

Supplementary Data

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, Vincent Gardeux, Qike Li, Ikbel Achour, Haiquan Li,
Walter W. Piegorsch, Yves A. Lussier

Supplement Table S1	ii
Supplement Table S2	iii
Supplement Figure S1	iv
Supplement Figure S2	v
Supplement Figure S3	vi
Supplement Figure S4	vii
Supplement references	viii

Supplement Tables

Table S1. Patient clinical data of MD-derived clusters and diametric extreme patients indicated in Figure 4. Columns 1 and 2 contain additional information regarding the two patient clusters identified by unsupervised clustering of the 2130 MD scores from deregulated pathways in at least one patient (**Methods 2.8**). Columns 3 and 4 present clinically relevant information for the diametric extreme phenotypes (TGCA breast cancer, RNA-seq, **Dataset II, Table 1**).

	Cluster 1	Cluster 2	Disease-free > 4 years	Death of Disease < 2.5 years
Num. of Patients	52	28	9	5
Stage I	10 (19%)	7 (25%)	3 (33%)	0
Stage II	22 (42%)	18 (64%)	3 (33%)	2 (40%)
Stage III	16 (31%)	4 (14%)	3 (33%)	2 (40%)
Stage IV	1 (2%)	0 (0%)	0	1 (20%)
Age range	34-90	30-80	40-78	45-90
Age median	57	46	59	80

Table S2. Ranked predictive pathway signatures identified using Support Vector Machine (SVM).

Utilizing the pathway MD CRMs in a different manner, a binary classification ('good' versus 'poor' prognosis) model for the 80 breast cancer patients in **Dataset II, Table 1** was constructed using *svm* in the *e1071* R package. 20 out of 80 patients were given a 'good' clinical outcome, defined by disease-free patients surviving more than 2.5 years. The features were the 2130 MD CRMs for pathways found deregulated in at least one patient. A linear kernel was specified and leave-on-out cross-validation (CV) was used to tune the cost parameter. The model with the lowest CV error rate (0.25) was retained. 57 patients were labeled as the support vectors or in the margin. Weights were assigned to each feature by multiplying the SVM model coefficients times the features from the 57 patients. The features (pathways) were ranked by absolute value of the weights. The top 20 pathways are listed below:

GO ID	Description
0016339	calcium-dependent cell-cell adhesion
0016254	preassembly of GPI anchor in ER membrane
0046633	alpha-beta T cell proliferation
0030949	positive regulation of vascular endothelial growth factor receptor signaling pathway
0016079	synaptic vesicle exocytosis
0001522	pseudouridine synthesis
0046640	regulation of alpha-beta T cell proliferation
0050853	B cell receptor signaling pathway
0021696	cerebellar cortex morphogenesis
0055117	regulation of cardiac muscle contraction
0001782	B cell homeostasis
0007617	mating behavior
0032786	positive regulation of DNA-dependent transcription, elongation
0061437	renal system vasculature development
0061440	kidney vasculature development
0000722	telomere maintenance via recombination
0045909	positive regulation of vasodilation
0042255	ribosome assembly
0072012	glomerulus vasculature development
0000413	protein peptidyl-prolyl isomerization

