Network Aggregation to Enhance Results Derived from Multiple Analytics

Diane Duroux, Héctor Climente-González, Lars Wienbrandt, Kristel Van Steen

Network aggregation to enhance results derived from multiple analytics

Diane Duroux[1], Héctor Climente-González[2,3,4], Lars Wienbrandt[5], and Kristel Van Steen[1,6]

[1] BIO3 - GIGA-R Medical Genomics, University of Liège, Liège, Belgium;
[2] Institut Curie, PSL Research University, F-75005 Paris, France;
[3] INSERM, U900, F-75005 Paris, France;
[4] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France;
[5] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany;
[6] WELBIO researcher, University of Liège, Liège, Belgium;
diane.duroux@uliege.be

Abstract. The more complex data are, the higher the number of possibilities to extract partial information from those data. These possibilities arise by adopting different analytic approaches. The heterogeneity among these approaches and in particular the heterogeneity in results they produce are challenging for follow-up studies, including replication, validation and translational studies. Furthermore, they complicate the interpretation of findings with wide-spread relevance. Here, we take the example of statistical epistasis networks derived from genome-wide association studies with single nucleotide polymorphisms as nodes. Even though we are only dealing with a single data type, the epistasis detection problem suffers from many pitfalls, such as the wide variety of analytic tools to detect them, each highlighting different aspects of epistasis and exhibiting different properties in maintaining false positive control. To reconcile different network views to the same problem, we considered 3 network aggregation methods and discussed their performance in the context of epistasis network aggregation. We furthermore applied a latent class method as best performer to real-life data on *inflammatory bowel disease (IBD)* and highlighted its benefits to increase our understanding about IBD underlying genetic architectures.

Keywords: Networks · Aggregation · Latent Class Methods · Epistasis

1 Introduction

Analyses carried out with different analytic tools often lead to inconsistent conclusions that are difficult to unify. In biology, integrative analyses usually aim at identifying the driving factors of a biological process by the joint exploration of several datasets, possibly reduced in dimension, or by obtaining a single solution per dataset prior to aggregation. All of these settings often involve a single analytical modelling framework to address the main question of interest.

Several aggregation methods exist and have been discussed in different contexts within human complex genetics [21]. Restricting attention to omics data, we mention the context of multi-omics analyses with supervised methods [13] for association or for prediction [17], and unsupervised methods for disease subtyping [29]. A returning common approach is the exploitation of network representations of the data. Here, nodes either represent samples (individuals) or biological features and edges represent interactions. Features may be directly measured or synthetic (modules); edges may be functional, biological or analytically derived via statistical and machine learning models.

In *genome-wide association interaction studies (GWAIS)* thousands of individuals, typed for genome-wide sets of genetic variants, are mined to identify interacting loci in association with a characteristic, such as disease state. The most popular genetic variants in these studies are *Single Nucleotide Polymorphisms (SNPs)*. In this paper, the main question of interest is how to derive unified conclusions from GWAIS with SNPs that have been typed out on the same dataset, yet with different analytic tools or protocols. The motivation for this question is multi-fold. In Bessonov et al. [2], it was demonstrated that slightly different GWAIS analysis protocols may lead to highly different analysis results. At the same time, different analytics are believed to highlight only particular aspects of the genetic architecture underlying complex traits under investigation. Hence, in order to aid in generating robust genetic interaction findings, that can be used as input to replication and experimental validation studies, there is a need for novel approaches to prioritize interactions obtained by different analytic workflows [32]. To our knowledge, the presented study is the first that explores the utility of network aggregation in deriving an aggregated statistical epistasis network across different epistasis detection analysis protocols.

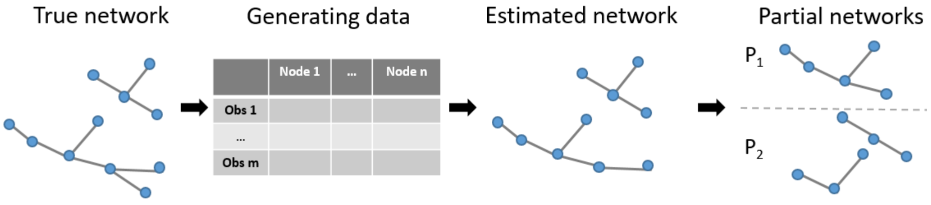
The remainder of this paper is organized as follows. We present in silico data and a case study on inflammatory bowel disease in Sect. 2. In Sect. 3 we outline the aggregation methods included in a comparative study. We report results in Sect. 4. Finally, in Sect. 5 we discuss and conclude this work.

