Reusing microarray clinical data from a complex disease with bioinformatics tools

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

[1] Department of Sciences, University of Basilicata, Via dell'Ateneo Lucano 10, 83100 Potenza, Italy

[2] National Research Council, Institute of Food Science (CNR-ISA), via Roma 64, 85100 Avellino, Italy

[3] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD Cambridge, UK

Email of Corresponding author: eugenio.delprete@unibas.it; eugenio.delprete@isa.cnr.it

Abstract

Clinical bioinformatics, translational bioinformatics and personalised medicine are connected by the common task of analysing and integrating clinical data and results, in order to find important biomarkers related to pathologies and facilitate their prediction, diagnosis and treatment. New technologies provides the possibility to have more and more clinical data available in online databases. This data can be reused for studying complex disease from novel point of views. This work show how it is possible considering online microarray data from coeliac disease and some of its comorbidities, combining both the data and the results. The main goal is the extraction of common evidences among the selected pathologies, from genes to different kinds of functional annotation, showing which biological processes are more involved in these autoimmune disorders and quantifying the similarity between coeliac disease and its comorbidities. The pipeline of the work is developed in R language, and it is semi-automated. Methodologically, the advantage of this work is the possibility of performing the entire analysis starting from a different pathology; clinically, scientists can have the possibility of using data already published to highlight old and new evidences, with the possibility of improve the knowledge on a complex disease according to the availability of new microarray data.

Introduction

Clinical bioinformatics deals with the elaboration of clinical data by using bioinformatics methodologies. The data are a set of miscellaneous information related to the patients, collected and made available from physicians and researchers. The main goal is the extraction of latent knowledge about a disease that is hard to obtain with standard techniques of analysis. The underlying features can improve not only the strategy to cope with a selected pathology, but even provide some branches to other diseases connected to the primary [1]. Clinical bioinformatics mostly concerns the theoretical aspects about the integration of clinical data, by means of bioinformatics methodologies and tools, to understand biological mechanisms and design suitable therapies. Starting from this prospective, clinical bioinformatics should help physicians in dealing with omics data and support the researchers in reusing them to find new evidences. The design of the pipeline and the choice of the bioinformatics tools are, therefore, two of the main key points in clinical bioinformatics. Translational bioinformatics and personalised medicine are terms strictly related to clinical bioinformatics. Without entering into details, translation bioinformatics can be interpreted as synonymous of clinical bioinformatics, with particular emphasis on storage, analysis and interpretation of biomedical data, e.g. genomic data, in order to enhance all the fields of health management [2]. An intelligent application of bioinformatics on healthcare, essentially, reduces costs and improves outcomes, facilitating a predictive and preventive medicine [3]

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27398v1 | CC BY 4.0 Open Access | rec: 30 Nov 2018, publ: 30 Nov 2018

The availability of data from new high-throughput technologies and the increasing computational power provide the possibility of improve tools and methods that clean, aggregate, integrate and analyse multi-omics data. The idea is considering each omic field as a layer, searching for the principal interconnections inter/intra layers, in order to model biological system, and generate findings, such as clinical outcomes, useful for improving personalised medicine [4]. Several omics challenges cope with the comparison of two biological states, usually two main phenotypes, which need to be clustered by their features data, in order to understand underlying differences. Determining these states means revealing the causation of a system response, a very big challenge in bioinformatics and biostatistics analyses [5].

