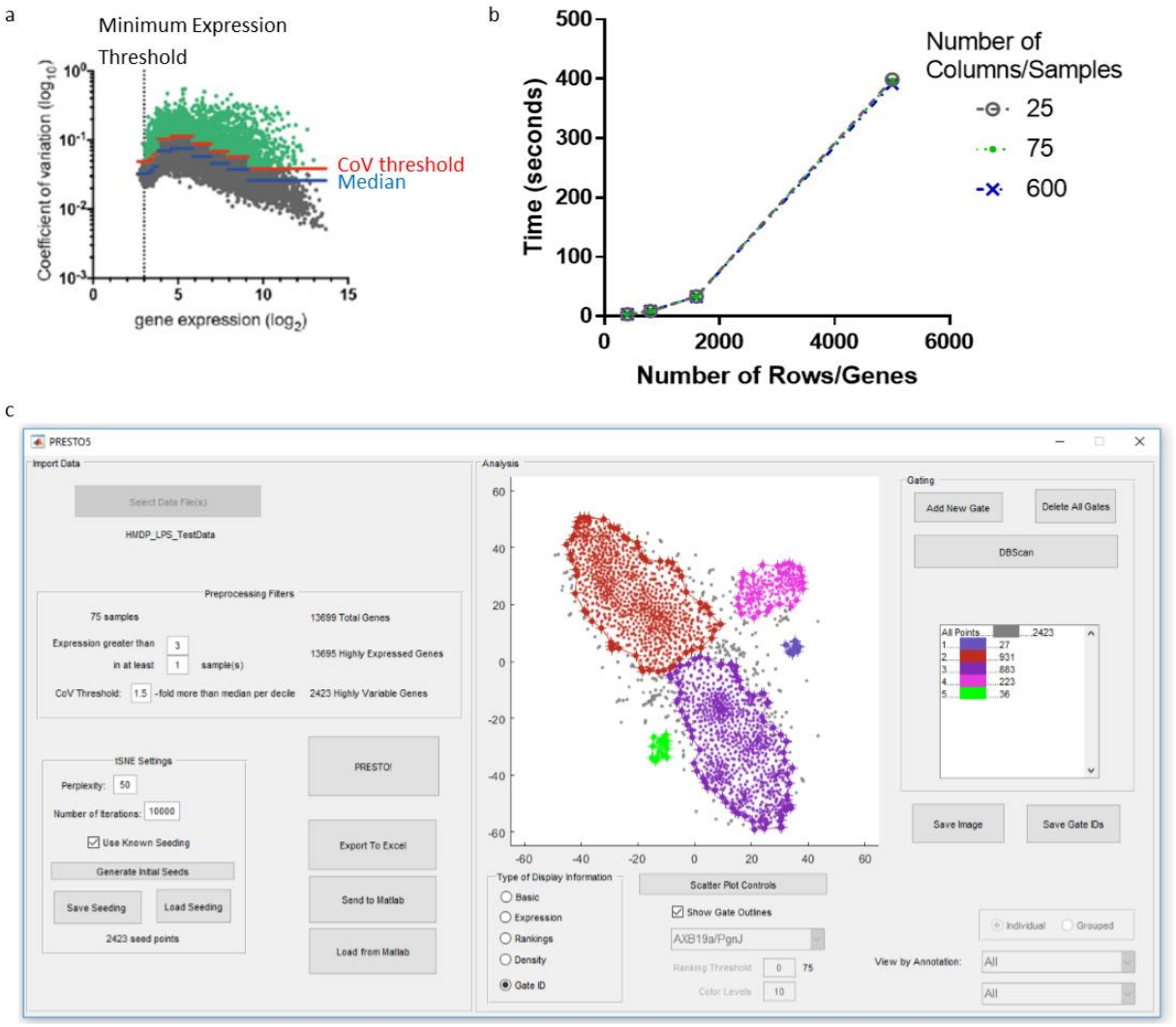
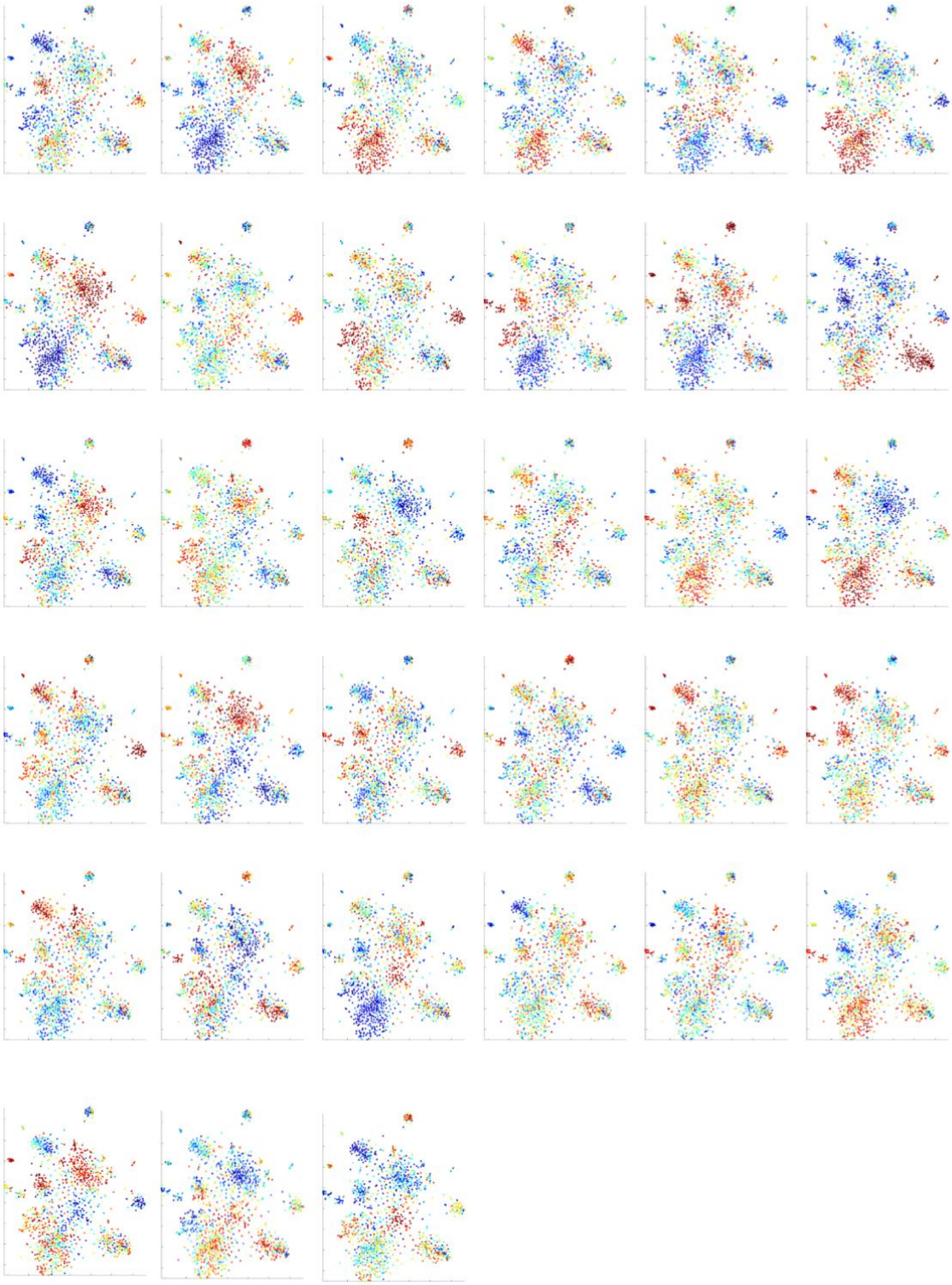
