Title
Computational approaches to cell type and interindividual variation in autoimmune disease

Author
Targ, Sasha

Publication Date
2018

Peer reviewed|Thesis/dissertation

Computational approaches to cell type and interindividual variation in autoimmune disease

by

Sasha Kiang Targ

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

[redacted]

Acknowledgements

The work contained in this thesis would not have been possible without the help of many individuals along the way. First and foremost, I would like to thank my advisor, Jimmie Ye, who gave me the freedom and flexibility to pursue my scientific interests during graduate school, and introduced me to computational genetics and methods development. I would also like to thank the UCSF MSTP for administrative help. Finally, I would like to thank my friends and family for their support and encouragement throughout the many phases of the MD/PhD training program.

The chapter entitled “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation” was published in Nature Biotechnology (PMID: 29227470, doi: 10.1038/nbt.4042).

Computational approaches to cell type and interindividual variation in autoimmune disease

Sasha Kiang Targ

Abstract

Computational approaches offer substantial ability to improve annotation and interpretation of a range of genomic datasets collected with the advent of next generation sequencing technologies, providing an avenue to further understand the impact of changes in genomic data which might contribute to disease. Decoding the genome using deep learning is a promising approach to identify the most important sequence motifs in predicting functional genomic outcomes. In the first part of this work, we develop a search algorithm for deep learning architectures that finds models which succeed at using only RNA expression data to predict gene regulatory structure, learn human-interpretable visualizations of key sequence motifs, and surpass state-of-the-art results on benchmark genomics challenges.

We also develop a computational tool, demuxlet, for droplet-based single-cell RNA-sequencing (dscRNA-seq) that harnesses natural genetic variation to determine the sample identity of each cell and detect droplets containing two cells. These capabilities enable multiplexed dscRNA-seq experiments in which cells from unrelated individuals are pooled and captured at higher throughput than in standard workflows. Using simulated data, we show that 50 SNPs per cell are sufficient to assign 97% of singlets and identify 92% of doublets in pools of up to 64 individuals. Given genotyping data for each of 8 pooled samples, demuxlet correctly recovers the sample identity of >99% of singlets and identifies doublets at rates consistent with previous estimates.

We apply demuxlet to assess cell type-specific changes in gene expression in 8 pooled lupus patient samples treated with IFN- and perform eQTL analysis on 23 pooled samples.

Table of Contents

List of Figures

List of Tables

Chapter 2:

Chapter 1: Introduction

Computational approaches offer substantial ability to improve annotation and interpretation of a range of genomic datasets collected with the advent of next generation sequencing technologies, providing an avenue to further understand the impact of changes in genomic data which might contribute to disease. More than 90% of disease-associated genetic variants discovered by genome-wide association studies (GWAS) fall outside coding regions (introns and intergenic regions), making the assignment of their function challenging (Maurano et al. 2012). With a large number of whole genome sequences being collected for various clinical applications, including in autoimmune diseases, the number of rare (< 0.01 minor allele frequency (MAF)) or private non-coding variants for which impact is not clearly known is also increasing (Lek et al. 2016; Mudge and Harrow 2016; Ashley 2016; Goodwin, McPherson, and McCombie 2016). Thus, there is a significant need for computational tools that use functional genomics data (e.g. RNA-seq measuring gene expression or ATAC/ChIP-seq measuring chromatin state) collected in large cohorts to predict the impact of non-coding sequence variants on gene regulation and disease (Y. I. Li et al. 2016).

Unlike disease-associated variants that fall within coding sequences, where the genetic code and evolutionary conservation enable prediction of functional effects, the interpretation of disease-associated variants in non-coding sequences is more difficult. Previously, these regions have been described as ‘junk DNA’ with unknown function; however, more recently, high-throughput, unbiased characterization of functional genomic markers such as histone modification and enhancer or promoter associated marks from the ENCODE Consortium shows 80% of the

genome demonstrates biochemical activity suggestive of a potential gene regulatory role (ENCODE Project Consortium, 2012). Enrichment of disease variants in candidate regulatory regions of the genome within particular cell types (for example, immune cell subsets) suggests interindividual variation impacting gene regulation programs in specific tissues could play a major role in mediating disease risk (Farh et al. 2015). Thus, annotating the function of non-coding variants is a critical open problem that impacts our ability to understand and treat disease.

One approach to annotate non-coding variants is by associating genetic differences with variability in functional genomic traits. When a variant is simultaneously associated with a molecular trait and disease, a causal relationship can be inferred. In lymphoblastoid cell lines (LCLs), it was found that more than 85% of QTLs were shared from one form of functional genomic data to another. (Y. I. Li et al., 2016). The overlap between GWAS hits and several types of quantitative trait loci, including expression (eQTLs), chromatin accessibility (ATAC-QTLs), histone modification (hmQTLs), and transcription factor binding (bQTLs) suggests that that effects of genetic variation on disease can be mediated through these intermediate phenotypes (Schaub et al. 2012; Banovich et al. 2014; McVicker et al. 2013; Battle et al. 2015).

Previous work that employs functional genomics data to annotate noncoding SNPs includes the Ensembl Variant Effect Predictor (VEP), RegulomeDB, and Functional Identification of SNPs (FunciSNP) (McLaren et al. 2010, Boyle et al. 2012, Coetzee et al. 2012). In RegulomeDB, an average of 56% of variants from whole genome sequences intersected regulatory annotations of eQTLs, dsQTLs, ChIP-exo, TF ChIP-seq, FAIRE, and DNase I hypersensitive site data, allowing insight into the potential function of these SNPs. Another method for annotation of non-coding

variants is the use of conservation data to prioritize SNPs that are most likely to be causative. Examples of work that uses this information to analyze noncoding variation include ANNOVAR, HaploReg, GWAS3D, and fitCons (Wang et al. 2010, Ward et al. 2012, Li et al. 2013, Gulko et al. 2015). These methods have the strength that they make use of substitution rates across organisms to infer functional constraint; however, a limitation to conservation-based methods is that they do not account for more rapid adaptation that may be specific to humans.

Machine learning algorithms that incorporate functional and conservation data for prioritization of functional variants include genome-wide annotation of variants (GWAVA), combined annotation-dependent depletion (CADD), FATHMM-MKL, and deltaSVM (Kircher et al. 2014, Ritchie et al. 2014, Shihab et al. 2015, Lee et al. 2015). Most models used in functional and population genetics to map genetic variants and predict phenotypic outcome have been linear models (Yang et al. 2011; Lee et al. 2015), which can capture marginal effect sizes for the variants that are used as features. These types of models have the advantage that they have closed form solutions, can be learned by convex optimization, and they are less likely to overfit (bias-variance tradeoff); but, they have the downside that the model capacity is limited to linear functions, so do not model any interactions between multiple input variables that could exist in the true function to be learned.

Recently, more complex models based on deep learning methods have been developed by several groups for tasks including transcription factor binding site classification, chromatin accessibility and variant prediction (Alipanahi et al. 2015; Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016; Deming et al. 2016; Angermueller et al. 2016). The DeepBind method uses a neural

network model to distinguish transcription factor binding sites in transcription factor ChIP-seq data from randomly shuffled control sequences, outperforming the previous state of the art FeatureREDUCE and BEEML-PBM (Alipanahi et al. 2015). DeepSEA uses a deep neural network trained on transcription factor binding, DNaseI hypersensitivity sites and histone-mark data to prioritize functional SNPs based on their effect on predict chromatin features (Zhou and Troyanskaya 2015). Bassett similarly uses a neural network trained on cell type specific DNaseI hypersensitivity sites to identify GWAS SNPs that are likely to be causal (Kelley, Snoek, and Rinn 2016). These models have the advantage that they can capture nonlinear behavior among the input features as could be expected for biological systems in which multiple components interact to lead to a given output (van Dijk et al. 2015; Samee, Bruneau, and Pollard 2017). While these models are nonconvex so a global optimum is not guaranteed, deep learning models trained using backpropagation and stochastic gradient descent optimization have been widely successful in reaching low generalization error across a range of domains (Goodfellow, Bengio, and Courville 2016; Schmidhuber 2015). Thus, deep learning models could be a powerful method to make predictions and inference about the effect of genetic variants on the variability in functional outputs between cell types or individuals. In this work, we develop attention-based models for deep learning in transcription factor binding and cell type specific gene expression within immune cells, providing a proof of concept for this approach to interpretation in the field of computational genomics.

The resolution at which functional genomic data is collected offers another direction through which to improve understanding of noncoding variants identified through GWAS (Tanay and Regev, 2017). Droplet-based single cell RNA-sequencing (dscRNA-seq) allows parallel measurement of transcriptomes at the single cell level across thousands of individual cells,

yielding a comprehensive profile of heterogeneity across cells within a sampled population (Kolodziejczyk et al. 2015). Single cell transcriptional data enables unbiased characterization of frequency and identity of cellular subpopulations, providing a more scalable approach to investigating cell type specific contributions to disease than single cell sorting (Shalek et al. 2014). Identification and tracking of shifts in cell states that occur within a population or in response to a stimulation constitutes another promising area of exploration facilitated by single cell RNA-sequencing techniques (Pollen et al. 2014). The basic framework of single cell RNA-sequencing involves capture and lysis of the contents of a single cell, followed by reverse transcription to obtain barcoded cDNA from mRNA transcripts within the cell and amplification of the cDNA for sequencing library preparation (Kolodziejczyk et al. 2015).

The ability to multiplex samples for analysis by single cell RNA-sequencing offers the additional capability to compare samples across individuals or different treatment conditions without batch effects due to separate processing (Gehring et al. 2018). Previous work on sample multiplexing such as combinatorial indexing by split pool synthesis or synthetically introduced barcodes allows for higher throughput and comparison across samples, but requires experimental manipulation of samples that could lead to artifacts in the resulting data (Cao et al. 2017, Dixit et al. 2016, Adamson et al. 2016, Jaitin et al. 2016, Datlinger et al. 2017). We therefore present a computational algorithm that improves on these methods and makes use of naturally occurring genetic variation to assign each cell in a mixed population to the most likely sample of origin based on the read overlapping SNPs within sequencing data and reference genotypes. We demonstrate the ability of this method to estimate cell type proportion, reveal transcriptional responses to stimulation and identify cell type specific eQTLs in autoimmune disease samples.

References

1. Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." Nature Biotechnology 33 (8): 831–38.
2. Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. 2016. "Deep Learning for Computational Biology." Molecular Systems Biology 12 (7): 878.
3. Ashley, Euan A. 2016. "Towards Precision Medicine." Nature Reviews. Genetics 17 (9): 507–22.
4. Banovich, Nicholas E., Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F. Degner, John D. Blischak, Julien Roux, Jonathan K. Pritchard, and Yoav Gilad. 2014. "Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels." PLoS Genetics 10 (9): e1004663.
5. Battle, Alexis, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. 2015. "Genomic Variation. Impact of Regulatory Variation from RNA to Protein." Science 347 (6222): 664–67.
6. Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, Julie Park, Benjamin C. Hitz, Shuai Weng, J. Michael Cherry, and Michael Snyder. "Annotation of functional variation in personal genomes using RegulomeDB". Genome Res. 2012;22:1790–1797.
7. Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic

Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." Nature Methods 10 (12): 1213–18.

8. Cheng, Christine S., Rachel E. Gate, Aviva P. Aiden, Atsede Siba, Marcin Tabaka, Dmytro Lituiev, Ido Machol, et al. 2016. "Genetic Determinants of Chromatin Accessibility and Gene Regulation in T Cell Activation across Human Individuals." bioRxiv. doi:10.1101/090241.
9. Churchill, G. A., and R. W. Doerge. 1994. "Empirical Threshold Values for Quantitative Triat Mapping." Genetics 138 (3): 963–71.
10. Coetzee, Simon G., Rhie, Suhn K., Berman, Benjamin P., Coetzee, Gerhard A.* and Noushmehr, Houtan*. "FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs." Nucleic Acids Res. 2012;40:e139.
11. Dabelea, Dana, Elizabeth J. Mayer-Davis, Sharon Saydah, Giuseppina Imperatore, Barbara Linder, Jasmin Divers, Ronny Bell, et al. 2014. "Prevalence of Type 1 and Type 2 Diabetes among Children and Adolescents from 2001 to 2009." JAMA: The Journal of the American Medical Association 311 (17): 1778–86.
12. Delong, Thomas, Timothy A. Wiles, Rocky L. Baker, Brenda Bradley, Gene Barbour, Richard Reisdorph, Michael Armstrong, et al. 2016. "Pathogenic CD4 T Cells in Type 1 Diabetes Recognize Epitopes Formed by Peptide Fusion." Science 351 (6274): 711–14.
13. Deming, Laura*, Sasha Targ*, Nate Sauder, Diogo Almeida, and Chun Jimmie Ye. 2016. "Genetic Architect: Discovering Genomic Structure with Learned Neural Architectures." arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1605.07156.

