Title: High-throughput quality assessment of sequencing data

Contact: t.abeel@tudelft.nl

Student: Student in computer science

Research question: What are key summary statistics from whole genome DNA resequencing data-sets to assess quality, and can we develop methods to quickly identify sub-par samples within a large population.

Context: 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) some key QC aspects are only evident within context of other samples (e.g. GC ratio), so reading QC for one sample may ignore some important problems.

We believe that QC analysis for BAM and Fastq files may be simplified by reducing the current QC statistics into a limited set of metrics. These key metrics may then be stored in a database system so that samples may be analyzed within the context of a larger population and over time.

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

Aims: Your job will be to:

- Identify and implement summary metrics that can be used to describe the quality a sample: mean read coverage?
- Classify good from bad samples by selecting good separating metrics
- Report and visualize selected metrics across large sequencing collections

Partner companies: Bayer Vegetable Seeds (Nunhems) is a leading innovative vegetable seeds company with over 1,200 seed varieties in ~25 vegetable crops. Rijk Zwaan Zaadteelt en Zaadhandel B.V. is a Dutch vegetable breeding and seed production company.

References:

1. Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D. and Davey, R.P. (2013) Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front. Genet.*, **4**, 1–5.

2. Trivedi, U.H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M. and Gharbi, K. (2014) Quality control of next-generation sequencing data without a reference. *Front. Genet.*, **5**, 1–13.