Computational Biology and Chemistry 63 (2016) 73–82

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Characterizing mutation–expression network relationships in multiple cancers

Shila Ghazanfar[a,b,*], Jean Yee Hwa Yang[a]

[a] School of Mathematics and Statistics at the University of Sydney, F07, The University of Sydney, NSW 2006, Australia
[b] Data61, CSIRO, Locked Bag 17, North Ryde, NSW 2113, Australia

ARTICLE INFO

Article history:
Received 18 January 2016
Accepted 1 February 2016
Available online 12 February 2016

Keywords:
Mutation
Gene expression
Protein–protein interaction network
Pan-cancer

ABSTRACT

Background: Data made available through large cancer consortia like The Cancer Genome Atlas make for a rich source of information to be studied across and between cancers. In recent years, network approaches have been applied to such data in uncovering the complex interrelationships between mutational and expression profiles, but lack direct testing for expression changes via mutation. In this pan-cancer study we analyze mutation and gene expression information in an integrative manner by considering the networks generated by testing for differences in expression in direct association with specific mutations. We relate our findings among the 19 cancers examined to identify commonalities and differences as well as their characteristics.
Results: Using somatic mutation and gene expression information across 19 cancers, we generated mutation–expression networks per cancer. On evaluation we found that our generated networks were significantly enriched for known cancer-related genes, such as skin cutaneous melanoma ($p < 0.01$ using Network of Cancer Genes 4.0). Our framework identified that while different cancers contained commonly mutated genes, there was little concordance between associated gene expression changes among cancers. Comparison between cancers showed a greater overlap of network nodes for cancers with higher overall non-silent mutation load, compared to those with a lower overall non-silent mutation load.
Conclusions: This study offers a framework that explores network information through co-analysis of somatic mutations and gene expression profiles. Our pan-cancer application of this approach suggests that while mutations are frequently common among cancer types, the impact they have on the surrounding networks via gene expression changes varies. Despite this finding, there are some cancers for which mutation-associated network behaviour appears to be similar: suggesting a potential framework for uncovering related cancers for which similar therapeutic strategies may be applicable. Our framework for understanding relationships among cancers has been integrated into an interactive R Shiny application, PAn Cancer Mutation Expression Networks (PACMEN), containing dynamic and static network visualization of the mutation–expression networks. PACMEN also features tools for further examination of network topology characteristics among cancers.

1. Background

It is known among the field of cancer biology that there is a genetic basis behind the growth of cancers. This is manifested fundamentally through mutations in the DNA within tumour cells, which can impact gene transcript levels potentially leading to phenotypic changes in the cell via protein translation or other mechanisms. Thus, it is important to characterize the relationships between mutations in the DNA and the way in which the genes behave, in terms of their expression within the cell.

Cancer research efforts, especially integrative studies using various -omics and external information sources have been enriched by the growing availability of large datasets through consortia

Abbreviations: CGC, Cancer Gene Census; DE, differential expression; FDR, false discovery rate; GEO, Gene Expression Omnibus; HGNC, HUGO Gene Nomenclature Committee; KEGG, Kyoto Encyclopedia of Genes and Genomes; NCG, Network of Cancer Genes; NMF, non-negative matrix factorization; PPI, Protein–protein interaction; UPPI, union protein–protein interaction; TCGA, The Cancer Genome Atlas; WGCNA, Weighted Gene Correlation Network Analysis; WGS, whole genome sequencing; WES, whole exome sequencing.

* Corresponding author at: School of Mathematics and Statistics at the University of Sydney, F07, The University of Sydney, NSW 2006, Australia.
E-mail address: s.ghazanfar@maths.usyd.edu.au (S. Ghazanfar).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.009
1476-9271/© 2016 Elsevier Ltd. All rights reserved.

such as The Cancer Genome Atlas (TCGA). The availability of this information gives unprecedented opportunity to explore these relationships both within and between cancers. At the same time, there is also a multitude of information available in biological databases pertaining to protein, gene interaction and regulatory networks. From manually curated pathway databases, e.g. KEGG (Ogata et al., 1999) and MetaCore^{TM} (Ekins et al., 2007) to primary (Rolland et al., 2014; Keshava Prasad et al., 2009; Breitkreutz et al., 2007) and secondary (Turner et al., 2010) repositories of protein–protein interaction measured through yeast 2 hybrid assays; there is a wealth of knowledge available for both interrogation and integration with computational biology methods.

However, among this host of data comes the challenge of extracting meaningful information. Network analysis has been previously employed in the context of gene expression (Taylor et al., 2009), identifying ‘party’ and ‘date’ hubs along a protein–protein interaction (PPI) network, determined by differing levels of correlation coexpression in genes among samples between two conditions. In addition, methods have aimed to build networks in silico based on gene–gene coexpression (Fuller et al., 2007), circumventing the use of network information based on prior knowledge. Methods involving network approaches have uncovered insightful regulatory modules in disease networks (Ideker et al., 2002; Ma et al., 2011; Dittrich et al., 2008) using curated networks and available data to identify subnetworks worth pursuing for further biological analysis (Barabási et al., 2011). In sum, network methods have aimed to shed light on interesting regions of the biological space and have hinted towards meaningful explanations of the phenomena observed.

