A novel approach for data integration and disease subtyping

Supplementary Material

Tin Nguyen[1], Rebecca Tagett[2], Diana Diaz[2], and Sorin Draghici[2,3,*]

[1] Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557
[2] Department of Computer Science, Wayne State University, Detroit, MI 48202
[3] Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48202

Contents

1 Methods details

The overall workflow of the algorithm is shown in Fig. S1. The input is a dataset (matrix) $\mathbf{E} \in \mathbf{R}^{N \times M}$, where N is the number of patients and M is the number of measurements for each patient. In the example of gene expression, N is the number of samples and M is the number of genes (or probes) measured in each sample. In short, the rows of the matrix $\mathbf{E}$ represent the patients and the columns represent the components (features). The algorithm parameters are the maximum number of clusters K (default 10) and the number of iterations H (default 200).

2 Simulation studies

In this section, we will demonstrate that the proposed approach: i) does not produce spurious clusters when the data does not contain any true classes, and ii) is able to find the correct subtypes when the data consists of distinct classes. In the first case, when the data has no structure, we show that any partitioning is unstable. In the second case, when the data consists of distinct classes, we show that the connectivity between samples is stable if and only if the partitioning is identical to the true classes.

In order to do this, we constructed 10 datasets: Gaussian1, Dataset2, ..., Dataset10, where the number in each name represents the number of classes of the dataset. Each dataset has 100 samples and $1,000$ genes. The samples are equally divided among the classes. For example, Dataset2 has two classes of size 50 and Dataset3 has three classes of size 33, 33, and 34. The dataset Gaussian1 has no distinct classes and thus will be used to demonstrate that PINS does not report false clusters. We will show that the pair-wise connectivity between samples are very unstable when the data is perturbed, regardless of the number of k. In consequences, the perturbed connectivity matrices are very different from the original connectivity matrices. This results in low AUC values for all values of k. Each of the other 9 datasets, Dataset2, ..., Dataset10, has distinct classes and thus will be used to demonstrate PINS' ability to retrieve the correct number of clusters in a mixture of data. We will show that for each of these datasets, the pair-wise connectivity is stable only when the number of clusters equals to the true number of classes.

The distribution of gene expression for the dataset Gaussian1 is shown in Figure S2A. The expression values of all genes follow a Gaussian distribution $\mathcal{N}(0,1)$ with mean 0 and variance 1. We note that the variance of the normal distribution for each gene has no impact on the result of PINS because the noise variance is set to be the median variance of the genes. For each value of k, the algorithm partitions the original data and then builds an original connectivity matrix. It then calculates the variance of each gene and the median variance σ^2. Since $\sigma_i^2 \approx 1$, $\forall i \in [1..1000]$, we have the median variance σ^2 is approximately 1. This median variance is used as the noise variance to construct 200 perturbed datasets. From the perturbed dataset, the algorithm constructs 200 connectivity matrices $\mathbf{G}_k^{(h)}$ ($h \in [1..200]$) for each value of k. The perturbed connectivity matrix is then calculated as the average of these 200 matrices, $\mathbf{A}_k = \frac{\sum_{h=1}^{200} \mathbf{G}_k^{(h)}}{200}$. For each value of $k \in [2..10]$, we have one *original* and one *perturbed connectivity matrix*.

Figure S2B shows several of the original connectivity matrices (upper row) with their corresponding perturbed connectivity matrices just below. Using the original data, when $k = 2$, the algorithm forms two clusters of approximately equal size. Perturbation of the data moves each data point around its original location, allowing it to be grouped together with any other point with the same probability. Visually, the perturbed connectivity matrix $\mathbf{A}_2$ in panel (B) shows that data points are randomly connected. This is also true for other values of $k \in [2..10]$. Thus, the perturbed connectivity greatly diverges from the original connectivity, for any value of $k \in [2..10]$, using dataset Gaussian1.

Figure S2C shows the CDF curves obtained from the difference matrices $\mathbf{D}_k$ for all values of $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the vertical axis represents $\mathbf{F}_k$ values. Figure S2D shows the area under the curve (AUC) of the CDFs. The horizontal axis shows different values of k as the numbers of clusters and the vertical axis shows the AUC values. These AUC values monotonically increase with k, and they range from 0.5 to 0.85.

To understand the variability of the AUC values, we repeat the whole process 20 times. Each time we regenerate the gene expression of the dataset Gaussian1 and recalculate the AUC values for $k \in [2..10]$. The vertical lines of Figure S2D show the 95% confidence interval of the AUC values at each value of k. We also plot the AUC values for another simulated dataset, in which the expression values of all genes are uniformly distributed on the interval $[0..1]$. The figure shows that both uniform data and Gaussian data have very similar AUC values.

Having demonstrated the behavior of PINS using data without structure, we next show that PINS determines the correct clusters using simulated datasets with separable classes. Dataset2 is created to have two classes, each with 50 samples. As shown in Figure S3A, the first class has the genes $1-100$ up-regulated while the second class has the genes $101-200$ up-regulated. Figure S3B shows several original connectivity matrices (upper row) and their corresponding perturbed connectivity matrices (lower row). When $k = 2$, the algorithm correctly separates the two classes using the original data. We see that the perturbed connectivity matrix is identical to the original connectivity matrix when $k = 2$, but when $k > 2$, the algorithm further splits each group into smaller groups of patients. For example, when $k = 3$, the upper-left cluster from the $k = 2$ result is split into two smaller groups. When the data are perturbed, however, the connectivity between data points of the same class tend to recover.

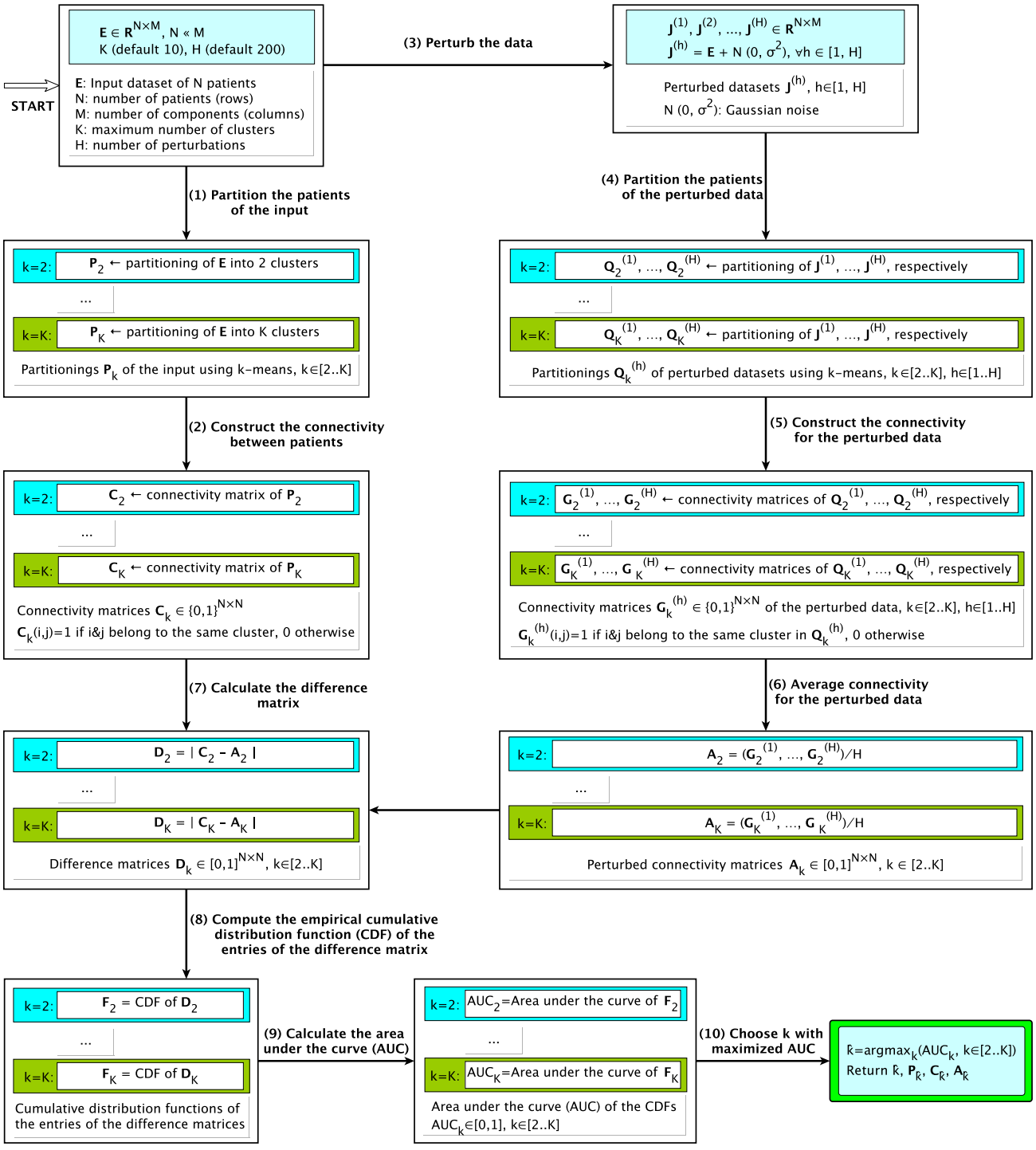


Figure S1. Perturbation clustering algorithm for high dimensional data. The data are first partitioned with different values of k (number of clusters). For each value of k, we construct the pair-wise connectivity matrix. To identify the number of clusters we add noise to the data and then build the pair-wise connectivity for the perturbed data. We calculate the discrepancy in pair-wise connectivity between before and after data perturbation. We choose $\hat{k}$ as the optimal number of clusters for which the pair-wise connectivity is the most stable.

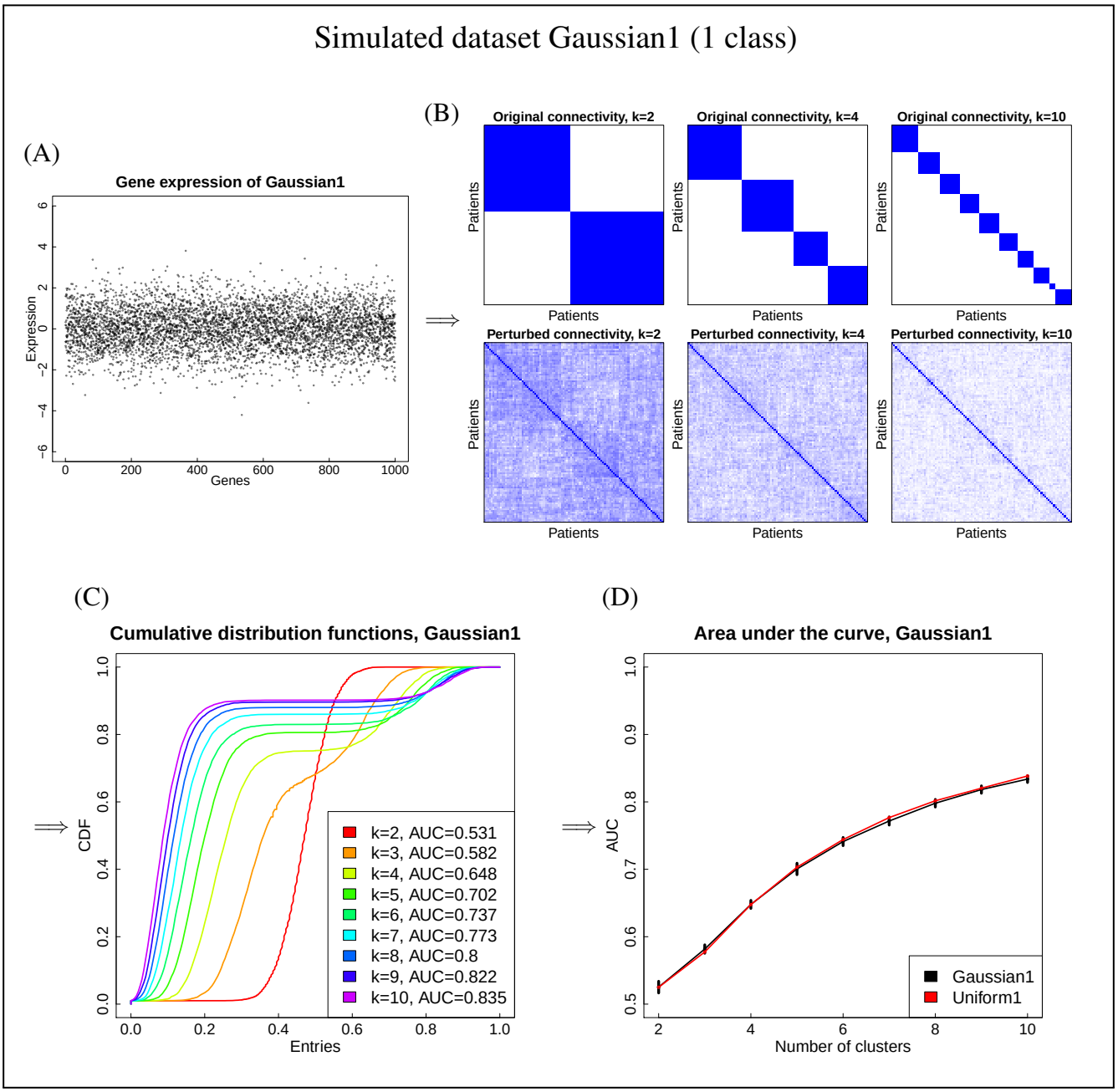


Figure S2. PINS workflow for the simulated dataset Gaussian1. The dataset consists of 100 samples and 1,000 genes. Panel (A) shows the expression profile of the dataset, in which all patients belong to one class. All gene expression values follow a normal distribution $\mathcal{N}(0,1)$ with mean 0 and variance 1. Panel (B) shows the original connectivity matrices (upper row) and perturbed connectivity matrices (lower row), for different numbers of clusters. The two left-most matrices show the original and the perturbed connectivity matrices for $k = 2$. For $k = 2$, the algorithm divides the original data into two clusters. When the data are perturbed, each data point is randomly moved around its original location and can be grouped together with any other point with the same probability. The perturbed connectivity matrix shows that the connectivity between any two data points is random, without any structure. Similarly, the perturbed connectivity matrices for $k = 4$ and $k = 10$ have no structure either. Panel (C) displays the empirical cumulative distribution functions (CDF) $\mathbf{F}_k$ of the difference matrix $\mathbf{D}_k$, $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the horizontal axis displays the values of the function (the number of elements in $\mathbf{D}_k$ smaller than or equal to each entry). Panel (D) shows the area under the curve (AUC) for each value of k. The horizontal axis shows the number of clusters and the vertical axis shows the AUC values. To assess the variability of the AUC values, we repeat the whole process 20 times with different simulated datasets having normally distributed gene expression. The vertical lines show the 95% confidence interval of the AUCs at each value of k. We also plot the AUC for a simulated dataset with uniformly distributed expression values. The figure shows that when the data are random, regardless of the distribution, the AUC values vary only slightly. In addition, the AUC values monotonically increase with k, and range from 0.5 to 0.85.

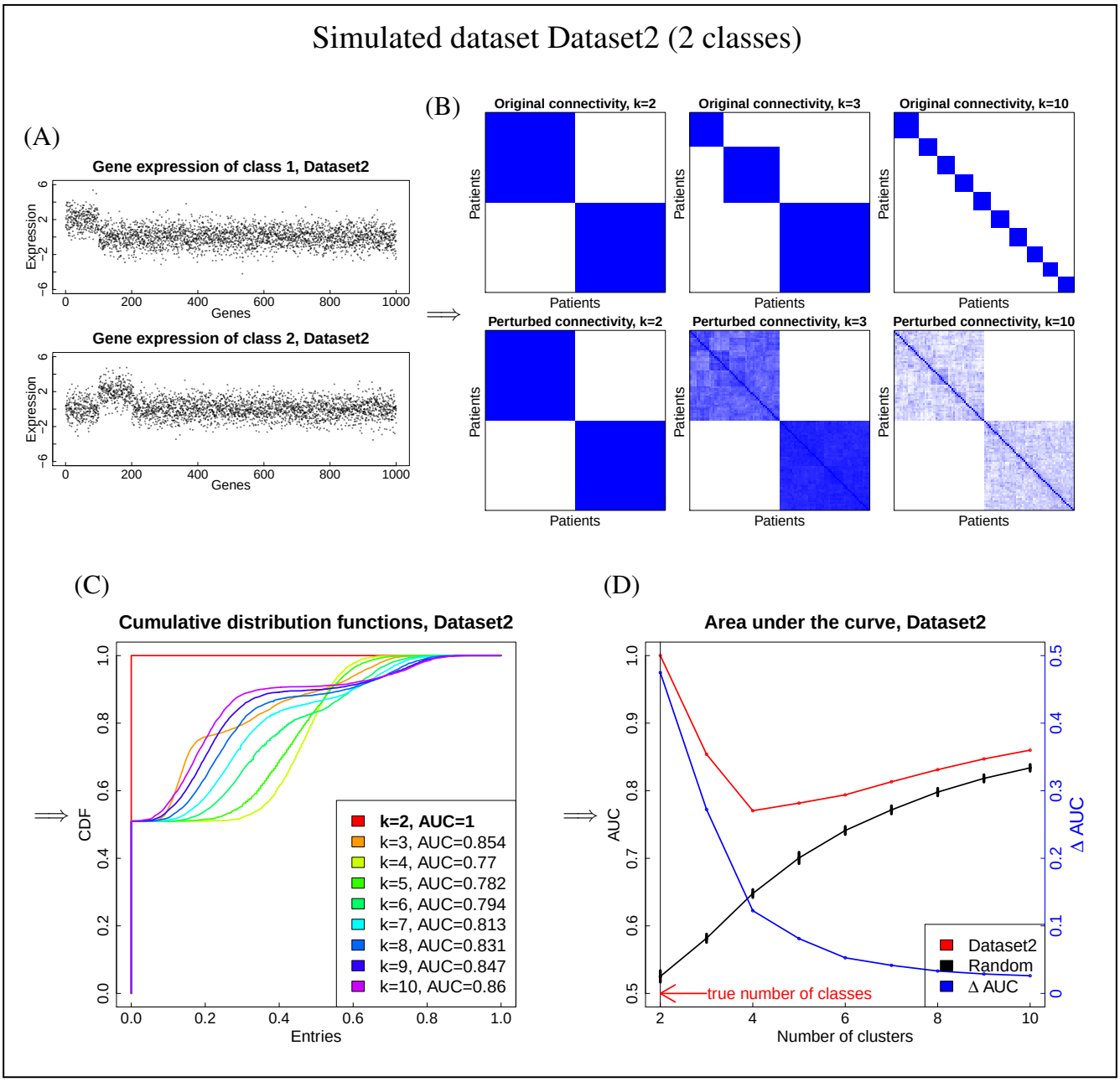


Figure S3. PINS workflow for the simulated dataset Dataset2. The dataset consists of 100 samples and 1,000 genes. Panel (A) shows the expression of the two classes. Each class has 50 samples. The first class has the genes $1-100$ up-regulated while the second class has the genes $101-200$ up-regulated. The expression of the up-regulated genes follow the distribution $\mathcal{N}(2,1)$ with mean two while the expression of other genes follow the distribution $\mathcal{N}(0,1)$ with mean 0. Panel (B) shows the original connectivity matrices (upper row) and perturbed connectivity matrices (lower row). For $k=2$, the algorithm correctly separates the two classes using the original data. As we perturb the data, each data point moves around its original position but still stays close to its own cluster. Therefore, samples of the same class are still grouped together, making the perturbed connectivity matrix identical to the original connectivity matrix. For $k>2$, the algorithm further splits each group into smaller groups. However, when the data are perturbed, samples of the same class tend to connect to each other. Regardless of k value being used, the perturbed connectivity matrices clearly suggest that the data consists two groups of samples, which is the true structure of Dataset2. Panel (C) displays the empirical cumulative distribution functions (cdf) $\mathbf{F}_k$ of the difference matrix $\mathbf{D}_k$, $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the vertical axis displays the values of the function (the number of elements in $\mathbf{D}_k$ smaller than or equal to each entry). Panel (D) shows the AUC values for Dataset2 (red curve), Gaussian1 (black curve) and the difference (blue) between the two curves. Since the original and perturbed connectivity matrices are identical for $k=2$, $\mathbf{F}_2(0)=1$ and $AUC_2=1$. The AUC curve shows that only the partitioning $\mathbf{P}_2$ is stable against data perturbation, i.e. $\hat{k}=2$. PINS correctly and deterministically discovers the true classes of the dataset Dataset2.

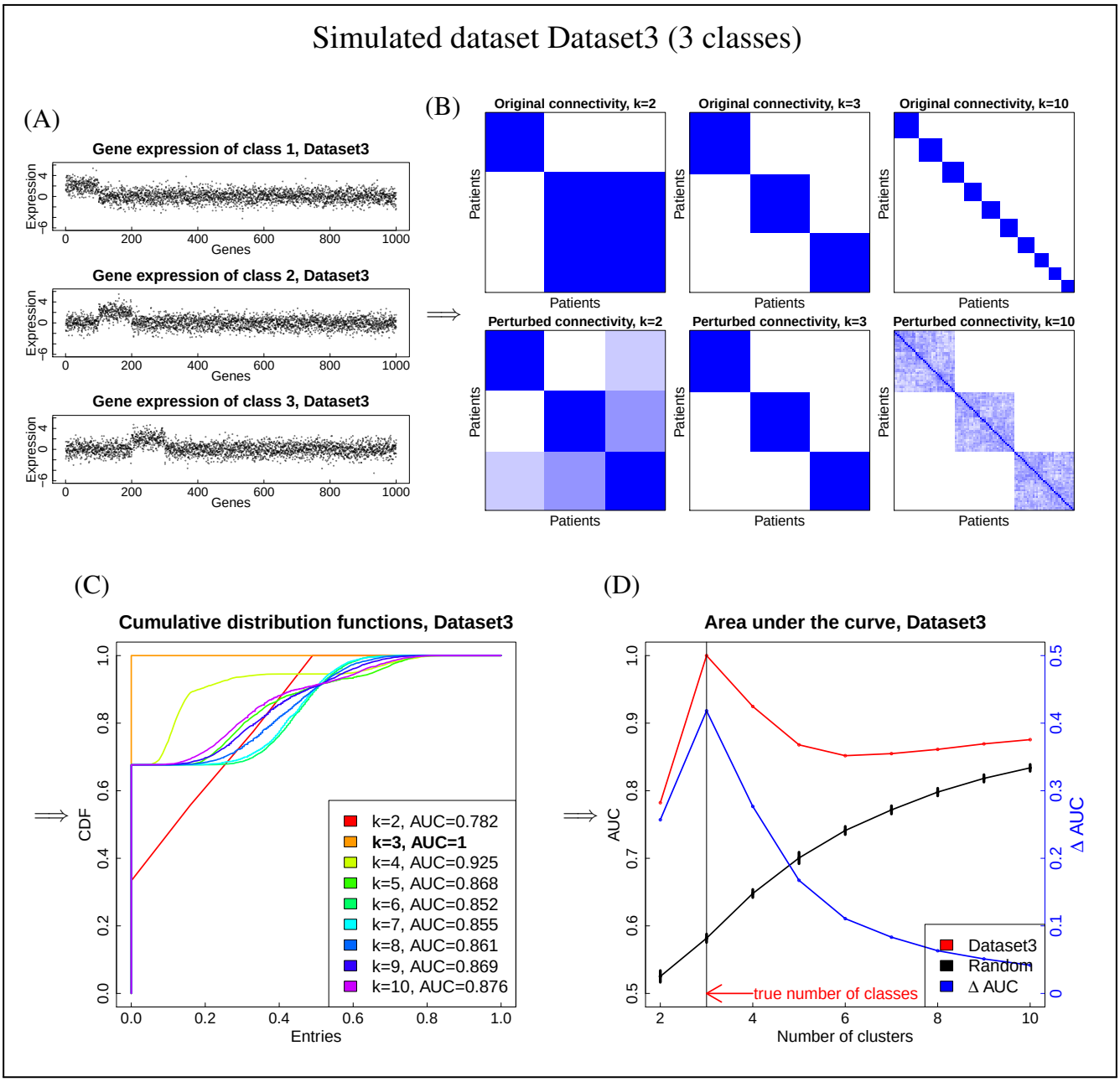


Figure S4. PINS workflow for the simulated dataset Dataset3. The dataset consists of 100 samples and 1,000 genes. Panel (A) shows the expression of the three classes. Each of the first and second classes have 33 samples while the third class has 34 samples (totally 100 samples). The first class has the genes $1-100$ up-regulated; the second class has the genes $101-200$ up-regulated; the third class has the genes $201-300$. The up-regulated genes' expression follow the distribution $\mathcal{N}(2,1)$ with mean two while other genes' expression follow the distribution $\mathcal{N}(0,1)$ with mean 0. Panel (B) shows the original connectivity matrices (upper row) and perturbed connectivity matrices (lower row). For $k=3$, the algorithm correctly separates the three classes using the original data. As we perturb the data, samples of the same class are still grouped together, making the perturbed connectivity matrix identical to the original connectivity matrix. For $k>3$, the algorithm further splits each class into smaller groups. However, when the data are perturbed, samples of the same class tend to connect to each other. For $k=2$, the original connectivity matrix $\mathbf{C}_2$ shows that two of the three classes are merged but the connectivity between them is not stable when the data are perturbed. The perturbed connectivity matrices clearly suggest that the data consists three groups of samples, which is the true structure of Dataset3. Panel (C) displays the empirical cumulative distribution functions (CDF) $\mathbf{F}_k$ of the difference matrix $\mathbf{D}_k$, $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the vertical axis displays the values of the function (the number of elements in $\mathbf{D}_k$ smaller than or equal to each entry). Panel (D) shows the AUC values for Dataset3 (red curve), Gaussian1 (black curve) and the difference (blue) between the two curves. The AUC curve shows that only the partitioning $\mathbf{P}_3$ is stable against data perturbation, i.e. $\hat{k}=3$. PINS correctly and deterministically discovers the true classes of the dataset Dataset3.

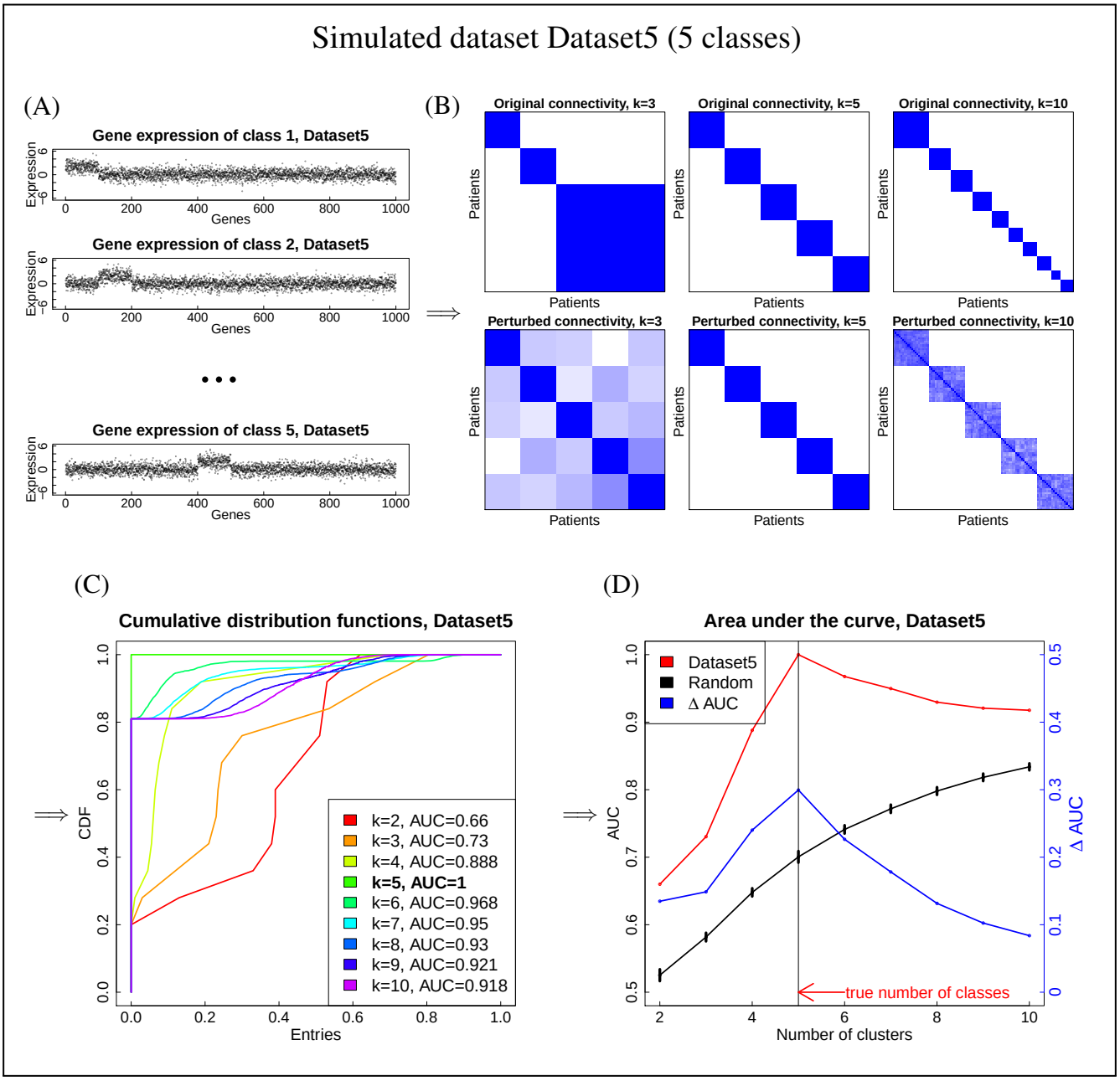


Figure S5. PINS workflow for the simulated dataset Dataset5. The dataset consists of 100 samples and 1,000 genes. Panel (A) shows the expression of the 5 classes. Each class consists of 20 samples. The i^{th} class has genes the i^{th} 100 genes up-regulated, e.g. genes $1-100$ are up-regulated in the first class and genes $401-500$ are up-regulated in the fifth class. These up-regulated genes follow the distribution $\mathcal{N}(2,1)$ with mean 2. Other genes follow the distribution $\mathcal{N}(0,1)$ with mean 0. Panel (B) shows the original connectivity matrices (upper row) and perturbed connectivity matrices (lower row). For $k=5$, the algorithm correctly separates the 5 classes using the original data. As we perturb the data, samples of the same class are still grouped together, making the perturbed connectivity matrix identical to the original connectivity matrix. For $k>5$, the algorithm further splits each class into smaller groups but samples of the same class tend to connect to each other when the data are perturbed. For $k<5$, some classes are merged together, but the connectivity between samples of different classes is not stable against data perturbation. The perturbed connectivity matrices clearly suggest that the data consists 5 groups of samples, which is the true structure of Dataset5. Panel (C) displays the empirical cumulative distribution functions (CDF) $\mathbf{F}_k$ of the difference matrix $\mathbf{D}_k$, $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the vertical axis displays the values of the function (the number of elements in $\mathbf{D}_k$ smaller than or equal to each entry). Panel (D) shows the AUC values for Dataset5 (red curve), Gaussian1 (black curve) and the difference (blue) between the two curves. The AUC curve shows that only the partitioning $\mathbf{P}_5$ is stable against data perturbation, i.e. $\hat{k}=5$. PINS correctly and deterministically discovers the true classes of the dataset Dataset5.

