Discovery and Functional Interpretation of Genetic Risk in Autoimmune Diseases

Discovery and Functional Interpretation of Genetic Risk in Autoimmune Diseases

A dissertation presented

by

Xinli Hu

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

May 2015

Dissertation Advisor: Dr. Soumya Raychaudhuri Xinli Hu

Discovery and Functional Interpretation of Genetic Risk in Autoimmune Diseases

Abstract

Autoimmune diseases are chronic and debilitating conditions arising from abnormal immune responses directed against normal body tissues; they collectively affect the lives of 5-10% of the world population. These diseases often show familial clustering, suggesting strong genetic heritability. For many of autoimmune diseases, variation in the human leukocyte antigen (HLA) genes is the primary modulator of genetic risk. Recently, genome-wide association studies (GWAS) identified hundreds of genomic regions outside the HLA that harbor additional risk-conferring variants. The ultimate goal is to identify the precise causal variants and understand the mechanisms by which they lead to autoimmunity, which is challenged by complexities of the genome and the immune system.

In this work, my colleagues and I developed and applied experimental and computational tools to reveal critical clues from multiple genetic and biological data types. First, we devised a statistical algorithm to identify the critical cell types involved in different autoimmune diseases. Two strongly heritable and common diseases, rheumatoid arthritis (RA) and type 1 diabetes (T1D), both involve the adaptive immune system, specifically the $CD4^+$ T cells. We then conducted focused studies in $CD4^+$ T cells using high-throughput genomic and proteomic technologies, and showed that immunological phenotypes and functions varied with genetic differences across individuals. To facilitate this study, we developed an automated computational tool to efficiently and reliably analyze the large-scale data. Finally, the HLA genes, which encode a family of highly variable antigen-recognition proteins, are the longest-known and strongest modulators of genetic risk in T1D. However, the extraordinary level of polymorphism and complex structure in the HLA region

largely hindered precise localization and functional investigation of the causal mutations. We used dense-genotyping and robust statistical analyses to pinpoint the amino acid residue changes at a few key amino acid positions that explained the majority of disease risk within the HLA.

The work presented in this dissertation revealed the specific immune cell populations, genetic variants, and cellular functions that affect RA, T1D, and other autoimmune diseases. Furthermore, it offers a rational framework, as well as powerful open-source computational tools, that can be applied in future functional genomic studies.

TABLE OF CONTENTS

ACKNOWLEDGMENT

First and foremost, I thank my amazing parents for their unconditional love and support. You encourage me to do my best professionally; more importantly, you remind me to remain balanced and humble. I would not have come this far in school or in life without your inexhaustible wisdom and forgiveness. If I had achieved anything at all, all credit goes to you.

I have been extremely fortunate to have Dr. Soumya Raychaudhuri as my PhD advisor. Thank you for teaching me everything I know about genetics. You are not only an incredible physician-scientist and intellect, but also a marvelous person and a great friend. Your knowledge and compassion have made every aspect of the past six years incredibly fulfilling and truly enjoyable. Every fellow trainee who has been lucky enough to receive your mentorship appreciates and admires your special ability to "keep it real"; I am especially grateful for your patience, understanding, and a great sense of humor, which were invaluable and kept me going through even the toughest times.

I would like to thank all past and current members of the Raychaudhuri lab for their support and companionship. You are a special crowd, and made NRB255 the most relaxed, friendly, fun, and productive work environment.

I would also like to thank members of my dissertation advisory committee, Drs. Joel Hirschhorn, Vijay Kuchroo, and George Church, for your time and helpful discussions. In addition, I would like to thank Drs. Jill Mesirov, David Hafler, and Philip De Jager, who introduced me to this wonderful community of researchers at Harvard/MIT. Working with you, I first experienced the ins and outs of multidisciplinary research, and came to love it.

Thank you to my family and friends near and far, who supported me over the years. Thank you to Yin-Yin Wang, Jennifer Wei, and my cousin Zhang Ying, for your companionship, for listening to me and keeping me going through thick and thin. Thank you to Kuai Letian for sharing with me

your passion for science and life. Thank you to Zhang Zhourui for being a great friend and sister for the last 15 years. Thanks to all of my childhood friends in Nanjing for keeping the seemingly distant parts of my life always colorful.

Finally, I would like to thank HMS/HST for giving me this rare opportunity, and my classmates for being such inspiring and generous people. All across the world, there are many with talent and potential, but for whom opportunity is a luxury. I am grateful for all that has been bestowed upon me; it has been a humbling experience.

ATTRIBUTIONS

Chapter 2

Xinli Hu designed and wrote the algorithm, performed data analysis, and wrote the manuscript.

Hyun Kim performed data analysis and edited the manuscript.

Eli Stahl and Robert Plenge provided statistical guidance for data analysis and edited the manuscript.

Mark Daly conceived of the study and provided guidance for data analysis.

Soumya Raychaudhuri conceived of the study, designed and wrote the algorithm, performed data analysis, and wrote the manuscript.

Chapter 3

Xinli Hu conceived of the study, designed and wrote the algorithm, performed data analysis, and wrote the manuscript.

Hyun Kim performed all experiments and manual analyses, and edited the manuscript.

Patrick J. Brennan designed experiments and assays for the iNKT cell studies, provided guidance on experimental protocols and data analysis, and edited the manuscript.

Buhm Han provided critical feedback on data analysis and edited the manuscript.

Clare Baecher-Allan provided critical guidance on the design and experimental protocols of the T cell study, and edited the manuscript.

Philip L. De Jager provided samples for the study, and edited the manuscript.

Michael B. Brenner provided critical guidance on experimental design and edited the manuscript.

Soumya Raychaudhuri conceived of the study, designed the experiments, provided guidance on data analysis, and edited the manuscript.

Chapter 4

Xinli Hu designed the study, performed data analyses, and wrote the manuscript.

Hyun Kim performed all experiments, conducted data analyses, and wrote the manuscript.

Towfique Raj ,Gosia Trynka, Kamil Slowikowski and Nick Teslovich performed data analyses, and edited the manuscript.

Patrick J. Brennan designed the experiments, provided feedback on data analysis, and edited the manuscript.

Wei-Min Chen and Suna Onengut provided genotype data and edited the manuscript.

Clare Baecher-Allan provided critical guidance on the design and experimental protocols of the T cell study, and edited the manuscript.

Philip L. De Jager provided genotype data and blood samples.

Stephen S. Rich provided genotype data, critical guidance on data analyses, and edited the manuscript.

Barbara E. Stranger provided feedback on data analyses and edited the manuscript.

Michael B. Brenner provided guidance on experimental protocol, data analyses, and edited the manuscript.

Soumya Raychaudhuri conceived the study, designed the experiments, performed and provided guidance on data analyses, and wrote the manuscript.

Chapter 5

Xinli Hu and Aaron J. Deutsch designed and performed the analyses, and wrote the manuscript.

Tobias L. Lenz performed data analysis, and edited the manuscript.

Suna Onengut-Gumuscu and Wei-Min Chen provided genotype data and guidance on data analysis.

Buhm Han provided scripts and critical guidance on data analysis, and edited the manuscript.

Joanna M. M. Howson, John A. Todd and Paul I. W. de Bakker provided critical feedback on data analysis, and edited the manuscript.

Paul I. W. de Bakker provided critical guidance on data analysis, and edited the manuscript.

Stephen S. Rich provided genotype data, guidance on data analysis, and edited the manuscript

Soumya Raychaudhuri designed the study, provided guidance on data analysis, and edited the manuscript.

CHAPTER 1

Introduction

What have we learned from GWAS in autoimmune diseases; and what is next?

Xinli Hu[1-4], Mark Daly*[4,5]

1. Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology, Boston, MA USA
2. Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
3. Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142
5. Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114

Corresponding author: Mark Daly (mjdaly@atgu.mgh.harvard.edu)

Originally published as:

Hu, X. and M. Daly (2012). "What have we learned from six years of GWAS in autoimmune diseases, and what is next?" Curr Opin Immunol **24**(5): 571-575.

ABSTRACT

Genome-wide association studies (GWAS) have discovered hundreds of common genetic variants that predispose humans to autoimmune diseases, opening up unprecedented potential for elucidating the pathways and processes of disease. To understand the role of these variants in susceptibility, we need to derive mechanistic insight by integration of genetic results with other biological data types and also with careful functional studies. In many cases, such studies have highlighted coherent biological processes at a high level and elucidated specific mechanisms that contribute to autoimmunity and inflammation. The understanding of the genetic component of autoimmune etiology will become more complete as fine-mapping and sequencing data become readily available. A comprehensive catalog of human immune phenotypes could provide a functional basis for assessing genetic influence on immune function and variation in response to therapeutic interventions, as well as for rationally designing new targeted therapeutics.

INTRODUCTION

Autoimmune diseases are a clinically diverse group of diseases caused by inappropriate or hyperactive immune responses against tissues and substances normally considered "self." Although malfunction of the immune system is clearly indicated, the mechanism of autoimmune pathogenesis is still far from elucidated. Family and twin studies have long suggested that genetics contributes significantly to autoimmunity, and earlier linkage studies identified a handful of strong genetic variants primarily in the major histocompatibility (MHC) region encoding human leukocyte antigens (HLA). Since the completion of the HapMap Project and development of massively parallel array-based genotyping technologies, genome-wide association studies (GWAS) in search for common variants have been extensively carried out in complex diseases. In this review, we examine approaches that have been taken and progress made to understanding of autoimmune genetics through GWAS, with examples of mechanistic insight derived from follow-up studies. We also suggest areas where association studies can be improved to yield further insights.

Since 2006, hundreds of single-nucleotide polymorphisms (SNPs) have been associated with rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), multiple sclerosis (MS), Crohn's disease, ulcerative colitis (UC), type 1 diabetes (T1D), celiac disease, psoriasis and other autoimmune diseases [1]. While a significant accomplishment given the limited historical successes in complex trait genetics, these studies by themselves have limited value without subsequent steps to link emergent genetics to function by pinpointing causal variants, deriving biological relevance, and elucidating pathogenic mechanisms. Although occasionally these risk loci harbor functional coding variants in genes known to be involved in immune processes that alter protein sequences, most others implicate genes with unknown function or are simply non-coding. Furthermore all the currently validated variants have moderate effect sizes and account for only a fraction of the heritability [2]. Therefore, even with an explosion of genetics data, a reasonable question arose: are

GWAS results truly useful for either understanding etiology, or for clinical diagnosis and management of autoimmune diseases?

HOW DO WE LEARN ABOUT AUTOIMMUNE DISEASES FROM GWAS?

One critical challenge in defining disease biology from associated alleles is the absence of an understanding of the specific genes implicated and the molecular perturbations created by associated alleles. However what GWAS potentially provides is the genetic architecture that underlies autoimmunity, around which the relevant physiological and pathological pathways can be constructed. Following SNP discovery, investigators have devised statistical methods and carried out functional studies to elucidate relevant biology. The approaches taken by these studies can be broadly grouped into two major categories, as described in **Figure 1.1**: 1) a traditional "bottom-up" approach in which individual alleles, genes and proteins are examined to gain insight into molecular mechanisms of disease and 2) a "top-down" approach, where lists of disease-associated regions are examined together to outline relevant biological systems and pathways.

"Bottom-up" approach: focused follow-up studies directly shed light on molecular mechanisms of pathogenesis.

In traditional Mendelian genetics, the identification of disease causing mutations in a single gene would be immediately followed by targeted follow-up efforts to understand the function of the gene and how its alteration precipitates disease. Similarly, conducting focused functional and statistical follow-up studies in well-validated loci have quite directly suggested specific molecular mechanisms through which allelic variants confer risk or protection in autoimmune diseases [3-13]. For example, Pidascheva et al. determined that the R381Q (rs11290926) variant in *IL23R*, which is associated with reduced risk of

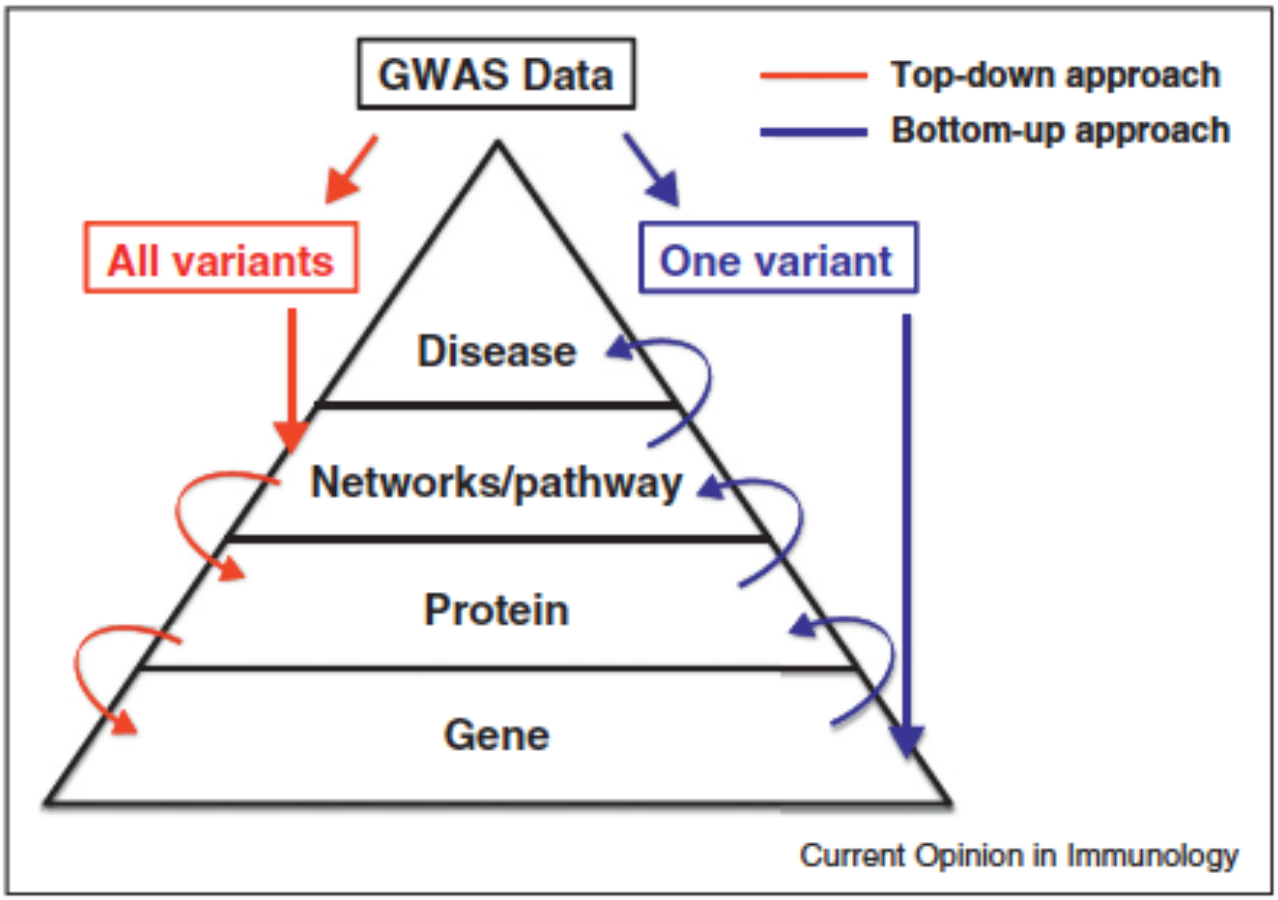


Figure 1.1 Two broad approaches taken by GWAS follow-up studies. In a more traditional 'bottom-up' approach, a specific variant/gene is selected from the GWAS results to investigate its functions and the effect of perturbation in the context of the disease. In a 'top-down' approach, the entire list of variants identified by GWAS is analyzed together, often incorporating existing biological data types, to highlight processes and pathways pertinent to disease, which helps to prioritize and focus additional follow-ups such as targeted functional assays and/or discovery of more risk variants.

inflammatory bowel disease, results in a loss of function which reduces T-cell activation in response to IL-23, lending evidence that the IL-23 pathway is pathogenic [5]. Leucine-rich repeat kinase 2 (*LRRK2*) is another major risk gene for Crohn's disease. Liu et al. [6] investigated the molecular

mechanisms by which LRRK2 negatively regulates NFAT1 which mediates T-cell activation as well as cytokine production in antigen presenting cells, and showed that LRRK2-deficiency conferred increased risk to colitis in mouse model. And in one of the clearest examples of novel insights gained from GWAS, the role of autophagy in Crohn's disease, previously unsuspected before the first GWAS studies, has now become a major therapeutic target in disease owing to the elucidation of the role of variation at *ATG16L1* and *IRGM* [described in detail by Garder and Xavier in this issue]. In these follow-up studies that characterize specific candidate genes that emerged from GWAS, the functional effect of variants on immune response has been clearly illustrated.

In other examples, using detailed statistical approaches Raychaudhuri et al. [8] delved into one of the longest-known susceptibility loci associated with anti-CCP-positive rheumatoid arthritis, namely the MHC region. They imputed alleles in the region including *HLA-DRB1* using existing GWAS genotypes, followed by conditional analysis, and successfully fine-mapped the causal variants to sequences of five amino acids located the binding groove of three MHC molecules. This analysis improved upon the "shared epitope" theory that had existed for more than 30 years [14], and is a successful example of delineating the causal variants from GWAS data. Okada et al. focused on the HLA-Cw*1202-B*5201-DRB1*1502 haplotype and were able to tease apart its opposite effects on the two closely-related inflammatory bowel diseases in the Japanese population, specifically increasing susceptibility to ulcerative colitis but reducing risk of Crohn's disease [7]. This observation supports the hypothesis that Crohn's disease is a Th1-like disease (preferentially producing a repertoire of Th1 cytokines such as interferon-γ) whereas UC is more Th2-like; and susceptibility difference between the two are manifestations of Th1/Th2 imbalance in immune response induced by differential helper T-cell responses to specific pathogens.

"Top-down" approaches: examining GWAS results in aggregate outlines coherent biological processes

While the bottom-up approach is more tractable and traditional, the realization emerging from GWAS that autoimmune diseases are the result of many, likely hundreds, of simultaneously contributing genes and variants has borne an alternate approach to the problem by considering all associations together. Many analytic methods have therefore been developed to interpret GWAS results by examining groups of risk regions or implicated genes in aggregate, incorporating information from other biological data types such as gene and protein expression [15-24]. Encouragingly, these analyses have often convincingly shown that disease-associated regions implicate coherent biological processes and molecular pathways. For example, for many autoimmune diseases, significant enrichment of physical interactions among genes in associated risk loci has been demonstrated, implicating organized networks of biological pathways [23]. Similarly, the evaluation of cell-specific expression of autoimmune SNPs-implicated genes showed enrichment in particular immune cells, providing an unbiased method for using raw GWAS results to identify likely pathogenic cell types [19]. Such analyses must be undertaken with care – biases in both the genetic discovery data (e.g., GWAS arrays preferentially discover associations that are both gene and SNP rich, certain gene families are systematically much larger than others) and in the biological data to be integrated (e.g., certain families of genes are more well-studied in protein interaction experiments and more well-represented on expression arrays) require careful attention in the evaluation of statistical significance.

If managed properly, these issues can be overcome, and when GWAS results are shown to implicate regions that contain genes that are compellingly non-randomly drawn from specific pathways or which are unusually closely related in the context of independent functional data, such relationships clearly indicate a convincing biological insight into disease has emerged from the collective GWAS results set. In one study, Lee et al. constructed a genome-scale human gene network using genomic and proteomic data. Studying genes regions associated with Crohn's disease, they observed the interaction of genes in the TNF pathway, the Th17-differentiation network, as

well as genes involved in autophagy. Moreover, they were able to combine this approach to boost the detection power of GWAS, and suggest more candidate genes based on their interaction partners in highlighted pathways, such as *GRB2* and *SHC1* [20].

A number of overlapping HLA and non-HLA risk alleles have emerged across autoimmune diseases, implicating shared pathogenic pathways among phenotypically heterogeneous diseases. It is intuitive that immune response pathways would be broadly involved in all autoimmune diseases; as well as diverging pathways should eventually lead to variation in clinical presentations. Systematically examining the shared alleles provides one means for comparing and contrasting autoimmune diseases in terms genetic backgrounds and implications on shared pathogenic mechanisms [25-31]. As one example, Ramos et al. investigated the genetic connections across 17 autoimmune diseases, and noted the strongest correlations between RA and T1D as well as between Crohn's disease and UC, suggesting common pathogenic mechanistic pathways. They also noted that SLE possesses a distinct genetic background, supporting partial pleiotropy among autoimmune diseases [28].

The "top-down" approach often does not directly yield validated risk genes and variants, however it is powerful in highlighting coherent biological processes relevant to the diseases, and in turn providing guidance to focus statistical and functional follow-ups on the most interesting candidate genomic regions and the most relevant molecular and cellular models.

HOW CAN AUTOIMMUNE GWAS BE IMPROVED?

Identifying the remaining genetic architecture

It is estimated that all the validated SNPs in aggregate explain a fraction hovering around 10-40% of total heritability in most complex-trait diseases [32, 33], making missing heritability a glaring issue with existing GWAS results. There are several oft-discussed possible sources for what has been termed this "missing heritability", including the potentially large-effect rare and structural

variants, gene-gene interactions (epistasis), gene-environment interactions, as well as many more common variants that simply have not been discovered due to the lack of power as limited by sample size [34, 35]. Stahl et al. modeled the polygenic architecture based on Bayesian inference in several complex-trait diseases, and suggested that thousands of undiscovered common SNPs could explain an additional 20% and 43% of disease risk in RA and celiac disease, respectively [36]. Similarly Park et al. predicted that for Crohn's disease, more than 140 risk loci with effect sizes comparable to the first 32 validated loci together would account for 20% of the heritability [37].

Next-generation deep sequencing in candidate regions has been successful in discovering a handful of rare variants associated with several diseases [38-42]. Examples include a splice variant in *CARD9* (OR = 0.29) associated with reduced risk of Crohn's disease [40], four protective rare variants in *IFIH1* in T1D [41], as well as several rare variants contributing to the risk of celiac disease discovered by Trynka et al. using data from pilot 1000 Genome Project studies and other resequencing data [42]. Much more than increasing heritability explained by identifying rare variation missed by GWAS, such studies can provide clearer functional insights into the roles of individual genes and in the case of strongly protective variants, articulate clear therapeutic hypotheses that can be pursued.

Harnessing the value of genetics: profiling the human immunophenotypes is necessary to link genetics to function in autoimmune diseases

The value of genetic knowledge we have gathered through GWAS is one piece of the puzzle for elucidating normal physiology and disease mechanisms of autoimmunity. One major hindrance to delineating the functional role of genetic variants as well as in autoimmune research in general lies in the lack of a comparable functional catalog of the human immune system [43]. Much of the current immunological knowledge has come from mouse models. While it has been a valuable source, the mouse immune system only emulates the human system partially. Without clear

definitions and comprehensive descriptions of the normal immune system, it will be difficult for sporadic and focused manipulative assays to capture patterns of functional variance with respect to genetic or environmental changes in such a complex system. Also clinically, diagnosis of autoimmune diseases is largely descriptive with few reliable biomarkers, and it is difficult to design, assess, and improve efficacy of therapeutic agents. It is therefore crucial to establish the phenotypes and functional profiles of the components of the human immune system. Then, the relationship between function and genetics can be systematically assessed, and therapeutic interventions can be designed and modified accordingly.

CONCLUSION

For years the primary bottleneck in understanding autoimmune disease was the inability of genetics to provide genuine, replicable insights – this has been overcome, but simply places the emphasis now on challenges greater than expected, the interpretation of hundreds of contributing genetic risk factors into a coherent and actionable biological insights. Now a large volume of statistical and functional follow-up studies to GWAS results is starting to emerge. But, despite the amount of raw data and follow-up statistical and functional studies, many mechanisms of genetic contribution to autoimmune diseases remain unclear.

It is important to keep in mind that in the effort to ultimately understand disease mechanisms, GWAS and other statistical approaches do not replace *in vitro* or *in vivo* functional studies. Rather, GWASs have proven to serve as a powerful first step to guide and complement functional studies, as they 1) outline relevant and coherent systems and processes involved in disease pathogenesis; 2) suggest specific molecular pathways and appropriate cellular/organismal models for focused functional studies; and 3) offer a promising path to discovering potential leads of targeted therapy. It is particularly encouraging that progress is being made beyond GWAS via the integration of many different types of biological data such as gene expression, immunophenotypes,

epigenetics, protein interactions and clinical data. These efforts are helping to elucidate the complex biology that association studies outline and make further progress towards realizing the promise of human genetics.

REFERENCES

1. Hindorff LA, M.J.E.B.I., Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA., *A Catalog of Published Genome-Wide Association Studies.* .

2. Anderson, C.A., et al., *Synthetic associations are unlikely to account for many common disease genome-wide association signals.* PLoS Biol, 2011. **9**(1): p. e1000580.

3. Gorlatova, N., et al., *Protein characterization of a candidate mechanism SNP for Crohn's disease: the macrophage stimulating protein R689C substitution.* PLoS One, 2011. **6**(11): p. e27269.

4. Guerini, F.R., et al., *A functional variant in ERAP1 predisposes to multiple sclerosis.* PLoS One, 2012. **7**(1): p. e29931.

5. Pidasheva, S., et al., *Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q.* PLoS One, 2011. **6**(10): p. e25038.

6. Liu, Z., et al., *The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease.* Nat Immunol, 2011. **12**(11): p. 1063-70.

7. Okada, Y., et al., *HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease.* Gastroenterology, 2011. **141**(3): p. 864-871 e1-5.

8. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis.* Nat Genet, 2012. **44**(3): p. 291-6.

9. Rossi, F., et al., *The Cannabinoid Receptor type 2 Q63R variant increases the risk of celiac disease: Implication for a novel molecular biomarker and future therapeutic intervention.* Pharmacol Res, 2012. **66**(1): p. 88-94.

10. Tejasvi, T., et al., *TNFAIP3 gene polymorphisms are associated with response to TNF blockade in psoriasis.* J Invest Dermatol, 2012. **132**(3 Pt 1): p. 593-600.

11. Wang, S., et al., *A functional haplotype of UBE2L3 confers risk for systemic lupus erythematosus.* Genes Immun, 2012.

12. Wiede, F., et al., *T cell protein tyrosine phosphatase attenuates T cell signaling to maintain tolerance in mice.* J Clin Invest, 2011. **121**(12): p. 4758-74.

13. Zhou, J., et al., *A20-binding inhibitor of NF-kappaB (ABIN1) controls Toll-like receptor-mediated CCAAT/enhancer-binding protein beta activation and protects from inflammatory disease.* Proc Natl Acad Sci U S A, 2011. **108**(44): p. E998-1006.

14. Stastny, P., *HLA-D and Ia antigens in rheumatoid arthritis and systemic lupus erythematosus.* Arthritis Rheum, 1978. **21**(5 Suppl): p. S139-43.

15. Bakir-Gungor, B. and O.U. Sezerman, *A new methodology to associate SNPs with human diseases according to their pathway related context.* PLoS One, 2011. **6**(10): p. e26277.

16. Chen, M., J. Cho, and H. Zhao, *Incorporating biological pathways via a Markov random field model in genome-wide association studies.* PLoS Genet, 2011. **7**(4): p. e1001353.

17. Cleynen, I., et al., *Genetic evidence supporting the association of protease and protease inhibitor genes with inflammatory bowel disease: a systematic review.* PLoS One, 2011. **6**(9): p. e24106.

18. Nakaoka, H., et al., *A systems genetics approach provides a bridge from discovered genetic variants to biological pathways in rheumatoid arthritis.* PLoS One, 2011. **6**(9): p. e25389.

19. Hu, X., et al., *Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets.* Am J Hum Genet, 2011. **89**(4): p. 496-506.

20. Lee, I., et al., *Prioritizing candidate disease genes by network-based boosting of genome-wide association data.* Genome Res, 2011. **21**(7): p. 1109-21.

21. Bergholdt, R., et al., *Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein-protein interactions, and human pancreatic islet gene expression.* Diabetes, 2012. **61**(4): p. 954-62.

22. Akula, N., et al., *A network-based approach to prioritize results from genome-wide association studies.* PLoS One, 2011. **6**(9): p. e24220.

23. Rossin, E.J., et al., *Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology.* PLoS Genet, 2011. **7**(1): p. e1001273.

24. Fehrmann, R.S., et al., *Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.* PLoS Genet, 2011. **7**(8): p. e1002197.

25. Corona, E., J.T. Dudley, and A.J. Butte, *Extreme evolutionary disparities seen in positive selection across seven complex diseases.* PLoS One, 2010. **5**(8): p. e12236.

26. Eyre, S., et al., *Overlapping genetic susceptibility variants between three autoimmune disorders: rheumatoid arthritis, type 1 diabetes and coeliac disease.* Arthritis Res Ther, 2010. **12**(5): p. R175.

27. Suzuki, A., et al., *Insight from genome-wide association studies in rheumatoid arthritis and multiple sclerosis.* FEBS Lett, 2011. **585**(23): p. 3627-32.

28. Ramos, P.S., et al., *A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap.* PLoS Genet, 2011. **7**(12): p. e1002406.

29. Menon, R. and C. Farina, *Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes.* PLoS One, 2011. **6**(4): p. e18660.

30. Lewis, S.N., et al., *Prediction of disease and phenotype associations from genome-wide association studies.* PLoS One, 2011. **6**(11): p. e27175.

31. Cotsapas, C., et al., *Pervasive sharing of genetic effects in autoimmune disease.* PLoS Genet, 2011. **7**(8): p. e1002254.

32. So, H.C., M. Li, and P.C. Sham, *Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study.* Genet Epidemiol, 2011. **35**(6): p. 447-56.

33. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.

34. Gibson, G., *Hints of hidden heritability in GWAS.* Nat Genet, 2010. **42**(7): p. 558-60.

35. Lee, S.H., et al., *Estimating missing heritability for disease from genome-wide association studies.* Am J Hum Genet, 2011. **88**(3): p. 294-305.

36. Stahl, E.A., et al., *Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis.* Nat Genet, 2012. **44**(5): p. 483-9.

37. Park, J.H., et al., *Estimation of effect size distribution from genome-wide association studies and implications for future discoveries.* Nat Genet, 2010. **42**(7): p. 570-5.

38. Feng, T. and X. Zhu, *Genome-wide searching of rare genetic variants in WTCCC data.* Hum Genet, 2010. **128**(3): p. 269-80.

39. Lehne, B., C.M. Lewis, and T. Schlitt, *Exome localization of complex disease association signals.* BMC Genomics, 2011. **12**: p. 92.

40. Rivas, M.A., et al., *Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.* Nat Genet, 2011. **43**(11): p. 1066-73.

41. Nejentsev, S., et al., *Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.* Science, 2009. **324**(5925): p. 387-9.

42. Trynka, G., et al., *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nat Genet, 2011. **43**(12): p. 1193-201.

43. Blumberg, R.S., et al., *Unraveling the autoimmune translational research process layer by layer.* Nat Med, 2012. **18**(1): p. 35-41.

CHAPTER 2

Genes in Autoimmune Risk Loci Are Specifically Expressed in Critical Immune Cell Types

Integrating Autoimmune Risk Loci with Gene Expression Data Identifies Specific Pathogenic Immune Cell Subsets.

Xinli Hu,[1-4] Hyun Kim,[1,2] Eli Stahl,[1-3] Robert Plenge,[1-3] Mark Daly,[3,6] Soumya Raychaudhuri,[1-3,5,*]

1. Division of Genetics, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

2. Division of Rheumatology, Immunology, and Allergy, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

3. Medical and Population Genetics Group, Broad Institute, Cambridge, MA 02142, USA

4. Health Science and Technology MD Program, Harvard University and Massachusetts Institute of Technology, Boston, MA 02115, USA

5. Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA 02115, USA

6. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

Correspondence: Soumya Raychaudhuri (soumya@broadinstitute.org)

Originally published as:

Hu, X., H. Kim, E. Stahl, R. Plenge, M. Daly and S. Raychaudhuri (2011). "Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets." Am J Hum Genet **89**(4): 496-506.

ABSTRACT

While genome-wide association studies have implicated many individual loci for complex diseases, identifying the exact causal alleles and the cell-types within which they act remains greatly challenging. To ultimately understand disease mechanism, carefully conceived functional studies in relevant pathogenic cell-types will be necessary to demonstrate the cellular impact of disease-associated genetic variants. This challenge is highlighted in autoimmune diseases, such as rheumatoid arthritis, where any of a broad range of immunological cell-types might potentially be impacted by genetic variation to cause disease. To this end, we developed a statistical approach to identify potentially pathogenic cell-types in autoimmune diseases using a gene expression data set of 223 sorted murine immune cells from the Immunological Genome Consortium. We found enrichment of *Transitional B-cell* genes in systemic lupus erythematosus (p=5.9x10^{-6}) and *Epithelial-associated Stimulated Dendritic Cells* genes in Crohn's disease (p=1.6x10^{-5}). Finally, we demonstrated enhancement *of CD4+ Effector Memory T-cell* genes within rheumatoid arthritis loci (p<10^{-6}). To further validate the role of CD4+ effector memory T-cells within rheumatoid arthritis, we identified 436 loci not yet known to be associated with disease but with statistically suggestive association in a recent GWAS meta-analysis (p_{GWAS}<0.001). Even among these putative loci, we noted a significant enrichment for CD4+ effector memory T-cell gene expression (p=1.25x10^{-4}). These cell-types are primary candidates for future functional studies to reveal the role of risk alleles in autoimmunity. Our approach has application in other phenotypes, outside of autoimmunity, where many loci have been discovered and high-quality cell-type specific gene expression is available.

INTRODUCTION

Autoimmune diseases are complex traits with many scores of common variants throughout the genome that might subtly impact disease risk[1-4]. But, using these loci to elucidate mechanisms from common variants has proven to be a challenging task, particularly since many of them do not directly alter coding sequences, but potentially impact gene regulation modestly in a cell-specific manner[5]. If the critical immune cell subsets were known for a given disease, then investigators could derive relevant cellular model systems for focused functional studies to understand pathogenic mechanisms. These studies might include broad genomics approaches, such as cell-type specific expression quantitative trait loci (eQTL) screens to identify alleles that act to alter gene expression[6-8], or epigenetic screens to identify key active regulatory elements[9, 10]. Additionally, investigators could pursue focused mechanistic studies to understand the role of individual disease alleles within that tissue.

But for most autoimmune diseases the immune cell-types specifically impacted by common risk variants are not defined. Past mechanistic studies in autoimmune model systems have often led to confusing results that may not easily translate to human disease. For example, separate influential studies in rheumatoid arthritis (RA [MIM 180300]) have implicated a wide range of pathogenic cell-types including B and T lymphocyte subsets[11], neutrophils[12], mast cells[13], macrophages[14], platelets[15], and synoviocytes[16, 17]. The importance of pursing mechanistic studies in the appropriate cell-type is highlighted by the fact that common variants can have conflicting functions in different closely related immune tissues. For example a deletion of the promoter region of *IRGM*, associated with Crohn's disease (MIM 266600), might either increase or decrease allelic gene expression depending on the tissue[18]. Similarly an *IL2RA* autoimmune variant impacts different intermediate phenotypes, even in closely related immune cells[19].

Here, we hypothesize that predisposing autoimmune risk alleles impact a small number of pathogenic tissues or cell-types. If this is the case, then the subset of genes with critical functions in

those pathogenic cell-types are likely to be within disease loci. However, in practice, a comprehensive and unbiased catalog of cell-type specific gene function is simply not available. As an alternative, compendia of gene expression data are available for many tissues. These compendia can serve as objective proxies for tissue-specific gene function. Practically, gene expression profiles have been used to identify cell-types of origin in malignancies[20, 21]. In addition, investigators commonly use gene expression profiling of presumed pathogenic tissues to screen risk alleles and to prioritize genes for followup within complex trait loci. An orthogonal approach is to broadly consider a large collection of potential cell-types and to then identify the single tissue that specifically expresses genes within loci that contain disease risk alleles. To our knowledge, no such systematic approach has yet been devised.

We developed a statistical method that, given a collection of disease-associated SNPs and a compendium of gene expression profiles from a broad set of tissue types, scores tissues for enrichment of specifically expressed genes in LD with the SNPs (see **Figure 2.1** and **Methods** for details). For such a method to be effective, it is critical to use high quality cell-specific expression data with minimal contamination and including replicates to reduce noise. To this end we use the Immunological Genome Project (ImmGen) data set assaying 223 mouse immune tissues individually double-sorted by FACS to ensure high purity and profiled in at least triplicate[22]. Also, it is critical for the methodology to be robust to key confounders. Therefore, in our method we (1) use non-parametric expression specificity scores to avoid confounding by the inherently skewed nature of expression levels, (2) correct for number of genes per SNP to avoid multiple-hypothesis testing biases, and (3) assess significance of disease associated SNP sets using matched SNP sets to avoid confounding by correlations in gene size and cell-specific expression[23], correlations in expression between proximate genes, and genomic biases in gene density and genetic variation across the genome.

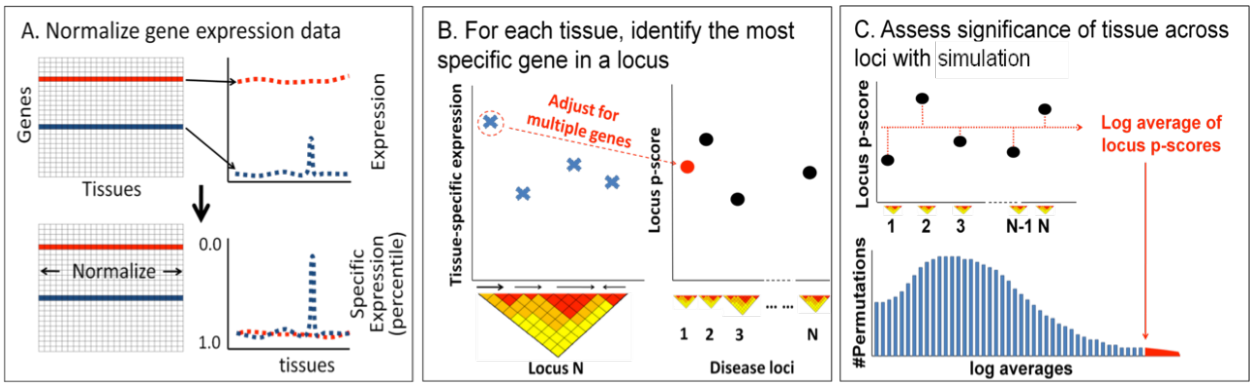


Figure 2.1. A) *Normalize gene expression data.* We normalized the expression profile by dividing the expression value of each gene in each tissue by the Euclidean norm of the gene's expression across all tissues in order to emphasize tissue specificity of each gene's expression. Scores were converted to non-parametric percentiles. B) *For each tissue, identify the most specific gene in a locus.* For each SNP, we first defined genes implicated by the SNP based on LD. For a specific tissue, we identified the most specifically expressed gene in the locus and then scored the SNP based on that gene's nonparametric specific expression score after adjusting for multiple genes within the locus. C) *Assess significance of tissue across loci with permutations.* We first calculated a score for the tissue by taking the average of the log adjusted-percentiles across loci from B. Then, we randomly selected matched SNP sets and score them similarly. The proportion of random SNP sets with tissue scores exceeding that of the actual set of SNPs being tested was reported as the p-value of the tissue.

METHODS

Summary of Statistical Method

First, after standard quality control and quantile-normalization of expression data[24], we transform expression values into non-parametric "tissue-specific expression" scores for each gene (**Figure 2.1A**). In order to do this, we first divide raw expression values by the Euclidean norm of values for each gene across all tissue types. Then, for a given tissue, we order genes by normalized expression values and assign each gene a percentile. These uniformly distributed percentiles

constitute non-parametric tissue-specific expression scores. Genes with low percentile scores for a tissue are highly specifically expressed in that tissue while genes with high percentile scores are either not expressed in that tissue at all or ubiquitously expressed. Second, for a given tissue, we assign each disease-associated SNP a "locus p-score" (**Figure 2.1B**). To do this, we first identify genes that are in LD with a disease-associated SNP by using standard methods[25, 26]. Then, we identify the single proximate gene most specifically expressed within that tissue. The SNP's locus p-score for that tissue is determined to be the tissue-specific percentile score of that gene after correcting for multiple genes tested within that locus. These locus p-scores should be roughly uniformly distributed under the null. Finally, we assign an overall significance score for each tissue by taking the average of the log of locus p-scores across all disease-associated SNPs (see **Figure 2.1C**). While an analytical p-value can be calculated, to avoid realistic confounders that may inadvertently inflate theoretical p-values, we calculate statistical significance scores by comparing the actual log average of locus p-scores to that of random SNPs matched for total number of genes.

Gene Expression Datasets

For this project we used two separate data sets. The Genomics Institute of the Novartis Research Foundation (GNF) tissue atlas is a set of human expression profiles of 79 human tissues and cells types including immune cell types measured in duplicate (http://biogps.gnf.org/)[27]. The Immunological Genome Project (ImmGen) consists of gene expression profiles of 223 immunological tissue/blood-sorted cell types obtained from mice[22]. Each sample is sorted with at least three biological replicates.

Preprocessing and Normalizing Gene Expression Datasets

For each dataset, after applying standard quantile normalization[24], we averaged expression values from replicates for each probe set. To obtain the single most robust expression value for genes with multiple probe sets, we selected the single probe set within each gene transcript that had the highest minimal expression value across all tissues. The GNF dataset then consisted of measurements on 17,581 unique genes in 79 tissue types. The ImmGen dataset contained 21,968 unique Mus musculus genes. We used HomoloGene (March 2010) to map the Mus musculus genes to 14,623 unique human homologs.

We then transformed both datasets into non-parametric tissue-specific expression scores for genes. First, we normalized expression level of each gene to reflect the specificity of expression in each tissue type. To do so, for each gene in each tissue, we divided the raw expression value by the Euclidean norm of values across all tissues:

$$X'_{i,j} = X_{i,j}/\text{norm}(X_{i,})$$

where $X_{i,j}$ is the expression value of gene *i* at tissue *j*, and $X'_{i,j}$ is the specificity score. Thus, each gene and tissue received a score between 0 and 1 where a score of 1 means the gene is exclusively expressed in this tissue. Ubiquitously expressed genes have low normalized scores across tissue types.

Next, for a given tissue, we transformed these normalized scores, $X'_{i,j}$, into non-parametric tissue-specificity percentile scores for each gene, $P_{i,j}$, where a low percentile represents high specificity relative to other genes for a given tissue and a high percentile represents low specificity.

Mapping SNPs to genes

Disease-associated SNPs are linked to proximate genes in LD with them, using a previously described approach[25, 26]. First, for each SNP, we defined genes implicated by the SNP by defining a disease region. To do so, we identified the furthest neighboring SNPs in LD with the SNP in the 3'

and 5' directions (r^2>0.5, CEU HapMap). We then extended outward in each direction to the nearest recombination hotspot[28]. This region would include the disease-associated SNP and all SNPs in LD. All genes that overlapped with this region were considered implicated by the SNP. If no genes were found in the region, we extended an additional 250kb in each direction. If two SNPs contained overlapping genes, they were merged as one single locus.

Testing tissue for enrichment

Given our list of SNPs connected to genes and our non-parametric expression tissue-specificity percentiles, we scored the list of disease-associated SNPs for enrichment of genes specifically expressed in each individual tissue type.

To score each tissue *j*, we first identified the most specifically expressed gene near each SNP *S* in tissue *j*. We will refer to that gene as $g_{S,j}$. We applied a Bonferroni correction to adjust the tissue-specificity percentile for testing of the multiple genes near each SNP:

$$P_{S,j} = 1 - (1 - P_{g_S,j})^{n_s}$$

where n_s is the number of genes implicated by SNP *S*. The $P_{S,j}$ values are referred to in the main text as the "locus *p*-score". They should be roughly uniformly distributed. For each tissue, we scored for enrichment by summing the $P_{S,j}$ values of all SNPs:

$$T_j = -\sum_{S \in all\ SNPs} \log\left(P_{S,j}\right)$$

Under the null, if $P_{S,j}$ scores were randomly distributed, then T_j should be distributed according to the gamma distribution:

$$D \sim \Gamma(\alpha, \beta)$$

where α is the shape parameter and is equal to the number of SNPs and β is the rate parameter and is set to 1. In this case, the p-value for the tissue is calculated as:

$$p(D \leq T_j, D \sim \Gamma(1, N_{SNPs}))$$

However, analytical p-values are not robust to realistic biological factors.

Significance scores are based on random SNP sets.

To estimate the significance in a more robust and unbiased manner, we calculated *p*-values empirically by comparing observed T_j values to empirical values from random sets of SNPs. Given a set of disease-associated SNPs, we create a matched SNP set with exactly the same number of SNPs and approximately similar numbers of genes for each permutation. We drew random SNPs for permutation from a pool of 45,265 independent Hapmap SNPs that were "clumped" to insure minimal correlation[29]. To create a matched SNP set with approximately similar gene numbers, for each disease-associated SNP that implicated <11 genes we selected a random SNP that implicated exactly the same number of genes and for SNPs that implicated >10 genes, we selected a random SNP that also implicated >10 genes. To ensure a comparable number of genes, the total number of genes implicated by all random SNPs must be within 10% of that implicated by disease SNPs. We then scored each of matched SNP sets for enrichment of genes in tissue *j* and calculated T_j. The proportion of randomly selected matched SNP sets whose T_j is less than the T_j for the disease-associated SNPs set was reported as the p-value.

In practice, we varied the number of random SNP sets that we evaluated for a tissue from 250 to 1,024,000. We started by evaluating each tissue with 250 SNP sets. For those tissues where at least 25 sets were observed to be more significant than the observed SNP set, we accepted the p-value and did not evaluate for any more SNP sets. For those tissues for which fewer than 25 sets

were more significant than the observed SNP set (ie $p<0.1$), we doubled the number of SNP sets. The number of SNP sets was doubled until at least 25 events were more significant than the observed SNP set or until we reached 1,024,000 permutations. This insured a variance of <20% of the reported p-value for p-values $>2.5x10^{-5}$.

Assessing the significance of individual SNPs

For each SNP and tissue, we calculated an "empirical locus *p*-value", which assessed the degree to which an individual gene within a locus is contributing to enriched specific gene expression within a tissue. This value was calculated by comparing the locus *p*-score for the actual disease-associated SNP, based on the most specifically expressed gene within a tissue, to that of the matched SNP in randomly selected SNP sets during the permutation process, as described above. The empirical locus *p*-value was reported as the fraction of randomly selected matched SNPs with more extreme locus *p*-scores than the actual locus *p*-score.

Adjusting for Expression Profiles

In order to assess enrichment across tissues after accounting for the effect of tissues that have already been identified as significant from the dataset, we have devised an adjusted analysis framework. Briefly, we used the X' matrix of tissue specificity scores, then removed the component of each tissue expression profile that was correlated with the tissue that we are conditioning on.

Let the expression scores of the most significant tissue be vector *v*. We subtracted the components of v from another tissue's expression profile, *u*, in order to obtain a new profile u' which is independent from *v*:

$$u' = u - v * \sum \frac{u}{|u|} * \frac{v}{|v|}$$

The new profile scores were used to recalculate tissue-specific percentiles *P*, which can then be reused with the same statistical framework as above.

Scoring nominally associated RA SNPs

In order to score RA SNPs not yet associated with RA, we used the p-value results from a recently published meta-analysis of six GWAS consisting of 5,539 autoantibody positive RA cases and 20,169 controls of European descent. We selected all SNPs that had an association *p*-value of <0.001. After excluding SNPs within the MHC region (ranging from 25.8-3.4 MB on chromosome 6 in HG 18 coordinates), we grouped the resulting SNPs into independent loci. We grouped two SNPs within the same locus if they had r^2>0.1 in HapMap or shared a common gene. For each locus, we selected the single SNP with the most significant association to RA. We excluded any of these SNPs that were in LD with a known RA-associated risk loci (r^2<0.1) or implicated a gene that was also implicated by a known RA SNP. We tested these loci for enrichment of specifically expressed genes in each of the individual cell types in RA. Significance for each tissue was determined by selecting matched SNP sets as described above. Given the large number of SNPs, we allowed for the total gene number to be outside the +/- 10% criteria described above.

