Data supplement

"N-of-1-*pathways*" unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine

Vincent Gardeux, Ikbel Achour, Jianrong Li, Mark Maienschein-Cline, Haiquan Li, Lorenzo Pesce, Gurunadh Parinandi, Neil Bahroos, Robert Winn, Ian Foster, Joe G.N. Garcia, Yves A. Lussier

Supplement Methods

Gene Sets Enrichment Analysis of GO-BP (GSEA). Gene set enrichment analysis between normal and tumor samples was conducted on the exploration dataset as well as on the external validation datasets (studies II & III) using GSEA v2.0.10 software [1]. The default parameters were used, except the permutation parameter selection, which was set to "gene_set" instead of "phenotype". Gene set permutation was chosen to achieve enough statistical power for permutation resampling due to the small number of samples.

Differentially Expressed Genes enriched in GO-BP (DEG Enrichment). Enrichments of GO-BP genesets with differentially expressed genes (DEG) were conducted in the R statistical software using the Fisher's Exact Test (FET) based on the following contingency table: (DE genes, non-DE genes) X (In Pathway, Not in Pathway). Adjustment for multiple comparisons was performed using FDR, and pathways with FDR≤5% were considered significantly enriched. Of note, the up-regulated and down-regulated genes were enriched independently to generate significant "upregulated" and significant "downregualted" GO terms. DEGs were available from validation studies II and III except for study I for which neither the dataset nor DEGs were available. DEG of the exploration dataset was calculated in the following way: (i) genes whose average expression differs by at least a factor of four between normal and tumor samples were selected for analysis, (ii) then a non-parametric Wilcoxon test was applied between the two groups, and p-values were adjusted with Benjamini and Hochberg method (False Discovery Rate; FDR). Only DE genes with FDR≤5% were retained.

Single-Sample GSEA (ssGSEA). ssGSEA [2] is an extension of the GSEA method. It has two different modes of application: 1) "ssGSEA Projection" which is directly applied on single sample data and projects gene-level expressions to pathway-level scores, and 2) "ssGSEA Pre-ranked" which is applied on a pre-ranked list of genes and is able to compute a permutation-based p-value for each pathway of a geneset database. We used the "Pre-ranked" variant as a possible alternative of N-of-1-*pathways,* since it's the only ssGSEA variant providing geneset-level p-values. Here, we pre-ranked the genes according to their fold-change between normal and tumor samples, for every single patient. We used the "GseaPreranked" tool in GSEA v2.0.10 software [1] with default parameters. Of note, this software is rate-limiting, as it requires processing samples individually via a Java display interface. "ssGSEA Projection" variant (but not "ssGSEA Pre-Ranked") is also implemented in the R software and in GenePattern [3]. We did not find other enrichment-type methodologies specifically designed to provide p-values on paired samples analyses.

"Proxy" Gold Standard for the Internal Validation within the Exploration Dataset (Figure 2). In order to objectively assess the accuracy of the significantly deregulated mechanisms identified by N-of-1-*pathways* statistical component in the exploration dataset (**Methods: Table 1**), a gold standard comprising the true deregulated mechanisms should have been used. However, such a gold standard does not exist and published enrichment studies

that generate large lists of candidate mechanisms could not be thoroughly validated experimentally in their entirety because of the rate limiting nature and cost of such an endeavor. Nonetheless, since a sufficient subset of individual predictions of deregulated mechanisms from previously published enrichment and/or GSEA studies have been confirmed experimentally, we proceeded in using these two conventional enrichment methods as "proxy-gold standards". Specifically, GO-BPs were statistically prioritized by four above-mentioned methods: two established cohort-level ones (GSEA and DEG-Enrichment), an alternative single-sample one (ssGSEA), and the one we propose (N-of-1-*pathways*). Thus, the accuracy of the N-of-1-*pathways* could systematically be compared to one of the conventional methods (eg. DEG Enrichment) while the other serves as a proxy-gold standard (GSEA).

