Defining tissue specific gene sets using consensus clustering

GEN80436

Stephen Coleman
940309160050

supervised by

Bas Zwaan
Laboratory of Genetics, Wageningen University
and
Chris Wallace
Department of Medicine, Cambridge University

A thesis presented for the degree of
Master's in Bioinformatics

Laboratory of Genetics
Wageningen University

Abstract

A priori defined gene sets are key to gene set enrichment analysis [24] a powerful tool in genetic analysis. Gene sets are constructed through linking genes by some common feature. This can be a function, the location of the gene product, the participation of the product in some metabolic or signalling pathway, the protein structure, the presence of transcription-factor-binding sites or other regulatory elements, the participation in multiprotein complexes, or any one of several other definitions [25][24][12][1]. However, all of these criteria are tissue agnostic. We propose to produce tissue specific gene sets by applying multiple dataset integration [13] (a Bayesian unsupervised clustering method) to the gene expression data from the Correlated Expression and Disease Association Research cohort [26], a dataset of 9 tissue / cell types.

We show that problems with convergence and dependence upon initialisation common in high dimensionality settings can be overcome by means of consensus clustering [18]. We then use Multiple Dataset Integration by means of consensus clustering to produce gene sets.

1 Introduction

With the onset of microarrays and RNAseq, producing gene expression data in large quantities for a wide number of genes is increasingly enabled. Unfortunately the large amount of data gifted onto the genomics community by these methods is difficult to interpret and analyse. Gene Set Enrichment Analysis (GSEA) attempts to overcome some of these issues by using prior knowledge to define groups of genes linked through their biological function [9]. The set is defined using knowledge external to the current analysis; a common method is using the manually annotated pathways available on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [5].

Analysing pre-defined gene sets and changes in the expression of the full set rather than considering each constituent member on an individual basis has more statistical power [19]. Consider, in analysing gene sets as a group the degree of perturbation in the expression of the full gene set due to the disease state / alternative phenotype that is required to be considered significant is much less than that required in analysing each of its constituent members individually [4][28].

The problem of how to define gene sets is non-trivial, with many variations present in the literature. There exist many databases of gene sets [1][12][25]. The Molecular Signature Database [24] (MSigDB) is one of the most popular resources for GSEA and encompasses many different gene sets defined under various criteria or generated from separate resources.

However, none of these definitions of a "set" incorporate tissue specific information. This seems an oversight. Cell-type specific gene pathways are pivotal in differentiating tissue function, implicated in hereditary organ failure, and mediate acquired chronic disease [11]. More and more evidence is being accrued to highlight the cell-type specific level of gene expression [7][21][16]. Thus we propose defining tissue specific gene sets.

To describe gene sets within the data, some clustering method is required. Applied on expression values or some transformed variation thereof, groups of genes are created based on some concept of similarity (or alternatively on some concept of dissimilarity or distance). Depending on the choice of transformation and clustering method further questions might arise such as defining the number of clusters (required for instance with K-means clustering) or the type of distance to use (for instance within hierarchical clustering and the methods that integrate this method such as Weighted Gene Correlation Network Analysis). For clustering within a dataset we choose *Dirichlet processes* as the method as the number of clusters is learnt from the data and the concept of distance in these models is based upon the likelihood of the Gaussian distributions describing the sub-populations, an intuitive measure for continuous data. Specifically we use Bayesian Dirichlet processes as these capture uncertainty of membership which is appropriate in this application. Genes can be members of multiple sets or else their membership might be poorly defined; thus the model uncertainty represents biological uncertainty.

Within the CEDAR cohort there are multiple datasets containing information about the same genes for different tissues or cell types. Ideally a model could integrate information about common clustering structure across the datasets to reduce uncertainty within making assumptions that could impose false structure upon the data or in some other way reduce the signal unique to each tissue. Such methods are referred to as *integrative clustering methods*. Of this field we choose to use *Multiple Dataset Integration* (MDI) [13] as this method is Bayesian (and thus has principled quantification of uncertainty) and is an extension of Dirichlet processes.

As we have a large number of variables (p ¿ 250), we implement *consensus clustering* to overcome the problem of describing multiple modes in high dimension space. This is a recurring problem with Bayesian clustering methods as they rely upon *Markov Chain Monte Carlo* (MCMC) methods to describe the posterior distribution. These methods have the nice property that they guarantee describing the correct distribution given infinite time. Unfortunately any amount of finite time is very small in comparison to infinite time and no MCMC method guarantees convergence within finite time. This problem is particular present as the number of dimensions scales and the posterior distribution is multi-modal. In this case the algorithm tends to describe the space within a mode, but as the mode is far denser than the surrounding space (particularly in high dimensions), the probability of escaping the mode and exploring the full space is very low. Thus to describe the full space in finite time requires use of multiple unique initialisations. The number of initialisations is required to be sufficiently high that each mode is described. In this way the full space can be explored by the combinations of models. Combining clustering models in this way is referred to as *consensus clustering*. It is a natural extension to the concept of ensemble methods (such as Random Forest citeBREIMANrandomforest). The final model is a strong method built of many weak methods that depend upon the instability of the model similarly to Bagging citeBREIMANbagging. This consensus clustering uses many short MCMC chains aggressively thinned, thus it is both a better description of the target distribution than an individual model, but also far quicker to run as each chain is both embarrassingly parallel and short.

2 Theory

2.1 Bayesian inference

Bayesian inference is an alternative paradigm to frequentist methods that has several attractive properties.

1. Principled error qualification;
2. Integration of prior knowledge and beliefs; and
3. Variables are treated as stochastic rather than the data (in comparison to frequentist methods).

In this project it is points 1 and 3 that makes this framework attractive. As stated previously, model uncertainty can represent biological uncertainty and treating variables within the model as random is more intuitive than treating the data as stochastic realisations of some process.

The keystone of Bayesian inference is Bayes' rule which defines how one can update a hypothesis as more information is made available. For observations X and a random variable θ where Θ is the entire sample space for θ:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_{\Theta} p(X|\theta')p(\theta')d\theta'} \tag{1}$$

- We refer to $p(\theta|X)$ as the *posterior* distribution of θ as it is the distribution associated with θ *after* observing X.
- $p(\theta)$ is the *prior* distribution of θ and captures our beliefs about θ before we observe X.
- $p(X|\theta)$ is the *likelihood* of X given θ, the probability of data X being generated given our model is true. It is the criterion we focus on in our model if we would use a frequentist approach to the inference (and hence why the frequentist philosophy treats the data as random); maximising this quantity in our model generates the manifold that best describes the observed data.
- $\int_{\Theta} p(X|\theta')p(\theta')d\theta'$ is the *normalising constant*. This quantity is also referred to as the *evidence* [15] or *marginal likelihood* and is normally represented by Z. It is referred to as the marginal likelihood as we marginalise the parameter θ by integrating over its entire sample space.

In terms of sampling, the prior is very useful as a clever choice of prior can ensure that the posterior is always solvable, that we do not encounter singularities in our distribution.

2.2 Clustering

Given data $X = (x_1, \ldots, x_n)$, we define a *clustering* or partition of the data by:

$$
\begin{aligned}
Y &= \{Y_1, \ldots, Y_K\} && (2)\\
Y_k &= \{x_{1_k}, \ldots, x_{n_k}\} && (3)\\
Y_i \cap Y_j &= \emptyset \ \forall \ i, j \in \{1, K\}, i \neq j && (4)\\
n_k &= \mid Y_k \mid \ \geq 1 \ \forall \ k \in \{1, \ldots, K\} && (5)\\
\sum_{k=1}^{K} n_k &= n && (6)
\end{aligned}
$$

In short we have K nonempty disjoint sets of data, each of which is referred to as a *cluster*, the set of which form a *clustering*. A label $c_i = k$ states that point x_i is assigned to cluster Y_k. We define the collection of labels $c = (c_1, \ldots, c_n)$ as denoting the membership of each point.

2.3 Mixture models

Given some data $X = (x_1, \ldots, x_n)$, we assume K unobserved subpopulations generate the data and that these subpopluations can be revealed by imposing a clustering $Y = \{Y_1, \ldots, Y_K\}$ on the data.

It is assumed that each of the K subpopulations can be modelled by a parametric distribution, $f(\cdot)$ with associated parameters θ and that the full model density is then the weighted sum of these probability density functions where the weights are the component proportions, π_k:

$$p(x_i | c_i = k) = \sum_{k=1}^{K} \pi_k f(x_i | \theta_k) \tag{7}$$

We use a Multivariate Gaussian distribution to describe each subpopulation for three reasons:

1. Convention: the Gaussian distribution is extremely common within the literature;
2. Pragmatism: the Gaussian distribution is easy to work with; and
3. Conservatism: if the only statements we are willing to make about a distribution over real numbers are its mean and variance, than the Gaussian distribution maximises the entropy and is thus the most conservative choice of distribution.

We use a large K (specifically, 50) to imitate a Dirichlet process. Strictly speaking a Dirichlet process sets $K = \infty$, but if we use a sufficiently large K such that the data empties the majority of clusters, then we have the desired property of Dirichlet processes. This property is that the number of inhabited clusters is not fixed and can increase or decrease depending on the data.

2.3.1 Bayesian mixture models

We carry out Bayesian inference of this model using MCMC methods. Specifically, we employ is Gibbs sampling which can be summarised as iterating between the following steps:

1. For each of K clusters sample θ_k and π_k from the associated distributions based on current memberships, c_i; and
2. For each of n individuals sample c_i based on the new θ_k and π_k.

The output consists of a matrix n_{iter} rows and p columns (fro p genes). The i^{th} row describes the cluster the genes are assigned to in the i^{th} iteration of the Gibbs sampler. To summarise this information we use a posterior similarity matrix (PSM). The (i, j) cell of the PSM contains there is the fraction of recorded iterations for which the i^{th} and j^{th} genes have common labelling. One can see that this implies the PSM is symmetric and has diagonal entries of 1.

From this PSM a single clustering estimate, $\hat{c}$, can be described from the PSM by maximising the posterior expected adjusted Rand index citeFRITSCH. Other methods such as minimisation of Binder's loss function or minimization of Dahl's criterion are based on the original unadjusted for chance Rand index. We prefer the adjusted Rand index for reasons mentioned in section 2.6 and thus choose to use the method described by citetFRITSCH.

2.4 Multiple dataset integration

Consider the case when we have observed paired datasets $X_1 = (x_{1,1}, \ldots, x_{n,1})$, $X_2 = (x_{1,2}, \ldots, x_{n,2})$, where observations in the ith row of each dataset represent information about the same individual. We would like to cluster individuals using information common to both datasets. One could concatenate the datasets, adding additional covariates for each individual. However, if the two datasets have different clustering structures this would reduce the signal of both clusterings and probably have one dominate. If the two datasets have the same structure but different signal-to-noise ratios this would reduce the signal in the final clustering. In both these cases independent models on each dataset would be preferable. Kirk et al. [13] suggest a method to carry out clustering on both datasets where common information is used but two individual clusterings are outputted. This method is driven by the allocation prior:

$$p(c_{i1}, c_{i2}|\phi) \propto \pi_{i1}\pi_{i2}(1 + \phi\mathbb{I}(c_{i1} = c_{i2})) \tag{8}$$

Here $\phi \in \mathbb{R}_+$ controls the strength of association between datasets. (8) states that the probability of allocating individual i to component $c_{i,1}$ in dataset 1 and to component $c_{i,2}$ in dataset 2 is proportional to the proportion of these components within each dataset and up-weighted by ϕ if the individual has the same labelling in each dataset. Thus as ϕ grows the correlation between the clusterings grow and we are more likely to see the same clustering emerge from each dataset. Conversely if $\phi = 0$ we have independent mixture models.

The generalised case for L datasets, $X_1 = (x_{1,1}, \ldots, x_{n,1}), \ldots, X_L = (x_{1,l}, \ldots, x_{n,l})$ for any $L \in \mathbb{N}$ is simply a matter of combinatorics. In this case, (8) extends to:

$$p(c_{i1}, \ldots, c_{iL}|\boldsymbol{\phi}) \propto \left[\prod_{l_1=1}^{L} \pi_{c_{il_1}l_1}\right] \left[\prod_{l_2=1}^{L-1} \prod_{l_3=l_2+1}^{L} (1 + \phi_{l_2l_3}\mathbb{1}(c_{il_2} = c_{il_3}))\right] \tag{9}$$

Here $\boldsymbol{\phi}$ is the $\binom{L}{2}$-vector of all ϕ_{ij} where ϕ_{12} is the variable ϕ in (8).

Thus MDI is an extension of mixture models to multiple datasets where correlated clustering structure is used to "upweigh" similar clusters across datasets. MDI has been applied to precision medicine, specifically glioblastoma sub-typing [23], in the past showing its potential as a tool.

2.5 Consensus clustering

In the scenario that MDI struggles to explore the entire posterior distribution from any given initialisation for any realistic number of iterations of MCMC, we propose use of a "consensus" clustering [18]. In this scenario we draw samples of clusterings from MCMC chains with different initialisations and use these clusterings to describe the posterior distribution. In practice this involves running n_{seeds} different chains of MDI for a smaller number of iterations (some $n_{iter} \in [500, 1000]$), burning out the first $n_{iter} - 1$ iterations and saving the clustering from the final iteration. We then combine the clusterings from all n_{seeds} within a posterior similarity matrix (PSM), a $N \times N$ matrix where the (i, j) entry is the proportion of times genes i and j are in the same cluster. This means that the PSM is not affected by label-flipping and that it is a symmetric matrix with 1's along the diagonal and all entries in the unit interval. From this PSM a summary clustering may be calculated. The combination of different initialisations explores all the possible likelihood maixma and thus provides a more informed clustering. As the algorithm is not exploring the full space in any given iteration, we expect that the uncertainty quantification is optimistic, however we argue that an estimate made using insufficient data is better than one made using none at all and that this method is the best currently available to us for quantifying the uncertainty and exploring the posterior distribution.