In order to calculate an overall association to CD4+ effector memory T-cell association, we averaged all four X' specificity score profiles of each of the CD4+ effector memory T-cell subsets together to calculate significance of association and empirical locus *p*-values.

RESULTS AND DISCUSSION

We wanted to ensure that out statistical method was robust to realistic biological factors (e.g. neural tissues tend to express larger genes[23]), which can inadvertently inflate theoretical p-

values in certain cell-types (see **Figure 2.2**). Thus, we scored 10,000 sets of 20 random SNPs, each in LD with at least one gene, from a larger set of independent SNPs from the HapMap project. Applying our approach to assess gene expression enrichment in both the 79 tissues from the Genomics Institute of the Novartis Research Foundation data (GNF) and to the 223 cell-types from the ImmGen demonstrate appropriate type I error rate (see **Figure 2.3A and 2.3B**). We also note that error rates are consistent across all cell-types, with no evidence of inflation of significance scores at any given tissue. Furthermore, our method demonstrates little evidence of statistical inflation in 500 sets of 20 random SNPs in either of those two data sets (see **Figure 2.3C**,**2.3D**).

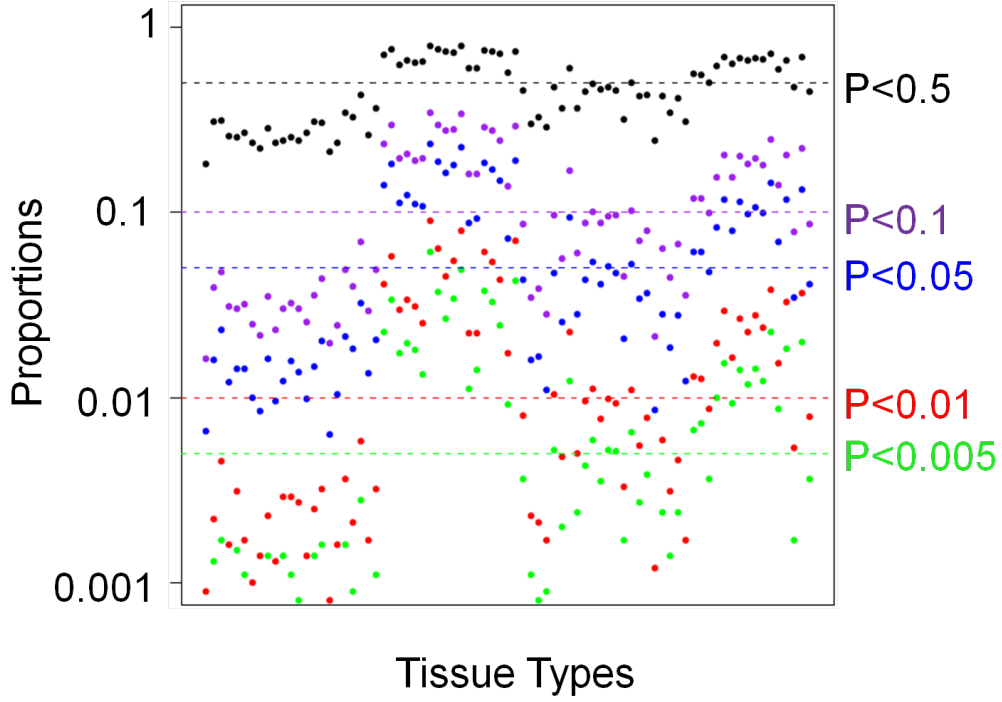
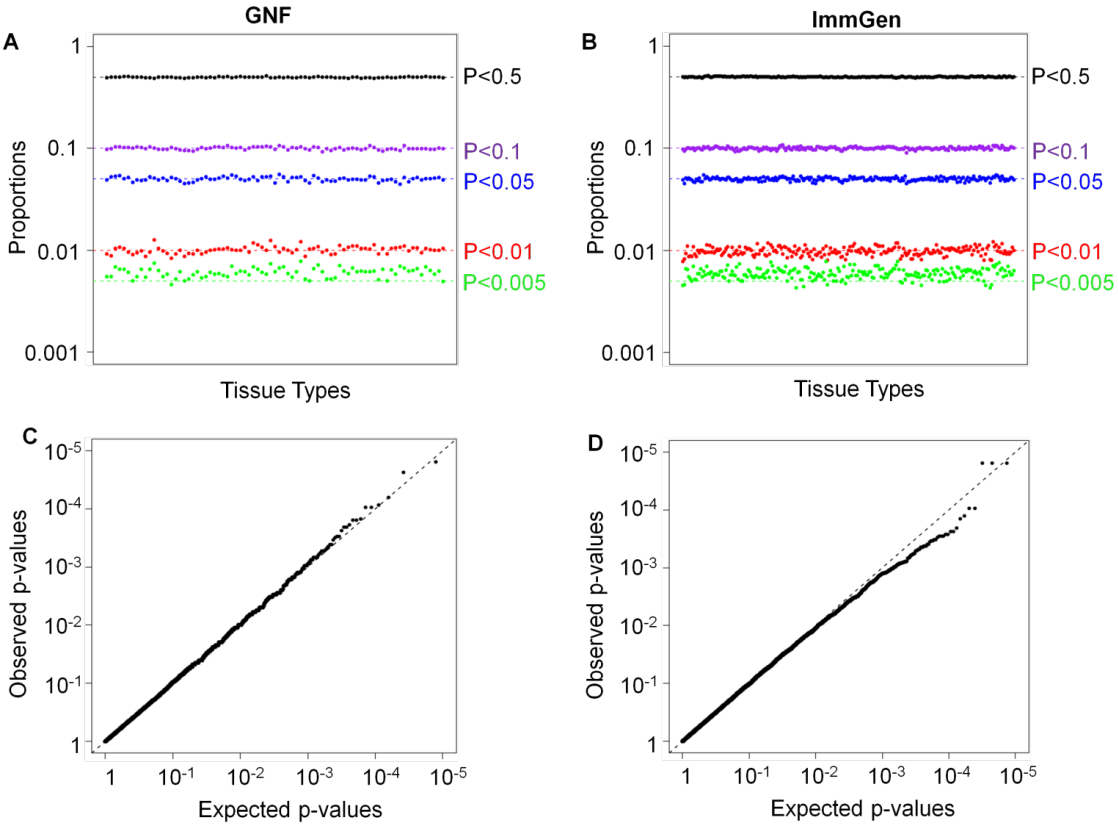


Figure 2.2. Analytical p-values, calculated without permutations show inflation in neural tissues. We evaluated 10,000 sets of 20 random independent SNPs across the genome for tissue specific gene enrichment using the GNF human tissue expression data set. The subset of tissues enriched in the middle demonstrating marked inflation are all central or peripheral nervous system tissues. Here we present analytical p-values using the gamma distribution (see **Supplementary Methods**).

Figure 2.3 Permutation scheme produces appropriate type I error. A) To test the statistical properties of our approach, we selected 10,000 random SNP sets of 20 independent SNPs across the genome. We tested 79 tissues expression profiles from the GNF dataset for enrichment of specifically expressed genes. For each tissue type, we plotted the proportion of sets that obtained specific p-value thresholds. B) Similar results for the 223 tissue profiles in the ImmGen dataset. C) We tested 500 random SNP sets and tested 79 tissues expression profiles from the GNF dataset for enrichment. After aggregating p-values for all tissues, we plotted observed p-values as a function of expected p-values in a Q-Q plot. D) Similar results for the ImmGen dataset.

Figure 2.3 Permutation scheme produces appropriate type I error (Continued).

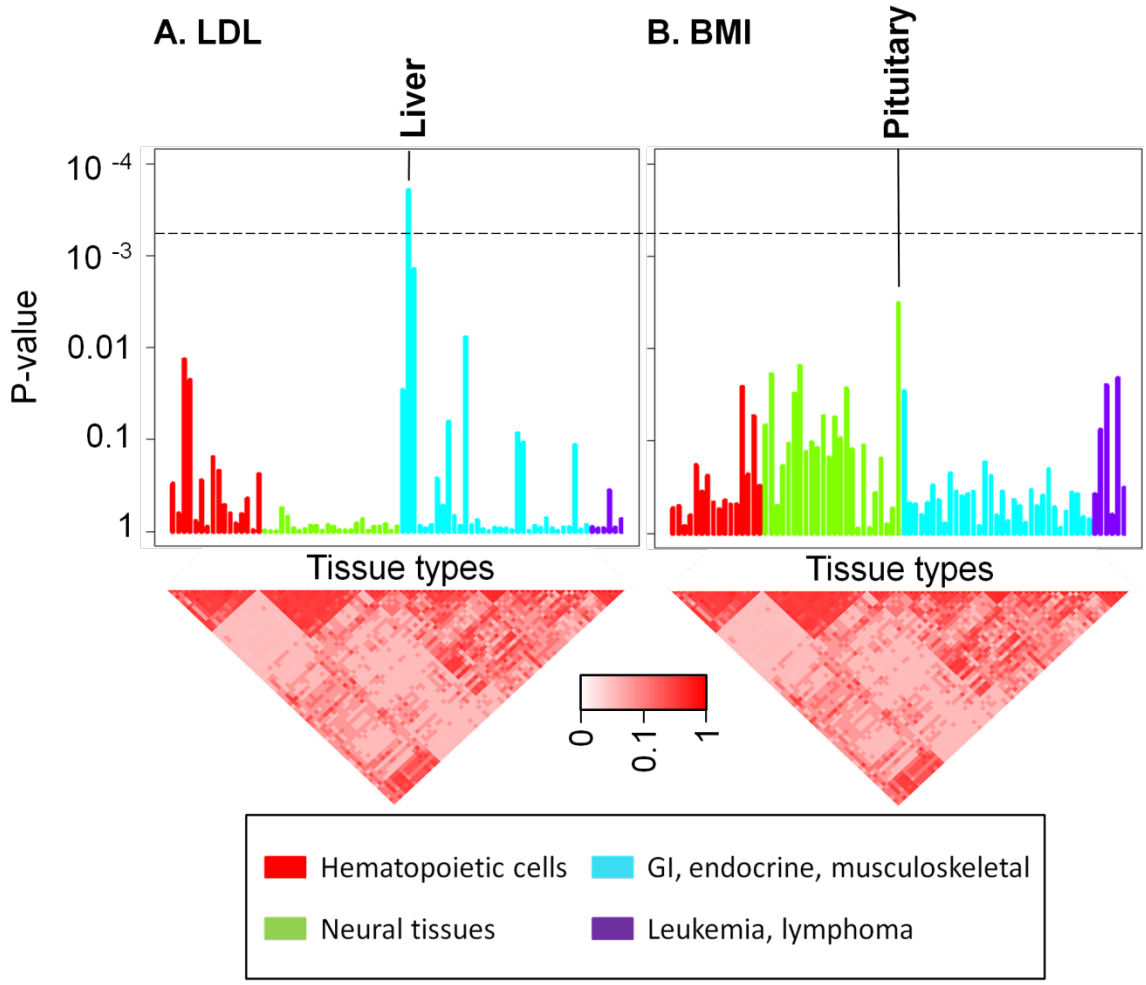


As a positive control, we examined common variants from two phenotypes. First, we applied our method to 37 SNPs associated with serum LDL cholesterol from a recent large genome-wide SNP association meta-analysis[30]. We hypothesized that these genes would be most specifically enriched in the liver, since the liver is the primary organ where LDL is regulated[31] and known mutations impact hepatocyte cellular function[32, 33]. In aggregate, these SNPs implicated 165 genes in LD. When we tested each of the 79 tissue expression profiles from GNF for specific expression of genes in LD with these SNPs, we did indeed observe that only the liver showed highly specific expression of genes in LD with cholesterol metabolism SNPs (p=2.0x10^{-4}, see **Figure 2.4A**). Other tissues that obtained nominal significance at p<0.01, fetal liver (p=0.0014) and the adipocyte (p=0.0077), were no long significant after adjusting for the liver expression profile. This suggests that the other observed associations were the consequence of correlated expression (see **Methods**). In certain cases, loci harbored genes that were specifically expressed within the liver, and in these cases, these genes were often compelling candidate genes. Next, we applied our method to the 32 obesity associated SNPs[34], that in aggregate implicate 91 genes. When we tested 79 tissues from the GNF for specific expression of genes within obesity loci, we observed that only the pituitary gland obtained nominal significance at p=0.0032 (see **Figure 2.4B**). While this was not statistically significant after accounting for 79 independent tests, we were encouraged that it emerged as the most significant tissue, since pituitary dysfunction, from trauma or rare familial mutations, in a known cause of obesity[35, 36]. Furthermore, authors of recent genome-wide genetic studies have speculated that obesity SNPs act on the hypothalamus-pituitary axis[34, 37]. Potentially, a more targeted expression data set of the brain with carefully dissected human tissues might have resulted in a more powerful analysis.

While there is concern that multiple inter-correlated gene expression profiles might compromise power, we found that even in extreme circumstances that the power loss is minimal.

Figure 2.4. Cell-specific gene expression enrichment in metabolic diseases. A) 37 SNPs associated with LDL metabolism were evaluated for gene enrichment in 79 human tissue types. The Bonferroni-corrected *p*-value is shown by a dotted line. Only the liver showed statistically significant specific expression of genes in LD with cholesterol metabolism SNPs (p= 1.95×10^{-4}). We have plotted a heat map along the bottom to depict the *p*-value correlation between tissue types among random SNP sets. B) 32 SNPs associated with obesity were evaluated for gene enrichment in 79 human tissue types. The pituitary achieved the most significant *p*-value (p=3.25×10^{-3}).

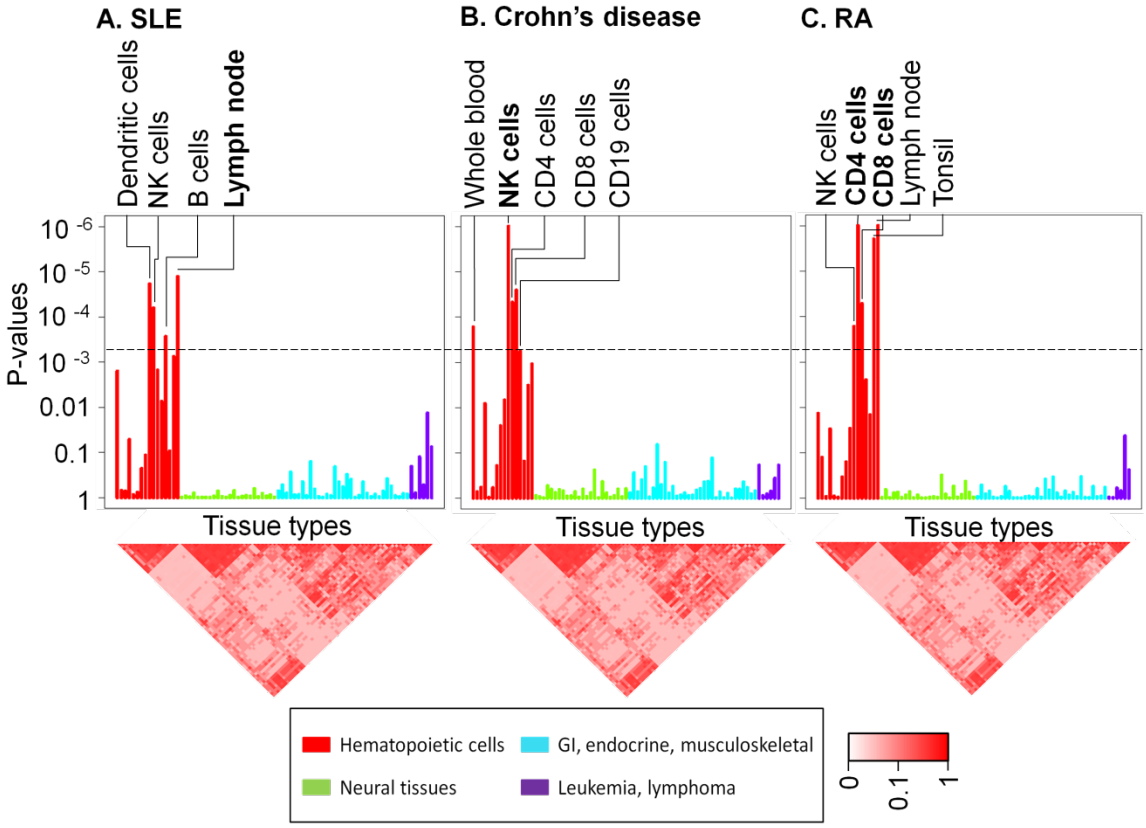
Figure 2.4. Cell-specific gene expression enrichment in metabolic diseases (Continued).



Our non-parametric approach relies on the ordering of a gene's specific expression within a tissue, rather than its magnitude of the specificity. The addition of inter-correlated tissue profiles impact the magnitude of specific expression scores for a tissue, but has a minimal impact on the ordering of the genes themselves. To assess the robustness of our method to multiple inter-correlated expression profiles, we repeated our analysis of LDL SNPs after adding in 1, 10, 50, and 100, copies of the identical liver expression profile. In each case the liver showed the exact same highly specific expression of genes in LD with cholesterol metabolism SNPs (p=2.0x10^{-4}). As a second test we added 1, 10, 50, and 100 copies of liver expression profiles, where we permuted expression values of 50% genes independently; so that each added profile was correlated with but also had substantial differences from the original liver expression profile. In each case the liver showed highly specific expression of genes in LD with cholesterol metabolism SNPs (p ranging from 9.8x10^{-5} to 2.9x10^{-4}). In instances where the correlational structure of the data is more complex, and power is impacted, dimensional reduction approaches to simplify the expression data is useful[38]. Convinced that this approach was statistically robust and could detect potentially pathogenic cell-types, we applied it to autoimmune disease SNPs. We focused on three separate autoimmune diseases. For systemic lupus erythematosus (SLE [MIM 152700]), we identified 30 SNPs, implicating 27 independent loci with a total of 136 genes [39-43]. For Crohn's disease, we identified 71 SNPs, implicating 69 independent loci with a total of 316 genes [3]. Finally, for RA, we identified 40 SNPs, in aggregate implicating 39 independent loci with a total of 132 genes [44, 45].
Testing each of these three autoimmune disease SNP sets against the 79 GNF tissues implicated only immune tissues (in each case multiple tissues with p<2x10^{-5}, **Figure 2.5**). But given the limited number of immunological tissues and the high degree of correlation between them, we could not pinpoint the causal immune cell-types. We speculated that the ImmGen dataset could more clearly demonstrate the key immune cell-types for each of the different diseases since it was collected to represent a very broad view of transcriptional profiles in mouse immune cell-types across many

Figure 2.5. SNPs associated to SLE, Crohn's Disease, and RA evaluated for cell-specific gene enrichment in 79 human tissue types. Subsets of hematopoietic cells showed statistically significant enrichment in each of the three diseases, while none of the other tissues showed significant enrichment. A) In SLE, B-cells, dendritic cells, NK cells, and lymph node tissue showed significant enrichment p-values after adjusting for multiple hypothesis testing. B) NK cells, CD8+ and CD4+ T-cells, whole blood, as well as CD19+ cells achieved statistical significance in Crohn's Disease. C) In RA CD4+ T-cells, CD8+ T-cells, tonsil, lymph node, and NK cells showed significant enrichment p-values after adjusting for multiple hypothesis testing.

Figure 2.5. SNPs associated to SLE, Crohn's Disease, and RA evaluated for cell-specific gene enrichment in 79 human tissue types (Continued).

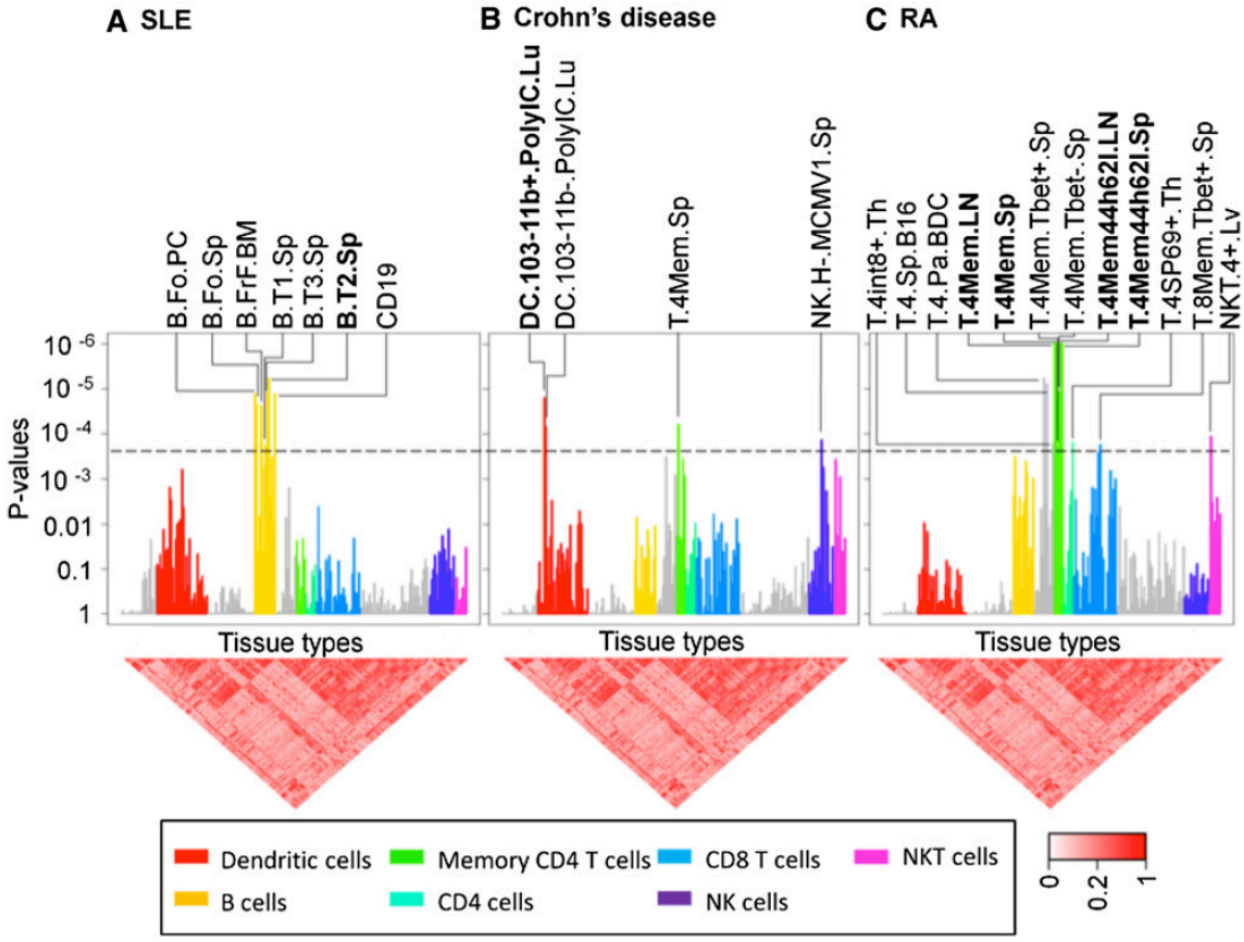


lineages, developmental stages, and target organs. It includes hematopoietic stem, myeloid and lymphoid cells, and both innate and adaptive immune cells.

When we tested SLE loci for enrichment of specifically expressed genes within the 223 expression profiles contained within ImmGen data set, the single most significant immune cell-type was transitional B-cells (stage T3) collected from the spleen (p=5.9x10^{-6}, see **Table 2.1, Figure 2.6A).** Strikingly, all of the other statistically significant associations were other B-cell subsets, including other closely related splenic transitional B-cell subsets (p<2x10^{-4} =0.05/223). All of the B-cell associations are obviated (see **Table 2.1**), when we repeated our analysis after adjusting for three splenic transitional B cell profiles (B.T1.Sp , B.T2.Sp, B.T3.Sp). This strongly suggests that other observed B-cell associations are the result of expression correlation with transitional B cells, and not representative of independent effects. The implication of transitional B-cells by associated loci is consistent with much of the known pathobiology of SLE, which has implicated B-cells more broadly. The pathogenic nature of antibodies produced by B-cells in lupus has been long established and is supported by mouse models[46], and by the demonstration of the efficacious nature of B-cell targeted therapies in SLE[47]. These results implicating transitional B-cells specifically offer a finer resolution on this commonly accepted hypothesis.

Figure 2.6 Cell-specific gene expression enrichment in autoimmune diseases. We evaluated SNPs associated with systemic lupus erythematous, rheumatoid arthritis, and Crohn's disease for cell-specific gene enrichment in 223 murine immune cell types. The Bonferroni-corrected p-value is shown by a dotted line. In each case, we listed cell types that are significant after multiple hypothesis testing ($p<2.2x10^{-4}$), and we bold the single most significant cell type. A) In lupus, B-cells, especially transitional B-cells in the spleen (B.T2.Sp), showed significant enrichment of genes within disease loci. B) In Crohn's disease, epithelial-associated stimulated CD103- dendritic cells (DC.103-11b+PoluIC.Lu) achieved the highest statistical significance. C) In rheumatoid arthritis, the four CD4+ effector memory T-cell subsets in both the spleen and lymph nodes (T.4Mem.LN, T.4Mem.Sp, T.4Mem44h62l.LN, and T.4Mem44h62l.Sp) showed the most significant gene enrichment ($p<10^{-6}$).

Figure 2.6 Cell-specific gene expression enrichment in autoimmune diseases (Continued).



Some of the most significant loci may harbor compelling candidate genes. For example, the rs13385731 locus (empirical locus p=0.0017 for stage T3 transitional B-cells) harbors the RAS pathway gene, *RASGRP3*, which has been shown to potentially play a role in downstream signaling from the B-cell receptor[48]. In other cases, we are able to identify specifically expressed genes that are not yet well characterized but might warrant further examination. For example, the rs6445975 SNP locus (p=1.6x10^{-5}) contains *PXK*, encoding a transcription factor, whose role in immunology is not yet well characterized but is highly and specifically expressed in transitional B-cells.

When we tested Crohn's loci for enrichment of specifically expressed genes within the 223 cell-types of the ImmGen data set, the single most significant cell-type was an epithelial associated stimulated dendritic cell subset (lung CD11b+ dendritic cells stimulated by polyinosinic:polycytydylic acid**,** p=1.6x10^{-5} , see **Table 2.1, Figure 2.6B**). In Crohn's disease, other cell types, including a single CD4+ memory T-cell subset and natural killer cell subset demonstrate statistical significance after multiple hypothesis testing. Moreover, these effects are independent of the DC effect, as their signals are maintained after adjusting for dendritic cell contributions (see **Table 2.1**). Dendritic cells in the intestinal mucosa play a key role in mediating the intestinal inflammation associated with Crohn's disease and have long been thought of as key mediators of disease activity[49, 50]. For example, *NOD2* Crohn's disease risk variants have been shown to disrupt autophagy in dendritic cells[51]. The potential role of dendritic cells has been further highlighted in a mouse model where defective TGF-beta activation can result in spontaneous colitis[52].

When we tested RA loci for enrichment of specifically expressed genes within the 223 cell-types of the ImmGen data set, we observed that each of the four CD4+ effector memory cell subsets emerge as the most highly significant subset (p<10^{-6} for all four CD4+ effector memory cell subsets; see **Table 2.1, Figure 2.6C**). Strikingly, most of the other cell-types achieving statistically significant

Table 2.1. Summary of autoimmune disease loci cell-specific expression enrichment in 22 ImmGen cell types. Here, we list all 22 out of 223 cell types from ImmGen that obtained nominally significant association (p<0.01) in at least one of the three autoimmune phenotypes tested. For each disease, we listed results of our analysis without any conditioning, as well as results after removing the contributions of the most significant tissues. For each tissue, we listed an association significance p-value for each phenotype.

	SLE		Crohn's Disease		Rheumatoid Arthritis	
Cell type	Unconditional	Adjusting for Trans. B subtypes	Unconditional	Adjusting for DC Subtypes	Unconditional	Adjusting for CD4+ TEM cell subtypes
B.T1.Sp	0.00018	-	0.06	0.12	0.0032	0.005
B.T2.Sp	0.000013	-	0.013	0.029	0.00041	0.0015
B.T3.Sp	0.0000059	-	0.03	0.036	0.00072	0.0055
B.Fo.PC	0.000013	0.23	0.06	0.04	0.0024	0.001
B.Fo.Sp	0.000022	0.11	0.0071	0.014	0.00032	0.0015
B.FrF.BM	0.000023	0.82	0.031	0.021	0.0025	0.0025
CD19Control	0.000013	0.7	0.011	0.026	0.00099	0.0025
T.4Mem.LN	0.024	0.18	0.018	0.012	<0.0000010	-
T.4Mem.Sp	0.08	0.27	0.000061	0.00075	<0.0000010	-
T.4Mem44h62l.LN	0.021	0.05	0.00037	0.009	<0.0000010	-

Table 2.1. Summary of autoimmune disease loci cell-specific expression enrichment in 22 ImmGen cell types (Continued).

Cell type	SLE		Crohn's Disease		Rheumatoid Arthritis	
	Unconditional	Adjusting for Trans. B subtypes	Unconditional	Adjusting for DC Subtypes	Unconditional	Adjusting for CD4+ TEM cell subtypes
T.4Mem44h62l.Sp	0.062	0.14	0.00087	0.032	<0.0000010	-
T.4.Pa.BDC	0.0073	0.019	0.00032	0.00025	0.0000059	0.017
T.4.Sp.B16	0.0016	0.0045	0.01	0.009	0.0000078	0.0085
T.4int8+.Th	0.05	0.33	0.00083	0.0015	0.00017	0.18
T.4Mem.Tbet..Sp	0.61	0.75	0.02	0.043	0.00021	0.7
T.4Mem.Tbet+.Sp	0.27	0.43	0.023	0.068	0.000012	0.46
T.4SP69+.Th	0.08	0.23	0.0094	0.013	0.00016	0.17
T.8Mem.Tbet+.Sp	0.27	0.58	0.041	0.1	0.00018	0.64
DC.103-11b+.PolyIC.Lu	0.013	0.026	0.000016	-	0.0094	0.021
DC.103+11b-.PolyIC.Lu	0.043	0.062	0.000069	-	0.13	0.09
NKT.4+.Lv	0.16	0.29	0.00038	0.019	0.00012	0.01
NK.H-.MCMV1.Sp	0.018	0.0061	0.00014	0.0073	0.076	0.33

association (but at a more modest level) are closely related CD4+ T-cell subsets. Adjusting for the four CD4+ effector memory T-cell profiles obviates the significance of all of these cell types (see **Table 2.1**) strongly suggesting that the associations found in these other T-cell subsets are due to their high correlation in expression with CD4+ effector memory T-cells.

Certain SNPs containing highly specifically expressed genes in CD4+ effector memory cells were particularly significant. In many cases these SNPs pointed to well-described candidate genes known to play key roles broadly in CD4+ T-cell biology. As examples, we note multiple genes that are specifically expressed in CD4+ effector memory cells: *PTPN22* (rs2476601, empirical locus p=0.056 for CD4+ memory T-cells), *CD2* (rs11586238, p=0.040)*, PTPRC* (rs10919563, p=0.043), *CD28* (rs1980422 p=0.0045), *IL2RA* (rs2104286, p=0.0010), and *CTLA4* (rs3087243, p=0.0010). The rs2104286 SNP has already been shown to correlate with surface expression of the protein product of *IL2RA* in CD4+ memory T-cells[19] and likely has CD4+ effector memory T-cell function. However, in at least one instance, we identified a candidate gene that has not been specifically connected to T-cell function. For example, the *ANKRD55* (rs6859219, p=0.017) has currently unknown biological function with respect to the immune system, but is highly and specifically expressed in CD4+ effector memory cells.

To assess whether results were influenced by loci that overlap multiple diseases, we repeated our analyses for all three diseases excluding those loci that are implicated in more than one disease. This decreased the number of loci per disease substantially; the number of loci was reduced in SLE to 11 (from 27), in Crohn's to 56 (from 69), and in RA to 23 (from 39). However, the pattern of tissue specific enrichment was not altered.

In order to independently validate the role of CD4+ effector memory T-cells in RA we examined a second set of loci that were nominally associated to RA but not yet considered validated risk loci. Using a polygenic modeling approach, we have separately demonstrated that SNPs with nominal significance at a threshold of p_{GWAS}<0.001 in the latest RA GWAS meta-analysis are

significantly associated with RA risk in aggregate in independent validation samples (Stahl et al *in review*). We estimated that 5-15% of the SNPs that define this polygenic signal represented true RA risk alleles (see Stahl *et al* Supplementary Table 2 for estimates) while the majority (>85%) of them represented statistical fluctuation. We hypothesized that if these SNPs were indeed enriched for true RA risk loci and if our result that CD4+ effector memory T-cells are important for RA holds true, then the nominally associated SNPs should also be modestly enriched for genes specifically expressed in CD4+ effector memory T-cells.

To test this hypothesis we obtained the latest results of an RA GWAS meta-analysis and identified all SNPs with $p_{GWAS}<0.001$. To ensure independence, we combined SNPs in LD ($r^2>0.1$) into individual loci and for each locus, we picked the single most significant SNP for each locus. To ensure that our results were independent of previously known RA loci, we removed all loci in LD with ($r^2>0.1$) or sharing implicated genes with a known RA risk locus. In aggregate, we obtained a total of 436 loci implicating 1,037 genes.

The most significant cell-type was a subcutaneous lymph nodes CD4+ effector memory T-cell subset ($p=8.2\text{x}10^{-5}$, T.4Mem44h62l.LN CD4+, see **Figure 2.7A**). This was the only cell-type that obtained significance after correcting for multiple hypothesis testing. Indeed, each of the CD4+ effector memory cell subsets demonstrated at least nominally significant association at $p<0.008$. To identify the contribution of the individual loci toward the effector memory T-cell enrichment, we averaged the specificity profiles of all four primary CD4+ effector memory cell subsets together and again tested the aggregate effector memory T-cell profile for association among these nominally associated loci. We again observed an association ($p=1.3\text{x}10^{-4}$). Looking at the individual loci and genes, we note that there are 68 loci that show specificity for CD4+ effector memory T-cell populations at a $p<0.1$ level while by chance alone, we would expect only 43.6 (see **Figure 2.7B)**. Based on these results, we might expect that as many as 25 true RA risk loci are embedded within this set. Of the loci tested, we list those with the most significant specific expression in CD4+

effector memory T-cells (**Table 2.2**). We predict that subsequent ongoing genetic association studies for RA will eventually clarify which of these are true RA loci.

We assessed the degree of enrichment at different more liberal GWAS significance thresholds. In order to do this, we grouped SNPs into 50 p_{GWAS} bins, each of size 0.001, ranging from 0 to 0.05. Then for each group we quantified the degree to which genes implicated by those SNPs were enriched for CD4+ effector memory cell specific expression. We observed at least nominally significant enrichment for bins up to $p_{GWAS}<0.005$, with very little evidence of any enrichment at $p_{GWAS}>0.02$ (see **Figure 2.7C**).

In the present study, we looked at gene expression data alone to ascertain the key cell types impacted by autoimmune loci. Previously, the potential value of using gene expression data, and other external information sources, in integrative analysis to understand relationships between disease-associated genes and to identify candidate genes for follow-up study has been demonstrated. For example, we have separately integrated protein-protein interaction data with expression data to identify specific pathways in disease[26]. As another example, Prioritizer uses a large compendium of gene expression data, along with a multitude of other data sources, to identify likely candidate genes within loci[53]. Chen used a large-scale gene expression compendium to look for genes that vary most dramatically across Gene Expression Omnibus and to identify potential candidate genes[54]. Our approach is contingent on the quality and availability of a high-quality gene expression database. A comprehensive dataset containing all of the necessary human tissue types would be most ideal. While the GNF dataset is reasonably comprehensive, important immune cell-types are not always present. On the other hand, ImmGen offers the highest quality and most comprehensive immunological dataset that we are aware of. It does lack certain important derived

Figure 2.7 Cell-specific expression of genes in nominally associated RA loci. We evaluated 436 loci containing SNPs nominally associated to rheumatoid arthritis for cell-specific gene enrichment in 223 murine immune cell types. The Bonferroni-corrected p-value is shown by a dotted line. A) We listed cell types that are significant after multiple hypothesis testing ($p<2.2x10^{-4}$). Only one of the four CD4+ effector memory cells (T.4Mem44h62l.LN) is significant. B) We aggregated the specificity scores for the four different CD4+ effector memory T-cell types and calculated empirical locus *p*-values for each of the 436 loci. These *p*-values assessed the degree of specificity that the most highly specific CD4+ effector memory T-cell in each locus achieved. In red, we plotted the histogram of these empirical locus *p*-values while in grey, we plotted the expected histogram of empirical locus *p*-values. We plotted the ratio of those two values at each *p*-value interval. We noted modest deflation at higher values ($p>0.5$) and inflation at lower *p*-values ($p<0.1$). C) We grouped loci by their association statistics (p_{GWAS}) into 50 bins ranging from $p_{GWAS}<0.001$ (as in pane A and B) to 0.049 $p_{GWAS}<0.05$. Then, using aggregated specificity scores for CD4+ effector memory T-cell types we evaluated these groups to see if they were enriched for specifically expressed genes. For each bin we plot the observed to expected ration of loci with lower empirical locus *p*-values ($p<0.1$, blue, left axis), and the statistical significance of enrichment (red, right axis).

Figure 2.7 Cell-specific expression of genes in nominally associated RA loci (Continued).

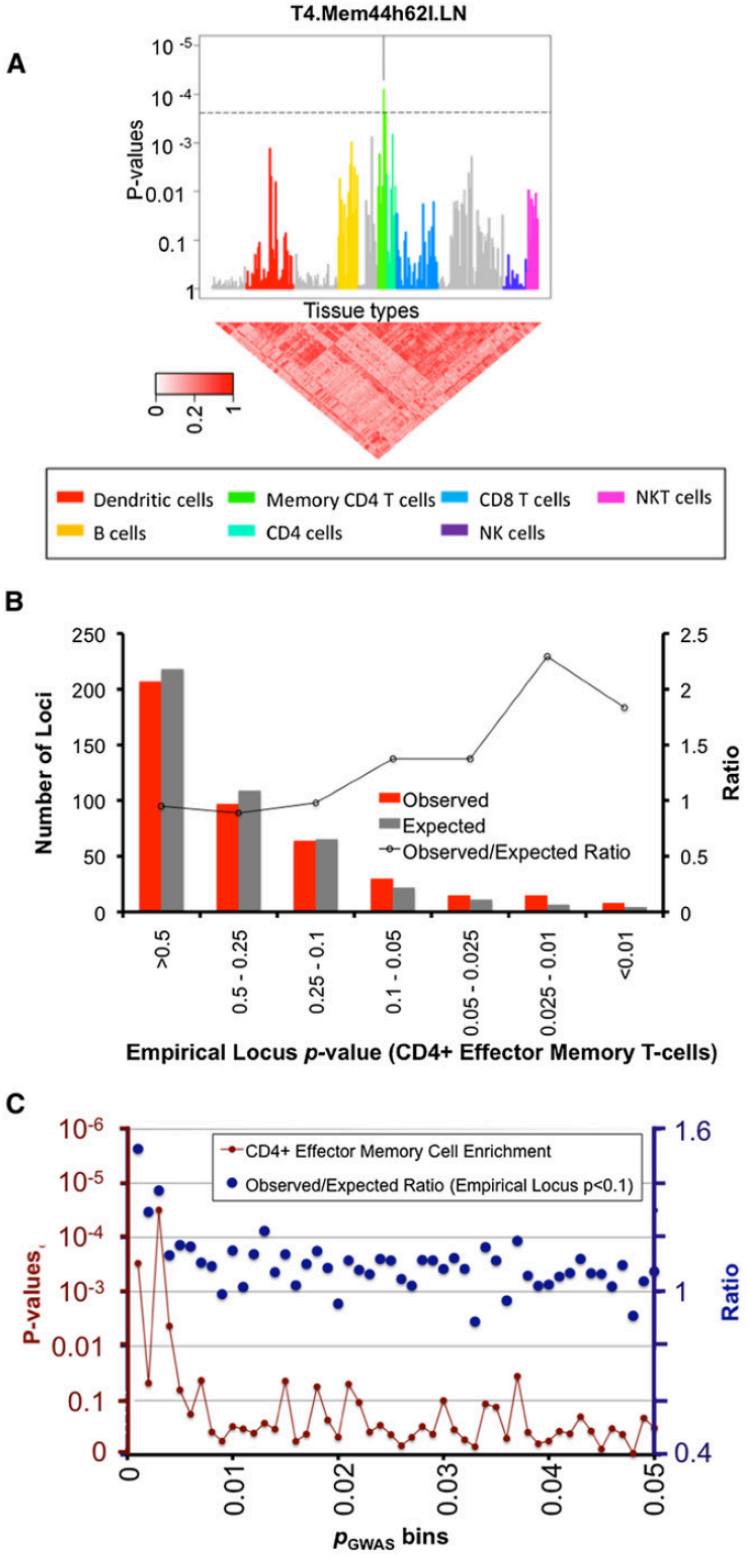


Table 2.2. Nominally Associated RA alleles near genes specifically expressed in CD4+ effector memory cells. Here, we listed nominally associated SNPs and their p-values from RA GWAS (p<0.001, first, and last columns respectively), their genomic coordinates (second and third column), a significance score suggesting enrichment for a single proximate gene for specific CD4+ effector memory T-cells (fourth column), and the candidate gene with the most specific expression in CD4+ effector memory T-cells (fifth column).

SNP	CHR	hg18 Position	Empirical locus p ($CD4^+$ T_{EM} cell)	Most specifically expressed gene (Entrez ID)	Best assoc. RA p-val from GWAS
rs6683027	chr1	204663522	0.0011	*CTSE* (1510)	4.26E-04
rs11867591	chr17	62021413	0.0032	*PRKCA* (5578)	5.06E-04
rs2023628	chr8	17091505	0.0043	*ZDHHC2* (51201)	1.32E-04
rs10937694	chr4	5979650	0.0054	*CRMP1* (1400)	7.64E-04
rs7155603	chr14	75030289	0.0084	*BATF* (10538)	1.53E-05
rs17215817	chr8	131488842	0.0094	*DDEF1* (50807)	8.22E-05
rs16898297	chr8	101453401	0.0096	*RNF19A* (25897)	7.58E-04
rs7046901	chr9	20236894	0.0097	*MLLT3* (4300)	6.43E-04
rs10468137	chr15	86012950	0.011	*NTRK3* (4916)	7.32E-04
rs735684	chr5	141465117	0.013	*NDFIP1* (80762)	9.71E-05

Table 2.2. Nominally Associated RA alleles near genes specifically expressed in CD4+ effector memory cells (Continued).

SNP	CHR	hg18 Position	Empirical locus *p* ($CD4^+$ T_{EM} cell)	Most specifically expressed gene (Entrez ID)	Best assoc. RA *p*-val from GWAS
rs6021275	chr20	49588531	0.013	*NFATC2* (4773)	6.30E-04
rs7579944	chr2	30298530	0.014	*LBH* (81606)	1.08E-04
rs9907505	chr17	73250509	0.017	*SEPT9* (10801)	1.94E-05
åsrs2939931	chr10	121626396	0.018	*INPP5F* (22876)	9.94E-04
rs9366347	chr6	20474041	0.019	*MBOAT1* (154141)	6.16E-04
rs1422673	chr5	150419181	0.02	*TNIP1* (10318)	9.51E-05

cell types of potential interest. For example, derived helper T- cell subgroups such as Th1, Th2, and Th17 cells are not individually profiled. One additional limitation of ImmGen is that it is based on mouse, and not human, tissues. While the immune systems of the mouse and human are very similar in lineage and structure, there are also important differences. But the breadth of data collected for ImmGen would be impractical to obtain in human.

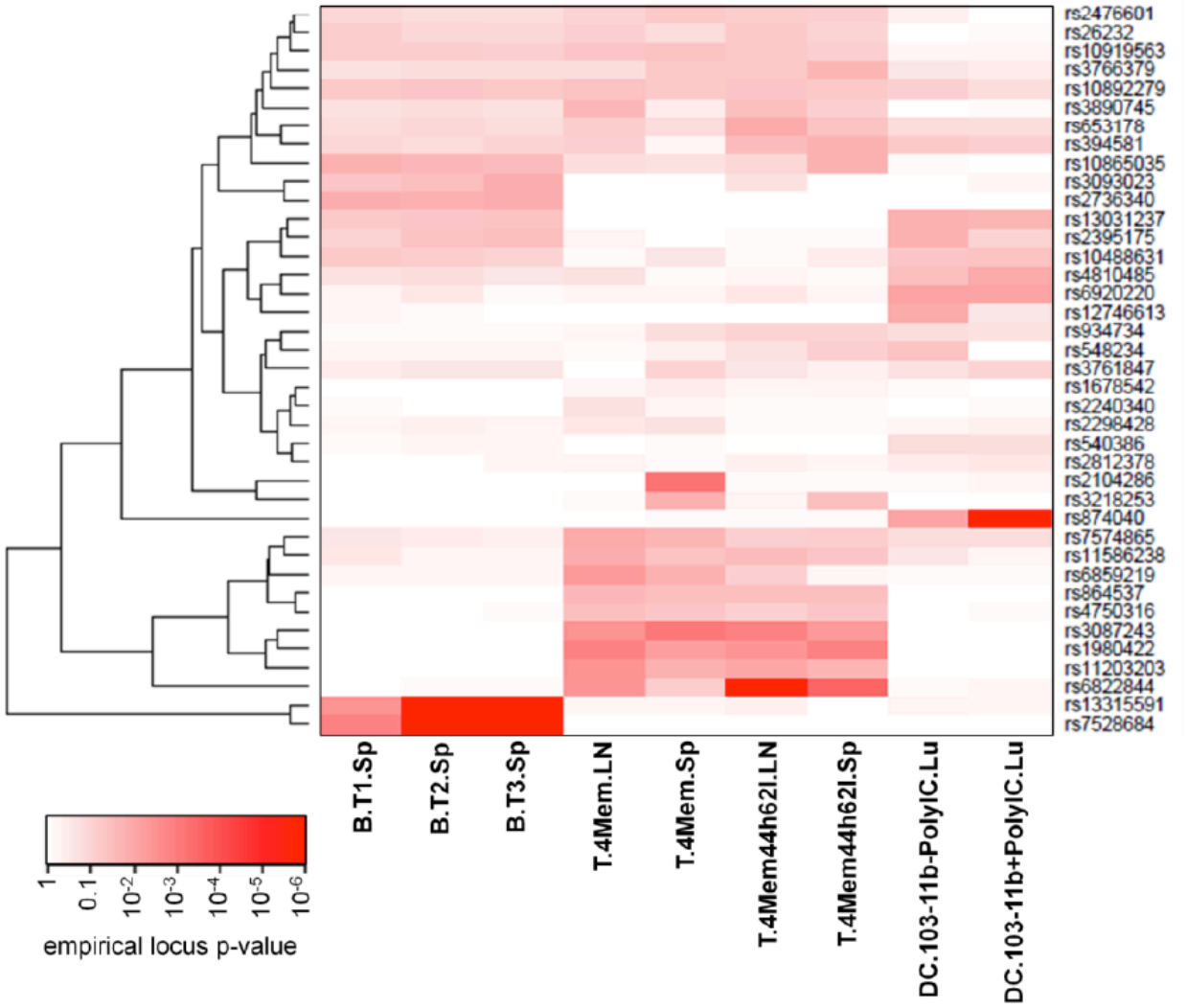
For each of the autoimmune diseases we are able to identify very specific subsets of immune cells that may play a critical role in disease, that go well beyond broad immunological categories. For example for RA we are able to not only establish that CD4+ T cells express genes within RA loci, but we are able to go beyond that and specifically implicate the very specific effector memory subset. All four of the CD4+ effector memory T-cell subsets achieve the greatest significance in this data set, and adjusting for their effects obviates the other less significant observations. In this case, we validate our results by looking at independent SNP sets with more nominal disease association.

