# Heterogeneous Sub-Population Proportion Estimations of Immune Cells

Benjamin LeRoy*
Advisor: Max G'Sell*
Collaborators: Tania Dubovik† and Shai Shen-Orr†

**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, 9, 10]. 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 [4, 8] . This diverse dataset also provides complications is missing data, stemming from certain genetic strains preventing a few specific protein markers from binding on any cell from mice with this genetic makeup. This missingness results in an inability to estimate proportions of specific types of immune cells. We set out solving this problem by extending analysis ran on mice without this missingness to those with missingness. To do so we develop procedures to standardize variable expression, classify cells, and estimate proportion of cell types.

# 1 Introduction

Immunology studies how the immune system works and how the body responds to infection and disease, with a specific focus on giving insight into links between the immune system response and the individual. Important research continues to explore links between proportions of different immune cells and immune system response, like Strauss-Albee et al's [10] recent linkage of individual's proportions of NK cells to HIV susceptibility and Stoop et al's [9] ability to related differences in relative quantities of T cell to how humans deal with Hepatitis B. Additionally, there has been a long history of immunologists like Benacerraf and McDevitt [2] linking individuals immune system response to genetic characteristics of individuals.

Although there has been a lot of work relating the immune response to the proportion of immune cell types and to the genetic make-up of the individual, few works have looked at the relationship between immune cell proportions and genetics. This lack of exploration was generally associated with difficulties capturing a large amount of genetic variation

*Department of Statistics and Data Science, Carnegie Mellon University
†Department of Immunology, Faculty of Medicine, Technion - Israel Institute of Technology

and controlling for environmental factors that could make examining such relationships harder to identify [3]. Collaborators at Technion - Israel Institute of Technology were able to address these issues in a collection of genetically diverse young mice bred using Collaborative Cross [4] and recorded tens of thousands of cells per mouse using mass cytometry [1].

## 1.1 Statistical Motivation

Mass cytometry allows for our collaborators to record the amount of specific proteins on a cell through the use of antibodies with rare metallic markers that are expected to bind to such proteins before the cell is vaporized, where-up the amount of rare metals are recorded (see figure 2(a) for more details) [1]. Generally, immunologists manually apply a series of gating one and two dimensional gating schemes based on cells' protein expressions to classify the cells (see figure 1 for example). Problematically, the genetic diversity of mice in our sample lead to certain marker antibodies failing to bind for all cells of certain mice (see figure 2(b) for visual), thereby leading to missingness in protein expressions that immunologists use to define different types of immune cells. This leads of our collaborators to an inability to estimate proportions of certain types of immune cells. We develop a way to estimate these proportions for mice with genetics that cause missingness by combining a random forest to classify manually unclassifiable cells and mixture models to estimate the proportions of these cells. We also present a constrained mixture model technique to align protein expression distributions before applying classification of cells in order to allow for better transference of the classification trained on one mouse to be applied to classify cells from other mice. Our collaborators use our proportion estimates in downstream analysis to gain 13.8% more significant associations between granular genetic information (loci) and immune cell proportions.

