

REPORT



Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires

Samuel Schmitz ^a, Cinque Soto ^{b,c}, James E. Crowe Jr. ^{b,c,d}, and Jens Meiler ^{a,e}

^aDepartment of Chemistry, Vanderbilt University, Nashville, TN, USA; ^bDepartment of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA; ^cThe Vaccine Center, Vanderbilt University Medical Center, Nashville, TN, USA; ^dDepartment of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA; ^eInstitute for Drug Development, Leipzig University Medical School, Leipzig, SAC, Germany

ABSTRACT

The antibody (Ab) germline gene rearrangement of variable (V), diversity (D), and joining (J) gene segments, as well as somatic hypermutation, give rise to the human Ab variable gene sequence repertoire. It is common to characterize single nucleotide frequencies of the variable region by alignment to species-specific wildtype germline genes. The increasing application of next-generation sequencing to immune repertoire studies has led to the compilation of increasing large adaptive immunome receptor repertoire datasets. We have developed a method that maps the sequence of a target Ab onto an immunome dataset of 326 million human Ab sequences. For this purpose, we created a position- and gene-specific scoring matrix (PGSSM) and its corresponding antibody similarity score. We characterized our PGSSM score and found that it strongly correlated with the phylogenetic distance of 181,355 Ab sequences from GenBank across 20 species. The most likely human nucleotide back-translation was obtained given only PGSSMs and the amino acid sequence of an Ab achieving a nucleotide sequence recovery of 95.9% and 97.2% for human heavy and light chains, respectively. In conclusion, the scoring of our back-translation is a valuable estimate for the similarity of an Ab sequence to the natural human repertoire. As expected, Ab therapeutic molecules developed from a human source showed a higher similarity to the repertoire than engineered Abs. Thus, the PGSSM metric introduced here can be used to engineer human-like Ab therapeutics.

ARTICLE HISTORY

Received 8 January 2020
Revised 13 April 2020
Accepted 14 April 2020

KEYWORDS



Single nucleotide polymorphism; immunoglobulin variable region; gene rearrangement; antibody diversity; high-throughput nucleotide sequencing; sequence analysis; biostatistics


Introduction

Antibodies (Abs) bind to epitopes on the surface of microbial pathogens like bacteria and viruses. Abs are produced by B lymphocytes that use genetic mechanisms to increase sequence diversity of the expressed repertoire. These genetic mechanisms include recombination of variable (V), diversity (D), and joining (J) gene segments as well as enzymatic modification and addition of non-templated (N) or palindromic (P) nucleotides in the V-D, D-J and V-J junction regions.¹ The variable domain of an antibody is encoded by the three genes (V, D, and J) for heavy chain sequences, and two genes (V, and J) for light chain sequences. The variable domain can further be divided into framework regions (FR) and complementarity determining regions (CDR). The introduction of somatic mutations in the variable domains occurs in recombined genes during the secondary immune responses.^{2,3} The resulting sequence space of the combined set of naïve and mature sequences of the V domain in an individual organism depends on general characteristics of the Ab genes for a species and on the prior experience of the individual including pathogen exposures. We previously determined the immunome (adaptive immunome receptor repertoire) comprising Ab sequences for three healthy human blood donors using very deep next-generation sequencing (NGS).⁴ The Ab

sequences of this dataset either cover the full variable domain or start midway into the FR region.

The analysis of human Ab sequences usually comprises the partitioning into V, D, and J gene-encoded domains, and the determination of the FR and CDR as well as somatic mutations. Various computational tools are available to assign inferred genes and domains to portions of Ab sequences by making species-specific germline gene calls.^{5–10} Germline genes also may vary in individuals and ethnic subgroups, potentially biasing the maturation process in ways that may be of clinical relevance.¹¹ The increasing availability of large immunome datasets^{4,12–15} was leveraged to create a position- and gene-specific scoring matrix (PGSSM) for datasets in order to describe the human Ab sequence space. For this study we used the sequencing dataset from the Soto et al.⁴ dataset composed of the antibody sequencing from the blood compartment of three healthy human donors. The PGSSMs were derived from this dataset and consisted of 326 million unique antibody sequences. The PGSSM was used to model the single nucleotide frequencies (SNFs) per position in the germline gene, allowing us the estimation of similarity of an Ab sequence to a given immunome repertoire collection. SNFs can arise from different sources such as: allelic differences, hypermutation, or sequencing errors. The method

CONTACT Jens Meiler  jens@meilerlab.org  Departments of Chemistry, Stevenson Center, Vanderbilt University, Station B 351822, Room 7330, Nashville, TN 37235, USA

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

developed in this study attempts to capture frequencies caused by hypermutations by grouping all SNFs to their respective germline gene. The size of immune repertoire dataset ensures that any errors that arise from sequencing are minimized.

Our PGSSMs are germline gene-specific¹⁶ for templated regions, and length-dependent for the heavy chain complementarity-determining region three (CDRH3). This approach allows us to model SNFs that exclude insertions, but include non-templated (N) and palindromic (P) nucleotide additions that bracket the CDR3. This feature enables us to derive the nucleotide sequence that maximizes the nucleotide frequencies in the PGSSM model so that the resulting nucleotide has a high human likeness. In this study, we attributed each optimized nucleotide sequence with a score for the variable (V) and joining (J) domain (PGSSM_{VJ}) and characterized the properties of the PGSSM_{VJ}. We show that the PGSSM_{VJ} represents a similarity measure between an amino acid sequence and a given immune repertoire. Thus, the PGSSM_{VJ} could in principle be used to engineer an antibody sequence to make it more human-like in the future.¹⁷

Methods for detecting human-likeness in antibody amino acid sequences support the screening and engineering of antibodies with immunogenic effects, which tend to reduce the efficacy of Abs in a clinical setting. The H-Score method to estimate human-likeness developed by Abhinadan et al. in 2007 was based on pairwise sequence identity calculations.¹⁸ The method evolved by replacing pairwise sequence calculations with Basic Local Alignment Search Tool (BLAST) databases. The resulting T20 score was also derived from a dataset of about 38,700 sequences.¹⁹ To take germline gene family specificity of immunogenic effects into account, the germline gene aware G-Score was developed.²⁰ Seeliger et al.²¹ demonstrated the usefulness of a heuristic scoring function to increase human-likeness and reduce immunogenic effects. The heuristic scoring function is capable of suggesting mutations to reduce immunogenicity and increase human-likeness based on a pairwise probabilistic model.

The Human String Content (HSC) is an alternative method to decrease immunogenic effects by increasing the germline similarity to 9-mer fragments of germline genes in order to reduce the class II MHC binding affinity.²² The HSC has successfully been combined with structure-based antibody design to produce humanized antibodies with high affinity.²³ The methods H-Score, T20 and the heuristic scoring function have been developed from small amino acid sequence datasets of several thousand sequences. Recent advances of deep-learning methods enabled Wollacott et al. to precisely capture human-likeness of antibody sequences using a Long-Short-Term-Memory (LSTM) model trained on 25,000 sequences.²⁴ Human likeness scores are usually derived from small datasets, and are primarily concerned with the question of how to separate human from non-human antibodies instead of developing a sequence model that explains how an Ab can emerge from a repertoire.

In this study, we developed the algorithm IgReconstruct, which draws conclusions about Ab human-likeness that are distinctly different from other methods. Firstly, our method is based on single nucleotide frequencies. Secondly, to estimate the similarity of a target Ab amino acid sequence to a given

repertoire, a germline gene rearrangement tailored to the nucleotide frequency observations made in the repertoire is generated. Thirdly, the target Ab amino acid sequence is back-translated to the nucleotide sequence to allow a fine-grained comparison with the observed immune repertoire nucleotide frequencies. IgReconstruct scales well with large repertoires consisting of hundreds of millions of sequences, and will be useful for computational antibody engineering.

Results

We calculated position- and gene-specific PGSSM matrices (Figure S1) from a publicly available human immunome repertoire of 326 million antibody Ab sequences.⁴ The PGSSM matrices encode the observed single nucleotide frequencies in the repertoire. The PGSSM matrices were used to calculate the PGSSM_{VJ} score (Figure 1, Equation 1) for any given antibody sequence, which essentially represents the similarity of a given antibody sequence to the immunome repertoire. We then curated a set of in total of 181,355 GenBank²⁵ sequences from 20 different species (see Material and Methods for a sequence breakdown by species). To measure the performance of our PGSSM method with an independent dataset, we used the GenBank sequences and estimated the similarity to the human immunome repertoire of 326 million naturally occurring antibody Ab sequences.

Human Likeness was assessed by calculating the Z-Score of the PGSSM_{VJ} score (Equation 2), for which we used the distribution of PGSSM_{VJ} scores of human GenBank sequences as reference. As expected, human GenBank antibody sequences were most similar to the antibody sequences in our human immunome repertoire.

We demonstrated that our statistical PGSSM model captures a human-like antibody sequence space by recovering the human-like nucleotide sequences. We further were able to calculate a score of the V and J gene-encoded regions to quantify the similarity of an antibody sequence to a given immunome repertoire. The PGSSM_{VJ} score is the average of SNFs in the V and J gene-encoded region of the optimized sequence (Equation 1). We successfully used the score to distinguish between human, non-human, and engineered antibodies. We assessed the scores for 475 antibodies in clinical trials or approved by the U.S. Food and Drug Administration (FDA), indicating a high level of human likeness, but distinguishable difference from natural human antibody sequences.

