IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Modeling and Visualizing Uncertainty in Gene Expression Clusters using Dirichlet Process Mixtures

Carl Edward Rasmussen, Bernard J de la Cruz, Zoubin Ghahramani and David L Wild,

Carl Edward Rasmussen is with the University of Cambridge, and the Max Planck Institute for Biological Cybernetics in Tübingen

Bernard J de la Cruz is with the Beckman Research Institute of the City of Hope Hospital, Duarte, CA

Zoubin Ghahramani is with the University of Cambridge

David L Wild is with the University of Warwick

Digital Object Indentifier 10.1109/TCBB.2007.70269

Abstract

Although the use of clustering methods has rapidly become one of the standard computational approaches in the literature of microarray gene expression data, little attention has been paid to uncertainty in the results obtained. Dirichlet process mixture models provide a non-parametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Most previously published applications of Bayesian model based clustering methods have been to short time series data. In this paper we present a case study of the application of non-parametric Bayesian clustering methods to the clustering of high-dimensional non-time series gene expression data using full Gaussian covariances. We use the probability that two genes belong to the same cluster in a Dirichlet process mixture model as a measure of the similarity of these gene expression profiles. Conversely, this probability can be used to define a dissimilarity measure, which, for the purposes of visualization, can be input to one of the standard linkage algorithms used for hierarchical clustering. Biologically plausible results are obtained from the Rosetta compendium of expression profiles which extend previously published cluster analyses of this data.

Index Terms

Clustering, classification, and association rules, Biology and genetics, Bioinformatics (genome or protein) databases, Statistical computing, Stochastic processes, Monte Carlo

I. INTRODUCTION

The use of clustering methods has rapidly become one of the standard computational approaches to understanding microarray gene expression data [1]–[3]. In clustering, the patterns of expression of different genes across time, treatments, and tissues are grouped into distinct clusters (perhaps organized hierarchically) in which genes in the same cluster are assumed to be potentially functionally related or to be influenced by a common upstream factor. Such cluster structure can be used to aid the elucidation of regulatory networks. For example, a compendium of gene expression profiles corresponding to mutants and chemical treatments can be used as a systematic tool to identify gene functions because mutants or drug targets that display similar profiles are likely to share cellular functions [4]. It would also be expected that gene knockouts/mutations or treatments that have impact on the same signaling or metabolic pathway or affect the same organelle would exhibit some overlap in altered gene expression profiles.

Agglomerative hierarchical clustering [1] is one of the most frequently used methods for

clustering gene expression profiles. However, commonly used methods for agglomerative hierarchical clustering rely on the setting of some score threshold to distinguish members of a particular cluster from non-members, making the determination of the number of clusters arbitrary and subjective. The algorithm provides no guide to choosing the "correct" number of clusters or the level at which to prune the tree. It is often difficult to know which distance metric to choose, especially for structured data such as gene expression profiles. Moreover, these approaches do not provide a measure of uncertainty about the clustering, making it difficult to compute the predictive quality of the clustering and to make comparisons between clusterings based on different model assumptions (e.g. numbers of clusters, shapes of clusters, etc.). In this paper we use statistical inference to overcome these limitations. An important issue that must be addressed in any clustering method is the question of how many clusters to use. Bayesian statistics and model based approaches can provide elegant solutions to model selection questions of this kind. With these approaches there is no need to make arbitrary choices about how many clusters there are in the data; nevertheless, after modeling one can still ask questions such as "how probable is it that two genes belong to the same cluster?"

Within a Bayesian framework, all assumptions are presented in terms of *priors* and the choice of likelihood function. Since it seems unreasonable to assume that complex gene expression data have been generated by some small finite number of causes, an elegant nonparametric approach is to assume that the data was in fact generated from an *infinite number of Gaussian clusters*. In a Gaussian clustering model each gene expression profile represents a multidimensional vector of measurements and the probability distribution for each cluster is assumed to be a multivariate Gaussian. We describe an approach to the problem of automatically clustering microarray gene expression profiles based on the theory of infinite Gaussian mixtures (or Dirichlet process mixtures (DPM)) [5], [6]. This theory is based on the observation that the mathematical limit of an infinite number of components in an ordinary finite mixture model (i.e. clustering model) corresponds to a Dirichlet process prior [5]–[7]. In an infinite Gaussian mixture model there is no need to make arbitrary choices about how many clusters there are in the data. Although in theory the infinite mixture model has an infinite number of parameters, surprisingly, it is possible to do exact inference in these infinite mixture models efficiently using Markov chain Monte Carlo (MCMC) methodology, since only the parameters of a finite number of the mixture components need to be represented explicitly. The theory of Dirichlet process mixture models

has a long history, going back to [7]–[9], and has recently become popular with the availability of fast MCMC inference, see [6], [10] for early examples. We first proposed and implemented the application of DPMs to clustering gene expression profiles in an extended conference abstract in 2002 [11] . Although this work is not widely known and cited, many groups have subsequently independently rediscovered the value of a fully Bayesian analysis based on DPMs to this problem [12]–[16]. We have also subsequently applied the approach to the clustering of protein sequences [17].

In this paper we illustrate our methods in detail, with a practical application to a well studied data set: the Rosetta compendium of expression profiles corresponding to 300 diverse mutations and chemical treatments in *S. cerevisiaie* [4]. We describe a simple, but novel method of visualizing the results which facilitates comparison with the dendrograms obtained by the usual hierarchical clustering approach to this type of data. Whilst our results confirm many of the previously published clusters identified in this data set, they also provide new biological insights by revealing a finer level of granularity in the clustering. These results are consistent with recent literature which suggests that distinct functions may share proteins and have overlapping regulatory mechanisms.

II. METHODS

A. Dirichlet Process Mixture Models

Although hierarchical clustering is the most widely used method for clustering gene expression data, model-based non-hierarchical methods have also been explored. One commonly used computational method of non-hierarchical clustering based on measuring Euclidean distance between gene expression profiles is given by the k-means algorithm [18], [19]. However, the k-means algorithm is inadequate for describing clusters of unequal size or shape [20]. A generalization of k-means can be derived from the theory of maximum likelihood estimation of Gaussian mixture models [21], [22]. In a Gaussian mixture model, the data (e.g. gene expression profiles, which can be arranged into p-dimensional vectors $\mathbf{y}$) is assumed to have been generated from a finite number (k) of Gaussians,

$$P(\mathbf{y}) = \sum_{j=1}^{k} \phi_j P_j(\mathbf{y}) \tag{1}$$

where ϕ_j is the mixing proportion for cluster j (fraction of population belonging to cluster j; $\sum_j \phi_j = 1$; $\phi_j \geq 0$) and $P_j(\mathbf{y})$ is a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j. The clusters can be found by fitting the maximum likelihood Gaussian mixture model as a function of the set of parameters $\theta = \{\phi_j, \mu_j, \Sigma_j\}_{j=1}^k$ using the EM algorithm [21]. Euclidean distance corresponds to assuming that the Σ_j are all equal multiples of the identity matrix.

Starting from a finite mixture model (1), we define a prior over the mixing proportion parameters ϕ. The natural conjugate prior for mixing proportions is the symmetric Dirichlet distribution, with concentration parameter α/k:

$$P(\phi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \phi_j^{\alpha/k-1} \tag{2}$$

where α controls the distribution of the prior weight assigned to each cluster, and Γ is the gamma function.

We then explicitly include indicator variables c_i for each data point (i.e. gene expression profile) which can take on integer values $c_i = j$, $j \in \{1, \ldots, k\}$, corresponding to the hypothesis that data point i belongs to cluster j. Under the mixture model, by definition, the prior probability is proportional to the mixing proportion: $P(c_i = j|\phi) = \phi_j$. A key observation is that we can compute the conditional probability of one indicator variable given the setting of all the other indicator variables after *integrating over* all possible settings of the mixing proportion parameters:

$$P(c_i = j|\mathbf{c}_{-i}, \alpha) = \int P(c_i = j|\mathbf{c}_{-i}, \phi) P(\phi|\mathbf{c}_{-i}, \alpha)\, d\phi = \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha} \tag{3}$$

where $\mathbf{c}_{-i}$ is the setting of all indicator variables except the i^{th}, n is the total number of data points, and $n_{-i,j}$ is the number of data points belonging to cluster j not including i. By Bayes rule,

$$P(\phi|\mathbf{c}_{-i}, \alpha) = P(\phi|\alpha)/P(\mathbf{c}_{-i}|\alpha) \prod_{\ell \neq i} P(c_\ell|\phi) \tag{4}$$

which is also a Dirichlet distribution, making it possible to perform the above integral analytically. We can now take the limit of k going to infinity, obtaining a Dirichlet Process with differing conditional probabilities for clusters with and without data: for clusters where $n_{-i,j} > 0$: $p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}$. For all other clusters combined: $p(c_i \neq c_{i'} \text{ for all } i' \neq i|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}$. This shows that the probabilities are proportional to the occupation numbers, $n_{-i,j}$. Using these

conditional probabilities one can Gibbs sample from the indicator variables efficiently, even though the model has infinitely many Gaussian clusters. Having integrated out the mixing proportions one can also Gibbs sample from all of the remaining parameters of the model, i.e. $\{\mu, \Sigma\}_j$, or one can integrate these out as well. The details of these procedures can be found in [6].