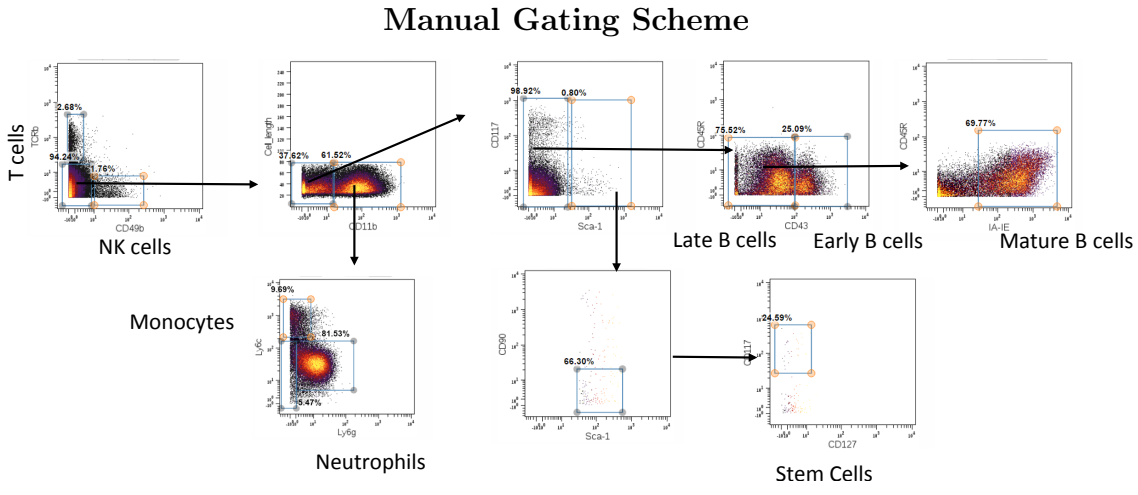


**Figure 1:** Visualization of manual gating scheme to classify immune cells using protein expressions. Figure provided by Tania Dubovik and Shai Shen-Orr.

# 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 Collaborative Cross [4]. The number of cells scanned per mouse ranges from 1,260 to 48,130 with a median of 12,760 cells scanned (only two mice had with less that 3,000 cells collected).

Each cell was scanned use mass cytometry, a technique developed in 2009 [1, 8]. 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 [8]. 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 Figure 2(a), 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 particles 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 [5].

## 2.2 Missingness

The antibodies used in mass cytometry were developed to bind to are 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, antibodies designed to bind to the specific proteins fail to do so. This causes random noise for certain proteins to be recorded for all cells in these mice from the mass cytometry procedure, visualized in Figure 2(a), and means that for some mice we must approach certain proteins as if we did not record amounts for all cells of certain mice, visualized in Table 1(b).

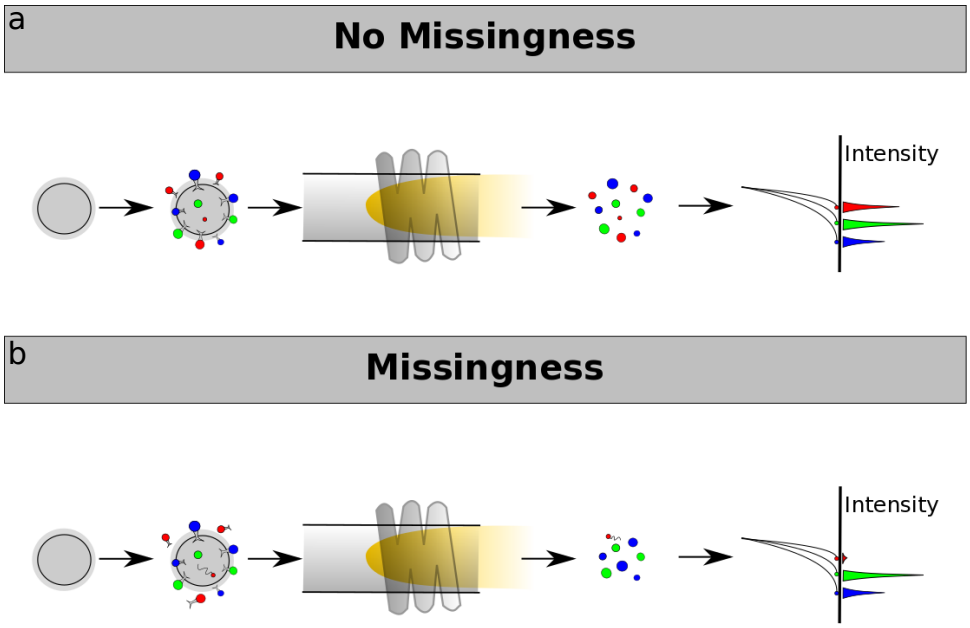


**Figure 2:** Mass Cytometry Procedure visualized for cells from mice that do not have any missing and for those mice that observe missingness of specific protein markers.

