Computational and Structural Biotechnology Journal 15 (2017) 425–432

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Meta-analysis of Liver and Heart Transcriptomic Data for Functional Annotation Transfer in Mammalian Orthologs

Pía Francesca Loren Reyes, Tom Michoel, Anagha Joshi*, Guillaume Devailly*

The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form 10 August 2017
Accepted 11 August 2017
Available online 26 August 2017

Keywords:
Gene function
Transcriptomics
Liver
Heart
Orthologs
Paralogs
Co-expression
Gene networks

ABSTRACT

Functional annotation transfer across multi-gene family orthologs can lead to functional misannotations. We hypothesised that co-expression network will help predict functional orthologs amongst complex homologous gene families. To explore the use of transcriptomic data available in public domain to identify functionally equivalent ones from all predicted orthologs, we collected genome wide expression data in mouse and rat liver from over 1500 experiments with varied treatments. We used a hyper-graph clustering method to identify clusters of orthologous genes co-expressed in both mouse and rat. We validated these clusters by analysing expression profiles in each species separately, and demonstrating a high overlap. We then focused on genes in 18 homology groups with one-to-many or many-to-many relationships between two species, to discriminate between functionally equivalent and non-equivalent orthologs. Finally, we further applied our method by collecting heart transcriptomic data (over 1400 experiments) in rat and mouse to validate the method in an independent tissue.

1. Introduction

Annotation of gene function is a crucial step to understand the DNA sequencing data currently generated at an unprecedented rate. The lack of functional annotation forms a major bottleneck in analyses across diverse fields, including de novo genome sequencing [1], Genome Wide Association Studies (GWAS) in model and non-model organisms [2], and metagenomics [3]. An experimental validation of each gene is impractical to this end as it demands high financial and time cost. It is estimated that only 1% of proteins have experimental functional annotations [4]. Bioinformatic approaches therefore provide an attractive alternative [5]. The most widely used and successful gene annotation strategy has been the annotation transfer between homologous genes. Automated annotation pipelines from sequence alone are widely used, including GOtcha [6] and BlastGO [7]. They allow fast annotation of thousands of genes for newly sequenced genomes [8]. This approach can be used within a species, where gene families (paralogs), might share common functions, or across species, where known function(s) of a gene in one species are used to infer functions of the homologous gene(s) in another species.

Despite being widely used, fast computational annotation comes at a cost of misannotation, which is present at high levels (over 10%) and is believed to be increasing [9] due to misannotation transfer. The most common misannotation is over-annotation, where a gene is assigned a specific but incorrect function [10]. This is partly because one of the major challenges in functional annotation transfer across species is that the orthology relationships are not always one-to-one. Specifically, a single gene in one species can be homologous to multiple paralogs in another (one-to-many homologies), after gene duplication or gene loss event(s). After a gene duplication, the two paralogs can have redundant functions, and thus should share similar functional annotations, or one copy might diverge (lose functionality, or gain new functionalities, or change cellular localisation or tissue specificity), and thus paralogs should have different functional annotations despite their homology. Similarly, multigene families (with many-to-many homologies) are highly prone to over-annotation errors.

Protein structure information can act as source for functional distinction within multigene family proteins [4]. Protein-protein interaction networks have also been successfully used to identify functional orthologs [11]; two orthologs interacting with the same proteins in each species are likely to share similar functions.

* Corresponding authors.
E-mail addresses: Anagha.Joshi@roslin.ed.ac.uk (A. Joshi), Guillaume.Devailly@roslin.ed.ac.uk (G. Devailly).

http://dx.doi.org/10.1016/j.csbj.2017.08.002

Similar strategy has been applied to biochemical pathway information [12]. Co-expression gene networks have also been used in this context [13–15], as they offer two main advantages over protein-protein interactions and biochemical pathways. First, they can be inferred from transcriptomic datasets, which are more abundant than protein-protein interaction datasets. Second, they allow functional annotation of the various classes of RNA genes. We have previously shown that multi-species information improves gene network reconstruction [16].

In order to further explore the potential of co-expressed gene networks to identify functional equivalents in complex homologous families, we collected transcriptomic data from mouse and rat liver samples. To minimise technical variation, we collected datasets generated using a single microarray platform in each species, resulting into 920 experiments in mouse and 620 experiments in rat. We firstly identified clusters of co-expressed genes using hierarchical clustering and found biologically relevant clusters. We applied an hyper-graph clustering method, SCHype [17] to simultaneously cluster co-expressed orthologous genes between species. We then focussed on 18 complex (one-to-many or many-to-many) homology groups, where at least one member in mouse and in rat where present in similar co-regulated gene clusters providing an independent source of evidence for shared functionality amongst orthologous genes in complex homologous families. We successfully applied the same method on heart transcriptomic data from mouse and rat, and investigated functional relevance of 11 other orthologous groups. Our results show the potential of this method to use co-expression as an independent measure to evaluate shared functionality amongst orthologs and limit over-zealous annotation transfers.

