The use of non-functional clonotypes as a natural spike-in for multiplex PCR bias correction in immune receptor repertoire profiling

Alexander Komkov[1,2,*], Anastasia Smirnova[1], Anna Miroshnichenkova[2], Egor Volchkov[2], Yulia Olshanskaya[2], Michael Maschan[2], Yuri Lebedev[1,3], Dmitriy Chudakov[1,3,4], Ilgar Mamedov[1,2,3,4,*]

[1] Department of Genomics of Adaptive Immunity, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia.

[2] Dmitry Rogachev National Medical and Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia.

[3] Department of Molecular Technologies, Pirogov Russian National Research Medical University, Moscow, Russia

[4] Central European Institute of Technology, Masaryk University, Brno, Czech Republic

*Correspondence: **alexandrkomkov@yandex.ru** and **imamedov@mx.ibch.ru**

Abstract

High-throughput sequencing of immune receptor repertoires is a valuable tool for receiving insights in adaptive immunity studies. Several powerful methods for TCR/BCR repertoire reconstruction and analysis have been developed in the past decade. However, detection and correction of the discrepancy between real and experimentally observed lymphocyte clone frequencies are still challenging. Here we formulated a quantitative measure of V- and J-genes frequency bias driven by multiplex PCR during library preparation called Over Amplification Rate (OAR). Based on OAR concept, we developed an original software for multiplex PCR-specific bias evaluation and correction named iROAR: Immune Repertoire Over Amplification Removal (https://github.com/smiranast/iROAR). The iROAR algorithm was successfully tested on previously published TCR repertoires obtained using both 5' RACE (Rapid Amplification of cDNA Ends)-based and multiplex PCR-based approaches. The developed tool can be used to increase the accuracy and consistency of repertoires reconstructed by different methods making them more applicable for comparative analysis.

Introduction

Immune receptor (TCR – T-cell receptor and BCR – B-cell receptor) repertoire is usually defined as a set of TCR or BCR sequences obtained from blood, bone marrow or specific lymphocyte population of an individual. Reflecting the T/B cells clonal composition the repertoire is characterized by high degree of specificity for each individual and substantial variation in clone frequencies. Accuracy of both sequences and frequencies of TCR/BCR genes in the obtained repertoire is essential to receive the correct biological information from immune repertoire analysis.

High-throughput sequencing (HTS) of immune receptor repertoires is widely used in immunological studies (reviewed in (Minervina et al. 2019)) for the investigation of immune response to vaccines (Pogorelyy et al. 2018; Sycheva et al. 2018), tumor-infiltrating lymphocytes (Gee et al. 2018), new therapeutic agents (Huang et al. 2019; Wang et al. 2018), leukemia clonality and minimal residual disease monitoring (Brüggemann et al. 2019; Komkov et al. 2020; Nazarov et al. 2016; Wood et al. 2018). HTS-based methods for immune repertoire profiling use either RNA or DNA as a starting material and in most cases use PCR for selective enrichment of receptor sequences. DNA-based methods generally use two-side multiplex PCR with primers annealing to multiple V- and J-genes of the rearranged receptor (Brüggemann et al. 2019; Komkov et al. 2020; Robins et al. 2009). RNA-based methods start with cDNA synthesis, usually with TCR/BCR C(constant)-genes specific oligonucleotides, followed by one-side multiplex amplification with a set of V-gene specific primers and a universal C-gene specific primer (Wang et al. 2010). Alternatively, two universal primers are used for amplification if an artificial sequence is added to the 5' end during synthesis using template-switch (5'-RACE) (Mamedov et al. 2013) or ligation (Oakes et al. 2017). DNA-based methods protect the repertoire from gene transcription bias and provide more comprehensive results (Barennes et al. 2020) which include non-functional (out-of-frame) as well as functional (in-frame) rearrangements, but produce high amplification bias in the course of multiplex PCR. Additionally, each T/B cell contains a single DNA copy (i.e. two target strands) of the receptor molecule in contrast to tens of single-stranded RNA copies. RNA-based methods using 5'-RACE or ligation are characterized by the lowest PCR bias as they need a single primer pair for the amplification. However, the low efficiency of adding a universal oligo to the 5'-end makes its sensitivity comparable or even lower than DNA-based methods. The compromise between these two approaches is the RNA-based method with one-side multiplex that has moderate

amplification bias yet sufficient sensitivity (Ma et al. 2018). Most bias in one-side multiplex RNA-based approaches could be removed by using unique molecular identifiers (UMI) (Ma et al. 2018). Unfortunately, for DNA-based methods efficient incorporation of UMIs into the initial molecule before PCR is still challenging. The only method for DNA multiplex bias correction (Carlson et al. 2013) is undirected and cost-ineffective due to the utilization of an expensive synthetic spike-in control repertoire. Here we propose an orthogonal solution for this challenge: the first fully computational algorithm for amplification bias detection and correction in immune receptor repertoires named iROAR (immune Repertoire Over Amplification Removal).

