Published in final edited form as:

Drug Dev Res. 2014 September ; 75(6): 343–347. doi:10.1002/ddr.21220.

Correcting Transcription Factor Gene Sets for Copy Number and Promoter Methylation Variations

Komal S. Rathi¹, Daria A. Gaykalova², Patrick Hennessey³, Joseph A. Califano^{2,4}, and Michael F. Ochs⁵

¹ Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

² Department of Otolaryngology—Head and Neck Surgery, Johns Hopkins Medical Institutions, Baltimore, MD, USA

³ Mid-Michigan Ear, Nose, and Throat, P.C., East Lansing, MI, USA

⁴ Milton J. Dance Head and Neck Center, Greater Baltimore Medical Center, Baltimore, MD, USA

⁵ Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ, USA

Abstract

Gene set analysis provides a method to generate statistical inferences across sets of linked genes, primarily using high-throughput expression data. Common gene sets include biological pathways, operons, and targets of transcriptional regulators. In higher eukaryotes, especially when dealing with diseases with strong genetic and epigenetic components such as cancer, copy number loss and gene silencing through promoter methylation can eliminate the possibility that a gene is transcribed. This, in turn, can adversely affect the estimation of transcription factor or pathway activity from a set of target genes, since some of the targets may not be responsive to transcriptional regulation. Here we introduce a simple filtering approach that removes genes from consideration if they show copy number loss or promoter methylation and demonstrate the improvement in inference of transcription factor activity in a simulated data set based on the background expression observed in normal head and neck tissue.

Keywords

Gene set analysis; transcription factor gene sets; copy number variations; promoter methylation; simulated data set

INTRODUCTION

Gene set analysis (GSA) has a substantial history as a tool in bioinformatics, beginning with work by Draghici et al., [2003] and Coombes et al., [2002]. The fundamental concept is to borrow strength across genes for inference of activity that involves the entire set of genes. For instance, the activation of a metabolic pathway typically requires the creation of a set of

enzymes that together convert one metabolic compound into another through a set of intermediaries. Often, the genes encoding these enzymes are regulated as a group, so that all the necessary enzymes can be created from a single driving event in a straightforward manner. Each individual gene may show a noisy signal related to the activation of the driving event, but together they may provide a substantial signal if treated as a group [Mootha et al. 2003].

The gene sets with the strongest logical basis in eukaryotic organisms are those generated from the targets of transcription factors (TFs). Since the TF binds upstream of each target and regulates its expression, the genes known to be regulated by an individual TF provide a natural gene set. In practice, TF regulation is more complicated, as each target gene may be regulated contextually, so that it is only transcribed in certain situations. The most obvious example of regulation of this type is provided by the epigenetic silencing of developmental genes, so that they are expressed only during development and later activation of their transcriptional regulators does not lead to potentially deleterious activation in the adult organism. We have demonstrated the value of gene set analysis on TF targets in inferring the on- and off-target effects of small molecule targeted (SMT) therapeutics [Ochs et al. 2009].

In many cases in disease, especially in cancer with its genetic variation and epigenetic changes, the target genes of a TF may be unavailable for transcription even in contexts where transcription would be expected. Retaining the genes in the calculation of the gene set statistic then leads to errors in estimation of the activity level of a TF, as the change in expression expected is not observed. This can lead, especially in cancers with substantial loss of heterozygosity (LOH), chromosomal loss, or epigenetic changes, to faulty estimation of TF activity and consequent loss of inferential power concerning signaling changes and transcriptional reprogramming.

METHODS

Here we provide a brief overview of GSA and the modifications we have made to the approach, as well as details on our simulation method and the analysis of results. We limit our simulations to the effects of promoter methylation for simplicity, however the extension to copy number variation (CNV) is straightforward.

Filtered Gene Set Analysis (FGSA)

GSA relies on the creation of gene-based statistics comparing two groups, as in differential expression analysis. Traditional tests, such as the *t*-test for comparing the means of the expression level in two classes, have been shown to be inadequate due to the low sample count and high noise in transcription [Tusher et al. 2001]. As such, most methods rely on more robust methods to derive initial gene-level statistics, such as empirical Bayes estimates [Smyth 2004], or on methods that provide the statistics in a single step on the set, such as GSEA [Subramanian et al. 2005].

For this work, we had to address the issue that in an individual tumor, the methylation and CNV will typically vary from the mean, so that a simple statistic comparing two classes is not feasible. Instead, we created a pseudo-*t*-statistic for each gene taking into account

promoter methylation and CNV. First, we treated the expression level for a gene with high methylation ($\beta > t_{\text{meth}}$, with t_{meth} provides a threshold) or low copy number ($\text{CNV} < t_{\text{cnv}}$) as unmeasured, effectively removing these values from the expression data. Second, we calculated the mean and standard deviation of all normal samples (μ_N, σ_N), thus taking into account only expression levels where methylation and CNV have not negatively impacted expression. Third, we set the statistic for each tumor and each gene as

$$S_{gp} = \frac{e_{gp} - \mu_N}{\sigma_N}$$

where s_{gp} provides the statistic and e_{gp} is the value of expression of the g^{th} gene in the p^{th} patient. As such, s_{gp} can be viewed as the Z-score of the individual tumor measurement relative to the distribution of normal samples and is thus related to the probability that the expression in the tumor is from the same distribution as found in normal tissue.