"External" Gold Standard derived from the External Validation Studies (Figures 3-5). Deregulated GO-BP terms in the three External validation studies served as External Gold Standards (GS) to evaluate the GO-BP of individual patients in the Exploration dataset. **Figure 3** used the GO-BP deregulated in each external study (FDR<5%) as three distinct External GS, while Figure 4 used the union of all deregulated GO-BP (FDR<5%) as one aggregated External GS. In studies II and III, significant GO-BPs were successively calculated by two previously described methods: GSEA and DEG-Enrichment. Of note, significant GO-BPs published in the supplementary table of External validation study I were utilized as the authors did not provide an original expression dataset.

Precision-Recall curves (Figures 2-3). Using the R statistical software, we computed two types of Precision-Recall curves: (i) internal validations (Figure 2) and (ii) external validations (Figure 3) of the GO-BP mechanism predicted by the N-of-1-*pathways* statistical analysis component (Cross-Patient; see above). *INTERNAL VALIDATION (**Figure 2**)*: Precision-recall curves of the "internal validation" compared the N-of-1-*pathways* predictions of GO-BP from the exploration dataset with the GO-BPs predicted on the same dataset by ssGSEA, GSEA and DEG-Enrichment. The latter two alternatively served as "Proxy Gold Standard". *EXTERNAL VALIDATION (**Figure 3**)*: The GO-BPs predicted in the exploration dataset from all three methods (N-of-1-*pathways*, GSEA and DEG-Enrichment) were compared to those obtained in each of the external datasets (considered as External Gold Standards). *STANDARD PRECISION-RECALL CURVE:* The Proxy Gold Standards GO-BPs were fixed, while each precision and recall point of each GO-BP prediction method was ranked either according to its p-values (GSEA and DEG-enrichment) or the number of patients (N-of-1-*pathways*). The precision and recall values were calculated using different cutoffs of the ranked GO-BPs from the prediction methods. In this case, a true positive was calculated as an overlap between a prediction and the gold standard. A true negative corresponded to a GO-BP neither predicted nor found in the Gold Standard. A false positive was a predicted GO-BP not found in the Gold Standard, and a false negative was a Gold Standard not predicted GO-BP. *INFORMATION-THEORY SIMILARITY IN PRECISION-RECALL CURVE:* in this type of precision-recall curve, we considered a true positive prediction if the predicted GO-BP was similar to a GO-BP from the Gold Standard or from the Proxy Gold Standard (GO-ITS $\geq$ 0.7). We previously showed that an GO-ITS score ≥ 0.7 robustly corresponded to highly similar GO terms using different computational biological validations: protein interaction [4, 5], human genetics [6], and Genome-Wide Association Studies [7].

Concordance of GO-BP Predicted in External Studies (Figure 3, Venn diagram). The concordance of predicted GO-BPs at FDR $\leq$ 5% between the three external studies was compared using the overlap drawn as a Venn Diagram. GO-BPs of external validation study I were taken directly from the manuscript, as the authors did not provide either deregulated genes or expression data (link broken). For studies II and III, significant GO-BPs were calculated by GSEA and DEG Enrichment adjusted at FDR $\leq$ 5%.

Gene Ontology annotations of Biological Processes (GO-BP). Hierarchical GO terms were retrieved using the *org.Hs.eg.db* package [8] of *Bioconductor* [9], available for R statistical software [10]. We used the *org.Hs.egGO2ALLEGS* database (downloaded on 03/15/2013), which contains a list of genes annotated to that GO term (*geneset*) along with all of its child nodes according to the hierarchical ontology structure. As stated in **Figure 1**, the genesets were filtered so that only those sized between 15 and 500 are kept in the studies.

