Published online 22 June 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 20* ***6831–6840***
doi:10.1093/nar/gkq550

Predictive response-relevant clustering of expression data provides insights into disease processes

Lisa E. M. Hopcroft[1], Martin W. McBride[2], Keith J. Harris[1], Amanda K. Sampson[2], John D. McClure[2], Delyth Graham[2], Graham Young[3], Tessa L. Holyoake[3], Mark A. Girolami[1,*] and Anna F. Dominiczak[2,*]

[1]Inference Group, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, [2]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA and [3]Paul O'Gorman Leukaemia Research Centre, Section of Experimental Haematology, Faculty of Medicine, University of Glasgow, Gartnavel General Hospital, 1053 Great Western Road, Glasgow G12 0YN, UK

Received January 27, 2010; Revised May 25, 2010; Accepted May 29, 2010

ABSTRACT

This article describes and illustrates a novel method of microarray data analysis that couples model-based clustering and binary classification to form clusters of 'response-relevant' genes; that is, genes that are informative when discriminating between the different values of the response. Predictions are subsequently made using an appropriate statistical summary of each gene cluster, which we call the 'meta-covariate' representation of the cluster, in a probit regression model. We first illustrate this method by analysing a leukaemia expression dataset, before focusing closely on the meta-covariate analysis of a renal gene expression dataset in a rat model of salt-sensitive hypertension. We explore the biological insights provided by our analysis of these data. In particular, we identify a highly influential cluster of 13 genes—including three transcription factors (*Arntl*, *Bhlhe41* and *Npas2*)—that is implicated as being protective against hypertension in response to increased dietary sodium. Functional and canonical pathway analysis of this cluster using Ingenuity Pathway Analysis implicated transcriptional activation and circadian rhythm signalling, respectively. Although we illustrate our method using only expression data, the method is applicable to any high-dimensional datasets. Expression data are available at ArrayExpress (accession number E-MEXP-2514) and code is available at http://www.dcs.gla.ac.uk/inference/metacovariateanalysis/.

INTRODUCTION

Microarray experiments allow the simultaneous expression measurements of tens of thousands of genes in a biological sample and have been employed extensively to investigate human disease since the early 90s (1). Despite almost two decades of research, challenges regarding the analysis of these data remain. Typically, the number of variables (or probes) measured vastly outnumbers the number of replicate experiments: over 30 000 probes might be measured in only three or four samples, making good predictive performance possible by chance, irrespective of whether the data contain relevant patterns. In addition, many variables will exhibit similar patterns across the samples; we require methods that identify which of these correlations are the result of genuine functional relationships and/or co-regulation and which are merely observed by chance. Taken together, these features make microarray analysis statistically demanding, prone to variability in model parameter estimates and ultimately susceptible to inaccurate prediction.

Our meta-covariate method is a novel approach to analysing microarray data, which overcomes and, in the case of correlated expression patterns, exploits the statistical properties of gene expression data, with a view to improving prediction and identifying biologically relevant structure in the data (2). It is, however, applicable

*To whom correspondence should be addressed. Tel: +44(0) 141 330 1623; Fax: +44(0) 141 330 2673; Email: girolami@dcs.gla.ac.uk
Correspondence may also be addressed to Anna F. Dominiczak. Tel: +44(0) 141 330 2738; Fax: +44(0) 141 330 6997; Email: A.Dominiczak@clinmed.gla.ac.uk

The authors wish it to be know that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

to any high-dimensional dataset (including proteomics, sequencing and miRNA datasets) where informative correlations exist between the variables. Initially, the D probes are grouped into K clusters, using gene expression similarity across the N samples and a standard Gaussian mixture model. An N-dimensional meta-covariate vector is then generated from each cluster and predictions are made by weighting these meta-covariates in a probit regression model. We then take the novel step of using the prediction performance to update the clustering structure, the meta-covariates and the regression weights. This iterative procedure is repeated until convergence (Figure 1).

The approach of reducing microarray data dimensionality by forming clusters (independent of predictions) and making subsequent predictions using cluster summaries has been adopted previously by Hanczar *et al.* (3) and Park *et al.* (4) amongst others. Where our method improves upon existing methods is that inter-predictor correlations are coupled with predictor-outcome correlations to inform the clustering structure, the cluster summaries (meta-covariates) and the regression weights (indicated by the turquoise arrow in Figure 1). The advantages of our method are three-fold. First, the clustering component of the model identifies response-relevant structure in the data, aiding biological interpretation. Second, the regression coefficients allow the identification of influential clusters: the greater the weight assigned to a cluster in the regression model, the more 'informative' that cluster is when discriminating between the outcomes of the response variable. And finally, using the predictor-response correlations to fine-tune the clustering structure in the model potentially improves prediction performance.

In this article, we will first demonstrate how the meta-covariate method works using the well-known leukaemia dataset described by Golub *et al.* (5). We will then employ the method to analyse gene expression data in the rat kidney to investigate the genetics of salt-sensitive hypertension.

MATERIALS AND METHODS

Leukaemia data

In the Golub *et al.* (5) dataset, bone marrow or peripheral blood samples were taken from 25 acute myeloid leukaemia (AML) and 47 acute lymphoid leukaemia (ALL) patients. The training data contain 38 samples, of which 11 are AML and 27 are ALL samples. The test data contain 34 samples, of which 14 are AML and 20 are ALL. RNA extracted from these samples was tagged and subsequently hybridized to a high-density Affymetrix oligoneuclotide microarray (Hu6800/HuGeneFL). The expression data were obtained from the Broad Institute Website and pre-processed as recommended in Dudoit *et al.* (6) (see Supplementary Data for details), leaving 3571 probes for analysis.

Animal strains

Inbred colonies of SHRSP and WKY have been maintained at the University of Glasgow since 1991, as previously described (7). From weaning, all rats were maintained on normal rat chow (rat and mouse No.1 maintenance diet, Special Diet Services) and at 18 weeks of age rats were given a salt challenge (1% NaCl in drinking water) for 3 weeks.

The congenic strain SP.WKYGla2a (D2Rat13-D2Rat157) was generated using a marker-assisted 'speed' congenic strategy (8) where a WKY (donor strain) segment was introgressed into the SHRSP (recipient strain) genetic background.