We then used s_{gp} to rank the genes for GSA. In order to minimize the potential influence of individual genes where $|s_{gp}|$ is large, potentially due to noise in the single tumor measurement or the assumptions of normality inherent in a Z-score, the Wilcoxon test was used for calculating p -values for the gene sets, as implemented in the limma R package [Smyth 2005].

Key parameters in the determination of s_{gp} are obviously the thresholds, t_{meth} and t_{cnv} . In other work we have explored the impact of varying the thresholds, however here for simulations we will treat the cutoffs as fixed at $t_{\text{meth}} = 0.15$ and $t_{\text{cnv}} = 1.2$. Detailed studies on patient data showing the impact of varying thresholds in related work on outlier statistics can be found elsewhere [Ochs et al. 2013].

Simulations

With the goal of mirroring the true complexity of gene expression, we based our simulations on measurements of 44 tumor samples and 25 normal tissue samples, which we have previously submitted to the Gene Expression Omnibus, GSE33232. We first identified a natural gene-specific noise level by calculating the standard deviation for each gene in all tumors and all normals, obtaining separate estimates for normal and tumor samples.

We next created a quantile regression model, with $\tau = 0.9$, relating the expression of a gene to its methylation level. We used a large τ in order to capture only the case where the gene was actually expressed, thus assuming, naturally, that we could have low methylation but no expression due to other regulatory mechanisms. For each gene we then obtained a model of expression for when the gene was activated by a process, as corrected for promoter methylation.

The regression estimates were used to create a model of expression that included baseline expression levels for a gene based on the intercept, additional TF driven expression if a TF was active, and a correction for methylation based on a simulated methylation matrix. The simulation included eight TFs with targets identified from TRANSFAC [Matys et al. 2006],

and activity due to activation adding log2 expression from a uniform distribution between 8 and 12. The expression change was assumed to be the same each time a TF was active.

Multiple methylation states were simulated, where for each case a model for a gene based on methylation in normal samples was modified by assuming no additional methylation, 10% of genes, 25% of genes, or 50% of genes methylated randomly in tumors. For methylated genes, a uniform distribution with β between 0.6 and 0.95 was used to generate a new value for the methylation state.

Finally, based on our estimates of the noise levels in tumors and normals, noise was added to the simulated expression data at three levels, equivalent to measured variation and two and four times greater. The final model for log2 expression was then generated for each of the four methylation levels (0, 0.1, 0.25, 0.50) and was given by

$$e_{gp} = e_g^0 + \beta_g^0 m_{gp} + k\mathbb{N}(\mu_{gp}, \sigma_{gp})$$

where e_{gp} is the log2 expression, e_g^0 is the baseline expression level for gene g with no methylation, m_{gp} is the methylation level for gene g in patient p , β_g^0 is the coefficient from the regression model for the impact of methylation on expression, k is the noise level (1, 2, or 4), and μ_{gp}, σ_{gp} give the mean and standard deviation for the noise for gene g in patient p , where two values are used, one across all normal patients and the other across all cancer patients. For tumors, e_g^0 is modified by the addition of expression due to TF activity. TF activity is modeled as on or off from a binary random distribution with probability of 0.5. The final values in the simulated expression matrices were limited to the log2 range [5, 14] to capture the actual dynamic range in the expression array data.

Analysis

The final result was then a set of 12 simulated expression data sets representing four methylation levels and three noise levels. These were analyzed using GSA and FGSA applying a Wilcoxon rank sum test, and for FGSA, the threshold $t_{\text{meth}} = 0.15$. Results were summarized as the number of correct estimates of TF activity, given the known distribution of active TFs in the tumor samples. Since there were 44 simulated tumors and 8 TFs, this provided 352 binary calls per simulation. Two-by-two tables were used to compare the correct calls from FGSA and GSA.

In order to clarify the impact of FGSA, we also generated some heatmaps demonstrating the overall bias introduced by failure to correct for methylation when making estimates of TF activity.

RESULTS

In Table 1, the correct and incorrect calls are summarized for all simulations. The correct call was assigned to any TF in a tumor where the TF was active in the simulation and the p -value was less than 0.5 based on the gene set statistic. As is clear, FGSA outperforms GSA

in all cases, and importantly shows poorer behavior as noise levels and methylation levels increase.

Table 2 provides the accuracies for the calls in Table 1. The decline in accuracy at the highest methylation levels can be explained by the loss of genes in the gene sets. As gene sets become smaller, it becomes more difficult to generate a reasonable p -value, simply due to the small size of the gene set.

