frontiers in Immunology

METHODS
published: 26 September 2019
doi: 10.3389/fimmu.2019.02159

A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies

Mikhail V. Pogorelyy[1,2] *and Mikhail Shugay*[1,2,3]*

[1] *Genomics of Adaptive Immunity Department, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia,* [2] *Institute of Translational Medicine, Pirogov Russian Medical State University, Moscow, Russia,* [3] *Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia*

Recently developed molecular methods allow large-scale profiling of T-cell receptor (TCR) sequences that encode for antigen specificity and immunological memory of these cells. However, it is well-known that the even unperturbed TCR repertoire structure is extremely complex due to the high diversity of TCR rearrangements and multiple biases imprinted by VDJ rearrangement process. The latter gives rise to the phenomenon of "public" TCR clonotypes that can be shared across multiple individuals and non-trivial structure of the TCR similarity network. Here, we outline a framework for TCR sequencing data analysis that can control for these biases in order to infer TCRs that are involved in response to antigens of interest. We apply two previously published methods, ALICE and TCRNET, to detect groups of homologous TCRs that are enriched in samples of interest. Using an example dataset of donors with known HLA haplotype and CMV status, we demonstrate that by applying HLA restriction rules and matching against a database of TCRs with known antigen specificity, it is possible to robustly detect motifs of epitope-specific responses in individual repertoires. We also highlight potential shortcomings of TCR clustering methods and demonstrate that highly expanded TCRs should be individually assessed to get the full picture of antigen-specific response.

Keywords: T-cell receptor, antigen, motif inference, immune repertoire, high-throughput sequencing

OPEN ACCESS

Edited by:
Thomas Herrmann, Julius Maximilian University of Würzburg, Germany

Reviewed by:
Debbie Van Baarle, National Institute for Public Health and the Environment, Netherlands
Sarina Ravens, Hannover Medical School, Germany

***Correspondence:**
Mikhail Shugay
mikhail.shugay@gmail.com

Specialty section:
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology

Received: *23 June 2019*
Accepted: *28 August 2019*
Published: *26 September 2019*

Citation:
Pogorelyy MV and Shugay M (2019) A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies. Front. Immunol. 10:2159. doi: 10.3389/fimmu.2019.02159

INTRODUCTION

Immune repertoire profiling technology [AIRR-Seq (1)] is an efficient technique that can be employed to study the structure and dynamics of the adaptive immune system. AIRR-Seq makes it possible to characterize the structure of both naive and antigen-experienced T-cell receptor (TCR) repertoires (2–4), tumor infiltrating T-cells (5), and TCRs related to autoimmunity (6), leading to numerous downstream applications in both basic and applied immunological research (7). While novel single-cell RNA sequencing methods allow coupling individual T-cell clones to their phenotype and function using their gene expression profiles (8), the actual antigen specificity (i.e., the set of antigens that can be potentially recognized by a given TCR) remains a mystery for most of the T-cells observed by high-throughput profiling. Even with deep repertoire profiling, the number of unique TCR variants obtained from MHC-multimer positive T-cell fraction is usually below 10^4 (9), dwarfed by the highly conservative estimate of 10^8 for the diversity of TCR beta chain (10).

Recent developments in the field of bioinformatic analysis of AIRR-Seq data are aimed at providing a mean for annotation of TCR repertoires with predicted antigen specificities. For example, the McPAS-TCR database (11) lists pathogen- and disease-associated TCRs and the VDJdb database (12) features a large set of TCRs with experimentally verified epitope specificities and their MHC restrictions. Existing computational methods for TCR repertoire annotation allow both matching against a database of known antigen specificities (12, 13) and clustering of TCR sequences for *de novo* motif detection (4, 14). Annotation of a large number of TCR repertoires from healthy donors (15, 16) demonstrates both high variance of frequencies of epitope-specific T-cells and the imprint of past and ongoing pathogen encounters. Thus, *de novo* discovery of T-cells associated with antigens of interest or certain disease appears to be a hard problem, complicated by the biases in the structure of the naive (unperturbed) TCR repertoire (17), the presence of existing clonal expansions specific to unrelated pathogens, and the high number of false positives that result from the extremely high diversity of the TCR repertoire.

In the present paper, we describe a general framework that can be used to infer sets of T-cells specific to antigens of interest using AIRR-Seq data and TCR neighborhood enrichment algorithms (ALICE and TCRNET). Throughout the study, we apply the TCRNET algorithm for most of the analysis relying on controls generated from healthy donors and switch to the ALICE method that uses a built-in control (VDJ rearrangement model) for the analysis of hematopoietic stem cell transplant time course as there is no feasible control for this dataset. The notable difference between ALICE and TCRNET methods is that while the former utilizes a common VDJ rearrangement model and controls for intrinsic biases of the VDJ rearrangement process, the latter relies on a user-provided set of samples additionally controlling for thymic selection and common pathogen-specific T-cell expansions.

We discuss how various biases of AIRR-Seq datasets can be handled using proper experimental design and give a theoretical basis for the proper application of methods that are based on the probabilistic model of VDJ rearrangement. Using an example dataset of individual human TCR repertoires, we demonstrate the capability of the framework to infer HLA-restricted antigen-specific responses, discuss possible modifications of the proposed method, and expose potential shortcomings of the existing methodology that should be taken into account when running antigen-specific TCR inference.