Microarray data analysis Affymetrix Rat 230-2

Affymetrix GeneChip renal expression analysis was used to identify differentially expressed probe sets (representing

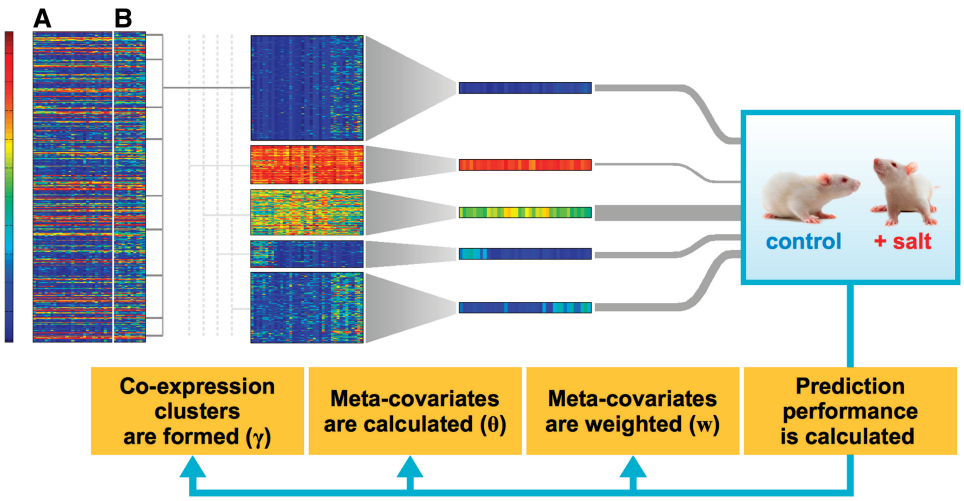

Figure 1. The meta-covariate method. Expression data are used to form clusters of probes (clustering is represented by the $D \times K$ matrix of responsibilities γ). N-dimensional meta-covariates ($\boldsymbol{\theta}_k$) are calculated from these clusters and used to make predictions in a probit regression model (with regression coefficients w). The novelty of our method is highlighted in turquoise: the prediction performance is used to update γ, $\boldsymbol{\theta}_k$ and w, thereby iteratively improving the cluster structure and the prediction performance.

unique gene or expressed sequence tag sequence on the Affymetrix GeneChip) between male, 21-week old, age-matched salt-loaded and non-treated SHRSP, SP.WKYGla2a, and WKY rats. Whole kidneys (harvested between 10 a.m. and 12 noon and snap-frozen in liquid nitrogen) were homogenized and total RNA extracted from three rats from each strain by using the maxi RNeasy kit according to the manufacturers protocol (Qiagen). Biotinylated amplified target cRNA was prepared and hybridized to the Affymetrix Rat Rat230-2 gene chips as described by Affymetrix. After hybridization, microarray chips were washed, stained and scanned. The microarray data were normalized in R using RMA (9). To assess the statistical significance of pairwise intergroup differences, Rank products (RP) (10) was used, corrected for multiple testing using a false discovery rate of 5% (11). The microarray dataset has been submitted to ArrayExpress (Accession Number E-MEXP-2514).

Quantitative real-time polymerase chain reaction

Renal total RNA was extracted from 21-week-old salt-loaded and non-treated male rats using RNeasy kits (Qiagen), treated with DNase-Free RNase (Ambion), and accurately quantified. qRT–PCR was performed using Taqman (Applied Biosystem, UK) *Actb* (β-actin) labelled Vic, as a normalization control and either *Arntl* (Rn00577590_m1), *Npas2* (Rn01438224_m1), *Nfil3* (Rn01434874_s1) and *Bhlhe41* (Rn00591084_m1) labeled FAM. *Arntl*, *Npas2*, *Nfil3* and *Bhlhe41* were normalized to *Actb*, expressed relative to SHRSP (non-salt treated) in each sample using the comparative ($\Delta\Delta CT$) method.

Description of method

As described in the 'Introduction' section, the novelty of this method lies in the coupling of the clustering and prediction components (as depicted by the turquoise arrow in Figure 1). These components are coupled by optimising all the parameters (i.e. the parameters pertaining to both components) simultaneously, rather than optimising the clustering parameters before the prediction parameters. Here, we have chosen a Gaussian mixture model as the clustering model (12, Section 9.2) and probit regression (12, Section 4.3.5) as the prediction method. We optimize the parameters of these models using the Expectation-Maximisation (EM) algorithm (12, Section 9.4), which finds the most likely parameter estimates in a probabilistic model by updating them over a number of iterations. Our model updates are derived by merging the standard EM updates for the clustering and regression parameters.

Intuitively, the meta-covariate model can be thought of as follows: (A) all probes on the array are clustered into K groups, and each group is then represented by some average of its members; (B) these cluster averages (which we call the 'meta-covariates') are then used to predict the response by assigning each meta-covariate a weight in a regression model; (C) we update the cluster structure (Step A) and the regression weights (Step B) depending on how well the meta-covariate regression model predicts the response. It is also important to appreciate that this method can be used as an exploratory tool as well as a prediction algorithm.

The significant parameters in this model are θ, π, Σ, γ and w. w is a vector, containing the weights assigned to each meta-covariate (and therefore each cluster) in the regression model. Each value in w indicates how much influence each cluster has in determining the value of the response and therefore how informative it is when discriminating between different values of the response (in the hypertension dataset, the response is salt-loaded or non-salt-loaded, while in the leukaemia dataset, the response is AML or ALL). The other four parameters are relevant to the clustering model. θ is a matrix containing the meta-covariate representations of the clusters and Σ is a matrix that describes the variance within each cluster in the model; i.e. $\boldsymbol{\theta_k}$ and $\Sigma_k = diag(\sigma^2_{k_1},...,\sigma^2_{k_N})$ are the mean and covariance of the k-th cluster. π is a vector containing the proportion of probes in the dataset that are assigned to each cluster, which are the 'mixing coefficients'. γ is a matrix containing the 'responsibilities' that each cluster k takes for *explaining* each probe; each element of γ can be interpreted as the probability that a particular probe belongs to a particular cluster (the γ values for any probe will sum to 1). To generate assignments, a probe is assigned to the cluster to which it has the highest probability of belonging. Using such 'soft' clustering (rather than 'hard' clustering, where each probe is assigned to a cluster with a probability of 1), aids the interpretation of the model.

Our EM procedure iteratively updates the values of π, θ, Σ, γ and w (and others, see Supplementary Data) until the model converges. More specifically, given some number of clusters K, the goal is to maximise the log joint distribution with respect to the parameters, π, θ, Σ, γ and w, until the model converges. Here, the convergence criterion is an increase in the log joint distribution of ≤ 0.00001 or some maximum number of iterations. Note that the value of K must be set before optimisation, necessitating a model selection step that identifies which K is best for a given dataset.

Full details of our method are given in the Supplementary Data, Sections 1.2–1.3 and MATLAB code is available at http://www.dcs.gla.ac.uk/inference/metacovariateanalysis/.

Mapping and ingenuity pathway functional analyses

All probe to gene mappings; gene to pathway mappings and network analysis tools were taken from Ingenuity Pathway Analysis software (IPA, http://www.ingenuity.com/) as of October 2009. Molecular interactions between genes were mapped to a common pathway using the Pathway Explorer function within IPA software.

RESULTS AND DISCUSSION

A well-established leukaemia dataset containing expression data for AML and ALL was used initially to illustrate our method (2). Our method was then applied to a novel dataset of renal gene expression data with a view to providing insight into salt-sensitive hypertension.

Throughout this section, clusters will be represented as C_n^D where n gives the ID of that cluster in the dataset D ($D \in \{leuk, salt\}$) where *leuk* denotes the Golub *et al*. dataset and *salt* denotes the hypertension dataset.

The leukaemia data analysis

Leukaemia is a broad term to describe cancer of the blood or bone marrow. Haemopoiesis, the process of blood production, is organized hierarchically with the haemopoietic stem cell at the apex. The first major lineage diversion is between myeloid and lymphoid progenitors. In AML there is a block to differentiation with a rapid accumulation of abnormally proliferating myeloid blasts. This process is mirrored in ALL, but in this case, the blasts are of lymphoid morphology (13, Chapter 12).