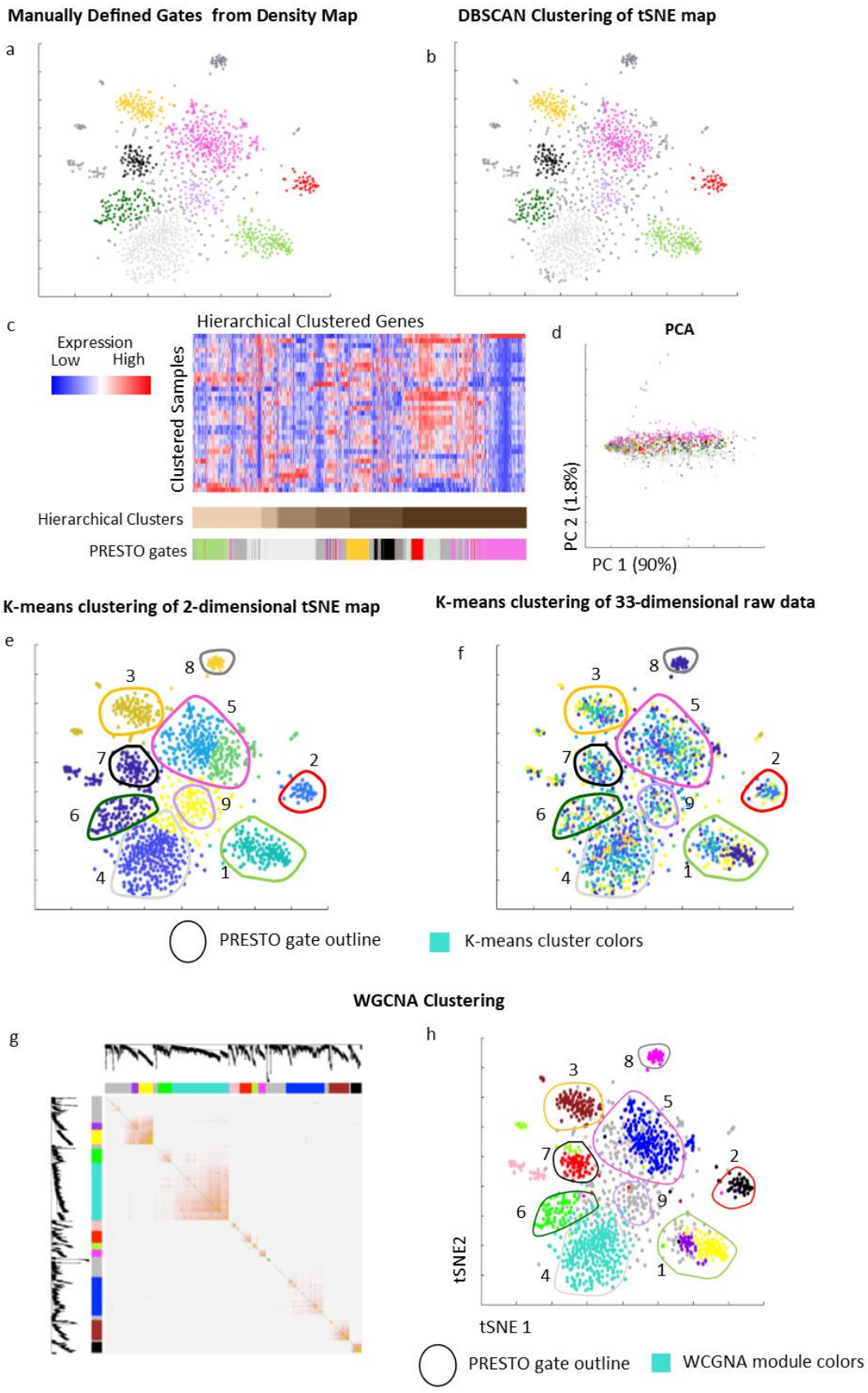


