Title: Profiling adaptive immune repertoires across multiple human tissues by RNA Sequencing

Authors: Serghei Mangul*[#1,2], Igor Mandric*[3], Harry Taegyun Yang[1], Nicolas Strauli[4], Dennis Montoya[5], Jeremy Rotman[1], Will Van Der Wey[1], Jiem R. Ronas[6], Benjamin Statz[1], Douglas Yao[5], Alex Zelikovsky[3], Roberto Spreafico[2], Sagiv Shifman**[7], Noah Zaitlen**[8], Maura Rossetti**[9], K. Mark Ansel**[10], Eleazar Eskin**[#1,11]

Affiliations:

[1]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA

[2]Institute for Quantitative and Computational Biosciences**,** University of California Los Angeles, Los Angeles, CA, USA

[3]Department of Computer Science, Georgia State University, Atlanta, USA

[4]Biomedical Sciences Graduate Program, University of California, San Francisco, CA, USA

[5]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA, USA

[6]Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA

[7]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

[8]Department of Medicine, University of California, San Francisco, CA, USA

[9]Immunogenetics Center, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

[10]Department of Microbiology and Immunology, Sandler Asthma Basic Research Center, University of California, San Francisco, San Francisco, CA, USA

[11]Department of Human Genetics, University of California Los Angeles, Los Angeles, USA

Correspondence to: Serghei Mangul smangul@ucla.edu and Eleazar Eskin eeskin@cs.ucla.edu

*Equal contribution

**Equal contribution

Abstract

Assay-based approaches provide a detailed view of the adaptive immune system by profiling T and B cell receptor repertoires. However, these methods carry a high cost and lack the scale of standard RNA sequencing (RNA-Seq). Here we report the development of ImReP, a novel computational method for rapid and accurate profiling of the adaptive immune repertoire from regular RNA-Seq data. We applied our novel method to 8,555 samples across 544 individuals from 53 tissues from the Genotype-Tissue Expression (GTEx v6) project. ImReP is able to efficiently extract TCR- and BCR-derived reads from RNA-Seq data. ImReP can also accurately assemble the complementary determining regions 3 (CDR3s), the most variable regions of B and T cell receptors, and determine their antigen specificity. Using ImReP, we have created a systematic atlas of immunological sequences for B and T cell repertoires across a broad range of tissue types, most of which have not been studied for B and T cell receptor repertoires. We also compared the GTEx tissues to track the flow of T- and B-clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites, and we examined the compositional similarities of clonal populations between these tissues. The atlas of T and B cell receptors, freely available at https://sergheimangul.wordpress.com/atlas-immune-repertoires/, is the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future immunology studies and advance development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/ .

Introduction

A key function of the adaptive immune system, which is composed of B-cells and T-cells, is to mount protective memory responses to a given antigen. B and T cells recognize their specific antigens through their surface antigen receptors (B and T cell receptors, BCR and TCR, respectively), which are unique to each cell and its progeny. BCR and TCR are diversified through somatic recombination, a process that randomly combines variable (V), diversity (D), and joining (J) gene segments, and inserts or deletes non-templated bases at the recombination junctions[1] (Figure 1a). The resulting DNA sequences are then translated into the antigen receptor proteins. This process allows for an astonishing diversity of the lymphocyte repertoire (i.e., the collection of antigen receptors of a given individual), with $>10^{13}$ theoretically possible distinct immunological receptors[1]. This diversity is key for the immune system to confer protection against a wide variety of potential pathogens[2]. In addition, upon activation of a B-cell, somatic hypermutation further diversifies BCRs in their variable region. These changes are mostly single-base substitutions occurring at extremely high rates (10^{-5} to 10^{-3} mutations per base pair per generation)[3]. Isotype switching is another mechanism that contributes to B-cell functional diversity. Here, antigen specificity remains unchanged while the heavy chain VDJ regions join with different constant (C) regions, such as IgG, IgA, or IgE isotypes, and alter the immunological properties of a BCR.

High-throughput technologies enable unprecedented accuracy when profiling the BCR and TCR repertoires. Commonly used assay-based approaches provide a detailed view of the adaptive

immune system with deep sequencing of amplified DNA or RNA from the variable region of BCR or TCR loci (Rep-Seq)[4]. Those technologies are usually restricted to one chain, with the majority of studies focusing on the beta chain of TCRs and the heavy chain of BCRs. Recent studies[2] successfully applied assay-based approaches to characterize the immune repertoire of the peripheral blood. However, little is known about the immunological repertoires of other human tissues, including barrier tissues like skin and mucosae. Studies involving assay-based protocols usually have small sample sizes, thus limiting analysis of intra-individual variation of immunological receptors across diverse human tissues.

RNA Sequencing (RNA-Seq) traditionally uses the reads mapped onto human genome references to study the transcriptional landscape of both single cells and entire cellular populations. In contrast to assay-based protocols that produce reads from the amplified variable region of BCR or TCR loci, RNA-Seq is able to capture the entire cellular population of the sample, including B and T cells. However, due to the repetitive nature of loci encoding for BCRs and TCRs, as well as the extreme level of diversity in BCR and TCR transcripts, most mapping tools are ill equipped to handle immune repertoire sequences. Despite this, BCR and TCR transcripts often occur in sufficient numbers within the transcriptome of many tissues to characterize their respective immunological repertoires[5].

In this study, we developed ImReP, a novel computational method for rapid and accurate profiling of the adaptive immune repertoire from regular RNA-Seq data. We applied it to 8,555 samples across 544 individuals from 53 tissues obtained from Genotype-Tissue Expression study

(GTEx v6)[6]. The data was derived from 38 solid organ tissues, 11 brain subregions, whole blood, and three cell lines. ImReP is able to efficiently extract TCR- and BCR- derived reads from the RNA-Seq data and accurately assemble the complementarity determining regions 3 (CDR3s). CDR3 are the most variable regions of B and T cell receptors and determine the antigen specificity. Using ImReP, we created a systematic atlas of immunological sequences for B- and T-cell repertories across a broad range of tissue types, most of which were not previously studied for B- and T-cell repertoires. We also examined the compositional similarities of clonal populations between the tissues to track the flow of T and B clonotypes across immune-related tissues, including secondary lymphoid and organs that encompass mucosal, exocrine, and endocrine sites. Our proposed approach is not superior in comparison to targeted TCR or BCR; rather, it provides a useful tool for mining large-scale RNA-Seq datasets for study of adaptive immune repertoires.

Results

ImReP: a two stage approach for adaptive immune repertoires reconstruction

We applied ImReP to 0.6 trillion RNA-Seq reads (92 Tbp) from 8,555 samples to assemble CDR3 sequences of B and T cell receptors (Table S1). The RNA-Seq data was generated by the Genotype-Tissue Expression Consortium (GTEx v6). First, we mapped RNA-Seq reads to the human reference genome using a short-read aligner (performed by GTEx consortium[6]) (Figure 1). Next, we identify reads spanning the V(D)J junction of B and T cell receptors and assemble clonotypes (a group of clones with identical CDR3 amino acid sequences). Here ImReP used 0.02 trillion high quality reads that successfully mapped to BCR genes, successfully mapped to TCR genes, or were unmapped reads that failed to map to the human reference genome (Figures 1a and S1).

