A BIOINFORMATIC PIPELINE FOR THE ANALYSIS OF PROTEIN MICROARRAY DATA WITH APPLICATIONS TO MALARIA AND LUNG CANCER STUDIES

by

Sophie Christine Bérubé

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
April, 2021

© 2021 Sophie Christine Bérubé
All rights reserved
Abstract

Bioinformatic pipelines are steps taken to transform data from a raw measurement to a form that enables direct biological inference. These steps vary across assays and different methods can have important impacts on downstream analyses and subsequent inference. While there has been substantial work on optimizing methods for many types of assays including DNA microarrays, relatively few methods have been developed and evaluated specifically for protein microarrays. Due to the high levels of technical variation, and relative measurements obtained from protein microarray data methods specifically suited to these assays are especially important to ensure that biological questions of interest can be directly answered with these data. Here, we propose a bioinformatic pipeline for protein microarray data that contains three main steps: a pre-processing pipeline to quantify and address technical variation, a Bayesian model to produce full posterior distributions of signal, and ranking methods that use information from full posterior distributions. In Chapter 2 we use Bland-Altman plots and associated analysis show that the pre-processing pipeline reduces technical variation in two previously published data sets that use protein microarrays to investigate lung cancer and malaria. In Chapter 3 we show that our proposed Bayesian model fits well
to these same two data sets and produces estimates of signal that are well suited to downstream inference, specifically to ranking methods that pay attention to uncertainty. Finally, in Chapter 4 we show how the use of our bioinformatic pipeline, can impact downstream inference. In particular, using protein microarray data from a previously published malaria study, we show how our pipeline identifies potential biomarkers of past malaria infection that were not identified with previous analysis methods.
Thesis Committee

Primary Readers

Thomas A. Louis (Primary Advisor)
Professor Emeritus
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

William J. Moss
Professor, Chair
W. Harry Feinstone Department of Molecular Microbiology and Immunology
Department of Epidemiology
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Douglas E. Norris
Professor
W. Harry Feinstone Department of Molecular Microbiology and Immunology
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Ingo Ruczinski
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Amy Wesolowski  
Assistant Professor  
Department of Epidemiology  
Johns Hopkins Bloomberg School of Public Health

**Alternate Readers**

Abhirup Datta  
Assistant Professor  
Department of Biostatistics  
Johns Hopkins Bloomberg School of Public Health

Matthew Ippolito  
Assistant Professor  
Department of Medicine Division of Infectious Diseases and Clinical Pharmacology  
Department of Medicine Division of Clinical Pharmacology  
Johns Hopkins University School of Medicine  
W. Harry Feinstone Department of Molecular Microbiology and Immunology  
Johns Hopkins Bloomberg School of Public Health
Acknowledgments

First, to my advisor Dr. Tom Louis, I am profoundly grateful for the sheer volume of time you spent teaching and mentoring me, but also, and perhaps even more importantly, for the kindness and generosity you brought to the process. Being your student has been a true joy and will cherish the experience for the rest of my career. To my committee members, Drs. Bill Moss, Doug Norris, Ingo Ruczinski, and Amy Wesolowski, I feel very lucky to have worked closely with all of you throughout my training; I have benefited tremendously from the time you invested in teaching me. I would also like to thank my two alternate committee members, Drs. Abhi Datta and Matt Ippolito, for agreeing to serve. To the Biostatistics department staff members, Mary Joy Argo, Maria Beeson, Marti Gilbert, Debra Moffitt, and Nanette Bell, I appreciate all of your work. To my fellow students in the department, Ben Ackerman, Athena Chen, Jacob Fiksel, Lamar Hunt, Ji Soo Kim, Andrew Leroux, Emily Scott and Bonnie Smith, you have all enriched my experience, and I am so glad for your friendship. To my friends outside the department, Mary Gebhardt, Kyra Grantz, Sonia Hegde, Chenery Lowe, Hannah Meredith, Julia Pringle, and Jessica Schue, getting to know each of you has been a highlight of my time in this program.
I gratefully acknowledge funding from the Southern and Central Africa ICEMR, and to the members of this team, I am appreciative of your extensive efforts in supporting trainees. In particular to Dr. Tamaki Kobayashi, you are a wonderful epidemiologist, lab scientist, and researcher and your guidance has been invaluable. To our collaborators Drs. Heng Zhu and Jianbo Pan at JHU School of Medicine, and Huw Davies, Phil Felgner, Aarti Jain, and Li Liang at UC Irvine, thank you for sharing your data and insights. To Joshua Kaminski, thank you for your computational assistance, not only are you extremely competent but you approached all of the problems I brought you with a calm demeanor; it was very much appreciated. To Drs. Karen Bandeen-Roche, Marie Diener-West, Leah Jager, and Margaret Taub, being your teaching assistant has been a pleasure, your commitment to student well-being both in and out of the classroom is inspiring and makes our community so much stronger.

To my pre-doctoral mentors at the University of Ottawa, Drs. James Butcher and Annika Flint, your patience and generosity are a large part of what prompted me to pursue a PhD. While I don’t think I was really ever cut out to be a laboratory scientist, though I did try (!), the training you provided has served me throughout my degree. To Professor Jennifer Taback, I would never have considered pursuing graduate school in a math related field were it not for your encouragement. Your efforts at Bowdoin and in the larger Brunswick community to make math accessible to everyone are both inspiring and necessary for our field. I am lucky to have been one of the people who benefited from those efforts and hope to pay it forward one day. To Charlie,
thank you for the constant support, and the unyielding belief that I would finish this degree, most especially when I doubted myself. To my parents, you have supported me every step of the way and this experience was no exception, thank you for everything.
Dedication

À Luc et Odette
Table of Contents

Abstract ii

Acknowledgements vi

Dedication ix

Table of Contents x

List of Tables xiv

List of Figures xviii

1 Introduction 1

2 A Pre-Processing Pipeline to Quantify, Visualize and Reduce Technical Variation in Protein Microarray Studies 10

2.1 Introduction ............................................. 10

2.1.1 Technical Variation and Bland-Altman Plots .... 10

2.1.2 Protein Microarrays ................................ 11

2.1.3 A Pre-Processing Pipeline ......................... 13
2.2 Methods ................................................................. 14
  2.2.1 The Pre-Processing Pipeline ................................. 14
    2.2.1.1 Step 1: Log-Transformed Readings .................. 14
    2.2.1.2 Step 2: Robust Linear Model for Normalization 15
    2.2.1.3 Step 3: Standardization .............................. 16
  2.2.2 Simulation ....................................................... 17
  2.2.3 Protein Microarray Data ...................................... 20
2.3 Results ................................................................. 21
  2.3.1 Within Array Variation ....................................... 21
  2.3.2 Between Array Variation ................................... 22
  2.3.3 A Evaluation of the Pipeline Using Simulated Data .... 26
  2.3.4 The Impact of Normalization Techniques on Protein Ranks 29
2.4 Discussion ............................................................. 29
2.5 Supplement ........................................................... 33
2.6 Pipeline Overview ................................................... 33
2.7 Bland-Altman Analysis .............................................. 33
2.8 Sources of Technical Variation ................................... 37
2.9 Protein Ranks .......................................................... 38

3 A Bayesian Hierarchical Model for Signal Extraction from Protein Microarrays 51
  3.1 Introduction ......................................................... 51
  3.2 Methods .............................................................. 54
3.2.1 Protein Microarray Components ........................................ 54
3.2.2 Normalized Data .......................................................... 56
3.2.3 Proposed Models .......................................................... 56
  3.2.3.1 Complete Model ...................................................... 56
  3.2.3.2 Simplified Models ................................................... 59
  3.2.3.3 Estimation Procedure .............................................. 61
  3.2.3.4 Evaluating Model Fit .............................................. 62
3.3 Results ................................................................................ 62
  3.3.1 Estimation and Fit Using Model 1 ..................................... 63
  3.3.2 Estimation and Fit Using Models 2 and 3 .......................... 64
  3.3.3 Comparing Estimation Across the two Datasets ................. 67
3.4 Discussion and Future Work ................................................ 70
3.5 Supplement .......................................................................... 76
  3.5.1 Estimating $\alpha$ .......................................................... 76

4 Using Full Posterior Distributions of Normalized Signal From Pro-
tein Microarrays to Perform Differential Analysis: Data Example 101
4.1 Introduction ........................................................................ 101
4.2 Methods ............................................................................ 105
  4.2.1 Protein Microarray Components .................................... 105
  4.2.2 Pre-Processing Pipeline ............................................... 107
  4.2.3 Bayesian Model .......................................................... 107
4.3 Ranking .............................................................................. 107
List of Tables

2.1 Intercepts and slopes ($\hat{\beta}_0$ and $\hat{\beta}_1$) of OLS regression on between-array Bland-Altman plots. Sample mean difference ($\mu$) and sample concordance correlation ($\hat{\rho}_c$) of pairs of measurements across arrays. .......................................................... 26

2.2 Mean squared error (MSE) of difference in means and variances across pairs of arrays, and mean differences of spike-in controls across arrays. Sample mean of 1000 simulations and, in subscript, the sample standard deviation. Scenario 1: data generating mechanism with a log-normal distribution and no difference in distributions across arrays. Scenario 2: data generating mechanism with a log-normal distribution and a difference in distributions across arrays. Scenario 3: data generating mechanism with a normal distribution and squared values and no difference in distributions across arrays. Scenario 4: data generating mechanism with a normal distribution and squared values and a difference in distributions across arrays. .......................... 27
S1 Intercepts and slopes ($\hat{\beta}_0$ and $\hat{\beta}_1$) of OLS regression on between-array Bland-Altman plots. Sample mean difference ($\mu$) and sample concordance correlation ($\hat{\rho}_c$) of pairs of measurements across arrays. 

S2 Mean squared error (MSE) of difference in means and variances across pairs of arrays, and mean differences of spike-in controls across arrays. Sample mean of 1000 simulations and, sample standard deviation. Scenario 1: data generating mechanism with a log-normal distribution and no difference in distributions across arrays. Scenario 2: data generating mechanism with a log-normal distribution and a difference in distributions across arrays. Scenario 3: data generating mechanism with a normal distribution and squared values and no difference in distributions across arrays. Scenario 4: data generating mechanism with a normal distribution and squared values and a difference in distributions across arrays.

3.1 Running times for the three models on individual malaria and HuProt\textsuperscript{TM} arrays (x1) as well as both datasets (x503 and x100 respectively).

S1 Values of $t(p)$ on the HuProt\textsuperscript{TM} array.

S2 Values of $t(p)$ on the malaria array.
4.1 Minimum posterior probability of antigens with high ranks on the malaria arrays. Number of proteins ranking ($\tilde{Q}_{p(i,j)}$) in the 97th percentile or above, in the 95th percentile or above, or in the 90th percentile or above, that also have a high (0.75 and 0.5) posterior probability ($\tilde{\pi}_{p(i,j)}$) of being in the top 10, 5 and 3 percent of proteins. ................................. 115

4.2 Minimum posterior probability of proteins with high ranks on the 20 HuProtTM arrays spotted with serum from cancer-free individuals. Number of proteins ranking ($\tilde{Q}_{p(i,j)}$) in the 97th percentile or above, in the 95th percentile or above, or in the 90th percentile or above, that also have a high (0.75 and 0.5) posterior probability ($\tilde{\pi}_{p(i,j)}$) of being in the top 10, 5 and 3 percent of proteins. ................................. 116

4.3 Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\tilde{\pi}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\tilde{\pi}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed among adult and child samples across three locations. ................................. 122

S1 Demographic characteristics of the study sample from Kobayashi et al. (2019). ................................. 128
S2 Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\bar{\pi}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\bar{\pi}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across three locations with ($\bar{\pi}_{\text{Honde},p(i,j)}$, $\bar{\pi}_{\text{Macha},p(i,j)}$, and $\bar{\pi}_{\text{Nchelenge},p(i,j)}$). ... 128

S3 Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\bar{\pi}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\bar{\pi}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across two age categories with ($\bar{\pi}_{\text{adults},p(i,j)}$, and $\bar{\pi}_{\text{children},p(i,j)}$). ... 129

S4 Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\bar{\pi}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\bar{\pi}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across two RDT result categories with ($\bar{\pi}_{\text{pos},p(i,j)}$, and $\bar{\pi}_{\text{neg},p(i,j)}$). ... 130

S5 Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\bar{\pi}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\bar{\pi}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across two RDT result categories and three regions. 131
List of Figures

1.1 Steps of the proposed bioinformatic pipeline. . . . . . . . . . . 5

2.1 Bland-Altman Plots of HuProt\textsuperscript{TM} arrays (Pan et al., 2017) (Arrays 1, 2, 33) and malaria arrays (Kobayashi et al., 2019) (Arrays 33, 43, 504). We show $M$ vs $|D|$ with $\tilde{Y}$s of duplicated probes from a single array. The ordinary least squares line is shown in red. Outlying pairs of probes have been removed. . . . . . . 23

2.2 Bland-Altman Plots of HuProt\textsuperscript{TM} arrays spotted with cancer-free samples (Pan et al., 2017) (Arrays (1,2), (1,3), and (2,3)) and malaria arrays spotted with samples from U.S. residents (Kobayashi et al., 2019) (Arrays (498,500),(498,504) and (500,504)). We show $M$ vs $D$ with $\tilde{Y}$s from pairs of arrays. The ordinary least squares line is shown in red. . . . . . . . . . . . . . . . . . . . . . . . . . 25

2.3 Ranks of top 10 and top 25 – 35 antigens on malaria and HuProt\textsuperscript{TM} (cancer) arrays are compared under three different conditions: raw intensities ($Y'$), log-transformed and RLM corrected intensities ($\tilde{Y}$) and log-transformed, RLM corrected and standardized intensities ($\hat{Y}$) all as described in Section 2.2.1. . . . . . . 30
S1  Schematic of pre-processing pipeline, starting with values obtained directly from the protein microarray scanner. . . . . . . . 34

S2  The distribution of lambda values for the Box-Cox transform of the readings (R) estimated separately for each array in each study using a maximum likelihood approach. . . . . . . . . . . 35

S3  Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) (Kobayashi et al., 2019). We show M vs D with Ys from pairs of arrays. The ordinary least squares line is shown in red. . . . . . . . . . . 39

S4  Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) (Kobayashi et al., 2019). We show M vs D with Ys standardized using the median and IQR from pairs of arrays. The ordinary least squares line is shown in red. . . . . . . . . . . . . . 40

S5  Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) Kobayashi et al. (2019). We show M vs D with Ys standardized using the Winsorized mean and standard deviation (bottom and top 5% of data points replaced) from pairs of arrays. The ordinary least squares line is shown in red. . . . . . . . . . . . . . . . . . . . 41
Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) (Kobayashi et al., 2019). We show $M$ vs $D$ with values normalized using the quantile method (Bolstad et al., 2003) from pairs of arrays. The ordinary least squares line is shown in red.

Density of intensities from 10 randomly selected arrays from each study for values standardized values, $Y$, and for quantile normalized values.

3.1 Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria arrays (A,B,C,D) fit with Model 1. Vertical dashed lines are at the value $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$.

3.2 Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria HuProt$^{TM}$ arrays (E,F,G,H) fit with Model 1. Vertical dashed lines are at mean of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$.

3.3 $P^\dagger$ for all 503 malaria arrays across Models 1, 2, and 3. The red line is $y = x$.

3.4 $P^\dagger$ for a subset of four HuProt$^{TM}$ arrays across Models 1, 2, and 3. The red line is $y = x$. 
3.5 Posterior distributions for $\sigma_i$, $\mu_i$, and $\tau$ for four malaria (A,B,C,D) arrays fit with Model 2. Vertical dashed lines are at the value $\hat{\sigma}_i$. ................................................................. 70

3.6 Posterior distributions for $\sigma_i$, $\mu_i$, and $\tau$ for four HuProt$^\text{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the value $\hat{\sigma}_i$. ................................................................. 71

S1 Observed distributions of $e_{i,t(p),p(i,j)}$ for negative, positive, and combined positive and negative controls. ......................... 78

S2 Log likelihood values of equation 3.7 for each of the three sets of proteins in the malaria arrays, negative control (A and D), positive control (B and E) and combined positive and negative control (C and F) proteins across values of $\alpha$ in the grid. The value of $\alpha$ that produces the maximum log likelihood value is shown with a red vertical line. ................................. 79

S3 Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria arrays (A,B,C,D) fit with Model 1. Vertical dashed lines are at the means of values $Y_{i,t(p),j}$ corresponding to the proteins with type $t(p)$. ................................. 80

S4 Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four HuProt$^\text{TM}$ arrays (E,F,G,H) fit with Model 1. Vertical dashed lines are at the means of values $Y_{i,t(p),j}$ corresponding to the proteins with type $t(p)$. ................................. 81

S5 Trace plots for estimated parameters, $S$, $\mu_{i,t(p)}$, $\mu_i$, $\tau$, $\sigma_i$ obtained by fitting Model 1 to four malaria arrays (A,B,C,D). ................................. 82
S6 Trace plots for estimated parameters, $S_i, \mu_{i,t(p)}, \mu_{i,a_1}, \tau_{i,a_1}, \sigma_i$ obtained by fitting Model 1 to four HuProt$^{TM}$ (E,F,G,H) arrays.  

S7 Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria (A,B,C,D) and HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the value or means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i, j)$.  

S8 Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria (A,B,C,D) and HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 3. Vertical dashed lines are at the value or means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i, j)$.  

S9 Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria (A,B,C,D) and HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the proteins with type $t(p)$.  

S10 Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria (A,B,C,D) and HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 3. Vertical dashed lines are at the means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the proteins with type $t(p)$.  

S11 Trace plots for estimated parameters, $S_i, \mu_{i,t(p)}, \mu_{i,a_1}, \tau_{i,a_1}, \sigma_i$ obtained by fitting Model 2 to four malaria arrays (A,B,C,D).  

S12 Trace plots for estimated parameters, $S_i, \mu_{i,t(p)}, \mu_{i,a_1}, \tau_{i,a_1}, \sigma_i$ obtained by fitting Model 2 to four HuProt$^{TM}$ (E,F,G,H) arrays.
S13 Trace plots for estimated parameters, $S, \mu_{i,t(p)}, \mu_{i,a_1}, \tau_{i,a_1}, \sigma_i$ obtained by fitting Model 3 to four malaria arrays (A,B,C,D).  

S14 Trace plots for estimated parameters, $S, \mu_{i,t(p)}, \mu_{i,a_1}, \tau_{i,a_1}, \sigma_i$ obtained by fitting Model 3 to four HuProt$^TM$(E,F,G,H) arrays.  

S15 Posterior distributions for $\sigma_i, \mu_{i,a_1}$, and $\tau_{i,a_1}$ for four malaria (A,B,C,D) arrays fit with Model 1.  

S16 Posterior distributions for $\sigma_i, \mu_{i,a_1}$, and $\tau_{i,a_1}$ for four HuProt$^TM$(E,F,G,H) arrays fit with Model 1.  

S17 Posterior distributions for $\mu_{i,a_1}$, and $\tau_{i,a_1}$ for four malaria (A,B,C,D) arrays fit with Model 3.  

S18 Posterior distributions for $\mu_{i,a_1}$, and $\tau_{i,a_1}$ for four HuProt$^TM$(E,F,G,H) arrays fit with Model 3.  

4.1 $\bar{\xi}_{p(i,j)}$ values averaged across subsets of arrays with samples from Honde Valley, Macha, and Nchelenge are on the ordinate and associated percentiles of ranks $(\frac{2^{\bar{W}_{p(i,j)}-1}}{2^{p}-1})$ are on the abscissa. The red vertical line is at the 90th percentile.  