| cell idx | Ly6g expression | Ly6c expression | ... | CD44 expression | CD43 expression | CD45R expression | ... |
|---|---|---|---|---|---|---|---|
| 1 | 0.15 | 0.38 | ... | 0.15 | 0.78 | 0.59 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1053 | 1.96 | 0.25 | ... | 0.84 | 1.16 | 0.3 | ... |
| 1052 | 0.70 | 1.67 | ... | 0.26 | 1.19 | 0.77 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

**(a)** Data mxample for mouse without any missingness

| cell idx | Ly6g expression | Ly6c expression | ... | CD44 expression | CD43 expression | CD45R expression | ... |
|---|---|---|---|---|---|---|---|
| 1 | 0.71 | X | ... | 0.65 | X | 0.30 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1053 | 1.0 | X | ... | 0.50 | X | 0.26 | ... |
| 1052 | 0.85 | X | ... | 0.51 | X | 0.56 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

**(b)** Data example for mouse with missingness

**Table 1:** Data example visuals for mice with and without missingness.

# 3 Methods

The missingness of certain protein expressions in some mice means that, for specific cell types, our collaborators cannot estimate the proportions of these cell types. This is because immunologists use a sequence of manually-defined gating schemes (decision boundaries) on one-and-two dimensional distributions of protein expressions to classify cells; if protein expression uses in these gating are missing, then cells defined by these gating schemes cannot be classified. In the following methods we develop an approach to estimate proportions of of Early and Late B cell subtypes for mice that do not have the CD43 protein expression, and therefore cannot be manually classified.

## 3.1 Batch (Mouse Specific) Correction

Before estimating proportions of cell types and trying to classify cells we must first deal with misaligned protein expression distributions across each mouse (batch). Even with bead normalization mentioned in Section 2, across each mouse the cells' protein expressions do not align well (for an example see figure 3 (a)). It has been hypothesized that additional variation, beyond that corrected by bead normalization, comes from differences induced by staining (time of stain, time of scan, dilution factor, etc). Motivated by manual gating schemes being a set of binary cut-offs, we only need to align so that these cutoffs match across mice. As such, we model the final misalignment, per protein expression, with a linear shift for each mouse. Moreover, standard mean-based alignments wouldn't be able to account for the differences across mice of proportions of cells in each of the immune cell class. Differences in proportions of cells with similar protein expression across mice can be seen in figure 3 (a).

We attempt to capture the linear shift and proportional differences across mice by fitting a constrained mixture model defined as

$$X_{imc} \sim \sum_{k=1}^{K_c} \pi_{mkc} \mathcal{N}(\mu_{kc} + \delta_{mc}, \sigma_{kc}^2)$$

where $X_{imc}$ is the $c^{\text{th}}$ protein expression fo the $i^{\text{th}}$ cell of the $m^{\text{th}}$ mouse, and where we have a mixture of $K_c$ gaussians for the specific protein expression $c$. For each protein, across mice we preserve each mixture's spread (fixed $\sigma_{kc}$) and the distance between mixtures (fixed $\mu_{kc}$), but allow for linear shifts of the means to vary across mice ($\delta_{mc}$) and different proportions of mixture components to vary across mice ($\pi_{mkc}$).

We fit this constrained mixture model on each protein expression (on all cells of all mice) with one, two or three mixtures depending upon the protein distribution using an EM (expectation maximization) algorithm from the `R` package `flexmix`. Visual checks of the EM output were preformed to assess convergence to the expected maximum, with initialization points updated and the model rerun if a suboptimal local maximum was obtained. A few distributions that didn't appear like a mixture of gaussians were later converted to a discrete variable with well defined levels - done manually for each mouse. After fitting this constrained mixture model across all mice for for a single protein expression (conditional on the number of observed modes), we subtract $\delta_{mc}$ from each cell's protein expression for mouse $m$ (this process is visualized in Figure 3).