B. Data preprocessing

All gene expression profile data was obtained from the web site `http://www.rii.com/tech/pubs/cell_hughes.htm`. Data from the treatment and mutant experiments were concatenated with the control ("wild-type") experiments. To facilitate direct comparison of our results with previously published work, profiles were selected from the raw data to include only experiments with 2 or more genes up- or down-regulated by more than 3-fold, and significant at $P \leq 0.01$ under a gene-specific error model, as described by Hughes et al. [4]; and to include only genes that were up- or down-regulated more than 3-fold, significant at $P \leq 0.01$ in 2 or more experiments. Following Hughes et al. [4], missing data was replaced by row (column) means[1]. The final data set comprised 636 genes and 194 experiments (including controls).

C. Computational Experiments

For all data sets the dimensionality of the data was first reduced by projecting the data onto the 10 leading eigen-directions of the correlation coefficient matrix. These 10 directions captured most of the variance in the data. This 10 dimensional projection of the data, $\mathbf{y}$, was then modeled with the Dirichlet process mixture model. A fully Bayesian approach to choosing the number of dimensions of the low dimensional projection is beyond the scope of this paper, however one possibility would be based on defining a Dirichlet process mixture of factor analyzers, which combines clustering with dimensionality reduction [23]. We have experimented with using 5 and 15 directions in the projection; in both cases the inference algorithm discovers fewer represented mixture components.

The parameters of the model were assigned prior distributions following [6]. The priors on the parameters of the Gaussian mixtures were *conditionally conjugate*, specifically Gaussian for

[1]We note that a full Bayesian treatment of missing data would involve integrating over the missing values.

the means and Wishart for the covariances (with top level parameters set to the moments of the data, such that the entire procedure is insensitive to translation, rotation and rescaling of the data). The prior on the concentration parameter was chosen to be vague, identical with [6].

The mixture model was initialized with all data belonging to a single Gaussian, and a large number of Gibbs sampling sweeps are performed, updating all variables and parameters, i.e. $\{\{\mu_j, \Sigma_j\}, \{c_i\}, \alpha\}$, in turn by sampling from the conditional distributions derived in the previous sections and described in more detail in [6]. To assess the mixing time, we examined the auto-correlation coefficients for the number of represented components, see Figure 1. We chose the number of represented components as a diagnostic, as this is one of the properties of the state which changes most slowly. We estimated the mixing time as the sum of the auto-correlation coefficients from a large negative lag to large positive lag. For the transcript response clustering experiment, shown in Figure 1, the mixing time is about 200. We then ran the final MCMC to generate 100 roughly independent samples, by using a burn-in of 10,000 samples, and then saving every 1000'th sample for the next 100,000 samples. This took 34 minutes on a desktop computer. For the clustering of experimental conditions, a similar strategy reveals a somewhat slower mixing time of 60,000. We thus ran the chain initially for 100,000 iterations for burn in, and then for 11,000,000 samples, keeping every 100,000th to get 100 roughly independent samples. This takes about 11 hours on a desktop, but the results of a 100 times shorter run (6 minutes) are virtually indistinguishable.

D. Visualization of Results

We wish to determine the probability that two genes belong to the same cluster, i.e. have similar functional roles or are influenced by a common upstream factor. Unlike methods based on a single clustering of the data, the approach described in this paper computes this probability while taking into account all sources of model uncertainty (including number of clusters and location of clusters). Specifically, we use the probability p_{ij} that two genes i and j belong to the same cluster in the Dirichlet process mixture model as a measure of the similarity of these gene expression profiles. Conversely $1 - p_{ij}$ defines a *dissimilarity measure*, which for the purposes of visualization, can be input to one of the standard linkage algorithms used for hierarchical clustering (Figure 6). We can easily compare the dendrograms thus obtained to the usual hierarchical clustering approach, which computes distance metrics directly on the gene

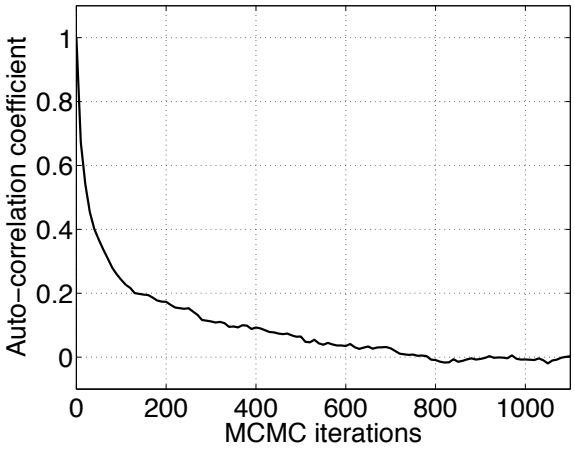


Fig. 1. Auto-correlation coefficient function for the clustering of genes experiment. The auto-correlation coefficient function for the number of represented components for the clustering of genes experiment. The function is only shown for positive lags, but is symmetric. The area under the curve (including both sides) is about 200.

expression profiles or correlation coefficients between profiles [1]. Clustering is done in both directions: both by gene transcripts and by experimental profiles.

E. Annotation of clusters by Gene Ontology

An important first step towards obtaining a functional profile of a gene list is to cluster the genes in terms of a comprehensive, well-structured set of functional categories such as that defined by the Gene Ontology (GO) Database. GO provides three structured ontologies of defined terms to describe gene product attributes: biological process, molecular function and cell component. Groups annotated at the highest level in the GO hierarchy (biological process) are likely to contain genes involved in related pathways. In order to find statistically significant GO annotations related to a given cluster of genes, we looked for annotation terms that are over-represented in this cluster. The probability that this over-representation is not found by chance can be calculated by the use of a hypergeometric test. Because of the effects of multiple testing, a subsequent correction of the p-values is necessary, and we used the SGD GO Term Finder `http://db.yeastgenome.org/cgi-bin/GO/goTermFinder` [24], which applies a Bonferroni correction.

After identifying clusters and their members, the SGD GO Term Finder was used to determine whether clusters were overrepresented by particular cellular localization, molecular function, or

molecular process GO terms. Absolute p-value depends on size of clusters and the size of the reference list, in this case all yeast ORFs with an assigned GO term. The set of experimental clusters shrinks when we exclude double mutants, chemical treatments, and wild type profiles. It should also be noted that SGD GO Term Finder does not calculate underrepresented GO terms and this has not been considered here. It can be seen for some clusters that the assigned GO term may be either too specific or too general. For example, cluster 15 of the clustering of experiments has as its top molecular process GO term "physiological process", a high-level ontology but not insightful. For the same cluster, the best molecular function GO term is given as "hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides" – this is a low-level, highly specific function yet also not immediately insightful. Rather than focussing on the best hit alone, all significant GO terms are used to provide insight (see Supplemental Material, Tables 1-6).

III. Results and Discussion

A. *Clustering by transcript response*

In all, 636 transcripts were found to meet the prefiltering criteria described in the Methods section. That is, these genes are those most affected by the gene knockouts/treatments which constitute the experimental conditions. In Figure 2 we show the relative frequency of the number of represented components over the MCMC samples. It shows that between 40 and 70 components are likely. This wide range of number of clusters underline our premise, that the individual clusterings found are associated with substantial uncertainties. Rather than picking one particular clustering, in the following we always visualize properties averaged over all states sampled by MCMC.

In Figure 3 we show the number of times, out of 100 samples, that the *indicator variables* for two genes were equal. As described in the Methods section, this may be interpreted as the probability p_{ij} that two genes i and j belong to the same cluster, and the different colours represent this probability. We refer to p_{ij} as the *co-occurrence probability* of genes i and j. The granularity of this clustering is determined by the data and not by some user-defined threshold. Large solid blocks of color along the diagonal correspond to homogeneous clusters. Note that in our method, sequences may partially belong to more than one cluster; off-diagonal elements indicate such 'cross-clustering' or overlapping clusters. These off-diagonal blocks (such as cluster

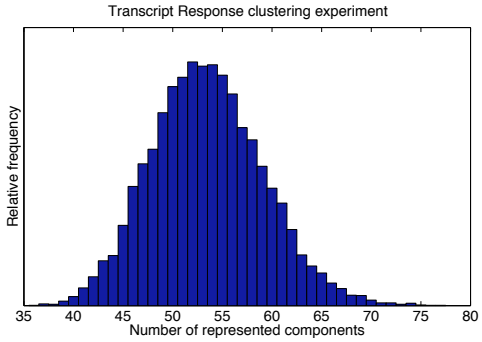


Fig. 2. Number of represented components for the clustering of transcript responses experiment. Figure showing that the relative number of components along the MCMC run varies between about 40 and 70.

2 or 4 in Figure 3) may indicate one of two possibilities; it may mean that there is *uncertainty* in whether a set of genes should be assigned to one of the two clusters, or it may indicate a set of genes which should really belong *simultaneously* to two clusters. In this latter case the fundamental assumption that a gene belongs to only one cluster does not apply, and suggests the existence of overlapping regulatory pathways. We focus on 17 transcript response clusters (TCs) represented as blocks of color along the diagonal (cluster members are given in Table I). Of these, 11 clusters form a single group along the diagonal, whilst in 5 cases, the clusters are broken into subclusters (clusters 2, 4, 9, 12 and 15). These are seen as mirrored bands above the background color (dark blue) and off the diagonal. The subclusters indicate that, while their members are most closely linked, there is also simultaneously a weaker affinity for other clusters. Using the SGD GO Term Finder, we determined overrepresented GO terms for each of the 17 transcript clusters. The top GO term and the p-value for each TC is given in Table III. Significance is defined as $p < 10^{-2}$.