Processing of immune repertoire data and counting SNFs in V, D, J gene-encoded, and CDR3

Our NGS sequence dataset was annotated with IgBLASTn results comprising germline gene alignments (Figure 1, A1). We only considered Ab sequences without sequencing ambiguity that contain nonstandard nucleotide letters. A collection of 196,755,218 heavy chain and 128,815,779 light chain sequences was used to create PGSSMs (325,570,997 in total). The dataset was processed with IgBLASTn and inferred germline gene alignments were assigned. We generated a full-length PGSSM for each of the 287 V_H, 79 V_K, 72 V_L, 37 D, 13 J_H, 9 J_K, and 9 J_L germline gene alleles. In-frame (+open

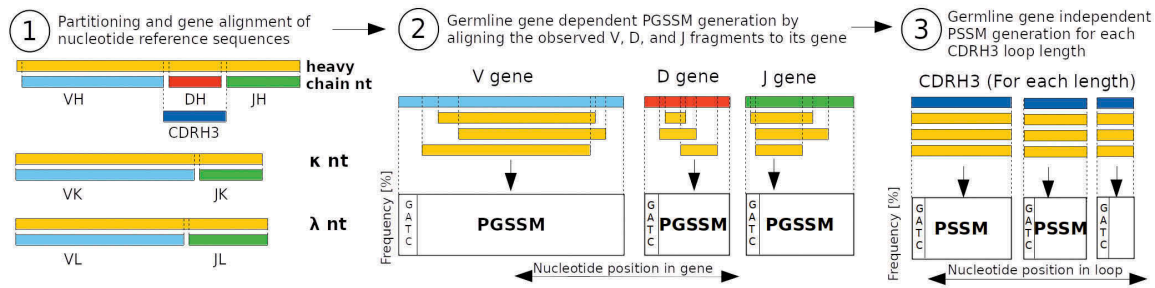
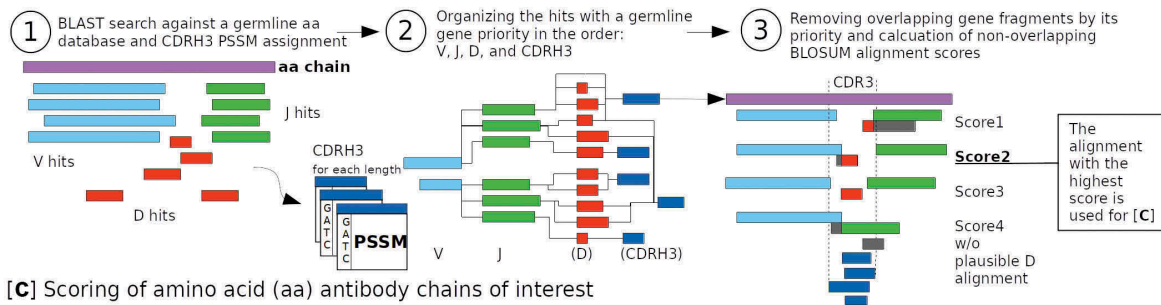
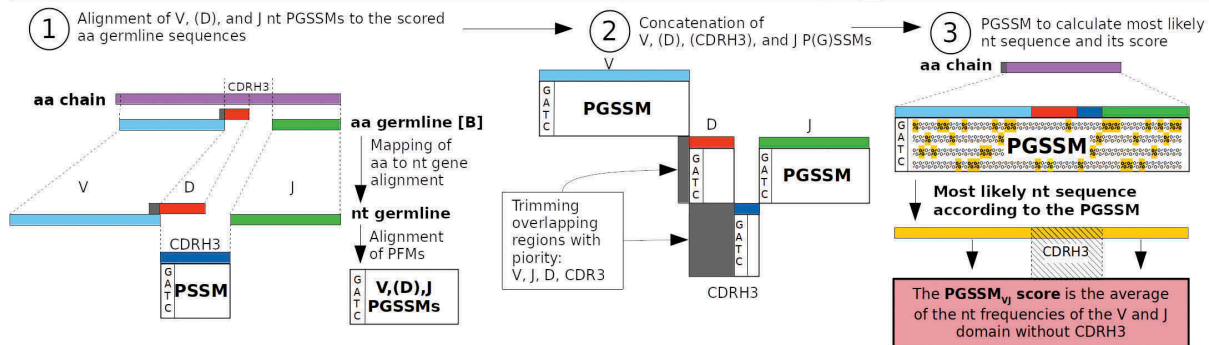
[a] Generation of Position and Gene Specific Frequency Matrices (PGSSM) from an immunome repertoire of nt sequences**[b] Partitioning of amino acid (aa) antibody chains of interest****[c] Scoring of amino acid (aa) antibody chains of interest**

Figure 1. Flowchart of scoring Ab sequences with IgReconstruct. The algorithm can be divided into three tasks (a-c) with three steps (1–3) in each task. (a) The IgReconstruct algorithm starts with the generation of Position and Gene Specific Scoring Matrices (PGSSM) for the variable (V, light blue bars), diversity (D, red bars), joining (J, green bars) and CDR3 (dark blue bars) regions of the Ab nucleotide sequence (yellow bars). In this study, nucleotide sequences were obtained from a large immunome repertoire dataset. (b) For a given amino acid Ab sequence (purple bars), the V, D, and J germline gene rearrangement is determined from the alignment to the PGSSMs by creating a hierarchic tree of aligned nucleotide PSSMs. (c) The highest scoring rearrangement then is mapped to germline gene-dependent V, D, J and germline gene independent CDR3 PSSMs. The resulting nucleotide model is used to determine a back-translation which maximizes the observed nucleotide frequencies in the repertoire. The V and J regions of back-translated sequence is then scored ($PGSSM_{VJ}$) after the observed nucleotide frequencies in the repertoire.

reading frame (ORF)) germline reference sequences that are pre-annotated with CDR and FR start positions were pulled from IMGT/GENE-DB.²⁶ Each of the matrices ultimately contains the frequency of observed G, A, T or C nucleotides for each position in each human germline gene (SNF). Here, we defined the CDR3 sequence as the sequence that starts with the first untemplated position after the V germline gene-encoded alignment and stops one position before the first J germline gene-encoded residue. For each observed heavy chain CDR3 loop (CDRH3) length, we created a germline gene independent PGSSM.

Calculation of PGSSMs from single nucleotide counts

To generate the PGSSMs, we first counted nucleotide observations in each germline gene as well as CDR3 loops. We extracted the V, D, and J gene alignments for each sequence as well as the untemplated region of the CDR3 loops. For some

light chains and heavy chain sequences with high mutation frequency, no unambiguous D gene assignment was possible, whereas V, and J alignments are present for all analyzed sequences. Here, we refer to this D gene segment uncertainty with (D). IgBLASTn generates alignments that contain in some cases overlaps of a few residues between V, (D), and J genes. In this case, we prioritized the alignments in the following descending order: V, J, (D). Each column of a PGSSM matrix corresponds to a nucleotide position in a germline gene. We then incremented either the G, A, T, C or gap cell in each aligned column of the PGSSM, avoiding double counts caused by gene overlaps (Figure 1, A2). We converted the observed counts into frequencies for each column after adding one pseudo-count to each cell, which resembles the SNFs. In addition to germline gene dependent V, D, and J PGSSMs, we generated germline gene independent CDR3 PGSSMs for each observed loop-length in the same manner (Figure 1, A3).

BLAST database generation and searches for creating a plausible amino acid germline gene rearrangement

In order to construct a PGSSM for a given amino acid target Ab sequence, we create a germline gene rearrangement as the first step (Figure 1, B1). For this purpose, we translated all human nucleotide germline genes using the reference sequences in the ImMunoGeneTics information system® (IMGT) database²⁶ in all reading frames, allowing non-productive sequences, and generated separate BLAST databases²⁷ containing V, D, and J genes while not distinguishing between heavy, kappa, or lambda chains. For each target Ab amino acid sequence, our algorithm conducts three independent BLAST searches with e-value thresholds of 20 (V), 100 (D), or 50 (J). The number of alignments was limited to 3 (V), 100 (D), or 10 (J). Word sizes were 4 (V), 2 (D), or 3 (J). BLAST hits were discarded if a stop codon was observed in the aligned region or if a corresponding PGSSM was not available. The length and position of the CDR3 is defined by the V, and J germline gene alignments. For each combination of V, and J BLAST hits, we assigned its distinct CDRH3 PGSSM, which is solely chosen by the length of the non-templated part of the CDRH3.

Assignment of a plausible V(D)J rearrangement for an amino acid target sequence

Our algorithm chooses a plausible V(D)J rearrangement for an amino acid sequence by scoring the combinations of BLAST hits. First, we create a V-J-D-CDRH3 tree hierarchy in the form of a nested data structure for each possible V(D)J alignment (Figure 1, B2). We prevented incorrect alignments from being added to the tree, such as D alignments that were not overlapping with the CDR3, and J alignments not overlapping with the FR4 region. Both regions were calculated for each V germline gene dynamically following the IMGT Unique Numbering scheme,^{28,29} which encodes the positions of FR and CDR as fixed positions in gapped germline genes. The pattern [WF] GXG in the J gene-encoded region marks the end of the CDR3. We also ensured the rearrangements were consistent regarding chain type (heavy, kappa, or lambda).

Second, to choose a final V(D)J rearrangement from the tree, we rescored all recombinations of V, (D), and J alignments after trimming all overlapping regions (Figure 1, B3). We calculated the BLOSUM62 scores for each alignment after pruning the aligned region from overlaps. Overlapping alignments were trimmed or kept with the following descending priority: V, J, D. For example, a D gene alignment overlapping with N residues of a J gene alignment shortens the scoring area of the D gene alignment by N residues. The remaining V(D)J recombinations then were sorted after summing the scores of the individual alignments. We discarded all rearrangements but the one with the highest score. This process does not require D germline gene alignments, since BLAST D germline genes could not be aligned in about 50% of all cases.

It is important to point out, that the germline gene rearrangement tree is individually generated for each antibody and depends on the unique SNF of the repertoire.

A rearrangement in the tree is preferred if a compatible and optional CDRH3 PSSM has been found. A CDRH3 PSSM is compatible if it can bridge the distance between the last aligned V residue and the first J residue. Hence, the chosen V, J, D, CDRH3 rearrangement is dependent on observed CDRH3 lengths in the repertoire.

