A Gene Set Enrichment Analysis of multiomic celiac disease data

Eugenio Del Prete[1,2,3], Angelo Facchiano[2] and Pietro Liò[3]

[1] Department of Sciences, University of Basilicata, Viale dell'Ateneo Lucano 10, 85100, Potenza (Italy)
[2] Institute of Food Sciences, National Research Council, Via Roma 64, 83100, Avellino (Italy)
[3] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD (UK)

Corresponding Author Email: eugenio.delprete@{isa.cnr.it; unibas.it}

Introduction

Celiac disease is a chronic condition, which can be described as inflammatory and autoimmune. Two factors are significant: the genetic and the environmental components. It occurs in predisposed individuals with genetic variants, which are related to an autoimmune response from the small intestine after the contact with gluten, a storage protein fraction found in different cereals, such as wheat, rye, barley, oat. Gliadins and glutenins are the main components of gluten, responsible for triggering of the autoimmune response after the crossing of the small intestine epithelium.
From a genetic point of view, celiac disease is strictly associated with genes related to HLA-DQ2 and HLA-DQ8 alleles, directly involved in the molecular mechanism of recognition of celiac disease-inducing gluten peptides. Nevertheless, there is a genetic risk due to an unknown number of non-HLA genes. Celiac disease is associated to many symptoms and conditions. A significant amount of individuals is asymptomatic at diagnosis. Furthermore, endoscopic and histological evaluation are the best ways of assessing small intestine healing, despite of theirs invasiveness and costs. The well-known treatment is a lifelong gluten-free diet, but it can be not totally effective for a high percentage of the patients [1].
The aim of this work is to approach the celiac disease complexity from a bioinformatics point of view. The idea is to analyse the state of the art from GEO online repository and revisit the works, by integrating gene expression data and Gene Ontology (GO) terms. Finally, the new evidence will give a direction to improve the knowledge on celiac disease.

Methods

One of the most important bioinformatics online repositories is the Gene Expression Omnibus (GEO) database. From the webpage *https://www.ncbi.nlm.nih.gov/gds/*, writing 'celiac disease' in the query field, it is possible to get 427 related results. In order to eliminate study with a low number of samples, the selected choice is in the range of 30-3000 samples (adding the filter from show additional filters), thus the number of results decreases to 18, from which only 6 are strictly related to the disease or not redundant. The selected studies are more robust in terms of download, data coherence and feasibility of analysis [2].
Gene Set Enrichment Analysis (GSEA) is a set of statistical methods to classify genes in groups, which are related to common biological function, chromosomal location or regulation. In a nutshell, genes from microarrays or NGS can be analysed by their differential expression among several conditions, selected as representative for a certain disease and correlated to GO terms, finding possible pathways linked to phenotypical changes, e.g. disease development. From this perspective, it is not important the single gene, because its behaviour can be masked, but the subset of genes taken into account [3].
The work is developed in R environment. The packages are downloaded by the online repository Bioconductor, open source and open development [4]: *GEOquery* for getting data from NCBI GEO; *limma* for data analysis, linear models and differential expression on microarray data; *genefilter* for some basic functions on filtering genes; *topGO* for testing GO terms and arranging topology in the GO graph.
The steps of the pipeline are: 1. the GEO data are selected by code, downloaded with the platform and matrices information, and converted in Expression Set class for the analysis; 2. the design model is decided by the user and, after a linear and a Bayesian filtering, the results are tabulated, sorted by logFC values; 3. the topGOdata class is created by selecting the GO domain, a filtering function to establish selected genes (out of genes universe) and the annotation for the mapping; 4. the

Fisher's exact test is performed in order to obtain a filtering on GO terms and the relationships with the selected genes; 5. two output are available: a GO graph with the hierarchy of the terms according to the selected genes, and a text file with the GO terms-selected genes correspondence. For completeness, the design model is celiac vs control or treated. About the selection of candidate genes, the choice is made on three parameters: p-value, adjusted p-value (Benjamini-Hochberg correction) and logFC.