Supplement Tables

Supplement Table S1. Subset of GO-BPs predicted by N-of-1-*pathways* in the exploration dataset that are unrelated (GO-ITS < 0.3) to the Gold Standard (GS) derived from the union of the three validation datasets (Methods: External GS). Some of these GO-BP mechanisms are common to up 10 patients, and thus might be relevant to lung adenocarcinoma, which were overlooked by conventional cross-patient enrichment studies. GO-BP individual mechanisms (unique to a patient) found in **Supp. Table S2** are colored in green, those not found in **Supp. Table S2** are colored in red, 4 out of 6 non-reproducible individual mechanisms are reported in both tables.

Curated classes	GO ID	GO Description	Max GO-ITS to GS	#Patients sharing this pathway
Immune response	GO:0031341	regulation of cell killing	0.20	1
	GO:0032640	tumor necrosis factor production	0.22	2
	GO:0032609	interferon-gamma production	0.22	2
	GO:0032635	interleukin-6 production	0.23	2
	GO:0071706	tumor necrosis factor superfamily cytokine	0.25	1
Metabolic process/ transport	GO:0006091	generation of precursor metabolites and energy	0.27	3
	GO:0006732	coenzyme metabolic process	0.27	2
	GO:0006790	sulfur compound metabolic process	0.28	1
	GO:0051186	cofactor metabolic process	0.29	3
Organ/tissue development	GO:0060021	palate development	0.25	1
	GO:0048771	tissue remodeling	0.25	1
Reproduction development	GO:0048610	cellular process involved in reproduction	0.11	10
	GO:0007283	spermatogenesis	0.24	3
	GO:0048232	male gamete generation	0.24	3
	GO:0007276	gamete generation	0.29	3
Not curated	GO:0032259	methylation	0.23	2
	GO:0008037	cell recognition	0.25	1

Supplement Table S2. GO-BP Terms predicted by N-of-1-*pathways* for only a single patient in RNA-Seq exploration dataset that are unrelated to the other patients (GO-ITS < 0.3). All the pathways found in this table were already found in **Supp. Table S1**.

Patient Id	GO ID	GO Description	Max GO-ITS to another patient
TCGA-44-2655	GO:0060021	palate development	0.246823639
TCGA-55-6972	GO:0048771	tissue remodeling	0.249372422
TCGA-55-6972	GO:0008037	cell recognition	0.250935749
TCGA-91-6836	GO:0006790	sulfur compound metabolic process	0.28389005

Supplement Table S3. GO-BP terms deregulated (from FAIME score perspective) for survival outcome. This table lists the GO-BP terms presented in **Supp. Figure S3** along with their curated classes and complete GO Description.

