Heterogeneous Sub-Population Proportion Estimations of Immune Cells

Benjamin LeRoy
Advisor: Max G'Sell
Department of Statistics
Carnegie Mellon University
bpleroy@andrew.cmu.edu

Abstract

The associations between genetic markers and immune system response has be well documented, as has the association of proportions of immune cell types and immune system response [2, 7]. Recent methods to create highly genetically distinct sets of mice in 2004 and the development of mass cytometry in 2010 has provided collaborators with the ability to attempt to understand links between genetic markers and proportions of different cell types [3, 6] . This diverse dataset also provides complications is missing data, stemming from certain genetic strains not allows certain protein markers to bind on any cell for certain mice. We set out solving this problem by extending analysis ran on mice without missingness to those with missingness. To do so we develop procedures to standardize variable expression, and estimate proportion of cell types.

1 Introduction

The associations between genetic markers and immune system response has be well documented, as has the association of proportions of immune cell types and immune system response [2, 7]. Proportions of immune cell types are obtained after the classification of immune cells through gating based on protein expressions of the cell. It is thought that the quantity of different types of protein receptors of an individual cell captures information about the processes inside and goals of the cell.

These associations between immune cell proportions to immune system response and genetic makeup of the individual to immune system response naturally begets the question: "Are these two things, immune cell proportions and the genetic makeup of the individual also associated?". In order to analyze the potential for associations between genetics and proportions of different types of immune cells, our collaborators have collected a large set of protein expression data for individual cells across a range of very genetically different mice using mass cytometry (see Section 2 for more details).

Our collaborators' use of genetically diverse mice instead of the standard genetically similar mice was done to be able to look at more genetic structure and relationships to proportions of cell types. As the genetics of our collaborators' mice are much more diverse, we observe instances where a specific mouse's cells' proteins don't bind correctly to the indicator antibody. The anibodies used were developed with the standard genetically similar mice in mind. As poor binding occurs for all of the mouse's cells, we are unable to apply the standard manual gating to classify cells for a large proportion of the data, which directly leads to an inability for immunologists to classify cells and therefore estimate proportions of different cell types. In order to correct for missing data, we leverage the high dimensional protein expressions per cell from mass cytometry to estimate cell types and use mixture fitting on class membership probabilities to estimate proportions of different cell types.

To truly extend our classification approach to the cells that are missing certain protein expresses we have to address the differences in the recording of data across each mouse (which we refer to as a "batch"). Some of the batch effects, specifically created by the mass cytometry machine can be corrected with normalization techniques (see Section 2 for more details). Across these batches we also see other batch differences appear mainly as linear shifts in hyperbolic sinusoidal transformation of the data. Because we develop techniques to extend cell class proportion estimation and cell classification across mice these batch effects create an interesting challenge that also needs to be addressed.

2 Dataset

Our data is a collection of cell level recordings for 84 unique mice ranging from 38 different genetic strains. These mice's genetics come for 8 founder strains and provided large genetic differentiation's between groups using techniques developed in [3]. The number of cells scanned per mouse ranges from 1,260 to 48,130 with a median of 12,760 cells scanned (only 2 with less that 3,000 cells).

Each cell was scanned use mass cytometry, a technique developed in 2009 [1, 6]. This technique is able to measure protein expressions for individual cells and allows scientists to collect around 40 protein expression levels compared to other techniques that allow for much more limited collect of approaches, like flow cytometry that tends to be only able to capture at most 10 protein expression levels [6]. This gain in number of features is accomplished by using rare metals to serve as markers for different proteins of the cell. As visualized in the top visual of Figure 1, mass cytometry works by staining each cell with antibodies that are bound to rare metals. These antibodies have been developed to bind to specific proteins on the outside of the cell. The machine the vaporizes the cells, removes biological components and just leaving just the rare metals. These particulars are projected across a electromagnetic field which perturb the projection of different rare metals, and due to the differences in molecular mass, the proportion of each metal is recorded at a specific location on a receptor.

2.1 Protein Expression Collection

Inside the mass cytometry machines, two things occurs before scientists obtain the data. First, the machines that collect the protein expressions record a lack of observed markers (zero expression) by randomly selecting a value between 0 and -1 uniformly. This was done to help biologists explore simple distribution structure and it conforms with output of other method's expressions. Secondly, to correct for natural degradation of the intensity of the protein expression reading from the cell vaporization obtained from the machine, immunologists use bead normalization to make the data more consistent. This involves having collected a bead reading initially (which is marker that should be consistent in reading across scans and batches) and then apply a scaling to the protein expressions that mirrors a correction in intensity. Our collaborators used beads and normalization procedures that mirror work explained in [4].

