THE UNIVERSITY OF CHICAGO

INVESTIGATING SUSCEPTIBILITY TO TUBERCULOSIS USING FUNCTIONAL GENOMICS APPROACHES

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

COMMITTEE ON GENETICS

BY
JOHN D. BLISCHAK

CHICAGO, ILLINOIS
DECEMBER 2016

Table of Contents

LIST OF FIGURES vi

LIST OF TABLES viii

ACKNOWLEDGMENTS ix

ABSTRACT xi

1 INTRODUCTION 1
1.1 Human Genetics and the search for the genetic basis of human phenotypes . 1
1.2 Functional genomics and the investigation of the non-coding regions of the genome 4
1.3 Tuberculosis and the genetic basis of susceptibility 7
1.4 Single cell sequencing technology and the future of functional genomics . . . 11

2 MYCOBACTERIAL INFECTION INDUCES A SPECIFIC HUMAN INNATE IMMUNE RESPONSE 13
2.1 Abstract 13
2.2 Introduction 13
2.3 Results 17
2.3.1 Bacterial infection induces large changes in gene expression 17
2.3.2 Joint analysis identifies bacteria-specific response genes 18
2.3.3 Infection-induced response eQTLs are shared across bacterial infections 24
2.4 Discussion 26
2.4.1 Bayesian analysis identified mycobacteria-specific response genes . . . 26
2.4.2 Little evidence for strain-specific transcriptional response to infection 31
2.4.3 Differences in response to virulent versus attenuated pathogens are not mycobacteria-specific 32
2.4.4 Previously identified response eQTLs affect response to bacterial infection in general 34
2.4.5 Conclusions 34
2.5 Methods 35
2.5.1 Ethics Statement 35
2.5.2 Sample collection and macrophage differentiation 35
2.5.3 Bacterial infection 36
2.5.4 RNA extraction, library preparation, and sequencing 36
2.5.5 Mapping, counting, and normalization 37
2.5.6 Differential expression analysis 37
2.5.7 Analysis using previously identified response eQTLs 39
2.5.8 Data and code availability 40
2.6 Acknowledgments 40
2.7 Author Contributions 40

2.8 Supplementary Information . . . 40
2.8.1 Supplementary Figures . . . 41
2.8.2 Supplementary Tables . . . 52

3 PREDICTING SUSCEPTIBILITY TO TUBERCULOSIS BASED ON GENE EXPRESSION PROFILING . . . 54
3.1 Abstract . . . 54
3.2 Introduction . . . 54
3.3 Results . . . 56
3.3.1 Susceptible individuals have an altered transcriptome in the non-infected state . . . 56
3.3.2 Differentially expressed genes are enriched with TB susceptibility loci 57
3.3.3 Susceptibility status can be predicted based on gene expression data . 59
3.4 Discussion . . . 60
3.5 Methods . . . 65
3.5.1 Ethics Statement . . . 65
3.5.2 Sample collection . . . 65
3.5.3 Isolation and infection of dendritic cells . . . 65
3.5.4 RNA extraction and sequencing . . . 66
3.5.5 Read mapping . . . 66
3.5.6 Quality control . . . 66
3.5.7 Differential expression analysis . . . 67
3.5.8 Combined analysis of gene expression data and GWAS results . . . 68
3.5.9 Classifier . . . 69
3.5.10 Software implementation . . . 70
3.5.11 Data availability . . . 71
3.6 Acknowledgments . . . 71
3.7 Author Contributions . . . 71
3.8 Supplementary Information . . . 72
3.8.1 Supplementary Figures . . . 72
3.8.2 Supplementary Data . . . 84

4 BATCH EFFECTS AND THE EFFECTIVE DESIGN OF SINGLE-CELL GENE EXPRESSION STUDIES . . . 86
4.1 Introduction . . . 86
4.2 Results . . . 89
4.2.1 Study design and quality control . . . 89
4.2.2 Batch effects associated with UMI-based single cell data . . . 92
4.2.3 Measuring regulatory noise in single-cell gene expression data . . . 97
4.3 Discussion . . . 100
4.3.1 Study design and sample size for scRNA-seq . . . 100
4.3.2 The limitations of the ERCC spike-in controls . . . 103
4.3.3 Outlook . . . 104
4.4 Methods . . . 104

4.4.1 Ethics statement . . . 104
4.4.2 Cell culture of iPSCs . . . 105
4.4.3 Single cell capture and library preparation . . . 105
4.4.4 Illumina high-throughput sequencing . . . 106
4.4.5 Read mapping . . . 107
4.4.6 Filtering cells and genes . . . 108
4.4.7 Calculate the input molecule quantities of ERCC spiked-ins . . . 109
4.4.8 Subsampling . . . 109
4.4.9 A framework for testing individual and batch effects . . . 110
4.4.10 Estimating variance components for per-gene expression levels . . . 111
4.4.11 Normalization . . . 112
4.4.12 Removal of technical batch effects . . . 113
4.4.13 Measurement of gene expression noise . . . 114
4.4.14 Identification of genes associated with inter-individual differences in regulatory noise . . . 115
4.4.15 Gene enrichment analysis . . . 116
4.4.16 Individual assignment based on scRNA-seq reads . . . 116
4.4.17 Data and code availability . . . 117
4.5 Acknowledgments . . . 117
4.6 Author Contributions . . . 117
4.7 Supplementary Information . . . 118
4.7.1 Supplementary Figures . . . 118
4.7.2 Supplementary Tables . . . 127

5 CONCLUSION . . . 130
5.1 A joint Bayesian model provides a general framework for analyzing functional genomics studies with many conditions . . . 130
5.2 Initial success classifying individuals susceptible to tuberculsosis and future directions . . . 135
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 137
5.4 The importance of mitigating batch effects in any genomics experiment . . . 139
5.5 Concluding remarks . . . 141

REFERENCES . . . 142

List of Figures

2.1 Differential expression analysis. . . . 19
2.2 Joint Bayesian analysis. . . . 22
2.3 Joint Bayesian analysis - 18 hours post-infection. . . . 23
2.4 Joint Bayesian analysis - 48 hours post-infection. . . . 25
2.5 Response eQTLs at 18 hours post-infection. . . . 27
2.6 Study design. . . . 41
2.7 Principal components analysis (PCA) of uncorrected and batch-corrected expression values. . . . 42
2.8 Joint Bayesian analysis with 14 expression patterns. . . . 43
2.9 Expression of genes involved in phagosome maturation. . . . 45
2.10 Expression of genes involved in vitamin D signaling. . . . 46
2.11 Expression of *DUSP14* at 18 hours post-infection. . . . 47
2.12 Little difference in transcriptional response to infections with different MTB strains. 48
2.13 Response of example cytokines to infection with different MTB strains. . . . 49
2.14 Distribution of the number of exonic reads and RNA quality scores (RIN) across variables of interest. . . . 50
2.15 Comparison to Tailleux et al., 2008. . . . 51

3.1 Differential expression analysis. . . . 58
3.2 Comparison of differential expression and The Gambia GWAS results. . . . 59
3.3 Classifying TB susceptible individuals using a support vector machine model. . 61
3.4 Batch processing. . . . 72
3.5 Gene expression distributions before and after filtering genes and samples. . . . 73
3.6 Heatmap of correlation matrix of samples. . . . 74
3.7 Heatmap of correlation matrix after removing outliers. . . . 75
3.8 Principal components analysis (PCA) to identify outliers. . . . 76
3.9 Check for technical batch effects using principal components analysis (PCA). . . 77
3.10 Check for confounding effect of infection batch. . . . 78
3.11 Effect of treatment with MTB. . . . 79
3.12 Comparison of differential expression and Ghana GWAS results. . . . 80
3.13 Normalizing gene expression distributions. . . . 80
3.14 Principal components analysis (PCA) of combined data sets. . . . 81
3.15 Comparing the classification results of different methods and number of input genes. . . . 81
3.16 Classifying TB susceptible individuals using an elastic net model. . . . 82
3.17 Classifying TB susceptible individuals using a random forest model. . . . 82
3.18 Comparing gene expression between the two studies. . . . 83

4.1 Experimental design and quality control of scRNA-seq. . . . 90
4.2 The effect of sequencing depth and cell number on single cell UMI estimates. . . 93
4.3 Batch effect of scRNA-seq data using the C1 platform. . . . 96
4.4 Normalization and removal of technical variability. . . . 98

4.5 Cell-to-cell variation in gene expression. 101
4.6 Removal of low quality samples. 119
4.7 Removal of samples with multiple cells. 120
4.8 Sources of cell-to-cell variance in per-gene expression profile. 121
4.9 The gene-specific dropout rate. 121
4.10 Permutation-based P-value. 122
4.11 Inter-individual differences in regulatory noise. 123
4.12 Cell-to-cell variation of pluripotency genes. 124
4.13 Proposed study design for scRNA-seq using C1 platform. 125
4.14 The proportion of genes detected in single cell samples. 125
4.15 Coefficients of variation (CV) before and after adjusting for gene mean abundance. 126

List of Tables

2.1 Description of bacteria. 17
2.2 Gene expression matrix. 52
2.3 Differential expression results. 52
2.4 Joint Bayesian analysis results. 52
2.5 Joint Bayesian analysis results with gene descriptions. 53
2.6 Gene ontology results. 53
2.7 RNA quality. 53
2.8 Number of differentially expressed genes from intersecting gene lists. 53
2.9 Number of differentially expressed genes from pairwise tests. 53
2.10 Concordance in direction of effect for genes in each expression pattern. 53

3.1 Sample information. 84
3.2 Gene expression matrix. 84
3.3 Differential expression results. 85
3.4 Data for combined analysis of gene expression data and GWAS results. 85
3.5 Classifier results. 85

4.1 Data collection. 128
4.2 High quality single cell samples. 129
4.3 Genes associated with inter-individual differences in regulatory noise. 129
4.4 Gene ontology analysis of the genes associated with inter-individual differences in regulatory noise. 129

ACKNOWLEDGMENTS

I am truly grateful for all the assistance and guidance I have received during my PhD studies. None of my projects would have been feasible without the help of my collaborators.

I am very thankful to my advisor, Yoav Gilad. Not only has he had the largest influence on the development of my scientific thinking, but just as important, he made my graduate school experience enjoyable. He found the right balance of letting me explore ideas but also reminding me to stay on track. I could not have asked for a better PhD advisor.

I am thankful to my committee members. They have also been instrumental in my path to graduation. My chair, Joe Thornton, consistently reminded me to remember the "big picture" when describing my research, a very important lesson since I have the tendency to focus on the details. Matthew Stephens has always had great patience when explaining statistical concepts to me, from the basic to the advanced. John Novembre was always ready with a great piece of advice to improve my projects (e.g. the figures in Chapter 2 are much more interpretable after incorporating his suggestions).

I really enjoyed being a member of the Gilad lab. I learned so much from my labmates, past and present, and we also had a lot of fun over the years. I have to specifically thank Darren Cusanovich and Irene Gallego Romero, who invested a lot of time training me how to think like a scientist when I was a junior graduate student.

Even more broadly, I have greatly benefited from the Human Genetics community in Cummings Life Science Center. From the formal collaborations of TAing with Mark Abney and doing a project with Silvia Kariuki and Anna Di Rienzo, to the informal interactions with other professors, postdocs, and graduate students. I am so grateful to have had the opportunity to spend my PhD years in such a collegial environment.

My projects on tuberculosis would not have been possible without the help of Ludovic Tailleux and Luis Barriero. I am thankful to Ludo for expertly performing the many bacterial infections required for our studies, and for all his work handling IRB applications and patient

recruitments. I am thankful to Luis for his insightful advice and enthusiasm.

The single cell sequencing project was also a group effort. I am thankful for PoYuan Tung for expertly performing all the experimental work and her extensive knowledge of the single cell literature, for Joyce Hsiao for her ability to tackle tough statistical problems, and to Dave Knowles and Jonathan Pritchard for their insightful advice.

I was very fortunate to have made some great friends during my time at University of Chicago. I wasn't expecting to have so much fun during my PhD. I am going to really miss them as we all graduate and move on to other things.

My family has been super supportive of me my whole life. I am especially appreciative of my parents. They've supported me even during those times I was not enjoying graduate school.

Lastly, I cannot thank my wife enough. She has made so many sacrifices so that I could complete my research and earn my PhD. None of this would have been possible without her.

ABSTRACT

A major goal of human genetics is to characterize the role of genetic variation on complex, polygenic phenotypes. With the discovery from genome-wide association studies (GWAS) that many associated variants have a small effect size and are located in non-coding regions of the genome, there has been a large effort to collect functional genomics data. The hope is that a better understanding of how the genome functions in diverse developmental states and environments will provide insight into the context-specific activity of associated non-coding variants. My research applies this paradigm to the complex phenotype of susceptibility to develop tuberculosis (TB). It has been estimated that 10% of individuals infected with *Mycobacterium tuberculosis* (MTB) progress to active disease. Despite being heritable, very few genetic variants have been associated with susceptibility to TB. For my studies, I use RNA sequencing (RNA-seq) to interrogate genome-wide transcript levels in *in vitro* cellular models. In Chapter 2, I use a joint Bayesian model to idenitfy genes which are differentially expressed in macrophages only after infection with MTB and related mycobacteria, but not other bacterial pathogens. In Chapter 3, I build a support vector machine model to classify individuals as susceptible or resistant to TB based on the gene expression levels in their dendritic cells. In Chapter 4, I characterize the technical variation introduced by batch processing of single cell RNA-seq (scRNA-seq) and propose an effective study design that accounts for technical variation while minimizing replication. In addition to providing insight into the genes important for the innate immune response to MTB infection, my work is informative for the design and analysis of future functional genomics experiments.

CHAPTER 1

INTRODUCTION

1.1 Human Genetics and the search for the genetic basis of human phenotypes

The field of Human Genetics aims to discover the genetic basis of the variation observed in human phenotypes [288]. The difficulty of this goal depends on the genetic architecture of the phenotype [66, 26, 262]. On the one hand are monogenic (or Mendelian) traits, which are caused by mutations in a single gene. Isolating the causal gene is relatively tractable. On the other hand are polygenic (or complex) traits, which are the result of many genes acting in concert. Furthermore, the genes can interact with each other in a non-additive manner (gene-gene interactions, or epistasis) [55, 204] and the environment can also play a significant role (gene-environment interactions) [299, 226].

As an example, consider human height. A disabling mutation in just one gene, the growth hormone receptor (GHR), nullifies the effect of growth hormone leading to very short stature and other metabolic abnormalities (Laron Syndrome) [174]. Because of its easily identifiable phenotype and single gene origin, the genetic basis of Laron Syndrome was discovered in the late 1980s using a candidate gene approach in a small number of pedigrees [8, 102, 268] (the correct candidate gene was known from previous physiology experiments [85]). In contrast, the considerable variability in height in the human population is not caused by rare mutations in a single or a few genes, but instead is due to the aggregate effect of many mutations (or variants) of small effect size interacting with each other and the environment (e.g. diet, pollution, etc.) [179, 307]. Thus although height has been determined to be highly heritable, genetic studies involving hundreds of thousands of individuals that identified thousands of associated variants which affect height still only explain a small percentage of the heritable variation [170, 330]. Larger studies with increased power will only continue to find associated

variants with even smaller effect size (or otherwise they would have already been discovered), thus the genetic basis of a highly polygenic trait subject to millennia of natural selection may unsatisfyingly be that most of the genome makes a minute contribution to the height of an individual.

The current state-of-the-art technique for mapping genetic variants that affect a polygenic trait is the genome-wide association study (GWAS) [118]. This technique was made possible by the sequencing of the human genome [167, 314, 130, 168] and the cataloging of the common genetic variation segregating in the human population (the latter done via the International HapMap Project [128, 129, 127] and 1000 Genomes Project [1, 2]). For a GWAS, individuals are phenotyped (e.g. height is measured) and genotyped at millions of common variants, referred to as single nucleotide polymorphisms (SNPs). Then each SNP is tested individually for an association with the trait measurements via a linear regression or related statistical technique [16, 287, 339]. Similarly, for a binary trait such as cases with a disease versus controls without a disease, the phenotype is the presence or absence of a disease and each SNP is tested for association with a logistic regression or related statistical technique [42]. GWAS have identified many genetic variants affecting a diverse set of human polygenic traits, especially as the sample sizes for GWAS increased into the hundreds of thousands [327]. Nevertheless, their results have several limitations.

As mentioned above, one of the main issues with GWAS results is the small effect size of the associated SNPs on the trait of interest [199]. The hope of finding these SNPs is that they will be useful for predicting the trait (e.g. how likely are you to develop diabetes). However, with such small effect sizes, they have little predictive power and thus are generally not clinically actionable [303, 333]. These disappointing results could be due to limitations in our knowledge when designing the study and modeling the data. For example, when recruiting study participants, it is impossible to record every possible environmental factor that could have contributed to each persons trait value [226]. Furthermore, in case-control

studies, the controls will likely include a subset of individuals that have yet to develop the disease. Similarly, when modeling the genetic associations, most models assume an isolated additive effect of each variant on the trait [55]. This simplifying assumption is made such that the statistical test is tractable and interpretable. However, it is missing the contribution of any gene-gene or gene-environment interactions [55, 299, 173, 3, 125, 284]. On the other hand, the disappointing results of GWAS may not be due to limitations of the approach, but simply reflect the actual biology of polygenic traits [66, 26, 262, 241]. Mutations with strong effect, such as those that disable the GHR and cause Laron Syndrome, are often disruptive to the complex network of biochemical reactions that sustain a living individual. For this reason, they face strong negative selection and are often rare in the population. In contrast, mutations with small effect on a trait are more likely to be neutral or slightly favorable, and thus are able to rise to higher allele frequencies in the population. Over millions of years of evolution, the many variants of small effect could give rise to the large variation in phenotypes observed today, e.g. the difference in height between a 5 foot person and a 7 foot person. Supporting this view, when all SNPs assayed in an experiment are used to explain heritability, known as the "chip" heritability, this estimate is closer to the observed heritability (this has been demonstrated for height and other polygenic traits) [338, 177]. This suggests that highly polygenic traits like human height are indeed the result of thousands of variants of small effect size [330].

Beyond the ability to predict a disease outcome or trait value, another goal of GWAS is to elucidate the underlying biological mechanisms which ultimately determine the trait. This has proven difficult because most GWAS hits do not affect the protein-coding sequence of a gene, for which it would be straightforward to predict and test the effect this would have on gene function, but instead the associated SNPs are located in non-coding regions of the genome [117, 199]. It is much more difficult to predict the effect of these variants because there is no simple code to translate changes in non-coding sequence. This has motivated the

study of gene regulation in the field of Human Genetics [172, 304, 49, 171].

1.2 Functional genomics and the investigation of the non-coding regions of the genome

Gene regulation refers to how cells control which genes are turned on and to what extent [61, 217, 288, 230, 305]. This is critical because all cells in the human body contain the same genomic material (ignoring the complications of somatic recombination in certain immune cells and somatic mutations in general). Thus in order for a liver cell to function differently than a skin cell, the two cells must have different gene expression levels. These gene regulatory differences are established during development as an organism grows from an initial single cell. Signaling molecules, initially from the mother but subsequently produced by the offspring's cells, bind to the receptors of a cell to initiate signal transduction cascades that ultimately lead to activation of transcription factors which bind to DNA at their degenerate binding sites across the genome to modulate the expression of many genes. As development continues and cells differentiate into their final tissue type, the gene expression levels are maintained by the gene regulatory network established by the transcription factors active in that cell type.

Just as differences in gene regulation generate extreme diversity in cellular function among cells with identical genomes in a single organism [217], a long standing hypothesis is that differences among humans and the differences between humans and our closest evolutionary relatives, the great apes, are due to mutations that affect not the protein-coding sequence but instead mutations which affect the spatiotemporal expression of genes [30, 159, 36]. This theory was originally proposed because of the high similarity of protein-coding sequences between humans and chimpanzees [159], and is supported by the finding of mainly non-coding SNPs from GWAS [327].

Understanding which transcription factors establish and maintain a given cellular identity

is quite difficult [313, 23, 342]. However, even without this knowledge, it is possible to learn about the regulatory state of a given cell type [119]. First, it is possible to measure genome-wide gene expression levels using technologies like microarrays or RNA sequencing (RNA-seq; described in more detail below) [322, 229, 318]. Second, it is possible to interrogate the non-coding regions of the genome by measuring chromatin marks [234, 169]. Chromatin marks are deposited by chromatin-remodeling enzymes which are recruited by the transcription factors active in the cell. The most common are methylation of the cysteine base in CpG dinucleotides (DNA methylation) or chemical modification of the tails of the protein octamers (histones) which DNA is wrapped around. These marks signal the state of the region, e.g. active or repressed, and may help to maintain the current state. Histone marks can be assayed with chromatin immunoprecipitation followed by sequencing (ChIP-seq), and DNA methylation can be assayed with specialized microarrays or bisulphite sequencing. Using these technologies, it is possible to learn about the function of the non-coding SNPs discovered by GWAS.

As an aside, it should be noted that there is a lot of confusion about the role of chromatin marks and their effect on gene expression [112]. Chromatin marks are not causal. Instead, they are signs of a given chromatin state, and at best help maintain that state. As an analogy, consider viewing a stretch of highway from a helicopter. If you observe orange signs and barrels, you can conclude that this section of the highway is a construction zone. Furthermore, because they notify the motorists to slow down and to merge into one lane, you can conclude that the construction signs and barrels help this section to maintain the characteristics of a construction zone. However, you would not conclude that the signs and barrels caused this section of highway to be a construction zone. The decision to work on this section of road was made by local government officials and contractors after observing the conditions of the road and receiving complaints from citizens. In gene regulation, the chromatin marks are the construction signs and barrels. If you observe activating chromatin

marks, you can conclude that the nearby gene is expressed and that the chromatin marks are helping maintain this transcriptional activity. However, it is the result of transcription factors receiving input from outside the cell that caused these active chromatin marks to be established and the gene to be expressed [217].

Thus using these chromatin marks enables the deciphering of the non-coding regions of the genome. While not as easily readable as the initially envisioned "histone code" [141], much progress has been made. The ENCODE Project [80, 82, 81, 119, 155], Roadmap Epigenomics Project [254], and independent laboratories [206] have assayed gene expression and many chromatin marks in a large variety of cell types. Using a hidden Markov model (HMM), one group was able to define distinct regions of the genome in each of the cell types they collected [83]. This now provides the context-specificity required to predict and test the effect of non-coding SNPs identified in GWAS [304]. For example, a GWAS hit for type II diabetes could be potentially affecting gene expression in the liver, adipose tissue, brain, or beta cells of the pancreas. If chromatin profiling reveals that SNP is located in an enhancer region in only one of those tissues, that would inform the follow-up experiments to perform. Encouragingly, this sort of relationship is observed generally. That is, GWAS hits for given disease are more likely to be found in gene regulatory regions of the genome specific to tissues relevant to the disease pathogenesis [83, 304, 87, 254]. Furthermore, knowledge of these genomic annotations has been successfully used as prior information to increase the power to detect associations in GWAS [238, 321].

While knowing that an associated SNP is located in an enhancer region in a particular cell type is extremely helpful for generating testable hypotheses, it still leaves many unanswered questions. While it is usually assumed that a variant is affecting the most nearby gene, there is no guarantee this is true. And even if that assumption is true, it is unknown which allele is associated with higher expression. A direct method for addressing these uncertainties is expression quantitative trait loci (eQTL) mapping (note that the name is a misnomer; early

eQTL studies were performed using linkage in pedigrees, but current eQTL studies are tests of association in unrelated individuals like a typical GWAS) [211, 73, 92, 171, 233]. In this approach the phenotype of interest is the expression level of a gene. To reduce the multiple testing burden (and also because regulatory variants are often closer to the gene they affect [19]), most eQTL studies test for eQTLs nearby the transcription start site of each gene. To date, eQTL studies have been performed in many cell types [221, 108]. Reassuringly, eQTLs are more likely to be GWAS associated SNPs, consistent with the idea that GWAS hits in non-coding regions are affecting gene expression [79, 220, 222, 247, 108]. Furthermore, by combining eQTL results from many tissues collected by the GTEx Consortium [107] with GWAS results, it is possible to determine the tissue(s) most affecting a given disease by finding which tissue is enriched for tissue-specific eQTLs that are also GWAS hits for the disease [228].

A common functional genomics technique is RNA-seq [322, 229, 318]. It is an efficient method for interrogating cellular function by measuring genome-wide gene expression levels. RNA-seq has multiple advantages over its predecessor, gene expression microarrays. For example, it is not as limited by genome annotations and has a higher dynamic range [201, 344]. Most importantly for Human Genetics applications, any polymorphisms present in the coding regions in a population being studied will be present in the RNA-seq reads. These can be used to verify the identity of the individual being sequenced (i.e. avoid sample swaps and contamination) [147], and also to increase power in eQTL studies by comparing the allele-specific expression measurements to the eQTL effects [39, 309].

1.3 Tuberculosis and the genetic basis of susceptibility

A major subfield of Human Genetics focuses on understanding the genetic basis of susceptibility to infectious diseases [37, 43, 225, 200]. There are multiple reasons that infectious diseases are of particular interest. First, from a pragmatic standpoint, infectious diseases

are a major public health concern, responsible for the deaths of millions annually. Thus any increased understanding of who is likely to be susceptible or potential drug targets has great potential to reduce human suffering worldwide. Second, from a theoretical perspective, because hosts and pathogens engage in a constant co-evolutionary arms race, natural selection on any mutations affecting the response to a pathogen are strong and more likely to be detected via statistical tests (in contrast to the example of height above) [110, 243, 225, 95, 198]. Indeed, genome-wide scans of selection have found an enrichment of immune-related genes [247, 96, 95]. Third, because the immune system is responsible for fighting all pathogens it encounters but also results in less desirable functions like allergic reactions and auto-immune disorders, understanding the genetic basis of the susceptibility to one pathogen informs susceptibility to other pathogens and also other immune-based phenotypes of interest [312, 247]. As an example, a GWAS of the auto-immune disorder inflammatory bowel disease (IBD) found significant overlaps between the IDB susceptibility loci and not only other immune-related disorders, but also with susceptibility loci for infection with *Mycobacterium leprae* (which causes leprosy) [146].

My dissertation research addresses susceptibility to a different mycobacterium, *Mycobacterium tuberculosis* (MTB), which causes tuberculosis (TB). TB has been an extremely deadly disease throughout human history and continues to this day [101]. Specifically, the latest statistics released by the World Health Organization estimated that in 2014 there were 9.6 million new cases of TB and 1.5 million deaths caused by TB [332, 331]. TB is contracted by inhaling MTB in air droplets [216]. If uncontained, MTB proliferates in the lungs (and also sometimes spreads to other organs) leading to coughing, weight loss, and degradation of the lung tissue (and any other organ it colonizes) [189]. The induced coughing is how MTB spreads to other hosts, and thus it is very contagious. 70% of untreated individuals die from TB [332]. The current treatment regimen involves 6 months of cocktail antibiotic therapy [280]. Because of the difficulty of adhering to intense antibiotic therapy for such

a prolonged period of time, multi-drug resistant strains of MTB have evolved as patients stop taking their medicine once their symptoms resolve [265]. With the concurrent spread of human immunodeficiency virus (HIV), which weakens the immune system, MTB (and especially multi-drug resistant MTB) has the potential to kill many millions more in the future [32]. International efforts to prevent the spread of MTB and reduce the number of cases of TB worldwide have successfully led to improvements in diagnosis and treatment of TB and stimulated further research [101, 332]. Unfortunately progress is slow; the incidence rate of TB has decreased by an average of only 1.5% every year since 2000 [332].

TB is an ancient disease. It is estimated that MTB began infecting humans before the out-of-Africa migration [52]. This provides an explanation for the peculiar combination of features of MTB transmission. On the one hand, because MTB induces coughing and spreads through aerosol transmission, its ability to quickly spread throughout a population is typical of so called crowd diseases which emerged when humans began living in dense populations (e.g. smallpox) [328]. On the other hand, only approximately 10% of individuals infected with MTB will go on to develop active TB, with the most critical window within the first 2 years [224, 227]. The majority of individuals will have what is called a latent TB infection, in which MTB persists in a dormant state inside alveolar macrophages [18, 213]. This suggests an adaptation to a low density population where it would be disastrous for a pathogen to kill the entire population.

Given that only approximately 10% of individuals are susceptible to TB and that this variation is heritable [149, 54, 50, 210], there has been much interest in understanding its genetic basis. Being able to predict who is susceptible to TB would inform who to monitor most closely after a TB outbreak and also reduce the need of treating individuals with latent TB infections (the current standard is to give 3 months of antibiotic therapy to a latently infected individual [213]). Unfortunately, multiple GWAS to date have only identified a few loci with small effect size [302, 194, 301, 239, 47, 59, 278]. Thus they are not useful for

predicting TB susceptibility. These results suggest that TB susceptibility is also a highly polygenic disease, perhaps because of its long evolutionary history with humans. Encouragingly, however, some of these GWAS signals have identified genes potentially important for fighting TB susceptibility. A recent GWAS in a Russian population identified associated SNPs nearby the gene *ASAP1*, and further functional experiments revealed that decreased expression of *ASAP1* leads to decreased migration of dendritic cells (DCs) [59].

Because the small number of GWAS hits to date suggest a highly polygenic architecture, and furthermore since most hits have been in non-coding regions, there has been motivation to use functional genomics techniques on innate immune cells [300, 17, 232]. When MTB enters the lungs, the first immune cells it encounters are the alveolar macrophages [293, 84, 270]. Importantly, these are the cells that MTB persists inside during a latent infection. Another important part of the innate immune response to MTB is the DC. DCs phagocytose MTB in the lungs and then subsequently travel to the draining lymph nodes to stimulate T cell maturation [293, 84, 270]. Thus DCs are a critical connection between innate and adaptive immunity. The adaptive immune response is necessary for fighting an MTB infection [91]. However, most functional genomics studies of MTB infection focus on innate immune cells because 1) MTB replicates inside macrophages [270] (not DCs however [292]) , 2) the adaptive immune response does not begin until well after MTB has time to propagate [156], and 3) attempts to prime the adaptive immune response via vaccinations have been largely unsuccessful [319].

Focusing on innate immune cells, both the chromatin profiling and eQTL mapping strategies described above have been used to study MTB infection. Specifically, DNA methylation, histone marks, and chromatin accessibility have been assayed in DCs 18 hours post-infection with MTB (or control) [232]. Furthermore, hundreds of eQTLs were identified that were specific either to MTB-infected or non-infected DCs [17]. My dissertation work continued this functional genomics approach. In Chapter 2, I identified gene expression changes in

macrophages infected with MTB and related mycobacteria, but not other bacterial pathogens [25]. These genes likely harbor genetic variation underlying TB susceptibility. In Chapter 3, I discovered gene expression changes in DCs isolated from individuals known to be susceptible or resistant to TB. Furthermore, I used these gene expression differences to build a classifier of TB susceptibility. Both of these studies increased our understanding of the role of the innate immune response in TB susceptibility, and more generally demonstrate the utility of a functional genomics approach to decipher polygenic traits.

1.4 Single cell sequencing technology and the future of functional genomics

The previous functional genomics techniques discussed above, e.g. RNA-seq and ChIP-seq, measure averages across many cells. Thus they are unable to detect the cell-to-cell heterogeneity due to stochastic noise, different subpopulations of cells, or differences in the surrounding environment. In recent years, many new technologies have been developed for assaying genome-wide measurements in single cells [182, 191, 260, 106, 285, 15]. Most enable measuring gene expression levels via single cell RNA-seq (scRNA-seq), but techniques have also been developed to measure DNA methylation [274, 10], transcription factor binding [259], chromatin accessibility [34, 60], protein levels [98], and 3D genome architecture [215] in single cells. The increased resolution obtained with these single cell technologies has great potential for improving our understanding of gene regulatory mechanisms.

While there are many applications of single cell technology, one exciting area of research is investigating the innate immune response to infection at the resolution of single cells [261, 242]. The gene expression differences I have observed in my bulk RNA-seq experiments of bacterial infection could have arisen from any combination of the following explanations: a global difference in the innate immune response to a pathogen, a difference in only a subset of cells, or a difference in the fraction of cells that are infected. Interestingly, initial studies

of the innate immune response to bacterial infection have observed substantial heterogeneity among single cells [266, 137, 267, 12].

As with any new functional genomics technique, it is important to investigate and understand the technical biases to avoid during study design [11, 178, 100]. The initial studies which investigated the technical noise in scRNA-seq were interested in differentiating between the biological and technical variation affecting cell-to-cell differences in gene expression [29, 105, 134, 68, 308]. On the other hand, scRNA-seq studies of gene expression differences across multiple conditions were generating data from many batches of single cell RNA-seq. Because many of the new technologies are limited to sorting cells from just one condition in each batch, many of these studies confounded the biological conditions of interest with the technical batch processing [116]. With this confounded study design (each biological condition being represented in only one batch), any technical variation will be contributed to differences between the biological conditions of interest. In Chapter 4, I discuss my research that measured the technical variation in scRNA-seq introduced by batch processing and recommended an efficient study design to account for technical variation while minimizing replication [306].

CHAPTER 2

MYCOBACTERIAL INFECTION INDUCES A SPECIFIC HUMAN INNATE IMMUNE RESPONSE

2.1 Abstract

The innate immune system provides the first response to infection and is now recognized to be partially pathogen-specific. *Mycobacterium tuberculosis* (MTB) is able to subvert the innate immune response and survive inside macrophages. Curiously, only 5-10% of otherwise healthy individuals infected with MTB develop active tuberculosis (TB). We do not yet understand the genetic basis underlying this individual-specific susceptibility. Moreover, we still do not know which properties of the innate immune response are specific to MTB infection. To identify immune responses that are specific to MTB, we infected macrophages with eight different bacteria, including different MTB strains and related mycobacteria, and studied their transcriptional response. We identified a novel subset of genes whose regulation was affected specifically by infection with mycobacteria. This subset includes genes involved in phagosome maturation, superoxide production, response to vitamin D, macrophage chemotaxis, and sialic acid synthesis. We suggest that genetic variants that affect the function or regulation of these genes should be considered candidate loci for explaining TB susceptibility.

2.2 Introduction

The innate immune system provides the first line of defense against microbial pathogens. Broadly speaking, innate immune cells recognize foreign molecules through pattern recognition receptors (PRRs), e.g. Toll-like receptors (TLRs), which bind to highly-conserved pathogenic motifs known as pathogen-associated molecular patterns (PAMPs) [121, 209]. In addition, innate immune cells recognize damage-associated molecular patterns (DAMPs)

of host molecules released by infected cells [45]. The initial innate response involves the release of proinflammatory cytokines and lipids to recruit and activate other immune cells, phagocytosis of the pathogen, and apoptosis [138]. If the infection persists, the phagocytes stimulate the adaptive immune system by presenting antigens to activate T and B cells. In contrast to the highly specific adaptive immune response, the innate immune response has traditionally been viewed as a general response to infection.

Yet, more recent work revealed that the innate immune system also produces a pathogen-specific response in addition to the general response [124, 27, 218, 140]. Furthermore, this pathogen-specific innate response can in turn affect the specificity of the adaptive immune response by directing the differentiation of T cells into distinct subtypes [135]. That said, though we developed an appreciation for the importance of the specific innate immune response, we still do not know the extent to which the innate immune response differs between infections nor fully understand the consequences of specific innate immune responses for fighting pathogens. One of the first challenges is to distinguish the unique immune response to a specific pathogen from the large core more general response.

