BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 22 2009, pages 2906–2912
doi:10.1093/bioinformatics/btp543

Genome analysis

# Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis

Ronglai Shen[1,*], Adam B. Olshen[2] and Marc Ladanyi[3]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, [2]Department of Epidemiology and Biostatistics and Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA and [3]Department of Pathology and Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Received June 22, 2009; revised on August 25, 2009; accepted on September 9, 2009

Advance Access publication September 16, 2009

Associate Editor: Alex Bateman

## ABSTRACT

**Motivation:** The molecular complexity of a tumor manifests itself at the genomic, epigenomic, transcriptomic and proteomic levels. Genomic profiling at these multiple levels should allow an integrated characterization of tumor etiology. However, there is a shortage of effective statistical and bioinformatic tools for truly integrative data analysis. The standard approach to integrative clustering is separate clustering followed by manual integration. A more statistically powerful approach would incorporate all data types simultaneously and generate a single integrated cluster assignment.

**Methods:** We developed a joint latent variable model for integrative clustering. We call the resulting methodology iCluster. iCluster incorporates flexible modeling of the associations between different data types and the variance–covariance structure within data types in a single framework, while simultaneously reducing the dimensionality of the datasets. Likelihood-based inference is obtained through the Expectation–Maximization algorithm.

**Results:** We demonstrate the iCluster algorithm using two examples of joint analysis of copy number and gene expression data, one from breast cancer and one from lung cancer. In both cases, we identified subtypes characterized by concordant DNA copy number changes and gene expression as well as unique profiles specific to one or the other in a completely automated fashion. In addition, the algorithm discovers potentially novel subtypes by combining weak yet consistent alteration patterns across data types.

**Availability:** R code to implement iCluster can be downloaded at http://www.mskcc.org/mskcc/html/85130.cfm.

**Contact:** shenr@mskcc.org

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

In recent years genomic profiling of multiple data types in the same set of tumors has gained prominence. In a breast cancer study relating DNA copy number to gene expression, (Pollack *et al.*, 2002) estimated that 62% of highly amplified genes demonstrate moderately or highly elevated gene expression, and that DNA copy number aberrations account for ∼10–12% of the global gene expression changes at the messenger RNA (mRNA) level. Hyman *et al.* (2002) observed similar results in breast cancer cell lines. MicroRNAs, which are small non-coding RNAs that repress gene expression by binding mRNA target transcripts, provide another mechanism of gene expression regulation. Over 1000 microRNAs are predicted to exist in humans, and they are estimated to target one-third of all genes in the genome (Lewis *et al.*, 2005). The NCI/NHGRI-sponsored Cancer Genome Atlas (TCGA) pilot project is a coordinated effort to explore the entire spectrum of genomic alternations in human cancer to obtain an integrated view of such interplays. The group recently published an interim analysis of DNA sequencing, copy number, gene expression and DNA methylation data in a large set of glioblastomas (TCGA, 2008).

In this study, we will refer to any genomic dataset involving more than one data type measured in the same set of tumors as multiple genomic platform (MGP) data. Identifying tumor subtypes by simultaneously analyzing MGP data is a new problem. The current approach to subtype discovery across multiple types is to separately cluster each type and then to manually integrate the results. An ideal integrative clustering approach would allow joint inference from MGP data and generate a single integrated cluster assignment through simultaneously capturing patterns of genomic alterations that are: (i) consistent across multiple data types; (ii) specific to individual data types; or (iii) weak yet consistent across datasets that would emerge only as a result of combining levels of evidence. Therefore, the goal of this study is to develop such an integrative framework for tumor subtype discovery.

There are two major challenges to the development of a truly integrative approach. First, to capture both concordant and unique alterations across data types, separate modeling of the covariance between data types and the variance–covariance structure within data types is needed. Most of the existing deterministic clustering methods cannot be easily adapted in this way. For example, Qin (2008) performed a hierarchical clustering of the correlation matrix between gene expression and microRNA data. Similarly,

[*]To whom correspondence should be addressed.

Lee *et al.* (2008) applied a biclustering algorithm on the correlation matrix to integrate DNA copy number and gene expression data. In both the cases, the goal was to identify correlated patterns of change given the two data types. While identifying correlated patterns is sufficient for studying the regulatory mechanism of gene expression via copy number changes or epi-genomic modifications, it is not suitable for integrative tumor subtype analysis where both concordant and unique alteration patterns may be important in defining disease subgroups. The importance of capturing both concordant and unique alterations across data types will be demonstrated in our data examples. In addition, properly separating covariance between data types and variance within data types facilitates probabilistic inference for data integration.

Second, dimension reduction is a key to the feasibility and performance of integrative clustering approaches. Methods that rely on pairwise correlation matrices are computationally prohibitive with today's high-resolution arrays. Dimension reduction techniques such as principal component analysis (PCA; Alter *et al.*, 2000; Holter *et al.*, 2000) and non-negative matrix factorization (NMF; Brunet *et al.*, 2004) have been proposed for use in combination with clustering algorithms. These methods work well for a single data type. However, simultaneous dimension reduction of multiple correlated datasets is beyond the capabilities of these algorithms.