MATERIALS AND METHODS

AIRR-Seq Data Analysis

Six samples were selected from a large TCRbeta repertoire sequencing dataset published by Emerson et al. (18). The samples were chosen based on HLA matching; they include CMV^- and CMV^+ donors and four controls with CMV^- status (see **Figure 3** for sample IDs). Short nucleotide sequences covering the CDR3 region together with Variable (V) and Joining (J) gene parts were then re-aligned with MiXCR software (19) to produce clonotype tables compatible with VDJtools software (20) and to resolve cases with missing V/J allele calls.

VDJ Rearrangement Simulation, Network Analysis, and Repertoire Annotation

Random TCRbeta sequences were simulated using OLGA software (21) with default VDJ rearrangement model parameters and V/J allele sequences. TCR similarity networks were constructed by allowing a single substitution (a Hamming distance of 1) in CDR3 amino acid sequences. Neighborhood size (degree) enrichment of TCR similarity network nodes was tested against the VDJ rearrangement model using ALICE algorithm (4). The minimal number of neighbors was set to 2, and Q selection factor was set to 1 (no thymic selection) for the analysis of sequences generated with OLGA (see **Figure 2**) and to 9.41 (default) for the analysis of the HSCT dataset (see **Figure 5**). Node neighborhood enrichment test against a pooled control dataset of real TCR repertoires was performed using the TCRNET algorithm implemented in the VDJtools software (2, 20). TCR repertoire annotation was performed using the VDJdb database (12) with a single substitution allowed in the CDR3 amino acid sequence using VDJmatch software (https://github.com/antigenomics/vdjmatch).

Customizing the Framework for Analysis of User-Provided Datasets

Users can modify the core R markdown notebook of the framework located at https://github.com/antigenomics/tcr-annotation-methodology/blob/master/tutorial.Rmd to make use of their own sets of control and test samples. For example, the control dataset (control.txt.gz) can be replaced with a pooled dataset of healthy controls relevant for the user's experimental setup, while CMV^+ and $B35^+$ datasets can be replaced with a set of samples of interest (test samples). Datasets should be stored in VDJtools format (https://vdjtools-doc.readthedocs.io/en/master/input.html#vdjtools-format) and it is recommended to pool samples using VDJtools "PoolSamples" routine with "-i strict" parameter to ensure that separate VDJ rearrangement events (V+J+CDR3 nucleotide sequence) are considered. After that, users can re-run the entire notebook with minor modifications.

RESULTS

Considerations for Experimental Design and the Analysis Pipeline

There are several factors that should be controlled for when searching for TCR motifs associated with a certain treatment or disease (**Figure 1A**). Firstly, HLA restriction is the major factor that shapes the entire response: TCR motifs result from a response targeting certain epitope and are very likely to be absent in case some of the donors do not have a specific HLA haplotype even when there are no other differences between donor phenotypes. Thus, HLA typing is a prerequisite for any AIRR-Seq study that aims at detecting TCR motifs, and both test

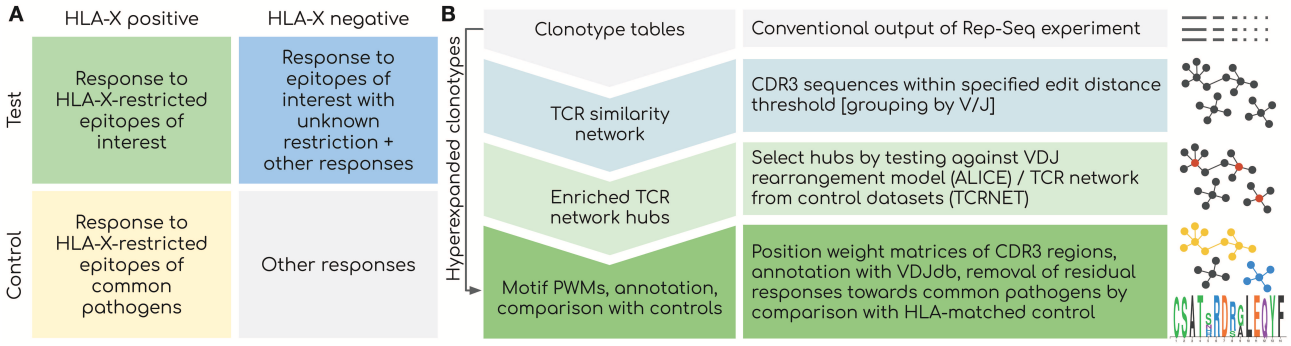

FIGURE 1 | Overview of the proposed framework. **(A)** Considerations for experimental design. Controlling for HLA type is critical as the response toward epitopes linked to the disease of interest is shaped by HLA binding restrictions: donors with distinct HLA can have a completely different response linked to the same disease. Using a control set with matched HLAs is also critical as a response to present and past common infections will be present among all donors. **(B)** Outline of the bioinformatic analysis pipeline. Network-based analysis is aimed at detecting groups of T-cell clonotypes with homologous TCRs that were expanded during antigen-specific response. Statistical testing offered by ALICE/TCRNET methods is required to control for the intrinsic structure of TCR repertoire shaped by the process of VDJ rearrangement and presence of common pathogens (TCRNET case). Downstream annotation with VDJdb can be used to infer actual antigen specificities of TCR clusters and to remove common antigen-specific responses in case of a study aimed at detecting a response to novel/rare pathogens and self-antigens. Hyperexpanded clonotype should be considered even in the case they are not included in homologous TCR clusters for the reasons described later in the following sections of the paper.

and control cohorts should be carefully balanced according to HLA frequency.

