RESEARCH

Consensus clustering for Bayesian mixture models

Stephen Coleman[1*], Paul D.W. Kirk[1,2] and Chris Wallace[1,2]

Correspondence:
tephen.coleman@mrc-
su.cam.ac.uk
MRC Biostatistics Unit,
Jniversity of Cambridge,
Cambridge, UK
ull list of author information is
vailable at the end of the article

Abstract

Background: Cluster analysis is an integral part of precision medicine and systems biology, used to define groups of patients or biomolecules. However, problems such as choosing the number of clusters and issues with high dimensional data arise consistently. An ensemble approach, such as consensus clustering, can overcome some of the difficulties associated with high dimensional data, frequently exploring more relevant clustering solutions than individual models. Another tool for cluster analysis, Bayesian mixture modelling, has alternative advantages, including the ability to infer the number of clusters present and extensibility. However, inference of these models is often performed using Markov-chain Monte Carlo (MCMC) methods which can suffer from problems such as poor exploration of the posterior distribution and long runtimes. This makes applying Bayesian mixture models and their extensions to 'omics data challenging. We apply consensus clustering to Bayesian mixture models to address these problems.

Results: Consensus clustering of Bayesian mixture models successfully finds the generating structure across our simulation studies and captures multiple modes in the likelihood surface. This approach also offers significant reductions in runtime compared to traditional Bayesian inference when a parallel environment is available. We propose a heuristic to decide upon ensemble size and then apply consensus clustering to Bayesian integrative clustering method, showing consensus clustering can be applied to any MCMC-based clustering method. We perform an integrative analysis of three 'omics datasets for budding yeast and find clusters of co-expressed genes with shared regulatory proteins. We validate these clusters using data external to the analysis. These clusters can help assign likely function to understudied genes, for example *GAS3* clusters with histones active in S-phase, suggesting a role in DNA replication.

Conclusions: Consensus clustering enables use of existing implementations of MCMC-based clustering methods on high-dimensional datasets, performing meaningful, reproducible inference where traditional approaches fail and faster inference where they do not. This enables researchers to use state-of-the-art Bayesian clustering methods on modern 'omics datasets, methods that can jointly model multiple datasets and can infer the number of clusters present.

Keywords: cluster analysis; ensemble methods; integrative clustering; Bayesian

Background

From defining a taxonomy of disease to creating molecular sets, grouping items can help us to understand and make decisions using complex biological data. For example, grouping patients based upon disease characteristics and personal omics data may allow the identification of more homogeneous subgroups, enabling stratified medicine approaches. In systems biology, defining and studying molecular sets can improve our understanding of biological systems as these sets are more interpretable than their constituent members [1], and study of their interactions and perturbations may have ramifications for diagnosis and drug targets [2, 3]. The act of identifying such groups is referred to as "cluster analysis", and has been traditional been done using tools such as K-means clustering [4, 5] or hierarchical clustering. However, these methods have various problems. For example, in K-means clustering, its sensitivity to initialisation means multiple runs are required, with that which minimises some metric such as the within-cluster sum of squared errors (**SSE**) used [6]. This problem arises as the algorithm has no guarantees on finding the global minimum of SSE. Another common problem is that traditional methods offer no measure of the uncertainty in the final clustering, a quantity of interest in many analyses. Returning to the stratified medicine example of clustering patients, there might be individuals with almost equal probability of being allocated between several clusters which might influence decisions made; however if only a point estimate is obtained this information is not available to the decision-maker. Ensemble methods offer a solution to this problem as well as reducing sensitivity to initialisation. These approaches have had great success in supervised learning, most famously in the form of Random Forest [7] and boosting [8]. In clustering, consensus clustering [9] is a popular ensemble method which has been implemented in R [10] and applied to a variety of methods [11, 12] and problems such as cancer subtyping [13, 14] and identifying subclones in single cell analysis [15]. Consensus

clustering uses W runs of some base model or learner (such as K-means clustering) and compiles the W proposed partitions into a *consensus matrix*, the $(i,j)^{th}$ entries of which contain the proportion of model runs for which the i^{th} and j^{th} individuals co-cluster (for this and other definitions see section 1 of Additional file 1). This proportion represents some measure of confidence in the co-clustering of any pair of items. Furthermore, ensembles can offer reductions in computational runtime. This is as the individual learners can be weaker (and thus use either less of the available data or stop before full convergence) and because the learners in most ensemble methods are independent of each other and thus enable use of a parallel environment for each of the quicker model runs [16].

Traditional clustering methods condition upon a fixed choice of K, the number of clusters. Choosing K is a difficult problem that haunts many analyses with researchers often relying on rules of thumb to decide upon a final model choice. For example, different choices of K are compared under some metric such as silhouette or SSE as a function of K. [9] proposed some methods for choosing K using the consensus matrix, but this means that any of the uncertainty about K is not represented in the final clustering and each model run uses the same, fixed, number of clusters. An alternative clustering approach, model-based clustering or mixture models, embeds the cluster analysis within a formal, statistical framework [17]. This means that models can be compared formally, and problems such as the choice of K can be addressed as a model selection problem with all the associated tools. Mixture models are also attractive, as they have great flexibility in the type of data they can be applied to due to different choice of densities. Bayesian mixture models can directly infer K, treating this as another random variable that is inferred from the data. This means that the final clustering is not conditional upon a user chosen value, but K is jointly modelled along with the clustering. Such inference can be performed through use of a Dirichlet Process [18], a mixture of finite mixture mod-

els [19, 20] or an over-fitted mixture model [21]. These models and their extensions have a history of successful application to a diverse range of biological problems such as finding clusters of gene expression profiles [22], cell types in flow cytometry [23, 24] or scRNAseq experiments [25], and estimating protein localisation [26]. Bayesian mixture models can be extended to jointly model the clustering across multiple datasets [27, 28] (section 2 of Additional file 1).

However, performing inference of Bayesian mixture models is a difficult task. Variational inference [29] (**VI**) may be used to perform approximate inference [30], but, while VI is powerful, it can struggle with multi-modality, underestimates the variance in the posterior distribution [31] and it has been shown to have a very computationally heavy initialisation cost to have good results [32]. Implementation is difficult, requiring either complex derivations (see the Appendix Supplementary Methods of [33] for an example) or black-box, approximate solutions [34]. Markov chain Monte Carlo (MCMC) methods are the most common tool for performing Bayesian inference. In Bayesian clustering methods, they are used to construct a chain of clusterings and an assessment of the convergence of this chain is made to determine if its behaviour aligns with the expected asymptotic theory. However, in practice individual chains often fail to explore the full support of the posterior distribution despite the ergodicity of MCMC methods (see, e.g., the Supplementary Materials of [35]) and can experience long runtimes. Some MCMC methods make efforts to overcome the problem of exploration, often at the cost of increased computational cost per iteration. See, e.g., [36, 37, 38] for examples of problems and attempted solutions for MCMC methods.

We propose that applying consensus clustering to Bayesian mixture models can overcome some of the issues endemic in high dimensional Bayesian clustering. [9] suggest this application as part of their original paper, but no investigation has been attempted to our knowledge. This ensemble approach sidesteps the problems

of convergence associated MCMC methods and offers computational gains through using shorter chains run in parallel. Furthermore, this approach could be directly used on any existing MCMC based implementation of Bayesian mixture models or their extensions and would avoid the re-implementation process that changing to newer MCMC methods or VI would entail.