2. Methods

2.1. Data Collection and Normalisation

Microarray data for liver and heart samples in mouse and rat were collected from GEO, where data for mouse was generated using Affymetrix Mouse Genome 430 2.0 Array, and data for rat was generated using Affymetrix Rat Genome 230 2.0 Array as they were the platforms with a large number of experiments available for each species. Liver experiments came from 62 (mouse) and 28 (rat) independent studies or GEO series. Heart experiments came from 20 (mouse) and 19 (rat) independent studies or GEO series. The GEO accession numbers for individual studies are provided in Supplementary Table 1. Processed data was not directly comparable between studies, as different studies used different normalisation methods, leading to different distribution of values (Supplementary Fig. 1, A and B, Supplementary Fig. 3, A and B). As some datasets had a trimmed lower quartile for reduction in noise by limiting the variability of lowly expressed genes, we applied lower quartile trimming on all datasets (Supplementary Fig. 1, C and D, Supplementary Fig. 3, C and D). Specifically, we set the expression value of all probes belonging to the lower quartile to the value of the 25 percentile. We then applied quantile normalisation resulting into a uniform distribution of values for each experiment. To facilitate the comparison between mouse and rat data, we used liver mouse data as a target for quantile normalisation of heart mouse data and liver and heart rat data, using preprocessCore functions `normalize.quantiles.determine.target` and `normalize.quantiles.use.target` [18]. Liver mouse data was selected as the target because it contained more experiments than the liver rat dataset. Thus, after our normalisation steps, the distribution of values was identical for each experiment in both species.

2.2. Data Clustering

We selected genes with variable expression across experiments by selecting probes with a standard deviation greater than one across experiments. As shown in Fig. 1, such probes included genes of low as well as high expression levels, and largely excluded probes showing very low expression in all experiments. Microarray data being already log-transformed, log fold change over the average values were obtained by subtracting the mean expression of each probes.

Hierarchical clustering was done on the log fold change matrices using R functions `dist` ad `hclust` with default parameters (euclidean distance, complete linkage). Dendrogram branches were reordered using the function `order.optimal` from the cba package [19]. Both rows (probes) and columns (experiments) were clustered using this approach.

Gene homology information was retrieved from the Homologen database [20], and probe orthology information was obtained using the R package annotationTools [21]. Due to one-to-many homologs, rat probes and mouse probes intersections resulted into slightly different numbers for each species. Average of the two numbers was used to obtain Jaccard indexes. Jaccard index significance was obtained using the hypergeometric test, and P-values were corrected for multiple testing using Bonferroni correction.

SCHype takes as input a list of conserved interactions which was generated as follows. First Spearman correlation coefficient between each pair of probes was obtained independently for both Mouse and Rat expression data. Pairs of probes with a correlation coefficient ≥ 0.5 were selected. Then if orthologs of two connected probes were connected in the other species, they were kept as an SCHype input. SCHype was run using default parameters. In liver, SCHype identified 132 clusters of homologous genes co-expressed both in mouse and in rat, which included 825 nodes in mouse and 778 nodes in rat. SCHype allows probes to be included in multiple clusters. The different number of probes in mouse and rat is due to the presence of one-to-many and many-to-many orthologs, as well as the presence of gene measured by multiple probes on the array.

2.3. Gene Ontology Analysis

Gene ontology analysis was performed using PantherDB [22], using as a control gene set the genes analysed by the microarray, or only the variable gene sets previously defined.

2.4. Scripts and Data Availability

R scripts used for this analysis are available in a Github repository https://github.com/gdevailly/liver_mouse_rat. Normalised expression matrices, fold change matrices, as well as probe clusters (hierarchical clustering and SCHype clustering) are available through

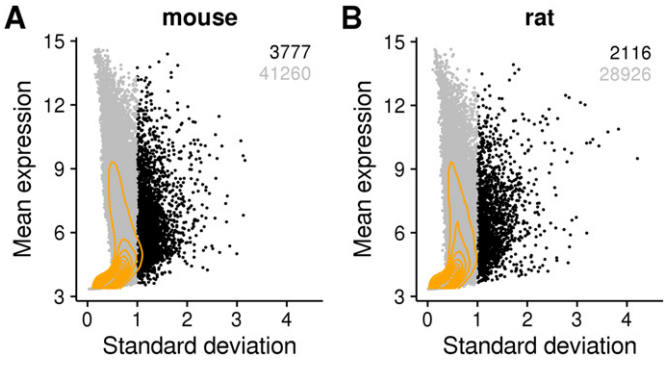

Fig. 1. Identification of variable probes in mouse (A) and rat (B) datasets. Each dot represents a single probe. X axis: standard deviation across experiments. Y-axis: mean expression values across experiments (in arbitrary units). In black the probes with a standard deviation ≥ 1, in grey the probes with a standard deviation < 1. Orange lines: 2D kernel density. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)