THE UNIVERSITY OF CHICAGO

FUNCTIONAL GENOMICS APPROACHES TO UNDERSTANDING HUMAN COMPLEX TRAITS AND DISEASES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON HUMAN GENETICS

BY LAUREN ELIZABETH BLAKE

CHICAGO, ILLINOIS
DECEMBER 2019

Table of Contents

List of Figures

List of Tables [1]

1. Note: Due to the large size of some tables, the tables have been provided in a supplementary file accompanying the dissertation. In such cases, the page number provided below directs the reader to a table's caption.

ACKNOWLEDGMENTS

They say it takes a village to raise a child. In much the same way, it takes a village to get a PhD. I am extremely grateful for my "village" of supporters, including my mentors, collaborators, friends, and family.

I am very thankful to my advisor, Yoav Gilad. He has been extremely influential on me as a scientist, teaching me how to creatively solve challenging problems in human genetics with rigor and focus. His support along the path to graduation has been critical.

I am thankful to my committee members, Marcelo Nobrega, Matthew Stephens, and Oni Basu. I am also grateful for the opportunity to work with top eating disorder specialists, including Jennifer Wildes, Jessica Baker, and Cindy Bulik. This work would not have been possible without the Human Genetics community in Cummings Life Science Center. Additionally, I have enjoyed working with with Chris Porras, Dan Rice, Eric Friedlander, and Roy Morgan on the Computational STEM Lab (CSL) outreach program.

Throughout my PhD, I have been financially supported by the Genetics and Regulation Training Grant (T32GM007197) from the National Institutes of Health, the National Science Foundation (Graduate Research Fellowship and Innovation Corps), and the University of North Carolina Center of Excellence for Eating Disorders. John Novembre also supported the CSL. Sue Levison started as my "grants person" and evolved into a friend. Sue made sure that I didn't run away during my first week of graduate school, and has seen me all the way through graduation.

After about 2 weeks of my rotation in the Gilad lab, I knew that this was where I wanted to do my thesis research. After 5 years, I am still confident that I made the right decision. This feeling is in large part due to my fantastic labmates, past and present. In particular, thank you to Nick Banovich, Seb Pott, Genevieve Housman, Kenneth Barr, Benjamin Fair, Reem Elorbany, Katie Rhodes, Sidney Wang, Po-Young Tung, John Blischak, Briana Mittleman, Ittai Eres, Natalia Gonzales, Jonathan Burnett, Emily Davenport, and

Samantha Thomas. Michelle Ward has influenced my scientific thinking tremendously. I am also thankful to the technicians in the lab, particularly Amy Mitrano, Marsha Myrthil, and Claudia Chavarria, who also helped with my thesis work. I'd also like to thank my close collaborators, particularly Julien Roux and Bryan Pavlovic, without whom Chapters 2 and 3 in this thesis would not have been possible. Joyce Hsiao and Abhishek Sarkar also provided key statistical advice for my work.

Thank you to my friends Unjin Lee, Charlie Lang, Marissa Pelot, Hillary Childs, Sierra Smart, Sam Lohnes, Tatiana Birgisson, Janet S., and Aarti Venkat. Marcus Soliai is the best coffee buddy anyone could ask for. It is an honor to call Manny Vazquez my friend and get to work under his leadership. Finally, it would be impossible to list all the ways that Bryce van de Geijn supported me during my Phd.

My family has been incredibly supportive throughout my life. In particular, my grandparents, Nannie and Papa, encouraged me via weekly phone calls. Emily and Kate Blake are both amazing people, and I am lucky to call them my sisters. Thank you to my Mom and Dad, who have loved and encouraged me during my entire life.

ABSTRACT

A primary aim of the human genetics is to determine how genetic variation impacts phenotypic variation, including in complex traits and diseases. Understanding this relationship will ultimately allow the field understand the molecular basis of complex traits and better diagnose and treat human diseases. To dissect this relationship, human geneticists have leveraged comparisons between humans and other primates, as well as between different groups of humans. To maximize the utility of these functional genomics studies, proper study design must be deployed. Indeed, the primate comparative studies in Chapters 2 and 3 highlight study design challenges and potential solutions. Furthermore, this work demonstrates that adherence to key study design principles helps elucidate biological insight. In Chapter 4, I apply these solutions to a new problem, distinguishing individuals with eating disorders at high risk of rehospitalization from those with lower risk. In the final chapter, I discuss lessons learned and next steps for using functional genomics to study eating disorders.

CHAPTER 1

INTRODUCTION

1.1 Using genetics to understand complex traits and disease

The overarching goal of human genetics is to understand how genetic variation influences phenotypes– including complex traits and diseases– within our species. One approach to connect genotypic to phenotypic variation is to compare humans with non-human primates. This approach is particularly compelling given that human share over 99% of their DNA sequence with chimpanzees (REF). Efforts have been made to understand the 1%, that is to say the impact of these approximately 30 million single nucleotide polymorphisms (SNPs) differences (REF). Importantly, the majority of these SNPs are in the non-protein coding regions of the genome (REF). Overall this high degree of sequence similarity led Mary Claire King and Greg Wilson to hypothesize that it is the way these DNA sequences are regulated that leads to phenotypic differences (REF King and Wilson). Consequently, comparisons between humans and non-human primates can help to determine how these DNA sequences function and to reveal the underlying mechanisms that act on or as a result of sequence variation. This desire for for mechanistic understanding has inspired a number of studies to compare gene regulation levels (REF), epigenetic marks (REF), protein levels, and various other molecular phenotypes. These efforts have resulted in large comparative genomic catalogs of similarities and differences in gene regulation between humans and other primates. These catalogs have great potential to help us better understand the evolutionary processes that led to adaptations in humans [3, 4, 6, 13, 16-26], establish informed models of the relative importance of changes in different molecular mechanisms to regulatory evolution [27, 28], and identify molecular pathways that may be functionally important in the context of complex diseases (REF Gallego Romero review). However, the genetic variants that drive phenotypic diversity often have small effects, which are difficult to detect even in well-

powered studies (REF). It is therefore essential to design comparative studies that allow us to isolate the variables of interest while minimizing the effects of unwanted biological and technical differences. Yet, all too often, various aspects of study design are overlooked, to the detriment of the field. In extreme cases, poor study design leads to erroneous inference and incorrect interpretations (REF Mouse Encode; Gilad and Man 2015). The majority of the time, a poor study design limits the accuracy of the study and by extension, the biological insight that can be drawn from it (REF Pai et al 2011). Compounding the issue is that study design considerations are usually not explicitly discussed in comparative genomic papers.

The majority of my work has focused on primate comparative genomics - an area which presents many opportunities to discuss how an effective study design can affect the results of a study and aid in its interpretation. In this Introduction, I focus on two principles critical to identifying robust biological differences between species: minimizing confounders and careful sample collection. However, the principles of study design that I will discuss extend beyond primate comparative genomics to any study that makes comparisons between groups, including case-control studies in humans.

1.2 Study design challenges in comparative primate genomics

Confounders and other potential biases. As the technology used in genomic studies has progressed, so too has our understanding of the widespread nature and impact of confounders and other potential biases. Furthermore, using multiple species in functional genomic studies has increased the difficulty of minimizing these confounders. For example, early comparative studies using gene expression and DNA methylation microarrays often did not account for the attenuation of hybridization caused by sequence mismatches, which differ between species [6, 17, 18, 31]. More recently, common confounders of sequence-based comparative studies include individual sampling schemes that are unbalanced across species, and sample processing steps that are segregated by species [2, 10, 21, 32-34]. Systemic differences inher-

ent to the samples, such as differences in material quality between species [2, 35], also remain a concern. Similarly, a primate comparative framework brings forth analytical challenges. For example, analyses that do not use orthologous sequences or effective normalization procedures can result in bias [10, 21, 35]. Yet, most comparative genomic studies of humans and non-human primates that we are aware of, including previous studies from our own group [10, 21, 35], suffer from one or more of these weaknesses and caveats. Very few people would disagree that it is important to use good study design. However, the number of potential confounders are vast, including sex, date of death, age, RNA concentration, RIN score, RNA extraction date, library concentration, index sequence, sequencing pool, sequencing location, sequencing lane, total sequencing reads. Therefore, it can be difficult to detect confounding factors and bias introduced during sample processing or data analysis. Unfortunately, these factors are often neither accounted for nor discussed. Sometimes, this can lead to erroneous conclusions, which could have major implications for biological research. For example, a recent paper claimed that global gene expression levels were driven by species rather than tissue type (Yue et al. 2014). Upon re-analysis, it was uncovered that the human and mouse samples used in the study were sequenced in different batches (Gilad and Mizrahi-Man). With this study design, the biological variable of interest (species) is confounded with the technical variable of sequence batch. Therefore, it is unknown to what extent the technical variable drove the biological results reported in the original paper. Sorting out which results were driven by biological, rather than technical differences, is often led to the reader. Such a task can be quite challenging, as the comparative genomics field?and indeed, the larger genomics community?lacks consensus regarding meta-data collection and study documentation, particularly around sample and study design reporting.

Opportunistic study collection. Sample type, size, and collection techniques are critical study design considerations in comparative studies of primates. Until recently, flash frozen tissues were one of the only options for comparing primate biological material (REF Gal-

lego Romero). Because of the difficulty of obtaining samples (both for logistic and ethical reasons), we could typically sample only a small number of individuals from each species. Consequently, some primate comparative studies only have only a handful of individuals per species (REF Pai et al Tomas? papers, Browand et al). Particularly problematic is when there is only 1 individual per species, as individual and tissue are confounded (REF tomas? papers). Even when there are 3 or more individuals per species, tissues are often subject to high environmental variances because the donors are in an uncontrolled environment, and also flash frozen and shipped post mortem. The necessity of collecting samples opportunistically, together with small sample sizes, can lead to incomplete power to detect regulatory differences between species in any given study, and hence to relatively large apparent differences between studies. Furthermore, it was nearly impossible to obtain multiple tissue samples from the same individual. For example, to date, there have been no published comparative studies in primates that have analyzed multiple tissues sampled from the same individuals across multiple species in a balanced design [30]. Consequently, regulatory differences between tissues are always confounded with regulatory differences between individuals. In turn, relative measures of tissue-specific regulatory differences between species are confounded with inter-tissue differences in regulatory variation within species.

1.3 Study design challenges in psychiatric genetics

Confounders and other potential biases. Unfortunately, these study design challenges are not limited to the field of primate comparative genomics. For example, these are issues in psychiatric genetics and are highly prevalent in the nascent field of eating disorders genomics. Confounders and other potential biases. When performing eating disorder studies, there are many differences between groups, both in comparing individual cases or across cases and controls. Problematically, many biological variables are confounded with the biological variable of interest, disease state. For example, in case-control studies of individuals with

anorexia nervosa (AN), disease state is confounded with BMI. Even within cases, clinical variables can be confounded, including medication types, age of diagnosis, and number of rehospitalizations. In addition to the large number of biological variables, there are also technical considerations, such as RIN score. As discussed earlier, the larger genomics community lacks consensus regarding meta-data collection and study documentation. This area has also been discussed in the eating disorder field but have not been widely acted upon (REF Topher and Cindy?s methylation paper).

Opportunistic study collection. Similar to primates, tissues from humans are collected opportunistically. For many complex psychiatric trait, it is difficult to link genetic variation or gene regulatory differences, to phenotypes. While gene regulation in the brain is a logical place to start, it is difficult to access brains from living patients and a brain bank for eating disorders has only recently been established (REF). [Another point you could mention here is that the brain is extremely heterogeneous ? which regions are most relevant? It?s difficult to know.] Moreover, while recent research suggests tissues besides the brain may be relevant to disease state in AN (Watson et al 2019), it is hard to decide which tissue to study. Blood is relatively easy to collect from individuals with eating disorders, but the clinical utility of whole blood is unknown. Since individuals are collected opportunistically, collecting large sample sizes, particularly in longitudinal studies are difficult. Furthermore, individuals are almost always recruited during a state of illness, so it can be difficult to predict patient trajectory and therefore the outcomes represented in the cohort. This issue is compounded by the fact that many eating-disorder phenotypes, including treatment outcomes, are nebulous (Kahlsa et al 2017). Indeed, although the field of eating disorder genomics faces many of the same obstacles as primate comparative genomics and psychiatric genomics more generally, it also presents a unique set of challenges. .

1.4 Conclusion

Much of my thesis work has been devoted to addressing these challenges of potential biases and opportunistic study collection. in Chapters 2 and 3, I present examples of how to approach these challenges in gene regulatory studies across primates. Furthermore, I demonstrate that adherence to these key study design principles helps elucidate biological insight. In Chapter 4, I apply these lessons learned in a new context: comparing individuals with eating disorders at high risk of rehospitalization versus those with lower risk. In the final chapter, I discuss opportunities and challenges when using functional genomics to study eating disorders.

CHAPTER 2

A COMPARISON OF GENE EXPRESSION AND DNA METHYLATION PATTERNS ACROSS TISSUES AND SPECIES

2.1 Abstract[1]

Previously published comparative functional genomic data sets from primates using frozen tissue samples, including many data sets from our own group, were often collected and analyzed using non-optimal study designs and analysis approaches. In addition, when samples from multiple tissues were studied in a comparative framework, individual and tissue were confounded. We designed a multi-tissue comparative study of gene expression and DNA methylation in primates that minimizes confounding effects, by using a balanced design with respect to species, tissues, and individuals. We also developed a comparative analysis pipeline that minimizes biases due to sequence divergence. Thus, we present the most comprehensive catalog of similarities and differences in gene expression and methylation levels between livers, kidneys, hearts, and lungs, in humans, chimpanzees, and rhesus macaques. We estimate that overall, only between 7 to 11% (depending on the tissue) of inter-species differences in gene expression levels can be accounted for by corresponding differences in promoter DNA methylation. However, gene expression divergence in conserved tissue-specific genes can be explained by corresponding inter-species methylation changes more often. Finally, we show that genes whose tissue-specific regulatory patterns are consistent with the action of natural selection are highly connected in both gene regulatory and protein-protein interaction networks.

1. Citation for chapter: Blake LE*, Roux J*, Hernando-Herraez I, Banovich NE, Garcia-Perez R, Hsiao CJ, Eres I, Chavarria C, Marques-Bonet T, Gilad Y. A comparison of gene expression and DNA methylation patterns across tissues and species. bioRxiv. doi: https://doi.org/10.1101/487413. * denotes equal contribution.

2.2 Introduction

Gene regulatory differences between humans and other primates are hypothesized to underlie human-specific traits [97]. Over the past decade, dozens of comparative genomic studies focused on characterizing mRNA expression level differences between primates in a large number of tissues (e.g., [6, 14, 147, 173, 90]), typically focusing on differences between humans and other primates. A few studies have also characterized inter-primate differences in regulatory mechanisms and phenotypes other than gene expression levels, such as DNA methylation levels, chromatin modifications and accessibility, and protein expression levels [23, 66, 67, 68, 140, 175, 195, 199, 210]. These studies often construct catalogs of gene expression levels and other mechanisms. These catalogs have been useful to better understand the evolutionary processes that led to adaptations in humans [3, 13, 14, 22, 24, 173, 58, 87, 90, 91, 94, 107, 136, 140, 162] and ancestral or derived phenotypes that may be relevant to human diseases [117, 154]. One caveat that is shared among practically all comparative studies in primates is related to difficulty in obtaining multiple tissue samples from the same individual. To date, there have been no published comparative studies in primates that have analyzed multiple tissues sampled from the same individuals across multiple species in a balanced design [154]. As a result, regulatory differences between tissues are always confounded with regulatory differences between individuals [14, 147, 29, 140]. In turn, catalogs from these studies can not be used to compare tissue-specific regulatory differences between species to inter-tissue differences in regulatory variation within species (see Discussion in [140]). Our group and others often use previously published catalogs of comparative data in primates in our different studies. While we do not expect previously observed patterns to be erroneous, we are aware that data on gene-specific inter-species regulatory differences, and especially data that pertain to comparisons of divergence across tissues, may be inaccurate for the reasons we discussed above. We thus designed the current study to produce a new comprehensive catalog of comparative gene

expression and DNA methylation data from humans, chimpanzees, and rhesus macaques, attempting to minimize possible confounders. The goal of our study is not to challenge previous conclusions or document specific differences between the current and previous data. Instead, we aim to provide a new and more accurate comparative catalog of inter-tissue and inter-species differences in gene regulation between humans and other primates, with substantial sample and study design documentation. Overall, we believe that this catalog can be useful for many future applications and can serve as a new benchmark for regulatory divergence in primates.

2.3 Results

2.3.1 Study design and data collection

To comparatively study gene expression levels and DNA methylation patterns in primates, we collected primary heart, kidney, liver and lung tissue samples from four human, four chimpanzee, and four rhesus macaque individuals (Figure 1A, Supplemental Table S1A). From these 48 samples, we harvested RNA and DNA in parallel (Methods). After confirming that the RNA from all samples was of acceptable quality (Supplemental Table S1B; Supplemental Fig. S1A), we performed RNA-sequencing to obtain estimates of gene expression levels. Additional details about the donors, tissue samples, sample processing, and sequencing information can be found in the Methods and Supplemental Table S1. We estimated gene expression levels using an approach designed to prevent biases driven by sequence divergence across the species (similar to the approach of [11]). Briefly, we first mapped RNA-sequencing reads to each species' respective genome, and to compare gene expression levels across species, we only calculated the number of reads mapping to exons that can be classified as clear orthologs across all three species (Supplemental Table S1B). We excluded data from genes that were lowly expressed in over half of the samples as well as data from one human

heart sample that was an obvious outlier, probably due to a sample swap (Supplemental Fig. S2A-B). We normalized the distribution of gene expression levels to remove systematic expression differences between species (maximizing the number of genes with invariant expression levels across species corresponds to our null hypothesis; see Methods). Through this process, we obtained TMM- and cyclic loess-normalized log2 counts per million (CPM) values for 12,184 orthologous genes to be used in downstream analyses (Supplemental Table S2). Elements of study design, including sample processing, have previously been shown to impact gene expression data (Gilad and Mizrahi-Main 2015). Consequently, we tested the relationship between a large number of technical factors recorded throughout our experiments and the biological variables of interest in our study, namely tissue and species (Methods, Supplemental Materials, Supplemental Table S3A-B). We found that there were no technical confounders with tissue but two technical factors were confounded with species: time postmortem until collection and RNA extraction date (Supplemental Fig. 1B-1C). Due to the opportunistic nature of sample collection, these confounders are practically impossible to avoid in comparative studies in primates (especially apes). We discuss possible implications of these confounders throughout the paper.

2.3.2 Gene expression varies more across tissues than across species

We first examined broad patterns in the gene expression data. A principal component analysis (PCA) indicated that, as expected [7, 147, 125], the primary sources of gene expression variation are tissue (Figure 1B, regression of PC1 by tissue = 0.81; P ¡ 10-14; regression of PC2 by tissue = 0.70; P ¡ 10-10; Supplemental Tables S1A-B and S3A-B), followed by species (regression of PC2 by species = 0.27; P ¡ 10-3; Supplemental Tables S1A-B and S3A-B). This pattern is also supported by a clustering analysis based on the correlation matrix of pairwise gene expression estimates across samples (Supplemental Fig. S3). We then confirmed that, globally, gene expression levels across tissues from the same individual are more highly

correlated than gene expression levels across tissues from different individuals (Supplemental Fig. S2C). This observation supports the intuitive notion that collecting and analyzing multiple tissues from the same individual is highly desirable in functional genomics studies. We sought further explicit evidence that a study design incorporated balanced collection of multiple tissues from the same individuals is more effective. To do so, we used data from the GTEx Consortium (The GTEx Consortium 2017) for lung and heart. We first identified differentially expressed (DE) genes between lung and heart using all of the available GTEx data; we designated these classifications, which are based on hundreds of samples, as the "truth" (Supplemental Materials; Supplemental Table S3E). Next, we repeatedly identified DE genes between lung and heart using GTEx data from randomly chosen sets of just 4 samples from each tissue, and compared the results to DE genes identified from an equivalent analysis of sets of 4 samples from each tissue, in which the tissue samples originated from the same donor. Compared to the ?true classification? based on the entire GTEx dataset, DE analyses using data from samples of tissues that are matched for donors result in a higher ratio of true positives to false positives than analyses using samples from tissues that are unmatched for donors ($P = 0.03$; Supplemental Table S3F). Given the small number of false positives in both datasets, study design is unlikely to impact large-scale, highly robust trends across species. However, this study design choice is particularly important if one is interested in individual genes (as demonstrated by an example in Figure 1C).

2.3.3 Putatively functional tissue-specific gene expression patterns

To analyze the pairwise regulatory differences across tissues and species, we used the framework of a linear model (see Methods). We first identified (at FDR = 1%) 3,695 to 7,027 (depending on the comparison we considered) differences in gene expression levels between tissues, within each species (Table 1; Supplemental Table S4). Overall, the patterns of inter-tissue differences in gene expression levels are similar in the three species, significantly more

so than expected by chance alone (P ¡ 10-16, hypergeometric distribution; Supplemental Materials; Supplemental Table S5). A range of 17 to 26% of inter-tissue DE genes have conserved inter-tissue expression patterns in all three species (Supplemental Table S5). Regardless of species, we found the fewest inter-tissue DE genes when we considered the contrast between liver and kidney, and the largest number of DE genes between liver and either heart or lung (Table 1; Supplemental Table S4B). Unfortunately, since our data were produced from bulk RNA-sequencing, we were unable to determine the impact of cell composition on the number of inter-tissue DE genes. We used the same framework of linear modeling to identify gene expression differences between species, within each tissue (Supplemental Table S4A). Depending on the tissue and species we considered, we identified between 805 to 4,098 inter-species DE genes (at FDR = 1%; Table 1, Supplemental Table S4C). As expected given the known phylogeny of the three species, within each tissue, we classified far fewer DE genes between humans and chimpanzees than between either of these species and rhesus macaques (Supplemental Table S4B). It is a common notion that genes with tissue-specific expression patterns may underlie tissue-specific functions. Previous catalogs of such patterns in primates were always confounded by the effect of individual variation (because each tissue was sampled from a different individual). To classify tissue-specific genes using our data, we focused on genes that are either up-regulated or down-regulated in a single tissue relative to the other three tissues (within one or more species). We define such genes as having a "tissue-specific" expression pattern, acknowledging that this definition may only be relevant in the context of the four tissues we considered here. Using this approach and considering the human data across all tissue comparisons, we identified 5,284 genes with tissue-specific expression patterns (FDR 1%, Figure 2). By performing similar analyses using the chimpanzee and rhesus macaque data, we found that the degree of conservation of tissue-specific expression patterns is higher than expected by chance (P ¡ 10-16; Figure 2). This observation is robust with respect to the statistical cutoffs we used to classify tissue-specific expression

patterns (Supplemental Table S6), indicating that many of these conserved tissue-specific regulatory patterns are likely of functional significance. To broadly analyze the biological function of genes with conserved tissue-specific expression, we performed a Gene Ontology enrichment analysis (GO, see Supplemental Materials). We found these genes are indeed highly enriched with functional annotations that are relevant to the corresponding tissue (Supplemental Tables S7A-D, S8). For example, genes with conserved heart-specific expression patterns were enriched in GO categories related to muscle filament sliding (e.g. ACTA1, MYL2) and cardiac muscle contraction (e.g. MYBPC3, TNNI3).

2.3.4 Functional Analysis of Gene Regulatory Differences

We sought further evidence that the classification of genes with conserved tissue specific expression patterns is meaningful. To do so, we considered transcription co-expression networks [176, 209] based on GTEx data from heart and lung [143]. We found that genes with conserved tissue specific expression patterns are more likely to appear as nodes in the networks than genes without tissue-specific expression patterns, or genes whose tissue-specific expression patterns are not conserved (P ¡ 10-5). When we only considered genes that do appear as nodes in the network, we found that genes with conserved tissue specific expression patterns are more likely to be classified as hubs in the networks than genes without tissue-specific expression patterns, or genes whose tissue-specific expression patterns is not conserved (P ¡ 0.007). Motivated by these findings, we focused on gene expression patterns that are consistent with the action of natural selection (as described in [14]; see Supplemental Materials and Supplemental Table S7E). We found that genes whose expression patterns are consistent with the action of either stabilizing or directional selection (top 10%; Supplemental Table S7F) have more interactions with other genes in the network than genes whose expression patterns are not consistent with the action of natural selection (bottom 10%; P ¡ 0.05 for all comparisons; Figure 2E). This observation is fairly robust with respect to percentile cutoff

(Supplemental Table 7F). We repeated a similar analysis by using protein-protein interaction data from the Human Protein Atlas [115, 187, 193, 207] in all four tissues. We again found that genes whose expression patterns are consistent with selection have more annotated protein-protein interactions (P ¡ 0.05 in all 8 comparisons, Figure 2F; Supplemental Table 7G). These interaction results suggest that functionally important genes are carefully regulated. Furthermore, this tight regulation occurs at both the gene expression and protein levels in primates.

2.3.5 Variation in DNA methylation across tissues and species

As we mentioned above, we collected both RNA and DNA from each sample in our study. We used the DNA samples to study DNA methylation patterns through low-coverage whole genome bisulfite sequencing (BS-seq). The bisulfite conversion reaction efficiency was higher than 99.4% for all samples (Supplemental Table S1C). Following sequencing, we mapped the high-quality BS-seq reads to in silico bisulfite-converted genomes of the corresponding species and removed duplicate reads. We were able to measure methylation level in 12.5M to 22.9M CpG sites per sample, with a minimum coverage of two sequencing reads per site (Supplemental Table S1C). We estimated local methylation levels by smoothing the data across nearby CpG sites (see Supplemental Materials; Supplemental Figs. S4-S6; [122]). To facilitate a comparison of methylation levels across species, we annotated 10.5M orthologous CpGs in the human and chimpanzee genomes, as well as a smaller set of 2.4M orthologous CpGs in all three primate genomes (Supplemental Table S1C-E). To identify differences in methylation levels between tissues and species we again employed a linear model framework (Methods). Focusing on methylation patterns across tissues within species, we identified between 7,026 to 41,280 differentially methylated regions between tissues, within species (T-DMRs), depending on the pairwise tissue comparisons we considered (Table 2; Supplemental Table S9A; [21]). Pairwise comparisons between hearts and lungs showed the lowest

number of DMRs, regardless of species (7,026 in rhesus macaques, 8,524 in chimpanzees, 14,208 in humans), while comparisons involving heart and liver showed the largest number of DMRs (22,561 in humans, 28,767 in chimpanzee and 41,280 in rhesus macaques; Table 2). We found that human T-DMRs overlapped genic and regulatory features significantly more than expected by chance. In particular, there is an enrichment of T-DMRs in intergenic regions, introns, 5'UTRs, 3'UTRs, and active enhancers (as defined by [133]; P ¡ 0.04 for all tests; Supplemental Table S9B). We found strong evidence for T-DMR conservation across all three species (P ¡ 10-16 across all comparisons; Supplemental Table S10A). Though this level of conservation is higher than expected by chance, we recognize that in each tissue comparison we performed, we had incomplete power to identify T-DMRs and so the true conservation of T-DMR is expected to be even higher. To sidestep this challenge and compare T-DMRs across species more effectively, we considered methylation data from all T-DMR orthologous regions that were classified as such in at least one species. When we performed hierarchical clustering using orthologous DNA methylation data from these T-DMRs, the data clustered first by tissue than by species (Supplemental Fig. S7). This trend is robust with respect to the species used to initially locate T-DMRs (Supplemental Fig. S8-S9). Thus, our results suggest that in general, inter-tissue methylation differences within a species tend to be conserved, consistent with the observations of previous studies [67, 68, 123, 129, 140]. We next focused specifically on tissue-specific DMRs, as these may contribute to tissue-specific function. In contrast to differences in methylation between any pair of tissues, a tissue-specific DMR is defined as having a similar methylation level in three of the tissues we considered, but a significantly different methylation level in the remaining tissue. We found that there were more DMRs specific to liver (3,278 to 11,433 DMRs depending on the species) than to kidney (2,300 to 3,957 DMRs), heart (1,597 to 2,969 DMRs), or lung (453 to 5,018 DMRs, Figure 3; Supplemental Table S10B). Tissue-specific DMRs are highly conserved regardless of the comparisons we made (P ¡ 10-13 for all com-

parisons, at least 25% bp overlap was required to be considered shared). In all four tissues, over 59% of conserved DMRs are hypo-methylated in a tissue-specific manner. We evaluated the overlap between tissue-specific DMRs and genomic regions marked with H3K27ac, a mark often associated with active gene expression (The ENCODE Project 2012). We found that conserved hypo-methylated tissue-specific DMRs were annotated with H3K27ac more frequently than tissue-specific DMRs identified only in humans (P ¡ 0.001, difference of proportions test; Supplemental Table S10C; Supplemental Materials). We then asked about the potential impact of these conserved hypo-methylated tissue-specific DMRs on the expression of nearby genes. We found that genes with the closest TSSs to conserved tissue-specific DMRs are highly enriched with relevant functional annotations in hearts and livers (the tissues with the largest numbers of conserved hypo-methylated tissue-specific DMRs; Figure 3E-F, Supplemental Table S10D) [180]. For example, conserved heart-specific DMRs are closest to genes in cardiovascular-related pathways, including ventricular cardiac muscle cell development, canonical Wnt signaling pathway, and ERK5 cascade. Overall, these observations suggest that conserved tissue-specific DMRs are likely to underlie tissue-specific gene regulation in primates.

2.3.6 Inter-species differences in gene expression and DNA methylation levels

Our comparative catalog can be used to identify DNA methylation differences that could potentially explain gene expression differences across species and tissues. To do so, we first identified the 7,725 orthologous genes with expression data and corresponding promoter DNA methylation data in humans and chimpanzees, and the 4,155 orthologous genes with the same information for all three species. We then determined to what extent divergence in DNA methylation levels could potentially underlie interspecies differences in gene expression by comparing the gene expression effect size associated with ?species? before and after

accounting for methylation levels. To determine significant effect size differences, we applied adaptive shrinkage [155], a flexible Empirical Bayes approach for estimating false discovery rate (Methods). We note that this mediation approach does not consider the possibility that a third, unobserved event, may be causally responsible for both the methylation and expression patterns. Considering differentially expressed genes between humans and chimpanzees (in at least one tissue), we found that between 11% and 25% of genes (depending on tissue) showed a difference in the effect of species on gene expression levels once average promoter methylation levels were accounted for (significant difference in effect size classified at FSR 5% and are represented by red in Supplemental Fig. 10; Supplemental Table S11A; Supplemental Fig. S10). As a control analysis, we considered only the genes that were not originally classified as differentially expressed between humans and chimpanzees, and found that the difference in the effect size of species on gene expression levels was reduced in less than 1% of genes once methylation data were accounted for (FSR 5%, Supplemental Fig. 10; Supplemental Table S11A); thus, our approach is well calibrated. We applied the same approach to the human and rhesus macaque data, and found that the percentage of genes for which gene expression differences could potentially be explained by methylation differences ranges from 21% in the lung to 40% in the liver (Supplemental Fig. 11; Supplemental Table S11B). This observation may reflect the more extreme gene expression differences between humans and rhesus macaques than between humans and chimpanzees (prior to accounting for DNA methylation levels, P ¡ 0.003 in all tissues, t-test comparing the absolute values of the effect sizes for both groups of DE genes;). Next, we examined the genes in which DNA methylation differences may underlie inter-tissue gene expression differences (example in Figure 4A-C). Using adaptive shrinkage, we found that 7-25% of inter-tissue gene expression differences could potentially be explained by DNA methylation differences across tissues (FSR 5%; Supplemental Table S11C-E). When we performed the control analysis and considered only data from genes that were not differentially expressed between tissues,

less than 1% of effect sizes differed once we accounted for the methylation data (Figure 4F; Supplemental Table S11C-E). Finally, we focused on regulatory patterns that are most likely to be functional; namely, conserved inter-tissue gene regulatory differences. These differences were more likely to be explained by variation in methylation levels than non-conserved inter-tissue gene expression differences (minimum difference is 7%, P ¡ 0.005 for all comparisons; at FDR ¡ 5% and FSR ¡ 5%; Figure 4E-4F; Supplemental Table S11C-E). This observation is robust with respect to the FDR and FSR cutoff used (Supplemental Table S11C-E). Indeed, the correlation between methylation and gene expression data is higher for genes with conserved inter-tissue expression patterns compared to genes whose expression patterns were not conserved (Figures 1E-1F). One way to maintain conserved inter-tissue expression differences could be through DNA methylation level differences. We compared the genes whose variation in inter-tissue gene expression can potentially be explained by variation in DNA methylation levels (assuming no independent effect on an unobserved factor) to all genes with conserved inter-tissue expression differences (13-19% of genes across all pairwise tissue comparisons, Supplemental Table S11C). We found that these genes are enriched for ?essential tissue functions? (Supplemental Table S11F). For example, the heart genes are enriched for cardiac and smooth muscle contraction, whereas those in liver are enriched for regulation of cholesterol transport and hormone secretion (Figure 4G, Supplemental Table S11F). These observations suggest that DNA methylation levels may mark or even drive differences in the expression levels in functionally relevant genes.

2.4 Discussion

We designed a comparative study of gene regulation in humans, chimpanzees, and rhesus macaques that minimized confounding effects and bias. Consistent with previous studies, we found a high degree of conservation in gene expression levels when we considered the same tissue across species [8, 147, 57, 113, 161]. We also found evidence for conservation of tissue-

specific DMRs. Our observations are qualitatively consistent with those of previous studies that mostly used microarrays to measure methylation levels [67, 140, 191], however, the high resolution of our sequence-based DMR data allowed us to examine a much larger number of CpG sites. Thus, we were able to show that while DNA methylation can potentially explain a modest proportion of expression differences between tissues [140], it is more likely to play a role in underlying conserved tissue-specific gene expression levels. We created and made available the most comprehensive, and likely most accurate comparative catalog of gene expression and methylation levels in humans, chimpanzees, and rhesus macaques. Comparative functional genomic studies in primates, including from our own lab, often are not designed to test for specific hypotheses. Rather, many of these comparative genome-scale studies aim to build catalogs of similarities and differences in gene regulation between humans and other primates. These catalogs have been shown to be quite useful; for example, they can be used to identify inter-species regulatory changes that have likely evolved under natural selection [3, 13, 14, 22, 24, 173, 58, 87, 90, 91, 94, 107, 136, 140, 162], and thereby help us better understand the evolutionary processes that led to adaptations in humans. These catalogs are also used to establish informed models of the relative importance of changes in different molecular mechanisms to regulatory evolution [92, 200], and to inform us about ancestral or derived phenotypes that may be relevant to human diseases [117, 57]. Ultimately, comparative catalogs of gene regulatory phenotypes are used to develop and test specific hypotheses regarding the connection between inter-species regulatory changes and physiological, anatomical, and cognitive phenotypic difference between species. In this study, we used a comparative catalog to identify species-specific and, in particular, tissue specific regulatory patterns, as these genes are often drug targets [38] and are likely important for the evolution of human traits [14]. We showed that genes with conserved tissue-specific regulatory patterns have more regulatory interactions and protein-protein interactions than genes whose regulatory patterns are not conserved or are not tissue-specific. These patterns

became even more pronounced when we focused on genes whose expression patterns are consistent with the action of natural selection. Put together, these observations consistently support the inference that when genes perform an important function that needs to be carefully regulated, evolution can act on multiple levels of the regulatory cascade in primates. Focusing on species-specific patterns of tissue-specific gene regulation, our observations can help formulate specific functional hypotheses regarding human-specific adaptations. For example, genes with tissue-specific gene regulation identified in humans only are enriched for GO pathways that may contribute to human-specific features, including the sodium ion import across plasma membrane in kidneys (e.g. SLC9A3 and TRPM4), the glycogen biosynthetic process in livers (e.g. PGM1 and AKT1), and paraxial mesoderm morphogenesis in lungs (e.g. MST1R and MAP9).

2.4.1 Consideration of study design and record keeping

Regardless of the model system used and the types of data that are collected, study design considerations are always critical. Perhaps because comparative studies in primates typically rely on opportunistic sample collection, there are no recognized study design standards that are kept and consistently reported in most existing studies (including many earlier studies from our own group). We thus believe that it is worthwhile to explicitly discuss a few important considerations regarding study design and the recording of meta-data. Without a balanced study design, it would have been impossible to independently estimate the effects of individual, tissue, and species on our data. Because the sources of confounding factors are difficult to predict in advance, we strongly recommend that samples are collected using a balanced design with respect to as many parameters as possible. These include the distribution of tissue samples per individual, the number of individuals from each species, sex, age range, cause of death and collection time (in the case of post-mortem tissues), or sample collection and cell culturing (in the case of iPSC-based models). All steps of sample

processing (RNA extraction, library preparation, etc.) should be done in batches that are randomized or balanced with respect to species, tissue, and any other variables of interest. Most importantly, all sample processing steps should be recorded in a sample history file that includes anything that happened to any sample. We have documented many of these steps in Supplemental Tables S1A-E. This documentation can help provide evidence that a phenomenon is driven by biological rather than technical factors. It may also benefit future studies by facilitating effective meta-analysis of multiple data sets, which would help to address the problems of tissue availability and small sample sizes. We believe that, moving forward, it should be a requirement that these meta-data are available with every published comparative genomic data set.

2.5 Methods

2.5.1 Sample Description

We collected heart, kidney (cortex), liver and lung tissues from four individual donors in human (Homo sapiens, all of reported Caucasian ethnicity), chimpanzee (Pan troglodytes), and Indian rhesus macaque (Macaca mulatta), for a total of 48 samples (3 species * 4 tissues * 4 individuals; Figure 1A). The choice of these particular tissues was guided by their relative homogeneity with respect to cellular composition (e.g. [**?**]), which do not change substantially across primate species. In contrast, other tissues, such as brain subparts, differ substantially in cellular composition across primates [**?**], which could potentially confound the analyses. Human samples were obtained from the National Disease Research Interchange (IRB protocol 14378B). Non-human samples were obtained from several sources, including the Yerkes primate center and the Southwest Foundation for Biomedical Research, under IACUC protocol 71619. When possible, samples were collected from adult individuals whose cause of death was unrelated to the tissues studied.

2.5.2 RNA library preparation and sequencing

In total, we prepared 48 unstranded RNA-sequencing libraries as previously described [35, 185]. Twenty-four barcoded adapters were used to multiplex different samples on two pools of libraries. RNA-sequencing libraries were sequenced on 26 lanes on 4 different flow-cells on an Illumina HiSeq 2500 sequencer in either the Gilad lab or at the University of Chicago Genomics Facility (50bp single end reads, Supplemental Table S1; Supplemental Materials).

2.5.3 Quantifying the number of RNA-seq reads from orthologous genes

We used FastQC (version 0.10.0; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) to generate read quality report and TrimGalore (version 0.2.8), a wrapper based on cutadapt (version 1.2.1)[18], to trim adaptor sequences from RNA-seq reads. We trimmed using a stringency of 3, and to cut the low-quality ends of reads, using a quality threshold (Phred score) of 20. Reads shorter than 20 nt after trimming were eliminated before mapping (Supplemental Table S1). For each sample, we used TopHat2 (version 2.0.8b) [95] to map the reads to the correct species? genome: human reads to the hg19 genome, chimpanzee reads to the panTro3 genome, and the rhesus macaque reads to the rheMac2 genome (Supplemental Materials). Expression level estimates may be biased across the species due to factors such as mRNA transcript size and different genome annotation qualities. To circumvent these issues, we only retained reads that mapped to a set of 30,030 Ensembl gene orthologous metaexons available for each of the 3 genomes, as described and used previously [13**? ?**]. We defined the number of reads mapped to orthologous genes as the sum of the reads mapping to the orthologous metaexons of each gene. We quantified gene expression levels using the program coverageBed from the BEDtools suite and then performed TMM and cyclic loess normalization (Supplemental Materials). The reason we are using hg19 is that this is still, by far, the dominant genome build in the community. In particular, all 3 releases of GTEx use hg19 and only a fraction of the 1000 genome data are available in GRCh38

coordinates. Second, to demonstrate that the results we report would not change much if we used the GRCh38 build, we leveraged the fact that differential expression analysis compares gene expression levels from groups of samples (e.g. human liver samples to human lung samples). Therefore, we compared the ranks of the normalized gene expression levels in the 15 human samples mapped using hg19 to the same samples mapped to GRCh38. The correlations of these ranks were extremely high (median Pearson?s correlation = 0.96). These strong correlations suggest that our general conclusions?and indeed, many genes we identified as DE?would remain if we had used the GRCh38 build.

