A data science approach for exploring differential expression profiles of genes in transcriptomic studies-Application to the understanding of ageing in obese and lean rats in the FIGHT-HF project

Emmanuel Bresso, Claire Lacomblez, Anne Pizard, Patrick Rossignol, Faiez Zannad, Malika Smaïl-Tabbone, Marie-Dominique Devignes

To cite this version:

Emmanuel Bresso, Claire Lacomblez, Anne Pizard, Patrick Rossignol, Faiez Zannad, et al.. A data science approach for exploring differential expression profiles of genes in transcriptomic studies-Application to the understanding of ageing in obese and lean rats in the FIGHT-HF project. JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Jul 2018, Marseille, France. hal-01928421

HAL Id: hal-01928421
https://hal.inria.fr/hal-01928421

Submitted on 20 Nov 2018

A data science approach for exploring differential expression profiles of genes in transcriptomic studies – Application to the understanding of ageing in obese and lean rats in the FIGHT-HF project.

Emmanuel Bresso, Claire Lacomblez, Anne Pizard, Patrick Rossignol, Faiez Zannad, Malika Smaïl-Tabbone and Marie-Dominique Devignes

EB, CL, MS-T, M-DD : Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

AP, PR, FZ : CIC-P, Centre Hospitalier Régional Universitaire de Nancy, F-54500 Vandoeuvre les Nancy.

M-DD : BIOBASE, Service de Support à la Recherche, Centre Hospitalier Régional Universitaire de Nancy, F-54500 Vandoeuvre les Nancy.

FIGHT-HF project: Fight Heart-Failure, RHU project 2015-2020, ANR-15-RHU-004,

Keywords: differential expression profiles; heterogeneous graphs; complex network; transcriptomics

Introduction

Transcriptomic studies are known to produce huge amounts of information about differentially expressed genes in various situations. The expression levels measured for several thousands of genes in contrasting pairs of situations (different tissues or organs, different physiological state, different age, etc.) allow to calculate fold-change ratios and false discovery rates. Today, data science should be able to derive valuable knowledge units from all these data, by extracting relevant lists of differentially expressed genes and interpreting them. However, in many cases, only a small proportion of all transcriptomic results is finally exploited.

Here we revisit the concept of differential expression profile (DEP) and we propose a combined database - network approach to extract relevant lists of differentially expressed genes based on DEP sharing and to interpret these lists using heterogeneous graph settings before visualisation.

Methods

Differential Expression Profile definition.

A transcriptomic study is usually performed on various biological samples derived from tissues or organs under certain conditions. In most cases, expression profiles are defined in a non-supervised manner by clustering the genes on the basis of their expression values across all situations in a study. In other cases, the study involves contrasting situations that lead to the calculation of fold-change (FC) ratio with false discovery rates (FDR) for each gene. Such a setting allows to define differential expression profiles (DEPs).

Let a transcriptomic study be represented by a set of contrastsC_i, $i \in \{1...\text{n}\}$, defined as ordered pairs of situations, and by the FC ratios and FDR values obtained for each gene $g_k$in each contrast C_i. Four discrete statuses and two modalities have been defined to represent the behaviour of a given gene in a given contrast. The two modalities correspond to *stringent* (FC<-2 or FC>2 and FDR <0.05) or *not stringent* (any FC and FDR <0.05) definitions of differential expression. Under *stringent* modality, status $Str1$ means differential expression, status $Str2$ and $Str3$ are for up (FC>2) and down (FC<-2) regulation respectively and status

$Str4$means no differential expression. Under *non stringent* modality, similar statuses can be defined and are prefixed by $NStr$.

Formally, a DEP is described by a definition and an extension. The definition of a DEP can be represented as a set of pairs $(C_i, status)$, $i \in \{1...\mathrm{n}\}$ and $status \in \{Str1...Str4, NStr1...NStr4\}$ describing the differential expression status of a gene across a set of contrasts C_i. The extension of a DEP is the set of genes $g_k$that match the definition. It should be noted that a DEP can also be defined from the differential expression statuses of a given gene across a set of contrasts and then used to select other genes matching the DEP definition.

DEP Relational Data Model.

A relational database named DEPdb was built to store and query transcriptomic results w.r.t. DEPs. The data model covers the contrast definitions between pairs of situations, the transcripts involved and their corresponding genes along with human orthologs when necessary, the FC ratios and associated FDR values, calculated for each transcript and each contrast.

