Heterozygote calling in significantly inbred populations

Rohan Shah, Alex Whan

November 28, 2018

1 Introduction

Realising the potential of large genetic resources requires the ability to perform genotyping efficiently. In populations with tens or hundreds of thousands of SNP markers, it is infeasible to assign genotypes manually; this must be done automatically. This problem is generally solved by the application of mixture models [Xiao et al., 2007, Teo et al., 2007].

We describe a marker calling method able to correctly identify marker heterozygotes, in large populations generated according to some experimental design; e.g. the Collaborative Cross [Threadgill and Churchill, 2012] or MAGIC [Huang et al., 2015]. Identifying marker heterozygotes is of substantial value in these populations, which are substantially inbred with a small amount of residual heterozygosity. If genetic lines can be identified which are heterozygous at a QTL locus, these lines can be used to develop a follow-up population suitable for fine-mapping of that QTL. However, as the amount of residual heterozygosity is low, these lines cannot be identified by an unsupervised method.

We describe a model-based clustering method for identifying marker heterozygotes in significantly inbred populations. Like Xiao et al. [2007] and Teo et al. [2007], our method is also based on mixture models. Unlike those methods, we include a constraint that the heterozygote cluster lies between the two marker homozygote clusters (Figure 1). This constraint accurately represents the structure of the heterozygote cluster in the case of data generated by the Illumina Infinium genotyping platform. The proposed model-based clustering algorithm was motivated by an eight-parent MAGIC population in spring bread wheat, genotyped using the Illumina Infinium genotyping platform, and we demonstrate its use in that population. The proposed method is implemented in the R package `magicCalling`.

2 Model

Assume that we have genotyping data at a single marker, for N individuals. This data takes the form of normalized hybridization intensities, which can be represented in a number of different coordinate systems. We assume every hybridization intensity is represented as a strength r and a contrast θ . As an example, every point in Figure 1 is a hybridization intensity.

Our model of this bivariate data is a four-component normal mixture model, with a hierarchical prior. The first two components correspond to the homozygous marker alleles, and have bivariate normal distributions $\mathbf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. The third component represents the marker heterozygotes, and has distribution $\mathbf{N}\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}, \boldsymbol{\Sigma}_3\right)$; that is, the third component lies between the two heterozygote clusters, and does not have its own independent mean component, although it does have its own covariance matrix. The fourth component represents errors, and has distribution $\mathbf{N}(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4)$. It is intended to be much flatter, and capture all points not represented in the first three components.

We now specify the priors for the mean parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_4$ and covariance matrices $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_4$ using multivariate normal and Wishart distributions. The priors for the mean parameters are

$$\begin{aligned}\boldsymbol{\mu}_1 &\sim \mathbf{N}((0.25, 1), \mathbf{D}_{hom}), & \boldsymbol{\mu}_2 &\sim \mathbf{N}((0.75, 1), \mathbf{D}_{hom}), \\ \boldsymbol{\mu}_4 &\sim \mathbf{N}((0.5, 1), \mathbf{D}_{err}),\end{aligned}$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_4$ are independent. Matrix $\mathbf{D}_{hom}$ is a hyperparameter for the homozygote clusters, and $\mathbf{D}_{err}$ is a hyperparameter for the error cluster. The priors for the covariance matrices $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_4$ are Wishart distributions:

$$\begin{aligned}\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 &\sim \text{Wishart}(n_{hom} \mathbf{V}_{hom}, n_{hom}), \\ \boldsymbol{\Sigma}_3 &\sim \text{Wishart}(n_{het} \mathbf{V}_{het}, n_{het}), \\ \boldsymbol{\Sigma}_4 &\sim \text{Wishart}(n_{err} \mathbf{V}_{err}, n_{err}).\end{aligned}$$

The matrix parameters in these distributions are the expected values of the covariance matrices. So for example, $\mathbb{E}[\boldsymbol{\Sigma}_1] = \mathbf{V}_{hom}$, and $\mathbf{V}_{hom}$ represents the expected shape of the covariance matrix. Similarly for $\boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$. The real parameters n_* represent the degree to which the covariance matrix distribution is concentrated around its expectation; a large value indicates a distribution highly concentrated around the matrix parameter. A small value indicates a flatter distribution, and greater a priori uncertainty about the shape of one of the clusters.

