Investigating Coordinated Architectures Across Clusters in Integrative Studies: a Bayesian Two-Way Latent Structure Model

David M. Swanson[1,*], Tonje Lien[2], Helga Bergholtz[2,4],
Therese Sørlie[2,4], Arnoldo Frigessi[1,3]
[1]Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital
[2]Department of Cancer Genetics,
Institute for Cancer Research, Oslo University Hospital
[3]Oslo Centre for Biostatistics and Epidemiology, University of Oslo
[4]Institute of Clinical Medicine, University of Oslo, Norway

Abstract

We propose a new Bayesian parametric model for integrative, unsupervised clustering across data sources. Such approaches are important in disease subtyping. In our method, we cluster samples in relation to each data source, distinguishing it from methods like Cluster of Cluster Assignments and iCluster, but cluster labels have across-dataset meaning, allowing cluster information to be shared between data sources. We propose a two-way latent structure of cluster assignment variables, within each dataset and across datasets (per sample), which allows borrowing strength across data sources. Importantly, a common scaling across data sources is not required. Inference is obtained by a Gibbs Sampler, which we improve to better cope with sparsity of unoccupied clusters and speed of convergence, and fit models with Gaussian and more general densities which influences across-dataset cluster label sharing. Uniquely, our formulation makes no nestedness assumptions of samples across data sources.

We apply our model to a Norwegian breast cancer cohort of ductal carcinoma in-situ and invasive tumors, comprised of somatic copy-number alteration, methylation and expression datasets. We find enrichment in the Her2 subtype and ductal carcinoma among those observations exhibiting greater cluster correspondence across expression and CNA data.

1 Introduction

Interest in integrative genomic analysis has grown in recent years as the ability to measure a range of genomic features at reasonable cost has increased [1–5]. Integrating different genomic data sources is motivated by understanding the data at the level of biological systems, rather than discrete spaces to only be understood in isolation of one another [6, 7, 5, 8, 9]. It is apparent that the study of functional genomics benefits from acknowledgement of this interplay between data layers [10].

*david.swanson@medisin.uio.no, ORCID: 0000-0003-3174-1656

Clustering algorithms have an important role to play in integrative genomics. One example is that disease subtypes often manifest themselves as distinct clusters in one or multiple datasets [11–13]. As awareness of tumor heterogeneity and grows and data granularity improves, one might also begin using these methods to understand the genomic landscape within tumor [14–16]. Sometimes cluster boundaries are shared across data sources, and sometimes each data source has distinct sets of clusters [17]. Since many of these subtypes are still relatively unknown, unsupervised clustering algorithms have a unique role to play in increasing biological understanding of disease heterogeneity–unsupervised clustering orients itself more towards discovery than those methods which implicitly encode a priori assumptions with class labels.

Integrative clustering approaches have taken different tacks both in terms of underlying assumptions of the biological mechanisms and fitting procedures of the algorithms. An important distinction in the former relates to relative placement of clusters; many models assume that clusters and their boundaries are held in common across data sources and try to best increase statistical power for finding these shared boundaries under this assumption [18–23]. When common boundaries are assumed, some models additionally assume common scaling across different datasets, necessitating "pan-normalization" of them [24, 25]. One disadvantage of the assumption is that even when normalization is done well the discrete nature of the data of certain genomic platforms (eg., methylation data) is in fact not amenable to direct comparison to that which is continuous (eg, RNAseq data). Other models make no assumption about relative placement of cluster boundaries across genomic data sources, but parametrize the model to find common boundaries if they exist [26–29]. These methods additionally tend to draw boundaries in a probabilistic way, such that cluster assignment has a distribution rather than definite label. In some cases, each observation has a single cluster distribution, rather than the observation-dataset pair.

Another distinction in integrative clustering algorithms relates to frequentist, Bayesian, and algorithmic formulations of the models. This distinction has implications on the fitting procedures and their scalability. Frequentist formulations of clustering models have generally been those that assume common cluster boundaries across data sources [30, 4, 18, 23]. Algorithms with similar aims that decompose aggregations of data matrices into lower dimension spaces have also been developed [20–22]. **(author?)** [31]. A Bayesian model makes the common boundary assumption and additionally encodes sparsity into the model via use of priors on variable inclusion. These models have many attractive properties including power gains for finding cluster boundaries when they exist. An important drawback of these models is that due to the iterative nature of model fitting with algorithms like EM and, in other cases, large matrix inversions often involved, they do not scale well to thousands of covariates from many different genomic platforms. An additional drawback is that the number of clusters must generally be specified a priori based on exploratory analysis or post-fitting checks of model fitness criteria, rendering the fitting process multi-stage (eg, [31]). Bayesian models draw probabilistic cluster boundaries and additionally tend to benefit from learning the number of clusters [26–29]. While these approaches often sample closed form conditional posteriors quickly, sometimes iterating involves computationally costly operations (eg, [28]).

We propose a Bayesian, unsupervised, integrative clustering model that attempts to combine many of the strengths of approaches outlined above. Our model is based on two sets of cluster assignment variables of each sample in each dataset: the first set follows a priori a multinomial distribution, for each dataset independently, so that all cluster assignment variables share the same prior within each dataset; the second set of cluster assignment variables is such that each such variable has the same (sample dependent) multinomial probability in each datasets. In this case each sample will have a priori the same probability to be assigned to a cluster label in all datasets. The intuition behind this construction is that the first construction

allows learning cluster assignments within each dataset, the second follows the sample across all datasets. By conditioning on coherence of these two assignments, we are able a posteriori to learn in a natural way within and across datasets. The final cluster assignments depend on three components: a priori cluster probabilities within-dataset (ie, across observations), a priori cluster probabilities within-observation (ie, across dataset), and the likelihood model (ie, how similar the feature vector is to other feature vectors in the same cluster). With appropriate Dirichlet conjugate priors on cluster probabilities, we obtain conditional posteriors for each observation-dataset cluster assignment variable. We describe a Gibbs sampler implementation which performs inference on all parameters.

Our two-way latent structure model (TWL) bears some resemblance to a fully-parametric, Cluster-of-Cluster Assignments (COCA) approach–cluster parameters are dataset specific (thereby making no unreasonable assumption of common scaling or centering across disparate genomic data platforms), but cluster information is shared across data sources [13, 32]. Unlike COCA, a cluster posterior for the TWL model exists for each observation-dataset, though depending on how cluster boundaries align across datasets, this cluster posterior may be dataset-invariant. Our method is not a two step approach, where uncertainty of the first step is ignored, but produces a coherent posterior uncertainty quantification.

Our model is more flexible than many integrative clustering models in that we do not assume that the same samples are measured in the different datasets (nestedness of observations across data sources). Our literature review of data integration methods suggests this is a unique feature of our model. The special case of no common samples across data sources simplifies to fitting independent models on each dataset independently. An additional benefit of the model's formulation is run-time that scales linearly in the number of features. We also learn the number of clusters. One can influence the amount datasets share information through manipulation of hyperparameters as we demonstrate in simulation.

We propose ways to post-process the estimated posterior probabilities for discovering clustering within datasets and identifying that subset of samples whose cluster assignments may span across all datasets. This last metric allows us to examine enrichment in clinical annotation of those samples with greater pan-dataset cluster assignment correspondence. Though we call this "cluster correspondence" because of the context of our model, its calculation is identical to [26]'s "cluster fusing" probabilities.

Our paper is organized as follows: in Section 2 we introduce the model, followed by practical considerations and modifications of it. In Section 3, we describe post-processing metrics and simulation studies to develop intuition. We then perform a data analysis of in-situ ductal carcinoma (DCIS) and invasive (IDC) breast cancer using our TWL model and find 2 copy number clusters, 7 distinct methylation clusters, and confirm the 5 known breast cancer subtype expression clusters [11]. Posterior metrics reveal little correspondence between these disparate clusterings, suggesting that models assuming common cluster boundaries across our datasets could be missleading. We also find modest stratification of the DCIS and invasive samples depending on data source, indicating that the distinction in tumor state may not be reflected equally in genomic data sources. In the Supplementary Material we include a discussion of our modified Gibbs sampler, comparisons with other, common approaches including iCluster and COCA [18], how we address the challenge of label switching, and supporting tables and Figures referenced in-text.

2 Methods

2.1 Model

To give context to description of the TWL model, we use language assuming a clinical study setting where we seek to cluster subjects on whom we have different genomic data sources. However, the model is equally applicable to a setting in which we seek to cluster genes, on which we have different genomic data sets. We also assume each genomic data source has a common set of subjects and no missing values for notational and conceptual clarity, but formulate the more general model (used for our data analysis), which makes no such assumption, in the Supplementary Material.

Let $Y_{i,j}$ be a vector of features for observation i and data set j and of dimension d_j. So the jth data source has information on d_j features. We assume $j \in \{1, \ldots, J\}$, for a total of J data sources, and $i \in \{1, \ldots, N\}$. For example, the first data set (j=1) could be the expression values of d_1 genes, and the second data set (j=2), the copy number variation of d_2 loci.

Consider viewing the $Y_{i,j}$ vectors of length d_j as the atomic units of our model. In doing so, we can think of the TWL model as having 2 separate axes, rows (each row corresponding to a sample) and columns (datasets). We work with a different clustering of the samples for each data set so that every sample is assigned to J clusters. We assume that samples tend to be grouped in the same cluster in the various data sets so that a certain cluster coordination is hypothesized across data sets (see Figure 1).

We start with a model for the clustering within each dataset. The cluster assignment variables corresponding to the *column* (ie, data source) models are

$$\mathbf{C} \equiv (C_{1,1}, \ldots, C_{i,j}, \ldots, C_{N,J})$$

where $C_{ij} \in \{1, 2, \ldots, K\}$ assigns sample i in data set j to one of K clusters, where K is the fixed upper bound on the number of clusters. We assume a multinomial prior distribution for the C_{ij} as

$$C_{(i,j)} \sim Multinom(p_j^{(1)}, p_j^{(2)} \ldots, p_j^{(K)}) \quad \forall \text{ valid } (i,j)$$

The interpretation of $p_j^{(k)}$ is the probability of a draw of cluster label k in dataset j for sample i.

We emphasize that though $C_{(i,j)}$ is subscripted by i and j, the multinomial probabilities are only subscripted by j. Therefore, all observations within dataset j draw from this dataset-specific j^{th} multinomial model.

Traditionally, within each dataset, we could model the data vectors Y_{ij} as a mixture, with one component per cluster and the vector of probabilities $(p_j^{(1)}, \ldots, p_j^{(k)}, \ldots, p_j^{(K)})$ used as mixing parameters. Assuming a parametric model $f(\cdot)$ for the density of Y_{ij}, with a cluster dependent parameter $\theta_j^{(k)}$, this model would be written as $Y_{ij} \propto \sum_{k=1}^{K} p_j^{(k)} f(Y_{ij}|\theta_j^{(k)})$.

However, we want to allow and favor alignment in cluster assignments across datasets. For this purpose we propose a new mixture model with different mixing parameters. This is the intuition: in order to help maintain the samples into the same clusters across data sets, we use a second vector of parameters, $(\rho_i^{(1)}, \ldots, \rho_i^{(k)}, \ldots, \rho_i^{(K)})$, which follow the sample i in all data sets. Cluster assignment of a sample in each data set is therefore influenced by both the p_j and the ρ_i parameters. This second parameter ρ_i helps align clusters across datasets because it is

the same in each data set. We will combine these two models and use as mixing parameters $p_j^{(k)}$ and $\rho_j^{(k)}$ to assign sample i to cluster k in dataset j, as in $Y_{ij} \propto \sum_{k=1}^{K} \rho_i^{(k)} p_j^{(k)} f(Y_{ij}|\theta_j^{(k)})$.

More formally, we now introduce a second set of cluster assignment variables parametrized by these ρ parameters just introduced, which are $\mathbf{R} \equiv (R_{1,1}, \ldots, R_{i,j}, \ldots, R_{N,J})$, where $R_{ij} \in \{1, 2, \ldots, K\}$ assigns sample i in data set j to one of K clusters. We assume a multinomial prior distribution for the cluster assignment variables R_{ij} corresponding to rows (samples)

$$R_{(i,j)} \sim Multinom(\rho_i^{(1)}, \rho_i^{(2)}, \ldots, \rho_i^{(K)}) \quad \forall \text{ valid } (i,j),$$

with K the upper bound on the number of clusters. Note that though R is subscripted by i and j, the parameters are only subsrcipted by i, the sample id. We can loosely think of the cluster models on the i's, cutting across datasets, as models on the rows of the aggregated data sources if we arranged them next to one another (see Figure 1 as an example). The cluster models for $\mathbf{R}$ influence cluster alignment along that axis and, in doing so, giving across-dataset meaning to cluster labels. This occurs despite cluster density parameters not being directly informed by observations of the same cluster label in different data sets.

It is because cluster labels for the ith row of dataset j, which we denote with $R_{i,j}$ and $C_{i,j}$, must agree to produce a coherent and well-defined clustering of samples within each dataset that we condition our model on their equivalence; ie, we condition our likelihood on $R_{i,j}$=$C_{i,j}$ for all valid (i,j) pairs. Simultaneous use of multinomial models along samples (or "rows") and within dataset (or "columns"), the resultant necessary conditioning event of C=R for cluster coherence within each cluster, and the subsequent inferential procedure, are the methodological novelties of the TWL model. This framework results in a joint posterior cluster distribution whose interpretation has granularity at the level of a sample-dataset, rather than just sample. We advocate for this more rich and descriptive, if complex, interpretation in Section 3 below.

Conditional on C=R, and on the dataset- and cluster-specific density parameters $\theta_j^{(k)}$ introduced below, Y_{ij} follows a typical Gaussian mixture model, with cluster probabilities being normalized products of row and column model specific cluster probabilities. For $f(\cdot)$ the Gaussian density, the mixture model is

$$Y_{ij} \mid (C{=}R), \theta_j^{(k)} \sim \kappa \sum_k p_j^{(k)} \, \rho_i^{(k)} \cdot f(Y_{ij} \mid \theta_j^{(k)})$$

with κ the normalizing constant.

