BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2231–2238
doi:10.1093/bioinformatics/btr363

Gene expression Advance Access publication June 17, 2011

Genomic data integration using guided clustering

Matthias Maneck[1], Alexandra Schrader[2], Dieter Kube[2] and Rainer Spang[1,*]

[1]Institut für funktionelle Genomik, Universität Regensburg and [2]Klinik für Hämatologie, Universitätsmedizin Göttingen, Germany

Associate Editor: David Rocke

ABSTRACT

Motivation: In biomedical research transcriptomic, proteomic or metabolomic profiles of patient samples are often combined with genomic profiles from experiments in cell lines or animal models. Integrating experimental data with patient data is still a challenging task due to the lack of tailored statistical tools.

Results: Here we introduce *guided clustering*, a new data integration strategy that combines experimental and clinical high-throughput data. Guided clustering identifies sets of genes that stand out in experimental data while at the same time display coherent expression in clinical data. We report on two potential applications: The integration of clinical microarray data with (i) genome-wide chromatin immunoprecipitation assays and (ii) with cell perturbation assays. Unlike other analysis strategies, guided clustering does not analyze the two datasets sequentially but instead in a single joint analysis. In a simulation study and in several biological applications, guided clustering performs favorably when compared with sequential analysis approaches.

Availability: *Guided clustering* is available as a R-package from http://compdiag.uni-regensburg.de/software/guidedClustering.shtml. Documented R code of all our analysis is included in the Supplementary Materials. All newly generated data are available at the GEO database (GSE29700).

Contact: rainer.spang@klinik.uni-regensburg.de

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

Received on March 16, 2011; revised on June 8, 2011; accepted on June 13, 2011

1 INTRODUCTION

Today, molecular profiling technologies allow the characterization of clinical biopsies which capture complementary cellular properties including genotype, methylation state, transcriptome, proteome and metabolome snapshots. Moreover, the same technologies are used in experiments with cell lines or primary cell cultures where they are complemented by chromatin immunoprecipitation assays like ChIP-seq, protein–protein interaction screens or protein localization assays. Each profiling technology sheds different, and partly complementary light on the functioning and malfunctioning of cells. However, their joint full potential can only be realized when the two information sources are combined. Thus, integrating data from different sources is an important part of modern biomedical research. An important aspect of data integration is the development of tailored statistical methods that are able to leverage knowledge contained within a diverse range of data sources and at the same time, being able to provide evidence to answer the types of question being posed by the research community as a whole. While the concept of statistical data integration is self evident, its realization in genomics is challenging. Obstacles include the heterogeneity of experimental setups, study designs, profiling platforms, sample handling and data management. Furthermore, missing metadata and insufficient documentation of heuristic and complex multistep analysis procedures complicate the endeavor.

Heterogeneities can in part be overcome by data integration on the level of gene lists [Beissbarth and Speed (2004), Pavlidis *et al.* (2002), Subramanian *et al.* (2005)]. All datasets are analyzed individually and gene sets or ranked list of genes are produced. In a subsequent meta-analysis step, the studies are integrated by detecting significant overlaps of gene-sets (Beissbarth and Speed, 2004), enrichment of genes sets in ranked lists (Subramanian *et al.*, 2005) or similarities of multiple ranked lists (Lottaz *et al.*, 2006). This strategy avoids the need for joint quantitative data models that describe the dependencies between individual quantitative profiles. A quantitative approach is the concatenation of feature vectors from different platforms in the context of classification problems. If several types of high-dimensional readouts are available for the same group of samples, predictive signatures can be constructed by combining selected features across all data types, thus exploring potential complementary information. Somewhat surprisingly, several authors observed only marginal improvements in classification accuracy resulting from data integration [Boulesteix *et al.* (2008), Lu *et al.* (2005)].

In complementary work, data integration has been used to aggregate information across clinical and experimental sample populations rather than platforms. Bild *et al.* (2006) combined data generated experimentally by overexpression of active oncogenes in non-malignant breast epithelial cells with tumor samples of various carcinomas. Signatures of pathway activation were learnt on the primary cell culture data and applied to tumor profiles for predicting pathway activation status, outcome and treatment efficiency of the cancer samples. The exact opposite sequential analysis strategy is described by Läuter *et al.* (2009). The authors start their analysis on the clinical data by identifying clusters of strongly coexpressed genes, which they subsequently test for joint differential expression between experimental conditions. Here, the experimental data cannot influence the formation, but the selection of gene clusters. The first analysis approach to combine both data sets from the outset is described in Bentink *et al.* (2008). Applying a class discovery method by von Heydebreck *et al.* (2001) in a semisupervised setting, the authors find classifications of patient

*To whom correspondence should be addressed.

samples based on coherently expressed genes that simultaneously separate experimental conditions.

Here, we complement and extend the approach of Bentink *et al.* We introduce *guided clustering* a new data integration strategy that combines experimental and clinical high-throughput data of possibly different genomic data types. Guided clustering is tailored to analysis scenarios, where the construction of a diagnostic signature is not driven by class labels on the clinical data, for instance disease types or clinical outcomes, but by a biological focus, for example the activity of a transcription factor or an entire pathway. The biological focus of the signature is established by a complementing experimental study e.g. a cell perturbation experiment. Guided clustering identifies sets of genes that stand out in the experimental data while at the same time display coherent expression in clinical data. Different to the semisupervised approach of Bentink *et al.*, guided clustering is unsupervised. However, the feature selection that drives the clustering of patient samples is guided by complementing experimental data. Moreover, it extends the framework of the previous method in that it can incorporate data from different genomic platforms and provide quantitative predictions of pathway activation. Further the balance between guiding and clinical data can be calibrated and tested explicitly.

