Computational Biology and Chemistry 59 (2015) 67–77

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Transcriptional master regulator analysis in breast cancer genetic networks

Hugo Tovar[a], Rodrigo García-Herrera[a], Jesús Espinal-Enríquez[a,b], Enrique Hernández-Lemus[a,b,*]

[a] Computational Genomics Department, National Institute of Genomic Medicine (INMEGEN), Mexico
[b] Center for Complexity Sciences, National Autonomous University of Mexico (UNAM), Mexico

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 17 August 2015
Accepted 17 August 2015
Available online 22 August 2015

Keywords:
Transcriptional master regulators
Breast cancer
Gene regulatory networks
Systems biology

ABSTRACT

Gene regulatory networks account for the delicate mechanisms that control gene expression. Under certain circumstances, gene regulatory programs may give rise to amplification cascades. Such transcriptional cascades are events in which activation of key-responsive transcription factors called *master regulators* trigger a series of gene expression events. The action of transcriptional master regulators is then important for the establishment of certain programs like cell development and differentiation. However, such cascades have also been related with the onset and maintenance of cancer phenotypes. Here we present a systematic implementation of a series of algorithms aimed at the inference of a gene regulatory network and analysis of transcriptional master regulators in the context of primary breast cancer cells. Such studies were performed in a highly curated database of 880 microarray gene expression experiments on biopsy-captured tissue corresponding to primary breast cancer and healthy controls. Biological function and biochemical pathway enrichment analyses were also performed to study the role that the processes controlled – at the transcriptional level – by such master regulators may have in relation to primary breast cancer. We found that transcription factors such as AGTR2, ZNF132, TFDP3 and others are master regulators in this gene regulatory network. Sets of genes controlled by these regulators are involved in processes that are well-known hallmarks of cancer. This kind of analyses may help to understand the most upstream events in the development of phenotypes, in particular, those regarding cancer biology.

1. Introduction

Cancer is a pathway-disease (Hanahan and Weinberg, 2000). The main hallmarks of cancer are associated to the action of pathways related to cell proliferation, apoptosis evasion, cell differentiation and in general, to the dysregulation of cell cycle and the alteration of DNA-repairing processes (Hanahan and Weinberg, 2000, 2011). The phenotype of a cell is determined by the activity of a large number of genes and proteins (Basso et al., 2005). Hence, transcriptional regulation lies at the heart of many of the intricate molecular relationships driving the activity of biological pathways (Emmert-Streib et al., 2014).

It has been observed that a number of large scale transcriptional cascades behind such complex cellular processes are actually triggered by the action of a relatively small number of transcription factor molecules that have been called Transcriptional Master Regulators (TMRs) (Han et al., 2004; Sun-Kin Chan and Kyba, 2013; Mullen et al., 2011). It has been argued that these genes control the entire transcriptional regulatory program for specific cellular phenotypes (in eukaryotic cells; Han et al., 2004; Basso et al., 2005; Affara et al., 2013). However, TMRs are also able to act on general cellular processes at the same time (Hinnebusch and Natarajan, 2002; Medvedovic et al., 2011; Affara et al., 2013). A proper understanding of the organization of these TMR-mediated highly-regulated events is thus crucial to elucidate normal cell physiology as well as complex pathological phenotypes (Basso et al., 2005).

Given the complex mechanisms underlying transcriptional regulations on eukaryotes, the identification of TMRs is often based on the (inferred or observed) relationship among them and their cascade of RNA targets in gene regulatory networks (Hernández-Lemus and Siqueiros-García, 2013). Being a primary upstream event in the cell regulatory program, dysregulation of TMRs may have a high impact on cancer-related phenotypes, since under genetic instability conditions, uncontrolled synthesis of these

* Corresponding author at: Computational Genomics Department, National Institute of Genomic Medicine (INMEGEN), Mexico.
E-mail address: ehernandez@inmegen.gob.mx (E. Hernández-Lemus).

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.007
1476-9271/

molecules could originate the activation/amplification of several transcriptional cascades (Basso et al., 2005; Baca-Lopez et al., 2014; Baca-López et al., 2012).

A TMR is a transcription factor (TF) that is expressed at the early onset of the development of a particular phenotype or cell type (Sun-Kin Chan and Kyba, 2013). It also participates in the specifications of such a phenotype by regulating multiple downstream genes, either directly or by means of genetic cascades. Transcription factors are hence key cellular components that control gene expression: their activities may determine how cells function and respond to the environment (Vaquerizas et al., 2009).

Transcription factors may act in two opposite directions: either activating or repressing transcriptional activity of their targets. Based on the initial estimations of the whole human genome sequence, it was calculated that the transcriptional machinery could be composed of 200 to 300 genes and there could exist between 2000 to 3000 specific union sites for transcription factors (Lander et al., 2001; Venter et al., 2001). In Vaquerizas et al. (2009) it is stated that in the http://amigo.geneontology.orgGene Ontology database 1052 TFs were defined and just 6% (62 cases) of them had experimental corroboration. Six years later, the same database recognized 1846 TFs and 14% (260) of them had experimental evidence. This is indicative of the fast progress on documenting the transcription mechanisms, but this also points to the overwhelming complexity of the mechanisms of genomic control.

