Additional background and discussion:

T cells are defined by a surface T cell receptor that mediates recognition of pathogen-associated epitopes, generally via interactions with peptide-major histocompatibility complexes (pMHC). T cell receptors are generated by germline recombinase activating gene (RAG)-mediated rearrangements of the genomic TCR locus, a process termed V(D)J recombination. This process has the potential to generate a staggering diversity of TCRs, with estimates ranging from 10^{15} [1] to as high as 10^{61} [2] possible receptors that could be generated by recombination, although only a relatively small portion of these is thought to appear in any individual (~10^{6}-10^{8})[3–6]. In mammals, two types of TCRs are possible, αβ and γδ, and different species produce different ratios of cells bearing these receptors. In humans and mice, αβ T cells dominate, representing up to 90% of the T cell compartment.

The pool of T cells that recognizes a specific epitope expresses diverse T cell receptors. The size of these naïve precursor repertoires has been estimated for various epitopes by limiting dilution techniques and, more recently, by a tetramer-based magnetic enrichment approach, the latter of which finds pool sizes ranging between 50 and 500 naïve cells per epitope, on average[7–9]. Due to the rounds of expansion that T cells undergo in the thymus during development, it has been assumed that there are multiple naïve cells with identical TCRs. However, sequencing the naïve repertoire of epitope-specific responses in mice has instead shown that most naïve cells contain a unique receptor, with a very low rate of duplicates among cells[10–12].

Sequencing the T cell receptor requires identifying the specific V-region utilized by the α or β chain and obtaining the complete sequence of the hypervariable CDR3 region, the site of RAG-mediated V(D)J junctional diversity. Due to the availability of TCR Vβ staining reagents in the human and mouse, analyses of the repertoire initially focused solely on the TCRβ chain. Subsequently, two broad approaches to sequencing the TCR repertoire emerged: single-cell based methods that permit direct pairing of the α and β chains[13–16], and deep sequencing-based methods that amplify single chains from pools of cells[5,17,18] where pairing can be achieved through specific sort conditions and algorithmic imputation[19]. Using these two methods, both of which usually focus on the CDR3s, significant amounts of data from bulk and epitope-specific populations of naïve, activated, and memory T cells have been published and are being actively collected. This work has demonstrated that the recruited TCR repertoire has a direct impact on several features of the immune response, including its size, efficacy, and memory potential[5,6,8,20–27]. Furthermore, it has become clear that the epitope-specific immune repertoire typically contains receptors that are overrepresented relative to their representation in the naïve pool[10,28,29].

Another striking feature of epitope-specific immune repertoires is that, despite the vast potential size of the naïve repertoire, the same receptors (at least at single chain resolution) are frequently identified in multiple individuals in both mice and humans (termed “public” receptors). Public TCRs represent an extreme form of bias in V(D)J recombination known as type III bias[30]. Many of the identified public receptors are similar to germline sequences and therefore can be generated without multiple insertions or deletions and/or by multiple theoretical recombination

mechanisms[31,32]. Importantly, most of this work has targeted β chain sequences, with little characterization of public paired αβ sequences. Less characterized are receptors that, despite differences in amino acid sequence, share TCR motifs and/or exhibit relative enrichment of certain V- and J-regions[33–36]. Despite these characterizations of TCR overrepresentation both within and among individuals, there is little known about the selection processes that lead to such preferences.