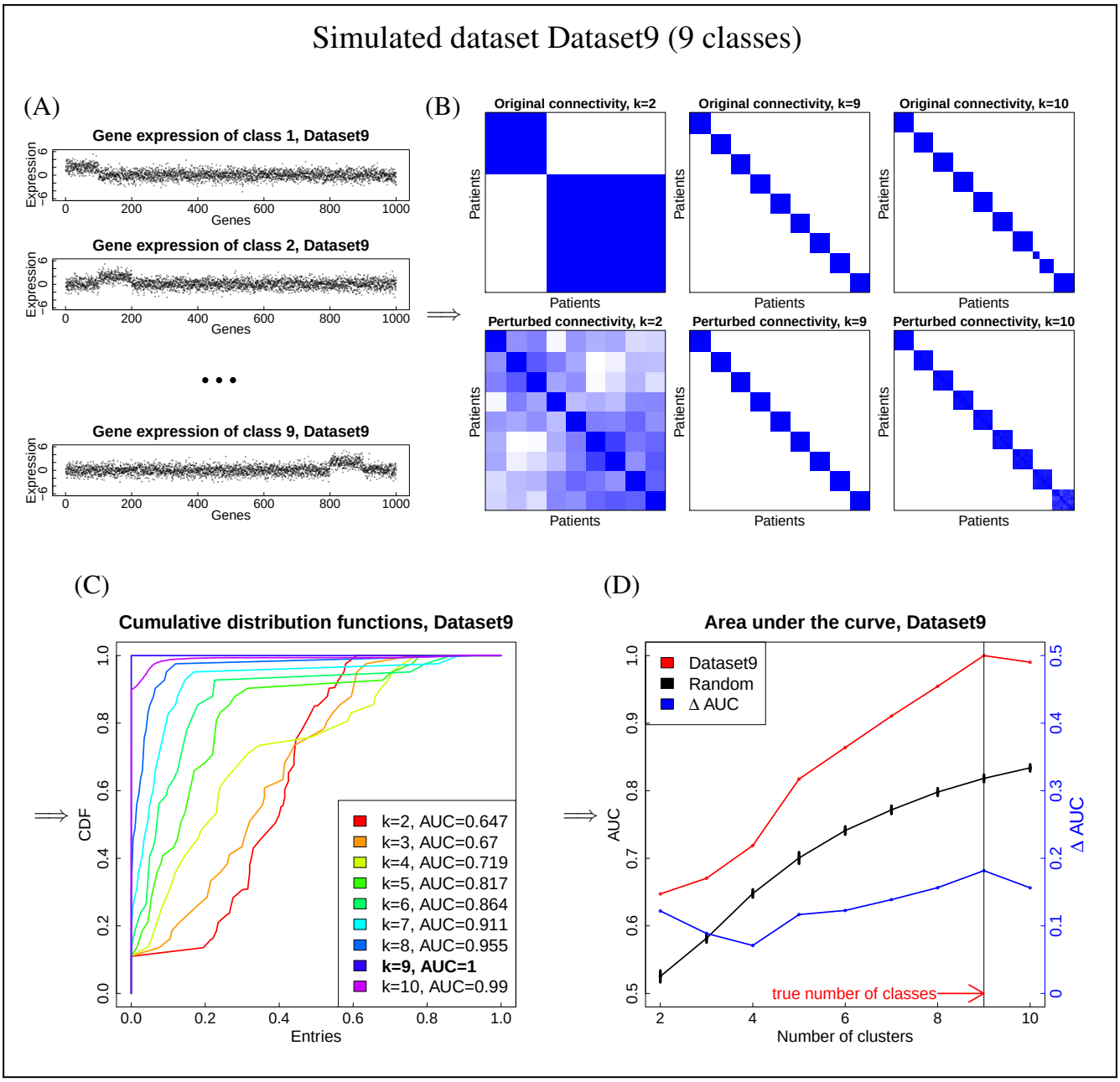


Figure S6. PINS workflow for the simulated dataset Dataset9. The dataset consists of 100 samples and 1,000 genes. Panel (A) shows the expression of the 9 classes. Each of the 8 first classes consists of 11 samples and ninth class consists of 12 samples (totally 100). The i^{th} class has genes the i^{th} 100 genes up-regulated, e.g. genes $1-100$ are up-regulated in the first class and genes $801-900$ are up-regulated in the 9^{th} class. These up-regulated genes are normally distributed with mean 2 and variance 1. Other genes are normally distributed with mean 0 and variance 1 ($\mathcal{N}(0,1)$). Panel (B) shows the original connectivity matrices (upper row) and perturbed connectivity matrices (lower row). For $k=9$, the algorithm correctly separates the 9 classes using the original data. As we perturb the data, samples of the same class are still grouped together, making the perturbed connectivity matrix identical to the original connectivity matrix. For $k=10$, the algorithm further splits a class into two smaller groups but samples of the same class tend to connect to each other when the data are perturbed. For $k<9$, some classes are merged together, but the connectivity between samples of different classes is not stable against data perturbation. The perturbed connectivity matrices clearly suggest that the data consists 9 groups of samples, which is the true structure of Dataset9. Panel (C) displays the empirical cumulative distribution functions (cdf) $\mathbf{F}_k$ of the difference matrix $\mathbf{D}_k$, $k \in [2..10]$. The horizontal axis represents the entries of the difference matrix while the vertical axis displays the values of the function (the number of elements in $\mathbf{D}_k$ smaller than or equal to each entry). Panel (D) shows the AUC values for Dataset9 (red curve), Gaussian1 (black curve) and the difference (blue) between the two curves. The AUC curve shows that only the partitioning $\mathbf{P}_9$ is stable against data perturbation, i.e. $\hat{k}=9$. PINS correctly and deterministically discovers the true classes of the dataset Dataset9.

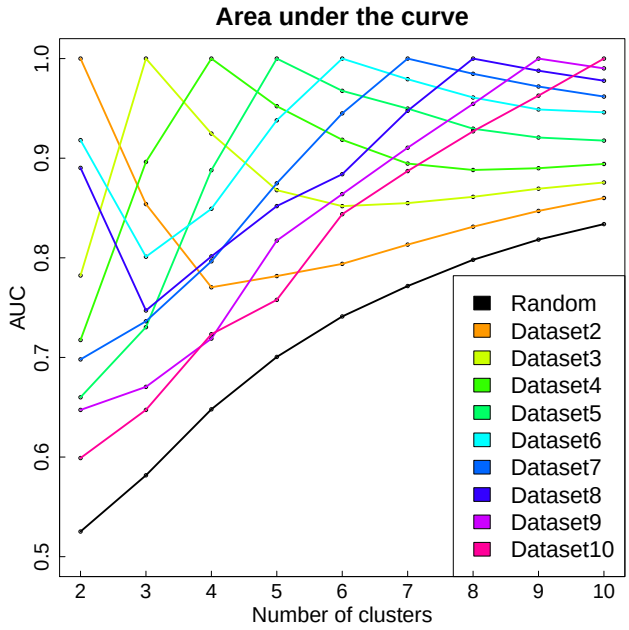


Figure S7. Area under the curve (AUC) of the 10 simulated datasets. The horizontal axis shows the number of clusters while the vertical axis shows the AUC values. The AUC values of Gaussian1 (random data) are the lowest for all values of k, and range from 0.5 to 0.85. For all other datasets, PINS correctly identifies the true number of clusters $\hat{k}$ ($AUC_{\hat{k}} = 1$). These optimal AUC values are much higher than the AUC values of the purely random dataset (Gaussian1).

Regardless of the value of k being used, the perturbed connectivity matrices clearly show that there are only two groups of strongly connected patients, reflecting the true structure of the dataset. Panel (C) shows the CDF curves obtained from the difference matrices while panel (D) shows the AUC values. Since the original and perturbed connectivity matrices are identical for $k = 2$, we have $\mathbf{F}_2(0) = 1$ and $AUC_2 = 1$. In other words, $\mathbf{P}_2$ is the only partitioning that is stable against data perturbation, and therefore $\hat{k} = 2$ is the optimal number of subtypes for the dataset Dataset2. PINS correctly and deterministically recovers the true classes of the dataset Dataset2.

Similarly, Dataset3 is created to have three classes, with 33, 33, and 34 samples, totaling 100, as before. Each class has 100 up-regulated genes, as shown in Figure S4A: gene numbers $1-100$ for the first class, $101-200$ for the second, and $201-300$ for the third. Original and perturbed connectivity matrices are shown for $k = 2$, $k = 3$, and $k = 10$ in Figure S4B. When $k = 3$, the algorithm correctly separates the data into three classes using the original data or the perturbed data. As k increases beyond $k = 3$, the non-perturbed data is split into smaller groups by the algorithm. However, when $k \neq 3$, data perturbation allows samples of the same class to connect to each other with higher probability, producing a shadow image of the correct number of classes in Figure S4B. When $k = 2$, the original connectivity matrix $\mathbf{C}_2$ shows that the second and third classes are merged, but the connectivity between them is not stable against data perturbation. All perturbed connectivity matrices clearly suggest that the data consists of three groups of samples, which is the true structure. Panels (C, D) display the CDF curves and the AUC values for different values of k. $\mathbf{P}_3$ is the only partitioning that is stable against data perturbation with $AUC_3 = 1$. PINS deterministically discovers the true classes of the dataset Dataset3.

Finally, Figures S5 and S6 display the PINS results for the simulated datasets Dataset5 (5 classes) and Dataset9 (9 classes). In both cases, the perturbed connectivity matrices clearly show the true structure of the data and PINS correctly discovers the true classes of each dataset. A plot of the AUC values for all of the 10 datasets are shown in Figure S7. When the data have no structure as in Gaussian1, the AUC values monotonically increase with k, and range between 0.5 and 0.85. When the data consist of at least two classes, the AUC values greatly increase. For any value of k, the AUC value of Gaussian1 is always smaller than the AUC value of any other dataset. PINS correctly identifies the optimal number of clusters $\hat{k}$ with $AUC_{\hat{k}} = 1$ for all 9 datasets.

3 Experimental studies

3.1 Implementation and settings

PINS was implemented in the R programming language. For Consensus Clustering (CC),[1] we used the R package ConsensusClusterPlus (version 1.24.0),[2] downloaded from the Bioconductor website. ConsensusClusterPlus returns a graph that shows the change of the area under the curve $\Delta(k)$ as the number of clusters k increases. According to the original CC manuscript,[1] the optimal number of clusters $\hat{k}$ is chosen where the area under the curve levels off and $\Delta(\hat{k})$ approaches zero. For Similarity Network Fusion (SNF), we used the R package SNFtool (version 2.1), downloaded from the website of the authors (compbio.cs.toronto.edu/SNF/SNF/Software.html). We calculate the number of clusters for SNF using the function *estimateNumberOfClustersGivenGraph*. This function returns four possible choices, in order of preference. We select the first as the best choice for the number of clusters. For iClusterPlus, we use the R package iClusterPlus (version 1.2.0), downloaded from the Bioconductor website. To choose the best k, iClusterPlus first computes the deviance ratio which is the ratio of the fitted log-likelihood - null model's log likelihood divided by the full model's log-likelihood - null mode's log-likelihood. It then chooses the value of k where the ratio levels off. For all four algorithms (PINS, CC, SNF, and iClusterPlus), we set the range for the number of clusters k to $[2..10]$.

3.2 Subtyping gene expression data

For this single data type analysis, we downloaded 8 gene expression datasets, from a variety of human cancers with known classes (subtypes). Details of the 8 datasets are described in Table S1. The 5 datasets GSE10245,[3] GSE19188,[4] GSE43580,[5] GSE15061,[6] and GSE14924[7] were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). For these datasets, the subtypes and the number of samples per subtype were collected from the description of each dataset and from the corresponding reference manuscripts. GSE10245 has a total of 58 lung cancer samples (40 adenocarcinomas and 18 squamous cell carcinomas). GSE19188 consists of 91 tumor samples (45 adenocarcinomas, 19 large cell carcinomas, and 27 squamous cell carcinomas). GSE43580 includes 150 tumor samples (77 adenocarcinomas and 73 squamous cell carcinomas). GSE15061 include 366 leukemia related samples (202 acute myeloid leukemias and 164 myelodysplastic syndromes). The fifth dataset, GSE14924, includes 20 leukemia samples (10 CD4 T cells and 10 CD8 T cells).

The other three datasets were downloaded from the Broad Institute. The dataset AML2004[8,9] was downloaded from https://archive.broadinstitute.org/cancer/pub/nmf/. Subtype information of AML2004 is described in Brunet et al.,[9] and is available in the file "ALL_AML_samples.txt" on the website. AML2004 includes 38 leukemia samples (11 acute myeloid leukemia, 19 acute lym- phoblastic leukemia B cell, and 8 T cell). The dataset Lung2001 was downloaded from http://archive.broadinstitute.org/mpr/lung/. Subtype information of Lung2001[10] is available in the file "datasetA_scans.txt" on the website. This dataset consists of 237 lung cancer samples (190 adenocarcinomas, 21 squamous cell car- cinomas, 20 carcinoid, and 6 small-cell lung carcinomas). The dataset Brain2002[11] was downloaded from https://archive.broadinstitute.org/mpr/CNS/. The subtype information of this dataset is described in Pomeroy et al.[11] (data set A) and is available in the file "Brain_samples_clinical_table.xls" on the website. This dataset consists of 42 samples (10 meduloblastomas, 10 malignant gliolas, 10 atypical teratoid/rhaboid tumors, 4 normal cerebellums, and 8 primitive neuroectodermal tumors). The dataset AML2004 was already processed and normalized and thus no further data processing was needed. For the other 7 datasets, Affymetrix *CEL* files containing raw expression data were downloaded and processed and normalized using the *threestep* function from the package *affyPLM*.[12]

Table S1. Description of the 8 gene expression datasets used in the experimental studies. The top 5 datasets were downloaded from Gene Expression Omnibus. The bottom 3 datasets were downloaded from the Broad Institute website.

Datasets	#Classes	#Samples	#Components	Platform	Description
GSE10245[3]	2	58	19851	hgu133plus2	40 adenocarcinomas and 18 squamous cell carcinomas
GSE19188[4]	3	91	19851	hgu133plus2	45 adenocarcinomas, 19 large cell carcinomas, and 27 squamous cell carcinomas
GSE43580[5]	2	150	19851	hgu133plus2	77 adenocarcinomas and 73 squamous cell carcinomas
GSE14924[7]	2	20	19851	hgu133plus2	10 acute myeloid leukemia CD4 T cell and 10 CD8 T cell
GSE15061[6]	2	366	19851	hgu133plus2	202 acute myeloid leukemia samples and 164 myelodyplastic syndrome samples
Lung2001[10]	4	237	8641	hgu95a	190 adenocarcinomas, 21 squamous cell carcinomas, 20 carcinoid, and 6 small-cell lung carcinomas
AML2004[8,9]	3	38	5000	hgu6800	11 acute myeloid leukemia, 19 acute lymphoblastic leukemia B cell, and 8 T cell
Brain2002[11]	5	42	5299	hgu6800	10 meduloblastomas, 10 malignant gliolas, 10 atypical teratoid/rhaboid tumors, 4 normal cerebellums, and 8 primitive neuroectodermal tumors

3.3 Stability of clustering methods

To investigate the stability of PINS regarding noise and the distance between the true subtypes, we performed more simulations. We used the case of Dataset9 shown in Figure S6. Each simulation dataset has 100 samples and 1,000 genes. The samples are equally divided into 9 classes. The variance of the expression level for each gene is 1. Without loss of generality, the genes can

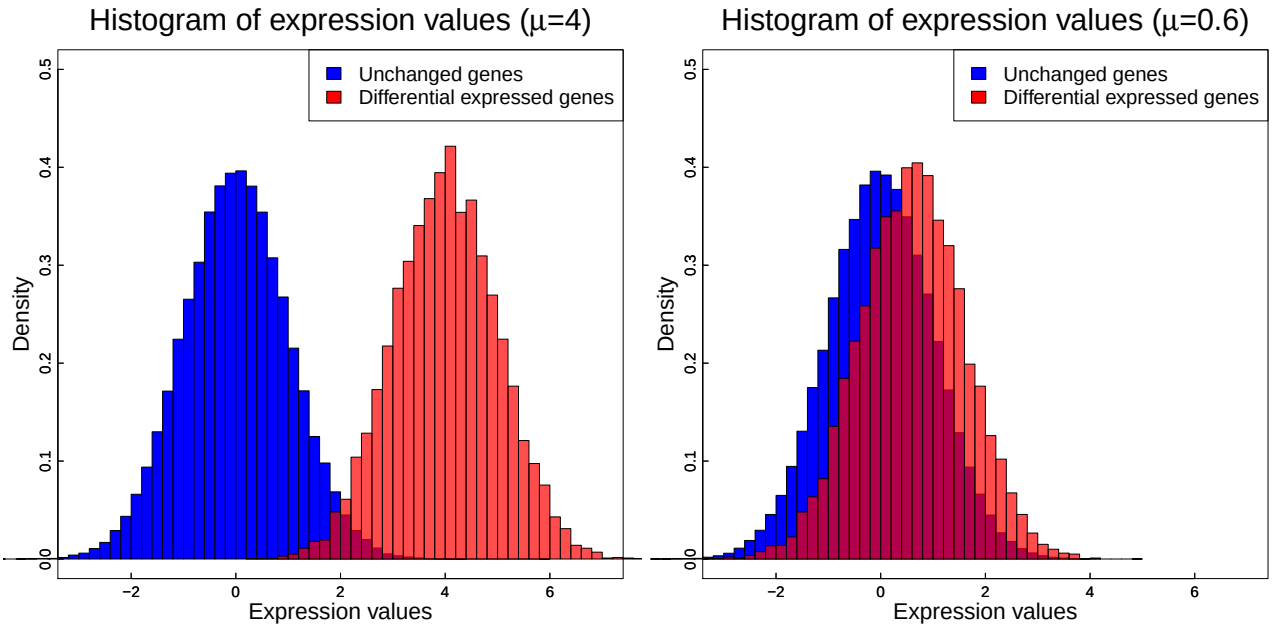


Figure S8. Histogram of expression values for up-regulated and regular genes. In each of the panel, the blue histogram represents the density of unchanged genes while the red one represents the density of up-regulated genes. The left panel displays the case when the mean difference (distance) between the up-regulated genes and unchanged genes is 4, which is much higher than gene variability (gene variance =1). The right panels shows the case when this distance is 0.6, smaller than gene variability (gene variance =1).

Table S2. Adjusted Rand Index of classes discovered by PINS, CC, SNF, and iClusterPlus for the simulation data. μ is the expression mean of the genes that are up-regulated, which is also the difference in expression between the up-regulated genes and the rest. the noise variance is the variance used for data perturbation. This noise variance is equal to the median of gene expression variances. Among all 4 methods, PINS is the most robust against changes in differential expression.

Distance (μ)	Noise variance	Clustering method			
		PINS	CC	SNF	iClusterPlus
4	2.577	1	1	0.06	0.968
3	1.877	1	1	0.88	0.966
2	1.384	1	1	0.262	0.964
1	1.08	1	1	1	-0.002
0.9	1.06	1	1	1	-0.002
0.8	1.05	1	0.908	1	-0.011
0.7	1.03	0.94	0.799	0.897	-0.0005
0.6	1.02	0.647	0.345	0.221	-0.0005

be reordered such that the genes 1–100 are up-regulated for samples in class 1, genes 101–200 are up-regulated for samples in class 2, etc. The expression of the up-regulated genes follow the distribution $\mathcal{N}(\mu,1)$ with mean μ while the expression of other genes follow the distribution $\mathcal{N}(0,1)$ with mean 0. We investigate the performance of PINS, CC, SNF, and iClusterPlus using different values of μ: 4, 3, 2, 1, 0.9, 0.8, 0.7, and 0.6. In brief, $\mu = 4$ describes a situation in which the difference between the mean expression levels of the differentially expressed genes (DEGs) and the rest of the genes is larger than the variance from one individual to another (see the left panel in Fig S8). The case $\mu = 0.6$ describes a situation in which the true differences in gene expression are smaller than the variance of the genes due to the individual differences (see the right panel in Fig S8).

Table S2 shows the adjusted Rand Index of the clustering results for these additional simulations. The data shows that the ability to discover the true subgroups degrades as their average differences (μ) become smaller compared to the intrinsic variability. This is true for PINS, CC and iClusterPlus. Unexpectedly, SNF's performance is also degrading as the distance between clusters becomes either much smaller or much larger than the gene variance. We hypothesize that this is because SNF uses a kernel-based distance with a "hyper-parameter" (authors' term). In conclusion, the data in Table S2 shows that PINS is by far the most robust among the 4 methods.

We also analyzed the gene expression datasets using different settings for SNF and iClusterPlus. SNF allows users to set several parameters, including a hyper-parameter named *alpha* that is used to compute the similarity between patients. By default, this parameter is set to 0.5, but even a slight change in this parameter will likely change the outcome of the analysis. The adjusted Rand index values for the gene expression datasets are shown in Table S4 (we note that SNF returns NA values for GSE14924) with *alpha* set to different values: 0.45, 0.47, 0.5, 0.53, 0.55. For every single dataset, the results change when we slightly alter this alpha parameter (plus/minus 0.05). For iClusterPlus, we re-run the scripts using different number of most-variable genes in all gene expression datasets, ranging from the top 2,000 most variable genes up to and including all

genes. The ARI values are shown in Suppl. Table S3 for different number of pre-selected genes. For every single dataset, the results change when the number of selected genes changes. Regardless of how we select the number genes (for iClusterPlus) and hyper-parameter (for SNF), PINS continues to outperform iClusterPlus and SNF in identifying the known subtypes for the 8 gene expression datasets.

Table S3. Adjusted Rand Index (ARI) calculated for iClusterPlus subtypes using different number of selected genes on the 8 gene expression datasets used in the manuscript.

Dataset \ #Genes	2000	3000	4000	All
GSE10245	0.43	0.13	0.25	0.34
GSE19188	0.33	0.23	0.22	0.23
GSE43580	0.19	0.2	0.19	0.34
GSE15061	0.17	0.163	0.161	0.18
GSE14924	0.73	0.38	0.47	0.25
Lung2001	0.11	0.16	0.13	0.16
AML2004	0.21	0.21	NA	NA
Brain2002	0.35	0.24	0.24	0.16

Table S4. Adjusted Rand Index (ARI) calculated for SNF subtypes using different values of alpha.

Dataset \ Alpha	0.45	0.47	0.5	0.53	0.55
GSE10245	0.374	0.333	0.375	0.334	0.334
GSE19188	0.159	0.159	0.121	0.171	0.171
GSE43580	0.177	0.177	0.154	0.154	0.154
GSE15061	0.259	0.259	0.051	0.078	0.105
Lung2001	0.296	0.279	0.279	0.287	0.283
AML2004	0.069	0.069	0.171	0.171	0.171
Brain2002	0.151	0.134	0.134	0.134	0.134

In order to understand PINS behavior when the perturbation magnitude changes, we subtype the gene expression data while setting the noise parameter (σ^2) to different values in the spectrum of gene variances. Considering a gene expression dataset with M genes. Without loss of generality, assume that the gene variances are sorted in an increasing order: $\sigma_1^2 < \sigma_2^2 < ... < \sigma_M^2$. By default the variance of the noise is set to the median, i.e. $\sigma^2 = \frac{\sigma^2_{\lfloor \frac{M}{2} \rfloor} + \sigma^2_{\lceil \frac{M}{2} \rceil}}{2}$. To demonstrate the robustness of PINS, we show here the results obtained with noise variances chosen across the entire range of gene variances: $\sigma^2_{[M/4]}$ (first quartile), $\sigma^2_{[0.3M]}$, $\sigma^2_{[0.35M]}$, $\sigma^2_{[0.4M]}$, $\sigma^2_{[0.45M]}$, $\sigma^2_{[M/2]}$ (median), $\sigma^2_{[0.55M]}$, $\sigma^2_{[0.6M]}$, $\sigma^2_{[0.65M]}$, $\sigma^2_{[0.7M]}$, $\sigma^2_{[3M/4]}$ (third quartile). The resulted ARI values are reported in Table S5. When the noise parameter varies from the first to the third quartile of the spectrum, only three ARI values change out of 88 such ARI values. This demonstrates the robustness of PINS to the perturbation magnitude.

Table S5. Adjusted Rand Index (ARI) calculated for PINS subtypes using different noise parameters. When the noise parameter varies from the first to the third quartile of the gene variance spectrum, none but 3 ARI values changed (3 out of 88). This demonstrates the robustness of PINS to the perturbation magnitude.

Dataset \ Percentile	25%	30%	35%	40%	45%	50 (median)%	55%	60%	65%	70%	75%
GSE10245	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
GSE19188	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
GSE43580	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
GSE15061	0.39	0.39	0.39	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
GSE14924	1	1	1	1	1	1	1	1	1	1	1
Lung2001	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
AML2004	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
Brain2002	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61

3.4 Subtyping TCGA data

In this section, we demonstrate ability of PINS to simultaneously integrate and subtype multiple types of data. The performance of the clustering is assessed by measuring the significance of the differences in survival between the discovered groups. Since Consensus Clustering (CC) is not designed to integrate multiple data types, we concatenate the three data types for the integrative analysis. For some cancer datasets, iClusterPlus is unable to cluster miRNA data and therefore NA values are reported. We show that PINS outperforms CC, SNF, and iClusterPlus by identifying subtypes that have more significant differences in survival profiles. We included patients that have measurements across all the three data types. The number of components for a data type is the number of measurements for a patient for that data type. The expression values of DNA methylation fall between 0 and 1 and the expression values of microarray measurements (gene expression) fall between 2 and 14. We use these data as they are without any processing or filtering. For sequencing data, since the values are too large (up to millions), we use log transformation (base 2) to re-scale the data.

Using the survival data from TCGA, we calculate the Cox log-rank test p-values[13–15] for the results of the four clustering algorithms. We note that the same Cox p-log-rank test was used to demonstrate the abilities of SNF.[16] We report the number of

Table S6. Concordance Index (CI) of discovered subtypes. The results for the integrated data are displayed in bold. The cells highlighted in green have the highest CI. After data integration, PINS finds subtypes with highest CI for five out of the six cancers (KIRC, GBM, LUSC, BRCA, and COAD).

TCGA dataset			PINS		CC		SNF		iClusterPlus		maxSilhouette	
Name	#Patients	Data type	k	CI	k	CI	k	CI	k	CI	k	CI
KIRC	124	mRNA	2	0.527	6	0.575	2	0.585	9	0.506	2	0.527
		Methylation	3	0.62	6	0.621	3	0.571	10	0.567	3	0.62
		miRNA	2	0.532	5	0.583	2	0.532	NA	NA	2	0.532
		Integration	**4**	**0.696**	**6**	**0.495**	**2**	**0.532**	**6**	**0.529**	**2**	**0.527**
GBM	273	mRNA	2	0.51	5	0.511	2	0.483	10	0.502	2	0.51
		Methylation	2	0.542	6	0.505	2	0.528	10	0.57	2	0.542
		miRNA	4	0.525	6	0.489	2	0.515	10	0.547	2	0.512
		Integration	**3**	**0.569**	**7**	**0.509**	**4**	**0.536**	**5**	**0.493**	**2**	**0.51**
LAML	164	mRNA	5	0.619	5	0.529	2	0.519	6	0.567	2	0.555
		Methylation	6	0.58	7	0.521	2	0.513	10	0.528	2	0.561
		miRNA	2	0.522	6	0.531	3	0.531	NA	NA	2	0.522
		Integration	**4**	**0.57**	**8**	**0.579**	**3**	**0.553**	**5**	**0.514**	**3**	**0.549**
LUSC	110	mRNA	3	0.576	5	0.528	3	0.573	7	0.591	2	0.559
		Methylation	8	0.622	9	0.511	2	0.529	10	0.558	2	0.504
		miRNA	2	0.549	7	0.52	2	0.609	NA	NA	3	0.552
		Integration	**5**	**0.632**	**6**	**0.554**	**3**	**0.519**	**4**	**0.529**	**2**	**0.541**
BRCA	172	mRNA	2	0.486	8	0.402	2	0.551	9	0.67	2	0.486
		Methylation	4	0.632	8	0.614	5	0.498	10	0.577	2	0.509
		miRNA	3	0.603	5	0.54	2	0.575	NA	NA	2	0.628
		Integration	**7**	**0.728**	**7**	**0.684**	**2**	**0.618**	**10**	**0.54**	**2**	**0.486**
COAD	146	mRNA	2	0.582	8	0.707	2	0.575	6	0.64	2	0.582
		Methylation	2	0.544	8	0.463	2	0.547	10	0.496	2	0.544
		miRNA	4	0.641	7	0.457	3	0.613	NA	NA	2	0.511
		Integration	**5**	**0.605**	**5**	**0.555**	**2**	**0.56**	**10**	**0.526**	**2**	**0.582**

Table S7. Silhouette index (SI) values of discovered subtypes. The results for the integrated data are displayed in bold. The cells highlighted in green have the highest Silhouette.

