# From Comorbidities to Gene Expression Fingerprints and Back

Beatriz Urda-García[1,2], Alfonso Valencia[1,3*]

[1]*Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain*

[2*] *Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain,*[3*] *ICREA, Barcelona, Spain*

[1]beatriz.urda@bsc.es, [3]alfonso.valencia@bsc.es

***Keywords*— Comorbidity, gene expression, RNA-seq**

## EXTENDED ABSTRACT

Epidemiological evidence shows that some diseases tend to co-occur more than expected by chance and that patient-specific trends are observed. However, the molecular processes underlying these phenomena remain unclear.

Here we exploit the accumulating RNA-seq data on human diseases to calculate disease similarities at the transcriptomic level. We build a disease similarity network that significantly captures almost half of the medically known comorbidities, substantially outperforming previously published methods and providing biological explanations for such co-occurrences. Additionally, we group patients from a given disease with a similar expression profile into meta-patients and calculate their molecular similarities with the analyzed diseases, highlighting the need to study disease comorbidities within a personalized medicine scope. Finally, we provide a web application in which the networks and their underlying molecular mechanisms can be easily inspected.

### *A. Introduction*

Comorbidity, defined as the co-occurrence of two or more diseases in the same patient, is a complex medical problem that has become a key research area due to the associated increased Disability-Adjusted Life-Years (DALYs), complex clinical management and health care cost [1].

Accumulating evidence from epidemiological studies indicates that some diseases co-occur more than expected by chance and that patients suffering from the same disease present different risks of developing secondary conditions [2].

To tackle this problem, a better understanding of the molecular processes driving comorbidity relationships is essential. In line with this, several studies have analyzed disease similarities using molecular information (disease-associated genes in protein-protein interaction networks (PPINs) [3], microbiome, miRNA or microarrays [4]). Although these efforts were able to meaningfully capture interesting examples, they were unable to recapitulate what is known at the medical level in a considerable manner.

Here, we have reformulated the problem and we show, for the first time, that actually gene expression data – RNA-seq data – is able to reproduce medical interactions in a substantial and improved way. Additionally, we introduce the concept of meta-patients (molecularly similar patients from a given disease), that allows for the exploration of subgroup-specific patterns.

### *A. Methods*

First, we collected RNA-seq studies comprising 72 human diseases from the Gene Expression Omnibus. Then, we developed an RNA-seq pipeline destined to the parallel processing of a collection of RNA-seq studies for a given set of diseases. Afterwards, we performed Gene Set Enrichment Analyses to obtain the significantly altered gene sets and pathways for each disease.

Next, we defined a Disease Similarity Network (DSN) in which we connected diseases based on the similarities of their differential gene expression profiles. Specifically, for each disease pair, we computed the Spearman's correlation between the logFC values of the genes in the union of their significantly differentially expressed genes (sDEGs). We kept the interactions that were significant after correcting for multiple testing (FDR <= 0.05).

Since epidemiological networks only describe positive comorbidity relationships, we evaluated the overlap of the positive interactions in our DSN with the ones described by Hidalgo *et al*.[2] (based on medical records). To do so, we transformed our disease names into the International Code of Diseases, version 9 (ICD9 codes), computed the overlap of the networks and assessed its significance by shuffling the interactions while preserving the degree distribution. Next, we followed the same methodology to compare our overlap with the ones obtained with other disease-disease networks based on molecular information (microbiome, miRNAs and disease-associated genes in PPINs [3]).

Going into a deeper detail, we stratified diseases into subgroups of patients with similar expression profiles (meta-patients) by applying clustering algorithms to the normalized and batch effect corrected gene expression matrix. Both PAM (k-medoids) and Ward2 algorithms were applied independently. Next, we performed differential expression analyses and functional enrichment to the obtained meta-patients, and built a Stratified Similarity Network (SSN) by connecting meta-patients and diseases in the previously described manner.

### *B. Results and discussion*

First, we collected published studies analyzing human diseases with RNA-seq data. After quality filtering, 58% of the samples were kept, corresponding to 2.705 samples from 62 studies and comprising 45 diseases. We performed differential expression analyses to obtain sDEGs for each disease and functional enrichment analyses to better understand the transcriptomic alterations associated with them. We showed that the diseases' altered molecular processes match their known pathophysiology. We also discussed cases in which such processes can be involved in the existence of medically known comorbidities.

Next, we built a disease-disease similarity network (DSN) connecting diseases based on the similarity or dissimilarity of their gene expression profiles. The resulting network contains one single connected component and a higher percentage of positive than negative interactions (63.37% versus 36.63%). The DSN captures many known disease comorbidities, like the relationship between Chron's disease, ulcerative colitis and colorectal cancer; comorbidities between neoplasms, like lung and liver cancer; and multiple described relationships among mental and nervous system disorders, such as the one of schizophrenia with bipolar disorder, autism or Parkinson's and Huntington's diseases (HD). Interestingly, we also

observe some negative correlations that reflect known inverse comorbidity patterns, defined as a lower than expected risk of disease co-occurrence. For instance, the decreased risk of developing different types of cancer (liver, lung, breast and chronic lymphocytic leukemia) in HD patients is corroborated by a negative correlation in our DSN. Moreover, since we have the gene expression fingerprint of all the diseases at different levels of granularity (genes and pathways), we can inspect the molecular mechanisms that may underlie the observed relationships. We should consider that the presence of shared molecular mechanisms does not always reflect a comorbid relationship. However, we have included detailed examples in which the dysregulation of key physiopathological pathways is shared between comorbid diseases and shows an opposite pattern for inverse comorbidities, revealing crucial aspects of such disease relationships.