Intriguingly, we note for the autoimmune diseases that while a single cell-type is most strongly associated, there is often evidence that more than one immune cell-type is involved. For example, for RA, there is a nominally significant cell-type association for B-cell subsets led by follicular B cells (B.Fo.Sp, p=0.00032), stage 2 transitional B cells (B.T2.Sp, p=0.00041), and nine other B-cell subsets obtained $p<0.01$. The loci driving the B-cell subset association are distinct from those driving the CD4+ effector memory cell association (see **Figure 2.8**). Thus, adjusting for CD4+ effector memory T-cell profiles does not completely remove the B-cell association signal. Similarly for Crohn's disease, after adjusting for the main effects of dendritic cells, there are remaining nominal signals in NK and CD4+ T-cell subsets. While these associations are not significant after multiple hypothesis testing, they may suggest possible separate roles of other cell-types in disease that might become more apparent as additional SNP discoveries accumulate. Since risk alleles across autoimmune diseases are known to overlap, diseases may be best understood by

Figure 2.8. Patterns of cell-specific expression of RA loci

Here we plot the association between specific SNPs associated with RA (right) and selected tissues (bottom). Redness in each box correlates with the significance as measured by the empirical locus p-value; red boxes indicate that the SNP is in LD that is highly expressed in the tissue based on ImmGen, after accounting for the number of genes within the locus. SNPs are hierarchically clusters (left). Some of the SNPs toward the top are uninformative either because they lack a gene that is specifically expressed, or have genes with specific expression in multiple displayed cells. SNPs toward the bottom have the most informative expression patterns. A large number of the most informative SNPs have signal for the CD4+ effector memory cell subsets, but a small number have specific expression for transitional B-cells as well. These sets are mutually exclusive.

Figure 2.8. Patterns of cell-specific expression of RA loci (Continued).



considering individual immune cell types. While the distribution of immune cell types that are critical to particular diseases may vary, overlapping loci between different diseases might be explained by overlapping pathogenic cell types that may play a common role in the different diseases.

This algorithm has been implemented as *SNPsea* by Kamil Slowikowski [55] and made freely available here: http://www.broadinstitute.org/mpg/snpsea/.

ACKNOWLEDGEMENTS

This work benefitted from data assembled by the ImmGen Consortium. This work was supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH-NIAMS) Development Award [1K08AR055688] and by the Harvard Health Sciences and Technology Program. We thank Joel Hirschorn, Michael B. Brenner, David Altshuler, and Lude Franke for helpful discussions. The authors declare no conflict of interests.

WEB RESOUCES

http://www.omim.org/

http://www.immgen.org/index_content.html

http://www.broadinstitute.org/mpg/snpsea/

REFERENCES

1. Barrett, J.C., et al., *Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes.* Nat Genet, 2009. **41**(6): p. 703-7.

2. Raychaudhuri, S., *Recent advances in the genetics of rheumatoid arthritis.* Curr Opin Rheumatol, 2010. **22**(2): p. 109-18.

3. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.

4. International Schizophrenia, C., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.

5. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

6. Zhong, H., et al., *Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes.* PLoS Genet, 2010. **6**(5): p. e1000932.

7. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.

8. Price, A.L., et al., *Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals.* PLoS Genet, 2011. **7**(2): p. e1001317.

9. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.

10. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.

11. Firestein, G.S., *Evolving concepts of rheumatoid arthritis.* Nature, 2003. **423**(6937): p. 356-61.

12. Wipke, B.T. and P.M. Allen, *Essential role of neutrophils in the initiation and progression of a murine model of rheumatoid arthritis.* J Immunol, 2001. **167**(3): p. 1601-8.

13. Lee, D.M., et al., *Mast cells: a cellular link between autoantibodies and inflammatory arthritis.* Science, 2002. **297**(5587): p. 1689-92.

14. Kinne, R.W., et al., *Macrophages in rheumatoid arthritis.* Arthritis Res, 2000. **2**(3): p. 189-202.

15. Boilard, E., et al., *Platelets amplify inflammation in arthritis via collagen-dependent microparticle production.* Science, 2010. **327**(5965): p. 580-3.

16. Pap, T., et al., *Fibroblast biology. Role of synovial fibroblasts in the pathogenesis of rheumatoid arthritis.* Arthritis Res, 2000. **2**(5): p. 361-7.

17. Lefevre, S., et al., *Synovial fibroblasts spread rheumatoid arthritis to unaffected joints.* Nat Med, 2009. **15**(12): p. 1414-20.

18. McCarroll, S.A., et al., *Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease.* Nat Genet, 2008. **40**(9): p. 1107-12.

19. Dendrou, C.A., et al., *Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource.* Nat Genet, 2009. **41**(9): p. 1011-5.

20. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.

21. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.

22. Hyatt, G., et al., *Gene expression microarrays: glimpses of the immunological genome.* Nat Immunol, 2006. **7**(7): p. 686-91.

23. Raychaudhuri, S., et al., *Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function.* PLoS Genet, 2010. **6**(9): p. e1001097.

24. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

25. Raychaudhuri, S., et al., *Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions.* PLoS Genet, 2009. **5**(6): p. e1000534.

26. Rossin, E.J., et al., *Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology.* PLoS Genet, 2011. **7**(1): p. e1001273.

27. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

28. Myers, S., et al., *A fine-scale map of recombination rates and hotspots across the human genome.* Science, 2005. **310**(5746): p. 321-4.

29. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature, 2010. **467**(7317): p. 832-8.

30. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.

31. Berg, J., J. Tymoczko, and L. Stryer, *Biochemistry*. 6th ed. 2006, New York: WH Freeman.

32. Hobbs, H.H., M.S. Brown, and J.L. Goldstein, *Molecular genetics of the LDL receptor gene in familial hypercholesterolemia.* Hum Mutat, 1992. **1**(6): p. 445-66.

33. Musunuru, K., et al., *From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.* Nature, 2010. **466**(7307): p. 714-9.

34. Speliotes, E.K., et al., *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nat Genet, 2010. **42**(11): p. 937-48.

35. Clement, K., et al., *A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction.* Nature, 1998. **392**(6674): p. 398-401.

36. Pinkney, J., et al., *Hypothalamic obesity in humans: what do we know and what can be done?* Obes Rev, 2002. **3**(1): p. 27-34.

37. Willer, C.J., et al., *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.

38. Raychaudhuri, S., J.M. Stuart, and R.B. Altman, *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pac Symp Biocomput, 2000: p. 455-66.

39. Gateva, V., et al., *A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus.* Nat Genet, 2009. **41**(11): p. 1228-33.

40. Han, J.W., et al., *Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus.* Nat Genet, 2009. **41**(11): p. 1234-7.

41. International Consortium for Systemic Lupus Erythematosus, G., et al., *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.* Nat Genet, 2008. **40**(2): p. 204-10.

42. Hom, G., et al., *Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX.* N Engl J Med, 2008. **358**(9): p. 900-9.

43. Kozyrev, S.V., et al., *Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(2): p. 211-6.

44. Stahl, E.A., et al., *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci.* Nat Genet, 2010. **42**(6): p. 508-14.

45. Zhernakova, A., et al., *Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci.* PLoS Genet, 2011. **7**(2): p. e1002004.

46. Chan, O.T., M.P. Madaio, and M.J. Shlomchik, *The central and multiple roles of B cells in lupus pathogenesis.* Immunol Rev, 1999. **169**: p. 107-21.

47. Navarra, S.V., et al., *Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial.* Lancet, 2011. **377**(9767): p. 721-31.

48. Zheng, Y., et al., *Phosphorylation of RasGRP3 on threonine 133 provides a mechanistic link between PKC and Ras signaling systems in B cells.* Blood, 2005. **105**(9): p. 3648-54.

49. Cho, J.H., *The genetics and immunopathogenesis of inflammatory bowel disease.* Nat Rev Immunol, 2008. **8**(6): p. 458-66.

50. Baumgart, D.C. and S.R. Carding, *Inflammatory bowel disease: cause and immunobiology.* Lancet, 2007. **369**(9573): p. 1627-40.

51. Cooney, R., et al., *NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation.* Nat Med, 2010. **16**(1): p. 90-7.

52. Travis, M.A., et al., *Loss of integrin alpha(v)beta8 on dendritic cells causes autoimmunity and colitis in mice.* Nature, 2007. **449**(7160): p. 361-5.

53. Franke, L., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011-25.

54. Chen, R., et al., *FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease.* Genome Biol, 2008. **9**(12): p. R170.

55. Slowikowski, K., X. Hu, and S. Raychaudhuri, *SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci.* Bioinformatics, 2014. **30**(17): p. 2496-7.

CHAPTER 3

X-Cyt: Automated Cytometric Data Analysis

User-guided cytometric data analysis for large-scale immunoprofiling studies: application to invariant natural killer T cells

Xinli Hu*[1-5], Hyun Kim*[1-3], Patrick J. Brennan[1], Buhm Han[1-4], Clare Baecher-Allan[6], Philip L. De Jager[4,7], Michael B. Brenner[1], Soumya Raychaudhuri[1-4,8]

1. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
2. Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
3. Partners Center for Personalized Genetic Medicine, Boston, MA, USA.
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
5. Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA
6. Department of Dermatology/Harvard Skin Disease Research Center, Brigham and Women's Hospital, Boston, MA, USA
7. Program in Translational Neuropsychiatric Genomics, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
8. Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK

*These authors contributed equally to this manuscript

Correspondence:

Michael B. Brenner mbrenner@research.bwh.harvard.edu

Soumya Raychaudhuri (soumya@broadinstitute.org)

Originally published as:

Hu, X., H. Kim, P. J. Brennan, B. Han, C. M. Baecher-Allan, P. L. De Jager, M. B. Brenner and S. Raychaudhuri (2013). "Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells." Proc Natl Acad Sci U S A **110**(47): 19030-19035.

ABSTRACT

Defining and characterizing pathologies of the immune system require precise and accurate quantification of abundances and functions of cellular subsets via cytometric studies. Currently, data analysis relies on manual gating, which is a major source of variability in large-scale studies. We devised an automated, user-guided method, *"X-Cyt"*, which specializes in rapidly and robustly identifying targeted populations of interest in large datasets. We first applied *X-Cyt* to quantify $CD4^+$ effector and central memory T cells in 236 samples, demonstrating high concordance to manual analysis (r = 0.91 and 0.95, respectively), and superior performance to other available methods. We then characterized the population dynamics of invariant natural killer T (iNKT) cells, a particularly rare peripheral lymphocyte, in 110 individuals by assaying 19 markers. We demonstrated that while iNKT cell numbers and marker expression are highly variable in the population, the iNKT abundance correlates with gender and age, and the expression of phenotypic and functional markers correlate closely with CD4 expression.

INTRODUCTION

Flow cytometry is a technology widely used in clinical practice and in research, particularly in the field of immunology. It is capable of interrogating a wide variety of markers on many different cell types on a single-cell basis using fluorophore-conjugated antibodies. While molecular as well as genomic studies have advanced understanding of immunological processes and autoimmune diseases, the components of the human immune system and their functions have yet to be comprehensively described. Without such a reference "catalog" of the immune system, it is ultimately difficult to interpret the pathogenic significance of genetic, molecular, or phenotypic variants observed in diseases.

Immunoprofiling is emerging as a means to establish the constituents, physiological roles, and population dynamics of the immune system [1]. Specifically, it aims to define A) the cellular components of the immune system, B) the developmental processes and lineage relationship among the cell types, and C) the phenotypes and functions of each cell type at different physiological states. To profile such a complex and dynamic system in large sample sizes, high-throughput cytometric studies have become crucial.

Cytometric technologies are quickly advancing and outpacing analytical approaches. Currently, flow cytometers can measure up to 17 markers [2]. Next-generation cytometers, such as Cytometry by Time of Flight (CyTOF), will soon to be able to assay hundreds of markers [3]. However, data analysis largely relies on manual gating by expert analysts. It is a simple but slow process dependent on one- or two-dimensional visualization and sequential gating using software such as FlowJo™. As the number of samples and markers in a study increase, gating becomes increasingly time consuming, inconsistent, and does not fully exploit the power of high-dimensional information contained in these complex studies.

In recent years, a number of automated methods for cytometric data handling, particularly for cell population identification, have emerged and demonstrated power to harness the rich

information in large-scale data, minimize inconsistencies, and reduce analysis time [4]. Current methods use parametric [5-7] or nonparametric [8-13] clustering to partition high-dimensional data. Some methods specialize in capturing difficult (such as rare or convex) cell populations [6, 7, 13], and delineate developmental and functional relationships among cell types [14]. These methods make no assumptions about the underlying structure of the data and primarily aim to discover all discernible populations *de novo* in each sample. Consequently, they have been used primarily for exploratory studies.

In contrast to exploratory studies, the goal of many immunoprofiling studies is to reliably and consistently identify the target cell population across many individuals. For example, a profiling study may aim to quantify regulatory T (Treg) cells in healthy controls and patients with autoimmune diseases, using antibodies specifically selected for identifying Tregs. In this case, the goal of the analysis is to accurately extract Tregs from all samples using a standardized definition. Automating this type of analysis is challenging because accurate inter-sample alignment of cell populations is required in addition to the partitioning of the cell populations within each sample.

We developed a user-guided analytical tool, "*X-Cyt*", for automating targeted population identification in immunoprofiling studies. X-Cyt uses multivariate mixture modeling for partitioning cytometric data. Unlike unsupervised methods, X-Cyt allows the user to set up the optimal partitioning scheme. By applying a uniform scheme to all samples in a cohort, X-Cyt consistently identifies and aligns the targeted cell populations.

In this study, we aimed to identify and characterize invariant natural killer T (iNKT) cells. iNKT cells are lymphocytes with a non-diverse T cell receptor repertoire that recognizes CD1d-presented lipid antigens [15-17], and in humans normally make up less than 0.5% of circulating peripheral blood mononuclear cells (PBMCs) [18]. They play important roles in host defense, autoimmunity, allergy, and cancer [19]. Functional characterization of iNKT cells requires comprehensive assessment of surface expression of homing receptors, lectins, cell adhesion

molecules, as well as cytokine production. Immunoprofiling studies have yet to assay such a comprehensive set of markers in primary iNKT cells in a sufficiently large cohort [18, 20-24]. Here, we profiled iNKT cells in 110 subjects with 19 surface and intracellular markers.

RESULTS

Overview of the X-Cyt method.

X-Cyt identifies the populations of interest in a given sample by partitioning all events into clusters following a user-designed partitioning scheme. When more than one marker is used to define populations, X-Cyt partitions the data using multivariate mixture modeling via an expectation-maximization (EM) algorithm, as described in **Methods**.

We make the assumption that in profiling studies, samples within a cohort share a general cell population structure. That is, similar cell populations are present in all samples and their relative spatial configuration is conserved. X-Cyt therefore aims to follow the same user-defined partitioning scheme to analyze all samples, while allowing for biological and technical variations. Population identification by X-Cyt is therefore accomplished in two major steps: 1) a user-guided "trial" analysis to set up the partitioning scheme, and 2) a template-guided cohort analysis. Markers that describe the phenotype and function of cells are analyzed separately downstream of population identification.

Step 1. Set up the partitioning scheme. The goal of the initial "trial" phase of the analysis is to set up a partitioning scheme by optimizing two parameters for mixture modeling: 1) a parsimonious combination of differentiation markers for defining the population(s) of interest and 2) the number of clusters to adequately and intuitively partition the events. The user test-partitions a few representative samples using different input parameters, evaluates the results, and then chooses one optimal scheme. The ideal resulting configuration is one that most accurately captures each target population as one coherent cluster of events (see **Methods** for detailed description of

parameter selection). The user-approved configuration (parameters of the mixture model components) is passed onto Step 2 as the template (see **Figure 3.1A**).

Step 2. Template-guided cohort analysis. X-Cyt initializes the mixture model parameters of each new sample to that of the template. The EM algorithm then iteratively updates the parameters describing the location, shape (covariance matrix), and the proportion of each cluster. The EM algorithm indexes each emerging cluster according to the template, which automatically aligns across all samples simultaneously to clustering (**Figure 3.1B**). Downstream to population extraction, markers that describe the phenotype and function of cells are analyzed separately (**Figure 3.1C**).

We have made X-Cyt, along with a sample dataset and user input files, available for download at http://www.broadinstitute.org/mpg/xcyt/.

Demonstration of X-Cyt's performance in two datasets

We first assessed the performance of X-Cyt in identifying common cell populations by querying the proportions of memory cell subsets in $CD4^+$ T cells. We isolated $CD4^+$ T cells from PBMCs via magnetic-activated cell sorting (MACS) depletion from a cohort of 236 healthy donors (**Table 3. 1**), and labeled them with antibodies against CD45RA, CD45RO, and CD62L (see **Methods** and **SI Appendix** for experimental methods).

To identify effector memory (T_{EM}) and central memory (T_{CM}) T cells, we partitioned each sample in two steps: 1) bivariate normal mixture modeling using forward- (FSC) and side-scatter (SSC) to obtain a purer $CD4^+$ T cell population, and 2) 3-dimensional normal

Table 3. 1. Demographics of enrolled subjects

	T Cell Study		iNKT Cell Study	
	Females	Males	Females	Males
Total Enrollment	128	88	62	48
# with Repeat Visit	16	18	5	6
Mean Age (range)	28.8 (19-57)	34.9 (19-54)	27.6 (19-52)	34.6 (19-53)

mixture modeling using CD45RA, CD45RO, and CD62L. To determine the optimal partitioning scheme, an expert analyst assessed different sets of partitioning parameters in ten random samples. In Step 1, a two-component mixture model captured the $CD4^+$ T cell population, which was extracted from each sample. In Step 2, the analyst evaluated a range of four to nine clusters and selected the 7-cluster model as it most accurately captured the T_{EM} and T_{CM} subsets (**Figure 3.2A**). X-Cyt applied this partitioning scheme to all 236 samples and consistently identified the T_{EM} and T_{CM} subsets (representative samples shown in **Figure 3.2B**). We compared X-Cyt results to proportions defined by an independent expert cytometry analyst with manual gating in FlowJo™. We observed that the proportions for both populations were highly concordant with the manual analysis ($r = 0.91$ and $r = 0.95$, $p < 10^{-15}$, Pearson correlation test; **Figure 3.2C**). We wanted to quantitatively compare the performance of X-Cyt to automated methods that were the top five performers in the FlowCAP consortium challenge [4]: FLoCK, ADICyt, flowMeans, FLAME, and SamSPECTRAL. We were unable to run FLoCK since it was not able to use standard FCS3.0 format files. We ran each method with their default parameters to identify $CD4^+$ T_{EM} cells from lymphocytes.

We analyzed all 236 samples using FLAME, and compared the $CD4^+$ T_{EM} cell percentages to those procured by an expert user via gating in FlowJo™. FLAME achieved more modest

Figure 3.1. Schematic of X-Cyt's analytical process (synthetic samples). A) In a few representative samples, the user adjusts analytical parameters and evaluates the clustering outcome. Adjustable parameters include the differentiation markers to be used, the number of clusters in mixture modeling (g), distribution type, and standard deviation cutoffs for continuous markers. The user selects one optimal set of parameters that most accurately identifies the cell populations of interest (here the blue cluster using g=3). The clustering result of the representative sample is chosen as the template (dashed circles in the g=3 panel). B) X-Cyt applies the template to guide the partitioning of all samples in the study. The population of interest (shown in red dots and blue dashed circle) is consistently identified across all samples. C) Downstream to population extraction, random samples are pooled to establish the distribution of phenotypic/functional markers. The percentage of cells positive for each marker is reported based on either mixture modeling (top) or standard deviation cutoff (bottom).

Figure 3.1. Schematic of X-Cyt's analytical process (synthetic samples) (Continued).

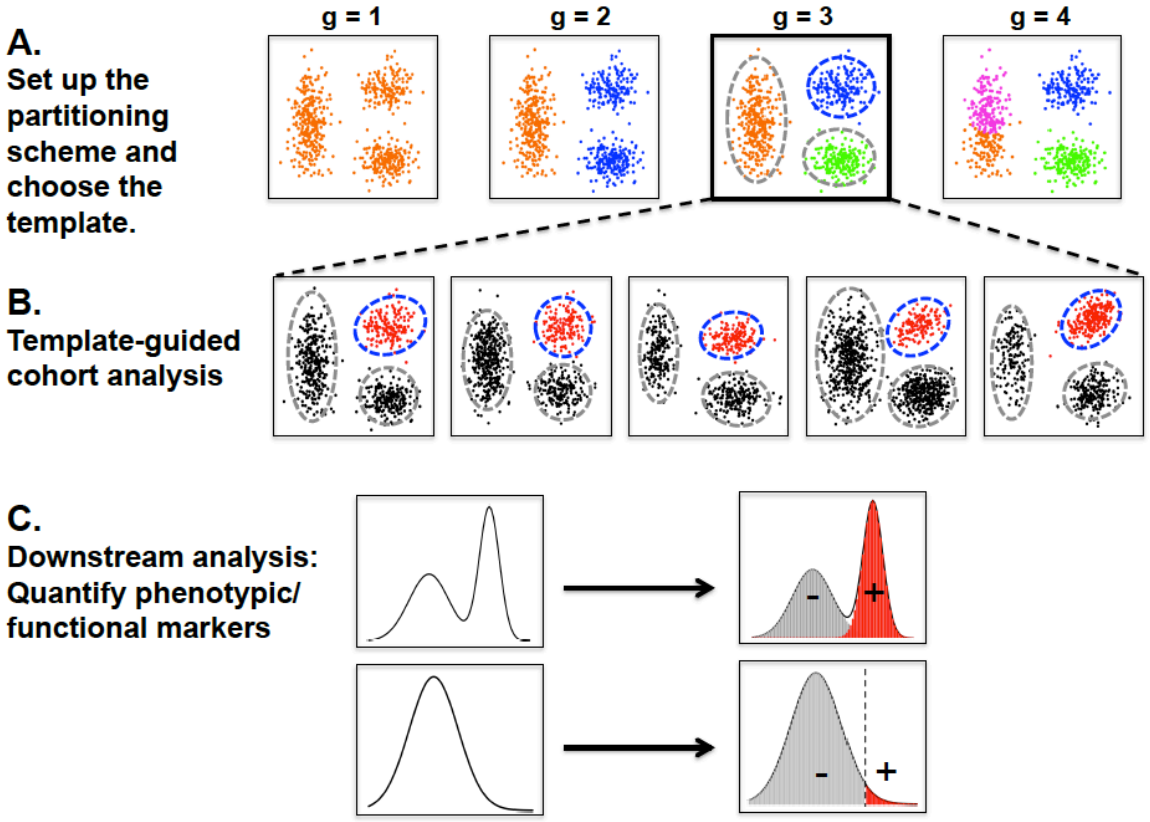
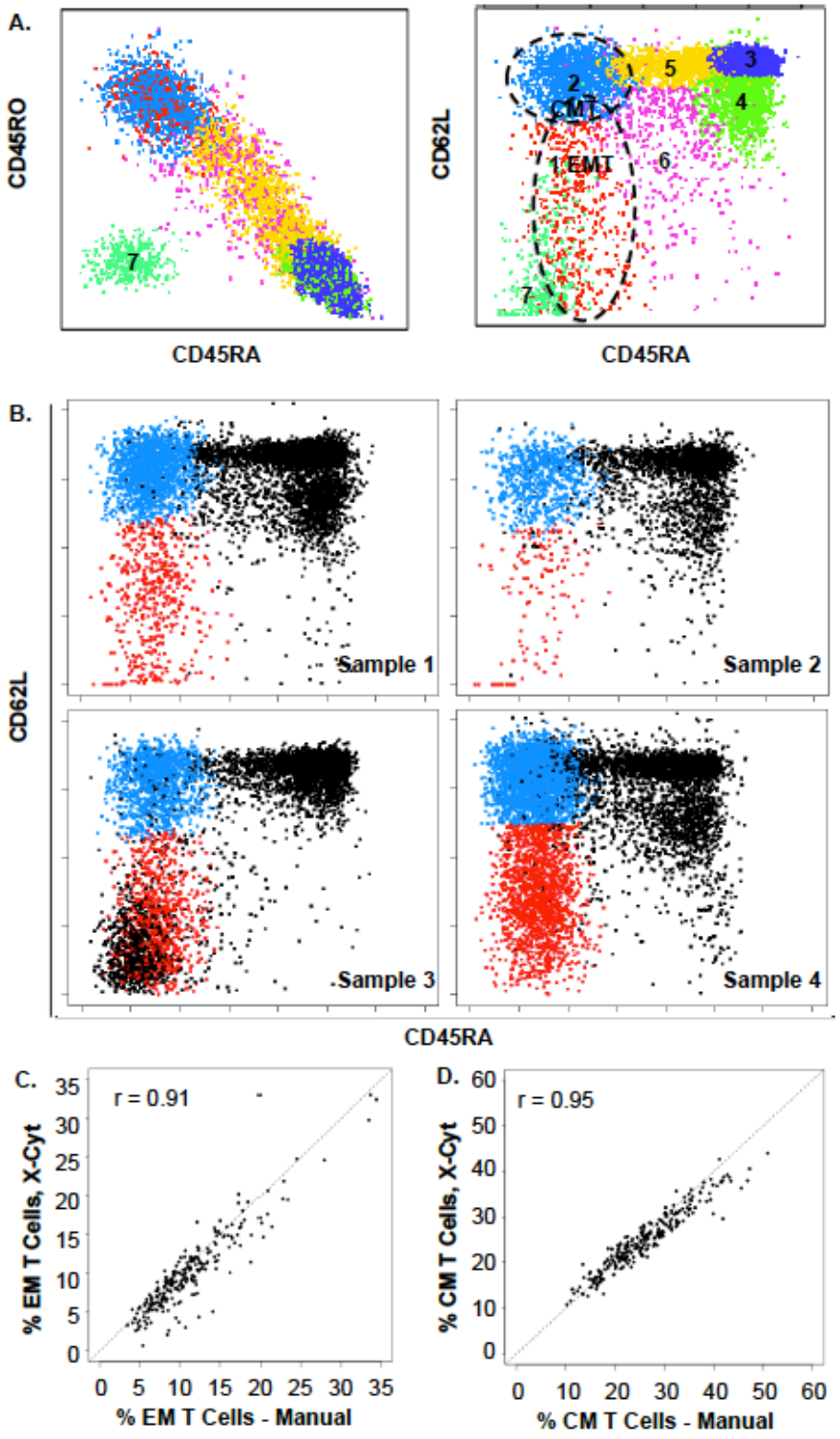


Figure 3.2. CD4+ memory T cell subset identification. A) An optimal model of seven clusters using CD45RA, CD45RO, and CD62L identified T_{EM} (red, cluster1) and T_{CM} (blue, cluster2) cells. Other clusters include naïve, and intermediate CD4+ T cells, as well as impurities. B) X-Cyt consistently identified the T_{EM} (red) and T_{CM} (blue) populations in all samples. Four random samples are shown here. C) X-Cyt and manual gating in FlowJo returned highly concordant proportions of T_{EM} and T_{CM} in 236 samples.

Figure 3.2. CD4+ memory T cell subset identification (Continued).



concordance (r = 0.50), compared to that achieved by X-Cyt (r = 0.91, see **Figure 3.3**). Because flowMeans and SamSPECTRAL do not align clusters across samples, comparison of results to X-Cyt was not possible without manual intervention. Therefore we manually inspected a random subset (20 samples) of the clustering results, and in each sample selected the cluster most closely representing the T_{EM} cells to obtain a concordance. Even after manual selection of clusters, flowMeans and SamSPECTRAL achieved limited concordances of only 0.57 and 0.44, respectively. ADICyt had a high sample failure rate. In three separate attempts with the same 20 samples, we observed that on average 50% of samples failed to cluster with different random seeds. We note however, that ADICyt achieved high performance on the limited samples that it did successfully analyze (r = 0.98, average of three runs). Representative clustering results by each method are shown in **Figure 3.4**.

Next we challenged X-Cyt to identify a rare population, mucosal associated invariant T (MAIT) cells, from PBMCs for 35 subjects. We labeled cells with antibodies against CD3, CD45, Vα7.2, and CD161. Following convention, we defined MAIT cells as $CD3^{+}CD45^{+}V\alpha7.2^{+}CD161^{+}$. We first partitioned PBMCs into four clusters using FSC and SSC to obtain lymphocytes. Subsequently, we partitioned in CD3 and CD45 dimensions to obtain a double-positive T cell population. In five random samples, the analyst evaluated three to seven clusters, and selected the six-cluster model as the best to identify MAIT cells. We then applied the template to all 35 samples. Comparing proportions obtained by X-Cyt with those procured by an independent manual analyst, we again observed high concordance (r = 0.98, $p < 10^{-15}$; **Figure 3.5**).

Figure 3.3. Comparison between FLAME and X-Cyt. FLAME and X-Cyt were compared by their abilities to identify the effector memory T (EMT) cell subset in 236 samples. X-Cyt achieved a significantly higher concordance with manually obtained cell proportions than FLAME.

Figure 3.3. Comparison between FLAME and X-Cyt (Continued).

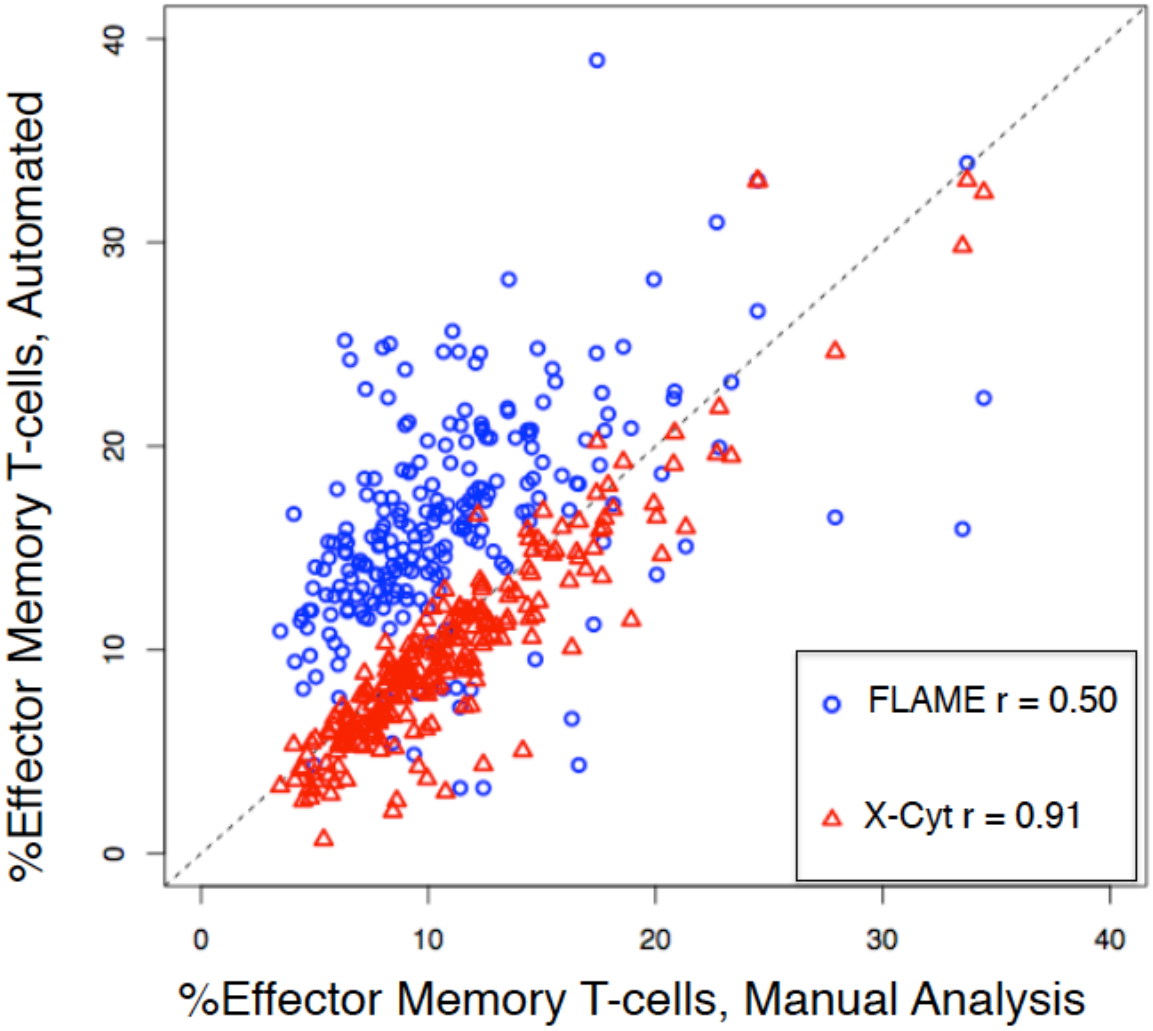


Figure 3.4. Four automated methods were compared for their performance in quantifying CD4+ effector memory T cells in 20 samples. Four random samples A – D are shown here. X-Cyt and FLAME return aligned clusters across samples. For flowMeans, SamSPECTRAL and ADICyt, which do not provide cluster alignment across all samples, the expert analyst manually selected the cluster that most closely agrees with the TEM population in each sample. ADICyt performs a different mathematical transformation of the raw data than other methods, therefore produces different but equivalent plots. Due to high sample failure rate, the concordance of ADICyt is calculated by averaging three runs, excluding the (seven to sixteen) failed samples in each. The fourth sample (marked by *) failed in the first two rounds of clustering. The five methods' concordances (Pearson's r) with manual analysis are 0.93, 0.86, 0.57, 0.44, and 0.98, respectively.

Figure 3.4. Four automated methods were compared for their performance in quantifying CD4+ effector memory T cells in 20 samples (Continued).

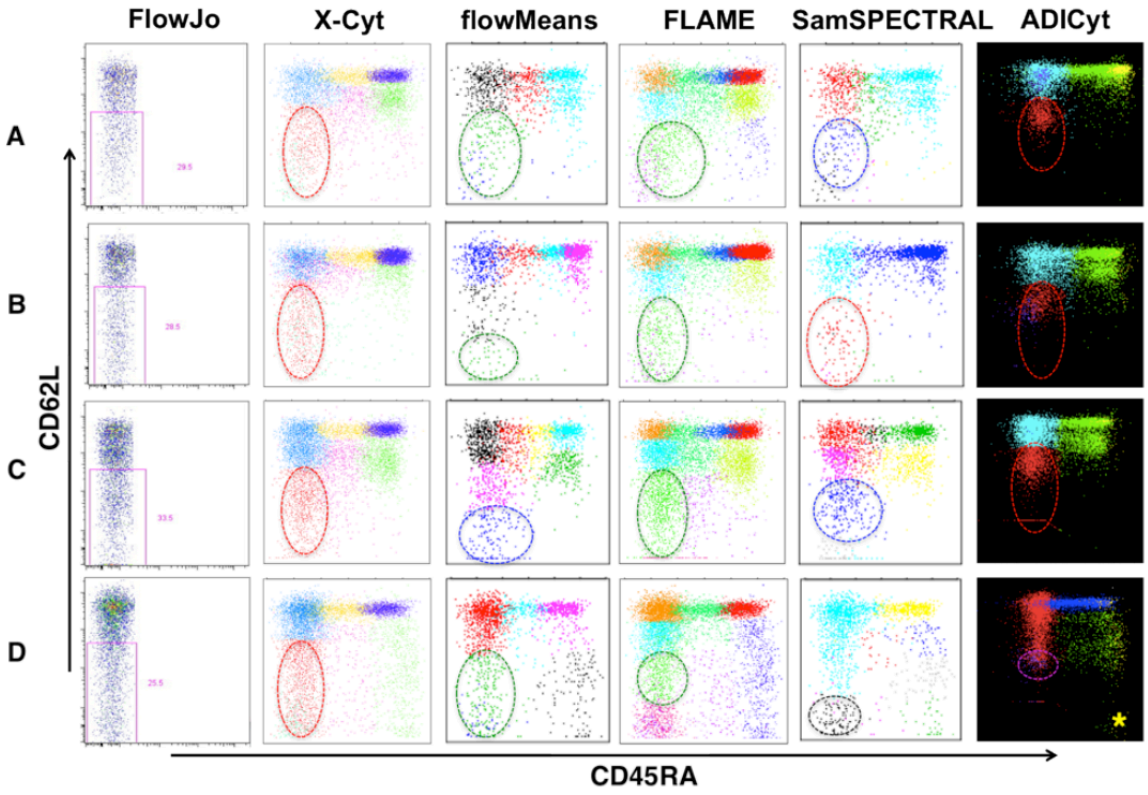
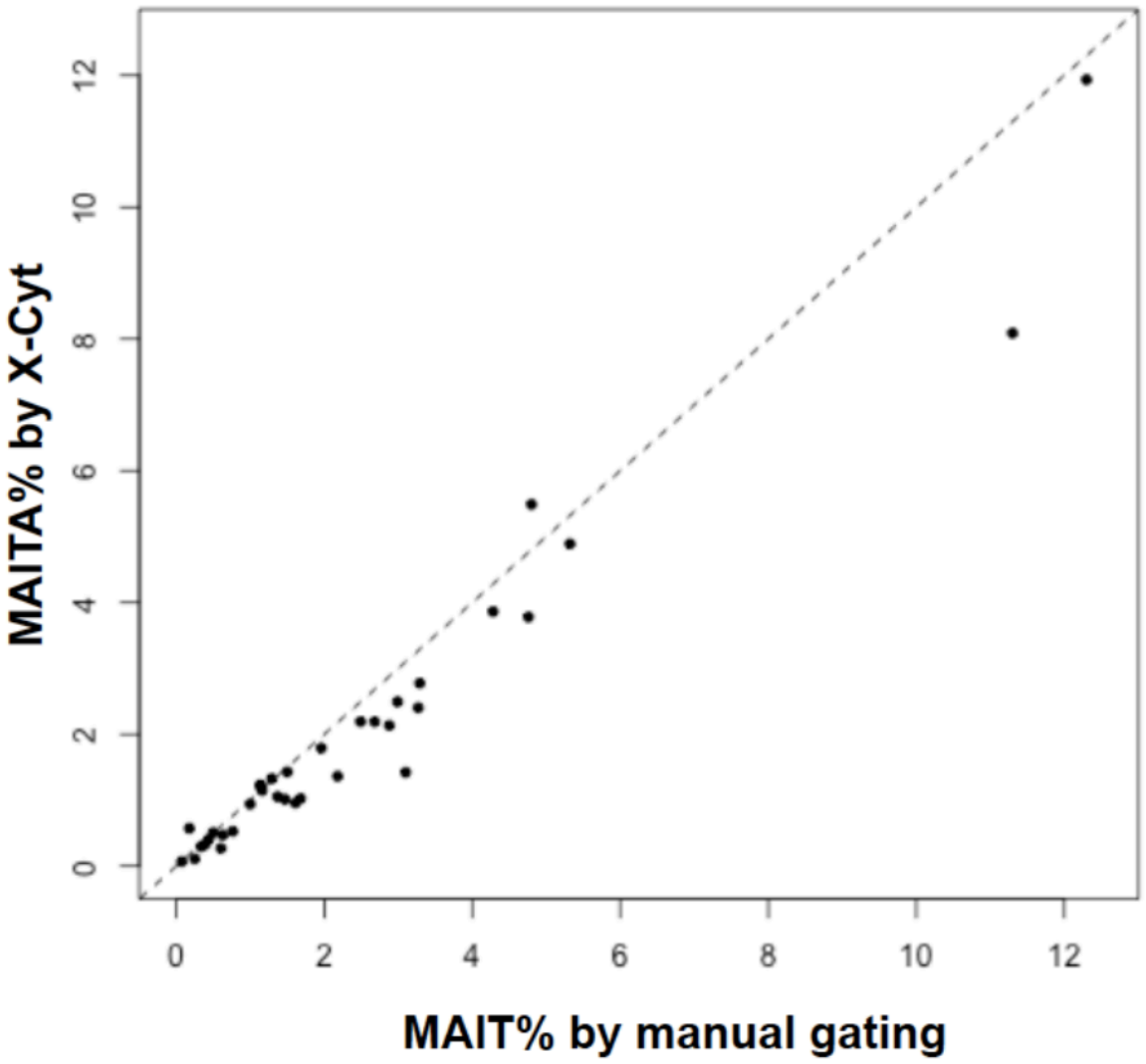


Figure 3.5. **Identification of MAIT cells.** We used X-Cyt to identify the mucosal associated invariant T (MAIT) cell population from PBMCs of 35 samples, and extracted its proportion as a percentage of all CD3+ cells. X-Cyt achieved a concordance of r = 0.98 ($p < 10^{-15}$, Pearson correlation) with manual gating in FlowJo®.

Figure 3.5. Identification of MAIT cells (Continued).



Characterizing rare iNKT cells

We applied X-Cyt to identify iNKT cell subsets from the peripheral blood sample in 110 individuals (**Table 3. 1**). We labeled PBMCs with a total of 19 surface and intracellular markers in nine separate panels. Each panel included the antibodies for CD3ε, CD4, and α-galactosylceramide (α-GalCer)-loaded CD1d tetramer, which are the standard markers used to identify iNKT cells [25], as well as two to three phenotypic or functional markers.

We configured X-Cyt to identify iNKT cells in three steps: 1) a 3-component bivariate normal mixture modeling using FSC and SSC to extract lymphocytes from PBMCs, 2) a 3-component bivariate normal mixture modeling using CD3ε and CD4 to identify $CD4^+$ and $CD4^-$ T cells, and 3) a threshold cutoff of five standard deviations above the mean of all lymphocytes in CD1d tetramer to identify the iNKT cells (**Figure 3.6A**).

We observed that iNKT cells were present in individuals at extremely low but highly variable abundances, ranging from 0.0033% to 0.89% of all $CD3\varepsilon^+$ cells (mean = 0.072%, median = 0.031%). The proportion of iNKT cells that are $CD4^+$ also ranged dramatically from 1.4% to 87% (mean = 39.5%). For comparison, an expert user manually gated and quantified iNKT cells and the $CD4^+$ subset in 36 of the 110 subjects using FlowJo™. Automated and manual results were almost perfectly concordant for the percentages of both iNKT cells (r = 0.99; **Figure 3.6B**) and the $CD4^+$ subset (r = 0.99).

Rapid and robust processing of cytometric data makes it feasible to discover population dynamics of immune cell subsets from profiling studies. We examined our cohort of 110 samples for interesting population dynamics of iNKT cellular subsets. First, we note that 11 of the 110 subjects had two visits separated by at least two months. In these subjects, we observed stable iNKT abundances and $CD4^+$ proportions over time (r = 0.99 and 0.98, respectively). We observed a negative correlation between the proportion of $CD4^+$ iNKT cells and the ($\log_{10}$) proportion of total

iNKT cells (r = -0.48, p = 8.2×10^{-8}, Pearson correlation; **Figure 3.6C)**. Also, women had significantly higher amounts of iNKT cells than men ($median_{female}$ = 0.038%, $median_{male}$ = 0.022%, p = 8.7×10^{-3}; Wilcoxon test). Finally, we observed that iNKT cell abundance correlated negatively with age (p = 0.014, Pearson correlation). The correlations between iNKT cell abundance and age, gender and $CD4^+$ proportion are independent of each other; they remain significant in a multivariate regression (**Table 3. 2)**. However, the proportion of $CD4^+$ iNKT cells was not correlated with gender (p = 0.12) or age (p = 0.75). Some of these trends had been observed in previous datasets [24]. With a larger sample size, we confirmed the correlations with statistical significance.

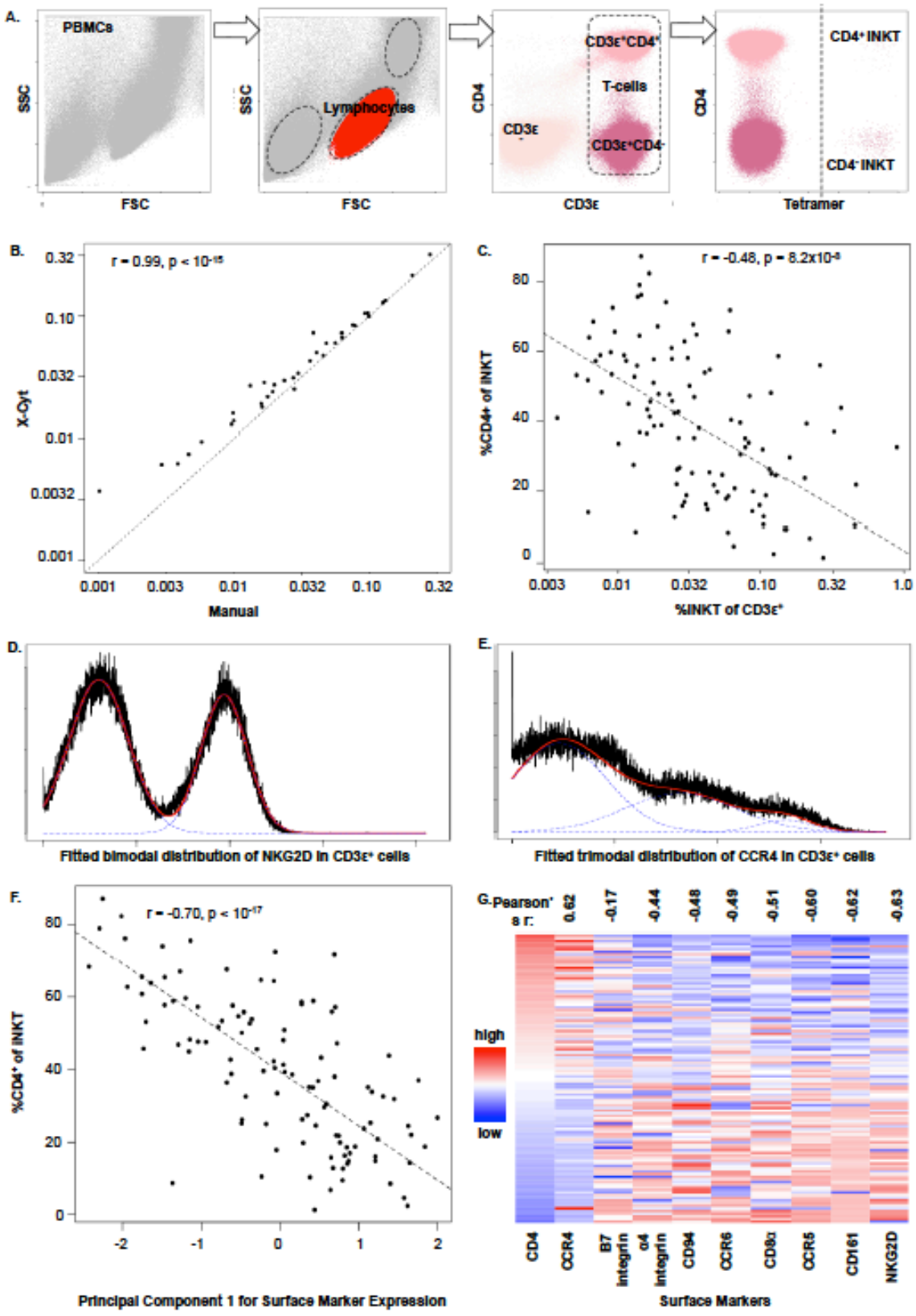
Downstream of successful identification and quantification of $CD4^+$ and $CD4^-$ iNKT cell subsets, we characterized the expression pattern of phenotypic markers in each. We quantified the expression of each marker in each subset by measuring the proportion of events with positive expression. We randomly sampled and pooled $CD3\varepsilon^+$ cells from all subjects to display the natural intensity distribution of each marker. Two examples of phenotypic markers are shown in **Figure 3.6D and 6E**. Eight of the 11 surface markers (α4, β7, CCR6, CCR5, CD8α, CD94, CD161, and NKG2D) and two of five cytokines (TNFα and IFNγ)

Table 3. 2. Multivariate regression for $\log_{10}$ iNKT cell proportion

	Estimate (95% CI)	***p*-value**
Intercept	-0.55 (-0.86 – -0.24)	5.6×10^{-4}
Age (per year)	-0.010 (-0.019 – -0.0021)	1.4×10^{-2}
Male Gender	-0.18 (-0.34 – -0.016)	3.2×10^{-2}
CD4 fraction	-1.24 (-1.61 – -0.86)	2.3×10^{-9}

Figure 3.6. iNKT cell identification. A) The partitioning scheme: FSC and SSC were clustered into three components to identify the lymphocyte population (red). Lymphocytes were subsequently clustered using CD3ε and CD4 into three components, namely the $CD3\varepsilon^-$, $CD3\varepsilon^+CD4^+$, and $CD3\varepsilon^+CD4^-$ populations. A cutoff of five standard deviations above the mean in αGalCer-loaded CD1d tetramer isolated the $tetramer^+$ iNKT cells (either $CD4^+$ or $CD4^-$). B) X-Cyt returned iNKT cell proportions highly concordant with manual gating (Pearson's r = 0.99). C) The $CD4^+$ proportion of iNKT cells correlates negatively with total iNKT abundance. Randomly sampled $CD3\varepsilon^+$ cells from all 110 samples were pooled to establish the intensity distribution of each phenotypic marker. Fitted distributions of D) NKG2D (bimodal) and E) CCR4 (trimodal) are shown. F) The first principal component of expression levels of the nine surface markers, which captured 31.8% of total variation, correlates strongly with the proportion of $CD4^+$ iNKT cells. G) The heatmap shows the correlations (also indicated by Pearson's r on top) of the nine surface markers' expression with $CD4^+$ proportion in iNKT cells. Each row represents one sample.

