:45: [~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/unmapped_assembly/spades_assembly]$ cat quast/report.txt
All statistics are based on contigs of size >= 500 bp, unless otherwise
noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include
all contigs).

Assembly      contigs
# contigs (>= 0 bp)    400
# contigs (>= 1000 bp) 14
Total length (>= 0 bp) 340803
Total length (>= 1000 bp) 131904
# contigs      280
Largest contig 67492
Total length   286988
GC (%)         43.33
N50            795
N75            551
L50            28
L75            144
# N's per 100 kbp 0.00
```

Try to interpret the information in the light of what we were trying to do. Because we are assembling unaligned reads we are not expecting a whole chromosome to pop out. We are expecting bits of our strain that does not exist in the reference we aligned against; possibly some contamination; various small contigs made up of reads that didn't quite align to our reference.

The N50 and L50 measures are very important in a normal assembly and we will visit them later, they are not really relevant to this assembly.

You will notice that we have 1 contig 67kb long - what do you think this might be? And 14 other contigs longer than 1kb. We need to find out what this stuff is.

Analysing the de novo assembled reads

Now that we have assembled the reads and have a feel for how much (or in this case, how little) data we have, we can set about analysing it. By analysing, we mean identifying which genes are present, which organism they are from and whether they form part of the main chromosome or are an independent unit (e.g. plasmid).

We are going to take a 3-prong approach. The first will simply search the nucleotide sequences of the contigs against the NCBI non-redundant database. This will enable us to identify the species to which a given contig matches best (or most closely). The second will call open reading frames

within the contigs and search those against the Swissprot database of manually curated (i.e. high quality) annotated protein sequences.

Why not just search the NCBI blast database? Well, remember nearly all of our biological knowledge is based on homology – if two proteins are similar they probably share an evolutionary history and may thus share functional characteristics. Metrics to define whether two sequences are homologous are notoriously difficult to define accurately. If two sequences share 90% sequence identity over their length, you can be pretty sure they are homologous. If they share 2% they probably aren't. But what if they share 30%? This is the notorious twilight zone of 20-30% sequence identity where it is very difficult to judge whether two proteins are homologous based on sequence alone.

To help overcome this searching more subtle signatures may help – this is where Pfam comes in. Pfam is a database which contains protein families identified by particular signatures or patterns in their protein sequence. These signatures are modeled by Hidden Markov Models (HMMs) and used to search query sequences. These can provide a high level annotation where BLAST might otherwise fail. It also has the advantage of being much faster than BLAST. In the interest of time, we aren't going to run this today but think about it for your own data.

Task 7: Search contigs against NCBI non-redundant database

Firstly we can filter out low coverage and very short contigs using a perl script:

```
filter_low_coverage_contigs.pl < contigs.fasta > contigs.goodcov.fasta
```

The following command executes a nucleotide BLAST search (blastn) of the sequences in the contigs.fasta file against the non-redundant database. As this takes over half an hour to process, the results have been precomputed in

~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/blast_precompute/unmapped_reads/ but the command would be as follows:

```
blastn -db ~/workshop_data/genomics_tutorial/db/blast/nt -query contigs.goodcov.fasta  
-evalue 1e-06 -num_threads 2 -show_gis -num_alignments 10 -num_descriptions 10 -out  
contigs.fasta.blastn
```

There are a lot of options in this command, let's go through them,

- **-db** is the prepared blast database to search
- **-evalue** apply an e-value (expectation value) cutoff (<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>) cutoff of 1e-06 to limit ourselves to statistically significant hits (i.e. in this case 1 in 1 million likelihood of a hit to a database of this size by a sequence of this length).
- **-num_alignments** and **-num_descriptions** flags tell blastn to only display the top 10 results for each hit,
- **-num_threads** that it should use 2 CPU cores
- **-show_gis** that it should include general identifier (GI) numbers in the output.
- **-out** file in which to place the output.

There is lots of information on running blast from the command line at <http://www.ncbi.nlm.nih.gov/books/NBK1763/>

Open the results file

```
pluma contigs.fasta.blastn
```

BLASTN 2.2.28+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Database: Nucleotide collection (nt)
29,442,065 sequences; 84,823,117,434 total letters

Query= NODE_2_length_18538_cov_20.5315_ID_3

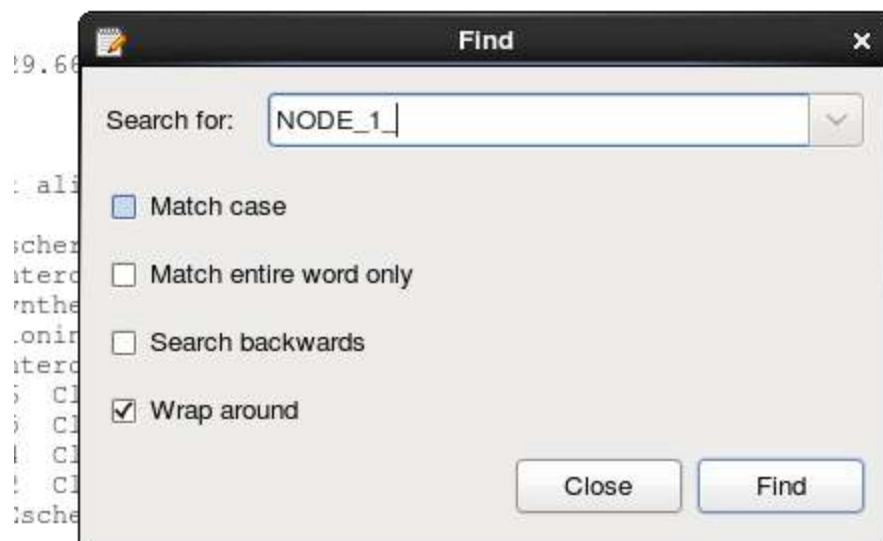
Length=18538

Sequences producing significant alignments:			Score (Bits)	E Value
gi 392506193 gb JQ182735.1	Enterobacteria phage HK629, complete...		3.422e+04	0.0
gi 383395315 gb JQ086376.1	Enterobacteria phage HK630, complete...		3.422e+04	0.0
gi 313848522 emb AM946981.2	Escherichia coli BL21(DE3), complet...		3.422e+04	0.0
gi 296142109 gb CP001509.3	Escherichia coli BL21(DE3), complete...		3.422e+04	0.0
gi 194021541 gb EU078592.1	Enterobacteria phage DE3, complete g...		3.422e+04	0.0
gi 253322479 gb CP001665.1	Escherichia coli 'BL21-Gold(DE3)pLys...		3.422e+04	0.0
gi 215104 gb J02459.1 LAMCG	Enterobacteria phage lambda, complet...		3.422e+04	0.0
gi 126032369 gb AC198536.1	Monosiga brevicollis clone JGIACYI-5...		3.421e+04	0.0
gi 126012625 gb AC198467.1	Monosiga brevicollis clone JGIACYI-5...		3.421e+04	0.0
gi 68532089 gb AC150248.3	Uncultured bacterium clone zdt-9n2, c...		3.421e+04	0.0

...

Search for our largest contig - SPAdes names the contigs by increasing size, so

click on “Search” and then “Find” and enter NODE_1_



Query= NODE_1_length_67492_cov_565.407_ID_1

Length=67492

Sequences producing significant alignments:			Score (Bits)	E Value
gi 664682453 gb CP008801.1	Escherichia coli KLY, complete genome		79013	0.0
gi 8918823 dbj AP001918.1	Escherichia coli K-12 plasmid F DNA, ...		78976	0.0
gi 619497957 gb KJ170699.1	Escherichia coli strain K-12 plasmid...		65330	0.0
gi 665821556 gb KJ484626.1	Escherichia coli plasmid pH2332-166,...		65302	0.0
gi 665821958 gb KJ484628.1	Escherichia coli plasmid pH2291-144,...		65213	0.0
gi 28629230 gb AF550679.1	Escherichia coli plasmid p1658/97, co...		64591	0.0
gi 4874241 gb U01159.2	Escherichia coli F sex factor transfer r...		61474	0.0

gi 665822931 gb KJ484636.1	Escherichia coli plasmid pC59-153, c...	41227	0.0
gi 301130432 gb CP002090.1	Salmonella enterica subsp. enterica ...	41026	0.0
gi 301130304 gb CP002089.1	Salmonella enterica subsp. enterica ...	41026	0.0

There are a number of good hits; notice from the contig header line that the average coverage is >500 and the coverage of our genome was around 50 - does this give you a clue to what it is?

Task 8: Obtain open reading frames

The first task is to call open reading frames within the contigs. These are designated by canonical start and stop codons and are usually identified by searching for regions free of stop codons. We will use the EMBOSS package program getorf to call these.

We will use codon table 11 which defines the bacterial codon usage table (<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>) and state that the sequences we are dealing with are not circular (they are nowhere near long enough!). We will also restrict the ORFs to just those sequences longer than 300 nucleotides (i.e. 100 amino acids). We will store the results in file contigs.orf.fa.

```
getorf -table 11 -circular N -minsize 300 -sequence contigs.goodcov.fasta -outseq
contigs.orf.fasta
```

If we look at the output file we can see that it is a FASTA formatted file containing the name of the contig on which the ORF occurs, followed by an underscore and a number (e.g. _1) to indicate the number of the ORF on that contig. The numbers in square brackets indicate the start and end position of the ORF on the contig (i.e. in nucleotide space). So the first ORF occurs on NODE 2 and is between position 371 and 682. The second ORF occurs between positions 1 and 1641.

Also note that many ORFs do not start with a Methionine. This is because by default the getorf program calls ORFs between stop codons rather than start and stop codons. Primarily this is to avoid spurious ORFs due to Met residues within a protein sequence and to ensure untranslated regions are captured.

