Supporting Information

Mo et al. 10.1073/pnas.1208949110

SI Methods

Monte Carlo Newton–Raphson Algorithm. In this section, we discuss a Monte Carlo Newton–Raphson algorithm for maximizing a penalized log-likelihood under the iCluster+ model. The joint log-likelihood of (x_{ijt}, z_i) can be written as

$$\ell\left(x_{ijt}, \mathbf{z}_i; \alpha_{jt}, \boldsymbol{\beta}_{jt}\right) = \sum_{i=1}^{n}\sum_{t=1}^{m}\sum_{j=1}^{p_t}\left\{\log f\left(x_{ijt}|\mathbf{z}_i, \alpha_{jt}, \boldsymbol{\beta}_{jt}\right) + \log f(\mathbf{z}_i)\right\},$$

where the summation is due to the conditional independence assumption of x_{ijt} given $\mathbf{z}_i$. Here, $f(\mathbf{z}_i)$ is the density function of the standard multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_k)$, and the conditional density $f(x_{ijt}|\mathbf{z}_i, \alpha_{jt}, \boldsymbol{\beta}_{jt})$ has the form of normal, Bernoulli, multinomial, or Poisson density function depending on the type of genomic variable. Specifically,

$$f\left(x_{ijt}|\mathbf{z}_i, \alpha_{jt}, \boldsymbol{\beta}_{jt}\right) \propto \begin{cases} \sigma_{jt}^{-1}\exp\left(-\left(x_{ijt}-\mu_{jt}\right)^2/\left(2\sigma_{jt}^2\right)\right) \\ \left(\exp\left(\eta_{jt}\right)\right)^{x_{ijt}}\left(1+\exp\left(\eta_{jt}\right)\right)^{-1} \\ \prod_{c=1}^{C}\left(\exp\left(\phi_{jct}\right)/\sum_{\ell=1}^{C}\exp\left(\phi_{jlt}\right)\right)^{I\left(x_{ijt}=c\right)} \\ \xi_{jt}^{x_{ijt}}\exp\left(-\xi_{jt}\right) \end{cases},$$

if x_{ijt} is a continuous, binary, categorical, or count variable, respectively. In the above, $\mu_{jt} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i$, $\xi_{tj} = \exp(\alpha_{jt} + \boldsymbol{\beta}_{jt}z_i)$, $\eta_{jt} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i$, $\phi_{jct} = \alpha_{jct} + \boldsymbol{\beta}_{jct}\mathbf{z}_i$, and $I(x_{ijt} = c) = 0$ if $x_{ijt} = c$, and 0 otherwise.

To identify genomic variables that make important contribution to the latent variables, we apply the L_1-norm penalty (1) and consider the following penalized likelihood estimation:

$$\max_{\alpha_{jt}, \boldsymbol{\beta}_{jt}} \ell\left(x_{ijt}, \mathbf{z}_i; \alpha_{jt}, \boldsymbol{\beta}_{jt}\right) - \sum_{t=1}^{m}\sum_{j=1}^{p_t}\lambda_t\left\|\boldsymbol{\beta}_{jt}\right\|_1,$$

where $\|\boldsymbol{\beta}_{jt}\| = |\beta_{j1t}| + \ldots + |\beta_{jkt}|$ is the L_1-norm penalty and λ_ts are nonnegative tuning parameters. Due to the singularity of the L_1-norm penalty at $\beta_{jst} = 0$, some estimated β_{jst} will be exactly zero. If the entire vector β_{jt} is zero, then the corresponding genomic variable is effectively removed from the model. In addition to variable selection, the lasso-type procedures have also been shown to have good prediction ability in both finite samples and asymptotic situations (2, 3).

To estimate the parameters α_{jt} and $\boldsymbol{\beta}_{jt}$, we apply a modified Monte Carlo Newton–Raphson algorithm (4, 5). Conditional on $\mathbf{z}_i$, the penalized estimation can be divided into $\sum_{t=1}^{m}p_t$ separable optimization problem. Thus, for ease of presentation, we use generic notations by omitting index j and t for ℓ_{jt}, α_{jt}, $\boldsymbol{\beta}_{jt}$ and index t for λ_t in what follows. Specifically, we solve a constrained minimization problem of the following form going through $j = 1, \cdots, p_t, t = 1, \cdots, m$:

$$\min_{\alpha, \boldsymbol{\beta}} \ell + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\ell = \sum_{i=1}^{n}\log f(x_i|\mathbf{z}_i; \alpha, \boldsymbol{\beta})$. Let $\tilde{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ be the current parameter estimates.

Let $\eta_i = \alpha + \boldsymbol{\beta}\mathbf{z}_i$, $\tilde{\eta}_i = \tilde{\alpha} + \tilde{\boldsymbol{\beta}}\mathbf{z}_i$, $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_n)'$, $u_i = \partial\ell(\tilde{\boldsymbol{\eta}})/\partial\eta_i$, and $w_i = \partial\ell^2(\tilde{\boldsymbol{\eta}})/\partial\eta_i^2$. Following ref. 4, we form a quadratic approximation to the log-likelihood ℓ (Taylor expansion around current estimates) and consider the following optimization problem:

$$\min_{\alpha, \boldsymbol{\beta}} \sum_{i=1}^{n} w_i\left(y_i - \alpha - \sum_{s=1}^{k}\beta_s z_{is}\right)^2 + \lambda\sum_{s=1}^{k}|\beta_s|,$$

where $y_i = \tilde{\alpha} + \tilde{\boldsymbol{\beta}}\mathbf{z}_i - \sum_{i=1}^{n}u_i/w_i$. Following ref. 4, we apply the coordinate descent algorithm to solve the optimization problem. To be specific, we iterate among the following updates for α, β_s:

$$\hat{\beta}_s = \left(\frac{\sum_i w_i z_{is} y_i^{(s)}}{\sum_i w_i z_{is}^2} - \frac{\lambda}{2}\right)_+, s = 1, \ldots, k$$

$$\hat{\alpha} = \frac{\sum_i w_i y_i - \sum_{s=1}^{k}\hat{\beta}_s z_{is}}{\sum_i w_i},$$

where $y_i^{(s)} = y_i - \alpha - \sum_{l\neq s}\beta_l z_{il}$ is the partial residual for fitting β_s. Notice that, in this update, $y_i^{(s)}$ also depends on $\mathbf{z}_i$.