The pathogen-specific innate immune response is determined, at least in part, by the specificity of the PRRs of the host immune cell. Each PRR binds to its specific targets and activates certain downstream signaling pathways [153]. For example, treatment of mouse dendritic cells with lipopolysaccharide (LPS), which is found on the outer membrane of gram-negative bacteria, or with PAM3CSK4 (PAM), which is a synthetic lipoprotein that mimics those found on both gram-negative and gram-positive bacteria, induce different transcriptional response programs, because the two antigens are bound by TLR4 and TLR2, respectively [7]. Different pathogens not only stimulate different PRRs, but they have also evolved different evasion mechanisms to manipulate the innate immune response [209, 122, 31, 67]. These evasion strategies likely also contribute to the specificity of the response to different pathogens.

In the context of evasion strategies, the case of *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis (TB), is especially interesting. In order to increase its success inside alveolar macrophages - the primary cells that target MTB upon infection - MTB subverts the immune response through various mechanisms. MTB disrupts phagosomal maturation, thus preventing acidification by vesicular proton pumps and lysosomal fusion [289, 122, 113], and delays stimulation of the adaptive immune system by inducing host expression of anti-inflammatory cytokines [311, 99]. In order to achieve these manipulations, MTB must be able to secrete bacterial effectors from the phagosome into the cystosol where they can interact with host factors [282]. For this reason, the ESX-1 secretion system of MTB is critical for virulence because it permeabilizes the phagosome membrane [310, 272]. Not only does this membrane permeabilization provide a means for bacterial molecules to access the cytosol, but at later timepoints MTB has been observed to have escaped into the cytosol [282]. One well-studied consequence of phagosomal permeability is the detection of MTB DNA in the cytosol by the host sensor cGAS (*MB21D1*) and subsequent activation of the STING (*TMEM173*) pathway [65, 51, 324, 323]. These signaling events result in immune responses that are both beneficial and detrimental to MTB survival. On the one hand, the expression of anti-viral type I interferons are increased, a response which has been observed to benefit the growth of MTB and other bacteria [283]. On the other hand, MTB is targeted for destruction via autophagy, a key defense mechanism for fighting intracellular pathogens [325]. Thus the survival or destruction of MTB inside the macrophage depends on complex interactions between secreted bacterial effectors and host immune factors.

While the adaptive immune system is needed to prevent the spread of MTB and subsequent onset of TB, infected individuals do not become immunized against future MTB infections. This property may be related to the difficulty to develop an effective vaccine for adult TB (the current vaccine, bacillus Calmette–Gurin, BCG, is partly effective in children, much less so in adults) [319].

Interestingly from a human genetics viewpoint, there are large inter-individual differences in susceptibility to developing TB. While it is estimated that roughly a third of the human population is latently infected with MTB, only approximately 10% of healthy infected individuals will develop active TB (immunocompromised individuals, e.g. HIV-infected, develop active TB at a much higher frequency) [224]. Despite an inference for a strong individual genetic component to TB susceptibility, the genetic architecture remains largely unknown [149, 54, 50, 210]. There have been quite a few reports of candidate-gene associations, but genome wide scans have only identified two weak associations with disease susceptibility [302, 341, 301].

To begin addressing this gap, we have previously investigated genetic variation that is associated with inter-individual differences in the transcriptional response of human phagocytes to infection with MTB [17]. We found 102 and 96 genes that were associated with an expression QTL (eQTL) only pre- or post-infection, respectively. We refer to these loci as response eQTLs since their association with gene expression is affected by MTB infection. Interestingly, these response eQTLs were enriched for significant signal in a genome wide association study of TB susceptibility [302]. However, it is unknown if the genes associated with these response eQTLs are induced specifically in response to infection with MTB or are a part of the core innate immune response.

In order to characterize the innate immune response specific to MTB infection and better understand the role of the response eQTL-associated genes in the innate immune response, we infected macrophages isolated from a panel of six healthy individuals with a variety of bacteria. In addition to MTB, we chose both related mycobacteria and more distantly related bacteria.

Abbr.	Name	Description	Gram staining*
none	control	Mock infection	N/A
Rv	MTB H37Rv	A common laboratory strain of MTB	acid-fast
Rv+	heat-inactivated MTB H37Rv	Dead MTB H37Rv	acid-fast
GC	MTB GC1237	More virulent strain of MTB	acid-fast
BCG	bacillus Calmette-Gurin	Vaccine (attenuated *M. bovis*)	acid-fast
Smeg	*Mycobacterium smegmatis*	Non-pathogenic mycobacterium	acid-fast
Yers	*Yersinia pseudotuberculosis*	Facultative intracellular pathogen	Negative
Salm	*Salmonella typhimurium*	Facultative intracellular pathogen	Negative
Staph	*Staphylococcus epidermidis*	Extracellular pathogen	Positive

Table 2.1: **Description of bacteria.** *Mycobacteria are unable to be gram stained due to the low permeability of their cell walls. They are more closely related evolutionarily to gram-positive bacteria than gram-negative. However, their thick cells walls share features of gram-negative bacteria, e.g. a "pseudoperiplasm" similar to the gram-negative periplasm [114].

2.3 Results

2.3.1 Bacterial infection induces large changes in gene expression

To learn about the immune response to infection with different bacteria, with a particular emphasis on *Mycobacterium tuberculosis* (MTB), we investigated the *in vitro* gene regulatory response of macrophages to infection with multiple MTB stains, related mycobacterial species, and other bacterial species (Table 2.1). Specifically, we infected cultured macrophages with either MTB H37Rv, which is a common strain often used in laboratory experiments [253]; MTB GC1237, which is a strain of the highly virulent Beijing family [6]; bacillus Calmette-Gurin (BCG), which is attenuated *Mycobacterium bovis* used for vaccinations; *Mycobacterium smegmatis*, which is non-pathogenic; or heat-killed MTB H37Rv. In order to compare the response to infection with mycobacteria to the response to infection with other bacteria, we also included infection treatments with *Yersinia pseudotuberculosis* (gram-negative), *Salmonella typhimurium* (gram-negative), or *Staphylococcus epidermidis* (gram-positive).

We infected monocyte-derived macrophages from six individuals with the bacteria described above (including a non-infected control) and extracted RNA at 4, 18, and 48 hours post-infection (see Methods; Supplementary Fig. 2.6). We assessed RNA quality using the Agilent Bioanlyzer (Supplementary Table 2.7) and sequenced the RNA to estimate gene expression levels. Detailed descriptions of our data processing, quality control analyses, and statistical modeling are available in the Methods section. Briefly, we mapped the short RNA-seq reads to the human genome (hg19) using the Subread algorithm [183], discarded reads that mapped non-uniquely, and counted the number of reads mapped to each protein-coding gene. We normalized the read counts using the weighted trimmed mean of M-values algorithm (TMM) [256], corrected for confounding "batch" effects (Supplementary Fig. 2.7), and used limma+voom [276, 277, 175] to test for differential expression (DE) between cultures infected with each bacteria and their time-matched controls (Supplementary Table 2.3). Using this approach we initially observed the following general patterns: at four hours post-infection, only *Y. pseudotuberculosis* and *S. typhimurium* elicitepd a strong transcriptional response (Fig. 2.1A); at 18 hours post-infection, all the bacteria had elicitepd a strong immune response (Fig. 2.1A-B); and at 48 hours post-infection, all the bacteria continued stimulating the immune response (Fig. 2.1A), however, many of the DE genes were not shared between the 18 and 48 hour timepoints (Fig. 2.1C). Of note, at 48 hours post-infection we were unable to collect RNA from macrophages infected with *S. epidermidis* (see Methods).

2.3.2 Joint analysis identifies bacteria-specific response genes

In order to learn about variation in the innate immune response to bacterial infection, we identified genes whose regulation was altered by treatment with specific bacteria at specific timepoints. We first used a nave approach whereby we determined all the pairwise overlaps between lists of DE genes across treatments (Supplementary Table 2.8). The caveat of

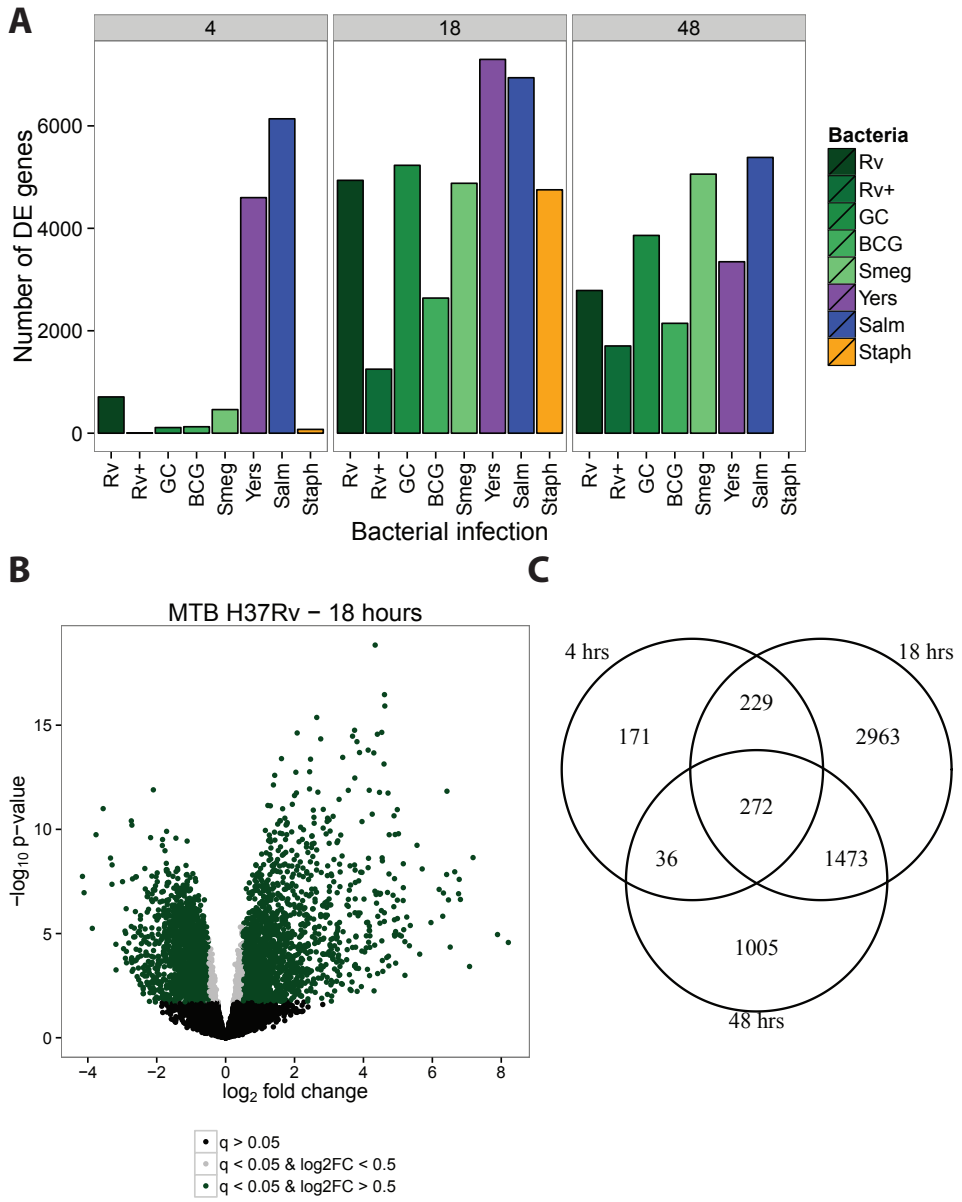


Figure 2.1: **Differential expression analysis.** We tested for differentially expressed genes for each bacterial infection by comparing it to its time-matched control. (A) We classified genes with q-value $< 5\%$ as differentially expressed. The mycobacteria are labeled in shades of green. (B) As expected, there were large transcriptional changes 18 hours post-infection with MTB H37Rv. Genes with q-value $< 5\%$ and absolute $\log_2$ fold change greater than 0.5 are labeled green, those with q-value $< 5\%$ and absolute log 2 fold change less than 0.5 are labeled grey, and non-differentially expressed genes are labeled black. (C) The overlap in differentially expressed genes identified at 4, 18, and 48 hours post-infection with MTB H37Rv.

this strategy is that incomplete power can result in overestimating the difference between treatments. In order to account for incomplete power to detect DE genes when performing multiple pairwise comparisons, we performed a joint Bayesian analysis, which we implemented using the R/Bioconductor package Cormotif [326] (see Methods for more details). Using this approach, we classified genes into regulatory patterns based on their expression levels following each of the bacterial infections.

First, we examined the data across all the bacteria-time combinations. Initially we built a model that classified genes into one of 14 separate patterns based on their expression levels after each infection relative to their expression level in the non-infected control (Supplementary Fig. 2.8). However, we found that a model with only six expression patterns (Fig. 2.2; Supplementary Tables 2.4,2.5), where a subset of the original 14 patterns are combined, is more intuitive from a biological perspective; thus we proceeded with the reduced model. Broadly speaking, we classified genes as responding in the early, middle, or late stages of infection, and we characterized the response as temporary or sustained. Pattern "non-DE" includes 4,245 genes whose expression levels were unchanged in all the experiments. Pattern "Yers-Salm" includes 1,414 early response genes whose expression levels changed at four hours post-infection with either *Y. pseudotuberculosis* or *S. typhimurium*, but not after infection with other bacteria. The genes in this pattern are enriched for gene ontology (GO) annotations related to type I interferon signaling (e.g. *SP100*, *IFI35*, *STAT2*), antigen presentation (*HLA-A*, *PSME1*, *CTSS*), and apoptosis (*CASP8*, *TRADD*, *FADD*) (Supplementary Table 2.6). Pattern "18 h" includes 3,201 middle response genes whose expression levels changed exclusively at 18 hours post-infection in response to all bacteria and is enriched for GO annotations related to apoptosis (e.g. *E2F1*, *TP53*, *WWOX*). Pattern "48 h" includes 1,204 late response genes whose expression levels changed at 48 hours and is enriched for GO annotations related to phagocytosis (e.g. *MFGE8*, *COLEC12*) and tumor necrosis factor-mediated signaling (e.g. *STAT1*, *TRAF2*, *TNFRSF14*). Pattern "18 & 48 h"

includes 1,926 middle-sustained response genes whose expression levels changed at 18 and 48 hours and is enriched for GO annotations related to the regulation of phagocytosis (e.g. *CD36*) and TLR signaling (*TLR1*, *TLR2*, *MYD88*). Lastly, pattern "All" includes 738 early-sustained genes whose expression levels changed after infection with all the bacteria across all three timepoints and is enriched for GO annotations related to type I interferon signaling (e.g. *IRF1*, *SOCS1*, *IFIT3*), cytokine secretion (*TNF*, *IL10*, *LILRB1*), and apoptosis (e.g. *IRF7*, *BCL2A1*, *MCL1*).

Next, we tested for more specific patterns by performing Cormotif separately on the data from the middle (18 h) and late (48 h) stages of infection. At 18 hours post-infection, we identified five separate expression patterns (Fig. 2.3; Supplementary Tables 2.4,2.5). Pattern "non-DE" includes 5,268 genes whose expression levels were unchanged across all infections. Pattern "All" includes 4,424 genes whose expression levels were affected by all infections (e.g. *IL24*, *IRF2*, *TLR2*). Pattern "MTB" includes 177 genes whose expression levels changed specifically in response to infection with mycobacteria (e.g. *NCF2*, *TNFSF13*, *CSF1*). These genes had a high posterior probability of being DE 18 hours after infection with MTB H37Rv, heat-killed H37Rv, MTB GC1237, and BCG. Furthermore, the gray shading for *M. Smegmatis* (Fig. 2.3A) signified an intermediate posterior probability for DE. In essence, this pattern is a merger of two sets of genes that were not large enough to be separated: one set that was DE across all five mycobacteria and another that was only DE after infection with the MTB strains and the closely-related BCG, but not *M. Smegmatis*. Pattern "Virulent", in contrast, includes 1,165 genes whose expression levels were less strongly changed after infection with heat-inactivated MTB H37Rv or the attenuated vaccine strain BCG compared to the other bacteria (e.g. *IL1R1*, *IRF1*, *PILRB*). Also the genes in this category only have an intermediate probability of responding to the non-pathogenic *M. smegmatis*. Lastly, pattern "Yers-Salm" includes 1,694 genes whose expression levels changed preferentially after infection with *Y. pseudotuberculosis* or *S. typhimurium* (e.g. *TLR8*, *TGFB1*, *IL18*).

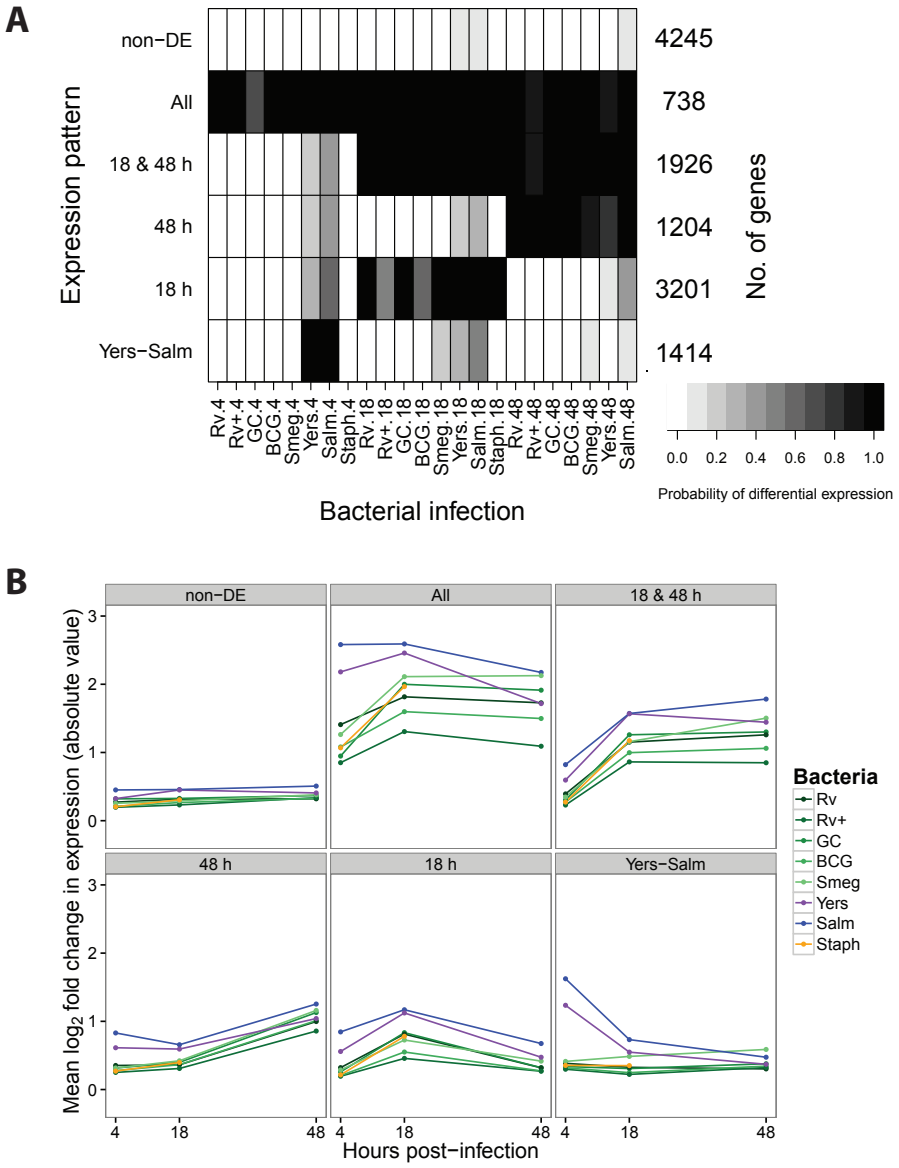


Figure 2.2: **Joint Bayesian analysis.** (A) Joint analysis of gene expression data from all three timepoints with Cormotif [326] identified six expression patterns: "non-DE", "Yers-Salm", "18 h", "48 h", "18 & 48 h", and "All". The shading of each box represents the posterior probability that a gene 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. (B) Each data point is the mean $\log_2$ fold change in expression (absolute value) in response to infection with the given bacteria for all the genes assigned to the particular expression pattern.

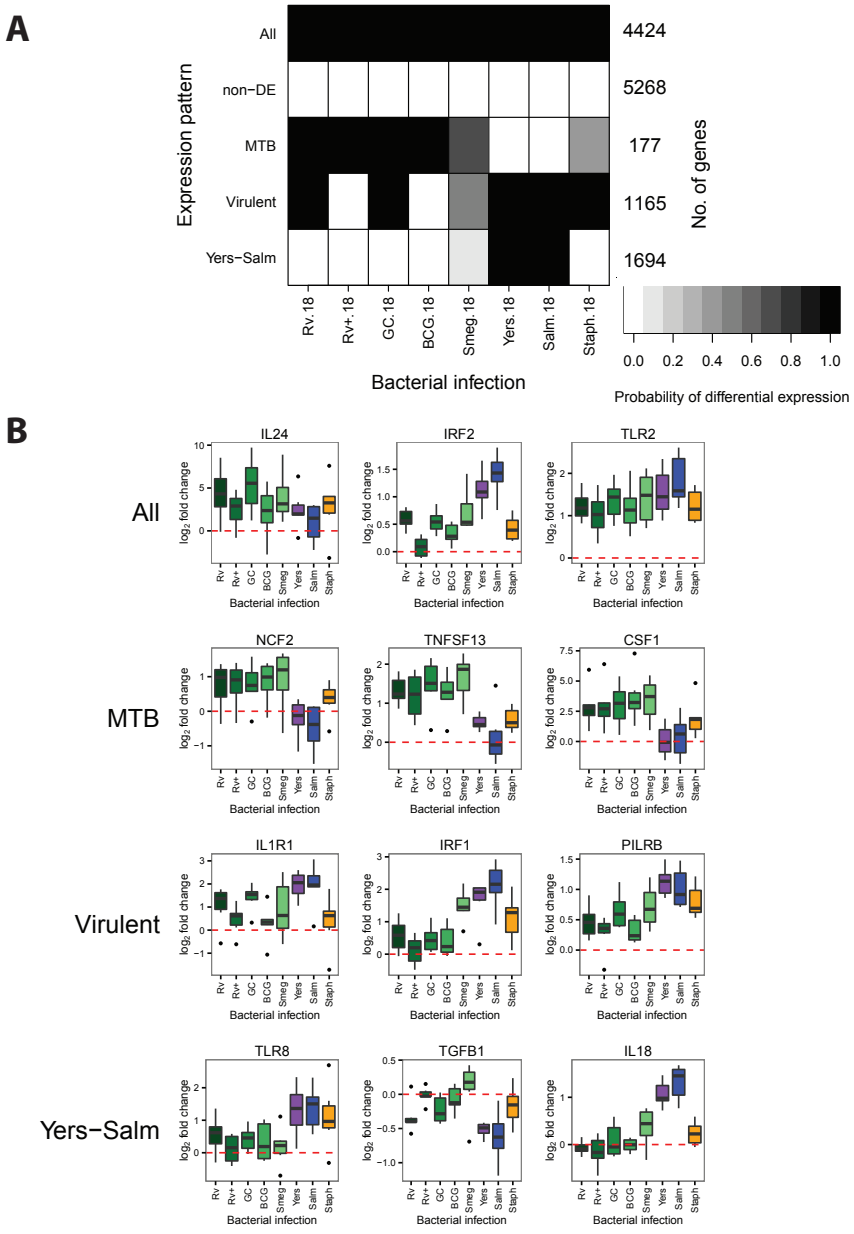


Figure 2.3: **Joint Bayesian analysis - 18 hours post-infection.** (A) Joint analysis of gene expression data from 18 hours post-infection with Cormotif identified five expression patterns: "Yers-Salm", "Virulent", "MTB", "non-DE", and "All". (B) Example genes from the different expression patterns.

At 48 hours post-infection, we also discovered five expression patterns (Fig. 2.4; Supplementary Tables 2.4,2.5). While many of the patterns have similar specificities to those observed at 18 hours post-infection, there is only little overlap across timepoints with respect to the genes comprising the patterns. For example, pattern "Yers-Salm" at 48 hours includes 1,582 genes whose expression levels changed strongly after infection with *Y. pseudotuberculosis* or *S. typhimurium* (e.g. *HLA-DPB1*, *IL10RB*, *CD248*), but only 263 of these genes are also in the corresponding pattern when we considered the data from the 18 hour timepoint. Similarly, at the 48 hour timepoint, pattern "MTB" includes 288 genes whose expression levels changed preferentially after infection with mycobacteria (e.g. *CCL1*, *ATP6V1A*, *IL27RA*), but only 33 of these genes are in the corresponding pattern at the 18 hour timepoint. Pattern "Virulent" includes 14 genes whose expression levels were not changed after infection with heat-inactivated MTB H37Rv or the attenuated vaccine strain BCG (e.g. *MAP3K4*, *SEMA4G*, *BTG1*), and only one of these also belongs to the pattern "Virulent" at 18 hours post-infection.

2.3.3 Infection-induced response eQTLs are shared across bacterial infections

Using the gene expression patterns we identified by applying the joint analysis approach, we investigated the specificity of previously identified response eQTLs to infection with MTB H37Rv [17]. Since the response eQTLs were identified at 18 hours post-infection, we investigated the distribution of genes associated with response eQTLs among the five patterns we found at that timepoint (Fig. 2.5A). Only one gene associated with a response eQTL was also DE specifically in response to MTB (*CMAS*). Otherwise, most of the response eQTL-associated genes were classified as either DE following infection with all bacteria or not DE in any infection. That a large proportion of the genes associated with response eQTLs were not DE in any of these experiments is likely due to the fact that the eQTL study was

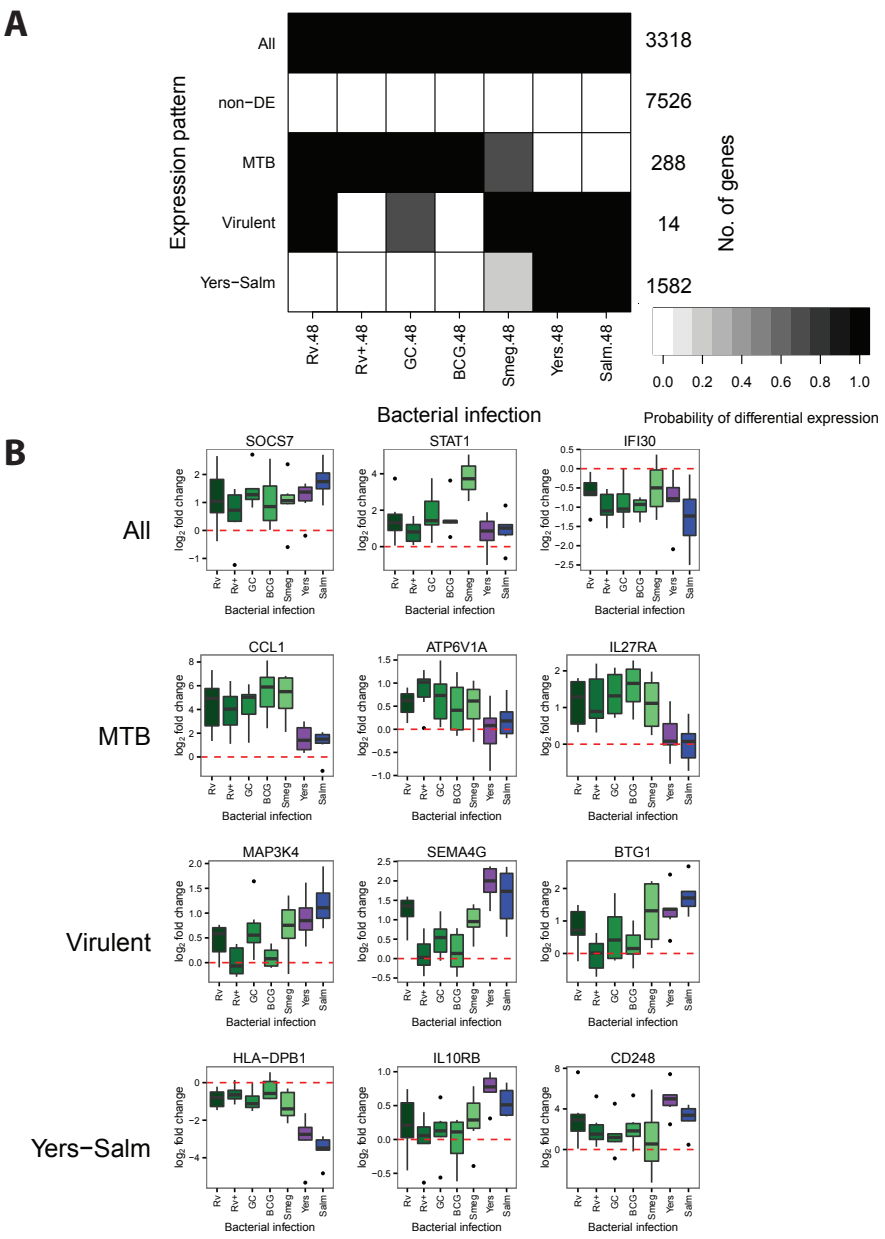


Figure 2.4: **Joint Bayesian analysis - 48 hours post-infection.** (A) Joint analysis of gene expression data from 48 hours post-infection with Cormotif identified five expression patterns: "Yers-Salm", "Virulent", "MTB", "non-DE", and "All". (B) Example genes from the different expression patterns.

performed in dendritic cells whereas our data were collected from macrophages. Overall, our observations suggest that most of the previously identified response eQTLs are genetic variants that affect the human innate immune response to bacterial infection in general, and not specifically the response to MTB H37Rv.

To provide further broad support for the interpretation that the response eQTL genes are important for the innate response to bacterial nfection in general, we considered the $\log_2$ fold change in expression values following infection (Fig. 2.1). For each gene, we calculated the mean $\log_2$ fold change in expression level across the eight bacterial infections at the 18 hour timepoint. Next, we compared the absolute values of the $\log_2$ fold change in expression between genes associated with response eQTLs, genes associated with general eQTLs (i.e. genes associated with an eQTL pre- and post-infection), and genes not found to be associated with an eQTL. Since there was a large difference in the number of genes in these three classes, we subsampled genes from each and calculated the mean of the absolute values (and repeated this process 1000 times). We found that the expression level of genes associated with response eQTLs is altered to a larger degree (significantly higher effect size; $P < 2.2 \text{ x } e^{-16}$; Fig. 2.5B) following infection compared to the genes in the other two classes.

2.4 Discussion

2.4.1 Bayesian analysis identified mycobacteria-specific response genes

In order to identify general and treatment-specific gene regulatory responses, we performed a joint Bayesian analysis of the data using Cormotif [326]. By jointly analyzing the data, as opposed to comparing overlaps between independent lists of differentially expressed genes generated using an arbitrary cutoff, we minimized the identification of specific responses due to false negatives (i.e. genes that appear to be differentially expressed in response to a subset of bacterial infections when in reality the response is similar across all the infections). Similar

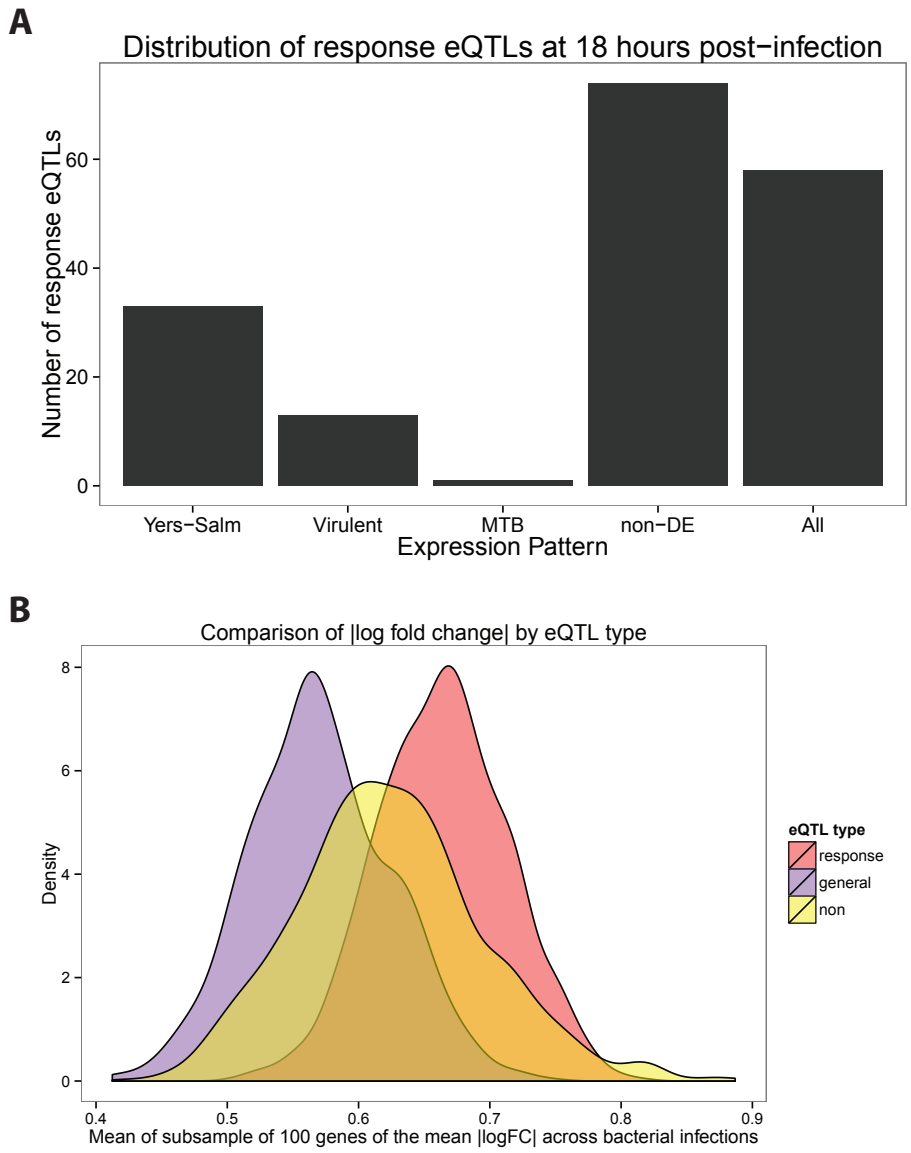


Figure 2.5: **Response eQTLs at 18 hours post-infection.** (A) We counted the number of response eQTLs from Barreiro et al. [17] (179 out of the 198 were also expressed in our study) in each of the five gene expression patterns at 18 hours post-infection (Fig. 2.3). (B) We compared the mean $\log_2$ fold change in expression across the 8 bacterial infections at the 18 hour timepoint for three classes of genes: response eQTL genes (red), general eQTL genes (purple), and non-eQTL genes (yellow) (see Methods for details).

to previous observations [124, 27], we found a large core transcriptional response to infection. However, we also identified a novel subset of genes whose regulation is preferentially altered in response to infection with mycobacteria but not to the other bacteria we tested. Since these responses are unique to infection with mycobacteria (at least in the context of our study design), they may be promising candidates for future studies that focus on the mechanisms by which mycobacteria successfully subvert the human innate immune response. Since this study does not extend to investigation of mechanisms, we do not have empirical data with which to prioritize such possible candidate genes. Yet, the reported functions of many of these genes often suggest mechanisms that are relevant, and often quite specific, to MTB infection. Prioritizing candidate genes in this way is not statistically valid, and one can argue (and indeed, this has been shown [235]) that any list of genes can be scrutinized to yield "interesting relevant stories". We therefore offer these details in the context of a discussion (rather than "results"), to provide one set of alternative explanations for our findings, and generate ideas for further investigations.