```
ubuntu@ip-10-37-155-238:16:59: [~/workshop_data/genomics_tutorial/data/se
quencing/ecoli_exeter/unmapped_assembly/spades_assembly]$ head contigs.o
rf.fasta
>NODE_2_length_18538_cov_20.5315_ID_3_1 [371 - 682]
EAPVRLSVQPVNPRILCVFMLPARIAGRSSILNLATKRRRLASNGRRMTPPACFISASIM
PASSASRSWTLMLPVISAKRPGSGPVMAFSGFRHPVKRLSHLTV
>NODE_2_length_18538_cov_20.5315_ID_3_2 [1 - 1641]
EHKQRTLIWLPTDGDENFMKTHVEPTIRDIPSLALAPWYGKKHRDNTLTMKRFTNGR
GFWCLGGKAAKNYREKSVVDVAGYDELAADFDDIEQEGSPTFLGDKRIEGSVWPKSIRGST
PKVRGTCQIERAASESPHFMRFHVACPHCGEEQYLKFGDKETPFGWKWTPDDPSSVFYLC
EHNACVIRQQELDFTDARYICEKGTGIWTRDGILWFSSSGEEIEPPDSVTFHIWTAYSPFT
TWVQIVKDWMMKTKGDTGKRKTFVNTTLGETWEAKIGERPDAEVMAERKEHYSAPVPDRVA
YLTAGIDSQLDREYMRVWGWGPGEESWLIDRQIIMGRHDDEQTLRLRVDEAINKTYTRNG
```

Task 9: Search open reading frames against NCBI non-redundant database

The first thing we can do with these open reading frames is to search them against the NCBI non-redundant database of protein sequences to see what they may match.

Here we will perform a BLAST search using the non-redundant (nr) database, using the blastp program and store the results in contigs.orf.blastp. We'll apply an e-value (expectation value) (<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>) cutoff of 1e-06 to limit ourselves to statistically significant hits (i.e. in this case 1 in 1 million likelihood of a hit to a database of this size by a sequence of this length). The `-num_alignments` and `num_descriptions` flags tell blastp to only display the top 10 results for each hit, the `num_threads` tells blastp to use 2 CPU cores and `-show_gis` tells blastp it should include general identifier (GI) numbers in the output.

However this command takes several hours to complete, therefore the results have been pre-computed and are available for you in

`~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/blast_precompute/unmapped_reads`

If you were going to run the command, you would type (all on one line):

```
blastp -db ~/workshop_data/genomics_tutorial/db/blast/nr -query contigs.orf.small.fasta  
-evalue 1e-06 -num_threads 2 -show_gis -num_alignments 10 -num_descriptions 10 -out  
contigs.orf.blastp
```

Task 10: Review the BLAST format

Open the results file with pluma and search for plasmid in the text. You should find a number of hits to plasmid related proteins - one example is below - can you find any others? This evidence is not conclusive, but combined with the high coverage over, it is starting to look like this contig is a plasmid.

Query= NODE_1_length_67492_cov_565.407_ID_1_32 [31455 - 31889]

Length=145

Sequences producing significant alignments:	Score	E
	(Bits)	Value
gi 446834068 ref WP_000911324.1 MULTISPECIES: pirin	275	3e-92
gi 446834058 ref WP_000911314.1 pirin	273	1e-91
gi 446834061 ref WP_000911317.1 pirin	271	1e-90
gi 446834059 ref WP_000911315.1 pirin	269	6e-90
gi 545289568 ref WP_021572485.1 hypothetical protein	269	6e-90
gi 446834062 ref WP_000911318.1 MULTISPECIES: pirin	269	6e-90
gi 585223672 ref WP_024168023.1 plasmid maintenance protein	268	9e-90
gi 723058272 ref WP_033552985.1 plasmid maintenance protein	268	9e-90
gi 446834056 ref WP_000911312.1 plasmid maintenance protein	268	1e-89
gi 446834060 ref WP_000911316.1 pirin	268	1e-89

```
>gi|446834068|ref|WP_000911324.1| MULTISPECIES: pirin [Escherichia]
gi|32470009|ref|NP_862949.1| plasmid maintenance protein [Escherichia coli]
gi|689926354|ref|YP_009060131.1| PIN domain protein [Escherichia coli]
gi|691230621|ref|YP_009070585.1| VapC toxin protein [Escherichia coli]
gi|28629266|gb|AA049546.1| hypothetical protein [Escherichia coli]
gi|323184064|gb|EFZ69443.1| PIN domain protein [Escherichia coli OK1357]
gi|325495739|gb|EGC93600.1| plasmid maintenance protein [Escherichia fergusonii ECD227]
gi|385154377|gb|EIF16391.1| plasmid maintenance protein [Escherichia coli O32:H37 str. P4]
```

Additional checks:

Task 11: Check that the contigs do not appear in the reference sequence

In theory, the unmapped reads used to generate the contigs should not assemble into something which will map against the genome. However, it is always possible that this might happen. Especially with more complex genomes with large numbers of repetitive elements or high/low GC-content. To double check:

```
blastn -subject ~/workshop_data/genomics_tutorial/data/reference/U00096/U00096.fna  
-query contigs.goodcov.fasta | more
```

Here we use the BLAST+ package in a different mode to compare two sequences against each other. Unlike the previous examples where we have searched against a database of sequences, here we are doing a simple search of the contigs against the reference genome we are using. Scroll down a little...

Query= NODE_18_length_917_cov_10.3076_ID_35

Length=917

Subject= U00096.3

Length=4641652

Score = 193 bits (104), Expect = 3e-49
Identities = 186/227 (82%), Gaps = 0/227 (0%)
Strand=Plus/Plus

Query	68	ACGGCATCCACGAAGGCGACAGAGGCTGCGGGAAGTGC GGTATCAGCATCGCAGAGCAAA	127
Sbjct	1430285	ACGGCATCCACGAAGGCGACAGAGGCTGCTGGCAGTGCGCACGGCGGCAGCTCAGAGCAAA	1430344

You can see that some of the contigs that have been assembled hit the reference sequence. In the record above the eval is $3e-49$ which may look like a significant hit; however, the eval is calculated as the chance of a hit this close against a random sequence of the same size. Since our subject sequence is now very small and we know it is related to our strain it is not surprising that there are some hits. We are concerned about whole contigs that map closely to the reference genome. Overall there are only three hits and none cover the entire length of the contig, so we can be happy that this is not happening.

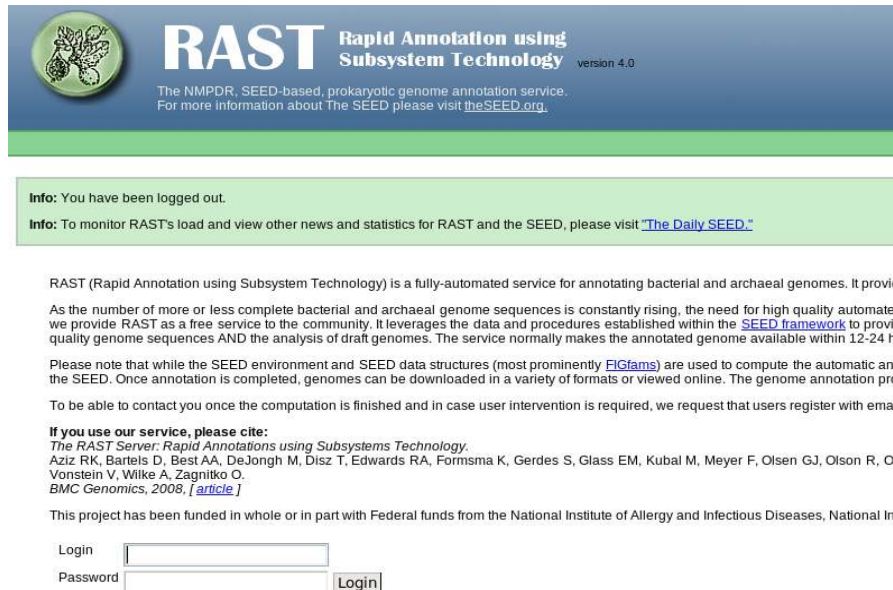
Analysing the results in RAST

By now you should be able to see that analysing results for de novo assembled reads of any sort can be difficult and time-consuming. Bear in mind that we have only been faced with a single contig of 3kb. Quite often you may find yourself dealing with hundreds, if not thousands of contigs. Some will be a few 100kb long. Others may only be 200-300bp. How should we go about analysing these in a more efficient manner? There are a number of options here. For eukaryotes I would suggest looking at MAKER (<http://www.yandell-lab.org/software/maker.html>). For prokaryotes the situation is somewhat easier and we can use a web-based service known as RAST. This is not the only service (Xbase is another), but it is one of the most common.

RAST is a website where you upload the results of your de novo assembly and RAST will attempt to provide annotation in commonly used GFF and Genbank formats. This can be used to load up the annotation in Artemis or Apollo. Alternatively RAST has its own in-built viewer.

Task 12: Log in to RAST

Within your instance, go to <http://rast.nmpdr.org/> If you already have an account, log-in with the details RAST provided to you. If you do not have one, you may need to wait several days for your login to be issued by RAST. Please skip ahead and come back to this section.



RAST Rapid Annotation using Subsystem Technology version 4.0
The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

Info: You have been logged out.
Info: To monitor RAST's load and view other news and statistics for RAST and the SEED, please visit ["The Daily SEED."](#)

RAST (Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating bacterial and archaeal genomes. It provides the number of more or less complete bacterial and archaeal genome sequences is constantly rising, the need for high quality automated we provide RAST as a free service to the community. It leverages the data and procedures established within the [SEED framework](#) to provide quality genome sequences AND the analysis of draft genomes. The service normally makes the annotated genome available within 12-24 hours.

Please note that while the SEED environment and SEED data structures (most prominently [FIGfams](#)) are used to compute the automatic annotation the SEED. Once annotation is completed, genomes can be downloaded in a variety of formats or viewed online. The genome annotation provided.

To be able to contact you once the computation is finished and in case user intervention is required, we request that users register with email.

If you use our service, please cite:
The RAST Server: Rapid Annotations using Subsystems Technology.
Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Ost Vonstein V, Wilke A, Zagnitko O.
BMC Genomics, 2008, [[article](#)]

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

Login
Password

Task 13: Upload the assembled contigs and annotate using RAST

Click on Your jobs->Upload New Job

Upload a Genome

A prokaryotic genome in one or more contigs should be uploaded in either a single [FASTA](#) format file or in a Genbank format file. Our pipeline will use the taxonomy identifier as a handle for the genome. Therefore i and genus, species and strain in the following upload workflow.

Please note, that only if you submit all relevant contigs (i.e. all chromosomes, if more then one, and all plasmids) that comprise the genomic information of your organism of interest in one job, Features like *Metabolic* picture.

Confidentiality information: Data entered into the server will not be used for any purposes or in fact integrated into the main SEED environment, it will remain on this server for 120 days or until deleted by the subm

If you use the results of this annotation in your work, please cite:

The RAST Server: Rapid Annotations using Subsystems Technology.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pi A, Zagnitko O.

BMC Genomics, 2008, [[article](#)]

File formats: You can either use [FASTA](#) or Genbank format.

- If in doubt about FASTA, [this service](#) allows conversion into FASTA format.
- Due to limits on identifier sizes imposed by some of the third-party bioinformatics tools that RAST uses, we limit the size of contig identifiers to 70 characters or fewer.
- If you use Genbank, you have the option of preserving the gene calls in the options block below. By default, genes will be recalled.

Please note: This service is intended for complete or nearly complete prokaryotic genomes. For now we are not able to reliably process sequence data of very small size, like small plasmid, phages or fragments.

File Upload:

Sequences File

Browse...

Use this data and go to step 2

Upload the contigs.fasta file obtained by the de novo assembly of unmapped reads. Click on “Use this data and go to step 2”.

Review genome data

We have analyzed your upload and have computed the following information.

Contig statistics

Statistic	As uploaded	After splitting into scaffolds
Sequence size	3326	3326
Number of contigs	1	1
GC content (%)	43.4	43.4
Shortest contig size	3326	3326
Median sequence size	3326	3326
Mean sequence size	3326.0	3326.0
Longest contig size	3326	3326

Please enter or verify the following information about this organism:

Required information:

Taxonomy ID: (leave blank if NCBI Taxonomy ID unknown)
[Look up taxonomy ID at NCBI.](#)
Find the taxonomy id for your organism by searching for its name in the [NCBI taxonomy browser](#).

Taxonomy string:

Domain: ☒ Bacteria ☐ Archaea ☐ Virus

Genus:

Species:

Strain:

Genetic Code: ☒ 11 (Archaea, most Bacteria, most Virii, and some Mitochondria)
☐ 4 (Mycoplasmaeae, Spiroplasmaeae, Ureoplasmaeae, and Fungal Mitochondria)

[Use this data and go to step 3](#)

We know this is an *E.coli* genome so we can enter 562 as the Taxonomy ID and click on 'Fill in form based on NCBI taxonomy-ID'. If you're dealing with a different organism, be sure to change this number. RAST will automatically split any scaffolds (i.e. contigs with bits missing in the middle – denoted by Ns). Then click “Use this data and go to step 3”.

Upload a Genome

Complete Upload

Please consider the following options for the RAST annotation pipeline:

RAST Annotation Settings:		
Choose RAST annotation scheme	Classic RAST ▼	Choose "Classic RAST" for the current production RAST, or "RASTtk" for the new modular RAS
Select gene caller	RAST ▼	Please select which type of gene calling you would like RAST to perform. Note that using GLIM backfilling of gaps.
Select FIGfam version for this run	Release70 ▼	Choose the version of FIGfams to be used to process this genome.
Automatically fix errors?	☒ Yes	The automatic annotation process may run into problems, such as gene candidates overlappin these problems (even if that requires deleting some gene candidates), please check this box.
Fix frameshifts?	☐ Yes	If you wish for the pipeline to fix frameshifts, check this option. Otherwise frameshifts will not
Build metabolic model?	☒ Yes	If you wish RAST to build a metabolic model for this genome, check this option.
Backfill gaps?	☒ Yes	If you wish for the pipeline to blast large gaps for missing genes, check this option.
Turn on debug?	☐ Yes	If you wish debug statements to be printed for this job, check this box.
Set verbose level	0	Set this to the verbosity level of choice for error messages.
Disable replication	☐ Yes	Even if this job is identical to a previous job, run it from scratch.

Finish the upload

Replicate the settings above and click on 'Finish the upload'.

Your job may take several hours to run. In the meantime, proceed to the next section and come back to this later.

Once complete, RAST should email you a message. You can then view the results or download them in standardized formats (e.g. GFF3, Genbank, EMBL etc).

On the start page click on view details for your annotation

Progress bar color key:

- not started
- queued for computation
- in progress
- requires user input
- failed with an error
- successfully completed

Jobs you have access to :

Job ▲▼	Owner ▲▼	ID ▲▼	Name ▲▼	Num contigs ▲▼	Size (bp) ▲▼	Creation Date	Annotation Progress	Status
205173	O'Neill, Paul	562.4461	Escherichia coli	15	129530	2014-12-04 10:33:46	[view details]	complete

You will get a summary of the sequence you uploaded and you have the ability to download the annotations to your computer

Job Details #205173

» [Browse annotated genome in SEED Viewer](#)

» [View metabolic model](#)

» Available downloads for this job: GFF3 ▼ Download Update download files

Download the GFF3 annotation and open it in a text editor (this may be in your Downloads folder)

NOTE: your output may be different.

Scan down the list of annotations do any themes stand out?