In principal, guided clustering can be applied to any data integration problem that combines a sample clustering problem with a feature selection problem driven by a second dataset. Here, we report on two exemplary applications: (i) the prediction of transcription factor activity in clinical samples guided by a chromatin immunoprecipitation experiment and (ii) the prediction of pathway activity guided by cell culture perturbation experiments.

2 ALGORITHM

For clarity, we first describe guided clustering in the application context of oncogenic pathway activation in tumor samples, and later describe a series of modifications that adapt the method to different applications.

Let T_{ij} be a set of tumor expression profiles and G_{ij} a guiding dataset. Rows i denotes genes while columns j denotes samples. For simplicity, we assume that the same profiling platform is used and hence the same genes are monitored in both datasets.

The guiding dataset G consists of two types of samples: cell lines where a pathway is perturbed and the corresponding unperturbed control. Thus it is labeled, since we know in which profiles the pathway was experimentally perturbed. The label is stored in a binary vector L , where a one marks the profiles from perturbed cell lines and a zero control samples. We assume that the activity of the pathway varies across the tumors in T due to different genotypes. However, *a priori* we do not know in which tumors the pathway is active and to what extend. The dataset T is thus unlabeled. The activity of the pathway affects the expression of its target genes. In general, targets can be either induced or repressed by the pathway. Consequently, the guiding data contains genes that are upregulated in perturbed samples relative to controls and those that are downregulated. To simplify computation, we multiply the expression of pathway repressed genes by -1 in both datasets, such that numerically all targets display 'high' expression upon pathway activation.

2.1 Fusion of similarity matrices

Guided clustering starts by computing a matrix of pairwise gene similarities. The gene similarities are high if two genes are strongly correlated in T and simultaneously both genes are differentially expressed in G. This matrix fuses the information of two other matrices that capture gene similarity for both datasets separately. For two genes g and h, we define A_T a similarity matrix on T by the Gaussian smoothing kernel:

$$A_T(g,h)=\exp\left(\frac{-(1-\omega)d(g,h)^2}{2\sigma^2}\right), \tag{1}$$

where $d(g,h)=1-\max(\rho(g,h),0)$ and $\rho(g,h)$ the Spearman's correlation of the expression vectors of genes g and h. The bandwidth of the Gaussian smoothing function is specified by σ, and $\omega\in[0\ldots1]$ is a tuning parameter that will help us balance the information from both datasets. Parameter calibration will be discussed in Section 2.4. Note that anticorrelated pairs of genes are dissimilar. Similarly, we use G to define a diagonal similarity matrix

$$A_G(g,g)=\exp\left(\frac{-\omega d(g,L)^2}{2\sigma^2}\right) \tag{2}$$

where $d(g,L)=1-\varrho(g,L)$ and $\varrho(g,L)$ the Pearson's correlation. This matrix captures the correlation of gene expression with the class label vector L. High values correspond to genes that respond to pathway activation. Again σ is the bandwidth of the smoothing function and ω a weighting parameter that will be discussed later. The two individual matrices are then fused by a simple matrix multiplication:

$$W=A_G^{1/2}A_TA_G^{1/2} \tag{3}$$

W is a symmetric similarity matrix that holds high values only for pairs of genes that show consistent expression in T and simultaneously respond to pathway activation in G.

Figure 1 schematically explains the effect of matrix fusion. The points in the left panel show a set of genes embedded in the 2D plane. Their distance reflects similarity according to A_T. The gray tone encodes information from the guiding data: black points do not show differential expression in G while gray points are targets and the brighter the point the stronger the gene responds to pathway perturbation. Note that all genes fall into clusters. This is typical for expression data. Barely is any one gene regulated independently from any other genes. The right panel shows the same genes again. However this time, the distances are based on the fused similarity matrix. The dark points moved out of the clusters and distribute uniformly. The only remaining dense areas consist of bright genes that were already close to each other in the left panel. The effect of the matrix fusion can be viewed as a magnetic repulsion between genes. The less a gene responds to pathway activation, the stronger is its repulsion from all other genes. Genes remaining in clusters are potential pathway target genes that are consistently regulated in tumors. We will use their consensus expression as a surrogate for pathway activity in tumors. Figure 1 shows that the matrix fusion induces similarities that makes clustering genes a hard problem, since many genes are no longer in clusters but on their own. However, we do not need to assign all genes to clusters. Instead we aim to detect the top most densest modules of genes thus leaving the majority of genes unassigned to any cluster.

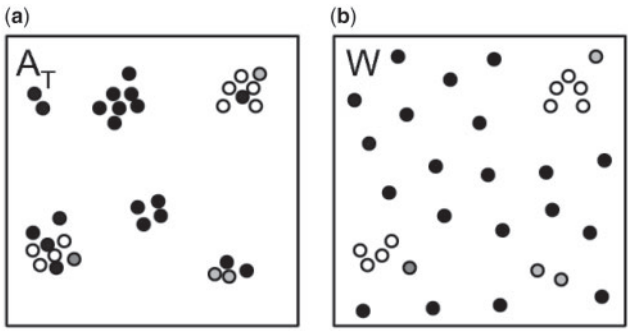

Fig. 1. Matrix fusion induces a magnetic repulsion of genes: each point represents one gene. The distance between points reflects the similarity of genes while gray shades represent the genes response to pathway perturbation. The brighter a point, the stronger the gene responds to the perturbation. (**a**) Gene similarities based on coexpression in the tumor data only. All genes fall in clusters, since genes are regulated in concert. (**b**) The same genes as in (a), but distances are based on the fused similarity matrix W. Genes that do not respond to the pathway moved out of the clusters and distribute uniformly across the plane.

