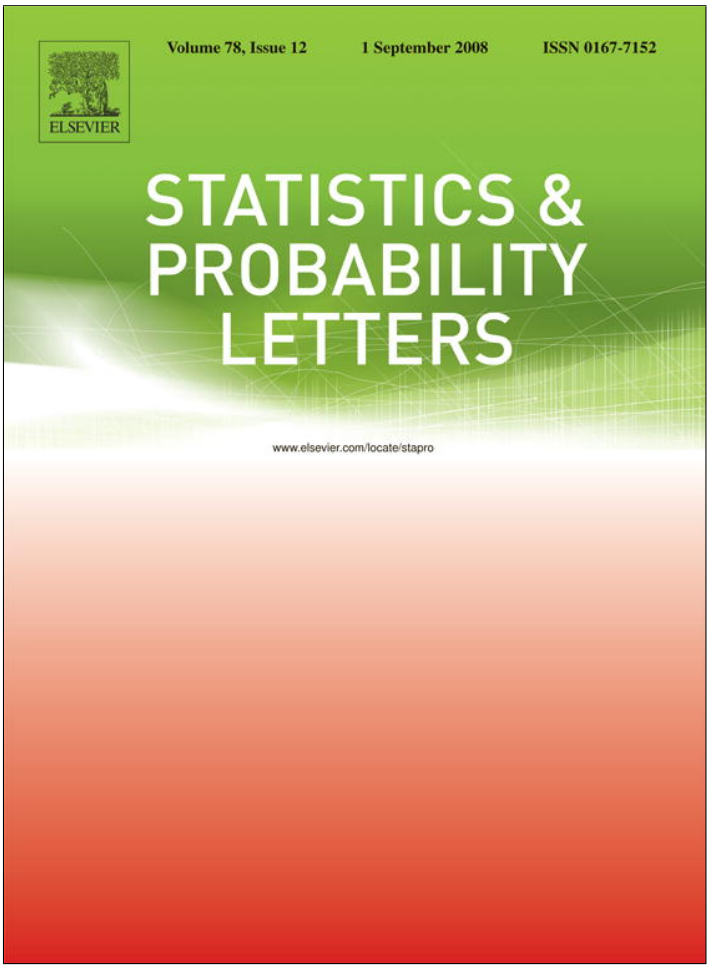

Available online at www.sciencedirect.com
ScienceDirect
Statistics and Probability Letters 78 (2008) 1490–1497
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Clustering gene expression profile data by selective shrinkage

Hemant Ishwaran*, J. Sunil Rao

Cleveland Clinic and Case Western Reserve University, United States

Received 16 October 2007; received in revised form 14 December 2007; accepted 1 January 2008
Available online 15 January 2008

Abstract

Clustering of gene expression profiles is a widely used approach for finding macroscopic data structure. A complication in such analyses is that not all genes are informative for forming clusters and different clusters might have different transcription regulation. Driven by these considerations, we present a novel two-stage clustering approach. The first stage identifies informative genes by adaptive variable selection using pseudo-samples modeled by a high dimensional multigroup ANOVA model. Variables are selected using a rescaled spike and slab Bayesian hierarchical model having a special selective shrinkage property. The second stage uses output from the first stage for clustering. We demonstrate why selective shrinkage occurs, and by extension, why it is useful for the clustering paradigm. We analyze a human gene atlas expression dataset where the question of interest is to look for tissue-specific transcription regulation and investigate whether tissues can be grouped together due to similar genomic control.

1. Introduction

Cluster analysis is a popular tool often used as a preliminary analysis in the interpretation of data from gene expression profiling experiments. The interest lies in identifying groups of samples that have a similar expression pattern, and also in genes appearing similar across samples. Supervised clustering may also be used where the interest lies in associating sample clusters with an outcome of interest (disease progression, survival, etc.). Additionally, external information, such as biological annotation, can be used to help tune distances between objects (see Boratyn et al., 2007, for details). Dissimilarity between observations or genes is typically measured by a distance metric (see Gower, 1971; Joly and Le Calve, 1996, for detailed discussions).

