Supplementary Materials for "Correlation-based iterative clustering methods for time course data: the identification of temporal gene response modules to influenza infection in humans"

Michelle Carey[a,b], Shuang Wu[a,c], Guojun Gan[d], Hulin Wu[a,e,*]

[a] *Department of Biostatistics and Computational Biology, Crittenden Blvd, Rochester, NY 14642, United States*
[b] *Department of Mathematics and Staistics, McGill University, 805 Sherbrooke Street West, Montreal, Canada.*
[c] *Biogen, 250 Binney Street, Cambridge, MA, USA.*
[d] *Department of Mathematics, University of Connecticut,196 Auditorium Road U-3009, Storrs, USA.*
[e] *Department of Biostatistics, University of Texas Health Science Center School of Public Health at Houston, 1200 Pressler Street, Houston, USA*

S.1. Details of the Design of the H3N2 study

RNA was extracted at Expression Analysis (Durham, NC) from whole blood using the PAXgene 96 Blood RNA Kit (PreAnalytiX, Valencia, CA) employing the manufacturer's recommended protocol. While whole blood RNA is initially extracted, a secondary procedure (B-globin reduction) was then employed to remove the contribution of red blood cell (RBC) RNA to the total RNA. A set of four peptide nucleic acid (PNA) oligomers whose sequences are complementary to the 3 portions of the alpha and beta hemoglobin RNA transcripts were added to reduce globin RNA transcription due to RBC. The inhibition of globin cDNA synthesis dramatically reduces the relative amount of anti-sense, biotin-labeled cRNA corresponding to the hemoglobin transcripts. Hybridization and microarray data collection was performed using the Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA) and expression profiles were pre-processed using robust multi-array (RMA) method [1].

S.2. Convergence of IHC and IPC methods

As the data have considerable measurement error, the index of the clusters rarely converges exactly or reaches full convergence (i.e., the cluster index is identical for consecutive iterations). However, the majority of the genes are clustered into the same cluster at each iteration except for a few genes that are not robust to a slight change in the cluster centre and hence at each

*Corresponding author
Email address: Hulin.Wu@uth.tmc.edu (Hulin Wu)

iteration are assigned to different clusters. Here we define convergence as the majority of cluster indices have converged. Typically, the convergence is around 95% (i.e., >95% of the genes are consecutively being clustered into the same clusters).

S.3. Clustering quality measures

S.3.1. Within-cluster correlation (WCC)

The within cluster correlation (WCC) is defined as

$$WCC = \frac{1}{N}\sum_{i=1}^{N}\bar{d}_i$$

where $\bar{d}_i$ is the average of one minus the pairwise Spearman rank correlation between each gene in the i^{th} cluster. WCC measures the similarity of the genes contained in each cluster.

S.3.2. Between-cluster correlation (BCC)

The between cluster correlation (BCC) is defined as

$$BCC = \frac{1}{N}\sum_{i,j=1,j<i}^{N} d_{i,j}$$

where $r_{i,j}$ is one minus the sample Spearman's rank correlation between the centre of the i^{th} gene and the centre of the j^{th} gene. BCC measures the similarity of the average time course patterns within each cluster.

S.3.3. Davies-Bouldin criterion (DB)

The Davies-Bouldin criterion (DB) is defined as

$$DB = \frac{1}{N}\sum_{i,j=1,j<i}^{N} \max_{j\neq i}\{\frac{\bar{d}_i + \bar{d}_j}{d_{i,j}}\},$$

where $\bar{d}_i$ is the average of one minus the pairwise Spearman rank correlation between each gene in the i^{th} cluster, $\bar{d}_j$ is one minus the average of the pairwise Spearman rank correlation between each gene in the j^{th} cluster, and $d_{i,j}$ is one minus the Spearman rank correlation between the centres of the i^{th} and j^{th} clusters.