Creation of the final PGSSM model and scoring of an amino acid target sequence

We used the V(D)J rearrangement chosen earlier and mapped the aligned amino acids corresponding to V, (D) or J genes to their nucleotide counterparts. In addition, we assigned one CDR3 PSSM depending on the length of the loop (Figure 1, C1). We concatenated each V, (D), J and (CDR3) PGSSM such that overlapping parts were discarded. We again respected the domain priority in the descending order V, J, D, CDR3 (Figure 1, C2). Despite the important role of the CDRH3 PSSM for back-translation as well as scoring of the germline gene rearrangement, we chose to not include the untemplated CDRH3 region in the score calculation for two reasons. Firstly, the germline D gene and CDRH3 PSSMs cannot always be assigned. Success depends on the chain type and the availability of CDRH3 PSSMs of a certain length, i.e., the CDRH3 must be observed in the repertoire. Secondly, the CDRH3 PSSM contains all CDRH3 loops of 128,815,779 heavy chain sequences, solely grouped by length. As a result, we do not expect predictive capabilities to the PSSM regarding human-likeness (Figure S3b), even though it supports the generation of a back-translated sequence in this region (Figure S3a).

We therefore restricted calculation of the PGSSM score to V and J PGSSMs, whereas residues without assigned V or J PGSSM remain unscored (Equation 1). Mann-Whitney statistics were used to assess the significance between PGSSM_{VJ} scores of human, non-human Abs and Ab drugs.

To assess the human likeness of the PGSSM_{VJ} score, we calculated the Z-Score using mean and standard deviation of PGSSM_{VJ} scores obtained for all human GenBank antibody sequences separated by heavy or light chain type (Equation 2).

Strategy to reconstruct nucleotide sequences from Ab amino acid sequences

The concatenated nucleotide PGSSM (Figure 1, C2 and Figure S1) aligned and cropped to fit the amino acid target sequence was used to calculate the PGSSM_{VJ} score. Naturally, this approach also can deduce a nucleotide sequence that maximizes the SNFs (Figure 1, C3). Such a nucleotide back-translation is codon-optimized and exhibits the highest possible similarity to the PGSSM and its underlying immune repertoire data. Creating an optimized nucleotide sequence eliminates a potential sequence bias of reported nucleotide sequence and increases the robustness of our method in scenarios where only amino acid sequences are available. This situation occurs frequently in artificial computational protein Ab design in which typically the design process is performed without regard to germline gene rearrangements or nucleotide sequences.^{30,31} The generation of our nucleotide sequence comprises two steps. First, we interrogated for each amino acid the aligned nucleotide PGSSM and

chose the triplet with the smallest hamming distance to the wild-type germline gene. For the untemplated CDRH3, we skipped this step. Second, if multiple triplets after step one are available, we chose the triplet, which maximizes the cumulative SNF.

Figure 1 depicts the complete strategy from amino acid Ab target sequence to nucleotide reconstruction. This method presents per-nucleotide frequency statistics for almost the complete Ab variable domain, including the junction areas of the CDR3 loop and the loop itself. The few exceptions to this assignment are N and C termini without alignments, short light chain junctions, or residues encoded by insertions in the templated regions. Figure S1 shows the complete PGSSM rearrangement of the heavy chain with GenBank accession number EU620063.

The PGSSM_{V_J} acts as a human likeness score in the context of immunomes from healthy humans

We calculated the PGSSM_{V_J} (Equation 1) for all reconstructed nucleotide sequences in the context of three human healthy immunome repertoires (Figure 2b). The scores for human heavy and light sequences were significantly higher with $93.6\% \pm 3.5\%$ (heavy chain) and $93.7\% \pm 2.9\%$ (light chain), respectively, than the scores for other species.

The non-human primates *Callithrix jacchus* ($91.1 \pm 2.2\%/90.9 \pm 2.9\%$), *Chlorocebus sabaeus* ($89.1 \pm 2.4\%/91.5 \pm 2.7\%$) and *Macaca fascicularis* ($89.2 \pm 2.4\%/91.7 \pm 2.1\%$) scored significantly lower with *P* values from a Mann-Whitney test $\square 10^{-7}$. The lowest scoring species include *Gallus gallus* (Red junglefowl) and *Salmo salar* (Atlantic salmon) with $78.6 \pm 1.9\%/82.0 \pm 1.5\%$ and $79.3\% \pm 3.7\%/N.A$ (heavy chain/light chain). The lower bound of PGSSM_{V_J} as well as sequence recovery is constrained by the chance to guess

nucleotides of a fixed amino acid sequence correctly, which is approximately 73.68% (Appendix). Scores around the value of 73.68% are strong indicators for sequence alterations such as engineered sequences.

The PGSSM_{V_J} score can be used to identify engineered and atypical antibodies

Some sequences of the species *Homo sapiens* are outliers in that they score significantly lower than the 95% confidence interval. For Abs annotated with *Mus musculus*, a number of high-scoring outliers outside the 95% confidence interval occurred (Figure S2b). These findings can be attributed to engineered or other non-natural Abs. For the case of *Mus musculus*, sequences often can be associated to studies involving transgenic mice with human Ab loci.^{32–36}

A large number of low scoring human sequences are annotated with patents related to engineering and or animal Ab sources (US20050002930A1, JP2007524605A, EP2150565A2) often directed to human cancer and immune disorder treatments (JP2009221224A, EP2150565A2, WO2005063299A3, WO2004085474A2) like prostate cancer (WO0173032A2, JP2003528591A), or patents evolving in the vicinity of anti-human Abs (WO2005067477A3). Another possible explanation for the low scoring GenBank entries are their annotations designating them as unpublished or having incomplete publication records (e.g., GenBank IDs: EU620060, FW576479, DQ187727). Our observations match previously reported concerns of incorrectly annotated Abs.³⁷

Heavy chain/light chain sequences of structures from the Protein Database (PDB)³⁸ with IDs 1GAF (79.9%/86.3%), 1AXS (80%/83.9%), 1BBJ (81.9%/84.7%) 4UOK (88.0%/82.8%), and 4UOM (80.7%/90.0%) were scored. These PDBs

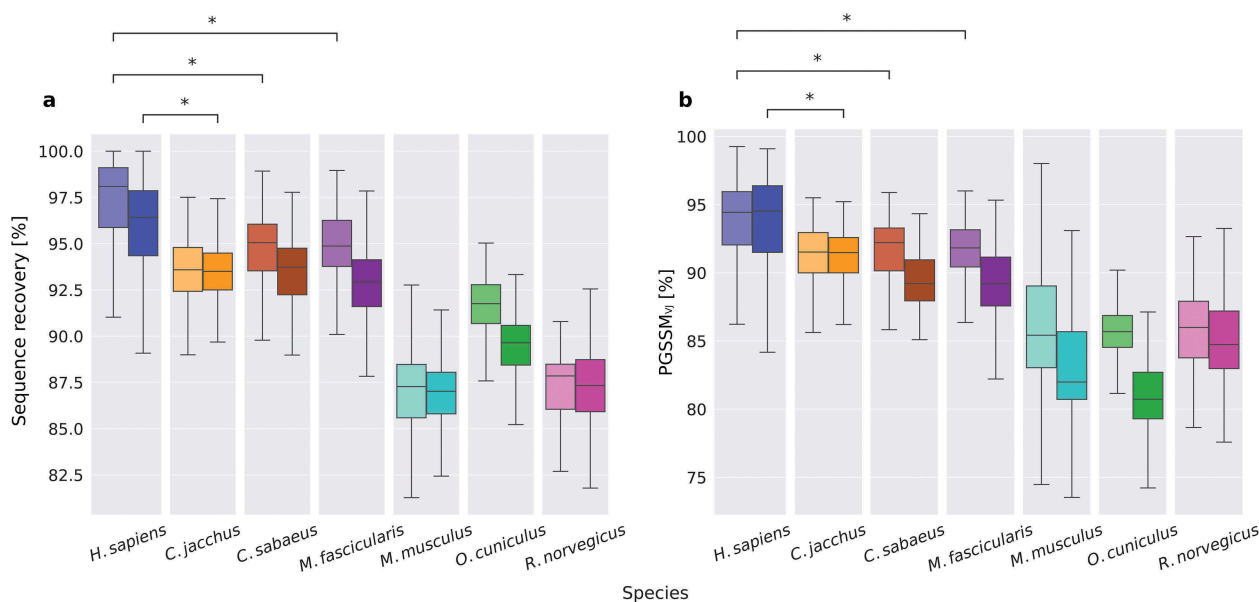


Figure 2. Native nucleotide sequence recovery and PGSSM_{V_J} score for Ab sequences taken from GenBank. Amino acid sequences were downloaded from GenBank²⁵ and then back-translated to nucleotide sequences using IgReconstruct. (a) The sequence recovery rate after back-translation with IgReconstruct is highest for human (*H. sapiens*) sequences when compared to that for sequences from non-human primates (*C. jacchus*, *C. sabaeus*, *M. fascicularis*), mouse (*M. musculus*), rat (*R. norvegicus*) or rabbit (*O. cuniculus*). (b) The PGSSM_{V_J} score for the same set of back-translated nucleotide sequences also scores highest for amino acid sequences derived from humans. Light colors (left bar in each subplot) represent light chain sequences, dark colors (right bar in each subplot) represent heavy chain sequences. A Mann-Whitney test shows statistically significant (*, $p \square 10^{-7}$) recovery rates and scores for human sequences compared to the other species.

were reported previously as incorrectly annotated with human origin.³⁷ The low PGSSM_{VJ} scores (< 1 σ of GenBank sequences assigned as human) also underlines the probable non-human origin of all heavy chains and most light chains.

One shotgun sequenced human light chain of the transcriptome with ORF expressed sequence tags described in 2000³⁹ exhibits two insertions and a region of five deletions, dropping the sequence score to 77.16%. Other examples for sequences with presumably human background but atypical mutation patterns are broadly neutralizing HIV Abs^{40,41} like VRC01 and its derivatives that occurred after long-term lineage evolution.⁴² These highly matured Abs can indicate sensitivity to the progress in sequencing methods. Low-scoring HIV mAbs may highlight the challenge for the human system to generate the right combination of rare mutations against the highly variable sequences of HIV envelope protein.⁴³

Another example of Abs with rare mutations are fetal lymphocyte progenitors,⁴⁴ highly mutated Abs of tonsillar IgD-cells,⁴⁵ or expanded multiple sclerosis associated lineages in immortalized B cells.⁴⁶ Some of these Abs are related to tissue location or to autoimmune diseases, and might therefore not be typical of Abs found circulating in the peripheral blood, which is the current context of our Ab analysis method.

