OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–14
https://doi.org/10.1093/bib/bbab566
Problem Solving Protocol

TCRpower: quantifying the detection power of T-cell receptor sequencing with a novel computational pipeline calibrated by spike-in sequences

Shiva Dahal-Koirala †, Gabriel Balaban †, Ralf Stefan Neumann, Lonneke Scheffer, Knut Erik Aslaksen Lundin, Victor Greiff, Ludvig Magne Sollid, Shuo-Wang Qiao‡ and Geir Kjetil Sandve‡

Corresponding authors: Gabriel Balaban, Department of Computational Physiology, Simula Research Laboratory, Oslo, Norway. Tel.: +4767828200; E-mail: gabrib@simula.no. Shiva Dahal-Koirala, Department of Immunology, University of Oslo, Norway. Tel.: +4723072721; E-mail: shiva.dahal-koirala@medisin.uio.no; Geir Kjetil Sandve, Department of Informatics, University of Oslo, Norway. Tel.: +4722840861; E-mail: geirksa@ifi.uio.no; Shuo-Wang Qiao, Department of Immunology, University of Oslo, Norway. Tel.: +4722850533; E-mail: s.w.qiao@medisin.uio.no

†Shared first authors
‡Shared senior authors

Abstract

T-cell receptor (TCR) sequencing has enabled the development of innovative diagnostic tests for cancers, autoimmune diseases and other applications. However, the rarity of many T-cell clonotypes presents a detection challenge, which may lead to misdiagnosis if diagnostically relevant TCRs remain undetected. To address this issue, we developed TCRpower, a novel computational pipeline for quantifying the statistical detection power of TCR sequencing methods. TCRpower calculates the probability of detecting a TCR sequence as a function of several key parameters: *in-vivo* TCR frequency, T-cell sample count, read sequencing depth and read cutoff. To calibrate TCRpower, we selected unique TCRs of 45 T-cell clones (TCCs) as spike-in TCRs. We sequenced the spike-in TCRs from TCCs, together with TCRs from peripheral blood, using a 5′ RACE protocol. The 45 spike-in TCRs covered a wide range of sample frequencies, ranging from 5 per 100 to 1 per 1 million. The resulting spike-in TCR read counts and ground truth frequencies allowed us to calibrate TCRpower. In our TCR sequencing data, we observed a consistent linear relationship between sample and sequencing read frequencies. We were also able to reliably detect spike-in TCRs with frequencies as low as one per million. By implementing an optimized read cutoff, we eliminated most of the falsely detected sequences in our data (TCR α-chain 99.0% and TCR β-chain 92.4%), thereby improving diagnostic specificity. TCRpower is publicly available and can be used to optimize future TCR sequencing experiments, and thereby enable reliable detection of disease-relevant TCRs for diagnostic applications.

Keywords: T-cell receptor, bulk T-cell receptor sequencing, spike-in standards, computational model, TCRpower and adaptive immune receptor repertoire sequencing

Introduction

The adaptive immune system records all past and ongoing immune responses in the form of immune memory (e.g. principle of vaccination), stored in the immune receptors of adaptive immune cells, such as T-cells. Each person has a unique repertoire of T-cell receptors (TCRs), with a high genetic sequence diversity. The number of TCR beta (TRB) clonotypes in an individual has been estimated to be 10^6–10^8 [1, 2], whereas the potential diversity of the paired TCR alpha (TRA) and TRB repertoire, was

Shiva Dahal-Koirala is a postdoctoral researcher in the Department of Immunology at the University of Oslo. Her main research focus is human T cell immunology and T cell receptor sequencing.
Gabriel Balaban is a postdoctoral researcher at Simula Research Laboratory, in Oslo, Norway. He was previously at the Department of Informatics at the University of Oslo and was funded by the PharmaTox Strategic Research Initiative. His research focuses on data science and biophysics-based approaches to studying human health and disease.
Ralf Stefan Neumann is a postdoctoral researcher in the Department of Immunology at the University of Oslo. His research focuses on developing a bioinformatics pipeline to analyze antigen receptor sequencing data.
Lonneke Scheffer is a PhD candidate in computational immunology in the Department of Informatics at the University of Oslo. Her research focuses on quantifying the impact of germline gene variation on immune receptor repertoires.
Knut Erik Aslaksen Lundin is a professor in the Institute of Clinical Medicine at the University of Oslo. He is a gastroenterologist with a research focus on gastrointestinal disorders.
Victor Greiff is an associate professor in the Department of Immunology at the University of Oslo. His research focuses on developing novel computational and experimental strategies by combining antigen receptor sequencing with artificial intelligence and high-dimensional statistics.
Ludvig Magne Sollid is a professor in the Department of Immunology at the University of Oslo. His research focuses on pathogenesis of celiac disease.
Shuo-Wang Qiao is an associate professor in the Department of Immunology at the University of Oslo. Her main research focus is human T cell immunology as well as comparative immunology.
Geir Kjetil Sandve is a professor in the Department of Informatics at the University of Oslo. His current research is focused on development of machine learning methodology to learn sequence patterns in immune cells indicative of disease.

