EXPERIMENT: 1

WORKING WITH NGS DATABASES AND NGS FILE FORMATS

AIM: To retrieve files of the from SRA database, obtain .fastq files using SRA toolkit and interpret results

PROCEDURE: Downloading data from SRA database: [DNA]

- Go to the website https://www.ncbi.nlm.nih.gov/sra
- Type any organism name or known accession number or (SRR)
- Collect library information

Name	Penicillium chrysogenum
Instrument	Illumina NovaSeq 6000
Strategy	WGS
Source	GENOMIC
Selection	RANDOM
Layout	PAIRED

- Click on the SRR number under Run. It takes us to the next page

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR12825363	1.2M	363.7M	110.8MB	48.4%	2020-10-18	public

- Click on reads tab take print screen
- Download the data.
- fastq-dump: Convert SRA data into fastq format
- The below command should be used in the SRA toolkit folder

fastq-dump --split-files SRRXXXXXXX.sra

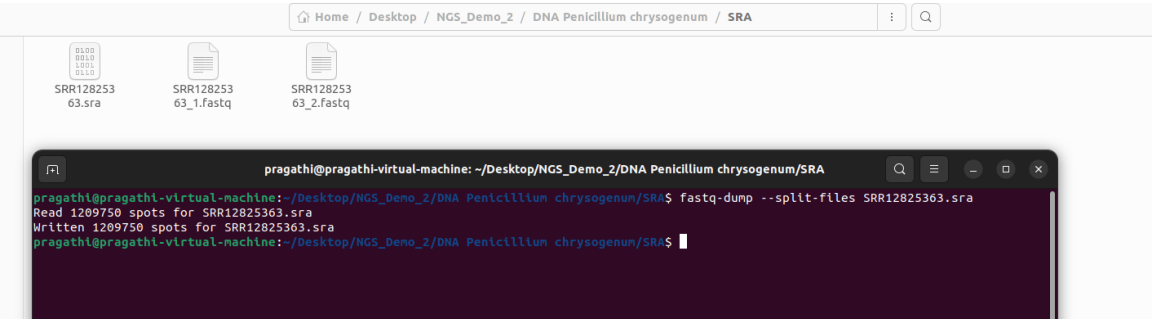


Fig: Conversion from .sra to .fastq files

Quality Control analysis for raw data

Quality control or QC of short refers to the quality of the data before starting the experiment. The objective of the quality control is to find the Basic statistics, per base sequence quality, per sequence quality score, per base sequence content, per sequence GC content, per sequence GC content, sequence length distribution, sequence duplication level, overrepresented sequence.

FastQC

Modern high throughput sequencers can generate tens of millions of sequences in a single run. Before analysing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it. FastQC can be run in one of two modes. It can either run as a standalone interactive application for the immediate analysis of small numbers of FastQ files, or it can be run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files.

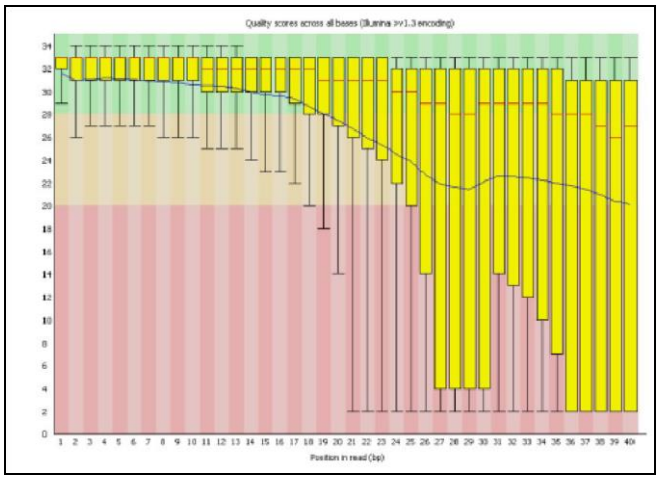
Basic Statistics

The Basic Statistics module generates some simple composition statistics for the file analyzed.

- **Filename:** The original filename of the file which was analyzed
- **File type:** Says whether the file appeared to contain actual base calls or colorspace data which had to be converted to base calls
- **Encoding:** Says which ASCII encoding of quality values was found in this file.

