A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data

Yaxuan Yu, Rhodri Ceredig and Cathal Seoighe

*J Immunol* 2017; 198:2202-2210; Prepublished online 23 January 2017;
doi: 10.4049/jimmunol.1601710
http://www.jimmunol.org/content/198/5/2202

**Supplementary Material**
http://www.jimmunol.org/content/suppl/2017/01/20/jimmunol.1601710.DCSupplemental

**References**
This article cites 32 articles, 14 of which you can access for free at:
http://www.jimmunol.org/content/198/5/2202.full#ref-list-1

**Subscription**
Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

**Permissions**
Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

**Email Alerts**
Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data

Yaxuan Yu,* Rhodri Ceredig, † and Cathal Seoighe*

High-throughput sequencing data from TCRs and Igs can provide valuable insights into the adaptive immune response, but bioinformatics pipelines for analysis of this data are constrained by the availability of accurate and comprehensive repositories of TCR and Ig alleles. We have created a analytical pipeline to recover immune receptor alleles from genome sequencing data. Applying this pipeline to data from the 1000 Genomes Project we have created Lym1K, a collection of immune receptor alleles that combines known, well-supported alleles with novel alleles found in the 1000 Genomes Project data. We show that Lym1K leads to a significant improvement in the alignment of short read sequences from immune receptors and that the addition of novel alleles discovered from genome sequence data are likely to be particularly significant for comprehensive analysis of populations that are not currently well represented in existing repositories of immune alleles. The Journal of Immunology, 2017, 198: 2202–2210.

The online version of this article contains supplemental material.

Abbreviations used in this article: G1K, 1000 Genomes Project; IMGT, international ImMunoGeneTics information system; SNP, single nucleotide polymorphism; SRA, National Center for Biotechnology Information Sequence Read Archive; VCF, variant call format.

Copyright © 2017 by The American Association of Immunologists, Inc. 0022-1767/17/$30.00

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1601710

The Journal of Immunology

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1601710

The Journal of Immunology
studies. We have developed a bioinformatics pipeline that can be used to generate a TCR/Ig reference database using data from population resequencing studies. We applied this pipeline to data from the 1000 Genomes Project (G1K) (29), which contains the most comprehensive collection of global human variation currently available in the public domain, to create the Lym1K database. We aligned short read sequence data from human immune receptors to Lym1K and obtained significantly improved alignments compared with what could be achieved using the IMGT database as reference. We also report evidence that the diversity of immune receptor alleles observed in non-African populations is particularly underrepresented in the existing database.

Materials and Methods

Immune receptor sequencing dataset

We obtained two datasets from the National Center for Biotechnology Information Sequence Read Archive (SRA). The first dataset was for the TCR β-chain (SRA index: PRJNA310731) and consisted of 34 samples. The number of reads ranged from 4,571,596 to 13,742,272. Each read was 200-bp long and included part of the V, all of the D, and part of the J region. The second dataset contained both Ig H and L chain sequences (SRA index: PRJNA179099). We used four samples from this dataset (all samples that included ≥1 million reads). The number of reads in the samples ranged from 1,046,575 to 23,191,224. Each read was 250 bp long and included part of V (all of the D for the H chain) and part of the J region.

Simulated dataset

Each simulated sequence was created based on the process of VDJ recombination of TCR and Ig and included mismatches to stimulate the combined effects of sequencing errors and mutations/polymorphisms. First, we randomly chose a V, D (only for the H chain of Ig and the β-chain of TCRs), and J gene assuming uniform gene usage. Then we decided the particular allele of the chosen V gene based on the frequency of its occurrence in the chosen population. Second, we “mutated” the chosen V and J segments by introducing 0–10 mismatches on the V region and 0–5 mismatches on the J region. To simulate junction diversity, we inserted zero to six randomly generated nucleotides into the V-D and D-J junction (V-J junction only for the Ig L chain and TCR α-chain).

We applied our simulation pipeline for TCRB, TCRA, IGH, and IGL genes separately. In all, we produced 27 datasets, with separate datasets corresponding to each of the 26 populations included in G1K and 1 dataset derived from the pool of all populations. Each dataset consisted of 20 simulated samples, and each sample contained 200,000 simulated sequences.

G1K data

We retrieved variant cell format (VCF) files corresponding to the TCRB, TCRA, IGH, and IGL regions from phase 3 of G1K. We used the R package SNPRelate (30) to perform principal component analysis separately on variant data from each of the four regions.

