Supplementary Data for "pyNVR: Investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction"

Bob Chen[1,2], Charles A. Herring[1,3], Ken S. Lau[1,2,3]

[1] Epithelial Biology Center, Vanderbilt University Medical Center, Nashville, TN 37232
[2] Program in Chemical and Physical Biology, Vanderbilt University School of Medicine, Nashville, TN 37232
[3] Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN 37232

Corresponding author:
Ken S. Lau
Epithelial Biology Center
Vanderbilt University Medical Center
2213 Garland Ave
10475 MRB IV
Nashville, TN 37232-0441
email: ken.s.lau@vanderbilt.edu

Supplementary Method S1.1 NVR

This algorithm (Welch *et al.*, 2016) generates a connected graph based on the Euclidean distances of cell to cell gene expression. Based on this graph, the algorithm compares the variance of gene expression within neighborhoods and the variance of gene expression globally on a cell to cell basis. It then assumes that if the neighborhood variance is lower than the global variance, there exists some meaningful and controlled gene expression. The formalization of this neighborhood variance, in the context of genes, is described as follows, where n is the sample number, k_c is the minimum number of neighbors in the connected graph, g is the gene of interest, and N(i , j) is the nearest neighbor j of the sample i:

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} \left(e_{ig} - e_{N(i,j)g} \right)^2$$

An example of this phenomenon would be the expression of some gene that changes monotonically along the progression of a given developmental lineage. Neighborhood variation would be low given the gradual change of gene expression, and global variation would be higher given the differences in expression between end states of a transition. Due to the calculation of neighborhood variance, the time complexity of this algorithm is O(n) where n is the product of the number of cells and the number of genes. The following is the pseudocode for the algorithm:

Determine the minimum number of connections, k, that will generate a connected graph.
    Calculate the pairwise distances between each element of the input matrix
    Convert this vector into squareform
    Generate an adjacency matrix based on this squareform
    Permit k number of connections and generate a graph based on the adjacency matrix
        Count the number of connected components, c
        If C>1, add 1 to k and repeat until C=1
Use this number of connections, k, to generate a connected graph
For each gene, calculate the mean variance of some n neighbors based on the generated graph
    Repeat for all possible neighborhoods
    Calculate the mean of this neighborhood variance
For each gene, calculate the global variance in the context of all cells
If the global variance of a gene divided by the average neighborhood variance of that same gene is greater than 1, select that gene.

Supplementary Method S1.2 dpFeature (dpF)

This feature selection method, developed by the Trapnell group (Qiu *et al.*, 2017), utilizes density peak clustering (Rodriguez and Laio, 2014) on a t-SNE (t-distributed stochastic neighbor embedding) dimension-reduced representation of transcriptomic data (Van Der Maaten and Hinton, 2008). For t-SNE, our study used the monocle R package using the parameters max_components=2, num_dim=6, and check_duplicates=FALSE. Using this representation of the data, density peak clustering was performed. A generalized linear model was then used to test for the most significantly differentially expressed genes between clusters. As this model calculates the significance for every gene and does not output a discrete number of genes like NVR, we selected the n most significant genes with respect to q-value. For the sake of set similarity calculations, this n is simply the number of genes that NVR selected.

Supplementary Method S1.3 Closeness Thresholding and Resampling

This resampling method is detailed in Herring et al. (Herring *et al.*, 2018) and involves three primary steps: Down-sampling, density-based k-NN construction, and closeness thresholding. Down-sampling was performed on datasets with dimensionality reduced by PCA (Principal Component Analysis) to first normalize rare versus common events. The down-sampling procedure takes into consideration the local density of each cell, given a user determined metrics: space radius, target noise, and target cell number. An undirected density-based k-NN graph was then generated using the down-sampled dataset. This graph's weighted edges were calculated as a product of node Euclidean distances and their minimum local density values. Given this connected graph, node closeness metrics were calculated by taking the normalized mean graph distance from a node x to all other nodes y in the graph given N nodes. This distance is the shortest path determined by Dijkstra's algorithm. Resampling of the data involved setting a closeness threshold and randomly selecting a constant number of cells that satisfy this threshold.

$$C(x) = \frac{N}{\sum_y d(y,x)}$$

Supplementary Method S1.4 Jaccard Index calculations

Set similarities were calculated based on the Jaccard index (Levandowsky and Winter, 1971) formally defined below where A and B are sets of genes selected by distinct algorithms such as NVR or dpFeature:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Set similarities, in the context of algorithm robustness, *r*, were calculated based on the Jaccard index given the subset *a*:

$$J_r(A,a) = \frac{|A \cap a|}{|A \cup a|} \text{ where A} \supset a$$

Supplementary Method S1.5 p-Creode

p-Creode is an unsupervised trajectory reconstruction algorithm which utilizes the hierarchical placement of putative cell states to organize state transition trajectories (Herring *et al.*, 2018). It incorporates a graph dissimilarity scoring metric built upon the Gromov-Hausdorff distance. p-Creode assesses trajectories generated from resampled datasets using this metric to identify the most representative graph topology. Given these trajectories are represented as graphs, we overlaid heatmaps of gene expression across the representative nodes. We used the Python package, https://github.com/KenLauLab/pCreode, to perform this analysis.

Supplementary Method S1.6 Gene Ontology Term Enrichment through WebGestalt

As described by the Zhang group, WebGestalt is a web-based platform for gene ontology term enrichment analysis (Wang *et al.*, 2017). We performed overrepresentation enrichment analysis because of the nature of the outputs from our feature selection algorithms, were lists of genes without expression values. Below is an overview of the parameters we used for our analysis:

Parameter	**Value**
Select Organism of Interest	Mmusculus
Select Method of Interest	Overrepresentation Enrichment Analysis
Select Functional Database	geneontology/Biological_Process_noRedundant
Select Gene ID Type	Genesymbol
Upload Gene List	.txt file of genes selected by algorithms
Select Reference Set for Enrichment Analysis	Genome
All advanced parameters	default values

The returned values using these parameters are detailed in Supplementary Figure 4 and Supplementary Table 4. The following is a description of the abbreviated column names:

Column	**Description**
C	The number of reference genes in this category
O	The number of genes in the user gene list and also in the category
E	The expected number in the category
R	Ratio of enrichment (Observed/Expected)
Pvalue	P value from hypergeometric test
FDR	False discovery rate generated from the Benjamini-Hochberg (BH) procedure

Supplementary Method S1.7 findVariableGenes (FVG)

findVariableGenes is a feature selection algorithm included as part of the Seurat R package (Butler *et al.*, 2018). First, it calculates a normalized measure of gene expression, taking mean expression and dispersion into account. The genes are then binned (Bins=20). Finally, z-scores for dispersion are calculated given these bins, and the top N genes are returned based on these scores. We used an N equivalent to the number of genes selected by NVR for our similarity analyses.

Supplementary Method S1.8 PCA-based Feature Extraction (PCAFE)

Principal component analysis (PCA)-based unsupervised feature extraction (FE) is a another method used to select biologically relevant genes (Taguchi, 2018). This method starts by scaling the raw count data and performing a principal component analysis. For the first three principal components, the gene weights are then scaled and summed. These sums are used for a Chi-squared test. Finally, an adjusted p-value threshold is set and genes that meet that threshold are selected. Although the provided examples of this method are dependent on a direct adjusted p-value threshold, we modified the method to return the top N genes based on the most significant adjusted p-values.

Supplementary Table S1. Native sequencing datasets used

Dataset	Biological Context	scRNA-seq Platform	Cell #	Transcripts/Genes
GSE 60781	Flow sorted dendritic cells collected from mouse bone marrow	Fluidigm C1	251	29779
GSE 52529	Human skeletal muscle myoblasts collected over 4 days in a differentiation time course	Fluidigm C1	271	47192
GSE 102698	Dissociated epithelial cells collected from mouse colon	inDrop	1597	25507

Supplementary Table S2. Implementation Runtime Measurements

Dataset	Computer	Method	Time (h)	Cell Count	Genes	Selected
GSE72857	1	R	17.14	4000	27297	379
GSE72857	1	Py	1.38	4000	27297	379
GSE72857	2	R	29.25	4000	27297	379
GSE72857	2	Py	1.74	4000	27297	379
GSE102698	1	R	4.86	1597	25507	529
GSE102698	1	Py	0.47	1597	25507	529
GSE102698	2	R	8.74	1597	25507	529
GSE102698	2	Py	0.57	1597	25507	529
GSE60781	1	R	0.033833	251	29779	100
GSE60781	1	Py	0.075552	251	29779	100
GSE52529	1	R	0.071078	271	47192	318
GSE52529	1	Py	0.130833	271	47192	318
s1_pan	1	R	0.886138	329	25494	74
s1_pan	1	Py	0.091389	329	25494	74
s1_pan	2	R	1.877549	329	25494	74
s1_pan	2	Py	0.109365	329	25494	74
s2_pan	1	R	1.745	614	25494	56
s2_pan	1	Py	0.170409	614	25494	56
s2_pan	2	R	3.619634	614	25494	56
s2_pan	2	Py	0.205921	614	25494	56

Supplementary Table S3. Linear Regressions

Coefficients						
Methods	**Dataset**	**Coefficient**	**Estimate**	**Std. Error**	**T value**	**Pr(>|t|)**
dpF + FVG	Closeness	Intercept	0.221254	0.005550	39.865	< 2e-16
dpF + FVG	Closeness	Slope	0.045821	0.005307	8.634	1.63e-08
dpF + FVG	Size	Intercept	0.279249	0.006737	41.448	< 2e-16
dpF + FVG	Size	Slope	0.058447	0.010396	5.622	8.9e-07
NVR + dpF	Closeness	Intercept	0.41371	0.01540	26.871	< 2e-16
NVR + dpF	Closeness	Slope	0.09717	0.01472	6.601	1.23e-06
NVR + dpF	Size	Intercept	0.558264	0.007767	71.88	<2e-16
NVR + dpF	Size	Slope	0.159985	0.011985	13.35	<2e-16
NVR + FVG	Closeness	Intercept	0.174128	0.008139	21.395	3.24e-16
NVR + FVG	Closeness	Slope	0.065332	0.007782	8.395	2.63e-08
NVR + FVG	Size	Intercept	0.24727	0.00833	29.684	< 2e-16
NVR + FVG	Size	Slope	0.11066	0.01285	8.609	2.27e-11
NVR + PCAFE	Closeness	Intercept	0.16849	0.01270	13.264	5.68e-12
NVR + PCAFE	Closeness	Slope	0.07862	0.01215	6.473	1.64e-06
NVR + PCAFE	Size	Intercept	0.272186	0.004610	59.042	<2e-16
NVR + PCAFE	Size	Slope	0.012361	0.007113	1.738	0.0886

Summary						
Methods	**Dataset**	**Residual Standard Error**	**Multiple RSquared**	**Adjusted RSquared**	**F-Statistic**	**p-value**
dpF + FVG	Closeness	0.01489 on 22 degrees of freedom	0.7721	0.7618	74.55 on 1 and 22 DF	1.626e-08
dpF + FVG	Size	0.01819 on 49 degrees of freedom	0.3921	0.3797	31.61 on 1 and 49 DF	8.895e-07
NVR + dpF	Closeness	0.04131 on 22 degrees of freedom	0.6645	0.6492	43.57 on 1 and 22 DF	1.227e-06
NVR + dpF	Size	0.02097 on 49 degrees of freedom	0.7843	0.7799	178.2 on 1 and 49 DF	< 2.2e-16
NVR + FVG	Closeness	0.02184 on 22 degrees of freedom	0.7621	0.7513	70.48 on 1 and 22 DF	2.627e-08
NVR + FVG	Size	0.02248 on 49 degrees of freedom	0.602	0.5939	74.12 on 1 and 49 DF	2.272e-11
NVR + PCAFE	Closeness	0.03409 on 22 degrees of freedom	0.6557	0.64	41.89 on 1 and 22 DF	1.641e-06
NVR + PCAFE	Size	0.01244 on 49 degrees of freedom	0.05805	0.03882	3.02 on 1 and 49 DF	0.08855