4.2 $\bar{\xi}_{p(i,j)}$ values averaged across subsets of arrays with samples from adults (over 5) and children (under 5) are on the ordinate and associated percentiles of ranks $(\frac{2^{\bar{W}_{p(i,j)}-1}}{2^{p}-1})$ are on the abscissa. The red vertical line is at the 90th percentile.
4.3 \( \xi_{p(i,j)} \) values averaged across subsets of arrays with samples from RDT positive and RDT negative individuals are on the ordinate and associated percentiles of ranks \( \left( \frac{2 \times \hat{W}_{p(i,j)} - 1}{2 \times p} \right) \) are on the abscissa. The red vertical line is at the 90th percentile.

4.4 \( \bar{\pi}_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among adults (older than 5) are on the ordinate and associated percentiles of ranks \( \left( \frac{2 \times \hat{Q}_{p(i,j)} - 1}{2 \times p} \right) \) are on the abscissa. The red vertical line is at the 90th percentile.

4.5 \( \bar{\pi}_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among RDT negative samples are on the ordinate and associated percentiles of ranks \( \left( \frac{2 \times \hat{Q}_{p(i,j)} - 1}{2 \times p} \right) \) are on the abscissa. The red vertical line is at the 90th percentile.

S1 Each row is an antigen on the original malaria array, each column is the rank of that antigen under different pre-processing conditions, using a different ranking technique. We include published ranks from Kobayashi et al. (2019), ranks of \( Y \) in Chapter 2, and \( \hat{T}_{p(i,j)} \) as well as \( \hat{Q}_{p(i,j)} \) ranks with \( \gamma = 3 \).

S2 Each row is a protein on the original HuProt\textsuperscript{TM} array, each column is the rank of that protein considering 20 arrays spotted with serum from cancer-free individuals under different pre-processing conditions, using a different ranking technique. We include ranks of \( Y \) in Chapter 2, and \( \hat{T}_{p(i,j)} \) as well as \( \hat{Q}_{p(i,j)} \) ranks with \( \gamma = 3 \).
S3 \( \tilde{\pi}_{p(i,j)} \) values across all malaria arrays are on the ordinate and associated percentiles of ranks \( \frac{2 \cdot \bar{Q}_{p(i,j)} - 1}{2 \cdot p} \) are on the abscissa. The red vertical line is at the percentile determined by the value of \( \gamma \) (90, 95, 97).

S4 \( \tilde{\pi}_{p(i,j)} \) values across the 20 HuProt\textsuperscript{TM} arrays spotted with serum form cancer-free individuals are on the ordinate and associated percentiles of ranks \( \frac{2 \cdot \bar{Q}_{p(i,j)} - 1}{2 \cdot p} \) are on the abscissa. The red vertical line is at the percentile determined by the value of \( \gamma \) (90, 95, 97).

S5 Boxplots showing the distribution of differences of antigen ranks \( \tilde{T}_{\text{group},p(i,j)} \) across groups of samples. The x-axis label indicates the order of subtraction, in other words the X-axis label Group 1 vs. Group 2 indicates that the values on the Y axis are \( \hat{D}_{\text{group1},\text{group2},p(i,j)} = \tilde{T}_{\text{group1},p(i,j)} - \tilde{T}_{\text{group2},p(i,j)}. \)

S6 \( \tilde{\pi}_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among children (5 and under) are on the ordinate and associated percentiles of ranks \( \frac{2 \cdot \bar{Q}_{p(i,j)} - 1}{2 \cdot p} \) are on the abscissa. The red vertical line is at the 90th percentile.

S7 \( \tilde{\pi}_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among RDT positive samples are on the ordinate and associated percentiles of ranks \( \frac{2 \cdot \bar{Q}_{p(i,j)} - 1}{2 \cdot p} \) are on the abscissa. The red vertical line is at the 90th percentile.
Chapter 1

Introduction

The concept of a microarray was first put forward in scientific literature by Chang (1983). A small array of miniaturized chemical reactions bound to a glass slide, provided many advantages and opportunities for scientific discovery, including running multiple tests on multiple samples in parallel and with relatively small quantities of reactants. The concept of microarrays took off with the introduction of the DNA, or gene expression array (Schena et al., 1995). Gene expression arrays gave rise to many important scientific discoveries, however, given that proteins are the functional units of the cell and that there is not always a clear relationship between gene expression and protein level or function, information about gene expression alone is not sufficient to fully elucidate mechanisms behind important biological phenomena. The protein microarray, first proposed in the literature by MacBeath and Schreiber (2000) and H. Zhu et al. (2000), offers the ability to measure protein levels and interactions in a way not dissimilar to the gene expression array. Protein microarrays accomplish this in a variety of ways, but the most common method of detection is a sandwich method, whereby target proteins
are printed onto the glass slide, a serum or plasma sample is added to the slide and incubation time and conditions allow proteins in the sample to bind the target proteins on the glass slide. A secondary protein with a fluorescent tag is then added to the slide and binds the proteins from the sample that were captured by target proteins bound to the slide. Binding events on the array are measured by quantifying the level of fluorescence at each target protein through imaging or scanning. Protein microarrays have been used for a broad range of advancements, including biomarker discovery, drug targets, and vaccine development (Hartmann et al., 2009; Huang & Zhu, 2017; Nagele et al., 2011; Ramachandran et al., 2008; X. Zhu et al., 2006). However, as with any laboratory assay, pre-processing and analyzing measurements in a manner that ensures direct inferences can be made about biological questions of interest is key.

To this end, a multitude of bioinformatic pipelines have been developed for a wide range of assays (Bao et al., 2014; Hwang et al., 2018; Roy et al., 2018). In particular, a robust statistical literature on normalizing and analyzing gene expression array data has developed in parallel with the technology, allowing ever more robust and nuanced information to be extracted from these assays (Barlund et al., 2008; Bolstad et al., 2003; Efron et al., 2001; Gottardo et al., 2003; Noma et al., 2010; Quackenbush, 2002; Yang et al., 2002). Each technology though, poses its own challenges to the analytical pipeline, and in particular, many of the assumptions and properties of the pre-processing steps developed for and evaluated with DNA microarray data is not suitable for protein microarrays. But despite the nuances of each assay, bioinformatic
or pre-processing pipelines often have common goals, and can be evaluated using similar benchmarks. We identify three important goals of a pipeline:

1. To remove as much technical variation as possible while leaving true biological variation unperturbed.

2. To produce outputs that are well suited as inputs to statistical analysis methods and that allow analysts to report appropriate levels of uncertainty.

3. To enable nuanced biological inference.

In this thesis, we propose a bioinformatic pipeline for protein microarray data and demonstrate through simulation and application to two datasets, that it accomplishes the foregoing three goals. In particular, Chapter 2 builds upon an existing normalization method proposed by Sboner et al. (2009) by leveraging controls to reduce technical variation between the individual microarrays that compose a study. We evaluate our proposed method by examining sources and levels of technical variation by using Bland-Altman plots and concordance correlation coefficients to examine transformed data from two previously published protein microarray studies. One study uses HuProt™ arrays commercially manufactured by CDI Labs to identify novel biomarkers for the early diagnosis of lung cancer (Pan et al., 2017). The other uses protein microarrays made by an academic research laboratory to measure the human antibody responses to Plasmodium falciparum and P. vivax infections, two causes of human malaria (Kobayashi et al., 2019). Additionally, we use a simulation to show that the pipeline reduces technical variation while
preserving biological signals of interest. We then demonstrate the potential impact of our proposed changes to the pre-processing pipeline on downstream analysis in both studies.

Chapter 3 uses the outputs of the pipeline in Chapter 2 as the inputs of a Bayesian model. Our proposed model extracts full posterior distributions for normalized measurements from protein microarrays by using repeated measurements from control probes available on all protein microarrays to inform an error structure that is then subtracted from observed values to estimate true underlying signal. We evaluate the computational efficiency of our model, with proposed simplifications, and show that our model produces estimates of true underlying signal that are useful for downstream inference, in particular for ranking proteins by fluorescent intensity.

Finally, Chapter 4 uses the full posterior distributions output by the Bayesian model to rank the proteins measured in a study of humoral immune response to malaria antigens (Kobayashi et al., 2019) and a study of lung cancer (Pan et al., 2017). We demonstrate that our pipeline produces different protein ranks across the two studies, suggesting an impact on downstream inference. Additionally, the ranking methods we propose add information about differences in response to malaria antigens across three malaria endemic regions in Zambia and Zimbabwe, different age groups, and different stages of infection (previous exposure vs. ongoing infection). We identify a subset of antigens that may be associated with differences in past malaria exposure, and that may therefore be good candidates for the development of serological targets for malaria surveillance in low transmission settings. In doing so,
we demonstrate that our pipeline has the potential to extract more nuanced information from protein microarray data sets, and therefore to further ability of this technology to answer important biological questions.

Figure 1.1: Steps of the proposed bioinformatic pipeline.
References


Chapter 2

A Pre-Processing Pipeline to Quantify, Visualize and Reduce Technical Variation in Protein Microarray Studies

Revisions invited in Proteomics and Systems Biology.

2.1 Introduction

2.1.1 Technical Variation and Bland-Altman Plots

Technical variation is the total variability across a laboratory assay that does not arise from biological processes. Correcting for this variation is a key step that enables focused inferences on biological questions of interest. In order to make corrections, it is important to understand and account for the magnitude and sources of technical variation. For example, quantifying measurement agreement across replicated or similar measurements in an assay (e.g., technical replicates and controls) reveals magnitudes and statistical features of
Bland and Altman proposed a method designed to display measurement agreement in a laboratory assay with multiple sources of measurement. (Altman & Bland, 1999) Their method, now called the Bland-Altman plot, compares pairs of measurements by plotting their mean on the horizontal axis and their difference on the vertical axis. Lin supplements information obtained from the Bland-Altman plot with the concordance correlation coefficient, which quantifies measurement agreement from multiple sources. (Lin, 1989)

Bland-Altman plots and the concordance correlation coefficient are broadly informative for a biological assay with multiple sources of replicated measurements. They have been used in the analytic pipelines of assays in other contexts including diagnostic tests, medical equipment evaluation and DNA microarrays. (Barlund et al., 2008; Bolstad et al., 2003; Liao et al., 2007; Misyura et al., 2018; Montenij et al., 2016; Quackenbush, 2002; Yang et al., 2002) Here, we focused on their role in structuring normalization approaches for protein microarrays and detecting technical variation that may be difficult to correct with normalization.

### 2.1.2 Protein Microarrays

Protein microarrays are used to measure quantities of antibodies and other proteins in human or animal sera. They are extremely versatile and have been employed in numerous areas of medicine such as cancer research, infectious disease diagnostics and biomarker identification (Hartmann et al., 2009;
Huang & Zhu, 2017; Nagele et al., 2011; Ramachandran et al., 2008; X. Zhu et al., 2006). Their first documented use was in 2000 (H. Zhu et al., 2000), and subsequently many different types of protein microarrays have become available through both commercial and academic laboratories using a variety of laboratory protocols and thus generating a wide range of technical effects. Characterizing these technical effects in a systematic and quantitative manner is necessary to develop best practices for normalization methods. Effective normalization ensures that inferences made with measurements from protein arrays do not arise from technical variation. While many methods exist for normalizing measurements from DNA arrays, few have been optimized specifically for protein microarrays (Diez et al., 2012; Eckel-Passow et al., 2005; Rosenberg & Utz, 2015; Sboner et al., 2009) despite the fact that the variable binding affinities and high sample-to-sample variability of proteins violate assumptions relied upon by normalization methods optimized for DNA arrays (Bolstad et al., 2003; Irizarry et al., 2003). Currently, studies that use protein microarrays to investigate various disease processes use a wide variety of normalization methods and pipelines (Turewicz et al., 2016; Turewicz et al., 2013). Different normalization methods have an impact on the degree to which technical variation is reduced and therefore on the validity of inferences on biological questions. In particular, many biological questions are answered using analysis that involves ranking proteins based on the magnitude of their fluorescent intensity, and then using these ranks to infer which proteins may be associated with observable phenotypes (Kricka & Master, 2008; Sundaresh et al., 2006). Proteins with the lowest fluorescent intensities are often excluded from analysis, therefore, choice of normalization method impacts inference by
changing the ranks of proteins, and can change which proteins are included in a final analysis (Baum et al., 2013; Kobayashi et al., 2019; Nakajima et al., 2018; Pan et al., 2017).

2.1.3 A Pre-Processing Pipeline

Here we introduce a pre-processing pipeline for protein microarrays that leverages information from assay controls to improve upon existing normalization methods. We envision it as a precursor to a modeling approach (such as a Bayesian hierarchical model) that is used to extract true signal from measured signal in subsequent analyses. Using regression analysis of Bland-Altman plots and Lin’s concordance correlation coefficient, we evaluate the performance of this pipeline on protein microarray data from two studies. One study uses HuProt\textsuperscript{TM} arrays commercially manufactured by CDI Labs to identify novel biomarkers for the early diagnosis of lung cancer (Pan et al., 2017). The other uses protein microarrays made by an academic research laboratory to measure the human antibody responses to \textit{P. falciparum} and \textit{P. vivax} infections, two causes of human malaria (Kobayashi et al., 2019). We measure sources of technical variation in the two studies using the Bland-Altman plots and associated regression analysis at various stages of our proposed normalization pipeline in order to evaluate the degree to which each step corrects technical variation. We suggest some sources of technical variation and features of the data that should be taken into account in further normalization steps and others that normalization may not be able to correct, but could be accommodated in subsequent analyses. Using simulation, we show that the pipeline reduces
technical variation while preserving true underlying signals in arrays that have technical variation and in those that do not. Finally, we demonstrate that the global ranks of protein levels are sensitive to which normalization methods are employed.

2.2 Methods

We provide a description of the proposed pipeline (Figure S1), a simulation designed to test the pipeline and a description of the protein microarray components. A description of the Bland-Altman analysis can be found in Section S2.7, and a description of ranking procedures can be found in Section 2.3.4.

2.2.1 The Pre-Processing Pipeline

2.2.1.1 Step 1: Log-Transformed Readings

Let $R_{ai}$ be the measured fluorescent intensity (the reading) from source “a” and $R_{bi}$ the reading from source “b” with “i” indicating pairing. The inference related to Bland-Altman plots relies on the assumption that the vertical axis ($R_{ai} - R_{bi}$), the difference of the paired readings, is normally distributed, (Altman & Bland, 1999) or at least symmetric and unimodal with no edge effects. But, the R-values are non-negative and so we use natural log-transform values
as inputs. Specifically, we plot $D_i$ (ordinate) versus $M_i$ (abscissa), where:

\[
Y'_{ai} = \log(R_{ai}); \quad Y'_{bi} = \log(R_{bi})
\]

\[
M_i = \frac{Y'_{ai} + Y'_{bi}}{2}
\]

\[
D_i = Y'_{ai} - Y'_{bi}
\]

We also evaluate a data-informed Box-Cox transform as an alternative to the log transform, however, in most cases the best transform is close to the logarithm (see Figure SS2).

### 2.2.1.2 Step 2: Robust Linear Model for Normalization

In order to account for probe-level, sub-array level and array level fixed effects, Sboner et al. (Sboner et al., 2009) suggest a linear model with robust approach (using an M-estimate) to estimate these effects and use model residuals as inputs for downstream analysis. We follow that approach, expand the subscripts, and apply the linear model:

\[
Y'_{ijkr} = \alpha_i + \beta_j + \tau_k + \epsilon_{ijkr}
\]

to control probes only, where $Y'_{ijkr}$ is the log-transformed measured fluorescent intensity (the reading) on array $i$, in sub-array $j$ for replicate $r$ (generally, at most 2) of probe $k$. Therefore, $\alpha_i$ is the effect for array $i$, $\beta_j$ the effect for sub-array $j$, $\tau_k$ the probe effect for protein $k$, and $\epsilon_{ijkr}$ is the spot specific unexplained variation for the control probes, assumed to be approximately
Gaussian distributed with mean 0 and variance that doesn’t depend on subscripts.

Estimates of $\alpha$, $\beta$ and $\tau$ are obtained by fitting the model to the set of control probes ($Y_{ictrl}'$) across all arrays and sub-arrays in a study. That is, all array, sub-array and probe level effects are jointly estimated for a study. We then compute residuals for all data (control and active probes), producing,

$$\hat{Y}_{ijkr} = Y_{ijkr}' - \hat{\alpha}_i - \hat{\beta}_j$$  \hspace{1cm} (2.3)

and use the values $\hat{Y}_{ijkr}$, with array and sub-array effects removed, in subsequent analyses. While we include $\tau_k$ in the model (Equation 2.2) to ensure that $\hat{\alpha}_i$ and $\hat{\beta}_i$ are adjusted for probe level effects, given that the model is only fit with control probes, it would not be appropriate to subtract these control probe-level effects from all test probes, therefore we only subtract $\hat{\alpha}_i$ and $\hat{\beta}_i$ to obtain $\hat{Y}_{ijkr}$ (Equation 2.3).

2.2.1.3 Step 3: Standardization

Sets of control probes are often comprised of blank spots and proteins that are known to be ubiquitously present at high concentrations in the samples being studied. Therefore, regardless of the sample’s phenotype, we anticipate strong measurement agreement of control probes across arrays if technical variation has essentially been removed. After removal of array level-effects described in Section 2.2.1.2, the sample mean over the set of RLM corrected control probes on array $i$, $\hat{Y}_{i,ctrl}$, should be approximately zero (because the model is estimated robustly, it won’t be exactly zero). However, the sample
variances of the $\tilde{Y}_{i,ctrl}$ may still be different. Therefore, we apply and evaluate a correction to ensure that the sample variance of all corrected control probes $\tilde{Y}_{ctrl}$, across all arrays is nearly constant. Specifically, we let,

$$Y_{ijkr} = \frac{\tilde{Y}_{ijkr} - \hat{\mu}_i}{\hat{\sigma}_i},$$  \hspace{1cm} (2.4)$$

where $\hat{\mu}_i$ is the sample mean over the set of RLM corrected control probes form array $i$, $\tilde{Y}_{i,ctrl}$, and $\hat{\sigma}_i$ is the sample standard deviation over this same set. We evaluate two other methods of standardization: one with the sample median and the “pseudo-variance” (inter-quartile range (IQR) divided by 1.35); the other using the 10% (5% on each end) Winsorized mean and sample standard deviation. Note that while $\hat{\mu}_i$ and $\hat{\sigma}_i$ are computed using only control probe intensities, the correction in equation 2.4 is applied to all probes.

2.2.2 Simulation

We generate measurements that mimic the malaria arrays and the HuProt$^{TM}$ arrays in four different simulation scenarios. Each simulation scenarios compares measurements across 5 generated malaria arrays and 5 generated HuProt$^{TM}$ arrays. Generated measurements are denoted by $X_{ijkl}$, where $i$ is the array type, HuProt$^{TM}$ (H) or malaria (M), $j$ is the simulation scenario 1, 2, 3 or 4, $k$ is the array number A,B,C,D or E, and $l$ is the type of probe, test (T) or control (C). We further consider a subscript $C_p$ where $p$ is the type of control spot, $p \in \{1, \ldots, 15\}$ for the malaria arrays (M) and $p \in \{1, \ldots, 17\}$ for the HuProt$^{TM}$ arrays (H).

The set of generated malaria control probe measurements $X_{MjkC_p}$ contains
108 individual measurements, drawn from 15 different continuous uniform
distributions \((U_1, \ldots, U_{15})\). The set of generated HuProt\textsuperscript{TM} control probe mea-
urements \(X_{HjkC_p}\) contains 4656 individual measurements. These measure-
urements are drawn from 17 continuous uniform distributions \((U_1, \ldots, U_{17})\). The
bounds of the uniform distributions vary depending on the scenario \(j\) and the
array number \(k\).