ImReP is a two-stage approach to assemble CDR3 sequences and detect corresponding V(D)J recombinations (Figure 1b). In the first stage, ImReP utilizes reads that simultaneously overlap V and J gene segments to infer the CDR3 sequences. We define the CDR3 as the sequence of amino acids between the cysteine on the right of the junction and phenylalanine (for all TCR chains and immunoglobulin light chains) or tryptophan (for IGH) on the left of the junction. In the second stage, ImReP utilizes reads that overlap a single gene segment containing a partial CDR3 sequence. ImReP then uses a suffix tree to perform pairwise comparison of the reads and join the reads based on overlap in the CDR3 region. Further, ImReP uses a CAST clustering technique[7] to accurately assemble clonotypes for PCR and sequencing errors. We map D genes (for IGH, TCRB, and TCRG) onto assembled CDR3 sequences and infer corresponding V(D)J recombination. A detailed description of the methodology implemented with ImReP is provided in the Extended

Experimental Procedures Section. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/.

To validate the feasibility of using RNA-Seq to study the adaptive immune repertoire, we simulated RNA-Seq data as a mixture of transcriptomic reads and reads derived from BCR and TCR transcripts (Figure S3). BCR and TCR transcripts are simulated based on random recombination of V and J gene segments (obtained from IMGT database[8]) with non-template insertion at the recombination junction (Figure S2). We assessed the ability of ImReP to extract CDR3-derived reads from the RNA-Seq mixture by applying ImReP to a simulated RNA-Seq mixture. While our simulation approach may not completely summarize the various nuances and eccentricities of actual immune repertoires, it allows us to assess the accuracy of our tool. ImReP is able to identify 99% of CDR3-derived reads from the RNA-Seq mixture, suggesting it is a powerful tool for profiling RNA-Seq samples of immune-related tissues. Details about the simulation data are provided in the Extended Experimental Procedures section.

Next, we compared ImReP with other methods designed to assemble immune repertoires. We also investigated the sequencing depth and read length required to reliably assemble TCR and BCR sequences from RNA-Seq data. Our simulations suggest that both read length and sequencing depth have a major impact on precision-recall rates of CDR3 sequence assembly. ImReP is able to maintain an 80% precision rate for the majority of simulated scenarios. Average CDR3 coverage that is higher than 8 allows ImReP to archive a recall rate close to 90% for a read length above 75bp (Figure 2a). Increasing coverage has a positive effect on the number of

assembled clonotypes achieved by ImReP for both B and T cell receptors. In general, we observe higher precision-recall rates of CDR3 sequence assembly for TCRs in comparison to BCRs (Figure 2a-b).

We compared the performance of ImReP to MiXCR (RNA-Seq mode)[9], TRUST[10], TraCeR[11], V'DJer[12], IgBlast-based pipeline[13], and iSSAKE[14]. These tools were developed to assemble the hypervariable sequences in the T and B cell receptors directly from RNA-Seq data. We supplied each of those tools with the original RNA-Seq reads as raw or mapped reads, depending on the software developers' recommendations.). TRUST and TraCeR do not support the analysis of BCR sequences and were excluded from the comparison based for the IGH data. iSSAKE is no longer supported and was not recommended for use. Unfortunately, we obtained empty output after running V'DJer, and increasing coverage in the simulated data did not solve the problem. Alternative approaches, such as IMSEQ[15], cannot be applied directly to RNA-Seq reads because they were originally designed for targeted sequencing of B or T cell receptor loci. Thus, to independently assess and compare accuracy with ImReP, we only ran IMSEQ with the simulated reads derived from BCR or TCR transcripts (Figure S1). Scripts and commands to run all tools used in this study are provided in the Extended Experimental Procedures and are available online at https://github.com/smangul1/Profiling-adaptive-immune-repertoires-across-multiple-human-tissues-by-RNA-Sequencing. ImReP consistently outperformed existing methods on IGH data in both recall and precision rates for the majority of simulated parameters. ImReP and MiXCR show similar performance on TCRA data and outperform other methods. Notably, ImReP was the only

method with acceptable performance on IGH data at 50bp read length, reconstructing with a higher precision rate significantly more CDR3 clonotypes than other methods.

To demonstrate the feasibility of applying non-specific RNA Sequencing to assemble immune repertoire sequences, we used the TCRB-Seq data prepared from three samples of kidney renal clear cell carcinoma (KIRC) by Li, Bo, et al. [10]. We downloaded matching RNA-Seq samples from the TCGA portal. In total, we obtained 301 million 2x50bp reads from three RNA-Seq samples. First, we prepared the CDR3 sequences obtained from TCRB-Seq and considered only complete CDR3s, which we defined as a sequence of amino acids starting with cysteine (C) and ending with phenylalanine (F). We considered the prepared, complete CDR3s obtained from TCRB-Seq as total immune repertoire.

We used ImReP, MIXCR, TRUST, and IMSEQ to assemble CDR3s of the TCRB chain. We excluded V'DJer, because it only supports immunoglobulin chains and is not suitable to assemble CDR3s from T cell receptors. ImReP, MIXCR, and TRUST assembled comparable numbers of complete CDR3s that fully match CDR3s from the total immune repertoire obtained by TCRB-Seq; IMSEQ assembled none of the CDR3 sequences (Figure 2c and Figure S4). On average, 54% of CDR3s assembled by ImReP fully match CDR3s from TCRB-Seq. Our method was able to recover 0.1-0.9% of the total immune repertoire obtained by TCRB-Seq from kidney tissue. Other tissues, including spleen and whole blood, contain a higher fraction of T cells and allow RNA-Seq to capture a higher fraction of the total immune repertoire. One should note, the number of complete CDR3s fully matching CDR3s obtained by TCRB-Seq in our study (reported in Figure 2c

and Figure S4) are not fully comparable with the results reported in Li, Bo, et al. [10], where CDR3 sequences are considered to match CDR3s from TCRB-Seq if at least 6 amino acids are matched. Scripts and commands utilized to process the data and run repertoire assembly tools are provided in Extended Experimental Procedures and are available online at https://github.com/smangul1/Profiling-adaptive-immune-repertoires-across-multiple-human-tissues-by-RNA-Sequencing.

We further validated the ability of ImReP to accurately infer the proportion of immune cells in the sampled tissue. We hypothesized that the fraction of B- and T-cells in the sample will be proportional with the fraction of receptor-derived reads in our RNA-Seq data. We used a transcriptome-based computational method, SaVant[16], which uses cell-specific gene signatures (independent of BCR or TCR transcripts) to infer the relative abundance of B or T cells within each tissue sample. We found that B and T cell signatures inferred by SaVant showed positive correlation with the amount of BCR ($r = 0.77$, $P < 0.001$) or TCR ($r = 0.86$, $P < 0.001$) transcripts, respectively (Figure 2c,d). An exception to this correlation was for tissues that contain the highest density of B or T cells: spleen, whole blood, small intestine (terminal ileum), lung, and EBV-transformed lymphocytes (LCLs).

