Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation [1]

A. Grant Schissler[1,4], **Walter W. Piegorsch**[2–5], **and Yves A. Lussier**[4–6]

[1]Department of Mathematics & Statistics, University of Nevada, Reno, NV, USA

[2]Interdisciplinary Program in Statistics, [3]Department of Mathematics, [4]Center for Biomedical Informatics and Biostatistics (CB2), [5]BIO5 Institute, [6]Department of Medicine, University of Arizona, Tucson, AZ, USA.

1 Introduction

We seek to assess a patient's differential RNA expression, but *not* for individual genes, for collections of genes (a gene set or *pathway*). We call this detecting Differentially Expressed Pathways (DEPs) as opposed to Differentially Expressed Genes (DEGs).

1.1 Motivating data & background

The table below contains paired RNA-seq (log_2-normalized) expression data derived from a triple negative breast cancer patient [2]. The genes listed here are contained within a Gene Ontology [3] pathway.

Gene	Tumor expression	Normal expression	Difference
CYP4A11	0.00	3.71	-3.71
AGTR1	6.13	7.86	-1.73
OR51E2	2.90	1.54	1.36
CYP11B2	0.00	0.00	0.00
PTPRO	3.72	6.22	-2.50
CYP4F2	0.00	0.40	-0.40
AGT	8.40	7.89	0.52
...	...	...	...
SERPINF2	6.38	9.57	-3.19

1.2 The N-of-1-*pathways* clinical framework [4]

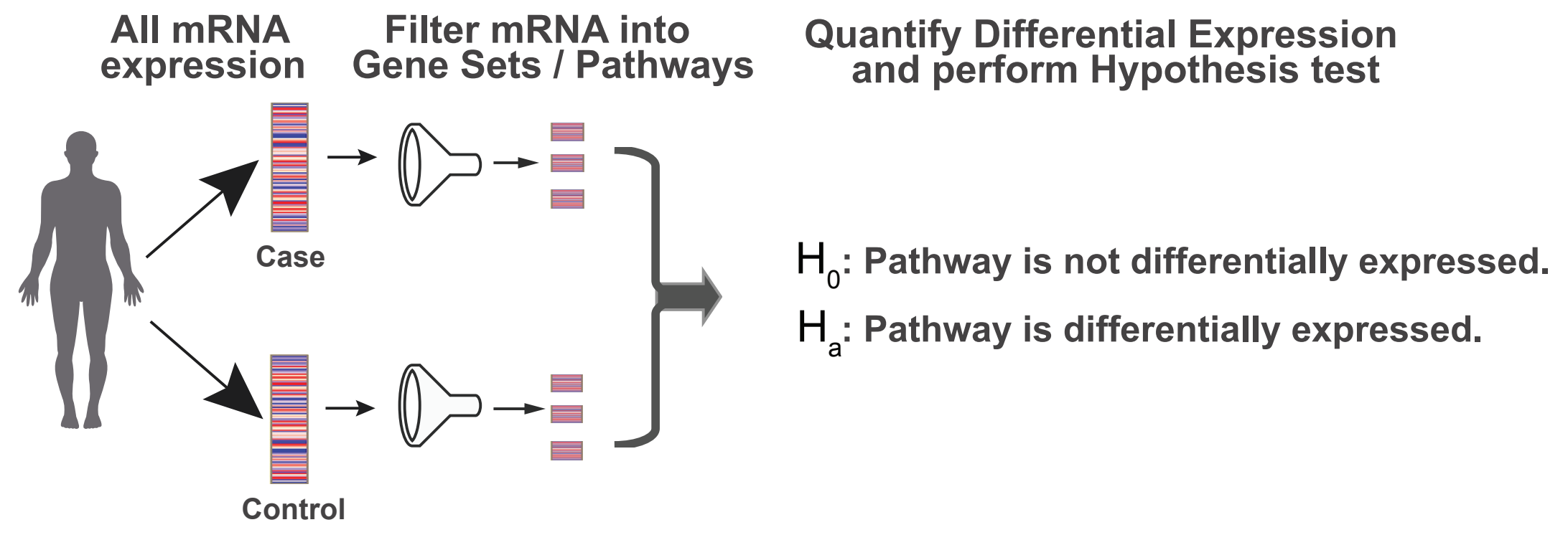
All mRNA expression
Filter mRNA into Gene Sets / Pathways
Quantify Differential Expression and perform Hypothesis test
Case
Control
H_0: Pathway is not differentially expressed.
H_a: Pathway is differentially expressed.