A complex disease is a phenotype caused by many individual gene events, with an important influence from the environment. Briefly, complex disease can be resumed with few points: a) caused by a combination of genetic, environmental and lifestyle factors; b) heritable, even if it has not simple patterns of inheritance; c) the insurgence or the transmission are difficult to predict; d) complicated way of treatment. Difficulties in complex diseases are in characterising genes and their interaction in a pathophysiology context [6]. The identification of important genes in complex diseases can provide precious evidences about disease pathogenesis and help in therapeutic treatment, indexing the primary cause of the disease and not only the surface symptoms. The discovery of genetic factors can be transformed in the use of biomarkers for patients clusterization and prognostic categorization. Moreover, if these genetic factors are present at the birth, it can be possible to take immediately action [7]. Coeliac disease is a pathology identified in the small bowel that can be described with five principal keywords: 1) autoimmune, the immune system has an abnormal reaction against a normal body part; 2) inflammatory, the immune system tries to block the inflammation triggering and repair the damage; 3) systemic, there is a high probability of co-occurrence (comorbidity) of other autoimmune disease localised in many parts of the body; 4) complex, as previously explained; 5) multicomponent, with an environmental (triggering) component, the gluten, and a genetic component, particular kinds of haplotypes in HLA-DQ cell surface receptor protein [8,9].

The main goal of the work is reusing coeliac disease online data together with some of its comorbidities online data, in order to strengthen known evidences and find new ones, with focus on which genes and ontology terms can be assumed as the most important biomarkers for the selected pathologies.

Methods

Microarray

Nowadays, microarray technology is still one of the most used in molecular biology, with the advantage of analyse more than one gene at time. In a nutshell, microarray is a technology which analyses the expression of thousands of genes, in a quantitative way. Each spot in a rectangular microarray contains pieces of DNA from a particular gene. Samples to analyse consist of mRNA copies that can bind to DNA pieces, i.e. the gene from the transcribed mRNA is highly expressed. The quantification is possible by using fluorescence or radioactive tags, which highlights the bind of multiple copies [10]. The Gene Expression Omnibus (GEO) is one of the most important public online repository for microarray, next generation sequencing (NGS) and other high-throughput data, with a prevalence of gene expression data [11]. The cooperation among microarray data, statistical methods and R language is reported in many scientific work, such as in [12].

Gene Set Enrichment Analysis and Semantic Similarity

Gene Set Enrichment Analysis (GSEA) is a set of statistical methods to establish if a group of differential expressed genes (DEG), among two or more phenotypes, are effectively describing differences in expression. In other words, GSEA provides the statistical significance on the genes expression between two conditions or states, usually sick individuals and healthy control, and the identification of gene sets previously considered unrelated [13]. A list of advices is presented: getting and cleaning data process is necessary before performing GSEA; select a suitable annotation (e.g., pathways, PPI network or GO), coherent with biological expectations; integrate different annotations for a robust knowledge in gene function, checking if they are upgraded; filter genes in a proper way (small or large gene set); control the mapping between the original source (e.g. SNPs, miRNAs) and the genes; choose a suitable GSEA method and control the statistical evaluations; provide details on gene sets and statistical algorithms; report the biological context to facilitate the interpretation of gene set annotation; select good visual representation.

A complex disease can be thought as system with a number of redundant components: if a subset of these components are defective, disease state occurs. The genes are the components, and a subset of them is more important than single genes in analysing a disease. One perspective of similarity concept among diseases is to share underlying molecular process. Relation between symptoms and shared genes are reported in scientific literature, but many diseases with different symptoms can have gene sets in common. Functional-based methods of semantic similarity generally associate a set of biomolecules (e.g. genes) to each disease and compare the sets for quantifying the similarity [14]. Wang's method determines the semantic similarity of two terms (genes, Gene Ontology or Disease Ontology) based on both the locations of these terms in a hierarchical graph and their relations with their ancestor terms [15]. This method is a hybrid approach, because each edge is weighted according to reported relationships and the semantic contribution of all the common ancestors are summed for each term (normalised with total semantic contribution of ancestors) [16].

Overview on the pipeline

The pipeline proposed in Figure 1 has been developed in R language, with the libraries (packages) downloaded from Bioconductor online repository [17].