However, $\mathbf{z}_i$ is not observed in our model. A Markov chain simulation is particularly suitable for the latent variable formulation (6). The basic idea is to replace the expression in the parameter updates shown above by its expectation with respect to $\mathbf{z}_i$ given x_{ijt} by repeatedly sampling (typically we use 1,000 draws) the latent variable $\mathbf{z}_i$ from its joint posterior distribution

$$\mathbf{z}_i|\cdot \propto f(\mathbf{z}_i)\prod_{t=1}^{m}\prod_{j=1}^{p_t}f\left(x_{ijt}|\alpha_{jt}, \boldsymbol{\beta}_{jt}, \mathbf{z}_i\right), \quad i = 1, \cdots, n$$

using a random-walk Metropolis–Hasting algorithm (7, 8). We then calculate parameter updates by their sample averages over the repeated draws.

Clustering the Latent Variables. Sample clusters are assigned by the values of the latent variables. We use K-means clustering to divide the n samples into g clusters using k latent variables where $g = k + 1$; in the null model case where $k = 0$ (intercept only), this implies that all samples belong to one cluster. Following a general principle (9) for separating g clusters among the n data points, a rank-k approximation where $k \leq g - 1$ is sufficient.

Gene-Centric Identification of Concordant Copy Number and Expression Alterations. Somatic copy number alterations that characterize a particulate integrated cluster as identified by our method often span broad regions covering up to thousands of genes. To gain additional insights, we performed a gene-centric integration for each cluster for cancer gene identification. For each gene, we applied independent two-sample t tests on its copy number and on its mRNA expression between patient samples in cluster k vs. the rest. We then use Fisher's method to combine the P values as $-2\sum_{i=1}^{r}\log P_i$, which has a χ^2 distribution under the null with $2r$ (r is the number of independent tests being combined) degrees of freedom. A large χ^2 statistic provides strong evidence for concordant events (e.g., copy number-induced expression changes) and indicates candidate cancer genes that characterize a distinct molecular subgroup.

Data Processing. In the Cancer Cell Line Encyclopedia (CCLE) dataset, we integrated $m = 3$ data types: somatic mutation by

massively parallel sequencing (t = 1), chromosomal copy number by Affymetrix SNP Array 6.0 (t = 2), and gene expression by Affymetrix Human Genome U133 Plus 2.0 array (t = 3) measured in a total of $n = 729$ cancer cell lines with all three data types available. For a given data type t, it consists of p_t genomic variables, which can be annotated as genes, exons, genomic regions, CpG sites, coding or noncoding mRNA, and other genetic markers, depending on the data type. For somatic mutation data (t = 1), a gene by sample matrix of binary values (1, mutated; 0, not mutated) was generated for clustering, which included $p_1 =$ 1,670 genes that were mutated in ≥1% of the cell lines.

For the copy number data (t = 2), the circular binary segmentation (CBS) segmented data based on Affymetrix SNP Array 6.0 was used. Nonredundant copy number regions were first obtained by adapting a method that is described in ref. 10. Briefly, we first form genomic "neighborhoods" (regions) along a chromosome defined by consecutive positions with a maximum Euclidean distance (based on copy number log-ratio segmented values) between any adjacent two probes smaller than 0.01; this resulted in a total of $p_2 = 8{,}432$ copy number regions. Each region was then represented by its medoid signature, reducing the dimension of the data from close to 2 million genomic positions on SNP Array 6.0 to 8,432 unique copy number regions.

For the gene expression data (t = 3), we included $p_3 = 4{,}000$ most-variable genes based on their expression profile across the cell lines. The order of data types t = 1, 2, 3) is imposed in the algorithm flow for the purpose of specifying the corresponding type of generalized linear model (11). In the dataset of application in this study and in ref. 12, we directly modeled the CBS (13, 14) segmented copy number measures (logR values from array comparative genomic hybridization or genotyping arrays) to avoid any information loss by generating discrete calls. Therefore, copy number-segmented log-ratio values, gene expression, and methylation β-values after proper transformation and normalization were modeled using linear regression with normal errors in this study. However, the iCluster+ framework can easily accommodate discretized calls (gain, loss, expressed, unexpressed, methylated, unmethylated), depending on user preference, by simply switching to binomial or multinomial regression for these data types.

Similar data preprocessing procedures were applied to the Cancer Genome Atlas (TCGA) colorectal cancer (CRC) dataset with slight variation. For mutation data generated by whole-exome sequencing (15), a gene by sample matrix of binary values (1, mutated; 0, WT) was generated from a multiple alignment format data. Significantly mutated genes ($q \leq 0.1$), as identified by the MutSig algorithm (15), were included for clustering. The copy number region method was applied to segmented log-ratio data from Affymetrix SNP Array 6.0 and a data matrix of 6,587 copy number regions by 189 CRC samples was used as the input data. Gene expression was measured by RNA-sequencing platform. Transcript levels were quantified in reads per kilobase of exon model per million mapped reads (RPKM) (16). Our method uses Poisson regression for count variables. However, we found it was more effective to model the logarithm-transformed RPKM data with normal distribution with independent mean and variance parameters. Median absolute deviation was used to select the top 2,000 most-variable genes for clustering. For DNA methylation, the methylation β-values (on logit scale) from the HumanMethylation27 array platform was used. Median absolute deviation was used to select the top 2,000 most-variable CpG sites for clustering.