Hughes et al. [4] applied agglomerative hierarchical clustering using a correlation coefficient based distance metric [1]. They identified eight main transcript response clusters: PAU; RNR2,3,4; ergosterol; amino acid biosynthesis; calcineurin/PKC; mitochondrial function, mating, and S/C (general stress response and carbohydrate metabolism). The PAU cluster includes a family of genes noted only for their lack of serine residues, and for being induced during anaerobic growth, but which otherwise do not have a known function [25], [26]. The RNR cluster represents genes that respond to DNA damage. The following TCs in Figure 3 appear to

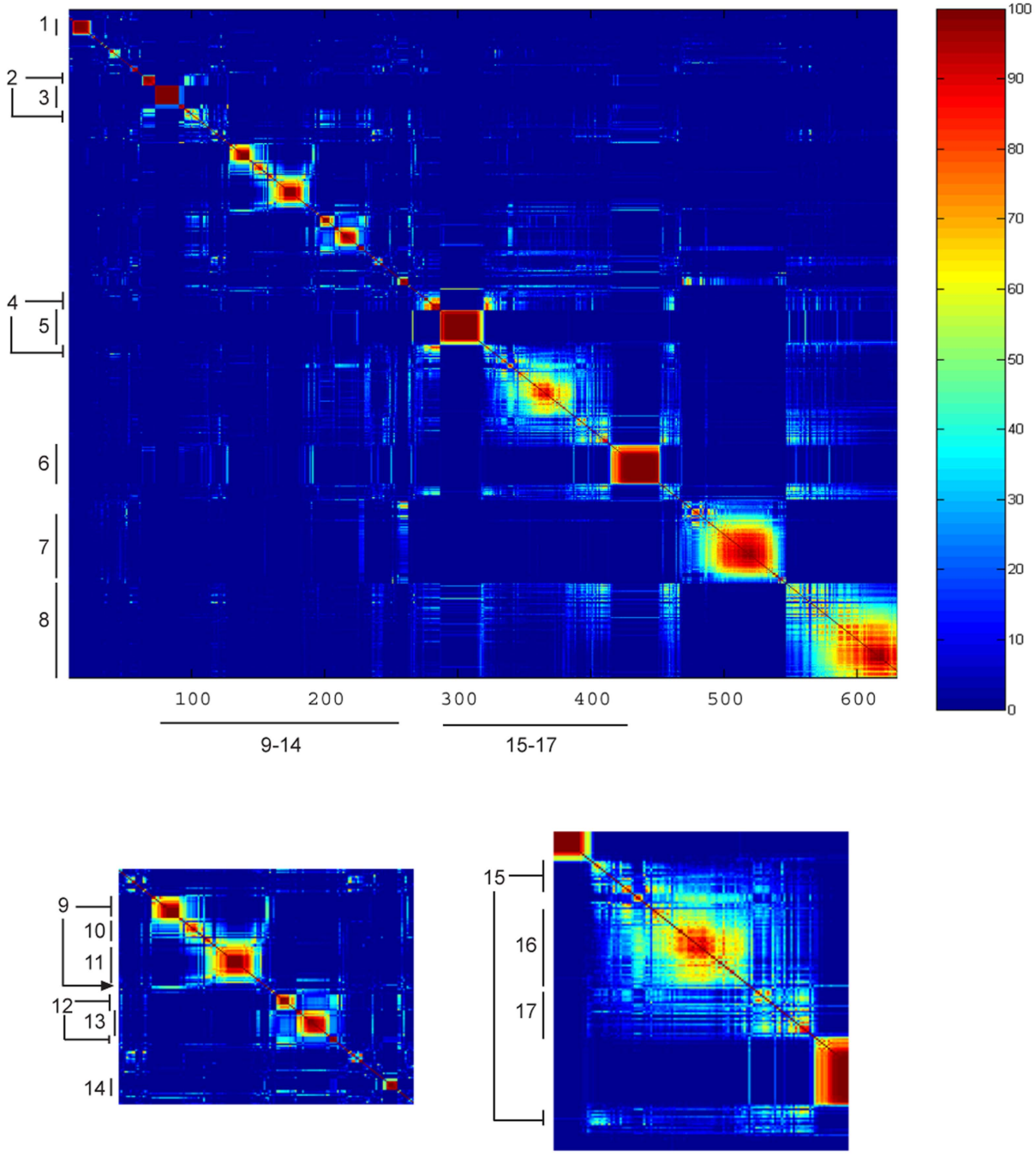


Fig. 3. Co-occurrence probabilities of the 636 transcript response clusters. Figure showing the number of times, out of 100 samples, that the *indicator variables* for two genes were equal. This may be interpreted as the probability p_{ij} that two genes i and j belong to the same cluster, and the different colors represent this probability. Numbers 1–17 indicated in the margins refer to the Transcript Clusters (TC's) discussed in detail in the text. Sub-figures represent a magnified view of portions of the larger figure. A larger version of this figure is available in the Supplementary Materials.

Table 1. List of ORFs in transcript clusters (TCs)

Cluster 1	
YPR002W	PDH1
YGR035C	
YMR102C	
YOR136W	IDH2
YDR406W	PDR15
YOR135C	
YLR304C	ACO1
YOR153W	PDR5
YNL037C	IDH1
YLR346C	
YOR049C	RSB1
YAL061W	
YBL043W	ECM13
YNR056C	BIO5
Cluster 2	
YLR042C	
YOR247W	SRL1
YKR013W	PRY2
YPL163C	SVS1
YIL123W	SIM1
YPL256C	CLN2
YJL158C	CIS3
YOR248W	
YGR014W	MSB2
YDR309C	GIC2
YGR189C	CRH1
YLR194C	
YKR091W	SRL3
YKR061W	KTR2
YDR077W	SED1
YPL067C	
YHR030C	SLT2
YLR121C	YPS3
YPR078C	
YHR209W	
YDR085C	AFR1
YNL034W	
YJL027C	
YGR156W	PTI1
YKL163W	PIR3
YOL011W	PLB3
YEL021W	URA3
YLR391W-A	
Cluster 3	
YLR343W	GAS2
YBL101W-B	
YCL019W	
YIL060W	
YBL005W-B	
YAR009C	
YER138C	
YBR012W-B	
YMR050C	
YMR045C	

Cluster 3 (continued)	
YOL106W	
YER160C	
YLR334C	
YML045W	
YJR029W	
YJR027W	
YHR214C-B	
YML039W	
YLR035C-A	
YFL002W-A	
YDR170W-A	
YHR213W	
Cluster 4	
YDL037C	BSC1
YDL039C	PRM7
YDL038C	
YCR021C	HSP30
YDR516C	EMI2
YER066C-A	
YDR343C	HXT6
YFR053C	HXK1
YDR342C	HXT7
YER067W	
YPR160W	GPH1
YOL150C	
YFL060C	SNO3
YEL011W	GLC3
YBR183W	YPC1
YBL049W	MOH1
YPL230W	
YDR277C	MTH1
YBL064C	PRX1
YML128C	MSC1
YLL026W	
HSP104	
Cluster 5	
YHR092C	HXT4
YBR066C	NRG2
YLL025W	PAU17
YDR213W	UPC2
YOR394W	PAU21
YJR150C	DAN1
YPL282C	PAU22
YHR139C	SPS100
YJL223C	PAU1
YIR041W	PAU15
YKL224C	PAU16
YCR104W	PAU3
YLL064C	PAU18
YEL049W	PAU2
YGR294W	PAU12
YIL176C	PAU14
YOR010C	TIR2
YLR461W	PAU4

Cluster 5 (continued)	
YLR037C	DAN2
YFL020C	PAU5
YBR301W	DAN3
YGL261C	PAU11
YOL161C	PAU20
YDR542W	PAU10
YHL046C	PAU13
YNR076W	PAU6
YMR325W	PAU19
YOR009W	TIR4
YAL068C	PAU8
YGR213C	RTA1
YJL114W	
YIL175W	
YMR316C-B	
YGR144W	THI4
Cluster 6	
YGL183C	MND1
YNL180C	RHO5
YGR040W	KSS1
YKL178C	STE3
YHR145C	
YOL104C	NDJ1
YLR040C	
YJR004C	SAG1
YPL192C	PRM3
YIL082W	
YCL027W	FUS1
YCL055W	KAR4
YDR124W	
YHR005C	GPA1
YMR065W	KAR5
YIL015W	BAR1
YLR452C	SST2
YNR044W	AGA1
YBL016W	FUS3
YJL157C	FAR1
YGL032C	AGA2
YCR089W	FIG2
YBR083W	TEC1
YCLX07W	
YDR461W	MFA1
YFL026W	STE2
YKL209C	STE6
YCRX18C	
YML048W-A	
YNL279W	PRM1
YNL145W	MFA2
YMR232W	FUS2
YIL080W	
YIL082W-A	
YIL037C	PRM2
YJL170C	ASG7
YIL011W	TIR3
YBR250W	