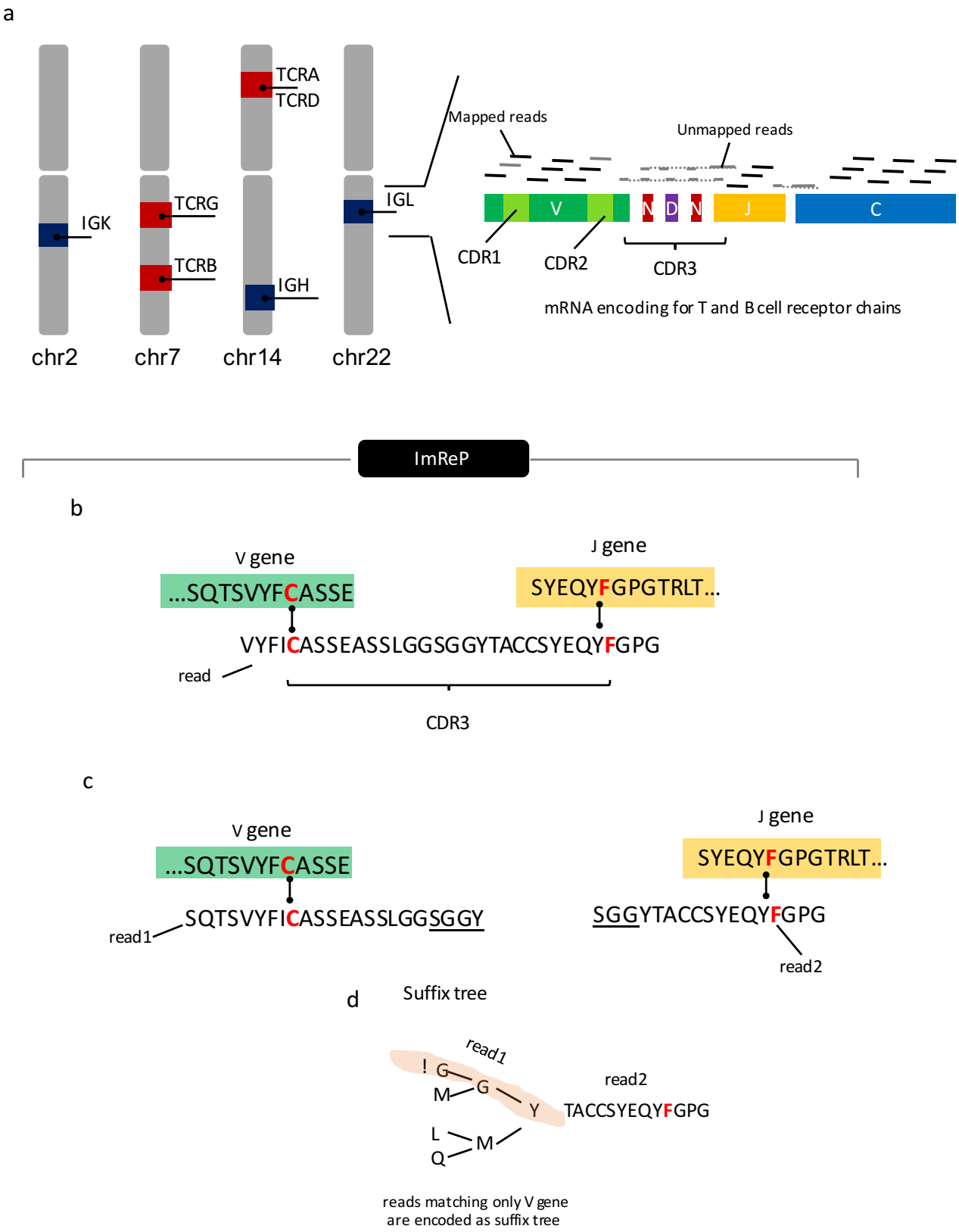


Figure 1. Overview of ImReP. (a) Schematic representation of human adaptive immune repertoire. Adaptive immune repertoire consists of four T-cell receptor loci (blue color, T cell receptor alpha locus (TCRA); T-cell receptor beta locus (TCRB); T-cell receptor delta locus (TCRD); and T-cell receptor gamma locus (TCRG)) and three immunoglobulin loci (red color, Immunoglobulin heavy locus **(**IGH); Immunoglobulin kappa locus (IGK); Immunoglobulin lambda locus (IGL). Alternative name – BCR, B cell receptor). B- and T-cell receptors contain multiple variable (V, green color), diversity (D, present only in IGH, TCRB, TCRG, violet color), joining (J, yellow color) and constant (C, blue color) gene segments. V(D)J gene segments are randomly jointed and non-templated bases (N, dark red color) are inserted at the recombination junctions. The resulting spliced T- or B-cell repertoire transcript incorporates the C segment and is translated into the antigen receptor proteins. RNA-Seq reads are derived from the rearranged immunoglobulin IG and TCR loci. Reads entirely aligned to genes of B- and T-cell receptors are inferred from mapped reads (black color). Reads with extensive somatic hypermutations and reads spanning the V(D)J recombination are inferred from the unmapped reads (grey color). Complementarity determining region 3 (CDR3) is the most variable region of the three CDR regions and is used to identify T/B-cell receptor clonotypes—a group of clones with identical CDR3 amino acid sequences. **(b)** Receptor derived reads spanning V(D)J recombinations are identified from unmapped reads and assembled into the CDR3 sequences. We first scan the amino acid sequences of the read and determine the putative CDR3 boundaries defined by last conserved cysteine encoded by the V gene and the conserved phenylalanine (for TCR) or tryptophan (for BCR) of J gene. Given the putative CDR3 boundaries, we check the prefix and suffix of the read to match the suffix of V and prefix of J genes, respectively. **(c-d)** In case a read overlaps with only the V or J gene, we perform the second stage of ImReP to match such reads based on the overlap of CDR3 sequence using suffix tree.

