Directional association test reveals high-quality putative cancer driver biomarkers including noncoding RNAs

Hua Zhong$^{1[0000-0003-1962-2603]}$ and Mingzhou Song$^{1,2[0000-0002-6883-6547]}$

[1] Department of Computer Science
[2] Molecular Biology Graduate Program
New Mexico State University, Las Cruces, NM 88003, USA
huazhong@nmsu.edu joemsong@nmsu.edu

Extended Abstract

Statistical methods to identify cancer genes are often either biased due to the choice of models or insensitive to directional causal gene-to-cancer relationships. To overcome such issues, a recent functional chi-square test (FunChisq) uses model-free functional dependency to detect directional, nonlinear, non-monotonic, and functional relationships. The outstanding performance of FunChisq at HPN-DREAM Breast Cancer Network Inference Challenges [2] supports its practicality in causal inference.

Meanwhile, the Functional Annotation of Mammalian Genome 5 (FANTOM5) project [4] surveyed the expression at over 200,000 transcription start sites (TSSs) in nearly all human tissue types, primary cells, and cell lines of many cancer types. The data cover TSSs originated from both coding and noncoding genes. The unique advantage of FunChisq motivated us to study gene-to-cancer relationships using the FANTOM5 data.

We first evaluated the performance of FunChisq and six other methods on FANTOM5 data in distinguishing the most transcribed TSSs of 719 curated cancer genes [1] from the same number of randomly picked non-cancer TSSs. The six other methods include Pearson's chi-square test [6], Wilcoxon test [9], *t*-test [7], logistic regression [3], DESeq2 [5] and edgeR [8]. DESeq2 and edgeR were tested on read count data, while the other methods on transcript abundance data. We repeated the same evaluation on 100 different sets of randomly selected non-cancer TSSs. The performance of FunChisq is markedly better than all six other methods, giving the largest areas under the receiver operating characteristic (ROC) (Figure 1a) and precision-recall (PR) (Figure 1b) curves.

FunChisq reveals non-monotonic patterns, to which typical differential analysis method such as *t*-test is not sensitive. We observed strong non-monotonic directional association from the abundance of many TSSs to the cancer status, such as the TSS of known cancer gene *BRAF* (Figure 1c).

We further applied FunChisq on unannotated TSSs in FANTOM5, and predicted 1108 putative cancer driver noncoding RNAs. Their directional association to cancer is stronger than 90% of the curated cancer driver genes.

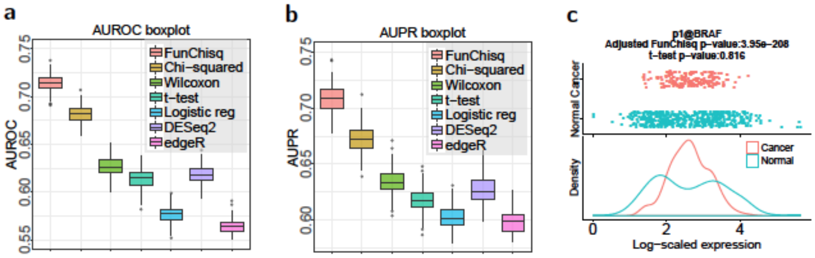

Fig. 1. FunChisq outperformed six other methods and detected non-monotonic patterns. (a) Areas under ROC curves. (b) Areas under PR curves. (c) Non-monotonic association is observed from known cancer gene *BRAF* to cancer.

Next, we compared samples from two leukemia subtypes (lymphoid and myeloid leukemia) against other samples in FANTOM5. FunChisq predicted 332/79 potential biomarkers for lymphoid/myeloid leukemia, stronger than the TSSs of all 87/100 known lymphoid/myeloid leukemia driver genes. Our findings of the biomarker locations are consistent with known chromosomal abnormalities in both leukemia subtypes.

Using the powerful FunChisq method to detect directional association that can be either monotonic or non-monotonic, our study contributes a catalog of novel biomarker candidates that may enable a deeper understanding of cancer.

References

1. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R.: A census of human cancer genes. Nature Reviews Cancer **4**(3), 177–183 (2004)
2. Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al.: Inferring causal molecular networks: empirical assessment through a community-based effort. Nature Methods **13**(4), 310 (2016)
3. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. John Wiley & Sons (2013)
4. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al.: Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology **16**(1), 22 (2015)
5. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(12), 550 (2014)
6. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 **50**(302), 157–175 (1900)
7. Rice, J.: Mathematical Statistics and Data Analysis. Nelson Education (2006)
8. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
9. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)