2.5.4 Analysis of Technical Variables

To assess whether the study?s biological variables of interest?tissue and species– were confounded with the study?s recorded sample and technical variables, we used an approach described in [11]. For the 12 RNA-seq related technical variables that were the most highly correlated with tissue or species, we assessed which technical variables constitute the ?best set? of independent variables to be included in a linear model for gene expression levels. Because of the partial correlations between the variables, we applied lasso regression using the package glmnet [54]. Before performing the analysis, we also protected our variables of interest, tissue and species, in the model for each gene. We summarized each technical variable?s influence across the genes by counting the number of times each technical variable was included in the ?best set? of the gene models. We found that none of the technical variables appeared in more than 25% of the best sets (i.e. more than 25% of the gene models). Therefore, we chose not to include these technical variables in our model for testing differential expression. Finally, during our analysis of technical factors, we discovered that RNA extraction date was confounded with species. In 2012, we extracted RNA from the chimpanzee samples on March 8, from the human samples on three days between March 12-29, and from the rhesus samples on March 6. To test the relationship between the date of

RNA extraction and gene expression PCs in humans, we performed individual linear models on PCs 1-5 using RNA extraction date as a predictor. None of the models were statistically significant at FDR 10%, suggesting that tissue type is more highly associated with gene expression levels than RNA extraction date.

2.5.5 Differential expression analysis using a linear model-based framework

To perform differential expression analysis, we used the same approach as in [11]. We applied a linear model-based empirical Bayes method [166**?**] that accounts for the mean-variance relationship of the RNA-seq read counts, using weights specific to both genes and samples [103]. To be considered a "tissue-specific DE gene" under our stringent definition, the gene must be in the same direction and statistically significant in all pairwise comparisons including the given tissue but not significant in any comparison without that tissue. For example, for a gene to be classified as having heart-specific upregulation in a given species, the gene needed to be upregulated (a significant, positive effect size) in heart versus liver, heart versus lung, heart versus kidney, but not significantly different between the liver versus lung, liver versus kidney, or kidney versus lung, in the same species. Under the more lenient definition of tissue-specific DE genes, we compared the gene expression level of one tissue to the mean of the other three tissues. To do so, we grouped the three tissues together and again used the limma+voom framework to identify significant differences in one tissue versus the group of the other tissues. To identify inter-species differences in gene expression patterns across tissues within species (tissue-by-species interactions), we used the limma+voom framework and looked for the significance of tissue-by-group interactions. In one analysis, the groups were Great Ape versus Rhesus macaque and in another analysis, the groups were Human versus non-human primates. To minimize the number of interactions, we compared one tissue relative to a group of the other 3 tissues (e.g. Great Ape versus rhesus macaque heart versus non-heart). Significant tissue-specific interactions were detected using the adaptive

shrinkage method, ashr [155]. Specifically, for each test, we input the regression estimates from limma to ashr: regression coefficients, posterior standard errors, and posterior degrees of freedom. We used the default settings in ashr to calculate the shrunken regression coefficients (called the ?Posterior Mean? in ashr), false discovery rate (FDR or also known as q-value), and false sign rate (FSR or also known as s-value: the probability that sign of the estimated effect size is wrong in either direction). We assigned directionality based on the sign of the posterior mean and determined significance based on the false sign rate.

2.5.6 The impact of matched tissue samples on DE results

To determine the impact of matched tissue samples on DE results, we compared intertissue DE analysis results when using tissues from the same or different individuals in GTEx v7 data (The GTEx Consortium 2017). We first subset the GTEx raw gene expression counts data to only individuals for which there was gene count information in the heart and lung tissues, for genes included in all 3 tissues. (There were the most GTEx samples in heart and lung; we decided to focus on these samples). Furthermore, to minimize the number of confounders needed in the linear model, we decided to only analyze individuals of the same sex and whose samples were sequenced on the same platform (sex = 1 and platform = 1 from the GTEx documentation). We then normalized the data and performed DE analysis using a voom+limma pipeline. In the linear model, we included tissue and 3 GTEx-provided covariates (covariates 1 and 2 and inferred covariate 1 from the covariate file for each tissue) as fixed effects and individual as a random effect. We chose to renormalize the raw counts data rather than use the normalized counts from GTEx because the voom+limma pipeline requires raw counts to assign voom weights. We considered the output of the intertissue DE analysis for all individuals (DE versus non-DE genes at FDR 5%) as the ?ground truth?. To evaluate the impact of our study design, we then subset the gene expression information to the individuals for which there is information in all 3 tissues. We obtained gene expression

level information from 4 randomly selected individuals and used the voom+limma pipeline to identify intertissue DE genes. Next, we compared the list of DE genes from this analysis to the "ground truth" list. We performed this downsampling procedure for tissues from the same 4 individuals as well as different 4 individuals 10 times each and compared the number of true and false positives from the tests. For the analysis with 8 different individuals, there were no repeated individuals, so we did not use the ?duplicateCorrelation? function in voom.

2.5.7 BS-seq library preparation, sequencing, and mapping

We prepared a total of 48 whole-genome BS-seq libraries from extracted DNA as previously described [5, 186]. We aligned the trimmed reads to the human (hg19, February 2009), chimpanzee (panTro3, October 2010), or rhesus macaque (rheMac2, January 2006) genomes, and to the lambda phage genome using the Bismark aligner (version 0.8.1)[196]. We estimated the percentage of methylation at an individual cytosine site by the ratio of the number of cytosines (unconverted) found in mapped reads at this position, to the total number of reads covering this position (sequenced as cytosine or thymine, i.e., converted or unconverted) using the methylation extractor tool from Bismark (version 0.8.1). We additionally collapsed information from both DNA strands (because CpG methylation status is highly symmetrical on opposite strands [53]) to achieve better precision in methylation estimates across the genome. To obtain CpGs that mapped to multiple species, the chimpanzee and rhesus macaque CpG sites were mapped to human coordinates (hg19) using chain files from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/ and the liftOver tool from the UCSC Genome Browser [**?**]. These files had previously been filtered for paralogous regions and repeats, but we also removed positions that were not remapped to their original position when we mapped from human back to their original genome. Chimpanzee and rhesus macaque CpG sites were mapped to human, even if their orthologous positions were not a CpG site in human.

2.5.8 Identifying differentially methylated regions (DMRs)

We were interested in identifying regions exhibiting consistent differences between pairs of tissues or pairs of species, taking biological variation into account. To identify DMRs we used the linear model-based framework in the Bioconductor package bsseq (version 0.10.0)[122]. For a given pairwise comparison (e.g., human liver vs. human heart), the bsseq package produces a signal-to-noise statistic for each CpG site similar to a t-test statistic, assuming that methylation levels in each condition have equal variance. As recommended by the authors of the package, we used a low-frequency mean correction to improve the marginal distribution of the t-statistics. Similar to previous studies using this methodology, a t-statistics cutoff of ?4.6,4.6 was used for significance, [**?** **?**]. DMRs were defined as regions containing at least three consecutive significant CpGs, an average methylation difference of 10% between conditions, and at least one CpG every 300 bp [122]. We used BEDTools (version 2.26.0) [148] to calculate the number of overlapping DMRs across tissues and/or species [21]. We required overlapping DMRs to have a minimum base pair overlap of at least 25%, unless otherwise stated. To be considered a tissue-specific DMR, the region was required to be a significant tDMR in 1 tissue compared to the other 3 tissues pairwise (in the same direction) but not a significant tDMR across any of the other 3 tissues in pairwise comparisons. We again used BEDTools to ensure a minimum base pair overlap of at least 25%. Once the tissue-specific DMRs were identified within a species, we then classified them as species-specific or conserved. To be considered conserved (across humans and chimpanzees or all three species), the tissue-specific DMR had to be significant in all species in the comparison and have a minimum base pair overlap of the tDMR of at least 25%.

2.5.9 Calculating the average methylation levels of conserved promoters

To calculate the DNA methylation levels of orthologous CpGs around the transcription start site (TSS) of orthologous genes, we first had to determine the orthologous TSSs. We began with the 12,184 orthologous genes in our RNA-seq analysis. Of these, we found that 11,131 of these orthologous genes had an hg19 RefSeq TSS annotation. We used liftOver to find orthologous sites in the chimpanzees and rhesus macaque genomes in 9,682 of those 11,131 genes. We then determined which of the hg19 RefSeq TSS annotations were closest to the first hg19 orthologous exon, and repeated this process with the other two species and their respective genomes. We found that 9,604/9,682 of the closest TSS annotations in humans at the same liftOver coordinates in the other two species. We then calculated the distance between the first orthologous exon to the TSS site in all 3 species individually. To minimize this difference between the 3 species, we filtered all genes with a maximum distance difference across the species of larger than 2,500 bp. (For reference, the 75th percentile of the maximum difference in distance was 2,078 bp.) 7,263 autosomal genes remained after this filtering step. 4,155 genes had at least 2 orthologous CpGs 250 bp upstream and 250 bp downstream of the orthologous TSS. We chose a 250 bp window around the TSS based on DNA methylation levels around the promoter in [5] and calculated the average of orthologous CpGs within this window for the 4,155 genes. Using the same method but in humans and chimpanzees only, we found and calculated the average of orthologous CpGs within this window for 7,725 genes.

2.5.10 Joint analysis of promoter DNA methylation and gene expression levels

To determine whether DNA methylation may underlie interspecies differences in gene expression levels, we used a joint analysis method as described below. For each gene, we analyzed the gene expression levels, along with the accompanying average methylation level

250 bp upstream and downstream of the TSS (found above). For a given tissue, we first determined the effect of species on gene expression levels using a linear model, with species and RIN score as fixed effects (Model 1). Next, we parameterized a linear model attempting to predict expression levels exclusively from methylation levels. We refer to these residuals as "methylation-corrected" gene expression values. We then used these values to again determine the effect of species, this time on gene expression levels ?corrected? for methylation, using a linear model with species and RIN score as fixed effects (Model 2). To determine the contribution of DNA methylation levels to inter-species differences in gene expression, we computed the difference in the species effect size between Model 1 and Model 2 for each gene, as well as the standard error of the difference. Large effect size differences between Models 1 and 2 for a given gene suggest that methylation status may be a significant driver of DE. To assess the significance of this difference, we used adaptive shrinkage (ashr) [155] to compute the posterior mean of the differences in the effect sizes, using vashr, with the degrees of freedom equal to the number of samples in the linear model minus 2. The shrunken variances from vashr were used in the ashr posterior mean computation. From this procedure, we obtained the number of genes where species has a significant difference in effect sizes before and after regressing out methylation. We assessed significance using the s-value statistic (false sign rate, FSR [155]). Using the s-values, rather than the q-values, not only takes significance into account but also has the added benefit of assessing our confidence in the direction of the effect. We performed the above analysis separately for inter-species DE genes and non-DE genes, and in each tissue individually. We identified inter-species DE genes in our tissue of interest as those with a significant species term in the model of species and RIN score as fixed effects. We assessed significance of DE genes at FDR 5%, unless otherwise noted. We also applied the same analysis framework to determine whether DNA methylation may underlie inter-tissue differences in gene expression levels. For the inter-tissue DE genes and non-DE genes, we replaced ?species? with ?tissue? as a fixed

effect in models 1 and 2. We assessed significance with various FDR and FSR thresholds, as specified in the text.

2.5.11 Data and code availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI?s Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) using GEO Series accession number GSE112356. Data and scripts used in this paper are available at https://github.com/Lauren-Blake/Reg_Evo_Primates. The results of our scripts can be viewed at https://lauren-blake.github.io/Regulatory_Evol/analysis/.

2.6 Acknowledgments

Members of the Gilad, Marques-Bonet, Robinson-Rechavi, Stephens, and Pritchard labs provided helpful discussions and comments on the manuscript. In particular, Matthew Stephens provided guidance on integrating the DNA methylation and gene expression data, and Michelle Ward and Natalia Gonzales provided helpful comments on a draft of the manuscript. We thank Kasper Hansen, Jenny Tung, and Luis Barreiro for discussions regarding multispecies methylation analysis. We acknowledge Athma Pai for sharing a methylation protocol, Bryce van de Geijn for help with the WASP pipeline, Charlotte Soneson, Jacob Degner, and Unjin Lee. We also thank the Yerkes Primate Center and Southwest Foundation for Biomedical Research, Anne Stone and Jssica Hernndez Rodrguez for providing and/or helping to ship the tissue samples. L.E.B. was supported by the National Science Foundation Graduate Research Fellowship (DGE-1144082). Additionally, L.E.B. and I.E. were funded by the Genetics and Gene Regulation Training Grant (T32 GM07197). J.R. was funded by a Swiss NSF postdoc mobility fellowship (PBLAP3-134342) and the Marie Curie International Outgoing Fellowship PRIMATE_REG_EVOL. This project was funded in part by the ORIP/OD P51OD011132 grant. T.M.B. is supported by BFU2017-86471-P

(MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya. The content presented in this article is solely the responsibility of the authors and does not necessarily reflect the official views of the funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.7 Supplementary Information

2.7.1 Supplementary Figures

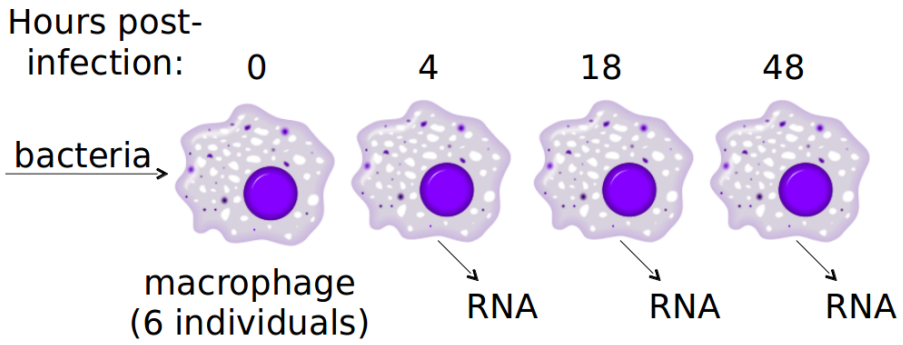


Figure 2.1: **Study design.** We infected monocyte-derived macrophages isolated from six healthy donors with the bacteria described in Table **??**. We isolated RNA for sequencing at 4, 18, and 48 hours post-infection.

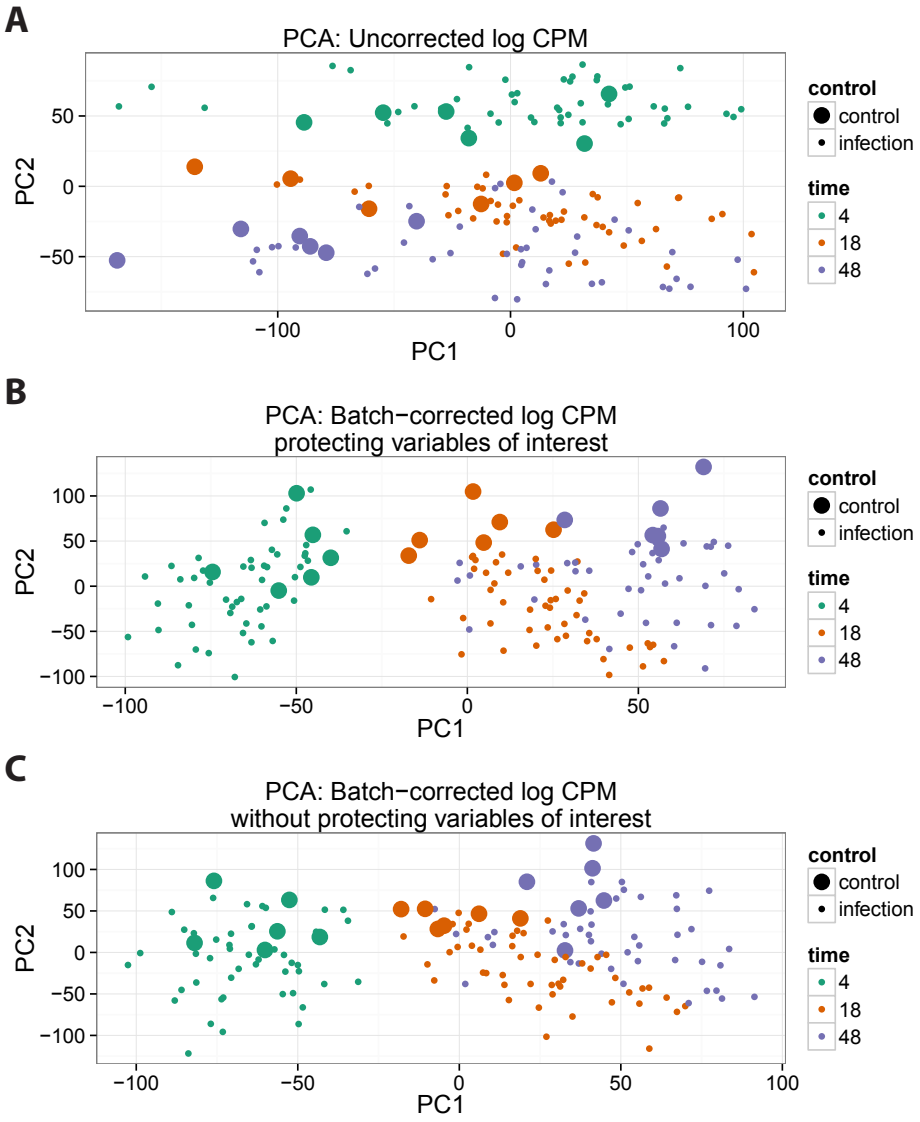


Figure 2.2: **Principal components analysis (PCA) of uncorrected and batch-corrected expression values.** (A) PCA of the TMM-normalized $\log_2$-transformed counts per million (CPM). Infected and control samples are not well separated. PC2 separates the samples by timepoint. (B) PCA of the TMM-normalized $\log_2$-transformed CPM after removing the effects of RIN score and processing batch. PC1 separates the samples by timepoint. PC2 separates the infected and control samples. (C) PCA of the TMM-normalized $\log_2$-transformed CPM after removing the effects of RIN score and processing batch without protecting the variables of interest (individual, bacteria, timepoint).

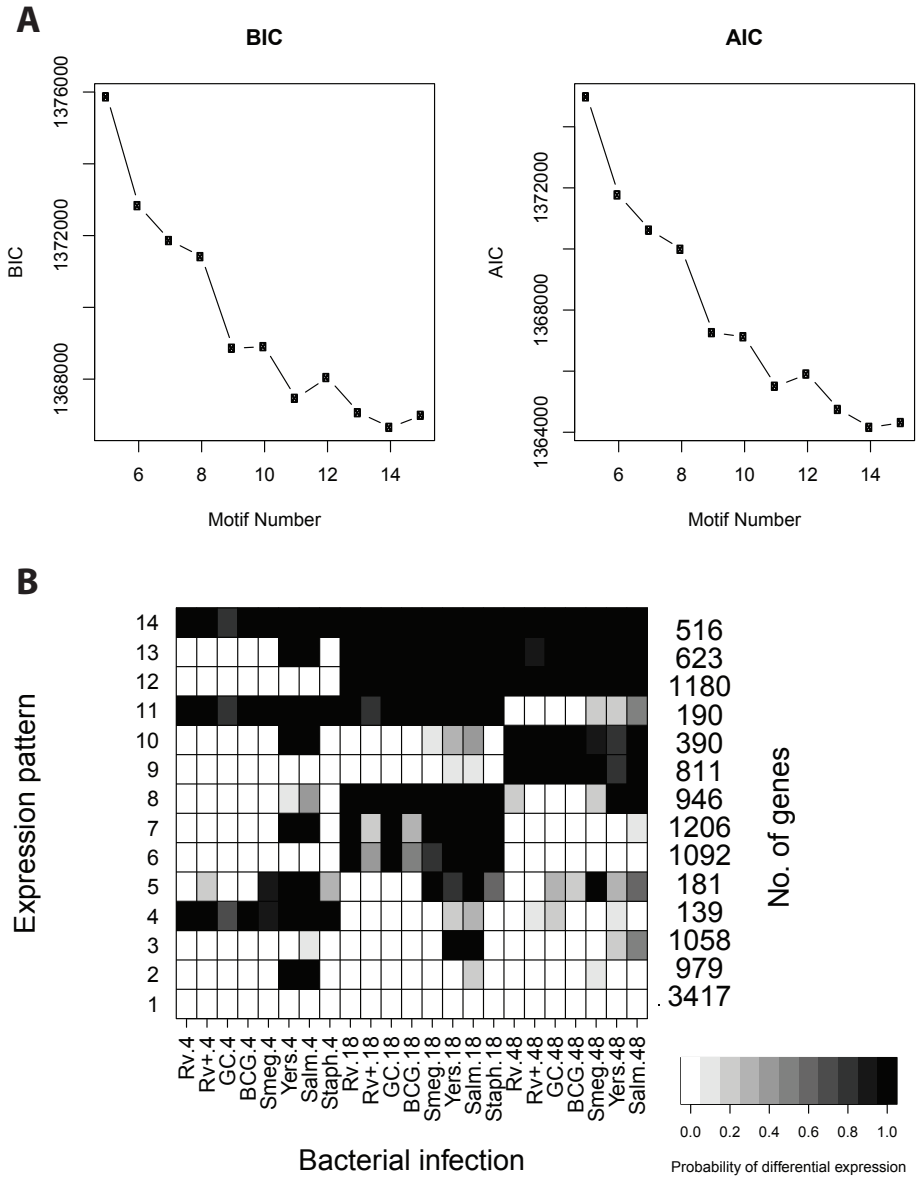


Figure 2.3: **Joint Bayesian analysis with 14 expression patterns.** (A) Cormotif [201] estimates the number of expression patterns (i.e. motifs, using their terminology) to use by calculating the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). These criteria penalize models for additional parameters to avoid overfitting. The model with the lowest BIC/AIC is considered the best fit, which in this context is the model with 14 expression patterns. (B) Joint analysis with Cormotif. The shading of each box represents the posterior probability that a gene assigned to the expression pattern (row) is differentially expressed in response to infection with a particular bacteria (column), with black representing a high posterior probability and white a low posterior probability.

Figure 2.3: (continued) The expression patterns have the following interpretations: "non-DE" - Genes that do not respond to infection; "Yers-Salm-4h" - Genes that respond 4 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "Yers-Salm-18h" - Genes that respond 18 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "4h" - Genes that respond to 4 hours post-infection with any bacteria; "non-MTB" - Genes that respond at 4, 18, and 48 hours post-infection to bacteria that are not MTB or BCG (attenuated *M. bovis*); "Virulent-18h" - Genes that respond 18 hours post-infection with virulent bacteria; "Virulent-18h+Yers-Salm-4h" - Genes that respond 18 hours post-infection with virulent bacteria and 4 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "18h+Yers-Salm-48h" - Genes that respond 18 hours post-infection with any bacteria and 48 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "48h" - Genes that respond 48 hours post-infection with any bacteria; "48h+Yers-Salm-4h" - Genes that respond 48 hours post-infection with any bacteria and 4 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "4&18h" - Genes that respond 4 and 18 hours post-infection with any bacteria; "18&48h" - Genes that respond 18 and 48 hours post-infection with any bacteria; "18&48h+Yers-Salm-4h" - Genes that respond 18 and 48 hours post-infection with any bacteria and 4 hours post-infection with *Y. pseudotuberculosis* or *S. typhimurium*; "All" - Genes that respond at 4, 18, and 48 hours post-infection with any bacteria.

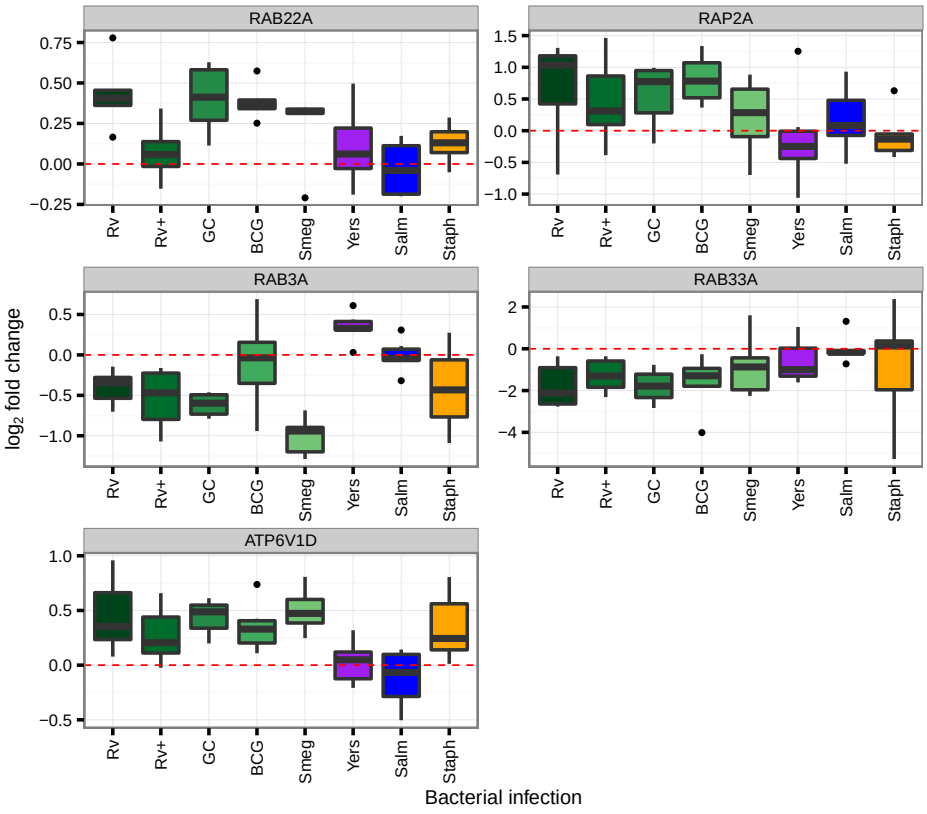


Figure 2.4: **Expression of genes involved in phagosome maturation.** *RAB22A*, *RAP2A*, and *ATP6V1D* are upregulated in response to infection with mycobacteria at 18 hours; whereas, *RAB3A* and *RAB33A* are downregulated (pattern "MTB" in Fig. **??**).

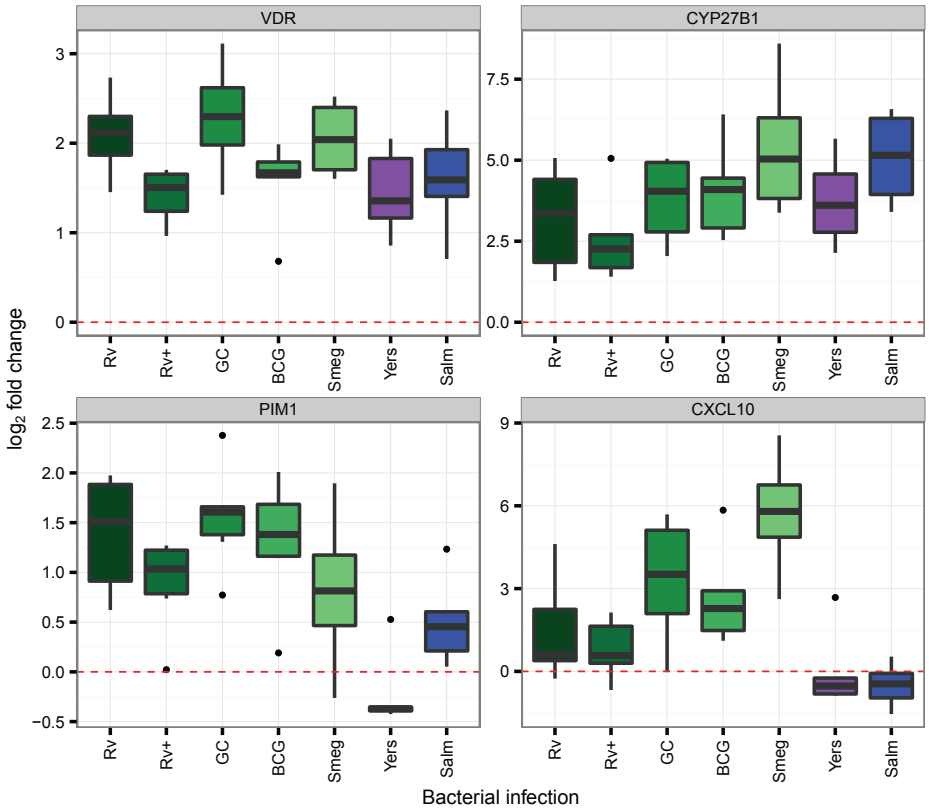


Figure 2.5: **Expression of genes involved in vitamin D signaling.** *VDR* and *CYP27B1* are upregulated at 48 hours post-infection with all bacteria (pattern "All" in Fig. **??**). *PIM1* and *CXCL10* are upregulated at 48 hours post-infection with the mycobacteria (pattern "MTB" in Fig. **??**).

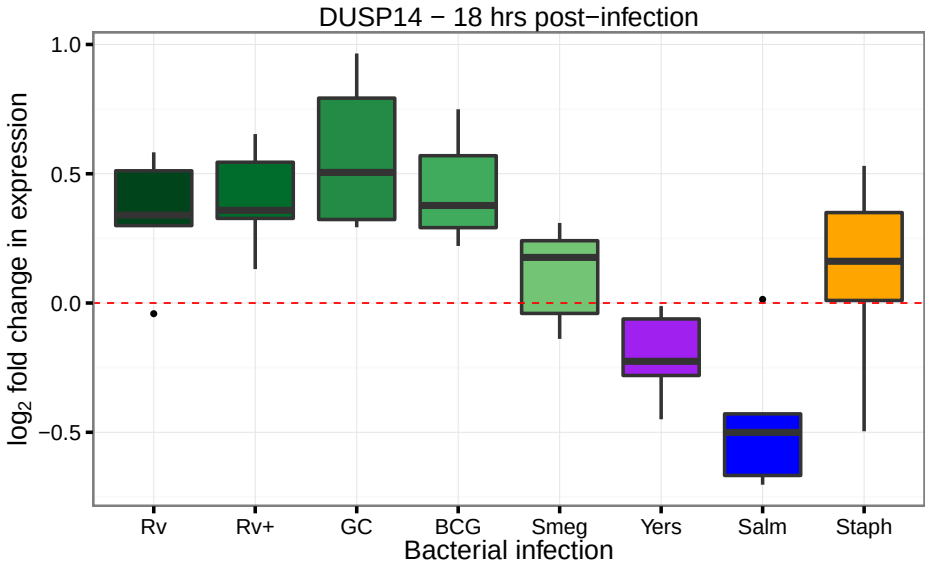


Figure 2.6: **Expression of *DUSP14* at 18 hours post-infection.** *DUSP14* is an example of an interesting gene not identified with our approach. At 18 hours, it is upregulated after infection with MTB H37Rv (q-value: 16%), MTB GC1237 (q-value: 3%), and BCG (q-value: 9%) (the change in heat-inactivated MTB H37Rv had a q-value of 26%); and downregulated post-infection with *S. typhimurium* (q-value: 9%). Because it did not fit well into one of the main patterns of gene expression identified at 18 hours post-infection, it was classified as the "non-DE" pattern.

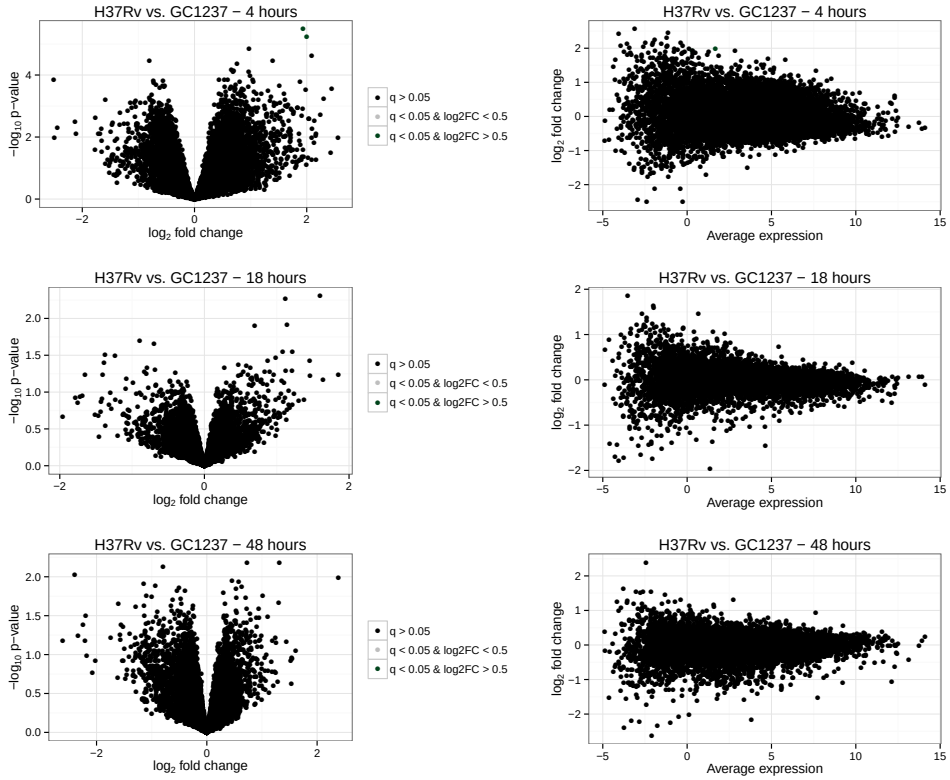


Figure 2.7: **Little difference in transcriptional response to infections with different MTB strains.** Few statistically significant differences were identified when explicitly testing gene expression levels post-infection with MTB H37Rv and MTB GC1237 at 4, 18, or 48 hours post-infection (top, middle, and bottom panels, respectively). The volcano plots (left) display the -$\log_{10}$ transformed p-value versus the $\log_2$ fold change in expression level. The MA plots (right) display the $\log_2$ fold change in expression level versus the average expression level. Most of the genes are labeled black indicating that their FDR value is greater than 5%. Two genes (labeled green) at 4 hours post-infection had a q-value < 0.05 and $\log_2$ fold change greater than 0.5 (see Supplementary Table 2.8).

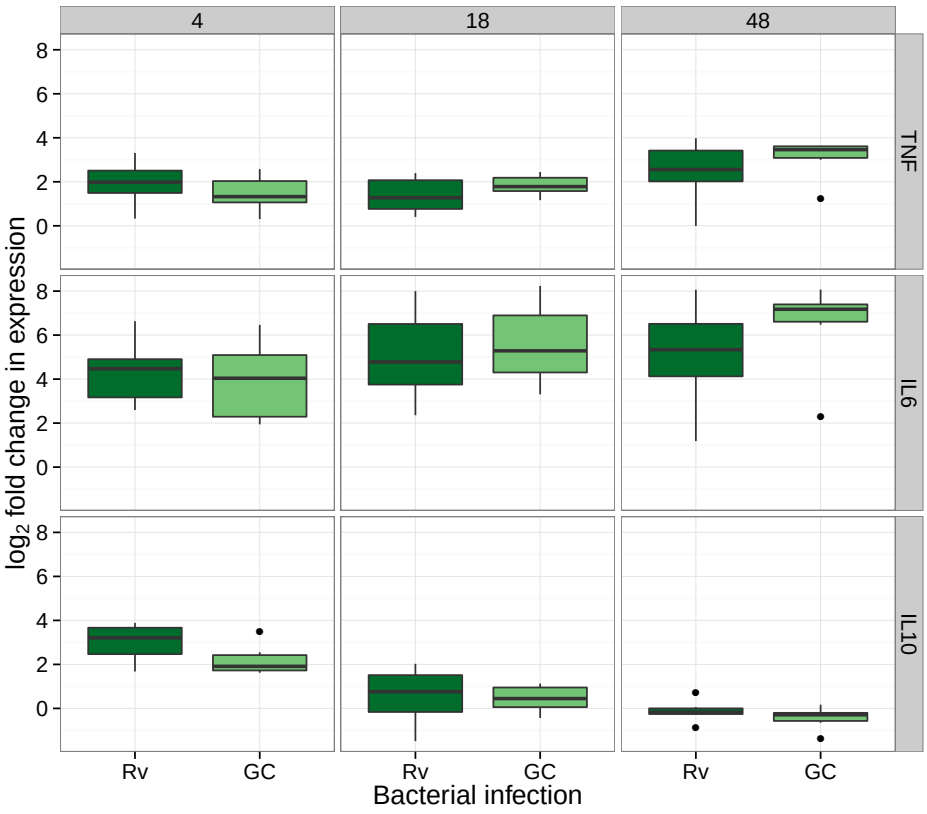


Figure 2.8: **Response of example cytokines to infection with different MTB strains.** The $\log_2$ fold change in expression of the pro-inflammatory cytokines *TNF* and *IL6* and the anti-inflammatory cytokine *IL10* is similar post-infection with MTB H37Rv or MTB GC1237. *TNF* and *IL6* are upregulated at all three timepoints; whereas, *IL10* is upregulated only at 4 hours post-infection.

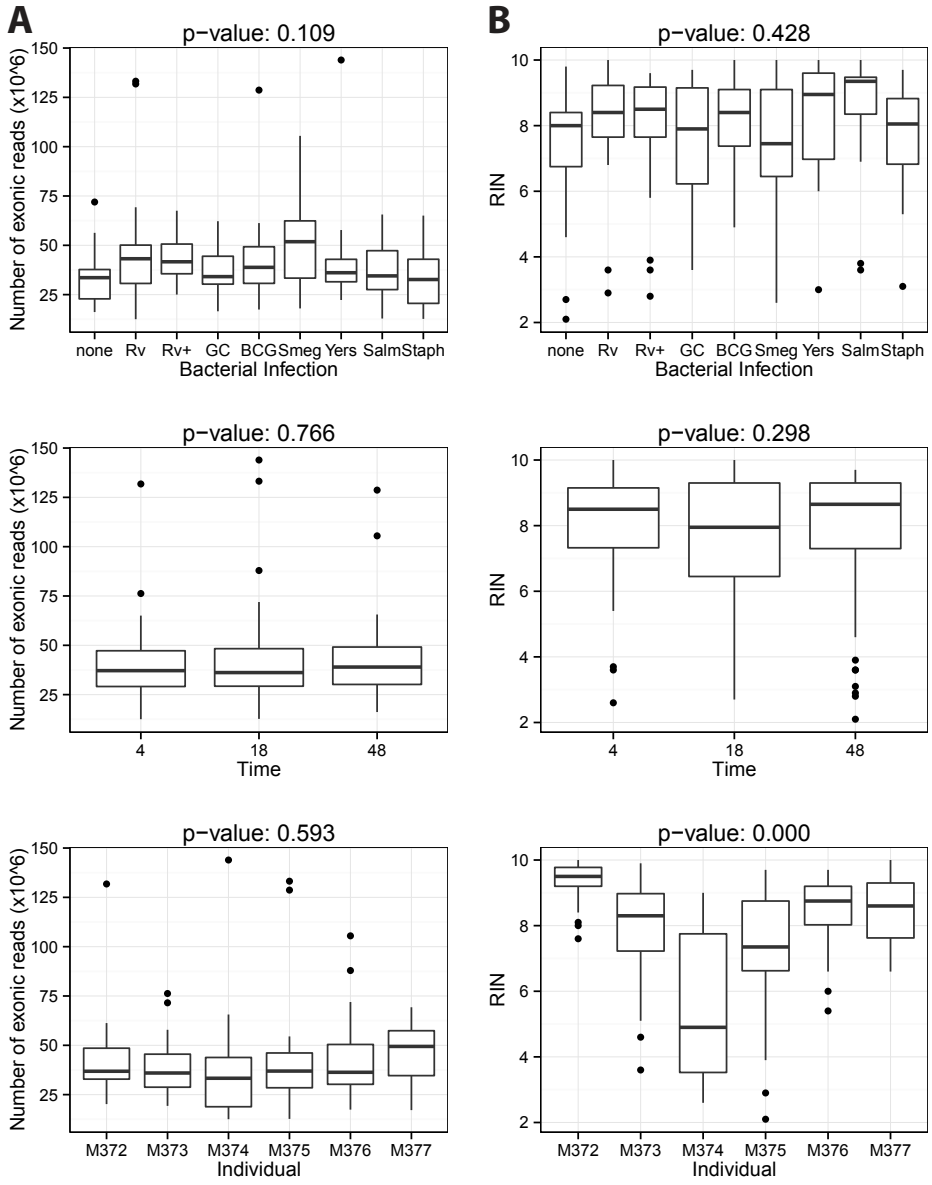


Figure 2.9: **Distribution of the number of exonic reads and RNA quality scores (RIN) across variables of interest.** (A) The number of exonic reads is evenly distributed across the bacterial infections, timepoints, and individuals. (B) The RIN scores are evenly distributed across the bacterial infections and timepoints; however, the RIN does vary between the individuals. The p-values are from an F-test.