Figure 1: Conceptual workflow to enable precision medicine from within-patient paired expression data.

2 The Clustered-T test statistic

We seek to improve upon the first N-of-1-*pathways* testing procedure, a Wilcoxon signed-rank test [4]. The major issue is that inter-gene correlations invalid modeling assumptions [5].

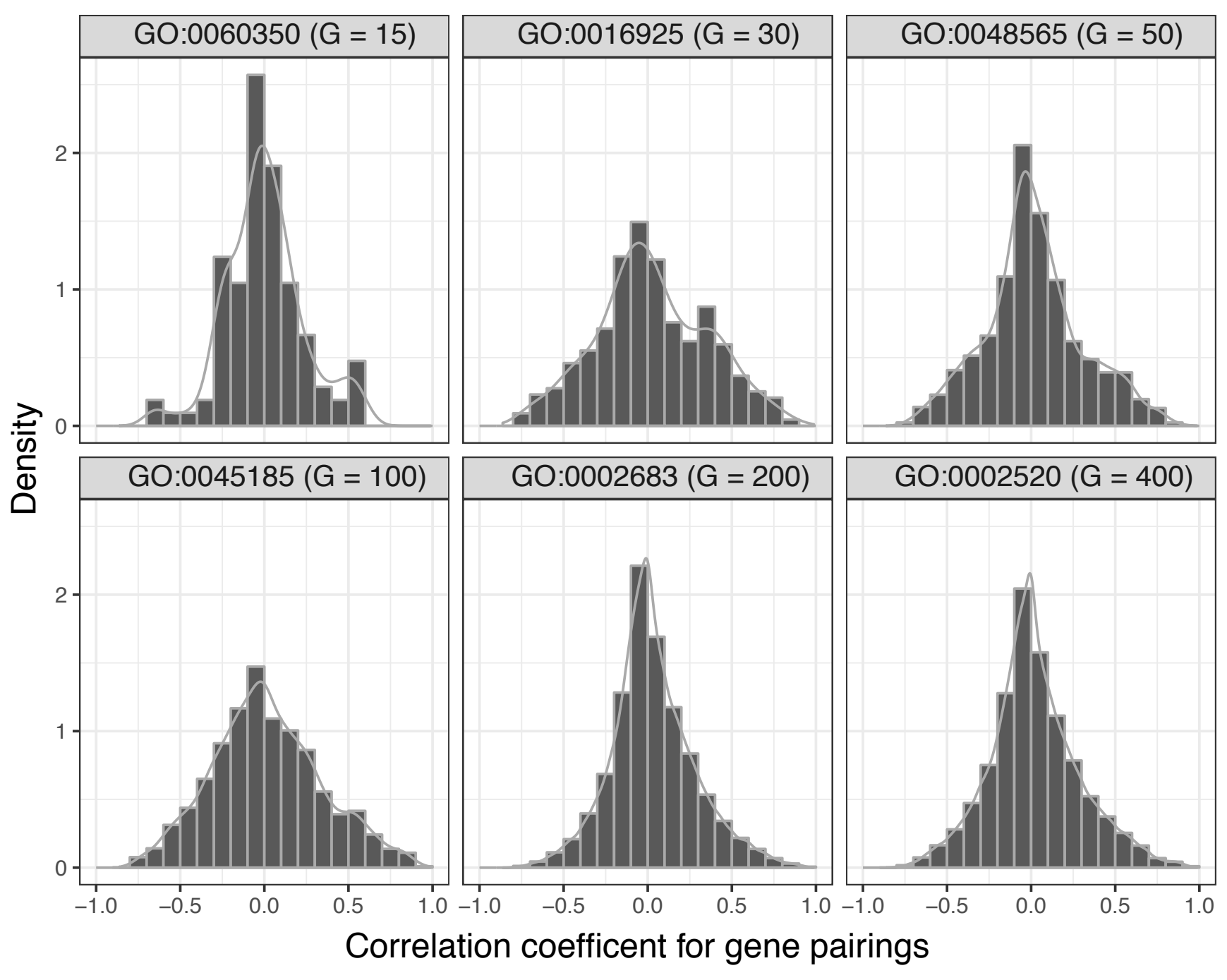

Figure 2: Evidence of gene co-expression within pathways: Pearson correlation estimates across all pairs of genes within 6 Gene Ontology pathways, derived from 111 breast cancer patients.

2.1 Clustering of positively co-expressed genes

Accounting for inter-gene correlatinot is difficutlt with only limtied samples. Instead of estimating covariance directly, we use a robust cluster-correlated variance estimator [6]. This requires that we cluster genes *a priori*, using a database of samples. We omit the clustering procedure here, see Reference [1] for details.

2.2 Notation

Once clustered, we develop our statistic using the following notation:

Concept	Symbol	Definition
Observation index	k	$k = 1, 2, \ldots, n_j$
Cluster index	j	$j = 1, 2, \ldots, m$
Gene-wise difference	d_{jk}	$c_{jk} - b_{jk}$
Total number of genes	G	$G = \sum_j n_j$
Grand sum	d_{++}	$\sum_j \sum_k d_{jk}$
Grand mean	$\bar{\bar{d}}$	d_{++}/G
Cluster sum	d_{j+}	$\sum_k d_{jk}$
Cluster mean	$\bar{d}$	$\sum_j d_{j+}/m$
Sample variance	S_d^2	$\frac{1}{m-1}\sum_j (d_{j+} - \bar{d})^2$

2.3 Williams' robust variance estimator [6]