```
##gff-version 3
NODE_10_length_3324_cov_22.7003_ID_19      FIG    CDS    249    1163    .    -    0
ID=fig|562.4461.peg.1;Name=FIG010773: NAD-dependent epimerase/dehydratase
NODE_10_length_3324_cov_22.7003_ID_19      FIG    CDS    1160   2782    .    -    2
ID=fig|562.4461.peg.2;Name=FIG022758: Long-chain-fatty-acid--CoA ligase (EC
6.2.1.3);Ontology_term=KEGG_ENZYME:6.2.1.3
NODE_1_length_67492_cov_565.407_ID_1 FIG    CDS    151    927     .    -    1
ID=fig|562.4461.peg.3;Name=FIG00638373: hypothetical protein
NODE_1_length_67492_cov_565.407_ID_1 FIG    CDS    973    1407    .    -    1
ID=fig|562.4461.peg.4;Name=YcgB
NODE_1_length_67492_cov_565.407_ID_1 FIG    CDS    1421   1642    .    -    2
ID=fig|562.4461.peg.5;Name=putative cytoplasmic protein
NODE_1_length_67492_cov_565.407_ID_1 FIG    CDS    1643   2326    .    -    2
ID=fig|562.4461.peg.6;Name=Adenine-specific methyltransferase (EC
2.1.1.72);Ontology_term=KEGG_ENZYME:2.1.1.72
NODE_1_length_67492_cov_565.407_ID_1 FIG    CDS    2434   2559    .    +    1
ID=fig|562.4461.peg.7;Name=hypothetical protein
```

```

NODE_1_length_67492_cov_565.407_ID_1 FIG      CDS      2573    2761    .      +      2
ID=fig|562.4461.peg.8;Name=FIG00639560: hypothetical protein
NODE_1_length_67492_cov_565.407_ID_1 FIG      CDS      2712    3149    .      -      0
ID=fig|562.4461.peg.9;Name=FIG01048508: hypothetical protein

```

From the job details page:

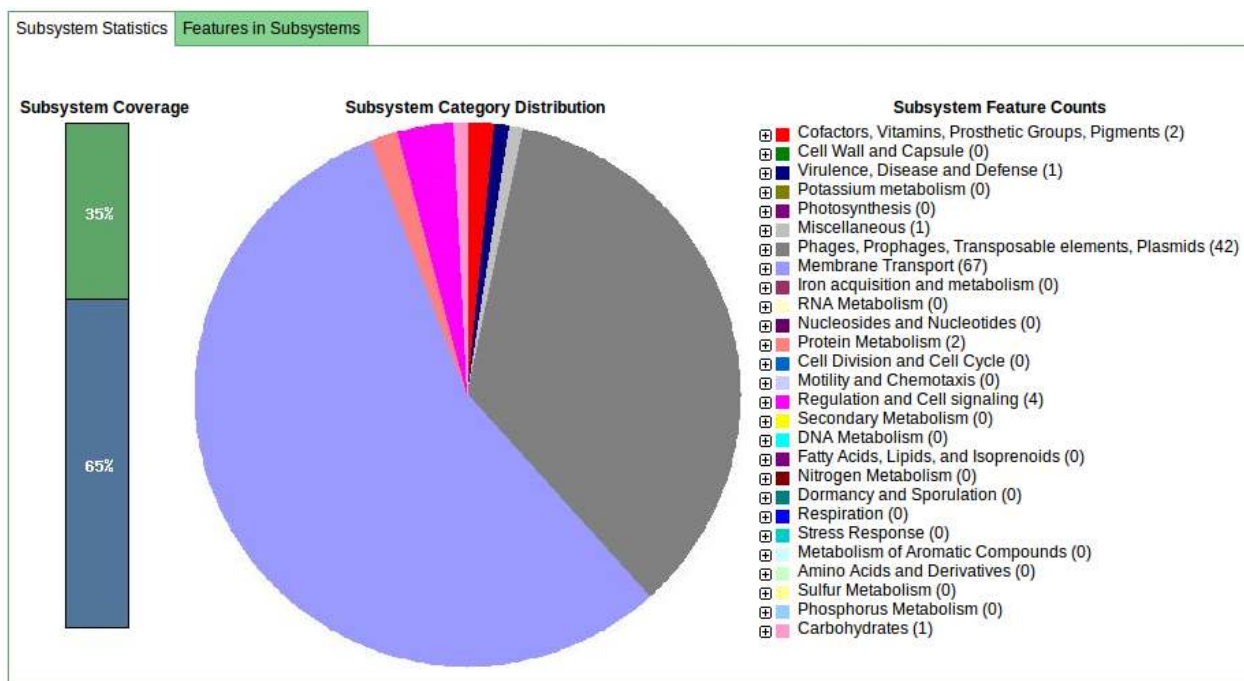
Job Details #205173

» [Browse annotated genome in SEED Viewer](#)

» [View metabolic model](#)

» Available downloads for this job:

Click on 'Browse annotated genome in SEED viewer'



This gives you a hierarchical view of the subsystems.

Browse the rest of the RAST server and get a feel for the possibilities the platform may offer you.

When you're ready, move on to (or back to) the de novo assembly part of the lab.

2015 Workshop on Genomics

Part 4

Short read genomics:

De-novo assembly

Instructors:

- Konrad Paszkiewicz k.h.paszkiewicz@exeter.ac.uk

Objectives:

By the end of this section you will be expected to be able to:

- Perform QC and adaptor-trim Illumina reads.
- Assemble these reads de novo using SPAdes.
- Generate summary statistics for the assembly.
- Understand how to incorporate long PacBio reads into the assembly.
- Identify open reading frames within the assembly.
- Search for matches within the NCBI database via BLAST.
- Visualize species distribution of potential matches.

4.1 Introduction

In this section of the lab we will continue the analysis of a strain of *E.coli*. In the previous section we extracted those reads which did not map to the reference genome and assembled them. However, it is often necessary to be able to perform a de novo assembly of a genome. In this case, rather than doing any remapping, we will start with the filtered reads we obtained in part 3 of the lab.

To do this we will use a program called SPAdes to try to get the best possible assembly for a given genome. We will then generate assembly statistics and then produce some annotation via BLAST.

Task 1 Start the Assembly

The assembly takes a while so the results have been prepared for you. See the `~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/denovo_assembly` directory.

However, if you did want to run the command, you would type the following (on one line):