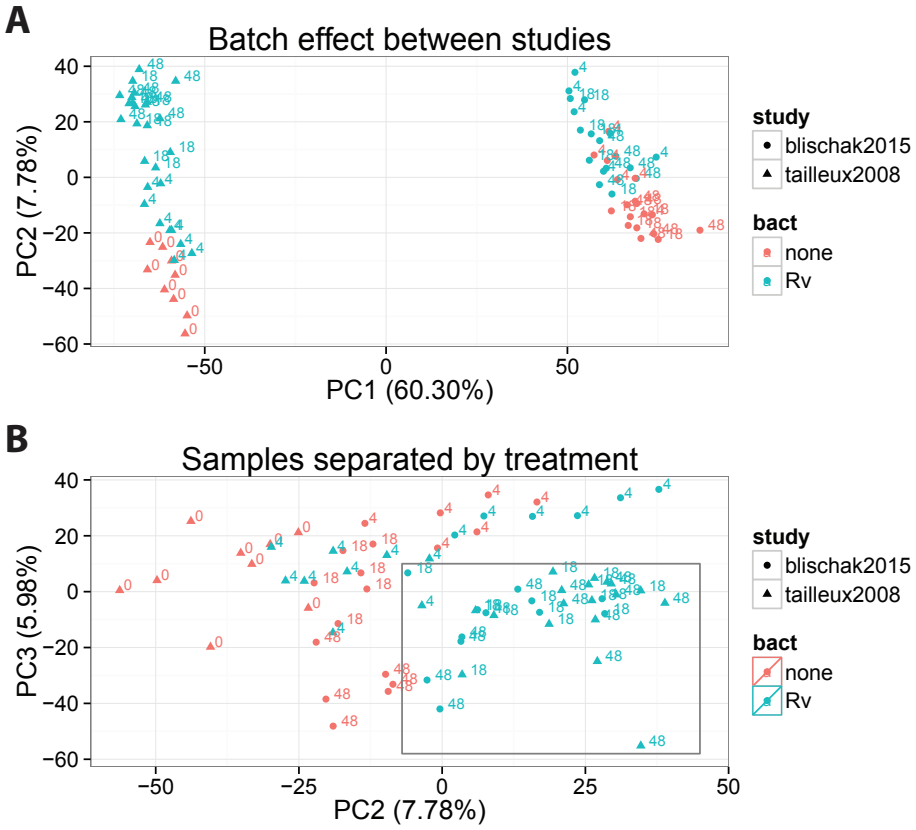


Figure 2.10: **Comparison to Tailleux et al., 2008.** We compared our RNA-seq data to the microarray data of Tailleux et al., 2008 [181] to confirm a consistent signature of infection. From our experiment, we used the batch-corrected TMM-normalized $\log_2$-transformed counts per million (CPM) from the MTB H37Rv infected macrophages at 4, 18, and 48 hours post-infection and their time-matched controls. From their experiment, we used the $\log_2$-transformed quantile-normalized data from the MTB H37Rv infected macrophages at 4, 18, and 48 hours post-infection as well as the zero timepoint non-infection control. In addition to the difference in technology, the macrophages were isolated via positive selection in our study and negative selection in theirs. Despite these differences, we still observe a common transcriptional signature of infection when performing principal components analysis (PCA). (A) PC1 is the expected batch effect between the two experiments. (B) Plotting PC2 versus PC3, the infected samples at 18 and 48 hours (when there is a strong transcriptional response; see Fig. **??**A) from the two different studies cluster together. The quantile-normalized data from Tailleux et al., 2008 [181] is available at `https://bitbucket.org/jdblischak/tb-data`.

2.7.2 Supplementary Tables

Table 2.1: **Gene expression matrix.** (see supplementary file associated with this dissertation) Contains the batch-corrected $\log_2$ counts per million for the 12,728 Ensembl genes analyzed in this study for each of the 156 samples. The column names are in the format "individual.infection.time". It can also be downloaded from `http://giladlab.uchicago.edu` or `https://bitbucket.org/jdblischak/tb-data`.

Table 2.2: **Differential expression results.** (see supplementary file associated with this dissertation) Contains the differential expression statistics from limma. This includes the $\log_2$ fold change (logFC), average expression level (AveExpr), t-statistic (t), p-value (P.Value), q-value (adj.P.Val), and log-odds (B). The column names also contain the infection and timepoint for the given comparison. It can also be downloaded from `http://giladlab.uchicago.edu` or `https://bitbucket.org/jdblischak/tb-data`.

Table 2.3: **Joint Bayesian analysis results.** (see supplementary file associated with this dissertation) Contains the assigned expression patterns for the 12,728 Ensembl genes analyzed in this study for each of the three analyses in Fig. **??**, **??**, and **??**. The columns "full_time_course", "time_18h", and "time_48h" correspond to Fig. **??**, **??**, and **??**, respectively.

Table 2.4: **Joint Bayesian analysis results with gene descriptions.** (see supplementary file associated with this dissertation) Contains the assigned expression patterns for the 12,728 Ensembl genes analyzed in this study for each of the three analyses in Fig. **??**, **??**, and **??**. It is the same information as Supplementary Table 2.3, but with the genes from each pattern from each of the three figures in its own sheet of the workbook. Furthermore, it contains the gene descriptions from Ensembl.

Table 2.5: **Gene ontology results.** (see supplementary file associated with this dissertation) Contains the gene ontology results for each of the expression patterns for the three analyses in Fig. **??**, **??**, and **??**.

Table 2.6: **RNA quality.** (see supplementary file associated with this dissertation) Contains the RNA Integrity Number (RIN) and molarity (nmol/L) measured with a Bioanalyzer (Agilent) for each of the 156 samples.

Table 2.7: **Number of differentially expressed genes from intersecting gene lists.** (see supplementary file associated with this dissertation) Contains the results of intersecting lists of differentially expressed genes for all pairwise comparisons (within each of the three timepoints).

Table 2.8: **Number of differentially expressed genes from pairwise tests.** (see supplementary file associated with this dissertation) Contains the number of differentially expressed genes when performing all pairwise tests between bacterial infections for each of the three timepoints.

Table 2.9: **Concordance in direction of effect for genes in each expression pattern.** (see supplementary file associated with this dissertation) Cormotif does not distinguish between the direction of the effect when assigning a gene to a given expression pattern. For example, a gene that is upregulated in one infection but downregulated in another is indistinguishable from a gene that is upregulated in response to both infections. However, in this data set, this is a rare effect. We calculated the percent concordance for the genes in the expression patterns from the three separate analyses. For example, for the expression pattern "MTB", 100% would indicate the gene is regulated in the same direction in the five mycobacterial infections, 80% would indicate that the gene is regulated in the same direction for four of the five mycobacterial infections, etc. "num_concord" is the number of genes in that expression pattern that are 100% concordant across the infections. "num_discord" is the number of genes in that expression pattern that are not 100% concordant. "mean_perc_concord" is the mean percent concordance of all the genes in that expression pattern.

CHAPTER 3

PREDICTING SUSCEPTIBILITY TO TUBERCULOSIS BASED ON GENE EXPRESSION PROFILING

3.1 Abstract[1]

Tuberculosis is a deadly infectious disease, which kills millions of people every year. The causative pathogen, *Mycobacterium tuberculosis* (MTB), is estimated to have infected up to a third of the world's population; however, only approximately 10% of healthy individuals progress to active TB disease. Despite evidence for heritability, it is not currently possible to predict whether a healthy person is susceptible to TB. To explore approaches to classify susceptibility to TB, we infected dendritic cells (DCs) from individuals known to be susceptible or resistant to TB with MTB, and measured genome-wide gene expression levels in infected and uninfected cells. We found hundreds of differentially expressed genes between susceptible and resistant individuals in the non-infected cells. We further found that genetic polymorphisms in proximity to the differentially expressed genes between susceptible and resistant individuals are more likely to be associated with TB susceptibility in published GWAS data. In particular, we identified two promising candidate genes: *CCL1* and *UNC13A*. Lastly, we trained a classifier based on the gene expression levels in the non-infected cells, and demonstrated decent performance on our data and an independent data set. Overall, our promising results from this small study suggest that training a classifier on a larger cohort may enable us to accurately predict TB susceptibility.

1. Citation for chapter: Blake LE*, Thomas SM*, Blischak JD, Hsiao CJ, Chavarria C, Myrthil C, Gilad Y, Pavlovic BJ. 2017. A comparative study of endoderm differentiation in humans and chimpanzees. Genome Biology 19(1):162. * denotes equal contribution.

3.2 Introduction

Tuberculosis (TB) is a major public health issue. Worldwide, over a million people die of TB annually, and millions more currently live with the disease [203, 202, 60]. Successful treatment requires months of antibiotic therapy [170], and the difficulty of adhering to the full treatment regimen has lead to the emergence of drug-resistant strains of *Mycobacterium tuberculosis* (MTB) [159].

Approximately a third of the world's population has been infected with MTB, but most are asymptomatic. While these naturally resistant individuals are able to avoid active disease, MTB persists in a dormant state inside their innate immune cells, known as a latent TB infection [131]. In contrast, approximately 10% of individuals will develop active TB after infection with MTB [135, 137]. Unfortunately, we are currently unable to predict if an individual is susceptible. While twin and family studies have indicated a heritable component of TB susceptibility [85, 33, 32, 130], genome wide association studies (GWAS) have only identified a few loci with low effect size [190, 121, 189, 144, 30, 36, 169]. Due to the highly polygenic architecture, it may be informative to examine differences between susceptible and resistant individuals at a higher level of organization, e.g. gene regulatory networks. Using this approach, previous studies have characterized gene expression profiles in innate immune cells isolated from individuals known to be susceptible or resistant to infectious diseases, including tuberculosis [188] and acute rheumatic fever [20].

We hypothesized that gene expression profiles in innate immune cells may be used to classify individuals with respect to their susceptibility to develop an active TB infection. To test this hypothesis, we isolated innate immune cells from individuals that are resistant or susceptible to TB and infected them with MTB. We discovered that the gene expression differences between resistant and susceptible innate immune cells were present primarily in the non-infected state, that these differentially expressed genes were enriched for nearby SNPs with low p-values in TB susceptibility GWAS, and furthermore, that these gene expression

levels could be used to classify individuals based on their susceptibility status.

3.3 Results

3.3.1 Susceptible individuals have an altered transcriptome in the non-infected state

We obtained whole blood samples from 25 healthy individuals (Supplementary Table 3.1). Six of the donors had recovered from a previous active TB infection, and are thus susceptible. The remaining 19 tested positive for a latent TB infection without ever experiencing symptoms of active TB, and are thus resistant. We isolated dendritic cells (DCs) and treated them with *Mycobacterium tuberculosis* (MTB) or a mock control for 18 hours. To measure genome-wide gene expression levels in infected and non-infected samples, we isolated and sequenced RNA using a processing pipeline designed to minimize the introduction of unwanted technical variation (Supplementary Fig. 3.4). We obtained a mean ($\pm$ SEM) of 48 $\pm$ 6 million raw reads per sample. We performed quality control analyses to remove non-expressed genes (Supplementary Fig. 3.5; Supplementary Table 3.2), identify and remove outliers (Supplementary Fig. 3.6, 3.7, 3.8), and check for confounding batch effects (Supplementary Fig. 3.9, 3.10). Ultimately 6 samples failed the quality checks and were removed from all downstream analyses (Supplementary Fig. 3.8).

We performed a standard differential expression analysis using a linear modeling framework (Supplementary Table 3.3), defined in equation (3.1). As expected, there was a strong response to infection with MTB in both resistant and susceptible individuals (Supplementary Fig. 3.11). Considering the resistant individuals, we identified 3,486 differentially expressed (DE) genes between the non-infected and infected states at a q-value of 10% and an arbitrary absolute log-fold change greater than 1. Similarly, 3,789 genes were DE between the non-infected and infected states for susceptible individuals at a q-value of 10% and an abso-

lute log fold change greater than 1. The DE genes included the important immune response factors *IL12B*, *REL*, and *TNF*. While the treatment effect was obvious in all individuals, of most interest were the patterns of gene expression differences between susceptible and resistant individuals in either the non-infected or infected states (Fig. 3.1). We identified 645 DE genes between resistant and susceptible individuals in the non-infected state at a q-value of 10%, including *ATPV1B2*, *FEZ2*, *PSMA2*, *TNFRSF25*, and *TRIM38*. In contrast, no genes were DE between resistant and susceptible individuals in the infected state (q-value of 10%).

3.3.2 Differentially expressed genes are enriched with TB susceptibility loci

We next sought evidence that genes classified as DE in our *in vitro* experimental system play a role in determining susceptibility to TB. To do this, we intersected our results with those from a TB susceptibility GWAS conducted in The Gambia and Ghana [190]. To perform a combined analysis of the both data sets, we coupled each gene in our expression data with the GWAS SNP with the lowest p-value among all tested SNPs located within 50 kb of the gene's transcription start site (Supplementary Table 3.4). We then asked whether the GWAS SNPs coupled with the genes we classified as DE between susceptible and resistant individuals in our experiment are enriched for low GWAS p-values compared to SNPs coupled to randomly chosen genes. Specifically, we calculated the fraction of SNPs with a GWAS p-value less than 0.05 among SNPs coupled with ranked subsets of genes whose expression profiles show increasing difference between susceptible and resistant individuals (the effect size was the absolute value of the log fold change in our experiment). In order to assess the significance of the observations, we performed 100 permutations of the enrichment analysis to derive an empirical p-value (Fig. 3.2b). Using this approach, we observed a clear enrichment (empirical $P < 0.01$) of low GWAS p-values for SNPs coupled with the genes classified as DE between susceptible and resistant individuals (Fig. 3.2a). We obtained similar results for the Ghana GWAS; see Supplementary Fig. 3.12).

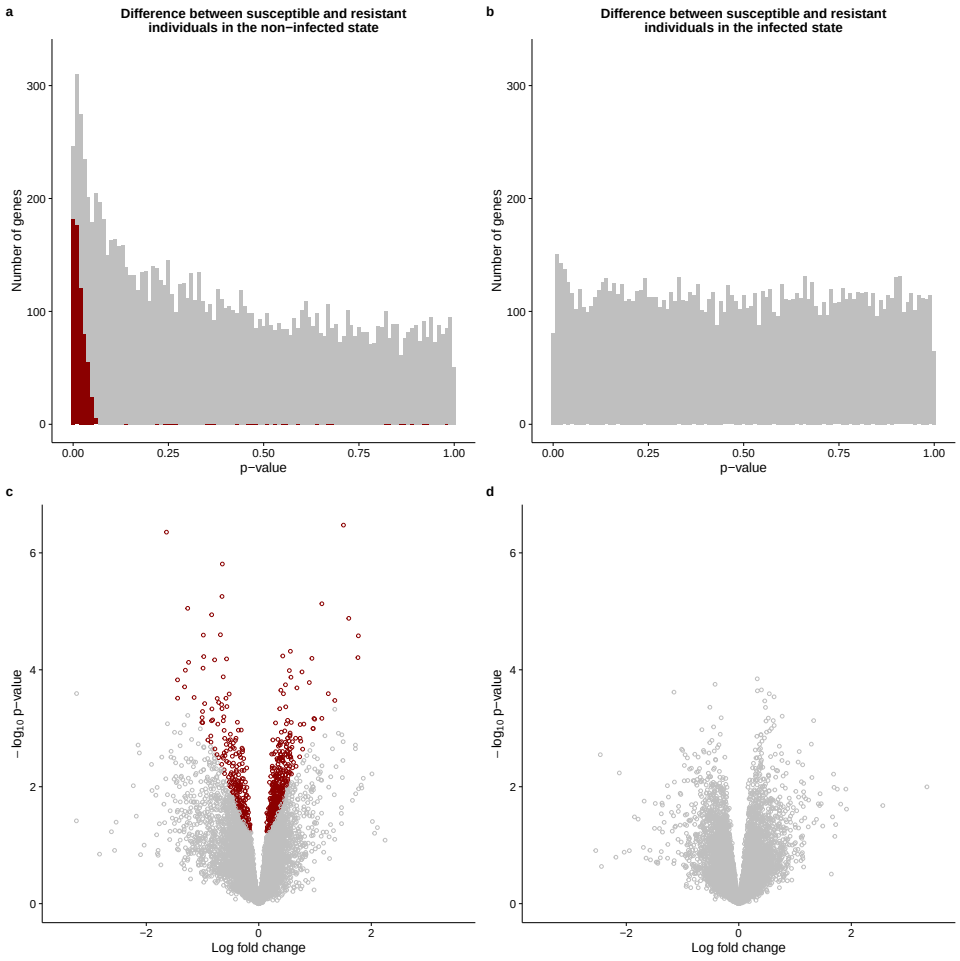


Figure 3.1: **Differential expression analysis.** The top panel contains the distribution of unadjusted p-values after testing for differential expression between susceptible and resistant individuals in the (a) non-infected or (b) infected state. The bottom panel contains the corresponding volcano plots for the (c) non-infected and (d) infected states. The x-axis is the log fold change in gene expression level between susceptible and resistant individuals and the y-axis is the $\log_{10}$ p-value. Red indicates genes which are significant differentially expressed with a q-value less than 10%.

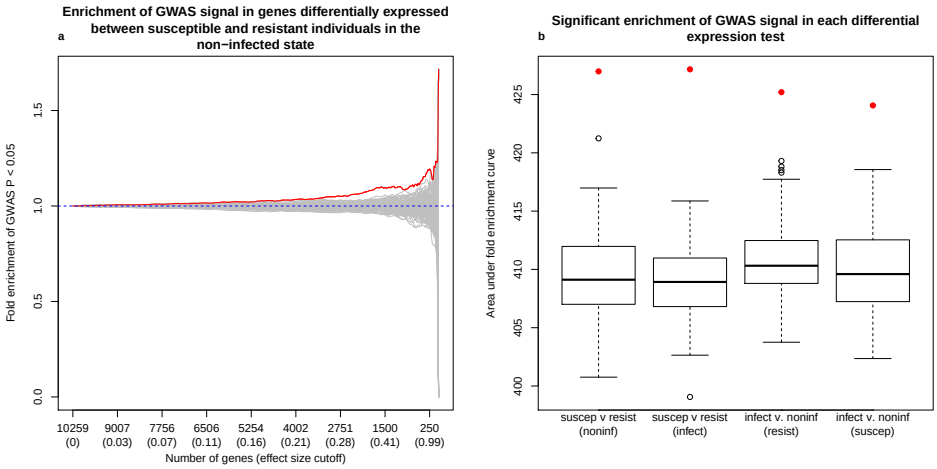


Figure 3.2: **Comparison of differential expression and The Gambia GWAS results.** (a) The y-axis is the fold enrichment (y-axis) of genes assigned a SNP with p-value less than 0.05 from the GWAS in The Gambia [190]. The x-axis is bins of genes with increasingly stringent effect size cutoffs of the absolute log fold change between susceptible and resistant individuals in the non-infected state. The effect size cutoffs were chosen such that each bin from left to right contained approximately 25 fewer genes. The red line is the results from the actual data. The grey lines are the results from 100 permutations. The dashed blue line at y=1 is the null expectation. (b) The x-axis is each of the 4 differential expression tests performed. The y-axis is the area under the curve of the fold enrichment. The boxplot is the result of the 100 permutations, and the red point is the result from the actual data. As a reference, the leftmost boxplot corresponds to the enrichment plot in (a).

We used this combined expression and GWAS data set to identify genes potentially involved in TB susceptibility. Only two genes, *CCL1* and *UNC13A*, were associated with a p-value less than 0.01 in both The Gambia and Ghana GWAS and had an absolute log fold change greater than 2 between susceptible and resistant individuals in the non-infected state (these arbitrary cutoffs were chosen to be stringent; see Supplementary Table 3.4 for the results with various cutoffs). Interestingly, these two genes were previously shown to play a role in MTB infection.

3.3.3 Susceptibility status can be predicted based on gene expression data

Next we attempted to build a gene expression-based classifier to predict TB susceptibility status (Supplementary Table 3.5). We focused on the gene expression levels measured in

the non-infected state both because this is where we observed the largest gene regulatory differences between susceptible and resistant individuals (Fig. 3.1ac), and also because, from the perspective of a translational application, it is more practical to obtain gene expression data from non-infected DCs. We trained a support vector machine using the 99 genes that were differentially expressed between resistant and susceptible individuals in the non-infected state at a q-value less than 5% (see Methods for a full description of how we selected this model). Encouragingly, we observed a clear separation between susceptible and resistant individuals when comparing the predicted probability of being resistant to TB for each sample obtained from leave-one-out-cross-validation (Fig. 3.3a). Using a cutoff of 0.75 for the predicted probability of being resistant to TB, we obtained a sensitivity of 100% (5 out of 5 susceptible individuals classified as susceptible) and a specificity of ~88% (15 out of 17 resistant individuals were classified as resistant).

Unfortunately our current data set is too small to properly split into separate training and testing sets (it is challenging to collect samples from previous TB patients, who are healthy and have no medical reason to go back for a GP visit). To our knowledge, there are also no other similar data sets available. Thus, in order to further assess the plausibility of our model, we applied the classifier to an independent study, which collected genome-wide gene expression levels in DCs from 65 healthy individuals [9], none with a previous history of TB. Using the cutoff of 0.75 for the probability of being resistant to TB (determined to be optimal in the training set), ~11% (7 of 65) of the individuals were classified as susceptible to TB. This result is intriguing similar to the estimate that roughly 10% of the general population is susceptible to TB (Fig. 3.3b).

3.4 Discussion

We obtained dendritic cells (DCs) from individuals that were known to be susceptible or resistant to developing active tuberculosis (TB) and measured genome-wide gene expression

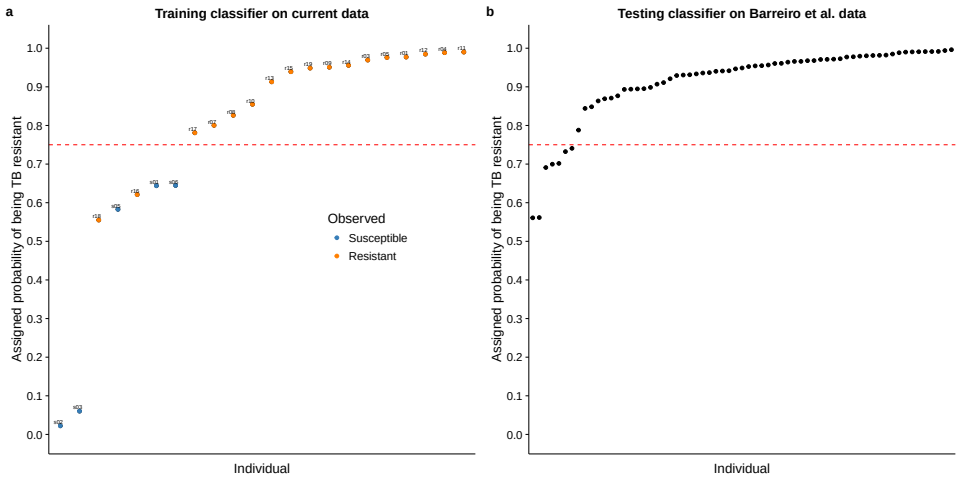


Figure 3.3: **Classifying TB susceptible individuals using a support vector machine model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, and orange those resistant to TB. The horizontal dashed red line at a probability of 0.75 separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals [9].

levels in non-infected DCs and DCs infected with *Mycobacterium tuberculosis* (MTB) for 18 hours. As expected, there were large changes in gene expression due to MTB infection in both resistant and susceptible individuals (Supplementary Fig. 3.11). We identified 645 genes, which were differentially expressed (DE) between susceptible and resistant individuals in the non-infected state; whereas, we did not observe any DE genes between susceptible and resistant individuals in the infected state (Fig. 3.1). This suggests that the differences in the transcriptomes between DCs of resistant and susceptible individuals are present pre-infection, and affect the initial response to MTB. Yet, 18 hours after infection gene expression profiles in both susceptible and resistant individuals have converged to the same gene regulatory network to fight the active infection. We chose to measure gene expression 18 hours post-infection because this time point was previously associated with a large change in genome-wide gene expression levels [181]. Given our observations, however, future studies investigating the difference in the innate immune response between individuals resistant and

susceptible to TB may want to focus on earlier time points post infection.

Among the 645 DE genes between resistant and susceptible individuals in the non-infected state, there were many interesting genes involved in important innate immune activities critical for fighting MTB and other pathogens such as autophagy [37, 26], phagolysosomal acidification, and antigen processing. In particular, *FEZ2*, a suppressor of autophagosome formation [171], was down-regulated when DCs were infected with MTB; however, in the non-infected DCs, this gene has elevated expression level in susceptible compared with resistant individuals. In turn, *ATP6V1B2*, a gene coding for a subunit of the proton transporter responsible for acidifying phagolysosomes [177, 71, 69], has increased expression in susceptible individuals compared to resistant in the non-infected state. Lastly, genes coding for nine subunits of the proteasome, which is critical for processing of MTB antigens to be presented via major histocompatibility complex (MHC) class I molecules [52, 61, 62, 114], have increased expression in susceptible individuals compared to resistant in the non-infected state. These genes are candidates for future functional studies investigating the mechanisms of TB susceptibility.

To our knowledge, our study was only the second to collect data from *in vitro* MTB infected innate immune cells isolated from individuals known to be susceptible to MTB (Thuong et al., 2008). However, there were substantial differences between our study and that of Thuong et al., 2008 [188]. First, they isolated and infected macrophages, the primary target host cell in which MTB resides; whereas, we infected DCs, which play a larger role in stimulating the adaptive immune response to MTB. Second, the susceptible individuals in Thuong et al., 2008 had an active TB infection at the time the cells were isolated; whereas, our individuals had recovered from a past TB infection. Third, we collected samples from a larger number of resistant individuals (19 versus 4), increasing our power to distinguish between the gene expression profiles of susceptible and resistant individuals.

We observed that DE genes in our *in vitro* experimental system were enriched for lower

GWAS p-values (Fig. 3.2). This suggests that such *in vitro* approaches are informative for interrogating the genetic basis of disease susceptibility. That being said, we recognized multiple caveats with this analysis. First, assigning SNPs to their nearest gene on the linear chromosome is problematic because regulatory variants can have longer range effects. Second, the fold enrichments we calculated, albeit statistically significant, were modest, indicating there were also many SNPs with low p-values nearby genes with low effect sizes in our experiment. It is possible that these variants contribute to TB susceptibility by affecting gene expression in other cell types or environmental conditions. Third, the individuals in our study were Europeans; whereas, the GWAS were conducted in Africans. Nevertheless, considering these limitations, it was encouraging that we were able to detect evidence of the genetic basis of TB susceptibility in this system.

Not only did this analysis identify a global enrichment of TB susceptibility loci, but by intersecting the expression and GWAS data, we were able to identify two genes (*CCL1* and *UNC13A*) which were marginally significant in both. Interestingly, both of these genes were previously shown to play important roles in MTB infection. *CCL1* is a chemokine that stimulates migration of monocytes [127]. In our study, it was upregulated in susceptible individuals compared to resistant in both the non-infected and infected states (but did not reach statistical significance in either) and was statistically significantly upregulated with MTB treatment. The previous differential expression study of TB susceptibility mentioned above found that *CCL1* was upregulated to a greater extent 4 hours post MTB-infection in macrophages isolated from individuals with an active TB infection (i.e. susceptible) compared to individuals with a latent TB infection (i.e. resistant) [188]. Additionally they performed a candidate gene association study and found that SNPs nearby *CCL1* were associated with TB susceptibility. In a previous study from our lab, we discovered that *CCL1* was one of only 288 genes that were differentially expressed in macrophages 48 hours post-infection with MTB and related mycobacterial species but not unrelated virulent bacteria

[15]. *UNC13A* is involved in vesicle formation [178]. In our study, it was downregulated in susceptible individuals compared to resistant in both the non-infected and infected states (but did not reach statistical significance in either) and was statistically significantly up-regulated with MTB treatment. In our past study mapping expression quantitative trait loci (eQTLs) in DCs 18 hours post-infection with MTB, *UNC13A* was one of only 98 genes which was associated with an eQTL post-infection but not pre-infection, which we called an MTB-specific eQTL [9]. Thus our new results increased the evidence that *CCL1* and *UNC13A* play important roles in TB susceptibility.

Previous attempts to use gene expression based classifiers in the context of TB have focused on predicting the status of an infection rather than the susceptibility status of an individual [10, 137, 12]. In other words, the goal of most previous study was to detect individuals in an early stage of an active TB infection when antibiotic intervention would be most effective or to monitor the effectiveness of a treatment regimen [120]. In contrast, our goal was not to distinguish between an active or latent infection, but instead to be able to determine susceptibility status before individuals have an active TB infection. Even with our small sample size, we were able to successfully train a classier with high sensitivity and decent specificity. Because such a classification of susceptibility status could affect the decision of whether or not to take antibiotics to treat a latent TB infection [131], false negatives (susceptible individuals mistakenly classified as resistant) would be much more harmful than false positives (resistant individuals mistakenly classified as susceptible), which is why we emphasized sensitivity over specificity.

At this time, we are not aware of any other data set from healthy individuals known to be sensitive to TB, with which we can further test our classifier. When we applied our classifier to an independent set of non-infected DCs isolated from healthy individuals of unknown susceptibility status, our model predicted that ~11% of the individuals were susceptible TB, which reassuringly is similar to the average in the general population (10%). Despite

this success, our results must be interpreted cautiously as a proof-of-principle due to our very small sample size of only 5 susceptible individuals. That said, our promising results in this small study suggest that collecting blood samples from a larger cohort of susceptible individuals would enable building a gene expression based classifier able to confidently assess risk of TB susceptibility. By reducing the number of resistant individuals receiving treatment for a latent TB infection, we can eliminate the adverse health effects of a 6 month regimen of antibiotics for these individuals and also reduce the selective pressures on MTB to develop drug resistance.

3.5 Methods

3.5.1 Ethics Statement

We recruited 25 subjects to donate a blood sample for use in our study. All methods were carried out in accordance with relevant guidelines and regulations. The experimental protocols were approved by the Institutional Review Boards of the University of Chicago (10-504-B) and the Institut Pasteur (IRB00006966). All study participants provided written informed consent.

3.5.2 Sample collection

We collected whole blood samples from healthy Caucasian male individuals living in France. The putatively resistant individuals tested positive for a latent TB infection in an interferon-γ release assay, but had never developed active TB. The putatively sensitive individuals had developed active TB in the past, but were currently healthy.

3.5.3 Isolation and infection of dendritic cells

We performed these experiments as previously described [9]. Briefly, we isolated mononuclear cells from the whole blood samples using Ficoll-Paque centrifugation, extracted monocytes via CD14 positive selection, and differentiated the monocytes into dendritic cells (DCs) by culturing them for 5 days in RPMI 1640 (Invitrogen) supplemented with 10% heat-inactivated FCS (Dutscher), L-glutamine (Invitrogen), GM-CSF (20 ng/mL; Immunotools), and IL-4 (20 ng/mL; Immunotools). Next we infected the DCs with *Mycobacterium tuberculosis* (MTB) H37Rv at a multiplicity of infection of 1-to-1 for 18 hours.

3.5.4 RNA extraction and sequencing

We extracted RNA using the Qiagen miRNeasy Kit and prepared sequencing libraries using the Illumina TruSeq Kit. We sent the master mixes to the University of Chicago Functional Genomics Facility to be sequenced on an Illumina HiSeq 4000. We designed the batches for RNA extraction, library preparation, and sequencing to balance the experimental factors of interest and thus avoid potential technical confounders (Supplementary Fig. 3.4).

3.5.5 Read mapping

We mapped reads to human genome hg38 (GRCh38) using Subread [110] and discarded non-uniquely mapping reads. We downloaded the exon coordinates of 19,800 Ensembl [206] protein-coding genes (Ensembl 83, Dec 2015, GRCh38.p5) using the R/Bioconductor [73] package biomaRt [43, 44] and assigned mapped reads to these genes using featureCounts [111].

3.5.6 Quality control

First we filtered genes based on their expression level by removing all genes with a transformed median $\log_2$ counts per million (cpm) of less than zero. This step resulted in a set of 11,336 genes for downstream analysis (Supplementary Fig. 3.5, Supplementary Table 3.2). Next we used principal components analysis (PCA) and hierarchical clustering to identify and remove 6 outlier samples (Supplementary Fig. 3.6, 3.7, 3.8). We did this systematically, by removing any sample whose data projections did not fall within two standard deviations of the mean for any of the first six PCs (for the first PC, which separated the samples by treatment, we calculated a separate mean for the non-infected and infected samples).

After filtering lowly expressed genes and removing outliers, we performed the PCA again to check for any potential confounding technical batch effects (Supplementary Fig. 3.9). Reassuringly, the major sources of variation in the data were from the biological factors of interest. PC1 was strongly correlated with the effect of treatment, and PCs 2-6 were correlated with inter-individual variation. The only concerning technical factor was the infection experiments, which were done in 12 separate batches (Supplementary Fig. 3.4). Infection batch correlated with PCs 3 and 5; however, we verified that this variation was not confounded with our primary outcome of interest, TB susceptibility (Supplementary Fig. 3.10).

3.5.7 Differential expression analysis

We used limma+voom [167, 104, 153] to implement the following linear model to test for differential expression:

$$Y \sim \beta_0 + X_{treat}\beta_{treat} + X_{status}\beta_{status} + X_{treat,status}\beta_{treat,status} + I + \epsilon \tag{3.1}$$

where β_0 is the mean expression level in non-infected cells of resistant individuals, β_{treat} is the fixed effect of treatment in resistant individuals, β_{status} is the fixed effect of susceptibility status in non-infected cells, $\beta_{treat,status}$ is the fixed interaction effect of treatment in susceptible individuals, and I is the random effect of individual. The random individual effect was implemented using the limma function duplicateCorrelation [168]. To jointly model the data with voom and duplicateCorrelation, we followed the recommended best practice of running both voom and duplicateCorrelation twice in succession [116].

We used the model to test different hypotheses (Supplementary Data S3). We identified genes which were differentially expressed (DE) between infected and non-infected DCs of resistant individuals by testing $\beta_{treat} = 0$, genes which were DE between infected and non-infected DCs of susceptible individuals by testing $\beta_{treat} + \beta_{treat,status} = 0$, genes which were DE between susceptible and resistant individuals in the non-infected state by testing $\beta_{status} = 0$, and genes which were DE between susceptible and resistant individuals in the infected state by testing $\beta_{status} + \beta_{treat,status} = 0$. We corrected for multiple testing using q-values estimated via adaptive shrinkage [174] and considered differentially expressed genes as those with a q-value less than 10%.

3.5.8 Combined analysis of gene expression data and GWAS results

The GWAS p-values were from a study of TB susceptibility conducted in The Gambia and Ghana [190]. To perform a combined analysis of the gene expression and GWAS data, we assigned each gene to the SNP with the minimum GWAS p-value out of all the SNPs located within 50 kb up or downstream of its transcription start site. Specifically, we obtained the genomic coordinates of the SNPs with the R/Bioconductor [73] package SNPlocs.Hsapiens.dbSNP144.GRCh38 and matched SNPs to nearby genes using GenomicRanges [105]. 10,260 of the 11,336 genes were assigned an association p-value (Supplementary Table 3.4). For each of the 4 hypotheses we tested, we performed an enrichment

analysis. To do so, we calculated the fraction of genes assigned a GWAS SNP with p-value less than 0.05 for bins of genes filtered by increasingly stringent cutoffs for the observed differential expression effect size (the absolute value of the log fold change) between susceptible and resistant individuals. The effect size cutoffs were chosen such that on average each subsequent bin differed by 25 genes. To measure enrichment, we calculated the area under the curve using the R package flux [84]. In order to assess significance, we calculated the area under the curve for 100 permutations of the data. All differential expression tests were statistically significantly enriched for SNPs low GWAS p-values in both the The Gambia (Fig. 3.2b) and Ghana (Supplementary Fig. 3.12) data sets.

3.5.9 Classifier

The training set included data from the 44 high-quality non-infected samples from this study with known susceptibility status. The test set included the 65 non-infected samples from one of our previous studies in which the susceptibility status is unknown [9], and thus assumed to be similar to that in the general population (~10%). Because the two studies are substantially different, we took multiple steps to make them comparable. First, we subset to include only those 9,450 genes which were assayed in both. Second, because the dynamic range obtained from RNA-seq (current study) and microarrays (previous study [9]) were different, we normalized the gene expression levels to a standard normal with $\mu = 0$ and $\sigma = 1$ (Supplementary Fig. 3.13). Third, we corrected for the large, expected batch effect between the two studies by regressing out the first PC of the combined expression data using the limma function removeBatchEffect [153] (Supplementary Fig. 3.14).

To identify genes to use in the classifier, we performed a differential expression analysis on the normalized, batch-corrected data from the current study using the same approach described above (with the exception that we no longer used voom [104] since the data were no longer counts). Specifically, we tested for differential expression between susceptible and re-

sistant individuals in the non-infected state and identified sets of genes to use in the classifier by varying the q-value cutoff. Cutoffs of 5%, 10%, 15%, 20%, and 25% corresponded to gene set sizes of 99, 385, 947, 1,934, and 3,697, respectively. We used the R package caret [102] to train 3 different machine learning models: elastic net [55], support vector machine [88], and random forest [112] (the parameters for each individual model were selected using the Kappa statistic). To assess the results of the model on the training data, we performed leave-one-out-cross-validation (LOOCV). In order to choose the model with the best performance, we calculated the difference between the mean of the LOOCV-estimated probabilities of being TB resistant for the samples known to be TB resistant and the corresponding mean for the samples known to be TB susceptible. This metric emphasized the ability to separate the susceptible and resistant individuals into two separate groups. Using this metric, the best performing model was the support vector machine with the 99 genes that are significantly differentially expressed at a q-value of 5% (Supplementary Fig. 3.15, Supplementary Table 3.5); however, both the elastic net (Supplementary Fig. 3.16) and random forest (Supplementary Fig. 3.17) had similar performance. Lastly, we tested the classifier by predicting the probability of being TB resistant in the 65 healthy samples (Fig. 3.3b). For evaluating the predictions on the test set of individuals with unknown susceptibility status, we used a relaxed cutoff of the probability of being TB resistant of 0.75, which was based on the ability of the model at this cutoff to classify all TB susceptible individuals in the training set as susceptible with only 2 false positives. As expected, the 99 genes used in the classifier had similar normalized, batch-corrected median expression levels in the non-infected state across both studies (Supplementary Fig. 3.18).