In scenario 1, the set of control probes \(X_{M1kC_p}\) and \(X_{H1kC_p}\) are drawn from
the same uniform distributions across all five arrays A,B,C,D, and E, and the
set of 1000 malaria test probes \(X_{M1kT}\) and 10000 HuProt\textsuperscript{TM} test probes \(X_{H1kT}\)
are drawn from

\[
X_{i1kT} = S \cdot e, \quad e = \exp\{Z\}
\]

where

\[
Z = \Phi_{0,1}^{-1}(B), \quad B \sim \text{beta}(2, 2), \quad S \sim \text{log-normal}(\mu_S = 0, \sigma^2_S = 1.5)
\]

across all five arrays A,B,C,D, and E. In scenario 2, the set of control probes
\(X_{M2kC_p}\) and \(X_{H2kC_p}\) are drawn from different uniform distributions across each
of the five arrays A,B,C,D, and E, (the uniform distributions from scenario
1 are multiplied by a different constant for each of the arrays B,C,D and E)
and the set of 1000 malaria test probes \(X_{M2kT}\) and 10000 HuProt\textsuperscript{TM} test probes
\(X_{H2kT}\) are drawn from

\[
X_{i2kT} = S \cdot e, \quad e = \exp\{Z\}
\]

where

\[
Z = \Phi_{0,1}^{-1}(B), \quad B \sim \text{beta}(2, 2), \quad S \sim \text{log-normal}(\mu_S, \sigma^2_S)
\]

where \(\mu_S = 0\) and \(\sigma^2_S = 1.5\) for Array A, \(\mu_S = 0.25\) and \(\sigma^2_S = 2\) for Array
B, \( \mu_S = 0.5 \) and \( \sigma^2_S = 2 \) for Array C, \( \mu_S = 0.75 \) and \( \sigma^2_S = 2 \) for Array D, and \( \mu_S = 1 \) and \( \sigma^2_S = 2 \) for Array E. In scenario 3, the set of control probes \( X_{M3kC_p} \) and \( X_{H3kC_p} \) are drawn in the same way as scenario 1 and the set of 1000 malaria test probes \( X_{M1kT} \) and 10000 HuProt\textsuperscript{TM} test probes \( X_{H1kT} \) are drawn from

\[
X_{i3kT} = S^2 \cdot e, \quad e = \exp^Z
\]

\[
Z = \Phi_{0,1}^{-1}(B), \quad B \sim \text{beta}(2,2), \quad S \sim \text{normal}(\mu_S = 0, \sigma^2_S = 1.5)
\]

across all five arrays A, B, C, D, and E. Finally, in Scenario 4, the set of control probes \( X_{M3kC_p} \) and \( X_{H3kC_p} \) are drawn in the same way as scenario 2 and the set of 1000 malaria test probes \( X_{M1kT} \) and 10000 HuProt\textsuperscript{TM} test probes \( X_{H1kT} \) are drawn from

\[
X_{i3kT} = S^2 \cdot e, \quad e = \exp^Z
\]

\[
Z = \Phi_{0,1}^{-1}(B), \quad B \sim \text{beta}(2,2), \quad S \sim \text{normal}(\mu_S, \sigma^2_S)
\]

where \( \mu_S \) and \( \sigma^2_S \) are the same as in scenario 2 for arrays A, B, C, D and E. In each simulation scenario, for the malaria and HuProt\textsuperscript{TM} we coerce 12 randomly selected test probes probes to have exactly the same value for arrays A, B, C, D and E, these 12 probes serve as spike-in controls. With the generated values \( X_{ijkl} \) we perform log transformation, remove fixed effects estimated by the robust linear model outlined in Section 2.2.1.2 and standardize as described in Section 2.2.1.3. We also perform a simulation with two alternative standardization methods using the median and IQR of each array and
using the Winsorized sample mean and standard deviation as described in 
Section 2.2.1.3.

To measure the degree to which our pipeline removes technical variation, 
we measure mean squared error (MSE) defined as 
$$(\bar{X}_{ijkT} - \bar{X}_{ijk'T})^2$$ for $k$ not 
equal to $k'$, we report the sample mean and sample variance over all $k$ not 
equal to $k'$ (or all possible pairwise differences across arrays A, B, C, D and E) 
of the MSE. We also reported the mean difference of the 12 spike-in probes 
across all five arrays.

### 2.2.3 Protein Microarray Data

The HuProt\textsuperscript{TM} arrays contain 43776 test probes corresponding to approx-
imately 16000 different proteins in the human proteome. The arrays also 
contain 4608 control probes corresponding to multiple replicates of negative 
controls such as empty spots as well as positive controls comprising proteins 
to which antibodies are commonly found at high levels in human serum. This 
study analyzed the antibody profiles of 100 individuals. Of these individuals, 
20 were healthy and the remaining 80 had a form of lung cancer. (Pan et al., 
2017)

The malaria arrays contain 500 \textit{Plasmodium falciparum} and \textit{P. vivax} specific 
antigens referred to as test probes, as well as multiple control probes similar 
to those on the Huprot\textsuperscript{TM}. In this study, 429 samples from 290 individuals 
were collected across three sites in malaria endemic regions of Zambia and 
Zimbabwe. Across all three study sites, a random stratified sampling scheme 
was employed for household selection and every individual present in the
household at the time of visit was eligible for enrollment meaning that not all individuals in the study had malaria. For each individual enrolled in the study, at least one serum sample was collected and spotted on an array. Additionally, protein microarrays were spotted with sera from adults residing in the USA who had never traveled to a malaria endemic region to serve as controls. (Kobayashi et al., 2019)

2.3 Results

We apply all three steps of the normalization procedure to the malaria array and HuProt\textsuperscript{TM} datasets, evaluate the impact on within and between array variability using the Bland-Altman analysis described in Section S 2.7 and concordance correlation. We also evaluate the pipeline’s ability to remove technical variation while preserving true underlying signal across four simulation scenarios as described in Section 2.2.2, and we show the impact of the pipeline steps on protein ranks.

2.3.1 Within Array Variation

Using three HuProt\textsuperscript{TM} arrays (Pan et al., 2017) and three malaria arrays (Kobayashi et al., 2019), we observe a general pattern of within-array technical variation that remains in the data after RLM normalization but the level of remaining variation is different from array to array within a study and between studies. We measure the level of technical variation remaining after RLM normalization within each array by computing slopes and intercepts of OLS lines on modified Bland-Altman plots of probe-level technical replicates on a single
array. Given that the slopes and intercepts of OLS lines in Figure 2.1 are non-zero, we conclude that every array has residual technical variation after RLM normalization. However, the modified Bland-Altman plot for each array has a different slope and intercept, suggesting that each array may have a different level of technical variation even after fixed sub-array effects have been removed using estimates from the linear model. (Sboner et al., 2009)

Beyond the different levels of technical variation, the shape of the modified Bland-Altman plot also reveals different characteristics of technical variation across the two studies. For example, Figure 2.1 shows a strong edge effect in the distribution of the absolute value of differences of paired measurements in the malaria arrays (Array 333, Array 43, Array 504). This edge effect is not present in the HuProt\textsuperscript{TM} arrays (Array 1, Array 2, Array 33).

### 2.3.2 Between Array Variation

Using pairwise comparisons across control serum from the HuProt\textsuperscript{TM} and malaria arrays, we observe a lack of measurement agreement in pairwise comparisons of arrays after RLM normalization in Figure 2.2 and Table 2.1. This finding provides evidence that an additional pre-processing step is necessary. After standardization, which is aimed at homogenizing both the mean and variance of RLM corrected control probes across arrays, we observe increased measurement agreement between control serum arrays through the higher concordance correlation coefficients ($\hat{\rho}_c$) in Table 2.1 of the standardized measurements. In comparing standardization with median and $\text{IQR}_\text{T35}$ and standardization with Winsorized mean and standard deviation, the results
Figure 2.1: Bland-Altman Plots of HuProt® arrays (Pan et al., 2017) (Arrays 1, 2, 33) and malaria arrays (Kobayashi et al., 2019) (Arrays 333, 43, 504). We show \( M \) vs \(|D|\) with \( \tilde{Y} \)'s of duplicated probes from a single array. The ordinary least squares line is shown in red. Outlying pairs of probes have been removed.
are generally very similar to those obtained with traditional standardization (mean and standard deviation), and for some pairs of arrays such as (1, 2) and (1, 3) from the cancer study, the traditional method of standardization appears to slightly outperform the other two methods (Figures S3, S4, S5, and Table S1).

Breaking down the separate goals of homogenizing both the mean and variance across arrays, a comparison of the slope, intercept and mean difference in measurements in Table 2.1 across malaria arrays suggests that the standardization procedure did little to homogenize the variance of measurements: the slopes are similar for RLM corrected measurements and standardized measurements. However, standardization did successfully homogenize mean values of fluorescent intensities: the intercepts and mean differences are closer to 0 in the standardized measurements than in the RLM corrected measurements. In the cancer arrays, the slopes in Table 2.1 suggest that the standardization succeeded in homogenizing the variance across control serum arrays. In all cases though, the intercept and the mean differences are further from zero for the standardized values than for the RLM corrected values and in both cases, even after standardization, there is still between-array variation.

An analysis of Bland-Altman plots in Figure 2.2 further reveals a strong positive correlation between the mean and variance of measurements and an edge effect, as statistical features of technical variation. Both trends are more pronounced in the malaria arrays than in the cancer arrays. While additional normalization steps and analytic methods can account for the between-array technical variation remaining after standardization, some arrays, such as
Figure 2.2: Bland-Altman Plots of HuProt\textsuperscript{TM} arrays spotted with cancer-free samples (Pan et al., 2017) (Arrays (1,2), (1,3), and (2,3)) and malaria arrays spotted with samples from U.S. residents (Kobayashi et al., 2019) (Arrays (498,500), (498,504) and (500,504)). We show $M$ vs $D$ with $\tilde{Y}$s from pairs of arrays. The ordinary least squares line is shown in red.
Table 2.1: Intercepts and slopes ($\hat{\beta}_0$ and $\hat{\beta}_1$) of OLS regression on between-array Bland-Altman plots. Sample mean difference ($\hat{\mu}$) and sample concordance correlation ($\hat{\rho}_c$) of pairs of measurements across arrays.

<table>
<thead>
<tr>
<th>Arrays</th>
<th>Method</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\rho}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>498, 500</td>
<td>RLM</td>
<td>0.086</td>
<td>0.194</td>
<td>0.033</td>
<td>0.618</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>0.107</td>
<td>0.215</td>
<td>0.095</td>
<td>0.606</td>
</tr>
<tr>
<td>498, 504</td>
<td>RLM</td>
<td>-1.530</td>
<td>-0.518</td>
<td>-1.892</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>-0.828</td>
<td>-0.525</td>
<td>-1.112</td>
<td>0.196</td>
</tr>
<tr>
<td>500, 504</td>
<td>RLM</td>
<td>-1.421</td>
<td>-0.740</td>
<td>-1.920</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>-0.821</td>
<td>-0.768</td>
<td>-1.210</td>
<td>0.132</td>
</tr>
<tr>
<td>1, 2</td>
<td>RLM</td>
<td>0.438</td>
<td>0.416</td>
<td>0.560</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>0.739</td>
<td>0.358</td>
<td>0.939</td>
<td>0.342</td>
</tr>
<tr>
<td>1, 3</td>
<td>RLM</td>
<td>0.533</td>
<td>0.477</td>
<td>0.560</td>
<td>0.240</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>0.916</td>
<td>0.365</td>
<td>0.939</td>
<td>0.264</td>
</tr>
<tr>
<td>2, 3</td>
<td>RLM</td>
<td>0.092</td>
<td>0.038</td>
<td>0.091</td>
<td>0.841</td>
</tr>
<tr>
<td></td>
<td>RLM+Std.</td>
<td>0.153</td>
<td>-0.007</td>
<td>0.153</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Arrays 1 and 504, are clear outliers and may need to be discarded from future analysis.

2.3.3 A Evaluation of the Pipeline Using Simulated Data

In order to further compare RLM normalization alone to RLM normalization with subsequent standardization, we perform a simulation to test the performance of the pipeline under four different data generating mechanisms: when the analysis model is correctly specified for the data generating mechanism we consider a scenario with (Scenario 1) and without (Scenario 2) technical variation, and when the analysis model is misspecified for the data generating
Table 2.2: Mean squared error (MSE) of difference in means and variances across pairs of arrays, and mean differences of spike-in controls across arrays. Sample mean of 1000 simulations and, in subscript, the sample standard deviation. Scenario 1: data generating mechanism with a log-normal distribution and no difference in distributions across arrays. Scenario 2: data generating mechanism with a log-normal distribution and a difference in distributions across arrays. Scenario 3: data generating mechanism with a normal distribution and squared values and no difference in distributions across arrays. Scenario 4: data generating mechanism with a normal distribution and squared values and a difference in distributions across arrays.

<table>
<thead>
<tr>
<th>Array Type</th>
<th>Scenario</th>
<th>Method</th>
<th>MSE Mean</th>
<th>MSE Var</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria</td>
<td>1</td>
<td>RLM</td>
<td>0.013,0.009</td>
<td>0.017,0.013</td>
<td>−0.001,0.098</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.003,0.002</td>
<td>0.006,0.004</td>
<td>0.001,0.033</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>RLM</td>
<td>157.760,5,865</td>
<td>1.145,0.162</td>
<td>4.357,0.128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.015,0.006</td>
<td>0.296,0.047</td>
<td>0.357,0.035</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>RLM</td>
<td>1.840,1,353</td>
<td>0.282,0.200</td>
<td>−0.010,1.367</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.002,0.001</td>
<td>0.005,0.004</td>
<td>−0.001,0.037</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>RLM</td>
<td>2739.451,122,545</td>
<td>0.289,0.212</td>
<td>70.966,3,276</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.018,0.006</td>
<td>0.022,0.012</td>
<td>0.291,0.037</td>
</tr>
<tr>
<td>HuProtTM</td>
<td>1</td>
<td>RLM</td>
<td>0.000,0.000</td>
<td>0.001,0.001</td>
<td>−0.000,0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.001,0.000</td>
<td>0.022,0.015</td>
<td>−0.000,0.014</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>RLM</td>
<td>2.580,0.068</td>
<td>0.588,0.028</td>
<td>1.072,0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.029,0.004</td>
<td>3.358,0.317</td>
<td>0.772,0.017</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>RLM</td>
<td>0.002,0.001</td>
<td>0.016,0.011</td>
<td>0.003,0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.000,0.000</td>
<td>0.012,0.008</td>
<td>0.000,0.018</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>RLM</td>
<td>17.377,0.362</td>
<td>0.026,0.017</td>
<td>3.406,0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLM+Std.</td>
<td>0.175,0.011</td>
<td>1.748,0.256</td>
<td>0.822,0.037</td>
</tr>
</tbody>
</table>
mechanism we consider a scenario with (Scenario 3) and without (Scenario 4) technical variation (Section 2.2.2). In the case where there is no technical variation between Arrays (Scenario 1), there is little difference between standardization and RLM correction. On the other hand, when we simulate technical variation across arrays (Scenario 2), standardization substantially increaseds measurement agreement between arrays, and preserves true differences between measurements, evidenced by the mean difference of the 12 spike-in controls across arrays. Furthermore, when log transformation is not an appropriate step for the data, as is the case when the probes are generated from a squared normal distribution (Scenarios 3 and 4), the standardized values show strong measurement agreement, evidenced by the near zero mean squared error for differences of means and variances (Table 2.2), and preserves true differences between arrays, something RLM correction alone does not accomplish, most markedly when there is both model misspecification and technical variation between arrays (Scenario 4). In cases where there is no technical variation between arrays, adding standardization to the pipeline leaves results relatively unchanged, when there is between-array technical variation, standardization reduces it.

Generally, traditional standardization with mean and standard deviation outperforms standardization with median and $\frac{IQR}{1.35}$, and standardization with Winsorized means and standard deviations. While there are certain scenarios (Scenarios 2 and 4) where standardization with the median and IQR outperforms traditional standardization for the malaria arrays, the traditional method of standardization outperforms the two other methods in all four
scenarios for the HuProt\textsuperscript{TM} arrays (Table S2).

### 2.3.4 The Impact of Normalization Techniques on Protein Ranks

We compare ranks when different normalization procedures are used. For both the malaria and HuProt\textsuperscript{TM} arrays the ordering of the top 10 proteins changes only slightly as a function of which normalization technique is used, with only one or two proteins changing ranks at each stage of normalization. In contrast, when considering the top 25 – 35 proteins, there are several changes in ordering at each stage of normalization, these changes could impact which proteins are targeted for analysis (Figure 2.3). These results suggest that including standardization in the pre-processing pipeline produces different protein ranks which could lead to different inference about biological questions of interest.

### 2.4 Discussion

Given that protein microarrays are a relatively new technology, methods available to address technical variation present in this assay are relatively new and untested. Performing effective normalization enables analysts to directly answer biological questions of interest and the choices made in the normalization stage of analysis impact subsequent analysis. Using Bland-Altman plots, associated regression analyses, and concordance correlations we evaluate a pipeline for normalizing measurements from protein microarrays. We show that within-array and between-array technical variability are present in both protein microarray studies after RLM normalization (Sboner et al.,...
Figure 2.3: Ranks of top 10 and top 25 – 35 antigens on malaria and HuProt\textsuperscript{TM} (cancer) arrays are compared under three different conditions: raw intensities ($Y'$) , log-transformed and RLM corrected intensities ($\tilde{Y}$) and log-transformed, RLM corrected and standardized intensities ($\bar{Y}$) all as described in Section 2.2.1.
Standardization conducted using information from control probes reduces some between-array technical variation in both the malaria and cancer arrays, this finding is further supported by results of a simulation study.

Even after standardization, technical variation across arrays, as well as within arrays, remains. We briefly investigate quantile normalization optimized for DNA microarrays (Bolstad et al., 2003), as an alternative to standardization. While this method produces strong measurement agreement across control serum arrays (Figure S6), the method achieves this by coercing the distribution of all intensities across arrays to be identical. However, the differences in the shapes of distributions across arrays in both studies (Figure S7), could result from biological signal necessary for answering questions of scientific interest. Therefore using quantile normalization and thereby erasing these differences is likely inappropriate. Moreover, certain features of the data revealed by the Bland-Altman plots, such as the positive correlation between the mean and variance of fluorescent intensities and edge effects in the distributions of signals are important to account for in future model-based normalization and analysis methods.

Flexible model-based normalization approaches are the best candidates to address the differences in technical variation we observe between arrays and across studies. These normalization approaches are likely to have the greatest impact on data from arrays that are manufactured on a small scale by academic labs. However, given the differences in the number of replicated probes and the target proteins between the HuProt\textsuperscript{TM} and malaria arrays, we cannot directly compare the impact of our pipeline on the results of the
two studies. Furthermore, even those flexible normalization procedures that take into account different levels and characteristics of technical variation within-and between arrays are unlikely to fully account for some extreme outliers both at the array and probe levels. Therefore, identifying these outliers through Bland-Altman analysis and removing them from subsequent analysis is likely the wisest course of action.

In addition to data-based limitations, we discuss the limitations of our proposed pipeline and evaluation methods. Generally, ranking procedures that take into account differential variability perform better (Louis & Ruczinski, 2010; Louis & Shen, 1999). A model based approach to normalization, such as a Bayesian hierarchical model referenced in Section 2.1 that uses the information about correlation and edge effects obtained through the Bland-Altman analysis to optimize estimation, has many advantages and produces an output that is easily used by various robust ranking methods. Additionally, we evaluate our normalization procedures for two published studies and one simulated dataset, representing a fairly narrow sample of the currently available protein microarrays. It is clear that technical variability is fairly specific to the manufacturing process, therefore a more comprehensive evaluation of normalization procedures on different types of protein microarrays is warranted. Finally, we use simulated spike-in controls to assess how standardization preserves true underlying measurements, but evaluating the pre-processing pipeline using laboratory data from an experiment with spike in probes would be an important additional evaluation.