S.4. Simulation studies

Figure S.2shows a portion of the of the simulated clusters with $\sigma = 0.1$; this confirms that the simulated clusters do resemble the observed gene expression clusters from the influenza study.

S.5. Semantic similarity

The GO ontologies for gene functional annotation facilitate the comparison of genes by quantifying the similarity of their annotation. Several semantic similarity measures have been proposed [2, 3, 4]. In this paper, we use the popular node-based method attributable to [2]. This calculates the similarity of two GO ontology terms based on the information content of their closest common ancestor term. Let an_t be the closest common ancestor GO term, the information content of an_t is defined as $IC(t) = -log(\frac{|G_{an_t}|}{|G_{root}|})$, where G_{an_t} and G_{root} are the sets of genes annotated to an_t and the root GO term (and all its descendants) respectively. This methodology is implemented using the Matlab software developed by [5] and is available from the following website: http://www.cs.rhul.ac.uk/home/haixuan/GOSIM.html.

S.6. Power Law for the size of temporal gene response modules

Let x be the size of temporal gene response modules (the number of genes in a cluster), we hypothesize that the distribution of x follows $p(x) \propto x^{-\beta}$. Following from [6], we used the Kolmogorov-Smirnov goodness-of-fit statistic to test the hypothesis and the maximum likelihood method to estimate the parameters x_{min} and β for the power-law model, where x_{min} is the minimum size of the clusters.

S.7. Sensitivity analysis

For the IHC, IPC and MCL methods the correlation threshold α is crucial in determining the number of temporal gene response modules. Thus, it is important to assess the robustness and sensitivity of the results to this critical parameter. We repeat the clustering analysis for the time course gene response data from the influenza study with α being set as $0.70, 0.75$ and 0.80, respectively. Table S.8 illustrates the sensitivity of the three clustering methods to either a 5% or 10% change in the threshold parameter α by providing the adjusted rand index of the cluster indexes before and after the change in the threshold parameter. Overall, the three clustering methods are not very sensitive to changes in the correlation threshold. For a 5% change in the correlation threshold, the average adjusted rand index is 0.78, 0.65 and 0.64 for the MCL, IHC, and IPC methods respectively. Similarly, for a 10% change in the correlation threshold, the average adjusted rand index is 0.62, 0.56 and 0.58 for the MCL, IHC, and IPC methods respectively.

Figures S.4-S.6 show the number of genes in the cluster versus the percentage of clusters containing that number of genes across all subjects for each of the three threshold parameters considered in each of the clustering methods.

For the IPC and IHC methods, a small change in the threshold parameter causes a small change in the size and number of clusters produced, while a small change in the threshold parameter has a large effect on the size and number of clusters for the MCL approach. Interestingly Table S.8 showed that the MCL method had the largest average adjusted rand index for either a 5% or 10% change in the threshold parameter yet Tables **??-??** show that the sizes of the clusters does change considerably. Suggesting that decreasing the threshold parameter for the MCL method is essential dividing the existing clusters into smaller groups rather than re-clustering the data.

