**Open terminal and create a directory to save your files.**

*Run in terminal:*

> mkdir *exercise*

> cd *exercise*

*(this command only works in a linux terminal)*

*> echo $PWD*

**Get the Docker images with the necessary tools**

*Run in terminal:*

> docker pull ncbi/sra-tools:2.11.1

> docker pull biocontainers/fastqc:v0.11.9\_cv8

> docker pull biocontainers/bwa:v0.7.17-3-deb\_cv1

> docker pull biocontainers/samtools:v1.9-4-deb\_cv1

> docker pull maxulysse/freebayes:1.3

> docker pull ummidock/breseq:0.32.1

**Download sample files and convert from SRA format to fastq.**

We will download and process samples with the SRA accession numbers SRR14773779 and SRR17960576.

*Run in terminal:*

> docker run -v $PWD:/exercise/ --rm -i -t ncbi/sra-tools:2.11.1

(you may need to adjust folder names for your case)

(you should now be inside the docker image)

The docker image may not start in your folder (unless you specify it when calling docker, so, you should move to your folder inside the docker machine)

> cd /exercise/

(you should now be inside the shared folder)

> fasterq-dump SRR17960576

* A fastq file should have been produced

> fasterq-dump --split-files SRR14773779

* How many files were produced? Why?

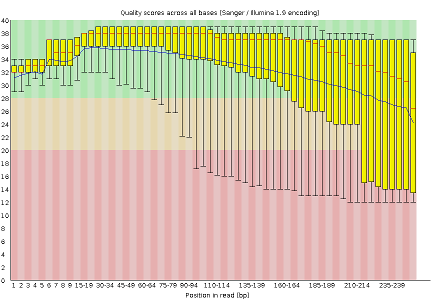
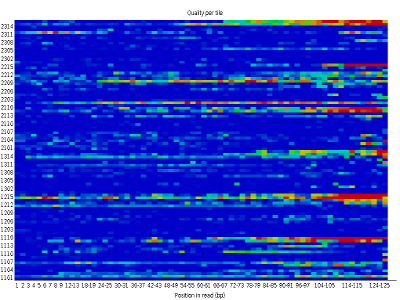
> exit

(you should now be outside the sratools docker image)

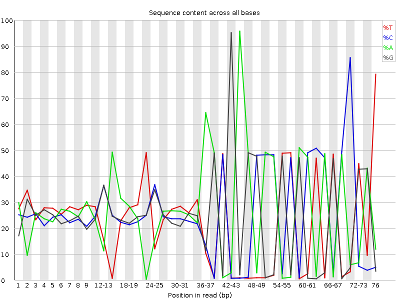
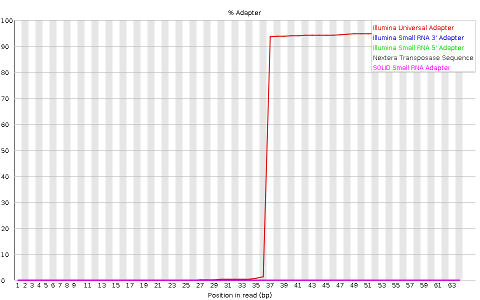
**Assess read quality.**

High Throughput Sequencing machines read thousands or millions of sequences in parallel. As you can imagine, this usually generates large fastq files, with millions of lines. Manually inspecting the quality of each read is out of the question. Specialized software has been developed to provide quality measures for fastq files generated by HTS machines. [FastQC](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a popular program to generate quality reports on fastq data. In fact, this is usually the first thing you should do once you receive a new dataset. FastQC reports provide a series of plots that allow the user to assess the overall quality of their raw data and detect potential biases and problems.

Some plots indicate distribution of base qualities along the length of reads. At least for illumina data, on average the quality of each base tends to decrease along the length of the read.

[](https://github.com/dsobral/ELB/blob/master/images/base_quality.png) [](https://github.com/dsobral/ELB/blob/master/images/tile_quality.png)

Other plots indicate biases in nucleotidic content of reads, either globally (such as %GC plots), or positionally. Global bias in nucleotidic content can be useful to search for signs of contaminants. On the other hand, positional bias are useful to detect presence of artefactual sequences in your reads such as adaptors. Another insight you may obtain from this information are potential biases in the preparation of your library. For example, random hexamer priming is not truly random, and preferentially selects certain sequences. The currently popular transposase-based enzymatic protocol, although reasonably random, is also not completely random, and you can see this through positional bias, particularly in the beginning of reads. The presence of adaptors is a relatively common event, and therefore specific plots exist to detect the presence of the most commonly used adaptors. Finally, the presence of repetitive sequences can also suggest contaminants, pcr artifacts, or other types of bias.