2.2 Extraction of tight expression modules

Along the lines of kernel density estimation, we calculate the neighborhood density for each gene as

$$K(g)=\sum_{i=1}^{n} W_{g,i} \tag{4}$$

where n is the total number of genes in the dataset. A gene g with a high value $K(g)$ is located in a large and dense cluster. Guided clustering starts by selecting the gene g_0 that maximizes $K(g)$ as a seed gene. Next a module of genes C is grown around g_0 using average linkage by iteratively adding genes g_k that maximize

$$\gamma(g_0,g_1,\ldots,g_{k-1},g_k)=\frac{\sum_{i,j\leq k} W_{g_i,g_j}}{|C|+1} \tag{5}$$

where $|C|$ is the number of genes in C. The iteration is terminated, if no gene g_k exists, such that $\gamma(g_0,g_1,\ldots,g_{k-1},g_k) > \gamma(g_0,g_1,\ldots,g_{k-1})$. In case we want to extract more than one dense cluster, we remove all genes selected in the current iteration, recompute $K(g)$ and proceed as described above.

2.3 Condensing the joint expression of genes in a module to a consensus expression index

By construction, the expression levels of genes in a module are tightly correlated across tumor samples. In any tumor, they are either unanimously up- or downregulated. This allows us to condense their expression into a single number per tumor. This index can be used as a surrogate for pathway activity in the tumor. A high index points to an active pathway and a low index to an inactive pathway. Although genes in a module correlate strongly, their raw expression can deviate highly due to scale differences. A module can be composed of genes, where some have a greater expression level than others. To compute the consensus index, we use a standard additive model that accounts for scale differences on the log scale:

$$y_{ij}=\mu+\alpha_i+\beta_j+\epsilon_{ij}, \tag{6}$$

where y_{ij} is the observed expression of a gene, α_i a gene-specific scale coefficient, β_j the sample specific index of pathway activation and ϵ_{ij} the error term. We fit the additive model using Tuckey's robust median polish procedure (Hoaglin *et al.*, 1977). In the following, we will refer to the index as pathway activation index (PAI). Note that our approach for summarizing the expression of module genes to a PAI is identical to a method used to compute probe-set summaries in the popular normalization package RMA for Affymetrix gene chips.

2.4 Balancing both datasets

Guided clustering uses the parameter ω to balance the influence between both datasets. While the tumor data are driving within cluster strength, the guiding data forces genes to have a high response to perturbation. ω shifts the focus from the clinical data ($\omega=0$) to the guiding data ($\omega=1$). *Guided clustering* evaluates within cluster strength by the average pairwise correlation:

$$\phi(\omega)=1/|C_\omega|^2 \sum_{g,h\in C_\omega} \max(\rho(g,h),0)$$

where C_ω is a gene cluster retrieved for a specific ω. The response to perturbation is assessed by the average gene activation:

$$\varphi(\omega)=1/|C_\omega| \sum_{g\in C_\omega} \varrho(g,L).$$

We calculate $\phi(\omega)$ and $\varphi(\omega)$ for $\omega\in[0,0.1,0.2,\ldots,1]$. Since $\phi(\omega)$ and $\varphi(\omega)$ are on different scales, we rescale them such that they both range form 0 to 1. The algorithm chooses the ω that maximizes the sum $\phi(\omega)+\varphi(\omega)$.

The smoothing parameter σ specifies the bandwidth of the smoothing kernel and influences the sensitivity of the method. Larger bandwidths result in larger clusters that may include genes with low responses to perturbation. Smaller bandwidth enforce more rigorous restrictions on the genes from the guiding data. We recommend to tune σ manually starting from a large value and decreasing it in several steps while at the same time monitoring the cluster tightness and the distribution of perturbation responses in the guiding data. For our analysis, we varied σ between $1/3$ and $0.1/3$ in steps of $0.1/3$. In the next section, we give an example of parameter tuning.

2.5 Extensions to other experimental settings

So far, we have discussed guided clustering in the context of a pathway perturbation experiment with a labeled guiding dataset G. It can be easily adapted to other application scenarios by tailoring the similarity function in Equation (2) to the application. The similarity values need to quantitatively rank genes that should be preferentially used to build gene clusters. The strongest preference possible is encoded by a value of 1. Smaller values gradually reduce the influence of a gene. The preference scores need to be calculated from guiding data. Many application settings are possible. The preference score can e.g. reflect the connectivity of a gene in a protein–protein interaction network, thus guiding the formation of gene clusters seeded around hub genes. They can also reflect the binding abundance of a transcription factor assessed in a chromatin immunoprecipitation experiment, thus guiding the gene clusters to be build from targets of a specific transcription factor. We will demonstrate the use of guided clustering in this context in Section 4.1

2.6 Runtime

To analyze the runtime of the *guided clustering* algorithm, we dissect the algorithm into its three main parts: (i) the calculation of the pairwise gene distance matrix has a runtime of $O(n^2)$, where n is the number of genes. (ii) Transformation of the pairwise gene distances into affinities needs additional $n_\omega O(n^2)$ operations, where n_ω is the number of values used for ω when choosing the optimal weighting between both datasets. (iii) Extracting k gene modules has a complexity of $kO(n)$. The total complexity $O(n^2)+n_\omega O(n^2)+kO(n)=O(n^2)$ is dominated by the number of input genes (Supplementary Fig. 1).