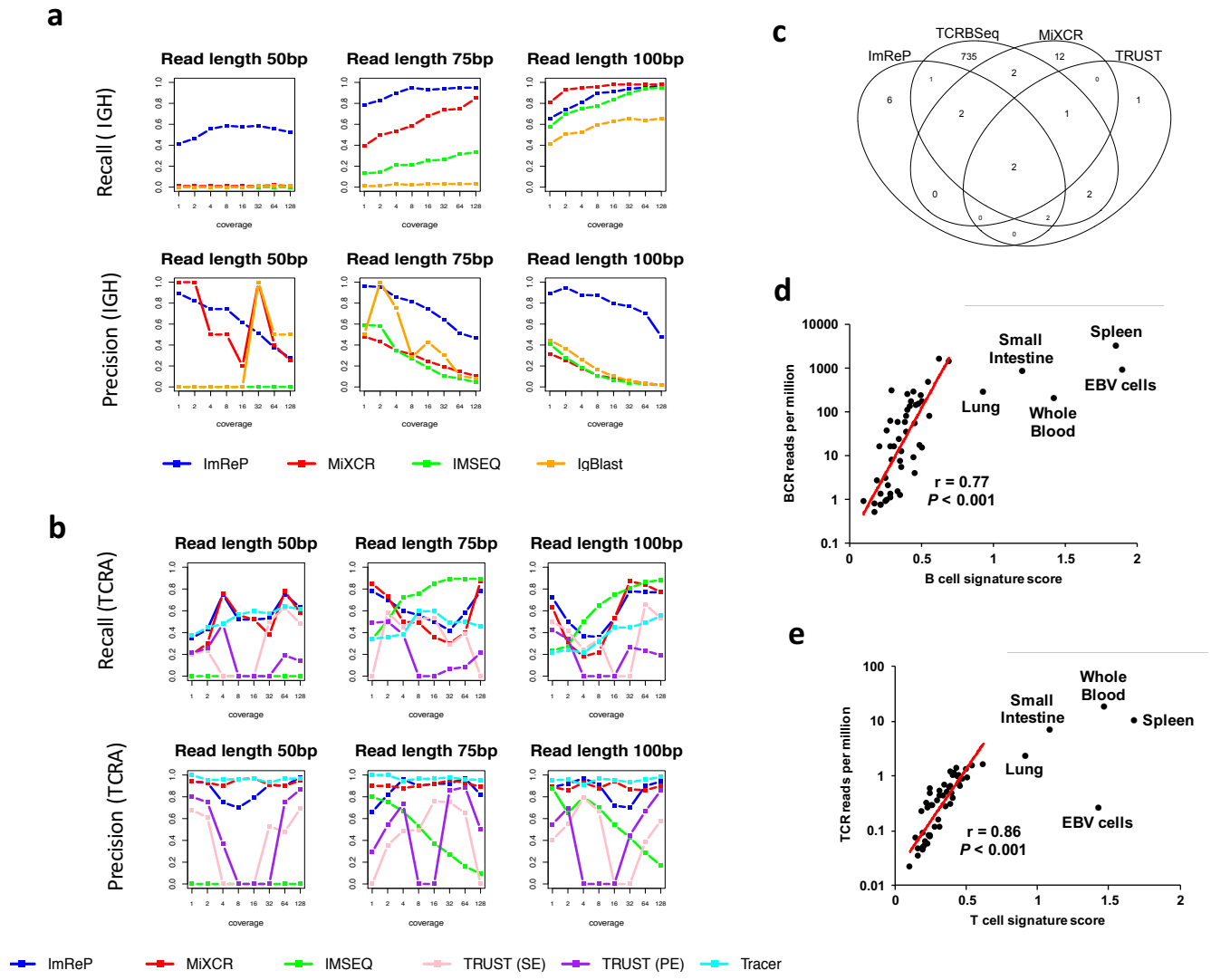


Figure 2. Evaluation of ImReP. (a-b) Evaluation of ImReP based on the number of assembled CDR3 sequences and comparison to existing methods. **(c)** Concordance of targeted TCRB-Seq and non-specific RNA-Seq. **(d-e)** Correspondence of ImReP-derived reads from B-cell (BCR) and T-cell (TCR) receptors to the relative abundance of B- and T-cells inferred from cell-specific gene expression profiles. **(a)** Precision and recall rates for ImReP (blue), MiXCR (RNA-Seq mode) (blue), IMSEQ (green), and IgBlast (orange) on simulated data for immunoglobulin heavy (IGH) transcripts are reported for various reads length (separate plots) and per transcript coverages (1,2,4,8,16,32,64,128) (x-axis). TRUST and TraCeR do not support the analysis of BCR sequences and were excluded from the comparison

based for the IGH data. **(b)** Precision and recall rates for ImReP (blue), MiXCR (RNA-Seq mode) (red), TRUST (default, paired-end mode) (pink), TRUST (single-end mode) (violet), IMSEQ (green), and TraCeR (aqua) on simulated data for T cell receptor alpha (TCRA) transcripts are reported for various reads length (separate plots) and per transcript coverages (1,2,4,8,16,32,64,128) (x-axis). **(c)** Concordance of targeted TCRB-Seq and non-specific RNA-Seq performed on three TCGA samples (only one is shown) from kidney renal clear cell carcinoma (KIRC). Venn diagram on TCGA-CZ-5463 sample presents number of matching CDR3s reported by immunoSEQ Analyzer (http://www.adaptivebiotech.com/) and CDR3 sequences assembled from non-specific RNA-Seq data by ImReP, MiXCR (RNA-Seq mode), TRUST (default, paired-end mode). Results on other samples are presented in Figure S4. **(d)** Scatterplot of the number of all BCR reads per 1 million RNA-Seq reads (y-axis) and B-cell signature score inferred by SaVant (x-axis). **(e)** Scatterplot of the number of all TCR reads per 1 million RNA-Seq reads (y-axis) and B-cell signature score inferred by SaVant (x-axis). Pearson correlation coefficient (r) and *P*-value are reported.

Characterizing the adaptive immune repertoire across 53 GTEx tissues

ImReP identified over 26 million reads overlapping 3.8 million distinct CDR3 sequences that originate from diverse human tissues. The majority of assembled CDR3 sequences derived from BCRs, with 1.7 million from the immunoglobulin heavy chain (IGH), 0.9 million from the immunoglobulin kappa chain (IGK), and 1.0 million from the immunoglobulin lambda chain (IGL). A smaller fraction of CDR3 sequences derived from TCRs, with 0.2 million sequences from alpha and beta TCRs (TCRA and TCRB). The vast majority of all assembled CDR3s had a low frequency in the data. 98% of CDR3 sequences had a count of less than 10 reads, and the median CDR3 sequence count was 1.4. CDR3 sequences derived from IGK were the most abundant across all tissues, accounting on average for 54% of the entire B-cell population (Figure S5). In the T-cell

population, alpha and beta jointly accounted for 83% of the population. Delta T-cell population was the rarest, accounting for less than one percent of the entire T cell population (Figure S6).

We compared the length and amino acid composition of the assembled CDR3 sequences of immunoglobulin and T-cell receptor chains (Figure 3a-g). Consistent with previous studies, we observed that immunoglobulin light chains have notably shorter and less variable CDR3 lengths compared to heavy chains[17] (Fig. 3h). The tissue type appears to have no effect on the length distribution of CDR3 sequences of BCRs (Figure S7). Differences in the CDR3 length distributions for TCRs cannot be estimated due to the small number of available TCRs. Sequencing composition[18] of CDR3 regions of beta T-cells assembled by ImReP recapitulates one detected by TCR sequencing[2] from the whole blood with two dominant CASS and CSAR motifs in the beginning of the sequence (Fig. 3f). In line with other studies[19], both light chains exhibited a reduced amount of sequencing diversity (Fig. 3b-c).

