OXFORD

(GIGA)^n SCIENCE

GigaScience, 2023, **12**, 1–16

DOI: 10.1093/gigascience/giad074

Technical Note

simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods

Chakravarthi Kanduri [1,2,*], Lonneke Scheffer [1], Milena Pavlović [1,2], Knut Dagestad Rand [1], Maria Chernigovskaya [3], Oz Pirvandy [4], Gur Yaari [4], Victor Greiff [3] and Geir K. Sandve [1,2,*]

[1] Centre for Bioinformatics, Department of Informatics, University of Oslo, 0373 Oslo, Norway
[2] UiORealArt Convergence Environment, University of Oslo, 0373 Oslo, Norway
[3] Department of Immunology and Oslo University Hospital, University of Oslo, 0373 Oslo, Norway
[4] Faculty of Engineering, Bar-Ilan University, 5290002, Israel
*Correspondence address. Chakravarthi Kanduri, Centre for Bioinformatics, Department of Informatics, University of Oslo, 0373 Oslo, Norway. E-mail: skanduri@uio.no; Geir K. Sandve, Centre for Bioinformatics, Department of Informatics, University of Oslo, 0373 Oslo, Norway. E-mail: geirksa@ifi.uio.no

Abstract

Background: Machine learning (ML) has gained significant attention for classifying immune states in adaptive immune receptor repertoires (AIRRs) to support the advancement of immunodiagnostics and therapeutics. Simulated data are crucial for the rigorous benchmarking of AIRR-ML methods. Existing approaches to generating synthetic benchmarking datasets result in the generation of naive repertoires missing the key feature of many shared receptor sequences (selected for common antigens) found in antigen-experienced repertoires.

Results: We demonstrate that a common approach to generating simulated AIRR benchmark datasets can introduce biases, which may be exploited for undesired shortcut learning by certain ML methods. To mitigate undesirable access to true signals in simulated AIRR datasets, we devised a simulation strategy (simAIRR) that constructs antigen-experienced-like repertoires with a realistic overlap of receptor sequences. simAIRR can be used for constructing AIRR-level benchmarks based on a range of assumptions (or experimental data sources) for what constitutes receptor-level immune signals. This includes the possibility of making or not making any prior assumptions regarding the similarity or commonality of immune state–associated sequences that will be used as true signals. We demonstrate the real-world realism of our proposed simulation approach by showing that basic ML strategies perform similarly on simAIRR-generated and real-world experimental AIRR datasets.

Conclusions: This study sheds light on the potential shortcut learning opportunities for ML methods that can arise with the state-of-the-art way of simulating AIRR datasets. simAIRR is available as a Python package: https://github.com/KanduriC/simAIRR.

Keywords: simulation of AIRR data, shortcut learning, benchmarking of machine learning methods, adaptive immune receptor repertoires, AIRR, ML

Background

High-throughput sequencing of adaptive immune receptor (AIR) repertoires (AIRRs), including B-cell and T-cell receptors (BCRs and TCRs), can provide a snapshot of ongoing and past immune responses [1–7]. Decoding the information specific to various immune responses embedded in AIRRs has recently seen a surge in interest because of its potential to aid the development of immunodiagnostics and therapeutics [1–7]. Receptor sequence sharing between repertoires in a population (so-called public responses) can be due to antigen selection or antigen-independent mechanisms (like convergent recombination and other recombination biases) [6–12] and can vary further depending on the species, cell types, cell subsets, chains, and pairedness of sequences [13]. Previous studies have shown that sequences selected for a common antigen share similarities in sequence patterns and can be detected in multiple individuals who experienced the antigen (e.g., see Table 1 in reference [7] for examples of public AIRs in various diseases in humans). Such public response sequences (defined by sequence identity or similarity [14]) have aided in the classification of diseases or immune states [5].

The pattern recognition capacity of machine learning (ML) methods has been increasingly utilized to learn the sequence patterns associated with immune states [5, 15, 16]. Many studies continue to develop and apply classical ML (supervised and unsupervised) and modern deep learning methods to learn complex sequence patterns that can distinguish immune states [17–35]. The continued rise in the development and application of AIRR-ML methods warrants rigorous benchmarking to compare the performance of methods. This requires a combination of suitable real-world experimental datasets and simulated benchmark datasets with known true signals [36, 37] (hereafter signal refers to sequence patterns in AIRR-seq data that distinguish immune states). There are currently only a few large-scale datasets of immune state–associated donor repertoires available, and even these are of limited size (<1,000 donors) and offer limited knowledge of true signals (only donor-level annotation, with no

Received: February 21, 2023. **Revised:** July 20, 2023. **Accepted:** August 29, 2023

© The Author(s) 2023. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

true information at the individual receptor level) [19, 38–40]. Therefore, simulated AIRR datasets with artificially introduced discriminative sequence patterns at the individual receptor level play a central role in the rigorous evaluation of AIRR-ML methods [36, 41–46].

A principal observation regarding the failure of modern ML methods is that many are related to unintended "shortcut" strategies adopted by ML methods [47]. Shortcuts can be defined as decision rules that work well on selected benchmark datasets but fail to generalize to other real-world datasets [47]. In the context of AIRR-ML, the field will benefit not only by avoiding shortcut strategies in discriminative learning but also by generating synthetic AIRR datasets devoid of shortcut opportunities. Below we describe a notable shortcut opportunity in simulated AIRR datasets that arise as a result of the state-of-the-art simulation approaches but is absent in real-world experimental datasets.

The AIRRs of a study cohort with a common immune state can be categorized into 2 components: (i) private sequences, which are seen only in one individual of the cohort, and (ii) public sequences, which are observed in more than one individual in the cohort. Previous studies have suggested 2 main mechanisms that determine the interindividual sharing of sequences: (i) convergent recombination, where owing to the biases of current stochastic V(D)J recombination models [48, 49], the probability of generating certain sequences is high and, thus, such sequences are observed in multiple individuals [6, 50–53], and (ii) selection/bias in receptor usage, where identical or similar sequences are observed in numerous individuals that share a common immune state due to being selected for a common antigen [17, 18, 22, 30, 54, 55]. The observed publicness of sequences in a cohort (how frequent they are) is known to depend on the sampled cohort size and sequencing depth [55, 56].

The probability of generating a specific CDR3 sequence (often called *generation probability* in the AIRR context) is the sum of the probabilities of all recombination events that can generate the specific sequence [57]. The generation probability distributions of private sequences and public sequences in any sampled cohort differ considerably [55], where the population incidence of sequences increases monotonically with an increase in generation probability (Fig. 1A). By comparing the population incidence and generation probability of sequences, it may be possible to identify sequences that are observed with unlikely high population incidence given their generation probability (Fig. 1A) [20, 30]. Such sequences can potentially be immune state associated and will be referred to as outlier sequences throughout this article. Since antigen-experienced repertoires may carry several immune state–associated receptor sequences accumulated over time (specific to distinct antigens), mining public sequences in antigen-experienced repertoire cohort may reveal several outlier sequences irrespective of any particular immune state (Fig. 1B). In a real-world setting, there is thus no trivial relation between the presence of outlier sequences in a repertoire and a given immune state of interest. However, constructing synthetic AIRR datasets by sampling from known V(D)J recombination models alone [41, 43–45] will result in naive repertoires that have not experienced any immune events and thus do not carry multiple outlier sequences like antigen-experienced repertoires do (Fig. 1B). When a selected group of the simulated naive repertoires is enriched for sequence patterns to represent an immune state for prediction methods, the publicity versus generation probability relation can in itself make these repertoires stand out through being the only repertoires in the simulated dataset containing outlier sequences and thus provide shortcut opportunities for ML methods. By learning to directly connect the presence of outlier sequences to a particular immune state, a predictive method could perform very well on a benchmark based on a prediction strategy that exploits simulation artifacts instead of learning immune state–associated signals that are relevant for real-world applications. This issue, which is referred to as shortcut learning in the machine learning field, is known to lead to a lack of generalization and unintuitive failures of ML methods and has been suggested to be one of the main barriers to robust, fair, trustworthy, and deployable machine learning [47]. For the sake of convenience, we hereafter refer to the bias of discordance between introduced signal and baseline repertoires in simulated AIRR data as *generation probability discordance bias*.

In this study, we investigate the shortcut opportunity arising from generation probability discordance bias in simulated AIRR datasets and show that such shortcuts are absent in tested real-world experimental datasets. To mitigate the demonstrated shortcut opportunity in AIRR datasets, we present the simulation strategy, simAIRR. simAIRR provides a systematic approach for simulating AIRR datasets according to the assumptions that immune receptor binding is determined either by (i) the full CDR3 sequences and is best addressed by learning appropriate similarity metrics for full CDR3s or by (ii) subsequence patterns such as *k*-mers. We also present case studies to demonstrate the utility of simAIRR, where simulated AIRR datasets are generated using both full-CDR3 assumption and sub-CDR3 motif assumption (*k*-mers) and subsequently used for evaluating suitable ML methods as a function of different witness rates and sample sizes. Here, witness rate refers to the rate at which signal occurs in the positive class *examples*. Note that the italicized term *examples* commonly used in ML literature refers to repertoires throughout this article.