In 1999, Golub *et al*. published work in which previously unseen samples could be classified according to their gene expression profiles; using a weighted vote of 50 probes, they successfully classified all but one of the samples in the test set of 34 samples (14 AML and 20 ALL samples). This dataset has been subject to extensive analysis in the past decade and predictions made from these data are consistently of good quality, regardless of the approach taken: using a sparse Bayesian classification model, Bae and Mallick (14) misclassified two test samples; Lee and Lee (15) used support vector machines to analyse an extended multinomial version of the Golub *et al*. dataset and achieved a misclassification rate of 1; Tibshirani *et al*. (16) used the nearest shrunken centroids and misclassified two samples; and using a hierarchical Bayesian model, Lee *et al*. (17) misclassified only one sample.

Although AML and ALL are both forms of leukaemia, they cause accumulation of different types of cell (5). As such, there will be many differences between the two sets of samples in this dataset that are attributable to cell type, rather than the molecular pathology of the two diseases. These cellular differences may be responsible for the ease with which the AML and ALL samples are discriminated in the literature. It must also be noted that there are subtypes of AML and ALL (18)—in the process of haemopoiesis, myeloid and lymphoid progenitor cells give rise to further cell lineages, where subtypes of AML and ALL describe cancers exhibiting variable levels of differentiation towards mature myeloid and lymphoid cells—and that the Golub *et al*. dataset pools all AML and ALL subtypes together. In addition to the heterogeneity inherent in the disease, the samples in the dataset vary with respect to the age of the patient and with respect to sample type (e.g. both bone marrow aspirates and peripheral blood mononuclear samples are used). As such, we expect that any biology captured by our model would represent very 'general' characterizations of myeloid and lymphoid cells.

The meta-covariate analysis of the leukaemia data

The Golub data were pre-filtered as described by Dudoit *et al*. (6). In our representation, AML samples have been encoded as 1 and ALL samples have been encoded as 0; therefore, positively weighted clusters are predictive of AML samples (these clusters will be described as AML+) and negatively weighted clusters are predictive of ALL (such clusters will be described as ALL+). A model selection step identified $K = 22$ as the best model using the criterion of minimum average test error (the model selection step performed 1000 iterations of the EM algorithm, where $2 < K < 50$).

Table 1. The 22 clusters obtained by meta-covariate analysis of the leukaemia data [clusters are ordered by $abs(w)$]

Cluster	Probes	w	π	$\overline{\sigma^2}$
10	62	−5.32	1.79×10^{-2}	7.34×10^{-1}
12	96	2.55	2.70×10^{-2}	6.69×10^{-1}
21	177	2.43	4.93×10^{-2}	3.94×10^{-1}
14	214	−2.30	5.95×10^{-2}	2.10×10^{-1}
5	142	2.17	3.96×10^{-2}	2.59×10^{-1}
19	37	−1.91	1.06×10^{-2}	4.88×10^{-1}
22	124	1.73	3.49×10^{-2}	9.13×10^{-1}
3	179	−1.71	5.02×10^{-2}	1.18×10^{-1}
4	190	−1.65	5.42×10^{-2}	2.37×10^{-1}
8	263	1.25	7.33×10^{-2}	1.31×10^{-1}
15	143	1.04	3.99×10^{-2}	2.39×10^{-1}
1	75	1.00	2.12×10^{-2}	1.92×10^{-1}
7	52	−0.85	1.45×10^{-2}	2.68×10^{-1}
16	339	−0.79	9.54×10^{-2}	2.99×10^{-1}
11	111	0.56	3.09×10^{-2}	6.57×10^{-1}
20	162	0.53	4.50×10^{-2}	1.72×10^{-1}
13	202	0.50	5.62×10^{-2}	1.42×10^{-1}
2	191	−0.30	5.35×10^{-2}	2.61×10^{-1}
9	210	−0.27	5.78×10^{-2}	2.16×10^{-1}
18	265	−0.17	7.48×10^{-2}	1.45×10^{-1}
6	98	0.15	2.71×10^{-2}	2.88×10^{-1}
17	239	−0.04	6.69×10^{-2}	1.12×10^{-1}

w is the regression coefficient of the cluster, π is the size of the cluster (as a percentage of the whole dataset), $\overline{\sigma^2}$ is the mean variance of the cluster.

The *maximum a posteriori* (MAP; 12, pp. 30) solution for this model discriminates perfectly between AML and ALL samples, in both the training and test set, providing evidence that our meta-covariate model is able to make good predictions and suggesting that the clusters formed are response relevant and, therefore, potentially biologically relevant.

Cluster morphology. The meta-covariate model algorithm was run to convergence—the criterion being a difference in the joint posterior of <0.0001 or a maximum of 5000 iterations—on the leukaemia data, partitioning the probes into 22 clusters. These clusters and their associated regression coefficients (w), dataset proportion (π) and mean variance ($\overline{\sigma^2}$, the variance in expression of cluster members, averaged over samples) are described in Table 1. There is a marginal trend for $|w_k|$ to decrease with cluster size ($\rho = -0.37$, $P = 0.09$). However, there is a significant correlation between the mean variance in the cluster and its influence ($\rho = 0.54$, $P = 0.01$). This is perhaps contrary to expectation. It might be expected that the most influential clusters would identify transcriptionally tight clusters of genes corresponding to specific sub-functionality; however, the opposite is true: the more influential clusters are more variable.

This can be explained by considering how $\boldsymbol{\theta_k}$ is calculated (see Equation 4 in the Supplementary Data).