Cluster 7	
YOR107W	RGS2
YLR237W	THI7
YNR064C	
YFR047C	BNA6
YBR045C	GIP1
YKL218C	SRY1
YJL217W	
YIL164C	NIT1
YBR294W	SUL1
YOL064C	MET22
YJL089W	SIP4
YGL009C	LEU1
YNR069C	BSC5
YER081W	SER3
YPL135W	ISU1
YBR105C	VID24
YIL056W	VHR1
YNR058W	BIO3
YOR130C	ORT1
YJR130C	STR2
YGL180W	ATG1
YKL120W	OAC1
YLR162W	
YKL121W	
YOR303W	CPA1
YDL170W	UGA3
YJR154W	
YOR337W	TEA1
YOR339C	UBC11
YHR208W	BAT1
YGR239C	PEX21
YIL165C	
YGL125W	MET13
YJR155W	AAD10
YER091C	MET6
YDL198C	GGC1
YNL104C	LEU4
YDR127W	ARO1
YJR109C	CPA2
YOL140W	ARG8
YNR050C	LYS9
YHR162W	
YER024W	YAT2
YLR267W	BOP2
YER073W	ALD5
YHR029C	YHI9
YBR248C	HIS7
YDR158W	HOM2
YOR203W	
YLR152C	
YDR035W	ARO3
YMR097C	MTG1
YJR111C	
YMR108W	ILV2
YPL250C	ICY2
YER052C	HOM3

TABLE I

LIST OF ORFS IN TRANSCRIPT CLUSTERS (TCS)

Table 2. List of mutants in experimental conditions clusters (ECs)

Cluster 1	**Cluster 8**	**Cluster 13/14**	**Cluster 16 (continued)**
ssn6	aj307-vs.-aj308	2-deoxy-D-glucose	ca1105-vs.-ca1106
tup1	ca721-vs.-ca702	anp1-	ca1107-vs.-ca1108
	fus3-, kss1- (haploid)	AUR1 (TET promoter)	ca1109-vs.-ca1110
Cluster 2	med2- (haploid)	clb2	ca1133-vs.-ca1134
CDC42 (TET promoter)	sgt2	CMD1 (TET promoter)	ca1135-vs.-ca1136
KAR2 (TET promoter)	sod1- (haploid)	erg4- (haploid)	ca1167-vs.-ca1168
	ste11- (haploid)	fks1- (haploid)	ca1169-vs.-ca1170
Cluster 3	ste12- (haploid)	FKS1 (TET promoter)	ca1171-vs.-ca1172
HU	ste12- (haploid)	gas1	ca1189-vs.-ca1190
rad6	ste18- (haploid)	Glucosamine	ca1191-vs.-ca1192
rnr1	ste4- (haploid)	hst3	ca1290-vs.-ca1289
swi6	ste5- (haploid)	kin3	ca1296-vs.-ca1295
	ste7- (haploid)	rad57	ca1332-vs.-ca1331
Cluster 4	yjl107c- (haploid)	she4	ca1334-vs.-ca1333
rpd3		spf1	ca1369-vs.-ca1368
sin3	**Cluster 9**	swi4	ca1408-vs.-ca1407
	cup5	swi5	ca1410-vs.-ca1409
Cluster 5	qcr2- (haploid)	yar014c	ca1448-vs.-ca1447
dig1	rip1		ca1450-vs.-ca1449
dig1, dig2	vma8	**Cluster 15**	ca1488-vs.-ca1487
dig1-, dig2- (haploid)		aep2	ca1490-vs.-ca1489
fus3- (haploid)	**Cluster 10**	afg3- (haploid)	ca1492-vs.-ca1491
hda1	Tunicamycin	ard1	ca1547-vs.-ca1546
hog1- (haploid)	yer083c	ase1	ca1549-vs.-ca1548
sst2- (haploid)		bub3- (haploid)	ca1601-vs.-ca1600
yor080w	**Cluster 11**	ca719-vs.-ca700	ca753-vs.-ca752
	ade2	cem1	ca755-vs.-ca754
Cluster 6	aj307-vs.-aj308	cyt1	ca775-vs.-ca774
dot4	bim1	imp2	ca789-vs.-ca788
mrt4	bub1	kim4	ca791-vs.-ca790
rpl27a-	bub3	mac1	ca827-vs.-ca826
rps24a-	bul1	mrp133	ca841-vs.-ca840
rps24a- (haploid)	cka2	msu1	ca843-vs.-ca842
rps27b-	erg4	rml2-	ca926-vs.-ca927
rrp6	pfd2	yap1	ca931-vs.-ca930
sir4	rtg1	yer050c	ca994-vs.-ca993
yel033w	rts1	yhl029c	cs1412vs.-ca1411
yel044w	vac8	yhr011w-	ds1242-vs.-ds1241
yhr034c	vps8	ymr031w-a	ds1244-vs.-ds1243
ymr014w		ymr293c	ds1286-vs.-ds1285
ymr269w	**Cluster 12**		ds1288-vs.-ds1287
	isw1	**Cluster 16**	ds1308-vs.-ds1307
Cluster 7	isw1, isw2	aj318-vs.-aj317	ds1316-vs.-ds1315
ca884-vs.-ca881	isw2	aj324-vs.-aj323	ds720-vs.-ds719
Calcofluorwhite	ras2- (haploid)	aj338-vs.-aj337	ds798-.vs.-ds797
ERG11- (TET promoter)		arg80	ds800-vs.-ds799
erg2		ca1047-vs.-ca1048	ds866-vs.-ds865
erg3- (haploid)		ca1081-vs.-ca1082	ds904-vs.-ds903
hmg1- (haploid)		ca1083-vs.-ca1084	ds906-vs.-ds905
HMG2 (TET promoter)			ecm10
imp2'			gln2
Itraconazole			npr2
Lovastatin			nta1
sir2			pex12
Terbinafine			ppr1
top3- (haploid)			sir3
yer044c			

TABLE II

LIST OF MUTANTS IN EXPERIMENTAL CONDITIONS CLUSTERS (ECS)

Table 3. Clustering by transcript profiles. 636 transcript profiles used. 515 placed in clusters.

Cluster	#ORFs	Function	p	Process	p	Component	p
1	14	isocitrate dehydrogenase (NAD+) activity	6.83E-06	glutamate biosynthesis	2.03E-06	mitochondrial nucleoid	8.800E-04
2	28	structural constituent of cell wall	5.78E-07	cell wall organization and biogenesis	1.79E-05	cell wall	8.140E-12
3	22	RNA-directed DNA polymerase activity	1.02E-27	Ty element transposition	2.23E-28	retrotransposon nucleocapsid	1.860E-28
4	21	fructose transporter activity	8.60E-04	monosaccharide transport	2.27E-05	plasma membrane	3.144E-02
5	34	molecular function unknown	3.85E-08	biological process unknown	2.40E-07	cellular component unknown	1.170E-06
6	38	cell adhesion molecule binding	5.81E-07	conjugation	2.14E-23	mating projection tip	1.750E-08
7	82	catalytic activity	1.78E-06	amino acid biosynthesis	1.74E-31	carbamoyl-phosphate synthase complex	2.400E-04
8	91	sugar transporter activity	2.79E-11	carbohydrate transport	3.42E-10	cellular component unknown	3.550E-06
9	19	iron ion transporter activity	8.93E-11	siderophore transport	4.02E-19	endosome	3.100E-05
10	10	cyclin-dependent protein kinase regulator activity	4.40E-04	regulation of cyclin dependent protein kinase activity	1.60E-04	endoplasmic reticulum	7.396E-02
11	32	hydrolase activity, hydrolyzing O-glycosyl compounds	4.53E-06	cytokinesis, completion of separation	1.61E-08	cell wall (sensu Fungi)	2.510E-08
12	14	endopeptidase activity	6.14E-03	cell wall organization and biogenesis	3.19E-02	plasma membrane	7.754E-02
13	18	monooxygenase activity	7.14E-07	steroid biosynthesis	6.23E-19	endoplasmic reticulum	4.400E-11
14	8	oxidoreductase activity, acting on sulfur group of donors	2.48E-07	sulfur utilization	2.71E-13	sulfite reductase complex (NADPH)	2.100E-06
15	17	oxidoreductase activity, acting on the aldehyde or oxo group of donors	1.54E-03	vitamin metabolism	1.40E-02	storage vacuole	4.414E-02
16	42	protease inhibitor activity	1.86E-06	beta-alanine biosynthesis	6.43E-05	cytoplasm	1.090E-03
17	25	polyamine transporter activity	1.57E-03	polyamine transport	5.50E-04	vacuole	3.830E-05

TABLE III

SUMMARY OF SGD GO ANNOTATIONS FOR TRANSCRIPT CLUSTERS

match with the following groups found by Hughes et al.: PAU (TC 5), RNR (TC 3), ergosterol (TC 13), mitochondrial function (TC1), and mating (TC 6). The other clusters described by Hughes et al., in particular the S/C cluster and amino acid biosynthesis cluster, are distributed over several TC clusters. In particular TC4 (monosaccharide transport), TC7 (general amino acid biosynthesis), TC 8 (carbohydrate transport) TC 14 (sulfer metabolism), TC15 (vitamin metabolism), TC16 (beta-alanine biosynthesis), and TC 17 (polyamine transport). As such, the DPM method was able to distribute the general S/C and amino acid biosynthesis groups into more specific clusters.

TC2, TC11, and TC12 all exhibit significance for "cell wall", "plasma membrane", and "cytokinesis" GO terms. Examination of the cluster members suggest TC2 is involved in the formation of the mating bud. The best process GO term associated with TC11 is "cytokinesis, completion of separation". TC12 is associated with process GO term "cell wall organization and biogenesis". We note that for TC 5, the best hit for all three GO categories is "unknown". Cluster 5 is a large group (32 transcripts) and contains 20 out of 21 PAU genes (PAU7 appears in TC 8). TC5 also contains five DAN/TIR mannoproteins genes, which are typically part of the cell wall. This is in agreement with work indicating the importance of these sets of ORFs in cell wall integrity [27], suggesting that TC5 is yet another "cell wall" cluster. This identification of a new cluster of "cell wall" transcripts makes sense in light of the clustering of experimental conditions described below. While Hughes et al. identified a group of profiles collectively related to "cell wall", the DPM clustering suggests that this large group forms smaller, distinctly regulated subclusters. Recent literature looking at cell wall proteins suggests that distinct functions – for example, controlling osmotic pressure, responding to physical stress, maintaining cell wall integrity and providing a protein scaffold – may share proteins and have overlapping regulatory mechanisms [27]. Furthermore, the signaling pathways involve crosstalk among MAPK kinase pathways [28]. For example, sets of cell wall proteins, such as the PAU family, are activated by pheremone signaling, by global stress signaling, as well as the calcineurin-mediated signaling, suggesting multiple modes of regulation.