We need to specify the prior model further. The prior probabilities for $\mathbf{p_j}$ and $\boldsymbol{\rho}_i$ are

$$\mathbf{p_j} \sim Dirichlet(\beta_1, \ldots, \beta_K) \;\; \forall \;\; j \quad \text{and} \quad \boldsymbol{\rho}_i \sim Dirichlet(\alpha_1, \ldots, \alpha_K) \;\; \forall \;\; i$$

respectively, with hyperparameters α_i and β_j constant in i and j, for which hyperpriors could further be assumed. Instead we choose α_i as a function of the average number of data sets per sample (or simply number of data sets if all samples are present in all data sets and unique within them), and β_j as a function of the average number samples within each data set (or simply the number of unique sample ids if again all samples are present in all data sets and unique within them). As a result, we have $\alpha \equiv \alpha_1 = \alpha_2 = \cdots = \alpha_K$ and $\beta \equiv \beta_1 = \beta_2 = \cdots = \beta_K$.

Define the Gaussian cluster density parameters specific to dataset j and cluster k with $\theta_j^{(k)} \equiv (\mu_j^{(k)}, \tau_j^{(k)})$, where $\mu_j^{(k)}$ is the mean vector of dimension d_j. We assume a priori independence for elements of vector $Y_{i,j}$ so that precision parameter $\tau_j^{(k)}$ is also of dimension d_j. The

assumption increases computation efficiency significantly, and filtering correlated, redundant features before analysis makes the assumption reasonable. We generalize beyond the Gaussian likelihood in Section S1.1.2.

We place a Gaussian prior on $\mu_j^{(k)}$ with mean and precision parameters $\mu_{0,j}$ and $\psi_{0,j}$, respectively. We take an Empirical Bayes approach to choosing these hyperparameters, setting $\mu_{0,j}$ to the mean of $Y_{\cdot,j}$ (ie, dataset j), and the corresponding precision parameter to

$$\psi_{0,j} = \left(\frac{1}{2} \cdot \frac{\sigma_G^{2(j)}}{N/20} \cdot 15\right)^{-1}$$

for all j, where $\sigma_{G,j}^2$ is a vector of variances of the features of $Y_{\cdot,j}$. Setting $\psi_{0,j}$ in this way makes the assumption that if half of marginal variation in data set j, $\sigma_{G,j}^2$, is attributable within-cluster and the size of clusters are 1/20 of the total sample size, N, then there will be 1/15 cluster mean shrinkage to the global mean under such circumstances. If one fits the model with an upper bound of more than 20 clusters, this will result in greater mean shrinkage earlier in the MCMC chain since the chain begins with random cluster assignment. If in truth there are fewer than 20 clusters in the data, it will result in less shrinkage of the respective cluster means once convergence has occurred.

Such a prior also serves a practical purpose: without it, small clusters, especially those with one observation, will be resistant to becoming unoccupied since the mean parameter $\mu_{k,j}$ will tend to be close or identical to the corresponding (few) observation(s) and thus inflate the likelihood. This is even more true if additionally cluster variance parameters $\tau_{k,j}$'s were cluster-specific, as estimated Gaussian densities will diverge to infinity. By shrinking $\mu_{k,j}$ to a global mean, especially when the cluster size is small, the posterior of C, $R \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C}{=}\mathbf{R}$ will exhibit greater cluster sparsity.

With this same goal of convergence, we assumed a common variance parameter $\tau_j(k) = \tau_j$ across clusters and one that is considered fixed. Doing so leads to more efficient estimation in the parameter. While the variance across clusters within dataset may not always be constant, since the model has no effective upper bound on the number of clusters, the posterior tends to favor the partition of more heterogeneous clusters into smaller, more homogeneous ones. This characteristic is likely helpful for interpretation.

The full conditional posterior for C (which we only consider since we condition on C=R), is

$$C_{ij}|Y_{ij}, \mathbf{C} = \mathbf{R}, \rho_i, p_j, \mu_j^{(\cdot)}, \tau_j \sim Multinom\Big(p_j^{(1)}\rho_i^{(1)} \cdot f(Y_{ij}|\mu_j^{(1)}, \tau_j),\ \ldots,\ p_j^{(K)}\rho_i^{(K)} \cdot f(Y_{ij}|\mu_j^{(K)}, \tau_j)\Big) \quad (1)$$

The full conditional posterior for $\mu_j^{(k)}$ is

$$\mu_j^{(k)} \mid \mathbf{Y}_{\cdot,\mathbf{j}}, \boldsymbol{\mu}_{0,j}, \hat{\boldsymbol{\tau}}_{\boldsymbol{j}}, \boldsymbol{\psi}_{\mathbf{0},\boldsymbol{j}} = N\left(\mu_{k,j} \,\middle|\, \left[N\,\widehat{\tau}_j \cdot \bar{\mathbf{Y}}_{\cdot,\mathbf{j}} + \psi_{0,j} \cdot \mu_{0,j}\right] \cdot \left[\psi_{0,j} + N\,\widehat{\tau}_j\right]^{-1}, \left[\psi_{0,j} + N\,\widehat{\tau}_j\right]^{-1}\right)$$

where $\widehat{\boldsymbol{\tau}}_{\boldsymbol{j}}$ is estimated from the data according to an empirical Bayes approach, assuming $\boldsymbol{\tau}_j \equiv \tau_j^{(1)} = \cdots = \tau_j^{(k)} = \cdots = \tau_j^{(K)}$.

The full conditional posteriors for p_j and ρ_i are, respectively,

$$p_j|C_{\cdot j}\,,\ \alpha \sim Dirichlet\Big((p_j^{(1)})^{\alpha+\sum_{i|j} I(C_{i,j}=1)}, (p_j^{(2)})^{\alpha+\sum_{i|j} I(C_{i,j}=2)}, \ldots, (p_j^{(K)})^{\alpha+\sum_{i|j} I(C_{i,j}=K)}\Big)$$

and

$$\rho_i | R_{i\cdot}\,,\, \beta \sim Dirichlet\Big((\rho_i^{(1)})^{\beta+\sum_{j|i} I(R_{i,j}=1)}, (\rho_i^{(2)})^{\beta+\sum_{j|i} I(R_{i,j}=2)}, \ldots, (\rho_i^{(K)})^{\beta+\sum_{j|i} I(R_{i,j}=K)}\Big)$$

We write the posterior of ρ_i with R_{ij} to emphasize ρ_i's connection to the "row" multinomial models. However, since $C_{ij} = R_{ij}$ for all (i, j) by assumption, using C_{ij} instead is an identical formulation. Additionally, we use the notation $i \,|\, j$ and $j \,|\, i$ to denote valid values of i and j within strata of j and i, respectively. We use this notation despite assuming a common set of subject ids for all datasets in this simpler formulation of the TWL model to emphasize that no such assumption is necessary.

We report the general model in the Supplementary Material which assumes dataset-specific sets of sample ids, along with modifications of our Gibbs sampler and generalization of the density function beyond the Gaussian case.

2.1.1 Hyperparameter tuning

Choice of α and β can be a function of the total number of samples N and number of datasets D, respectively, and the specified maximum number of clusters. If we consider α, increasing its value dilutes the effect of cluster distribution on the posterior of $\mathbf{p_j}$, and subsequently $C_{\cdot j}$, by increasing the unnormalized probabilities of all labels equally. Similar reasoning holds for β and its effect on ρ_i and $R_{i\cdot}$. The special case of $\beta \to \infty$ results in a lack of cluster information sharing across datasets, effectively fitting independent Gaussian mixture models on each dataset. By choosing α and β so that the proportions $N \cdot K/\alpha$ and $D \cdot K/\beta$ are constant in applying the TWL model in different settings, one additionally keeps their influence on the model posterior constant. We discuss choice of α and β more in the supplementary materials.

3 Results

3.1 Simulation

We generated several different data sets to demonstrate different properties of the TWL model. First, we generated 2 different data sets with nested cluster patterns across data sets (See Figures 2 and 3). Second, to demonstrate how common cluster sharing across data sets can be used to learn structure of cluster patterns in the data sets, we generated "progressively misaligned" clusters, whose diagram can be seen in Figure 1. These simulations illustrate how cluster assignment in one data set influences that in other ones, depending on alignment pattern. They also resemble real genomic data sets whose cluster patterns are not often perfectly aligned across genomic datasets and which exhibit nestedness patterns to some degree [17]. For example, for RNAseq and CNA tumor data on a common set of subjects, clusters defined solely by RNAseq and solely by CNAs would likely not align. Additionally, within, say, a CNA cluster, there very well may be multiple RNAseq clusters. Ability to find such structure is an emphasis of our model.

3.1.1 First nested scenario

We see in Figure 2 our first nested data scenario, a collection of five data sets. Each data sets consists of 200 observations, each with 10 features. In each of the five data sets, observations in blocks denoted in the figure are drawn from the same multivariate normal distribution. The multivariate normal distributions corresponding to the different blocks have the same

covariance matrix but different means. Means were generated randomly with each sampled element in the mean vector having a standard deviation of 1.4. The standard deviation of observations from respective means was 1 in all 10 feature dimensions, with features generated independently. We make this independence assumption in simulation and later analysis to significantly increase computational efficiency. The assumption is valid in analysis if data is processed so that highly correlated features are excluded. Moving from left to right over the 5 data sets in Figure 2, the number of clusters varies from 2 to 6, with each additional cluster in the next data set resulting from splitting the bottom cluster in the previous data set in half and introducing slight misalignment. Such a telescoping pattern helps us understand how well our model identifies the small clusters in the right-most data sets and how common cluster labels are shared across data sets. There is a trade-off between fitting larger, "umbrella" clusters, with a single cluster label, and using multiple, different labels for the "enclosed" clusters in successive data sets because of information sharing across them all. Analyzing how our model resolves these trade-offs yields information on relative signal clarity and size of clusters, and their alignments.

3.1.2 Second nested scenario

Figure 3 shows our second nested data scenario with five data sets, each 200 observations by 10 features. Now all clusters added in subsequent data sets are of the same size. Such a pattern allows us to again examine the effect of nested clusters in posterior cluster labels, but no one cluster becomes especially small in the furthest right data sets, unlike the first nested scenario.

3.1.3 Progressively misaligned scenario

Figure 1 shows a set of five data sets of the same dimension as those above exhibiting no nestedness in its cluster patterns. In this scenario, however, clusters are progressively misaligned. The first data set consists of 10 clusters each consisting of 20 observations. The next data set to the right also has 10 clusters, though those clusters are misaligned by 2 observations (10% of the size of each cluster). The next data set to the right again has 10 clusters, again misaligned by 2 observations as compared to the second data set. This pattern continues until the 5th dataset is almost entirely misaligned as compared to the first data sets, by 8 observations total. Data sets were generated in this way in order to examine how posterior cluster labels are held in common in less (e.g., the third and fourth data sets) and more (the first and fifth data sets) cluster-misaligned datasets.

3.2 Posterior interpretation

Unlike many other integrative clustering models, the TWL model yields cluster assignments for each sample in each dataset. One can therefore examine for each observation a measure of cluster membership across datasets, which we call "cluster correspondence". We propose this measure and four related metrics to understand clustering patterns in our data sets.

Metric 1 ("Within"): This metric looks at cluster co-assignment of observations, within datasets. We calculated the matrix of elements

$$P(C_{ij} = C_{i^*j}) \ \ \forall \text{ pairs } i, i^* \text{ with } i \neq i^*, \forall\ j$$

which we estimate by the proportion of samples in our chain after burn-in where observations i and i^* have a common cluster label. These proportions result in J symmetric $N \times N$ matrices. Figures 6 and 7 show two examples of heatmaps of such output matrices.

Metric 2 ("Between"): This metric looks at cluster assignment between datasets, within observations. We calculate the matrix of

$$P(C_{ij} = C_{ij^*}) \ \forall \text{ pairs } j, j^* \text{ with } j \neq j^*, \forall \ i$$

There are N such matrices, one for each observation, each of dimension $J \times J$ with $J * (J - 1)$ unique elements. If we consider the more general TWL model where samples are not assumed nested across datasets, some elements in a subset of these matrices may be NA if the corresponding observation does not have data on both data sources associated with that element. The interpretation of an element of a particular matrix is the proportion of common cluster assignments for that pair of data sets for that observation.

Hierarchical clustering: We perform hierarchical clustering on each of the $N \times N$ matrices from Metric 1. The recommended workflow is to first examine heatmaps of the matrices generated in Metric 1, determine the number of clusters, and then cut the dendrogram at the appropriate place to generate that number of clusters and make a determination of cluster membership if that is needed and desired. We make additional suggestions about this post-processing analysis in Section 3.

Between-dataset cluster correspondence: Consider the N matrices from Metric 2: we can average element-wise over all such N matrices. The interpretation of entry (j, j^*) in the resulting symmetric $J \times J$ matrix with elements in $[0, 1]$ is the degree to which cluster boundaries are similar in datasets j and j^*. Higher numbers indicate more similar boundaries.

The best example of the use of this measure is for the misaligned cluster simulation (see Figure 1) and demonstrates that the more clusters do not align, the lower are elements in the matrix. Indeed, we observe a decreasing cascade of proportion of common clusters for more "distal" datasets, as the misalignments in datasets increases.

Sample-specific cluster correspondence: Consider again the N matrices from Metric 2: we can average over the $J(J - 1)$ unique elements in each matrix, resulting in a length N vector. The interpretation of the entry corresponding to observation i in the vector, between 0 and 1, is the degree to which cluster membership is similar across the J datasets for that observation, what we call sample-specific "cluster correspondence". A higher number indicates greater common cluster membership across datasets for the observation. One can also calculate cluster correspondence just for subsets of datasets for more specific hypotheses.

3.3 Simulation Results

We show the PSMs for our 2 nestedness simulations in Figures 6 and 7. The figures reveal that the TWL model can identify the clusters with a high degree of fidelity. As one might expect from the telescoping nested scenario, the cluster labels in the lowermost block from the first dataset in that simulation scenario are not as stable as the uppermost block, due to the "splintering" of clusters in subsequent datasets and across-dataset influence of cluster membership. One can see a similar phenomenon in the smaller cluster blocks of (c), (d), and (e), and also the larger blocks of Figure 7 (a) and (b) – while their outline and therefore recognition is clear, common labelling of those observations is not as stable, as should be.