[](https://github.com/dsobral/ELB/blob/master/images/base_bias.png) [](https://github.com/dsobral/ELB/blob/master/images/adaptor.png)

Check the quality of the sequences we obtained using FastQC.

*Run in terminal:*

> docker run -v $PWD:/exercise/ --rm -i -t biocontainers/fastqc:v0.11.9\_cv8

> cd /exercise/

> fastqc SRR17960576.fastq

> fastqc SRR14773779\_1.fastq

> fastqc SRR14773779\_2.fastq

> exit

Open the HTML reports that should have been generated

* + What are major differences between SRR17960576 and SRR14773779
  + Can you find differences between the paired fastqs of SRR14773779

*Optional*: look at other fastq examples here

<https://github.com/dsobral/ELB/tree/master/fastq_examples>

**Align to a genome.**

You will need the SARS-COV2 genome, which you can get at NCBI (reference NC\_045512 – latest version is 2). Download as fasta file and put it in the same folder as the fastq files.

<https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

*Run in terminal:*

> docker run -v $PWD:/exercise/ --rm -i -t biocontainers/bwa:v0.7.17-3-deb\_cv1

> cd /exercise/

> bwa index NC\_045512.2.fasta

* + What was the output of running the command?

> bwa mem NC\_045512.2.fasta SRR17960576.fastq -o SRR17960576.\_NC\_045512.2.sam

> bwa mem NC\_045512.2.fasta SRR14773779\_1.fastq SRR14773779\_2.fastq -o SRR14773779\_NC\_045512.2.sam

> exit

The [Sequence Alignment/Map (SAM) format](https://samtools.github.io/hts-specs/SAMv1.pdf) is a tabular text file format, where each line (after header section) contains information for one alignment.

Open the file SRR14773779\_NC\_045512.2.sam. It can be in a text editor, but it will be easier to interpret in a spreadsheet software (excel-like). You can also use the linux command line (eg. cat | head; nano; etc…).

* + *How many alignments are there? What is the position of the first alignment (position is in the 4th column). In what order do they appear?*

(optional): if you want to further try to interpret the content of the file, read the SAM format specification: <https://samtools.github.io/hts-specs/SAMv1.pdf>

**Visualize in the Integrative Genome Browser.**

*You will need to get IGV for your system*

<https://software.broadinstitute.org/software/igv/download>

IGV, like many other programs, require BAM (Binary SAM) files, which are compressed versions of SAM files. These BAM files can then be indexed (do not confuse this indexing with the indexing of the reference genome) to allow direct access to alignments in any arbitrary region of the genome. Several tools such as IGV only work with index BAM files. To transform a SAM into an indexed BAM, you need to process the SAM with samtools. Namely, you need to sort the SAM by position, compress it into a BAM and then index the BAM.

Run in terminal:

> docker run -v $PWD:/exercise/ --rm -i -t biocontainers/samtools:v1.9-4-deb\_cv1

> cd /exercise/

> samtools sort -o SRR14773779\_NC\_045512.2.sorted.sam SRR14773779\_NC\_045512.2.sam

* *What is the position of the first alignment now? How many alignments do you have? In what order do the alignments appear now?*

> samtools view -Sb -o SRR14773779\_NC\_045512.2.sorted.bam SRR14773779\_NC\_045512.2.sorted.sam

> samtools index SRR14773779\_NC\_045512.2.sorted.bam

You now have the indexed BAM file necessary to visualize in IGV

> exit

Next, open IGV

Load the Sars-Cov2 genome in IGV. You can load the fasta file, but to also have access to the gene annotations you should download a genbank file.

<https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

Graphical user interface, text, application

Description automatically generated

Open the indexed BAM file you generated previously.

Graphical user interface, text, application

Description automatically generated

Apply the same procedure to visualize the data from SRR17960576

* *Can you see differences in the alignments?*

**Perform variant calling using freebayes with the reads aligned previously.**

Freebayes is a bayesian-based algorithm to estimate SNVs and small indels

<https://github.com/freebayes/freebayes>

Run in terminal:

> docker run -v $PWD:/exercise/ --rm -i -t maxulysse/freebayes:1.3

> cd /exercise/

> freebayes -0 -f NC\_045512.2.fasta SRR14773779\_NC\_045512.2.sorted.bam > SRR14773779\_NC\_045512.2.sorted.bam.freebayes.vcf

(-0 is a parameter that indicates to use a set of recommended parameter values)

* *What is the result of this command?*

> exit

Open the file SRR14773779\_NC\_045512.2.sorted.bam.freebayes.vcf. It can be in a text editor, but it will be easier to interpret in a spreadsheet software (excel-like). You can also use the linux command line (eg. cat | head; nano; etc…).