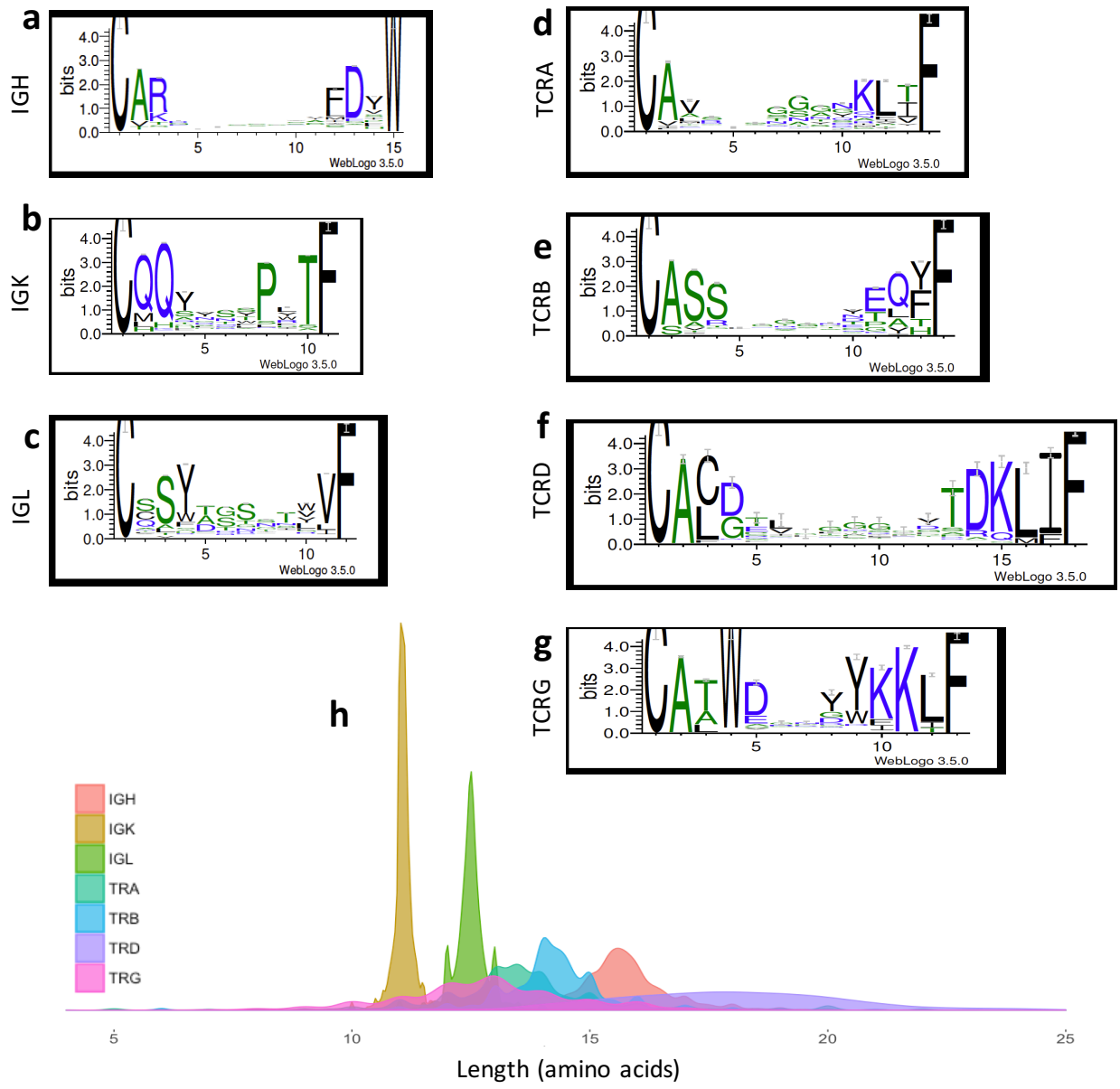


Figure 3. Length and amino acid composition of the assembled CDR3 sequences of immunoglobulin and T-cell receptor chains. The sequence logo (using WebLogo) of amino acid composition representation for CDR3 sequences with mean length. The height of the amino acid within the stack indicates the relative frequency. Distribution of CDR3 sequence length is estimated using s kernel density. **(a)** Sequence logo of 15-amino-acid CDR3 sequence of IGH. **(b)** Sequence logo of 11-amino-acid CDR3 of IGK. **(c)** Sequence logo of 12-amino-acid CDR3 sequence of IGL. **(d)** Sequence logo of 14-amino-acid CDR3 sequence of TCRA. **(e)** Sequence logo of 14-amino-acid CDR3 sequence of TCRB. (f) Sequence logo of 17-amino-acid CDR3 sequence of TCRD. **(g)** Sequence logo of 13-amino-acid CDR3

sequence of TCRG. (h) Distribution of CDR3 sequence length is estimated using s kernel density separately for each chain of T- and B-cell receptors.

We observed per sample an average of 1331 distinct clonotypes for BCRs and 20 distinct clonotypes sequences for TCRs. We normalized the number of distinct clonotypes by the total number of RNA-Seq reads, which we call number of clonotypes per one million reads (CPM). As the number of distinct clonotypes does not increase linearly with the sequencing depth, CPM metric should not be used in studies comparing clonotype diversity across various phenotypes. Instead, CPM is intended to be an informative measure of clonal diversity adjusted for sequencing depth.

We used per sample alpha diversity (Shannon entropy) to incorporate the total number of distinct clonotypes and their relative frequencies into a single diversity metric. Among all tissues, spleen has the largest B-cell population, with a median of 1301 BCR-derived reads per one million RNA-Seq reads. It also has the most diverse population of B cells with median per sample alpha diversity of 7.6 corresponding to 1025 CPM (Figure 4 and Table S1). Whole blood has both the largest and most diverse T-cell population (Figure 4 and Table S1). Organs that possess mucosal, exocrine, and endocrine sites (n=24) harbor a rich clonotype population with a median of 87 CPM per sample. Minor salivary glands have the highest immune diversity in the group (alpha=7.1) and surpass the diversity of the terminal Ileum containing Peyer's Patches, which are secondary lymphoid organs (Table S1).

Tissues not related to the immune system, including adipose, muscle, and the organs from the central nervous system, contained a median of 6 CPM per sample, which are most likely due to the blood content of the tissues[20]. The highest number of distinct CDR3 sequences among non-lymphoid organs was present in the omentum, a membranous double layer of adipose tissue containing fat-associated lymphoid clusters. As expected[21], Epstein bar virus (EBV)-transformed lymphocytes (LCL) harbored a large homogeneous population of B cell clonotypes (Table S1 and Figure S7).