TCGA dataset			PINS		CC		SNF		iClusterPlus		maxSilhouette	
Name	#Patients	Data type	k	SI	k	SI	k	SI	k	SI	k	SI
KIRC	124	mRNA	2	0.369	6	0.051	2	0.09	9	0.051	2	0.369
		Methylation	3	0.1	6	-0.032	3	0.007	10	-0.032	3	0.1
		miRNA	2	0.32	5	0.002	2	0.32	NA	NA	2	0.32
		Integration	**4**	**0.025**	**6**	**-0.013**	**2**	**0.33**	**6**	**0.05**	**2**	**0.365**
GBM	273	mRNA	2	0.353	5	0.009	2	0.078	10	0.009	2	0.353
		Methylation	2	0.239	6	0.009	2	0.035	10	0.009	2	0.239
		miRNA	4	0.097	6	0.004	2	0.203	10	0.004	2	0.163
		Integration	**3**	**0.105**	**7**	**-0.016**	**4**	**0.065**	**5**	**0.029**	**2**	**0.337**
LAML	164	mRNA	5	0.097	5	0.043	2	0.108	6	0.043	2	0.109
		Methylation	6	0.071	7	-0.01	2	0.093	10	-0.01	2	0.127
		miRNA	2	0.316	6	0.038	3	0.108	NA	NA	2	0.316
		Integration	**4**	**0.062**	**8**	**0.032**	**3**	**0.087**	**5**	**0.062**	**3**	**0.11**
LUSC	110	mRNA	3	0.056	5	0.048	3	0.04	7	0.048	2	0.063
		Methylation	8	0.051	9	-0.002	2	0.022	10	-0.002	2	0.079
		miRNA	2	0.144	7	-0.006	2	0.033	NA	NA	3	0.153
		Integration	**5**	**0.037**	**6**	**0**	**3**	**0.029**	**4**	**0.017**	**2**	**0.056**
BRCA	172	mRNA	2	0.157	8	-0.011	2	0.101	9	-0.011	2	0.157
		Methylation	4	0.061	8	-0.035	5	0.016	10	-0.035	2	0.08
		miRNA	3	0.078	5	0.025	2	0.065	NA	NA	2	0.093
		Integration	**7**	**0.002**	**7**	**-0.003**	**2**	**0.133**	**10**	**0.006**	**2**	**0.156**
COAD	146	mRNA	2	0.213	8	-0.019	2	0.219	6	-0.019	2	0.213
		Methylation	2	0.179	8	-0.013	2	0.013	10	-0.013	2	0.179
		miRNA	4	0.07	7	0.027	3	0.055	NA	NA	2	0.081
		Integration	**5**	**0.086**	**5**	**-0.035**	**2**	**0.082**	**10**	**0.007**	**2**	**0.199**

Table S8. Subtyping results of PINS, CC, SNF, iClusterPlus, and maxSilhouette for the 6 cancer diseases. For each disease, the first row displays the results using mRNA, methylation data, and miRNA while the other three rows display the results using two types of data. Since iClusterPlus is unable to subtype miRNA data for KIRC, LAML, LUSC, BRCA, and COAD, the results for any combination with miRNA is shown as NA. The cells highlighted in green have Cox p-values smaller than 0.01. Cells highlighted in yellow have Cox p-values between 0.01 and 0.05.

TCGA dataset			PINS		CC		SNF		iClusterPlus		maxSilhouette	
Name	Patients	Data type	k	Cox p	k	Cox p	k	Cox p	k	Cox p	k	Cox p
KIRC	124	All	4	1.3×10^{-4}	6	0.104	2	0.138	NA	NA	2	0.176
		mRNA, methyl.	5	1.4×10^{-4}	6	0.21	2	0.4	6	0.077	2	0.176
		mRNA, miRNA	3	9.3×10^{-3}	7	0.016	2	0.138	NA	NA	2	0.176
		miRNA, methyl.	5	10^{-3}	9	0.633	2	0.492	NA	NA	2	0.138
GBM	273	All	3	8.7×10^{-5}	7	0.039	4	0.062	5	0.076	2	0.408
		mRNA, methyl.	3	9.4×10^{-4}	7	0.018	3	0.04	10	0.021	2	0.408
		mRNA, miRNA	2	0.408	8	0.211	2	0.563	7	0.117	2	0.408
		miRNA, methyl.	2	10^{-4}	6	0.058	3	0.105	5	3×10^{-4}	2	10^{-4}
LAML	164	All	4	2.4×10^{-3}	8	0.035	2	0.037	NA	NA	3	0.032
		mRNA, methyl.	10	0.029	6	0.108	3	0.004	5	0.017	2	0.058
		mRNA, miRNA	7	0.013	5	0.014	2	0.011	NA	NA	3	0.027
		miRNA, methyl.	4	0.191	7	0.022	4	0.001	NA	NA	2	0.072
LUSC	110	All	5	9.7×10^{-3}	6	0.794	3	0.428	NA	NA	2	0.172
		mRNA, methyl.	4	0.205	5	0.549	2	0.849	4	0.36	2	0.522
		mRNA, miRNA	3	0.125	6	0.435	2	0.569	NA	NA	2	0.241
		miRNA, methyl.	8	0.037	8	0.69	2	0.942	NA	NA	2	0.117
BRCA	172	All	7	3.4×10^{-2}	7	0.667	2	0.398	NA	NA	2	0.902
		mRNA, methyl.	10	1.9×10^{-3}	5	0.565	2	0.479	10	0.416	2	0.902
		mRNA, miRNA	7	0.208	8	0.376	2	0.337	NA	NA	2	0.902
		miRNA, methyl.	7	0.037	7	0.668	2	0.737	NA	NA	2	0.883
COAD	146	All	5	0.201	5	0.225	2	0.296	NA	NA	2	0.113
		mRNA, methyl.	5	0.266	5	0.225	2	0.606	10	0.445	2	0.113
		mRNA, miRNA	3	0.66	5	0.751	3	0.091	NA	NA	2	0.113
		miRNA, methyl.	3	0.66	8	0.355	2	0.108	NA	NA	2	0.678

discovered subtypes and Cox p-values for each data type as well as for the integrated data in Table 3 in the main text. The subgroups of patients obtained by PINS integration are more significantly different in their survival profiles than those obtained by CC, SNF, and iClusterPlus integration in every case. For each of the six diseases, a comparison of the survival curves using the four algorithms is shown in Figures S9–S14. Both KIRC and GBM include subtypes that were obtained by splitting a large cluster obtained from the first stage of the algorithm into smaller subtypes using the second stage of the algorithm. More details on the clinical significance of the PINS subtypes for the three diseases with the most significantly different survival profiles, KIRC, GBM, and LAML, are included in Section 4. The results using different combinations of data types are shown in Table S8.

We further compared the different methods in term of the coherence of the groups discovered may be undertaken using the concordance index (CI)[17] and silhouette scores.[18] The concordance indexes (CI) of the discovered subtypes are shown in Table S6. The concordance indexes of the subtypes discovered by PINS are better than those of the subtypes identified by SNF and iClusterPlus in all 6 datasets. PINS is also superior to CC in 5 out of the 6 datasets. The silhouette scores of the subtypes discovered by all methods are compared in Table S7. PINS scores are better for every single data set and every single data type compared to the scores of iClusterPlus and CC.

Even though we included the comparison using the concordance index and the silhouette score, we feel that the most significant comparison is that provided by the survival analysis of the subgroups discovered. Ultimately, we are interested in the ability to discover the subtypes that have the potential to make a difference in the clinical practice and from that perspective we are first and foremost interested in an approach that can distinguish between patients with the more and less aggressive disease subtypes.

To investigate how stable PINS is with respect to the agreement parameter, we re-ran our analysis using 5 different cutoffs: 0.4, 0.45, 0.5, 0.6, and 0.7. The Cox p-values are shown in Table S9 in this document. In 4 out of the 6 datasets (GBM, LAML, LUCS, COAD), there is no change whatsoever, when this threshold varies from 0.4 to 0.7. in the remaining two datasets (KIRC and BRCA), the results remain the same in 7 out of 10 cases. For KIRC, when the cutoff changes from 0.5 to 0.6, (i.e. increases our requirement for agreement), PINS does not split the female group in stage II anymore. The second case is BRCA, when the cutoff changes from 0.45 to 0.4. The low agreement cutoff made PINS cluster the patients using the strong similarity matrix when this matrix is not supported by the majority of patient pairs. Overall, the data shows a very good stability of the results with respect to the choice of this parameter. Furthermore, for all choices of this parameter, the results obtained continue to be better than those obtained with CC, SNF and iClusterPlus.

We studied the clinical information available for BRCA and we realized that most patients are estrogen receptor positive.

Table S9. Cox p-values obtained from PINS subtypes using different values of agreement cutoff. Cox p-values are highlighted in yellow if they are different from the Cox p-values obtained by using the default agreement cutoff (0.5). When the cutoff varies from 0.4 to 0.7, the results change only 3 out of 30 cases.

Dataset \ Cutoff	0.4	0.45	0.5	0.6	0.7
KIRC	1.3×10^{-4}	1.3×10^{-4}	1.3×10^{-4}	0.158	0.158
GBM	8.7×10^{-5}	8.7×10^{-5}	8.7×10^{-5}	8.7×10^{-5}	8.7×10^{-5}
LAML	2.4×10^{-3}	2.4×10^{-3}	2.4×10^{-3}	2.4×10^{-3}	2.4×10^{-3}
LUCS	9.7×10^{-3}	9.7×10^{-3}	9.7×10^{-3}	9.7×10^{-3}	9.7×10^{-3}
BRCA	0.05	0.034	0.034	0.034	0.034
COAD	0.201	0.201	0.201	0.201	0.201

Out of 172 patients, there are 34 ER-negative (ER-), 134 ER-positive (ER+) and 4 not evaluated. Tables S10–S13 show the comparisons between ER subtypes and subtypes discovered by PINS, CC, SNF, and iClusterPlus. These approaches perform poorly on this breast cancer dataset (Cox p-value=0.034, 0.667, 0.398, 0.416 for PINS, CC, SNF, iClusterPlus, respectively) partially because most patients belong to the ER+ subtype.

Table S10. Confusion matrix between ER subtypes and groups discovered by PINS.

ER \ PINS groups	1	2	3	4	5	6	7
ER-	27	1	0	0	1	1	4
ER+	4	20	13	39	23	16	19

Table S11. Confusion matrix between ER subtypes and groups discovered by CC.

ER \ CC groups	1	2	3	4	5	6	7
ER-	30	0	0	0	1	0	3
ER+	15	33	27	24	16	3	16

Table S12. Confusion matrix between ER subtypes and groups discovered by SNF.

ER \ SNF groups	1	2
ER-	31	3
ER+	8	126

Table S13. Confusion matrix between ER subtypes and groups discovered by iClusterPlus.

ER \ iClusterPlus groups	1	2	3	4	5	6	7	8	9	10
ER-	8	1	7	0	0	1	0	16	0	1
ER+	2	12	7	18	20	26	19	1	11	18

3.5 Subtyping METABRIC data

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) breast cancer dataset[19] consists of a discovery cohort (997 patients) and a validation cohort (995 patients). For each of these patients, matched DNA and RNA were subjected to copy number analysis and transcriptional profiling on the Affymetrix SNP 6.0 and Illumina HT 12 v3 platforms, respectively. We downloaded the normalized data from the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/) with accession IDs: EGAD00010000210 (expression data, discovery), EGAD00010000214 (CNV, discovery), EGAD00010000211 (expression data, validation), and EGAD00010000216 (CNV, validation). The only preprocessing step we did is to map CNVs to genes using the CNTools package.[20]

We also downloaded high quality follow up clinical data from cBioPortal (http://www.cbioportal.org/). There are patients that were followed up to almost 30 years. The clinical data include PAM50 subtypes, overall survival, as well as disease free survival (DFS) information. For the discovery set, the clinical data of all of the 997 patients are available. For the validation set, there are high quality clinical data for 983 patients. Among these 983 patients, there are 6 that are not classified (NC) by PAM50.[19] For PINS, CC, and SNF, we analyze the data without gene filtering. For, iClusterPlus, we used the 2000 features with largest median absolute deviation for each data type as we did throughout the data analysis.

3.6 Silhouette index for high-dimensional data

Silhouette index offers valuable information for unsupervised clustering, to measure how well the resulted clusters are separated. However, the usefulness of the Silhouette index (SI) is somewhat limited for high-dimensional data due to noise. In particular, we will show here that when the number of dimensions increases, the silhouette values decreases and approach zero, regardless of the number of clusters. As a consequence, the silhouette values are not a good criterion to use for real multi-dimensional

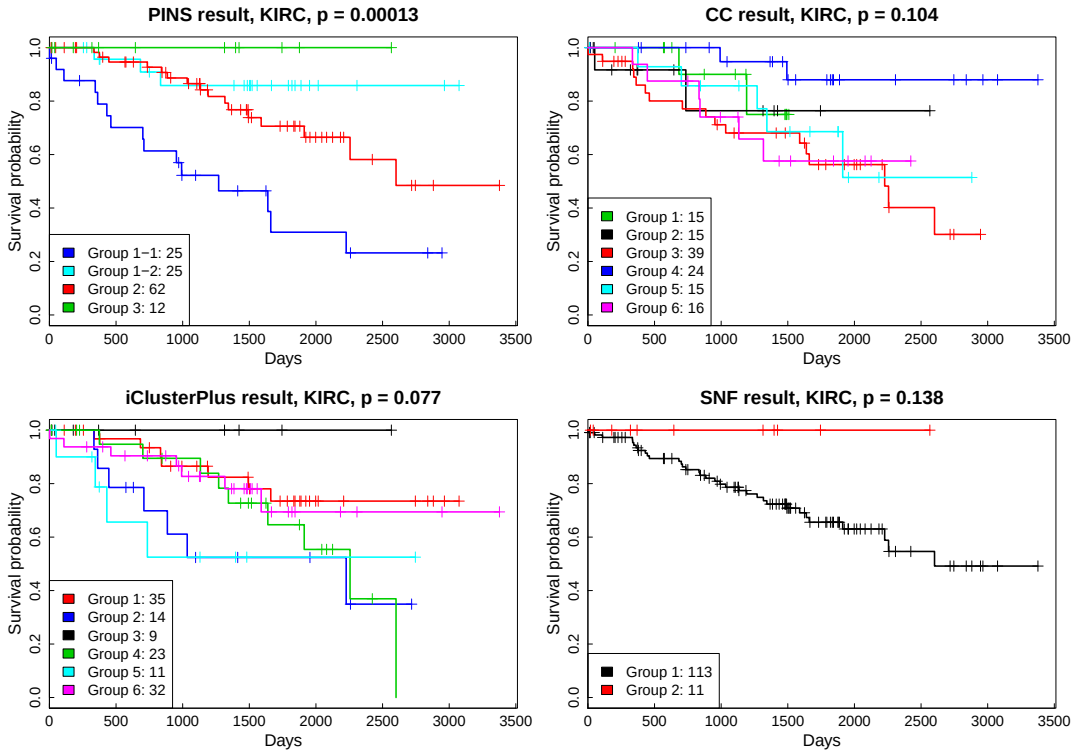


Figure S9. Kaplan-Meier survival analysis for kidney renal clear cell carcinoma (KIRC). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. SNF finds two groups while CC and iCluster find 6 groups. The survival profiles of the groups discovered by each of the three methods are not significantly different. In contrast, PINS discovers 4 groups with very different survival profiles ($p = 1.3 \times 10^{-4}$).

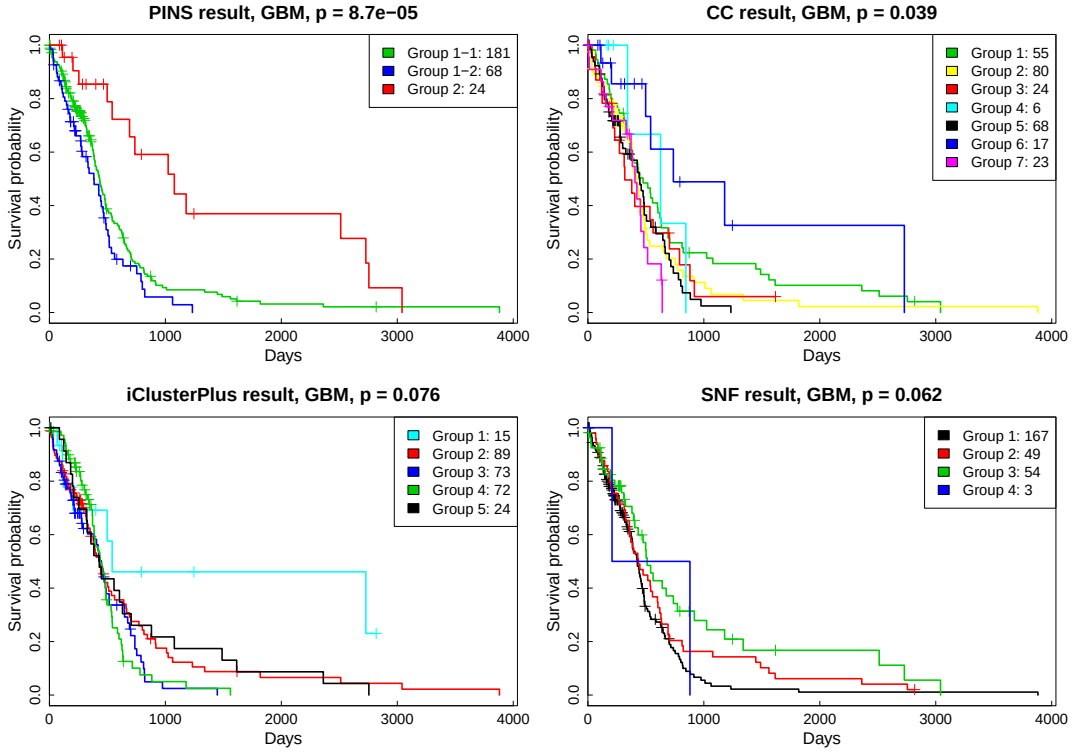


Figure S10. Kaplan-Meier survival analysis for glioblastoma multiforme (GBM). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. SNF and iClusterPlus discover 4 and 5 groups, respectively, with no significantly different survival profiles. CC finds 7 groups with significant different survival profiles ($p = 0.039$). PINS discovers three different groups with very different survival profiles ($p = 8.7 \times 10^{-5}$).

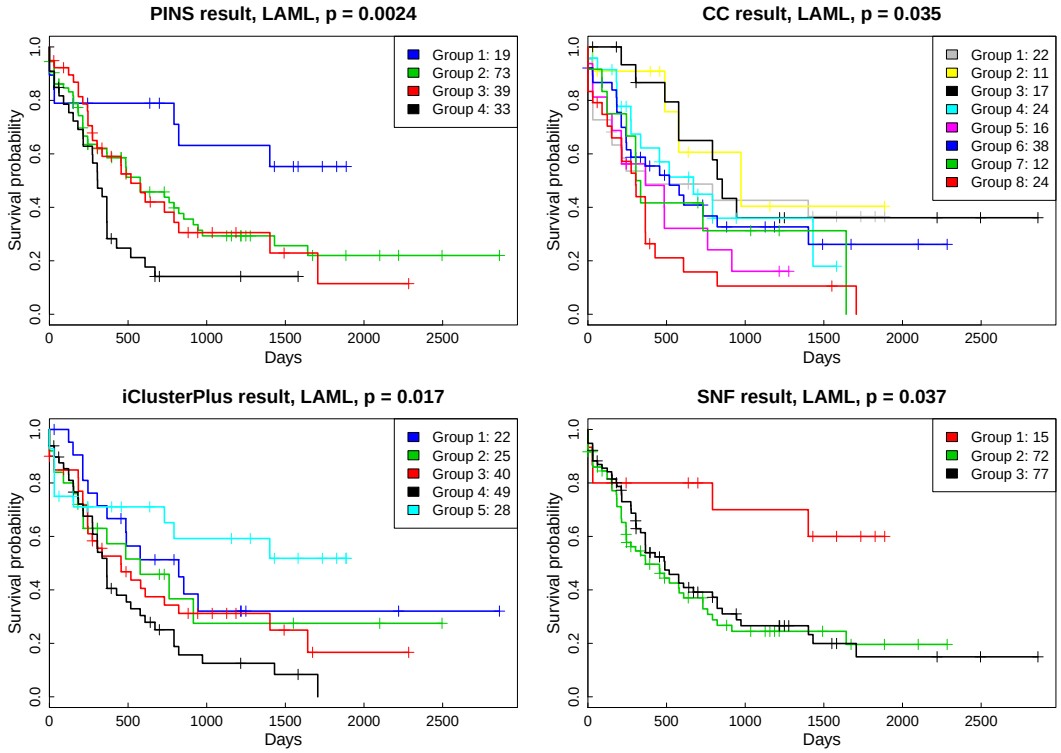


Figure S11. Kaplan-Meier survival analysis of acute myeloid leukemia (LAML). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. All the four methods discover groups of patients that have different survival profiles.

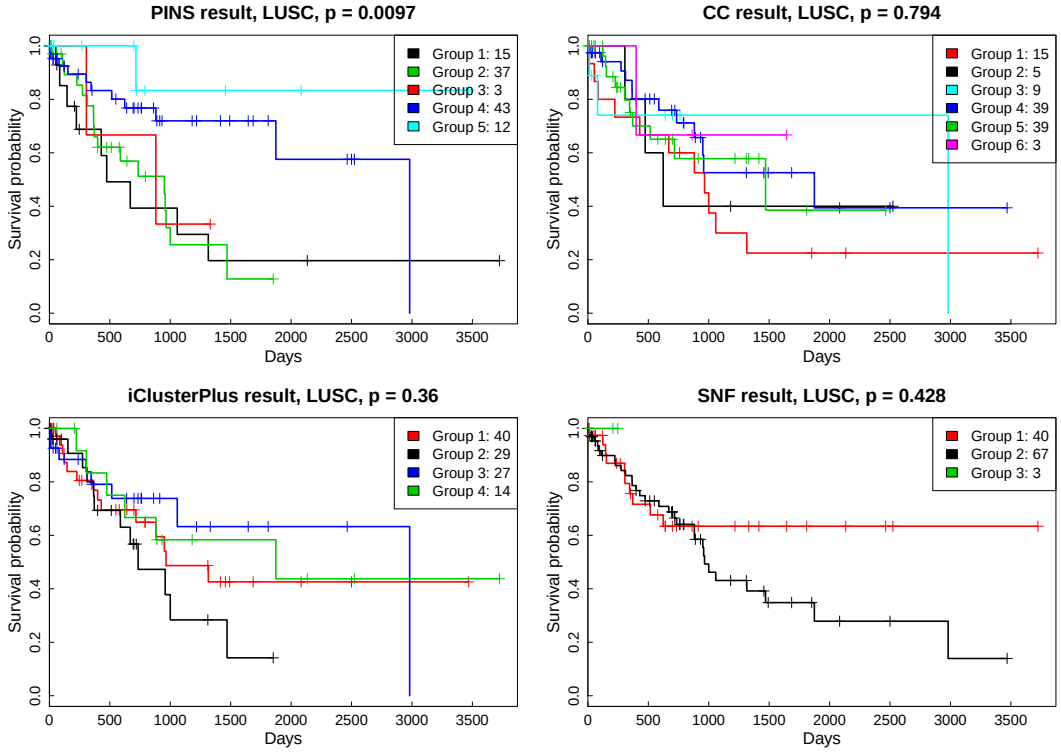


Figure S12. Kaplan-Meier survival analysis for lung squamous cell carcinoma (LUSC). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. CC, iClusterPlus, and SNF finds 6, 4, and 3 groups, respectively, with no significantly different survival. In contrast, PINS discovers 5 different groups with different survival profiles ($p = 9.7 \times 10^{-3}$).

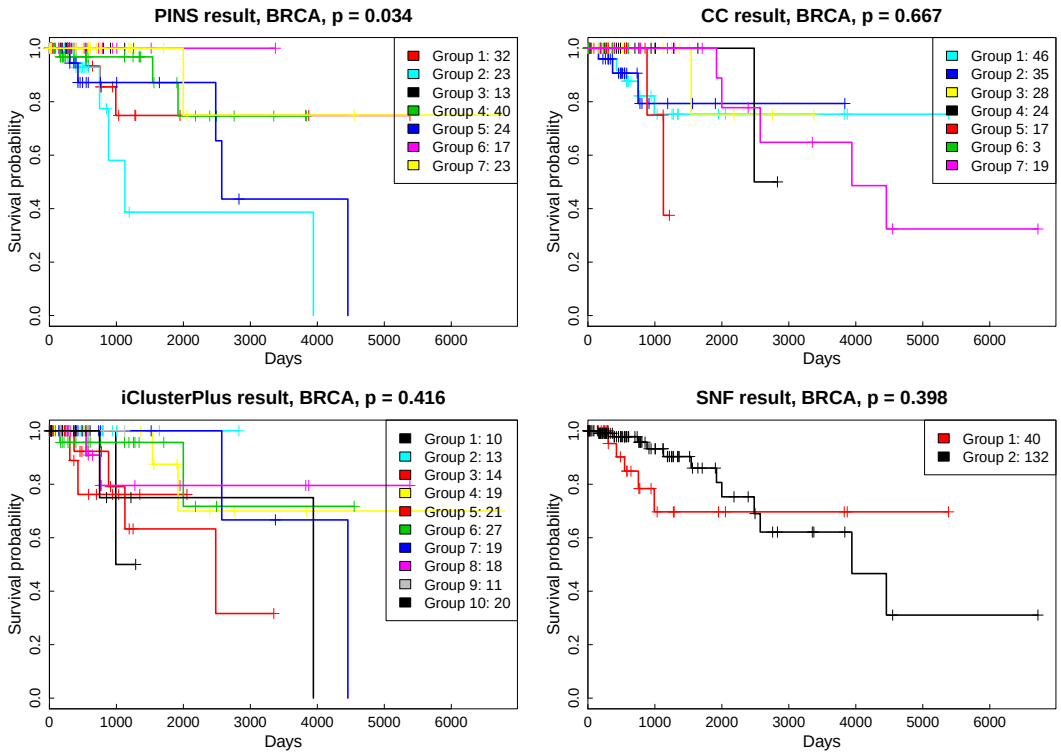


Figure S13. Kaplan-Meier survival analysis for breast invasive carcinoma (BRCA). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. Only PINS discovers subtypes with different survival profiles ($p = 0.034$).

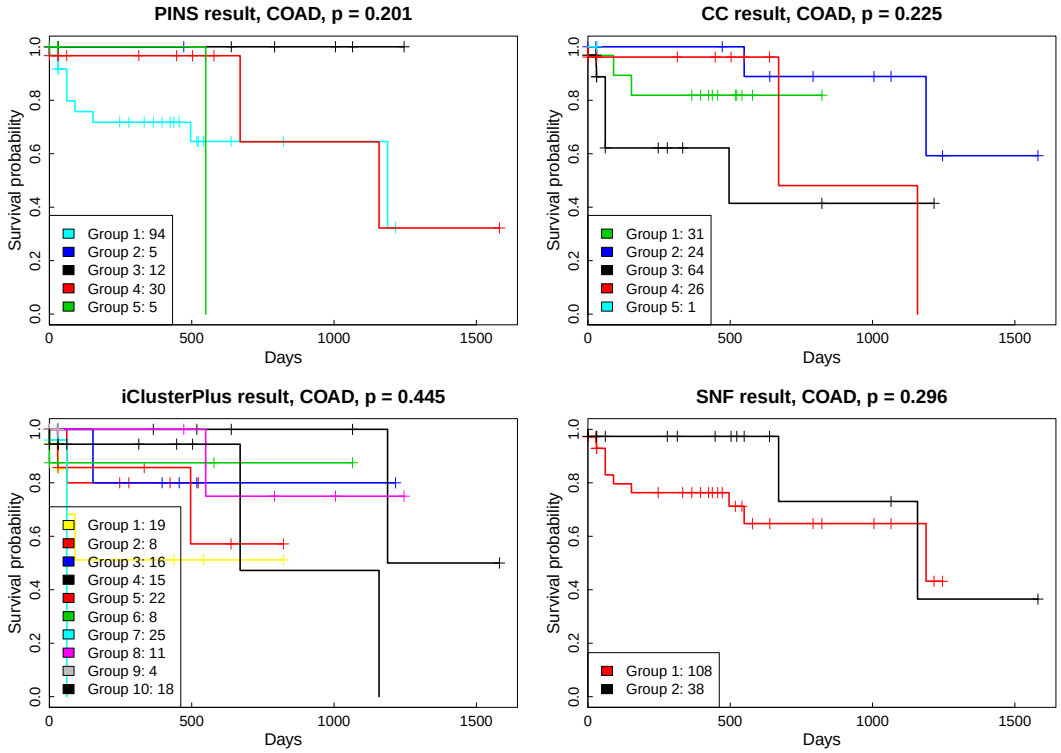


Figure S14. Kaplan-Meier survival analysis for colon adenocarcinoma (COAD). The horizontal axis represents the time passed after entry into the study while the vertical axis represents estimated survival percentage. For all four methods, the discovered groups do not exhibit significant differences in survival.

datasets. We demonstrate this by showing the results obtained when by maximizing the silhouette scores. For all TCGA datasets analyzed, this maxSilhouette approach yielded the highest silhouette scores but the clusters obtained in each case did not show significant differences in survival and were significantly inferior to the results obtained with PINS.

3.6.1 Behaviour of the silhouette scores with increased dimensionality.

We used the examples similar to Dataset9 shown in simulation studies (Section 2). The simulation dataset has 200 samples and 1,000 genes. The samples are equally divided into 9 classes. The variance of the expression level for each gene is 1. Without loss of generality, the genes can be reordered such that the genes 1–100 are up-regulated for samples in class 1, genes 101–200 are up-regulated for samples in class 2, etc. The expression of the up-regulated genes follow the distribution $\mathcal{N}(\mu,1)$ while the expression of other genes follow the distribution $\mathcal{N}(0,1)$.