Gene expression profiles present special problems for cluster analysis. Beyond the shear dimensionality (i.e. number of genes) which can far exceed the sample size, there are many other subtle issues at play. For instance, it is typical that not all genes are informative in forming subgroups — the so-called feature selection problem. In addition, it is also quite likely that different subgroups might cluster on different numbers of genes. This clustering on a subset of attributes has been studied by others. For example, see the Clustering Objects on Subsets of Attributes (COSA) algorithm of Friedman and Meulman (2004) or the gene shaving algorithm of Hastie et al. (2000). These algorithms embed the feature selection step into the clustering algorithm but may be computationally difficult to implement

* Corresponding author.
E-mail address: hemant.ishwaran@gmail.com (H. Ishwaran).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2008.01.003

requiring specification of tuning parameters for which there is little theory to suggest good values (Friedman and Meulman, 2004).

We cast the clustering problem into two stages. The first stage analysis focuses on whether a given samples' gene expression differs relative to other samples. Informative genes with expression differences correspond to genes with non-zero parameter estimates from a high-dimensional multigroup ANOVA model. Parameter estimates are obtained from a special rescaled spike and slab hierarchical model designed for selecting differentially expressed genes in multigroup microarray experiments (Ishwaran and Rao, 2003, 2005a). Rescaled spike and slab models have a selective shrinkage property which shrinks ordinary least squares (OLS) estimates of truly non-informative genes towards zero, thereby making delineation of informative genes easier.

The output of the stage one analysis is a vector of Bayesian parameters (one for each sample), in which each element is an estimate of signal for a gene for a given sample relative to other samples. The second-stage analysis uses these to rank genes and to cluster samples.

The paper is organized as follows. Section 2.1 introduces multigroup microarray data and reviews the use of rescaled spike and slab models for finding differentially expressing genes in such settings. Gene clustering is then recast in terms of a no-baseline ANOVA model (Section 2.2). A rescaled spike and slab Bayesian hierarchy is introduced as a means for accurate parameter estimation (Section 2.3). Section 3 discusses adaptive selection for the resulting Bayesian parameter estimates, and is the basis for assessing informativeness of a gene used in the clustering algorithm. Section 4 illustrates the clustering procedure on a large data set made up of differing human tissue types.

2. Model selection for gene expression profile clustering

We show in this section how clustering gene expression profiles can be approached using a two-stage analysis. The first stage involves hunting for differentially expressed genes in a multigroup microarray data setting; that is, an experimental design where expression data involves multiple biologic groups. Following this, we show how the problem can be mapped to the problem of clustering gene expression profile data.

2.1. Spike and slab gene selection

The problem of finding differentially expressed genes in multigroup settings can be described mathematically as follows. Let $Y_{i,j}$ be the expression value for individual i, for $i = 1, \ldots, n$, from gene j, for $j = 1, \ldots, p$. Each individual belongs to a specific group which we indicate by a class label C_i. To study whether there is a group difference for a gene j, a standard approach is to construct a statistic T_j for testing a null hypothesis H_j of no difference. Genes differentially expressed are found using the p-value for each test, then making an adjustment to account for false detections; for example by using false discovery rate (FDR) error control (Efron et al., 2001; Storey, 2002; Tusher et al., 2001) or family-wise error rate (FWER) (Datta and Datta, 2005).

Alternatively, differentially expressing genes can be detected using a Bayesian variable selection approach called Bayesian ANOVA for Microarrays (BAM). This approach rests on the use of rescaled spike and slab models, a new class of Bayesian hierarchical models for high dimensional variable selection and prediction. BAM was first applied to microarray data from two-groups. Performance relative to FDR and FWER methods was studied and it was shown BAM was better able to maintain a balance between low false detection and high statistical power (Ishwaran and Rao, 2003). BAM was later extended to multigroup microarray data settings by recasting the problem as a high dimensional orthogonal model selection problem (Ishwaran and Rao, 2005a). In orthogonal settings, rescaled spike and slab models were shown to possess a *selective shrinkage* property (Ishwaran and Rao, 2005a,b). This property allows the posterior mean for coefficients, used to select genes, to shrink towards zero for truly zero coefficients, while for non-zero coefficients, posterior estimates are similar to OLS estimates. In Ishwaran and Rao (2005a), selective shrinkage was shown to be a sufficient condition for superior total misclassification of genes relative to the OLS. Further, gains in misclassification was shown to increase as number of biologic groups increased in sparse settings. A finite sample adaptive method for selecting genes, exploiting the property of selective shrinkage, was given.