Likewise, rather than finding a single large group of transcripts specific to the PKC/calcineurin as in [4], we find this group split amongst other TCs. Hughes et al. identified this group as comprising genes activated when yeast are treated with FK506 or cyclosporin-A. Both compounds affect calcineurin, a serine/threonine phosphatase implicated in intracellular ion homeostasis, adaptation to mating pheromone treatment, and mitosis. However, the two compounds are thought to act through different pathways. Hughes et al. list 42 transcripts as part of this PKC/calcineurin gene cluster. Of these, we find 31 in five different TCs. 10 transcripts are found in TC2 (cell wall), 11 in TC12 (cell wall), 8 in TC 16 (beta-alanine biosynthesis), and one each in TC4 (monosaccharide transport) and TC5. It is known that PKC is part of a MAPK cascade involved in cell wall integrity. It has crosstalk with other MAPK cascades including pheromone response, osmolarity, and filamentous growth. All told, five of the 17 TCs are associated with the cell wall. Recent work indicates that beyond providing structural support, components of the cell

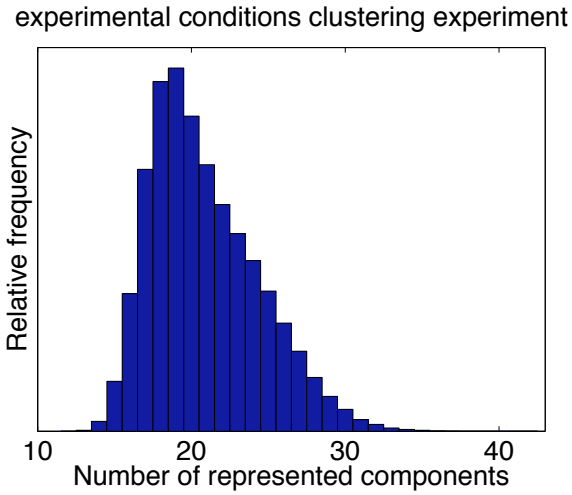


Fig. 4. Number of represented components for the clustering of the experimental conditions. Figure showing that the relative number of components along the MCMC run varies between about 15 and 30.

wall are involved in diverse functions from uptake of nutrients/metabolism to energy generation [28]. Likewise, formation of the shmoo during mating involves not only signal transduction by mating factor but rearrangement of the cytoskeleton and cell wall.

Finally, we identified a cluster (TC9) that does not appear to be covered by those defined by Hughes et al. The best GO term matches are “siderophore transport” (process GO), “iron ion transporter activity” (function GO), and “endosome” (component GO).

B. Clustering by Experimental Conditions

Clustering of the expression profiles by experimental conditions identifies those yeast mutants or compounds that have similar effects on all transcripts. In Figure 4 we show that a minimum of about 15 components are necessary, and the data supports up to about 30.

Figure 5 shows the clustering of the experimental conditions, which has an interpretation similar to that of Figure 3. After prefiltering the 300 compendium experiments, 194 expression profiles including 60 “wild types” remained. “Wild types” represent control experiments testing neither chemical treatment nor gene knockout, but yet had at least one ORF whose expression changed more than 2-fold. (These were explicitly excluded from the cluster analysis of Hughes et al.)

From Figure 5, 16 experimental condition clusters (ECs) are apparent. This is in contrast to

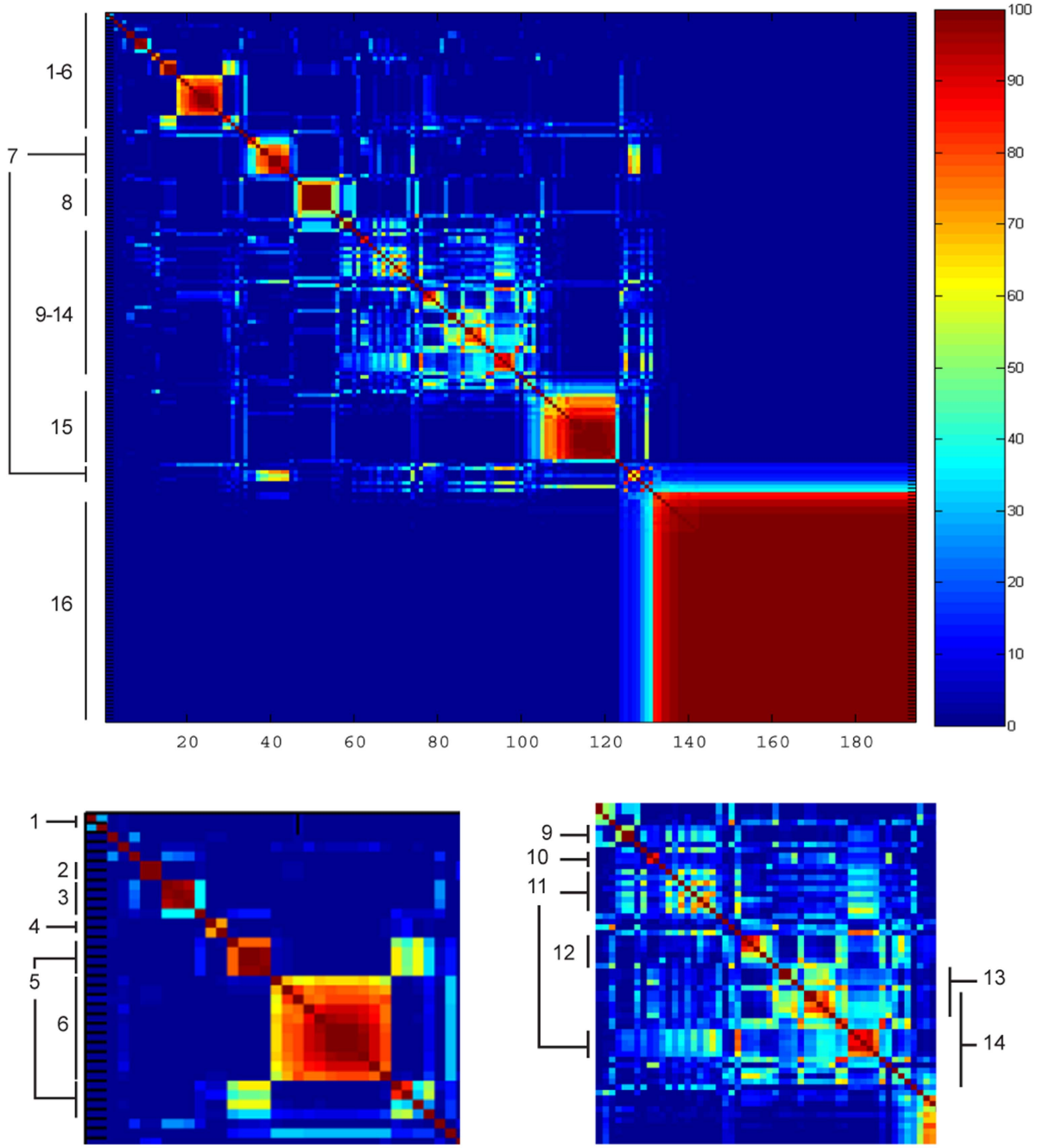


Fig. 5. Co-occurrence probabilities of the 194 experimental conditions clusters. Figure showing the number of times, out of 100 samples, that the *indicator variables* for two experimental conditions were equal. This may be interpreted as the probability p_{ij} that two experimental conditions i and j belong to the same cluster, and the different colors represent this probability. Numbers 1–16 indicated in the margins refer to the Experimental Condition Clusters (EC's) discussed in detail in the text. Sub-figures represent a magnified view of portions of the larger figure. A larger version of this figure is available in the Supplementary Materials.

Table 4. Clustering by experiment/condition. 194 experiment profiles used. 143 profiles placed in clusters.

Cluster	#ORFs	Function	p	Process	p	Component	p
1	2	general transcriptional repressor activity	7.52E-08	nucleosome spacing	1.69E-07	nucleus	7.40E-02
2	2	nucleoside-triphosphatase activity	1.03E-03	conjugation	2.00E-04	none	
3	3	-	-	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	9.15E-03	protein complex	9.97E-02
4	2	histone deacetylase activity	5.43E-06	chromatin silencing at rDNA	3.68E-06	histone deacetylase complex	7.52E-06
5	7	MAP kinase activity	9.85E-06	invasive growth (sensu Saccharomyces)	1.63E-08	nucleus	1.88E-02
6	12	structural molecule activity	2.11E-03	rRNA processing	2.42E-03	non-membrane-bound organelle	3.90E-04
7	8	hydroxymethylglutaryl-CoA reductase (NADPH) activity	2.10E-06	ergosterol metabolism	4.52E-14	endoplasmic reticulum	2.38E-07
8	12	receptor signaling protein activity	8.75E-09	invasive growth (sensu Saccharomyces)	3.72E-14	mating projection	5.76E-07
9	5	hydrogen ion transporter activity	4.01E-06	hydrogen ion homeostasis	9.88E-05	hydrogen-translocating V-type ATPase complex	4.21E-05
10		-	-	-	-	-	-
11	12	protein binding	7.70E-04	spindle checkpoint	4.45E-06	kinetochore	8.50E-05
12	4	nucleoside-triphosphatase activity	1.20E-04	chromatin remodeling	1.94E-03	chromatin remodeling complex	5.90E-04
13	10	transferase activity, transferring hexosyl groups	4.74E-03	protein amino acid glycosylation	3.83E-03	incipient bud site	1.01E-03
14		transferase activity	4.47E-02	protein amino acid glycosylation	3.08E-03	-	-
15	23	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	5.75E-03	physiological process	6.27E-05	mitochondrion	2.23E-05
16	41			wild type			

- : no value determined.