3.5.10 Software implementation

We automated our analysis using Python (`https://www.python.org/`) and Snakemake [101]. Our processing pipeline used the general bioinformatics software FastQC (`http://`

www.bioinformatics.babraham.ac.uk/projects/fastqc/), MultiQC [48], samtools [108], and bioawk (https://github.com/lh3/bioawk). We used R [149] for all statistics and data visualization. We obtained gene annotation information from the Ensembl [206] and Lynx [179] databases. The computational resources were provided by the University of Chicago Research Computing Center. All code is available for viewing and reuse at https://github.com/jdblischak/tb-suscept.

3.5.11 Data availability

The raw fastq files will be deposited in NCBI's Gene Expression Omnibus [45] before official publication. The RNA-seq gene counts and other summary data sets are included as Supplementary Data and are also available for download at https://github.com/jdblischak/tb-suscept/data.

3.6 Acknowledgments

We thank T. Thye for sharing the GWAS data with us. This study was funded by National Institutes of Health (NIH) Grant AI087658 to YG and LT. JDB was supported by NIH T32GM007197. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

3.7 Author Contributions

YG, LT, and LBB conceived of the study and designed the experiments. LT coordinated sample collection and performed the infection experiments. MM extracted the RNA and prepared the sequencing libraries. JDB analyzed the results. LBB and YG supervised the project. JDB wrote the paper with input from all authors.

3.8 Supplementary Information

3.8.1 Supplementary Figures

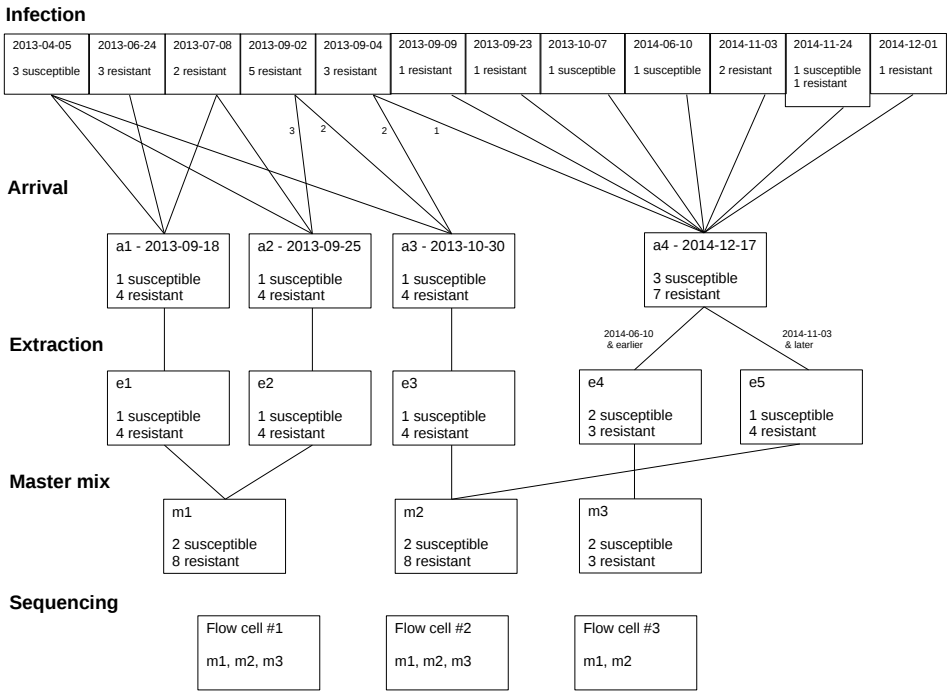


Figure 3.4: **Batch processing.** We designed the processing of the samples to minimize the introduction of technical batch effects. Specifically, we attempted to balance the processing of samples obtained from susceptible and resistant individuals. In the diagram, each box represents a batch. "Infection" labels the batches of the infection experiments, "Arrival" labels the batch shipments of cell lysates arrived in Chicago, USA from Paris, France, "Extraction" labels the batches of RNA extraction, "Master Mix" labels the batches of library preparation, and "Sequencing" labels the batches of flow cells. Each master mix listed in a flow cell batch was sequenced on only one lane of that flow cell.

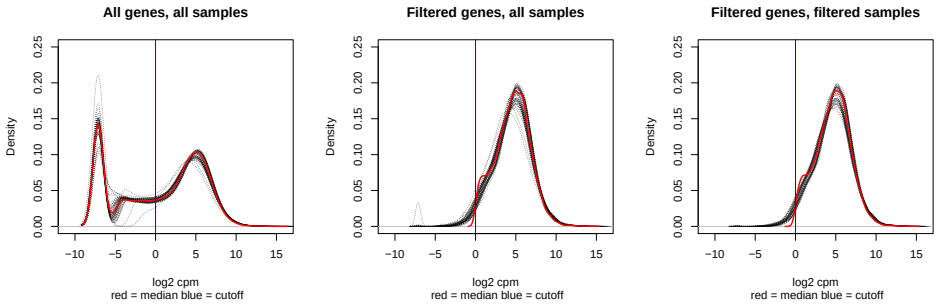


Figure 3.5: **Gene expression distributions before and after filtering genes and samples.** The $\log_2$ counts per million (cpm) of each sample is plotted as a dashed gray line. The solid red line represents the median value across all the samples. The vertical solid blue line at $x = 0$ represents the cutoff used to filter lowly expressed genes based on their median $\log_2$ cpm. The left panel is the data from all 19,800 genes and 50 samples, the middle panel is the data from the 11,336 genes remaining after removing lowly expressed genes, and the right panel is the data from 11,336 genes and the 44 samples remaining after removing outliers.

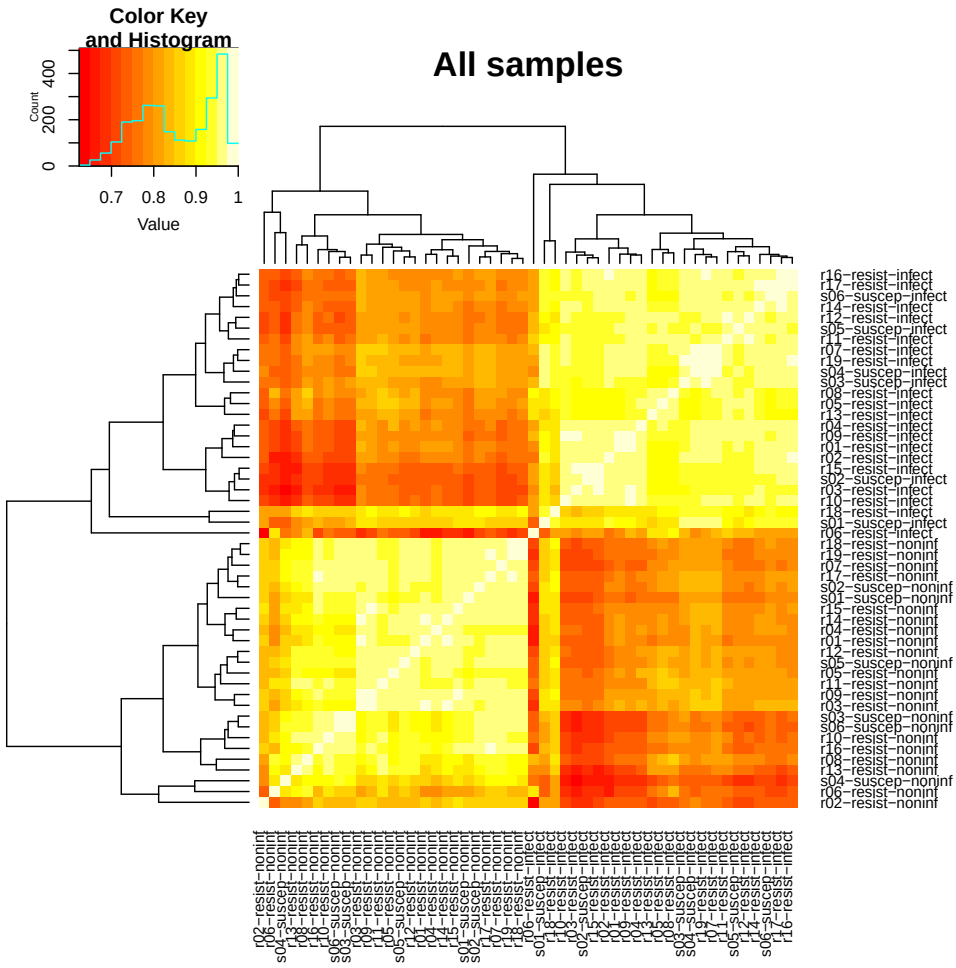


Figure 3.6: **Heatmap of correlation matrix of samples.** Each square represents the Pearson correlation between the $\log_2$ cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of hierarchical clustering with the complete linkage method. The outliers of the non-infected samples are s04-suscept-noninf, r02-resist-noninf, and r06-resist-noninf. The outliers of the infected samples are s01-suscep-infect, r06-resist-infect, and r18-resist-infect.

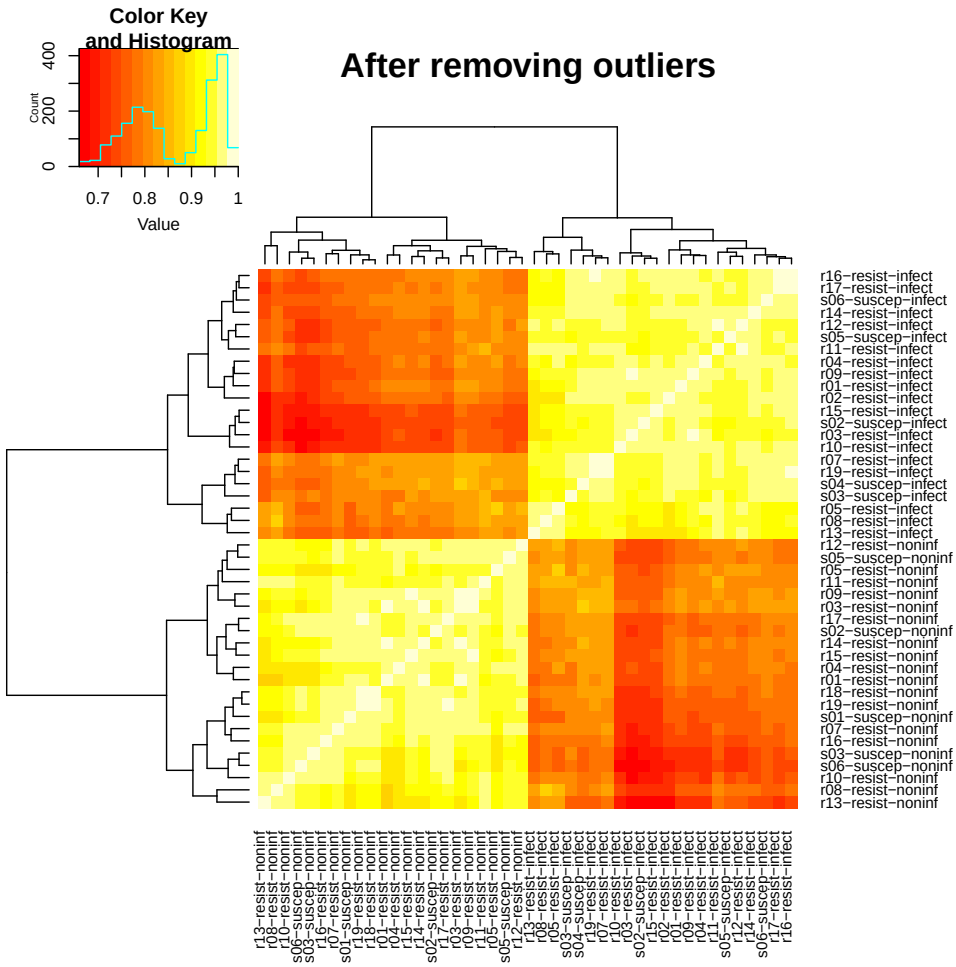
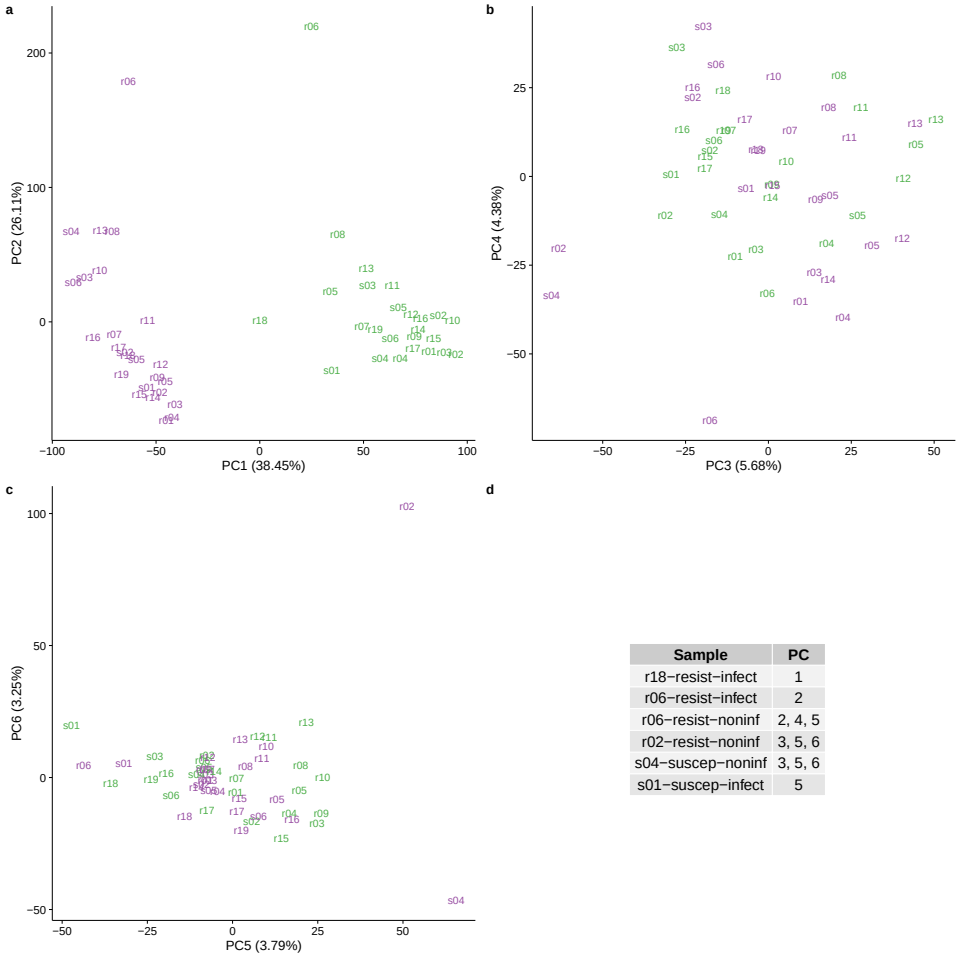


Figure 3.7: **Heatmap of correlation matrix after removing outliers.** Each square represents the Pearson correlation between the $\log_2$ cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of hierarchical clustering with the complete linkage method.



Sample	PC
r18−resist−infect	1
r06−resist−infect	2
r06−resist−noninf	2, 4, 5
r02−resist−noninf	3, 5, 6
s04−suscep−noninf	3, 5, 6
s01−suscep−infect	5

Figure 3.8: **Principal components analysis (PCA) to identify outliers.** PC1 versus PC2 (a), PC3 versus PC4 (b), and PC5 versus PC6 (c). Each sample is represented by its 3-letter ID. "s" stands for susceptible and "r" for resistant, and the text is colored on the basis of treatment status (purple is non-infected; green is infected). The value is parentheses in each axis is the percentage of total variation accounted for by that PC. The outliers are listed in (d). These samples do not fall within 2 standard deviations of the mean value of the PCs listed in the right column. Note that a separate mean was calculated for the non-infected and infected samples for PC1 only.

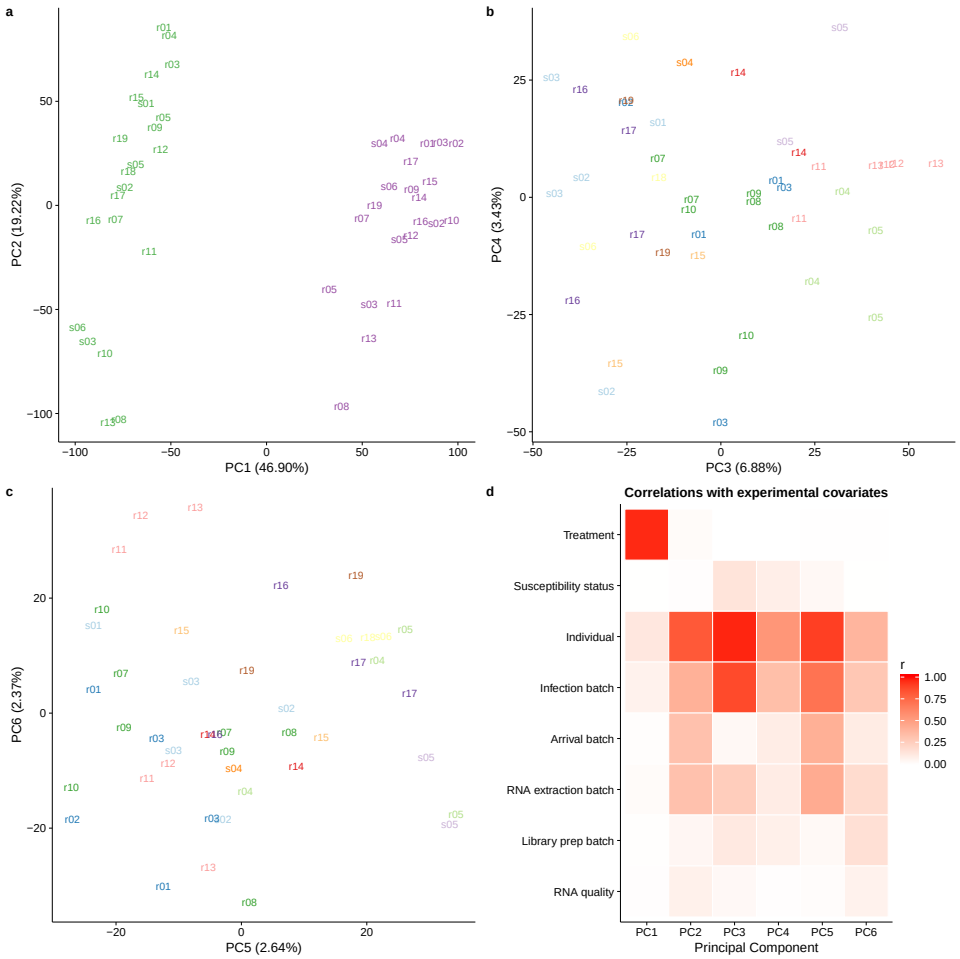


Figure 3.9: **Check for technical batch effects using principal components analysis (PCA).** (a) PC1 versus PC2. The text labels are the individual identifiers. Purple indicates non-infected samples and green indicates infected. (b) PC3 versus PC4. The colors indicate the different infection batches. (c) PC5 versus PC6. The colors indicate the different infection batches. (d) The Pearson correlation of PCs 1-6 with each of the recorded biological and technical covariates. The correlations vary from 0 (white) to 1 (red).

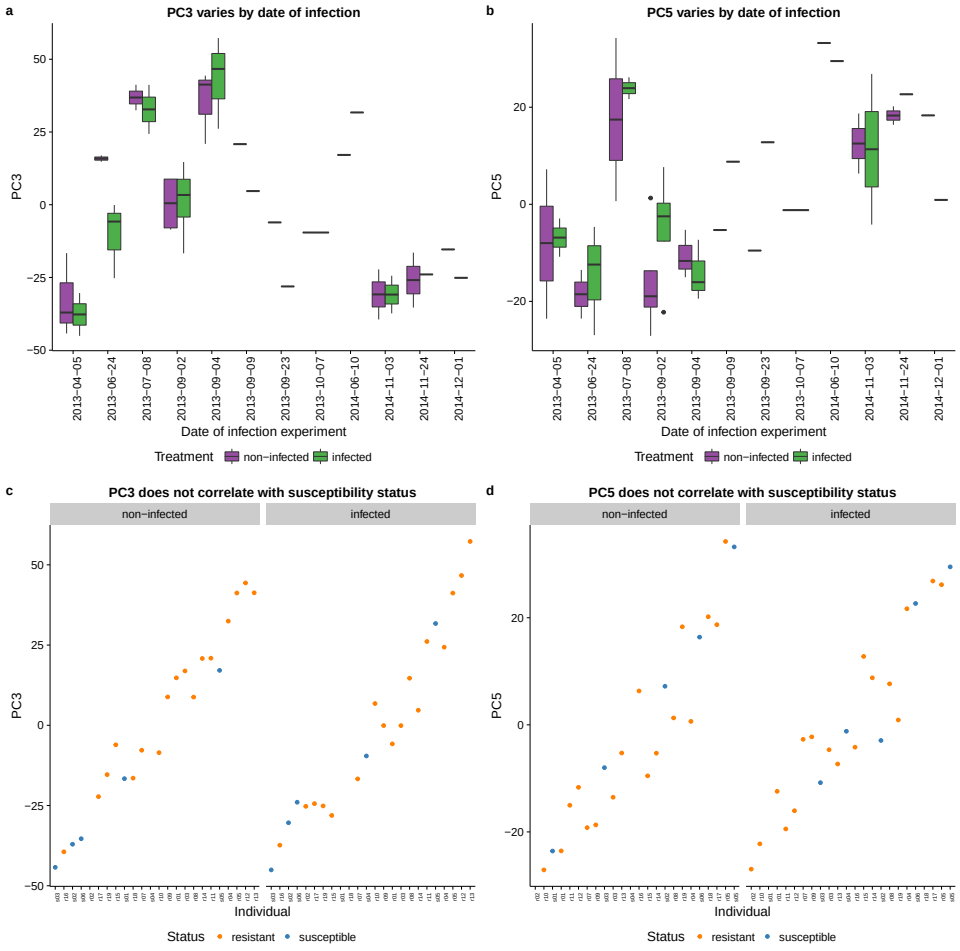


Figure 3.10: **Check for confounding effect of infection batch.** PC3 (a) and PC5 (b) varied by the date of infection. Non-infected samples are in purple and infected samples in green. Importantly, however, this technical variation arising from infection batch did not correlate with the susceptibility status of the individuals (c and d). Resistant individuals are in orange and susceptible individuals in blue.

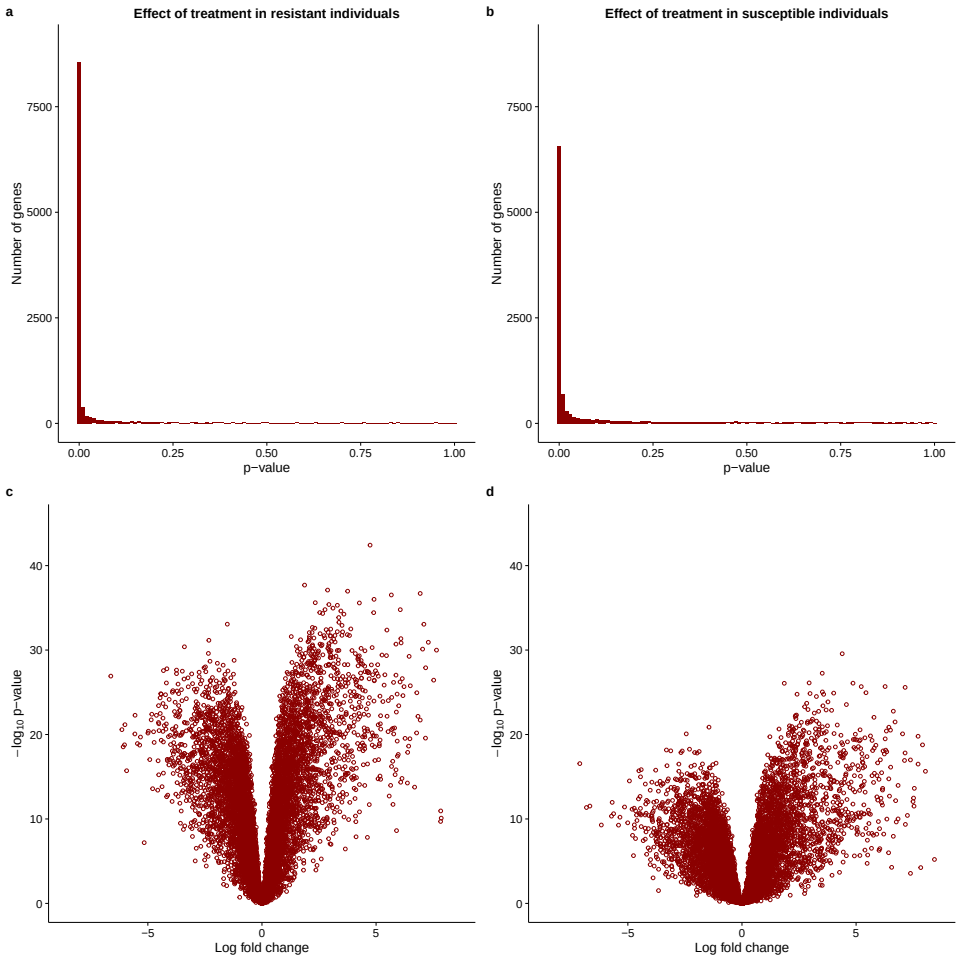


Figure 3.11: **Effect of treatment with MTB.** The top panel contains the distribution of unadjusted p-values after testing for differential expression between the non-infected and infected states in (a) resistant and (b) susceptible individuals. The bottom panel contains the corresponding volcano plots for the (c) resistant and (d) susceptible individuals. The x-axis is the log fold change in gene expression level between susceptible and resistant individuals and the y-axis is the $\log_{10}$ p-value. Red indicates genes which are significant differentially expressed with a q-value less than 10%. Because of the extremely skewed p-value distribution, all genes are significantly differentially expressed at this false discovery rate.

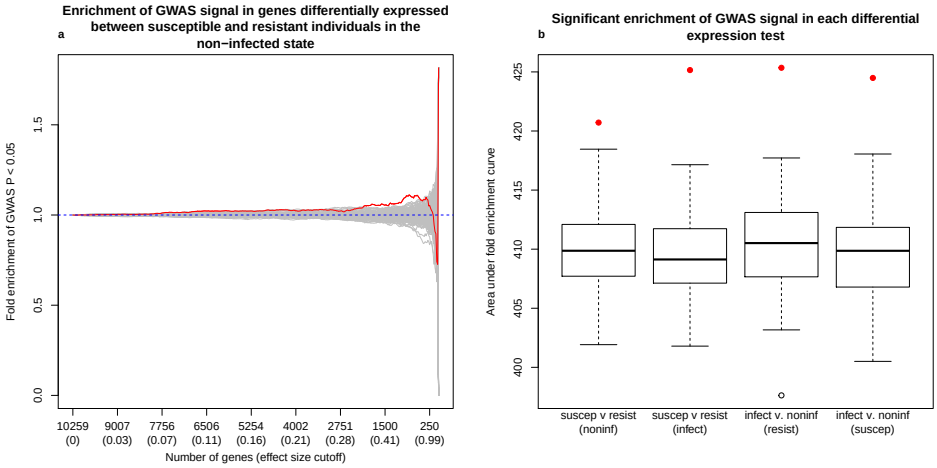


Figure 3.12: **Comparison of differential expression and Ghana GWAS results.** (a) The y-axis is the fold enrichment (y-axis) of genes assigned a SNP with p-value less than 0.05 from the GWAS in Ghana [190]. The x-axis is bins of genes with increasingly stringent effect size cutoffs of the absolute log fold change between susceptible and resistant individuals in the non-infected state. The effect size cutoffs were chosen such that each bin from left to right contained approximately 25 fewer genes. The red line is the results from the actual data. The grey lines are the results from 100 permutations. The dashed blue line at y=1 is the null expectation. (b) The x-axis is each of the 4 differential expression tests performed. The y-axis is the area under the curve of the fold enrichment. The boxplot is the result of the 100 permutations, and the red point is the result from the actual data. As a reference, the leftmost boxplot corresponds to the enrichment plot in (a).

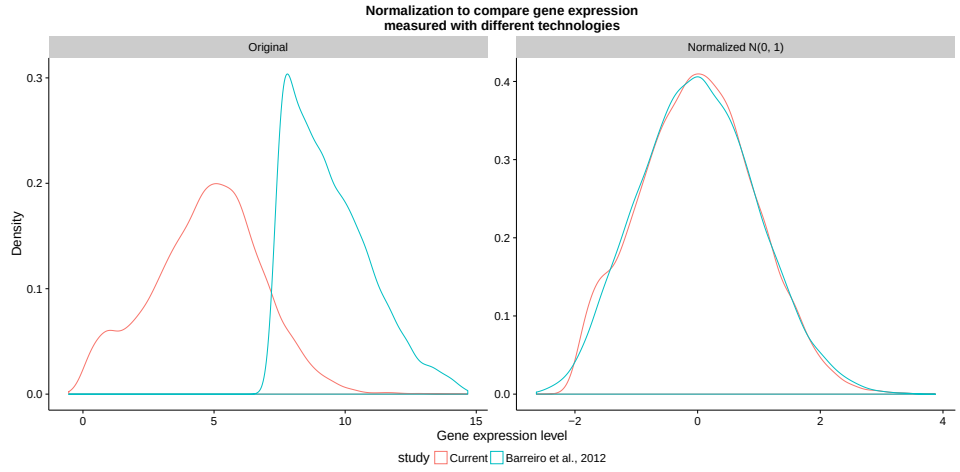


Figure 3.13: **Normalizing gene expression distributions.** (left) The distribution of the median log2 cpm of the RNA-seq data from the current study in red compared to the distribution of the median gene expression levels of the microarray data from Barreiro et al., 2012 [9] in blue. (right) The distributions of the same data sets after normalizing each sample to a standard normal distribution.

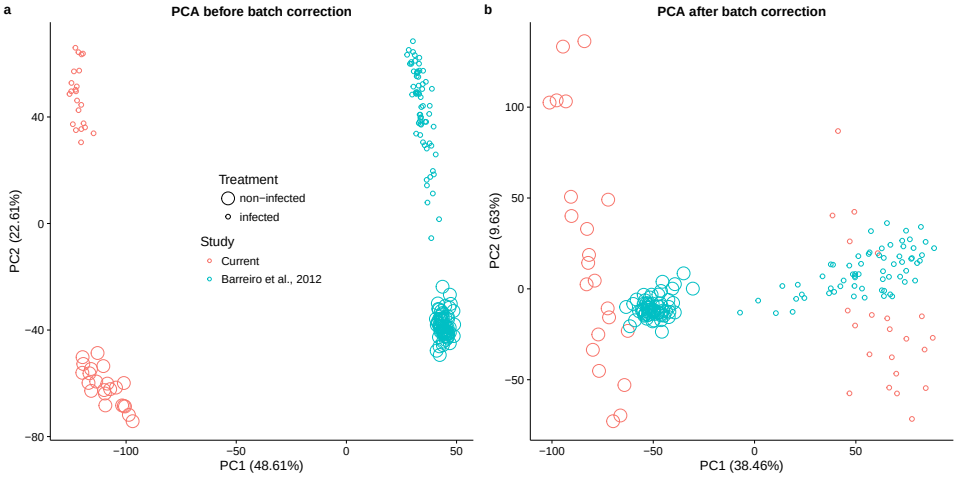


Figure 3.14: **Principal components analysis (PCA) of combined data sets.** (a) PC1 versus PC2 of the combined data set of the RNA-seq data from the current study (red) and the microarray data from Barreiro et al., 2012 [9] (blue). The large circles are non-infected samples, and the small circles are infected samples. The value in parentheses is the percentage of the total variation accounted for by that PC. (b) The same data after regressing the original PC1 in (a).

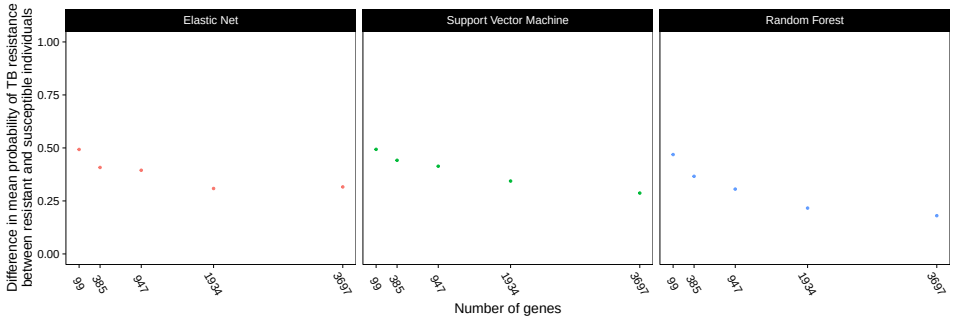


Figure 3.15: **Comparing the classification results of different methods and number of input genes.** We compared 3 different machine learning methods (elastic net, support vector machine, random forest) and used 5 different sets of input genes. The input genes (x-axis) were obtained by varying the q-value cutoff for differential expression between susceptible and resistant individuals in the non-infected state from 5% to 25%. The evaluation metric (y-axis) was the difference of the mean assigned probability of being TB resistant between the known resistant and susceptible individuals in the current study.

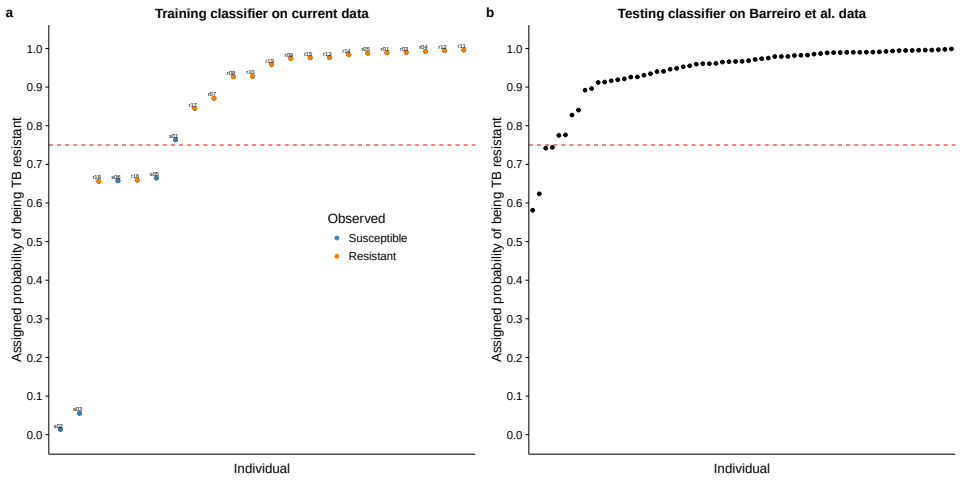


Figure 3.16: **Classifying TB susceptible individuals using an elastic net model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, and orange those resistant to TB. The horizontal blue line at a probability of 0.75 almost separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals [9].

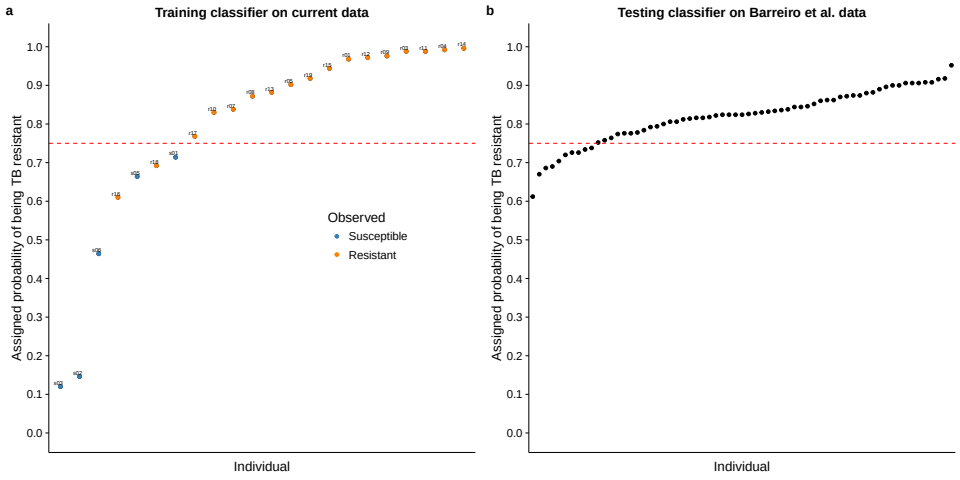


Figure 3.17: **Classifying TB susceptible individuals using a random forest model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, and orange those resistant to TB. The horizontal blue line at a probability of 0.75 separates susceptible and resistant individuals. (b) The estimates of predicted probability of TB resistance from applying the classifier trained on the data from the current study to a test set of independently collected healthy individuals [9].

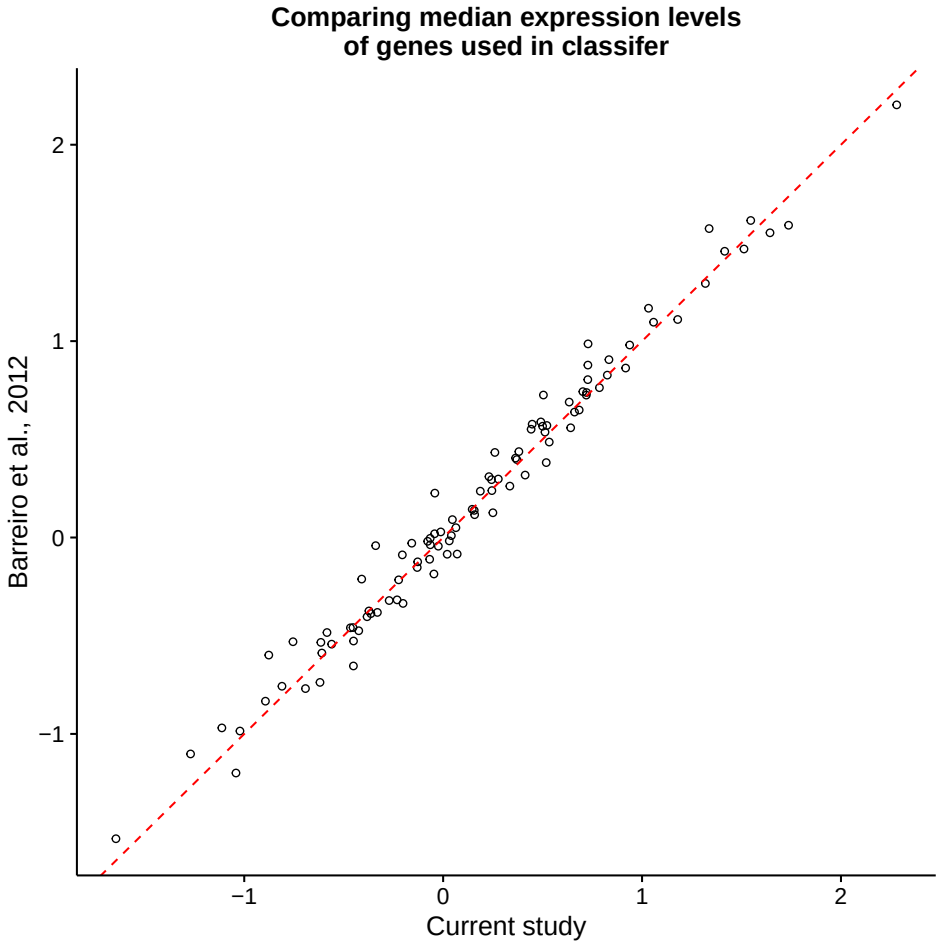


Figure 3.18: **Comparing gene expression between the two studies.** After normalization and batch-correction, the median expression levels of the 99 genes used in the classifier were similar between the samples in the current study and those in Barreiro et al., 2012 [9]. The dashed red line is the 1:1 line.