We developed the misaligned cluster simulation in part to examine between-dataset cluster correspondence, which is shown in Figure 1 for two different values of the α hyperparameter. In both (a) and (b), we see decreasing values in across-dataset cluster correspondence as one moves away from the diagonal, in a Toeplitz or auto-regressive pattern as one would expect.

We also see the influence of α on this measure and again recommend some normalization of that parameter with respect to the total number of dataset used in analysis.

We developed the nestedness cluster simulation in part to examine sample-specific cluster correspondence, shown in Figure 4. The horizontal bar chart in (a) is the observation specific cluster correspondence. We observe greater cluster correspondence in observations belonging to clusters in the first dataset that are not split in subsequent data sets. The lowermost observations, whose cluster membership in the left-most datasets are split frequently into smaller clusters in subsequent right-most datasets, have much smaller cluster correspondence measures.

We also calculated the sample-specific cluster correspondence on the misaligned cluster scenario, shown in Figure 5. We observe greater cluster correspondence in those observations whose own row is less likely to be crossed by a cluster boundary in one of the datasets. Since this feature varies by observation and all clusters in all datasets are the same size, one observes growing and shrinking peaks of the cluster correspondence measure as one moves down the horizontal barplot (Figure 5 (b)).

3.4 Breast cancer subtyping in DCIS and IDC tumors

3.4.1 Preprocessing

We performed an analysis of ductal carcinoma in-situ and invasive breast cancer tumors comprising all intrinsic subtypes coming from a cohort of Norwegian and Italian women. The cohort includes 370 women, 57 in the DCIS tumor state, and 313 IDC, and the data has been described in greater detail elsewhere [33, 34]. We used Agilent 60K microarrays to measure gene expression on 370 observations and 21887 genes from this cohort. Quantile normalization was performed, and probes collapsed over genes by mean value [35].

We used the Illumina 450K array to measure methylation of CpG sites on 314 of the 370 women, normalized according to the method described in [36]. We set as missing probes whose detection values had p-values above 0.05. Probes targeting a given CpG were all removed if more than 25% of probes were denoted as missing and performed imputation on those remaining with k-nearest neighbors. Probes were grouped by gene with flanking regions of 50kb, and we collapsed the data to gene-level granularity by using the score according to the first principal component [37].

We had copy number alteration data on 338 observations and 19089 genes, the women being a subset of the 370 on whom we had expression data and not a superset of those on whom we had methylation data. We used the SNP 6.0 array and GC-content corrected raw log ratio values. Segmentation was performed using the PCF algorithm in the R package *copynumber* [38]. We collapsed to gene-level data by determining which copy number segment overlapped the gene most.

3.4.2 Analysis

We performed TWL analysis in the statistical programming language R [39]. We ran our Gibbs Sampler for 4000 iterations for different feature draws from the approximately 20000 genes available, in order to investigate reproducibility. Our datasets consisted of approximately 1790 of the same genes from all genomic data sources. We used $\alpha = 0.4$ and $\beta = 7$ as hyperparameters. We ran also parallel, independent chains on identical draws to assess convergence and Monte Carlo error. Convergence of cluster membership occurred relatively early, though we considered the first 2000 iterations burn-in.

PSMs and corresponding heatmaps were generated to give Figure 8. We performed hierarchical clustering analysis on these heatmaps, manually inspected the tree, and cut it at a height so that approximately 95% of the samples had been grouped with at least one other observation (Figure S1). Of those which had been grouped, we gave an identifying label to the M largest clusters (where M was 5, 2 or 3, and 7, based on inspection of the expression, CNA, and methylation heatmaps, respectively), and the rest grouped into a more heterogeneous cluster we called "unknown". The idea behind such a procedure was to label those observations clearly belonging together in the same cluster as such, and to keep those more heterogeneous observations in their own separate group to not dilute signal clarity. One can use different thresholds for M and the tree height depending on proportion of observations falling into a clear cluster. For example, the PSM for CNA suggests that there is a greater proportion of observations with unknown cluster membership, and so we used a dendrogram height translating to approximately 90% of samples already grouped with at least one other observation.

We calculated the between-dataset cluster correspondence on two different feature draws from our datasets (Table 2), indicative of cluster boundary alignment between data sources as a whole, and found little suggestion of significant boundary overlap in both cases. This stands in contrast to scenarios like our simulations (e.g., Figure 2). Despite the different draws from the data, there was stability in estimation of the measure.

We found considerable stability in clustering between parallel MCMC chains (Tables S2 and S3), and relatively high stability in cluster assignment (Tables 3 and S1). In the case of either draw from the data, cross tabulations of at least expression-based cluster membership and breast cancer intrinsic subtypes (PAM50) showed that misclassification, where it occurred, mainly was between more closely related subtypes (e.g., Luminal A and Luminal B, see Table S4, [12, 40]). While there is some correspondence between CNA or methylation clusters and subtypes, it is not as strong (Table S5). The observations assigned to the "unknown" cluster were relatively evenly spread over the different subtypes, reflective of their heterogeneity. An exception to that pattern was the Basal subtype, which was disproportionately represented among the unknown cluster label for CNAs (Table S12 and Figure S1b). Most Basal samples could not be categorized into a CNA cluster.

We discuss the results of the iCluster and COCA analyses in the Supplementary Material, but have included annotation from these analyses in Figure 8. We say briefly that because both methods assume common cluster boundaries across all data sources, the degree to which consensus clusters reflect those of the particular data source varies considerably.

In general cross tabulations of DCIS and Invasive tumor state with posterior cluster labels from the different data sources was not as compelling as that for expression and intrinsic subtypes. An important exception however was a relatively small expression cluster that contained many of the DCIS samples (Table S13). We additionally found that nearly all DCIS tumors were contained by the large CNA cluster identified in our analysis (Table S9). The second, smaller cluster identified in the CNA analysis contained almost entirely invasive tumors. There was a similar finding in the methylation dataset where DCIS tumors were significantly overrepresented in one of the clusters, though a small one in this case.

Because our between-dataset cluster correspondence measure suggested relatively little cluster overlap between datasets on the whole, sample-specific cluster correspondence becomes more interesting: subsets of samples with greater correspondence can be enriched for clinical annotation. We calculated sample-specific cluster correspondence on expression and CNAs and examined the 50 observations with the largest such cluster correspondence. We observed

enrichment in the Her2 subtype and DCIS tumor state as compared to the marginal distribution of these variables. These features were observed across parallel MCMC chains and different draws from the data and so appear as robust findings (see Tables S7 and S10). We found similar enrichment in those posterior cluster labels that roughly correspond to the Her2 subtype and DCIS, which may be considered proxies for these types. In either case, there is an important caveat in that both Her2 and its proxy cluster label, and DCIS and its proxy cluster label, are almost entirely contained in the very large CNA cluster one sees in Figure 8b. Thus, this enrichment may be an inevitable result of that cluster's size and a large proportion of observations with those annotations falling within it. However, the normal subtype is also almost entirely found in this large CNA cluster, but we observe no such enrichment in it when looking at expression-CNA cluster correspondence (Table S12). The finding should therefore be considered with caution, though not discounted.

We observe a lack of cluster correspondence between CNA and methylation clusters among the Basal subtype in particular, across parallel MCMC chains and draws from the data (Table S11). The interpretation of this phenomenon is difficult, in part because of the Basal subtype's aforementioned presence in the large CNA cluster. One hypothesis is that the alignment of the Basal observations between the expression-methylation and expression-CNA pairs is higher than under the null hypothesis of independent clusterings on each dataset, but on disjoint sets of Basal observations, pushing the alignment for CNA-methylation low, as observed. However examination of the data suggests that this hypothesis likely does not fully explain the lack of CNA-methylation correspondence observed in the Basal subtype. Another unexpected feature of the Basal subtype was its presence in primarily only 2 methylation clusters (Figure S1c).

When looking at pan-genomic cluster correspondence (that measure shown for our simulations in Figures 4a and 5a), we observed no enrichment in subtypes or tumor state. Since we observe enrichment for specific pairs of data sources, the lack of cluster correspondence on all three is either because the enrichment we do see does not carry over to the excluded source, or that there is some bias in types of samples on whom certain data platforms could be measured.

4 Discussion

We developed an innovative integrative clustering model called TWL in which we envisioned the existence of multinomial models on clusters along both the "column" (or dataset) and "row" (or observation) of the aggregated datasets. For model coherence, we conditioned on the event that these two otherwise disparate sets of clusterings were equal to one another, resulting in closed form posteriors. The model shares cluster assignment information across datasets, giving across-dataset meaning to the labels, though fits cluster parameters within-dataset. The model learns the total number of clusters with the analyst specifying an upper bound, and posteriors exist on the level of the sample-dataset pair, rather than only on sample which leads to a loss of information. Run time of cluster estimation scales linearly in the total number of covariates in the datasets being integrated.

We argue that forcing a clustering on samples, invariant to data source, seriously oversimplifies diseases subtyping. In our analysis each observation appears best characterized by its full set of clusters, one for each data set. Subtyping appears more precise and potentially more useful for treatment and prognoses by duly recognizing this complexity. Our approach is flexible, in the sense that if a single cluster assignment would emerge from the integrative clustering, it would appear in the TWL model, which however is able to capture deviations from it. Our analysis on the breast cancer data shows clearly that a single, unique clustering of all patients cannot account for pan-genomic sample heterogeneity and can therefore be misleading. We

believe that a TWL analysis of breast cancer patients would be the best starting point to study precision treatments.

Our integrative clustering data analysis is unique in that it includes 57 women with ductal carcinoma in-situ, in addition to 313 women with invasive breast cancer. The Cancer Genome Atlas exclusively includes invasive tumors, and so our dataset and unsupervised TWL analysis is positioned to characterize the genomics of DCIS and invasive cancers in a way that is impossible in the well-analyzed TCGA samples. We leverage our integrative clustering analysis to better understand breast cancer subtypes in both DCIS and invasive breast cancer cases.

The TWL model is built to achieve robust and stable results. Because of the dimensionality of the genomic datasets, we proposed a modified likelihood with lower bounding tail values of the Gaussian model, thus coarsening the data, to achieve greater mixing in the MCMC. We experimented with simulated annealing [41], but found convergence improved under our alternative approach. Our Gibbs sampler was additionally modified to avoid excessive sampling of clusters close to the global mean of features in each dataset. Our ability to clearly find 5 distinct expression clusters, which align to a high degree with the known breast cancer subtypes, without specification of the number of clusters and across parallel chains and draws from the data serves as validation of our model fitting procedure on this dataset. We additionally found 2 distinct CNA clusters and 7 methylation clusters, and a relative lack of "cluster correspondence" across datasets as judged by our proposed measure.

We did not observe a high degree of pan-genomic cluster correspondence, and primarily found modest enrichment in the Her2 subtype and DCIS tumor state in those samples with greater expression-CNA cluster correspondence. There are different reasons we may not have seen significant pan-genomic cluster correspondence. According to the analysis of [10], tumor purity and cell type composition confound the relationship between expression and methylation genomic data sources, and we are unable to control for these factors in our analysis. [10] also finds that CNAs affect expression and methylation independently. If this is the case, one would expect weaker expression-methylation associations, and stronger associations between the expression-CNA and CNA-methylation pairs.

[8] performs a similar analysis to examine correspondence across genomic platforms, calculating Spearman correlation on gene-annotated expression, copy number, and proteomic probes both within subtypes and also marginally. Direct comparison with their results is difficult however, as they do not perform a genome-wide analysis, but examine the PI3K/Akt pathway. They do observe variation by subtype in the degree of correlation between measures on different platforms for specific genes. However, examination of all genes in the pathway does not reveal broad trends of greater correlation in one subtype over another on a specific pair of data sources. In particular, the Her2 subtype does not seem to exhibit greater expression-CNA Spearman correlation for the genes interrogated than the other subtypes.

We found that nearly all DCIS tumors inhabit the large CNA cluster identified in our analysis in a robust fashion. There were similar and perhaps more significant findings in the expression dataset, where one cluster seemed to contain most DCIS observations (see Table S13). The large CNA cluster also contained a large proportion of invasive tumors, however, and the second cluster in the CNA analysis contained almost entirely invasive tumors. The finding of a single CNA cluster for all DCIS and some IDC seems to stand in contrast to findings of [42], who did not identify genomic loci that consistently separated the two tumor states. The difference could result from the accumulation of small differences between the two tumor states. This in turn could lead to the fairly characteristic differences one sees in the two CNA clusters in our analysis and to which other methods are not sensitive. Also unclear was why

we observed so little CNA cluster homogeneity in the Basal intrinsic subtype so that most of its samples had the "unknown" cluster classification. The finding suggests that nearly all Basal samples are unlike one another and unlike any other subtype with respect to CNAs.

It may not be surprising that we observe relatively little commonality between expression and methylation clusterings. [43] finds that the correlations that do exist in breast and other cancers are in certain, tissue-specific sets of genes. While their analysis only includes invasive tumors from TCGA, genome-wide analysis therefore may very well not detect such a specific subset. We also may not expect a lot a correlation between expression clusters and those of CNAs, manifested either through posterior cross-tabulations of cluster labels from those data sources or the more global measure of between-dataset cluster correspondence. Indeed, when [7] integrates the two sources of information, they find CNAs stratify subtypes in some cases and also group subsets from disparate ones together, resulting in 10 total integrated clusters.

The volume of genomic data will only increase, and integrative clustering models have an important role to play in giving insight into underlying biology. The TWL model, because of its scalability and flexibility, can aid in this understanding.

5 Data Availability

Data is available within our published R package on CRAN as "twl" [39].

6 Acknowledgements

We acknowledge useful discussions with Sylvia Richardson and Paul Kirk. Funding for this research was also provided by the Research Council of Norway, BigInsight, the Norwegian Cancer Society, and the South-Eastern Norway Regional Health Authority. We thank Dr. Maria Grazia Daidone (Fondazione IRCCS Instituto Nazionale dei Tumori, Milan, Italy) for the contribution of patient samples to the project.

References

[1] Claudia Cava, Gloria Bertoli, and Isabella Castiglioni. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. *BMC Systems Biology*, 9(1), December 2015.

[2] N. Huang, P. K. Shah, and C. Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13(3):305–316, May 2012.

[3] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, October 2012.

[4] Nora K. Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, June 2015.

