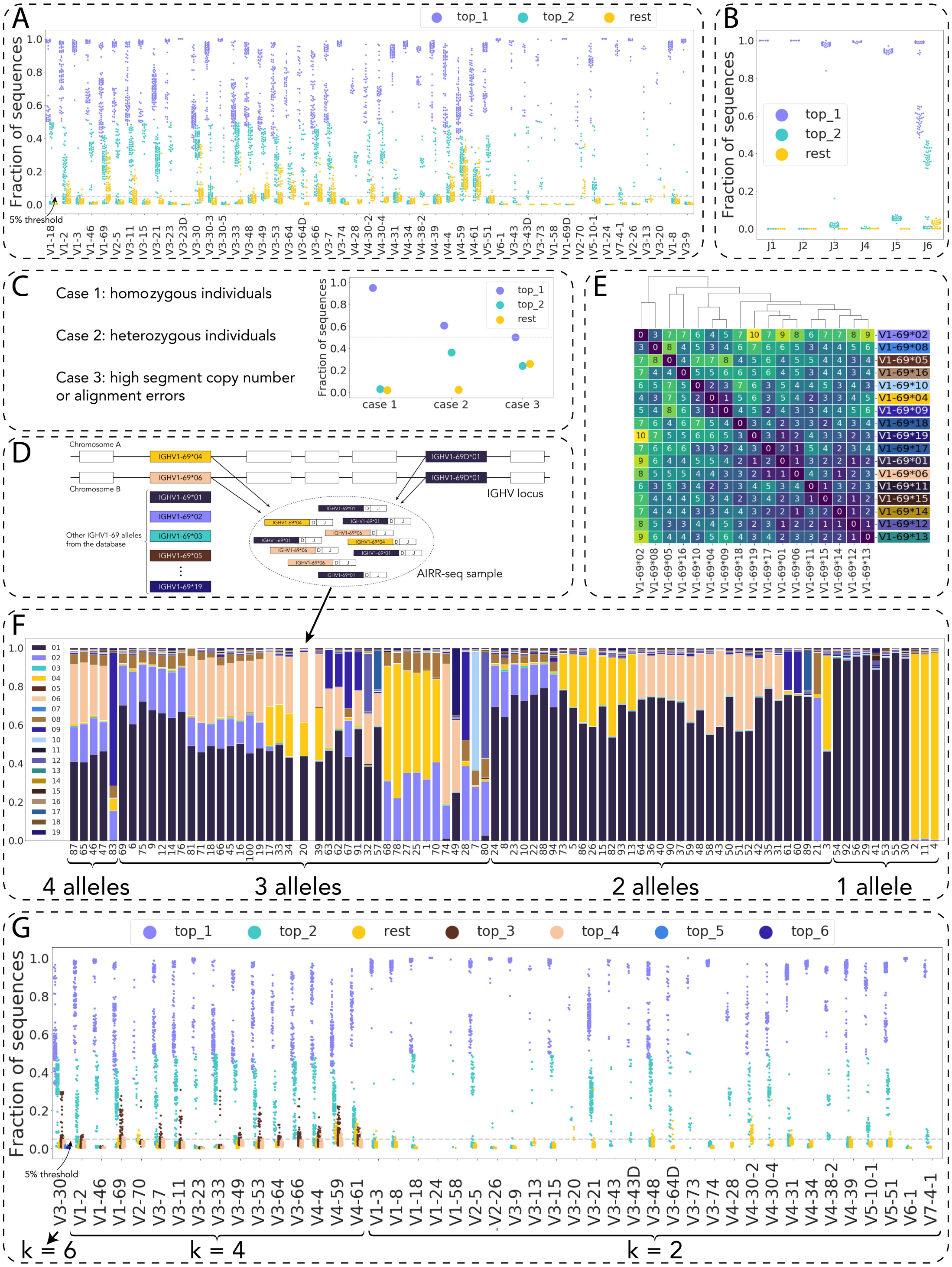
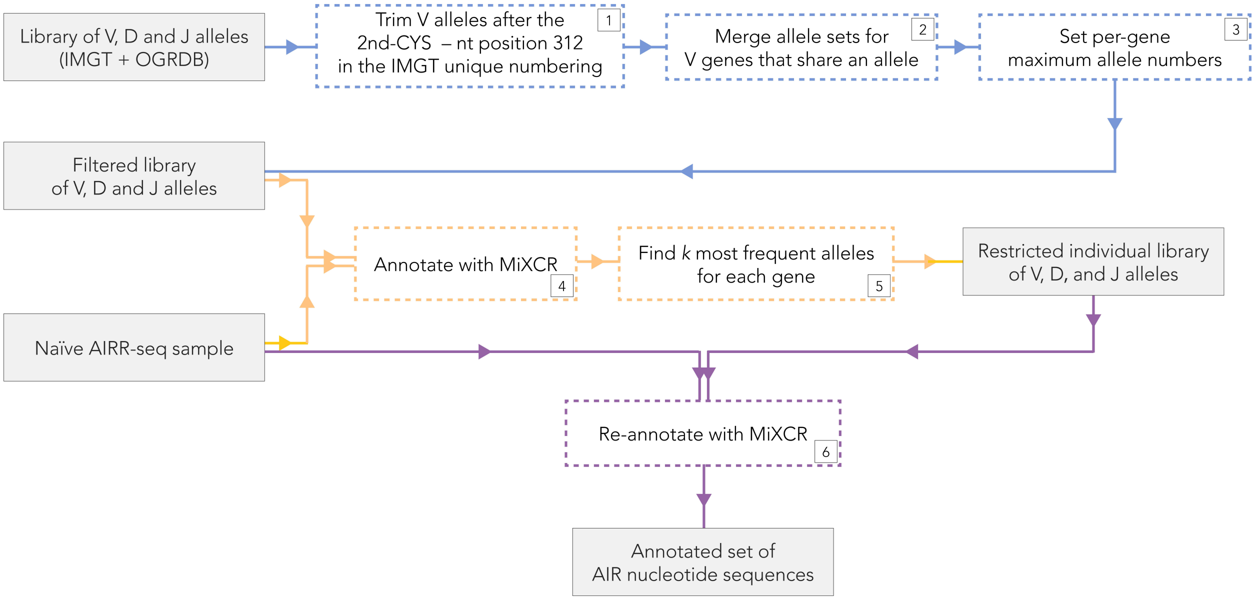
Supplemental Materials



V1-69*02	V1-69*08	V1-69*05	V1-69*16	V1-69*10	V1-69*04	V1-69*09	V1-69*18	V1-69*19	V1-69*17	V1-69*01	V1-69*06	V1-69*11	V1-69*15	V1-69*14	V1-69*12	V1-69*13	
0	3	7	7	6	4	5	7	10	7	9	8	6	7	7	8	9	V1-69*02
3	0	8	4	5	3	4	6	7	6	6	5	3	4	4	5	6	V1-69*08
7	8	0	4	7	7	8	4	5	6	4	5	5	4	4	3	4	V1-69*05
7	4	4	0	5	5	6	6	5	6	4	5	3	4	4	3	4	V1-69*16
6	5	7	5	0	2	3	7	6	3	5	4	4	5	3	4	3	V1-69*10
4	3	7	5	2	0	1	5	6	3	5	4	2	3	3	4	5	V1-69*04
5	4	8	6	3	1	0	4	5	2	4	3	3	4	4	5	6	V1-69*09
7	6	4	6	7	5	4	0	3	4	2	3	3	2	4	3	4	V1-69*18
10	7	5	5	6	6	5	3	0	3	1	2	4	3	3	2	3	V1-69*19
7	6	6	6	3	3	2	4	3	0	2	1	5	4	2	3	4	V1-69*17
9	6	4	4	5	5	4	2	1	2	0	1	3	2	2	1	2	V1-69*01
8	5	5	5	4	4	3	3	2	1	1	0	4	3	1	2	3	V1-69*06
6	3	5	3	4	2	3	3	4	5	3	4	0	1	3	2	3	V1-69*11
7	4	4	4	5	3	4	2	3	4	2	3	1	0	2	1	2	V1-69*15
7	4	4	4	3	3	4	4	3	2	2	1	3	2	0	1	2	V1-69*14
8	5	3	3	4	4	5	3	2	3	1	2	2	1	1	0	1	V1-69*12
9	6	4	4	3	5	6	4	3	4	2	3	3	2	2	1	0	V1-69*13

Supplemental Figure 1. Impact of individualized restriction of germline gene database on allelic assignment in AIRR-seq samples.

(A) The proportion of the two most frequent alleles for 100 individuals from the HUMAN3 dataset. Each column on the plot corresponds to one *IGHV* gene. For each sample (individual), three data points are shown: the fraction of sequencing reads assigned to the two most frequent alleles (violet: top_1, turquoise: top_2) and the total fraction of all remaining alleles (yellow: rest). The dashed gray line corresponds to the 5% threshold above which we consider allele fractions as not due to sequencing/PCR errors. Alignments were performed using MiXCR v3.0.12. (B) Analogous to A but for IGHJ genes. (C) Per-gene alignment scenarios can be divided into three general cases. Case 1: a high fraction (> 90%) of sequencing reads assigned to the most frequent allele, and negligible fractions (<5%) of sequencing reads assigned to the remaining ones (which suggests a homozygous individual for a given gene). Case 2: comparable fractions of sequencing reads assigned to the first and the second most frequent alleles and a low (< 5%) total fraction of the remaining ones (suggesting a heterozygous individual). Case 3: a high (>5%) total fraction of the remaining alleles, suggesting a systematic alignment bias likely caused by gene duplication, which may lead to the presence of three or more alleles of the current gene. (D) Illustration of the generation of an AIRR-seq sample that contains more than two alleles of the same gene. (E) Example: the sequence diversity of *IGHV1-69* alleles as quantified by the edit (Levenshtein) distance. (F) Example: the proportion of the *IGHV1-69* alleles across the 99 individuals of the HUMAN3 dataset: one column corresponds to one individual. The columns are ordered by the number of alleles present, and within each category for allelic diversity (4 alleles, 3 alleles, etc.) by the most frequent alleles. (G) The fraction of the remaining alleles is low and is below the 5% threshold for sequencing and PCR error, for most of the genes and most of the individuals after performing the full pipeline, i.e., with realignment of the sequences against the restricted set of alleles (see Supplemental Fig. S2).