We propose a heuristic for deciding upon the ensemble width (the number of learners used, W) and the ensemble depth (the number of iterations run within each chain, D), inspired by the use of scree plots in Principal Component Analysis [39] (**PCA**).

We show via simulation that ensembles consisting of short chains can be sufficient to successfully recover generating structure. We also show that consensus clustering explores as many or more modes of the likelihood surface than either standard Bayesian inference or `Mclust`, a maximum likelihood method, all while offering improvements in runtime to traditional Bayesian inference.

We use consensus clustering of Multiple Dataset Interation (**MDI**), a Bayesian integrative clustering method, to analyse multiple 'omics datasets relating to the cell cycle of *Saccharomyces cerevisiae* to show that consensus clustering can applied to more complex MCMC-based clustering methods and real datasets.

Methods

Consensus clustering for Bayesian mixture models

We apply consensus clustering to MCMC based Bayesian clustering models using the method described in algorithm 1. Our application of consensus clustering has two main parameters at the ensemble level, the chain depth, D, and ensemble width, W. We infer a point clustering from the consensus matrix using the `maxpear` function [47] from the R package `mcclust` [48] to (section 3 of Additional file 1 for details).

Data: $X = (x_1, \ldots, x_N)$
Input:
The number of chains to run, W
The number of iterations within each chain, D
A clustering method that uses MCMC methods to generate samples of clusterings of the data *Cluster*(X, d)
Output:
A predicted clustering, $\hat{Y}$
The consensus matrix $\mathbf{M}$
begin
  /* initialise an empty consensus matrix */
  $\mathbf{M} \leftarrow \mathbf{0}_{N \times N}$;
  for $w = 1$ **to** W **do**
    /* set the random seed controlling initialisation and MCMC moves */
    $set.seed(w)$;
    /* initialise a random partition on X drawn from the prior distribution */
    $Y_{(0,w)} \leftarrow Initialise(X)$;
    for $d = 1$ **to** D **do**
      /* generate a markov chain for the membership vector */
      $Y_{(d,w)} \leftarrow Cluster(X, d)$;
    end
    /* create a coclustering matrix from the D^{th} sample */
    $\mathbf{B}^{(w)} \leftarrow Y_{(D,w)}$;
    $\mathbf{M} \leftarrow \mathbf{M} + \mathbf{B}^{(w)}$;
  end
  $\mathbf{M} \leftarrow \frac{1}{W}\mathbf{M}$;
  $\hat{Y} \leftarrow$ partition X based upon $\mathbf{M}$;
end

Algorithm 1: Consensus clustering for Bayesian mixture models.

Determining the ensemble depth and width

As our ensemble sidesteps the problem of convergence within each chain, we need an alternative stopping rule for growing the ensemble in chain depth, D, and number of chains, W. We propose a heuristic based upon the consensus matrix to decide if a given value of D and W are sufficient. We suspect that increasing W and D might continuously improve the performance of the ensemble, but we observe in our simulations that these improvements will become smaller and smaller for greater values, approaching some asymptote for each of W and D. We notice that this behaviour is analogous to PCA in that where for consensus clustering some improvement might always be expected for increasing chain depth or ensemble width, more variance will always be captured by increasing the number of components used in PCA. However, increasing this number beyond some threshold has diminishing returns, diagnosed in PCA by a scree plot. Following from this, we recommend, for some set of ensemble parameters, $D' = \{d_1, \ldots, d_I\}$ and $W' = \{w_1, \ldots, w_J\}$, find the mean absolute difference of the consensus matrix for the d_i^{th} iteration from w_j chains to that for the $d_{(i-1)}^{th}$ iteration from w_j chains and plot these values as a function of chain depth, and the analogue for sequential consensus matrices for increasing ensemble width and constant depth.

If this heuristic is used, we believe that the consensus matrix and the resulting inference should be stable (see, e.g., [49, 50]), providing a robust estimate of the clustering. In contrast, if there is still strong variation in the consensus matrix for varying chain length or number, then we believe that the inferred clustering is influenced significantly by the random initialisation and that the inferred partition is unlikely to be stable for similar datasets or reproducible for a random choice of seeds.

Simulation study

We use a finite mixture with independent features as the data generating model within the simulation study. We include "irrelevant features" [51] that have global parameters rather than cluster specific parameters and use the generating model:

$$
\begin{aligned}
&p(X, c, \theta, \pi | K) \\
&\quad = p(K)p(\pi|K)p(\theta|K) \prod_{i=1}^{N} p(c_i|\pi, K) \prod_{p=1}^{P} p(x_{ip}|c_i, \theta_{c_i p})^{\phi_p} p(x_{ip}|\theta_p)^{(1-\phi_p)}
\end{aligned}
$$

for data $X = (x_1, \ldots, x_N)$, cluster label or allocation variable $c = (c_1, \ldots, c_N)$, cluster weight $\pi = (\pi_1, \ldots, \pi_K)$, K clusters and the relevance variable, $\phi \in \{0, 1\}$ with $\phi_p = 1$ indicating that the p^{th} feature is relevant to the clustering. We used a *Gaussian* density, so $\theta_{kp} = (\mu_{kp}, \sigma^2_{kp})$. We defined three scenarios and simulated 100 datasets in each (Figure 1 and Table 1) Additional details of the simulation process and additional scenarios are included in section 4.1 of Additional file 1s.

Table 1 Parameters defining the simulation scenarios as used in generating data and labels. $\Delta\mu$ is the distance between neighbouring cluster means within a single feature. The number of relevant features (P_s) is $\sum_p \phi_p$, and $P_n = P - P_s$.

Scenario	N	P_s	P_n	K	$\Delta\mu$	σ^2	π
2D	100	2	0	5	3.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Small N, large P	50	500	0	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	100	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$

In each of these scenarios we apply a variety of methods (listed below) and compare the inferred point clusterings to the generating labels using the Adjusted Rand Index [52] (**ARI**).

- `Mclust`, a maximum likelihood implementation of finite mixture models (for a range of modelled clusters, K),
- 10 chains of 1 million iterations, thinning to every thousandth sample for the overfitted Bayesian mixture model, and