2.2 Missingness

The antibodies used in mass cytometry a developed to bind to a part of a protein on a cell's surface [1]. With a more genetically diverse dataset than normally used, certain genetic strains of mice have different enough proteins that, although one expects the cells to work the same, certain types of antibody expected to bind to them. This causes noise and non-useful readings for certain proteins from the mass cytometry procedure, visualized in Figure 1, and means that for some mice we must approach certain proteins as if we failed to record amounts for all cells of certain mice, visualized in Table 2.

cell idx	**Ly6g expression**	**Ly6c expression**	**...**	**CD44 expression**	**CD43 expression**	**CD45R expression**	**...**
1	0.74	0.99	...	0.91	0.17	0.100	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1053	1.67	0.98	...	0.40	1.5	0.51	...
1052	0.85	1.13	...	0.99	1.94	0.91	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Data Example for Mouse without any missingness

cell idx	**Ly6g expression**	**Ly6c expression**	**...**	**CD44 expression**	**CD43 expression**	**CD45R expression**	**...**
1	0.43	X	...	0.95	X	0.96	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1053	1.55	X	...	0.97	X	0.74	...
1052	0.66	X	...	0.94	X	0.96	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2: Data Example for Mouse with missingness

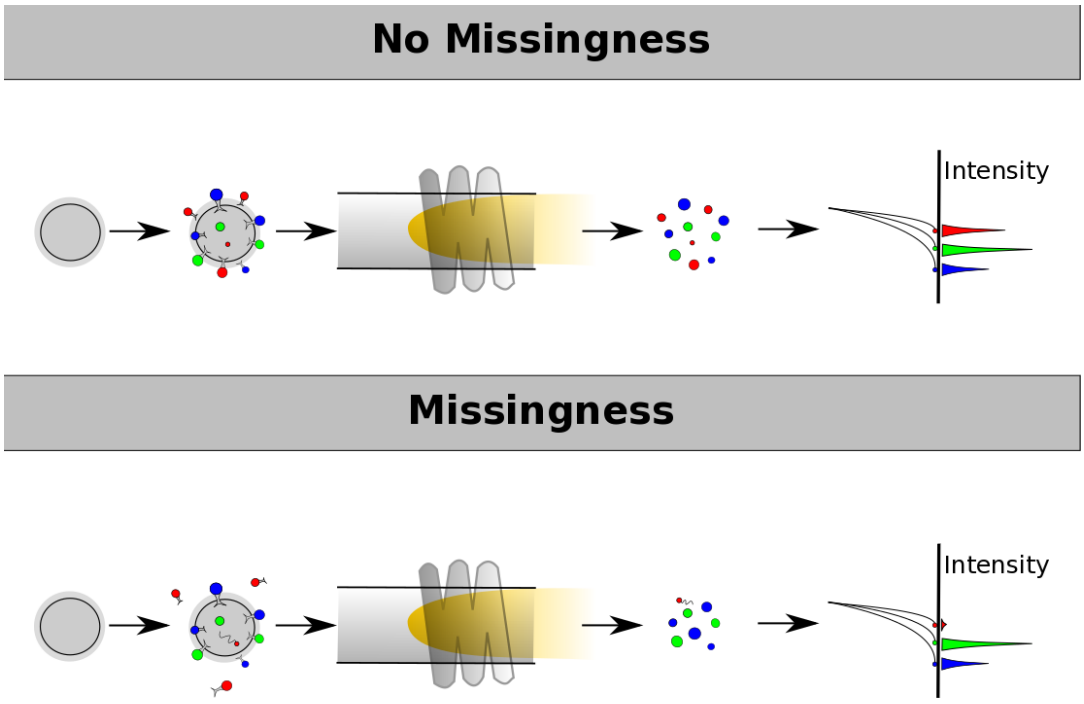

Figure 1: Mass Cytometry Procedure visualized for cells from mice that do not have any missing and for those mice that observe missingness of specific protein markers

3 Methods

back ref good

Due to the missingness in our dataset we set out to provide a way to estimate proportions of cell types that are usually defined by manually gating using a missing protein expression. Manual gating, the way immunologists classify cell types, is usually done on a 1 or 2 dimensional joint distribution level creating blocks of cells to be defined as certain times, Figure 2 provide a motivating example for gating B cells. In order for our collaborators to do downstream analysis relating genetics of mice to the immune cell proportions the methods below set out to provide estimates for the proportions of cell types (specifically Early and Late B cells) that are unable to obtained from manual gating due to mice missing the CD43 protein expression.

togh to parese

3.1 Batch Correction

3.1.1 Gaussian Mixture: Linear shift

Even with bead normalization mentioned in Section 2, across each batch the cells' protein expressions do not well align. To correct for distinctions of distinctions across batches due to time of stain, time of scan, and mixture dilation factor, which can be assumed to be a consistent effect for each batch, we attempt to model these alterations as a linear shift in the protein expression. This decision is motivated by exploratory data analysis but the full model we use is motivated by biological concerns and is described below.