2 Synthetic and Real-life Data

2.1 In Silico Data

We created several imperfect networks with binary edges (i.e., an edge is present or not) that partially represented a true network. In particular, we used the function *huge.generator* from the package *huge* [39] in R [26] to generate data with random graph structures. Essentially, we applied the number of observations ($n = 200$) and the number of variables ($d = 50$) as parameters. The adjacency matrix θ with probability $3/d$ that a pair of nodes is connected was computed via *huge.generator*. In other words, each pair of off-diagonal elements were randomly set to $\theta[i,j] = \theta[j,i] = 1$ for $i \neq j$ with probability $3/d$, and 0 otherwise. Then, a precision matrix was calculated from the adjacency matrix and was used to compute a covariance matrix in order to create the generating data. It led to a true baseline binary network with 50 nodes and a random graph structure, and the associated generating data.

Next, we created 5 so-called partial networks in the following way. We first applied the graphical lasso estimator (*glasso* option in the function *huge*) on the data of 200 samples and 50 nodes that we previously generated. The employed function carries out undirected graph estimation using a lambda sequence of size 10 to control the regularization. It returns a list of precision matrices corresponding to the lambdas. Second, the function *huge.select* was applied to select the regularization parameter. We applied the *stability approach to regularization selection (stars)*, which selects the optimal network by variability of subsamplings and gives a supplementary estimated network by merging the corresponding subsampled networks using the frequency counts. Then, to actually build a partial network, we randomly selected 50% of the edge values of the estimated graph and kept them as is. The remaining edge values were set to zero, i.e. representing the lack of interaction between corresponding nodes. This selection of 50% of the edges was performed five times, to give rise to 5 partial networks. Several variations to the baseline network were considered as detailed in Fig. 2. For each of the considered configurations we created 1,000 replicates. We highlight that the partial networks constructed in this way are in line with the hypothesis that *statistical epistasis networks (SENs)* derived from multiple analytics only partially reflect a true underlying interaction network.



	Node 1	...	Node n
Obs 1			
...			
Obs m			

Fig. 1. Generation of the data

2.2 Inflammatory Bowel Disease Data

IBD defines several chronic idiopathic inflammatory conditions. *Crohn's disease (CD)* and *ulcerative colitis (UC)* are the two main forms of IBD. UC is related to the colon, whereas CD affects the whole gastrointestinal tract and especially the terminal ileum and colon [9]. To date, identified independent loci associated with human complex diseases such as IBD only explain a small part of the disease heritability. As previous studies indicate, genetic interactions may have a significant role in this missing heritability [40], yet only a handful of replicable and clinically actionable interactions have been discovered [32].

Using data as part of the International IBD Consortium, we performed a first data *quality control (QC)* check as in Ellinghaus et al. [6]. Then, additional QC measures were taken, specifically related to large-scale GWAIS, as motivated in Gusareva and Van Steen [10]. In particular, only common variants (MAF $> 5\%$) and those in Hardy-Weinberg equilibrium (p-value > 0.001) were considered. Also, we pruned out SNPs that were in Linkage Disequilibrium (LD $r^2 > 0.75$) with the option "`--indep-pairwise 50 5 0.75`" in PLINK [25]. Since this LD filtering is based on sliding windows, LD was not tested exhaustively among all possible pairs. This may induce redundant epistasis signals due to LD and requires taking additional measures post interaction analysis (see Sect. 4.1). Lastly, to enrich the data for known risk loci, all risk SNPs described in Liu et al. [20] were included. In addition, we adjusted phenotypes to correct for population structure using the top 7 principal components. These adjusted phenotypes were obtained as residuals from a logistic regression model by subtracting model-fitted values from observed phenotype values. Submitting the phenotype adjusted traits to analytic tools may reduce power but is a pragmatic choice when the analytics do not accept covariates or explanatory variables other than the SNPs under investigation. Overall, the obtained dataset contained 38,225 SNPs and 66,280 individuals, partitioned in 32,622 cases and 33,658 controls.