```
spades.py -t 2 -o denovo_assembly_2 -1 E_Coli_CGATGT_L001_R1_001.filtered.fastq -2  
E_Coli_CGATGT_L001_R2_001.filtered.fastq
```

This will create a directory called `denovo_assembly` to hold the results.

Assembly theory

We are using SPAdes (<http://bioinf.spbau.ru/en/spades>) to perform our assembly. It is a de Bruijn graph based assembler, similar to other short read assemblers like velvet (<https://www.ebi.ac.uk/~zerbino/velvet/>). The advantage of SPAdes is that it does a lot of error correction and checking before and after the assembly which improve the final result. A downside of SPAdes is that it was designed for assembling reads from a single cell and although it does a good job with DNA prepared from a community it can leave in some low coverage sequences which are likely to be artifacts.

You can read more about the comparison here

<http://thegenomefactory.blogspot.co.uk/2013/08/how-spades-differs-from-velvet.html>

SPAdes is also very easy to use - apart from telling it where your input files are the only parameter that you might want to choose is the length of k-mer.

K-mer length. Rather than store all reads individually which would be unfeasible for Illumina type datasets, de Bruijn assemblers convert each read to a series of k-mers and stores each k-mer once, along with information about how often it occurs and which other k-mers it links to. A short k-mer length (e.g. 21) reduces the chance that data will be missed from an assembly (e.g. due to reads

being shorter than the k-mer length or sequencing errors in the k-mer), but can result in shorter contigs as repeat regions cannot be resolved. .

When using the Velvet assembler it is necessary to try a large combination of parameters to ensure that you obtain the 'best' possible assembly for a given dataset. There is even a program called VelvetOptimiser which does it for you. However, what 'best' actually means in the context of genome assembly is ill-defined. For a genomic assembly you want to try to obtain the lowest number of contigs, with the longest length, with the fewest errors. However, although numbers of contigs and longest lengths are easy to evaluate, it is extremely difficult to know what is or isn't an error when sequencing a genome for the first time.

SPAdes allows you to choose more than one k-mer length - it then performs an assembly for each k-mer and merges the result - trying to get the best of both worlds. It actually has some pre-calculated k-mer settings based on the length of reads you have, so you don't even have to choose that.

Let's look at the assembly process in more detail:

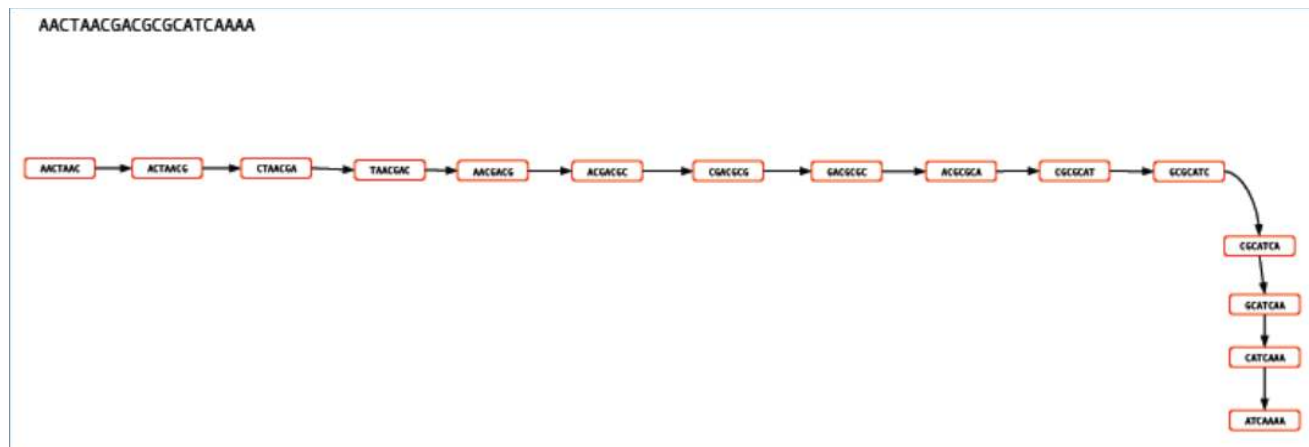
Description of k-mers:

What are they? Let's say you have a single read:

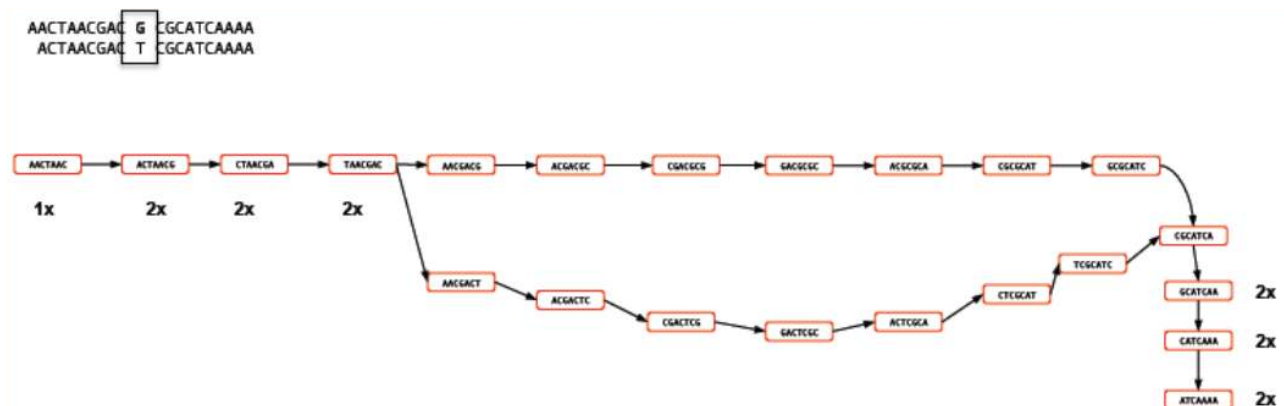
AACTAACGACGCGCATCAAAA

The set of k-mers obtained from this read with length 6 (i.e. 6-mers) would be obtained by taking the first six bases, then moving the window along one base, taking the next 6 bases and so-on until the end of the read. E.g:





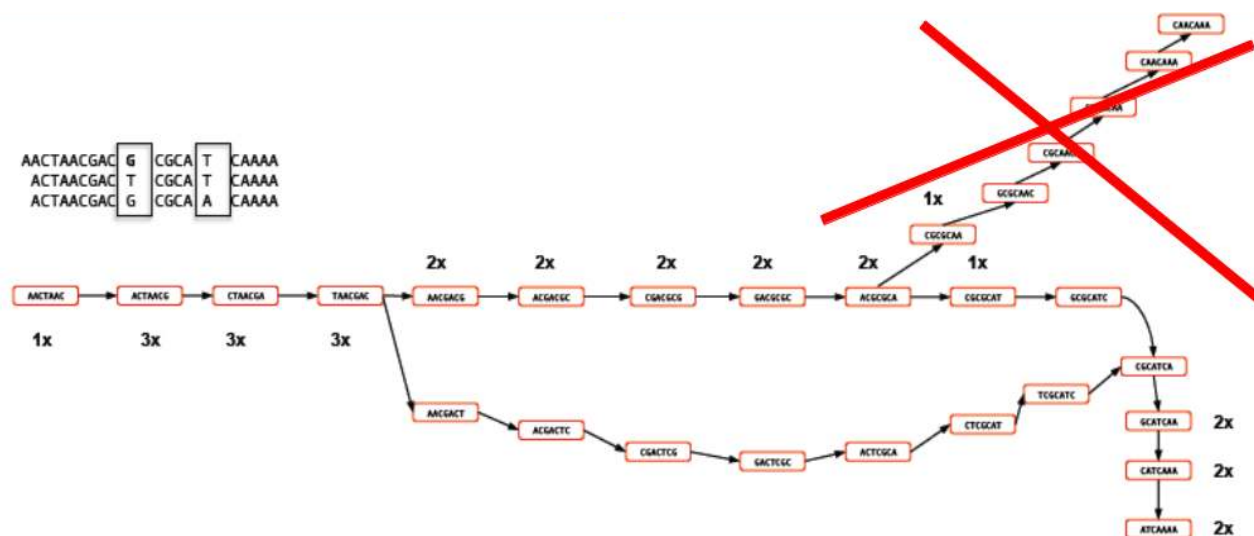
You may well ask, “So what? How does that help”? For a single read, it really doesn't help. However let's say that you have another read which is identical except for a single base:



Rather than represent both reads separately, we need only store the k-mers which differ and the number of times they occur. Note the 'bubble' like structure which occurs when a single base-change occurs. This kind of representation of reads is called a 'k-mer graph' (sometimes inaccurately referred to as a de-bruijn graph).

Now let's see what happens when we add in a third read. This is identical to the first read except for a change at another location. This results in an extra dead-end being added to the path.

The job of any k-mer based assembler is to find a path through the k-mer graph which correctly represents the genome sequence.



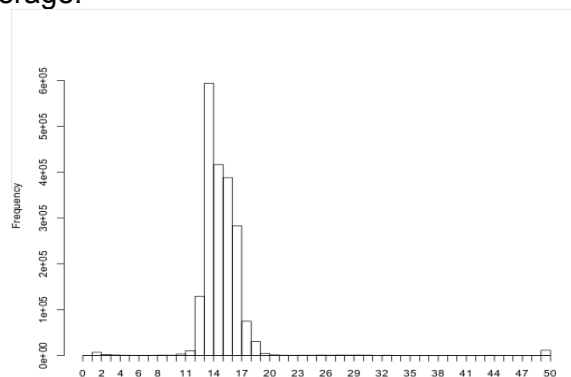
Images courtesy of Mario Caccamo

Description of coverage cutoff:

In the figure above, you can see that the coverage of various k-mers varies between 1x and 3x. The question is which parts of the graph can be trimmed or removed so that we avoid any errors. As the graph stands, we could output three different contigs as there are three possible paths through the graph. However, we might wish to apply a coverage cutoff and remove the top right part of the graph because it has only 1x coverage and is more likely to be an error than a genuine variant.

In a real graph you would have millions of k-mers and thousands of possible paths to deal with. The best way to estimate the coverage cutoff in such cases is to look at the frequency plot of contig (node) coverage, weighted by length. In the example below you can see that contigs with a coverage below 7x or 8x occur very infrequently. As such it is probably a good idea to exclude those contigs which have coverage less than this – they are likely to be errors.

Description of expected coverage:

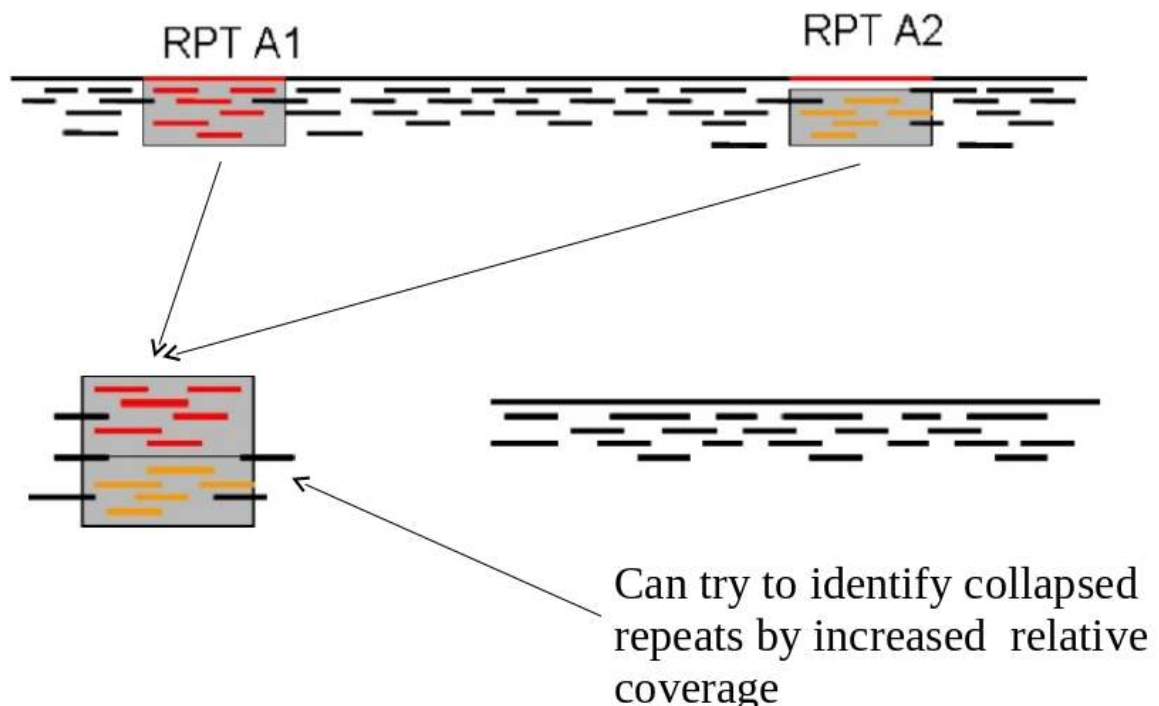


In the example below you can see a stretch of DNA with many reads mapping to it. There are two repetitive regions A1 and A2 which have identical sequence. If we try to assemble the reads without

any knowledge of the true DNA sequence, we will end up with an assembly that is split into two or more contigs rather than one.

One contig will contain all the reads which did not fall into A1 and A2. The other will contain reads from both A1 and A2. As such the coverage of the repetitive contig will be twice as high as that of the non-repetitive contig.

If we had 5 repeats we would expect 5x more coverage relative to the non-repetitive contig. As such, provided we know what level of coverage we expect for a given set of data, we can use this information to try and resolve the number of repeats we expect.



A commonly used metric to describe the effectiveness of the assembly is called N50 - see http://en.wikipedia.org/wiki/N50_statistic for details.

When SPAdes has finished move onto the next section.