Tipping and Bishop (1999) showed that the principal components can be computed through maximum-likelihood estimation of parameters under a Gaussian latent variable model. In their framework, the correlations among variables are modeled through the latent variables of a substantially lower dimension space, while an additional error term is added to model the residual variance. Using the connection between PCA and latent variable models as a building block, we propose a novel integrative clustering method called iCluster that is based on a joint latent variable model. The main idea behind iCluster is that tumor subtypes can be modeled as unobserved (latent) variables that can be simultaneously estimated from copy number data, mRNA expression data and other available data types. It is a conceptually simple and computationally feasible model that allows simultaneous inference on any number and type of genomic datasets. Furthermore, we develop a sparse solution of the iCluster model through optimizing a penalized complete-data log-likelihood using the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977). A lasso-type regularization method (Tibshirani, 1996) is used in the penalized complete-data likelihood. The resulting model continuously shrink the coefficients for non-informative genes toward zero, and thus leading to reduced variance and better clustering performance. Moreover, a variable selection strategy emerges (since the coefficients for some of the genes will be exactly zero under lasso penalty), which helps to pinpoint important genes.

The article is organized as follows. In Section 2.1, we discuss the $K$-means clustering algorithm and a global optimal solution for the $K$-means problem through PCA. In Section 2.2, we formulate the $K$-means problem as a Gaussian latent variable model and show the maximum likelihood-based solution and its connection with the PCA solution. Then in Section 2.3, we extend the latent variable model to allow multiple data types for the purpose of integrative clustering. A sparse solution is derived in Section 2.4. We demonstrate the method using two datasets from published studies in Section 3.

## 2 METHODS

### 2.1 Eigengene $K$-means algorithm

We start the investigation with the $K$-means clustering algorithm. In standard $K$-means, given an initial set of $K$ cluster assignments and the corresponding cluster centers, the procedure iteratively moves the centers to minimize the total within-cluster variance. For purposes of exposition, we assume that the data are gene expression, although they could be any type of genomic measurements. Let $\mathbf{X}$ denote the mean-centered expression data of dimension $p \times n$ with rows being genes and columns being samples. Given a partition $C$ of the column space of $\mathbf{X}$ and the corresponding cluster mean vectors $\{\mathbf{m}_1, \cdots, \mathbf{m}_K\}$, the sample vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are assigned cluster membership such that the sum of within-cluster squared distances is minimized:

$$\min \sum_{k=1}^{K} \sum_{C(i)=k} \|\mathbf{x}_i - \mathbf{m}_k\|^2. \tag{1}$$

The cluster centers are subsequently recalculated successively based on the current partition. The algorithm iterates until the assignments do not change.

One of the main criticisms of $K$-means clustering is that the algorithm is sensitive to the choice of starting points; it can iterate to local minima rather than the global maximum. However, it has been recently shown that a better optimization scheme for $K$-means arises through PCA (Zha *et al.*, 2001). To see this, let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)'$ with the $k$-th row being the indicator vector of cluster $k$ normalized to have unit length:

$$\mathbf{z}_k' = (0, \ldots, 0, \underbrace{\frac{1}{\sqrt{n_k}}, \ldots \frac{1}{\sqrt{n_k}}}_{n_k}, 0, \ldots, 0), \tag{2}$$

where $n_k$ is the number of samples in cluster $k$ and $\sum_{k=1}^{K} n_k = n$. The objective is to obtain an optimal solution of the cluster assignment matrix $\mathbf{Z}$ such that the within-cluster variance is minimized. Let $\mathbf{X}'\mathbf{X}$ be the Gram matrix of the samples. The $K$-means loss function in (1) can be expressed as

$$\text{trace}(\mathbf{X}'\mathbf{X}) - \text{trace}(\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{Z}'),$$

which is the total variance minus the between-cluster variance. Since the total variance is a constant given the data, it follows that minimizing (1) is equivalent to maximizing the between-cluster variance

$$\max_{\mathbf{Z}\mathbf{Z}'=I_K} \text{trace}(\mathbf{Z}\mathbf{X}'\mathbf{X}\mathbf{Z}'). \tag{3}$$

Now consider a continuous $\mathbf{Z}^*$ that satisfies all the conditions of $\mathbf{Z}$ except for the discrete structure. In other words, $\mathbf{z}_k^*$ is no longer restricted to take values of either zero or one (scaled by the square-root of the cluster size). Then the above is equivalent to the eigenvalue decomposition of $\mathbf{S}$. Therefore, a closed-form solution of (3) is $\hat{\mathbf{Z}}^* = \mathbf{E}$, where $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_K)'$ are the eigenvectors corresponding to the $K$ largest eigenvalues from the eigenvalue decomposition of $\mathbf{S}$. As a result, $\hat{\mathbf{Z}}^*$ is the solution to the relaxed trace maximization problem of (3). A later publication by Ding and He (2004) pointed out the redundancy in $\mathbf{Z}$ such that the $K$-means solution can be defined by the first $K-1$ eigenvectors. The eigenvectors lie in a low-dimensional latent space where the original data are projected onto each of the first $K-1$ principal directions such that the total variance is maximized. As a result, any distinct subgroup structures will be automatically embedded in this set of orthogonal directional vectors.

Note that although the continuous parameterization of $\mathbf{Z}$ causes some loss in interpretability of the cluster indicator matrix, it is a necessary condition for the closed-form optimal solution to the $K$-means problem. The discrete structure in $\mathbf{Z}$ and its interpretability can be easily restored by a simple mapping by a pivoted QR decomposition or a standard $K$-means algorithm invoked on $\mathbf{Z}^*$. Zha *et al.* (2001) found similar performance by the two methods for recovering the class indicator matrix. For simplicity, in what comes later we use $K$-means for this final step. Finally, since we are in the genomic data context, we refer to the algorithm described in this section as eigengene $K$-means, and it yields the eigengene solution $\hat{\mathbf{Z}}^E$.