Figure 3.6. iNKT cell identification (Continued).



followed bimodal distributions. For each of these ten markers, we fitted a two-component mixture model. Using the mean and standard deviation of the pooled distribution, we calculated the proportions of iNKT cells belonging to the positive component in each sample using maximum *a posteriori* estimation. CCR4 followed a trimodal distribution, which we fitted with a three-component mixture model; we considered the sum of the higher two components to be the positive portion. Two surface markers (CD103 and IL23R) and three intracellular markers (IL4, IL13, and IL17A) showed negligible staining in all CD3ε$^+$ cells. These five markers were excluded from subsequent expression analyses.

After assessing the global pattern of phenotypic marker expression among the 110 subjects, we then applied principal component analysis to look for general trends. We observed that the first principal component captured 31.8% of the total variation correlated tightly with the proportion of CD4$^+$ iNKT cells ($r = -0.70$, $p < 10^{-17}$; see **Figure 3.6F**). We then examined the expression level of individual markers in all iNKT cells and confirmed that each was correlated with the proportion of CD4$^+$ iNKT cells, indicating biased expression in either the CD4$^+$ or the CD4$^-$ subset (**Figure 3.6G**, **Table 3. 3**). Specifically, CCR4 was preferentially expressed by the CD4$^+$ subset while all other surface markers were CD4$^-$- biased. Similarly, functional markers also showed iNKT subtype bias, where CD4$^-$ iNKT cells released much higher levels of TNFα and IFNγ upon PMA-ionomycin stimulation**.** These results suggest that variation in iNKT cell abundance, phenotypic marker expression, and functional response are all captured by CD4 expression, which is therefore a critical biomarker for iNKT function.

DISCUSSION

In this study, we profiled human iNKT cells, a rare immune cell type, in 110 samples of peripheral blood. In this large cohort, we showed that the quantity of iNKT cells was low

Table 3. 3. Differential expression of surface markers in $CD4^+$ and $CD4^-$ iNKT cells

Marker	Δ($CD4^-$ - $CD4^+$)	*p*-value
NKG2D	66.50%	2.8×10^{-19}
α4 integrin	58.40%	8.6×10^{-17}
CCR5	58.00%	2.7×10^{-18}
CCR6	38.40%	4.6×10^{-15}
CD161	36.80%	9.5×10^{-17}
CD8	31.50%	2.0×10^{-19}
CD94	30.10%	3.6×10^{-18}
β7 integrin	21.60%	1.5×10^{-7}
CCR4	-33.40%	4.9×10^{-18}
ΔTNFα*	27.8	2.6×10^{-9}
ΔIFNγ*	23.30%	1.4×10^{-7}

*Δ denotes the differential expression upon administration of PMA-ionomycin vs. DMSO (PMA-ionomycin – DMSO).

but variable in the population, showing increased quantity in females and a decreased quantity with age. Subsequently, we extracted patterns of expression of surface phenotypic markers and intracellular cytokines, observing differences between $CD4^+$ and $CD4^-$ iNKT subsets. By applying X-Cyt to characterize iNKT cells, we demonstrated the potential for robust and efficient automated population identification in a large-scale immunoprofiling study.

X-Cyt reliably discovers targeted populations with important advantages in terms of consistency and speed, which result from user-guidance and template-guided partitioning. We make the distinction between the goal of X-Cyt and that of existing automated cytometric analysis tools that are, in general, designed for exploratory studies. In exploratory studies, for example those

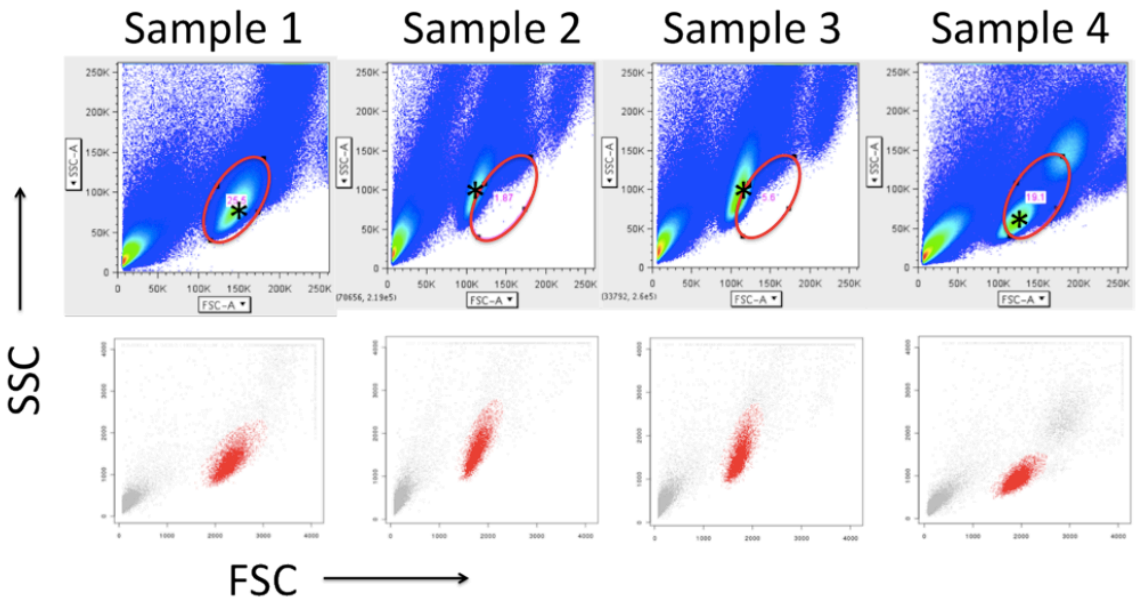
aiming to map developmental lineage of cells, populations are defined *de novo* in each sample. However, targeted studies focus on a specific cell type, often in large sample sizes. In such studies, we can assume samples in a cohort share a population structure defined by selected markers. X-Cyt allows the user to choose markers for defining cell types, the sequence of partitioning, and the resolution at which to partition, thus catering the analysis to the original intent of the experiment.

Both biological and technical variations often create notable shifts in fluorescence intensities, which complicate batch data analysis. However, the shifts rarely alter the relative spatial arrangement of populations. X-Cyt uses a template to capture this conserved structure, and uses expectation-maximization (EM) algorithm to optimize the fit for each sample independently, which gives the method substantial tolerance for intensity shifts. Via EM, the corresponding populations across samples are allowed to vary from the template, in terms of the "site" (the location parameter), "shape" (the covariance matrix), and "size" (the mixing proportion). In **Figure 3.7**, we illustrate samples in which a gate (i.e. in FlowJo®) requires manual adjustment in each sample, but X-Cyt automatically detects the shifted location via parameter optimization.

The use of a template confers two additional advantages over *de novo* clustering. First, the template serves as a guide for indexing emerging clusters (e.g. the T_{EM} and T_{CM} clusters are indexed as Clusters 1 and 5, respectively, in every sample), which eliminates the need for a separate alignment step that could potentially introduce additional error. If a population is present in the template but missing from a given sample, no event in the sample will be assigned and its proportion in that sample becomes "0". Next, by initializing the parameters to a close approximation of the optimal solution, the number of iterations needed to reach

Figure 3.7. The location of lymphocyte population among all PBMCs is expected to shift from sample to sample. In the top row, we show four samples displayed in forward and side-scatter in FlowJo®. The bright cluster of events in the middle of each panel, indicated by *, is the lymphocyte population. The red ovals are at the location of the gate determined by manually gating the first sample, and subsequently applied to all four samples, without manual adjustment. The true lymphocyte population often escapes the gate. Contrastingly, in the bottom row, the lymphocyte population in each sample is easily identified via clustering by expectation-maximization algorithm.

Figure 3.7. The location of lymphocyte population among all PBMCs is expected to shift from sample to sample (Continued).



convergence in the EM algorithm decreases by several orders of magnitude, substantially reducing computation time. To demonstrate, we compared the runtimes of clustering using X-Cyt with and without initialization by a template. Using approximately 200 megabytes of physical memory, X-Cyt was able to partition the $CD4^+$ T cell subset (~8,000 cells) into four clusters using three markers (CD62L, CD45RA, and CD45RO) in about 5 seconds per sample, compared to about 0.4 seconds with a template. In the MAIT cell study containing 500,000 cells per sample, X-Cyt partitioned a random subset of 100,000 cells into four clusters in two dimensions in approximately 45 seconds without a template. On the other hand, clustering a full sample of 500,000 cells required about 10 seconds when guided by a template.

Emerging cytometric technologies, such as CyTOF, can simultaneously measure more than 30 markers in a cell [26, 27], and facilitate precise characterization of the human immune system. For these studies, robust and versatile analytical methods will become indispensible. In addition to algorithms well suited for exploratory studies, there is a strong need for tools to replace gating-based manual analysis when conducting focused characterization of targeted cell types. X-Cyt presents an efficient and robust method for analyzing such high-throughput immunoprofiling datasets.

METHODS

Overview of flow cytometry datasets

$CD4^+$ memory T cell subset study. PBMCs were isolated from the whole blood of 236 healthy volunteers and depleted of non-$CD4^+$ T cells using magnetic-activated cell sorting kits. Cells were then stained with fluorophore-conjugated antibodies against CD45RO, CD45RA, and CD62L.

Mucosal associated invariant T cell study. PBMCs were isolated from the whole blood of 40 healthy. Cells were then stained with fluorophore-conjugated antibodies against CD3, CD45, and Vα7.2, and CD161.

iNKT cell study. PBMCs were obtained from the blood of 110 healthy volunteers. From PBMCs, iNKT cells were stained with fluorophore-conjugated antibodies against 14 cell surface markers, and five intracellular cytokines following PMA-ionomycin administration.

Detailed experimental protocols can be found in the **Experimental Methods**. Flow cytometric data were exported from FlowJo™ as text files after compensation and transformation in the "channel number" format.

Implementation of X-Cyt.

X-Cyt utilizes the EMMIX-skew package developed by Wang *et al.* previously written and published (1, 2) for mixture modeling. X-Cyt is currently implemented in R. There are two distinct modules for cell population identification in X-Cyt:

1. ***MixtureModel*** models cytometry data with multivariate mixture distributions of user-specified markers. For this, the user may choose to run a set of samples each *de novo* (i.e. for a trial run), or specify a template to guide batch analysis.
2. ***StandardDeviationCutoff*** defines and applies univariate threshold cutoff for selected marker(s).

To run each module, the user specifies input parameters as text files. Currently, R packages for X-Cyt, as well as a sample iNKT and $CD4^+$ T_{EM} cell datasets and user input files, are available for download at http://www.broadinstitute.org/mpg/xcyt/. X-Cyt is currently under development as a graphical user interface to facilitate easier use of the software.

Multivariate normal distribution.

The multivariate normal distribution is described by Y ~ N(μ, Σ), where μ is the mean vector, and Σ the covariance matrix. The probability density function of an observation is given by

$$f(y;\Theta) = \sum_{g=1}^{G} \pi_g f_g(y;\mu_g,\Sigma_g)$$

where $\pi_{1,\dots,}\pi_g$ denote non-negative mixing proportions of G components, and sum to 1. Θ denotes the collection of unknown parameters $(\mu_1, \dots, \mu_g, \Sigma_1, \dots, \Sigma_g)$, and is estimated by the maximum likelihood method via EM.

Estimation-Maximization (EM) algorithm.

For the application of the EM algorithm, the observed-data vector $(y_1^T, \dots, y_n^T)^T$ is regarded as incomplete. The component-label indicator variables z_{jh} are introduced, where z_{jh} is defined to be one or zero based on weather y_j arose from the g^{th} component of the mixture model $(g = 1, \dots, G; j = 1, \dots . n)$. Letting $z_j = (z_{j1} \dots, z_{jG})^T$, the complete-data vector x_c is given by $x_c = (x_1^T \dots, x_n^T)$, where $x_1 = (y_1^T, z_1^T)^T, \dots, x_n = (y_n^T, z_n^T)^T$ are taken to be independent and identically distributed with $z_1, \dots, z_n$ being independent realizations from a multinomial distribution consisting of one draw on G categories with respective probabilities $p_1, \dots, p_G$. That is,

$$p_1, \dots, p_G \sim Mult_G(1, p), \text{ where } p = (p_1, \dots, p_G)^T.$$

For this specification, the complete-data log likelihood is

$$l_c = \sum_{j=1}^{n} \sum_{g=1}^{G} -\frac{1}{2}[k \log(2\pi) + \log|\sum_h| + (y_j - \mu_g)^T \sum_g^{-1}(y_j - \mu_g)].$$

The EM algorithm proceeds iteratively in two steps: estimation (E) and maximization (M).

The E step comprises of computing the following conditional expectations, using the current fit for the vector of unknown parameters Θ:

$$E(z_{jg}|y_j) = \tau_{jg} = \frac{p_g f_g(y_j;\mu_g,\Sigma_g)}{\sum_g p_g f_g(y_j;\mu_g,\Sigma_g)},$$

While the M step updates the estimates of the parameters, using the equations,

$$p_g = \frac{1}{n}\sum_{j=1}^{n} \tau_{jg},$$

$$p_g = \sum_{j=1}^{n}(y_j - \mu_g)(y_j - \mu_g)^T e_{jg}\tau_{jg} / \sum_{j=1}^{n} \tau_{jg},$$

$$\mu_g = \sum_{j=1}^{n} y_j \, e_{jg}\tau_{jg} / \sum_{j=1}^{n} e_{jg}\tau_{jg}.$$

The E and M steps are alternated repeatedly until the likelihood changes by a predefined arbitrarily small amount, and the process has reached convergence.

Phenotypic and functional marker expression.

For a phenotypic marker with a multimodal distribution, we assume that it is comprised of multiple components, where each component, *c*, follows the normal distribution with a mean μ_c, and standard deviation σ_c. X-Cyt fits a one-dimensional n-component mixture model. For each individual sample, X-Cyt then calculates the proportion of cells in each component by maximum a posteriori estimation. Given the n components each described by $N(\mu, \sigma^2)$, and the vector of observed data, D, the vector of mixing proportions, $p_1, \ldots \; p_C$, is estimated by minimizing:

$$\sum_{c=1}^{C} -\log\left[p_c \Phi(D, \mu_c, \sigma_c)\right]$$

where Φ is the probability density function defined by $\Phi(\mathrm{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.

INKT cell abundance and CD4$^+$INKT percentage calculations.

Each of the 110 individuals has nine measurements, one from each surface marker panel (including the lineage marker-only panel). Each vector of nine observed measurement for an individual, i, was considered to be the linear combination of $\alpha_i S + \beta_i P$, where S and P are categorical factors subject, and panel, respectively. The weight averages, the vector of $\{\alpha_1, \ldots, \alpha_{110}\}$, were then estimated through regression model.

The concordance between manual analysis by an expert flow user and automated analysis by X-Cyt were tested using correlation coefficients.

We calculated the significance of Pearson's correlations of iNKT cell abundance with CD4 proportion and age were correlation coefficients with the cor.test() function in R. P-values for Pearson test are calculated based the standard test statistic *s*, where $s = r * \sqrt{n-2}/\sqrt{1-r^2}$, where *r* is the correlation coefficient, and *n* is the number of pairs of data. The *x* be the cumulative probability of *s* under a *t* distribution with *n*-2 degrees of freedom, and the p-value is 2*(1-*x*). This is also implemented by the same R function. We calculated the difference of iNKT cell abundance in men and women with the 2-sample Wilcoxon test, using wilcox.test() function in R.

Additionally, we constructed a multivariate regression model of INKT ~ Age + (factor) Gender + CD4%.

Characterization of phenotypic and functional markers in INKT cells

Surface markers. We pooled random CD3ε$^+$ T cells from all 110 subjects in order to establish the overall distribution of each phenotypic marker. Eight markers (α4, β7, CCR4, CCR5, CD8, CD94, CD161, and NKG2D) showed clear bimodal distributions of intensities; CCR6 showed a trimodal distribution; two markers (CD103 and IL23R) followed unimodal distributions. For bimodal markers, X-Cyt fitted one-dimensional 2-component mixture models to estimate the means of the negative and positive subsets. Based on the estimated parameters, X-Cyt estimated the proportion of cells positive each cell population (CD3ε$^+$ T cells, INKT cells, as well as CD4$^+$ and CD4$^-$ subsets of INKT cells) in each sample. For CD103 and IL23R, which were unimodal, we considered cells with expression higher than three standard deviations above the mean as positive.

Intracellular cytokines. IFNγ and TNFα were bimodally distributed; we fitted a two-cluster mixture model to each based on post-PMA-ionomycin samples. IL4, IL13, and IL17a followed

unimodal distributions with rare outliers; we imposed a threshold of three standard deviations from the mean of the DMSO-stimulated samples to isolate positive expressers.

Principal component analysis (PCA) of iNKT cell surface markers.

CD103 and IL23R were omitted from PCA analysis, as they showed negligible staining in all cells. We constructed a 9x110 matrix of proportions of INKT cells positive for the surface markers in all samples. Each expression vector was normalized to have a mean of 0 and variance of 1. We reported the Pearson correlation between the first principal component and the proportions of $CD4^+$ INKT cells.

Marker expression bias in $CD4^+$ and $CD4^-$ INKT cells. The percentages of in $CD4^+$ and $CD4^-$ INKT cells that express each of the 16 phenotypic markers in each individual were compared using a paired nonparametric Wilcoxon test. The "CD4-biased expression" reported in Table 3. 1 was calculated as the difference of the medians. We omitted markers with negligible staining.

Data partitioning

X-Cyt partitions the data with user-designated differentiation markers. At each step of partitioning, the user can opt to use multivariate mixture-modeling or univariate cutoffs to identify outliers.

Multivariate mixture modeling. X-Cyt fits a given number of multivariate components to a sample via expectation-maximization (EM) algorithm, as previously described in "Automated high-dimensional flow cytometric data analysis" [5]. The user can specify three input parameters: 1) the markers used for clustering, 2) the number of expected clusters, and 3) the distribution type (multivariate normal, skew-normal, *t*, or skew-*t*; default is normal). Given m differentiation markers, and g clusters, X-Cyt models a given sample as an m-variate mixture of g components using EM algorithm initiated by k-means clustering. Upon convergence, each cluster is described by a location

parameter, a covariance matrix that describes its multidimensional shape, as well as a mixing proportion. Each event in the sample is assigned membership to one of the *g* clusters.

Trial analysis in representative samples

Using a small test set of random samples, the user sets up the partitioning scheme, optimizes input parameters, and chooses a template.

Select test samples. The user randomly selects a small subset of samples from the cohort to serve as test samples. Assuming that a target population is present in *f*% of all samples, the chances of encountering this population at least once among *N* random test samples at a 95% confidence is described by $(1-f)^N = 0.05$. Therefore, there is a 95% chance that a 20% population will be observed at least once in 14 samples; a 50% population will be observed at least once in 5 samples; and a 90% population will be observed at least once in 2 samples. A table of recommended size of test-sets is available in **Table 3. 4**.

Select differentiation markers. The user should select the subset of markers that most efficiently distinguishes the cell type(s) of interest from the rest of the events. The user often already has selected a set of differentiating markers while designing the marker panel for the experiment. For example, one would use CD3, CD45 (RA/RO), and CD62L to identify naïve T cells in PBMCs. On the other hand, if certain markers in the panel are assayed for the purpose of characterizing the phenotype and function rather than differentiating cell types, (*e.g.* certain intracellular cytokines and chemokine receptors), they should be excluded in this step.

Select the number of clusters (g). In the trial analysis, the user should evaluate the partitioning result of each sample from testing a range of *g*. For example, given *k* differentiation markers, it is reasonable to test a range of *k* to 2^k clusters. The user reviews the output clusters and defines the optimal *g* as one that most accurately captures the population of interest as one cluster

without including undesired events or spuriously splitting the population into more than one cluster. Often, the target population remains stable as one coherent cluster over a small range of g. In this case, the lowest g is recommended to minimize computation time.

Table 3. 4. Recommended size for test-sets (95% CI)

Prevalence of population (%)	Test samples needed* (>=N)
99	1
95	1
90	2
85	2
80	2
75	3
70	3
65	3
60	4
55	4
50	5
45	6
40	6
35	8
30	9
25	11
20	14
15	19
10	29
5	59

Standard deviation thresholds. For rare cell types with extreme intensities in one marker *M*, it is most efficient to first partition the sample to coarser-grained clusters, based on other differentiation markers, and then distinguish the rare events in *M* using a cutoff by standard deviation threshold. For example, to identify Tregs from PBMCs using a panel of CD3, CD4, CD25, and Foxp3, one may first partition all events with CD3, CD4, and CD25 to extract $CD3^+CD4^+CD25^+$ activated T cells, and then apply a threshold cutoff in Foxp3 to extract the Tregs.

Guided cohort analysis by template

X-Cyt uses a user-approved template selected from the trial analysis to guide the partitioning of all subsequent samples. The template serves as the initial parameters and as the indexing guide. Instead of using a *k*-means initialization, X-Cyt initializes each sample's mixture model parameters to that of the template's, upon which EM algorithm iterates and converges quickly.

Phenotypic and functional marker characterization.

For each marker, we report the percentage of cells with positive expression. We construct a pooled sample of random events from all samples in the dataset, thus establishing a "reference" fluorescence intensity distribution for each marker.

For multimodal markers, we assume the intensity distribution comprises of n normal components. We fit a one-dimensional mixture model on the pooled sample and then estimate the proportion of cells in each cell population positive for the marker in each sample.

For a unimodal marker, we specify a standard deviation threshold. We report the proportion of cells that express the marker above the threshold.

EXPERIMENTAL METHODS

Subjects.

In the main iNKT cell profiling study, 110 individuals (62 females, 48 males) were consented and enrolled into the study through the Phenogenetics Project at Brigham and Women's Hospital. Subjects' ages ranged from 19 to 53 years of age with an average age of 31 years. Eleven subjects returned for a second study visit at least 3 months after their initial visits.

For the automated method validation study for identifying $CD4^+$ T_{EM} cells, a separate enrollment of 216 individuals (128 females, 88 males) was conducted. These subjects' ages ranged from 19 to 57 years of age with an average age of 31.3 years. Thirty-four subjects returned for a second study visit at least 3 months after their initial visits.

For the validation study for identifying MAIT cells, a separate enrollment of 30 individuals was conducted.

All subjects for all studies were healthy individuals of Caucasian, non-Hispanic descent **(Table 3. 1).**

Buffers and media.

Peripheral blood mononuclear cells (PBMCs) were washed with a cold, divalent cation-free Hycole Dulbeccos (Thermo Scientific) phosphate buffered solution (PBS). PBMCs were cultured in "Basic Human Media", which is RPMI 1640 (Gibco) containing 10% Hyclone fetal bovine serum (Thermo Scientific), 5% BenchMark fetal bovine serum (Gemini Bio-Products), and supplemented with the following items and their final concentrations or volumes : 30 mM HEPES, 100 U/mL penicillin, 100 μg/mL streptomycin, 1 mM L-glutamine, 0.5 mM sodium pyruvate, 0.055 mM β-mercaptoethanol, 2.5 mL of an essential amino acid solution (Gibco; catalog #11130), and 2.5 mL of a non-essential amino acid solution (Gibco; catalog #11140). PBMCs were stained in "FACS buffer", which is the aforementioned PBS solution with 0.5% BenchMark fetal bovine serum (Gemini Bio-Products) and 2 mM EDTA (Gibco).

Blood collection and PBMC isolation

For the iNKT cell study, subjects were instructed to fast overnight prior to the blood draw. 50 mL of blood was collected from each subject into plastic tubes spray-coated with sodium heparin (BD). For the validation study, 30 mL of non-fasting blood was collected from each subject into plastic tubes spray-coated with EDTA (BD). In both studies, the blood was carefully layered over Ficoll-Paque PLUS (GE Healthcare) at a ratio of 4:3 and centrifuged at 2000 rpm for 30 minutes to isolate peripheral blood mononuclear cells (PBMCs). PBMCs were washed twice with cold PBS and filtered through a 70 μm nylon mesh. The time from blood collection to the Ficoll procedure was always less than 6 hours. PBMCs for the INKT cell study were resuspended at a final concentration of $3x10^6$ cells/mL in Basic Human Media and kept on ice until labeling with antibodies. PBMCs for the validation study proceeded directly to MACS depletion.

MACS depletion

In the validation study, filtered PBMCs were magnetically depleted of non-$CD4^+$ cells, such as $CD8^+$ T cells, monocytes, neutrophils, eosinophils, B cells, dendritic cells, NK cells, granulocytes, γ/δ T cells, and red blood cells using a $CD4^+$ T cell isolation MACS kit (Miltenyi). The enriched $CD4^+$ T cells were then kept in FACS buffer on ice until labeling with antibodies

iNKT cell stimulation

Aliquots of each donor's PBMCs were cultured in Basic Human Media in the presence of 40nM IL-2 overnight at 37°C. The following morning, cells were given monensin (Golgi-Stop; BD) and either PMA (25 ng/mL) and ionomycin (1 μg/mL) dissolved in DMSO or DMSO only for 4 hours at 37°C. Following incubation, cells were removed from the plates, washed with FACS buffer, and then labeled with the cell surface lineage markers for 40 minutes on ice in the dark. After two washes with FACS buffer, the cells were fixed with 4% paraformaldehyde for 20 minutes on ice in

the dark. After two washes with FACS buffer, the cells were then permeabilized with a solution of Perm/Wash (BD) buffer for 15 minutes on ice in the dark. Cells were washed twice with deionized, distilled water. Cells were then labeled with intracellular cytokine staining antibodies for 40 minutes on ice in the dark. Cells were washed twice with Perm/Wash buffer and then resuspended in FACS buffer immediately prior to analysis.

Fluorescence antibodies and iNKT tetramer

In the iNKT cell study, iNKT cells were identified by positive binding to allophycocyanin (APC)-conjugated, α-galactosylceramide (α-GalCer)-loaded Cd1d tetramers (NIH Tetramer Facility) as well as positive labeling by phycoerythrin-Cy7 (PE-Cy7)-conjugated anti-CD3ε antibodies (BD). The following cell surface marker antibodies were used: Pacific Blue (PacBlue)-conjugated anti-CD4 (BD), fluorescein isothiocyanate (FITC)-conjugated anti-α4 integrin (BD), anti-CD161 (Biolegend), anti-CD8α (eBioscience) and anti-CCR5 (BD), PE-conjugated anti-β7 integrin (eBioscience), anti-NKG2D (eBioscience), anti-CD103 (BD), and anti-CD94 (eBioscience), and peridinin-chlorophyll-protein-Cy5.5 (PerCP-Cy5.5)-conjugated anti-CCR6 (Biolegend), anti-IL23R (R&D), and anti-CCR4 (Biolegend). The following intracellular cytokine staining antibodies were used: FITC-conjugated anti-IL13 (eBioscience), PE-conjugated anti-interferon-γ (eBioscience), and anti-IL4 (eBioscience), and PerCP-Cy5.5-conjugated anti-IL17A (eBioscience), and TNF-α (eBioscience). Due to the large number of markers assayed, the markers were divided into several panels. The following antibodies were used in the validation study: eFluor450-conjugated anti-CD45RA (eBioscience), APC-conjugated anti-CD45RO (eBioscience), and PE-conjugated anti-CD62L (eBioscience). All staining was done on ice for 40 minutes in FACS buffer.

Data acquisition and manual analysis

For the iNKT cell study, samples were analyzed on a FACSCantoII cytometer (BD). Since INKT cells were a rare cell population, at least 500,000 events were recorded for each sample. Manual single-color and unstained compensation was performed before each set of experiments using spare PBMCs. For the validation study, samples were analyzed on a FACSAriaII SORP cytometer (BD). 10,000 events were recorded for each sample. All manual analysis of flow cytometry data from both studies were performed using FlowJo (version 8.8.7; Treestar). Positivity for each marker analyzed was based on an empty, unstained control or DMSO-only samples in the case of the INKT cell stimulation portion of the study.

Acknowledgements

X.H. is partially supported by the IDEA^2 program at the Massachusetts Institute of Technology. P.J.B. is supported by a career development award from the American Academy of Allergy, Asthma & Immunology ARTrust. M.B.B. is supported by research grants from the National Institutes of Health (AI063428, AI028973, and DK057521) and the American Diabetes Association (7-12-IN-07). SR is supported by research grants from the Arthritis Foundation, the National Institutes of Health (U01HG0070033 and 5K08AR055688), and the Harvard University Milton Foundation. We would like to acknowledge the PhenoGenetic Project at Brigham and Women's Hospital for providing blood samples. We thank Dr. Joshua Randall for assistance in preparing the software package. We thank Dr. Vijay Kuchroo for helpful discussions.

REFERENCES

1. Blumberg, R.S., et al., *Unraveling the autoimmune translational research process layer by layer.* Nat Med, 2012. **18**(1): p. 35-41.

2. Perfetto, S.P., P.K. Chattopadhyay, and M. Roederer, *Seventeen-colour flow cytometry: unravelling the immune system.* Nat Rev Immunol, 2004. **4**(8): p. 648-55.

3. Cheung, R.K. and P.J. Utz, *Screening: CyTOF-the next generation of cell detection.* Nat Rev Rheumatol, 2011. **7**(9): p. 502-3.

4. Aghaeepour, N., et al., *Critical assessment of automated flow cytometry data analysis techniques.* Nat Methods, 2013.

5. Pyne, S., et al., *Automated high-dimensional flow cytometric data analysis.* Proc Natl Acad Sci U S A, 2009. **106**(21): p. 8519-24.

6. Lo, K., R.R. Brinkman, and R. Gottardo, *Automated gating of flow cytometry data via robust model-based clustering.* Cytometry A, 2008. **73**(4): p. 321-32.

7. Finak, G., et al., *Merging mixture components for cell population identification in flow cytometry.* Adv Bioinformatics, 2009: p. 247646.

8. Chan, C., et al., *Statistical mixture modeling for cell subtype identification in flow cytometry.* Cytometry A, 2008. **73**(8): p. 693-701.

9. Qian, Y., et al., *Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data.* Cytometry B Clin Cytom, 2010. **78 Suppl 1**: p. S69-82.

10. Aghaeepour, N., et al., *Rapid cell population identification in flow cytometry data.* Cytometry A, 2011. **79**(1): p. 6-13.

11. Sugar, I.P. and S.C. Sealfon, *Misty Mountain clustering: application to fast unsupervised flow cytometry gating.* BMC Bioinformatics, 2010. **11**: p. 502.

12. Naumann, U., G. Luta, and M.P. Wand, *The curvHDR method for gating flow cytometry samples.* BMC Bioinformatics, 2010. **11**: p. 44.

13. Zare, H., et al., *Data reduction for spectral clustering to analyze high throughput flow cytometry data.* BMC Bioinformatics, 2010. **11**: p. 403.

14. Qiu, P., et al., *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.* Nat Biotechnol, 2011. **29**(10): p. 886-91.

15. Bendelac, A., P.B. Savage, and L. Teyton, *The biology of NKT cells.* Annu Rev Immunol, 2007. **25**: p. 297-336.

16. Brigl, M. and M.B. Brenner, *CD1: antigen presentation and T cell function.* Annu Rev Immunol, 2004. **22**: p. 817-90.

17. Kronenberg, M., *Toward an understanding of NKT cell biology: progress and paradoxes.* Annu Rev Immunol, 2005. **23**: p. 877-900.

18. Gumperz, J.E., et al., *Functionally distinct subsets of CD1d-restricted natural killer T cells revealed by CD1d tetramer staining.* J Exp Med, 2002. **195**(5): p. 625-36.

19. Lawson, V., *Turned on by danger: activation of CD1d-restricted invariant natural killer T cells.* Immunology, 2012. **137**(1): p. 20-7.

20. Snyder-Cappione, J.E., et al., *A comprehensive ex vivo functional analysis of human NKT cells reveals production of MIP1-alpha and MIP1-beta, a lack of IL-17, and a Th1-bias in males.* PLoS One, 2010. **5**(11): p. e15412.

21. Carvalho, K.I., et al., *Skewed distribution of circulating activated natural killer T (NKT) cells in patients with common variable immunodeficiency disorders (CVID).* PLoS One, 2010. **5**(9).

22. O'Reilly, V., et al., *Distinct and overlapping effector functions of expanded human CD4+, CD8alpha+ and CD4-CD8alpha- invariant natural killer T cells.* PLoS One, 2011. **6**(12): p. e28648.

23. Pariente, B., et al., *Activation of the receptor NKG2D leads to production of Th17 cytokines in CD4+ T cells of patients with Crohn's disease.* Gastroenterology, 2011. **141**(1): p. 217-26, 226 e1-2.

24. Montoya, C.J., et al., *Characterization of human invariant natural killer T subsets in health and disease using a novel invariant natural killer T cell-clonotypic monoclonal antibody, 6B11.* Immunology, 2007. **122**(1): p. 1-14.

25. Kawano, T., et al., *CD1d-restricted and TCR-mediated activation of valpha14 NKT cells by glycosylceramides.* Science, 1997. **278**(5343): p. 1626-9.

26. Ornatsky, O., et al., *Highly multiparametric analysis by mass cytometry.* J Immunol Methods, 2010. **361**(1-2): p. 1-20.

27. Bendall, S.C., et al., *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.* Science, 2011. **332**(6030): p. 687-96.

CHAPTER 4

Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in $CD4^+$ effector memory T cells

Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells

Xinli Hu[1-6*], Hyun Kim[1-3*], Towfique Raj[2,4,7], Patrick J. Brennan[1], Gosia Trynka[1-4], Nikola Teslovich[1,2], Kamil Slowikowski[1-5], Wei-Min Chen[8], Suna Onengut[8], Clare Baecher-Allan[9], Philip L. De Jager[4,7], Stephen S. Rich[8], Barbara E. Stranger[10-11], Michael B. Brenner[1], Soumya Raychaudhuri[1-4,12]

1. Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
2. Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
3. Partners Center for Personalized Genetic Medicine, Boston, MA, USA
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
5. Harvard Medical School, Boston, MA USA
6. Harvard-MIT Division of Health Sciences and Technology, Boston, MA USA
7. Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA
8. Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
9. Department of Dermatology/Harvard Skin Disease Research Center, Brigham and Women's Hospital, Boston, MA, USA
10. Section of Genetic Medicine, University of Chicago, Chicago, IL USA
11. Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL USA
12. Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK

* These authors contributed equally to this work.

Correspondence:

Soumya Raychaudhuri (soumya@broadinstitute.org)

Originally published as:

Hu, X., H. Kim, T. Raj, P. J. Brennan, G. Trynka, N. Teslovich, K. Slowikowski, W. M. Chen, S. Onengut, C. Baecher-Allan, P. L. De Jager, S. S. Rich, B. E. Stranger, M. B. Brenner and S. Raychaudhuri (2014). "Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells." PLoS Genet **10**(6): e1004404.

ABSTRACT

Genome-wide association studies (GWAS) and subsequent dense-genotyping of associated loci identified over a hundred single-nucleotide polymorphism (SNP) variants associated with the risk of rheumatoid arthritis (RA), type 1 diabetes (T1D), and celiac disease (CeD). Immunological and genetic studies suggest a role for CD4-positive effector memory T (CD^+ T_{EM}) cells in the pathogenesis of these diseases. To elucidate mechanisms of autoimmune disease alleles, we investigated molecular phenotypes in $CD4^+$ effector memory T cells potentially affected by these variants. In a cohort of genotyped healthy individuals, we isolated high purity $CD4^+$ T_{EM} cells from peripheral blood, then assayed relative abundance, proliferation upon T cell receptor (TCR) stimulation, and the transcription of 215 genes within disease loci before and after stimulation. We identified 46 genes regulated by *cis*-acting expression quantitative trait loci (eQTL), the majority of which we detected in stimulated cells. Eleven of the 46 genes with eQTLs were previously undetected in peripheral blood mononuclear cells. Of 96 risk alleles of RA, T1D, and/or CeD in densely genotyped loci, eleven overlapped *cis*-eQTLs, of which five alleles completely explained the respective signals. A non-coding variant, $rs389862^A$, increased proliferative response ($p = 4.75x10^{-8}$). In addition, baseline expression of seventeen genes in resting cells reliably predicted proliferative response after TCR stimulation. Strikingly, however, there was no evidence that risk alleles modulated $CD4^+$ T_{EM} abundance or proliferation. Our study underscores the power of examining molecular phenotypes in relevant cells and conditions for understanding pathogenic mechanisms of disease variants.

INTRODUCTION

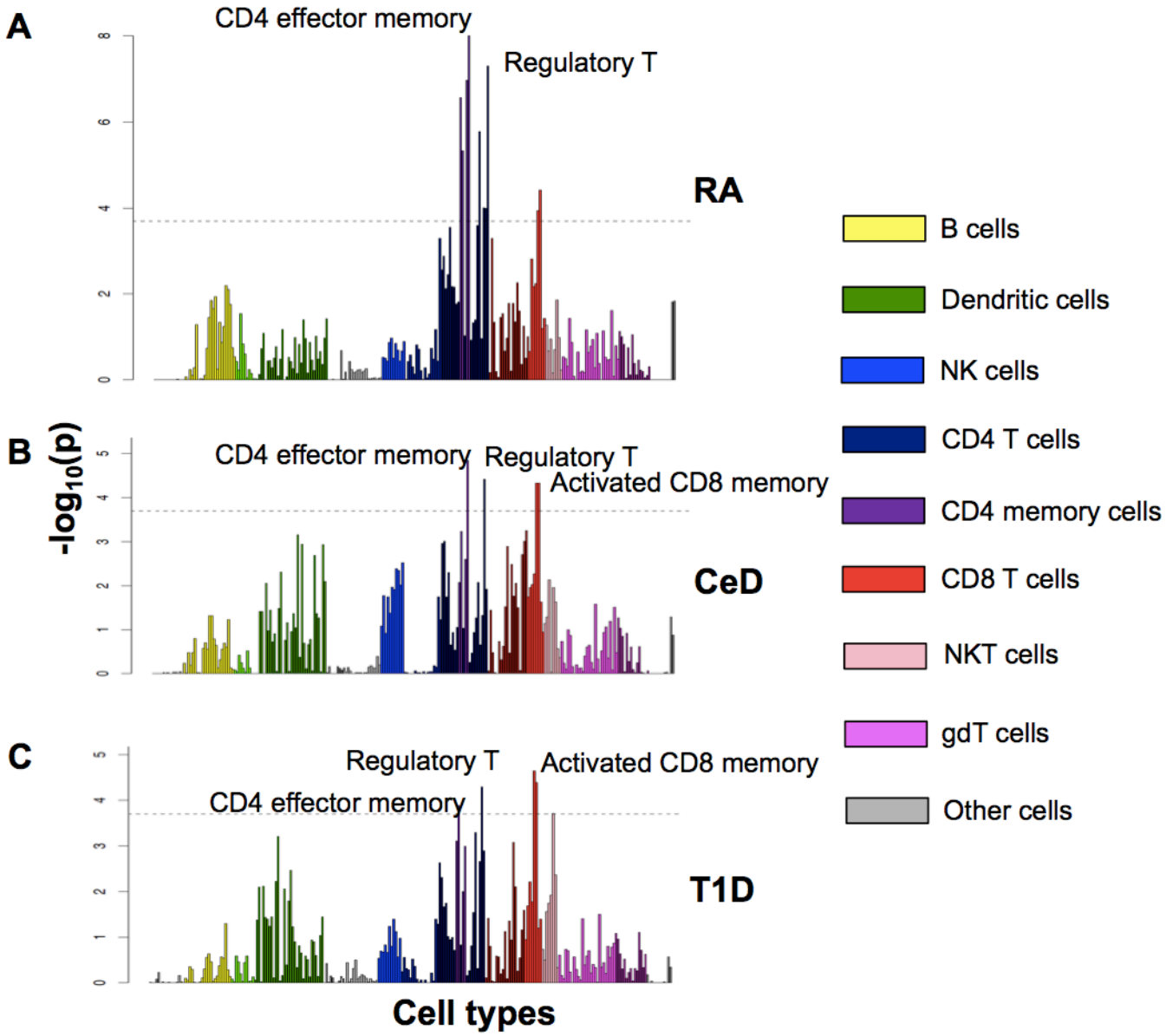
Memory T cells are an important component of the adaptive immune system. They circulate between lymphoid organs, blood, and peripheral tissues, and facilitate faster and more aggressive immune response to antigens after re-exposure. CD4-positive effector memory T ($CD4^+$ T_{EM}) cells are known to migrate to peripheral sites of inflammation upon activation, and rapidly produce both Th1 and Th2 cytokines [1]. Investigators have long suggested their involvement in autoimmune diseases including rheumatoid arthritis (RA), type I diabetes (T1D), and celiac disease (CeD) [2-5]. However, whether changes in cell population subsets and functions are causal or reactive to disease is uncertain. One strategy to answer this question is to examine potential intermediate molecular phenotypes, and identify those modulated by genetic variants. In order to understand the pathogenic roles of $CD4^+$ T_{EM} cells in autoimmunity, we aimed to characterize the variation in their phenotypic and functional markers in a healthy population, and to identify whether these markers intersect with the genetic basis for autoimmunity.

The majority of autoimmune disease risk variants are located in non-coding regions of the genome. It is reasonable to hypothesize that a subset of them causes disease by altering gene regulatory mechanisms as expression quantitative trait loci (eQTL) [6-9]. So far, studies of gene regulation have largely been carried out in cell lines and primary resting blood cells including undifferentiated $CD4^+$ T cells, B cells, monocytes, and dendritic cells [8, 10-12]. However, to understand the pathogenic mechanisms of risk variants, especially when studying the immune system where cells are highly diverse and functionally specialized, it is crucial to focus on relevant cell types and stimulated cellular states.

We have previously shown that genes within RA risk loci were most specifically expressed in $CD4^+$ T_{EM} cells, compared to more than 200 other immune cell types of various lineages and developmental stages ($p = 1.00 \times 10^{-8}$; **Figure 4.1**) [13]. Celiac disease and T1D loci were also enriched for genes specifically expressed in $CD4^+$ T_{EM} cells ($p = 1.43 \times 10^{-5}$ and 1.29×10^{-4}, respectively; **Figure 4.1**) [13]. Non-coding single nucleotide polymorphisms (SNPs) associated with RA significantly overlap chromatin marks of

Figure 4.1. Enrichment of cell-specific expression of genes within risk loci. As described in Hu et al. *AJHG* 2011, **A)** genes within risk loci of RA were the most specifically expressed in $CD4^+$ T_{EM} cells ($p = 1.00 \times 10^{-8}$) followed by signal in regulatory T cells ($p = 5.00 \times 10^{-8}$). **B)** Genes within CeD were also the most strongly enriched in CD4 TEM cells ($p = 1.43 \times 10^{-5}$) followed by regulatory T cells ($p = 3.78 \times 10^{-5}$). **C)** In T1D, CD8 memory T cells showed the strongest enrichment ($p = 2.26 \times 10^{-5}$), followed by regulatory T cells ($p = 5.13 \times 10^{-5}$) and $CD4^+$ T_{EM} cells ($p = 1.29 \times 10^{-4}$).

Figure 4.1. Enrichment of cell-specific expression of genes within risk loci (Continued).



trimethylation of histone H3 at lysine 4 (H3K4me3) specifically in $CD4^+$ regulatory and memory T cells (p = $1.3x10^{-4}$ and $7.0x10^{-4}$, respectively) [14].

We hypothesized that the risk alleles of these conditions might influence $CD4^+$ T_{EM} quantitative molecular phenotypes: 1) the expression of immune-related genes; 2) the relative abundance of $CD4^+$ T_{EM} cells in peripheral blood; and 3) proliferative response to T cell receptor (TCR) stimulation. To this end, we undertook a large immunoprofiling study in a healthy population of 174 European-descent individuals, by cross-analyzing genotype, transcription, abundance, and proliferative response in primary $CD4^+$ T_{EM} cells. Because the post-stimulation activation of $CD4^+$ T_{EM} cells is presumably crucial for their autoimmune response, we assayed cells not only at rest, but also after T cell receptor (TCR) stimulation with anti-CD3/CD28 beads. As such, this study is the first to our knowledge to map expression quantitative trait loci and examine immunological cellular traits in primary $CD4^+$ T_{EM} cells under multiple states.

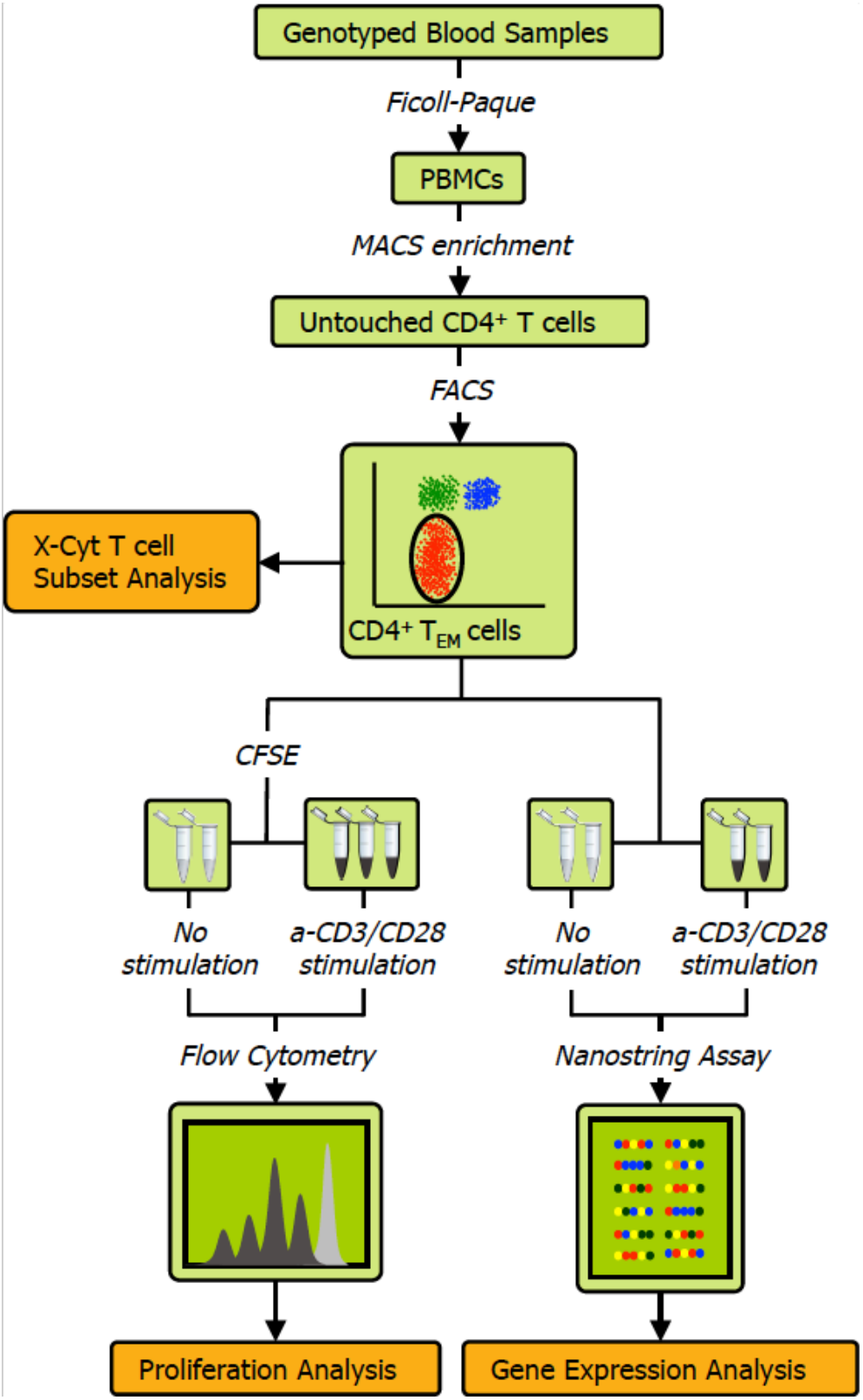
Using the ImmunoChip platform, investigators recently densely genotyped 186 loci disease that originally arose through genome-wide association studies (GWAS) in case-control samples for RA, CeD, and inflammatory bowel disease [15-17], as well as T1D (unpublished data). Dense genotyping allowed localization of association signals within these disease loci to a set of alleles that are very likely to be causal. Within these loci, we have a greater ability to identify co-localization between alleles driving variation in molecular phenotypes (such as eQTLs) and the disease risk alleles. However, in instances where multiple variants are in perfect linkage, we cannot pinpoint the exact causal variant without functional evaluation.