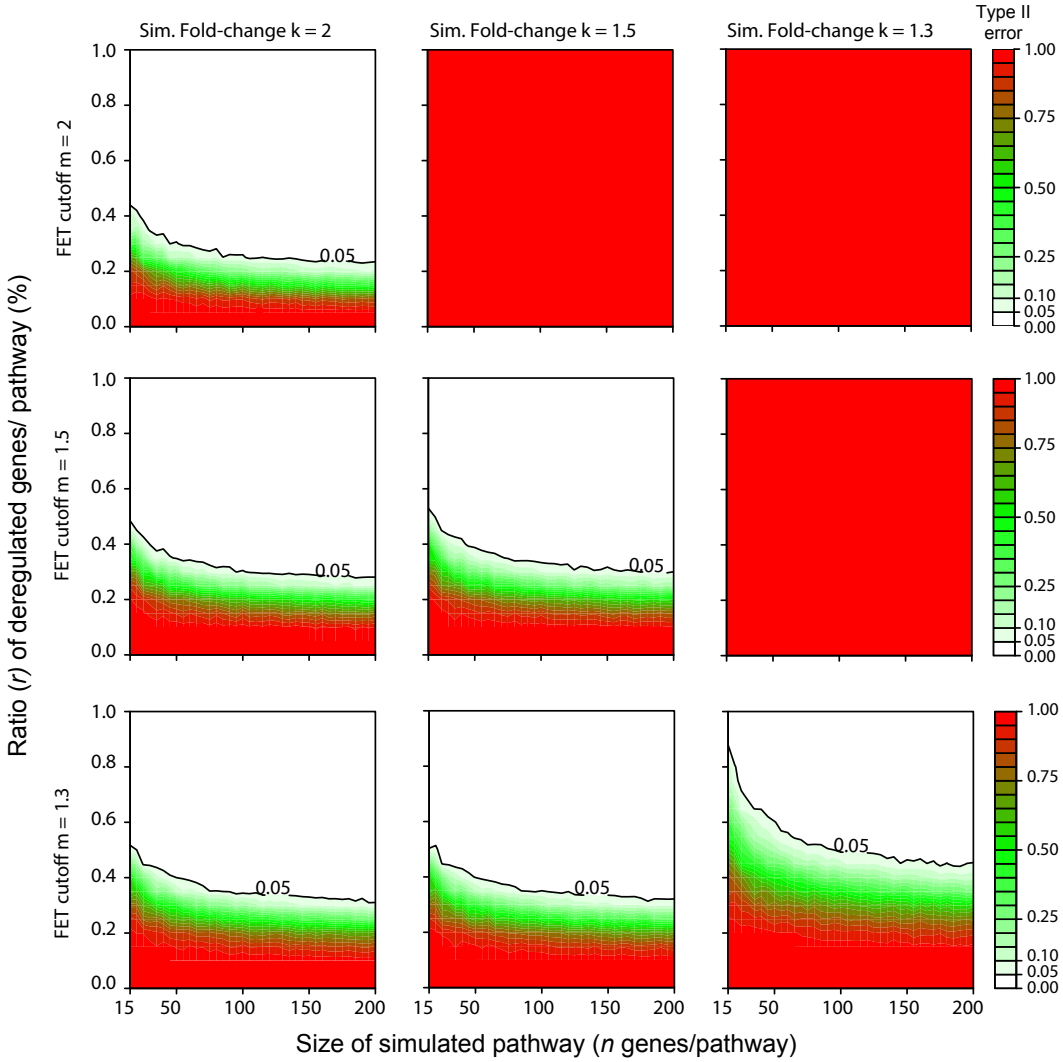
Curated classes	GO ID	GO Description
ND	GO:0015695	Organic cation transport
Chromosome localization	GO:0050000	Chromosome localization
	GO:0051303	Establishment of chromosome localization
ND	GO:0016079	Synaptic vesicle exocytosis
ND	GO:0015669	Gas transport
homeostasis	GO:0050891	Multicellular organismal water homeostasis
	GO:0003091	Renal water homeostasis
	GO:0055092	Sterol homeostasis
	GO:0042632	Cholesterol homeostasis
ND	GO:0006111	Regulation of gluconeogenesis
DNA/chromatin assembly	GO:0065004	Protein DNA complex.assembly
	GO:0043486	Histone exchange
	GO:0006323	DNA packaging
	GO:0030261	Chromosome condensation
Hormone secretion/transport	GO:0030072	Peptide hormone secretion
	GO:0090276	Regulation of peptide hormone secretion
	GO:0009914	Hormone transport
	GO:0032024	Positive regulation of insulin secretion

ND: Not Determined.