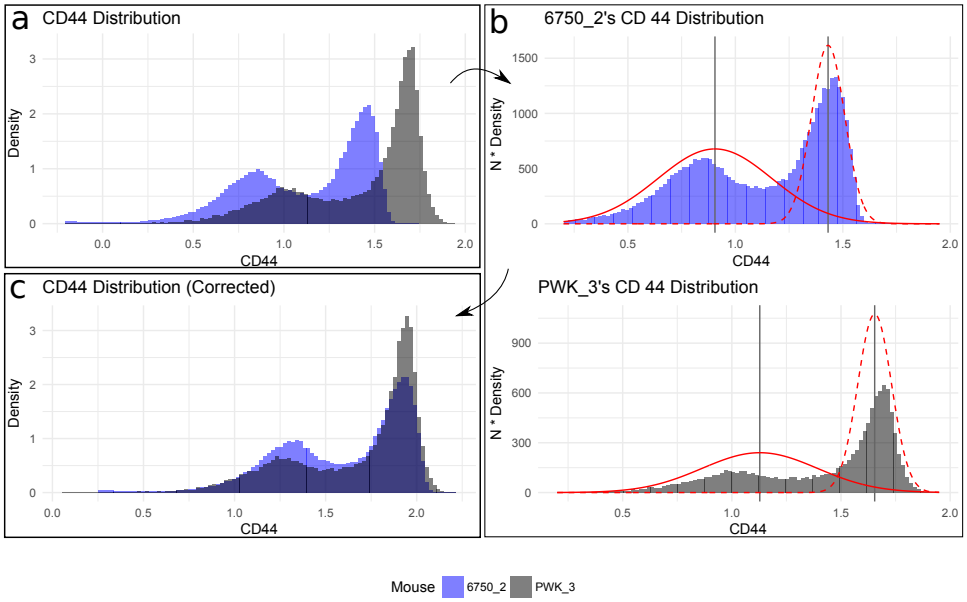


**Figure 3:** (a) Unaligned distributions of CD44 in two mice. (b) A two-component mixture model fit to both mice, with the component mean spacings and the component variances identical, but an overall mean shift allowed. (c) Aligned distributions resulting from this mixture-based approach.

## 3.2 Estimation of Cell Classes via Random Forest

After aligning protein distributions across batches we focus on transferring information about cell classification across different mice and estimation of cell type proportions. Mimicking a manual gating classification we train a random forest on a single mouse with the goal to extending classification to other mice. Specifically, we created a random forest with 1000 trees on only B cells to classify these cells as either Early Bcell, Late Bcell or Uncategorized without protein expressions that are missing for any mouse. Within the final random forest we also transform a few variables to ordered categorical variables (below 0, medium, and extreme expression levels) when the protein expression doesn't appear similar to a mixture of gaussians as mentioned in section 3.1.

## 3.3 Estimating Proportions of Cell Sub-Populations

After finding a good random forest classifier that extends well to classifying cells in other mice we can approach estimating the proportions of these classified cells. In order to estimate the proportions we model logit probabilities of being an Early B cell (conditional on being a B cell) as a mixture of 2 gaussians (respectively Late B cells and Early B cells by order of means) and took the associated proportions as estimates of the true proportion (see figure 6 as an example of the fit).

Beyond estimation, we used the estimation of cell proportions of other mice with the true proportion of Early and Late B cells known to understand the errors in the procedure. Our collaborators use the empirical distribution of these errors in the downstream analysis comparing the proportions to genetic makeup.

# 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. Using the random forest trained on 3609_1 we estimate the proportions of the two cells types of interest.