The goal of normalization in a laboratory assay analysis is to remove
technical variation from measurements in order to make direct inferences about biological processes. Understanding how technical variation manifests in an assay is therefore a key step in designing a normalization procedure that comprehensively corrects this variation. Given the relative novelty of protein microarrays, their versatility and the variety of manufacturing processes for the assay, a systematic and quantitative investigation of technical variation across studies is relevant for developing broadly usable analysis pipelines. This is an important investigation especially in light of the sensitivity of protein ranks to the normalization steps employed. Adding standardization to existing methods corrects some technical variation, but further normalization models that incorporate features of the data revealed by our Bland-Almtan analysis, such as edge effects, are necessary and will likely have an important impact on biological inference.

2.5 Supplement

2.6 Pipeline Overview

We evaluate a Box-Cox transform of raw measurements, $R, \frac{R^\lambda - 1}{\lambda}$ for $\lambda \neq 0$, and $\log(R)$ for $\lambda = 0$. We estimate $\lambda$ separately for each array in each study using a maximum likelihood approach.

2.7 Bland-Altman Analysis

In order to characterize measurement agreement across protein microarrays, we use observed signals on the array described in Section 2.2.3. Specifically, $R_{ai}$
Figure S1: Schematic of pre-processing pipeline, starting with values obtained directly from the protein microarray scanner.

is the ratio of observed foreground to observed background of a probe $i$ from source $a$ and $R_{bi}$ is the ratio of observed foreground to observed background of this same probe from source $b$.

We consider the following notation for each ordered pair on a Bland-Altman plot where values can be raw intensities, log-intensities or corrected
Figure S2: The distribution of lambda values for the Box-Cox transform of the readings \((R)\) estimated separately for each array in each study using a maximum likelihood approach.

log-intensities (see Section 2.2.1):

\[
(M_i = \frac{R_{ai} + R_{bi}}{2}, D_i = R_{ai} - R_{bi}), \quad (2.9)
\]

In order to understand the utility of the Bland-Altman plot, we compute the relationship between \(M\) and \(D\) via the Ordinary Least Squares (OLS) fit obtained by regressing \(D\) on \(M\). With \(R_a\) and \(R_b\) the measurements for sources \(a\) and \(b\), \(s_a^2\) and \(s_b^2\) the sample variances for \(R_a\) and \(R_b\), \(\text{Cov}(R_a, R_b)\) the sample covariance of \(R_a\) and \(R_b\) and \(\bar{M}\) and \(\bar{D}\) the means of \(M\) and \(D\) from equation 2.9, the slope and intercept are:

\[
\hat{\beta}_1 = \frac{2(s_a^2 - s_b^2)}{s_a^2 + 2\text{Cov}(R_a, R_b) + s_b^2}, \quad \hat{\beta}_0 = \bar{D} - \hat{\beta}_1 \bar{M} \quad (2.10)
\]
Thus, the $\hat{\beta}_1$ measures the difference in variability across the two sources of measurement relative to the variability of the sum of measurements from the two sources. In the case of perfect agreement across both sources of measurement, the difference in variability is zero and thus the slope of the OLS line is also zero. More realistically, measurement agreement across sources in an assay will not be absolute, but slopes close to zero can still be observed indicating strong measurement agreement (see Bland and Altman p. 309 for an example of this (Bland & Altman, 1986)).

Since $\hat{\beta}_0$ is a function of both slope and the mean difference of $R_a$ and $R_b$, this parameter gives the analyst information about both differences in variability, and average difference in magnitude of measurements from sources $a$ and $b$. With $\hat{\beta}_1 = 0$, $\hat{\beta}_0$ is exactly the average difference in measurement magnitude between source $a$ and $b$.

The OLS line provides information about the linear relationship between $M$ and $D$, however, the shape of points around this line also provides insight about measurement agreement. Specifically, the association between the magnitude of measurements and their variability can be inferred by observing the shape of points on the Bland-Altman plot. In cases where the magnitude and the standard deviation have a strong positive correlation, points should fan away from the OLS line as $M$ increases. Additionally, points in a diamond shape around the OLS line suggest an upper and lower bound on the difference between measurements. We described how transformations of $R$, $M$ and $D$ can impact the shape of points around the OLS line in Section 2.2.1.

Bland-Altman plots can also diagnose outlying sources of measurements
through multiple pairwise comparisons across sources of measurement either at a cursory glance or by analyzing residuals of a simple linear regression of $D$ on $M$.

### 2.8 Sources of Technical Variation

We consider two main sources of technical variation in evaluating the pipeline: within and between array technical variability. In both instances, duplicated probes between (control probes) and within (replicated probes) arrays can be used to assess and correct technical variation.

Within array technical variability arises when laboratory conditions, such as solvent concentration, differ within sub-regions of a single array (sub-array effects), and is measured by observing differences across replicated probes on a single array. Therefore Bland-Atlman plots aimed at measuring within array technical variability use as inputs replicated probes within an array:

$$
(M_i = \frac{R_{1i} + R_{2i}}{2}, D_i = R_{1i} - R_{2i}),
$$

(2.11)

where $i$ refers to the array index, and 1 is the first replicate of a particular probe, and 2 is the second replicate of that same probe. In addition to within-array technical variability, elements of the laboratory procedure can induce between array technical variability. Analyzing differences across arrays that are spotted with serum intended to be non-reactive to the probes on arrays, or control serum, allows us to characterize between-array technical variation. It is standard to include arrays with control serum in a study, such as cancer-free individuals in the cancer study and individuals residing in non-malarious areas.
regions in the malaria study. Given that we expect measurements from control serum samples to have high degrees of measurement agreement in the absence of technical variability, characterizing measurement agreement across the same probe on different control serum arrays, reveals between array technical variation. Bland-Atlman plots aimed at measuring between array technical variability used as inputs

\[
M_i = \frac{R_{li} + R_{l'i'}}{2}, \quad D_i = R_{li} - R_{l'i'},
\]

(2.12)

where \(i\) and \(i'\) are the indices of two different arrays, and \(l\) is the index of a particular probe at a particular location on the array, typically the layout of probes do not change from array to array within a study.

### 2.9 Protein Ranks

Central to developing broadly applicable normalization techniques is an evaluation of the impact of pre-processing procedures on downstream analyses. A common component of protein microarray analysis is identifying a pared down list of target proteins, often the 30 to 50 proteins with the highest fluorescent intensities, to directly compare across observable phenotypes, such as disease status. Identifying target proteins for analysis based on ranking fluorescent intensities is often done after normalization, but ranking procedures are sensitive to which normalization techniques the analyst applied. For both example datasets, we consider the median fluorescent intensity for each probe across all arrays, these median probe-level intensities are then ranked as point estimates.
**Figure S3**: Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) (Kobayashi et al., 2019). We show $M$ vs $D$ with $Y$s from pairs of arrays. The ordinary least squares line is shown in red.
Figure S4: Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500),(498,504) and (500,504)) (Kobayashi et al., 2019). We show $M$ vs $D$ with $Y$s standardized using the median and $\frac{IQR}{1.35}$ from pairs of arrays. The ordinary least squares line is shown in red.
Figure S5: Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500), (498,504) and (500,504)) Kobayashi et al. (2019). We show $M$ vs $D$ with $Y$s standardized using the Winsorized mean and standard deviation (bottom and top 5% of data points replaced) from pairs of arrays. The ordinary least squares line is shown in red.
Table S1: Intercepts and slopes ($\hat{\beta}_0$ and $\hat{\beta}_1$) of OLS regression on between-array Bland-Altman plots. Sample mean difference ($\hat{\mu}$) and sample concordance correlation ($\hat{\rho}_c$) of pairs of measurements across arrays.

<table>
<thead>
<tr>
<th>Arrays</th>
<th>Method</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\rho}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>498, 500</td>
<td>Std Mean+SD</td>
<td>0.107</td>
<td>0.215</td>
<td>0.095</td>
<td>0.606</td>
</tr>
<tr>
<td></td>
<td>Std Med +IQR</td>
<td>0.118</td>
<td>0.267</td>
<td>0.162</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.116</td>
<td>0.214</td>
<td>0.107</td>
<td>0.605</td>
</tr>
<tr>
<td>498, 504</td>
<td>Std Mean+SD</td>
<td>-0.828</td>
<td>-0.525</td>
<td>-1.11</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>Std Med+IQR</td>
<td>-0.531</td>
<td>-0.456</td>
<td>-0.833</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>-0.905</td>
<td>-0.551</td>
<td>-1.26</td>
<td>0.192</td>
</tr>
<tr>
<td>500, 504</td>
<td>Std Mean+SD</td>
<td>-0.821</td>
<td>-0.768</td>
<td>-1.21</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>Std Med+IQR</td>
<td>-0.557</td>
<td>-0.754</td>
<td>-0.996</td>
<td>0.176</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>-0.901</td>
<td>-0.792</td>
<td>-1.36</td>
<td>0.129</td>
</tr>
<tr>
<td>1,2</td>
<td>Std Mean+SD</td>
<td>0.739</td>
<td>0.358</td>
<td>0.939</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td>Std Med+IQR</td>
<td>3.32</td>
<td>0.297</td>
<td>4.26</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>1.39</td>
<td>0.205</td>
<td>1.59</td>
<td>0.374</td>
</tr>
<tr>
<td>1,3</td>
<td>Std Mean+SD</td>
<td>0.916</td>
<td>0.365</td>
<td>0.939</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>Std Med+IQR</td>
<td>4.03</td>
<td>0.390</td>
<td>4.26</td>
<td>0.246</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>1.85</td>
<td>0.075</td>
<td>1.59</td>
<td>0.299</td>
</tr>
<tr>
<td>2,3</td>
<td>Std Mean+SD</td>
<td>0.153</td>
<td>-0.007</td>
<td>0.153</td>
<td>0.847</td>
</tr>
<tr>
<td></td>
<td>Std Med+IQR</td>
<td>0.797</td>
<td>0.070</td>
<td>0.893</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.325</td>
<td>-0.117</td>
<td>0.322</td>
<td>0.842</td>
</tr>
</tbody>
</table>
Figure S6: Bland-Altman Plots of arrays spotted with cancer-free samples (Arrays (1,2), (1,3), and (2,3)) (Pan et al., 2017) and U.S. residents (Arrays (498,500), (498,504) and (500,504)) (Kobayashi et al., 2019). We show M vs D with values normalized using the quantile method (Bolstad et al., 2003) from pairs of arrays. The ordinary least squares line is shown in red.
Figure S7: Density of intensities from 10 randomly selected arrays from each study for values standardized values, $Y$, and for quantile normalized values.
Table S2: Mean squared error (MSE) of difference in means and variances across pairs of arrays, and mean differences of spike-in controls across arrays. Sample mean of 1000 simulations and, sample standard deviation. Scenario 1: data generating mechanism with a log-normal distribution and no difference in distributions across arrays. Scenario 2: data generating mechanism with a log-normal distribution and a difference in distributions across arrays. Scenario 3: data generating mechanism with a normal distribution and squared values and no difference in distributions across arrays. Scenario 4: data generating mechanism with a normal distribution and squared values and a difference in distributions across arrays.

<table>
<thead>
<tr>
<th>Array Type</th>
<th>Scenario</th>
<th>Method</th>
<th>MSE Mean</th>
<th>MSE Var</th>
<th>Mean Dif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria</td>
<td>1</td>
<td>Std Mean+SD</td>
<td>0.027±0.018</td>
<td>0.055±0.040</td>
<td>−0.001±0.033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>1.365±0.225</td>
<td>7.186±1.045</td>
<td>0.001±0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.037±0.024</td>
<td>0.092±0.067</td>
<td>0.001±0.044</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Std Mean+SD</td>
<td>0.151±0.062</td>
<td>2.950±0.494</td>
<td>0.356±0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.127±0.060</td>
<td>1.823±0.433</td>
<td>0.312±0.051</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.184±0.080</td>
<td>4.148±0.757</td>
<td>0.387±0.046</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Std Mean+SD</td>
<td>0.020±0.015</td>
<td>0.053±0.039</td>
<td>0.001±0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.031±0.022</td>
<td>0.052±0.037</td>
<td>0.003±0.055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.031±0.023</td>
<td>0.081±0.060</td>
<td>0.001±0.051</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Std Mean+SD</td>
<td>0.172±0.059</td>
<td>0.219±0.115</td>
<td>0.291±0.037</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.103±0.054</td>
<td>0.174±0.099</td>
<td>0.237±0.060</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.200±0.075</td>
<td>0.320±0.172</td>
<td>0.312±0.050</td>
</tr>
<tr>
<td>HuProtTM</td>
<td>1</td>
<td>Std Mean+SD</td>
<td>0.006±0.004</td>
<td>0.215±0.153</td>
<td>−0.001±0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.015±0.011</td>
<td>1.368±0.988</td>
<td>−0.001±0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.017±0.011</td>
<td>1.235±0.958</td>
<td>−0.001±0.019</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Std Mean+SD</td>
<td>0.293±0.043</td>
<td>33.703±3.337</td>
<td>0.771±0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.593±0.098</td>
<td>138.530±16.448</td>
<td>1.096±0.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.827±0.120</td>
<td>266.116±21.497</td>
<td>1.295±0.026</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>STD Mean+SD</td>
<td>0.004±0.003</td>
<td>0.110±0.078</td>
<td>−0.000±0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>0.013±0.009</td>
<td>0.669±0.490</td>
<td>−0.001±0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>0.011±0.008</td>
<td>0.760±0.564</td>
<td>−0.000±0.028</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Std Mean+SD</td>
<td>1.758±0.118</td>
<td>17.511±2.488</td>
<td>0.826±0.040</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Med+IQR</td>
<td>5.058±0.345</td>
<td>257.858±28.043</td>
<td>1.215±0.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Std Wins Mean+SD</td>
<td>8.492±0.427</td>
<td>603.182±46.494</td>
<td>1.499±0.086</td>
</tr>
</tbody>
</table>
References


for large autoimmune biomarker discovery studies with ProtoArrays.

Proteomics and Systems Biology, 13, 2083–2087.


Chapter 3

A Bayesian Hierarchical Model for Signal Extraction from Protein Microarrays

3.1 Introduction

Pre-processing pipelines for laboratory assays are a series of transformations applied to raw data in order to ensure the data has features enabling specific statistical analyses, these are then leveraged to answer scientific questions of interest. Examples of these pipelines range in complexity from log transformation to reduce skew in a distribution of observations that are then used as outcomes in a linear regression, to the GATK pipeline for discovering novel genetic variants from high-throughput genome sequencing data (DePristo et al., 2011). Despite the wide range of pipelines, certain features are essential for a pipeline to be useful. First and foremost, the pipeline should ensure that the estimates used in a statistical analysis enable analysts to make direct inference about biological questions of interest, this often involves removing technical variation, from measurements. However, even within a single assay, a wide
range of pipelines can often achieve this goal and perform similarly (Bao et al., 2014; Hwang et al., 2018; Roy et al., 2018). Oftentimes, despite the presence of robust literature about best practices for pre-processing assay measurements, studies continue employ a wide range of bioinformatic pipelines, and differences in these pre-processing steps can substantially change downstream analysis and conclusions (Arora et al., 2020; Marizzoni et al., 2020; Siegwald et al., 2019). Incorporating appropriate levels of uncertainty in statistical analysis methods is one way to mitigate the sensitivity of conclusions to choices made in the pre-processing stage.

Protein microarrays are used to measure quantities of antibodies and other proteins in human or animal serum. They are extremely versatile and have been employed in numerous areas of medicine such as cancer research, infectious disease diagnostics and biomarker identification (Hartmann et al., 2009; Huang & Zhu, 2017; Nagele et al., 2011; Ramachandran et al., 2008; X. Zhu et al., 2006). Measurements from protein microarrays can be used to quantify relative levels of proteins in serum samples but are not an absolute measurement of protein concentration. Therefore, much of the analysis used to make biological inferences from protein microarray data is based on ranking the fluorescent intensity of tagged proteins within and across samples in a study. Ranks are particularly sensitive to data transformations and incorporating uncertainty in ranking procedures is a well known way of making ranks more robust to pre-processing choices and ranking procedures that incorporate uncertainty often produce optimal ranks (Laird & Louis, 1989; Lin et al., 2006; Shen & Louis, 1998). Bayesian frameworks that allow analysts to obtain
full posterior distributions for measurements from assays, have been used
to remove technical variation while also propagating uncertainty in several
bioinformatic pipelines (Cantarel et al., 2014; Gu et al., 2014; Shiraishi et al.,
2013). DNA microarray studies specifically have used empirical Bayes models
to improve the ranks of normalized gene expression levels, and maximize
the sensitivity in selecting differentially expressed genes with the largest ef-
teffects (Efron et al., 2001; Gottardo et al., 2003; Noma et al., 2010). However these
models rely heavily on the assumption that the majority of genes investigated
have similar expression levels across samples; it is unreasonable to assume
that the majority of proteins investigated in a protein microarray study are
present at similar levels across samples.

Bayesian models extract full posterior distributions for normalized mea-
measurements from protein microarrays, availability of these distributions enables
improved rankings and other inferences. Therefore, we develop and apply
a Bayesian hierarchical model that extracts the full posterior distribution of
normalized signal from protein microarrays. Our model uses information
from repeated measurements and controls across the assay to inform an er-
ror structure, that is then subtracted from observed measurements to obtain
an estimate of true underlying signal. We show that the model, with some
simplifications to improve computational efficiency, fits well to two protein
microarray datasets that use protein microarrays with different numbers of
probes, different proteins, and different manufacturing procedures, suggest-
ing that our model has the potential for applications to a wide variety of
protein microarray studies. Obtaining full posterior distributions from protein
microarray measurements will enable analysts to use ranking procedures that incorporate uncertainty not only reducing the sensitivity of ranks to pre-processing steps, but also producing optimal ranks enabling robust and accurate biological inference.

3.2 Methods

3.2.1 Protein Microarray Components

Protein microarrays are comprised of proteins spotted on glass slides and organized into rows and columns. A serum sample is added to the slide and antibodies in this sample bind to specific proteins on the array. Using fluorescent tags, binding events on the array are detected by a scanner that produces two observable signals. These are background signal, which is a measure of the fluorescent intensity in an outer ring surrounding a particular spot, and foreground signal which is the measure of the fluorescent intensity at the center of a particular spot (H. Zhu & Snyder, 2001). Protein array-based studies seek to characterize proteins in multiple human or animal samples, meaning that the number of individual samples analyzed is typically equal to the number of individual arrays in the study. We denote fluorescent intensity measurements from the scanner in a particular study with the notation $R_{i,t(p),p(i,j),j}$ or its corresponding transformed measurement $Y_{i,t(p),p(i,j),j}$, with the following subscripts:

- array number $i \in \{1, \ldots, I\}$
- spot $j \in \{1, \ldots, J_i\}$, where $J_i$ is the number of spots on array $i$ in a study,
since the number of spots on an array does not usually change within a study, we assume that \( J_i \) is the same for all \( i \in \{1, ..., I\} \).

- protein \( p(i, j) \in \{1, ..., P\} \), each individual spot \( j \) is spotted with a protein (or buffer if it is in the negative control category).