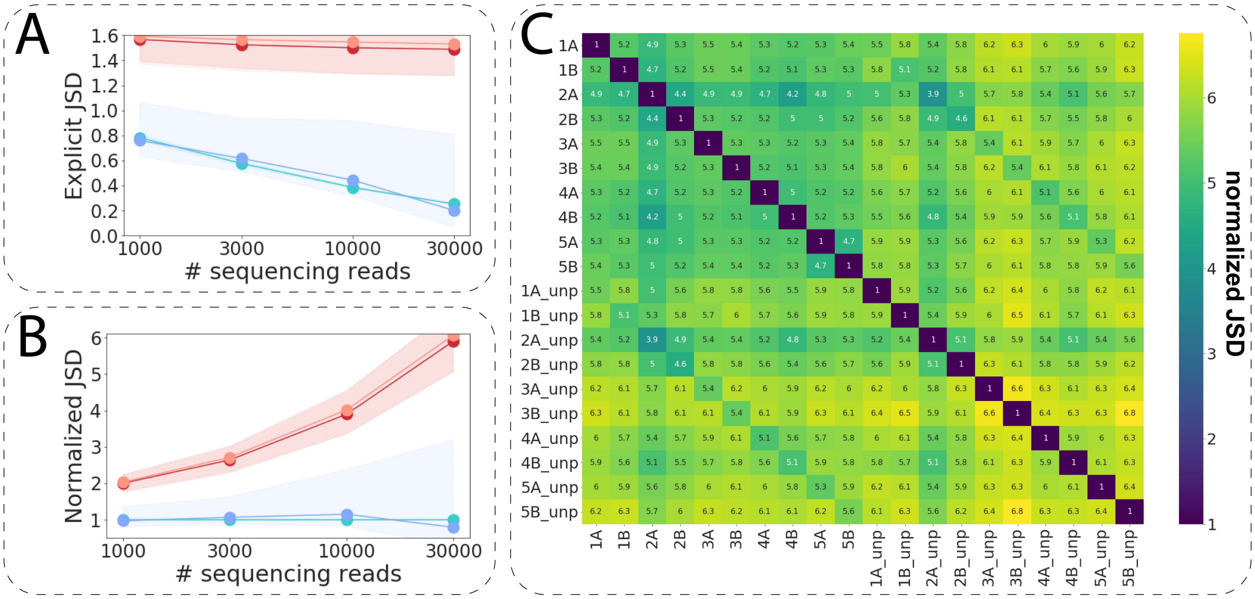


Supplemental Figure 2 (relates to Supplemental Fig. S1). Pipeline for allele-level annotation of AIR sequencing reads using validated allele sets deposited in publicly available databases.

Diploidy and structural variations in the *IGHV* locus, along with the high similarity between the *IGHV* segments render annotation of AIRR-seq data a complex problem. The majority of the currently available annotation tools (Bolotin et al. 2015; Ye et al. 2013; Brochet et al. 2008; Ralph and Matsen 2016) may suggest several V segments for each sequencing read. Each sequencing read is processed independently, which may result in the presence of any number of alleles per gene in the annotations for a given sample. Thus, there is a need (see Supplemental Fig. S1) for a method to accurately restrict the set of *IGHV* alleles in AIRR-seq data for each individual. To address this need, we employed the following strategy: Step 1: we constructed a library of V allele nucleotide sequences based on IMGT (Lefranc 2001, downloaded on 2020-05-29) and OGRDB (Lees et al. 2020) downloaded on 2020-05-29), databases, and trimmed all sequences after the second Cysteine, nt position 312 according to the IMGT unique numbering scheme (Lefranc et al. 2003). We performed trimming to ensure the absence of alignment errors which may be caused by nucleotides located after position 312, i.e., in the VD junction (this problem was also noted by (Mikocziova et al. 2020)). Step 2: for those alleles of the same gene that became identical after trimming (namely, *IGHV1-46*01=*03*, *IGHV3-7*02=*04*, *IGHV3-30*03*=*18, *IGHV3-33*01=*06*, *IGHV3-66*01=*04*, *IGHV4-28*01=*03*), we deleted all but one copy. If there were different genes that shared identical alleles (such as *IGHV2-70*04* and *IGHV2-70D*04*), we considered these genes operationally indistinguishable (Luo et al. 2019) and combined their allele sets (again, leaving only one copy of the duplicated allele). Step 3: for each gene, we set a number *k* – the maximum number of alleles that we expect to observe in a sample: default, *k* = 2, for diploid species. However, for some individuals, there may exist more than two alleles for those genes whose allele sets were combined on the previous step (Supplemental Fig. S1D). For those genes, we set k = 4 (and we set k = 6 for *IGHV3-30* as its allele set originates from 3 genes: *IGHV3-30*, *IGHV3-30-3* and *IGHV3-30-5*). We also set k = 4 for those genes that are known to have multiple copies (Watson and Breden 2012; Ford et al. 2020; Luo et al. 2019), see Supplemental Table 1), Step 4: with the modified allele library, we annotated the AIR nucleotide sequencing reads of the given sample using MiXCR (version 3.0.12). Step 5: for each gene, we detected the *k* most frequent alleles, restricted the allele library to these alleles, thus obtaining an individually restricted allele library. Step 6 consisted of annotating all AIR sequencing reads again using the individually restricted allele library. Note that the blue path (steps 1–3) corresponds to preprocessing the allele library and hence those steps need to be performed only once if the original allele database has not been updated. The orange (steps 4–5, computing the individually restricted allele library) and violet (step 6, annotating the sequencing reads) paths, however, must be repeated for each AIRR-seq sample.

For murine data, we only cleaned the sequencing reads (i.e., assembled the contigs) using MiXCR (Bolotin et al. 2015): we used the MiXCR-computed clones, each clone multiplied by its abundance. We did not apply the allele preprocessing workflow, since all considered subjects were inbred mice, so by definition, they share identical germline gene sets.

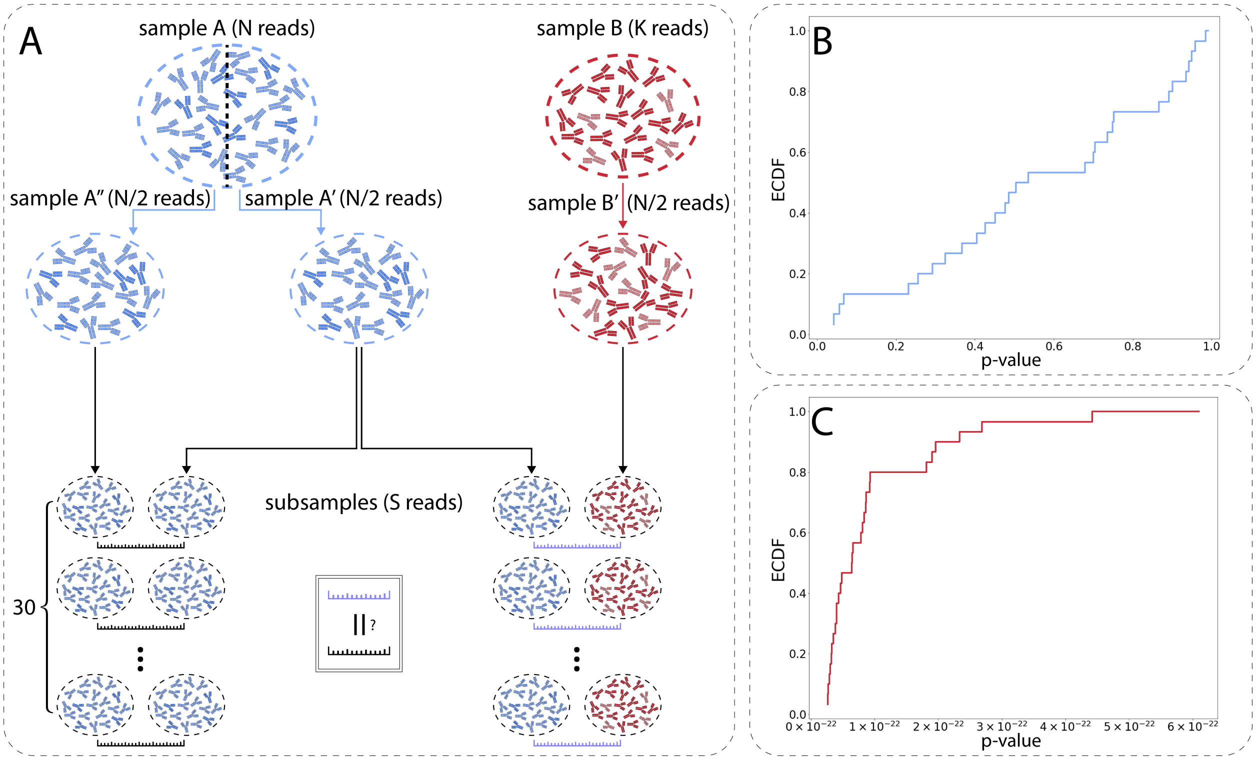
Of note, when aligning the sequences with IGoR for the subsequent model inference, we used an increased D gene alignment score threshold Supplemental Fig. S8): with the default threshold, the alignment module of IGoR produced a huge amount of alignments that did not significantly impact nor the fitted model parameter values nor the sequence Pgens evaluated using this model.



