Immune Fingerprinting through Repertoire Similarity

Thomas Dupic,[1] Meriem Bensouda Koraichi,[2] Anastasia Minervina,[3]
Mikhail Pogorelyy,[3,4] Thierry Mora,[2,*] and Aleksandra M. Walczak[2,*]

[1]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA*
[2]*Laboratoire de physique de l'École Normale Supérieure, CNRS, Sorbonne Université, Université de Paris, and École normale supérieure (PSL), 24 rue Lhomond, 75005 Paris, France*
[3]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia*
[4]*Pirogov Russian National Research Medical University, Moscow, Russia*

Immune repertoires provide a unique fingerprint reflecting the immune history of individuals, with potential applications in precision medicine. However, the question of how personal that information is and how it can be used to identify individuals has not been explored. Here, we show that individuals can be uniquely identified from repertoires of just a few thousands lymphocytes. We present "Immprint," a classifier using an information-theoretic measure of repertoire similarity to distinguish pairs of repertoire samples coming from the same versus different individuals. Using published T-cell receptor repertoires and statistical modeling, we tested its ability to identify individuals with great accuracy, including identical twins, by computing false positive and false negative rates $< 10^{-6}$ from samples composed of 10,000 T-cells. We verified through longitudinal datasets and simulations that the method is robust to acute infections and the passage of time. These results emphasize the private and personal nature of repertoire data.

INTRODUCTION

Personalized medicine is a frequent promise of next-generation sequencing. These high-throughput and low-cost sequencing technologies hold the potential of tailored treatment for each individual. However, progress comes with privacy concerns. Genome sequences cannot be anonymized: a genetic fingerprint is in itself enough to fully identify an individual, with the rare exception of monozygotic twins. The privacy risks brought by these pseudonymized genomes have been highlighted by multiple studies [1–3], and the approach is now routinely used by law enforcement. Sequencing experiments that focus on a limited number of expressed genes should be less prone to these concerns. However, as we will show, B- and T-cell receptor (BCR and TCR) genes are an exception to this rule.

BCR and TCR are randomly generated through somatic recombination [4], and the fate of each B- or T-cell clone depends on the environment and immune history. The immune T-cell repertoire, defined as the set of TCR expressed in an individual, has been hailed a faithful, personalized medical record, and repertoire sequencing (RepSeq) as a potential tool of choice in personalized medicine [5–9]. In this report, we describe how, from small quantities of blood (blood spot or heel prick), one can extract enough information to uniquely identify an individual, providing an immune fingerprint. The "Immprint" classifier analyzes this immune fingerprint to decide whether two samples were sampled from the same individual.

* Corresponding authors. These authors contributed equally.

RESULTS

Given two samples of peripheral blood containing respectively M_1 and M_2 T-cells, we want to distinguish between two hypothetical scenarios: either the two samples come from the same individual ("autologous" scenario), or they were obtained from two different individuals ("heterologous" scenario), see Fig. 1a.

TCR are formed by two protein chains α and β. They each present a region of high somatic variability, labeled CDR3α and CDR3β, randomly generated during the recombination process. These regions are coded by short sequences (around 50 nucleotides), which are captured by RepSeq experiments. The two chains are usually not sequenced together so that the pairing information between α and β is lost. Most experiments focus on the β chain, and when not otherwise specified, the term "receptor sequence" in this paper will refer only to the nucleotide sequence of the TRB gene coding for this β chain (which include CDR3β). Similarly, as most cells expressing the same beta chain are clonally related, we will be using the terms "clone" and "clonotype" to refer to set of cells with the same nucleotide TRB sequence, even if they were produced in separate generation events and are not real biological clones (since we have no means of distinguishing the two cases). CDR3β sequences are very diverse, with more than 10^{40} possible sequences [10]. For comparison, the TCRβ repertoire of a given individual is composed of 10^8 to 10^{10} unique clonotypes [11, 12]. As a result, most of the sequences found in a repertoire are "private".

Immprint scores

To discriminate between the autologous and heterologous scenarios, one can simply count the number of nu-

cleotide receptor sequences shared between the two samples, which we call $\mathcal{S}$. Samples coming from the same individual should have more receptors in common because T-cells are organized in clones of cells carrying the same TCR. By contrast, $\mathcal{S}$ should be low in pairs of samples from different individuals, in which sharing is due to rare convergent recombinations. Appropriately setting a threshold to jointly minimize the rates of false positives and false negatives (Fig. 1b), we can use $\mathcal{S}$ as a classifier to distinguish autologous from heterologous samples. $\mathcal{S}$ is not normalized for sequencing depth and values of $\mathcal{S}$ should not be compared between samples of different size.

The $\mathcal{S}$ score can be improved upon by exploiting the fact that some receptors are much more likely than others to be generated during V(D)J-recombination, with variations in generation probability (P_{gen}, [13–15]) spanning 15 orders of magnitude. Public sequences (with high P_{gen}) are likely to be found in multiple individuals [16], while rare sequences (low P_{gen}) are unlikely to be shared by different individuals, and thus provide strong evidence for the autologous scenario when found in both samples. To account for this information, we define the score:

$$\mathcal{I} = \sum_{\text{shared } s} \left[\ln\left(1/P_{\text{gen}}(s)\right) - \gamma\right], \tag{1}$$

which accounts for Shannon's "surprise" $\ln(1/P_{\text{gen}})$—a measure of unexpectedness—associated with each shared sequence s, so that rare shared sequences count more than public ones. The constant γ depends on the repertoire's clonal structure and is set to 12 in the following (see Methods for an information-theoretic derivation). P_{gen} is computed using models previously trained on data from multiple individuals [14]. Small differences reported between the P_{gen} of distinct individuals justify the use of a universal model [15].

Measuring error rates

We tested the classifiers based on the $\mathcal{S}$ and $\mathcal{I}$ scores on TCRβ RepSeq datasets from 656 individuals [17] – labeled according to their cytomegalovirus (CMV) serological status. Sequences were downsampled to mimic experiments where $M_1 = M_2 = M$ cells were sampled and their receptors sequenced. The frequency of a particular clonotype in the sample (the proportion of cells expressing a particular beta chain) was estimated using the read counts of unique TCRβ sequences, and the mean values of $\mathcal{S}$ and $\mathcal{I}$ computed with a procedure designed to correct for the limited diversity of the sampled repertoire relative to the full repertoire, see Methods . Similar results may be obtained when M_1 and M_2 are different (see Methods). The clones most often shared between two autologous samples are also the most clonally expanded – and hence are probably antigen experienced. We verified that the sequence statistics of those expanded clonotypes did not differ from generic ones (Fig. S1).

In Fig. 1c, we plot the mean value of $\mathcal{S}$ (over many draws of $M = 5000$ cells) for each individual (autologous scenario, in blue) and between pairs of different individuals (heterologous scenario, in yellow). The two scenarios are clearly discernable under both scores. This result holds for pairs of monozygotic twins obtained from a distinct dataset [18] (green dots), consistent with previous reports that twins differ almost as much in their repertoires as unrelated individuals [18–20]. Heterologous scores (yellow dots) vary little, and may be bounded from above by a theoretical prediction (dashed line) based on a model of recombination [21] (see Methods). On the other hand, autologous scores (blue dots) show several orders of magnitude of variability across individuals. These variations stem from the clonal structure of the repertoire, and correlates with measures of diversity (Fig. S2), which is known to vary a lot between individuals and correlates with age [22], serological status, and infectious disease history [23, 24]. To explore the worst case scenario of discriminability, hereafter we will focus on the individual with the lowest autologous $\mathcal{S}$ found in the dataset.

The sampling process introduces an additional source of variability within each individual. Two samples of blood from the same individual do not contain the exact same receptors, and the values of $\mathcal{S}$ and $\mathcal{I}$ is expected to vary between replicates. Examples of these variations are shown in Fig. 1d. The blue (respectively yellow) curves correspond to the sample distributions in the autologous (heterologous) scenario for different individuals (pairs of individuals). The distribution of $\mathcal{S}$ is well-approximated by a Poisson distribution, while $\mathcal{I}$ follows approximately a compound distribution of a normal and Poisson distributions (see Methods for details). Armed with these statistical models of variations, we can predict upper bounds for the false negative and false positive rates. As seen from the detection error trade-off (DET) graph Fig. 1e, the Immprint classifier performs very well for a few thousand receptors with an advantage for $\mathcal{I}$.

With $10,000$ cells, corresponding to $\sim 10\ \mu$L of blood, Immprint may simultaneously achieve a false positive rate of $< 10^{-16}$ and false negative rate of $< 10^{-6}$, allowing for the near-certain identification of an individual based on the $\mathcal{I}$ score in pairwise comparisons against the world population $\sim 10^{10}$. When a large reference repertoire has been collected ($M_1 = 1,000,000$, corresponding to $\sim$ 1mL of blood), an individual can be identified with just 100 cells (Fig. S3).