RESULTS

The experimental protocol (**Figure 4.2**) is described in detail in **Methods.** Briefly, we obtained peripheral blood mononuclear cells (PBMCs) from the whole blood of healthy individuals via Ficoll-Paque® centrifugation, and then used magnetic- and fluorescence-activated cell sorting to isolate $CD4^+$ T_{EM} cells at a high degree of purity (>90%). We acquired genome-wide genotype data of about 640,000 SNPs on Illumina Infinium Human OmniExpress Exome BeadChips [18]. For each individual we then measured

Figure 4.2. Schematic of the experimental workflow. We collected four types of data from each individual: 1) quality-controlled genome-wide SNP data containing 638,347 markers collected on Illumina Infinium Human OmniExpress Exome BeadChips, 2) abundance of CD4 T_{EM} cells as a percentage of all CD4 T cells obtained by FACS and quantified by X-Cyt, 3) average cell division upon T cell receptor stimulation by anti-CD3/CD28 commercial beads, measured using a CFSE (carboxyfluorescein succinimidyl ester) dye dilution assay, and 4) expression of 215 genes measured by NanoString nCounter. We repeated each proliferation assay in two-three technical replicates.

Figure 4.2. Schematic of the experimental workflow (Continued).



three quantitative phenotypes: 1) the expression of 215 genes before and after T cell receptor (TCR) stimulation by anti-CD3/CD28 antibody beads; 2) the relative abundance of $CD4^+$ T_{EM} cells ($CD45RA^-$/$CD45RO^+$/$CD62L^{-/low}$) as a proportion of total $CD4^+$ T cells; and 3) proliferation upon stimulation. Since we had low numbers of primary cells for expression profiling, we used the highly sensitive NanoString nCounter assay to avoid biases potentially induced by cDNA preparation. Out of the 215 genes assayed, 115 were within densely genotyped disease risk loci. We quantified $CD4^+$ T_{EM} cell abundance with X-Cyt, an automated statistical method that accurately identifies cell populations in cytometry data [19].

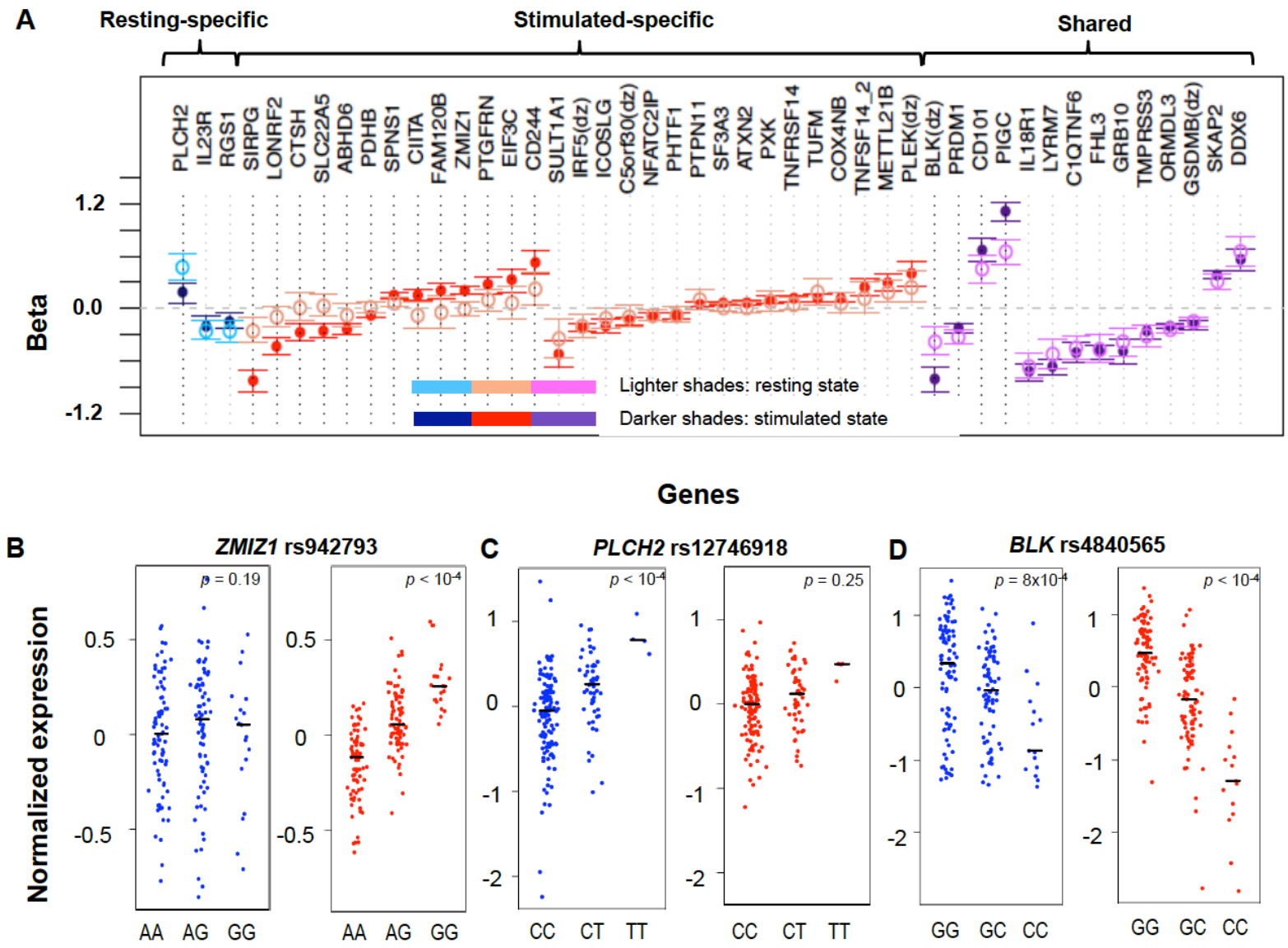
Mapping cis-eQTLs that regulate genes in risk loci

We first aimed to identify SNP variants that regulated expression of genes in *cis*. To best localize eQTL signals, we imputed 1000 Genomes variants within 250kb from the transcription start site (TSS) of each gene (excluding five HLA genes and five long non-coding RNAs). We tested SNPs in gene-coding and non-coding regions in both resting and stimulated $CD4^+$ T_{EM} cells. We included gender and the top five principal components of the genotype data (calculated by EIGENSTRAT) as covariates in regression. To adjust for multiple hypothesis testing, we conducted 10,000 permutations within each gene region to calculate empirical *p*-values, and then reported associations at a false discovery rate of 5%.

In total, we observed 46 genes (22.4%) with *cis*-eQTL signals, including 17 in resting cells and 43 in stimulated cells (**Tables 4.1 and 4.2, Figure 4.3A**). For 14 of the 46 genes (30.4%), we detected eQTL signals in both resting (14/17, 82.4%) and stimulated (14/43, 32.6%) states. In four of these 14 genes (*FHL3, GRB10, IL18R1, and PIGC*), the lead eQTL SNPs across resting and stimulated states were identical. In another five genes (*C1QTNF6, PRDM1, SKAP2, DDX6, and LYRM7)*, the lead SNPs are in tight LD (r^2 = 0.80~1; based on 1000 Genomes Release 2, European samples). For the remaining five genes (*BLK, TMPRSS3, CD101, ORMDL3, and GSDMB*), the lead SNPs from the two states were in partial LD ($0.42 < r^2 < 0.56$). In these five cases, we could not be confident that the eQTL SNPs across stimulation states were tagging the same variant.

Figure 4.3. State-specific effects of eQTL SNPs. A) For a subset of genes, the correlation effects (β) of the top associated SNP across resting and stimulated cells differed. The genes shown with a black-dotted vertical lines had significantly different effect sizes across states. Black horizontal segments in **B)-D)** denote median values. Blue panels show resting-state (normalized) expression values; red panels show stimulated expression values. **B)** rs942793^{G} significantly increased the expression of *ZMIZ1* only in stimulated cells. **C)** rs12746918^{T} was correlated with increased expression of *PLCH2* in resting cells only. **D)** rs4840565^{C} decreased *BLK* expression in stimulated cells nearly twice as much as in resting cells [β_{rest}(SE) = -0.366(0.085), β_{stim} = -0.805(0.071)].

Figure 4.3. State-specific effects of eQTL SNPs (Continued).



Three genes (*IL23R, PLCH2, and RGS1*) had statistically significant eQTLs exclusively in the resting state, while 29 genes had statistically significant eQTLs exclusively in stimulated cells, such as rs942793 associated with *ZMIZ1* expression (**Figure 4.3B**). One possibility is that some SNPs failed to reach significance threshold due to the small sample size or low expression levels in resting cells. However, we observed many genes with truly state-specific eQTLs, where the estimated effect sizes (β) of the eQTL SNP differed significantly across resting and stimulated states. To systematically compare the β_{rest} and β_{stim} for each gene, we used a *z*-statistic to quantify the probability that they differ. We then reported the *p*-value (two-tailed) assuming that *z* is distributed as standard normal, considering $p < 0.05$ to be significantly different ("state-specific"; see **Tables 4.1 and 4.2**). For example, rs12746918^{T} increased the expression of *PLCH2* significantly only in resting cells; and β_{rest} was approximately twice as large as β_{stim} (**Figure 4.3C**). We note that 1 of the 3 eQTLs in resting cells was state-specific ($p<0.05$), and 13 out the 29 eQTLs seen in stimulated cells were state-specific ($p<0.05$). Of the 14 eQTLs that were shared between resting and stimulated cells, only 4 of them, *BLK* (**Figure 4.3D**)*, CD101, PIGC,* and *PRDM1*, had different *β's* across states. The abundance of eQTLs detected exclusively in stimulated cells underscores the importance of studying cells in different cellular states.

We wanted to assess whether the eQTLs might act by altering gene regulatory elements in $CD4^+$ T_{EM} cells. To this end we asked whether the eQTL SNPs co-localized with marks of active promoters or enhancers. We utilized H3K4me3 marks from the NIH Roadmap Epigenomics Mapping Consortium [20] measured by ChIP-seq in primary $CD4^+$ memory T cells. For the SNP with the strongest association to each gene, we queried the distance of the nearest H3K4me3 mark to this SNP or its LD partners ($r^2 > 0.8$). We compared this distance measure between two sets of SNPs: the 46 SNPs with significant eQTL associations (FDR < 5%, resting or stimulated), and the SNPs most strongly correlated with the other 159 genes but did not reach significance threshold. Indeed, the 46 significant eQTL SNPs were located at smaller distances to H3K4me3 marks ($p = 1.10 \times 10^{-7}$, one-sided Mann-Whitney test). In addition, we queried the height of each H3K4me3 mark's peak, which reflected the number of reads at a given position compared to genomic

Table 4.1. Cell-state specific eQTLs. For each row we list a SNP and the gene transcript for which it is a *cis*-eQTL. We indicate whether the effect is observed in resting or stimulated CD4+ effector memory T cells. For each SNP we indicate the fold change in expression conferred per allele, assuming an additive model and the false discovery rate estimate. Finally, we indicate whether the effect is specific for CD4+ effector memory T cells; we define specificity as the absence of any cis-eQTL effect in PBMCs at FDR>50% [8].

			eQTL Effect		Resting			Stimulated				
Chr	**Top SNP**	**Gene**	**Rest**	**Stim**	**Fold Δ per Allele**	**Assoc. p-value**	**FDR**	**Fold Δ per Allele**	**Assoc. p-value**	**FDR**	**pΔ***	**CD4 T_{EM}-specific**
1	rs12746918	*PLCH2*	X		1.61	4.75E-09	<0.001	0.18	2.35E-03	> 0.20	2.70E-03	
1	rs10789226	*IL23R*	X		0.78	1.05E-05	0.022	-0.21	2.34E-03	> 0.20	6.78E-01	Yes
1	rs6704162	*RGS1*	X		0.77	2.96E-05	0.036	-0.14	1.71E-03	> 0.20	1.25E-01	
20	rs3746721	*SIRPG*		X	-0.24	1.67E-03	> 0.20	0.44	3.29E-28	<0.001	1.91E-09	
2	rs11123823	*LONRF2*		X	-0.09	1.52E-01	> 0.20	0.65	7.09E-17	<0.001	4.26E-06	
10	rs942793	*ZMIZ1*		X	0.00	9.85E-01	> 0.20	1.23	3.67E-21	<0.001	7.70E-07	Yes
5	rs10058074	*SLC22A5*		X	0.04	5.35E-01	> 0.20	0.78	1.81E-09	<0.001	1.11E-04	
16	rs143150526	*SULT1A1*		X	-0.35	2.60E-03	> 0.20	0.59	4.04E-10	<0.001	2.08E-01	Yes
21	rs2847224	*ICOSLG*		X	-0.11	5.22E-02	> 0.20	0.82	1.44E-06	0.005	1.90E-01	
5	rs39984	*C5orf30*		X	-0.09	1.77E-01	> 0.20	0.87	1.90E-08	<0.001	4.36E-01	Yes
3	rs1399754	*ABHD6*		X	-0.08	2.61E-01	> 0.20	0.79	1.11E-12	<0.001	4.07E-02	

Table 4.1. Cell-state specific eQTLs (Continued).

Chr	Top SNP	Gene	eQTL Effect		Resting			Stimulated			pΔ*	CD4 T_{EM}-specific
			Rest	Stim	Fold Δ per Allele	Assoc. p-value	FDR	Fold Δ per Allele	Assoc. p-value	FDR		
1	rs12138115	*SF3A3*		X	0.02	3.29E-01	> 0.20	1.06	1.63E-08	<0.001	7.10E-02	
1	rs6671426	*TNFRSF14*		X	0.06	1.18E-01	> 0.20	1.12	1.28E-09	<0.001	1.89E-01	
16	rs7140	*SPNS1*		X	0.07	8.07E-03	> 0.20	1.16	2.73E-12	<0.001	3.10E-02	
12	rs1021469	*METTL21B*		X	0.20	2.72E-03	> 0.20	1.35	1.22E-09	<0.001	2.09E-01	Yes
2	rs10167650	*PLEK*		X	0.25	6.30E-03	> 0.20	1.49	4.98E-08	<0.001	2.11E-01	
1	rs11265501	*CD244*		X	0.23	1.50E-02	> 0.20	1.70	1.46E-12	<0.001	1.06E-02	
15	rs7183668	*CTSH*		X	0.02	7.63E-01	> 0.20	0.76	2.19E-07	0.001	1.55E-03	
16	rs6498114	*CIITA*		X	-0.07	3.70E-01	> 0.20	1.17	6.38E-07	0.001	7.36E-03	
3	rs149241987	*PDHB*		X	0.01	6.61E-01	> 0.20	0.93	1.94E-07	0.001	1.28E-02	
16	rs11639897	*TUFM*		X	0.20	1.21E-02	> 0.20	1.13	3.05E-07	0.001	3.43E-01	
1	rs4659344	*PTGFRN*		X	0.10	1.27E-01	> 0.20	1.31	1.18E-07	0.001	3.09E-02	
16	rs146435192	*EIF3C*		X	0.07	4.84E-01	> 0.20	1.38	4.06E-06	0.002	2.96E-02	Yes
1	rs971173	*PHTF1*		X	-0.06	1.24E-01	> 0.20	0.92	7.76E-06	0.004	7.24E-01	
3	rs11711261	*PXK*		X	0.10	1.12E-01	> 0.20	1.11	5.72E-06	0.004	9.36E-01	
12	rs11066028	*ATXN2*		X	0.01	6.64E-01	> 0.20	1.07	1.86E-05	0.005	1.63E-01	

Table 4.1. Cell-state specific eQTLs (Continued).

Chr	Top SNP	Gene	eQTL Effect		Resting			Stimulated			pΔ*	CD4
			Rest	Stim	Fold Δ per Allele	Assoc. p-value	FDR	Fold Δ per Allele	Assoc. p-value	FDR		T_{EM}-specific
12	rs3858706	*PTPN11*		X	0.11	2.81E-01	> 0.20	1.05	2.17E-05	0.009	3.65E-01	Yes
19	rs2291668	*TNFSF14*		X	0.12	9.21E-02	> 0.20	1.28	5.16E-06	0.017	2.44E-01	
16	rs8587	*COX4NB*		X	0.07	1.96E-01	> 0.20	1.13	8.44E-06	0.018	4.28E-01	
16	rs7498329	*NFATC2IP*		X	-0.07	3.85E-03	> 0.20	0.91	9.22E-05	0.037	4.53E-01	Yes
6	rs7453655	*FAM120B*		X	-0.04	7.02E-01	> 0.20	1.22	3.83E-05	0.042	2.41E-02	

* pΔ measures the difference between the effect sizes across resting and stimulated states

Table 4.2. Genes with resting and stimulated eQTLs. As in **Table 4.4**, for each row we list a SNP and the gene transcript for which it is a *cis*-eQTL. For each SNP we indicate the fold change (per allele) in expression conferred per allele, assuming an additive model and the false discovery rate estimate. Finally, we indicate whether the effect is specific for CD4+ effector memory T cells [8].

		Resting				Stimulated						
Gene	**Chr**	**SNP_rest**	**Fold Δ**	**p**	**FDR_rest**	**SNP_stim**	**Fold Δ**	**p**	**FDR_stim**	**R^2**	**pΔ****	**Cell-specific**
BLK	8	rs4840565	0.69	2.75E-05	0.012	rs4840565	0.45	1.35E-22	<0.001	1*	7.38E-05	
C1QTNF6	22	rs229515	0.63	6.10E-10	<0.001	rs229522	0.61	8.75E-15	<0.001	1	5.41E-01	
CD101	1	rs4620527	1.8	3.06E-09	<0.001	rs9332416	1.96	1.76E-19	<0.001	0.547	3.23E-02	yes
DDX6	11	rs500254	1.92	2.33E-13	<0.001	rs4938544	1.75	1.48E-16	<0.001	0.834	3.40E-01	yes
FHL3	1	rs67631072	0.63	6.43E-08	<0.001	rs67631072	0.61	9.53E-26	<0.001	1*	7.05E-01	
GRB10	7	rs12536500	0.68	4.10E-06	0.016	rs12536500	0.61	4.92E-11	<0.001	1*	3.11E-01	
GSDMB	17	rs36038753 rs12936409	0.86	2.20E-07	<0.001	rs36038753/rs12936409	0.83	2.64E-12	<0.001	1*	3.65E-01	
IL18R1	2	rs3771164	0.52	6.35E-16	<0.001	rs3771164	0.48	3.79E-31	<0.001	1*	4.35E-01	
LYRM7	5	rs12522164	0.59	1.05E-08	<0.001	rs12517633	0.51	1.38E-29	<0.001	0.792	1.36E-01	
ORMDL3	17	rs35222145	0.75	1.16E-19	<0.001	rs2290400	0.82	4.44E-28	<0.001	0.482	2.75E-01	

Table 4.2. Genes with resting and stimulated eQTLs (Continued).

Gene	Chr	Resting				Stimulated						
		SNP_rest	Fold Δ	p	FDR_rest	SNP_stim	Fold Δ	p	FDR_stim	R^2	pΔ**	Cell-specific
PIGC	1	rs1063412	1.91	1.36E-14	<0.001	rs1063412	3.04	2.65E-46	<0.001	1*	6.79E-07	
PRDM1	6	rs811925	0.72	2.61E-13	<0.001	rs578653	0.81	3.30E-13	<0.001	0.944	2.78E-02	
SKAP2	7	rs4719882	1.37	1.85E-12	<0.001	rs3801813	1.46	6.08E-39	<0.001	0.862	1.17E-01	
TMPRSS3	21	rs7283281	0.76	1.08E-06	0.005	rs2277798	0.77	7.74E-10	<0.001	0.55	5.42E-01	

* denotes identical SNPs

** In cases where the lead resting and stimulated SNPs differ, pΔ compares the cross-state effect sizes of only one SNP. For each gene, the SNP with the stronger association across the two states was considered.

controls as defined by the MACS software package. A tall peak gives us confidence that the mark is present in a large proportion of cells. Comparing the marks nearest to the two sets of SNPs, we saw that the 46 eQTL SNPs were also located near taller peaks ($p = 9.56 \times 10^{-8}$**).**

Many eQTLs are CD4$^+$ T_{EM} cell-specific

We compared the *cis*-eQTLs we discovered to those found in heterogeneous peripheral blood mononuclear cells (PBMC) in a large genome-wide eQTL meta-study (n = 5,331) conducted by Westra *et al*. [8]. At 5% FDR, eleven of the 46 eQTL genes we identified showed no detectable signal in PBMCs at 50% FDR. We saw significant associations in 131 genes at 50% FDR, 53 of which had no signal in PBMCs at 50% FDR (**Tables 4.1 and 4.2**). We hypothesized that these genes tended to be more specifically expressed in CD4$^+$ T_{EM} cells, thus making eQTLs readily detectable in the purified cell population. To assess this, we examined cell-specific expression of the genes the ImmGen dataset, which assayed the genome-wide expression in 247 murine mouse immunological cell types [13, 21]. We found that the genes with CD4$^+$ T_{EM} cell-specific eQTLs (at 50% FDR) were more specifically expressed in CD4$^+$ T_{EM} cells than genes with eQTLs detected in both datasets ($p = 0.044$, one-sided Mann-Whitney test).

Autoimmune disease alleles affect the transcription of genes in cis

We then focused on 115 genes near 96 risk alleles of RA, T1D, and/or CeD in densely genotyped loci (182 gene-SNP pairs, including two risk alleles shared by at least two diseases). We discovered that eleven (11.4%) disease-associated SNPs (6 of 24 RA SNPs, 5 of 37 T1D SNPs, and 3 of 37 CeD SNPs) correlated significantly with the expression of ten genes in either resting or stimulated state **(Table 4.1).** In addition, there was substantial enrichment of nominally significant associations ($p < 0.05$) among disease SNPs. By random chance, we expected about nine SNP-gene pairs to reach nominal association in each stimulation state. However, we observed 26 pairs (14.2%) with nominal association in resting cells ($p = 4.67 \times 10^{-7}$, one-

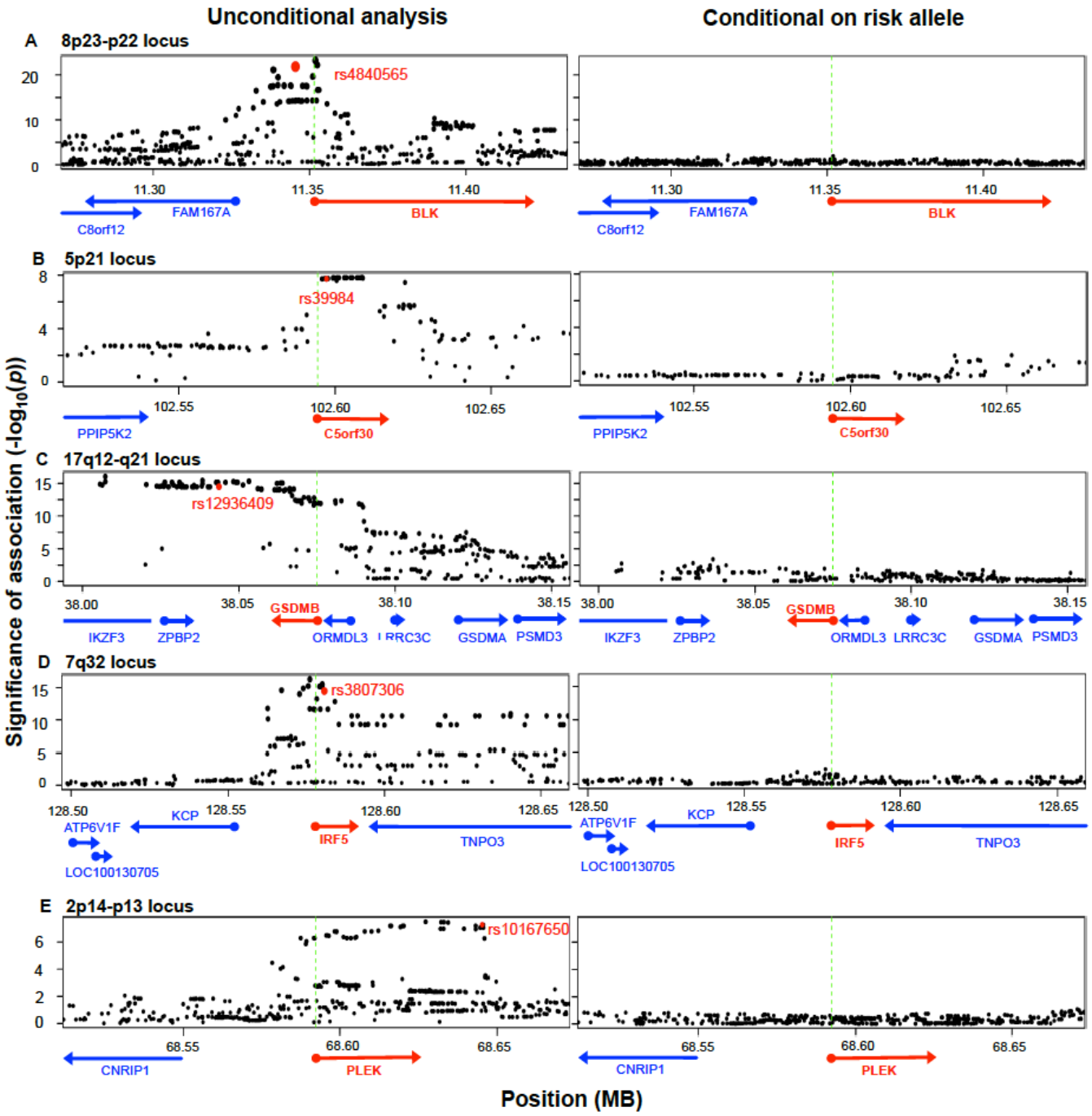
tailed binomial test). Even more strikingly, we observed 45 pairs (24.7%) with nominal association in stimulated cells ($p < 10^{-15}$, one-tailed binomial test).

To identify those instances where the disease-associated SNP could explain the entire eQTL signal in the gene region, we applied conditional analysis to identify any residual signals after controlling for the disease SNP. In five of the ten genes (*BLK, C5orf30, GSDMB, IRF5, PLEK*), conditioning on the disease SNP obviated any remaining eQTL signal in the region (no SNP with permutation *p*-value < 0.05; **Figure 4.4**), suggesting that there was a single variant (the disease-associated SNP or one in very high LD to it) that drove variation in expression. Interestingly, as previously noted, the lead SNPs in resting and stimulated states for *BLK* and *GSDMB* were in partial linkage to each other. The absence of residual eQTL signal upon conditioning on the same risk allele might suggest that the lead SNPs were indeed tagging the same causal SNP in each of these genes. In each of the other five genes (*ORMDL3*, *SKAP2*, *TMPRSS3, TNFRSF14*, and *ZMIZ1*), evidence of independent eQTL effect remained after conditional analysis. In these instances the disease-associated SNP and remaining lead signal are in partial linkage disequilibrium (r^2 = 0.36-0.73). In these cases, we could not conclude whether the disease SNPs drove the alteration in expression, or whether the true causal SNPs were in partial linkage and caused spurious associations. It is probable that disease risk alleles were indeed causal, yet we could not confidently fine-map the effect due to experimental noise in expression assays or inadequate sampling.

We note that another 26 genes within disease loci associated contained *cis*-eQTL signals, but that these *cis*-eQTL signals did not co-localize with RA, T1D, or CeD alleles. As these loci had been fine-mapped using Immunochip, the lack of overlap strongly suggested that these *cis*-eQTLs and disease-causing variants were distinct. For example, rs798000 is an RA risk allele located in a non-coding region upstream of *CD2*, *CD58*, and *PTGFRN*. However, it was not associated with the expression of any of these genes ($p > 0.5$). Another example was rs6911690, an RA allele located about 60kb 5' of *PRDM1*, that was not associated with the expression of the gene at rest or after stimulation ($p > 0.5$). The lead eQTL SNP associated to *PRDM1* was rs578653 (FDR < 10^{-3}), which was not in LD with the disease allele ($r^2 < 0.05$).

Figure 4.4. Five disease risk alleles explained the eQTL associations with five genes. The left-sided panels show unconditional SNP-expression association results. Green dashed lines mark the TSS of the eQTL gene. The red dots indicate the risk alleles associated with the expression of respective genes shown as red arrows. The right-sided panels show adjusted association results after conditioning for the respective risk alleles. In each of the five loci, conditioning on the disease SNP obviated signals in the entire region, such that no association more significant than $p = 0.05$ remains.

igure 4.4. Five disease risk alleles explained the eQTL associations with five genes (Continued).



The genetic basis of CD4+ T_{EM} cell proliferation

The relative peripheral abundance of CD4+ T_{EM} cells varied between individuals (mean = 9.57%; SD = 4.85%), and was reproducible 35 individuals with two separate blood draws more than one month apart (Pearson's r = 0.87, p = 1.77x10^{-11}, see also **Figure 4.5**). Consistent with other studies, we observed that the relative proportion of CD4+ T_{EM} cells increased with age by 0.11% per year (p_{age} = 1.92x10^{-3}) [22]. We also observed that on average men had 2.22% more CD4+ T_{EM} cells than women (p_{gender} = 3.80x10^{-2}; see **Figure 4.6**). Upon anti-CD3/CD28 stimulation, there was a substantial inter-individual variation in proliferation measured by both division index (DI, average number of divisions undergone by all cells; mean = 1.46, SD = 0.35), and proliferation index (PI, average number of divisions undergone only by dividing cells; mean = 2.16, SD = 0.21). Proliferation metrics were also reproducible in the 35 individuals (Pearson's r_{DI} = 0.57; Pearson's r_{PI} = 0.62, **Figures 4.5C and 4.5D**). Interestingly, proliferation was negatively correlated to the proportion of CD4+ T_{EM} cells (p_{DI} = 1.28x10^{-3}, p_{PI} = 1.93x10^{-3}), but was not associated to age or gender (p > 0.3). This negative correlation needs to be replicated in an independent dataset. Effector functions of T_{EM} cells with higher proliferative capacities need to be examined to understand whether they represent a hyperactive subset whose abundance is controlled to maintain immune homeostasis. Possibly individuals with a lower proportion of T_{EM} cells are relatively enriched for these subsets.

We tested genome-wide SNPs for association to relative abundance, division index, and proliferation index, considering p < 5x10^{-8} as the threshold for significance. For abundance, we included gender, age, and the top five principal components of genotypes as covariates. Given the correlation with proliferation, we also included the measured CD4+ T_{EM} relative abundance as an additional covariate. We observed associations to division index in several loci, including 13q34 led by rs389862 (p = 4.75x10^{-8}; **Figure 4.7**). This SNP is a non-coding variant located 30kb upstream of *RASA3*, and 70kb upstream from *CDC16.* Both genes have known roles in regulating cell proliferation or differentiation [23, 24]. This SNP was also strongly associated with proliferation index (p = 2.75x10^{-7}). Additionally, there was a strongly suggestive association to rs3775500 on chromosome 4, located in the intron of *DAPP1*, which encodes the

Figure 4.5. Purity and reproducibility. A) Using a combination of magnetic and fluorescence-activated cell sorting (MACS and FACS), $CD4^+$ T cells were isolated to a high degree of purity. The isolated population contained ~97% $CD3^+$ cells, ~90% $CD4^+$ cells, ~0.4% $CD8^+$ cells, and ~0.03% $CD19^+$ cells. **B)** The relative abundance (as a percentage of all sorted lymphocytes), **C)** division index (average division of all cells), and **D)** proliferation index (average division of all cells that went into division), were reproducible in 35 individuals with two blood draws at least one month apart. Pearson's r = 0.87, 0.57, and 0.62, respectively.

Figure 4.5. Purity and reproducibility (Continued).

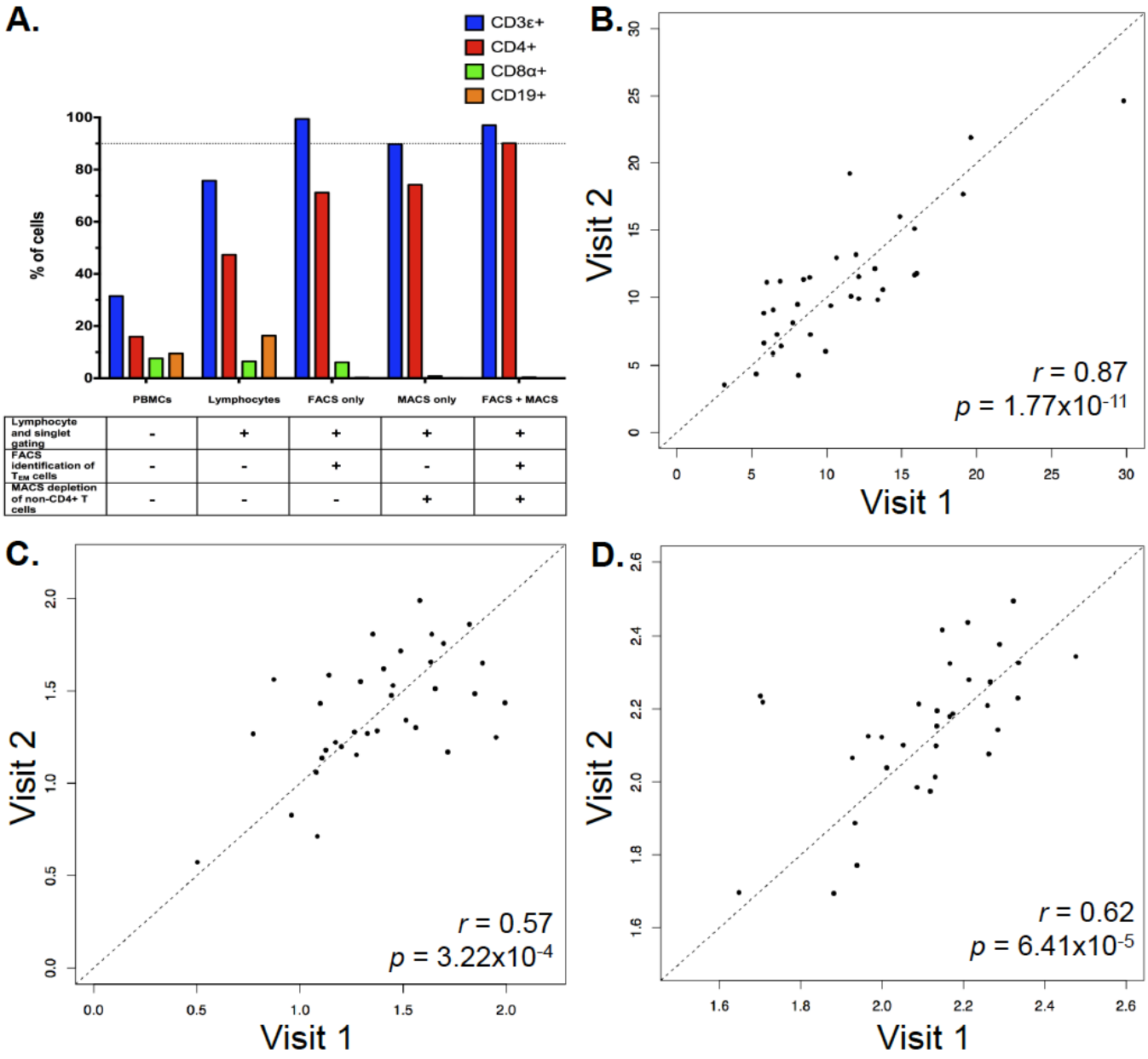


Figure 4.6. The relative abundance of CD4$^+$ T_{EM} cells. CD4$^+$ T_{EM} abundance as the percentage of CD4 T cells **A)** increased with age, at 0.11% per year; and **B)** was correlated with gender, where men on average as 2.2% more CD4 T_{EM} cells than women. The associations remained significant in a multivariate linear regression.

Figure 4.6. The relative abundance of $CD4^+$ T_{EM} cells (Continued).

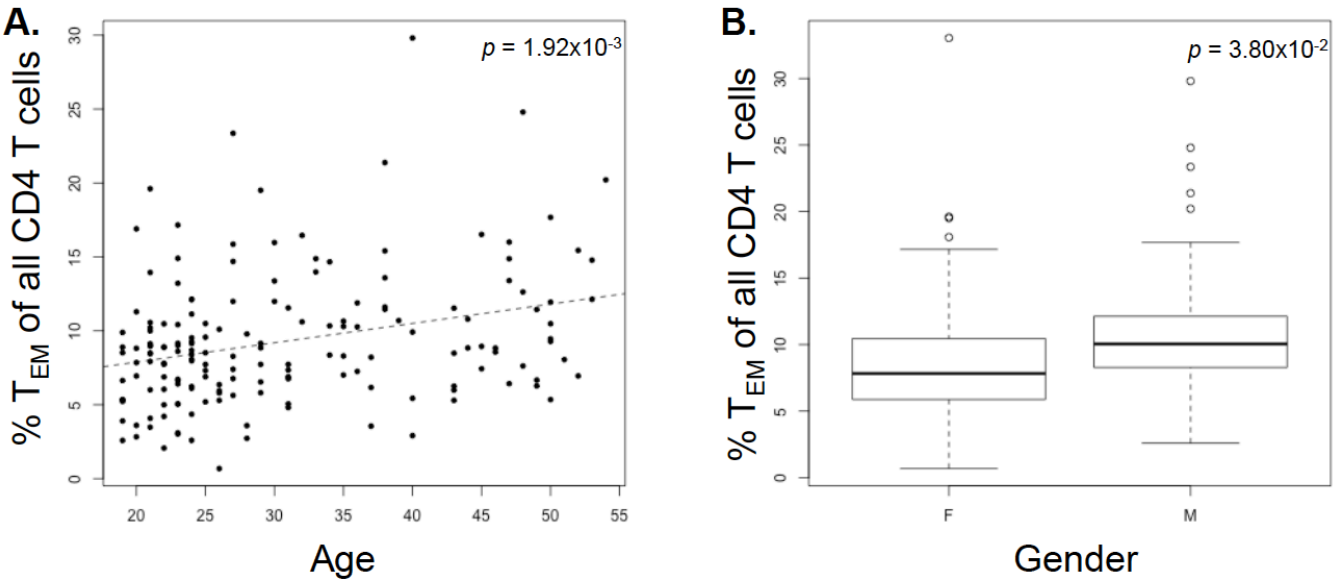
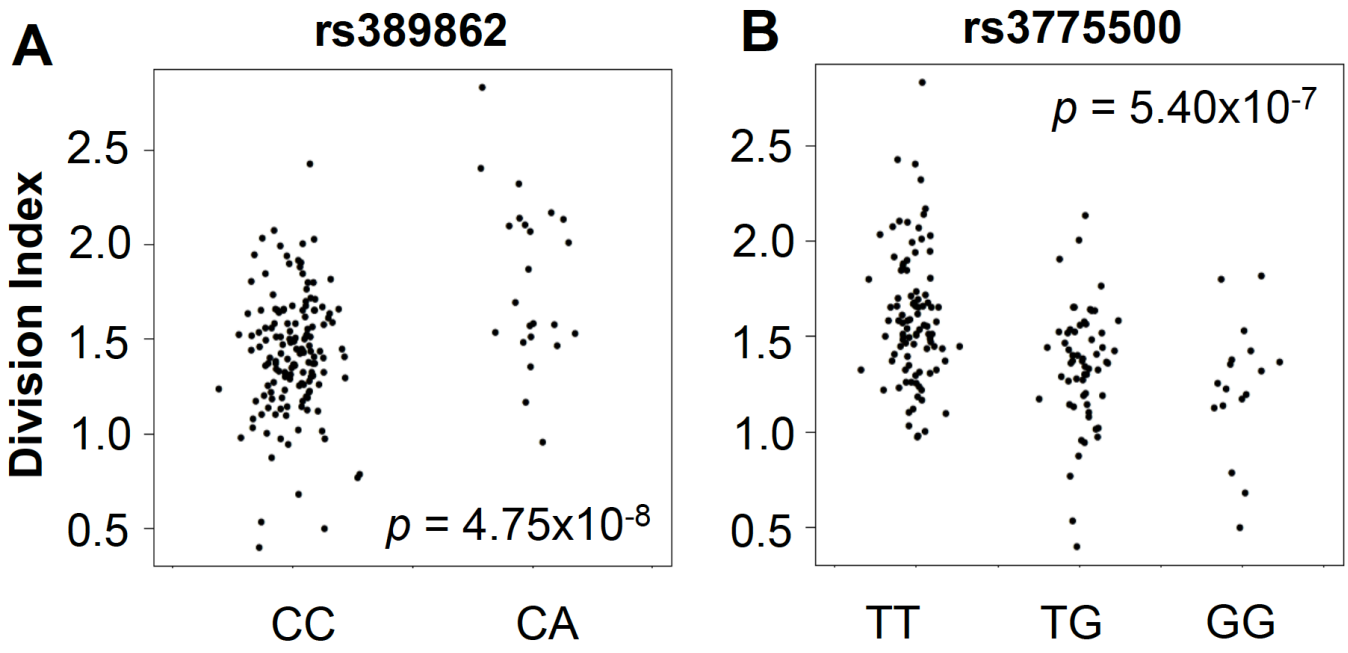


Figure 4.7. Genome-wide association to division index (the average number of division undergone by all cells). A) rs389862^{A} on chromosome 13 was significantly associated to increased division index at $p = 4.75 \times 10^{-8}$, and is located in a non-coding region 30kb upstream of *RASA3*, and 70kb upstream from *CDC16*. **B)** rs3775500^{G} on chromosome 4 shows a strongly suggestive association at $p = 5.40 \times 10^{-7}$, and is located within the *DAPP1 (*Bam32*)* gene.

Figure 4.7. Genome-wide association to division index (the average number of division undergone by all cells) (Continued).



Bam32 protein (p = 5.40x10^{-7}; **Figure 4.7B**), which is an adaptor protein expressed solely in antigen presenting B cells. Interestingly, mutations in this gene have been shown by several groups to affect T cell activation [25, 26], suggesting the possibility that B cells may indirectly regulate T cell function in autoimmunity. We did not observe any significant association with the relative abundance of CD4$^+$ T_{EM} cells.

When we extracted the association statistics of 118 densely genotyped risk alleles of CeD, RA, and/or T1D, they showed no inflation in association p-values for relative abundance of CD4$^+$ T_{EM} cells (**Figure 4.8A**). This suggested that risk variants did not modify risk via modulation of CD4$^+$ T_{EM} peripheral abundance. We recognized that the power to detect significant associations might have been limited in our study by the sample size. However, this negative finding was corroborated by results from a recently published study with data from ~2800 individuals, in which the same set of risk alleles also showed no significant association to CD4$^+$ T_{EM} (see **Figure 4.9**) [27]. Similarly, the same set of risk alleles did not show significant association to proliferative response (**Figure 4.8B).** Based on these data, it was unlikely that SNP variants associated to RA, T1D, or CeD conferred risk through modulation of CD4$^+$ T_{EM} cell abundance or proliferation.

Gene expression in resting cells predicted post-TCR stimulation proliferation

After stimulation we observed that 122 genes showed significant changes in expression in response to stimulation, including 78 whose expression at least doubled or decreased by 50%. The gene with the greatest post-stimulation induction was *GZMB* (average fold change = 93.48), which encodes granzyme B, a protein involved in the apoptosis of target cells during cell-mediated immune response in cytotoxic and memory lymphocytes. The most significantly down-regulated gene was *GRB10* (average fold change = 0.18), which is near rs6944602 associated with T1D and encodes growth factor receptor-bound protein 10, whose function in the immune system is unclear.

Figure 4.8. Risk alleles of CeD, RA, and T1D, showed no significant association to CD4 T_{EM} cell abundance or proliferation. A) The 118 SNPs with association to diseases in densely genotyped regions on Immunochip platform were not significantly associated to CD4 T_{EM} cell abundance. The shaded region shows 95% confidence interval. See also **Figure S5. B)** The same set of 118 risk alleles also showed no inflation in association with proliferative response measured as division index.

Figure 4.8. Risk alleles of CeD, RA, and T1D, showed no significant association to CD4 T_{EM} cell abundance or proliferation (Continued).