- **Total Sequences:** A count of the total number of sequences processed.
- **Filtered Sequences:** If running in Casava mode sequences flagged to be filtered will be removed from all analyses.
- **Sequence Length:** Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
- **%GC:** The overall %GC of all bases in all sequences.

Per Base Sequence Quality



- This view shows an overview of the range of quality values across all bases at each position in the FastQ file for each position a BoxWhisker type plot is drawn.
- The elements of the plot are as follows: The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%) ● The upper and lower whiskers represent the 10% and 90% points ● The blue line represents the mean quality.
- The y-axis on the graph shows the quality scores. The higher the score the better the base call.
- The background of the graph divides the y axis into ○ Very good quality (green), Reasonable quality (orange), Poor quality (red).

Per Sequence Quality Scores

The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will

have universally poor quality, often because they are poorly images (on the edge of the field of view etc)

Per Base Sequence Content

Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other.

Per Base N Content

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call

THE ONLINE SERVERS AND TOOLS USED IN THIS EXERCISE ARE:

FASTQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

PROCEDURE

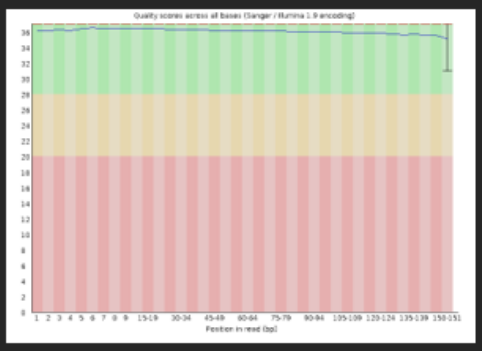
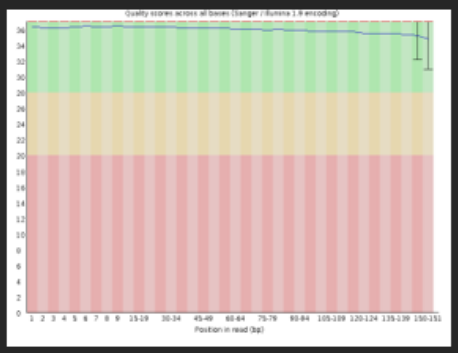
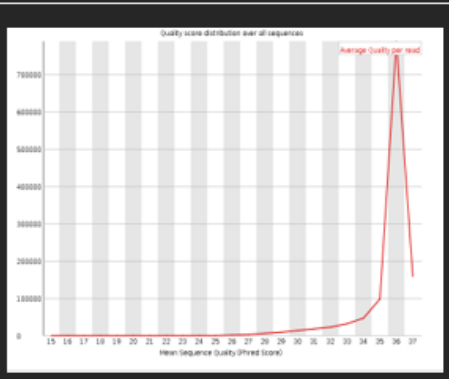
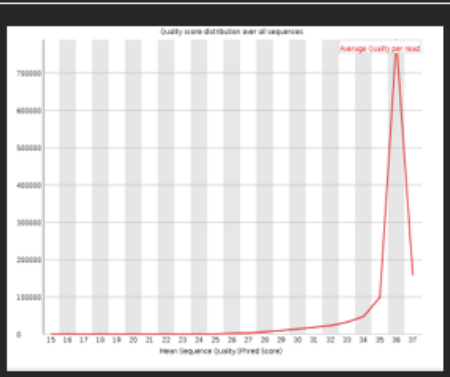
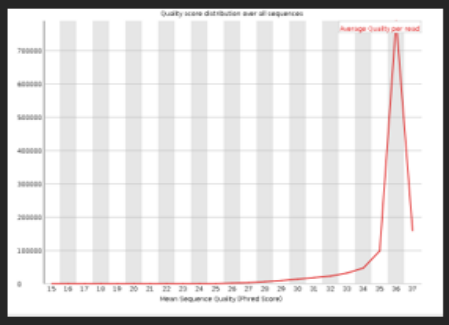
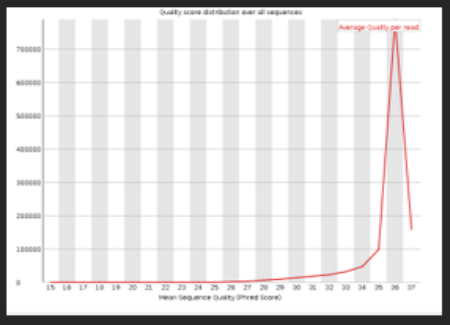
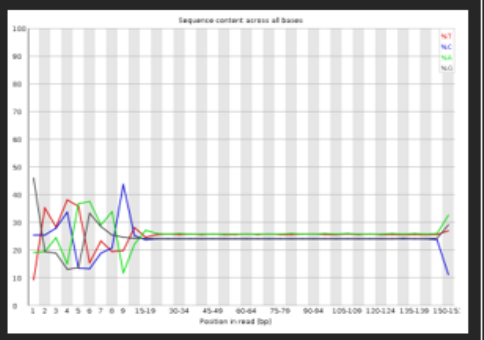
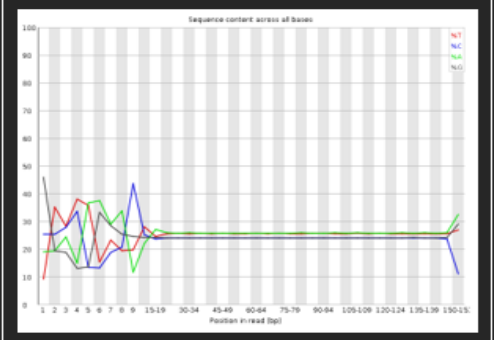
Collect basic statistics for selected data [DAN Set]