We first set $\mu = 4$ and the number of genes to $1{,}000$. Figure S15A shows the expression values for each class. The histogram of the gene expression values are shown in Figure S15B. We use k-means to cluster the data using different number of clusters, and then compute the silhouette values of the resulted partitionings. The silhouette values are shown in the top left panel in Figure S16. In this particular case, the silhouette score is indeed highest when the number of cluster equals to the true number of the classes, as expected.

We next increase the number of dimensions by adding more unchanged genes to the dataset. The expression values of the added genes follow the standard normal distribution $\mathcal{N}(0,1)$. The higher the number of dimensions, the more noise we have in the data, and the harder it is to separate the true subtypes. We generated 2 more cases with 3,000 and 5,000 genes. The histograms of the expression values for 5,000 genes are show in Figure S15C. These distributions are similar to those used for 1,000 genes. Again, we use k-means to partition the data and compute the silhouette values, which are shown in the top row of Figure S16. Overall, **the silhouette values decrease when the number of dimensions increases**. Furthermore, **the maximum values of the silhouette does not correspond anymore to the true number of clusters**. For the same difference in means between the upregulated and non-regulated genes ($\mu = 4$ vs $\mu = 0$, respectively), the maximum silhouette indicates 10 clusters for 3,000 genes and 8 clusters for 5,000 genes, when the true number of clusters is 9.

With the same procedure, we set μ to other values: 3, 2, and 1. The distributions of the expression values for $\mu = 1$ are shown in panels D, E, and F of Figure S15. The silhouette values are shown in Figure S16. We see that, for any given difference in means, the silhouette values decrease when the number of dimensions increases from 1,000 to 5,000 genes. When $\mu = 1$, the silhouette values are also very low even with 1,000 genes.

In the TCGA data, the number of dimensions of each data type can be as large as 24,454. The low silhouette values for the clustering results (see Table S7) are mostly due to the noisy nature of the high-dimensional data. As illustrated above, silhouette scores become less reliable when the distance between the true subtypes are eclipsed by the noisy nature of high-dimensional data. We note that in all of the datasets involved in the simulations described above, PINS is able to recover the true structure of the data and identify the correct number of clusters, while the maxSilhouette does not.

Regarding data integration, we follow the same strategy. We try to identify pair-wise connectivities that are not only robust against noise, but also consistent across multiple data types and multiple clustering algorithms (ensemble of hierarchical clustering, partitioning around medoids,[21] and dynamic tree cut[22]). This strategy does not necessarily maximize the silhouette. Our hypothesis is that groups of patients that are strongly connected from many perspectives (data types) might be correlated with some important clinical variables. As shown in both TCGA and METABRIC datasets, the groups of patients identified by PINS have significantly different survival profiles.

3.6.2 Maximizing the silhouette index does not translate to survival differences.

To further demonstrate our point, we investigate a new clustering method named "maxSilhouette" using the TCGA data. For this method, we use k-means as the clustering algorithm and the silhouette index as the objective function to identify the optimal number of clusters. For the omics data, we integrate different types of data by concatenating the data types. The resulted number of clusters and silhouette values are shown in Table S7. In terms of silhouette value, maxSilhouette outperforms all existing methods in all but one case (23/24). This is expected because maxSilhouette aims to maximize the silhouette values. However, higher silhouette values do not necessarily translate into better clinical correlation, especially for data integration. As shown in Table 3 in main text, PINS finds subtypes with significantly different survival for five out of the six cancers while the maxSilhouette method succeeds for only one. Similarly, in terms of concordance index (see Table S6), PINS outperforms maxSilhouette in all of the six cancers.

3.7 Time complexity

The data analysis is done on a Linux server X8OBNF, Intel E7-8837 that has 1TB RAM (64 X 16GB DDR3, 1067MHz) and multi-core CPU (64 cores, 8 chips, 8 cores/chip, Intel Xeon E7-8837, 2.67GHz). The running time of each subtyping method for the 14 datasets is reported in Table S14. PINS, CC, and SNF were run using only 1 core/CPU whereas iClusterPlus were run using 60 cores. Both SNF and CC are the fastest among the four methods. SNF needs less than 1 minute while CC requires

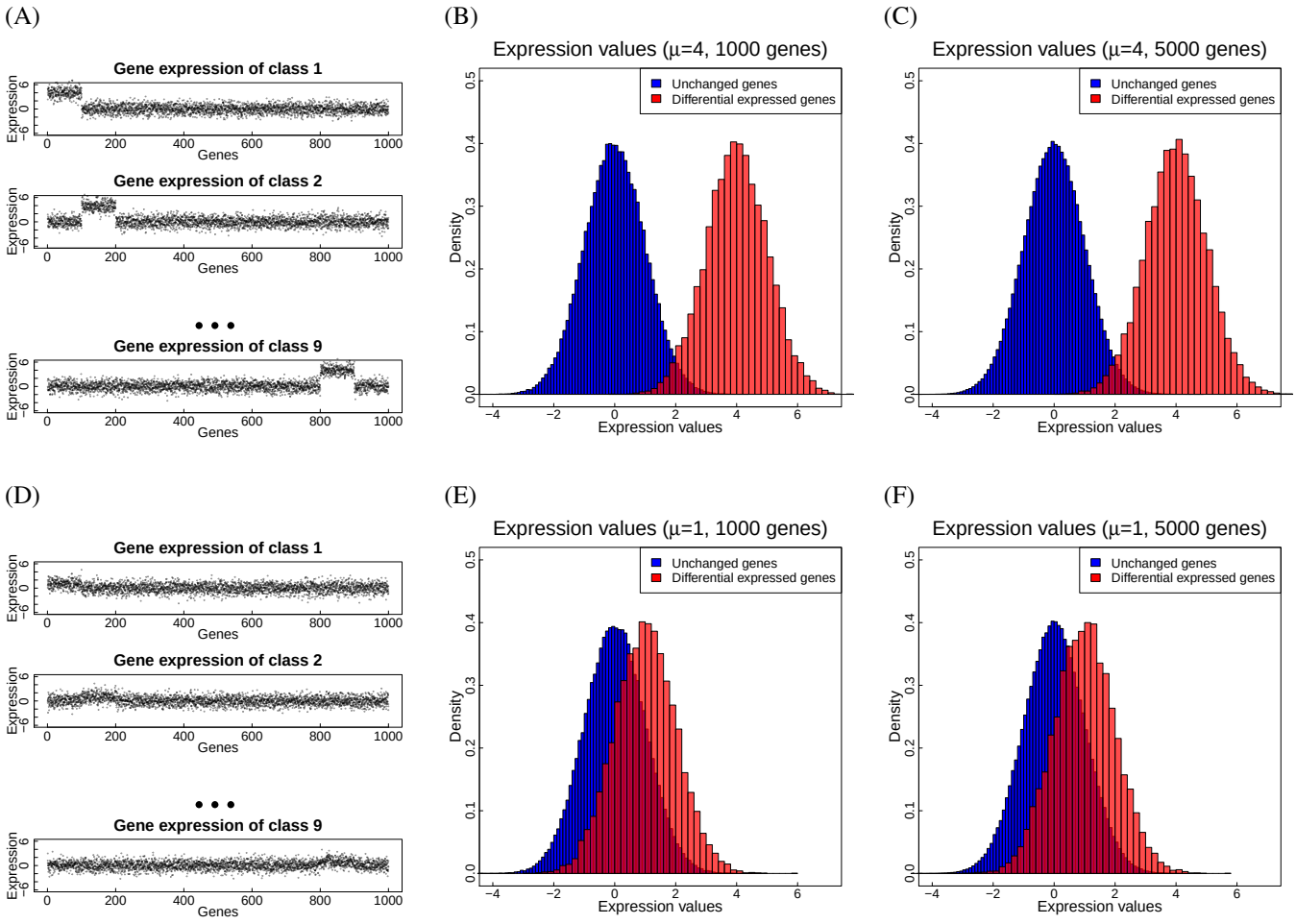


Figure S15. Expression values for simulated data. (A) The dataset has 200 samples and 1,000 genes. Without loss of generality, the genes can be reordered such that the genes 1–100 are up-regulated for samples in class 1, genes 101–200 are up-regulated for samples in class 2, etc. The expression of the up-regulated genes follow the distribution $\mathcal{N}(4,1)$ with mean $\mu = 4$ while the expression of other genes follow the distribution $\mathcal{N}(0,1)$ with mean 0. (B) Histogram of gene expression values. The blue histogram represents the density of unchanged genes while the red one represents the density of up-regulated genes. (C) Histogram of gene expression values when the number of genes increases to 5,000. The density functions of the genes are similar with different number of dimensions. (D, E, F) Gene expression values and their histogram for $\mu = 1$.

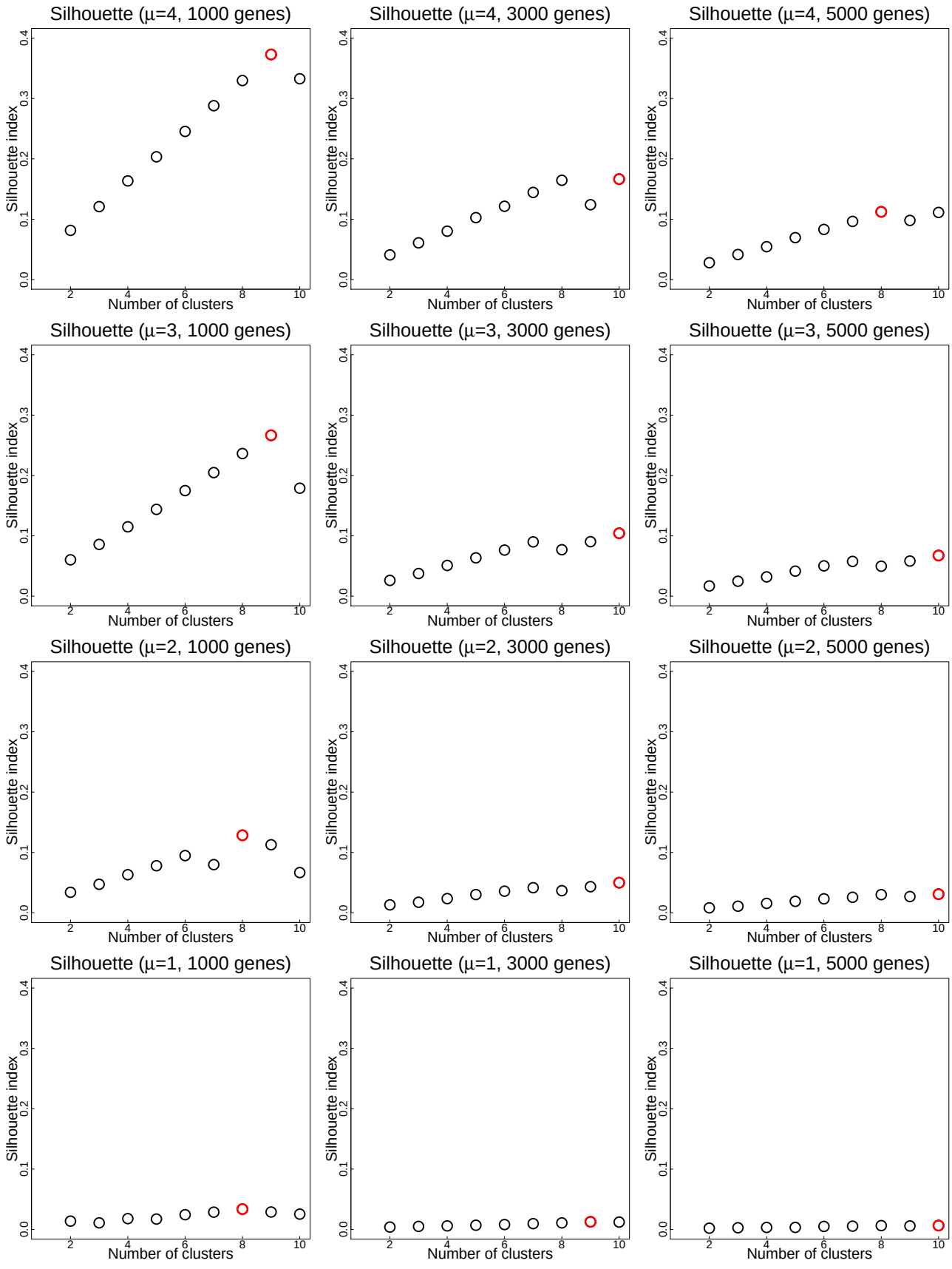


Figure S16. Silhouette index obtained from 16 simulated datasets. The red circles indicate the maximum silhouette value in each case.

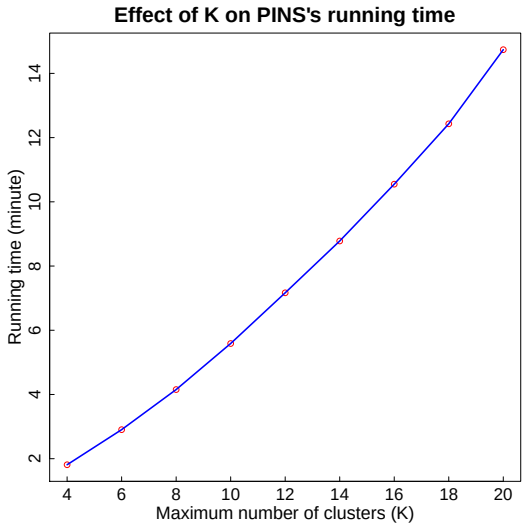


Figure S17. Running time of PINS for different settings of maximum number of clusters. The simulated dataset has 200 samples and 10,000 genes. The horizontal axis shows the number of clusters while the vertical axis shows the running time in minutes.

several minutes to subtype the patients in each of the datasets analyzed. The running time of PINS ranges from 1 minute to 13-14 hours. The running time of iClusterPlus may be up to 12-13 hours even with 60 cores. Among the four tools, only iClusterPlus allows for parallel computing.

By default, we set K (maximum number of clusters) to 10. We did this because we feel it is highly unlikely that more than 10 meaningful subtypes will be present in most cases. However, users are free to set K to other values. To understand how this setting would influence the runtime of the algorithm, we tested the algorithm with different values of K on a simulated dataset that has 300 patients and 10,000 genes. Fig. S17 shows the runtime of PINS when the maximum number of clusters varies from 4 to 20. We can see that the running time increases almost linearly when K increases from 4 to 20. The running time increases from 6 minutes for K=10 to 14 minutes for K=20.

Table S14. Running time of each subtyping method. PINS, CC, and SNF were run using only 1 core whereas iClusterPlus were run using 60 cores. The time is rounded to minutes (m).

Datasets	#Patients	PINS	CC	SNF	iClusterPlus (60 cores)
GSE10245	58	2m	< 1m	< 1m	23m
GSE19188	91	4m	< 1m	< 1m	36m
GSE43580	150	9m	< 1m	< 1m	60m
GSE15061	366	75m	2m	< 1m	146m
GSE14924	14	< 1m	< 1m	< 1m	8m
Lung2001	237	27m	< 1m	< 1m	93m
AML2004	38	< 1m	< 1m	< 1m	15m
Brain2002	42	< 1m	< 1m	< 1m	17m
KIRC	124	25m	2m	< 1m	298m
GBM	273	175m	4m	< 1m	750m
LAML	164	35m	2m	< 1m	390m
LUSC	110	18m	1m	< 1m	268m
BRCA	172	42m	2m	< 1m	516m
COAD	146	29m	2m	< 1m	348m
METABRIC discovery	997	836m	42m	3m	895m
METABRIC validation	983	853m	44m	3m	876m

4 Functional analysis of TCGA subgroups

We choose three of the TCGA diseases with the best Cox p-values, Kidney Renal Clear Cell Carcinoma (KIRC), Glioblastoma Multiforme (GBM), and Acute Myeloid Leukemia (LAML), and report the results of clinical correlations and functional analyses for the discovered subtypes. TCGA provides clinical parameters for the samples; sometimes there are many, sometimes they are sparse. We focus mostly on gene expression data because it is the most comprehensive of the data types.

PINS subgroups are investigated in pairs selected according to the data distributions, to avoid confounding data bias. As a caveat, note that, since we are making comparisons between disease subtypes, and not with respect to normal control tissue, the hypotheses that we generate are relative (between the subtypes). For example, if subtypes 'A' and 'B' are being compared, and we find N genes "up-regulated" in 'A', we simply mean that these genes are higher in 'A' than in 'B', without necessarily meaning that these genes or any subset of them are (i) up-regulated compared to normal, (ii) down-regulated compared to normal, or (iii) up in 'A' but down in 'B' compared to normal. Therefore, throughout the text, we refer to "up-regulated" and "down-regulated", to simply mean gene expression in the context of 'A' relative to 'B'. Table S15 shows the numbers of genes that are up-regulated and down-regulated in the survival groups that we compared. Note that some comparisons are gender specific, because as we will show, gender is a confounding variable in these cases. The table also shows the comparisons that we made for the three diseases, and the number of differentially expressed genes used in the group comparisons.

Table S15. Number of differentially expressed genes (FDR-corrected p-value $< 1\%$) in the pairwise subtype comparisons. Note that some comparisons are gender specific; that will be discussed in the sections below. Since these are comparisons of subtypes, we assign the subtype with the poorer survival to "case", and the better survival subtype to "control". The log fold change is calculated as $case - control$, therefore "up-regulation" here means that log fold changes are positive, and expression for an up-regulated gene is higher in the worse-survival subtype.

TCGA Disease	Gender	Group "case"	Group "control"	Gene regulation	#Genes at FDR < 0.01
KIRC	F	1-1	1-2	Up	257
KIRC	F	1-1	1-2	Down	2880
KIRC	M	2	3	Up	3504
KIRC	M	2	3	Down	1856
GBM	M	1-2	1-1	Up	1036
GBM	M	1-2	1-1	Down	80
GBM	M	1-2	2	Up	1112
GBM	M	1-2	2	Down	613
GBM	M	1-1	2	Up	594
GBM	M	1-1	2	Down	798
LAML	M/F	1	(2,3,4)	Up	1856
LAML	M/F	1	(2,3,4)	Down	1105
LAML	M/F	2	(1,3,4)	Up	2796
LAML	M/F	2	(1,3,4)	Down	1396
LAML	M/F	3	(1,2,4)	Up	1837
LAML	M/F	3	(1,2,4)	Down	4409
LAML	M/F	4	(2,3,4)	Up	1878
LAML	M/F	4	(2,3,4)	Down	760

In the next 3 sub-sections, we discuss each disease one at a time. Only statistically significant results are presented. There were many more clinical parameters available for AML (including mutations and blood counts) than for GBM or KIRC, and thus we are able to provide more detailed and significant results for AML clinical parameters. Section 4.4 explains the procedures and software that are used for functional analysis.

4.1 KIRC subtypes

Clear cell renal carcinoma is already a subtype of renal cell carcinoma, and has not been further subtyped, to our knowledge. The KIRC subtypes discovered by PINS include two exclusively female groups, one 98% male group, and a high-survivor group with 75% males. The significant Cox p-values for the survival rates of the PINS subgroups implies that they are actual disease subtypes related to gender and not due to a purely gender based signal. However, when performing functional analysis of subtypes using differentially expressed genes, it is important not to compare groups that are confounded by gender. Therefore, for KIRC and GBM, we are obliged to compare same-gender subsets of subtypes. Comparing the two female groups, we find that the poor survivors have higher grade tumors, overall down regulation of genes, and damage to the brush border membrane of the proximal tubules. Comparing the males in the two groups which are predominantly male, there is more gene up-regulation in the poor survival group, and these genes are correlated to metastasis and inflammation. However, the down-regulated genes overwhelmingly indicate that there is a mitochondrial malfunction in the poor survivors, potentially linked to the X-chromosome.

Table S16 shows the numbers and percentages of the 124 patients as they are partitioned into the survival clusters and clinical categories, nominally significant (uncorrected p-value less than 0.01) for at least one comparison. Apart from gender, these include: histologic grade, pathologic tumor stage, serum calcium level, hemoglobin level, platelet count, and age. Column (A) gives the actual number of samples in each category. Note that many measurements are not available for all patients, so

Table S16. Three columns with sets of tables (A, B, and C) show the distribution of 124 KIRC patients, in the four survival clusters, in each of the phenotypic categories. Note that there are differing numbers of missing values in each phenotypic category, so the sum of the number of patients will not be the same in every sub-table. The survival clusters are ordered from good to poor survival. A category is shown if there was at least one survival group (or pair of groups) significant for that category. Phenotypic categories are not shown if they are nearly the same as another that is shown (i.e., 'AJCC pathologic pt' and 'AJCC pathologic tumor stage'). The first column **(A)**, gives the actual number of patients in each survival group per phenotypic category. The second column **(B)**, gives the percentage of a each phenotypic subcategory in each of the survival groups (horizontal/column sum is 100). The third column **(C)**, gives the percentage of each of the survival groups in each of the phenotypic subcategories (vertical/row sum is 100). For example: survival group '1-1' has 25 females and no males; it is 100% female, and includes 46% of all of the females in the study. Percentages greater than 50% are highlighted.

		(A) Number in each group				**(B)** % phenotype in each group				**(C)** % group in each phenotype			
Survival		3	1-2	2	1-1								
Group		(12)	(25)	(62)	(25)	3	1-2	2	1-1	3	1-2	2	1-1
Gender	Female	3	25	1	25	6	46	2	46	25	**100**	2	**100**
	Male	9	0	61	0	13	0	**87**	0	**75**	0	**98**	0
tumor	G1	0	1	1	1	0	33	33	33	0	4	2	4
grade	G2	6	15	28	12	10	25	46	20	**75**	**60**	45	48
	G3	2	9	28	5	5	20	**64**	11	25	36	45	20
	G4	0	0	5	7	0	0	42	**58**	0	0	8	28
AJCC	I	7	17	31	5	12	28	**52**	8	**58**	**68**	50	20
pathologic	II	4	3	9	2	22	17	50	11	33	12	15	8
tumor stage	III	1	4	18	12	3	11	**51**	34	8	16	29	48
	IV	0	1	4	6	0	9	36	**55**	0	4	6	24
serum	Low	1	9	27	9	2	20	**59**	20	14	**60**	**68**	43
calcium	Normal	6	6	12	9	18	18	36	27	**86**	40	30	43
level	Elevated	0	0	1	3	0	0	25	**75**	0	0	3	14
hemoglobin	Low	3	8	34	18	5	13	**54**	29	38	38	**68**	**75**
level	Normal	5	13	16	6	13	33	40	15	**63**	**62**	32	25
platelet	Normal	7	17	42	16	9	21	**51**	20	**88**	**81**	**88**	**67**
count	Elevated	1	3	2	8	7	21	14	57	13	14	4	33

these sets of numbers should not be expected to produce the same sum. Column (B) gives the percentage of each phenotypic category in each of the survival groups, and column (C) gives the percentage of each of the survival groups in each of the phenotypic categories. Age is the only continuous variable and it is shown as box plots in Figure S18. We proceed with the analysis comparing the female groups to each other, and the males in groups '2' and '3' to each other separately. There is no significant confounding or interaction between the clinical variables.

4.1.1 Female subgroups

The low survival female group includes 86% of the Stage IV cases, while the high survival female group includes 77% of the Stage I cases, representing an FDR-corrected p-value of less than 5%. Other parameters that are significant at 5% include tumor grade (poor survivors had a higher incidence of grade 4 tumors, FDR-corrected p-value 2%), tumor status (poor survivors were 'with tumor', FDR-corrected p-value 2%), hemoglobin level (poor survivors had low levels, FDR-corrected p-value 2%), and metastasis (FDR-corrected p-value 2%).

There are 3137 genes are differentially expressed between long-term and short-term survivors (Table S15). Ninety-two percent (2880) of these were down-regulated in the poor survivors. Functional analysis using WebGestalt (Table S17) shows that the poorest surviving female group had damage to the brush border membrane of the kidney proximal tubules, acute phase reaction, decreased transmembrane ion transport, and elevated response to erythropoietin, compared to the females with better survival. The significant Cellular Component terms are related to plasma membrane, in particular 'brush border membrane'. The Biological Process and pathway terms concern known proximal tubule functions: metabolic processes and transmembrane and ionic transport. The Molecular Function term "glycosides activity" is also related, since alpha-glucosidase precursor has been localized to the proximal tubule brush border, where it is secreted into the urine.[23] Another process which is highly significant among the genes down-regulated in poor survivors is protein folding and the ability to dispose of incorrectly folded proteins.

Functional analysis of all 3137 genes using iPathwayGuide[24] also points to damaged proximal tubules in the nephrons of women with poor outcome. Sixteen pathways are significant with FDR-corrected p-values less than 0.01; the most significant signaling pathway is "Mineral Absorption" at $FDR = 0.002$. Several differentially expressed solute carriers on the Mineral Absorption Pathway are located in 'brush border membrane', shown in Figure S19a. In kidney, brush border membranes are found in the proximal tubules, which carry filtrate away from the glomerulus in the nephron, and support the secretion and absorption of charged molecules into and out of the filtrate. Other pathways with FDR-corrected p-values < 0.01 were all metabolic, except the PPAR signaling pathway, shown in Figure S19b (adjusted p-value 0.005). PPAR signaling is down regulated in poor surviving women, and may reflect the advanced age of this group.[25] The significant metabolic pathways include 'Fatty acid degradation', 'Butanoate metabolism', and 'Valine, leucine and isoleucine degradation', with FDR-corrected p-values of e^{-7}, e^{-6}, and e^{-8}, respectively. Notably, almost all DE genes on the mentioned pathways are down regulated in the poor survival group.

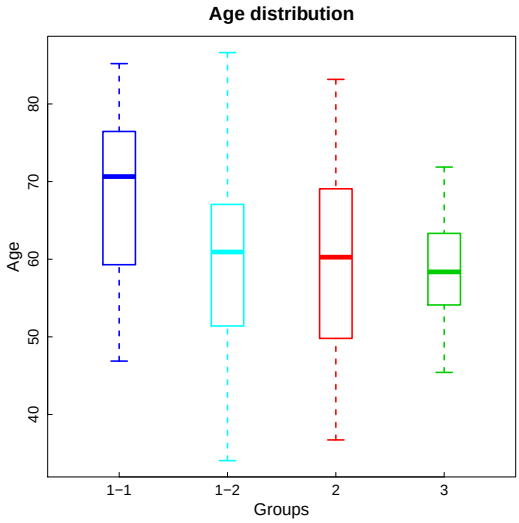
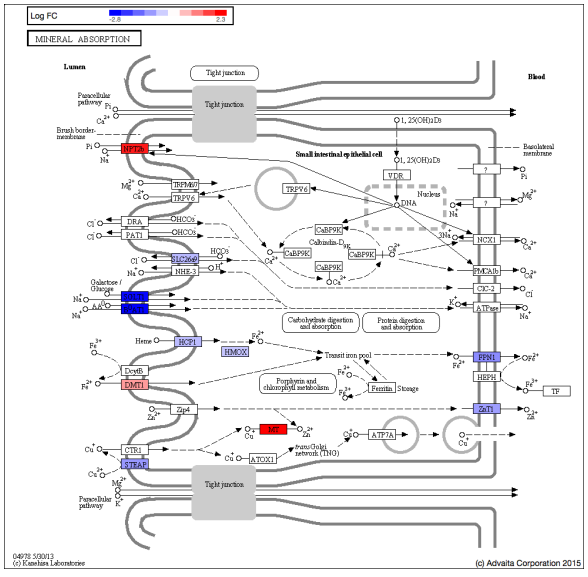


Figure S18. Age distribution of the discovered subtypes for kidney renal clear cell carcinoma (KIRC). The ages of subgroup '1-1' are significantly higher than any of other groups ($p < 0.01$). The range of ages for the other 3 groups are very similar even though their survival profiles are significantly different.

(a) Mineral Absorption (FDR p-value = $2e^{-3}$)



(b) PPAR Signaling (FDR p-value = $5e^{-3}$)

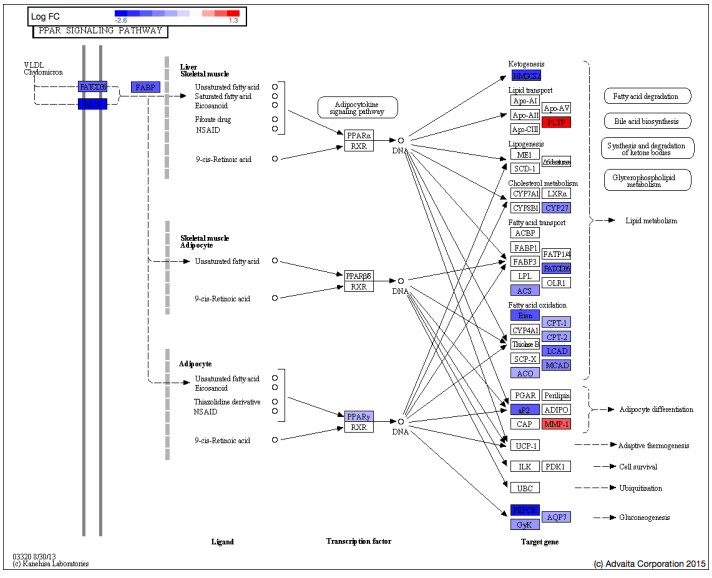


Figure S19. Significant KEGG signaling pathways, showing genes that are differentially expressed in KIRC females '1-1' vs. '1-2'. This figure is based on genes with maximum FDR adjusted p-value of 1%. Differentially expressed genes which are down-regulated are blue, and up-regulated are red. Panel (a) is a model of a brush border cell, with microvilli, from the intestine. Brush border membranes are also present in the proximal tubules of the kidneys. Panel (b) shows the PPAR signaling pathway. Peroxisome proliferator-activated receptors (PPARs) have been implicated in a variety of cancers, as well as in metabolic processes in the kidney. Figures obtained with iPathwayGuide.[24]

Females '1-1' vs '1-2'		
Database	257 genes up in '1-1' (down in '1-2')	2880 genes down in '1-1' (up in '1-2')
GO Biological Process	Acute phase response(e^{-2}). Response to erythropoietin(e^{-2}).	Establishment of protein localization (e^{-6}). Cellular metabolic process (e^{-10}). Protein ubiquitination (e^{-5}). Organic anion transport (e^{-3}).
GO Molecular Function	Enzyme inhibition activity(e^{-3}).	Small molecule binding (e^{-6}), Ion binding (e^{-6}), Cofactor binding (e^{-9}), Nucleotide binding (e^{-6}). Catalytic activity (e^{-15}), Ligase activity (e^{-7}). Protein transporter activity (e^{-3}). Phosphatase activity (e^{-3}). Glycosidase activity(e^{-2}).
GO Cellular Component	Extracellular matrix (e^{-3}).	Mitochondrial matrix (e^{-12}). Endosome (e^{-10}). Golgi apparatus (e^{-7}). Peroxisome (e^{-12}). Apical plasma membrane(e^{-7}), Brush border membrane(e^{-6}).
Pathway Commons		Branched-chain amino acid catabolism (e^{-5}). Metabolism of lipids and lipoproteins (e^{-5}). Citric acid cycle (TCA cycle) (e^{-4}). Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds(e^{-3}). SLC-mediated transmembrane transport(e^{-3}). plus 39 more signaling pathways at (e^{-4}) significance, all involving the same set of approximately 230 genes.
best PPI module	Hsapiens_Module_19(e^{-6}) BP:collagen catabolic process. MF:anchoring collagen. CC:collagen binding.	Hsapiens_Module_41(e^{-18}) BP:protein folding. MF: small conjugating protein ligase activity. CC: Cytoplasm.
Cytogenetic Band		14q (e^{-24}), 14q24 (e^{-6}). 4q (e^{-9}), 4q21 (e^{-4}).
Disease	Neoplastic Processes (e^{-4}). Collagen Diseases (e^{-4}). Cancer or viral infections (e^{-4}). Acute-Phase Reaction (e^{-4}). Neoplasms(e^{-3}).	Metabolism, Inborn Errors (e^{-12}). Zellweger Syndrome (e^{-7}).
Drug	Collagenase (e^{-3}).	cyanocobalamin (e^{-4}).
Phenotype		Abnormality of the urinary system physiology (e^{-6}). Abnormality of amino acid metabolism (e^{-6}). Abnormality of blood glucose concentration (e^{-5}). Neurophysiological abnormality (e^{-6}). Abnormality of movement (e^{-5}). Decreased liver function (e^{-6}). Muscular hypotonia (e^{-5}).