3 Comparative Study – Towards Network-based Aggregation Methods for Statistical Epistasis Networks

In this project, we compared three unsupervised network aggregation methods. Given a set of edges and several networks, aggregation process was used to partition edges into two clusters [7], edge present or not, based on edge similarity across partial networks. The variables to assess similarity are the value of the edges in the different partial networks. The input matrix for clustering takes edges for rows and partial networks for columns. Matrix entries are 1 when an edge is present and 0 otherwise. First we selected one of the most popular unsupervised learning algorithms, k-means, using the function *kmeans* in R [26]. In particular, the *kmeans* function was applied to group the edges such that edges within the same cluster were as similar as possible, whereas edges in different clusters were as different as possible in order to maximize intra-cluster similarity, and minimize inter-cluster similarity. The algorithm of Hartigan and Wong [11] was applied. The total within-cluster variation was set as the sum of squared Euclidean distances between edges and the mean of edges in this cluster, called center. Each edge was associated with a cluster so the total within-cluster variation is minimized. In practice, two edges were picked randomly as cluster center. Then, each edge was assigned to their closest center based on the Euclidean distance. For each cluster, the center was updated by computing the mean values of all the edges in the cluster. The process was repeated 10 times to iteratively minimize the total within-cluster variation.

Second, we used the *Latent Class Modelling (LCM)* approach for clustering with the R [26] function *poLCA* [19]. It allows a dataset to be partitioned into exclusive groups called *latent classes*. The main latent class model is $P(y_n|\theta) = \sum_{j=1}^{S} \pi_j P_j(y_n|\theta_j)$ where y_n is the observation n (edge pair) of the variables (partial networks), S is the number of clusters (2), and π_j is the prior probability (random) of belonging to cluster j. P_j is the cluster specific probability of y_n given the cluster specific parameters θ_j. Expectation-Maximization algorithm was used to maximize the latent class mode log-likelihood function with *poLCA*. The output included a vector of predicted cluster memberships for each edge.

We also adapted the *Similarity Network Fusion (SNF)* approach from Wang et al. [35] to handle unweighted graphs. Note, that SNF was originally created for aggregating data types on a genomics scale so as to create an aggregated similarity matrix between individuals and that aggregation was based on normalized similarity matrices with continuous values. In our approach no normalization was performed and the partial networks were iteratively updated with information from the other networks to build an aggregated graph via the R library *SNFtool* [34] and the function *SNF* therein. We then set the diagonal of the aggregated adjacency matrix to 0. Since the outputted consensus network was continuous, it was binarized again by testing a variety of thresholds ranging from 0 to 1 with a step of 0.01 and selecting the threshold with maximal performance (simulation setting dependent). An edge was considered to be present in the final aggregated network if and only if the optimal threshold was surpassed.

Because most edges in the synthetic and real-life data are absent, we chose the F1-score to evaluate the performance of aggregation methods. It is defined as $F_1 = 2 \times \frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ and seeks a balance between precision (true positives divided by the number of true positives and false positives) and recall (true positives divided by the number of true Positives and false negatives). The F1-score ranges from 0 to its best value 1. First, we measured the initial F1-scores for each partial network compared to the true base network and selected the partial network with the highest score ($max(F1_{\text{Inital}})$). Then we computed the gain in F1 score ($F1_{\text{Gain}}$) defined as $F1_{\text{Gain}} = F1_{\text{AggregatedNetwork}} - max(F1_{\text{Inital}})$. The aggregation method for which $F1_{\text{Gain}}$ was the highest was our best performer.

Using the real-life data of Section.2.2, and by means of illustration, we applied 3 analytic methods to identify pairwise genetic interactions (hereafter referred to as epistasis). For each of these we subsequently constructed a statistical epistasis network with connected nodes representing SNPs involved in a significant interaction. The best performing network aggregation method was applied to obtain a single network comprising epistasis results from the 3 analytic methods.

The first method was regression-based and belongs to the most popular methods used in this context, as it is easy to implement and interpret. With PLINK 1.9 we fitted the linear regression model $E[Y|A, B] = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B$, where Y is the phenotype adjusted for population structure as described in Sect.2.2 and is assumed to follow a normal distribution, with g_A (g_B) representing genotype information for SNP A (B) under an additive encoding scheme, and with $\beta_i, i = 0, \ldots, 3$ the regression coefficients. The null hypothesis tested in PLINK [25] was H0: $\beta_3 = 0$ versus H1: $\beta_3 \neq 0$, i.e. a 1 degree of freedom test. Multiple testing corrected significance was assessed by creating permutation null samples while permuting Y values 400 times. We then produced a "top p-value" distribution with the smallest p-value of each permutation and we set an overall threshold at 5% of these top p-values. From that, we defined an overall p-value threshold with guarantee of 5% *Family Wise Error Rate (FWER)*, as in Hemani et al. [12].

