# THE SEQUENCING QUALITY ATLAS

**Type of project*:**
> Bachelor End Project
> Master End Project (thesis or internship)

**Student background**:** Bachelor/Master CS

**Responsible supervisor:** Thomas Abeel – t.abeel@tudelft.nl (Pattern Recognition and Bioinformatics)

**Background**

DNA sequencing is an ever more accessible commodity, which is effectively becoming available to anyone. New entrants in the field however face the challenge to assess how well their experiments went. A next-generation sequencing dataset (Fastq / BAM) is effectively the aggregated results of a very large array of independent chemical chain reactions, each resulting into a single 'read' of DNA. As with any chemical reaction, the quality of each read depends on multiple factors, such as DNA purity, library preparation or GC-content of the DNA itself.

Currently available tools to assess the quality of Fastq and BAM files (FastQC or samstat) provide a plethora of statistics. This is problematic for two reasons: (i) Manually analyzing QC reports for many samples is labor intensive, and (ii) many key QC aspects are only evident within context of other samples (e.g. GC ratio), so QC for one sample ignores important problems.

To address both problems we want to build further on our high-throughput QC platform and create an atlas of QC metrics for all existing public DNA sequencing data.

This project builds upon a pilot project that was executed as part of the Contextproject course 2017-2018

Your job will be to:
- Design and implement a high-throughput system that can calculate QC metrics for all existing DNA sequencing data
- Design and implement a stream-like system that keeps QC metrics up-to-date for newly published data in the future.
- Investigate trends over time with evolving technology

**This project is complementary to another DBL project: "**High-throughput quality assessment of sequencing data"

---