A DEP query interface was implemented on DEPdb to retrieve lists of genes matching a given DEP definition or retrieve a DEP definition for a given gene.

Network-based interpretation of lists of genes.

Various tools exist already to help biologists interpret a list of genes in the light of pathways and interaction networks available in various integrated and curated public sources. Hence, complex graphs are produced showing the multiple interactions existing between the genes of interest and their interactants. Network science should help end-users to interpret such complex networks.

We present here an analysis strategy based on the notion of « heterogeneous graph » in which different types of nodes are interconnected by various types of relationships. Such heterogeneous graph is the basis of the EdgeBox, a « graph knowledge box » constructed by the EdgeLeap company, using the Neo4J graph database system and available public resources [1]. The EdgeBox currently contains seven types of nodes: protein/gene, disease, pathway, drug, metabolite, gene ontology term and miRNA, and fourteen types of relationships between these nodes (five monopartite, such as protein-protein interactions, and nine bipartite ones). The 2017 version for the FIGHT-HF project concerns human proteins/genes and counts about 211,000 nodes and almost 22 million of relationships.

To interpret a DEP extension, *i.e.* a list of genes corresponding to a given DEP definition, in the light of the EdgeBox, a first approach consists in querying the EdgeBox for pathways that interconnect at least two genes/proteins of the list. The query result can be completed with the genes/proteins that interact with at least two genes/proteins of the list. When the gene list is short (less than 10 genes), the resulting graph can be interpreted manually and reveals at a glance which pathways or interactions possibly explain why genes share a given DEP definition. This network-query tool has been implemented onto DEPdb connected to the graph knowledge box. It displays the heterogeneous graph using the Cytoscape program for further analyses.

When the gene list grows, the resulting graph becomes extremely complex and impossible to interpret manually. We propose to filter the nodes for enhancing user interpretation. Enrichment analyses are first carried out on the list of genes to select a small number of nodes corresponding to the top10 of significantly ($p<0.001$) enriched pathways and biological process GO terms. In parallel, it reveals useful to select subgroups of gene/protein nodes of interest based on their neighbourhood in the EdgeBox. Such selection may be driven by user knowledge. After node reduction, a collection of smaller heterogeneous subgraphs is produced that usually become tractable.

Our combined database-network approach can function iteratively. Starting from a first DEP definition involving a given set of contrasts, the database will return a list of genes that is subsequently displayed as a heterogeneous network. In some cases, the user will select from this network a gene of interest and will query the database to retrieve its DEP across a different set of contrasts. The database will then return on demand all other genes matching this second DEP definition. The new returned extension can then in turn be analysed as a heterogeneous graph, etc.

Results and discussion: Differentially expressed genes upon ageing in heart and kidney of obese and lean SHHF rats

Description of the study

A transcriptomic study aimed at characterizing simultaneously metabolic syndrome and cardiac, vascular and renal phenotypes in ageing lean and obese SHHF (Spontaneously Hypertensive Heart Failure) rats has been described previously [2]. Obesity is induced in SHHF rats by homozygous inactivation of the leptin receptor gene. Rats have been monitored during 11 months (from the age of 1.5 to 12.5 months). In the frame of the FIGHT-HF project, transcriptomic results of this study are newly investigated using our coupled database/network approach.

Definition of two DEPs and extraction of corresponding gene lists

Transcriptomic results have been stored in the DEPdb database as described above. To illustrate our approach we focus on a simple comparison between ageing in obese versus lean rats, in both heart and kidney tissues. We therefore use four different contrasts from our database: heart samples from « old versus young » lean and obese rats (contrasts $C_{13}$and C_{14}respectively in DEPdb), and kidney samples from « old versus young » lean and obese rats (contrasts $C_{21}$and C_{22}respectively in DEPdb).