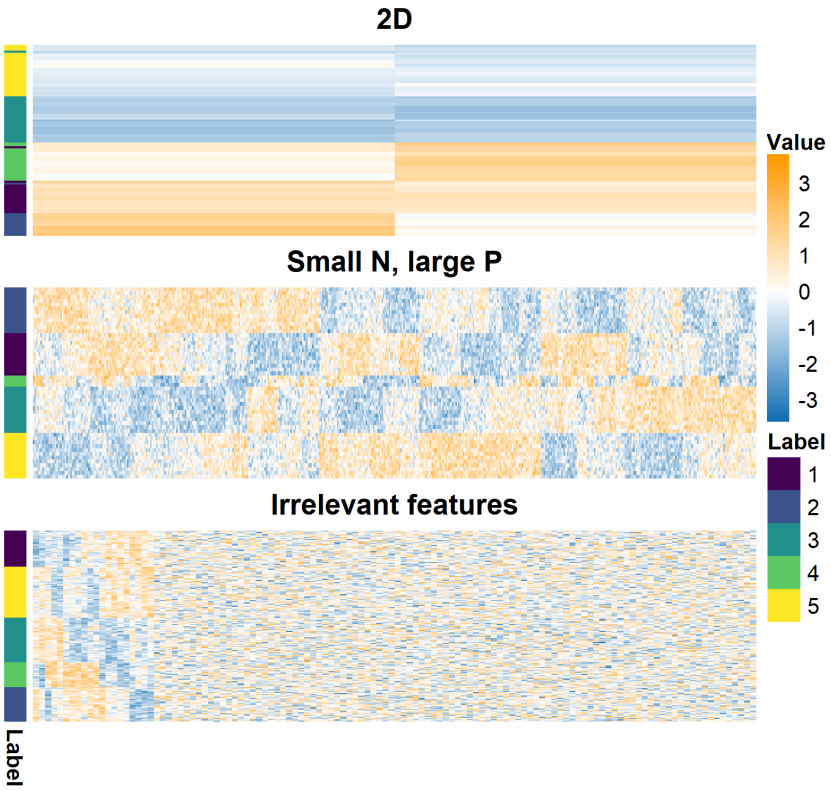

Figure 1 Example of generated datasets. Each row is an item being clustered and each column a feature of generated data. The rows and relevant features of each dataset are ordered by hierarchical clustering. For the irrelevant features dataset, only the 20 relevant features are used to order rows. The 2D dataset should enable proper mixing of chains in the MCMC. The small N, large P case has clear structure (observable by eye). This is intended to highlight the problems of poor mixing due to high dimensions even when the generating labels are quite identifiable. In the irrelevant features case, the structure is clear in the relevant features (on the left-hand side of this heatmap). This setting is intended to test how sensitive each approach is to noise.

- A variety of consensus clustering ensembles defined by inputs of W chains and D iterations within each chain (see algorithm 1) with $W \in \{1, 10, 30, 50, 100\}$ and $D \in \{1, 10, 100, 1000, 10000\}$.

The ARI is a measure of similarity between two partitions,c_1, c_2, corrected for chance, with 0 indicating c_1 is no more similar to c_2 than a random partition would be expected to be and a value of 1 showing that c_1 and c_2 perfectly align. Details of the methods in the simulation study can be found in sections 4.2, 4.3 and 4.4 of Additional file 1.

Mclust

`Mclust` [53] is a function from the R package `mclust`. It estimates Gaussian mixture models for K clusters based upon the maximum likelihood estimator of the parameters. it initialises upon a hierarchical clustering of the data cut to K clusters. A range of choices of K and different covariance structures are compared and the "best" selected using the Bayesian information criterion, [54] (details in section 4.2 of Additional file 1).

Bayesian inference

To assess within-chain convergence of our Bayesian inference we use the Geweke Z-score statistic [55]. Of the chains that appear to behave properly we then asses across-chain convergence using $\hat{R}$ [56] and the recent extension provided by [57]. If a chain has reached its stationary distribution the Geweke Z-score statistic is expected to be normally distributed. Normality is tested for using a Shapiro-Wilks test [58]. If a chain fails this test (i.e., the associated p-value is less than 0.05), we assume that it has not achieved stationarity and it is excluded from the remainder of the analysis. The samples from the remaining chains are then pooled and a posterior similarity matrix (**PSM**) constructed. We use the `maxpear` function to infer a point clustering. For more details see section 4.3 of Additional file 1.

Analysis of the cell cycle in budding yeast

Datasets

The cell cycle is crucial to biological growth, repair, reproduction, and development [41, 42, 43] and is highly conserved among eukaryotes [43]. This means that understanding of the cell cycle of *S. cerevisiae* can provide insight into a variety of cell cycle perturbations including those that occur in human cancer [44, 42] and ageing [45]. We aim to create clusters of genes that are co-expressed in the cell cycle, have common regulatory proteins and share a biological function. To achieve this, we use

three datasets that were generated using different 'omics technologies and target different aspects of the molecular biology underpinning the cell cycle process.

- Microarray profiles of RNA expression from [40], comprising measurements of cell-cycle-regulated gene expression at 5-minute intervals for 200 minutes (up to three cell division cycles) and is referred to as the **time course** dataset. The cells are synchronised at the START checkpoint in late G1-phase using alpha factor arrest [40]. We include only the genes identified by [40] as having periodic expression profiles.
- Chromatin immunoprecipitation followed by microarray hybridization (**ChIP-chip**) data from [59]. This dataset discretizes p-values from tests of association between 117 DNA-binding transcriptional regulators and a set of yeast genes. Based upon a significance threshold these p-values are represented as either a 0 (no interaction) or a 1 (an interaction).
- Protein-protein interaction (**PPI**) data from BioGrid [60]. This database consists of of physical and genetic interactions between gene and gene products, with interactions either observed in high throughput experiments or computationally inferred. The dataset we used contained 603 proteins as columns. An entry of 1 in the $(i, j)^{th}$ cell indicates that the i^{th} gene has a protein product that is believed to interact with the j^{th} protein.

The datasets were reduced to the 551 genes with no missing data in the PPI and ChIP-chip data, as in [27].

Multiple dataset integration

We applied consensus clustering to MDI for our integrative analysis. Details of MDI are in section 2.2 of Additional file 1, but in short MDI jointly models the clustering in each dataset, inferring individual clusterings for each dataset. These partitions are informed by similar structure in the other datasets, with MDI learning this similarity as it models the partitions. The model does not assume global structure.

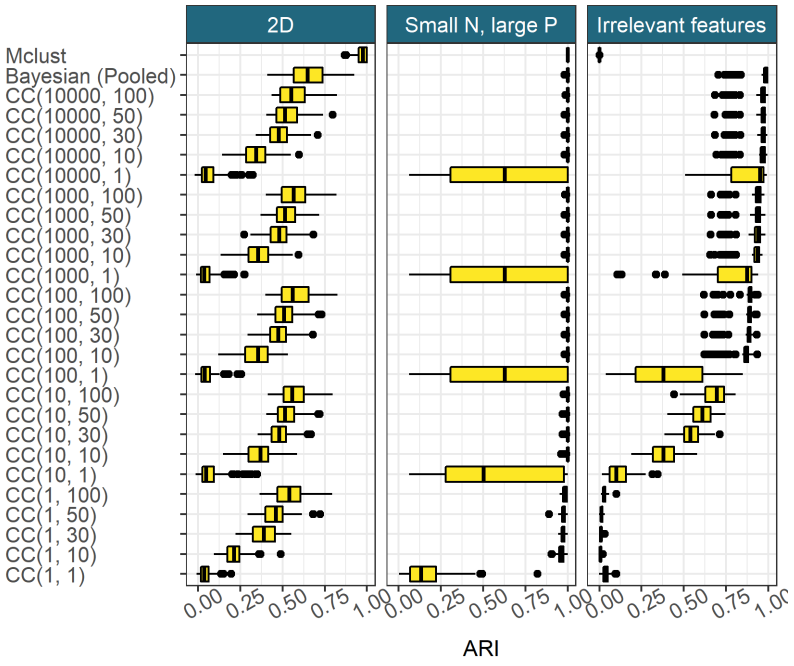

Figure 2 Model performance in the 100 simulated datasets for each scenario, defined as the ARI between the generating labels and the inferred clustering. $CC(d, w)$ denotes consensus clustering using the clustering from the d^{th} iteration from w different chains.

