Exploring the Importance of Cancer Pathways by Meta-Analysis of Differential Protein Expression Networks in Three Different Cancers

Sinjini Sikdar, Somnath Datta, Susmita Datta[1]

Department of Bioinformatics and Biostatistics, University of Louisville, KY 40202, USA

1 INTRODUCTION

Landscape of most cancers involve twelve important pathways ("target pathways") that regulate three core cellular processes "cell fate", "cell survival" and "genome maintenance"; the "driver" genes, which are responsible for the formation of tumors, function through these signaling pathways [1]. We undertake a novel investigation of the roles of these pathways using a differential network analysis of the protein expression datasets on three cancers (Head and Neck Squamous Cell Carcinoma, Lung Adenocarcinoma and Kidney Renal Clear Cell Carcinoma). These datasets were available to us from International Cancer Genomic Consortium (ICGC) as part of the CAMDA 2015 challenge data. We pursue a meta-analysis of protein expressions to investigate whether each of these target pathways plays a significant role in these three cancers in the sense that the proteins associated in these pathways interact differently between two clinical groups ("progression" or "complete remission") of patients. From our analysis of the protein expression data, overall, RAS and PI3K signaling pathways appear to play the most significant roles in these three cancers. This analysis suggests that these two signaling pathways should be investigated further for their roles in cancers. It is interesting to note that these two main pathways are related to "cell survival" function.

2 DATASETS

We have analyzed the preprocessed challenge datasets for CAMDA 2015 provided by the International Cancer Genomic Consortium (ICGC). For our study, we have considered the protein expression and the clinical profiles of the patients for three cancers, Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), and Kidney Renal Clear Cell Carcinoma (KIRC). A set of 132 proteins is present in the protein expression profiles of each of the three cancers; the patient sample sizes of HNSC, LUAD and KIRC were 212, 237 and 454 patients, respectively. The clinical profile of each of the cancer type represents the disease status (progression or complete remission) of each patient. In summary, we have two groups of patients for each cancer type and the set of recorded protein expression values of 132 proteins on each of them.

[1] to whom correspondence should be addressed (susmita.datta@louisville.edu)

3 METHODOLOGY

3.1 Pathway analysis: From a recent study [1], it has been found that there are 140 "driver" genes/proteins which can promote the formation of tumors if affected by intragenic mutations. These "driver" genes can be classified into twelve signaling pathways which are: TGF − β, MAPK, STAT, PI3K, RAS, Cell Cycle/Apoptosis, NOTCH, HH, APC, Chromatin modification, Transcriptional regulation and DNA damage control. Among these, TGF − β, MAPK, STAT, PI3K, RAS and Cell Cycle/Apoptosis regulate "cell survival"; NOTCH, HH, APC, Chromatin modification and Transcriptional regulation regulate "cell fate"; while the DNA damage control signaling pathway regulates "genome maintenance". We refer to these twelve signaling pathways as "target pathways".

We separately analyze the protein profiles of the three cancer types using "GO" clustering tool [2, 3], and group the proteins according to their biological pathways. Out of the pathways obtained, we only considered the proteins included in the "target pathways" for our analysis.

3.2 Differential network analysis: In order to identify whether the network structures of the "target pathways" have changed from the complete remission group to the progression group, we performed differential network analysis [4] using the R package *dna* [5]. This differential network analysis for each pathway is conducted based on connectivity scores between the proteins in these target pathways. Initially, to get an idea about the network structures in each of the two groups, graphical networks are constructed by connecting each pair of proteins for which the connectivity scores exceed a threshold. The difference in connectivity between the two groups (progression versus complete remission) is computed mathematically, using the following statistic:

$$\Delta(\mathcal{F}) = \frac{1}{k(k-1)} \sum_{p \neq p' \in \mathcal{F}} \left| s_{pp'}^{pr} - s_{pp'}^{cr} \right|, \qquad (1)$$

