# Network based multi-omics data integration and cluster analysis of inflammatory bowel disease patients

Tim Hensen
i6217294

Network Biology research report
Maastricht University

## 1 Abstract

Inflammatory bowel disease (IBD) is a group of inflammatory diseases with a complex and not completely understood pathogenesis. Currently, correctly diagnosing inflammatory bowel disease subtypes is difficult sometimes leads to inaccurately classified or unclassified patients. This study explores a novel way of describing IBD disease presentations by integrating metagenomics and metatranscriptomics data from IBD patients and producing 11 clusters using spectral clustering. Cluster analysis is done by comparing the 4 largest clusters by the average age of diagnosis and the male/female ratio. Interpretation has proven difficult due to the low sample size of the clusters and the lack of molecular data needed for GO and pathway analysis. In spite of these limitations, this study demonstrates a first step towards describing IBD patients, not only by subtype but also by multi-omics molecular profiles.

## 2 Introduction

Inflammatory bowel disease (IBD) is a group of two chronic diseases characterized by inflammation of the gastrointestinal tract leading to pain in the abdomen, chronic diarrhoea, fever and bloody stools. Most western countries have a disease prevalence rate of 0.3% which has been stable for the past decade. Newly industrialized countries, mainly in Asia and Africa however, are seeing a sharp rise in the prevalence of IBD cases. This rise in newly industrialized countries is thought to be linked with an increased urbanization and relatively more people living in cities (1). In spite of decades of research being done on IBD, the disease is still not completely understood, as its pathogenesis contains many factors including genetic, metabolic, immunologic and environmental factors (2). IBD consists of two subtypes, Crohn's disease (CD) and Ulcerative Colitus

(UC). Both subtypes have a similar disease presentation but generally, but not always differ in their location and inflammation patterns (3) (4). Although differences between CD and UC are well established, subtype diagnosis has remained to be difficult and error prone. Between 4 and 15% of patients undergo reclassification while between 5 and 15% of IBD patients remain unclassified (5). Modern diagnosis is focused primarily on a combination of imaging techniques like colonoscopy and simple biomarkers. Recently, research has started to tackle IBD subtype diagnosis using an omics approach. Kohlo et. al. 2016 has build metabolic profiles of classified IBD patients to find metabolites that correlate with each subtype. These metabolites might serve as future biomarkers (6). Santoru et. al. 2017 used 16 rRNA to determine how changes in the microbial makeup of the gastrointestinal tract correlate with each subtype (5). This project is build on the idea of using a multi-omics approach to better characterize IBD patients. In contrast to earlier omics based IBD research, the goal of this project is not to better classify patients in the two subtypes but to gain information about the state of IBD patients in a more holistic sense. This approach aims to bypass the vaguely defined molecular differences between CD and UC patients in favour of a more unbiased molecular profile of patient specific characteristics of IBD patients. In this report I demonstrate a novel approach to describe IBD patients by obtaining patient clusters with similar characteristics. Network based data integration of metatranscriptomics and metagenomics datasets is carried out. Subsequently, clusters are produced using spectral clustering. These clusters were analyzed using information from metadata on the samples present in the clusters.

# 3 Methods

Multi-omics data of IBD patients is acquired from the: "Inflammatory Bowel Disease Multi-Omics Database" (7). The metagenome and metatranscriptomics pilot datasets were chosen for data integration. Both datasets contain pathway abundances of human gut microbes of IBD patients, determined by gene or RNA read counts respectively. Pilot datasets were chosen because of their smaller size as the full datasets were too large to analyze within the given time limit. After data acquisition, the datasets were checked for strange and missing values. Although both datasets were made in the same study, the sample sizes differed significantly with 316 samples for the metagenomics and 82 samples for the metatranscriptomics dataset. Taking the intersection of these two datasets by sample, resulted in 82 overlapping samples. The metagenomic dataset was filtered to contain only overlapping samples.

## 3.1 Data integration

After filtering, distance matrices were produced based on the euclidean distances between the samples. The distance matrices were then used to produce similarity matrices. To determine the similarity matrix, the number of neighbours of a node (K) was set to 20, while a hyper parameter alpha was set to 0.5. These are

standard parameters, used and recommended in Wang et. al. 2014 (8). The produced similarity matrices were then used as input for the similarity network fusion (SNF) algorithm which was run with 10 iterations. Here, the two similarity matrices were integrated into a single network consisting of the overlapping samples for the metagenomics and metatranscriptomics pilot datasets.