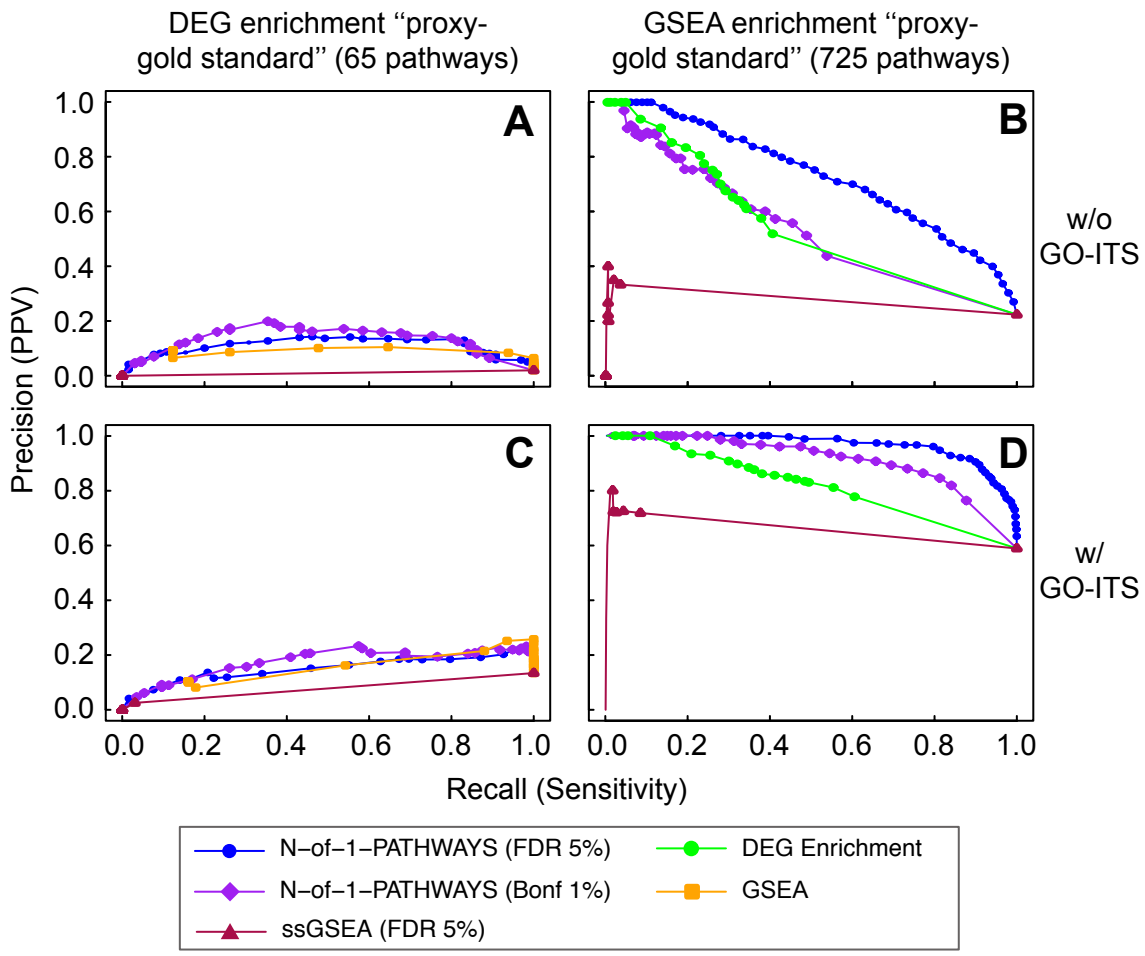
Supplement Table S4. Area Under Curve (AUC) for Precision-Recall curves shown in Figure 2 and Supp. Figure S2.

	GSEA as Proxy GS (Figure 2)		**DEG Enrichment as Proxy GS (Supp. Figure S2)**	
	Without Semantic Similarity	With Semantic Similarity	Without Semantic Similarity	With Semantic Similarity
N-of-1-*pathways* (FDR 5%)	**0.729**	**0.963**	0.109	0.151
N-of-1-*pathways* (Bonf. 1%)	0.532	0.904	**0.131**	**0.175**
ssGSEA (FDR 5%)	0.279	0.658	0.010	0.078
DEG Enrichment	0.541	0.813		
GSEA			0.081	0.140