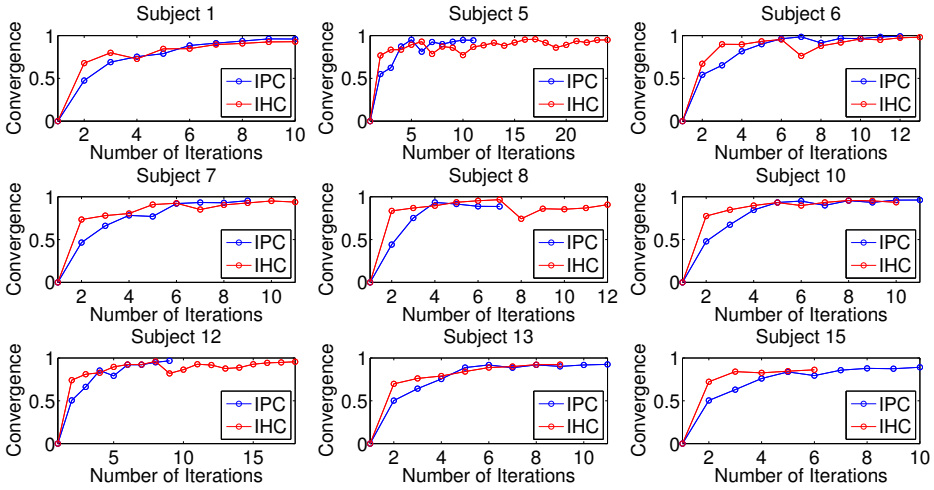

Figure S.1: The convergence of the cluster index for each iteration of the IHC and IPC methods for all 9 symptomatic subjects. A convergence of one implies that the cluster indices are identical. All 9 subjects converge at around 95% (i.e. 95% of the genes are consecutively being clustered into the same clusters).

[1] C. W. Woods, M. T. McClain, M. Chen, A. K. Zaas, B. P. Nicholson, J. Varkey, T. Veldman, S. F. Kingsmore, Y. Huang, R. Lambkin-Williams, A. G. Gilbert, A. O. Hero, III, E. Ramsburg, S. Glickman, J. E. Lucas, L. Carin, G. S. Ginsburg, A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza h1n1 or h3n2, PLoS ONE 8 (1) (2013) e52198. doi:10.1371/journal.pone.0052198. URL http://dx.doi.org/10.1371%2Fjournal.pone.0052198

[2] P. Resnik, et al., Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res.(JAIR) 11 (1999) 95–130.

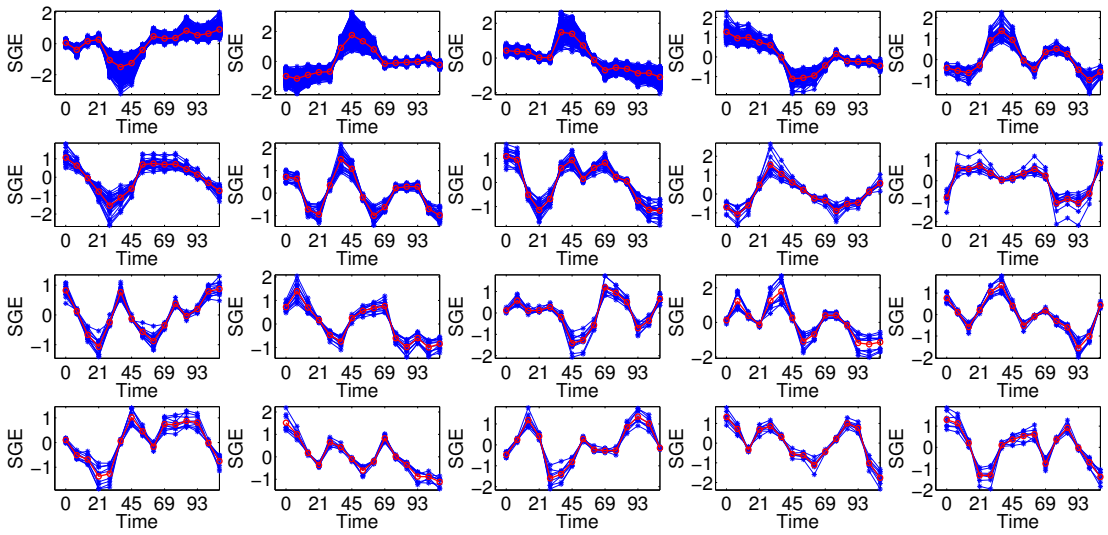

Figure S.2: An example of a proportion of the simulated gene expression (SGE) data sets.