Supplementary Table S4. Gene Ontology Term Enrichment

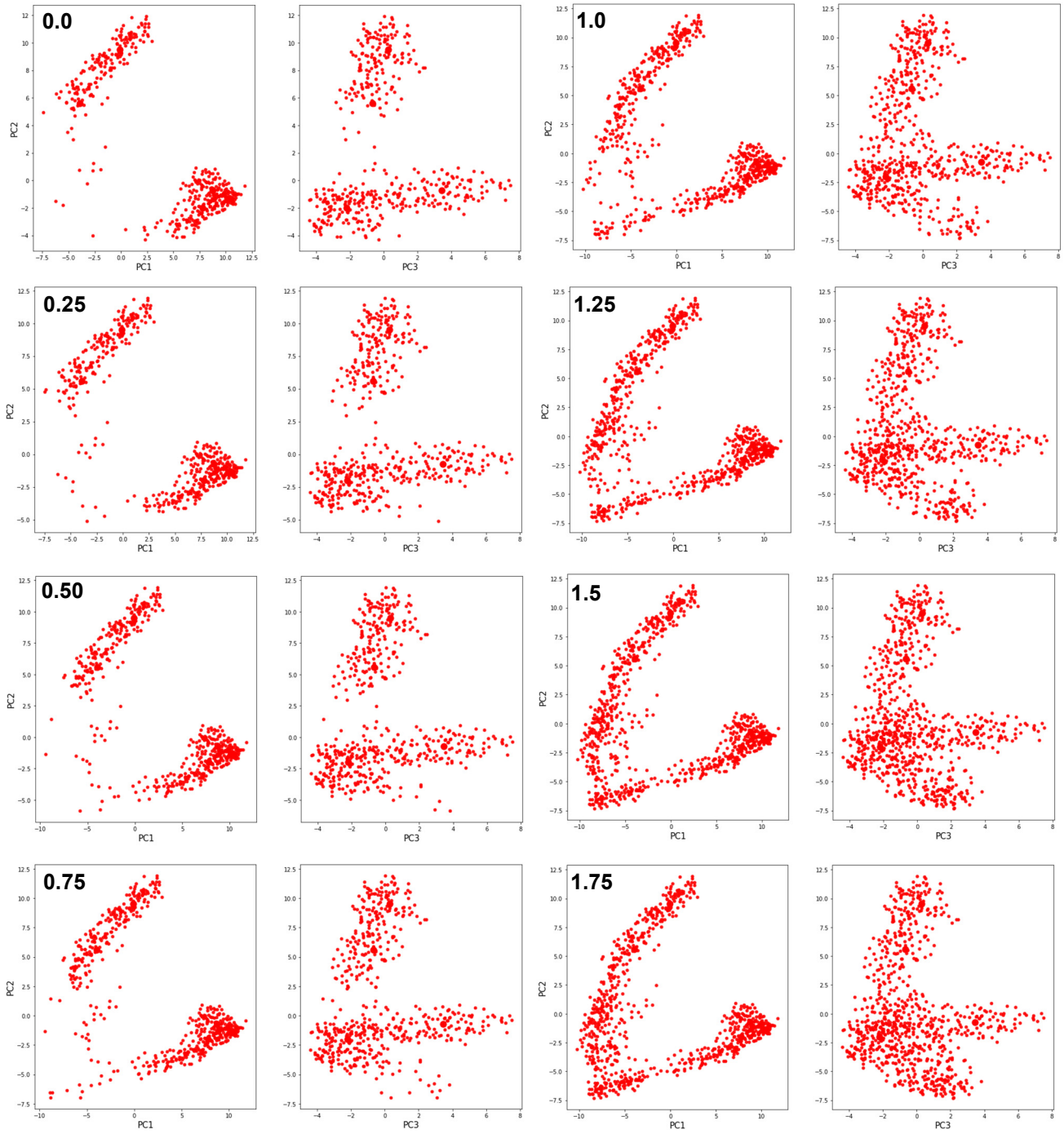
Feature Selection	Dataset	ID GO	GO Category	C	O	E	R	PValue	FDR
dpF	Native GSE102698	0022613	ribonucleoprotein complex biogenesis	358	32	8.17	3.92	3.76E-11	2.69E-08
dpF	Native GSE102698	0045454	cell redox homeostasis	64	12	1.46	8.21	1.84E-08	6.55E-06
dpF	Native GSE102698	0002181	cytoplasmic translation	49	10	1.12	8.94	1.24E-07	2.79E-05
dpF	Native GSE102698	0071826	ribonucleoprotein complex subunit organization	180	18	4.11	4.38	1.56E-07	2.79E-05
dpF	Native GSE102698	0006457	protein folding	165	14	3.77	3.72	2.53E-05	3.62E-03
dpF	Native GSE102698	0009636	response to toxic substance	201	14	4.59	3.05	2.14E-04	2.54E-02
dpF	Native GSE102698	0006575	cellular modified amino acid metabolic process	139	11	3.17	3.47	3.43E-04	3.49E-02
dpF	Native GSE102698	0042493	response to drug	380	20	8.68	2.31	4.66E-04	4.16E-02
dpF	Native GSE102698	0034660	ncRNA metabolic process	415	21	9.47	2.22	5.66E-04	4.47E-02
dpF	Native GSE102698	0018196	peptidyl-asparagine modification	31	5	0.71	7.07	6.26E-04	4.47E-02
FVG	Native GSE102698	0002237	response to molecule of bacterial origin	315	24	5.59	4.29	1.87E-09	1.33E-06
FVG	Native GSE102698	0098542	defense response to other organism	434	27	7.7	3.51	1.41E-08	5.02E-06
FVG	Native GSE102698	0071216	cellular response to biotic stimulus	184	15	3.26	4.6	9.74E-07	1.93E-04
FVG	Native GSE102698	0010035	response to inorganic substance	441	24	7.82	3.07	1.08E-06	1.93E-04
FVG	Native GSE102698	0046683	response to organophosphorus	114	11	2.02	5.44	5.60E-06	8.00E-04
FVG	Native GSE102698	0014074	response to purine-containing compound	131	11	2.32	4.73	2.12E-05	2.52E-03
FVG	Native GSE102698	0034341	response to interferon-gamma	88	9	1.56	5.77	2.53E-05	2.58E-03
FVG	Native GSE102698	0031347	regulation of defense response	456	21	8.09	2.6	6.34E-05	5.31E-03
FVG	Native GSE102698	0042044	fluid transport	25	5	0.44	11.27	6.69E-05	5.31E-03
FVG	Native GSE102698	0072348	sulfur compound transport	29	5	0.51	9.72	1.41E-04	1.01E-02
NVR	Native GSE102698	0022613	ribonucleoprotein complex biogenesis	358	30	7.99	3.76	4.36E-10	3.11E-07
NVR	Native GSE102698	0002181	cytoplasmic translation	49	11	1.09	10.06	7.76E-09	2.77E-06
NVR	Native GSE102698	0045454	cell redox homeostasis	64	11	1.43	7.71	1.47E-07	3.50E-05
NVR	Native GSE102698	0071826	ribonucleoprotein complex subunit organization	180	17	4.01	4.23	5.66E-07	1.01E-04
NVR	Native GSE102698	0006457	protein folding	165	16	3.68	4.35	8.59E-07	1.23E-04
NVR	Native GSE102698	0043900	regulation of multi-organism process	228	16	5.09	3.15	5.32E-05	6.33E-03
NVR	Native GSE102698	0070670	response to interleukin-4	32	6	0.71	8.41	6.51E-05	6.64E-03
NVR	Native GSE102698	0034976	response to endoplasmic reticulum stress	205	14	4.57	3.06	2.06E-04	1.73E-02
NVR	Native GSE102698	0006818	hydrogen transport	135	11	3.01	3.65	2.18E-04	1.73E-02
NVR	Native GSE102698	0044419	interspecies interaction between organisms	296	17	6.6	2.57	3.56E-04	2.54E-02
PCAFE	Native GSE102698	0022613	ribonucleoprotein complex biogenesis	358	28	5.6	5	1.89E-12	1.35E-09
PCAFE	Native GSE102698	0002181	cytoplasmic translation	49	12	0.77	15.65	8.69E-12	3.10E-09
PCAFE	Native GSE102698	0006818	hydrogen transport	135	17	2.11	8.05	3.34E-11	7.95E-09
PCAFE	Native GSE102698	0009123	nucleoside monophosphate metabolic process	247	19	3.86	4.92	1.11E-08	1.98E-06
PCAFE	Native GSE102698	0071826	ribonucleoprotein complex subunit organization	180	15	2.82	5.33	1.46E-07	1.74E-05
PCAFE	Native GSE102698	0009141	nucleoside triphosphate metabolic process	233	17	3.64	4.66	1.46E-07	1.74E-05
PCAFE	Native GSE102698	0019693	ribose phosphate metabolic process	435	23	6.81	3.38	3.30E-07	3.37E-05
PCAFE	Native GSE102698	0015672	monovalent inorganic cation transport	460	23	7.2	3.2	8.75E-07	7.81E-05
PCAFE	Native GSE102698	0006414	translational elongation	49	7	0.77	9.13	1.02E-05	8.09E-04
PCAFE	Native GSE102698	1901657	glycosyl compound metabolic process	323	17	5.05	3.36	1.30E-05	8.68E-04

Feature Selection	Dataset	ID GO	GO Category	C	O	E	R	PValue	FDR
dpF	Closeness 0 GSE102698	0006820	anion transport	475	20	6.51	3.07	8.11E-06	3.64E-03
dpF	Closeness 0 GSE102698	0055067	monovalent inorganic cation homeostasis	128	10	1.75	5.7	1.02E-05	3.64E-03
dpF	Closeness 0 GSE102698	0006818	hydrogen transport	135	10	1.85	5.41	1.63E-05	3.88E-03
dpF	Closeness 0 GSE102698	0015672	monovalent inorganic cation transport	460	18	6.3	2.86	6.14E-05	1.10E-02
dpF	Closeness 0 GSE102698	0006575	cellular modified amino acid metabolic process	139	9	1.9	4.73	1.24E-04	1.77E-02
dpF	Closeness 0 GSE102698	0070085	glycosylation	247	12	3.38	3.55	1.53E-04	1.82E-02
dpF	Closeness 0 GSE102698	0043270	positive regulation of ion transport	208	10	2.85	3.51	5.89E-04	6.01E-02
dpF	Closeness 0 GSE102698	0009100	glycoprotein metabolic process	338	13	4.63	2.81	7.79E-04	6.95E-02
dpF	Closeness 0 GSE102698	0045454	cell redox homeostasis	64	5	0.88	5.7	1.81E-03	1.35E-01
dpF	Closeness 0 GSE102698	0072521	purine-containing compound metabolic process	466	15	6.38	2.35	1.89E-03	1.35E-01
FVG	Closeness 0 GSE102698	0098542	defense response to other organism	434	18	4.22	4.26	2.05E-07	1.45E-04
FVG	Closeness 0 GSE102698	0002237	response to molecule of bacterial origin	315	15	3.07	4.89	4.07E-07	1.45E-04
FVG	Closeness 0 GSE102698	0071216	cellular response to biotic stimulus	184	11	1.79	6.14	1.76E-06	4.18E-04
FVG	Closeness 0 GSE102698	0002440	production of molecular mediator of immune response	158	8	1.54	5.2	1.54E-04	2.29E-02
FVG	Closeness 0 GSE102698	0019221	cytokine-mediated signaling pathway	351	12	3.42	3.51	1.60E-04	2.29E-02
FVG	Closeness 0 GSE102698	0034341	response to interferon-gamma	88	6	0.86	7.01	2.13E-04	2.54E-02
FVG	Closeness 0 GSE102698	0010035	response to inorganic substance	441	13	4.29	3.03	3.62E-04	3.70E-02
FVG	Closeness 0 GSE102698	0019932	second-messenger-mediated signaling	195	8	1.9	4.22	6.35E-04	5.67E-02
FVG	Closeness 0 GSE102698	0035821	modification of morphology or physiology of other organism	79	5	0.77	6.5	1.02E-03	7.70E-02
FVG	Closeness 0 GSE102698	0002697	regulation of immune effector process	273	9	2.66	3.39	1.39E-03	7.70E-02
NVR	Closeness 0 GSE102698	0002181	cytoplasmic translation	49	8	0.66	12.12	2.62E-07	1.87E-04
NVR	Closeness 0 GSE102698	0006818	hydrogen transport	135	10	1.82	5.5	1.41E-05	5.05E-03
NVR	Closeness 0 GSE102698	0015672	monovalent inorganic cation transport	460	18	6.2	2.9	4.94E-05	1.18E-02
NVR	Closeness 0 GSE102698	0022613	ribonucleoprotein complex biogenesis	358	15	4.82	3.11	1.01E-04	1.81E-02
NVR	Closeness 0 GSE102698	0070085	glycosylation	247	12	3.33	3.61	1.31E-04	1.87E-02
NVR	Closeness 0 GSE102698	0009100	glycoprotein metabolic process	338	14	4.55	3.07	1.95E-04	2.20E-02
NVR	Closeness 0 GSE102698	0045454	cell redox homeostasis	64	6	0.86	6.96	2.16E-04	2.20E-02
NVR	Closeness 0 GSE102698	0042493	response to drug	380	14	5.12	2.73	6.36E-04	5.68E-02
NVR	Closeness 0 GSE102698	0055067	monovalent inorganic cation homeostasis	128	7	1.72	4.06	1.73E-03	1.26E-01
NVR	Closeness 0 GSE102698	0006457	protein folding	165	8	2.22	3.6	1.77E-03	1.26E-01
PCAFE	Closeness 0 GSE102698	0006818	hydrogen transport	135	10	1.07	9.34	1.11E-07	7.93E-05
PCAFE	Closeness 0 GSE102698	0019693	ribose phosphate metabolic process	435	15	3.45	4.35	1.66E-06	5.91E-04
PCAFE	Closeness 0 GSE102698	0009123	nucleoside monophosphate metabolic process	247	11	1.96	5.61	4.08E-06	9.72E-04
PCAFE	Closeness 0 GSE102698	0009141	nucleoside triphosphate metabolic process	233	10	1.85	5.41	1.58E-05	2.60E-03
PCAFE	Closeness 0 GSE102698	0072521	purine-containing compound metabolic process	466	14	3.7	3.79	1.82E-05	2.60E-03
PCAFE	Closeness 0 GSE102698	0006091	generation of precursor metabolites and energy	304	11	2.41	4.56	2.88E-05	3.42E-03
PCAFE	Closeness 0 GSE102698	0050878	regulation of body fluid levels	317	11	2.52	4.37	4.22E-05	4.30E-03
PCAFE	Closeness 0 GSE102698	1901657	glycosyl compound metabolic process	323	10	2.56	3.9	2.43E-04	2.01E-02
PCAFE	Closeness 0 GSE102698	0034109	homotypic cell-cell adhesion	73	5	0.58	8.63	2.80E-04	2.01E-02
PCAFE	Closeness 0 GSE102698	0015672	monovalent inorganic cation transport	460	12	3.65	3.29	2.82E-04	2.01E-02

