# Assignment Omics, GWAS and inheritance (5BD002 HT25)

This assignment will be presented in group format; you have **maximum 20 minutes** (see schedule for each group's assigned time) to present the results from your group's investigations according to assigned phenotype and for both of the two different sections below.

## Methods for family data

You will get a simulated dataset, same format as in practical, downloadable from Canvas, see "Modules -> Omics, GWAS and inheritance (Week 43/44)", Assessment section (here you will also find your assigned phenotype).

Help for theory and code is available in lecture notes and the distributed "Cheat sheet_Family_methods.R".

***The overall aim is to assess if trait X (i.e., your assigned phenotype) is familial and heritable.***

Suitable steps are

1. Assess the similarity between siblings in trait X
    a. Reformat data from standard format to suitable format for assessing familiality
        i. What data shape did you reformat to?
    b. Use suitable methods for the data type
        i. Which method did you use
2. Assess the associations in trait X between siblings of different types,
    a. Choose suitable siblings to contrast to assess if the trait is likely heritable (i.e., affected by additive genetics)
        i. Which siblings did you compare?
    b. Choose suitable methods to assess the associations
        i. Which method did you choose
3. Estimate heritability using twins
    a. Sub-set data to twins
        i. How did you subset the data?
        ii. How many twins and twin pairs are there?
    b. Use suitable model and function for data type
        i. What type of model is it?
        ii. How did you find the estimate of heritability?

## GWAS for a (simulated) real trait

You will be given a `plink` binary format genotype dataset with a phenotype column included in the .fam file (download the folder matching your group's assigned trait from the GWAS data folder (GWAS_assessment)). The phenotype can be continuous or binary. Note that continuous phenotypes in this exercise can be trusted to follow a standard normal distribution reasonably well (this will not be case for any GWAS situation in the wild).

- Describe the genotype dataset (number of samples, number of variants), and the phenotype (distribution).
- Perform a principal component analysis to generate ancestry covariates. Visualize the population structure in your dataset. Could the genotype-phenotype relation be confounded by ancestry?
- Perform a genome-wide association analysis on the given data, first without any covariates, then adjusting for ancestry principal components.
- Show Manhattan and QQ plots of the GWAS results with and without PCA adjustment.
- Use `plink2` to perform LD clumping of the GWAS results, and annotate the clumps with gene names. Use clumping parameters

```
--bfile your_dataset_prefix \
--clump cols=-sp2,+bounds,-bins your_gwas_result_file \
--clump-p1 1e-5 \
--clump-p2 0.01 \
--clump-r2 0.1 \
--clump-kb 500 \
--clump-range glist-hg38 \
--clump-range-border 35
```

Are there any genes in your most strongly associated region? If there are, search online (for example GeneCards, OMIM) for a short description of gene function and any associated diseases or traits.

Search the literature for a reliable estimate of the heritability for your trait, and important genes/pathways if any, and present your findings. Do they agree with the results from your simulated datasets?