The model selection approach (Ishwaran and Rao, 2005a) for multigroup microarray data rests implicitly on an ANOVA model, defined as follows:

$$\begin{aligned} Y_{i,j} &= \theta_{j,0} && \textit{Baseline effect for gene } j \\ &\quad + \sum_{g=1}^{G-1} \beta_{g,j,0}\mathbb{I}\{C_i = g\} && \textit{Differential effect} \\ &\quad + \varepsilon_{i,j}, && \textit{Error: distributed as } \mathscr{D}_{i,j}(0,\sigma_j^2) \\ &\quad i = 1,\ldots,n,\ j = 1,\ldots,p. \end{aligned} \tag{1}$$

The value $\theta_{j,0}$ in (1) represents a baseline effect for gene j. The baseline group is typically taken to be group G, where the different biologic groups are coded as $\{1,\ldots,G\}$. The parameters $\beta_{g,j,0}$ represent the gene differential expression effect for group g relative to the baseline group G. For example, in two-group problems, $G = 2$ and $\beta_{1,j,0}$ is a measure of differential expression between the two groups. The ANOVA model makes minimal assumptions about the nature of the data. A distribution free approach is taken where it is assumed only that $\varepsilon_{i,j}$ are independent such that $\mathbb{E}(\varepsilon_{i,j}) = 0$ and $\mathbb{E}(\varepsilon_{i,j}^2) = \sigma_j^2$.

2.2. Clustering as a two-stage analysis from a multigroup ANOVA setup

The sample size of individuals within a group plays a crucial role in the effectiveness of BAM in identifying differentially expressing genes. If sample sizes are small within group, then the posterior mean (which is used for finding differentially expressing genes) will have high variance and the method suffers. Clustering gene expression profiles falls under this paradigm.

We modify and extend BAM to this setting. The key idea is to increase the sample size by using each sample as a baseline value. For each sample i, we create "pseudo-data" $Y_{i,j,k}$, where

$$Y_{i,j,k} = Y_{k,j} - Y_{i,j}, \quad \text{for } k = 1,\ldots,i-1,i+1,\ldots,n.$$

Thus, each expression value $Y_{i,j}$ is used as a baseline value that is subtracted from all other expression values for gene j. In place of $Y_{i,j}$, we substitute $n-1$ pseudo-observations

$$Y_{i,j,1},\ldots,Y_{i,j,i-1},Y_{i,j,i+1},\ldots,Y_{i,j,n}.$$

The group membership label is removed in unsupervised clustering. All observations in the pseudo-data above are simply referred to as belonging to sample i. Pseudo-data is created for each sample i for a gene j. In total there are n samples with $n-1$ pseudo-observations within each sample for each gene.

We assume a modified ANOVA framework for the pseudo-data, defined as follows:

$$Y_{i,j,k} = \beta_{i,j,0} + \varepsilon_{i,j,k} \quad i = 1,\ldots,n,\ j = 1,\ldots,p,\ k = 1,\ldots,i-1,i+1,\ldots,n. \tag{2}$$

Here $\beta_{i,j,0}$ represent a sample-gene differential effect. Since $Y_{i,j,k}$ has subtracted $Y_{i,j}$, a baseline correction dependent on i, there is no fixed baseline group in this analysis, and $\beta_{i,j,0}$ represents a differential effect for gene j, for sample i, *relative to all other samples*.

The interpretation of the $\beta_{i,j,0}$ is interesting. Clustering based on these values implies clustering samples with similar "baseline effects". The advantage of this approach is samples that are close to one another are with respect to $\beta_{i,j,0}$, a summarized relative distance from all other samples, rather than based on simple pairwise distances as in typical similarity measures. Further, using the multigroup setup in tandem with a rescaled spike and slab hierarchical model for estimation (the details of which are provided in the next section), allows the identification of only useful genes in determining similarity of baseline effects across samples.