The last prior to be specified is the prior on the mixing proportions $\mathbf{p} = (p_1, p_2, p_3, p_4)$ of the four components. Our prior knowledge is that most lines will be homozygotes, due to inbreeding, and relatively few lines will be heterozygotes or errors. This means we have a prior belief that $p_1 + p_2$ should be much larger than $p_3 + p_4$. However *one of* p_1 or p_2 may be small, as the marker allele distribution is unknown. For example, in our target eight-parent MAGIC population, we will quite often have $p_1 \sim 0.125$ and $p_2 \sim 0.875$.

To account for this, we define our prior as follows. Let p_{hom} be the proportion of homozygotes, and let

$$(p_{hom}, 1 - p_{hom}) \sim \text{Dirichlet}(600, 30).$$

That is, most lines are homozygotes. Let p_{hom}^a be the proportion of homozygotes that fall into the first cluster, and let

$$(p_{hom}^a, 1 - p_{hom}^a) \sim \text{Dirichlet}(10, 10).$$

That is, there is no prior belief that one homozygote cluster is bigger than the other. Let p_{het} be the proportion of heterozygotes, among all lines that are

either heterozygotes or errors, and let

$$(p_{het}, p_{err}) \sim \text{Dirichlet}(2, 2).$$

That is, our prior belief is that any line that is not a homozygote is equally likely to be an error or a heterozygote. Combining all these priors, the mixing proportions are

$$\mathbf{p} = (p_{hom}p_{hom}^a, p_{hom}(1 - p_{hom}^a), (1 - p_{hom})p_{het}, (1 - p_{hom})p_{err}).$$

3 Implementation

Previous mixture model approaches have used the EM algorithm [Teo et al., 2007]. Our choice of conjugate priors has the advantage that standard software tools can sample from the the posterior distribution. We use the JAGS (Just Another Gibbs Sampler) package [Plummer, 2015] to perform Gibbs sampling for the model described in Section 2. Technically, our use of Gibbs sampling means that the resulting Markov Chain mixes poorly, as the two homozygote clusters will never swap. However, given our clustering context, this type of mixing is irrelevant.

For multi-parent experimental designs, the number of lines is generally large enough that the posterior variance for the model parameters is small, so significantly different parameter estimates from Markov chains with different starting points indicates MCMC non-convergence. This holds in particular for our motivating population, which has 3,412 lines. Because of this, our approach is to run the Gibbs sampler for a certain number of steps, and to use the parameter values at the final step of the Markov chain as estimated parameters.

Given a large enough number of markers, and therefore applications of the model, it is inevitable that some Markov chains will fail to converge. There will also be cases where the model is inappropriate, as the marker is either monomorphic, or has more than two alleles. These cases can be detected using heuristics.

To determine whether a fitted model is acceptable, it is useful to consider:

1. $\delta \Sigma_1^{-1} \delta$ and $\delta \Sigma_2^{-1} \delta$, where $\delta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. These represent the separation between the homozygote clusters; if these are small, the homozygote clusters are not well separated.
2. $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2$. If this is small, the homozygote clusters are not well separated.
3. The number of lines in the smaller of the homozygote clusters.
4. The number of lines in the error component.

Given multiple acceptable parameter estimates from multiple runs of the Markov chain, we select the estimates with the smallest value of $\max\{|\Sigma_1|, |\Sigma_2|\}$. This heuristic is effective, because chains that have become stuck in a local maximum and not converged, generally (incorrectly) group multiple clusters into a single (excessively large) cluster. This results in covariance matrices with large determinants. Markers with more than two homozygous alleles must be analysed using some other method, e.g. DBSCAN [Ester et al., 1996].