Subject	MCL vs. IHC	MCL vs. IPC	MCL vs. GMM	IHC vs. IPC	IHC vs. GMM	IPC vs. GMM
1	0.4210	0.3743	0.4020	0.5244	0.4716	0.4020
5	0.5105	0.5367	0.5335	0.7016	0.5342	0.5335
6	0.5473	0.5129	0.5556	0.7772	0.5736	0.5556
7	0.6859	0.7273	0.5210	0.7968	0.4662	0.5210
8	0.6801	0.6396	0.6574	0.6369	0.7347	0.6574
10	0.6639	0.6483	0.6988	0.8223	0.7128	0.6988
12	0.6149	0.5011	0.3696	0.6604	0.3951	0.3696
13	0.4720	0.5193	0.5980	0.7190	0.5562	0.5980
15	0.7820	0.7190	0.6915	0.6986	0.7093	0.6915

Table S.1: The adjusted Rand indexes for the GMM, MCL, IHC and IPC methods.

[3] D. Lin, An information-theoretic definition of similarity., in: ICML, Vol. 98, 1998, pp. 296–304.

[4] J. J. Jiang, D. W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, arXiv preprint cmp-lg/9709008.

[5] H. Yang, T. Nepusz, A. Paccanaro, Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty, Bioinformatics 28 (10) (2012) 1383–1389.

[6] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data, SIAM review 51 (4) (2009) 661–703.