The clustering procedure clusters genes and samples using posterior mean values from our Bayesian model. Let $\hat{\beta}_{i,j}$ denote the posterior mean for $\beta_{i,j}$. Using an adaptive selection rule (Section 3), $\hat{\beta}_{i,j}$ is mapped to a 3-bit pixel, $P_{i,j} \in \{-1,0,+1\}$. A pixel value of $P_{i,j} = 0$ indicates $\beta_{i,j} = 0$ and sample i is similar to all other samples for gene j. A value of $P_{i,j} = -1$ indicates sample i is down-regulated, whereas a value of $P_{i,j} = +1$ indicates up-regulation. The $\hat{\beta}_{i,j}$ are clustered by gene and sample.

The algorithm for clustering $\hat{\beta}_{i,j}$ is as follows:

1. Column (gene) clustering: Rank $\hat{\beta}_{i,j}$ by gene j. The top gene is that gene j with the most number of pixels $P_{i,j} \neq 0$, for $i = 1,\ldots,n$. The second highest ranked gene is that gene with second largest number of non-zero pixels, etc. Sort genes.

2. Row (sample) clustering: Compute the rank for $|\hat{\beta}_{i,j}|$ for $j = 1, \ldots, p$. Large absolute values are associated with high rank. For those $\hat{\beta}_{i,j}$ with pixels $P_{i,j} = 0$ set the rank to zero. This yields a p-dimensional rank vector for each sample i with rank values from $\{0, 1, \ldots, p\}$. From this compute a $n \times n$ correlation matrix. Use this for the similarity matrix in a clustering procedure. Cluster the rows.

Step 2 is unusual. Clustering on samples typically would proceed using a dissimilarity matrix defined by a distance metric applied to $\hat{\beta}_{i,j}$. Using ranked values of $|\hat{\beta}_{i,j}|$ to cluster samples, however, gives a more refined view of closeness of samples. We do not necessarily expect samples to cluster because they have gene effects of similar magnitude. Rather, it is more likely that similarly ranked gene effects identify samples with similarly perturbed biological states and these samples should be deemed close.

2.3. Rescaled spike and slab model hierarchy

Here we present the details of our Bayesian hierarchy. We deviate slightly from the presentation given in Ishwaran and Rao (2003, 2005a) and exclude a Bayesian parameter for the variance σ^2 for the error distributions. Such a parameter is unnecessary as we shall assume $\varepsilon_{i,j,k}$ in (2) satisfy $\mathbb{E}(\varepsilon_{i,j,k}) = 0$ and $\mathbb{E}(\varepsilon_{i,j,k}^2) = 1$. The assumption of equal variance is unrealistic for microarray data, but the issue is easily handled by transforming the data using a pre-processing variance stabilization technique (Ishwaran and Rao, 2005a; Papana and Ishwaran, 2006). We assume the data has been variance stabilized to the value of 1.

With this caveat in mind, we model (2) using a rescaled spike and slab model (Ishwaran and Rao, 2005b). A rescaled spike and slab model is a spike and slab model model where the response value is scaled by the square-root of the sample size. Rescaling further involves incorporating a variance inflation factor equal to sample size into the Bayesian hierarchy. Because of the special setting considered here, no variance inflation factor is needed, and one can show that rescaling is equivalent to replacing $Y_{i,j,k}$ with rescaled values $Y_{i,j,k}^* = (n-1)^{-1/2} Y_{i,j,k}$. The rescaled spike and slab model is defined as follows:

$$\begin{aligned}
&(Y_{i,j,k}^* | \beta_{i,j}) \overset{\text{ind}}{\sim} \mathrm{N}\left(\beta_{i,j}, 1\right) \\
&(\beta_{i,j} | \gamma_{i,j}) \overset{\text{ind}}{\sim} \mathrm{N}(0, \gamma_{i,j}) \\
&\gamma_{i,j} \overset{\text{iid}}{\sim} \pi(d\gamma_{i,j}) \\
&\quad i = 1, \ldots, n,\ j = 1, \ldots, p,\ k = 1, \ldots, i-1, i+1, \ldots, n.
\end{aligned} \tag{3}$$