Results

The rationale for the Over Amplification Rate measure

Since out-of-frame TCR/BCR rearrangements do not form a functional receptor they are not subjected to any specific clonal expansions and selection (Murugan et al. 2012). Being a passenger genomic variation, they change their initial (recombinational) clonal frequencies just randomly following the frequency changes of the second functional (in-frame) TCR/BCR allele present in the same T/B cell clone. According to the mechanism of TCR/BCR loci rearrangement, the formation of in-frame and out-of-frame allele combinations in the same cell is also a stochastic and independent process in terms of V- and J-genes usage. It leads to the conclusion that V- and J-gene frequencies among out-of-frame rearrangements are generally constant during a lifetime and are equal to the initial recombination frequencies despite repertoire changes caused by various immune challenges (Fig 1 and Suppl Fig 1). Thus, reproducible deviation of out-of-frame V- and J-gene usage and frequencies from the initial recombinational frequencies observed in the sequenced repertoire dataset is a result of artificial aberration caused by PCR amplification rather than immune repertoire evolution. Thus out-of-frame clonotypes can be considered as a natural spike-in that can be used to measure amplification bias and quantitatively correct immune repertoire data.

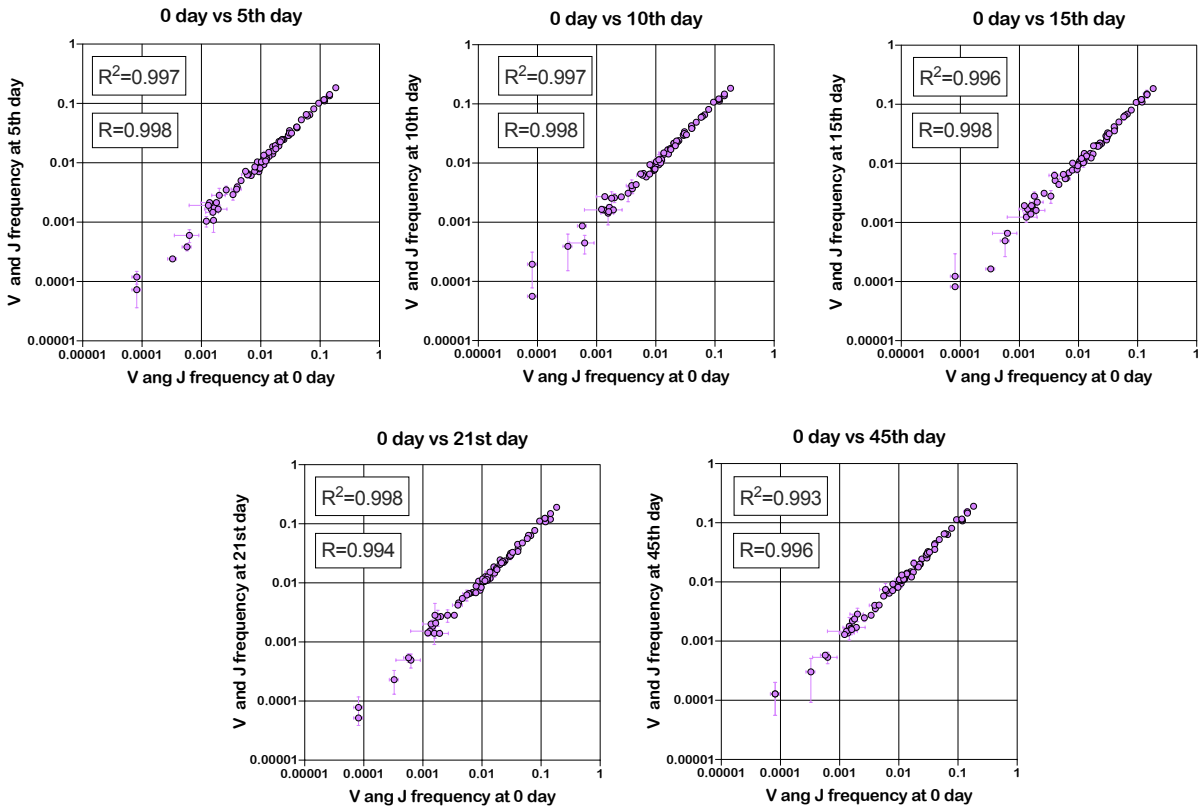

Figure 1. *Stability of TRBV and TRBJ genes frequencies (usage) calculated based on unique out-of-frame rearrangements after Yellow fever vaccination (model of acute viral infection). Out-of-frame clonotypes for frequencies calculation were extracted from low-biased 5' RACE TRB repertoires of PBMC samples obtained in two replicates for 6 time points: 0, 5, 10, 15, 21 and 45 days after YFV injection (donor M1, SRA accession number PRJNA577794* (Minervina et al. 2020)*).*