Here we present a comprehensive analysis of ten epitope-specific TCR repertoires, complete with paired α and β chain sequences obtained using a single-cell, PCR-based approach[13,37]. In addition to extensively characterizing these TCR repertoires, we also identify the conserved features of TCRs that convey epitope-specificity in both humans and mice. This dataset of seven mouse and three human epitope-specific responses encompasses over 4600 in-frame, paired, single-cell amplified TCR sequences from 78 mice and 32 humans. In order to identify key parameters that characterize receptors that bind to the same epitope, we report the development of a comprehensive repertoire analysis framework with multiple unique features that can be applied to any collection of TCR repertoire data. Although several TCR repertoire analysis tools have been published, many of them are directed towards processing sequence data[38–41], with only a few addressing the post processing analysis of the repertoire, such as V(D)J usage, sharing, diversity, and spectratyping[42–44]. Our analysis pipeline includes several advances over these existing tools: (i) TCRdist - a simple and intuitive metric defining the distance between any two receptors, which can be used for TCR clustering and visualizing TCR landscapes via dimensionality reduction; (ii) a CDR3 motif discovery algorithm corrected for biases of the gene rearrangement process; (iii) TCRdiv - a robust repertoire diversity measure that generalizes Simpson's Diversity Index by weighting diversity based on sequence similarity (as measured by TCRdist) rather than clonotypic identity; (iv) TCRdist nearest-neighbor classifier capable of accurately discriminating receptors specific for a particular epitope from a background naive repertoire and from other epitope-specific responses, which we demonstrate with an independent validation data set; and (v) novel visualizations and information-theoretic measures of gene usage and covariation, CDR3 sequence patterns, and repertoire structure. By utilizing sequence data to facilitate predictions of which TCRs are epitope-specific, these analyses provide important insights into the analysis of polyclonal repertoires (such as tumor infiltrating lymphocytes), where clusters of related receptors with unknown specificities need to be identified and segregated to identify relevant antigenic targets.

By analyzing the landscape of ten different epitope-specific repertoires using single-cell, paired TCRαβ sequencing, we have quantified multiple core features of adaptive immune recognition. All repertoires contained at least one cluster of motif-associated receptors of varying size and proportion to the rest of the repertoire, along with a population of dispersed, outlier receptors of varying size and diversity. The development of a novel distance metric, TCRdist, allowed us to compare repertoire sequences in a manner that facilitated the visualization of recognition landscapes, the quantification of defining motifs, and the definition of a sampling score that could be used to identify and discriminate novel epitope-specific receptors with robust sensitivity and specificity.

One immediate application of these findings is the analysis of the abundance of mixed repertoire data being actively generated in clinical settings where the number or identity of the antigen-specific targets is unknown. Tumor-infiltrating lymphocytes (TILs) can be isolated from solid tumors and sequenced, but the targets of those T cells have in the past proven exceptionally difficult to identify[33,45–47]. Our analyses provide a way of grouping related receptors and selecting representative members of these clusters for experimental interrogation of specificity. For example, after sequencing a group of TILs, TCRdist could be employed to identify multiple clusters of related receptors. If epitope-specificity to a tumor-associated antigen were assigned to one member of the cluster, it would be likely that other members of the cluster will also respond to the same antigen; the identification of these putative associations would thus permit the testing of a small number of representative receptors to cover the range of TCR reactivities among the TILs.

Considered as a whole, these analyses furthermore demonstrate the necessity of paired TCRαβ data for properly characterizing a TCR repertoire. The small numbers of paired public receptors observed across individuals similarly emphasize the need for continued focus on obtaining paired data, as it appears unlikely that specific alpha-beta pairs will be consistently maintained across individuals at the same rates as public single chains. As single-cell paired techniques[13,14] algorithmic pairing techniques from single-chain NGS sequencing[19,48] have advanced, paired methods have become more tractable. Additionally, with the advent of engineered cell based therapies, the ability to isolate complete receptors against pathogen or tumor-associated antigens will be necessary for rapid therapeutic application[49,50].

A study by Birnbaum et al. examined the preference of an individual TCR for peptide antigen using an unbiased phage display approach to screen for targets. The authors concluded that the antigen originally used to isolate the receptor was a dominant target of this receptor, as the screen largely produced similar antigens[51]. It would be interesting to determine where this receptor falls in the distribution of the associated repertoire; we might hypothesize that it should fall within the main cluster of receptors as a representative of a dominant part of the response to this epitope.

The dispersed receptors that fail to cluster in each repertoire are an interesting area for further exploration. This organization of receptors into motif-sharing clusters and dispersed outliers suggests that this latter group may represent an alternative means of recognizing the antigen. We hypothesize that this large diversity of receptor types, even if not used at high levels, may be an evolutionary development to prevent epitope escape, as the extremely diverse sequences indicate unique means of peptide recognition. Structural studies will be useful in determining whether clustered and dispersed receptors employ distinct binding modalities to recognize the same epitope, as has recently been shown for some receptors present in the naive repertoire that fail to expand efficiently after antigen stimulation[52]. Similarly, close examination of cross-reactive responses will allow us to determine from which part of the repertoire such cross-reactivity is prone to arise (i.e., the clustered or the dispersed TCRs), which will in turn help test the hypothesis that this mode of repertoire selection serves to prevent epitope escape. These types of studies may also be useful for testing theoretical models of the fundamental principles of adaptive immunity[53].