2.6 Rand index

A popular metric for comparing the similarity of two clusterings of the data is the *Rand index* [22]. If one assumes that all points are of equal importance in determining clusterings, then in combination with the discrete nature of clusters and the fact that a cluster is defined as much by what it does not contain as that which it does, Rand [22] proposes a metric to measure similarity between clusterings. Between clusterings Y and Y' for any two points x_i and x_j there can exist one of a number of scenarios regarding their labeling. Let γ_{ij} be a measure between the two points x_i and x_j. For the two points, they can have:

1. the same label in both clusterings ($c_i = c_j \land c'_i = c'_j$) ($\gamma_{ij} = 1$);
2. different labels in both ($c_i \neq c_j \land c'_i \neq c'_j$) ($\gamma_{ij} = 1$); or
3. the same label in one but not in the other ($c_i \neq c_j \land c'_i = c'_j \lor c_i = c_j \land c'_i \neq c'_j$) ($\gamma_{ij} = 0$).

Thus Rand [22] propose counting the number of times any two points have one of 1 or 2 from list 2.6 and finding the proportion of these compared to the number of all possible point combinations. More formally, this is:

$$A\binom{n}{2}^{-1} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \gamma_{ij} \tag{10}$$

This can be envisioned as a $K \times K'$ contingency table of the count of overlapping points, as shown in table 1. Table 1 uses the following notation:

- n_{ij} is the number of points that have membership in Y_i in clustering Y and Y'_j in clustering Y';
- $n_{\cdot j}$ is the number of points in cluster Y'_j in clustering Y';
- $n_{i\cdot}$ is the number of points in cluster Y_i in clustering Y; and
- $n_{\cdot\cdot} = n$ is the number of points in clusterings Y and Y'.

$Y \backslash Y'$	Y'_1	Y'_2	$\cdots$	$Y'_{K'}$	Sums
Y_1	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1\cdot}$
Y_2	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Y_K	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K'}$	$n_{\cdot\cdot} = n$

Table 1: Contingency table used by Rand [22] to calculate a measure of similarity between clusterings Y and Y'.

One can restate equation 10 in terms of the notation from table 1 [2]:

$$\begin{aligned} A &= \binom{n}{2} + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right) && (11) \\ &= \binom{n}{2} + 2\sum_{i=1}^{K}\sum_{j=1}^{K'} \binom{n_{ij}}{2} - \left[\sum_{i=1}^{K}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\right] && (12) \end{aligned}$$

Hubert and Arabie [10] extend the Rand index to account for chance. They include a null hypothesis and assume that there is a probability of some points having a γ value of 1 by chance. Consider the scenario where a point x_i has the same label as another point x_j under clustering Y. For another clustering Y', there a non-zero is a probability $c'_i = c'_j$ purely by chance and does not represent a similarity between Y and Y'. If one generates two clusterings Y and Y' by sampling from the integers in the closed interval $[1, K]$ (i.e. by sampling from discrete unfirom distribution $\mathcal{U}\{1, K\}$), then the contingency table generated is constructed from the generalised hyper-geometric distribution [10]. It can be shown that the expected number of points with common membership in both clusters is non-zero. Specifically:

$$\mathbb{E}\left(\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{n_{ij}}{2}\right) = \frac{\sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K}\binom{n_{\cdot j}}{2}}{\binom{n}{2}} \tag{13}$$

This is the product of the number of distinct pairs that can be formed from rows and the number of distinct pairs that can be constructed from columns, divided by the total number of pairs.

For a particular cell of the contingency table, the expected number of entries of the type described in point 1, is the product of number of pairs in its row and in its column divided by the total number of possible pairs:

$$\mathbb{E}\left(\binom{n_{ij}}{2}\right) = \frac{\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}}{\binom{n}{2}} \tag{14}$$

One can see that as each component of equation 11 is some transformation of $\sum_{i,j}\binom{n_{ij}}{2}$, one can directly state the expected value of the Rand index by combining equations 11 and 14:

$$\mathbb{E}\left(A\binom{n}{2}^{-1}\right) = 1 + 2\sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\binom{n}{2}^{-2} - \left[\sum_{i=1}^{K}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\right]\binom{n}{2}^{-1} \tag{15}$$

Defining an index corrected for chance as:

$$\text{Corrected index} = \frac{\text{Index} - \text{Expected index}}{\text{Maximum index} - \text{Expected index}} \tag{16}$$

Assuming a maximum value of 1 for the Rand index then gives a corrected Rand index:

$$\frac{\sum_{i=1}^{K}\sum_{j=1}^{K'}\binom{n_{ij}}{2} - \sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\binom{n}{2}^{-1}}{\frac{1}{2}\left[\sum_{i=1}^{K}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\right] - \sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}\binom{n}{2}^{-1}} \tag{17}$$

We define this quantity described in equation 17 as the *adjusted Rand index* and we use it as our measure of choice for similarity between clusterings.

We describe an explicit example motivating the adjusted Rand index in section 2.6.1.

2.6.1 Motivating example: adjusted Rand index

Consider the case of n labels Y and Y' generated from $\mathcal{U}\{1,3\}$ where n is some arbitrarily large number. Then as n tends to infinity we can expect that our contingency table has entries of $\frac{n}{9}$ in each cell. If one calculates the Rand index on these random partitions where any similarity is purely by chance one finds, it comes to (approximately) 0.56. This suggests there is some similarity between Y and Y', but this is misleading as we know any similarity is stochastic. In the same scenario the adjusted Rand index between the partitions is 0. This seems preferable. Based on this, one could argue that the Rand index has inflated values. Consider the case that we have n points in total, but we let the first $\frac{7n}{16}$ have a common label (say $(c_1, \ldots, c_{n_1}) = 1$ for $n_1 = \frac{7n}{16}$) and then draw the remaining $\frac{9n}{16}$ points from $\mathcal{U}\{1,3\}$. Then, as n tends to infinity, our contingency table tends to that described in table 2. One feels that the high Rand index for such a clustering, 0.64, is misleading in its magnitude. In such a scenario we feel one has to consider this 0.64 in the context of the 0.56 for a purely random similarity - this is difficult to do without explicitly checking what the Rand index is for a random partitioning for a given K and K'. Thus the use of the full unit interval

$Y \diagdown Y'$	Y'_1	Y'_2	Y'_3	Sums
Y_1	$\frac{n}{2}$	$\frac{n}{16}$	$\frac{n}{16}$	$\frac{10n}{16}$
Y_2	$\frac{n}{16}$	$\frac{n}{16}$	$\frac{n}{16}$	$\frac{3n}{16}$
Y_3	$\frac{n}{16}$	$\frac{n}{16}$	$\frac{n}{16}$	$\frac{3n}{16}$
Sums	$\frac{10n}{16}$	$\frac{3n}{16}$	$\frac{3n}{16}$	$\frac{16n}{16} = n$

Table 2: Contingency table for the non-random clustering described in section 2.6.1.

in comparing similarity by a corrected index such as the adjusted Rand index requires less vigilance on the part of the analyst. In the second scenario, the adjusted Rand index is 0.28.

2.7 Gene sets

We know from Genome Wide Association Studies (GWAS) that many diseases are polygenic in nature [19]. This suggests that it is natural to be considering sets of genes in analysis of many diseases. Subramanian et al. [24] highlight the importance of gene sets, claiming that within a single metabolic pathway an increase of 20% in all the associated gene products may be more important than a 20-fold increase in a single gene.

Thus clustering genes into groups known as "gene sets" is natural and useful from both a biological and statistical perspective - it can increase the interpretability and the power of an analysis [20][27].

2.7.1 Tissue specificity

Previous attempts to achieve this have used the Genotype Tissue Expression (GTEx) [8] database [14], but here the profiles are for human donors post-mortem. We suspect that the data derived from these cells may not contain the same information as that collected from living, active cells. Furthermore, the GTEx data is across many different tissues (144 are used in [14]), but we focus on cell types relevant to autoimmune disease in general (i.e. blood cells) and Inflammatory Bowel Disease in particular (intestinal samples).

Gene sets should contain sets of genes that have correlated expression. If this is the case, it is often assumed that the genes are common members of some metabolic pathway and that their products interact. As this correlated expression is represented by a common variation across people rather than in the magnitude of expression, we will standardise the expression data as described in section 2.8. We describe a small example to highlight our reasoning

Genes	Person 1	Person 2	Person 3	Person 4
A	5.1	5.2	4.9	5.0
B	5.1	4.9	5.2	5.4
C	1.4	1.5	1.2	1.3
D	1.4	1.2	1.5	1.7
E	1.4	1.5	1.4	1.5

Table 3: Example gene expression data.

in section 2.8.1.

2.8 Standardisation

For a p-vector of observations, $X_i = (x_{i1}, \ldots, x_{1p})$, we define *standardisation* of X_i as the mapping from X_i to $X_i' = (x_{i1}', \ldots, x_{ip}')$ defined by the *sample mean*, $\bar{x}_i$, and *sample standard deviation*, s_i:

$$\bar{x}_i = \frac{1}{p} \sum_{j=1}^{p} x_{ij} \tag{18}$$

$$s_i^2 = \frac{1}{p-1} \sum_{j=1}^{p} \left(x_{ij} - \bar{x}_i\right)^2 \tag{19}$$

$$x_{ij}' = \frac{x_{ij} - \bar{x}_i}{s_i} \quad \forall \quad j \in (1, \ldots, p) \tag{20}$$

We refer to X_i' as the standardised form of X. If we are given a dataset $X = (X_1, \ldots, X_n)$ where each X_i is a p-vector of observations of the form referred to above, then in referring to the standardised form of X, we mean the dataset $X' = (X_1', \ldots, X_n')$ where each X_i' is the standardised form of X_i.

Standardisation moves the values observed for each X_i to a common scale where each vector has an observed mean and standard deviation of 0 and 1 respectively.

2.8.1 Motivating example: Standardising gene expression data

If one considers table 3 which contains an example of expression data for some genes A, B, C, D and E across people 1 to 4. One can see that genes A and C have similar patters in variation across the people, as do genes B and D. Gene E is not consistent with any other gene here. However, as this relative variation is of interest rather than the magnitude of expression, one can see that standardising the data is required. If one were to cluster the data as represented in table

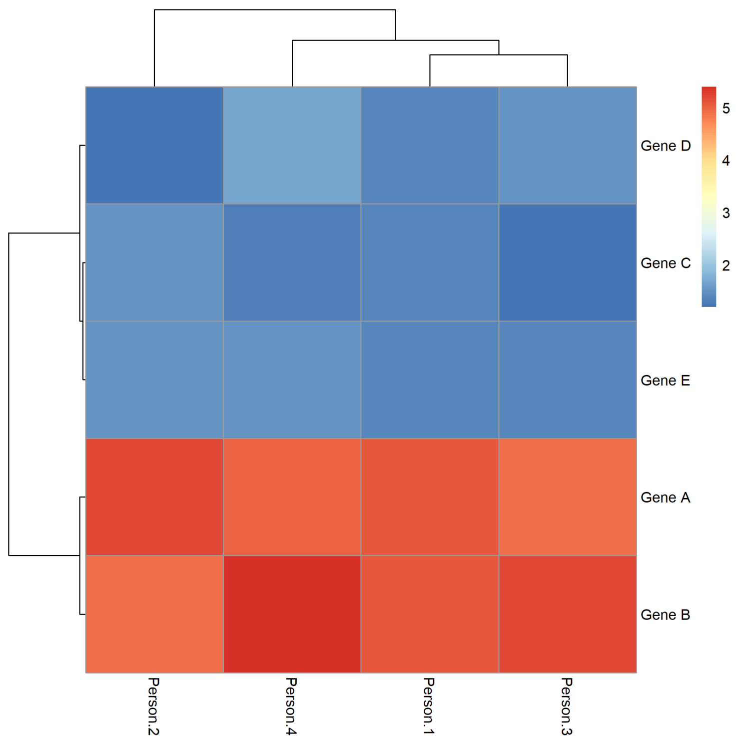

Figure 1: Heatmap of expression data in table 3 showing the clusters based upon magnitude of expression.

3, one would place genes A and B in one cluster and genes C, D and E in another as their absolute expression levels are similar (as can be seen in figure 1). However, if the expression level of each gene is standardised as per section 2.8, the data is then as represented in table 4. As the data are now on the same scale the characteristic that will determine a clustering is the variation of expression across people. As we want genes with similar patterns of variation (i.e. that are co-expressed) this enables us to cluster under our objective of defining gene sets. In this case genes A and C are one cluster, genes B and D another with gene E in a cluster alone, as can be seen in figure 2. As this is the type of data we wish to cluster across, we therefore most standardise our expression data before clustering can be implemented.

Genes	Person 1	Person 2	Person 3	Person 4
A	0.39	1.16	-1.16	-0.39
B	-0.24	-1.20	0.24	1.20
C	0.39	1.16	-1.16	-0.39
D	-0.24	-1.20	0.24	1.20
E	-0.87	0.87	-0.87	0.87

Table 4: Example standardised gene expression data.