The AUROC estimator (Area Under the Curve of the Receiver Operating Characteristic), a typical measure of a binary classifier performance, can be used to score the quality of the classifier with a number between 0.5 (chance) and 1 (perfect classification). The $\mathcal{I}$ score outperforms the $\mathcal{S}$ score (Fig. 1f), particularly above moderate sample sizes ($M \approx 5000$). Both scores can be readily generalized to the case of paired receptors TCR$\alpha\beta$, when the pairing of the two chains is available (through single-cell sequencing [25–27] or computational pairing [28]), using $P_{\text{gen}}(\alpha, \beta) = P_{\text{gen}}(\alpha) \times P_{\text{gen}}(\beta)$ [29] for the gener-

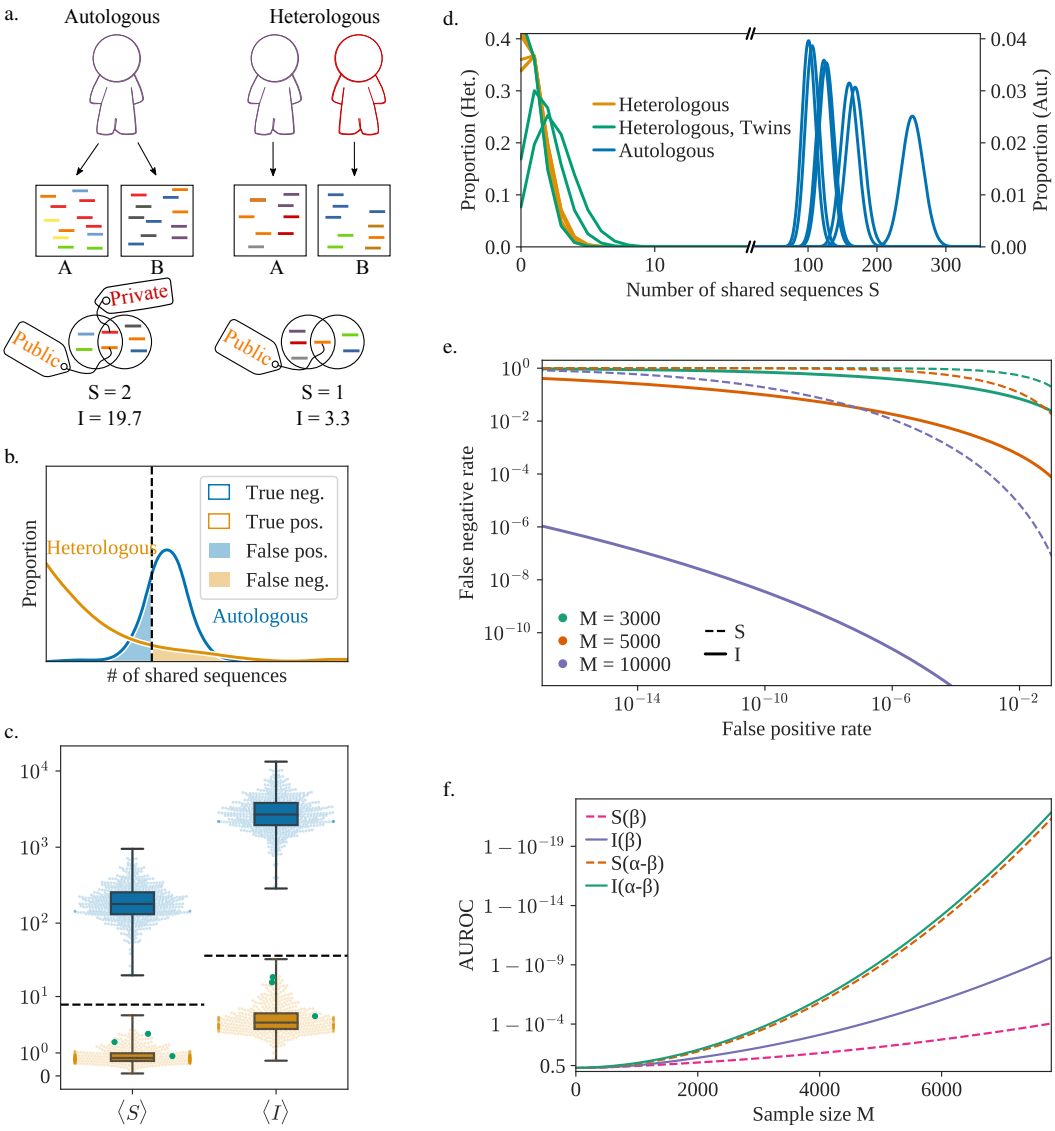

FIG. 1. a) The two samples A and B can either originate from the same individual (autologous) or two different individuals (heterologous). In both scenarios, sequences can be shared between the two samples, but their quantity and quality vary. b) Schematic representation of the distribution of the $\mathcal{S}$ or $\mathcal{I}$ scores for multiple pairs of samples extracted from the same individual (in the autologous scenario) or the same pair of individuals (heterologous). The dashed vertical line represents the threshold value. c) Expected value of $\mathcal{S}$ and $\mathcal{I}$ for different pairs of samples, sampled from the same individual (in blue) or different ones (yellow). Green dots represent samples extracted from pairs of identical twins. The dashed lines represents the theoretical upper bound for heterologous repertoires (see Methods) for both $\mathcal{S}$ and $\mathcal{I}$ ($\gamma = 12$). d) Sampling distributions of $\mathcal{S}$ for 6 different patients (autologous case, each blue curve is one patient) or 6 different pairs of patients (heterologous, each yellow or green curve is a pair of patients) for $M = 5000$. The y-axis scale on the left is adapted to the heterologous distributions while the scale on the right corresponds to the (much wider) autologous ones. The 3 sampling distributions in green correspond to a pair of samples extracted from identical twins. e) Detection Error Trade-off (DET) graph for both summary statistics and different sample sizes M. $\mathcal{I}$ ($\gamma = 12$) outperforms $\mathcal{S}$ in all scenarios. f) AUROC (Area Under Receiver Operating Characteristic), as a function of M. The AUROC is a traditional measure of the quality of a binary classifier (a score closer to one indicates a better classifier). The results are shown for $\mathcal{S}$ and $\mathcal{I}$ both in the default case (only the β chain considered) or for the full (α-β) receptor.

ation probability of the full TCR. Because coincidental sharing of both chains is substantially rarer than with the β chain alone, using the paired chain information greatly improves the classifier.

While this paper focuses on T-cells and TCR sequences, the structure of the B-cell receptors (BCR) repertoire is very similar to the TCR repertoire and we expect to find qualitatively similar results. As an example we use the dataset obtained in Ref. [30] to measure $\mathcal{S}$ and $\mathcal{I}$ for IGH chains (forming half of the BCR receptor), Fig S4. We see that 5000 IgG+ memory B-cells are enough to reliably identify the individuals in the study. However, B-cells are 5 times less common in peripheral blood than T-cells, and somatic hypermutation tends to distort the statistics of receptors, reducing the reliability of our classifier. Hence, for practical applications, T-cells are a better means of identification.

Evolution with time

The previous results used samples obtained at the same time. However, immune repertoires are not static: interaction with pathogens and natural aging modify their composition. The evolution of clonal frequencies will decrease Immprint's reliability with time, especially if the individual has experienced immune challenges in the meantime.

To study the effect of short-term infections, we analyzed an experiment where 6 individuals were vaccinated with the yellow fever vaccine, which is regarded as a good model of acute infection, and their immune system was monitored regularly through blood draws [18]. We observe an only moderate drop in $\mathcal{I}$ caused by vaccination (Fig. 2 a). This is consistent with the fact that infections lead to the strong expansion of only a limited number of clones, while the rest of the immune system stays stable [32–35]. While other types of infections, auto-immune diseases, and cancers may affect Immprint in more substantial ways, our result suggests that it is relatively robust to changes in condition.

We then asked how stable Immprint is over long times. Addressing this issue is hampered by the lack of longitudinal datasets over long periods, so we turn to mathematical models [12, 36–39] to describe the dynamics of the repertoire. Following the model of fluctuating growth rate described in Ref. [38], we define two typical evolutionary timescales for the immune system: τ, the typical turnover rate of T-cell clones, and θ, which represents the typical time for a clonotype to grow or shrink by a factor two as its growth rate fluctuates. The model predicts a power-law distribution for the clone-size distribution, with exponent $-1 - \tau/2\theta$. This exponent has been experimentally measured to be ≈ -2 [10], which leaves us with a single parameter τ, and $\theta = \tau/2$. An example of simulated evolution of Immprint with time is shown in Fig. 2 b. The highlighted histogram represents a data point at two years obtained from [18]. While a fit is possible for this specific individual ($\tau = 0.66$ in Fig. 2 b), the τ parameter is not universal, and we expect it to vary between individuals, especially as a function of age. In Fig. 2c we explore the effect on the stability of Immprint for a range of reasonable values for the clone turn-over rate τ, from 6 months to 10 years, encompassing both previous estimates of the parameter [38] and measured turnover rates for different types of T-cells [40]. While Fig. 2 focuses on $\mathcal{I}$, the behaviour is similar for $\mathcal{S}$ (see Fig. S5). We observe that under this model, for most individuals and bar exceptional events, Immprint should conserve its accuracy for years or even decades.