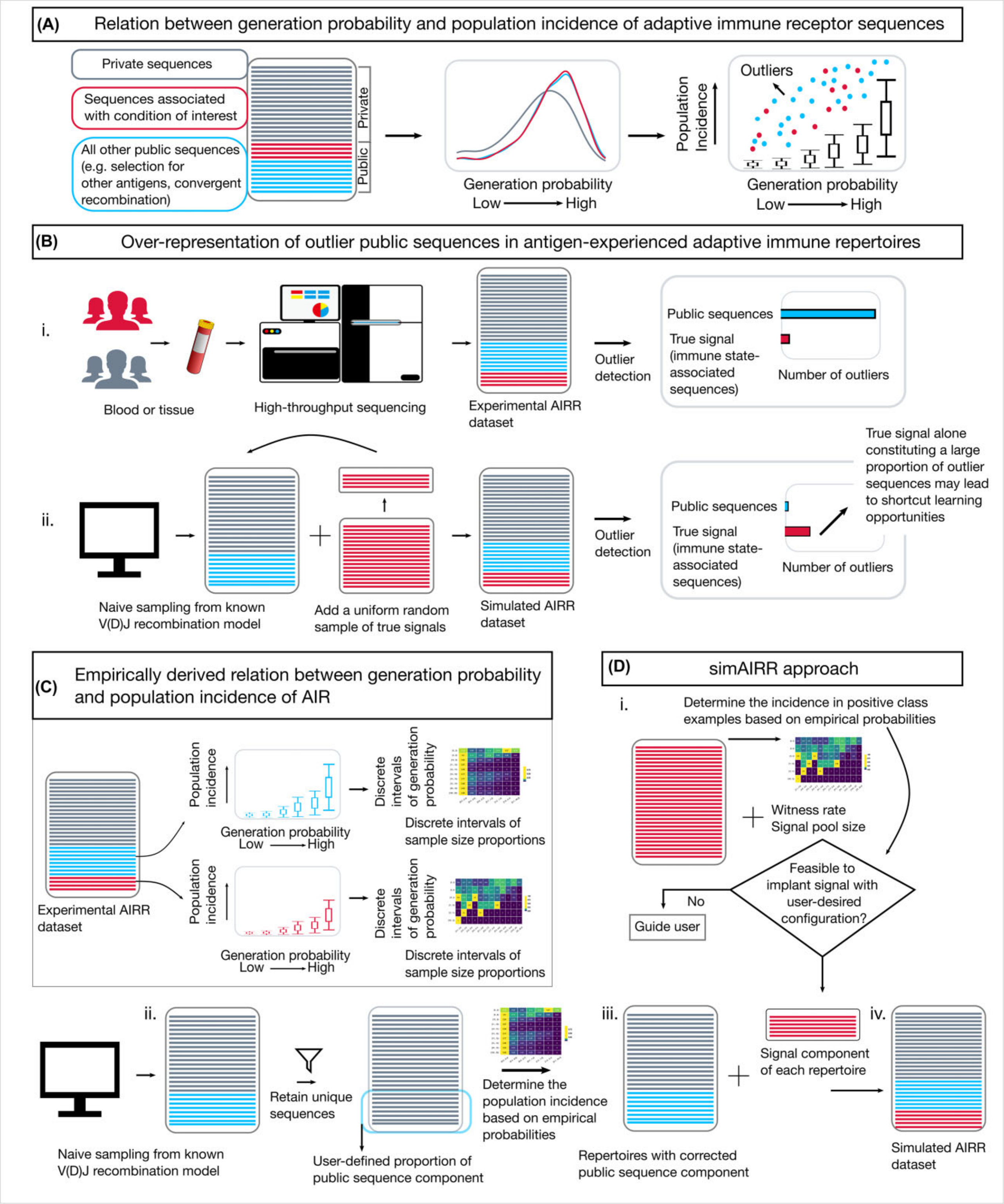
Analyses

Generation probability discordance bias in simulated AIRR datasets leads to unintended shortcut opportunities for ML methods

Throughout the article, we refer to AIRs that are observed with unlikely high population incidence in a study cohort given their generation probability as outlier sequences. As an example of unlikely high population incidence given generation probability, consider a case where an AIR has a very low probability of occurring in >1% of a population sample but is rather observed in 10% of the population sample. We refer to the disparity in generation probability distributions between true signals and remaining public sequences in synthetic AIRR datasets as generation probability discordance bias. Here, true signals refer to the sequences that from the outset are known to be immune state associated. In simulated datasets, true signals are those sequences introduced into the repertoires that distinguish the immune states. In the real-world experimental dataset used in this study, we refer to the original study-reported sequences as the true signals, for which the rationale is provided further below.

We hypothesized that, unlike real-world experimental datasets of antigen-experienced repertoires, simulated AIRR datasets may not carry several outlier sequences. Thus, the subsequent introduction of true signals representing an immune state will lead to shortcut opportunities that ML methods can exploit. To investigate whether generation probability discordance bias exists in real-world experimental and simulated AIRR datasets, we computed an outlier score for each public sequence of the experimental and simulated AIRR datasets irrespective of the immune state

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024


(A) Relation between generation probability and population incidence of adaptive immune receptor sequences
Private sequences
Sequences associated with condition of interest
All other public sequences (e.g. selection for other antigens, convergent recombination)
Private
Public
Generation probability
Low → High
Population Incidence
Outliers
Generation probability
Low → High
(B) Over-representation of outlier public sequences in antigen-experienced adaptive immune repertoires
i.
Blood or tissue
High-throughput sequencing
Experimental AIRR dataset
Outlier detection
Public sequences
True signal (immune state-associated sequences)
Number of outliers
ii.
Naive sampling from known V(D)J recombination model
Add a uniform random sample of true signals
Simulated AIRR dataset
Outlier detection
Public sequences
True signal (immune state-associated sequences)
Number of outliers
True signal alone constituting a large proportion of outlier sequences may lead to shortcut learning opportunities
(C) Empirically derived relation between generation probability and population incidence of AIR
Experimental AIRR dataset
Population incidence
Generation probability
Low → High
Discrete intervals of generation probability
Discrete intervals of sample size proportions
Population incidence
Generation probability
Low → High
Discrete intervals of generation probability
Discrete intervals of sample size proportions
(D) simAIRR approach
i.
Determine the incidence in positive class examples based on empirical probabilities
Witness rate
Signal pool size
Feasible to implant signal with user-desired configuration?
No
Guide user
ii.
Naive sampling from known V(D)J recombination model
Retain unique sequences
User-defined proportion of public sequence component
Determine the population incidence based on empirical probabilities
iii.
Repertoires with corrected public sequence component
Signal component of each repertoire
iv.
Simulated AIRR dataset

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

Figure 1: Rationale and simAIRR workflow. (A) The relation between the generation probability and population incidence of AIR sequences in a study cohort could be used to gauge outlier sequences that are observed with unlikely high population incidence given their generation probability. (B) (i) Because of accumulating multiple immune state–associated signals over time by being exposed to common antigens, repertoires in real-world experimental AIRR datasets may harbor many outlier public sequences. (ii) However, the same phenomenon is not naturally occurring in synthetic AIRR datasets. Any label-associated signal introduced into synthetic repertoires as shown in ii may alone stand out as outlier sequences, unlike in experimental repertoires. This provides unintended shortcut opportunities for AIRR-ML methods to detect the introduced signal. (C) To mitigate unintended shortcut learning opportunities, we devised a simulation approach that relies on empirically derived relation between generation probability and population incidence of AIRs calibrated separately for signal sequences and remaining public sequences. (D) (i) simAIRR assesses the feasibility of introducing the user-supplied signal into the repertoires at the desired witness rates and, if deemed feasible, generates AIRR datasets with realistic receptor sequence sharing between repertoires. If the signal introduction was deemed infeasible in (i), simAIRR provides descriptive information to the user to act as guidance in reconfiguring the simulation. simAIRR could be used to execute the whole workflow (i, ii, iii, iv) in a sequence or exclusively to perform (i) or (ii) or (iii).

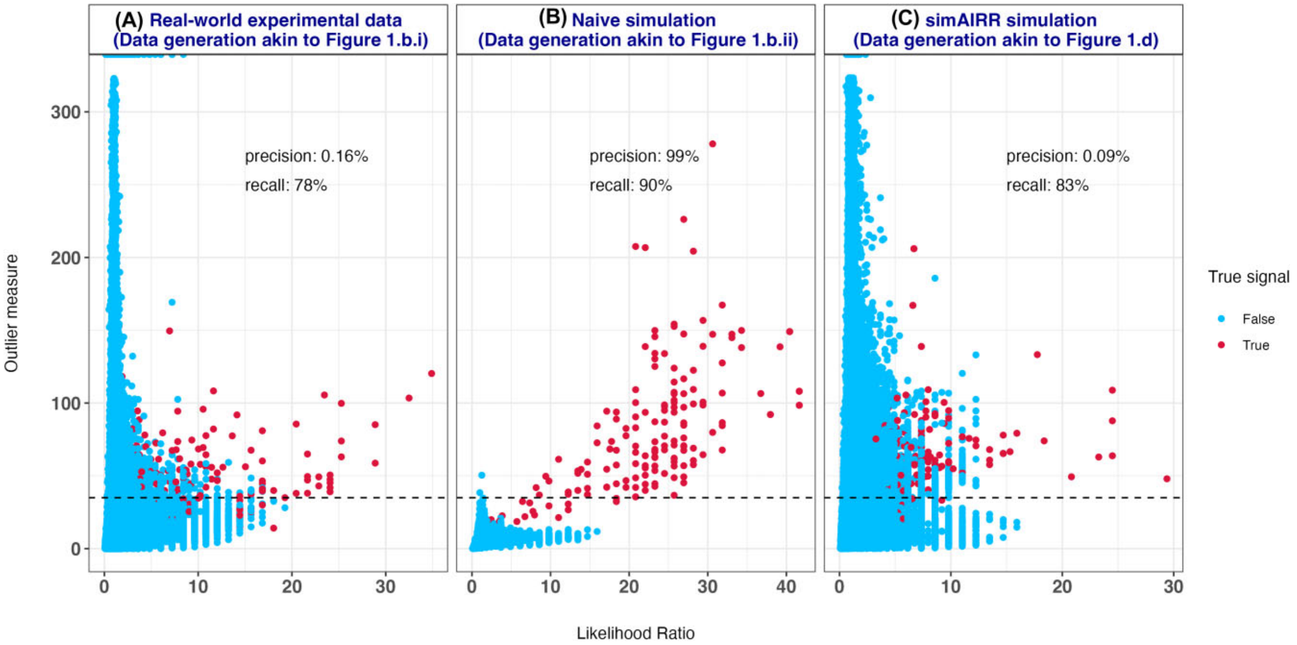


Figure 2: Generation probability discordance bias in the naive simulation of AIRR datasets. (A) In a real-world experimental AIRR dataset ($n = 683$ repertoires; 307 positive class examples and 376 negative class examples, average unique number of TCRβs per repertoire ~200,000), the outlier measure (y-axis) alone is a poor classifier of presumed true signals as reported in the original study. Here, each point represents a unique TCRβ CDR3 sequence and the colors represent whether the sequence is a true signal e. Red points are the true signal sequences. (B) On the contrary, in a synthetic dataset generated through a naive but intuitive simulation approach as shown in Fig. 1B.ii ($n = 683$ repertoires; 310 positive class and 373 negative class examples), the outlier measure alone has a very high precision (99%). This behavior of the simulation approach can lead to shortcut learning opportunities for ML methods. (C) Our novel simulation approach that is intended to mitigate the shortcut learning opportunity makes the repertoires antigen-experienced-like to behave more like the real-world experimental data, where there can be many outlier sequences because of accumulating many immune state–associated sequences over the lifetime of repertoires. On a dataset with matched sample size ($n = 683$ repertoires; 310 positive class and 373 negative class examples), our simulation approach reduces shortcut learning opportunities because thresholding on the outlier measure alone has a precision of 0.09%, which is comparable to the 0.16% precision of the real-world experimental data.

label (see Methods). The outlier score that we computed is qualitatively similar to the methodology by Pogorelyy et al. [20]. In addition to the outlier score, we computed a likelihood ratio for each public sequence (see Methods) that compares the probability of incidence of a sequence in contrasting immune states (positive and negative class labels in ML terminology). For this analysis, we specifically used 3 different datasets: (i) a real-world experimental T-cell repertoire dataset with known cytomegalovirus (CMV) serostatus [19], (ii) a human TCRβ sequence dataset simulated using a naive simulation approach (here, naive simulation refers to sampling sequences from known V(D)J recombination models to construct synthetic repertoires and subsequent introduction of signals in a fraction of the repertoires to represent contrasting immune states; see Fig. 1B for a depiction), and (iii) a human TCRβ sequence dataset simulated using the simulation approach that we developed, simAIRR.

In simulated datasets, we know, by construction, the true signals that differ between immune states. In real-world experimental data, we are aware that experimental artifacts and other study design aspects could impact the selection of immune state–associated signals, thus affecting what can be perceived as true signals. However, we considered the list of signals reported in the original study [19] as true signals for this analysis, as the purpose of this analysis is only to obtain an indication of the disparity between true signals and the remaining public sequences in terms of the degree of being outliers.