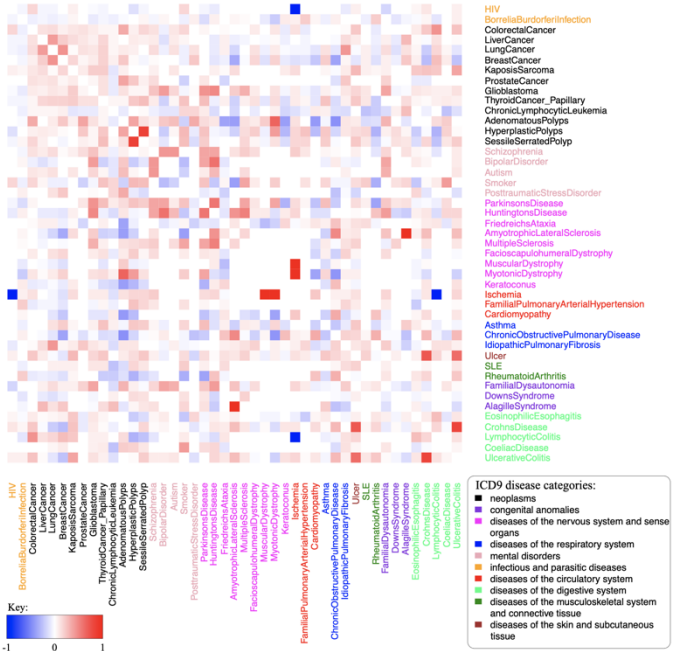


Fig. 1 Disease-disease Similarity Network (DSN). Pairwise disease correlations were computed based on the Spearman's correlation of the union of the sDEGs of each pair of diseases. A disease-disease network was built, containing the significantly positive and negative correlations (FDR <= 0.05), where the edge weights correspond to the Spearman's correlations. The heatmap shows the positive and negative disease interactions, in red and blue respectively. Diseases are coloured by ICD9 disease category.

Subsequently, we evaluated to what extent our DSN is able to capture medically known comorbidities. We found that our DSN significantly overlaps 46.53% (p-value = 0.001) of the interactions in Hidalgo et al. (based on medical records) and up to 60.48% (p-value = 0.0076) with a more stringent approach.

Next, we compared our overlap with the ones derived from previous disease-disease networks based on other molecular data. Both, the microbiome and the miRNA networks yielded non-significant overlaps with the epidemiological network. The network derived from PPINs [3] presented significant yet small overlaps with the epidemiology (8.71% for the entire network and 18.52% over the diseases in our DSN), and the one generated by Sánchez-Valle et al. using microarray presents a significant overlap of 25% [4]. This implies, for the first time, that molecular -transcriptomic- similarities can capture and meaningfully explain a sizeable percentage of medically known comorbidities.

Additionally, since patient-specific patterns are observed at the epidemiological level, we introduced the concept of meta-patients as groups of patients from a given disease with a similar expression profile. Then, we calculated the similarities between meta-patients and diseases, in an attempt to identify subgroup-specific similarities potentially reflecting comorbidity relations. Our results show that some known disease associations that are difficult to reproduce at the disease level become evident when considering disease subtypes. In fact, we observe that some diseases present meta-patients that vary greatly on their disease links. This highlights the importance of studying comorbidities within a personalized medicine scope.

A current limitation of this study is the lack of information about the patient's relevant features (e.g., sex or age). Importantly, we provide a web application in which the networks at the disease and meta-patient level, as well as the molecular mechanisms that may explain their relationships, can be easily inspected. Furthermore, the automatization of the presented analysis allows for the future integration of the fast-growing and publicly available RNA-seq studies.

### *C. ACKNOWLEDGEMENTS*

This work was supported by a Ph.D. Fellowship and funded by the Spanish Ministry of Economics and Competitiveness. We recognize Jon Sánchez Valle and Rosalba Lepore as authors of the present study.

### *References*

[1] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: Implications for understanding health and health services," *Ann. Fam. Med.*, 2009, doi: 10.1370/afm.983.

[2] C. A. Hidalgo, N. Blumm, A. L. Barabási, and N. A. Christakis, "A Dynamic Network Approach for the Study of Human Phenotypes," *PLoS Comput. Biol.*, 2009, doi: 10.1371/journal.pcbi.1000353.

[3] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science (80-. ).*, 2015, doi: 10.1126/science.1257601.

[4] J. Sánchez-Valle *et al.*, "Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-16540-x.

## *Author biography*

**Beatriz Urda** was born in Almería, Spain, in 1993. She received the BSc in Biochemistry from the University of Granada in 2018 and the MSc in Bioinformatics for the Health Sciences from Pompeu Fabra University in 2020, Barcelona, Spain. She joined Alfonso Valencia's Computational Biology group as a master's student in 2019 and has recently started her PhD with a fellowship from the Spanish Ministry of Economics and Competitiveness.