My dictionary

* **Transcriptomic** = analysis of RNA data (bulk or single cell)
* **Metadata** = all the descriptive information about the RNAseq samples
* **Phage** = virus that infects bacteria
* **Translation** = production of proteins from mRNAs
* **Infection** = phage attach the bacteria cell surface, inject genetic material and (integrate) replicate (lytic or lysogenic)
* **Lytic cycle** = phage replicates rapidly, makes new viral particles, lyses host
* **Lysogenic cycle** = phage genome integrates into host DNA and stays dormant and may produce virions
* **Prophage** = phage genome in lysogenic state
* **Lysogen** = host carrying prophage
* **Superinfective** = infection of bacteria carrying prophage
* **Superinfection exclusion** = bacterium that is already infected by one phage becomes resistant to infection by another phage
* **Hotspots** = genomic region with clusters of defence genes – defence islands
* **Defence system** = systems that produce small antiphage molecules, systems that rely on intracellular signal transduction through production of signalling molecules and systems that recognize conserved structural patterns of viral proteins to trigger immune responses
* **Domain** (Bacteria) > **Phylum** (Proteobacteria) > **Class** (Gammaproteobacteria) > **Order** (Enterobacterales) > **Family** (Enterobacteriaceae) > **Genus** (Escherichia) > **Species** (Escherichia coli) > **Strain** (E. coli K-12, E. coli O157:H7, …)
* **mOTUs** = autre forme de taxonomie créée à partir de gènes ultra conservés
* **Taxa** = any group of organisms in biological classification (can be anything in the hierarchy)
* **Transcript length** = RNA size in nucleotide (bp)
* **Sequencing depth** = number of time a specific base in read during the sequencing process
* **Illumina sequencing** = sequencing of DNA/RNA (fragmentation, adapters, denaturation, cluster generation by bridge amplification, DNA polymerase incorporates fluorescently labelled nucleotides, camera to capture signal)
* MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq = different Illumina sequencing platforms
* **Gram-positive** bacteria (+): thick peptidoglycan layer +no outer membrane + teichoic acids and lipoteichoic acids in the wall
* **Gram-negative** bacteria (–): thin peptidoglycan layer + surrounded by an outer membrane with lipopolysaccharides (LPS) +periplasmic space between membranes
* **Biofilm** = communities of microorganisms embedded into a self-produced matrix consisting of exopolysaccharides, extracellular DNA, vesicles, and proteins
* **Virion** = the extracellular, complete particle that spreads infection (genome, capsid, envelope)
* **GCA** = genome assembly GenBank (first published version)
* **GCF** = genome assembly RefSeq (validated and corrected by NCBI)
* **RNA extraction** = cell suspension centrifuged, pellet resuspended in TRIzol Reagent, cells lysed, total RNA extracted, DNA removed, total RNA quantified with nanodrop, rRNA depletion -> Illumina for sequencing
* **Latent period** = time between the phage infecting a bacterial cell and the lysis of that cell, when newly formed phage particles are released into the environment.
* **LB broth** = 5 g/L yeast extract, 10 g/L tryptone, and 0,5 / 5 / 10 g/L NaCl

RNAseq normalization methods

**CPM** = Counts per Million

Adjust for differences in sequencing depths across samples and provide relative expression values on a comparable scale by scaling the raw read counts of each gene by a sample-specific sequencing depth (total counts) and multiplying by a scaling factor of one million (to obtain counts per million).

**RPKM** = Reads per Kilobase Million = for single end RNA-seq

1. Normalize for read depth (calculate total number of reads per run and divide each read by this total number \* million)
2. Normalize for gene length (divide each read by length of the gene kb)

**FPKM** = Fragments per Kilobase Million = for paired end RNA-seq (sequencing possible from both ends)

**TPM** = Transcript per Million (similar to RPKM and FPKM but in a different order) -> gives proportions of expression of each gene

1. Normalize for gene length
2. Normalize for read depth

Bibliography

* DESeq2 documentation <https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>
* Defense Finder  <https://defensefinder.mdmlab.fr/wiki/>
* Diversity of defence systems <https://doi.org/10.1038/s41579-023-00934-x>
* Normalisation in R documentation <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>
* Normalisation methods explanations <https://pluto.bio/resources/Learning%20Series/navigating-rna-seq-data-a-guide-to-normalization-methods>
* RPKM, FPKM, TPM explanation <https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>

How to get the count data from the row data?

1. RNA-seq
2. Adapter trimming (removal of adapters used in Illumina) + QC (BBDuck)
3. Mapping (alignment with ref sequences GCF) + Filter (low reads <0.97 complementarity with ref + <40bp (Samtools))
4. Count # genes / genes (feature counts)
5. # / genes immunity
6. QC post analysis
   1. # reads / SRR
   2. # genes phages and bacteria detected