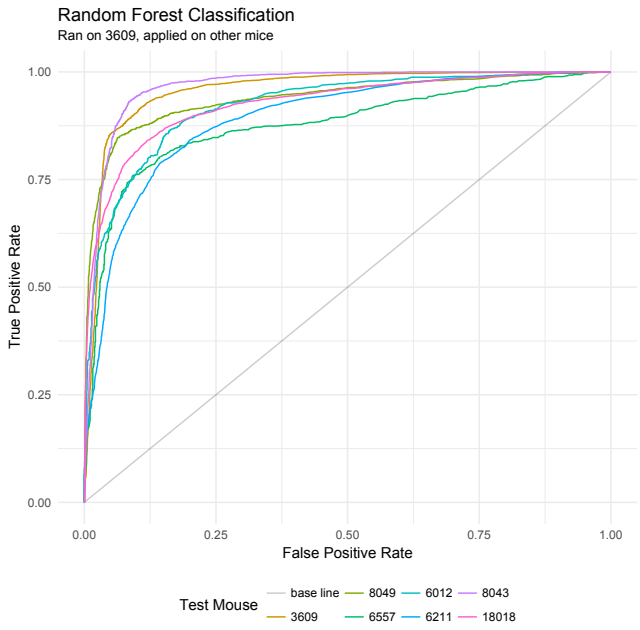


**Figure 4:** A random forest was trained on 60% of B cells from mouse 3609_1, and then applied to other mices' test B cells before batch corrections. The ROC curves from those test set of mice's cells are presented above.

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 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 (namely proteins __, __ , __). 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 2 show the change on model performance after corrected the data. This extendability of the model built on 3609_1 is observed 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.

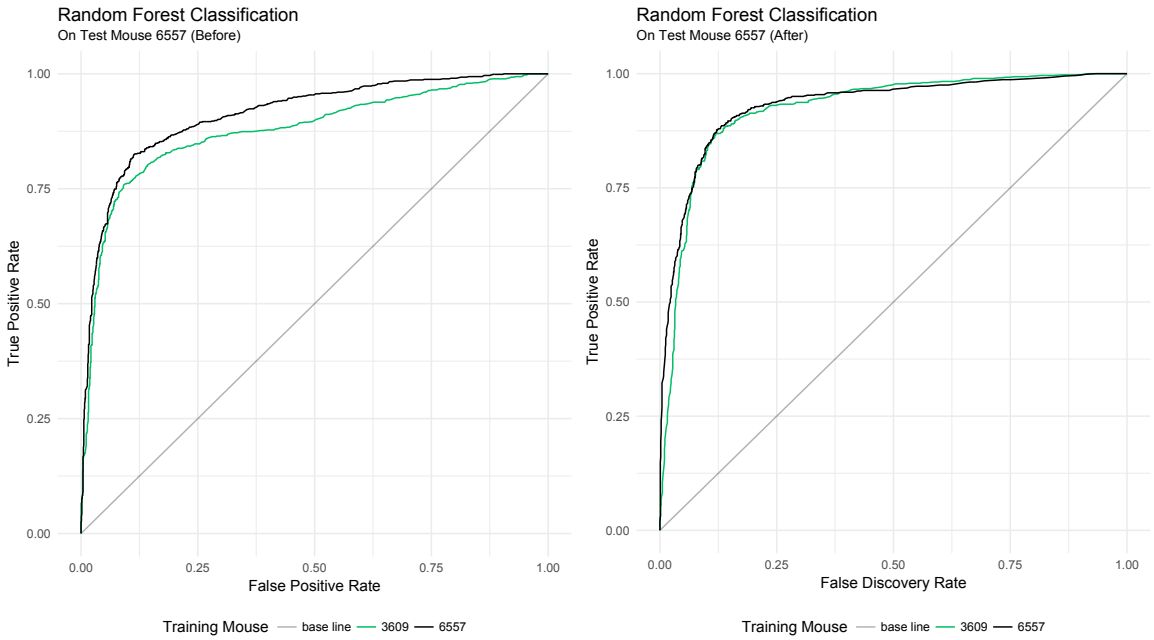


**Figure 5:** Random Forest Models applied on test mouse 6557_1 either developed with training cells from 3609_1 or 6557_1. The left figure uses the raw dataset, and the right uses data after batch correction and discretization of 3 protein expression distributions

## 4.2 Estimating Proportions

Using the random forest based on 3609_1 with batch correction and 3 discrete protein expressions we produce probabilities of each cell being an Early B cell given it is a B cell. Four distinct mices' log odd probabilities of being an Early B cell conditional on being a B cell are 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

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

**Table 2:** AUC values from ROC curves from random forest models developed on either 3609_1 or 6557_1 applied to 6557_1. "Before" is the AUC from the model without batch corrections and discretization and "After" is from the model ran on data with batch corrections and and the 3 discretized protein expressions.