All calculations have been performed on a machine containing 16 Quad-Core AMD Opteron 8354 processors with 2.2 GHz each and 132 GB main memory. At the current state of development, *guided clustering* is a single thread method using 1 of 16 processors available. For the analysis of simulated datasets used in Section 3, each run needed 13 s on average. Analysis of the lymphoma samples together with BCL6 and LPS data in Sections 4.1 and 4.2 took about 890 and 902 s, respectively.

3 SIMULATION-BASED VALIDATION AND COMPARISON TO COMPETING APPROACHES

Prior to testing guided clustering in real data integration contexts, we study its performance on data that is artificially generated and fulfills the underlying assumptions of our algorithm. This allows us to better understand its limitations alongside those of competing strategies. In simulations, the data generating process defines a ground truth, against which any analysis result can be evaluated. In real applications, we often do not have a ground truth result. Moreover, focused simulations allow us to study individual difficulties in the analysis independently from each other, while real data usually comprises many of them in parallel. Finally, the difficulty of clustering problems can be scaled freely in simulations. Here, we compare guided clustering to the two competing sequential analysis concepts described in the literature, which select genes sets only using the clinical or the guiding data, respectively.

We simulate artificial data that mimics the application of guided clustering in the context of pathway activation prediction.The data consist of an artificial clinical dataset T with 80 samples and a guiding dataset G with 20 control and 20 *perturbed* samples. The datasets hold 1500 features. Both datasets are generated by adding a signal component and a noise component ϵ_{ij}:

$$T_{ij}=F_{ij}+\omega_T\epsilon_{ij}, \text{ and } G_{ij}=I_{ij}+\omega_G\epsilon_{ij}.$$

F_{ij} and I_{ij} are the signal components that contain the ground truth of simulated effects, and the tuning parameters ω_T and ω_G are used to calibrate the signal to noise ratio. The noise component ϵ is simulated using a multivariate normal distribution with a block structured covariance matrix following Guo *et al.* (2007).

For T_{ij}, we generate signals in three clusters $E_1,\ldots,E_3$ of 200 features. To generate traces of pathway activity, we draw for each gene in a cluster a random number α_i uniformly from the interval $[0,1]$, which represents the strength with which the gene responds to pathway activation. Moreover, for every sample we draw a uniformly distributed random number β_j from $[-1,1]$ which represents the strength of the pathway activation in this sample. F_{ij} is then set to $\alpha_i+\beta_j$. For genes that do not fall in any of the three clusters, F_{ij} are set to zero.

The simulation of the guiding data G_{ij} includes a set B_d of 600 responding genes. For each of these, we draw a random number γ_i uniformly from $[0,1]$ and set $I_{ij}=-\gamma$ for control samples and $I_{ij}=\gamma$ for perturbation samples. For the remaining genes, we set I_{ij} to zero. The size of the intersection of the three clusters E_i with the set of responding genes varies across clusters (200 for E_1, 100 for E_2, and 50 for E_3). More details are given in the Supplementary Materials. The simulation setup is summarized in Figure 2a. Note that we not only simulate signals and noise, but also confounding structures including clusters of correlated genes in T that do not correspond to induced genes in G, as well as induced genes in G that do not form tight clusters in T.

The goal of guided clustering is to reconstruct the hidden pathway activation signal β_j for all samples and all three clusters accurately. We ran guided clustering on this series of simulated datasets with increasing difficulty by varying the signal-to-noise ratio of T between 0.5 and 2 in steps of 0.1. The signal-to-noise ratio for G is kept constant at 2. To evaluate accuracy of the estimated signals, we calculated the maximal correlation of the top three estimated pathway indices $\hat{\beta}_j$ to any of the β_j underlying the simulation.

We give a detailed description of analysis results obtained for a signal-to-noise ratio of 1. Analysis starts with calibrating the parameter σ such that the resulting clusters are strongly enriched for genes with strong signals from the guiding data. We start with a large value of σ and consecutively lower it while monitoring the distribution of signals from the guiding data. Figure 2b shows that with $\sigma=0.23$, a good balance was found. σ was kept constant for all other analyzes.

A characteristic feature of guided clustering is the simultaneous use of both datasets. In order to evaluate its benefits, we compare guided clustering to the two possible sequential analysis concepts, namely gene selection based on experimental data followed by pathway activity prediction on the clinical data as described by Bild *et al.* (2006), and gene selection via identification of strongly correlated gene-sets in the clinical data followed by multivariate tests in the experimental data as described by Läuter *et al.* (2009).

Figure 2d shows the results for estimating the pathway indices of the three simulated effects (E_1 solid, E_2 dotted, E_3 dashed line). *Guided clustering* reconstructs the pathway indices correctly for signal-to-noise ratio ≥ 1. However, the smaller the overlap between the correlated clusters in T and the differentially expressed genes in G, the more difficult the reconstruction. The approach by Läuter *et al.* reconstructs the signals perfectly since we provided it with the correct gene-sets. Using the approach of Bild *et al.* results in a poor reconstruction quality. The method is only able to coarsely reconstruct the cluster with the biggest overlap. We believe that the ignorance of the correlation structure in T that this method exercises when constructing gene-sets compromises its performance in signal reconstruction.