Results

The studies are not standardized: three of them have a raw filter on microarray genes, based on unused or defected probes, whereas the second three have been uploaded after manipulation. The decision on results seems hard to take about searching for differential expressed genes. The scores (adjusted p-value, p-value and logFC) are strictly related to dataset dimension; furthermore, logFC is dependent from the kind of dataset. Moreover, the application of the Fisher's exact test seems to be biased for the selection of the GO terms, because of the dimension of the dataset and, probably, the performed manipulations. In these circumstances, the candidate genes subset is chosen with a trade-off among all the scores, thus the creation of a GO graph eludes the Fishers exact test, keeping its biological importance to define process clusters, but losing in statistical power. The most important genes and associated GO terms from the Biological Process (BP) domain are shown in Tab. 1.

Table 1: List of genes and related GO terms. The genes direct-related to celiac disease are in bold, the genes indirect-related to celiac disease are in italic (by means of secondary pathways or connection with other diseases). The most important five GO terms from GSEA are shown. In summary, the three datasets GSE72625, GSE61849a e GSE76168 are the most connected to the pathology.

Dataset	Genes	GO terms (BP)
GSE11501	**HLA-DRB1**, CAMK1D, **EP300**, FAM129A, SEMA4D, CBL, **COMMD1**, PSMC2	GO:0050732, GO:0014068, GO:0034976, GO:0031400, GO:0031399
GSE87629	NUSAP1, *CCNB2*, **CCNA2**, **UHRF1**, MELK, CSRP2, **ROCK1**, ASPM, PCLAF, SNCA, FBXO4	GO:0007067, GO:0000226, GO:0051726, GO:0031648, GO:0007049
GSE72625	**UBD**, **CYP3A4**, **HMGCS2**, **CXCL10**, **IFI27**, **STAT1**, **LCN2**,**HLA-A**, **SPINK4**, **GBP1**, **SLC19A1**, **GUCA2B**, **ITLN1**, *WARS*, **AQP10**	GO:0060337, GO:0060333, GO:0071357, GO:0034341, GO:0034340
GSE61849a	**PRM1**, **CTLA4**, **ICOS**, **FASLG**, **CCR4**	GO:0002520, GO:0048731, GO:0007275, GO:0006955, GO:0002376
GSE61849b	**PRM1**, **MMEL1**, **TREH**, **TNFSF18**, *CLK3*	GO:0006468, GO:0006915, GO:0012501, GO:0045785, GO:0008219
GSE76168	**TNFSF18**, **FASLG**, **CCR6**, **JAK2**	GO:0006139, GO:0006725, GO:0046483, GO:0034641, GO:0006807

A little framework on the biological processes involved in each study on celiac disease is suggested: GSE11501, peptidyl-tyrosine phosphorylation, phosphatidylinositol 3-kinase signaling, and response to endoplasmic reticulum stress; GSE87629, mitosis regulation, microtubule cytoskeleton organisation, and protein destabilization; GSE72625, signaling pathway and cellular response about interferon-gamma; GSE61849a, immune response and immune system development; GSE61849b, protein phosphorylation, apoptotic process, and regulation of cell adhesion; GSE76168, cytokine mediate signaling pathways.

References

1. M. Ostensson, et al. "A Possible Mechanism behind Autoimmune Disorders Discovered By Genome-Wide Linkage and Association Analysis in Celiac Disease". *PLoS ONE*, vol.8 (8), e70174, 2013.
2. S. E. Wilhite, T. Barrett. "Strategies to Explore Functional Genomics Data Sets in NCBIs GEO Database". *Methods Mol Biol*, vol.802, pp. 729-743, 2012.
3. A. Subramanian, et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *PNAS*, vol.102 (43), pp.15545-15550, 2005.
4. W. Huber, et al. "Orchestrating high-throughput genomic analysis with Bioconductor". *Nat Methods*, vol.12 (2), pp. 115-121, 2015.