## 2.2 A Gaussian latent variable model representation

Now we consider a Gaussian latent variable model representation of the eigengene $K$-means clustering:

$$\mathbf{X}=\mathbf{W}\mathbf{Z}+\boldsymbol{\varepsilon}, \tag{4}$$

where $\mathbf{X}$ is the mean-centered expression matrix of dimension $p\times n$ (no intercept), $\mathbf{Z}=(\mathbf{z}_1,\dots,\mathbf{z}_{K-1})'$ is the cluster indicator matrix of dimension $(K-1)\times n$ as defined in Section 2.1, $\mathbf{W}$ is the coefficient matrix of dimension $p\times(K-1)$, and $\boldsymbol{\varepsilon}=(\varepsilon_1,\dots,\varepsilon_p)'$ is a set of independent error terms with zero mean and a diagonal covariance matrix $\mathbf{Cov}(\boldsymbol{\varepsilon})=\boldsymbol{\Psi}$ where $\boldsymbol{\Psi}=\mathrm{diag}(\psi_1,\dots,\psi_p)$. The fundamental concepts of model (4) are: (i) it differs from a regular regression model in that $(\mathbf{z}_1,\dots,\mathbf{z}_{K-1})$ are treated as latent variables representing the true molecular tumor subtypes to be discovered; and (ii) in dimension reduction terms, $\mathbf{W}$ is the projection matrix that maps the gene×array space of the original data matrix $\mathbf{X}$ onto an eigengene×eigenarray subspace spanned by the first $K-1$ principal directions.

Now consider a continuous parameterization $\mathbf{Z}^*$ of $\mathbf{Z}$ and make the additional assumption that $\mathbf{Z}^*\sim N(0,\mathbf{I})$ and $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\boldsymbol{\Psi})$. Then a likelihood-based solution to the $K$-means problem is available through model (4). The inference will be based on the posterior mean of $\mathbf{Z}^*$ given the data. Tipping and Bishop (1999) established a connection between the Gaussian latent variable model and PCA under an *isotropic* error model with a scalar covariance matrix $\boldsymbol{\Psi}=\sigma^2\mathbf{I}$. Then it was shown that by plugging in the maximum likelihood estimate of $\mathbf{W}$ and $\sigma^2$, the posterior mean is represented through the principal axes of the data vectors. In particular,

$$\hat{E}[\mathbf{Z}^*|\mathbf{X}]=(\boldsymbol{\Lambda}-\sigma^2\mathbf{I})^{1/2}\boldsymbol{\Lambda}^{-1/2}\mathbf{E}, \tag{5}$$

where $\mathbf{E}$ denotes the eigengene matrix as defined before. It is clear that the posterior mean yields the same eigengene $K$-means solution $\hat{\mathbf{Z}}^*=\mathbf{E}$ if the residual error $\sigma^2$ is assigned the value zero. However, the subspace $\hat{E}[\mathbf{Z}^*|\mathbf{X}]$ obtained through maximum likelihood approach will *not* generally correspond to the principal subspace obtained through PCA. Such a link occurs only under the isotropic error model.

The motivation for formulating the $K$-means problem as a Gaussian latent variable model is 2-fold: (i) it provides a probabilistic inference framework; and (ii) the latent variable model has a natural extension to multiple data types. In the next section, we propose a joint latent variable model for integrative clustering.

## 2.3 iCluster: a joint latent variable model-based clustering method

The basic concept of iCluster is to jointly estimate $\mathbf{Z}=(\mathbf{z}_1,\dots,\mathbf{z}_{K-1})'$, the latent tumor subtypes, from, say, DNA copy number data (denoted by $\mathbf{X}_1$, a matrix of dimension $p_1\times n$), DNA methylation data (denoted by $\mathbf{X}_2$, a matrix of dimension $p_2\times n$), mRNA expression data (denoted by $\mathbf{X}_3$, a matrix of dimension $p_3\times n$) and so forth (Fig. 1). The mathematical form of the integrative model is

$$\begin{aligned}\mathbf{X}_1&=\mathbf{W}_1\mathbf{Z}+\boldsymbol{\varepsilon}_1\\ \mathbf{X}_2&=\mathbf{W}_2\mathbf{Z}+\boldsymbol{\varepsilon}_2\\ &\vdots\\ \mathbf{X}_m&=\mathbf{W}_m\mathbf{Z}+\boldsymbol{\varepsilon}_m,\end{aligned} \tag{6}$$

where $m$ is the number of genomic data types available for the same set of samples. We assume each dataset is row centered and therefore intercept terms are not included in the models.

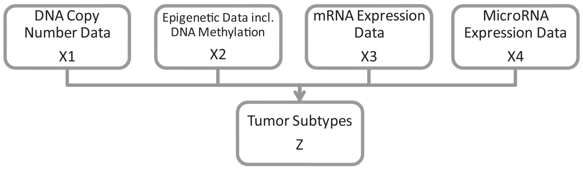

**Fig. 1.** The integrative model. The concept is to formulate the tumor subtypes as the joint latent variable $\mathbf{Z}$ that needs to be simultaneously estimated from multiple genomic data types measured on the same set of tumors.