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, and appear decently symmetric.

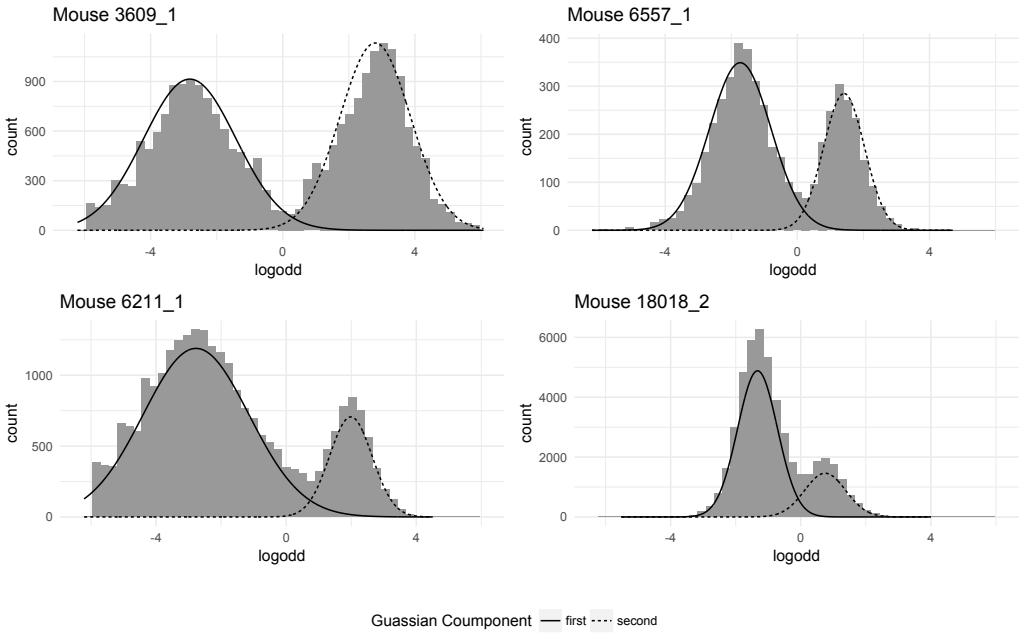


**Figure 6:** Distribution of log odd probabilities from a Random Forest trained onf 3609_1 applied to other mice. Overlaid with distributions of 2 fitted Gaussian components.

## 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. We then explain how the predicted cell proportions and empirical distribution of errors are used.