AlleleMiner pipeline

The goal of AlleleMiner is to infer all the possible alleles of the target genes from the input VCF files. For the Lym1K database, we first retrieved the location information of TCR and Ig genes from BioMart (GRCh38.p2) and phased VCF files from G1K (GRCh38; phase 3). Note that the current phase 3 VCF files from G1K were generated based on the GRCh37 human genome assembly; however, VCF files in the coordinates of the GRCh38 assembly are also provided.

For each gene, AlleleMiner extracts the corresponding reference genome sequence from the University of California Santa Cruz database and single nucleotide polymorphisms (SNPs) from the input VCF file. Subsequently, AlleleMiner retrieves all alleles of each gene from the haplotypes
spanning the gene of interest. Identical haplotypes are merged, and the number of times a particular haplotype appears is counted. Users can define a threshold for the haplotype occurrence times to eliminate rare haplotypes, some of which may be the result of sequencing error; only the haplotypes that occur more than the threshold number of times are stored as potential alleles in the new TCR/Ig database.

**Software resource**

The pipelines were implemented in JAVA and are freely available at https://sourceforge.net/projects/alleleminer/. The Lym1K database of immune receptor alleles can be obtained from http://maths.nuigalway.ie/biocluster/database/.

**Results**

**Construction of TCR/Ig reference database and performance comparison**

The sequencing of increasing numbers of whole human genomes and improvements in the reference human genome assembly give rise to opportunities to expand substantially the reference database of human immune receptor alleles and to profile the diversity of immune receptor alleles across global populations. Currently, G1K represents the most extensive collection of whole genomes sampled from global human populations in the public domain. Focusing

---

**Table I. Shared allele counts between IMGT database and inferred alleles with minimum repeat threshold of one**

<table>
<thead>
<tr>
<th></th>
<th>IGHV</th>
<th>IGLV</th>
<th>TRBV</th>
<th>TRAV</th>
<th>IGHJ</th>
<th>IGLJ</th>
<th>TRBJ</th>
<th>TRAJ</th>
<th>IGHD</th>
<th>TRBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of alleles in IMGT</td>
<td>208</td>
<td>133</td>
<td>88</td>
<td>56</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>50</td>
<td>34</td>
<td>3</td>
</tr>
<tr>
<td>Number of alleles retrieved from G1K</td>
<td>3746</td>
<td>3917</td>
<td>370</td>
<td>568</td>
<td>100</td>
<td>160</td>
<td>22</td>
<td>162</td>
<td>249</td>
<td>2</td>
</tr>
<tr>
<td>Number of putative novel alleles</td>
<td>3609</td>
<td>3740</td>
<td>318</td>
<td>511</td>
<td>90</td>
<td>141</td>
<td>8</td>
<td>114</td>
<td>219</td>
<td>0</td>
</tr>
<tr>
<td>Number of alleles shared</td>
<td>83</td>
<td>99</td>
<td>54</td>
<td>46</td>
<td>6</td>
<td>12</td>
<td>13</td>
<td>48</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Proportion of IMGT alleles found in G1K, %</td>
<td>39.9</td>
<td>74.4</td>
<td>98.1</td>
<td>100</td>
<td>100</td>
<td>92.86</td>
<td>96</td>
<td>88.24</td>
<td>66.67</td>
<td></td>
</tr>
<tr>
<td>Proportion of IMGT alleles not found in G1K, %</td>
<td>60.1</td>
<td>25.6</td>
<td>1.9</td>
<td>0</td>
<td>53.85</td>
<td>0</td>
<td>7.14</td>
<td>4</td>
<td>11.76</td>
<td>33.33</td>
</tr>
</tbody>
</table>

The compared alleles were restricted to the genes that are GRCh37 and GRCh38. The IMGT alleles retrieved from non-germline sequences were not included in the comparison.
specifically on TCRs and IgS, most human immune receptor genes have been sequenced to sufficiently high coverage to recover novel alleles; the median coverage of IGHV, IGKV, TRBV, and TRAV are accordingly 121, 120.5, 76, and 53 (Supplemental Fig. 1). We retrieved known and novel TCR/Ig alleles from 2504 humans in the phase 3 cohort of G1K. We then tested the performance of the reference database we derived from population resequencing data and compared it to the existing IMGT reference database of human immune receptors using simulated data as well as short read receptor sequencing datasets (Fig. 1).