In previous cancer research, mutation information has been coupled with network information to identify network modules within a particular cancer (Ciriello et al., 2012; Leiserson et al., 2013). These tools make use of external networks such as PPI networks in order to restrict the testing space and as a frame to then identify subnetworks of interest. An overview of these methods and others is available by Creixell et al. (2015).

In recent years, pan-cancer network research has come to the forefront with the use of large-scale matrix summarization techniques such as non-negative matrix factorization (NMF) and clustering to identify commonalities across different cancers (Jia et al., 2014; Hofree et al., 2013). Leiserson et al. (2015) conducted a pan-cancer analysis by considering the mutations over multiple cancers and implementing the HotNet2 algorithm to identify significantly mutated subnetworks.

More recently methods have aimed to incorporate both gene expression and mutation information in a network setting to identify genes and networks of interest. For example, in a direct interrogation of a dozen genes harbouring somatic mutations, Gerstung et al. (2015) examined the mutation and expression network characteristics in myelodysplastic syndromes, leading to improved accuracy in the prediction of patient outcome. Similar studies have also been conducted in certain disease settings (Bashashati et al., 2013). On a larger scale, more suited to high-throughput studies, DriverNet (Bashashati et al., 2012) identifies a set of driver genes by building bipartite graphs between mutated genes and genes with outlying gene expression values. Similar algorithms make use of a combination of mutation, gene expression and external network information (Hou and Ma, 2014; Jia and Zhao, 2014; Paull et al., 2013). However most of these algorithms do not analyse the data in a direct manner, and do not enable incorporation of sample-specific information such as prognosis among other clinical variables. To this end, we introduce a framework that performs direct testing, accounting for such information via covariates introduced in linear models. We build directed networks resulting from differing gene expression levels by partitioning via mutation, and build bipartite graphs between mutated genes and pathways for which pathways are significantly affected according to mutation status in samples. This enables us to explore the potential mutational basis behind gene expression changes, on a single gene and pathway basis.

2. Materials and methods

2.1. Data acquisition

2.1.1. Mutation

Mutation information was downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) between 14 November 2013 and 9 July 2015, listed in Supplementary Table S1. For each cancer, non-silent mutations were identified from the full list of mutations, using the ‘Variant_Classification’ parameter in those datasets. For each cancer dataset analyzed, a binary non-silent mutation incidence matrix M was formed by setting

$$M_{ij} = 1_{\text{at least 1 non-silent mutation in gene } i \text{ and sample } j}$$

for gene $i = 1, 2, \ldots, I$ and sample $j = 1, 2, \ldots, J$, where 1 is the indicator function.

2.1.2. Gene expression

Gene expression and clinical information was downloaded directly onto R using the R package `AnnotationHub` (Morgan et al.) obtaining data from Gene Expression Omnibus (GEO) submission ID GSE62944. The gene expression data contained raw RNA-Seq read counts obtained using the package Rsubread (Liao et al., 2013). We converted the RNA-Seq counts to log2-CPM (counts per million) via the `voom` function in the `limma` package. For each cancer dataset analyzed, a continuous gene expression matrix Y was formed where Y_{ik} indicates the expression level in sample $i = 1, 2, \ldots, I$ and gene $k = 1, 2, \ldots, K$.

2.2. Data processing

Sets of identified somatic mutations and gene expression data originated from TCGA as described. A sample was retained in a cancer cohort if both mutation and gene expression information was available. A total number of 4443 tumour samples were analyzed from 19 different cancers, listed in Table 1. Sample sizes ranged from 66 to 665, with a median of 207 samples. To enable downstream differential expression analysis, genes with non-silent mutations for fewer than 3 samples were removed, resulting in between 26 and 18,190 genes with mutations, with a median of 4353 genes. Lowly expressed genes were removed, i.e. those with fewer than 20 mapped reads for at least 50% of the samples. This resulted in numbers of genes with gene expression information ranging from 13,660 to 16,410, with a median of 15,670 genes.

2.3. Protein–protein interaction (PPI) networks

A union PPI network, denoted UPPI, was built by constructing the union of the five PPI networks listed in Table 2, with 68,832 edges shared among 12,237 nodes. This network effectively limits the search space in testing differential expression with respect to mutation, allowing statistically significant and potentially more biologically relevant relationships to be observed. It is entirely feasible that genes' expression may be affected by mutation that are multiple steps away in the network, thus we also considered a larger search space named UPPI2, defined by drawing edges between nodes that share are least one interacting partner.

We compared UPPI and UPPI2 with an existing consolidation of protein–protein interaction networks, namely HIPPIE (Schaefer et al., 2012), using the high-confidence cut-off of 0.68 as previously