Computing Time. The tuning procedure for the CCLE dataset is computationally intensive because it involves 729 samples over 20 cancer types; it is a highly heterogeneous dataset, and we expect that the number of distinct clusters is large. Therefore, the model selection included a series of 19 k's ($k = 1$–19). For a dataset that involves a single cancer type (TCGA CRC), we typically run $k =$ 1–5, which requires much less computing time with the same computing capacity. By parallel processing using 30 cores on a 3.2 GHz Xeon Linux computing cluster, the model-tuning procedure finished in 5 d for the CCLE dataset, and in 16 h for the TCGA CRC data.

Stability. The computational complexity of our method that implements a modified Monte Carlo Newton–Raphson algorithm is high, and it is further compounded by the large sampling space of the lasso parameters. In our analysis procedure, we excluded the uniform-design sampled vectors of lasso parameter λ (rescaled to be between 0 and 1 in our method) that lead to models either too dense [close to the full model, including all genomic features ($\lambda \geq 0.95$)], or too sparse, including too few features ($\lambda \leq 0.05$)] in any data dimension to avoid the extreme cases. This strategy is in part to reduce computational complexity, and by doing so, the variability in the outcome (both the selected number of clusters and the cluster membership assignment) is kept sufficiently small.

To assess variation in the selected number of clusters, we repeated the same run for the CCLE dataset five times for $k = 1 -$ 19 using our setup by filtering out the extreme cases. Fig. S8 shows the average of the five repeats with error bars (some error bars do not show on the graph because they are close to zero), suggesting the variation is small.

Due to the stochastic nature of Markov chain Monte Carlo sampling and the iterative Monte Carlo Newton–Raphson procedure, each independent run of iCluster+ may lead to a slightly different solution. To assess the degree of such variation in cluster assignment, we repeated the 12-cluster run 10 times (given selected lasso parameter values obtained from the tuning process). The average Rand index (17) between any random pair of repeats that measures the agreement for the two partitions is 93% (±1%). We repeated this for the 14-cluster run and obtained similarly high agreement (93% ± 1%). We also observed that part of this small variation was due to the K-means step in our algorithm, which assigns sample cluster membership on the basis of the latent variables (a matrix of $K \times n$).

The silhouette statistic (18) measures the strength of cluster membership assignment for each individual sample. A sample that clusters tightly to its corresponding cluster will have a high silhouette width s_i. A sample that clusters loosely will have a low silhouette width. Fig. S9 shows the silhouette profile for the 729 cancer cell line samples. The average silhouette indicates the strength of each cluster. The silhouette profile can be used to selected "core" samples most representative of the clusters and thus most reproducible.

Note that the choice of 12-cluster vs. other possible choices (e.g., 14-cluster) is not mutually exclusive. Table S1 compares the cluster assignment between 12-cluster and 14-cluster assignments. It is clear that the 14-cluster solution primarily involves further subdivisions and regrouping among clusters that have relatively low average silhouette measures (Fig. S9) and heterogeneity in tissue composition, which include 7, 8, and 12. Cluster 6 is subdivided into large intestine vs. endometrial because of biological difference.

Subsampling. We conducted a subsampling experiment using the TCGA CRC dataset to illustrate the stability of our integrative clustering algorithm. Specifically, we generated 100 random subsamples (without replacement) of the original data with a sampling ratio of 0.8 to preserve the general structure (19). We ran iCluster+ for $k = 1 - 5$ (number of latent variables) on each of the subsampled datasets using the lasso parameter tuning procedure as described previously. For a randomly chosen pair of subsampled datasets, a measure of agreement between the clustering is computed using the adjusted Rand index (17, 20). When two partitions agree perfectly, the adjusted Rand index is 1. When the agreement is random (the degree of agreement

equals the expected value under a model of randomness from the generalized hypergeometric distribution), the adjusted Rand index is zero. We exhausted all possible pairs, and the box plots of adjusted Rand indices computed for the pairs for $g = 2 - 6$ (number of clusters) are shown in Fig. S10. The mean adjusted Rand index averaged over all possible pairs for $g = 2 - 6$ is 0.92 0.90, 0.68, 0.62, and 0.54. For three-cluster solutions, cluster assignments of any random subsamples of the colorectal cancer dataset are on average 90% concordant, suggesting stability of the results. Based on the stability plot, three-cluster is the demarcating point, which is consistent with our proposed criteria based on percent explained variation (Fig. S1).

1. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58(1):267–288.
2. Bickel P, Ritov Y, Tsybakov A (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann Stat* 37(4):1705–1732.
3. Greenshtein E, Ritov Y (2004) Linear predictor-selection and the virtue of over-parametrization. *Bernoulli* 10(6):971–988.
4. Friedman J, Hastie T, Tibshirani R (2010) Regularized paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22.
5. McCulloch C (1997) Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc* 92(437):162–170.
6. Tanner MA, Wong WH (2010) From EM to data augmentation: The emergence of MCMC Bayesian computation in the 1980s. *Stat Sci* 25(4):506–516.
7. Robert CP, Casella G (2004) *Monte Carlo Statistical Methods* (Springer, New York).
8. Liu J (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York).
9. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Learning: Data Mining, Inference, and Prediction* (Springer, New York).
10. van de Wiel MA, Wieringen WN (2007) CGHregions: Dimension reduction for array CGH data with minimal information loss. *Cancer Inform* 3:55–63.
11. McCullagh P, Nelder J (1994) *Generalized Linear Models* (Chapman & Hall, London).
12. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25(22):2906–2912.
13. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557–572.
14. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23(6):657–663.
15. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.
16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
17. Rand WM (1971) Objective criteria for the evaluation of clustering method. *J Am Stat Assoc* 66(336):846–885.
18. Kaufman L, Rousseeuw P (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York).
19. Ben-Hur A, Elisseeff A, Guyon I (2002) *A stability based method for discovering structure in clustered data. Pac Symp Biocomput* 7:6–17.
20. Hubert L, Arabie P (1985) Comparing partitions. *J Classification* 2:193–218.