Additionally, we tested the gene-sets for joint differential expression on G. Here, the approach of Läuter *et al.* (2009) only identifies effects in gene-sets where the majority of contributing genes are differently expressed within the guiding data. It does not filter non-responding genes since it does not access the guiding data during gene-set formation. *Guided clustering* reliably detected gene-sets with differential expression in G. The approach of Bild *et al.*

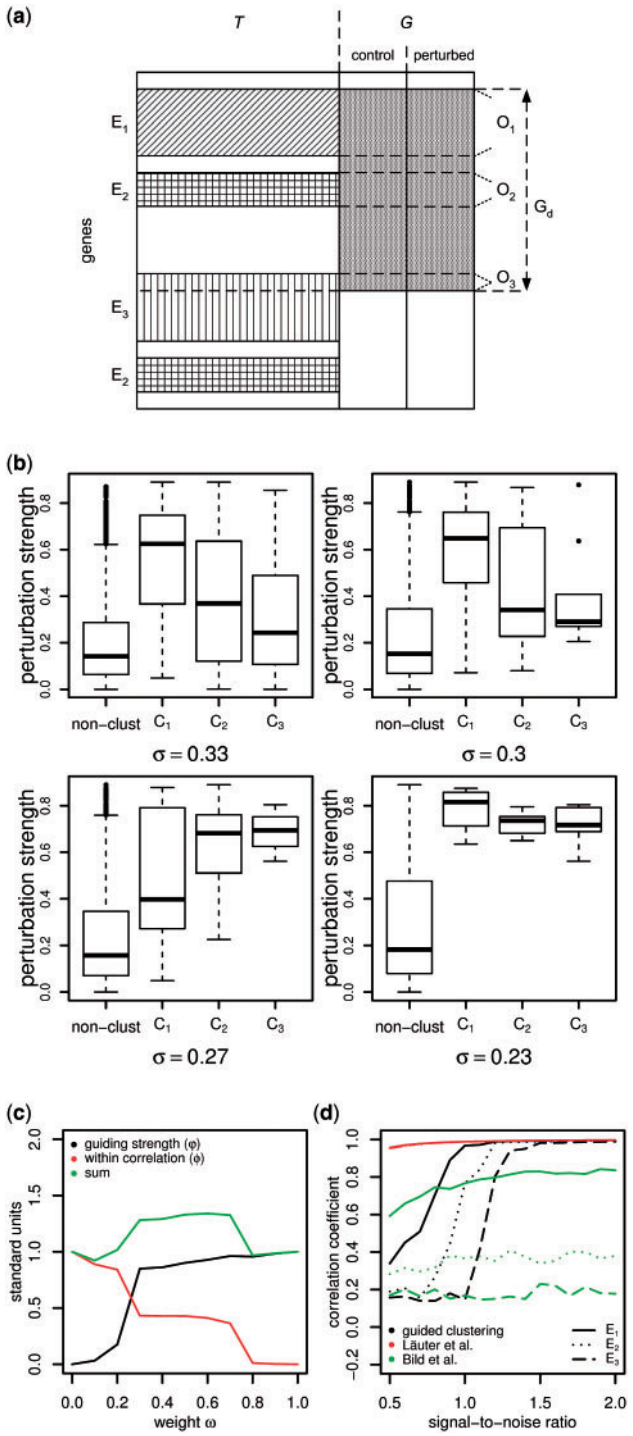

Fig. 2. (**a**) Structure of the simulated data. The three clusters with non-zero signals F_{ij} in T are named E_1, E_2 and E_3. Each effect overlaps with a certain number of differentially expressed genes in $B_d \subset G$. The overlaps are named O_1, O_2 and O_3. (**b**) Guiding strength of top three clusters extracted with different choices of σ. (**c**) Trade-off between within cluster correlation and guiding strength for the selection of ω. (**d**) Accuracy of estimated effects depending on the signal-to-noise ratio with $\sigma \simeq 0.23$: black, guided clustering; red, Läuter *et al.*; green, Bild *et al.*

directly uses the top differentially expressed genes in G. The results are summarized in Supplementary Table S1.

4 APPLICATIONS

Guided clustering identifies sets of genes that stand out in the experimental data while at the same time display coherent expression in clinical data. This general setting allows for the application of guided clustering in many different situations. Here, we choose two of them: (i) the detection of functional targets of a transcription factor and the quantitative estimation of its activity in individual clinical samples; (ii) the estimation of pathway activities in tumor samples as introduced by Bild *et al.* (2006) using a cell line stimulation experiment.

4.1 Guided clustering identifies BCL6 activity as a strong prognostic factor in diffuse large B-cell lymphomas and suggests a link between BCL6 and the activity of Toll-like receptor pathways

BCL6 is a transcription factor that predominantly acts as a repressor of transcription. In mature B-cells, BCL6 is required for the formation of germinal centers (GC). Hence, its function is critical for the working of the acquired immune system. However in diffuse large B-cell lymphomas (DLBCL), BCL6 is frequently translocated, hypermutated and hypermethylated. Its dysfunctional activity is assumed to contribute to oncogenesis in a subset of DLBCL and potentially also in different malignancies (Iqbal *et al.*, 2007). DLBCL is a morphologically, genetically and clinically heterogeneous lymphoma entity (WHO, 2008) with several defined subentities (Bentink *et al.*, 2008; Rosenwald *et al.*, 2002) and most likely more. Whether differential BCL6 activity contributes to the heterogeneity is not fully known.