The second method was a non-parametric dimensionality reduction method. In particular, we fitted *Model-Based Multifactor Dimensionality Reduction (MB-MDR)* [31] with default options that exhaustively explores the association between each SNP pair and Y adjusted for population structure as before. The method is non-parametric in the sense that no assumptions are made regarding the modes of interaction inheritance. Unlike the regression method above, MB-MDR is fairly robust to deviations from the normal distributions for Y, even though the final MB-MDR test for non-binary traits is by default the result of a sequence of t-tests. The Model-Based part of MB-MDR assumes the default of adjusting two-locus testing for main effects (SNP A, SNP B) and thus the considered MB-MDR alternative hypothesis was H1: the joint effect of SNP A and SNP B goes beyond additivity. Significance assessment with multiple testing correction was achieved by the default MB-MDR options of carrying out 999 permutations and gammaMAXT at a FWER of 5%.

The third method we considered was epiHSIC [16], as implemented in R's gpuEpiScan [15]. It searches for genomic interactions in a regression framework by efficiently scanning high-dimensional datasets. Efficiency is based on prescreening by HSIC, a statistical measure of non-independence between two variables: e.g. the larger HSIC value, the more likely it is that the correlation between SNP A and SNP B is independent from Y. Such independencies are believed to be indicative for potential epistasis. Significance was assessed via comparing obtained Bonferroni corrected p-values to 0.05.

PLINK analyses were performed on a cluster running Scientific Linux release 7.2 (Nitrogen), using 6 threads and the total runtime was 2 h 35 m. MBMDR analysis were implemented on the same computing system, with 100 threads and the total runtime was 2 h 05 m. EpiHSIC analysis was performed on CentOS Linux release 7.7.1908 (Core) cluster with 1 GPU (V100 GPU 2 x 12-Core Intel Xeon Gold 6126 2.6GHz 192GB RAM) and the total runtime was 20 minutes.

4 Results

4.1 Simulation Study

The average F1 gain is 0.18 for LCA, 0.13 for k-means and 0.01 for SNF with the baseline simulation scenario. Also, the more knowledge the partial network contains (i.e. percentage of edges overlapping with the estimated base network), the higher the initial F1 scores and the less beneficial the aggregation (Fig. 2A), which is in line with intuition.

In addition, when the number of observations in the generating data exceeds 500, then the F1 gain stabilizes (Fig. 2B). A too small number of nodes (here below 50) or a too high number of nodes (here above 100) shows to be suboptimal for both LCA and k-means. For SNF, the less nodes, the more beneficial the aggregation is, although its F1 gain is still the smallest of the 3 considered aggregation methods (Fig. 2C). The results of Fig. 2C may have repercussions for ”true” epistasis networks that would be too large in terms of numbers of SNPs. Part of the problem can be alleviated by deriving gene-based SENs from SNP-based SENs. In fact, to date, there is little evidence that the number of gene-based interactions would be extremely large, especially when ruling out spurious interactions due to major gene effects. The situation may be different for other interactome networks such as protein-protein interaction networks.