3.8.2 Supplementary Data

Table 3.1: **Sample information.** (see supplementary file associated with this dissertation) Contains information on the 50 samples. Most variables describe the batch processing steps outlined in Supplementary Fig. 3.4. "id" is a unique identifier for each sample, "individual" is the individual identifier ("s" = susceptible, "r" = resistant), "status" is the susceptibility status, "treatment" is if the sample was infected or non-infected, "infection" is the date of the infection experiment (12 total), "arrival" is the identifier for the arrival batch (4 total), "extraction" is the batch for RNA extraction (5 total), "master_mix" is the batch for library preparation (3 total), "rin" is the RNA Integrity Number from the Agilent Bioanalyzer, and "outlier" is a Boolean variable indicating if the sample was identified as an outlier (Supplementary Fig. 3.8) and removed from the analysis. (txt)

Table 3.2: **Gene expression matrix.** (see supplementary file associated with this dissertation) Contains the gene expression counts for the 11,336 genes after filtering lowly expressed genes for all 50 samples (Supplementary Fig. 3.5). Each row is a gene labeled with its Ensembl gene ID. Each column is a sample. Each sample is labeled according to the pattern "x##-status-treatment", where x is "r" for resistant or "s" for susceptible, ## is the ID number, status is "resist" for resistant or "suscep" for susceptible, and treatment is "noninf" for non-infected or "infect" for infected. (txt)

Table 3.3: **Differential expression results.** (see supplementary file associated with this dissertation) Contains the results of the differential expression analysis with limma (Fig. 3.1). The workbook contains 4 sheets corresponding to the 4 tests performed. "status_ni" is the test between resistant and susceptible individuals in the non-infected state, "status_ii" is the test between resistant and susceptible individuals in the infected state, "treat_resist" is the test between the non-infected and infected states for resistant individuals, and "treat_suscep" is the test between the non-infected and infected states for susceptible individuals. Each sheet has the same columns. "id" is the Ensembl gene ID, "gene" is the gene name, "logFC" is the log fold change from limma, "AveExpr" is the average log expression from limma, "t" is the t-statistic from limma, "P.Value" is the p-value from limma, "adj.P.Val" is the adjusted p-value from limma, "qvalue" is the q-value calculated with adaptive shrinkage, "chr" is the chromosome where the gene is located, "description" is the description of the gene from Ensembl, "phenotype" is the associated phenotype(s) assigned my Ensembl, "go_id" is the associated GO term(s) assigned by Ensembl, and "go_description" is the corresponding name(s) of the GO term(s). (xlsx)

Table 3.4: **Data for combined analysis of gene expression data and GWAS results.** (see supplementary file associated with this dissertation) Contains the results of the GWAS comparison analysis (Fig. 3.2). The first sheet "input-data" contains the data for the 10,260 genes which were assigned a SNP in the studies from The Gambia and Ghana. "gwas_p_ghana" is the minimum p-value from the GWAS in Ghana, "gwas_p_gambia" is the minimum p-value from the GWAS in The Gambia, and "n_snps" is the number of GWAS SNPs within 50 kb of the transcription start site. The columns status_ni, status_ii, treat_resist, and treat_suscep refer to the tests described for Supplementary Table 3.3 and contain the absolute log fold changes for each comparison. All the other gene annotation columns are the same as described for Supplementary Table 3.3. The second sheet "top-genes" contains the results of stringently filtering the combined differential expression and GWAS results. "GWAS P cutoff" is the p-value cutoff used for both the The Gambia and Ghana GWAS, "Effect size cutoff" is the cutoff of the absolute log fold change for the test between susceptible and resistant individuals in the non-infected state (Fig. 3.1a), "Number of genes" is the number of genes which satisfied these thresholds, and "Names" is the corresponding official gene names (sorted alphabetically). (xlsx)

Table 3.5: **Classifier results.** (see supplementary file associated with this dissertation) Contains the results of the classifier analysis. Specifically it contains the results from the support vector machine using the genes with a qvalue less than 0.05 (Fig. 3.3). The sheet "gene-list" contains information about the genes used for the classifier (the columns are described in the section for Supplementary Table 3.3). The sheet "training-input" contains the input gene expression data for training the model. The sheet "training-results" contains the results of the leave-one-out-cross-validation when training the model on the samples from the current study. The sheet "testing-input" contains the input gene expression data for testing the model. The sheet "testing-results" contains the results from testing the model on the samples from Barreiro et al., 2012 [9]. The column "prob_tb_resist" is the probability of being resistant to TB assigned by the model. (xlsx)

CHAPTER 4

BATCH EFFECTS AND THE EFFECTIVE DESIGN OF SINGLE-CELL GENE EXPRESSION STUDIES

4.1 Abstract[1]

Single-cell RNA sequencing (scRNA-seq) can be used to characterize variation in gene expression levels at high resolution. However, the sources of experimental noise in scRNA-seq are not yet well understood. We investigated the technical variation associated with sample processing using the single-cell Fluidigm C1 platform. To do so, we processed three C1 replicates from three human induced pluripotent stem cell (iPSC) lines. We added unique molecular identifiers (UMIs) to all samples, to account for amplification bias. We found that the major source of variation in the gene expression data was driven by genotype, but we also observed substantial variation between the technical replicates. We observed that the conversion of reads to molecules using the UMIs was impacted by both biological and technical variation, indicating that UMI counts are not an unbiased estimator of gene expression levels. Based on our results, we suggest a framework for effective scRNA-seq studies.

4.2 Introduction

Single-cell genomic technologies can be used to study the regulation of gene expression at unprecedented resolution [118, 156]. Using single-cell gene expression data, we can begin to effectively characterize and classify individual cell types and cell states, develop a better understanding of gene regulatory threshold effects in response to treatments or stress, and address a large number of outstanding questions that pertain to the regulation of noise and

1. Citation for chapter: Po-Yuan Tung*, John D Blischak*, Chiaowen Hsiao*, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. bioRxiv, page 062919, 2016. * denotes equal contribution. Accepted with minor revisions in Scientific Reports.

robustness of gene expression programs. Indeed, single cell gene expression data have already been used to study and provide unique insight into a wide range of research topics, including differentiation and tissue development [119, 65, 41], the innate immune response [160, 78], and pharmacogenomics [128, 96].

Yet, there are a number of outstanding challenges that arose in parallel with the application of single cell technology [172]. A fundamental difficulty, for instance, is the presence of inevitable technical variability introduced during sample processing steps, including but not limited to the conditions of mRNA capture from a single cell, amplification bias, sequencing depth, and variation in pipetting accuracy. These (and other sources of error) may not be unique to single cell technologies, but in the context of studies where each sample corresponds to a single cell, and is thus processed as a single unrepeatable batch, these technical considerations make the analysis of biological variability across single cells particularly challenging.

To better account for technical variability in scRNA-seq experiments, it has become common to add spike-in RNA standards of known abundance to the endogenous samples [19, 63]. The most commonly used spike-in was developed by the External RNA Controls Consortium (ERCC) [80]; comprising of a set of 96 RNA controls of varying length and GC content. A number of single cell studies focusing on analyzing technical variability based on ERCC spike-in controls have been reported [19, 63, 39, 194]. However, one principle problem with spike-ins is that they do not 'experience' all processing steps that the endogenous sample is subjected to. For that reason, it is unknown to what extent the spike-ins can faithfully reflect the error that is being accumulated during the entire sample processing procedure, either within or across batches. In particular, amplification bias, which is assumed to be gene-specific, cannot be addressed by spike-in normalization approaches.

To address challenges related to the efficiency and uniformity with which mRNA molecules are amplified and sequenced in single cells, unique molecule identifiers (UMIs) were intro-

duced to single cell sample processing [98, 56, 25, 163]. The rationale is that by counting molecules rather than the number of amplified sequencing reads, one can account for biases related to amplification, and obtain more accurate estimates of gene expression levels [78, 77, 63]. It is assumed that most sources of variation in single cell gene expression studies can be accounted for by using the combination of UMIs and a spike-in based standardization [77, 194]. Nevertheless, though molecule counts, as opposed to sequencing read counts, are associated with substantially reduced levels of technical variability, a non-negligible proportion of experimental error remains unexplained.

There are a few common platforms in use for scRNA-seq. The automated C1 microfluidic platform (Fluidigm), while more expensive per sample, has been shown to confer several advantages over platforms that make use of droplets to capture single cells [204, 119]. In particular, smaller samples can be processed using the C1 (when cell numbers are limiting), and the C1 capture efficiency of genes (and RNA molecules) is markedly higher. Notably, in the context of this study, the C1 system also allows for direct confirmation of single cell capture events, in contrast to most other microfluidic-based approaches [119, 99]. One of the biggest limitations of using the C1 system, however, is that single cell capture and preparation from different conditions are fully independent [70]. Consequently, multiple replicates of C1 collections from the same biological condition are necessary to facilitate estimation of technical variability even with the presence of ERCC spike-in controls [172]. To our knowledge, to date, no study has been purposely conducted to assess the technical variability across batches on the C1 platform.

To address this gap, we collected scRNA-seq data from induced pluripotent stem cell (iPSC) lines of three Yoruba individuals (abbreviation: YRI) using C1 microfluidic plates. Specifically, we performed three independent C1 collections per each individual to disentangle batch effects from the biological covariate of interest, which, in this case, is the difference between individuals. Both ERCC spike-in controls and UMIs were included in our sample

processing. With these data, we were able to elucidate technical variability both within and between C1 batches and thus provide a deep characterization of cell-to-cell variation in gene expression levels across individuals.

4.3 Results

4.3.1 Study design and quality control

We collected single cell RNA-seq (scRNA-seq) data from three YRI iPSC lines using the Fluidigm C1 microfluidic system followed by sequencing. We added ERCC spike-in controls to each sample, and used 5-bp random sequence UMIs to allow for the direct quantification of mRNA molecule numbers. For each of the YRI lines, we performed three independent C1 collections; each replicate was accompanied by processing of a matching bulk sample using the same reagents. This study design (Fig. 4.1A and Supplementary Table 4.1) allows us to estimate error and variability associated with the technical processing of the samples, independently from the biological variation across single cells of different individuals. We were also able to estimate how well scRNA-seq data can recapitulate the RNA-seq results from population bulk samples.

In what follows, we describe data as originating from different samples when we refer to data from distinct wells of each C1 collection. Generally, each sample corresponds to a single cell. In turn, we describe data as originating from different replicates when we refer to all samples from a given C1 collection, and from different individuals when we refer to data from all samples and replicates of a given genetically distinct iPSC line.

We obtained an average of 6.3 +/- 2.1 million sequencing reads per sample (range 0.4-11.2 million reads). We processed the sequencing reads using a standard alignment approach (see Methods) and performed multiple quality control analyses. As a first step, we estimated the proportion of ERCC spike-in reads from each sample. We found that, across samples,

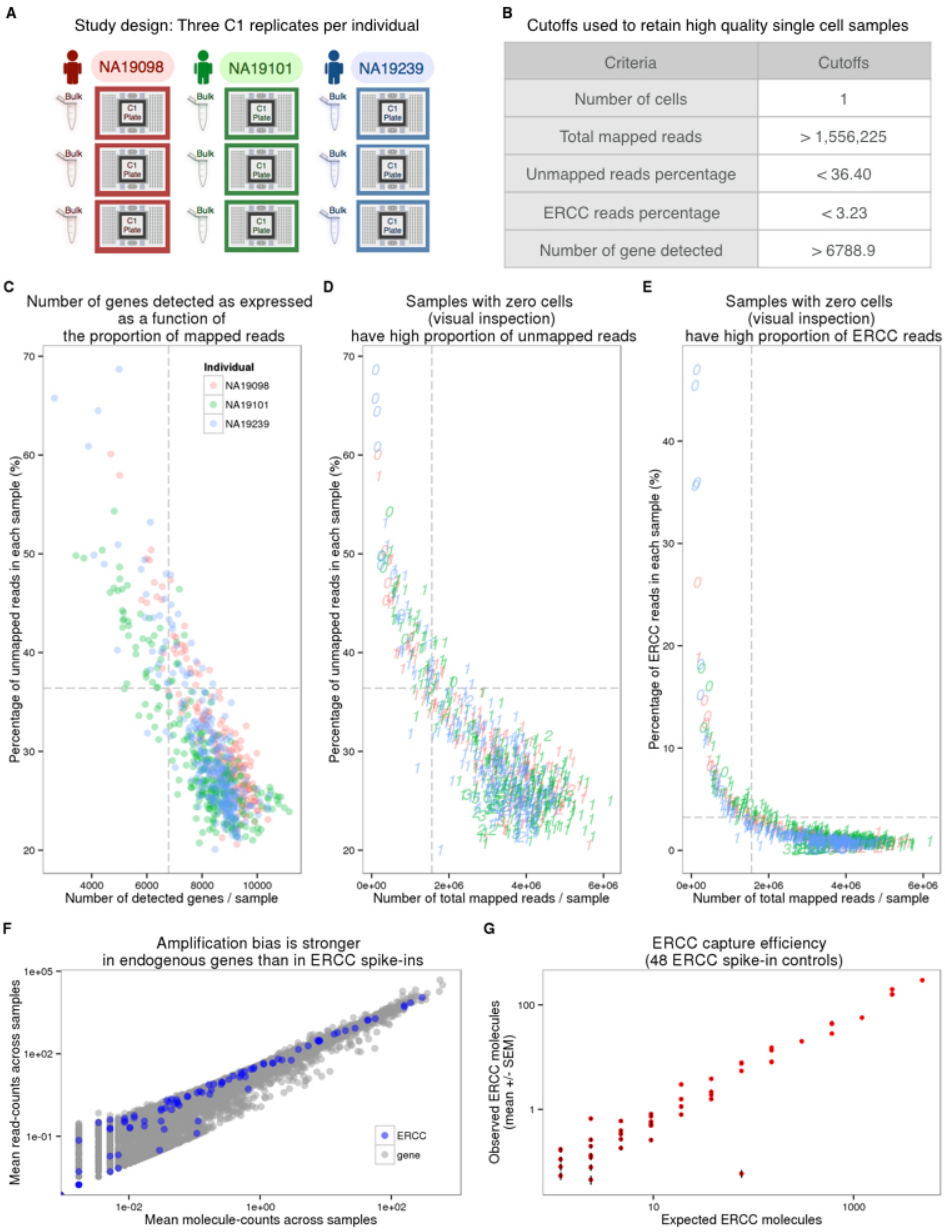


Figure 4.1: **Experimental design and quality control of scRNA-seq.** (A) Three C1 96 well-integrated fluidic circuit (IFC) replicates were collected from each of the three Yoruba individuals. A bulk sample was included in each batch. (B) Summary of the cutoffs used to remove data from low quality cells that might be ruptured or dead (See Supplementary Fig. 4.6 for details). (C-E) To assess the quality of the scRNA-seq data, the capture efficiency of cells and the faithfulness of mRNA fraction amplification were determined based on the proportion of unmapped reads, the number of detected genes, the numbers of total mapped reads, and the proportion of ERCC spike-in reads across cells. The dash lines indicate the cutoffs summarized in panel (B). The three colors represent the three individuals (NA19098 in red, NA19101 in green, and NA19239 in blue), and the numbers indicate the cell numbers observed in each capture site on C1 plate.

Figure 4.1: (continued) (F) Scatterplots in log scale showing the mean read counts and the mean molecule counts of each endogenous gene (grey) and ERCC spike-ins (blue) from the 564 high quality single cell samples before removal of genes with low expression. (G) mRNA capture efficiency shown as observed molecule count versus number of molecules added to each sample, only including the 48 ERCC spike-in controls remaining after removal of genes with low abundance. Each red dot represents the mean +/- SEM of an ERCC spike-in across the 564 high quality single cell samples.

sequencing reads from practically all samples of the second replicate of individual NA19098 included unusually high ERCC content compared to all other samples and replicates (Supplementary Fig. 4.6). We concluded that a pipetting error led to excess ERCC content in this replicate and we excluded the data from all samples of this replicate in subsequent analyses. With the exception of the excluded samples, data from all other replicates seem to have similar global properties (using general metrics; Fig. 4.1C-E and Supplementary Fig. 4.6).

We next examined the assumption that data from each sample correspond to data from a single cell. After the cell sorting was complete, but before the processing of the samples, we performed visual inspection of the C1 microfluidic plates. Based on that visual inspection, we flagged 21 samples that did not contain any cell, and 54 samples that contained more than one cell (across all batches). Visual inspection of the C1 microfluidic plate is an important quality control step, but it is not infallible. We therefore filtered data from the remaining samples based on the number of total mapped reads, the percentage of unmapped reads, the percentage of ERCC spike-in reads, and the number of genes detected (Fig. 4.1B-E). We chose data-driven inclusion cutoffs for each metric, based on the 95th percentile of the respective distributions for the 21 libraries that were amplified from samples that did not include a cell based on visual inspection (Supplementary Fig. 4.6). Using this approach, we identified and removed data from 15 additional samples that were classified as originating from a single cell based on visual inspection, but whose data were more consistent with a multiple-cell origin based on the number of total molecules, the concentration of

cDNA amplicons, and the read-to-molecule conversion efficiency (defined as the number of total molecules divided by the number of total reads; Supplementary Fig. 4.7). At the conclusion of these quality control analyses and exclusion steps, we retained data from 564 high quality samples, which correspond, with reasonable confidence, to 564 single cells, across eight replicates from three individuals (Supplementary Table 4.2).

Our final quality check focused on the different properties of sequencing read and molecule count data. We considered data from the 564 high quality samples and compared gene specific counts of sequencing read and molecules. We found that while gene-specific reads and molecule counts are exceptionally highly correlated when we considered the ERCC spike-in data (r = 0.99; Fig. 4.1F), these counts are somewhat less correlated when data from the endogenous genes are considered (r = 0.92). Moreover, the gene-specific read and molecule counts correlation is noticeably lower for genes that are expressed at lower levels (Fig. 1F). These observations concur with previous studies [77, 63] as they underscore the importance of using UMIs in single cell gene expression studies.

We proceeded by investigating the effect of sequencing depth and the number of single cells collected on multiple properties of the data. To this end, we repeatedly subsampled single cells and sequencing reads to assess the correlation of the single cell gene expression estimates to the bulk samples, the number of genes detected, and the correlation of the cell-to-cell gene expression variance estimates between the reduced subsampled data and the full single cell gene expression data set (Fig. 2). We observed quickly diminishing improvement in all three properties with increasing sequencing depth and the number of sampled cells, especially for highly expressed genes. For example, a per cell sequencing depth of 1.5 million reads (which corresponds to ~50,000 molecules) from each of 75 single cells was sufficient for effectively quantifying even the lower 50% of expressed genes. To be precise, at this level of subsampling for individual NA19239, we were able to detect a mean of 6068 genes out of 6097 genes expressed in the bulk samples (the bottom 50%; Fig. 4.2B); the estimated single

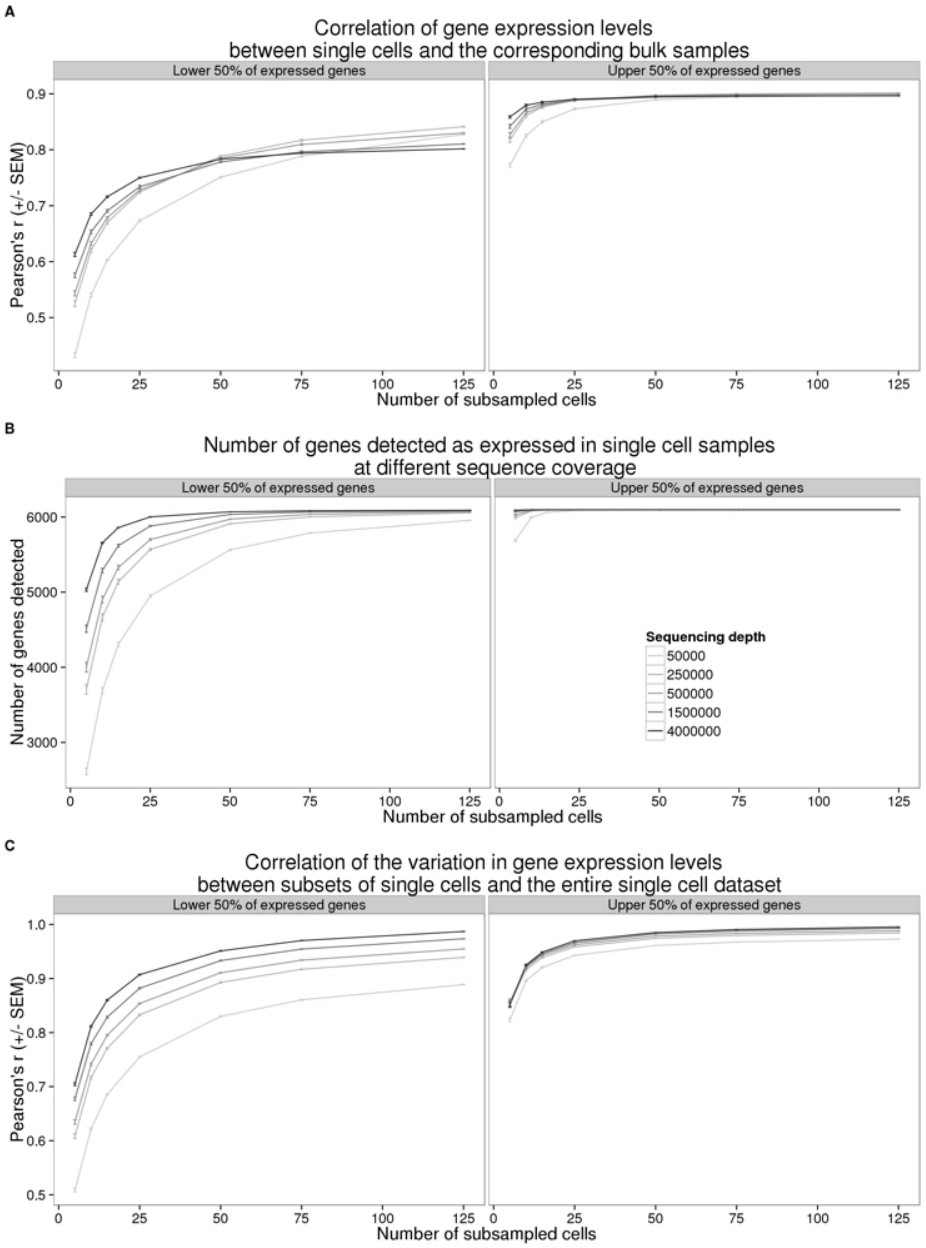


Figure 4.2: **The effect of sequencing depth and cell number on single cell UMI estimates.** Sequencing reads from the entire data set were subsampled to the indicated sequencing depth and cell number, and subsequently converted to molecules using the UMIs. Each point represents the mean +/- SEM of 10 random draws of the indicated cell number. The left panel displays the results for 6,097 (50% of detected) genes with lower expression levels and the right panel the results for 6,097 genes with higher expression levels. (A) Pearson correlation of aggregated gene expression level estimates from single cells compared to the bulk sequencing samples. (B) Total number of genes detected with at least one molecule in at least one of the single cells. (C) Pearson correlation of cell-to-cell gene expression variance estimates from subsets of single cells compared to the full single cell data set.

cell expression levels of these genes (summed across all cells) correlated with the bulk sample gene expression levels with a mean Pearson coefficient of 0.8 (Fig. 4.2A), and the estimated cell-to-cell variation in gene expression levels was correlated with the variation estimated from the full data set with a mean Pearson coefficient of 0.95 (Fig. 4.2C).

4.3.2 Batch effects associated with UMI-based single cell data

In the context of the C1 platform, typical study designs make use of a single C1 plate (batch/replicate) per biological condition. In that case, it is impossible to distinguish between biological and technical effects associated with the independent capturing and sequencing of each C1 replicate. We designed our study with multiple technical replicates per biological condition (individual) in order to directly and explicitly estimate the batch effect associated with independent C1 preparations (Fig. 4.1A).

As a first step in exploring batch effects, we examined the gene expression profiles across all single cells that passed our quality checks (as reported above) using raw molecule counts (without standardization). Using principal component analysis (PCA) for visualization, we observed – as expected - that the major source of variation in data from single cells is the individual origin of the sample (Fig. 4.4A). Specifically, we found that the proportion of variance due to individual was larger (median: 8%) than variance due to C1 batch (median: 4%; Kruskal-Wallis test; $P < 0.001$, Supplementary Fig. 4.8; see Methods for details of the variance component analysis). Yet, variation due to C1 batch is also substantial - data from single cell samples within a batch are more correlated than that from single cells from the same individual but different batches (Kruskal-Wallis test; $P < 0.001$).

Could we account for the observed batch effects using the ERCC spike-in controls? In theory, if the total ERCC molecule-counts are affected only by technical variability, the spike-ins could be used to correct for batch effects even in a study design that entirely confounds biological samples with C1 preparations. To examine this, we first considered the

relationship between total ERCC molecule-counts and total endogenous molecule-counts per sample. If only technical variability affects ERCC molecule-counts, we expect the technical variation in the spike-ins (namely, variation between C1 batches) to be consistent, regardless of the individual assignment. Indeed, we observed that total ERCC molecule-counts are significantly different between C1 batches (F-test; $P < 0.001$). However, total ERCC molecule-counts are also quite different across individuals, when variation between batches is taken into account (LRT; $P = 0.08$; Fig. 4.3A). This observation suggests that both technical and biological variation affect total ERCC molecule-counts. In addition, while we observed a positive relationship between total ERCC molecule-counts and total endogenous molecule-counts per sample, this correlation pattern differed across C1 batches and across individuals (F-test; $P < 0.001$; Fig. 4.3B).

To more carefully examine the technical and biological variation of ERCC spike-in controls, we assessed the ERCC per-gene expression profile. We observed that the ERCC gene expression data from samples of the same batch were more correlated than data from samples across batches (Kruskal-Wallis test; Chi-squared $P < 0.001$). However, the proportion of variance explained by the individual was significantly larger than the variance due to C1 batch (median: 9% vs. 5%, Chi-squared test; $P < 0.001$, Supplementary Fig. 4.8), lending further support to the notion that biological variation affects the ERCC spike in data. Based on these analyses, we concluded that ERCC spike-in controls cannot be used to effectively account for the batch effect associated with independent C1 preparations.

We explored potential reasons for the observed batch effects, and in particular, the difference in ERCC counts across batches and individuals. We focused on the read-to-molecule conversion rates, i.e. the rates at which sequencing reads are converted to molecule counts based on the UMI sequences. We defined read-to-molecule conversion efficiency as the total molecule-counts divided by the total reads-counts in each sample, considering separately the reads/molecules that correspond to endogenous genes or ERCC spike-ins (Fig. 4.3C and

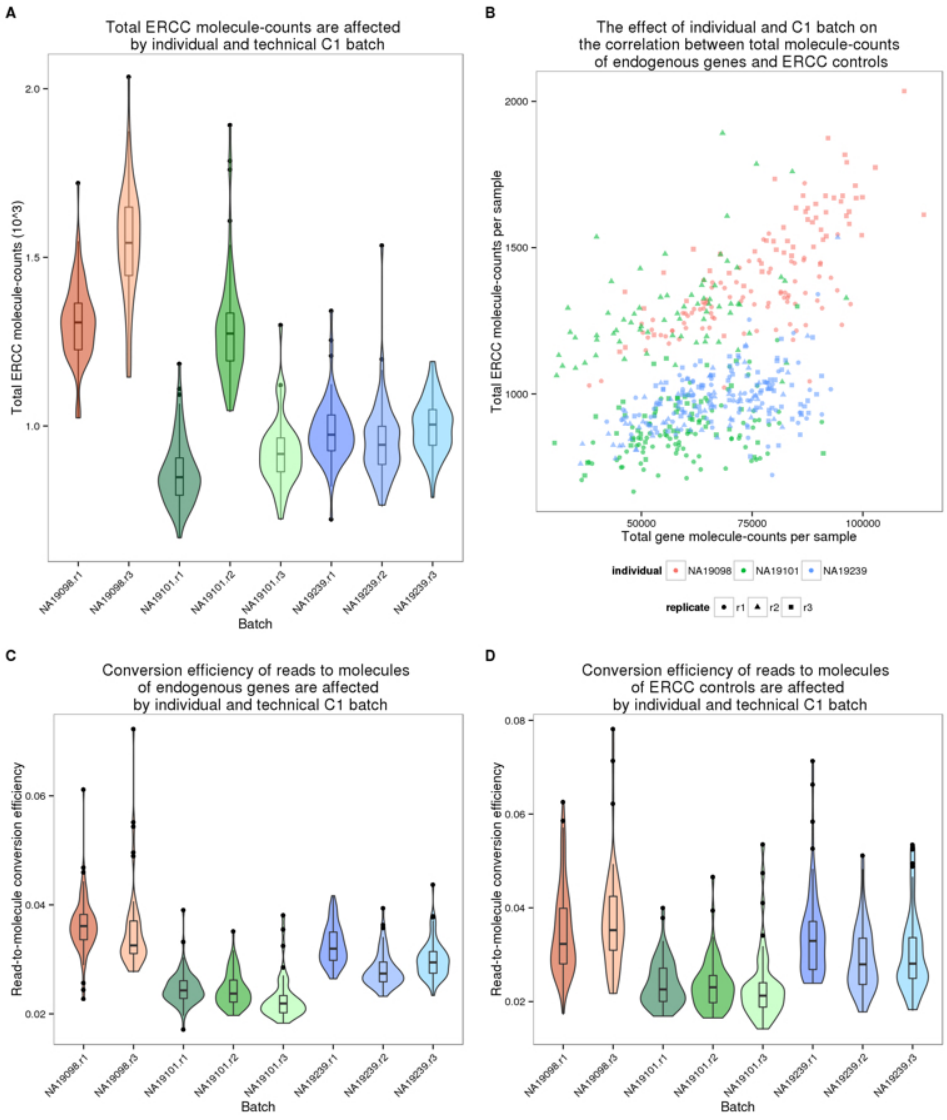


Figure 4.3: **Batch effect of scRNA-seq data using the C1 platform.** (A) Violin plots of the number of total ERCC spike-in molecule-counts in single cell samples per C1 replicate. (B) Scatterplot of the total ERCC molecule-counts and total gene molecule-counts. The colors represent the three individuals (NA19098 is in red, NA19101 in green, and NA19239 in blue). Data from different C1 replicates is plotted in different shapes. (C and D) Violin plots of the reads to molecule conversion efficiency (total molecule-counts divided by total read-counts per single cells) by C1 replicate. The endogenous genes and the ERCC spike-ins are shown separately in (C) and (D), respectively. There is significant difference across individuals of both endogenous genes ($P < 0.001$) and ERCC spike-ins ($P < 0.05$). The differences across C1 replicates per individual of endogenous genes and ERCC spike-ins were also evaluated (both $P < 0.01$).

4.3D). We observed a significant batch effect in the read-to-molecule conversion efficiency of both ERCC (F-test; $P < 0.05$) and endogenous genes (F-test; $P < 0.001$) across C1 replicates from the same individual. Moreover, the difference in read-to-molecule conversion efficiency across the three individuals was significant not only for endogenous genes (LRT; $P < 0.01$, Fig. 4.3C) but also in the ERCC spike-ins (LRT; $P < 0.01$, Fig. 4.3D). We reason that the difference in read to molecule conversion efficiency across C1 preparations may contribute to the observed batch effect in this platform.

4.3.3 Measuring regulatory noise in single-cell gene expression data

Our analysis indicated that there is a considerable batch effect in the single cell gene expression data collected from the C1 platform. We thus sought an approach that would account for the batch effect and allow us to study biological properties of the single-cell molecule count-based estimates of gene expression levels, albeit in a small sample of just three individuals. As a first step, we adjusted the raw molecule counts by using a Poisson approximation to account for the random use of identical UMI sequences in molecules from highly expressed genes (this was previously termed a correction for the UMI 'collision probability' [56]). We then excluded data from genes whose inferred molecule count exceeded 1,024 (the theoretical number of UMI sequences) – this step resulted in the exclusion of data from 6 mitochondrial genes.

We next incorporated a standardization step by computing log transformed counts-per-million (cpm) to remove the effect of different sequencing depths, as is the common practice for the analysis of bulk RNA-seq data (Fig. 4.4A and 4.4B). We used a Poisson generalized linear model to normalize the endogenous molecule $\log_2$ cpm values by the observed molecule counts of ERCC spike-ins across samples. While we do not expect this step to account for the batch effect (as discussed above), we reasoned that the spike-ins allow us to account for a subset of technical differences between samples, for example, those that arise from differences

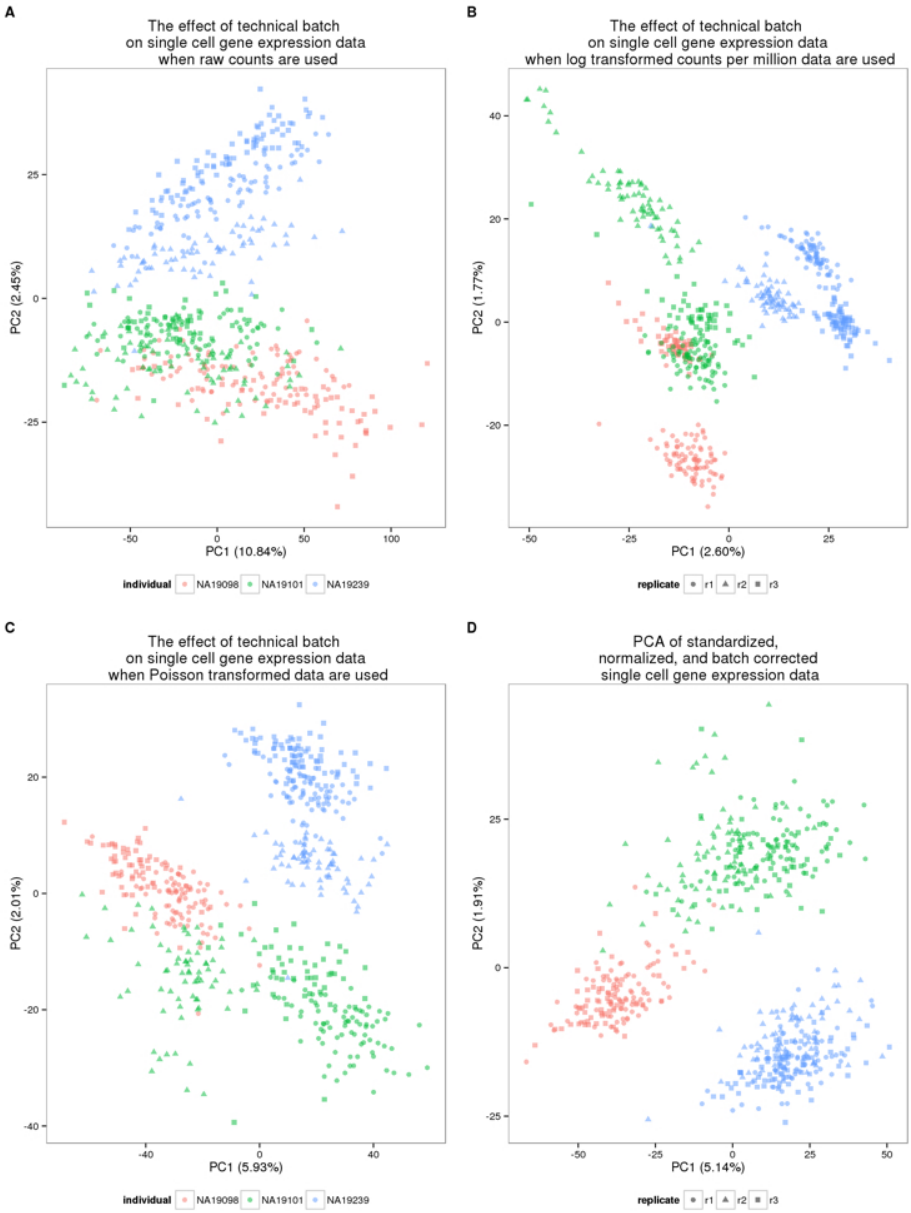


Figure 4.4: **Normalization and removal of technical variability.** Principal component (PC) 1 versus PC2 of the (A) raw molecule counts, (B) $\log_2$ counts per million (cpm), (C) Poisson transformed expression levels (accounting for technical variability modeled by the ERCC spike-ins), and (D) batch-corrected expression levels. The colors represent the three individuals (NA19098 in red, NA19101 in green, and NA19239 in blue). Data from different C1 replicates is plotted in different shapes.

in RNA concentration (Fig. 4.4C).

Finally, to account for the technical batch effect, we modeled between-sample correlations in gene expression within C1 replicates (see Methods). Our approach is similar in principle to limma, which was initially developed for adjusting within-replicate correlations in microarray data [168]. We assume that samples within each C1 replicate share a component of technical variation, which is independent of biological variation across individuals. We fit a linear mixed model for each gene, which includes a fixed effect for individual and a random effect for batch. The batch effect is specific to each C1 replicate, and is independent of biological variation across individuals. We use this approach to estimate and remove the batch effect associated with different C1 preparations (Fig. 4.4D).

Once we removed the unwanted technical variability, we focused on analyzing biological variation in gene expression between single cells. Our goal was to identify inter-individual differences in the amount of variation in gene expression levels across single cells, or in other words, to identify differences between individuals in the amount of regulatory noise [151]. In this context, regulatory noise is generally defined as the coefficient of variation (CV) of the gene expression levels of single cells [49]. In the following, we used the standardized, normalized, batch-corrected molecule count gene expression data to estimate regulatory noise (Fig. 4.4D). To account for heteroscedasticity from Poisson sampling, we adjusted the CV values by the average gene-specific expression level across cells of the same individual. The adjusted CV is robust both to differences in gene expression levels, as well as to the proportion of gene dropouts in single cells.

To investigate the effects of gene dropouts (the lack of molecule representation of an expressed gene [19, 160]) on our estimates of gene expression noise, we considered the association between the proportion of cells in which a given gene is undetected (namely, the gene-specific dropout rate), the average gene expression level, and estimates of gene expression noise. Across all genes, the median gene-specific dropout was 22 percent. We found

significant individual differences (LRT; $P < 10^{-5}$) in gene-specific dropout rates between individuals in more than 10% (1,214 of 13,058) of expressed endogenous genes. As expected, the expression levels, and the estimated variation in expression levels across cells, are both associated with gene-specific dropout rates (Supplementary Fig. 4.9). However, importantly, adjusted CVs are not associated with dropout rates (Spearman's correlation = 0.04; Supplementary Fig. 4.9), indicating that adjusted CV measurements are not confounded by the dynamic range of single-cell gene expression levels.

We thus estimated mean expression levels and regulatory noise (using adjusted CV) for each gene, by either including (Fig. 4.5A) or excluding (Fig. 4.5B) samples in which the gene was not detected/expressed. We first focused on general trends in the data. We ranked genes in each individual by their mean expression level as well as by their estimated level of variation across single cells. When we considered samples in which a gene was expressed, we found that 887 of the 1,000 most highly expressed genes in each individual are common to all three individuals (Fig. 4.5C). In contrast, only 103 of the 1,000 most highly variable (noisy) genes in each individual were common to all three individuals (Fig. 4.5D). We found similar results when we considered data from all single cells, regardless of whether the gene was detected as expressed (Fig. 4.5E and 4.5F).

Next, we identified genes whose estimated regulatory noise (based on the adjusted CV) is significantly different between individuals. For the purpose of this analysis, we only included data from cells in which the gene was detected as expressed. Based on permutations (Supplementary Fig. 4.10), we classified the estimates of regulatory noise of 560 genes as significantly different across individuals (empirical $P < .0001$, Supplementary Fig. 4.11 for examples; Supplementary Table 4.3 for gene list). These 560 genes are enriched for genes involved in protein translation, protein disassembly, and various biosynthetic processes (Supplementary Table 4.4). Interestingly, among the genes whose regulatory noise estimates differ between individuals, we found two pluripotency genes, *KLF4* and *DPPA2* (Supplementary

Fig. 4.12).