3 Data

3.1 Simulation: Case 1

3.2 Simulation: Case 2

3.3 CEDAR dataset

We use the gene expression data from the CEDAR cohort [26]. This data is available in a processed form online. This consists of 9 .csv files, one for each tissue / cell type present of normalised gene expression data for 323 individuals. These are healthy individuals of European descent; the cohort consists of 182 women and 141 men with an average age of 56 years (but ranging from 19 to 86). None of the individuals are suffering from any autoimmune or inflammatory disease and were not taking corticosteroids or non-steroid anti-inflammatory drugs (with the exception of aspirin).

With regards to tissue types, samples from six circulating immune cells types (followed in brackets by the abbreviation for the associated dataset):

- CD4+ T lymphocytes (CD4);
- CD8+ T lymphocytes (CD8);
- CD14+ monocytes (CD14);
- CD15+ granulocytes (CD15);
- CD19+ B lymphocytes (CD19); and
- platelets (PLA).

Data from intestinal biopsies are also present, with samples taken from three distinct locations:

- the illeum (IL);
- the rectum (RE); and
- the colon (TR).

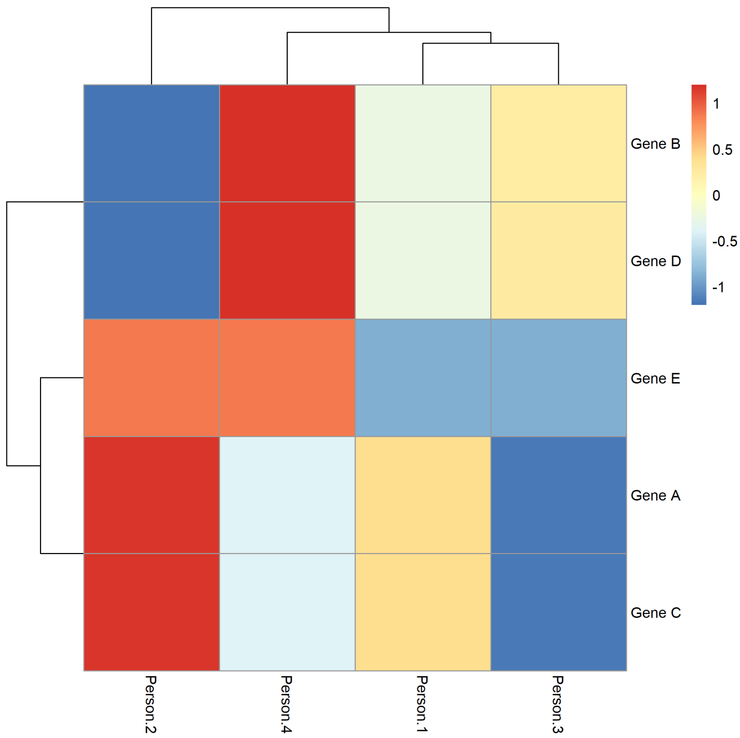

Figure 2: Heatmap of expression data in table 4 showing the clusters based upon variation of expression across people.

Not every individual is present in every dataset. However, as we are clustering genes this should not present a problem.

Whole genome expression data were generated using HT-12 Expression Beadchips following the instructions of the manufacturer (Illumina). There are 18,524 probes present between the 9 datasets. The fluorescence intensities are available after undergoing a $\log_2$ transformation and being Robust Spline Normalized (a method that is designed to normalize variance-stabilized data).

It should be noted that there are differing degrees of missingness between the datasets (for instance the platelets dataset has 6,564 probes present in comparison to an average of 12,838 probes present per dataset, see figure 3).

Due to exponential increase in computational cost for each additional dataset, we use only the 7 most informative datasets, dropping PLA and CD15 from our analysis.

From a biological perspective we also expect PLA to be the least rich as platelets have no nucleus [29] and therefore any gene expression is an artefact from before they differentiated into platelets.

With regards to CD15 granulocytes, (mast cells, basophils, neutrophils and

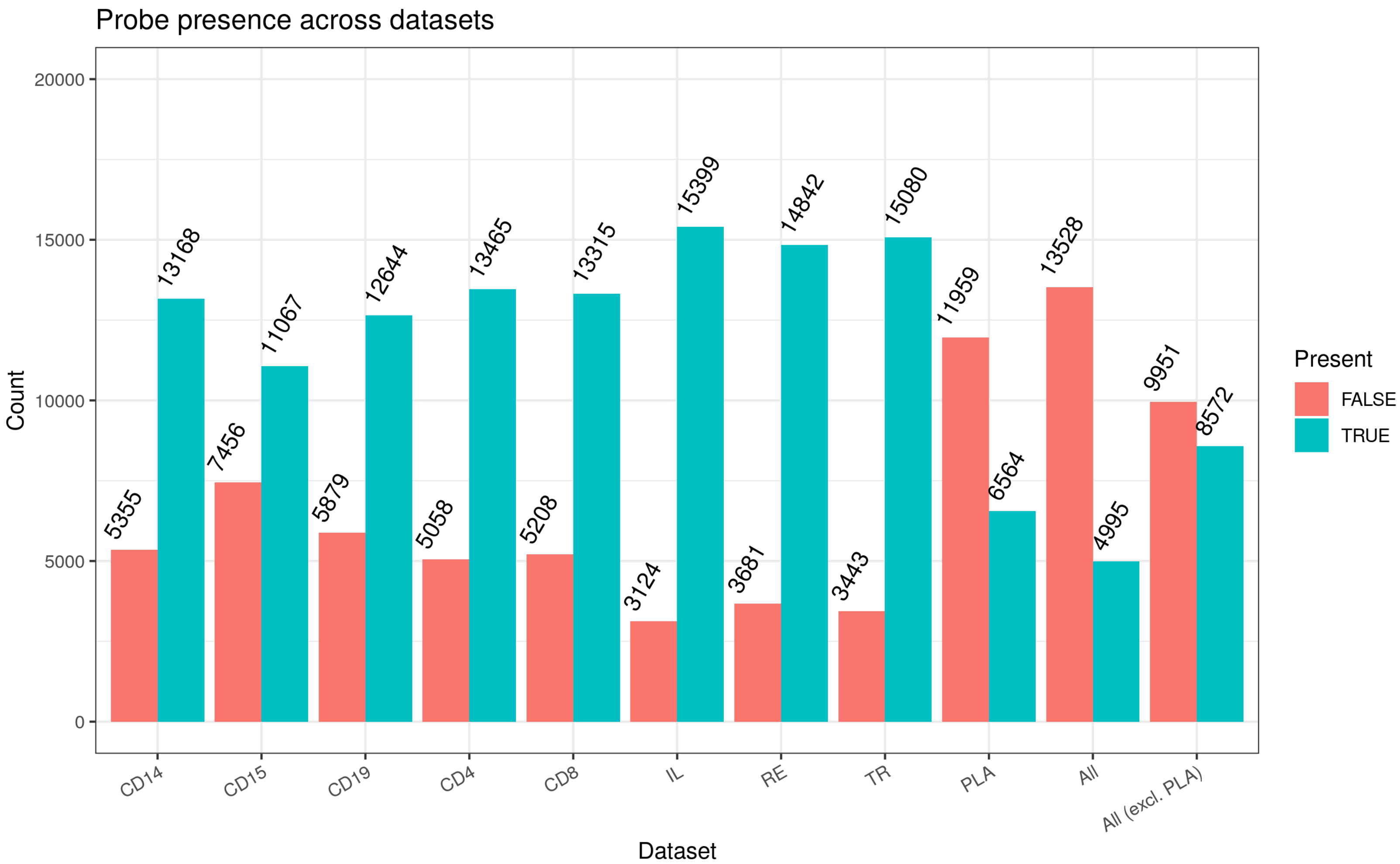

Figure 3: Probe presence across datasets. Under “All” we have the number of probes present in every dataset, under “All (excl. PLA)” we have the number of probes present in every dataset bar PLA. Note how there is greater missingness in the PLA dataset in comparison to the others.

eosinophils), these are quite distinct from B and T lymphocytes (see figure 4). Based on this we expect there to be less common information pertinent to clustering genes in other datasets. Arguably monocytes are equally distant, but the level of missingness in the CD15 dataset is greater than that in the CD14 dataset; thus CD15 is elimanted from our analysis.

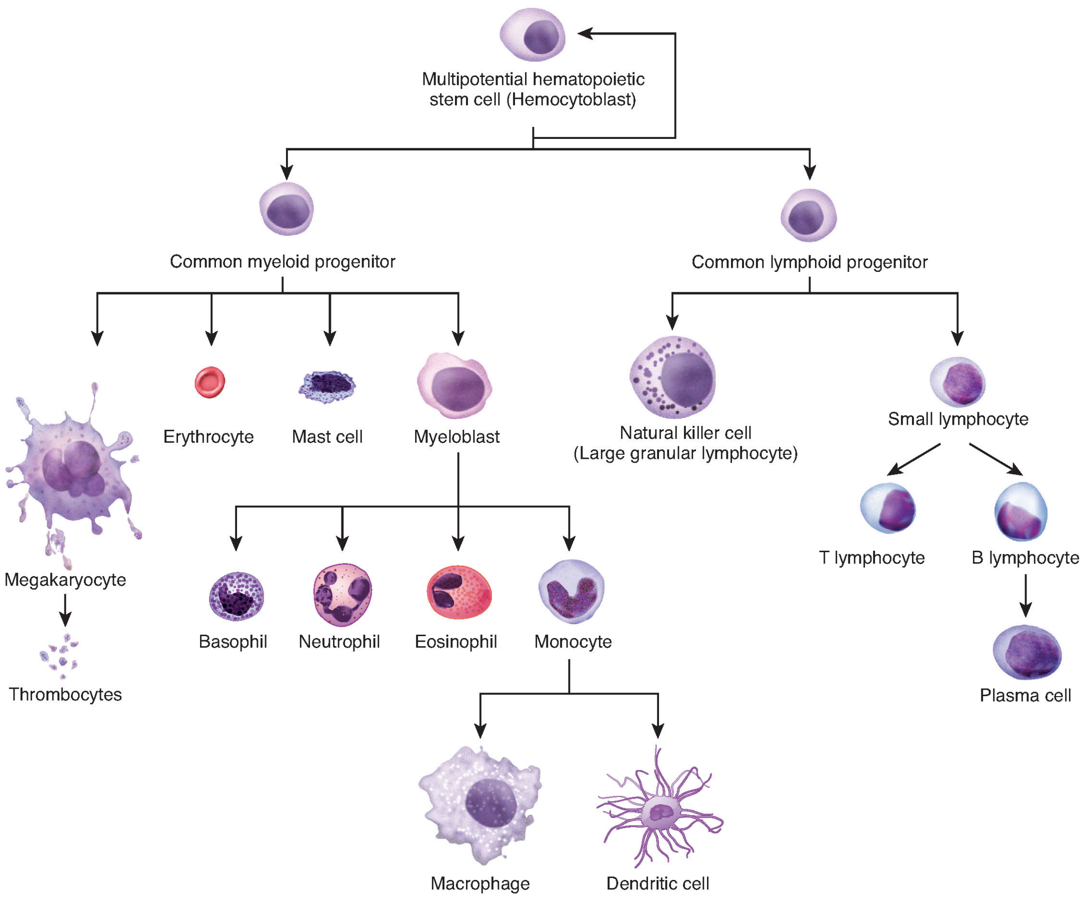

Figure 4: The differentiation of multipotent cells into blood and immune cells.

4 Methods

We first show via simulated data that MDI can cluster appropriately and that the consensus clustering does produce similar results to a converged single run.

We then simulate data where individual chains of MDI will struggle to converge and possibly will not converge in finite time. We show that consensus clustering explores a wider space than any individual chain and appears to describe something similar to the space described by the union of the chains.

Finally we apply consensus clustering to 1,000 probes for 8 datasets from the CEDAR dataset. An initial set of probes are chosen based on the members of 3 KEGG pathways:

1. Inositol phosphate metabolism (a broad biological pathway);
2. NOD-like receptor signaling pathway (a specific biological pathway with known involvement in IBD [3][6]); and
3. Inflammatory bowel disease (IBD) (a pathological pathway).

The union of these sets corresponds to 169 unique genes (or 287 probes as the mapping from the space of probes to that of genes is non-injective) that are present in the CEDAR dataset. The remaining probes are randomly selected from the total possible space (18,524 probes) less those corresponding to these genes (leaving 18,287 possible candidate probes). We then expect that the genes from the sets mentioned above (list 4) should cluster together. We use this as a test of our final clustering.

4.1 CEDAR data pipeline

For the CEDAR data, we follow this pipeline to prepare the data for clustering:

1. Transpose the data to have rows associated with gene probes and columns associated with individuals;
2. Remove NAs either imputing values using the minimum expressed value (as missingness is not random) or if above a threshold of missingness removing the column;
3. Standardise the data;
4. To apply MDI we require that each dataset have the same row names in the same order, so we re-arrange our datasets to have common order of probes;
5. For probes entirely missing from a given dataset we generate expression from a standard normal distribution for each probe. Then these probes are expressed as noise in the dataset and any clustering imposed upon them should be due to information about these probes present in other datasets; and
6. Apply MDI [17].

5 Results

5.1 Case 1: Proof of consensus

Generated data based on data from original MDI paper with some additional noisy clusters that MDI succeeds with.

5.2 Case 2: Overcoming multiple modes

Generated data that MDI struggles with.

5.3 Case 3: CEDAR data

Actual data.

6 Conclusion

We conclude, and finish dispute.