Table S17. WebGestalt enrichment summary for KIRC all-female survival groups '1-1' vs. '1-2'. The input gene set is defined by an FDR-corrected p-value of 0.01. The poorest surviving all-female group, '1-1', is down-regulated for genes in the brush border membrane of the kidney proximal tubules, and transmembrane ion transport, but up-regulated for acute phase reaction and elevated response to erythropoietin, compared to the females with better survival. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Males '2' vs '3'		
Database	3504 genes up in '2' (down in '3')	1856 genes down in '2' (up in '3')
GO Biological Process	Cell migration(e^{-22}). Immune response (e^{-49}). Lymphocyte activation (e^{-26}). Leukocyte activation(e^{-26}). Signal transduction(e^{-23}). Vasculature development(e^{-21}).	Hydrogen transport(e^{-10}). Cellular respiration(e^{-52}). ATP biosyntetic process(e^{-7}). Mitochondrial ATP synthesis coupled electron transport(e^{-21}). Small molecule metabolic process(e^{-20}). Mitochondrial transport(e^{-8}).
GO Molecular Fucntion	Binding: kinase(e^{-7}), actin(e^{-8}), ephrin receptor(e^{-6}), lipid(e^{-10}), anion(e^{-8}), ribonucleotide(e^{-5}). Kinase activity(e^{-5}). Small GTPase regulator activity(e^{-5}). Cytokine receptor activity(e^{-6}).	Catalytic activity(e^{-12}). Cofactor binding(e^{-6}). Oxyreductase activity(e^{-18}). Hydrogen ion transmembrane transporter activity(e^{-20}).
GO Cellular Component	Plasma membrane part(e^{-17}). MHC class II protein complex(e^{-6}). Lamellipodium(e^{-7}). Membrane raft(e^{-9}). Adherens junction(e^{-9}). Cytosol(e^{-16}). Endocytic vescicle(e^{-6}). Focal adhesion(e^{-9}).	Mitochondrial inner membrane(e^{-43}). NADH dehydrogenase complex(e^{-22}). Mitochondrial respiratory chain(e^{-30}). Proton-transporting two-sector ATPase complex(e^{-16}).
Pathway Commons	Integrin family cell surface interactions(e^{-46}). VEGF and VEGFR signaling network(e^{-46}). PAR1-mediated thrombin signaling events(e^{-46}). Plasma membrane estrogen receptor signaling(e^{-46}). Sphingosine 1-phosphate (S1P) pathway(e^{-46}). TRAIL signaling pathway(e^{-46}). IFN-gamma pathway(e^{-46}).	The citric acid (TCA) cycle and respiratory electron transport(e^{-58}). Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins(e^{-45}). Pyruvate metabolism and Citric Acid (TCA) cycle(e^{-13}). Citric acid cycle (TCA cycle)(e^{-11}).
microRNAs that target geneset	miR-29a/b/c(e^{-12}). miR-30a-5p/b/c/d/e-5p(e^{-11}). miR-200b/c, miR-429 (e^{-10}). miR-506 (e^{-10}). miR-17-5p, miR-20a/b, miR-106a, miR-106b, miR-519d (e^{-8}).	
Best PPI module	Hsapiens_Module_111 (e^{-12}) BP:actin polymerization or depolymerization MF:non-membrane spanning protein tyrosine kinase activity CC:lamellipodium	Hsapiens_Module_49 (e^{-47}) BP:mitochondrial ATP synthesis coupled proton transport MF:proton-transporting ATP synthase complex CC:cytochrome-c oxidase activity
Cytogenetic Band	1q(e^{-14}). 1p(e^{-7}) 2p(e^{-5}). 5q(e^{-5}).	3p(e^{-27}), 3p21(e^{-12}), 3p25(e^{-6}).
Disease	Immune system diseases (e^{-40}). Virus diseases(e^{-31}). Necrosis(e^{-26}). Infection(e^{-30}). Adhesion(e^{-29}). Leukemia(e^{-26}). Neovascularization, pathologic(e^{-24}). Inflammation(e^{-22}).	Mitochondrial diseases(e^{-25}). Acidosis(e^{-14}). Mitochondrial encephalomyopathies(e^{-7}).
Drug	Immune globin(e^{-26}). Heparin(e^{-7}). Collagenase(e^{-7}). Glutatione(e^{-6}). Fluorouracil(e^{-6})	Nadh(e^{-11}). Lipoic acid(e^{-4}) Iron(e^{-3}). Gabapentin(e^{-3}).
Phenotype	Abnormality of the immune system(e^{-6}). Localized skin lesion(e^{-4}). Abnormality of blood and blood-forming tissues(e^{-7}). Abnormality of the gingiva(e^{-3}). Cellulitis(e^{-3}).	Mitochondrial inheritance(e^{-9}). X-linked dominant inheritance(e^{-3}). Sensorineural hearing impairment(e^{-5}). Acute encephalopathy(e^{-7}). Coma(e^{-5}). Increased CSF lactate(e^{-8}). Acidosis(e^{-15}). Abnormality of the mitochondrial metabolism(e^{-9}). Hepatic necrosis(e^{-5}). Cardiomyopathy(e^{-5}). Exercise intolerance(e^{-5}). Macrocephaly(e^{-5}). Vomiting(e^{-5}).

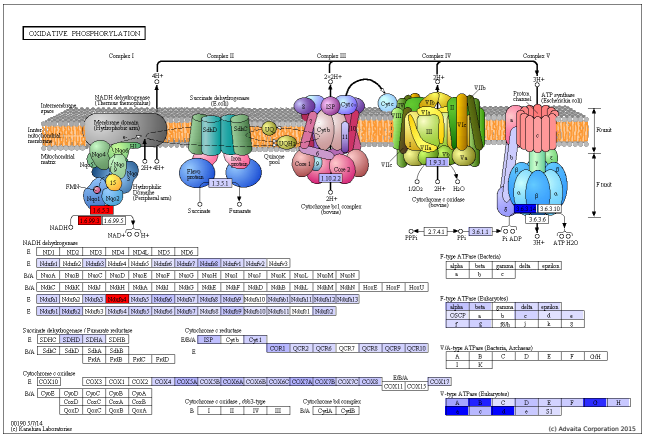
Table S18. WebGestalt enrichment summary for male members of the KIRC, groups '2' (intermediate survivors) vs. '3' (all survive). The input gene set is defined by an FDR-corrected p-value of 0.01. Up-regulated genes in the poor survivors are associated with activation of the immune system, vascularization, and hematological disease. Down-regulated genes in the poor survivors are overwhelmingly associated with mitochondrial problems. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

4.1.2 Male subgroups

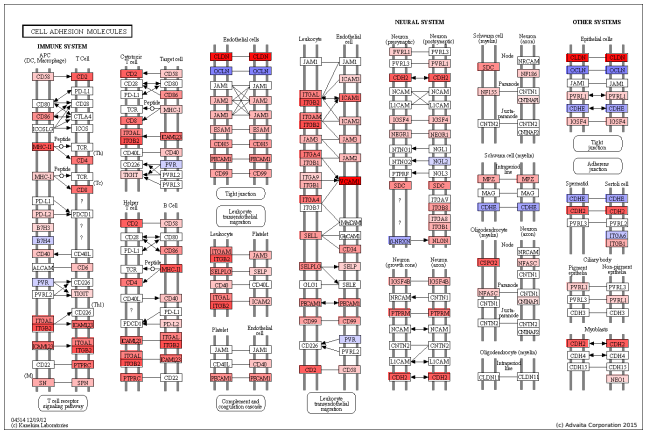
There are many differentially expressed genes between the high mortality males (61 members in group '2'), and the males that survive renal cell carcinoma (9 members in group '3'). The majority of these genes are more highly expressed in the poor-survival group. WebGestalt gene set enrichment is shown in Table S18. Not surprisingly, up-regulated genes in the poor survivors are associated with immune response and known metastatic processes, including cell migration and vascularization. Significant up-regulated pathways support these observations: Cell Adhesion, Leukocyte Transendothelial Migration, and Focal Adhesion are up-regulated (Figures S20b through S20d).

Among terms that are enriched with genes down-regulated in poor surviving males, mitochondrial terms are found across the board. The only significant pathway, Oxidative Phosphorylation, shown in Figure S20a, is down-regulated, supporting these observations. In addition, there is an association with the phenotype "X-linked dominant inheritance", which is interesting given that all samples in this comparison are male. A GO enrichment using the subset of 59 down-regulated genes that are on the X-chromosome, shows that 10 are associated with "mitochondrial part" ($FDR = e^{-3}$), and 8 with "the mitochondrial inner membrane" ($FDR = 7e^{-4}$). In contrast, the subset of 92 up-regulated genes that are on the X-chromosome is not significant for any GO terms. The 59-gene subset is not enriched for microRNAs or any specific transcription factors, but NetGestalt[26] analysis identifies 5 hub genes at FDR < 0.001: COX7B, CA5B, NDUFB11, IDH3G, and NDUFA1. The related Biological Process given by NetGestalt for these 5 genes is "mitochondrial ATP synthesis coupled proton transport". The full set of genes down-regulated in poor survivors is also highly enriched on Chromosome 3p (129 genes, $FDR = e^{-27}$). Cytoband 3p21 ($FDR = e^{-12}$) is known for an abundance of tumor suppressor genes, and is involved in the development of epithelial cancers, including renal cell carcinoma.[27] Cytoband 3p25 ($FDR = e^{-6}$) is noted for deletions and loss of heterozygosity in several tumors, including renal.[27]

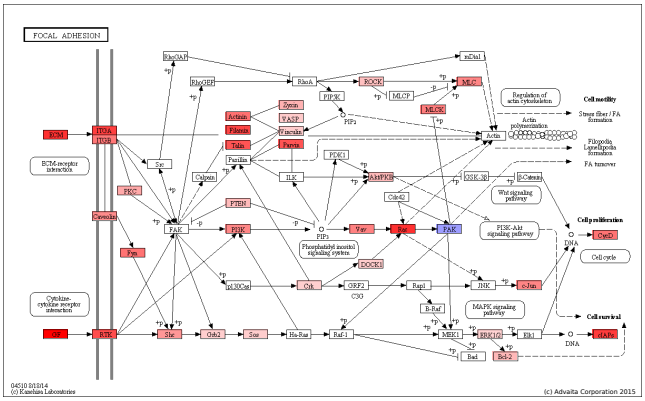
(a) Oxidative Phosphorylation (FDR p-value = $3e^{-19}$)



(b) Cell Adhesion (FDR p-value = $2e^{-11}$)



(c) Focal Adhesion (FDR p-value = $1e^{-6}$)



(d) Leukocyte Transendothelial Migration (FDR p-value = $4e^{-7}$)

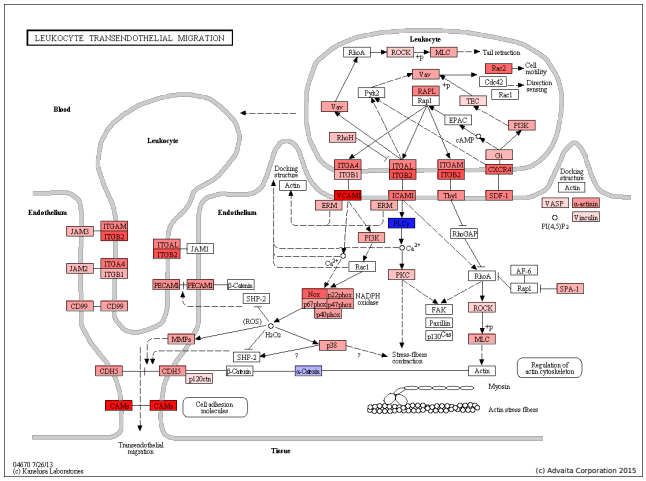


Figure S20. Significant KEGG pathways, showing genes that are differentially expressed in KIRC males between short term surviving males (group '2') and 100% survival males (group '3'). The pathway "Oxidative Phosphorylation" is down-regulated in poor surviving males, implying a mitochondrial connection. The pathways "Cell adhesion", "Focal Adhesion", and "Leukocyte Transendothenial Migration" are up-regulated, reflecting increased metastasis and inflammation.

4.2 GBM subtypes

There are three PINS subtypes for GBM, two of which have approximately equally poor survival, while the third has a much better survival. We find that the GBM subgroups found by PINS are roughly in agreement with already published proposed subtypes; we compare our results with these, and point out novelties in PINS subtypes. The primary original finding is that glycine and serine metabolism is not a hallmark of all GBM, but is associated with a particular poor-survival subtype. Glycine and serine metabolism support DNA and histone methylation, and may explain the strong influence of the methylation data in PINS clustering.

Previous attempts to subtype GBM have given generally consistent results. Using consensus clustering on TCGA GBM data, Verhaak et. al.[28] found four subtypes which they refer to as proneural, neural, classical, and mesenchymal. Using Similarity Network fusion (SNF), Wang et. al.[16] reported 3 clusters, including the IDH subtype (characteristic of proneural), and a hypermethylated subtype. Phillips et. al.[29] report 3 clusters of high grade glioma based on microarray expression data: proneural, proliferative, and mesenchymal. Chinnaiyan et. al.[30] investigated grade 2, 3, and 4 glioma and identified three metabolic signatures: energetic, anabolic, and phospholipid catabolism, the last of which included the majority of high grade gliomas (GBM). Within the high grade tumors, they found a particularly aggressive subgroup with accumulation of phosphoenolpyruvate and decreased pyruvate kinase activity, which they correlated to the mesenchymal subtype. They also identified accumulation of glycine and serine in grade 4 tumors. Synthesis of serine and glycine represents one way to divert

Table S19. Distribution of the 273 GBM patients among the clinical variables that are nominally enriched with p-value < 0.01 for at least one of the comparisons between the three survival clusters. Groups are shown in the order of low to high survival. Note that there are missing values in the 'Had New Tumor Event' category, so the sum of the number of patients is less than 273. The first column **(A)**, gives the actual number of patients in each survival group per phenotypic category. The second column **(B)**, gives the percentage of each phenotypic subcategory in each of the survival groups (horizontal/column sum is 100). The third column **(C)**, gives the percentage of each of the survival groups in each of the phenotypic subcategories (vertical/row sum is 100). For example: survival group '1-1' has 99 females and 82 males; it is 55% female, and includes 91% of all of the females in the study. Percentages greater than 50% are bold and underlined.

		(A) Number in each group			**(B)** % phenotype in each group			**(C)** % group in each phenotype		
Survival group (total number)		1-1 (181)	1-2 (68)	2 (24)	1-1	1-2	2	1-1	1-2	2
Gender	Female	99	2	8	**91**	2	7	**55**	3	33
	Male	82	66	16	50	40	10	45	**97**	**67**
Age	< 50	48	10	18	**63**	13	24	27	15	**75**
	50-60	40	22	3	**62**	34	5	22	32	12
	60-70	53	20	2	**71**	27	3	29	29	8
	>70	40	16	1	**70**	28	2	22	24	4
Had New Tumor Event	no	45	20	4	**65**	29	6	38	**57**	27
	yes	75	15	11	**74**	15	11	**62**	43	**73**

carbon from glycolysis through the pentose phosphate pathway, metabolically reprogramming the tumor cells. PINS subtypes can be correlated to some of these subtypes described by other authors.

Each of the 14 different glioblastoma clinical parameters is analyzed for enrichment in each of the three survival clusters, using the hypergeometric test. Every combination of the three survival cluster sets is compared against every other, and all parameters significant at nominal p-value < 0.01 are summarized in Table S19, which shows the numbers and percentages of the 273 patients, distributed into each of the survival clusters and clinical categories. Only age, gender, and "Had new tumor event" qualified. Column (A) gives the actual number of samples in each category. Column (B) gives the percentage of each phenotypic category in each of the survival groups, and column (C) gives the percentage of each of the survival groups in each of the phenotypic categories. Age distributions in subtypes are also portrayed as continuous variables in Figure S21.

Validated mutation data[28] (from https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/) confirms that IDH1 mutations are specific to group '2' (Fisher's exact test $p = 2 \times 10^{-8}$). Among the 45 patients that have the IDH1 mutation information, all 7 mutated samples belong to group '2' and all 38 wild-type samples belong to other groups.

To summarize, GBM clusters found by PINS can be correlated to subtypes described by other authors. Our data shows that GBM clusters are highly influenced by methylation profiles. We see that although group '2' patients are younger, they tend toward recurrence events. Since the best surviving group ('2') consists of young patients with a tendency for recurrent tumor events, and with disease tissue rich in IDH1 mutations, it is similar to the proneural subtype,[28] and may respond to temozolomide,[16,29] a drug that interferes with DNA replication. Cluster '1-1' has a wide range of ages, and also tends to have recurrence events. Groups '1-2' and '2' are majority male, so gender bias may arise if either of these is compared to group '1-1'. Therefore, all differentially expressed genes input into WebGestalt and iPathway guide are determined for male patients only.

Enrichment analysis of both clusters '1-1' and '1-2', compared to the good survivors ('2'), shows high invasiveness and vascularization, as should be expected. Like the proliferative and mesenchymal subgroups identified by,[29] these clusters have close, parallel survival curves. Pathway analysis contrasting these two reveals that subtype '1-1' is more collagenous than '1-2', with more extracellular matrix and calcium ion binding and thus may be more mesenchymal than proliferative. Collagen and extracellular matrix terms are associated with invasiveness in GBM.[31,32] GO analysis suggests that '1-2' is a subtype with strong regulation of glial and astrocyte differentiation, and thus may be more proliferative than mesenchymal. In addition, '1-2' is significantly enriched in glycine and serine metabolism compared to '1-1', a phenomenon reported in aggressive glioma.[30] Serine and glycine metabolism are implicated in oncogenesis, and notably, provide methyl groups for DNA and histone methylation,[30,33] a possible explanation for the dominant influence of methylation profile on our subtyping results.

4.2.1 Short term versus medium term survival

Over 90% of the genes that are differentially expressed between the short term surviving males (group '1-2') and the medium term surviving males (group '1-1'), are up-regulated in the poor survivors. Figure S22 shows the only significant KEGG pathway, "Glycine, serine, and threonine metabolism" (FDR-corrected p-value $< 2e^{-4}$), which is up-regulated in the poor survival group, '1-2'. Chinnaiyan et. al.[30] identified an abundance of serine and glycine in glioblastoma, a sign of the metabolic reprogramming that is a hallmark in many cancers. This "glycolytic shunt" is characterized by overproduction of the gene PHGDH,[34] and indeed, PHGDH overexpression is observed in group '1-2' compared to '1-1' (FDR-corrected p-value < 0.001).

Table S20 summarizes WebGestalt results separately for the relatively up-regulated and down-regulated genes. Genes more highly expressed in '1-2' than in '1-1' enrich Biological Processes and diseases associated with the differentiation of astrocytes, gliogenesis, regulation of gene expression, and astrocytoma. Many microRNAs are more highly expressed in '1-2'. The most significant up-regulated microRNAs are in the family including miR-200B and C (FDR-corrected p-value $= e^{-10}$). MiR-200C is known to associate with high grade gliomas, and the miR-200 family is implicated in GBM for the epithelial-mesenchymal

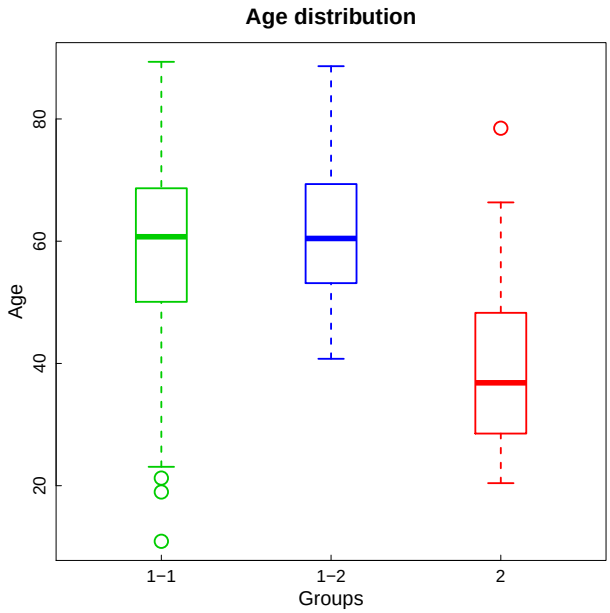


Figure S21. Age distribution of the discovered subtypes for glioblastoma multiforme (GBM). The ages of subgroup 2 are significantly lower than any of other groups ($p < 10^{-6}$).

transition.[35]

Genes from the cytogenetic bands 14q and 19q are also enriched in'1-2', or lacking in group '1-1' as the case may be, since the differential expression is relative. Loss of heterozygosity (LOH) in the cytogenetic region 14q, at several sites, has been observed to correlate with glioblastoma development. The sub-region 14q23-31 is suspected to be rich in tumor suppressor genes,[36,37] and is represented in our results by the enriched cytoband 14q24 (FDR-corrected p-value $= e^{-5}$). Thus, we may assume that it is down-regulated in group '1-1' as opposed to '1-2' as well as normal tissue. The cytoband sub-region 19q13 (FDR-corrected p-value $= e^{-6}$) has been observed to be amplified in some glioblastomas at 19q13.2,[38] but deletions within the region 19q13.33-q13.41 are also reported in astrocytic tumors.[39]

4.2.2 Short term versus long term survival

The only pathways significant at $FDR < 0.01$ for short term surviving males (group '1-2') versus long term surviving males (group '2'), are "Focal Adhesion" and "ECM-receptor interaction", shown in Figures S23a and S23b. Both pathways are activated in poor survivors, as indicated by the red (up-regulated) genes in Figures S23b and S23a. As the figures show, "ECM-receptor interaction" is upstream of "Focal Adhesion", at the level of the transmembrane integrins ITGA and ITGB. The importance of the extracellular matrix (ECM) and focal adhesion in GBM are discussed by Bellail et. al.[40]

Table S21 summarizes the gene set enrichment results from WebGestalt. Many terms up-regulated in the poor survivors are related to invasiveness. Poor survivors are up-regulated in genes on Chromosome 7, but down-regulated in genes on Chromosome 10, therefore, we can identify group '1-2' as a previously described poor survival subtype with Chromosome 7 polysomy, together with loss of, or monosomy of Chromosome 10.[41,42] Table S21 also alludes to an issue regarding unfolded proteins. Translation is down in group '1-2', but response to stress, endoplasmic reticulum, golgi apparatus, and lytic vacuole terms are enriched.

4.2.3 Medium term versus long term survival

The gene sets for medium term surviving males (group '1-1) versus long term surviving males (group '2') are essentially the same as those for short term ('1-2') versus long term ('2'). Not surprisingly, the only resulting significant pathways are the same, "Focal Adhesion" and "ECM-receptor interaction", with the same fold-change directions on each. Table S22 gives the WebGestalt enriched terms for intermediate versus longer term survivors. We see a similar pattern to that in Section 4.2.2 - poor versus good survivors. Very slight differences can be observed, but they are not quantifiable. In order to distinguish the two comparisons, we take the genes that are up-regulated in one but not the other, and down-regulated in one but not the other. We perform WebGestalt analyses for the two gene sets. The numbers in brackets below are minus log FDR-corrected p-values.

The size of the intersection of the up-regulated sets for each comparison is 516 representing 89% of the genes up-regulated in '1-1' versus '2'. The size of the intersection of the down-regulated sets for each comparison is 435 representing 71% of the genes down-regulated in '1-2' versus '2'. The 596 up-regulated genes in the '1-2 vs '2' comparison, but not in the '1-1' vs

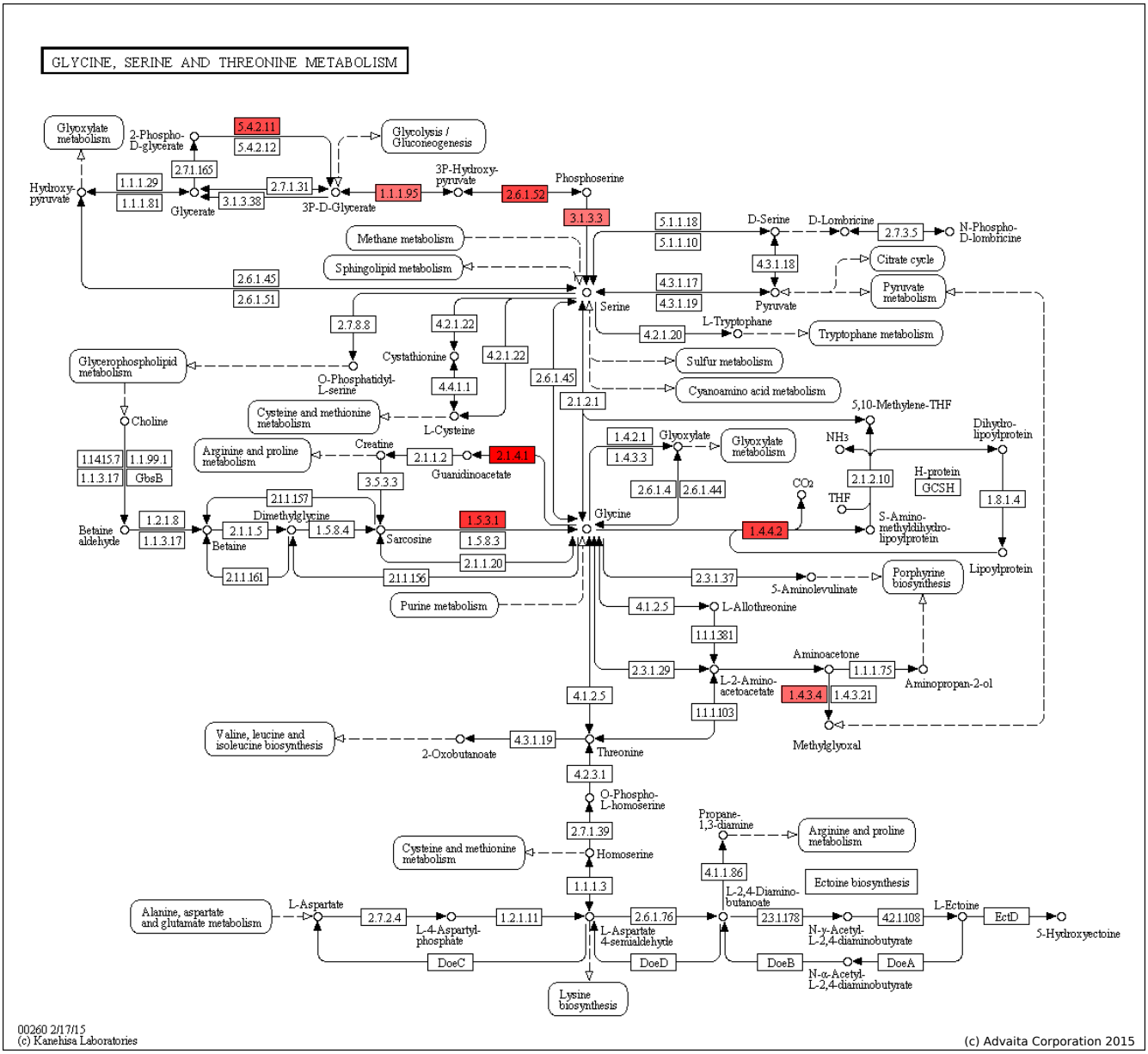
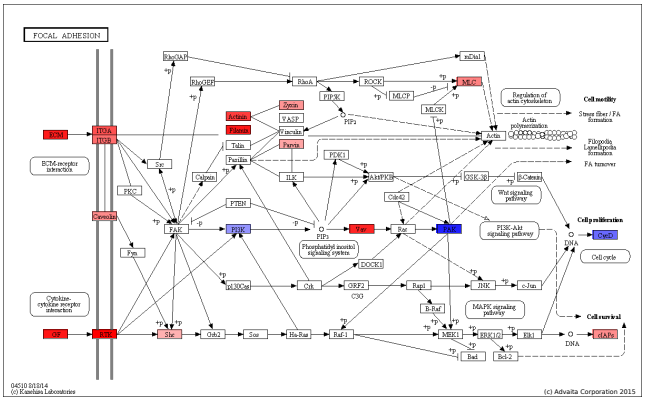


Figure S22. KEGG "Glycine, serine, and threonine metabolism", FDR-corrected p-value $< 2e^{-4}$, the only pathway significant using 1% FDR-corrected p-value cutoff. Genes that are differentially expressed in GBM between short term surviving males (group '1-2') and slightly longer term surviving males (group '1-1'). Glycine and serine metabolism is up-regulated in group '1-2'.