The PGSSM_{VJ} score correlates with the phylogenetic distance to human V germline genes

We further interrogated the PGSSM_{VJ} properties and estimated their correlation with the phylogenetic distances between human and non-human species. The phylogenetic distance was calculated as the sum of the branch length between the two closest germline genes of the same class (heavy, kappa, lambda) of two species. We calculated a phylogenetic tree between the available IMGT reference germline sequences. Nucleotide frequencies in V and J gene-encoded domains are on average low in number and guide the overall sequence space of a species. This germline gene preference of nucleotides is directly captured in the PGSSM frequencies and ultimately in the PGSSM_{VJ} score.

The average PGSSM_{VJ} score for all studied sequences is plotted against the phylogenetic distance from the assigned human V gene to its closest V gene of the organism of origin separately for heavy chain (Figure 3a) and light chain V genes (Figure 3b). GenBank sequences of the species *Mus musculus* are frequently the subject of lineage evolution and of engineering studies, and such sequences exhibit highly artificial mutation patterns, which causes a low correlation between phylogenetic distance and score. We therefore separated *Mus musculus* sequences and highlighted these in red color. The correlation of heavy chains remains less affected due to the higher number of datapoints.

Single nucleotide frequencies in Abs roughly recapitulate phylogenetic distances. One can thus use the PGSSM_{VJ} to confirm or question the Ab species annotation. The PGSSM_{VJ} therefore can be used as a measure of the degree of recombinant engineering with known phylogenetic relations.

PGSSM_{VJ} allows for the recovery of nucleotide sequences for human Abs

We performed a nucleotide sequence recovery benchmark to demonstrate that triplet independent observations of single nucleotide frequencies can approximate the human Ab sequence space. 181,335 GenBank sequences of 20 different species were translated with IgBLASTn.⁵ The nucleotide sequence was optimized by maximizing the PGSSM_{VJ} score.

Back-translation recovery rates peak for human sequences, with an average heavy and light chain recovery of 95.9 \pm 2.6% or 97.2 \pm 2.8%, respectively (Figure 2a, Figure S2a). As expected, when we leveraged the human PGSSM_{VJ} score to determine the most likely human nucleotide sequence for Abs of different species, correct nucleotide identification dropped, labeling these Abs as non-human. For non-human primates, recovery rates were *Callithrix jacchus* (93.5 \pm 1.5%/93.3 \pm 2.2%), *Chlorocebus sabaeus* (93.4 \pm 1.9%/94.5 \pm 2.7%) and *Macaca fascicularis* (92.8 \pm 1.9%/94.7 \pm 1.9%). The lowest scoring species included *Gallus* (Red junglefowl) and *Salmo salar* (Atlantic salmon) with heavy/light chain scores as low as 82.7 \pm 1.1%/82.9 \pm 1.4% and 82.6 \pm 2.2%/N.A. A comparison of PGSSM_{VJ} scores with sequence recovery rates (Figure 2) shows striking similarity, suggesting that the PGSSM_{VJ} score is a predictor of sequence recovery. Figure S2 depicts the similarity of sequence recovery (a) with PGSSM_{VJ} score (b) for all 20 species.

The sequence recovery frequency strongly correlates with the PGSSM_{VJ}

A third property of PGSSM_{VJ} is the ability to estimate the nucleotide sequence recovery rate. We calculated the correlation between average nucleotide mutation frequency (PGSSM_{VJ} score) with the sequence identities determined in our sequence recovery benchmark. The recovered sequence is of importance to determine the minimal distance to its context for Ab-dataset comparisons. With a Mann-Whitney correlation coefficient of $R = 0.92$, $P = 0$ for heavy chains (Figure 3c) and $R = 0.86$, $P = 0$ for light chains (Figure 3d), the PGSSM_{VJ} is approximately the sequence recovery rate for human sequences \pm 5%.

Ab therapeutics in context of the Ab repertoire of healthy humans

We used 475 unique Abs that are either approved by the U.S. FDA or are in clinical trials.^{47,48} All biologics were either annotated with the INN designations⁴⁹ HU, ZU, XI, and XIZU as reported by Jain et al.⁴⁷ or annotated with Human, Humanized, Chimeric, and Mouse in case of antibodies taken from IMGT/mAb-DB.⁴⁸ For this study, we chose appropriate labels for HU (Human), ZU (Humanized), XI (Chimeric), and XIZU (Humanized Chimeric Hybrid) to match the designations used in IMGT/mAb-DB. The sequences were treated the same way independent from its labeling in the algorithm. We investigated the Ab sequences in the context of our three individual immunome repertoires and in the context of one large merged repertoire. For Z-Score calculation, mean and

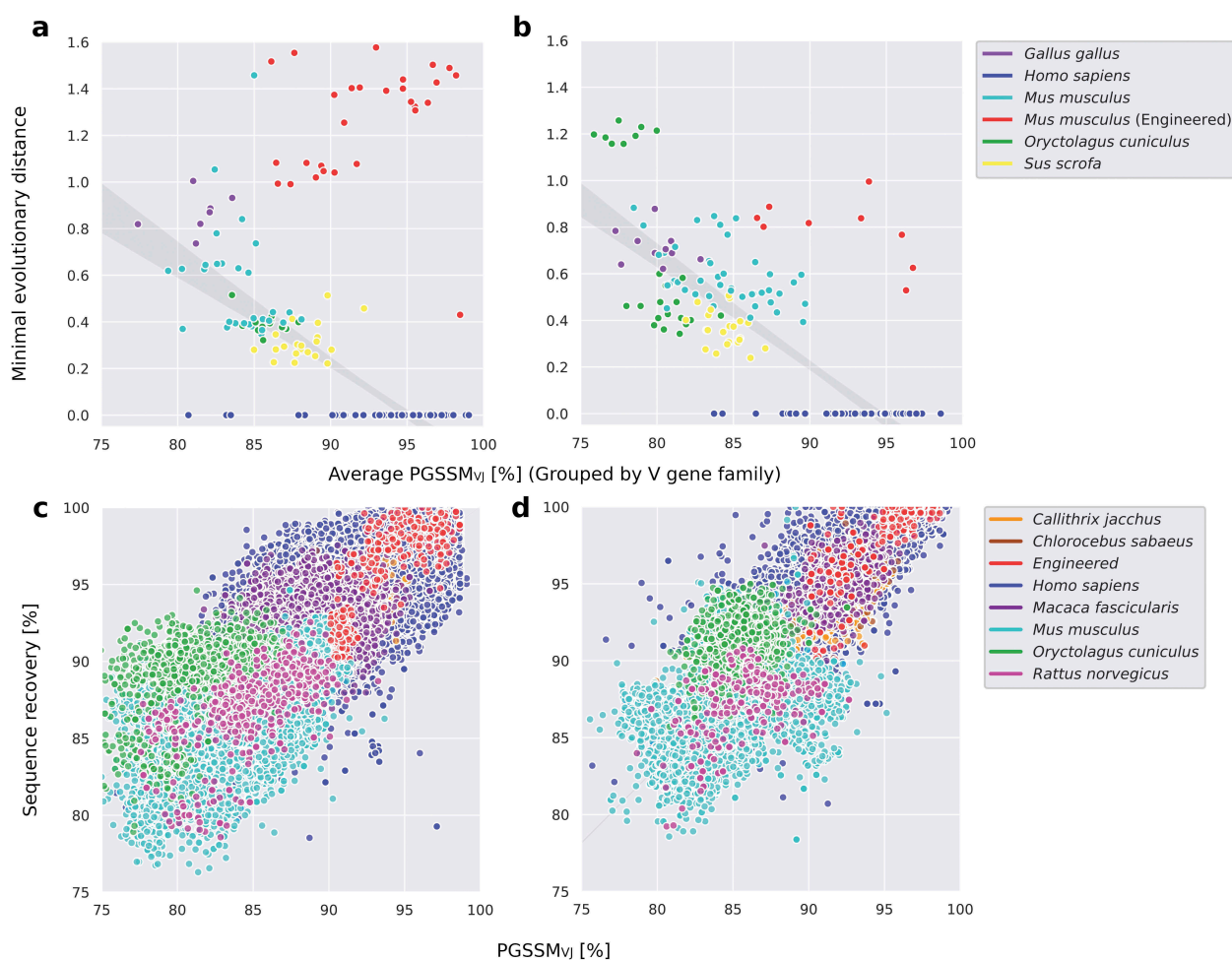


Figure 3. The PGSSM_{VJ} score approximates the evolutionary distance from human immunoglobulin germline genes to immunoglobulin germline genes belonging to 20 species. Amino acid sequences were downloaded from GenBank²⁵ and then back-translated to nucleotide sequences using IgReconstruct. (a) The average PGSSM_{VJ} scores for heavy chain Ab sequences or (b) light chain Ab sequences are plotted against the phylogenetic distance from the assigned human germline gene using IgReconstruct (see Methods section for details). The PGSSM_{VJ} scores correlate with the phylogenetic distance with a Spearman rank correlation coefficient of $\rho = -0.83$ ($P = 2e-41$, $\alpha = 0.01$) for heavy chain Ab sequences and $\rho = -0.83$ ($P = 2e-37$, $\alpha = 0.01$) for light chains Ab sequences. (c) Sequence recovery between native heavy chain sequences and back-translated nucleotide sequences, made using IgReconstruct, gave a Spearman rank correlation coefficient of $\rho = 0.92$ ($P = 0$, $\alpha = 0.01$). (d) Sequence recovery between native light chain sequences and back-translated nucleotide sequences using IgReconstruct gave a Mann-Whitney correlation coefficient of $\rho = 0.86$ ($P = 0$, $\alpha = 0.01$). Mouse (*M. musculus*) Abs engineered to be human-like are colored red (top right corner of subplot a and b).

standard deviation (σ) from GenBank sequences (Figure 2b) were used (Equation 2).

