# Introduction to sequencing: Sequencing technology and preprocessing of sequencing data

Koen Van den Berge

6/25/2021

# The study of gene expression

The first part of this course has focussed on proteomics, studying the concentration of proteins in biological samples. We have seen that identification of proteins and measuring their respective concentrations are extremely challenging, leading to many technological and statistical challenges in order to interpret these data. In the second part of the course, we will now focus on measuring gene expression, i.e., measuring the concentration of mRNA molecules, that may eventually be translated into proteins, but may also have functions on their own.

# Sequencing technology

- Measuring mRNA molecules typically happens through sequencing.
- The technology continues to evolve at an incredible speed. The data output of so-called 'next generation sequencing' machines has more than doubled each year! Simultaneously, the cost of sequencing (in terms of $ per Gigabase) is dropping. Each year, we're able to sequence more for less money.
- This tremendous technological revolution has revolutionized biology, and genomic sequencing is now a core component of the modern-day biologist's toolkit.
- The large majority of sequencing data is generated using sequencing-by-synthesis using machines produced by the company Illumina. While new players such as Pacific Biosciences and Oxford Nanopore have entered the scene, these are typically most useful for DNA sequencing rather than gene expression studies, owing to their capability of sequencing long molecules.

# Preprocessing of raw sequencing data

After sequencing, we typically do a quality control (QC) check to verify the quality of the samples. During QC check, aberrant samples due to e.g. degraded mRNA can be detected.

The sequencing reads on their own contain a lot of information, but are most useful if we would be able to assign sequencing reads to genomic features (genes, exons, transcripts, etc.), i.e., for each sequencing read we will try to derive the (set of) feature(s) that could have plausibly produced the fragment through the process of gene expression. This process is called 'mapping'. Most often we map reads to genes.

# References