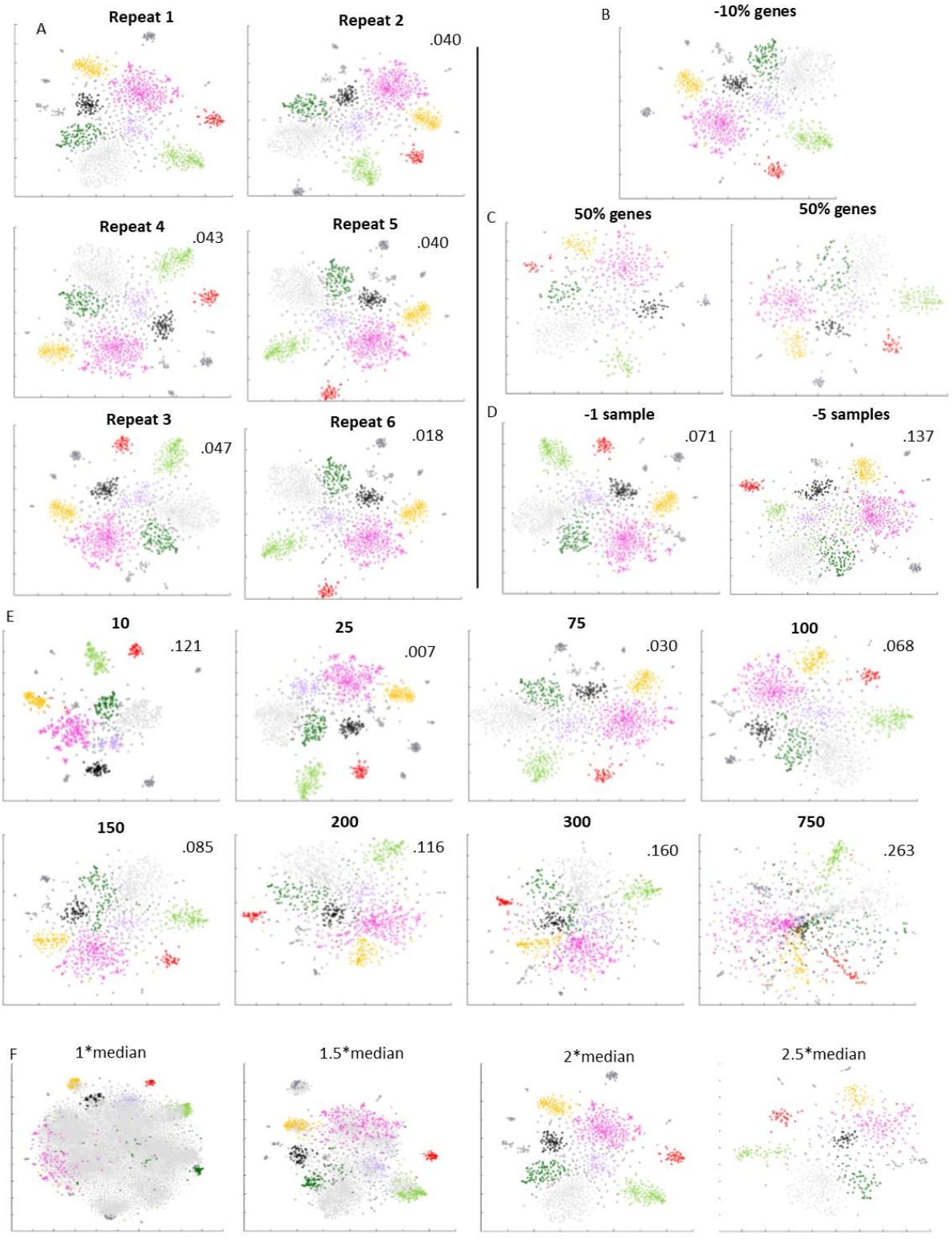
Supplemental Figure 1: PRESTO implementation. a) PRESTO pre-processing divides the data into deciles based on mean expression across the samples. For each decile the median coefficient of variation (CoV) is calculated (blue line). A minimum variation threshold is set by multiplying the median values by a user-controlled factor (red). b) Simulated data of varying sizes was run to 1,000 iterations on a high-end desktop PC with 128 GB RAM. The processing time increases exponentially with increasing number of rows (genes or proteins) and negligibly for increasing number of columns (samples). The "perplexity" is 20. c) Screen shot of the PRESTO user interface.