We attempt to correct for potential linear shifts by by fitting a specialized Gaussian Mixture for each batch grouping of the form

$$X[c] \sim \sum_{k=1}^{K_c} \pi_{mk} \mathcal{N}(\mu_k + \delta_m, \sigma_k^2)$$

Where $X[c]$ is the protein expression of the protein with index c for a single cell, k is the index of the Gaussian mixture and m is the index to account for different batch effects.

Under this model, we allow different proportion π_{mk} of the specific Gaussian density specifically to allow for mice have different proportions of certain cell types directly related to the assumptions that different mice have different

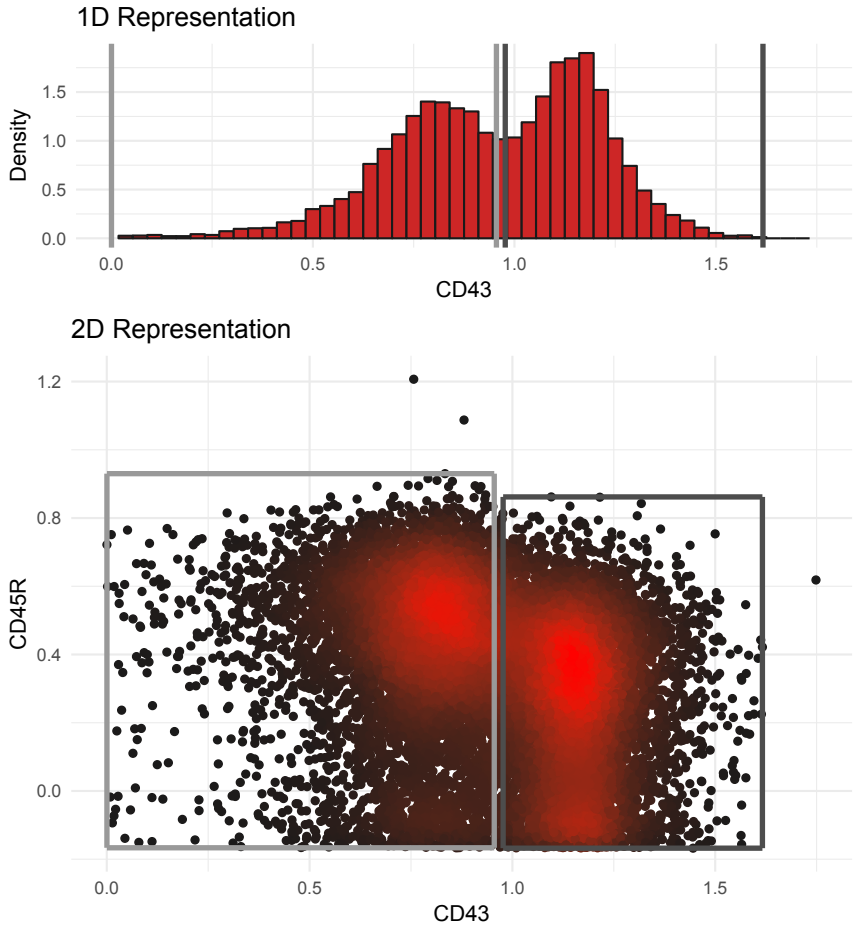

Figure 2: Manual gating schemes for Early and Late B cells are usually done on the CD43 and CD45R expression levels. The different color grey boxes and lines seperate the two cell types. In our dataset some mice are missing CD43.

proportions of cell sub-populations. This that the component weights vary considerably across components, which would confound simple mean-based alignment. The model allows for a single linear δ_m shift for each batch effect group. This decision is based of understanding that manual gating schemes tend to focus on particular parts of a distribution, like a trough in the density as see in Figure 2. After fitting this constrained mixture model across all mice for for a single protein expression (conditional on the number of observed modes), we subtract δ_m from each cell's protein expression for mouse m. This process is see in Figure 3.

3.2 Estimation of Cell Classes via Random Forest

In order to estimate cell types without certain protein expression crucial for human classification we attempt to replicate human classification approach and uncertainty. In order to mirror manual gating schemes that look at 1 or 2 protein expressions at a time we trained a random forest on a single mouse and test it on other mice. Specifically we created a forest to classify cells as either Early Bcell, Late Bcell or Uncategorized given a cell was classified as a Bcell, without protein expressions that are missing for any mouse. This was done on the cellular data after protein expressions corrected by the batch corrections. Additionally for 3 protein expressions **(mention which)** we discretize the protein expression into 3 groups (0, 1, 2) and have the random forest use these order to create the trees. This was done when we observed non-guassian structure and that the original trained random forest so nice breaks in the marginal distribution of classifications on these variables also encouraged such splits **(Max - what is the name of this again?)**.