AlleleMiner workflow
AlleleMiner is the core component of the pipeline and is used to infer immune receptor alleles from genomic data. The algorithm used in AlleleMiner is described in detail in Materials and Methods. In brief, AlleleMiner first enumerates all possible haplotypes of each TCR/Ig gene, using as input genome-wide sequence variants in VCF. Subsequently, the inferred haplotypes that occur in more than a user-defined minimum number of individuals are merged with known alleles (here we merged discovered variants with alleles in the IMGT database to yield the Lym1K database of known and inferred immune receptor alleles). A filtering step is then applied to assess the support for each nonreference SNP allele found among the observed haplotypes (Fig. 2).

Performance assessment
To assess the extent to which known immune receptor alleles could be recovered from population resequencing data, we set the initial inclusion threshold to 1 (i.e., we considered all alleles found at least once in the G1K data). For IGL, TCRA, and TCRB most alleles in IMGT could be recovered (Table I, Supplemental Table I). We then examined the alleles that were not recovered. In the case of TRAV and TRBV, many of these alleles were annotated as originating from rearranged genomic DNA or cDNA, rather than originating from unrearranged germline DNA. Rearranged sequences may have additional somatic variants that we do not expect to find among the alleles we recovered from the genomic sequence. Many of the remaining alleles are from the 22 TRBV genes that are not included in either the GRCh37 or GRCh38 human genome assemblies (GRCh37 was used in G1K, and the current human genome assembly, GRCh38, is used as a reference genome by AlleleMiner). Excluding these genes and the alleles corresponding to rearranged sequences, we were able to retrieve all of the remaining TRAV and all but one allele (TRBV19*02) of the remaining TRBV alleles in IMGT (Table I).

In contrast, although most IGLV alleles were recovered, fewer than half of the IGHV and IGHJ alleles from IMGT were found among the inferred alleles (Table I). This could reflect the greater diversity of Ig genes (Supplemental Fig. 2) or the presence of false-positive alleles in IMGT, particularly in IGHV genes. Indeed, a high rate of false-positive IGHV alleles in IMGT has previously been reported (26). We found evidence for both of these effects. The greater diversity of Ig genes is apparent from the relationship between the number of inferred alleles and the inclusion threshold used in AlleleMiner (Supplemental Fig. 3A). The number of shared IGHV alleles dropped significantly more rapidly than the other three gene groups, suggesting that IMGT includes IGHV alleles that are found at relatively low frequencies in the G1K populations (Supplemental Fig. 3B). To explore the

FIGURE 3. Improvements in alignment performance using the Lym1K database.
impact of false-positive alleles in IMGT on the extent to which known alleles could be recovered from genomic data, we used the validation classes from Wang et al. (26). We found that most of the IMGT alleles that were classified as high confidence (levels 1 and 2) were successfully recovered. In contrast, there were only a few IMGT alleles labeled as low confidence (levels 3, 4, and 5) found among the inferred alleles (Supplemental Fig. 3D).

In addition to the known alleles, we recovered large numbers of novel putative immune receptor alleles from the population sequence data not found in IMGT. These numbered in the hundreds for TRVA and TRBV and in the thousands for IGLV and IGHV, when we applied the minimal threshold of a single occurrence (Table I). The number of putative novel alleles decreased rapidly, as this threshold was increased in the case of Ig, but far more slowly in the TCR case, suggesting saturation with respect to the number of samples in G1K of the TCR but not of the Ig allelic diversity (Supplemental Fig. 1).

Comparison of IMGT and Lym1K on real datasets

We compared the performance of the IMGT and Lym1K databases on real TCR and Ig short read sequence datasets. Because there are no high-throughput sequencing datasets for TCRA currently in the public domain (to our knowledge), this comparison was restricted to

FIGURE 4. Alignment performance difference between IMGT and Lym1K. The aligned reads were classified into 13 categories starting from 0 mismatches with increments of 2 mismatches, to 20 mismatches and >20 mismatches. For example, for IGHV in sample SRR1056423 (top left panel) the number of alignments with zero mismatches increased by 33.2% using Lym1K, relative to IMGT.
IGH, IGL, and TCRB. For each input sequence, when the alignment score of the best match of the sequence is higher using Lym1K rather than IMGT alone as the reference database, it demonstrates that the Lym1K database contains additional allele(s) that are more similar to the input read than the most similar allele from IMGT. Consequently, this corresponds to a sequence for which an improved match can be found when alleles recovered from genomic sequence data are included in the reference database. The median length of the IGHD genes is only 19 nucleotides, and the entire D gene is within the CDR3 region with high mutation rates. During alignment, the median of the perfect matching nucleotides of the D gene is only six, which hampered the accuracy of the assessments on IGHD genes. Therefore we focused our comparison on V and J gene alleles.