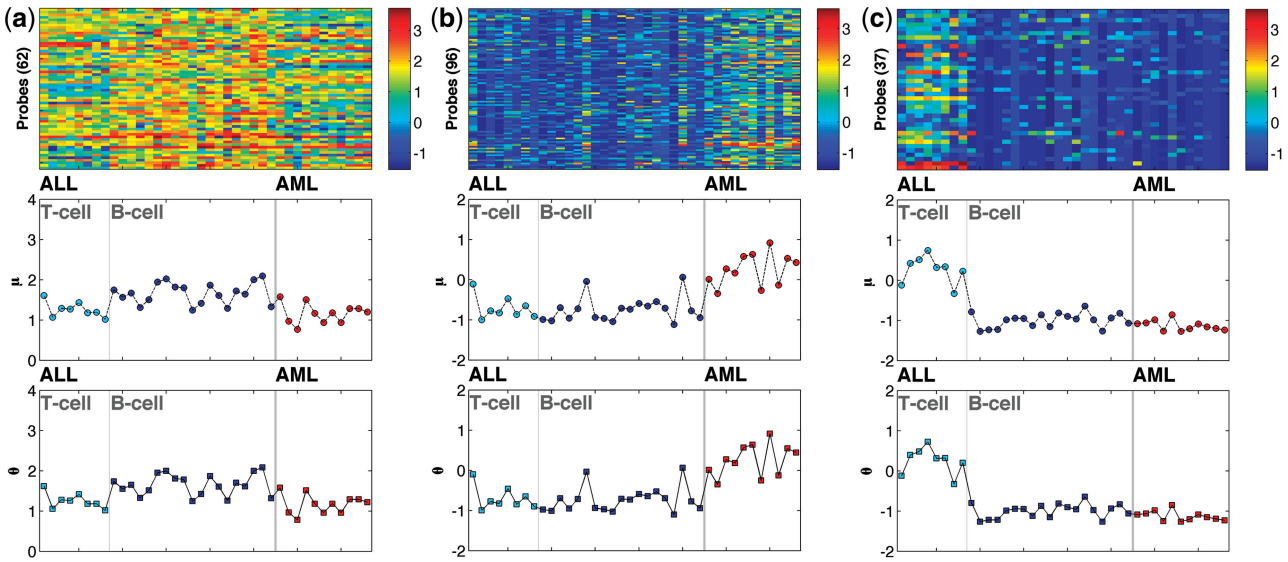

Figure 2. Expression, mean expression (μ) and $\boldsymbol{\theta_k}$ vectors for three clusters generated by the meta-covariate method, when analysing the leukaemia data. (**a**) ALL+C_{10}^{leuk}($w = -5.32$); (**b**) AML+C_{12}^{leuk}($w = 2.55$) and (**c**) ALL+C_{19}^{leuk}($w = -1.91$). Extended versions with sample IDs are available in Supplementary Figures S14–S16.

$\boldsymbol{\theta_k}$ is comprised of both a model mismatch component, which describes how well the current classification model matches the response data, and a standard clustering component (12, Section 9.2.2). As the cluster size decreases, that is, as $\boldsymbol{\gamma_k}$ becomes more sparse (where $\boldsymbol{\gamma_k}$ is the vector of clustering responsibilities for cluster $\boldsymbol{k}$), the model mismatch terms will dominate the calculation of $\boldsymbol{\theta_k}$ as the standard clustering component, dependent on $\boldsymbol{\gamma_k}$, will diminish. Conversely, as the cluster becomes larger and $\boldsymbol{\gamma_k}$ becomes less sparse, the standard mixture modelling component will dominate the calculation. Furthermore, as the cluster becomes more influential and the value of $|w_k|$ increases, the model mismatch term will dominate further. Therefore, the model will tend to form smaller, influential, more variable clusters and larger, less influential and less variable clusters, thereby automatically inducing sparsity in the model.

Capturing large-scale, biological functionality. The model is capable of capturing large-scale biological functionality that is of relevance to the response. As expected, the biology captured by the most influential clusters in this dataset describes functions characteristic of myeloid and lymphoid cells.

C_{10}^{leuk} is the most influential cluster in the model generated from the leukaemia data (Figure 2a). The expression of the genes in this cluster is associated with ALL samples. C_{10}^{leuk} is enriched for elements in the 'MIF regulation of innate immunity' pathway, due to the inclusion of MIF and its cell surface receptor CD74 (19) in the cluster (Supplementary Figure S2). MIF is a lymphokine, a signalling molecule expressed by lymphocytes (http://www.ncbi.nlm.nih.gov/gene/4282), which has been shown to play a role in T-cell tumourigenesis (20) and lymphocyte proliferation (21,22). CD74 is expressed on malignant B cells (a form of lymphoid cell), but is expressed to a much lesser extent on non-malignant cell surfaces stein (23) (Supplementary Figure S3).

C_{12}^{leuk} is the most influential AML+ cluster (Figure 2b). The most over-represented IPA pathway in this cluster is the 'triggering receptor expression on myeloid cells 1' (or TREM1) signalling pathway (Supplementary Figure S4). TREM1 activation has various roles in both the adaptive and innate immune response, but critically, it is only expressed in myeloid cells (Supplementary Figure S5). This cluster is also enriched for 'acute myeloid leukaemia signalling' proteins; in fact, the top five AML+ clusters (C_{12}^{leuk}, C_{21}^{leuk}, C_{5}^{leuk}, C_{22}^{leuk} and C_{8}^{leuk}) are all enriched for this pathway (Supplementary Figure S6). This IPA canonical pathway describes the signalling pathways which, when disrupted by abnormalities (e.g. mutations to genes and/or transcription factors), can lead to increased proliferation and apoptosis resistance in AML. These two pathways are myeloid-specific, describing processes that occur exclusively in myeloid cells.

Clusters can also represent more specific, biological sub-functionality. The predictive ability of each cluster only exists within the meta-covariate model. Although some clusters may clearly discriminate between AML and ALL samples, others may be good at predicting subtypes of either disease. ALL samples can be further sub-classified as T- or B-cell ALL. C_{19}^{leuk} is an example of one of these 'subtype specific' clusters. It is ALL+, with a regression coefficient of -1.91. From the expression plot in Figure 2c, it is clear that this cluster is important when classifying specifically T-cell ALL samples: expression in these samples is visibly higher, while expression in the B-cell ALL and AML samples is similarly low. This cluster is enriched for several T-cell lymphocyte-specific canonical pathways, including the 'Calcium-induced T lymphocyte apoptosis'; 'iCOS-iCOSL' and 'CD28 Signalling in T Helper Cells'; 'cytotoxic T lymphocyte-mediated apoptosis of target cells' and 'T cell receptor signalling' IPA canonical pathways (Supplementary Figure S7).

Sub-type-specific clusters can arise in our model because complementary clusters, which are able to predict the other subtype(s) within a class, can exist. An example of a complementary cluster to C_{19}^{leuk} is C_{14}^{leuk} (Supplementary Figure S8); here, the cluster contains genes that are more highly expressed in B-cell ALL samples than T-cell ALL and AML samples.

Applying our meta-covariate method to novel renal, gene expression data

In the previous section, we illustrated the use of our meta-covariate method by applying it to a well-known leukaemia dataset. We observed that the influential clusters tend to be smaller and more variable than the less influential clusters and that the model is able to capture both large-scale biological characteristics and small-scale, more specific biological characteristics. In the next section, our method is applied to a dataset of renal gene expression profiles in a rat model of salt-sensitive hypertension.

The hypertension data analysis

Essential hypertension (chronically elevated blood pressure) is a genetically complex disease, currently affecting one quarter of adults worldwide and projected to affect almost 30% of adults within 15 years (24). One half of hypertensive patients are salt sensitive, exhibiting increased blood pressure with increased dietary sodium (25). Elucidating the genetics of hypertension would have far-reaching implications for global health. Animal models are useful functional models allowing the genetic dissection of complex, polygenic disease; the data described here are derived from a rat model of salt-sensitive hypertension (26).

The SHRSP, Wistar-Kyoto (WKY) and congenic SP.WKYGla2a strains of rat are distinct with respect to phenotype in response to salt, with the SHRSP demonstrating increased systolic blood pressure and circadian amplitude in response to salt, the WKY being largely unaffected by salt, and the SP.WKYGla2a demonstrating an intermediate increase in both systolic blood pressure and circadian amplitude in response to salt (27, Supplementary Figure S9).