The hypervariances $\gamma_{i,j}$ are specified using the continuous bimodal priors of Ishwaran and Rao (2003, 2005a). The prior π for $\gamma_{i,j}$ is induced by the following parameterization. Define $\gamma_{i,j}$ by $\gamma_{i,j} = I_{i,j}\tau_{i,j}$, where $I_{i,j}$ and $\tau_{i,j}$ are parameters with priors specified according to (we use $\delta_x(\cdot)$ to indicate a measure concentrated at x):

$$\begin{aligned}
&(I_{i,j} | v_0, w_i) \overset{\text{ind}}{\sim} (1 - w_i)\, \delta_{v_0}(\cdot) + w_i\, \delta_1(\cdot) \\
&(\tau_{i,j}^{-1} | a_1, a_2) \overset{\text{iid}}{\sim} \mathrm{Gamma}(a_1, a_2) \\
&w_i \overset{\text{iid}}{\sim} \mathrm{Uniform}[0, 1].
\end{aligned} \tag{4}$$

The choice for v_0 (a small near zero value) and a_1 and a_2 (the shape and scale parameters for a gamma density) are selected so $\gamma_{i,j}$ has a continuous bimodal distribution with a spike at v_0 and a right continuous tail (Ishwaran and Rao, 2003, 2005a).

3. Adaptive gene selection

Adaptive selection using shrinkage plots was discussed in Ishwaran and Rao (2005a) for rescaled spike and slab models under orthogonal design matrices. A shrinkage plot is a plot of $\hat{\beta}_{i,j}$ against $\hat{\sigma}_{i,j}^2$, the posterior variance for $\beta_{i,j}$. In Ishwaran and Rao (2005a), it was shown asymptotically that truly differentially expressing genes are those genes with posterior variances $\hat{\sigma}_{i,j}^2$ coalescing near the value of 1 and with large $|\hat{\beta}_{i,j}|$ values. This led to an adaptive selection rule whereby genes were selected only if they were found on the far left and right sides of a shrinkage plot and with posterior variances nearly equal to 1.

However, while the theory for shrinkage plots applies to orthogonal design matrices such as (2), there are subtle differences in the setup here requiring careful modification to the original theory. One subtle issue is the complicated correlation structure within pseudo-data. Within any given gene, correlation exists in two forms. There is correlation across samples, and there is correlation within sample.

Consider two samples i and i'. The pseudo-data for sample i for gene j is,

$$Y_{k,j} - Y_{i,j}, \quad \text{for } k = 1, \ldots, i-1, i+1, \ldots, n,$$

whereas the pseudo-data for sample i' for gene j is,

$$Y_{k,j} - Y_{i',j}, \quad \text{for } k = 1, \ldots, i'-1, i'+1, \ldots, n.$$

These two sets of data are clearly heavily correlated. This issue is dealt with by the sample-specific Bayesian parameters w_i used in (4). See Ishwaran and Rao (2005a), Section 2.9, for further rationale for the use of group-specific complexity parameters.

More subtle, though, is the issue of correlation within sample. This leads to a persistent correlation as sample size increases. The correlation between two observations within a sample i and gene j is non-zero because of the presence of the baseline value. In fact,

$$\text{Corr}(Y_{i,j,k}, Y_{i,j,k'}) = \frac{\text{Var}(Y_{i,j})}{\sqrt{\text{Var}(Y_{k,j}) + \text{Var}(Y_{i,j})}\sqrt{\text{Var}(Y_{k',j}) + \text{Var}(Y_{i,j})}}.$$

If variances are equal, as we would expect after a variance stabilizing pre-processing step, then this correlation is exactly $1/2$. This correlation does not disappear as n increases.