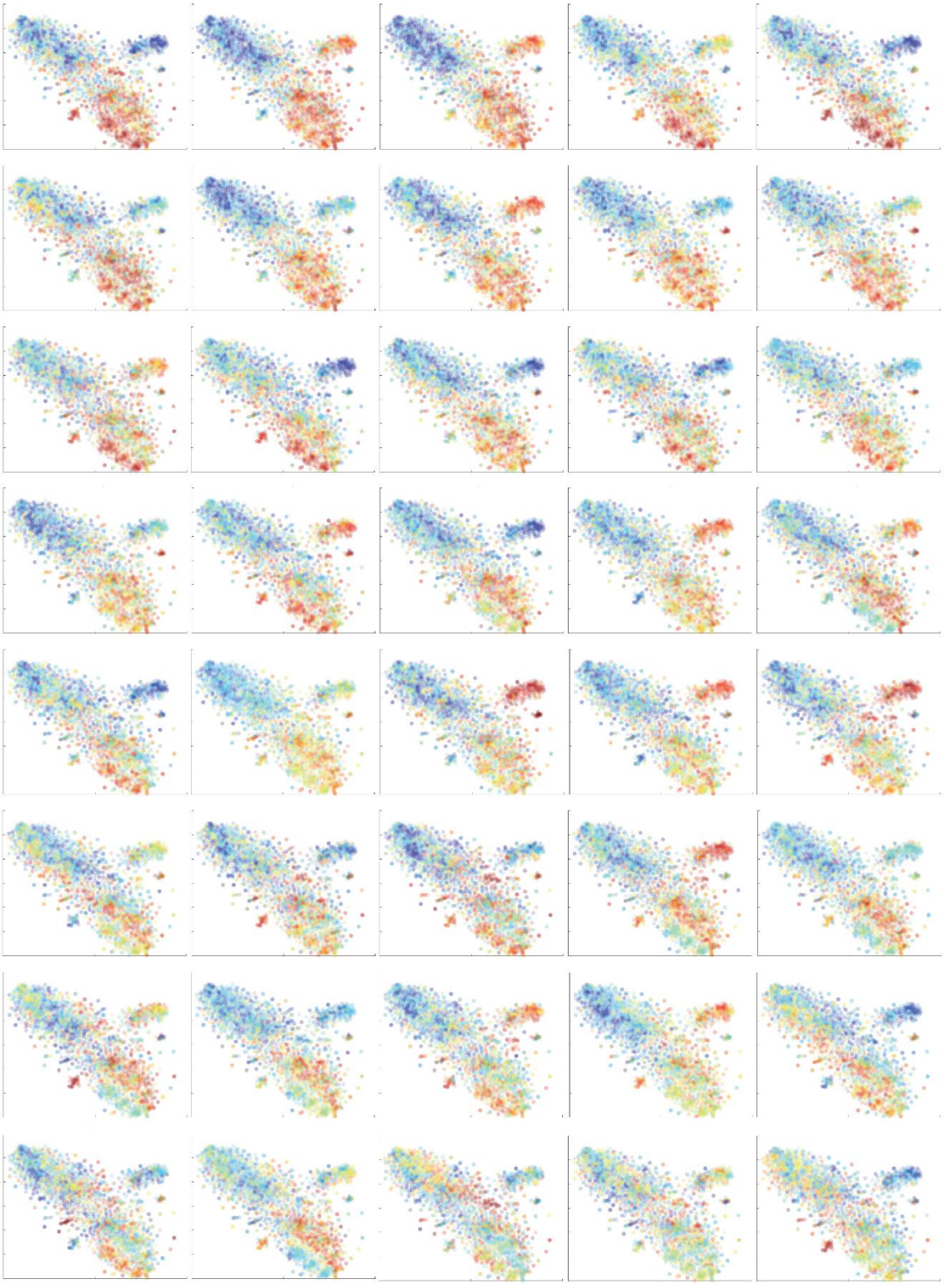
Supplemental Figure 2: Relative ranking of gene expression in 33 human subjects. For each gene (represented by a dot), expression was ranked from lowest (blue) to highest (red) by subject. Data for each of the 33 subjects shown in random order (GSE74816).