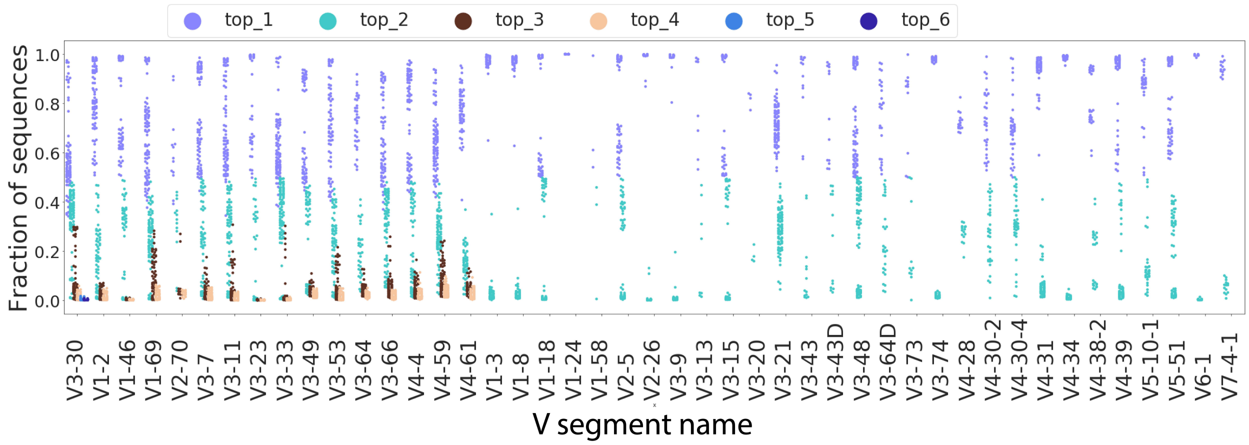
Supplemental Figure 3 (relates to Fig. 2). On the strategy not to filter unproductive sequences: RGMPs inferred from unproductive sequences exhibit similar patterns with higher variance compared to RGMPs inferred from all (unproductive and productive) sequences.

(A–B) Explicit and normalized JSDs between RGMPs inferred from unproductive sequences from the samples of the HUMAN 2 dataset show similar trends to the ones inferred from all (unproductive and productive) sequences (Fig. 2

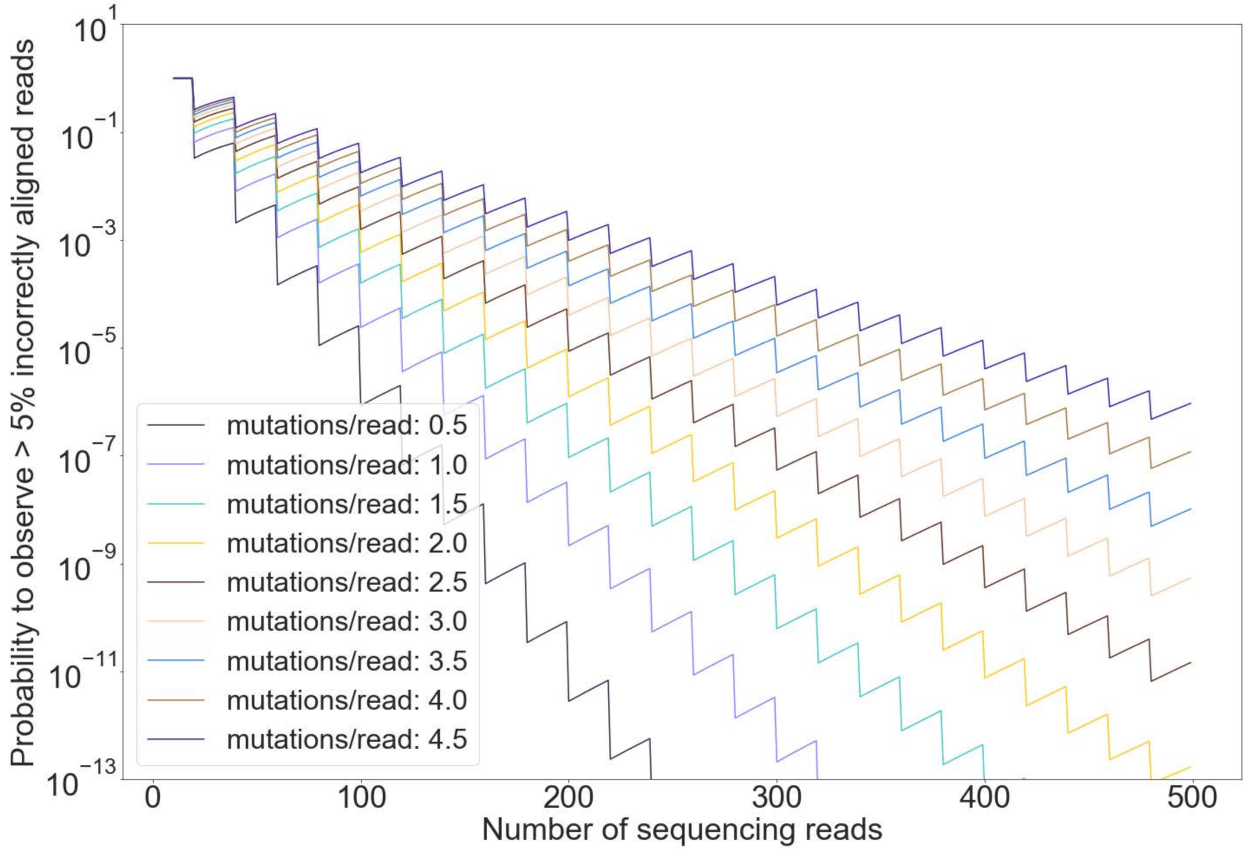
K,L).. However, the range of the distances between replicates (shaded area) is wider due to the lower number of sequences. (C) Pairwise normalized JSDs between RGMPs inferred from all sequences (1A-5B) and from unproductive sequences only (1A_unp-5B_unp). Full sample-inferred (from both productive and unproductive sequences) RGMP sets are generally closer to the "unproductive-inferred" RGMP sets of the same corresponding sample (see the green diagonal in the lower-left/upper-right part), while showing comparable pairwise distances between RGMPs inferred from different samples (upper left quarter). Although Marcou and colleagues suggest (Marcou et al. 2018) to filter out productive sequences for IGoR inference, we inferred RGM from both productive and unproductive reads. One of the reasons for this strategy is that we intended to fit the IGoR model to the most diverse available sequence space: if the VDJ recombination can produce a sequence (productive or unproductive), then we should take this sequence into consideration. To justify this decision, we reproduced part of our analysis for the distance between RGMP sets of human monozygotic twins and unrelated individuals from the HUMAN2 dataset with RGMPs inferred from unproductive sequences only): the data led to the same conclusion (i.e., the explicit and normalized JSDs between twin/unrelated subjects were consistently higher than the ones between replicates but have higher variance). Furthermore, we only worked with data from flow-cytometry sorted non-antigen-experienced cells (pre- and naïve B cells). As hypothesized in the studies in the field (Marcou et al. 2018; Sethna et al. 2020; Desponds et al. 2021), we argue that, although those cells have already undergone selection, the bias introduced by the selection is negligible when compared to the inter-individual differences: if selection impact on RGMP was too high, we would have seen that RGMP sets inferred from naïve B cell samples obtained from different subjects were indistinguishable (i.e., the distance would have been comparable to the distance caused by the noise). However, we see the opposite: although RGMP sets inferred from naïve B cell samples were closer to each other than RGMP sets inferred from pre-B cell samples, they were still clearly distinct (Fig. 2E–G).



Supplemental Figure 4 (relates to Methods 4).
(A) Subsampling from a pair of experimental samples to obtain the 30 independent measures of the explicit JSD values for the sample size of S sequencing reads (S in [1000, 3000, 10000, 30000]). (B) In the case when the null hypothesis holds (A and B are data replicates), for the sample size of 3000 sequencing reads, p-values of the Student's *t*-test for repertoire comparison replicated 30 times are distributed uniformly between 0 and 1, as shown by the empirical cumulative distribution function (ECDF). (C) In the case when the null hypothesis supposedly does not hold (A and B are samples obtained from twin subjects), for the sample size of 3000 sequencing reads, the p-value distribution is highly skewed towards 0.



Supplemental Figure 5 (relates to Methods 2). Allele fractions in the HUMAN3 dataset after applying the pipeline described in Supplemental Fig. S2, similar to Supplemental Fig. S1G.
The yellow points corresponding to the fractions of the remaining (i.e., not among the top_k) alleles were redistributed across other colors – as each of the misaligned sequencing reads was re-aligned to one of the top_k alleles.



Supplemental Figure 6 (relates to Supplemental Fig. S1 and Methods 2). The theoretical probability to observe more than 5% incorrect allele assignments for reads originating from a given V segment.
The probability is calculated for the worst-case scenario under the following assumptions: (i) there are two alleles that differ by a single nucleotide substitution on position *Pos_x.* (ii) nucleotide substitutions in the sequencing reads are independent for different positions (as we talk about naïve B cell data and we can be sure that these