4.4 Discussion

4.4.1 Study design and sample size for scRNA-seq

Our nested study design allowed us to explicitly estimate technical batch effects associated with single cell sample processing on the C1 platform. We found previously unreported technical sources of variation associated with the C1 sample processing and the use of UMIs, including the property of batch-specific read-to-molecule conversion efficiency. As we used a well-replicated nested study design, we were able to model, estimate, and account for the batch while maintaining individual differences in gene expression levels. We believe that our observations indicate that future studies should avoid confounding C1 batch and individual source of single cell samples. Instead, we recommend a balanced study design consisting of multiple individuals within a C1 plate and multiple C1 replicates (for example, Supplementary Fig. 4.13). The origin of each cell can then be identified using the RNA sequencing data. Indeed, using a method originally developed for detecting sample swaps in DNA sequencing experiments [83], we were able to correctly identify the correct YRI individual of origin for all the single cells from the current experiment by comparing the polymorphisms identified using the RNA-seq reads to the known genotypes for all 120 YRI individuals of the International HapMap Project [184] (Supplementary Fig. 4.13). The mixed-individual-plate is an attractive study design because it allows one to account for the batch effect without the requirement to explicitly spend additional resources on purely technical replication (because the total number of cells assayed from each individual can be equal to a design in which one individual is being processed in using a single C1 plate).

We also addressed additional study design properties with respect to the desired number of single cells and the desired depth of sequencing (Fig. 2). Similar assessments have been

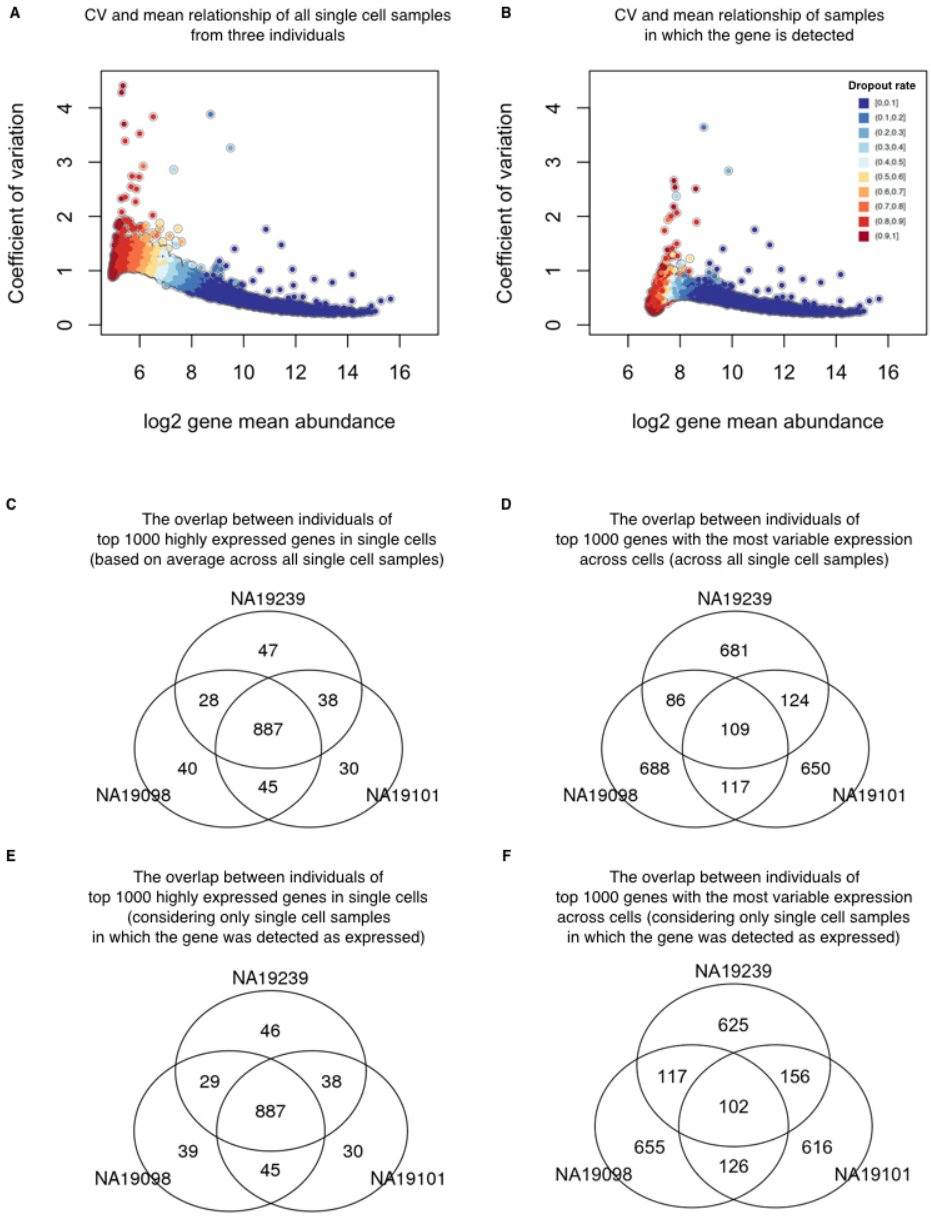


Figure 4.5: **Cell-to-cell variation in gene expression.** Adjusted CV plotted against average molecule counts across all cells in (A) and across only the cells in which the gene is expressed (B), including data from all three individuals. Each dot represents a gene, and the color indicates the corresponding gene-specific dropout rate (the proportion of cells in which the gene is undetected). (C and D) Venn diagrams showing the overlaps of top 1000 genes across individuals based on mean expression level in (C) and based on adjusted CV values in (D), considering only the cells in which the gene is expressed. (E and F) Similarly, Venn diagrams showing the overlaps of top 1000 genes across individuals based on mean expression level in (E) and based on adjusted CV values in (F), across all cells.

previously performed for single cell sequencing with the C1 platform without the use of UMIs [204, 145], but no previous study has investigated the effects of these parameters for single cells studies using UMIs. We focused on recapitulating the gene expression levels observed in bulk sequencing experiments, detecting as many genes as possible, and accurately measuring the cell-to-cell variation in gene expression levels. We recommend sequencing at least 75 high quality cells per biological condition with a minimum of 1.5 million raw reads per cell to obtain optimal performance of these three metrics.

4.4.2 The limitations of the ERCC spike-in controls

The ERCC spike-in controls have been used in previous scRNA-seq studies to identify low quality single cell samples, infer the absolute total number of molecules per cell, and model the technical variability across cells [19, 63, 39, 194]. In our experience, the ERCC controls are not particularly well-suited for any one of these tasks, much less all three. With respect to identifying low quality samples, we indeed observed that samples with no visible cell had a higher percentage of reads mapping to the ERCC controls, as expected. However, there was no clear difference between low and high quality samples in the percentage of ERCC reads or molecules, and thus any arbitrarily chosen cutoff would be associated with considerable error (Fig. 4.1E). With respect to inferring the absolute total number of molecules per cell, we observed that the biological covariate of interest (difference between the three YRI individuals), rather than batch, explained a large proportion of the variance in the ERCC counts (Supplementary Fig. 4.8), and furthermore that the ERCC controls were also affected by the individual-specific effect on the read-to-molecule conversion rate (Fig. 4.3D). Thus ERCC-based corrected estimates of total number of molecules per cell, across technical or biological replicates, are expected to be biased. Because the batch effects associated with the ERCC controls are driven by the biological covariate of interest, they will also impede the modeling of the technical variation in single cell experiments that confound batch and

the biological source of the single cells.

More generally, it is inherently difficult to model unknown sources of technical variation using so few genes [152] (only approximately half of the 92 ERCC controls are detected in typical single cell experiments), and the ERCC controls are also strongly impacted by technical sources of variation even in bulk RNA-seq experiments [158]. Lastly, from a theoretical perspective, the ERCC controls have shorter polyA tails and are overall shorter than mammalian mRNAs. For these reasons, we caution against the reliance of ERCC controls in scRNA-seq studies and highlight that an alternative set of controls that more faithfully mimics mammalian mRNAs and provides more detectable spike-in genes is desired. Our recommendation is to include total RNA from a distant species, for example using RNA from *Drosophila melanogaster* in studies of single cells from humans.

4.4.3 Outlook

Single cell experiments are ideally suited to study gene regulatory noise and robustness [17, 50]. Yet, in order to study the biological noise in gene expression levels, it is imperative that one should be able to effectively estimate and account for the technical noise in single cell gene expression data. Our results indicate that previous single cells gene expression studies may not have been able to distinguish between the technical and the biological components of variation, because single cell samples from each biological condition were processed on a single C1 batch. When technical noise is properly accounted for, even in this small pilot study, our findings indicate pervasive inter-individual differences in gene regulatory noise, independently of the overall gene expression level.

4.5 Methods

4.5.1 Ethics statement

The YRI cell lines were purchased from CCR. The original samples were collected by the HapMap project between 2001-2005. All of the samples were collected with extensive community engagement, including discussions with members of the donor communities about the ethical and social implications of human genetic variation research. Donors gave broad consent to future uses of the samples, including their use for extensive genotyping and sequencing, gene expression and proteomics studies, and all other types of genetic variation research, with the data publicly released.

4.5.2 Cell culture of iPSCs

Undifferentiated feeder-free iPSCs reprogrammed from LCLs of Yoruba individuals in Ibadan, Nigeria (abbreviation: YRI) [184] were grown in E8 medium (Life Technologies) [28] on Matrigel-coated tissue culture plates with daily media feeding at 37 C with 5% (vol/vol) CO2. For standard maintenance, cells were split every 3-4 days using cell release solution (0.5 mM EDTA and NaCl in PBS) at the confluence of roughly 80%. For the single cell suspension, iPSCs were individualized by Accutase Cell Detachment Solution (BD) for 5-7 minutes at 37 C and washed twice with E8 media immediately before each experiment. Cell viability and cell counts were then measured by the Automated Cell Counter (Bio-Rad) to generate resuspension densities of 2.5 X 105 cells/mL in E8 medium for C1 cell capture.

4.5.3 Single cell capture and library preparation

Single cell loading and capture were performed following the Fluidigm protocol (PN 100-7168). Briefly, 30 μl of C1 Suspension Reagent was added to a 70-μl aliquot of ~17,500 cells. Five μl of this cell mix were loaded onto 10-17 μm C1 Single-Cell Auto Prep IFC

microfluidic chip (Fluidigm), and the chip was then processed on a C1 instrument using the cell-loading script according to the manufacturer's instructions. Using the standard staining script, the iPSCs were stained with StainAlive TRA-1-60 Antibody (Stemgent, PN 09-0068). The capture efficiency and TRA-1-60 staining were then inspected using the EVOS FL Cell Imaging System (Thermo Fisher) (Supplementary Table 4.1).

Immediately after imaging, reverse transcription and cDNA amplification were performed in the C1 system using the SMARTer PCR cDNA Synthesis kit (Clontech) and the Advantage 2 PCR kit (Clontech) according to the instructions in the Fluidigm user manual with minor changes to incorporate UMI labeling [77]. Specifically, the reverse transcription primer and the 1:50,000 Ambion ERCC Spike-In Mix1 (Life Technologies) were added to the lysis buffer, and the template-switching RNA oligos which contain the UMI (5-bp random sequence) were included in the reverse transcription mix [75, 76, 77]. When the run finished, full-length, amplified, single-cell cDNA libraries were harvested in a total of approximately 13 μl C1 Harvesting Reagent and quantified using the DNA High Sensitivity LabChip (Caliper). The average yield of samples per C1 plate ranged from 1.26-1.88 ng per microliter (Supplementary Table 4.1). A bulk sample, a 40 μl aliquot of ~10,000 cells, was collected in parallel with each C1 chip using the same reaction mixes following the C1 protocol (PN 100-7168, Appendix A).

For sequencing library preparation, tagmentation and isolation of 5' fragments were performed according to the UMI protocol [77]. Instead of using commercially available Tn5 transposase, Tn5 protein stock was freshly purified in house using the IMPACT system (pTXB1, NEB) following the protocol previously described [142]. The activity of Tn5 was tested and shown to be comparable with the EZ-Tn5-Transposase (Epicentre). Importantly, all the libraries in this study were generated using the same batch of Tn5 protein purification. For each of the bulk samples, two libraries were generated using two different indices in order to get sufficient material for sequencing. All 18 bulk libraries were then pooled and

labeled as the "bulk" for sequencing.

4.5.4 Illumina high-throughput sequencing

The scRNA-seq libraries generated from the 96 single cell samples of each C1 chip were pooled and then sequenced in three lanes on an Illumina HiSeq 2500 instrument using the PCR primer (C1-P1-PCR-2: Bio-GAATGATACGGCGACCACCGAT) as the read 1 primer and the Tn5 adapter (C1-Tn5-U: PHO-CTGTCTCTTATACACATCTGACGC) as the index read primer following the UMI protocol [77].

The master mixes, one mix with all the bulk samples and nine mixes corresponding to the three replicates for the three individuals, were sequenced across four flowcells using a design aimed to minimize the introduction of technical batch effects (Supplementary Table 4.1). Single-end 100 bp reads were generated along with 8-bp index reads corresponding to the cell-specific barcodes. We did not observe any obvious technical effects due to sequencing lane or flow cell that confounded the inter-individual and inter-replicate comparisons.

4.5.5 Read mapping

To assess read quality, we ran FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`) and observed a decrease in base quality at the 3' end of the reads. Thus we removed low quality bases from the 3' end using sickle with default settings [82]. To handle the UMI sequences at the 5' end of each read, we used umitools [81] to find all reads with a UMI of the pattern NNNNNGGG (reads without UMIs were discarded). We then mapped reads to human genome hg19 (only including chromosomes 1-22, X, and Y, plus the ERCC sequences) with Subjunc [110], discarding non-uniquely mapped reads (option -u). To obtain gene-level counts, we assigned reads to protein-coding genes (Ensembl GRCh37 release 82) and the ERCC spike-in genes using featureCounts [111]. Because the UMI protocol maintains strand information, we required that reads map to a gene in the correct orientation

(featureCounts flag -s 1).

In addition to read counts, we utilized the UMI information to obtain molecule counts for the single cell samples. We did not count molecules for the bulk samples because this would violate the assumptions of the UMI protocol, as bulk samples contain far too many unique molecules for the 1,024 UMIs to properly tag them all. First, we combined all reads for a given single cell using samtools [108]. Next, we converted read counts to molecule counts using UMI-tools [165]. UMI-tools counts the number of UMIs at each read start position. Furthermore, it accounts for sequencing errors in the UMIs introduced during the PCR amplification or sequencing steps using a "directional adjacency" method. Briefly, all UMIs at a given read start position are connected in a network using an edit distance of one base pair. However, edges between nodes (the UMIs) are only formed if the nodes have less than a 2x difference in reads. The node with the highest number of reads is counted as a unique molecule, and then it and all connected nodes are removed from the network. This is repeated until all nodes have been counted or removed.

4.5.6 Filtering cells and genes

We performed multiple quality control analyses to detect and remove data from low quality cells. In an initial analysis investigating the percentage of reads mapping to the ERCC spike-in controls, we observed that replicate 2 of individual NA19098 was a clear outlier (Supplementary Fig. 4.6). It appeared that too much ERCC spike-in mix was added to this batch, which violated the assumption that the same amount of ERCC molecules was added to each cell. Thus, we removed this batch from all of our analyses.

Next, we kept data from high quality single cells that passed the following criteria:

- Only one cell observed per well
- At least 1,556,255 mapped reads
- Less than 36.4% unmapped reads

- Less than 3.2% ERCC reads
- More than 6,788 genes with at least one read

We chose the above criteria based on the distribution of these metrics in the empty wells (the cutoff is the 95th percentile, Supplementary Fig. 4.6). In addition, we observed that some wells classified as containing only one cell were clustered with multi-cell wells when plotting 1) the number of gene molecules versus the concentration of the samples, and 2) the read to molecule conversion efficiency (total molecule number divided by total read number) of endogenous genes versus that of ERCC. We therefore established filtering criteria for these misidentified single-cell wells using linear discriminant analysis (LDA). Specifically, LDA was performed to classify wells into empty, one-cell, and two-cell using the discriminant functions of 1) sample concentration and the number of gene molecules, and 2) endogenous and ERCC gene read to molecule conversion efficiency (Supplementary Fig. 4.7). After filtering, we maintained 564 high quality single cells (NA19098: 142, NA19101: 201, NA19239: 221).

The quality control analyses were performed using all protein-coding genes (Ensembl GRCh37 release 82) with at least one observed read. Using the high quality single cells, we further removed genes with low expression levels for downstream analyses. We removed all genes with a mean $\log_2$ cpm less than 2, which did not affect the relative differences in the proportion of genes detected across batches (Supplementary Fig. 4.14). We also removed genes with molecule counts larger than 1,024 for the correction of collision probability. In the end we kept 13,058 endogenous genes and 48 ERCC spike-in genes.

4.5.7 Calculate the input molecule quantities of ERCC spiked-ins

According to the information provided by Fluidigm, each of the 96 capture chamber received 13.5 nl of lysis buffer, which contain 1:50,000 Ambion ERCC Spike-In Mix1 (Life Technologies) in our setup. Therefore, our estimation of the total spiked-in molecule number was 16,831 per sample. Since the relative concentrations of the ERCC genes were provided by

the manufacturer, we were able to calculate the molecule number of each ERCC gene added to each sample. We observed that the levels of ERCC spike-ins strongly correlated with the input quantities (r = 0.9914, Fig. 4.1G). The capture efficiency, defined as the fraction of total input molecules being successfully detected in each high quality cell, had an average of 6.1%.

4.5.8 Subsampling

We simulated different sequencing depths by randomly subsampling reads and processing the subsampled data through the same pipeline described above to obtain the number of molecules per gene for each single cell. To assess the impact of sequencing depth and number of single cells, we calculated the following three statistics:

1. The Pearson correlation of the gene expression level estimates from the single cells compared to the bulk samples. For the single cells, we summed the gene counts across all the samples and then calculated the $\log_2$ cpm of this pseudo-bulk. For the bulk samples, we calculated the $\log_2$ cpm separately for each of the three replicates and then calculated the mean per gene.
2. The number of genes detected with at least one molecule in at least one cell.
3. The Pearson correlation of the cell-to-cell gene expression variance estimates from the subsampled single cells compared to the variance estimates using the full single cell data set.

Each data point in Fig. 4.2 represents the mean +/- the standard error of the mean (SEM) of 10 random subsamples of cells. We split the genes by expression level into two groups (6,097 genes each) to highlight that most of the improvement with increased sequencing depth and number of cells was driven by the estimates of the lower half of expressed genes. The data shown is for individual NA19239, but the results were consistent for individuals

NA19098 and NA19101. Only high quality single cells (Supplementary Table 4.2) were included in this analysis.

4.5.9 A framework for testing individual and batch effects

Individual effect and batch effect between the single cell samples were evaluated in a series of analyses that examine the potential sources of technical variation on gene expression measurements. These analyses took into consideration that in our study design, sources of variation between single cell samples naturally fall into a hierarchy of individuals and C1 batches. In these sample-level analyses, the variation introduced at both the individual-level and the batch-level was modeled in a nested framework that allows random noise between C1 batches within individuals. Specifically, for each cell sample in individual i, replicate j and well k, we used y_{ijk} to denote some sample measurement (e.g. total molecule-counts) and fit a linear mixed model with the fixed effect of individual α_i and the random effect of batch b_{ij}:

$$y_{ijk} = \alpha_i + b_{ij} + \epsilon_{ijk} \quad (1)$$

where the random effect b_{ij} of batch follows a normal distribution with mean zero and variance σ_b^2, and ϵ_{ijk} describes residual variation in the sample measurement. To test the statistical significance of individual effect (i.e., null hypothesis $\alpha_1 = \alpha_2 = \alpha_3$), we performed a likelihood ratio test (LRT) to compare the above full model and the reduced model that excludes α_i. To test if there was a batch effect (i.e., null hypothesis $\sigma_b^2 = 0$), we performed an F-test to compare the variance that is explained by the above full model and the variance due to the reduced model that excludes b_{ij}.

The nested framework was applied to test the individual and batch effects between samples in the following cases. The data includes samples after quality control and filtering.

1. Total molecule count (on the $\log_2$ scale) was modeled as a function of individual effect and batch effect, separately for the ERCC spike-ins and for the endogenous genes.

2. Read-to-molecule conversion efficiency was modeled as a function of individual effect and batch effect, separately for the ERCC spike-ins and for the endogenous genes.

4.5.10 Estimating variance components for per-gene expression levels

To assess the relative contributions of individual and technical variation, we analyzed per-gene expression profiles and computed variance component estimates for the effects of individual and C1 batch (Supplementary Fig. 4.8). The goal here was to quantify the proportion of cell-to-cell variance due to individual (biological) effect and to C1 batch (technical) at the per-gene level. Note that the goal here was different from that of the previous section, where we simply tested for the existence of individual and batch effects at the sample level by rejecting the null hypothesis of no such effects. In contrast, here we fit a linear mixed model per gene where the dependent variable was the gene expression level ($\log_2$ counts per million) and the independent variables were individual and batch, both modeled as random effects.

The variance parameters of individual effect and batch effect were estimated using a maximum penalized likelihood approach [31], which can effectively avoid the common issue of zero variance estimates due to small sample sizes (there were three individuals and eight batches). We used the blmer function in the R package blme and set the penalty function to be the logarithm of a gamma density with shape parameter = 2 and rate parameter tending to zero.

The estimated variance components were used to compute the sum of squared deviations for individual and batch effects. The proportion of variance due to each effect is equal to the relative contribution of the sum of squared deviations for each effect compared to the total sum of squared deviations per gene. Finally, we compared the estimated proportions of variance due to the individual effect and the batch effect, across genes, using a non-parametric

one-way analysis of variance (Kruskal-Wallis rank sum test).

4.5.11 Normalization

We transformed the single cell molecule counts in multiple steps (Fig. 4). First, we corrected for the collision probability using a method similar to that developed by Grn et al. [63]. Essentially we corrected for the fact that we did not observe all the molecules originally in the cell. The main difference between our approach and that of Grn et al. [63] was that we applied the correction at the level of gene counts and not individual molecule counts. Second, we standardized the molecule counts to $\log_2$ counts per million (cpm). This standardization was performed using only the endogenous gene molecules and not the ERCC molecules. Third, we corrected for cell-to-cell technical noise using the ERCC spike-in controls. For each single cell, we fit a Poisson generalized linear model (GLM) with the $\log_2$ expected ERCC molecule counts as the independent variable, and the observed ERCC molecule counts as the dependent variable, using the standard log link function. Next we used the slope and intercept of the Poisson GLM regression line to transform the $\log_2$ cpm for the endogenous genes in that cell. This is analogous to the standard curves used for qPCR measurements, but taking into account that lower concentration ERCC genes will have higher variance from Poisson sampling. Fourth, we removed technical noise between the eight batches (three replicates each for NA19101 and NA19239 and two replicates for NA19098). We fit a linear mixed model with a fixed effect for individual and a random effect for the eight batches and removed the variation captured by the random effect (see the next section for a detailed explanation).

For the bulk samples, we used read counts even though the reads contained UMIs. Because these samples contained RNA molecules from ~10,000 cells, we could not assume that the 1,024 UMIs were sufficient for tagging such a large number of molecules. We standardized the read counts to $\log_2$ cpm.

4.5.12 Removal of technical batch effects

Our last normalization step adjusted the transformed $\log_2$ gene expression levels for cell-to-cell correlation within each C1 plate. The algorithm mimics a method that was initially developed for adjusting within-replicate correlation in microarray data [168]. We assumed that for each gene g, cells that belong to the same batch j are correlated, for batches $j = 1, \ldots, 8$. The batch effect is specific to each C1 plate and is independent of biological variation across individuals.

We fit a linear mixed model for each gene g that includes a fixed effect of individual and a random effect for within-batch variation attributed to cell-to-cell correlation in each C1 plate:

$$y_{g,ijk} = \mu_g + \alpha_{g,i} + b_{g,ij} + \epsilon_{g,ijk}, \quad (2)$$

where $y_{g,ijk}$ denotes $\log_2$ counts-per-million (cpm) of gene g in individual i, replicate j, and cell k; $i = NA19098, NA19101, NA19239$, $j = 1, \ldots, n_i$ with n_i the number of replicates in individual i, $k = 1, \ldots, n_{ij}$ with n_{ij} the number of cells in individual i replicate j. μ_g denotes the mean gene expression level across cells, $\alpha_{g,i}$ quantifies the individual effect on mean gene expression, $b_{g,ij}$ models the replicate effect on mean expression level (assumed to be stochastic, independent, and identically distributed with mean 0 and variance $\sigma^2_{g,b}$). Finally, $\epsilon_{g,ijk}$ describes the residual variation in gene expression.

Batch-corrected expression levels were computed as

$$\widehat{y}_{g,ijk} = y_{g,ijk} - \widehat{b}_{g,ij}, \quad (3)$$

where $\widehat{b}_{g,ij}$ are the least-squares estimates. The computations in this step were done with the gls.series function of the limma package [153].

4.5.13 Measurement of gene expression noise

While examining gene expression noise (using the coefficient of variation or CV) as a function of mean RNA abundance across C1 replicates, we found that the CV of molecule counts among endogenous genes and ERCC spike-in controls suggested similar expression variability patterns. Both endogenous and ERCC spike-in control CV patterns approximately followed an over-dispersed Poisson distribution (Supplementary Fig. 4.15), which is consistent with previous studies [77, 19]. We computed a measure of gene expression noise that is independent of RNA abundance across individuals [100, 134]. First, squared coefficients of variation (CVs) for each gene were computed for each individual and also across individuals, using the batch-corrected molecule data. Then we computed the distance of individual-specific CVs to the rolling median of global CVs among genes that have similar RNA abundance levels. These transformed individual CV values were used as our measure of gene expression noise. Specifically, we computed the adjusted CV values as follows:

1. Compute squared CVs of molecule counts in each individual and across individuals.

2. Order genes by the global average molecule counts.

3. Starting from the genes with the lowest global average gene expression level, for every sliding window of 50 genes, subtract $\log_{10}$ median squared CVs from $\log_{10}$ squared CVs of each cell line, and set 25 overlapping genes between windows. The computation was performed with the rollapply function of the R zoo package [208]. After this transformation step, CV no longer had a polynomial relationship with mean gene molecule count (Supplementary Fig. 4.15).

4.5.14 Identification of genes associated with inter-individual differences in regulatory noise

To identify differential noise genes across individuals, we computed median absolute deviation (MAD) - a robust and distribution-free dissimilarity measure for gene g:

$$MAD_g = Median_{i=1,2,3} \left| \text{adjCV}_{g,i} - Median_{i=1,2,3}(\text{adjCV}_{g,i}) \right|. \quad (4)$$

Large values of MAD_g suggest a large deviation from the median of the adjusted CV values. We identified genes with significant inter-individual differences using a permutation-based approach. Specifically, for each gene, we computed empirical P-values based on 300,000 permutations. In each permutation, the sample of origin labels were shuffled between cells. Because the number of permutations in our analysis was smaller than the maximum possible number of permutations, we computed the empirical P-values as $\frac{b+1}{m+1}$, where b is the number of permuted MAD values greater than the observed MAD value, and m is the number of permutations. Adding 1 to b avoided an empirical P-value of zero [141].

4.5.15 Gene enrichment analysis

We used ConsensusPATHDB [86] to identify GO terms that are over-represented for genes whose variation in single cell expression levels were significantly difference between individuals.

4.5.16 Individual assignment based on scRNA-seq reads

We were able to successfully determine the correct identity of each single cell sample by examining the SNPs present in their RNA sequencing reads. Specifically, we used the method verifyBamID (`https://github.com/statgen/verifyBamID`) developed by Jun et al., 2012 [83], which detects sample contamination and/or mislabeling by comparing the polymor-

phisms observed in the sequencing reads for a sample to the genotypes of all individuals in a study. For our test, we included the genotypes for all 120 Yoruba individuals that are included in the International HapMap Project [184]. The genotypes included the HapMap SNPs with the 1000 Genomes Project SNPs [183] imputed, as previously described [124]. We subset to include only the 528,289 SNPs that overlap Ensembl protein-coding genes. verifyBamID used only 311,848 SNPs which passed its default thresholds (greater than 1% minor allele frequency and greater than 50% call rate). Using the option –best to return the best matching individual, we obtained 100% accuracy identifying the single cells of all three individuals (Supplementary Fig. 4.13).

4.5.17 Data and code availability

The data have been deposited in NCBI's Gene Expression Omnibus [45] and are accessible through GEO Series accession number GSE77288 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77288`). The code and processed data are available at `https://github.com/jdblischak/singleCellSeq`. The results of our analyses are viewable at `https://jdblischak.github.io/singleCellSeq/analysis`.

4.6 Acknowledgments

We thank members of the Pritchard, Gilad, and Stephens laboratories for valuable discussions during the preparation of this manuscript. This work was funded by NIH grant HL092206 to YG and HHMI funds to JKP. PYT is supported by NIH T32HL007381. JDB was supported by NIH T32GM007197. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4.7 Author Contributions

YG and JKP conceived of the study, designed the experiments, and supervised the project. PT and JEB performed the experiments. PT, JDB, CH, and DAK analyzed the results. PT, JDB, CH, and YG wrote the original draft. All authors reviewed the final manuscript.

4.8 Supplementary Information

4.8.1 Supplementary Figures

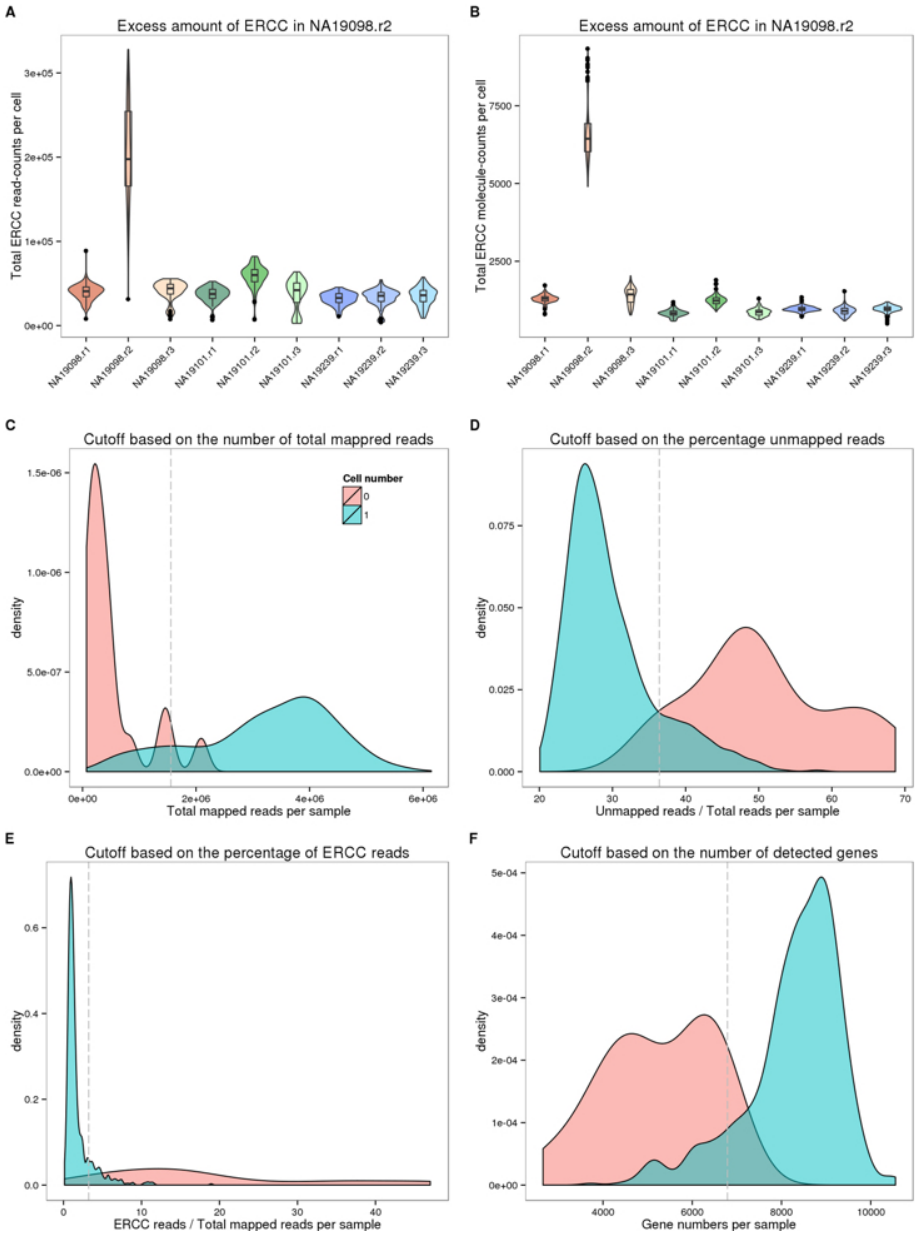


Figure 4.6: **Removal of low quality samples.** Violin plots of the total read-counts of ERCC spike-in controls in (A) and the total molecule-counts in (B) in single cell samples. The three colors represent the three individuals (NA19098 in red, NA19101 in green, and NA19239 in blue). (C-F) Density plots of the distributions of the total mapped reads in (C), the percentage of unmapped reads in (D), the percentage of ERCC reads in (E), and the number of detected genes in (F). The dash lines indicate the cutoffs based on the 95th percentile of the samples with no cells.

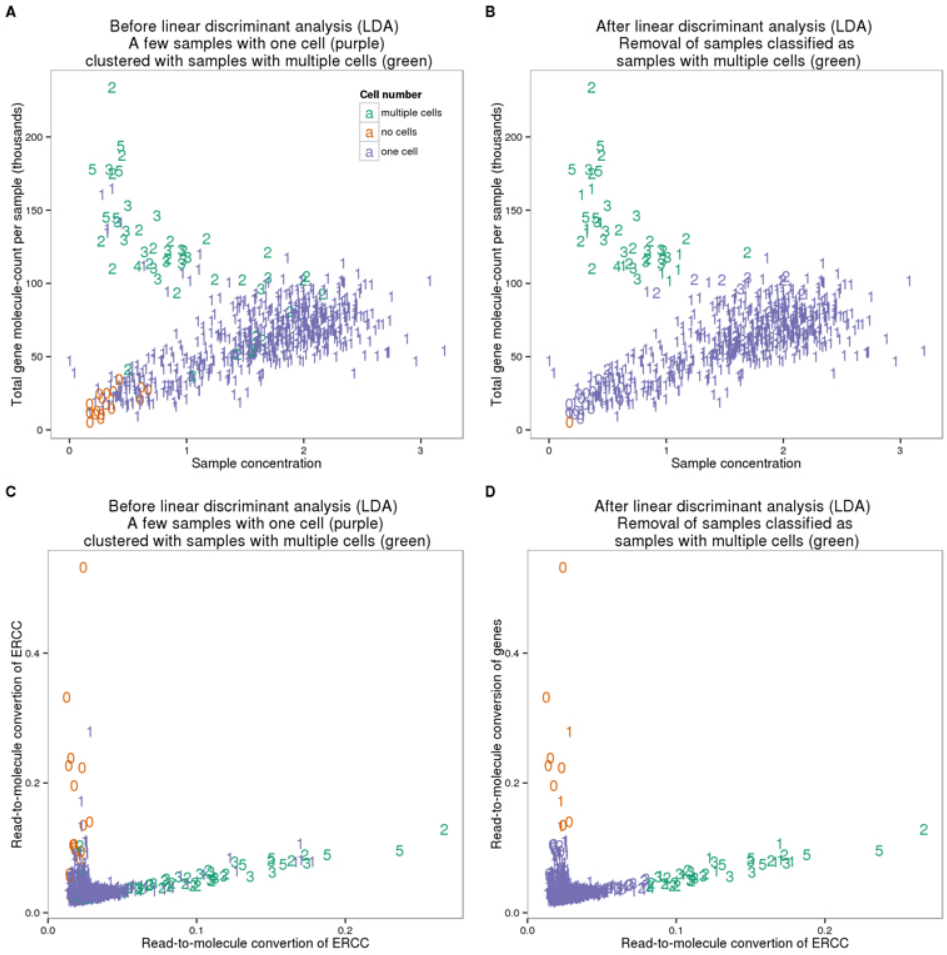


Figure 4.7: **Removal of samples with multiple cells.** Scatterplots of the three groups of samples (no cell in green, single-cell in orange, and two or more cells in purple) before (A) and after (B) the linear discriminant analysis (LDA) using sample concentration of cDNA amplicons (ng/μl) and the number of detected genes. (C and D) Similarly, LDA was performed to identify potential multi-cell samples using the read-to-molecule conversion efficiency (total molecule-counts divided by total read-counts per sample) of endogenous genes and ERCC spike-in controls. Scatterplots of before and after the LDA in (C) and (D), respectively. The numbers indicate the number of cells observed in each cell capture site.

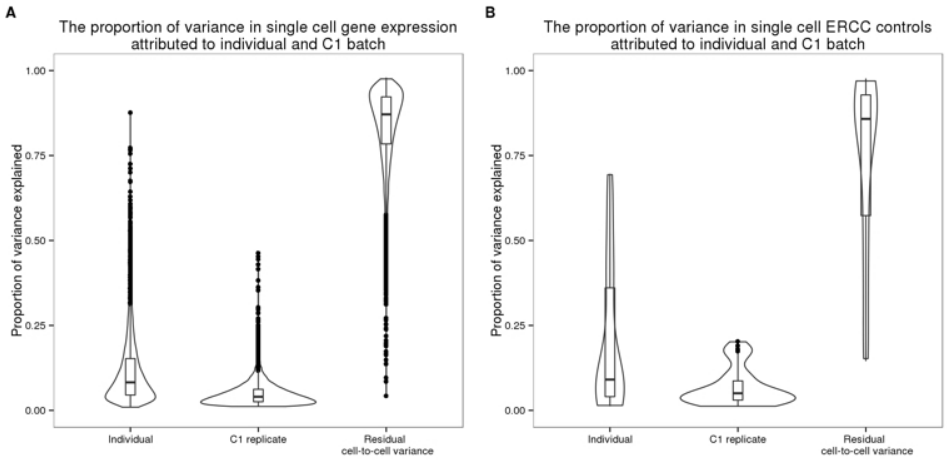


Figure 4.8: **Sources of cell-to-cell variance in per-gene expression profile.** Violin plots of the proportion of per-gene cell-to-cell variance that was due to individual sample of origin, different C1 replicates, and other single cell sample differences. These results were calculated from the molecule counts before normalization and batch correction. Endogenous genes are shown in (A) and the ERCC spike-in controls in (B).

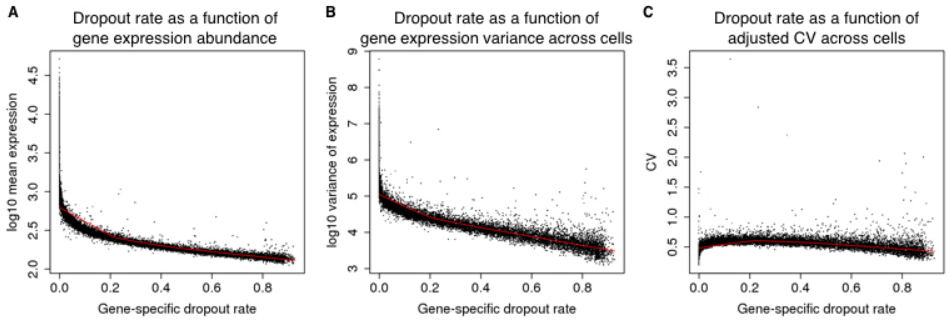


Figure 4.9: **The gene-specific dropout rate.** The gene-specific dropout rate (the proportion of cells in which the gene is undetected) and its relationship with $\log_{10}$ mean expression in (A), with $\log_{10}$ variance of expression in (B), and with the CV in (C) of the cells in which the gene is expressed (cells in which at least one molecule of the given gene was detected). Each point represents a gene, and red lines indicate the predicted values using locally weighted scatterplot smoothing (LOESS).