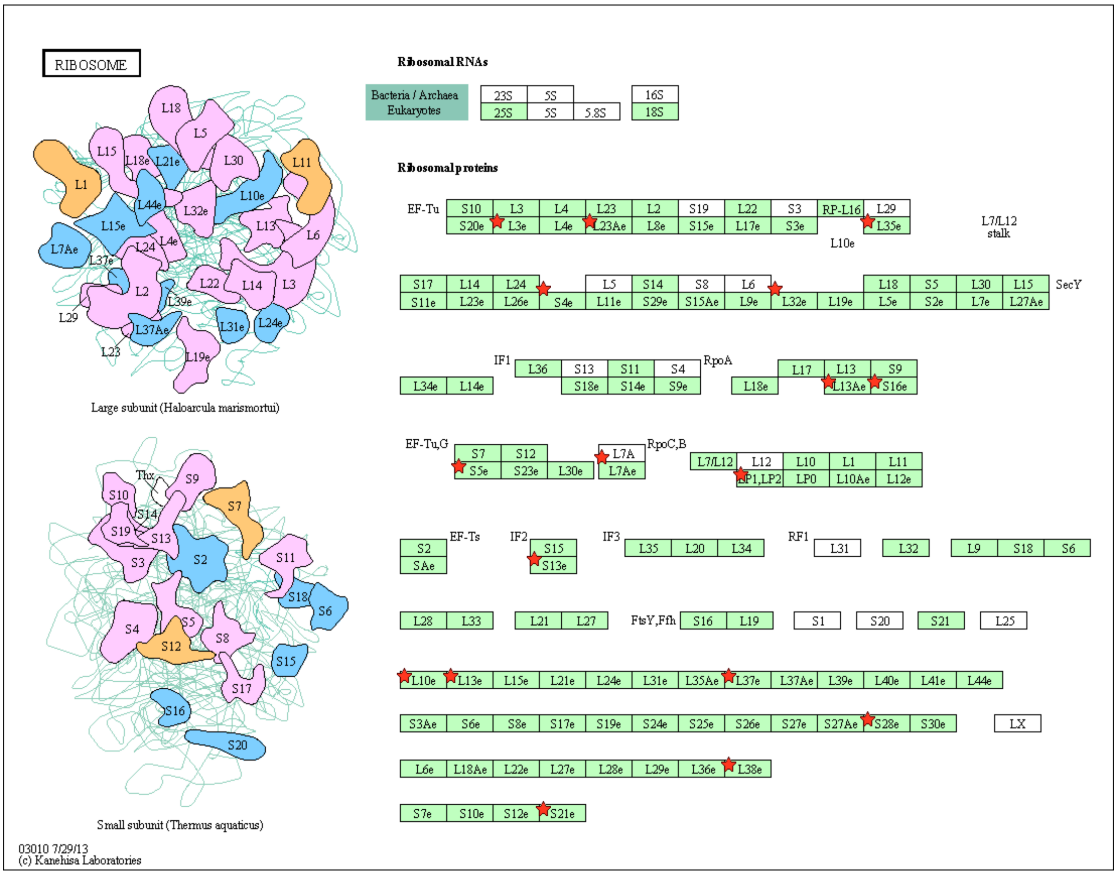

Figure S.3: The enriched Ribosome pathway: 18 genes, P-value 2.2E-5

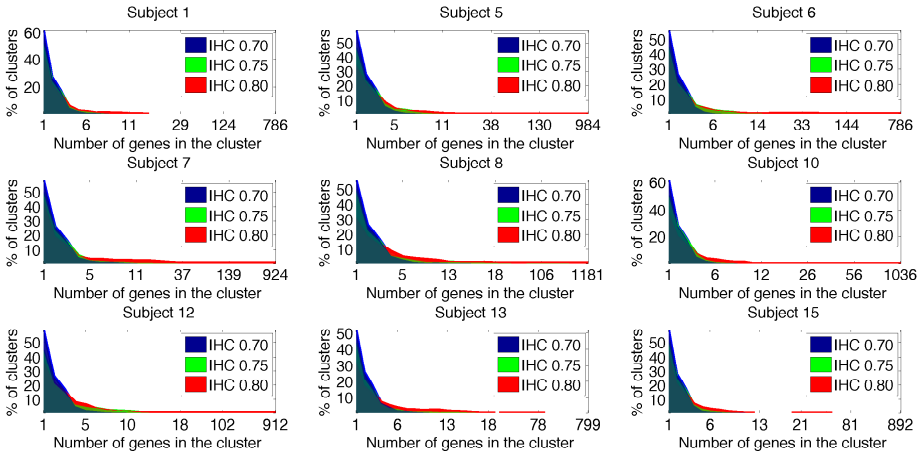

Figure S.4: The distribution of the cluster sizes for IHC method. The number of genes in the cluster versus the percentage of clusters containing that number of genes for each threshold parameter across all nine subjects for the IHC method

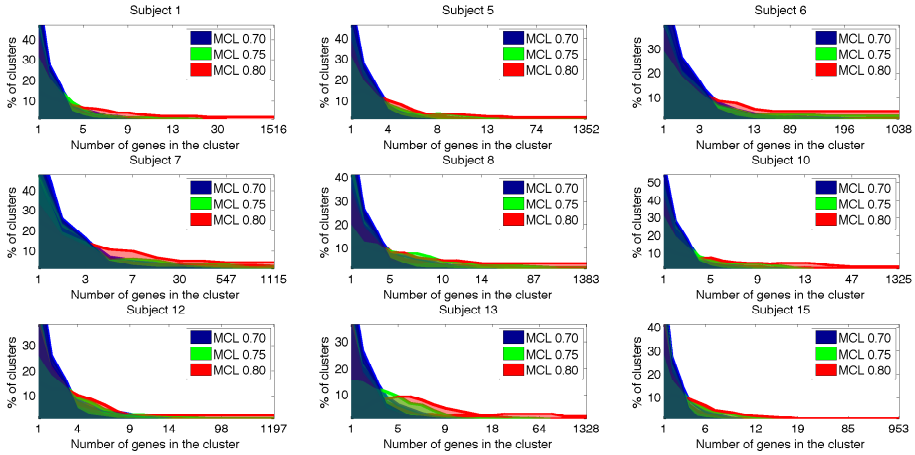

Figure S.5: The distribution of the cluster sizes for MCL method. The number of genes in the cluster versus the percentage of clusters containing that number of genes for each threshold parameter across all nine subjects for the MCL

