Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-*pathways* Mahalanobis distance within pathways of single subjects predicts breast cancer survival

A. Grant Schissler[1,2,3,*], Vincent Gardeux[1,3,*], Ikbel Achour[1,3,*], **Qike Li**[1,2,3], Haiquan Li[1,3], Walter W. Piegorsch[1,2,3,§], Yves A. Lussier[1,3,§]; *AGS, VG, and IA contributed equally; § Corresponding Authors

Abstract

Motivation: The conventional approach to personalized medicine relies on molecular data analytics across multiple patients. The path to precision medicine lies with molecular data analytics that can discover interpretable single-subject signals (N-of-1). We developed a global framework, N-of-1-*pathways*, for a mechanistic-anchored approach to single-subject gene expression data analysis. We previously employed a metric that could prioritize the statistical significance of a dysregulated pathway in single subjects, however, it lacked in quantitative interpretability (e.g., the equivalent to a gene expression fold-change).

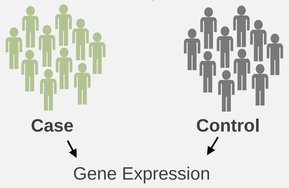
Results: In this study, we extend our previous approach with the application of statistical Mahalanobis distance to quantify personal pathway-level dysregulation. We demonstrate that this approach, N-of-1-*pathways* Mahalanobis Distance (MD), detects dysregulated pathways (empirical simulations), while not inflating false positive rates using a study with biological replicates. Finally, we establish that N-of-1-*pathways* MD scores are biologically significant, clinically relevant, and are predictive of breast cancer survival ($p<0.05$, n=80 invasive carcinoma; TCGA RNA-sequences).

Conclusion: N-of-1-*pathways* MD provides a practical approach towards precision medicine. The method generates the magnitude and the biological significance of personal dysregulated pathways results derived solely from the patient's transcriptome. These pathways offer the opportunities for deriving clinically actionable decisions that have the potential to complement the clinical interpretability of personal polymorphisms obtained from DNA acquired or inherited polymorphisms and mutations. In addition, it offers an opportunity for applicability to diseases in which DNA changes may not be relevant, and thus expand the "interpretable 'omics" of single subjects (e.g. personalome).

Availability: http://www.lussierlab.net/publications/N-of-1-pathways

Problem Statement

Population/cohort



Common
Gene / pathway signature

Conventional Methods
GSEA, DE gene enrichment

- 99% of molecular biomarkers derived from large patient samples predictors fail to be reproducible
- Pathway signatures built on cohort miss individual specificities

''ONE DRUG FITS ALL''

Individual ?

Control / Case
A single paired Sample

Gene Expression

Common
Pathway signature

Individual
Pathway signature

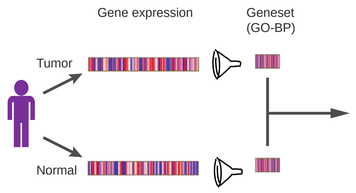
METHODS for personal transcriptome ?

"N-OF-1-PATHWAYS"
Bridging the gap

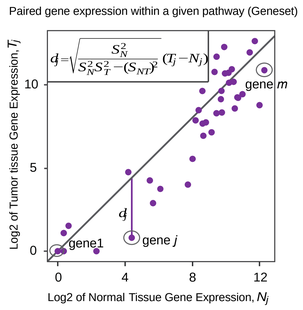
INDIVIDUALIZED TREATEMENT

Method: N-of-1-*pathways* Mahalanobis Distance

A Single Subject Input



B N-of-1-*pathways* MD: Calculation I

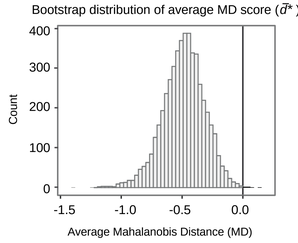


C N-of-1-*pathways* MD: Calculation II

Gene	Normal N_j	Tumor T_j	Mahalanobis Distance (MD)
1	0	0	d_1 = 0
2	0.36	1.39	d_2 = 0.48
...	...	...	...
j	4.41	0.81	d_j = -2.37
...	...	...	...
m	12.32	10.88	d_m = -0.95
Mean	8.084	7.37	$\vec{d}$ = -0.47

indices of genes within a geneset

D N-of-1-*pathways* MD: Calculation III



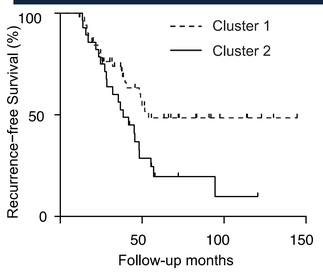
E Single Subject Output

Output

Single subject ID	GO-BP ID	GO-BP terms	GO-BP: Gene Count	Deregulated	MD CRM ($\vec{d}$)	Direction
Jane Smith	GO:0008380	RNA splicing	323	Yes	-0.279	-
Jane Smith	GO:0071363	Cellular response to growth factor stimulus	264	Yes	0.307	+
Jane Smith	...	...	...	...	...	...
Jane Smith	GO:0060350	Endochondral bone morphogenesis	37	No	0.001	+