We have described the general features of epitope-specific T cell repertoires and quantified the key elements that determine membership within a particular response. By parameterizing the elements of immune repertoires of known specificities consistently associated with a particular reactivity, we are closer to building a general model of TCR:pMHC recognition. While such a model will require more data and further algorithmic development, the implications of predicting epitope-specificity based solely on paired TCR sequences are substantial. In addition to providing insight into one of the key interfaces of the immune system, the clinical applications include identifying promising targets in cancer immunotherapy[54], rapid identification of vaccine target antigens, and assessment of efficacy after vaccination[55].

1. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334,** 395–402 (1988).
2. Mora, T. & Walczak, A. M. Quantifying lymphocyte receptor diversity. *bioRxiv* 046870 (2016). doi:10.1101/046870
3. Casrouge, A. *et al.* Size Estimate of the TCR Repertoire of Naive Mouse Splenocytes. *The Journal of Immunology* **164,** 5782–5787 (2000).
4. Arstila, T. P. *et al.* A direct estimate of the human alphabeta T cell receptor diversity. *Science* **286,** 958–961 (1999).
5. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114,** 4099–4107 (2009).
6. Warren, R. L. *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* **21,** 790–797 (2011).
7. Obar, J. J., Khanna, K. M. & Lefrançois, L. Endogenous naive CD8+ T cell precursor frequency regulates primary and memory responses to infection. *Immunity* **28,** 859–869 (2008).
8. La Gruta, N. L. *et al.* Primary CTL response magnitude in mice is determined by the extent of naive T cell recruitment and subsequent clonal expansion. *J. Clin. Invest.* **120,** 1885–

1894 (2010).

9. Moon, J. J. *et al.* Naive CD4 T Cell Frequency Varies for Different Epitopes and Predicts Repertoire Diversity and Response Magnitude. *Immunity* **27,** 203–213 (2007).
10. Thomas, P. G., Handel, A., Doherty, P. C. & La Gruta, N. L. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naive and immune T-cell populations. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 1839–1844 (2013).
11. Cukalac, T. *et al.* Paired TCRαβ analysis of virus-specific CD8(+) T cells exposes diversity in a previously defined 'narrow' repertoire. *Immunol. Cell Biol.* **93,** 804–814 (2015).
12. Lythe, G., Callard, R. E., Hoare, R. L. & Molina-París, C. How many TCR clonotypes does a body maintain? *J. Theor. Biol.* **389,** 214–224 (2016).
13. Dash, P. *et al.* Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J. Clin. Invest.* **121,** 288–295 (2011).
14. Wang, G. C., Dash, P., McCullers, J. A., Doherty, P. C. & Thomas, P. G. T cell receptor αβ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Sci. Transl. Med.* **4,** 128ra42 (2012).
15. Kim, S.-M. *et al.* Analysis of the paired TCR α- and β-chains of single human T cells. *PLoS One* **7,** e37338 (2012).
16. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32,** 684–692 (2014).
17. Weinstein, J. A., Jiang, N., White, R. A., 3rd, Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324,** 807–810 (2009).
18. Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. & Holt, R. A. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19,** 1817–1824 (2009).
19. Howie, B. *et al.* High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* **7,** 301ra131 (2015).

20. Tubo, N. J. *et al.* Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell* **153,** 785–796 (2013).
21. La Gruta, N. L. *et al.* A virus-specific CD8 T cell immunodominance hierarchy determined by antigen dose and precursor frequencies. *Proceedings of the National Academy of Sciences* **103,** 994–999 (2006).
22. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13139–13144 (2014).
23. Mamedov, I. Z. *et al.* Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.* **3,** 201–207 (2011).
24. Heather, J. M. *et al.* Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and following Antiretroviral Therapy. *Front. Immunol.* **6,** 644 (2015).
25. Klarenbeek, P. L. *et al.* Human T-cell memory consists mainly of unexpanded clones. *Immunol. Lett.* **133,** 42–48 (2010).
26. Kim, C., Wilson, T., Fischer, K. F. & Williams, M. A. Sustained interactions between T cell receptors and antigens promote the differentiation of $CD4^+$ memory T cells. *Immunity* **39,** 508–520 (2013).
27. Britanova, O. V. *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192,** 2689–2698 (2014).
28. Gerlach, C. *et al.* Heterogeneous Differentiation Patterns of Individual CD8 T Cells. *Science* **340,** 635–639 (2013).
29. Neller, M. A. *et al.* Naive $CD8^+$ T-cell precursors display structured TCR repertoires and composite antigen-driven selection dynamics. *Immunol. Cell Biol.* **93,** 625–633 (2015).
30. Turner, S. J., Doherty, P. C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6,** 883–894 (2006).
31. Li, H. *et al.* Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. *J. Immunol.* **189,** 2404–2413 (2012).