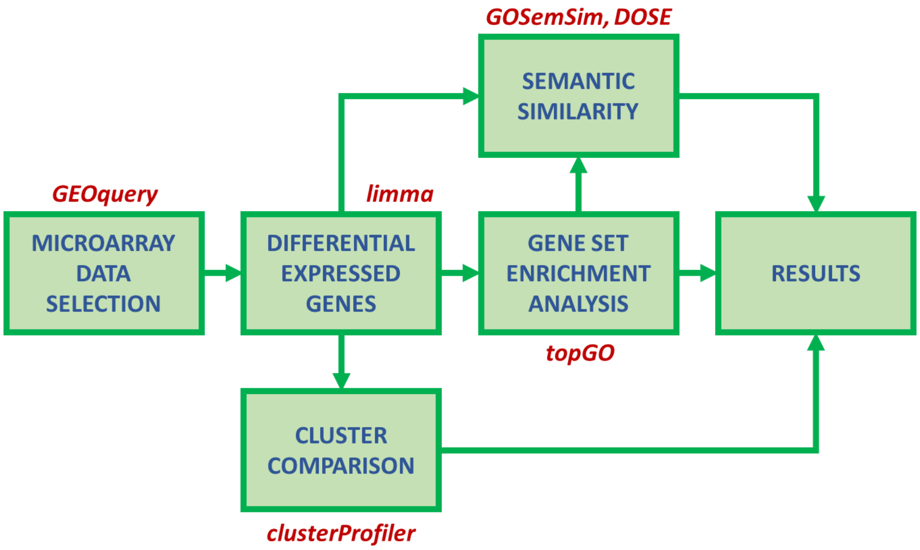

Figure 1. Pipeline of the work on a complex disease. The R packages downloaded from Bioconductor online repository and used for the integration and analysis of data are reported in red for each step.

Briefly, the selection of datasets is up to the user, but they can be downloaded inside the pipeline or recall from a local folder. Selection criteria should be discussed: in this work, microarray data with at least 8 samples (sick, healthy and treated) have been chosen. The second step is the extraction of the DEG from the comparison of the two states, and the third step is the application of GSEA to these genes. The fourth step is the calculation of the similarity among the datasets, by using DEG, related GO terms and Disease Ontology terms (previously provided inside the script). Moreover, the DEG are used for an enrichment with functional annotations from KEGG database [18]. Finally, the results are collected in form of table, GO terms tree, similarity matrix and bubble plot.

Results

Selection of datasets

The microarray datasets selected from GEO on coeliac disease are 7, divided into 10 different comparisons. The studies are of array type, simple or from reverse transcriptase polymerase chain reaction (RT-PCR), with samples extracted from peripheral blood or intestinal biopsy. The two condition considered as control are healthy individuals or gluten free diet (GFD) treated individuals. The microarray datasets selected from GEO on the comorbidities are 14, divided into 17 different comparisons. The selected comorbidities are: alopecia areata, arteritis, autoimmune thyroid disease, dermatomyositis, primary biliary cirrhosis, peripheral neuropathy, rheumatoid arthritis, and vitiligo. The studies are of array type, with samples extracted from skin biopsy, temporal artery biopsy, peripheral blood, skeletal muscle biopsy, and miRNA-mRNA profile. In order to perform the semantic similarity not only with genes and Gene Ontology terms, but also among all the entire functional hierarchies of the pathologies, Disease Ontology terms are extracted from the website *http://disease-ontology.org/*, with reference to the abovementioned diseases.

Common evidences among coeliac disease and comorbidities

The comparison of the differential expressed genes between the coeliac disease datasets and the autoimmune diseases shows that a group of genes is in common. A cluster network has been generated on the list of genes taking into account co-expression, consolidated pathways, co-localization, shared protein domain, predicted and physical interactions, using the online tool GeneMania [19]. The most important clusters are related to the pathways, involved in the immune system, chemokine and cytokine signalling, to pathologies, such as rheumatoid arthritis and influenza A, or to genes, such as CTLA4 and IL12. The comparison between Gene Ontology terms from coeliac disease list and autoimmune disorders list highlights the presence of terms in common, related to immune system, response to virus, and cytokine pathways (with focus on type I interferon).