where $\mathcal{F}$ denotes the set of proteins present in a "target pathway" and k denotes the number of proteins in $\mathcal{F}$. Here $s_{pp'}^{pr}$ and $s_{pp'}^{cr}$ are the connectivity scores between the proteins p and p' in the progression and complete remission groups, respectively. For our analysis, the connectivity score of a protein pair in a network is taken to be the Pearson's correlation coefficients of the expression values of the two proteins in the corresponding sample data. A permutation test is then carried out using the test statistic $\Delta(\mathcal{F})$ and the corresponding observed level of significance (p-value) is obtained.

In addition to testing the overall pathway significance, we also test whether the connectivity of each single protein has changed between the two groups (progression versus complete remission) using the following statistic:

$$d(p) = \frac{1}{f-1} \sum_{p' \in \mathcal{G},\ p' \neq p} \left| s^{pr}_{pp'} - s^{cr}_{pp'} \right|, \qquad (2)$$

where $\mathcal{G}$ denotes the set of all proteins and f is the number of proteins in $\mathcal{G}$. Once again, a permutation test is carried out for each protein using the test statistic $d(p)$ and the corresponding p-value is obtained.

3.3 Rank Aggregation: The p-values, obtained using the test statistic given in (1), are used to obtain ranked lists of the "target pathways" for each cancer type. Here, ranking is done in such a way that the "target pathway" with the lowest p-value gets rank 1, the next one gets rank 2 and so on. Since, these ranked lists vary according to the cancer type; we need to aggregate them in a meaningful way to get an overall ranked list which would then rank the pathways by their global order of importance. In other words, this overall ranked list may provide us with the most important "target pathways" in all the three cancers. The R package RankAggreg [6], which is based on Cross-entropy Monte Carlo algorithm [7], is used to get this overall ranked list.

For our second analysis at the individual protein level, the p-values obtained using the test statistic given in (2), are used to rank the set of 132 individual proteins. An overall ranked list of these proteins is obtained using the R package RankAggreg [6].

4 RESULTS

We find representation of five out of twelve "target pathways" in our sample of 132 proteins; they are the PI3K signaling pathway, Cell Cycle, Apoptosis, RAS signaling pathway and MAPK signaling pathway. Based on our differential network analysis [4, 5] between the two groups of patients (progression vs complete remission) using the test statistic given in (1), with Pearson's correlation coefficients as scores and absolute distance measure carried out for each of the 3 cancer types, we have the following findings: the RAS signaling pathway is highly significant (p-value = 0.026) and MAPK signaling pathway is marginally significant (p-value = 0.082) in HNSC; for LUAD, PI3K signaling pathway is highly significant (p-value = 0.013). Table 1 shows the overall ordering of the 5 "target pathways" for the three cancers along with the rank aggregated list. Thus overall, the RAS signaling pathway appears to be most important followed by the PI3K signaling pathway, based on our meta-analysis of the available data on three cancers.

Table 1: Target pathways ordered by statistical significance (p-values) for each cancer type along with the overall ordering by rank aggregation.

Cancer Type	Pathway Ordering by p-values
HNSC	R, M, P, A, C
LUAD	P, C, A, R, M
KIRC	R, A, M, P, C
Overall	R, P, M, A, C

R: RAS Signaling pathway
M: MAPK Signaling pathway
P: PI3K Signaling pathway
A: Apoptosis
C: Cell Cycle

A graphical representation of the network structure of the proteins in the two groups of patients for RAS signaling pathway in HNSC is shown in Figure 1. In this figure, two proteins are connected if the connectivity score between them is significantly large. Different colors and shades in the figure represent positive or negative correlations and the thickness of the lines represents the strength of the associations. A visual inspection reveals some obvious differences in the network connectivity between the two groups of patients. Notably, GAB2, MAPK1, MET, and BAD show noticeably different activities in the two networks. The corresponding genes are known oncogenes; e.g., GAB2 – melanoma, MAPK1 – multiple cancers, MET - papillary carcinoma, BAD - pancreatic cancer, prostate cancer.