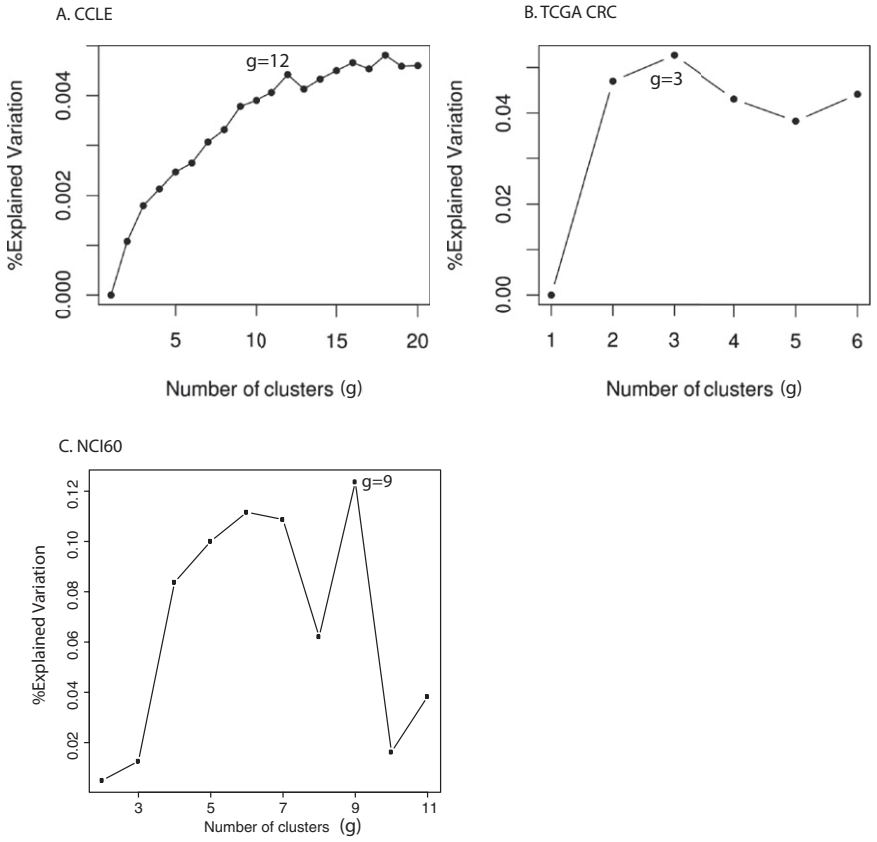

Fig. S1. Model selection in (*A*) CCLE dataset, (*B*) TCGA CRC dataset, and (*C*) NCI60 dataset. The optimal *g* (number of sample clusters) is chosen as the transition point beyond which the size of the increase in percent explained variation is diminishing. The jaggedness of the curve observed in the NCI60 dataset may reflect its smaller sample size compared with the other two datasets.

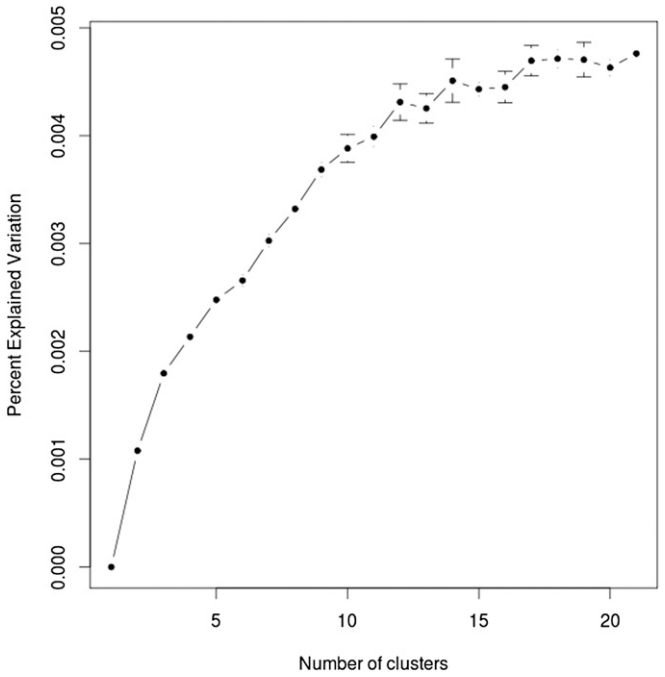

Fig. S2. Repeated run of the CCLE dataset.

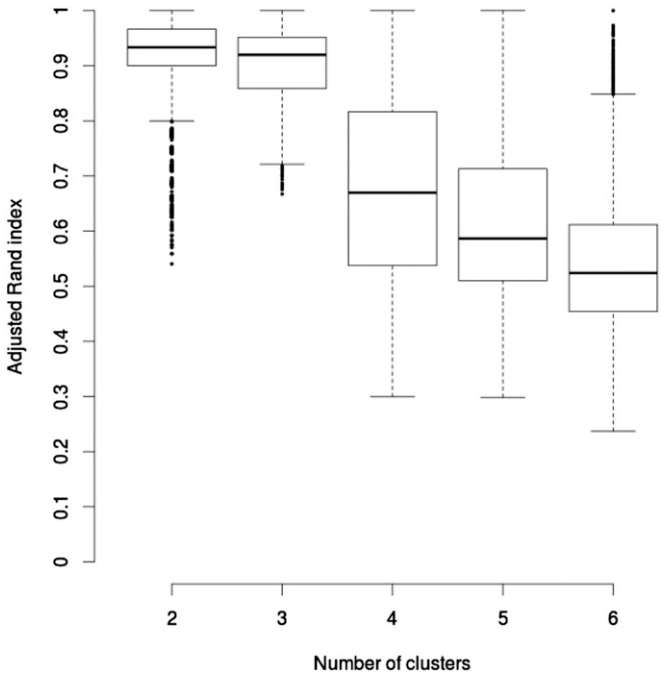

Fig. S3. Subsampling of the TCGA CRC dataset.