Microarray experiments were conducted to measure renal gene expression in male, age-matched, 21-week-old salt-loaded and non-salt-loaded animals. The resulting dataset was analysed using our meta-covariate method. Genes contained in influential clusters will be informative when discriminating between salt-loaded samples and non-salt-loaded samples. Furthermore, identifying gene expression changes between SP.WKYGla2a and SHRSP will highlight chromosome 2-dependent processes involved in blood pressure regulation.

The sample size ($n = 18$) is small; as such, we used all the data to build the model, rather than making predictions on a test set. However, here we can demonstrate the second use of our method, by employing it as a valuable supervised clustering tool to generate response-relevant clusters within the given dataset, rather than using it primarily to build a classifier (as demonstrated in the previous section when analysing the leukaemia data).

Table 2. The 20 clusters obtained by meta-covariate analysis of the salt data [clusters are ordered by *abs*(*w*)]

Cluster	Probes	w	π	$\overline{\sigma^2}$
13	13	−46.76	2.71×10^{-3}	2.27×10^{-1}
8	7	−9.59	1.51×10^{-3}	2.26×10^{-1}
5	14	−7.13	3.14×10^{-3}	2.33×10^{-1}
3	70	4.27	$1.54 \times 10-^{2}$	5.59×10^{-2}
15	301	−3.32	6.61×10^{-2}	2.37×10^{-2}
6	317	−3.08	6.96×10^{-2}	2.31×10^{-2}
14	284	−2.69	6.23×10^{-2}	2.51×10^{-2}
2	408	2.65	8.95×10^{-2}	1.52×10^{-2}
16	329	2.02	7.19×10^{-2}	1.95×10^{-2}
4	454	1.85	9.95×10^{-2}	1.58×10^{-2}
19	28	−1.54	6.18×10^{-3}	1.13×10^{-1}
17	336	−1.44	7.29×10^{-2}	2.13×10^{-2}
10	361	1.35	7.93×10^{-2}	2.05×10^{-2}
1	163	1.00	3.62×10^{-2}	3.07×10^{-2}
7	316	−0.88	6.91×10^{-2}	2.29×10^{-2}
11	231	0.82	5.07×10^{-2}	2.46×10^{-2}
20	90	0.80	1.96×10^{-2}	4.53×10^{-2}
9	310	−0.58	6.85×10^{-2}	2.24×10^{-2}
12	282	−0.28	6.15×10^{-2}	2.38×10^{-2}
18	248	−0.10	5.45×10^{-2}	2.09×10^{-2}

w is the regression coefficient of the cluster, π is the size of the cluster (as a percentage of the whole dataset), $\overline{\sigma^2}$ is the mean variance of the cluster.

The 4562 probes on the array were identified as significant using the Wilcoxon rank-sum test ($P \leq 0.05$). K was set to 20, following 1000 iterations of the method for each value of K where $2 < K < 50$, and subsequent analysis using Bayesian information criterion (BIC); for $K = 20$, BIC = 85 731. BIC is a regularized model selection technique, which identifies the most likely values of the model parameters, whilst penalising unnecessary model complexity (12, Section 4.4.1). Upon completion, the meta-covariate model successfully partitioned the dataset with respect to salt.

Cluster morphology. The 20 clusters that are formed are described in Table 2. Here, there is an imbalance of positively and negatively weighted clusters—12 negative to 8 positive—unlike in the leukaemia model (Table 1), where there were equal numbers of positively and negatively weighted clusters.

This model is dominated by heavily, negatively weighted clusters: the three most influential clusters (C_{13}^{salt}, $w = -46.76$; C_{8}^{salt}, $w = -9.59$; C_{5}^{salt}, $w = -7.13$) are all negatively weighted (note also that these three clusters have similar variance). This suggests the dominant biology captured by this model is reduced expression in the salt-loaded animals. That is, the biology that contributes most significantly to the discrimination between the salt-loaded and non-salt-loaded samples is that of lower expression in the salt-loaded samples.

Cluster size is significantly inversely correlated with regression weight ($\rho = -0.46$, $P = 0.04$) and significantly correlated with average variance ($\rho = 0.67$, $P < 0.01$). Therefore, as observed in the leukaemia dataset, the method has generated both small, variable (with respect

to member gene expression), influential clusters and large, tight, non-influential clusters.

This feature of our model is particularly useful in this dataset, where all of the 4562 probes are significantly correlated with the response. The induced sparsity allows identification of the most relevant features, in a congested dataset where all features are relevant by traditional, univariate methods. Furthermore, there is no correlation between the Wilcoxon P-value and regression coefficient in this dataset ($\rho = -1.47 \times 10^{-3}$, $P = 9.21 \times 10^{-1}$), indicating that the most valuable predictors (as defined by the meta-covariate model) would not be selected on the basis of P-value alone.

With a view to establishing (i) how sensitive our method is to variation in the data and (ii) how robust these clusters are, we performed leave one out cross-validation (LOOCV) and compared the models generated from the LOOCV folds to each other and to the model generated from the full dataset using two metrics—adjusted rand index (ARI) (28) and adjusted mutual information (AMI) (29)—both of which measure concordance between clustering structures, while accounting for chance. The results are very encouraging: despite the small sample size, clustering concordance is high (see Supplementary Figures S10–S11 and Supplementary Tables S1–S2). The mean concordance between the clustering structures of the LOOCV-fold models and the clustering structure of the model built from the dataset in its entirety (i.e. the clustering described in Table 2) is 0.96 ($\sigma = 0.011$) and 0.96 ($\sigma = 0.0070$) for the ARI and AMI metrics, respectively (all values rounded to 2 s.f.). This convincingly demonstrates that a similar clustering structure is observed across LOOCV folds and, therefore, that the method is insensitive to variation in the input data. This is particularly encouraging given that an initial motivation for this method was to avoid such sensitivity. We can now progress with the analysis of these data, with confidence in the clustering structure.

An influential cluster of thirteen genes. C_{13}^{salt} is the most influential cluster: its regression coefficient is five times larger than the second most influential cluster. Classification using this cluster and its regression coefficient in isolation results in only one misclassification (the SHRSP+salt animal, C3996) using the decision boundary ($y = 0$). Although we should be cautious of reading too much into cluster performance in isolation, given that clusters are only relevant as part of the model as a whole, it is a useful indicator of how informative a cluster is in the model.

The negative regression coefficient indicates that the genes in this cluster are, largely, more highly expressed in the non-salt-loaded samples than the salt-loaded samples, as is evident in the graph of the mean expression values (μ) in Figure 3. Comparing the mean expression values to the θ values illustrates the effect of incorporating an outcome-specific component in the calculation of $\boldsymbol{\theta_k}$: the difference between the non-salt-loaded and salt-loaded samples is exaggerated in the graph of θ_{13} ($\boldsymbol{\theta_k}$ where $k = 13$) in Figure 3.