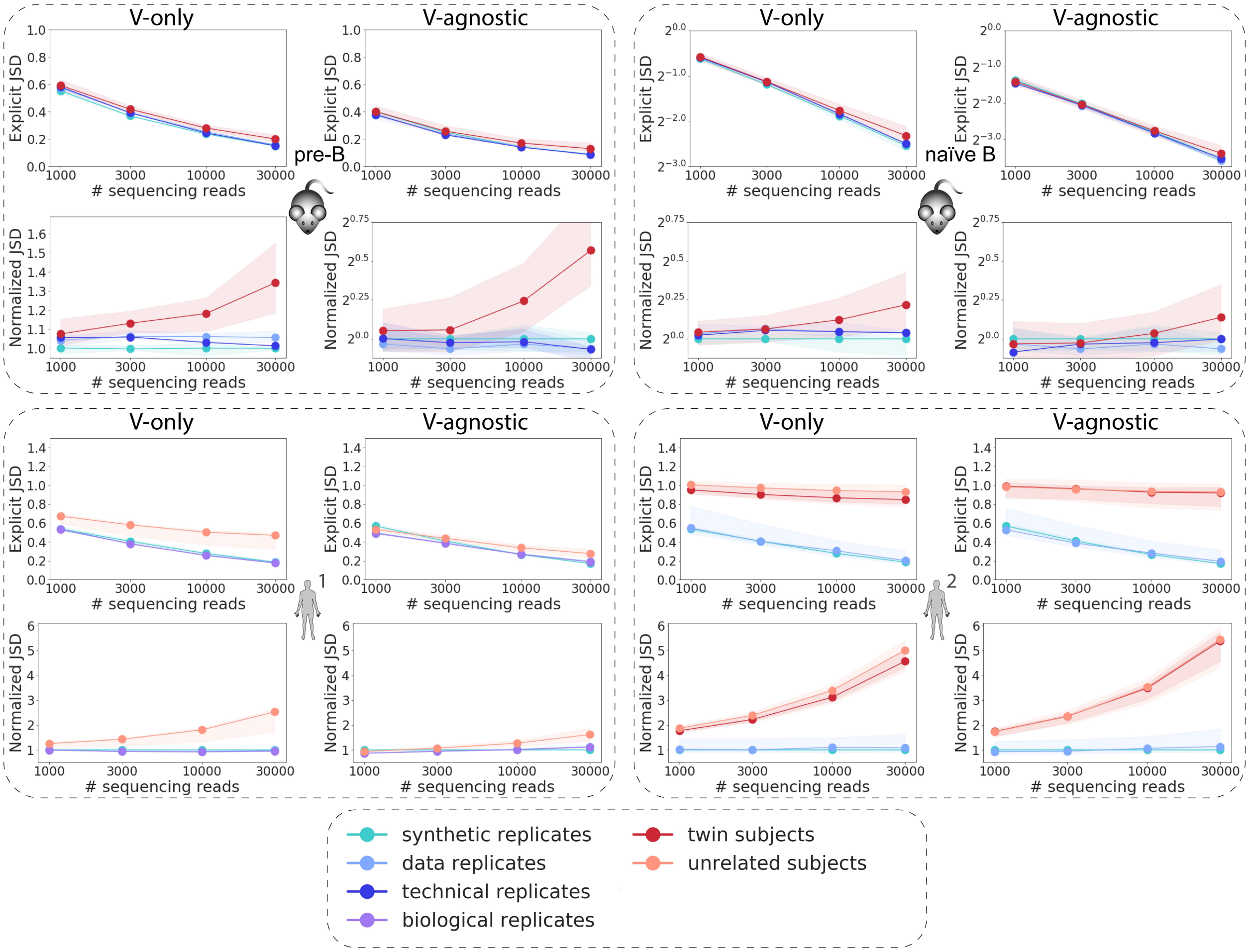
substitutions are caused by sequencing/PCR errors and not somatic hypermutations), (ii) the number of errors per read is below 4.5 (Wardemann and Busse 2019). In such conditions, the probability to misalign a read is the probability for the nucleotide on position *Pos_x* to be mutated. Since two alleles differing by a single nucleotide substitution is the worst case, the calculated probability is an upper bound for a probability to misalign an arbitrary sequencing read. If there are n sequencing reads in the sample and the probability to observe a SHM on a given position is p_{mut} (which is proportional to the average number of mutations per sequencing read), then the probability to observe more than 5% of misassigned alleles is $1 - \sum_0^{n \cdot 0.05} \binom{n}{k}(1 - p_{mut})^{n-k} p_{mut}^k$. The zigzag shape of the plot is due to the number of terms in the sum in the formula: every time the number of reads increases by twenty, the upper limit for k increases by one. The calculations show that the threshold of 500 reads ensures a reliable probability of observing less than 5% misalignments even for the upper bound of the error rate.

V name	Duplications	References	Max #alleles
IGHV1-2	yes	(Watson and Breden 2012)	4
IGHV1-3	no		2
IGHV1-8	no		2
IGHV1-18	suspected	(Luo et al. 2016)	2
IGHV1-24	no		2
IGHV1-45	no		2
IGHV1-46	yes	(Watson and Breden 2012)	4
IGHV1-58	no		2
IGHV1-69	yes	IMGT, (Watson and Breden 2012)	4
IGHV1-69-2	no		2
IGHV2-5	suspected	(Watson and Breden 2012)	2
IGHV2-26	no		2
IGHV2-70	yes	IMGT	4
IGHV3-7	no	inferred from the data	4
IGHV3-9	no		2
IGHV3-11	yes	(Watson and Breden 2012)	4
IGHV3-13	no		2
IGHV3-15	no		2
IGHV3-20	suspected	(Luo et al. 2016)	2
IGHV3-21	suspected	(Luo et al. 2016)	2
IGHV3-23	yes	IMGT	4
IGHV3-30	yes	IMGT	6
IGHV3-33	no	inferred from the data	4
IGHV3-43	yes	IMGT	2
IGHV3-48	no		2
IGHV3-49	no	inferred from the data	4
IGHV3-53	yes	(Watson and Breden 2012)	4

IGHV3-62	no		2
IGHV3-64	yes	(Watson and Breden 2012; Ford et al. 2020), IMGT	4
IGHV3-66	no	inferred from the data	4
IGHV3-72	no		2
IGHV3-73	no		2
IGHV3-74	no		2
IGHV3-NL1	no		2
IGHV4-4	no		2
IGHV4-28	no		2
IGHV4-30-2	no		2
IGHV4-30-4	no		2
IGHV4-31	no		2
IGHV4-34	no		2
IGHV4-38-2	no		2
IGHV4-39	no		2
IGHV4-59	yes	(Watson and Breden 2012; Ford et al. 2020)	4
IGHV4-61	no		2
IGHV5-10-1	no		2
IGHV5-51	no		2
IGHV6-1	no		2
IGHV7-4-1	no		2

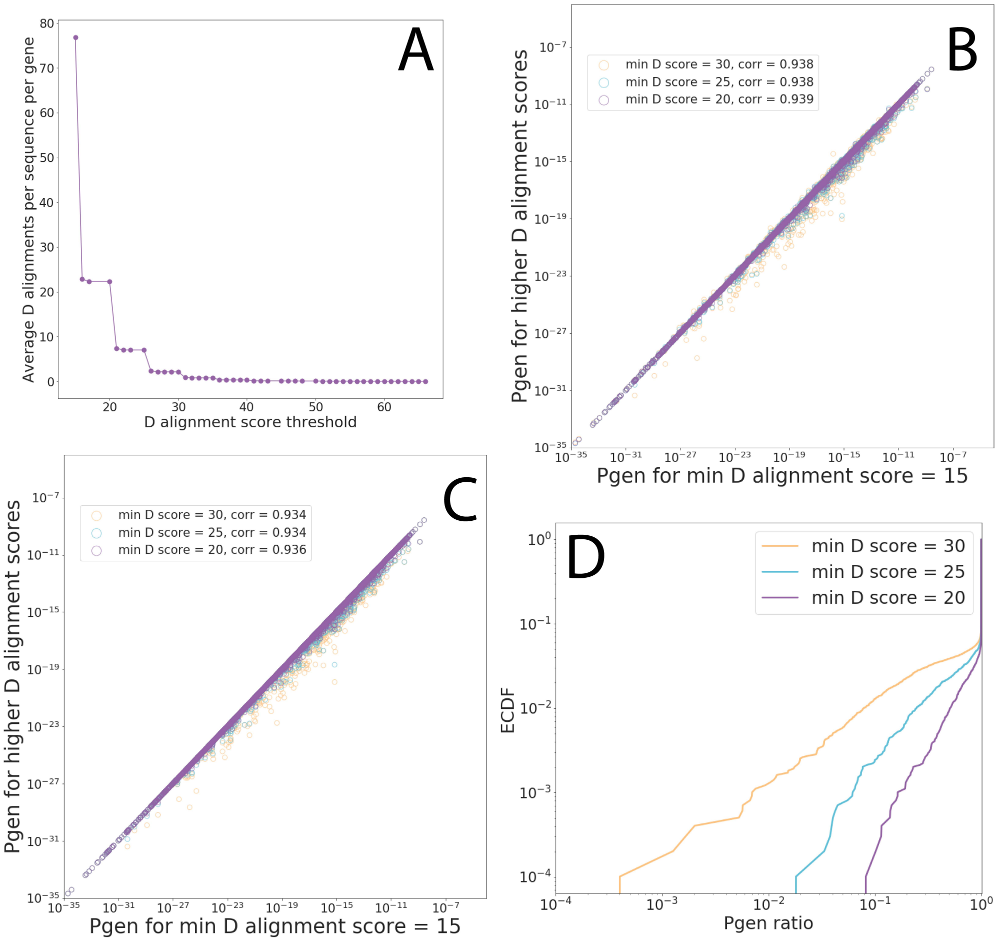
Supplemental Table 1 (relates to Supplemental Fig. S1 and Methods 2). Maximum number of alleles set for each *IGHV* gene.

These numbers are based on the previously known information about *IGHV* copy numbers and on the high fractions of remaining alleles (not top_1 or top_2 alleles in terms of frequency) in the HUMAN3 dataset (Supplemental Fig. S1A).



Supplemental Figure 7 (relates to Fig. 2). JSD-based distance for V segment-only and V segment-agnostic explicit and normalized JSD.

Analogous to Fig. 2 except for here the JSD was computed either using only the V choice probability distribution and V deletion profiles (columns 1 and 3) or in a V segment-agnostic way (columns 2 and 4) for MOUSE_PRE (upper left), MOUSE_NAIVE (upper right), HUMAN1 (lower left) and HUMAN2 (lower right) datasets.



Supplemental Figure 8 (relates to Supplemental Fig. S1 and Methods 2). IGoR D gene alignment score threshold can be increased without substantial change in the inferred RGMPs.