```
===== SPAdes pipeline finished WITH WARNINGS!

=== Pipeline warnings:
* Default k-mer sizes were set to [21, 33, 55, 77, 99, 127] because estimated read length (301) is equal to or greater than 250
===== Warnings saved to /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/denovo_assembly/warnings.log

SPAdes log can be found here: /home/ec2-user/genomics_tutorial/data/sequencing/ecoli_exeter/denovo_assembly/spades.log

Thank you for using SPAdes!
```

Task 2 Checking the assembly

Firstly we can filter out low coverage and very short contigs using a perl script:

```
filter_low_coverage_contigs.pl < contigs.fasta > contigs.goodcov.fasta
```

We will use QUAST again (<http://bioinf.spbau.ru/quast>) to generate some statistics on the assembly.

```
quast.py --output-dir quast contigs.goodcov.fasta
```

This will create a directory called quast and create some statistics on the assembly you produced.

```
cat quast/report.txt
```

```
Assembly                                contigs.goodcov
# contigs (>= 0 bp)                     81
# contigs (>= 1000 bp)                  67
Total length (>= 0 bp)                  4689514
Total length (>= 1000 bp)               4679794
# contigs                               81
Largest contig                          293215
Total length                            4689514
GC (%)                                  50.72
N50                                      136627
N75                                      95318
L50                                      12
L75                                      21
# N's per 100 kbp                       0.00
```

You can see that there are 81 contigs in the assembly - so it is still far from complete. The N50 is 136K and the N75 is 95K so most of the assembly is in quite large contigs.

This is fairly normal for a short read assembly - don't expect complete chromosomes.

A good check at this point is to map the original reads back to the contigs.fasta file and check that all positions are covered by reads. Amazingly it is actually possible for de-novo assemblers to generate contigs to which the original reads will not map.

Task 3: Map reads back to assembly

Here we will use BWA again to index the contigs.fasta file and remap the reads. This is almost identical to the procedure we followed in Part 3, the only difference is that instead of aligning to the reference genome, we are aligning to our newly created reference.

Make sure you are in the following directory:

```
~/workshop_data/genomics_tutorial/data/sequencing/ecoli_exeter/denovo_assembly/
```

Let's create a subdirectory to keep our work separate

```
mkdir remapping_to_assembly
cd remapping_to_assembly
cp ../contigs.goodcov.fasta .
```

Let's start by indexing the contigs.fasta file. Type:

```
bwa index contigs.goodcov.fasta
```

```
[ec2-user@ip-10-181-110-211 remapping_to_assembly]$ bwa index contigs.fasta
[bwa_index] Pack FASTA... 0.07 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 2.04 seconds elapse.
[bwa_index] Update BWT... 0.06 sec
[bwa_index] Pack forward-only FASTA... 0.04 sec
[bwa_index] Construct SA from BWT and Occ... 0.68 sec
[main] Version: 0.7.10-r789
[main] CMD: bwa index contigs.fasta
[main] Real time: 6.961 sec; CPU: 2.901 sec
[ec2-user@ip-10-181-110-211 remapping_to_assembly]$ █
```

Once complete we can start to align the reads back to the contigs. Type (all on one line):

```
bwa mem -t 4 contigs.goodcov.fasta ../../E_Coli_CGATGT_L001_R1_001.filtered.fastq
../../E_Coli_CGATGT_L001_R2_001.filtered.fastq > E_Coli_CGATGT_L001_filtered.sam
```

Once complete we can convert the SAM file to a BAM file:

```
samtools view -bS E_Coli_CGATGT_L001_filtered.sam >
E_Coli_CGATGT_L001_filtered.bam
```

And then we can sort the BAM file:

```
samtools sort E_Coli_CGATGT_L001_filtered.bam E_Coli_CGATGT_L001_filtered.sorted
```

Once completed, we can index the BAM file:

```
samtools index E_Coli_CGATGT_L001_filtered.sorted.bam
```

We can then (at last!) obtain some basic summary statistics using the samtools flagstat command:

```
samtools flagstat E_Coli_CGATGT_L001_filtered.sorted.bam
```

```
1269457 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
1264905 + 0 mapped (99.64%:-nan%)
1269457 + 0 paired in sequencing
634808 + 0 read1
634649 + 0 read2
1252624 + 0 properly paired (98.67%:-nan%)
1263794 + 0 with itself and mate mapped
1111 + 0 singletons (0.09%:-nan%)
8382 + 0 with mate mapped to a different chr
7369 + 0 with mate mapped to a different chr (mapQ>=5)
```

We can see here that very few of the reads do not map back to the contigs. Importantly 98% of reads are properly paired which gives us some indication that there are not too many mis-assemblies.

Run qualimap to get some more detailed information (and some images)

```
qualimap bamqc -outdir bamqc -bam E_Coli_CGATGT_L001_filtered.sorted.bam
```

```
firefox bamqc/qualimapReport.html
```

In the Chromosome stats section:

Chromosome stats

Name	Length	Mapped bases	Mean coverage	Standard deviation
NODE_1_length_302987_cov_26.9107_ID_1	302987	16219315	53.53	12.6
NODE_2_length_293215_cov_26.7425_ID_3	293215	15963219	54.44	11.38
NODE_3_length_235405_cov_26.009_ID_5	235405	12492827	53.07	10.9
NODE_4_length_229124_cov_26.8518_ID_7	229124	11965221	52.22	10.47

The larger of our contigs have a mean coverage of around 50 - which is what we would expect from our original alignment.

NODE_25_length_67492_cov_567.168_ID_49	67492	78077990	1,156.85	224.3
--	-------	----------	----------	-------

There is one contig which has the size of 67492 - this is exactly the same as the contig we found in the unmapped reads - that is pretty good indication that it is a separate sequence (remember we suspected a plasmid) and not integrated into the chromosome.

Let's double check that by blasting these contigs against the unmapped assembly contigs from part 3:

```
blastn -subject contigs.goodcov.fasta -query  
../unmapped_assembly/spades_assembly/contigs.fasta > check_plasmid.blastn
```

Open the file in a text editor:

```
pluma check_plasmid.blastn
```

and about 30% of the way down the file you should find: (hint use edit/find)

```
Query= NODE_1_length_67492_cov_565.407_ID_1  
  
Length=67492  
  
Subject= NODE_25_length_67492_cov_567.168_ID_49  
  
Length=67492  
  
Score = 1.246e+05 bits (67492), Expect = 0.0  
Identities = 67492/67492 (100%), Gaps = 0/67492 (0%)  
Strand=Plus/Plus
```

This shows us that this contig exactly matches that in the unmapped assembly, strongly supporting that this is a plasmid sequence and not integrated into the chromosomes.

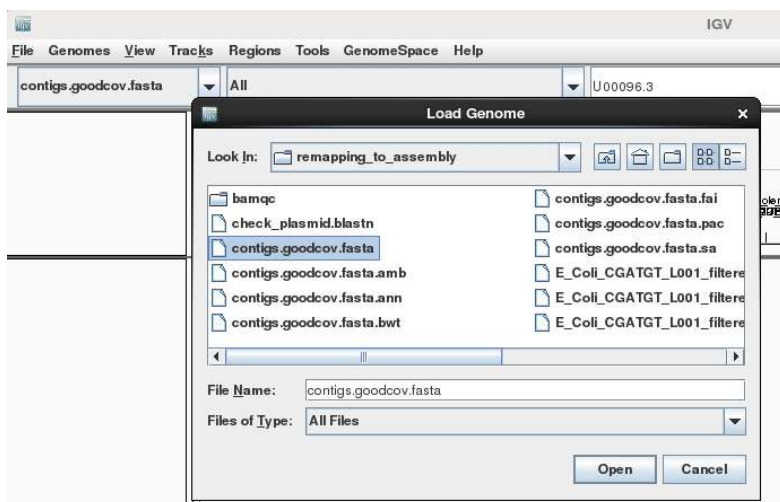
Task 4: View assembly in IGV

Load up IGV
`igv.sh`

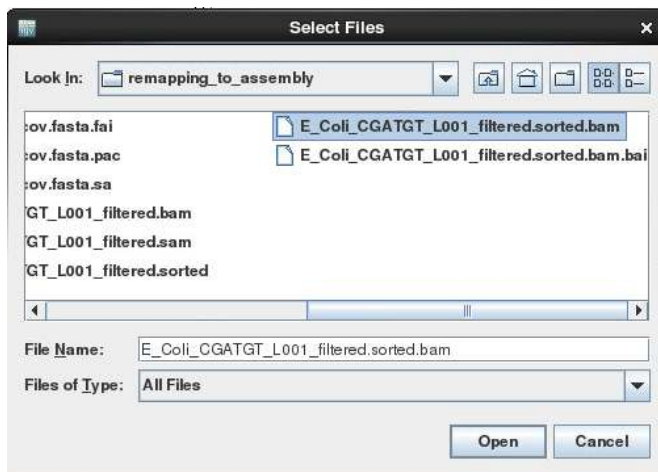
Click Genomes -> Load Genome from File....

Give the genome a sensible unique identifier and descriptive name.

We are going to import the contigs we have assembled as the reference. Unlike the reference genome though, we have no annotation available. Make sure you select the `contigs.goodcov.fasta` file for the complete de novo assembly (not the unmapped reads assembly).



Once loaded, click on File->Load From File... select the `E_Coli_CGATGT_L001_filtered.sorted.bam` file. Again, make sure you load the file in the **remapping_to_assembly** directory.

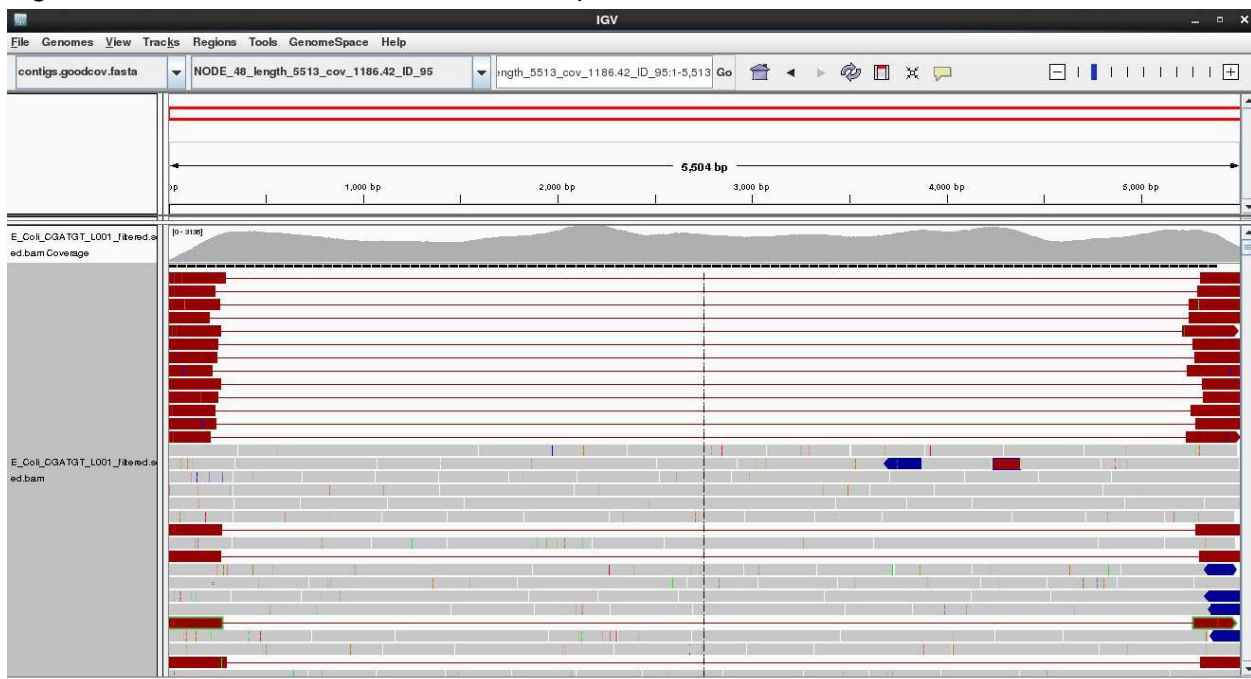


Once loaded, explore some of the contigs in IGV.
See if you can find anything unusual in any of the contigs.

Here is one to get you started.

Select NODE_49.....

Right click on the reads and select view as pairs



What do you think is going on here??

Annotation of de-novo assembled contigs

We will now annotate the contigs using Prokka. This is already installed on the instance, but if you would like to learn more you can find details at

<http://www.vicbioinformatics.com/software/prokka.shtml>. Prokka is actually a 'wrapper' for many

other tools which we have already seen in this lab (e.g. Blast). Prokka will also perform gene prediction (as well as calling open reading frames) and do some additional annotation using SignalP to predict signal peptides and Infernal to predict ncRNA and Barnap to predict conserved rRNA and tRNA. It also searches the PFAM database to identify proteins based on similarity of sequence, structure or hidden-markov profile. It also outputs files required for submission to NCBI Genbank which is very useful when publishing!

Remember that genome annotation relies primarily on homology and the identification of functional motifs. No genome annotation system will be able to produce an accurate functional prediction for a protein which has not been previously identified. Any annotation errors in existing databases will also adversely impact annotation. You have been warned!

N.B. You could also run Prokka on the contigs produced from unmapped reads

Unfortunately Prokka is only useful for bacteria and archaea. Eukaryote genome annotation is a much harder proposition and no comparable packages exist to serve the eukaryote community. The best package for such organisms is Maker (<http://www.yandell-lab.org/software/maker.html>).

Task 5: Annotate the de-novo assembled contigs using Prokka

Prokka is straightforward to execute. A single command will start gene prediction, Blast searches, SignalP and a whole host of other useful annotation tools. This saves us an awful lot of work compared to our previous look at the unmapped reads.

Again, we will use codon table 11 which defines the bacterial codon usage table (<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>).

Make sure you are in the denovo_assembly directory containing the scaffolds.fa file and type (again all on one line):