Feature Selection	Dataset	ID GO	GO Category	C	O	E	R	PValue	FDR
dpF	Closeness 1.75 GSE102698	0022613	ribonucleoprotein complex biogenesis	358	31	7.58	4.09	2.55E-11	1.82E-08
dpF	Closeness 1.75 GSE102698	0002181	cytoplasmic translation	49	11	1.04	10.6	4.53E-09	1.62E-06
dpF	Closeness 1.75 GSE102698	0071826	ribonucleoprotein complex subunit organization	180	17	3.81	4.46	2.74E-07	6.51E-05
dpF	Closeness 1.75 GSE102698	0045454	cell redox homeostasis	64	10	1.36	7.38	8.57E-07	1.53E-04
dpF	Closeness 1.75 GSE102698	0006457	protein folding	165	14	3.5	4.01	1.10E-05	1.57E-03
dpF	Closeness 1.75 GSE102698	0070085	glycosylation	247	16	5.23	3.06	7.48E-05	8.90E-03
dpF	Closeness 1.75 GSE102698	0032528	microvillus organization	23	5	0.49	10.26	1.01E-04	1.03E-02
dpF	Closeness 1.75 GSE102698	0006575	cellular modified amino acid metabolic process	139	11	2.94	3.74	1.80E-04	1.61E-02
dpF	Closeness 1.75 GSE102698	0009100	glycoprotein metabolic process	338	18	7.16	2.51	3.20E-04	2.54E-02
dpF	Closeness 1.75 GSE102698	0018196	peptidyl-asparagine modification	31	5	0.66	7.61	4.45E-04	3.18E-02
FVG	Closeness 1.75 GSE102698	0046683	response to organophosphorus	114	7	0.93	7.53	3.98E-05	2.84E-02
FVG	Closeness 1.75 GSE102698	0014074	response to purine-containing compound	131	7	1.07	6.55	9.64E-05	3.44E-02
FVG	Closeness 1.75 GSE102698	0010035	response to inorganic substance	441	12	3.6	3.34	2.49E-04	5.92E-02
FVG	Closeness 1.75 GSE102698	0098542	defense response to other organism	434	11	3.54	3.11	8.22E-04	1.47E-01
FVG	Closeness 1.75 GSE102698	0042493	response to drug	380	10	3.1	3.23	1.08E-03	1.54E-01
FVG	Closeness 1.75 GSE102698	0061614	pri-miRNA transcription from RNA polymerase II promoter	29	3	0.24	12.68	1.65E-03	1.97E-01
FVG	Closeness 1.75 GSE102698	0070670	response to interleukin-4	32	3	0.26	11.49	2.21E-03	2.25E-01
FVG	Closeness 1.75 GSE102698	0055067	monovalent inorganic cation homeostasis	128	5	1.04	4.79	3.94E-03	3.31E-01
FVG	Closeness 1.75 GSE102698	0002237	response to molecule of bacterial origin	315	8	2.57	3.11	4.23E-03	3.31E-01
FVG	Closeness 1.75 GSE102698	0034341	response to interferon-gamma	88	4	0.72	5.57	5.78E-03	3.31E-01
NVR	Closeness 1.75 GSE102698	0022613	ribonucleoprotein complex biogenesis	358	30	6.14	4.89	5.37E-13	3.83E-10
NVR	Closeness 1.75 GSE102698	0002181	cytoplasmic translation	49	12	0.84	14.29	2.54E-11	9.08E-09
NVR	Closeness 1.75 GSE102698	0071826	ribonucleoprotein complex subunit organization	180	16	3.09	5.19	7.98E-08	1.90E-05
NVR	Closeness 1.75 GSE102698	0045454	cell redox homeostasis	64	8	1.1	7.29	1.28E-05	2.28E-03
NVR	Closeness 1.75 GSE102698	0006414	translational elongation	49	7	0.84	8.33	1.85E-05	2.64E-03
NVR	Closeness 1.75 GSE102698	0034660	ncRNA metabolic process	415	20	7.11	2.81	3.12E-05	3.72E-03
NVR	Closeness 1.75 GSE102698	0006457	protein folding	165	11	2.83	3.89	1.28E-04	1.24E-02
NVR	Closeness 1.75 GSE102698	0006575	cellular modified amino acid metabolic process	139	10	2.38	4.2	1.39E-04	1.24E-02
NVR	Closeness 1.75 GSE102698	0070670	response to interleukin-4	32	5	0.55	9.12	1.95E-04	1.55E-02
NVR	Closeness 1.75 GSE102698	0034976	response to endoplasmic reticulum stress	205	12	3.51	3.42	2.18E-04	1.56E-02
PCAFE	Closeness 1.75 GSE102698	0022613	ribonucleoprotein complex biogenesis	358	29	5.47	5.31	1.60E-13	1.14E-10
PCAFE	Closeness 1.75 GSE102698	0006818	hydrogen transport	135	18	2.06	8.73	2.12E-12	7.57E-10
PCAFE	Closeness 1.75 GSE102698	0002181	cytoplasmic translation	49	12	0.75	16.04	6.53E-12	1.55E-09
PCAFE	Closeness 1.75 GSE102698	0009123	nucleoside monophosphate metabolic process	247	19	3.77	5.04	7.45E-09	1.33E-06
PCAFE	Closeness 1.75 GSE102698	0071826	ribonucleoprotein complex subunit organization	180	16	2.75	5.82	1.57E-08	2.10E-06
PCAFE	Closeness 1.75 GSE102698	0009141	nucleoside triphosphate metabolic process	233	18	3.56	5.06	1.76E-08	2.10E-06
PCAFE	Closeness 1.75 GSE102698	0015672	monovalent inorganic cation transport	460	24	7.02	3.42	1.45E-07	1.48E-05
PCAFE	Closeness 1.75 GSE102698	0019693	ribose phosphate metabolic process	435	23	6.64	3.46	2.13E-07	1.90E-05
PCAFE	Closeness 1.75 GSE102698	0006414	translational elongation	49	8	0.75	10.69	6.79E-07	5.39E-05
PCAFE	Closeness 1.75 GSE102698	0072521	purine-containing compound metabolic process	466	22	7.12	3.09	2.63E-06	1.88E-04

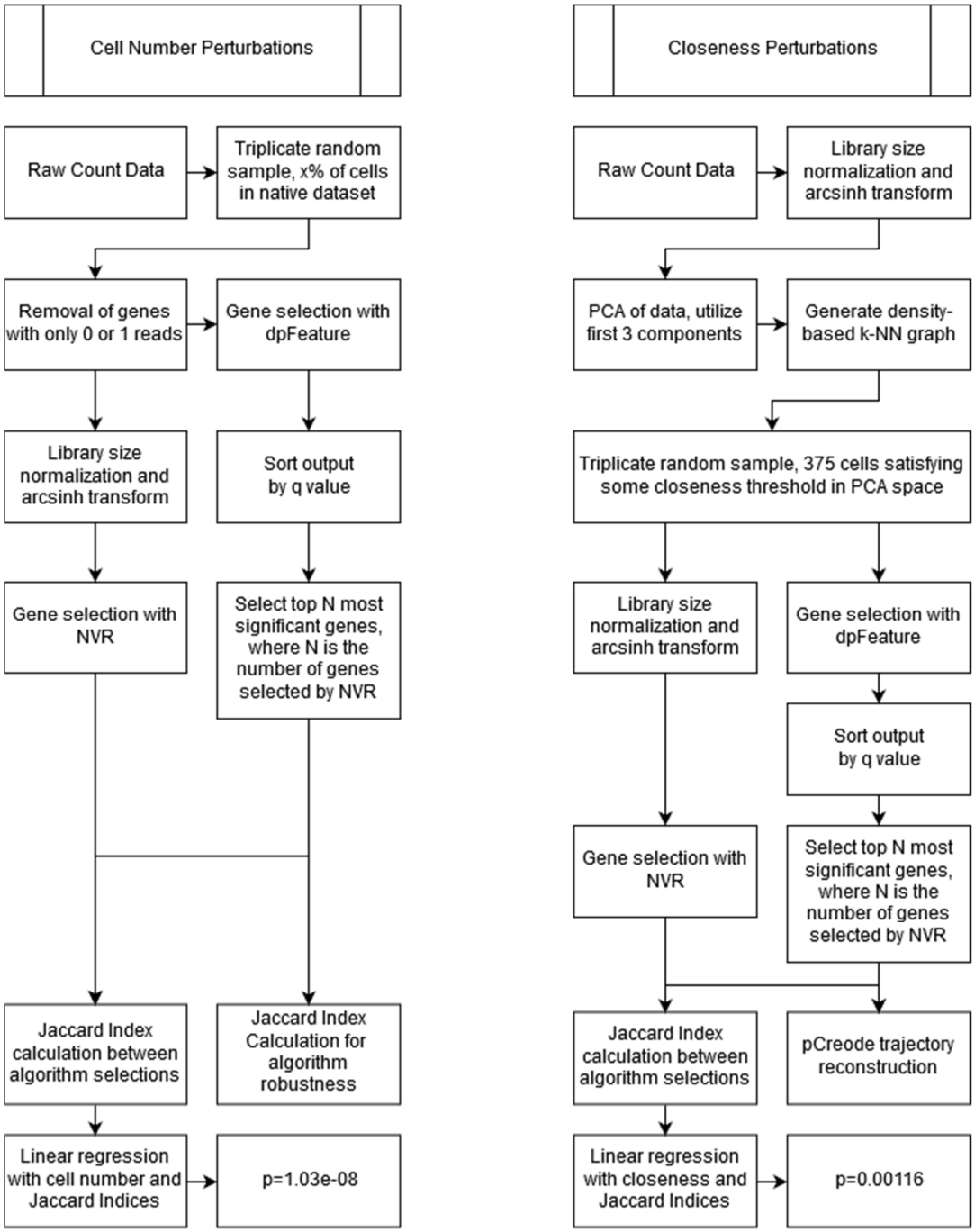
Supplementary Figure S1. Closeness Threshold Sampling

Visualization of the closeness threshold sampling procedure with closeness threshold indicated on top left-hand corners of the plots. This visualization is in principal component space where the axes are the first and second components or the second and third components.