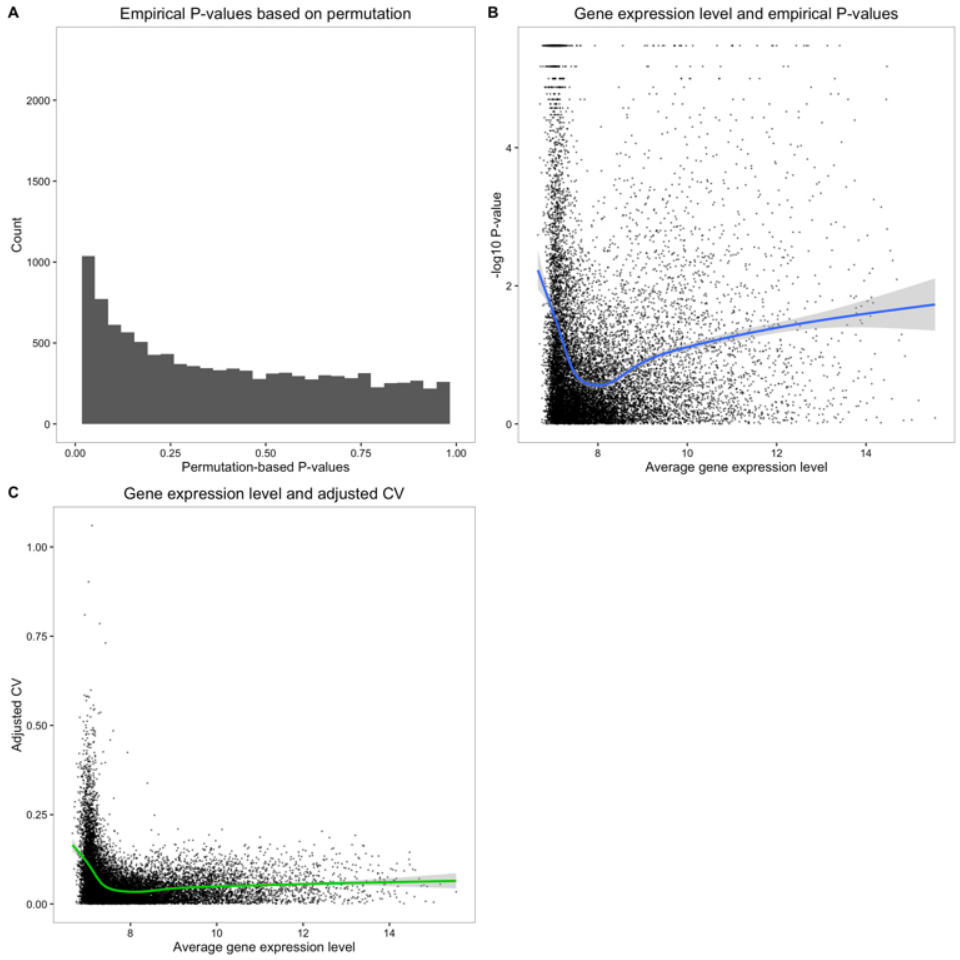


Figure 4.10: **Permutation-based *P*-value.** (A) Histogram of empirical P-values based on 300,000 permutations. (B) -$\log_{10}$ empirical P-values are plotted against average gene expression levels. Blue line indicates the fitted relationship between -$\log_{10}$ P-values and average $\log_2$ gene expression levels of cells that were detected as expressed, using locally weighted scatterplot smoothing (LOESS). (C) Median of Absolute Deviation (MAD) of genes versus average gene expression levels. Green line indicates the fitted relationship (LOESS) between the MAD values and average $\log_2$ gene expression levels of cells in which the gene was detected as expressed.

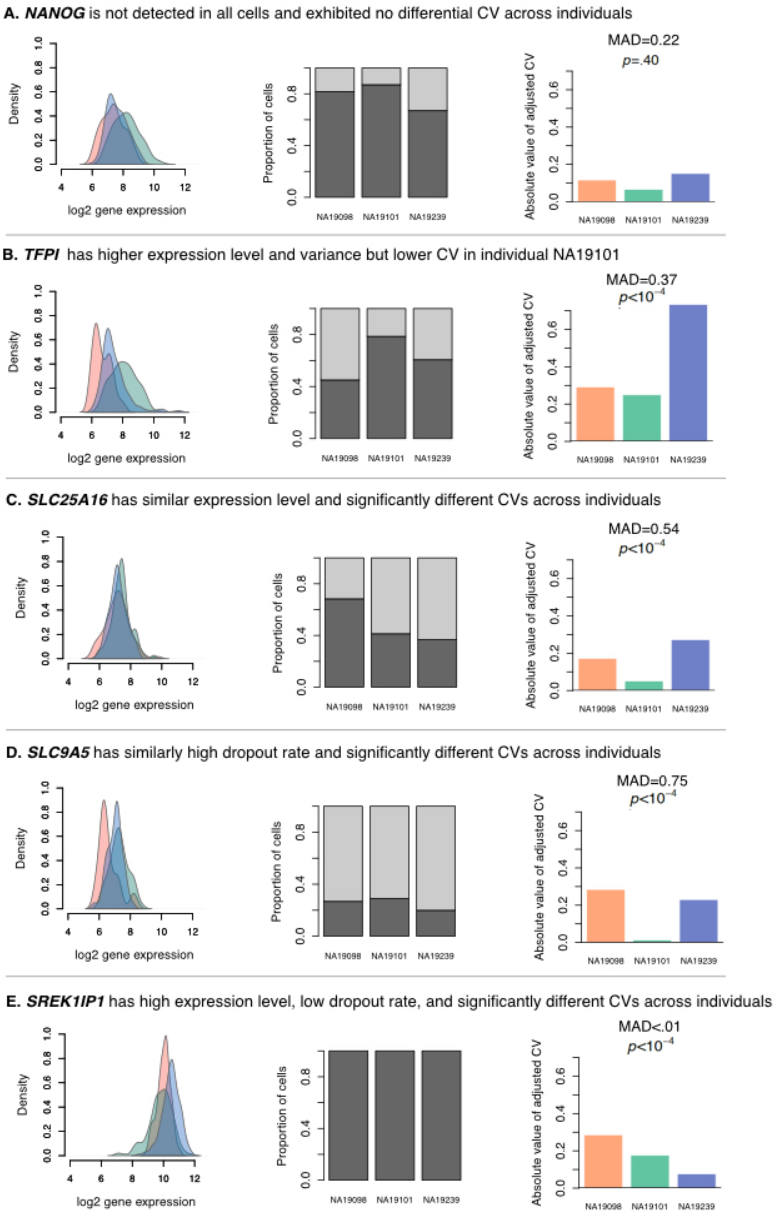


Figure 4.11: **Inter-individual differences in regulatory noise.** These 5 example genes illustrate various patterns of cell-to-cell gene expression variance. For each gene, the left panel shows the distribution of the $\log_2$ gene expression levels (considering only cells in which the gene is detected as expressed), the middle panel shows the proportion of cells in which the gene is detected as expressed (dark grey) and the dropout rate (light grey) for each individual, and the right panel shows the absolute value of adjusted CV for each individual, along with the corresponding gene-specific MAD (median of absolute deviation) value and P-value. The three colors in the upper and lower panel represent the individuals (NA19098 in red, NA19101 in green, and NA19239 in blue).

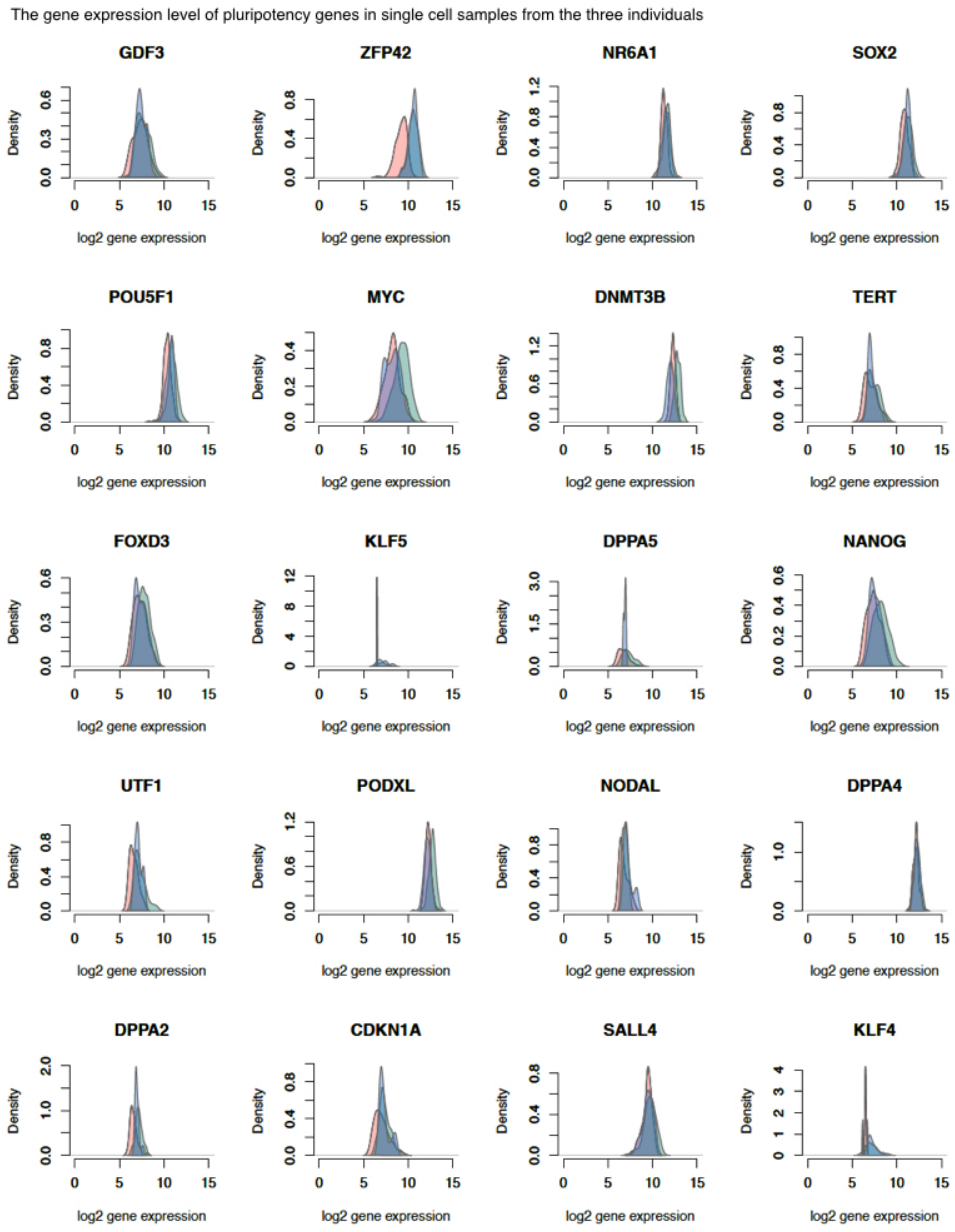
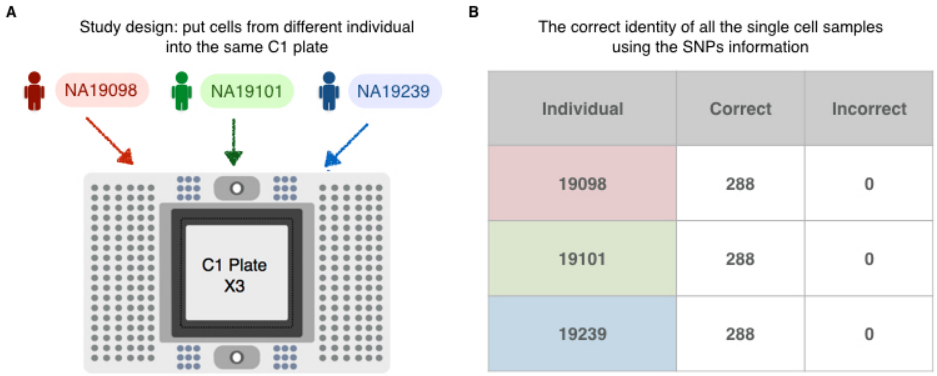


Figure 4.12: **Cell-to-cell variation of pluripotency genes.** Density plots of the distribution of $\log_2$ gene expression of key pluripotency genes across all single cells by individual. The peaks with lower gene expression values (log2 around 4) represent the cells in which the gene is undetected. The three colors represent the three individuals (NA19098 is in red, NA19101 in green, and NA19239 in blue).



Individual	Correct	Incorrect
19098	288	0
19101	288	0
19239	288	0

Figure 4.13: **Proposed study design for scRNA-seq using C1 platform.** (A) A balanced study design consisting of multiple individuals within a C1 plate and multiple C1 replicates to fully capture the batch effect across C1 plates and further retrieve the maximum amount of biological information. (B) The correct identity of each single cell sample was determined by examining the SNPs present in their RNA sequencing reads.

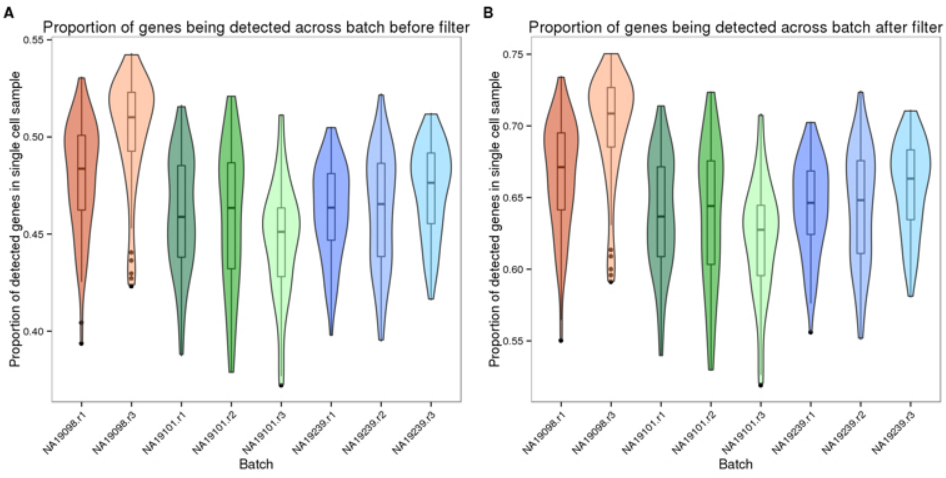


Figure 4.14: **The proportion of genes detected in single cell samples.** Violin plots of the proportion of genes detected, computed by the total number of detected genes in each single cell divided by the total number of genes detected across all single cells, before in (A) and after in (B) the removal of genes with low expression. The three colors represent the three individuals (NA19098 is in red, NA19101 in green, and NA19239 in blue).

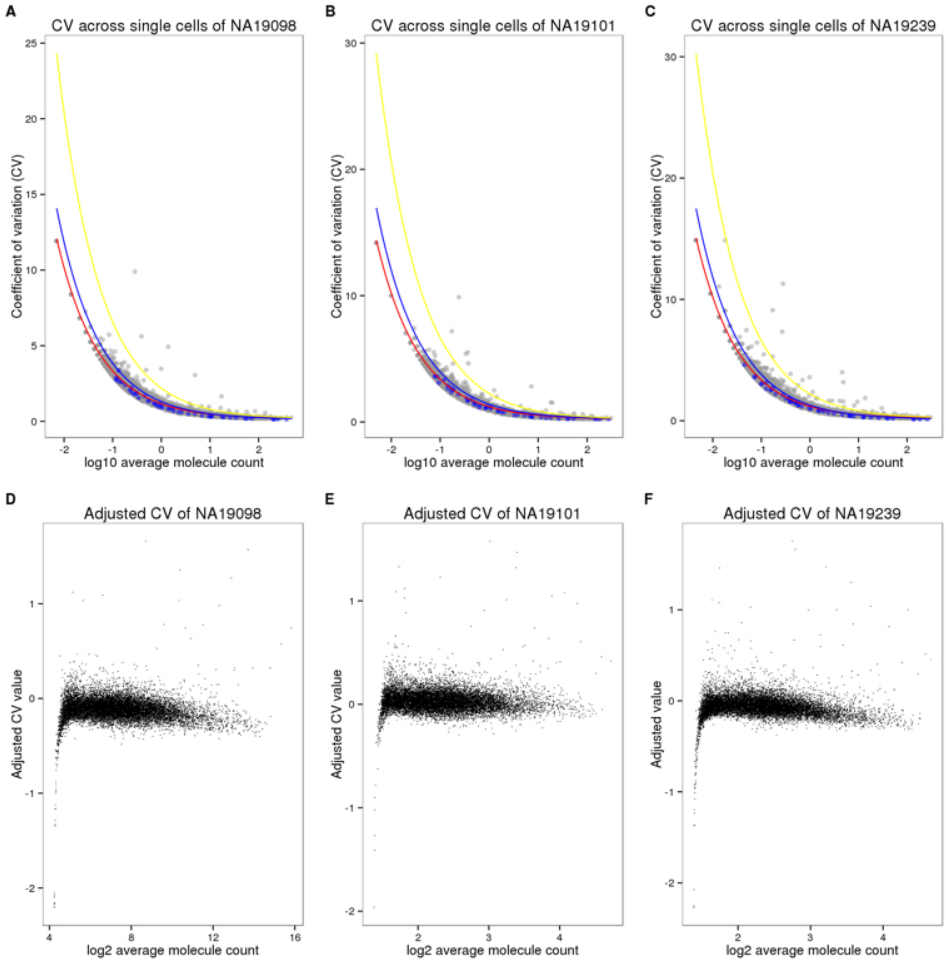


Figure 4.15: **Coefficients of variation (CV) before and after adjusting for gene mean abundance.** (A-C) CV plotted against average molecule counts across all cells for each individual [77]. Grey points represent endogenous genes, and blue points represent ERCC spike-in controls. The curves indicate the expected CV under three different scenarios. Red curve depicts the expected CV of the endogenous genes while assuming a Poisson distribution with no over-dispersion. Likewise, blue curve depicts the expected CVs of the ERCC spike-in controls under the Poisson assumption. Yellow curve depicts the expected CVs of an over-dispersed Poisson distribution for which standard deviation is three times the ERCC spike-in controls. (D-F) Adjusted CV values of each gene including all cells are plotted against $\log_{10}$ of the average molecule counts for each individual.

4.8.2 Supplementary Tables

A

Information and results of each C1 collections

Cell line	Passage	Input viability	No cell	Multiple cells	% single cell occupancy	TRA1-60 negative	% TRA1-60	Date	Ave cDNA con (ng/ul)	Replicate for seq
19098	12+15	89%	2	2	95.83	2	97.92	11/07/2014	N/A	(1)
19098	12+18	90%	2	4	93.75	0	100.00	11/14/2014	1.88	(2)
19098	12+20	83%	3	10	86.46	0	100.00	11/22/2014	1.40	(3)
19101	12+16	70%	2	3	94.79	9	90.63	11/13/2014	1.81	(1)
19101	12+19	94%	5	3	91.67	0	100.00	11/23/2014	1.38	(2)
19101	12+19	69%	1	19	79.17	0	100.00	11/24/2014	1.26	(3)
19239	12+16	85%	1	2	96.88	4	95.83	11/11/2014	1.60	(1)
19239	12+18	75%	1	6	92.71	5	94.79	11/17/2014	1.55	(2)
19239	12+19	93%	5	7	87.50	2	97.92	11/21/2014	1.70	(3)

B

The arrangement of samples for sequencing on four flowcells

Flowcell 1	Flowcell 2	Flowcell 3	Flowcell 4
Bulk	19098 (2)	19098 (3)	19239 (1)
19098 (1)	19239 (3)	19101 (1)	19101 (2)
19239 (2)	19098 (1)	19098 (2)	19098 (3)
19101 (3)	19239 (2)	19239 (3)	19101 (1)
19239 (1)	19101 (3)	Bulk	19098 (2)
19101 (2)	19239 (1)	19098 (1)	19239 (3)
19098 (3)	19101 (2)	19239 (2)	Bulk (all 9)
19101 (1)	Bulk	19101 (3)	

Table 4.1: **Data collection.** (A) iPSCs were sorted using the 10-17 μm IFC plates with the staining of the pluripotency marker, TRA1-60. Single cell occupancy is the percentage of occupied capture sites containing one single cell. The average cDNA concentration was measured by the HT DNA high sensitivity LabChip (Caliper). (B) The 96 single cell libraries from one C1 plate were pooled and sequenced in three HiSeq lanes. The pooled samples were assigned across the four 8-lane flowcells.

Table 4.2: **High quality single cell samples.** (see supplementary file associated with this dissertation) List of the 564 high quality single cell samples.

Table 4.3: **Genes associated with inter-individual differences in regulatory noise.** (see supplementary file associated with this dissertation) List of genes that we classified the estimates of regulatory noise as significantly different across individuals (empirical permutation $P < 10^{-4}$). There are a total of 560 genes.

Table 4.4: **Gene ontology analysis of the genes associated with inter-individual differences in regulatory noise.** (see supplementary file associated with this dissertation)

CHAPTER 5

CONCLUSION

Traditional genetics approaches have been unable to identify variants which can be used to predict susceptiblity to tuberculosis (TB), likely due to the highly polygenic architecture of this complex trait [190, 121, 189, 144, 30, 36, 169]. Thus I performed experiments to interrogate a higher level phenotype, gene expression levels, for which the effect of many variants of small effect size can manifest in aggregrate. In my first approach, I identified genes in innate immune cells whose gene expression levels change in response to infection with *Mycobacterium tuberculosis* (MTB) but not other bacteria, highlighting their potential importance for mycobacterial diseases [15]. In my second approach, I measured gene expression levels in innate immune cells from individuals either susceptible or resistant to develop active TB and built a classifier to predict susceptibility to TB. These first two experiments measured average gene expression levels across many cells, and thus they missed any cell-to-cell heterogeneity in the innate immune system [157, 146]. In my third approach, I established principles for the effective design of studies to measure gene expression levels in single cells [192]. Given the success of my first two experiments, I expect many more discoveries will be made by interrogating gene expression measurements in single cells of the innate immune system.

5.1 A joint Bayesian model provides a general framework for analyzing functional genomics studies with many conditions

In Chapter 2, I described my work investigating the innate immune response to MTB [15]. It is known that the innate immune response is important for fighting MTB infections [93]. Alveolar macrophages are the primary target of MTB, and they initiate the formation of granulomas to sequester MTB [164]. Furthermore, vaccines against TB have had limited

efficacy [198]. To identify human genes which are important for the response to MTB infection, we isolated macrophages from six healthy donors, infected them with MTB and other bacteria, and measured genome-wide gene expression levels using RNA-seq at 4, 18, and 48 hours post-infection.

Previous studies had identified genes which are differentially expressed upon infection with MTB [46, 150, 132, 27, 197, 181], and some have even compared the differences between the reponse to strains of MTB that vary in their virulence [34, 205]. The first novelty of our study was to include many other bacteria in the infection experiments. Specifically, we included the following mycobacteria: two strains of virulent MTB, avirulent (heat-inavtivated) MTB, bacillus Calmette-Guérin (BCG; attentuated *Mycobacterium bovis* used as a vaccine), and the avirulent *Mycobacterium smegmatis*. The non-mycobacteria species we included were *Yersinia pseudotuberculosis*, *Salmonella typhimurium*, and *Staphylococcus epidermidis*. This allowed us to distinguish between the innate immune response to MTB versus other virulent bacteria, MTB versus avirulent mycobacteria, amd MTB versus deceased MTB.

This novel study design comparing many bacterial infections to isolate the innate immune respone to MTB also posed analytical challenges. The goal was to identify differences between the innate immune response to each of the eight bacterial infections compared to the non-infected control condition. Standard differential expression analyses (or in general any large scale testing of thousands or more genomic features) are well-suited for experiments with a few conditions [138, 1, 153]. For example, the most common approach is to perform pairwise differential expression tests and then overlap the lists of differentially expressed genes. In this instance, that would have meant peforming eight pairwise tests to compare each bacterial infection to the control. These results are always biased by incomplete power [40, 51]. Because hypothesis testing uses an arbitrary p-value threshold to determine statistical significance, a gene with a p-value below this threshold for one comparison but a p-value slightly above this threshold for a separate comparison will be classified as specific

to the first when in reality the gene is behaving similarly in both. As the number of pairwise comparisons increases, the problem of incomplete power is exacerbated, i.e. a gene is more likely to be statistically significant for some susbset of comparisons. This increase in comparisons also decreases the ability to interpret the results. A 3-way Venn diagram (and perhaps a 4- or 5-way) can be interpreted, but this approach breaks down with additional comparisons.

Another approach would be to directly compare the effect of infection between two different groups of bacteria, e.g. compare the mean effect of infection with mycobacteria versus the mean effect of infection with non-mycobacteria (or virulent versus non-virulent bacteria). The advantage of this approach is that it explicitly models the comparison and returns a p-value, unlike the Venn diagram overlap approach. However, there are two main downsides. First, statistical significance can be driven by outliers. For example, in my study most of the significantly differentially expressed genes between mycobacteria and non-mycobacteria were actually genes which were simply differentially expressed in response to infection with *S. typhimurium* and *S. epidermidis*. Second, this limits the potential results to the *a priori* ideas of the analyst and are not driven by the patterns in the actual data.

On the other end of the spectrum, a very data-driven approach would be to use a clustering method such as hierarchical or k-means clustering [47, 126]. These multivariate methods are able to find the patterns of gene expression in the data, both expected and unexpected; however, since they are not accompanied by any formal hypothesis test, it is difficult to interpret which clusters of co-expressed genes are the most interesting to report.

Since none of the standard genomics approaches were adequate for properly comparing 8 bacterial infections, I instead used a joint Bayesian model, implemented in the software package Cormotif, to analyze the data [201]. Conceptually, Cormotif combines the clustering and pairwise testing approaches described above. Just like the pairwise testing approach, the input to Cormotif are the pairwise comparisons between each bacterial infection and

the control condition. However, to account for incomplete power, Cormotif models the gene expression levels across all the pairwise comparisons to identify the main gene expression patterns, conceptually similar to a clustering analysis.

The Cormotif results for my study were informative. Most of the genes were either differentially expressed or not after infection with any of the bacteria (Fig. **??**). The two most interesting patterns in regards to understanding the innate immune response to MTB were "MTB" and "Virulent" (Fig. **??**,**??**). The "MTB" pattern included those genes which had a high posterior probability of being differentially expressed in response to infection with MTB or closely related species and a medium posterior probability of being differentially expressed in response to ifnection with *M. smegmatis*, the nonvirulent mycobacteria. The "Virulent" pattern included genes which had a high posterior probability of being differentially expressed in response to infection with any of the bacteria except heat-inactivated MTB or BCG.

In terms of better understanding TB susceptibility, the main takeaway from this study was the identification of hundreds of genes which are differentially expressed in response specifically to infection with MTB and related species but not other virulent bacteria. These genes are candidates for containing genetic variants which affect TB susceptibility. Furthermore, these genes could be targets for future functional studies of how the innate immune system fights MTB and also could give context to future results from genetic and functional genomics studies of MTB infection. More generally, our methods are informative to all future functional genomics studies. We were only able to confidently isolate the effects of MTB infection by including multiple other bacterial infections as comparison. Had we only infected the macrophages with MTB and heat-inactivated MTB, we would have made multiple misclassifications. We would have assigned differences between the two infections as specific to a live, virulent MTB; however, these gene expression changes were also induced by other live bacteria. Similarly, we would never have known that a subset of the genes which were differentially expressed in response to both MTB and heat-inactivated MTB

were actually specific to mycobacteria in general. Not only was it important to include multiple bacterial infections, but it was also critical to properly analyze the results. Because the innate immune system is largely a general response to infection, we expected most of the induced gene expression changes to be similar across bacteria [72, 16, 132, 79]. Had we performed the straight-forward approach of overlapping lists of differentially expressed genes from comparing the individual infections to their controls, we would have had identified lots of spurious differences in the innate immune response caused by incomplete power. In contrast, by jointly modeling the data with Cormotif [201], we were able to identify the shared gene expression patterns in response to related bacteria. In support of the generality of this approach, the Cormotif approach was successfully applied to distinguish the effects of vitamin D and bacterial lipopolysaccharide on the innate immune response between individuals of African-American and European-American ancestry (note: I was a co-author of the study) [89].

It should be noted that this method also has its caveats. First, its strength of sharing information across the pairwise comparisons can also be a negative because it will not identify genes with unique expression patterns (Fig. 2.6). While useful for projects with the aim of broadly characterizing the genome-wide gene expression patterns for a given phenomenon, it is not well-suited for identifying outlier genes. Second, because the algorithm is not deterministic, Cormotif must be run multiple times to obtain the model with the highest log likelihood. Because of this added complexity, using Cormotif is more difficult to implement than more standard differential expression approaches.

5.2 Initial success classifying individuals susceptible to tuberculsosis and future directions

In Chapter 3, I described my work investigating the role of gene regulation in the innate immune system on TB susceptibility (not yet published). Specifically, in order to investigate

how the innate immune cells of susceptible individuals function compared to those of resistant individuals, we collected primary dendritic cells (DCs) from individuals that had recovered from TB (i.e. susceptible) and individuals that tested positive for latent TB infections but had not developed TB (i.e. resistant). We infected the DCs with MTB and performed RNA sequencing (RNA-seq) on the infected and non-infected cells.

There were three main conclusions from this work. First, the differences in gene expression levels between resistant and susceptible individuals were primarily present in the non-infected state and not 18 hours post-infection with MTB (Fig. 3.1). This suggests that these gene expression differences primarily affect the very early response to MTB infection. Second, we discovered that the effect sizes measured in our *in vitro* experiment, whether comparing between resistant and susceptible individuals or between the infected and non-infected states, were negatively correlated with lower p-values from two genome wide association studies (GWAS) of TB susceptibility [190] (Fig. 3.2). This suggests that our *in vitro* system is a useful model for investigating the genetic basis of TB susceptibility. Third, we trained a classifier based on the gene expression levels in the non-infected state (Fig. 3.3). Using the threshold required to obtain a 100% sensitivity (zero false negatives) in the training data, we found that 11% of healthy individuals from an independent study [9] were predicted to be susceptible to TB, very close to the estimated population average of 10% [135, 137]. This suggests that isolating innate immune cells and performing gene expression profiling could be a feasible test for TB susceptibility. The most obvious extension of this work is to conduct a larger study with more susceptible individuals. Our current results are only a proof-of-principle. With a larger study, we could properly split the data into training and test sets to assure that the model is not overfitting the data. On the one hand, since we identified that the gene expression differences are only present in the non-infected state and that these are sufficient for the performance of the classifier, this future study would be simplified by not having to perform the MTB infections. On the other hand, collecting a

large number of patient samples is always difficult, and it is even worse when the individuals are currently healthy and thus not regularly visiting the doctor like those recovered from a past case of active TB. Hopefully these initial successful results will provide the impetus for larger scale sample collection.

Another fruitful direction for future experiments would be to further investigate the role of *CCL1* in the innate immune response to MTB and its role in TB susceptibility. While studies of this gene have had mixed results [188, 182, 139], all the studies, including my own [15], have had small sample sizes. In the first study of *in vitro* differences in gene expression between susceptible and resistant individuals, *CCL1* was found to be differentially expressed based on susceptibility status [188]. Furthermore, the same study found that SNPs nearby *CCL1* were associated with TB susceptibility in an independent cohort. In Chapter 2, I found that *CCL1* was one of the genes which changed expression level specifically in response to infection with mycobacteria [15]. In Chapter 3, I found that *CCL1* was one of two genes which had an effect size greater than 2 between susceptible and resistant individuals in the non-infected state and also a p-value less than 0.01 in two GWAS of TB susceptibility. There were many differences between these studies (e.g. cell type, ethnicity of donors, timepoints RNA was collected post-infection), yet *CCL1* was still a top hit in all three analyses. I believe this warrants further investigation. As an example, one could use CRISPR/Cas [42] to modify individual SNPs in THP-1 cells (a common cell line model of monocytes) and test for differences in the response to MTB infection. Another idea, since *CCL1* is a secreted chemokine [127], would be to add varying amounts of exogenous *CCL1* to the *in vitro* system to detect an effect on the innate immune response.

5.3 Incorporating lessons from single cell pilot study for future studies of the genetic basis of gene expression noise and the response to bacterial infection

In Chapter 4, I described my work on single cell RNA-seq (scRNA-seq) [192]. scRNA-seq is a relatively new technique [109, 118, 156, 64, 172, 4] that enables the investigation of gene regulatory changes at a much finer resolution than the bulk RNA-seq projects I performed in Chapters 2 and 3. While this new technology is exciting, we must exercise the same caution as when performing any large-scale genomics experiment [2, 106, 59]. Early studies of the Fluidigm C1 system for scRNA-seq that focused on the technical sources of variation largely focused on the variation from well-to-well within just one C1 chip [19, 63, 77, 39, 194] ; whereas, the studies investigating biological phenomena tended to use multiple C1 chips without addressing the obvious confounding batch effects (this problem is nicely highlighted by [70]). Before conducting large scale scRNA-seq experiments, we first aimed to better understand the technical factors affecting the design of such studies. To do so, we performed scRNA-seq of three C1 chip replicates of three HapMap [74] Yoruba individuals.

From these data, we learned many important lessons. First, by performing subsampling analyses, we determined that sequencing approximately 1.5 million reads for at least 75 cells from a given individual was sufficient for detecting most expressed genes, achieving a high correlation between the sum of the gene expression levels across the single cells and the gene expression levels from bulk sequencing of 10,000 cells, and achieving a high correlation between the cell-to-cell variance in the gene expression levels across the subset of single cells and the cell-to-cell variance measured in all the single cells we collected for an individual (which ranged from 142 to 221) (Fig. 4.2). Second, we observed technical variation introduced from the processing of the C1 batches (Fig. 4.3). While this was expected, we also observed unexpected aspects of this batch effect. The ERCC spike-in controls which

were added to each well and could potentially be used to correct for this effect across C1 chips was affected not only by technical variation but also by the biological variation (differences between individuals) (Fig. 4.8). This entanglement of the technical and biological sources of variation renders the spike-ins insufficient for modeling technical variation between C1 chips (however they can still be used to model technical variation between wells of the same C1 chip). This confounder occurred despite our use of unique molecular identifiers (UMIs) to account for the bias introduced by amplifying RNA from a small original source of just one cell [98, 77]. In fact, we found that the conversion of reads to unique molecules was affected by inter-individual differences (Fig. 4.3). Third, even with our small sample size of only three individuals, we were able to identify inter-individual differences in the cell-to-cell gene expression variance, or gene expression noise (Fig. 4.5). This lends further support to the notion that gene expression noise is a relevant factor that can affect biological processes. Fourth, we demonstrated that we can use the single nucleotide polymorphisms (SNPs) present in the RNA-seq reads to identify the individual of origin for a given single cell [83] (Fig. 4.13). This enables us to use a crossed-design where single cells from multiple individuals are included on the same C1 chip and later each well is assigned to each individual based on the RNA-seq reads obtained. Our initial nested design was inefficient because we collected hundreds of single cells per individual across the multiple technical replicates. From our subsampling we knew that collecting 75 high quality single cells was sufficient. With a crossed design, we can obtain about one C1 chip worth of wells (96) while still modeling the technical variation across C1 chips.

Given the promising results from our first study, our next study will aim to further investigate the impact of genetic variation on gene expression noise by measuring single cell gene expression levels in 60 individuals. The design of the study is informed by our previous findings. First, we will put single cells from multiple individuals on each C1 chip because we know we can identify the individual based on the RNA-seq reads. Second, we will repeat each

individual across C1 chips until they obtain on average 96 wells (e.g. one C1 chip) because this will get us close to our target of 75 single cells after removing low quality cells. Third, we will replace the ERCC spike-ins with RNA from a distantly related model organism. With many more technical spike-ins gene to measure, these will be more useful for modeling technical variation [152]. Using this study design, we'll be able to efficiently measure gene expression noise from many individuals while still properly accounting for technical variation.

Returning to the *in vitro* models of bacterial infection from my other studies, I can imagine future single cell studies that shed further light on the innate immune response. While we infect the cells at a multiplicity of infection of 1:1, some cells will still be infected by multiple bacteria and others not infected at all. Furthermore, there could be variation in this distribution of the number of bacteria per cell across individuals. In order to efficiently measure single cell gene expression in response to infection, I would put uninfected cells from one individual on the same C1 chip as the infected cells from a different individual. Also, since the MTB H37Rv strain we typically use has a GFP tag, we could use high-throughput fluorescence microscopy of each well to count the number of bacteria per cell. With this high resolution data, we could differentiate between inter-individual differences in the innate immune response due to differences in the number of infected cells (or the number of bacteria per cell) or differences in the innate immune response in the infected cells.

5.4 The importance of mitigating batch effects in any genomics experiment

A common theme of all my projects is accounting for technical biases. Although only Chapter 4 has a main focus on mitigating batch effects, all my projects required close attention to this problem. This is because all genomics studies need to account for batch effects in both the design and analysis of the data, otherwise the results are meaningless [2, 106, 59]. There will be signal in any large data set, but it will only inform biological insight if the signal

arises from the biological processes being studied.

In Chapter 2, we collected a total of 156 RNA-seq samples. During the batch processing, we ensured that the biological factors of interest (bacterial infection, timepoint, individual) were balanced to avoid introducing spurious signal. Furthermore, upon data exploration, we observed that the processing batch and the RNA quality score (RIN) were correlated with the first principal component (PC) (Fig. 2.2). After regressing these two variables, the first PC was the effect of timepoint and the second PC was the effect of infection. Importantly, we obtained similar results with or without protecting the variables of interest in the linear model when regressing out the technical variables. This was a result of the careful planning of the batch processing to avoid confounding biological and technical variables.

In Chapter 3, I once again designed the batch processing to balance the biological factors of interest (susceptibility status, treatment, individual) (Fig. 3.4). Conveniently, this project did not have large scale batch effects (Fig. 3.9), likely due to the smaller overall sample size of 48. However, accounting for a batch effect was critical for training a classifier on the current data set and testing it on an independent data set [9]. Not only were the studies performed years apart, but the gene expression levels were measured using different technologies. Thus I was only able to compare the two studies after normalizing each sample (Fig. 3.13) and removing the large batch effect by regressing the first PC of the combined data set (Fig. 3.14). Testing the classifier without accounting for the batch effect would have given poor results simply due to technical reasons.

In Chapter 4, one of the main motivations for the study was understanding the magnitude of the batch effect of collecting scRNA-seq on separate C1 chips. While the technical effect of C1 batch was smaller in magnitude compared to the biological effect of individual (assessed using variance components analysis) (Fig. 4.8), not including technical replicates would attribute the substantial technical effect to the biological effect. Just as we require replication for established genomic protocols, it is also necessary to replicate scRNA-seq experiments,

especially since the standard ERCC spike-ins appear to be affected by both biological and technical factors. Fortunately, we were able to devise a strategy to reduce the required number of C1 replicates by combining single cells from multiple individuals onto each C1 chip and then replicating the multiple individuals across multiple chips (Fig. 4.13). This crossed design accounts for batch effects while minimizing the required replication.

In summary, technical batch effects need to be considered from the initial design of a genomics experiments through to the data analysis and interpretation of the results.

5.5 Concluding remarks

First, I have identified hundreds of genes specifically involved in fighting MTB infections. More broadly, I have demonstrated that a joint Bayesian model is an effective tool for analyzing the results of genomic studies with many conditions. Second, I have demonstrated that the gene expression levels in non-infected DCs may be able to predict susceptiblity to TB. Third, I have determined an effective study design for future single cell studies that accounts for technical batch effects while simultaneously decreasing the necessary sample size. Overall my results are informative not only for understanding how differences in the innate immune response confer susceptibility or resistance to TB, but also inform the design and analysis of any functional genomics experiment.

References

[1] Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols*, 8(9):1765–86, 2013.

[2] Paul L Auer and R W Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–16, 2010.

[3] C. C. Babbitt, O. Fedrigo, A. D. Pfefferle, A. P. Boyle, J. E. Horvath, T. S. Furey, and G. A. Wray. Both noncoding and protein-coding rnas contribute to gene expression evolution in the primate brain. *Genome Biol Evol*, 2:67–79, 2010.

[4] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome biology*, 17(1):63, 2016.

[5] N. E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, 10(9):e1004663, 2014.