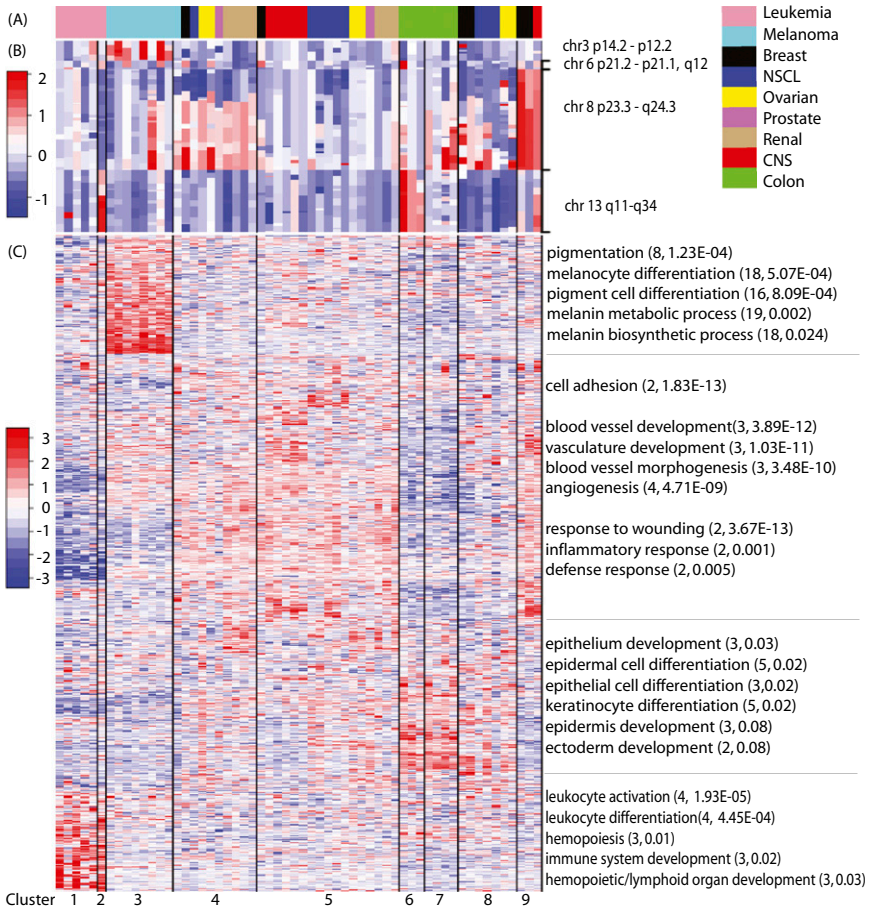

Fig. S4. Integrative clustering found nine cell line clusters in the NCI60 dataset (DNA copy number and gene expression data). (*A*) The 58 tumor cell lines representing nine different tumor types, which are arranged by their cluster ID. (*B* and *C*) Identified genomic regions and genes that contribute to cell-line clustering. As shown in the figure, the cluster assignment was determined by the joint patterns of the copy number and gene expression. For example, five of six leukemia cell lines had similar copy number and gene expression patterns, and were thus assigned to the same cluster. The remaining leukemia cell line had a different copy number pattern (amplification on chr3 q11–q34), and was thus assigned to another cluster. Similarly, the colon cell lines were divided to two subclusters because of the different copy number patterns. The top enriched biological processes are shown for each subset of genes on *C*. The values in parentheses are the fold enrichment and adjusted *P* value.

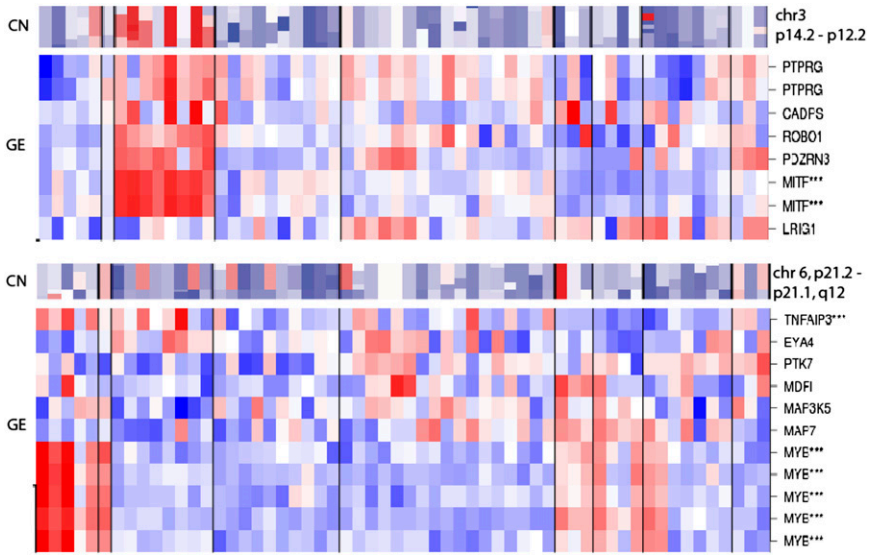

Fig. S5. Selective amplification and overexpression of *MITF* in melanoma cell lines in the NCI60 dataset (*Upper*). Leukemia cell lines show overexpression of MYB (*Lower*). CN, copy number; GE, gene expression.