Another factor to account for is the imprint of past infections that is subject to HLA restriction (15, 16). Multiple clonal expansions related to common pathogens can be detected across a partially HLA-matched set containing patients and healthy controls that are unrelated to the studied case. Therefore, a large HLA-matched cohort of healthy donors is required to filter out TCR motifs associated with common infections such as EBV.

Finally, the features of the unperturbed TCR repertoire structure itself should be considered, as the repertoire is heavily shaped by the process of VDJ rearrangement. The fact that the VDJ rearrangement process can be described by a relatively simple probabilistic model makes it possible to accurately predict population frequencies of specific TCRs (15). However, huge differences in epitope-specific TCR frequencies can lead to the detection of potentially irrelevant high-frequency (public) TCRs in case the cohort size is relatively small as illustrated in Bagaev et al. (22). Moreover, those public TCRs are the basis for large hubs of a TCR similarity network (17); therefore, they can be mistaken for real homologous clonal expansions. The means to handle this factor are described in the next section and form the basis of the proposed TCR motif inference framework (**Figure 1B**).

Theoretical Basis for TCR Sequence Motif Inference

There are three assumptions that make it possible to detect a set of homologous TCRs that are involved in an ongoing antigen-specific response from AIRR-Seq data:

1a. TCR rearrangement process follows the probabilistic model of Murugan et al. (23). This assumption allows one to compute the expected incidence rate of a given TCR, which is in good agreement with observed results (15) in case a single amino acid substitution is allowed in the CDR3 sequence.

1b. As it follows from Murugan et al. (23), average TCR sequence is extremely rare: median rearrangement probability across all TCRs is $\sim 10^{-12}$. Thus, given one sample's $\sim 10^{6}$ T-cells from a total of $\sim 10^{11}$ T-cells in human peripheral blood, the expected count for a TCR clonotype is $\ll 1$. This allows us to model TCR sampling using Poisson distribution and is required for point 3.

2. There are multiple T-cells with homologous TCRs that recognize the same epitope in an individual. As can be observed in MHC-multimer sequencing studies [see, e.g., Neller et al. (24)], it is typical for the set of multimer+ cells to contain groups of homologous TCRs, although it is still possible that the response to a given epitope is either monoclonal or utilizes a set of distinct TCR sequences, in which case it is not possible to infer any motif.

3. Antigen-specific T-cells expand upon antigen exposure and rare variants go above the detection boundary 1b. This effect allows us to run a neighborhood enrichment test to detect the antigen-driven response without relying on clonotype frequency statistics and pre-exposure control samples.

Assumption 1 can be utilized to build a control TCR similarity network that recaptures biases of the VDJ rearrangement process. An example of a network of 10,000 randomly generated TCRs constructed by connecting CDR3aa regions that differ by a single substitution is given in **Figure 2A**. This network reveals a complex structure with multiple hubs that, as previously shown in Madi et al. (17), are enriched in "public" TCR variants. It is necessary to note that disconnected hubs may arise, in part, due to the fact that TCRs with different CDR3 lengths are not connected when using Hamming distances. Allowing for indels or more substitutions, on the other hand, leads to larger hubs at