For example, when we focused on the mycobacteria-specific regulatory response 18 hours post-infection, we noticed an intriguing number of genes that are involved in phagosome maturation (Supplementary Fig. 2.9). Broadly speaking, phagocytosed bacteria are killed by vesicular proton pumps, which lower the pH inside the phagosome, and lysosomal fusion. This process occurs once a phagosome has matured through a series of steps mediated by the exchange of Rab GTPases [315, 212]. A unique property of mycobacteria is their ability to survive inside the macrophage by inhibiting phagosome maturation [113]. As part of this strategy, the bacterium recruits *RAB22A* to MTB-containing phagosomes [255]. Indeed, we found the *RAB22A* gene to be upregulated in response to infection with mycobacteria (Supplementary Table 2.5). Similar GTPases whose regulation was altered following mycobacterial infections include *RAP2A* (upregulated), *RAB3A* and *RAB33A* (both downregulated). In addition, the vesicular (v)-ATPase subunit *ATP6V1D* was exclusively upregulated in re-

sponse to mycobacterial infection. Thus, the mycobacteria-specific response we identified includes genes putatively involved in mycobacteria-specific survival mechanisms.

An additional intriguing example involves the *NCF2* gene. This is a potential candidate gene whose expression level was affected specifically by infection with mycobacteria at 18 hours post-infection. Neutrophil cytosolic factor 2 (*NCF2*, also known as *p67phox*) is a subunit of the phagocyte NAPDH oxidase, which is responsible for generating reactive oxygen species used to fight intracellular pathogens [77, 214, 14, 35, 157, 62]. These reactive oxygen species may also serve a signaling role in activating other immune cell types to ensure proper granuloma formation and killing of mycobacteria [63]. Loss-of-function mutations in subunits of the NAPDH oxidase cause chronic granulomatous disease (CGD) [62], which is characterized by the formation of granulomas throughout the body due to the inability of phagocytes to kill the ingested pathogens. In contrast to wild type animals, mice with mutations in subunits of the phagocyte NAPDH oxidase develop tuberculosis after infection with the vaccine strain, BCG [63]. Humans who are administered the vaccine before being diagnosed with CGD also develop the disease [62].

At 48 hours post-infection (Fig. 2.4), the mycobacteria-specific response was enriched with genes annotated (based on GO) as having a role in "response to vitamin D" (Supplementary Table 2.6; Supplementary Fig. 2.10). Individuals with low circulating levels of vitamin D are more susceptible to developing tuberculosis [346, 223], and vitamin D has been investigated as a supplemental therapy for the treatment of tuberculosis, though with mixed results [202, 190, 336, 154]. Vitamin D has been found to be important for innate immune cells to fight MTB [187, 316, 337]; however, it is also an important pathway for generic bactericidal activity [115]. Consistent with its role in the innate immune response, both the enzyme that converts vitamin D to its active form (*CYP27B1*) and its receptor (*VDR*) are upregulated in response to any of the infections (pattern "All"; Fig. 2.4). Yet, the regulation of other genes involved in the response to vitamin D was only affected by

infection with MTB. *PIM1*, a serine/threonine kinase that binds the VDR and enhances transcription of its target genes [195], is upregulated in response to the mycobacteria (pattern "MTB", Fig. 2.4). Interestingly, the increased expression level of *PIM1* in T-cells was successfully used in a six-gene classifier of patients with active versus latent TB infections [136]. Another gene, the chemokine *CXCL10* (also known as interferon gamma-induced protein 10 or IP-10), is also upregulated in response to mycobacterial infection (pattern "MTB", Fig. 2.4). Discordant with the observed increase in expression of *CYP27B1*, *VDR*, and *PIM1* in response to infection, treatment with vitamin D usually leads to the reduction of *CXCL10* expression and secretion in multiple cell types [109, 4, 263]. In fact, supplementation with vitamin D decreased serum levels of CXCL10 in TB patients [57]. This suggests that the immunosuppresive effect of vitamin D signaling is insufficient to overcome the pro-inflammatory response to mycobacterial infection. This observation is in concordance with past studies which found increased expression of *CXCL10*, as well as increased secretion level from macrophages, following infection with MTB [345, 316]. Interestingly, a polymorphism in *CXCL10* was found to be associated with susceptibility to tuberculosis in a Chinese population [295, 13]. Overall, these observations provide support for the importance of vitamin D signaling for specifically fighting mycobacterial infections.

Another gene of interest from the mycobacteria-specific expression pattern at 48 hours post-infection is chemokine (C-C motif) ligand 1 (*CCL1*), which stimulates migration of human monocytes [207] (Fig. 2.4B). Thuong et al. identified *CCL1* as being induced to a greater extent in MTB-infected macrophages (4 hours post-infection) isolated from individuals with pulmonary TB compared to macrophages from individuals with latent TB infections [300]. Put together, our observations and those of Thuong and colleagues suggest that *CCL1* is involved in the pathogenesis of TB. Further supporting this notion, Thuong et al. also found a genetic association between variants in the *CCL1* region and TB susceptibility [300]. However to date, subsequent genetic association studies investigating *CCL1* have reported

mixed results [294, 231].

One caveat of the joint Bayesian analysis is that we were not able to classify genes into unusual patterns (because this approach can only discover expression patterns shared by a large number of genes and, by definition, only few genes fall into "unusual" patterns). For example, unusual patterns of interest include changes in expression specifically in response to some but not all of the mycobacterial infections. One gene that satisfied this pattern is the dual specificity phosphatase 14 (*DUSP14*). We specifically examined the expression data for this gene because it was previously associated with an MTB infection response eQTL in dendritic cells [17], and consequently when the eQTL results were used as a prior - *DUSP14* was found to be significantly associated with TB susceptibility. Moreover, knocking down *DUSP14* expression via siRNA in murine macrophages resulted in a lower bacterial load 90 hours post-infection with MTB H37Rv [139]. In our joint Bayesian analysis, *DUSP14* was not classified as one of the genes whose regulation was altered in response to infection with mycobacteria. Yet, *DUSP14* was upregulated at 18 hours post-infection with MTB H37Rv (q-value: 16%), MTB GC1237 (q-value: 3%), and BCG (q-value: 9%); and downregulated post-infection with *S. typhimurium* (q-value: 9%) (Supplementary Fig. 2.11). Thus, our data lends further support for the role of *DUSP14* as a TB susceptibility gene.

2.4.2 Little evidence for strain-specific transcriptional response to infection

There are six major families of MTB that differ in their geographic distribution and virulence [97, 53]. Strains from these families are known to differ in their growth rates inside macrophages [181], expression levels of bacterial genes [120, 258], and cell wall lipid composition [165]. Previous studies have found that different MTB strains induce different innate immune responses in human cell lines and other infection models [56]. A dominate narrative is that MTB strains from East Asia, referred to as the Beijing family (Gagneux et al. classified it as MTB lineage 2 [97]), are more virulent because they induce a lower proinflamma-

tory immune response compared to the common laboratory strains [196, 197, 250, 296, 320]. However, other studies have reported the opposite, namely that Beijing strains induce a larger proinflammatory response [41], or a conflicting response in which various pro- and anti-inflammatory cytokines are differentially regulated [257, 163] compared to laboratory strains.

In our study, albeit with a small sample size, we found no marked differences between the transcriptional response to infection with MTB H37Rv or MTB GC1237, a Beijing strain (Supplementary Fig. 2.12; Supplementary Table 2.9). Furthermore, the pro-inflammatory cytokines *TNF* and *IL6* and the anti-inflammatory cytokine *IL10* were strongly upregulated in response to both strains of MTB (Supplementary Fig. 2.13). This observation is in concordance with Wu et al., who also reported no apparent difference in the transcriptional response of THP-1 cells to infection with MTB H37Rv versus multiple Beijing strains [335]. Thus the increased virulence of the Beijing family of MTB strains may be due to mechanisms not assayed in this study such as post-transcriptional effects, cell-cell signaling, and environmental stimuli. It should be noted, however, that not all Beijing strains are equally virulent [70, 273] and that MTB H37Rv is a laboratory-adapted strain that has evolved independently in different laboratories [131].

2.4.3 Differences in response to virulent versus attenuated pathogens are not mycobacteria-specific

To better understand the interaction between MTB and macrophages, we included in our study both virulent mycobacteria (MTB strains H37Rv and GC1237) and attenuated mycobacteria (heat-inactivated MTB H37Rv and the vaccine strain BCG). Overall, the response to infection with either virulent or attenuated mycobacteria was similar (Fig. 2.3,2.4). This observation was unsurprising because it has been previously demonstrated that infections with inactivated pathogens (in fact, even individual pathogen components) are able to largely

recapitulate the transcriptional response to infection [124, 27, 218, 140]. In other words, as expected, the transcriptional response to infection is largely driven by the antigens present.

Yet, the responses to inactivated pathogens or individual pathogen components in past studies were not identical to the responses to live pathogens, suggesting a potential role for bacterial manipulation of the immune response. For example, it is known that BCG lacks the locus containing the ESX-1 secretion system, which is critical for MTB virulence [20, 244, 123, 271]. In our study we also observed differences between the response to virulent and attenuated mycobacteria. Specifically, there are 1,165 genes in the expression pattern "Virulent" at 18 hours post-infection (Fig. 2.3) and 14 genes that comprise of the "Virulent" pattern at 48 hours post-infection (Fig. 2.4). Importantly, these genes are also differentially expressed in response to the other virulent infections in our study, and thus they are not specifically due to the manipulations of the host cell by virulent mycobacteria.

We attempted to identify a gene expression pattern that specifically represented differences in virulence only in the mycobacteria, yet we never saw such a pattern. It is important to note that had we simply performed a simple pairwise analysis of the overlap of DE genes between MTB and BCG infections, our results would be quite different (Supplementary Table 2.8). Yet, a pairwise analysis is misleading in the context of the entire study. Indeed, by accounting for incomplete power by using the joint Bayesian model and including other bacterial species, we avoided attributing many differentially expressed genes specifically to the differences in the immune evasion mechanisms used by MTB and BCG. We conclude that either a larger sample size or a different experimental system is required to find specific differences between the response to infection with MTB and BCG.

2.4.4 Previously identified response eQTLs affect response to bacterial infection in general

In a previous study, we identified response eQTLs that were associated with gene expression levels in MTB-infected human dendritic cells. We investigated the expression pattern of genes associated with the response eQTLs in our study. Using the five expression patterns identified by the joint Bayesian analysis at 18 hours post-infection, we examined the distribution of response eQTL genes and discovered that these genes were not enriched in the mycobacteria-specific expression pattern (Fig. 2.5A). Instead, many were differentially expressed across all the infections (pattern "All"). Thus, response eQTLs modulate the inter-individual response to infection with diverse types of bacteria. That said, one gene was both associated with a response eQTL and specifically differentially expressed following mycobacterial infection. Though this result does not represent a significant enrichment of response eQTL genes among those whose regulation was affected specifically by infection with MTB, the identity of the gene renders the observation intriguing. *CMAS* (cytidine monophosphate N-acetylneuraminic acid synthetase), is an enzyme that is involved in the processing of sialic acid, which is then added to cell surface glycoproteins and glycolipids. Glycoproteins are known to be important in many functions of the immune response, including initial pathogen detection (e.g. TLRs) and antigen presentation (e.g. major histocompatibility complex (MHC) molecules) [329, 144, 58]. We suggest that this gene is an interesting candidate for further understanding both MTB pathogenesis and inter-individual susceptibility to tuberculosis.

2.4.5 Conclusions

By jointly considering data from multiple infection treatments, using a variety of bacteria, we have classified distinct innate immune transcriptional response patterns. The most inclusive pattern was a response to all the bacterial infections, indicating that the receptors that bind

the diverse antigens present on the different bacteria converge to largely similar signaling pathways. We also found an expression response pattern specific to mycobacterial infections, the main focus of the current study. At 18 hours post-infection, the mycobacteria response pattern includes genes involved in phagosome maturation and the NAPDH oxidase subunit *NCF2*. At 48 hours post-infection, it includes genes involved in the response to vitamin D and the chemokine *CCL1*. We found that the response to infection with different MTB strains was highly similar. Furthermore, the differences we identified between the response to MTB and the vaccine strain BCG were not mycobacteria-specific, but likely represent a difference between the innate immune response to virulent and non-virulent (or attenuated) pathogens. Lastly, we identified a single gene, *CMAS*, which is both associated with a response eQTL to MTB infection, and whose regulation is altered specifically when we infected the cells with mycobacteria. This gene is thus an especially promising candidate for future studies of TB susceptibility.

2.5 Methods

2.5.1 Ethics Statement

Buffy coats were obtained from healthy donors after informed consent. The blood collection protocols were approved by both the French Ministry of Research and a French Ethics Committee under the reference DC-2008-68 collection 2. The blood collection was carried out in accordance with these approved protocols by the Etablissement Franais du Sang.

2.5.2 Sample collection and macrophage differentiation

We collected buffy coats (˜50 mL) from six healthy donors. Next we isolated peripheral blood mononuclear cells (PBMCs) via Ficoll-Paque centrifugation [253] and enriched for monocytes via positive selection with beads containing CD14 antibodies [17]. Then we differentiated

the monocytes into macrophages by culturing for 6-7 days in RPMI buffer supplemented with macrophage colony-stimulating factor (M-CSF) [292].

2.5.3 Bacterial infection

For each bacterial infection (Table 2.1), we treated the macrophages with a multiplicity of infection (MOI) of 2:1. After one hour, we washed the macrophages five times with phosphate-buffered saline (PBS) and treated them with gentamycin (50 mug/muL) to kill all extracellular bacteria. After one hour of antibiotic treatment, we changed the medium to a lower concentration of gentamycin (5 mug/muL), which marked the zero timepoint of the study. We allowed the cells to grow for 4, 18, or 48 hours before lysing them with QIAzol Lysis Reagent and then storing them at -80 C. We chose these timepoints based on a previous analysis of the human transcriptional response to infection with MTB [293]. No data is available for 48 hours post-infection with *S. epidermidis*. After escaping the macrophages upon cell death, sufficient *S. epidermidis* were able to proliferate in the gentmycin-supplemented medium to contiminate the entire well by 48 hours post-infection.

2.5.4 RNA extraction, library preparation, and sequencing

We extracted RNA using the QIAgen miRNeasy kit. There were a total of 13 batches of 12 samples each (6 individuals x 9 conditions x 3 timepoints, minus 48 hours post-infection with *S. epidermidis*). We designed the batches to maximally partition the variables of interest (individual, condition, timepoint) in order to minimize the introduction of biases due to batch processing [11]. To assess RNA quality, we measured the RNA Integrity Number (RIN) with the Agilent Bioanalyzer (Supplementary Table 2.7). Importantly, there were no significant differences in the RIN (mean of 7.8 $\pm$ 2.0) between the bacterial infections or between the timepoints (Supplementary Fig. 2.14B). In batches of 12 samples, we added barcoded adapters (Illumina TruSeq RNA Sample Preparation Kit v2) and sequenced 50

base pairs single end over multiple flow cells on the Illumina HiSeq 2500.

2.5.5 Mapping, counting, and normalization

We mapped the short reads to the human genome (hg19) using the Subread algorithm [183] and discarded those that mapped non-uniquely. Next, we obtained the read counts for each Ensembl protein-coding gene (biotype: "protein_coding") with the featureCounts algorithm, which sums the reads falling in the union of all exons of a gene and discards reads mapping to more than one gene [184]. There were no significant differences in the number of mapped exonic reads (mean of 41.8 $\pm$ 21.2 million per sample) between the bacterial infections or between the timepoints (Supplementary Fig. 2.14A). We removed genes with fewer than one count per million exonic reads in fewer than six samples. To account for differences in the read counts at the extremes of the distribution, we normalized the samples using the weighted trimmed mean of M-values algorithm (TMM) [256].

2.5.6 Differential expression analysis

To assess the quality of the data, we performed principal components analysis (PCA) of the TMM-normalized $\log_2$-transformed counts per million (CPM). PC2 separated the samples by timepoint, but PC1 was associated with the RIN score and the processing batch (Supplementary Fig. 2.7A). After the effects of RIN score and processing batch were removed with the function removeBatchEffect from the limma package [252], PC1 separated the samples by timepoint and PC2 separated the infected and control samples (Supplementary Fig. 2.7B). We protected the variables of interest (individual, bacteria, timepoint) when regressing the effects of RIN score and processing batch by including them in the linear model used by removeBatchEffect. However, the result was similar if they were not protected since the variables of interest were partitioned across the processing batches (Supplementary Fig. 2.7C). All figures displaying expression data were generated using the batch-corrected data.

To confirm that the transcriptional response to MTB infection in our study was consistent with previous observations, we compared our MTB infected samples and their time-matched controls to the MTB infected samples and zero timepoint control from Tailleux et al., 2008 [293]. Despite differences in the technology used to assay gene expression (RNA-seq versus microarray) and the method used to isolate the macrophages (positive versus negative selection), we still observed a common transcriptional signature of infection using PCA (Supplementary Fig. 2.15).

For the standard analysis, we tested for differential expression using limma+voom [276, 277, 175] because it has been shown to perform well with sufficient sample size (n >= 3 per condition) [248, 279]. Based on the PCA results, we included RIN score and processing batch as covariates in the model. We corrected for multiple testing with the Benjamini & Hochberg false discovery rate (FDR) [21] and considered genes with q-value less than 5% to be differentially expressed.

Since we were interested in the shared and differential response to infection with the different bacteria, we performed a joint Bayesian analysis using the Cormotif algorithm [326]. Cormotif shares information across experiments, in this case infections, to identify the main patterns of differential gene expression (which it refers to as *correlation motifs*) and assigns each gene to one of these gene expression patterns. One caveat of the Cormotif algorithm is that is does not distinguish the direction of the effect across infections. In other words, a gene that is assigned to an expression pattern could be differentially expressed in different directions across the infections. However, in this data set, this was rarely observed (Supplementary Table 2.10).

In practice, we had to make several modifications when using Cormotif. First, since the method was developed for microarray data, we used the batch-corrected TMM-normalized $\log_2$-transformed CPM as input. Second, the method assumes independence between the experiments, and we only have one control per timepoint. However, since this dependence

will cause genes to be more likely to be either uniformly differentially expressed across all the infections or uniformly unchanged, this caveat is conservative to our results of gene expression patterns that are specific to subgroups of the bacterial infections. Third, the current version of the method (v1.14.0) does not return the cluster likelihoods, i.e. the likelihood that a gene belongs to each of the gene expression patterns. To facilitate downstream analyses with these sets of genes, we modified the original code to additionally return this information. Lastly, Cormotif is non-deterministic. Thus to obtain consistent results, we ran each test 100 times and kept the result with the largest maximum likelihood estimate.

We tested for enrichment of gene ontology (GO) biological processes among the genes in the gene expression patterns using topGO [5]. We tested for significance with the Fisher's Exact Test, used the weight01 algorithm from topGO to account for the correlation among GO categories due to its graph structure, and considered significant any category with p-value less than 0.01.

2.5.7 Analysis using previously identified response eQTLs

We downloaded the list of response eQTL genes from Supplementary Table 3 from Barreiro et al. [17]. Of the 198 response eQTL genes discovered in the dendritic cells in that study, 179 of the genes were also expressed in the macrophages from this study. In order to compare the differential expression results of the response eQTL genes to other genes, we used the $\log_2$ fold changes in expression estimated by limma [252]. First, we calculated the mean $\log_2$ fold change at 18 hours post-infection for each gene across the eight bacteria. Second, we converted these mean estimates to their absolute values. Third, we subsampled 100 genes from each of the three categories (response eQTL, general eQTL, and non-eQTL genes) and calculated the mean of the absolute values. We performed this subsampling 1000 times (Fig. 2.5B). Fourth, we performed t-tests to compare the distribution of response eQTL genes to either that of the general eQTL genes or the non-eQTL genes.

2.5.8 Data and code availability

The data have been deposited in NCBI's Gene Expression Omnibus [76] and are accessible through GEO Series accession number GSE67427 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67427`). Supplementary Table 2.2, which contains the gene expression data, and Supplementary Table 2.3, which contains the differential expression results from limma, are available from our lab website: `http://giladlab.uchicago.edu`. The code is available at `https://bitbucket.org/jdblischak/tb`.

2.6 Acknowledgments

We thank Matthew Stephens and Bryce van de Geijn for advice on the statistical analyses, and all members of the Gilad lab for helpful discussions. This work was supported by grant AI087658 to YG and LT. JDB was partially supported by National Institutes of Health Grant T32 GM007197.

2.7 Author Contributions

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

2.8 Supplementary Information

2.8.1 Supplementary Figures

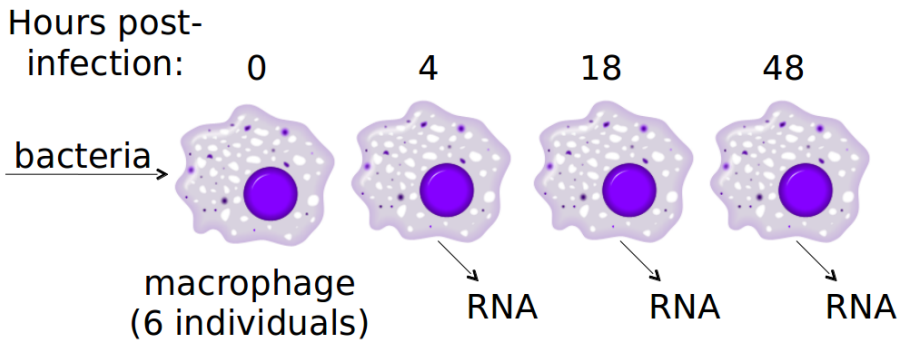


Figure 2.6: **Study design.** We infected monocyte-derived macrophages isolated from six healthy donors with the bacteria described in Table 2.1. We isolated RNA for sequencing at 4, 18, and 48 hours post-infection.

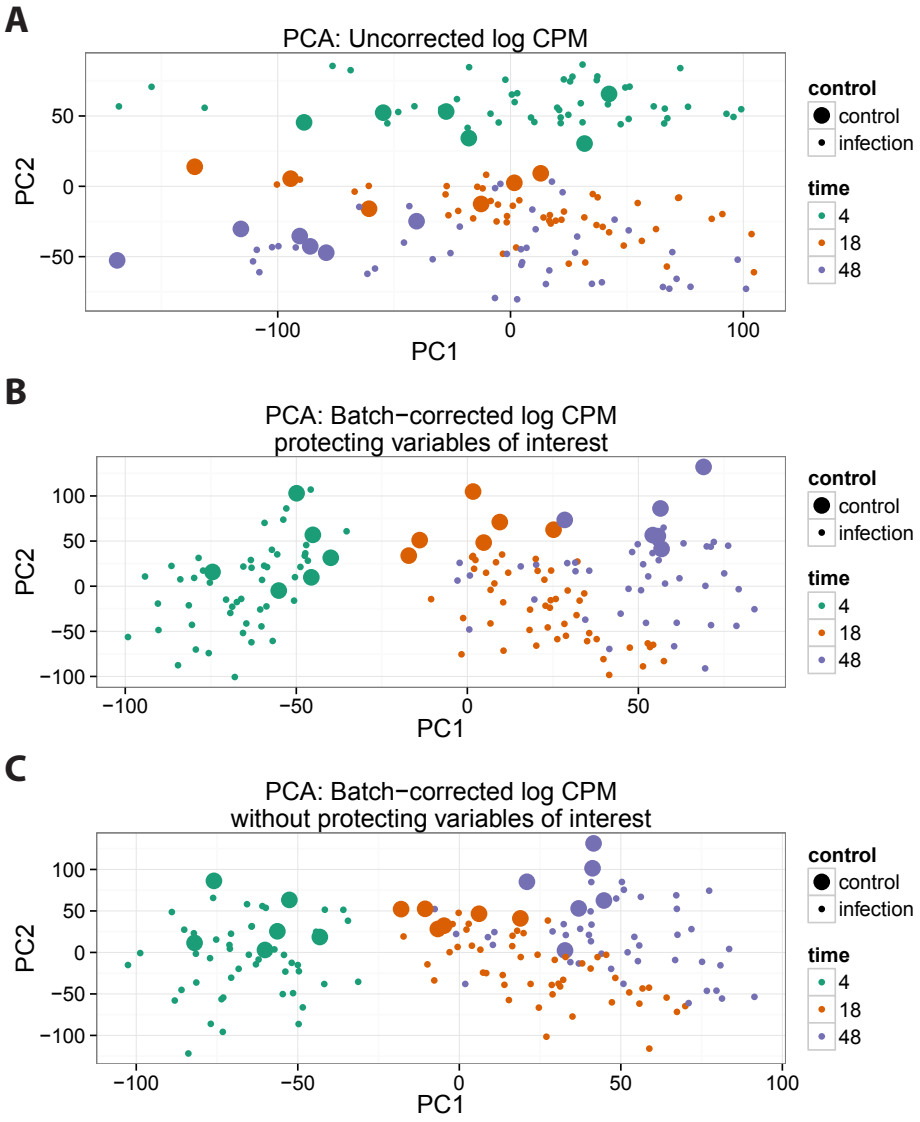


Figure 2.7: **Principal components analysis (PCA) of uncorrected and batch-corrected expression values.** (A) PCA of the TMM-normalized $\log_2$-transformed counts per million (CPM). Infected and control samples are not well separated. PC2 separates the samples by timepoint. (B) PCA of the TMM-normalized $\log_2$-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).

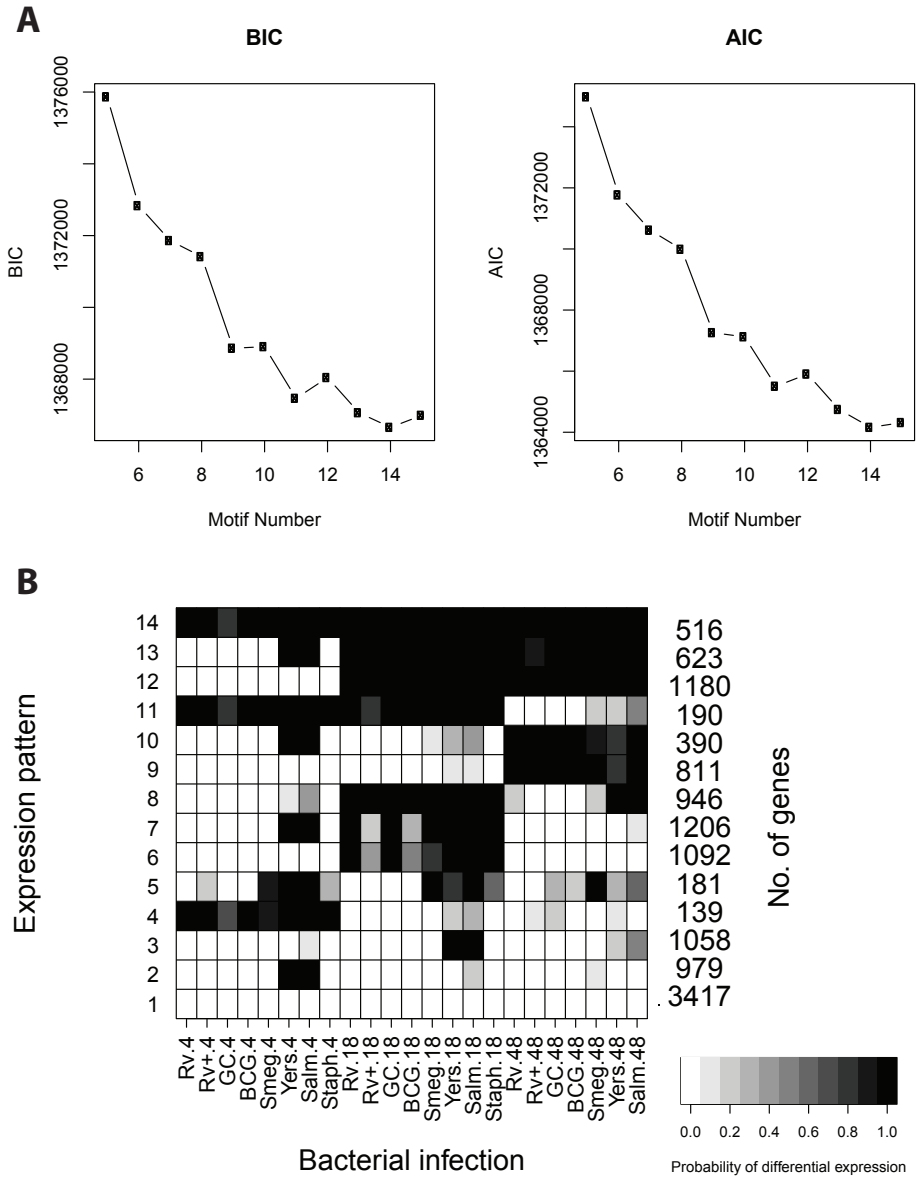


Figure 2.8: **Joint Bayesian analysis with 14 expression patterns.** (A) Cormotif [326] estimates the number of expression patterns (i.e. motifs, using their terminology) to use by calculating the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). These criteria penalize 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 2.8: (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.

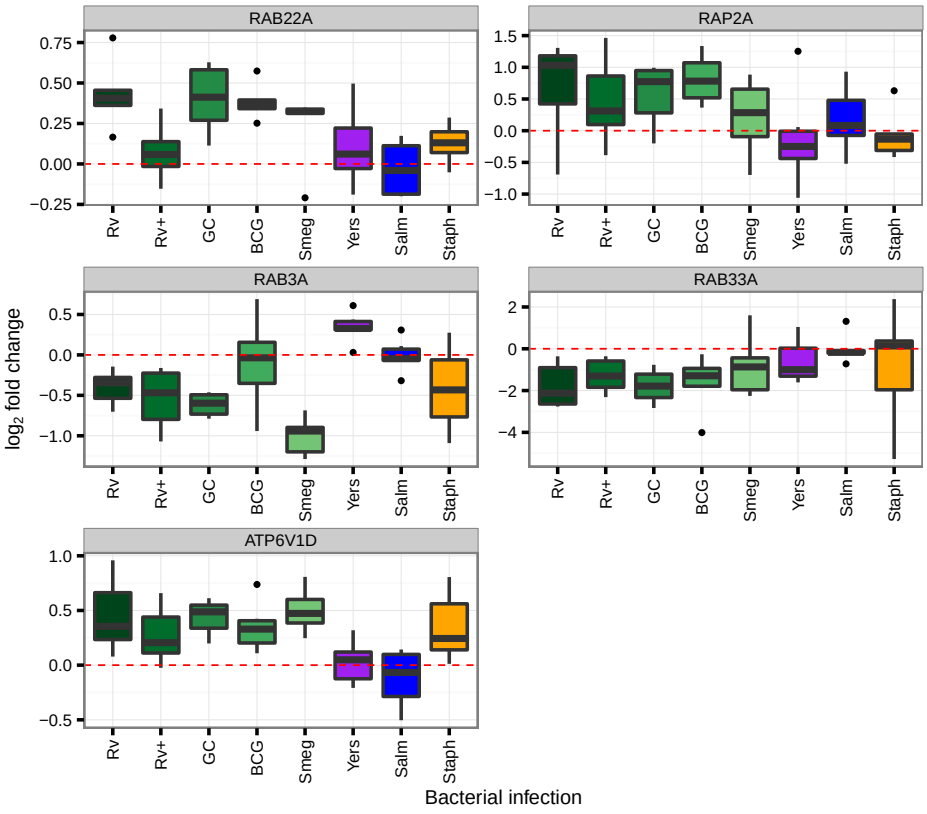


Figure 2.9: **Expression of genes involved in phagosome maturation.** *RAB22A*, *RAP2A*, and *ATP6V1D* are upregulated in response to infection with mycobacteria at 18 hours; whereas, *RAB3A* and *RAB33A* are downregulated (pattern "MTB" in Fig. 2.3).

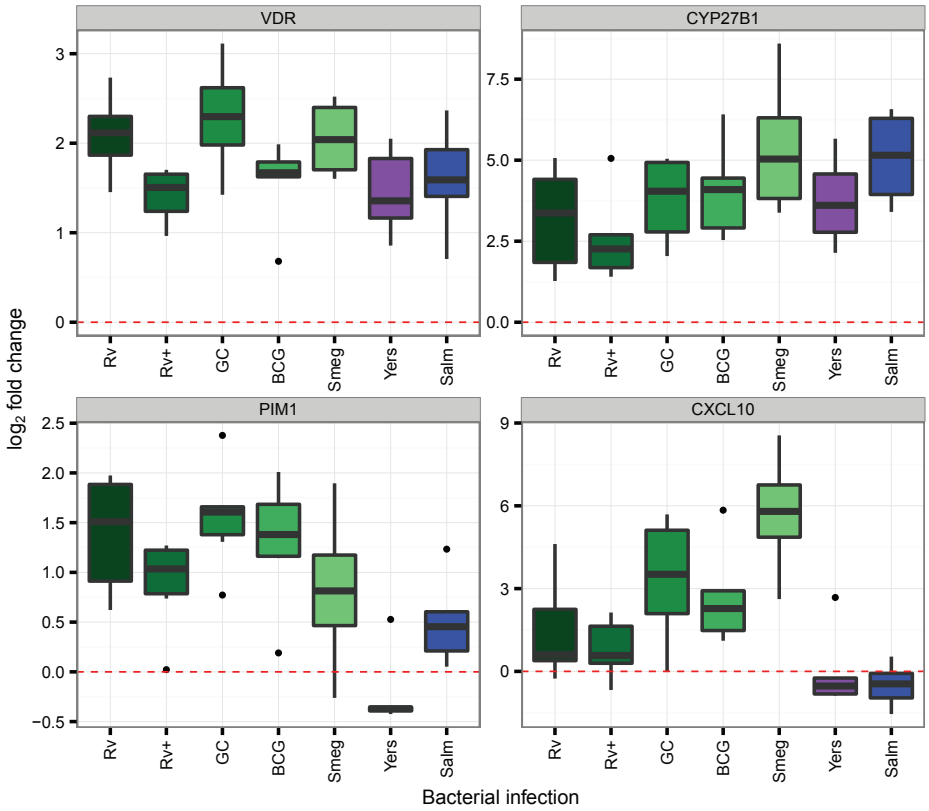


Figure 2.10: **Expression of genes involved in vitamin D signaling.** *VDR* and *CYP27B1* are upregulated at 48 hours post-infection with all bacteria (pattern "All" in Fig. 2.4). *PIM1* and *CXCL10* are upregulated at 48 hours post-infection with the mycobacteria (pattern "MTB" in Fig. 2.4).