14. De Jager, Philip L., Nir Hacohen, Diane Mathis, Aviv Regev, Barbara E. Stranger, and Christophe Benoist. 2015. “ImmVar Project: Insights and Design Considerations for Future Studies of ‘healthy’ Immune Variation.” Seminars in Immunology 27 (1): 51–57.
15. Dijk, D. van, E. Sharon, M. Lotan-Pompan, and A. Weinberger. 2015. “Competition between Binding Sites Determines Gene Expression at Low Transcription Factor Concentrations.” bioRxiv. biorxiv.org. http://biorxiv.org/content/early/2015/12/06/033753.abstract.
16. ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” Nature 489 (7414): 57–74.
17. Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. “Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants.” Nature 518 (7539): 337–43.
18. Feng, Jianxing, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. 2012. “Identifying ChIP-Seq Enrichment Using MACS.” Nature Protocols 7 (9): 1728–40.
19. Gamazon, Eric R., Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, et al. 2015. “A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data.” Nature Genetics 47 (9): 1091–98.
20. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. “Deep Learning. Book in Preparation for MIT Press.” http://www.deeplearningbook.org.
21. Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” Nature Reviews. Genetics 17 (6): 333–51.

22. Gusev, Alexander, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, et al. 2016. “Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies.” Nature Genetics 48 (3): 245–52.
23. Han, Buhm, and Eleazar Eskin. 2011. “Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-Wide Association Studies.” American Journal of Human Genetics 88 (5): 586–98.
24. Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the next Generation of Genome-Wide Association Studies.” PLoS Genetics 5 (6): e1000529.
25. Kelley, David R., Jasper Snoek, and John Rinn. 2016. “Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks.” bioRxiv. doi:10.1101/028399.
26. Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. “Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq.” Nature Genetics 48 (2): 206–13.
27. Kurachi, Makoto, R. Anthony Barnitz, Nir Yosef, Pamela M. Odorizzi, Michael A. DiIorio, Madeleine E. Lemieux, Kathleen Yates, et al. 2014. “The Transcription Factor BATF Operates as an Essential Differentiation Checkpoint in Early Effector CD8+ T Cells.” Nature Immunology 15 (4): 373–83.
28. Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” Genome Biology 10 (3): R25.

29. Lee, Dongwon, David U. Gorkin, Maggie Baker, Benjamin J. Strober, Alessandro L. Asoni, Andrew S. McCallion, and Michael A. Beer. 2015. “A Method to Predict the Impact of Regulatory Variants from DNA Sequence.” Nature Genetics 47 (8): 955–61.
30. Lee, Mark N., Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, et al. 2014. “Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells.” Science 343 (6175): 1246980.
31. Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” Nature 536 (7616): 285–91.
32. Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” BMC Bioinformatics 12 (August): 323.
33. Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” Bioinformatics 25 (14): 1754–60.
34. Li, Peng, Rosanne Spolski, Wei Liao, Lu Wang, Theresa L. Murphy, Kenneth M. Murphy, and Warren J. Leonard. 2012. “BATF-JUN Is Critical for IRF4-Mediated Transcription in T Cells.” Nature 490 (7421): 543–46.
35. Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. “RNA Splicing Is a Primary Link between Genetic Variation and Disease.” Science 352 (6285): 600–604.
36. Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology 15 (12): 550.

37. Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” Science 337 (6099): 1190–95.

38. McLaren William, Pritchard Bethan, Rios Daniel, Chen Yuan, Flicek Paul, Cunningham Fiona.

39. “Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.” Bioinformatics. 2010 Aug 15;26(16):2069-70.

40. McVicker, Graham, Bryce van de Geijn, Jacob F. Degner, Carolyn E. Cain, Nicholas E. Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K. Pritchard. 2013. “Identification of Genetic Variants That Affect Histone Modifications in Human Cells.” Science 342 (6159): 747–49.

41. Mudge, Jonathan M., and Jennifer Harrow. 2016. “The State of Play in Higher Eukaryote Gene Annotation.” Nature Reviews. Genetics 17 (12): 758–72.

42. Murphy, Theresa L., Roxane Tussiwand, and Kenneth M. Murphy. 2013. “Specificity through Cooperation: BATF-IRF Interactions Control Immune-Regulatory Networks.” Nature Reviews. Immunology 13 (7): 499–509.

43. Onengut-Gumuscu, Suna, Wei-Min Chen, Oliver Burren, Nick J. Cooper, Aaron R. Quinlan, Josyf C. Mychaleckyj, Emily Farber, et al. 2015. “Fine Mapping of Type 1 Diabetes Susceptibility Loci and Evidence for Colocalization of Causal Variants with Lymphoid Gene Enhancers.” Nature Genetics 47 (4): 381–86.

44. Pickrell, Joseph K. 2014. “Joint Analysis of Functional Genomic Data and Genome-Wide Association Studies of 18 Human Traits.” American Journal of Human Genetics 94 (4): 559–73.

45. Pociot, Flemming, Beena Akolkar, Patrick Concannon, Henry A. Erlich, Cécile Julier, Grant Morahan, Concepcion R. Nierras, John A. Todd, Stephen S. Rich, and Jørn Nerup. 2010. “Genetics of Type 1 Diabetes: What’s Next?” Diabetes 59 (7): 1561–71.

46. Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” American Journal of Human Genetics 81 (3): 559–75.

47. Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” Bioinformatics 26 (6): 841–42.

48. Raj, Towfique, Katie Rothamel, Sara Mostafavi, Chun Ye, Mark N. Lee, Joseph M. Replogle, Ting Feng, et al. 2014. “Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes.” Science 344 (6183): 519–23.

49. Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” Genome Biology 11 (3): R25.

50. Samee, Md. Abul Hassan, Benoit Bruneau, Katherine Pollard. 2017. “Transcription Factors Recognize DNA Shape Without Nucleotide Recognition.” bioRxiv, May. doi:10.1101/143677.

51. Samstein, Robert M., Aaron Arvey, Steven Z. Josefowicz, Xiao Peng, Alex Reynolds, Richard Sandstrom, Shane Neph, et al. 2012. “Foxp3 Exploits a Pre-Existent Enhancer Landscape for Regulatory T Cell Lineage Specification.” Cell 151 (1): 153–66.

52. Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. “Spatial Reconstruction of Single-Cell Gene Expression Data.” Nature Biotechnology 33 (5): 495–502.

53. Schaub, Marc A., Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. 2012. “Linking Disease Associations with Regulatory Information in the Human Genome.” Genome Research 22 (9): 1748–59.

54. Schmidhuber, Jürgen. 2015. “Deep Learning in Neural Networks: An Overview.” Neural Networks: The Official Journal of the International Neural Network Society 61 (January): 85–117.

55. Storey, John D., and Robert Tibshirani. 2003. “Statistical Significance for Genomewide Studies.” Proceedings of the National Academy of Sciences of the United States of America 100 (16): 9440–45.

56. Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “From the Cover: Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” Proceedings of the National Academy of Sciences 102 (January): 15545–50.

57. Tehranchi, Ashley K., Marsha Myrthil, Trevor Martin, Brian L. Hie, David Golan, and Hunter B. Fraser. 2016. “Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk.” Cell 165 (3): 730–41.

58. Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. “The Accessible Chromatin Landscape of the Human Genome.” Nature 489 (7414): 75–82.

59. Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” Bioinformatics 25 (9): 1105–11.

60. Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” Nature Protocols 7 (3): 562–78.

61. Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. “GCTA: A Tool for Genome-Wide Complex Trait Analysis.” American Journal of Human Genetics 88 (1): 76–82.

62. Ye, Chun Jimmie, Ting Feng, Ho-Keun Kwon, Towfique Raj, Michael T. Wilson, Natasha Asinovski, Cristin McCabe, et al. 2014. “Intersection of Population Variation and Autoimmunity Genetics in Human T Cell Activation.” Science 345 (6202): 1254665.

63. Zhou, Jian, and Olga G. Troyanskaya. 2015. “Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model.” Nature Methods 12 (10): 931–34.

Chapter 2: Genetic Architect: Discovering Genomic Structure with Learned Neural Architectures

Laura Deming*, Sasha Targ*, Nathaniel Sauder, Diogo Almeida, Chun Jimmie Ye

1 Introduction

Deep learning demonstrates excellent performance on tasks in computer vision, text and many other fields. Most deep learning architectures consist of matrix operations composed with non-linearity activations. Critically, the problem domain governs how matrix weights are shared. In convolutional neural networks – dominant in image processing – translational equivariance ("edge/color detectors are useful everywhere") is encoded through the use of the convolution operation; in recurrent networks – dominant in sequential data – temporal transitions are captured by shared hidden-to-hidden matrices. These architectures mirror human intuitions and priors on the structure of the underlying data. Genomics is an excellent domain to study how we might learn optimal architectures on poorly- understood data because while we have intuition that local patterns and long-range sequential dependencies affect genetic function, much structure remains to be discovered.

The genome is a very challenging data type, because although we have tens of thousands of whole genome sequences, we understand only a small subset of base pairs within each sequence. While the genetic code allows us to annotate the 5% of the genome encoding proteins (~20,000 genes in the human genome), we do not have a "grammar" for decoding the rest of the non-coding sequences (90-95% of the mouse and human genomes) important for gene regulation, evolution of species and susceptibility to diseases. The availability of a wealth of genomic assays

(PBM, CHIP-seq, Hi-C) allows us to directly measure the function of specific regions of the genome, creating an enormous opportunity to decode non-coding sequences. However, the overwhelming volume of new data makes our job as decoders of the genome quite complex. The design and application of new domain-specific architectures to these datasets is a promising approach for automating interpretation of genomic information into forms that humans can grasp.

2 Related Work

Inspired by human foveal attention where global glances drive sequential local focus, attention components have been added to neural networks yielding state-of-the-art results on tasks as diverse as caption generation, machine translation, protein sublocalization, and differentiable programming. There are two main architectural implementations: hard attention, where the network's focus mechanism non-differentiably samples from the available input, and soft attention, where the component outputs an expected glimpse using a weighted average. Beyond biological inspiration, these components enable improved performance and excellent intepretability. Other techniques have been applied for interpreting neural networks without changing their architectures (Simonyan et al. [2013], Zeiler and Fergus [2014], Springenberg et al. [2014]), but these are simply heuristics for finding the relevant regions of an input and do not work with all existing modern neural network components.

Previous groups have demonstrated excellent progress applying deep learning to genomics. Both Alipanahi et al. [2015] and Lanchantin et al. [2016] provide initial results on the task of learning which sequences a transcription factor (a biological entity which affects gene expression) can bind using convolutional architectures. This problem appears suited for convolution, as motifs

determining binding are expected to be modular (~7-10 base pair units) and the setup of the task (preselected input sequences of fixed short length) does not allow for learning significant long-term dependencies. In particular, Alipanahi et al. [2015] demonstrated that a single-layer convolutional neural network, DeepBind, outperformed 26 other tested machine learning approaches in predicting probe intensities on protein binding microarrays from the DREAM5 PBM challenge, and then showed that the same architecture generalized to the related task of predicting transcription factor binding sites (TFBSs) from sequencing measurements of bound DNA. Subsequently, Lanchantin et al. [2016] showed that a deeper network with the addition of highway layers improved on DeepBind results in the majority of cases tested [Srivastava et al., 2015]. In addition, Basset [Kelley et al., 2015], an architecture trained to predict motifs of accessible DNA from sequencing regions of open chromatin, was able to map half of the first layer convolutional filters to human TFBSs.

3 Development of Genetic Architect

Deep learning algorithm development is often dependent on the knowledge of human domain experts. Researchers in domains such as computer vision and natural language processing have spent much more time tuning architectures than in genomics. The challenge in genomics is that our insufficient understanding of biology limits our ability to inform architectural decisions based on data. Early genomic deep learning architectures have shown promising results but have undertaken only limited exploration of the architectural search space over possible components. In addition, not all components work well together, and there is evidence optimal component choice is highly dependent on the domain. Accordingly, we design a novel road-map for

applying deep learning to data on which we have limited prior understanding, by developing an iterative architecture search over standard and cutting-edge neural net building blocks.

Prior approaches to architecture search focus on finding the best architecture in a single step, rather than sequentially learning more about the architecture space and iteratively improving models (Bergstra et al. [2011], Bergstra and Bengio [2012], Snoek et al. [2012]). Our framework understands the results allowing us to sequentially narrow the search space and learn about which combinations of components are most important. Since our algorithm limits the most important hyperparameters to their best ranges, they no longer dominate the search space and we discover additional hyperparameters that are most important and can help us create a highly tuned architecture. The sequential nature allows us to fork our architectural search into independent subspaces of coadapted components, thus enabling further search in each parallel branch to be exponentially more efficient than considering the union of all promising architectures.

The heart of the framework is an interactive visualization tool (Figure 4). Given any hyperparameter optimization run, it produces common patterns for the best few datapoints and presents this information in highly-interpretable decision trees showing effective architectural subspace and plots of the interactions between the most significiant hyperparameters, informing general domain intuition and guiding future experiments. The framework is general enough to be applied to other domains, and is orthogonal to existing hyperparameter optimization algorithms. These algorithms can be applied in the inner loop of the sequential search of our tool, which then interprets the results and informs the user about the domain and how to manually prune the search space.

We employ Genetic Architect to discover an optimal architecture for a novel genome annotation task, regression to predict lineage-specific gene expression based on genomic sequence inputs, for which six stages of architecture search were required. Figure 1A shows the sequential process of architecture search, the most important findings at each stage of the process, and tool-guided division of the search into two separate promising architectures. By splitting effective architectures into separate branches for further optimization, our tool identifies high-performing but architecture-specific choices that may be difficult to notice when architectures are mixed together.