We compared the Z-Score of PGSSM_{VJ} either grouped by clinical stage (Figure 6) or source subsystem, which indicates the origin and type of engineering of the biologics (Figure 5).⁴⁹ Drugs with a human source scored highly similar to GenBank sequences (Z-Score around 0), followed by humanized, chimeric and murine Abs. This trend was consistent for both drug datasets processed. Scores of sequences from mice still score in a similar range of GenBank *Mus musculus* sequences. This finding shows that antibody sequences from IMGT/mAb-DB with a murine background remain distinguishable from biologics with human origin. On the other hand, humanized and chimeric sequences populate a scoring range closer to human and non-human primate sequences. Pooling drugs by their clinical status shows that drugs in Phase 2 to 3 clinical trials and approved Abs have an average Z-score of -0.56 ± 1.05 (Phase 2), -0.77 ± 1.35 (Phase 3), and -1.18 ± 1.45 (Approved). On average, human drugs appear human-like with a Z-Score greater than -2 , caused by the

high number of human (57) and humanized (68) drugs compared to 13 chimeric. The low number of available sequences aggravates the challenge to draw reliable conclusions. The PGSSM_{VJ} indicates that there is a non-human sequence space compatible with the human system. However, we hereby choose a Z-Score cutoff of -2 or greater to roughly group the majority of clinical stage antibodies (Figure 6, horizontal red line). For our next experiment, we used this cutoff to distinguish between biologics/human antibodies, and non-human antibodies.

To further investigate the role of public and private repertoires on the eligibility of Abs as drugs, we calculated PGSSM_{VJ} scores using each of the three individual immunome repertoires. The majority of staged antibodies exhibit a cutoff of -2 or greater (Figure 6). Hence, we roughly defined any of the three scores as human-like as long as the Z-Score of the PGSSM_{VJ} was greater or equal to -2 . Figure 7 depicts the number of human-like scores for non-human (orange), human GenBank Abs (blue), and biologics (green), separated by light chains (a) and heavy chains (b). We observed high

agreement between the three scores for human and therapeutic Abs. We also observed high agreement rates between all three repertoires, including 70.0% of all biologics and 92.3% of all human GenBank heavy chain sequences and 81.8% of all biologics and 94.6% of all human GenBank light chain sequences. In contrast only 8.8% light chain and 8.8% heavy chain sequences of biologics and 1.3% of light chain biologics and 2.6% of heavy chain human GenBank sequences were scored as non-human in all three cases.

Performance and robustness

The initial release of our algorithm requires amino acid Ab sequences that cover at least a fraction of the V and J gene-encoded region, which can be successfully aligned via BLAST. The algorithm then places optional D PGSSMs as well germline gene CDR3 loop PGSSMs in the appropriate locations if available. Templated regions as well CDR3 junctions are modeled statistically; insertions are represented in the statistical SNF model as gaps.

We compared the germline gene families with the top five germline gene families assigned by IgBLASTp, the IgBLAST tool for protein sequences (Table 1). Our method reliably assigns germline V genes to our sequences when IgBLASTp is taken as reference.

Output

We provide a webservice called IgReconstruct (<http://meilerlab.org/index.php/servers/IgReconstruct>), which takes amino acid sequences of Ab variable domain in FASTA format as input. The output is presented graphically in a downloadable PDF file (Figure 4), and a spreadsheet with equivalent machine-readable information. The PDF report presents the query amino acid sequence aligned to its reconstructed nucleotide sequence, V, (D), and J germline gene alignments. The germline gene alignments indicate sequence identity with a dot and residue type replacements with a one-letter code. The variable region is annotated in the form of branches for the predicted IMGT-CDR1-3. V(D)J domains are colored blue, red, and green and match the colors used in the IgReconstruct flowchart (Figure 1). In case of overlapping alignments, the region is colored according to the hierarchy of the rearrangement tree.

Discussion

We have shown that statistics of SNFs of the variable region using large human immunome repertoires are capable of modeling the human Ab sequence space by predicting nucleotide sequences from amino acid sequences (Figure 2). With

more and more large NGS nucleotide sequence datasets becoming publicly available,^{4,12–15} IgReconstruct resembles an approach to link the nucleotide sequence space with resources of Abs where primarily amino acid information is available, like de-novo computational models or structural databases.^{38,50} Approaches of structural modeling of Abs³⁰ have been made to include amino acid sequence profiles of V and CDR3. IgReconstruct may pave the way to completely model the germline gene rearrangement of an amino acid sequence at the nucleotide level and provide full access to large-scale human immunome repertoire statistics.

We demonstrated that the PGSSM_{VJ} score, derived from the SNF statistics of an individual Ab, is an appropriate distance measure of a particular chosen Ab to a nucleotide immunome repertoire or arbitrary large set of sequences (context). For this, we fulfilled the requirement to find the minimal distance by suggesting the most probable nucleotide sequence for a given repertoire (context-dependent). The PGSSM_{VJ} then can be used to estimate the likelihood to observe a context-dependent nucleotide sequence in the dataset. Finally, the PGSSM_{VJ} strongly correlates with the phylogenetic distance between human and non-human germline genes (Figure 3). These combined properties allowed us to estimate the similarity of a variable domain to a dataset and to interpret it as a distance value. For example, further studies might conclude that infections like HIV exhibit a greater distance to the human sequence space, which results in less effective immune responses.

A current shortcoming of our method is that our CDRH3 statistics, which include the heavy chain junctions, are only length dependent. As a result, the major domain that diversifies an immunome repertoire^{51–53} is merged into relatively small bins, disregarding the sequence similarity and function. As a result, our PGSSM_{VJ} score is currently exclusively calculated from V and J gene templated regions. We do not anticipate or observe sufficient performance using solely CDRH3 PSSMs to distinguish between non-human, human, and biologics only using CDRH3 sequences due to high variability (Figure S4). However, CDRH3 PSSMs can be used to support the back-translation of amino acid sequences to nucleotides (Figure S3).

We evaluated Ab sequences from 20 species and were able to distinguish sequence origins between human primates, non-human primates and other species reliably. While doing this, we found that the prior species annotation in deposited sequences often was not reliable. The signal that allows us to distinguish between human vs. non-human persisted while studying the IgReconstruct results of clinical-stage and FDA-approved Abs (Figure 5). A non-human source could reliably be detected in murine, chimeric, humanized chimeric and humanized Abs. Due to the higher count of therapeutic Abs

Table 1. Recovery frequency of germline gene families for each species in the dataset. We display the frequency with which the germline gene rearrangement families of our algorithm can be found in the top five IgBLASTp hits.

Antibody gene type	Agreement of inferred V germline gene family between IgReconstruct and IgBLASTp for indicated species							
	Homo sapiens	Callithrix. jacchus	Chlorobaeus. sabaeus	Macaca fascicularis	Mus musculus	Oryctolagus cuniculus	Rattus norvegicus	
IGHV	96.4	98.5	97.6	93.5	94.4	83.7	93.2	
IGKV	98.3	98.5	97.8	97.4	92.7	94.3	94.5	
IGLV	96.1	96.8	92.6	97.1	99.7	99.7	96.0	

IgReconstruct PDF Report

User: samschmitz || Date: 07-28-19 00:29

Version: ABL/341310e7b2

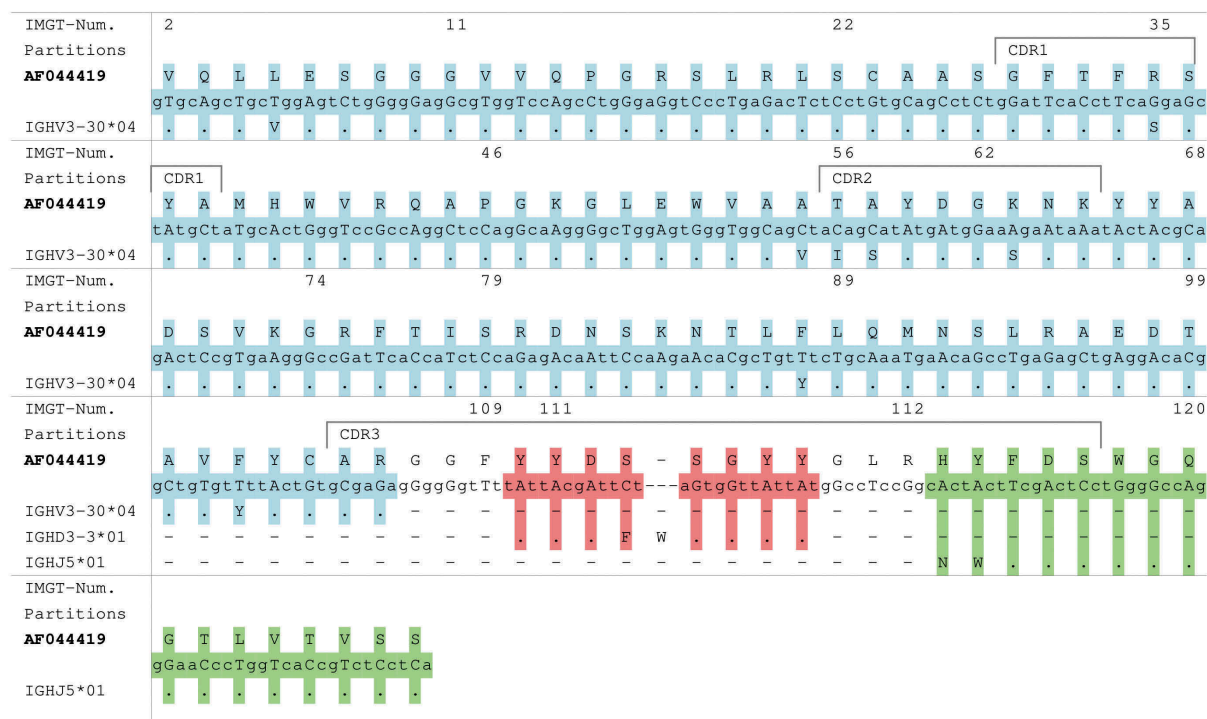


Figure 4. Alignment report generated by IgReconstruct. An example alignment report for the human heavy chain Ab sequence with the GenBank accession number AF044419. Reports generated by IgReconstruct provide information on the query amino acid sequence (first row), the back-translation (second row) and alignments to the germline gene sequences (third and following row if applicable). The color code blue (V gene), red (D gene), and green (J gene) refers to the aligned germline PSSMs which were used to create the back-translated sequence. Columns without color are not aligned to a specific germline gene. Dots represent the germline sequence; mutations are shown using the one-letter amino acid code. CDR loops 1 to 3 are inferred based on alignments to the V and J germline genes. The numbers on top of the amino acid sequence was implemented using the IMGT numbering scheme.²⁸ Non-templated regions at the V-D and D-J junctions flanking the D gene alignment (red) are covered by the CDRH3 PSSM, but are not visualized in the color scheme. The PDF report gives a quick insight into the nature of the germline gene rearrangement which is used to generate the back-translation and the human-likeness score.