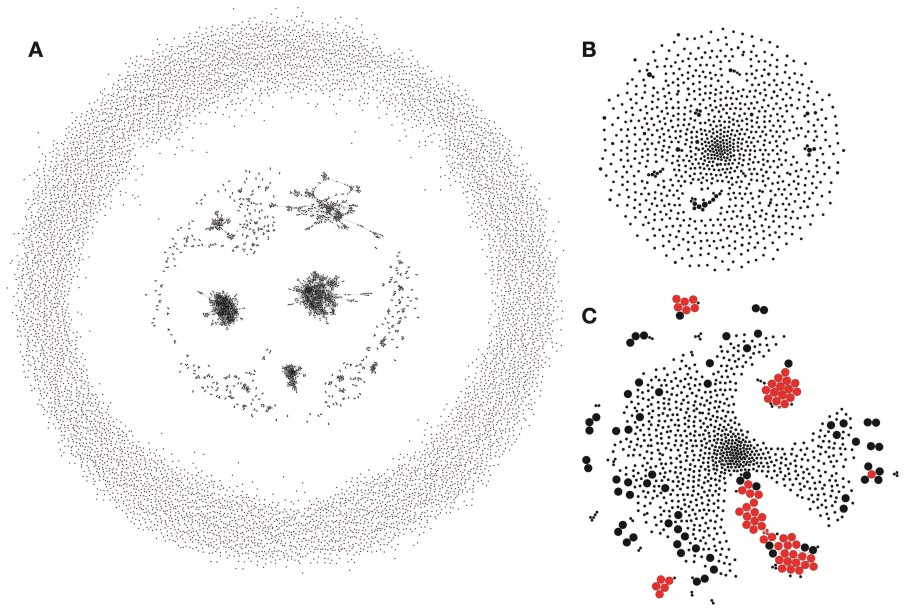

FIGURE 2 | Rep-Seq sample simulated according to the VDJ rearrangement model. **(A)** The similarity network of 10,000 randomly generated TRBV7-6/TRBJ1-4 TCRs. Each vertex shows an individual clonotype, and edges indicate a Hamming distance of 1 or less between CDR3 amino acid sequences. **(B)** A TCR similarity network of 1,000 clonotypes randomly sampled from **(A)** modeling uniform selection from the repertoire. ALICE algorithm identifies no hits (clusters) in this network. **(C)** Modeling an antigen-driven selection by a 100-fold increase in the sampling probability of 50 randomly selected clonotypes and all their observed neighbors from **(A)**. ALICE hits are shown in red. Vertex size shows antigen-driven expansion fold in the initial repertoire. ALICE algorithm identified as significant hits 60 clonotypes (true positives, large red circles) out of 126 expanded in this simulation and four unexpanded clonotypes belonging to enriched clusters (false positives, small red circles); the rest of the 870 clonotypes are thus true negatives. While there are some clusters of cooperatively expanded similar clonotypes, there are also a lot of expanded singletons, which have no neighbors (large black circles, false negatives).

A

Sample	*ID*	*a1*	*a2*	*b1*	*b2*	*status*
B35+	HIP02877	A*26	A*33	B*14	B*35	CMV-
CMV+	HIP13994	A*02	A*02	B*07	B*44	CMV+
Control-1	HIP03484	A*02	A*02	B*07	B*58	CMV-
Control-2	HIP03592	A*02	A*32	B*07	B*39	CMV-
Control-3	HIP04532	A*02	A*24	B*07	B*51	CMV-
Control-4	HIP04576	A*02	A*30	B*07	B*18	CMV-

B

Sample	*ID*	*a1*	*a2*	*b1*	*b2*	*status*
B35+	HIP02877	A*26	A*33	B*14	B*35	CMV-
CMV+	HIP13994	A*02	A*02	B*07	B*44	CMV+
Control-1	HIP03484	A*02	A*02	B*07	B*58	CMV-
Control-2	HIP03592	A*02	A*32	B*07	B*39	CMV-
Control-3	HIP04532	A*02	A*24	B*07	B*51	CMV-
Control-4	HIP04576	A*02	A*30	B*07	B*18	CMV-

FIGURE 3 | The design of a benchmark experiment. **(A)** A case study of responses restricted to HLA-B*35 allele. B*35-positive donor is highlighted with orange, and control samples are selected in a way that there is no B*35 allele and are highlighted with blue. Note that the B*35-positive donor is CMV-negative. **(B)** A case of CMV-specific response linked to HLA-A*02 and HLA-B*07 allele. A CMV^{+} donor (orange) is compared to CMV^{-} controls (blue), with A*02 and B*07 alleles matched across all donors. In this case, one expects to find specific TCRs recognizing CMV epitopes presented by A*02 and B*07. Selected samples from the Emerson et al. (18) study were used.

the cost of greatly increasing the number of false positives [see Shugay et al. (12)].

These theoretic assumptions provide a basis for usage of TCR neighborhood enrichment tests implemented in ALICE and TCRNET algorithms (2, 4). The difference between the latter two is that ALICE uses the VDJ rearrangement model described by Murugan et al. as a control, while TCRNET utilizes pools of healthy control samples as a background (negative) set of clonotypes. Thus, the former tests against VDJ rearrangement biases and thymic selection biases defined by the Q-factor (4), while the latter implicitly tests against thymic and peripheral selection biases and common pathogen-specific responses based on the structure of the control TCR set.

In a random sample of 1,000 TCRs, one can still observe some of those hubs; however, those are smaller and lack any TCRs enriched with neighbors according to the ALICE algorithm (**Figure 2B**). We next simulate antigen-specific clonal expansion by sampling 50 random TCRs and their observed one mismatch neighbors at 100× higher rate. This simulation leads to **Figure 2C** in which ALICE detects several enriched clusters, though there are multiple expanded TCRs having no neighbors (monoclonal expansions). Note that the latter are naturally occurring and were previously observed in multiple studies that involve tetramer-based enrichment of antigen-specific T-cells (13); a detailed description of such cases is provided in the following sections.

Inferring CMV-Specific and B35-Restricted Responses

Straightforward annotation of selected samples (CMV^+, $B35^+$, and four pooled controls, see **Figure 3**) by querying the VDJdb database with one substitution allowed (see Materials and Methods) results in a huge variety of antigen specificities (**Supplementary Figure 1**). The latter include HIV- and HCV-specific clonotypes that are not expected for systematically healthy individuals, and it is hard to tell the overall difference observed in CMV and B35 samples with respect to control. Applying HLA (restricting to HLA-B*35 for B35 sample and HLA-A*02/HLA-B*07 for CMV sample) and pathogen restriction rules (CMV for CMV^+ sample) that follow from our experimental design greatly reduces the complexity of observed results (**Supplementary Figure 2**). However, while the presence of EBV-related B*35:EPL clonal expansions is evident for the B*35 donor, it is hard to tell whether the CMV-specific