Table S20. GBM, males only. Enrichment summary, based on genes selected at FDR-corrected p-value < 0.01. Poor survivors ('1-2') vs medium survivors ('1-1') are compared. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Males '1-2' vs '1-1'		
Database	1036 genes up in '1-2' (down in '1-1')	80 genes down in '1-2' (up in '1-1')
GO Biological Process	Astrocyte differentiation(e^{-5}), gliogenesis(e^{-4}). Negative regulation of neuron differentiation(e^{-4}). Regulation of neural precursor cell proliferation(e^{-4}). Regulation of gene expression(e^{-4}). Cellular macromolecule biosynthetic process(e^{-4}). Nucleic acid metabolic process(e^{-4}).	
GO Molecular Function	Protein binding(e^{-7}). DNA binding(e^{-7}).	NAD(P)H oxidase activity(e^{-3}). Interleukin-1 receptor activity(e^{-4}). Melanocyte-stimulating hormone receptor activity(e^{-3}).
GO Cellular Component	Nucleoplasm(e^{-6}). Dendrite(e^{-3}). Mitochondrion(e^{-4}).	Plasma membrane(e^{-3}). Golgi lumen(e^{-3}). Endoplasmic reticulum lumen(e^{-3}). Extracellular space(e^{-3}). Collagen(e^{-3}).
Pathway Commons	Syndecan-3-mediated signaling events(e^{-3}). Global Genomic NER (GG-NER)(e^{-3}). DNA Repair(e^{-3}). Regulation of Telomerase(e^{-3}). Nucleotide Excision Repair(e^{-3}).	Peptide GPCRs(e^{-3}). GPCRs, Class A Rhodopsin-like(e^{-3}).
microRNAs that target geneset	miR-200B, miR-200C, miR-429 (e^{-10}). miR-524(e^{-9}). miR-124A(e^{-6}). miR-527(e^{-6}). miR-374(e^{-6}). miR-203(e^{-5}). miR-369-3p(e^{-5}).	
Best PPI module	Hsapiens_Module_929(e^{-3}) BP: protein deneddylation. MF: Ran guanyl-nucleotide exchange factor activity. CC: signalosome.	Hsapiens_Module_19(e^{-4}) BP: collagen catabolic process. MF: collagen binding. CC: collagen binding.
Cytogenetic band	14q(e^{-17}), 14q11(e^{-5}), 14q24(e^{-5}). 19q(e^{-11}), 19q13(e^{-6}). 19p(e^{-7}), 19p13(e^{-6}).	
Disease	Central Nervous System Diseases(e^{-5}). Alagille Syndrome(e^{-5}). Mental Disorders(e^{-5}). Astrocytoma(e^{-5}).	Bronchial Diseases(e^{-3}). Asthma(e^{-3}). Chorioamnionitis(e^{-3}). Inflammatory Bowel Diseases(e^{-3}).
Drug	leucovorin(e^{-3}).	finasteride(e^{-3}).
Phenotype	Aplasia/Hypoplasia of the cerebrum(e^{-5}). Microcephaly(e^{-4}).	

(a) Focal Adhesion (FDR p-value = $2e^{-4}$).



(b) ECM-receptor interaction (FDR p-value = $2e^{-4}$).

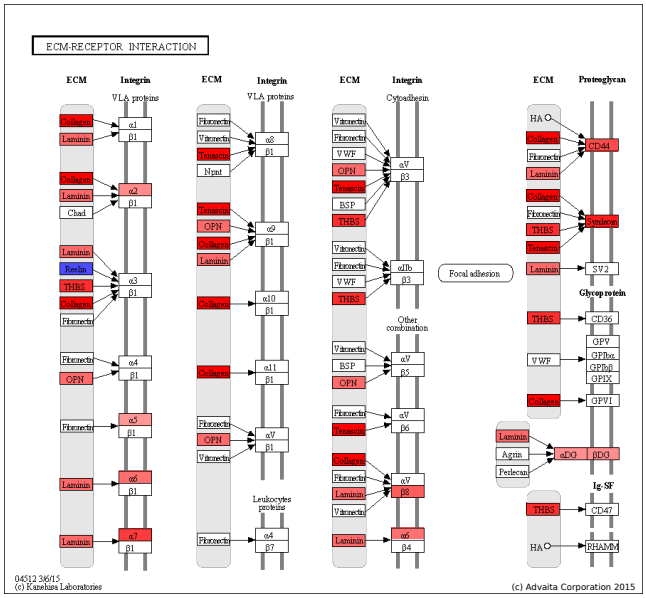


Figure S23. Significant KEGG signaling pathways, showing genes that are differentially expressed in GBM males '1-2' vs. '2'.

Table S21. GBM WebGestalt results, based on genes selected at FDR < 0.01. Poor survivors ('1-2') vs good survivors ('2') are compared, only for male patients. There were no significant phenotype terms. The many significant Biological Processes for the 1112 up-regulated genes were dominated by 'extracellular matrix' GO terms, and diseases including Glioblastoma. The significant GO terms for the 613 down-regulated terms infer a decrease in translation.

Males '1-2' vs '2'		
	1112 genes	613 genes
Database	up in '1-2' (down in '2')	down in '1-2' (up in '2')
GO Biological Process	Extracellular matrix organization(e^{-11}). Intracellular protein kinase cascade(e^{-8}). Response to external stimulus(e^{-7}). Response to stress(e^{-12}). Oxoacid metabolic proecss(e^{-7}). Vascular development(e^{-11}). Cell morphogenesis(e^{-6}). Antigen processing and presentation of peptide via MHC class I (e^{-6}).	Negative regulation of metabolic process(e^{-5}). Negative regulation of transcription(e^{-6}). Chromatin organization(e^{-5}). Translational initiation (e^{-5}). Generation of neurons(e^{-5}). Cell surface receptor signaling(e^{-5}). Regionalization(e^{-5}). Macromolecular complex disassembly(e^{-4}). mRNA catabolic process(e^{-5}).
GO Molecular Function	Integrin binding(e^{-4}). Extracellular matrix binding (e^{-3}). Growth factor binding (e^{-4}). Hereupon binding (e^{-3}). Catalytic activity(e^{-3}).	Nucleic acid binding(e^{-3}). Transcription corepressor activity(e^{-4}).
GO Cellular Component	Basement membrane(e^{-11}). Plasma membrane(e^{-6}). Sarcolemma(e^{-5}). Cytplasmic vescicle(e^{-14}). Cytoplasmic membrane-bounded vescicle(e^{-12}). Endoplasmic reticulum (e^{-11}). Golgi apparatus (e^{-6}). Lytic vacuole (e^{-8}).	Nucleoplasm part(e^{-4}). Cytosolic ribosome(e^{-4}).
Pathway Commons	Integrin family cell surface interactions(e^{-17}). Proteoglycan syndecan-mediated signaling events(e^{-15}).	Formation of a pool of free 40S subunits(e^{-8}). Translation(e^{-8}). GTP hydrolysis and join gin of the 60S ribosomal subunit(e^{-8}).
microRNAs that target geneset		miR-15a/b, miR195, miR-424, miR-497(e^{-10}). miR-9(e^{-6}). miR-145(e^{-6}).
Best PPI module	Hsapiens_Module_19(e^{-8}) BP: collagen catabolic process. MF: collagen binding. CC: anchoring collagen.	Hsapiens_Module_560(e^{-5}). BP: translational termination. MF:structural constituent of ribosome. CC:Component:cytosolic large ribosomal subunit.
Cytogenetic band	7p(e^{-7}), 7p15(e^{-3}). 7q(e^{-4}), 7q22(e^{-3}). 11p15 (e^{-3}).	10q(e^{-24}), 10q22(e^{-9}), 10q24(e^{-5}), 10q11(e^{-3}). 10p(e^{-19}), 10p15(e^{-6}), 10p14(e^{-5}), 10p11(e^{-4}).
Disease	Neoplasm invasiveness(e^{-17}). Adhesion(e^{-14}). Cancer or viral infections(e^{-12}). Neoplasm metastasis(e^{-12}). Neovascularization, pathologic(e^{-11}). Neoplastic Processes(e^{-10}). Astrocytoma(e^{-10}). Collagen diseases(e^{-10}). Fibrosis(e^{-10}). Neoplasms(e^{-10}). Glioblastoma(e^{-10}). Glioma(e^{-10}).	
Drug	Herapin(e^{-13}). Urokinase(e^{-9}). Netilmicin(e^{-9}). Cefotaxime(e^{-9}). Cefacetrile(e^{-9}).	

'2' comparison are most highly enriched for GO cellular components: Endoplasmic reticulum [6], Golgi apparatus [5], and Vescicle[6]. Other enriched terms include cytoband 7p [4], and metabolic disease[4].

The 363 down-regulated genes in the '1-1' vs '2' comparison, but not the '1-2 vs '2' comparison are most highly enriched for GO Biological Process Regulation of transcription, DNA-dependent[13], and Cellular Component Nucleus[17]. Enriched diseases included Autism[4] and ADHD[4], enriched cytogenetic bands are 14q[3] and 14q11[3], enriched microRNAs include miR-493[9], miR-249[8] and miR-200b/c[8]. The significant PPI modules[5] have the following related functions: Notch signaling, Nucleosome assembly, and Methylated histone residue binding.

4.3 AML subtypes

In the case of AML, PINS discovered four subtypes. In this case, there is no potential confounding with gender, therefore, each subgroup is compared to the set of the other three. The best surviving group discovered by PINS can be classified as Acute Promyelocytic Leukemia (APL), based on clinical data alone, and supported by molecular data. The worst surviving group has the largest variety of mutations, and includes lymphocytic signals. Accordingly, this group appears to include what is generally referred to as "mixed phenotype acute leukemia" (MPAL), subtypes known to be associated with high mortality. The two intermediate groups are quite distinct. One of them is dominated by myelocyte and monocyte lineages. The other is not associated with any specific hematopoietic lineages. However, the set of genes that are differentially expressed in this subtype (with respect to the union of the others) is strongly enriched in terms for mitochondrial translation. The antibiotic tigecycline has been reported to stop proliferation of AML cells by blocking mitochondrial translation. These results provide some evidence suggesting that this drug may be useful only for a subgroup of AML patients.

Unlike GBM and KIRC, many clinical variables are provided for LAML by TCGA. Each of the 64 different clinical parameters was analyzed for enrichment in each of the four survival clusters, using the hypergeometric test. Every combination of the four survival cluster sets was compared against every other. Although we considered 64 clinical parameters for assessment, and made all possible comparisons, there were still many very significant categories after correction for multiple hypotheses.

Strict definitions of clinical subtypes for Acute Myeloid Leukemia remain controversial.[43,44] However, there are two traditional classifications which are provided by TCGA, and they are significant here: Cancer and Leukemia Group B (CALBG) risk, and the French-American-British classification (FAB). FAB classes leukemia based on the cell type of origin and the maturity of the diseased cells.[45] The main criticism for FAB is that while it accounts for the morphological heterogeneity of AML, it does not sufficiently account for the clinical and genetic diversity of AML,[43] and in particular does not consider

Table S22. GBM. Gene Ontology, disease, and phenotype terms, based on genes selected at FDR < 0.01. Intermediate survivors ('1-1') and good survivors ('2') are compared, only for male patients. The many significant Biological Processes for the 594 up-regulated genes were dominated by 'extracellular matrix', vascularization, and 'binding' GO terms, and diseases terms including 'neoplasm invasiveness'. The significant GO terms for the 798 down-regulated terms point to transcription regulation.

Males '1-1' vs '2'		
Database	594 genes up in '1-1' (down in '2')	798 genes down in '1-1' (up in '2')
GO Biological Process	Vasculature development(e^{-18}). Extracellular matrix organization(e^{-14}). Angiogenesis(e^{-13}). Cell adhesion(e^{-9}). Amino glycerin catabolic process(e^{-6}). Locomotion(e^{-7}). Cell migration(e^{-6}). Regulation of phosphorylation(e^{-6}). Response to lipid(e^{-7}). Response to stress(e^{-13}). Oxireductase activity(e^{-3}). Phospholipase inhibitor activity(e^{-3}).	RNA metabolic process(e^{-15}), regulation of transcription(e^{-13}). Chromatin organization(e^{-13}).
GO Molecular Function	Herapin binding(e^{-4}). L-ascorbic acid binding(e^{-3}). Calcium ion binding(e^{-4}). Protease binding(e^{-3}). Integrin binding(e^{-8}). Platelet derived Growth factor binding(e^{-6}).	Nucleic acid binding(e^{-11}), protein binding(e^{-10}). Deacetylase activity(e^{-3}). Histone-lysine N-methyltransferase activity(e^{-3}). Sequence-specific DNA binding transcription factor activity(e^{-3}). Transcription factor binding transcription factor activity(e^{-7}).
GO Cellular Component	Endoplasmic reticulum lumen(e^{-10}). lysosomal lumen(e^{-10}). Extracellular matrix(e^{-13}). Collagen(e^{-6}). Basement membrane(e^{-10}). Melanosome(e^{-5}). Platelet alpha granule lumen(e^{-5}).	Nuclear lumen(e^{-15}). Neuron projection(e^{-4}), dendrite(e^{-5}). Cytosol(e^{-4}), cytosolic large ribosomal subunit(e^{-4}).
Pathway Commons	Integrin family cell surface interactions(e^{-23}). Beta1 integrin cell surface interactions(e^{-21}). Proteoglycan syndecan-mediated signaling events(e^{-16})	PDGFR-beta signaling pathway(e^{-5}). Formation of a pool of free 40S subunits(e^{-5}). Arf6 downstream pathway(e^{-5}). Sphingosine 1-phosphate (S1P) pathway(e^{-5}). S1P1 pathway(e^{-5}). Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling(e^{-5})
microRNAs that target geneset	miR17p, miR20a/b, miR106a/b(e^{-10}).	miR-493(e^{-14}). miR-15a/b, miR-16,miR195, miR-424, miR-497(e^{-13}).
Best PPI module	Hsapiens_Module_19(e^{-14}) BP:collagen catabolic process. MF:collagen binding. CC:anchoring collagen.	Hsapiens_Module_39(e^{-5}) BP:nuclear-transcribed mRNA catabolic process . MF:DNA-directed RNA polymerase activity. CC:proteasome accessory complex.
Cytogenetic band	7p(e^{-5}), 7p15(e^{-5}). 10q(e^{-14}), 10q22(e^{-5}).	10p(e^{-11}), 10p15(e^{-4}), 10p11(e^{-4}).
Disease	Neoplasm invasiveness(e^{-19}). Neovascularization, pathologic(e^{-18}). Collagen diseases(e^{-15}). Fibrosis(e^{-13}). Adhesion(e^{-15}). Carcinoma(e^{-14}). Metaplasia(e^{-14}). Vascular skin diseases (e^{-11})	Brain Neoplasms(e^{-4}). Glioma(e^{-4}). Mental disorders(e^{-3}).
Drug	Collagenase(e^{-14}). Herapin(e^{-13}). Alteplase(e^{-9}). Urokinase(e^{-8}).	
Phenotype	Premature rupture of membranes(e^{-4}). Joint laxity(e^{-4}). Molluscoid pseudotumors(e^{-4}). Workman bones(e^{-4}). Dilation of the ascending aorta(e^{-4}). Soft skin(e^{-4}).	Abnormality of skeletal maturation(e^{-3}). Abnormality of palate(e^{-4}), Abnormality of lip(e^{-4}), Abnormal hair pattern(e^{-3}). ADHD(e^{-5}). Autism(e^{-5}). Rhabdomyoma(e^{-3}).

multilineage dysplasia as a separate category. CALGB[46] is based on specific cytogenetic abnormalities known for prediction of outcome. The WHO classification is considered more relevant now because it places greater emphasis on prognostic factors.[47] However, it is complex and not provided with this data.

Gender, CALGB, and FAB distributions among the subtypes are shown in Figure S23. Summaries for mutations in the different subgroups are shown in Table S24. Summaries for AML biomarkers in the different subgroups are shown in Table S25. Violin plots of blood counts, are shown in Figure S24. Interacting and confounding variables are shown in figures S26 and S27.

We find that gender is not a factor for distinguishing PINS survival groups in LAML. Since there is no confounder with an influence on the selection of groups to compare, we perform four comparisons of each subgroup against the union of all others. For example, DE genes are calculated for group '1' compared to the union of groups '2', '3', and '4'. As before, we use these gene sets to perform pathway impact analysis with the KEGG pathway database, and gene set enrichment using the databases in WebGestalt, independently for genes that are relatively up-regulated and down-regulated.

The FDR-corrected p-values of the pathways that are the most significant in the four comparisons are shown in Table S28. The majority of these pathways were significant in more than one comparison. Several of the pathways are selected to show the contrast in fold-change for the different PINS subtypes, in 2x2 format, a graph for each subtype, shown in figures S25 through S29: Hematopoietic Cell Lineage, Tuberculosis, Antigen Processing and Presentation, Primary Immunodeficiency, and Phagosome. WebGestalt results are shown in tables S29 through S32.

4.3.1 High survival group - APL

PINS subgroup '1' has the best survival, and our results suggest that it may represent the Acute Promyelocytic Leukemia (APL) subtype. Subgroup '1' is characterized by younger patients, lower percent bone marrow blasts, and higher percent bone marrow lymphocytes (Figure S24). All FAB M3 cases are in group '1', and group '1' consists of 83% M3 cases (Table S23). FAB M3 is the Acute Promyelocytic Leukemia (APL) subtype, caused by the fusion of part of the RAR-alpha gene from Chromosome 17 to the PML region on Chromosome 15. All members of subgroup '1' are PML-RAR positive and 93% of all PML-RAR cases are in group '1'. APL has favorable prognosis, and subgroup '1' patients are seen to be in better CALB risk groups (Table S24). Table S25 shows that group '1', is associated with negative CD34 and negative HLA-DR (Human Leukocyte antigen) - negativity of both together is highly indicative of the APL subtype.[48]

4.3.2 Intermediate survival group - mitochondrial

Highest in percent bone marrow blasts, PINS subtype '2' is not defined by any specific hematopoietic lineages (Figure S25), and does not actively process and present antigens (Figure S27). WebGestalt results, shown in Table S30, show that genes

Table S23. Distribution of the 164 LAML patients among gender, FAB classification (www.cancer.org, the American Cancer Society) and CALGB (Cancer and Leukemia Group B) risk group (www.calgb.org). Gender is shown specifically to emphasize that there is an almost equal number of males and females in each category. FAB classification and CALGB were significant with an FDR adjusted p-value < 0.05 in at least one of the comparisons between the four survival clusters. The first column **(A)**, gives the actual number of patients in each survival group per phenotypic category. There are differing numbers of missing values in each category, so the sum of the number of patients will not be the same in column (A) of every sub-table. The second column **(B)**, gives the percentage of each phenotypic subcategory in each of the survival groups (horizontal/column sum is 100). The third column **(C)**, gives the percentage of each of the survival groups in each of the phenotypic subcategories (vertical/row sum is 100). For example, survival group '1' has 15 members with FAB=M3; 100% of FAB=M3 are in group '1', and 83% of the members of group '1' have FAB=M3. Percentages greater than 50% are highlighted. FAB classifications: M0 - Undifferentiated acute myeloblastic leukemia; M1 - Acute myeloblastic leukemia with minimal maturation; M2 - Acute myeloblastic leukemia with maturation; M3 - Acute promyelocytic leukemia (APL); M4 -Acute myelomonocytic leukemia (AMML); M5 -Acute monocytic leukemia; M6 -Acute erythroid leukemia; M7 - Acute megakaryoblastic leukemia. We see that group '1', with the best survival, is strongly populated with FAB M3, APL patients, which is also a good CALGB risk group. Survival group '3' is predominantly FAB M4, AMML. There are very few patients in the cohorts with FAB M6 or M7, but all are in group '4'.

		(A) Number in each group				**(B)** % phenotype in each group				**(C)** % group in each phenotype			
Survival Group		1 (19)	2 (73)	3 (39)	4 (33)	1	2	3	4	1	2	3	4
gender	female	10	31	21	16	13	40	27	21	53	42	54	48
	male	9	42	18	17	10	49	21	20	47	58	46	52
French -American -British (FAB) classification	M0	0	6	0	9	0	40	0	60	0	8	0	27
	M1	1	23	5	6	3	**66**	14	17	6	32	13	18
	M2	2	24	2	8	6	**67**	6	22	11	33	5	24
	M3	15	0	0	0	**100**	0	0	0	**83**	0	0	0
	M4	0	5	25	5	0	14	**71**	14	0	7	**64**	15
	M5	0	14	7	0	0	**67**	33	0	0	19	18	0
	M6	0	0	0	2	0	0	0	**100**	0	0	0	6
	M7	0	0	0	3	0	0	0	**100**	0	0	0	9
CALGB risk group	Good	14	8	9	1	44	25	28	3	**74**	11	23	3
	Med	3	52	27	13	3	**55**	28	14	16	**73**	**69**	39
	Poor	2	11	3	19	6	31	9	**54**	11	15	8	**58**

up-regulated in cluster '2' are dominated by mitochondrial and translation terms. The antimicrobial tigecycline kills the majority of AML cells in vitro and in xenograft models,[49] through mitochondrial translation inhibition, and has finished phase I clinical safety trials for treatment of AML.[50] While no clinical trial results are posted at the time of this writing, our results suggest that specifically members of cluster '2' may benefit from this treatment, as opposed to AML patents in general.

4.3.3 Intermediate survival group - monocytic

Subgroup '3' has poor survival, and is dominated by myelocyte (neutrophil) and monocyte (macrophage) lineages, inflammation and phagocytosis terms. It is highest in bone marrow monocytes, as shown in Figure S25. Table S23 shows that 71% of the FAB M4 cases are in group'3' and 64% of group '3' cases are in FAB M4. FAB M4 is Acute Myelomonocytic leukemia (AMML). Table S25 shows that group '3' includes 80% of the patients that are CD14 positive. CD14 is a marker for dendritic cell differentiation[51] and presence of monocytes and macrophages. CD14 indicates myeloid lineage, is often positive in FAB M4 and M5.[52] Group '3' includes the largest proportion of positive staining NSE (nonspecific esterase), which indicates the presence of cells of monocytic origin, but can be positive across several FAB subtypes (www.pathologystudent.com). The KEGG Tuberculosis pathway depicts a macrophage-dendritic cell. It is particularly up-regulated in cell surface proteins in group '3', shown in Figure S26c. Tuberculosis infects macrophages in the lungs and obstructs phagosome activity and antigen presentation.[53] Patients with hematological disorders, especially AML, are very susceptible to tuberculosis.[54,55] WebGestalt results, in Table S31, show a dominance of terms for endocytosis and phagocytosis. Of note, subgroup '3' has a relative lack of genes on Chromosome 19.

4.3.4 Poor survival group - MPAL

Group '4' has the worst survival and includes the patients with the greatest variety of mutations. All representatives of FAB M6 (Erythroleukemia) and FAB M7 (Acute megakaryoblastic leukemia) are in survival cluster '4', although there are also members of other FAB subtypes (except M3 and M5). Up-regulation of genes on the KEGG pathway "Hematopoietic Lineage", in Figure S25, show that it has higher lymphoid markers than the other groups, and therefore may be "mixed phenotype acute leukemia", or (MPAL).[47] Patients with MPAL present with a large number of cytogenetic abnormalities, are difficult to treat, and have high mortality rate.[56] MPAL accounts for between 2% and 5% of AML cases,[56,57] although there are other AML classes with MPAL phenotype. Group '4' comprises 20% of the AML cases in this study, and therefore is probably not purely MPAL. However, 64% of group '4' have FISH abnormalities, which is consistent with,[58] who tested 92 patients with MPAL and showed that 64% had cytogenic abnormalities. HLA-DR and CD34 tend to be positive in MPAL, but MPAL is heterogeneous, and may not be a distinct entity. Table S24 shows the highest number of 5q and 7q deletions, the poor risk in Table S23, a high significance of several cytogenetic abnormalities, high interaction of CALGB risk group with these cytogenetic abnormalities in Table S26, and confounding of several cytogenetic abnormalities with other clinical variables in Table S27. WebGestalt results, in Table S32, support the strong presence of T-cell leukemia (ALL) along with B-cell leukemia (AML). In addition, we note that there is a highly significant overabundance of genes on Chromosomes 22, 11, and 19, but significant loss of genes on Chromosomes 5 and 7.

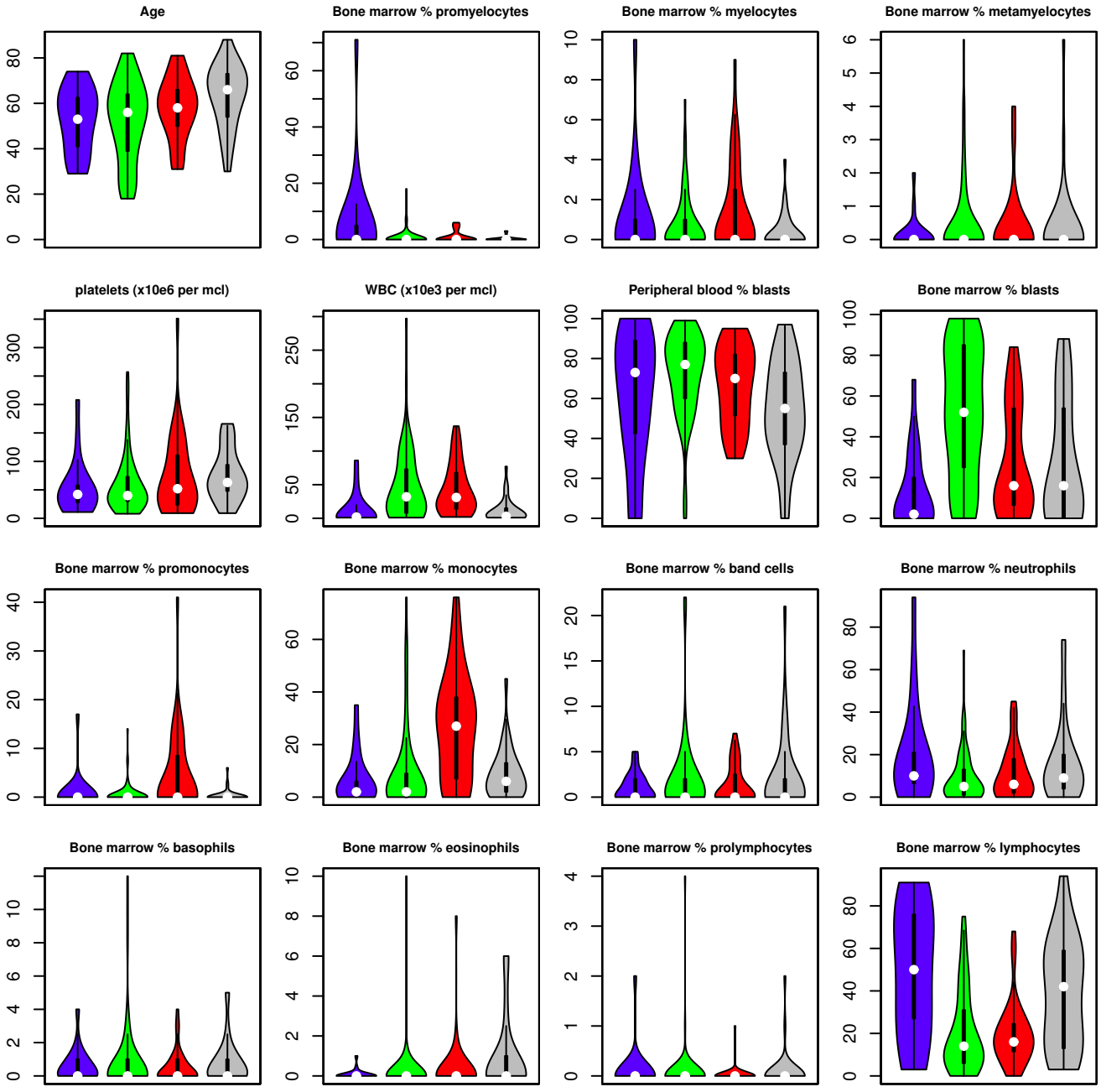


Figure S24. Violin plots showing the distribution of the 164 LAML patients among the continuous variables for clinical variables for blood counts and age. There were very few missing values for blood counts. Group '1' is shown in blue, group '2' in green, group '3' in red, and group '4' in grey. We see that increasing age correlates to the survival groups. Group '1' includes patients with higher promeylocyte and lymphocyte counts, but a lower number of blasts in the bone marrow. Group '2' is much higher than the others in percent of bone marrow blasts, includes patients with high white blood cell counts (WBC), and several patients with high basophils, eosinophils, and prolymphocytes in the bone marrow. Group '3' is much higher than the others in percent of bone marrow monocytes. Group '4' includes the oldest patients, and has the lowest percent of blasts in the peripheral blood.

Table S24. LAML mutations. Several chromosomal aberrations were reported in the clinical data. Many were too sparse to provide signifiant results. The clinical data presented here were sufficiently informative and were significant with an FDR adjusted p-value < 0.05 in at least one of the comparisons between the four survival clusters. Data provided as counts were recoded as pos (> 0), and neg (zero). We combined the binary (pos/neg) data and the continuous data for 'PML RAR' (we use greater than 20% as positive). FISH abnormality indicates whether any chromosomal abnormality is present at all according to fluorescence in-situ hybridization. FISH abnormalities occur in the majority of group '1', and these mutations always include the classic APL rearrangement PML-RAR. Group '4' includes the majority of 5q and 7q deletion events. Most of the members of groups '2' and '3' had detectable FISH abnormalities, and although the specific mutational tendencies for these groups were not distinguishable, they often include 5q and/or 7q deletions.