with a human sequence background, the combined population of sequences scores at the lower end of “human-like” (Figure 6). A more comprehensive therapeutic Ab and immune repertoire relationship might be developed in the future, when our statistical Ab model incorporates a more sophisticated CDRH3 model. The results indicate that there is a non-human sequence space, which is compatible with human biology (i.e., is associated with a manageable frequency of adverse effects). Abs from that space can be used as therapeutics. These sequences remain unlike the repertoire in our study with low human likeness scores, despite humanization efforts. However, the majority of Z-Scores of antibody biologics in clinical phases appears to be -2 or greater (red horizontal line). For our next experiment, we used this cutoff to distinguish between biologics/human antibodies, and non-human antibodies.

Krawczyk et al. used amino acid alignments of variable and CDR regions to show that sequences with high similarity to therapeutic Abs can emerge in the human antibody repertoire, whereas chimeric and humanized antibodies tend to be slightly more dissimilar.⁵⁴ This observation could be reproduced using SNFs mapped onto germline genes instead of amino acid sequence alignments. In addition, a Z-score cutoff

of -2 was chosen, which enables us to separate between non-human and human as well as biologics. The ability to separate drugs from non-human antibodies is hypothesized to support antibody drug development in the future.

The human-likeness score in this study is distinctly different from previously published methods, where typically the ability of the separation of real human and non-human sequences was being maximized. Recent advances in deep-learning have shown excellent classification capabilities.²⁴ Here, we devised a method that generates a nucleotide frequency model based on repertoire observations, which represents the plausibility that an Ab sequence arises from a particular repertoire. The results of a previous study could be confirmed, which has shown that biologics can be distinguished from human sequences.⁵⁴ On the one hand, this study does not aim to maximize the separation between truly human and non-human sequences, resulting in less clear boundaries between human and, for example, macaque sequences.²⁰ On the other hand, the approach could hypothetically be used to capture the biologically relevant question of the immunogenicity of an Ab, which cannot strictly be answered by separating human from non-human sequences. Consequently, a slightly worse separation performance

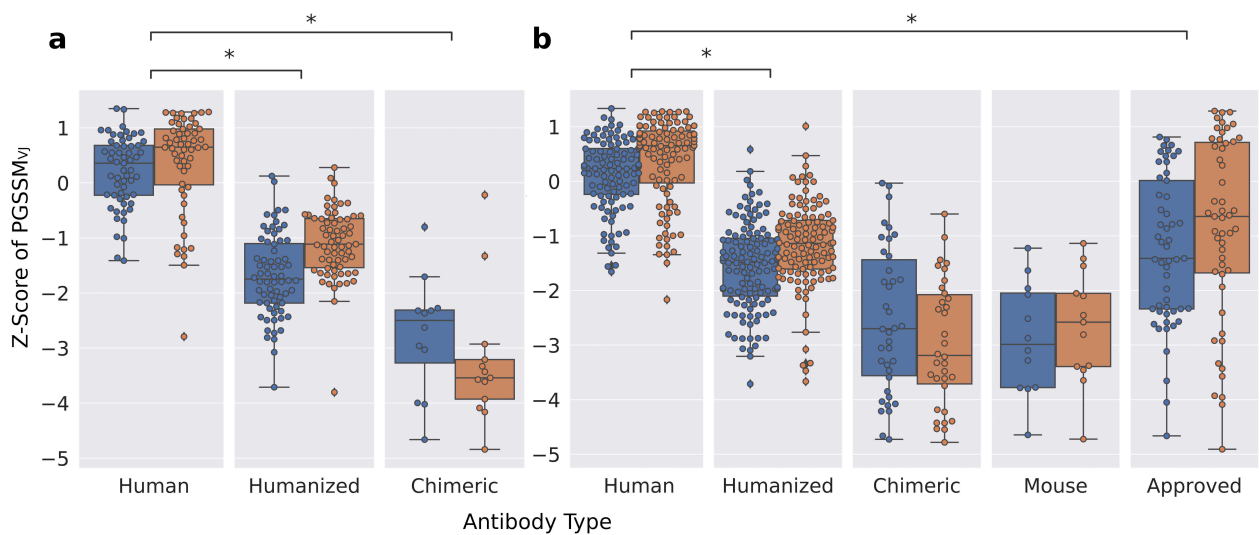


Figure 5. The PGSSM_{VJ} score ranks human Abs highest when compared to either chimeric or mouse Abs used as biologics. Ab sequences for biologics were obtained from IMGT/mAb-DB⁴⁸ separated by heavy chain (blue) and light chain (orange). All PGSSM_{VJ} scores were transformed into Z-scores and ranked within each group. (a) Biologics analyzed from the Jain *et al.*⁴⁷ study show that human Abs rank highest when compared to either chimeric or mouse Abs. Humanized Abs also rank higher than either chimeric or mouse Abs. (b) Biologics from the IMGT monoclonal Ab database show a similar picture, with human sequences scoring higher than biologics with a non-human origin. Approved Biologics are distinguishable from human antibodies. Mann-Whitney significance tests show statistical significance ($p \leq 10^{-7}$) and are labeled with a star (*).

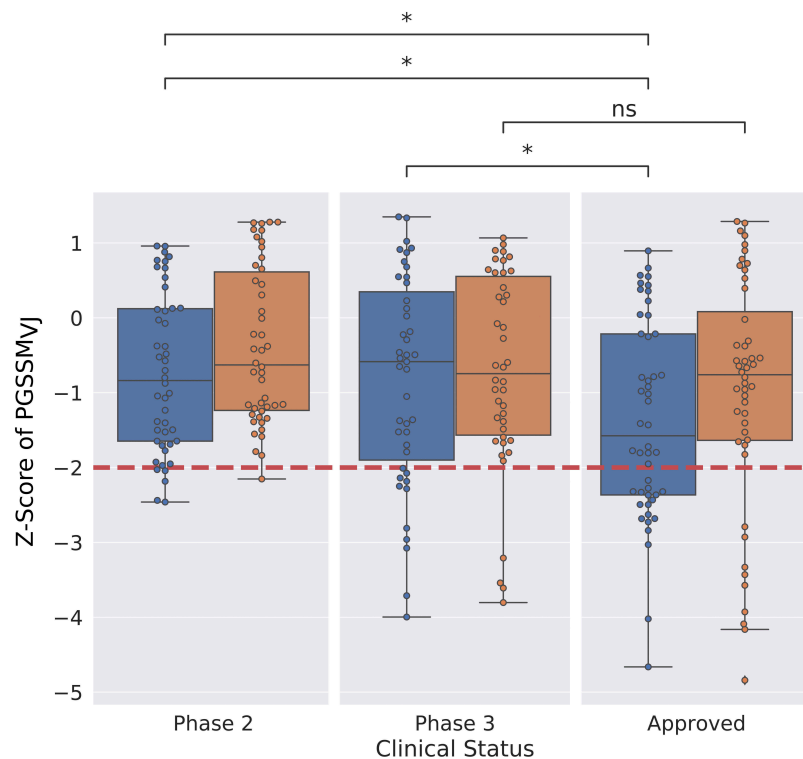


Figure 6. PGSSM_{VJ} score cannot discriminate between clinical stage and FDA-approved biologics. The Z-Scores of heavy chains (blue) and light chains (orange) were calculated using the distribution of GenBank sequences annotated as human. PGSSM_{VJ} scores of biologics from Jain *et al.*⁴⁷ grouped by their clinical phase, show an overall picture of human-like sequences (within one standard deviation of human GenBank sequences) and a smaller population of low scoring sequences. A Mann-Whitney test between clinical trial Phase 2, 3 and FDA-approved Abs revealed no significance (ns) to very weak statistical significance ($p < 5 \times 10^{-2}$, *).

compared to the deep-learning approach of Wollacott *et al.* could be observed with an Area Under the Curve (AUC) of 0.94 compared to 0.97 (Figure S6). At the same time, IgReconstruct is able to leverage the substantial sizes of the largest repertoires with hundreds of million to billion sequences like the Observed Antibody Space¹⁵ by using

nucleotide germline gene rearrangements as reference, as opposed to using smaller datasets in the ranges of ten thousands of sequences of previous methods.^{18–21,24,19} IgReconstruct provides an alternative to extrapolating sequence landscapes from a small representative set of sequences in favor of leveraging large repertoires to its full