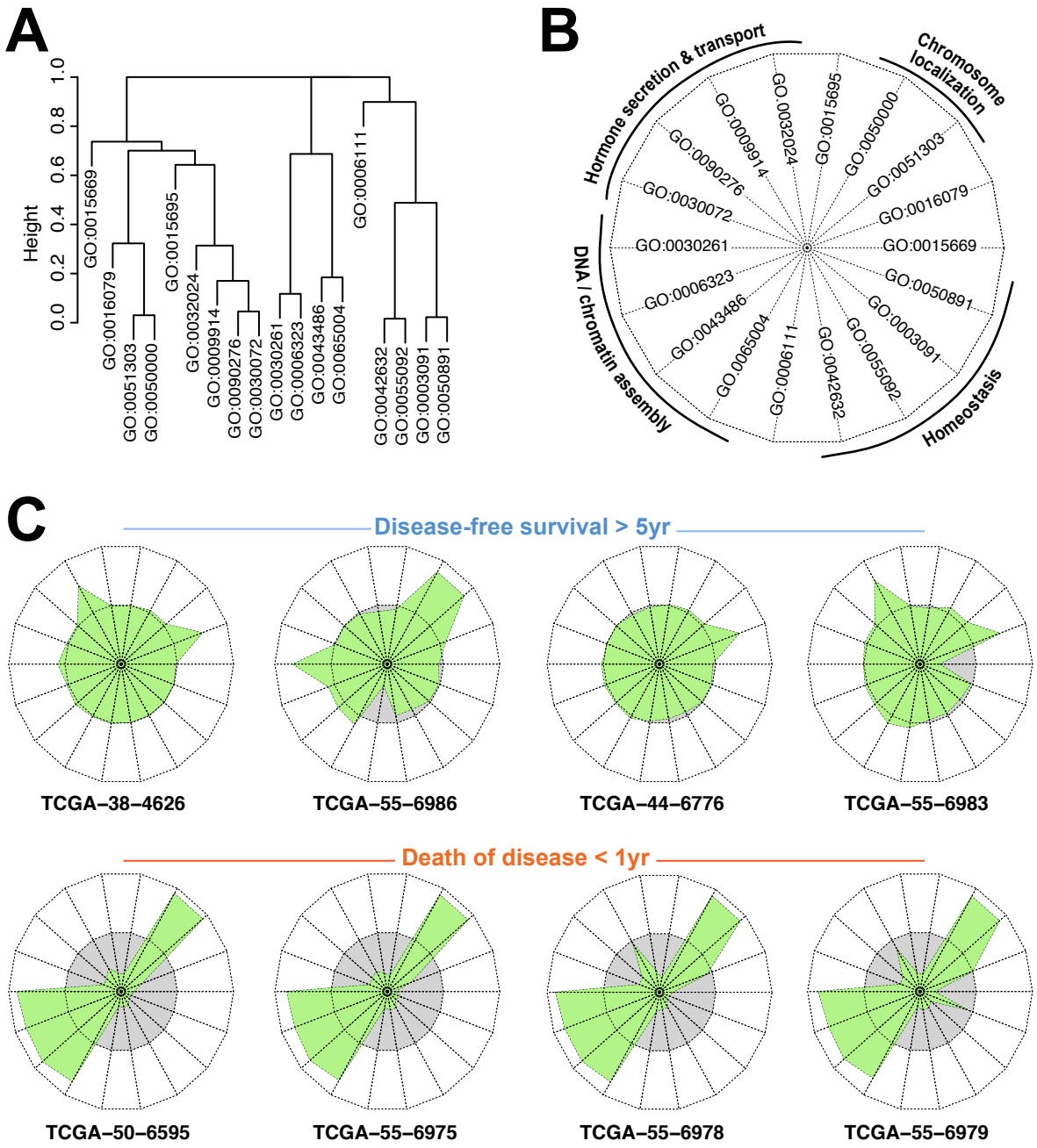
Supplement Figures



Supplement Figure S1. Evaluation of size and ratio of concordant deregulated genes within a pathway required to be found deregulated in the *Fisher-Exact Test Enrichment* performed on genes at a certain fold-change level (*k*). We tested an alternate approach to the Wilcoxon test, however as shown, it does not perform as well. Each point represents one size of a simulated pathway generated by randomly selecting *n* genes and a ratio *r* of the deregulated genes within the pathway. The ratio *r* is artificially increased by a *k*-fold change in a simulated pathway seeded in the exploration dataset ($k \in \{1.3, 1.5, 2\}$). We then applied the Fisher-Exact Test Enrichment (FET Enrichment) model. For each value (*n,r,k,m*), we repeated this procedure 1,000 times in order to estimate the false negative rate (type II error β). Of note, the utilized FET enrichment requires specifying a fold change cutoff "*m*", and differs from the conventional DEG Enrichment test in that differentially expressed genes cannot be calculated with a p-value between the two samples of exploration dataset (DEG requires two groups of n>2). Legend: "sim"=simulated (**Methods**). This Supplement Figure S1 illustrates that the N-of-1-*pathways* shown in **Figure 1** performs better in two ways: (i) the type II error is lower (less false negative), and (ii) it does not require the *m* fold change threshold to be specified. In other words, if the selected *k* required for performing FET Enrichment is higher than the threshold where the signal is visible, then the pathway is undetected by FET-Enrichment (e.g. panels all red above) while it is detected as significant by N-of-1-*pathways* **(Figure 1).**



Supplement Figure S2. Concordant deregulated pathways (genesets) between N-of-1-*pathways*, ssGSEA, DEG enrichment and GSEA methods. To evaluate the GO-BP associated terms yielded by the N-of-1-*pathways* method, we compared these pathways to those found by a single sample method: ssGSEA, and two well-established cohort-based methods: DEG enrichment and GSEA. We then generated precision-recall curves based on the perfect GO overlap (**Panels A and B;** noted without "w/o" GO-ITS), and GO semantic similarity overlap (**Panels C and D**; GO-ITS ≥ 0.7; **Methods: GO-ITS**). When GSEA is chosen as the Proxy Gold Standard (**Methods: Proxy GS**), N-of-1-*pathways* method uncovered deregulated pathways comparable, or better, to those of DEG enrichment analysis, with or without GO-ITS analysis respectively (**Panels C and D**). When DEG Enrichment is chosen as the Proxy Gold Standard, N-of-1-*pathways* performed marginally better than GSEA (**Panels A and C**).