It is interesting that under these simulations GSA performs essentially randomly at the lowest noise levels, with a small hint of improvement as noise increases. The reason for this is clarified in Figure 1. GSA essentially assigns a low p -value to all gene sets at all times. Under our simulation, activity is made positive in 50% of tumor-TF pairs, so that the result is a 50% correct call rate from a test that always is positive.

DISCUSSION

The ultimate goal of statistical analysis in the age of high-throughput molecular measurements is the correct estimation of biological process activity and the identification of specific molecular changes that drive activity change. The ability to measure multiple molecular species genome-wide offers a new opportunity for methods to integrate the data into a single view of the system, which allows for greater statistical power.

The most powerful tools are those that most correctly integrate our understanding of the biology and technology that generate the measurements into the statistical framework. Examples include Cancer Outlier Profile Analysis (COPA), which matches the fact that cancer is highly heterogeneous to an outlier statistic needed to identify variants [MacDonald and Ghosh 2006], Top-Scoring Pair (TSP), which removes normalization issues when defining molecular biomarkers [Geman et al. 2004], and Outlier Gene Set Analysis (OGSA), which refines outlier statistics with biologically motivated thresholds [Ochs et al. 2013]. FGSA fits within this framework as it attempts to refine expression gene sets based on biological models of gene expression and additional molecular measurements.

Filtering gene sets does have serious limitations. The loss of genes from the set leads to difficulty in obtaining significant p -values in a rank statistical test, where low gene number cannot be overcome by strong differential expression. We therefore believe that a more robust model that weights the genes in the set based on their methylation and copy number status are likely to provide a more powerful statistical test in the future.

Acknowledgments

MFO was supported by the National Library of Medicine (R01 LM001100). DAG, PH, and JAC were supported by NIH/NIDCR (DE020324) and the Johns Hopkins University Head and Neck Cancer SPORE (5P50DE019032).

REFERENCES

Coombes, KR.; Baggerly, KA.; Stivers, DN.; Wang, J.; Gold, D.; Sung, HG.; Lee, SJ. Biology-Driven Clustering of Microarray Data: Applications to the NCI60 Data Set. In: Johnson, K.; Lin, S., editors. Methods of Microarray Data Analysis II. Kluwer Academic Publishers; Boston: 2002. p. 65-79.

- Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* 2003; 31(13):3775–3781. [PubMed: 12824416]
- Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparison. *Statistical Applications in Genetics and Molecular Biology.* 2004; 3(1):19.
- MacDonald JW, Ghosh D. COPA--cancer outlier profile analysis. *Bioinformatics.* 2006; 22(23):2950–2951. [PubMed: 16895932]
- Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006; 34(Database issue):D108–110. and others. [PubMed: 16381825]
- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics.* 2003; 34(3):267–273. and others. [PubMed: 12808457]
- Ochs MF, Farrar JE, Considine M, Wei Y, Meshinchi S, Arceci RJ. Outlier Analysis and Top Scoring Pair for Integrated Data Analysis and Biomarker Discovery. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM.* 2013
- Ochs MF, Rink L, Tarn C, Mburu S, Taguchi T, Eisenberg B, Godwin AK. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer research.* 2009; 69(23):9125–9132. [PubMed: 19903850]
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004; 3 Article3.
- Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, RV.; Dudoit, S.; Irizarry, R.; Huber, W., editors. *Bioinformatics and Computational Biology Solutions using R and BioConductor.* Springer; New York: 2005. p. 397-420.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005; 102(43):15545–15550. and others. [PubMed: 16199517]
- Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001; 98(9):5116–5121. [PubMed: 11309499]

Figure 1.

A comparison of the simulated TF activity on the left, the p-value estimate from GSA in the middle, and the p-value estimate from FGSA on the right, for the case of no additional methylation and noise level = 1. Darker shading indicates activity and low p-value. It is clear that FGSA does a much better job of reflecting the structure in the simulation. A color version is available online with red indicating activity.

Table 1

Correct and incorrect transcription factor activity calls at different levels of methylation and expression noise. Each call is made in an individual tumor on the simulated activity. Noise levels are based on normal and tumor gene level estimates generated from the head and neck data set.

Methylation		Noise Level = 1		Noise Level = 2		Noise Level = 4	
		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
None	FGSA	304	48	264	88	219	133
	GSA	177	175	184	168	181	171
10%	FGSA	282	70	248	104	207	145
	GSA	175	177	184	168	197	155
25%	FGSA	262	90	228	124	206	146
	GSA	177	175	180	172	182	170
50%	FGSA	213	139	203	149	192	160
	GSA	182	170	184	168	187	165

Table 2

The accuracies for the values in Table 1.

Methylation		Noise Level = 1	Noise Level = 2	Noise Level = 4
None	FGSA GSA	0.864 0.503	0.750 0.523	0.622 0.514
10%	FGSA GSA	0.801 0.497	0.705 0.523	0.588 0.560
25%	FGSA GSA	0.744 0.503	0.648 0.511	0.585 0.517
50%	FGSA GSA	0.605 0.517	0.577 0.523	0.546 0.531