Supplementary Figure S2. Dataset Distribution Analytical Workflow

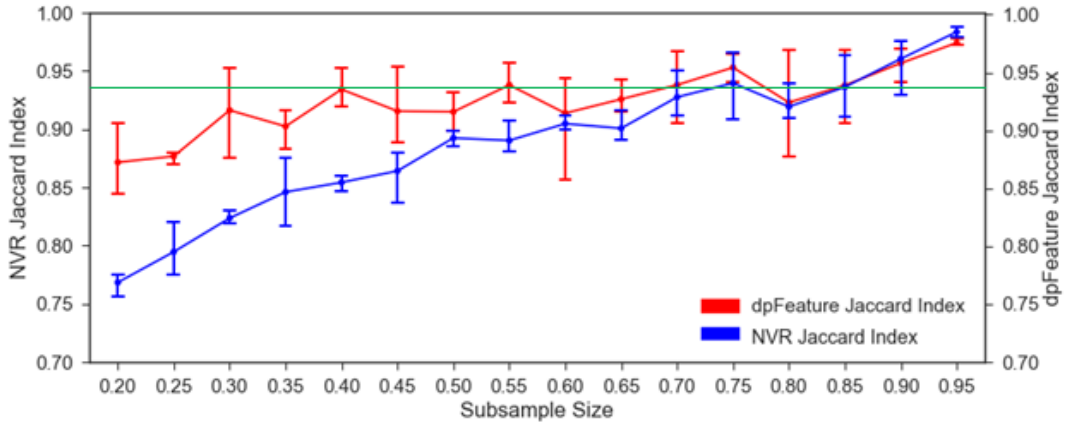
To quantify the effects of cell number, we performed feature selection on triplicate random cell samplings of the datasets with replacement. This random sampling progressed through different granularities from 20% to 95%, in increments of 5%, of the full dataset. Jaccard indices were then calculated to compare algorithm performance. To quantify the effects of cell closeness, we performed feature selection on triplicate random cell samplings given a closeness threshold in principal component space with replacement. We perturbed closeness because of its effect on dataset distribution. The tested closeness thresholds progressed from 0 to 1.75, in increments of 0.25. Jaccard indices were also calculated given these perturbations to compare algorithm performance. The analyses of FVG and PCAFE performance followed a similar workflow to dpFeature, with the number of genes selected based on a significance threshold cutoff.



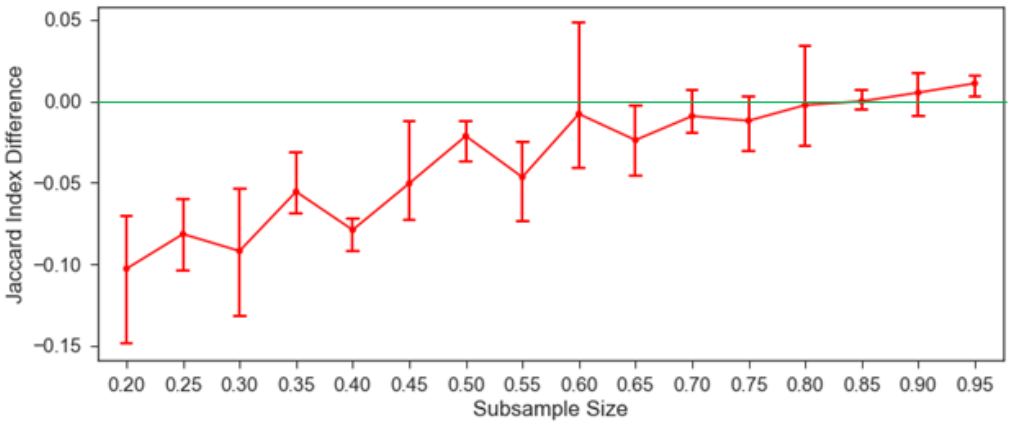
Supplementary Figure S3. Algorithm Robustness

The robustness of the algorithms by overlap between the sets of genes selected by the same algorithm given different dataset sample sizes. (A) Comparison of Jaccard Indices in the context of robustness to random sampling. (B) Representation of the difference in Jaccard index given NVR minus dpFeature. The green lines indicate the points where the mean NVR Jaccard Index becomes greater than the mean dpFeature Jaccard Index.

A.



B.



Supplementary Figure S4. Over-Representation Analysis

Given that the gene sets returned by selected feature selection algorithms vary significantly, we examined the gene ontological annotations of the respective sets. We did this through WebGestalt, as described by the Zhang lab for over-representation analysis. We examined the gene sets generated from GSE102698 using four feature selection algorithms. Over-representation analysis considers the expected and observed number of genes falling within a given GO category. The height of the bars in the plots generated represent the ratio of the number of observed genes over the number of expected genes within some given category. The coloring represents the p-value as determined through hypergeometric testing. Given these p-values, we examined the top ten gene ontology categories by significance.

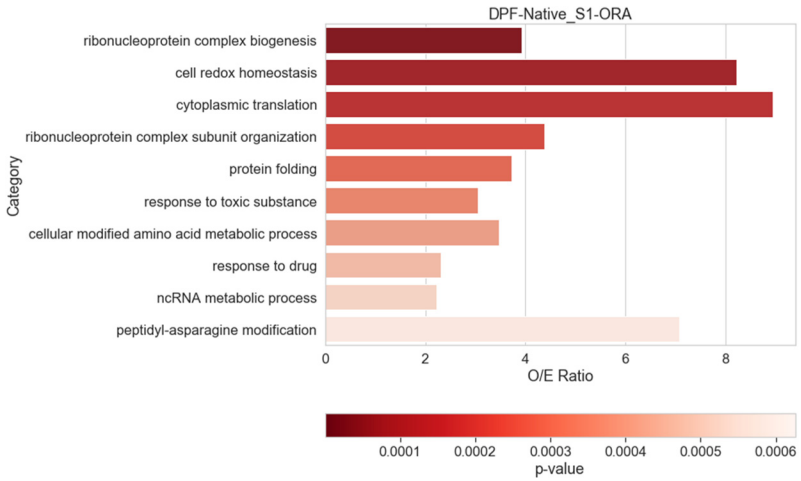
(A) Gene sets generated by the four feature selection algorithms from the native GSE102698 dataset. We observed significantly enriched categories. Though there were similarities, we observed distinctly different categories across the gene sets selected by the four algorithms tested. Notably, NVR and FVG gene sets were associated with categories presumably related to the microbiome. Unlike NVR and FVG, categories similar to “response to molecule of bacterial origin” and “interspecies interaction between organisms” were not associated with dpF and PCAFE gene sets. In total only 8 of 31 unique GO categories were associated with more than one gene set.

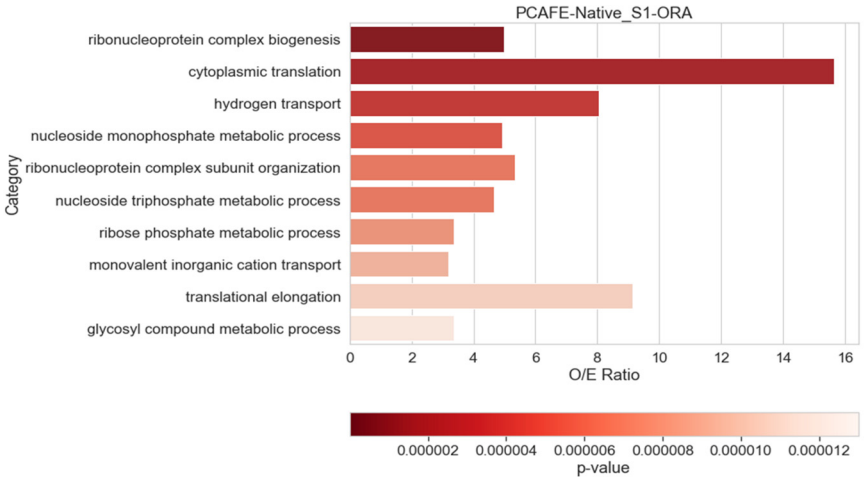
(B) Gene sets generated by the same feature selection algorithms, using resampled GSE102698 datasets with closeness threshold set 0.0 (Supplementary Figure S1). To consolidate the replicates, we performed our analyses on the intersection of the generated gene sets. Interestingly, FVG found associations with an entirely unique set of categories, sharing no categories with the other algorithms given the dataset. 7 of 31 unique GO categories were associated with more than one gene set.

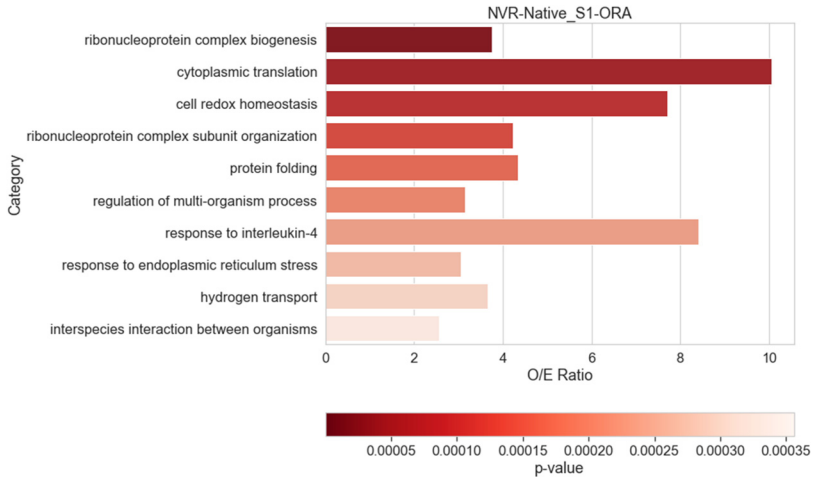
(C) Gene sets generated by the same feature selection algorithms, using resampled GSE102698 datasets with closeness threshold set 1.75 (Supplementary Figure S1). Replicates of resampled datasets were also consolidated through finding replicate set intersections. FVG consistently produced gene sets that had some association with an immune response or bacterial interaction, with categories such as “response to molecule of bacterial origin” and “response to interleukin-4”. A notable category detected using this dataset includes “microvillus organization” associated with dpF’s gene set and unobserved elsewhere in this analysis. 8 of 29 unique GO categories were associated with more than one gene set.

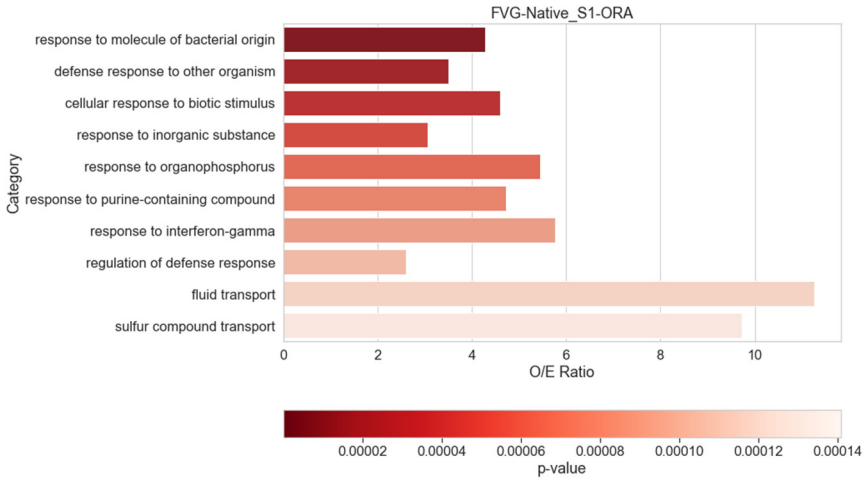
The observed variation of GO categories associated with these four feature selection algorithms demonstrate quantifiable inconsistencies in genes selected with different biological contexts.