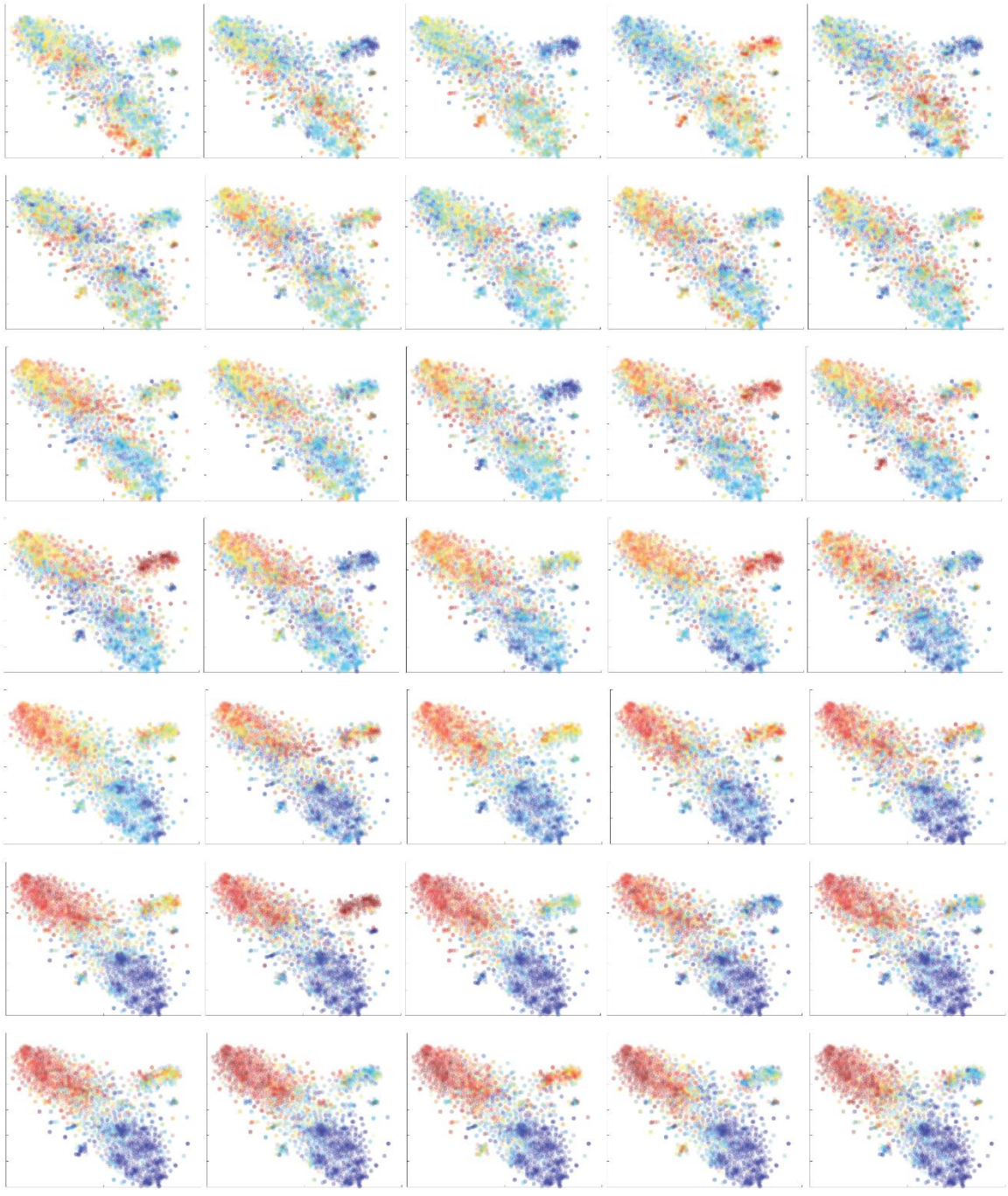

Manually Defined Gates from Density Map
DBSCAN Clustering of tSNE map
a
b
c
Hierarchical Clustered Genes
Expression
Low High
Clustered Samples
Hierarchical Clusters
PRESTO gates
d
PCA
PC 2 (1.8%)
PC 1 (90%)
K-means clustering of 2-dimensional tSNE map
K-means clustering of 33-dimensional raw data
e
f
PRESTO gate outline
K-means cluster colors
WGCNA Clustering
g
h
tSNE2
tSNE 1
PRESTO gate outline
WGCNA module colors

Supplemental Figure 3: Comparison of PRESTO results with other methods. A) Scatter plot showing the gene groupings based on the manual gating of the human PBMC transcriptomes in Figure 2. B) DBScan was used to automatically cluster the points on the tSNE map, where the parameters were altered so that the results best matched the visually perceived gene groupings. C) K-means clustering of the tSNE map was performed, with k=9 visible groups. Comparison of the PRESTO gates (outlines) and assigned cluster (dot color) shows weak similarity between the methods. D) K-means clustering of the 1,298 genes before dimensionality reduction (based on expression in all 33 samples) was performed and shows no similarity to the PRESTO gates. E) Hierarchical clustering of the 1,298 genes that passed the PRESTO filters produced 7 groups (shown below the heat map). For comparison, the group identity for each gene is shown. F) Principal component analysis of the 1,298 filtered genes shows no discernible pattern. Each gene is colored based on the scheme in (a). G) WGCNA was performed on the filtered genes and 11 groups were automatically delineated. H) Comparison of the PRESTO gates (outlines) and assigned cluster (dot color) shows weak similarity between the methods.