In the real-world experimental dataset [19], a decision rule based on thresholding the outlier measure that we computed (e.g., outlier measure >35) is observed to have very low precision (0.16%) and 78% recall (Fig. 2A) in retrieving the true signals. This indicates that the outlier measure alone is a poor classifier of the perceived true signals in real-world experimental data. On the contrary, in a naive-simulated AIRR dataset of TCRs, the same outlier measure-based decision rule is found to have 99% precision and 90% recall (Fig. 2B) in retrieving the true signals. This indicates that the outlier measure alone can capture a large fraction of the true signals in naive-simulated data, which can be used for shortcut learning by ML methods. This trend of generation probability discordance bias persisted at sample sizes as low as 50 repertoires in naive-simulated datasets, while subsampled real-world experimental data were devoid of such bias (Supplementary Fig. S1). The phenomenon of the absence of generation probability discordance bias in real-world experimental datasets was also observed on TCRβ data of a much smaller sample size (compared to that of [19]) of 79 cases and 13 controls with/without pediatric COVID-19 from [58] (Supplementary Fig. S2). We also observed a similar phenomenon in the CMV-negative cohort from [19] indicating the persistence of aforementioned findings in CMV-negative cohorts (Supplementary Fig. S3). To overcome the generation probability discordance bias demonstrated in Fig. 2B, we devised a novel simulation approach to construct benchmark AIRR datasets for the immune state prediction problem. In AIRR TCR datasets simulated based on our novel simulation approach, the same outlier measure-based decision rule was found to have 0.09% precision and 83% recall (Fig. 2C). Overall, the very low precision equivalent to that observed in a real-world experimental dataset mitigates the shortcut opportunities that can arise through naive simulation approaches.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

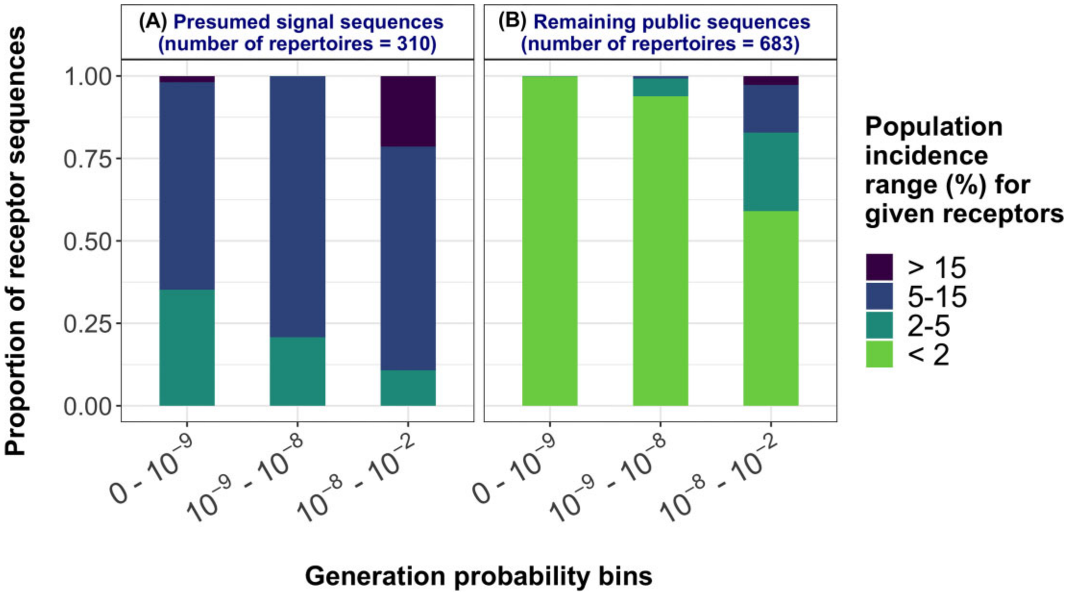


Figure 3: Empirical distribution of population incidence for receptors within different generation probability intervals. The relation between generation probability and population incidence of public AIR sequences was determined based on a previously published large cohort study of TCR repertoires [19] separately for presumed signal sequences (A) and all the remaining public sequences (B). In both A and B, receptors are along the x-axis split into 3 distinct bins according to their generation probability. The stacked bars along the y-axis represent the proportion of receptors within a given generation probability interval that have population incidence falling within a particular range (with a distinct color representing each population incidence range). Note that the population incidence is the proportion of repertoires among a total of the 310 positive repertoires in A, while it is the proportion of repertoires among a total of 683 repertoires in B. Also, note that the bins of both generation probability and sample size proportions are half-open intervals that include the left endpoint but exclude the right endpoint. The stacked bar charts show that a large fraction of the sequences among the nonsignal sequences is observed in a small proportion of the population irrespective of their generation probability (population incidence below 2%). A substantial fraction of the sequences in the higher intervals of generation probability was observed with increased population incidence in signal and nonsignal sequences. The presumed signal sequences were observed with increased population incidence when compared to the remaining public sequences.

A novel simulation approach to mitigate shortcut opportunities for AIRR-ML methods

To mitigate unintended shortcut opportunities for AIRR-ML, we devised a novel simulation approach to generate AIRR datasets that rely on the empirical relation between generation probability and population incidence of public sequences calibrated from real-world experimental datasets separately for signal and other public sequences (Fig. 1C). We hereafter refer to this simulation approach as simAIRR. Below we briefly describe the simAIRR approach (see also Fig. 1D).

simAIRR approach

simAIRR accepts a user-supplied set of AIRs as a potential pool of immune state–associated signals and (i) determines whether it is feasible to introduce the signal sequences into the baseline repertoires at the user-desired witness rate (Fig. 1D). Note that the feasibility of introducing signal sequences depends on the user-supplied set of AIRs, desired witness rate, and the learned models of realistic population incidence of AIR sequences. There may be scenarios where the models of realistic population incidence of AIR sequences may not allow reaching a user-desired witness rate given the set of user-supplied signal sequences. If deemed feasible, simAIRR proceeds to (ii) generate baseline repertoires and (iii) adjust the proportion of public sequences (the sequences that will be shared across repertoires in a dataset) and their population incidence levels. The public component correction is needed because datasets generated through naive sampling lack sequence sharing between repertoires to the degree that is observed in experimental datasets containing antigen-experienced repertoires (see Methods for details). In the repertoires generated with the corrected public sequence component, the relation between generation probability and population incidence is respected. (iv) simAIRR further introduces signal components into the desired number of repertoires, where again the relation between generation probability and population incidence is respected. If the signal introduction was deemed infeasible in (i), simAIRR provides descriptive information to the user to act as guidance in reconfiguring the simulation. simAIRR could be used to execute the whole workflow (i, ii, iii, iv) or exclusively to perform (i) or (ii) or (iii).

To determine the population incidence of public sequences (including both signal sequences and other public sequences), simAIRR relies on the empirical relation between generation probability and the population incidence of AIRs. For this, the user could either calibrate the aforementioned relation based on a real-world dataset of their choice and supply the learned relation to simAIRR or use the default choice that is supplied with simAIRR. Hereafter, we refer to these as models of realistic population incidence of AIR sequences. In simAIRR, the default models of realistic population incidence of AIRs are based on a previously published large cohort study of TCR repertoires [19] calibrated separately for signal sequences (detected in the original study) and other public sequences. See the Methods section for details on learning models of realistic population incidence of AIR sequences. Figure 3 shows the empirical distribution models learned for the signal and other public sequences from the dataset of Emerson et al. [19]. The empirical probability distributions differed between the presumed signal sequences (Fig. 3A) and remaining public sequences (Fig. 3B), where the presumed signal sequences (full AIRs) occurred in a relatively higher fraction of the sampled population, unlike a large majority of the public sequences. The stacked bars show that a large fraction of

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

the public sequences is present in a small fraction of the sampled population. A small fraction of the sequences with high generation probability are unsurprisingly observed in higher fractions of the sampled population. A more granular representation of the learned empirical distribution models is shown in Supplementary Figures S4 and S5.

By analyzing the empirical probability distributions of population incidence for receptors based on datasets of different sample sizes ($n = 50$, 100, 200 repertoires), we observed that the approximation of the true population estimates of realistic receptor sequence sharing improved with increased sample size. We also noticed that a smaller sample size can spuriously inflate the learned population incidence of public sequences (Supplementary Fig. S6). Thus, when simulating moderate to large simulated datasets (e.g., $n \geq 200$) that are typically used when developing/benchmarking ML methods, learning the models based on as large sample sizes as possible can increase the realism of population incidence of public sequences.

Case Studies

To demonstrate how simAIRR may be used to generate benchmark datasets for AIRR-ML predictions, we performed 2 separate case studies.

Case study 1: Prior assumptions on the similarity of immune state–associated sequences

In the first case study, we made prior assumptions on how the immune state–associated sequences that differentiate the positive and negative class labels are similar. Specifically, we assumed that immune state–associated sequences share sequence similarity in the form of shared contiguous amino acid subsequences of size 4 (4-mers).

First, we generated independent AIRR datasets each containing 200 repertoires using simAIRR, where the average sequence count was 119,633 ± 1,313. The datasets were simulated in such a way that 100 repertoires carried condition-associated sequences, hence labeled as a positive class, whereas the negative-labeled repertoires did not receive any condition-associated sequences, although they were not checked for carrying those specific sequences just by chance. We varied the average witness rate in different experiments to observe how the performance of the tested ML method varies depending on the average witness rate. We assumed that the signal sequences carry any one of the three 4-mers: *WKDY*, *YREV*, and *ERFY*. For this, instead of the implantation of k-mers as in our previous study [59], we selected sequences enriched for these 4-mers by querying a large set of reference sequences for matching patterns. To make such a pattern-matching process easy for users if need be, we provide a simple Python script and a corresponding tutorial that shows how to generate a large set of reference sequences and retrieve the pattern-matched sequences of interest with minimal effort and with very few lines of code specification [60]. Notably, we here did not exclude the low probability events of a signal sequence carrying 2 or all the 3 k-mers of interest by chance, but the users can impose such additional sanity checks if need be. The rationale behind the choice of the selected k-mers was to compare the performance of ML methods on benchmark datasets with similar characteristics (witness rates and sample sizes) in our previous study that used the same k-mers. The signal sequences were added to the positive-class repertoires using the simAIRR approach, where the observed frequency of signal sequences in the dataset depends on the generation probability distributions calibrated based on real-world experimental data. Since the signal sequences are assumed to share any of the 3 chosen 4-mers, we used a suitable ML method that matches the assumption that the signal is in the form of 4-mers as in our previous study [59]. Specifically, we used a highly regularized logistic regression model that is well optimized for hyperparameters on a 4-mer encoded representation of the amino acid sequences. To compare with the reported performance metrics of Emerson et al. [19], immuneML case studies [61], and Motifbooster [62], we here chose to present the area under receiver operating characteristic curve (ROC AUC) obtained through nested cross-validation. We observed that the performance of the suitable ML method was close to perfect when the witness rate was equal to or above 10 sequences per 10^5 sequences (Fig. 4A). When the witness rate was 5 sequences per 10^5 sequences, we observed an ROC AUC of around 0.8 on average (Fig. 4A). The observed performance metrics of the ML method at the respective witness rates were in strong alignment with our prior expectations. This is because we previously profiled the performance of the same ML method on datasets across a wide range of witness rates (including the witness rates explored in this case study) and similar assumptions of signal sequences (k-mer sharing) [59] and thus knew beforehand which level of performance to expect. Notably, the performance of ML methods that solely focus on the desired signal and ignore simulation artifacts does not necessarily have to be different between simAIRR-generated data and data generated by other simulation approaches. We have elaborated on the comparable/similar performance of ML models on simAIRR-generated data vs data generated with other simulators in the Discussion section.