DISCUSSION

In summary, the T-cells present in small blood samples provide a somatic and long-lived barcode of human individuality, which is robust to immune challenges and stable over time. While the uniqueness of the repertoire was a well known fact, we demonstrated that the most common T-cells clones are still diverse enough to uniquely define an individual and frequent enough to be reliably sampled multiple times. Unlike genome sequencing, repertoire sequencing can discriminate monozygotic twins with the same accuracy as unrelated individuals. However, a person's unique immune fingerprint can be completely wiped out by a hematopoietic stem cell transplant [41]. The different datasets used cover a range of different sequencing methods , however different approaches may lead to slightly different threshold choices. In particular, in practical implementations, sequencing depth is an important concern. One needs enough coverage to sequence TCRβ genes from as many as possible of the T-cells present in the sample, in order to measure a more precise a immune fingerprint. In addition, the specific calculations presented here only apply to peripheral blood cells. Specific cell types or cells extracted from tissue samples may have different clonal distributions and potentially different receptor statistics. For example the value of $\mathcal{S}$ in the autologous case varies between CD4+ and CD8+ T-cells (Fig. S6), different individuals remain distinguishable using each subset.

Immprint is implemented in a python package and webapp (see Methods) allowing the user to determine the autologous or heterologous origin of a pair of repertoires. Beyond identifying individuals, the tool could be used to check for contamination or labelling errors between samples containing TCR information. The repertoire information used by Immprint can be garnered not only from RepSeq experiments, but also from RNA-Seq experiments, which contain thousands of immune receptor transcripts [42, 43]. Relatively small samples of immune repertoires are enough to uniquely identify an individual even among twins, with potential forensics applications. At the same time, unlike genetic data from genomic or mRNA sequencing, Immprint provides no information

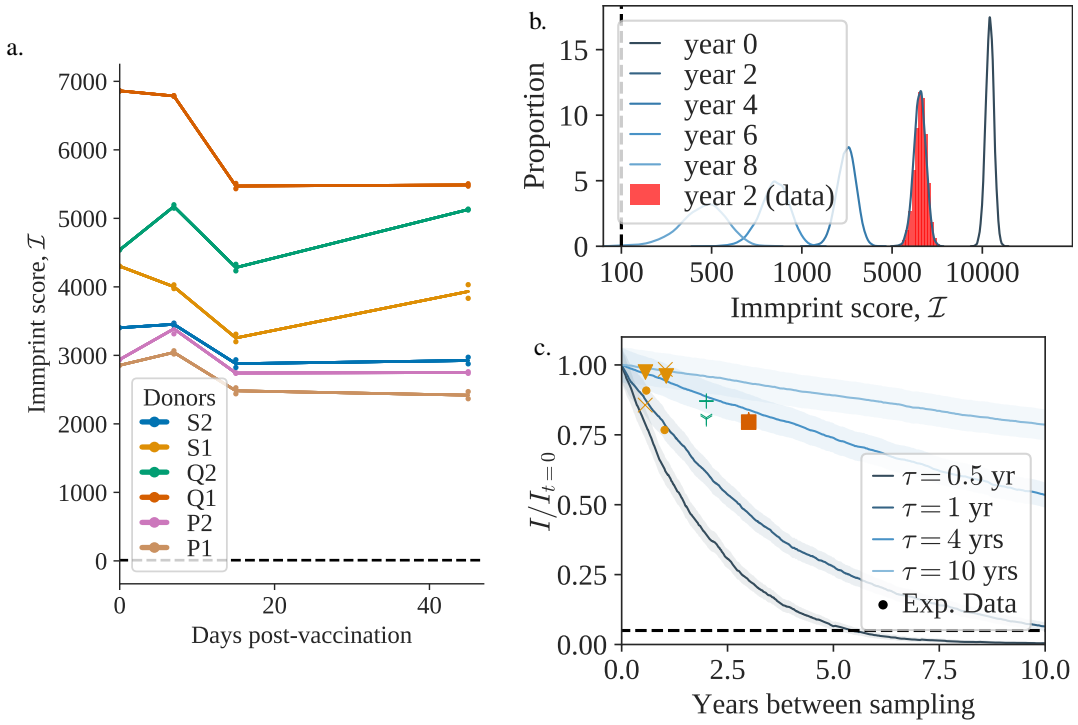

FIG. 2. a) Evolution of $\mathcal{I}$ (M=5000) during vaccination, between a sample taken at day 0 (vaccination date) and at a later timepoint. Each color represents a different individual. Each pair timepoint/individual has two biological replicates. The dashed line represents the threshold value. b) Evolution of $\mathcal{I}$ between a sample taken at year 0 and a later timepoint. The red histogram corresponds to one of the individuals sampled in [18] and the blue curves show theoretical estimates, fitted to match ($\tau = 0.66$). c) Evolution of the (normalized) mean of $\mathcal{I}$ (M=5000) as a function of time for different values of the turnover rate τ. The dashed line represents the threshold value divided by the smallest value of $\mathcal{I}_{t=0}$ (M=5000) in the data. The data points were obtained from the datasets [31] (yellow), [18] (green) and [22] (orange). Different markers indicate different individuals.

about kin relationships, very much like classical fingerprints, and avoids privacy concerns about disclosing genetic information shared with non consenting relatives.

ACKNOWLEDGMENTS

The study was supported by the European Research Council COG 724208. A.A.M and M.V.P. are supported by RSF-20-15-00351. The authors are grateful for the help of Natanael Spizak with the analysis of BCR repertoire datasets.

[1] Homer N, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLOS Genetics* 4:e1000167.

[2] Naveed M, et al. (2015) Privacy in the Genomic Era. *ACM computing surveys* 48.

[3] Sweeney L, Abu A, Winn J (2013) Identifying Participants in the Personal Genome Project by Name., (Social Science Research Network, Rochester, NY), SSRN Scholarly Paper ID 2257732.

[4] Hozumi N, Tonegawa S (1976) Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci USA* 73:3628–3632.

[5] Robins H (2013) Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology* 25:646–52.

[6] Attaf M, Huseby E, Sewell AK (2015) $\alpha\beta$ T cell receptors as predictors of health and disease. *Cellular & molecular immunology* 12:391–399 Edition: 2015/01/26 Publisher: Nature Publishing Group.

[7] Woodsworth DJ, Castellarin M, Holt Ra (2013) Sequence analysis of T-cell repertoires in health and disease. *Genome medicine* 5:98.

[8] Bradley P, Thomas PG (2019) Using T Cell Receptor Repertoires to Understand the Principles of Adap-

tive Immune Recognition. *Annual Review of Immunology* 37:547–570.

[9] Davis MM, Boyd SD (2019) Recent progress in the analysis of $\alpha\beta$ T cell and B cell receptor repertoires. *Current Opinion in Immunology* 59:109–114.

[10] Mora T, Walczak A (2016) Quantifying lymphocyte receptor diversity. *é*.

[11] Qi Q, et al. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America* 111:13139–44.

[12] Lythe G, Callard RE, Hoare RL, Molina-París C (2016) How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology* 389:214–224.

[13] Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences* 109:16161–16166.

[14] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. *Nature Communications* 9:561.

[15] Sethna Z, et al. (2020) Population variability in the generation and thymic selection of T-cell repertoires. *bioRxiv* p 2020.01.08.899682.

[16] Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nature Reviews Immunology* 8:231–238.

[17] Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics* 49:659–665.

[18] Pogorelyy MV, et al. (2018) Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proceedings of the National Academy of Sciences* p 201809642.

[19] Zvyagin IV, et al. (2014) Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 111:5980–5985.

[20] Tanno H, et al. (2020) Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proceedings of the National Academy of Sciences* 117:532–540.

[21] 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. *Immunological Reviews* 284:167–179.

[22] Britanova OV, et al. (2016) Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *The Journal of Immunology* 196:5005–5013.

[23] Sylwester AW, et al. (2005) Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. *Journal of Experimental Medicine* 202:673–685.

[24] Khan N, et al. (2002) Cytomegalovirus Seropositivity Drives the CD8 T Cell Repertoire Toward Greater Clonality in Healthy Elderly Individuals. *The Journal of Immunology* 169:1984–1992.

[25] Dash P, et al. (2011) Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *Journal of Clinical Investigation* 121:288–295.

[26] Redmond D, Poran A, Elemento O (2016) Single-cell TCRseq: Paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Medicine* 8:80.

[27] Grigaityte K, et al. (2017) Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire. *bioRxiv:213462*.

[28] Howie B, et al. (2015) High-throughput pairing of T cell receptor a and b sequences. *Science translational medicine* 7:301ra131.

[29] Dupic T, Marcou Q, Walczak AM, Mora T (2019) Genesis of the $A\beta$ T-cell receptor. *PLOS Computational Biology* 15:e1006874.

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

[31] Chu ND, et al. (2019) Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunology* 20:19.

[32] DeWitt WS, et al. (2015) Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection. *Journal of Virology* 249:JVI.03474–14.

[33] Wolf K, et al. (2018) Identifying and Tracking Low-Frequency Virus-Specific TCR Clonotypes Using High-Throughput Sequencing. *Cell Reports* 25:2369–2378.e4.

[34] Qi Q, et al. (2016) Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Science Translational Medicine* 8:332ra46–332ra46.