Formulating this observation, we developed the Over Amplification Rate (OAR) measure which we define as a ratio of the observed and expected frequency of a V- (OAR(Vi)) or a J-gene (OAR(Ji)) among identified out-of-frame rearrangements. Observed frequency represents a value calculated as a percentage of read counts (RC) for each V- and J-gene divided by the sum of all out-of-frame clones read count in the obtained repertoire sequencing dataset. The expected frequency is a value before clonal expansions calculated as a number of unique clones (UCN) having each V- or J-gene divided by the total number of unique out-of-frame clones in the repertoire.

$$OAR(\boldsymbol{Vi}) = \frac{RC(Vi) \big/ \sum_1^N RC(Vi)}{UCN(Vi) \big/ \sum_1^N UCN(Vi)}$$

$$OAR(\boldsymbol{Ji}) = \frac{RC(Ji) \big/ \sum_1^N RC(Ji)}{UCN(Ji) \big/ \sum_1^N UCN(Ji)}$$

OAR value tends to be equal to 1 under ideal conditions (low or no amplification bias) and deviates from 1 as amplification bias increases in line: 5'-RACE with a single universal primer pair, one-side multiplex PCR (VMPlex) and two-side multiplex PCR (VJMPlex) (Fig 2).

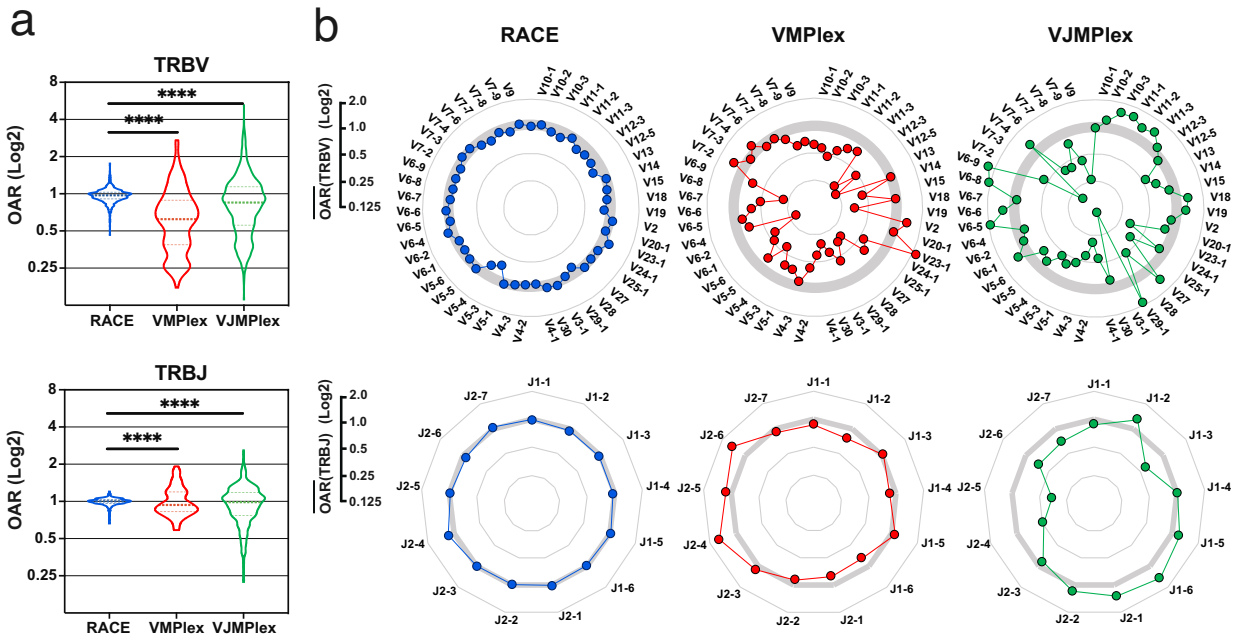

Figure 2. *a. Comparison of OAR values distribution for TRB repertoires obtained with 5'-RACE, one-side multiplex (VMPlex) and two-side multiplex (VJMPlex) PCR. The F-test was performed to compare OAR distributions: ****P<0.0001. b. Average OAR values for TRBV and TRBJ genes in repertoires obtained with 5'-RACE, one-side multiplex (VMPlex) and two-side multiplex (VJMPlex) PCR (for each method type 15 PBMC TCR repertoires were chosen randomly from PRJNA214848, PRJEB27352, PRJNA493983, PRJNA577794, PRJNA550004, PRJNA427746, PRJEB31283, PRJEB33490, https://clients.adaptivebiotech.com/pub/emerson-2013-jim, https://zenodo.org/record/3703311#.Xqm7WpMzZQI, https://zenodo.org/record/27483#.XpCuQ1MzZQI)*

The versatility of OAR measure

OAR measurement is a universal approach and can be applied for different types of immune repertoire data. To demonstrate this versatility, we calculated OAR values for low-biased (5' RACE) repertoires of different immune receptor chains obtained from bulk human PBMC: TCR alpha (TRA), TCR beta (TRB), and BCR heavy chains (Fig. 3a). The results show that OARs for both TCR and BCR repertoires obtained by 5' RACE are close to 1 and stay within the range of 0.5 to 2, which is much narrower than OAR for multiplex PCR-based repertoires (see main text Fig. 2). We also analyzed OARs for low-biased (5' RACE) TCR repertoires of different T cell subpopulations, including T-helper, cytotoxic, central memory, effector memory, and naïve T cells. As shown in Fig. 3, the OAR values demonstrate almost no differences between analyzed T cell types and are close to 1 as for the repertoire of bulk T cells mix obtained from PBMC. Thus, our results demonstrate that OAR is a universal measure of repertoires and can be applied to most immune receptors and cell types.