Case study 2: No prior assumptions on the immune state–associated sequences

In the second case study, we did not make any prior assumptions regarding the immune state–associated sequences, specifically in what way they are similar. We retrieved sequences that were reported to be specific to the hepatitis C virus (HCV) from VDJdb [54], thereby avoiding making any prior assumptions regarding the similarity of signal sequences. We used a uniform random sample of the HCV-associated sequences as the pool of signal sequences introduced into the positive-labeled repertoires.

Unlike in the first case study, we fixed the witness rate as constant in all the experiments (to have on average 10 immune state–associated sequences per approximately 10^5 sequences). Rather, we varied the sample size (number of examples) of the class-balanced AIRR datasets (200, 400, or 600 examples per dataset) to observe how the performance of the tested ML method varies depending on the sample size. The class balance and signal introduction were similar to case study 1, where 50% of the repertoires carried signal sequences. Since we do not have prior knowledge of the sequence similarity patterns of the signal sequences, we used a suitable ML method that assumes the full sequence identity as a potential signal representation [19]. Specifically, we used a probabilistic binary classifier based on phenotype burden [19] implemented in immuneML [61], where the hyperparameters were selected through nested cross-validation. To compare with the reported performance metrics of Emerson et al. [19], immuneML case studies [61], and Motifbooster [62], we here chose to present ROC AUC obtained through nested cross-validation. The findings of the performance metrics align well with the known behavior of the probabilistic binary classifier method [19], where prior studies reported a drop in the performance of the method at lower sample

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

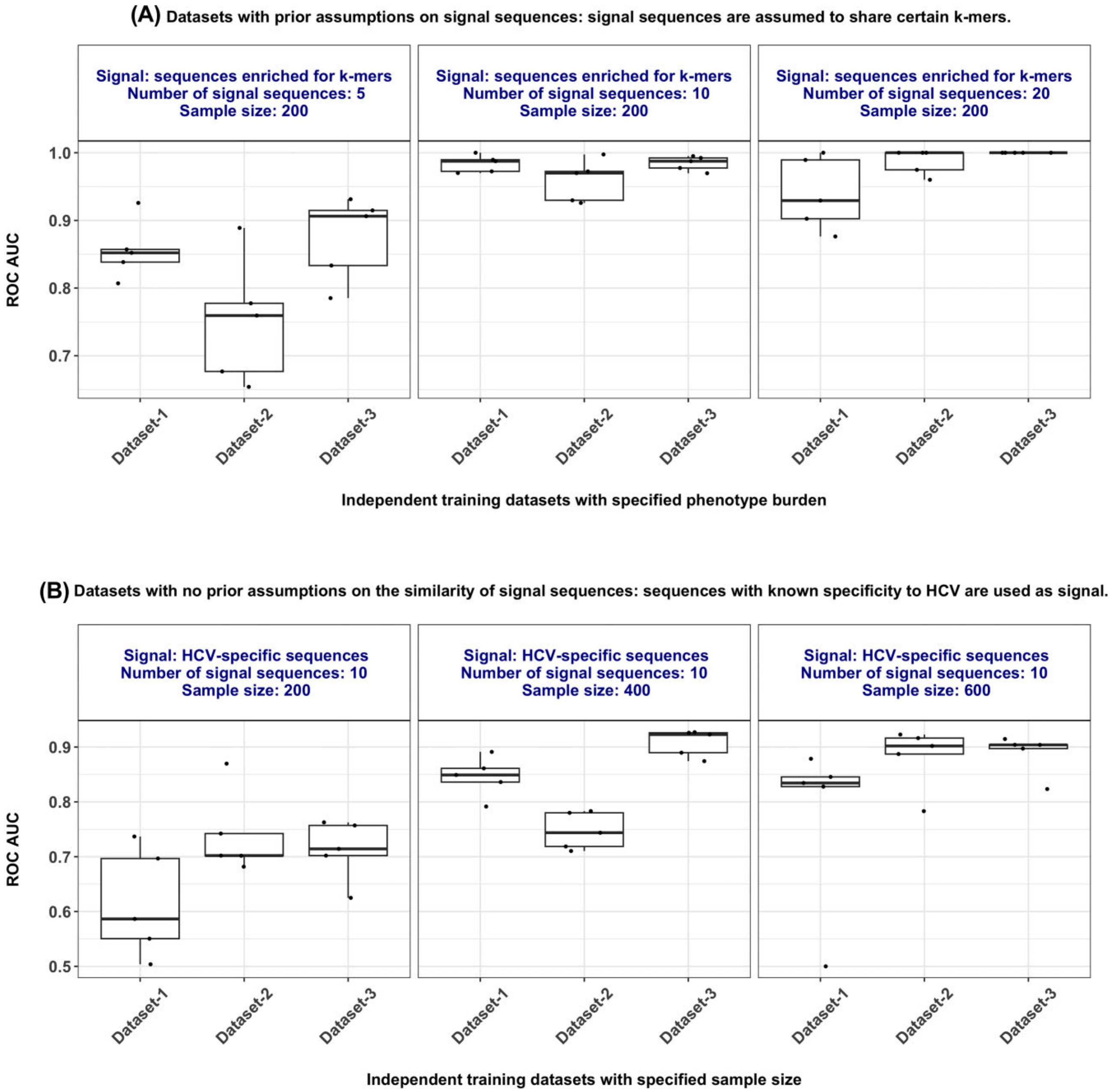


Figure 4: Performance of ML methods on benchmark datasets generated by simAIRR. (A) Performance estimates (ROC AUC) of a highly penalized logistic regression model (on the y-axis) that is optimized well for hyperparameters in a binary classification of balanced, labeled AIRR datasets (number of examples = 200) encoded as 4-mers, where the signal in positive class examples is composed of full sequences that share any of the three 4-mers: *WKDY*, *YREV*, and *ERFY*. The rationale behind the choice of the selected *k*-mers was to compare the performance of ML methods on benchmark datasets with similar characteristics (witness rates and sample sizes) in our previous study that used the same *k*-mers. The full sequences sharing any of the 3 chosen 4-mers were obtained by generating a large number of sequences and retrieving only those sequences that carry the 4-mers through pattern matching. The subpanels represent witness rates, where we vary whether the positive-labeled repertoires carry 5, 10, or 20 immune state–associated sequences per approximately 10^5 sequences. In each subpanel, we simulated 3 independent datasets (named Dataset-1, Dataset-2, and Dataset-3) to gauge the variation in performance on similar dataset characteristics. The spread of the performance metrics on the y-axis shows the variation of performance obtained through nested cross-validation. Overall, the findings indicate that when the witness rate is ≥10 sequences per 10^5 receptor sequences, the performance of a suited ML method was close to perfect, while it dropped when the witness rate was smaller than 10 sequences per 10^5 receptor sequences. A similar trend of performance drop was observed in our previous study at such low witness rate (see Fig. 3a in [59]). (B) Performance estimates (ROC AUC) of a probabilistic binary classifier [19] implemented in immuneML [61] (on the y-axis) in a binary classification of balanced, labeled AIRR datasets (constant witness rate of 10 immune state–associated sequences per approximately 10^5 sequences). The signal in positive class examples is composed of full sequences that are reported to have specificity for HCV. The HCV-specific sequences were downloaded from VDJdb [54]. The subpanels represent datasets of different sample sizes of balanced AIRR datasets. In each subpanel, we simulated 3 independent datasets (named Dataset-1, Dataset-2, and Dataset-3) to gauge the variation in performance on similar dataset characteristics. The spread of the performance metrics on the y-axis shows the variation of performance obtained through nested cross-validation. Overall, the findings indicate that the performance dropped steadily with a decrease in sample size, and a similar trend was shown earlier by Pavlović et al. on datasets with similar study design characteristics (see Fig. 2b in [61]).

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

sizes [61, 62]. When the sample size reached a similar level as in Emerson et al. [19] ($n = 600$), the ROC AUC reached a comparable level as reported in the original study [19] and other studies that reanalyzed the same dataset [61, 62] (Fig. 4B).

Methods

Outlier measure and likelihood ratio

We computed 2 different quantitative measures for each public sequence that could potentially aid in the identification of immune state–associated sequences independently. First, an outlier score is computed using a 2-step process: (i) given the generation probability ($\mathbf{p_{gen}}$) and the average unique number of sequences in a repertoire (number of trials), we computed the probability of observing a public sequence at least once in a repertoire ($\mathbf{p_{obs}}$) using the cumulative density function of the binomial distribution. (ii) Given the probability of observing a public sequence at least once in a repertoire ($\mathbf{p_{obs}}$) and the number of positive labeled repertoires, we further computed the probability of observing a public sequence in the same or higher number of repertoires as it was observed for a given dataset ($\mathbf{p_{count}}$). A negative $\log_{10}$ of $\mathrm{p}_{\mathrm{count}}$ is referred to as an outlier score throughout the article. Second, a likelihood ratio is computed as the ratio of the empirical probabilities of incidence in positive class examples (repertoires) to negative class examples. Given $\boldsymbol{P}$ positive class examples and $\boldsymbol{N}$ negative class examples, $\mathbf{c_P}$ number of occurrences in positive class examples, and $\mathbf{c_N}$ number of occurrences in negative class examples, the likelihood ratio (LR) is defined as $LR = \frac{\frac{c_P}{P}}{\frac{c_N}{N}}$.

Models of realistic population incidence of AIR sequences

To learn the empirical relation between generation probability and population incidence of sequences from a real-world experimental dataset [19], we first computed the relative frequencies and generation probabilities of each unique sequence, including V and J gene masks, in a public experimental dataset [19]. For counting the population sequence frequencies, we used CompAIRR [35], and for computing generation probabilities, we used OLGA [44], as described further below. The relative frequencies refer to the proportion of the population carrying a particular sequence. We derived the relation between generation probability and population incidence of sequences separately for the perceived true signal sequences reported by the original study [19] and for all the remaining public sequences. For this, we discretized the entire range of generation probabilities and population incidences into discrete intervals, and we placed each unique sequence into its corresponding bin of generation probability and population incidence. With that, we obtained empirical probability distributions describing what fraction of the total unique sequences with a certain generation probability distribution occurs at certain population incidence levels in a sample.