Supplement Figures

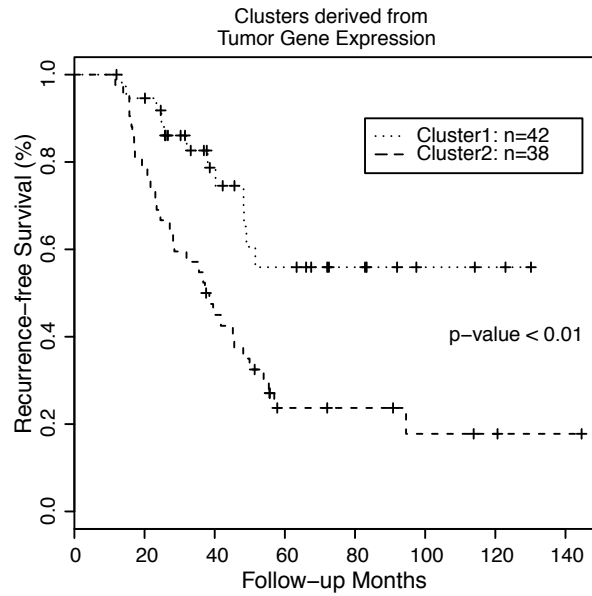
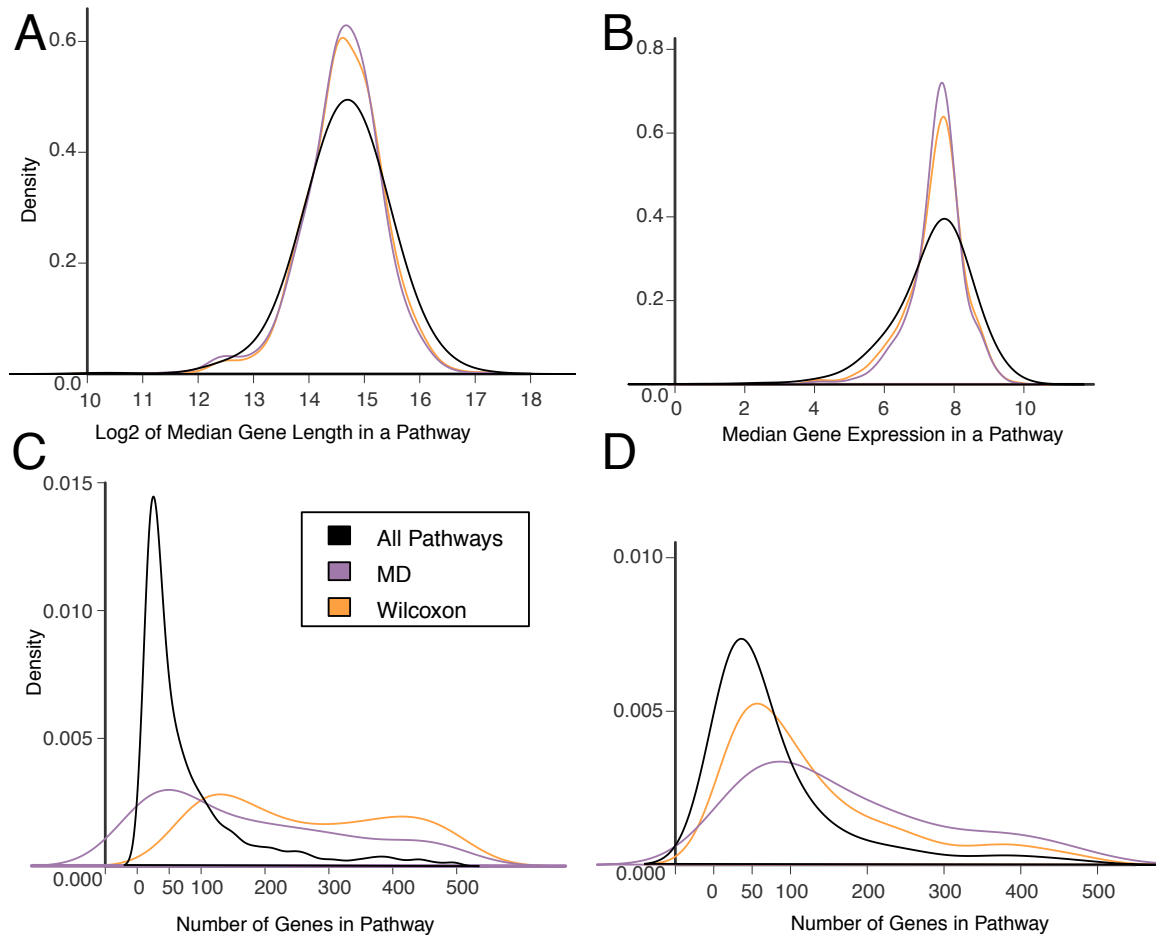
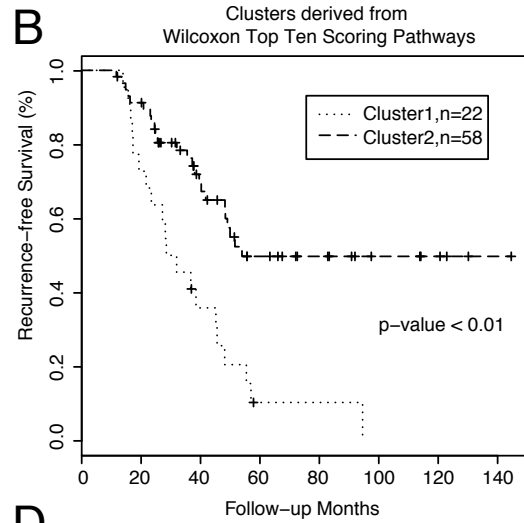
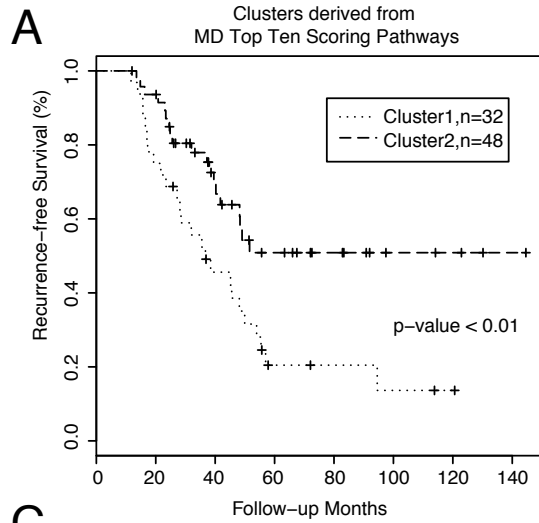