The application of our tool demonstrates the power in refining architectural components that dominate results to uncover additional hyperparameter combinations that perform well together. Several examples we encounter during use of the tool for design of architectures for genomics follow: 1) removal of batch normalization demonstrated clear superiority of exponential linear units, 2) dimensionality reduction in the middle of the convolutional network module was beneficial to the recurrent-based architectures (perhaps since it decreased the distance of long-range dependencies), and 3) in contrast, non-recurrent architectures required wider layers (likely to enable processing of long-range dependencies in final dense layers). In our search over architectures using soft attention, we found that fully-connected layers were preferred to convolutional layers as it made processing global information more important. Finally, only by proceeding through several steps of optimization did we find the unintuitive result that bidirectional LSTMs did not help with attentional models (perhaps because the preceding layer effectively attends to a single location, making it difficult to combine information from both

directions).

The final models learned by Genetic Architect consist of several initial layers of convolutions, residual blocks, an LSTM layer in the case of the PromoterNet architecture, and an attention-based dimensionality reducing step followed by fully-connected layers. Previous approaches to genome annotation use convolutional networks, which are ideal for detecting local features. However, more closely approximating the structure of genomic information would take into account that a real genome is a sequence, not a disjointed set, of local features – an input type on which recurrent architectures generally excel. In addition, with larger sequences to analyze (identifiable promoter sequences reach hundreds of base pairs in length), a neural network must learn to focus on the most important parts of the sequence and integrate new information derived from each part with the contextual information of the previously-seen sequence. As such, long genomic sequences seem an ideal fit for the recurrent attentional models learned.

4 Experimental Results

4.1 Tasks

4.1.1 Transcription factor binding site (TFBS) classification

The TFBS binary classification task was proposed by Alipanahi et al. [2015] and used as a benchmark by [Lanchantin et al., 2016]. The basic motivation is to learn a classifier that correctly predicts, from empirical binding data on a training sample of short DNA sequences, which sequences in a separate test set are TFBS (likely to be bound by biological entities, in this case, a given transcription factor protein).

The input and target data for the TFBS classification task consists of 108 datasets with an average of ~31,000 sequences of 101 characters per dataset. Each sequence is a string of base pairs (A, C, G, or T) and is transformed into an array with one-hot encoding. Each sequence has an associated label (1 or 0) which indicates if this sequence is a TFBS. Each dataset represents a different chromatin immunoprecipitation sequencing (ChIP-seq) experiment with a specified transcription factor, and each sequence in the dataset a potential binding site. For each positive example, a negative example is generated. The data included in the TFBS classification task derive from ENCODE CHIP-seq experiments performed in K562 transformed human cell lines [Consortium, 2012].

4.1.2 ImmGen lineage-specific expression prediction (ILSEP) regression

In addition to the TFBS classification problem, neural network architectures could be extended to treat a much broader and complex variety of problems to do with interpreting biological data. Here, we develop a novel genomic benchmark task, ILSEP, which requires regression to predict empirically-determined related target data, namely, prediction of the amount of various biological entities produced in different cellular contexts given an input genomic sequence.

The input dataset for the ILSEP task is 14,116 one-hot encoded (4,2000) input promoter sequences and corresponding (243,) floating point gene expression outputs ranging between 2.60 and 13.95 (see appendix for details). We split the dataset using 10-fold cross validation to obtain predictions for all promoter gene expression pairs.

4.2 Results on TFBS

4.2.1 Model performance

We benchmark the performance of AttentionNet models learned by hyperparameter optimization described above against published state-of-the-art neural network models on the TFBS task, DeepBind [Alipanahi et al., 2015] and DeepMotif [Lanchantin et al., 2016]. To compare the architectures, we train models for each of 108 datasets, as in Lanchantin et al. [2016]. In a head-to-head comparison on each dataset, AttentionNet outperforms DeepMotif in 67.6% of cases and the mean AUC across datasets for AttentionNet is 0.933, improving over both DeepMotif (0.927) and DeepBind (0.904) (Table 1).

4.2.2 Prediction and visualization of genomic information

Interpretable information about sequence features is an important consideration for genomic learning tasks where fundamental understanding of biology is as important as prediction power. We hypothesize that a net which performed well on the TFBS classification task would be able to make biologically meaningful inferences about the sequence structure. We show that the mean attention weights across all positive sequences show a distinct "footprint" of transcription factor (TF) binding consistent with known nucleotide preferences within each sequence (Figure 2B). Further, visualizing the attention mask (with the addition of Gaussian blur) across input sequences for 10 representative TFs showed the net focusing its attention on parts of the sequence known to be regulatory (Figure 2C).

To see if we could directly obtain motif sequences from the net, we took 10 nucleotides surrounding the position with highest attention for each of the top 100 sequences of a TF and averaged across the motifs. We took the maximum score for each nucleotide per position and queried the results against JASPAR, the "gold standard" TFBS database (with $q < 0.5$)

[Mathelier et al., 2015]. 30/57 motifs possible to check (i.e. in JASPAR) were correct, and 39/57 corresponded to at least one transcription factor. By additionally searching the top 3 recurring sequences attended to for each TF, we recover a total of 42/57 correct motifs.

4.3 Results on ILSEP

4.3.1 Model performance

The PromoterNet architecture demonstrates a marked gain in performance over DeepBind and Deep- Motif architectures adapted to the ILSEP regression task, achieving an average Pearson r correlation value of 0.587 between out-of-sample predictions and target expression values across lineages, com- pared to 0.506 and 0.441 for DeepBind [Alipanahi et al., 2015] and DeepMotif [Lanchantin et al., 2016] respectively (Figure 3A). We also train PromoterNet architectures on single task regression with a separate model for each of the 11 lineages and on cell type specific multi-task regression with one output unit for each of 243 cell types, which obtains similar improvements in average Pearson r correlation value of 0.592 over 0.502 for DeepBind and 0.498 for DeepMotif.

4.3.2 Promoter element recovery and visualization of proximal regulatory elements

Visualization of attention mask weights from the PromoterNet model reveals attended locations over promoter sequences of 32 genes selected for highest mean expression across lineages are enriched directly adjacent to the TSS, suggesting that properties of the core promoter sequence constitute the most informative features for genes that do not show differences in expression across lineages (Figure 3B) (see appendix for list of genes). In contrast, attended locations over promoter sequences of 32 genes with maximal variance in expression across lineages span a

much greater range of positions. This indicates that in genes with the greatest degree of lineage-specific expression, informative sequence features can occur throughout the promoter sequence. This observation merits follow up given previous reports that the performance of (non-deep) classifiers for cell type specific expression tasks trained only on TSS proximal promoter information is close to that of a random classifier [Natarajan et al., 2012]. Consistent with accepted understanding that TSS proximal regions contain genomic elements that control gene expression levels, we observe the maximum of average attention mask weights across all promoters occurs at the center of input sequences, which corresponds to core promoter elements required for recruitment of transcriptional machinery [Maston et al., 2006] (Figure 3C). PromoterNet models trained for multi-task regression result in a global attention mask output across all lineages. To investigate whether the PromoterNet architecture is capable of learning distinct features for each lineage, we also visualize attention weights for a given promoter sequence from separate models, each trained on expression data for a single lineage. We find that genes selected for maximal variance in expression demonstrate distinct patterns of learned attention across lineages, while a shared pattern of attention is learned for a control gene with high mean expression in all lineages even when each lineage was trained on a separate model (Figure 3D).

5 Conclusions

We tackle the problem of discovering architectures on datasets where human priors are not available. To do so we create a novel architecture search framework that is domain agnostic, is capable of sequential architectural subspace refinement and informing domain understanding, and is composable with existing hyperparameter optimization schemes. Using this search

algorithm, we create state- of-the art architectures on significant challenges in the domain of genomics utilizing a combination of standard and cutting-edge components. In particular, the learned architecture is capable of simultaneous discovery of local and non-local patterns, important subsequences, and sequential composition thereby capturing substantial genomic structure.

6 Appendix

6.1 ILSEP data processing

For the input sequences in the ILSEP task, we use sequences spanning 1 kilobase upstream and downstream of transcriptional start sites (TSS), the region of the promoter at which production of the gene is initiated, for 17,565 genes from the Eukaryotic Promoter Database [epd]. For the corresponding labels, we obtain expression data (log2 normalized microarray intensities) for each of these genes in each of 243 immune cell types from the ImmGen Consortium April 2012 release, which contains data for 21,755 genes x 243 immune cell types after quality control according to published methods [Ericson et al.]. Intersection of these two datasets by gene results in a dataset of 14,116 input promoter sequences and expression value target pairs.

To create lineage-specific gene expression value targets, we combine cell types into 11 groups following the lineage tree outlined in previous work: B cells (B), dendritic cells (DC), gamma delta T cells (γδ), granulocytes (GN), macrophages (Mφ), monocytes (MO), natural killer cells (NK), stem and progenitor cells (SP), CD4+ T cells (T4), and CD8+ T cells (T8), and average expression values across samples within each group [Jojic et al., 2013].

6.2 Selected genes (Figure 3B)

Highest mean expression across lineages: Rac2, Rpl28, Pfn1, Rpl9, Ucp2, Tmsb4x, Tpt1, Rplp1, Hspa8, Srgn, Rpl27a, Rpl13a, Cd53, Eef2, Rps26, Cfl1, Ppia, Gm9104, Rps2, Rps27, Actg1, Laptm5, Rpl21, Eef1a1, Rplp0, Gm15427, Pabpc1, B2m, Gapdh, Actb, Rpl17, Rps6

Highest variance in expression across lineages: Plbd1, Tlr13, Tyrobp, Ifitm2, Pld4, Pla2g7, Gda, Cd96, Gzma, Nkg7, Ctsh, Klrb1c, Ccl6, Prkcq, Itgam, Sfpi1, Itk, Ms4a4b, Alox5ap, Ly86, Cd2, Fcer1g, Gimap3, Il2rb, Gimap4, Ifitm6, Cybb, Ifitm3, Mpeg1, H2-Aa, Cd3g, Lyz2

Random control: Krt84, Lrrc8b, 8030411F24Rik, Syngr2, Spint3, Slc17a4, Slc22a23, Thoc6, AF529169, Phf5a, Yif1b, 4930467E23Rik, Pgam1, Pcdhb1, Bak1, Neu3, Plcb2, Fabp4, Srgap1, Olfr1339, Sox12, Atg7, Gdf10, 1810008A18Rik, 1700011A15Rik, Anks4b, Magea2, Pygb, Spc25, Rras2, Slc28a3, 9130023H24Rik

Table 1: Mean and median AUC of models and percentage of datasets on which each model outperforms DeepBind or DeepMotif.

model	mean AUC	median AUC	vs DeepBind	vs DeepMotif
DeepBind	0.904	0.932	-	-
DeepMotif	0.927	0.949	85.2	-
AttentionNet	0.933	0.952	92.6	67.6

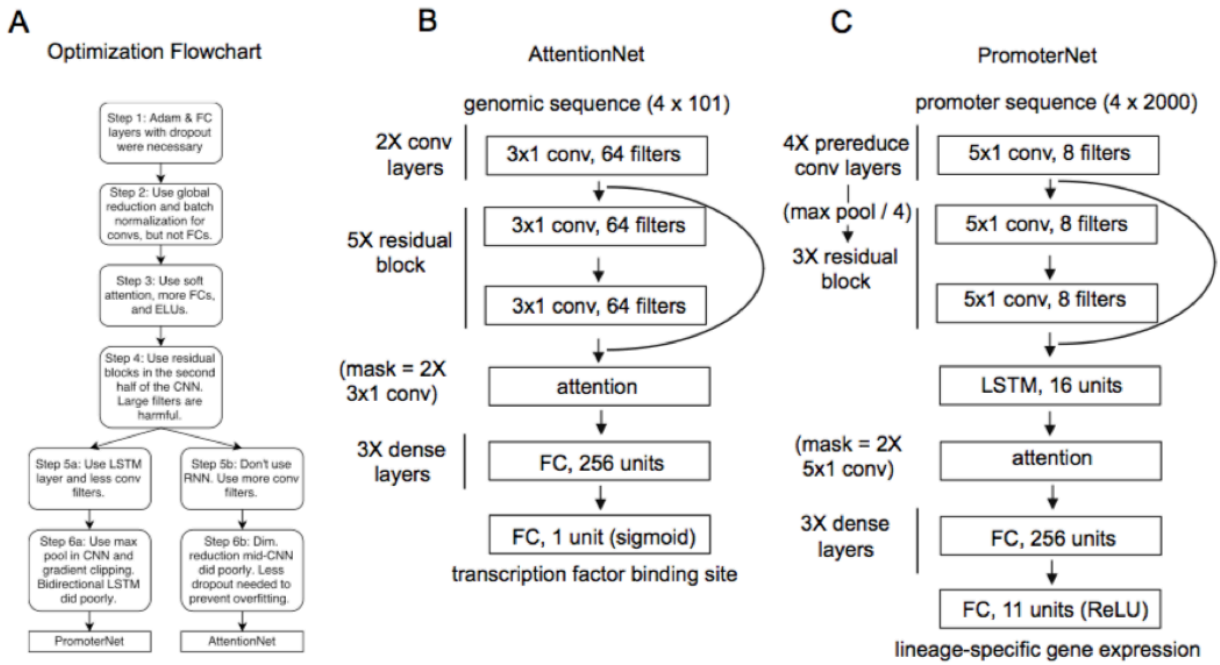

Figure 1: Schematic of hyperparameter optimization and final architecture designs. A) Overview of steps taken in hyperparameter optimization to generate AttentionNet and PromoterNet. B) AttentionNet architecture. C) PromoterNet architecture.