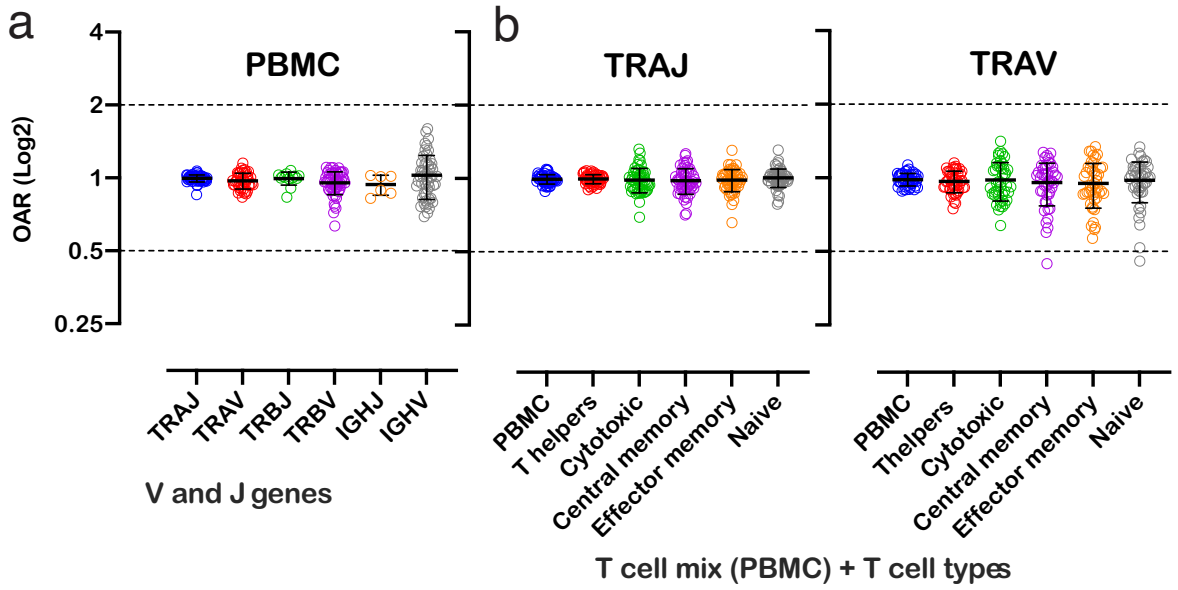

Figure 3. *Distribution of average Over Amplification Rates of V- and J-genes in RACE-based repertoires of various types of TCR and BCR (TCR repertoires: SRA accession numbers: PRJNA577794, PRJNA316572, PRJEB27352; BCR repertoires: SRA accession number: PRJNA297771, PRJNA494572). B. Over Amplification Rates of V- and J-genes of TCR alpha chains in RACE-based repertoires of different types of T-cells (donors M1 and P30, 45th day after booster vaccination, SRA accession number PRJNA577794* (Minervina et al. 2020)*)*

Factors affecting OAR measure accuracy

Importantly, in the case of insufficient sequencing coverage (see Suppl Fig 2 and 3) high PCR bias can lead to dramatic loss of clones and thus incorrect measurement of V- and J-genes frequencies. To prevent this issue, we compare samples' UCN-based frequencies with average frequencies calculated for the population. If the first one significantly differs from populational frequencies, the last one is used for OAR calculation (as long as its dispersion is significantly less than the error of original frequencies) (Suppl. Fig 2). To reduce the impact of enormously amplified (either due to clonal expansion or extreme amplification bias) clones on OAR value we excluded the outliers from OAR calculation using z-test.

It is also important to note, that errors in nucleotide sequences occurring during library preparation and sequencing could lead to an artificial increasing of both in-frame and out-of-frame clone diversity. Single nucleotide substitutions generate artificial clones as a branch of real most abundant clones inside of each in-frame and out-of-frame group independently. Single nucleotide indels lead to cross-generation of artificial clones between groups: real in-frame clones generate false out-of-frame clones and vice versa. Artificial clones compromise the accuracy of both repertoire itself and OAR value. To eliminate such clones generated by single-nucleotide substitutions we filtered them out by VDJTOOLS software (see Methods section). To eliminate artificial clones produced by indels we searched for in-frame and out-of-frame clone pairs which differ by one indel (Levenshtein distance = 1). In such pairs the smaller clone is discarded and its count is added to the count of the larger clone if their ratio is less than 1:500 (to guaranteed discard most sequencing errors present in 1 per 1000 nucleotides average).