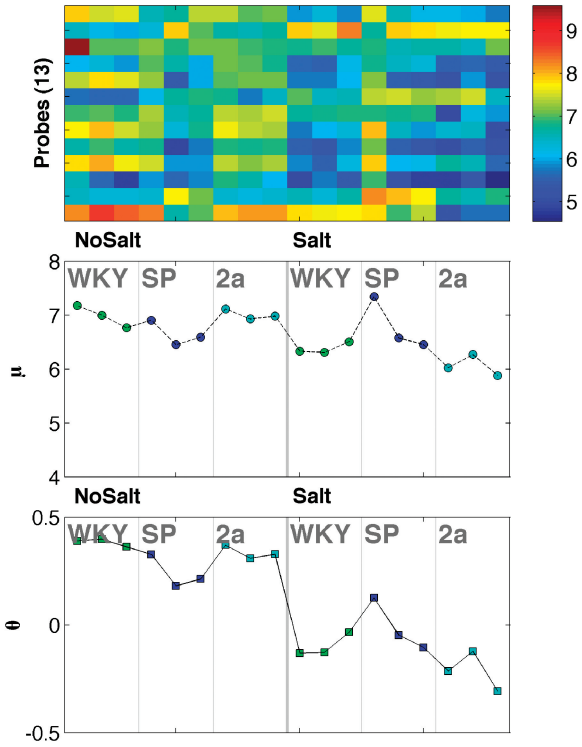

Figure 3. Expression, mean expression (μ) and $\boldsymbol{\theta_k}$ vectors for $C_{13}^{salt}(w = -46.76)$, generated by the meta-covariate method when analysing the hypertension data. Extended version with sample IDs is given in Supplementary Figure S17.

Note that the difference between μ and θ in the renal dataset is greater than in the leukaemia dataset (Figure 2a–c and Supplementary Figure S8). This suggests that there is greater discriminative power in the unaltered leukaemia data than in the unaltered hypertension data. This is not surprising, given the known heterogeneity in the leukaemia samples and the comparative homogeneity of the inbred rats. It is encouraging that the model is able to use patterns that exist in the mean gene expression data to build the model, but that it is also able to alter the cluster representation (i.e. alter θ) to find more complex informative patterns.

Strain-specific expression of C_{13}^{salt} genes. Figure 4a–c show the results of a RP analysis (10) within each strain, between the salt-loaded and non-salt-loaded datasets (Supplementary Tables S3–S5, chromosome mappings given in Supplementary Table S6). Most of the genes are significantly differentially expressed between the salt-loaded and non-salt-loaded datasets in both the WKY and SP.WKYGla2a (Figure 4a and b, respectively). However, the same genes are not differentially expressed when comparing the salt-loaded and non-salt-loaded SHRSP animals (Figure 4c) giving rise to the hypothesis that changes in expression levels of the genes in C_{13}^{salt} are protective against hypertension in response to an increase in dietary sodium. These results are corroborated by a Rosetta Resolver analysis (http://www.rosettabio.com/products/resolver; data not shown) and the differential expression of the four transcription factors (*Arntl*,

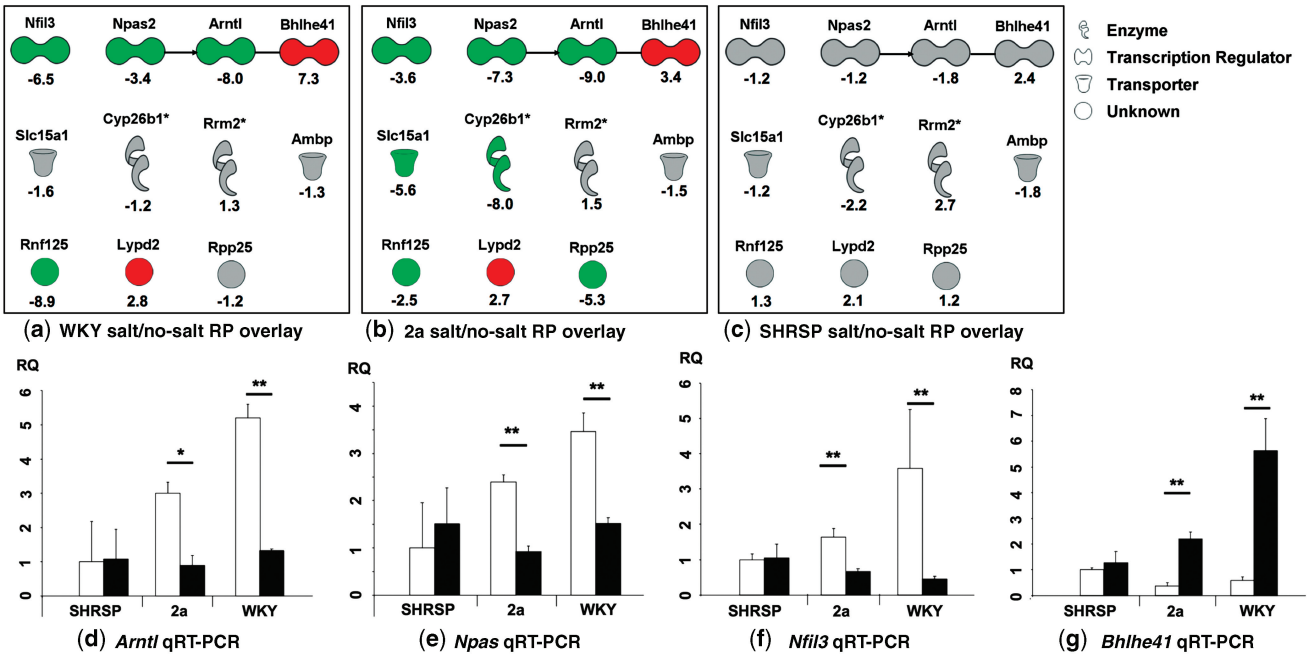

Figure 4. The contents of C_{13}^{salt} overlaid with the salt-loaded comparisons in the (**a**) WKY (salt/no-salt RP overlay), (**b**) SP.WKY Gla2a (salt/no-salt RP overlay) and (**c**) SHRSP (salt/no-salt RP overlay) animals. Green indicates significant down-regulation and red indicates significant up-regulation in the salt dataset. RP-fold change is indicated below each molecule. Direct relationships are indicated by a solid line. qRT–PCR of the four transcription factors identified in C_{13}^{salt} confirming significant differences in SP.WKYGla2a and WKY salt-loaded animals (filled) compared to age-matched animals not exposed to salt (open) for (**d**) *Arntl*, (e) *Npas*, (**f**) *Nfil3* and (**g**) *Bhlhe41* ($*P < 0.05$, $**P < 0.001$).

Npas2, *Nfil3* and *Bhlhe41*) have been confirmed by qRT–PCR (Figure 4d–g, respectively).

Circadian rhythm genes are implicated. Eleven of the thirteen probes in this cluster were mapped to genes using IPA. A canonical pathway analysis of these eleven genes shows that the cluster is enriched for circadian rhythm signalling genes (Supplementary Figures S12–S13). This is relevant as all three rat strains demonstrate circadian patterns of systolic blood pressure: these nocturnal animals have a higher blood pressure during the night than during the day and this difference and the circadian amplitude is exacerbated on salt-loading in the SHRSP (Supplementary Figure S9).