In (6), $\mathbf{Z}$ is the latent component that connects the $m$-set of models, inducing dependencies across the data types measured on the same set of tumors. On the other hand, the independent error terms $(\boldsymbol{\varepsilon}_1,\dots,\boldsymbol{\varepsilon}_m)$, in which each has mean zero and diagonal covariance matrix $\boldsymbol{\Psi}_i$, represent the remaining variances unique to each data type after accounting for the correlation across data types. Lastly, $(\mathbf{W}_1,\dots,\mathbf{W}_m)$ denote the coefficient matrices. In dimension reduction terms, they embed a simultaneous data projection mechanism that maximizes the correlation between data types.

To derive a likelihood-based solution of (6), we use a latent continuous parameterization that further assumes $\mathbf{Z}^*\sim N(\mathbf{0},\mathbf{I})$. The error term is $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\boldsymbol{\Psi})$, which has a diagonal covariance matrix $\boldsymbol{\Psi}=\mathrm{diag}(\psi_1,\dots,\psi_{\sum_i p_i})$. The marginal distribution of the integrated data matrix $\mathbf{X}=(\mathbf{X}_1,\dots,\mathbf{X}_m)'$ is then multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma}=\mathbf{W}\mathbf{W}'+\boldsymbol{\Psi}$, where $\mathbf{W}=(\mathbf{W}_1,\dots,\mathbf{W}_m)'$. The corresponding log-likelihood function of the data is

$$\ell(\mathbf{W},\boldsymbol{\Sigma})=-\frac{n}{2}\left(\sum_{i=1}^{m}p_i\ln(2\pi)+\ln\det(\boldsymbol{\Sigma})+tr(\boldsymbol{\Sigma}^{-1}\mathbf{G})\right), \tag{7}$$

where $\mathbf{G}$ is the sample covariance matrix of the following form

$$\mathbf{G}=\begin{pmatrix}\mathbf{G}_{11}&\mathbf{G}_{12}&\cdots&\mathbf{G}_{1m}\\ \mathbf{G}_{21}&\mathbf{G}_{22}&\cdots&\mathbf{G}_{2m}\\ \vdots&\vdots&\ddots&\vdots\\ \mathbf{G}_{m1}&\mathbf{G}_{m2}&\cdots&\mathbf{G}_{mm}\end{pmatrix}. \tag{8}$$

We employ the EM algorithm to obtain the maximum likelihood estimates of $\mathbf{W}$ and $\boldsymbol{\Psi}$. In the EM framework, we deal with the complete-data log-likelihood

$$\begin{aligned}\ell_c(\mathbf{W},\Psi)=&-\frac{n}{2}\left\{\sum_{i=1}^{m}p_i ln(2\pi)+\ln\det(\Psi)\right\}\\ &-\frac{1}{2}\left\{tr((\mathbf{X}-\mathbf{W}\mathbf{Z}^*)'\Psi^{-1}(\mathbf{X}-\mathbf{W}\mathbf{Z}^*))+tr(\mathbf{Z}^{*'}\mathbf{Z}^*)\right\}.\end{aligned} \tag{9}$$

This is a much more efficient approach than directly maximizing the marginal data likelihood in (7). It does not require explicit evaluation of the sample covariance matrices in (8), which would call for $O(n\sum_i p_i^2)$ operations and thus be computationally prohibitive.

Finally, the problem of $p>>n$ is exacerbated in our model by the multiple high-dimensional datasets. A sparse solution to $\mathbf{W}$ is desirable. In the next section, we derive a sparse solution to solve the iCluster model via penalizing the complete-data log-likelihood.

## 2.4 A sparse solution

We write the penalized complete data log-likelihood as

$$\ell_{c,p}(\mathbf{W},\boldsymbol{\Psi})=\ell_c(\mathbf{W},\boldsymbol{\Psi})-J_\lambda(\mathbf{W}), \tag{10}$$

where $J_\lambda(\mathbf{W})$ is a penalty term on $\mathbf{W}$ with a non-negative regularization parameter $\lambda$. Various types of penalties can be employed. In this study, we use a lasso type ($L_1$-norm) penalty (Tibshirani, 1996) that takes the form

$$J_\lambda(\mathbf{W})=\lambda\sum_{i=1}^{m}\sum_{k=1}^{K-1}\sum_{j=1}^{p_i}|w_{ikj}|. \tag{11}$$

We derive the E- and M-step with respect to the penalized complete-data log-likelihood. The E-step involves computing the objective function

$$Q_p(\mathbf{W},\boldsymbol{\Psi}|\mathbf{W}^{(t)},\boldsymbol{\Psi}^{(t)})=E_{\mathbf{Z}^*|\mathbf{X},\mathbf{W}^{(t)},\boldsymbol{\Psi}^{(t)}}[\ell_{c,p}(\mathbf{W},\boldsymbol{\Psi})],$$

which is the expected value of the complete-data log-likelihood with respect to the distribution of $\mathbf{Z}^*$ given $\mathbf{X}$ under the current estimates $(\mathbf{W}^{(t)},\boldsymbol{\Psi}^{(t)})$.