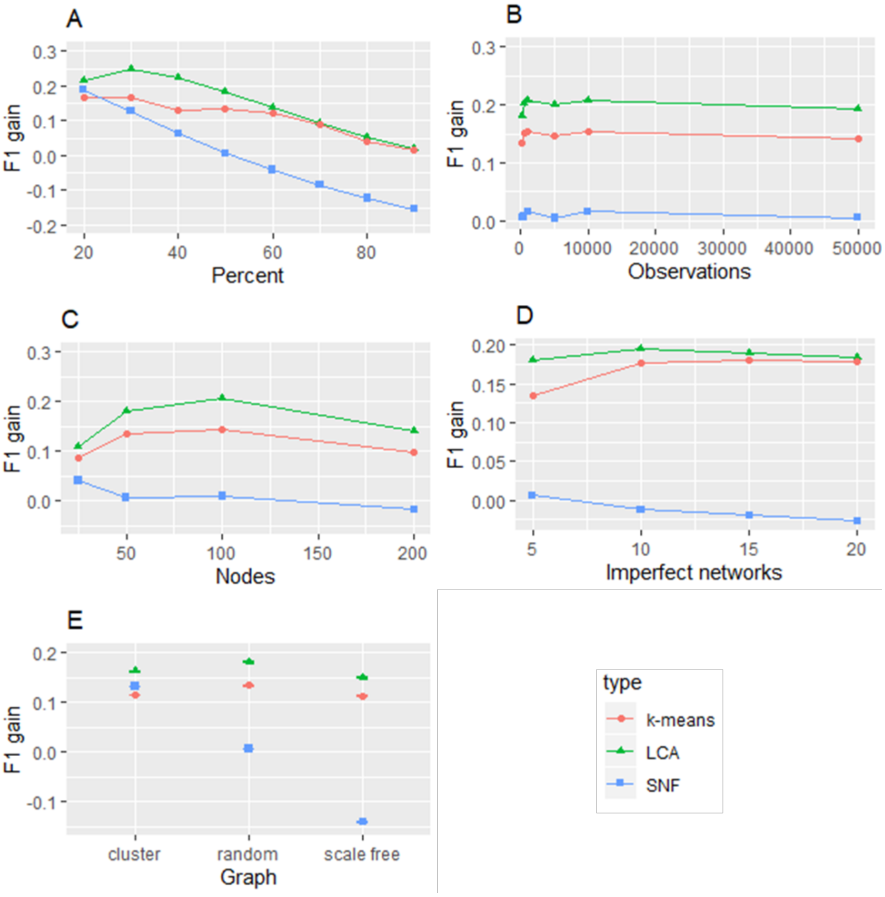


Fig. 2. Average F1 gain per simulation scenario. [A] Variation of the percentage of nodes of the estimated network used to create each partial network; [B] Number of observations in the generating dataset; [C] Number of vertices in the true network. [D] Number of partial networks aggregated to estimate the true network; [E] Graph type

Furthermore, with the SNF algorithm, F1 gain decreases with the inclusion of more partial networks in the aggregation process (Fig. 2D). In contrast, with k-means, the F1 gain increases as the number of partial networks varies from 5 to 20, whereas it remains stable with LCA when at least 10 networks are aggregated. Such information is relevant to have an idea about the number of epistasis networks (e.g. derived from different analytic protocols) to include in the aggregation process, with minimal loss of information compared to the "true" (unknown) underlying epistasis network. Notably, in real-life it is expected that non-random heterogeneity exists between partial epistasis networks. Not properly accounting for this may jeopardize the reliability of the aggregated network. Unfortunately, intrinsic differences between epistasis detection tools are often hard to assess based on the supporting literature that underlies each tool: to date there is no consensus about sufficiently advanced gold standard in silico datasets on human interactomes. Hence, a pragmatic way to deal with different forms of heterogeneity is to act at the level of the epistasis networks themselves, and includes accommodating potential scale differences in SEN edge weights across networks. We are currently working on a strategy around a notion of statistically significant differences between (groups of) SENs and clustering that combines the ideas of consensus clustering with meta clustering [4].

Finally, k-means and LCA, are quite stable across network structures, whereas SNF performs extremely poor on scale-free networks (Fig. 2E). The future will show what the implications are for the aggregation of SENs. Indeed, whether or not SENs or genetic interaction networks are scale-free is still under debate [3].

Based on all of the above, we selected LCA as SEN aggregation method of choice and compared the two LCA-derived clusters on the synthetic data, in more detail. In particular, we computed the average distance between and within clusters using the Manhattan distance for each of the 1000 runs. Overall, the average distance between clusters is 2.7 (standard error 0.005) and the average distance within groups is 0.15 (standard error 0.0001). Also, for each partial network and cluster, we calculated the frequency of 1's (i.e. edges present), to generate 1000 times two 5-dimensional vectors. Permutational multivariate analysis of variance with 1000 permutations (using R library *vegan* [24] and function *adonis*) shows that the clustering is significantly associated to edge abundance across partial networks (p-value of 0.001).