TABLE IV

SUMMARY OF SGD GO ANNOTATIONS FOR EXPERIMENTAL CONDITIONS CLUSTERS

the 13 identified by Hughes et al. [4]. As with the transcript response clustering, it can be seen that some clusters are bipartite (eg., ECs 5, 7, 11), and there is a region of diffuse clusters (ECs 9-14). Closer examination suggests there may be smaller clusters within this region. Also, two clusters (EC13 and 14) may be considered to be overlapping. In addition, a dendrogram using the dissimilarity measure defined above is shown in Figure 6, which may be compared to Figure 3B in the supplementary material of [4].

Apart from EC1, other ECs correspond closely, although not exactly, to those identified by Hughes et al. For example, the Hughes et al. cluster *rnr1/HU* overlaps with our EC 3

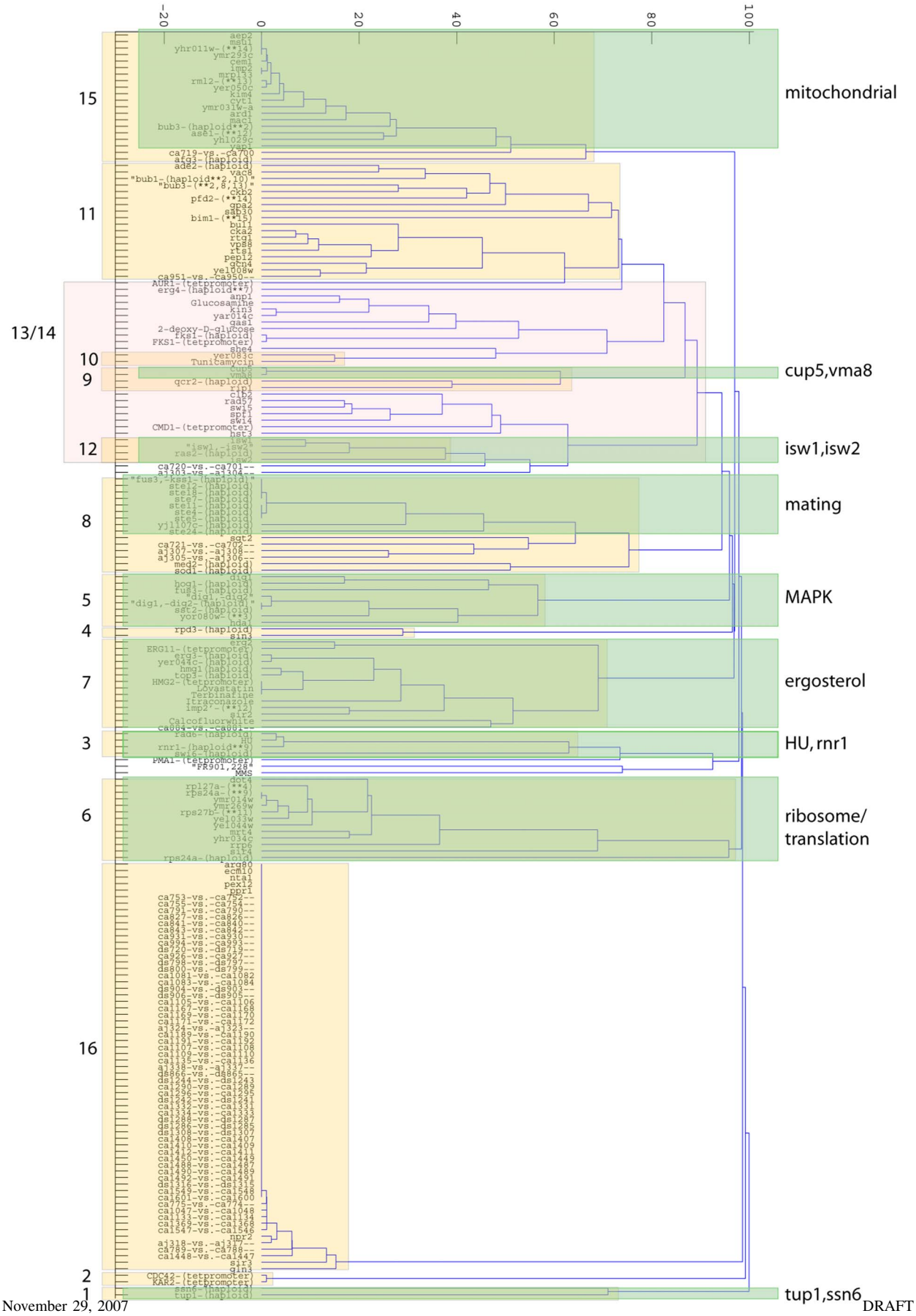


Fig. 6. Dendrogram of the dissimilarity measure of the 194 experimental conditions clusters. $1 - p_{ij}$ defines a

Table 5. SGD GO annotation of Hughes clustering by experiment. 77 experiment profiles placed in clusters.

	Hughes assignment	#ORFs	Function	p	Process	p	Component	p
1	mitochondrial function	17	oxidoreductase activity, acting on diphenols and related substances as donors, cytochrome as acceptor	2.00E-04	mitochondrial electron transport, ubiquinol to cytochrome c	1.26E-06	mitochondrion	2.38E-11
2	cell wall	9	catalytic activity	2.05E-03	budding cell apical bud growth	8.13E-05	actin cap	1.50E-04
3	protein synthesis	12	structural constituent of ribosome	4.00E-04	ribosome biogenesis and assembly	5.40E-04	intracellular non-membrane-bound organelle	3.80E-05
4	ergosterol biosynthesis	8	hydroxymethylglutaryl-CoA reductase (NADPH) activity	1.12E-06	ergosterol metabolism	1.62E-15	endoplasmic reticulum	9.24E-09
5	mating	8	receptor signaling protein activity	1.24E-09	invasive growth (sensu Saccharomyces)	3.82E-16	cell projection	8.31E-08
6	MAPK activation	6	MAP kinase activity	7.04E-06	filamentous growth	3.56E-07	plasma membrane	1.55E-02
7	rnr1/HU	3	none		nucleobase, nucleoside, nucleotide and nucleic acid metabolism	9.15E-03	protein complex	9.97E-02
8	histone deacetylase	3	histone deacetylase activity	1.26E-08	chromatin silencing at rDNA	7.07E-09	histone deacetylase complex	2.06E-08
9	isw	3	ATPase activity	1.68E-03	chromatin remodeling	6.07E-06	chromatin remodeling complex	2.90E-04
10	vacuolar ATPase/iron regulation	3	hydrogen ion transporter activity	1.60E-04	cation homeostasis	1.94E-06	hydrogen-translocating V-type ATPase complex	1.26E-05
11	sir	3	histone binding	1.65E-10	loss of chromatin silencing during replicative cell aging	1.32E-09	nuclear telomeric heterochromatin	8.84E-10
12	tup1/ssn6	2	general transcriptional repressor activity	7.52E-08	nucleosome spacing	1.69E-07	nucleus	7.40E-02

TABLE V

SUMMARY OF SGD GO ANNOTATIONS FOR EXPERIMENTAL CONDITIONS CLUSTERS DESCRIBED BY HUGHES ET AL. (2000)

with the exception of MMS. We both find a histone deacetylase group (EC4), an ergosterol biosynthesis group (EC 7), a mating group (EC 8), a V-ATPase/iron regulation group (EC9), and a mitochondrial group (EC15). The “ribosome/translation” group identified by Hughes et al. overlaps with EC 6, which is associated with the molecular process GO term of “rRNA processing”.

A major difference between Hughes et al. and our DPM results involves profiles identified as “cell wall”. Hughes et al. identified 13 expression profiles as part of a “cell wall” group. However we find three distinct clusters within this group. Knockouts for two tetracycline-driven

genes, *tet-KAR2* and *tet-CDC42*, cluster together as EC2 with a co-occurrence probability close to 100%; this cluster does not overlap with any other. In addition, tunicamycin and yer083 form a cluster (identified as EC10) with a co-occurrence probability around 85%, clearly apart from other profiles. Tunicamycin is thought to disrupt protein glycosylation in yeast [29] while yer083c has recently been identified as localized to the ER and involved in trafficking cell wall proteins [30], [31]. The remaining members appear in EC13 which is associated with "incipient bud site" as its best component GO term. Thus while all 13 members do involve proteins associated with the cell wall, it may be seen that multiple processes or functions are being affected. Recent work has indicated the cell wall stress influences many genes through diverse signaling pathways and different transcription factors [27], [32].