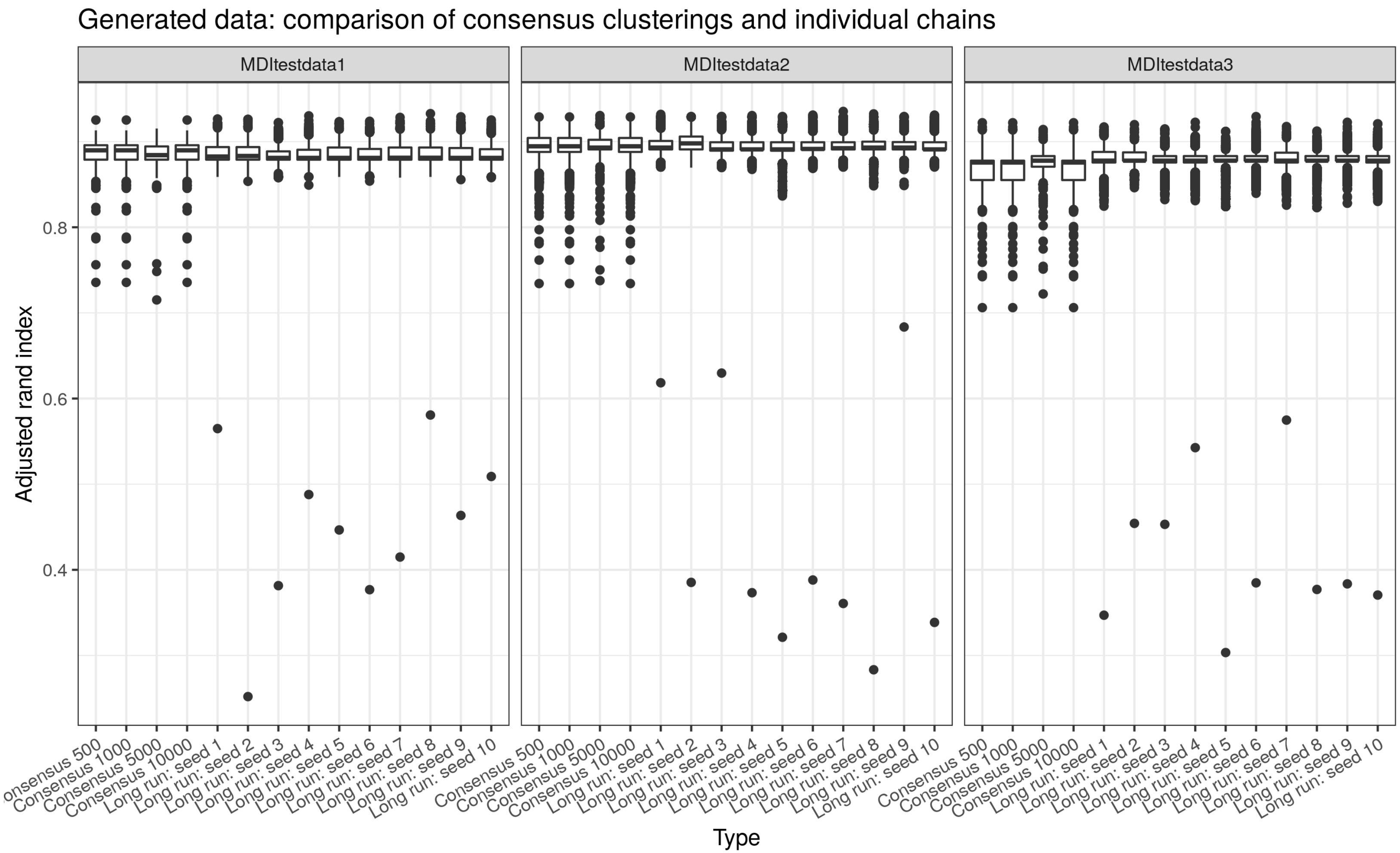

Figure 5: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and different initialisation of long chains.

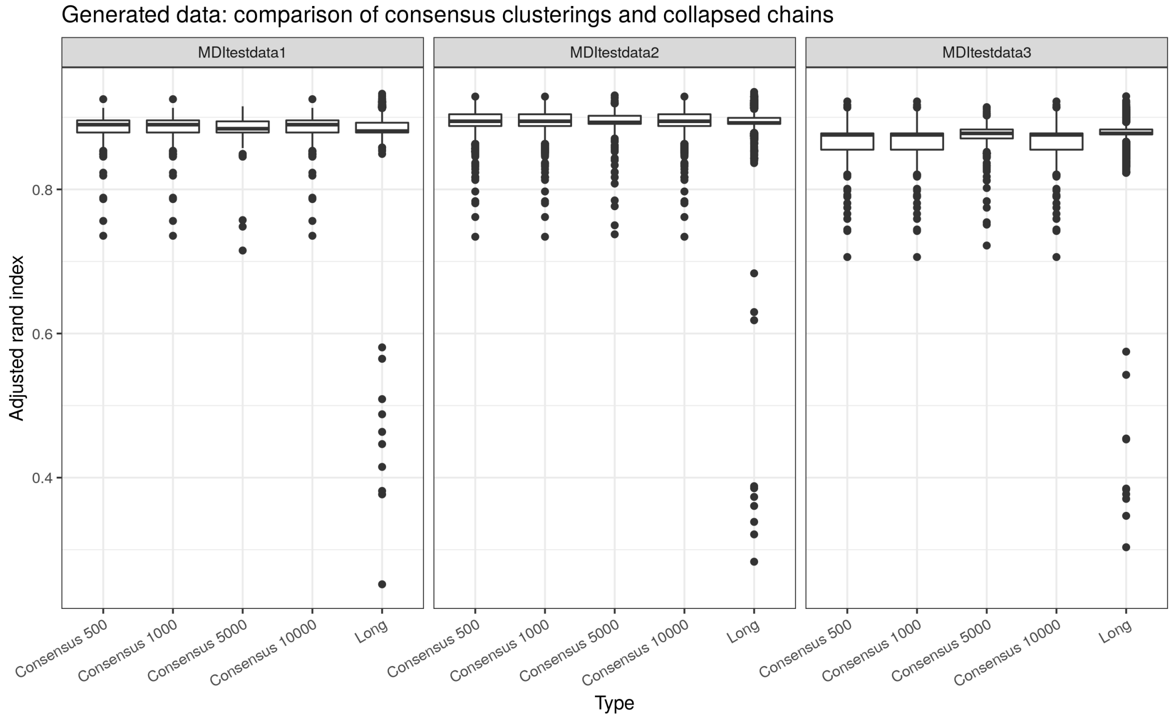

Figure 6: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

References

[1] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036, 1546-1718. doi: 10.1038/75556.

[2] Robert L. Brennan and Richard J. Light. MEASURING AGREEMENT WHEN TWO OBSERVERS CLASSIFY PEOPLE INTO CATEGORIES NOT DEFINED IN ADVANCE. *British Journal of Mathematical and Statistical Psychology*, 27(2):154–163, November 1974. ISSN 00071102. doi: 10.1111/j.2044-8317.1974.tb00535.x.

[3] Lam Carneiro, Jg Magalhaes, I Tattoli, Dj Philpott, and Lh Travassos. Nod-

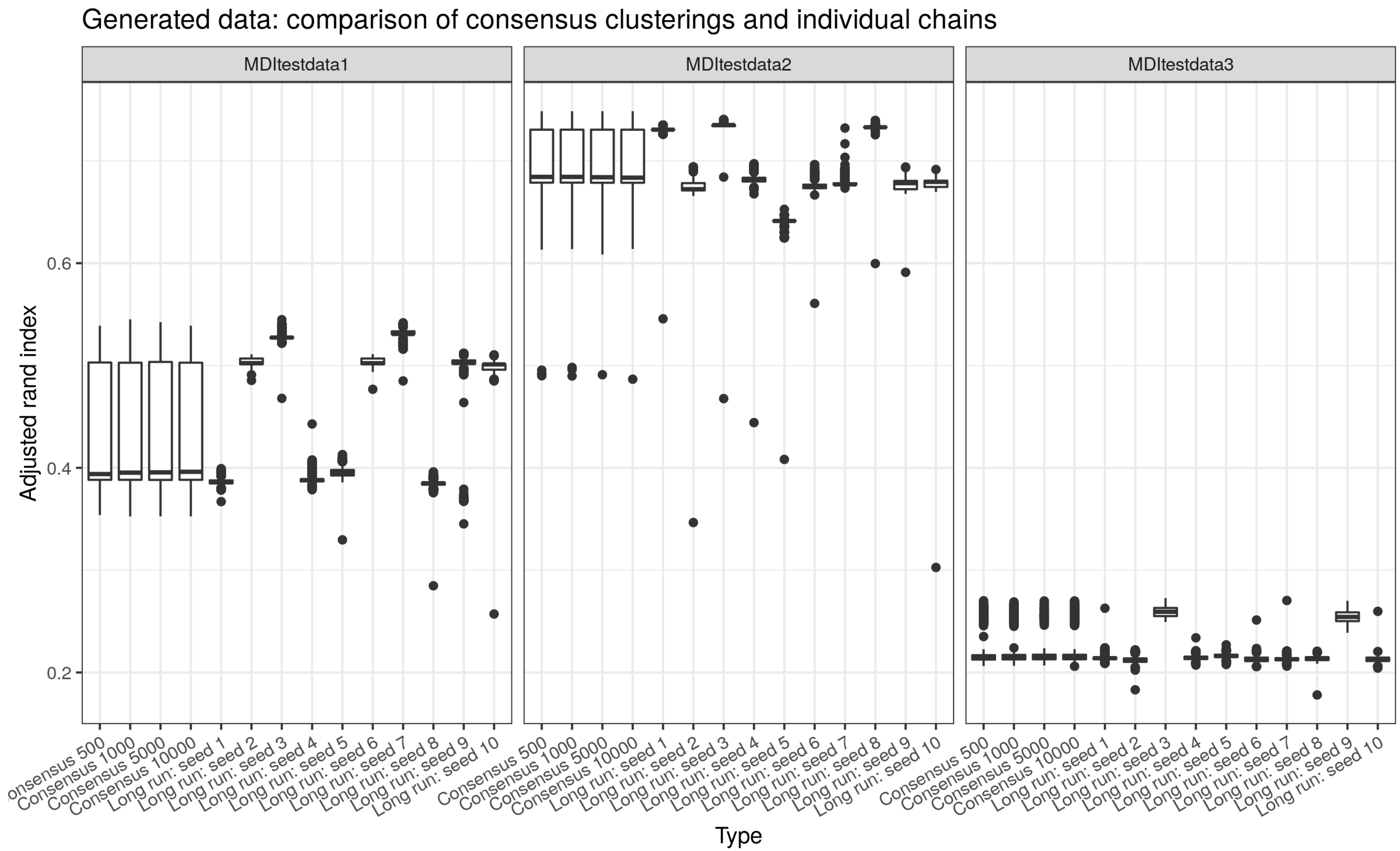

Figure 7: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and different initialisation of long chains.

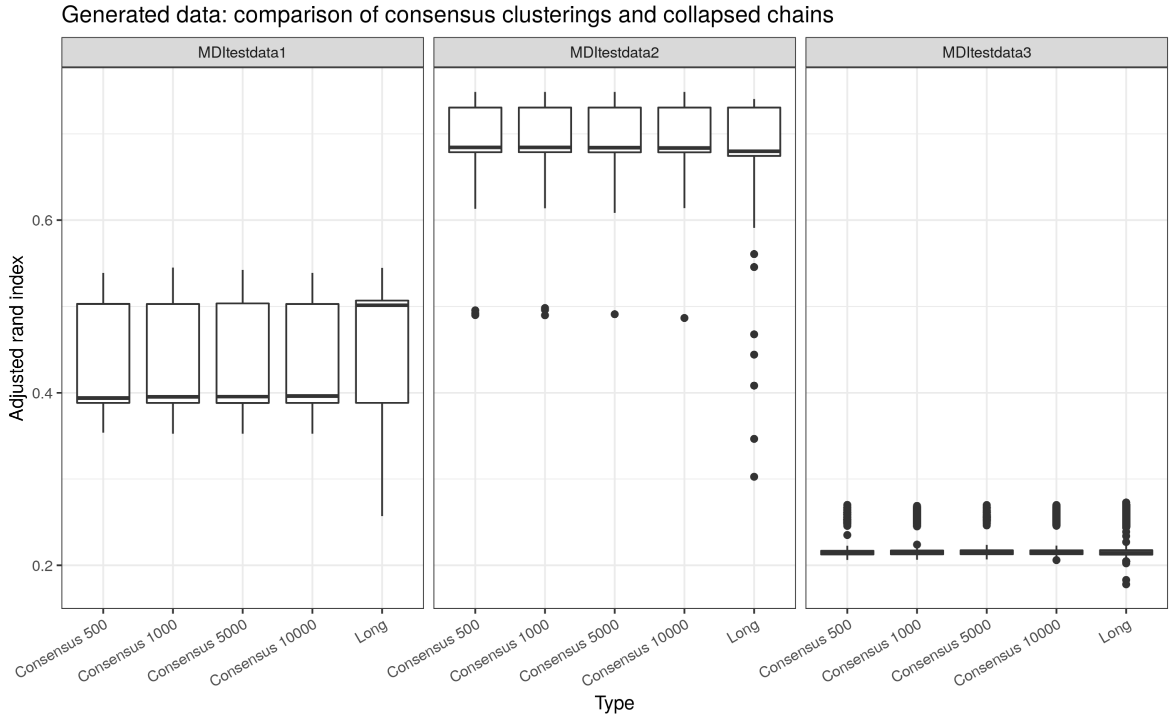

Figure 8: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

like proteins in inflammation and disease. *The Journal of Pathology*, 214(2): 136–148, January 2008. ISSN 00223417, 10969896. doi: 10.1002/path.2271.