4.2 Inflammatory Bowel Disease Aggregated Statistical Epistasis Network

Here, the aim is to use knowledge derived from our simulation study to uncover the "true" epistasis network underlying inflammatory bowel disease, via multiple partial epistasis networks that are obtained from different analytic protocols on the same real-life data. As LCA performed best in Sect. 4.1, we applied it to combine 3 statistical epistasis networks for IBD, after further manipulation of the networks. We reduced the size of the networks while minimizing spurious edges. In particular, SNP pairs where both SNPs resided in the HLA region were deleted, as for this region it is notoriously hard to distinguish between main and additional non-additive effects [30]. Significant SNP pairs exhibiting strong LD ($r^2 > 0.75$) were eliminated as well. The resulting SENs are depicted in Fig. 3. The LCA aggregated SEN counts 193 nodes, 203 interactions and 12 modules. The size of the largest connected component (LCC) is 163.

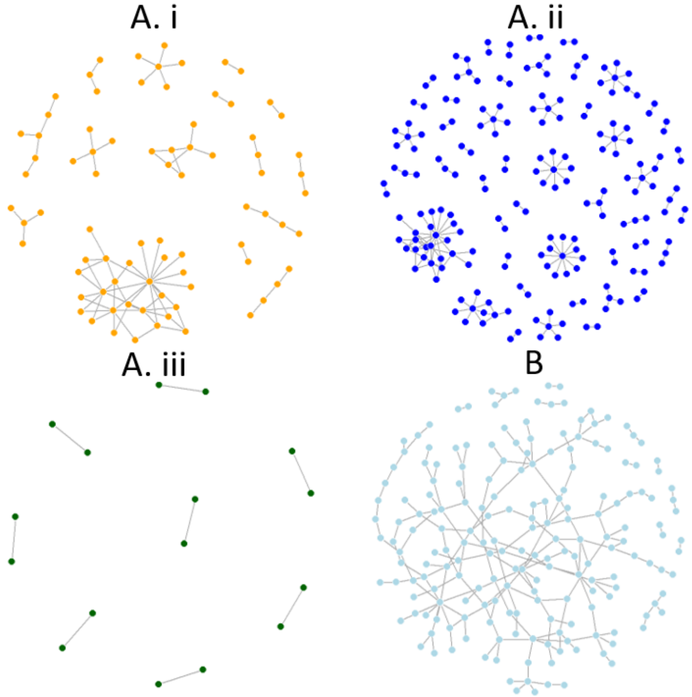


Fig. 3. SNP-based statistical epistasis networks (SENs). IBD SEN derived from [A.i] linear regression, [A.ii] MB-MDR, [A.iii] epiHSIC, [B] LCA aggregation.

To address the question whether the aggregated network gives added value over the contributing SENs to understanding underlying genetic architectures of IBD, we carried out several pathway enrichment analyses. To this end, we first mapped all SNPs of the LCA aggregated network to genes. This was done location-wise with FUMA [36] and its function *SNP2GENE*: SNPs were mapped to a gene whenever the SNP was located in that gene's region, i.e. including 10kb before and after the gene. Second, we ran pathway over-representation analyses of LCC containing at least 3 SNPs in R [26] using the library *clusterProfiler* [38] and the function *enrichKEGG*. FDR was controlled using the Benjamini-Hochberg procedure [1].

Since the LCC obtained with epiHSIC contained only 2 SNPs, no enriched pathway is obtained. The pre-screening approach from epiHSIC combined with stringent Bonferroni correction for multiple testing may not be a good choice since too much information gets lost, as we also saw with the small size of the associated network compared to the two other networks. For linear regression and MB-MDR the same 5 significant pathways were detected. This larger overlap between MB-MDR and linear regression is not surprising as neither of these methods involved a pre-screening, in contrast to epiHSIC. Also, filtering or not increases the heterogeneity in epistasis results [2]. The 5 pathways referred to cytokine-cytokine receptor interaction, JAK/STAT signaling pathway, Inflammatory Bowel Disease, Th17 cell differentiation and Th1 and Th2 cell differentiation and were already linked to IBD in earlier work [5, 8, 23]. Pathway enrichment analysis applied to the LCA aggregated SEN identified 12 significant pathways, including the 5 mentioned before. Therefore, in this case study, aggregation highlighted more pathways than the union of the pathways detected with each epistasis detection method. Note, that epiHSIC network contributed to the aggregated network. In fact, without including it, the LCA aggregated SEN lost 5 nodes, 1 module and 2 enriched pathways. The 7 unique pathways to the LCA aggregated SEN were viral protein interaction with cytokine, C-type lectin receptor, TNF, Yersinia infection, allograft rejection, intestinal immune network for IgA production and autoimmune thyroid disease. They seemed to be coherent with earlier work in relation to IBD [14, 18, 27, 33].