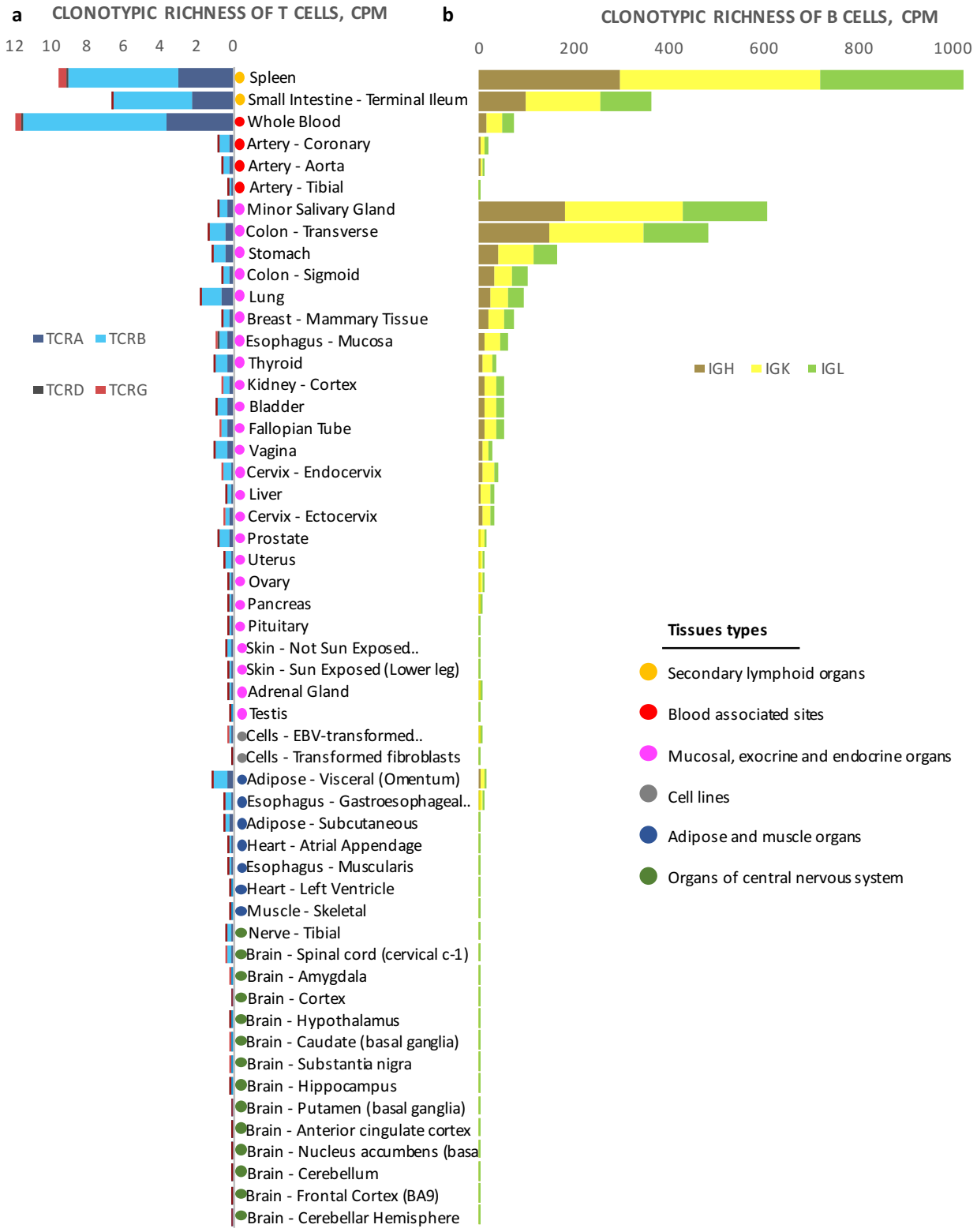


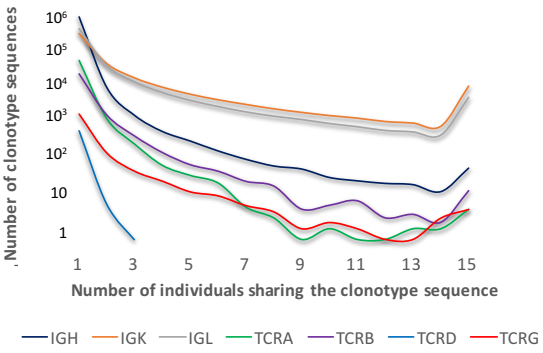
Figure 4. Adaptive immune repertoires across multiple human tissues. Adaptive immune repertoires of 8,555 samples across 544 individuals from 53 body sites obtained from Genotype-Tissue Expression study (GTEx v6). We group the tissues by their relationship to the immune system. The first group includes the lymphoid tissues (n=2, red colors). The second group includes blood associated sites including whole blood and blood vessel (n=4, red color). The third group are the organs that encompasses mucosal, exocrine and endocrine organs (n=21, lavender color). The fourth group are cell lines (n=3, grey color). The filth group are adipose or muscle tissues and the gastroesophageal junction (n=7, blue color). The sixth group are organs from central nervous system (n=14, green color). Histogram reports clonotypic richness of T and B cells, calculated as number of distinct amino acid sequences of CDR3 per one million RNA-Seq reads (CPM). (**a)** Median CPMs are presented individually for T cell receptor alpha chain (TCRA), T-cell receptor beta chain (TCRB), T-cell receptor delta chain (TCRD), and T-cell receptor gamma chain (TCRG). (**b)** Median number of distinct amino acid sequences of CDR3 are presented individually for immunoglobulin heavy chain (IGH), immunoglobulin kappa chain (IGK), immunoglobulin lambda chain (IGL).

Individual- and tissue-specific T- and B-cell clonotypes

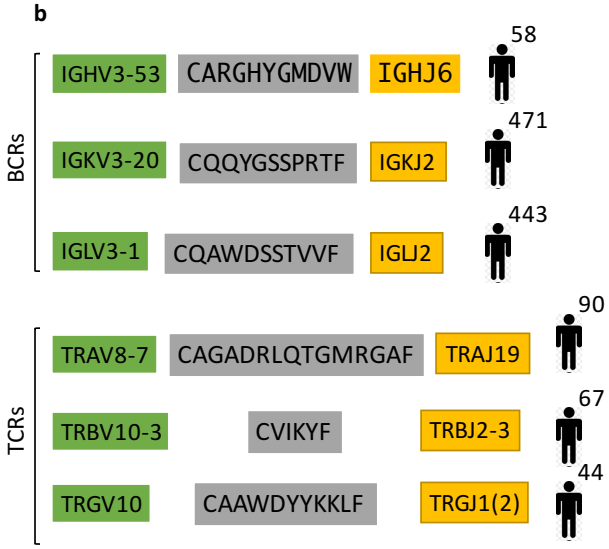
Amino acid sequences of clonotypes exhibited extreme inter-individual dissimilarity, with 88% of clonotypes unique to a single individual (private) (Figure 5a). The remaining ~400,000 clonotypes were shared by at least two individuals (public). The number of individuals sharing clonotypes varied across T and B cell receptors, with immunoglobulin light chains having the highest number of public clonotypes. Twenty-five percent of all IGK clonotypes were public, and the number of individuals sharing the IGK clonotype sequences can be as high as 471 (Figure 5b). T cell clonotypes had on average 7% public clonotypes, with the highest number of public clonotypes among the delta chain of gamma-delta TCRs (12%). The limited capacity of RNA-Seq to cover low

abundant clonotypes may misclassify public clonotypes as private. Consistent with previous studies[10,22], we observe public clonotypes to be significantly shorter in length than the private ones (p-value<$2x10^{-16}$). For example, IGH chain public clonotypes had an average length of 13 amino acids, and private clonotypes had an average length of 16. We also examined whether the public clonotypes were more often shared across tissues within an individual. Only 14% of the ~240,000 clonotypes shared across tissues were public. The majority of clonotypes were individual- and tissue-specific (Figure 5c). The full list of public clonotypes is distributed with the 'Atlas of T- and B-cell repertoires' that accompanies this manuscript.

a



b



c

	Tissue specific	Shared across tissues
Individual specific (private)	3.1×10^6	0.2×10^6
Shared across individuals (public)	0.4×10^6	0.03×10^6

Figure 5. Private and public T and B cell clonotypes. (a) Distribution of frequencies of private (n=1) and public (n>1) clonotypes across 544 individuals. We collect clonotypes from all tissues of the same individual into a single set corresponding to that individual. **(b)** The most public clonotypes (shared across maximum number of individuals) and corresponding VJ recombination are presented for IGH, IGK, IGL, TCRA, TCRB, and TCRG. **(c)** Clonotypes sequences are classified into public clonotypes (shared across individuals), private (individual specific), tissue-specific, and clonotypes shared across multiple tissues. The number of clonotypes falling into each pair of categories is reported across all T and B cell receptor chains.

Flow of T and B cell clonotypes across human GTEx tissues

The large number of individuals available through the study allow us to establish a pairwise relationship between the tissues and track the flow of T and B clonotypes across human tissues. W observed a significant increase of the CDR3 sequences shared across pairs of tissues from the same individuals. Further, we observed this pattern consistently for all chains of B and T cell receptors (p-value<$2x10^{-16}$) (Figure 6a and Table S2). We observe a different amount of shared CDR3 sequences across different types of BCRs and TCRs with an increase in immunoglobulin light chains. Decreased number of TCR reads compared to the BCRs makes it unfeasible to compare the number of shared CDR3 across T and B cells. On average, we observe 21.0 CDR3 sequences to be shared across a pair of tissues from the same individuals. Pairs of tissues from different individuals share on average 10.6 CDR3 sequences (Figure 6a and Table S2).

To establish the flow of B- and T-cell clonotypes across various tissues, we compared clonotype populations between and within the same individuals. We limited this analysis to pairs of tissues for which we had at least 10 individuals (870 pairs of tissues out of 1378 possible pairs). We used beta diversity (Sørensen–Dice similarity index) to measure compositional similarities between the tissues in terms of gain or loss of CDR3 sequences (Figure 6b-c). For the majority of the 870 available tissue pairs, we observe no BCR or TCR sequences in common, which corresponds to beta diversity of 0.0.