Identifying a transcriptional network within C_{13}^{salt}. Also shown in Figure 4a–c are the relationships between the genes in C_{13}^{salt}, as described in the Ingenuity Pathway Knowledge Base. Of note are the four transcription factors, three of which, neuronal PAS domain protein 2 or *Npas2* (also known as *Bhlhe9*); aryl hydrocarbon receptor nuclear translocator-like or *Arntl* (also known as *Bmal1* and *Bhlhe5*); and basic helix–loop–helix family member e41 *Bhlhe41* (also known as *Dec2*), are known to form a transcriptional network and, as seen in a previous section, are potentially protective against hypertension, being differentially expressed on salt in the SP.WKYGla2a and WKY strains. These three transcription factors are central components of the circadian clock (Supplementary Figure S12). Aryl hydrocarbon receptor nuclear translocator-like (*Arntl*) forms a heterodimer with *Clock* and is required for E-box-dependent transactivation activating the transcription of the *Per* genes from the E-box elements in its promoter region (30,31). Protein products of *Per* act together with *Cry* proteins to inhibit *Per* transcription, thus closing the autoregulatory feedback loop. It has been shown that the basic helix–loop–helix transcription factors (*Bhlhe41*) can repress *Clock*/*Bmal1*-induced transactivation of the mouse *Per1* promoter through direct protein-protein interactions with *Bmal1* and/or competition for E-box elements. Disruption of the key molecular oscillators (*Arntl*, *Npas2*) and autoregulatory feedback loops (*Bhlhe41*, *Per*, *Dbp*, *Cry*), have recently been shown to be involved in hypertension (32) and salt sensitivity in both mice (33,34) and rats (35).

Identifying a significant transcriptional network. The IPA network generation algorithm was used to form networks of genes known to be functionally related, as defined by the Ingenuity Pathway Knowledge Base. This algorithm generates small (at most 35 genes), densely connected networks from a set of 'focus genes'; IPA is able to 'fill in the gaps' with linker genes to maximise connectivity in the networks. Constructing networks around the significantly differentially expressed genes identified by RP (10) in the salt data, we can identify networks of functionally related genes that are relevant to salt-loaded animals.

The same three transcriptional regulators that form the transcriptional network in C_{13}^{salt} are present in the most significant networks generated from the SP.WKYGla2a ($P = 1 \times 10^{-48}$) and WKY ($P = 1 \times 10^{-46}$) RP gene

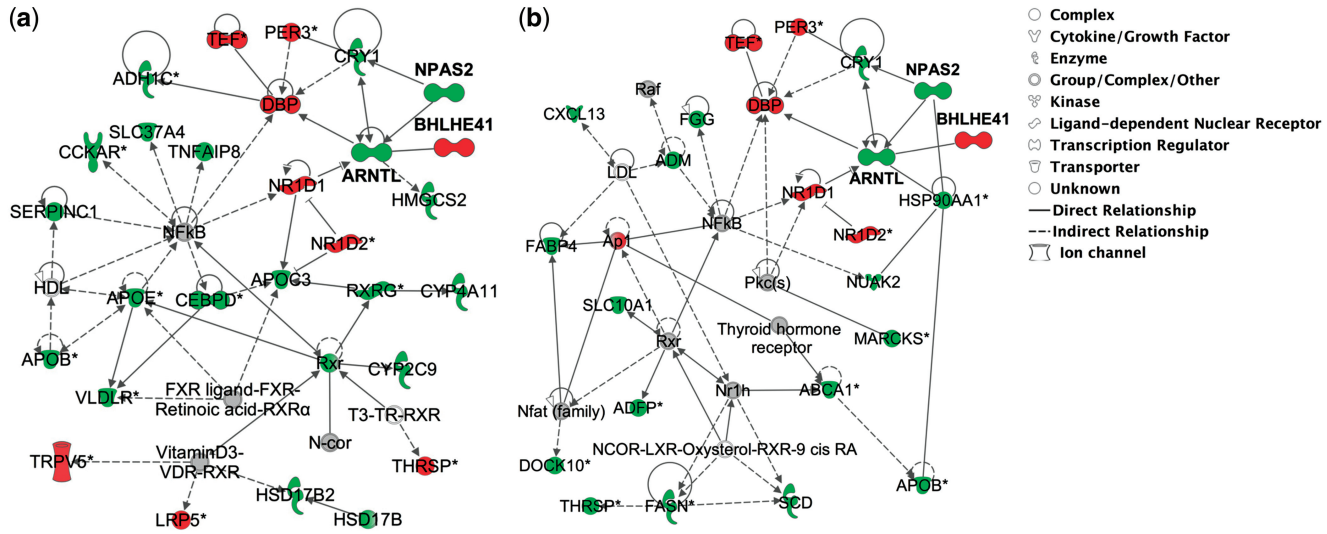

Figure 5. (**a**) and (**b**) the most significant networks identified in the salt-loaded/non-salt-loaded data, across the SP.WKYGla2a and WKY strains, respectively. Red indicates up-regulation in the salt-dataset compared to no-salt treatment, green indicates down-regulation in the dataset, as defined by RP (10); legend shown above.

expression microarray data: *Arntl*, *Npas2* and *Bhlhe41* (Figure 5a–b and Supplementary Tables S7–S8). To have arrived at a similar conclusion both by way of IPA network analysis and by our meta-covariate method is encouraging. Further, our meta-covariate method identifies a much smaller set of genes, allowing more concise interpretation of the data.

Further investigation and validation experiments are underway, with the priority being the elucidation of how these genes are linked to chromosome 2 and how they are involved in sodium homeostasis. In addition, a major focus will be to investigate why the genes in C_{13}^{salt} vary similarly with the response; this may be due to shared transcriptional regulation.

CONCLUSION

In this article, we describe and illustrate a novel method of microarray analysis using the Golub *et al.* (5) leukaemia dataset, before applying the same analysis to a novel dataset of renal gene expression data in rat models of salt-sensitive hypertension. It was demonstrated that the prediction performance of our meta-covariate method is competitive in the Golub *et al.* dataset. Although we refrain from drawing any additional conclusions from these data, beyond the identification of general patterns, we would like to stress that further analysis of these data could be informative, provided the caveats with respect to the experimental design are kept in mind.

Although we were not able to evaluate prediction performance in an independent test set given the small size of the hypertension dataset, the model generated from the training set was able to perfectly discriminate between salt-loaded and non-salt-loaded samples. However, it must also be noted that it is perfectly valid to use the meta-covariate method as a supervised clustering technique with a view to identifying response-relevant gene clusters, as well as a classification model.

Both datasets demonstrated that the model tends to form small, variable, influential clusters and larger, tighter, less influential clusters. This is particularly useful in a congested, homogeneous dataset, such as the hypertension dataset, where many, if not all, variables are significantly correlated with the response. The flexibility of the model was evident in that discrimination patterns were identified in the mean gene expression data where possible, but where these data were not informative, complex patterns were identified by alternative representations of the clusters.

We are currently developing a fully Bayesian implementation of this meta-covariate method—which will provide a range of clustering structures for a dataset rather than a single clustering scheme—while carrying out further biological validation of our findings.

SUPPLEMENTARY DATA

Supplementary Data available at NAR Online.