- spot type \( t(p) \), we consider three broad categories of proteins, negative controls, positive controls, and active spots and within these three categories there are multiple types of spots, \( t(p) \), that depend on the protein (or lack of protein in the case of negative controls). For the control protein types (negative and positive controls), \( t(p) \in \{1, ..., T\} \) and for active spots \( t(p) \in \{a_1, ..., a_A\} \). The designations of \( t(p) \) for the malaria and HuProt\textsuperscript{TM} arrays are shown in tables S2 and S1 respectively.

The HuProt\textsuperscript{TM} arrays used by Pan et al. (2017) contain 43776 active probes corresponding to approximately 16000 different proteins in the human proteome. The arrays also contain 4608 control proteins corresponding to multiple replicates of negative and positive controls, these are fully described in Table S1. In their study, Pan et al. (2017) analyze the antibody profiles of 100 individuals. Of these individuals, 20 were healthy and the remaining 80 had a form of lung cancer.

The arrays used by Kobayashi et al. (2019) contain 500 \textit{Plasmodium falciparum} and \textit{P. vivax} specific antigens referred to as active proteins, as well as multiple control proteins, fully described in S2. In this study, 429 samples from 290 individuals were collected across three sites in malaria endemic regions of Zambia and Zimbabwe. Across all three study sites, a random stratified
sampling scheme was employed for household selection and every individual present in the household at the time of visit was eligible for enrollment meaning that not all individuals in the study had malaria. For each individual enrolled in the study, at least one serum sample was collected and spotted on an array. Additionally, protein microarrays were spotted with sera from adults residing in the USA who had never traveled to a malaria endemic region to serve as controls.

### 3.2.2 Normalized Data

We use the output of the pre-processing pipeline described in Chapter 1 (Figure S1) as the data that will be fed into the Bayesian model. Briefly, the ratio of observed foreground signal to background signal at each probe \( Y' \) is log transformed. Then, using a robust linear model described by Sboner et al. (2009), array and subarray effects are estimated and subtracted from \( \log(Y') \) to produce \( \tilde{Y} \). Finally, these values are standardized using array specific sample means and sample standard deviations of control probes to produce \( Y \). We use the same subscripts on \( Y \) as in Section 3.2.1.

### 3.2.3 Proposed Models

#### 3.2.3.1 Complete Model

We use the following model that describes the relationship among observed, and then transformed data, \( Y \) as described in Section 3.2.2 and true underlying signal \( S \). We assume that true underlying signal, \( S \), adds to error \( e \), which is the technical variation remaining in the measurements after the pre-processing
pipeline (Figure S1), to produce the normalized observation $Y$. Equation 3.1
is the full hierarchical Bayesian model describing this relationship. In order to
leverage the control proteins present on all arrays to inform the distribution of
the errors $e_{i,t(p),p(i,j)}$ for $t(p) \in \{1, ..., T\}$, we consider the empirical
distributions of mean centered control probes $Y_{i,t(p),p(i,j),j} - \hat{\mu}_{i,t(p),p(i,j)}$ where
$\hat{\mu}_{i,t(p),p(i,j)}$ is the sample mean taken over all spots of type $t(p), \in \{1, ..., T\}$ on
each array (Figure S1). The observed densities of $e_{i,t(p),p(i,j)}$ for $t(p) \in \{1, ..., T\}$
are similar across all proteins, therefore it is reasonable to assume that the
errors in equation 3.1 are identically distributed, and well approximated by a
generalized beta-generated distribution (Alexander et al., 2012).

**Model 1**

\[
\sigma_i \sim iid \text{ Uniform}(0, 10^3), \ U|\alpha \sim Beta(\alpha, \alpha)
\]

\[
[e_{i,t(p),p(i,j),j} | \sigma_i, \alpha] = \sigma_i \frac{1}{c(\alpha)} \Phi^{-1}(U)
\]

\[
\mu_{i,a_1} \sim \text{Uniform}(-10^2, 10^2)
\]

\[
\tau_{i,a_1} \sim \text{Uniform}(0, 10^3)
\]

\[
[S_{i,a_1,p(i,j)} | \mu_{i,a_1}, \tau_{i,a_1}] \sim iid \ N(\mu_{i,a_1}, \tau_{i,a_1}^2)
\]

\[
\mathbb{1}_{t(p)=t} = 1 \text{ if } t(p) = t, \ 0 \text{ otherwise}
\]

\[
\mu_{i,t(p)\in\{1,\ldots,T\}} \sim \text{Uniform}(-10^2, 10^2)
\]

\[
[Y_{i,t(p),p(i,j),j} | \mu_{i,t(p),p(i,j),j}, S_{i,a_1,p(i,j),j}, \alpha, \sigma_i] = \sum_{t(p)\in\{1,\ldots,T\}} \mu_{i,t(p)} \times \mathbb{1}_{t(p)=t} + \sum_{t(p)\in\{1,\ldots,T\}} S_{i,a_1,p(i,j),j} \mathbb{1}_{t(p)=a_1} + e_{i,t(p),p(i,j),j}
\]

Importantly, \( c(\alpha) \) is a constant, used to ensure that the standard deviation of \( e_{i,t(p),p(i,j),j} \) is approximately \( \sigma_i \) for an array \( i \), and we approximate \( c(\alpha) \) in the following way:

\[
c(\alpha) = \text{sd}(\Phi^{-1}\{\text{Beta}^{-1}(\text{mesh}, \alpha, \alpha)\})
\]

(3.2)
where the mesh is a sequence of values between 0 and 1 increasing by increments of $\frac{1}{10000}$. We also use a numerical approximation procedure to estimate the maximum likelihood value of $\alpha$ given $Y_{i,t(p),p(i,j)}$ for $t(p) \in \{1, ..., T\}$ at roughly 2.5, supplementary section 3.5.1 details this procedure.

3.2.3.2 Simplified Models

Additionally, we propose and evaluate two simplifications to the complete model in Equation 3.1 to increase computational efficiency. First, letting $\alpha = 1$ is equivalent to assuming a Gaussian distribution for the errors (see
Model 2

\[ \sigma_i \sim iid \quad \text{Uniform}(0, 10^3) \]

\[ [e_{i,t(p),p(i,j),j}|\sigma_i, \alpha] \sim N(0, \sigma_i^2) \]

\[ \mu_{i,a_1} \sim \text{Uniform}(-10^2, 10^2) \]

\[ \tau_{i,a_1} \sim \text{Uniform}(0, 10^3) \]

\[ [S_{i,a_1,p(i,j)}|\mu_{i,a_1}, \tau_{i,a_1}] \sim iid \quad N(\mu_{i,a_1}, \tau_{i,a_1}^2) \]

\[ \mathbb{I}_{t(p)=t} = 1 \text{ if } t(p) = t, \ 0 \text{ otherwise} \]

\[ \mu_{i,t(p)\in\{1,...,T\}} \sim \text{Uniform}(-10^2, 10^2) \quad (3.3) \]

\[ [Y_{i,t(p),p(i,j),j}|\mu_{i,t(p)}, S_{i,a_1,p(i,j),j}, \alpha, \sigma_i] = \sum_{t(p)\in\{1,...,T\}} \mu_{i,t(p)} \times \mathbb{I}_{t(p)=t} + S_{i,a_1,p(i,j)} \mathbb{I}_{t(p)=a_1} + e_{i,t(p),p(i,j),j} \]

Second, we use an empirical point estimate of \( \sigma_i \) (see Equation 3.5), the variance of the error distribution, \( \hat{\sigma}_i^2 \):

\[ \hat{\sigma}_i^2 = \frac{\sum_{t(p)\in\{1,...,T\}} (n_{i,t(p)} - 1) (S_{i,t(p)}^2)}{\sum_{t(p)\in\{1,...,T\}} (n_{i,t(p)} - 1)} \]

\[ S_{i,t(p)}^2 = \frac{\sum_{j\in n_{i,t(p)}} (y_{i,t(p),p(i,j),j} - \bar{y}_{i,t(p)})}{(n_t(p) - 1)} \quad (3.4) \]
where \( n_{i,t(p)} \) is the number of spots of type \( t(p) \) on an array \( i \) and \( y_{i,t(p)} \) is the mean of \( y_{i,t(p),p(i,j),j} \) across all spots of type \( t(p) \) in array \( i \).

**Model 3**

\[
\begin{align*}
[e_{i,t(p),p(i,j),j} | \hat{\sigma}_r, \alpha] & \sim N(0, \sigma_i^2) \\
\mu_{i,a_1} & \sim \text{Uniform}(-10^2, 10^2) \\
\tau_{i,a_1} & \sim \text{Uniform}(0, 10^3) \\
[S_{i,a_1,p(i,j)} | \mu_{i,a_1}, \tau_{i,a_1}] & \sim \text{iid} N(\mu_{i,a_1}, \tau_{i,a_1}^2) \\
1_{t(p) = t} & = 1 \text{ if } t(p) = t, \ 0 \text{ otherwise} \\
H_{i,t(p) \in \{1, \ldots, T\}} & \sim \text{Uniform}(-10^2, 10^2) \quad (3.5) \\
[Y_{i,t(p),p(i,j),j} | \mu_{i,t(p)}, S_{i,a_1,p(i,j)}, \alpha, \sigma] & = \sum_{t(p) \in \{1, \ldots, T\}} \mu_{i,t(p)} \times 1_{t(p) = t} \\
& + S_{i,a_1,p(i,j)}1_{t(p) = a_1} + e_{i,t(p),p(i,j),j}
\end{align*}
\]

**3.2.3.3 Estimation Procedure**

We aim primarily to estimate the full posterior distribution \( S_{i,t(p),p(i,j)} | Y_{i,t(p),p(i,j)} \) for all active spots on each array, and will also evaluate the full posterior distributions of the other parameters, namely, \( \mu_{i,t(p)} \) for \( t(p) \in \{1, \ldots, T\} \), \( \mu_{i,a_1}, \tau_{i,a_1}, \) and \( \sigma \), where relevant. We use Markov Chain Monte Carlo (MCMC) estimation implemented through the Rjags package. We run each array \( i \) in a study independently but jointly estimate all parameters for all spots \( j \) and proteins.
Within each array, for each array, we use a burn-in of 5,000 draws, and posterior distributions are estimated using 10,000 draws after burn-in with no thinning.

3.2.3.4 Evaluating Model Fit

We evaluate the fit of our model using percentile based residuals mapped to the associated quantile of a standard Gaussian distribution as described in Berube et al. (2019). Specifically, for each \( y_{i,t(p),p(i,j)} \) in our dataset, we compute the transformed residual \( P_{i,t(p),p(i,j)}^\dagger \), as follows:

\[
P_{i,t(p),p(i,j)}^\dagger = \phi^{-1}\{D_i(y_{i,t(p),p(i,j),j})\}
\]

where \( D_i \) is the combination of posterior distributions across all \( S_{i,p(i,j)} \) and \( D_i(y_{i,t(p),p(i,j),j}) \) is the percentile location of the observed \( y_{i,t(p),p(i,j),j} \) in \( D_i \).

3.3 Results

We fit the model described in Model 1 (Equation 3.1) to the protein microarray data from the lung cancer (Pan et al., 2017) and malaria (Kobayashi et al., 2019) studies. We show that posterior distributions of the parameter \( S_{i,t(p),p(i,j)} \) for active \( (t(p) = a_1) \) proteins reveal clear separation across proteins with different corresponding \( Y \) values while also displaying shrinkage towards the mean \( \mu_{i,a_1} \), making them good candidates for downstream inference. The intercepts \( \mu_{i,t(p),p(i,j),j} \) and additional parameters \( \mu_{i,a_1}, \tau_{i,a_1} \) and \( \sigma_i \) all reveal good updating and convergence. We show that the simplifications in, Models 2 and 3 (Equations 3.3 and 3.5) also converge, provide similar posterior
distributions for $S_{i,t(p),p(i,j)}$ to those obtained with Model 1, and fit well to both datasets. Finally, we show that our models have the built-in flexibility to accommodate data from arrays with different control probes, different manufacturing processes, and different active proteins.

### 3.3.1 Estimation and Fit Using Model 1

The posterior distributions for the estimates of true underlying signals in the active spots ($t(p) = a_1$), $S_{i,t(p),p(i,j)}$ obtained by fitting Model 1 to the HuProt™ and malaria arrays, generally show good separation for proteins with different observed ($Y$) values in both datasets (Figures 3.1 and 3.2), these four arrays from each study are representative of all arrays), suggesting that these posterior distributions are good candidates for downstream analysis and inference. The posterior distributions for $S$ also reveal shrinkage towards the mean $\mu_{i,a_1}$ that is particularly pronounced for extreme values of $Y$, which is evidence that the model stabilizes these estimates, and provides good tradeoff between bias and variance for estimates of key parameters of interest, another useful quality for downstream analysis. Additionally, the posterior densities for the means of control proteins ($t(p) \in \{1, ..., T\}$), $\mu_{i,t(p)}$ are centered exactly at the observed and transformed means $Y_{i,t(p),p(i,j),j}$ for each type of control protein $t(p)$ in both studies (Figures S3 and S4), which reveals that the model performed accurate estimation and that the error structure was well approximated by a generalized beta-generated distribution. Trace plots for all parameters show good convergence (Figure S5 and S6) and the observed quantiles of percentile-based residuals ($P^4$, Equation 3.6) for all arrays are close to
those of a standard Gaussian distribution with a few deviations (Figures 3.3 and 3.4), suggesting a good overall fit for both studies.

![Malaria Arrays Posterior Distributions for S, Model 1](image)

**Figure 3.1:** Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria arrays (A,B,C,D) fit with Model 1. Vertical dashed lines are at the value $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$.

### 3.3.2 Estimation and Fit Using Models 2 and 3

Models 2 and 3 substantially reduce the computation time involved in obtaining full posterior distributions for all parameters, with Model 3 being slightly more efficient than Model 2 (Table 3.1), but both models show similar fit to both datasets when compared to Model 1. Despite the numerical computation of the maximum likelihood estimate in Section 3.5.1 revealing that the empirical error distribution for signal on both protein microarray
Figure 3.2: Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria HuProt™ arrays (E,F,G,H) fit with Model 1. Vertical dashed lines are at mean of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$.

Table 3.1: Running times for the three models on individual malaria and HuProt™ arrays (x1) as well as both datasets (x503 and x100 respectively).

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Malaria Array (x1)</th>
<th>Malaria Dataset (x503)</th>
<th>HuProt™ (x1)</th>
<th>HuProt™ Dataset (x100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9 mins</td>
<td>4577 mins</td>
<td>236 mins</td>
<td>23600 mins</td>
</tr>
<tr>
<td>2</td>
<td>2 mins</td>
<td>956 mins</td>
<td>52 mins</td>
<td>5160 mins</td>
</tr>
<tr>
<td>3</td>
<td>1 min</td>
<td>604 mins</td>
<td>49 mins</td>
<td>4880 mins</td>
</tr>
</tbody>
</table>
datasets \((e_{i,t(p),p(i,j)}, j)\) is well approximated by a generalized Beta-generated distribution with \(\alpha = 2.5\), assuming a Gaussian distribution for the error structure (Model 2) and using a point estimate \(\hat{\sigma}_i\) (Equation 3.4) for the standard deviation of this error structure (Model 3), produces similar overall fit to Model 1 in both datasets. Specifically, the quantiles of percentile residuals, \(P^{\dagger}\) (described in Section 3.2.3.4) computed for all \(S\) values across all arrays in the malaria study and a subset of arrays in the lung cancer study match fairly closely to those of a standard Gaussian distribution across all three models. The deviations from the standard Gaussian quantiles occur in the extremes, low and high \(Y\) values, but these deviations are similar across Models 1 and 2 (Figures 3.3 and 3.4). In the HuProt\(^{TM}\) arrays, Model 3 has larger deviations from the standard Gaussian quantiles, which can generally be attributed to the more extreme shrinkage of \(S\) posteriors towards the mean \(\mu_{i,d_1}\) in Model 3 than Model 2. The discrepancies between Model 2 fit and Model 3 fit in the HuProt\(^{TM}\) arrays is not observed in the malaria arrays. Therefore the most computationally efficient model, Model 3, can be used instead of Model 1 with almost identical results for the malaria arrays, and Model 2, still more computationally efficient than Model 1, can be used with almost identical results in the HuProt\(^{TM}\) arrays. However, the differences in model fit between Models 1 and 3 in the HuProt\(^{TM}\) arrays are small and the posterior estimates of \(S\) achieve important goals for downstream analysis, such as good separation for proteins with different observed \(Y\) values, therefore it may still be worthwhile to consider fitting Model 3 to the HuProt\(^{TM}\) arrays with appropriate sensitivity analyses. Importantly, the posterior distributions for \(\mu_{i,t(p)}\) in models 2 and 3 look nearly identical to those of model 1 for both the HuProt\(^{TM}\) and
malaria arrays (Figures S9, and S10) and trace plots reveal good convergence (Figures S11, S12, S13, and S14).

**Figure 3.3:** \( p^\dagger \) for all 503 malaria arrays across Models 1, 2, and 3. The red line is \( y = x \).

### 3.3.3 Comparing Estimation Across the two Datasets

Differences in the posterior distributions of \( \sigma_i, \mu_{i,t_1} \), and \( \tau_{i,t_1} \) across the two datasets for all three models reveal the flexible nature of the models, which has enabled the model to fit well and provide good estimates of \( S \) for downstream analysis across two vastly different protein microarray datasets. Estimation of the standard deviation of the error distribution, \( \sigma_i \) is largely informed by the control probe sets \( (t(p) \in \{1, \ldots, T\}) \), on each array. Both \( \hat{\sigma}_i \) in model 3 and the
Figure 3.4: $P^\dagger$ for a subset of four HuProt\textsuperscript{TM} arrays across Models 1, 2, and 3. The red line is $y = x$. 
posterior distributions of $\sigma_i$ in models 1 and 2, capture the differences in the control probe sets; the estimates are generally larger for the HuProt$^{TM}$ arrays that have more control probe spots than the malaria arrays. Furthermore, the point estimate $\hat{\sigma}_i$ is generally close to the mode of the posterior distribution for $\sigma_i$ (Models 1 and 2), which is typically narrower for HuProt$^{TM}$ arrays than for malaria arrays, suggesting that models 2 and 3 are able to use the information from different types of control probe sets to provide good estimates of the error structure which, in turn, informs estimates of the main parameter of interest, $S$ (Figures 3.5, 3.6, S15, and S16). In addition to differences across control probe sets, the two arrays also have different numbers of active spots ($t(p) = a_1$) with different target proteins, these are used to inform the estimation of hyper-parameters $\mu_{i,a_1}$ and $\tau_{i,a_1}$, and by extension $S$. The spread of posterior distributions of $\mu_{a_1}$ and $\tau_{a_1}$ for all three models fit to the HuProt$^{TM}$ arrays is smaller than those obtained from a fit to the malaria arrays, which is a result of the 15,402 active proteins ($t(p) = a_1$) on the HuProt$^{TM}$ as compared to the 1,038 active proteins ($t(p) = a_1$) on the malaria arrays (Figures 3.5, 3.6, S15, S16, S17, and S18). The increased precision of the posterior distributions for $\mu_{a_1}$ and $\tau_{a_1}$ reflect the models ability to accommodate major differences in array composition, particularly in the number of active spots, and reflect these differences in the level of uncertainty in estimating certain parameters. Both the differences in the standard deviation of the error distribution as well as in the posterior distributions of hyper parameters $\mu_{a_1}$ and $\tau_{a_1}$ reveal that Models 1, 2, and 3 are able to use different sets of control and active spots to produce posterior distributions for $S$ with desirable qualities, such as clear separation of posteriors corresponding to different $Y$ values, across both datasets without
adjustments to the model.