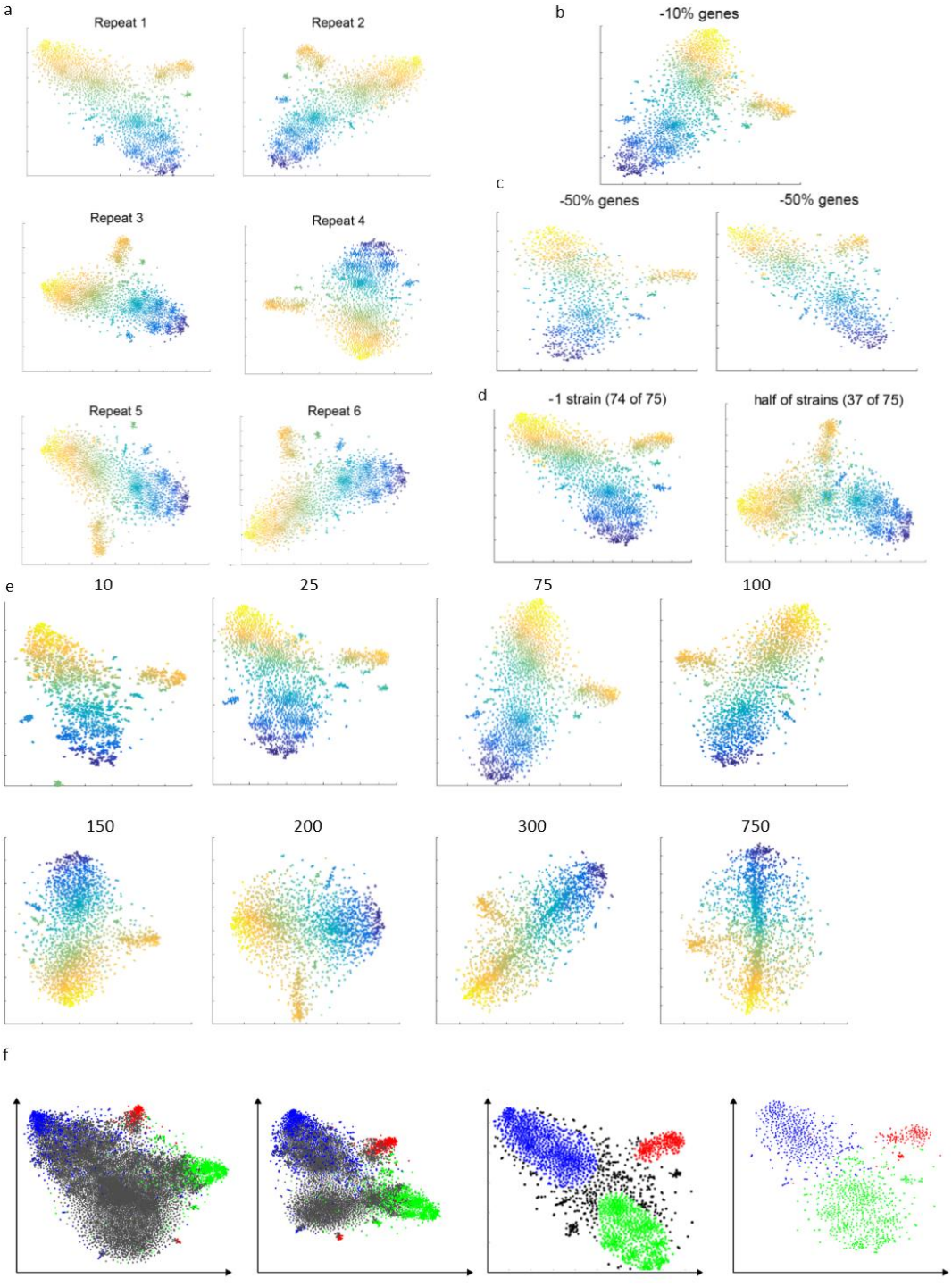

A
Repeat 1
Repeat 2
.040
Repeat 4
.043
Repeat 5
.040
Repeat 3
.047
Repeat 6
.018
B
-10% genes
C
50% genes
50% genes
D
-1 sample
.071
-5 samples
.137
E
10
.121
25
.007
75
.030
100
.068
150
.085
200
.116
300
.160
750
.263
F
1*median
1.5*median
2*median
2.5*median

Supplemental Figure 4: Sensitivity analysis shows PRESTO produces consistent results. Dimensionality reduction was performed for the 1,298 highly variable genes from the human PBMC transcriptome data set (Figure 2, GSE74816) using a "perplexity" of 50 and 6 different random initial seedings. The color of each gene in each plot was defined by its group ID (Figure 2b). The Jenson-Shannon divergence (JSD) between each plot and the "standard" plot (Repeat 1) is shown in the upper right corner. Randomly generated data yield an average JSD of 0.334. The grouping of the genes minimally changes but the plot rotates randomly. B-D) Subsampling was performed to determine the sensitivity of the 2 dimensional organization to small changes in the input data. Dimensionality reduction was performed after B) 10% of the filtered genes, C) 50% of the filtered genes, D) 1 patient out of 33 or 5 patients out of 33 were removed, using the same initial seeding for the remaining genes. F) The "perplexity" was varied between 10-750. The boundaries between regions becomes less clear with a "perplexity" values that is too high or too low, but the basic organizational scheme does not change until very high "perplexity" values. E) Varying CoV cutoff values from 0*median to 2.5*median were used in the pre-processing of transcriptome data from human PBMCs (Figure 2). The color of each gene in each plot was defined by its group ID, with genes that were not in (a) shown in dark grey.