[35] Sycheva AL, et al. (2018) Quantitative profiling reveals minor changes of T cell receptor repertoire in response to subunit inactivated influenza vaccine. *Vaccine* 36:1599–1605.

[36] Borghans JAM, De Boer RJ (2007) Quantification of T-cell dynamics: From telomeres to DNA labeling. *Immunological Reviews* 216:35–47.

[37] Thomas-Vaslin V, Altes HK, de Boer RJ, Klatzmann D (2008) Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. *The Journal of Immunology* 180:2240–2250.

[38] Desponds J, Mora T, Walczak AM (2016) Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences* 113:274–279.

[39] de Greef PC, et al. (2020) The naive t-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife* 9:1–24.

[40] De Boer RJ, Perelson AS (2013) Quantifying T Lymphocyte Turnover. *Journal of theoretical biology* 327:45–87.

[41] Buhler S, et al. (2020) Genetic T-cell receptor diversity at 1 year following allogeneic hematopoietic stem cell transplantation. *Leukemia* 34:1422–1432.

[42] Li B, et al. (2017) Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. *Nature Genetics* 49:482–483.

[43] Bolotin DA, et al. (2017) Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology* 35:908–911.

[44] Shugay M, et al. (2014) Towards error-free profiling of immune repertoires. *Nature Methods* 11:653–655.

[45] Bolotin DA, et al. (2015) MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods* 12:380–381.

[46] Sethna Z, Elhanati Y, Callan C, 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.

[47] Touzel MP, Walczak AM, Mora T (2020) Inferring the immune response from repertoire sequencing. *PLoS Computational Biology* 16:1–21.

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

[49] Oakes T, et al. (2017) Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Frontiers in Immunology* 8:1–17.

METHODS

Datasets & Pre-processing

We use five independant RepSeq datasets in this study: *(i)* genomic DNA from Peripheral blood mononuclear cells (PBMCs) from 656 healthy donors [17]; *(ii)* cDNA of PBMCs sampled from three pairs of twins, before and after a yellow-fever vaccination [18]; *(iii), (iv)* two longitudinal studies of healthy adults [22, 31];*(v)* cDNA dataset of IGH genes (B-cells) from 9 individuals, with multiple replicates [30].

CDR3 nucleotide sequences were extracted with MIGEC [44] (for the second dataset) coupled with MiXCR [45]. We also extract the frequency of reads from the three datasets. The non-productive sequences were discarded (out-of-frame, non-functional V gene, or presence of a stop codon). The generation probability (P_{gen}) was computed using OLGA [46], with the default TCRβ model. The frequency of each clone was estimated by summing the frequencies of all reads that shared the same nucleotide CDR3 sequence and identical V,J genes.

The preprocessing code is distributed on the Git repository associated with the paper. We also developed a command-line tool (`https://github.com/statbiophys/immprint`) that discriminates between sample origins, and a companion webapp (`https://immprint.herokuapp.com`).

Discrimination scores

To discriminate between the autologous and heterologous scenarios, we introduce a log-likelihood ratio test between the two possibilities:

$$\mathcal{I} = \sum_s \ln \frac{P(y_1(s), y_2(s)|\text{autologous})}{P(y_1(s), y_2(s)|\text{heterologous})}, \tag{2}$$

where $y_1(s) = 1$ if the sequence s is found in sample 1, and 0 otherwise; likewise $y_2(s) = 1$ if s is in sample 2. The sum runs over all potential sequences s, including unseen ones. To be present in a sample, a sequence s first has to be present in the repertoire. This occurs with probability $1 - (1 - p(s))^{N_c}$, where N_c is the total number of clonotypes in the repertoire, and $p(s)$ is the probability of occurence of sequence s (resulting from generation and selection, see below). Second, it must be picked in a sample of size M, with probability $1 - (1 - f)^M \approx Mf$ (assuming $Mf \ll 1$) depending on its frequency f, which is distributed according to the clone size distribution $\rho(f)$. We checked that $f(s)$ and $P_{\text{gen}}(s)$ were not correlated (Fig. S1) and that the effects of a shared infection between different individuals were limited to a handful of clones (Fig. S7). Then one can write

$$P(y_1(s) = 1, y_2(s) = 1 \mid \text{autologous}) \approx \left(1 - e^{-N_c p(s)}\right) M_1 M_2 \int df\, \rho(f)\, f^2, \tag{3}$$

$$P(y_1(s) = 1, y_2(s) = 0 \mid \text{autologous}) \approx \left(1 - e^{-N_c p(s)}\right) \frac{M_1}{N_c} \text{ and } 1 \leftrightarrow 2, \tag{4}$$

$$P(y_1(s) = 0, y_2(s) = 0 \mid \text{autologous}) \approx 1 - \left(1 - e^{-N_c p(s)}\right) \frac{M_1 + M_2}{N_c}, \tag{5}$$

where we've used $\int df\, \rho(f)\, f = 1/N_c$. For the heterologous case the probability factorizes as:

$$P(y_1(s), y_2(s) \mid \text{heterologous}) = P_1(y_1(s)) P_2(y_2(s)), \tag{6}$$

with

$$P_a(y_a(s) = 1) \approx \left(1 - e^{-N_c p(s)}\right) \frac{M_a}{N_c}, \quad a = 1, 2. \tag{7}$$

Since only the term $y_1(s) = y_2(s) = 1$ (shared sequences) is different between the autologous and heterologous cases, we obtain:

$$\mathcal{I} = \sum_{\text{shared } s} \left[\ln(N_c^2 \langle f^2 \rangle) - \ln\left(1 - e^{-N_c p(s)}\right)\right]. \tag{8}$$

Further assuming $N_c p(s) \ll 1$, and $p(s) = P_{\text{gen}}(s) q^{-1}$ (where q accounts for selection [21] and $P_{\text{gen}}(s)$ is the probability of sequence generation [14]), the score simplifies to Eq. 1, with $\gamma = -\ln(q N_c \langle f^2 \rangle) = \ln(q^{-1} \langle f \rangle / \langle f^2 \rangle)$. The factor γ depends on unknown parameters of the model, but can be estimated assuming a power-law for the clone size distribution [47], $\rho(f) \propto f^{-2}$ extending from $f = 10^{-11}$ to $f = 0.01$, and $q = 0.01$ [21], yielding $\gamma \approx 12.24$. Alternatively we optimized γ to minimize the AUROC, yielding $\gamma \approx 15$ (SI Fig. S8). Since performance degrades quickly for larger values, we conservatively set $\gamma = 12$.

Estimating mean scores from RepSeq datasets

The sampling of M cells from blood is simulated using large repertoire datasets. In a bulk repertoire sequencing dataset, the absolute number of cells for each clonotype (cells with a specific receptor) is unknown, but the fraction of each clonotype can be estimated using the proportion of reads that are associated with this specific receptor. To estimate the autologous $\mathcal{S}$ and $\mathcal{I}$ of two samples of size M_1 and M_2 in the absence of true replicates, we computed their expected values from a single dataset containing N reads, from which two random subsamples of sizes M_1 and M_2 were taken. The mean value of $\mathcal{S}$ is equal to $\langle\mathcal{S}\rangle = \sum_s (1-(1-f(s))^{M_1})(1-(1-f(s))^{M_2})$, where $f(s)$ is the true (and unknown) frequency of sequence s. A naive estimate of $\langle mS\rangle$ may be obtained by repeatly resampling subsets of sizes M_1 and M_2 from the observed repertoire, calculate $\mathcal{S}$ for each draw, and average. One get the same result by replacing $f(s)$ by $\hat{f}_s = n(s)/N$ in the previous formula, where $n(s)$ is the number of s reads in the full dataset, and $N = \sum_s n(s)$. However, this naive estimate leads to a systematic overestimate of the sharing (visible when compared with biological replicates, see Fig. S9), simply because this procedure overestimates the probability of resampling rare sequences, in particular singletons whose true frequency may be much lower that $1/N$. A similar bias occurs when computing $\mathcal{I}$. To correct for this bias, we look for a function $h(n)$ that satisfies for all f:

$$\langle h(n)\rangle \equiv \sum_n \binom{N}{n} f^n (1-f)^{N-n} h(n) = (1-(1-f)^{M_1})\ (1-(1-f)^{M_2}), \tag{9}$$

so that $\langle\mathcal{S}\rangle$ and $\langle\mathcal{I}\rangle$ can be well approximated by:

$$\langle\mathcal{S}\rangle \approx \sum_s h(n(s)), \tag{10}$$

$$\langle\mathcal{I}\rangle \approx -\sum_s h(n(s))\left[\ln(1/P_{\text{gen}}(s)) - \gamma\right]. \tag{11}$$

Expanding the right-hand side of Eq. 9 into 4 terms, we find that $h(n) = 1 - g_{M_1}(n) - g_{M_2}(n) + g_{M_1+M_2}(n)$ satisfies Eq. 9 provided that:

$$\sum_n \binom{N}{n} f^n (1-f)^{N-n} g_M(n) = (1-f)^M. \tag{12}$$

Under the change of variable $x = f/(1-f)$, the expression becomes:

$$\sum_n \binom{N}{n} x^n g_M(n) = (1+x)^{N-M} = \sum_n \binom{N-M}{n} x^n. \tag{13}$$

Identifying the polynomial coefficients in x^n on both sides yields:

$$g_M(n) = \binom{N-M}{n} \Big/ \binom{N}{n}. \tag{14}$$

These corrected estimates agree with the direct estimates using biological replicates (Fig. S9).