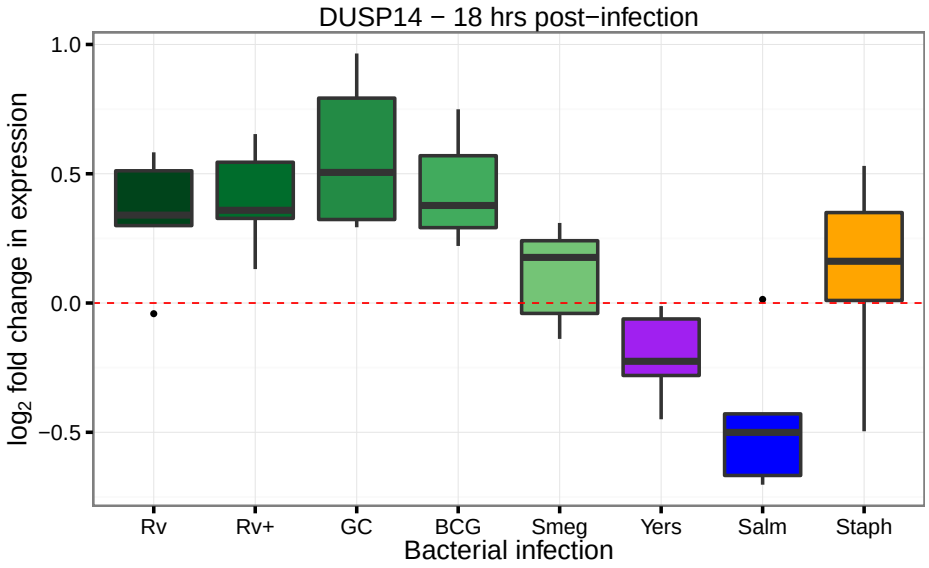


Figure 2.11: **Expression of *DUSP14* at 18 hours post-infection.** *DUSP14* is an example of an interesting gene not identified with our approach. At 18 hours, it is upregulated after infection with MTB H37Rv (q-value: 16%), MTB GC1237 (q-value: 3%), and BCG (q-value: 9%) (the change in heat-inactivated 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.

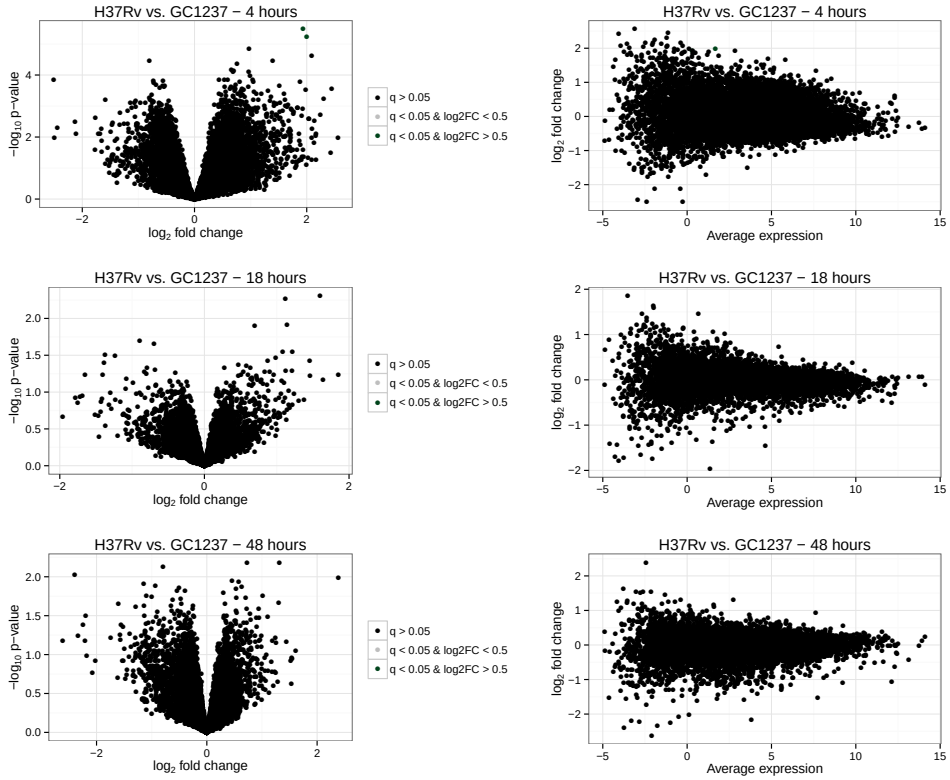


Figure 2.12: **Little difference in transcriptional response to infections with different MTB strains.** Few statistically significant differences were identified when explicitly testing gene expression levels post-infection with MTB H37Rv and MTB GC1237 at 4, 18, or 48 hours post-infection (top, middle, 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 2.9).

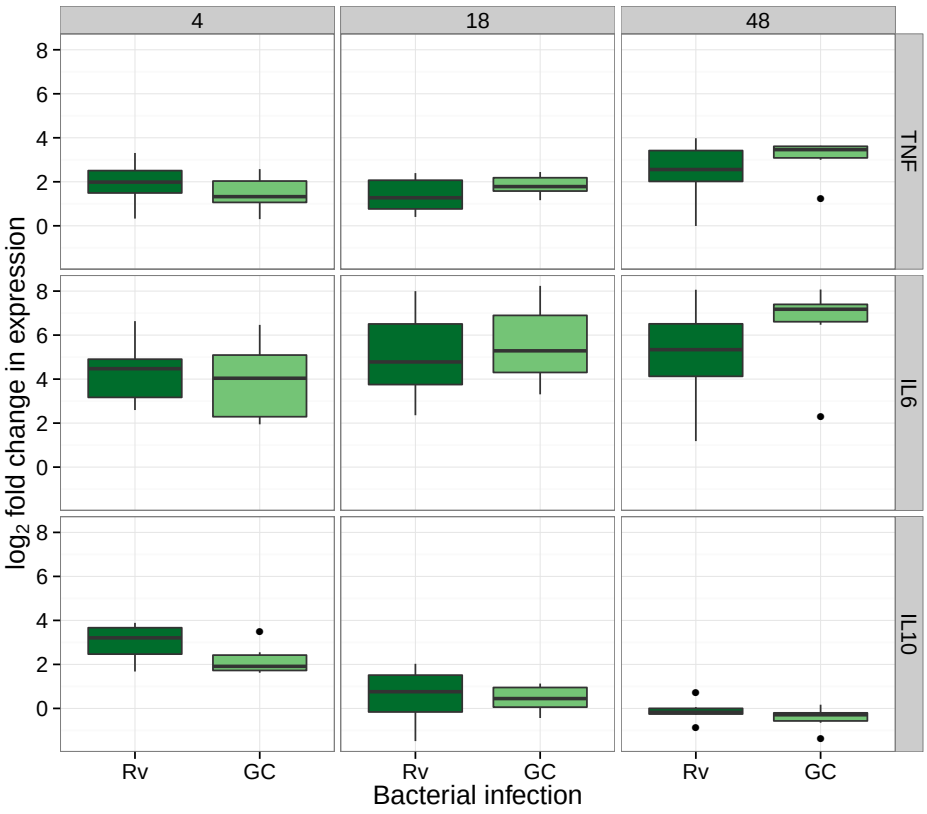


Figure 2.13: **Response of example cytokines to infection with different MTB strains.** The $\log_2$ fold change in expression of the pro-inflammatory cytokines *TNF* and *IL6* and the anti-inflammatory cytokine *IL10* is similar post-infection with MTB H37Rv or MTB GC1237. *TNF* and *IL6* are upregulated at all three timepoints; whereas, *IL10* is upregulated only at 4 hours post-infection.

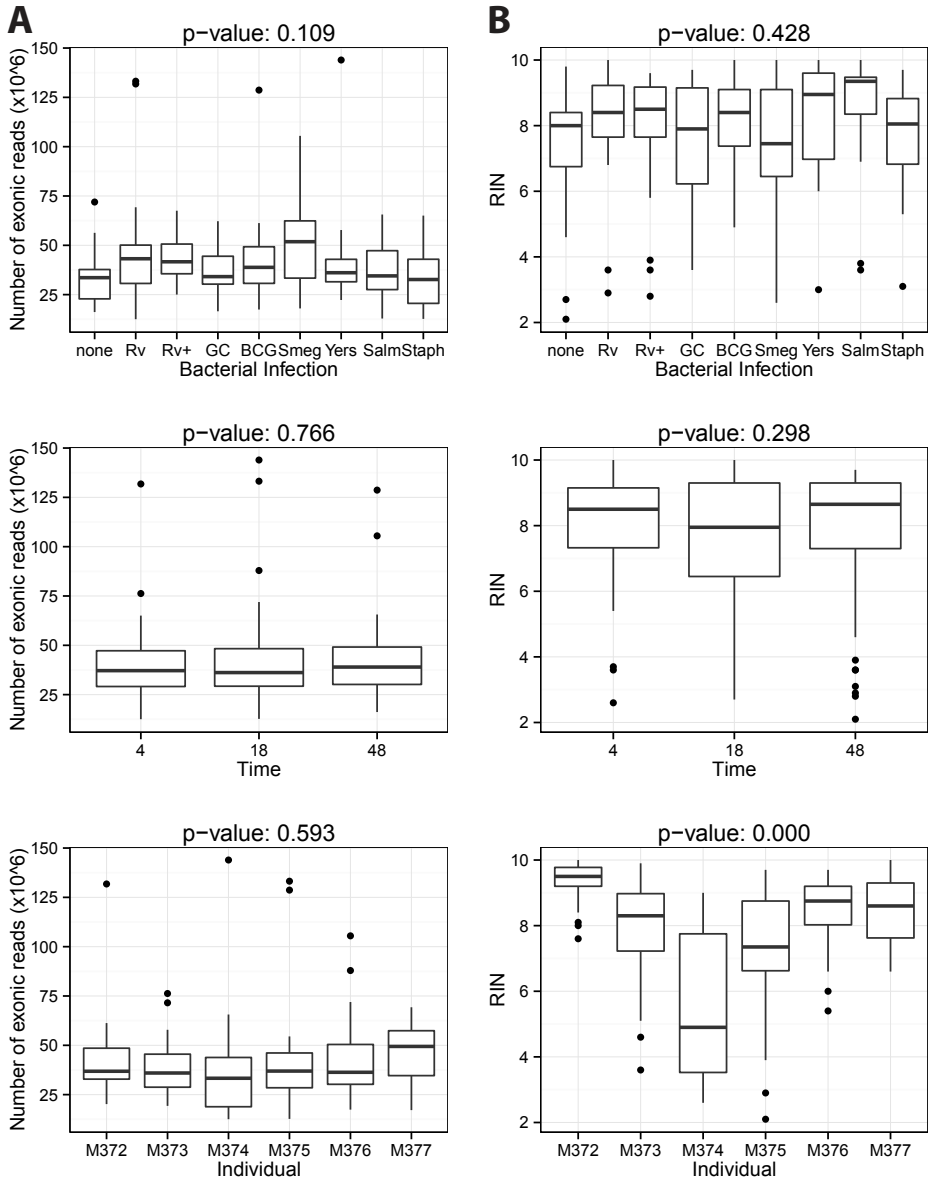


Figure 2.14: **Distribution of the number of exonic reads and RNA quality scores (RIN) across variables of interest.** (A) The number of exonic reads is evenly distributed across the bacterial infections, timepoints, and individuals. (B) The RIN scores are evenly distributed across the bacterial infections and timepoints; however, the RIN does vary between the individuals. The p-values are from an F-test.

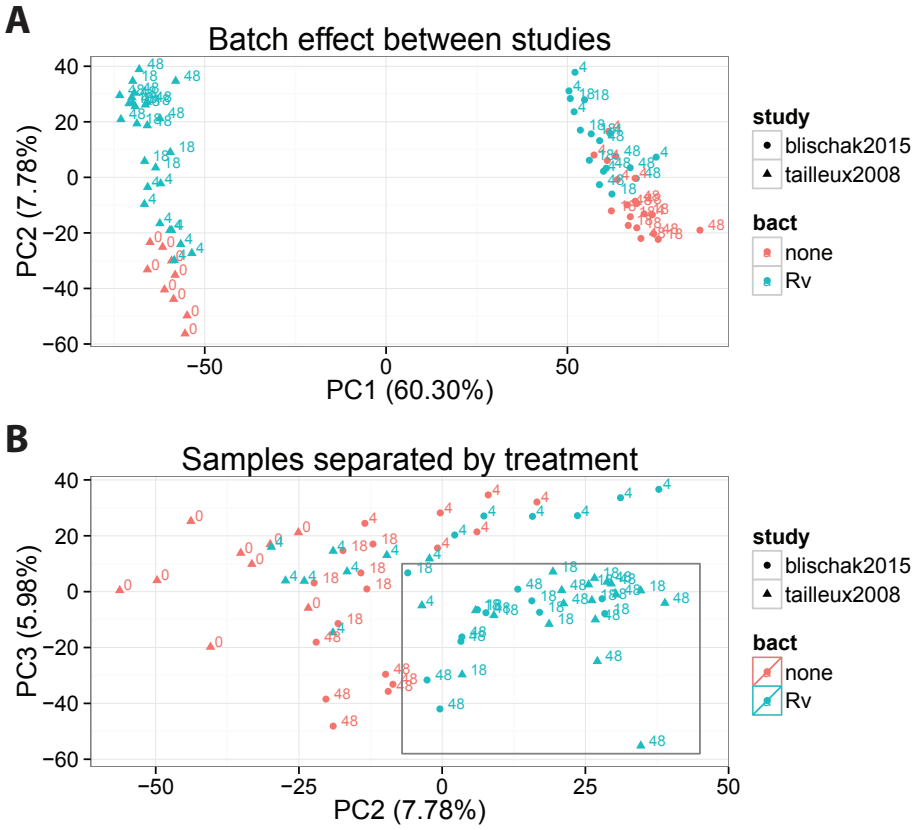


Figure 2.15: **Comparison to Tailleux et al., 2008.** We compared our RNA-seq data to the microarray data of Tailleux et al., 2008 [293] to confirm a consistent signature of infection. From our experiment, we used the batch-corrected TMM-normalized $\log_2$-transformed counts per million (CPM) from the 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. 2.1A) from the two different studies cluster together. The quantile-normalized data from Tailleux et al., 2008 [293] is available at `https://bitbucket.org/jdblischak/tb-data`.

2.8.2 Supplementary Tables

Table 2.2: **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 2.3: **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 2.4: **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. 2.2, 2.3, and 2.4. The columns "full_time_course", "time_18h", and "time_48h" correspond to Fig. 2.2, 2.3, and 2.4, respectively.

Table 2.5: **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. 2.2, 2.3, and 2.4. It is the same information as Supplementary Table 2.4, 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 2.6: **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. 2.2, 2.3, and 2.4.

Table 2.7: **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 2.8: **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 2.9: **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 2.10: **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

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 worlds 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.

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 [332, 331, 101]. Successful

treatment requires months of antibiotic therapy [280], and the difficulty of adhering to the full treatment regimen has lead to the emergence of drug-resistant strains of *Mycobacterium tuberculosis* (MTB) [265].

Approximately a third of the worlds 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 [213]. In contrast, approximately 10% of individuals will develop active TB after infection with MTB [224, 227]. 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 [149, 54, 50, 210], genome wide association studies (GWAS) have only identified a few loci with low effect size [302, 194, 301, 239, 47, 59, 278]. 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 be susceptible or resistant to infectious diseases, including tuberculosis [300] and acute rheumatic fever [33].

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 absolute 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 [302]. 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 genes 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).

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

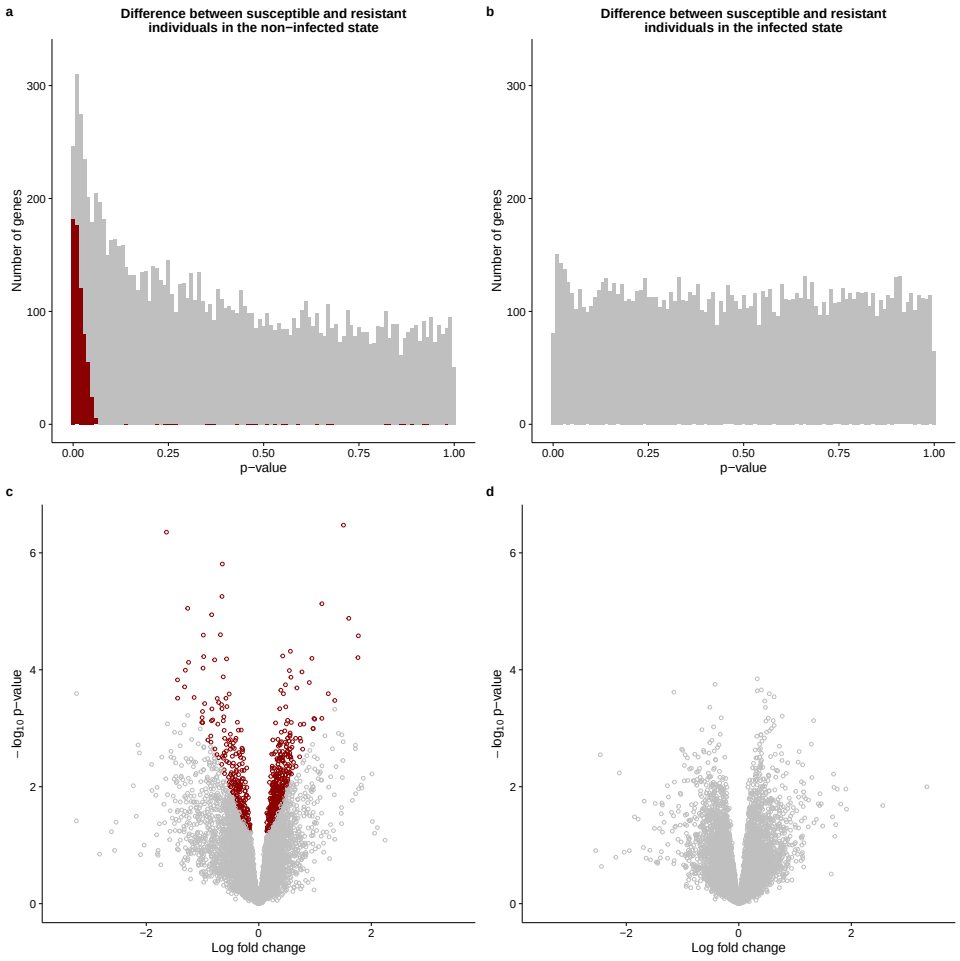


Figure 3.1: **Differential expression analysis.** The top panel contains the distribution of unadjusted p-values after testing for differential expression between susceptible and resistant individuals in the (a) non-infected or (b) infected state. The bottom panel contains the corresponding volcano plots 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%.

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.

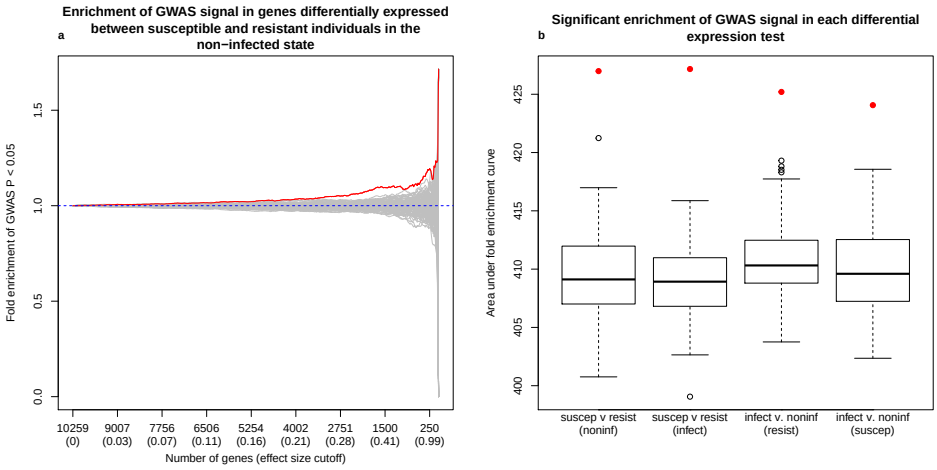


Figure 3.2: **Comparison of differential expression and The Gambia GWAS results.** (a) The y-axis is the fold enrichment (y-axis) of genes assigned a SNP with p-value less than 0.05 from the GWAS in The Gambia [302]. The x-axis is bins of genes with increasingly stringent effect size cutoffs of the absolute 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).

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 [17], 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 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

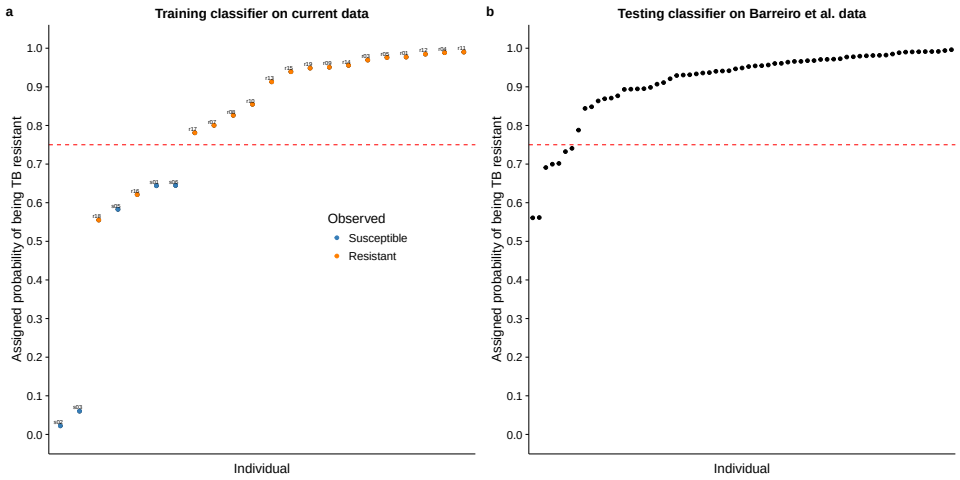


Figure 3.3: **Classifying TB susceptible individuals using a support vector machine model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, 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 [17].

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 [293]. 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 [64, 40], phagolysosomal acidification, and antigen processing. In particular, *FEZ2*, a suppressor of autophagosome formation [281], 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 [289, 122, 113], 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 [91, 103, 104, 186], 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 [300]. 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 [207]. 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) [300]. 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 [25]. *UNC13A* is involved in vesicle formation [290]. 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 [17]. 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 [22, 227, 24]. 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 [193]. 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 [213], 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 $\sim$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 [17]. 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 [183] and discarded non-uniquely mapping reads. We downloaded the exon coordinates of 19,800 Ensembl [340] protein-coding genes (Ensembl 83, Dec 2015, GRCh38.p5) using the R/Bioconductor [126] package biomaRt [74, 75] and assigned mapped reads to these genes using featureCounts [184].

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 [276, 175, 252] 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 [277]. To jointly model the data with voom and duplicateCorrelation, we followed the recommended best practice of running both voom and duplicateCorrelation twice in succession [188].

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 [286] 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 [302]. 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 [126] package SNPlocs.Hsapiens.dbSNP144.GRCh38 and matched SNPs to nearby genes using GenomicRanges [176]. 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 [148]. 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 [17], and thus assumed to be similar to that in the general population ($\sim$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 [17]) 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 [252] (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 [175] since the data were no longer counts). Specifically, we tested for differential expression between susceptible and resistant 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 [166] to train 3 different machine learning models: elastic net [93], support vector machine [151], and random forest [185] (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 [164]. Our processing pipeline used the general bioinformatics software FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`), MultiQC [86], samtools [180], and bioawk (`https://github.com/lh3/bioawk`). We used R [245] for all statistics and data visualization. We obtained gene annotation information from the Ensembl [340] and Lynx [291] 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 [76] 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

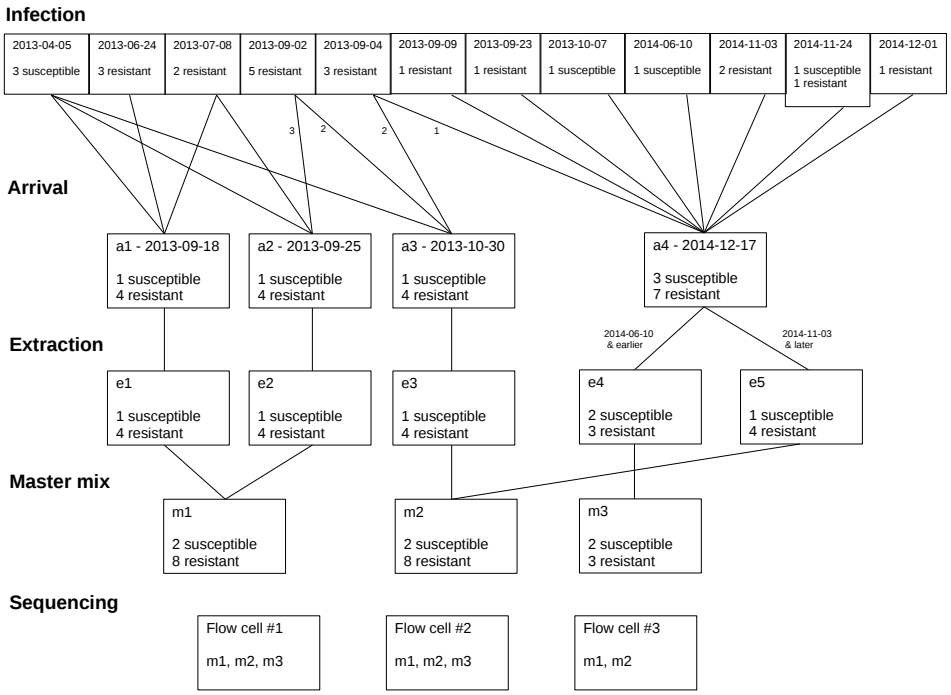


Figure 3.4: **Batch processing.** We designed the processing of the samples to minimize the introduction of technical batch effects. Specifically, we attempted to balance the processing of samples obtained from susceptible and resistant individuals. In the diagram, each box represents a batch. Infection 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.

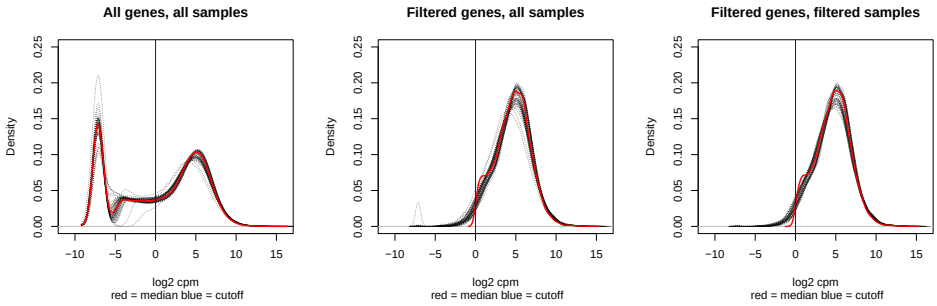


Figure 3.5: **Gene expression distributions before and after filtering genes and samples.** The $\log_2$ counts per million (cpm) of each sample is plotted as a dashed gray line. The solid red line represents the median value across all the samples. The vertical solid blue line at $x = 0$ represents 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.

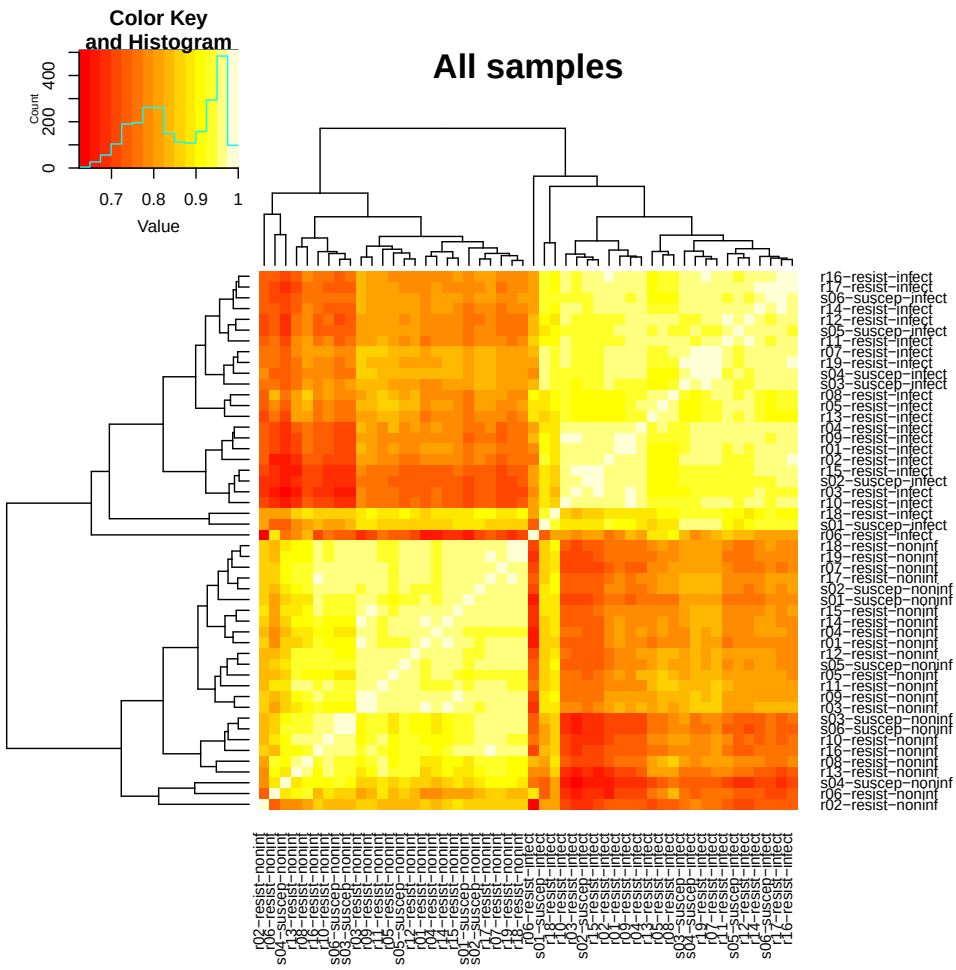


Figure 3.6: **Heatmap of correlation matrix of samples.** Each square represents the Pearson correlation between the $\log_2$ cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of hierarchical 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.

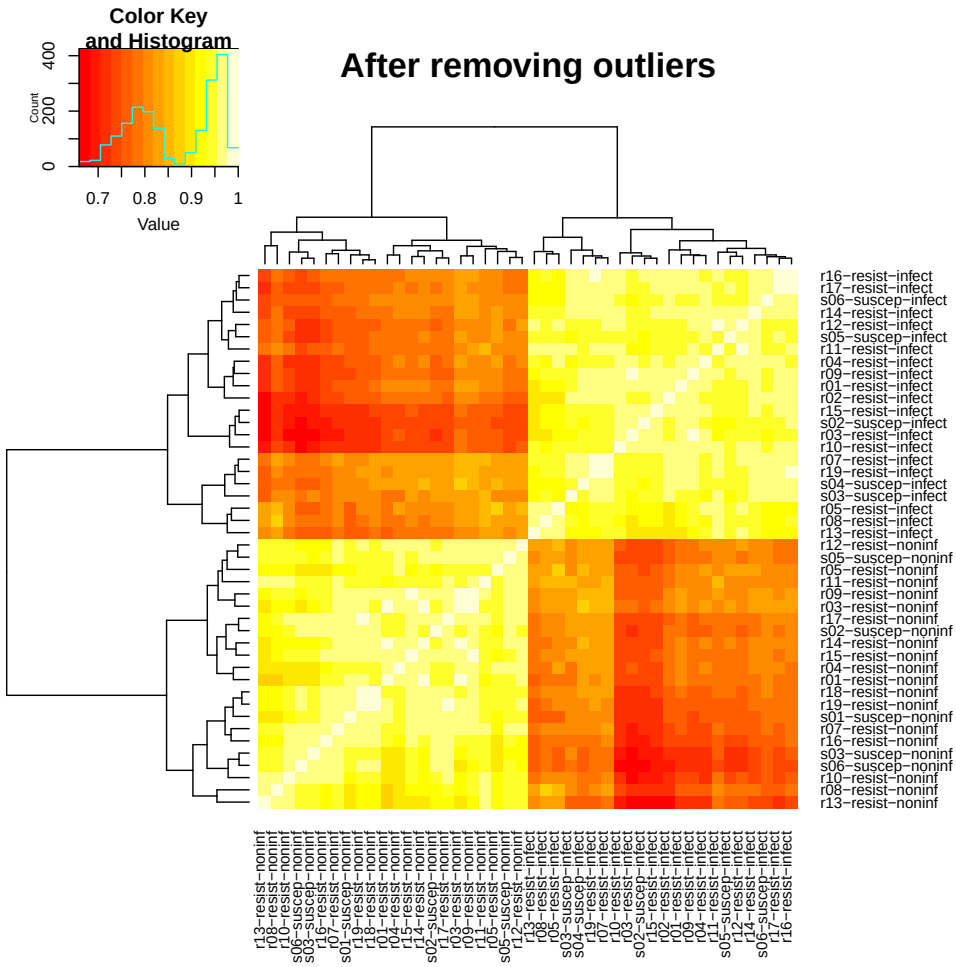
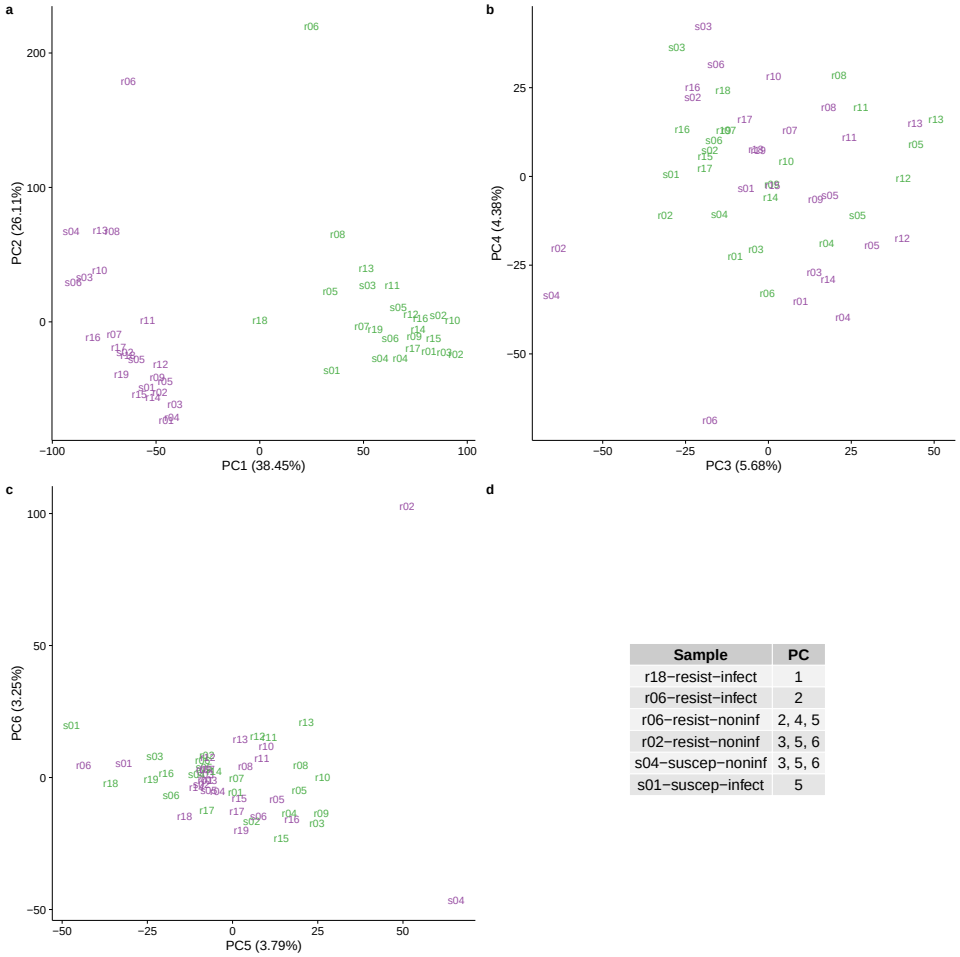


Figure 3.7: **Heatmap of correlation matrix after removing outliers.** Each square represents the Pearson correlation between the $\log_2$ cpm expression values of two samples. Red indicates a low correlation of zero and white represents a high correlation of 1. The dendrogram displays the results of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

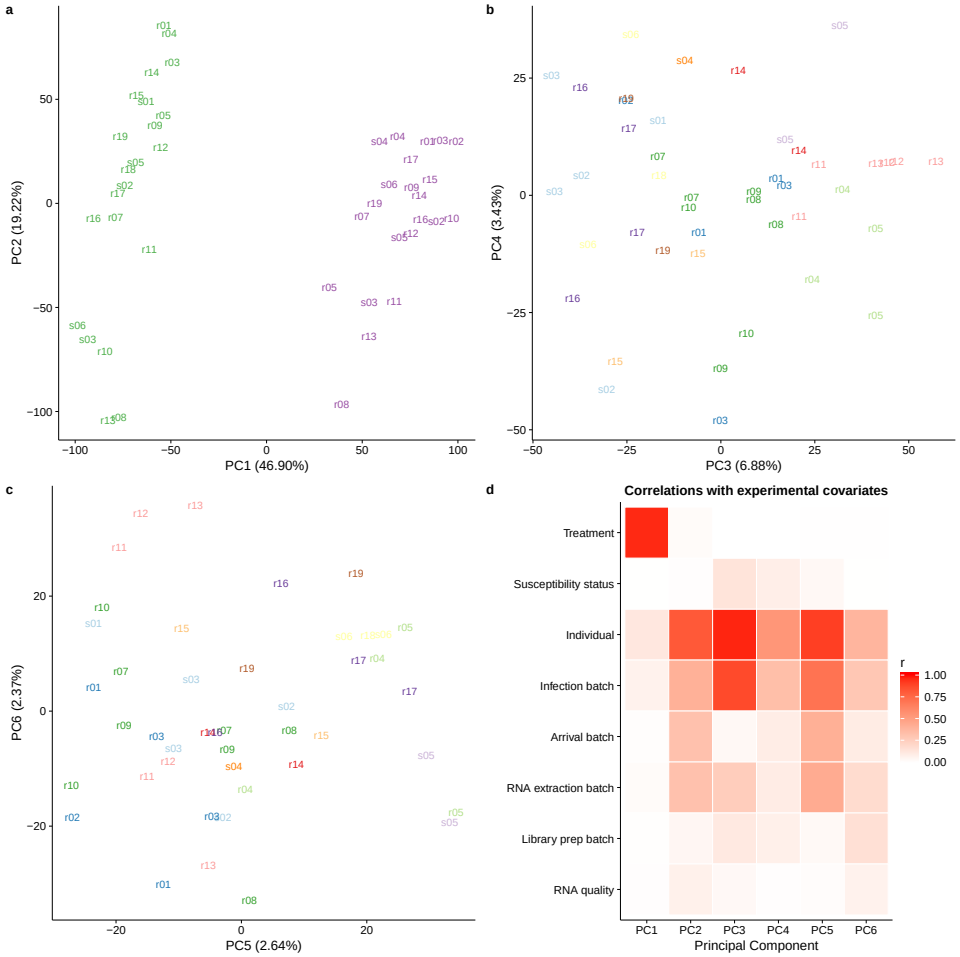


Figure 3.9: **Check for technical batch effects using principal components analysis (PCA).** (a) PC1 versus PC2. The text labels are the individual identifiers. Purple indicates non-infected samples and green indicates infected. (b) PC3 versus PC4. The colors indicate the different infection batches. (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).