This means that the similarity between datasets is not strongly assumed in our model; individual clusters or genes that align across datasets are based solely upon the evidence present in the data and not due to strong modelling assumptions. Thus, datasets that share less common information can be included without fearing that this will warp the final clusterings in some way.

The datasets were modelled using a mixture of Gaussian processes in the time course dataset and Multinomial distributions in the ChIP-chip and PPI datasets.

Results

Simulated data

We use the ARI between the generating labels and the inferred clustering of each method to be our metric of predictive performance. In Figure 2, we see `Mclust` performs very well in the 2D and Small N, large P scenarios, correctly identifying

the true structure However, the irrelevant features scenario sees a collapse in performance, `Mclust` is blinded by the irrelevant features and identifies a clustering of $K = 1$.

The pooled samples from multiple long chains performs very well across all scenarios and appears to act as an upper bound on the more practical implementations of consensus clustering.

Consensus clustering does uncover some of the generating structure in the data, even using a small number of short chains. With sufficiently large ensembles and chain depth, consensus clustering is close to the pooled Bayesian samples in predictive performance. It appears that for a constant chain depth increasing the ensemble width used follows a pattern of diminishing returns. There are strong initial gains for a greater ensemble width, but the improvement decreases for each successive chain. A similar pattern emerges in increasing chain length for a constant number of chains (Figure 2).

We see very little difference between the similarity matrix from the pooled samples and the consensus clustering (Figure 3). Similar clusters emerge, and we see comparable confidence in the pairwise clusterings. For the PSMs from the individual chains, all entries are 0 or 1. This means only a single clustering is sampled within each chain, implying very little uncertainty in the partition. However, three different modes emerge across the chains showing that the chains are failing to explore the full support of the posterior distribution of the clustering and are each unrepresentative of the uncertainty in the final clustering. This shows that consensus clustering is exploring more possible clusterings than any individual chain and, as it explores a similar space to the pooled samples which might be considered more representative of the posterior distribution than any one chain, it suggests it better describes the true uncertainty present than any single chain. It also shows that pooling chains offers robustness to multi-modality (as expected for an ensemble) and the ARI for

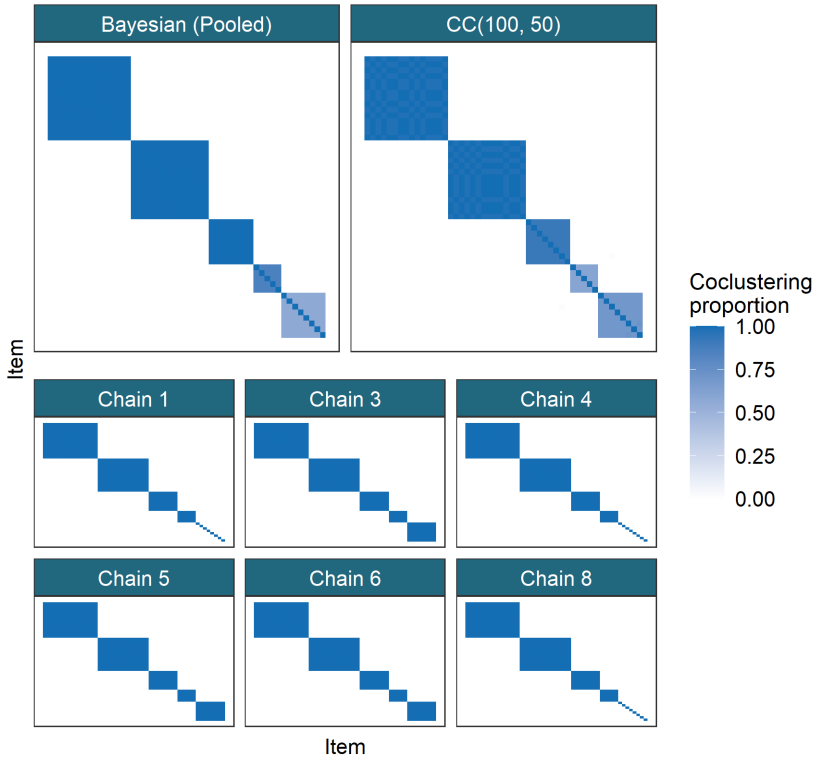

Figure 3 Comparison of similarity matrices from a dataset for the Small N, large P scenario. In each matrix, the $(i, j)^{th}$ entry is the proportion of clusterings for which the i^{th} and j^{th} items co-clustered for the method in question. In the first row the PSM of the pooled Bayesian samples is compared to the CM for CC(100, 50), with a common ordering of rows and columns in both heatmaps. In the following rows, 6 of the long chains that passed the tests of convergence are shown.

the pooled samples is an upper bound on the performance for the individual long chains.

Figure 4 shows that chain length is directly proportional to the time taken for the chain to run. This means that using an ensemble of shorter chains, as in consensus clustering, can offer large reductions in the time cost of analysis when a parallel environment is available compared to standard Bayesian inference. Even on a laptop of 8 cores running an ensemble of 1,000 chains of length 1,000 will require approximately half as much time as running 10 chains of length 100,000 due to par-

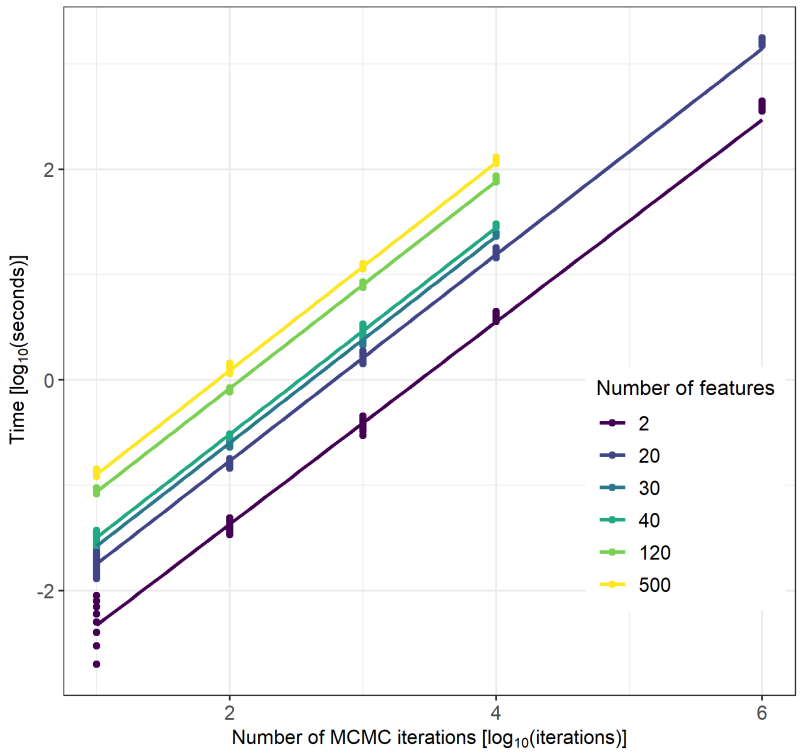

Figure 4 The time taken for different numbers of iterations of MCMC moves in $\log_{10}(seconds)$. The relationship between chain length, D, and the time taken is linear (the slope is approximately 1 on the $\log_{10}$ scale), with a change of intercept for different dimensions. The runtime of each Markov chain was recorded using the terminal command `time`, measured in milliseconds.

allelisation, and the potential benefits are far greater when using a large computing cluster.