```
prokka --outdir prokka_annotation --genus Escherichia --species coli --strain UTI89
--kingdom Bacteria --gcode 11 --gram neg --cpus 2 scaffolds.fa
```

This will take around 20 minutes to complete. Once it has finished you will see the files Prokka generated in the prokka_annotation directory:

```
ubuntu@ip-10-61-164-199:16:06: [~/workshop_data/genomics_tutorial/data/sequencing/
/ecoli_exeter/denovo_assembly]$ ls prokka_annotation/
PROKKA_01102015.err  PROKKA_01102015.fsa  PROKKA_01102015.sqn
PROKKA_01102015.faa  PROKKA_01102015.gbk  PROKKA_01102015.tbl
PROKKA_01102015.ffn  PROKKA_01102015.gff  PROKKA_01102015.txt
PROKKA_01102015.fna  PROKKA_01102015.log
```

In summary:

The .faa file contains the amino acid sequence of each protein found in the assembly

The .ffn file contains the nucleotide level sequence for each gene

The .fna file contains the full nucleotide sequence of the contigs

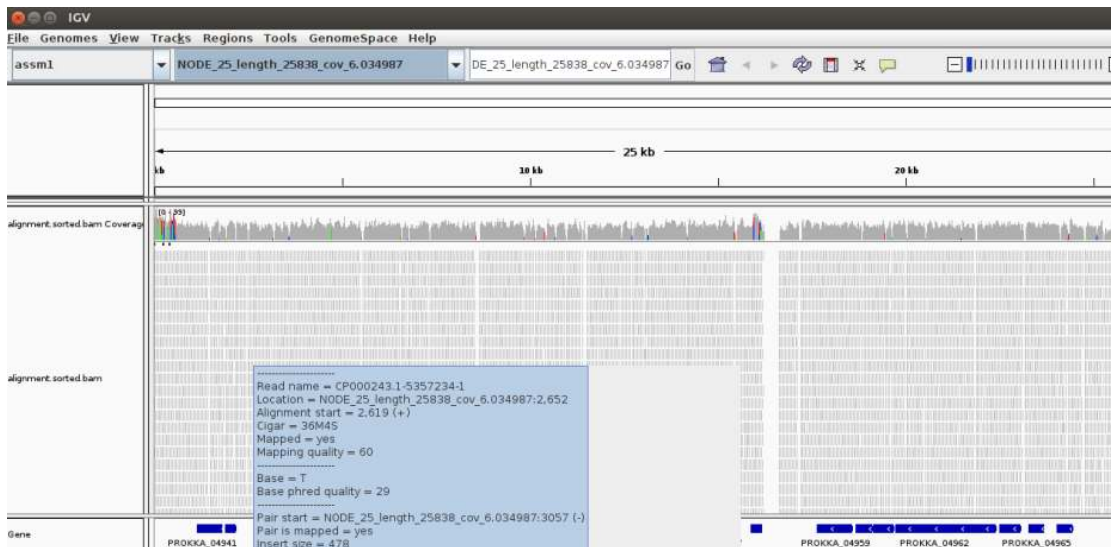
The .fsa file contains the full nucleotide sequence of the contigs with the full headers including species name and codon usage table

The annotation files in Genbank and GFF format

The .txt file with a summary of how many features were found

The sqn and tbl files are needed for submission to genbank

Task 6: Load the annotation into IGV (as per Part 2)



Task 7 (Optional): Run the contigs through the RAST server and import the resulting GFF annotation into IGV (refer back to Part 3 for instructions).

Visually compare the RAST annotation to the Prokka annotation.

Hybrid de-novo assembly

You will have seen that even with good coverage and a relatively long (300bp) paired end Illumina dataset - the assembly we get is still fairly fragmented. Our *E.coli* example assembles into 78 contigs and the largest contig is around 10% of the genome size.

Why is this?

One possible reason would be that regions of the original genome were not sequenced, or sequenced at too low coverage to assemble correctly. Regions of the genome will occur with different frequencies in the library that was sequenced - You can see this in the variation of coverage when you did the alignment. This can be due to inherent biases in the preparation and the random nature of the process.

However as coverage increases the chances of not sequencing a particular region of the genome reduces and the most significant factor becomes the resolution of repeats within the assembly process. If two regions contain the same or very similar sequences the assembler cannot reliably detect that they are actually two or more distinct sequences and incorrectly 'collapses' the repeat into a single sequence. The assembler is now effectively missing a sequence and therefore breaks in the assembly occur.

One resolution to this is to use a sequencing technology like PacBio or Sanger which can produce longer reads - the reads are then long enough to include the repeated sequence, plus some unique sequence, and the problem can be resolved. Unfortunately getting enough coverage using Sanger sequencing is expensive and PacBio - although relatively inexpensive has a high error rate.

An approach becoming more and more popular is to combine technologies. For example: high quality Illumina sequencing to get the accuracy of reads combined with low quality PacBio sequencing to enable the repeats to be spanned and correctly resolved.

Our exercise will be to use Illumina and PacBio datasets to assemble a species of pseudomonas. These are subsets of data used in "Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences" Utturkar et al., 2014. (<http://www.ncbi.nlm.nih.gov/pubmed/24930142>). This paper also contains a good explanation of the process and different approaches that are available.

You can also refer to this paper "One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly" for more information surrounding this topic:

<http://www.sciencedirect.com/science/article/pii/S1369527414001817>

Task 9: QC the data

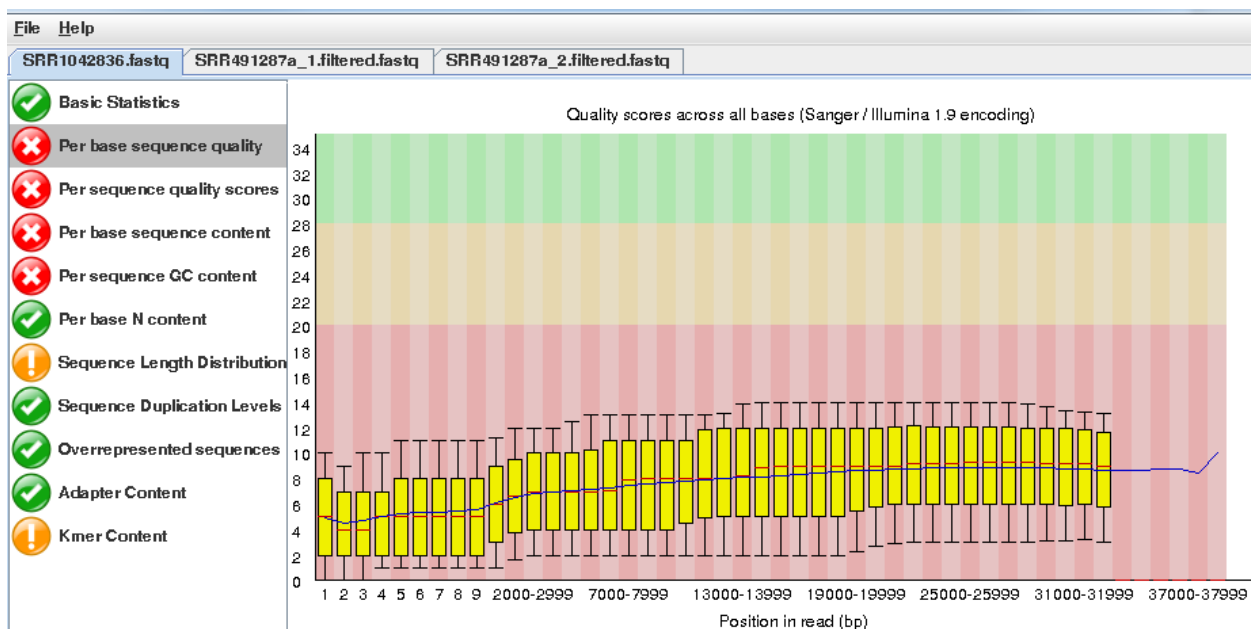
It is always important to check and understand the quality of the data you are working with:

Change to the directory and run fastqc

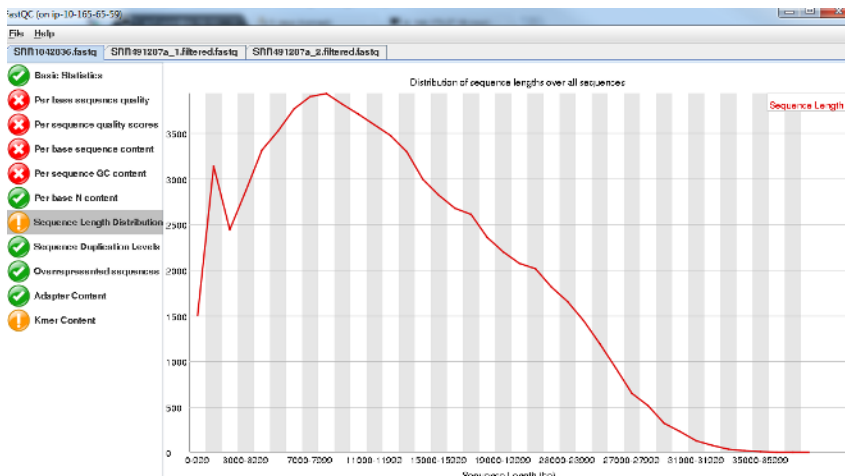
```
cd ~/workshop_data/genomics_tutorial/data/sequencing/pseudomonas_gm41
```

