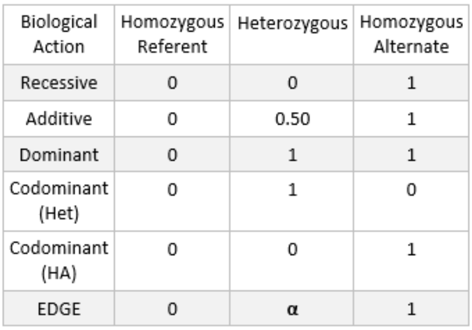
Individual-specific assignment

- Process and summarize the key reference paper linked to your assignment theme. The key paper is indicated in bold.
    - When processing the paper, pay attention to the actual assignment (see below)
    - In addition, pay attention to whether LD (measure of association between genetic markers) is allowed or necessary; and check whether missingness is allowed and/or dealt with
    - Also extract information about the protocol that allows optimal use the method on real-life data and collect information about IT demands (software installation, computational burden/ computation time, ancillary software that is needed, etc)
- Create presentation slides, that show how you processed the paper.
- At the time of your presentation, we will further explain the connections between the paper and the assignment theme. We will also identify gaps of knowledges that we will deal with in follow-up personalized sessions.
- Even though final reports will be individual, we invite you to maintain close contact to each other and to take advantage of available complementary backgrounds.

Student	Assignment	Group
ESCORCIO REBOLO Maria (20208021) - Leonor	**Background:** GWAIS studies try to identify genetic markers (single nucleotide polymporphisms – SNPs) to an outcome of interest, such as Inflammatory Bowel Disease (IBD). The key models in such studies are regression-based association models. Hence, these models do not necessarily predict IBD well. However, based on the effect sizes obtained from GWAS, it is possible to derive a so-called risk prediction score. When based on GWAS markers, these are called polygenic risk scores (PRS). An overview of PRS construction methods is given in **https://www.nature.com/articles/s41596-020-0353-1.pdf?origin=ppub**, which at the same time provide a recipe for their construction. **The problem**: Use IBD data https://www.prsice.info/ (or another method via the review above) and to compute PRS. Take IBD as dichotomous trait and default options in prsice. Follow aforementioned references to check how the PRS can be optimized (this may include pruning the data further, or expanding the data with all possible SNPs that are available, etc). Using different PRS versions, we will adjust the binary trait and will provide you with code to set up epistasis runs in PLINK with an IBD trait that has been adjusted for your computed PRS scores and population structure. These runs will take their time. Hence, it is best to set them up, once you have an optimized PRS score. In the end, compare the differences in the epistasis outcomes. Additional reference: application to Alzheimer https://www.frontiersin.org/articles/10.3389/fdgth.2020.00014/full	31
FARIA PEREIRA Lara Isabel (20207997)	**Background:** Multiple methods exist to carry out a GWAIS study that aims to identify or capture gene interactions (usually via SNP-SNP interactions). It is not unusual for different methods to give rise to different results. One explanation may be that different methods highlight different genetic architectures, simply	31

	by the nature of those methods. EDCF is one such method: **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3244765/** **The problem**: Apply the simple, fast and effective EDCF algorithm to detect genome-wide multi-locus epistatic interactions based on the clustering of relatively frequent items. Extensive experiments on simulated data show that the EDCF algorithm is fast and more powerful in general compared to methods that have been proposed earlier. Once applied to the two available IBD data sets, compare and try to understand the results. You may also want to compare your results with epistasis results obtained by fellow students.	
WLASOW-WLASOWSKI Nicolas (20151047) -	**Background:** Multiple methods exist to carry out a GWAIS study that aims to identify or capture gene interactions (usually via SNP-SNP interactions). It is not unusual for different methods to give rise to different results. One explanation may be that different methods highlight different genetic architectures, simply by the nature of those methods. GWGGI is one such method: **https://bmcgenomdata.biomedcentral.com/articles/10.1186/s12863-014-0101-z** **The problem**: The C++ package, GWGGI, for high-dimensional gene-gene interaction analyses, comprises two major functions, TAMW and LRMW, each of which can be used for genome-wide gene-gene interaction analyses without requiring a filter algorithm. In addition, each approach has its own uniqueness. While LRMW is suitable for the identification of gene-gene interactions among a few moderate-marginal-effect genetic variants, TAMW is designed for detecting gene-gene interactions involving hundreds low-marginal-effect genetic variants. Once applied to the two available IBD data sets, compare and try to understand the results.	31
LE Ba (20161106)	**Background**: Multiple methods exist to carry out a GWAIS study that aims to identify or capture gene interactions (usually via SNP-SNP interactions). It is not unusual for different methods to give rise to different results. One explanation may be that different methods highlight different genetic architectures, simply by the nature of those methods. The machine learning approach RF has often been said to be able to identify epistasis. In fact, this is not necessarily the case. It depends on which importance scores are implemented in the algorithm. For instance, classic RF approaches will generate a variable importance score (one for each variable), but not an importance score for pairs of variables. Hence classic RF approaches will at best be able to “capture” epistasis, never to “detect” epistasis. The following RF implementations are able to “detect” interactions in the aforementioned sense: Ranger, randomForestSRC, SNPInterForest, iRF, pRF, … **The problem**: Select RF implementations that are able to “detect” interactions. Understand the issues related to RF detection or capturing (**https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0995-8)** while applying the selected RF algorithms to the available data. on real-life data. You will need to work with imputed data (available to you) as RF algorithms typically do not handle missing data. Tuning parameters will require	30