An enrichment with KEGG pathways applied on candidate genes highlights that some functional annotations have been already claimed, that is the involvement of natural kill cell and T cell, the signalling pathways from chemokines and cytokines, or the relationships with autoimmune thyroid disease and influenza A. Further functional annotations are: hepatitis C, herpes simplex infection, measles, NOD-like receptors signalling pathway, and prolactin signalling pathway. It is not straightforward to associate coeliac disease with hepatitis C [20], even if there can be a connection by means of HLA-DQ2, a secondary pathway with Sjogren's syndrome, or an amino-acid sequence homologous to a gliadin epitope. In [21], herpes simplex has a co-occurence in a patient with CD, but the correlation is not claimed in term of causality: probably the malabsorption of the nutrients is the trigger for a sort of immunodeficiency. A direct relationship between CD and measles is lacking in scientific literature, but pathway analysis and PPI network have highlighted common functional annotations between measles and rheumatoid arthritis, that is cytokine-cytokine receptor interaction, Jak-STAT signalling, T cell receptor signalling, and cell adhesion molecules [22], and all of them are involved in CD. Nucleotide-binding and oligomerization domain (NOD)-like receptors (NLRs) are related to infections and immunity, by recognizing pathogen-associated and damage-associated molecular patterns: some SNPs and polymorphisms in NLRs subfamilies have a direct connection with CD [23]. Finally, prolactin acts not only as a hormone, but also as a cytokine, and prolactin levels are positively correlated with CD, indeed a GFD reduces these levels [24].

Finally, the semantic similarity applied on genes, Gene Ontology terms and Disease Ontology terms, after the selection of an appropriate threshold, highlights that coeliac disease is mainly connected to dermatomyositis and vitiligo, with further connections with primary biliary cirrhosis and alopecia areata, dictated by specific datasets and fold change thresholds.

Discussion and conclusion

The reuse of microarray datasets, with further bioinformatics and statistics tools, such as GSEA and semantic similarity, is important in finding new evidences by the integration and analysis of both data and results. In this case, functional annotations from biological process, pathways and disease have been extracted from the comparison between coeliac disease and selected comorbidities. In summary, the results concern immune system, inflammatory pathways and virus-related pathologies: all the three topics are coherent with the selected autoimmune diseases. Furthermore, the semantic similarity, applied on three different kinds of term (genes, Gene Ontology and Disease Ontology), shows how dermatomyositis and vitiligo are the closest pathologies to coeliac disease. The methodology and details on the results of this work are widely explained in [25].

It is important to underline that this work is data-driven, thus more datasets are taken into account, more robust are the results. Therefore, the implementation of other datasets, from coeliac disease and its comorbidities, is a suggested addition at the beginning of the pipeline. Furthermore, the pipeline is entirely in R language and can be modularised in a bigger workflow, in order to completely automate the process. For the clinicians and scientists, the availability of new omics data will offer the opportunity to apply the described bioinformatics pipeline for a more complete view of molecular mechanisms of other complex disease, with the ultimate aim of highlighting biomarkers from different levels (gene, ontology and disease) useful for the improvement of prediction, diagnosis and treatment of pathologies.

References

[1] Chang PL., et al. (2005). Clinical bioinformatics. Chang Gung Medical Journal, 28(4), 201-211.

[2] Tenenbaum JD. (2016). Translational Bioinformatics: Past, Present, and Future. Genomics, proteomics & bioinformatics, 14(1), 31-41.

[3] Sethi P., et al. (2009). Translational bioinformatics and healthcare informatics: computational and ethical challenges. Perspectives in health information management, 6(Fall):1h.

[4] Zhu B., et al. (2017). Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. Scientific Reports, 7:16954.