Similarly, $\langle\mathcal{S}\rangle$ and $\langle\mathcal{I}\rangle$ in heterologous samples can be estimated using:

$$\langle\mathcal{S}\rangle \approx \sum_s [1-g_{M_1}(n(s))][1-g_{M_2}(n'(s))], \tag{15}$$

$$\langle\mathcal{I}\rangle \approx \sum_s [1-g_{M_1}(n(s))][1-g_{M_2}(n'(s))]\left[\ln(1/P_{\text{gen}}(s)) - \gamma\right]. \tag{16}$$

where $n(s)$ and $n'(s)$ are the empirical counts of sequence s in the two samples.

Theoretical upper bound on heterologous scores

When the two samples were extracted from two different individuals (heterologous scenario), we can use the universality of the recombination process to give upper bounds on the values of $\mathcal{S}$ and $\mathcal{I}$. These bounds are represented by the dashed lines in Fig1c). If two samples of respectively M_1 and M_2 cells, containing $T_1 \leq M_1$ and $T_2 \leq M_2$

unique sequences are extracted from two different individuals, the number of shared sequences between them is given by [21]:

$$\langle \mathcal{S} \rangle_{\text{heterologous}} \leq \sum_s \left(1-(1-p(s))^{T_1}\right)\left(1-(1-p(s))^{T_2}\right) \tag{17}$$

$$\lessapprox T_1 T_2 \sum_s p(s)^2 = T_1 T_2 \langle p(s) \rangle \leq M_1 M_2 \langle p(s) \rangle. \tag{18}$$

$p(s)$ is the probability of finding a sequence s in the blood. Following [21], we make the approximation $p(s) = P_{\text{gen}}(s) q^{-1}$, where the $q = 0.01$ factor is the probability that a generated sequences passes selection. Then $\langle p(s) \rangle$ can be estimated from the mean over generated sequences. Similarly, we can estimate $\mathcal{I}$ as

$$\langle \mathcal{I} \rangle_{\text{heterologous}} \lessapprox T_1 T_2 \sum_s p(s)^2 \left[\ln\left(1/P_{\text{gen}}(s)\right) - \gamma\right] \tag{19}$$

$$= -T_1 T_2 \langle p(s)[\gamma + \ln(q p(s))] \rangle \leq -M_1 M_2 \langle p(s)[\gamma + \ln(q p(s))] \rangle, \tag{20}$$

which is also estimated from the mean over generated sequences.

Error rate estimates

To make the quantitative predictions shown in Fig. 1, we need to constrain the tail behavior of the distributions of $\mathcal{S}$ and $\mathcal{I}$, for the two scenarios.

The $\mathcal{S}$ statistic can be rewritten as a sum of Bernouilli variables over all possible sequences, each with a parameter corresponding to its probability of being present in both samples, either in the autologous or the heterologous case. Therefore $\mathcal{S}$ is a Poisson binomial distribution, a sum of independent Bernouilli variables with potentially different parameters. The variance and tails of that distribution are bounded by those of the Poisson distribution with the same mean, denoted by m_a for the autologous case, and m_h for the heterologous case (Fig. S10).

Thanks to that inequality, the rates of false negatives and false positives for a given threshold r are bounded by:

$$P(\mathcal{S} < r | \text{autologous}) \leq Q(r+1, m_a), \qquad P(\mathcal{S} > r | \text{heterologous}) \leq 1 - Q(r+1, m_h), \tag{21}$$

where Q is the regularized gamma function, which appears in the cumulative distribution function of the Poisson distribution. The mean autologous score m_a is estimated from experimental data: we use the smallest value of $\langle \mathcal{S} \rangle$ in the Emerson dataset and Eq. 10. To compute m_h, we use the semi-theoretical prediction made using the universality of the recombination process, Eq. 17.

Similarly, $\mathcal{I}$ can be viewed as a sum of $\mathcal{S}$ independent random variables, all following the distribution of $\ln(1/P_{\text{gen}}) - \gamma$. However, this distribution differs in the two scenarios. Sequences shared between more than one donor have an higher probability of being generated, their $\ln(P_{\text{gen}})$ distribution has higher mean and smaller variance (Fig. S11).

The sum is composed of a relatively large number of variables in most realistic scenarios. Hence, we rely on the central limit theorem to approximate it by a normal distribution, of mean and variance proportional to $\mathcal{S}$. Explicitly:

$$P(\mathcal{I} < r | \text{autologous}) = \frac{1}{2} \sum_{\mathcal{S}=0}^{\infty} \frac{(m_a)^{\mathcal{S}} e^{-m_a}}{\mathcal{S}!} \left(1 + \text{erf}\left(\frac{r - \mathcal{S}\langle \ln(1/P_{\text{gen}}) - \gamma \rangle}{\sqrt{2\mathcal{S}\text{Var}[\ln(1/P_{\text{gen}}) - \gamma]}}\right)\right), \tag{22}$$

$$P(\mathcal{I} > r | \text{heterologous}) = \frac{1}{2} \sum_{\mathcal{S}=0}^{\infty} \frac{(m_h)^{\mathcal{S}} e^{-m_h}}{\mathcal{S}!} \left(1 - \text{erf}\left(\frac{r - \mathcal{S}\langle \ln(1/P_{\text{gen}}) - \gamma \rangle_{\text{shared}}}{\sqrt{2S\text{Var}[\ln(1/P_{\text{gen}}) - \gamma]_{\text{shared}}}}\right)\right). \tag{23}$$

The AUROC are computed based on these estimates, by numerically integrating the true positive rate $P(\mathcal{S}, \mathcal{I} < r | \text{heterologous})$ with respect to the false negative rate $P(\mathcal{S}, \mathcal{I} < r | \text{autologous})$ as the threshold r is varied.

Modeling the evolution of autologous scores

We use the model of Ref. [38] to describe the dynamics of individual T-cell clone frequencies f under a fluctuating growth rate reflecting the changing state of the environment and the random nature of immune stimuli:

$$\frac{df}{dt} = \left[-\frac{1}{\tau} + \frac{1}{2\theta} + \frac{1}{\sqrt{\theta}}\eta(t)\right] f(t), \tag{24}$$

where $\eta(t)$ is a Gaussian white noise with $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$.

With the change of variable $x = \ln(f)$, these dynamics simplify to a simple Brownian motion in log-frequency: $\partial_t x = -\tau^{-1} + \theta^{-1/2}\eta(t)$. In that equation, τ appears as the decay rate of the frequency, while θ is the timescale of the noise, interpreted as the typical time it takes for the frequency to rise or fall by a logarithmic unit owing to fluctuations. Considering a large population of clone, each with their independent frequency evolving according to Eq. 24, and a source term at small f corresponding to thymic exports, one can show that the steady-state probability density function of f follows a power-law [38], $\rho(f) \propto f^{-\alpha}$, with exponent $\alpha = 1 + 2\theta/\tau$. α was empirically found to be ≈ 2 in a wide variety of immune repertoires [10, 47–49], implying $2\theta \approx \tau$. The turn-over time τ is unknown, and was varied from 1/2 year to 10 years in the simulations.

We simulated the evolution of human TRB repertoires by starting with the empirical values of the frequencies of each observed clones, $f(s, 0) = \hat{f}(s, 0) = n(s, 0)/N$ from the analysed datasets. A sample of size M was randomly selected with respect to these frequencies, and the frequencies of the clones captured in that sample were then evolved with a time-step of 2 days using Euler-Maruyama's method, which is exact in the case of Brownian motion. Clones with frequencies falling below 10^{-11} (corresponding to a single cell in the organism) were removed. At each time $t > 0$, we measured the mean value of $\mathcal{S}$ with the formula $\sum_s (1 - (1 - f(s, t))^M)$ where the sum runs over the sequences captured in the initial sample.

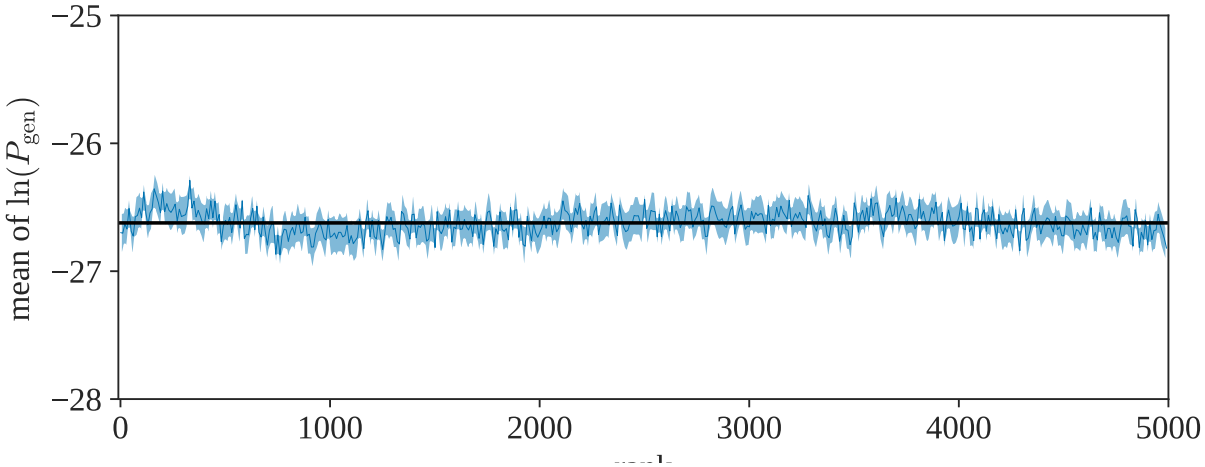

FIG. S1. Mean value of $\log P_{\text{gen}}$ as a function of the rank of the clonotype (from most abundant to least abundant) from [17]. The black line represents the mean of $\log P_{\text{gen}}$ for naive clones. The statistic for the top-clones (low rank) is similar to the one for the naive clones.

