# Genetic variants and genetic data.

Deadline submission: 04.02.2019

**NB!** Part of the grading for this assignment will be based on how reproducible your analysis is. The answer to each task must include the command used or a full description of how the answer was found.

## Software

Recommended software for this exercise is to use either Plink (Plink v1.9, <https://www.cog-genomics.org/plink/1.9/>) or VCFtools (<https://vcftools.github.io/index.html>).

## Dataset

Download the provided dataset from: <https://drive.google.com/file/d/1KuWL2sLqJF8BaEsMkjFXt-pQ-n3heY7V/view?usp=sharing>

## Task 1

In addition to the original samples, the dataset also contains 10 samples that were artificially created. These samples all have one or more features that indicates they do not belong in the dataset. Use the following QC step(s) and any additional tests to identify these samples. Report each sample together with a description of why you think this sample does not belong.

## Task 2

Inspection of the dataset.

How many variants are there and how many samples?

Which chromosomes are present and how many variants are there for each chromosome?

## Task 3

Create histograms showing the distribution of missing values, MAF and HWE values for all SNPs. Also create a histogram of missing values for the samples.

NB! Plink does not provide plotting functionality so you have to use another program (e.g. R, Python, Excel).

## Task 4

The transition/transversion ratio for single base mutations are usually consistent for a given species (in humans it is about 2.1). However, this value is not consistent across the whole genome and in coding regions the ratio is often higher (>3). Describe why the codon table for amino acid translations can be used to explain why the Tv/Ts ratio is higher in coding regions of the genome.

## Task 5

Create a cleaned version of the dataset, removing all SNPs with MAF < 1%, missing values > 5% and p-value for HWE < 1e-5, and all samples with more than 5% missing values. How many SNPs and samples are left in the cleaned dataset?