FUNDING

Engineering and Physical Sciences Research Council (EP/F009429/1, EP/E052029/1) awarded to M.A.G.; British Heart Foundation Chair and Programme grant funding (CH98001, RG/07/005); Wellcome Trust Cardiovascular Functional Genomics Initiative (066780/Z/01/Z); European Union Sixth Framework Programme Integrated Project (LSHG_CT 2005-019015 EURATools) awarded to A.F.D. Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. de Snoo,F., Bender,R., Glas,A. and Rutgers,E. (2009) Gene expression profiling: decoding breast cancer. *Surgical Oncology*, **18**, 366–378.
2. Harris,K., McMillan,L. and Girolami,M. (2009) Inferring meta-covariates in classification. *Pattern Recogn. Bioinform. LNBI.*, **5780**, 150–161.
3. Hanczar,B., Courtin,M., Benis,A., Henegar,C., Clement,K. and Zucker,J.D. (2003) Improving classification of microarray data using prototype-based feature selection. *SIGKCC Explorations*, **5**, 23–30.
4. Park,M.Y., Hastie,T. and Tibshirani,R. (2007) Averaged gene expressions for regression. *Biostatistics*, **8**, 212–227.
5. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
6. Dudoit,S., Fridlyand,J. and Speed,T.P. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
7. Clark,J.S., Jeffs,B., Davidson,A.O., Lee,W.K., Anderson,N.H., Bihoreau,M.T., Brosnan,M.J., Devlin,A.M., Kelman,A.W., Lindpaintner,K. *et al.* (1996) Quantitative trait loci in genetically hypertensive rats. Possible sex specificity. *Hypertension*, **28**, 898–906.
8. Jeffs,B., Negrin,C.D., Graham,D., Clark,J.S., Anderson,N.H., Gauguier,D. and Dominiczak,A.F. (2000) Applicability of a "speed" congenic strategy to dissect blood pressure quantitative trait loci on rat chromosome 2. *Hypertension*, **35**, 179–187.
9. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
10. Breitling,R., Armengaud,P., Amtmann,A. and Herzyk,P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.
11. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate–a practical and powerful approach to multiple testing. *JRSS-B*, **57**, 289–300.
12. Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York, 2006.
13. Hoffbrand,A.V., Moss,P.A.H. and Pettit,J.E. (2006) *Essential Haematology*, 5th edn. Blackwell Publishing, Malden, MA, USA.
14. Bae,K. and Mallick,B.K. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423–3430.
15. Lee,Y. and Lee,C.-K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.
16. Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
17. Lee,K.E., Sha,N., Dougherty,E.R., Vannucci,M. and Mallick,B.K. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.
18. Bennett,J.M., Catovsky,D., Daniel,M.T., Flandrin,G., Galton,D.A., Gralnick,H.R. and Sultan,C. (1976) Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br. J. Haematol*, **33**, 451–458.
19. Bach,J.-P., Rinn,B., Meyer,B., Dodel,R. and Bacher,M. (2008) Role of MIF in inflammation and tumorigenesis. *Oncology*, **75**, 127–133.
20. Abe,R., Peng,T., Sailors,J., Bucala,R. and Metz,C.N. (2001) Regulation of the CTL response by macrophage migration inhibitory factor. *J. Immunol.*, **166**, 747–753.
21. Ziino,O., D'Urbano,L.E., De Benedetti,F., Conter,V., Barisone,E., De Rossi,G., Basso,G. and Arica,M. (2005) The MIF-173G/C polymorphism does not contribute to prednisone poor response in vivo in childhood acute lymphoblastic leukemia. *Leukemia*, **19**, 2346–2347.
22. Gore,Y., Starlets,D., Maharshak,N., Becker-Herman,S., Kaneyuki,U., Leng,L., Bucala,R. and Shachar,I. (2008) Macrophage migration inhibitory factor induces B cell survival by activation of a CD74-CD44 receptor complex. *J. Biol. Chem.*, **283**, 2784–2792.
23. Stein,R., Mattes,M.J., Cardillo,T.M., Hansen,H.J., Chang,C.-H., Burton,J., Govindan,S. and Goldenberg,D.M. (2007) CD74: a new candidate target for the immunotherapy of B-cell neoplasms. *Clin. Cancer Res.*, **13**, 5556s–5563s.
24. Kearney,P.M., Whelton,M., Reynolds,K., Muntner,P., Whelton,P.K. and He,J. (2005) Global burden of hypertension: analysis of worldwide data. *Lancet*, **365**, 217–223.
25. Weinberger,M.H. (1996) Salt sensitivity of blood pressure in humans. *Hypertension*, **27**, 481–490.
26. McBride,M.W., Graham,D., Delles,C. and Dominiczak,A.F. (2006) Functional genomics in hypertension. *Curr. Opin. Nephrol. Hypertens.*, **15**, 145–151.
27. Graham,D., McBride,M.W., Gaasenbeek,M., Gilday,K., Beattie,E., Miller,W.H., McClure,J.D., Polke,J.M., Montezano,A., Touyz,R.M. *et al.* (2007) Candidate genes that determine response to salt in the stroke-prone spontaneously hypertensive rat: congenic analysis. *Hypertension*, **50**, 1134–1141.
28. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
29. Vinh,N.X., Epps,J. and Bailey,J. (2009) Information theoretic measures for clusterings comparison: Is a correction for chance necessary? *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.
30. Bunger,M.K., Wilsbacher,L.D., Moran,S.M., Clendenin,C., Radcliffe,L.A., Hogenesch,J.B., Simon,M.C., Takahashi,J.S. and Bradfield,C.A. (2000) Mop3 is an essential component of the master circadian pacemaker in mammals. *Cell*, **103**, 1009–1017.
31. Huffman,J.L., Mokashi,A., Bächinger,H.P. and Brennan,R.G. (2001) The basic helix-loop-helix domain of the aryl hydrocarbon receptor nuclear transporter (ARNT) can oligomerize and bind E-box DNA specifically. *J. Biol. Chem.*, **276**, 40537–40544.
32. Woon,P.Y., Kaisaki,P.J., Bragança,J., Bihoreau,M.-T., Levy,J.C., Farrall,M. and Gauguier,D. (2007) Aryl hydrocarbon receptor nuclear translocator-like (BMAL1) is associated with susceptibility to hypertension and type 2 diabetes. *Proc. Natl Acad. Sci. USA*, **104**, 14412–14417.
33. Doi,M., Takahashi,Y., Komatsu,R., Yamazaki,F., Yamada,H., Haraguchi,S., Emoto,N., Okuno,Y., Tsujimoto,G., Kanematsu,A. *et al.* (2010) Salt-sensitive hypertension in circadian clock-deficient cry-null mice involves dysregulated adrenal hsd3b6. *Nat. Med.*, **16**, 67–74.
34. Zuber,A.M., Centeno,G., Pradervand,S., Nikolaeva,S., Maquelin,L., Cardinaux,L., Bonny,O. and Firsov,D. (2009) Molecular clock is involved in predictive circadian adjustment of renal function. *Proc. Natl Acad. Sci. USA*, **106**, 16523–16528.
35. Mohri,T., Emoto,N., Nonaka,H., Fukuya,H., Yagita,K., Okamura,H. and Yokoyama,M. (2003) Alterations of circadian expressions of clock genes in Dahl salt-sensitive rats fed a high-salt diet. *Hypertension*, **42**, 189–194.