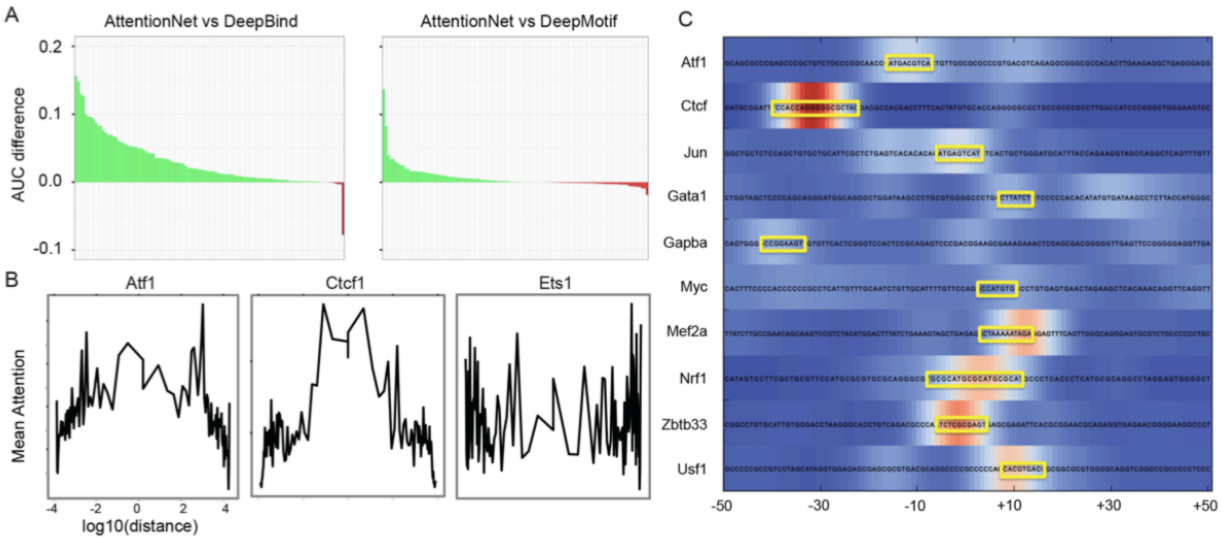

Figure 2: Results of AttentionNet on transcription factor binding site (TFBS) task. A) AttentionNet models outperform DeepMotif and DeepBind models trained on corresponding datasets. Each bar represents the difference in AUC for one of 108 different datasets. B) Mean of attention mask over all sequences in experiment. C) Recovery of transcription factor motifs by visualization of attention masks produced by AttentionNet over example sequences.

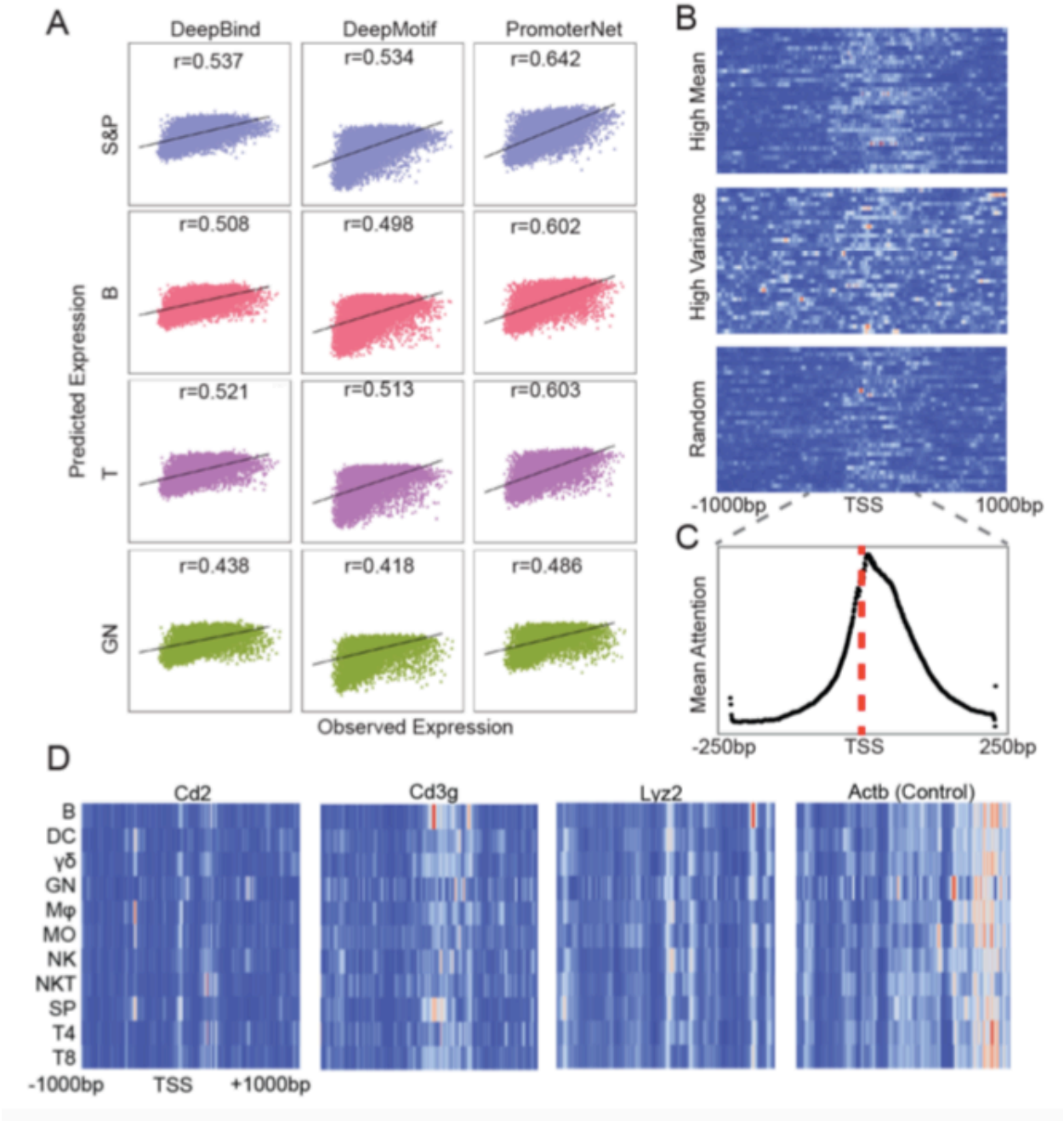

Figure 3: Results of PromoterNet on ImmGen lineage-specific expression prediction (ILSEP) task. A) Comparison of predicted versus observed gene expression for DeepBind, DeepMotif, and PromoterNet architectures. B) Visualization of attention mask over selected promoter sequences. C) Mean attention mask over all promoters. D) Visualization of attention masks learned by models trained on data from single lineages.

Table 2: Search space explored for AttentionNet and PromoterNet architectures, including techniques from Maas et al. [2013], Graham [2014], Shah et al. [2016], Ioffe and Szegedy [2015], He et al. [2015], Hochreiter and Schmidhuber [1997], Kingma and Ba [2014], Sutskever et al. [2013], Srivastava et al. [2014].

Hyperparameter	Values
conv filter size	3, 5, 7, 9
nonlinearity	ReLU, Leaky ReLU, Very Leaky ReLU, ELU
using batch normalization for convs	True, False
number of conv filters	8, 16, 32, 64
number of convs before dim. reduction	1, 2, 3, 4, 5
using residual blocks before dim. reduction	True, False
type of dim. reduction	None, Max Pool, Mean Pool, Strided Conv
dim. reduction stride	2, 4, 8
number of convs after dim. reduction	1, 2, 3, 4, 5
using residual blocks after dim. reduction	True, False
number of RNN layers	0, 1, 2
number of units in RNN	16, 32, 64
RNN type	Simple RNN, LSTM
using bidirectional RNNs	True, False
RNN gradient clipping	0, 1, 5, 15
global reduction type	None, Attention, Max Pool, Mean Pool
number of FC layers	0, 1, 2, 3
number of units in FC	64, 128, 256, 512, 1024
using batch normalization for FCs	True, False
dropout probability for FCs (after)	0., 0.1, 0.2, 0.3, 0.4, 0.5, 0.6
L2 regularization	0, 1e-3, 1e-4, 1e-5
optimizer	Adam, SGD w/ Nesterov Momentum
learning rate scale	0.03, 0.1, 0.3, 1.0, 2.0, 10.
batch size	25, 50, 125, 250

Table 3: AttentionNet architecture

Layer
3x1 conv 64 filters + BN + ELU
3x1 conv 64 filters + BN + ELU
residual block (w/ 3x1 conv 64 filters + BN + ELU + 3x1 conv 64 filters + BN)
residual block (w/ 3x1 conv 64 filters + BN + ELU + 3x1 conv 64 filters + BN)
residual block (w/ 3x1 conv 64 filters + BN + ELU + 3x1 conv 64 filters + BN)
residual block (w/ 3x1 conv 64 filters + BN + ELU + 3x1 conv 64 filters + BN)
residual block (w/ 3x1 conv 64 filters + BN + ELU + 3x1 conv 64 filters + BN)
attention (w/ 3x1 conv 64 filters + BN + tanh + 3x1 conv 1 filter + BN + softmax)
FC 256 units + ELU + 0.2 dropout
FC 256 units + ELU + 0.2 dropout
FC 256 units + ELU + 0.2 dropout
FC 1 unit + sigmoid

Table 4: PromoterNet architecture

Layer
5x1 conv 8 filters + BN + ELU
5x1 conv 8 filters + BN + ELU
5x1 conv 8 filters + BN + ELU
5x1 conv 8 filters + BN + ELU
4x1 maxpool, stride 4
residual block (w/ 5x1 conv 8 filters + BN + ELU + 5x1 conv 8 filters + BN)
residual block (w/ 5x1 conv 8 filters + BN + ELU + 5x1 conv 8 filters + BN)
residual block (w/ 5x1 conv 8 filters + BN + ELU + 5x1 conv 8 filters + BN)
LSTM 16 units
attention (w/ 5x1 conv 8 filters + BN + tanh + 5x1 conv 1 filter + BN + softmax)
FC 256 units + ELU + 0.3 dropout
FC 256 units + ELU + 0.3 dropout
FC 256 units + ELU + 0.3 dropout
FC 1 unit + sigmoid

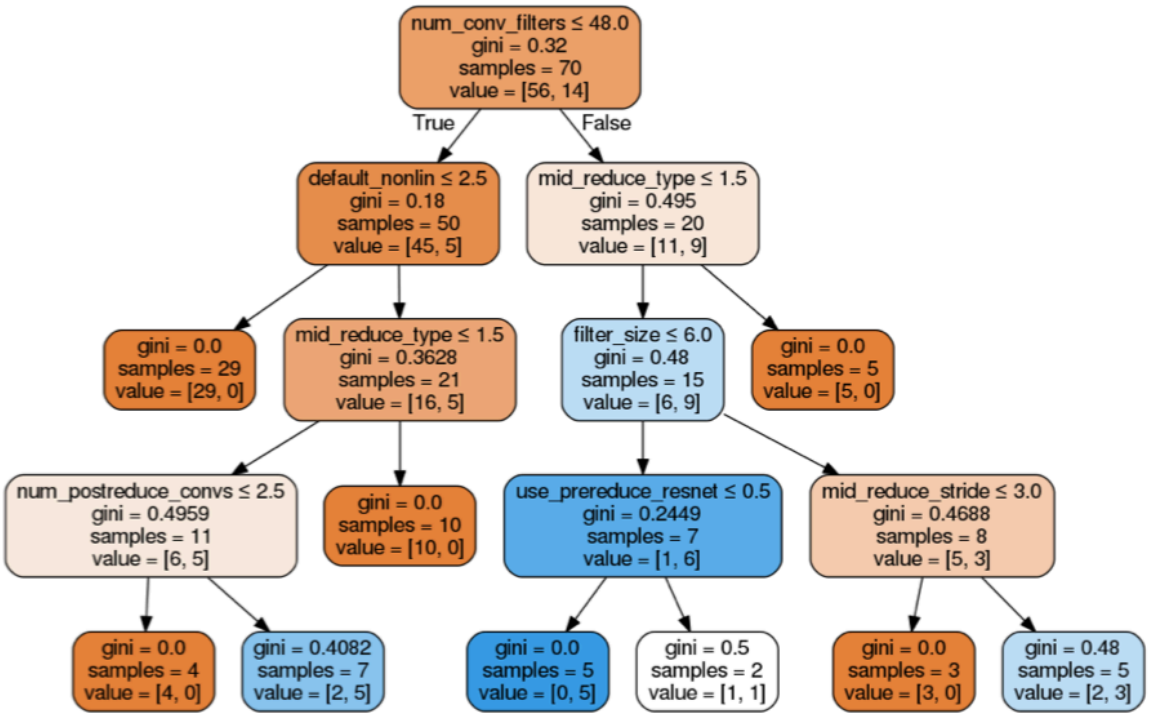

Figure 4: Example decision tree output from Genetic Architect visualization tool depicting significant hyperparameters for models with top 20% performance.