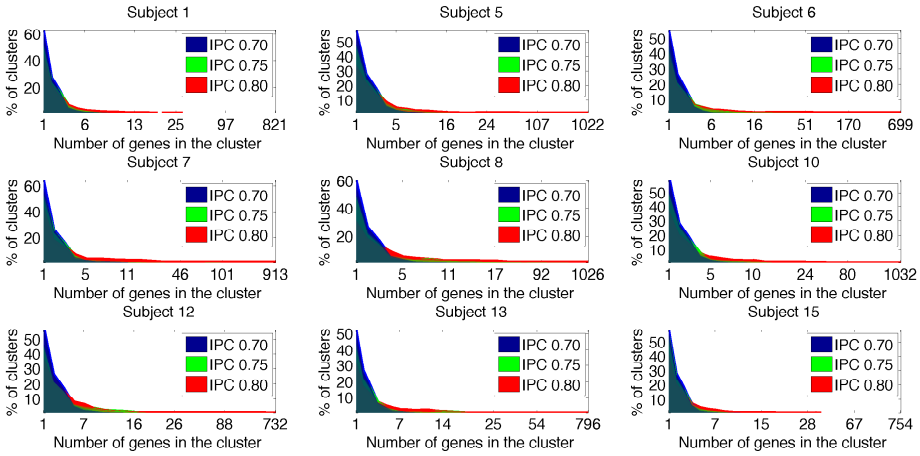

Figure S.6: The distribution of the cluster sizes for IPC method. The number of genes in the cluster versus the percentage of clusters containing that number of genes for each threshold parameter across all nine subjects for the IPC method

Subject	LSM	MSM	SSM	SGM
	%clusters (%genes)	%clusters (%genes)	%clusters (%genes)	%clusters (%genes)
1	6.00% (86.65%)	23.00% (8.51%)	56.00%(4.60%)	15.00% (0.24%)
5	6.00% (86.83%)	17.00% (8.50%)	63.00% (4.43%)	14.00% (0.24%)
6	26.00% (94.93%)	13.00% (3.93%)	39.00% (0.96%)	22.00% (0.18%)
7	17.00% (96.90%)	4.00% (1.00%)	54.00% (1.90%)	25.00% (0.20%)
8	13.00% (89.86%)	30.00% (7.56%)	47.00% (2.46%)	10.00% (0.12%)
10	7.00% (86.26%)	30.00% (9.76%)	50.00% (3.76%)	13.00% (0.22%)
12	10.00% (88.93%)	18.00% (7.70%)	57.00% (3.16%)	15.00% (0.21%)
13	6.00% (83.26%)	26.00% (11.23%)	62.00% (5.40%)	6.00% (0.11%)
15	6.00% (72.90%)	23.00% (18.60%)	62.00% (8.23%)	9.00% (0.27%)

Table S.2: The number of clusters and number of genes (in parenthesis) in each category of modules (LSM, MSM, SSM and SGM) for the MCL method for each of the 9 subjects.

Subject	LSM	MSM
	%clusters (%genes)	%clusters (%genes)
1	43.75% (84.47%)	56.25% (15.53%)
5	53.33% (89.73%)	46.67% (10.27%)
6	66.67% (93.37%)	33.33% (6.63%)
7	75.00% (96.13%)	25.00% (3.87%)
8	41.67% (87.70%)	58.33% (12.30%)
10	31.25% (82.27%)	68.75% (17.73%)
12	58.33% (94.17%)	41.67% (5.83%)
13	42.86% (85.57%)	57.14% (14.43%)
15	57.14% (85.67%)	42.86% (14.33%)

Table S.3: The number of clusters and number of genes (in parenthesis) in each category of modules (LSM, MSM) for the GMM method for each of the 9 subjects. There are no SSM and SGM.