A.

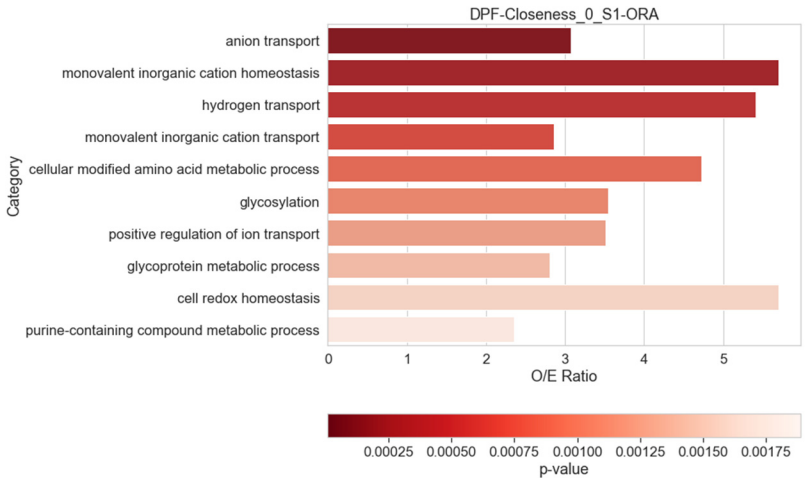

DPF-Native_S1-ORA
ribonucleoprotein complex biogenesis
cell redox homeostasis
cytoplasmic translation
ribonucleoprotein complex subunit organization
protein folding
response to toxic substance
cellular modified amino acid metabolic process
response to drug
ncRNA metabolic process
peptidyl-asparagine modification
Category
0
2
4
6
8
O/E Ratio
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
p-value

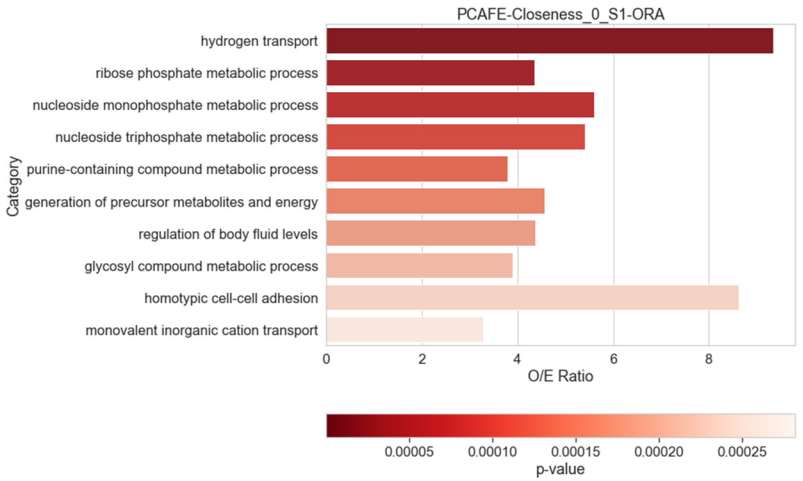

PCAFE-Native_S1-ORA
ribonucleoprotein complex biogenesis
cytoplasmic translation
hydrogen transport
nucleoside monophosphate metabolic process
ribonucleoprotein complex subunit organization
nucleoside triphosphate metabolic process
ribose phosphate metabolic process
monovalent inorganic cation transport
translational elongation
glycosyl compound metabolic process
Category
0
2
4
6
8
10
12
14
16
O/E Ratio
0.000002
0.000004
0.000006
0.000008
0.000010
0.000012
p-value

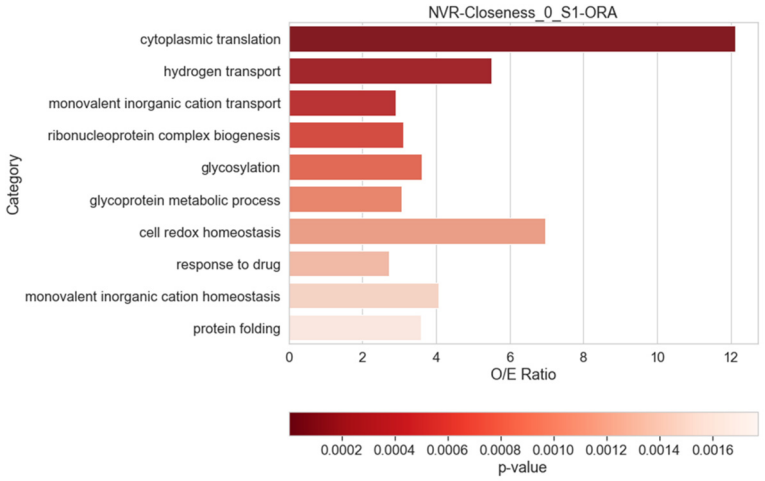

NVR-Native_S1-ORA
ribonucleoprotein complex biogenesis
cytoplasmic translation
cell redox homeostasis
ribonucleoprotein complex subunit organization
protein folding
regulation of multi-organism process
response to interleukin-4
response to endoplasmic reticulum stress
hydrogen transport
interspecies interaction between organisms
Category
0
2
4
6
8
10
O/E Ratio
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
0.00035
p-value

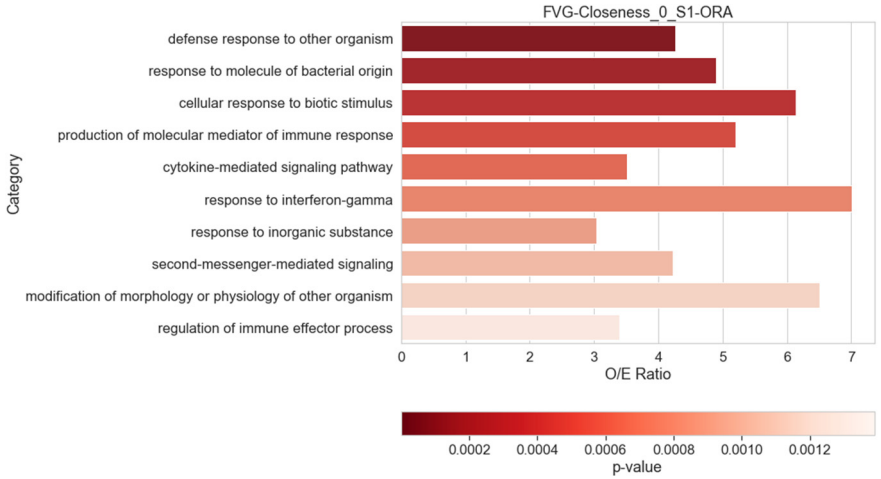

FVG-Native_S1-ORA
response to molecule of bacterial origin
defense response to other organism
cellular response to biotic stimulus
response to inorganic substance
response to organophosphorus
response to purine-containing compound
response to interferon-gamma
regulation of defense response
fluid transport
sulfur compound transport
Category
0
2
4
6
8
10
O/E Ratio
0.00002
0.00004
0.00006
0.00008
0.00010
0.00012
0.00014
p-value

B.

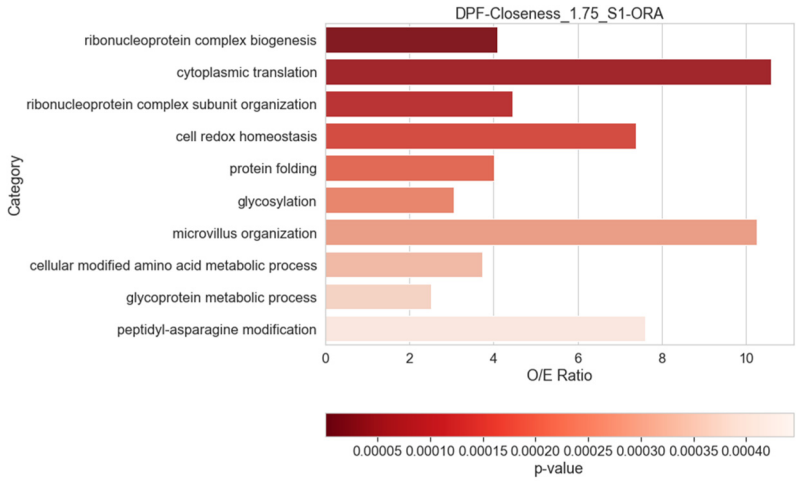

DPF-Closeness_0_S1-ORA
anion transport
monovalent inorganic cation homeostasis
hydrogen transport
monovalent inorganic cation transport
cellular modified amino acid metabolic process
glycosylation
positive regulation of ion transport
glycoprotein metabolic process
cell redox homeostasis
purine-containing compound metabolic process
Category
0
1
2
3
4
5
O/E Ratio
0.00025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
p-value

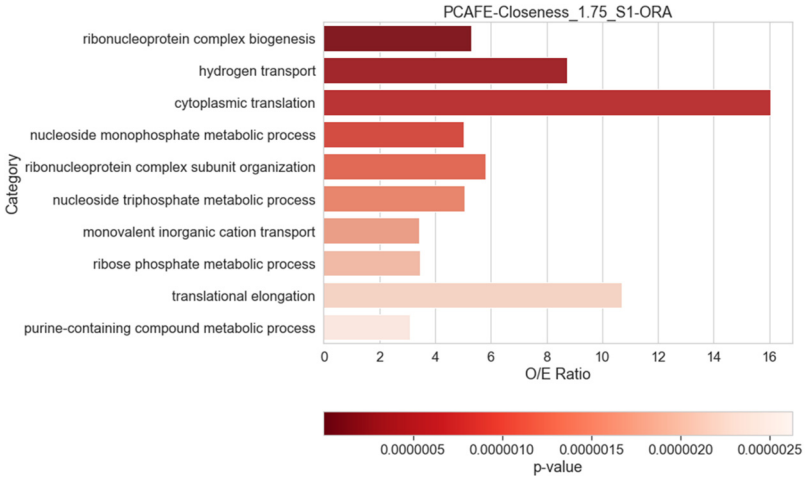

PCAFE-Closeness_0_S1-ORA
hydrogen transport
ribose phosphate metabolic process
nucleoside monophosphate metabolic process
nucleoside triphosphate metabolic process
purine-containing compound metabolic process
generation of precursor metabolites and energy
regulation of body fluid levels
glycosyl compound metabolic process
homotypic cell-cell adhesion
monovalent inorganic cation transport
Category
0
2
4
6
8
O/E Ratio
0.00005
0.00010
0.00015
0.00020
0.00025
p-value

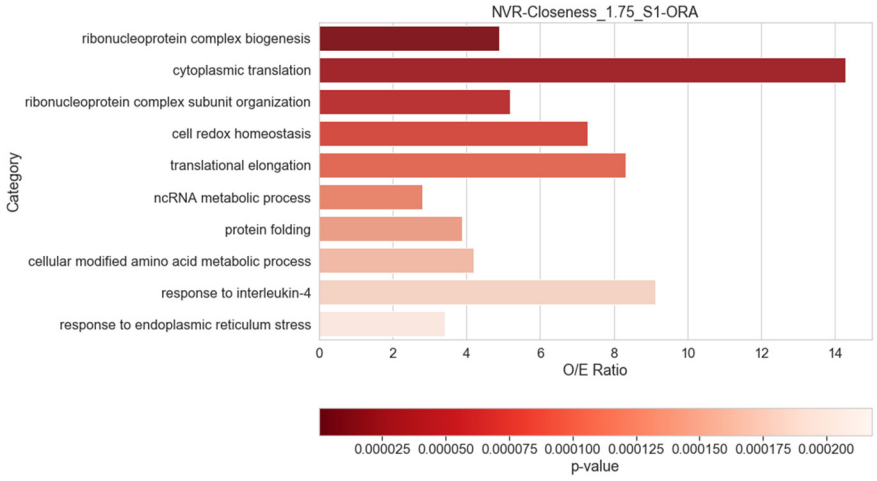

NVR-Closeness_0_S1-ORA
cytoplasmic translation
hydrogen transport
monovalent inorganic cation transport
ribonucleoprotein complex biogenesis
glycosylation
glycoprotein metabolic process
cell redox homeostasis
response to drug
monovalent inorganic cation homeostasis
protein folding
Category
0
2
4
6
8
10
12
O/E Ratio
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
0.0016
p-value