Recently, the influence of BCL6 on the transcriptional program of B cells was investigated (Ci *et al.*, 2009). The authors performed a ChIP-on-chip screen with primary human GC B-cells and the DLBCL cell lines OCI-Ly1 and OCI-Ly7 using high-density oligonucleotide promotor microarrays. These data were combined with gene expression profiles from DLBCL. The authors followed a strictly sequential analysis strategy: first only the ChIP data were used to identify distinct groups of genes bound by BCL6 exclusively in GC B cells, DLBCL or both. Second, a look up of these genes in the tumor data revealed that a large number of BCL6 target genes are silenced in primary human centroblasts compared with naïve B cells. But only half of those genes are also silenced in DLBCL. Additionally, BCL6 target genes have a lower expression in GC B cell-like (GCB) lymphomas compared with activated B cell-like (ABC) lymphomas. DLBCL and its subentities GCB and ABC were analyzed as lymphoma populations. The variability of BCL6 target expression across individual lymphomas was not addressed. In fact, we observed that the expression of the top ranking genes in DLBCL is rather disperse (Fig. 3a).

This observation is not surprising: genome-wide chromatin immunoprecipitation assays identify binding sites of transcription factors in the neighborhood of genes and thus generate lists of potential targets of the transcription factor. Clearly, binding does not imply regulation. In general, several regulators and cofactors are needed for transcription and their presence depends on the cellular context. However, if a transcription factor like BCL6 actively

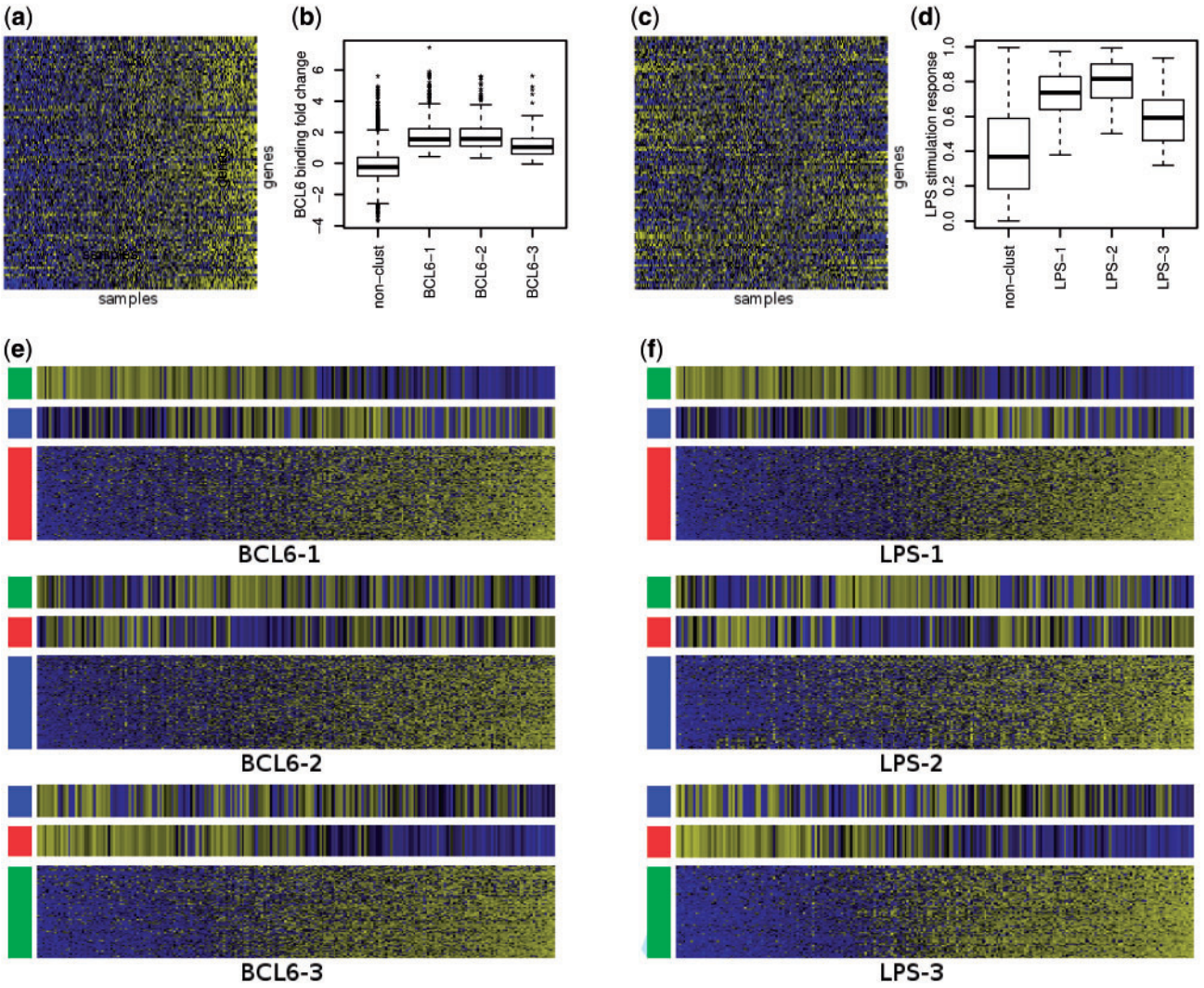

Fig. 3. (**a** and **c**) Gene expression of top 100 genes with highest BCL6 binding fold change (a) or LPS activation (c), respectively. Yellow indicates high and blue low expression. The samples were ordered according to their mean expression. (**b** and **d**) BCL6 binding fold change (b) or LPS activation (d) of the extracted gene clusters compared with non-cluster genes. (**e** and **f**) Gene expression of extracted PAI gene clusters across the lymphoma samples (first, red; second, blue; third, green). Samples are ordered increasingly with respect to the PAI. Yellow indicates high and blue low expression. On top of each gene cluster, the two other PAIs are shown.