[4] Frank Dudbridge. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3):e1003348, March 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003348.

[5] Brooke L Fridley and Joanna M Biernacka. Gene set analysis of SNP data: Benefits, challenges, and future directions. *European Journal of Human Genetics*, 19(8):837–843, August 2011. ISSN 1018-4813, 1476-5438. doi: 10.1038/ejhg.2011.57.

[6] Wendy S. Garrett, Jeffrey I. Gordon, and Laurie H. Glimcher. Homeostasis and Inflammation in the Intestine. *Cell*, 140(6):859–870, March 2010. ISSN 00928674. doi: 10.1016/j.cell.2010.01.023.

[7] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri,

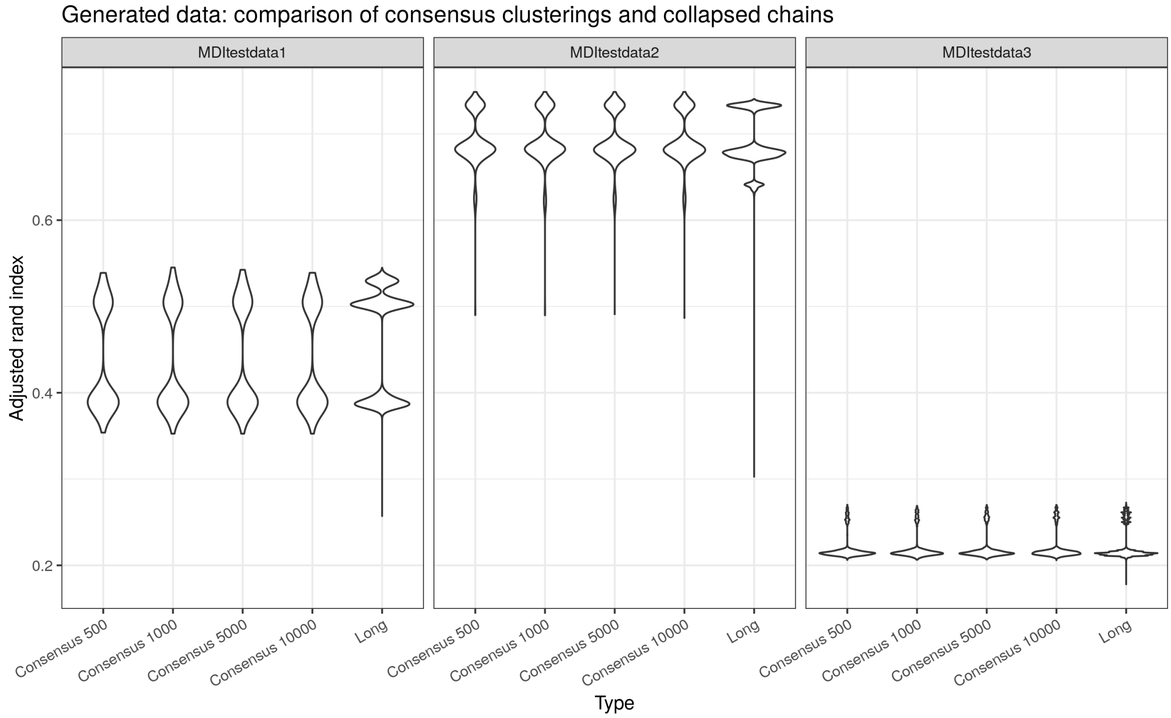

Figure 9: Violoin plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains. We can see that the consensus clustering approximates the modes described across chains quite well.

Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E Lowe, Paola Di Meglio, Stephen B Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M Lindgren, Krina T Zondervan, Kourosh R Ahmadi, Eric E Schadt, Kari Stefansson, George Davey Smith, Mark I McCarthy, Panos Deloukas, Emmanouil T Dermitzakis, and Tim D Spector. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, October 2012. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2394.

[8] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017. ISSN 0028-0836, 1476-4687.

doi: 10.1038/nature24277.

[9] Boris P. Hejblum, Jason Skinner, and Rodolphe Thiébaut. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. *PLOS Computational Biology*, 11(6):e1004310, June 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004310.

[10] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985. ISSN 0176-4268, 1432-1343. doi: 10.1007/BF01908075.

[11] Wenjun Ju, Casey S. Greene, Felix Eichinger, Viji Nair, Jeffrey B. Hodgin, Markus Bitzer, Young-suk Lee, Qian Zhu, Masami Kehata, Min Li, Song Jiang, Maria Pia Rastaldi, Clemens D. Cohen, Olga G. Troyanskaya, and Matthias Kretzler. Defining cell-type specificity at the transcriptional level in human disease. *Genome Research*, 23(11):1862–1873, November 2013. ISSN 1088-9051. doi: 10.1101/gr.155697.113.

[12] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky962.

[13] 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. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/bts595.

[14] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant,

Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2653.

[15] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. September 2003.

[16] T Maniatis, S Goodbourn, and J. Fischer. Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806):1237–1245, June 1987. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.3296191.

[17] Samuel A. Mason, Faiz Sayyid, Paul D.W. Kirk, Colin Starr, and David L. Wild. MDI-GPU: Accelerating integrative modelling for genomic-scale data using GP-GPU computing. *Statistical Applications in Genetics and Molecular Biology*, 15(1), January 2016. ISSN 1544-6115, 2194-6302. doi: 10.1515/sagmb-2015-0055.

[18] Stefano Monti. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. page 28.

[19] Michael A. Mooney and Beth Wilmot. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(7):517–527, October 2015. ISSN 15524841. doi: 10.1002/ajmg.b.32328.

[20] A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362–20120362, May 2013. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2012.0362.

[21] Chin-Tong Ong and Victor G. Corces. Enhancer function: New insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293, April 2011. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2957.

[22] William N. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[23] Richard S. Savage, Zoubin Ghahramani, Jim E. Griffin, Paul Kirk, and David L. Wild. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv:1304.3577 [q-bio, stat]*, April 2013.

[24] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0506580102.

[25] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, January 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1131.

[26] The International IBD Genetics Consortium, Yukihide Momozawa, Julia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, François Crins, Elisa Docampo, Mahmoud Elansary, Ann-Stephan Gori, Christelle Lecut, Rob Mariman, Myriam Mni, Cécile Oury, Ilya Altukhov, Dmitry Alexeev, Yuri Aulchenko, Leila Amininejad, Gerd Bouma, Frank Hoentjen, Mark Löwenberg, Bas Oldenburg, Marieke J. Pierik, Andrea E. vander Meulen-de Jong, C. Janneke van der Woude, Marijn C. Visschedijk, Mark Lathrop, Jean-Pierre Hugot, Rinse K. Weersma, Martine De Vos, Denis Franchimont, Severine Vermeire, Michiaki Kubo, Edouard Louis, and Michel Georges. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, 9(1):2427, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04365-8.

[27] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, Natalia Pervjakova, Isabel Alvaes, Marie-Julie Fave, Mawusse Agbessi, Mark Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez, Andrew A. Andrew, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg-Guzman, Johannes Kettunen, Joseph Powell, Bernett Lee, Futao Zhang, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, Harm Brugge, i2QTL Consortium, Julia Dmitrieva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T Heijmans, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Jonathan Pritchard, Olli Raitakari, Olaf Rotzschke, Eline P. Slagboom, Coen D.A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A.C. 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan Veldink, Uwe Völker, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon Pierce, Terho Lehtimäki, Dorret Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem H. Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Jian Yang, Peter M. Visscher, Markus Scholz, Gregory Gibson, Tõnu Esko, and Lude Franke. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. Preprint, Genomics, October 2018.

[28] Naomi R. Wray, Sang Hong Lee, Divya Mehta, Anna A.E. Vinkhuyzen, Frank Dudbridge, and Christel M. Middeldorp. Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10):1068–1087, October 2014. ISSN 00219630. doi: 10.1111/jcpp.12295.

[29] James Homer Wright. The histogenesis of the blood platelets. *Journal of Morphology*, 21(2):263–278, July 1910. ISSN 0362-2525, 1097-4687. doi: 10.1002/jmor.1050210204.