Additional results for these and other simulations are in section 4.4 of Additional file 1.

Multi-omics analysis of the cell cycle in budding yeast

We use the stopping rule proposed in to determine our ensemble depth and width. In Figure 5, we see that the change in the consensus matrices from increasing the ensemble depth and width is diminishing in keeping with results in the simulations. We see no strong improvement after $D = 6,000$ and increasing the number of learners from 500 to 1,000 has small effect. We therefore use the largest ensemble

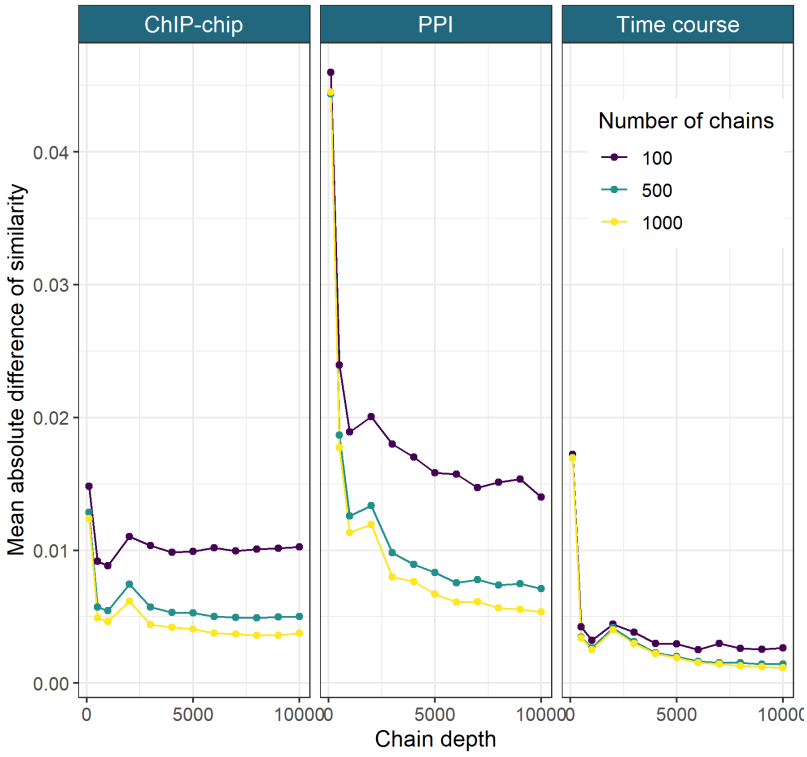

Figure 5 The mean absolute difference between the sequential Consensus matrices. For a set of chain lengths, $D' = \{d_1, \ldots, d_I\}$ and number of chains, $W' = \{w_1, \ldots, w_J\}$, we take the mean of the absolute difference between the consensus matrix for (d_i, w_j) and (d_{i-1}, w_j) (here $D' = \{101, 501, 1001, 2001, \ldots, 10001\}$ and $W' = \{100, 500, 1000\}$).

available, a depth $D = 10001$ and width $W = 1000$, believing this ensemble is stable (additional evidence in section 5.1 of Additional file 1).

We focus upon the genes that tend to have the same cluster label across multiple datasets. More formally, we analyse the clustering structure among genes for which $\hat{P}(c_{nl} = c_{nm}) > 0.5$, where c_{nl} denotes the cluster label of gene n in dataset l. In our analysis it is the signal shared across the time course and ChIP-chip datasets that is strongest, with 261 genes (nearly half of the genes present) in this pairing tending to have a common label, whereas only 56 genes have a common label across all three datasets. Thus, we focus upon this pairing of datasets in the results of the analysis performed using all three datasets. We show the gene expression and regulatory proteins of these genes separated by their cluster in Figure 6. In Figure

6, the clusters in the time series data have tight, unique signatures (having different periods, amplitudes, or both) and in the ChIP-chip data clusters are defined by a small number of well-studied transcription factors (**TFs**) [61] (see Table 2 of Additional file 1).

As an example, we briefly analyse clusters 9 and 16 in greater depth. Cluster 9 has strong association with MBP1 and some interactions with SWI6, as can be seen in Figure 6. The Mbp1-Swi6p complex, MBF, is associated with DNA replication [62]. The first time point, 0 minutes, in the time course data is at the START checkpoint, or the G1/S transition. The members of cluster 9 begin highly expressed at this point before quickly dropping in expression (in the first of the 3 cell cycles). This suggests that many transcripts are produced immediately in advance of S-phase, and thus are required for the first stages of DNA synthesis. These genes' descriptions ([found using `org.Sc.sgd.db`, [63], and shown in Table 3 of Additional file 1) support this hypothesis, as many of the members are associated with DNA replication, repair and/or recombination. Additionally, *TOF1*, *MRC1* and *RAD53*, members of the replication checkpoint [64, 65] emerge in the cluster as do members of the cohesin complex. Cohesin is associated with sister chromatid cohesion which is established during the S-phase of the cell cycle [66] and also contributes to transcription regulation, DNA repair, chromosome condensation, homolog pairing [67], fitting the theme of cluster 9.

Cluster 16 appears to be a cluster of S-phase genes, consisting of *GAS3*, *NRM1* and *PDS1* and the genes encoding the histones H1, H2A, H2B, H3 and H4. Histones are the chief protein components of chromatin [68] and are important contributors to gene regulation [69]. They are known to peak in expression in S-phase [40], which matches the first peak of this cluster early in the time series. Of the other members, *NRM1* is a transcriptional co-repressor of MBF-regulated gene expression acting at the transition from G1 to S-phase [70, 71]. Pds1p binds to and inhibits the Esp1

