Transcriptome Analysis of Differential Gene Expression in Disease

Jose Miguel Ramirez[#1], Raquel García-Pérez[#2], Marta Melé[#3]

#*Life Science Department, Barcelona Supercomputing Center (BSC), Eusebi Güell 1-3, Barcelona, Spain*

[1]jose.ramirez1@bsc.es, [2]raquel.garcia@bsc.es [3]marta.mele@bsc.es

Keywords— **transcriptomics, disease, demographic traits, diabetes**

EXTENDED ABSTRACT

Many diseases strongly impact the human transcriptome at the gene expression level [1]. However, previous work has focused on accessible tissues [2], and has not incorporated the effect of demographic traits, known risk factors for complex diseases [3]. Here, we leveraged the GTEx dataset to investigate the gene expression changes associated with different diseases. By studying the transcriptomes from an organismal perspective -across tissues and individuals- we can gain deeper insights into disease biology and help preventing complex diseases.

A. The GTEx dataset

The GTEx dataset [4] was originally intended to represent a healthy part of the population, but its latest version (v8) has been able to include diseased samples. This project has collected RNA-seq data from 49 human tissues in more than 838 individuals.

We studied 20 diseases in this dataset and controlled for their different effects across demographic groups, depending on age, ancestry, sex, and body mass index.

B. Gene Expression Changes with Disease

We performed differential gene expression analysis using linear models per tissue accounting for all available diseases and demographic traits. The largest number of differentially expressed genes was found in Hashimoto's thyroiditis in the thyroid, followed by pneumonia in the lung and atherosclerosis in the tibial artery.

Some of these diseases alter the transcriptome in specific directions, for instance, most of the genes affected by type 1 diabetes in the pancreas were downregulated, consistent with the known beta-cell loss in long-standing diabetes [5].

C. Contribution of Disease to Expression Variability

We then asked how much of the expression variability in the whole population can be explained by our diseases compared to demographic traits. In most cases, diseases explain a larger proportion of tissue expression variability than age, ancestry, sex, and body mass index combined, indicating profound alterations in the tissue transcriptome. This means that the thyroid of a healthy young female is more similar to that of a healthy old male than to a young female with Hashimoto's thyroiditis, as we expected.

D. Similar Effects Across Tissues in Diabetes

We benefited from the multi-tissue data to study how diseases affect 49 human tissues for the first time. Although most of the effects are tissue-specific, we find 79 genes dysregulated across tissues with type 1 diabetes and 309 with type 2 diabetes. Some of these genes have previously been associated with the disease, like OXCT1 [6], downregulated in 6 tissues with type 2 diabetes, including both subcutaneous and visceral adipose tissues.

E. Similar Effects Across Diabetes in Nerve

Surprisingly, the most affected tissue in type 1 and type 2 diabetes was the tibial nerve with 491 and 2955 differentially expressed genes, respectively, and neither the pancreas nor the adipose tissue, as we could have expected. This could be explained by diabetic neuropathy, a pathology originating due to continuous high blood sugar levels and known to occur independently with the type of diabetes [7]. We checked if the affected genes were the same in both diabetes and found 327 genes in which this was the case. Not only this number is higher than expected by chance, but all the genes were dysregulated in the same direction.

F. Machine Learning on Nerve Images

To validate the finding in the nerve at the transcriptomic level, we benefitted from the histology images available for the GTEx samples to explore whether the diabetic status could be distinguished at the image level or not.

We extracted Haralick features from the GTEx images and trained a support vector machine. With this simple model, we have already been able to predict diabetic status with an accuracy of 74% and an area under the ROC curve of 73%. This result validates our previous finding, suggesting that type 1 and type 2 diabetes have major impacts on the tibial nerve, a signal that can be detected both at the transcriptome and image level.

G. Additive Effects

Demographic traits often influence disease prevalence and progression [3], so next, we set out to investigate the interplay between demographic traits and diseases.

33.6% of the genes affected by a disease were also affected by a demographic trait, being age and sex the traits overlapping the most with diseases. The most prominent examples were both diabetes in the nerve, in which the overlap with age was significantly higher than expected by chance, but also in particular directionalities. In both cases, the genes upregulated with diabetes were upregulated in older individuals, and the ones downregulated with diabetes were also downregulated in older individuals. This suggests that diabetes impacts the tibial nerve similarly to biological ageing.

H. Conclusion

Some diseases affect more tissues than just the tissue of origin, highlighting the importance of multi-tissue studies to gain insight into disease mechanisms and advance towards prevention and cure.

This work also shows that diseases often alter genes with underlying expression differences in the human population, highlighting the importance of considering additive effects when studying inter-individual gene expression variation in human disease. Understanding the relationship between diseases and demographic traits is crucial to advance towards personalized medicine, as the second ones also tend to influence disease risk and progression.

I. Future Enhancements

In this work, we have explored the effect of disease on gene expression. However, diseases can also affect the transcriptome at the alternative splicing level. The GTEx dataset could also allow this type of analysis that would help us increase the knowledge about diseases and their interplay with demographic traits.

We could also validate the signal observed at the gene expression level with the histology images of other tissues. And, potentially, we could also correlate gene expression matrices with features extracted from the images to pinpoint genes to specific changes in tissue morphology.

J. ACKNOWLEDGEMENTS

References

[1] Casamassimi, A. et al. "Transcriptome Profiling in Human Diseases: New Advances and Perspectives." *International journal of molecular sciences* vol. 18,8 1652 (2017)

[2] Piasecka, B. et al. "Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges." *Proceedings of the National Academy of Sciences of the United States of America* vol. 115,3 E488-E497 (2018)

[3] Huypens P, et al. "Epigenetic germline inheritance of diet-induced obesity and insulin resistance." *Nature Genetics* vol. 48,5 497-9 (2016)

[4] Lonsdale, J., Thomas, J., Salvatore, M. *et al. "*The Genotype-Tissue Expression (GTEx) project." *Nature Genetics* 45 580-585 (2013)

[5] Richardon, Sarah J, and Alberto Pugliese. "100 YEARS OF INSULIN: Pancreas pathology in type 1 diabetes: an evolving story." *Journal of Endocrinology* 252,2 R41-R57 (2022)

[6] MacDonald, MJ. et al. "Decreased levels of metabolic enzymes in pancreatic islets of patients with type 2 diabetes." *Diabetologia* vol. 52,6 1087-91 (2009)

[7] Iqbal, Z. et al. "Diabetic Peripheral Neuropathy: Epidemiology, Diagnosis, and Pharmacotherapy." *Clinical Therapeutics* vol. 40,6 828-849 (2018)

Author biography

Jose Miguel Ramirez is a bioinformatician passionate about data science, machine learning, RNA and the role this molecule plays in disease. He studied a bachelor's degree in Bioinformatics by the UPF, UB, and UPC, and he did his bachelor's thesis on machine learning applied to Oxford Nanopore Technologies. He did a master's in Data Science by the UPC and studied abroad for an exchange at Aalto in Helsinki to focus on high throughput Bioinformatics. He did his master's thesis at the Transcriptomics and Functional Genomics Lab to work with transcriptomics and stayed to work for an extra year. Now, he is starting his PhD at the same lab where he is expanding the work started as a master student.