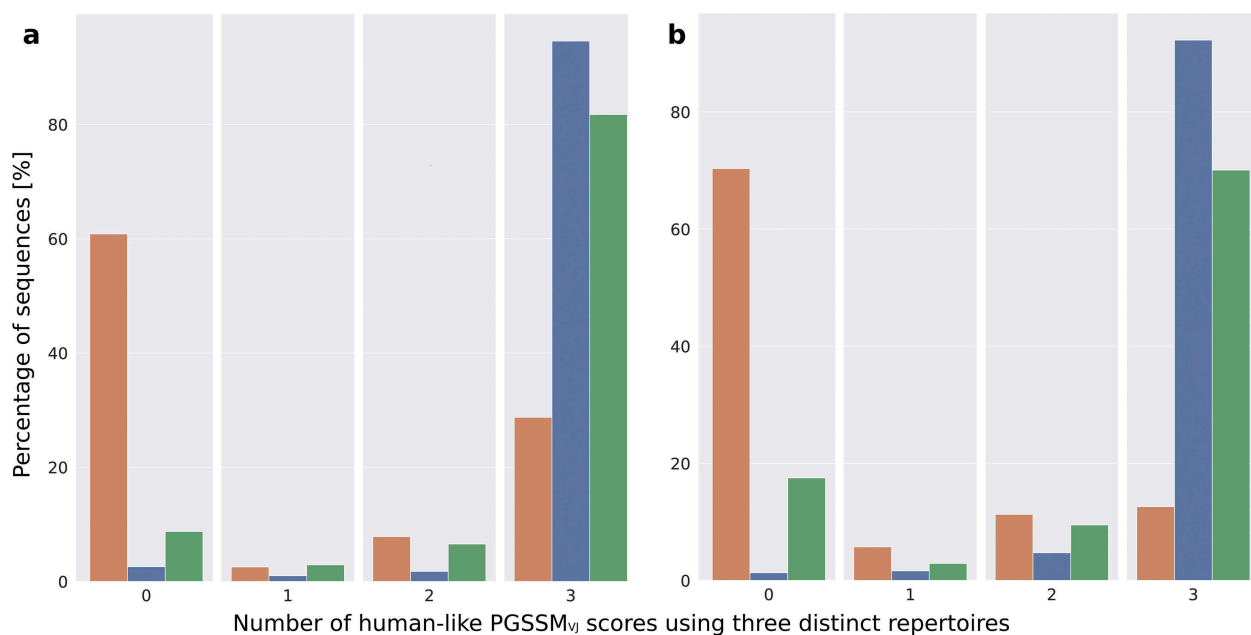


Figure 7. Scoring medically relevant Abs using sequencing data from three individual human immunome repertoires. PGSSM_{VJ} scores of biologics (Jain et al.⁴⁷) versus human and non-human Ab sequences from GenBank. All Ab sequences were scored using sequencing data from three separate immunome repertoires.⁴ Z-Scores of the PGSSM_{VJ} was calculated using GenBank sequences annotated as human. A binary score was used to indicate if an amino acid sequence was human-like. A score of 1 indicates a human-like sequence with a Z-score of -2 or greater. A score of 0 indicates a non-human-like sequence with a Z-score less than -2 (see human data in Figure 6 for Z-score cutoff value). Each sequence was scored against each repertoire and summed up. Thus, a maximum number of three scores can be achieved for any individual sequence which signifies that the sequence is human-like according to comparison with all three individual repertoires. Using the cutoff of -2 allows to roughly separate between non-human (orange), human (blue) GenBank sequences and biologics (green). In case of light chains (a) the cutoff of -2 classifies a larger amount ($\sim 30\%$) of non-human antibodies as human than in the case of heavy chains ($\sim 12\%$) (b).

extent. The definition of human-likeness in this study is a novel approach with the potential to support Ab engineering and explain immunogenic effects in future studies.

Materials and methods

We developed the PGSSM method and supplementary tools for repertoire processing in Python-3.7.1. We provide a webserver that generates germline gene rearrangements for amino acid Ab sequences in text or PDF form, and numeric information in a spreadsheet format.

Curation of sequences from three sources

We curated a set of 181,355 Ab sequence from GenBank.²⁵ 119,827 heavy chains Ab were from the following species: *Bos indicus* (5), *Bos taurus* (1,520), *Callithrix jacchus* (328), *Camelus dromedarius* (388), *Canis lupus familiaris* (253), *Chlorocebus sabaeus* (82), *Equus caballus* (427), *Felis catus* (94), *Gallus gallus* (157), *Homo sapiens* (76,728), *Lama glama* (499), *Macaca fascicularis* (3,592), *Mus musculus* (27,863), *Oryctolagus cuniculus* (1,253), *Ovis aries* (1719), *Rattus norvegicus* (544), *Salmo salar* (109), *Sus scrofa* (4029), *Vicugna pacos* (237). 61,528 light chain sequences were from the species *Anas platyrhynchos* (298), *Bos indicus* (191), *Bos taurus* (353), *Callithrix jacchus* (874), *Camelus dromedarius* (32), *Canis lupus familiaris* (417), *Chlorocebus sabaeus* (74), *Equus caballus* (319), *Felis catus* (76), *Gallus* (301), *Homo sapiens* (41,347), *Lama glama* (15), *Macaca fascicularis* (673), *Mus musculus* (13,249), *Oryctolagus*

cuniculus (1,099), *Ovis aries* (583), *Rattus norvegicus* (299), and *Sus scrofa* (1,328). We applied our method on the translated variable domains reported by IgBLASTn. To estimate the performance, we calculated the nucleotide sequence identity of the complete variable region and compared the germline gene families assigned with our method with the results from IgBLASTp for protein sequences.

In addition to GenBank, we used a dataset of 137 Ab drugs⁴⁷ and extracted 382 Abs for clinical use from IMGT/mAb-DB.⁴⁸ In total, we had sequences for 475 unique Ab drugs available for analysis.

Calculation of PGSSM_{VJ} scores and assessment of human-likeness

We developed a method that creates position- and gene-dependent scoring matrices for a given immunome repertoire (Figure 1). Our PGSSM_{VJ} score assesses the similarity of any given amino acid antibody sequence to the repertoire by averaging the observed single nucleotide frequencies over the Ab V and J gene-encoded regions. The single nucleotide frequencies were looked up in the PGSSM matrix that was generated for each antibody individually (Figure S1). Equation 1 was used to calculate the similarity score using a specific sequence and PGSSM matrix.

Equation 1 Calculation of the PGSSM_{VJ} score for the variable and joining region calculated as an average of the observed single nucleotide frequencies

$$PGSSM_{VJ} = \sum_{resi}^N PGSSM_{VJ}(resi, resn) / N$$

N: = Sequence Length

resi: = Residue Position *i*

resn: = Residue Type at position $i \in \{G, A, T, C\}$

$PGSSM_{VJ}(\text{resi}, \text{resn})$: = Observed frequency of nucleotide resn at position resi if aligned to V or J

The Z-Score of the $PGSSM_{VJ}$ score was used to estimate the human likeness of an antibody. For Z-Score calculation, we used the average and standard deviation of $PGSSM_{VJ}$ scores we calculated for 76,728 human GenBank sequences (Equation 2). We also defined an antibody as human-like as long as its Z-Score was -2 or greater.

Equation 2. Z-Score calculation to assess the human likeness of $PGSSM_{VJ}$ scores using average and standard distributions from human GenBank antibodies

$$Z = (PGSSM_{VJ} - \mu) / \sigma$$

μ : = Mean of $PGSSM_{VJ}$ scores of human GenBank sequences

σ : = Standard deviation of $PGSSM_{VJ}$ scores of human GenBank sequences

Phylogenetic tree construction and the evolutionary distance of germline genes

To characterize the $PGSSM_{VJ}$ score, we correlated scores of 20 species with the phylogenetic distance to human germline genes. For this purpose, we constructed a phylogenetic tree from the complete set of IMGT reference sequences⁵⁵ of all species available using the program PhyML.⁵⁶ For each human V germline gene allele, we calculated the minimal phylogenetic distance to each genus of the same chain class (heavy, lambda, kappa) by summing up the branch lengths of the closest path. We averaged the sequence recovery rate and $PGSSM_{VJ}$ score for each germline gene in the tree.

Availability

IgReconstruct is available as a webservice, hosted by Meiler Lab with no restrictions for sequence files up to 4 MB. (<http://meilerlab.org/index.php/servers/IgReconstruct>)

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This work was supported by National Institutes of Health under grant numbers U19 AI117905, U01 AI150739, R01 AI141661 and supported through the Vanderbilt Trans-Institutional Program (TIPs) “Integrating Structural Biology with Big Data for next Generation Vaccines”.