A Data generation explained

B Additional convergence plots

B.1 Case 1: Convergence diagnostics

B.1.1 Geweke plots

B.1.2 Estimated burn in

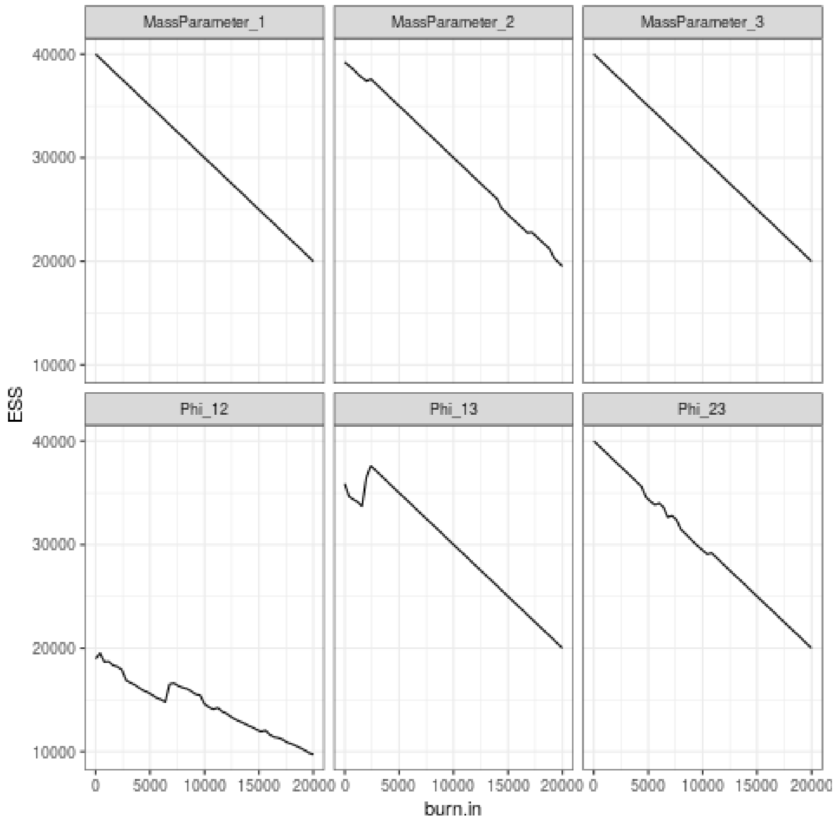

Figure 10: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

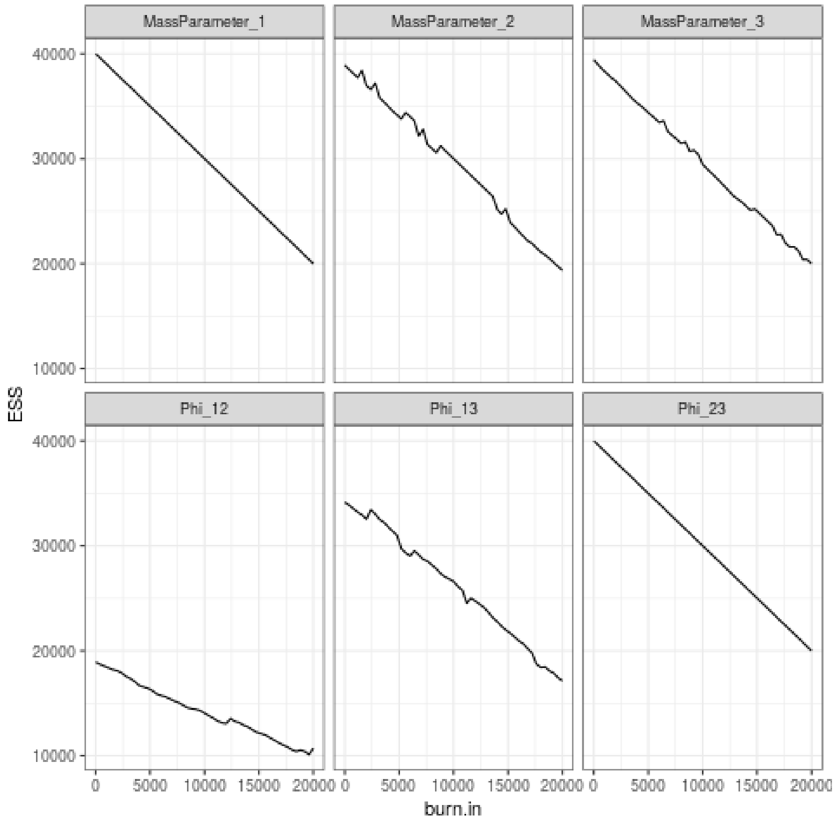

Figure 11: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

B.2 Case 2: Convergence diagnostics

B.2.1 Geweke plots

B.2.2 Estimated burn in

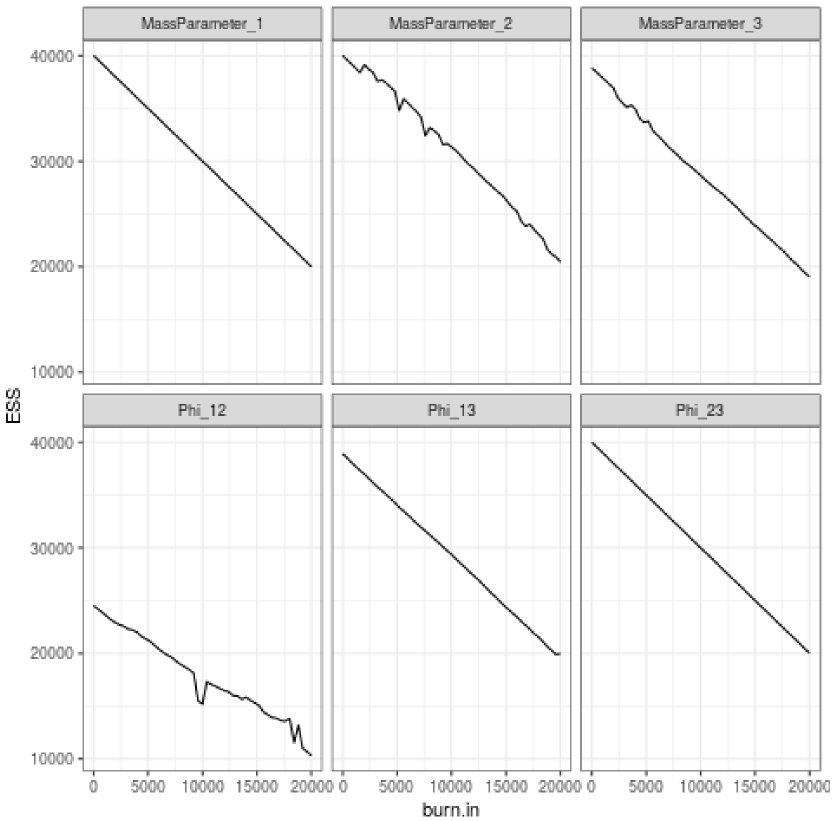

Figure 12: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

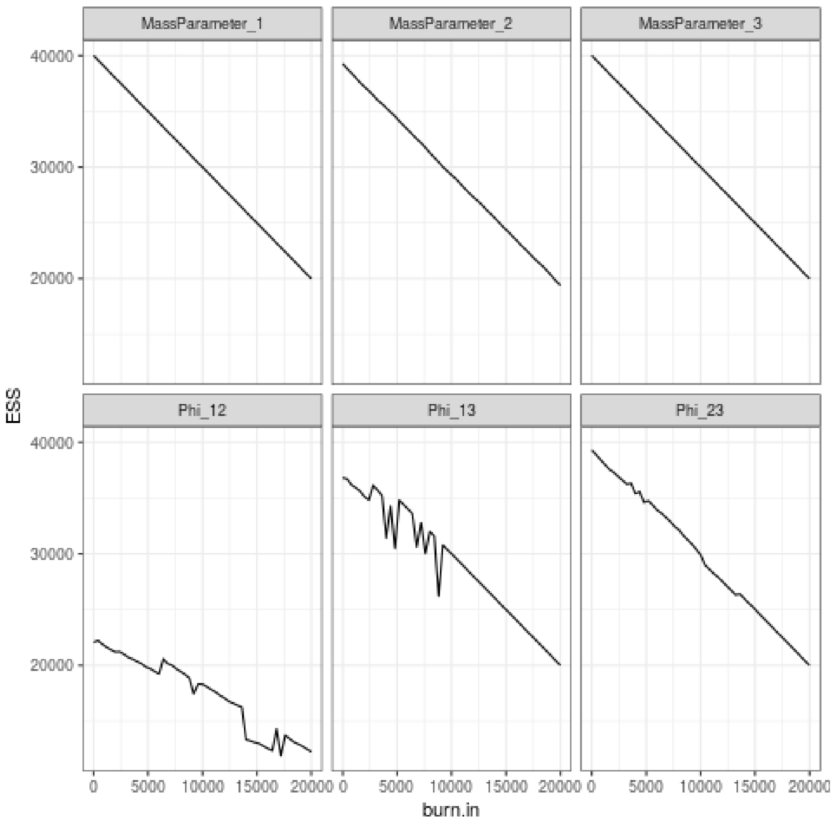

Figure 13: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

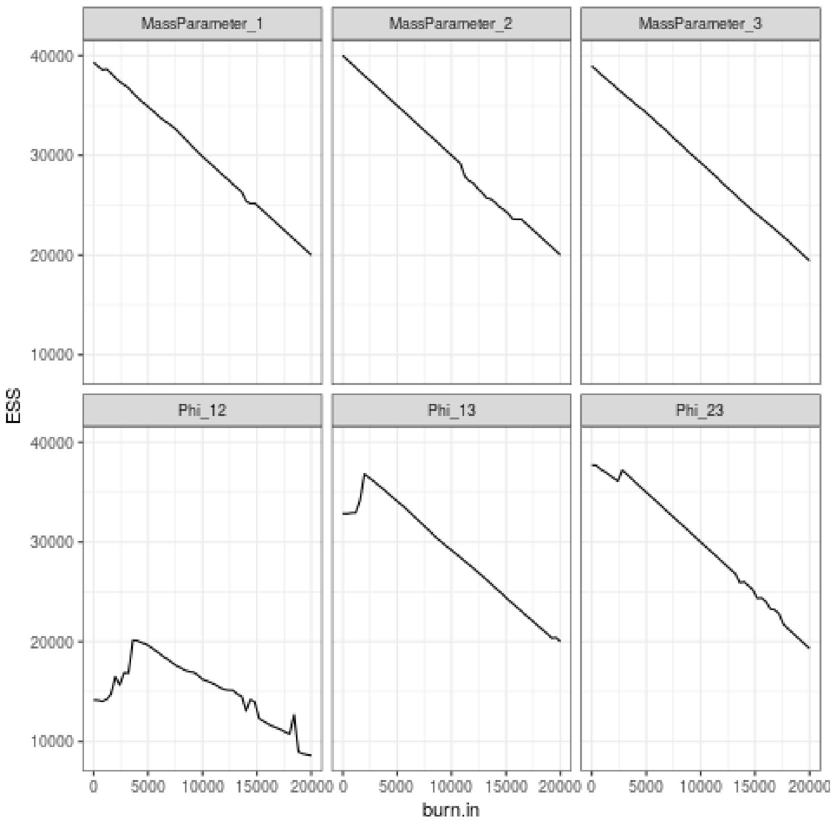

Figure 14: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

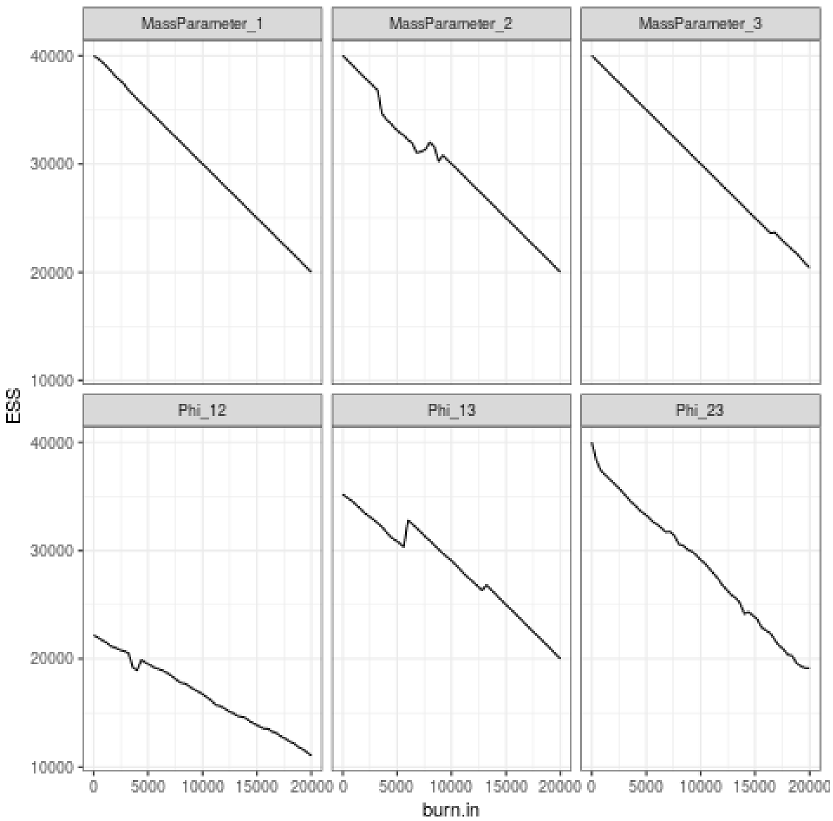

Figure 15: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

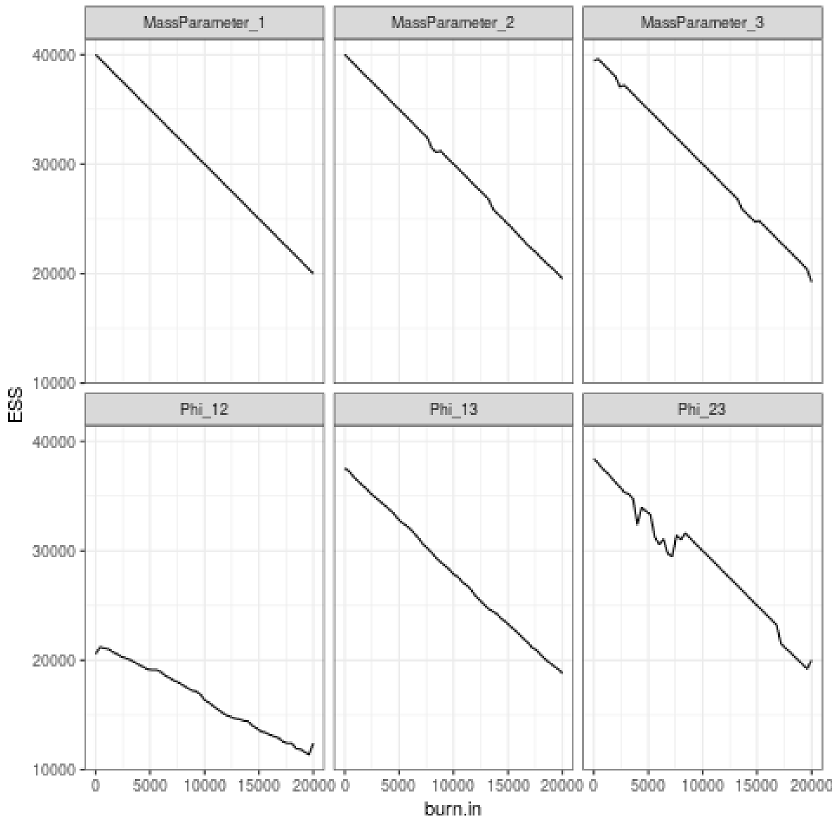

Figure 16: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