[5] Vessela N. Kristensen, Ole Christian Lingjærde, Hege G. Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, May 2014.

[6] H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, Samuel AJR Aparicio, and Carlos Caldas. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology*, 15(8):431, 2014.

[7] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, Christina Curtis†, Sohrab P. Shah, Carlos Caldas, Samuel Aparicio, James D. Brenton, Ian Ellis, David Huntsman, Sarah Pinder, Arnie Purushotham, Leigh Murphy, Carlos Caldas, Samuel Aparicio, Carlos Caldas, Helen Bardwell, Suet-Feung Chin, Christina Curtis, Zhihao Ding, Stefan Gräf, Linda Jones, Bin Liu, Andy G. Lynch, Irene Papatheodorou, Stephen J. Sammut, Gordon Wishart, Samuel Aparicio, Steven Chia, Karen Gelmon, David Huntsman, Steven McKinney, Caroline Speers, Gulisa Turashvili, Peter Watson, Ian Ellis, Roger Blamey, Andrew Green, Douglas Macmillan, Emad Rakha, Arnie Purushotham, Cheryl Gillett, Anita Grigoriadis, Sarah Pinder, Emanuele di Rinaldis, Andy Tutt, Leigh Murphy, Michelle Parisien, Sandra Troup, Carlos Caldas, Suet-Feung Chin, Derek Chan, Claire Fielding, Ana-Teresa Maia, Sarah McGuire, Michelle Osborne, Sara M. Sayalero, Inmaculada Spiteri, James Hadfield, Samuel Aparicio, Gulisa Turashvili, Lynda Bell, Katie Chow, Nadia Gale, David Huntsman, Maria Kovalik, Ying Ng, Leah Prentice, Carlos Caldas, Simon Tavaré, Christina Curtis, Mark J. Dunning, Stefan Gräf, Andy G. Lynch, Oscar M. Rueda, Roslin Russell, Shamith Samarajiwa, Doug Speed, Florian Markowetz, Yinyin Yuan, James D. Brenton, Samuel Aparicio, Sohrab P. Shah, Ali Bashashati, Gavin Ha, Gholamreza Haffari, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, April 2012.

[8] Simen Myhre, Ole-Christian Lingjaerde, Bryan T. Hennessy, Miriam R. Aure, Mark S. Carey, Jan Alsner, Trine Tramm, Jens Overgaard, Gordon B. Mills, Anne-Lise Børresen-Dale, and Therese Sørlie. Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Molecular Oncology*, 7(3):704–718, June 2013.

[9] David J Reiss, Nitin S Baliga, and Richard Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, page 22, 2006.

[10] Wei Sun, Paul Bunn, Chong Jin, Paul Little, Vasyl Zhabotynsky, Charles M Perou, David Neil Hayes, Mengjie Chen, and Dan-Yu Lin. The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Research*, February 2018.

[11] Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt Van De Rijn, and Stefanie S. Jeffrey. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[12] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, and Stephanie Geisler. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003.

[13] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Noreen Dhalla, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, and Nam H. Pho. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, September 2012.

[14] Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E. Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, August 2014.

[15] T. A. Yap, M. Gerlinger, P. A. Futreal, L. Pusztai, and C. Swanton. Intratumor Heterogeneity: Seeing the Wood for the Trees. *Science Translational Medicine*, 4(127):127ps10–127ps10, March 2012.

[16] So Yeon Park, Mithat Gönen, Hee Jung Kim, Franziska Michor, and Kornelia Polyak. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *Journal of Clinical Investigation*, 120(2):636–644, February 2010.

[17] Dvir Netanely, Ayelet Avraham, Adit Ben-Baruch, Ella Evron, and Ron Shamir. Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Research*, 18(1), December 2016.

[18] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, November 2009.

[19] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, 7(1):269–294, March 2013.

[20] Prabhakar Chalise and Brooke L. Fridley. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS one*, 12(5):e0176278, 2017.

[21] Prabhakar Chalise, Devin C. Koestler, Milan Bimali, Qing Yu, and Brooke L. Fridley. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research*, 3(3):202, 2014.

[22] Yanhua Chen, Manjeet Rege, Ming Dong, and Jing Hua. Non-negative matrix factorization for semi-supervised data clustering. *Knowledge and Information Systems*, 17(3):355–379, December 2008.

[23] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[24] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, March 2013.

[25] Kristoffer H. Hellton and Magne Thoresen. Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*, 17(3):537–548, July 2016.

[26] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, December 2012.

[27] David B. Dunson and Amy H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25, 2005.

[28] Eric F. Lock and David B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, October 2013.

[29] Evelina Gabasova, John Reid, and Lorenz Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLOS Computational Biology*, 13(10):e1005781, October 2017.

[30] Matthias Kormaksson, James G. Booth, Maria E. Figueroa, and Ari Melnick. Integrative Model-based clustering of microarray methylation and expression data. *The Annals of Applied Statistics*, 6(3):1327–1347, September 2012.

[31] Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2017.

[32] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D.M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A. Margolin, Laura J. van't Veer, Nuria Lopez-Bigas, Peter W. Laird, Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, Christopher C. Benz, Charles M. Perou, and Joshua M. Stuart. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944, August 2014.

[33] R Lesurf, MR Aure, HH Mørk, V; Oslo Breast Cancer Research Consortium (OSBREAC) Vitelli, S Lundgren, AL Børresen-Dale, V Kristensen, F Wärnberg, M Hallett, and T Sørlie. Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Research*, 19(1), December 2017.

[34] Aslaug Aamodt Muggerud, Michael Hallett, Hilde Johnsen, Kristine Kleivi, Wenjing Zhou, Simin Tahmasebpoor, Rose-Marie Amini, Johan Botling, Anne-Lise Børresen-Dale, Therese Sørlie, and Fredrik Wärnberg. Molecular diversity in ductal carcinoma *in situ* (DCIS) and early invasive breast cancer. *Molecular Oncology*, 4(4):357–368, August 2010.

[35] Dhammika Amaratunga and Javier Cabrera. Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, 96(456):1161–1170, December 2001.

[36] Nizar Touleimat and Jörg Tost. Complete pipeline for Infinium ® Human Methylation 450k BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–341, June 2012.

[37] C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey, C J Marsit, E A Houseman, and R Brown. Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6):1394–1402, September 2013.

[38] Gro Nilsen, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Marianne B. Eide, Oscar M. Rueda, Suet-Feung Chin, Roslin Russell, Lars O. Baumbusch, and Carlos Caldas. Copynumber: Efficient algorithms for single-and multi-track copy number segmentation. *BMC genomics*, 13(1):591, 2012.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[40] Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J.S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, March 2009.

[41] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[42] Robert Lesurf, Miriam Ragle Aure, Hanne Håberg Mørk, Valeria Vitelli, Steinar Lundgren, Anne-Lise Børresen-Dale, Vessela Kristensen, Fredrik Wärnberg, Michael Hallett,

Therese Sørlie, Torill Sauer, Jürgen Geisler, Solveig Hofvind, Elin Borgen, Anne-Lise Børresen-Dale, Olav Engebråten, Øystein Fodstad, Øystein Garred, Gry Aarum Geitvik, Rolf Kåresen, Bjørn Naume, Gunhild Mari Mælandsmo, Hege G. Russnes, Ellen Schlichting, Therese Sørlie, Ole Christian Lingjærde, Vessela Kristensen, Kristine Kleivi Sahlberg, Helle Kristine Skjerven, and Britt Fritzman. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Reports*, 16(4):1166–1179, July 2016.

[43] Matahi Moarii, Valentina Boeva, Jean-Philippe Vert, and Fabien Reyal. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 16(1), December 2015.

[44] Carlos E. Rodríguez and Stephen G. Walker. Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, January 2014.

[45] Arno Fritsch and Katja Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, June 2009.

[46] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, December 2011.

[47] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236, March 1963.

Table 1: Tables of the between-dataset cluster correspondence for the misaligned cluster simulation, shown in Figure 1, for different values of α. In either table, entry j, j^* is the average cluster overlap of datasets j and j^*. One expects more cluster overlap when clusters align to a greater degree, as they do in adjacent datasets according to Figure 1. For either value of α, as one moves further from the diagonal, one notices a cascading decrease in common clusters across data sets as expected. One also observes larger numbers in b) as compared to a), due to the influence of higher values of α diluting the cluster label information sharing across datasets.

a) Using $\alpha = 4.5$ for the misaligned cluster simulation.

	dat1	dat2	dat3	dat4	dat5
dat1	1.000	0.392	0.364	0.230	0.201
dat2	0.392	1.000	0.461	0.272	0.230
dat3	0.364	0.461	1.000	0.314	0.299
dat4	0.230	0.272	0.314	1.000	0.247
dat5	0.201	0.230	0.299	0.247	1.000

b) Using $\alpha = 0.4$ for the misaligned cluster simulation.

	dat1	dat2	dat3	dat4	dat5
dat1	1.000	0.744	0.690	0.643	0.565
dat2	0.744	1.000	0.773	0.717	0.640
dat3	0.690	0.773	1.000	0.838	0.745
dat4	0.643	0.717	0.838	1.000	0.869
dat5	0.565	0.640	0.745	0.869	1.000

Table 2: The between-dataset cluster correspondence from our breast cancer data analysis. Subfigures (a) and (b) come from two MCMC chains. The interpretation of an element in either table is the average cluster overlap in the corresponding pair of datasets. We note that there is similar cluster overlap between all pairwise comparisons of genomic platforms.

a)

	Expr	CN	Methyl
Expr	1.000	0.158	0.193
CN	0.158	1.000	0.218
Methyl	0.193	0.218	1.000

b)

	Expr	CN	Methyl
Expr	1.000	0.152	0.209
CN	0.152	1.000	0.144
Methyl	0.209	0.144	1.000

Table 3: Post-processed cluster labels for different feature draws from the expression dataset.

	clust 1	clust 2	clust 3	clust 4	clust 5	clust unknown	Sum
clust 1	2	52	0	6	10	15	85
clust 2	0	0	0	0	0	25	25
clust 3	4	0	0	102	7	4	117
clust 4	0	1	36	1	6	0	44
clust 5	24	9	3	1	15	2	54
clust unknown	3	2	2	4	13	21	45
Sum	33	64	41	114	51	67	370

7 Figures

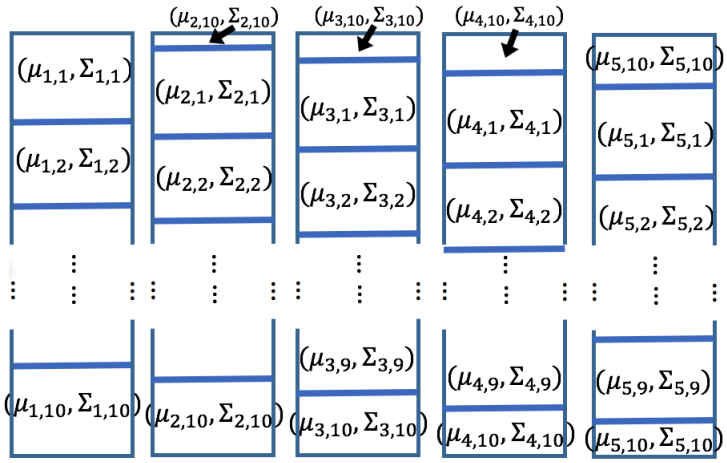


Figure 1: Diagram of progressively misaligned clusters simulation that results in the half-matrix cluster probabilities shown in Table 1. The five long rectangles represent the five generated datasets, each with 200 observations, 10 features, and 10 equally-sized clusters within. Only the first and last few of the 10 clusters can be depicted with boxes in each data set, and the parameters associated with the boxes are those indexing the multivariate normal distributions from which observations in each cluster are sampled. The Σ variance parameters are identical.

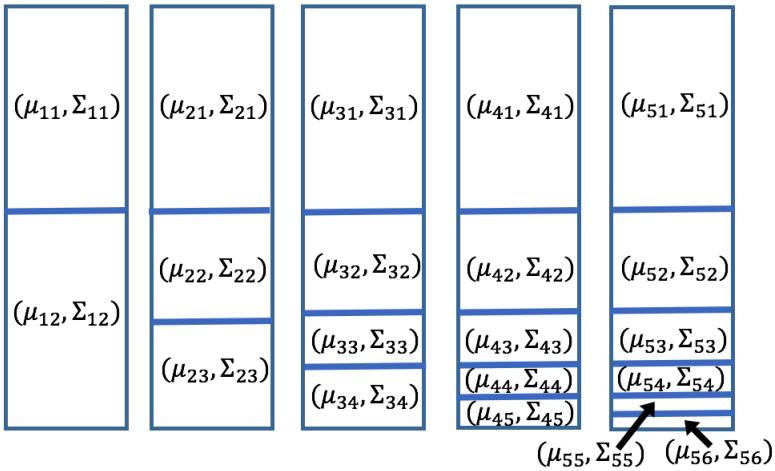


Figure 2: Diagram of nested clusters simulation. The five long rectangles represent datasets, each with 200 observations and 10 features, and boxes within them the unique clusters composing the datasets. The parameters in each box index the multivariate normal distributions used to generate samples in each cluster. As one moves from left to right along datasets, one new cluster is added relative to the previous dataset by halving the bottommost cluster in that dataset. The Σ variance parameters are equivalent.