References

1. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.
2. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In Computer vision–ECCV 2014, pages 818–833. Springer, 2014.
3. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014.
4. Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology, 2015.
5. Jack Lanchantin, Ritambhara Singh, Zeming Lin, and Yanjun Qi. Deep motif: Visualizing genomic sequence classifications. arXiv preprint arXiv:1605.01133, 2016.
6. Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In Advances in Neural Information Processing Systems, pages 2368–2376, 2015.
7. David R Kelley, Jasper Snoek, and John Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. bioRxiv, page 028399, 2015.
8. James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems, pages 2546–2554, 2011.

9. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1):281–305, 2012.
10. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In Advances in neural information processing systems, pages 2951–2959, 2012.
11. ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. Nature, 489(7414):57–74, 2012.
12. Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, et al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic acids research, page gkv1176, 2015.
13. Anirudh Natarajan, Galip Gürkan Yardımcı, Nathan C. Sheffield, Gregory E. Crawford, and Uwe Ohler. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Research, 22(9):1711–1722, 2012. doi: 10.1101/gr.135129.111. URL http://genome.cshlp. org/content/22/9/1711.abstract.
14. Glenn A. Maston, Sarah K. Evans, and Michael R. Green. Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics, 7(1):29–59, 2006. doi: 10.1146/annurev.genom.7.080505.115623. URL http://dx.doi.org/10.1146/annurev. genom.7.080505.115623. PMID: 16719718.
15. Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In Proc. ICML, volume 30, page 1, 2013.
16. Benjamin Graham. Spatially-sparse convolutional neural networks. arXiv preprint arXiv:1409.6070, 2014.

17. Anish Shah, Eashan Kadam, Hena Shah, and Sameer Shinde. Deep residual networks with exponential linear unit. arXiv preprint arXiv:1604.04112, 2016.
18. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.
19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.
20. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.
21. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
22. Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th international conference on machine learning (ICML-13), pages 1139–1147, 2013.
23. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15:1929–1958, 2014.
24. Eukaryotic promoter database. URL http://epd.vital-it.ch/mouse/mouse_database.php.
25. Jeff Ericson, Scott Davis, Jon Lesh, Melissa Howard, Diane Mathis, and Christophe Benoist. Immgen microarray gene expression data: Data generation and quality control pipeline. URL www.immgen. org/Protocols/ImmGenQCDocumentation_ALL-DataGeneration_0612.pdf.
26. Vladimir Jojic, Tal Shay, Katelyn Sylvia, Or Zuk, Xin Sun, Joonsoo Kang, Aviv Regev, Daphne Koller, Immunological Genome Project Consortium, et al. Identification of

transcriptional regulators in the mouse immune system. Nature immunology, 14(6):633–643, 2013.

Chapter 3: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation

Hyun Min Kang*$[1], Meena Subramaniam$[2-6], Sasha Targ$[2-6,11], Michelle Nguyen[7-9], Lenka Maliskova[3,10], Elizabeth McCarthy[11], Eunice Wan[3], Simon Wong[3], Lauren Byrnes[12], Cristina Lanata[13,14], Rachel Gate[2-6], Sara Mostafavi[15], Alexander Marson[7-9,16,17], Noah Zaitlen[3,13,18], Lindsey A Criswell[3,13,14,19], Chun Jimmie Ye[3-6]*

Introduction

Droplet single cell RNA-sequencing (dscRNA-seq) has increased substantially the throughput of single cell capture and library preparation[1, 10], enabling the simultaneous profiling of thousands of cells. Improvements in biochemistry[11, 12] and microfluidics[13, 14] continue to increase the number of cells and transcripts profiled per experiment. But for differential expression and population genetics studies, sequencing thousands of cells each from many individuals would better capture inter-individual variability than sequencing more cells from a few individuals. However, in standard workflows, dscRNA-seq of many samples in parallel remains challenging to implement. If the genetic identity of each cell could be determined, pooling cells from different individuals in one microfluidic run would result in lower per-sample library preparation cost and eliminate confounding effects. Furthermore, if droplets containing multiple cells from different individuals could be detected, pooled cells could be loaded at higher concentrations, enabling additional reduction in per-cell library preparation cost.

Here we develop an experimental protocol for multiplexed dscRNA-seq and a computational algorithm, demuxlet[32], that harnesses genetic variation to determine the genetic identity of each

cell (demultiplex) and identify droplets containing two cells from different individuals (**Fig. 1a**). While strategies to demultiplex cells from different species[1, 10, 17] or host and graft samples[17] have been reported, simultaneously demultiplexing and detecting doublets from more than two individuals has not been possible. Inspired by models and algorithms developed for detecting contamination in DNA sequencing[18], demuxlet is fast, accurate, scalable, and compatible with standard input formats[17, 19, 20].

Demuxlet implements a statistical model for evaluating the likelihood of observing RNA-seq reads overlapping a set of single nucleotide polymorphisms (SNPs) from a single cell. Given a set of best-guess genotypes or genotype probabilities obtained from genotyping, imputation or sequencing, demuxlet uses maximum likelihood to determine the most likely donor for each cell using a mixture model. A small number of reads overlapping common SNPs is sufficient to accurately identify each cell. For a pool of 8 individuals and a set of uncorrelated SNPs each with 50% minor allele frequency (MAF), 4 reads overlapping SNPs are sufficient to uniquely assign a cell to the donor of origin (**Fig. 1b**) and 20 reads overlapping SNPs can distinguish every sample with >98% probability in simulation (Supplementary Fig. 1). We note that by multiplexing even a small number of individuals, the probability that a doublet contains cells from different individuals is very high (1 – 1/N, e.g., 87.5% for N=8 samples) (**Fig. 1C**). For example, if a 1,000-cell run without multiplexing results in 990 singlets with a 1% undetected doublet rate, multiplexing 1,570 cells each from 63 samples can theoretically achieve the same rate of undetected doublets, producing up to a 37-fold more singlets (36,600) if the sample identity of every droplet can be perfectly demultiplexed (Supplementary Fig. 2, see Methods for details). To minimize the effects of sequencing doublets, profiling 22,000 cells multiplexed from

26 individuals generates 23-fold more singlets at the same effective doublet rate (Supplementary Fig. 3).

Results

We first assess the performance of multiplexed dscRNA-seq through simulation. The ability to demultiplex cells is a function of the number of individuals multiplexed, the depth of sequencing or number of read-overlapping SNPs, and relatedness of multiplexed individuals. We simulated 6,145 cells (5,837 singlets and 308 doublets) from 2 – 64 individuals from the 1000 Genomes Project[21]. We show that 50 SNPs per cell allows demultiplexing of 97% of singlets and identification of 92% of doublets in pools of up to 64 individuals (Supplementary Figs. 4-5, see Methods for details). Simulating a range of sequencing depths, we determined that 50 SNPs can be obtained with as few as 1,000 unique molecular identifiers (UMIs) per cell (Supplementary Fig. 6)**,** and recommended sequencing depths of standard dscRNA-seq workflows would capture hundreds of SNPs. To assess dependence on the relatedness of multiplexed individuals, we simulated 6,145 cells from a set of 8 related individuals from 1000 Genomes[21]. In this simulation, 50 SNPs per cell would allow demuxlet to correctly assign over 98% of cells (Supplementary Fig. 7). These results suggest optimal multiplexed designs where cells from tens of unrelated individuals should be pooled, loaded at concentrations 2-10x higher than standard workflows, and sequenced to at least 1,000 UMIs per cell.

We evaluate the performance of demuxlet by analyzing a pool of peripheral blood mononuclear cells (PBMCs) from 8 lupus patients. By sequential pairwise pooling, three pools of equimolar concentrations of cells were generated (W1: patients S1-S4, W2: patients S5-S8 and W3: patients

S1-S8) and each loaded in a well on a 10X Chromium instrument (**Fig. 2a**). 3,645 (W1), 4,254 (W2) and 6,205 (W3) cell-containing droplets were sequenced to an average depth of 51,000, 39,000 and 28,000 reads per droplet.

In wells W1, W2 and W3, demuxlet identified 91% (3332/3645), 91% (3864/4254), and 86% (5348/6205) of droplets as singlets (likelihood ratio test, L(singlet)/L(doublet) > 2), of which 25% (+/- 2.6%), 25% (+/- 4.6%) and 12.5% (+/- 1.4%) mapped to each donor, consistent with equal mixing of individuals in each well. From wells W1 and W2, each containing cells from two disjoint sets of 4 individuals, we estimated a demultiplexing error rate (number of cells assigned to individuals not in the pool) of less than 1% of singlets (W1: 2/3332, W2: 0/3864) (**Fig. 2b**).

We next assess the ability of demuxlet to detect doublets in both simulated and real data. 466/3645 (13%) droplets from W1 were simulated as synthetic doublets by setting the cellular barcodes of 466 cells each from individuals S1 and S2 to be the same. Applied to simulated data, demuxlet identified 91% (426/466) of synthetic doublets as doublets or ambiguous, correctly recovering the sample identity of both cells in 403/426 (95%) doublets (Supplementary Fig. 8). Applied to real data from W1, W2 and W3, demuxlet identified 138/3645, 165/4254, and 384/6205 doublets corresponding to doublet rates of 5.0%, 5.2% and 7.1%, consistent with the expected doublet rates estimated from mixed species experiments (**Fig. 2c**).

Demultiplexing of pooled samples allows for the statistical and visual comparisons of individual-specific dscRNA-seq profiles. Singlets identified by demuxlet in all three wells cluster into

known immune cell types (**Fig. 2d**) and are correlated with bulk RNA-sequencing of sorted cell populations (R=0.76-0.92) (Supplementary Fig. 9). For the same individuals from different wells, t-distributed stochastic neighbor embedding (t-SNE) of dscRNA-seq data are qualitatively consistent, and estimates of cell type proportions are highly correlated (R = 0.99) (**Fig. 2e** and Supplementary Fig. 10). Further, t-SNE projections of the pool and each individual are not confounded by well-to-well effects (Supplementary Fig. 11a). While 6 genes were differentially expressed between wells W1 and W2 (DESeq2 on pseudobulk counts, FDR < 0.05), only 2 genes were differentially expressed between W1 and W2 individuals in well W3 (FDR < 0.05) (Supplementary Fig. 11b), suggesting multiplexing reduces technical effects due to separate sample processing[22, 23].

We used multiplexed dscRNA-seq to characterize the cell type specificity and inter-individual variability of response to IFN-β, a potent cytokine that induces genome-scale changes in the transcriptional profiles of immune cells[24, 25]. From each of 8 lupus patients, PBMCs were activated with recombinant IFN-β or left untreated for 6 hours, a time point we previously found to maximize the expression of interferon-sensitive genes (ISGs) in dendritic cells (DCs) and T cells[26, 27]. Two pools, IFN-β-treated and control, were prepared with the same number of cells from each individual and loaded onto the 10X Chromium instrument.

We obtained 14,619 (control) and 14,446 (stimulated) cell-containing droplets, of which demuxlet identified 83% (12,138) and 84% (12,167) as singlets. The estimated doublet rate of 10.9% in each condition is consistent with predicted rates (**Fig. 2C**) and the observed and expected frequencies of doublets for each pair of individuals are highly correlated (R=0.98)

(Supplementary Fig. 12). Detected doublets form distinct clusters near the periphery of other clusters defined by cell type (Supplementary Fig. 13).

Demultiplexing individuals enables the use of the 8 individuals within each pool as biological replicates to quantitatively assess cell type-specific IFN-β responses in PBMCs. Consistent with previous reports from bulk RNA-sequencing data, IFN-β stimulation induces widespread transcriptomic changes observed as a shift in the t-SNE projections of singlets[24] (**Fig. 3A**). As expected, IFN-β did not affect cell type proportions between control and stimulated cells (Supplementary Fig. 14), and these were consistent with flow cytometry measurements (R=0.88) (Supplementary Fig. 15). Estimates of abundances for ~2000 homologous genes in each cell type and condition correlated with similar data from mice (Supplementary Fig. 16). We identified 3,055 differentially expressed genes (logFC > 2, FDR < 0.05) in at least one cell type (Supplementary Table 1). For 709 genes, estimates of fold change in response to IFN-β stimulation in myeloid and $CD4^+$ cells are consistent with estimates in monocyte derived dendritic cells[28] and $CD4^+$ T cells[27], respectively (Supplementary Fig. 17) and correlated with qPCR results of sorted $CD4^+$ T cells (Supplementary Fig. 18). Differentially expressed genes cluster into modules of cell type-specific responses enriched for distinct gene regulatory programs (**Fig. 3B,** Supplementary Table 2). For example, genes upregulated in all leukocytes (Cluster III: 401 genes, logFC > 2, FDR < 0.05) or only in myeloid cells (Cluster I: 767 genes, logFC > 2, FDR < 0.05) are enriched for general antiviral response (e.g. KEGG Influenza A: Cluster III $P < 1.6 \times 10^{-5}$), chemokine signaling (Cluster I $P < 7.6 \times 10^{-3}$) and pathways active in systemic lupus erythematosus (Cluster I $P < 4.4 \times 10^{-3}$). The five clusters of downregulated genes are enriched for antibacterial response (KEGG Legionellosis: Cluster II monocyte down P <

$5.5x10^{-3}$) and natural killer cell mediated toxicity (Cluster IV NK/Th cell down: $P < 3.6x10^{-2}$). The analysis of multiplexed dscRNA-seq data recovers cell type-specific gene regulatory programs affected by interferon stimulation consistent with published IFN-β signatures in mouse and humans[29].