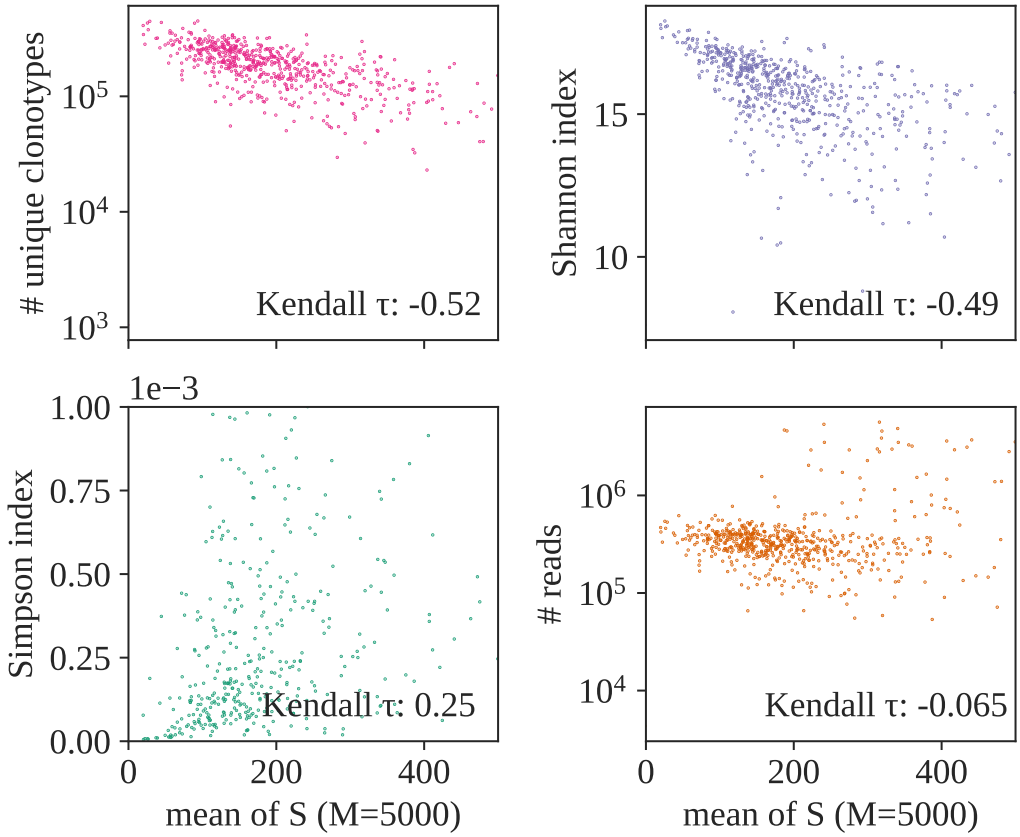

FIG. S2. Comparison between the mean of $\mathcal{S}$ (autologous case), and three common diversity measures: the number of unique sequences found in the dataset (top left), the Shannon index, $-\sum \hat{f}_s \ln \hat{f}_s$ (top right), the Simpson index (bottom left), and the total number of reads in each datasets (bottom right). All the diversity measures show a strong correlation with $\mathcal{S}$, but the correlation with the sequencing depth is low.

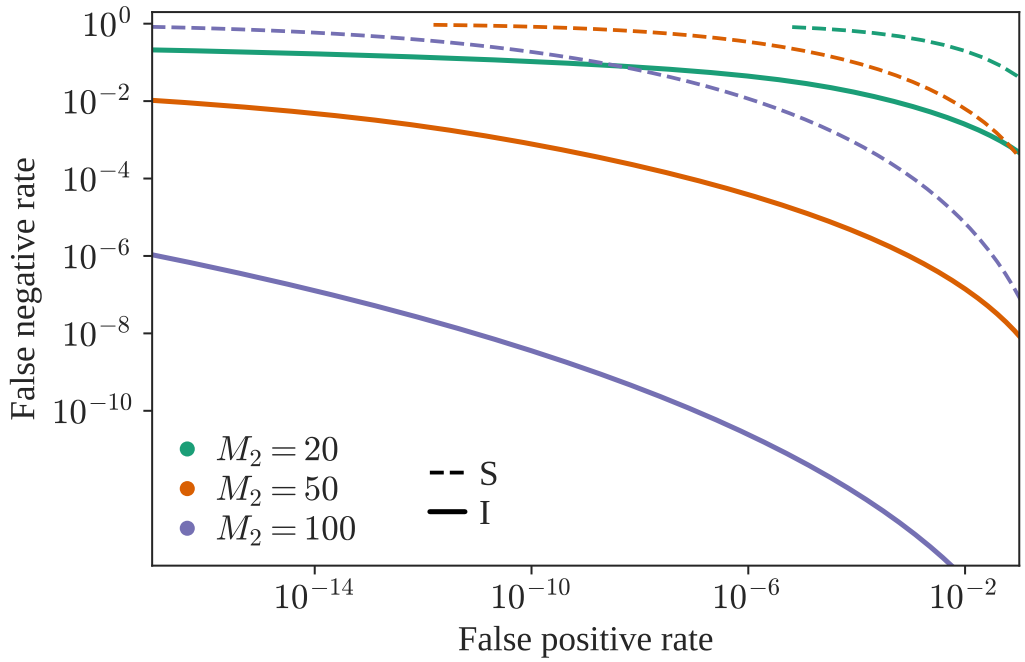

FIG. S3. Detection Error Trade-off (DET) graph for both summary statistics, between a large sample (full dataset, $M_1 = 10^6$) and a smaller one, of size $M_2 = M$.

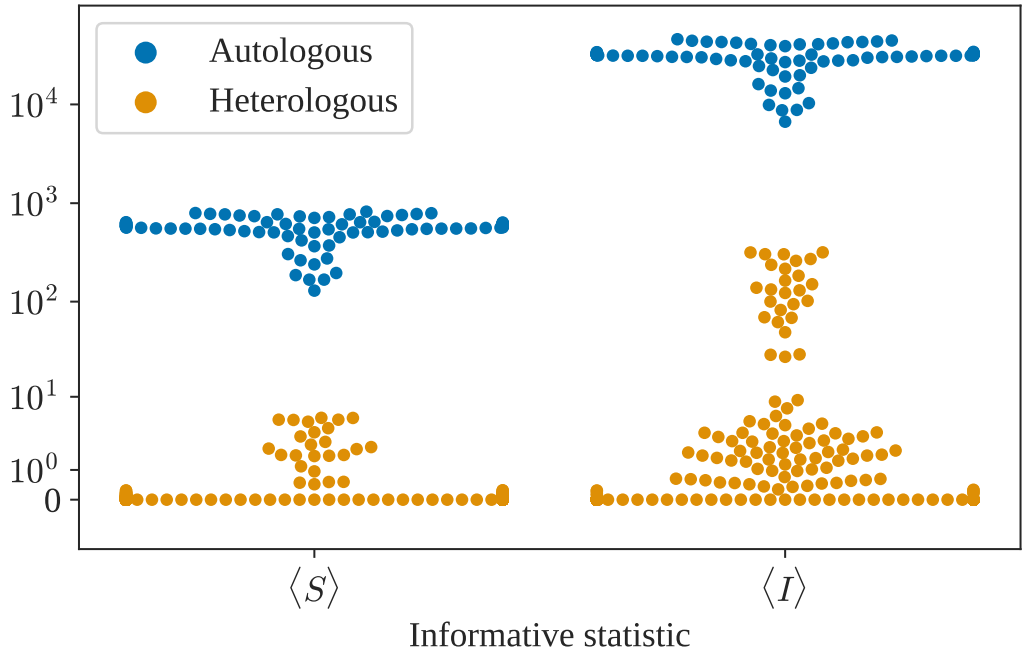

FIG. S4. Mean value of $\mathcal{S}$ (left) and $\mathcal{I}$ (right) for the IGH chain, part of the B-cell receptor, for sample sizes $M = 5000$. The IGH sequences used are restricted to IgG+ B-cells (selected according to their CH gene). The sequences were obtained from 8 different individuals (6 biological replicates each) in the dataset from [30]. Autologous (blue) and heterologous (yellow) are well separated.