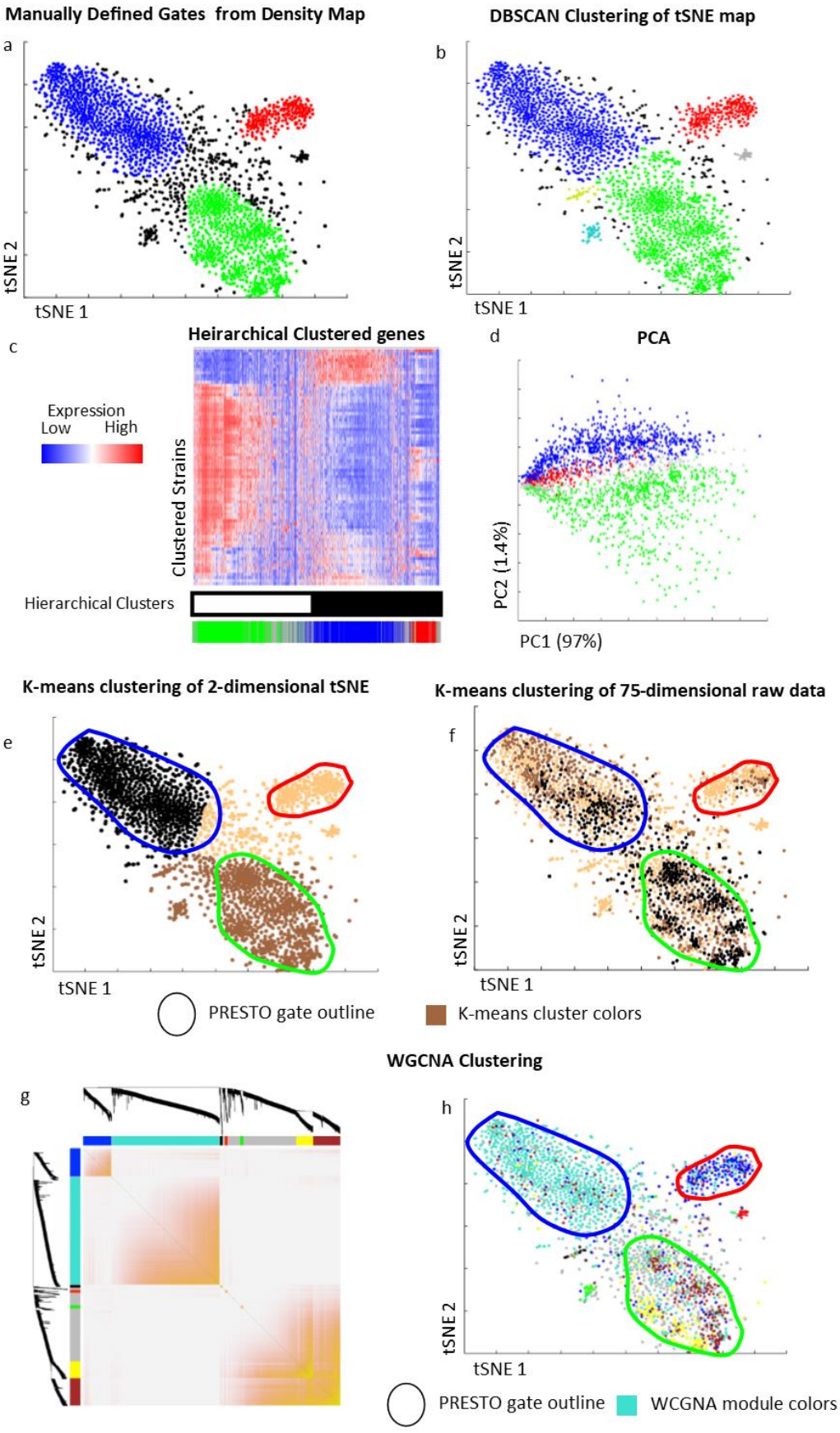




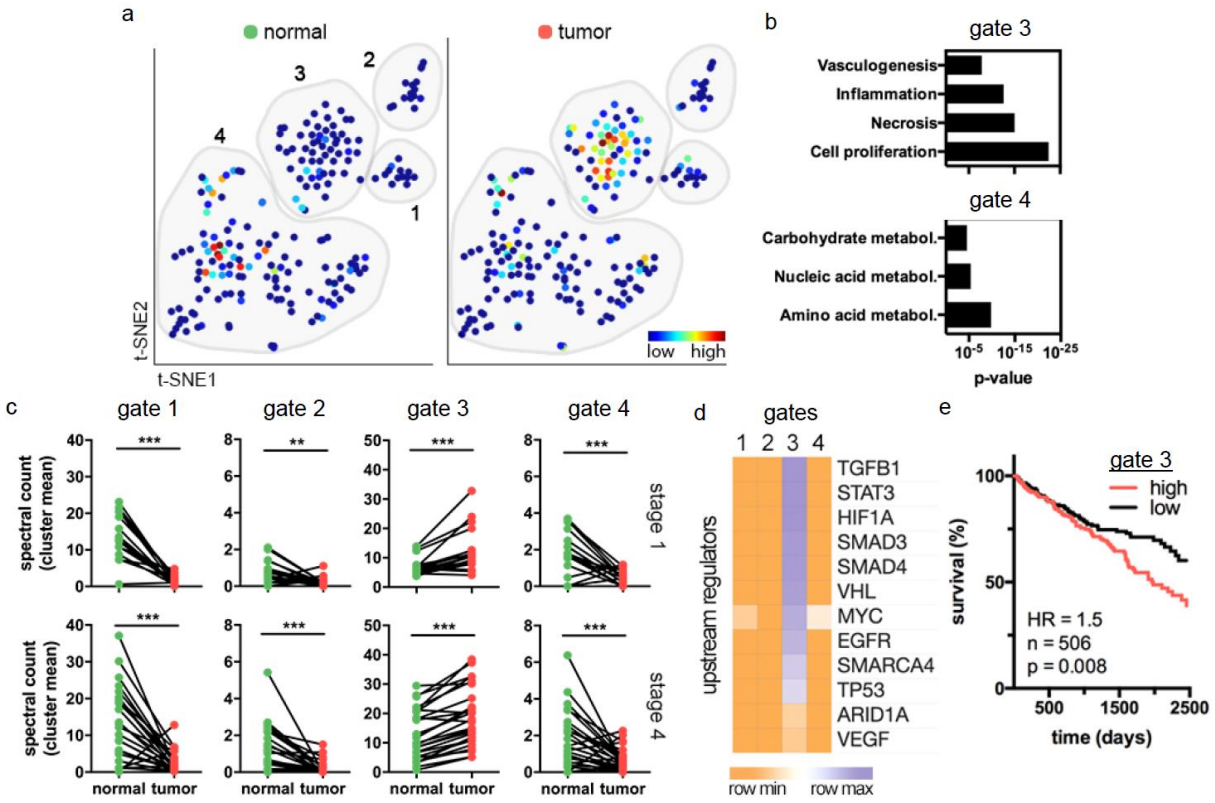
Supplementary Figure 5: Relative ranking of gene expression in 75 mouse strains. For each gene (dot), the strains were ranked from that which expresses that gene the highest to the lowest. Then, the rankings of a particular strain for all genes can be displayed so that the genes that it expresses at relatively high amounts are colored red and the relatively low expressed genes are color blue. Mouse strains were ordered based on decreasing group mean expression of the bottom right region, similar to Supplemental Table 2. Data from GSE38705[32]. See Supplemental Movie 2.


a
Repeat 1
Repeat 2
Repeat 3
Repeat 4
Repeat 5
Repeat 6
b
-10% genes
c
-50% genes
-50% genes
d
-1 strain (74 of 75)
half of strains (37 of 75)
e
10
25
75
100
150
200
300
750
f