32. Venturi, V. *et al.* Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 18691–18696 (2006).
33. Li, B. *et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* **48,** 725–732 (2016).
34. Serana, F. *et al.* Identification of a public CDR3 motif and a biased utilization of T-cell receptor V beta and J beta chains in HLA-A2/Melan-A-specific T-cell clonotypes of melanoma patients. *J. Transl. Med.* **7,** 21 (2009).
35. Yang, J. *et al.* Profiling the repertoire of T-cell receptor beta-chain variable genes in peripheral blood lymphocytes from subjects who have recovered from acute hepatitis B virus infection. *Cellular and Molecular Immunology* **11,** 343–354 (2014).
36. Billam, P. *et al.* T Cell receptor clonotype influences epitope hierarchy in the CD8+ T cell response to respiratory syncytial virus infection. *J. Biol. Chem.* **286,** 4829–4841 (2011).
37. Dash, P., Wang, G. C. & Thomas, P. G. Single-Cell Analysis of T-Cell Receptor αβ Repertoire. *Methods Mol. Biol.* **1343,** 181–197 (2015).
38. Giraud, M. *et al.* Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics* **15,** 409 (2014).
39. Alamyar, E., Giudicelli, V., Li, S. & Duroux, P. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunomethods* (2012).
40. Bolotin, D. A. *et al.* MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* **10,** 813–814 (2013).
41. Gerritsen, B., Pandit, A., Andeweg, A. C. & de Boer, R. J. RTCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data. *Bioinformatics* **32,** 3098–3106 (2016).
42. Nazarov, V. I. *et al.* tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* **16,** 175 (2015).

43. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput. Biol.* **11,** e1004503 (2015).
44. Bagaev, D. V. *et al.* VDJviz: a versatile browser for immunogenomics data. *BMC Genomics* **17,** 453 (2016).
45. Parkhurst, M. R. *et al.* Isolation of T cell receptors specifically reactive with mutated tumor associated antigens from tumor infiltrating lymphocytes based on CD137 expression. *Clin. Cancer Res.* (2016). doi:10.1158/1078-0432.CCR-16-2680
46. Pasetto, A. *et al.* Tumor- and Neoantigen-Reactive T-cell Receptors Can Be Identified Based on Their Frequency in Fresh Tumor. *Cancer Immunol Res* **4,** 734–743 (2016).
47. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **344,** 641–645 (2014).
48. Lee, E. S., Thomas, P. G., Mold, J. E. & Yates, A. J. Identifying T Cell Receptors from High-Throughput Sequencing: Dealing with Promiscuity in TCRα and TCRβ Pairing. *PLoS Comput. Biol.* **13,** e1005313 (2017).
49. Barrett, A. J. & Bollard, C. M. The coming of age of adoptive T-cell therapy for viral infection after stem cell transplantation. *Ann Transl Med* **3,** 62 (2015).
50. Restifo, N. P., Dudley, M. E. & Rosenberg, S. A. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nat. Rev. Immunol.* **12,** 269–281 (2012).
51. Birnbaum, M. E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157,** 1073–1087 (2014).
52. Gras, S. *et al.* Reversed T Cell Receptor Docking on a Major Histocompatibility Class I Complex Limits Involvement in the Immune Response. *Immunity* **45,** 749–760 (2016).
53. Mayer, A., Balasubramanian, V., Mora, T. & Walczak, A. M. How a well-adapted immune system is organized. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 5950–5955 (2015).
54. Yang, J. C. & Rosenberg, S. A. Adoptive T-Cell Therapy for Cancer. *Adv. Immunol.* **130,** 279–294 (2016).

55. Furman, D. & Davis, M. M. New approaches to understanding the immune response to vaccination and infection. *Vaccine* **33,** 5271–5281 (2015).