We query DEPdb consecutively with two DEP definitions, namely:

$DEP_{lean} = \{(C_{13}, Str1), (C_{14}, Str4), (C_{21}, Str1), (C_{22}, Str4)\}$,

and $DEP_{obese} = \{(C_{13}, Str4), (C_{14}, Str1), (C_{21}, Str4), (C_{22}, Str1)\}$. In other words, DEP_{lean} will retrieve from DEPdb all genes differentially expressed (stringent defintion) upon aging in lean ($(C_{13}, Str1)$and$(C_{21}, Str1)$) but not in obese ($(C_{14}, Str4)$and $(C_{22}, Str4)$) rats, in heart and kidney samples respectively, and DEP_{obese}will retrieve from DEPdb all genes differentially expressed upon aging in obese ($(C_{14}, Str1)$and $(C_{22}, Str1)$) but not in lean ($(C_{13}, Str4)$and$(C_{21}, Str4)$) rats, in heart and kidney samples respectively. Note that the expression statuses in all other contrasts are not considered here. We retrieved 7 and 55 genes for DEP_{lean}andDEP_{obese}definitions, respectively. We also determined with a third appropriate DEP definition that 11 genes are differentially expressed upon aging in both lean and obese rats in heart and kidney samples.

Interpretation with heterogeneous graphs

We subsequently analysed the two extensions of the $DEP_{lean}$and DEP_{obese}profiles in the light of our network knowledge box. Please note that these gene lists first need to be converted to their human orthologs. The DEP_{lean}extension includes 7 genes and is short enough to be directly analyzed in DEPdb connected with the EdgeBox. We observe that all gene nodes are interconnected together through pathways and GO terms related to cell cycle and cell proliferation, metabolism of proteins and DNA, apoptosis, rhythmic processes, signal transduction and immune system.

Because of its complexity, theDEP_{obese}extension of 55 genes underwent the filtering process described above before heterogeneous graph retrieval from the EdgeBox. Interestingly the resulting network clearly reveals two modules of interconnected genes, with 4 genes forming bridges between the two modules. One of these modules involves 22 genes interconnected with the neutrophil degranulation pathway and several GO terms related to inflammation, whereas the other one involves 15 genes interconnected with two pathways : extracellular matrix organization and elastic fiber formation, and with GO terms related to either extracellular matrix organisation or epoxygenase P450 pathway that is involved in anti-inflammatory response.

In summary the comparison of $DEP_{lean}$and DEP_{obese}extensions reveals that a group of 7 genes involved in particular in cell cycle and cell proliferation is dysregulated in heart and kidney upon ageing in lean rats but not in obese rats. On the contrary, genes that are dysregulated in heart and kidney upon ageing in obese rats but not in lean rats fall into two modules, one related to inflammatory response and the other one related to extracellular matrix organisation. This analysis brings new precise molecular support to the general statement that ageing can have quite different outcomes in lean or obese individuals. The four proteins bridging these two modules (CD44, INTBD2, ANXA2 and CP2E1) can be further investigated to better understand the complementarity between the two groups of biological processes dysregulated upon ageing in the heart and kidney of obese rats. Interestingly, context analysis of the two lists of genes in a more classical framework such as CPDB returned similar results [3].

Conclusion and Perspectives

A formal definition of DEPs has been proposed for transcriptomic studies involving contrasting situations. A coupled database/network approach has been designed and implemented to explore any desired DEP from a study, by extracting the corresponding list of genes and interpreting them using a network knowledge box and heterogeneous graphs. The ageing case study shows that this approach provides useful support for manual interpretation of gene lists and can lead to new hypotheses generation. Obviously, a large number of different DEPs remain to be tested in the same way from the SHHF study.

We are currently searching to design and implement automatic filters to reduce heterogeneous graph complexity in a knowledge-based manner for assisting user interpretation of heterogeneous networks.

Reference

[1] Pinet F, CuvelliezvM , Kelder T, Amouyel P, Radonjic M, Bauters C. (2017)
Integrative network analysis reveals time-dependent molecular events underlying left ventricular remodeling in post-myocardial infarction patients.
Biochimica et Biophysica Acta (BBA) – Molecular Basis of Disease 1863(6): 1445-53.

[2] Youcef G, Olivier A, Nicot N, Muller A, Deng C, Labat C, Fay R, Rodriguez-Guéant RM, Leroy C, Jaisser F, Zannad F, Lacolley P, Vallar L, Pizard A. (2016)
Preventive and chronic mineralocorticoid receptor antagonism is highly beneficial in obese SHHF rats.
Br J Pharmacol. 173(11):1805-19.

[3] Kamburov A, Wierling C, Lehrach H, Herwig R. (2009)
ConsensusPathDB--a database for integrating human functional interaction networks.
Nucleic Acids Research 37(Database issue):D623-D628.