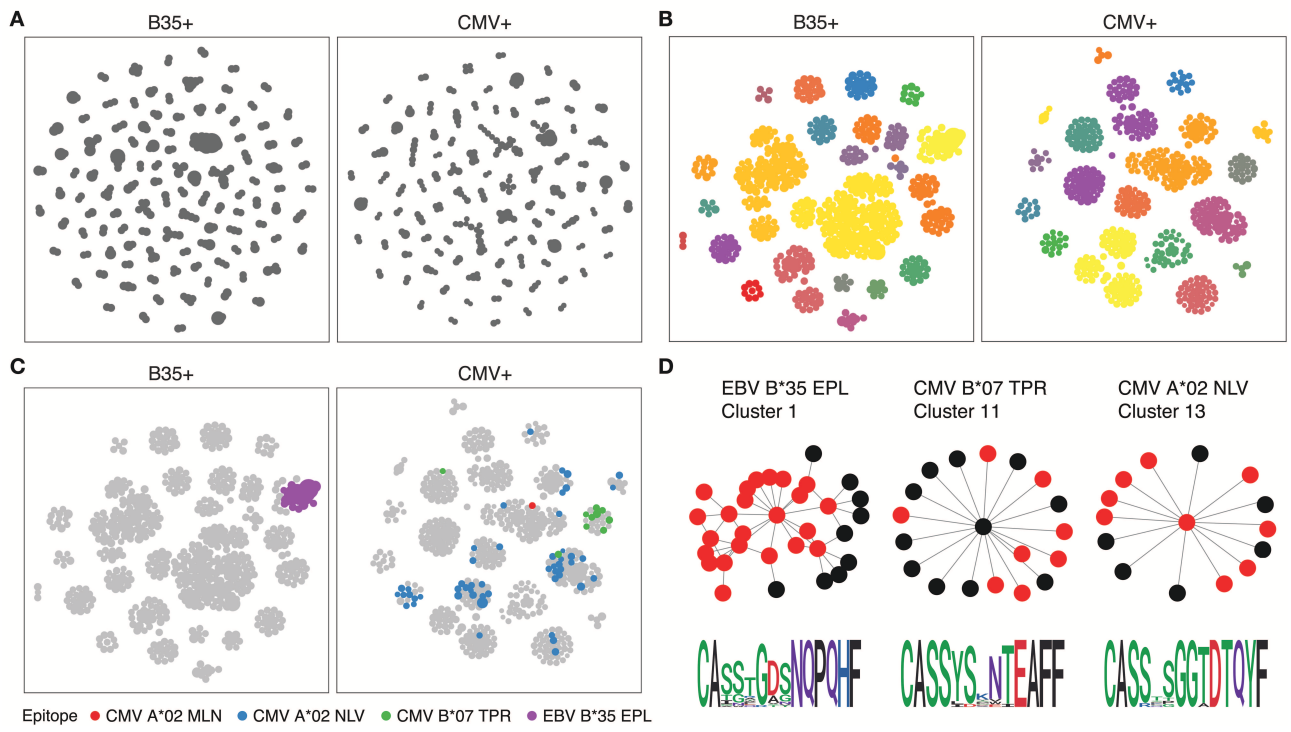

FIGURE 4 | Results of proposed framework when applied to the benchmark experiment. **(A)** A TCR similarity network built for top 3,000 TCRs present in $B35^+$ and CMV^+ samples. **(B)** Selected network hubs for the TCR similarity network built on top of TCRs enriched in neighbor sequences. Each connected component is colored with a distinct color. **(C)** VDJdb annotations of network hubs based on donor status and HLA restriction. Color shows TCR epitope specificity. **(D)** TCR motifs inferred for selected TCRNET clusters. Graph shows the structure of the connected component of a given cluster, and red nodes represent TCRs annotated with a given epitope specificity according to VDJdb (1 amino acid substitution allowed for CDR3aa). Sequence logos of CDR3aa sequences are given at the bottom of each cluster. TCR similarity networks are created using a Hamming distance threshold of 1 for CDR3aa sequences. Multidimensional scaling (MDS)-based graph layout is used for graph visualization in two dimensions.

clonal expansions are significantly different from control for the CMV sample.

We therefore ran the *de novo* motif discovery procedure for B35 and CMV samples using the TCRNET algorithm and specifying the background dataset to be the pool of four control samples (see Materials and Methods). Notably, there is a correlation of TCR neighbor enrichment rate with the overall expansion of those TCRs (**Supplementary Figure 3**). As one can see from **Figures 4A,B**, the TCR similarity network selected by TCRNET is substantially different from the one built using the top (by frequency) 1,000 clonotypes from those samples. Namely, while the latter resembles the unperturbed network of random VDJ rearrangements with power law distribution of hub sizes, the network observed in **Figure 4B** has a more uniform hub size distribution. By annotating the resulting network against VDJdb (**Figure 4C**), one can see the presence of network hubs that have a huge fraction of associated clonotypes annotated with the same antigen specificity (**Table 1**). We can, therefore, derive position weight matrices for specific responses (**Figure 4D**) and build corresponding TCR motifs. Notably, this way, we do re-assign many clonotypes of unknown specificity to the predicted specificity of their hubs and can further extend our knowledgebase of TCRs of known specificity with these predictions.

As there is a huge variance in CMV responder phenotypes and there is no predominant CMV-specific TCR motif compared to response to EBV B*35:EPL, we have additionally assayed all A*02+B*07+ donors from the Emerson et al. dataset (18). Our results (**Supplementary Figures 4**, **5**) show that the proposed methodology can indeed enhance the detection of CMV-specific clonotypes on an extended set of 28 donors. As can be seen in **Supplementary Figure 4**, mean TCR frequency is significantly higher in VDJdb-annotated datasets than in raw datasets for both A*02:NLV and B*07:TRP epitopes. Moreover, applying TCRNET algorithm and selecting TCRs with an adjusted *P*-value for neighbor enrichment of <0.05 lead to enrichment of unique CMV-specific TCRs as can be seen in **Supplementary Figure 5**.