Figure S1. Tumor gene expression predicts survival in breast cancer patients. Gene expression measured from the tumoral samples in (TCGA, RNA-seq, **Table 1 Dataset II**) was analyzed to form a comparison to N-of-1-*pathways* Mahalanobis Distance (MD) results. (A) Two patient clusters were found using the $\log_2(\text{gene expression} + 1)$ values via PAM clustering in the *cluster* R package (**Methods 2.8**). These two clusters produce distinct Kaplan-Meier curves (log-rank p-value < 0.01). Additionally, none of the 1st five principal components of gene expression from the tumor samples separated the diametric extreme phenotypes.



Supplement Figure S2. MD and Wilcoxon methods exhibit bias towards finding larger genesets deregulated, but not towards higher intensity or longer genes. We examined the relationship between identifying a geneset deregulated and the typical gene intensity, gene length, and number of genes within the 3243 GO-BP terms studied. For each quantity of interest, we computed a reference distribution, “All Pathways.” This is the distribution of the values for all pathways regardless of deregulation status. **(A) Gene length analysis.** We computed gene lengths using the UCSC Genome Browser (vers. hg19) by subtracting the end position from the start position. Then we summarized the gene length by finding the median of the log2 gene lengths within the pathway. Using **Dataset II (Table 1)**, the distributions of log2 median gene lengths of the deregulated pathways identified by both methods were found. Both methods are slightly more peaked and less variable than the reference distribution. There is not a bias towards longer genes as reported for some geneset tests (Young, et al., 2010). **(B) Gene intensity analysis.** A measure of pathway gene intensity was assigned to each GO-BP term by first averaging the gene expression from the 80 tumor samples in **Dataset II, Table 1**. Then find the median average gene expression to summarize the pathway gene intensity. Again we see that there is a tendency towards central intensity values in the pathways identified as deregulated in **Dataset II**, but do not observe a bias towards higher intensity as reported in other geneset methodology (Young, et al., 2010). **(C, D) Pathway size analysis.** The reference distribution is all GO-BP terms with 15 to 500 genes. **Panel C** depicts the comparison between pathways found deregulated in the TCGA breast cancer patients, **Dataset I**. We see that there is a tendency to identify larger pathways deregulated. To ensure that this is an artifact of the methodology and not reflecting true biology, we employed technical replicates of breast cancer samples captured by microarray, (MAQC-II, GEO: [GSE20194](#), Shi, et al., 2010). We treat pairs of technical replicates as “normal” and “tumor” samples. Any identified pathways would then be considered a false positive. Indeed, we find the distribution of identified pathways in technical replicates is biased towards larger genesets. **Legend: “All pathways” = reference distribution.**