### 3.2 Clustering

After network integration, spectral clustering was done. This clustering technique is chosen as it works well on similarity graphs and often outperforms other clustering methods like k-means and hierarchical clustering (9). Spectral clustering has the disadvantage that the number of clusters need to be set manually. To find the optimal number of clusters, the clustering performance is determined by calculating the normalized mutual information score (NMI). NMI is a normalized measure of the overlap between clusters. A higher score corresponds with higher overlap between clusters (10). The NMI score is determined for 1 to 82 clusters. By visualizing the NMI score over the number of clusters, a cluster number of 11 was chosen as this resulted in not too much clusters to analyze and a higher NMI score than would be obtained with less clusters, as can be seen in figure 1 in section 6. This part of the methods is done in Rstudio and can be found on github (11). After cluster creation, visualization was done by loading the samples with their cluster labels in Cytoscape 3.8.1 (12).

### 3.3 Cluster analysis

After clustering, cluster characterization was done. First, the metadata from the metagenomics and metatranscriptomics pathway abundances datasets was obtained. The metadata consisted of the patient ID's, the obtained datatype and various psychometric and demographic data. The patient sex and age of diagnosis were selected to analyze the found clusters. Next, the size of each of the found clusters was determined. Because of time constraints, the 4 largest clusters were taken and the entries of each cluster were used to filter the metadata. This resulted in tables for each cluster with the age of diagnosis and the sex for each patient in the cluster. In order to compare the 4 clusters, the mean age and female to male rate were taken. Cluster characterization was done in a Python Jupyter notebook (13). The source code can be found on github (11).

## 4 Results

The NMI score for each number of clusters is displayed in figure 1. This score varies between 0.04 for 2 clusters to 0.4 for 82 clusters. The found curve is roughly linear with a flattening between 35 and 60 clusters. 11 clusters are chosen and visualized. These clusters are displayed in figure 2. The cluster size is displayed in figure 3. After importing the clusters in Python, the four largest clusters were compared and characterized by

their mean age of diagnosis and male/female ratio. This resulted in table 1. This table shows that cluster 6 contains patients with on average a higher age of diagnosis compared to, especially, cluster 10 and 2. Another difference found between cluster 6 and the other clusters was a much lower male to female ratio of 0.08 compared to 0.71 in cluster 10, 0.29 in cluster 2 and 1.25 in cluster 9.

| | Average age of diagnosis | Male/Female ratio | Cluster size |
|---|---|---|---|
| Cluster 6 | 33.2 | 0.08 | 13 |
| Cluster 10 | 21.3 | 0.71 | 12 |
| Cluster 2 | 26.6 | 0.29 | 9 |
| Cluster 9 | 29 | 1.25 | 9 |

Table 1: Characteristics of the largest 4 clusters

# 5 Discussion

This project showed a new approach to classify and describe IBD patients. Network based data integration using SNF has been carried out and clustering has been done using spectral clustering. Information about the age of diagnosis and the male/female ratio of the clusters has been inferred using metadata from the used datasets. The clusters shown in table 1 are derived from the metagenomic and metatranscriptomic datasets and reflect molecular differences in pathways from gut microbes in IBD patients.

The onset of Crohn's disease peaks between the age of 20 and 30 while the onset of Ulcerative Colitus peaks between the ages of 30 and 40 (14). As the IBD patients are unclassified this makes cluster 6 (average age of diagnosis = 32) more likely to contain mostly UC patients while cluster 10 (average age of diagnosis = 21) might contain mostly CD patients. Previous demographic research to IBD patients also found differences between CD and UC sex ratio's. CD patients were found to have a male/female ratio of 1 while UC patients were found to have a male/female ratio of 1.7 (15). Table 1 shows cluster 6, 10 and 2 to have sex ratio of about 1 or lower, making them more likely to be classified as CD patients based on this metric. Cluster 9 has a m/f ratio of 2.2, making this cluster more likely to consist of mostly UC patients.

Interpreting these results should done very carefully. The number of clusters is chosen arbitrary and as this number impacts the makeup of these clusters, the results will differ when a different number of clusters is chosen. Also, the network integration is based only on metagenomic and metatranscriptomic data which refers only to microbial abundances in the human gut. Host genetics, proteomics or metabolomics data is not included making the obtained network from SNF integration not ideal for further analysis.