ORCID

Samuel Schmitz  <http://orcid.org/0000-0001-5314-6095>

Cinque Soto  <http://orcid.org/0000-0002-3997-6217>

James E. Crowe  <http://orcid.org/0000-0002-0049-1079>

Jens Meiler  <http://orcid.org/0000-0001-8945-193X>

References

- Jung D, Alt FW. Unraveling V(D)J recombination; insights into gene regulation. *Cell*. 2004;116(2):299–311. doi:10.1016/S0092-8674(04)00039-X.
- Jung D, Giallourakis C, Mostoslavsky R, Alt FW. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol*. 2006;24:541–70. doi:10.1146/annurev.immunol.23.021704.115830.
- Murphy K. Chapter 5, Janeway's Immunobiology. 8th ed. Garland Science; 2012: 116–9.
- Soto C, Robin GB, Andre B, et al. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*. 2019;566(7744):398. doi:10.1038/s41586-019-0934-8.
- Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013;41(W1):W34–40. doi:10.1093/nar/gkt382.
- Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*. 2015;12(5):380. doi:10.1038/nmeth.3364.
- Russ DE, Ho K-Y, Longo NS. HTJoinSolver: human immunoglobulin VDJ partitioning using approximate dynamic programming constrained by conserved motifs. *BMC Bioinformatics*. 2015;16(1):170. doi:10.1186/s12859-015-0589-x.
- Liu XH, Jian Z, Jun SL, Bo L, Xiaole S. Evaluation of immune repertoire inference methods from RNA-seq data. *Nat Biotechnol*. 2018;36(11):1034. doi:10.1038/nbt.4294.
- Brochet X, Lefranc MP, Giudicelli V. IGMT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*. 2008;36(Web Server issue):W503–8. doi:10.1093/nar/gkn316.
- Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*. 2007;23(13):1580–87. doi:10.1093/bioinformatics/btm147.
- Brovkina OI, Shigapova L, Chudakova DA, Gordiev MG, Enikeev RF, Druzhkov MO, Khodyrev DS, Shagimardanova EI, Nikitin AG, Gusev OA, et al. The ethnic-specific spectrum of germline nucleotide variants in DNA damage response and repair genes in hereditary breast and ovarian cancer patients of tatar descent. *Front Oncol*. 2018;8(421). doi:10.3389/fonc.2018.00421.
- DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One*. 2016;11(8):e0160853. doi:10.1371/journal.pone.0160853.
- Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*. 2019;566(7744):393. doi:10.1038/s41586-019-0879-y.
- Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, Knoetze N, Breden FMW, Christley S, Scott JK, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev*. 2018;284(1):24–41. doi:10.1111/imr.12666.
- Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *J Immunol*. 2018;201(8):2502. doi:10.4049/jimmunol.1800708.
- Sheng Z, Schramm CA, Kong R, Mullikin JC, Mascola JR, Kwong PD, Shapiro L. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol*. 2017;8:537. doi:10.3389/fimmu.2017.00537.
- Olimpieri PP, Marcatili P, Tramontano A. Tabhu: tools for antibody humanization. *Bioinformatics*. 2015;31(3):434–35. doi:10.1093/bioinformatics/btu667.

18. Abhinandan KR, Martin AC. Analyzing the “degree of human-ness” of antibody sequences. *J Mol Biol.* 2007;369(3):852–62. doi:10.1016/j.jmb.2007.02.100.
19. Gao SH, Huang K, Tu H, Adler AS. Monoclonal antibody human-ness score and its applications. *BMC Biotechnol.* 2013;13(1):55. doi:10.1186/1472-6750-13-55.
20. Thullier P, Huish O, Pelat T, Martin ACR. The humanness of macaque antibody sequences. *J Mol Biol.* 2010;396(5):1439–50. doi:10.1016/j.jmb.2009.12.041.
21. Seeliger D, Borg J, Stricher F, Nys R, Rousseau F. Development of scoring functions for antibody sequence assessment and optimization. *PLoS One.* 2013;8:e76909. doi:10.1371/journal.pone.0076909.
22. Lazar GA, Desjarlais JR, Jacinto J, Karki S, Hammond PW. A molecular immunology approach to antibody humanization and functional optimization. *Mol Immunol.* 2007;44(8):1986–98. doi:10.1016/j.molimm.2006.09.029.
23. Choi Y, Hua C, Sentman CL, Ackerman ME, Bailey-Kellogg C. Antibody humanization by structure-based computational protein design. *mAbs.* 2015;7(6):1045–57. doi:10.1080/19420862.2015.1076600.
24. Wollacott AM, Xue C, Qin Q, Hua J, Bohnuud T, Viswanathan K, Kolachalama VB. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Eng Des Sel.* 2019. doi:10.1093/protein/gzz031.
25. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2012;41(D1):D36–42. doi:10.1093/nar/gks1195.
26. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 2005;33(Database issue):D256–61. doi:10.1093/nar/gki010.
27. Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu YK. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* 2009;37(3):815–24. doi:10.1093/nar/gkn981.
28. Lefranc M-P, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol.* 2003;27(1):55–77. doi:10.1016/S0145-305X(02)00039-3.
29. Lefranc MP. Unique database numbering system for immunogenetic analysis. *Immunol Today.* 1997;18(11):509. doi:10.1016/S0167-5699(97)01163-8.
30. Adolf-Bryfogle J, Kalyuzhnyi O, Kubitz M, Weitzner BD, Hu X, Adachi Y, Schief WR, Dunbrack RL. RosettaAntibodyDesign (RABD): a general framework for computational antibody design. *PLoS Comput Biol.* 2018;14(4):e1006112. doi:10.1371/journal.pcbi.1006112.
31. Sircar A, Kim ET, Gray JJ. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* 2009;37(Web Server):W474–9. doi:10.1093/nar/gkp387.
32. Sok D, Briney B, Jardine JG, Kulp DW, Menis S, Pauthner M, Wood A, Lee E-C, Le KM, Jones M, et al. Priming HIV-1 broadly neutralizing antibody precursors in human Ig loci transgenic mice. *Science.* 2016;353(6307):1557–60. doi:10.1126/science.aah3945.
33. Longo NS, Rogosch T, Zemlin M, Zouali M, Lipsky PE. Mechanisms that shape human antibody repertoire development in mice transgenic for human Ig H and L chain loci. *J Immunol.* 2017;198(10):3963–77. doi:10.4049/jimmunol.1700133.
34. Suarez E, Magadan S, Sanjuan I, Valladares M, Molina A, Gambon F, Diazspada F, Gonzalezfernandez A. Rearrangement of only one human IGHV gene is sufficient to generate a wide repertoire of antigen specific antibody responses in transgenic mice. *Mol Immunol.* 2006;43(11):1827–35. doi:10.1016/j.molimm.2005.10.015.
35. Tian M, Cheng C, Chen X, Duan H, Cheng H-L, Dao M, Sheng Z, Kimble M, Wang L, Lin S, et al. Induction of HIV neutralizing antibody lineages in mice with diverse precursor repertoires. *Cell.* 2016;166(6):1471–84.e18. doi:10.1016/j.cell.2016.07.029.
36. Protopapadakis E, Kokla A, Tzartos SJ, Mamalaki A. Isolation and characterization of human anti-acetylcholine receptor monoclonal antibodies from transgenic mice expressing human immunoglobulin loci. *Eur J Immunol.* 2005;35(6):1960–68. doi:10.1002/eji.200526173.
37. Martin ACR, Rees AR. Extracting human antibody sequences from public databases for antibody humanization: high frequency of species assignment errors. *Protein Eng Des Sel.* 2016;29(10):403–08. doi:10.1093/protein/gzw018.
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235–42. doi:10.1093/nar/28.1.235.
39. Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MRS, Nagai MA, da Silva W, Zago MA, Bordin S, Costa FF, Goldman GH, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci U S A.* 2000;97(7):3491–96. doi:10.1073/pnas.97.7.3491.
40. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, Chen X, Longo NS, Louder M, McKee K, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science.* 2011;333(6049):1593–602. doi:10.1126/science.1207532.
41. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature.* 2013;496(7446):469–76. doi:10.1038/nature12053.
42. Wu X, Zhang Z, Schramm CA, Joyce M, Do Kwon Y, Zhou T, Sheng Z, Zhang B, O’Dell S, McKee K, et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell.* 2015;161(3):470–85. doi:10.1016/j.cell.2015.03.004.
43. Bhatti AB, Usman M, Kandi V. Current scenario of HIV/AIDS, treatment options, and major challenges with compliance to antiretroviral therapy. *Cureus.* 2016;8(3):e515–e. doi:10.7759/cureus.515.
44. Kolar GR, Yokota T, Rossi MI, Nath SK, Capra JD. Human fetal, cord blood, and adult lymphocyte progenitors have similar potential for generating B cells with a diverse immunoglobulin repertoire. *Blood.* 2004;104(9):2981–87. doi:10.1182/blood-2003-11-3961.
45. Seifert M, Steimle-Grauer SA, Goossens T, Hansmann ML, Brauning A, Kuppers R. A model for the development of human IgD-only B cells: genotypic analyses suggest their generation in superantigen driven immune responses. *Mol Immunol.* 2009;46(4):630–39. doi:10.1016/j.molimm.2008.07.032.
46. Fraussen J, Vrolix K, Claes N, Martinez-Martinez P, Losen M, Hupperts R, Van Wijmeersch B, Espiño M, Villar LM, De Baets MH, et al. Autoantigen induced clonal expansion in immortalized B cells from the peripheral blood of multiple sclerosis patients. *J Neuroimmunol.* 2013;261(1–2):98–107. doi:10.1016/j.jneuroim.2013.05.002.
47. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y, et al. Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci.* 2017;114(5):944–49. doi:10.1073/pnas.1616408114.
48. Poirion C, Wu Y, Ginestoux C, Ehrenmann F, Douroux P, Lefranc M-P. IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies. 2010 [accessed 2020 04 04]. <http://www.jobim2010.fr/indexe662.html?q=en/node/56>.
49. Parren PWHI, Carter PJ, Plückthun A. Changes to International Nonproprietary Names for antibody therapeutics 2017 and beyond: of mice, men and more. *mAbs.* 2017;9(6):898–906. doi:10.1080/19420862.2017.1341029.
50. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SABDab: the structural antibody database. *Nucleic Acids Res.* 2014;42(D1):D1140–D6. doi:10.1093/nar/gkt1043.
51. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GMR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into

- the human immunoglobulin repertoire. *Proc Natl Acad Sci.* 2009;106(48):20216. doi:[10.1073/pnas.0909775106](https://doi.org/10.1073/pnas.0909775106).
52. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. *Immunol Cell Biol.* 2007;85(4):323–32. doi:[10.1038/sj.icb.7100055](https://doi.org/10.1038/sj.icb.7100055).
 53. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. 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.* 2011;21(5):790–97. doi:[10.1101/gr.115428.110](https://doi.org/10.1101/gr.115428.110).
 54. Krawczyk K, Raybould MIJ, Kovaltsuk A, Deane CM. Looking for therapeutic antibodies in next-generation sequencing repositories. *mAbs.* 2019;11(7):1197–205. doi:[10.1080/19420862.2019.1633884](https://doi.org/10.1080/19420862.2019.1633884).
 55. Lefranc MP. IMGT (ImMunoGeneTics) locus on focus. A new section of experimental and clinical immunogenetics. *Exp Clin Immunogenet.* 1998;15(1):1–7. doi:[10.1159/000019049](https://doi.org/10.1159/000019049).
 56. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59(3):307–21. doi:[10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010).