Profiling immunoglobulin repertoires across multiple human tissues using RNA Sequencing

Anonymous Authors[1]

Abstract

Profiling immunoglobulin (Ig) receptor repertoires with specialized assays can be cost-ineffective and time-consuming. Here we report ImReP, a computational method for rapid and accurate profiling of the Ig repertoire, including the complementary-determining region 3 (CDR3s), using regular RNA sequencing data such as those from 8,555 samples across 53 tissues types from 544 individuals in the Genotype-Tissue Expression (GTEx v6) project. Using ImReP and GTEx v6 data, we generate a collection of 3.6 million Ig sequences, termed the atlas of immunoglobulin repertoires (TAIR), across a broad range of tissue types that often do not have reported Ig repertoires information. Moreover, the flow of Ig clonotypes and inter-tissue repertoire similarities across immune-related tissues are also evaluated. In summary, TAIR is one of the largest collections of CDR3 sequences and tissue types, and should serve as an important resource for studying immunological diseases.

1. Introduction

Igs 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. The resulting DNA sequences are then translated into antigen receptor proteins. This process enables the Ig repertoire to develop astonishing diversity of antigen receptors from any given individual, with over $> 10^{13}$ theoretically possible distinct Ig receptors. Ig repertoire diversity is key for an individual's immune system to confer protection against a wide variety of potential pathogens.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

Studies involving assay-based protocols usually have small sample sizes, thus limiting analysis of intra-individual variation of immunological receptors across diverse human tissues. In contrast to assay-based protocols that produce reads from the amplified variable region of the Ig locus, RNA-Seq is able to capture the entire cellular population of the sample, including B cells. However, due to the repetitive nature of the Ig locus, and the extremely high level of diversity in Ig transcripts, most mapping tools are ill-equipped to handle Ig sequences.

In this study, we develop ImReP, an alignment-free computational method for rapid and accurate profiling of the Ig repertoire from regular RNA-Seq data. ImReP is capable of efficiently extracting receptor-derived reads from RNA-Seq data and accurately assembling Ig clonotypes, defined as distinct amino acid sequences of complementarity-determining region 3 (CDR3).

2. Results

2.1. Existing tools for profiling the Ig receptor repertoire

Most methods use alignment or assembly to infer CDR3s and align reads to V and J genes. In contrast, the ImReP procedure provides a match between the read prefix and the read suffix to the prefix of J genes and suffix of V genes, respectively, without a need for alignment. In avoiding alignment, ImReP is able to significantly decrease running time and minimize required computational resources. Average CPU time reported for ImReP is 44 minutes, a runtime substantially shorter than the average 10 hours required for MiXCR. On average, per sample, ImReP consumes 3G of CPU while MiXCR requires 10G of CPU.

2.2. ImReP is a method for profiling of Ig repertoire

We apply ImReP to 0.6 trillion RNA-Seq reads (92 Tbp) from 8,555 samples to assemble CDR3 sequences of Ig receptors. The RNA-Seq data was generated by the Genotype-Tissue Expression Consortium (GTEx v6). ImReP is a two-stage alignment-free approach to assembling CDR3 sequences and detecting corresponding V(D)J recombinations. In the first stage, we prepare the candidate receptor reads

from mapped and unmapped RNA-Seq reads. In the second stage, ImReP utilizes reads that contain a partial CDR3 sequence and overlap a single gene segment (V or J). ImReP performs matching with a suffix tree technique; matched reads with an overlap of at least 15 nucleotides are used to assemble full-length CDR3s.

2.3. Feasability of using RNA-Seq to study the Ig repertoire

To validate the feasibility of using RNA-Seq to study the Ig receptor repertoire, we simulate RNA-Seq data as a mixture of transcriptomic reads and reads derived from Ig transcripts. We assess the ability of ImReP to extract CDR3-derived reads from the RNA-Seq mixture by applying ImReP to a simulated RNA-Seq mixture. ImReP is able to maintain an 80% precision rate for the majority of simulated scenarios. Average CDR3 coverage that is higher than eight allows ImReP to archive a recall rate close to 90% for a read length above 75bp. We compare the performance of ImReP to MiXCR (RNA-Seq mode), IgBlast-based pipeline, and IMSEQ. ImReP consistently outperforms existing methods in both recall and precision rates.

To further demonstrate the feasibility of applying non-specific RNA sequencing techniques to profile Ig receptor repertoires, we use 18 tumor biopsies sequenced by BCR-Seq and RNA-Seq. Using RNA-Seq, ImReP is able to capture on average 53.3% of the IGH repertoire, estimated as the sum of detected BCR-seq-confirmed clonotypes; MiXCR is able to capture 40.1%. Both methods are able to accurately estimate the relative frequencies of assembled clonotypes (ImRep: r=0.97, p-value=4.4x10-40; MiXCR r=0.87, p-value=5.1x10-15).

2.4. Characterizing the Ig repertoire across 53 GTEx tissues

ImReP identifies over 26 million reads overlapping 3.6 million distinct CDR3 sequences that originated from diverse human tissues. To account for various sequencing depths we further normalized the detected number of clonotypes by the total number of RNA-Seq reads. We refer to this measure as clonotypes per one million raw RNA-Seq reads (CPM). We use per sample alpha diversity (Shannon entropy) to incorporate into a single diversity metric the total number of distinct clonotypes and their relative frequencies. Among all tissues, spleen has the largest B-cell population. Spleen also has the most diverse population of B cells with median per sample alpha diversity rate of 7.6, corresponding to 1025 CPM. Organs that possess mucosal, exocrine, and endocrine sites (n=24) harbor a rich clonotype population with a median of 87 CPM per sample.

2.5. Ig clonotypes specific to an individual or a tissue type

Amino acid sequences of clonotypes exhibit extreme inter-individual dissimilarity, with 88% of clonotypes unique to a single individual (private). The remaining 400,000 clonotypes are shared by at least two individuals (public). Twenty-five percent of all IGK clonotypes are public, and the number of individuals sharing the IGK clonotype sequences can be as high as 471. Overall, 14% of the 240,000 clonotypes from both light and heavy chains shared across tissues are public. The full list of public clonotypes is distributed with the Atlas of Immunoglobulin Repertoires (TAIR), which is publically available at https://github.com/Mangul-Lab-USC/TAIR.

2.6. The flow of Ig clonotypes across human GTEx tissues

We observe a significant increase in the number of CDR3 sequences shared across pairs of tissues obtained from the same individual. Further, we consistently observe this pattern for all chains of Ig receptors. We examine the flow of IGH clonotypes across tissues. Among 870 available tissue pairs, we identify 56 tissue pairs with a beta diversity score above .001. The spleen has the most highly connected tissue (17 connections), followed by lung (16 connections).

2.7. ImReP identifies tissue samples with lymphocyte infiltration

We observe a significant increase in the number of distinct IGH clonotypes in samples from individuals with Hashimoto's thyroiditis. We also observe a significant increase in the number of distinct IGH clonotypes in positive correlation with the noted severity of Hashimoto's thyroiditis. We observe no difference in clonal diversity in males and females across the tissue types, except in breast tissues.

3. Availability

This manuscript has been recently accepted to Nature Communications, pending minor changes. The preprint of the full manuscript is currently available on bioRxiv via the following link: https://www.biorxiv.org/content/10.1101/089235v3.

ImReP is freely available at https://github.com/Mangul-Lab-USC/imrep. ImReP is distributed under the terms of the General Public License version 3.0 (GPLv3). All code required to produce the figures and analysis performed in this paper are freely available at https://github.com/Mangul-Lab-USC/ImReP_publication.