We examined the flow of IGH clonotypes across tissues and presented it as a network (Figure 5b). Among 870 available tissue pairs, we have identified 56 tissue pairs with beta diversity above .001. Spleen was the most highly connected tissue, with 17 connections, followed by lung, with 16 connections. Clonotypes represents one connected component, meaning that every two nodes are connected directly or via other nodes. Clonotype populations of spleen and lung are the most similar (0.02 beta diversity), other pairs include minor salivary gland and esophagus mucosa, terminal ileum (small intestine) and transverse colon. We observe above 200 pairs of tissues with beta diversity above .001 for immunoglobulin light chains (Figure S9-S10). The most similar tissue pairs for IGK chain were spleen and transverse colon (0.15 beta diversity).

We also examined the flow of TCRB clonotypes across tissues. TCRB clonotype sequences were shared across 10 pairs of tissues with average beta diversity of 0.02. Similar to B cells, the network of T cell clonotypes is a single connected component, with spleen being the most highly connected tissue (Figure 6c). For the TCR gamma chain, we observe beta diversity of 0.0 for all tissue pairs. At the same time 12% of TCR gamma clonotypes are public, showing the highest rate among all TCR genes. At the sequencing depth provided by RNA-Seq, we are unable to observe TCR delta clonotype sharing across individuals and tissues.

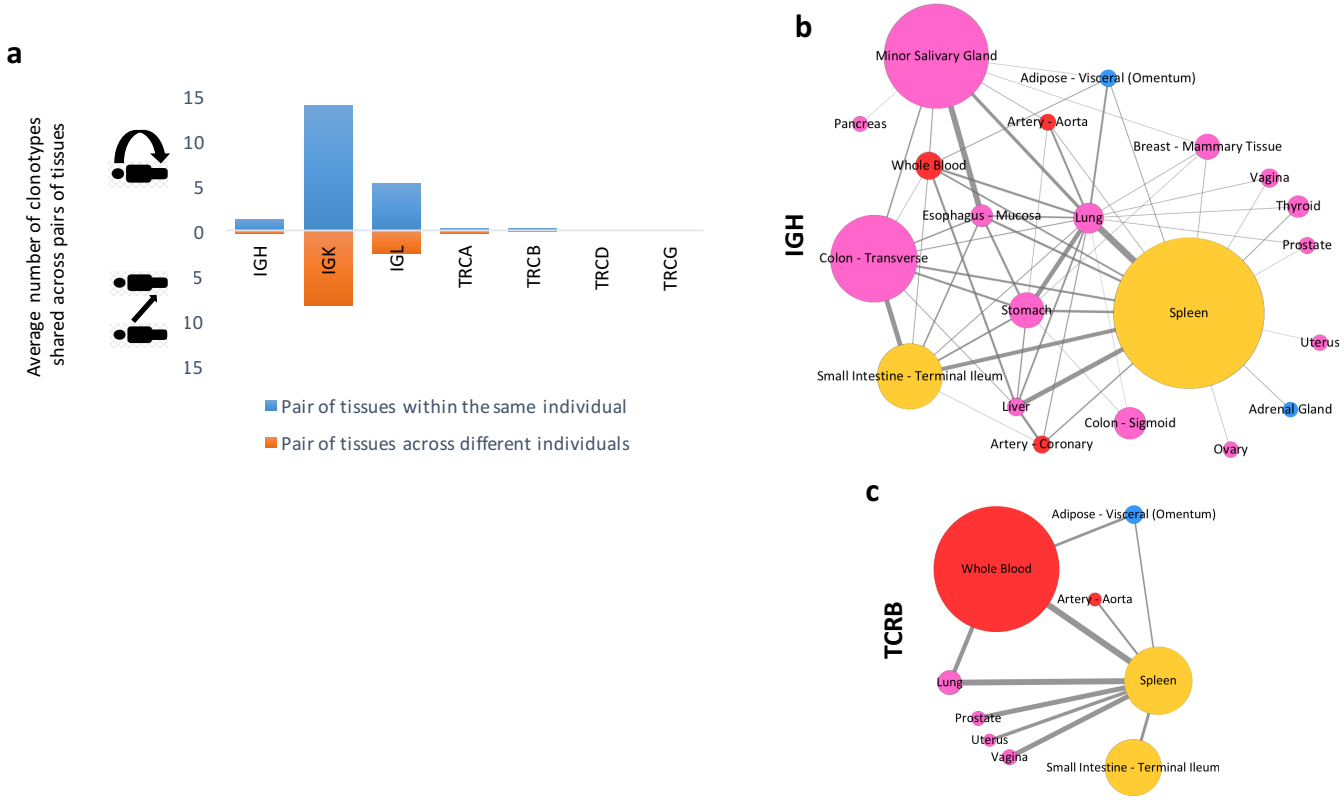


Figure 6. Flow of T and B cell clonotypes across diverse human tissues. Results are based on pairs of tissues with at least 10 individuals. **(a)** The number of clonotype sequences shared across pairs of tissues from the same individuals (blue color) and from different individuals (orange color) are presented. Number of clonotypes shared across tissues from the same individuals for TCRs is <.1. Number of clonotypes shared across tissues from different individuals for TCRs is <.001. **(b)-(c)** Flow of clonotypes across diverse human tissues is presented as a network. Each node is a tissue with the size proportional to a median number of clonotypes of the tissue. The color of the node corresponds to a type of the tissue type: lymphoid tissues (yellow colors), blood associated sites (red color), organs that encompasses mucosal, exocrine and endocrine organs (lavender color). Compositional similarities between the tissues in terms of gain or loss of CDR3 sequences are measured across valid pairs of tissues using beta diversity (Sørensen–Dice similarity index). Edges are weighted according to the beta diversity. **(b)** Flow of IGH clonotypes across diverse

human tissues is presented as a network. Edges with beta diversity >.001 are presented. **(c)** Flow of TCRB clonotypes across diverse human tissues is presented as a network. Edges with beta diversity >.001 are presented.

ImReP identifies tissue samples with lymphocyte infiltration

Histological images of tissue cross-sections and pathologists' notes have been used to validate the ImReP's ability to detect the samples with a high lymphocyte content, which often corresponds to a disease state. We examined the IGH clonotype populations from thyroid tissue across individuals. The median number of inferred distinct CDR3 sequences per sample was 20, though 14.5% of the samples had more than 500 distinct CDR3 sequences. We observed the highest number of CDR3 sequences among all the thyroid samples in an individual with late stage Hashimoto's thyroiditis, an autoimmune disease characterized by lymphocyte infiltration and T-cell mediated cytotoxicity. According to pathologists' notes, Hashimoto's disease was present in 11.2% of thyroid samples, with varying degrees of severity. First, we used pathologists' notes to annotate samples as healthy or bearing Hashimoto's disease, and then we compared the adaptive repertoire diversity between these groups. We observed a significant increase in the number of distinct IGH clonotypes in samples with Hashimoto's thyroiditis (p-value= 1.5×10^{-5}) (Figure S11). The number of clonotypes varied from 113 for focal Hashimoto's thyroiditis to 5621 for late stage Hashimoto's thyroiditis (Figure 7a). In addition, high clonotype diversity in kidney samples indicated the presence of glomerulosclerosis. In lung samples, high clonotype diversity corresponded to inflammatory diseases such as sarcoidosis and bronchopneumonia.

We observed no difference in clonal diversity in males and females across the tissues, except in breast tissues (p-value<3.2×10^{-12}, BCRs). Increased clonotype diversity of breast tissue in male individuals corresponded to gynecomastia, a common disorder of non-cancerous enlargement of male breast tissue (Figure 7b).