Over all PBMCs, the variance of mean expression across individuals is higher than the variance across synthetic replicates whose cells were randomly sampled (Lin's concordance = 0.022, Pearson correlation= 0.69, **Fig. 3C**). The variance across synthetic replicates whose cells were sampled matching for cell type proportions is more concordant with the variance across individuals (Lin's concordance = 0.54, Pearson correlation = 0.78, **Fig. 3C-D**), suggesting a contribution of cell type composition on expression variability. However, for each cell type, the variance across individuals[22, 30] is also higher than the variance across synthetic replicates (Lin's concordance = 0.007-0.20) suggesting additional inter-individual variability not due to cell type composition (Supplementary Fig. 19). In $CD14^{+}CD16^{-}$ monocytes, the correlation of mean expression between pairs of synthetic replicates from the same individual (>99%) is greater than from different individuals (~97%), further indicating inter-individual variation beyond sampling (**Fig. 3E**). We found between 15 to 827 genes with statistically significant inter-individual variability in control cells and 7 to 613 in stimulated cells (Pearson correlation, FDR < 0.05), with most found in classical monocytes (cM) and $CD4^{+}$ helper T (Th) cells. Inter-individual variable genes in stimulated cM and to a lesser extent in Th cells ($P < 9.3x10^{-4}$ and $4.5x10^{-2}$, hypergeometric test, **Fig. 3F**) are enriched for differentially expressed genes, consistent with our previous discovery of more IFN-β response-eQTLs in monocyte-derived dendritic cells than $CD4^{+}$ T cells[26, 27]. Comparing to 407 genes previously profiled in bulk monocyte-derived

dendritic cells, the proportion of variance explained by inter-individual variability is more correlated in myeloid cells after stimulation (R = 0.26 – 0.3) than before (R = 0.05 – 0.19).

To map genetic variants associated with cell type proportions and cell type-specific expression using multiplexed dscRNA-seq, we sequenced an additional 15,250 (7 donors), 22,619 (8 donors) and 25,918 cells (15 donors; 8 lupus patients, 5 rheumatoid arthritis patients, and 2 healthy controls). Demuxlet identified 71% (10,766/15,250), 73% (16,618/22,619) and 60% (15,596/25,918) of droplets as singlets, correctly assigning 99% of singlets from the first two pools, W1 and W2 (10,740/10,766 and 16,616/16,618). The estimated doublet rates of 18%, 18% and 25% are consistent with the increased concentrations of loaded cells (**Fig. 2C**). Similar to the IFN-β stimulation experiment, we found that expression variability was determined by variability in cell type proportion (**Fig. 4A**) and reproducible between batches (Supplementary Fig. 20). Associating >150,000 genetic variants (MAF > 20%) with the proportion of 8 major immune cell populations, we identified a SNP (chr10:3791224) significantly associated ($P = 1.03 \times 10^{-5}$, FDR < 0.05) with the proportion of NK cells (**Fig. 4B**).

Across 23 donors, we conducted an expression quantitative trait loci (eQTL) analysis to map genetic variants associated with expression variability in each major immune cell type. We found a total of 32 local eQTLs (+/- 100kb, FDR < 0.1), 22 of which were detected in only one cell type (**Fig. 4C**, Supplementary Table 3). Previously reported local eQTLs from bulk $CD14^+$ monocytes, $CD4^+$ T cells and lymphoblastoid cell lines are more significantly associated with gene expression in the most similar cell types (cM, Th and B cells, respectively) than other cell types (**Fig. 4D**). We used an inverse variance weighted meta-analysis to identify genes with pan-

cell type eQTLs, including those in the major histocompatibility complex (MHC) class I antigen presentation pathway including *ERAP2* ($P < 3.57x10^{-32}$, meta-analysis), encoding an aminopeptidase known to cleave viral peptides[34], and *HLA-C* ($P < 1.74x10^{-29}$, meta-analysis), which encodes the MHC class I heavy chain (**Fig. 4E**). *HLA-DQA1* has local eQTLs only in some cell types ($P < 2.11x10^{-15}$, Cochran's Q) while *HLA-DQA2* has local eQTLs in all antigen presentation cells ($P < 1.02e10^{-43}$, Cochran's Q). Among other cell type-specific local eQTLs are *CD52*, a gene ubiquitously expressed in leukocytes that only has eQTLs in monocyte populations, and *DIP2A*, a gene with an eQTL only in NK cells that is associated with immune response to vaccination in peripheral blood[35]. These results demonstrate the ability of multiplexed dscRNA-seq to characterize inter-individual variation in immune response and when integrated with genetic data, reveal cell type-specific genetic control of gene expression, which would be undetectable when bulk tissues are analyzed.

Discussion

The capability to demultiplex and identify doublets using natural genetic variation reduces the per-sample and per-cell library preparation cost of single-cell RNA-sequencing, does not require synthetic barcodes or split-pool strategies[36-40], and captures biological variability among individual samples while limiting unwanted technical variability. We find the optimal number of samples to multiplex is approximately 20, based on sample processing time and empirical doublet rates of current microfluidic devices and anticipate that number to increase with automated sample handling and lower doublet rates.

Compared to sorting known cell types followed by bulk RNA-seq, multiplexed dscRNA-seq is a more efficient and unbiased method for obtaining cell type-specific immune traits[41]. Demuxlet enables reliable estimation of cell type proportion, recovers cell type-specific transcriptional response to stimulation, and could facilitate further genetic and longitudinal analyses in relevant cell types and conditions across a range of sampled individuals, including between healthy controls and disease patients[42-44]. While demuxlet could in principle be applied to sequencing solid tissue, standardizing sample processing and preservation remain major challenges. Although we developed demuxlet specifically for RNA-sequencing, we anticipate that the computational framework could be easily extended to other single cell assays where synthetic barcodes or natural genetic variation are measured by sequencing.

Contributions: HMK and CJY conceived the project. MS, ST, LM, RG, LB, EW, SW, and MN performed all experiments. HMK, MS, ST, EM, SM, and CJY analyzed the data. CL and LAC provided the patient samples. NZ and AM provided helpful comments and discussion. HMK, MS, ST, and CJY wrote the manuscript.

Methods

Identifying the sample identity of each single cell.

We first describe the method to infer the sample identity of each cell in the absence of doublets. Consider RNA-sequence reads from *C* barcoded droplets multiplexed across *S* different samples, where their genotypes are available across *V* exonic variants. Let d_{cv} be the number of unique reads overlapping with the *v*-th variant from the *c*-th droplet. Let $b_{cvi} \in \{R, A, O\},\ i \in \{1, \cdots, d_{cv}\}$ be the variant-overlapping base call from the *i*-th read, representing reference (R),

alternate (A), and other (O) alleles respectively. Let $e_{cvi} \in \{0,1\}$ be a latent variable indicating whether the base call is correct (0) or not (1), then given $e_{cvi} = 0, b_{cvi} \in \{R = 0, A = 1\}$ and $\sim \text{Binomial}\left(2, \frac{g}{2}\right)$ when $g \in \{0,1,2\}$ is the true genotype of sample corresponding to *c*-th droplet at *v*-th variant. When $e_{cvi} = 1$, we assume that $\Pr(b_{cvi}|g, e_{cvi})$ follows Supplementary Table 4. e_{cvi} is assumed to follow $\text{Bernoulli}\left(10^{-\frac{q_{cvi}}{10}}\right)$ where q_{cvi} is a phred-scale quality score of the observed base call. We use the standard 10X pipeline to process the raw reads which estimates the phred-scale quality score based on the alignment of each read to the reference human transcriptome using the STAR aligner[49].

We allow uncertainty of observed genotypes at the *v*-th variant for the *s*-th sample using $P_{sv}^{(g)} = \Pr(g|\text{Data}_{sv})$, the posterior probability of a possible genotype g given external DNA data Data_{sv} (e.g. sequence reads, imputed genotypes, or array-based genotypes). If genotype likelihood $\Pr(\text{Data}_{sv}|g)$ is provided (e.g. unphased sequence reads) instead, it can be converted to a posterior probability scale using $P_{sv}^{(g)} = \Pr(\text{Data}_{sv}|g)\Pr(g)$ where $\Pr(g) \sim \text{Binomial}(2, p_v)$ and p_v is the population allele frequency of the alternate allele. To allow errors ε in the posterior probability, we replace it with $(1-\varepsilon)P_{sv}^{(g)} + \varepsilon\Pr(g)$. The overall likelihood that the *c*-th droplet originated from the *s*-th sample is

$$L_c(s) = \prod_{v=1}^{V}\left[\sum_{g=0}^{2}\left\{\prod_{i=1}^{d_{cv}}\left(\sum_{e=0}^{1}\Pr(b_{cvi}|g,e)\right)P_{sv}^{(g)}\right\}\right] \quad (1)$$

In the absence of doublets, we use the maximum likelihood to determine the best-matching sample as $\text{argmax}_s[L_c(s)]$.

Screening for droplets containing multiple samples.

To identify doublets, we implement a mixture model to calculate the likelihood that the sequence reads originated from two individuals, and the likelihoods are compared to determine whether a droplet contains cells from one or two samples. If sequence reads from the *c*-th droplet originate from two different samples, s_1, s_2 with mixing proportions $(1-\alpha):\alpha$, then the likelihood in (1) can be represented as the following mixture distribution[18],

$$L_c(s_1, s_2, \alpha) = \prod_{v=1}^{V}\left[\sum_{g_1,g_2}\left\{\prod_{i=1}^{d_{cv}}(\sum_{e=0}^{1}(1-\alpha)\Pr(b_{cvi}|g_1,e) + \alpha\Pr(b_{cvi}|g_2,e))P_{sv}^{(g_1)}P_{sv}^{(g_2)}\right\}\right]$$

To reduce the computational cost, we consider discrete values of $\alpha \in \{\alpha_1, \cdots, \alpha_M\}$, (e.g. 5 - 50% by 5%). We determine that it is a doublet between samples s_1, s_2 if and only if $\frac{\max_{s_1,s_2,\alpha} L_c(s_1,s_2,\alpha)}{\max_s L_c(s)} \geq t$ and the most likely mixing proportion is estimated to be $\text{argmax}_\alpha L_c(s_1, s_2, \alpha)$. We determine that the cell contains only a single individual *s* if $\frac{\max_{s_1,s_2,\alpha} L_c(s_1,s_2,\alpha)}{\max_s L_c(s)} \leq \frac{1}{t}$, and less confident droplets are classified as ambiguous. While we consider only doublets for estimating doublet rates, we remove all doublets and ambiguous droplets to conservatively estimate singlets. Supplementary Fig. 8 illustrates the distribution of singlet, doublet likelihoods and the decision boundaries when t = 2 was used.

Theoretical expectation of deconvoluting singlets.

The theoretical distribution of expected singlets with multiplexing (presented in Supplementary Fig. 2) is as follows. Let d_o (e.g. 0.01) be the proportion of true multiplets when x_o (1,000) cells are loaded when multiplexing was not used. Then the expected multiplet rates when *x* cells are loaded can be modeled exponentially as $d(x) = 1 - (1-d_0)^{\frac{x}{x_0}}$. Let □ be the fraction of true singlets incorrectly classified as non-singlets (i.e. doublet or ambiguous), and β be the fraction of multiplets correctly classified as non-singlets. When multiplexing x cells equally from n

samples, the expected multiplet rates are $d(x)$, and $\frac{1}{n}d(x)$ are expected to be undetectable doublets mixed between the cells from the same sample. Therefore, the overall effective multiplet rate is $\left[\frac{n-(n-1)\beta}{n}\right]d(x)$. Similarly, the expected number of correctly identified singlets becomes $\frac{(1-\alpha)[1-d(x)]x_0 d(x)}{-\log(1-d_0)}$. Given α, β the expected number of singlets can be calculated by fixing the multiplet rate $d(x) = d_0$. We used $d_0 = 0.01, x_0 = 1000$ for the simulation in Supplementary Fig. 2.

Dependence of demultiplexing performance on experimental design parameters.

The demuxlet 'plp' option was used to generate a pileup format of 6,145 cells from one well of PBMC 10x data. The reads in the pileup were then modified to reflect the genotypes of individuals sampled from the 1000 Genomes Phase 3 cohort. The pileup was downsampled to obtain different numbers of read-overlapping exonic SNPs (ranging from 5,000 to 100,000) for the whole cohort. To create simulated doublets, we randomly sampled and merged pairs of barcodes within a dataset, resulting in a 5% doublet rate in the original data. For simulations with related individuals, we simulated transcriptomes from 8 individuals in 1000 Genomes with varying degrees of relatedness, ranging from unrelated to parent-child (HG00146, HG00147, HG00500, HG00501, HG00502, HG00512, HG00514, and HG00524).

Isolation and preparation of PBMC samples.