The approach chosen in this study could be improved in a few ways. First, by choosing of the number of

cluster based on the NMI score didn't result in a clear point to take the optimal number of clusters. The optimal number of clusters could be determined automatically using other methods like DBSCAN (16). This distance based algorithm is related to spectral clustering but uses only the distance metrics for each node and the minimal cluster size as parameters. The distance measures could be acquired from the fused similarity network. Another improvement could come from using more types and more relevant omics datasets. Using host genomics, transcriptomics and metabolomics datasets could improve the obtained network after data integration, resulting in more reliable and robust results. Also, when gene/protein reads are available for these datasets, extensive gene ontology and pathway analysis could be carried out. This would improve the precision of the cluster analysis a lot.

In conclusion, this project attempted to find and characterize clusters of IBD patients using multi-omics data integration in order to explore an novel way of describing different IBD disease presentations. Problems in study design and data acquisition hampered the depth and reliability of the cluster analysis. Nevertheless, this project managed to be a first small step towards making accurate clustering based molecular profiles of IBD feasible.

# 6 Figures

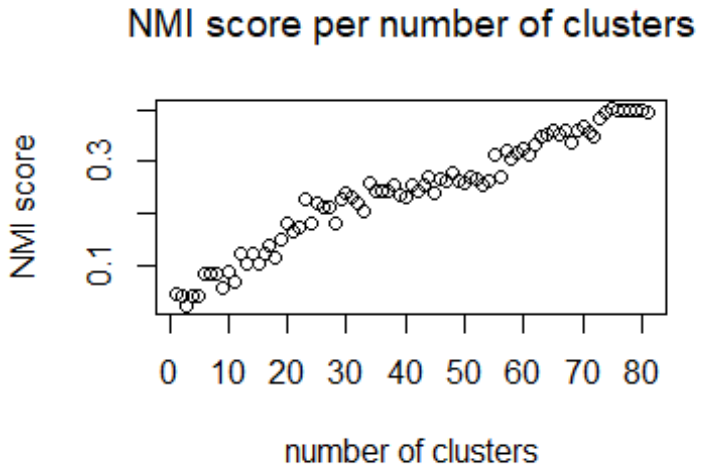


Figure 1: NMI scores for each number of clusters

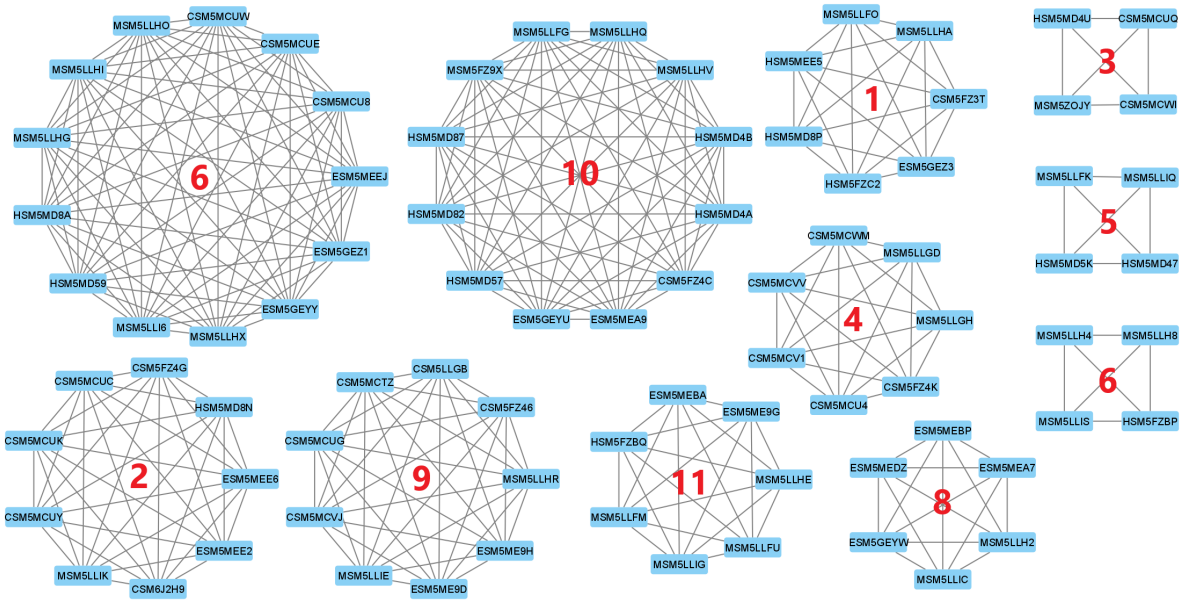


Figure 2: The 11 found clusters after spectral clustering. The number seen each cluster corresponds to the cluster number in figure 3.