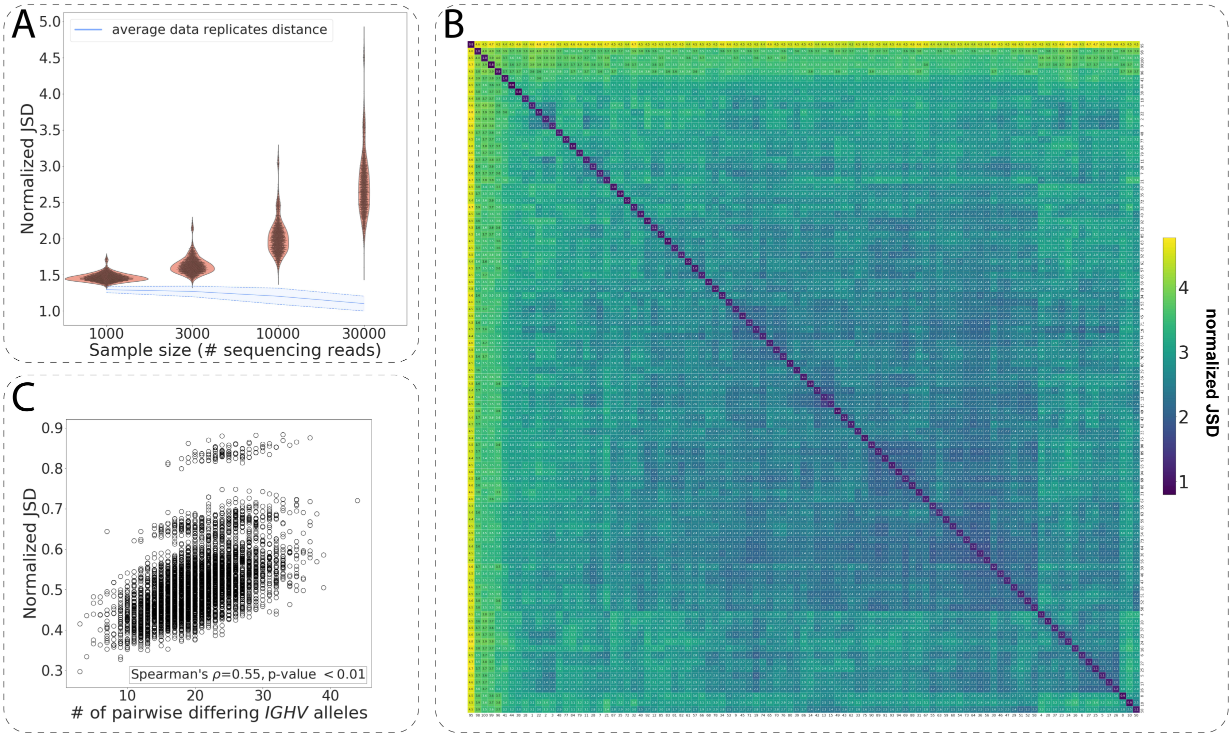
Here, we show the reasoning behind increasing the D gene alignment score thresholds in IGoR from 15 to 30.

(A) A synthetic dataset of 10000 sequencing reads generated using ImmuneSim was aligned using IGoR. With the default value for D gene alignment score threshold (15), IGoR produces an enormous number of alignments: more than 80 alignment variants per sequence per gene on average. We filtered the alignments, leaving only alignments with scores higher than 20 (violet), 25 (blue) and 30 (yellow). Filtering the alignments reduces the number of alignment variants per sequence per gene to the more realistic 8 (for threshold=25) or 2 (for threshold=30).

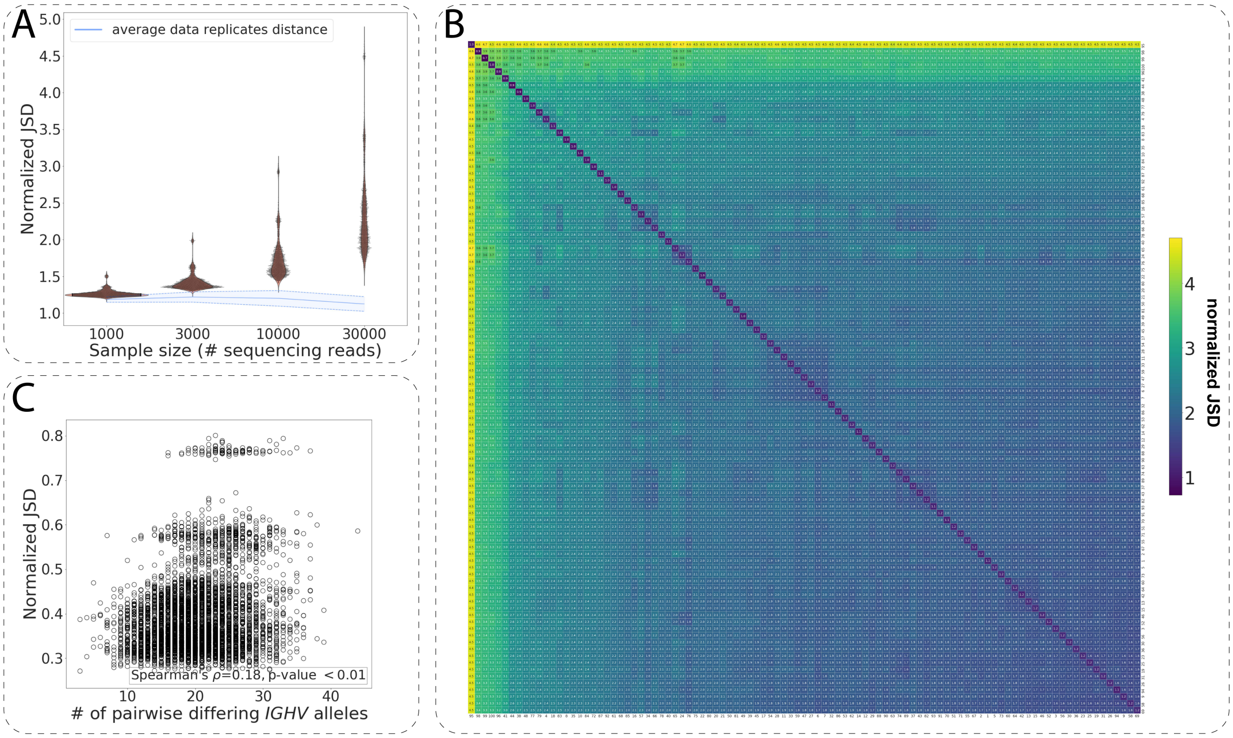
(B) After that, we inferred the RGMP for all four variants of alignments and used the four inferred RGMP sets to evaluate the sequence Pgens. To plot the Pgens estimated using the modified alignment score threshold as a function of the ones estimated using the default threshold, we showed them on a scatter plot: X-axis stands for the D gene alignment score threshold = 15, Y-axis stands for the higher ones. The Pgens are highly consistent (Pearson's r > 0.93), suggesting that one can use IGoR with an increased D gene alignment score threshold. Out of 10000 sequencing reads, 107 Pgens equaled zero (108 for threshold=30).

(C) The same as B, but all Pgens were computed using the same model (inferred from alignments computed with the default parameters) but using different alignments during the evaluation step. The Pgens are also consistent in this case (Pearson's r > 0.93).

(D) The Pgen ratio distribution in a log-log scale (Pgens for higher score thresholds divided by Pgens for threshold=15).

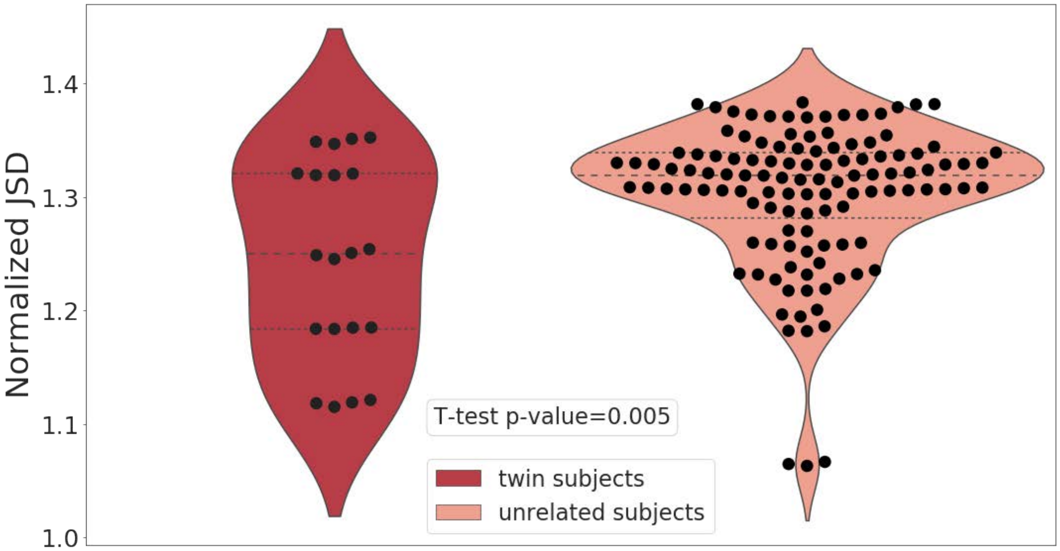


Supplemental Figure 9 (relates to Fig. 3). The normalized V segment-only JSD applied to the HUMAN3 dataset. Analogous to Fig. 3 but here the normalized JSD was calculated using only the V segment-related components (V segment choice and V deletion). The results support those obtained with the full model (Fig. 3).