This involves computing the following quantities given the current parameter estimates:

$$E[\mathbf{Z}^*|\mathbf{X}]=\mathbf{W}'\Sigma^{-1}\mathbf{X} \text{ and}$$
$$E[\mathbf{Z}^*\mathbf{Z}^{*'}|\mathbf{X}]=\mathbf{I}-\mathbf{W}'\Sigma^{-1}\mathbf{W}+E[\mathbf{Z}^*|\mathbf{X}]E[\mathbf{Z}^*|\mathbf{X}]'. \quad (12)$$

The E-step provides a *simultaneous dimension reduction* by mapping the original data matrices of joint dimensions $(p_1,\dots,p_m)\times n$ to a substantially reduced subspace represented by $\mathbf{Z}^*$ of dimension $(K-1)\times n$.

The M-step is to update the parameter estimates by maximizing $Q_p$ subject to $\|\mathbf{w}_k\|=1$ for all $k$. This leads to the following estimate of $\Psi$:

$$\boldsymbol{\Psi}^{(t+1)}=\frac{1}{n}\text{diag}\left\{\mathbf{X}\mathbf{X}'-\mathbf{W}^{(t)}E[\mathbf{Z}^*|\mathbf{X}]\mathbf{X}'\right\} \quad (13)$$

and the lasso estimate of $\mathbf{W}$:

$$\mathbf{W}_{\text{lasso}}^{(t+1)}=\text{sign}(\mathbf{W}^{(t+1)})\left(|\mathbf{W}^{(t+1)}|-\lambda\right)_+, \quad (14)$$

where $\mathbf{W}^{(t+1)}=\left(\mathbf{X}E[\mathbf{Z}^*|\mathbf{X}]'\right)(E[\mathbf{Z}^*\mathbf{Z}^{*'}|\mathbf{X}])^{-1}$. This is followed by a normalization step $\mathbf{w}_k/\|\mathbf{w}_k\|_2$ for all $k$, where $\|\mathbf{w}_k\|_2$ denotes the $L_2$ norm of the vector $\mathbf{w}_k$ that takes the form $\sqrt{\sum_j w_{jk}^2}$. The algorithm iterates between the E- and M-step until convergence. Once $\hat{E}[\mathbf{Z}^*|\mathbf{X}]$ is obtained, a final step to recover the class indicator matrix is to invoke a standard $K$-means on $\hat{E}[\mathbf{Z}^*|\mathbf{X}]$. We denote this solution as $\hat{\mathbf{Z}}_{\text{iCluster}}$.

The lasso-type penalty results in sparse estimates of $\mathbf{W}$ in which many of the coefficients are shrunken toward zero. The variance of the model is thus reduced, leading to better clustering performance though the bias-variance trade-off. The lasso also renders a variable selection mechanism owing to the $L_1$ penalty that shrinks some coefficients to *exactly* zero. As a result, one can pinpoint which genes contribute to which subtype by finding the genes with non-zero loadings on the $k$-th latent factor $\mathbf{z}_k$. This will be demonstrated in the data example.

### 2.5 Model selection based on cluster separability

Let $\hat{\mathbf{B}}^*=\hat{E}[\mathbf{Z}^*|\mathbf{X}]'\hat{E}[\mathbf{Z}^*|\mathbf{X}]$ be ordered such that samples belonging to the same clusters are adjacent. Then $\hat{\mathbf{B}}^*$ has a diagonal block structure and can be used to assess cluster separability. We standardize the elements of $\hat{\mathbf{B}}^*$ to be $b_{ij}/\sqrt{b_{ii}b_{jj}}$ for $i=1,\dots,n$ and $j=1,\dots,n$, and impose a non-negative constraint by setting negative values to zero. Then perfect cluster separability (non-overlapping subclasses) would lead to an exact diagonal block matrix with diagonal blocks of ones for samples belonging to the same cluster and off-diagonal blocks of zeros for samples in different clusters. As cluster separability decreases, $\hat{\mathbf{B}}^*$ increasingly deviates from the 'perfect' diagonal block structure. We thus define a deviance measure $d$ as the sum of absolute differences between $\hat{\mathbf{B}}^*$ and a 'perfect' diagonal block matrices of 1s and 0s. The proportion of deviance (POD) is defined as $d/n^2$ so that POD is between 0 and 1. Small values of POD indicate strong cluster separability, and large values of POD indicate poor cluster separability. In the data examples, we show the utility of $\hat{\mathbf{B}}^*$ matrix plots (we call them cluster separability plots) and associated the POD statistic for model selection, which includes estimating the number of clusters $K$ and the lasso parameter $\lambda$.

## 3 RESULTS

### 3.1 Subtype discovery in breast cancer

Pollack *et al.* (2002) studied 37 primary breast cancers and four breast cancer cell lines for DNA copy number and mRNA expression on the same cDNA microarrays that contain 6691 genes. Figure 2A shows the pair of heatmaps displaying the alteration patterns in the DNA (left panel) and in the mRNA (right panel) on chromosome 17. Samples are arranged by separate hierarchical clustering output. Clearly, the two dendrograms are substantially different. Although the leftmost clusters share members that carry the *HER2/ERBB2* amplicon profile near 17q12, they are not identical. This is a problem inherent to separate clustering approaches that fail to account for the correlation between the two datasets. On the other hand, mixing breast tumors and cell line samples, the four cell line samples (BT474, T47D, MCF7 and SKBR3, indicated in red text) should be distinguished as a separate 'subtype' from the rest of the tumor samples. This is clearly the case in the gene expression data, but it is not recapitulated in the DNA copy number data. This contrast shows the importance of capturing unique patterns specific to one data type.