Correction of population incidence of public sequences and construction of antigen-experienced-like repertoires

Individuals accumulate immune events over a lifetime. Thus, snapshots of AIRRs, as acquired through targeted immune receptor sequencing from donor blood samples, capture antigen-experienced repertoires that share not only public sequences that are easier to generate but also other pools of common immune event–associated sequences. However, *in silico*–generated synthetic repertoires from a method like OLGA behave like naive repertoires that did not experience any immune events. Synthetic naive repertoires share fewer unique sequences and thus tend to carry a lower proportion of public sequences compared to repertoires from experimental datasets. To correct the proportion of public sequences and their population incidence levels, we used the following procedure: we first generate a large number of AIR sequences using the V(D)J recombination model chosen by the user and retain only unique sequences. Notably, the user can choose from any one of the V(D)J recombination models supplied by default with OLGA [44]. We then make a user-desired proportion of sequences public (10% of the sequences is the default option to match experimental datasets). The population incidence levels (how frequent each unique sequence will be) for the public sequences follows the learned models of realistic population incidence of AIR sequences based on a previously published large cohort study of TCR repertoires [19]. As the publicness of sequences can vary between different species, chains, cell type and their subsets, and pairedness of sequences, the users need to calibrate the relation between generation probability and population incidence of sequences when intending to simulate datasets other than human TCRβ chain sequences, which is the default models supplied with simAIRR (RRID: SCR_023956). The user can supply custom models for the dependence between generation probability and population incidence levels calibrated based on the experimental datasets of their choice.

Assessing the feasibility of a user-desired witness rate

How simAIRR assesses the feasibility of a user-desired witness rate is best explained with an example. For the sake of an illustrative example, we use small numbers for simulation parameters. If the user chooses a pool of 3 sequences as the signal that separates immune state labels and wants them to be introduced into the positive-labeled repertoires ($n = 100$) of a repertoire dataset ($n = 200$) such that each positively labeled repertoire carries a total of 5 signal sequences on average (desired witness rate), one should be able to introduce a total of 500 instances of the signal (100 × 5) from the pool of 3 sequences. Based on the empirical knowledge of dependence between generation probability and population incidence, if each signal sequence cannot be seen in more than 30% of the total sample size, the pool of 3 sequences together cannot be observed more than 180 times (3 × 200 × 0.3) even if the sequences have a high generation probability. In such a case, it is considered infeasible to meet the user-desired witness rate (of 5 signal sequences per repertoire, amounting to a desired total of 500). simAIRR provides detailed statistics in such a case to help the user in reconfiguring the simulation parameters. This could mean that either the user supplies a larger pool of potential signal sequences or modifies the desired witness rate. To avoid expensive computations, users could first use the feasibility assessment mode to make sure that the simulations are feasible given the user-supplied simulation parameters.

Of particular note, the pool of potential signal sequences that one starts with plays a significant role in determining the feasibility of achieving the desired witness rate of signal in the simulated datasets using simAIRR's approach because of the reliance of possible incidence level of sequences on the generation probability of the individual sequences. Thus, unlike in naive simulation approaches, the job of carefully selecting a pool of signal sequences is delegated to the user.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

Construction of synthetic baseline repertoires

When constructing synthetic baseline repertoires as a reasonable proxy for real-world experimental repertoires, we ensured the nativeness of the simulated AIR sequences in terms of positional biases, amino acid usage, and sequence length distributions. For this, we generated AIR sequences according to recombination models provided by OLGA [44]. Note that for the analyses of this article, we generated human T-cell beta chain receptor datasets, while for simAIRR simulations in general, any of the 4 default V(D)J models supplied by OLGA [44] (humanTRB, humanTRA, humanIGH, mouseIGH) can be used. Future versions of simAIRR will also allow the usage of user-supplied AIR sequences to construct the synthetic baseline repertoires.

Computation of generation probability

The generation probabilities of AIR sequences are computed using OLGA [44] with the default generative models of the V(D)J recombination model while including both the masks for V and J genes for each sequence. For the analyses of this article, we computed the generation probabilities using the default V(D)J recombination model of the human T-cell beta chain receptor. Notably, any of the 4 default V(D)J recombination models supplied with OLGA [44] can be used as mentioned above.

ML models, training, selection, and evaluation

We used 2 different ML methods in the case studies. In the first case study, the signal sequences share any of the 3 chosen 4-mers. For those datasets, where the signal can be captured by 4-mers, we used an ML method that matches with the described at length in our previous study [59]. Briefly, we used a highly regularized logistic regression model on a 4-mer encoded representation of the amino acid sequences. The hyperparameters for the model were chosen through nested cross-validation. For details on the implementations and hyperparameter optimizations, see relevant descriptions in [59]. In the second case study, we did not have prior knowledge of the sequence similarity patterns of the signal sequences. Therefore, we used an ML method that assumes the full sequence identity as a potential signal representation [19]. Specifically, we used a probabilistic binary classifier based on phenotype burden [19] implemented in immuneML [61]. We used 5-fold nested cross-validation and an exhaustive grid search for hyperparameter optimization as in our previous study [59] for both ML methods. Balanced accuracy was used as the performance metric for optimization during training, and the ROC AUC was reported for the sake of comparison with previous studies that used the same metric.

Querying sequences enriched for *k*-mer-like patterns

In the first case study, we assumed that the true signal sequences share a similarity in terms of shared *k*-mers. Specifically, the signal sequences were required to carry any 1 of the 3 chosen *k*-mers: *WKDY*, *YREV*, and *ERFY*. In our previous study [59], we implanted *k*-mers in the central portion of the CDR3 amino acid sequences to obtain such signal sequences. However, the implantation of *k*-mers can introduce additional artifacts by destroying the biological properties of the sequence, which the ML methods can exploit as another way of shortcut learning. To avoid that, in this study, we queried a large set of reference sequences to retrieve all those sequences that carry a *k*-mer of interest. To make such a pattern-matching process easy for the users if need be, we provide a simple Python script based on the bionumpy library [63] and a corresponding tutorial that shows how to generate a large set of reference sequences and retrieve the pattern-matched sequences of interest with minimal effort and with very few lines of Python code [60]. We also provided corresponding examples using Unix grep. In the recipe, we have shown how to construct a complex set of signal sequences enriched for multiple criteria like the presence of multiple subsequence patterns within the sequences and gene usage. The code recipes not only require a minimal coding effort but are also efficient and have a minimal runtime. For instance, the wall time for generating 10 million sequences when using 40 processes was less than 1 minute. Similarly, the execution wall time for querying the 10 million sequences for selected *k*-mer patterns was less than 1 minute. Note that the query time can increase with the number of queried patterns for any pattern-matching tool, including Unix grep. A key benefit of the approach described here is that once a large number of reference sequences are generated and stored on disk, they can be queried multiple times.

Customizable simulation parameters

All simulation parameters of simAIRR are customizable. Some of the key customizable parameters include 1 of the 4 possible V(D)J recombination models, the number of repertoires and the proportion of positive class-labeled repertoires, average sequencing depth, the proportion of public sequences, and witness rate. In addition, the users are required to supply a pool of sequences that will be considered true signal sequences. It is also possible to control the number of sequences that will be used as a true signal.

Docker container to improve reproducibility

To ease the installation issues, allow quick testing, and improve portability, we supply a Docker image [64] with a predefined computing environment maintaining all the dependencies required for the execution of simAIRR workflows with minimal overhead. The Docker image is hosted on DockerHub and can be accessed at [64].

Discussion

One major challenge in using simulated datasets for benchmarking ML methods is to prevent shortcut learning opportunities [65]. Shortcut learning may lead to biased benchmarking of AIRR-ML methods, reduced generalizability of the methods, and reproducibility crisis [66]. Our study shows that an intuitive and state-of-the-art approach to generating simulated AIRR benchmark datasets can introduce signal artifacts that can be exploited for undesired shortcut learning by AIRR-ML methods. We refer to the introduced signal artifact as the generation probability discordance bias, where a large disparity exists in the generation probability distributions between the introduced true signals and the remaining public sequences. This can allow ML methods to exploit this bias and learn shortcuts instead of the true sequence patterns associated with immune states in the data. An analogy of such shortcut learning from the image recognition domain is an ML model trained to detect different types of animals in pictures (like cats and dogs) identifying the animals based on the background color of the images rather than learning the patterns associated with animal objects.

To mitigate this problem, we developed simAIRR, a novel simulation strategy that constructs antigen-experienced-like baseline repertoires in the sense of the publicity-generation probability relation. It introduces signals in the repertoires by following the models of realistic population incidence of AIR sequences

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

calibrated from real-world experimental datasets. This approach ensures that the simulated datasets are not biased and are instead representative of real-world scenarios. Our findings suggest that this novel simulation approach effectively mitigates the shortcut opportunities that can arise through naive simulation approaches.

One key benefit of simAIRR's approach is the possibility of not making any prior assumptions regarding the similarity or commonality of immune state–associated sequences that will be used as true signals. By utilizing known antigen-specific sequences from public databases (e.g., VDJdb [54, 67, 68]) or other experimentally determined antigen-specific sequences as the pool of signal sequences, users can refrain from making prior assumptions. This will be useful in representing real-world scenarios well because there is currently very limited knowledge of sequence similarity patterns between sequences selected for a common antigen. Existing AIRR-ML methods make varying assumptions regarding the similarity patterns of sequences selected for a common antigen. While some methods assume that signal sequences share subsequence patterns (*k*-mers) [21, 23, 29, 59, 69–71], other studies assume that immune receptor binding is determined by the full CDR3 sequence and is best approached by learning appropriate similarity metrics for full CDR3s [19, 38, 72]. Using case studies, we demonstrated the utility of simAIRR in generating benchmark datasets for immune state prediction problems following varying assumptions regarding true signals: that is, it could be assumed that the true signal sequences share subsequence patterns (*k*-mers, gapped *k*-mers, hamming distance, etc.) or known antigen-associated sequence pool could be used with no *a priori* assumptions.

The cases presented in this study demonstrate that simAIRR can generate AIRR benchmarking datasets for immune state prediction problems that represent real-world scenarios well. The behavior of ML methods on the generated benchmarking datasets matches the behavior of the same methods on real-world experimental datasets at different sample sizes and witness rates. This suggests that the generated benchmarking datasets sufficiently match the real-world experimental datasets in terms of signal and noise. Often, benchmarking of computational methods published as part of articles presenting novel methods identifies the proposed novel method(s) as a winner(s), largely owing to benchmarking not being fully neutral [73, 74]. However, neutral benchmarking is especially valuable for the scientific community to make the evaluation of the methods more rational and to establish standards on a scientific basis [73, 74]. We encourage AIRR-ML researchers to evaluate the performance of novel ML methods on neutral benchmarking datasets developed by other researchers to improve the generalizability of methods. For instance, the benchmarking datasets used in the case studies and similar simulation parameters at smaller sample sizes and lower witness rates represent suited benchmarking datasets for unbiased performance evaluation of novel AIRR-ML methods. These scenarios also represent the cases where novel AIRR-ML method development is needed because the performance of state-of-the-art ML methods drops compared to easier scenarios. Notably, the empirical performance maps of baseline ML models across a wide range of study design and ML challenges as profiled in our previous study [59] need to be reevaluated with more realistic simulated datasets (such as the ones generated in this study) given the strong prior assumptions on what constitutes a signal.