[6] S. Barbash and T. P. Sakmar. Brain gene expression signature on primate genomic sequence evolution. *Sci Rep*, 7(1):17329, 2017.

[7] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–93, 2012.

[8] D. P. Barlow. Methylation and imprinting: from host defense to gene regulation? *Science*, 260(5106):309–10, 1993.

[9] Luis B Barreiro, Ludovic Tailleux, Athma A Pai, Brigitte Gicquel, John C Marioni, and Yoav Gilad. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1204–9, 2012.

[10] Matthew P R Berry, Christine M Graham, Finlay W McNab, Zhaohui Xu, Susannah a a Bloch, Tolu Oni, Katalin A Wilkinson, Romain Banchereau, Jason Skinner, Robert J Wilkinson, Charles Quinn, Derek Blankenship, Ranju Dhawan, John J Cush, Asuncion Mejias, Octavio Ramilo, Onn M Kon, Virginia Pascual, Jacques Banchereau, Damien Chaussabel, and Anne O'Garra. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, 466(7309):973–7, 2010.

[11] L. E. Blake, S. M. Thomas, J. D. Blischak, C. J. Hsiao, C. Chavarria, M. Myrthil, Y. Gilad, and B. J. Pavlovic. A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biol*, 19(1):162, 2018.

[12] Simon Blankley, Matthew Paul Reddoch Berry, Christine M Graham, Chloe I Bloom, Marc Lipman, and Anne O'Garra. The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1645):20130427, 2014.

[13] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*, 20(2):180–9, 2010.

[14] R. Blekhman, A. Oshlack, A. E. Chabot, G. K. Smyth, and Y. Gilad. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet*, 4(11):e1000271, 2008.

[15] John D Blischak, Ludovic Tailleux, Amy Mitrano, Luis B Barreiro, and Yoav Gilad. Mycobacterial infection induces a specific human innate immune response. *Scientific reports*, 5:16882, 2015.

[16] Jennifer C Boldrick, Ash a Alizadeh, Maximilian Diehn, Sandrine Dudoit, Chih Long Liu, Christopher E Belcher, David Botstein, Louis M Staudt, Patrick O Brown, and David a Relman. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):972–7, 2002.

[17] C. Borel, P. G. Ferreira, F. Santoni, O. Delaneau, A. Fort, K. Y. Popadin, M. Garieri, E. Falconnet, P. Ribaux, M. Guipponi, I. Padioleau, P. Carninci, E. T. Dermitzakis, and S. E. Antonarakis. Biased allelic expression in human primary fibroblast single cells. *American Journal of Human Genetics*, 96(1):70–80, 2015.

[18] K. Brennand, J. N. Savas, Y. Kim, N. Tran, A. Simone, K. Hashimoto-Torii, K. G. Beaumont, H. J. Kim, A. Topol, I. Ladran, M. Abdelrahim, B. Matikainen-Ankney, S. H. Chao, M. Mrksich, P. Rakic, G. Fang, B. Zhang, 3rd Yates, J. R., and F. H. Gage. Phenotypic differences in hipsc npcs derived from patients with schizophrenia. *Mol Psychiatry*, 20(3):361–8, 2015.

[19] P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods*, 10(11):1093–5, 2013.

[20] Penelope A. Bryant, Gordon K. Smyth, Travis Gooding, Alicia Oshlack, Zinta Harrington, Bart Currie, Jonathan R. Carapetis, Roy Robins-Browne, and Nigel Curtis. Susceptibility to acute rheumatic fever based on differential expression of genes involved in cytotoxicity, chemotaxis, and apoptosis. *Infection and immunity*, 82(2):753–61, 2014.

[21] C. M. Bulik, P. F. Sullivan, and K. S. Kendler. An empirical study of the classification of eating disorders. *Am J Psychiatry*, 157(6):886–95, 2000.

[22] M. Caceres, J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo, D. H. Geschwind, D. J. Lockhart, T. M. Preuss, and C. Barlow. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A*, 100(22):13030–5, 2003.

[23] C. E. Cain, R. Blekhman, J. C. Marioni, and Y. Gilad. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*, 187(4):1225–34, 2011.

[24] J. A. Capra, G. D. Erwin, G. McKinsey, J. L. Rubenstein, and K. S. Pollard. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci*, 368(1632):20130025, 2013.

[25] J. A. Casbon, R. J. Osborne, S. Brenner, and C. P. Lichtenstein. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res*, 39(12):e81, 2011.

[26] Nayeli Shantal Castrejón-Jiménez, Kahiry Leyva-Paredes, Juan Carlos Hernández-González, Julieta Luna-Herrera, and Blanca Estela García-Pérez. The role of autophagy in bacterial infections. *Bioscience trends*, 9(3):149–59, 2015.

[27] Damien Chaussabel, Roshanak Tolouei Semnani, Mary Ann McDowell, David Sacks, Alan Sher, and Thomas B. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 102(2):672–81, 2003.

[28] Guokai Chen, Daniel R Gulbranson, Zhonggang Hou, Jennifer M Bolin, Victor Ruotti, Mitchell D Probasco, Kimberly Smuga-Otto, Sara E Howden, Nicole R Diol, Nicholas E Propson, Ryan Wagner, Garrett O Lee, Jessica Antosiewicz-Bourget, Joyce M C Teng, and James a Thomson. Chemically defined conditions for human iPSC derivation and culture. *Nature methods*, 8(5):424–9, 2011.

[29] J. Chen, R. Swofford, J. Johnson, B. B. Cummings, N. Rogel, K. Lindblad-Toh, W. Haerty, F. D. Palma, and A. Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res*, 29(1):53–63, 2019.

[30] Emile R. Chimusa, Noah Zaitlen, Michelle Daya, Marlo Möller, Paul D van Helden, Nicola J Mulder, Alkes L. Price, and Eileen G. Hoal. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human molecular genetics*, 23(3):796–809, 2014.

[31] Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu. A non-degenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*, 78(4):685–709, 2013.

[32] Aurelie Cobat, Caroline J Gallant, Leah Simkin, Gillian F Black, Kim Stanley, Jane Hughes, T Mark Doherty, Willem a Hanekom, Brian Eley, Nulda Beyers, Jean-Philippe Jaïs, Paul van Helden, Laurent Abel, Eileen G Hoal, Alexandre Alcaïs, and Erwin Schurr. High heritability of antimycobacterial immunity in an area of hyperendemicity for tuberculosis disease. *The Journal of infectious diseases*, 201(1):15–9, 2010.

[33] G W Comstock. Tuberculosis in twins: a re-analysis of the Prophit survey. *The American review of respiratory disease*, 117(4):621–4, 1978.

[34] Mireilla Coscolla and Sebastien Gagneux. Does m. tuberculosis genomic diversity explain disease diversity? *Drug discovery today. Disease mechanisms*, 7(1):e43–e59, 2010.

[35] M. A. Crisafulli, A. Von Holle, and C. M. Bulik. Attitudes towards anorexia nervosa: the impact of framing on blame and stigma. *Int J Eat Disord*, 41(4):333–9, 2008.

[36] James Curtis, Yang Luo, Helen L Zenner, Delphine Cuchet-Lourenço, Changxin Wu, Kitty Lo, Mailis Maes, Ali Alisaac, Emma Stebbings, Jimmy Z Liu, Liliya Kopanitsa, Olga Ignatyeva, Yanina Balabanova, Vladyslav Nikolayevskyy, Ingelore Baessmann, Thorsten Thye, Christian G Meyer, Peter Nürnberg, Rolf D Horstmann, Francis Drobniewski, Vincent Plagnol, Jeffrey C Barrett, and Sergey Nejentsev. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nature genetics*, 47(5):523–7, 2015.

[37] Vojo Deretic. Autophagy in tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, 4(11):53–67, 2014.

[38] Z. Dezso, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. J. Brennan, A. Guryanov, K. Li, J. Blake, R. R. Samaha, and T. Nikolskaya. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol*, 6:49, 2008.

[39] B. Ding, L. Zheng, Y. Zhu, N. Li, H. Jia, R. Ai, A. Wildberg, and W. Wang. Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, 31(13):2225–7, 2015.

[40] Jun Ding, Johann E. Gudjonsson, Liming Liang, Philip E. Stuart, Yun Li, Wei Chen, Michael Weichenthal, Eva Ellinghaus, Andre Franke, William Cookson, Rajan P. Nair, James T. Elder, and Gonçalo R. Abecasis. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *American journal of human genetics*, 87(6):779–89, 2010.

[41] R. Drissen, N. Buza-Vidas, P. Woll, S. Thongjuea, A. Gambardella, A. Giustacchini, E. Mancini, A. Zriwil, M. Lutteropp, A. Grover, A. Mead, E. Sitnicka, S. E. Jacobsen, and C. Nerlov. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat Immunol*, 2016.

[42] Dan Du and Lei S Qi. An introduction to CRISPR technology for genome activation and repression in mammalian cells. *Cold Spring Harbor protocols*, 2016(1):pdb.top086835, 2016.

[43] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, 21(16):3439–40, 2005.

[44] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8):1184–91, 2009.

[45] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–10, 2002.

[46] S Ehrt, D Schnappinger, S Bekiranov, J Drenkow, S Shi, T R Gingeras, T Gaasterland, G Schoolnik, and C Nathan. Reprogramming of the macrophage transcriptome in response to interferon-gamma and Mycobacterium tuberculosis: signaling roles of nitric oxide synthase-2 and phagocyte oxidase. *The Journal of experimental medicine*, 194(8):1123–40, 2001.

[47] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–8, 1998.

[48] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, (June):btw354, 2016.

[49] S. Fehrmann, H. Bottin-Duplus, A. Leonidou, E. Mollereau, A. Barthelaix, W. Wei, L. M. Steinmetz, and G. Yvert. Natural sequence variants of yeast environmental sensors confer cell-to-cell expression variability. *Mol Syst Biol*, 9(1):695, 2013.

[50] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana MCElrath, Martin Prlic, and Peter S. Linsley. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(278):1–13, 2015.

[51] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS genetics*, 9(5):e1003486, 2013.

[52] J L Flynn, M M Goldstein, K J Triebold, B Koller, and B R Bloom. Major histocompatibility complex class I-restricted T cells are required for resistance to Mycobacterium

tuberculosis infection. *Proceedings of the National Academy of Sciences of the United States of America*, 89(24):12013–7, 1992.

[53] L. C. Fosburgh. Mack: an autobiography. *AANA J*, 64(6):583–6, 1996.

[54] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[55] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):2008–2010, 2010.

[56] G. K. Fu, J. Hu, P. H. Wang, and S. P. Fodor. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci U S A*, 108(22):9026–31, 2011.

[57] I. Gallego Romero, B. J. Pavlovic, I. Hernando-Herraez, X. Zhou, M. C. Ward, N. E. Banovich, C. L. Kagan, J. E. Burnett, C. H. Huang, A. Mitrano, C. I. Chavarria, I. Friedrich Ben-Nun, Y. Li, K. Sabatini, T. R. Leonardo, M. Parast, T. Marques-Bonet, L. C. Laurent, J. F. Loring, and Y. Gilad. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *Elife*, 4:e07103. doi: 10.7554/eLife.07103, 2015.

[58] Y. Gilad, S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res*, 15(5):674–80, 2005.

[59] Yoav Gilad and Orna Mizrahi-Man. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research*, 4(May 2015):121, 2015.

[60] Philippe Glaziou, Charalambos Sismanidis, Katherine Floyd, and Mario Raviglione. Global epidemiology of tuberculosis. *Cold Spring Harbor perspectives in medicine*, 5(2):445–461, 2015.

[61] Jeff E. Grotzke, Melanie J. Harriff, Anne C. Siler, Dawn Nolt, Jacob Delepine, Deborah A Lewinsohn, and David M. Lewinsohn. The Mycobacterium tuberculosis phagosome is a HLA-I processing competent organelle. *PLoS pathogens*, 5(4):e1000374, 2009.

[62] Jeff E Grotzke, Anne C Siler, Deborah A Lewinsohn, and David M Lewinsohn. Secreted immunodominant Mycobacterium tuberculosis antigens are processed by the cytosolic pathway. *Journal of immunology (Baltimore, Md. : 1950)*, 185(7):4336–43, 2010.

[63] D. Grn, L. Kester, and A. van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat Methods*, 11(6):637–40, 2014.

[64] D. Grn and A. van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

[65] A. E. Handel, S. Chintawar, T. Lalic, E. Whiteley, J. Vowles, A. Giustacchini, K. Argoud, P. Sopp, M. Nakanishi, R. Bowden, S. Cowley, S. Newey, C. Akerman, C. P. Ponting, and M. Z. Cader. Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. *Hum Mol Genet*, 25(5):989–1000, 2016.

[66] I. Hernando-Herraez, R. Garcia-Perez, A. J. Sharp, and T. Marques-Bonet. Dna methylation: Insights into human evolution. *PLoS Genet*, 11(12):e1005661, 2015.

[67] I. Hernando-Herraez, H. Heyn, M. Fernandez-Callejo, E. Vidal, H. Fernandez-Bellon, J. Prado-Martinez, A. J. Sharp, M. Esteller, and T. Marques-Bonet. The interplay between dna methylation and sequence divergence in recent human evolution. *Nucleic Acids Res*, 43(17):8204–14, 2015.

[68] I. Hernando-Herraez, J. Prado-Martinez, P. Garg, M. Fernandez-Callejo, H. Heyn, C. Hvilsom, A. Navarro, M. Esteller, A. J. Sharp, and T. Marques-Bonet. Dynamics of dna methylation in recent human and great ape evolution. *PLoS Genet*, 9(9):e1003763, 2013.

[69] Anne Lise K Hestvik, Zakaria Hmama, and Yossef Av-Gay. Mycobacterial manipulation of the host cell. *FEMS microbiology reviews*, 29(5):1041–50, 2005.

[70] Stephanie C Hicks, Mingxiang Teng, and Rafael A Irizarry. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-seq data. *bioRxiv*, 2015.

[71] Mathias W Hornef, Mary Jo Wick, Mikael Rhen, and Staffan Normark. Bacterial strategies for overcoming host innate and adaptive immune responses. *Nature immunology*, 3(11):1033–40, 2002.

[72] Q Huang, D Liu, P Majewski, L C Schulte, J M Korn, R a Young, E S Lander, and N Hacohen. The plasticity of dendritic cell responses to pathogens and their components. *Science (New York, N.Y.)*, 294(5543):870–5, 2001.

[73] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael a Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2):115–21, 2015.

[74] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[75] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*, 21(7):1160–7, 2011.

[76] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc*, 7(5):813–28, 2012.

[77] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(1):163–6, 2014.

[78] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–9, 2014.

[79] Richard G Jenner and Richard a Young. Insights into host responses against pathogens from transcriptional profiling. *Nature reviews. Microbiology*, 3(4):281–94, 2005.

[80] L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Res*, 21(9):1543–51, 2011.

[81] Joe Brown and Jay Hesselberth and John Blischak. umitools v2.1.1, 2015.

[82] NA Joshi and JN Fass. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33) [software]. *Available at https://github.com/najoshi/sickle*, 2011.

[83] Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics*, 91(5):839–48, 2012.

[84] Gerald Jurasinski, Franziska Koebsch, Anke Guenther, and Sascha Beetz. *flux: Flux rate calculation from dynamic closed chamber measurements*, 2014.

[85] Franz J Kallmann and David Reisner. Twin studies on genetic variations in resistance to tuberculosis. *Journal of Heredity*, 34(9):269–276, 1943.

[86] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(Database issue):D712–717, 2011.

[87] M. W. Karaman, M. L. Houck, L. G. Chemnick, S. Nagpal, D. Chawannakul, D. Sudano, B. L. Pike, V. V. Ho, O. A. Ryder, and J. G. Hacia. Comparative analysis of gene-expression patterns in human and african great ape cultured fibroblasts. *Genome Res*, 13(7):1619–30, 2003.

[88] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[89] Silvia N. Kariuki, John D. Blischak, Shigeki Nakagome, David B. Witonsky, and Anna Di Rienzo. Patterns of transcriptional response to 1,25-dihydroxyvitamin D3 and bacterial lipopolysaccharide in primary human monocytes. *G3 (Bethesda, Md.)*, 6(5):1345–55, 2016.

[90] P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–4, 2005.

[91] P. Khaitovich, B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H. H. Do, G. Weiss, W. Enard, F. Heissig, T. Arendt, K. Nieselt-Struwe, E. E. Eichler, and S. Paabo. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, 14(8):1462–73, 2004.

[92] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. A neutral model of transcriptome evolution. *PLoS Biol*, 2(5):E132, 2004.

[93] Nargis Khan, Aurobind Vidyarthi, Shifa Javed, and Javed N. Agrewala. Innate immunity holding the flanks until reinforced by adaptive immunity against Mycobacterium tuberculosis infection. *Frontiers in microbiology*, 7(MAR):328, 2016.

[94] Z. Khan, M. J. Ford, D. A. Cusanovich, A. Mitrano, J. K. Pritchard, and Y. Gilad. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, 342(6162):1100–4, 2013.

[95] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), 2013.

[96] K. T. Kim, H. W. Lee, H. O. Lee, S. C. Kim, Y. J. Seo, W. Chung, H. H. Eum, D. H. Nam, J. Kim, K. M. Joo, and W. Y. Park. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol*, 16(1):127, 2015.

[97] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–16, 1975.

[98] T. Kivioja, A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, 9(1):72–4, 2012.

[99] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–201, 2015.

[100] A. A. Kolodziejczyk, J. K. Kim, J. C. H. Tsang, T. Illicic, J. Henriksson, K. N. Natarajan, A. C. Tuke, X. Gao, M. Bhler, P. Liu, and J. C. Marioni. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 2015.

[101] Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2, 2012.

[102] Max Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[103] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

[104] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014.

[105] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[106] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael a Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, 11(10):733–9, 2010.

[107] B. Lemos, C. D. Meiklejohn, M. Caceres, and D. L. Hartl. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution*, 59(1):126–37, 2005.

[108] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, 2009.

[109] Jialong Liang, Wanshi Cai, and Zhongsheng Sun. Single-cell sequencing technologies: current and future. *Journal of genetics and genomics = Yi chuan xue bao*, 41(10):513–28, 2014.

[110] Yang Liao, Gordon K Smyth, and Wei Shi. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108, 2013.

[111] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–30, 2014.

[112] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[113] S. Lin, Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A*, 111(48):17224–9, 2014.

[114] Cecilia S Lindestam Arlehamn, David Lewinsohn, Alessandro Sette, and Deborah Lewinsohn. Antigens for CD4 and CD8 T cells in tuberculosis. *Cold Spring Harbor perspectives in medicine*, 4(7):89–103, 2014.

[115] C. Lindskog. The human protein atlas - an important resource for basic and clinical research. *Expert Rev Proteomics*, 13(7):627–9, 2016.

[116] Ruijie Liu, Aliaksei Z. Holik, Shian Su, Natasha Jansz, Kelan Chen, Huei San Leong, Marnie E. Blewitt, Marie-Liesse Asselin-Labat, Gordon K. Smyth, and Matthew E. Ritchie. Why weight? modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic acids research*, 43(15):e97, 2015.

[117] M. Logotheti, O. Papadodima, N. Venizelos, A. Chatziioannou, and F. Kolisis. A comparative genomic study in schizophrenic and in bipolar disorder patients, based on microarray expression profiling meta-analysis. *ScientificWorldJournal*, 2013:685917, 2013.

[118] I. C. Macaulay and T. Voet. Single cell genomics: advances and future perspectives. *PLoS Genet*, 10(1):e1004126, 2014.

[119] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–14, 2015.

[120] Jeroen Maertzdorf, Stefan H.E. Kaufmann, and January Weiner. Toward a unified biosignature for tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, 5(1):183–95, 2015.

[121] Surakameth Mahasirimongkol, Hideki Yanai, Taisei Mushiroda, Watoo Promphittayarat, Sukanya Wattanapokayakit, Jurairat Phromjai, Rika Yuliwulandari, Nuanjun Wichukchinda, Amara Yowang, Norio Yamada, Patcharee Kantipong, Atsushi Takahashi, Michiaki Kubo, Pathom Sawanpanyalert, Naoyuki Kamatani, Yusuke Nakamura, and Katsushi Tokunaga. Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *Journal of human genetics*, 57(6):363–7, 2012.

[122] Gwas Consortium Major Depressive Disorder Working Group of the Psychiatric, S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, A. C. Heath, F. Holsboer, S. Lucae, P. A. Madden, N. G. Martin, P. McGuffin, P. Muglia, M. M. Noethen, B. P. Penninx, M. L. Pergadia, J. B. Potash, M. Rietschel, D. Lin, B. Muller-Myhsok, J. Shi, S. Steinberg, H. J. Grabe, P. Lichtenstein, P. Magnusson, R. H. Perlis, M. Preisig, J. W. Smoller, K. Stefansson, R. Uher, Z. Kutalik, K. E. Tansey, A. Teumer, A. Viktorin, M. R. Barnes, T. Bettecken, E. B. Binder, R. Breuer, V. M. Castro, S. E. Churchill, W. H. Coryell, N. Craddock, I. W. Craig, D. Czamara, E. J. De Geus, F. Degenhardt, A. E. Farmer, M. Fava, J. Frank, V. S. Gainer, P. J. Gallagher, S. D. Gordon, S. Goryachev, M. Gross, M. Guipponi, A. K. Henders, S. Herms, I. B. Hickie, S. Hoefels, W. Hoogendijk, J. J. Hottenga, D. V. Iosifescu, M. Ising, I. Jones, L. Jones, T. Jung-Ying, J. A. Knowles, I. S. Kohane, M. A. Kohli, A. Korszun, M. Landen, W. B. Lawson, G. Lewis, D. Macintyre, W. Maier, M. Mattheisen, P. J. McGrath, A. McIntosh, A. McLean, C. M. Middeldorp, L. Middleton, G. M. Montgomery, S. N. Murphy, M. Nauck, W. A. Nolen, D. R. Nyholt, M. O'Donovan, H. Oskarsson, N. Pedersen, W. A. Scheftner, A. Schulz, T. G. Schulze, S. I. Shyn, E. Sigurdsson, S. L. Slager, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry*, 18(4):497–511, 2013.

[123] D. I. Martin, M. Singer, J. Dhahbi, G. Mao, L. Zhang, G. P. Schroth, L. Pachter, and D. Boffelli. Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res*, 21(12):2049–57, 2011.

[124] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science (New York, N.Y.)*, 342(6159):747–9, 2013.

[125] J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–9, 2012.

[126] George S Michaels, Daniel B Carr, M Askenazi, Stefanie Fuhrman, Xiling Wen, and Roland Somogyi. Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 3:42–53, 1998.

[127] M D Miller and M S Krangel. The human cytokine I-309 is a monocyte chemoattractant. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7):2950–4, 1992.

[128] D. T. Miyamoto, Y. Zheng, B. S. Wittner, R. J. Lee, H. Zhu, K. T. Broderick, R. Desai, D. B. Fox, B. W. Brannigan, J. Trautwein, K. S. Arora, N. Desai, D. M. Dahl, L. V. Sequist, M. R. Smith, R. Kapur, C. L. Wu, T. Shioda, S. Ramaswamy, D. T. Ting,

M. Toner, S. Maheswaran, and D. A. Haber. RNA-seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science*, 349(6254):1351–6, 2015.

[129] A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, and A. D. Smith. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–41, 2011.

[130] Marlo Möller and Eileen G. Hoal. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 90(2):71–83, 2010.

[131] Laura Muñoz, Helen R Stagg, and Ibrahim Abubakar. Diagnosis and management of latent tuberculosis infection. *Cold Spring Harbor perspectives in medicine*, 5(11):517–529, 2015.

[132] Gerard J Nau, Joan F L Richmond, Ann Schlesinger, Ezra G Jennings, Eric S Lander, and Richard a Young. Human macrophage activation programs induced by bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1503–8, 2002.

[133] B. M. Neale, S. E. Medland, S. Ripke, P. Asherson, B. Franke, K. P. Lesch, S. V. Faraone, T. T. Nguyen, H. Schafer, P. Holmans, M. Daly, H. C. Steinhausen, C. Freitag, A. Reif, T. J. Renner, M. Romanos, J. Romanos, S. Walitza, A. Warnke, J. Meyer, H. Palmason, J. Buitelaar, A. A. Vasquez, N. Lambregts-Rommelse, M. Gill, R. J. Anney, K. Langely, M. O'Donovan, N. Williams, M. Owen, A. Thapar, L. Kent, J. Sergeant, H. Roeyers, E. Mick, J. Biederman, A. Doyle, S. Smalley, S. Loo, H. Hakonarson, J. Elia, A. Todorov, A. Miranda, F. Mulas, R. P. Ebstein, A. Rothenberger, T. Banaschewski, R. D. Oades, E. Sonuga-Barke, J. McGough, L. Nisenbaum, F. Middleton, X. Hu, S. Nelson, and Gwas Consortium Adhd Subgroup Psychiatric. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*, 49(9):884–97, 2010.

[134] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Nobel, J. L. DeRisi, , and J. S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[135] Robert J North and Yu-Jin Jung. Immunity to tuberculosis. *Annual review of immunology*, 22:599–623, 2004.

[136] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A*, 106(52):22358–63, 2009.

[137] Anne O'Garra, Paul S. Redford, Finlay W. McNab, Chloe I. Bloom, Robert J. Wilkinson, and Matthew P R Berry. The immune response in tuberculosis. *Annual review of immunology*, 31(1):475–527, 2013.

[138] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From RNA-seq reads to differential expression results. *Genome biology*, 11(12):220, 2010.

[139] Fethi Ahmet Özdemir, Deniz Erol, Hüseyin Yüce, Vahit Konar, Ebru Kara enli, Funda Bulut, and Figen Deveci. [investigation of CCL1 rs159294 T/A gene polymorphism in pulmonary and extrapulmonary tuberculosis patients]. *Tuberkuloz ve toraks*, 61(3):200–8, 2013.

[140] A. A. Pai, J. T. Bell, J. C. Marioni, J. K. Pritchard, and Y. Gilad. A genome-wide study of dna methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet*, 7(2):e1001316, 2011.

[141] Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):Article 39, 2010.

[142] Simone Picelli, Asa K Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research*, 24(12):2033–40, 2014.

[143] E. Pierson, G. TEx Consortium, D. Koller, A. Battle, S. Mostafavi, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol*, 11(5):e1004220, 2015.

[144] Eileen Png, Bachti Alisjahbana, Edhyana Sahiratmadja, Sangkot Marzuki, Ron Nelwan, Yanina Balabanova, Vladyslav Nikolayevskyy, Francis Drobniewski, Sergey Nejentsev, Iskandar Adnan, Esther van de Vosse, Martin L Hibberd, Reinout van Crevel, Tom H M Ottenhoff, and Mark Seielstad. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC medical genetics*, 13(1):1–9, 2012.

[145] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, 2nd Kemp, D. W., M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, and J. A. West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol*, 32(10):1053–8, 2014.

[146] Valentina Proserpio and Bidesh Mahata. Single-cell technologies to study the immune system. *Immunology*, 147(2):133–40, 2016.

[147] Gwas Consortium Bipolar Disorder Working Group Psychiatric. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nat Genet*, 43(10):977–83, 2011.

[148] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*, 2015.

[150] Silvia Ragno, Maria Romano, Steven Howell, Darryl J.C. Pappin, Peter J. Jenner, and Michael J. Colston. Changes in gene expression in macrophages infected with Mycobacterium tuberculosis: a combined transcriptomic and proteomic approach. *Immunology*, 104(1):99–108, 2001.

[151] J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–3, 2005.

[152] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*, 32(9):896–902, 2014.

[153] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47, 2015.

[154] Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516, 2012.

[155] K. G. Sagaser, S. Shahrukh Hashmi, R. D. Carter, J. Lemons, H. Mendez-Figueroa, S. Nassef, B. Peery, and C. N. Singletary. Spiritual exploration in the prenatal genetic counseling session. *J Genet Couns*, 25(5):923–35, 2016.

[156] A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res*, 42(14):8845–60, 2014.

[157] Rahul Satija and Alex K. Shalek. Heterogeneity in immune responses: from populations to single cells. *Trends in immunology*, 35(5):219–29, 2014.

[158] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology*, 32(9):903–14, 2014.

[159] Kwonjune J Seung, Salmaan Keshavjee, and Michael L Rich. Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis. *Cold Spring Harbor perspectives in medicine*, 5(9):579–598, 2015.

[160] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–40, 2013.

[161] A. J. Sharp, E. Stathaki, E. Migliavacca, M. Brahmachary, S. B. Montgomery, Y. Dupre, and S. E. Antonarakis. Dna methylation profiles of human active and inactive x chromosomes. *Genome Res*, 21(10):1592–600, 2011.

[162] Y. Shibata, N. C. Sheffield, O. Fedrigo, C. C. Babbitt, M. Wortham, A. K. Tewari, D. London, L. Song, B. K. Lee, V. R. Iyer, S. C. Parker, E. H. Margulies, G. A. Wray, T. S. Furey, and G. E. Crawford. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet*, 8(6):e1002789, 2012.

[163] K. Shiroguchi, T. Z. Jia, P. A. Sims, and X. S. Xie. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A*, 109(4):1347–52, 2012.

[164] Jonathan Kevin Sia, Maria Georgieva, and Jyothi Rengarajan. Innate immune defenses in human tuberculosis: An overview of the interactions between Mycobacterium tuberculosis and innate immune cells. *Journal of immunology research*, 2015:747543, 2015.

[165] Tom Sean Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification. *bioRxiv*, 2016.

[166] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[167] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):Article3, 2004.

[168] Gordon K. Smyth, Joëlle Michaud, and Hamish S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics (Oxford, England)*, 21(9):2067–75, 2005.

[169] Rafal S. Sobota, Catherine M. Stein, Nuri Kodaman, Laura B. Scheinfeldt, Isaac Maro, Wendy Wieland-Alter, Robert P. Igo, Albert Magohe, Lashaunda L. Malone, Keith Chervenak, Noemi B. Hall, Chawangwa Modongo, Nicola Zetola, Mecky Matee, Moses Joloba, Alain Froment, Thomas B. Nyambo, Jason H. Moore, William K. Scott, Timothy Lahey, W. Henry Boom, C Fordham von Reyn, Sarah A. Tishkoff, Giorgio Sirugo, and Scott M. Williams. A locus at 5q33.3 confers resistance to tuberculosis in highly susceptible individuals. *American journal of human genetics*, 98(3):514–24, 2016.

[170] Giovanni Sotgiu, Rosella Centis, Lia D'ambrosio, and Giovanni Battista Migliori. Tuberculosis treatment and drug regimens. *Cold Spring Harbor perspectives in medicine*, 5(5):505–516, 2015.

[171] Natalie Spang, Anne Feldmann, Heike Huesmann, Fazilet Bekbulat, Verena Schmitt, Christof Hiebel, Ingrid Koziollek-Drechsler, Albrecht M Clement, Bernd Moosmann,

Jennifer Jung, Christian Behrends, Ivan Dikic, Andreas Kern, and Christian Behl. RAB3GAP1 and RAB3GAP2 modulate basal and rapamycin-induced autophagy. *Autophagy*, 10(12):2297–309, 2014.

[172] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, 16(3):133–45, 2015.

[173] Jr. Stephens, Jonathan. *Prostitutes in the pews*. 2016.

[174] Matthew Stephens. False discovery rates: A new deal. *bioRxiv*, 2016.

[175] A. B. Stergachis, S. Neph, R. Sandstrom, E. Haugen, A. P. Reynolds, M. Zhang, R. Byron, T. Canfield, S. Stelhing-Sun, K. Lee, R. E. Thurman, S. Vong, D. Bates, F. Neri, M. Diegel, E. Giste, D. Dunn, J. Vierstra, R. S. Hansen, A. K. Johnson, P. J. Sabo, M. S. Wilken, T. A. Reh, P. M. Treuting, R. Kaul, M. Groudine, M. A. Bender, E. Borenstein, and J. A. Stamatoyannopoulos. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–70, 2014.

[176] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003.

[177] S Sturgill-Koszycki, P H Schlesinger, P Chakraborty, P L Haddix, H L Collins, A K Fok, R D Allen, S L Gluck, J Heuser, and D G Russell. Lack of acidification in Mycobacterium phagosomes produced by exclusion of the vesicular proton-ATPase. *Science (New York, N.Y.)*, 263(5147):678–81, 1994.

[178] Thomas C Sudhof. The synaptic vesicle cycle. *Annual review of neuroscience*, 27(6533):509–47, 2004.

[179] Dinanath Sulakhe, Bingqing Xie, Andrew Taylor, Mark D'Souza, Sandhya Balasubramanian, Somaye Hashemifar, Steven White, Utpal J Dave, Gady Agam, Jinbo Xu, Sheng Wang, T Conrad Gilliam, and Natalia Maltsev. Lynx: a knowledge base and an analytical workbench for integrative medicine. *Nucleic acids research*, 44(D1):D882–7, 2016.

[180] F. Supek, M. Bosnjak, N. Skunca, and T. Smuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, 2011.

[181] Ludovic Tailleux, Simon J Waddell, Mattia Pelizzola, Alessandra Mortellaro, Michael Withers, Antoine Tanne, Paola Ricciardi Castagnoli, Brigitte Gicquel, Neil G Stoker, Philip D Butcher, Maria Foti, and Olivier Neyrolles. Probing host pathogen cross-talk by transcriptional profiling of both Mycobacterium tuberculosis and infected human dendritic cells and macrophages. *PloS one*, 3(1):e1403, 2008.

[182] N. L S Tang, C. Y. Chan, C. C. Leung, C. M. Tam, and J. Blackwell. Tuberculosis susceptibility genes in the chemokine cluster region of chromosome 17 in hong kong chinese. *Hong Kong medical journal = Xianggang yi xue za zhi / Hong Kong Academy of Medicine*, 17 Suppl 6(6):22–5, 2011.

[183] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[184] The International Hapmap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[185] D. M. Thomas, J. E. Navarro-Barrientos, D. E. Rivera, S. B. Heymsfield, C. Bredlau, L. M. Redman, C. K. Martin, S. A. Lederman, M. Collins L, and N. F. Butte. Dynamic energy-balance model predicting gestational weight gain. *Am J Clin Nutr*, 95(1):115–22, 2012.

[186] PD Thomas, A Kejariwal, N Guo, HY Mi, MJ Campbell, A Muruganujan, and B Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mrna/protein expression analysis and coding snp scoring tools. *Nucleic Acids Research*, 34:W645–W650, 2006.

[187] P. J. Thul and C. Lindskog. The human protein atlas: A spatial map of the human proteome. *Protein Sci*, 27(1):233–244, 2018.

[188] Nguyen Thuy Thuong Thuong, Sarah J Dunstan, Tran Thi Hong Chau, Vesteinn Thorsson, Cameron P Simmons, Nguyen Than Ha Quyen, Guy E Thwaites, Nguyen Thi Ngoc Lan, Martin Hibberd, Yik Y Teo, Mark Seielstad, Alan Aderem, Jeremy J Farrar, and Thomas R Hawn. Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. *PLoS pathogens*, 4(12):e1000229, 2008.

[189] Thorsten Thye, Ellis Owusu-Dabo, Fredrik O Vannberg, Reinout van Crevel, James Curtis, Edhyana Sahiratmadja, Yanina Balabanova, Christa Ehmen, Birgit Muntau, Gerd Ruge, Jürgen Sievertsen, John Gyapong, Vladyslav Nikolayevskyy, Philip C Hill, Giorgio Sirugo, Francis Drobniewski, Esther van de Vosse, Melanie Newport, Bachti Alisjahbana, Sergey Nejentsev, Tom H M Ottenhoff, Adrian V S Hill, Rolf D Horstmann, and Christian G Meyer. Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nature genetics*, 44(3):257–9, 2012.

[190] Thorsten Thye, Fredrik O Vannberg, Sunny H Wong, Ellis Owusu-Dabo, Ivy Osei, John Gyapong, Giorgio Sirugo, Fatou Sisay-Joof, Anthony Enimil, Margaret a Chinbuah, Sian Floyd, David K Warndorff, Lifted Sichali, Simon Malema, Amelia C Crampin, Bagrey Ngwira, Yik Y Teo, Kerrin Small, Kirk Rockett, Dominic Kwiatkowski, Paul E Fine, Philip C Hill, Melanie Newport, Christian Lienhardt, Richard a Adegbola, Tumani Corrah, Andreas Ziegler, African TB Genetics Consortium, Wellcome Trust Case Control Consortium, Andrew P Morris, Christian G Meyer, Rolf D Horstmann, and Adrian V S Hill. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nature genetics*, 42(9):739–41, 2010.

[191] A. M. Tsankov, H. Gu, V. Akopian, M. J. Ziller, J. Donaghey, I. Amit, A. Gnirke, and A. Meissner. Transcription factor binding dynamics during human es cell differentiation. *Nature*, 518(7539):344–9, 2015.

[192] Po-Yuan Tung, John D Blischak, Chiaowen Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *bioRxiv*, page 062919, 2016.

[193] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[194] C. A. Vallejos, J. C. Marioni, and S. Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333, 2015.

[195] D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, and D. T. Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–66, 2015.

[196] I. Virgolini. Mack forster award lecture. receptor nuclear medicine: vasointestinal peptide and somatostatin receptor scintigraphy for diagnosis and treatment of tumour patients. *Eur J Clin Invest*, 27(10):793–800, 1997.

[197] Elisabetta Volpe, Giulia Cappelli, Manuela Grassi, Angelo Martino, Annalucia Serafino, Vittorio Colizzi, Nunzia Sanarico, and Francesca Mariani. Gene expression profiling of human macrophages at late time of infection with Mycobacterium tuberculosis. *Immunology*, 118(4):449–60, 2006.

[198] Charles C Wang, Bingdong Zhu, Xionglin Fan, Brigitte Gicquel, and Ying Zhang. Systems approach to tuberculosis vaccine development. *Respirology (Carlton, Vic.)*, 18(3):412–20, 2013.

[199] M. C. Ward, M. D. Wilson, N. L. Barbosa-Morais, D. Schmidt, R. Stark, Q. Pan, P. C. Schwalie, S. Menon, M. Lukk, S. Watt, D. Thybert, C. Kutter, K. Kirschner, P. Flicek, B. J. Blencowe, and D. T. Odom. Latent regulatory potential of human-specific repetitive elements. *Mol Cell*, 49(2):262–72, 2013.

[200] M. Warnefors and A. Eyre-Walker. A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. *PLoS One*, 7(4):e34935, 2012.

[201] Yingying Wei, Toyoaki Tenzen, and Hongkai Ji. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics (Oxford, England)*, 16(1):31–46, 2015.

[202] World Health Organization. Global TB facts 2015. 2015.

[203] World Health Organization. Global tuberculosis report 2015. 2015.

[204] A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke, and S. R. Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods*, 11(1):41–6, 2014.

[205] Kang Wu, Dandan Dong, Hai Fang, Florence Levillain, Wen Jin, Jian Mei, Brigitte Gicquel, Yanzhi Du, Kankan Wang, Qian Gao, Olivier Neyrolles, and Ji Zhang. An interferon-related signature in the transcriptional core response of human macrophages to Mycobacterium tuberculosis infection. *PloS one*, 7(6):e38367, 2012.

[206] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. Ensembl 2016. *Nucleic acids research*, 44(D1):D710–6, 2016.

[207] N. Y. Yu, B. M. Hallstrom, L. Fagerberg, F. Ponten, H. Kawaji, P. Carninci, A. R. Forrest, Consortium Fantom, Y. Hayashizaki, M. Uhlen, and C. O. Daub. Complementing tissue characterization by integrating transcriptome profiling from the human protein atlas and from the fantom5 consortium. *Nucleic Acids Res*, 43(14):6787–98, 2015.

[208] A. Zeileis and G. Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005.

[209] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4:Article17, 2005.

[210] X. Zhou, C. E. Cain, M. Myrthil, N. Lewellen, K. Michelini, E. R. Davenport, M. Stephens, J. K. Pritchard, and Y. Gilad. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol*, 15(12):547, 2014.