<table>
<tr><td></td><td>working with training and test sets (or cross-validation). Tuning is needed if one wants to achieve optimal prediction performance. We have asked for the original code related to “individual trait prediction using interactions” (see additional ref). Once available, the performance can be compared with MBMDR. You may also want to compare the prediction performance with the PRS scores developed by your fellow colleague.
Additional references: https://github.com/aehrc/VariantSpark (addressing computational issues, even with fast implementations of RF) and https://www.researchgate.net/publication/330672990_Empowering_individual_trait_prediction_using_interactions</td><td></td></tr>
<tr><td>LIBERT Robin (20186942)</td><td>**Background**: Multiple protocols exist to carry out a GWAIS study that aims to identify or capture gene interactions (usually via SNP-SNP interactions). It is not unusual for different protocols to give rise to different results. One element of choice in analysis protocols is how to encode the input variables. The most commonly used encoding for SNPs is additive (counting the number of rare allels: 0, 1, 2). This encoding scheme works well for GWAS but has been challenged for GWAIS.
The problem: First implement RFcouple (**https://www.nature.com/articles/ejhg201048**).
This requires developing a fast way to create new variables: for each pair of SNPs, an individual’s “measurement” on the two SNPs is replaced by the proportion of cases to controls that have that same measurement on the two SNPs. Second implement EDGE encoding and apply the same RF algorithm that you used in RFcouple. Compare the results when applied to the available data. You may also compare your results to other RF-driven approaches obtained by fellow students.
Additional reference: https://pubmed.ncbi.nlm.nih.gov/25939665/;
EDGE encoding:

<table>
<tr><th>Biological Action</th><th>Homozygous Referent</th><th>Heterozygous</th><th>Homozygous Alternate</th></tr>
<tr><td>Recessive</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>Additive</td><td>0</td><td>0.50</td><td>1</td></tr>
<tr><td>Dominant</td><td>0</td><td>1</td><td>1</td></tr>
<tr><td>Codominant (Het)</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>Codominant (HA)</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>EDGE</td><td>0</td><td>α</td><td>1</td></tr>
</table>
A. $Y \sim \beta_{Het}SNP_{Het} + \beta_{HA}SNP_{HA}$
B. $\alpha = \beta_{Het} / \beta_{HA}$</td><td>30</td></tr>
<tr><td>MANGO LOPA Christophe</td><td>**Background**: In STRING was shown to be among the networks with the best performance overall to recover disease gene sets from GWAS hits (**https://www.cell.com/cell-systems/pdf/S2405-4712(18)30095-4.pdf**).
However, the way the authors propagate the disease genes on the reference</td><td>15</td></tr>
</table>

(20131615)	network (f.i. STRING) is of particular interest. In principle, pairs of genes identified via epistasis analysis also give rise to a unique interesting set of genes that can be propagated on a reference network. In addition, GWAIS results can be depicted as gene-gene networks; thus genes in such networks can be “clustered” based on their connectivity in the network, and clusterings (one for each epistasis detection protocol) can be “compared”. **The problem**: First apply the propagation method of the aforementioned paper (Louvain algorithm or another of your choice – see additional reference) to propagate results of epistasis runs across a reference network (f.i. STRING). Significant gene pairs will be given based on analyses we have already performed on the available data. Second, per GWAIS network, create a similarity graph between the gene objects to cluster. Compute the first k eigenvectors of its Laplacian matrix to define a feature vector for each object. Run a simple k-means on these features to separate objects into k classes. Optimize k (see additional references). Compute a consensus clustering (see additional references). Additional references: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159161 https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/ (optimal number of clusters) https://www.semanticscholar.org/paper/Consensus-Clustering-%2B-Meta-Clustering-%3D-Multiple-Zhang-Li/43ee77c0d9ae2ac836ac35358a447e8c037c17cf; https://bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf (consensus clsutering)	