Subject	LSM	MSM	SSM	SGM
	%clusters (%genes)	%clusters (%genes)	%clusters (%genes)	%clusters (%genes)
1	4.80% (73.46%)	15.20% (18.66%)	37.6% (6.12%)	42.40% (1.76%)
5	6.25% (83.66%)	11.46% (10.06%)	42.70% (5.00%)	39.59% (1.28%)
6	12.00% (82.20%)	17.33% (13.70%)	32.00% (3.13%)	38.67% (0.97%)
7	9.23% (84.10%)	18.46% (12.16%)	38.46% (3.00%)	33.85% (0.70%)
8	6.57% (84.96%)	21.31% (11.70%)	36.06% (2.60%)	36.06% (0.74%)
10	5.95% (83.46%)	14.28% (10.76%)	36.92% (4.56%)	42.85% (1.22%)
12	5.88% (79.83%)	14.11% (14.86%)	38.82% (4.13%)	41.19% (1.18%)
13	5.10% (74.36%)	19.38% (19.26%)	34.69% (5.00%)	40.83% (1.38%)
15	3.70% (70.63%)	17.03% (21.36%)	30.39% (5.80%)	48.88% (2.30%)

Table S.4: The number of clusters and number of genes (in parenthesis) in each category of modules (LSM, MSM, SSM and SGM) for the IPC method for each of the 9 subjects.

Subject	Term	P-value
SGM		
1	Membrane fusion (BP)	1.0E-2
1	Extracellular region part (CC)	9.3E-1
1	Calcium-dependent protein binding (MF)	1.0E0
5	Angiogenesis (BP)	1.0E-2
5	Sequence-specific DNA binding (MF)	6.4E-2
6	Actin cytoskeleton organization (BP)	4.3E-3
6	Cytoskeletal part (CC)	4.4E-3
6	Cytoskeletal protein binding (MF)	5.5E-3
7	Response to organic nitrogen (BP)	9.8E-2
7	Intrinsic to endoplasmic reticulum membrane (CC)	9.1E-2
7	Peptide binding (MF)	4.6E-2
8	T cell activation (BP)	4.0E-4
8	Endomembrane system (CC)	9.3E-1
8	Monocarboxylic acid binding (MF)	6.5E-2
10	Regulation of transcription from RNA polymerase II promoter (BP)	1.0E-2
10	Nucleoplasm part (CC)	3.2E-3
10	Transcription regulator activity (MF)	1.6E-2
12	Positive regulation of T cell proliferation (BP)	2.0E-3
12	External side of plasma membrane (CC)	1.6E-3
12	MHC class I protein binding (MF)	2.7E-2
13	Regulation of blood pressure (BP)	9.9E-1
13	Cell junction (CC)	4.5E-2
13	Phosphoprotein phosphatase activity (MF)	4.3E-2
15	Chemical homeostasis (BP)	2.1E-3
15	Ribonucleoprotein complex (CC)	9.4E-2
15	Nucleotide diphosphatase activity (MF)	2.0E-2

Table S.5: The most enriched BP, MF and CC terms for the single gene modules (SGM) for each subject.

Subject	Term	P-value
SSM		
1	negative regulation of cell differentiation (BP)	2.2E-2
1	recycling endosome (CC)	8.6E-3
1	protein tyrosine kinase activity (MF)	2.3E-2
5	tube morphogenesis (BP)	2.1E-4
5	organelle lumen (CC)	2.2E-2
5	hydro-lyase activity(MF)	3.4E-3
6	macromolecular complex subunit organization (BP)	3.8E-4
6	microtubule (CC)	7.9E-4
6	protein N-terminus binding (MF)	7.7E-3
7	regulation of mitosis (BP)	3.2E-2
7	nuclear matrix (CC)	3.6E-2
7	Ras GTPase binding	1.3E-2
8	coenzyme metabolic process (BP)	1.4E-2
8	plasma membrane part (CC)	6.6E-3
8	GTPase activator activity (MF)	4.7E-2
10	phosphorus metabolic process (BP)	1.3E-3
10	cytosol (CC)	4.2E-3
10	protein kinase activity (MF)	7.8E-3
12	vesicle-mediated transport (BP)	4.1E-5
12	endosome (CC)	1.8E-3
12	phosphoprotein binding (MF)	1.5E-2
13	oxidation reduction (BP)	6.4E-3
13	integral to plasma membrane (CC)	2.4E-2
13	iron ion binding (MF)	1.3E-3
15	oxidation reduction (BP)	8.5E-4
15	endoplasmic reticulum lumen (CC)	1.3E-2
15	cofactor binding (MF)	8.3E-2