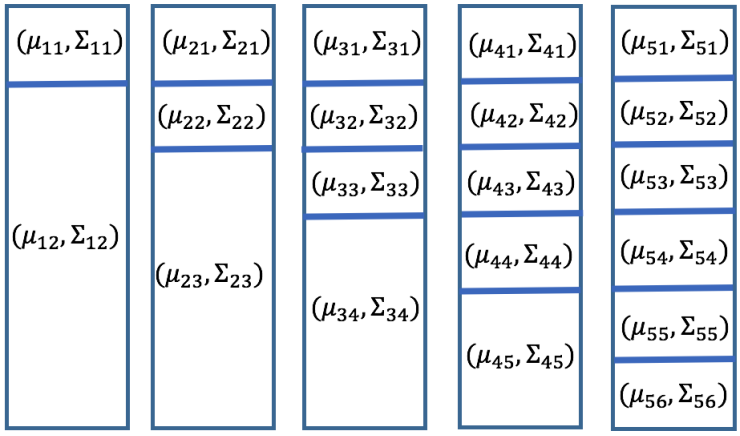
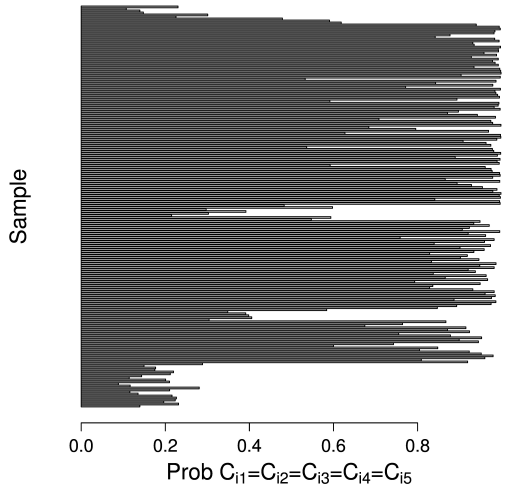
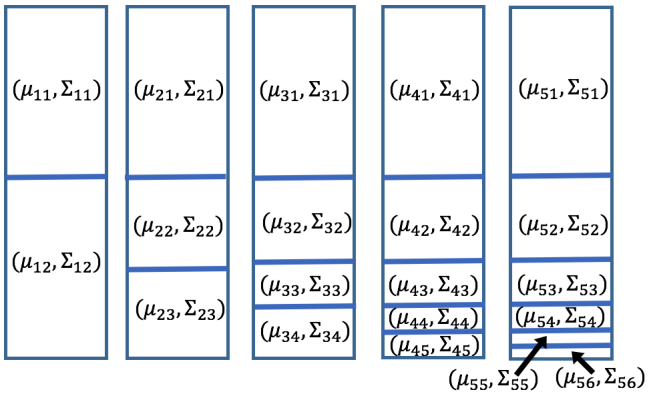


Figure 3: Diagram of a clusterings in the nested clusters simulation. The five long rectangles represent datasets, each with 200 observations and 10 features, and boxes within them the unique clusters composing the datasets. The parameters in each box index the multivariate normal distributions used to generate samples in each cluster. As one moves from left to right along datasets, one new cluster is added relative to the previous dataset. The Σ variance parameters are equivalent.

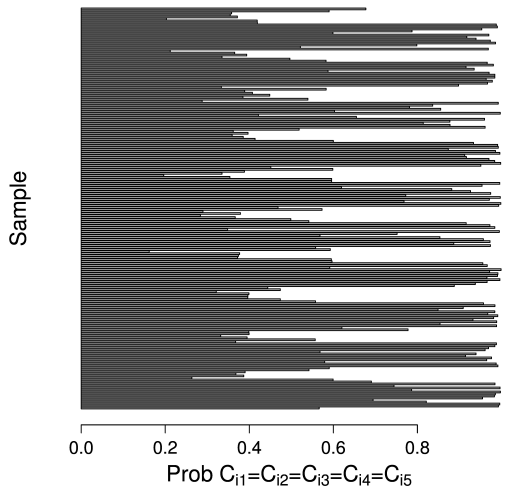


(a) Sample-specific cluster correspondence values, plotted as bars in this horizontal plot. This plot roughly aligns with the cluster boundary diagram in (b)

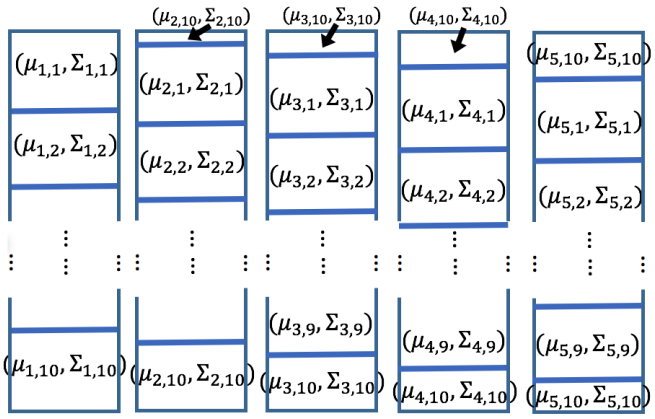


(b) Cluster boundary diagram as in Figure 2, inserted here to show alignment with sample-specific cluster correspondence plotted as a horizontal barplot in (a)

Figure 4: Sample-specific cluster correspondence for the telescoping nestedness simulation

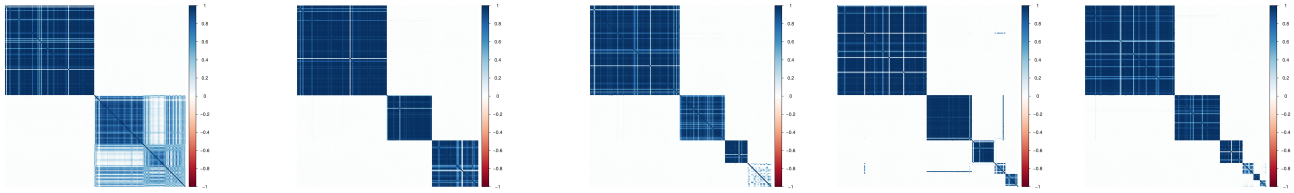


(a) Sample-specific cluster correspondence values, plotted as bars in this horizontal plot. This plot roughly aligns with the cluster boundary diagram in (b)



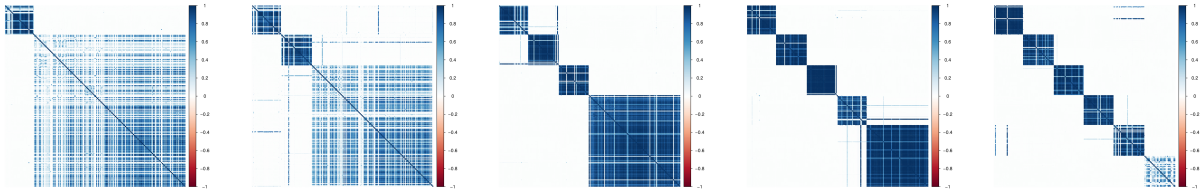
(b) Cluster boundary diagram as in Figure 1, inserted here to show alignment with sample-specific cluster correspondence plotted as a horizontal barplot in (a)

Figure 5: Sample-specific cluster correspondence for the misaligned cluster simulation



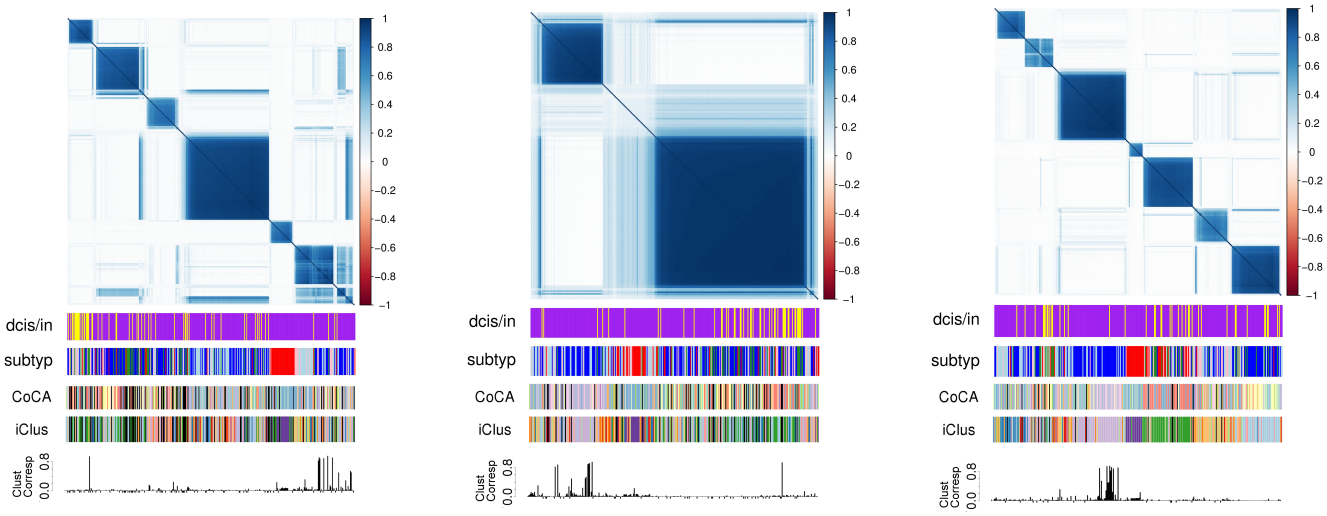
(a) Dataset 1 (b) Dataset 2 (c) Dataset 3 (d) Dataset 4 (e) Dataset 5

Figure 6: Metric 1 for the 5 datasets from the first nested cluster simulation, whose diagram is Figure 2. The heatmap depiction of the metric makes clear that simulated clusterings are generally identified with high fidelity. The last clusters in (a) and (c) are not identified as clearly, likely related to fracturing of clusters in adjacent datasets.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3 (d) Dataset 4 (e) Dataset 5

Figure 7: Metric 1 for the 5 datasets from the second nested cluster simulation, whose diagram is Figure 3. The heatmap depiction of the diagnostic shows that there is some difficulty identifying the largest clusters in (a) and (b), and the bottommost cluster in (e). At least in the case of the large clusters, this is likely due to cluster division in adjacent datasets per the simulation.



(a) Expression dataset (b) CNA dataset (c) Methylation dataset

Figure 8: Metric 1 for our breast cancer analysis. There is a high degree of alignment between the known breast cancer subtypes and those clusters shown in (a). CNAs (b) and Methylation (c) reveal 2 and 7 distinct clusters, respectively. Encoded in our model is cross-dataset cluster assignment influence, though in this analysis interestingly there is relatively little correspondence across datasets. The figures are annotated intrinsic subtype, DCIS/invasive tumor state, and sample-specific cluster correspondence (negative values indicate missing values in the sample in some sources). We also include results from the iCluster and COCA analyses, described in the Supplementary Material.

Supplemental Materials: TWL model

David Swanson, Tonje Lien, Helga Bergholtz, Therese Sørlie, Arnoldo Frigessi

S1.1 Gibbs Sampler implementation

S1.1.1 Tilting the log likelihood for unoccupied clusters

We begin now to describe aspects of the Gibbs Sampler implementation of our model. There was practical use for choosing default values of the log-likelihood for unoccupied clusters or those with only 1 observation. The suggested priors on the cluster means and the use of global variance parameters lead to full conditionals for moves to unoccupied clusters. However such a model favors observations whose feature vectors are close to the mean feature vector across all samples to be sampled into unoccupied clusters. Samples more distant from the mean feature vector across all samples are less likely to be sampled into unoccupied clusters, and when such samples are, the prior on the cluster mean parameters will exert relatively greater influence on the resulting posterior. As a result, sampling still more distant samples into the unoccupied cluster is rendered less likely.

We calculate what the log-likelihood of the Gaussian mixture model would be if in truth samples were generated from M clusters, but we performed random cluster assignment of them. That is, they are assigned a label not corresponding to the one from which they come, but any random integer in $\{1, \ldots, M\}$, and cluster mean and variance parameters are calculated assuming that random cluster assignment. One can think of this random assignment as that which would be "observed" in the absence of any information or as a first guess in an iterative or sampling fitting procedure. Suppose the marginal variance (i.e., the vector of variances of features across all samples in a data set irrespective of cluster assignment) is σ^2 and the assumed within-cluster and between-cluster variances are $c \cdot \sigma^2$ and $(1-c) \cdot \sigma^2$ for some $c \in (0, 1)$. (One must choose c, and 0.5 is suggested as a conservative choice. In practice, the calculated log likelihood is relatively insensitive to choice of c.)

Consider $E_{f_{\mu_2}}[E_{f_{\mu_1}}[E_g[log\, h(X)|\mu_1, \mu_2]|\mu_1]]$ where μ_1 and μ_2 are treated as random variables and represent the mean values at which the Gaussian mixture model likelihood is evaluated (i.e., where cluster-specific mean parameters assume the random cluster assignment) and from which the observation is generated (i.e., the mean parameter of the true cluster), respectively. This is

$$\int \left(\int \left(\int log(h(x))\, g(x)\, dx \right) f_{\mu_1}(\mu_1)\, d\mu_1 \right) f_{\mu_2}(\mu_2)\, d\mu_2$$

where h is a Gaussian pdf used to evaluate our samples with randomly assigned cluster labels. It is indexed by mean parameter μ_2 and variance σ^2. The Gaussian density g is the sample's true density, and f_{μ_1}, and f_{μ_2} are also Gaussian pdfs for the means of g and h, respectively. So informally, we are evaluating the log likelihood with h, which corresponds to the random (or "observed") cluster assignment, according to the density of the truth, g, where what determines truth is the indexing of these densities by μ_1 or μ_2. Additionally, g, f_{μ_1}, and f_{μ_2} are all indexed by variance parameter σ^2 and by mean parameters μ_1, 0, and 0, respectively (the 0's chosen arbitrarily since the result holds so long as these numbers are equal). So $h = h(\cdot|\, \mu_2, c \cdot \sigma^2)$, $g = g(\cdot|\, \mu_1, c \cdot \sigma^2)$, $f_{\mu_1} = f_{\mu_1}(\cdot|\, 0, (1-c) \cdot \sigma^2)$, and $f_{\mu_2} = f_{\mu_2}(\cdot|\, 0, (1-c) \cdot \sigma^2)$. We expand the expression above to

$$= \int \left(\int \left(\int -1/2 \cdot \log C - \frac{(x-\mu_2)^2}{2 \cdot c\, \sigma^2}\, g(x)\, dx \right) f_{\mu_1}(\mu_1)\, d\mu_1 \right) f_{\mu_2}(\mu_2)\, d\mu_2$$

where $C = 2\,\pi\sigma^2$, then have

$$= -1/2 \cdot \log C - 1/2(c) - \int \left(\int \frac{(\mu_1 - \mu_2)^2}{2 \cdot c\,\sigma^2} \, f_{\mu_1}(\mu_1)\, d\mu_1 \right) f_{\mu_2}(\mu_2)\, d\mu_2$$

$$= -1/2 \cdot \log C - 1/2(c) - \int \left(\int \left(\frac{\mu_1^2}{2 \cdot c\,\sigma^2} + \frac{\mu_2^2}{2 \cdot c\,\sigma^2} \right) f_{\mu_1}(\mu_1)\, d\mu_1 \right) f_{\mu_2}(\mu_2)\, d\mu_2$$

$$= -1/2 \cdot \log C - 1/2 \cdot (c + 2\,(1 - c))$$

This formula has been confirmed in simulation. We choose $c = 0.5$ though note that since $\log C$ dominates the expression, choice of c is not critical. While ad hoc, we found that using this formula for the default likelihood effectively allows for the repopulation of clusters once having become unoccupied, maintains cluster sparsity, and does not favor repopulating unoccupied clusters with observations closer to the global mean.