Over amplification rate index

To estimate the value of immune repertoire structure disruption by amplification bias we proposed the OAR-index which represents the mean square deviation of OARs for each V and J gene from the value characteristic for repertoire with no bias (OAR=1). OAR-index is directly proportional to the amplification bias and thus can be used for rapid estimation and comparison of immune repertoire bias. The less OAR index is the less PCR bias is with an ideally unbiased repertoire having OAR-index = 0.

$$OAR - index = \sqrt[2]{\frac{\sum_0^n (OARi - 1)^2}{n}}$$

Using OAR for the removal of amplification bias

Normalization coefficients for each V-J combination are estimated by multiplication of corresponding V- and J-gene OARs for two-side multiplex and V-gene OAR for one-side multiplex (Fig 2a). Corrected read count for each clonotype with the particular V-J gene combination is obtained simply by dividing the observed read count by the corresponding normalization coefficient. OAR of V- and J-genes could be co-dependent, what can be a reason of overcorrection. To avoid this issue the procedure can be recursively repeated with a modified normalization coefficient defined as described coefficient raised to the power of a number in the range from 0 to 1. The corrected read counts are used to estimate the real percentage of each clonotype in the repertoire. A detailed flowchart of OAR-based amplification bias correction algorithm named iROAR is shown in Fig 4a.