3.3 Estimating Proportions of Cell Sub-Populations

In order to estimate the proportions of Early Bcells and Late Bcells conditional on cells being a Bcell we modeled the log-odds of the probabilities of being an Early Bcell from the random forest as a mixture of two Gaussians, with no

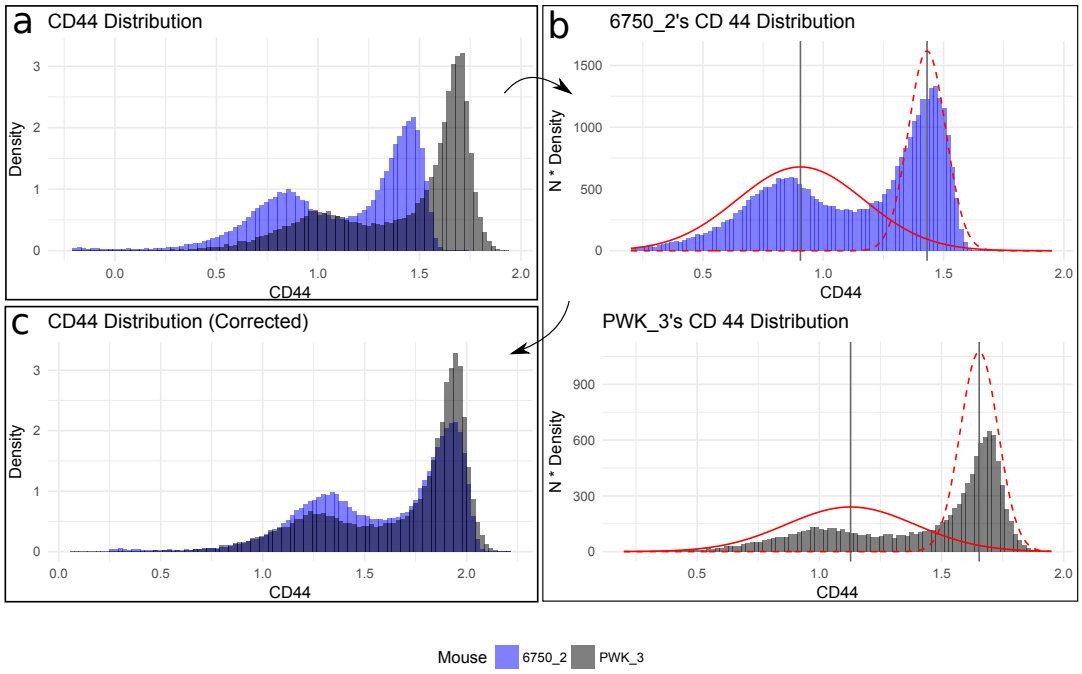

Figure 3: (a) Unaligned distributions of CD44 in two mice. (b) A two-component mixture model fit to both mice, with the component mean spacings and the component variances identical, but an overall mean shift allowed. (c) Aligned distributions resulting from this mixture-based approach.

constraints on this mixture model. After fitting this model we use the proportions of each cell class as our estimate of the proportion of the cell types, with the mixture with the lower center's proportion for the proportion of Late Bcells and the mixture with the higher center's proportion for the proportion of Early Bcells.

From these estimates we using those mice with no missingness (known as the training mice), that did not have the random forest trained on them to understand how much error we tend to have. Our collaborators use the empirical distribution of these errors for downstream comparing the proportions with genetic distributions.

4 Results

As our analysis was focused on subclasses of B cells, we only looked at the cells that were classified as B cells. All training mices' B cells were broken into a 60/40 split for training and test sets, although for the final estimation and checks we used all the B cells of the mice. We explored different models with 7 mice, denoted by 3609_1, 8049_1, 6557_1, 6012_1, 6211_1, 8043_1, and 18018_1.

Using ROC curves and AUC we selected mouse 3609_1 to be the mouse of the final prediction model was based of off. Before batch corrections 3609_1 also preformed well and seemed the most extendable.

4.1 Batch Corrections

To analyze extendability we examined the performance of random forest models, focusing on 1) if the ROC curves from the random forest trained on 60% of 3609_1's B cells and tested on 40% of each individual mouse's B cells shared similar structure and 2) if the ROC curves and AUC values for random forest trained on 3609_1 applied to each individual mouse and ROC curves and AUC values from random forest trained and tested on that same individual mouse is similar, with the hope that the random forest based on 3609_1 was better. Before transforming of the protein expression we observe, as seen in Figure 4, we observe ROC curves that, although they are pretty discriminate, do not share similar structure. Specifically, if you focus on the curves from testing on 3609_1 and 8043_1 you'll observe very different curves than that from testing on 6557_1.