Figure 2B–E shows the results of a unified set of cluster assignments from iCluster on the same data. Non-sparse ($\lambda=0$) and sparse solutions ($\lambda=0.01$ and 0.2) were generated. Figure 2B includes cluster separability plots described in Section 2.3 under the sparse solution given $\lambda=0.2$. Clearly, $K=4$ gives the best diagonal block structure. This is confirmed in Figure 2C where the four-cluster sparse solution ($\lambda=0.2$) minimized the POD statistic among a range of $K$ and $\lambda$ values. Figure 2D displays the heatmaps of the same data used in Figure 2A but with samples rearranged by their iCluster membership. In a completely automated fashion, the four cell lines were separated as cluster 1 (red). The *HER2/ERBB2* subtype emerged as cluster 2 (green) and showed coordinated amplification in the DNA and overexpression in the mRNA. This subtype was associated with poor survival as shown in Figure 2E. Cluster 3 was a potentially novel subtype derived only as a result of combining evidence across the two datasets. It represents a subset of tumors characterized by weak yet consistent amplifications toward the end of the q-arm of chromosome 17. Finally, cluster 4 did not show any distinct patterns, though a pattern may have emerged if there were additional data types. As mentioned in Section 2.4, the lasso-type penalty in the sparse iCluster solution renders variable selection as a part of the outcome. Supplementary Table 1 lists the selected subset of genes associated with each of the subtypes.

### 3.2 Lung cancer subtypes jointly defined by copy number and gene expression data

We also analyzed a set of 91 lung adenocarcinomas from Memorial Sloan–Kettering Cancer Center, which is a subset of the samples in Chitale *et al.* (2009). The iCluster method was applied to perform integrative clustering on copy number and gene expression data. The copy number data were segmented using the CBS algorithm (Olshen *et al.*, 2004; Venkatraman *et al.*, 2007). The segment means were used as the input for integration to reduce the noise level. Variance filtering based on gene expression was performed so as to focus on the most variable set of 2782 genes.

Using chromosomes 8 and 12 as examples, we compared the iCluster results with those obtained by separate hierarchical clustering. Cluster 1 in Figure 3A is characterized by a broad region of 8p loss evident in the copy number heatmap and the corresponding underexpression in the expression heatmap. In contrast, this 8p loss cluster is less well defined by separate clustering in Figure 3B. When annotated with somatic mutation status, this cluster shows significant enrichment of *EGFR* mutations (mutation panel on top of the heatmap). Specifically, 33% of the tumors in cluster 1 carry *EGFR* mutation, while 16%, 0% and 18% of the tumors in cluster 2, 3 and 4, respectively, are *EGFR* mutant samples (Fisher's exact test $P=0.03$). Another interesting observation made apparent by iCluster is that samples in cluster 4 show a similar but somewhat