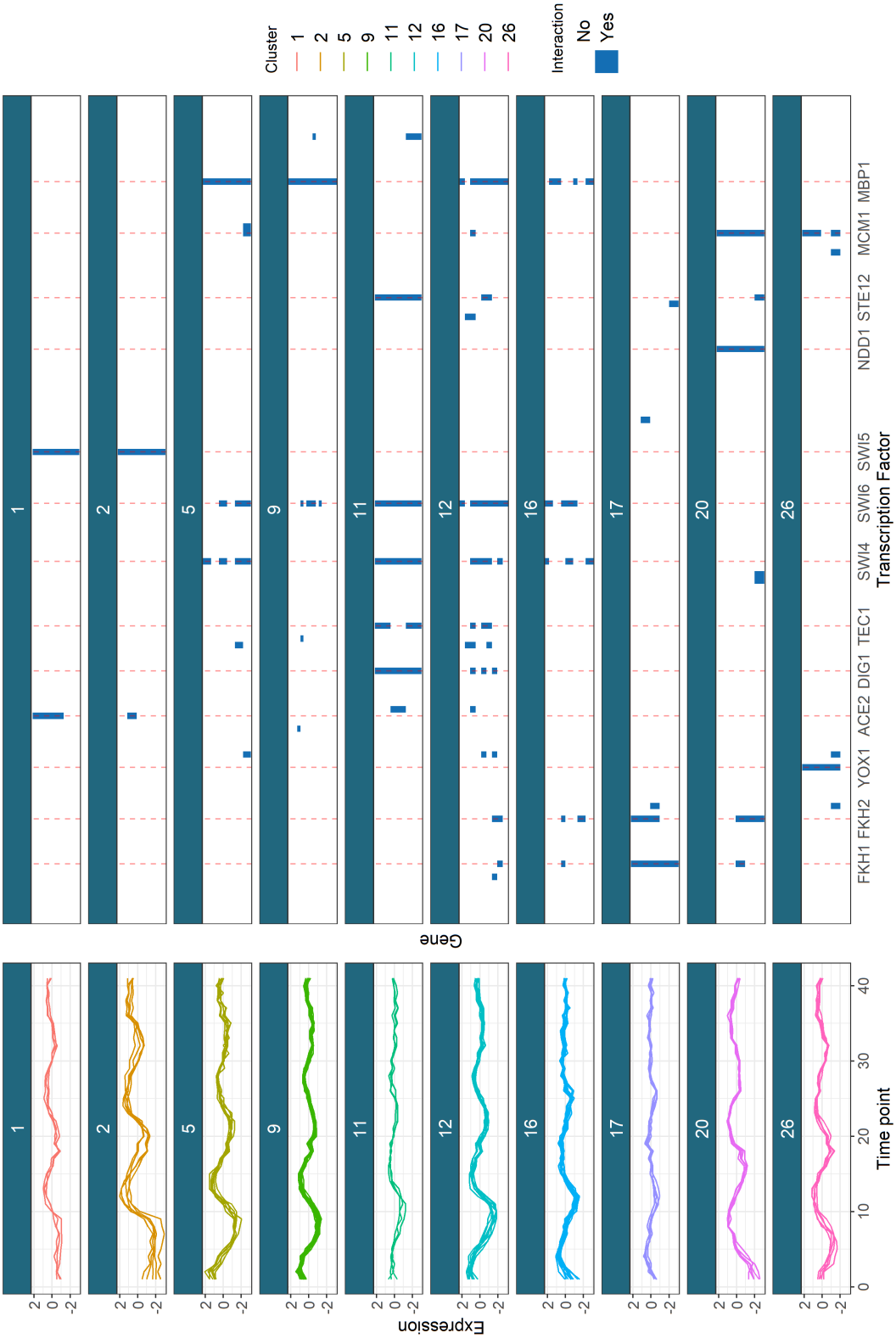

Figure 6 The gene clusters which tend to have a common label across the time course and ChIP-chip datasets, shown in these datasets. We include only the clusters with more than one member and more than half the members having some interactions in the ChIP-chip data. Red lines for the most common transcription factors are included.

class of sister separating proteins, preventing sister chromatids separation before M-phase [72, 66]. *GAS3*, is not well studied. It interacts with *SMT3* which regulates chromatid cohesion, chromosome segregation and DNA replication (among other things). Chromatid cohesion ensures the faithful segregation of chromosomes in mitosis and in both meiotic divisions [73] and is instantiated in S-phase [66]. These results, along with the very similar expression profile to the histone genes in the time course data, suggest that *GAS3* may be more directly involved in DNA replication or chromatid cohesion than is currently believed.

We attempt to perform a similar analysis using traditional Bayesian inference of MDI, but after 36 hours of runtime there is no consistency or convergence across the ten chains. We use the Geweke statistic and $\hat{R}$ to reduce to the five best behaved chains (none of which appear to be converged, Additional file 1, section 5.2). If we then compare the distribution of sampled values for the ϕ parameters for these long chains, the final ensemble used ($D = 10001$, $W = 1000$) and the pooled samples from the 5 long chains, then we see that the distribution of the pooled samples from the long chains (which might be believed to sampling different parts of the posterior distribution) is closer in appearance to the distributions sampled by the consensus clustering than to any single chain (figure 7). Further disagreement between chains is shown in the Gene Ontology term over-representation analysis in section 5.3 of Additional file 1.

Discussion

Our proposed method has demonstrated good performance on simulation studies, uncovering the generating structure and approximating Bayesian inference when the Markov chain is exploring the full support of the posterior distribution. However, we have shown that if a finite Markov chain fails to describe the full posterior and is itself only approximating Bayesian inference, our method has better ability to represent several modes in the data than individual chains and thus offers a

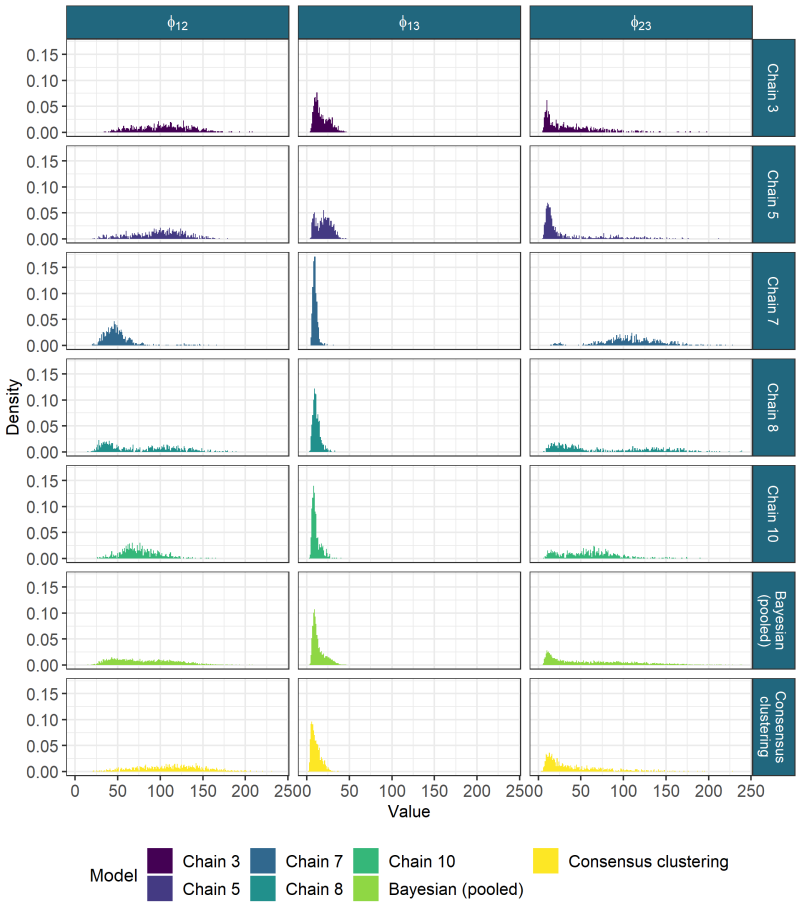

Figure 7 The sampled values for the ϕ parameters from the long chains, their pooled samples and the consensus using 1000 chains of depth 10,001. The long chains display a variety of behaviours. Across chains there is no clear consensus on the nature of the posterior distribution. The samples from any single chain are not particularly close to the behaviour of the pooled samples across all three parameters. It is the consensus clustering that most approaches this pooled behaviour.

more consistent and reproducible analysis. Furthermore, consensus clustering is significantly faster in a parallel environment than inference using individual chains, while retaining the ability to robustly infer the number of clusters present.