5 Conclusion

Genetic interactions, beyond effects of independent SNPs or genes, can further unravel the genetic underpinnings of human complex diseases. Such interactions contribute to epistasis, which has grown into a more general theory and applications framework for the analysis of interactions across and between multiple omics data. Many methods have been created to understand the true role of these interactions but findings are often inconsistent. This is in part due to different analytic protocols for epistasis detection giving rise to, at best, partially overlapping results. To this end, we first summarized the results of epistasis analyses in networks with nodes representing SNPs and edges representing binary evidence for a statistically significant interaction between corresponding SNPs. We second investigated the utility of network aggregation methods built on unsupervised machine learning to reconstruct the "true" disease underlying epistasis network. Unsupervised machine learning techniques have been used before in different contexts to unravel disease associated biological knowledge, for instance to derive multimodal biomarker signatures of disease risk [28], to identify subphenotypes for asthma [37], or to provide a molecular reclassification of Crohn's Disease [22]. Here, we used it to predict epistasis network links via the aggregation of partial networks. Our simulations revealed that *Latent Class Analysis (LCA)* outperformed k-means and a customized version of Similarity Network Fusion. We furthermore applied LCA to data for inflammatory bowel disease and underlined the benefits of an aggregated network via pathway enrichment analyses performed on the largest connected component of aggregated and contributing networks. These enrichment analyses revealed 7 pathways that could not be detected with either of the 3 considered statistical epistasis detection models. This pilot study suggests the potential of network aggregation in epistasis research and the need to investigate the added value of between-network heterogeneity in advanced network aggregation algorithms.

Acknowledgements

Data collections and processing were supported by funds to the International IBD Genetics Consortium. DD acknowledges the European Union's Horizon 2020 research and innovation programme under grant agreement No 813533.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
2. Bessonov, K., Gusareva, E.S., Van Steen, K.: A cautionary note on the impact of protocol changes for genome-wide association snp× snp interaction studies: an example on ankylosing spondylitis. Human genetics **134**(7), 761–773 (2015)
3. Broido, A.D., Clauset, A.: Scale-free networks are rare. Nature communications **10**(1), 1–10 (2019)
4. Caruana, R., Elhawary, M., Nguyen, N., Smith, C.: Meta clustering. In: Sixth International Conference on Data Mining (ICDM'06). pp. 107–118. IEEE (2006)
5. Coskun, M., Salem, M., Pedersen, J., Nielsen, O.H.: Involvement of jak/stat signaling in the pathogenesis of inflammatory bowel disease. Pharmacological research **76**, 1–8 (2013)
6. Ellinghaus, D., Jostins, L., Spain, S.L., Cortes, A., Bethune, J., Han, B., Park, Y.R., Raychaudhuri, S., Pouget, J.G., Hübenthal, M., et al.: Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nature genetics **48**(5), 510 (2016)
7. Faber, V.: Clustering and the continuous k-means algorithm. Los Alamos Science **22**(138144.21), 67 (1994)
8. Gálvez, J.: Role of th17 cells in the pathogenesis of human ibd. ISRN inflammation **2014** (2014)
9. Geboes, K., Dewit, O., Moreels, T.G., Faa, G., Jouret-Mourin, A.: Inflammatory bowel diseases. In: Colitis, pp. 107–140. Springer (2018)
10. Gusareva, E.S., Van Steen, K.: Practical aspects of genome-wide association interaction analysis. Human Genetics **133**(11), 1343–1358 (Nov 2014), 00015
11. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)
12. Hemani, G., Shakhbazov, K., Westra, H.J., Esko, T., Henders, A.K., McRae, A.F., et al.: Detection and replication of epistasis influencing transcription in humans. Nature **508**(7495), 249–253 (Apr 2014), 00162
13. Huang, S., Chaudhary, K., Garmire, L.X.: More is better: recent progress in multi-omics data integration methods. Frontiers in genetics **8**, 84 (2017)