![Figure 3.5: Posterior distributions for $\sigma_i$, $\mu_{i,t_1}$, and $\tau_{i,t_1}$ for four malaria (A,B,C,D) arrays fit with Model 2. Vertical dashed lines are at the value $\hat{\sigma}_i$.](image)

**3.4 Discussion and Future Work**

Protein microarray data have the ability to provide insight into important biological processes. However, between-array variation even after pre-processing, and the inability of the assay to directly measure protein concentration using known standards, only within-array comparisons of fluorescent intensities are meaningful. For this reason, relative intensity (ratio to a typical or baseline intensity), or use of protein-specific intensity ranks are necessary. All summaries are sensitive to pre-processing and normalization choices, but
Figure 3.6: Posterior distributions for $\sigma_i$, $\mu_{i,j}^{(0)}$, and $\tau_{i,j}^{(0)}$ for four HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the value $\hat{\sigma}_i$. 
methods that accommodate uncertainty, especially from Bayesian frameworks, produce optimal ranks that are less likely to change drastically under different pre-processing and normalization conditions.

We have developed and applied a model that extracts a full posterior distributions of normalized signal for every protein in every serum sample of a protein microarray study. We show that this model produces estimates that are useful for downstream inference, and fits well to two protein microarray datasets that use protein microarrays with substantially different manufacturing procedures. We also show that simplifications to this model that substantially decrease computation time, produce similar estimates and fit equally well to the two datasets. Finally, we show that the model has built-in flexibility that can accommodate vastly different protein microarrays. The full posterior distributions obtained from the models can be used as inputs for variety of ranking methods that incorporate uncertainty (Henderson & Newton, 2016; Laird & Louis, 1989; Lin et al., 2006; Shen & Louis, 1998).

Although investigation of percentile residuals for both datasets reveals a good fit for both the complete and simplified models, there are some deviations from the standard Gaussian distribution that warrant further investigation through simulation. Specifically, comparisons of percentile residuals for misspecified models is needed in order to better understand the implications of the observed deviations. Additionally, although both simplifications (Gaussian errors, and the use of \( \hat{\sigma}_i \)) to the complete model in Equation 3.1 produce very similar estimates and fit, the point estimate, \( \hat{\sigma}_i \) for the HuProt\textsuperscript{TM} arrays is either larger or smaller than most draws for the posterior distribution of \( \sigma_i \).
obtained from fitting the data with Models 1 and 2, which may indicate the need to consider a subset of control probes on either array to better inform our estimates of $\sigma_i$. This merits further investigation, both through simulations and through other datasets. In spite of the differences in manufacturing and features of the arrays used in the two datasets, our model fits quite well to both, suggesting the potential for our method to be effective across multiple protein microarray studies. However, we only evaluated two published studies, representing a fairly narrow sample of the currently available protein microarrays. It is clear that the amount of technical variability present is dependent at least in part on the manufacturing process, therefore a more comprehensive evaluation of our model on different types of protein microarrays is warranted.

There are a broad range of bioinformatic pipelines available for various assays, and many accomplish the goal of reducing technical variation and enabling researchers to make inference about questions of biological importance. However, choices made in pre-processing steps often change downstream results and conclusions. Propagating appropriate levels of uncertainty can mitigate the sensitivity of results to the choice of bioinformatic pipeline, and Bayesian models enable analysts to accomplish this goal while providing other advantages, for instance, the use of full posterior distributions for key parameters can be used to obtain optimal ranks. This is especially pertinent for protein microarrays, where ranks are the basis for most biological inference. While Bayesian models have been developed and evaluated for similar assays
Table S1: Values of $t(p)$ on the HuProt\textsuperscript{TM} array.

<table>
<thead>
<tr>
<th>Protein Category</th>
<th>Protein ($p(i, j)$)</th>
<th>Number of Spots ($j$)</th>
<th>Type $t(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Control</td>
<td>GST 10 ng/µl</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>GST 50 ng/µl</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>GST 100 ng/µl</td>
<td>48</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>GST 200 ng/µl</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mouse-anti-biotin</td>
<td>48</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Rabbit-anti-biotin</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>BSA</td>
<td>48</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Buffer</td>
<td>816</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Empty</td>
<td>3072</td>
<td>9</td>
</tr>
<tr>
<td>Positive Control</td>
<td>Histone 1</td>
<td>48</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Histone 2 (A+B)</td>
<td>48</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Histone 3</td>
<td>48</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Histone 4</td>
<td>48</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Alexa Fluor labeled IgG</td>
<td>48</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Rhodamine+ Alexa Fluor labeled IgG</td>
<td>48</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Biotin-BSA</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Mouse IgM</td>
<td>48</td>
<td>17</td>
</tr>
<tr>
<td>Active Proteins</td>
<td>Active Proteins</td>
<td>48384</td>
<td>$a_1$</td>
</tr>
</tbody>
</table>

such as DNA microarrays, to date, no such investigations have been published for protein microarrays. Our Bayesian model is developed specifically for this assay and therefore has the potential to improve the robustness and accuracy of conclusions drawn from protein microarray data by propagating appropriate levels of uncertainty through bioinformatic pipelines.
Table S2: Values of $t(p)$ on the malaria array.

<table>
<thead>
<tr>
<th>Control Type</th>
<th>Protein $p(i,j)$</th>
<th>Number of Spots $j$</th>
<th>Type $t(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Control</td>
<td>Blank1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Blank2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Empty</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>No DNA Reaction</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>TTBS</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>Positive Control</td>
<td>anti-human IgG 0.003</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>anti-human IgG 0.03</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>anti-human IgG 0.3</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>anti-human IgG 0.001</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>anti-human IgG 0.01</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>anti-human IgG 0.1</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.003</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.03</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.3</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.001</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.01</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>IgG mix 0.1</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Active Proteins</td>
<td>Active Proteins</td>
<td>1038</td>
<td>$a_1$</td>
</tr>
</tbody>
</table>
3.5 Supplement

3.5.1 Estimating $\alpha$

To estimate $\alpha$, we will use a maximum likelihood estimation procedure that uses an empirical estimate of the error distributions $e_{i,t(p),p(i),j}$ for $t(p) \in \{1, ..., T\}$. According to our proposed model in equation 3.1, the distribution of $e_{i,t(p),p(i),j}$ across all arrays can be estimated from the data by considering the distribution standardized control probes $Y_{i,t(p),p(i),j} - \hat{\mu}_{i,t(p),p(i),j}$, where $\hat{\mu}_{i,t(p),p(i),j}$ is the sample mean taken over all spots of type $t(p) \in \{1, ..., T\}$ on each array. We obtain an approximate maximum likelihood estimate (MLE) for $\alpha$ based on the following likelihood:

$$L(Y_{i,t(p),p(i),j}|\alpha) = \prod_{j} \frac{1}{B(\alpha, \alpha)} \times \Phi \left( \frac{y_{i,t(p),p(i),j} - \hat{\mu}_{i,t(p),p(i),j}}{\hat{\sigma}} \right)^{\alpha-1}$$

$$\times \left[ 1 - \Phi \left( \frac{y_{i,t(p),p(i),j} - \hat{\mu}_{i,t(p),p(i),j}}{\hat{\sigma}} \right) \right]^{\alpha-1}$$

$$\times \Phi \left( \frac{y_{i,t(p),p(i),j} - \hat{\mu}_{i,t(p),p(i),j}}{\hat{\sigma}} \right) \times \frac{c(\alpha)}{\hat{\sigma}} \quad (3.7)$$

Where $\hat{\sigma}$ is the sample standard deviation of the set $Y_{NC}$, or all observations of negative control probes across all arrays. Given that it is not possible to obtain a closed form equation for the maximum likelihood estimate (MLE) of $\alpha$ with the likelihood in Equation 3.7, we propose a numerical estimation that uses a fine-grained “grid” of possible $\alpha$ values ranging from 1 – 17 and increasing by increments of $\frac{1}{1000}$ to identify the value of $\alpha$ that produces the highest likelihood value. We replicate this estimation procedure for positive
control proteins and then for the combined positive and negative control proteins. Figure S2 shows the log likelihood values of equation 3.7 for each of the three sets of proteins in the malaria and Huprot\textsuperscript{TM} arrays. The value of $\alpha$ that produces the maximum log likelihood value is shown with a red vertical line.
Figure S1: Observed distributions of $e_{i,t(p),p(i,j),j}$ for negative, positive, and combined positive and negative controls.
Figure S2: Log likelihood values of equation 3.7 for each of the three sets of proteins in the malaria arrays, negative control (A and D), positive control (B and E) and combined positive and negative control (C and F) proteins across values of \( \alpha \) in the grid. The value of \( \alpha \) that produces the maximum log likelihood value is shown with a red vertical line.
Figure S3: Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria arrays (A,B,C,D) fit with Model 1. Vertical dashed lines are at the means of values $Y_{i,t(p)}$ corresponding to the proteins with type $t(p)$. 
Figure S4: Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four HuProt$^\text{TM}$ arrays (E,F,G,H) fit with Model 1. Vertical dashed lines are at the means of values $Y_{i,t(p)}$ corresponding to the proteins with type $t(p)$. 
Figure S5: Trace plots for estimated parameters, $S_i, \mu_{i,t(p)}, \mu_{i,v1}, \tau_{i,v1}, \sigma_t$ obtained by fitting Model 1 to four malaria arrays (A,B,C,D).
Figure S6: Trace plots for estimated parameters, $S$, $\mu_{i,j}(p)$, $H_{i,j}$, $T_j$, $a_j$, $c_j$ obtained by fitting Model 1 to four HuProt$^TM$ (E,F,G,H) arrays.
Figure S7: Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria (A,B,C,D) and HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the value or means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$.
Figure S8: Posterior distributions for $S_{i,t(p),p(i,j),j}$ for five different proteins on four malaria (A,B,C,D) and HuProt$^TM$ (E,F,G,H) arrays fit with Model 3. Vertical dashed lines are at the value or means of values $Y_{i,t(p),p(i,j),j}$ corresponding to the protein $p(i,j)$. 

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

85
Figure S9: Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria (A,B,C,D) and HuProt$^\text{TM}$ (E,F,G,H) arrays fit with Model 2. Vertical dashed lines are at the means of values $Y_{i,t(p)}$ corresponding to the proteins with type $t(p)$. 
Figure S10: Posterior distributions for $\mu_{i,t(p)}$ for five different types $t(p)$ if control proteins on four malaria (A,B,C,D) and HuProt$^\text{TM}$ (E,F,G,H) arrays fit with Model 3. Vertical dashed lines are at the means of values $Y_{i,t(p)}$ corresponding to the proteins with type $t(p)$. 
Figure S11: Trace plots for estimated parameters, $S, \mu_{i,t}(p), \mu_{t,1}, \tau_{i,1}, \sigma_i$ obtained by fitting Model 2 to four malaria arrays (A,B,C,D).
Figure S12: Trace plots for estimated parameters, $S_i, \mu_i, \mu_{i1}, \tau_i, a_1, \sigma_t$ obtained by fitting Model 2 to four HuProtTM (E,F,G,H) arrays.
Figure S13: Trace plots for estimated parameters, $S$, $\mu_{i,t}(p)$, $\mu_{i,1}$, $\tau_{i,1}$, $\sigma_{i}$ obtained by fitting Model 3 to four malaria arrays (A, B, C, D).
Figure S14: Trace plots for estimated parameters, $S$, $\mu_i(t(p))$, $\mu_i$, $\tau_i$, $a_1$, $\sigma_i$ obtained by fitting Model 3 to four HuProt$^TM$(E,F,G,H) arrays.
Figure S15: Posterior distributions for $\sigma_\ell$, $\mu_{i,\ell_1}$, and $\tau_{i,\ell_1}$ for four malaria (A,B,C,D) arrays fit with Model 1.
Figure S16: Posterior distributions for $\sigma_i$, $\mu_{i,\alpha_1}$, and $\tau_{i,\alpha_1}$ for four HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 1.
Figure S17: Posterior distributions for $\mu_{i,d_1}$ and $\tau_{i,d_1}$ for four malaria (A,B,C,D) arrays fit with Model 3.
Figure S18: Posterior distributions for $\mu_{i,d_1}$ and $\tau_{i,d_1}$ for four HuProt$^{TM}$ (E,F,G,H) arrays fit with Model 3.
References


DePristo, M., E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D.,


Kobayashi, T., Jain, A., Liang, L., Obiero, J., Hamapumbu, H., Stevenson, J., Thuma, P., Lupiya, J., Chaponda, M., Mulenga, M., Mamini, E., Mharakurwa, S., Gwanzura, L., Munyati, S., Matambu, S., Felgner,


Chapter 4

Using Full Posterior Distributions of Normalized Signal From Protein Microarrays to Perform Differential Analysis: Data Example

4.1 Introduction

Protein microarrays are a relatively new technology used to quantify levels of thousands of target proteins in a wide variety of samples and therefore have potentially broad utility across various areas of public health and medical research. Most commonly, studies involving protein microarrays seek to measure protein levels across groups of samples with specific characteristics, such as the presence or absence of cancer or exposure to infectious pathogens, and then to correlate differences in levels of certain proteins with those observable characteristics (Hartmann et al., 2009; Huang & Zhu, 2017; Nagele et al., 2011; Ramachandran et al., 2008; X. Zhu et al., 2006). These types of studies are typically useful in situations where the list of proteins that could be
associated with a phenotype numbers in the thousands or tens of thousands, and laboratory techniques that can more precisely correlate concentrations of proteins in samples with observable characteristics, such as enzyme-linked immunosorbent assays (ELISA) cannot accommodate such a high number of proteins (Kumble, 2003; Uhlen et al., 2016; Yang et al., 2016; H. Zhu et al., 2012). Therefore, protein microarrays are an important tool to narrow-down a list of candidate biomarkers for further study. The advantages afforded by the high throughput nature of protein microarrays, such as the ability to screen tens of thousands of proteins at once, also pose challenges. These assays are not constructed in a way that enable analysts to obtain objective measurements of protein concentration, as one might be able to do in an ELISA or a multiplex bead array assay (MBA) with a standard curve constructed with measurements from known protein concentrations (Elshal & J.P., 2006; Engvall & Perlmann, 1972; G. & Schreiber, 2000). Therefore measurements obtained from protein microarrays are only meaningful relative to other measurements in that same assay, for instance, in sample A protein X is present at higher levels than protein Y and in sample B protein X is present at lower levels than protein Y. The high levels of technical variation in these assays (described in Chapter 2) further complicate analysis and make comparisons of protein levels, even across samples within the same study, difficult.

In order to address some of these challenges, the measurements from protein microarrays are often pre-processed to remove as much technical variation as possible, and ranked by fluorescent intensity, so that conclusions can be based on relative differences within the assay (Diez et al., 2012; Eckel-Passow
et al., 2005; Rosenberg & Utz, 2015; Sboner et al., 2009; Turewicz et al., 2016; Turewicz et al., 2013). The pre-processing steps and ranking methods can therefore have a profound impact on conclusions drawn from these data. We focus on examples of how different pre-processing steps and ranking methods impact the results of a previously published protein microarray study. In their study, Kobayashi et al. (2019) measure levels of response to various malaria (*Plasmodium falciparum* and *P. vivax*) antigens in the serum of individuals who reside in malaria endemic regions of Zambia and Zimbabwe that have historically different levels of transmission (high, recent resurgence and near elimination) (Kanyangara et al., 2017; Mharakurwa et al., 2012; Moss et al., 2015). They use protein microarray data to identify markers of past malaria exposure that could be used for more sensitive malaria surveillance, particularly in low transmission settings, where capturing incident infections through traditional methods like rapid diagnostic tests, microscopy, and molecular testing is difficult and resource intensive. They rank antibody levels estimated from the protein microarray data using point estimates (median values) of fluorescent intensity and observe an association between the 30 antigens with the highest response and the region of residence, but only observe this association among children. Additionally, these authors observe a higher overall levels of response for individuals with a positive rapid diagnostic test (RDT) for malaria at the time of sampling than those with a negative RDT test and find some antibodies that are present at differential levels across these two groups.

Using the pre-processing pipeline in Chapter 2 and the outputs of the
Bayesian model in Chapter 3 (full posterior distributions of normalized measurements), we rank the level of protein in each sample and across studies using methods described in Lin et al. (2006), Laird and Louis (1989) and Shen and Louis (1998); these steps are different from those used in Kobayashi et al. (2019) and we discuss the advantages of our pre-processing pipeline and Bayesian model in Chapters 2 and 3. Using these methods we find that the ranks obtained using information from full posterior distributions are different than those obtained by Kobayashi et al. (2019), and from those in Chapter 2, in particular the list of 30 antigens with the highest response across samples changes and therefore our proposed ranking method offers a new set of antigens that could be further studied. We find similar differences when comparing the ranks of proteins on the HuProt\textsuperscript{TM} arrays after pre-processing to those obtained with methods that use information from full posterior distributions. Additionally, using these ranks, we find similar overall patterns in response to malaria antigens across regions of residence, age, and RDT test result, to previously published results, but additional antigens that may play a key role in differentiating the immunological response to malaria antigens across these groups. We also find that the differences in antigen ranks across regions of residence are also present in adults, and those with a negative RDT at the time of sampling, suggesting that further study of reactivity to subsets of antigens on these arrays could lead improved methods for malaria surveillance.
4.2 Methods

4.2.1 Protein Microarray Components

We denote fluorescent intensity measurements from the scanner in a particular study with the notation $R_{i,t(p),p(i,j),j}$ or its corresponding transformed measurement $Y_{i,t(p),p(i,j),j}$, with the following subscripts:

- array number $i \in \{1, \ldots, I\}$
- spot $j \in \{1, \ldots, J_i\}$, where $J_i$ is the number of spots on array $i$ in a study, since the number of spots on an array does not usually change within a study, we assume that $J_i$ is the same for all $i \in \{1, \ldots, I\}$.
- protein $p(i,j) \in \{1, \ldots, P\}$, each individual spot $j$ is spotted with a protein (or buffer if it is in the negative control category).
- spot type $t(p)$, we consider three broad categories of proteins, negative controls, positive controls, and active spots and within these three categories there are multiple types of spots, $t(p)$, that depend on the protein (or lack of protein in the case of negative controls). For the control protein types (negative and positive controls), $t(p) \in \{1, \ldots, T\}$ and for active spots $t(p) \in \{a_1, \ldots, a_A\}$.

Further details on the types $t(p)$ of proteins can be found in Chapter 3.

The arrays used by Kobayashi et al. (2019) contain 500 $P. falciparum$ and $P. vivax$ specific antigens referred to as active proteins, as well as multiple control proteins, fully described in S2. In this study, 429 samples from 290 individuals were collected across three sites in malaria endemic regions of Zambia.
and Zimbabwe: Macha in Choma District, Zambia, an area of consistently declining transmission for a decade, now nearing elimination, Honde Valley in Mutasa District, Zimbabwe, an area that experienced resurgence after a long period of low transmission and subsequently controlled that resurgence with effective interventions, and Nchelenge in Luapula Province, Zambia a region of consistently high transmission. Across all three study sites, a random stratified sampling scheme was employed for household selection and every individual present in the household at the time of visit was eligible for enrollment meaning that not all individuals in the study had malaria, demographic characteristics of the study sample are in Table S1. For each individual enrolled in the study, at least one serum sample was collected and spotted on an array. Additionally, protein microarrays were spotted with sera from adults residing in the USA who had never traveled to a malaria endemic region to serve as controls.

The HuProt\textsuperscript{TM} arrays used by Pan et al. (2017) contain 43776 active probes corresponding to approximately 16000 different proteins in the human proteome. The arrays also contain 4608 control proteins corresponding to multiple replicates of negative and positive controls, these are fully described in Table S1. In their study, Pan et al. (2017) analyze the antibody profiles of 100 individuals. Of these individuals, 20 were healthy and the remaining 80 had a form of lung cancer. We use the arrays spotted with the 20 healthy controls to compare ranking methods.
4.2.2 Pre-Processing Pipeline

We use the output of the pre-processing pipeline described in Chapter 2 (Figure S1).