Informed consent was obtained from all patients sequenced in this study. Peripheral blood mononuclear cells were isolated from patient donors, Ficoll separated, and cryopreserved by the UCSF Core Immunologic Laboratory (CIL). PBMCs were thawed in a 37°C water bath, and subsequently washed and resuspended in EasySep buffer (STEMCELL Technologies). Cells

were treated with DNAseI and incubated for 15 min at RT before filtering through a 40um column. Finally, the cells were washed in EasySep and resuspended in 1x PBMS and 0.04% bovine serum albumin. Cells from 8 donors were then re-concentrated to 1M cells per mL and then serially pooled. At each pooling stage, 1M cells per mL were combined to result in a final sample pool with cells from all donors.

IFN-β stimulation and culture.

Prior to pooling, samples from 8 individuals were separated into two aliquots each. One aliquot of PBMCs was activated by 100 U/mL of recombinant IFN-β (PBL Assay Science) for 6 hrs according to the published protocol[26]. The second aliquot was left untreated. After 6 hrs, the 8 samples for each condition were pooled together in two final pools (stimulated cells and control cells) as described above.

Fluorescence-activated cell sorting and analysis.

1M PBMCs from each donor were stained using standard procedure (30 min, 4 C) with the following surface antibody panel (CD3-PerCP clone SK7 (BioLegend), CD4-APC clone OKT4 (BioLegend), CD8-BV570 clone RPA-T8 (BioLegend), CD14-FITC clone 63D3 (BioLegend), CD19-BV510 clone SJ25C1 (BD), and Ghost dye A710 viability stain (Tonbo)) (Life Sciences Reporting Summary). Samples were then analyzed and sorted using a BD FACSAria Fusion instrument at the UCSF flow cytometry core. To calculate cell type proportions, the number of events in each of $CD3^+$ $CD4^+$ $CD8^-$ ($CD4^+$ T cells), $CD3^+$ $CD4^-$ $CD8^+$ ($CD8^+$ T cells), $CD3^-$ $CD19^+$ (B cells), and $CD3^-$ $CD14^+$ (monocytes) were divided by the sum of events in these gates (Supplementary Fig. 21).

Quantitative polymerase chain reaction analysis.

RNA was isolated from sorted $CD4^+$ T cells following the RNeasy micro kit protocol (QIAGEN), and cDNA was prepared using MultiScribe Reverse Transcriptase (Applied Biosystems cat #4368814). The qPCR primers were chosen from the PrimerBank reference when available [50]. Each sample was run in triplicate with the Luminaris HiGreen qPCR kit (Thermo Scientific #K0992) according to standard protocol using a Roche Light Cycler 96 instrument and fold change was calculated from ΔΔCT between control and stimulated samples with GAPDH as a reference gene.

Droplet-based capture and sequencing.

Cellular suspensions were loaded onto the 10x Chromium instrument (10x Genomics) and sequenced as described in Zheng et al[17]. The cDNA libraries were sequenced using a custom program on 10 lanes of Illumina HiSeq2500 Rapid Mode, yielding 1.8B total reads and 25K reads per cell. At these depths, we recovered >90% of captured transcripts in each sequencing experiment.

Bulk isolation and sequencing.

PBMCs from lupus patients were isolated and prepared as described above. Once resuspended in EasySep buffer, the EasyEights Magnet was used to sequentially isolate $CD14^+$ (using the EasySep Human CD14 positive selection kit II, cat #17858), $CD19^+$ (using the EasySep Human CD19 positive selection kit II, cat #17854), $CD8^+$ (EasySep Human CD8 positive selection kitII, cat#17853), and $CD4^+$ cells (EasySep Human CD4 T cell negative isolation kit (cat #17952)

according to the kit protocol. RNA was extracted using the RNeasy Mini kit (#74104), and reverse transcription and tagmentation were conducted according to Picelli et al. using the SmartSeq2 protocol[51, 52]. After cDNA synthesis and tagmentation, the library was amplified with the Nextera XT DNA Sample Preparation Kit (#FC-131-1096) according to protocol, starting with 0.2ng of cDNA. Samples were then sequenced on one lane of the Illumina Hiseq4000 with paired end 100bp read length, yielding 350M total reads.

Alignment and initial processing of single cell sequencing data.

We used the CellRanger v1.1 and v1.2 software with the default settings to process the raw FASTQ files, align the sequencing reads to the hg19 transcriptome, and generate a filtered UMI expression profile for each cell[17]. The raw UMI counts from all cells and genes with nonzero counts across the population of cells were used to generate t-SNE profiles.

Cell type classification and clustering.

To identify known immune cell populations in PBMCs, we used the Seurat package to perform unbiased clustering on the 2.7k PBMCs from Zheng et al., following the publicly available Guided Clustering Tutorial[17, 53]. The FindAllMarkers function was then used to find the top 20 markers for each of the 8 identified cell types. Cluster averages were calculated by taking the average raw count across all cells of each cell type. For each cell, we calculated the Spearman correlation of the raw counts of the marker genes and the cluster averages, and assigned each cell to the cell type to which it had maximum correlation.

Differential expression analysis.

Demultiplexed individuals were used as replicates for differential expression analysis. For each gene, raw counts were summed for each individual. We used the DESeq2 package to detect differentially expressed genes between control and stimulated conditions[54]. Genes with baseMean > 1 were filtered out from the DESeq2 output, and the qvalue package was used to calculate FDR < 0.05 [55].

Estimation of inter-individual variability in PBMCs.

For each individual, we found the mean expression of each gene with nonzero counts. The mean was calculated from the log2 single cell UMI counts normalized to the median count for each cell. To measure inter-individual variability, we then calculated the variance of the mean expression across all individuals. Lin's concordance correlation coefficient was used to compare the agreement of observed data and synthetic replicates. Synthetic replicates were generated by sampling without replacement either from all cells or cells matched for cell type proportion. Cell type-specific variability estimated as the correlation between synthetic replicates was compared to variability estimates from 23 biological replicates of bulk IFN-stimulated monocyte-derived dendritic cells. Protein coding genes (407/414) originally measured using Nanostring (a hybridization based PCR-free quantification method) were assessed, and variability in the bulk dataset was estimated as repeatability using a linear mixed model[56,26].

Estimation of inter-individual variability within cell types.

For each cell type, we generated two bulk equivalent replicates for each individual by summing raw counts of cells sampled without replacement. We used DESeq2 to generate variance-

stabilized counts across all replicates. To filter for expressed genes, we performed all subsequent analyses on genes with 5% of samples with > 0 counts. The correlation of replicates was performed on the log2 normalized counts. Pearson correlation of the two replicates from each of the 8 individuals was used to find genes with significant inter-individual variability.

Quantitative trait mapping in major immune cell types.

Genotypes were imputed with EAGLE[57] and filtered for MAF > 0.2, resulting in a total of 189,322 SNPs. Cell type proportions were calculated as number of cells for each cell type divided by the number of total cells for each person. Linear regression was used to test associations between each genetic variant and cell-type proportion with the Matrix eQTL software[58]. Cis-eQTL mapping was conducted in each cell type separately. All genes with at least 50 UMI counts in 20% of the individuals in all PBMCs were tested for each cell type, resulting in a total of 4,555 genes. Variance-stabilized and log-normalized gene expression was calculated using the 'rlog' function of the DESeq2 package[54]. All variants within a window of 100kbp of each gene were tested with linear regression using Matrix eQTL[58]. Batch information for each sample as well as the first 3 principal components of the expression matrix were used as covariates.

Single cell and bulk RNA-sequencing data has been deposited in the Gene Expression Omnibus under the accession number GSE96583. Demuxlet software is freely available at https://github.com/statgen/demuxlet

Fig. 1

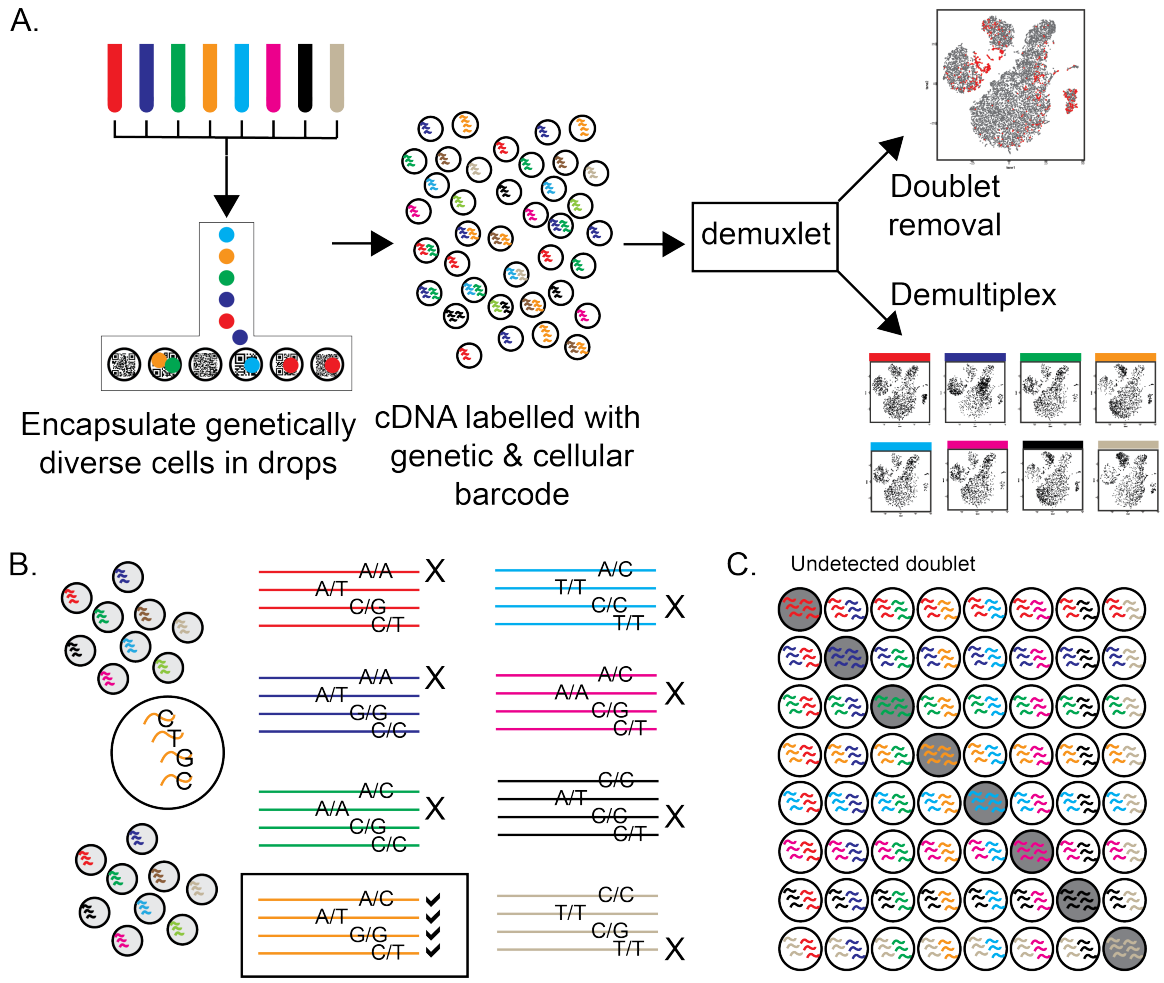

Figure 1 – Demuxlet: demultiplexing and doublet identification from single cell data. a) Pipeline for experimental multiplexing of unrelated individuals, loading onto droplet-based single-cell RNA-sequencing instrument, and computational demultiplexing (demux) and doublet removal using demuxlet. Assuming equal mixing of 8 individuals, b) 4 genetic variants can recover the sample identity of a cell, and c) 87.5% of doublets will contain cells from two different samples.