		(A) Number in each group				**(B)** % phenotype in each group				**(C)** % group in each phenotype			
		1	2	3	4								
Survival Group		(19)	(73)	(39)	(33)	1	2	3	4	1	2	3	4
FISH	neg	1	40	17	8	2	**61**	26	12	6	**68**	**53**	36
abnormality	pos	16	19	15	14	25	30	23	22	**94**	32	47	**64**
deletion 5q	neg	18	64	33	20	13	47	24	15	**95**	**97**	**97**	**67**
	pos	1	2	1	10	7	14	7	**71**	5	3	3	33
deletion 7q	neg	17	64	33	16	13	49	25	12	**89**	**97**	**97**	**53**
	pos	2	2	1	14	11	11	5	**74**	11	3	3	47
PML RAR	neg	0	45	21	10	0	**59**	28	13	0	**98**	**100**	**100**
	pos	14	1	0	0	**93**	7	0	0	**100**	2	0	0

Table S25. LAML biomarkers. CD14 is a marker for dendritic cell differentiation[51] and presence of monocytes and macrophages. It indicates myeloid lineage, is often positive in FAB M4 and M5,[52] and shorter survival in patients with secondary AML (non-de novo).[59] Only group '2' and '3' members have any CD14 positive disease, and group '3' includes 80% of patients with this marker. AML cells positive for CD34 myeloid antigen are believed to be more resistant to apoptosis,[60] and indicative of high relapse rate and poor survival.[61] APL (FAB M3), which is dominant in group '1', is associated with negative CD34 and negative HLA-DR (Human Leukocyte antigen) - negativity of both is highly indicative of the APL subtype.[48] Abnormalities in the nucleophosmin (NPM1) gene resulting in abnormal localization in the leukemic-cell cytoplasm are called $NPMc^+$. It occurs across AML subtypes, and is indicative of better survival in adults.[62] Only groups '2' and '3' include a percentage of patients with $NPMc^+$. Group '3' includes the largest proportion of positive staining NSE (nonspecific esterase), which indicates the presence of cells of monocytic origin, but can be positive across several FAB subtypes (www.pathologystudent.com).

		(A) Number in each group				**(B)** % phenotype in each group				**(C)** % group in each phenotype			
		1	2	3	4								
Survival Group		(19)	(73)	(39)	(33)	1	2	3	4	1	2	3	4
CD14	neg	7	17	12	13	14	35	24	27	**100**	**89**	**60**	**100**
	pos	0	2	8	0	0	20	**80**	0	0	11	40	0
CD34	neg	12	19	6	0	32	**51**	16	0	**75**	32	22	0
	pos	4	41	21	33	4	41	21	33	25	**68**	**78**	**100**
HLA DR	neg	11	8	1	1	**52**	38	5	5	**85**	18	4	5
	pos	2	37	22	18	2	47	28	23	15	**82**	**96**	**95**
$NPMc^+$	neg	18	45	24	33	15	38	20	28	**100**	**63**	**62**	**100**
	pos	0	26	15	0	0	**63**	37	0	0	37	38	0
NSE	neg	13	43	12	22	14	48	13	24	**87**	**68**	32	**88**
	pos	2	20	26	3	4	39	**51**	6	13	32	**68**	12

Table S26. LAML significant clinical variables and interactions. Only Group versus all other results are shown, and only if the nominal p-value is less than 10^{-3}.

Group vs. others	variable	coeff
1	bone marrow % lymphocytes	0.04
1	PML.RAR	-2.70
1	CALGB	-3.17
1	HLA-DR	-3.75
1	CD34	-2.43
2	bone marrow blast count	0.03
2	fish abnormality detected	-1.29
3	monocytes count	0.06
3	NSE (nonspecific esterase)	1.91
3	bone marrow promonocyte count result %	0.20
3	percent blasts peripheral blood*monocytes count	-0.003
4	trisomy 8*CALGB	3.58
4	MLL rearrangements*CALGB	3.78
4	bone marrow % promonocytes *CALGB	3.68
4	bone marrow % prolymphocytes*CALGB	3.61
4	bone marrow % promonocytes*CALGB	4.71
4	bone marrow % cellularity*CALGB	3.58
4	$>$ 3 distinct abnormalities	2.77
4	deletion 7q	2.99
4	deletion 5q	2.67
4	MPX (myeloperoxidase)	-1.63
4	bone marrow percent cellularity	-0.03

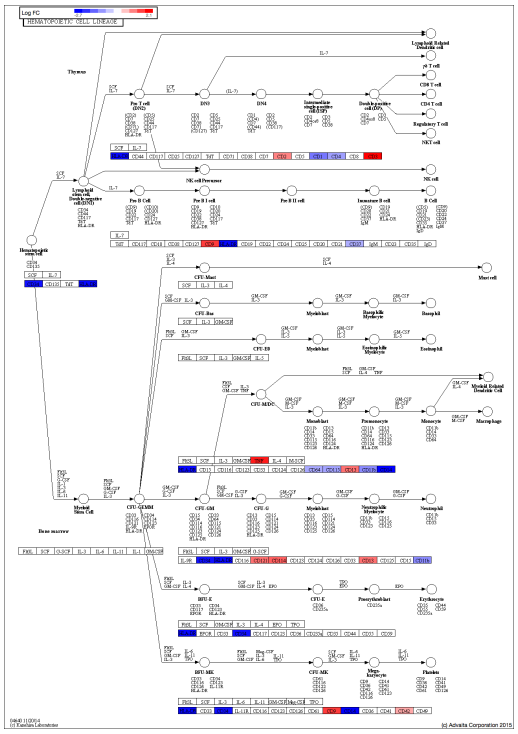
Table S27. AML confounding clinical variables and interactions. Only Group versus all other results are shown, and only if the absolute coefficients are greater than one, the p-values for both tests are < 0.01, and the percent difference before and after are $> 40\%$.

Gp1	Variable	Coef1	Coef2
1	history hematologic disorder / WBC count	2.89	4.09
1	history hematologic disorder / HLA-DR	2.89	4.34
1	history hematologic disorder / CALGB cyto risk group	2.89	4.47
1	history hematologic disorder / CD34	2.89	4.56
2	CALGB / CD33	1.29	1.84
2	CALGB / CD56	1.29	2.11
2	CALGB / transloc(8-21)	1.29	3.35
3	CBF-Beta / bone marrow % promonocytes	-1.33	-1.83
3	CBF-Beta / PML-RAR	-1.33	-1.89
3	history neoadjuvant hydroxyurea / bone marrow % blasts	1.07	1.52
3	NSE (nonspecific esterase) / deletion 5q	1.91	3.09
4	deletion 5q / HLA-DR	-2.25	-3.5
4	deletion 5q / peripheral blood % blasts	-2.25	-3.73
4	FAB category / CBF-Beta	-1.98	-3.63
4	MPX (myeloperoxidase) / AML1-ETO	-1.63	-2.62
4	MPX (myeloperoxidase) / deletion 5q	-1.63	-3.41
4	history other malignancy / HLA-DR	1.37	2.04
4	history other malignancy / trisomy 8	1.37	2.05
4	history other malignancy / NSE (nonspecific esterase)	1.37	2.12
4	history other malignancy / CBF-Beta	1.37	2.72
4	bone marrow % cellularity / CD56	1.66	3.24
4	trisomy 8 / gender	-1.86	-2.63
4	trisomy 8 / peripheral blood % blasts	-1.86	-2.63
4	trisomy 8 / bone marrow % band cells	-1.86	-3.37
4	trisomy 8 / CD33	-1.86	-2.89

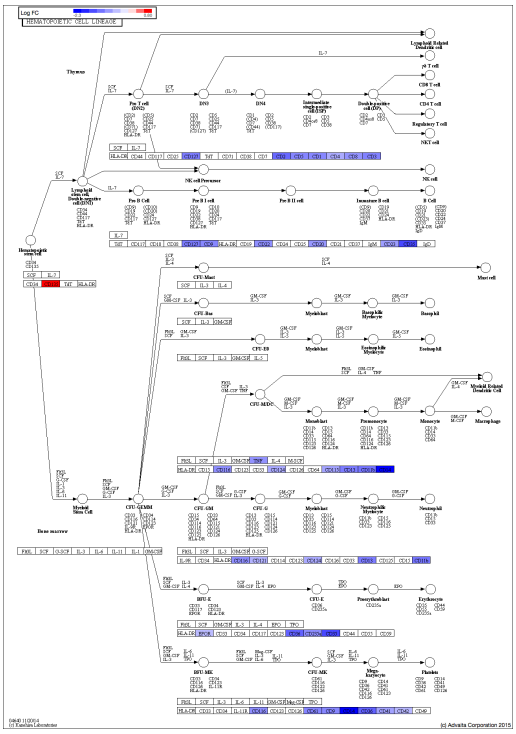
Table S28. All KEGG pathways significant at $FDR < e^{-4}$ for each AML survival cluster versus the set of all others.

Group vs. others	Pathway name	FDR
1	Antigen processing and presentation	e^{-5}
2	Antigen processing and presentation	e^{-5}
1	Cell adhesion molecules (CAMs)	e^{-8}
2	Cell adhesion molecules (CAMs)	e^{-5}
4	Cell adhesion molecules (CAMs)	e^{-6}
2	Hematopoietic cell lineage	e^{-14}
4	Hematopoietic cell lineage	e^{-9}
3	Leishmaniasis	e^{-6}
2	Leukocyte transendothelial migration	e^{-5}
2	Natural killer cell mediated cytotoxicity	e^{-5}
2	Osteoclast differentiation	e^{-5}
2	Phagosome	e^{-6}
3	Phagosome	e^{-8}
2	Primary immunodeficiency	e^{-5}
4	Primary immunodeficiency	e^{-5}
1	Rheumatoid arthritis	e^{-5}
3	Rheumatoid arthritis	e^{-5}
4	Staphylococcus aureus infection	e^{-5}
4	Systemic lupus erythematosus	e^{-5}
3	Toxoplasmosis	e^{-5}
1	Tuberculosis	e^{-5}
2	Tuberculosis	e^{-6}
3	Tuberculosis	e^{-6}

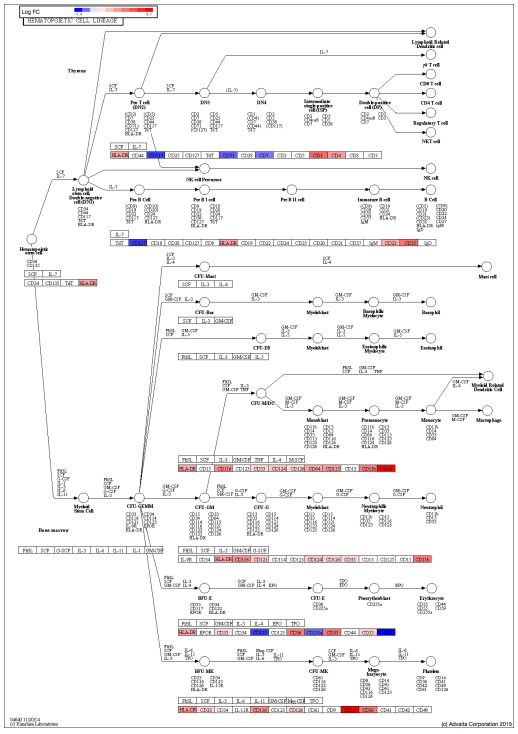
(a) Group '1' (FDR p-value = 0.001)



(b) Group '2' (FDR p-value = $2e^{-14}$)



(c) Group '3' (FDR p-value = $2e^{-4}$)



(d) Group '4' (FDR p-value = $5e^{-9}$)

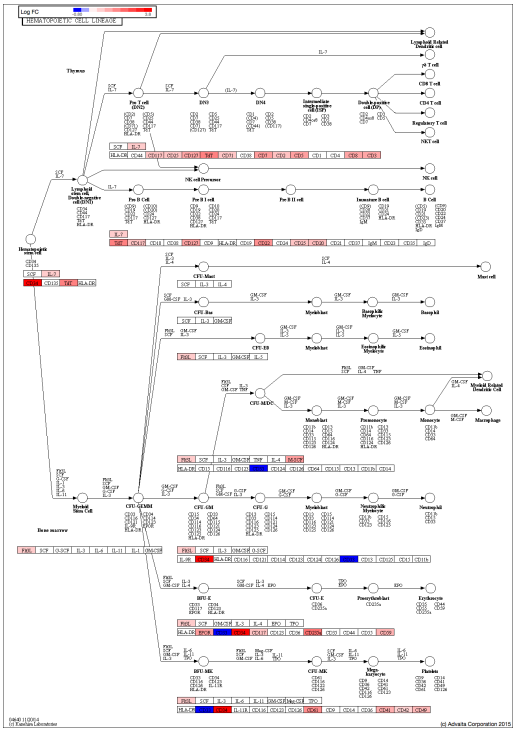
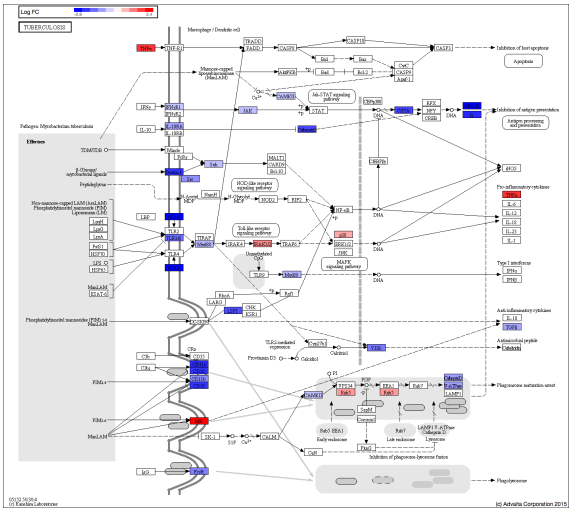
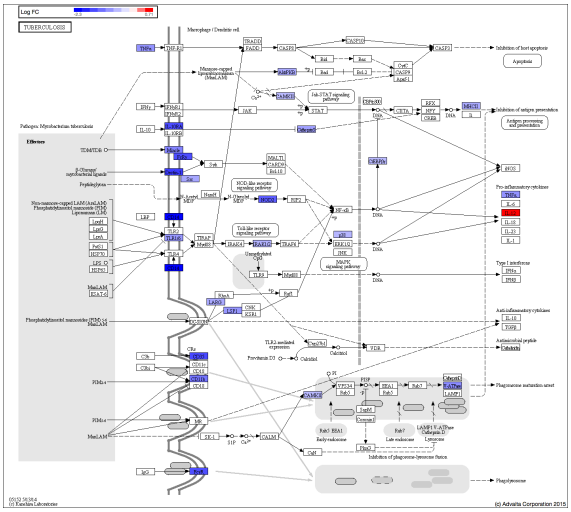


Figure S25. The KEGG pathway "Hematopoietic Cell Lineage" was significant for all four survival clusters when compared to the set of all others. (a) HLA-DR and CD34 are down as expected for APL. (b) All lineages markers are less represented than in other survival clusters. Multipotent progenitor marker CD135 (FLT gene) is high. (c) Macrophage and neutrophil lineages markers are most represented, notably CD14. (d) Lymphoid lineages markers are more present than myeloid.

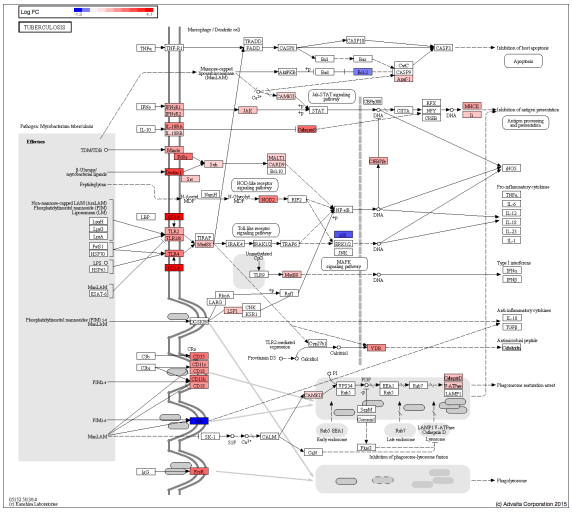
(a) Group '1' (FDR p-value = $3e^{-5}$)



(b) Group '2' (FDR p-value = $3e^{-6}$)



(c) Group '3' (FDR p-value = $2e^{-6}$)



(d) Group '4' (FDR p-value = 0.035)

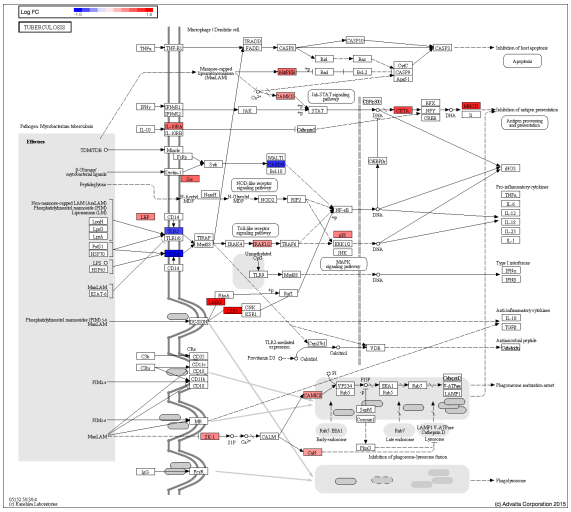
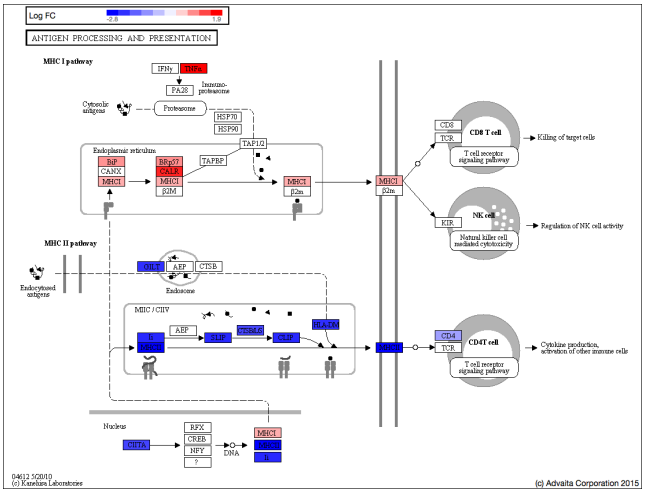
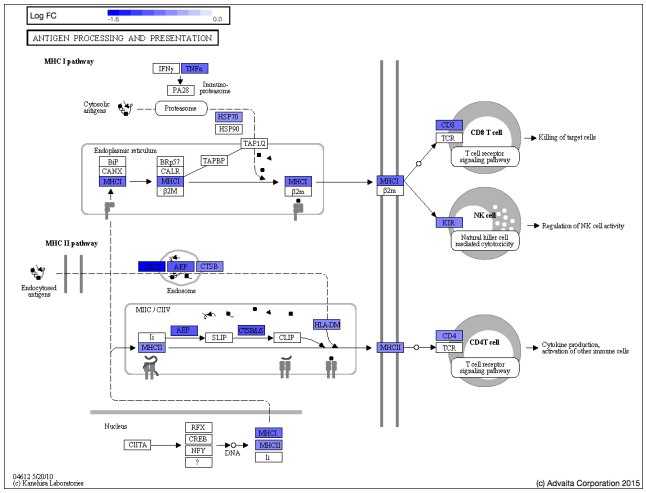


Figure S26. The KEGG pathway "Tuberculosis" was significant for survival clusters '1', '2', and '3' when compared to the set of all others. This KEGG pathway depicts a macrophage. It is particularly up-regulated in cell surface proteins in group '3'. Tuberculosis infects macrophages in the lungs and obstructs phagosome activity and antigen presentation.[53] Patients with hematological disorders, especially AML, are very susceptible to tuberculosis.[54,55]

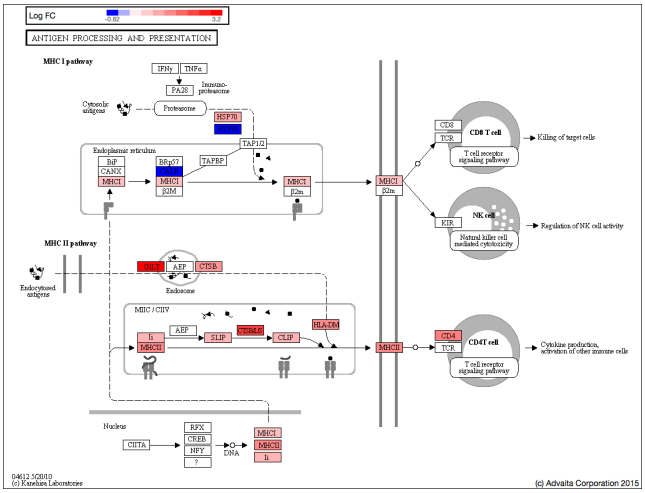
(a) Group '1' (FDR p-value = $2e^{-5}$)



(b) Group '2' (FDR p-value = $4e^{-5}$)



(c) Group '3' (FDR p-value = $8e^{-4}$)



(d) Group '4' (FDR p-value = 0.01)

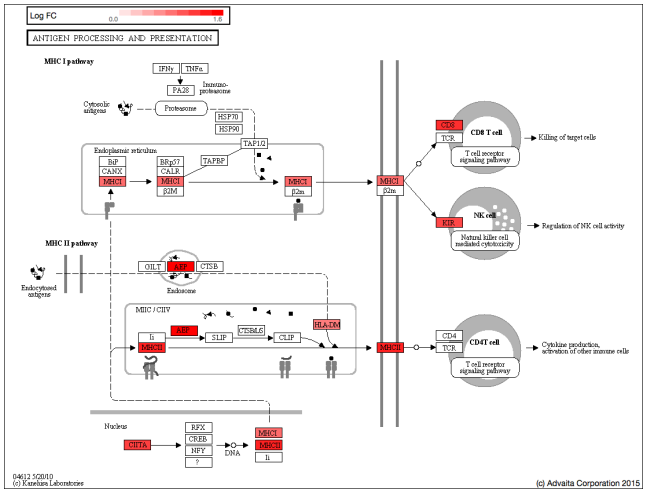
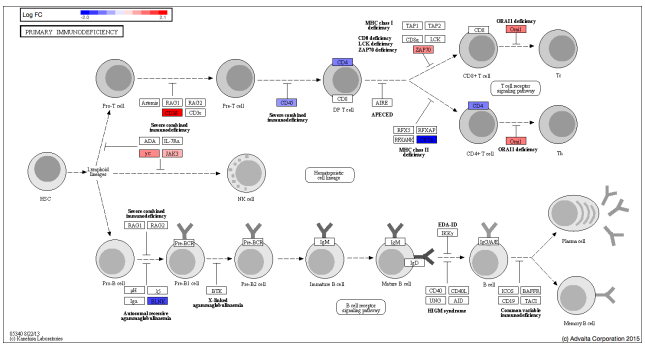
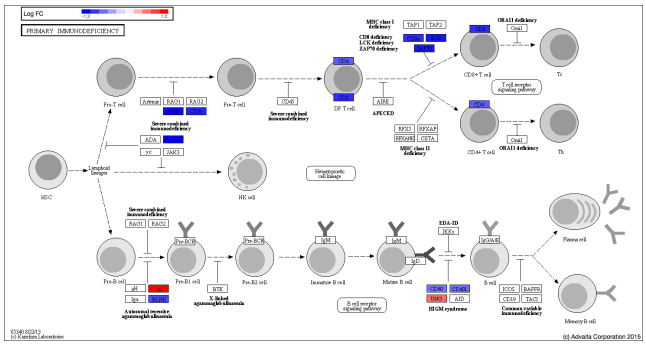


Figure S27. The KEGG pathway "Antigen Processing and Presentation" was significant for all four survival clusters when compared to the set of all others.

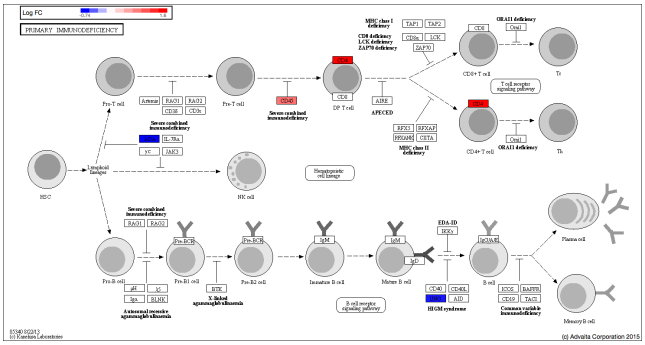
(a) Group ‘1’ (FDR p-value = 0.04)



(b) Group ‘2’ (FDR p-value = $2e^{-5}$)



(c) Group ‘3’ (FDR p-value = 0.9)



(d) Group ‘4’ (FDR p-value = $7e^{-5}$)

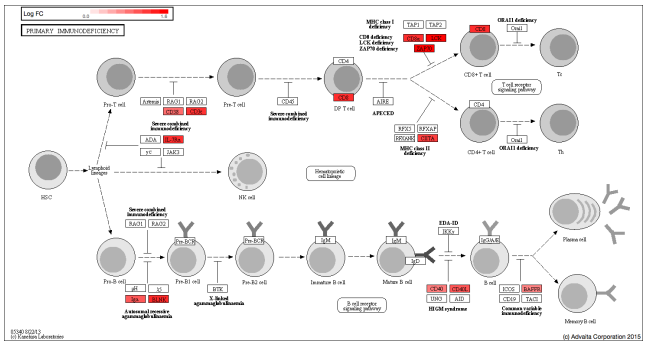
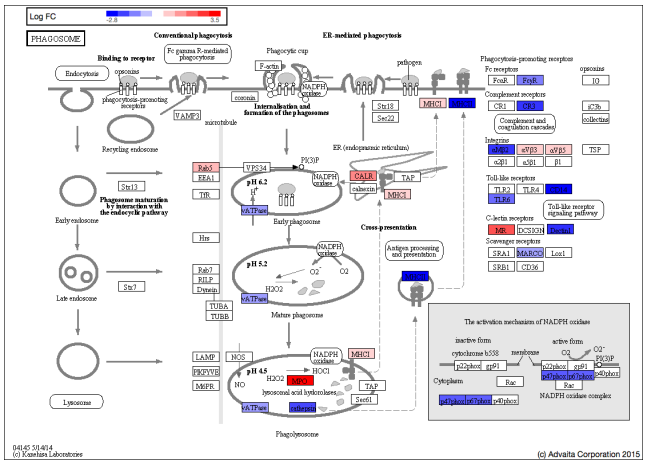


Figure S28. The KEGG pathway “Primary Immunodeficiency” was significant for survival clusters ‘2’, and ‘4’ when compared to the set of all others. This KEGG pathway depicts lymphoid lineages, and is particularly up-regulated in group ‘4’.

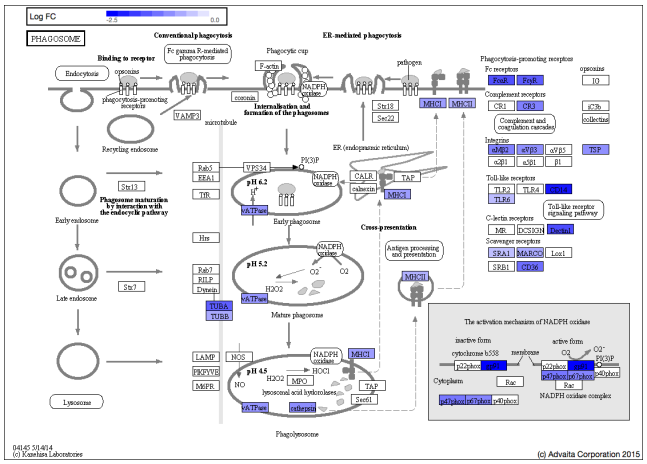
Table S29. Group ‘1’ vs ‘others’. The gene set ‘others’ consists of the union of groups ‘2’, ‘3’, and ‘4’. Gene Ontology and disease terms, based on genes selected at FDR $<$ 0.01. Many Pathway Commons results were signaling pathways significant at the same level, and due to the same genes. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Group ‘1’ vs ‘others’		
Database	1215 genes up in ‘1’ (down in others)	1105 genes down in ‘1’ (up in others)
GO Biological Process	Activation of signaling protein activity involved in unfolded protein response (e^{-6}). Regulation of nuclease activity(e^{-5}). Positive regulation of protein phosphorylation (e^{-3}). Glycoprotein biosynthetic process (e^{-6}). Extracellular matrix organization (e^{-4}).	Leukocyte activation (e^{-16}). Lymphocyte activation (e^{-14}). Regulation of immune response (e^{-19}). Response to stress (e^{-16}).
GO Molecular Function		Small molecule binding (e^{-3}). Lipid binding(e^{-5}). Protein binding (e^{-7}). Phosphoric ester hydrolase activity (e^{-4}). Enzyme activator activity (e^{-4}).
GO Cellular Component	Extracellular matrix (e^{-5}). Endoplasmic reticulum lumen (e^{-9}). Endoplasmic reticulum-Golgi intermediate compartment (e^{-3}). Cytoplasmic vesicle (e^{-4}). Primary cilium (e^{-3}). Tight junction (e^{-4}).	Endocytic vesicle (e^{-5}). ER to Golgi transport vesicle (e^{-4}). Vacuole (e^{-9}). Cytosol (e^{-10}). Endosome (e^{-11}). MHC protein complex (e^{-6}). Trans-Golgi network membrane (e^{-5}). Endoplasmic reticulum membrane (e^{-4}).
Pathway Commons	Unfolded Protein Response (e^{-5}). Asparagine N-linked glycosylation (e^{-4}). N-glycan trimming in the ER and Calnexin/Calreticulin cycle (e^{-4}). Diabetes pathways (e^{-4}). Post-translational protein modification (e^{-3}). Calnexin/calreticulin cycle (e^{-3}). Activation of Chaperones by IRE1alpha (e^{-3}).	Immune System (e^{-19}). PAR1-mediated thrombin signaling events (e^{-14}). Thrombin/protease-activated receptor (PAR) pathway (e^{-14}) . . . plus 33 more signaling pathways at (e^{-14}) significance, all involving the same set of approximately 140 genes.
microRNAs that target geneset		miR-506(e^{-5}). miR-182(e^{-4}).
Best PPI module		Hsapiens_Module_275 (e^{-6}). BP:T cell costimulation MF: peptide antigen binding CC:MHC class II protein complex
Cytogenetic band		10q(e^{-6}), 10q24 (e^{-5}), 10q22 (e^{-3}). 7p15 (e^{-4}). 6p21 (e^{-3}).
Disease	Abnormal axial skeleton morphology (e^{-4}).	Virus Diseases (e^{-18}). Infection (e^{-18}). Immune System Diseases (e^{-16}). Immunologic Deficiency Syndromes (e^{-12}).
Drug		Immune globulin (e^{-6}).