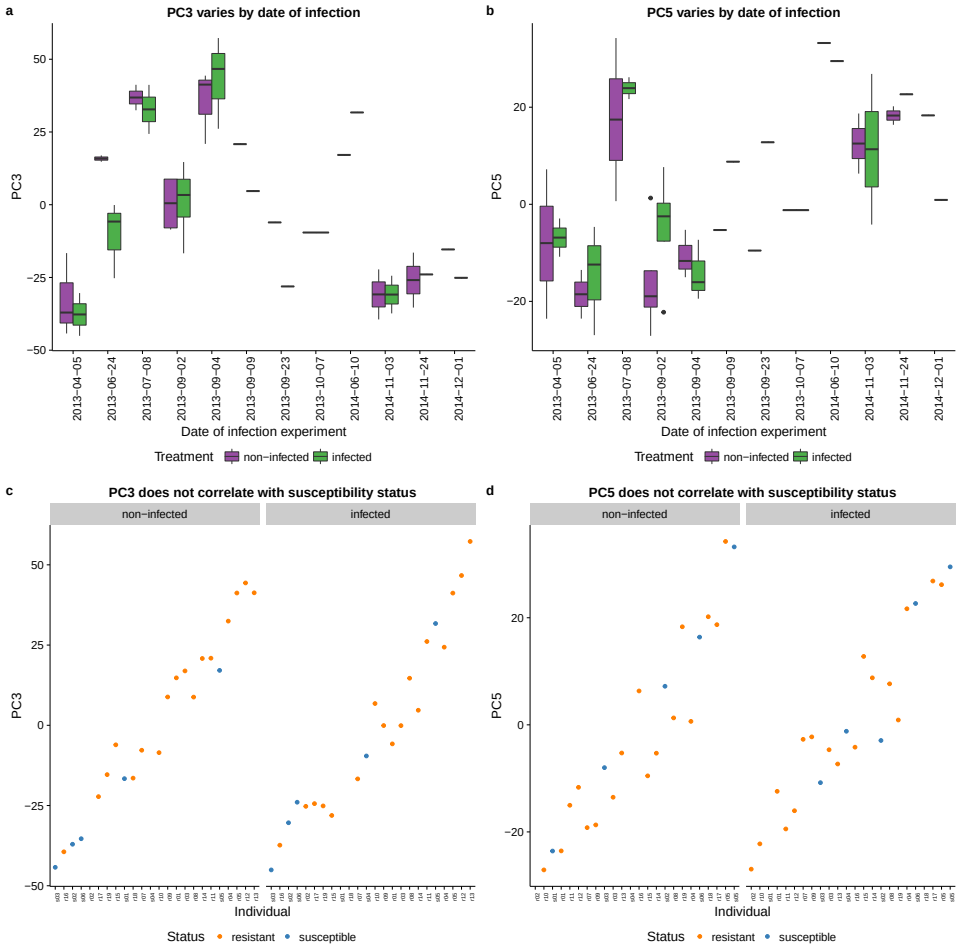


Figure 3.10: **Check for confounding effect of infection batch.** PC3 (a) and PC5 (b) varied by the date of infection. Non-infected samples are in purple and infected samples in green. Importantly, however, this technical variation arising from infection batch did not correlate with the susceptibility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

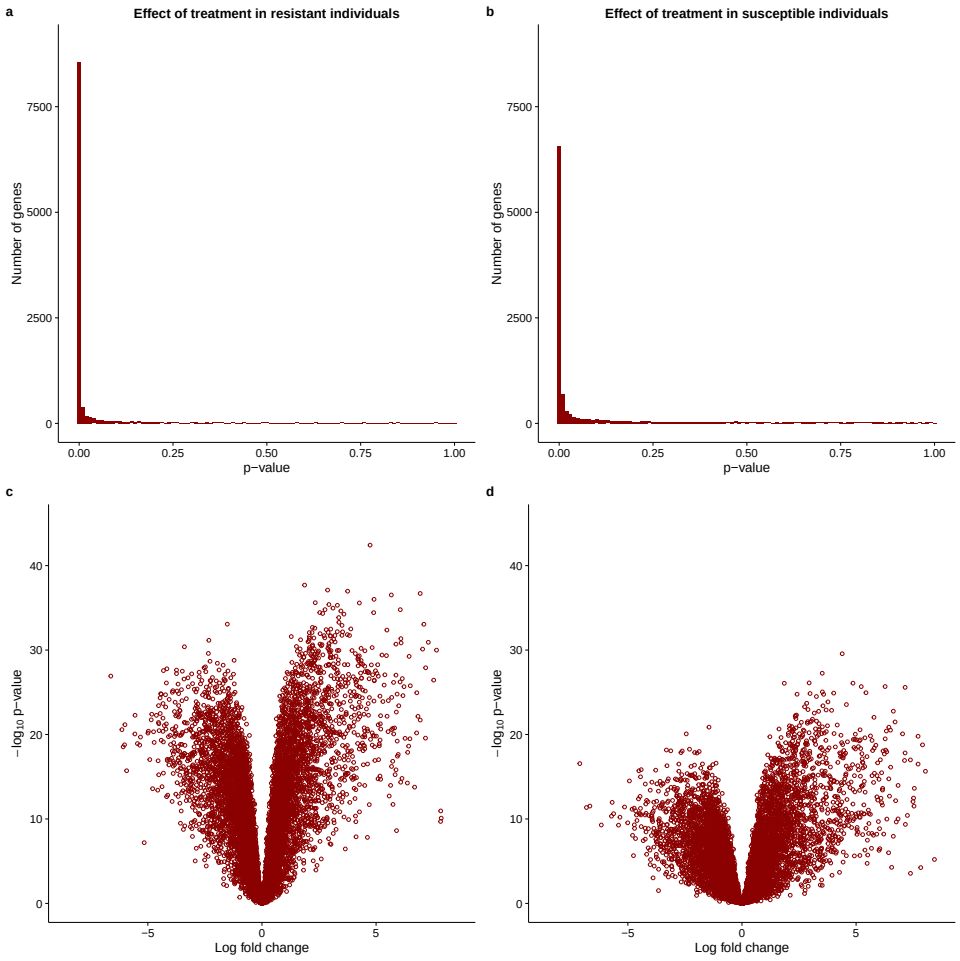


Figure 3.11: **Effect of treatment with MTB.** The top panel contains the distribution of unadjusted p-values after testing for differential expression between the non-infected and infected states in (a) resistant and (b) susceptible individuals. The bottom panel contains the corresponding volcano plots 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.

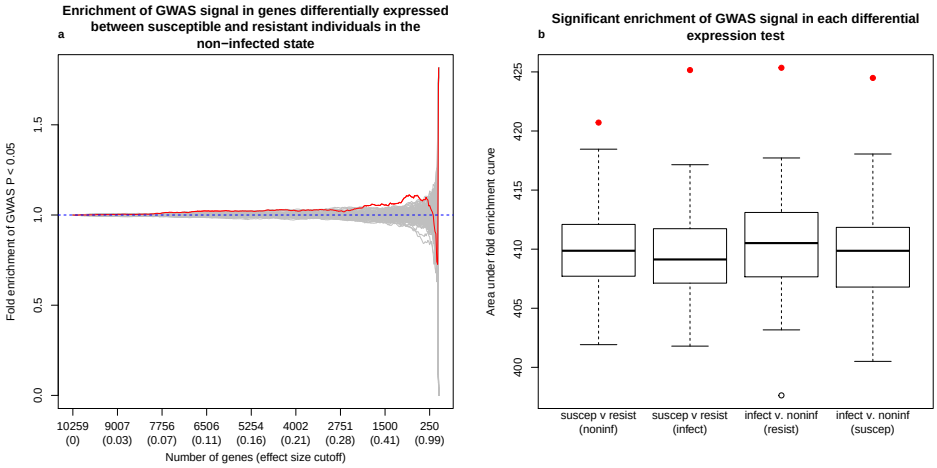


Figure 3.12: **Comparison of differential expression and Ghana GWAS results.** (a) The y-axis is the fold enrichment (y-axis) of genes assigned a SNP with p-value less than 0.05 from the GWAS in Ghana [302]. The x-axis is bins of genes with increasingly stringent effect size cutoffs of the absolute log 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).

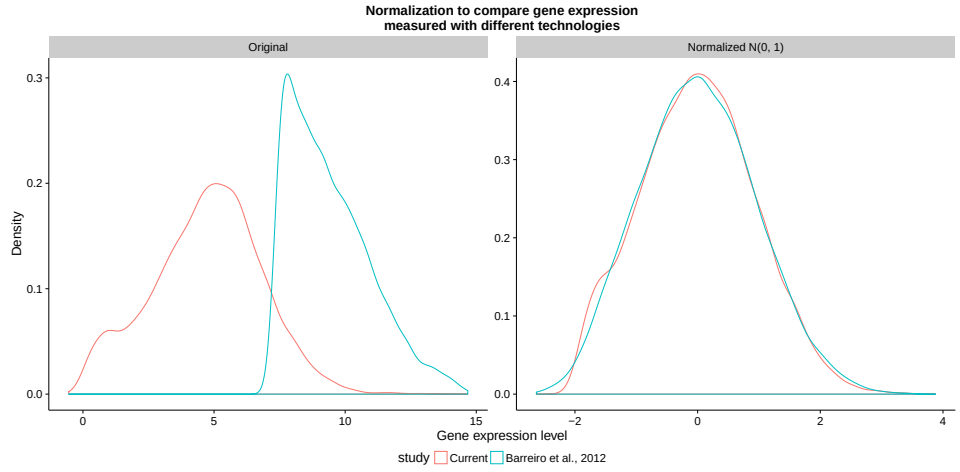


Figure 3.13: **Normalizing gene expression distributions.** (left) The distribution of the median log2 cpm of the RNA-seq data from the current study in red compared to the distribution of the median gene expression levels of the microarray data from Barreiro et al., 2012 [17] in blue. (right) The distributions of the same data sets after normalizing each sample to a standard normal distribution.

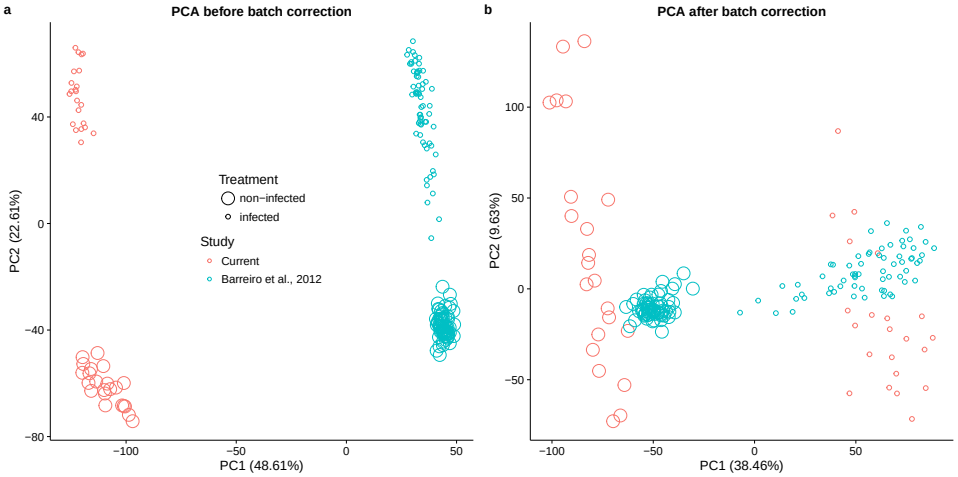


Figure 3.14: **Principal components analysis (PCA) of combined data sets.** (a) PC1 versus PC2 of the combined data set of the RNA-seq data from the current study (red) and the microarray data from Barreiro et al., 2012 [17] (blue). The large circles are non-infected samples, and the small circles are 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).

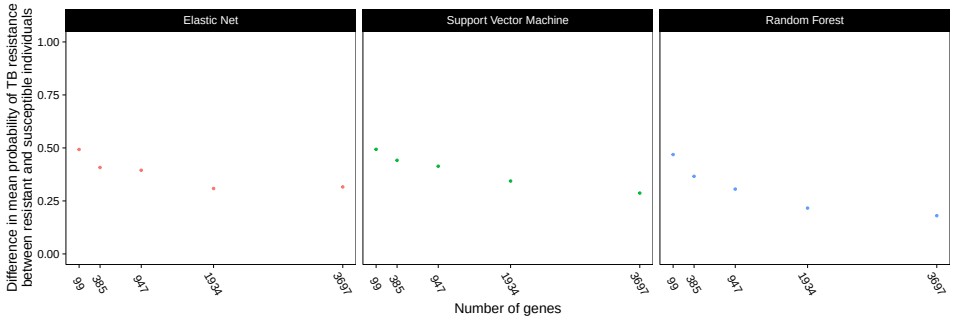


Figure 3.15: **Comparing the classification results of different methods and number of input genes.** We compared 3 different machine learning methods (elastic net, support vector machine, random forest) and used 5 different sets of input genes. The input genes (x-axis) were obtained by varying the q-value 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.

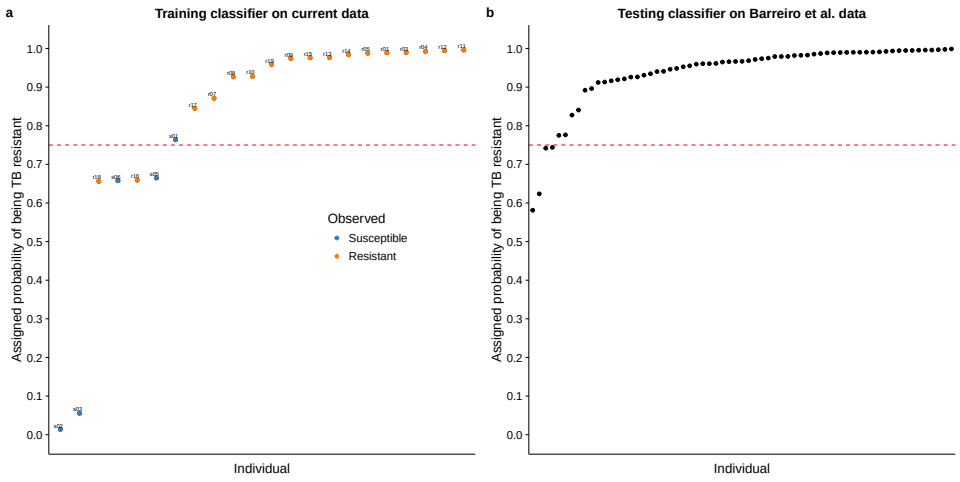


Figure 3.16: **Classifying TB susceptible individuals using an elastic net model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, and orange 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 [17].

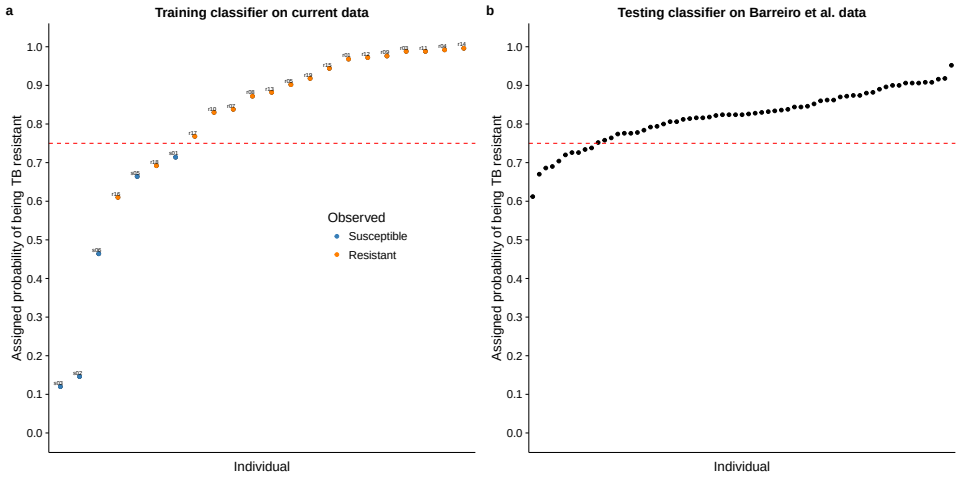


Figure 3.17: **Classifying TB susceptible individuals using a random forest model.** (a) The estimates of predicted probability of TB resistance from the leave-one-out-cross-validation for individuals in the current study. The blue circles represent individuals known to be susceptible to TB, and orange 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 [17].

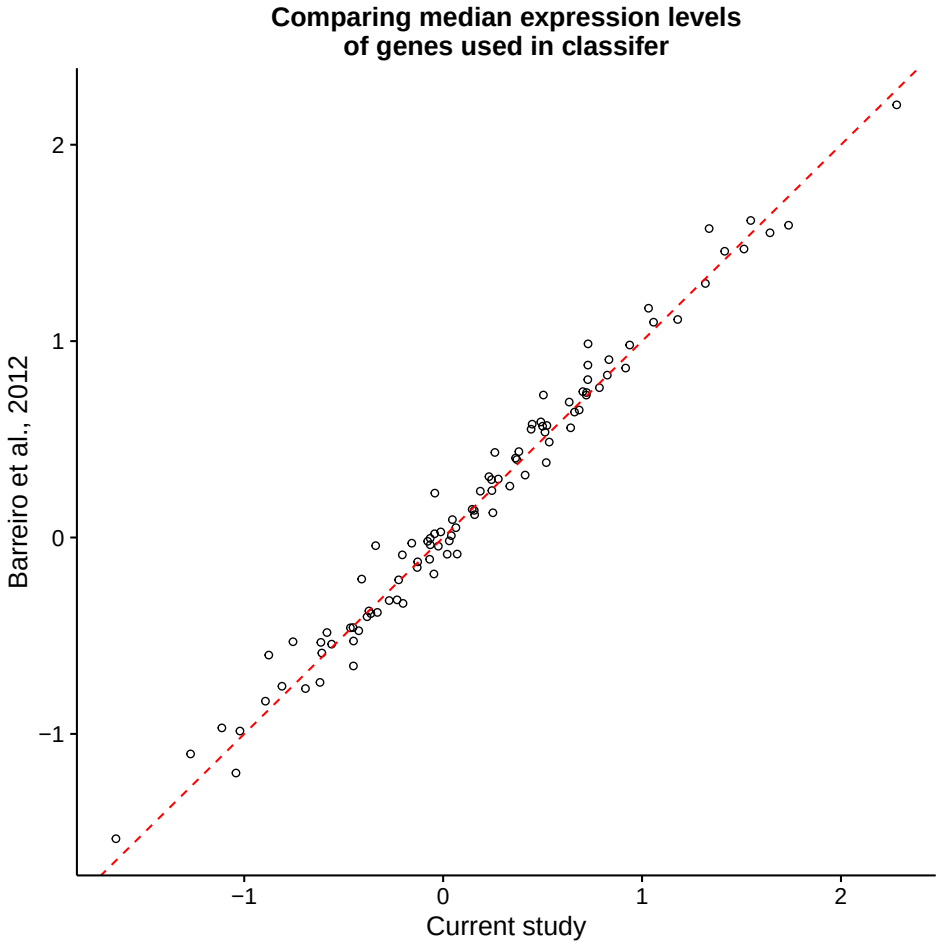


Figure 3.18: **Comparing gene expression between the two studies.** After normalization and batch-correction, the median expression levels of the 99 genes used in the classifier were similar between the samples in the current study and those in Barreiro et al., 2012 [17]. The dashed red line is the 1:1 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. (tds)

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. (tds)

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 [17]. 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

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.

4.1 Introduction

Single-cell genomic technologies can be used to study the regulation of gene expression at unprecedented resolution [191, 260]. 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 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 [192, 111, 71], the innate immune response [266, 137], and pharmacogenomics [208, 158].

Yet, there are a number of outstanding challenges that arose in parallel with the application of single cell technology [285]. 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 [29, 105]. The most commonly used spike-in was developed by the External RNA Controls Consortium (ERCC) [142]; 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 [29, 105, 68, 308]. 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 introduced to single cell sample processing [160, 94, 38, 269]. 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 [137, 134, 105]. It is assumed that most sources of variation in single cell gene expres-

sion studies can be accounted for by using the combination of UMIs and a spike-in based standardization [134, 308]. 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 [334, 192]. 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 [192, 161]. One of the biggest limitations of using the C1 system, however, is that single cell capture and preparation from different conditions are fully independent [116]. 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 [285]. 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.

4.2 Results

4.2.1 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, 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

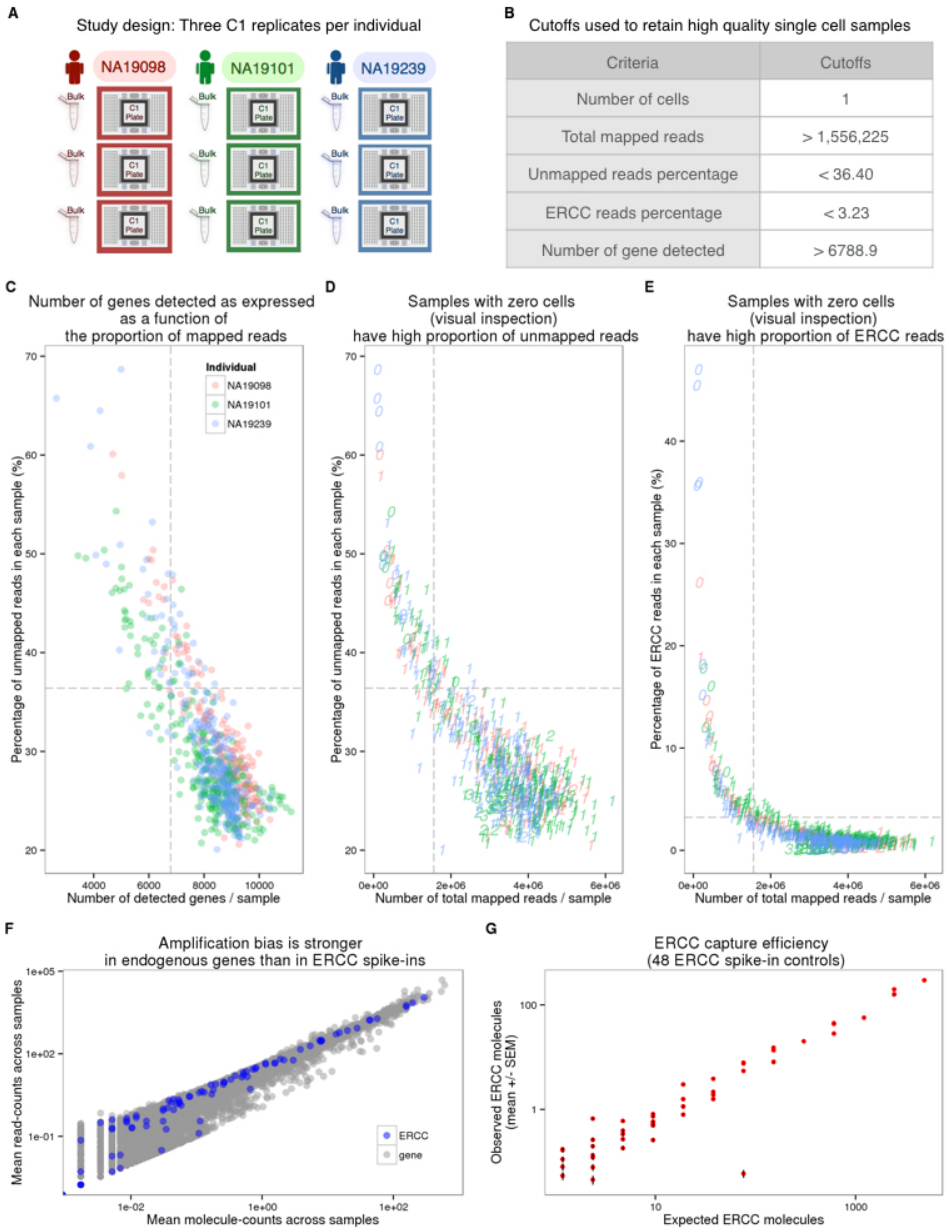


Figure 4.1: **Experimental design and quality control of scRNA-seq.** (A) Three C1 96 well-integrated fluidic circuit (IFC) replicates were collected from each of the three Yoruba individuals. A bulk sample was included in each batch. (B) Summary of the cutoffs used to remove data from low quality cells 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. (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.

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 [134, 105] 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. 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 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).

4.2.2 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 se-

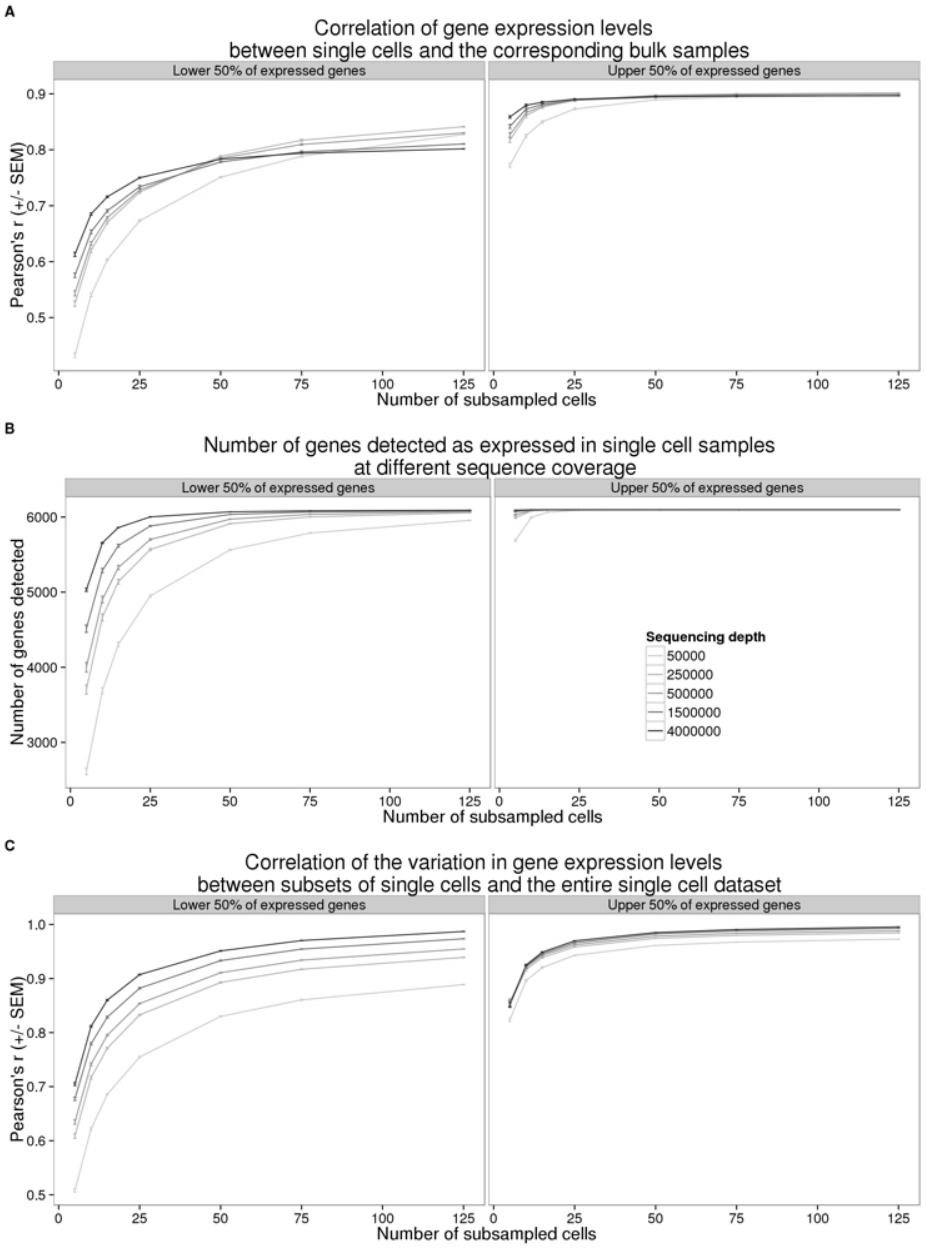


Figure 4.2: **The effect of sequencing depth and cell number on single cell UMI estimates.** Sequencing reads from the entire data set were subsampled to the indicated sequencing depth and cell number, and subsequently converted to molecules using the UMIs. Each point represents the mean +/- SEM of 10 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.

quencing 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 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.

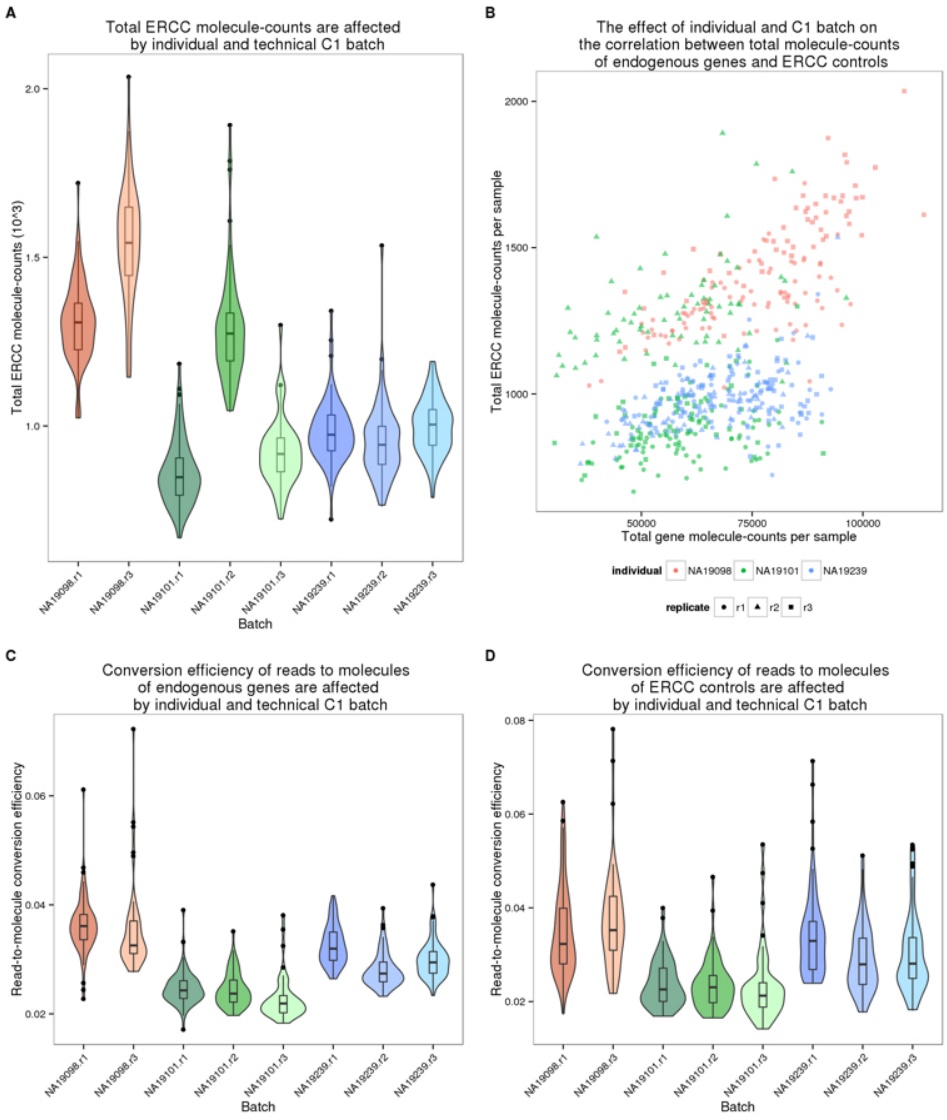


Figure 4.3: **Batch effect of scRNA-seq data using the C1 platform.** (A) Violin plots of the number of total ERCC spike-in molecule-counts in single cell samples per C1 replicate. (B) Scatterplot of the total ERCC molecule-counts and total gene molecule-counts. The colors represent the three individuals (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 $P < 0.01$).