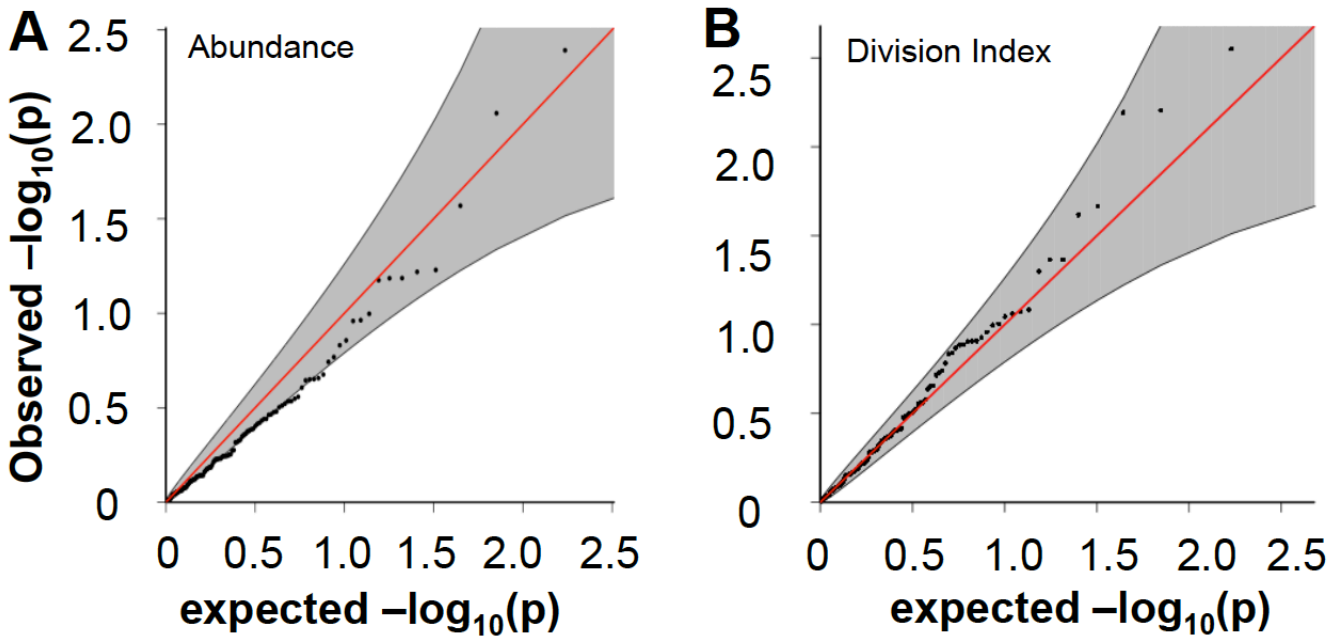
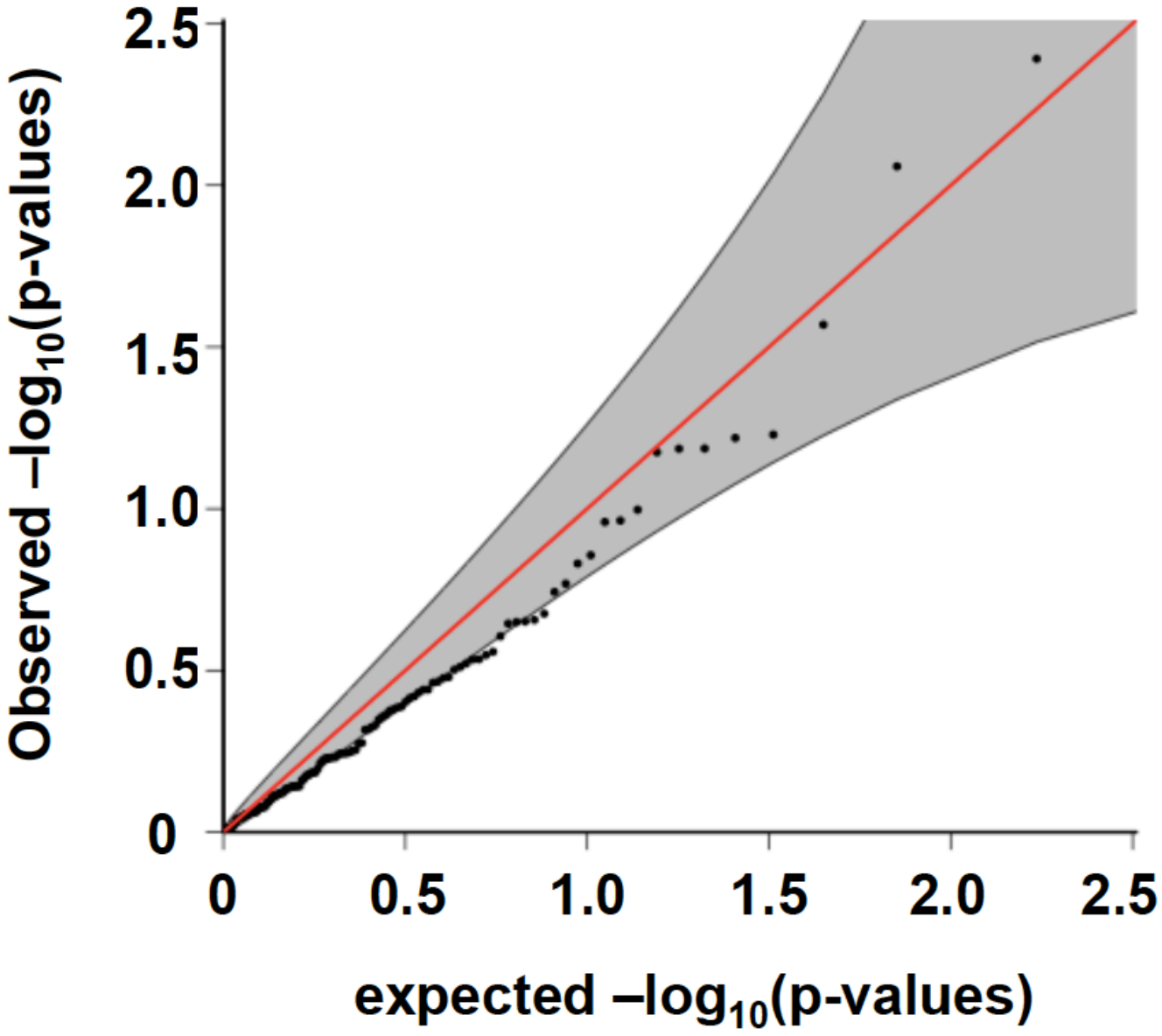


Figure 4.9. SNPs associated to CeD, RA, and T1D, showed no significant association to CD4 T_{EM} cell abundance. The 119 risk alleles within densely genotyped loci showed no significant association to CD4 T_{EM} abundance as a percentage of CD4 T cells in the study by Orru et al. The shaded area shows the 95% confidence interval.

Figure 4.9. SNPs associated to CeD, RA, and T1D, showed no significant association to CD4 T_{EM} cell abundance (Continued).



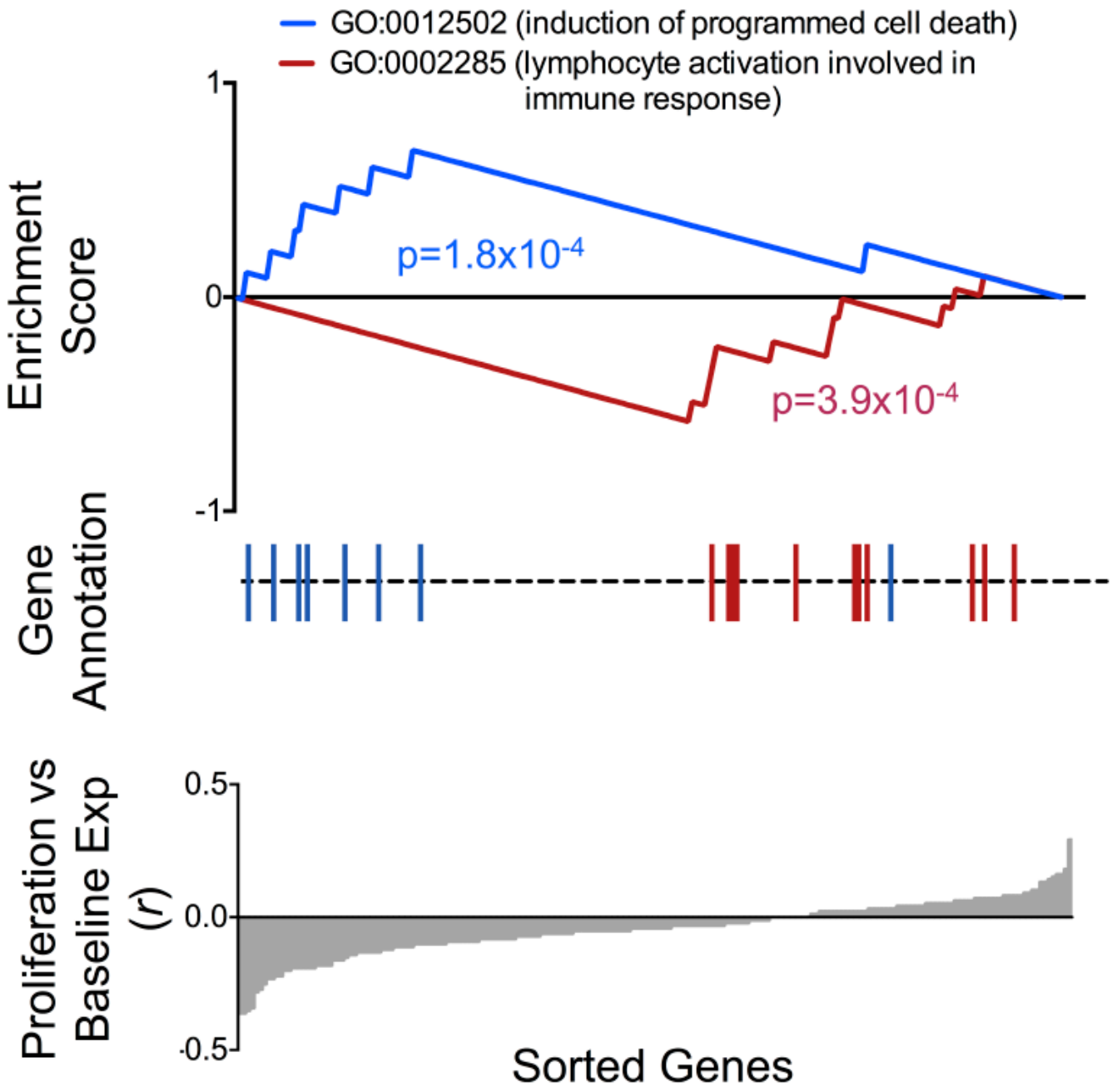
We observed that relative gene expression at rest predicted proliferative response. In 182 individuals with both proliferation and gene expression data, 17 of the 215 genes were associated with proliferation index ($p < 0.01$, two-tailed test by permuting proliferation data, **Figure 4.10**). Increased expression of 15 of the 17 genes including *CCR5*, *IL2RB, PRR5L, and TBX21,* were correlated with reduced proliferative response, while *CCR9* and the lncRNA XLOC_003479 showed significant correlation with increased proliferation. This number of correlated genes was far in excess of random chance based on a null distribution consisting of 1000 permutations ($p < 10^{-3}$, median 2, maximum 15). The weighted sum of the 17 genes served as a "proliferation potential signature", where we weighted the positively- and negatively-correlated genes as +1 and -1, respectively. This signature strongly predicted proliferation index (r = 0.55). We show the correlation between each of the 17 genes as well as the aggregate signature to proliferation as a heatmap (**Figure 4.10A**). To assess if we were overfitting the data, we applied a two way cross-validation, where we defined the proliferation signature based on genes from half of the individuals and tested correlation to proliferation in the remaining half of the individuals. In both instances we again observed significant prediction of proliferation (r=0.41,one tailed $p<10^{-3}$ by permutation; r=0.39, $p<10^{-3}$).

To search for biological pathways underlying genes correlated to proliferation, we applied gene set enrichment analysis (GSEA) to test for enrichment for 1,008 functional gene sets based on Gene Ontology codes [28] (**Figure 4.10B**). Genes correlated to reduced proliferation were most significantly enriched for GO:0012502 (induction of programmed cell death; one tailed $p = 1.8 \times 10^{-4}$); those correlated with increased proliferation were most significantly enriched for GO:0002285 (lymphocyte activation involved in immune response, one tailed $p = 3.9 \times 10^{-4}$).

Using data from 29 individuals each with two samples collected at least one month apart, we replicated the observed correlation. In these samples we performed a cross-visit analysis, and observed that the same 17-gene signature from the first visit significantly predicted proliferation indices on the second visit ($r = 0.65$, $p = 0.0006$, 1-tailed permutation), and *vice versa* ($r = 0.55$, $p = 0.0019$).

Figure 4.10. Relationship between baseline expression and post-stimulatory response. A) Baseline expression of 17 genes correlated with post-stimulation proliferation. Rows in the heatmap are ordered from top to bottom by ascending proliferation index. Genes/columns are ordered from the most negatively correlated (IL23RB) to the most positively correlated (CCR9). The 17-gene signature was calculated as the weighted sum of the 17 genes, where the negatively-correlated genes were given a weight of -1, and the positively-correlated genes were given a weight of +1. **B)** Genes correlated with proliferative response were enriched for apoptosis and lymphocyte activation pathways. Genes correlated to lower proliferative response (proliferative index) were enriched for Gene Ontology code GO:0012502 (induction of programmed cell death, $p = 1.8 \times 10^{-4}$). Conversely, genes correlated to higher proliferative response were enriched for GO:0002285 (lymphocyte activation, $p = 3.9 \times 10^{-4}$).

Figure 4.10. Relationship between baseline expression and post-stimulatory response (Continued).



DISCUSSION

To fine-map and link risk loci to their pathogenic mechanisms, we investigated molecular and immune phenotypes potentially leading to disease end-points. The immune system is particularly complex, and different cells under various activation states have specialized functions that may not be adequately captured by examining PBMCs. Therefore, we focused on one purified cell population that had been shown to be important for the pathogenesis of several autoimmune diseases. We quantified population variation in several traits, including peripheral abundance, proliferative response to TCR stimulation, and expression of genes within autoimmune disease loci at rest and after stimulation. In **Tables 4.1** and **4.2**, we provide significant cis-eQTLs and genome-wide association results.

To our knowledge this study was a first cross-examination of genetic-, transcriptional-, and cellular-level quantitative traits in $CD4^+$ T_{EM} cells. It demonstrated the importance of focusing functional studies in a purified cell population under relevant developmental and stimulation states. By examining the proliferative response upon TCR stimulation, we identified a subset of genes whose baseline expression predicted proliferative potential. Intriguingly, these genes were involved in programmed cell death and lymphocyte activation. Whether variation in proliferative abilities correlated with cytokine production and other signaling functions, thus affecting susceptibility to autoimmunity, remains a question to be addressed by future studies.

Of the 205 genes in disease loci that we examined, 46 had *cis*-eQTLs. Notably, eleven of these were specific to stimulated $CD4^+$ T_{EM} cells, and not previously found in PBMCs. We noted that approximately 10% of genes within risk loci of diseases had *cis*-eQTLs. However in many instances the lead eQTL SNPs were unrelated to the disease-associated SNPs. One example of a disease allele that functioned as *cis*-eQTL was rs39984, which was associated to lower risk of RA, and regulated the expression of *C5orf30* encoding an UNC119-binding protein. This SNP variant is located in the first intron of *C5orf30,* and indeed explained the entire *cis*-eQTL signal in this gene (see **Figure 4.3B**). This eQTL effect was previously undetected in PBMCs, and the protein's functional role in the immune system is largely unknown. However, a recent study

showed that rs26232 (the lead GWAS SNP prior to fine-mapping, r^2 = 0.988 to rs39984) was correlated with lower severity of radiologic damage in RA, independent of previously established biomarkers [29]. Another gene in the locus, *GIN1,* is located 140kb from rs39984; however its expression showed no correlation with the SNP ($p > 0.5$).

Another CD4$^+$ T_{EM} cell-specific eQTL gene was *DDX6*, which encodes DEAD-box RNA helicase 6. However, in this case, the lead eQTL SNP (rs4938544) associated to increased expression of *DDX6* in stimulated cells was not in LD with the known CeD risk allele (rs10892258, $r^2 < 0.1$) or the RA risk allele (rs4938573, $r^2 < 0.1$). Neither risk allele showed significant association to DDX6 expression (p = 0.19 and 0.26, respectively). Both risk alleles are also located near *CXCR5*, *BCL9L*, and *TREH;* none of these genes had reported *cis*-eQTLs in PBMCs [8]. However, we did not assay these three genes in this study, therefore could not confirm the role of disease alleles in regulating their expression in CD4$^+$ T_{EM} cells.

Although we did not assay all genes or test for *trans*-acting eQTLs, based on the level of co-localization between eQTL SNPs and risk alleles observed in the study, we found it unlikely that all non-coding risk variants caused disease by altering gene expression within resting or stimulated CD4$^+$ T_{EM} cells. In addition, while changes in proportions of lymphocyte subsets had been observed in patients of autoimmune disorders [30-35], we did not find evidence to support disease alleles' roles in directly modulating CD4$^+$ T_{EM} cell abundance or proliferative response. Ultimately, other cell states and cell types will need to be investigated.

We recognize several limitations to the current study. In order to conduct a focused study on a small amount of purified primary cells we used the NanoString nCounter assay system. This avoided potential biases and artifacts arising from cDNA synthesis required for microarray or RNA-seq studies, but restricted our analysis to a subset of candidate genes within risk loci of CeD, RA, and T1D, rather than a genome-wide expression analysis. Consequently we could not identify trans-eQTLs, splice variants, or epistatic effects on expression regulation. Additionally, anti-CD3/CD28 stimulation for memory T cells is

not antigenic, especially while in isolation from a "natural" multi-cellular environment, thus it was only partially physiological.

This and other cell-specific studies on population variation in molecular phenotypes are only a beginning of examining potential intermediate phenotypes. Post-activation cytokine production by $CD4^+$ T_{EM} cells are likely crucial in driving autoimmunity. Therefore, it is critical that future studies of molecular phenotypes include proteomic assays to quantify functional markers of immune response. Finally, functional experiments will need to be conducted in the future to determine whether these molecular phenotypes are indeed intermediary to disease.

MATERIALS AND METHODS

Ethics statement

All research was approved by our Institutional Review Board, and informed consent was obtained from each volunteer.

Study sample

We enrolled 225 healthy volunteers (134 females, 91 males) of non-Hispanic Caucasian descent that proved informed consent through the Phenogenetics Project at Brigham and Women's Hospital. Subjects' ages ranged from 19 to 57 years with average female and male ages of 28.8 years and 34.9 years, respectively. Thirty-five subjects (18 females, 17 males) returned for a second study visit one to nine months after their initial visits.

Genotyping

We genotyped each subject using the Illumina Infinium Human OmniExpress Exome BeadChip. In total, we genotyped 951,117 SNPs, of which 704,808 SNPs are common variants (minor allele frequency [MAF] >

0.01) and 246,229 are part of the exome. After quality control, 638,347 common SNPs remained. Of all subjects, 174 subjects had abundance, proliferation, gene expression, and quality controlled genotype data.

SNP Imputation

For each gene, we selected a 500kb region (250kb each in the 3' and 5' directions) around the transcription start site and imputed 1000 Genomes SNPs into the genome-wide SNP data using BEAGLE Version 3.3.2. We used the European samples from 1,000 Genomes as the reference panel. We excluded markers that had MAF < 0.05 in the reference panel as well as all insertion/deletions. After imputation, we excluded markers with a BEAGLE $R^2 < 0.4$ or MAF < 0.01 in the imputed samples.

Buffers and media

Peripheral blood mononuclear cells (PBMCs) were washed with a cold, divalent cation-free Hycole Dulbeccos (Thermo Scientific) phosphate buffered solution (PBS). Antibody staining of CD4 T cells was performed in "fluorescence activated cell sorting (FACS) buffer", which is PBS containing 0.5% BenchMark fetal bovine serum (Gemini Bio-Products) and 2mM EDTA (Gibco). CD4 T_{EM} cells were cultured in "basic human media", which is RPMI 1640 media (Gibco) containing 10% Hyclone fetal bovine serum (Thermo Scientific), 5% BenchMark fetal bovine serum (Gemini Bio-Products), and supplemented with the following items and their final concentrations or volumes: 30 mM HEPES, 100 U/mL penicillin, 100 μg/mL streptomycin, 1 mM L-glutamine, 0.5 mM sodium pyruvate, 0.055 mM β-mercaptoethanol, 2.5 mL of an essential amino acid solution (Gibco), and 2.5 mL of a non-essential amino acid solution (Gibco).

Blood collection and PBMC isolation

For each subject, 30 mL of non-fasting blood was collected into plastic tubes spray-coated with EDTA (BD). The blood was carefully layered over Ficoll-Paque PLUS (GE Healthcare) and centrifuged at 2,000 rpm for 30 minutes to isolate PBMCs. PBMCs were washed twice with cold PBS, resuspended in FACS

buffer, and filtered through a 70 μm nylon mesh. The time from blood collection to the Ficoll procedure was always less than 7 hours.

MACS enrichment for CD4 T cells

Magnetic activated cell sorting (MACS) was used to enrich PBMCs for CD4 T cells by depleting CD8+ T cells, monocytes, neutrophils, eosinophils, B cells, dendritic cells, NK cells, granulocytes, γ/δ T cells, and red blood cells using a CD4 T cell isolation MACS kit (Miltenyi). FACS buffer was used as the eluent.

FACS isolation of T_{EM} cells

FACS was used to isolate T_{EM} Cells from enriched CD4 T cells, which were labeled with phycoerythrin (PE)-conjugated anti-CD62L (eBioscience), eFluor450-conjugated anti-CD45RA (eBioscience), and allophycocyanin (APC)-conjugated anti-CD45RO antibodies (eBioscience) on ice for 40 minutes in FACS buffer. Labeled cells were then washed twice with and resuspended in FACS buffer. Cells were kept at 4°C overnight. The following morning, labeled cells were sorted on a BD FACSAria SORP flow cytometer for T_{EM} cells, which were defined as being $CD45RA^{-}CD45RO^{high}CD62L^{-/low}$. T_{EM} cells were sorted into two tubes of basic human media, one for 100,000 cells and one for 120,000 cells. The first tube was used for the Nanostring gene expression assay while the second tube was used for the proliferation assay. FCS files of the sorting data were saved for automated quantification of T_{EM} cell abundance.

T_{EM} cell stimulation

Sorted T_{EM} cells were plated into round-bottom, 96-well plates at 20,000 cells/well. Wells for the stimulated condition received 2,000 Dynabeads coated in anti-CD3 and anti-CD28 antibodies (Invitrogen) in basic human media for a bead:cell ratio of 1:10. Wells for the resting condition received an equal volume of basic human media only. The cells for the proliferation assay and the gene expression assay were plated on separate plates. The proliferation assay plate contained two resting replicates and three stimulated

replicates. The gene expression plate contained two resting replicates and two stimulated replicates. In both plates, the outer wells were avoided to minimize variability between the wells. All cells were incubated at 37°C for 72 hours.

Proliferation assay

Prior to plating the T_{EM} cells for the proliferation assay, cells were washed with and resuspended in room temperature PBS. They were then labeled with 0.5 μM carboxyfluorescein diacetate succinimidyl ester (CFSE; eBioscience) in room temperature PBS for two minutes. Cells were quenched with cold BenchMark fetal bovine serum (Gemini Bio-Products) and basic human media. Cells were then resuspended in basic human media and plated. Following the incubation period, cells were removed from wells and analyzed on a BD FACSCantoII flow cytometer. The two resting replicates for each sample were pooled to define the undivided cell population. FCS files of the proliferation assay were saved for downstream, automated analysis.

Selecting target and reference genes for custom codeset

A list of the known single nucleotide polymorphisms (SNPs) associated ($P < 5x10^{-8}$) with rheumatoid arthritis, celiac disease, and type 1 diabetes, via genome-wide association and/or Immunochip studies, as of May 2011, was compiled. For each associated SNP, its implicated genomic region of interest was first defined by the furthest SNPs in linkage disequilibrium in the 3' and 5' directions ($R^2 > 0.5$), then extended outward to the nearest recombination hotspot. All genes with any overlap with this region of interest were collected. A total of 931 unique genes were implicated by all associated SNPs. To prioritize these genes, they were annotated based on the following criteria: 1) distance to the SNP of interest; 2) Gene Ontology annotation; 3) known eQTL status; 4) a minimal expression specificity in CD4 T cells (based on mouse ImmGen data) or immune cells (based on human GNF dataset). In addition, 19 genes and ten long

non-coding RNAs (lncRNAs) were added to the codeset based on immunological interest, but were not implicated by the above-mentioned association SNPs,

15 reference genes were included for the purpose of controlling for cell numbers, total RNA quantity. Due to the large metabolic demand and cytoskeleton remodeling that occurs with stimulation-induced proliferation, housekeeping genes commonly used in other molecular biology assays, such as GAPDH and actin, were not used. Instead, genes that showed stable expression levels in resting and stimulated CD4 T cells were identified in two publically available microarray datasets in Gene Expression Omnibus, GSE32607 and GSE28726, which studied primary and cloned human T cells after stimulation. Genes that showed minimal fold change in both datasets and spanned the low, medium, to high expression ranges were selected.

In total, 314 target genes and 15 reference genes were included in the custom codeset.

Nanostring nCounter sample preparation and processing

Cells used in the gene expression assay were not labeled with CFSE prior to plating. Following the incubation period, plates were centrifuged at 2,000 rpm for 5 minutes and the media in the wells was aspirated. Cells were lysed with 5 μL of an RLT lysis buffer (Qiagen) solution containing 1% β-mercaptoethanol. Cell lysates were stored at -80°C for 2-14 months until analysis. The standard nCounter cell lysate gene expression assay protocol was used to process the samples. All replicates were processed separately.

$CD4^+$ T_{EM} cell isolation and stimulation

We isolated peripheral blood mononuclear cells (PBMC) from whole blood using a Ficoll density gradient (GE Healthcare). We then isolated $CD4^+$ effector memory T cells from PBMCs first by magnetic-activated cell sorting to enrich for $CD4^+$ T cells, followed by fluorescent-activated cell sorting using labeled antibodies against CD45RA, CD45RO, and CD62L.

We stimulated $CD4^+$ T_{EM} cells by incubation with commercial anti-CD3/CD28 beads for 72 hours. For proliferation studies, we labeled cells with carboxyfluorescein diacetate succinimidyl ester (CFSE; eBioscience), and measured proliferation by dye dilution.

Gene expression

We designed the NanoString codeset based on GWAS SNPs associated with CeD, RA, and T1D as of April 2012. As the numbers of associated loci with autoimmune diseases continuously expand, we refer the reader to ImmunoBase (https://www.immunobase.org) for up-to-date disease regions. For each locus, we defined a region of interest implicated by the GWAS lead SNP [36]. We identified the furthest SNPs in LD in the 3' and 5' directions ($r^2 > 0.5$). We then extended outward in each direction to the nearest recombination hotspot. If no genes were found in this region, we extended an additional 250kb in each direction. All genes overlapping this region were considered implicated by the locus. The final NanoString codeset (prior to expression data quality control) included 312 genes, including 270 genes near SNPs associated with 157 RA, CeD, and T1D through GWAS, 26 genes of immunological interest, and 15 reference genes with minimal change in expression after TCR stimulation.

After quality control, 215 genes remained. Of all 225 subjects in the study, 187 subjects passed gene expression quality control for both resting and stimulated cells..

Genotype principal component analysis

To control for any potential population stratification, we adjusted all association tests using the top five principal components of our genome-wide SNP data. Principal components were generated via EIGENSTRAT using unsupervised analysis (no reference populations were used). The top five PCs explained 6.88% (2.08%, 1.27%, 1.20%, 1.17%, and 1.16%, respectively) of the total variance. After controlling for these five PCs, the lambda GC for CD4 T_{EM} proportion association was 1.008; that of division index was 1.001.

Cis-eQTL analysis

For each gene-SNP pair, we applied linear regression using the first five principal components of the genotype data and gender as covariates. As such, normalized expression = $\beta_0 + \beta_1$*allelic dosage + $\beta_{2*}PC_1$ + $\beta_{3*}PC_2$ + $\beta_{4*}PC_3$ + $\beta_{5*}PC_4$ + $\beta_{6*}PC_5$ + β_7*(factor)gender. To adjust for multiple hypothesis testing while taking into consideration the correlation among SNPs within each locus, we calculated a permutation-based *p*-value for each SNP. We performed 10,000 permutations of the residual expression values. We reported each SNP's *p*-value the proportion of permutation *P* value smaller than the analytical *p*-value. For conditional analysis, the vector of allelic dosages of the disease-associated SNP was included as an additional covariate.

Statistical analysis

All linkage disequilibrium calculations (r^2) were based on 1000 Genomes Release 3 European samples. All association tests were performed using Plink v1.07. We considered $p < 5x10^{-8}$ to be genome-wide significant; $p < 5x10^{-5}$ was considered as suggestive. $CD4^+$ T_{EM} abundance and proliferation correlations with age and gender were calculated by multivariate linear model implemented in R-3.0. We calculated two-sample comparisons ($CD4^+$ T_{EM} cell-specific expression between genes, and H3K4me3 h/d scores between SNPs) with the Mann-Whitney test. Details of statistical analyses are described in **Text S1.**

Nanostring data analysis

Data quality control

Raw nCounter data consisted of 343 transcriptional measurements (314 target genes, 15 reference genes, 8 negative controls, and 6 positive controls). Data was available for 265 samples (including replicates) initially with both resting and stimulated data (*n*=530). First, we identified control genes with adequate signal intensity for normalization, we required that the signal intensity of the gene exceeded

double the median of negative control probe intensities in <10 samples; this resulted in 9 pre-defined control genes passing quality control. Then, we removed samples with low intensity by requiring for each sample that:

(1) The mean of the natural log of the 9 control genes >2 (525/530 passing).

(2) The median signal intensity of the of the 314 genes exceeded the median signal intensity of negative controls (512/530 passing).

(3) The mean of the natural log of the 314 measured genes >0.5 (523/530 passing).

(4) The standard deviation divided by the mean of the natural log of the 314 measured genes was >0.5 (509/530).

The resulting data set had 491 remaining samples. Finally, we applied stringent quality control to remove low intensity genes. In order to do this we required that:

(1) The intensity of the gene exceeded double the median of negative probe intensities in >80% of stimulated samples (246/314).

(2) The standard deviation divided by the mean of the natural log of the each gene across samples was >0.3 (292/314).

(3) The standard deviation of the natural log of the each gene across samples was >1 (245/314).

The resulting data set consisted of measurements on 215 genes.

To assess if stimulated and non-stimulated samples separate naturally in expression space, as we would expect with high quality sample measurements, we calculated principal components analysis, after normalizing each gene to a mean of 0 and standard deviation of 1.

Data normalization

For each gene, we normalized resting and stimulated conditions together, assuming that the observed signal was the composite of a true baseline expression value, and a stimulation effect (if the sample is indeed collected from stimulated cells). Fitting the observed intensity data (R_{ij}) in a log additive

model for each gene *j* individually, allows us to determine the residuals for each individual sample *i*, r_{ij}. Thus for each gene *i*, we fit the following model.

$$log(R_{i,j}) = \bar{X}_j + I(stim_i = 1) \cdot \bar{S}_j + r_{i,j}$$

where $stim_i$ represents a binary variable which is non-zero only for stimulated samples, $\bar{X}_j$, is the mean log expression of gene *j* across samples at baseline, $\bar{S}_j$ is the mean log fold change with stimulation, and r_{ij} is the residual expression of gene *j* for sample *i.* With this formalism, the log baseline expression, log fold change with stimulation, and statistical significance for each of these parameters being >0 can be estimated with a simple linear regression model. The r_{ij} residuals can be used to conduct association studies across individuals.

In addition to individual differences, we are cognizant residuals effects might be capturing variability in mRNA content, batch effects, reagent quality, and global shifts in expression for individual samples. In order to control for these effects, and maximize the extent to which residuals represented individual expression differences, we included additional confounder variables that might capture these effects:

$$log(R_{i,j}) = \bar{X}_j + I(stim_i = 1) \cdot \bar{S}_j + r_{i,j} log(R_{i,j}) + \sum \beta_c c_i$$

where c_i is a series of one or more confounder variables that influences gene expression in a log linear fashion. Here the confounders that we tested as covariates in this framework included the mean of the log positive control intensities, the mean of the log control gene intensities, the chip effect (12 samples are run together), the effect of the position of the chip (both row and column), and the principal components for stimulated data. To assess the impact of each confounder variable we assessed the sum square of the residual, with the aim to use confounders to reduce the total residual across all samples and genes:

$$\sum_i (r_{i,j})^2$$

Briefly, we observed that the mean of the log control gene intensities (cg_i) for each sample explained 52% of the sum-squared residuals – more than any other individual variable. Addition of log

positive control intensities, the chip effect (12 samples are run together), or the effect of the position of the chip (both row and column) did not reduce residuals substantially beyond the reduction of cg_i (<7%); we concluded that most of these effects are either minimal or captured by cg_i. However, we did not that the addition of principal components did reduce residuals further. Briefly normalizing expression data for each sample to have a mean of 0 and standard deviation of 1, we calculated principal components across resting and stimulated samples separately. We observed that adding the top two components for each stimulated (p^s) and non-stimulated (p^n) samples explained an additional 25% of the total sum-square residual, adding additional components only improved sum squared residual explained only incrementally (<2.3% per pair of components added).

In final form we implemented the following normalization scheme:

$$log\left(R_{i,j}\right) = \bar{X}_j + I(stim_i = 1) \cdot \bar{S}_j + r_{i,j} + \cdots$$

$$\gamma \cdot cg_i + I(stim_i = 1) \cdot \sum_{k=1}^{2} \pi_k^s . p_k^s + I(stim_i = 0) \cdot \sum_{k=1}^{2} \pi_k^n . p_k^n$$

where g is the linear effect for cg_i and p is the linear effect for each of the principal component variables, $p_k{}^n$. In aggregate the use of two pairs of principal components and the log average intensity of control genes explained 77% of the sum square of residuals after linear fit.

Assessing biological and technical reproducibility

After obtaining residual expression values for 215 genes for each individual under resting and stimulated conditions, correlations of residuals between technical and biological replicate pairs were assessed. First, a Pearson's r for each pair of normalized replicates was calculated. Then, technical replicates of samples collected from resting cells, technical replicates of samples collected from stimulated cells, biological replicates of samples collected from resting cells, and biological replicates of samples collected from stimulated cells were separately averaged.

To assess significance for each of these conditions, an equal number of pairs from the total pool of assayed samples, matching for stimulation status, were randomly identified. Pairs were restricted so that the data was not obtained from the same individual. For each of the four conditions, 1,000,000 sets of pairs were sampled. Significance was assessed by quantifying the number of instances the averaged correlation of randomly drawn pairs exceeded the observed averaged correlation.

Genotyping and imputation

Each subject was genotyped using the Illumina Infinium Human OmniExpress Exome BeadChips, which includes genome-wide genotype data as well as genotypes for rare variants from 12,000 exomes as well as common coding variants from the whole genome. In total, 951,117 SNPs were genotyped, of which 704,808 SNPs are common variants (minor allele frequency [MAF] > 0.01) and 246,229 are part of the exomes. The genotype success rate was greater than or equal to 97%. Rigorous quality control was applied that included 1) gender misidentification, 2) subject relatedness, 3) Hardy-Weinberg Equilibrium testing, 4) use concordance to infer SNP quality, 5) genotype call rate, 6) heterozygosity outlier, and 7) subject mismatches. 1,987 SNPs with a call rate < 95%, 459 SNPs with Hardy-Weinberg equilibrium $P < 10^{-6}$, and 63,781 SNPs with MAF < 0.01 were excluded.

For each gene, the 500kb region (250kb to the 3' and 5' direction) around the transcription start site (hg19) was selected and 1000 Genomes SNPs were imputed into the genome-wide SNP data using BEAGLE Version 3.3.2. The European samples from 1,000 Genomes were used as the reference panel. Markers that had MAF < 0.05 in the reference panel as well as all indels were excluded. After imputation, markers with a BEAGLE $R^2 < 0.4$ or MAF < 0.01 in the imputed samples were excluded.

***Cis*-eQTL analysis**

174 subjects had both genotyping and Nanostring expression data and were included in the eQTL analysis. Analyses were performed using R. For each gene, at rest and after simulation, each SNP within

250kb to the 3' or 5' direction of the transcription start site was assessed for *cis*-eQTLs using the residuals of the gene expression matrix. The imputed dosage, rather than the called minor allele number, was used to perform the linear regression. For each gene-SNP pair, a linear regression was perform, where normalized expression = $\beta_0 + \beta_1$*allelic dosage + $\beta_{2*}PC_1 + \beta_{3*}PC_2 + \beta_{4*}PC_3 + \beta_{5*}PC_4 + \beta_{6*}PC_5 + \beta_7$*(factor)gender. To adjust for multiple hypothesis testing and taking into consideration the correlation among SNPs within the loci, a permutation-based *P* value for each SNP was reported. We performed 10,000 permutations per gene. In each round, the residual expression values of the samples were permuted, and the lowest *P* value achieved by any of the SNPs was recorded. The proportion of permutation *P* value smaller than the analytical *P* value was reported.

We reported the lead SNP per gene with the most significant P-value. Based on locus-wide permutation p-values of all the top SNPs, we used a cut-off of false discovery rate < 0.05, and considered those passing this threshold to be significant.

Conditional analysis

For each gene near a SNP within a densely genotyped locus associated to CeD, RA, or T1D, conditional analysis was performed**.** The dosage of the associated SNP ("dzSNP") was used as a covariate, thus normalized expression = $\beta_0 + \beta_1$*allelic dosage + $\beta_{2*}PC_1 + \beta_{3*}PC_2 + \beta_{4*}PC_3 + \beta_{5*}PC_4 + \beta_{6*}PC_5 + \beta_7$*(factor)gender + β_8*$dosage_{dzSNP}$. If more than one disease-associated SNP reside in the same gene, each SNP is conditioned on separately. We repeated the linear regression and permutations to obtain any remaining eQTL signals (FDR < 0.05) independent of the associated SNP.

Comparison between eQTL effect sizes between resting and stimulated states

To systematically compare the β_{rest} and β_{stim} for each gene, we used a *z*-statistic to quantify the probability that they differ. The statistic was defined as $z = \frac{\beta_{stim} - \beta_{rest}}{\sqrt[2]{SE_{stim}{}^2 - SE_{rest}{}^2}}$, where β and SE are the mean and

standard error of the effect size estimate from regression analysis. We than reported the *p*-value (two-tailed) assuming that *z* is distributed as standard normal.

Enrichment of chromatin-mark overlap

For each SNP with the strongest association to each of the 158 genes in stimulated cells, we calculated an "*h/d*" score based on the distance to and the size of nearest H3K4me3 peak to the SNP in primary CD4 memory T cells. We first identify all SNP variants in LD ($R^2 > 0.8$) to the lead SNP, then locate the nearest H3K4me3 peak to any of the variants. "*H*" is the height of the peak, and "d" is the physical distance in units of base pairs to the peak. We calculated the ratio of *h/d* for each lead SNP.

Quantification of T_{EM} cell relative abundance

Enriched CD4 T cells labeled with antibodies against CD45RA, CD45RO, and CD62L were gated automatically in intensity space via clustering by mixture modeling. Each sample was clustered using forward- (FSC) and side-scatter (SSC) to extract a purer lymphocyte population. Subsequently, a three-dimensional mixture model was fitted to each sample with 7 clusters. The $CD45RA^{-}CD45RO^{high}CD62L^{-/low}$ cluster was annotated as the T_{EM} population. In a subset of samples, a small CD45RA/CD45RO/CD62L triple-negative population was identified, which was assumed to be non-lymphocytic debris and subtracted from the extracted lymphocyte population. T_{EM} abundance was calculated as the percentage of all extracted lymphocytes based on FSC/SSC (excluding any debris).

Quantification of T_{EM} cell proliferation

The CFSE intensity peak present in the pooled resting wells for each subject was modeled as a single Gaussian distribution. Its mean and variance were then used to initialize the location of the first component (undivided cells) and the variance of all components in the stimulated wells. The CFSE dilution peaks from stimulated wells were fitted using a one-dimensional mixture model of multiple Gaussian

components of equal peak-to-peak distance and equal variance via a gradient descent optimization algorithm. A maximum of six components (five divisions) was fitted to each stimulated well. All peaks were initialized as equal in weight. The location and variance of the first (undivided) was initialized to that of the single peak of unstimulated sample. The initial distance between peaks was initialized to 250. Each iteration updated three parameters of each component (mean, variance, and mixing proportion). The algorithm converged when the residual improved by an amount less than a precision threshold (0.1% of the previous iteration) or until a maximum of 1,000 iterations was reached.

Let the number of cells in each of the *N* components (mixing proportions x total cell count) of a stimulated sample be represented by the vector $\{G_0, G_1, G_2 \ldots G_{N-1}\}$, where G_0 is the number of cells that underwent zero divisions during the incubation period. Let A be the total number of cells at the start of the incubation period. Let B be the total number of divisions that all cells underwent during the incubation period. Let C be the total number of cells that underwent at least one division.

$$A = \sum_{i=0}^{N-1} G_i/2^i$$

$$B = \sum_{i=0}^{N-1} G_i/2^i \times i$$

$$C = A - G_0$$

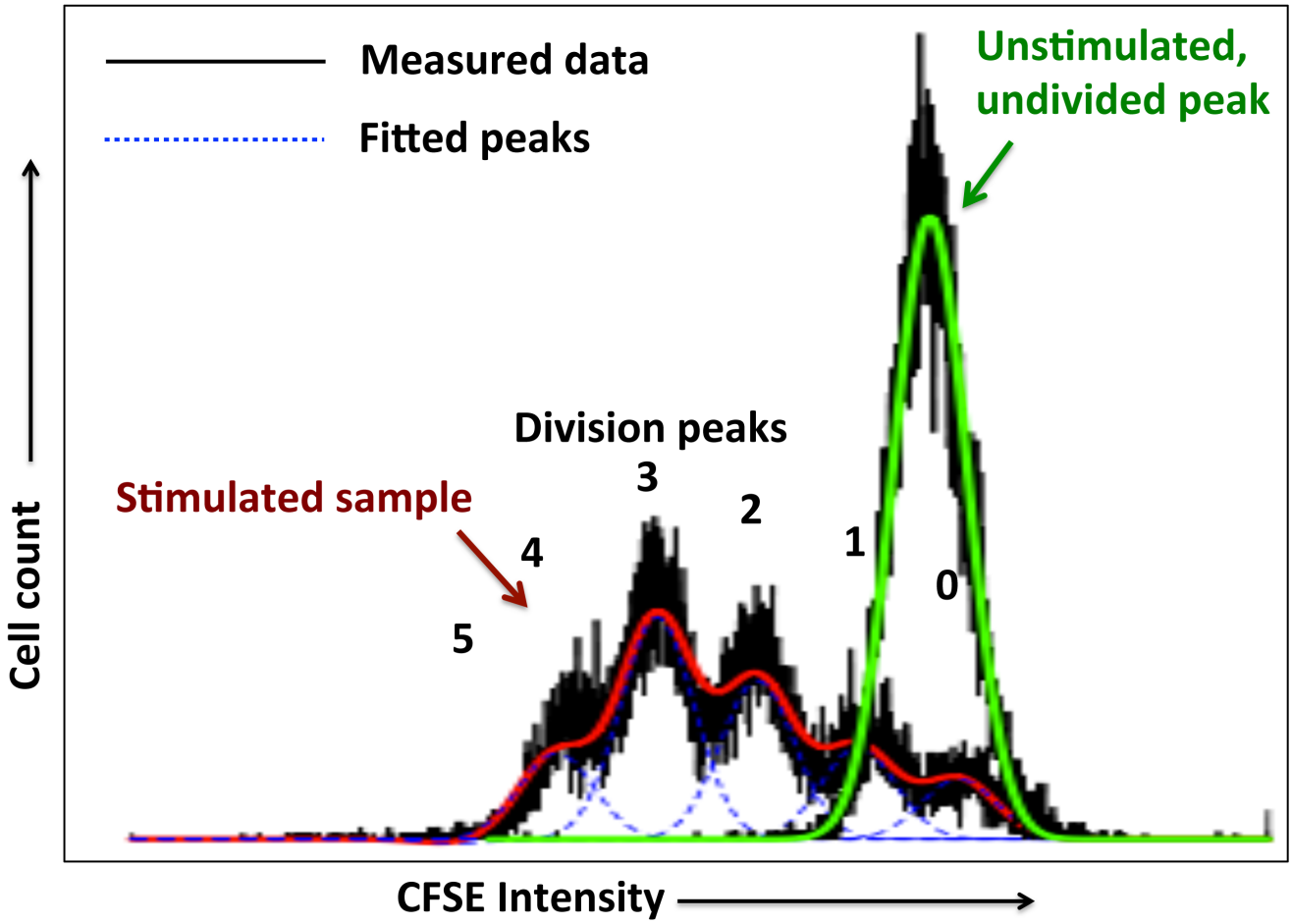
Division index = B/A. Proliferation index = B/C. Since each sample was assayed in three replicates, the average proliferation and division indices of all three replicates were reported. An example of fitted division peaks is shown in **Figure 4.11.**

Genome-wide association testing to CD4 T_{EM} abundance and proliferative response

Each genotyped SNP was tested for association with each quantitative trait using linear regression. For relative CD4 T_{EM} abundance, gender (as factor), age (per year), and the top five genotypic data principal components were included as covariates. For proliferation index and division index, relative CD4 T_{EM} cell

Figure 4.11. Modeling of the CFSE intensity peaks. The CFSE intensity peak present in the pooled resting wells for each subject (data underlying the green fitted curve) was modeled as a single Gaussian distribution. Its mean and variance were then used to initialize the location of the first component (undivided cells) and the variance of all components in each of the stimulated wells (data underlying the red fitted curve). The CFSE dilution peaks from stimulated wells were fitted using a one-dimensional mixture model of multiple Gaussian components of equal peak-to-peak distance and equal variance via a gradient descent optimization algorithm. A maximum of six components (five divisions) was fitted to each stimulated well; the weight of each component was allowed to be 0.

Figure 4.11. Modeling of the CFSE intensity peaks (Continued).



abundance and the top five genotypic data principal components were included as covariates. We considered $5x10^{-8}$ as the genome-wide significance threshold.

Resting gene expression association to proliferative response

We used a permutation-based framework to test whether individual gene transcript levels in resting CD4 Tem cells predict proliferative response. We calculated the correlation coefficient between individual residual differences for each gene, and for T cell proliferative response. In order to assess significance, we simply permuted data on T cell proliferation 10^6 times and calculated proliferative response. The significance *p*-value is simply the proportion of instances where the absolute value of the observed correlation coefficient was exceeded by the absolute value of a coefficient resulting from permutation.

Gene set enrichment analysis

In order to assess whether individual genes were enriched or depleted, we compiled and curated data on gene ontology (GO) code (reference). Briefly, for each gene we assigned it a GO code if the gene or one of its homologous genes was explicitly assigned the code, or if its descendants in the GO tree [36]. In total this resulted in a total of 20,687 genes. We conducted enrichment analysis in those genes that had at least 1 annotation. To select codes for subsequent analysis, we examined only those codes were present in >5, but absent in >5 genes in our data set. To assess enrichment we implemented GSEA as described in Subramanian *et al*. [28], with p=0.

Data access

We make all phenotypic data (expression, peripheral abundance, and proliferation) along with eQTL results publicly available online (http://immunogenomics.hms.harvard.edu/CD4eqtl.html). Genome-wide genotype data will become available through dbGAP and through the ImmVar project. These data are

potentially useful to investigators wishing to assess the potential of genetic variants in altering these molecular phenotypes.

Acknowledgments

We thank Drs. Yukinori Okada, Buhm Han, and Deepak Rao for helpful discussions.

REFERENCES

1. Masopust, D., et al., *Preferential localization of effector memory cells in nonlymphoid tissue.* Science, 2001. **291**(5512): p. 2413-7.

2. Fritsch, R.D., et al., *Abnormal differentiation of memory T cells in systemic lupus erythematosus.* Arthritis Rheum, 2006. **54**(7): p. 2184-97.

3. Zhou, X., et al., *Instability of the transcription factor Foxp3 leads to the generation of pathogenic memory T cells in vivo.* Nat Immunol, 2009. **10**(9): p. 1000-7.

4. Oling, V., et al., *Autoantigen-specific memory CD4+ T cells are prevalent early in progression to Type 1 diabetes.* Cell Immunol, 2012. **273**(2): p. 133-9.

5. Sattler, A., et al., *Cytokine-induced human IFN-gamma-secreting effector-memory Th cells in chronic autoimmune inflammation.* Blood, 2009. **113**(9): p. 1948-56.

6. Nica, A.C., et al., *Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations.* PLoS Genet, 2010. **6**(4): p. e1000895.

7. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.

8. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-43.

9. Trynka, G. and S. Raychaudhuri, *Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases.* Curr Opin Genet Dev, 2013. **23**(6): p. 635-41.

10. Fairfax, B.P., et al., *Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* Nat Genet, 2012. **44**(5): p. 502-10.

11. Lee, M.N., et al., *Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells.* Science, 2014.

12. Stranger, B.E., et al., *Patterns of cis regulatory variation in diverse human populations.* PLoS Genet, 2012. **8**(4): p. e1002639.

13. Hu, X., et al., *Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets.* Am J Hum Genet, 2011. **89**(4): p. 496-506.

14. Trynka, G., et al., *Chromatin marks identify critical cell types for fine mapping complex trait variants.* Nat Genet, 2013. **45**(2): p. 124-30.

15. Trynka, G., et al., *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nat Genet, 2011. **43**(12): p. 1193-201.

16. Eyre, S., et al., *High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis.* Nat Genet, 2012. **44**(12): p. 1336-40.

17. Jostins, L., et al., *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.* Nature, 2012. **491**(7422): p. 119-24.

18. Raj, T., et al., *Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes.* Science, 2014. **344**(6183): p. 519-523.

19. Hu, X., et al., *Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells.* Proc Natl Acad Sci U S A, 2013. **110**(47): p. 19030-5.

20. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.

21. Hyatt, G., et al., *Gene expression microarrays: glimpses of the immunological genome.* Nat Immunol, 2006. **7**(7): p. 686-91.

22. Saule, P., et al., *Accumulation of memory T cells from childhood to old age: central and effector memory cells in CD4(+) versus effector memory and terminally differentiated memory cells in CD8(+) compartment.* Mech Ageing Dev, 2006. **127**(3): p. 274-81.

23. Nafisi, H., et al., *GAP1(IP4BP)/RASA3 mediates Galphai-induced inhibition of mitogen-activated protein kinase.* J Biol Chem, 2008. **283**(51): p. 35908-17.

24. Tugendreich, S., et al., *CDC27Hs colocalizes with CDC16Hs to the centrosome and mitotic spindle and is essential for the metaphase to anaphase transition.* Cell, 1995. **81**(2): p. 261-8.