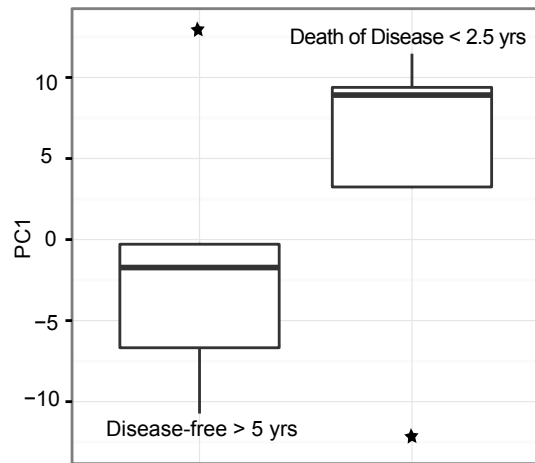
C

GO ID	Description
0008608	attachment of spindle microtubules to kinetochore
0006336	DNA replication-independent nucleosome assembly
0034080	CENP-A containing nucleosome assembly at centromere
0034724	DNA replication-independent nucleosome organization
0051313	attachment of spindle microtubules to chromosome
0034394	protein localization to cell surface
0055117	regulation of cardiac muscle contraction
0031055	chromatin remodeling at centromere
0051310	metaphase plate congression
0050000	chromosome localization

D

GO ID	Description
0000087	M phase of mitotic cell cycle
0048285	organelle fission
0000280	nuclear division
0007067	mitosis
0000279	M phase
0044057	regulation of system process
0000236	mitotic prometaphase
0001944	vasculature development
0000075	cell cycle checkpoint
0003012	muscle system process

Supplement Figure S3. Top ten scoring pathways predict breast cancer survival. (A) The ten deregulated pathways found from the 80 breast cancer patients in **Dataset II (Table 1)** with highest absolute MD scores were used to produce PAM clusters based on the MD scores. Restricting to only ten pathways maintains accurate survival prediction (log-rank p-value < 0.01). (B) Similarly, the top ten Wilcoxon-ranked pathway scores produced clusters with distinct survival curves. Panels C and D display the top ranked scoring pathways for MD and Wilcoxon approaches, respectively.



Supplement Figure S4. N-of-1-pathways MD GO-BP clinical importance metrics (CRMs) predict breast cancer patient survival. N-of-1-pathways MD was applied to n=80 invasive breast carcinoma patients (TCGA_BRCA, RNA-seq, **Table 1 dataset II**) resulting in 3225 clinical importance metrics. We present the first principal component for the 1344 MD CRMs corresponding to pathways found deregulated in at least one of the 14 diametric extreme phenotypes of: (i) death of disease in less than 2.5 years (n=5) and (ii) disease-free survival for more than four years (n=9). The boxplots display a trend for higher PC1 for the worst phenotype, but the outliers in each group muddle the association (Wilcoxon p-value > 0.02).

Supplement References

Shi, L., *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology* 2010;28(8):827-838.

Young, M.D., *et al.* Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11(2):R14.