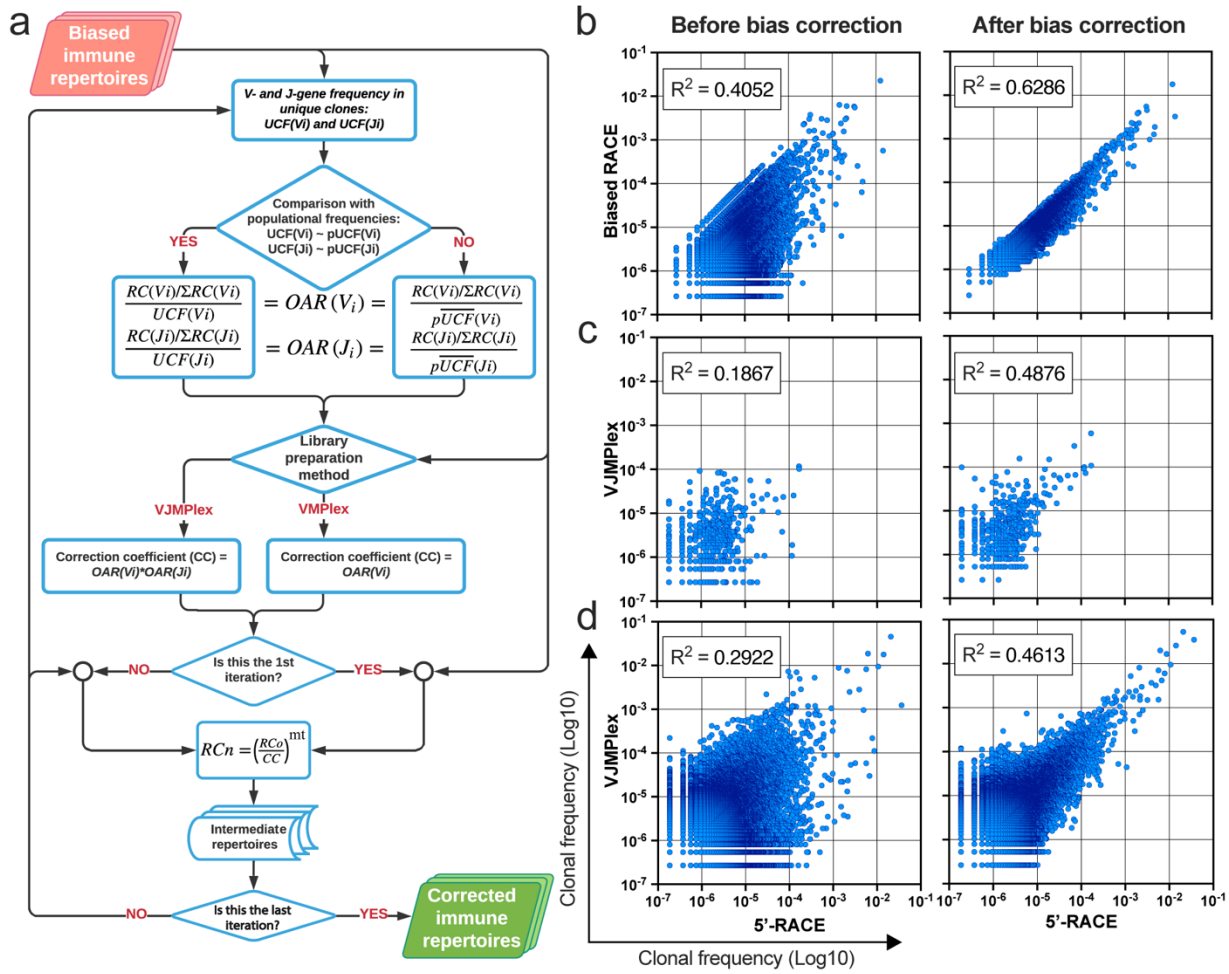

Figure 4. *a. Flowchart of iROAR algorithm. UCF – Frequency calculated using unique clones counts, pUCF – populational UCF, RC – read count, RCn – normalized RC, RCo – observed RC, mt – the number in the range from 0 to 1 for iterative procedure. b. Clone frequencies in low biased 5'-RACE-based repertoire (ENA database, accession number PRJEB27352) versus the same repertoire with introduced artificial bias: before and after iROAR processing. c. Out-of-frame (non-functional) clone frequencies in low biased 5'-RACE-based repertoire versus two-side multiplex (VJMPlex)-based repertoire obtained for the same RNA sample (SRA database, accession numbers SRR3129976 and SRR3129970): before and after iROAR processing. d. In-frame (functional) clone frequencies in low biased 5'-RACE-based repertoire versus two-side multiplex (VJMPlex)-based repertoire obtained for the same RNA sample: before and after iROAR processing.*

OAR-based approach validation

The validation of OAR-based amplification bias correction was performed on TRB dataset with *in silico* introduced bias generated from real (experimental) low-biased (5'-RACE) repertoire (Fig 4b). After correction, the OAR-index indicating general repertoire bias decreased from 1.81 to 0.76 whereas the R-square value of in silico biased and original repertoire correlation raised from 0.4052 to 0.6286. Both measures independently demonstrate substantial reducing of *in silico* introduced quantitative bias. Afterward, we tested our approach on real paired experimental datasets obtained from the same RNA sample by two different methods type: 5'-RACE and multiplex PCR (Liu et al. 2016; Barennes et al. 2020) (Fig 4c-d, Fig 5).

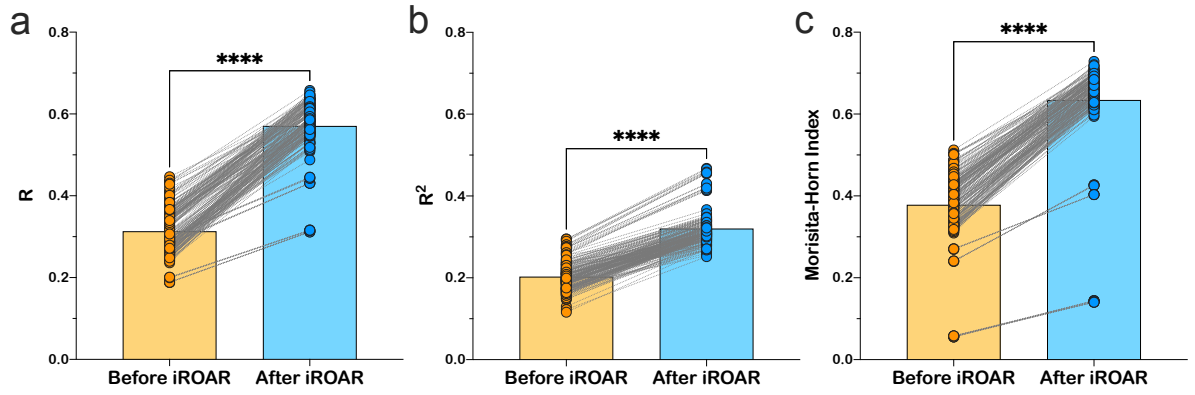

Figure 5. *Effect of iROAR-based PCR bias correction in MPlex repertoire on similarity with low-biased RACE-based repertoire obtained from the same RNA sample. a) Pearson correlation coefficient; b) R-square measure; c) Morisita-Horn similarity index. **** P<0.0001 (two-tailed Wilcoxon matched-pairs signed rank test). Dataset: PRJNA548335 (Donors 1 and 2, 100 ng RNA, 3 types of RACE (RACE-2, RACE-3, RACE-4) protocols and RNA-based MPlex (Multiplex-3) protocol; PRJNA309577 (Donors S01, S02 and S03).*

As a result of amplification bias correction OAR index for multiplex-based repertoire decreased 1.5-fold average. At the same time correlation of clonal frequencies obtained with RACE and multiplex significantly increased (Pearson correlation measure and R-square value increased 1.5-fold average each) with a significant rise of repertoires similarity (Morisita-Horn index increased 1.7-fold average) (Fig 5). Importantly, amplification bias decreased in both out-

of-frame and in-frame clone sets, despite the fact that normalization coefficients are calculated using out-of-frame ones only. Thus, we demonstrate the power of the iROAR algorithm on artificial and experimental immune repertoires.