25. Sommers, C.L., et al., *Bam32: a novel mediator of Erk activation in T cells.* Int Immunol, 2008. **20**(7): p. 811-8.

26. Al-Alwan, M., et al., *Bam32/DAPP1 promotes B cell adhesion and formation of polarized conjugates with T cells.* J Immunol, 2010. **184**(12): p. 6961-9.

27. Orru, V., et al., *Genetic variants regulating immune cell levels in health and disease.* Cell, 2013. **155**(1): p. 242-56.

28. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

29. Teare, M.D., et al., *Allele-dose association of the C5orf30 rs26232 variant with joint damage in rheumatoid arthritis.* Arthritis Rheum, 2013. **65**(10): p. 2555-61.

30. Faure, G., et al., *T cell subsets in the blood of rheumatoid arthritis patients in clinical remission.* Arthritis Rheum, 1982. **25**(12): p. 1507-9.

31. Henderson, L.A., et al., *A161: Novel 3-Dimensional Explant Method Facilitates the Study of Lymphocyte Populations in the Synovium and Reveals a Large Population of Resident Memory T cells in Rheumatoid Arthritis.* Arthritis Rheumatol, 2014. **66 Suppl 11**: p. S209.

32. Hussein, M.R., et al., *Alterations of the CD4(+), CD8 (+) T cell subsets, interleukins-1beta, IL-10, IL-17, tumor necrosis factor-alpha and soluble intercellular adhesion molecule-1 in rheumatoid arthritis and osteoarthritis: preliminary observations.* Pathol Oncol Res, 2008. **14**(3): p. 321-8.

33. Kotzin, B.L., et al., *Changes in T-cell subsets in patients with rheumatoid arthritis treated with total lymphoid irradiation.* Clin Immunol Immunopathol, 1983. **27**(2): p. 250-60.

34. Matsuki, F., et al., *CD45RA-Foxp3(high) activated/effector regulatory T cells in the CCR7 + CD45RA-CD27 + CD28+central memory subset are decreased in peripheral blood from patients with rheumatoid arthritis.* Biochem Biophys Res Commun, 2013. **438**(4): p. 778-83.

35. Syrjanen, S.M. and K.J. Syrjanen, *Enumeration of T cell subsets with monoclonal antibodies in minor salivary glands of patients with rheumatoid arthritis.* Scand J Dent Res, 1984. **92**(4): p. 275-81.

36. Raychaudhuri, S., et al., *Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions.* PLoS Genet, 2009. **5**(6): p. e1000534.

CHAPTER 5

Fine-mapping the HLA genetic associations in type 1 diabetes

Additive and interaction effects at three key amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes genetic risk

Xinli Hu[1-6]*, Aaron J. Deutsch[1-5]*, Tobias L. Lenz[2,7], Suna Onengut-Gumuscu[8], Buhm Han[2,4,9], Wei-Min Chen[8], Joanna M.M. Howson[10], John A. Todd[11], Paul I.W. de Bakker[12], Stephen S. Rich[8], Soumya Raychaudhuri[1-4,13†]

1. Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
2. Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
3. Partners Center for Personalized Genetic Medicine, Boston, MA, USA
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
5. Harvard-MIT Division of Health Sciences and Technology, Boston, MA USA
6. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA
7. Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Ploen, Germany
8. Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
9. Asan Institute for Life Sciences, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
10. Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK
11. JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research,

University of Cambridge, Wellcome Trust/MRC Building, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK

12. Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht 3584 CG, the Netherlands; Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht 3584 CG, the Netherlands
13. Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK

* These authors contributed equally to this work.

Correspondence:

Soumya Raychaudhuri (soumya@broadinstitute.org)

ABSTRACT

Variation in the human leukocyte antigen (HLA) genes within the major histocompatibility complex (MHC), particularly the class II *HLA-DRB1* and *HLA-DQB1* loci, accounts for one-half of the familial clustering in type 1 diabetes (T1D). Amino acid changes in the HLA-DR and -DQ molecules mediate most of the genetic component of this familial aggregation, but extensive linkage disequilibrium complicates precise localization of independent effects. Using 18,832 T1D case-control samples, we found that polymorphisms at three amino acid positions encoded by *HLA-DQB1* and *HLA-DRB1* parsimoniously explain the T1D genetic risk. The previously known position 57 in DQβ1 (p=10^{-1355}) alone explained 15.2% of the total phenotypic variance, or 55% of the variance captured by the *HLA-DRB1-DQA1-DQB*1 locus. Significantly, variation at positions 13 (p=10^{-721}) and 71 (p=10^{-95}) of DRβ1 further increased the explained proportion to 26.9% (90% of that explained by *HLA-DRB1-DQA1-DQB*1). Additionally, we observed statistically significant interactions in 11 of 21 pairs of common *HLA-DRB1-DQA1-DQB1* haplotypes (p=1.6x10^{-64}) that included and extended beyond the known interaction between *HLA-DR3* and *-DR4*. These results have important mechanistic significance: the two DRβ1 amino acid positions strongly implicate pocket 4 in the antigen-binding groove, thus pointing clearly to a protein structural feature, in addition to the DQ P9 pocket, critical in mediating T1D risk.

INTRODUCTION

Type 1 diabetes (T1D) is a highly heritable autoimmune disease that results from T cell-mediated destruction of the insulin-producing pancreatic β cells. The worldwide incidence of T1D ranges from 0.1 per 100,000 persons in China to >36 per 100,000 in parts of Europe, and has been steadily increasing[1]. Many autoimmune diseases, including T1D, rheumatoid arthritis (RA), celiac disease, and multiple sclerosis, have the majority of their genetic risk attributed to variants in the human leukocyte antigen (HLA) genes within the major histocompatibility complex (MHC) region located on chromosome 6p21.3[2-4]. HLA genes encode surface proteins that display antigenic peptides to effector immune cells in order to regulate self-tolerance and downstream immune responses. Autoimmune risk conferred by HLA is likely the result of variation in amino acid residues at specific positions within the antigen-binding grooves, which may then alter the repertoire of presented peptides[5-8]. In T1D, the largest allelic associations are in *HLA-DRB1-DQA1-DQB1*, a three-gene "superlocus" that encodes HLA-DR and -DQ proteins[9, 10], and additional associations have been identified in the *HLA-A, -B, -C, and -DP* genes[11-14].

Todd *et al.* initially identified strong T1D risk conferred by non-aspartate residues at position 57 of HLA-DQβ1[15]. However, this amino acid position alone does not fully explain the HLA risk in T1D. Subsequently, many amino acid positions in DQβ1 and DRβ1 have been hypothesized to modify T1D risk[16]; but extensive linkage disequilibrium (LD) spanning the 4 Mb MHC region makes it challenging to pinpoint the specific risk variants. Furthermore, certain heterozygous genotypes confer the greatest disease risk for T1D[13, 17-19], consistent with synergistic interactions between classical HLA alleles. Despite evidence of non-additive effects within the MHC on autoimmune disease risk, interactions have not been comprehensively examined in T1D. Mechanistic investigation of how autoantigens interact with HLA proteins could become feasible if specific amino acid positions and their interactions were understood. In this

study, we utilized recently established accurate genotype imputation methods to examine a large case-control sample, and rigorously identified independent amino acid positions as well as interactions within the HLA that account for T1D risk (see **Figure 5.1** for a schematic of analyses).

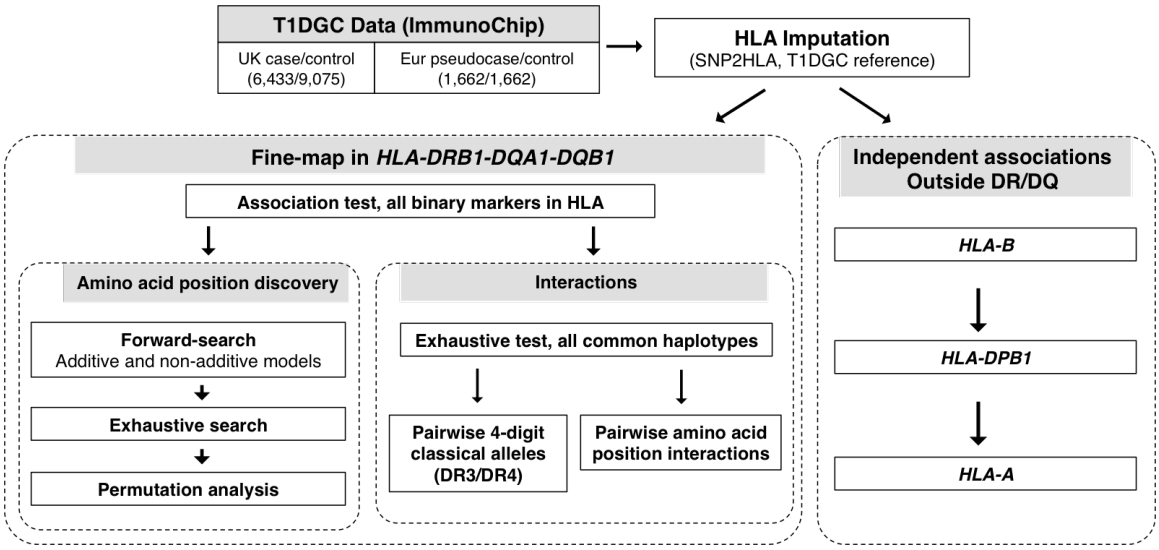


Figure 5.1. Schematic of analyses procedure followed in the study.

RESULTS

HLA Imputation and association testing

We fine-mapped the MHC region in a collection of 8,095 T1D cases and 10,737 controls genotyped with the ImmunoChip array, provided by the Type 1 Diabetes Genetics Consortium (T1DGC)[20-22]. The dataset included (1) case-control samples collected in the United Kingdom (UK), and (2) a pseudocase-control set derived from European families (Eur). Using a set of 5,225 individuals with HLA alleles genotyped by PCR as a reference[22], we accurately imputed 8,617 binary markers (with minor allele frequency > 0.05%) between ~29 Mb and ~33 Mb (the 4 Mb classical MHC region) on chromosome 6p21.3 with SNP2HLA software[21]. The resulting data included 7,242 SNPs, 260 2- and 4-digit classical alleles, and amino acid polymorphisms at 399 positions in eight HLA genes (*HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1,* and *-DPB1*. We have

previously independently benchmarked the strategy employed in this study for imputation accuracy of the HLA classical alleles and amino acid polymorphisms, using a set of 918 samples that were HLA typed. Starting with SNPs from the ImmunoChip genotyping platform and using the T1DGC reference panel, SNP2HLA obtained an accuracy of 98.4%, 96.7% and 99.3% for all 2-digit alleles, 4-digit alleles, and amino acid polymorphisms, respectively[21].

To test T1D association of a given variant, we used a logistic regression model, assuming the log-odds of disease to be proportional to the allelic dosage of the variant. We also included covariates to adjust for sex and region of origin. As expected, the strongest associations with T1D were within the *HLA-DRB1-DQA1-DQB1* locus. We confirmed that the leading risk variant was the presence of alanine at DQβ1 position 57 (DQβ1#57, $p=10^{-1090}$, OR=5.17; **Figure 5.2A**). In contrast, the single most significantly associated classical allele was *DQB1*03:02* ($p=10^{-840}$), which has an alanine at DQβ1#57, but was much more weakly associated than the amino acid residue itself. **Table 5.1** lists common classical alleles tagged by each residue at key amino acid positions.

Three amino acid positions independently drive T1D-HLA association

Given the strength and complexity of the association within *HLA-DRB1-DQB1-DQA1*, we aimed to first identify independent effects in this locus before examining the rest of the MHC. We assessed the significance of multi-allelic amino acid positions using conditional analysis by forward-search. The most strongly associated position with T1D was DQβ1#57 ($p=10^{-1355}$, **Figure 5.3**), At this position, alanine conferred the strongest risk (OR=5.17; **Figure 5.4**), while the most common residue in controls, aspartic acid, was the most protective (OR=0.16). Conditioning on DQβ1#57, the second independent association was at DRβ1#13 ($p=10^{-721}$). At this position, histidine (OR=3.64) and serine (OR=1.28) confer the strongest risk, while arginine (OR=0.08) and tyrosine (OR=0.28) were protective (**Figure 5.4**). The DRβ1#71 position was the third

Figure 5.2. Independently associated HLA loci to T1D. Each binary marker was tested for T1D association, using the imputed allelic dosage (between 0 and 2). In each panel, the horizontal dashed line marks p=5x10^{-8}. Color gradient of the diamond indicates LD (r^2) to the most strongly associated variant; the darkest shade is r^2=1. A) The strongest associations were located in *HLA-DRB1-DQA1-DQB1*. The single strongest risk variant was alanine at DQβ1#57 (OR=5.17; p=10^{-1090}). B) Adjusting for all *DRB1, DQA1,* and *DQB1* 4-digit classical alleles, the strongest independent signals were in *HLA-B*. The strongest association was to *B*39:06* (OR=6.64, p=10^{-75}). C) Adjusting for *HLA-DRB1-DQA1-DQB1* and *HLA-B,* the next associated variant was *DPB1*04:02 (*OR=0.48, p=10^{-55}). D) The final independent association was in *HLA-A*, led by glutamine at A#62 (OR=0.70, p=10^{-25}). E) We found no residual independent association in the *HLA-C* or *HLA-DPA1*.

Figure 5.2. Independently associated HLA loci to T1D (Continued).

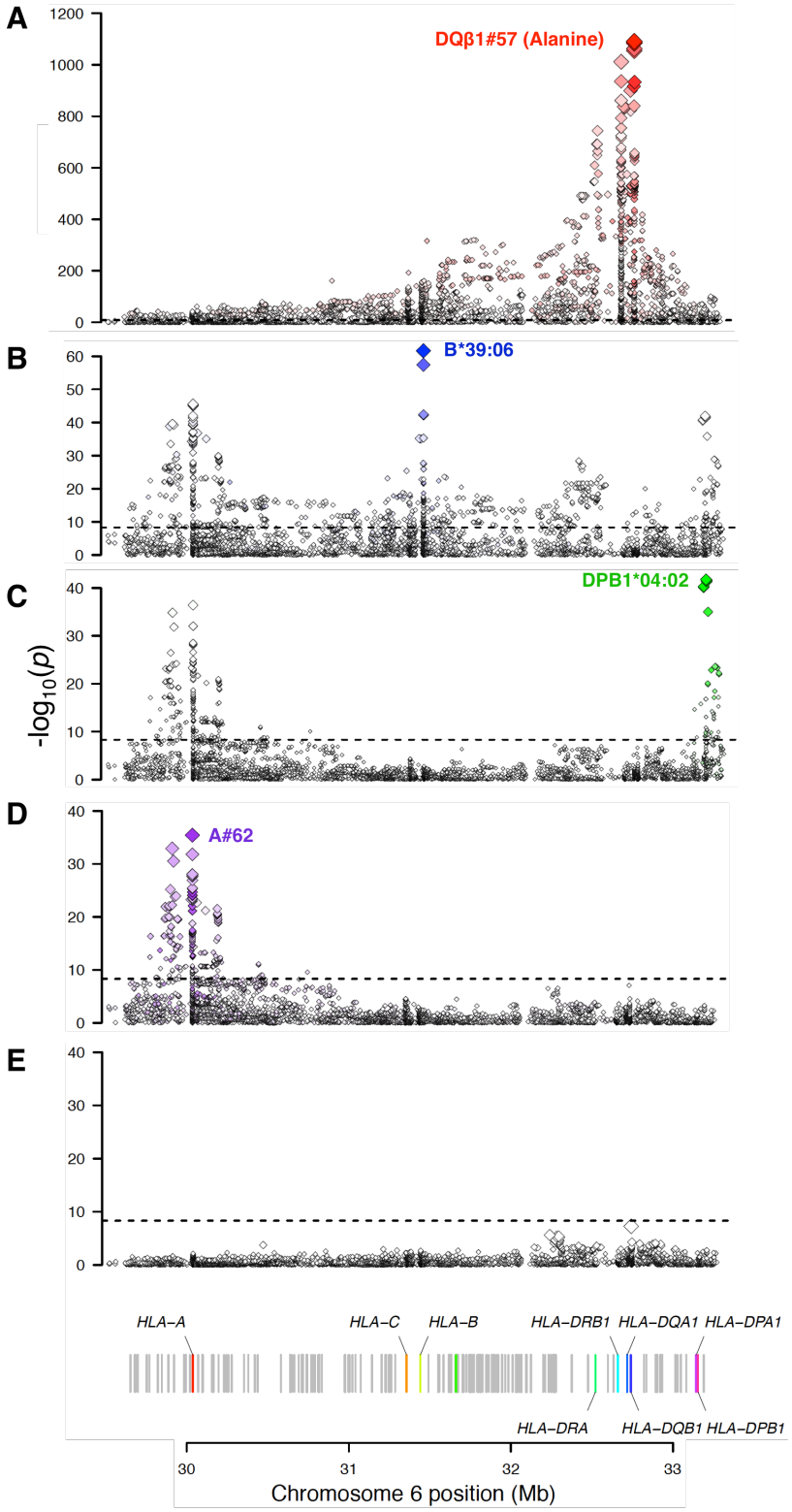
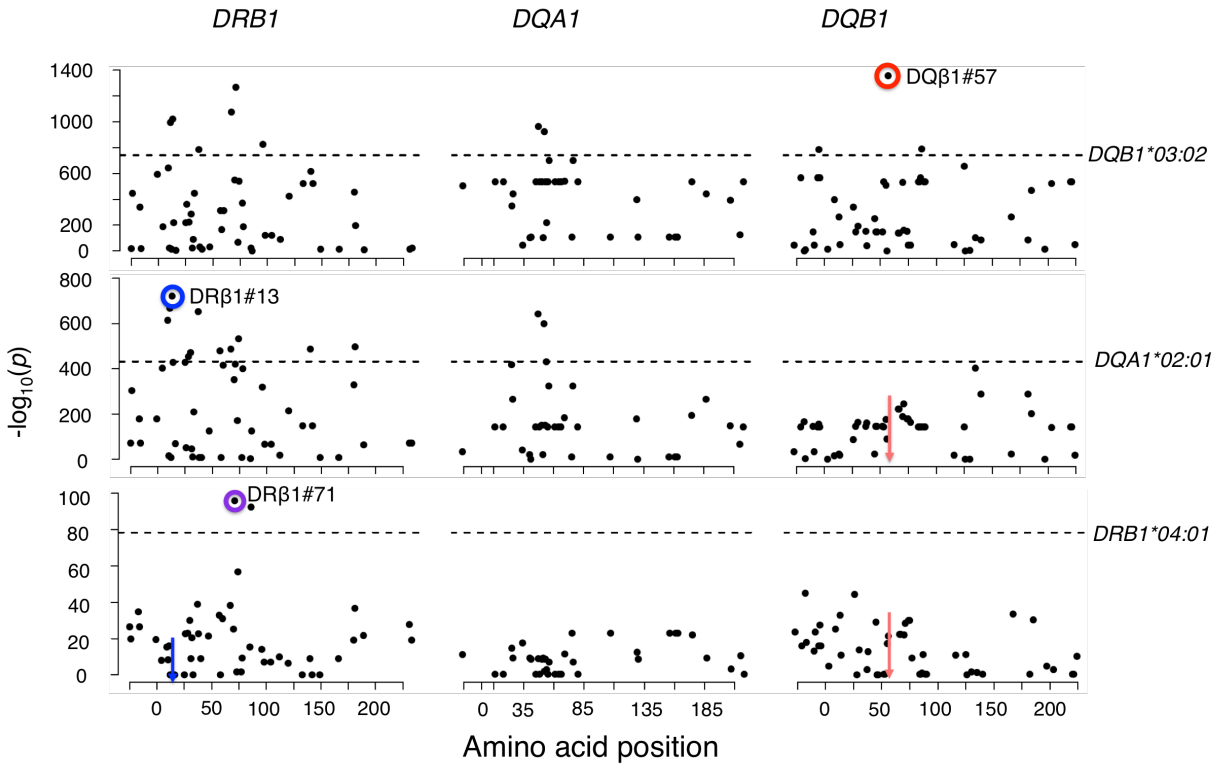


Figure 5.3. Amino acid positions DQβ1#57, DRβ1#13, and DRβ1#71 independently drive T1D risk associated to the *HLA-DRB1-DQA1-DQB1* locus. To identify each independently associated position, we used conditional haplotypic analysis by forward search, using the phased best-guess genotypes. In each panel, the dots mark amino acid positions along the gene (x-axis) and their $\log_{10}$ association *p*-values on the y-axis. The horizontal dashed line marks the $\log_{10}$ *p*-value of most strongly associated classical allele. The most strongly associated signals are circled. The colored arrows indicate positions that have been conditioned on. The most strongly associated position was DQβ1#57 ($p=10^{-1355}$). Conditioning on it, DRβ1#13 was the next independently associated position ($p=10^{-721}$), followed by DRβ1#71 ($p=10^{-95}$). Each position was much more strongly associated than the best classical allele (*DQB1*03:02, DQA1*02:01,* and *DRB1*04:01*, respectively).

Figure 5.3. Amino acid positions DQβ1#57, DRβ1#13, and DRβ1#71 independently drive T1D risk associated to the *HLA-DRB1-DQA1-DQB1* locus (Continued).



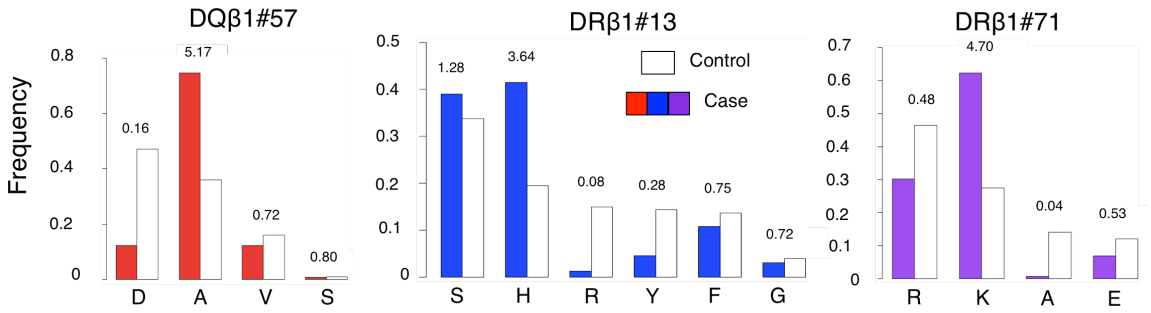


Figure 5.4. Amino acid residue effect sizes. Case (colored) and control (unfilled) frequencies, as well as unadjusted univariate odds ratio, of each residue, at DQβ1#57, DRβ1#13 and DRβ1#71.

independently associated signal (p=10^{-95}); lysine conferred strong risk (OR=4.70), and alanine was strongly protective (OR=0.04). We note that these positions, the risk-conferring amino acid residues indeed tag the DR3 and DR4 haplotypes, which confer the strongest risk among haplotypes. Histidine at position 13 tags *DRB1*04:01* and *04:04*, while serine tags *DRB1*03:01*. Lysine at position 71 tags both *DRB1*03:01* and *DRB1*04:01*. **Table 1** lists the classical alleles tagged by residues at each key amino acid position, and multivariate odds ratios of haplotypes defined by these positions.

Given the reported deviation from log-scale additivity of T1D risk effects in the HLA[19, 23], we wanted to confirm that their contribution did not alter the risk-driving amino acid positions. By repeating the forward-search analysis while including non-additive terms in the regression model, we confirmed that DQβ1#57, DRβ1#13, and DRβ1#71 were again the top three independent signals under the non-additive model as well as the additive model.

Conditioning on these three positions, more than 80 other positions and classical alleles remained highly significant ($p < 10^{-8}$), suggesting the presence of other independent associations. We tested all possible combinations of two, three, and four amino acid positions in *HLA-DRB1-DQA1-DQB1*, and confirmed that DQβ1#57, DRβ1#13, and DRβ1#71 were the best of all 457,450 combinations of three amino acids (p=10^{-2161}). DQβ1#-18 (located within the signal peptide)

Table 5.1. Haplotypes defined by DQβ1#57, DRβ1#13, and DRβ1#71 (control frequency > 0.1%). The three amino acid positions define 21 common haplotypes. We list their multivariate ORs, frequencies in controls and cases, as well as classical 4-digit alleles tagged by each haplotype. See **Table S5B** for multivariate ORs and p-values of all 31 haplotypes formed by DQβ1#57, DRβ1#13, and DRβ1#71.

Haplotype	OR	CtrlFreq	CaseFreq	Classical *DQB1* Alleles	Classical *DRB1* Alleles
A-H-K	2.13	0.050	0.248	0201,0202,0302,0304,0305	0401,0409
A-H-E	1.33	0.005	0.016	0201,0202,0302,0304,0305	0402,0437
A-S-K (Ref)	1.00	0.145	0.332	0201,0202,0302,0304,0305	0301,0302,0304,1303
A-H-R	0.89	0.054	0.107	0201,0202,0302,0304,0305	0403,0404,0405,0406,0407,0408,0410,0411
A-S-E	0.53	0.001	0.001	0201,0202,0302,0304,0305	1102,1103,1301,1302,1304
D-F-R	0.48	0.012	0.014	0301,0303,0401,0402,0503,0601,0602,0603	0101,0102,0901,1001
A-F-R	0.43	0.001	0.001	0201,0202,0302,0304,0305	0101,0102,0901,1001
S-R-R	0.37	0.008	0.007	0502,0504	1601,1602
V-F-R	0.35	0.106	0.085	0501,0604,0609	0101,0102,0901,1001
V-S-E	0.34	0.040	0.030	0501,0604,0609	1102,1103,1301,1302,1304
D-G-R	0.32	0.039	0.029	0301,0303,0401,0402,0503,0601,0602,0603	0801-0806,1201,1202,1404,1415
D-H-K	0.27	0.068	0.042	0301,0303,0401,0402,0503,0601,0602,0603	0401,0409

Table 5.1. Haplotypes defined by DQβ1#57, DRβ1#13, and DRβ1#71 (Continued).

Haplotype	OR	CtrlFreq	CaseFreq	Classical *DQB1* Alleles	Classical *DRB1* Alleles
V-F-E	0.24	0.013	0.006	0501,0604,0609	0103
A-Y-R	0.18	0.103	0.043	0201,0202,0302,0304,0305	0701
D-S-E	0.11	0.058	0.015	0301,0303,0401,0402,0503,0601,0602,0603	1102,1103,1301,1302,1304
D-F-E	0.08	0.004	0.001	0301,0303,0401,0402,0503,0601,0602,0603	0103
D-H-R	0.06	0.017	0.002	0301,0303,0401,0402,0503,0601,0602,0603	0403-0408,0410,0411
D-S-K	0.06	0.010	0.001	0301,0303,0401,0402,0503,0601,0602,0603	0301,0302,0304,1303
D-S-R	0.05	0.083	0.010	0301,0303,0401,0402,0503,0601,0602,0603	1101,1104,1106,1108,1305,1401,1402,1405,1406,1407
D-Y-R	0.03	0.041	0.003	0301,0303,0401,0402,0503,0601,0602,0603	0701
D-R-A	0.02	0.140	0.005	0301,0303,0401,0402,0503,0601,0602,0603	1501

emerged as the fourth most significant association (p=10^{-40}) through forward-search; however, in the exhaustive test, many other combinations of four amino acid positions exceeded the goodness-of-fit of DQβ1#57, DRβ1#13, DRβ1#71, and DQβ1#-18. Therefore, we do not report subsequent positions that emerged through conditional analysis, as we could not confidently claim additional positions as independent drivers of T1D risk.

We wanted to confirm that this combination of three amino acids was not simply tagging effects of specific haplotypes. To this end, we performed a permutation analysis in which we randomly reassigned amino acid sequences corresponding to each *HLA-DRB1, -DQB1,* and *-DQA1* classical allele, and retested for the best amino acid positions (see **Methods**). This approach preserved haplotypic associations; if certain amino acids were tagging associated haplotypes, equally significant amino acid associations would be found in the permuted data. After 10,000 permutations, no combination of permuted amino acids resulted in a model that equaled or exceeded the goodness-of-fit of DQβ1#57/DRβ1#13/DRβ1#71 in our data, as measured by either deviance or p-value (see **Figure 5.5**).

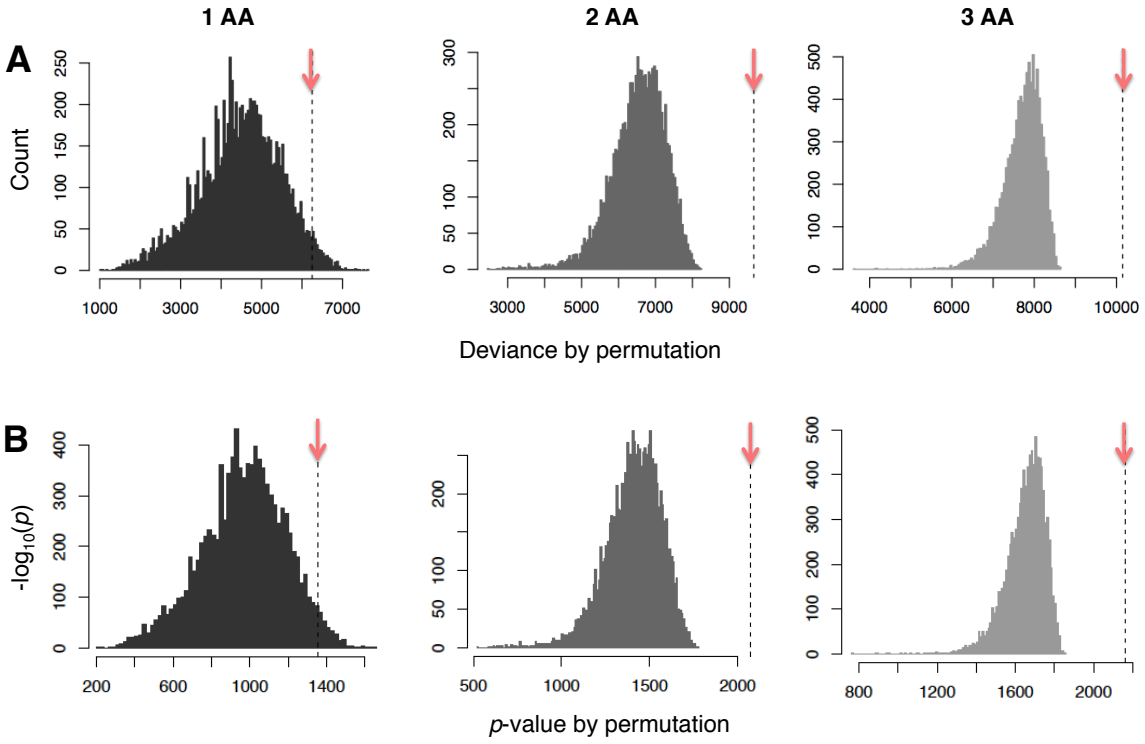
Finally, to ensure that the observed effects were not the results of heterogeneity between the UK and the European subsets, we separately repeated the association analysis in the two subsets. The two sets yielded highly correlated effects sizes for all binary markers (Pearson r=0.952), as well as for all haplotypes formed by residues at DQβ1#57, DRβ1#13, and DRβ1#71 (Pearson r=0.989).

Key amino acids are located in the peptide-binding grooves

DQβ1#57, DRβ1#13, and DRβ1#71 are each located in the peptide-binding grooves of the respective HLA molecule (**Figure 5.6**). DRβ1#13 and #71 line the P4 pocket of HLA-DR, which has been previously implicated in seropositive RA[2], seronegative RA[24], and follicular

Figure 5.5. Haplotype-amino acid sequence permutation analysis ensures DQβ1-57, DRβ1#13 and DRβ1#71 are the independent risk drivers. We performed 10,000 rounds of permutated association analysis; during each permutation, the amino acid sequence corresponding to each *DRB1*, *DQA1*, and *DQB1* classical allele was reassigned, before association analysis. A) Histogram of 10,000 **deviance** values (improvement upon the null model) while testing for the best combination of one, two, and three amino acid positions. Out of 10,000 trials, single position exceeded the deviance achieved by DQβ1#57 3% of the time. No combination of two and three amino acid positions out-performed the fit of DQβ1-57+DRβ1#13, and DQβ1-57+DRβ1#13+DRβ1#71, respectively. The best model achieved by the combination of any three amino acid positions obtained a Δdeviance of 8244.29 (p=10^{-1774}, df=41); in comparison, the model without permutation including DQβ1#57, DRβ1#13, and DRβ1#71 obtained a Δdeviance of 10148.53 (p=10^{-2161}, df=31). Red arrows indicate the deviance achieved by the best combination in actual data. B) Histogram of 10,000 **p-values** while testing for the best combination of one, two, and three amino acid positions. Similarly, 3% of the permuted amino acid positions achieved better p-values than DQβ1#57 in actual data. No combination of two and three amino acid positions out-performed the combinations of DQβ1-57+DRβ1#13, and DQβ1-57+DRβ1#13+DRβ1#71, respectively. Red arrows indicate the p-value achieved by the best combination in actual data.

Figure 5.5. Haplotype-amino acid sequence permutation analysis ensures DQβ1-57, DRβ1#13 and DRβ1#71 are the independent risk drivers (Continued).



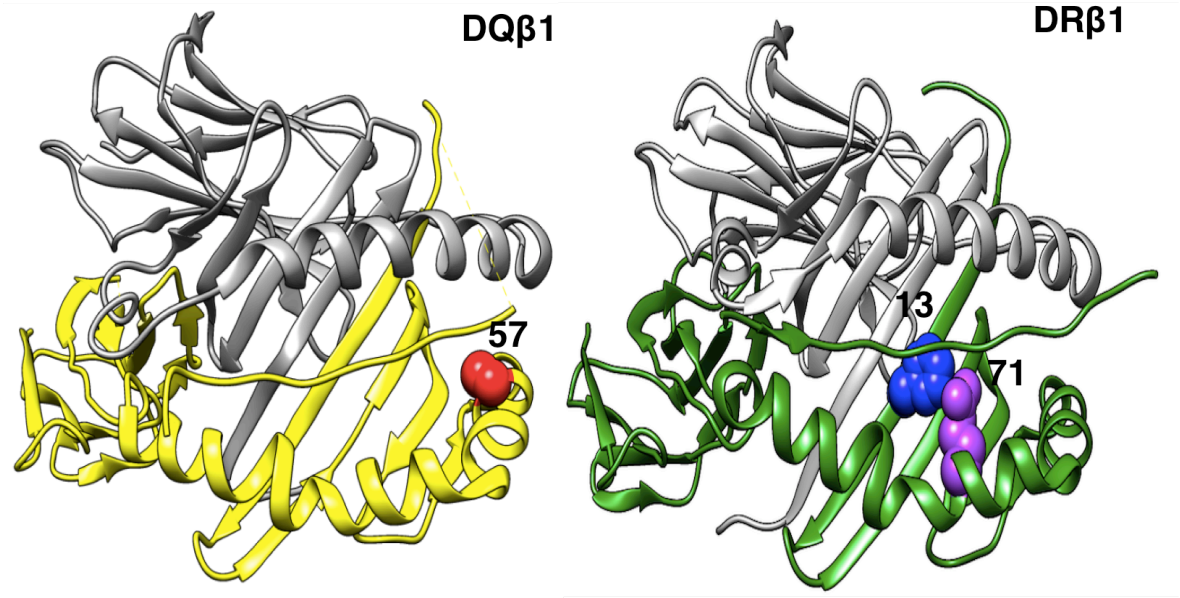


Figure 5.6. DQβ1#57, DRβ1#13 and DRβ1#71 are each located in the respective molecule's peptide-binding groove. DRβ1#13 and #71 line the P4 pocket of the DR molecule.

lymphoma[25]. While DRβ1#13 and DRβ1#71 are both involved in T1D and RA, the effects of individual residues at each position are discordant between the diseases ($p<10^{-232}$; see **Figure 5.7** and **Methods**).

The three amino acid positions explain over 90% of T1D phenotypic variance in *HLA-DRB1-DQA1-DQB1*

We quantified the proportion of phenotypic variance captured by the three amino acid positions using the liability threshold model[26] (see **Methods**). Assuming a T1D prevalence of 0.4%[27], the additive effects of all 67 haplotypes in *HLA-DRB1-DQA1-DQB1* explained 29.6% of the total phenotypic variance. DQβ1#57 alone explained 15.2% of the total variance, while the addition DRβ1#13, and DRβ1#71 increased the proportion explained by 11.7%. Therefore, these three amino acid positions together capture 26.9% of the total variance, which is over 90% of the T1D-

HLA association in this locus (**Figure 5.8**).

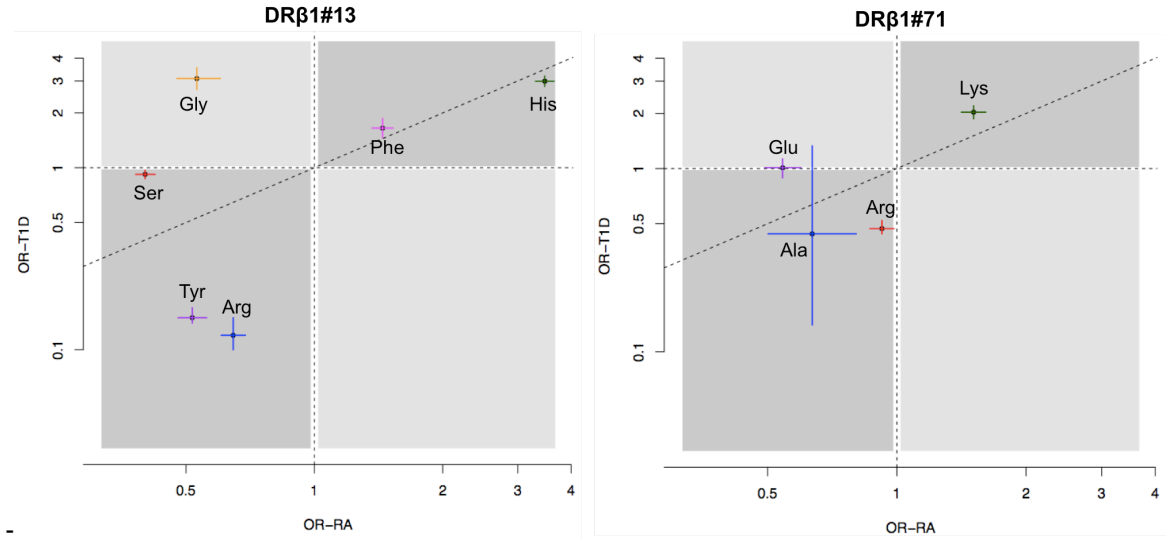


Figure 5.7. DRβ1#13 and #71 residues show discordant effect sizes in RA and T1D. DRβ1#13 and #71, which line the P4 pocket of the DR peptide-binding groove, are indicated in both rheumatoid arthritis (RA) and T1D. However, the individual amino acid residues at each position confer differential risk or protection toward each disease ($p<10^{-230}$). Each cross shows an individual residue's (adjusted) univariate OR (with 95% confidence interval) in RA and T1D. Darker areas indicate the same direction of effect (risk or protection) across the two diseases, while the lighter gray areas indicate opposite effects. The slanted dashed line indicates the identity line on which a residue's effect sizes in both diseases would be equal. At DRβ1#13, serine, tyrosine, and arginine confer relative protection toward each disease; however, they are located far away from the identity line. Glycine is protective toward RA while it confers strong risk toward T1D.

Independent HLA associations in *HLA-B, -DPB1, and -A*

We then sought to identify HLA associations to T1D independent of those in *HLA-DRB1-DQA1-DQB1.* We conservatively conditioned on all *HLA-DRB1, DQA1,* and *DQB1* 4-digit classical alleles to obviate all effects at that locus. We observed the next strongest association across the MHC in *HLA-B,* where the classical allele *HLA-B*39:06* was the most significant signal ($p=10^{-75,}$

Figure 5.8. DQβ1#57, DRβ1#13 and DRβ1#71 explain over 90% of the phenotypic variance explained by the *HLA-DRB1-DQA1-DQB1* locus. Assuming the liability threshold model and a global T1D prevalence of 0.4%, all haplotypes in *HLA-DRB1-DQA1-DQB1* together explain 29.6% of total phenotypic variance. DQβ1#57 alone explains 15.2% of the variance; the addition of DRβ1#13 and #71 increases the explained proportion to 26.9%. Therefore, these three amino acid positions together capture over 90% of the signal within *HLA-DRB1-DQA1-DQB1.* In contrast, variation in *HLA-A, -B,* and *-DPB1* together explain approximately 4% of total variance. Genome-wide independently associated SNPs outside the HLA together explain about 9% of variance; rs678 (in the *INS* gene) and rs2476601 (in *PTPN22*) each explain 3.3% and 0.78%, respectively.

Figure 5.8. DQβ1#57, DRβ1#13 and DRβ1#71 explain over 90% of the phenotypic variance explained by the *HLA-DRB1-DQA1-DQB1* locus (Continued).

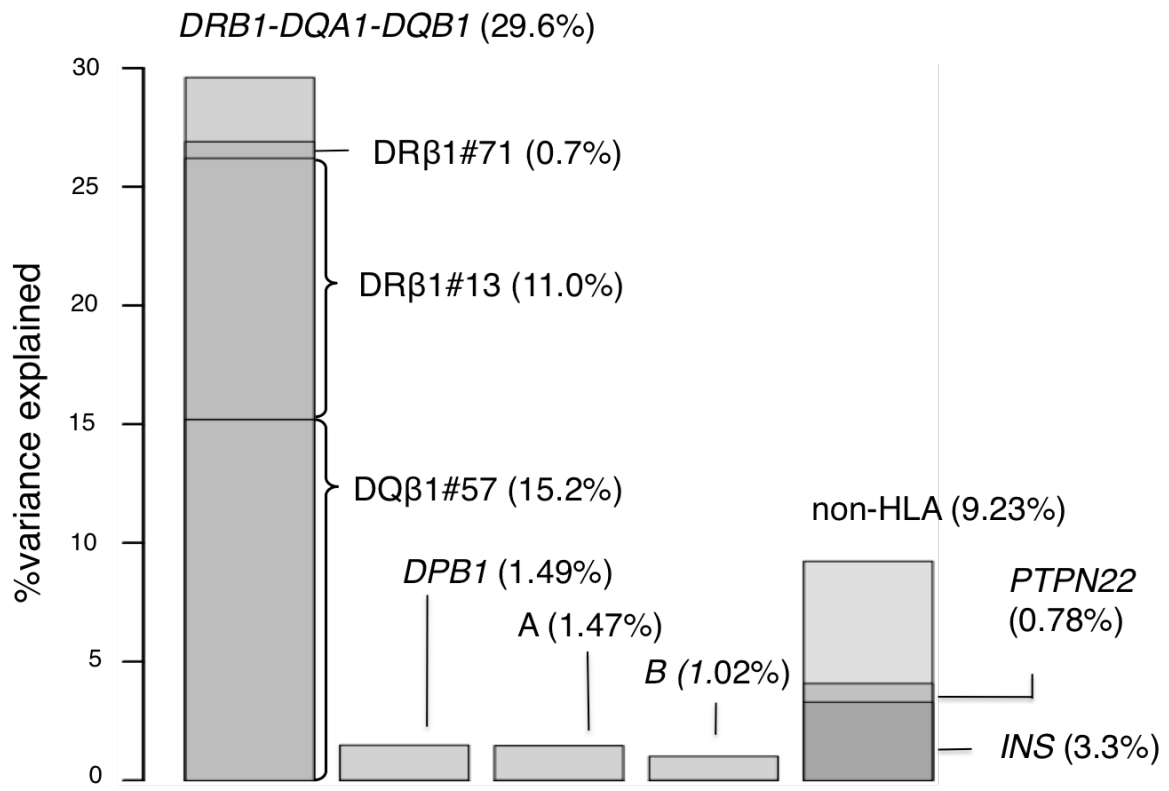


Figure 5.2B)[11]. After adjusting for *B*39:06*, other classical alleles and amino acid positions in *HLA-B* remained significantly associated, including *B*18:01* and *B*50:01.* Upon additionally adjusting for all *HLA–B* alleles, *HLA-DPB1*04:02* was the next strongest independent signal (OR=0.47, $p < 10^{-55}$, **Figure 5.2C**), which is nearly perfectly tagged by methionine at amino acid position 178. Conditioning on *DPB1*04:02*, additional associations were present in *HLA-DPB1*, including position 65 and *DPB1*01:01*. After conditioning on *DPB1* alleles as well, we observed independent effects in *HLA-A* led by position #62 (p=10^{-45}, **Figure 5.2D**); additional signals included *A*03* and *A*24:02*. We observed no independent association with T1D in *HLA-C* or *DPA1* (**Figure 5.2E**). The independent effects of all haplotypes in *HLA-B, -DPB1, and –A* together explained ~4% of the total phenotypic variance. The total T1D risk variance explained by additive effects in the eight HLA genes was ~34%, consistent with the estimates by Speed *et al.*[28].

HLA haplotypic interactions, beyond the DR3/DR4 heterozygote effect, are common in T1D

The previously observed excess risk of T1D in *HLA-DR3/DR4* (*DRB1*03/DRB1*04)* heterozygotes may represent a synergistic interaction between two distinct alleles[23]. Here, we conducted an unbiased search for interactions among all haplotypes within the *HLA-DRB1-DQA1-DQB1* locus. As interactions cannot be observed reliably with infrequent or rare genotypes, we focused this analysis on the seven *HLA-DRB1-DQA1-DQB1* haplotypes with frequencies > 5%; all of these haplotypes had very high imputation accuracies (INFO score > 0.98, see **Methods**).

We tested for interactive effects between all possible pairs of haplotypes using a global multivariate regression model that included 21 interactive terms as well as seven additive terms. The inclusion of interactions in the model produced a statistically significant improvement in fit over the additive model (p=$1.6\text{x}10^{-64}$). Of 21 potential interactions, 11 were significant after correcting for 21 tests (p<$0.05/21=2.4\text{x}10^{-3}$**; Figure 5.9, Table 5.2**). Consistent with previous reports[9, 19], we observed a significant interaction between the *HLA-DR3* haplotype (*DRB1*03:01-*

Figure 5.9. Interactions between common *HLA-DRB1-DQA1-DQB1* haplotypes lead to observed non-additive effects. We exhaustively tested the seven common haplotypes for pairwise interactions. Of the 21 possible pairs, eleven of them showed significant interactive effects. Along the perimeter, each segment represents one haplotype; red or blue color indicates risk or protective additive effect for each haplotype, respectively. Each arch connecting two haplotypes represents a significant interaction. Red indicates additional risk due to the interaction beyond the additive effects; while blue indicates reduced risk (protection) due to the interaction beyond the additive effects. Thickness of the arches represents the effect size of the interaction (thicker red means larger risk while thicker blue means more protective.)

Figure 5.9. Interactions between common *HLA-DRB1-DQA1-DQB1* haplotypes lead to observed non-additive effects (Continued).