4.2.3 Bayesian Model

We use the outputs of Model 3 described by Equations 3.5 and 3.4 to compute ranks.

Importantly, since we are interested in performing inference on the signal of active spots, we use the full posterior distributions of the $S_{i,a_1,p(i,j)}$ with three chains of 10,000 posterior draws each for ranking.

4.3 Ranking

Given that a high level of between array technical variation still remains after pre-processing steps, basing an analysis on within array comparisons mitigates the degree to which technical variation may confound or influence downstream results. Therefore, we propose one ranking method (method A) with two different loss functions (squared-error loss, and misclassification probability (Laird & Louis, 1989; Lin et al., 2006)), that computes ranks within each array first, and then combines these ranks across arrays. However, in order to make meaningful comparisons across arrays we consider two other methods of ranking. One method uses the misclassification probability loss function, but computes the ranks of proteins in each array relative to the overall distribution of $S$ across all proteins in all arrays (method B). The final
method (method C) uses the array specific ranks from method A but computes differences in rank of various proteins across subsets of samples.

4.3.1 Method A, Squared Error Loss (SEL)

We define the ranks of the true $S_{i,p(i,j)}$ to be:

$$T_{i,p(i,j)}(S_{i,p(i,j)}) = \text{rank of } S_{i,p(i,j)} \text{ in array } i := \sum_{k=1}^{P} I\{S_{i,p(i,j)} \geq S_{i,k}\} \quad (4.1)$$

so that the smallest $S$ will have rank 1 and the largest will have rank $P$. The estimator that minimizes squared error loss is the posterior mean:

$$\hat{T}_{i,p(i,j)}(Y) = E_{S|Y} \left[ T_{i,p(i,j)} \right] = \sum_{k=1}^{P} pr \left( S_{i,p(i,j)} \geq S_{i,k} \right) \quad (4.2)$$

We use burned in MCMC draws compute this quantity by computing the rank of each protein in each array for each MCMC draw, and then averaging those ranks across the MCMC draws. We compute $\hat{T}_{i,p(i,j)}$ as well as the integer version of these:

$$\hat{T}_{i,p(i,j)}(Y) = \text{within array ranks of } \hat{T}_{i,p(i,j)} \quad (4.3)$$

We report these values, combined across arrays, specifically the simple average of $\hat{T}_{i,p(i,j)}$ across arrays:

$$\hat{T}_{p(i,j)} = \frac{\sum_{i} \hat{T}_{i,p(i,j)}}{I} \quad (4.4)$$

As well as the integer version of these across arrays:

$$\hat{T}_{p(i,j)} = \text{rank of } \hat{T}_{p(i,j)} \quad (4.5)$$
4.3.2 Method A, Classification with above $\gamma$ Loss

We consider an estimator that will allow us to identify the top $\gamma \%$ of proteins within arrays. Specifically we let:

$$\pi_{i,p(i,j)}(\gamma) = Pr\{S_{i,p(i,j)} > k_i(\gamma)\}$$  \hfill (4.6)

where MCMC draws are indexed by $\nu \in \{1, ..., M\}$ and where $k_i(\gamma)$ is the value in the empirical distribution of $S_{i,p(i,j),\nu}$ over all $p(i,j), \nu$ in array $i$ that corresponds to the $(100 - \gamma)$ percentile. We estimate the value $\hat{\pi}_{i,p(i,j)}(\gamma)$ by defining the following indicator function:

$$I_{S_{i,p(i,j)},\nu} = \begin{cases} 0 & \text{if } S_{i,p(i,j),\nu} < k_i(\gamma) \\ 1 & \text{if } S_{i,p(i,j),\nu} \geq k_i(\gamma) \end{cases}$$

and therefore:

$$\bar{\pi}_{i,p(i,j)} = \frac{\sum_{\nu} I_{S_{i,p(i,j),\nu}}}{M}$$  \hfill (4.7)

We also consider the ranks of these $\pi$ values:

$$\hat{Q}_{i,p(i,j)} = \text{ranks within array of } \bar{\pi}_{i,p(i,j)}$$  \hfill (4.8)

We also define:

$$\tilde{\pi}_{p(i,j)} = \frac{\sum_i \bar{\pi}_{i,p(i,j)}}{I}$$  \hfill (4.9)

And the integer ranks of these values:

$$\tilde{Q}_{p(i,j)} = \text{ranks of } \tilde{\pi}_{p(i,j)}$$  \hfill (4.10)
We consider $\gamma = 3, 5$ and $10$ in our analysis. In the malaria arrays the top $3\%$ of proteins corresponds to roughly to the top $30$ proteins, the top $5\%$ to the top $50$ and the top $10\%$ to the top $100$.

4.3.3 Method B, Classification with above $\gamma$ Loss

We consider an estimator that will allow us to identify the top $\gamma\%$ of proteins across all arrays. Specifically we let:

$$\hat{\xi}_{i,p(i,j)}(\gamma) = Pr \left\{ S_{i,p(i,j)} > k(\gamma) \right\}$$  \hspace{1cm} (4.11)

where MCMC draws are indexed by $\nu \in \{1, ..., M\}$ and where $k(\gamma)$ is the value in the empirical distribution of $S_{i,p(i,j),\nu}$ over all $p(i,j), \nu$ across all arrays, that corresponds to the $(100 - \gamma)$ percentile. We estimate the value $\hat{\xi}_{i,p(i,j)}(\gamma)$ by defining the following indicator function:

$$I_{S_{i,p(i,j),\nu}} = \begin{cases} 0 & \text{if } S_{i,p(i,j),\nu} < k(\gamma) \\ 1 & \text{if } S_{i,p(i,j),\nu} \geq k(\gamma) \end{cases}$$

and therefore:

$$\bar{\xi}_{i,p(i,j)} = \frac{\sum_{\nu} I_{S_{i,p(i,j),\nu}}}{M}$$  \hspace{1cm} (4.12)

We also consider the ranks of these $\xi$ values:

$$\hat{W}_{i,p(i,j)} = \text{ranks within array of } \bar{\xi}_{i,p(i,j)}$$  \hspace{1cm} (4.13)
We also define:

\[ \tilde{\xi}_{p(i,j)} = \frac{\sum_i \tilde{\xi}_{i,p(i,j)}}{I} \]  

(4.14)

And the integer ranks of these values:

\[ \tilde{W}_{p(i,j)} = \text{ranks of } \tilde{\xi}_{p(i,j)} \]  

(4.15)

We average these \( \tilde{\xi}_{i,p(i,j)} \) values across a subset of arrays that share key characteristics, such as region of residence, age, and RDT result (negative or positive) status in the malaria arrays. Specifically we consider \( \tilde{\xi}_{\text{Macha},p(i,j)} \), the average of \( \tilde{\xi}_{i,p(i,j)} \) across all arrays \( i \) with samples from Macha and similar values for Nchelenge and Honde Valley (\( \tilde{\xi}_{\text{Nchelenge},p(i,j)}, \tilde{\xi}_{\text{Honde},p(i,j)} \)).

We also perform similar subsetting for RDT result and age: \( \tilde{\xi}_{\text{pos},p(i,j)} \), the average of \( \tilde{\xi}_{i,p(i,j)} \) across all arrays \( i \) with samples from RDT positive individuals and similar values for RDT negative samples (\( \tilde{\xi}_{\text{neg},p(i,j)} \)), and \( \tilde{\xi}_{\text{child},p(i,j)} \), the average of \( \tilde{\xi}_{i,p(i,j)} \) across all arrays \( i \) with samples from children aged 5 or under and similar values for people over the age of 5 (\( \tilde{\xi}_{\text{adult},p(i,j)} \)).

For each of these subsets of \( \tilde{\xi}_{\text{group},p(i,j)} \) we also compute the associated \( \tilde{W}_{\text{group},p(i,j)} \).

Importantly, since this ranking method requires compares protein ranks directly across all arrays without first computing within-array ranks, the technical variation still present between arrays after the steps outlined in Chapter 2 is likely influencing downstream conclusions. Therefore results obtained with this ranking method may be confounded by the presence of this technical variation; more so than results obtained from Methods A and C.
4.3.4 Method C, Squared Error Loss (SEL)

We use the ranks $T_{i,p(i,j)}$ and produce $\tilde{T}_{\text{group},p(i,j)}$ by averaging the $T_{i,p(i,j)}$ across groups of arrays that share certain characteristics, much like in Section 4.3.3, we consider the following subgroups: geographic location (Macha, $\tilde{T}_{\text{Macha},p(i,j)}$, Nchelenge $\tilde{T}_{\text{Nchelenge},p(i,j)}$, and Honde Valley $\tilde{T}_{\text{Honde},p(i,j)}$), age (adults, $\tilde{T}_{\text{adult},p(i,j)}$, and children, $\tilde{T}_{\text{child},p(i,j)}$), and RDT result (positive, $\tilde{T}_{\text{pos},p(i,j)}$, and negative, $\tilde{T}_{\text{neg},p(i,j)}$). In order to compare the ranks of antigens across groups we define

$$D_{\text{group 1},\text{group 2},p(i,j)} = \tilde{T}_{\text{group 1},p(i,j)} - \tilde{T}_{\text{group 2},p(i,j)}$$

(4.16)

This comparison could also be done using ranks $\check{Q}_{\text{group},p(i,j)}$.

4.3.4.1 Analysis of Ranks

In order to investigate the ranks with the misclassification loss function (Methods A and B), we propose one plot and some summary statistics. First, as described in Lin et al. (2006) we plot the values $\bar{\pi}_{i,p(i,j)}$ on the $y$-axis, and the quantile value of their associated ranks $\hat{\check{Q}}_{i,p(i,j)}$, computed as $\frac{2\times\check{Q}_{i,p(i,j)} - 1}{2 \times p}$. This plot is monotone increasing, and in particular, allows the analyst to highlight antigens that have both a high rank (high value of $\hat{\check{Q}}_{i,p(i,j)}$) and a high posterior probability of being classified in the top $\gamma\%$ of antigens ($\bar{\pi}_{i,p(i,j)}$), thereby providing a visual representation both of ranks and of the uncertainty associated with those ranks. Note that the same can be repeated with $\bar{\xi}_{i,p(i,j)}$ and associated ranks, $\hat{W}_{i,p(i,j)}$.

In order to further summarize the posterior probability of being classified
in the top $\gamma\%$ for the high-ranking antigens, we report the minimum $(\bar{\pi}_{i,p(i,j)})$ value of antigens whose rank ($\hat{Q}_{i,p(i,j)}$) is in the top $\gamma\%$ as well as the number of antigens whose rank ($\hat{Q}_{i,p(i,j)}$) is in the top $\gamma\%$ and who have a high (greater than 0.5 and greater than 0.75) posterior probability $(\bar{\pi}_{i,p(i,j)})$ of being classified in the top $\gamma\%$. In this way we not only identify high ranking antigens but high ranking antigens with high levels of certainty in terms of classification probability.

The process described on the array level can also be repeated for probabilities that are averaged across arrays, $(\bar{\pi}_{p(i,j)}$, and $\bar{\xi}_{p(i,j)})$ and associated ranks $(\bar{Q}_{p(i,j)}$, and $\bar{W}_{p(i,j)})$.

### 4.4 Results

Here, we find that using the ranks $\bar{T}_{p(i,j)}$ and $\bar{Q}_{p(i,j)}$ produce a different list of 30 high intensity proteins than previous methods in both the malaria and lung cancer arrays, and using the posterior probabilities of classification $\bar{\pi}_{p(i,j)}$, we show a high level of uncertainty among the antigens in this list, hi-lighting the importance of ranking methods that pay attention to uncertainty. Furthermore, we show that patterns of reactivity to antibodies across region of residence (Honde Valley, Macha and Nchelenge), age (under and over 5) and RDT status (positive and negative) elucidated from the ranks $\bar{W}_{\text{group},p(i,j)}$ are similar to previously published ones, but using these ranks, as well as differences across the ranks $\bar{T}$, that rely on within-array ranks and therefore may be less altered by between-array technical variation, we identify additional proteins of interest that may differentiate between these groups and merit
further investigation. Finally, we show that using information from the ranks $\hat{W}_{\text{group},p(i,j)}$, there are differences in reactivity to malaria antigens across the three regions in adults, these were previously visible only in children. These findings, coupled with differences among RDT negative samples across the three regions suggest that reactivity to groups of these antigens could be a marker of historical exposure to malaria and therefore that subsets of these antigens could useful for malaria surveillance.

4.4.1 Overall Ranks

The method A ranks minimizing both squared error loss and classification probability ($\hat{T}_{p(i,j)}$ and $\hat{Q}_{p(i,j)}$ reveal a different set of highly reactive antigens than those previously identified in Kobayashi et al. (2019) and in Chapter 2. Given that in this dataset, the top 3% of antigens corresponds roughly to the top 30 we compare the values $\hat{Q}_{p(i,j)}$ with $\gamma = 3$, to the point estimate ranks of $Y$ in Chapter 2, and to those published in Kobayashi et al. (2019). We see that even in the classification of the 30 antigens with the highest overall response, the method of ranking and pre-processing steps associated with it can drastically change ranks (Figure S1). This same result holds when comparing ranking methods for the HuProt$^{TM}$ arrays, though in this case, the top 3% of proteins corresponds to roughly 460 proteins, we only show how different ranking impacts the list of top 30 proteins (Figure S2). Given that the ranks $\hat{Q}_{p(i,j)}$ with $\gamma = 3$ minimize probability of misclassification of the top 3% of proteins these are the optimal ranks for identifying the top 30 antigens. However, even with these optimal ranks only 2 proteins in the
Table 4.1: Minimum posterior probability of antigens with high ranks on the malaria arrays. Number of proteins ranking ($\mathcal{Q}_{p(i,j)}$) in the 97th percentile or above, in the 95th percentile or above, or in the 90th percentile or above, that also have a high (0.75 and 0.5) posterior probability ($\tilde{\pi}_{p(i,j)}$) of being in the top 10, 5 and 3 percent of proteins.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>97th</th>
<th>95th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum $\tilde{\pi}_{p(i,j)}$</td>
<td>0.19</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>Number above 0.75 cutoff</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Number above 0.5 cutoff</td>
<td>2</td>
<td>4</td>
<td>21</td>
</tr>
</tbody>
</table>

malaria study rank in the top 3% with a posterior probability ($\tilde{\pi}_{p(i,j)}$) greater than 0.5, this number increases as we consider larger values of $\gamma$ (Figure S3 and Table 4.1). Similarly, in the HuProt™ arrays, only 194 proteins in the top 3% of proteins have a posterior probability ($\tilde{\pi}_{p(i,j)}$) greater than 0.5 of being classified as such (Figure S4 and Table 4.2). This finding further highlights the importance of reporting uncertainty throughout a bioinformatic pipeline; reporting uncertainty associated with classification allows researchers to potentially expand the number of proteins included in their analysis while balancing any conclusions with appropriate levels of uncertainty. Additionally, given that the analysis carried out in Kobayashi et al. (2019) relied on information from a different set of 30 antibodies than those identified as by $\mathcal{T}_{p(i,j)}$ and $\mathcal{Q}_{p(i,j)}$, repeating similar analysis techniques with these optimal ranks may reveal different trends and lead to different conclusions about this population’s reactivity to malaria antigens.
Table 4.2: Minimum posterior probability of proteins with high ranks on the 20 HuProtTM arrays spotted with serum from cancer-free individuals. Number of proteins ranking \( Q_{p(i,j)} \) in the 97th percentile or above, in the 95th percentile or above, or in the 90th percentile or above, that also have a high (0.75 and 0.5) posterior probability \( \tilde{\Pi}_{p(i,j)} \) of being in the top 10, 5 and 3 percent of proteins.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>97th</th>
<th>95th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum ( \tilde{\Pi}_{p(i,j)} )</td>
<td>0.23</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Number above 0.75 cutoff</td>
<td>73</td>
<td>111</td>
<td>213</td>
</tr>
<tr>
<td>Number above 0.5 cutoff</td>
<td>194</td>
<td>309</td>
<td>528</td>
</tr>
</tbody>
</table>

### 4.4.2 Patterns Across Geographic Location, RDT Status, and Age

Generally, the ranks \( \tilde{W}_{p(i,j)} \) reveal similar trends to previously published results about serological responses to malaria exposure, but identify different key antigens that may distinguish the serological response across different subsets of the population. Across regions of residence, samples from Nchelenge \( \tilde{\xi}_{\text{Nchelenge}, p(i,j)} \), have the highest number of antigens with a large posterior probability of being classified in the top 10\%, samples from Honde Valley have the second highest number \( \tilde{\xi}_{\text{Honde}, p(i,j)} \), and samples from Macha have the lowest number \( \tilde{\xi}_{\text{Macha}, p(i,j)} \) (Figure 4.1). This result is consistent with expected levels of reactivity to malaria antigens that accompany different levels of historical exposure to parasites; individuals in Nchelenge have the highest exposure and therefore are the most reactive, individuals in Honde Valley have higher levels of historical exposure due to recent resurgence but this exposure was over a short period, and those in Macha have the lowest level of exposure because of consistently declining transmission that is now nearing elimination. Some antigens that were not previously identified as
potentially differentiating across regions may warrant further study. In particular, two antigens that have high posterior probability of classification in the top 10% of antigens as well as high ranks in Honde Valley samples are not similarly positioned in Nchelenge samples, they are an asparagine-rich antigen and a WD domain G-beta repeat domain containing protein, neither of which were identified in the original analysis. Furthermore, in samples from Nchelenge, of the 123 antigens with a greater than 50% posterior probability of being classified in the top 10% of antigens, many were not previously studied and are not found in similar positions in samples from Honde Valley and Macha and may merit further investigation (Table S2). Additionally, comparing proteins that had extreme rank differences ($\tilde{D}_{\text{group1,group2},p(i,j)} < -250$, or $> 250$) across the three regions (Figure S5), reveals a group of 142 antigens that may be good targets for further research. In addition to the findings across locations, these ranks also reveal that samples from people over the age of 5 (\(\tilde{\xi}_{\text{adults},p(i,j)}\)) have a higher number of antigens with a large posterior probability of being classified in the top 10% than samples from people under the age of 5 (\(\tilde{\xi}_{\text{children},p(i,j)}\)) (Figure 4.2). This finding is consistent with the fact the older individuals have more exposure to parasites across all three regions. Further analysis of the antigens in the top 10% with greater than 50% posterior probability of classification in the 10% across these two groups reveals one antigen in the group of adults that was not previously studied that may be of further interest, a P vivax AMA1 Eoto monomer. This antigen also appears in the list of antigens with extreme rank differences ($\tilde{D}$) across children and adults, in addition to a circumsporozoite protein (CSP), merozoite surface protein 3, a sporozoite threonine and asparagine-rich protein, and QF122 antigen
Figure 4.1: $\tilde{\xi}_{p(i,j)}$ values averaged across subsets of arrays with samples from Honde Valley, Macha, and Nchelenge are on the ordinate and associated percentiles of ranks $(2Wp(i,j) - 1) / (2p)$ are on the abscissa. The red vertical line is at the 90th percentile.
Finally, individuals with a positive RDT result have more antibodies with a higher posterior probability of being classified in the top 10%, than those with a negative RDT result (Figure 4.3) which is explained by the fact that RDT positive individuals may have an ongoing infection, which is often associated with a heightened immune response and reactivity to malaria antigens. Among the 146 high ranking antigens with greater than 50% posterior probability of classification in the top 10% in the RDT positive samples, many were not
previously investigated and may have discriminatory power to differentiate individuals with and without ongoing infections (Table S4). Using the differences in protein ranks ($\tilde{D}$) across RDT positive and negative samples, five proteins suggest potential with discriminatory power across the two groups that were not previously investigated by (Kobayashi et al., 2019), they are subpellicular microtubule protein 1, EBA140, Rh1 PEP1, DNA topoisomerase II- putative, and exported protein 2.