regulates a module of target genes in a defined cellular context like the DLBCL context, its functional targets should display correlated expression across clinical samples. The heterogeneity of DLBCL poses an additional problem, since it is not clear whether there is only a single 'DLBCL context'. The expression and activation of transcriptional co-regulators might vary across DLBCL and we might find multiple modules of functional BCL6 targets, each expressed in different subsets of DLBCL.

Using *guided clustering*, we integrated BCL6 ChIP-on-chip data of GC B cells from Ci *et al.* (2009) GEO accession: GSE15179 with 220 expression profiles from DLBCL and Burkitt lymphoma samples from Hummel *et al.* (2006) GEO-accession: GSE4475. Affimetrix gene expression data were normalized using the variance stabilization method (Huber *et al.*, 2002) and probe-sets were summarized to gene expression values by fitting a standard additive model, employing Tuckey's median polish algorithm (Hoaglin *et al.*, 1977). For the ChIP-on-chip data, log2 ratios were averaged across all three replicates after truncating ratios above the 95% quantile of all positive log2 ratios (≈ 2.67). Locus by locus, we subtracted the log2 ratios from their maximum across samples and fed these values directly into the guided clustering algorithm. BCL6 binding loci were matched to HGU133a probe-sets using the accession numbers and refseq ids provided by the authors. Multiple probe-sets for the same locus were summarized using the median polish algorithm. After matching, the dataset consisted of 9648 genes and 220 samples.

On this data, we ran *guided clustering* and identified the top three modules of BCL6 targets each expressed predominantly in a different subset of lymphomas. The smoothing parameter was selected as described in Section 2.4 and set to $\sigma = 0.23$.

Figure 3e shows the expression of the extracted gene modules across lymphomas. Note that the genes are coherently expressed across lymphomas and there is a continuous gradient when lymphomas are arranged by increasing BCL6 indices. Figure 3b shows that these genes also have strikingly large log2 ratios, indicating that BCL6 indeed binds their promotor. Together, both plots confirm that the two datasets are in good balance. In agreement to the observations of Ci *et al.* (2009), the second extracted index (BCL6-index2) is higher in ABC than GCB type DLBCL ($P < 10^{-9}$, *t*-test).

To test for a clinical impact of BCL6 activity in DLBCL, we fitted a cox proportional hazard model including the BCL6-index from module 2 as a continuous covariate and established categorical prognostic factors like the ABC/GCB status, age >59 years and Ann Arbor stage. Survival analysis was restricted to a group of patients that received identical treatment [a combination of chemotherapy based on cyclophosphamide, doxorubicin, vincristine and prednisone (CHOP) or similar]. This was the case for 80 lymphoma patients in the study. We found the BCL6-index2 to be a significant independent predictor of survival ($P < 10^{-5}$). Patients with a high BCL6-index2 have a better outcome than patients with a low index. Notably, the hazard associated with the BCL6-index2 is higher than all other factors including the ABC/GCB status (Supplementary Table S2).

The BCL6-index2 accumulates the expression of 335 genes including several BCL6 targets that were also described in the primary analysis of the dataset by Ci *et al.* (2009) (Supplementary Table S4). We analyzed this gene-set for enrichment of genes involved in certain aspects of B-cell functionality or malignant transformation using the Gene Set Analysis Toolkit V2 by Duncan *et al.* (2010). Genes involved in Toll-like receptor signaling were significantly enriched ($P < 0.004$, hypergeometric test); similarly, we found enrichment of Jak-STAT signaling genes ($P < 0.001$, hypergeometric test). These observations support the findings of Basso and Dalla-Favera (2010) who hypothesize that BCL6 modulates signaling through Toll-like receptors.

4.2 LPS mediated Toll-like receptor signaling and BCL6 targets are coherently expressed in DLBCL

To support the hypothesis of Basso and Dalla-Favera (2010) experimentally, we stimulated cells from the BL-2 lymphoma cell line with lipopolysaccharide (LPS) for 6 h, thus stimulating Toll-like receptor signaling. Expression profiles of stimulated cells were compared to control profiles of unstimulated BL-2 cells. The expression profiles were generated on Affymetrix HGU133plus2 GeneChips and were normalized as described above and matched to both the lymphoma and BCL6 datasets through the accession numbers from Ci *et al.* (2009). Altogether six samples were hybridized, three independent biological replicates in each group. A detailed description of the sample preparation can be found in the Supplementary Material. These expression profiles are available at the GEO database (GSE29700).