Hughes et al. identify a single cluster containing the sir mutants. Sir proteins are involved in global gene regulation through chromatin restructuring. However by DPM clustering, we find each sir knockout in a different cluster: $sir2\Delta$ in EC7 (ergosterol), $sir3\Delta$ in EC16 (wild type), and $sir4\Delta$ in EC6 (rRNA processing). We note that assocation of sir2 with EC7 is at a co-occurrence probability of 60%, and association of $sir4\Delta$ with EC6 is at 30%. This suggests that while the SIR proteins are not strongly affiliated with any other group or each other globally, there may be a subset of specific transcripts that are strongly affected. It is possible that while there are few co-regulated transcripts, their regulation may be highly similar. The expression profile of the $sir2\Delta$ mutant is most similar to that of imp2' (YIL154C) at a co-occurrence probability close to 80%. Sir2p is involved in chromatin silencing; disruption causes problems with DNA repair while slight overexpression increases the lifespan of yeast and C. elegans [33], [34]. It is known that caloric restriction increases Sir2p activity. Imp2p is a transcription factor that activates galactose, maltose and raffinose utilization [35] as well as mediating oxidative damage to DNA [36]. Similarity in the expression profiles of these two mutants might be because the set of genes derepressed by the $sir2\Delta$ mutant overlap somewhat with those regulated by Imp2p. Alternately, both mutants might exhibit similar global effects.

The *isw1, isw2* group found by Hughes et al. contains four expression profiles (*isw1*, *isw2*, *isw1/2*, and *hst3*). We identify EC12 containing *isw1*, *isw1/isw2*, *isw2*, and *ras2* but instead put *hst3* as part of the larger EC13/14. The ISW proteins are ATPases and are likely part of a protein complex involved in chromatin remodeling [37]. Ras2p is a GTP-binding protein involved in nitrogen starvation response, sporulation, and filamentous growth [38]. Hst3p is part of the Sir

protein family of histone deacetylases and thought to be involved multiple functions including telomeric silencing [39]. As noted above, while Hughes et al. placed the Sir proteins into a single cluster, we find them distributed thought several clusters. However, examination of the dendrogram (Figure 6) indicates that EC12 may be considered a "subcluster" within the larger EC13/14 and is joined to the subcluster containing $hst3\Delta$.

C. Discussion

Although the use of clustering methods (in particular agglomerative hierarchical clustering) has rapidly become one of the standard computational approaches in the literature of microarray gene expression data, little attention has been paid to *uncertainty* in the results obtained.

Kerr and Churchill [40] have proposed the use of a bootstrap method to assess the results of clustering in a statistically quantifiable manner. However, their approach requires the fitting of a linear statistical (ANOVA) model to the microarray data to obtain least squares estimates of the differential expression of a given gene, which are then used as inputs to the bootstrap process. An alternative parametric bootstrap approach has been described by Zhang and Zhao [41] which uses estimates of the standard errors in gene expression measurements to simulate data from a log-normal distribution. Hughes et al. [4] describe a permutation procedure to calculate p-values for the significance of branching in a dendrogram produced by agglomerative hierarchical clustering, under the null hypothesis that the branching was not significant. However, hierarchical clustering is a bottom-up algorithm. It starts with each data point assigned to its own cluster and iteratively merges the two closest clusters together until all the data belongs to a single cluster. Consequently, the results presented by Hughes et al. (Figure 3B, supplementary information to [4]) only appear to show strong confidence for the branches at the lowest level of the dendrogram. In contrast, the dendrogram produced from the DPM procedure (Figure 6) represents a full probabilistic measure of the (dis)similarity of two gene expression profiles.

Dirichlet process mixture models provide a non-parametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Medvedovic and co-workers have applied infinite Gaussian (or Dirichlet process) mixture models to the clustering of time series gene expression data using spherical Gaussians with diagonal covariances [12], [13]. Similar approaches have also recently been described in [16]. However, these approaches do not explicitly model the correlations between subsequent time points which would be expected to

occur in time series data, and the use of diagonal covariances may result in more clusters than necessary to model such correlations. Lui et al. have recently extended their previous work to use full-covariance models for time series [14]. Since these authors are clustering short time series, inference in the space of the original data is feasible. In contrast, in the complementary approach we describe here, we apply the DPM method to high-dimensional non-time series data. Inference is carried out in a low dimensional projection of this space after dimensionality reduction by principal component analysis, which makes it possible to use Gaussians with full covariance matrices, which would be very computationally expensive in the original high dimensional space as each sampling step has a cubic computational dependency on the dimensionality.

Bayesian approaches to clustering gene expression data have recently received much attention. Heard, et al. [42] propose an agglomerative clustering procedure for gene expression time series curves based on a Bayesian merging score, but unrelated to DPMs. Heller and Ghahramani [43] proposed a different Bayesian hierarchical clustering (BHC) procedure which implements a non-MCMC inference procedure for DPMs. This BHC algorithm can be used to scale DPM learning and inference to very large data sets at the cost only partially representing the uncertainty in the cluster assignments. The MCMC procedure we present in this paper is more computationally demanding, but captures more completely the sources of uncertainty. In Lau and Green [44], model-based clustering procedures based on loss functions are derived. An integer program is identified for finding a single clustering that best matches the posterior co-occurrence probabilities.

Recently, Bidaut et al. [45] have re-analyzed the data of Hughes et al. using "Bayesian decomposition" to place the experimental profiles into patterns (clusters). The highest scoring (high persistence) genes in the patterns were annotated using the MIPS database [46] to assign the pattern to a cellular pathway. Fifteen patterns were discovered, six of which are assigned to MIPS pathways. Bidaut et al. find that $ssn6\Delta$ and $tup1\Delta$ appear in many of their patterns, albeit at low persistence. In contrast, with DPM modeling we find that $ssn6\Delta$ and $tup1\Delta$ cluster together although weakly (EC1 - co-occurrence probability of 30%) and apart from other experimental profiles. This is reinforced by the dendrogram (Figure 6) which shows while the $tup1\Delta$ and $ssn6\Delta$ profiles cluster away from the others, they are yet on very long branches from each other. Clustering of these two knockouts is supported by the fact that Tup1p and Ssn6p are thought to form a protein complex. As previously mentioned, both proteins are transcription

factors involved in glucose/catabolic repression although with different but overlapping sets of targets [39].

Patterns 13 and 15 identified by Bidaut et al. [45] are given significance as distinguishing between those genes involved in MAPK signaling mating versus those involved in filamentous growth. While these are two distinct cellular functions, they share signaling components. Bidaut et al. suggest these groups can be distinguished by whether the genes are regulated by Ste12p or the Ste12p-Tec1p complex. In our clustering of experimental conditions, all of the ste deletion mutants plus the $fus3\Delta, kss1\Delta$ double mutant cluster together (EC8 - component GO term: mating projection). The $fus3\Delta$ single mutant appears in EC8, together with other genes annotated by the GO molecular function term indicating MAPK activity. However, when we look at the top genes associated with the Bidaut patterns, 6 of the top 10 genes in pattern 13 are part of TC6 (component GO: mating projection tip) while 7 of the top 10 genes in pattern 15 are part of our TC3 (component GO: retrotransposon nucleocapsid).

IV. Conclusion

Dirichlet process mixture models provide a non-parametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Unlike methods based on a single clustering of the data, the approach computes the probability that two genes belong to the same cluster while taking into account the main sources of model uncertainty, including the number of clusters and the location of clusters. Biologically plausible results are obtained from the Rosetta compendium of expression profiles which extend previously published cluster analyses of this data. Our results confirm many of the previously published clusters identified in this data set, but also provide new biological insights by revealing a finer level of granularity in the clustering. In particular our method was able to distribute general stress response and carbohydrate metabolism and amino acid biosynthesis groups into more specific clusters. Whilst previous analyses have identified a group of profiles collectively related to cell wall functions, our results also suggest that this large group forms smaller, distinctly regulated subclusters. These results are consistent with recent literature on cell wall proteins which suggests that distinct functions – for example, controlling osmotic pressure, responding to physical stress, maintaining cell wall integrity and providing a protein scaffold – may share proteins and have overlapping regulatory mechanisms.

REFERENCES

[1] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression," *PNAS*, vol. 95, pp. 14 863–14 868, 1998.

[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl Acad. Sci*, vol. 96, pp. 6745–6750, 1999.

[3] G. McLachlan, R. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.

[4] T. Hughes, M. Marton, A. Jones, C. Roberts, R. Stoughton, C. Armour, H. Bennett, E. Coffey, H. Dai, Y. He, M. Kidd, A. King, M. Meyer, D. Slade, P. Lum, S. Stepaniants, D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. Friend, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, pp. 109–126, July 2000.

[5] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models'," *J. Comp. and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[6] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. MIT Press, 2000, pp. 554–560.

[7] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[8] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230., 1973.

[9] A. Y. Lo, "On a class of Bayesian nonparametric estimates. I. Density estimates." *Annals of Statistics*, vol. 12, pp. 351–357, 1984.

[10] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[11] D. L. Wild, C. E. Rasmussen, Z. Ghahramani, J. Cregg, B. J. de la Cruz, C.-C. Kan, and K. A. Scanlon, "A Bayesian approach to modelling uncertainty in gene expression clusters," in *3rd International Conference on Systems Biology, Stockholm, Sweden*, 2002.

[12] M. Medvedovic and S. Sivaganesan, "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206, 2002.

[13] M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner, "Bayesian mixture model based clustering of replicated microarray data," *Bioinformatics*, vol. 20, no. 8, pp. 1222–1232, 2004.