Modeling the differences as centered and cluster-correlated ($E[d_{jk}] = 0$, $cov[d_{jk}, d_{jk'}] = \sigma_{jkk'}$, and $cov[d_{jk}, d_{j'k'}] = 0$ when $j \neq j'$), we use an unbiased variance estimator for the grand sum:

$$\widehat{\text{Var}}[d_{++}] = \frac{m}{m-1}\sum_{j=1}^{m}(d_{j+} - \bar{d})^2. \tag{1}$$

Clearly, $\widehat{\text{Var}}[d_{++}] = mS_d^2$.

2.4 Hypotheses & reference distribution

Denote $\mu = E\left(\bar{\bar{D}}\right)$. Then the statistical hypotheses of interest are

$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0\,. \tag{2}$$

To build a reference distribution, model the cluster sums as $D_{j+} \sim N(0, \sigma^2)$. Then, conditional on the cluster assignments and under H_0, our Clustered-T statistic

$$T = \frac{\bar{d}}{S_d/\sqrt{m}} \tag{3}$$

follows a $t(m-1)$ distribution.

A (two-sided) P-value is $2 \times \Pr[t(m-1) \geq |T|]$.

3 Monte Carlo evaluation

We evaluate our methodology and compare to the leading alternatives, a Wilcoxon signed-rank test and an unadjusted (naïve) t-test.

3.1 Simulation settings

Variable	Description	Values
G	Number of genes in pathway	{15, 30, 50, 100, 200, 400}
p	the proportion of DEGs	{0, 0.3, 0.6, 0.9}
ψ	fold change of DEGs	{1.5, 2, 4}
$\mathbf{R}$	pathway correlation structure	{Independent, Block, All}

- 'Non-DEG ': $X_i \sim NegBin(\hat{\mu_i}, \hat{\delta_i})$
- 'DEG ': $X_i \sim NegBin(\psi \times \hat{\mu_i}, \hat{\delta_i})$

3.2 Simulating pathways via copulas

To induce correlation, we use copulas [7].