[5] Lorenzon R., et al. (2018). Clinical and multi-omics cross-phenotyping of patients with autoimmune and autoinflammatory diseases: the observational TRANSIMMUNOM protocol. BMJ Open, 8:e021037.

[6] Hofker MH., et al (2014). The genome revolution and its role in understanding complex diseases. Biochimica et Biophysica Acta, 1842(10), 1889-1895.

[7] Lowe WL., et al. (2015). Genomic approaches for understanding the genetics of complex disease. Genome Research, 25(10), 1432-141.

[8] Camarca A., et al. (2009). Intestinal T cell responses to gluten peptides are largely heterogeneous: implications for a peptide-based therapy in celiac disease. Journal of immunology, 182(7), 4158-4166.

[9] Gianfrani C. et al. (2007). Transamidation of wheat flour inhibits the response to gliadin of intestinal T cells in celiac disease. Gastroenterology, 133(3), 780-789.

[10] Marzancola MG., et al. (2016). DNA Microarray-Based Diagnostics. Methods in molecular biology, 1368, 161-178.

[11] Clough E., et al. (2016). The Gene Expression Omnibus Database. Methods in molecular biology, 1418, 93-110.

[12] Mutarelli M., et al. (2008). Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. BMC Bioinformatics, 9(Suppl 2):S12.

[13] Clark NR., et al. (2011). Introduction to Statistical Methods for Analyzing Large Data Sets: Gene-Set Enrichment Analysis. Science Signaling, 4(190):tr4.

[14] Pesaranghader A., et al. (2014) Gene Functional Similarity Analysis by Definition-based Semantic Similarity Measurement of GO Terms. In: Sokolova M., van Beek P. (eds) Advances in Artificial Intelligence. AI 2014. Lecture Notes in Computer Science, 8436. Springer, Cham.

[15] Wang JZ., et al. (2008). An Efficient Method to Measure the Semantic Similarity of Ontologies. In: Wu S., Yang L.T., Xu T.L. (eds) Advances in Grid and Pervasive Computing. GPC 2008. Lecture Notes in Computer Science, 5036, Springer, Berlin, Heidelberg.

[16] He W., et al. (2011) A Hybrid Approach for Measuring Semantic Similarity between Ontologies Based on WordNet. In: Xiong H., Lee W.B. (eds) Knowledge Science, Engineering and Management. KSEM 2011. Lecture Notes in Computer Science, 7091, Springer, Berlin, Heidelberg.

[17] Ramos M., et al. (2017). Software for the Integration of Multiomics Experiments in Bioconductor, Cancer Research, 77, e39-e42.

[18] Kanehisa M., et al. (2015). KEGG as a reference resource for gene and protein annotation. Nucleic acids research, 44(D1), D457-D462.

[19] Zuberi K., et al. (2013). GeneMANIA prediction server 2013 update. Nucleic Acids Research, 41(Web Server issue), W115-W122.

[20] Casella G., et al. (2016). Association between celiac disease and chronic hepatitis C. Gastroenterology and Hepatology from Bed to Bench, 9(3), 153–157.

[21] Chen A., et al. (2016). Celiac Crisis Associated with Herpes Simplex Virus Esophagitis. American College Of Gastroenterology Case Reports Journal, 3(4), e159.

[22] Liu G., et al. (2013). Measles Contributes to Rheumatoid Arthritis: Evidence from Pathway and Network Analyses of Genome-Wide Association Studies. PLoS One, 8(10), 1–9.

[23] Kim YK., et al. (2016). NOD-like receptors in infection, immunity, and diseases. Yonsei Medical Journal, 57(1), 5-4.

[24] Borba VV., et al. (2018). Prolactin and Autoimmunity. Frontiers in Immunology, 9:73, 2018.

[25] Del Prete E., et al. (2018). Bioinformatics methodologies for coeliac disease and its comorbidities. Briefings in Bioinformatics, bby109, https://doi.org/10.1093/bib/bby109.