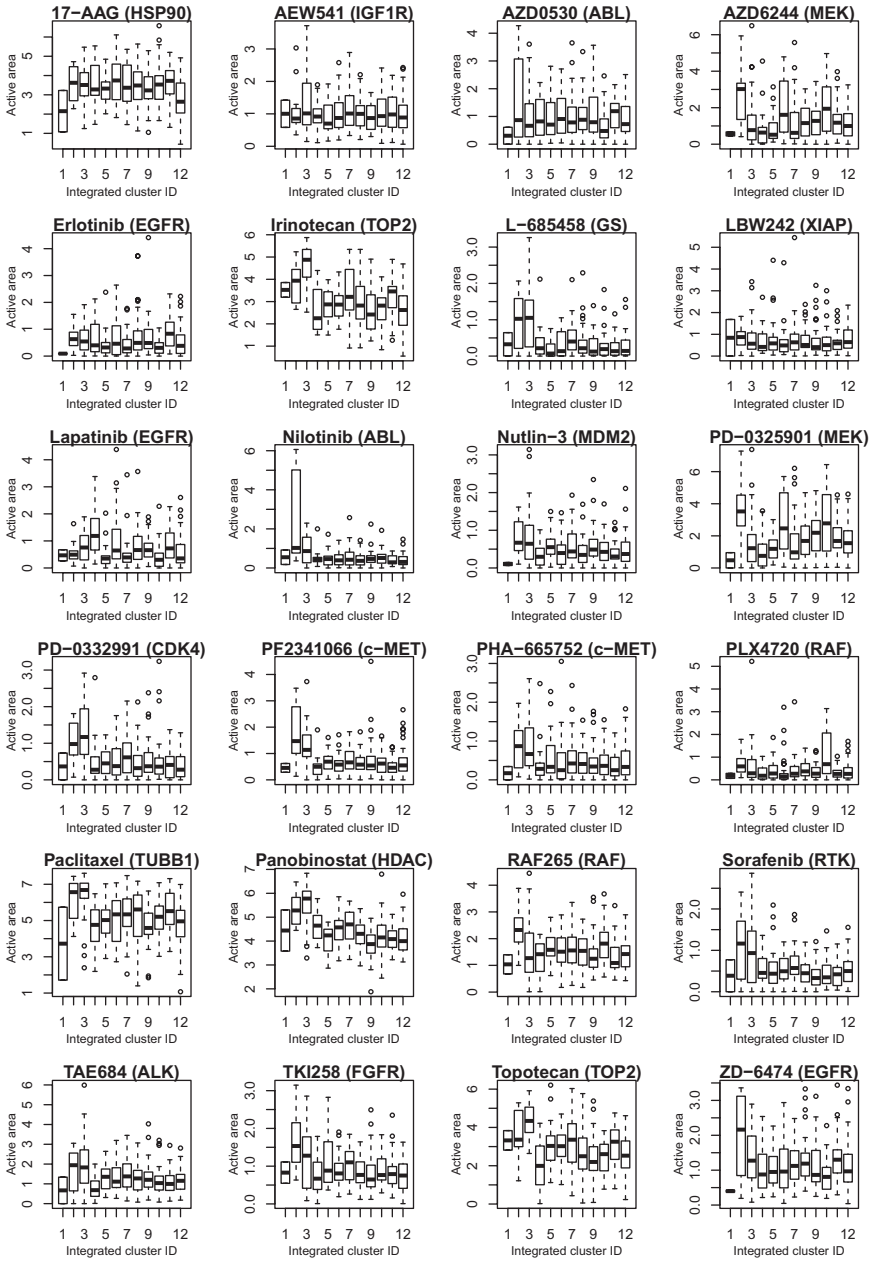

Fig. S6. The integrated cell line clusters show different drug sensitivity. The integrated cluster ID is the same as those assigned in Fig. 1 of the main text.