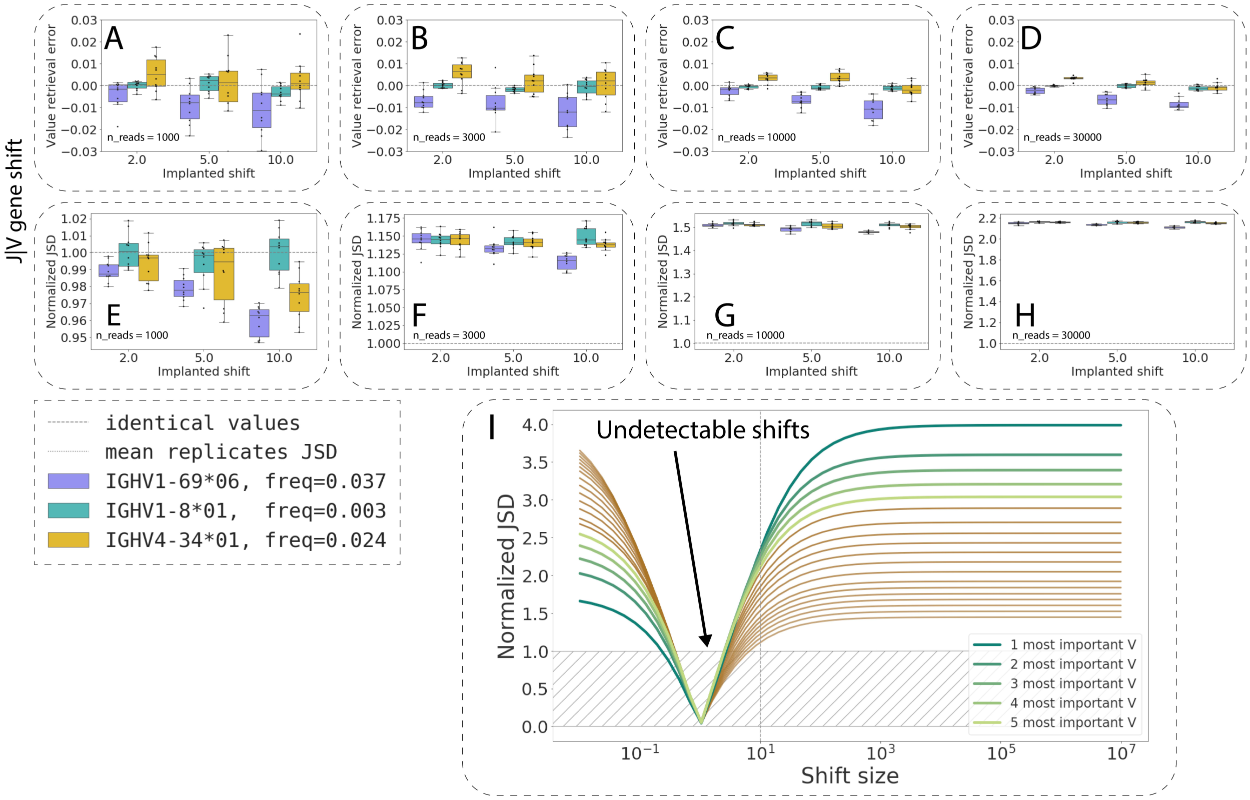
Supplemental Figure 10 (relates to Fig. 3). The normalized V segment-agnostic JSD applied to the HUMAN3 dataset.

Analogous to Fig. 3 but the normalized JSD was calculated only using V segment-agnostic components. The results support the ones obtained with the full model (Fig. 3).

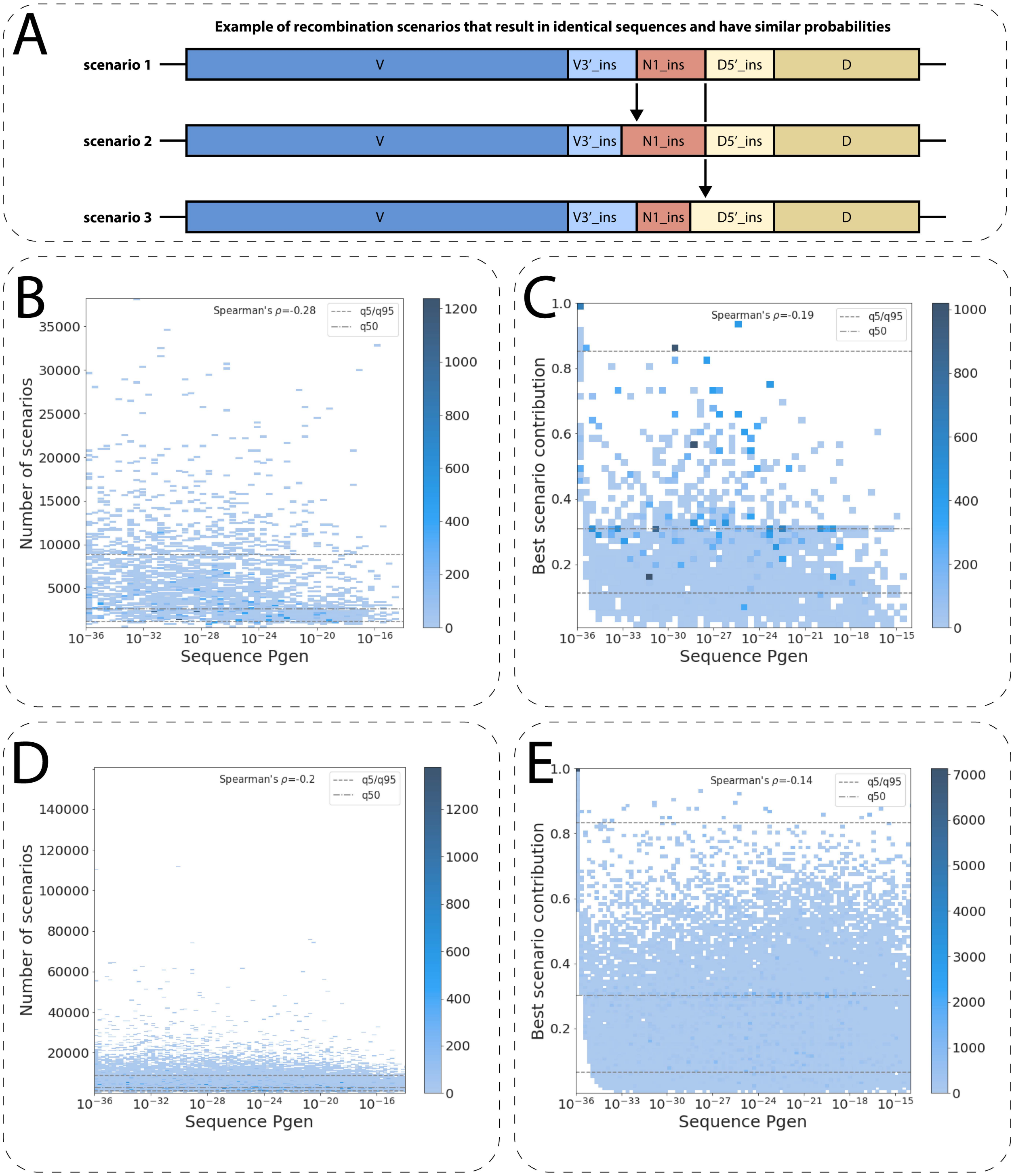


Supplemental Figure 11 (relates to Fig. 2). The normalized JSD distance for five twin pairs from the HUMAN2 dataset.

The normalized JSD within the pairs of twins (left) and across the pairs of twins, i.e., for unrelated individuals (right) in the HUMAN2 dataset, sample size=30000 sequencing reads. The dashed lines correspond to the quartiles of the normalized JSD distribution. The normalized JSD for unrelated individuals is on average higher than the normalized JSD between the twins. In some cases, the normalized JSD for twins is lower than for unrelated individuals: we speculate that this may be because genetic factors account for only a small portion of the JSD and, consequently, variation in non-genetic factors may be stronger than the impact of genetic factors.

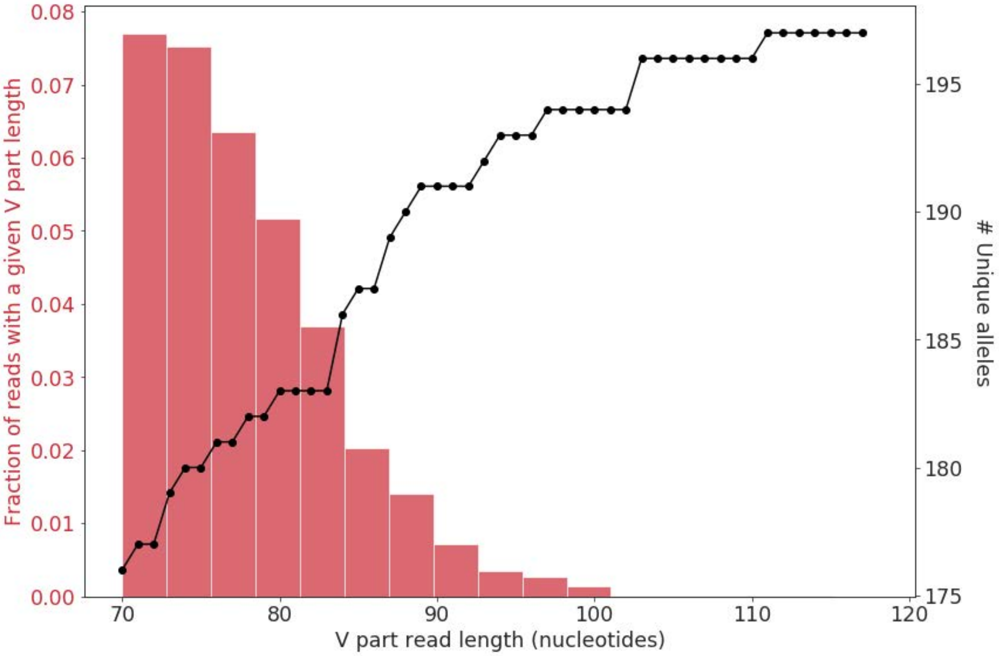


Supplemental Figure 12 (relates to Fig. 5). Sensitivity of IGoR and the normalized JSD to the V choice probabilities. Identical to Fig. 5 but computed for the V choice parameter (instead of the J|V conditional choice parameter).



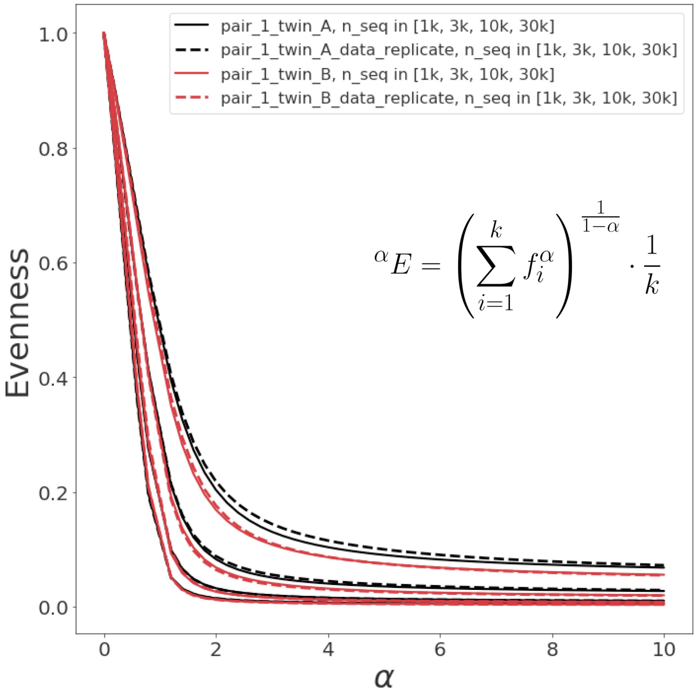
Supplemental Figure 13. The contribution of different recombination scenarios to the total sequence Pgen. (A) An example of similarly probable recombination scenarios that lead to the same sequence. The same nucleotide can be attributed to the V3' palindromic insertion (V3'_ins, scenario 1) or to the N1 insertion (N1_ins, scenario 2). Another nucleotide can be attributed to the N1 insertion (scenario 1) or to the D5' palindromic insertion (D5'_ins, scenario 3). (B) A heatmap showing the 2D-histogram: number of scenarios considered by IGoR for a sequence and its corresponding total Pgen. Sequences with higher Pgens do not tend to consist of more scenarios (computed for the pair_1_twin_A sample from the HUMAN2 dataset). The dashed lines show quantiles (5th, 50th, and 95th) of the number of scenarios distribution. (C) The contribution of the best scenario (best scenario probability divided by the total Pgen) and its corresponding total Pgen (computed for the pair_1_twin_A sample from the HUMAN2 dataset).