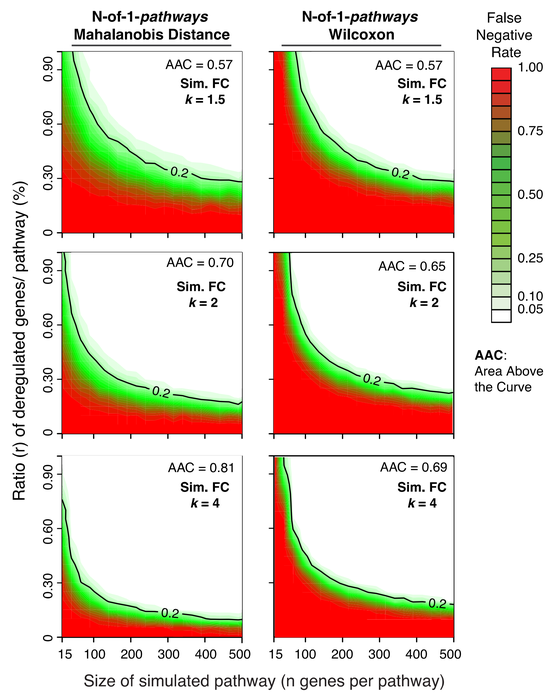
Method Overview of N-of-1-*pathways* Mahalanobis Distance. (**A**) The input is represented by the gene expression of single patient paired samples (e.g. tumor vs normal tissue) filtered into *a priori* defined genesets (e.g. Gene Ontology Biological Processes: GO-BP pathways). (**B**) Calculation I is visualized by the bivariate relationship between normal and tumor gene expression values for a given geneset (e.g. GO-BP pathway). The vertical, signed Mahalanobis distance (MD), d_j, is computed from each jth point (gene) to the diagonal line representing equal expression. (**C**) Calculation II: The mean MD represents the pathway-level dysregulation from normal to tumor expression where a negative value indicates down-regulation and a positive value represents up-regulation. The gene indices are randomly resampled and the "average MD score" is re-computed via bootstrapping to determine pathways with strong evidence of dysregulation. (**D**) Calculation III: The bootstrap distribution of "average MD scores". (**E**) The process results in pathway-level quantification of dysregulation, an approach to obtain a Clinically Relevant Metric (CRM).

Results: 1 MD Scores Predict Survival



N-of-1-*pathways* MD GO-BP clinical importance metrics predict breast cancer patient survival. N-of-1-*pathways* MD was applied to n=80 invasive breast carcinoma patients (TCGA_BRCA, RNA-seq) resulting in 3225 clinical importance metrics. Every patient has an N-of-1-*pathways* MD score for each of the identified dysregulated pathways (2130 pathways identified in at least one patient) and we performed principal component analysis and unsupervised clustering on these scores. Unsupervised partitioning around medoids (PAM) clustering reveals distinct Kaplan-Meier survival curves (log-rank p-value < 5%).

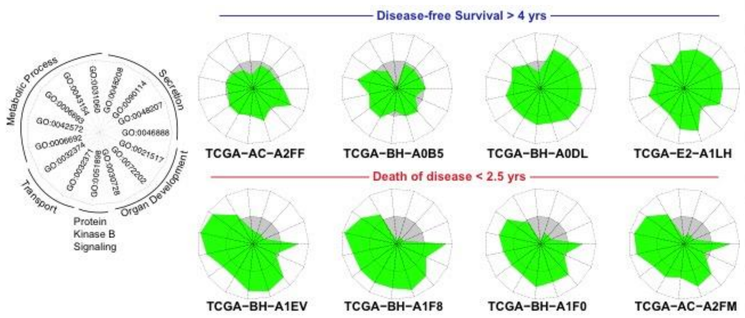
2 Simulated RNA-Seq datasets show that N-of-1-*pathways* MD Detects Dysregulation Powerfully with Infrequent False Positives



AAC: Area Above the Curve

Simulation study reveals that N-of-1-*pathways* MD powerfully detects artificially dysregulated pathways. Each point represents one size of a simulated pathway generated by randomly selecting n genes and a ratio r of the dysregulated genes within the pathway. The ratio r is artificially increased by a k-fold change in a simulated pathway generated from biological replicates (k=*1.5, 2, 4*). We then applied separately the N-of-1-*pathways* Wilcoxon (bottom) and N-of-1-*pathways* MD (top) methods to identify whether the truly dysregulated pathway is detected. We repeated the process 5000 times at each combination of (n, k, r) to estimate the false negative error rate. Area Above the Curve (AAC) quantifies the proportion of simulated pathway combinations with false negative error less than 0.20 (the black curve labeled 0.20 is the reference for this measure). Higher AAC indicates a greater number of scenarios with at least 80% power to detect dysregulated pathways. N-of-1-*pathways* MD outperforms N-of-1-*pathways* Wilcoxon at every fold-change, requiring fewer genes in the pathway and a smaller ratio of dysregulated genes. Notably, the simulated false positive rate (0.0% dysregulated genes; rate along the horizontal axis) is smaller for MD than Wilcoxon, averaging 0.14% and 0.94%, respectively. This rate can also be interpreted as the simulated rate of discovery when two non-tumor samples are paired. **Legend**, Sim. = simulated, AAC=area above curve

3 Visualization: Single Patient Dysregulated Pathways



N-of-1-*pathways* representation (star plot) of individual GO-BPs of diametric extreme patients. On the left lies the legend of the star plots, each edge corresponding to one GO term, each star reflects a single patient's dysregulation as measured by the MD CRM for each pathway. On the right lies a sample of eight patients' star plots (four from each extreme). The white zone represents up-regulated pathways (given the N-of-1-*pathways* direction of dysregulation), while the gray zone stands for down-regulation. The circle separating the gray and white areas represents non-dysregulation.

Conclusion

N-of-1-*pathways* MD provides a practical approach towards precision medicine. The method gives clinically actionable results derived solely from the patient. The entire transcriptome does not need to be measured, allowing for targeted experiments across multiple gene expression platforms, reducing cost and providing flexibility. The method generates the magnitude and the biological significance of personal dysregulated pathways results derived solely from the patient's transcriptome.

[1] Bio5 Institute, University of Arizona, Tucson, AZ, USA
[2] Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA
[3] Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ, USA