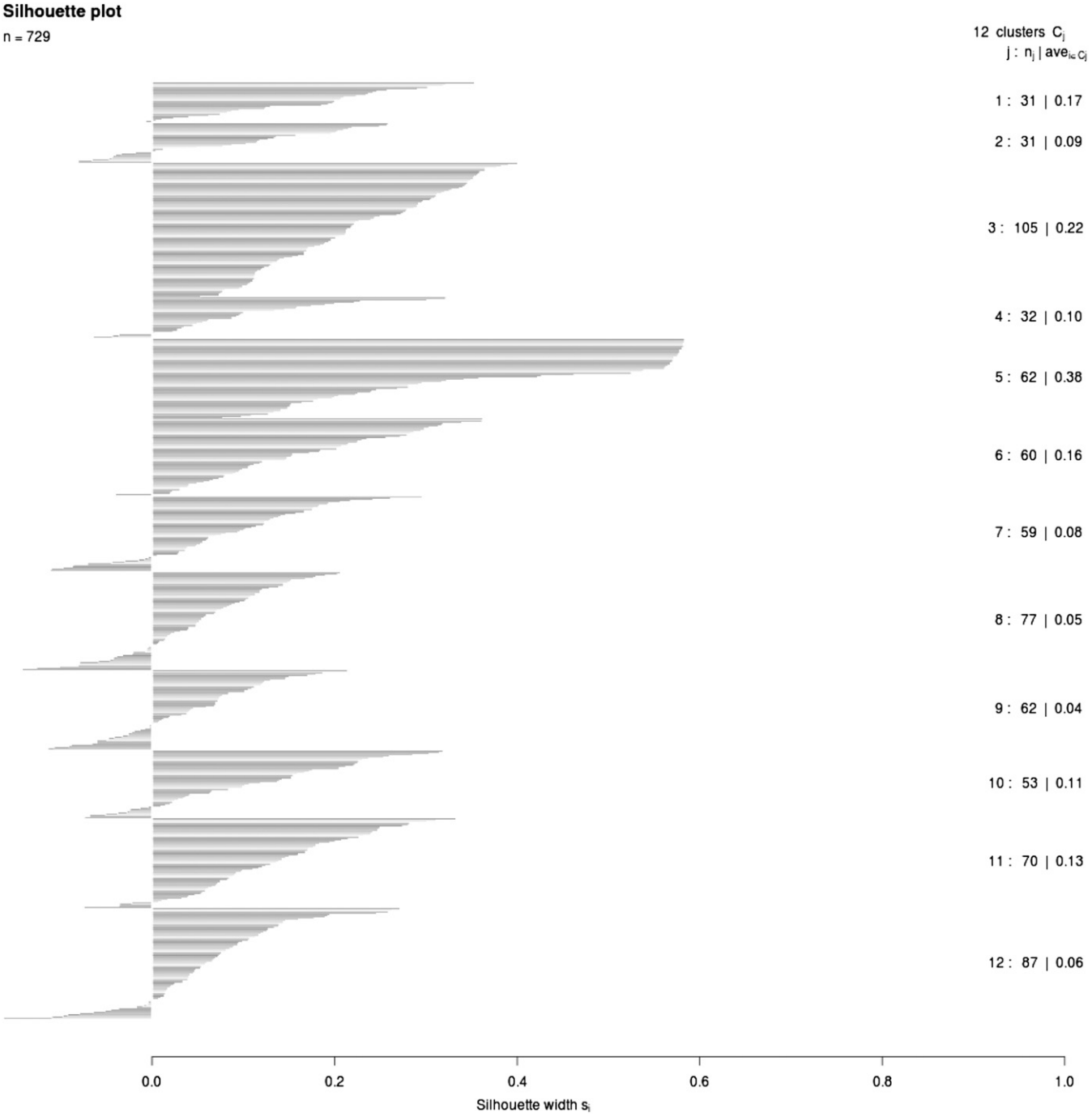

Fig. S7. Silhouette profile of the CCLE samples.

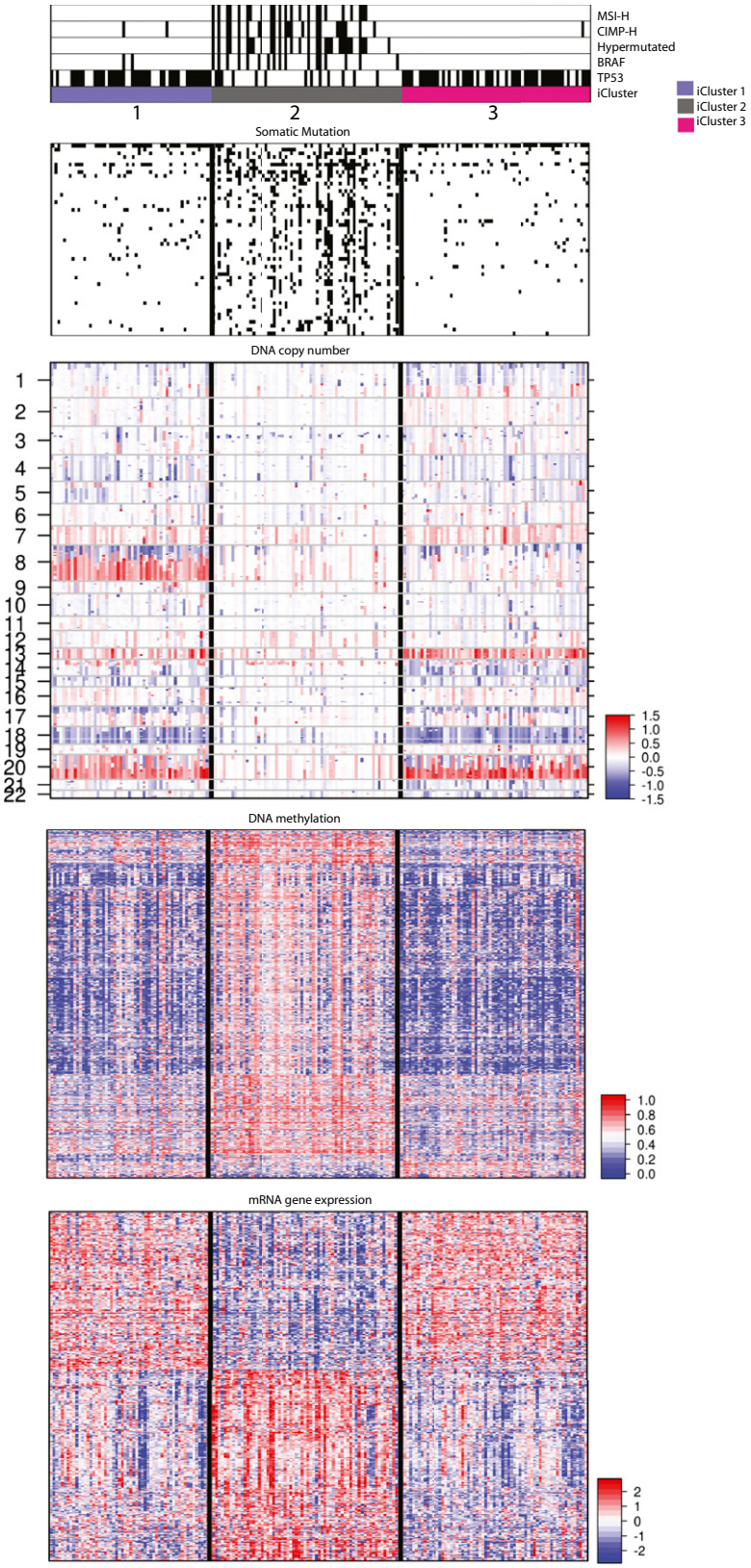

Fig. S8. Three-cluster solution among 189 TCGA colorectal cancer samples. Heat map display of lasso-selected cluster-discriminant features. Rows are features and columns are tumor samples. The first panel shows genes that are mutated (black) or not mutated (white) in each cluster; the second shows genomic regions amplified (red) or deleted (blue); the third shows genes hypermethylated (red) or hypomethylated (blue); and the fourth shows genes overexpressed (red) or underexpressed (blue) in each cluster.