4.2.3 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' [94]). 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 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 [277]. 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

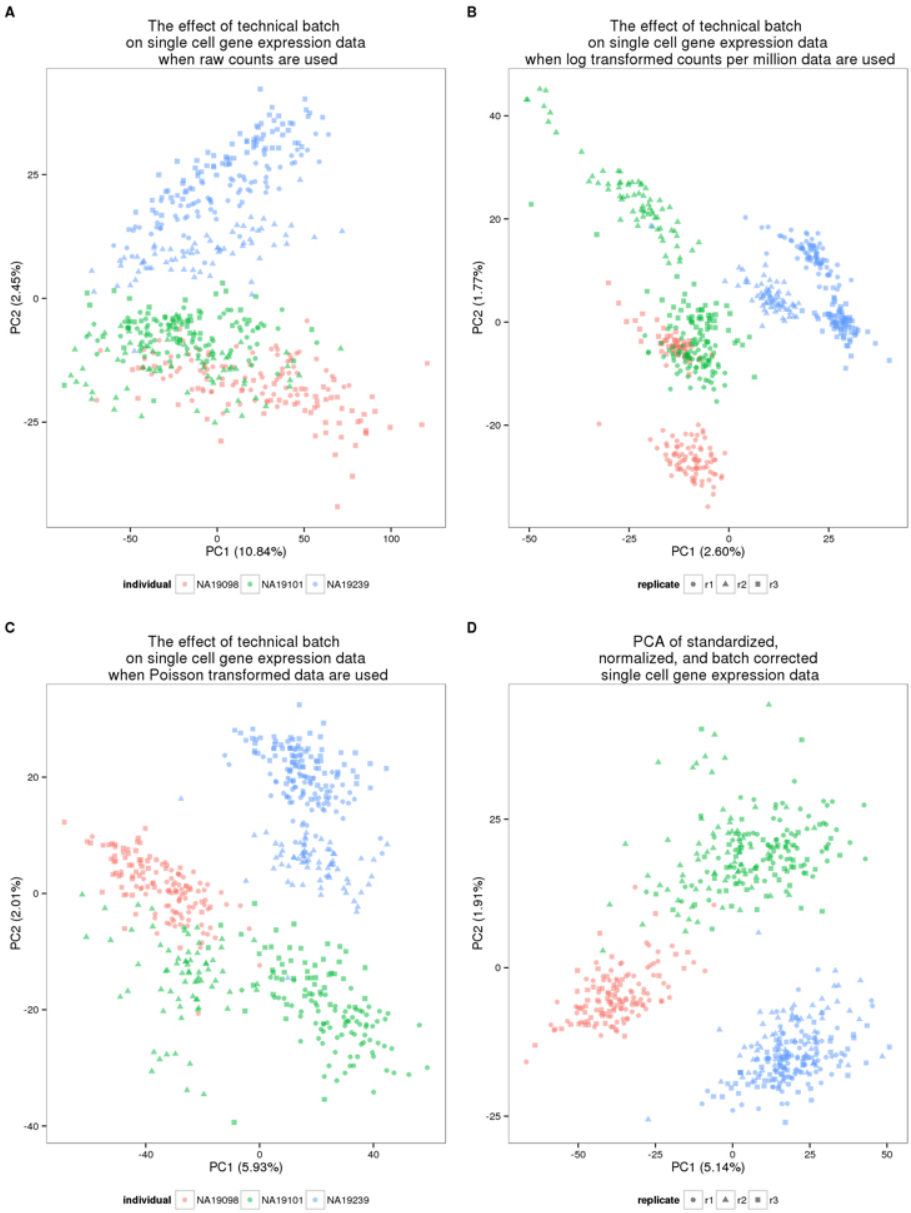


Figure 4.4: **Normalization and removal of technical variability.** Principal component (PC) 1 versus PC2 of the (A) raw molecule counts, (B) $\log_2$ counts per million (cpm), (C) Poisson transformed expression levels (accounting for technical variability modeled by the ERCC spike-ins), and (D) batch-corrected 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.

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 [249]. In this context, regulatory noise is generally defined as the coefficient of variation (CV) of the gene expression levels of single cells [88]. 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 [29, 266]) 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).

4.3 Discussion

4.3.1 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,

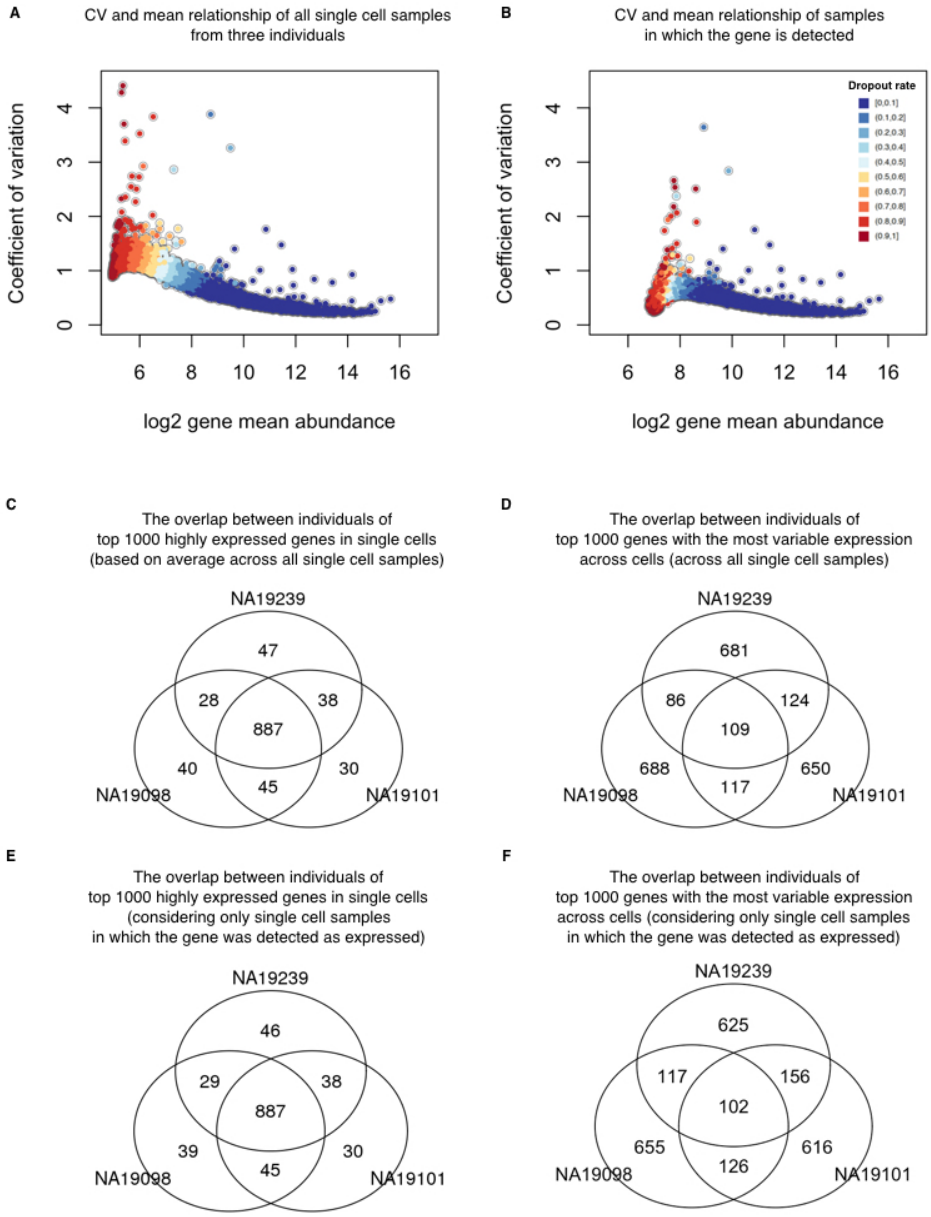


Figure 4.5: **Cell-to-cell variation in gene expression.** Adjusted CV plotted against average molecule counts across all cells in (A) and across only the cells in which the gene is expressed (B), including data from all three individuals. Each dot represents a gene, and the color indicates the corresponding 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.

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 [147], 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 [298] (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 previously performed for single cell sequencing with the C1 platform without the use of UMIs [334, 240], 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.

4.3.2 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 [29, 105, 68, 308]. 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 [251] (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 [264]. 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.

4.3.3 Outlook

Single cell experiments are ideally suited to study gene regulatory noise and robustness [28, 89]. 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.

4.4 Methods

4.4.1 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.

4.4.2 Cell culture of iPSCs

Undifferentiated feeder-free iPSCs reprogrammed from LCLs of Yoruba individuals in Ibadan, Nigeria (abbreviation: YRI) [298] were grown in E8 medium (Life Technologies) [46] 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.

4.4.3 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 [134]. 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 [132, 133, 134]. 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 [134]. 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 [237]. 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.

4.4.4 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 [134].

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.

4.4.5 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 [145]. To handle the UMI sequences at the 5' end of each read, we used umitools [143] 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 [183], 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 [184]. 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 [180]. Next, we converted read counts to molecule counts using UMI-tools [275]. 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.

4.4.6 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.

4.4.7 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%.

4.4.8 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.

4.4.9 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.

4.4.10 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 [48], 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).

4.4.11 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. [105]. 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. [105] 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.

4.4.12 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 [277]. 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 [252].

4.4.13 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 [134, 29]. We computed a measure of gene expression noise that is independent of RNA abundance across individuals [162, 219]. 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 [343]. After this transformation step, CV no longer had a polynomial relationship with mean gene molecule count (Supplementary Fig. 4.15).

4.4.14 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 [236].

4.4.15 Gene enrichment analysis

We used ConsensusPATHDB [150] to identify GO terms that are over-represented for genes whose variation in single cell expression levels were significantly difference between individuals.

4.4.16 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 [147], which detects sample contamination and/or mislabeling by comparing the polymorphisms 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 [298]. The genotypes included the HapMap SNPs with the 1000 Genomes Project SNPs [297] imputed, as previously described [203]. 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).

4.4.17 Data and code availability

The data have been deposited in NCBI's Gene Expression Omnibus [76] 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`.

4.5 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.

4.6 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.

4.7 Supplementary Information

4.7.1 Supplementary Figures

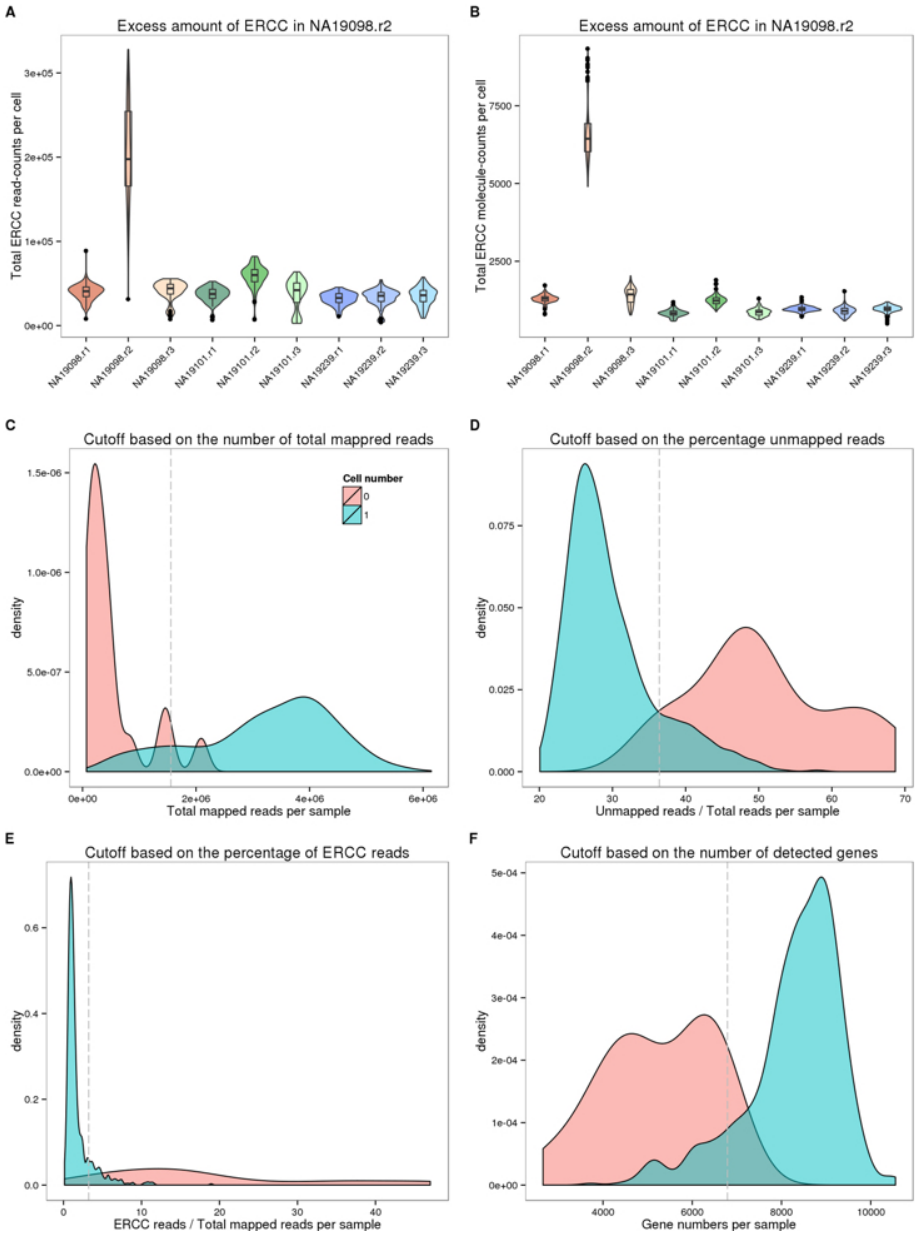


Figure 4.6: **Removal of low quality samples.** Violin plots of the total read-counts of ERCC spike-in controls in (A) and the total molecule-counts in (B) in single cell samples. The three colors represent the three individuals (NA19098 in red, NA19101 in green, and NA19239 in blue). (C-F) Density plots 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.

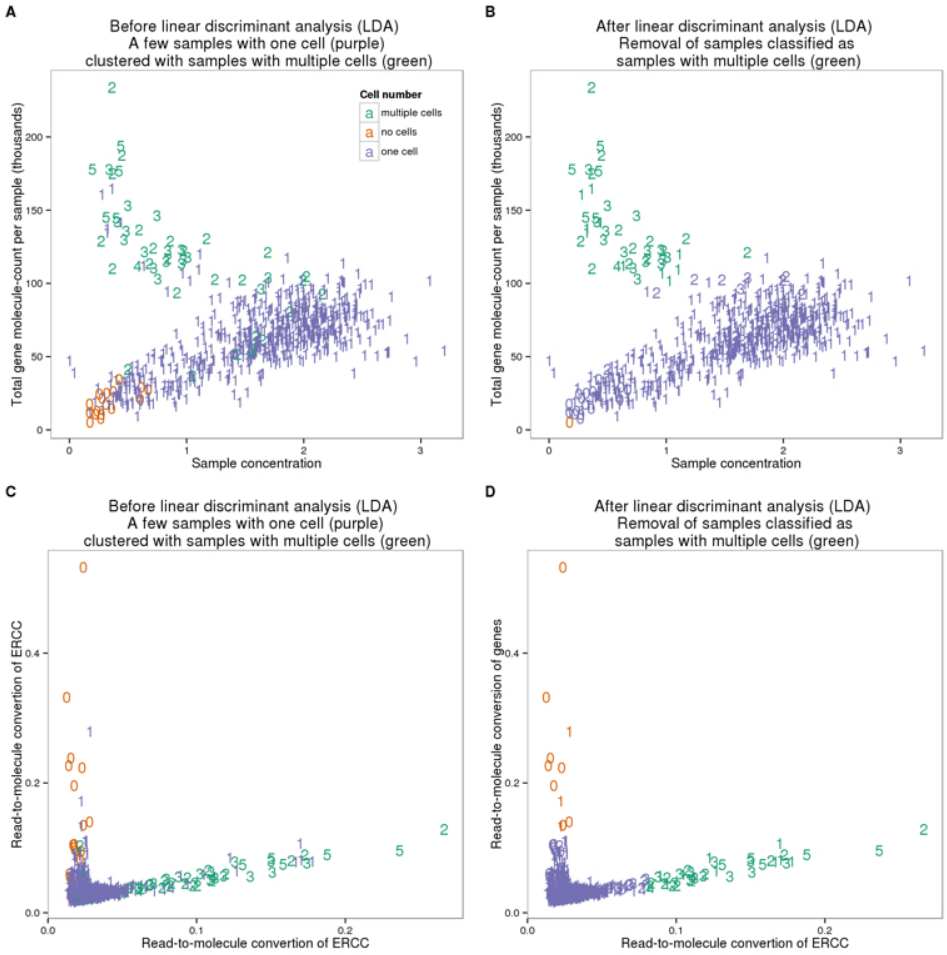


Figure 4.7: **Removal of samples with multiple cells.** Scatterplots of the three groups of samples (no cell in green, single-cell in orange, and two or more cells in purple) before (A) and after (B) the linear discriminant analysis (LDA) using sample concentration of cDNA amplicons (ng/μl) and the number 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.

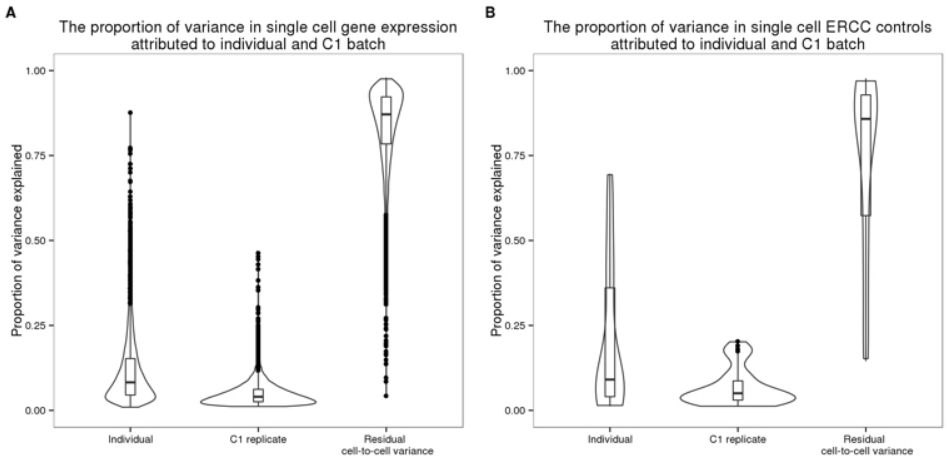


Figure 4.8: **Sources of cell-to-cell variance in per-gene expression profile.** Violin plots of the proportion of per-gene cell-to-cell variance that was due to individual sample of origin, different C1 replicates, and other single cell sample differences. These results were calculated from the molecule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

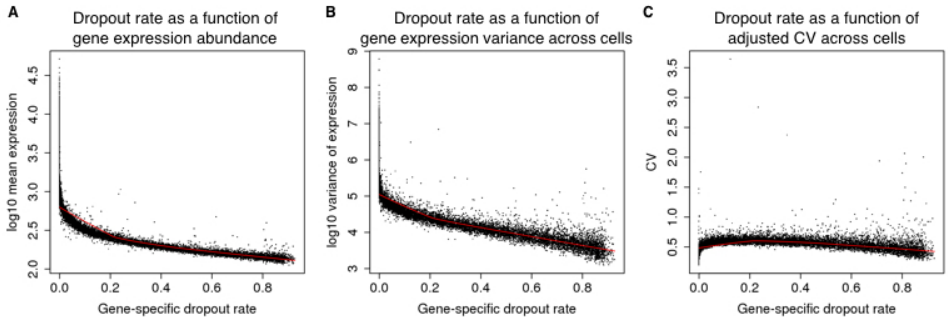


Figure 4.9: **The gene-specific dropout rate.** The gene-specific dropout rate (the proportion of cells in which the gene is undetected) and its relationship with $\log_{10}$ mean expression in (A), with $\log_{10}$ variance of expression in (B), and with the CV in (C) of the cells in which the gene 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).

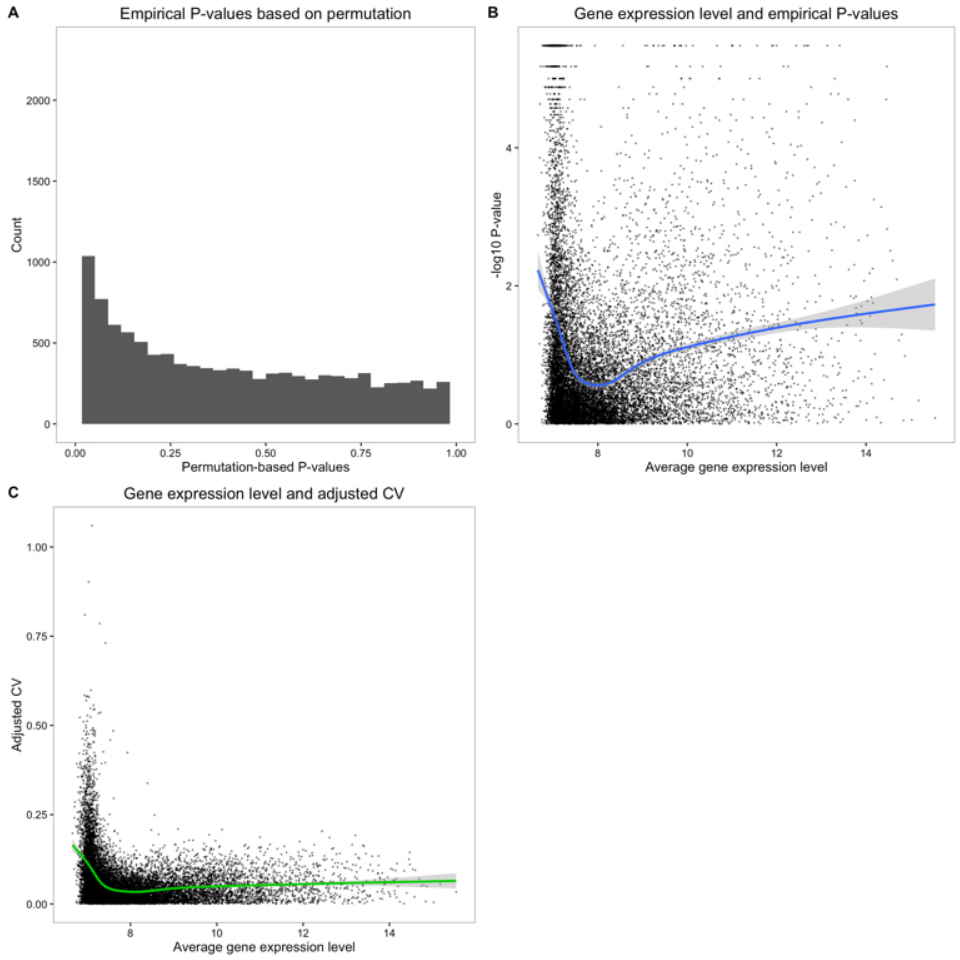


Figure 4.10: **Permutation-based *P*-value.** (A) Histogram of empirical P-values based on 300,000 permutations. (B) -$\log_{10}$ empirical P-values are plotted against average gene expression levels. Blue line indicates the fitted relationship between -$\log_{10}$ P-values and average $\log_2$ gene 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.

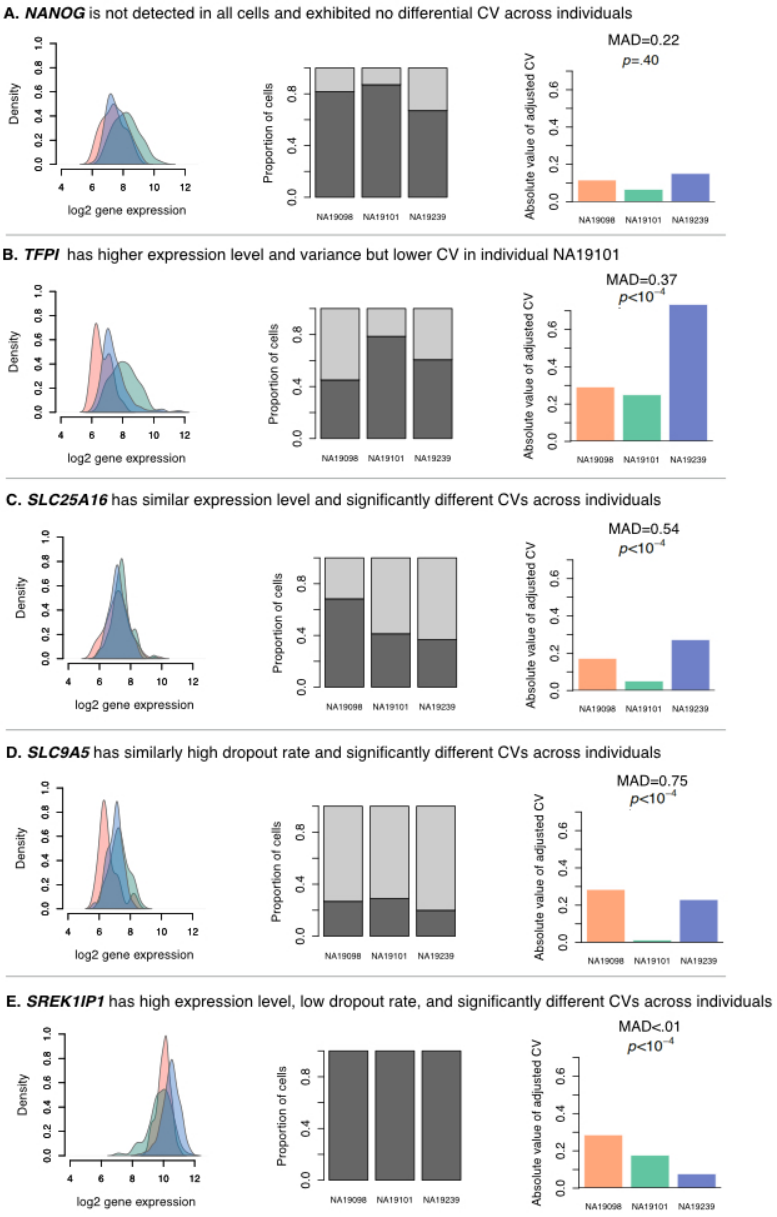


Figure 4.11: **Inter-individual differences in regulatory noise.** These 5 example genes illustrate various patterns of cell-to-cell gene expression variance. For each gene, the left panel shows the distribution of the $\log_2$ gene expression levels (considering only cells in which the gene is detected 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).

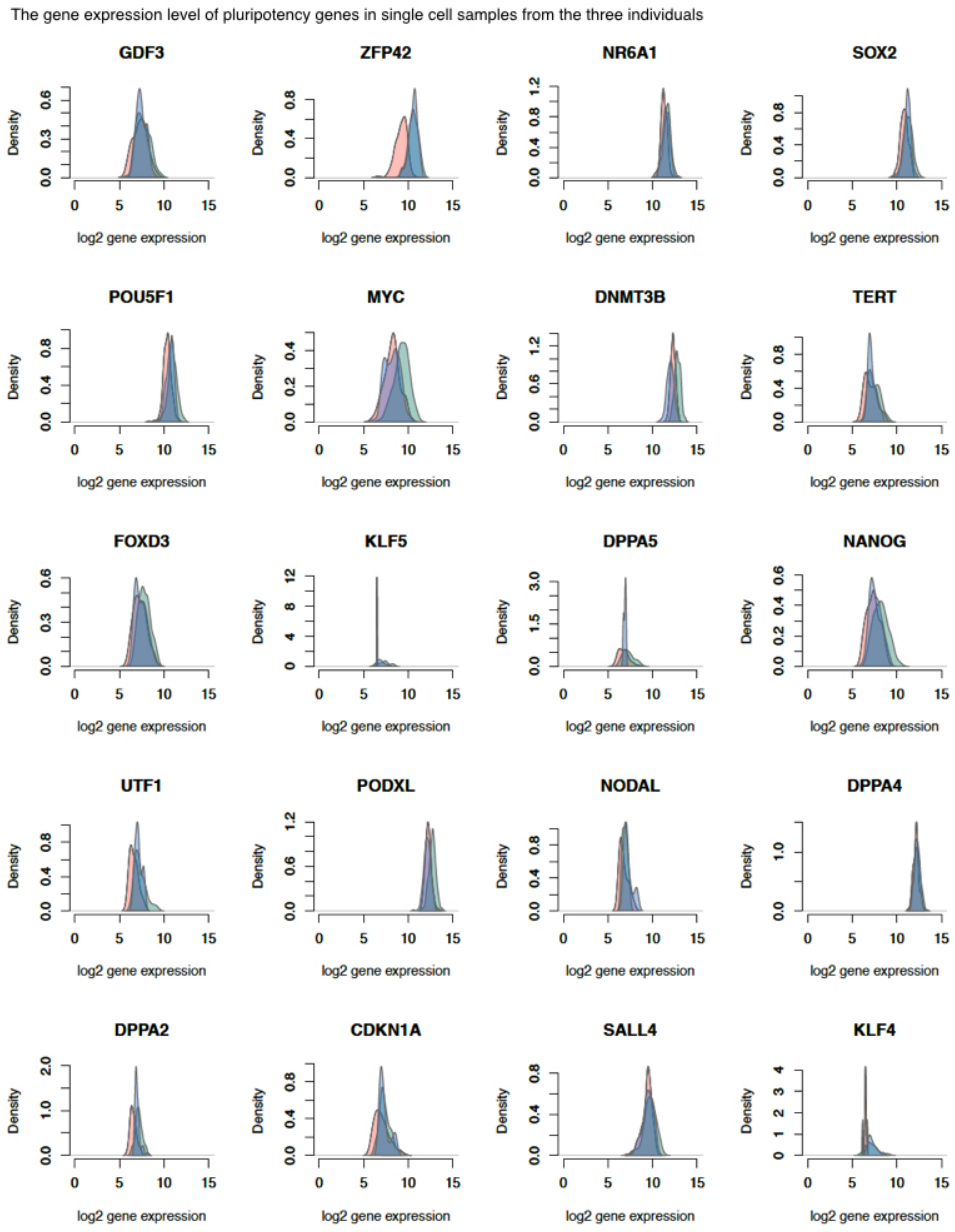
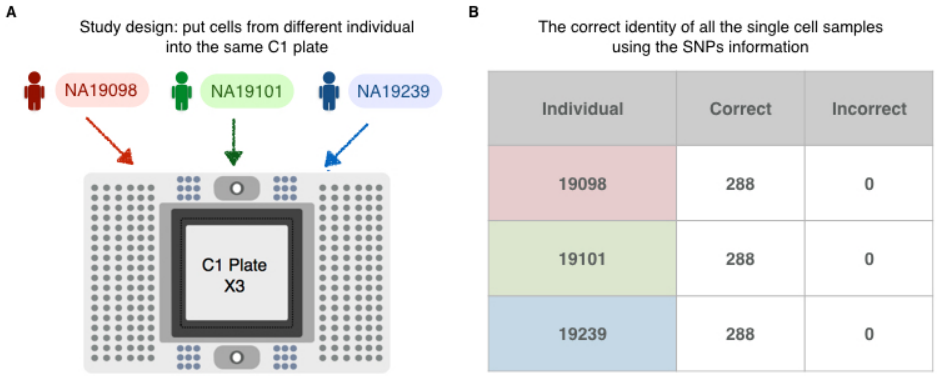


Figure 4.12: **Cell-to-cell variation of pluripotency genes.** Density plots of the distribution of $\log_2$ gene expression of key pluripotency genes across all single cells by individual. The peaks with lower gene expression values (log2 around 4) represent the cells in which the gene is undetected.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

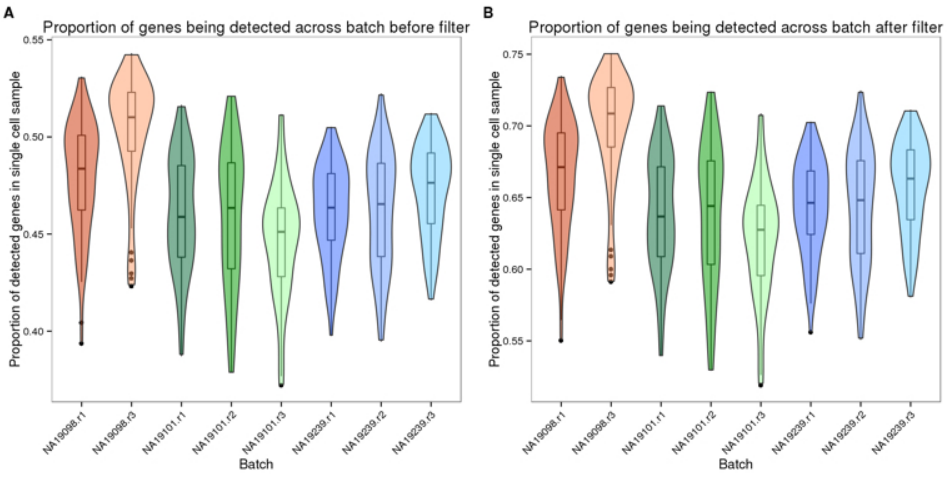


Figure 4.14: **The proportion of genes detected in single cell samples.** Violin plots of the proportion of genes detected, computed by the total number of detected genes in each single cell divided by the total number of genes detected across all single cells, before in (A) and after in (B) the removal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

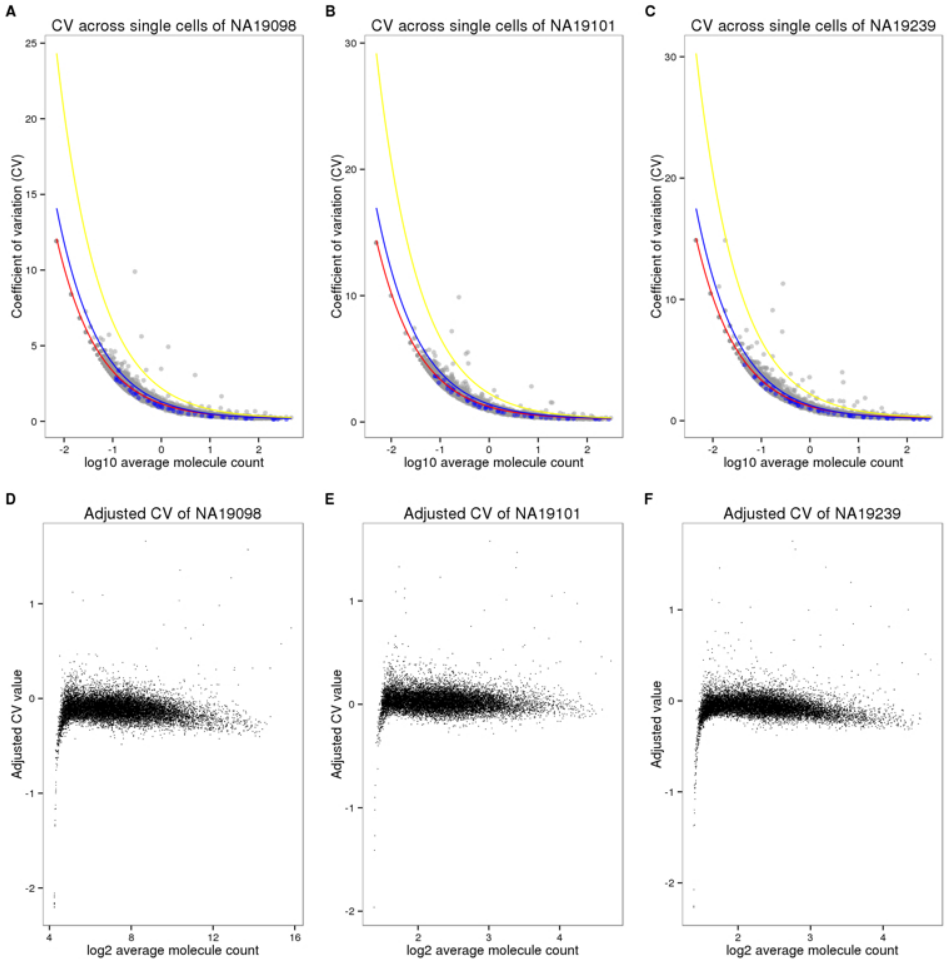


Figure 4.15: **Coefficients of variation (CV) before and after adjusting for gene mean abundance.** (A-C) CV plotted against average molecule counts across all cells for each individual [134]. Grey points represent endogenous genes, and blue points represent ERCC spike-in controls. The curves indicate 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.