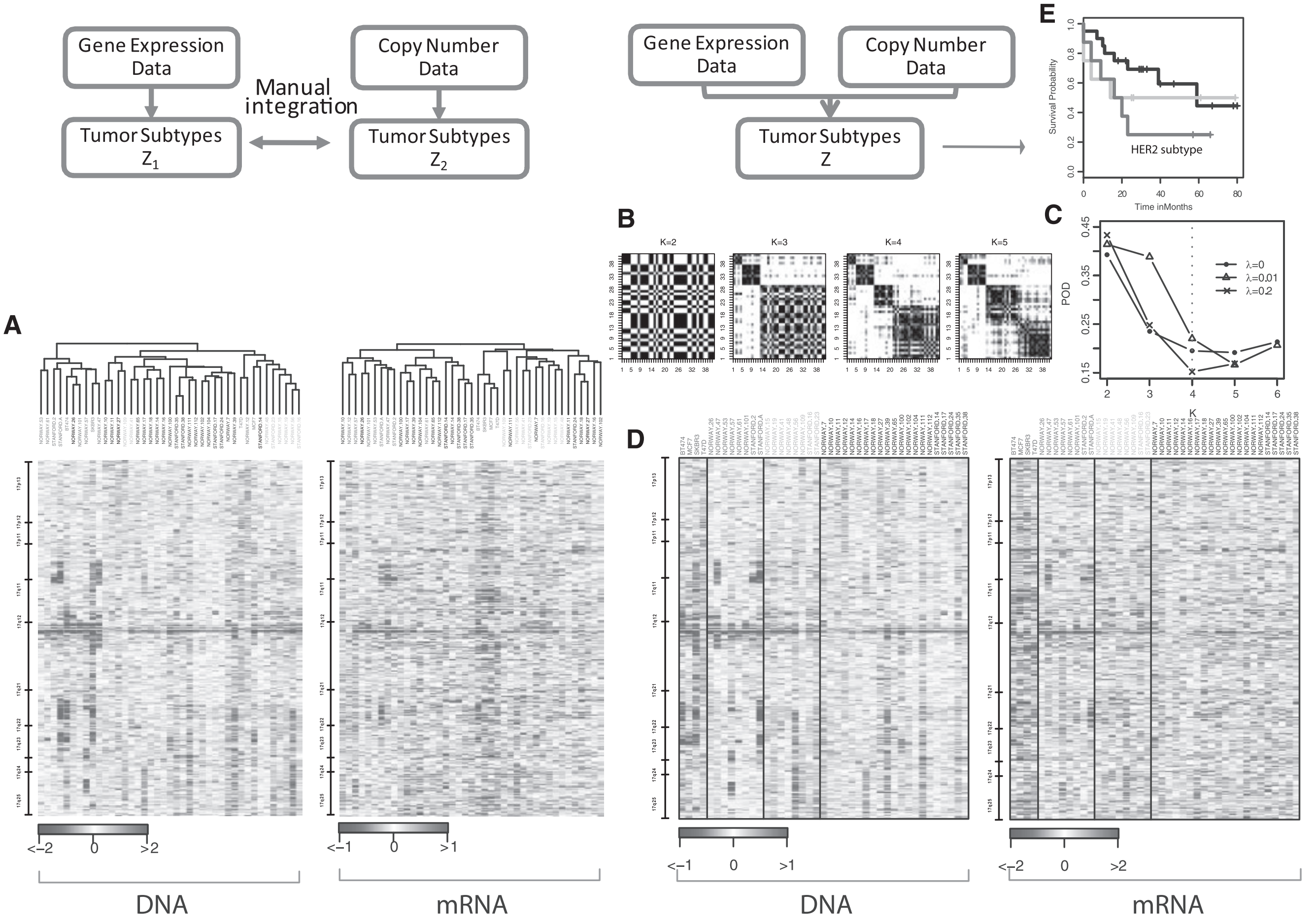

**Fig. 2.** Results from separate clustering (left panel) and integrative clustering (right panel) using the Pollack data. (**A**). Heatmaps of copy number (DNA) and gene expression (mRNA) on chromosome 17. Samples are arranged by separate hierarchical clustering on each data type. (**B**) Cluster separability plots. (**C**) Model selection based on POD measure. A four-cluster sparse solution ($\lambda = 0.2$) was chosen. (**D**) Heatmaps on the same data as in A with samples arranged by the integrated cluster assignment under the sparse iCluster model. (**E**) Kaplan–Meier plots of the subclasses identified via the integrative clustering. The *HER2/ERBB2* subtype showed poor survival.

diluted pattern of copy number aberrations when compared with cluster 1. These samples may be related to cluster 1 but with lower tumor content, which may account for the 18% EGFR mutations in this cluster, the second highest among the four clusters. Chitale *et al.* (2009) describe the association between chromosome 8p loss and *EGFR* mutation in further details. When studying the genes within the broad region of 8p loss, they discovered a striking association between *EGFR* mutation and concordant *DUSP4* deletion and underexpression. *DUSP4* is known to be involved in negative feedback control of *EGFR* signaling. Notably, the sparse solutions consistently showed better cluster separability than the non-sparse solution as evidenced by Figure 3C.

Chromosome 12 is another interesting example. Cluster 2 in Figure 3D is characterized by the well-known 12q14-15 amplicon that includes oncogenes such as *CDK4* and *MDM2*. Again, the sparse solution improves the cluster separability substantially from the non-sparse solution (Fig. 3F). Interestingly, the sparse model selected only 24 DNA probes that contributed to the clustering, which is consistent with the observation that there are relatively few aberrations other than the small region of 12q gain in the DNA. Note, however, that genomic alteration patterns are often chromosome specific (8p loss and 12q gain). They do not always occur in the same set of patients. Therefore, the results change when multiple chromosomes are combined (Supplementary Fig. 1).

## 4 DISCUSSION

Despite the ever-increasing volume of MGP, data resulting from the Cancer Genome Atlas project and other studies, there is a shortage of effective integrative methods. Researchers often resort to heuristic approaches where ‘manual integration’ is performed after separate analysis of individual data types, and it is unlikely that two investigators would perform manual integration in the same manner. Manual integration may require a considerable amount of prior knowledge about the underlying disease. In contrast, the iCluster method developed here generates a single integrated cluster assignment based on simultaneous inference from multiple data types. In both the breast and lung cancer data examples, we have