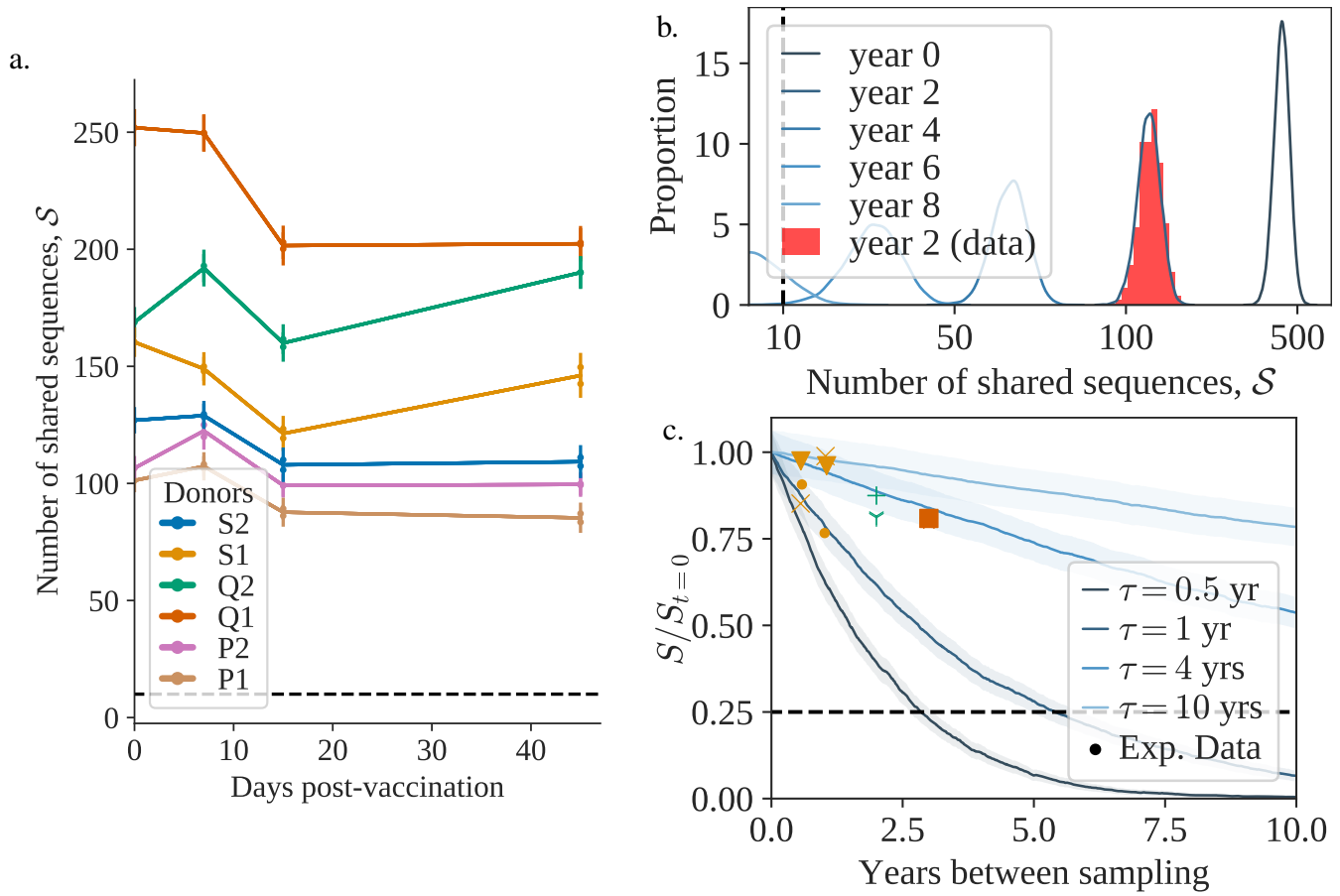

FIG. S5. a) Evolution of $\mathcal{S}$ ($M = 5000$) during vaccination, between a sample taken at day 0 (vaccination date) and at a later timepoint. Each color represents a different individual. Each pair timepoint/individual has two biological replicates. The dashed line represents the threshold value. b) Evolution of $\mathcal{S}$ between a sample taken at year 0 and a later timepoint. The red histogram corresponds to one of the individuals sampled in [18] and the blue curves show theoretical estimates, fitted to match ($\tau = 0.66$). c) Evolution of the (normalized) mean of $\mathcal{S}$ ($M = 5000$) as a function of time for different values of the turnover rate τ. The dashed line represents the threshold value divided by the smallest value of $\mathcal{S}_{t=0}$ ($M = 5000$) in the data. The data points were obtained from the datasets [31] (yellow), [18] (green) and [22] (orange). Different markers indicate different individuals.

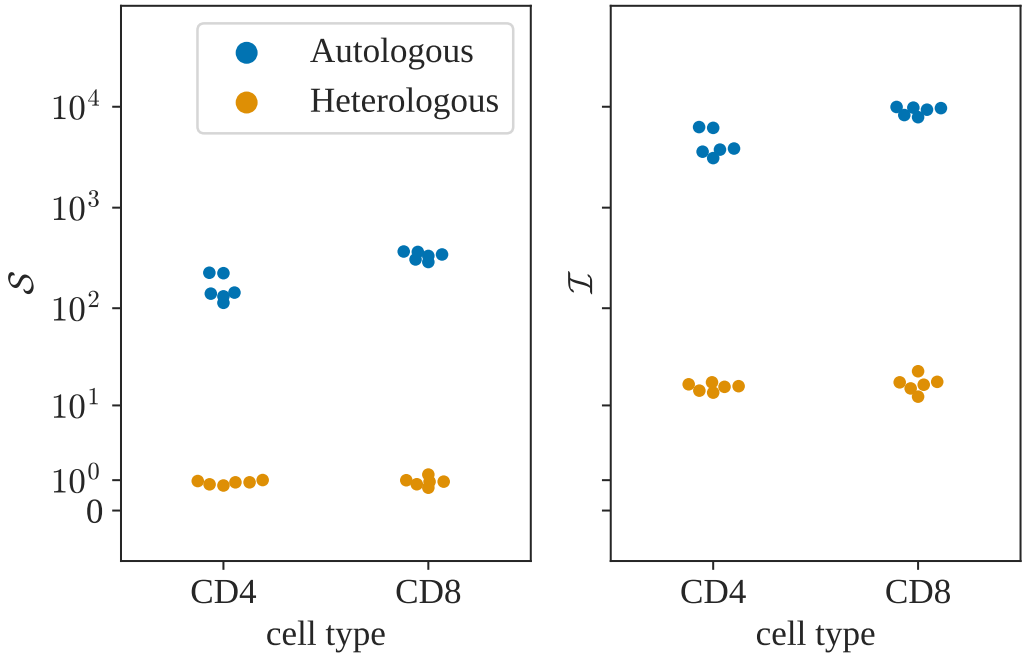

FIG. S6. Distribution of $\mathcal{S}$ (left) and $\mathcal{I}$ (right) for CD4+ and CD8+ T-cells in the dataset from [18] (6 individuals). In the autologous case (blue), there is a significant difference between the two cell types, caused by the differences between their clonal distributions.

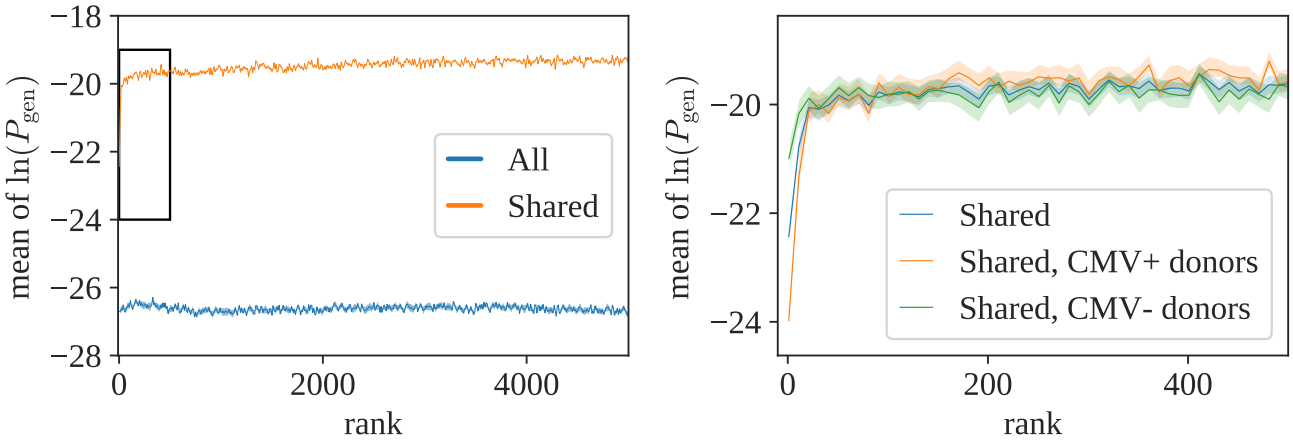

FIG. S7. Left: Mean value of P_{gen} as a function of the rank of the clonotype, for generic sequences (blue) and sequences shared between more than two donors (orange). The mean stays flat indicating that the probability of being generated does not generally depend on the clonotype size. There is an exception (black rectangle), shown as a close-up on the right panel. The top twenty clones, when shared between donors, have a smaller probability of being generated than expected by chance. This difference is likely to be driven by convergent selection against common pathogens, since CMV positive donors show a more prononced effect than CMV negative ones.