Machine learning models' performance on AIRR data generated by different simulation approaches (including simAIRR) can be comparable if the simulations avoid creating shortcut learning opportunities through broken realism. For a machine learning problem, the goal of any AIRR simulator, such as simAIRR, is to introduce specific signals (e.g., sequence patterns) into certain repertoires, but there is no consensus, to our knowledge, on what constitutes immune signals. Thus, simulation tools create counterfactual worlds through various assumptions of immune signals to enable the development of assumption-agnostic ML methods. ML methods that focus on the desired signal and ignore simulation artifacts should perform similarly on datasets generated by simAIRR and previous simulation approaches. However, if a simulation approach introduces artifacts by deviating from realistic biological properties, it can provide shortcut learning opportunities that certain ML methods may exploit. In conclusion, our viewpoint is that the performance of well-behaving ML methods does not necessarily have to be different between simAIRR-generated data and data generated by other simulation approaches, but previous approaches should be open for shortcut learning opportunities, where it can be hard to know whether or not they have been exploited by a complex ML model. Moreover, it is worth noting that simAIRR can serve as a corrective plug-in for other simulation approaches (e.g., [46]), effectively mitigating shortcut learning opportunities arising as a consequence of the perturbed degree of publicness of receptor sequences.

An important aspect to keep in mind is that the default models supplied with simAIRR with respect to the relation between generation probability and the population incidence of public sequences are calibrated based on human TCRβ chain sequences. However, the magnitude of public responses can vary for other AIRR loci depending on the species, cell types, cell subsets, chains, and pairedness of sequences [13]. Thus, when intending to simulate AIRR datasets of other AIRR loci, different species, cell types, and chains, users need to calibrate the relation between generation probability and population incidence of sequences on corresponding experimental datasets and supply such custom models to simAIRR. To make this process easier, we provided console scripts through another add-on Python package [75]. The file formats for the custom models are described in simAIRR's documentation. Note that the population incidence of sequences can be computed with CompAIRR [35], and the generation probability of sequences can be computed using OLGA [44] with minimal effort.

To learn models reflecting the realistic population incidence (of signal sequences and remaining public sequences), we used the largest TCRβ dataset currently publicly available [19] with the following desired properties: a case-control design in which there were no known protocol differences between the cases and controls and an optimal sample size to maximize the exactness of the estimates of population incidence of AIR sequences. We were not able to extend the testing to B-cell heavy chains because of the lack of availability of public IGH repertoire datasets of the aforementioned criteria in public databases like immuneACCESS [76] and iReceptor [77]. A limited number of datasets with a case-control design, protocol differences between cases and controls when available, small sample sizes, or larger portions of missing data were some of the problems that were encountered when trying to extend the testing to IGH repertoires. Note that the approximation of the true estimates of realistic population incidence of AIR sequences improves with increased sample size, and models learned on datasets of smaller sample size (e.g., 50 repertoires) can spuriously inflate population incidence of public sequences (Supplementary Fig. S6). IGH repertoires differ from TCRβ repertoires in terms of diversity introduced through somatic hypermutations (SHMs). When considering the consequence of SHM-induced diversity for an ML problem, as compared to the pres-

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

ence of identical public sequences in TCRβs within certain phenotypes, IGH repertoires are known to be characterized by clonal lineages and thus a "public clonotype" definition for BCRs has to rely on some form of similarity measure between sequences. For instance, in recent studies that observed convergent B-cell clonotypes as a response to SARS-CoV-2 [78, 79], 85% amino acid sequence similarity of CDR-H3 was used to identify convergent clonotypes. Preselection sequence generation tools (like OLGA [44] and its parent tool, IGoR [45]) can generate an arbitrary number of preselection sequences based on learned statistics of V(D)J recombination and SHM, and simAIRR can generate IGH repertoires that are akin to naive BCR without the additional diversity arising because of SHM. To introduce SHM-induced diversity into the repertoires, an extension to simAIRR is needed in the future to have the possibility of public clonotypes not being identical but similar by some distance as stated above.

A plethora of studies have compared repertoires of a common immune state and found identical or similar sequences detected in many individuals [6–11]. AIRR-ML methods rely on this observation that condition-associated sequences or similar sequences will be observed in a considerable fraction of the sample size that shares the immune state. This overlap of similar signal sequences across repertoires is a post-V(D)J recombination phenomenon. Although repertoire generation models of B-cell repertoires have been suggested to be individual-specific [80] (moderately mediated by a high degree of polymorphisms in *IGHV* genes), condition-associated similar sequences can be shared among individuals of a common immune state [6–11]. simAIRR simulates this exact phenomenon of shared similar sequences across repertoires while not allowing the sharing pattern to exceed what can be possible given a rough indication of the generation probability of the sequences. Previous research has suggested a high degree of concordance in repertoire generation models across individuals [45, 57, 81, 82], but recent observations provided evidence that the generation probability of sequences, in some cases, can vary up to or more than 3 orders of magnitude due to individualized recombination models [80]. It remains to be investigated how the individualized recombination models affect the overlap of public sequences at a population level and subsequently the pattern recognition capacity of ML models. If deemed necessary, future improvements of simAIRR should use individualized recombination models.

An important consideration for AIRR data simulation for ML-based prediction problems in the context of individualized repertoire generation models is what is the exact consequence of individualized repertoire general models at a population level. Although there exists limited knowledge on this aspect, we are allowing ourselves to speculate on the possible consequences. As individualized repertoire generation models and the generation probability of sequences are tightly related, some sequences that can be thought to be generated with high probability, in general, might not be occurring in individual repertoires with the same probability, leading to less overlap of sequences between repertoires at a population level. In simAIRR, users can customize the degree of sharing of public sequences across repertoires, meaning that the proportion of unique sequences in a repertoire dataset that is public can be controlled. A suitable default for this customizable parameter can be estimated from experimental datasets as a post-VDJ recombination model characteristic.

To minimize the effort needed to construct a complex set of signal sequences, if needed, we provided simple but efficient Python recipes in the documentation of simAIRR. The provided tutorials show how to generate a large set of reference sequences and query them to retrieve a complex set of signal sequences enriched for multiple criteria like the presence of multiple subsequence patterns within the sequences (e.g., *k*-mers) and gene usage. A major advantage of the described approach is that once a large number of reference sequences are generated and stored on disk, they can be queried many times. However, in some cases (e.g., when evaluating a particular behavior of a developed ML method), one might need to construct signal sequences enriched for very rare sequence patterns. To obtain a sufficiently large number of sequences sharing such rare sequence patterns, one may have to generate and query many millions of sequences. In this article, we did not measure the execution time for mining rare sequence patterns from a set of reference sequences. If the execution time is slow to mine rare sequence patterns, an alternative solution may be to implant sequence patterns (e.g., *k*-mers) while avoiding the positional biases using existing tools [41] or as we did in our previous study [59]. However, note that the implantation of sequence patterns may lead to another type of shortcut learning opportunities, and thus it is important to thoroughly assess the potential pitfalls before using implantation, particularly in benchmarking competitions.

simAIRR currently does not support the simulation of paired chain repertoires and clonal frequencies of receptors in repertoires. Notably, the clonal frequencies of AIRs have been suggested to follow a power law distribution [83]. Future improvements of simAIRR should also simulate clonal frequencies. However, further empirical evidence from independent studies on the usefulness of power law in describing clonal frequency distributions in multiple cell types, subsets, and species is needed in this connection. Also note that simAIRR focuses only on simulating datasets for immune state prediction problems at the repertoire level but does not focus on receptor–specificity prediction problems. Another notable unmodeled aspect of simAIRR is the possible dependence between AIR sequences in real-world repertoires. Particularly in the context of this study, there may be dependence between outlier sequences in real-world repertoires because of a possible association with the same immune stimulus. There may also be dependence between outlier and signal sequences where, for example, a particular antigen experienced in the past primes for a specific response to the condition being investigated. The lack of such dependence between public AIR sequences within simAIRR-generated repertoires does not impact simAIRR's objective of mitigating shortcut learning opportunities for ML methods when simulating repertoires of contrasting immune states. However, understanding and subsequent modeling of the dependence between AIR sequences within repertoires of real-world experimental AIRR datasets is warranted in future studies.

Conclusion

In summary, the contribution of this study is bringing to light not only the potential shortcut learning opportunity that can arise with the state-of-the-art way of simulating AIRR datasets but also a novel simulation approach implemented as a Python package that can help avoid potential shortcut learning opportunities for ML methods. Unlike state-of-the-art naive simulation approaches, a key benefit of simAIRR's approach is the possibility of not making any prior assumptions regarding the similarity or commonality of immune state–associated sequences that will be used as true signals. The AIRR datasets simulated using simAIRR were similar to real-world experimental datasets based on the performance of ML methods on both types of datasets. We

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

suggest testing new ML methods on neutral benchmark datasets like simAIRR's to aid unbiased evaluation.

Graphics

ggplot2 version 3.3.6 [84] was used for graphs, and Inkscape version 1.0.1 [85] was used for illustrations.

Availability of Source Code and Requirements

- Project name: simAIRR
- Project homepage: https://github.com/KanduriC/simAIRR [86]
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Python 3.8 or higher
- License: GNU AGPL version 3
- Research Resource Identification Initiative ID (RRID): SCR_023956
- biotoolsID: biotools:simairr

Data Availability

Snapshots of the frozen codebase with a permanent DOI are available on *Zenodo* database [87].

A Docker image of simAIRR is available on docker hub at [64], and configuration files to reproduce the simulations and ML models of the use cases are available on a separate repository on GitHub at [75]. Simulated datasets used in the case studies with permanent DOI are available on the NIRD database at [88]. A simple Python recipe and tutorial for generating sequences enriched for *k*-mer-like sequence patterns to be used as true signal are available on the simAIRR documentation at [60]. An archival copy of the code and supporting data is available via the *GigaScience* repository, GigaDB [89].

Additional Files

Supplementary Fig. S1. Generation probability discordance bias in the naive simulation of AIRR datasets at different sample sizes. We randomly subsampled repertoires of different sample sizes (n = 50, 100, 200) with balanced labels from a real-world experimental AIRR dataset ($n = 683$ repertoires; 307 positive class examples and 376 negative class examples, average unique number of TCRβs per repertoire $\sim$200,000) [1] and repeated the experiments shown in Fig. 2 of the main text. For comparison, we also included similar-sized data from naive simulations. The findings at different sample sizes are shown in different rows of the chart. Similar to the observations of Fig. 2, the outlier measure (y-axis) alone is a poor classifier of presumed true signals even in the subsampled real-world experimental data at different sample sizes (charts on the left column). On the contrary, the outlier measure alone has very high precision in synthetic datasets generated through the naive simulation approach described in Fig. 1B.ii of the main text.