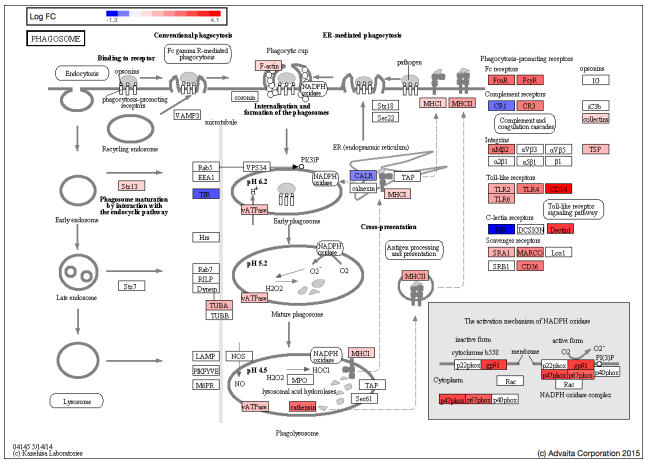
(a) Group '1' (FDR p-value = $3e^{-4}$)



(b) Group '2' (FDR p-value = $6e^{-6}$)



(c) Group '3' (FDR p-value = $9e^{-8}$)



(d) Group '4' (FDR p-value = 0.186)

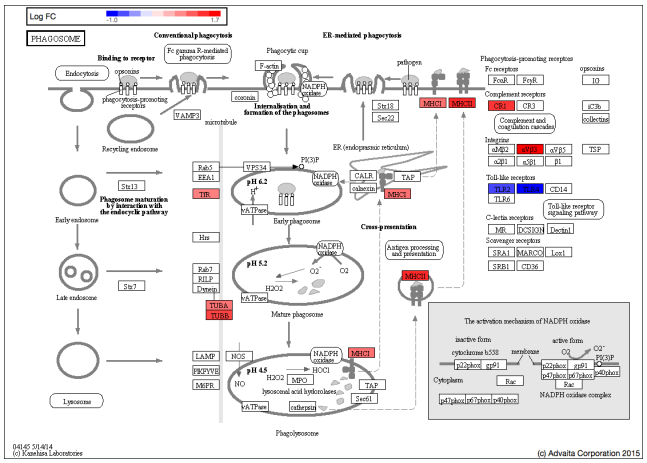


Figure S29. The KEGG pathway "Phagosome" was significant four survival clusters '1', '2', and '3' when compared to the set of all others.

Table S30. Group '2' vs 'others'. The gene set 'others' consists of the union of groups '1', '3', and '4'. Gene Ontology, disease and phenotype terms, based on genes selected at FDR < 0.01. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Group '2' vs 'others'		
	2796 genes	1396 genes
Database	up in '2' (down in others)	down in '2' (up in others)
GO Biological Process	Cellular macromolecular process(e^{-33}). ncRNA processing(e^{-29}). Ribosome biogenesis(e^{-36}). Translation initiation(e^{-39}), Translation elongation(e^{-47})). Nuclear-transcribed mRNA catabolic process, nonsense mediated decay(e^{-47}). Protein targeting to ER(e^{-37}). Cellular metabolic process(e^{-47}). Viral genome expression(e^{-33}).	Leukocyte activation(e^{-27}), lymphocyte activation(e^{-23}). Regulation of immune response(e^{-34}). Innate immune response(e^{-27}). Cytokine production(e^{-23}). Inflammatory response(e^{-22}).
GO Molecular Function	Binding (RNA(e^{-75}), Nucleotide(e^{-19}), Unfolded protein(e^{-11}), ATP(e^{-6})). Catalytic activity(e^{-12}). Helicase activity(e^{-7}). Structural constituent of ribosome(e^{-56}). Methyltransferase activity(e^{-46}).	Binding (Immunoglobulin(e^{-5}), Growth factor(e^{-5}), Enzyme(e^{-5}), Chemokine(e^{-5}), Actin(e^{-5}), Lipid(e^{-7}).). Kinase activity(e^{-5}). Enzyme regulator activity(e^{-6}). Chemokine receptor activity.(e^{-6}) MHC class I receptor activity.
GO Cellular Component	Ribosome(e^{-65}). Nucleolus(e^{-37}). Nuclear lumen(e^{-49}) Mitochondrion(e^{-83}).	Extracellular space(e^{-9}). Cytosol(e^{-6}). Endosome(e^{-7}). Integral to plasma membrane.(e^{-14}) Vacuole(e^{-8}). Phagocytic vesicle(e^{-7}). Lysosome(e^{-8}).
Pathway Commons	Gene Expression(e^{-60}). Metabolism of RNA(e^{-48}). Eukaryotic Translation Elongation(e^{-47}). 3' -UTR-mediated translational regulation(e^{-47}). Translation(e^{-47}). GTP hydrolysis and joining of the 60S ribosomal subunit(e^{-46}). Peptide chain elongation(e^{-46}). Eukaryotic Translation Termination(e^{-45}).	Integrin family cell surface interactions(e^{-37}). Proteoglycan syndecan-mediated signaling events(e^{-35}). Immune System(e^{-34}). Syndecan-1-mediated signaling events(e^{-34}). Sphingosine 1-phosphate (S1P) pathway(e^{-34}).
microRNAs that target geneset		miR-17-5p, Mir-20a/b, miR-10a/b, miR-519d(e^{-7}). miR-96(e^{-5}). miR-506(e^{-4}). miR-34a,/c, miR-449(e^{-4}).
Best PPI module	Hsapiens_Module_230(e^{-32}) BP:ribosomal small subunit biogenesis MF:structural constituent of ribosome. CC:cytosolic small ribosomal subunit.	Hsapiens_Module_111(e^{-15}) BP: actin polymerization or depolymerization. MF:non-membrane spanning protein tyrosine kinase activity. CC:lamellipodium.
Disease	Mitochondrial diseases(e^{-11}). Anemia (Diamond-Blackfan, Aplastic)(e^{-9}). Shock(e^{-4}).	Immune system diseases(e^{-44}). Autoimmune diseases(e^{-28}). Lymphoproliferative disorders.(e^{-25}) Inflammation(e^{-28}). Infection(e^{-25}). Connective tissue diseases(e^{-22}).
Drug	dactinomycin(e^{-48}). Tobramycin(e^{-12}). Kanamycin(e^{-9}). immune globulin(e^{-27}) Adenosine triphosphate(e^{-9}). collagenase(e^{-6}). sodium lauryl sulfate(e^{-5}).	
Phenotype	Abnormality of mitochondrial metabolism(e^{-5}). Acidosis(e^{-9}). Abnormality of amino acid metabolism(e^{-5}). Decreased liver function(e^{-5}). Microcytic anemia(e^{-5}). Lethargy(e^{-4}). Abnormality of the CNS(e^{-4}).	Abnormality of blood and blood forming tissues(e^{-6}). Abnormality of the lymphatic system(e^{-5}). Abnormality of the spleen(e^{-6}).

Table S31. Group '3' vs 'others'. Gene Ontology, disease and phenotype terms, based on genes selected at FDR < 0.01. The gene set 'others' consists of the union of groups '1', '2', and '4'. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Group '3' vs 'others'		
	1837 genes	4409 genes
Database	up in '3' (down in others)	down in '3' (up in others)
GO Biological Process	Innate immune response(e^{-39}). Positive regulation of immune system(e^{-28}). Signal transduction(e^{-33}) Leukocyte activation(e^{-25}). T-cell activation(e^{-17}). Endocytosis(e^{-16}). Inflammatory process(e^{-26}). Regulation of cytokine production(e^{-19}).	Translation termination(e^{-28}). Translation elongation(e^{-31}). RNA metabolic process(e^{-37}). ncRNA metabolic process(e^{-40}). Nuclear-transcribed mRNA cataboloc process, nonsense-mediated decay(e^{-29}). Nuclease-containing compound metabolic process(e^{-57}). Viral genome expression(e^{-27}).
GO Molecular Function	Binding (lipid(e^{-17}), actin(e^{-9}), immunoglobulin(e^{-6}), kinase(e^{-6})). Enzyme regulator activity(e^{-7}).	Binding (ion(e^{-8}), nucleic acid(e^{-60}), heterocyclic compound(e^{-47})). Aminoacyl-tRNA ligase activity(e^{-11}). Helicase activity(e^{-5}). Methyltransferase activity(e^{-19}).
GO Cellular Component	Vacuole(e^{-31}). Endosome(e^{-29}). Golgi apparatus(e^{-11}). Phagocytic vesicle(e^{-8}). Lysosome(e^{-32}). MHC protein complex(e^{-8}).	Ribonucleoprotein complex(e^{-48}), ribosomal subunit(e^{-40}). Mitochondrion(e^{-34}). Nuclear lumen(e^{-44}).
Pathway Commons	Immune System(e^{-31}). S1P1 pathway(e^{-29}). Internalization of ErbB1(e^{-29}). ErbB1 downstream signaling(e^{-29}). Arf6 trafficking events(e^{-29}). Thrombin/protease-activated receptor (PAR) pathway(e^{-29}). Sphingosine 1-phosphate (S1P) pathway(e^{-29}).	Gene expression(e^{-50}). Eukaryotic Translation Elongation(e^{-33}). Peptide chain elongation(e^{-31}). Metabolism of RNA(e^{-31}). Eukaryotic Translation Termination(e^{-31}).
microRNAs that target geneset	miR-506(e^{-9}). miR-19a/b(e^{-7}). miR-124a(e^{-7}).	
Best PPI module	Hsapiens_Module_111(e^{-10}) BP:actin polymerization or depolymerization MF:non-membrane spanning protein tyrosine kinase activity CC:lamellipodium.	Hsapiens_Module_39(e^{-28}) BP:nuclear-transcribed mRNA catabolic process. MF:DNA-directed RNA polymerase activity. CC:proteasome accessory complex.
Cytogenetic band		19q(e^{-7}), 19q13(e^{-7}), 19q34(e^{-5}). 19p(e^{-7}), 19p13(e^{-7}). 22q(e^{-5}).
Disease	Inflammation(e^{-24}). Immune system diseases(e^{-23}). Infection(e^{-23}). Necrosis(e^{-19}). Arthritis(e^{-14}).	Diamond-Blackfan anemia(e^{-5}). Mitochondrial diseases(e^{-3}).
Drug	Immune globulin(e^{-14}). Glutathione(e^{-9}).	dactinomycin(e^{-25}). tobramycin(e^{-12}). kanamycin(e^{-11}).
Phenotype	Abnormality of blood and blood forming tissues(e^{-5}). Abnormality of the lymphatic system(e^{-6}). Abnormality of the spleen(e^{-6}). Abnormality of macrophages(e^{-6}).	Abnormality of the cerebrum(e^{-8}). Microcephaly(e^{-7}). Morphological abnormality of the CNS(e^{-7}). Abnormality of the optic nerve(e^{-5}). Decreased liver function(e^{-5}). Anemia(e^{-5}). Increased serum lactate(e^{-5}). Muscular hypotonia(e^{-5}).

Table S32. Group '4' vs 'others'. The gene set 'others' consists of the union of groups '1', '2', and '3'. Gene Ontology, disease and phenotype terms, based on genes selected at FDR-corrected p-value < 0.01. Values in parentheses after each term are the FDR-corrected p-values for the enrichment.

Group '4' vs 'others'		
Database	1878 genes up in '4' (down in others)	760 genes down in '4' (up in others)
GO Biological Process	Blood coagulation(e^{-6}). T-cell activation(e^{-11}). Positive regulation of leukocyte activation(e^{-6}). Cell surface receptor signaling pathway(e^{-10}), Antigen receptor-mediated signaling pathway(e^{-6}). Immune system development(e^{-11}), Hemopoisis(e^{-10}).	Protein localization(e^{-5}), Protein transport(e^{-5}). Vacuole organization(e^{-3}). Glycolipid metabolic process(e^{-3}). Carbohydrate catabolic process(e^{-4}). Mitochondrial transport(e^{-3}), Respiratory electron transport chain(e^{-5}). Energy derivation by oxidation of organic compounds(e^{-4}). Protein folding(e^{-3}).
GO Molecular Function	Binding (metal ion(e^{-4}), growth factor(e^{-4})). Protein kinase activity(e^{-3}). GTPase regulator activity(e^{-3}). Oxygen transporter activity(e^{-5}).	Catalytic activity(e^{-6}), Oxidoreductase activity(e^{-5}). Hydrogen ion transmembrane transported activity(e^{-3}). Unfolded protein binding(e^{-5}).
GO Cellular Component	External side of plasma membrane(e^{-10}). Hemoglobin complex(e^{-5}). Plasma membrane(e^{-6}). Actin filament(e^{-3}). Membrane raft(e^{-3}).	Melanosome(e^{-4}). Vacuole(e^{-9}). Lysosome(e^{-7}). Golgi apparatus(e^{-4}). Endoplasmic reticulum(e^{-4}), Endoplasmic reticulum lumen(e^{-5}). Mitochondrion(e^{-15}), Mitochondrial inner membrane(e^{-12}), Respiratory chain(e^{-4}).
Pathway Commons	Integrin family cell surface interactions(e^{-13}). Proteoglycan syndecan-mediated signaling events(e^{-13}). LKB1 signaling events(e^{-12}). Thrombin/protease-activated receptor (PAR) pathway(e^{-12}). Plasma membrane estrogen receptor signaling(e^{-12}). PAR1-mediated thrombin signaling events(e^{-12}). Glypican pathway(e^{-12}). IFN-gamma pathway(e^{-12}).	The citric acid (TCA) cycle and respiratory electron transport(e^{-5}). Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins(e^{-5}). Glucose metabolism(e^{-4}).
microRNAs that target geneset	miR-518c(e^{-9}) miR-133a/b(e^{-4}).	
Best PPI module	Hsapiens_Module_25(e^{-10}) BP: JAK-STAT cascade involved in growth hormone signaling pathway. MF: non-membrane spanning protein tyrosine kinase activity. CC: platelet alpha granule .	Hsapiens_Module_201(e^{-4}) BP: cofactor metabolic process MF: catalytic activity CC:endoplasmic reticulum membrane
Cytogenetic band	22q(e^{-22}), 22q13(e^{-10}), 22q12(e^{-9}), 22q11(e^{-3}). 11q(e^{-4}), 11q13(e^{-7}). 19q(e^{-4}). 19q13(e^{-4}). 1p(e^{-4}),1p36(e^{-4}). 19p(e^{-3}).	7q(e^{-29}), 7q22(e^{-12}), 7q31(e^{-5}), 7q32(e^{-4}). 7p(e^{-7}), 7p15(e^{-3}). 5q(e^{-26}), 5q31(e^{-14}), 5q22(e^{-5}), 5q33(e^{-5}).
Disease	Immune system diseases(e^{-14}). Lymphoproliferative disorders(e^{-14}). Leukemia(e^{-14}). Leukemia T-cell(e^{-13}). Lymphatic diseases(e^{-12}). Lymphoma(e^{-12}). Lymphoma T-cell(e^{-12}). Hemolytic anemia(e^{-11}). Lymphoma B-cell(e^{-11}).	Nelson's syndrome(e^{-8}). Lysosomal Storage diseases(e^{-5}).
Drug	Immune globulin(e^{-10}). Epoprostenol(e^{-3}).	NADH(e^{-4}). adenosine triphosphate(e^{-4}). ciprofloxacin(e^{-4}).
Phenotype	Hemolytic anemia(e^{-7}). Abnormality of B cells(e^{-3}). Abnormality of the heme biosynthetic pathway(e^{-3}).	Dysostosis multiplex(e^{-3}).

4.4 Tools used in functional analysis

For functional analysis described above, we perform the following tests:

- Statistical enrichment of clinical variables in subgroups using the hypergeometric test (using the R package *phyper*),
- Gene set enrichment with each of the following databases: Biological Process, Molecular Function and Cellular Component (Gene Ontologies,[63] version 1.2, 11/11/2012), microRNA targets (MSigDB,[64] 11/11/2012), phenotype (Human Phenotype Ontology[65] 04/10/2013), cytogenetic band (NCBI Gene 10/26/2012), protein-protein interaction modules (NetGestalt[26] 11/11/2012), Pathway Commons (11/11/2012), and disease and drug associated genes (PharmKGB[66,67] 1/26/2013), using WebGestalt,[68] on differentially expressed mRNAs, separately for up-regulated and down-regulated genes,
- Pathway impact analysis on KEGG,[53] using iPathwayGuide.[24]

Enrichment of subgroups with different phenotypic parameters and clinical variables is tested by comparing each survival cluster to every other, and each to the set of all others, for each clinical variable. If a clinical parameter is significant, based on nominal p-value < 0.01, for any of these comparisons, it is included in a summary table. These tables have three sections: the first gives the number of patients represented for each group and each significant clinical parameter and state, the second gives percentages of each clinical state in each group, and the third gives percentages of each group in each clinical state.

Pathway analysis was performed using the topologically-based impact analysis[69,70] which calculates the significance of signaling pathways by taking into account not only the over-representation of genes in pathways, but also the relative positions of the genes, the type and direction of all their interactions on pathway networks, etc. The impact analysis is implemented in the iPathwayGuide[24] analysis package that we used here.

WebGestalt is a web-based tool that calculates enrichment using the hypergeometric test[71] for a large number of different functional genomics databases. We use the "BH" (Benjamini and Hochberg) option for multiple hypothesis correction, the significance level option is set to 0.01, and the "Minimum Number of Genes for a Category" option is set to two. For each WebGestalt-housed database that we use to test for comparative gene set enrichment between two disease subtypes, we provide a table of the most significant results, for up-regulated and down-regulated genes separately. For each result reported, minus log FDR adjusted p-values are shown in brackets after the term or group of terms. A term must be at a significance level of 1% or better, after FDR correction, to be reported.

MicroRNA results are provided by WebGestalt as families, so the members are separated by commas, with the minus log of the FDR-corrected p-value given at the end of the set. We report the protein-protein interaction (PPI) module ID with the best significance, followed by the "related function" Gene Ontology[63] terms provided for the module. Chromosome band results are grouped by chromosome arm.

For ontological databases such as the Gene Ontology and the Mammalian Phenotype Ontology, WebGestalt provides the results in a directed acyclic graph (DAG), with the root node and all significant and intermediate nodes connected, according to the user specified FDR-corrected p-value cutoff. To select which terms are reported for each ontology, we first locate the major subgraphs coming off the root node (e.g. Biological Process). For each subgraph, we report the most significant nodes (if the FDR-corrected p-value is < 0.01). If the subgraph is large, we identify connected groups of significant nodes and report the best node from each. When there are many nodes of equivalent significance, we report the node with the most genes, or if those are all the same too, we select a significant representative node in the middle of the hierarchy. The FDR-corrected p-value is given after each term in brackets.

References

1. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91–118 (2003).
2. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
3. Kuner, R. *et al.* Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* **63**, 32–38 (2009).
4. Hou, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE* **5**, e10312 (2010).
5. Tarca, A. L. *et al.* Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER diagnostic signature challenge. *Bioinformatics* **29**, 2892–2899 (2013).
6. Mills, K. I. *et al.* Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114**, 1063–1072 (2009).
7. Le Dieu, R., Taussig, D. C., Ramsay, A. G., Mitter, R., Miraki-Moud, F., Fatah, R., Lee, A. M., Lister, T. A. & Gribben, J. G. Peripheral blood T cells in acute myeloid leukemia (AML) patients at diagnosis have abnormal phenotype and genotype and form defective immune synapses with AML blasts. *Blood* **114**, 3909–3916 (2009).
8. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
9. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164–4169 (2004).
10. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* **98**, 13790–5 (2001).
11. Pomeroy, S. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
12. Bolstad, B. M. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. Ph.D. thesis, University of California (2004).
13. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748 (1959).
14. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187–220 (1972).
15. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model* (Springer, 2000).
16. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. & Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333–337 (2014).
17. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *Jama* **247**, 2543–2546 (1982).
18. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).
19. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
20. Zhang, J. *CNTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses* (2014).
21. Kaufman, L. & Rousseeuw, P. *Clustering by Means of Medoids* (Faculty of Mathematics and Informatics, 1987).
22. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

23. Klumperman, J., Oude Elferink, R., Fransen, J., Ginsel, L. & Tager, J. M. Secretion of a precursor form of lysosomal alpha-glucosidase from the brush border of human kidney proximal tubule cells. *European Journal of Cell Biology* **50**, 299–303 (1989).

24. Advaita Corporation. Pathway Analysis with iPathwayGuide. http://www.advaitabio.com/ipathwayguide.html.

25. Sung, B., Park, S., Yu, B. P. & Chung, H. Y. Modulation of PPAR in aging, inflammation, and calorie restriction. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **59**, B997–B1006 (2004).

26. Shi, Z., Wang, J. & Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nature Methods* **10**, 597–598 (2013).

27. Zabarovsky, E. R., Lerman, M. I., Minna, J. D. *et al.* Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene* **21**, 6915–6935 (2002).

28. Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

29. Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157–173 (2006).

30. Chinnaiyan, P., Kensicki, E., Bloom, G., Prabhu, A., Sarcar, B., Kahali, S., Eschrich, S., Qu, X., Forsyth, P. & Gillies, R. The metabolomic signature of malignant glioma reflects accelerated anabolic metabolism. *Cancer Research* **72**, 5878–5888 (2012).

31. Mammoto, T., Jiang, A., Jiang, E., Panigrahy, D., Kieran, M. W. & Mammoto, A. Role of collagen matrix in tumor angiogenesis and glioblastoma multiforme progression. *The American Journal of Pathology* **183**, 1293–1305 (2013).

32. Payne, L. S. & Huang, P. H. The pathobiology of collagens in glioma. *Molecular Cancer Research* **11**, 1129–1140 (2013).

33. Amelio, I., Cutruzzola, F., Antonov, A., Agostini, M. & Melino, G. Serine and glycine metabolism in cancer. *Trends in Biochemical Sciences* **39**, 191–198 (2014).

34. DeBerardinis, R. J. Serine metabolism: some tumors take the road less traveled. *Cell Metabolism* **14**, 285–286 (2011).

35. Lavon, I. *et al.* Gliomas display a microRNA expression profile reminiscent of neural precursor cells. *Neuro-Oncology* **12**, 422–433 (2010).

36. Hu, J., Jiang, C., Ng, H., Pang, J. & Tong, C. Chromosome 14q may harbor multiple tumor suppressor genes in primary glioblastoma multiforme. *Chinese Medical Journal* **115**, 1201–1204 (2002).

37. Misra, A., Pellarin, M., Nigro, J., Smirnov, I., Moore, D., Lamborn, K. R., Pinkel, D., Albertson, D. G. & Feuerstein, B. G. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clinical Cancer Research* **11**, 2907–2918 (2005).

38. Vranova, V., NeCesalova, E., Kuglik, P., Cejpek, P., PeSakova, M., Budinska, E., Relichova, J. & Veselska, R. Screening of genomic imbalances in glioblastoma multiforme using high-resolution comparative genomic hybridization. *Oncology Reports* **17**, 457–464 (2007).

39. Vogazianou, A. P., Chan, R., Backlund, L. M., Pearson, D. M., Liu, L., Langford, C. F., Gregory, S. G., Collins, V. P. & Ichimura, K. Distinct patterns of 1p and 19q alterations identify subtypes of human gliomas that have different prognoses. *Neuro-Oncology* **12**, 664–678 (2010).

40. Bellail, A. C., Hunter, S. B., Brat, D. J., Tan, C. & Van Meir, E. G. Microregional extracellular matrix heterogeneity in brain modulates glioma cell invasion. *The International Journal of Biochemistry & Cell Biology* **36**, 1046–1069 (2004).

41. Lopez-Gines, C. *et al.* Association of chromosome 7, chromosome 10 and EGFR gene amplification in glioblastoma multiforme. *Clinical Neuropathology* **24**, 209–218 (2004).

42. Inda, M. d. M. *et al.* Chromosomal abnormalities in human glioblastomas: gain in chromosome 7p correlating with loss in chromosome 10q. *Molecular Carcinogenesis* **36**, 6–14 (2003).

43. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood* **100**, 2292–2302 (2002).

44. Vardiman, J. W. *et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114**, 937–951 (2009).

45. Bennett, J. M., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D., Gralnick, H. t. & Sultan, C. Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group. *British Journal of Haematology* **33**, 451–458 (1976).

46. Schilsky, R. L., McIntyre, O. R., Holland, J. F. & Frei, E. A concise history of the cancer and leukemia group B. *Clinical Cancer Research* **12**, 3553s–3555s (2006).

47. The American Cancer Society. How is acute myeloid leukemia classified? http://www.cancer.org/cancer/leukemia-acutemyeloidaml/detailedguide/leukemia-acute-myeloid-myelogenous-classified (2014).

48. Promsuwicha, O. & Auewarakul, C. U. Positive and negative predictive values of HLA-DR and CD34 in the diagnosis of acute promyelocytic leukemia and other types of acute myeloid leukemia with recurrent chromosomal translocations. *Asian Pacific Journal of Allergy and Immunology* **27**, 209–216 (2009).

49. Skrtic, M. *et al.* Inhibition of mitochondrial translation as a therapeutic strategy for human acute myeloid leukemia. *Cancer Cell* **20**, 674–688 (2011).

50. Schimmer, A. Phase 1 study evaluating the tolerance and biologic activity of intravenous infusions of tigecycline in patients with relapsed or refractory aml. Tech. Rep., ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US)., University Health Network, Toronto (2011). Identifier NCT01332786.

51. Houtenbos, I., Westers, T., Ossenkoppele, G. & Van De Loosdrecht, A. Identification of CD14 as a predictor for leukemic dendritic cell differentiation in acute myeloid leukemia. *Leukemia* **17**, 1683–1684 (2003).

52. Abdul-Hamid, G. *Classification of Acute Leukemia* (INTECH Open Access Publisher, 2011).

53. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

54. Chen, C.-Y., Sheng, W.-H., Cheng, A., Tsay, W., Huang, S.-Y., Tang, J.-L., Chen, Y.-C., Wang, J.-Y., Tien, H.-F. & Chang, S.-C. Clinical characteristics and outcomes of mycobacterium tuberculosis disease in adult patients with hematological malignancies. *BMC Infectious Diseases* **11**, 324 (2011).

55. Al-Anazi, K. A., Al-Jasser, A. M. & Evans, D. A. Infections caused by mycobacterium tuberculosis in patients with hematological disorders and in recipients of hematopoietic stem cell transplant, a twelve year retrospective study. *Annals of Clinical Microbiology and Antimicrobials* **6**, 16 (2007).

56. Wolach, O. & Stone, R. M. How I treat mixed-phenotype acute leukemia. *Blood* **125**, 2477–2485 (2015).

57. Matutes, E. *et al.* Mixed-phenotype acute leukemia: clinical and laboratory features and outcome in 100 patients defined according to the WHO 2008 classification. *Blood* **117**, 3163–3171 (2011).

58. Yan, L., Ping, N., Zhu, M., Sun, A., Xue, Y., Ruan, C., Drexler, H. G., MacLeod, R. A., Wu, D. & Chen, S. Clinical, immunophenotypic, cytogenetic, and molecular genetic features in 117 adult patients with mixed-phenotype acute leukemia defined by WHO-2008 classification. *Haematologica* **97**, 1708–1712 (2012).

59. Choi, Y., Lee, J.-H., Kim, S.-D., Kim, D.-Y., Lee, J.-H., Seol, M., Kang, Y.-A., Jeon, M., Jung, A. R. & Lee, K.-H. Prognostic implications of CD14 positivity in acute myeloid leukemia arising from myelodysplastic syndrome. *International Journal of Hematology* **97**, 246–255 (2013).

60. van Stijn, A., van der Pol, M. A., Kok, A., Bontje, P. M., Roemen, G., Beelen, R., Ossenkoppele, G. J. & Schuurhuis, G. J. Differences between the CD34+ and CD34-blast compartments in apoptosis resistance in acute myeloid leukemia. *Haematologica* **88**, 497–508 (2003).

61. Feller, N., Schuurhuis, G., Van der Pol, M., Westra, G., Weijers, G., Van Stijn, A., Huijgens, P. & Ossenkoppele, G. High percentage of CD34-positive cells in autologous AML peripheral blood stem cell products reflects inadequate in vivo purging and low chemotherapeutic toxicity in a subgroup of patients with poor clinical outcome. *Leukemia* **17**, 68–75 (2003).

62. Brown, P., McIntyre, E., Rau, R., Meshinchi, S., Lacayo, N., Dahl, G., Alonzo, T. A., Chang, M., Arceci, R. J. & Small, D. The incidence and clinical significance of nucleophosmin mutations in childhood AML. *Blood* **110**, 979–985 (2007).

63. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

64. The GSEA Team. MSigDB. http://www.broadinstitute.org/ (2012).

65. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* gkt1026 (2013).

66. Whirl-Carrillo, M., McDonagh, E., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R. & Klein, T. E. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* **92**, 414–417 (2012).

67. Jourquin, J., Duncan, D., Shi, Z. & Zhang, B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics* **13**, S20 (2012).

68. Wang, J., Duncan, D., Shi, Z. & Zhang, B. Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic Acids Research* **41**, W77–W83 (2013).

69. Drăghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichiţa, C., Georgescu, C. & Romero, R. A systems biology approach for pathway level analysis. *Genome Research* **17**, 1537–1545 (2007).

70. Tarca, A. L., Drăghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P. & Romero, R. A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).

71. Zhang, B., Kirov, S. & Snoddy, J. Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research* **33**, W741–W748 (2005).