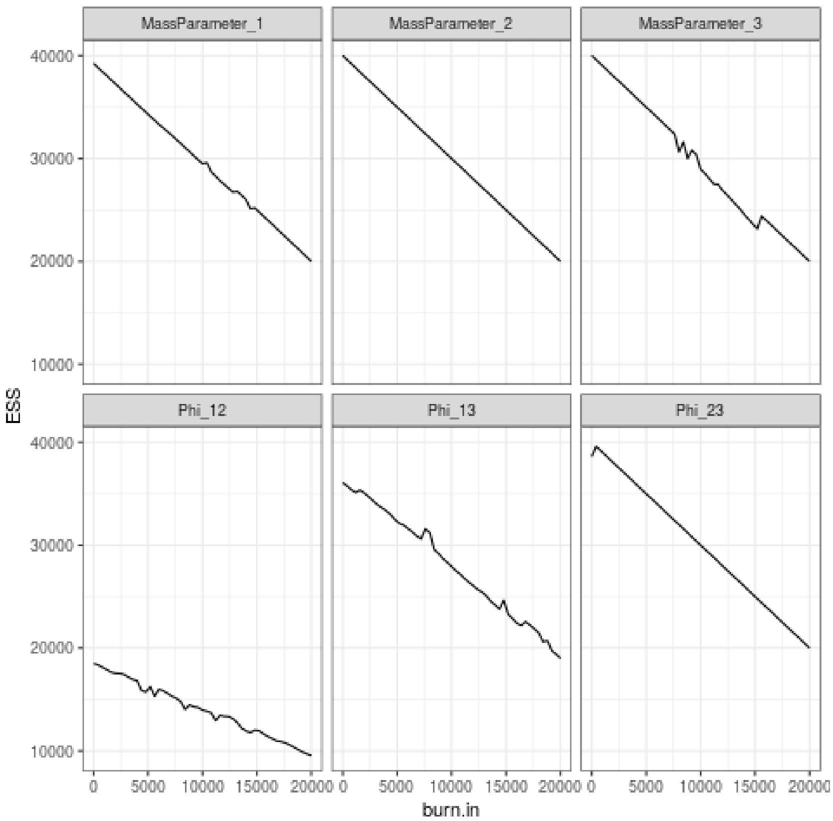

Figure 17: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

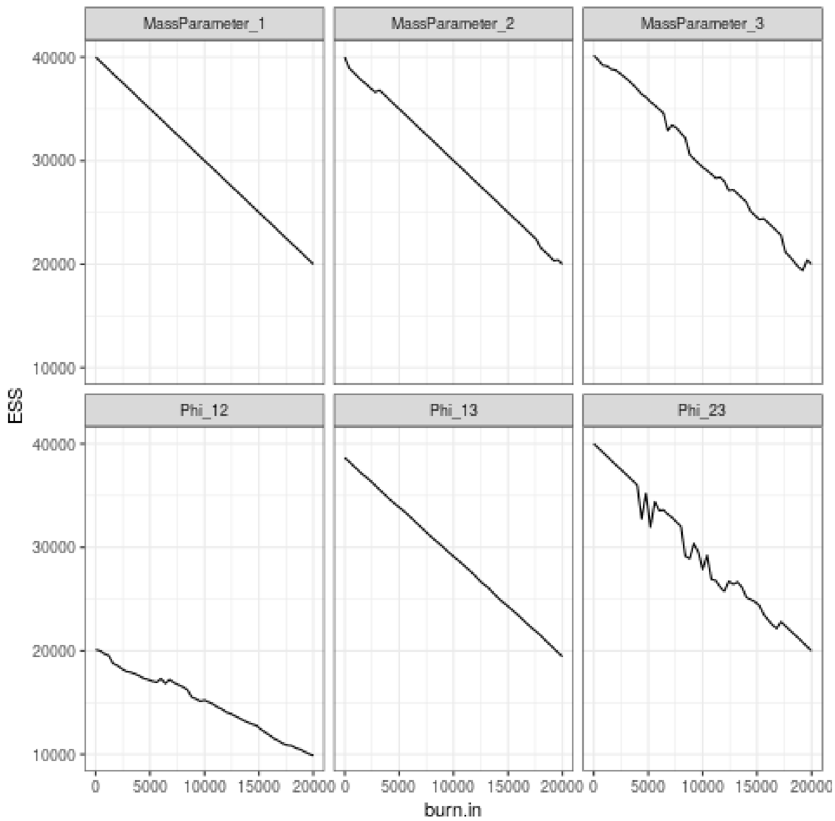

Figure 18: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

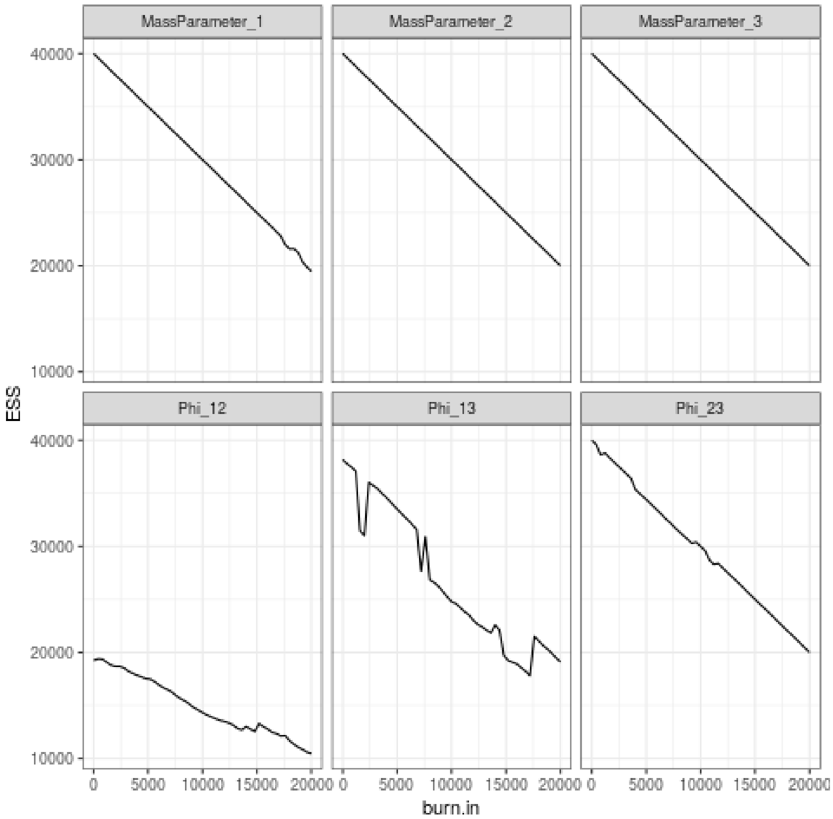

Figure 19: Box plots for distribution of adjusted rand index between the clustering at each iteration to the true clustering for different lengths of consensus clustering and the collapsed long chains.

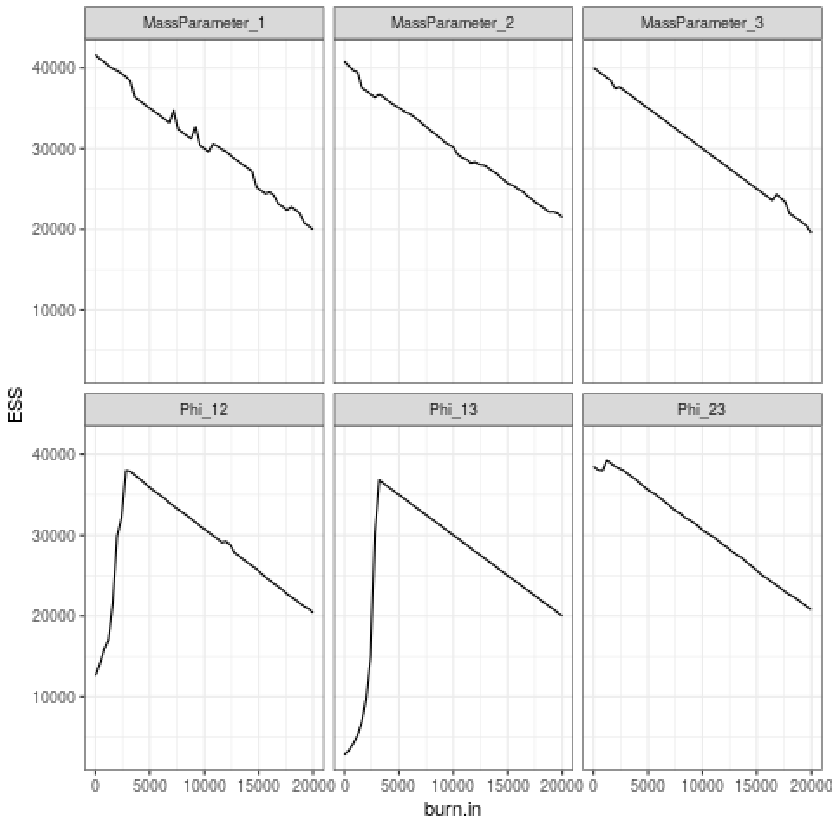

Figure 20: Plot of effective sample size (ESS) to burn-in for chain 1.

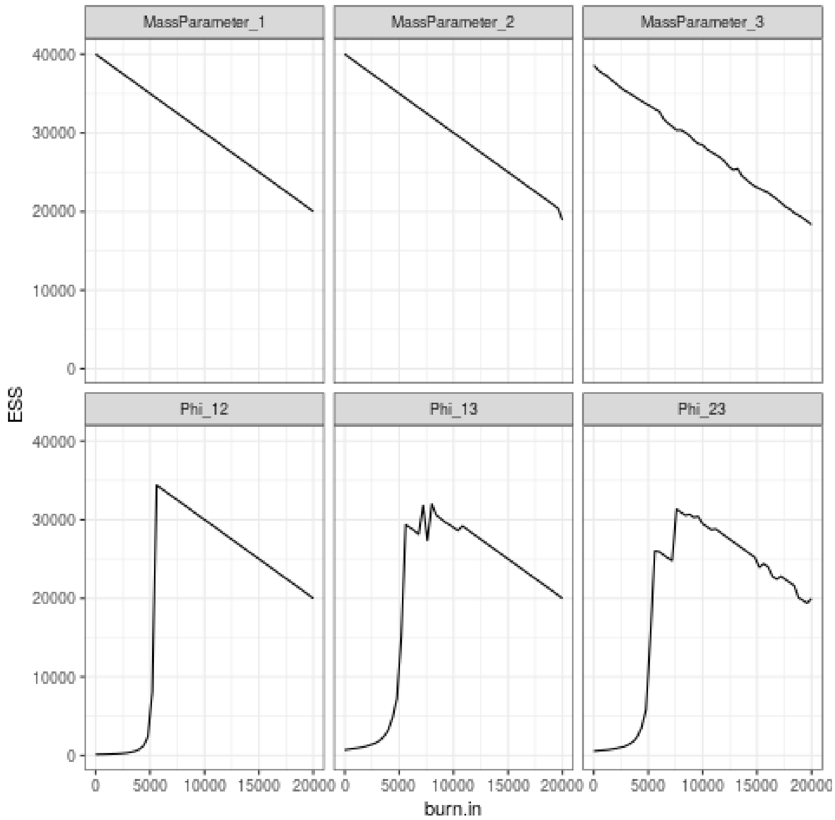

Figure 21: Plot of effective sample size (ESS) to burn-in for chain 2.

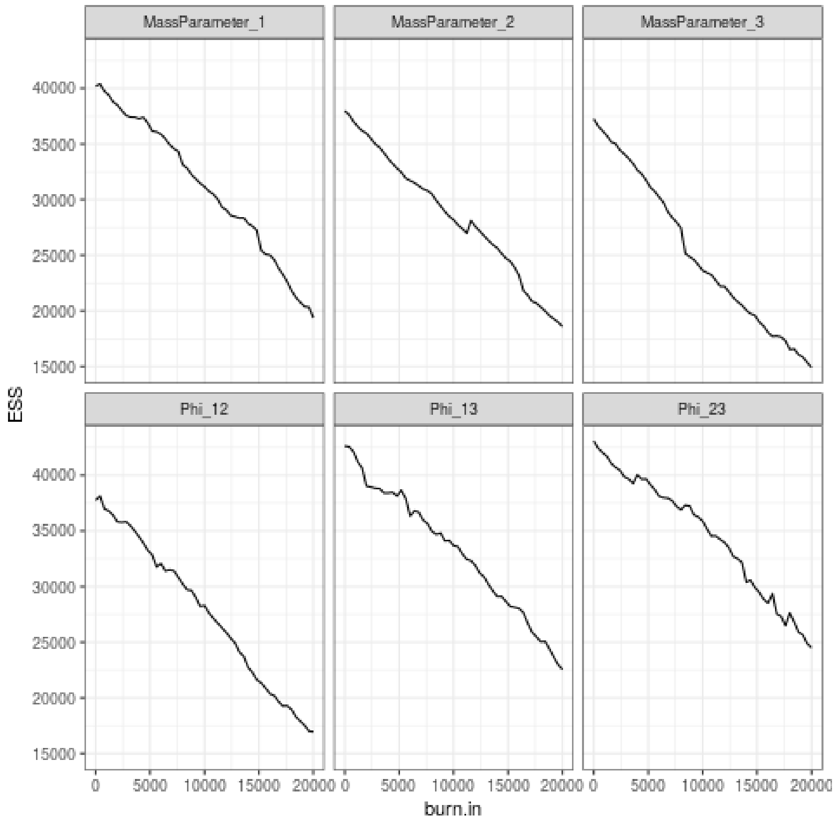

Figure 22: Plot of effective sample size (ESS) to burn-in for chain 3.

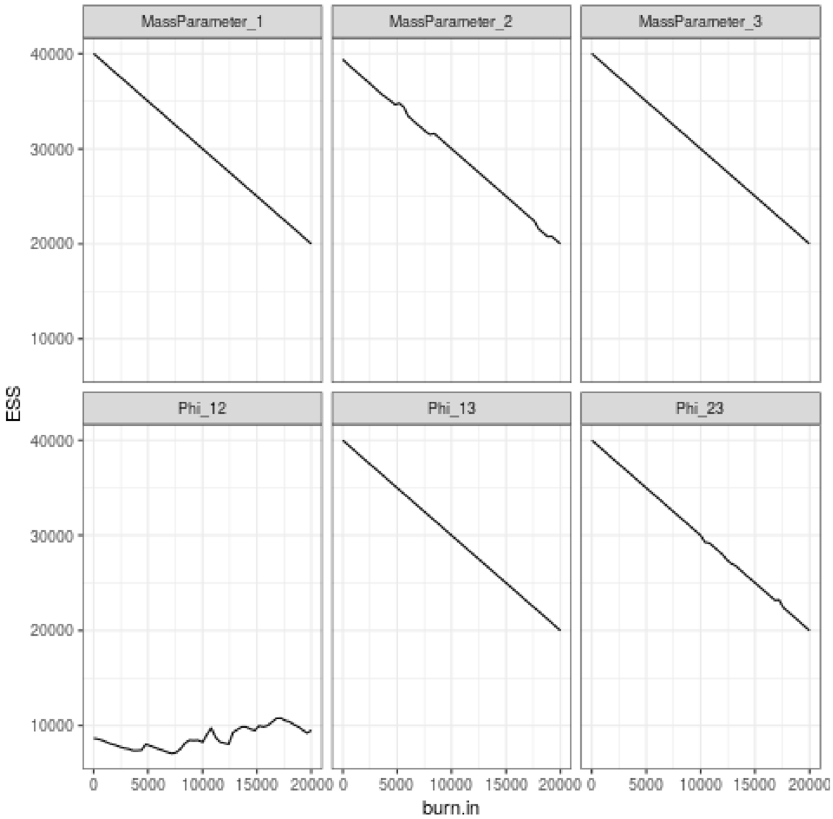

Figure 23: Plot of effective sample size (ESS) to burn-in for chain 4.