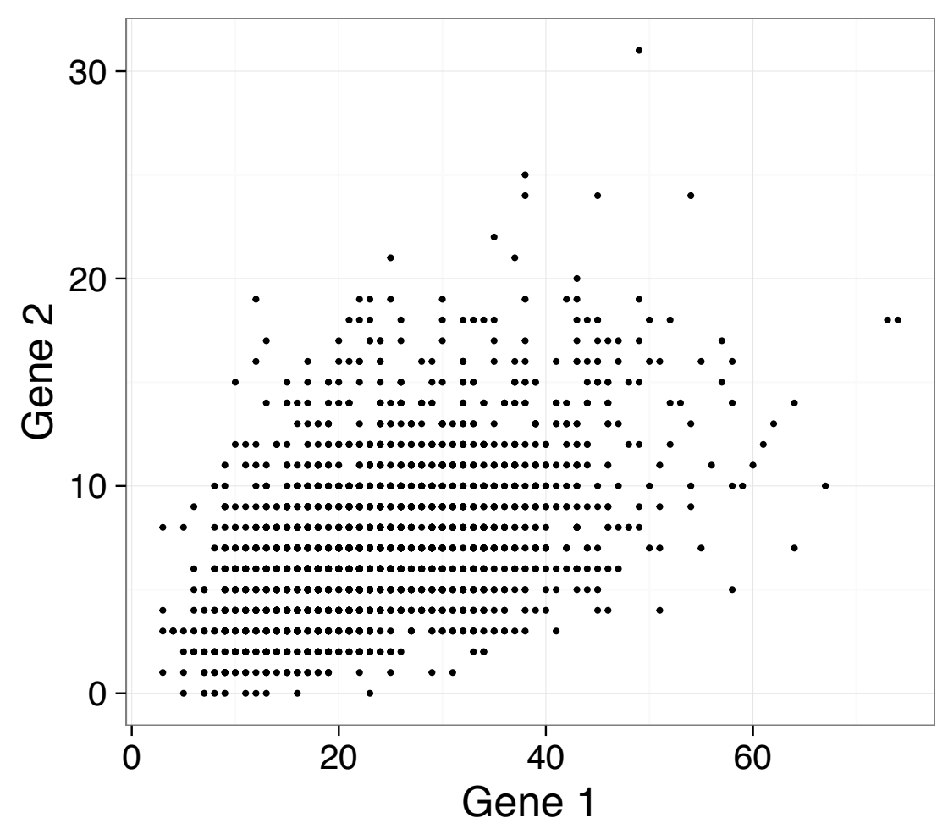

Figure 3: 2000 simulated bivariate gene counts with specified correlation of 0.49 and heterogeneous negative binomial marginals.