Supplementary Fig. S2. Absence of generation probability discordance bias in a real-world experimental dataset of a small sample size. The phenomenon of the absence of generation probability discordance bias in real-world experimental data was also observed on another TCRβ dataset with a much smaller sample size (compared to [1]) of 79 cases and 13 controls with/without pediatric COVID-19 from [2]. Sacco et al. assayed TCRβ data from pediatric patients with COVID-19 who either progressed to develop multisystem inflammatory syndrome or not (together 79 cases) and matched healthy controls (13 controls). In the original article of Sacco et al., TCRβ data of COVID-19–positive cases were not compared against healthy controls to identify potential public clones. However, here, for the sake of investigating the presence/absence of generation probability discordance bias in a dataset of as large a sample size as possible, we included all 79 cases and 13 controls in the analysis. Since the condition-associated public clones are not reported in the original study, we looked at the distribution of likelihood ratio (on the x-axis) and placed an arbitrary threshold to call sequences as presumed signals. Irrespective of the presumed signal sequences, the main observation from this analysis is that similar to the observations of Fig. 2 and Supplementary Fig. S1, in the real-world experimental datasets, many public sequences are observed with increased population incidence than what is expected based on their generation probability. In other words, thresholding based on the computed outlier score alone will be a poor classifier of signal sequences.

Supplementary Fig. S3. Presence of many receptor sequences at an unlikely high population incidence than what is expected based on their generation probability in the CMV-negative cohort of Emerson et al. [1]. We used the real-world experimental AIRR dataset with CMV-positive and CMV-negative repertoires that were used for the analyses in Fig. 2A ($n = 683$ repertoires; 307 CMV-positive examples and 376 CMV-negative examples, average unique number of TCRβs per repertoire $\sim$200,000). Instead of computing the outlier measure (y-axis) based on CMV-positive examples, we repeated the same experiment on a CMV-negative cohort as a control experiment to demonstrate that antigen-experienced repertoires carry many outlier sequences that occur at an unlikely high population incidence than what is expected based on their generation probability. Similar to the observations of Fig. 2A, even in the CMV-negative cohort, many sequences were observed with high outlier scores (on the y-axis). Here, each point represents a unique TCRβ CDR3 sequence. When computing the likelihood ratio (on the x-axis) as described in the Methods section of the main text, we treated the CMV-negative examples as a positive class and used the CMV-positive cohort as a negative class. The data points in red as shown in Fig. 2 are absent in this figure because we did not consider any sequences as condition-associated or true signals. Nevertheless, the main message from this analysis is the observation of many AIR sequences in antigen-experienced repertoires that occur at an unlikely population incidence level than what is expected given their generation probability as measured through the outlier score on the y-axis.

Supplementary Fig. S4. Empirical relation between generation probability and population incidence of public AIR sequences. The relation between generation probability and population incidence of public AIR sequences was determined based on a previously published large cohort study of TCR repertoires [1] separately for signal (3.b) and all the remaining public sequences (3.a). In both a and b, the x-axis represents the sample size proportion bins and the y-axis represents generation probability distribution bins in the $-\log_{10}$ scale. Note that the bins on both axes are half-open intervals that include the left endpoint but exclude the right endpoint. To explain one cell of the heatmaps, consider the top-left cell in panel a. The cell corresponding to the generation probability bin [8, 2) and sample size proportion bin [0.2, 0.4) tell that 18% of the total unique public sequences with a generation probability between [8, 2) (in $-\log_{10}$ scale) occurs between [0.2, 0.4) % of total repertoires in the sample (dataset). The row-sums of the heatmaps should sum to 1.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

Supplementary Fig. S5. Empirical cumulative distribution of population incidence (sample size percentages on the x-axis) for different generation probability bins (subpanels) for presumed signal sequences (b) and all the remaining public sequences (a) in a previously published large cohort study of TCR repertoires [1].

Supplementary Fig. S6. Empirical distribution of population incidence for receptors within different generation probability intervals at different sample sizes (rows of the chart). The relation between generation probability and population incidence of public AIR sequences was determined based on a previously published large cohort study of TCR repertoires [1] separately for presumed signal sequences (a) and all the remaining public sequences (b) at different sample sizes (rows of the chart)**.** In both a and b, receptors along the x-axis are split into 3 distinct bins according to their generation probability. The stacked bars along the y-axis represent the proportion of receptors within a given generation probability interval that have population incidence falling within a particular range (with a distinct color representing each population incidence range). A full sample size of $n = 683$ repertoires with approximately balanced labels (same chart as Fig. 3) was shown for comparison with other investigated sample sizes of $n =$ [200, 100, 50 repertoires] with balanced labels. Note that the bins of both generation probability and sample size proportions are half-open intervals that include the left endpoint but exclude the right endpoint. If the distribution of sequences at different population incidence levels (colors) within each generation probability interval (on the x-axis) at the full sample size of 683 repertoires (top row in the chart) is assumed as a proxy for the true population estimates, we notice that the true population estimates are well approximated with an increase in sample size (rows 2–4 in the chart).

Supplementary_Information.docx.

Abbreviations

AIR: adaptive immune receptor; AIRR: adaptive immune receptor repertoire; BCR: B-cell receptor; CDR3: complementarity determining region 3; CMV: cytomegalovirus; HCV: hepatitis C virus; IMGT: immunogenetics; LR: likelihood ratio; ML: machine learning; SHM: somatic hypermutation; TCR: T-cell receptor; TCRβ: T-cell receptor beta chain.

Competing Interests

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope, and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Specifica Inc, Roche/Genentech, immunai, and LabGenius. The other authors delcare no competing interests.

Funding

Supported by the Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to V.G.), UiO World-Leading Research Community (to V.G.), UiO: LifeScience Convergence Environment Immunolingo (to V.G. and G.K.S.), EU Horizon 2020 iReceptorplus (#825821) (to V.G.), a Norwegian Cancer Society Grant (#215817, to V.G.), Research Council of Norway projects (#300740, #331890 to V.G.), a Research Council of Norway IKTPLUSS project (#311341, to V.G. and G.K.S.), and Stiftelsen Kristian Gerhard Jebsen (K. G. Jebsen Coeliac Disease Research Centre, to G.K.S.).

Authors' Contributions

C.K. and G.K.S. conceived the overall study. C.K. designed and developed the code, performed analyses, and drafted the manuscript. L.S. and M.P. contributed with ML method implementations in immuneML [61] that were used in the case studies. K.R. contributed with the bionumpy [63] recipe [60] for retrieving sequences enriched for sequence patterns. M.C., O.P., G.Y., and V.G. participated in brainstorming and provided critical conceptual feedback. All authors read, provided critical edits, and approved the final manuscript.

Acknowledgments

Some of the analyses in this work were performed using the Immunohub eInfrastructure funded by the University of Oslo and operated by the authors in close collaboration with the University Senter for Information Technology (USIT), University of Oslo.

References

1. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol 2014;35:581–90. https://doi.org/10.1016/j.it.2014.09.004.
2. Georgiou G, Ippolito GC, Beausang J, et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 2014;32:158–68. https://doi.org/10.1038/nbt.2782.
3. Arnaout RA, Prak ETL, Schwab N, et al. Adaptive immune receptor repertoire community. The future of blood testing is the immunome. Front Immunol 2021;12. https://doi.org/10.3389/fimmu.2021.626793.
4. Miho E, Yermanos A, Weber CR, et al. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Front Immunol 2018;9. https://doi.org/10.3389/fimmu.2018.00224.
5. Greiff V, Yaari G, Cowell LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Curr Opin Syst Biol 2020;24:109–19. https://doi.org/10.1016/j.coisb.2020.10.010.
6. Venturi V, Price DA, Douek DC, et al. The molecular basis for public T-cell responses? Nat Rev Immunol 2008;8:231–8. https://doi.org/10.1038/nri2260.
7. Li H, Ye C, Ji G, et al. Determinants of public T cell responses. Cell Res 2012;22:33–42. https://doi.org/10.1038/cr.2012.1.
8. Madi A, Shifrut E, Reich-Zeliger S, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res 2014;24:1603–12. https://doi.org/10.1101/gr.170753.113.
9. Madi A, Poran A, Shifrut E, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. eLife 2017;6. https://doi.org/10.7554/eLife.22057.
10. Pogorelyy MV, Elhanati Y, Marcou Q, et al. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. PLoS Comput Biol 2017;13:e1005572. https://doi.org/10.1371/journal.pcbi.1005572.
11. Greiff V, Menzel U, Miho E, et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. Cell Rep 2017;19:1467–78. https://doi.org/10.1016/j.celrep.2017.04.054.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

12. Trück J, Ramasamy MN, Galson JD, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. J Immunol Baltim Md 1950 2015;194:252–61. https://doi.org/10.4049/jimmunol.1401405.
13. Jaffe DB, Shahi P, Adams BA, et al. Functional antibodies exhibit light chain coherence. Nature 2022;611:352–7. https://doi.org/10.1038/s41586-022-05371-z.
14. Greiff V, Miho E, Menzel U, et al. Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol 2015;36;738–49. https://doi.org/10.1016/j.it.2015.09.006.
15. Pertseva M, Gao B, Neumeier D, et al. Applications of machine and deep learning in adaptive immunity. Annu Rev Chem Biomol Eng 2021;12:39–62. https://doi.org/10.1146/annurev-chembioeng-101420-125021.
16. Katayama Y, Yokota R, Akiyama T, et al. Machine learning approaches to TCR repertoire analysis. Front Immunol 2022;13. https://doi.org/10.3389/fimmu.2022.858057.
17. Glanville J, Huang H, Nau A, et al. Identifying specificity groups in the T cell receptor repertoire. Nature 2017;547:94–98. https://doi.org/10.1038/nature22976.
18. Dash P, Fiore-Gartland AJ, Hertz T, et al. Quantifiable predictive features define epitope specific T cell receptor repertoires. Nature 2017;547:89–93. https://doi.org/10.1038/nature22383.
19. Emerson RO, DeWitt WS, Vignali M, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet 2017;49:659–65. https://doi.org/10.1038/ng.3822.
20. Pogorelyy MV, Minervina AA, Chudakov DM, et al. Method for identification of condition-associated public antigen receptor sequences. eLife 2018;7. https://doi.org/10.7554/eLife.33050.
21. Ostmeyer J, Christley S, Toby IT, et al. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. Cancer Res 2019;79:1671–80. https://doi.org/10.1158/0008-5472.CAN-18-2292.
22. Zhang H, Liu L, Zhang J, et al. Investigation of antigen-specific T-cell receptor clusters in Human cancers. Clin Cancer Res 2020;26:1359–71. https://doi.org/10.1158/1078-0432.CCR-19-3249.
23. Huang H, Wang C, Rubelt F, et al. Analyzing the mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol 2020;38:1194–202. https://doi.org/10.1038/s41587-020-0505-4.
24. Widrich M, Schäfl B, Pavlović M, et al. Modern Hopfield networks and attention for immune repertoire classification. Biorxiv. 2020. https://www.biorxiv.org/content/10.1101/2020.04.12.038158v3.full.
25. Beshnova D, Ye J, Onabolu O, et al. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. Sci Transl Med 2020;12. https://doi.org/10.1126/scitranslmed.aaz3738.
26. Shemesh O, Polak P, Lundin KEA, et al. Machine learning analysis of naïve B-cell receptor repertoires stratifies celiac disease patients and controls. Front Immunol 2021;12. https://doi.org/10.3389/fimmu.2021.627813.
27. Sidhom J-W, Larman HB, Pardoll DM, et al. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. Nat Commun 2021:12(1):1605. https://doi.org/10.1038/s41467-021-21879-w.
28. Greiff V, Weber CR, Palme J, et al. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. J Immunol 2017;199:2985–97. https://doi.org/10.4049/jimmunol.1700594.
29. Akbar R, Robert PA, Pavlović M, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. Cell Rep 2021;34:108856. https://doi.org/10.1016/j.celrep.2021.108856.
30. Pogorelyy MV, Minervina AA, Shugay M, et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. PLoS Biol 2019;17:e3000314. https://doi.org/10.1371/journal.pbio.3000314.
31. Jokinen E, Huuhtanen J, Mustjoki S, et al. Predicting recognition between T cell receptors and epitopes with TCRGP. PLoS Comput Biol 2021;17:e1008814. https://doi.org/10.1371/journal.pcbi.1008814.
32. Robert PA, Akbar R, Frank R, et al. Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. Nat Comput Sci 2022;2:845–65. https://doi.org/10.1038/s43588-022-00372-4.
33. Greiff V, Bhat P, Cook SC, et al. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med 2015;7(1):49. https://doi.org/10.1186/s13073-015-0169-8.
34. Sidhom J-W, Oliveira G, Ross-MacDonald P, et al. Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. Sci Adv 2022;8. https://doi.org/10.1126/sciadv.abq5089.
35. Rognes T, Scheffer L, Greiff V, et al. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. Bioinformatics 2022;38:4230–2. https://doi.org/10.1093/bioinformatics/btac505.
36. Sandve GK, Greiff V. Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. Bioinformatics 2022;38:4994–6. https://doi.org/10.1093/bioinformatics/btac612.
37. Chen V, Yang M, Cui W, et al. Best practices for interpretable machine learning in computational biology. Biorxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.10.28.513978v1.
38. Liu X, Zhang W, Zhao M, et al. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. Ann Rheum Dis 2019;78:1070–8. https://doi.org/10.1136/annrheumdis-2019-215442.
39. Nolan S, Vignali M, Klinger M, et al. 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 2020. https://doi.org/10.21203/rs.3.rs-51964/v1.
40. Zaslavsky ME, Ram-Mohan N, Guthridge JM, et al. Disease diagnostics using machine learning of immune receptors. Biorxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.04.26.489314v4.
41. Weber CR, Akbar R, Yermanos A, et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. Bioinformatics 2020;36:3594–6. https://doi.org/10.1093/bioinformatics/btaa158.
42. Sutherland C, Cowan GJM. AIRRSHIP: simulating human B cell receptor repertoire sequences. Bioinformatics 2023;39(6):btad365. https://doi.org/10.1093/bioinformatics/btad365.
43. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. Bioinformatics 2015;31:3213–5. https://doi.org/10.1093/bioinformatics/btv326.
44. Sethna Z, Elhanati Y, Callan CG, et al. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