```
fastqc
```

Open the files SRR1042836a.fastq SRR491287a_1.fastq -2 SRR491287a_2.fastq and look at the reports generated.



Note that the quality of the PacBio reads (SRR1042836a.fastq) is much lower than the Illumina reads with a greater than 1 chance in 10 of there being a mistake for most reads.



However, importantly, the length of the PacBio reads is much longer.

Trim the Illumina reads as before (again, this will take around 20 minutes so take a break):

```
fastq-mcf ../../reference/adaptors/adaptors.fasta SRR491287a_1.fastq SRR491287a_2.fastq  
-o SRR491287a_1.filtered.fastq -o SRR491287a_2.filtered.fastq -q 20 -p 10 -u -x 0.01
```

You can check the number of filtered reads using `grep -c` and the quality of trimmed reads with `fastqc` if you want.

For this exercise we want the long reads from PacBio even though they are low quality. We are relying on the assembler to use them appropriately.

Task 10: Illumina Only Assembly

Firstly let's construct an assembly using only the available Illumina data.

Make sure you are in the directory

```
~/workshop_data/genomics_tutorial/data/sequencing/pseudomonas_gm41
```

You would run the following command, however as it takes over an hour the results have been pre-computed and are available in `illumina_only_assembly/`

```
spades.py --threads 2 --careful -o illumina_only_assembly -1 SRR491287a_1.filtered.fastq  
-2 SRR491287a_2.filtered.fastq
```

Change to the directory:

```
cd illumina_only_assembly
```

Filter out low coverage and very short contigs using a perl script:

```
filter_low_coverage_contigs.pl < contigs.fasta > contigs.goodcov.fasta
```

Let's look at the metrics for the assembly.

```
quast.py --output-dir quast contigs.goodcov.fasta
```

```
cat quast/report.txt
```

Assembly	contigs.goodcov
# contigs (>= 0 bp)	162
# contigs (>= 1000 bp)	147
Total length (>= 0 bp)	6631076
Total length (>= 1000 bp)	6620316
# contigs	162
Largest contig	224897
Total length	6631076
GC (%)	59.00
N50	78980
N75	48621
L50	28
L75	54
# N's per 100 kbp	0.00

Task 11: Create Hybrid Assembly

Now will execute the same command, but this time include the longer PacBio reads to see the effect it has on our assembly.

Change back into the directory

```
~/workshop_data/genomics_tutorial/data/sequencing/pseudomonas_gm41
```

You would run the following command, however as it takes over an hour and the results have been pre-computed and are available in hybrid_assembly/

```
spades.py --threads 2 --careful -o hybrid_assembly --pacbio SRR1042836a.fastq -1  
SRR491287a_1.filtered.fastq -2 SRR491287a_2.filtered.fastq
```

Change to the directory:

```
cd hybrid_assembly
```

Filter out low coverage and very short contigs using a perl script:

```
filter_low_coverage_contigs.pl < contigs.fasta > contigs.goodcov.fasta
```

Let's look at the metrics for the assembly - this time we will compare it with the illumina only assembly (all on one line):

```
quast.py --output-dir quast contigs.goodcov.fasta  
../illumina_only_assembly/contigs.goodcov.fasta
```

```
cat quast/report.txt
```

Assembly	hybrid_assembly_contigs.good	illumina_only_assembly_contigs.good
# contigs (>= 0 bp)	90	162
# contigs (>= 1000 bp)	80	147
Total length (>= 0 bp)	6666636	6631076
Total length (>= 1000 bp)	6660077	6620316
# contigs	90	162
Largest contig	484701	224897
Total length	6666636	6631076
GC (%)	59.00	59.00
N50	122016	78980
N75	78853	48621
L50	16	28
L75	34	54
# N's per 100 kbp	0.00	0.00

You can also explore the interactive html report:

```
firefox quast/report.html
```

It seems that using the longer reads has improved the completeness of our assembly - reducing the number of contigs more or less in half.

Task 12: Align reads back to reference

Let's realign our original reads back to the assembly and see what we have - refer to previous notes if you are unsure of the steps.

Start in the hybrid assembly directory

```
~/workshop_data/genomics_tutorial/data/sequencing/pseudomonas_gm41/hybrid_assembly
```

```
mkdir remapping_to_assembly
```

```
cd remapping_to_assembly
```

```
cp ../contigs.goodcov.fasta .
```

```
bwa index contigs.goodcov.fasta
```

First remap the Illumina reads. Type all on one line:

```
bwa mem -t 2 contigs.goodcov.fasta ../../SRR491287a_1.filtered.fastq  
../../SRR491287a_2.filtered.fastq > gm41.illumina.sam
```

Process the output so that it is viewable in igv:

```
samtools view -bS gm41.illumina.sam > gm41.illumina.bam
```

```
samtools sort gm41.illumina.bam gm41.illumina.sorted
```

```
samtools index gm41.illumina.sorted.bam
```

```
samtools flagstat gm41.illumina.sorted.bam > illumina_mapping_stats.txt
```

4547338 + 0 in total (QC-passed reads + QC-failed reads)

0 + 0 duplicates

4532949 + 0 mapped (99.68%:-nan%)

4547338 + 0 paired in sequencing

2272510 + 0 read1

2274828 + 0 read2

4497413 + 0 properly paired (98.90%:-nan%)

4524747 + 0 with itself and mate mapped

8202 + 0 singletons (0.18%:-nan%)

21401 + 0 with mate mapped to a different chr

17653 + 0 with mate mapped to a different chr (mapQ>=5)

21367 + 0 in total (QC-passed reads + QC-failed reads)

We can also map the PacBio reads, but we need to tell bwa we are using PacBio reads

```
bwa mem -t 2 -x pacbio contigs.goodcov.fasta ../../SRR1042836a.fastq > gm41.pacbio.sam
```

```
samtools view -bS gm41.pacbio.sam > gm41.pacbio.bam
```

```
samtools sort gm41.pacbio.bam gm41.pacbio.sorted
```

```
samtools index gm41.pacbio.sorted.bam
```

```
samtools flagstat gm41.pacbio.sorted.bam > pacbio_mapping_stats.txt
```

17013 + 0 mapped (79.62%:-nan%)

0 + 0 paired in sequencing

0 + 0 read1

0 + 0 read2

0 + 0 properly paired (-nan%:-nan%)

0 + 0 with itself and mate mapped

0 + 0 singletons (-nan%:-nan%)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

You will notice that not such a high proportion of PacBio reads map back to the assembly.

Now start igv:

```
igv.sh
```

Load your assembled genome -

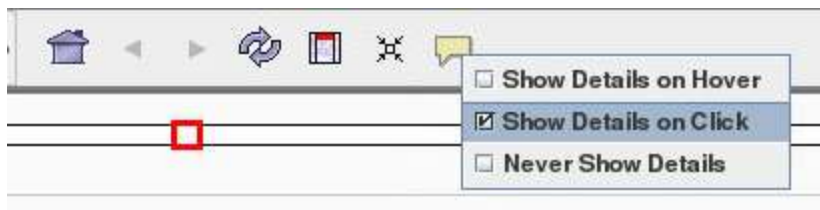
Click on "Genome - Load from genome from file"

Make sure you get the assembly from the hybrid_assembly (igv remembers the previous directory which may contain similar files.)

Now load your 2 alignment files:

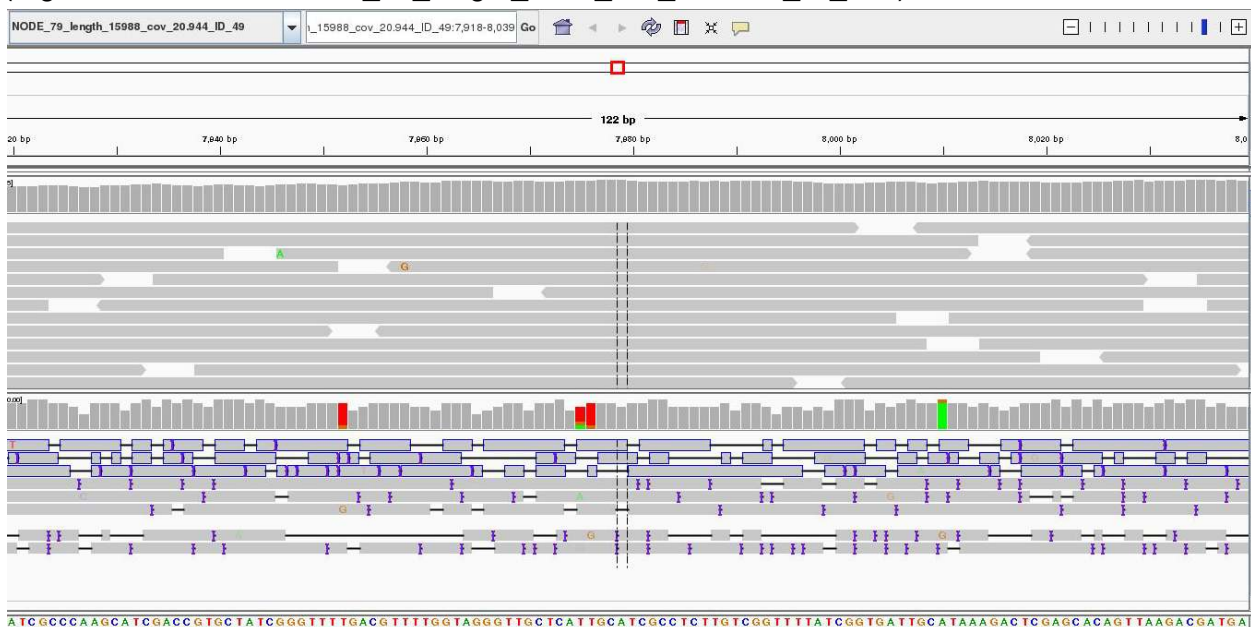
click on load from File and then select gm41.pacbio.sorted.bam and gm41.illumina.sorted.bam

On the toolbar select - "Show Details on Click"



Find a region that has decent coverage of both reads and zoom in.

(region shown here: NODE_79_length_2154_cov_37.1243_ID_157)



You can see that the PacBio reads are much longer, but the error rate particularly insertions and deletions is much higher than for the Illumina reads.

Explore a few other contigs to see if you can find something that looks like an error or mis-assembly. Remember the assembly process is difficult and far from perfect.

Task 13: Use Prokka to annotate the hybrid assembly (optional)

Follow similar instructions to Task 4.

Task 14: Rerun quast with the predicted genes from Prokka to evaluate the assemblies (optional)

If you wish, use quast with together with the Prokka annotation to evaluate the quality of the hybrid assembly compared to the Illumina-only assembly.

The quast manual can be found at <http://quast.bioinf.spbau.ru/manual.html>. We recommend you use options in quast to split the scaffolds.fasta into contigs (-s) and the -G option to include the predicted genes from Prokka. You will need to run quast twice - once for the Illumina-only assembly (with its prokka annotation) and once for the hybrid assembly (with its prokka annotation).

How do metrics such as number of mis-assemblies compare between the hybrid and Illumina-only assemblies?

Summary

You have seen that de-novo assembly of short reads is a challenging problem. Even for small genomes, the resulting assembly is fragmented into contigs and far from complete.

Incorporating longer reads to produce a hybrid assembly can be used to reduce the fragmentation of the genome. We have only used a single (perhaps the simplest) technique to incorporate long reads. You can read more about hybrid assembly techniques here:

<http://www.ncbi.nlm.nih.gov/pubmed/24930142>

2015 Workshop on Genomics

Part 5

Short read genomics:

Comparison of results between different strains

Instructors:

- Konrad Paszkiewicz k.h.paszkiewicz@exeter.ac.uk

Objectives:

By the end of this section you will be expected to be able to:

- Run parts 1-3 of the lab on up to new 6 datasets.
- Use pre-prepared scripts to compare SNPs and Indels between strains.
- Generate pseudo-sequences based on synonymous SNPs.
- Draw simple trees to illustrate the likely evolutionary relationship between strains.

Projects!

In the previous sections you have been taken through the steps required to:

1. QC and filter Illumina data
2. Remap Illumina short-read data to a reference sequence
3. View the results in IGV
4. Identify SNPs and Indels in an automated fashion using samtools and bcftools
5. Determine whether SNPs result in synonymous or non-synonymous changes in the corresponding amino acid
6. Extract unmapped reads
7. Assemble unmapped reads and obtain assembly statistics
8. Annotate unmapped reads RAST and/or BLAST
9. Assemble a bacterial genome de-novo using SPAdes
10. Obtain assembly statistics
11. Annotate as per the unmapped reads (where computationally feasible).

Now we want you to do the same on a set of *Vibrio parahaemolyticus* strains which have been isolated and sequenced. There are six strains available depending on how much time is available and enthusiasm you have - choose a number of strains (at least 2) as we want to run some comparisons.

The strains can be found in:

~/workshop_data/genomics_tutorial/data/sequencing/Vibrio_parahaemolyticus