Mouse 6557_1 is a good example of the problems the algorithm has before batch corrections. The plots of Figure 5 shows the ROC curves from a model trained on 3609_1 and one trained on 6557_1 in green and black respectively. The left plot shows random forests using the non-batch corrected data, and we can observe that there is a significant

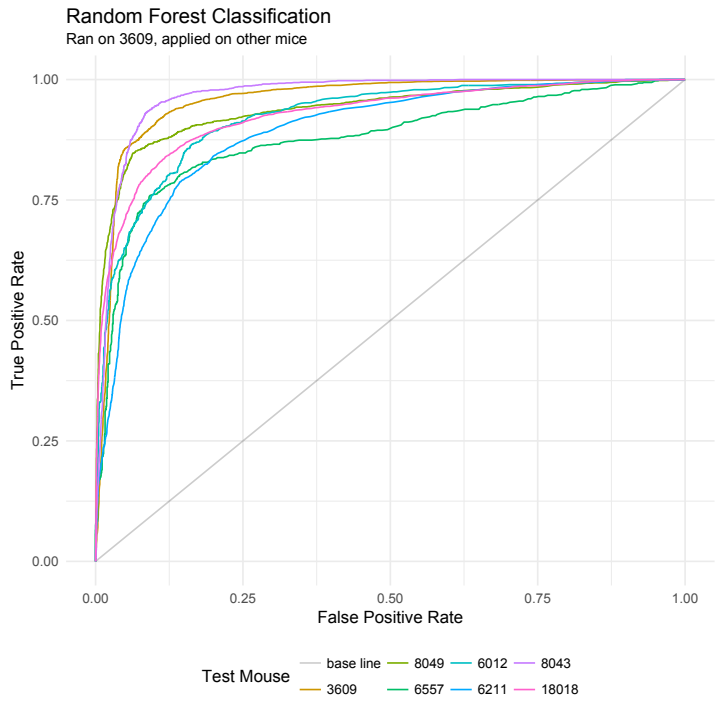

Figure 4: A random forest was trained on 60% of B cells from mouse 3609_1, and then applied to other mices' test B cells. The ROC curves from those test mice are presented above, one of which comes from the 40% of B cells held in the test set of mouse 3609_1.

amount of information not captured by the random forest made with 3609_1. After batch corrections and the use of a discretization of 3 proteins we observe decent extendability of the model based on 3609_1 to other mice. Figure 5's right plot where models are based on the corrected data has the model based on 3609_1 having a better ROC curve than the model based on 6557_1 when tested on 6557_1. AUC values presented in table 3 show the change on model performance after corrected the data. This observation occurs across all seven initial training mice we examined up to small strengths of the mouse that the random forest was tested on having a random forest modeled on it with similar to slightly better performance as the performance of the model developed on 3609_1.

	Before	After
3609_1:	.877	.923
6557_1:	.912	.931

Table 3: AUC values from ROC curves from random forest models developed on either 3609_1 or 6557_1 applied to 6557_1. Before is without batch corrections and discretization and After is when all the data has been batch corrected and discretized.

4.2 Estimating Proportions

Using the random forest based on 3609_1 with batch correctio nand 3 discrete protein expressions we produce probabilities of the cell being an Early B cell given it is a B cell. Four distinct mices' log odd probabilities for B cells to be Early B cells is visualized and overlay with a well fitted mixture of 2 Gaussian in Figure 6. Figure 7 visualizes that this separation into two mixtures appears to do a good job separating the cells. Across all training mice we observe an error in prediction the proportion of Early cells given the cell is a B cell ranging form -.08 to .07. The distribution of estimation errors for the training mice is visualized in Figure 8.

4.3 Use in Collaborator's Analysis

To understand how our estimates of the proportions of Early B cells were useful to our collaborators we first explain the general process our collaborators used to relate the genetic information of the mice and the immune cell proportions.