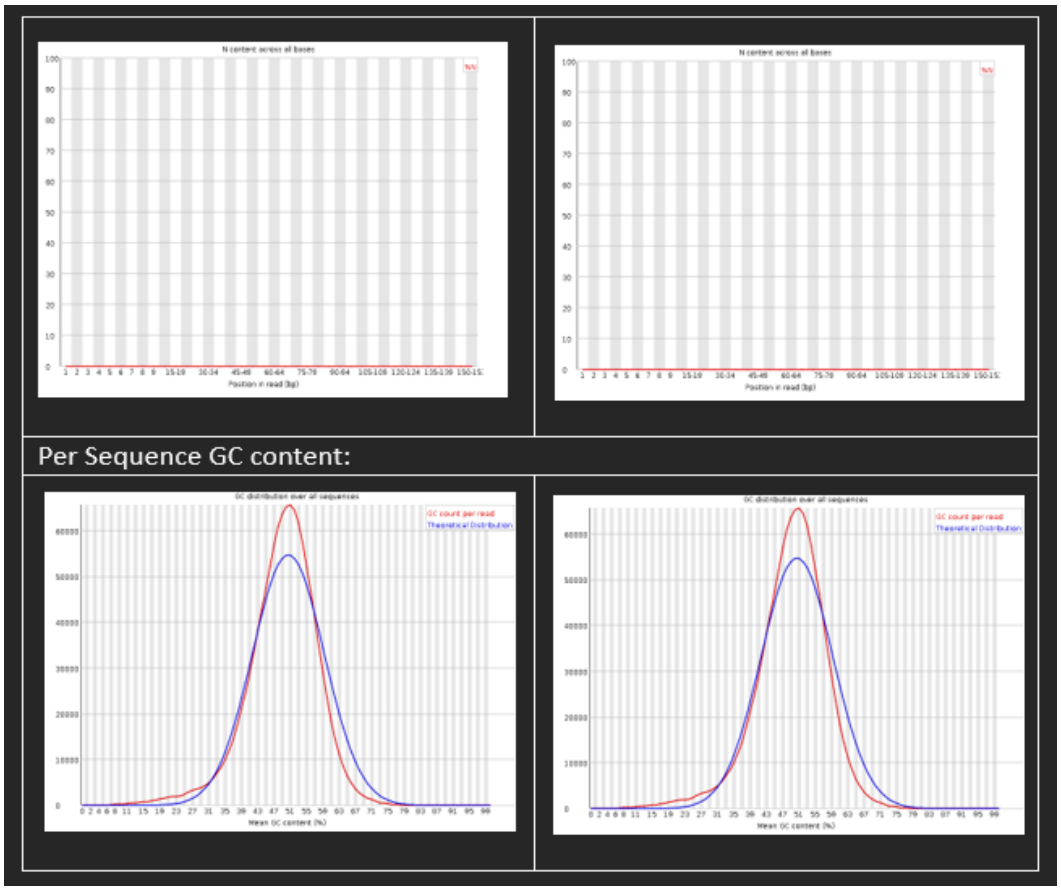
As the name implies Basic statistics is the primary characteristics of the sequence and that includes file name, file type, encoding total sequences, sequence flagged as poor quality, sequence length and %of GC content.