FIGURE 5. Example of a novel allele (IGHV1-2*75p) in Lym1K that matches a short read sequence from a real IGHV sequencing dataset (SRR611538). IGHV1-2*02 and IGHV1-2*03 are both found in the IMGT database, and the sequence (SRR611538.578794) was aligned to these alleles in the original study. These two alleles both have two mismatches with the observed sequence. The inferred allele differed from the two most similar alleles found in IMGT, but showed a perfect match with the input sequence read.

We calculated the portions of the input sequences with improved alignment performances (Fig. 3). To assess the differences in alignment performance in greater detail, we classified the aligned reads based on the number of mismatches with the best matching reference allele. Furthermore, to rule out the possibilities that Lym1K achieved improved alignment performance due to the larger number of included alleles, in each individual, we only kept the top two best aligned alleles (corresponding to the number of haplotypes per diploid individual) for each gene in the following comparison. We then compared the proportion of reads within each class between the two reference databases (Fig. 4). The proportions of unaligned reads were consistently lower using Lym1K whereas the proportions of reads with zero or one to two mismatches were consistently higher. This demonstrates that the Lym1K database contains alleles that are more similar to the repertoire of the input sample. For instance, IGHV1-2*75p is a novel allele that we inferred from the G1K data and included in Lym1K. For sample SRR611358 there were 67 input sequences aligned with IGHV1-2*75p with an improved alignment score compared with the best matching alleles (IGHV1-2*02 or IGHV1-2*03) in IMGT (Fig. 5).

In the case of TCRB we introduced 318 putative novel TRBV alleles in Lym1K (Table I). However, the improvements of the alignment performance of the TCRB sequences were small (Fig. 3). Among 34 samples, the highest proportion of reads with improved alignment was 4.1% (in sample SRR1033671) and the average proportion was only 2.4%. Therefore, it appears likely that IMGT contains adequate representation of the TCR diversity in the individuals from whom these samples were derived.

**Improved population coverage of Lym1K**

The Ig dataset consisted only of samples collected from the South African population (31). The TCRB samples were from whites and African Americans. This may have affected the results of the performance comparisons due to the under- or overrepresentation of certain populations in the IMGT database. For instance, if the IMGT database mainly consists of alleles from the white population, this could result in better alignment results for datasets that originate from white populations compared with datasets from African populations. In general, African populations display greater genetic diversity than do non-African populations (32), and this greater genetic diversity can be also be seen in the immune receptor

---

**FIGURE 7.** Comparison of the proportions of incorrectly mapped simulated reads among different populations using the IMGT database as the reference. Error bars corresponding to two SEs are shown in each plot. Population codes: ACB, African Caribbeans in Barbados; ASW, Americans of African ancestry in SW; BEB, Bengali from Bangladesh; CDX, Chinese Dai in Xishaungbanna; CEU, Utah residents (CEPH) with Northern and Western ancestry; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; CLM, Colombians from Medellin; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in U.K. and Scotland; GIH, Gujarati Indian from Houston; GWD, Gambian in western divisions in the Gambia; IBS, Iberian population in Spain; ITU, Indian Telugu from the U.K.; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City; LWK, Luhyia in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican ancestry from Los Angeles; PNL, Punjabi from Lahore; PEL, Peruvians from Lima; PUR, Puerto Ricans from Puerto Rico; STU, Sri Lankan Tamil from the U.K.; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria.
regions (Fig. 6). This suggests that a comprehensive collection of human immune alleles requires sampling from diverse global populations, particularly those from the African continent. We carried out simulations to assess the potential for differential performance of the existing IMGT database on datasets derived from different global populations. The simulation pipeline is described in detail in Materials and Methods. For each chain, we simulated separate datasets for each of the populations in G1K as well as one pooled dataset corresponding to all populations.

We used LymAnalyzer (19) to map simulated reads to reference alleles contained in IMGT and determined the proportion of reads that were mapped to the correct allele. Note that, in this case, our objective was not to compare IMGT to Lym1K, as the simulation is based on the same haplotypes used to construct Lym1K. Instead, we wanted to compare the accuracy of the alignments obtained when we used the existing IMGT database to map simulated reads from different human populations.