S1.1.2 Modified densities

Sampling from the conditional posterior for C_{ij} shown in Eqn. (1) will sometimes be dominated by one of the three terms of each normalized multinomial category. This is especially true for the likelihood model–the higher the dimension of the data, the more peaked one cluster likelihood score will tend to be relative to all others. And since the Gaussian model has thin tails, the corresponding log likelihood can become very small without lower bound for many clusters. This is unlike the priors on p_j and ρ_i, which effectively have thick "tails" because of hyperparameters α and β giving non-trivial probability mass to sampling even unoccupied clusters. As a result, even in settings where clusterings in other datasets are similar, the model will not tend to sample similar cluster labels across them. Instead, when sampling the posterior one will quickly find and stay near local maxima, and there will be little or no mixing of cluster assignments as the high-dimensional likelihood model renders the probability of sampling most cluster labels zero.

Since a primary feature of our model is the way in which cluster assignment information is shared across datasets and posterior assessment of across-dataset "cluster correspondence", this characteristic renders the basic model of little practical use. We can address this concern by choosing to lower bound the likelihood model, effectively censoring values that fall below the threshold to that value. One can choose the threshold as a function of the ceiling on the number of clusters (ie, K), and the minimum cluster label mixing desired per MCMC sample.

One can also interpret this strategy as a question of data granularity. If an observation Y_{ij} is beyond some distance from cluster center $\mu_j^{(k)}$, we set the value to what the density evaluates to at that distance. This approach is more helpful for achieving adequate cluster label mixing than use of thick-tailed densities, which again get stuck in local maxima, but only after a slightly larger number of iterations. Indeed, when notions of distance between cluster centers is only significant up to some threshold, use of modified densities is fitting. It is important to notice that use of such densities does not dull posterior peaks, whose functional form remains unaffected by lower bounding the tails. Our data analysis in Section 3 demonstrates that clusters still remain highly visible with use of these modified densities. One can also simply run an MCMC longer and examine statistical significance in a PSM if there is concern about slightly less visible clusters. We also experimented with simulated annealing to address the problem of the likelihood dominating posterior sampling, but found more robust results using modified densities.

S1.2 Comparison with Cluster of Cluster Assignments and iCluster data integreation methods

To compare our TWL model with those assuming a set of common boundaries across all data types, we fit iCluster and COCA models on our data [S18, S7]. We fit the iCluster model with a total of 10 clusters in keeping with previously used number of subtypes in breast cancer based on multitype data ([S32]). Examination of BIC additionally suggested this was an adequate number of clusters to represent the data. Consistent with previous use of the COCA procedure, we first performed non-negative matrix factorization (NMF) on each data type and chose the number of clusters in each one suggested by our heatmaps (5,2, and 7 for Expression, CN, and Methylation, respectively). We then used the weighting matrices from NMF as indicator matrices for subtype and specified a total of 12 consensus clusters using the *ConsensusClusterPlus* package in R [S39]. Because both iCluster and COCA require all datatypes for each sample, contrary to TWL, we could only analyze 299 of our total of 370 samples.

iCluster and COCA clusterings are shown in Figures 8a, 8b, and 8c as "barcode" annotation beneath the PSM heatmaps. Colors in the barcodes map to arbitrary cluster labels and were chosen to maximize color contrast in the annotation. Patients missing at least one datatype and therefore lacking a consensus cluster are denoted with black. Since the output of both methods is a single, consensus cluster, the different barcodes under each datatype are reorderings of the consensus cluster based on sample ordering of the heatmap.

We see considerable agreement between iCluster and COCA labels and clear association with certain heatmap clusters: the enrichment in iCluster and COCA cluster labels under the methylation heatmap clusters and the Basal expression heatmap cluster (denoted by red in the "subtyp" barcode annotation) are particularly striking. However, for other expression clusters and particularly the CNA datatype as a whole, the consensus clusterings of iCluster and COCA show little consistency with datatype clusters as revealed in the heatmaps. This phenomenon likely results from the models' attempts to find consistency across datatype clusterings while there is little. For COCA in particular, its greater association with the methylation heatmap clustering may come in part from the larger number of clusters in that datatype and resultant greater influence on the consensus clustering. Ultimately there seems to be much loss of datatype-specific clustering information by fitting a single consensus cluster for each sample.

S1.2.1 Label switching

Label switching due to posterior symmetry in label permutations is not a concern in our model [S44]. Cluster label is never of direct importance in post-processing our posterior – we only examine pair-wise common cluster assignment of observations, generating a dataset-specific $N \times N$ indicator matrix for each iteration in the MCMC chain after convergence, called posterior similarity matrix (PSM) [S45]. A "1" in the $(v,t)-th$ element of the indicator matrix associated with dataset j and iteration m indicates observations v and t in dataset j have a common cluster assignment, say "12", for iteration m. We can additionally average over all iterations for each dataset to generate a correlation matrix and use either hierarchical clustering or graphical lasso to define definite, fixed clusters, if one wants [S46, S47]. In practice, we have used hierarchical clustering for computational ease and subsequent flexibility of defining the number of clusters by simply "cutting" the dendrogram at different heights. We describe these post-processes more in Section 3.2.

We also recommend viewing the correlation matrix associated with each data set as a final product – the information loss associated with thresholding for definite cluster assignment can

negatively impact more subtle conclusions from analysis. Ambiguity in cluster assignment made evident in the matrix is relevant information, and alignment or its lack with clusters found in other datasets can be unstable when thresholding rules are used to define object-dataset cluster membership.

S1.3 The General Two-Way Latent Structure Model

Here we present our model in the important situation where samples have measurements only in some of the data sets. A_{ij} is the subject id associated with row i in dataset j. It is only annotation, or metadata, and serves as the index to which we apply multinomial models across datasets, within sample. If all data sources have identical sets of sample ids, the notion of A_{ij} is unnecessary and could be reduced to i since rows in each data set could be associated with the same sample if ordered properly. It is because we seek to fit our model in settings where some subjects only have a subset of data sources that we introduce A_{ij} notation. We assume $A_{ij} \in \{id_1, \ldots, id_n, \ldots, id_N\}$, the superset of ids over all datasets

Let j be the dataset index, and we assume $j \in \{1, \ldots, J\}$, for a total of J data sources.

$Y_{i,j}$ is a vector of features for id A_{ij} and data set j and of dimension d_j. Not all sample-dataset pairs (A_{ij}, j) need exist if a certain data source is not available for sample id_n. Assume that the cardinality of the ids in data set j is N_j. So $\forall\ j$, $N_j \leq N$, N again the size of the superset of ids. Dataset j therefore consists of N_j observation vectors $Y_{A_{ij},j}$ of length d_j.

Let $\mathbf{Y}$ be the set of all observation vectors $\{Y_{i,j}\}$ for all valid pairs of (i, j)

We seek to sample from the posterior:

$$P(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C} \,|\, \mathbf{Y}, \mathbf{C} == \mathbf{R})$$

where we now explain $\boldsymbol{\tau}$, $\mathbf{p}$, $\boldsymbol{\rho}$, $\mathbf{R}$, $\mathbf{C}$

Column (ie, data source) clusterings are defined with the random vector

$$\mathbf{C} \equiv (C_{1,1}, \ldots, C_{i,j}, \ldots, C_{N,J})$$

where J again is the total number of data sets, and

$$C_{(i,j)} \sim Multinom(p_j^{(1)}, p_j^{(2)} \ldots, p_j^{(K)}) \quad \forall \text{ valid } (i, j)$$

where K is the fixed upper bound on the number of clusters. The interpretation of p_j^k is the probability of a draw of cluster label k in dataset j. So the $1, 2, \ldots, K$ in parentheses here are superscripts, not exponents, and similarly for the model on the $R_{i,j}$'s below.

We emphasize that though C is subscripted by i and j, the parameters are only subscripted by j. Therefore, all observations within dataset j draw from this dataset-specific j^{th} multinomial model. We can loosely think of the cluster models on the j's as models on the *Columns* of the aggregated datasets if we arranged them next to one another.

Row (ie, sample id) clusterings are defined with the random vector

$$\mathbf{R} \equiv (R_{1,1}, \ldots, R_{i,j}, \ldots, R_{N,J})$$

Assuming that $A_{ij} = id_n$, we have

$$R_{(i,j)} \sim Multinom(\rho_{id_n}^{(1)}, \rho_{id_n}^{(2)} \ldots, \rho_{id_n}^{(K)}) \quad \forall \text{ valid } (i, j)$$

with K the upper bound on the number of clusters. Note that though R is subscripted by i and j, the parameters are only subsripted by id_n, subject id annotation. We can loosely think of the cluster models on the id_n's as models on the *Rows* of the aggregated data sources if we arranged them next to one another.

The model for $\mathbf{p_j}$ is

$$\mathbf{p_j} \sim Dirichlet(\beta_1, \ldots, \beta_K) \ \ \forall \ \ j$$

And for $\boldsymbol{\rho}_{id_n}$,

$$\boldsymbol{\rho}_{id_n} \sim Dirichlet(\alpha_1, \ldots, \alpha_K) \ \ \forall \ \ id_n$$

with hyperparameters α and β constant across id_n's and j's, respectively. $\boldsymbol{\alpha}$ should likely be chosen as a function of the average number of data sets per id (or simply number of data sets if all ids are present in all data sets and unique within them), and $\boldsymbol{\beta}$ as a function of the average number unique ids within each data set (or simply number of unique ids if again all ids are present in all data sets and unique within them). As a result, and as follows, we have

$$\alpha \equiv \alpha_1 = \alpha_2 = \cdots = \alpha_K$$

and

$$\beta \equiv \beta_1 = \beta_2 = \cdots = \beta_K$$

The parameters of the normal densities are defined with:

$$\boldsymbol{\mu} \equiv (\mu_{1,1} \ldots \mu_{K,1}, \ldots\ldots, \mu_{1,j} \ldots \mu_{K,j}, \ldots\ldots, \mu_{1,J} \ldots \mu_{K,J})$$

where $\mu_{k,j}$ is the mean vector and of dimension d_j, d_j again the number of features in data source j.

$$\boldsymbol{\tau} \equiv (\tau_{1,1} \ldots \tau_{K,1}, \ldots\ldots, \tau_{1,j} \ldots \tau_{K,j}, \ldots\ldots, \tau_{1,J} \ldots \tau_{K,J})$$

where $\tau_{k,j}$ is the precision vector and of dimension d_j. Independence is assumed for components of vector $Y_{i,j}$ and as a result information of its precision matrix is contained in a d_j dimensional vecor. The assumption increases computation efficiency, and filtering highly correlated features makes the assumption reasonable.

$\mathbf{C} = \mathbf{R}$ is shorthand notation for the event $\cup_{\{i,j\}} C_{i,j} = R_{i,j}$; i.e., for all valid pairs of (i, j), $C_{i,j} = R_{i,j}$ is true.

We factorize the posterior to the following and consider each factor in turn in the next sections.

$$P(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C} \,|\, \mathbf{Y}, \mathbf{C} = \mathbf{R}) \propto$$

$$P(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R}) \cdot P(\boldsymbol{\mu}, \boldsymbol{\tau} | \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R})$$

$$\cdot P(\mathbf{C}, \mathbf{R} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} = \mathbf{R}) \cdot P(\mathbf{p}, \boldsymbol{\rho} \,|\, \mathbf{C} = \mathbf{R})$$

S1.4 Considering Y

In the expressions below we have define $k^* \equiv C_{i,j}$ and use k^* for cleaner notation to avoid double subscripting. It functions as shorthand for subscripting according to the $C_{i,j}$ cluster label. We use $\{i,j\}$ to denote the set of valid (i,j) pairs.

$$P(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R})$$

$$= \prod_{\{i,j\}} P(Y_{i,j} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R})$$

$$= \prod_{\{i,j\}} P(Y_{i,j} \mid \mu_{k^*,j}, \tau_{k^*,j}, C_{i,j}, R_{i,j}, C_{i,j} = R_{i,j}) = \prod_{\{i,j\}} P(Y_{i,j} \mid \mu_{k^*,j}, \tau_{k^*,j}, C_{i,j})$$

$$= \prod_{\{i,j\}} P(Y_{i,j} \mid \mu_{k^*,j}, \tau_{k^*,j}) = \prod_{\{i,j\}} N(Y_{i,j} \mid \mu_{k^*,j}, \tau_{k^*,j})$$

where $N(\cdot)$ is used to denote the Gaussian density here and below. Each (i,j) term in this last expression can be written as $\prod_{k=1}^{K} N(Y_{i,j} \mid \mu_{k,j}, \tau_{k,j})^{I(k=C_{i,j})}$. For the purpose of combining terms when calculating posteriors on the $C_{i,j}$'s, we write the likelihood as

$$= \prod_{\{i,j\}} \prod_{k=1}^{K} N(Y_{i,j} \mid \mu_{k,j}, \tau_{k,j})^{I(k=C_{i,j})}$$

Notice we arbitrarily dropped the $R_{i,j}$ because it contains redundant information on $C_{i,j}$. The choice is not entirely arbitrary however because $C_{i,j}$ can be considered in some sense "more primary" for notation since $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ posteriors are sampled based only on within dataset (or loosely, column) cluster assignment. This is also evident in our choice of $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ subscripts.

S1.5 Considering C, R

Now consider

$$P(\mathbf{C}, \mathbf{R} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} = \mathbf{R})$$

$$= \prod_{\{i,j\}} P(C_{i,j}, R_{i,j} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} = \mathbf{R}) = \prod_{\{i,j\}} P(C_{i,j} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} = \mathbf{R})$$

$$= \prod_{\{i,j\}} \left(p_j^{(C_{i,j})} \cdot \rho_{A_{ij}}^{(C_{i,j})} / \left(\sum_k p_j^{(k)} \cdot \rho_{A_{ij}}^{(k)} \right) \right)$$

where again we drop $R_{i.j}$ to emphasize that it contains redundant information. The $C_{i,j}$'s and k's in this expression are superscripts for the cluster label and not exponents.