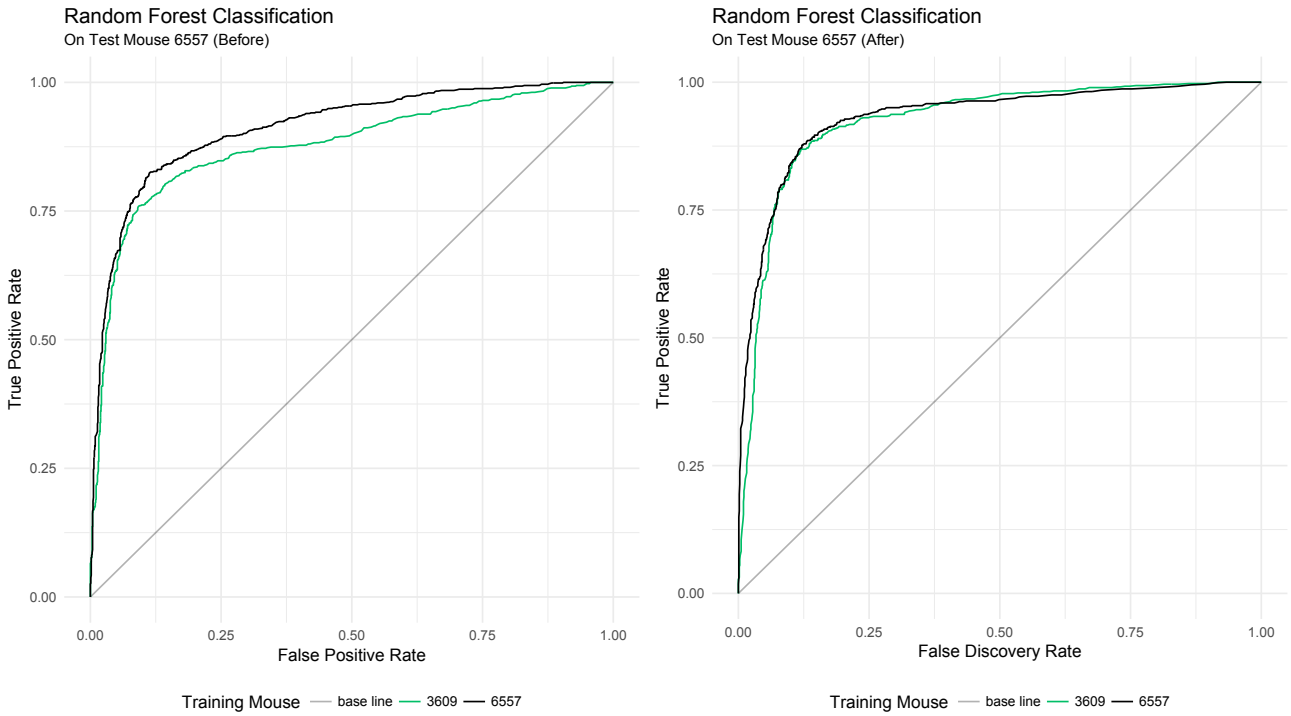

Figure 5: Random Forest Models applied on test mouse 6557_1 either developed with training cells from 3609_1 or 6557_1. The left figure uses the raw dataset, and the right uses data after batch correction and discretization.

We then follow with our of estimates of the missing proportions and the empirical distribution of errors from the training set were used to expand the our collaborators' results. Our Collaborators interest in associating proportions of different type of immune cells to genetics of the mice. On the genetic's side they looked a small pieces of the mices' DNA called loci, and select the loci to look at if the loci had been associated with immune system response in the past and if genetics across mice varied for this loci, while lead them to look at around 15 thousand of 75 thousand loci possible. For each loci, the examined the relationship between the proportion of a specific type of immune cell and that loci's genetic information, specifically using Quantitative trait locus (QTL) Mapping [5] where one regressions the proportions of the specific type of immune cell on a vector of length eight with each value in the vector corresponding to the probability that the mouse's genetic information matched the found strain 1-8. The hypothesis would test if the relationship was significant or not via an F test. This was done a lot of times (15 thousand times the number of immune cell types). To correct for multiple testing problem our collaborators used Significance Analysis of Microarrays [8] and selected an FDR control rate of $\alpha = .15$, which corresponded to a log odd probability threshold at 9.

To incorporate our predictions, our collaborators used the predicted proportions of Early B cells, added noise drawn from the empirical distribution of our estimated errors on the training set multiple times. For those loci that saw $\geq 95\%$ of their log odds probabilities above 9, our collaborators declared them significant. This increased our collaborators significant results relative to Early B cell proportions from 4 to 190 significant results, and overall saw and increase of significant results of 13.8%.

5 Discussion and Future Work

Current Gaussian mixture approaches to batch corrections provide a decnet amount of improvement in the extendability of models trained on one mouse and tested on another. Currently these models sometimes require to be manually initialized in order to find a logical local optimum, whereas we currently just initialize the components with mixture components fit on all the data together. Additionally some protein distributions do not cleanly fit into a multiple Gaussian mixture paradigm, with truncation, heavy skews, and very small potential mixture components across all mice. We used discretization of certain proteins to avoid these problems, but mixture models with a uniform component or truncated Gaussians may be a natural improvement in the model.