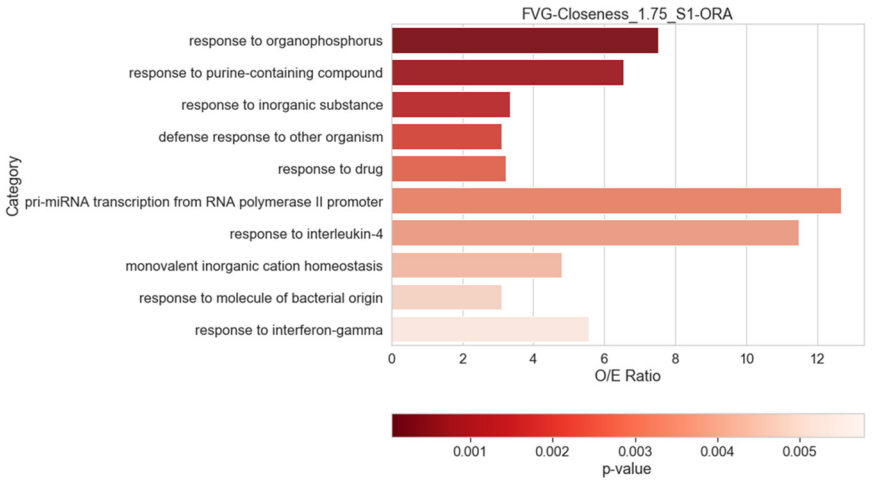

FVG-Closeness_0_S1-ORA
defense response to other organism
response to molecule of bacterial origin
cellular response to biotic stimulus
production of molecular mediator of immune response
cytokine-mediated signaling pathway
response to interferon-gamma
response to inorganic substance
second-messenger-mediated signaling
modification of morphology or physiology of other organism
regulation of immune effector process
Category
0
1
2
3
4
5
6
7
O/E Ratio
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
p-value

C.


DPF-Closeness_1.75_S1-ORA
ribonucleoprotein complex biogenesis
cytoplasmic translation
ribonucleoprotein complex subunit organization
cell redox homeostasis
protein folding
glycosylation
microvillus organization
cellular modified amino acid metabolic process
glycoprotein metabolic process
peptidyl-asparagine modification
Category
0 2 4 6 8 10
O/E Ratio
0.00005 0.00010 0.00015 0.00020 0.00025 0.00030 0.00035 0.00040
p-value

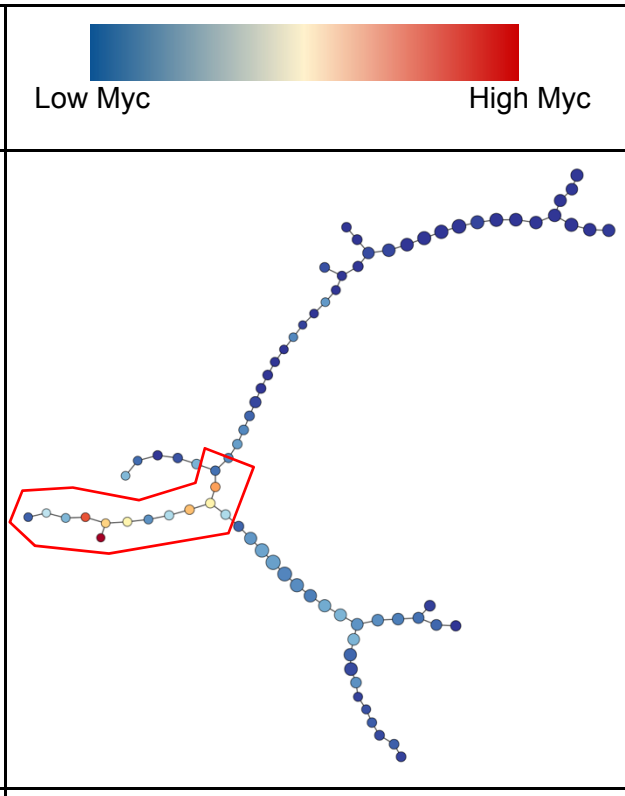
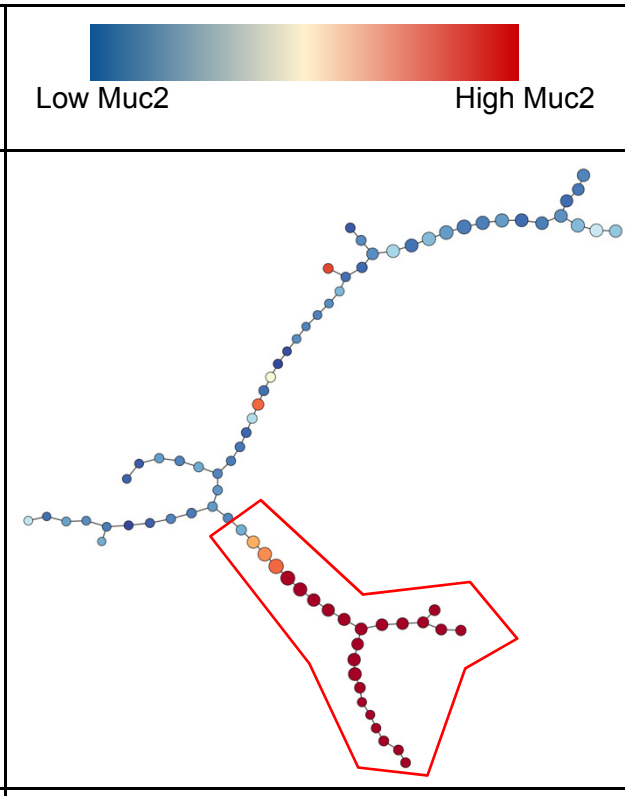
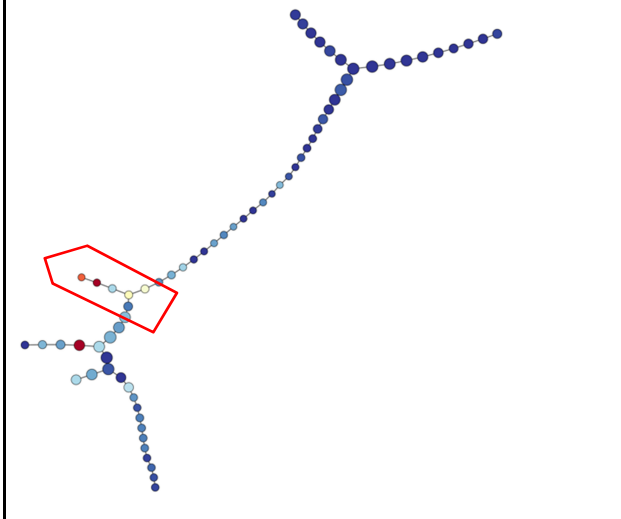
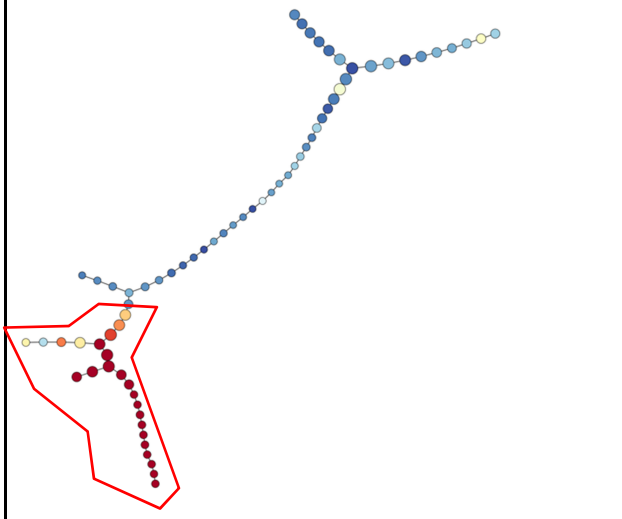

PCAFE-Closeness_1.75_S1-ORA
ribonucleoprotein complex biogenesis
hydrogen transport
cytoplasmic translation
nucleoside monophosphate metabolic process
ribonucleoprotein complex subunit organization
nucleoside triphosphate metabolic process
monovalent inorganic cation transport
ribose phosphate metabolic process
translational elongation
purine-containing compound metabolic process
Category
0 2 4 6 8 10 12 14 16
O/E Ratio
0.0000005 0.0000010 0.0000015 0.0000020 0.0000025
p-value


NVR-Closeness_1.75_S1-ORA
ribonucleoprotein complex biogenesis
cytoplasmic translation
ribonucleoprotein complex subunit organization
cell redox homeostasis
translational elongation
ncRNA metabolic process
protein folding
cellular modified amino acid metabolic process
response to interleukin-4
response to endoplasmic reticulum stress
Category
0 2 4 6 8 10 12 14
O/E Ratio
0.000025 0.000050 0.000075 0.000100 0.000125 0.000150 0.000175 0.000200
p-value

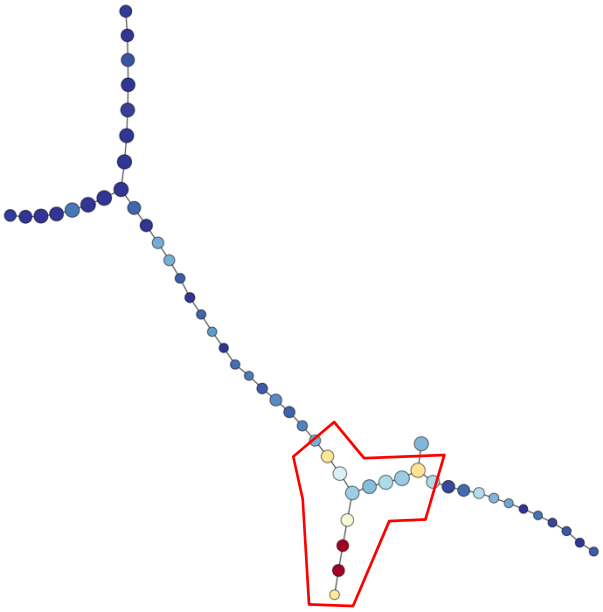
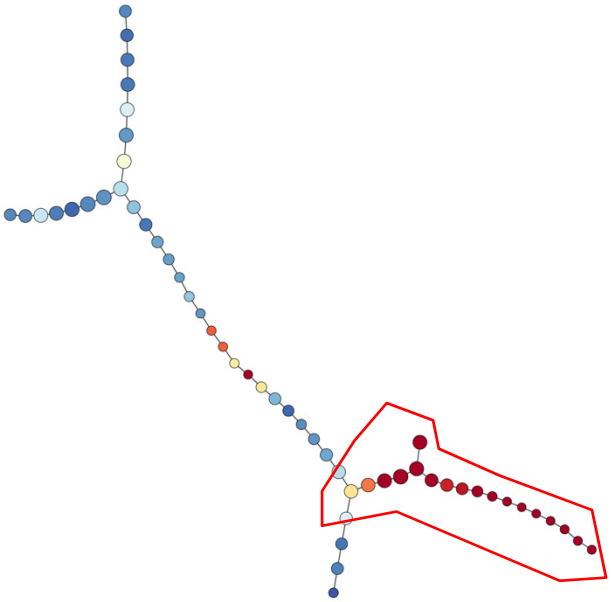
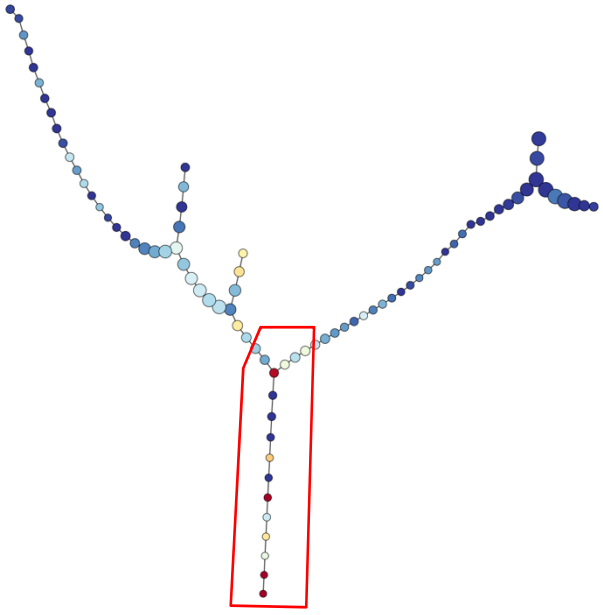
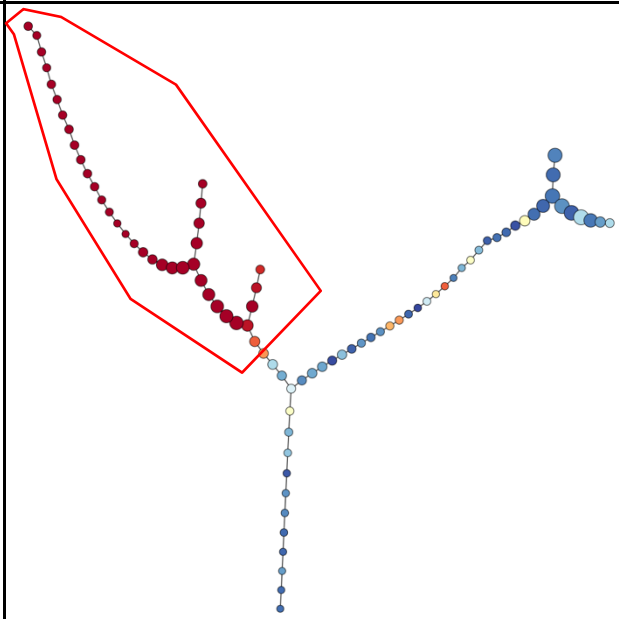