4.7.2 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 [302, 194, 301, 239, 47, 59, 278]. 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 [25]. 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 [261, 242]. In my third approach, I established principles for the effective design of studies to measure gene expression levels in single cells [306]. 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 [25]. It is known that the innate immune response is important for fighting MTB infections [156]. Alveolar macrophages are the primary target of MTB, and they initiate the formation of granulomas to sequester MTB [270]. Furthermore, vaccines against TB have had limited

efficacy [319]. 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 [77, 246, 218, 44, 317, 293], and some have even compared the differences between the reponse to strains of MTB that vary in their virulence [56, 335]. 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 [229, 9, 252]. 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 [69, 90]. 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 [78, 205]. 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 [326]. 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. 2.2). The two most interesting patterns in regards to understanding the innate immune response to MTB were "MTB" and "Virulent" (Fig. 2.3,2.4). 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 [124, 27, 218, 140]. 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 [326], 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) [152].

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. 2.11). 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 [302] (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 [17] were predicted to be susceptible to TB, very close to the estimated population average of 10% [224, 227]. 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 [300, 294, 231], all the studies, including my own [25], 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 [300]. 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 [25]. 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 [72] 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 [207], 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) [306]. scRNA-seq is a relatively new technique [182, 191, 260, 106, 285, 15] 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 [11, 178, 100]. 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 [29, 105, 134, 68, 308] ; 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 [116]). 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 [128] 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 [160, 134]. 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 [147] (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 [251]. Using this study design, well 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 [11, 178, 100]. There will 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. 2.7). 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 [17]. 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] 1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[2] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[3] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.

[4] Luciano Adorini. Intervention in autoimmunity: the potential of vitamin d receptor agonists. *Cellular immunology*, 233(2):115–24, 2005.

[5] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics (Oxford, England)*, 22(13):1600–7, 2006.

[6] Henar Alonso, Juan Ignacio Aguilo, Sofía Samper, Jose Antonio Caminero, María Isolina Campos-Herrero, Brigitte Gicquel, Roland Brosch, Carlos Martín, and Isabel Otal. Deciphering the role of is6110 in a highly transmissible mycobacterium tuberculosis beijing strain, gc1237. *Tuberculosis (Edinburgh, Scotland)*, 91(2):117–26, 2011.

[7] Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K Grenier, Weibo Li, Or Zuk, Lisa a Schubert, Brian Birditt, Tal Shay, Alon Goren, Xiaolan Zhang, Zachary Smith, Raquel Deering, Rebecca C McDonald, Moran Cabili, Bradley E Bernstein, John L Rinn, Alex Meissner, David E Root, Nir Hacohen, and Aviv Regev. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science (New York, N.Y.)*, 326(5950):257–63, 2009.

[8] S Amselem, P Duquesnoy, O Attree, G Novelli, S Bousnina, M C Postel-Vinay, and M Goossens. Laron dwarfism and mutations of the growth hormone-receptor gene. *The New England journal of medicine*, 321(15):989–95, 1989.

[9] 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.

[10] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien a Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle, and Wolf Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–32, 2016.

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

[12] Roi Avraham, Nathan Haseley, Douglas Brown, Cristina Penaranda, Humberto B. Jijon, John J. Trombetta, Rahul Satija, Alex K. Shalek, Ramnik J. Xavier, Aviv Regev, and Deborah T. Hung. Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell*, 162(6):1309–21, 2015.

[13] Abul K. Azad, Wolfgang Sadee, and Larry S. Schlesinger. Innate immune gene polymorphisms in tuberculosis. *Infection and immunity*, 80(10):3343–59, 2012.

[14] Bernard M. Babior. Nadph oxidase. *Current opinion in immunology*, 16(1):42–7, 2004.

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

[16] David J Balding. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10):781–91, 2006.

[17] 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.

[18] Clifton E Barry, Helena I Boshoff, Véronique Dartois, Thomas Dick, Sabine Ehrt, JoAnne Flynn, Dirk Schnappinger, Robert J Wilkinson, and Douglas Young. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature reviews. Microbiology*, 7(12):845–55, 2009.

[19] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, Alexander E Urban, Stephen B Montgomery, Douglas F Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, 24(1):14–24, 2014.

[20] M a Behr, M a Wilson, W P Gill, H Salamon, G K Schoolnik, S Rane, and P M Small. Comparative genomics of bcg vaccines by whole-genome dna microarray. *Science (New York, N.Y.)*, 284(5419):1520–3, 1999.

[21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995.

[22] 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.

[23] Mark D Biggin. Animal transcription networks as highly connected, quantitative continua. *Developmental cell*, 21(4):611–26, 2011.

[24] 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.

[25] 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.

[26] Walter Bodmer and Carolina Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, 40(6):695–701, 2008.

[27] 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.

[28] 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.

[29] 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.

[30] Roy J Britten and Eric H Davidson. Gene regulation for higher cells: a theory. *Science (New York, N.Y.)*, 165(3891):349–57, 1969.

[31] Igor E Brodsky and Ruslan Medzhitov. Targeting of immune signalling networks by bacterial pathogens. *Nature cell biology*, 11(5):521–6, 2009.

[32] Judith Bruchfeld, Margarida Correia-Neves, and Gunilla Källenius. Tuberculosis and HIV coinfection. *Cold Spring Harbor perspectives in medicine*, 5(7):a017871, 2015.

[33] 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.

[34] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–90, 2015.

[35] Jacinta Bustamante, Andres a Arias, Guillaume Vogt, Capucine Picard, Lizbeth Blancas Galicia, Carolina Prando, Audrey V Grant, Christophe C Marchal, Marjorie Hubeau, Ariane Chapgier, Ludovic de Beaucoudrey, Anne Puel, Jacqueline Feinberg, Ethan Valinetz, Lucile Jannière, Céline Besse, Anne Boland, Jean-Marie Brisseau, Stéphane Blanche, Olivier Lortholary, Claire Fieschi, Jean-François Emile, Stéphanie Boisson-Dupuis, Saleh Al-Muhsen, Bruce Woda, Peter E Newburger, Antonio Condino-Neto, Mary C Dinauer, Laurent Abel, and Jean-Laurent Casanova. Germline cybb mutations that selectively affect macrophages in kindreds with x-linked predisposition to tuberculous mycobacterial disease. *Nature immunology*, 12(3):213–21, 2011.

[36] Sean B Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008.

[37] Jean-Laurent Casanova and Laurent Abel. Human genetics of infectious diseases: a unified theory. *The EMBO journal*, 26(4):915–22, 2007.

[38] 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.

[39] Stephane E Castel, Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome biology*, 16(1):195, 2015.

[40] 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.

[41] R. Chacón-Salinas, J. Serafín-López, R. Ramos-Payán, P. Méndez-Aragón, R. Hernández-Pando, D. Van Soolingen, L. Flores-Romo, S. Estrada-Parra, and Iris Estrada-García. Differential pattern of cytokine expression by macrophages infected in vitro with different mycobacterium tuberculosis genotypes. *Clinical and experimental immunology*, 140(3):443–9, 2005.

[42] Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

[43] Stephen J Chapman and Adrian V S Hill. Human genetic susceptibility to infectious disease. *Nature reviews. Genetics*, 13(3):175–88, 2012.

[44] 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.

[45] Grace Y Chen and Gabriel Nuñez. Sterile inflammation: sensing and reacting to damage. *Nature reviews. Immunology*, 10(12):826–37, 2010.

[46] 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.

[47] 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.

[48] 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.

[49] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature reviews. Genetics*, 15(1):34–48, 2014.

[50] 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.

[51] Angela C. Collins, Haocheng Cai, Tuo Li, Luis H. Franco, Xiao-Dong Li, Vidhya R. Nair, Caitlyn R. Scharn, Chelsea E. Stamm, Beth Levine, Zhijian J. Chen, and Michael U. Shiloh. Cyclic gmp-amp synthase is an innate immune dna sensor for mycobacterium tuberculosis. *Cell host & microbe*, 17(6):820–8, 2015.

[52] Iñaki Comas, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian Parkhill, Bijaya Malla, Stefan Berg, Guy Thwaites, Dorothy Yeboah-Manu, Graham Bothamley, Jian Mei, Lanhai Wei, Stephen Bentley, Simon R Harris, Stefan Niemann, Roland Diel, Abraham Aseffa, Qian Gao, Douglas Young, and Sebastien Gagneux. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature genetics*, 45(10):1176–82, 2013.

[53] Iñaki Comas and Sebastien Gagneux. The past and future of tuberculosis research. *PLoS pathogens*, 5(10):e1000600, 2009.

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

[55] Heather J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, 10(6):392–404, 2009.

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

[57] Anna K. Coussens, Robert J. Wilkinson, Yasmeen Hanifa, Vladyslav Nikolayevskyy, Paul T. Elkington, Kamrul Islam, Peter M. Timms, Timothy R. Venton, Graham H. Bothamley, Geoffrey E. Packe, Mathina Darmalingam, Robert N. Davidson, Heather J. Milburn, Lucy V. Baker, Richard D. Barker, Charles a. Mein, Leena Bhaw-Rosun, Rosamond Nuamah, Douglas B. Young, Francis a. Drobniewski, Christopher J. Griffiths, and Adrian R. Martineau. Vitamin d accelerates resolution of inflammatory responses during tuberculosis treatment. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38):15449–54, 2012.

[58] Hélio J. Crespo, Joseph T Y Lau, and Paula a. Videira. Dendritic cells: a spot on sialic acid. *Frontiers in immunology*, 4(December):491, 2013.

[59] 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.

[60] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, 348(6237):910–4, 2015.

[61] Eric H Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–20, 2010.

[62] Christine Deffert, Julien Cachat, and Karl-Heinz Krause. Phagocyte nadph oxidase, chronic granulomatous disease and mycobacterial infections. *Cellular microbiology*, 16(8):1168–78, 2014.

[63] Christine Deffert, Michela G Schäppi, Jean-claude Pache, Julien Cachat, Dominique Vesin, Ruth Bisig, Xiaojuan Ma Mulone, Tiina Kelkka, Rikard Holmdahl, Irene Garcia,

Maria L Olleros, and Karl-heinz Krause. Bacillus calmette-guerin infection in nadph oxidase deficiency: defective mycobacterial sequestration and granuloma formation. *PLoS pathogens*, 10(9):e1004325, 2014.

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

[65] Bappaditya Dey, Ruchi Jain Dey, Laurene S Cheung, Supriya Pokkali, Haidan Guo, Jong-Hee Lee, and William R Bishai. A bacterial cyclic dinucleotide activates the cytosolic surveillance pathway and mediates innate resistance to tuberculosis. *Nature medicine*, 21(4):401–6, 2015.

[66] Anna Di Rienzo. Population genetics models of common diseases. *Current opinion in genetics & development*, 16(6):630–6, 2006.

[67] Lautaro Diacovich and Jean-Pierre Gorvel. Bacterial manipulation of innate immunity to promote infection. *Nature reviews. Microbiology*, 8(2):117–28, 2010.

[68] 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.

[69] 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.

[70] J. Dormans, M. Burger, D. Aguilar, R. Hernandez-Pando, K. Kremer, P. Roholl, S. M. Arend, and D van Soolingen. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different mycobacterium tuberculosis genotypes in a balb/c mouse model. *Clinical and experimental immunology*, 137(3):460–8, 2004.

[71] 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.

[72] 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.

[73] Shiwei Duan, RS Stephanie Huang, Wei Zhang, Wasim K Bleibel, Cheryl A Roe, Tyson A Clark, Tina X Chen, Anthony C Schweitzer, John E Blume, Nancy J Cox, and M Eileen Dolan. Genetic architecture of transcript-level variation in humans. *American journal of human genetics*, 82(5):1101–13, 2008.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G Gislason, Tryggvi Stefansson, Bjorn G Leifsson, Unnur Thorsteinsdottir, John R Lamb, Jeffrey R Gulcher, Marc L Reitman, Augustine Kong, Eric E Schadt, and Kari Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–8, 2008.

[80] ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science (New York, N.Y.)*, 306(5696):636–40, 2004.

[81] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[82] ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G

Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ula Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, Alexander E Urban, Peter Kraus, Sara Van Calcar, Nate Heintzman, Tae Hoon Kim, Kun Wang, Chunxu Qu, Gary Hon, Rosa Luna, Christopher K Glass, M Geoff Rosenfeld, Shelley Force Aldred, Sara J

Cooper, Anason Halees, Jane M Lin, Hennady P Shulha, Xiaoling Zhang, Mousheng Xu, Jaafar N S Haidar, Yong Yu, Yijun Ruan, Vishwanath R Iyer, Roland D Green, Claes Wadelius, Peggy J Farnham, Bing Ren, Rachel A Harte, Angie S Hinrichs, Heather Trumbower, Hiram Clawson, Jennifer Hillman-Jackson, Ann S Zweig, Kayla Smith, Archana Thakkapallayil, Galt Barber, Robert M Kuhn, Donna Karolchik, Lluis Armengol, Christine P Bird, Paul I W de Bakker, Andrew D Kern, Nuria Lopez-Bigas, Joel D Martin, Barbara E Stranger, Abigail Woodroffe, Eugene Davydov, Antigone Dimas, Eduardo Eyras, Ingileif B Hallgrímsdóttir, Julian Huppert, Michael C Zody, Gonçalo R Abecasis, Xavier Estivill, Gerard G Bouffard, Xiaobin Guan, Nancy F Hansen, Jacquelyn R Idol, Valerie V B Maduro, Baishali Maskeri, Jennifer C McDowell, Morgan Park, Pamela J Thomas, Alice C Young, Robert W Blakesley, Donna M Muzny, Erica Sodergren, David A Wheeler, Kim C Worley, Huaiyang Jiang, George M Weinstock, Richard A Gibbs, Tina Graves, Robert Fulton, Elaine R Mardis, Richard K Wilson, Michele Clamp, James Cuff, Sante Gnerre, David B Jaffe, Jean L Chang, Kerstin Lindblad-Toh, Eric S Lander, Maxim Koriabine, Mikhail Nefedov, Kazutoyo Osoegawa, Yuko Yoshinaga, Baoli Zhu, and Pieter J de Jong. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[83] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9, 2011.

[84] Joel D. Ernst. The immunological life cycle of tuberculosis. *Nature reviews. Immunology*, 12(8):581–91, 2012.

[85] R Eshet, Z Laron, A Pertzelan, R Arnon, and M Dintzman. Defect of human growth hormone receptors in the liver of two patients with laron-type dwarfism. *Israel journal of medical sciences*, 20(1):8–11, 1984.

[86] 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.

[87] Kyle Kai-how Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J H Ryan, Alexander a Shishkin, Meital Hatan, Marlene J Carrasco-Alfonso, Dita Mayer, C John Luckey, Nikolaos a Patsopoulos, Philip L De Jager, Vijay K Kuchroo, Charles B Epstein, Mark J Daly, David A Hafler, and Bradley E Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–43, 2015.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] Lude Franke and Ritsert C Jansen. eqtl analysis in humans. *Methods in molecular biology (Clifton, N.J.)*, 573:311–28, 2009.

[93] 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.

[94] 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.

[95] Matteo Fumagalli and Manuela Sironi. Human genome variability, natural selection and infectious diseases. *Current opinion in immunology*, 30:9–16, 2014.

[96] Matteo Fumagalli, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admetlla, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS genetics*, 7(11):e1002355, 2011.

[97] Sebastien Gagneux, Kathryn DeRiemer, Tran Van, Midori Kato-Maeda, Bouke C de Jong, Sujatha Narayanan, Mark Nicol, Stefan Niemann, Kristin Kremer, M Cristina Gutierrez, Markus Hilty, Philip C Hopewell, and Peter M Small. Variable host-pathogen compatibility in mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2869–73, 2006.

[98] Alex S Genshaft, Shuqiang Li, Caroline J. Gallant, Spyros Darmanis, Sanjay M. Prakadan, Carly G. K. Ziegler, Martin Lundberg, Simon Fredriksson, Joyce Hong, Aviv Regev, Kenneth J. Livak, Ulf Landegren, and Alex K. Shalek. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome biology*, 17(1):188, 2016.

[99] E Giacomini, E Iona, L Ferroni, M Miettinen, L Fattorini, G Orefici, I Julkunen, and E M Coccia. Infection of human macrophages and dendritic cells with mycobacterium tuberculosis induces a differential cytokine gene expression that modulates t cell response. *Journal of immunology (Baltimore, Md. : 1950)*, 166(12):7033–41, 2001.

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

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

[102] P J Godowski, D W Leung, L R Meacham, J P Galgani, R Hellmiss, R Keret, P S Rotwein, J S Parks, Z Laron, and W I Wood. Characterization of the human growth hormone receptor gene and demonstration of a partial gene deletion in two patients with laron-type dwarfism. *Proceedings of the National Academy of Sciences of the United States of America*, 86(20):8083–7, 1989.

[103] 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.

[104] 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.

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

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

[107] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–5, 2013.

[108] GTEx Consortium. Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348(6235):648–60, 2015.

[109] Conny a. Gysemans, Alessandra K. Cardozo, Hanne Callewaert, Annapaula Giulietti, Leen Hulshagen, Roger Bouillon, Décio L. Eizirik, and Chantal Mathieu. 1,25-dihydroxyvitamin d3 modulates expression of chemokines and cytokines in pancreatic islets: implications for prevention of diabetes in nonobese diabetic mice. *Endocrinology*, 146(4):1956–64, 2005.

[110] Martha T Hamblin, Emma E Thompson, and Anna Di Rienzo. Complex signatures of natural selection at the Duffy blood group locus. *American journal of human genetics*, 70(2):369–83, 2002.

[111] 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.

[112] Steven Henikoff and Ali Shilatifard. Histone modification: cause or cog? *Trends in genetics : TIG*, 27(10):389–96, 2011.

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

[114] Erik C Hett and Eric J Rubin. Bacterial growth and cell division: a mycobacterial perspective. *Microbiology and molecular biology reviews : MMBR*, 72(1):126–56, table of contents, 2008.

[115] Martin Hewison. Antibacterial effects of vitamin d. *Nature reviews. Endocrinology*, 7(6):337–45, 2011.

[116] 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.

[117] Lucia a Hindorff, Praveen Sethupathy, Heather a Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri a Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, 2009.

[118] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, 6(2):95–108, 2005.

[119] Joshua W. K. Ho, Youngsook L. Jung, Tao Liu, Burak H. Alver, Soohyun Lee, Kohta Ikegami, Kyung-Ah Sohn, Aki Minoda, Michael Y. Tolstorukov, Alex Appert, Stephen C. J. Parker, Tingting Gu, Anshul Kundaje, Nicole C. Riddle, Eric Bishop, Thea A. Egelhofer, Sheng'en Shawn Hu, Artyom A. Alekseyenko, Andreas Rechtsteiner, Dalal Asker, Jason A. Belsky, Sarah K. Bowman, Q Brent Chen, Ron A.-J. Chen, Daniel S. Day, Yan Dong, Andrea C. Dose, Xikun Duan, Charles B. Epstein, Sevinc Ercan, Elise A. Feingold, Francesco Ferrari, Jacob M. Garrigues, Nils Gehlenborg, Peter J. Good, Psalm Haseley, Daniel He, Moritz Herrmann, Michael M. Hoffman, Tess E. Jeffers, Peter V. Kharchenko, Paulina Kolasinska-Zwierz, Chitra V. Kotwaliwale, Nischay Kumar, Sasha A. Langley, Erica N. Larschan, Isabel Latorre, Maxwell W. Libbrecht, Xueqiu Lin, Richard Park, Michael J. Pazin, Hoang N. Pham, Annette Plachetka, Bo Qin, Yuri B. Schwartz, Noam Shoresh, Przemyslaw Stempor, Anne Vielle, Chengyang Wang, Christina M. Whittle, Huiling Xue, Robert E. Kingston, Ju Han Kim, Bradley E. Bernstein, Abby F. Dernburg, Vincenzo Pirrotta, Mitzi I. Kuroda,

William S. Noble, Thomas D. Tullius, Manolis Kellis, David M. MacAlpine, Susan Strome, Sarah C. R. Elgin, Xiaole Shirley Liu, Jason D. Lieb, Julie Ahringer, Gary H. Karpen, and Peter J. Park. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515):449–52, 2014.

[120] Susanne Homolka, Stefan Niemann, David G. Russell, and Kyle H. Rohde. Functional genetic diversity among mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS pathogens*, 6(7):e1000988, 2010.

[121] Philip a. Hopkins and S. Sriskandan. Mammalian toll-like receptors: to immunity and beyond. *Clinical and experimental immunology*, 140(3):395–407, 2005.

[122] 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.

[123] Tsungda Hsu, Suzanne M Hingley-Wilson, Bing Chen, Mei Chen, Annie Z Dai, Paul M Morin, Carolyn B Marks, Jeevan Padiyar, Celia Goulding, Mari Gingery, David Eisenberg, Robert G Russell, Steven C Derrick, Frank M Collins, Sheldon L Morris, C Harold King, and William R Jacobs. The primary mechanism of attenuation of bacillus calmette-guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12420–5, 2003.

[124] 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.

[125] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert R H Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, Crystal B Warner, Kerstin Blankenburg, Yi Han, Mehwish Javaid, Joy Jayaseelan, Shalini N Jhangiani, Donna Muzny, Fiona Ongeri, Lora Perales, Yuan-Qing Wu, Yiqing Zhang, Xiaoyan Zou, Eric a Stone, Richard a Gibbs, and Trudy F C Mackay. Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):15553–9, 2012.

[126] 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.

[127] International HapMap 3 Consortium, David M Altshuler, Richard a Gibbs, Leena Peltonen, David M Altshuler, Richard a Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Leena Peltonen, Emmanouil Dermitzakis, Penelope E Bonnen, David M Altshuler, Richard a Gibbs, Paul I W de Bakker, Panos Deloukas, Stacey B Gabriel, Rhian Gwilliam, Sarah Hunt, Michael Inouye, Xiaoming Jia, Aarno Palotie, Melissa Parkin, Pamela Whittaker, Fuli Yu, Kyle Chang, Alicia Hawes, Lora R Lewis, Yanru Ren, David Wheeler, Richard a Gibbs, Donna Marie Muzny, Chris Barnes, Katayoon Darvishi, Matthew Hurles, Joshua M Korn, Kati Kristiansson, Charles Lee, Steven a McCarrol, James Nemesh, Emmanouil Dermitzakis, Alon Keinan, Stephen B Montgomery, Samuela Pollack, Alkes L Price, Nicole Soranzo, Penelope E Bonnen, Richard a Gibbs, Claudia Gonzaga-Jauregui, Alon Keinan, Alkes L Price, Fuli Yu, Verneri Anttila, Wendy Brodeur, Mark J Daly, Stephen Leslie, Gil McVean, Loukas Moutsianas, Huy Nguyen, Stephen F Schaffner, Qingrun Zhang, Mohammed J R Ghori, Ralph McGinnis, William McLaren, Samuela Pollack, Alkes L Price, Stephen F Schaffner, Fumihiko Takeuchi, Sharon R Grossman, Ilya Shlyakhter, Elizabeth B Hostetter, Pardis C Sabeti, Clement a Adebamowo, Morris W Foster, Deborah R Gordon, Julio Licinio, Maria Cristina Manca, Patricia a Marshall, Ichiro Matsuda, Duncan Ngare, Vivian Ota Wang, Deepa Reddy, Charles N Rotimi, Charmaine D Royal, Richard R Sharp, Changqing Zeng, Lisa D Brooks, and Jean E McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010.

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

[129] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard a Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi,

Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur a Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement a Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard a Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 2007.

[130] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[131] Thomas R. Ioerger, Yicheng Feng, Krishna Ganesula, Xiaohua Chen, Karen M. Dobos, Sarah Fortune, William R. Jacobs, Valerie Mizrahi, Tanya Parish, Eric Rubin, Chris Sassetti, and James C. Sacchettini. Variation among genome sequences of h37rv strains of mycobacterium tuberculosis from multiple laboratories. *Journal of bacteriology*, 192(14):3645–53, 2010.

[132] 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.

[133] 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.

[134] 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.

[135] Akiko Iwasaki and Ruslan Medzhitov. Toll-like receptor control of the adaptive immune responses. *Nature immunology*, 5(10):987–95, 2004.

[136] M. Jacobsen, D. Repsilber, K. Kleinsteuber, A. Gutschmidt, S. Schommer-Leitner, G. Black, G. Walzl, and S. H E Kaufmann. Suppressor of cytokine signaling-3 is affected in t-cells from tuberculosistb patients. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 17(9):1323–31, 2011.

[137] 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.

[138] Charles a Janeway and Ruslan Medzhitov. Innate immune recognition. *Annual review of immunology*, 20(2):197–216, 2002.

[139] Shilpi Jayaswal, Md Azhar Kamal, Raina Dua, Shashank Gupta, Tanmay Majumdar, Gobardhan Das, Dhiraj Kumar, and Kanury V S Rao. Identification of host-dependent survival factors for intracellular mycobacterium tuberculosis through an sirna screen. *PLoS pathogens*, 6(4):e1000839, 2010.

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

[141] T Jenuwein and C D Allis. Translating the histone code. *Science (New York, N.Y.)*, 293(5532):1074–80, 2001.

[142] 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.

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

[144] Jenny L. Johnson, Mark B. Jones, Sean O. Ryan, and Brian a. Cobb. The regulatory power of glycans and their binding partners in immunity. *Trends in immunology*, 34(6):290–8, 2013.

[145] 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.

[146] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N Ananthakrishnan, Vibeke Andersen, Jane M Andrews, Leonard Baidoo, Tobias Balschun, Peter A Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C Lawrance, Charlie W Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y Ponsioen, Uros Potocnik, Natalie J Prescott, Miguel Regueiro, Jerome I Rotter, Richard K Russell, Jeremy D Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A Simms, Jurgita Sventoraityte, Stephan R Targan, Kent D Taylor, Mark Tremelling, Hein W Verspaget, Martine De Vos, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Sebastian Zeissig, Bin Zhang, Clarence K Zhang, Hongyu Zhao, International IBD Genetics Consortium (IIBDGC), Mark S Silverberg, Vito Annese, Hakon Hakonarson, Steven R Brant, Graham Radford-Smith, Christopher G Mathew, John D Rioux, Eric E Schadt, Mark J Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C Barrett, and Judy H Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–24, 2012.

[147] 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.

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

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

[150] 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.

[151] 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.

[152] 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 bacte-

rial lipopolysaccharide in primary human monocytes. *G3 (Bethesda, Md.)*, 6(5):1345–55, 2016.

[153] Taro Kawai and Shizuo Akira. The roles of tlrs, rlrs and nlrs in pathogen recognition. *International immunology*, 21(4):317–37, 2009.

[154] Malcolm D Kearns, Jessica A Alvarez, Natan Seidel, and Vin Tangpricha. Impact of vitamin d on infectious disease. *The American journal of the medical sciences*, 349(3):245–62, 2015.

[155] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, Ian Dunham, Laura L Elnitski, Peggy J Farnham, Elise A Feingold, Mark Gerstein, Morgan C Giddings, David M Gilbert, Thomas R Gingeras, Eric D Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D Lieb, Richard M Myers, Michael J Pazin, Bing Ren, John A Stamatoyannopoulos, Zhiping Weng, Kevin P White, and Ross C Hardison. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6131–8, 2014.

[156] 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.

[157] Bae-Hoon Kim, Avinash R Shenoy, Pradeep Kumar, Rituparna Das, Sangeeta Tiwari, and John D MacMicking. A family of ifn-γ-inducible 65-kd gtpases protects against bacterial infection. *Science (New York, N.Y.)*, 332(6030):717–21, 2011.

[158] 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.

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

[160] 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.

[161] 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.

[162] 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.

[163] Mi-Sun Koo, Selvakumar Subbian, and Gilla Kaplan. Strain specific transcriptional response in mycobacterium tuberculosis infected macrophages. *Cell communication and signaling : CCS*, 10(1):2, 2012.

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

[165] Nitya Krishnan, Wladimir Malaga, Patricia Constant, Maxine Caws, Thi Hoang Chau Tran, Jenifer Salmons, Thi Ngoc Lan Nguyen, Duc Bang Nguyen, Mamadou Daffé, Douglas B Young, Brian D Robertson, Christophe Guilhot, and Guy E Thwaites. Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. *PloS one*, 6(9):e23870, 2011.

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

[167] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R

Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[168] Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97, 2011.

[169] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li, Tao Liu, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–31, 2012.

[170] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R Wood, Robert J Weyant, Ayellet V Segrè, Elizabeth K Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L Heard-Costa, Joshua C Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi, Tomi Pastinen, Liming Liang, Iris M Heid, Jian'an Luan, Gudmar Thorleifsson, Thomas W Winkler, Michael E Goddard, Ken Sin Lo, Cameron Palmer, Tsegaselassie Workalemahu, Yurii S Aulchenko, Asa Johansson, M Carola Zillikens, Mary F Feitosa, Tõnu Esko, Toby Johnson, Shamika Ketkar, Peter Kraft, Massimo Mangino, Inga Prokopenko, Devin Absher, Eva Albrecht, Florian Ernst, Nicole L Glazer, Caroline Hayward, Jouke-Jan Hottenga, Kevin B Jacobs, Joshua W Knowles, Zoltán Kutalik, Keri L Monda, Ozren Polasek, Michael Preuss, Nigel W Rayner, Neil R Robertson, Valgerdur Steinthorsdottir, Jonathan P Tyrer, Benjamin F Voight, Fredrik Wiklund, Jianfeng Xu, Jing Hua Zhao, Dale R Nyholt, Niina Pellikka, Markus Perola, John R B Perry, Ida Surakka, Mari-Liis Tammesoo, Elizabeth L Altmaier, Najaf Amin, Thor Aspelund, Tushar Bhangale, Gabrielle Boucher, Daniel I Chasman, Constance Chen, Lachlan Coin, Matthew N Cooper, Anna L Dixon, Quince Gibson, Elin

Grundberg, Ke Hao, M Juhani Junttila, Lee M Kaplan, Johannes Kettunen, Inke R König, Tony Kwan, Robert W Lawrence, Douglas F Levinson, Mattias Lorentzon, Barbara McKnight, Andrew P Morris, Martina Müller, Julius Suh Ngwa, Shaun Purcell, Suzanne Rafelt, Rany M Salem, Erika Salvi, Serena Sanna, Jianxin Shi, Ulla Sovio, John R Thompson, Michael C Turchin, Liesbeth Vandenput, Dominique J Verlaan, Veronique Vitart, Charles C White, Andreas Ziegler, Peter Almgren, Anthony J Balmforth, Harry Campbell, Lorena Citterio, Alessandro De Grandi, Anna Dominiczak, Jubao Duan, Paul Elliott, Roberto Elosua, Johan G Eriksson, Nelson B Freimer, Eco J C Geus, Nicola Glorioso, Shen Haiqing, Anna-Liisa Hartikainen, Aki S Havulinna, Andrew a Hicks, Jennie Hui, Wilmar Igl, Thomas Illig, Antti Jula, Eero Kajantie, Tuomas O Kilpeläinen, Markku Koiranen, Ivana Kolcic, Seppo Koskinen, Peter Kovacs, Jaana Laitinen, Jianjun Liu, Marja-Liisa Lokki, Ana Marusic, Andrea Maschio, Thomas Meitinger, Antonella Mulas, Guillaume Paré, Alex N Parker, John F Peden, Astrid Petersmann, Irene Pichler, Kirsi H Pietiläinen, Anneli Pouta, Martin Ridderstråle, Jerome I Rotter, Jennifer G Sambrook, Alan R Sanders, Carsten Oliver Schmidt, Juha Sinisalo, Jan H Smit, Heather M Stringham, G Bragi Walters, Elisabeth Widen, Sarah H Wild, Gonneke Willemsen, Laura Zagato, Lina Zgaga, Paavo Zitting, Helene Alavere, Martin Farrall, Wendy L McArdle, Mari Nelis, Marjolein J Peters, Samuli Ripatti, Joyce B J van Meurs, Katja K Aben, Kristin G Ardlie, Jacques S Beckmann, John P Beilby, Richard N Bergman, Sven Bergmann, Francis S Collins, Daniele Cusi, Martin den Heijer, Gudny Eiriksdottir, Pablo V Gejman, Alistair S Hall, Anders Hamsten, Heikki V Huikuri, Carlos Iribarren, Mika Kähönen, Jaakko Kaprio, Sekar Kathiresan, Lambertus Kiemeney, Thomas Kocher, Lenore J Launer, Terho Lehtimäki, Olle Melander, Tom H Mosley, Arthur W Musk, Markku S Nieminen, Christopher J O'Donnell, Claes Ohlsson, Ben Oostra, Lyle J Palmer, Olli Raitakari, Paul M Ridker, John D Rioux, Aila Rissanen, Carlo Rivolta, Heribert Schunkert, Alan R Shuldiner, David S Siscovick, Michael Stumvoll, Anke Tönjes, Jaakko Tuomilehto, Gert-Jan van Ommen, Jorma Viikari, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael a Province, Manfred Kayser, Alice M Arnold, Larry D Atwood, Eric Boerwinkle, Stephen J Chanock, Panos Deloukas, Christian Gieger, Henrik Grönberg, Per Hall, Andrew T Hattersley, Christian Hengstenberg, Wolfgang Hoffman, G Mark Lathrop, Veikko Salomaa, Stefan Schreiber, Manuela Uda, Dawn Waterworth, Alan F Wright, Themistocles L Assimes, Inês Barroso, Albert Hofman, Karen L Mohlke, Dorret I Boomsma, Mark J Caulfield, L Adrienne Cupples, Jeanette Erdmann, Caroline S Fox, Vilmundur Gudnason, Ulf Gyllensten, Tamara B Harris, Richard B Hayes, Marjo-Riitta Jarvelin, Vincent Mooser, Patricia B Munroe, Willem H Ouwehand, Brenda W Penninx, Peter P Pramstaller, Thomas Quertermous, Igor Rudan, Nilesh J Samani, Timothy D Spector, Henry Völzke, Hugh Watkins, James F Wilson, Leif C Groop, Talin Haritunians, Frank B Hu, Robert C Kaplan, Andres Metspalu, Kari E North, David Schlessinger, Nicholas J Wareham, David J Hunter, Jeffrey R O'Connell, David P Strachan, H-Erich Wichmann, Ingrid B Borecki, Cornelia M van Duijn, Eric E Schadt, Unnur Thorsteinsdottir, Leena Peltonen, André G Uitterlinden, Peter M Visscher, Nilanjan Chatterjee, Ruth J F Loos, Michael Boehnke, Mark I

McCarthy, Erik Ingelsson, Cecilia M Lindgren, Gonçalo R Abecasis, Kari Stefansson, Timothy M Frayling, and Joel N Hirschhorn. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–8, 2010.