3.1. Asymptotics under an infinite number of samples paradigm

A naive study of asymptotics will not be fruitful because of this. While at first glance this might seem problematic, one has to question how realistic such an analysis is in the context of the problem considered. Implicit in our approach is there are an unknown, but fixed, number of true clusters within the data. By assuming n is increasing, we are assuming number of samples increases, yet we are only using one observation as a baseline value for clustering a sample. This is unrealistic. To study asymptotics under a paradigm where number of samples is increasing, information for the baseline should be improved. This is reasonable since a priori there are only a finite number of clusters and thus information about a sample should improve as n increases. Therefore, for the purpose of an asymptotic treatment, we make use of pseudo-data defined as follows:

$$Y_{i,j,k} = Y_{k,j} - Z_{i,j}, \quad \text{for } k = 1, \ldots, i-1, i+1, \ldots, n,$$

Here $Z_{i,j}$ are values drawn independently from a normal distribution, $\text{N}(\mathbb{E}(Y_{i,j}), \lambda_n^2)$. Thus, $Z_{i,j}$ are independent draws from a distribution centered around the mean for the baseline value $Y_{i,j}$. Using $Z_{i,j}$ in place of $Y_{i,j}$ improves the baseline value. At the same time, we allow for uncertainty in $Z_{i,j}$, as reflected by the use of a variance λ_n^2 that can change with n.

The use of $Z_{i,j}$ removes correlation within a gene for a sample. Therefore, it is reasonable to assume that $\varepsilon_{i,j,k}$ in (2) are independent within sample for a given gene. Under mild regularity conditions, such as those required for a triangular central limit theorem, it is reasonable to assume for each i and j that:

$$(n-1)^{-1/2} \sum_{k \neq i} \varepsilon_{i,j,k} = O_p(1). \tag{5}$$

With this assumption, we now establish selective shrinkage and adaptive selection:

Theorem 1. *Assume that the true data model is (2), such that (5) holds. Under the rescaled spike and slab model defined by (3) and (4):*

$$\left(|\hat{\beta}_{i,j}|, \hat{\sigma}_{i,j}^2\right) \xrightarrow{\text{p}} (+\infty, 1),$$

if and only if $\beta_{i,j,0} \neq 0$.

Theorem 1 establishes a selective shrinkage property for the posterior mean by showing that only those truly differentially expressing genes (those genes where $\beta_{i,j,0} \neq 0$) have infinitely large posterior means in the limit. In finite sample settings, truly expressing genes are those having large absolute posterior means and posterior variances nearly equal to 1. These are the genes adaptively selected.

An interesting corollary to Theorem 1 is correct complexity recovery within a sample. This follows from a similar result in Ishwaran and Rao (2005a). That is, selective shrinkage implies adaptive selection finds the correct number of informative genes within a given sample i, in probability. The consequence of this is that estimated ranks of truly non-zero coefficients will be reliable. Recall Step 2 of our clustering procedure derives a similarity matrix based on this ranking, and a stable ranking for non-zero coefficients within a sample provides a clean signal to work with. Further, ranks for genes identified as being non-informative within a sample are set to zero. Selective shrinkage ensures we can identify such genes asymptotically. Thus, ranking for zero coefficients is also reliable and both effects combine to produce a robust similarity matrix.

4. Human gene atlas

The following example illustrates our clustering procedure. We consider the human gene atlas data from Su et al. (2004). We use the Human U133A–GNF1H, MAS 5.0 processed data; one of several variants of the database found at http://symatlas.gnf.org. In total there are 33,689 probe sets for each of 158 chips in this dataset. The 33,689 probe sets represent a custom designed compilation of Affymetrix Human U133A probe sets (22, 283) as well as special GNF1H probe sets (22, 645). The 158 chips represent 82 groups of various human tissue types obtained from a diverse panel of 79 individuals. No one group (tissue type) had more than two chips within it. See Su et al. (2004) for details.

Given at most there were two observations for any tissue type, we slightly modified the procedure for creating pseudo-data. For each tissue group, we computed the median expression value for a gene (for tissues queried by only 1 chip, this amounts to using the expression value for a gene). The baseline median value was subtracted from the expression value for the gene from each of remaining chips not in the tissue group. This was repeated for all genes for each of the 82 tissues. This resulted in pseudo-data with $n = 82$ and a total of 12,797 observations on each of 33,689 probe sets (genes).