We used guided clustering for a joint analysis of our stimulation data and the lymphoma dataset by Hummel *et al.* (2006). This is a typical application of guided clustering in the context of integrating experimental cell perturbation data and clinical expression studies. In contrast to ChIP assays, cell perturbation experiments can identify functional targets of signaling pathways. However, they are not confined to direct targets. Transcriptional regulation is context specific and the molecular contexts of a cell culture significantly differs from that of a tumor. Nevertheless, if genes whose expression respond in the cell culture context also display a coherent expression across patient profiles, it is likely that their consensus expression reflects the activity of this pathway in individual patient probes (Bentink *et al.*, 2008). We applied guided clustering to identify transcriptional modules that are conserved between both cellular contexts. The smoothing parameter was selected as described in Section 2.4 and set to $\sigma = 0.17$. As in the BCL6 analysis, we extracted the top three modules and examined them for cluster tightness in the lymphoma data and differential expression in the guiding data. Figure 3f shows heatmaps of the extracted gene modules on the clinical data. The genes are coherently expressed across the lymphomas and form a continuous gradient when the lymphomas are arranged by increasing LPS indices. The distribution of correlations to the class label vector for module genes and non-module genes are shown in Figure 3d. The module genes stand out and are clearly enriched for LPS stimulation.

Strikingly, the LPS-index2 and our BCL6-index2 although derived from completely different guiding datasets and only intersecting in 73 of 198 genes (Supplementary Table S5) are almost perfectly correlated ($r > 0.98$). This further supports the hypothesis that BCL6, in fact, modulates Toll-like receptor signaling in DLBCL. Moreover, our analysis captures this link quantitatively. The higher a lymphoma expresses direct BCL6 targets, the higher it also expresses LPS-inducible genes. While many explanations for this correlation can be taken into account, the easiest of them is the modulation of Toll-like receptor signaling through BCL6.

5 CONCLUSION

We introduce guided clustering, a new method for the combined analysis of clinical microarray gene expression data and experimental data. Our method is controlled by two parameters ω and σ. While ω is selected automatically, σ has to be specified by the user. With the help of these parameters, guided clustering circumvents the need for crisp cutoffs of gene lists and clusters used by most competing methods. Even though both parameters are chosen based on the training data, they are not tuned to enhance any fit. Their purpose is to balance the influence of both datasets. In fact, we did not observe any overfitting phenomenon in simulations with separate test and training data (data not shown). Guided clustering is flexible and can be adapted to various data integration challenges. Here, we applied it in the context of a study on DLBCL that was guided to focus on aspects of BCL6 and Toll-like receptor signaling. We could establish a novel prognostic index in DLBCL, which holds more prognostic information than existing predictors of survival. The composition of the genes underlying this index point to a link between BCL6 activity and Toll-like receptor signaling. We experimentally strengthened this link by an LPS stimulation experiment combined with a second guided clustering analysis. We observed that targets of LPS-mediated Toll-like receptor signaling, and BCL6 targets are coherently expressed in a large collection of DLBCL, suggesting that BCL6 in fact influences the transcriptional program by Toll-like receptor signaling in DLBCL.

ACKNOWLEDGEMENTS

We would like to thank Stefan Bentink for fruitful discussions and Claudio Lottaz and Tully Ernst for carefully proofreading the manuscript. Further, we are grateful to Dido Lenze and Michael Hummel (Institute for Pathology, Campus Benjamin Franklin, Charité Berlin, Germany) for carrying out the hybridization of the LPS microarray data.

Funding: BMBF network Hematosys (BMBF Nr: 0315452E and 0315452H); Deutsche Forschungsgemeinschaft (GRK1034, FOR942); Bavarian Genome Network BayGene.

Conflict of Interest: none declared.

REFERENCES

Basso,K. and Dalla-Favera,R. (2010) BCL6: master regulator of the germinal center reaction and key oncogene in B cell lymphomagenisis. *Adv. Immunol.*, **105**, 193–210.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Bentink,S. *et al.* (2008) Pathway activation patterns in diffuse large B-cell lymphomas. *Leukemia*, **22**, 1746–1754.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Boulesteix,A.-L. *et al.* (2008) Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, **24**, 1698–1706.

Ci,W. *et al.* (2009) The BCL6 transcriptional program features repression of multiple oncogenes in primary B cells and is deregulated in DLBCL. *Blood*, **113**, 5536–5548.

Duncan,D. *et al.* (2010) WebGestalt2: an updated and expanded version of the web-based gene set analysis toolkit. *BMC Bioinformatics*, **11** (Suppl. 4), P10.

Guo,Y. *et al.* (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hoaglin,D.C. *et al.* (eds) (1977) *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York, USA.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1) , S96–S104.

Hummel,M. *et al.* (2006) A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.*, **354**, 2419.

Iqbal,J. *et al.* (2007) Distinctive patterns of BCL6 molecular alterations and their functional consequences in different subgroups of diffuse large B-cell lymphoma. *Leukemia*, **21**, 2332–2343.

Läuter,J. *et al.* (2009) High-dimensional data analysis: selection of variables, data compression and graphics–application to gene expression. *Biometr. J.*, **51**, 235–251.

Lottaz,C. *et al.* (2006) OrderedList–a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*, **22**, 2315–2316.

Lu,L.J. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.

Pavlidis,P. *et al.* (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, **9**, 401–411.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

von Heydebreck,A. *et al.* (2001) Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, **17** (Suppl. 1), S107–S114.

WHO (2008) *WHO Classification of Tumors of Haematopoietic and Lymphoid Tissues*. World Health Organization, WHO Press.