FVG-Closeness_1.75_S1-ORA
response to organophosphorus
response to purine-containing compound
response to inorganic substance
defense response to other organism
response to drug
pri-miRNA transcription from RNA polymerase II promoter
response to interleukin-4
monovalent inorganic cation homeostasis
response to molecule of bacterial origin
response to interferon-gamma
Category
0 2 4 6 8 10 12
O/E Ratio
0.001 0.002 0.003 0.004 0.005
p-value

Supplementary Figure S5. FVG and PCAFE p-Creode Analysis

p-Creode analysis on the gene sets selected by FVG and PCAFE similar to the analysis and interpretation performed with dpFeature and NVR. As we did for our over-representation analysis in Supplementary Figure S6, we consolidated genes selected from replicate dataset samplings by examining their intersections.

<table>
<tr><td colspan="3">FVG</td></tr>
<tr><td>Closeness threshold</td><td> 
</td><td> 
</td></tr>
<tr><td>0.0</td><td></td><td></td></tr>
<tr><td>1.75</td><td>  </td><td>  </td></tr>
</table>

PCAFE

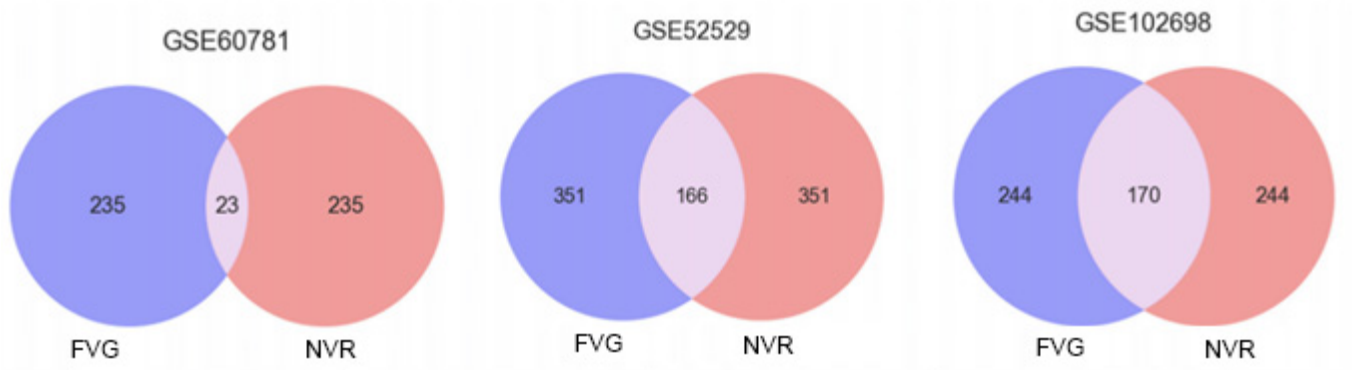
Closeness threshold	Low Myc — High Myc	Low Muc2 — High Muc2
0.0		
1.75		

Supplementary Figure S6. findVariableGenes Similarity Analysis

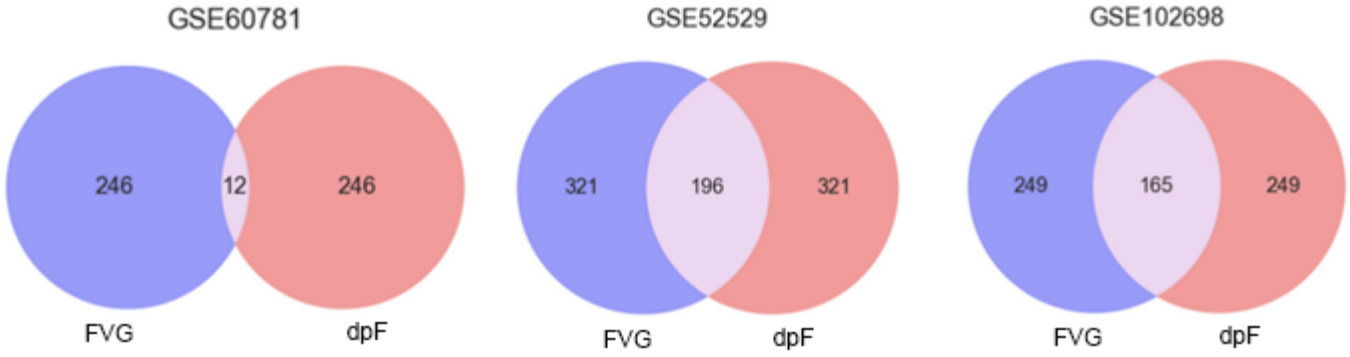
In addition to comparisons between dpFeature and NVR, we also examined findVariableGenes as a feature selection method. This algorithm is detailed in Supplementary Method S1.7. We observed relatively low similarity indices between the sets of genes selected from native datasets between findVariableGenes and NVR as well as dpFeature (A).

We performed the same gene set similarity analyses described in Figure 1 given different cell number and closeness samplings. (B) Between FVG and NVR, we observed significant, positive linear relationships (Supplementary Table S3) between gene set Jaccard index, cell number (p=2.27e-11), and cell closeness sampling thresholds (p=2.63e-8). (C) A similar relationship was observed between FVG and dpFeature; we observed significant, positive linear relationships (Supplementary Table S3) between gene set Jaccard index, cell number (p=8.90e-7), and cell closeness sampling thresholds (p=1.63e-8).

A.

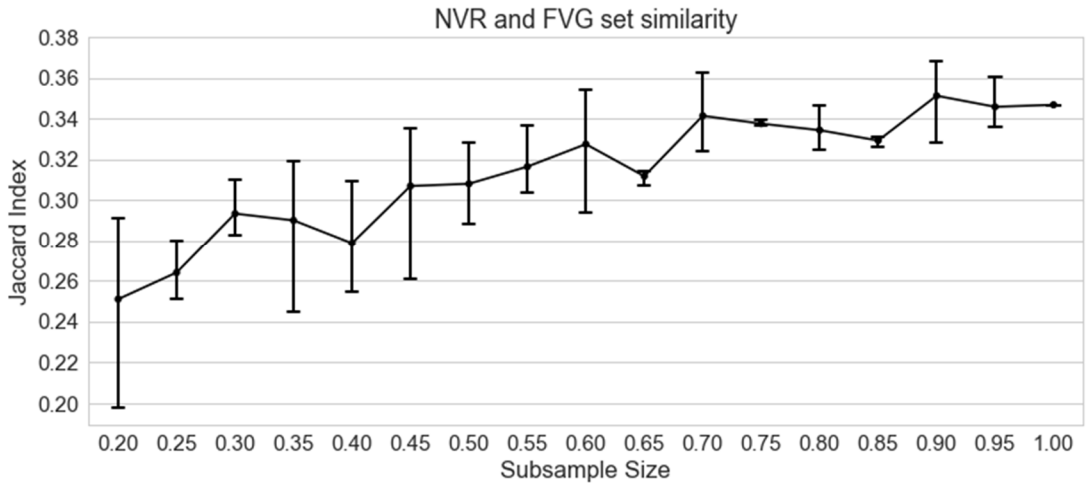


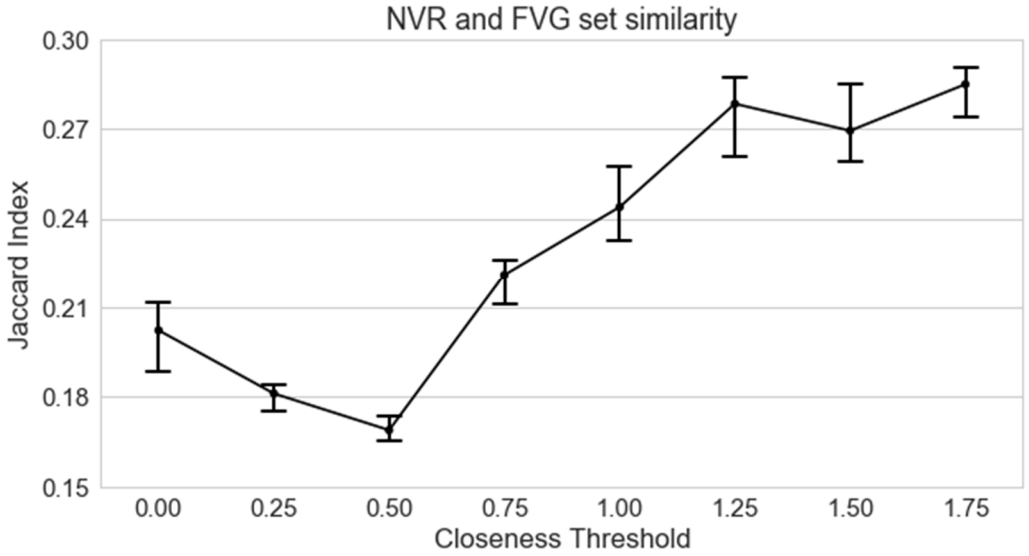
Dataset	GSE60781	GSE52529	GSE102698
Cell Number	251	271	1597
Jaccard Index	0.05	0.15	0.15



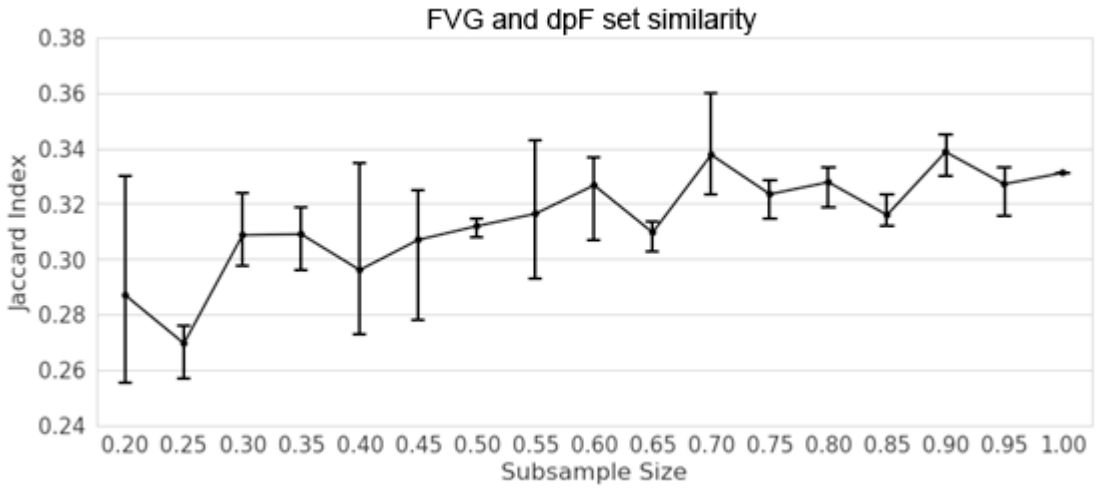
Dataset	GSE60781	GSE52529	GSE102698
Cell Number	251	271	1597
Jaccard Index	0.02	0.23	0.14