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

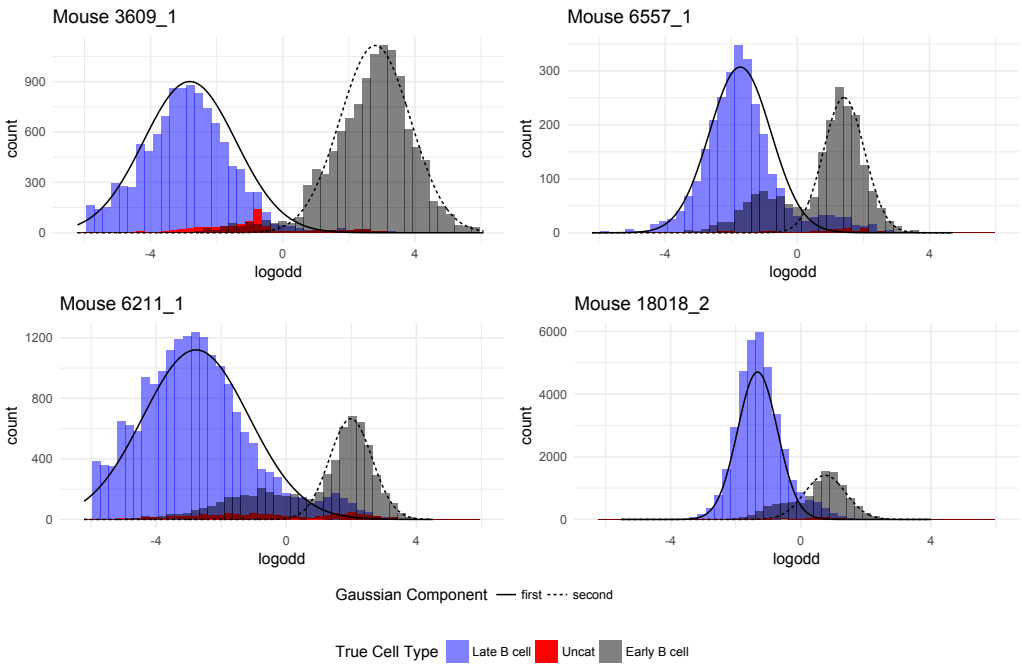


**Figure 7:** Distribution of log odd probabilities from a Random Forest trained on 3609_1 applied to other mice, for each B cell subclass. Overlaid with distributions of 2 Gaussian components fitted on all cell types.

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 [6] 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 [11] and selected an FDR control rate of $\alpha = .15$, which corresponded to a log odd probability threshold that depends on the cell type.

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 the required threshold, 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 an increase of significant results of 13.8%.

# 5 Discussion and Future Work

Current Gaussian mixture approaches to batch corrections provide a decent amount of improvement in the extendability of models trained on one mouse and tested on another. Currently these models can sometimes require manually initialization in order to find a

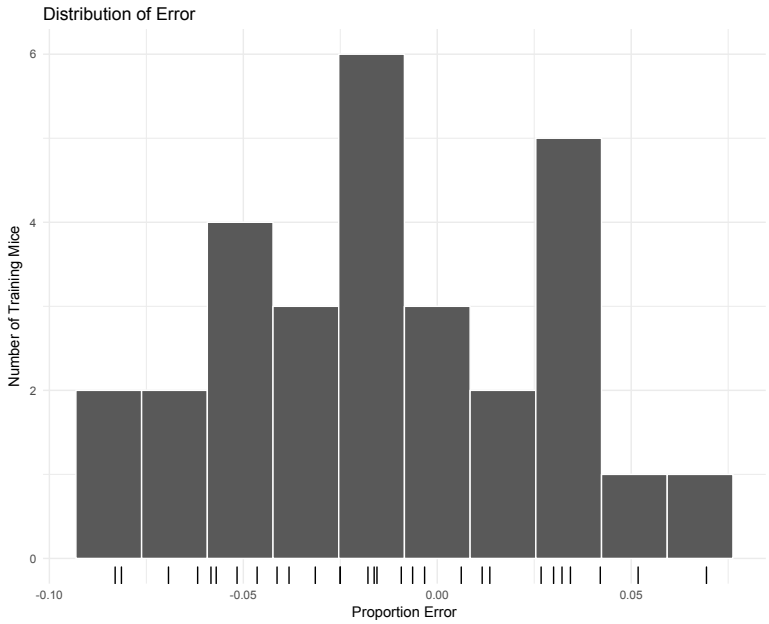


**Figure 8:** Distribution of errors in proportion estimation from mice we know the true proportion of Early B cells.

logical optimum. 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.

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 $\pi_i$. Work by Lipton et al suggest that more standard approaches are possible [7].

# 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] P. Brodin, V. Jojic, T. Gao, S. Bhattacharya, C. J. L. Angel, D. Furman, S. Shen-Orr, C. L. Dekker, G. E. Swan, A. J. Butte, et al. Variation in the human immune system is largely driven by non-heritable influences. *Cell*, 160(1-2):37–47, 2015.

[4] 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.

[5] 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.

[6] 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.

[7] Z. C. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*, 2018.

[8] 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.

[9] 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.

[10] D. M. Strauss-Albee, J. Fukuyama, E. C. Liang, Y. Yao, J. A. Jarrell, A. L. Drake, J. Kinuthia, R. R. Montgomery, G. John-Stewart, S. Holmes, et al. Human nk cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Science translational medicine*, 7(297):297ra115–297ra115, 2015.

[11] 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.