**Figure 4.3:** $\tilde{\xi}_{p(i,j)}$ values averaged across subsets of arrays with samples from RDT positive and RDT negative individuals are on the ordinate and associated percentiles of ranks ($\frac{2 + W_{p(i,j)} - 1}{2 + p}$) are on the abscissa. The red vertical line is at the 90th percentile.
4.4.3 Signal Across Geographic Locations in Adults, and RDT Negative Individuals

We further explore differences across the three regions of transmission by looking at differences across the ranks of antigens in Honde Valley, Macha, and Nchelenge among adults, and we find that adults as well as children across the three locations show differences in reactivity to antigens. Among adults, samples from Nchelenge \((\bar{e}_{\text{adult, Nchelenge}}, p(i,j))\), have the highest number of antigens with a large posterior probability of being classified in the top 10%, samples from Honde Valley have the second highest number \((\bar{e}_{\text{adult, Honde}}, p(i,j))\), and samples from Macha have the lowest number \((\bar{e}_{\text{adult, Macha}}, p(i,j))\) (Figure 4.4). This trend is also visible among children (Figure S6 and Table 4.3). Importantly, given that adults have had longer term exposure to past changes in transmission, the 146 high ranking antigens with greater than 50% posterior probability of classification in the top 10%, many of which were not previously studied, may allow analysts to draw conclusions about historical trends in transmission, and suggest that there may be long-lasting antibodies to malaria exposure that reveal historical patterns of transmission that may not have been fully captured with previous surveillance efforts. The potential value of using serology as a surveillance tool is further reinforced in Figure 4.5, where RDT negative individuals, that is those who may have been exposed to parasites in the past but are likely not actively harboring and infection, reveal clear differences in reactivity to malaria antigens across three regions, with Nchelenge having the highest overall reactivity, Honde Valley, the second highest and Macha the lowest.
Table 4.3: Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\pi_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\pi_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed among adult and child samples across three locations.

<table>
<thead>
<tr>
<th>Region</th>
<th>Honde</th>
<th>Macha</th>
<th>Nchelenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum $\pi_{p(i,j)}$ for Children</td>
<td>0.15</td>
<td>0.18</td>
<td>0.39</td>
</tr>
<tr>
<td>Minimum $\pi_{p(i,j)}$ for Adults</td>
<td>0.27</td>
<td>0.27</td>
<td>0.55</td>
</tr>
<tr>
<td>Number above 0.75 cutoff for Children</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Number above 0.5 cutoff for Children</td>
<td>8</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>Number above 0.75 cutoff for Adults</td>
<td>6</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Number above 0.5 cutoff for Adults</td>
<td>28</td>
<td>11</td>
<td>146</td>
</tr>
</tbody>
</table>

Therefore it is possible that the 104 high ranking proteins with greater than 50% posterior probability of being classified in the top 10% among adults in Nchelenge that are not similarly positioned in Macha and Honde Valley, may reveal markers of past infection that would allow analysts to perform malaria surveillance without having to capture individuals actively harboring an infection (Table S5). Interestingly the trend across regions is not visible among RDT positive individuals (Figure S7 and Table S5) which may suggest that response to these malaria antigens is differentiating regions based on historical transmission trends, and not present levels of transmission, though differences in sample size of RDT positive and negative individuals as well as the high overall reactivity of RDT positive individuals may be responsible for these observations and more investigation is required.
Figure 4.4: $\pi_{p(ij)}$ values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among adults (older than 5) are on the ordinate and associated percentiles of ranks $\left(\frac{2\times Q_{p(ij)}}{2+P} - 1\right)$ are on the abscissa. The red vertical line is at the 90th percentile.
Figure 4.5: $\bar{p}_{p(i,j)}$ values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among RDT negative samples are on the ordinate and associated percentiles of ranks $\left(\frac{2 \times Q_{p(i,j)}}{2 + p} - 1\right)$ are on the abscissa. The red vertical line is at the 90th percentile.
4.5 Discussion and Future Work

Measurements obtained from any laboratory assay, pre-processing choices and analysis methods have the potential to impact downstream analysis. For some technologies that have been extensively studied and used for decades, bioinformatic pipelines can be well established and ubiquitous. However, for newer technologies, such as protein microarrays, different techniques are used to get data from their raw format to an analyzable format, and the impacts of these different techniques on conclusions drawn from measurements have yet to be fully understood. Here, we used a novel pipeline described and evaluated in Chapters 2 and 3, to provide estimates of reactivity to malaria specific antigens obtained from a protein microarray across individuals from three malaria endemic regions of Zambia and Zimbabwe. We compared the results of our analysis using full posterior distributions of signal from these arrays and a ranking method that pays attention to uncertainty by using information from the full posterior distributions to a previously published analysis (Kobayashi et al., 2019). We performed a similar comparison with the protein microarray data from a lung cancer study (Pan et al., 2017). Broadly, we found that our method produces a different list of high ranking antigens from the one published, and we provide levels of uncertainty associated with these ranks that can further inform analytic techniques and alter conclusions drawn from these data. We also find that while our method reveals similar overall trends of reactivity to malaria antigens across regions, age groups and RDT status, we identify new sets of proteins that may be able to differentiate the serological response of individuals to malaria antigens across these groups.
Previously, differences in reactivity to these proteins were only visible across regions in children under the age of 5, we show that differences in reactivity are present in both children and adults, and RDT negative individuals. These findings suggest that among the antigens studied here, there may be a subset that are markers of past infection, and therefore good candidates for malaria surveillance.

This analysis is a first step in further exploring these data using optimal ranks and associated levels of uncertainty. We identify associations and antigens that merit further exploration; additional analyses are required to fully understand what, if any, antigens are markers of past infection or good candidates for malaria surveillance. Specifically, we compare numbers of high ranking proteins that also have a high posterior probability of classification in the top 10% of antigens across subsets of the population, but we do not statistically assess differences across these groups. In order more formally explore these relationships, classification models or algorithms such as random forest should be leveraged. Additionally, the between-array variance of protein specific ranks can be calibrated by the protein-specific posterior variance to assess the degree of between-array variability.

There are additional limitations in the analysis we present, for instance, we do not perform sensitivity analysis to the values of $\gamma$, the cutoff of 50% as a high posterior probability of classification in the top $\gamma\%$, or to the choice of model (Chapter 3) to obtain full posterior distributions of signal. These sensitivity analyses are important to ascertain whether the associations we see are robust to these choices. Furthermore, when performing subgroup
analysis, we do not take into account differences in sample size across groups, these could have an impact on relationships we observe, particularly across RDT positive and negative individuals, where sample size differences are substantial. Finally, given the fact that immune response to infection is known to be a complex process, with many influencing variables, a more formal assessment of potential confounders, such as sex, is warranted.

Our analysis of response to malaria antigens across three malaria endemic regions of Zambia and Zimbabwe, age groups, and RDT result is preliminary, but still reveals the potential benefits of using a bioinformatic pipeline that characterizes and removes (to the degree possible) technical variation, while propagating uncertainty. While other pipelines may accomplish these goals for protein microarray data, this analysis illustrates the impacts of choices made in the pre-processing stages on downstream analysis, and shows that while there may not be a single correct way to analyze these data, using normalization techniques that allow analysts to use ranks that incorporate uncertainty can change the conclusions of a data analysis. This is especially true for protein microarrays, since they provide relative measures of protein levels and typically display high quantities of technical variation. Further studies are required to fully develop and evaluate optimal bioinformatic pipelines for protein microarrays, however this analysis demonstrates the potential impact of this type of research on scientific advancements in proteomics and immunology.
4.6 Supplement

Table S1: Demographic characteristics of the study sample from Kobayashi et al. (2019).

<table>
<thead>
<tr>
<th>Region</th>
<th>Honde Valley (N=190)</th>
<th>Macha (N=167)</th>
<th>Nchelenge (N=122)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDT Positive</td>
<td>5</td>
<td>35</td>
<td>43</td>
</tr>
<tr>
<td>RDT Negative</td>
<td>185</td>
<td>132</td>
<td>79</td>
</tr>
<tr>
<td>Under 5 (N=78)</td>
<td>39</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>Over 5 (N=398)</td>
<td>151</td>
<td>148</td>
<td>99</td>
</tr>
</tbody>
</table>

Table S2: Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\pi_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\pi_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across three locations with ($\pi_{\text{Honde, } p(i,j)}$, $\pi_{\text{Macha, } p(i,j)}$, and $\pi_{\text{Nchelenge, } p(i,j)}$).

<table>
<thead>
<tr>
<th>Region</th>
<th>Honde Valley</th>
<th>Macha</th>
<th>Nchelenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum $\pi_{p(i,j)}$</td>
<td>0.24</td>
<td>0.27</td>
<td>0.52</td>
</tr>
<tr>
<td>Number above 0.75 cutoff</td>
<td>2</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Number above 0.5 cutoff</td>
<td>22</td>
<td>7</td>
<td>113</td>
</tr>
</tbody>
</table>
Figure S1: Each row is an antigen on the original malaria array, each column is the rank of that antigen under different pre-processing conditions, using a different ranking technique. We include published ranks from Kobayashi et al. (2019), ranks of \(Y\) in Chapter 2, and \(\hat{P}_{p(i,j)}\) as well as \(\hat{Q}_{p(i,j)}\) ranks with \(\gamma = 3\).

Table S3: Minimum posterior probability of antigens with high ranks. Number of proteins ranking \((\pi_{\bar{p}(i,j)})\) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability \((\hat{\pi}_{p(i,j)})\) of being in the top 10 percent of proteins. All values are computed across two age categories with \(\pi_{\text{adults},p(i,j)}\) and \(\pi_{\text{children},p(i,j)}\).

<table>
<thead>
<tr>
<th>Age</th>
<th>5 and Under</th>
<th>Over 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum (\hat{\pi}_{p(i,j)})</td>
<td>0.21</td>
<td>0.33</td>
</tr>
<tr>
<td>Number above 0.75 cutoff</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number above 0.5 cutoff</td>
<td>6</td>
<td>37</td>
</tr>
</tbody>
</table>
Figure S2: Each row is a protein on the original HuProt\textsuperscript{TM} array, each column is the rank of that protein considering 20 arrays spotted with serum from cancer-free individuals under different pre-processing conditions, using a different ranking technique. We include ranks of $Y$ in Chapter 2, and $T_p(i,j)$ as well as $Q_p(i,j)$ ranks with $\gamma = 3$.

Table S4: Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\pi_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\pi_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across two RDT result categories with ($\pi_{pos,p(i,j)}$ and $\pi_{neg,p(i,j)}$).

<table>
<thead>
<tr>
<th>RDT Result</th>
<th>RDT Positive</th>
<th>RDT Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum $\pi_{p(i,j)}$</td>
<td>0.54</td>
<td>0.26</td>
</tr>
<tr>
<td>Number above 0.75 cutoff</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Number above 0.5 cutoff</td>
<td>146</td>
<td>16</td>
</tr>
</tbody>
</table>
Figure S3: $\tilde{p}_{p(i,j)}$ values across all malaria arrays are on the ordinate and associated percentiles of ranks \(\frac{2\tilde{Q}_{p(i,j)} - 1}{2P}\) are on the abscissa. The red vertical line is at the percentile determined by the value of $\gamma$ (90, 95, 97).

Table S5: Minimum posterior probability of antigens with high ranks. Number of proteins ranking ($\tilde{p}_{p(i,j)}$) in the 90th percentile or above that also have a high (0.75 and 0.5) posterior probability ($\tilde{p}_{p(i,j)}$) of being in the top 10 percent of proteins. All values are computed across two RDT result categories and three regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Honde</th>
<th>Macha</th>
<th>Nchelenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum $\tilde{p}_{p(i,j)}$ for RDT Pos</td>
<td>0.86</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>Minimum $\tilde{p}_{p(i,j)}$ for RDT Neg</td>
<td>0.23</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>Number above 0.75 cutoff for RDT Pos</td>
<td>154</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Number above 0.5 cutoff for RDT Pos</td>
<td>317</td>
<td>158</td>
<td>137</td>
</tr>
<tr>
<td>Number above 0.75 cutoff for RDT Neg</td>
<td>2</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Number above 0.5 cutoff for RDT Neg</td>
<td>21</td>
<td>3</td>
<td>104</td>
</tr>
</tbody>
</table>
Figure S4: $\tilde{\pi}_{p(i,j)}$ values across the 20 HuProt$^{TM}$ arrays spotted with serum form cancer-free individuals are on the ordinate and associated percentiles of ranks $\left(\frac{2q_{p(i,j)}-1}{2\gamma p}\right)$ are on the abscissa. The red vertical line is at the percentile determined by the value of $\gamma$ (90, 95, 97).
**Figure S5:** Boxplots showing the distribution of differences of antigen ranks $\tilde{T}_{\text{group},p(i,j)}$ across groups of samples. The x-axis label indicates the order of subtraction, in other words the X-axis label Group 1 vs. Group 2 indicates that the values on the Y axis are $\tilde{D}_{\text{group1,group2},p(i,j)} = \tilde{T}_{\text{group1},p(i,j)} - \tilde{T}_{\text{group2},p(i,j)}$. 
Figure S6: \( \pi_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among children (5 and under) are on the ordinate and associated percentiles of ranks \( \left( \frac{2Q_{p(i,j)} - 1}{2 + P} \right) \) are on the abscissa. The red vertical line is at the 90th percentile.
Figure S7: \( \pi_{p(i,j)} \) values averaged across subsets of arrays from Honde Valley, Macha, and Nchelenge among RDT positive samples are on the ordinate and associated percentiles of ranks \( \left( \frac{2 + Q_{p(i,j)}}{2 + P} \right) \) are on the abscissa. The red vertical line is at the 90th percentile.
References


138
Chapter 5

Conclusion

In this thesis we first develop a pre-processing pipeline that minimizes technical variation in protein microarray data. We use metrics of measurement agreement such as concordance correlation and Bland-Altman plots to assess the degree to which the pipeline corrects technical variation on two previously published data sets and one simulation study. We then use the outputs of this pipeline as inputs for a Bayesian model that uses repeated measurements of control probes, found on each protein microarray, to estimate an error structure and we subtract this from observed measurements to produce full posterior distributions of normalized signal. We fit this model to two different types of protein microarrays, and show that in both cases, the model fits well and produces estimates of signal that are useful for downstream inference. Finally, we use ranking methods that incorporate information from these full posterior distributions to elucidate differences in protein levels that are associated with biological features of interest. In particular, we apply these methods to a protein microarray that measures humoral immune response to malaria antigens, and show that these methods reveal additional information.
about potential use of some malaria antigens as markers of recent infection, and tools for malaria surveillance.

These three chapters together make a bioinformatic pipeline that takes raw measurements from protein microarrays, and transforms them in a way that removes technical variation, to a degree, while leaving biological signal relatively unperturbed, provides outputs that are useful for downstream analysis and that allow analysts to report appropriate levels of uncertainty, and enables nuanced biological inference. While we focus on its applications to two datasets, several components of our pipeline are flexible and could be applied to multiple other datasets and types of protein microarrays, with appropriate validation studies. More generally, this pipeline has the potential to unlock additional information from existing protein microarray studies that is not available with currently used pre-processing and analysis pipelines. Given that protein microarrays are often the first step in narrowing down targets of interest for further study, the ability to extract more information from protein microarrays could provide direction about where to efficiently direct efforts to answer research questions of interest. Particularly in the field of immunology and proteomics, where several advances have been made but many mechanisms and biological processes remain unknown.

This research is a starting point for developing statistical methodology that address the issues involved in analyzing protein microarray data. In particular, refinements that better address technical variation across arrays could substantially improve our ability to compare information about protein levels across samples, and eventually across studies. Additionally, developing
analysis techniques that make use of full posterior distributions of normalized signals can further enhance the level of information we can extract from these data. Finally, creating an R package would be an important step to making this analytical pipeline available to a broader group of researchers. As was the case with gene expression arrays, the advances made with protein microarray data will continue to expand as bioinformatic and analytic tools are further developed and improved.
Sophie Christine Bérubé
E3032 615 N Wolfe Street
Baltimore, MD 21205
Updated March 31, 2021

Education

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
PhD Candidate, Biostatistics 2016-2021 (expected)

Bowdoin College, Brunswick, ME
A.B., Major in Mathematics 2012 - 2016

Peer-Reviewed Publications


Other Publications


Submitted for Publication

1. **Percentile-Based Residuals for Model Assessment**  

2. **Measuring Genetic Relatedness Across Plasmodium falciparum Parasite Populations Accounting for Multiple Infections**  

3. **A Hierarchical Bayesian Model for Background Correction of Protein Microarrays**  

4. **A Hierarchical Bayes Model for Background Correction of Protein Microarrays**  

5. **Higher rank lamplighter groups are graph automatic**  
   President’s Science Symposium, Bowdoin College, September 2015. Berube, S., Palnitkar, T., Taback, J.

6. **Post-translational Regulation of the Ferric Uptake Regulator in Campylobacter jejuni**  
   Ottawa Institute of Systems Biology Conference, June 2013. Berube, S., Butcher, J., Stintzi, A.

---

### Teaching

1. **Johns Hopkins University**  
   *Fall 2017*: Krieger School of Arts and Sciences, Undergraduate Biostatistics, Section Leader  
   *Spring 2018*: Bloomberg School of Public Health, Stat Methods in Public Health III and IV, Teaching Assistant  
   *Fall 2018-Present*: Bloomberg School of Public Health, Stat Methods in Public Health I, II and III, Lead Teaching Assistant and Lab Instructor

2. **Phillips Exeter Academy**  
   *Summer 2016*: Differential Calculus  
   *Summer 2016*: Algebra II

3. **Bowdoin College Mathematics Department**  
   *Fall 2015*: Grader for Integral Calculus  
   *Spring 2015 to Spring 2017*: Teaching Assistant and Grader for Introduction to Mathematical Reasoning
Research Support

1. **Summer 2017-present**  
   Bloomberg School of Public Health  
   Southern Africa International Center of Excellence in Malaria Research. Methodological and applied research done in statistical analysis of protein microarray data, including the application of Bayesian models and MCMC algorithms.  
   *Supervised by Professor T.A. Louis*

2. **Summer 2015**  
   Bowdoin College Mathematics Department  
   Summer research fellow, research done in group theory on automatic and graph automatic groups, specifically the higher rank lamplighter groups and the Diestel-Leader group.  
   *Supervised by Professor J. Taback*

   Stintzi Lab, Department of Biochemistry, Microbiology and Immunology, University of Ottawa  
   Research done on post-translational protein regulation of proteins and the analysis of RNA-seq data relating to *Campylobacter jejuni*.  
   *Supervised by Dr. James Butcher*

4. **Summer 2011**  
   Department of Biochemistry, Microbiology and Immunology, University of Ottawa  
   Research done in basic genomics, learning fundamental lab techniques including gel electrophoresis, PCR and sequencing.  
   *Supervised by Dr. Natalia Bunimov*

Awards & Honours

2021 Helen Abbey Award for Excellence in Teaching

2020 Louis I. Dublin and Thomas D. Dublin Award for the Advancement of Epidemiology and Biostatistics

2015 Student Faculty Research Grant Fellowship, Bowdoin College supported by NSF Grant awarded to Professor Jennifer Taback

2012 Bowdoin College Faculty Scholarship

Computing Skills

1. **Languages:** R, some experience in python

2. **Markup:** LaTeX, knitr, Sweave
3. Other: Git and GitHub

Languages