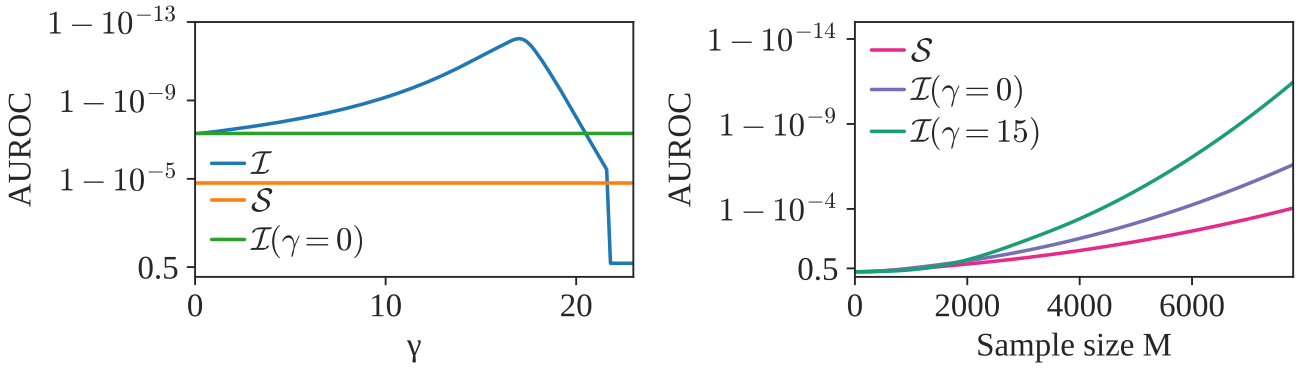

FIG. S8. Left panel: AUROC (Area Under Receiver Operating Characteristic) of $\mathcal{I}$, as a function of γ ($M = M_1 = M_2 = 5000$). We observe an optimum near $\gamma = 15$. Right panel: AUROC as a function of M, for $\mathcal{S}$, $\mathcal{I}(\gamma = 0)$, and $\mathcal{I}(\gamma = 15)$.

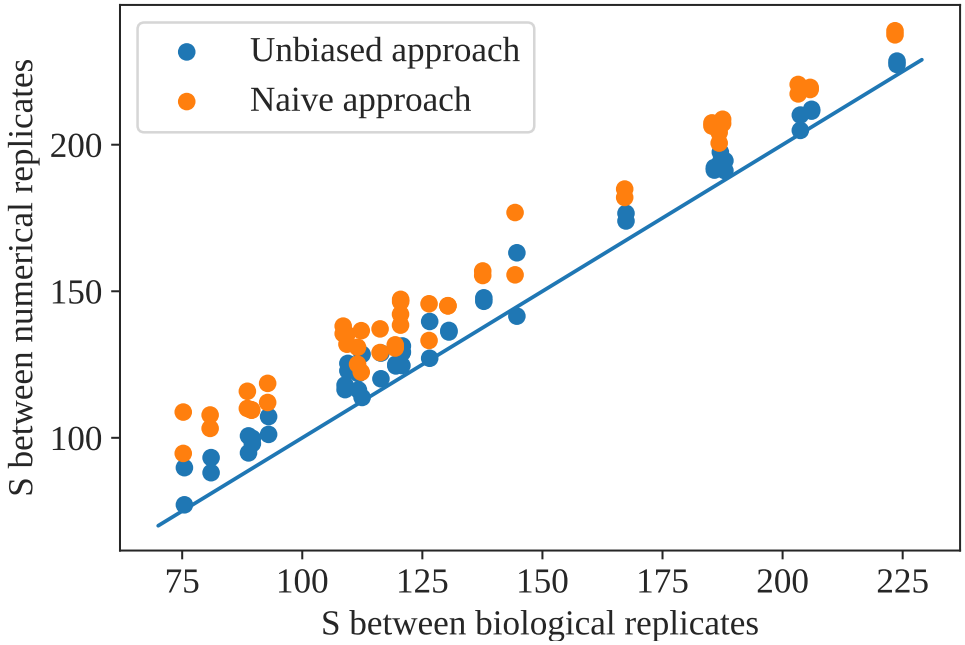

FIG. S9. Naive and corrected estimates of the autologous $\mathcal{S}$ from single datasets, versus its values computed using true biological replicates from Ref. [18].

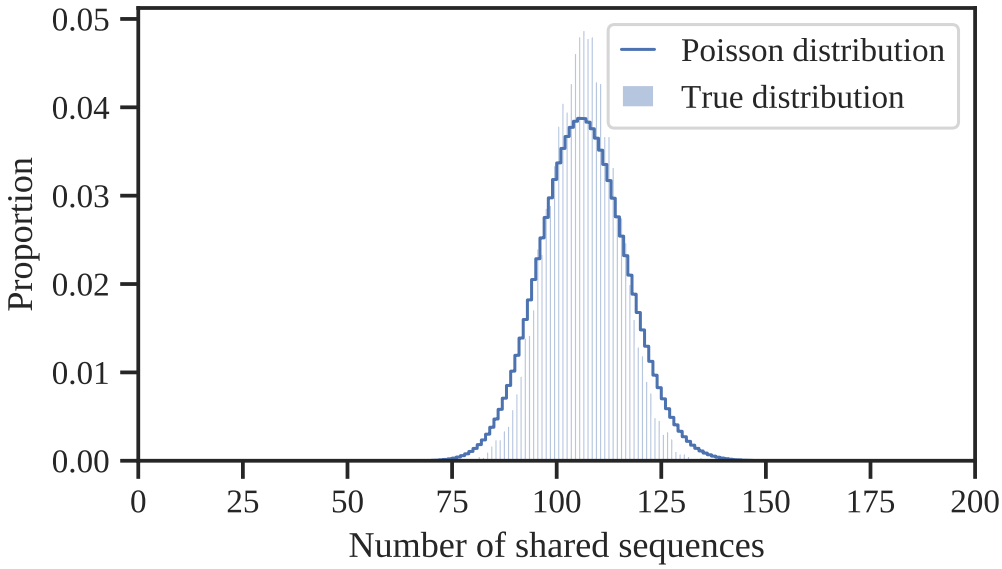

FIG. S10. Comparison between the distribution of $\mathcal{S}$ obtained by computationally and repeatedly downsampling a single repertoire from Ref. [17] with $M = 5,000$ (histogram), and a Poisson distribution of the same mean (full line).

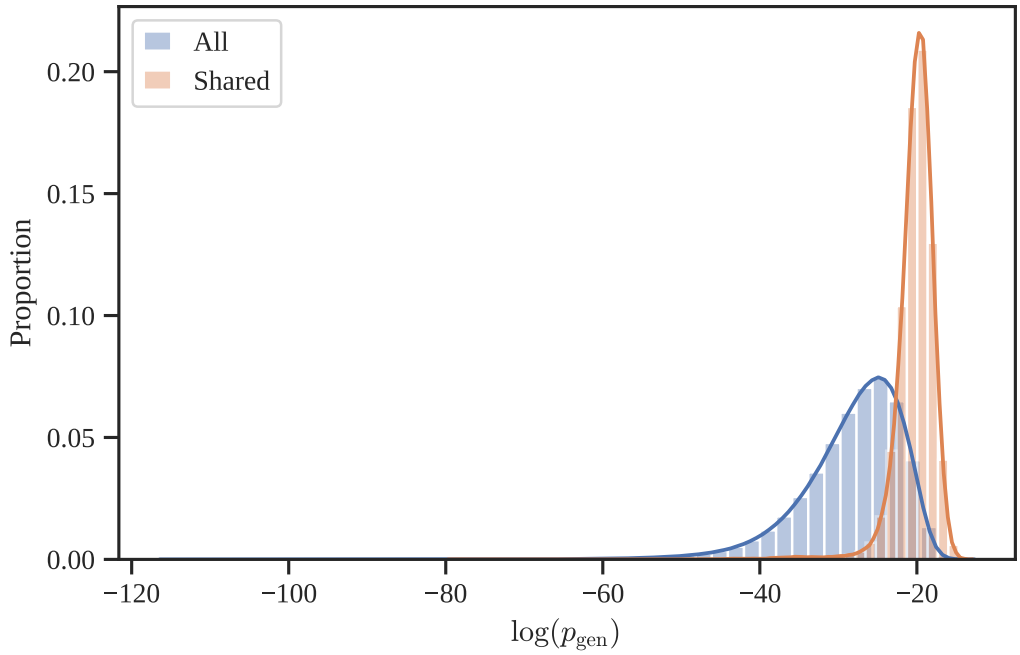

FIG. S11. Distribution of $\ln(P_{\mathrm{gen}})$ for generic sequences, and for sequences shared between heterologous samples.