TABLE 1 | VDJdb annotation results for TCRNET clusters.

Sample	Cluster ID	Cluster size	Epitope	Percent matched (%)	$P\text{-value}_{adj}$
$B35^+$	1	32	EBV B*35 EPL	71.9	1.9×10^{-26}
CMV^+	11	19	CMV B*07 TPR	42.1	6.9×10^{-10}
CMV^+	13	15	CMV A*02 NLV	60.0	1.4×10^{-06}
CMV^+	8	56	CMV A*02 NLV	26.8	1.2×10^{-04}
CMV^+	1	80	CMV A*02 NLV	15.0	0.08
CMV^+	3	7	CMV A*02 NLV	28.6	0.13
CMV^+	4	101	CMV A*02 MLN	1.0	0.17
CMV^+	4	101	CMV A*02 NLV	5.9	0.39
CMV^+	15	33	CMV B*07 TPR	3.0	0.45
CMV^+	6	28	CMV A*02 NLV	10.7	0.74
CMV^+	17	13	CMV A*02 NLV	7.7	1.00
CMV^+	7	39	CMV A*02 NLV	7.7	1.00
CMV^+	8	56	CMV B*07 TPR	1.8	1.00

The table shows the percentage of matches to a given epitope in a given cluster according to the VDJdb database (1 amino acid substitution allowed for CDR3aa).
P-values for the frequency of matches were computed based on binomial distribution using overall annotation rate across the sample, the number of specific TCR matches, and the size of a given cluster; P-values were adjusted for multiple testing. Clusters that have no VDJdb matched are not shown.

Exploring the Case of Dominant Clonal Expansions

While we were able to obtain a set of high-confidence TCR motifs in the previous section, a deeper exploration of the dataset, however, reveals that there is a huge fraction of T-cells that were missed by our analysis. Namely, when exploring VDJdb hits that are associated with large clonal expansions, one can see that there are certain large clonal expansions that exactly match CMV-specific clonotypes, yet do not fall into any of the listed motifs (**Supplementary Table 1**).

To further highlight this issue, we ran the pipeline for hematopoietic stem cell transplant (HSCT) time course data (25) that are known to result in large CMV-specific clonal expansions (**Figure 5**). We followed the fate of the A*02 NLV-specific TCRbeta CDR3aa sequence CASSLAPGATNEKLFF to trace the corresponding response. The corresponding TCR clonotype is detected in all time points, both prior to and after HSCT, where it reaches top three of expanded clonotypes. This clonotype is one of those that occupy homeostatic space following repertoire reset during HSCT and one would expect that homologous TCRs involved in CMV-specific response will follow it. However, while corresponding clonal expansion is evident both prior to HSCT and throughout the whole time course, homologous variants start to emerge and become detectable by the ALICE only after 10 months post-HSCT. This suggests that tracking of hyperexpanded clonotypes can be more sensitive than TCR neighborhood enrichment methods at early stages of response where the latter fail to detect a sufficient number of homologous TCRs.

DISCUSSION

As demonstrated using the example dataset, the proposed framework can successfully detect both TCR motifs associated with specific pathogens and TCRs associated with a response toward epitopes restricted by a given HLA. Noting that EBV-specific response detection is well-expected due to the fact that most individuals are EBV-positive, it is necessary to mention that a multitude of unknown responses are detected for $B35^+$ sample and most CMV-specific TCR annotations are restricted to individual clusters in the CMV donor, suggesting that the proposed method has both high precision in detecting known antigen-specific responses and high recall for detecting novel ones. As demonstrated above, the knowledge of donor HLAs can greatly simplify the analysis by narrowing down the list of potential specific TCR candidates. Currently, this seems the only way to combat the huge number of false positives that arise during VDJdb matching due to the immense diversity of TCR repertoires. On the other hand, the framework that we propose has the benefit of eliminating