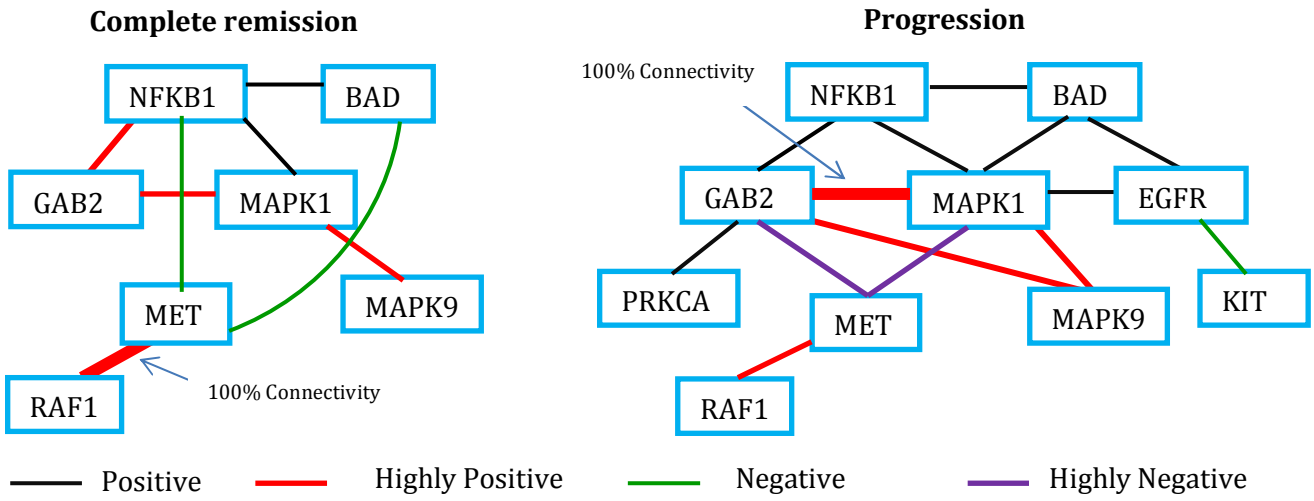

Figure 1: Network structure for RAS signaling pathway in Head and Neck Squamous Cell Carcinoma (HNSC)

Our analysis of individual proteins using the test statistic (2) produces Figure 2. The pie charts represent the proportions of top fifty differentially connected proteins for each of these pathways in the three cancers and in the overall aggregated list of proteins. Once again, PI3K and RAS take the top two most important spots in terms of differential network connectivity.

5 DISCUSSION

It is known that for most cancers with solid tumors the genes in the above mentioned "target pathways" display somatic mutations and change their protein products [1]. Here in this purely quantitative analysis of the existing protein expression data of three different cancers also reveals the significant alteration of the proteins in PI3K and RAS pathways. It is interesting to know that PI3K is a regulatory subunit, which binds to cell-surface receptors and to the RAS protein. Genes and proteins in PI3K and RAS have been investigated as therapeutic targets for many cancers ([8], [9] etc). Our findings are consistent with this and suggest that continued future efforts be made in this direction.