14. Hütter, J., Eriksson, M., Johannssen, T., Klopfleisch, R., von Smolinski, D., Gruber, A.D., Seeberger, P.H., Lepenies, B.: Role of the c-type lectin receptors mcl and dcir in experimental colitis. PLoS One **9**(7) (2014)
15. Jiang, B.: gpuEpiScan: GPU-Based Methods to Scan Pairwise Epistasis in Genome-Wide Level (2019), r package version 0.0.1
16. Kam-Thong, T., Putz, B., Karbalai, N., Muller-Myhsok, B., Borgwardt, K.: Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. Bioinformatics **27**(13), i214–i221 (Jul 2011), 00026
17. Kim, M., Tagkopoulos, I.: Data integration and predictive modeling methods for multi-omics datasets. Molecular omics **14**(1), 8–25 (2018)
18. Koelink, P.J., Bloemendaal, F.M., Li, B., Westera, L., Vogels, E.W., van Roest, M., et al.: Anti-tnf therapy in ibd exerts its therapeutic effect through macrophage il-10 signalling. Gut pp. gutjnl–2019 (2019)
19. Linzer, D.A., Lewis, J.: polca: Polytomous variable latent class analysis version 1. 4. J Stat Softw **42**, 1–29 (2011)
20. Liu, J.Z., Van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.: Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics **47**(9), 979 (2015)
21. López de Maturana, E., Pineda, S., Brand, A., Van Steen, K., Malats, N.: Toward the integration of omics data in epidemiological studies: still a “long and winding road”. Genetic epidemiology **40**(7), 558–569 (2016)
22. Maus, B., Jung, C., John, J.M.M., Hugot, J.P., Génin, E., Van Steen, K.: Molecular reclassification of crohn’s disease: a cautionary note on population stratification. PloS one **8**(10) (2013)
23. Nemoto, Y., Watanabe, M.: The th1, th2, and th17 paradigm in inflammatory bowel disease. In: Crohn’s Disease and Ulcerative Colitis, pp. 183–194. Springer (2012)
24. Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O’hara, R., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., et al.: Package ‘vegan’. Community ecology package, version **2**(9), 1–295 (2013)
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al.: Plink: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics **81**(3), 559–575 (2007)
26. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017), https://www.R-project.org/
27. Saebo, A., Vik, E., Lange, O.J., Matuszkiewicz, L.: Inflammatory bowel disease associated with yersinia enterocolitica o: 3 infection. European journal of internal medicine **16**(3), 176–182 (2005)
28. Shomorony, I., Cirulli, E.T., Huang, L., Napier, L.A., Heister, R.R., Hicks, M., Cohen, I.V., Yu, H.C., Swisher, C.L., Schenker-Ahmed, N.M., et al.: An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. Genome Medicine **12**(1), 1–14 (2020)
29. Tini, G., Marchetti, L., Priami, C., Scott-Boyer, M.P.: Multi-omics integration—a comparison of unsupervised clustering methodologies. Briefings in bioinformatics **20**(4), 1269–1279 (2019)
30. Traherne, J.: Human mhc architecture and evolution: implications for disease association studies. International journal of immunogenetics **35**(3), 179–192 (2008)

31. Van Lishout, F., Gadaleta, F., Moore, J.H., Wehenkel, L., Van Steen, K.: gammaxt: a fast multiple-testing correction algorithm. BioData mining **8**(1), 36 (2015)
32. Van Steen, K., Moore, J.: How to increase our belief in discovered statistical interactions via large-scale association studies? Human genetics **138**(4), 293–305 (2019)
33. Wadhwa, V., Lopez, R., Shen, B.: Crohn's disease is associated with the risk for thyroid cancer. Inflammatory bowel diseases **22**(12), 2902–2906 (2016)
34. Wang, B., Mezlini, A., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Snftool: Similarity network fusion. cran. 2014 (2014)
35. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods **11**(3), 333 (2014)
36. Watanabe, K., Taskesen, E., van Bochoven, A., Posthuma, D.: Functional mapping and annotation of genetic associations with FUMA. Nature Communications **8**(1) (Dec 2017). https://doi.org/10.1038/s41467-017-01261-5, 00139
37. Woodruff, P.G., Modrek, B., Choy, D.F., Jia, G., Abbas, A.R., Ellwanger, A., et al.: T-helper type 2–driven inflammation defines major subphenotypes of asthma. American journal of respiratory and critical care medicine **180**(5), 388–395 (2009)
38. Yu, G., Wang, L.G., Han, Y., He, Q.Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. Omics: a journal of integrative biology **16**(5), 284–287 (2012)
39. Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.: The huge package for high-dimensional undirected graph estimation in r. Journal of Machine Learning Research **13**(Apr), 1059–1062 (2012)
40. Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S.: The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences **109**(4), 1193–1198 (2012)