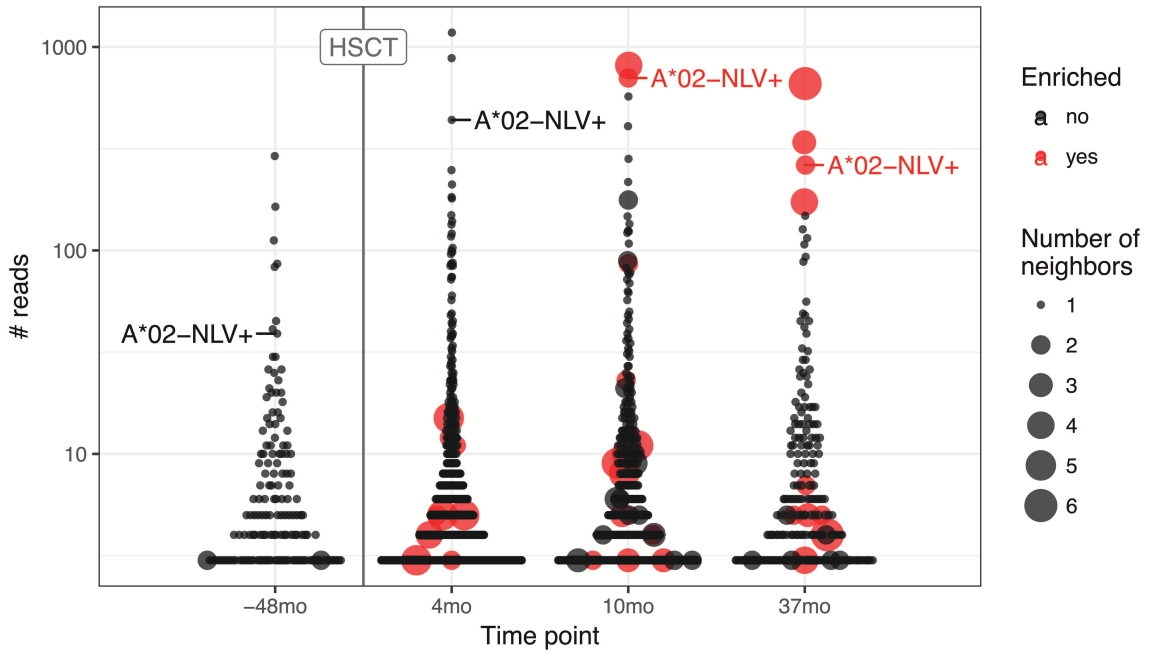

FIGURE 5 | TCR motifs inferred by ALICE and tracking of CMV-specific TCR expansion for the hematopoietic stem cell transplant (HSCT) experiment. Beeswarm plots show the distribution of the number of reads associated with TCR clonotypes prior and after HSCT. The A*02-NLV-specific TCR variant with "CASSLAPGATNEKLFF" CDR3aa sequence is highlighted for both the sample prior to HSCT (−38 months) and later samples (4, 10, and 37 months). Red dots show the clonotypes that have unexpectedly high number of neighbors as inferred by ALICE ($P < 0.001$ after Benjamini–Hochberg multiple testing correction), and point size reflects the number of neighbor sequences. Only TCRs supported by at least three sequencing reads are shown.

spurious matches that arise due to the presence of "public" clonotypes that can be shared across a wide range of samples by chance.

The main message of the present paper is that one should account for HLA restriction rules that govern epitope presentation and guide the response, while VDJdb-based annotation can highlight TCRs that were expanded in response to a certain antigen and thus facilitate TCR data analysis. Moreover, both ALICE and TCRNET can be used to select a fraction of repertoire enriched in responder T-cells and thus enhance the specificity of VDJdb-based annotation. We would also like to explicitly stress the fact that HLA restriction rules should be accounted for in an experimental design as T-cells cannot mount a response against epitopes that are not presented, and the presence of T-cells specific to a certain pathogen in an individual is uninformative unless these T-cells are targeted to epitopes that can be presented by donor HLAs.

The proposed approach can be further extended to a wide range of applications beyond previously reported detection of antigen-specific response in yellow fever virus vaccination studies (26). A successful application of a simpler VDJ rearrangement model-based approach that does not utilize TCR similarity networks to autoimmunity studies with strong HLA linkage such as ankylosing spondylitis (6) suggests that our approach can be utilized for detecting autoimmunity-related TCR motifs. One can also apply the proposed approach to cancer studies, in case an overexpression of certain oncogenes or oncogenic isoforms is expected. For the latter, one should expect the presence of an immunogenic neoantigen that is both restricted to a certain HLA allele and is characterized by overall low expression in healthy tissue.

We also suggest that the proposed framework can be used to expand the set of known antigen-specific TCRs that is currently negligibly small [$\sim 10^4$ variants (12)] compared to the overall repertoire diversity. Indeed, one can assign neighbors of enriched specific TCR clonotypes to the set of epitope-specific T-cells once corresponding homologous TCR clusters are detected in multiple donors. As some of the clusters have a statistically significant number of annotated TCRs even in the presence of a huge fraction of unannotated ones, one can expect to greatly increase the number of clonotypes in a database of TCRs with known specificity using those putative TCR variants.

One of the drawbacks of the proposed methodology follows from the fact that in case when the HLA disease association is relatively vague with no predominant HLA disease susceptibility [e.g., multiple sclerosis (27) or type 1 diabetes (28)], a huge cohort of donors featuring various HLAs should be used to detect potential TCR motifs for self-antigens. Another potential issue results from the presence of monoclonal expansions that are hard to cluster into inferred TCR motifs. The solution here is to treat all hyperexpanded

clonotypes separately and rely on a database of TCRs with known antigen specificities to annotate those clonotypes. The proposed framework is designed for single-chain data; however, it is rather straightforward to extend it to paired alpha-beta TCR analysis: in case certain pairing information is available via, say, scRNA-seq, one can try to pair individual alpha and beta TCR motifs based on their co-occurrence in single-cell data.

Both the methodology and the interactive analysis notebook provided as a supplement to this manuscript are easy to extend and adapt for post-analysis of various AIRR-Seq datasets; we hope that the described framework will be of high utility for future exploratory AIRR-Seq studies that aim at discovering novel antigen- and disease-specific TCR variants.

DATA AVAILABILITY STATEMENT

All data and code for this manuscript are available at https://github.com/antigenomics/tcr-annotation-methodology. Running the code requires installing several R packages listed in the notebook and Java 1.8+. The code was tested and runs without problems on a Unix system with a 4-core Intel CPU and 6 GB RAM.