Supplement Figure 3. Personal representation of top 18 survival-related deregulated mechanisms found by FAIME scores (Methods: Star plot). **Panel A** is the dendrogram representation of the clustering of the 18 GO Terms by GO-ITS similarity. **Panel B** is the legend of the star plots, each edge corresponding to one GO Term. Each cluster has been manually curated to a representative GO-BP category (**Supp. Table S3**). **Panel C** contains each extreme patient's own star plot representation of the 18 GO Terms. The green zone represents up-regulated pathways (given the fold change direction of FAIME scores), while the grey zone stands for down-regulation. The non-deregulated zone (Z-score = 0) is represented by a dotted line splitting the two color zones.

References

1 Subramanian A, Tamayo P, Mootha V, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;**102**:15545 - 15550.
2 Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009 Nov 5;**462**(7269):108-112.
3 Reich M, Liefeld T, Gould J, et al. GenePattern 2.0. *Nature genetics*. 2006 May;**38**(5):500-501.
4 Li H, Lee Y, Chen JL, et al. Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. *Journal of the American Medical Informatics Association : JAMIA*. 2012 Mar-Apr;**19**(2):295-305.
5 Tao Y, Sam L, Li J, et al. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*. 2007 Jul 1;**23**(13):i529-538.
6 Regan K, Wang K, Doughty E, et al. Translating Mendelian and complex inheritance of Alzheimer's disease genes for predicting unique personal genome variants. *Journal of the American Medical Informatics Association : JAMIA*. 2012 Mar-Apr;**19**(2):306-316.
7 Lee Y, Li J, Gamazon E, et al. Biomolecular Systems of Disease Buried Across Multiple GWAS Unveiled by Information Theory and Ontology. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*. 2010;**2010**:31-35.
8 Carlson M. org.Hs.eg.db: Genome wide annotation for Human.
9 Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*. 2004;**5**(10):R80.
10 R: Development core team. R: A language and enviroment for statistical computing.: R foundation for statistical computing. Vienna, Austria.; 2004.