For all but one of the genes (TCRA), a significantly lower mapping accuracy was achieved for African than for non-African populations, indicating that the additional variation in immune genes found in African populations has not been captured adequately in IMGT (Fig. 7). For the IGH datasets, the African populations (LWK, GWD, MSL, YRI, ESN, ASW, and ACB) contained significantly more incorrectly mapped reads than in non-African populations ($p < 2.2 	imes 10^{-16}$). The IGL datasets showed the largest portions of reads that were mapped inaccurately using IMGT. Again, using the IMGT database, there were noticeably larger portions of reads incorrectly mapped in the African populations than in non-African populations ($p < 2.2 	imes 10^{-16}$). The portions of incorrectly mapped reads in TRBV datasets ranged from 4.9% (KHV) to 9.9% (LWK), and there were also more incorrectly mapped reads in African populations compared with non-African populations ($p < 2.2 	imes 10^{-16}$). The TCRA datasets had the highest accuracy among the four chains using the IMGT database as the reference, with the smallest variation across populations. In contrast to other chains, the top three most incorrectly mapped populations in TCRA datasets are three East Asian populations (CDX, KHV, and CHS), which were among the least incorrectly mapped populations in other chains (Fig. 7).

**Discussion**

The advent of high-throughput sequencing technologies has enabled detailed studies of immune receptor repertoire diversity and changes in repertoire diversity over time, with many applications, including assessing the efficacy of leukemia treatment, understanding immune responses to disease, and determining the causes and consequences of changes in receptor diversity with age. Recently, a number of dedicated algorithms and bioinformatics tools have been developed to support such studies (11–19); however, in addition to accurate alignment algorithms, the results of bioinformatics pipelines for the analysis of receptor diversity depend on the accuracy and completeness of the reference database of immune receptors and their alleles.

The IMGT database is currently the most widely used global reference in TCR and Ig diversity analysis. However, multiple previous studies have shown that IMGT is incomplete (21–25), and one study suggested that some reported IGHV alleles might contain sequencing errors (26). This is mainly due to quality control issues arising from the inclusion of alleles from early studies and inadequate updating since the first release of the IMGT database. In this study, we provide evidence that alignment results obtained using IMGT as the reference source for TRBV, IGH, and IGL datasets derived from African populations are likely to be worse than those obtained for non-African populations, suggesting that IMGT does not adequately capture the global diversity of human immune receptors and that certain global populations are particularly underrepresented. This supports the value of incorporating immune alleles inferred from population sequencing studies that were specifically designed to profile the genetic diversity of global human populations.

To create a more comprehensive TCR/Ig reference database, we developed a novel bioinformatics pipeline that can be used to infer alleles from variant calling information, obtained from population sequencing projects. We applied this pipeline to data from G1K to create the Lym1K database and have made both Lym1K as well as...
the bioinformatics pipeline used to create it freely available (see “Software resource” above). Using real TCRB, IGH, and IGL datasets, we have shown that the incorporation of alleles inferred from population sequencing studies leads to improvements in alignment performance. These improvements are substantial in the case of Ig and result from novel alleles that provide a better match to the receptor repertoire of the input dataset than can be obtained using the alleles in the existing reference database.

There were in total 22 genes missing either from the current genome build (GRCh38) or GRCh37, and the depth of sequencing was unsatisfactory on several TRBV genes from G1K (Supplemental Fig. 3). This represents a limit to the comprehensiveness of receptor alleles inferred from the currently available population sequencing data; however, further improvements in the human reference genome and future population sequencing studies will enable alleles of these genes to be discovered.

Our bioinformatics pipeline is freely available and can be recovered completely with existing sample numbers (Supplemental Fig. 1). The availability of genomic variants from larger numbers of individuals sampled from global populations could lead to further improvements in the comprehensiveness of immune receptor alleles recovered from genomic data. Our bioinformatics pipeline is freely available and can be applied to recover additional immune alleles from further public or restricted access VCF files and updated genome builds, as they become available. Furthermore, we have designed our short read genomic data. Our bioinformatics pipeline is freely available and can become available. Furthermore, we have designed our short read genomic data. Our bioinformatics pipeline is freely available and can be downloaded from LymAnalyzer (19), so that Lym1K or any other TCR/Ig reference database for their analysis. Currently, G1K represents the largest public catalog of human genome variation. However, as new public as well as restricted-access population sequencing datasets become available, and with further improvements in the human genome assembly, users may want to replace or extend the reference allele database with their own dataset of reference alleles inferred from genomic variant data. Such users can apply AlleleMiner to their own VCF files to generate a customized TCR/Ig reference database, which can then be selected as the reference database for LymAnalyzer. Indeed, depending on the specific data access restrictions that apply, such users may be able to contribute alleles discovered by applying AlleleMiner to restricted-access data into public repositories (such as IMGT), without violating data protection rules.

Disclosures

The authors have no financial conflicts of interest.

References