S1.6 Considering $\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\psi}$

We now examine the term

$$P(\boldsymbol{\mu}, \boldsymbol{\tau} \,|\, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R})$$

$$= \prod_k \prod_j P(\mu_{k,j}, \tau_{k,j} \,|\, \mathbf{p}, \boldsymbol{\rho}, \mathbf{R}, \mathbf{C}, \mathbf{C} = \mathbf{R})$$

$$= \prod_k \prod_j N(\mu_{k,j} \,|\, \hat{\tau}_{k,j}, \mu_{0,j}, \psi_{0,j})$$

$$= \prod_k \prod_j N(\mu_{k,j} \,|\, \hat{\tau}_{k,j}, \mu_{0.j}, \psi_{0,j})$$

The mean and precision of $\mu_{k,j}$ are $\mu_{0,j}$ and $\psi_{0.j}$, respectively, where we set $\mu_{0,j}$ to the global mean vector of $Y_{\cdot,j}$, which is data set j. We set the corresponding precision parameter

$$\psi_{0,j} = \left(\frac{1}{2} \cdot \frac{\sigma^2_{G,j}}{n/20} \cdot 15 \right)^{-1}$$

for all j, where $\sigma^2_{G,j}$ is a vector of variances of the features of $Y_{\cdot,j}$. Setting $\psi_{0,j}$ in this way makes the assumption that half of global variation in data set j, $\sigma^2_{G,j}$, is attributable within-cluster, that the size of clusters will be 1/20 of the total sample size, n, and that there should be 1/15 cluster mean shrinkage to the global mean under such circumstances. If the analyst fits the model with an upper bound of more than 20 clusters, this will result in greater mean shrinkage earlier in the MCMC chain since the chain begins with random cluster assignment. If in truth there are fewer than 20 clusters in the data, will result in less shrinkage of the respective cluster means once convergence has occurred.

Such a prior also serves a practical purpose: without it, small clusters, especially those with one observation, will be resistant to becoming unoccupied since the mean parameter, $\mu_{k,j}$, will tend to be close or identical to the corresponding observation(s) and thus inflate the likelihood. This is more true if additionally cluster variance parameters, $\tau_{k,j}$'s, are cluster-specific, as estimated normal densities will diverge to infinity. By shrinking $\mu_{k,j}$ to a global mean, especially when the cluster size is smaller, the posterior of

$$P(\mathbf{C}, \mathbf{R} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} == \mathbf{R}) = P(\mathbf{C} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} == \mathbf{R})$$

will exhibit greater cluster sparsity.

Assuming a common variance parameter across clusters similarly keeps likelihoods from diverging and allows for more efficient estimation in the parameter. While addressing such a concern with a prior on the variance is also a reasonable choice, we found greater statistical efficiency and convergence by considering it fixed and common across all clusters. The parameter is estimated wth maximum likelihood. While the assumption of common cluster variance across clusters is unlikely to hold exactly, new clusters will form when a single one is too heterogeneous to be consistent with the common variance parameter.

Combining these priors with the likelihood given above, we can calculate posteriors for $\boldsymbol{\mu}$ and estimate $\tau_{1,j} = \tau_{2,j} = \cdots = \tau_{k,j} = \cdots = \tau_{K,j} \;\; \forall \;\; k$ via maximum likelihood, which we will denote $\widehat{\boldsymbol{\tau}}_{\cdot,\boldsymbol{j}}$. So $\widehat{\boldsymbol{\tau}}_{\cdot,\boldsymbol{j}}$ is a vector of length the number of features of data set j invariant to cluster label, and which is specific to each data set j.

Because of conjugacy relationships, the posterior for $\mu_{k,j}$ is then

$$P(\mu_{k,j} \,|\, \mathbf{Y}_{\cdot,\mathbf{j}}, \boldsymbol{\mu}_{0,j}, \hat{\boldsymbol{\tau}}_{\cdot,\boldsymbol{j}}, \boldsymbol{\psi}_{\mathbf{0},\boldsymbol{j}}) = N\left(\mu_{k,j} \,\middle|\, \left[n\,\widehat{\tau}_{\cdot,j}\cdot\bar{\mathbf{Y}}_{\cdot,\mathbf{j}}+\psi_{0,j}\cdot\mu_{0,j}\right]\cdot\left[\psi_{0,j}+n\,\widehat{\tau}_{\cdot,j}\right]^{-1}, \left[\psi_{0,j}+n\,\widehat{\tau}_{\cdot,j}\right]^{-1}\right)$$

parametrized as the mean and variance, not precision. The mean is seen to be a weighted average of the sample mean of the cluster and the prior, according to the respective precision parameters. Likewise, the variance is the inverse of the sum of the precision parameters. Both formula are familiar posterior parameters when using normal-normal conjugacy with variance $(1/\tau)$ considered fixed.

S1.7 Considering $\boldsymbol{p}, \boldsymbol{\rho}$

Lastly, we consider

$$P(\mathbf{p}, \boldsymbol{\rho} \,|\, \mathbf{C} = \mathbf{R})$$

$$= \prod_i \prod_j P(p_j, \rho_{A_{ij}} \,|\, \mathbf{C} = \mathbf{R}) \propto P(\mathbf{C} = \mathbf{R} \,|\, p_j, \rho_{A_{ij}}) \cdot \prod_i \prod_j P(p_j, \rho_{A_{ij}})$$

$$= \Big(\prod_{\{i,j\}} \sum_k p^(k)_j \cdot \rho^(k)_{A_{ij}}\Big) \cdot \Big(\prod_i \prod_j \big(p_j^{(1)}\, p_j^{(2)}\, p_j^{(3)} \dots p_j^{(K)}\big)^{\alpha} \cdot (\rho_{A_{ij}}^{(1)}\, \rho_{A_{ij}}^{(2)} \dots \rho_{A_{ij}}^{(K)})^{\beta}\Big)$$

where $1, 2, \dots, k, \dots, K$ denotes superscripts, not exponents, and we make use of identical hyperparameters across the prior cluster probabilities of α and β by taking each one as an exponent outside the respective parentheses.

We can collect terms for

$$P(\mathbf{C}, \mathbf{R} \mid \mathbf{p}, \boldsymbol{\rho}, \mathbf{C} = \mathbf{R}) \cdot P(\mathbf{p}, \boldsymbol{\rho} \,|\, \mathbf{C} = \mathbf{R})$$

and obtain

$$\propto \Big(\prod_{\{i,j\}} p_j^{(C_{i,j})} \cdot \rho_{A_{ij}}^{(C_{i,j})} / \Big(\sum_k p_j^{(k)} \cdot \rho_{A_{ij}}^{(k)}\Big)\Big) \cdot$$

$$\Big(\prod_{\{i,j\}} \sum_k p^(k)_j \cdot \rho^(k)_{A_{ij}}\Big) \cdot \Big(\prod_i \prod_j \big(p_j^{(1)}\, p_j^{(2)}\, p_j^{(3)} \dots p_j^{(K)}\big)^{\alpha} \cdot (\rho_{A_{ij}}^{(1)}\, \rho_{A_{ij}}^{(2)} \dots \rho_{A_{ij}}^{(K)})^{\beta}\Big)$$

$$= \Big(\prod_{\{i,j\}} p_j^{(C_{i,j})} \cdot \rho_{A_{ij}}^{(C_{i,j})}\Big) \cdot \Big(\prod_i \prod_j \big(p_j^{(1)}\, p_j^{(2)}\, p_j^{(3)} \dots p_j^{(K)}\big)^{\alpha} \cdot (\rho_{A_{ij}}^{(1)}\, \rho_{A_{ij}}^{(2)} \dots \rho_{A_{ij}}^{(K)})^{\beta}\Big)$$

$$= \prod_j \prod_{i|j} \Big((p_j^{(1)})^{\alpha+\sum_{i|j} I(C_{i,j}=1)} \cdot (p_j^{(2)})^{\alpha+\sum_{i|j} I(C_{i,j}=2)} \dots (p_j^{(K)})^{\alpha+\sum_{i|j} I(C_{i,j}=K)}\Big) \cdot$$

$$\Big((\rho_{A_{ij}}^{(1)})^{\beta+\sum_{j|i} I(R_{i,j}=1)} \cdot (\rho_{A_{ij}}^{(2)})^{\beta+\sum_{j|i} I(R_{i,j}=2)} \dots (\rho_{A_{ij}}^{(K)})^{\beta+\sum_{j|i} I(R_{i,j}=K)}\Big)$$

where we use the notation $i \,|\, j$ and $j \,|\, i$ to denote valid values of i and j within strata of j and i, respectively.

S1.7.1 Cluster posteriors

Using $C_{i,j} = R_{i,j}$ in the previous expression and incorporating it with the likelihood, we have

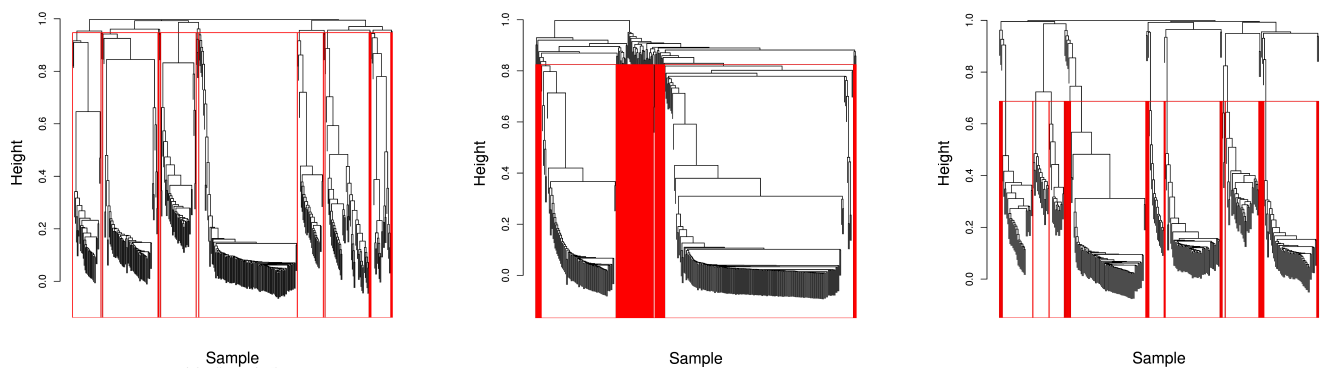
$$= \prod_j \prod_{i|j} \Big((p_j^{(1)})^{\alpha + \sum_{i|j} I(C_{i,j}=1)} \cdot (p_j^{(2)})^{\alpha + \sum_{i|j} I(C_{i,j}=2)} \dots (p_j^{(K)})^{\alpha + \sum_{i|j} I(C_{i,j}=K)} \Big) \cdot$$

$$\Big((\rho_{A_{ij}}^{(1)})^{\beta + \sum_{j|i} I(C_{i,j}=1)} \cdot (\rho_{A_{ij}}^{(2)})^{\beta + \sum_{j|i} I(C_{i,j}=2)} \dots (\rho_{A_{ij}}^{(K)})^{\beta + \sum_{j|i} I(C_{i,j}=K)} \Big)$$

$$\Big(N(Y_{i,j} \mid \mu_{k,j}, \tau_{k,j})^{I(C_{i,j}=1)} \cdot N(Y_{i,j} \mid \mu_{k,j}, \tau_{k,j})^{I(C_{i,j}=2)} \dots N(Y_{i,j} \mid \mu_{k,j}, \tau_{k,j})^{I(C_{i,j}=K)}$$

which admits the kernel of a multinomial probability mass function in $C_{i,j}$.

S1.8 Supplementary Figures



(a) Expression dataset (b) CNA dataset (c) Methylation dataset

Figure S1: Post-processed hierarchical clustering for our breast cancer analysis. Hierarchical clusters generally align well with the clearly identifiable clusters in Figure 8

Table S1: Post-processed cluster labels for different feature draws from the expression dataset. We see an even higher degree of stability in methylation cluster labels across feature draws.

	clus_1	clus_2	clus_3	clus_4	clus_5	clus_6	clus_7	clus_8	clus_unknown	Sum
clus_1	0	16	13	0	0	0	0	0	0	29
clus_2	2	0	1	0	0	1	0	34	0	38
clus_3	0	0	0	54	0	0	0	3	0	57
clus_4	0	0	0	6	0	0	19	2	0	27
clus_5	6	0	0	3	0	0	1	0	3	13
clus_6	13	0	0	0	0	0	0	0	0	13
clus_7	9	0	0	1	0	3	7	2	0	22
clus_8	0	0	0	0	8	45	0	0	1	54
clus_unknown	0	0	1	10	7	5	6	11	21	61
Sum	30	16	15	74	15	54	33	52	25	314

Table S2: Post-processed cluster labels from parallel MCMC chains for the expression dataset. Since cluster labels are arbitrary, one could permute them so that the largest cells would fall along the diagonal. We see stability in cluster labels, indicating model convergence.

	clus_1	clus_2	clus_3	clus_4	clus_5	clus_unknown	Sum
clus_1	0	0	37	1	11	1	50
clus_2	29	1	1	1	9	1	42
clus_3	2	57	1	0	2	5	67
clus_4	0	0	0	0	0	28	28
clus_5	0	4	0	109	28	17	158
clus_unknown	2	2	2	3	1	15	25
Sum	33	64	41	114	51	67	370

Table S3: Post-processed cluster labels from parallel MCMC chains for the methylation dataset. Since cluster labels are arbitrary, one could permute them so that the largest cells would fall along the diagonal. We see a high degree of stability in cluster labels, indicating model convergence.

	clus_1	clus_2	clus_3	clus_4	clus_5	clus_6	clus_7	clus_8	clus_unknown	Sum
clus_1	0	0	0	0	0	0	32	1	5	38
clus_2	26	0	0	2	0	0	0	0	1	29
clus_3	1	0	0	72	0	0	0	9	0	82
clus_4	2	0	0	0	0	1	0	34	0	37
clus_5	0	0	0	0	2	49	0	0	1	52
clus_6	0	0	15	0	0	2	0	0	0	17
clus_7	0	15	0	0	0	0	0	0	0	15
clus_unknown	1	1	0	0	13	2	1	8	18	44
Sum	30	16	15	74	15	54	33	52	25	314

Table S4: Post-processed cluster labels for the expression dataset tabulated against breast cancer subtypes. There is considerable association between the two sets of labels. Those observations with the more heterogeneous label ("clus_unknown") on the expression data source are more diffusely distributed across the subtypes than those with a clear cluster label.