The data was analyzed using BAMarray software (Ishwaran et al., 2006). By invoking a "no-baseline" analysis the software fits the rescaled spike and slab model defined by (3). The data was pre-processed using BAMarray's built in variance stabilizing algorithm (Papana and Ishwaran, 2006). We used the "unequal variance" option which variance stabilizes the data under an assumption of unequal variance across tissue types. Significant genes were identified using the adaptive selection rule employed by BAMarray. This rule makes use of a shrinkage plot as described in Section 3. We took the top 1000 genes identified using this rule and applied the clustering algorithm of Section 2.2 to this filtered subset (as our heatmap will show, there was not much to be gained using more than 1000 genes). In Step 2 of the algorithm we used PAM (partitioning around mediods) clustering (see Chapter 2 of Kaufman and Rousseauw (1990) for details). We tried other standard clustering algorithms with similar findings. The results were converted to a 3-bit pixel grid. See Fig. 1 for the resulting heatmap.

In Su et al. (2004), one of the goals of the analysis was to find tissue-specific transcription regulation by identifying genes expressed for a tissue. In Fig. 1, we have taken this analysis further by asking which tissues cluster together and then identifying genes important in driving this clustering. Note that all tissues (leave aside colorectal adenocarcinoma), represent the normal transcriptome, thus allowing examination of global trends in gene expression (Su et al., 2004). Examining Fig. 1, it is clear a reasonable clustering has been identified. First, the heatmap clearly identifies separated tissue clusters and genes that drive this clustering, as well as genes non-informative for clustering. Of immediate note is that certain tissue clusters are genomically controlled by a diverse set of genes while others seem to be tightly controlled by fewer genes. Also, some clusters are driven by up- or down-regulation of subsets of genes, while others are driven by both up- and down-regulation. There are some curious groupings. For instance, one which includes ganglion tissues, cardiac tissue, skin, skeletal muscle and interstitial testis (the second cluster from the bottom appearing all red). At the same time, tissue homogeneous subgroups also have been found. One example is the group including CD type cells and lymphoblasts. Interestingly, some tissues seem to sit off on their own. These include the colorectal adenocarcinoma and the fetal tissues — both of which a priori would not be expected to be close to the other tissues. Many such inferences can be gleaned and further workup can be done as