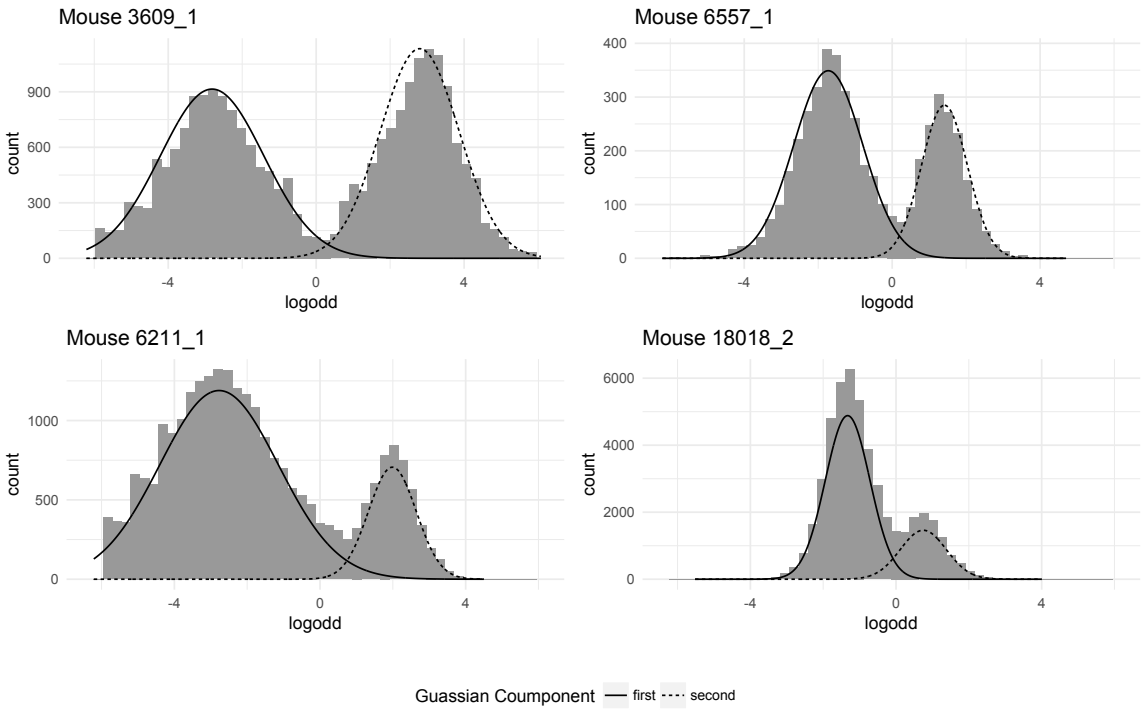

Figure 6: Distribution of log odd probabilities from a Random Forest trained onf 3609_1 applied to other mice. Overlaid with distributions of 2 fitted Gaussian components.

Our decisions to use Gaussian mixtures of log odd probabilities to estimate class proportions was based on the different proportions of the sub classes across mice effect the best threshold to separate groups. Threshold techniques either don't preserve the same rates of True Positive Rates / False positive rates and even when just using the estimation classification from the fitted mixtures preforms slightly worse than using the mixture component's proportion π_i.

References

[1] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem*, 81(16):6813–6822, 2009.

[2] B. Benacerraf and H. O. McDevitt. Histocompatibility-linked immune response genes. *Science*, 175(4019):273–279, 1972.

[3] G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, et al. The collaborative cross, a community resource for the genetic analysis of complex traits. *Nature genetics*, 36(11):1133–1137, 2004.

[4] R. Finck, E. F. Simonds, A. Jager, S. Krishnaswamy, K. Sachs, W. Fantl, D. Pe'er, G. P. Nolan, and S. C. Bendall. Normalization of mass cytometry data with bead standards. *Cytometry Part A*, 83(5):483–494, 2013.

[5] D. M. Gatti, K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, et al. Quantitative trait locus mapping methods for diversity outbred mice. *G3: Genes, Genomes, Genetics*, 4(9):1623–1633, 2014.

[6] O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M. A. Winnik, and S. Tanner. Highly multiparametric analysis by mass cytometry. *Journal of immunological methods*, 361(1):1–20, 2010.

[7] J. N. Stoop, R. G. van der Molen, C. C. Baan, L. J. van der Laan, E. J. Kuipers, J. G. Kusters, and H. L. Janssen. Regulatory t cells contribute to the impaired immune response in patients with chronic hepatitis b virus infection. *Hepatology*, 41(4):771–778, 2005.