3.3 Evaluation: operating characteristics

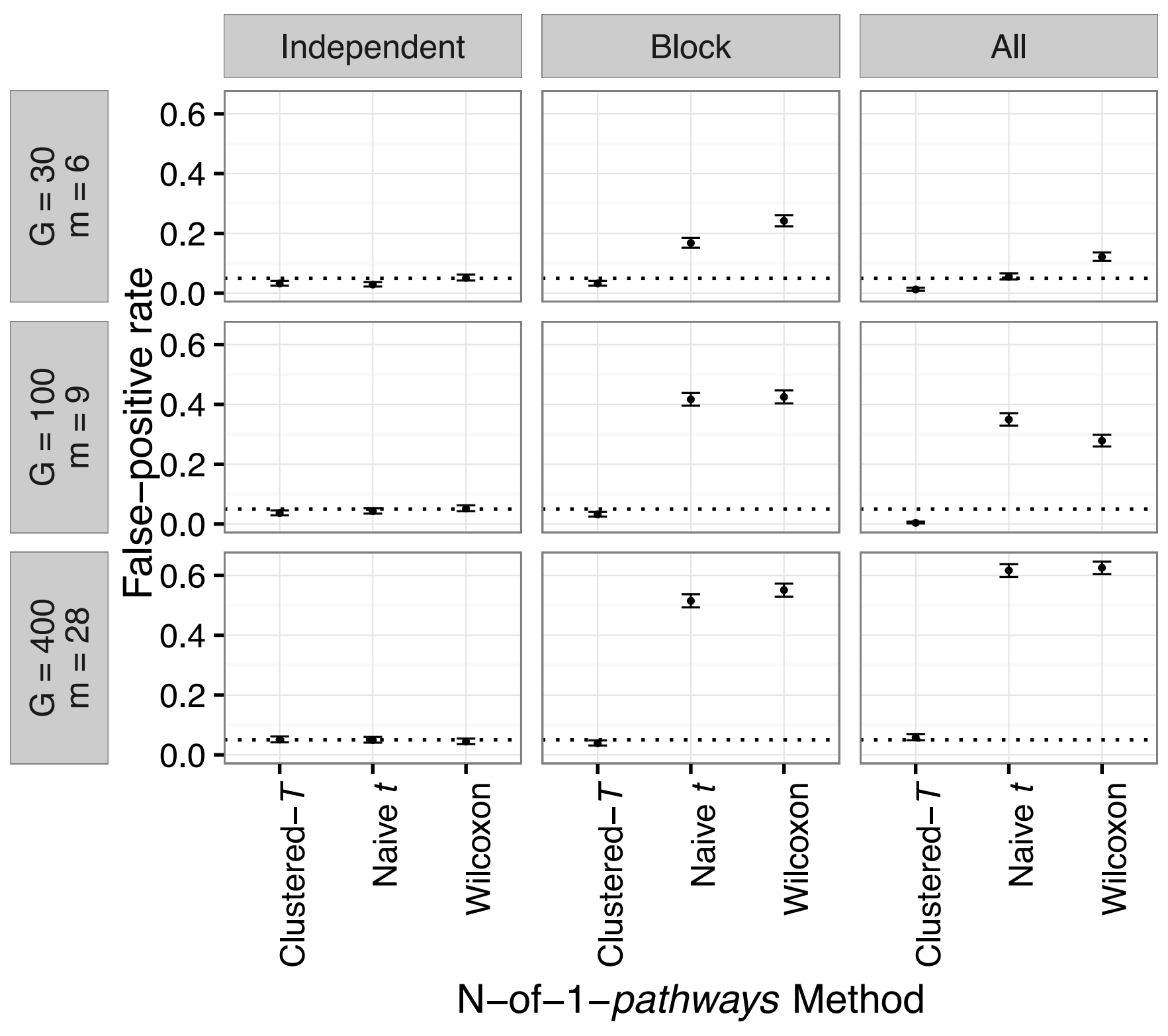

Figure 4: The Clustered-T statistic maintains a 5% size of the test under various correlation structures — 'Independent' simulates genes independently, 'Block' simulates under the clustering assumptions, 'All' allows all genes to be co-expressed.