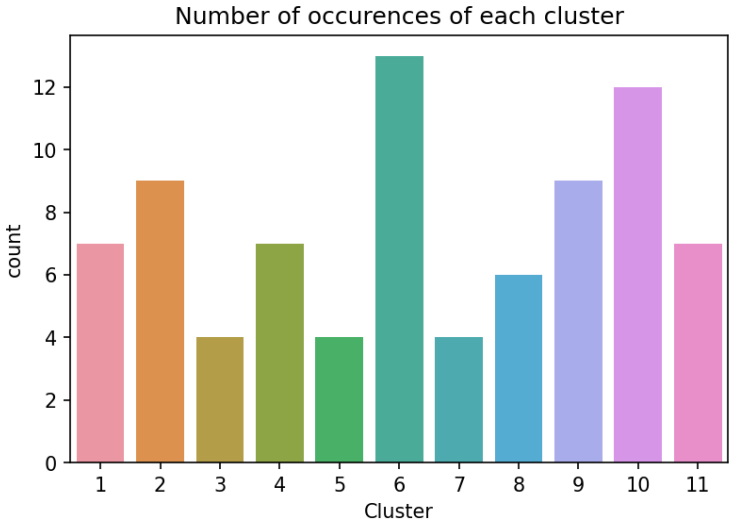


Figure 3: Number of samples per cluster

## References

[1] Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based

studies. The Lancet. 2017 Dec;390(10114):2769–2778. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673617324480.

[2] Atreya R, Neurath MF. Molecular pathways controlling barrier function in IBD. Nature Reviews Gastroenterology & Hepatology. 2015 Feb;12(2):67–68. Available from: http://www.nature.com/articles/nrgastro.2014.201.

[3] Head KA, Jurenka JS. Inflammatory Bowel Disease Part I: Ulcerative Colitis – Pathophysiology and Conventional and Alternative Treatment Options. Alternative Medicine Review. 2003;8(3):37.

[4] Baumgart DC, Sandborn WJ. Crohn's disease. The Lancet. 2012 Nov;380(9853):1590–1605. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673612600269.

[5] Santoru ML, Piras C, Murgia A, Palmas V, Camboni T, Liggi S, et al. Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. Scientific Reports. 2017 Dec;7(1):9523. Available from: http://www.nature.com/articles/s41598-017-10034-5.

[6] Kolho KL, Pessia A, Jaakkola T, de Vos WM, Velagapudi V. Faecal and serum metabolomics in paediatric inflammatory bowel disease. Journal of Crohn's and Colitis. 2016 Sep:jjw158. Available from: https://academic.oup.com/ecco-jcc/article-lookup/doi/10.1093/ecco-jcc/jjw158.

[7] IBDMDB Investigators, Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019 May;569(7758):655–662. Available from: http://www.nature.com/articles/s41586-019-1237-9.

[8] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 2014 Mar;11(3):333–337. Available from: http://www.nature.com/articles/nmeth.2810.

[9] von Luxburg U. A tutorial on spectral clustering. Statistics and Computing. 2007 Dec;17(4):395–416. Available from: http://link.springer.com/10.1007/s11222-007-9033-z.

[10] McDaid AF, Greene D, Hurley N. Normalized Mutual Information to evaluate overlapping community finding algorithms. arXiv:11102515 [physics]. 2013 Aug. ArXiv: 1110.2515. Available from: http://arxiv.org/abs/1110.2515.

[11] Tim H. https://github.com/trjhensen/Network-biology;.

[12] Su G, Morris JH, Demchak B, Bader GD. Biological Network Exploration with Cytoscape 3. Current Protocols in Bioinformatics. 2014 Sep;47(1):8.13.1–8.13.24. Available from: http://doi.wiley.com/10.1002/0471250953.bi0813s47.

[13] Kluyver T, Ragan-Kelley B, Pérez F, Bussonnier M, Frederic J, Hamrick J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows:4.

[14] Ruel J, Ruane D, Mehandru S, Gower-Rousseau C, Colombel JF. IBD across the age spectrum—is it the same disease? Nature Reviews Gastroenterology & Hepatology. 2014 Feb;11(2):88–98. Available from: http://www.nature.com/articles/nrgastro.2013.240.

[15] Tragnone A, Corrao G, Miglio F, Caprilli R, Lanfranchi GA. Incidence of Inflammatory Bowel Disease in Italy: A Nationwide Population-Based Study. International Journal of Epidemiology. 1996;25(5):1044–1052. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/25.5.1044.

[16] Tran TN, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometrics and Intelligent Laboratory Systems. 2013 Jan;120:92–96. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169743912002249.