[8] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

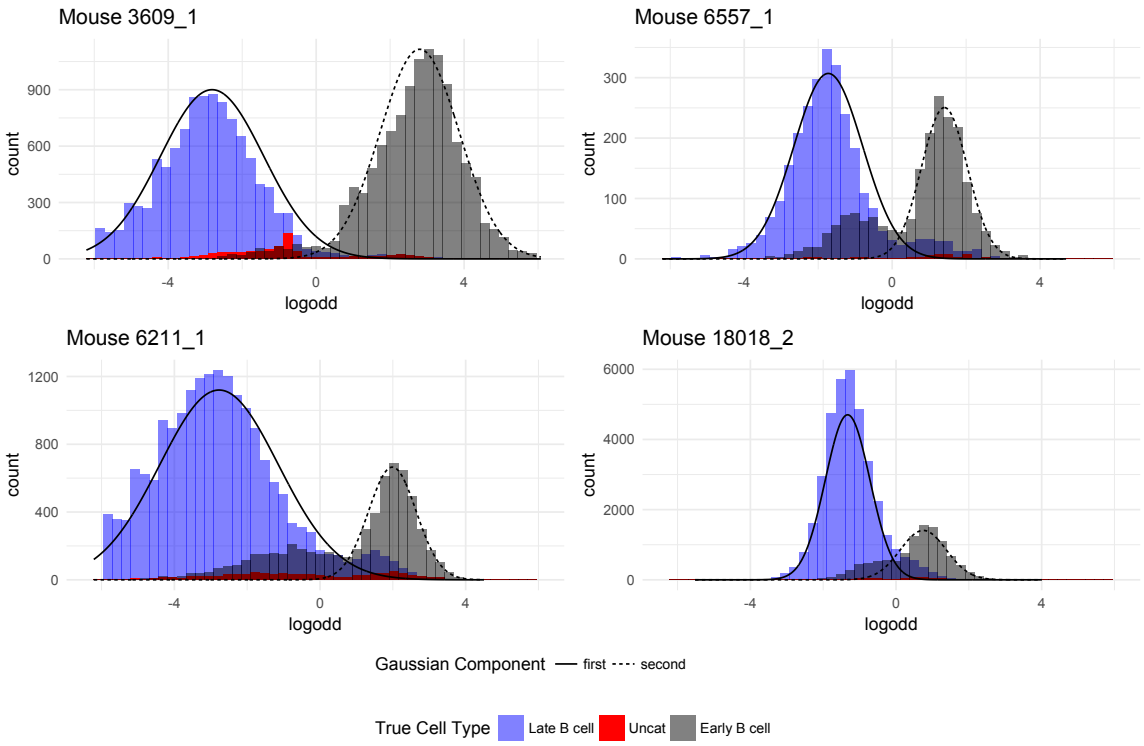

Figure 7: Distribution of log odd probabilities from a Random Forest trained on 3609_1 applied to other mice, for each B cell subclass. Overlaid with distributions of 2 Gaussian components fitted on all cell types.

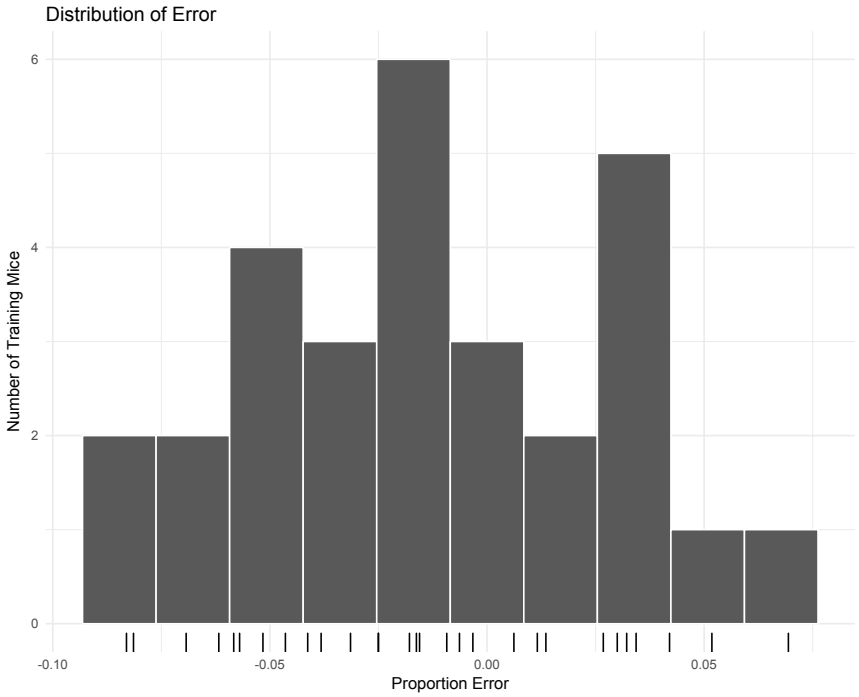

Figure 8: Distribution of errors in proportion estimation from mice we know the true proportion of Early B cells.