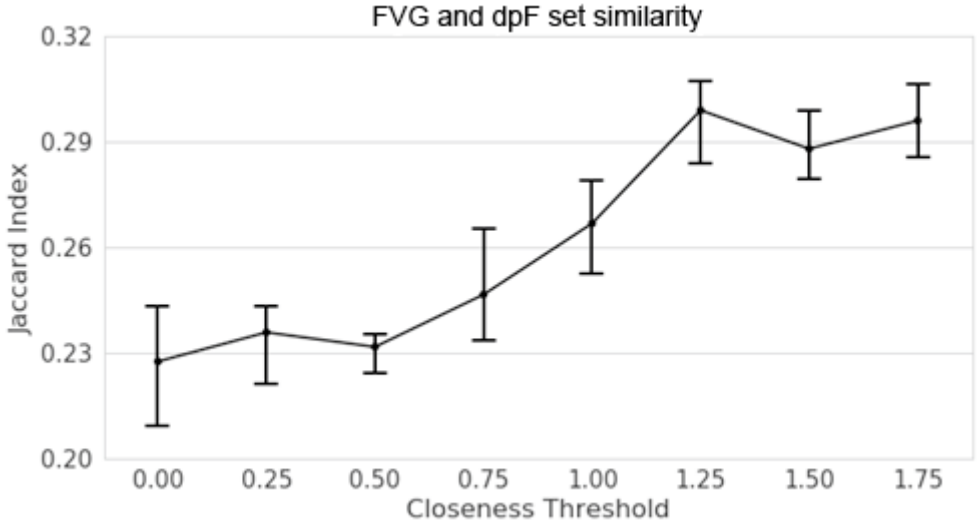
B.


NVR and FVG set similarity
Jaccard Index
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Subsample Size


NVR and FVG set similarity
Jaccard Index
0.30
0.27
0.24
0.21
0.18
0.15
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
Closeness Threshold

C.

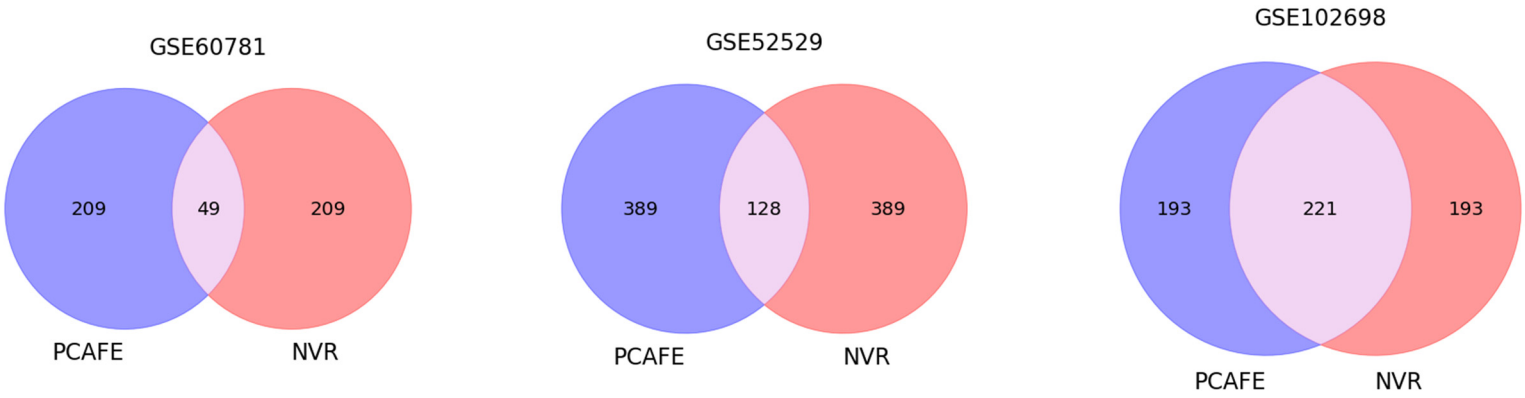

FVG and dpF set similarity
Jaccard Index
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Subsample Size


FVG and dpF set similarity
Jaccard Index
0.32
0.29
0.26
0.23
0.20
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
Closeness Threshold

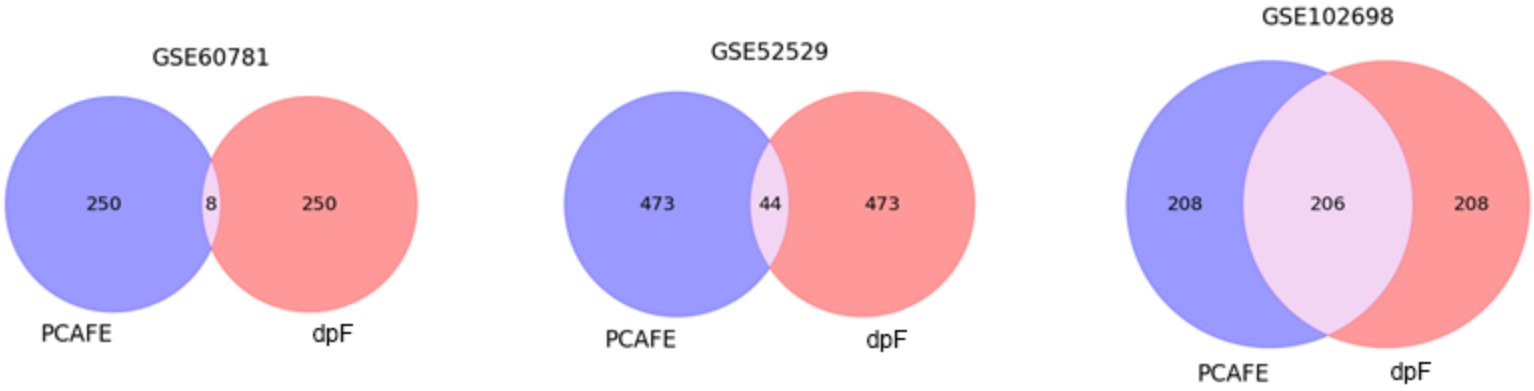
Supplementary Figure S7. PCAFE Similarity Analysis

We also performed analyses on PCAFE. This algorithm is detailed in Supplementary Method S1.8. We observed a range of similarity indices between the sets of genes selected from native datasets between PCAFE and NVR as well as dpFeature (A). We performed the same gene set similarity analyses described in Figure 1 given different cell number and closeness samplings. (B) Between PCAFE and NVR, we observed significant, positive linear relationships (Supplementary Table S3) between gene set Jaccard index and cell closeness sampling thresholds (p=1.64e-6A corresponding trend was not observed with respect to cell number sampling (p=0.088).

A.

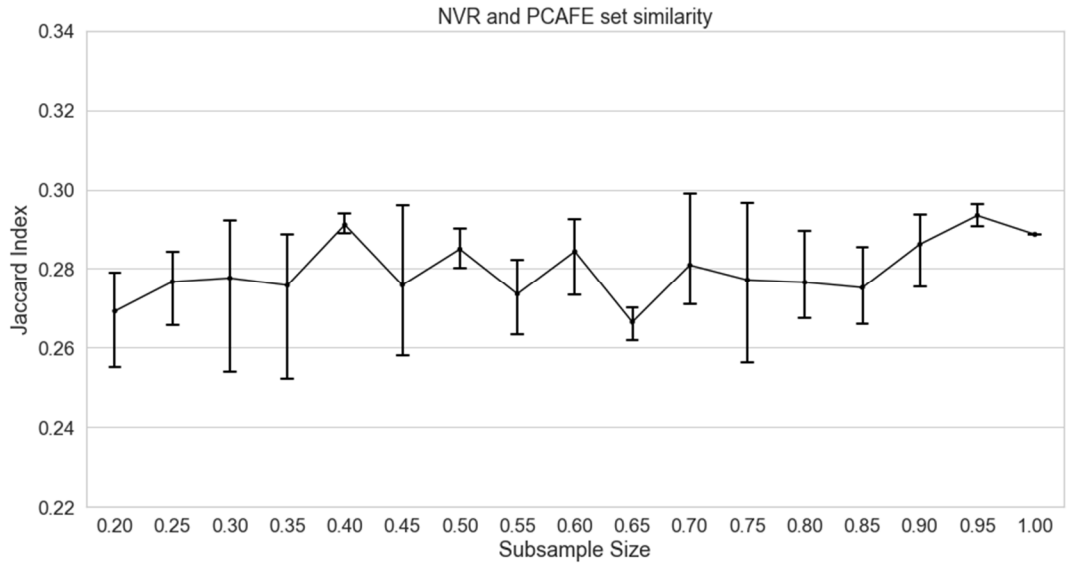


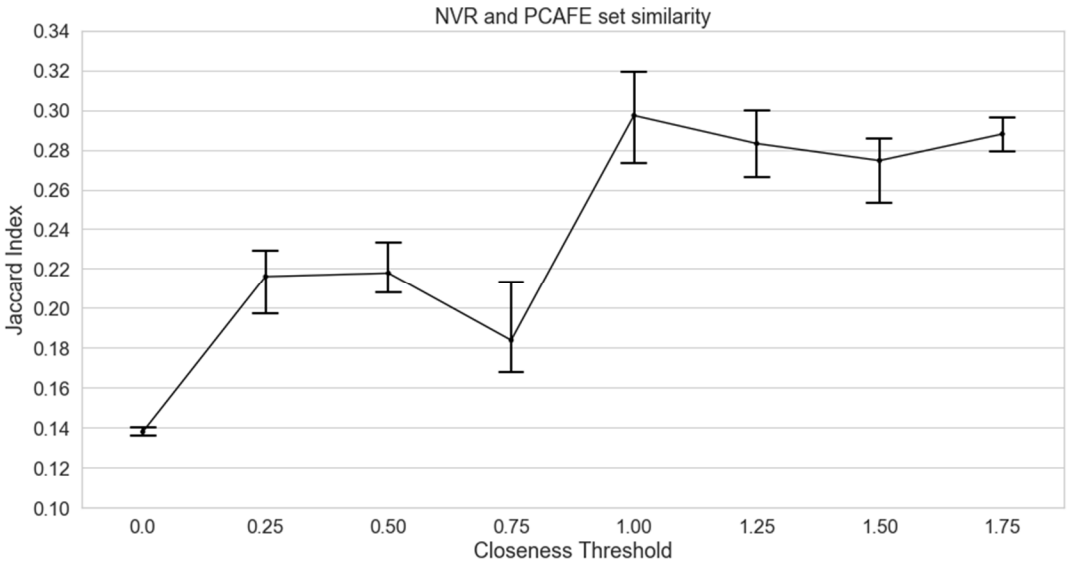
Dataset	GSE60781	GSE52529	GSE102698
Cell Number	251	271	1597
Jaccard Index	0.10	0.14	0.36



Dataset	GSE60781	GSE52529	GSE102698
Cell Number	251	271	1597
Jaccard Index	0.01	0.04	0.33

B.


NVR and PCAFE set similarity
Jaccard Index
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Subsample Size


NVR and PCAFE set similarity
Jaccard Index
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.0 0.25 0.50 0.75 1.00 1.25 1.50 1.75
Closeness Threshold

References

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Herring,C.A. *et al.* (2018) Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Syst.*, **6**, 37–51.e9.

Levandowsky,M. and Winter,D. (1971) Distance between Sets. *Nature*, **234**, 34–35.

Van Der Maaten,L.J.P. and Hinton,G.E. (2008) Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Macosko,E.Z. *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.

Qiu,X. *et al.* (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14**, 979–982.

Rodriguez,A. and Laio,A. (2014) Clustering by fast search and find of density peaks. *Science*, **344**, 1492–1496.

Taguchi,Y. (2018) Principal Component Analysis-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis. In, Huang,D. *et al.* (eds), *ICIC 2018: Intelligent Computing Theories and Application.* Springer, Cham, 816–826.

Wang,J. *et al.* (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.*, **45**, W130–W137.

Welch,J.D. *et al.* (2016) SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.*, **17**, 106.