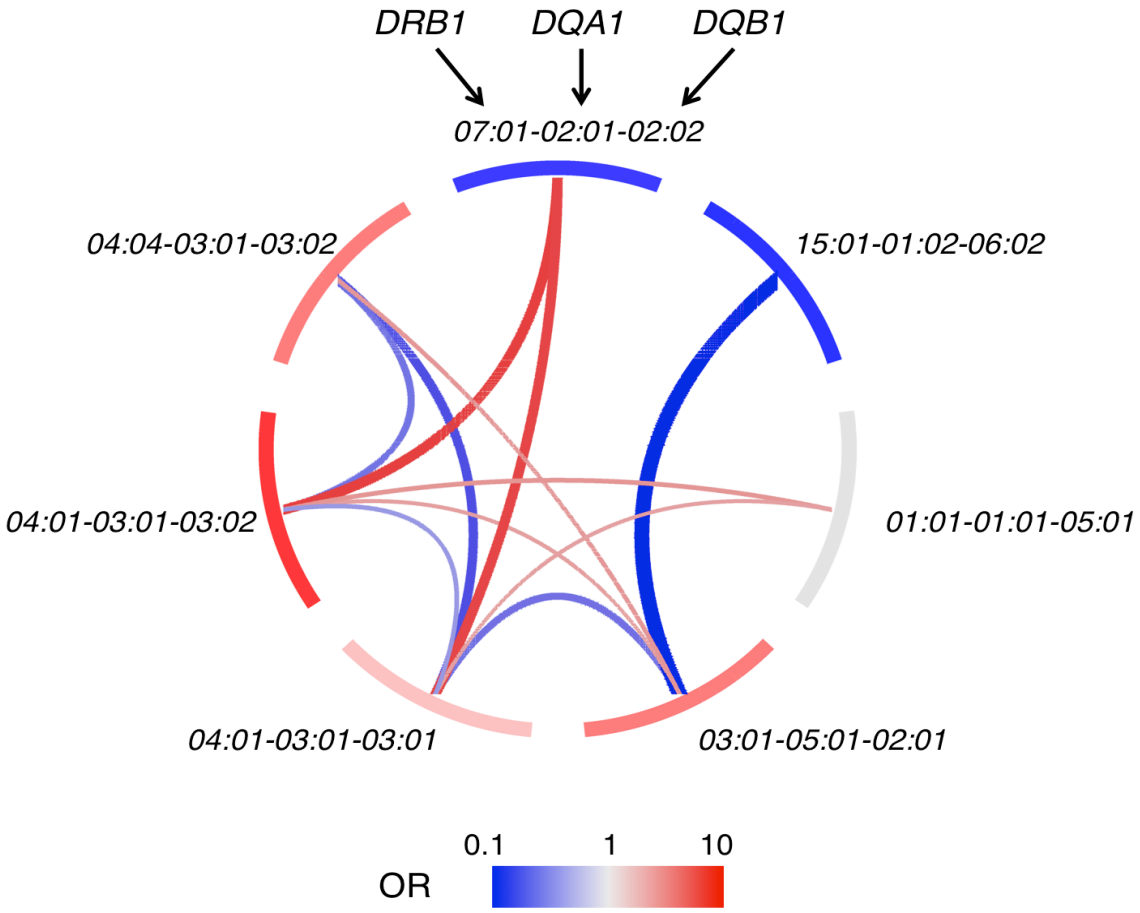


Table 5.2. Pairwise haplotypic interactions in *HLA-DRB1-DQA1-DQB1*. The table shows, for each given pair of haplotypes, the fold change in odds ratio (from additive effect-only) due to interaction. The "amino acids" column/row denote the residues at DQβ1#57, DRβ1#13, and DRβ1#71 corresponding to each haplotype. For each pair, the p-value of the interaction term is shown in parenthesis. Cells in **bold** indicate interactions that are significant after Bonferroni correction ($p < 0.05/21=0.0024$). Cells with underlines indicate the known *DR3/DR4* heterozygote effect. The odds ratio of a given diploid genotype is calculated as $\text{additive}_{\text{haplotype1}} \times \text{additive}_{\text{haplotype2}} \times \text{interaction}_{1,2}$.

Table 5.2. Significant pairwise haplotypic interactions in *HLA-DRB1-DQA1-DQB1* (Continued).

		DRB1 DQA1 DQB1	15:01 01:02 06:02	07:01 02:01 02:02	04:04 03:01 03:02	04:01 03:01 03:02	04:01 03:01 03:01	03:01 05:01 02:01	01:01 01:01 05:01
		Amino acids	D-R-A	A-Y-R	A-H-R	A-H-K	D-H-K	A-S-K	V-F-R
DRB1 DQA1 DQB1	Amino acids	Additive OR	0.16	0.19	2.77	5.49	1.40	2.83	1.00 (Ref)
01:01 01:01 05:01	V-F-R	1.00 (Ref)	0.14 (0.004)	2.32 (0.04)	0.71 (0.19)	**2.16** **(1.2×10^{-4})**	**1.95** **(7.7×10^{-4})**	1.04 (0.77)	
03:01 05:01 02:01	A-S-K	2.83	**0.09** **(1.2×10^{-5})**	2.24 (0.03)	**2.12** **(1.9×10^{-4})**	**1.96** **(1.2×10^{-5})**	**0.32** **(9.2×10^{-11})**		
04:01 03:01 03:01	D-H-K	1.40	0.26 (0.03)	**4.78** **(1.1×10^{-4})**	**0.23** **(2.4×10^{-6})**	**0.48** **(1.1×10^{-3})**			
04:01 03:01 03:02	A-H-K	5.49	0.36 (0.03)	**5.09** **(4.2×10^{-5})**	**0.33** **(3.5×10^{-5})**				
04:04 03:01 03:02	A-H-R	2.77	0.62 (0.03)	2.16 (0.08)					
07:01 02:01 02:02	A-Y-R	0.19	0.76 (0.74)						
15:01 01:02 06:02	D-R-A	0.16							

*DQA1*05:01-DQB1*02:01*) and a *DR4* haplotype (*DRB1*04:01-DQA1*03:01-DQB1*03:02*) (p=1.2 x 10^{-5}). This interaction resulted in an odds ratio of 30.42, compared to an expected odds ratio of 15.51 due to only additive contributions. Likewise, there was an independent interaction between *HLA-DR3* and another DR4 haplotype, *DRB1*04:04-DQA1*03:01-DQB1*03:02* (p=1.9x10^{-4}).

We observed many other significant haplotypic interactions beyond the well-studied *DR3/DR4* effect (**Table 5.2**). Most interactions increased T1D risk. For example, the combination of *DRB1*04:01-DQA1*03:01*DQB1*03:02* and *DRB1*07:01-DQA1*02:01-DQB1*02:02* dramatically increased risk by 5.09-fold (beyond the risk predicted by the additive model). Other pairs significantly reduced risk. Notably, while *DRB1*04:01-DQA1*03:01-DQB1*03:02* and *DRB1*04:04-DQA1*03:01-DQB1*03:02* each conferred risk, the heterozygote combination elicited a 3-fold reduction from the expected risk. Since we restricted our analysis to haplotypes with at least 5% allele frequency, other interactive effects are likely present but unobserved[19].

Interaction effects are mediated by DQβ1#57 and DRβ1#13

The HLA-DQ α/β *trans* heterodimer formed by proteins encoded by *DQA1*05:01* and *DQB1*03:02* may confer a particularly high risk for individuals with the *DR3/DR4* genotype due to its unique antigen binding properties[29]. In order to identify the possible drivers of this haplotypic interaction, we tested pairwise interactions among the *HLA-DRB1, -DQA1,* and *-DQB1* 4-digit alleles. We observed a significant interaction *between DQA1*05:01 and DQB1*03:02* (p=1.71x10^{-25}). However, due to a high degree of LD across the locus, several pairs of classical alleles (including *DQB1*02:01/DQB1*03:02 and DRB1*03:01/DQB1*03:02*, **Table 5.3**) achieved similarly significant p-values. Therefore, while our model was consistent with a risk-conferring interaction between *DQA1*05:01* and *DQB1*03:02*, we could not eliminate the possibility that interactions between other alleles within the two haplotypes are driving this specific interaction.

Table 5.3. DR3/DR4 pairwise allelic interactions

Allele 1	Allele 2	P value for interaction
DQB1*02:01	DQB1*03:02	5.57E-26
DRB1*03:01	DQB1*03:02	5.57E-26
DQA1*05:01	DQB1*03:02	1.71E-25
DQB1*02:01	DRB1*04:04	1.49E-22
DRB1*03:01	DRB1*04:04	1.49E-22
DQA1*05:01	DRB1*04:04	1.56E-21
DQB1*02:01	DQA1*03:01	5.74E-06
DRB1*03:01	DQA1*03:01	5.74E-06
DQA1*05:01	DQA1*03:01	1.69E-05
DQA1*05:01	DRB1*04:01	0.1455
DQB1*02:01	DRB1*04:01	0.19037
DRB1*03:01	DRB1*04:01	0.19037

We next assessed whether these haplotypic interactions could be explained by amino acid positions. We exhaustively tested for all pairwise interactions among amino acid residues in *HLA-DRB1-DQA1-DQB1*, again limiting the analysis to residues with at least 5% frequency. Of the 3,773 pairs of amino acid positions tested, we observed that interactions between DQβ1#57 and DRβ1#13 yielded the largest improvement over the additive model, as measured by delta-deviance (**Table 5.4**). We note that two other pairs of amino acid positions achieved similarly significant *p*-values. These analyses suggest that the same amino acid positions that explain the greatest proportion of the additive risk may also be the positions that mediate interactive effects within this locus.

Table 5.4. Top 50 (by deviance) pairwise amino acid interactions

Interaction	ΔDeviance	df	p value
DRB1_13 - DQB1_57	233.844	15	2.60E-41
DRB1_11 - DQB1_57	233.844	15	2.60E-41
DRB1_37 - DQB1_57	225.979	12	1.36E-41
DQA1_47 - DQB1_57	222.150	12	8.50E-41
DQA1_47 - DQB1_185	213.084	8	1.11E-41
DQB1_185 - DQB1_55	203.337	6	3.70E-41
DQA1_56 - DQB1_185	201.572	6	8.78E-41
DQA1_76 - DQB1_185	201.572	6	8.78E-41
DRB1_13 - DQB1_74	197.588	15	6.57E-34
DRB1_13 - DQB1_71	197.588	15	6.57E-34
DRB1_11 - DQB1_74	196.884	15	9.12E-34
DRB1_11 - DQB1_71	196.884	15	9.12E-34
DRB1_13 - DQA1_47	196.528	20	5.46E-31
DRB1_13 - DQB1_55	196.514	15	1.08E-33
DRB1_13 - DQB1_185	195.900	10	1.15E-36
DRB1_11 - DQA1_47	195.761	20	7.74E-31
DRB1_11 - DQB1_55	195.742	15	1.56E-33
DQB1_185 - DQB1_74	195.612	6	1.63E-39
DQB1_185 - DQB1_71	195.612	6	1.63E-39
DRB1_11 - DQB1_185	194.510	10	2.25E-36
DRB1_96 - DQA1_47	194.490	16	1.03E-32
DQA1_47 - DQB1_26	193.539	12	7.01E-35

Table 5.4. Top 50 (by deviance) pairwise amino acid interactions (Continued).

Interaction	ΔDeviance	df	p value
DRB1_13 - DQB1_30	193.359	15	4.73E-33
DRB1_13 - DQB1_26	193.135	15	5.25E-33
DRB1_11 - DQB1_26	193.135	15	5.25E-33
DRB1_37 - DQB1_185	192.721	8	2.18E-37
DRB1_11 - DQB1_30	192.630	15	6.65E-33
DQB1_185 - DQB1_30	191.826	6	1.04E-38
DQA1_47 - DQB1_74	191.305	12	2.02E-34
DQA1_47 - DQB1_71	191.305	12	2.02E-34
DRB1_37 - DQB1_26	191.202	12	2.12E-34
DRB1_96 - DQB1_74	190.159	12	3.48E-34
DRB1_96 - DQB1_71	190.159	12	3.48E-34
DRB1_96 - DQB1_55	190.032	12	3.70E-34
DRB1_37 - DQB1_74	189.266	12	5.32E-34
DRB1_37 - DQB1_71	189.266	12	5.32E-34
DRB1_140 - DQB1_57	188.684	6	4.85E-38
DRB1_9 - DQB1_57	188.546	6	5.19E-38
DQA1_47 - DQB1_30	188.144	12	9.05E-34
DQA1_47 - DQB1_167	187.903	8	2.25E-36
DQA1_47 - DQB1_13	187.380	8	2.90E-36
DQA1_52 - DQB1_57	187.101	9	1.66E-35
DRB1_13 - DQA1_56	186.834	15	9.91E-32
DRB1_13 - DQA1_76	186.834	15	9.91E-32

Table 5.4. Top 50 (by deviance) pairwise amino acid interactions (Continued).

Interaction	ΔDeviance	df	p value
DRB1_11 - DQA1_56	186.078	15	1.41E-31
DRB1_11 - DQA1_76	186.078	15	1.41E-31
DQA1_47 - DQB1_55	185.506	12	3.16E-33
DRB1_96 - DQB1_30	185.454	12	3.24E-33
DRB1_37 - DQB1_55	185.248	12	3.57E-33
DRB1_37 - DQB1_30	184.513	12	5.05E-33

DISCUSSION

Our fine-mapping of the MHC locus in T1D demonstrates that amino acid polymorphisms at DQβ1#57, DRβ1#13, and DRβ1#71 independently modulates T1D risk, and capture over 90% of the phenotypic variance explained by the *HLA-DRB1-DQA1-DQB1* locus (and 80% of the variance explained by all of the HLA). Previous studies have suggested that other amino acid positions within the HLA class II molecules confer T1D risk; for example, DRβ1#86, DRβ1#74, and DRβ1#57 in the P1, P4, and P9 pockets, respectively[16]. While our analysis highlights these three amino acid positions as the main contributors of T1D risk, there is also evidence of other allelic effects within *the HLA-DRB1-DQA1-DQB1* locus; however their relative effect sizes were very modest compared to the three leading positions identified. We note that our results are derived from cases and controls from a relatively homogeneous population (the United Kingdom), and our ability to interrogate rarer alleles in this population may be limited. For instance*, HLA-DRB1*04:03*, a common allele in the Sardinian population and highlighted by Cucca *et al.* as a protective allele[16], is rare in this dataset (allele frequency ~0.3%). As such, the observed effect of amino acid positions which best define this allele (DRβ1#74 and #86) may have been less pronounced than what might be observed in a more diverse dataset. Additional variants might be conclusively identified in the future with increased sample size. Finally, although coding variants contribute to the majority of the phenotypic variance in T1D, there is the possibility that there are other mechanisms, such as protein expression and structural stability, that modulate susceptibility[30, 31].

We find nine interactions between pairs of HLA haplotypes that contribute to T1D risk, in addition to the previously described *HLA-DR3/DR4* interactions, suggesting that non-additive effects are common within this locus. Notably, we showed that amino acid positions in DRβ1 and DQβ1 were the strongest contributors to both additive and interactive risk effects. Interestingly, the two strongest interacting amino acid positions were in separate HLA molecules (DQ and DR,

respectively). *HLA-DQA1*, which is in strong LD to *HLA-DRB1* and *HLA-DQB1*, appears to play a minimal role in modulating T1D risk. This suggests that the interactive effects are possibly due to the alteration in antigen-presentation repertoire created by the combination of different HLA molecules, rather than the consequence of specific DQα/β heterodimers with particular structural features that confer extreme binding affinities.

The HLA amino acid variants identified in our study may mediate recognition of one or more autoantigens and cause autoimmunity through different mechanisms. In particular, our findings implicate the HLA-DR P4 pocket in T1D in addition to the known role of the HLA-DQ P9 pocket; this is the first instance to our knowledge where the DR P4 pocket plays an important but secondary role to a different locus (DQβ1#57). The DR P4 pocket has been shown to play primary roles in other autoimmune diseases. For example, in RA, the risk-conferring amino acid residues in P4 likely facilitate the binding of citrullinated peptides[7]. In T1D, the anti-islet autoantibody reactivity in patients' sera is largely accounted for by four autoantigens: preproinsulin, glutamate decarboxylase (GAD), islet antigen 2 (IA-2), and ZnT8; although the identification of specific peptides that affect autoreactivity is still work in progress[8, 32-37]. Cucca *et al.* implicated signal peptide sequences of preproinsulin as potentially important in T1D, by modeling the associations of HLA class II alleles and their polymorphic amino acid positions with structural features of the peptide-binding pockets l[16]. The ability to focus on crucial variants that drive risk may enable functional investigations. Synthesis of HLA molecules containing single-residue alterations at risk-modulating positions may reveal their effects on the physical-chemical properties of the antigen-binding pockets. Furthermore, the use of peptide display or small molecule libraries may directly identify and characterize peptides that differentially bind to HLA molecules that differ at risk-modulating positions, thereby revealing the essential pathogenic peptides and the mechanisms through which they evoke autoimmunity.

METHODS

Sample collection

The dataset was provided by the Type 1 Diabetes Genetics Consortium[20], and consisted of (1) a UK case-control dataset (UK) and (2) a European family based dataset (Eur). The UK case-control dataset consisted of a total of 16,086 samples (6,670 cases and 9,416 controls) from 3 collections: (1) cases from the UK-GRID, (2) shared controls from the British 1958 Birth Cohort and (2) shared controls from Blood Services controls (data release February 4, 2012, hg18). The UK samples were collected from 13 regions. The European Family based dataset consisted of 10,791 samples (5,571 affected children and 5,220 controls) from 2,699 European-ancestry families (data release January 30, 2013, hg18). All samples were genotyped on the ImmunoChip array. After quality control, 6,223 and 6,608 markers, respectively, were genotyped in the MHC region between 29 and 45Mb on Chromosome 6 in the two datasets. From the family data we constructed 1,662 pairs of pseudocase and pseudocontrol samples.

Construction of pseudo-case/control samples

We constructed pseudo-controls from a set of 1,661 European families with at least 1 affected child and both healthy parents present. From each family, we selected one affected child (randomly selecting one if multiple affected children were present) to be the case (transmitted alleles).

We first determined the parent of origin for each of the child's two chromosomes, using heterozygous "checkpoints". Checkpoints consist of markers for which the child and only one parent are heterozygous. For example, if the mother's genotype at marker X is "AA", the father's is "AB", and the child's is "BA", we determine that the first allele at each marker came from the father. At each marker, the pseudo-case genotype is that of the affected child. Pseudo-control genotype consists of the two untransmitted allele from the parents, ordered by the parent of origin

determined above. For example, if at a given marker, the father is "AA", the mother is "AB", the affected child is "AA", and the allele order in the child is determined to be mother/father, then the pseudo-control genotype must be "BA".

After applying HLA imputation, we also used the checkpoint results to identify phasing error. We observed that in about 5% of the samples, there was jumping (such that the class I segment of a chromosome in the child is from one parent, while the class II segment is from the other); no significant phasing error occurred within genes, within the class I region or within the class II region.

HLA Imputation

We used SNP2HLA (default input parameters) to impute SNPs, amino acid residues, indels, and 2- and 4-digit classical alleles in eight HLA genes in the MHC between 29602876 and 33268403bp on Chromosome 6. We used the reference panel provided by T1DGC, which included 5,225 European samples classical typed for *HLA-A, B, C, DRB1, DQA1, DQB1, DPB1*, and *DPA1* 4-digit alleles[21, 22]. The imputed genotype dataset included 8,961 binary markers. For each marker and each individual, two types of output were produced: a phased best-guess genotype (*e.g.* "AA/AT/TT"); and a dosage, which accounts for imputation uncertainty and can be continuous between 0 (0 copies of the alternative allele) and 2 (2 copies of the alternative allele).

We imputed the UK case-control dataset and European family dataset independently; within each set, cases and controls were imputed together to avoid disparity in imputation quality. 4,604 and 5,125 SNPs in the MHC region were used for imputation in the UK and Eur datasets, respectively. After combining the UK and Eur datasets, we excluded a total of 344 binary markers due to allele missingness or rareness (allele frequency < 0.05%); we then removed individuals who carried the missing or rare alleles. The post-quality control final dataset consisted of 18,832 samples, including 8,095 cases (including 1,662 pseudo-cases) and 10,737 controls (including

1,662 pseudo-controls).

Statistical framework

We test a given variant's association to disease status using the logistic regression model:

$$\log(odds_i) = \beta_0 + \sum_{j=1}^{m-1} \beta_{1,j} x_{i,j} + \sum_{k=1}^{n-1} \beta_{2,k} y_{i,k} + \beta_3 z_i \quad \text{(Equation 1)}$$

where variant x_i may be the imputed dosage or the best-guess genotype for a SNP, classical allele, amino acid, or haplotype. β_0 is the logistic regression intercept and $\beta_{1,j}$ is the additive effects of allele j of variant x_i. The number of alleles at each variant is m; for a binary variant (presence or absence of x_i), m equals 2. The covariate $y_{i,k}$ denotes each region of sample collection (n=14). We included sex as covariate z. β_2 and β_3 are the effect sizes of the region and gender covariates, respectively.

To account for population stratification, we included the region codes as covariates. Samples from the Eur dataset were considered as the 14th region. To assess the statistical significance of a tested variant, we assessed the improvement of fit of a model over the null model (only region and gender covariates) when the test variant is added to the model. We calculate as the deviance defined by

$\Delta\text{deviance}_{alt\text{-}null} = -2\ln(\text{likelihood}_{alt}/\text{likelihood}_{null})$, which follows a χ^2 distribution with m-1 degrees of freedom, from which we calculate the p-value. We considered $p = 5\text{x}10^{-8}$ as the significance threshold.

Analysis of amino acid positions

To test amino acid effects within *HLA-DRB1-DQA1-DQB1*, we applied conditional haplotypic analysis. We tested each single amino acid position by first identifying the m amino acid residues occurring at that position, and then partitioning all samples into m groups with identical residues at

that position. We estimated the effect of each of the *m* groups using logistic regression model (including covariates as above), and assessed the significance of model improvement by Δdeviance compared to the null model, with *m-1* degrees of freedom. This is equivalent to testing a single multi-allelic locus for association with *m* alleles. To test the effect of a second amino acid position, while conditioning on the first, we further update the model to include all unique haplotypes created by residues at both positions. We then test whether the updated model improves upon the previous model based on Δdeviance, taking in consideration for the increased degrees of freedom.

Conditional analysis on entire locus

In order to condition out the effect of a whole gene or locus, we included all 4-digit classical alleles in the gene or locus as covariates in the regression model.

Exhaustive test

To ensure that the independently associated amino acids were not emerging only as the result of forward-search which might possibly converge on local minima, we exhaustively tested of all possible combinations of one, two, three, and four amino acid positions in *HLA-DRB1, DQA1,* and *DQB1*. For each number of amino acid combination, we select the best model based on Δdeviance from the null (gender and region covariates only).

Haplotype-amino acid permutation analysis

Given the polymorphic nature of the HLA genes and the strong effect sizes in the DRB1-DQA1-DQB1 locus, we wanted to assess whether the observed associations at DQβ1#57, DQβ1#13 and DQβ1#71 could emerge by chance, due to these positions' ability to tag classical alleles with different risk. To eliminate this possibility, we conducted a permutation test. In each permutation, for each of the three genes (e.g *HLA-DQB1, -DRB1,* and *-DQA1)* we preserved the sample's

case/control status and gender/region covariates. To preserve allelic associations, we preserved the groups of samples with the amino acid sequence (4-digit classical allele) at each gene. We then randomly reassigned the amino acid sequence corresponding to each classical allele in each permutation, and repeated the forward-search analysis. We repeated this permutation 10,000 times, each time selecting the combination of two, three, and four amino acid positions that produce the best model (as measured by deviance). If the amino acids were merely tagging the effects of certain haplotypes, the effects we observed in the real data would not be more significant compared to those generated from permutations. To obtain the permutation-based *p*-value, we calculate the proportion of permutated models that exceeds the goodness-of-fit of the best model in the unpermuted data.

Testing for non-additivity and interactions

We defined haplotypes across the *HLA-DRB1-DQA1-DQB1* locus based on unique combinations of amino acid residues across the three genes. As non-additive effects can be observed only when sufficient numbers of homozygous individuals are present, we limited the interaction analysis to a subset of common haplotypes or classical alleles with frequencies greater than 5%. We excluded all individuals with one or more haplotypes that fell below this threshold.

We constructed an interaction model, which included additive terms for each common haplotype and interaction terms between all possible pairs of common haplotypes.

$$\log(odds_i) = \beta_0 + \sum_{j=1}^{m-1} \beta_{1,j} x_{i,j} + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} \phi_{j,l} x_{i,j} x_{i,l} + \sum_{k=1}^{n-1} \beta_{2,k} y_{i,k} + \beta_3 z_i \quad \text{(Equation 2)}$$

where ϕ is the interaction effect size. We determined the improvement in fit with each successive model by calculating the change in deviance, and used a significance threshold of $p = 0.05/h$, where h is the total number of interactive parameters added to the original additive model.

HLA-DR3/DR4 classical allele interactions

To characterize the DR3/DR4 interaction, we defined 12 interaction terms, where each term represents a potential interaction between a classical allele on the DR3 haplotype (*DRB1*03:01*, *DQA1*05:01*, *DQB1*02:01*) and a classical allele on the DR4 haplotype (*DRB1*04:01* or *DRB1*04:04*, *DQA1*03:01*, *DQB1*03:02*). We only looked at *trans* interactions, since haplotype analyses already account for classical alleles that occur together in *cis*. We began with a null model that included additive effects for all haplotypes. Then, we individually tested each of the 12 interaction terms by adding each term to the null model separately. Once again, we used the change in deviance to assess the improvement in fit, using $p=0.05/21=2.4\text{x}10^{-3}$ as the threshold.

Amino acid interaction analysis

To determine whether amino acid positions can explain haplotypic interactions, we defined haplotypes across the *DRB1-DQA1-DQB1* locus based on the 141 amino acid positions imputed in this locus. To ensure that a significant number of homozygous individuals were present, we excluded all amino acid residues with less than 5% frequency prior to creating the haplotypes. We also excluded any individual who had one or more amino acids that fell below this threshold.

We began with a null model that included additive effects for each amino acid haplotype. Then, for each pair of amino acid positions [38], we added a set of $n_q \times n_r$ interaction terms, where each term specifies a *trans* interaction between one variant at each position, and n_p represents the total number of variants at position p. Each pair of amino acids was tested in a separate model, and we calculated the change in deviance to determine the improvement in fit. Monomorphic amino acid positions were excluded from this analysis, since they were constant across all individuals.

Analysis of amino acid positions considering non-additive effect

We wanted to show that the independently associated risk-modulating amino acid positions remained unchanged after including non-additive effects. To this end, we repeated the forward-search analysis after incorporating non-additive terms into the regression model. In this analysis, each variant is coded as 0/1/2 for allelic dosage; and an additional heterozygote factor is added, which equals 1 only if the individual is heterozygotic for this allele/haplotype.

Testing for discordance effect sizes between T1D and Rheumatoid Arthritis

DRB1#13 and #71 show strong independent effects in both T1D and RA. We tested whether the individual residues at each position confer differential risk or protection between the two diseases, using previously described method [3]. Given a multi-allelic amino acid position with *m* residues, we calculated the multivariate log-odds ratios (log-OR) of the residues by including in the logistic regression the binary markers corresponding to each residue [3]. We excluded the most common residue in controls as the reference (therefore the log-OR and variance for that residue are both 0). Let $a_1, a_2, \ldots a_{m-1}$ and $b_1, b_2, \ldots b_{m-1}$ be the multivariate log-ORs in the two diseases; and let $v_1, v_2, \ldots v_{m-1}$ and $u_1, u_2, \ldots u_{m-1}$ be the variances around the log-OR estimates. To test the discordance of effect sizes between the two diseases, we calculated the statistic

$$\sum_{i=1\ldots m-1} \frac{(a_i - b_i)^2}{v_i + u_i}$$

This is χ^2 distributed with $m-1$ degrees of freedom under the null.

We note that the RA and T1D datasets likely shared a proportion of control samples (from the British 1958 Birth Cohort and Blood Service). However, we do not expect inflated discordance, as had any bias been introduced by shared controls, it would tend toward artificial concordance, rather than discordance.

Proportion of phenotypic variance explained

We assumed the liability threshold model, and calculated the proportion of phenotypic variance explained (h^2) by a combination of variants using previously described methods [4,5].

We used a model based on the biometrical model from Fisher [6] and the liability threshold model from Pearson and Lee [7]. We assumed that disease risk is the consequence of an underlying liability score that is normally distributed with a mean of zero and a variance of one, and that individuals with a score above a pre-specified threshold get disease [8]. The value of h^2 is defined by the variance between genotypic groups (V_g) divided by the total population variance (V_t), which equals the sum of V_g and variance within group (V_i); or,

$$h^2 = V_g/(V_g + V_i)$$

V_g, the between-group variance, is defined as

$$\sum p_i(\overline{x_i} - \bar{\bar{x}})^2$$

Where p_i is the population frequency of a genotypic group i; $\bar{x}_i$ is the mean of the group; and $\bar{\bar{x}}$ is the grand mean of the population.

To estimate the h^2 explained by *HLA-DRB1-DQA1-DQB1*, we estimated the multivariate ORs of 67 haplotypes (each occurring at least four times in the dataset) defined by all amino acids in the locus, using the most common haplotype in controls as the reference. Similarly, to calculate h^2 explained by DQβ1#57, DRβ1#13, DRβ1#71, we calculated the control frequencies and multivariate ORs of 29 haplotypes (with at least four copies) defined by these three positions.

We next assumed that the alleles are in Hardy-Weinberg equilibrium, and calculated the genotype frequency (p_i) and prevalence of disease within each possible diploid genotype (f_i). Given that the disease is rare, relative risk approximately equals the odds ratio, therefore

$$f_i = \frac{RR_i \times F}{\sum p_i \times RR_i}$$

We assumed V_i within each genotypic groups to be 1. We then determined the liability threshold within each genotype (T_i) using the normal inverse cumulative distribution function.

Next, we could assume that the shift of liability threshold of the reference genotype group, $T_{ref,}$ from the population threshold is 0. Assuming equal within-genotype variance of 1, the shift in liability threshold equals the difference in the genotypic means; that is, $\bar{x}_i = T_i - T_{\text{ref.}}$ The grand mean of the population can then be updated as

$\bar{\bar{x}} = \sum p_i \times \bar{x}_i$.

INFO score calculation for *DRB1-DQA1-DQB1* haplotypes

To assess the imputation quality of a given DRB1-DQA1-DQB1 haplotype, we calculated the INFO score from the ratio of the observed variance in dosage to the expected variance under Hardy-Weinberg equilibrium [9]:

$$INFO = \frac{\text{var}(x)}{2(p)(1-p)}$$

where x is the imputed dosage and p is the frequency of the allele. An INFO score close to 0 indicates poor imputation quality, while a score closer to 1 indicates higher quality; a value greater than 1 is also possible. Due to the presence of non-additive effects that inflated the disease risk in heterozygotes, the allele distribution in disease cases deviated from Hardy-Weinberg equilibrium. Therefore, we calculated INFO scores using the variance and allele frequency in controls only.

Acknowledgement

This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. This work is supported in part by funding from the National

Institutes of Health (5R01AR062886-02 (PIdB), 1R01AR063759 (SR), 5U01GM092691-05 (SR), R01AR065183 (PIWdB)), a Doris Duke Clinical Scientist Development Award (SR), the Wellcome Trust (JAT) and the National Institute for Health Research (JAT and JMMH), and a Vernieuwingsimpuls VIDI Award (016.126.354) from the Netherlands Organization for Scientific Research (PIWdB). TLL was supported by the German Research Foundation (LE 2593/1-1 and LE 2593/2-1).

References

1. Maahs, D.M., et al., *Epidemiology of type 1 diabetes.* Endocrinol Metab Clin North Am, 2010. **39**(3): p. 481-97.

2. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis.* Nat Genet, 2012. **44**(3): p. 291-6.

3. Trynka, G., C. Wijmenga, and D.A. van Heel, *A genetic perspective on coeliac disease.* Trends Mol Med, 2010. **16**(11): p. 537-50.

4. Gourraud, P.A., et al., *The genetics of multiple sclerosis: an up-to-date review.* Immunol Rev, 2012. **248**(1): p. 87-103.

5. Lee, K.H., K.W. Wucherpfennig, and D.C. Wiley, *Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes.* Nat Immunol, 2001. **2**(6): p. 501-7.

6. Astill, T.P., et al., *Promiscuous binding of proinsulin peptides to Type 1 diabetes-permissive and -protective HLA class II molecules.* Diabetologia, 2003. **46**(4): p. 496-503.

7. Scally, S.W., et al., *A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis.* J Exp Med, 2013. **210**(12): p. 2569-82.

8. van Lummel, M., et al., *Posttranslational modification of HLA-DQ binding islet autoantigens in type 1 diabetes.* Diabetes, 2014. **63**(1): p. 237-47.

9. Erlich, H., et al., *HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families.* Diabetes, 2008. **57**(4): p. 1084-92.

10. Noble, J.A., et al., *The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families.* Am J Hum Genet, 1996. **59**(5): p. 1134-48.

11. Nejentsev, S., et al., *Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A.* Nature, 2007. **450**(7171): p. 887-92.

12. Cucca, F., et al., *The HLA-DPB1--associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1.* Diabetes, 2001. **50**(5): p. 1200-5.

13. Deschamps, I., et al., *HLA genotype studies in juvenile insulin-dependent diabetes.* Diabetologia, 1980. **19**(3): p. 189-93.

14. Howson, J.M., et al., *Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A.* Diabetes Obes Metab, 2009. **11 Suppl 1**: p. 31-45.

15. Todd, J.A., J.I. Bell, and H.O. McDevitt, *HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus.* Nature, 1987. **329**(6140): p. 599-604.

16. Cucca, F., et al., *A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins.* Hum Mol Genet, 2001. **10**(19): p. 2025-37.

17. Thomson, G., et al., *Genetic heterogeneity, modes of inheritance, and risk estimates for a joint study of Caucasians with insulin-dependent diabetes mellitus.* Am J Hum Genet, 1988. **43**(6): p. 799-816.

18. Svejgaard, A. and L.P. Ryder, *HLA genotype distribution and genetic models of insulin-dependent diabetes mellitus.* Ann Hum Genet, 1981. **45**(Pt 3): p. 293-8.

19. Koeleman, B.P., et al., *Genotype effects and epistasis in type 1 diabetes and HLA-DQ trans dimer associations with disease.* Genes Immun, 2004. **5**(5): p. 381-8.

20. Onengut-Gumuscu, S., et al., *Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers.* Nat Genet, 2015.

21. Jia, X., et al., *Imputing amino acid polymorphisms in human leukocyte antigens.* PLoS One, 2013. **8**(6): p. e64683.

22. Brown, W.M., et al., *Overview of the MHC fine mapping data.* Diabetes Obes Metab, 2009. **11 Suppl 1**: p. 2-7.

23. Noble, J.A. and H.A. Erlich, *Genetics of type 1 diabetes.* Cold Spring Harb Perspect Med, 2012. **2**(1): p. a007732.

24. Han, B., et al., *Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity.* Am J Hum Genet, 2014. **94**(4): p. 522-32.

25. Foo, J.N., et al., *Coding variants at hexa-allelic amino acid 13 of HLA-DRB1 explain independent SNP associations with follicular lymphoma risk.* Am J Hum Genet, 2013. **93**(1): p. 167-72.

26. Witte, J.S., P.M. Visscher, and N.R. Wray, *The contribution of genetic variants to disease depends on the ruler.* Nat Rev Genet, 2014. **15**(11): p. 765-76.

27. Sivertsen, B., et al., *Mental health in adolescents with Type 1 diabetes: results from a large population-based study.* BMC Endocr Disord, 2014. **14**: p. 83.

28. Speed, D., et al., *Improved heritability estimation from genome-wide SNPs.* Am J Hum Genet, 2012. **91**(6): p. 1011-21.

29. Reichstetter, S., W.W. Kwok, and G.T. Nepom, *Impaired binding of a DQ2 and DQ8-binding HSV VP16 peptide to a DQA1*0501/DQB1*0302 trans class II heterodimer.* Tissue Antigens, 1999. **53**(1): p. 101-5.

30. Ettinger, R.A., et al., *Exceptional stability of the HLA-DQA1*0102/DQB1*0602 alpha beta protein dimer, the class II MHC molecule associated with protection from insulin-dependent diabetes mellitus.* J Immunol, 1998. **161**(11): p. 6439-45.

31. Miyadera, H., et al., *Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA.* J Clin Invest, 2014.

32. Knight, R.R., et al., *A distinct immunogenic region of glutamic acid decarboxylase 65 is naturally processed and presented by human islet cells to cytotoxic CD8 T cells.* Clin Exp Immunol, 2015. **179**(1): p. 100-7.

33. Kronenberg, D., et al., *Circulating preproinsulin signal peptide-specific CD8 T cells restricted by the susceptibility molecule HLA-A24 are expanded at onset of type 1 diabetes and kill beta-cells.* Diabetes, 2012. **61**(7): p. 1752-9.

34. Nakayama, M., et al., *Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice.* Nature, 2005. **435**(7039): p. 220-3.

35. Baekkeskov, S., et al., *Identification of the 64K autoantigen in insulin-dependent diabetes as the GABA-synthesizing enzyme glutamic acid decarboxylase.* Nature, 1990. **347**(6289): p. 151-6.

36. Karlsen, A.E., et al., *Recombinant glutamic acid decarboxylase (representing the single isoform expressed in human islets) detects IDDM-associated 64,000-M(r) autoantibodies.* Diabetes, 1992. **41**(10): p. 1355-9.

37. Roep, B.O. and M. Peakman, *Antigen targets of type 1 diabetes autoimmunity.* Cold Spring Harb Perspect Med, 2012. **2**(4): p. a007781.

38. Aghaeepour, N., et al., *Critical assessment of automated flow cytometry data analysis techniques.* Nat Methods, 2013.

CHAPTER 6

Conclusion and discussion

SUMMARY

The understanding of genetic risk of human diseases and traits have advanced rapidly and expanded beyond the Mendelian models of rare, fully penetrant traits. Through a decade of genome-wide association studies, we have discovered hundreds of genomic loci associated with disease risk, and come to appreciate the polygenic architecture of common diseases. Autoimmune diseases are among the most successful disease groups that gained tremendous insight through GWAS; nearly 200 genomic loci have been found to be associated with dozens of these diseases, including rheumatoid arthritis (RA), systemic lupus erythematosus, inflammatory bowel diseases, type 1 diabetes, celiac disease, multiple sclerosis, and many more.

GWAS studies identify regions of the genome that harbor variants that influence disease risk, but cannot directly reveal the causal genetic changes or illustrate their pathogenic mechanisms. Therefore, ongoing efforts in the field must address at least two challenges: 1) localizing the association signals to pinpoint causal variants, and 2) functional interpretation of known variants' roles in pathogenesis. The work presented in this dissertation devised and applied both "top-down" and "bottom-up" approaches (see Chapter 1) to address both challenges. To do so, we considered two "categories" of risk loci – namely, those in the human leukocyte antigen (HLA) genes, and the non-HLA loci – separately.

The method presented in Chapter 2 was an example of the "top-down" approach to follow-up on genome-wide association studies, by integrating genetic data with other biological data types to deduct relevant biological pathways and other global properties implicated by the risk loci. In this study, we hypothesized that genes perturbed by risk variants of a given disease function specifically in tissues and cell types that are crucial in the pathogenic process; and that to identify the cell types, the cell-specific gene expression may serve as proxies to function. By examining compendia of cell-specific gene expression profiles, we showed that genes in autoimmune disease risk loci were indeed specifically expressed in relevant immunological cells. In particular, we found

$CD4^+$ effector memory T (T_{EM}) cells to be relevant in the pathogenesis of rheumatoid arthritis, celiac disease, and type 1 diabetes.

We next hypothesized that immunological phenotypes and functions vary among individuals under genetic influence. We profiled population variation in $CD4^+$ T_{EM} cells' peripheral abundance, proliferation, and expression of immune genes. In this study, we used multi-parametric flow cytometry to study blood samples of over 200 individuals using, producing over 1000 cytometric data samples. To analyze these samples efficiently, we developed X-Cyt, a user-guided automated data-partitioning tool based on parametric clustering. In Chapter 3, we demonstrated that X-Cyt rapidly and robustly analyzes large-scale profiling data samples. In contrast to previously developed automated analytical methods, X-Cyt allows the user to define a template according to which all samples are partitioned. The incorporation of user guidance ensures intuitive and interpretable partitioning outcome; in addition, the use of a template both allows simultaneous alignment across samples and dramatically decreases computational time.

In Chapter 4, we showed variation in immunological traits of $CD4^+$ effector memory T cells, including cell abundance, proliferative response to stimulation, and the expression of immune genes in disease risk loci. Variation in these traits in the population correlated with single nucleotide polymorphisms. Specifically, a quarter of the assayed immune genes were under expression regulation by nearby single nucleotide changes (expression quantitative trait loci, or "eQTL"). In particular, nearly a quarter of the eQTLs we observed were not previously detected in peripheral blood, which consists of heterogeneous cell types. We noted, however, that only a small percentage (<5%) of SNPs associated with RA, CeD, or T1D, were eQTLs of nearby genes. This study was an example of a "bottom-up" method to follow up on genetic data; in this case, rather than studying one specific locus, we used high-throughput technologies to examine many candidate regions.

Finally, the HLA genes have long been known to confer strong risk toward type 1 diabetes; however, the highly polymorphic nature and complex linkage structure in the HLA region challenge efforts to pinpoint causal variants. In Chapter 5, we used statistical imputation and fine-mapped the HLA genes, and identified individual amino acid sites in HLA molecules that drive risk toward type 1 diabetes. In addition to confirming the primary role of amino acid position 57 of the HLA-DQ β chain, we found that positions 13 and 71 of the HLA-DR β chain conferred strong independent risk. Polymorphisms at these three positions together explained 80% of the genetic risk harbored in the HLA region. Furthermore, we discovered multiple genotypes that exhibited non-additive risk effects through pairwise interactions.

DISCUSSION

The risk of common diseases follows a polygenic architecture, in that it is the result of multiple variants; they are therefore also referred to as complex-trait diseases. While linkage studies and genome-wide association studies facilely identify regions of the genome that harbor risk variants, fine-mapping and mechanistic follow-up are much more complicated. Linkage disequilibrium is one complicating factor, especially in the HLA, as the causal variant may be obscured while many nearby variants may appear statistically equivalent. Fine-mapping may benefit partially from dense-genotyping, or ultimately sequencing, large numbers of samples. For example, we fine-mapped the HLA signal in type 1 diabetes by increasing the sample size and genotype density through statistical imputation.

Complex-trait diseases and their causal variants contrast sharply with that of Mendelian diseases in several major ways. Mendelian diseases are often the results of single mutations that directly affect the synthesis, structure, or function of key proteins. In contrast, the causal variants of complex-trait diseases often localize to the vast parts of the genome that do not code for proteins, and currently have no annotated functional role. Unlike the fully penetrant Mendelian mutations,

each single complex-trait variant may carry a small effect size, and is neither necessary nor sufficient to cause diseases. To further complicate efforts to model and study these variants, they may exhibit non-additive effects (the risk of diseases is not directly proportional to the allelic dosage), either due to dominance or interactions among variants. Finally, environmental factors may significantly modify the effects of genetic variants.

Autoimmune diseases are very common in the world population, and have benefited tremendously from GWAS. Although the clinical presentations of autoimmune diseases are diverse and affect multiple organ systems, laboratory studies and clinical data clearly indicate the involvement of the immune system. The biology of immune system is as complex and intricate, if not more so, as that of the genome. This dissertation followed a systematic approach of first identifying the most likely main players in a given disease, *e.g.* $CD4^+$ effector memory T cells in rheumatoid arthritis, and then focusing on the relevant cell type(s) to investigate its functional pattern under genetic influence. Of course, this is an oversimplification; nevertheless, in-depth understanding of individual components is a relatively convenient, targeted, well-controlled, and arguably necessary initial step to understanding a complex system.

Many autoimmune diseases share a distinct feature – which is a blessing and a curse – that a large proportion of genetic risk is attributed to coding variants in the HLA genes. HLA molecules present antigenic peptides to effector cells of the immune system; the risk variants likely alter the repertoire of antigens that can be recognized and presented. For decades, however, statistical and functional fine-mapping have been stymied by the extremely polymorphic nature and linkage disequilibrium in the HLA region. Our discovery of amino acid sites that drive type 1 diabetes risk offer clear candidates for functional follow-up. Positions 13 and 71 of the HLA-DR β chain line the P4 pocket of the molecule, which is an amino acid binding pocket also implicated in rheumatoid arthritis and follicular lymphoma.

Many have suggested that the non-coding variants are expression quantitative trait loci (eQTL), and affect disease risk by regulating the expression levels of genes. There are many possible mechanisms for this regulation. For example, a single nucleotide change in the promoter motif near the gene body may change the binding affinity of transcription factors, thus act as a *cis* (local)-eQTL. Similarly, a mutation in a distant enhancer may increase or decrease the transcription of the gene, thus acting as a *trans* (distant)-eQTL. This hypothesis has led to many eQTL studies to search for risk variants that regulate gene expression. However, many genes are differentially expressed in different cell types, at different times, and depend on specific signaling cascades. Therefore, the presence of eQTLs may be difficult to detect in bulk blood samples without appropriate stimulation. To investigate the behavioral pattern of the most relevant cells, in this case, the $CD4^+$ effector memory T cells, we searched for *cis*-eQTLs of immune genes before and after T cell receptor stimulation. By doing so, we detected more *cis*-eQTLs for these genes than previously found in bulk peripheral blood. Strikingly, only about 5% of the risk variants associated to the diseases overlapped with *cis*-eQTL signals. The power to detect gene regulation by risk variants may have been limited by several factors. For example, we did not assay genome-wide transcription or examine *trans*-effects; we used generic T cell stimulation approach (anti-CD3/CD28) rather than antigen-specific stimulations; and furthermore, although the focus on stimulated $CD4^+$ effector memory T cell increased our ability to observe eQTLs in general, we may have missed variants that affect other cell types. Nevertheless, the low level of overlap between the observed eQTLs and disease-associated variants suggests that while regulation of gene expression is a plausible mechanism through which disease risk is modulated, it cannot explain the majority of associations in non-coding regions.

FUTURE DIRECTIONS

Functional follow-up of risk-conferring HLA amino acid sites

HLA variants explain the majority of genetic risk of both rheumatoid arthritis and type 1 diabetes and many other autoimmune diseases. Any amino acid residue change in these antigen-presenting proteins most likely leads to functional changes due to alteration of antigen-binding properties; however, only a few are relevant to each disease. Our discovery of the three relevant amino acid sites to type 1 diabetes, as well as the previous discovery of those relevant to rheumatoid arthritis, open the gate to effective functional validation.

Having knowledge of the specific amino acid residues that confer protection or risk, investigators can conduct targeted antigen-binding assays using purified HLA molecules with known peptide sequences, or recombinant proteins with designed mutations. Peptide libraries from tissue extract, display libraries, and small molecule libraries can all be used to profile the binding signatures of HLA molecules that differ at the relevant amino acid sites, and reveal the necessary or sufficient pathogenic antigens.

Immuno-profiling

One approach to understanding the functional role of non-HLA risk loci is to profile phenotypic and functional variation under genetic influence. Future studies may extend the studies presented in Chapter 4, by expanding upon several important parameters. First, as technologies become more affordable, whole genome microarray or RNA-sequencing can be applied to assay all genes (and isoforms), thus allowing the detection of risk variants that act as *trans*-eQTLs. Similarly, recently developed mass cytometry simultaneously interrogates dozens of proteins in single cells, and can provide an enormous library of functional profiles. Epigenetic changes are also likely to differ among individuals under genetic influence, thus contributing to variation in immune response and autoimmune risk. Therefore, future studies can incorporate assays that assess DNA methylation, histone modification, as well as allele-specific expression. In addition, any immune response to stimulation is expected to exhibit temporal dynamics that are not yet understood.

Future studies should assay epigenetic changes, transcription, and protein expression, at multiple time-points in order to capture this temporal pattern. Finally, in Chapter 4, we used generic T cell stimulation by anti-CD3/CD28 beads; however, effector immune cells may respond differently according to antigen-specific stimulation. Therefore, considering all these parameters together, future studies should strive to capture disease-causing functional variation under the most relevant conditions, which requires using the disease-specific antigenic stimulation and surveying relevant genes and proteins at the time of the strongest (or the most differential) response.

Automated analysis of large-scale, high dimensional data

Advances in large-scale genomic and proteomic technologies are allowing efficient functional characterization of cells under multiple conditions. For example, mass cytometry by time of flight (CyTOF) is able to assay dozens of surface and intracellular proteins in each cell. Contrastingly, analytical tools to utilize this high-dimensional data are lagging. Many software packages are available to analyze flow cytometry and mass cytometry data. However, there is currently no consensus on how to best extract high-dimensional information from these data. In addition, there is also no "golden standard" with which to benchmark the output of automated methods, other than to compare to that of manual analyses, which is not ideal as manual analyses is low-dimensional by nature. The method to analyze the data may depend on the biological hypothesis and experimental approach specific to a given study; therefore perhaps no single "best method" exists. Future development in the field should first aim to establish standard protocols for quality control and data normalization, so that large-scale, batched, and multi-center collections can produce comparable data.