Measure	Read 1	Read 2
Filename	SRR12825363_1_fastq	SRR12825363_2_fastq
File type	Conventional base calls	Conventional base calls
Encoding	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Total Sequences	1209750	1209750
Sequences flagged as poor quality	0	0
Sequence length	35-151	35-151

%GC	48	48

Save the report for read1 and read2 as *.html

SRR12825363_1_fastqc.html	SRR12825363_2_fastqc.html
Per Base Sequence Quality:	
	
Per Sequence Quality Score:	
	
Per Base Sequence Content:	
	
Per Base Sequence Content:	
	
Per Base N Content:	



CONCLUSION:

In read 1, all bases are in the green region, which signifies a good phred score and call quality. The average phred score is around 37, which indicates 99.98% base call accuracy with minimal errors, demonstrating excellent quality. N content is 0 throughout. Therefore, all these factors indicate trimming is not required

In read 2, all bases are in the green region, which signifies a good phred score and call quality. The average phred score is around 37, which indicates 99.98% base call accuracy with minimal errors, demonstrating excellent quality. N content is 0 throughout. Therefore, all these factors indicate trimming is not required

EXPERIMENT: 2

REFERENCES GENOME ALIGNMENT OF WHOLE GENOME (WGS) DATA SETS

AIM:

To perform references genome alignment for given DNA data

LITERATURE:

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

THE ONLINE SERVERS AND TOOLS USED IN THIS EXERCISE ARE:

BWA	https://sourceforge.net/projects/bio-bwa/files/
SAM TOOLS	git clone https://github.com/samtools/samtools

PROCEDURE

BWA:

- Raw data : WGS data for (**Saccharomyces cerevisiae) eukaryotic** ● Downloading References genome (**Saccharomyces cerevisiae)**
- Go to ncbi genome database and type “Saccharomyces cerevisiae”
- Click on Download sequence and annotation from **RefSeq**
- Can download from FTP
- references genome is available in GPU1080 system
- How many chromosome is available: **unknown**
- Save that references genome file in BWA folder

Running BWA

- Indexing

time bwa index ref.fna
Example: time bwa index /home/pc1/Ex8/bwa-0.7.17/Saccharomyces_cerevisiae.fna