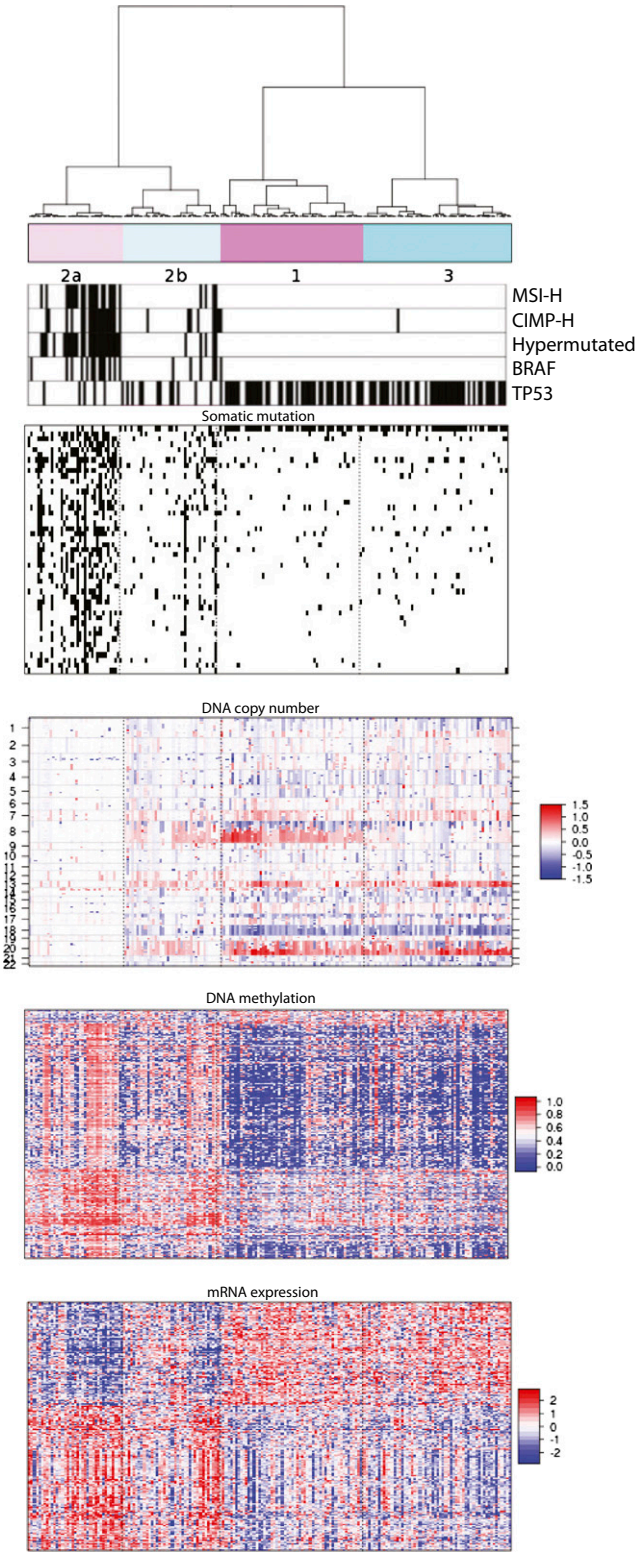

Fig. S9. Four-cluster solution among 189 TCGA colorectal cancer samples. (*Top*) Annotations of microsatellite instability (MSI-H), CpG-island methylation phenotype (CIMP-H), mutator phenotype associated with deficient DNA damage repair (hypermutated), and BRAF and TP53 mutation status. The second panel shows genes that are mutated (black) or not mutated (white) in each cluster; the third shows genomic regions amplified (red) or deleted (blue); the fourth shows genes hypermethylated (red) or hypomethylated (blue), and the fifth shows genes overexpressed (red) or underexpressed (blue) in each cluster.

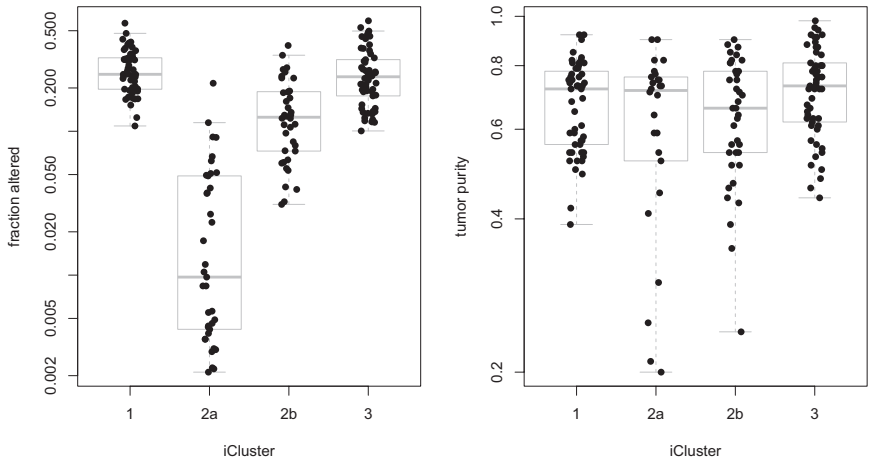

Fig. S10. Integrated colorectal cancer subtypes show great variation in chromosomal instability as measured by the fraction of the genome altered (*Left*). There are no significant differences in tumor purity as measured by the ABSOLUTE algorithm (1).

1. Carter SL, et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30(5):413–421.

Table S1. Comparison of 12-cluster and 14-cluster assignments of the CCLE samples

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	29	0	0	0	0	0	0	0	0	0	0	0	1	1
2	0	28	0	0	0	0	0	2	1	0	0	0	0	0
3	0	10	89	0	0	0	1	0	0	0	0	0	3	2
4	0	0	0	24	1	0	5	0	1	1	0	0	0	0
5	0	0	0	0	51	0	1	1	0	1	0	0	0	8
6	0	0	0	1	0	26	27	0	3	1	0	0	0	2
7	2	3	0	0	1	0	33	15	0	0	1	1	1	2
8	0	0	0	5	0	1	2	26	34	0	5	0	3	1
9	0	0	0	1	1	4	1	0	3	39	0	3	6	4
10	0	0	0	0	2	0	0	0	0	0	50	0	1	0
11	1	0	0	0	0	3	0	1	2	0	0	53	8	2
12	1	0	0	0	4	3	5	4	2	1	8	1	43	15

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)