The dashed lines show quantiles (5th, 50th, and 95th) of the distribution of the best scenario contribution. (D-E) Analogous to B-C but computed for all samples from the HUMAN2 dataset.



Supplemental Figure 14. Number of unique alleles depending on the length of the V segment end and the actual distribution of the lengths of V segment ends spanned by sequencing reads of the HUMAN1 dataset.

The approach described in Methods 2 was designed for full-length AIRR-sequencing reads. However, we directly applied it also to the HUMAN1 dataset, which was created using shorter reads (2×130nt reads), as we required the allele databases to be identical when comparing RGMPs across different human datasets. Thus, our approach differs from that of Omer and colleagues, who chose a very similar approach but did not have the requirement to compare RGMPs across datasets – and thus specifically modified their analysis for short-read data (Omer et al. 2021). In the preprocessing, we filtered out all reads that contained less than 70 nucleotides in the V segment. We found that only 33 out of 210 V-gene alleles become indistinguishable when this sequence cutoff was implemented. Moreover, there is no pair of alleles that belongs to different genes and becomes indistinguishable with the decreased read length – and hence no genes needed to be merged as 'potential duplications'. It is important to mention that, if two alleles are indistinguishable, we consistently discarded one in favor of the other for all related sequencing reads. Thus, although the pipeline designed for full-length reads may be too conservative for shorter reads, it is still directly applicable to them.



Supplemental Figure 15. Evenness profiles for twin individuals from the HUMAN2 dataset show that data replicates are not substantially affected by the variation of the clone size distribution.

If f_1, … f_i are the (normalized) clone sizes of a sample, then the evenness $^{\alpha}E$ value is determined as shown in the formula on the figure, where *k* is the number of clones in the sample (Greiff et al. 2015). We computed evenness for α in range from 0 to 10 with a step of 0.2 (except 1). For each experimental sample, we calculated the evenness profile for subsamples of size [1000, 3000, 10000, 30000]. Each curve corresponds to one subsample: the color indicates the original experimental sample the sequences were taken from, dashed lines correspond to data replicates, and the subsample size can be determined from the position of the curve: the higher the subsample size, the lower the evenness profile. Same as in the case of the JSD-RGMP-based distance between samples, for each sample, the evenness profile of its data replicate is closer to its own than to evenness profiles of other samples.

References

Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH, et al. 2021a. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep* 34. https://www.cell.com/cell-reports/abstract/S2211-1247(21)00170-4 (Accessed March 24, 2021).

Akbar R, Robert PA, Weber CR, Widrich M, Frank R, Pavlović M, Scheffer L, Chernigovskaya M, Snapkov I, Slabodkin A, et al. 2021b. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *bioRxiv* 2021.07.08.451480.

Arora R, Burke HM, Arnaout R. 2018. Immunological Diversity with Similarity. *bioRxiv* 483131.

Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, Qin K, Fu Y, Huang C-Y, Beigel JH, et al. 2016. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* 6: 20842.

Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, Uddin I, Ismail M, Oakes T, Chain B, et al. 2021. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol* 39: 236–245.

Bernat NV, Corcoran M, Nowak I, Kaduk M, Dopico XC, Narang S, Maisonasse P, Dereuddre-Bosquet N, Murrell B, Hedestam GBK. 2021. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. *Immunity* 0. https://www.cell.com/immunity/abstract/S1074-7613(20)30546-X (Accessed January 27, 2021).

Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. 2017. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* 18: 155.

Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, Chudakov DM. 2015. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 12: 380–381.

Briney B, Inderbitzin A, Joyce C, Burton DR. 2019. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566: 393–397.

Brochet X, Lefranc M-P, Giudicelli V. 2008. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36: W503-508.

Chi X, Li Y, Qiu X. 2020. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* 160: 233–247.

Collins AM, Yaari G, Shepherd AJ, Lees W, Watson CT. 2020. Germline immunoglobulin genes: Disease susceptibility genes hidden in plain sight? *Curr Opin Syst Biol* 24: 100–108.

Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, Martin M, Hedestam GBK. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* 7: 1–14.

Davidsen K, Olson BJ, DeWitt WS III, Feng J, Harkins E, Bradley P, Matsen FA IV. 2019. Deep generative models for T cell receptor protein sequences. *eLife* 8: e46935.

Desponds J, Mayer A, Mora T, Walczak AM. 2021. Population Dynamics of Immune Repertoires. In *Mathematical, Computational and Experimental T Cell Immunology* (eds. C. Molina-París and G. Lythe), pp. 203–221, Springer International Publishing, Cham https://doi.org/10.1007/978-3-030-57204-4_12 (Accessed March 19, 2021).

DeWitt WS III, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS. 2016. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLOS ONE* 11: e0160853.

Dupic T, Koraichi MB, Minervina AA, Pogorelyy MV, Mora T, Walczak AM. 2021. Immune fingerprinting through repertoire similarity. *PLOS Genet* 17: e1009301.

Elhanati Y, Sethna Z, Callan CG, Mora T, Walczak AM. 2018. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol Rev* 284: 167–179.

Emerson RO, DeWitt WS III, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, et al. 2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49: 659–665.

Ford M, Haghshenas E, Watson CT, Sahinalp SC. 2020. Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes using Long Reads. *iScience* 0. https://www.cell.com/iscience/abstract/S2589-0042(20)30067-5 (Accessed February 7, 2020).

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, et al. 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci* 102: 10604–10609.

Friedensohn S, Neumeier D, Khan TA, Csepregi L, Parola C, Vries ARG de, Erlach L, Mason DM, Reddy ST. 2020. Convergent selection in antibody repertoires is revealed by deep learning. *bioRxiv* 2020.02.25.965673.

Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, O'Connor KC, Yaari G, Kleinstein SH. 2019. Identification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. *Front Immunol* 10: 129.

Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, Sarna VK, Lundin KEA, Clouser C, Vigneault F, et al. 2019. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat Commun* 10: 628.

Giudicelli V, Chaume D, Lefranc M-P. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* 33: D256-261.

Glanville J, Kuo TC, Büdingen H-C von, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL, et al. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci* 108: 20066–20071.

Greef PC de, Boer RJ de. 2021. TCRβ rearrangements without a D segment are common, abundant, and public. *Proc Natl Acad Sci* 118. https://www.pnas.org/content/118/39/e2104367118 (Accessed September 29, 2021).

Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* 7: 49.

Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, Valai A, Lopes T, Radbruch A, Winkler TH, et al. 2017a. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep* 19: 1467–1478.

Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, Reddy ST. 2017b. Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *J Immunol* 199: 2985–2997.

Greiff V, Yaari G, Cowell LG. 2020. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr Opin Syst Biol* 24: 109–119.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Comput Sci Eng* 9: 90–95.

Isacchini G, Walczak AM, Mora T, Nourmohammad A. 2021. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proc Natl Acad Sci* 118. https://www.pnas.org/content/118/14/e2023141118 (Accessed April 6, 2021).

Kanduri C, Pavlović M, Scheffer L, Motwani K, Chernigovskaya M, Greiff V, Sandve GK. 2021. *Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification*. https://www.biorxiv.org/content/10.1101/2021.05.23.445346v2 (Accessed September 29, 2021).

Kenter AL, Watson CT, Spille J-H. 2021. Igh Locus Polymorphism May Dictate Topological Chromatin Conformation and V Gene Usage in the Ig Repertoire. *Front Immunol* 12. https://www.frontiersin.org/articles/10.3389/fimmu.2021.682589/full (Accessed May 18, 2021).

Khan TA, Friedensohn S, Vries ARG de, Straszewski J, Ruscheweyh H-J, Reddy ST. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* 2: e1501371.

Khatri I, Berkowska MA, van den Akker EB, Teodosio C, Reinders MJT, van Dongen JJM. 2021. Population matched (pm) germline allelic variants of immunoglobulin (IG) loci: Relevance in infectious diseases and vaccination studies in human populations. *Genes Immun* 22: 172–186.

Koraichi MB, Touzel MP, Mora T, Walczak AM. 2021. NoisET: Noise learning and Expansion detection of T-cell receptors with Python. *ArXiv210203568 Q-Bio*. http://arxiv.org/abs/2102.03568 (Accessed February 16, 2021).

Kullback S, Leibler RA. 1951. On Information and Sufficiency. *Ann Math Stat* 22: 79–86.

Lee JH, Toy L, Kos JT, Safonova Y, Schief WR, Havenar-Daughton C, Watson CT, Crotty S. 2021. Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells. *Npj Vaccines* 6: 1–12.

Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA IV, Yaari G, Watson CT, Collins A, Shepherd AJ. 2020. OGRDB: a reference database of inferred immune receptor genes. *Nucleic Acids Res* 48: D964–D970.

Lefranc M-P. 2001. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* 29: 207–209.

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 27: 55–77.

Luo S, Yu JA, Li H, Song YS. 2019. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Sci Alliance* 2: e201800221.

Luo S, Yu JA, Song YS. 2016. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. *PLOS Comput Biol* 12: e1005117.

Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, Chain B, Cohen IR, Friedman N. 2014. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* 24: 1603–1612.

Marcou Q, Mora T, Walczak AM. 2018. High-throughput immune repertoire analysis with IGoR. *Nat Commun* 9: 561.

Martins FR, Pontes LA de M, Mendes TA de O, Felicori LF. 2021. Discovery of 10,828 new putative human immunoglobulin heavy chain IGHV variants. *bioRxiv* 2021.01.15.426262.

Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, Cook SC, Pogson M, Reddy ST. 2014. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PloS One* 9: e96727.

Miho E, Roškar R, Greiff V, Reddy ST. 2019. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat Commun* 10: 1321.

Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. 2018. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front Immunol* 9. https://www.frontiersin.org/articles/10.3389/fimmu.2018.00224/full (Accessed October 3, 2020).

Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, Sollid LM. 2020. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res* 48: 5499–5510.

Mikocziova I, Greiff V, Sollid LM. 2021a. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun* 1–13.

Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. 2021b. Alternative splice variants and germline polymorphisms in human immunoglobulin light chain genes. *bioRxiv*

2021.02.05.429934.

Mitsunaga EM, Snyder MP. 2020. Deep Characterization of the Human Antibody Response to Natural Infection Using Longitudinal Immune Repertoire Sequencing. *Mol Cell Proteomics* 19: 278–293.

Müllner D. 2011. Modern hierarchical, agglomerative clustering algorithms. *ArXiv11092378 Cs Stat*. http://arxiv.org/abs/1109.2378 (Accessed April 2, 2021).

Nemazee D. 2017. Mechanisms of central tolerance for B cells. *Nat Rev Immunol* 17: 281–294.

Nolan S, Vignali M, Klinger M, Dines JN, Kaplan IM, Svejnoha E, Craft T, Boland K, Pesesky M, Gittelman RM, et al. 2020. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq* rs.3.rs-51964.

Olson BJ, Matsen FA IV. 2018. The Bayesian optimist's guide to adaptive immune receptor repertoire analysis. *Immunol Rev* 284: 148–166.

Omer A, Peres A, Rodrigues OL, Watson CT, Lees W, Polak P, Collins AM, Yaari G. 2021. T Cell Receptor Beta (TRB) Germline Variability is Revealed by Inference From Repertoire Data. *bioRxiv* 2021.05.17.444409.

Pal S, Tyler JK. 2016. Epigenetics and aging. *Sci Adv* 2: e1600584.

Parks T, Mirabel MM, Kado J, Auckland K, Nowak J, Rautanen A, Mentzer AJ, Marijon E, Jouven X, Perman ML, et al. 2017. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun* 8: 14946.

Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, Waagan K, Bernal FLM, Costa AA, Corrie B, et al. 2021. immuneML: an ecosystem for machine learning analysis of adaptive immune receptor repertoires. *bioRxiv* 2021.03.08.433891.

Perelson AS, Oster GF. 1979. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* 81: 645–670.

Peres A, Gidoni M, Polak P, Yaari G. 2019. RAbHIT: R Antibody Haplotype Inference Tool. *Bioinforma Oxf Engl* 35: 4840–4842.

Puelma Touzel M, Walczak AM, Mora T. 2020. Inferring the immune response from repertoire sequencing. *PLoS Comput Biol* 16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7213749/ (Accessed January 14, 2021).

Pulivarthy SR, Lion M, Kuzu G, Matthews AGW, Borowsky ML, Morris J, Kingston RE, Dennis JH, Tolstorukov MY, Oettinger MA. 2016. Regulated large-scale nucleosome density patterns and precise nucleosome positioning correlate with V(D)J recombination. *Proc Natl Acad Sci* 113: E6427–E6436.

Ralph DK, Matsen FA IV. 2016. Likelihood-Based Inference of B Cell Clonal Families. *PLOS Comput Biol* 12: e1005086.

Raposo B, Dobritzsch D, Ge C, Ekman D, Xu B, Lindh I, Förster M, Uysal H, Nandakumar KS, Schneider G, et al. 2014. Epitope-specific antibody response is controlled by immunoglobulin VH polymorphisms. *J Exp Med* 211: 405–411.

Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. 2020. CoV-AbDab: the coronavirus antibody database. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa739 (Accessed March 10, 2021).

Remmel JL, Ackerman ME. 2021. Rationalizing Random Walks: Replicating Protective Antibody Trajectories. *Trends Immunol* 42: 186–197.

Robert PA, Akbar R, Frank R, Pavlović M, Widrich M, Snapkov I, Chernigovskaya M, Scheffer L, Slabodkin A, Mehta BB, et al. 2021a. One billion synthetic 3D-antibody-antigen complexes enable unconstrained machine-learning formalized investigation of antibody specificity prediction. *bioRxiv* 2021.07.06.451258.

Robert PA, Arulraj T, Meyer-Hermann M. 2021b. Ymir: A 3D structural affinity model for multi-epitope vaccine simulations. *iScience* 0. https://www.cell.com/iscience/abstract/S2589-0042(21)00947-0 (Accessed August 18, 2021).

Robert PA, Kunze-Schumacher H, Greiff V, Krueger A. 2021c. Modeling the Dynamics of T-Cell Development in the Thymus. *Entropy* 23: 437.

Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, et al. 2020. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front Immunol* 11: 2136.

Roy B, Neumann RS, Snir O, Iversen R, Sandve GK, Lundin KEA, Sollid LM. 2017. High-Throughput Single-Cell Analysis of B Cell Receptor Usage among Autoantigen-Specific Plasma Cells in Celiac Disease. *J Immunol* 199: 782–791.

Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, M. Euskirchen G, Mamedov MR, Swan GE, Dekker CL, et al. 2016. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* 7: 11112.

Russell ML, Souquette A, Levine DM, Allen EK, Kuan G, Simon N, Balmaseda A, Gordon A, Thomas P, Matsen FA, et al. 2021. *Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities*. https://www.biorxiv.org/content/10.1101/2021.09.17.460747v1 (Accessed September 22, 2021).

Safonova Y, Pevzner PA. 2020. V(DD)J recombination is an important and evolutionarily conserved mechanism for generating antibodies with unusually long CDR3s. *Genome Res* 30: 1547–1558.

Sangesland M, Ronsard L, Kazer SW, Bals J, Boyoglu-Barnum S, Yousif AS, Barnes R, Feldman J, Quirindongo-Crespo M, McTamney PM, et al. 2019. Germline-Encoded Affinity for Cognate Antigen Enables Vaccine Amplification of a Human Broadly Neutralizing Response against Influenza Virus. *Immunity* 51: 735-749.e8.

Sangesland M, Yousif AS, Ronsard L, Kazer SW, Zhu AL, Gatter GJ, Hayward MR, Barnes RM, Quirindongo-Crespo M, Rohrer D, et al. 2020. A Single Human VH-gene Allows for a Broad-Spectrum Antibody Response Targeting Bacterial Lipopolysaccharides in the Blood. *Cell Rep* 32: 108065.

Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T. 2019. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* 35: 2974–2981.

Sethna Z, Isacchini G, Dupic T, Mora T, Walczak AM, Elhanati Y. 2020. Population variability in the generation and selection of T-cell repertoires. *PLOS Comput Biol* 16: e1008394.

Shemesh O, Polak P, Lundin KEA, Sollid LM, Yaari G. 2021. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. *Front Immunol* 12: 627813.

Sidhom J-W, Larman HB, Pardoll DM, Baras AS. 2021. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* 12: 1605.

Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, Huttner AJ, Laman JD, Nagra RM, Nylander A, et al. 2014. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* 6: 248ra107.

Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, Nielsen JH, Macindoe G, Hetherington J, Martin ACR. 2017. abYsis: Integrated Antibody Sequence and

Structure—Management, Analysis, and Prediction. *J Mol Biol* 429: 356–364.

Trück J, Eugster A, Barennes P, Tipton CM, Luning Prak ET, Bagnara D, Soto C, Sherkow JS, Payne AS, Lefranc M-P, et al. 2021. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling eds. L. Cowell and T. Taniguchi. *eLife* 10: e66274.

Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, Karlsson Hedestam GB. 2019. High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. *Front Immunol* 10. https://www.frontiersin.org/articles/10.3389/fimmu.2019.00660/full (Accessed March 30, 2021).

Venkataraman T, Valencia C, Mangino M, Morgenlander W, Clipman SJ, Liechti T, Valencia A, Christofidou P, Spector T, Roederer M, et al. 2021. Antiviral Antibody Epitope Selection is a Heritable Trait. *bioRxiv* 2021.03.25.436790.

Wardemann H, Busse CE. 2019. Expression Cloning of Antibodies from Single Human B Cells. In *Lymphoma: Methods and Protocols* (ed. R. Küppers), *Methods in Molecular Biology*, pp. 105–125, Springer, New York, NY https://doi.org/10.1007/978-1-4939-9151-8_5 (Accessed August 14, 2020).

Watson CT, Breden F. 2012. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun* 13: 363–373.

Watson CT, Glanville J, Marasco WA. 2017. The Individual and Population Genetics of Antibody Immunity. *Trends Immunol* 38: 459–470.

Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324: 807–810.

Yang J, Wang W, Chen Z, Lu S, Yang F, Bi Z, Bao L, Mo F, Li X, Huang Y, et al. 2020. A vaccine targeting the RBD of the S protein of SARS-CoV-2 induces protective immunity. *Nature* 586: 572–577.

Yang X, Zhu Y, Zeng H, Chen S, Guan J, Wang Q, Lan C, Sun D, Yu X, Zhang Z. 2021. Knowledge-based antibody repertoire simulation, a novel allele detection tool evaluation and application. *bioRxiv* 2021.07.01.450681.

Ye J, Ma N, Madden TL, Ostell JM. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 41: W34–W40.

Yu Y, Ceredig R, Seoighe C. 2017. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. *J Immunol Baltim Md 1950* 198: 2202–2210.

Zhang W, Wang I-M, Wang C, Lin L, Chai X, Wu J, Bett AJ, Dhanasekaran G, Casimiro DR, Liu X. 2016. IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data. *Front Immunol* 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5095119/ (Accessed March 30, 2021).