Implementation of computational methods to identify and analyze TMRs is relevant in the context of breast cancer, particularly at its earliest stages. We have probabilistically inferred the gene regulatory network associated with this phenotype, then a computational analysis has uncovered its active TMRs in the context of primary breast cancer. In our study we have considered such an analysis, as well as the resulting TMR-related phenomena in the context of transcriptional regulatory programs. We also discuss here some of the implications of our results in breast cancer biology. The article is structured as follows: Section 2 presents an overview of the materials and methods used in this work. This includes both the experimental datasets used, the network inference strategy and the molecular signature analysis, as well as the algorithm for the discovery of transcriptional master regulators. Section 3 presents some of the main results of the application of this pipeline in primary breast cancer microarray gene expression data. Finally, Section 4 presents some conclusions mainly related with the advantages of implementing a method such as MARINa (Lefebvre et al., 2010) in order to unveil some aspects of regulatory control that may lie behind the establishment of tumor phenotypes.

2. Materials and methods

2.1. Experimental datasets

For the analysis presented here, we obtained 880 microarray expression profiles from several experimental datasets that are available on the Gene Expression Omnibus site (http://www.ncbi.nlm.nih.gov/geo/GEO) (Edgar et al., 2002). All experiments were performed by using total mRNA on the microarray platform Affymetrix HGU133A (GPL96), which consists of 18,400 transcripts and variants, including 14,500 well-characterized human genes (Liu et al., 2003). From the total 880 samples, 819 correspond to primary breast cancer tissue, whereas the remaining 61 samples correspond to healthy breast tissue. In the case of experiments that included any kind of treatment or cell modification, we only used the unaltered samples (see Table 1).

A second dataset for comparing the results was obtained from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). We used 597 mRNA samples of invasive breast cancer, of which 534 correspond to tumor samples and the other 63 were non-tumor. All data used for this analysis correspond to level 3, which means they are already normalized.

2.2. Batch effect control

Batch effect is one of the most recurrent factors of error during data analysis from microarrays (Grass, 2009). Chen et al. (2011) tested six different algorithms to eliminate batch effect and found that the best results were obtained by using the empirical bayesian method known as ComBat (Combating Batch Effects When Combining Batches of Gene Expression Microarray Data) (Johnson et al., 2007). However, since seven out of the ten datasets corresponded to tumor tissue exclusively (i.e. there are no control samples), and the three remaining datasets had only healthy tissues, there is no intersection between those datasets. According to Leek et al. (2010), treatments and batches are completely confounded. Since currently there is no method to estimate the batch effect under these conditions (Leek et al., 2010), ComBat (Johnson et al., 2007) cannot perform the normalization of the whole dataset. Taking into account that ComBat does not eliminate batch effect with the conditions of our dataset, we decided to partially solve this issue as follows: After preprocessing all arrays with frma (McCall et al., 2010), and using summarization with robust weighted average with no background correction, we split the datasets into cases/controls, and then applied ComBat to both datasets separately. After that, we re-joined the two resulting datasets and re-normalized them together with the cyclic loess algorithm (Ballman et al., 2004), in such way that both conditions belong now to the *same dynamic range*

We needed to have a measure of the batch effect within the samples so that we could remove the corresponding bias as accurately as possible. To this end we resort to Principal Variance Component Analysis (PVCA) that is an algorithm that combines the advantages of the principal component analysis (reduction of dimensionality) with the components of the analysis of variance (Grass, 2009). Once the batch effect is reduced separately, a PVCA analysis corroborated that such a batch effect almost disappeared and the treatment effect was important enough. (Fig. 1).

Given our design conditions, it was not possible to eliminate batch effect completely. Since batch effect in such mixed experimental designs is an important topic of current research in computational genomics, we can envisage a scenario in which the present work may be revisited and some of its conclusions may need to be revised. In the meantime, the method described above aimed at reducing and estimating batch effects may be considered a first approximation for the purposes of the work presented here.

For the TCGA dataset, since we analyzed data level 3 samples, normalization had already been performed by the TCGA site. For batch effect correction, the data were computed using ComBat (Johnson et al., 2007), Median Polish and ANOVA.

2.3. Network inference

Gene regulatory networks (GRN) are models that describe the relationship between genes under certain given conditions. Network inference can be defined as the process of identifying gene interactions from experimental data by performing a computational analysis (Bansal et al., 2007). To infer the breast cancer transcription factor regulatory network (interactome), we proceeded as follows. First, we generated a network for every known human TF in the primary breast cancer gene expression dataset by using the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Basso et al., 2005; Margolin et al., 2006). ARACNE is a computational algorithm widely used to identify statistical relationships among genes, by calculating the mutual information (*MI*) between gene pairs from microarray expression data (Basso et al., 2005; Margolin et al., 2004).