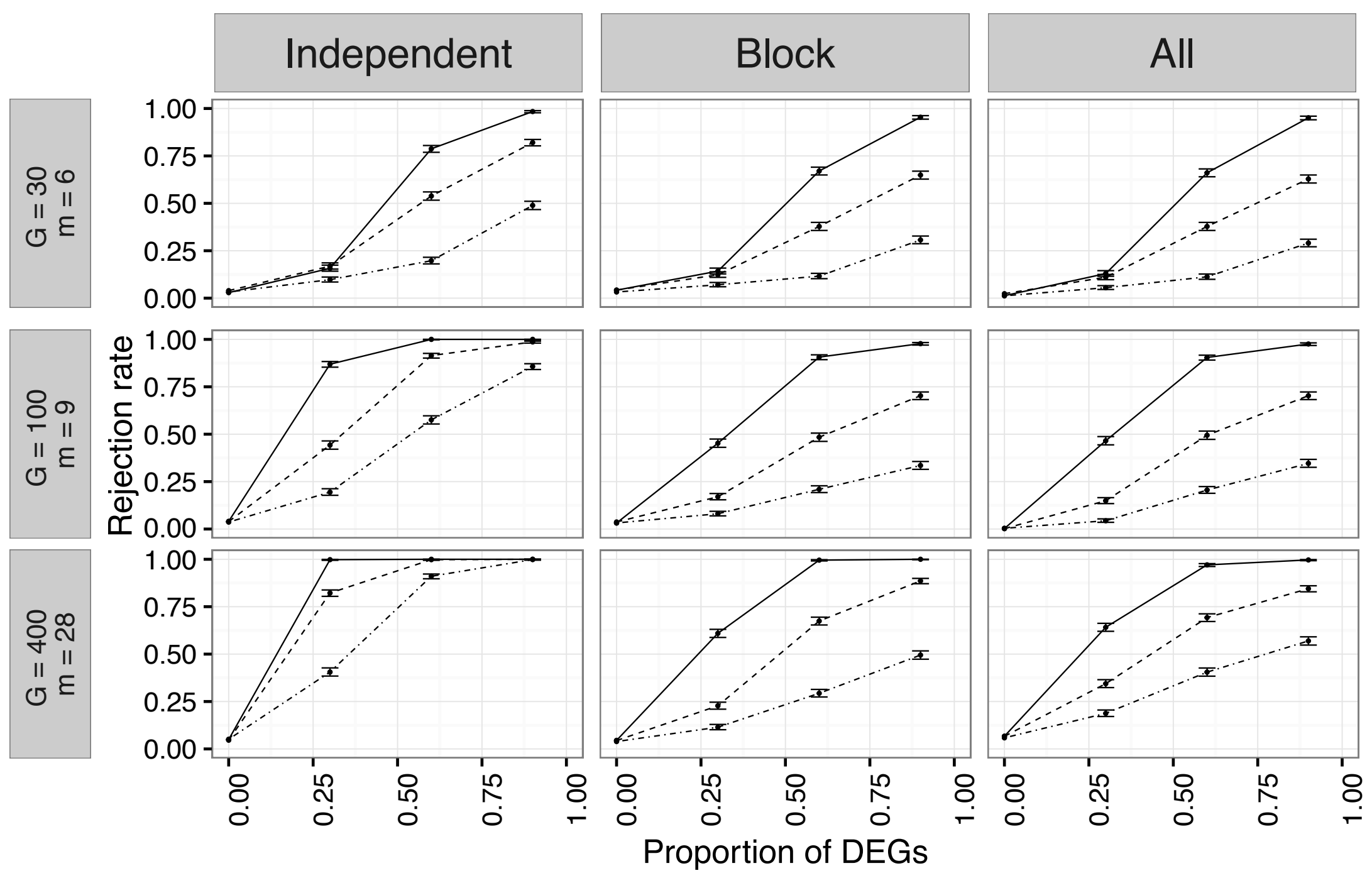

Figure 5: The Clustered-T displays adequate power while increasing fold change of DEGs (dash-dot = 1.5, dashed = 2, solid line = 4).

4 Application

We present our patient's four top-hit differentially expressed pathways when testing 3411 GO-BP pathways. These pathways represent potential therapeutic targets to enable precision medicine.

Gene set description	$\bar{\bar{D}}$	T-stat	P-value	G	m
pos reg of cell adhesion	-0.75	-4.92	0.00011	226	19
reg of resp to external stimulus	-0.47	-4.42	0.00015	458	28
mitochondrial translational initiation	0.28	7.55	0.00028	84	7
regulation of cell morphogenesis involved in differentiation	-0.51	-4.80	0.00028	168	15