4 Example

We applied this genotype calling method to data for 90,000 markers in a wheat population of 3,412 lines, generated by an eight-parent MAGIC design [Shah et al., 2019]. We used 200 MCMC samples for each Markov chain, and ran eight Markov chains per marker. Our choice of hyperparameters was

$$\begin{aligned} n_{hom} &= 30, & n_{het} &= 1500, & n_{err} &= 300 \\ \mathbf{V}_{hom} &= \begin{pmatrix} \frac{1}{600} & 0 \\ 0 & \frac{1}{30} \end{pmatrix}, & \mathbf{V}_{het} &= \frac{1}{120} \mathbb{I}_{2 \times 2}, & \mathbf{V}_{err} &= \frac{10}{3} \mathbb{I}_{2 \times 2}, \\ \mathbf{D}_{hom} &= 4\mathbb{I}_{2 \times 2}, & \mathbf{D}_{err} &= \mathbb{I}_{2 \times 2}. \end{aligned}$$

These parameters were chosen by checking their performance on a subset of 200 markers. These parameters capture prior information in the following ways:

1. Our choice of $\mathbf{V}_{hom}$ implies that homozygote clusters have more variability along the r -axis than the θ -axis, however the small value of n_{hom} implies that this information is weak.
2. Our choice of $\mathbf{V}_{het}$ implies that the heterozygote cluster is small. The large value of n_{het} implies a very strong prior belief.
3. The error cluster is large (flat distribution), with a moderate degree of certainty.

To achieve good results with few MCMC steps, the Gibbs sampler must have a sensible starting point. We initialized the r coordinates of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to $\bar{r} = N^{-1} \sum_{i=1}^N r_i$. We considered a model acceptable if all the following held:

1. $\delta \boldsymbol{\Sigma}_1^{-1} \delta < 25$ and $\delta \boldsymbol{\Sigma}_2^{-1} \delta < 25$, where $\delta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.
2. $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 > 0.06$.
3. Each homozygote cluster contained more than 200 lines.
4. The proportion of lines in the error component was smaller than 0.25.

The ‘best’ chain was selected as the chain with the smallest value of $\max\{|\boldsymbol{\Sigma}_1|, |\boldsymbol{\Sigma}_2|\}$. A correctly called marker is shown in Figure 1. Of 90,000 markers, only 6,743 needed to be manually reviewed. Some of these reviewed markers were determined to have more than two marker alleles, and were called using DBSCAN. Others were cases of MCMC non-convergence, for which the MCMC algorithm was run again.

References

- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.
- B. Emma Huang, Klara L. Verbyla, Arunas P. Verbyla, Chitra Raghavan, Vikas K. Singh, Pooran Gaur, Hei Leung, Rajeev K. Varshney, and Colin R. Cavanagh. MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics*, 128(6):999–1017, 2015.

Figure 1: Example of the hierarchical Bayesian clustering approach, applied to data from an 8-parent MAGIC population. Filled points represent lines classified as homozygotes. Crosses represent lines classified as heterozygotes. Hollow points represent outliers, or points not otherwise classified. Ellipses represent covariance matrices of components.

- Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2015. URL <https://CRAN.R-project.org/package=rjags>. R package version 4-4.
- Rohan Shah, B. E. Huang, Alex Whan, Klara Verbyla, M. K. Morell, and C. R. Cavanagh. Dissection of wheat genome structure through novel genetic resources. 2019.
- Yik Y. Teo, Michael Inouye, Kerrin S. Small, Rhian Gwilliam, Panagiotis Deloukas, Dominic P. Kwiatkowski, and Taane G. Clark. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23(20):2741, 2007.
- David W. Threadgill and Gary A. Churchill. Ten years of the collaborative cross. *Genetics*, 190(2): 291–294, Feb 2012.
- Yuanyuan Xiao, Mark R. Segal, Y.H. Yang, and Ru-Fang Yeh. A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, 23(12):1459, 2007.