Supplemental Figure 6: Sensitivity analysis shows PRESTO produces consistent results. a) Dimensionality reduction was performed for the 2,423 highly variable genes from the LPS-treated macrophage data set (Figure 4, GSE38705) using a "perplexity" of 50 and 30 different random initial seedings, six of which are shown. The color of each gene was defined based on its y-coordinate in "Repeat 1", and then was retained for all subsequent graphs. The relationship between the genes minimally changes despite global rotation of the plot. The JSD ranges from 0 to 0.036 for the repeats (average = 0.027). Randomly generated data yield an average JSD of 0.49. B-E) Subsampling was performed to determine the sensitivity of the 2 dimensional organization to small changes in the input data. Dimensionality reduction was performed after V) 10% of the filtered genes, C) 50% of the filtered genes, D) 1 strain out of 75 or 38 strains out of 75 were removed. E) The "perplexity" was varied between 10-750. The boundaries between regions becomes less clear with a "perplexity" values that is too high or too low, but the basic organizational scheme does not change. F) Varying CoV cutoff values from 1*median to 2.5*median were used in the pre-processing of transcriptome data. The color of each gene in each plot was defined by its group ID (Figure 3a), with genes that were not in the original plot shown in dark grey.


Manually Defined Gates from Density Map
a
tSNE 2
tSNE 1
DBSCAN Clustering of tSNE map
b
tSNE 2
tSNE 1
Heirarchical Clustered genes
c
Expression
Low High
Clustered Strains
Hierarchical Clusters
PCA
d
PC2 (1.4%)
PC1 (97%)
K-means clustering of 2-dimensional tSNE
e
tSNE 2
tSNE 1
PRESTO gate outline
K-means clustering of 75-dimensional raw data
f
tSNE 2
tSNE 1
K-means cluster colors
WGCNA Clustering
g
h
tSNE 2
tSNE 1
PRESTO gate outline
WCGNA module colors

Supplemental Figure 7: Comparison of PRESTO with other analysis methods. A) Scatter plot showing the gene groupings based on the manual gating of the mouse diversity panel transciptome (Figure 3A). B) DBScan was used to automatically cluster the points on the tSNE map, where the parameters were altered so that the results best matched the visually perceived gene groupings. C) Hierarchical clustering of the 2,423 genes that passed the PRESTO filters produced 2 groups (shown below the heat map). For comparison, the group identity for each gene is shown. D) Principal component analysis of the 2,423 filtered genes shows no discernible boundary between the groups. Each gene is colored based on the scheme in (a). E) K-means clustering of the tSNE map was performed, with k=3 visible groups. Comparison of the PRESTO gates (outlines) and assigned cluster (dot color) shows some similarity between the methods. F) K-means clustering of the 2,423 genes before dimensionality reduction (based on expression in all 75 strains) was performed and shows no similarity to the PRESTO gates. G) WGCNA was performed on the filtered genes and 7 groups were automatically delineated. H) Comparison of the PRESTO gates (outlines) and assigned cluster (dot color) shows weak similarity between the methods.



Supplemental Figure 8: Proteomic signatures identify functions and predict survival in patients with clear cell renal cell carcinoma. Proteomic data of tumor-site and adjacent normal tissue from paired biopsies of 84 patients with staged clear cell renal carcinoma (ccRCC) was analyzed using PRESTO. a) Example PRESTO relative ranking map for a normal (left) and tumor-bearing (right) biopsy with demarcation of 4 gates determined from density plots (not shown). For each dot, blue (low) and red (high) colors indicate the relative abundance of that protein in an ordinal scale across all patients. b) Functional annotation of groups 3 and 4 highlight tumor-associated responses. c) The group mean expression of all genes in each group, shown for tumor stages 1 and 4. Significant upregulation of proteins in group 3. Stages 2 – 3 show a similar pattern (not shown). Paired t-test. ***, p<.001. d) Upstream regulator prediction based on protein-corresponding genes using Ingenuity Pathway Analysis that determines a z-score to indicate activity (up- or downregulation). Known oncogenes are selectively associated with group 3. e) Genes (corresponding to proteins) of group 3 were used to analyze survival in an independent test set of ccRCC patients, showing favorable survival in patients with low expression of group 3. Data from TCGA. Log-rank test, hazard ratio, and number of patients is provided.

MOVIE LEGENDS

Supplemental Movie 1: Relative ranking of gene expression in 75 mouse strains. For each gene independently (represented by a dot), the strains were ranked from that which expresses that gene the highest to the lowest. Then, the rankings of a particular strain for all genes can be displayed so that the genes that it expresses at relatively high amounts are colored red and the relatively low expressed genes are color blue.). Mouse strains were ordered based on decreasing group mean expression of the bottom right region. Data from GSE38705. See Fig. 3 and Supplemental Figure 5.

Supplemental Movie 2: Change in co-expression networks between resting and LPS-treated macrophages. 2,423 high variance genes from LPS-treated and untreated peritoneal macrophages from 75 inbred mouse strains were processed concurrently. Each gene appears twice in the input data, gene_name_LPS or gene_name_untreated but PRESTO is blinded to labels during dimensionality reduction. The locations of the baseline or LPS-treated genes are plotted on separate maps (frame 1 and 20, respectively) and images showing the positions of each gene between the conditions are interpolated to create the appearance of motion. This reveals a cohesive dynamic of the blue and green groups, while the red genes are unaffected by LPS. This movie corresponds to Fig. 3.

Supplemental Movie 1: Change in co-expression networks between resting and LPS-treated macrophages. Transcriptomes from human PBMCs at days 0, 3, and 7 after flu vaccinations were processed concurrently. 1,875 genes from each time point were selected as highly variable. Each gene appears 3 times in the processed input data (gene_name_0, gene_name_3, gene_name_7) but PRESTO is blinded to labels during dimensionality reduction. The locations of the genes from each day are plotted on separate axes (frames 1, 21, and 41) and images showing the positions of each gene between the conditions are interpolated to create the appearance of motion. This movie corresponds to Figure 5.