* *How many variants are there? What is the position of the first variant? (position is in the 2nd column)? What was the genomic change of that variant (reference allele and alternate alleles are in 4th and 5th columns)? Can you know the potential effect of this variant?*

(optional): if you want to further try to interpret the content of the file, read the VCF format specification: <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

Visualize some variants using IGV (using the alignments you obtained before).

* *Compare the alignments in a position of a variant with bad quality (position 1440) with the alignments in the position of a “good” variant (position 3037). Even in the good variant, can you see in the alignments a possible problem?*

**Exercise: Perform variant calling and variant annotation using breseq.**

Breseq is a software package for finding mutations (both small SNVs and large structural variants) in microbial-size genomes (cannot cope with larger genomes), using high throughput sequencing.

<https://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing>

This software also performs variant annotation, and for this we need to provide not only the sequence, but also the annotation (genes).

Obtain the genbank file NC\_045512.2.gb (format Genbank (Full))

Open the link <https://www.ncbi.nlm.nih.gov/nuccore/1798174254> (or open the file NC\_045512.2.gb in a text editor)

* *How many genes are there? Do all genes contain only one protein product?*

Run in terminal:

> docker run -v $PWD:/exercise/ --rm -i -t ummidock/breseq:0.32.1

> cd /exercise/

> breseq -r NC\_045512.2.gb -o SRR14773779\_NC\_045512.2\_breseq SRR14773779\_1.fastq SRR14773779\_2.fastq

* *What was the result of this command?*

Open the file output/index.html (it will open in a browser).

* How many variants are detected? What are the values of the quality (6th column in the vcf file) of the variants that are present in freebayes and not in breseq? Do you have variants in the Spike (S) protein? Does this allow you to know which sars-cov2 variant it is?

*> exit*

**Structural Variants**

Structural variants (duplications, deletions, chromosomal rearrangements) are large variants, usually larger than the size of the small reads coming from machines such as illumina. There are two main types of evidence for structural variants: differences in coverage, and junction evidence. In targeted sequencing (such as whole exome sequencing) usually only coverage evidence can be used (and even then, with limitations coming from artifacts such as amplification bias)

Open the file of a breseq result of an *Escherichia coli sample*

<https://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing/IntroWorkshop/REL11392_Ara+1_50K_clone_A/>

* How may structural variants are there? What evidence is used for each?

(Optional to do at home) Read paper explaining how breseq detects structural variants: <https://pubmed.ncbi.nlm.nih.gov/25432719/>

(Optional) Exercise:

*Using breseq, find variants from an Ecoli sample from the paper*

*https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004182. Namely, use the sample SRR625891 and the reference NC\_000913.3*

* *How many variants do you detect, in which genes? What types of variants do you detect?*

**Clonal versus Subclonal Variants**

When assessing clonal variants, the main goal is to uncover the genotype of a single individual (which can be haploid, diploid, or more generally, n-ploid). This is different from the general goal of discovering if a variant is true (not due to technical issues such as the base quality)

Look again at the variants from the vcf

SRR14773779\_NC\_045512.2.sorted.bam.freebayes.vcf

* What seems to be the most relevant factor influencing the quality (QUAL field) of a variant (look in the unknown field – the data regarding the sample)?
* Knowing that this is a virus, what do you think is the meaning of the heterozygous genotype in some of the variants? Do you think the default parameter that assumes a diploid organism is a good choice in this case?

Open the file of a breseq result of an *Escherichia coli sample*

<https://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing/IntroWorkshop/REL11383_Ara+1_50K_population/>

* How many subclonal variants are there? Looking at the frequency of the subclonal variants, how many different clones are there?
* Look at the evidence for the variant in position 461,704. Is there a strand bias for that variant? What about a bias for low quality bases?

**Variant Annotation**

The most common variant annotation just checks whether a variant falls within a gene, and whether it can change its function. Evaluating the effect of variants in non-coding regions is usually much more complex. In the case of structural variants, they usually cause deletion or amplification of whole genes, although it can also cause partial alterations.

Diagram

Description automatically generated

Copy the vcf line corresponding to the variant in position 23,403 of Sars-Cov2

Paste the line in the ensemble Sars-cov2 VEP:

<https://covid-19.ensembl.org/Tools/VEP>

* What information do you obtain?

For humans in particular, variant annotation can become quite complex, with the incorporation of information of many sources, such as population structure, etc..

Look at the example of a well known variant:

https://www.ensembl.org/homo\_sapiens/Variation/Summary?v=rs713598