Figure 2

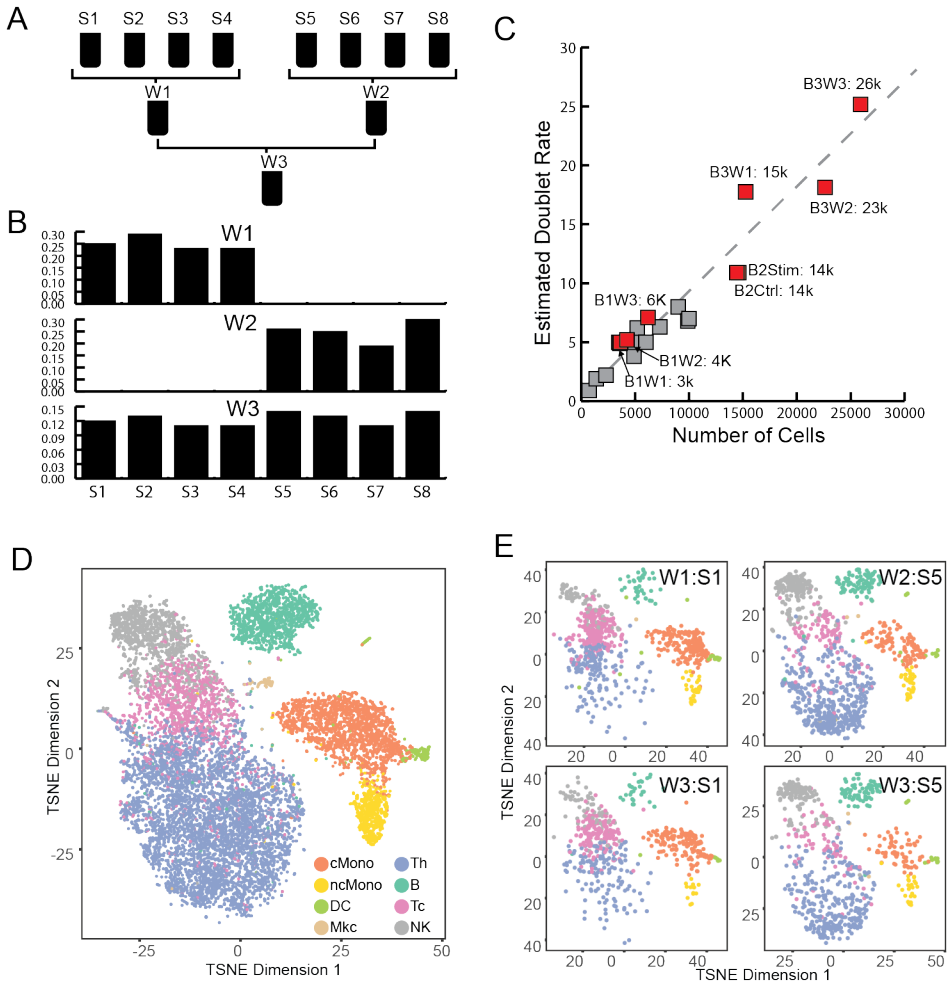

Figure 2 – Performance of demuxlet. a) Experimental design for equimolar pooling of cells from 8 unrelated samples (S1-S8) into three wells (W1-W3). W1 and W2 contain cells from two disjoint sets of 4 individuals. W3 contains cells from all 8 individuals. b) Demultiplexing single cells in each well recovers the expected individuals. c) Estimates of doublet rates versus previous estimates from mixed species experiments. d) Cell type identity determined by prediction to previously annotated PBMC data. e) t-SNE plot of two individuals (S1 and S5) from different wells are qualitatively concordant.

Figure 3

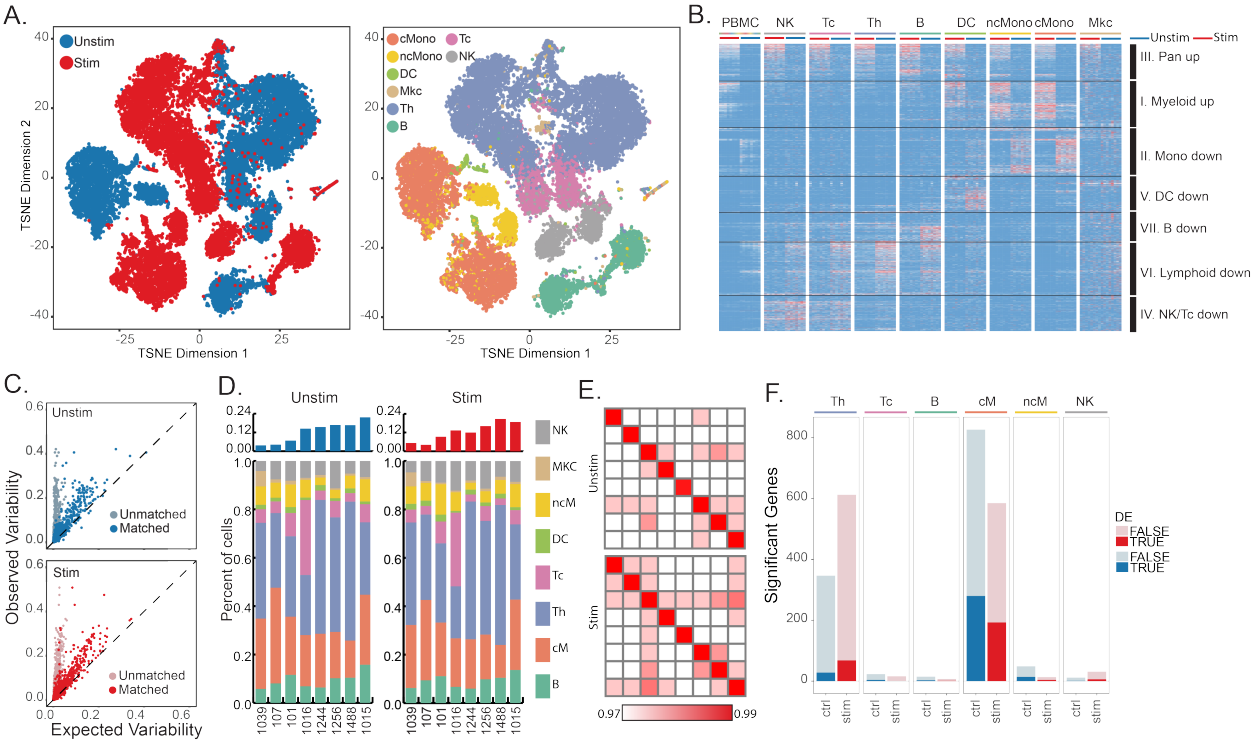

Figure 3 – Inter-individual variability in IFN-β response. a) t-SNE plot of unstimulated (blue) and IFN-β-stimulated (red) PBMCs and the estimated cell types. b) Cell type-specific expression in stimulated (left) and unstimulated (right) cells. Differentially expressed genes shown (FDR < 0.05, |log(FC)| > 1). Each column represents cell type-specific expression for each individual from demuxlet. c) Observed variance (y-axis) in mean expression over all PBMCs from each of the 8 individuals versus expected variance (x-axis) over synthetic replicates sampled across all cells (light blue, pink) or replicates matched for cell type proportion (blue, red). d) Cell type proportions for each individual in unstimulated and stimulated cells. e) Correlation between sample replicates in control and stimulated cells. f) Number of significantly variable genes in each cell type and condition.

Figure 4

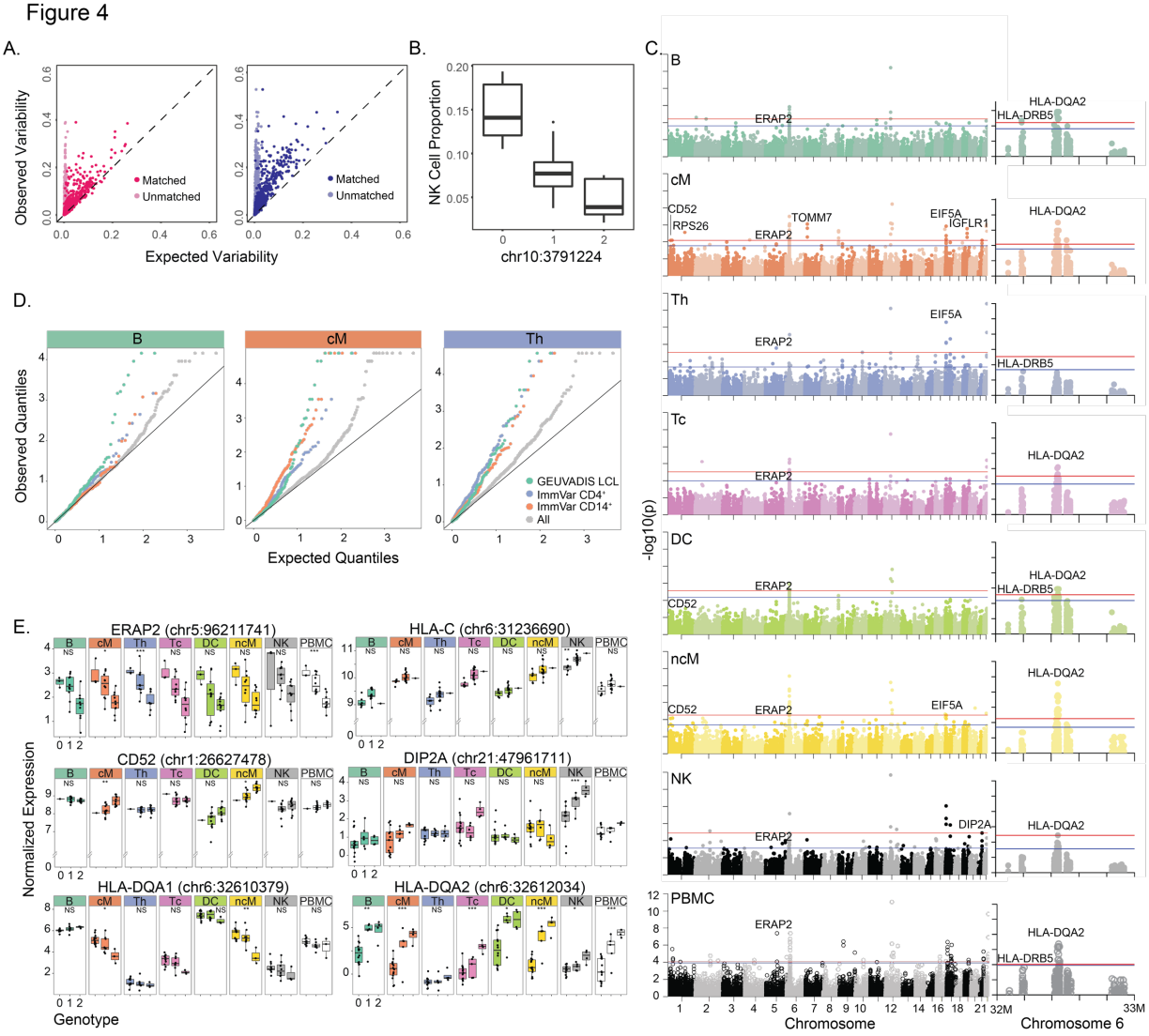

Figure 4 – Genetic control over cell type proportion and gene expression (N=23). a) Observed variance (y-axis) in mean expression over all PBMCs from each individual versus expected variance (x-axis) over synthetic replicates sampled across batch 1 (left, N=8) and batch 3 (right, N=15). b) Association of chr10:3791224 with NK cell type proportions. c) Genome-wide and chromosome 6 Manhattan plots across all major cell types. Horizontal lines correspond to FDR < 0.1 (blue) and FDR < 0.05 (red). d) Q-Q plots across all genes and subsets of previously published eQTLs in relevant cell types are shown for B, cM, and Th populations. e) Notable cis-eQTLs across all major immune cell types are marked with *(FDR < 0.25), ** (FDR < 0.1), and *** (FDR < 0.05). Lack of association is marked with NS (not significant).

References

1. Macosko, E.Z. et al. in Cell, Vol. 161 1202-1214 (2015).
2. Klein, A.M. et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201 (2015).
3. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145 (2015).
4. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
5. Streets, A.M. et al. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7048-7053 (2014).
6. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protocols* **12**, 44-73 (2017).
7. Zheng, G.X.Y. et al. in Nature Communications | doi:10.1038/ncomms9687, Vol. 8 14049 (Nature Publishing Group, 2017).
8. Jun, G. et al. in The American Journal of Human Genetics, Vol. 91 839-848 (2012).
9. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
10. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
11. The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

12. Aguirre-Gamboa, R. et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. *Cell Reports* **17**, 2474-2487.

13. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e1014 (2016).

14. Mostafavi, S. et al. Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164**, 564-578.

15. Stark, G.R., Kerr, I.M., Williams, B.R.G., Silverman, R.H. & Schreiber, R.D. in http://dx.doi.org/10.1146/annurev.biochem.67.1.227, Vol. 67 227-264 (Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 2003).

16. Lee, M.N. et al. in Science, Vol. 343 1246980-1246980 (2014).

17. Ye, C.J. et al. in Science, Vol. 345 1254665-1254665 (2014).

18. Andrés, A.M. et al. Balancing Selection Maintains a Form of ERAP2 that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. *PLOS Genetics* **6**, e1001157 (2010).

19. Mostafavi, S. et al. Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164**, 564-578 (2016).

20. Palmer, C., Diehn, M., Alizadeh, A.A. & Brown, P.O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 115 (2006).

21. Saveanu, L. et al. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat Immunol* **6**, 689-697 (2005).

22. Franco, L.M. et al. Integrative genomic analysis of the human immune response to influenza vaccination. *eLife* **2**, e00299 (2013).

23. Cao, J. et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv* (2017).

24. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016).

25. Adamson, B. et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e1821 (2016).

26. Jaitin, D.A. et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e1815 (2016).

27. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Meth* **14**, 297-301 (2017).

28. Farh, K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).

29. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotech* **33**, 155-160 (2015).

30. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* **7**, 39921 (2017).

31. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338 (2017).

32. Supplementary Code.

32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

33. Wang, X., Spandidos, A., Wang, H. & Seed, B. PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic Acids Research* **40**, D1144-D1149 (2012).

34. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth* **10**, 1096-1098 (2013).

35. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171-181 (2014).

36. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotech* **33**, 495-502 (2015).

37. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).

38. Dabney, A., Storey, J.D. & Warnes, G.R. qvalue: Q-value estimation for false discovery rate control. *R package version* **1** (2010).

39. Falconer, D.S., Mackay, T.F. & Frankham, R. Introduction to quantitative genetics (4th edn). *Trends in Genetics* **12**, 280 (1996).

40. Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-816 (2016).

41. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date 2018/06/15