	Basal	Her2	LumA	LumB	Normal	Sum
clus_4	27	0	0	0	0	27
clus_2	4	23	0	1	4	32
clus_5	4	7	72	62	7	152
clus_1	0	7	20	11	7	45
clus_3	2	2	37	1	23	65
clus_unknown	11	8	9	13	8	49
Sum	48	47	138	88	49	370

Table S5: Post-processed cluster labels for the methylation dataset tabulated against breast cancer subtypes. There is some association between the two sets of labels. There seems to be greater alignment between subtype and expression label than methylation label, which is expected since intrinsic subtype is defined with expression profiles.

	clus_1	clus_2	clus_3	clus_4	clus_5	clus_6	clus_7	clus_unknown	Sum
LumB	13	16	22	3	4	2	3	12	75
Her2	12	3	4	1	12	4	2	5	43
Normal	2	1	3	5	12	7	1	3	34
Basal	1	1	1	1	16	4	0	17	41
LumA	10	8	52	27	8	0	9	7	121
Sum	38	29	82	37	52	17	15	44	314

Table S6: We order observations according to sample-specific cluster correspondence and find enrichment in the invasive tumor state among those observations with greater expression-methylation cluster label sharing for different feature draws from the data (shown in (a) and (b)). The result suggests that there may be some degree of pan-genomic features acting in concert for tumors in a more advanced state.

a)

	DCIS	IDC
Most clust corresp samps	0.024	0.976
Least clust corresp samps	0.048	0.952
Marginal distribution	0.154	0.846

b)

	DCIS	IDC
Most clust corresp samps	0.048	0.952
Least clust corresp samps	0.048	0.952
Marginal distribution	0.154	0.846

Table S7: We order observations according to sample-specific cluster correspondence, and confirm using different feature draws enrichment in the HER2 intrinsic subtype among those observations with greater expression-CNA correspondence.

a)

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.100	0.350	0.250	0.125	0.175
Least clust corresp samps	0.075	0.100	0.350	0.200	0.275
Marginal distribution	0.130	0.127	0.373	0.238	0.132

b)

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.100	0.425	0.075	0.250	0.150
Least clust corresp samps	0.075	0.125	0.350	0.200	0.250
Marginal distribution	0.130	0.127	0.373	0.238	0.132

Table S8: We order observations according to sample-specific cluster correspondence, and find that there is slight enrichment in the HER2 subtype (a) and the LumA subtype (b) among those observations with greater cluster label sharing across the CNA-methylation and expression-methylation pairings, respectively. The result suggests that there is some degree of pan-genomic features acting in greater concert for tumors of the HER2 and LumA intrinsic subtypes.

a) CNA-methylation pairing

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.000	0.267	0.400	0.267	0.067
Least clust corresp samps	0.333	0.133	0.267	0.200	0.067
Marginal distribution	0.130	0.127	0.373	0.238	0.132

b) Expression-methylation pairing

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.017	0.033	0.617	0.283	0.050
Least clust corresp samps	0.117	0.067	0.350	0.217	0.250
Marginal distribution	0.130	0.127	0.373	0.238	0.132

Table S9: Ductal carcinoma is almost entirely found in the large CNA cluster across feature draws from the data (shown in (a) and (b)).

a)

	clus_1	clus_2	clus_3	clus_4	clus_unknown	Sum
DCIS	3	42	0	0	3	48
IDC	62	160	2	2	64	290
Sum	65	202	2	2	67	338

b)

	clus_1	clus_2	clus_3	clus_unknown	Sum
DCIS	3	1	40	4	48
IDC	63	24	129	74	290
Sum	66	25	169	78	338

Table S10: We order observations according to sample-specific cluster correspondence and confirm using different feature draws enrichment in the DCIS tumor state among those observations with greater cluster label sharing across the expression-CNA pairing. The result suggests that there is some degree of expression-CNA features acting in greater concert for tumors that have not yet become invasive.

a)

	DCIS	IDC
Most clust corresp samps	0.350	0.650
Least clust corresp samps	0.250	0.750
Marginal distribution	0.154	0.846

b)

	DCIS	IDC
Most clust corresp samps	0.425	0.575
Least clust corresp samps	0.275	0.725
Marginal distribution	0.154	0.846

Table S11: There is a lack of the Basal intrinsic subtype among those observations with greater CNA-methylation sample-specific cluster correspondence across feature draws.

a)

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.048	0.190	0.238	0.357	0.167
Least clust corresp samps	0.167	0.048	0.310	0.310	0.167
Marginal distribution	0.130	0.127	0.373	0.238	0.132

b)

	Basal	Her2	LumA	LumB	Normal
Most clust corresp samps	0.024	0.214	0.286	0.381	0.095
Least clust corresp samps	0.167	0.048	0.310	0.310	0.167
Marginal distribution	0.130	0.127	0.373	0.238	0.132

Table S12: Cross tabulation of intrinsic subtype and CNA clusters.

	clus_1	clus_2	clus_3	clus_unknown	Sum
LumB	17	15	22	27	81
Her2	5	5	22	11	43
Normal	3	1	35	1	40
Basal	1	1	17	26	45
LumA	40	3	73	13	129
Sum	66	25	169	78	338

Table S13: Cross tabulation of model defined expression subtype and tumor state. We observe significant enrichment in the DCIS tumor state in model defined cluster 1.

	clus_1	clus_2	clus_3	clus_4	clus_5	clus_unknown	Sum
DCIS	22	8	4	19	2	2	57
IDC	11	56	37	95	49	65	313
Sum	33	64	41	114	51	67	370

References

[S1] Claudia Cava, Gloria Bertoli, and Isabella Castiglioni. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. *BMC Systems Biology*, 9(1), December 2015.

[S2] N. Huang, P. K. Shah, and C. Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13(3):305–316, May 2012.

[S3] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, October 2012.

[S4] Nora K. Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, June 2015.

[S5] Vessela N. Kristensen, Ole Christian Lingjærde, Hege G. Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, May 2014.

[S6] H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, Samuel AJR Aparicio, and Carlos Caldas. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology*, 15(8):431, 2014.

[S7] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, Christina Curtis†, Sohrab P. Shah, Carlos Caldas, Samuel Aparicio, James D. Brenton, Ian Ellis, David Huntsman, Sarah Pinder, Arnie Purushotham, Leigh Murphy, Carlos Caldas, Samuel Aparicio, Carlos Caldas, Helen Bardwell, Suet-Feung Chin, Christina Curtis, Zhihao Ding, Stefan Gräf, Linda Jones, Bin Liu, Andy G. Lynch, Irene Papatheodorou, Stephen J. Sammut, Gordon Wishart, Samuel Aparicio, Steven Chia, Karen Gelmon, David Huntsman, Steven McKinney, Caroline Speers, Gulisa Turashvili, Peter Watson, Ian Ellis, Roger Blamey, Andrew Green, Douglas Macmillan, Emad Rakha, Arnie Purushotham, Cheryl Gillett, Anita Grigoriadis, Sarah Pinder, Emanuele di Rinaldis, Andy Tutt, Leigh Murphy, Michelle Parisien, Sandra Troup, Carlos Caldas, Suet-Feung Chin, Derek Chan, Claire Fielding, Ana-Teresa Maia, Sarah McGuire, Michelle Osborne, Sara M. Sayalero, Inmaculada Spiteri, James Hadfield, Samuel Aparicio, Gulisa Turashvili, Lynda Bell, Katie Chow, Nadia Gale, David Huntsman, Maria Kovalik, Ying Ng, Leah Prentice, Carlos Caldas, Simon Tavaré, Christina Curtis, Mark J. Dunning, Stefan Gräf, Andy G. Lynch, Oscar M. Rueda, Roslin Russell, Shamith Samarajiwa, Doug Speed, Florian Markowetz, Yinyin Yuan, James D. Brenton, Samuel Aparicio, Sohrab P. Shah, Ali Bashashati, Gavin Ha, Gholamreza Haffari, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, April 2012.

[S8] Simen Myhre, Ole-Christian Lingjaerde, Bryan T. Hennessy, Miriam R. Aure, Mark S. Carey, Jan Alsner, Trine Tramm, Jens Overgaard, Gordon B. Mills, Anne-Lise Børresen-Dale, and Therese Sørlie. Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Molecular Oncology*, 7(3):704–718, June 2013.

[S9] David J Reiss, Nitin S Baliga, and Richard Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, page 22, 2006.

[S10] Wei Sun, Paul Bunn, Chong Jin, Paul Little, Vasyl Zhabotynsky, Charles M Perou, David Neil Hayes, Mengjie Chen, and Dan-Yu Lin. The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Research*, February 2018.

[S11] Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt Van De Rijn, and Stefanie S. Jeffrey. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[S12] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, James Stephen Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, and Stephanie Geisler. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003.

[S13] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Noreen Dhalla, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, and Nam H. Pho. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, September 2012.

[S14] Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E. Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, August 2014.

[S15] T. A. Yap, M. Gerlinger, P. A. Futreal, L. Pusztai, and C. Swanton. Intratumor Heterogeneity: Seeing the Wood for the Trees. *Science Translational Medicine*, 4(127):127ps10–127ps10, March 2012.

[S16] So Yeon Park, Mithat Gönen, Hee Jung Kim, Franziska Michor, and Kornelia Polyak. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *Journal of Clinical Investigation*, 120(2):636–644, February 2010.

[S17] Dvir Netanely, Ayelet Avraham, Adit Ben-Baruch, Ella Evron, and Ron Shamir. Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Research*, 18(1), December 2016.

[S18] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, November 2009.

[S19] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, 7(1):269–294, March 2013.

[S20] Prabhakar Chalise and Brooke L. Fridley. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PloS one*, 12(5):e0176278, 2017.

[S21] Prabhakar Chalise, Devin C. Koestler, Milan Bimali, Qing Yu, and Brooke L. Fridley. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research*, 3(3):202, 2014.

[S22] Yanhua Chen, Manjeet Rege, Ming Dong, and Jing Hua. Non-negative matrix factorization for semi-supervised data clustering. *Knowledge and Information Systems*, 17(3):355–379, December 2008.

[S23] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[S24] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, March 2013.

[S25] Kristoffer H. Hellton and Magne Thoresen. Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*, 17(3):537–548, July 2016.

[S26] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, December 2012.

[S27] David B. Dunson and Amy H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25, 2005.

[S28] Eric F. Lock and David B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, October 2013.

[S29] Evelina Gabasova, John Reid, and Lorenz Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLOS Computational Biology*, 13(10):e1005781, October 2017.

[S30] Matthias Kormaksson, James G. Booth, Maria E. Figueroa, and Ari Melnick. Integrative Model-based clustering of microarray methylation and expression data. *The Annals of Applied Statistics*, 6(3):1327–1347, September 2012.

[S31] Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2017.

[S32] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D.M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A. Margolin, Laura J. van't Veer, Nuria Lopez-Bigas, Peter W. Laird, Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, Christopher C. Benz, Charles M. Perou, and Joshua M. Stuart. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944, August 2014.

[S33] R Lesurf, MR Aure, HH Mørk, V; Oslo Breast Cancer Research Consortium (OSBREAC) Vitelli, S Lundgren, AL Børresen-Dale, V Kristensen, F Wärnberg, M Hallett, and T Sørlie. Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Research*, 19(1), December 2017.

[S34] Aslaug Aamodt Muggerud, Michael Hallett, Hilde Johnsen, Kristine Kleivi, Wenjing Zhou, Simin Tahmasebpoor, Rose-Marie Amini, Johan Botling, Anne-Lise Børresen-Dale, Therese Sørlie, and Fredrik Wärnberg. Molecular diversity in ductal carcinoma *in situ* (DCIS) and early invasive breast cancer. *Molecular Oncology*, 4(4):357–368, August 2010.

[S35] Dhammika Amaratunga and Javier Cabrera. Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, 96(456):1161–1170, December 2001.

[S36] Nizar Touleimat and Jörg Tost. Complete pipeline for Infinium ® Human Methylation 450k BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–341, June 2012.

[S37] C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey, C J Marsit, E A Houseman, and R Brown. Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6):1394–1402, September 2013.

[S38] Gro Nilsen, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Marianne B. Eide, Oscar M. Rueda, Suet-Feung Chin, Roslin Russell, Lars O. Baumbusch, and Carlos Caldas. Copynumber: Efficient algorithms for single-and multi-track copy number segmentation. *BMC genomics*, 13(1):591, 2012.

[S39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[S40] Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J.S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, March 2009.

[S41] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[S42] Robert Lesurf, Miriam Ragle Aure, Hanne Håberg Mørk, Valeria Vitelli, Steinar Lundgren, Anne-Lise Børresen-Dale, Vessela Kristensen, Fredrik Wärnberg, Michael Hallett, Therese Sørlie, Torill Sauer, Jürgen Geisler, Solveig Hofvind, Elin Borgen, Anne-Lise Børresen-Dale, Olav Engebråten, Øystein Fodstad, Øystein Garred, Gry Aarum Geitvik, Rolf Kåresen, Bjørn Naume, Gunhild Mari Mælandsmo, Hege G. Russnes, Ellen Schlichting, Therese Sørlie, Ole Christian Lingjærde, Vessela Kristensen, Kristine Kleivi Sahlberg, Helle Kristine Skjerven, and Britt Fritzman. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Reports*, 16(4):1166–1179, July 2016.

[S43] Matahi Moarii, Valentina Boeva, Jean-Philippe Vert, and Fabien Reyal. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 16(1), December 2015.

[S44] Carlos E. Rodríguez and Stephen G. Walker. Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, January 2014.

[S45] Arno Fritsch and Katja Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, June 2009.

[S46] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, December 2011.

[S47] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236, March 1963.