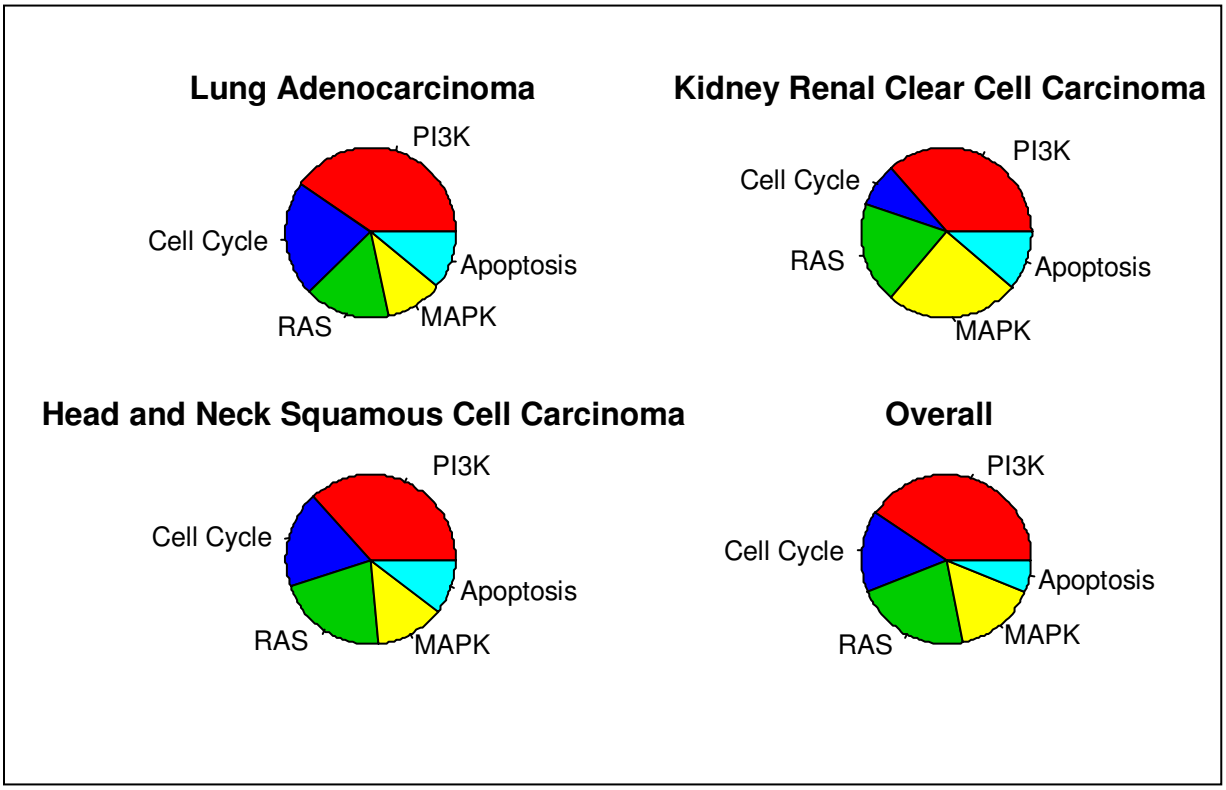

Figure 2: Relative contributions of the 5 "target pathways" in each of the three cancers separately as well as for all the three cancers combined.

References

[1] Vogelstein B, Papadopoulos N, Velculescu VE et al. Cancer Genome Landscapes. Science 2013; 339 (6127): 1546-58.
[2] Reimand J, Kull M, Peterson H et al. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucl. Acids Res. 2007; 35(2): W193-W200.
[3] Reimand J, Arak T, Vilo J. g: Profiler—a web server for functional interpretation of gene lists (2011 update). Nucl. Acids Res. 2011; 39(2): W307-W315.
[4] Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 2010; 11(1): 95.
[5] Gill R, Datta S, Datta S. dna: An R package for differential network analysis. Bioinformation 2014; 10(4): 233–34.
[6] Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. Bioinformatics 2007; 23(13): 1607-15.
[7] Rubinstein R. The cross-entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability 1999; 1: 127-90.
[8] Knight ZA, Shokat KM. Chemically targeting the PI3K family. Biochem Soc Trans.2007;35:245-249.
[9] Gysin, S., Salt, M., Young, A and McCormick, F. Therapeutic Strategies for Targeting Ras Proteins. Genes Cancer. 2011; 2(3): 359–372.