5 Concluding remarks

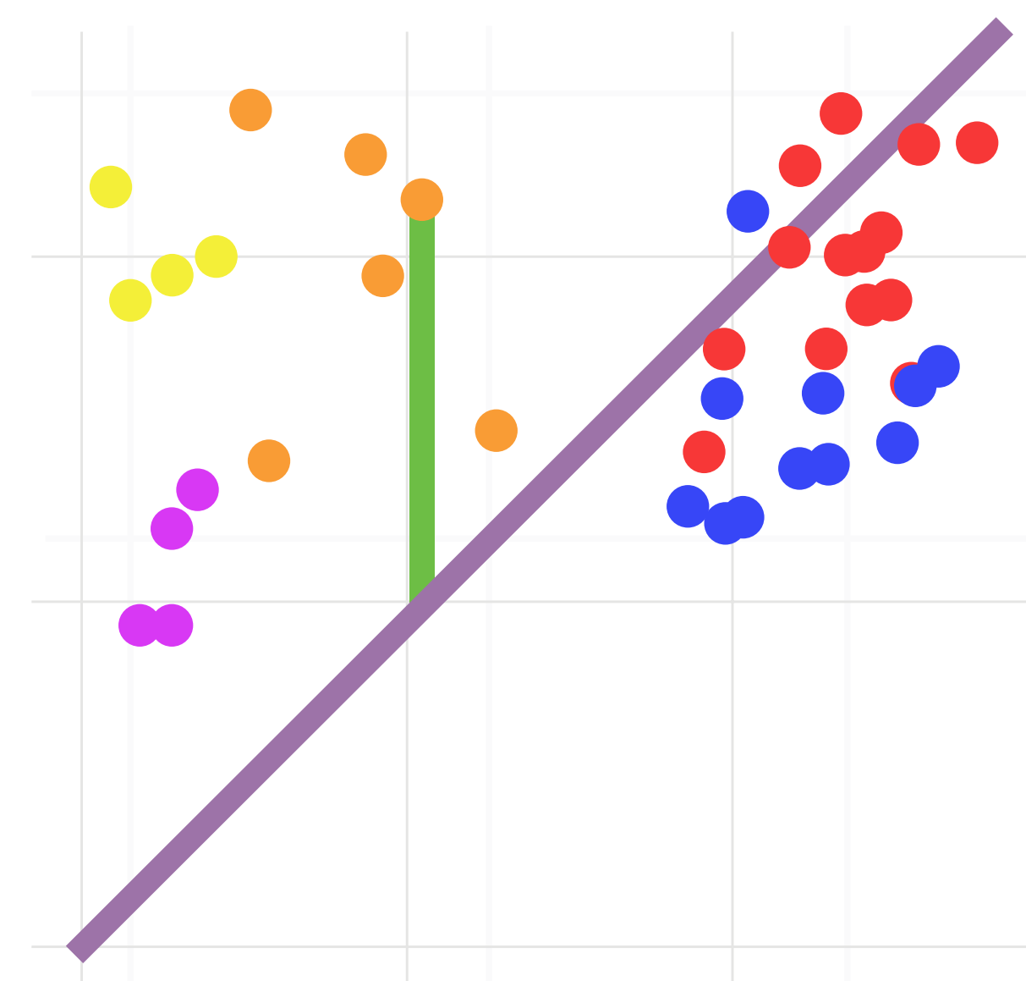

Figure 6: Illustrative summary. The axes represent baseline (horizontal) and tumor (vertical) expression within a pathway. The diagonal line visualizes equal expression. The coloring of each gene indicates cluster assignment. The vertical green line displays gene-wise differential expression. We use a clustered-correlated variance estimator to assess differential pathway expression.

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 1228509 and by the U.S. National Institutes of Health under Grant No. R03ES027394.

References

[1] Schissler, A. G., Piegorsch, W. W., and Lussier, Y. A. (2018) Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation. *Statistical Methods in Medical Research*.

[2] Weinstein, J. N., et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**, 1113.

[3] Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

[4] Gardeux, V., et al. (2014) 'n-of-1-pathways' unveils personal deregulated mechanisms from a single pair of rna-seq samples: towards precision medicine. *Journal of the American Medical Informatics Association*, **21**, 1015–1025.

[5] Tamayo, P., Steinhardt, G., Liberzon, A., and Mesirov, J. P. (2016) The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, **25**, 472–487.

[6] Williams, R. L. (2000) A note on robust variance estimation for cluster-correlated data. *Biometrics*, **56**, 645–646.

[7] Yan, J. (2007) Enjoy the joy of copulas: With a package copula. *Journal Of Statistical Software*, **21**, 1–21.