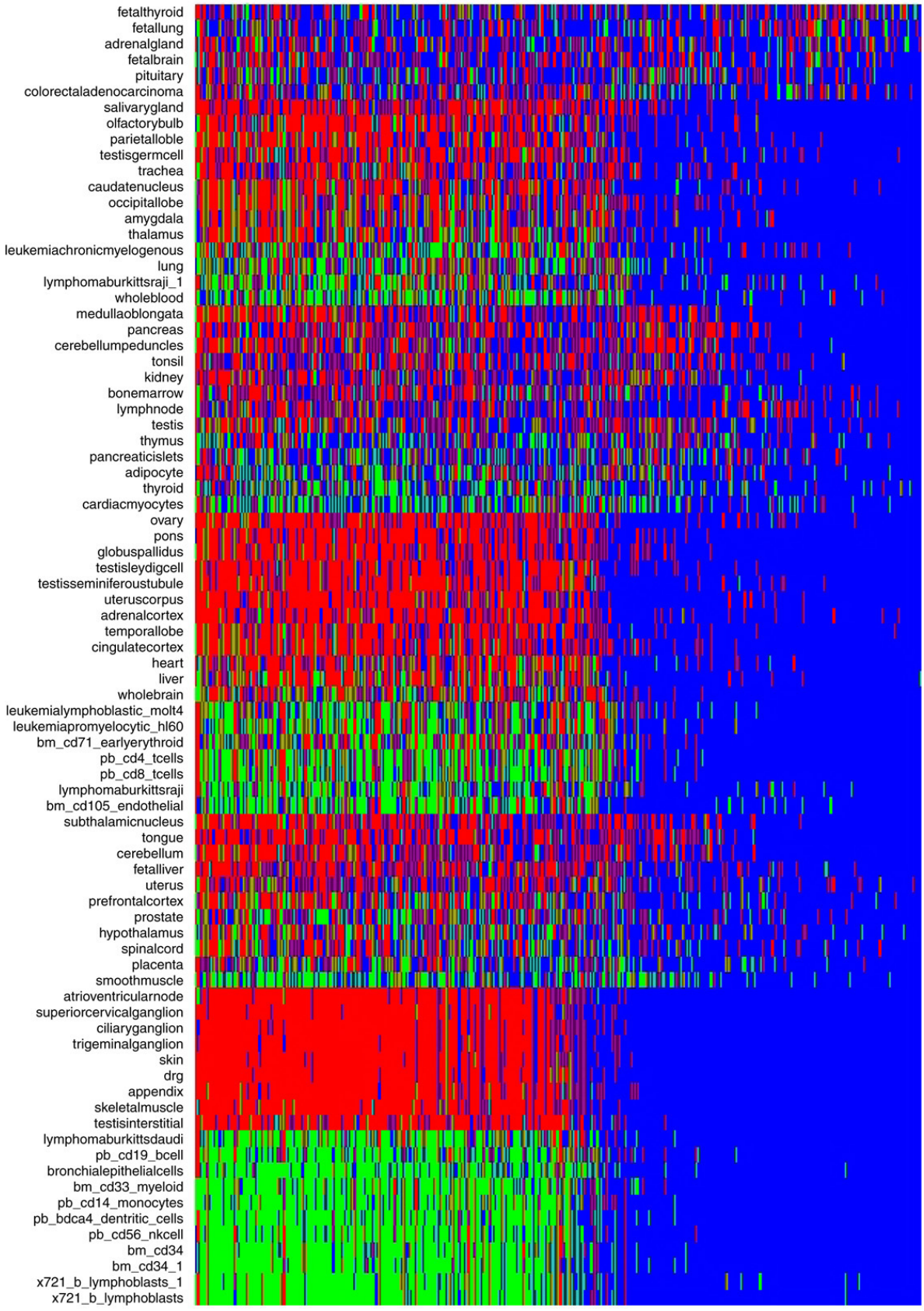

Fig. 1. Human gene atlas. Vertical axis corresponds to tissue type, horizontal axis is for genes (top 1000). Red points are genes up-regulated for a specific tissue relative to all other tissues; green points are genes down-regulated; blue points are non-significant genes. The large region of blue points on the right-hand side shows not much further structure can be found using more than 1000 genes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Su et al. (2004) which includes allelic control and evolutionary conservation (by looking at corresponding mouse tissues).

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.spl.2008.01.003.

References

Boratyn, G.M., Datta, S., Datta, S., 2007. Incorporation of biological knowledge into distance for clustering genes. Bioinformation 1, 396–405.
Datta, S., Datta, S., 2005. Empirical Bayes screening (EBS) of many p-values with applications to microarray studies. Bioinformatics 21, 1987–1994.
Efron, B., Tibshirani, R., Storey, J.D, Tusher, V.G., 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Stat. Assoc. 96, 1151–1160.
Friedman, J., Meulman, J., 2004. Clustering on subsets of attributes (COSA). J. Royal Statist. Society. B 66, 1–25.
Gower, J., 1971. A general coefficient of similarity and some of its properties. Biometrics 27, 857–871.
Hastie, T., Tibshirani, R., Eisen, M.B., Alizadesh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P., 2000. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology 1, 1–21.
Ishwaran, H., Rao, J.S., 2003. Detecting differentially expressed genes in microarrays using Bayesian model selection. J. Amer. Stat. Assoc. 98, 438–455.
Ishwaran, H., Rao, J.S., 2005a. Spike and slab gene selection for multigroup microarray data. J. Amer. Stat. Assoc. 100, 764–780.
Ishwaran, H., Rao, J.S., 2005b. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann. Statist. 33, 730–773.
Ishwaran, H., Rao, J.S., Kogalur, U.B., 2006. BAMarray: Java software for Bayesian analysis of variance for microarray data. BMC Bioinform. 7, 59.
Joly, S., Le Calve, G., 1996. Similarity functions, in Classification and Dissimilarity Analysis. In: Lecture Notes in Statistics, Springer-Verlag, New York.
Kaufman, L., Rousseauw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York.
Papana, A., Ishwaran, H., 2006. CART variance stabilization and regularization for high-throughput genomic data. Bioinformatics 22, 2254–2261.
Storey, J.D., 2002. A direct approach to false discovery rates. J. Royal Statist. Soc. B. 64, 479–498.
Su, A.I., et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. 101, 6062–6067.
Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. 98, 5116–5121.