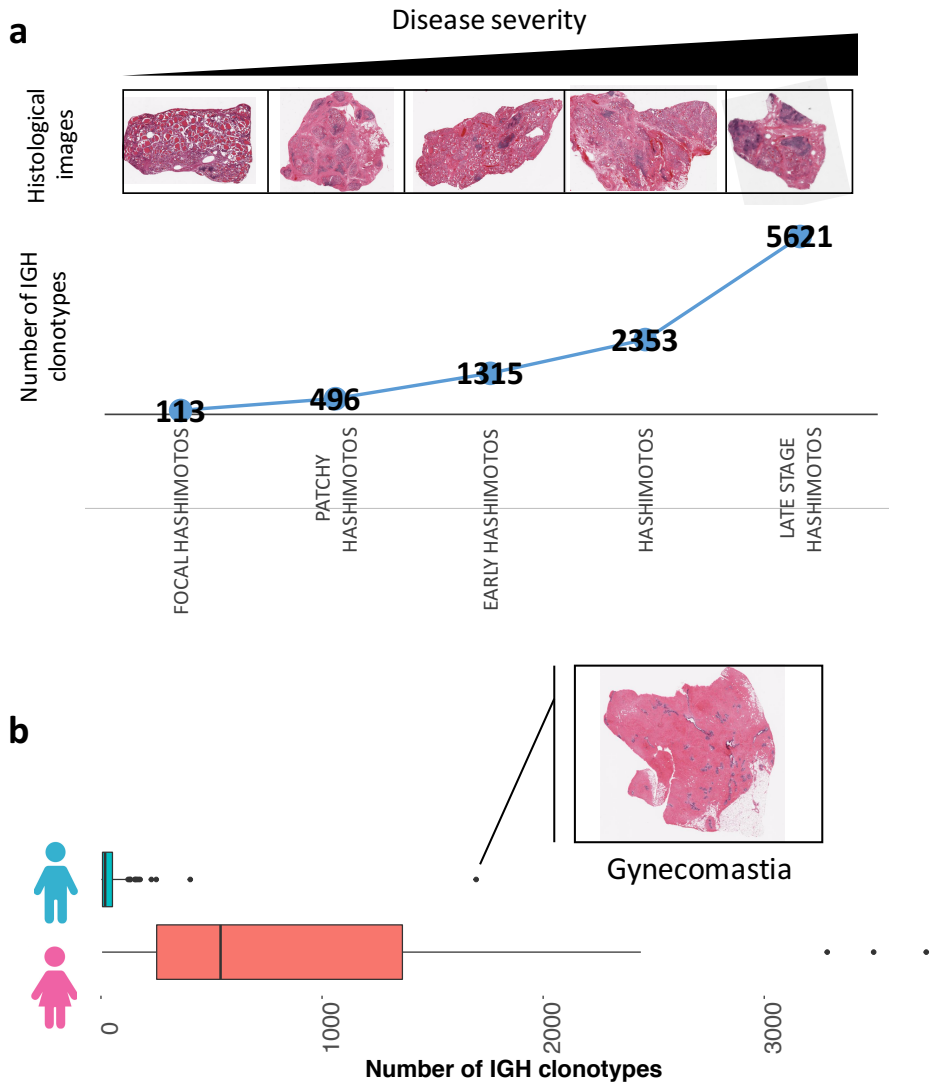


Figure 7. ImReP is able to identify samples with high activity of lymphocytes. Histological images of tissue cross-sections and pathologists' notes have been used to validate the ImReP's ability to detect the samples with a high activity of lymphocytes. **(a)** Samples were ordered by Hashimoto's thyroiditis severity, as reported by pathologists' notes. Histological images are provided to illustrate the disease state. Average number of clonotypes is reported for each disease group. **(b**) Boxplot reporting number of clonotypes in the breast tissues for males and females. Outlier among the male samples is illustrated with the histological image.

Discussion

We have developed a novel computational approach (ImReP) for reconstruction of adaptive immune repertoires using RNA-Seq data. We demonstrate the ability of ImReP to efficiently extract TCR- and BCR- derived reads from the RNA-Seq data and accurately assemble corresponding BCR and TCR clonotypes. The proposed algorithm can accurately assemble CDR3 sequences of immune receptors despite the presence of sequencing errors and short read length. Simulations generated using various read lengths and coverage depth show that ImReP consistently outperforms existing methods in terms of precision and recall rates.

We have demonstrated the feasibility of applying RNA-Seq to study the adaptive immune repertoire. Although RNA-Seq lacks the sequencing depth of targeted sequencing (Rep-Seq), it can compensate by examining a larger sample size. Using ImReP, we have created the first systematic atlas of immunological sequence for B- and T-cell receptor repertories across diverse human tissues. This provides a rich resource for comparative analysis of a range of tissue types, most of which have not been studied before. The atlas of T- and B-cell receptors, available with the paper, is the largest collection of CDR3 sequences and tissue types. We anticipate that this database will enhance future studies in areas such as immunology and contribute to the development of therapies for human diseases.

Using RNA-Seq to study immune repertoires has some advantages, including the ability to simultaneously capture both T and B cell clonotype populations during a single run. It also allows

simultaneous detection of overall transcriptional responses of the adaptive immune system, by comparing changes in the number of BCR and TCR transcripts to the much larger transcriptome. Given large number of large-scale RNA-Seq datasets becoming available, we look forward to scaling up the atlas of T- and B-cell receptors in order to provide valuable insights into immune responses across various autoimmune diseases, allergies, and cancers.

References

1. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–168 (2014).

2. 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).

3. Rajewsky, K., Forster, I. & Cumano, A. Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science (80-.).* **238,** 1088–1094 (1987).

4. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135,** 183–191 (2012).

5. Mangul, S. *et al.* Dumpster diving in RNA-sequencing to find the source of every last read. *bioRxiv* 53041 (2016).

6. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

7. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **6,** 281–297 (1999).

8. Lefranc, M.-P. *et al.* IMGT{®}, the international ImMunoGeneTics information system{®} 25 years on. *Nucleic Acids Res.* gku1056 (2014).

9. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat.*

Methods **12,** 380–381 (2015).

10. Li, B. *et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* (2016).
11. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* (2016).
12. Mose, L. E. *et al.* Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* btw526 (2016).
13. Strauli, N. & Hernandez, R. Statistical Inference of a Convergent Antibody Repertoire Response to Influenza Vaccine. *bioRxiv* 25098 (2015).
14. Warren, R. L., Nelson, B. H. & Holt, R. A. Profiling model T-cell metagenomes with short reads. *Bioinformatics* **25,** 458–464 (2009).
15. Kuchenbecker, L. *et al.* IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* **31,** 2963–2971 (2015).
16. Lefranc, M.-P. *et al.* SaVanT: A web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. *Nucleic Acids Res.* gku1056 (2014).
17. Philibert, P. *et al.* A focused antibody library for selecting scFvs expressed at high levels in the cytoplasm. *BMC Biotechnol.* **7,** 1 (2007).
18. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
19. Hoi, K. H. & Ippolito, G. C. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. *Genes Immun.* **14,** 271–276 (2013).
20. Yu, H.-P., Chiu, Y.-W., Lin, H.-H., Chang, T.-C. & Shen, Y.-Z. Blood content in guinea-pig

tissues: correction for the study of drug tissue distribution. *Pharmacol. Res.* **23,** 337–347 (1991).

21. De Rossi, A. *et al.* Infection of Epstein-Barr virus-transformed lymphoblastoid B cells by the human immunodeficiency virus: evidence for a persistent and productive infection leading to B cell phenotypic changes. *Eur. J. Immunol.* **20,** 2041–2049 (1990).

22. 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).