AUTHOR CONTRIBUTIONS

MP and MS processed and analyzed the data and wrote the manuscript. MS supervised the study.

FUNDING

This study was supported by Russian Science Foundation Grant No. 17-15-01495.

ACKNOWLEDGMENTS

Some parts of this work were originally presented during a tutorial session at the 2nd Meeting on Stochasticity and Control in Adaptive Immune Repertoires (Paris, 2018); the authors would like to thank the participants of this tutorial for their valuable feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2019.02159/full#supplementary-material

REFERENCES

1. Rubelt F, Busse CE, Bukhari SAC, Bürckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol.* (2017) 18:1274–8. doi: 10.1038/ni.3873
2. Ritvo PG, Saadawi A, Barennes P, Quiniou V, Chaara W, El Soufi K, et al. High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proc Natl Acad Sci USA.* (2018) 115:9604–9. doi: 10.1073/pnas.1808594115
3. Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, et al. Dynamics of individual T cell repertoires: from cord blood to centenarians. *J Immunol.* (2016) 196:5005–13. doi: 10.4049/jimmunol.1600005
4. Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T, et al. Detecting T-cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.* (2019) 17:e3000314. doi: 10.1371/journal.pbio.3000314
5. Ostmeyer J, Christley S, Toby IT, Cowell LG. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.* (2019) 79:1671–80. doi: 10.1158/0008-5472.CAN-18-2292
6. Komech EA, Pogorelyy MV, Egorov ES, Britanova OV, Rebrikov DV, Bochkova AG, et al. $CD8^+$ T cells with characteristic T cell receptor beta motif are detected in blood and expanded in synovial fluid of ankylosing spondylitis patients. *Rheumatology.* (2018) 57:1097–104. doi: 10.1093/rheumatology/kex517
7. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol.* (2017) 35:203–14. doi: 10.1016/j.tibtech.2016.09.010
8. Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single cell transcriptomics to explore the immune system in health and disease. *Science.* (2017) 358:58–63. doi: 10.1126/science.aan6828
9. Chen G, Yang X, Ko A, Sun X, Gao M, Zhang Y, et al. Sequence and structural analyses reveal distinct and highly diverse human $CD8^+$ TCR repertoires to immunodominant viral antigens. *Cell Rep.* (2017) 19:569–83. doi: 10.1016/j.celrep.2017.03.072
10. Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ, Chudakov DM. Huge overlap of individual TCR beta repertoires. *Front Immunol.* (2013) 4:466. doi: 10.3389/fimmu.2013.00466
11. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics.* (2017) 33:2924–9. doi: 10.1093/bioinformatics/btx286
12. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* (2018) 46(D1):D419–27. doi: 10.1093/nar/gkx760
13. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope specific T cell receptor repertoires. *Nature.* (2017) 547:89–93. doi: 10.1038/nature22383
14. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature.* (2017) 547:94–8. doi: 10.1038/nature22976
15. Pogorelyy MV, Fedorova AD, McLaren JE, Ladell K, Bagaev DV, Eliseev AV, et al. Exploring the pre-immune landscape of antigen-specific T cells. *Genome Med.* (2018) 10:68. doi: 10.1186/s13073-018-0577-7
16. DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife.* (2018) 7:28. doi: 10.7554/eLife.38358
17. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife.* (2017) 6:21. doi: 10.7554/eLife.22057
18. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet.* (2017) 49:659–65. doi: 10.1038/ng.3822
19. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods.* (2015) 12:380–1. doi: 10.1038/nmeth.3364
20. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: unifying post-analysis of T

cell receptor repertoires. *PLoS Comput Biol.* (2015) 11:e1004503. doi: 10.1371/journal.pcbi.1004503
21. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. *Nat Commun.* (2018) 9:561. doi: 10.1038/s41467-018-02832-w
22. Bagaev DV, Zvyagin IV, Putintseva EV, Izraelson M, Britanova OV, Chudakov DM, et al. VDJviz: a versatile browser for immunogenomics data. *BMC Genomics.* (2016) 17:453. doi: 10.1186/s12864-016-2799-7
23. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA.* (2012) 109:16161–6. doi: 10.1073/pnas.1212755109
24. Neller MA, Ladell K, McLaren JE, Matthews KK, Gostick E, Pentier JM, et al. Naive $CD8^+$ T-cell precursors display structured TCR repertoires and composite antigen-driven selection dynamics. *Immunol Cell Biol.* (2015) 93:625–33. doi: 10.1038/icb.2015.17
25. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, et al. Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol Med.* (2011) 3:201–7. doi: 10.1002/emmm.201100129
26. Pogorelyy MV, Minervina AA, Touzel MP, Sycheva AL, Komech EA, Kovalenko EI, et al. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc Natl Acad Sci USA.* (2018) 115:12704–9. doi: 10.1073/pnas.1809642115
27. Lang HL, Jacobsen H, Ikemizu S, Andersson C, Harlos K, Madsen L, et al. A functional and structural basis for TCR cross-reactivity in multiple sclerosis. *Nat Immunol.* (2002) 3:940–3. doi: 10.1038/ni835
28. Jacobsen LM, Posgai A, Seay HR, Haller MJ, Brusko TM. T cell receptor profiling in type 1 diabetes. *Curr Diab Rep.* (2017) 17:118. doi: 10.1007/s11892-017-0946-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.