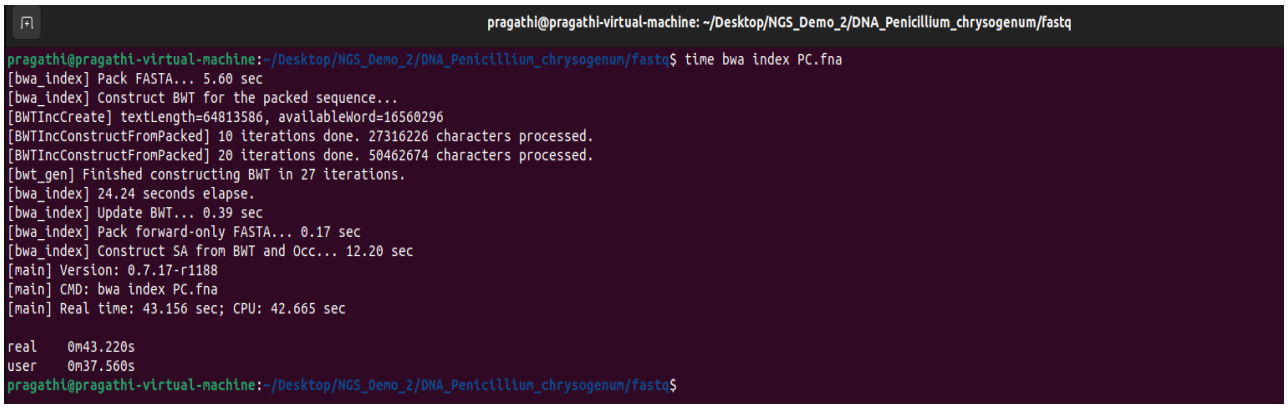


Fig: Indexing before running BWA MEM

- Once indexing is done, it will generate 5 files - .amb, .ann, .bwt, .pac, .sa

Running BWA MEM

Illumina/454/IonTorrent paired-end reads longer than ~70bp:

Use below code:

bwa mem ref.fna read1.fq read2.fq > aln-pe.sam
Example: time bwa mem -t 4 Saccharomyces_cerevisiae.fna SRR800826_1.fastq SRR800826_2.fastq > Rsult_MEM_ex8.sam

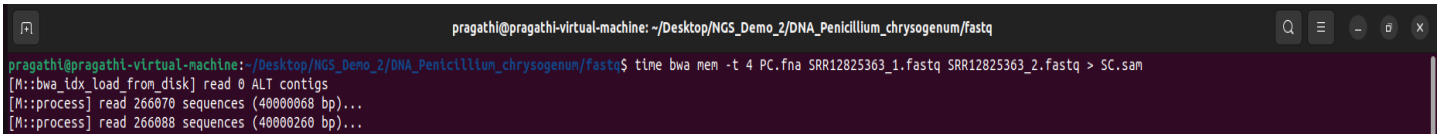


Fig: Running BWA MEM

- What is MEM? **It stands for "Maximal Exact Match." It is an algorithm used in bioinformatics for mapping sequence reads to a reference genome.**

- What is ref.fna? **It is the reference genome file in FASTA format. It contains the DNA sequence of the organism or genomic region against which the input reads will be aligned.**
- What is -t? **This option specifies the number of threads or CPU cores to be used for the alignment process.**
- What is sam file? **SAM stands for "Sequence Alignment/Map." It is a standard file format for storing biological sequence alignments. The output of the alignment process is written to a SAM file, which includes information about how each read aligns to the reference genome.**
- How much time it took? **43.220 seconds**
- What is the size of output file? **1.0 GB**

Running samtools:

- Go to BWA folder and run below command

samtools view -S -b sample.sam > sample.bam
Example: time samtools view -S -b Result_MEM_ex8.sam > Result_MEM_ex8.bam

- What is -S? **sam file**
- What is -b? **bam file**
- What is BAM? **Binary Alignment Mapping**

Sorting using samtools **samtools sort sample.bam -o sample.sorted.bam**
Example: **time samtools sort Result_MEM_ex8.bam -o Result_MEM_ex8_sort.bam**

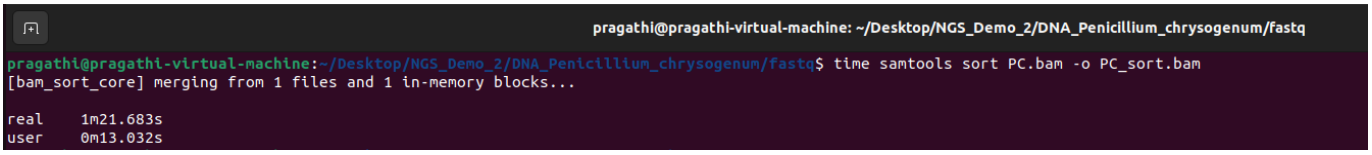


Fig: Sorting using SAM tools

samtools “indexing”

samtools index sample.sorted.bam

Example: time samtools time Result_MEM_ex8_sort.bam

```
pragathi@pragathi-virtual-machine:~/Desktop/NGS_Demo_2/DNA_Penicillium_chrysogenum/fastq$ time samtools index PC_sort.bam

real    0m1.570s
user    0m1.388s
sys     0m0.179s
```

Fig: Indexing using SAM tools

Fig: Visualisation of reference genome alignment using IGV