Discussion

Even a small difference in amplification efficiencies can lead to a massive bias after multiple amplification cycles due to the exponential nature of PCR. Thus, most of the existing methods for immune repertoire library preparation are subjected to amplification bias. The effect of distinct PCR bias-generating factors can be reduced experimentally by varying reaction mixture content and introducing special protocols (UMI, crafty primer structures, spike-in controls). However, the criteria for the estimation and removal of the residual bias after the application of these optimization approaches are lacking. Here we close this gap by introducing the OAR value and OAR-index which score PCR-bias for both V- and J-genes separately (OAR values) and for the whole repertoire dataset (OAR-index). Based on OAR values we developed the first fully computational approach to decipher and correct amplification bias in immune receptor repertoires produced by one-side or two-side multiplex PCR-based methods, using RNA or DNA as a template. Due to the inability to use UMI-based correction for DNA-based multiplex, the developed approach is the only currently available technique allowing to directly measure and correct PCR bias in such repertoires without additional experiments. In contrast to cell-lines spike-in (Knecht et al. 2019) or synthetic repertoire-based (Carlson et al. 2013; Wu et al. 2020) PCR bias correction the proposed approach operates with hundreds and even thousands of natural spike-ins (out-of-frame clones) for each V-J gene pair. It makes this method more reliable due to the ability to minimize the impact of CDR3 structure on PCR bias calculation. Moreover, similarly to previously described method (Carlson et al. 2013) the OAR-based approach can also be used for primer efficacy evaluation to optimize their structures and concentrations, which in turn will straighten the coverage of various V- and J-genes and minimize the number of experimentally lost clones. Being fully computational the developed PCR bias correction algorithm can be easily implemented in any TCR/BCR repertoire analysis pipeline noticeably improving the quantitative parts of the analysis. Being highly versatile the developed approach can provide the opportunity to compare the immune repertoire datasets generated using different library preparation methods.

Methods

Raw-data processing and immune repertoires reconstruction

All sequencing data used in this study represent human TCR and BCR repertoires. Immune repertoires from Adaptive Biotechnology server were used as they are. Other repertoires (see Suppl. Table 1) were reconstructed from fastq data using MiXCR v2 software (Bolotin et al. 2017, 2015) after primers and adapters trimming using FASTP software (Chen et al. 2018). For unification all obtained repertoires were converted to VDJTOOLS(Shugay et al. 2015) format. Erroneous clones generated by single nucleotide substitution were removed from the repertoire using "Correct" function from VDJTOOLS software package. Erroneous clones generated by single nucleotide indels were removed from repertoires using "Filter" function from developed iROAR software. V and J pseudogenes were removed from repertoires using "FilterBySegment" function of VDJTOOLS.

OAR evaluation and statistical analysis

For the OAR and OAR-index calculation as well as for amplification bias removal we used the command-line-based iROAR software designed in this study and freely available for nonprofit use at github (https://github.com/smiranast/iROAR). For the OAR comparison between 5'-RACE, one-side and two-side multiplex PCRs the equal number of out-of-frame clones (50,000) was randomly selected from TCR repertoires of 15 healthy individuals (for each approache). Average population V- and J-gene frequencies (unweighted) were calculated based on out-of-frame clones from 105 TRB repertoires obtained by two methods: 5'-RACE (95 repertoires) and single-cell TCR profiling (10X genomics) (10 repertoires) (Suppl. Table 1) using "CalcSegmentUsage" function "-u" parameter of VDJTOOLS. All statistical tests were performed using Prism9 GraphPad software (https://www.graphpad.com/).

Data access

All analyzed datasets were downloaded from open-source databases: NCBI SRA (https://www.ncbi.nlm.nih.gov/sra), ENA (https://www.ebi.ac.uk/ena), Adaptive Biotechnology Server (https://clients.adaptivebiotech.com/) and Zenodo project (https://zenodo.org/). A full list of web-links and accession numbers is summarized in Suppl. Table 1. The iROAR software and its documentation are available at the link: https://github.com/smiranast/iROAR. The additional

software used in this study is available in GitHub repository (https://github.com/milaboratory/mixcr, https://github.com/mikessh/vdjtools, https://github.com/OpenGene/fastp)

Funding

This work was supported by Russian Science Foundation [grant 20-75-10091 to A.K.], the analysis of TRA MPlex data was supported by Russian Foundation for Basic Research [grant 20-015-00462 to A.K.], the iROAR adaptation for repertoires containing abnormally proliferated clones was supported by Russian Foundation for Basic Research [grant 18-29-09132 to I.M.].

Acknowledgements

We thank Grigory Armeev and Valery Novoseletsky for their help with data storage and processing.

Competing interest statement

The authors declare that they have no competing interests.

Supplementary Data

Supplementary Table 1. XLSX. The file contains accession numbers and links to the datasets used in this study.

Supplementary note. PDF. The file contains supplementary figures and a description of the performed analysis and iROAR software parameters and characteristics.