We proposed a method of assessing ensemble stability and deciding upon ensemble size which we used when performing an integrative analysis of yeast cell cycle data using MDI, an extension of Bayesian mixture models that jointly models multiple datasets. We uncovered many genes with shared signal across several datasets and explored the meaning of some of the inferred clusters, using data external to

the analysis. We found sensible results as well as signal for possibly novel biology. In contrast, the traditional approach to Bayesian inference failed here. The lack of a consistent distribution across the chains made proceeding with the Bayesian analysis difficult as choosing the result of any single chain over the others would be arbitrary and thus prone to irreproducibility. The alternative of pooling the samples, which might be considered a reasonable compromise, appears to offer a very similar solution to consensus clustering, but with longer runtime and additional steps to reduce the chains to the "best-behaved" chains. We believe that the similarity between the sampled distribution of the parameters from the pooled long chains and the consensus clustering of short chains, figure 7, suggests that sufficiently deep chains within the ensemble can be used even to perform inference of continuous variables and not only the latent clustering of the data.

Consensus clustering loses the theoretical framework of true Bayesian inference. We attempt to mitigate this with our assessment of stability in the ensemble, but this diagnosis is heuristic and subjective, and while there is empirical evidence for its success, it lacks the formal results for the tests of model convergence for Bayesian inference. Nonetheless, the results of our simulations and the multi-omics analysis show that consensus clustering can be successfully used in a broad context, being applicable to any MCMC based clustering method. It offers computational gains and improves the exploration of the clustering space, overcoming the problem of becoming trapped in specific, local extrema of the likelihood surface that emerges in high-dimensional data. This enables the application of these methods in modern 'omics datasets and, attractively, consensus clustering can be applied to existing implementations, unlike improvements to the underlying MCMC methods or alternative methods for Bayesian inference such as VI which would require re-writing software.

Funding

This work was funded by the MRC (MC UU 00002/4, MC UU 00002/13) and supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This research was funded in whole, or in part, by the Wellcome Trust [WT107881]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials

The code and datasets supporting the conclusions of this article are available in the github repository, `https://github.com/stcolema/ConsensusClusteringForBayesianMixtureModels`.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SC designed the simulation study with contributions from PK and CW, performed the analyses and wrote the manuscript. PK and CW provided an equal contribution of joint supervision, directing the research and provided suggestions such as the stopping rule. All contributed to interpreting the results of the analyses. All authors revised and approved the final manuscript.

Author details

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. [2]Cambridge Institute of Therapeutic Immunology & Infectious Disease, University of Cambridge, Cambridge, UK.

References

1. Hejblum, B.P., Skinner, J., Thiébaut, R.: Time-course gene set analysis for longitudinal gene expression data. PLoS computational biology **11**(6), 1004310 (2015)
2. Bai, J.P., Alekseyenko, A.V., Statnikov, A., Wang, I.-M., Wong, P.H.: Strategic applications of gene expression: from drug discovery/development to bedside. The AAPS journal **15**(2), 427–437 (2013)
3. Emmert-Streib, F., Dehmer, M., Haibe-Kains, B.: Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Frontiers in cell and developmental biology **2**, 38 (2014)
4. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)
5. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics **21**, 768–769 (1965)
6. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Technical report, Stanford (2006)
7. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
8. Friedman, J.H.: Stochastic gradient boosting. Computational statistics & data analysis **38**(4), 367–378 (2002)
9. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning **52**(1-2), 91–118 (2003)

10. Wilkerson, D., M., Hayes, Neil, D.: ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics **26**(12), 1572–1573 (2010)
11. John, C.R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., Barnes, M.: M3C: Monte Carlo reference-based consensus clustering. Scientific reports **10**(1), 1–14 (2020)
12. Gu, Z., Schlesner, M., Hübschmann, D.: cola: an R/Bioconductor package for consensus partitioning through a general framework. Nucleic Acids Research (2020). doi:10.1093/nar/gkaa1146. gkaa1146. https://academic.oup.com/nar/advance-article-pdf/doi/10.1093/nar/gkaa1146/34695832/gkaa1146.pdf
13. Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., Pietenpol, J.A., *et al.*: Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of clinical investigation **121**(7), 2750–2767 (2011)
14. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., *et al.*: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer cell **17**(1), 98–110 (2010)
15. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., *et al.*: SC3: consensus clustering of single-cell RNA-seq data. Nature methods **14**(5), 483–486 (2017)
16. Ghaemi, R., Sulaiman, M.N., Ibrahim, H., Mustapha, N., *et al.*: A survey: clustering ensembles techniques. World Academy of Science, Engineering and Technology **50**, 636–645 (2009)
17. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association **97**(458), 611–631 (2002)
18. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The annals of statistics, 209–230 (1973)
19. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society: series B **59**(4), 731–792 (1997)
20. Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association **113**(521), 340–356 (2018)
21. Rousseau, J., Mengersen, K.: Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(5), 689–710 (2011)
22. Medvedovic, M., Sivaganesan, S.: Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics **18**(9), 1194–1206 (2002)
23. Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., Kepler, T.B.: Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A: The Journal of the International Society for Analytical Cytology **73**(8), 693–701 (2008)
24. Hejblum, B.P., Alkhassim, C., Gottardo, R., Caron, F., Thiébaut, R., *et al.*: Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. The Annals of Applied Statistics **13**(1), 638–660 (2019)
25. Prabhakaran, S., Azizi, E., Carr, A., Pe'er, D.: Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In: International Conference on Machine Learning, pp. 1070–1079 (2016)
26. Crook, O.M., Mulvey, C.M., Kirk, P.D., Lilley, K.S., Gatto, L.: A Bayesian mixture modelling approach for spatial proteomics. PLoS computational biology **14**(11), 1006516 (2018)
27. Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L.: Bayesian correlated clustering to integrate multiple datasets. Bioinformatics **28**(24), 3290–3297 (2012)
28. Gabasova, E., Reid, J., Wernisch, L.: Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS computational biology **13**(10), 1005781 (2017)