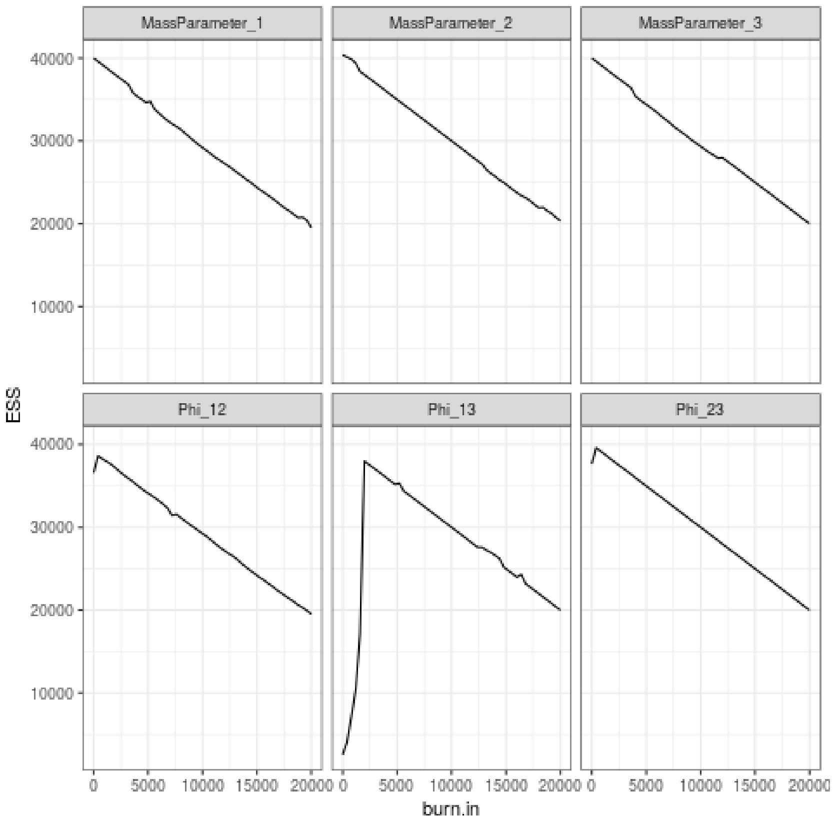

Figure 24: Plot of effective sample size (ESS) to burn-in for chain 5.

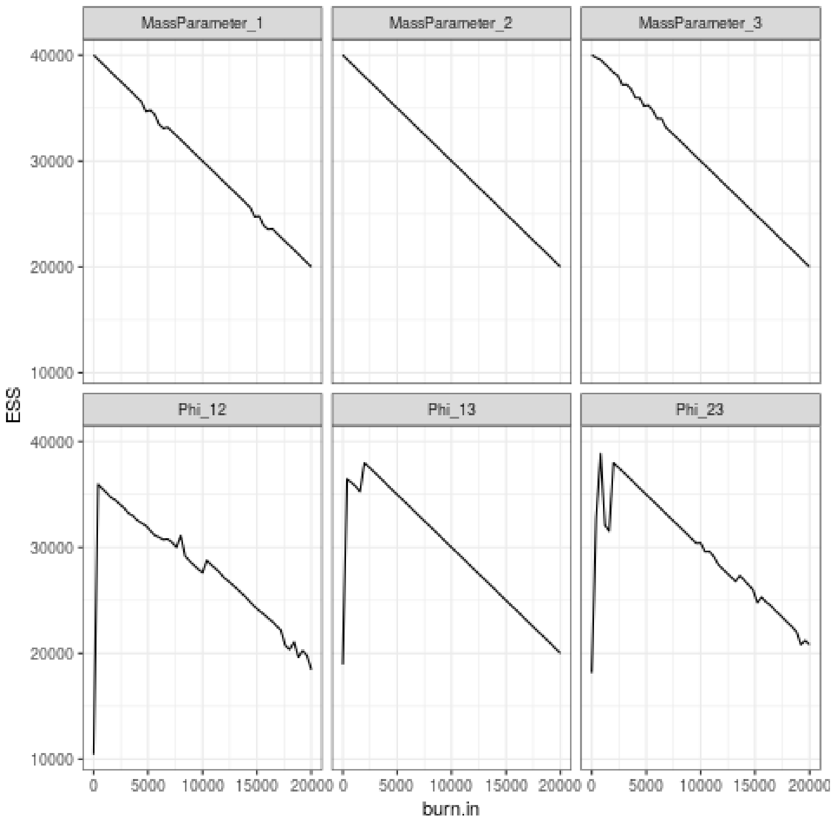

Figure 25: Plot of effective sample size (ESS) to burn-in for chain 6.

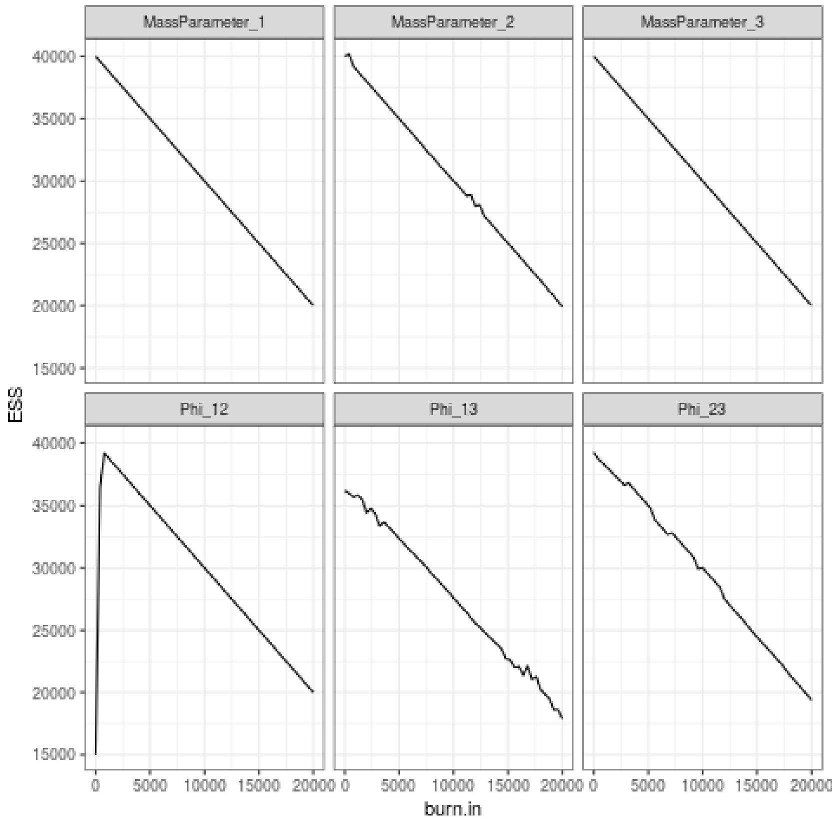

Figure 26: Plot of effective sample size (ESS) to burn-in for chain 7.

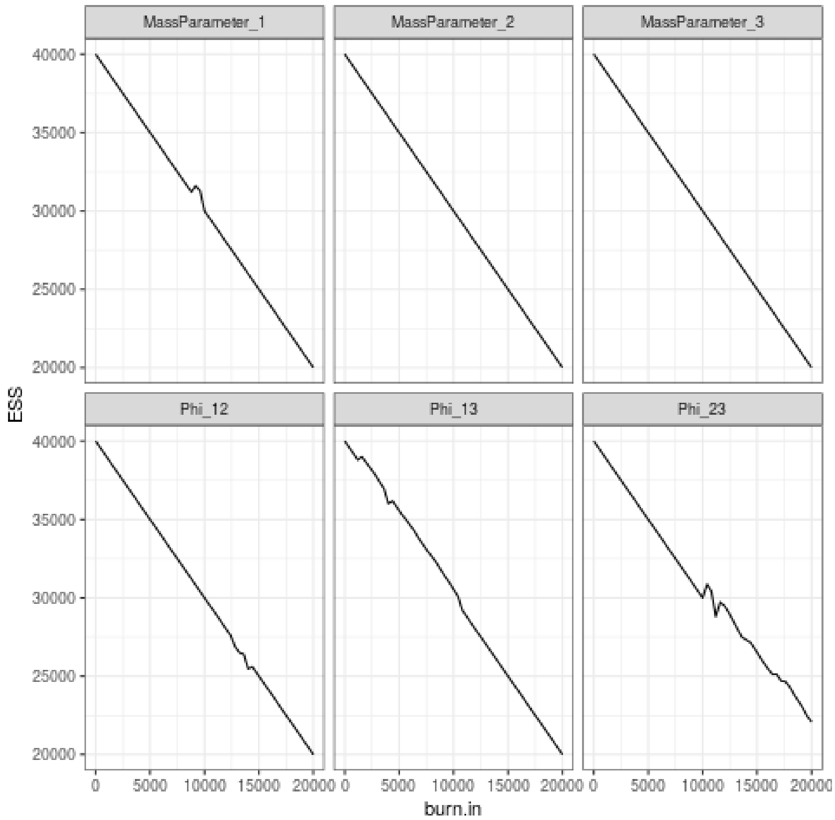

Figure 27: Plot of effective sample size (ESS) to burn-in for chain 8.

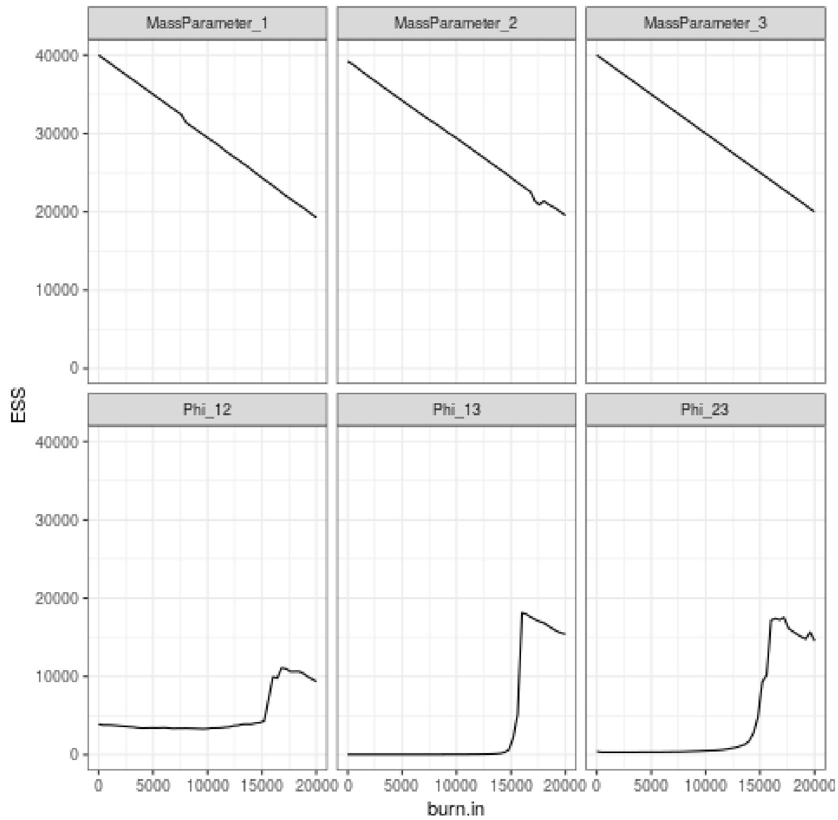

Figure 28: Plot of effective sample size (ESS) to burn-in for chain 9.

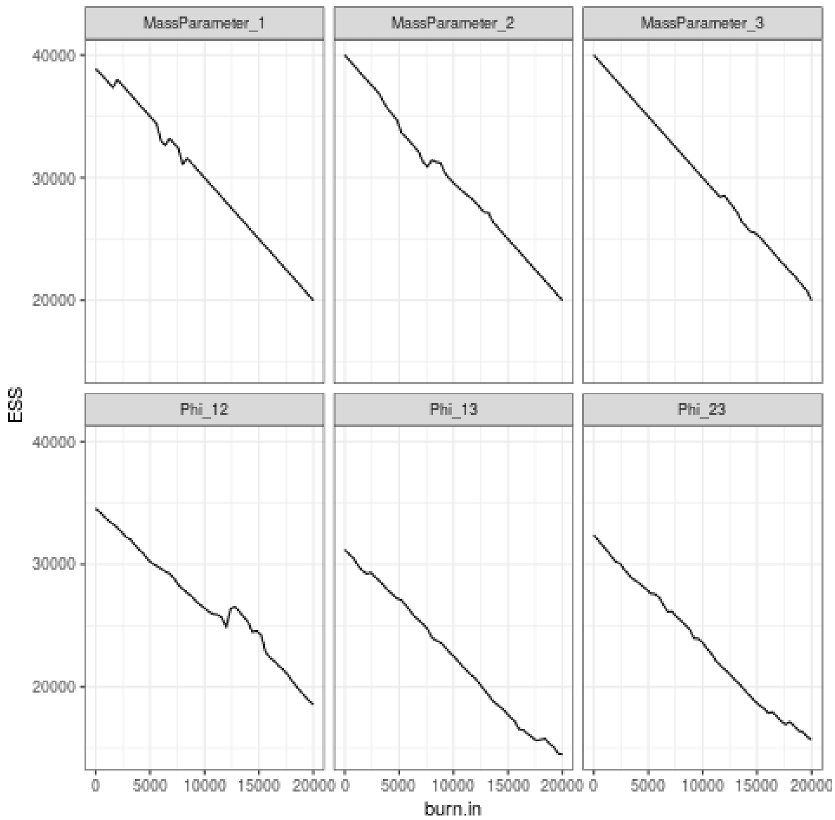

Figure 29: Plot of effective sample size (ESS) to burn-in for chain 10.