References

Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, Uddin I, Ismail M, Oakes T, Chain B, et al. 2020. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol*.

Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva EV, Obraztsova AS, Shugay M, et al. 2017. Antigen receptor repertoire profiling from RNA-seq data. *Nat Biotechnol* **35**: 908–911.

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

Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V, Fazio G, Froňková E, Giraud M, Grioni A, et al. 2019. Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia* **33**: 2241–2253.

Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung M-W, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, et al. 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* **4**: 2680.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinforma Oxf Engl* **34**: i884–i890.

Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, Bethune MT, Fischer S, Yang X, Gomez-Eerland R, et al. 2018. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell* **172**: 549-563.e16.

Huang J, Alexey S, Li J, Jones T, Grande G, Douthit L, Xie J, Chen D, Wu X, Michael M, et al. 2019. Unique CDR3 epitope targeting by CAR-T cells is a viable approach for treating T-cell malignancies. *Leukemia* **33**: 2315–2319.

Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V, Krejci A, Grioni A, Pal K, Stranska K, et al. 2019. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia* **33**: 2254–2265.

Komkov A, Miroshnichenkova A, Nugmanov G, Popov A, Pogorelyy M, Zapletalova E, Jelinkova H, Pospisilova S, Lebedev Y, Chudakov D, et al. 2020. High-throughput sequencing of T-cell receptor alpha chain clonal rearrangements at the DNA level in lymphoid malignancies. *Br J Haematol* **188**: 723–731.

Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, Cao H, Su Z, Wang C, Wu J, et al. 2016. Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire. *PloS One* **11**: e0152464.

Ma K-Y, He C, Wendel BS, Williams CM, Xiao J, Yang H, Jiang N. 2018. Immune Repertoire Sequencing Using Molecular Identifiers Enables Accurate Clonality Discovery and Clone Size Quantification. *Front Immunol* **9**: 33.

Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, Lebedev YB, Chudakov DM. 2013. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol* **4**: 456.

Minervina A, Pogorelyy M, Mamedov I. 2019. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transpl Int Off J Eur Soc Organ Transplant* **32**: 1111–1123.

Minervina AA, Pogorelyy MV, Komech EA, Karnaukhov VK, Bacher P, Rosati E, Franke A, Chudakov DM, Mamedov IZ, Lebedev YB, et al. 2020. Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife* **9**.

Murugan A, Mora T, Walczak AM, Callan CG. 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* **109**: 16161–16166.

Nazarov VI, Minervina AA, Komkov AY, Pogorelyy MV, Maschan MA, Olshanskaya YV, Zvyagin IV, Chudakov DM, Lebedev YB, Mamedov IZ. 2016. Reliability of immune receptor rearrangements as genetic markers for minimal residual disease monitoring. *Bone Marrow Transplant* **51**: 1408–1410.

Oakes T, Heather JM, Best K, Byng-Maddick R, Husovsky C, Ismail M, Joshi K, Maxwell G, Noursadeghi M, Riddell N, et al. 2017. Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile. *Front Immunol* **8**: 1267.

Pogorelyy MV, Minervina AA, Touzel MP, Sycheva AL, Komech EA, Kovalenko EI, Karganova GG, Egorov ES, Komkov AY, Chudakov DM, et al. 2018. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc Natl Acad Sci U S A* **115**: 12704–12709.

Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**: 4099–4107.

Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, Pogorelyy MV, Nazarov VI, Zvyagin IV, Kirgizova VI, et al. 2015. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol* **11**: e1004503.

Sycheva AL, Pogorelyy MV, Komech EA, Minervina AA, Zvyagin IV, Staroverov DB, Chudakov DM, Lebedev YB, Mamedov IZ. 2018. Quantitative profiling reveals minor changes of T cell receptor repertoire in response to subunit inactivated influenza vaccine. *Vaccine* **36**: 1599–1605.

Wang B, DeKosky BJ, Timm MR, Lee J, Normandin E, Misasi J, Kong R, McDaniel JR, Delidakis G, Leigh KE, et al. 2018. Functional interrogation and mining of natively paired human VH:VL antibody repertoires. *Nat Biotechnol* **36**: 152–155.

Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson JR, Davis RW, et al. 2010. High throughput sequencing reveals a complex

pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A* **107**: 1518–1523.

Wood B, Wu D, Crossley B, Dai Y, Williamson D, Gawad C, Borowitz MJ, Devidas M, Maloney KW, Larsen E, et al. 2018. Measurable residual disease detection by high-throughput sequencing improves risk stratification for pediatric B-ALL. *Blood* **131**: 1350–1359.

Wu J, Wang X, Lin L, Li X, Liu S, Zhang W, Luo L, Wan Z, Fang M, Zhao Y, et al. 2020. Developing an Unbiased Multiplex PCR System to Enrich the TRB Repertoire Toward Accurate Detection in Leukemia. *Front Immunol* **11**: 1631.