sequences and motifs. Bioinformatics 2019;35:2974–81. https://doi.org/10.1093/bioinformatics/btz035.
45. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nat Commun 2018;9(1):561. https://doi.org/10.1038/s41467-018-02832-w.
46. uio-bmi/ligo: LIgO is a tool for simulation of adaptive immune receptors and repertoires. https://github.com/uio-bmi/ligo. Accessed 2 July 2023.
47. Geirhos R, Jacobsen J-H, Michaelis C, et al. Shortcut learning in deep neural networks. Nat Mach Intell 2020;2:665–73. https://doi.org/10.1038/s42256-020-00257-z.
48. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature 1988;334:395–402. https://doi.org/10.1038/334395a0.
49. Tonegawa S. Somatic generation of antibody diversity. Nature 1983;302:575–81. https://doi.org/10.1038/302575a0.
50. Fazilleau N, Cabaniols J-P, Lemaître F, et al. Valpha and Vbeta public repertoires are highly conserved in terminal deoxynucleotidyl transferase-deficient mice. J Immunol 2005;174:345–55. https://doi.org/10.4049/jimmunol.174.1.345.
51. Venturi V, Kedzierska K, Price DA, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc Natl Acad Sci USA 2006;103:18691–6. https://doi.org/10.1073/pnas.0608907103.
52. Quigley MF, Greenaway HY, Venturi V, et al. Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. Proc Natl Acad Sci USA 2010;107:19414–9. https://doi.org/10.1073/pnas.1010586107.
53. Venturi V, Quigley MF, Greenaway HY, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J Immunol 2011;186:4285–94. https://doi.org/10.4049/jimmunol.1003898.
54. Shugay M, Bagaev DV, Zvyagin IV, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res 2018;46:D419–27. https://doi.org/10.1093/nar/gkx760.
55. Elhanati Y, Sethna Z, Callan CG, et al. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev 2018;284:167–79. https://doi.org/10.1111/imr.12665.
56. Shugay M, Bolotin DA, Putintseva EV, et al. Huge overlap of individual TCR beta repertoires. Front Immunol 2013;4. https://doi.org/10.3389/fimmu.2013.00466.
57. Murugan A, Mora T, Walczak AM, et al. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci USA 2012;109:16161–6. https://doi.org/10.1073/pnas.1212755109.
58. Sacco K, Castagnoli R, Vakkilainen S, et al. Immunopathological signatures in multisystem inflammatory syndrome in children and pediatric COVID-19. Nat Med 2022;28:1050–62. https://doi.org/10.1038/s41591-022-01724-3.
59. Kanduri C, Pavlović M, Scheffer L, et al. Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. Gigascience 2022;11:giac046. https://doi.org/10.1093/gigascience/giac046.
60. Kanduri C, Scheffer L, Pavlović M, et al. A simple Python recipe and tutorial for generating sequences enriched for k-mer-like sequence patterns to be used as true signal for “simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods.” SimAIRR Doc. 2023. https://kanduric.github.io/simAIRR/tutorials.html#querying-sequences-enriched-for-k-mer-like-patterns
61. Pavlović M, Scheffer L, Motwani K, et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. Nat Mach Intell 2021;3:936–44. https://doi.org/10.1038/s42256-021-00413-z.
62. Katayama Y, Kobayashi TJ. Comparative study of repertoire classification methods reveals data efficiency of k -mer feature extraction. Front Immunol 2022;13. https://doi.org/10.3389/fimmu.2022.797640.
63. Rand K, Grytten I, Pavlovic M, et al. BioNumPy: fast and easy analysis of biological data with Python. Biorxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.12.21.521373v1.
64. Kanduri C, Scheffer L, Pavlović M, et al. Docker image of simAIRR at docker hub. Docker Hub. 2023. doi: kanduric/simairr:latest.
65. Kaufman S, Rosset S, Perlich C, et al. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6:1–21. https://doi.org/10.1145/2382577.2382579.
66. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. arXiv. 2022. https://arxiv.org/abs/2207.07048.
67. Tickotsky N, Sagiv T, Prilusky J, et al. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 2017;33:2924–9. https://doi.org/10.1093/bioinformatics/btx286.
68. Fleri W, Paul S, Dhanda SK, et al. The Immune Epitope Database and Analysis resource in Epitope discovery and synthetic vaccine design. Front Immunol 2017;8. https://doi.org/10.3389/fimmu.2017.00278.
69. Cinelli M, Sun Y, Best K, et al. Feature selection using a one dimensional naïve Bayes’ classifier increases the accuracy of support vector machine classification of CDR3 repertoires. Bioinformatics 2017;33:951–5. https://doi.org/10.1093/bioinformatics/btw771.
70. Sun Y, Best K, Cinelli M, et al. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. Front Immunol 2017;8. https://doi.org/10.3389/fimmu.2017.00430.
71. Thomas N, Best K, Cinelli M, et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics 2014;30:3181–8. https://doi.org/10.1093/bioinformatics/btu523.
72. De Neuter N, Bartholomeus E, Elias G, et al. Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. Genes Immun 2019;20:255–60. https://doi.org/10.1038/s41435-018-0035-y.
73. Boulesteix A-L, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. PLoS One 2013;8:e61562. https://doi.org/10.1371/journal.pone.0061562.
74. Weber LM, Saelens W, Cannoodt R, et al. Essential guidelines for computational method benchmarking. Genome Biol 2019;20(1):125. https://doi.org/10.1186/s13059-019-1738-8.
75. Kanduri C, Scheffer L, Pavlović M, et al. Configuration files to reproduce the simulations and ML models of the use cases for “simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods.” GitHub. 2023. https://github.com/KanduriC/usecases_simairr.git.
76. immuneACCESS Data. https://clients.adaptivebiotech.com/immuneaccess. Accessed 1 July 2023.
77. Corrie BD, Marthandan N, Zimonja B, et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev 2018;284:24–41. https://doi.org/10.1111/imr.12666.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024

78. Nielsen SCA, Yang F, Jackson KJL, et al. Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Cell Host Microbe 2020;28:516–25. e5. https://doi.org/10.1016/j.chom.2020.09.002.
79. Galson JD, Schaetzle S, Bashford-Rogers RJM, et al. Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. Front Immunol 2020;11. https://doi.org/10.3389/fimmu.2020.605170.
80. Slabodkin A, Chernigovskaya M, Mikocziova I, et al. Individualized VDJ recombination predisposes the available ig sequence space. Genome Res 2021;31:2209–24. https://doi.org/10.1101/gr.275373.121.
81. Elhanati Y, Sethna Z, Marcou Q, et al. Inferring processes underlying B-cell repertoire diversity. Philos Trans R Soc Lond B Biol Sci 2015;370:20140243. https://doi.org/10.1098/rstb.2014.0243.
82. Sethna Z, Isacchini G, Dupic T, et al. Population variability in the generation and selection of T-cell repertoires. PLoS Comput Biol 2020;16:e1008394. https://doi.org/10.1371/journal.pcbi.1008394.
83. Desponds J, Mora T, Walczak AM. Fluctuating fitness shapes the clone-size distribution of immune repertoires. Proc Natl Acad Sci USA 2016;113:274–9. https://doi.org/10.1073/pnas.1512977112.
84. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2023.
85. Inkscape Project. Inkscape. https://inkscape.org. Accessed 27 September 2023.
86. Kanduri C, Scheffer L, Pavlović M, et al. Source code of simAIRR. GitHub. 2023. https://github.com/KanduriC/simAIRR.
87. Kanduri C, Scheffer L, Pavlović M, et al. Snapshot of the frozen codebase of simAIRR at Zenodo database. Zenodo. 2023. https://doi.org/10.5281/zenodo.7534810.
88. Kanduri C, Scheffer L, Pavlović M, et al. Simulated datasets used in the case studies for "simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods." NIRD Database. 2023. https://doi.org/10.11582/2023.00011.
89. Kanduri C, Scheffer L, Pavlović M, et al. Supporting data for "simAIRR: Simulation of Adaptive Immune Repertoires with Realistic Receptor Sequence Sharing for Benchmarking of Immune State Prediction Methods." GigaScience Database. 2023. https://doi.org/10.5524/102434.

Received: February 21, 2023. **Revised:** July 20, 2023. **Accepted:** August 29, 2023

© The Author(s) 2023. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad074/7319582 by guest on 21 February 2024