[171] Tuuli Lappalainen. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome research*, 25(10):1427–31, 2015.

[172] Tuuli Lappalainen and Emmanouil T Dermitzakis. Evolutionary history of regulatory variation in human populations. *Human molecular genetics*, 19(R2):R197–203, 2010.

[173] Tuuli Lappalainen, Stephen B Montgomery, Alexandra C Nica, and Emmanouil T Dermitzakis. Epistatic selection between coding and regulatory variation in human evolution and disease. *American journal of human genetics*, 89(3):459–63, 2011.

[174] Zvi Laron. *History of the Israeli Cohort of Laron Syndrome Patients (1958–2009)*, pages 3–7. 2011.

[175] 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.

[176] 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.

[177] Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics*, 88(3):294–305, 2011.

[178] 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.

[179] Guillaume Lettre. Recent progress in the study of the genetics of height. *Human genetics*, 129(5):465–72, 2011.

[180] 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.

[181] Qing Li, Christopher C Whalen, Jeffrey M Albert, Rhonda Larkin, Lynn Zukowski, M Donald Cave, and Richard F Silver. Differences in rate and variability of intracellular growth of a panel of mycobacterium tuberculosis clinical isolates within a human monocyte model. *Infection and immunity*, 70(11):6489–93, 2002.

[182] 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.

[183] 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.

[184] 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.

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

[186] 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.

[187] Philip T Liu, Steffen Stenger, Huiying Li, Linda Wenzel, Belinda H Tan, Stephan R Krutzik, Maria Teresa Ochoa, Jürgen Schauber, Kent Wu, Christoph Meinken, Diane L Kamen, Manfred Wagner, Robert Bals, Andreas Steinmeyer, Ulrich Zügel, Richard L Gallo, David Eisenberg, Martin Hewison, Bruce W Hollis, John S Adams, Barry R Bloom, and Robert L Modlin. Toll-like receptor triggering of a vitamin d-mediated human antimicrobial response. *Science (New York, N.Y.)*, 311(5768):1770–3, 2006.

[188] 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.

[189] Robert Loddenkemper, Marc Lipman, and Alimuddin Zumla. Clinical aspects of adult tuberculosis. *Cold Spring Harbor perspectives in medicine*, 6(1):a017848, 2016.

[190] Robyn M Lucas, Shelley Gorman, Sian Geldenhuys, and Prue H Hart. Vitamin d and immunity. *F1000prime reports*, 6(3):118, 2014.

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

[192] 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.

[193] 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.

[194] 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.

[195] Christina J Maier, Richard H Maier, Raphaela Rid, Andrea Trost, Harald Hundsberger, Andreas Eger, Helmut Hintner, Johann W Bauer, and Kamil Onder. Pim-1 kinase interacts with the dna binding domain of the vitamin d receptor: a further kinase implicated in 1,25-(oh)2d3 signaling. *BMC molecular biology*, 13(1):18, 2012.

[196] C Manca, L Tsenova, A Bergtold, S Freeman, M Tovey, J M Musser, C E Barry, V H Freedman, and G Kaplan. Virulence of a mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce th1 type immunity and is associated with induction of ifn-alpha /beta. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10):5752–7, 2001.

[197] Claudia Manca, Michael B Reed, Sherry Freeman, Barun Mathema, Barry Kreiswirth, Clifton E Barry, and Gilla Kaplan. Differential monocyte activation underlies strain-specific mycobacterium tuberculosis pathogenesis. *Infection and immunity*, 72(9):5511–4, 2004.

[198] Valentina D Mangano and David Modiano. An evolutionary perspective of how infection drives human genome diversity: the case of malaria. *Current opinion in immunology*, 30(Box 1):39–47, 2014.

[199] Teri a Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia a Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven a McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.

[200] Jérémy Manry and Lluis Quintana-Murci. A genome-wide perspective of human diversity and its implications in infectious disease. *Cold Spring Harbor perspectives in medicine*, 3(1):a012450, 2013.

[201] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–17, 2008.

[202] Adrian R. Martineau, Robert J Wilkinson, Katalin a. Wilkinson, Sandra M. Newton, Beate Kampmann, Bridget M. Hall, Geoffrey E. Packe, Robert N. Davidson, Sandra M.

Eldridge, Zoë J. Maunsell, Sandra J. Rainbow, Jacqueline L. Berry, and Christopher J. Griffiths. A single dose of vitamin d enhances immunity to mycobacteria. *American journal of respiratory and critical care medicine*, 176(2):208–13, 2007.

[203] 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.

[204] Leah E. Mechanic, Huann-Sheng Chen, Christopher I. Amos, Nilanjan Chatterjee, Nancy J. Cox, Rao L. Divi, Ruzong Fan, Emily L. Harris, Kevin Jacobs, Peter Kraft, Suzanne M. Leal, Kimberly McAllister, Jason H. Moore, Dina N. Paltoo, Michael A. Province, Erin M. Ramos, Marylyn D. Ritchie, Kathryn Roeder, Daniel J. Schaid, Matthew Stephens, Duncan C. Thomas, Clarice R. Weinberg, John S. Witte, Shunpu Zhang, Sebastian Zöllner, Eric J. Feuer, and Elizabeth M. Gillanders. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genetic epidemiology*, 36(1):22–35, 2012.

[205] 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.

[206] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-kyung Kim, Richard P Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S Lander, and Bradley E Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, 2007.

[207] 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.

[208] 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.

[209] Trine H Mogensen. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clinical microbiology reviews*, 22(2):240–73, Table of Contents, 2009.

[210] 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.

[211] S A Monks, A Leonardson, H Zhu, P Cundiff, P Pietrusiak, S Edwards, J W Phillips, A Sachs, and E E Schadt. Genetic inheritance of gene expression in human cell lines. *American journal of human genetics*, 75(6):1094–105, 2004.

[212] Alessandra Mortellaro, Lucy Robinson, and Paola Ricciardi-Castagnoli. Spotlight on mycobacteria and dendritic cells: will novel targets to fight tuberculosis emerge? *EMBO molecular medicine*, 1(1):19–29, 2009.

[213] 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.

[214] Jesse T Myers, Albert W Tsang, and Joel a Swanson. Localized reactive oxygen and nitrogen intermediates inhibit escape of listeria monocytogenes from vacuoles in activated macrophages. *Journal of immunology (Baltimore, Md. : 1950)*, 171(10):5447–53, 2003.

[215] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[216] Edward A Nardell. Transmission and institutional infection control of tuberculosis. *Cold Spring Harbor perspectives in medicine*, 6(2):a018192, 2016.

[217] Gioacchino Natoli. Maintaining cell identity through global control of genomic organization. *Immunity*, 33(1):12–24, 2010.

[218] 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.

[219] 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.

[220] Alexandra C Nica, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS genetics*, 6(4):e1000895, 2010.

[221] Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, Asa K Hedman, Veronique Bataille, Jordana Tzenova Bell, Gabriela Surdulescu, Antigone S Dimas, Catherine Ingle, Frank O Nestle, Paola di Meglio, Josine L Min, Alicja Wilk, Christopher J Hammond, Neelam Hassanali, Tsun-Po Yang, Stephen B Montgomery, Steve O'Rahilly, Cecilia M Lindgren, Krina T Zondervan, Nicole Soranzo, Inês Barroso,

Richard Durbin, Kourosh Ahmadi, Panos Deloukas, Mark I McCarthy, Emmanouil T Dermitzakis, Timothy D Spector, and MuTHER Consortium. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS genetics*, 7(2):e1002003, 2011.

[222] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888, 2010.

[223] Kelechi E. Nnoaham and Aileen Clarke. Low serum vitamin d levels and tuberculosis: a systematic review and meta-analysis. *International journal of epidemiology*, 37(1):113–9, 2008.

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

[225] John Novembre and Eunjung Han. Human population structure and the adaptive response to pathogen-induced selection pressures. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1590):878–86, 2012.

[226] Carole Ober and Donata Vercelli. Gene-environment interactions in human disease: nuisance or opportunity? *Trends in genetics : TIG*, 27(3):107–15, 2011.

[227] 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.

[228] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos Panousis, and Alexandra C Nica. Estimating the causal tissues for complex traits and diseases. *bioRxiv*, 2016.

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

[230] Renato Ostuni, Viviana Piccolo, Iros Barozzi, Sara Polletti, Alberto Termanini, Silvia Bonifacio, Alessia Curina, Elena Prosperini, Serena Ghisletti, and Gioacchino Natoli. Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1-2):157–71, 2013.

[231] 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.

[232] Alain Pacis, Ludovic Tailleux, Alexander M Morin, John Lambourne, Julia L MacIsaac, Vania Yotova, Anne Dumaine, Anne Danckaert, Francesca Luca, Jean-christophe Grenier, Kasper D Hansen, Brigitte Gicquel, Miao Yu, Athma Pai, Chuan He, Jenny Tung, Tomi Pastinen, Michael S Kobor, Roger Pique-Regi, Yoav Gilad, and Luis B Barreiro.

Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome research*, 25(12):1801–11, 2015.

[233] Athma A Pai, Jonathan K Pritchard, and Yoav Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857, 2015.

[234] Peter J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, 2009.

[235] Pavlos Pavlidis, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular biology and evolution*, 29(10):3237–48, 2012.

[236] 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.

[237] 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.

[238] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics*, 94(4):559–73, 2014.

[239] 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.

[240] 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.

[241] Jonathan K Pritchard and Anna Di Rienzo. Adaptation - not by sweeps alone. *Nature reviews. Genetics*, 11(10):665–7, 2010.

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

[243] Franck Prugnolle, Andrea Manica, Marie Charpentier, Jean François Guégan, Vanina Guernier, and François Balloux. Pathogen-driven selection and worldwide hla class i diversity. *Current biology : CB*, 15(11):1022–7, 2005.

[244] Alexander S. Pym, Priscille Brodin, Roland Brosch, Michel Huerre, and Stewart T. Cole. Loss of rd1 contributed to the attenuation of the live tuberculosis vaccines mycobacterium bovis bcg and mycobacterium microti. *Molecular microbiology*, 46(3):709–17, 2002.

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

[246] 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.

[247] Towfique Raj, Manik Kuchroo, Joseph M. Replogle, Soumya Raychaudhuri, Barbara E. Stranger, and Philip L. De Jager. Common risk alleles for inflammatory diseases are targets of recent positive selection. *American journal of human genetics*, 92(4):517–29, 2013.

[248] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome biology*, 14(9):R95, 2013.

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

[250] Michael B Reed, Pilar Domenech, Claudia Manca, Hua Su, Amy K Barczak, Barry N Kreiswirth, Gilla Kaplan, and Clifton E Barry. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature*, 431(7004):84–7, 2004.

[251] 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.

[252] 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.

[253] Octavio M Rivero-Lezcano. In vitro infection of human cells with mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 93(2):123–9, 2013.

[254] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom,

Matthew L. Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T. Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[255] Esteban a. Roberts, Jennifer Chua, George B. Kyei, and Vojo Deretic. Higher order rab programming in phagolysosome biogenesis. *The Journal of cell biology*, 174(7):923–9, 2006.

[256] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[257] Luz María Rocha-Ramírez, Iris Estrada-García, Luz María López-Marín, Erika Segura-Salinas, Patricia Méndez-Aragón, Dick Van Soolingen, Rubén Torres-González, Rommel Chacón-Salinas, Sergio Estrada-Parra, Carmen Maldonado-Bernal, Constantino López-Macías, and Armando Isibasi. Mycobacterium tuberculosis lipids regulate cytokines, tlr-2/4 and mhc class ii expression in human macrophages. *Tuberculosis (Edinburgh, Scotland)*, 88(3):212–20, 2008.

[258] Graham Rose, Teresa Cortes, Iñaki Comas, Mireia Coscolla, Sebastien Gagneux, and Douglas B. Young. Mapping of genotype-phenotype diversity among clinical isolates of mycobacterium tuberculosis by sequence-based transcriptional profiling. *Genome biology and evolution*, 5(10):1849–62, 2013.

[259] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph a Sperling, Alon Goren, David a Weitz, and Bradley E Bernstein. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33(11):1165–72, 2015.

[260] 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.

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

[262] Nicholas J Schork, Sarah S Murray, Kelly a Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–9, 2009.

[263] Sabino Scolletta, Marta Colletti, Luigi Di Luigi, and Clara Crescioli. Vitamin d receptor agonists target cxcl10: new therapeutic tools for resolution of inflammation. *Mediators of inflammation*, 2013:876319, 2013.

[264] 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.

[265] 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.

[266] 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.

[267] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–9, 2014.

[268] Orit Shevah and Zvi Laron. *Genetic Aspects*, pages 29–52. 2011.

[269] 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.

[270] 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.

[271] Roxane Simeone, Daria Bottai, and Roland Brosch. Esx/type vii secretion systems and their role in host-pathogen interaction. *Current opinion in microbiology*, 12(1):4–10, 2009.

[272] Roxane Simeone, Fadel Sayes, Okryul Song, Matthias I. Gröschel, Priscille Brodin, Roland Brosch, and Laleh Majlessi. Cytosolic access of mycobacterium tuberculosis: critical impact of phagosomal acidification control and demonstration of occurrence in vivo. *PLoS pathogens*, 11(2):e1004650, 2015.

[273] Daniel Sinsimer, Gaelle Huet, Claudia Manca, Liana Tsenova, Mi-Sun Koo, Natalia Kurepina, Bavesh Kana, Barun Mathema, Salvatore a E Marras, Barry N. Kreiswirth, Christophe Guilhot, and Gilla Kaplan. The phenolic glycolipid of mycobacterium tuberculosis differentially modulates the early host cytokine response but does not in itself confer hypervirulence. *Infection and immunity*, 76(7):3027–36, 2008.

[274] Sébastien a Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–20, 2014.

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

[276] 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.

[277] 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.

[278] 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.

[279] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91, 2013.

[280] 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.

[281] 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.

[282] Sarah A Stanley and Jeffery S Cox. Host-pathogen interactions during mycobacterium tuberculosis infections. *Current topics in microbiology and immunology*, 374(July):211–41, 2013.

[283] Sarah a Stanley, James E Johndrow, Paolo Manzanillo, and Jeffery S Cox. The type i ifn response to infection with mycobacterium tuberculosis requires esx-1-mediated secretion and contributes to pathogenesis. *Journal of immunology (Baltimore, Md. : 1950)*, 178(5):3143–52, 2007.

[284] Tyler N Starr and Joseph W Thornton. Epistasis in protein evolution. *Protein science : a publication of the Protein Society*, 25(7):1204–18, 2016.

[285] 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.

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

[287] Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, 10(10):681–90, 2009.

[288] T. Strachan and A.P. Read. *Human Molecular Genetics 4*. 2011.

[289] 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.

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

[291] 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.

[292] Ludovic Tailleux, Olivier Neyrolles, Stéphanie Honoré-Bouakline, Emmanuelle Perret, Françoise Sanchez, Jean-Pierre Abastado, Philippe Henri Lagrange, Jean Claude Gluckman, Michelle Rosenzwajg, and Jean-Louis Herrmann. Constrained intracellular survival of mycobacterium tuberculosis in human dendritic cells. *Journal of immunology (Baltimore, Md. : 1950)*, 170(4):1939–48, 2003.

[293] 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.

[294] 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.

[295] Nelson Leung-Sang Tang, Harris Pok Yin Fan, Kwok Chiu Chang, Jasmine Kuk Lai Ching, Kathy Pui Shan Kong, Wing Wai Yew, Kai Man Kam, Chi Chiu Leung, Cheuk Ming Tam, Jenefer Blackwell, and Chiu Yeung Chan. Genetic association between a chemokine gene cxcl-10 (ip-10, interferon gamma inducible protein 10) and susceptibility to tuberculosis. *Clinica chimica acta; international journal of clinical chemistry*, 406(1-2):98–102, 2009.

[296] Mahnaz Tanveer, Zahra Hasan, Akbar Kanji, Rabia Hussain, and Rumina Hasan. Reduced tnf-alpha and ifn-gamma responses to central asian strain 1 and beijing isolates of mycobacterium tuberculosis in comparison with h37rv strain. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(6):581–7, 2009.

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

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

[299] Duncan Thomas. Gene–environment-wide association studies: emerging approaches. *Nature reviews. Genetics*, 11(4):259–72, 2010.

[300] 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.

[301] 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.

[302] 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.

[303] J M Torres, N J Cox, and L H Philipson. Genome wide association studies for diabetes: perspective on results and challenges. *Pediatric diabetes*, 14(2):90–6, 2013.

[304] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–30, 2013.

[305] Alexander M Tsankov, Hongcang Gu, Veronika Akopian, Michael J Ziller, Julie Donaghey, Ido Amit, Andreas Gnirke, and Alexander Meissner. Transcription factor binding dynamics during human es cell differentiation. *Nature*, 518(7539):344–9, 2015.

[306] 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.

[307] Michael C Turchin, Charleston W K Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, and Joel N Hirschhorn. Evidence of widespread selection on standing variation in europe at height-associated snps. *Nature genetics*, 44(9):1015–9, 2012.

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

[309] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–3, 2015.

[310] Nicole van der Wel, David Hava, Diane Houben, Donna Fluitsma, Maaike van Zon, Jason Pierson, Michael Brenner, and Peter J. Peters. M. tuberculosis and m. leprae translocate from the phagolysosome to the cytosol in myeloid cells. *Cell*, 129(7):1287–98, 2007.

[311] T K VanHeyningen, H L Collins, and D G Russell. Il-6 produced by macrophages infected with mycobacterium species suppresses t cell responses. *Journal of immunology (Baltimore, Md. : 1950)*, 158(1):330–7, 1997.

[312] Fredrik O Vannberg, Stephen J Chapman, and Adrian V S Hill. Human genetic susceptibility to intracellular pathogens. *Immunological reviews*, 240(1):105–16, 2011.

[313] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63, 2009.

[314] J C Venter, M D Adams, E W Myers, P W Li, Richard J Mural, Granger G Sutton, H O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, Jeannine D Gocayne, Peter Amanatides, Richard M Ballew, Daniel H Huson, Jennifer Russo Wortman, Q Zhang, Chinnappa D Kodira, X H Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, Samuel Broder, Andrew G Clark, Joe Nadeau, V A McKusick, Norton Zinder, Arnold J Levine, Richard J Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher,

Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, J Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, V Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J Heiman, Maureen E Higgins, R R Ji, Zhaoxi Ke, Karen A Ketchum, Zhongwu Lai, Yiding Lei, Z Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V Merkulov, Natalia Milshina, Helen M Moore, Ashwinikumar K Naik, Vaibhav A Narayan, Beena Neelam, Deborah Nusskern, Douglas B Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Z Wang, A Wang, X Wang, J Wang, Ming-hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, W Zhang, Hongyu Zhang, Q Zhao, Liansheng Zheng, F Zhong, Wenyan Zhong, S Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, A Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, J Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, S McCawley, T McIntosh, I McMullen, M Moy, Linda Moy, B Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Y H Rogers, Deanna Romblad, Bob Ruhfel, R Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, G Wang, Jeremy Wetter, S Williams, Monica Williams, Sandra Windsor, E Winn-Deen, Keriellen Wolfe, J Zaveri, Karena Zaveri, Josep F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, J Carnes-Stine, Parris Caulk, Y H Chiang, My Coyne, Carl Dahlke, A Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, C Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, J McDaniel, Sean Murphy, Matthew Newman, T Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, D Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, 2001.

[315] Isabelle Vergne, Jennifer Chua, Sudha B Singh, and Vojo Deretic. Cell biology of mycobacterium tuberculosis phagosome. *Annual review of cell and developmental biology*, 20:367–94, 2004.

[316] Mark Verway, Manuella Bouttier, Tian-Tian Wang, Marilyn Carrier, Mario Calderon, Beum-Soo An, Emmanuelle Devemy, Fiona McIntosh, Maziar Divangahi, Marcel a. Behr, and John H. White. Vitamin d induces interleukin-1β expression: paracrine macrophage epithelial signaling controls m. tuberculosis infection. *PLoS pathogens*, 9(6):e1003407, 2013.

[317] 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.

[318] Karl Waern, Ugrappa Nagalakshmi, and Michael Snyder. Rna sequencing. *Methods in molecular biology (Clifton, N.J.)*, 759(1):125–32, 2011.

[319] 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.

[320] Chongzhen Wang, Pascale Peyron, Olga Mestre, Gilla Kaplan, Dick van Soolingen, Qian Gao, Brigitte Gicquel, and Olivier Neyrolles. Innate immune response to mycobacterium tuberculosis beijing and other genotypes. *PloS one*, 5(10):e13594, 2010.

[321] Xinchen Wang, Nathan R Tucker, Gizem Rizki, Robert Mills, Peter HL Krijger, Elzo de Wit, Vidya Subramanian, Eric Bartell, Xinh-Xinh Nguyen, Jiangchuan Ye, Jordan Leyton-Mange, Elena V Dolmatova, Pim van der Harst, Wouter de Laat, Patrick T Ellinor, Christopher Newton-Cheh, David J Milan, Manolis Kellis, and Laurie A Boyer. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *eLife*, 5:1–24, 2016.

[322] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

[323] Ruth Wassermann, Muhammet F. Gulen, Claudia Sala, Sonia Garcia Perin, Ye Lou, Jan Rybniker, Jonathan L. Schmid-Burgk, Tobias Schmidt, Veit Hornung, Stewart T. Cole, and Andrea Ablasser. Mycobacterium tuberculosis differentially activates cgas- and inflammasome-dependent intracellular immune responses through esx-1. *Cell host & microbe*, 17(6):799–810, 2015.

[324] Robert O. Watson, Samantha L. Bell, Donna A. MacDuff, Jacqueline M. Kimmey, Elie J. Diner, Joanna Olivas, Russell E. Vance, Christina L. Stallings, Herbert W. Virgin, and Jeffery S. Cox. The cytosolic sensor cgas detects mycobacterium tuberculosis dna to induce type i interferons and activate autophagy. *Cell host & microbe*, 17(6):811–9, 2015.

[325] Robert O. Watson, Paolo S. Manzanillo, and Jeffery S. Cox. Extracellular m. tuberculosis dna targets bacteria for autophagy by activating the host dna-sensing pathway. *Cell*, 150(4):803–15, 2012.

[326] 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.

[327] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, 2014.

[328] Nathan D Wolfe, Claire Panosian Dunavan, and Jared Diamond. Origins of major human infectious diseases. *Nature*, 447(7142):279–83, 2007.

[329] Margreet a Wolfert and Geert-Jan Boons. Adaptive immune activation: glycosylation does matter. *Nature chemical biology*, 9(12):776–84, 2013.

[330] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, Najaf Amin, Martin L Buchkovich, Damien C Croteau-Chonka, Felix R Day, Yanan Duan, Tove Fall, Rudolf Fehrmann, Teresa Ferreira, Anne U Jackson, Juha Karjalainen, Ken Sin Lo, Adam E Locke, Reedik Mägi, Evelin Mihailov, Eleonora Porcu, Joshua C Randall, André Scherag, Anna A E Vinkhuyzen, Harm-Jan Westra, Thomas W Winkler, Tsegaselassie Workalemahu, Jing Hua Zhao, Devin Absher, Eva Albrecht, Denise Anderson, Jeffrey Baron, Marian Beekman, Ayse Demirkan, Georg B Ehret, Bjarke Feenstra, Mary F Feitosa, Krista Fischer, Ross M Fraser, Anuj Goel, Jian Gong, Anne E Justice, Stavroula Kanoni, Marcus E Kleber, Kati Kristiansson, Unhee Lim, Vaneet Lotay, Julian C Lui, Massimo Mangino, Irene Mateo Leach, Carolina Medina-Gomez, Michael A Nalls, Dale R Nyholt, Cameron D Palmer, Dorota Pasko, Sonali Pechlivanis, Inga Prokopenko, Janina S Ried, Stephan Ripke, Dmitry Shungin, Alena Stancáková, Rona J Strawbridge, Yun Ju Sung, Toshiko Tanaka, Alexander Teumer, Stella Trompet, Sander W van der Laan, Jessica van Setten, Jana V Van Vliet-Ostaptchouk, Zhaoming Wang, Loïc Yengo, Weihua Zhang, Uzma Afzal, Johan Arnlöv, Gillian M Arscott, Stefania Bandinelli, Amy Barrett, Claire Bellis, Amanda J Bennett, Christian Berne, Matthias Blüher, Jennifer L Bolton, Yvonne Böttcher, Heather A Boyd, Marcel Bruinenberg, Brendan M Buckley, Steven Buyske, Ida H Caspersen, Peter S Chines, Robert Clarke, Simone Claudi-Boehm, Matthew Cooper, E Warwick Daw, Pim A De Jong, Joris Deelen, Graciela Delgado, Josh C Denny, Rosalie Dhonukshe-Rutten, Maria Dimitriou, Alex S F Doney, Marcus Dörr, Niina Eklund, Elodie Eury, Lasse Folkersen, Melissa E Garcia, Frank Geller, Vilmantas Giedraitis, Alan S Go, Harald Grallert, Tanja B Grammer, Jürgen Gräßler, Henrik Grönberg, Lisette C P G M de Groot, Christopher J Groves, Jeffrey Haessler, Per Hall, Toomas Haller, Goran Hallmans, Anke Hannemann, Catharina A Hartman, Maija Hassinen,

Caroline Hayward, Nancy L Heard-Costa, Quinta Helmer, Gibran Hemani, Anjali K Henders, Hans L Hillege, Mark A Hlatky, Wolfgang Hoffmann, Per Hoffmann, Oddgeir Holmen, Jeanine J Houwing-Duistermaat, Thomas Illig, Aaron Isaacs, Alan L James, Janina Jeff, Berit Johansen, Åsa Johansson, Jennifer Jolley, Thorhildur Juliusdottir, Juhani Junttila, Abel N Kho, Leena Kinnunen, Norman Klopp, Thomas Kocher, Wolfgang Kratzer, Peter Lichtner, Lars Lind, Jaana Lindström, Stéphane Lobbens, Mattias Lorentzon, Yingchang Lu, Valeriya Lyssenko, Patrik K E Magnusson, Anubha Mahajan, Marc Maillard, Wendy L McArdle, Colin A McKenzie, Stela McLachlan, Paul J McLaren, Cristina Menni, Sigrun Merger, Lili Milani, Alireza Moayyeri, Keri L Monda, Mario A Morken, Gabriele Müller, Martina Müller-Nurasyid, Arthur W Musk, Narisu Narisu, Matthias Nauck, Ilja M Nolte, Markus M Nöthen, Laticia Oozageer, Stefan Pilz, Nigel W Rayner, Frida Renstrom, Neil R Robertson, Lynda M Rose, Ronan Roussel, Serena Sanna, Hubert Scharnagl, Salome Scholtens, Fredrick R Schumacher, Heribert Schunkert, Robert A Scott, Joban Sehmi, Thomas Seufferlein, Jianxin Shi, Karri Silventoinen, Johannes H Smit, Albert Vernon Smith, Joanna Smolonska, Alice V Stanton, Kathleen Stirrups, David J Stott, Heather M Stringham, Johan Sundström, Morris A Swertz, Ann-Christine Syvänen, Bamidele O Tayo, Gudmar Thorleifsson, Jonathan P Tyrer, Suzanne van Dijk, Natasja M van Schoor, Nathalie van der Velde, Diana van Heemst, Floor V A van Oort, Sita H Vermeulen, Niek Verweij, Judith M Vonk, Lindsay L Waite, Melanie Waldenberger, Roman Wennauer, Lynne R Wilkens, Christina Willenborg, Tom Wilsgaard, Mary K Wojczynski, Andrew Wong, Alan F Wright, Qunyuan Zhang, Dominique Arveiler, Stephan J L Bakker, John Beilby, Richard N Bergman, Sven Bergmann, Reiner Biffar, John Blangero, Dorret I Boomsma, Stefan R Bornstein, Pascal Bovet, Paolo Brambilla, Morris J Brown, Harry Campbell, Mark J Caulfield, Aravinda Chakravarti, Rory Collins, Francis S Collins, Dana C Crawford, L Adrienne Cupples, John Danesh, Ulf de Faire, Hester M den Ruijter, Raimund Erbel, Jeanette Erdmann, Johan G Eriksson, Martin Farrall, Ele Ferrannini, Jean Ferrières, Ian Ford, Nita G Forouhi, Terrence Forrester, Ron T Gansevoort, Pablo V Gejman, Christian Gieger, Alain Golay, Omri Gottesman, Vilmundur Gudnason, Ulf Gyllensten, David W Haas, Alistair S Hall, Tamara B Harris, Andrew T Hattersley, Andrew C Heath, Christian Hengstenberg, Andrew A Hicks, Lucia A Hindorff, Aroon D Hingorani, Albert Hofman, G Kees Hovingh, Steve E Humphries, Steven C Hunt, Elina Hypponen, Kevin B Jacobs, Marjo-Riitta Jarvelin, Pekka Jousilahti, Antti M Jula, Jaakko Kaprio, John J P Kastelein, Manfred Kayser, Frank Kee, Sirkka M Keinanen-Kiukaanniemi, Lambertus A Kiemeney, Jaspal S Kooner, Charles Kooperberg, Seppo Koskinen, Peter Kovacs, Aldi T Kraja, Meena Kumari, Johanna Kuusisto, Timo A Lakka, Claudia Langenberg, Loic Le Marchand, Terho Lehtimäki, Sara Lupoli, Pamela A F Madden, Satu Männistö, Paolo Manunta, André Marette, Tara C Matise, Barbara McKnight, Thomas Meitinger, Frans L Moll, Grant W Montgomery, Andrew D Morris, Andrew P Morris, Jeffrey C Murray, Mari Nelis, Claes Ohlsson, Albertine J Oldehinkel, Ken K Ong, Willem H Ouwehand, Gerard Pasterkamp, Annette Peters, Peter P Pramstaller, Jackie F Price, Lu Qi, Olli T Raitakari, Tuomo Rankinen, D C Rao, Treva K Rice, Marylyn Ritchie, Igor Rudan, Veikko Salomaa, Nilesh J Samani, Jouko Saramies,

Mark A Sarzynski, Peter E H Schwarz, Sylvain Sebert, Peter Sever, Alan R Shuldiner, Juha Sinisalo, Valgerdur Steinthorsdottir, Ronald P Stolk, Jean-Claude Tardif, Anke Tönjes, Angelo Tremblay, Elena Tremoli, Jarmo Virtamo, Marie-Claude Vohl, Electronic Medical Records and Genomics (eMEMERGEGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, Philippe Amouyel, Folkert W Asselbergs, Themistocles L Assimes, Murielle Bochud, Bernhard O Boehm, Eric Boerwinkle, Erwin P Bottinger, Claude Bouchard, Stéphane Cauchi, John C Chambers, Stephen J Chanock, Richard S Cooper, Paul I W de Bakker, George Dedoussis, Luigi Ferrucci, Paul W Franks, Philippe Froguel, Leif C Groop, Christopher A Haiman, Anders Hamsten, M Geoffrey Hayes, Jennie Hui, David J Hunter, Kristian Hveem, J Wouter Jukema, Robert C Kaplan, Mika Kivimaki, Diana Kuh, Markku Laakso, Yongmei Liu, Nicholas G Martin, Winfried März, Mads Melbye, Susanne Moebus, Patricia B Munroe, Inger Njølstad, Ben A Oostra, Colin N A Palmer, Nancy L Pedersen, Markus Perola, Louis Pérusse, Ulrike Peters, Joseph E Powell, Chris Power, Thomas Quertermous, Rainer Rauramaa, Eva Reinmaa, Paul M Ridker, Fernando Rivadeneira, Jerome I Rotter, Timo E Saaristo, Danish Saleheen, David Schlessinger, P Eline Slagboom, Harold Snieder, Tim D Spector, Konstantin Strauch, Michael Stumvoll, Jaakko Tuomilehto, Matti Uusitupa, Pim van der Harst, Henry Völzke, Mark Walker, Nicholas J Wareham, Hugh Watkins, H-Erich Wichmann, James F Wilson, Pieter Zanen, Panos Deloukas, Iris M Heid, Cecilia M Lindgren, Karen L Mohlke, Elizabeth K Speliotes, Unnur Thorsteinsdottir, Inês Barroso, Caroline S Fox, Kari E North, David P Strachan, Jacques S Beckmann, Sonja I Berndt, Michael Boehnke, Ingrid B Borecki, Mark I McCarthy, Andres Metspalu, Kari Stefansson, André G Uitterlinden, Cornelia M van Duijn, Lude Franke, Cristen J Willer, Alkes L Price, Guillaume Lettre, Ruth J F Loos, Michael N Weedon, Erik Ingelsson, Jeffrey R O'Connell, Goncalo R Abecasis, Daniel I Chasman, Michael E Goddard, Peter M Visscher, Joel N Hirschhorn, and Timothy M Frayling. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–86, 2014.

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

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

[333] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature reviews. Genetics*, 14(7):507–15, 2013.

[334] 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.

[335] 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.

[336] Jingyan Xia, Liyun Shi, Lifang Zhao, and Feng Xu. Impact of vitamin d supplementation on the outcome of tuberculosis treatment: a systematic review and meta-analysis of randomized controlled trials. *Chinese medical journal*, 127(17):3127–34, 2014.

[337] Guanghua Xu, Jing Wang, George Fu Gao, and Cui Hua Liu. Insights into battles between mycobacterium tuberculosis and macrophages. *Protein & cell*, 5(10):728–36, 2014.

[338] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–9, 2010.

[339] Jian Yang, Noah a. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–6, 2014.

[340] 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.

[341] Jae-Joon Yim and Paramasivam Selvaraj. Genetic susceptibility in tuberculosis. *Respirology (Carlton, Vic.)*, 15(2):241–56, 2010.

[342] Kenneth S Zaret and Susan E Mango. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current opinion in genetics & development*, 37:76–81, 2016.

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

[344] Wei Zhao, Xiaping He, Katherine A Hoadley, Joel S Parker, David Neil Hayes, and Charles M Perou. Comparison of rna-seq by poly (a) capture, ribosomal rna depletion, and dna microarray for expression profiling. *BMC genomics*, 15(1):419, 2014.

[345] Xing Wu Zhu and Jon S Friedland. Multinucleate giant cells and the control of chemokine secretion in response to mycobacterium tuberculosis. *Clinical immunology (Orlando, Fla.)*, 120(1):10–20, 2006.

[346] Sanjay P Zodpey and Sunanda N Shrikhande. The geographic location (latitude) of studies evaluating protective effect of bcg vaccine and it's efficacy/effectiveness against tuberculosis. *Indian journal of public health*, 51(4):205–10, 2007.