29. Blei, D.M., Jordan, M.I., *et al.*: Variational inference for Dirichlet process mixtures. Bayesian analysis **1**(1), 121–143 (2006)
30. Martin, G.M., Frazier, D.T., Robert, C.P.: Computing Bayes: Bayesian Computation from 1763 to the 21st Century. arXiv preprint arXiv:2004.06425 (2020)
31. Turner, R.E., Sahani, M.: Two problems with variational expectation maximisation for time-series models. In: Barber, D., Cemgil, A.T., Chiappa, S. (eds.) Bayesian Time Series Models, 1st edn. Cambridge University Press, ??? (2011). Chap. 5
32. Wang, L., Dunson, D.B.: Fast Bayesian inference in Dirichlet process mixture models. Journal of Computational and Graphical Statistics **20**(1), 196–216 (2011)
33. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O.: Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology **14**(6), 8124 (2018)
34. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M.: Automatic differentiation variational inference. The Journal of Machine Learning Research **18**(1), 430–474 (2017)
35. Strauss, M.E., Kirk, P.D., Reid, J.E., Wernisch, L.: Gpseudoclust: deconvolution of shared pseudo-profiles at single-cell resolution. Bioinformatics **36**(5), 1484–1491 (2020)
36. Robert, C.P., Elvira, V., Tawn, N., Wu, C.: Accelerating mcmc algorithms. Wiley Interdisciplinary Reviews: Computational Statistics **10**(5), 1435 (2018)
37. Yao, Y., Vehtari, A., Gelman, A.: Stacking for non-mixing Bayesian computations: the curse and blessing of multimodal posteriors. arXiv preprint arXiv:2006.12335 (2020)
38. Bouchard-Côté, A., Doucet, A., Roth, A.: Particle Gibbs split-merge sampling for Bayesian inference in mixture models. Journal of Machine Learning Research **18**(28) (2017)
39. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
40. Granovskaia, M.V., Jensen, L.J., Ritchie, M.E., Toedling, J., Ning, Y., Bork, P., Huber, W., Steinmetz, L.M.: High-resolution transcription atlas of the mitotic cell cycle in budding yeast. Genome biology **11**(3), 1–11 (2010)
41. Tyson, J.J., Chen, K.C., Novák, B.: Cell cycle, budding yeast. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (eds.) Encyclopedia of Systems Biology, pp. 337–341. Springer, New York, NY (2013)
42. Chen, K.C., Calzone, L., Csikasz-Nagy, A., Cross, F.R., Novak, B., Tyson, J.J.: Integrative analysis of cell cycle control in budding yeast. Molecular biology of the cell **15**(8), 3841–3862 (2004)
43. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: The cell cycle and programmed cell death. Molecular biology of the cell **4**, 983–1027 (2002)
44. Ingalls, B., Duncker, B., Kim, D., McConkey, B.: Systems level modeling of the cell cycle using budding yeast. Cancer informatics **3**, 117693510700300020 (2007)
45. Jiménez, J., Bru, S., Ribeiro, M., Clotet, J.: Live fast, die soon: cell cycle progression and lifespan in yeast cells. Microbial Cell **2**(3), 62 (2015)
46. Juanes, M.A.: Methods of synchronization of yeast cells for the analysis of cell cycle progression. In: The Mitotic Exit Network, pp. 19–34. Springer, ??? (2017)
47. Fritsch, A., Ickstadt, K., *et al.*: Improved criteria for clustering based on the posterior similarity matrix. Bayesian analysis **4**(2), 367–391 (2009)
48. Fritsch, A.: Mcclust: Process an MCMC Sample of Clusterings. (2012). R package version 1.0. https://CRAN.R-project.org/package=mcclust

49. Von Luxburg, U., Ben-David, S.: Towards a statistical theory of clustering. In: Pascal Workshop on Statistics and Optimization of Clustering, pp. 20–26 (2005). Citeseer
50. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4), 417–473 (2010)
51. Law, M.H., Jain, A.K., Figueiredo, M.: Feature selection in mixture-based clustering. In: Advances in Neural Information Processing Systems, pp. 641–648 (2003)
52. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
53. Scrucca, L., Fop, M., Murphy, B.T., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 289–317 (2016)
54. Schwarz, G., *et al.*: Estimating the dimension of a model. The annals of statistics **6**(2), 461–464 (1978)
55. Geweke, J., *et al.*: Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments vol. 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, ??? (1991)
56. Gelman, A., Rubin, D.B., *et al.*: Inference from iterative simulation using multiple sequences. Statistical science **7**(4), 457–472 (1992)
57. Vats, D., Knudson, C.: Revisiting the gelman-rubin diagnostic. arXiv preprint arXiv:1812.09384 (2018)
58. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)
59. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., *et al.*: Transcriptional regulatory code of a eukaryotic genome. Nature **431**(7004), 99–104 (2004)
60. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic acids research **34**(suppl_1), 535–539 (2006)
61. Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., *et al.*: Serial regulation of transcriptional regulators in the yeast cell cycle. Cell **106**(6), 697–708 (2001)
62. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., Brown, P.O.: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409**(6819), 533–538 (2001)
63. Carlson, M., Falcon, S., Pages, H., Li, N.: Org. sc. sgd. db: Genome wide annotation for yeast. R package version **2**(1) (2014)
64. Bando, M., Katou, Y., Komata, M., Tanaka, H., Itoh, T., Sutani, T., Shirahige, K.: Csm3, Tof1, and Mrc1 form a heterotrimeric mediator complex that associates with DNA replication forks. Journal of Biological Chemistry **284**(49), 34355–34365 (2009)
65. Lao, J.P., Ulrich, K.M., Johnson, J.R., Newton, B.W., Vashisht, A.A., Wohlschlegel, J.A., Krogan, N.J., Toczyski, D.P.: The yeast DNA damage checkpoint kinase Rad53 targets the exoribonuclease, Xrn1. G3: Genes, Genomes, Genetics **8**(12), 3931–3944 (2018)
66. Tóth, A., Ciosk, R., Uhlmann, F., Galova, M., Schleiffer, A., Nasmyth, K.: Yeast cohesin complex requires a conserved protein, Eco1p (Ctf7), to establish cohesion between sister chromatids during DNA replication. Genes & development **13**(3), 320–333 (1999)
67. Mehta, G.D., Kumar, R., Srivastava, S., Ghosh, S.K.: Cohesin: functions beyond sister chromatid cohesion. FEBS letters **587**(15), 2299–2312 (2013)
68. Fischle, W., Wang, Y., Allis, C.D.: Histone and chromatin cross-talk. Current opinion in cell biology **15**(2), 172–183 (2003)
69. Bannister, A.J., Kouzarides, T.: Regulation of chromatin by histone modifications. Cell research **21**(3), 381–395

(2011)

70. de Bruin, R.A., Kalashnikova, T.I., Chahwan, C., McDonald, W.H., Wohlschlegel, J., Yates III, J., Russell, P., Wittenberg, C.: Constraining g1-specific transcription to late g1 phase: the mbf-associated corepressor nrm1 acts via negative feedback. Molecular cell **23**(4), 483–496 (2006)
71. Aligianni, S., Lackner, D.H., Klier, S., Rustici, G., Wilhelm, B.T., Marguerat, S., Codlin, S., Brazma, A., de Bruin, R.A., Bähler, J.: The fission yeast homeodomain protein Yox1p binds to MBF and confines MBF-dependent cell-cycle transcription to G1-S via negative feedback. PLoS Genet **5**(8), 1000626 (2009)
72. Ciosk, R., Zachariae, W., Michaelis, C., Shevchenko, A., Mann, M., Nasmyth, K.: An esp1/pds1 complex regulates loss of sister chromatid cohesion at the metaphase to anaphase transition in yeast. Cell **93**(6), 1067–1076 (1998)
73. Cooper, K.F., Mallory, M.J., Guacci, V., Lowe, K., Strich, R.: Pds1p is required for meiotic recombination and prophase I progression in Saccharomyces cerevisiae. Genetics **181**(1), 65–79 (2009)

Additional Files

Additional file 1 — Supplementary materials

Additional relevant theory, background and results. This includes some more formal definitions, details of Bayesian mixture models and MDI, the general consensus clustering algorithm, additional simulations and the generating algorithm used, steps in assessing Bayesian model convergence in both the simulated datasets and yeast analysis, a table of the transcription factors that define the clustering in the ChIP-chip dataset, a table of the gene descriptions for some of the clusters that emerge across the time course and ChIP-chip datasets and Gene Ontology term over-representation analysis of the clusterings from the yeast datasets. (PDF, 10MB)