Table S.6: The most enriched BP, MF and CC terms for the large size modules (SSM) for each subject.

Subject	Term	P-value
MSM		
1	Bone development (BP)	1.9E-5
1	intrinsic to plasma membrane (CC)	8.2E-3
1	calcium- and calmodulin-responsive adenylate cyclase activity (MF)	2.0E-2
5	regulation of protein kinase activity (BP)	6.0E-4
5	chromatin (CC)	9.7E-3
5	protein serine/threonine kinase activity (MF)	9.7E-4
6	cytoskeleton organization (BP)	2.6E-4
6	Cytosol (CC)	7.6E-3
6	cytoskeletal protein binding (MF)	4.0E-3
7	cell proliferation (BP)	3.3E-3
7	anchoring junction (CC)	7.3E-3
7	actin binding (MF)	5.5E-3
8	positive regulation of transport (BP)	3.3E-3
8	nuclear lumen (CC)	2.3E-6
8	calcium ion binding (MF)	9.6E-3
10	response to hypoxia (BP)	3.0E-4
10	integral to plasma membrane (CC)	1.8E-4
10	extracellular matrix structural constituent (MF)	1.8E-2
12	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (BP)	4.4E-3
12	nucleoplasm (CC)	3.0E-5
12	GTP binding (MF)	1.2E-2
13	hormone metabolic process (BP)	5.7E-4
13	plasma membrane part (CC)	1.5E-3
13	phosphatase activity (MF)	2.2E-4
15	M phase (BP)	5.6E-3
15	endomembrane system (CC)	1.3E-2
15	protein kinase binding (MF)	1.6E-2

Table S.7: The most enriched BP, MF and CC terms for the large size modules (MSM) for each subject.

Subject	70 vs. 75	75 vs. 80	70 vs. 80
MCL			
1	0.7874	0.7514	0.6052
5	0.7322	0.5903	0.4256
6	0.8114	0.8041	0.6662
7	0.8855	0.7800	0.7270
8	0.8643	0.8574	0.7451
10	0.7561	0.8626	0.6636
12	0.8789	0.6856	0.6205
13	0.6135	0.8541	0.5351
15	0.8215	0.8085	0.6773
IHC			
1	0.5920	0.6231	0.5150
5	0.6627	0.4918	0.4477
6	0.6163	0.6819	0.4880
7	0.7128	0.6556	0.6271
8	0.6652	0.7294	0.6423
10	0.8390	0.7580	0.6869
12	0.6242	0.6619	0.5791
13	0.6806	0.5605	0.5491
15	0.6562	0.6672	0.5876
IPC			
1	0.5389	0.4503	0.4732
5	0.5545	0.5233	0.3719
6	0.6855	0.7032	0.6027
7	0.7521	0.6893	0.7122
8	0.6564	0.6202	0.5770
10	0.7914	0.7720	0.7531
12	0.6273	0.6190	0.5649
13	0.6607	0.6309	0.6316
15	0.6044	0.6408	0.5570

Table S.8: Sensitivity of Clustering Methods to either 5% or 10% change in the parameter α: the adjusted rand index for $\alpha = 70$ vs. $\alpha = 75$, $\alpha = 75$ vs. $\alpha = 80$, and $\alpha = 70$ vs. $\alpha = 80$ for each of the clustering methods considered.