```
[ec2-user@ip-10-181-126-118 Vibrio_parahaemolyticus]$ ls -l
total 24
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:35 Sample_G35
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:44 Sample_PSU3384
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:37 Sample_T02347066
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:33 Sample_T024_47060
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:41 Sample_T0347070
drwxrwxr-x. 3 ec2-user ec2-user 4096 Dec  9 11:39 Sample_T0847053
```

Under each Sample directory is a subdirectory called raw_illumina_reads which contains the fastq files.

For remapping, the reference can be found in the folder:

~/workshop_data/genomics_tutorial/data/reference/Vibrio_parahaemolyticus_RIMD_2210633_uid57969

(Remember, you will need to create an index first).

For each strain, make a list of:

1. SNPs, Indels and their effects (from the remapping)

2. Missing genes (from the remapping)
3. Novel plasmids and/or genes (when performing the assembly - don't specify the k-mers SPAdes will choose appropriate ones.)

Once completed, see if you can predict what the phenotype of these bacteria might be. Then proceed to the final part of the tutorial where we will compare the results from all of these strains.

N.B.

It is recommended that you follow the same directory naming convention we have followed here (i.e. one for remapping to the reference, another for assembly of unmapped reads and a final one for the denovo assembly).

These tasks may take you several days. However, remember that all of the basic procedures are detailed in the previous sections – only the input FASTQ files will have changed. Feel free to refer to these previous tasks to remind yourself of the commands and parameters. By all means feel free to play around with different parameters if you wish, although remember that the results may differ from those you see here.

Just to give you some guidance:

You should find that strain Sample_T0347070 yields many more SNPs than other strains.

You may find that some scripts and programs run more slowly because of these extra differences and larger datasets.

Also, if you find the de novo assembly process causes your NX session to end, the chances are that SPAdes has caused your instance to run out of memory. If this happens, increase the minimum k-mer size in the spades.py command line.

Comparing variants between several samples and a reference genome:

Here we will use a script to compare the variants called in each sample. Ensure you are in the ~/genomics_tutorial/data/sequencing/Vibrio_parahaemolyticus directory

First of all, let's make a directory to store the results of the comparison:

```
mkdir snp_comparison/
```

We need a copy of all of the vcf4 files we created here. This is a quick way to do it - paste this in as one command

```
for sample in Sample*
do
cp -v $sample/remapping_to_reference/out.snps.vcf4
snp_comparison/$sample.out.snps.vcf4
done

cd snp_comparison/
```

Note that the copy commands may require modification depending on where you have saved the variant call results.

Our directory contents should look something like:

```
[ec2-user@ip-10-181-126-118 snp_comparison]$ ls -l
total 64388
-rw-rw-r--. 1 ec2-user ec2-user 7923228 Dec  9 17:16 Sample_G35.out.snps.vcf4
-rw-rw-r--. 1 ec2-user ec2-user 7377661 Dec  9 17:16 Sample_PSU3384.out.snps.vcf4
-rw-rw-r--. 1 ec2-user ec2-user 7766075 Dec  9 17:16 Sample_T02347066.out.snps.vcf4
-rw-rw-r--. 1 ec2-user ec2-user 7828043 Dec  9 17:16 Sample_T024_47060.out.snps.vcf4
-rw-rw-r--. 1 ec2-user ec2-user 27354227 Dec  9 17:16 Sample_T0347070.out.snps.vcf4
-rw-rw-r--. 1 ec2-user ec2-user 7666231 Dec  9 17:16 Sample_T0847053.out.snps.vcf4
```

We'll now set up some variables so we don't have to type long path names

```
ref=~/workshop_data/genomics_tutorial/data/reference/Vibrio_parahaemolyticus_RIMD_221
0633_uid57969/Vibrio_parahaemolyticus_RIMD_2210633_uid57969.fasta

gff=~/workshop_data/genomics_tutorial/data/reference/Vibrio_parahaemolyticus_RIMD_221
0633_uid57969/Vibrio_parahaemolyticus_RIMD_2210633_uid57969.gff

samples=`ls *.vcf4`
```

We can now use \$ref instead of the long path to our reference and \$gff for the feature file
e.g.

```
head $ref
echo $samples
```

When we are happy our variables are correct then run:

```
snp_comparator.pl 10 $ref $gff $samples > snp_comparison.txt
```

Looking at the snp_comparison.txt file (either in a text editor, or in a spreadsheet):
If you have chosen different samples - you will get different results of course.

```
## Table of SNP and Indel occurrences between these samples. Note that any comma-separated values (e.g. A,C indicate potential heterozygosity
and/or sample heterogeneity
Chrom  Pos  Ref  Sample_G35.out.snps.vcf4  Sample_PSU3384.out.snps.vcf4  Gene description  Status
NC_004603  1000  G  A  G  VP0002 tRNA modification GTPase TrmE ,silent atc -> atT;
NC_004603  1000051  T  A  T  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,non-silent ctg -> cAg;
NC_004603  1000065  G  A  G  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,non-silent gtc -> Atc;
NC_004603  1000067  C  T  C  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,silent gtc -> gtT;
NC_004603  1000080  C  C  A  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,non-silent ctc -> Atc;
NC_004603  100009  C  C  T  VP0092 hypothetical protein ,silent tcg -> tcA;
NC_004603  1000091  C  T  T  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,silent gac -> gaT; ,silent gac ->
gaT;
NC_004603  1000100  C  T  C  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,silent tgc -> tgT;
NC_004603  1000103  A  G  A  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,silent cca -> ccG;
NC_004603  100020  A  G  G  VP0092 hypothetical protein ,silent ttg -> CtG; ,silent ttg -> CtG;
NC_004603  1000219  T  C  C  VP0959 zinc/cadmium/mercury/lead-transporting ATPase ,non-silent ttg -> tCg; ,non-silent t
```

Here we can see the chromosome ID, the position in bp, the reference base and the base at each position as well as the gene (if any) the variant occurs in as well as the effect (silent, non-silent or indel).

Obtaining a phylogeny based on synonymous SNPs only:

How are the strains related on the basis of these variants? We can ask a number of questions, but if we are looking at the long-term evolutionary history of the strains we should only look at synonymous (i.e. silent) mutations as these should not confer a significant selective advantage to any strain. Using the data snp_comparison.txt file, we can form 'pseudo-sequences' using the script snp2tree_fullsequence.pl. These are concatenated bases consisting of only those positions which are silent across all strains. It is essentially the same as turning each column of each strain in the snp_comparison.txt file into a FASTA entry.

```
snp2tree_fullsequence.pl snp_comparison.txt > synonymous_tree.fasta
```

Examine the contents of the tree.fasta file. We can then treat this file as an alignment (since each base in each sequence is at the same position on the chromosome) and pass it to a phylogeny program called FastTree. FastTree will take an input alignment and output a Newick formatted tree (http://en.wikipedia.org/wiki/Newick_format).

```
FastTree -nt -gtr < synonymous_tree.fasta > synonymous_tree.newick
```

Now we can visually view this tree by using an online tool.

firefox <http://www.trex.uqam.ca/index.php?action=newick>

Either paste the contents of the .newick file into the window or select 'Sequences file' and load the file through the browser. Then select 'View Tree'.



Advanced task (optional):

Copy the `snp2tree_fullsequence.pl` script to this directory
(`~/workshop_data/genomics_tutorial/data/sequencing/Vibrio_parahaemolyticus/snp_comparison`)
and modify it so that it selects positions containing only non-silent mutations (not indels as these modify the alignment). Generate a new alignment and compare the resulting tree against the silent mutations.

Concluding remarks:

Well done! If you have reached this far, you deserve a round of applause. You have completed some of the most common tasks in short-read sequencing. You can use the same machine and the same scripts to perform analysis of any short-read dataset! All you need to do is transfer the FASTQ files to the server - if you have them on your personal desktop you can use WinSCP (windows) , Fugu or Cyberduck (Mac OSX) or any other SFTP program.

**A tutorial for WinSCP can be found at
http://www.siteground.com/tutorials/ssh/ssh_winscp.htm**