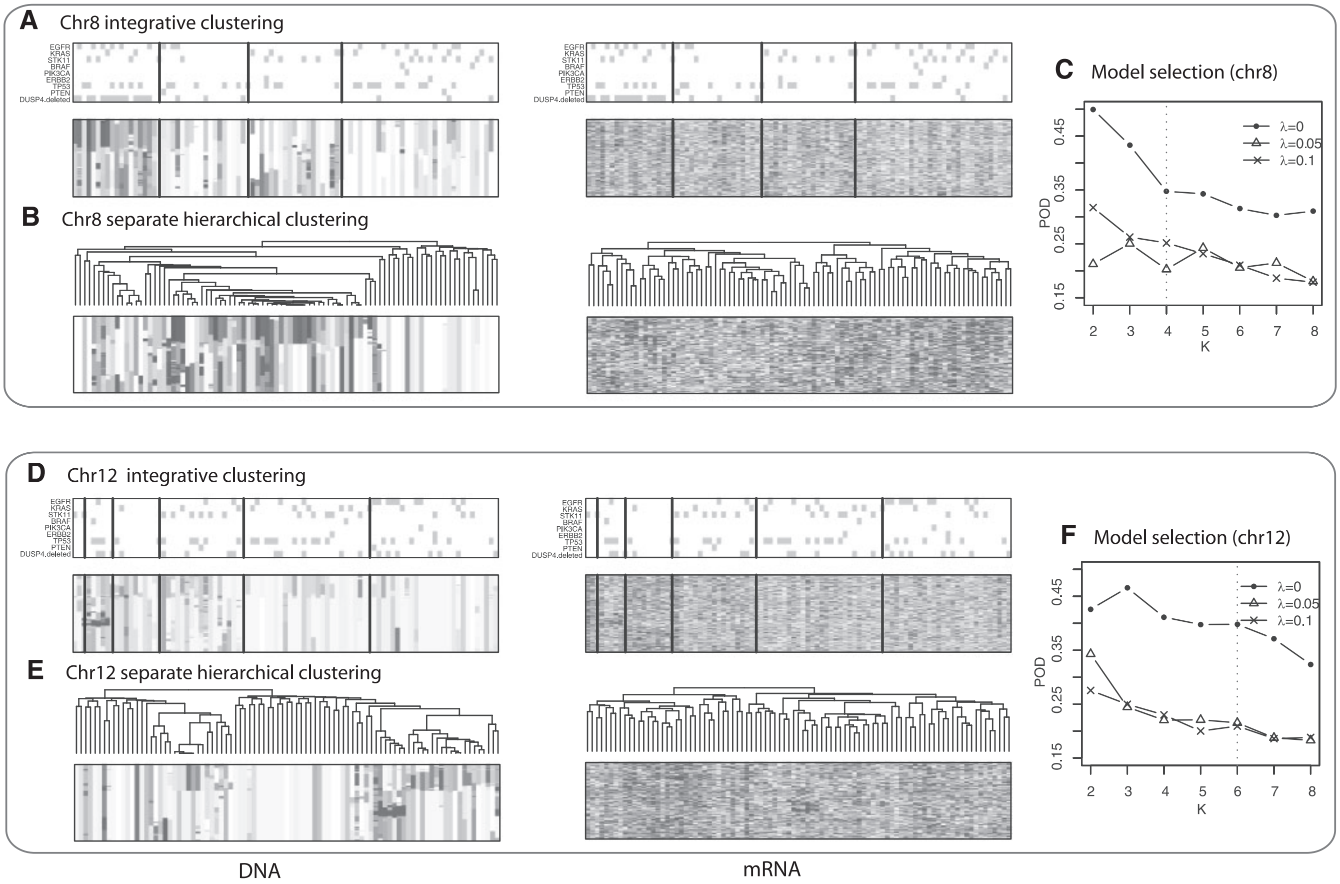

**Fig. 3.** Lung cancer subtypes for chromosomes 8 and 12. (**A**) Heatmap of DNA copy number (left) and mRNA expression (right) on chromosome 8. Columns are tumors arranged by the three subclasses obtained by iCluster. Rows are genes ordered by genomic position. On top of the heatmaps are gray-dot panels indicating mutation status of several well-known lung cancer genes. (**B**) Separate hierarchical clustering of the same data on chromosome 8 used in (A). (**C**) Model selection based on the POD measure. A four-cluster sparse solution ($\lambda = 0.05$) was chosen that selected 301 mRNA probes and 126 DNA probes from a total of 642 probes. (**D**) iCluster output on chromosome 12. Tumor samples are arranged by the six subclasses obtained by iCluster.(**E**) Separate hierarchical clustering of the same data on chromosome 12 used in (D). (**F**) Model selection based on the POD statistic. A six-cluster sparse solution ($\lambda = 0.1$) was chosen that selected 408 mRNA probes and 24 DNA probes from a total of 1038 probes.

shown that iCluster aligns concordant DNA copy number aberrations and gene expression changes. In some cases, potentially novel subclasses are revealed only by combining weak yet consistent evidence across data types.

In this study, we applied iCluster to integrate copy number and gene expression data. The joint latent variable model is completely scalable to include additional data types. Next-generation sequencing is emerging as an appealing alternative to microarrays for inferring RNA expression levels (mRNA-Seq), DNA–protein interactions (ChIP-Seq), DNA methylation and so on. Although we focus here on array data, our integrative framework could be generalized to next-gen sequencing data after proper modifications of the error terms to model count data based on mapped reads.

## ACKNOWLEDGMENTS

We thank Dr. Colin Begg, Dr. Glenn Heller and Dr. Richard Olshen for helpful comments. We thank the reviewers for their constructive comments, which we used to improve the manuscript.

*Funding*: Cancer Center Support Grant P30 CA008748 (in part).

*Conflict of Interest*: none declared.

## REFERENCES

Alter,O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Chitale,D. *et al.* (2009) An integrated genomic analysis of lung cancer reveals loss of *DUSP4* in *EGFR*-mutant tumors. *Oncogene*, **28**, 2773–2783.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **39**, 1–38.

Ding,C.H.Q. and He,X. (2004) *K*-means clustering via principal component analysis. In *ICML*, Vol. 69 of *ACM International Conference Proceeding Series*, ACM, Banff, Alberta, Canada.

Holter,N.S. *et al.* (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.

Hyman,E. *et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.

Kool,M. *et al.* (2008) Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PLoS ONE*, **3**, e3088–e2102.

Lee,H. *et al.* (2008) Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, **24**, 889–896.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, **120**, 15–20.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pollack,J.R. *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Qin,L.X. (2008) An integrative analysis of microRNA and mRNA expression - a case study. *Cancer Inform.*, **6**, 369–379.

TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Stat. Method.)*, **58**, 267–288.

Tipping,M.E. and Bishop,C.M. (1999) Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **61**, 611–622.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Zha,H. *et al.* (2001) Spectral relaxation for K-means clustering. In *Neural Information Processing Systems (NIPS 2001)*. Vol. 14. MIT Press, Vancouver, Canada, pp. 1057–1064.