[14] X. Liu, S. Sivaganesan, K. Y. Yeung, J. Guo, R. E. Bumgarner, and M. Medvedovic, "Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset," *Bioinformatics*, vol. 22, no. 14, pp. 1737–1744., 2006.

[15] D. Dahl, "Model-based clustering for expression data via a Dirichlet process mixture model," in *Bayesian Inference for Gene Expression and Proteomics*, K.-A. Do, P. Müller, and M. Vannucci, Eds. Cambridge: Cambridge University Press, 2006.

[16] Z. S. Qin, "Clustering microarray gene expression data using weighted chinese restaurant process," *Bioinformatics*, vol. 22, no. 16, pp. 1988–1997, 2006.

[17] A. Dubey, S. Hwang, C. Rangel, C. Rasmussen, Z. Ghahramani, and D. L. Wild, "Clustering protein sequence and structure space with infinite Gaussian mixture models." in *Pacific Symposium on Biocomputing 2004*, R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein, Eds. Singapore:World Scientific Publishing, 2004, pp. 399–410.

[18] J. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[19] K. Yeung, D. Haynor, and W. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, pp. 309–318, 2001.

[20] D. J. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press, 2003.

[21] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[22] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo, "Model based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, pp. 977–987, 2001.

[23] D. Görür, "Nonparametric Bayesian discrete latent variable models for unsupervised learning," Ph.D. dissertation, Max Planck Institute for Biological Cybernetics, Tübingen, 2007.

[24] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G. Sherlock, "Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[25] M. Viswanathan, G. Muthukumar, Y. S. Cong, and J. Lenard, "Seripauperins of saccharomyces cerevisiae: a new multigene family encoding serine-poor relatives of serine-rich proteins," *Gene*, vol. 148, no. 1, pp. 149–153, 1994.

[26] N. Rachidi, M. J. Martinez, P. Barre, and B. Blondin, "Saccharomyces cerevisiae pau genes are induced by anaerobiosis," *Mol. Microbiol.*, vol. 35, no. 6, pp. 1421–1430, 2000.

[27] F. Klis, A. Boorsma, and P. D. Groot, "Cell wall construction in saccharomyces cerevisiae," *Yeast*, vol. 23, no. 185–202, 2006.

[28] U. Jung and D. Levin, "Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway," *Mol. Microbiol.*, vol. 34, pp. 1049–1057, 1999.

[29] W. McDowell and R. Schwarz, "Dissecting glycoprotein biosynthesis by use of specific inhibitors," *Biochimie*, vol. 70, pp. 1535–1549, 1998.

[30] A. Enyenihi and W. Saunders, "Large-scale functional genomic analysis of sporulation and meiosis in saccharomyces cerevisiae," *Genetics*, vol. 163, no. 1, pp. 47–54, 2003.

[31] M. Schuldiner *et al.*, "Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile," *Cell*, vol. 123, no. 3, pp. 507–519, 2005.

[32] A. Boorsma, H. de Nobel, B. ter Riet, B. Bargmann, S. Brul, K. Hellingwerf, and F. Klis, "Characterization of the transcriptional response to cell wall stress in saccharomyces cerevisiae," *Yeast*, vol. 21, pp. 413–427, 2004.

[33] M. Kaeberlein, M. McVey, and L. Guarente, "The sir2/3/4 complex and sir2 alone promote longevity in saccharomyces cerevisiae by two different mechanisms," *Genes Dev.*, vol. 13, pp. 2570–2580, 1999.

[34] G. Blander and L. Guarente, "The sir2 family of protein deacetylases," *Annu. Rev. Biochem.*, vol. 73, pp. 417–435, 2004.

[35] J. Masson and D. Ramotar, "The saccharomyces cerevisiae imp2 gene encodes a transcriptional activator that mediates protection against dna damage caused by bleomycin and other oxidants," *Mol. Cell Biol.*, vol. 16, no. 5, pp. 2091–2100, 1996.

[36] C. Donnini *et al.*, "Imp2, a nuclear gene controlling the mitochondrial dependence of galactose, maltose and raffinose utilization in saccharomyces cerevisiae," *Yeast*, vol. 8, no. 2, pp. 83–93, 1992.

[37] J. Mellor and A. Morillon, "Iswi complexes in saccharomyces cerevisiae," *Biochim. Biophys. Acta*, vol. 1677, no. 1–3, pp. 100–112, 2004.

[38] T. Kataoka *et al.*, "Genetic analysis of yeast ras1 and ras2 genes," *Cell*, vol. 37, no. 2, pp. 437–445, 1984.

[39] R. L. Smith and A. D. Johnson, "Turning genes off by ssn6-tup1: a conserved system of transcriptional repression in eukaryotes," *Trends in Biochem. Sci.*, vol. 25, no. 325–330, 2000.

[40] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments," *Proc. Natl. Acad. Sci. U S A.*, vol. 98, no. 16, pp. 8961–8965., 2001.

[41] K. Zhang and H. Zhao, "Assessing reliability of gene clusters from gene expression data," *Funct. Integr. Genomics*, vol. 1, pp. 156–173, 2000.

[42] N. A. Heard, C. C. Holmes, and D. A. Stephens, "A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 18–29, 2006.

[43] K. A. Heller and Z. Ghahramani, "Bayesian Hierarchical Clustering," *Twenty-second International Conference on Machine Learning*, 1995.

[44] J. W. Lau and P. J. Green, "Bayesian model based clustering procedures," *Journal of Computational and Graphical Statistics*, in press.

[45] G. Bidaut, K. Suhre, J.-M. Claverie, and M. Ochs, "Determination of strongly overlapping signaling activity from microarray data," *BMC Bioinformatics*, vol. 7, pp. 99–111, 2006.

[46] H. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. S. V *et al.*, "Mips: analysis and annotation of proteins from whole genomes," *Nucleic Acids Res.*, vol. 32, pp. D41–D44, 2004.

ADDITIONAL FILES

Additional file 1 — Tab-delimited file of Component GO terms for Transcript Response Clusters : Supp_Table_1.txt

Additional file 2 — Tab-delimited file of Functional GO terms for Transcript Response Clusters : Supp_Table_2.txt

Additional file 3 — Tab-delimited file of Process GO terms for Transcript Response Clusters : Supp_Table_3.txt

Additional file 4 — Tab-delimited file of Component GO terms for Experimental Clusters : Supp_Table_4.txt

Additional file 5 – Tab-delimited file of Functional GO terms for Experimental Clusters : Supp_Table_5.txt

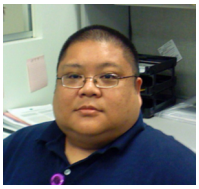
Additional file 6 – Tab-delimited file of Process GO terms for Experimental Clusters : Supp_Table_6.txt

Additional file 7 – Large version of Figure 3 : Fig03-largeNew.tif

Additional file 8 – Large version of Figure 5 : Fig05-largeNew.tif

Additional file 9 – Large version of Figure 6 : Fig06-largeNew.tif

Carl Edward Rasmussen is a lecturer in the Computational and Biological Learning Lab at the Department of Engineering, University of Cambridge and an adjunct research scientist at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. His main research interests are Bayesian inference and machine learning. He received his Masters in Engineering from the Technical University of Denmark and his PhD in Computer Science from the University of Toronto in 1996. Since then he has been a post doc at the Technical University of Denmark, a senior research fellow at the Gatsby Computational Neuroscience Unit at University College London from 2000-2002, and a junior research group leader at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany, 2002-2007.



Bernard J. de la Cruz earned his BA in biology from University of California Santa Cruz and PhD in biology from University of California San Diego. He was a postdoctoral research fellow at the Keck Graduate Institute studying the metabolism and genome of the yeast Pichia pastoris. Areas of interest include metabolic regulation of genes and analysis of microarray data. He is currently at the Beckman Research Institute at City of Hope.

Zoubin Ghahramani is Professor of Information Engineering at the University of Cambridge, UK, and is also Associate Research Professor of Machine Learning at Carnegie Mellon University, USA. He obtained BA and BSE degrees from University of Pennsylvania, and a PhD in 1995 from MIT working with Prof Mike Jordan. He did a postdoc in Computer Science at University of Toronto working with Prof Geoff Hinton, and was a faculty member at the Gatsby Unit, University College London from 1998 to 2005. His work has included research on human sensorimotor control, cognitive science, statistics, and machine learning. His current focus is on Bayesian approaches to statistical machine learning, non-parametric methods, graphical models, and approximate inference, He is also actively working on applications of machine learning to Bioinformatics and information retrieval. He has published over 100 peer reviewed papers, and serves on the editorial boards of several leading journals in the field, including JMLR, Annals of Statistics, JAIR, Machine Learning, Foundations and Trends in Machine Learning, and Bayesian Analysis. He is Associate Editor-in-Chief of IEEE Transactions on Pattern Analysis and Machine Intelligence, serves on the Board of the International Machine Learning Society, and was Program Chair of the 2007 International Conference on Machine Learning.

David L. Wild received a B.A. in physics from the University of York and a D.Phil. in molecular biophysics from the University of Oxford and has extensive experience in structural and computational molecular biology. He has worked at the European Molecular Biology Laboratory, the Salk Institute, and in industry with Allelix Biopharmaceuticals, Oxford Molecular and GlaxoWellcome. He is currently Professor of Bioinformatics at the University of Warwick Systems Biology Centre and an adjunct Research Professor at the Keck Graduate Institute of Applied Life Sciences. His research interests encompass bioinformatics, systems and structural biology.