Received: September 16, 2021. **Revised:** December 2, 2021. **Accepted:** December 11, 2021

found to be even higher, 2×10^{19} [3], only a few orders of magnitude less than the estimated number of stars in the universe [4]. Each TCR is specific to one or more antigens. This has allowed for the development of novel diagnostic and therapeutic applications: for autoimmune diseases [5], celiac disease [6], cancer [7] and infectious diseases [8], which are based on high-throughput, bulk TCR sequencing methods.

Several TCR sequencing methods have been developed for the analysis of T-cell populations (bulk sequencing) or of individual T cells (single-cell sequencing) by academics and industrial investigators [9]. These approaches can be broadly classified into DNA-based or RNA-based approaches, as well as multiplex PCR [using panels of V and J primers (RNA and DNA)] or rapid amplification of 5′ complementary DNA ends (RACE) followed by nested PCR based sequencing (RNA only) [9]. These different sequencing approaches have their own merits and limitations [9, 10] affecting the choice of sequencing approach for different applications. Single-cell TCR sequencing provides paired TRA and TRB sequencing, however the number of cells that can be sequenced (10^2–10^3) is much less than bulk TCR sequencing (10^2–10^6, 9). Recently developed commercial single-cell sequencing solutions (10× genomics) have revolutionized the field by providing full-length paired TRA and TRB sequencing of a large number of T cells. However, bulk TCR sequencing approaches are still typically employed for high-throughput analysis of immune cells in health and disease [9]. These bulk TCR sequencing approaches have different accuracies and intra- and inter-method reproducibility for detecting TRA and TRB chains [11].

Quantifying the detection power of TCR sequencing methods is crucial for TCR based diagnostics (e.g. for method selection, optimization and reproducibility). This is because the distribution of *in-vivo* TCR frequencies is long-tailed (akin to a power law; [12–14]) with many potentially disease-relevant TCRs appearing at frequencies as low as one per million [15, 16]. Thus, undetected low-frequency TCRs could potentially compromise the quality of TCR diagnostics, leading to misdiagnosis.

A pool of spike-in sequences at different frequencies allows for controlled experimentation and for the quantification of detection power. Such spike-in sequences have been previously considered in Ig sequencing [17, 18] to conduct error and bias correction. In both Ig studies, the spike-in pool contained different CDR3 sequences at different relative concentrations, thereby enabling the systematic study of sequence detection limits. Spike-in standards have also been used in TCR sequencing [11, 19]. By using synthetic DNA templates, Carlson *et al.* were able to account for amplification bias and computationally correct their sequencing library [19]. Similarly, Barennes *et al.* benchmarked different TCR sequencing methods with a single spike-in TCR clonotype, present at three different frequencies (1/10, 1/100 and 1/1000) [11]. However, unlike the Ig studies, the TCR studies [11, 19] did not consider the effects of spike-in sequence frequencies on sequence detection. Consequently, the effect of variable TCR clonal frequency on TCR sequence detection is an open question. Furthermore, previous TCR sequencing studies have not considered the crucial issue of detection reliability. That is, how can we estimate the probability of a disease-relevant TCR sequence being reliably detected by a given experimental design? By quantifying the effects of important sequencing parameters, computational models can thus provide precise detection power calculations, and thereby enable reliable TCR sequence detection for diagnostic applications.

In this study, we developed a combined experimental and computational framework to investigate the power of TCR sequencing methods to detect 45 unique spike-in TCRs across a wide range of frequencies (5×10^{-2} to 10^{-6}). We also investigated the effect of replicates (RNA and cDNA) and PCR amplification (combined TRA/TRB versus separate TRA/TRB) using a 5′RACE based protocol. We used the sequencing read counts to calibrate our computational model, which allowed us to calculate the detection power of our TCR sequencing methods. Based on our read count models, we developed a detection power calculator, TCRpower, which allows for the inference of TCR detection power as a function of TCR frequency, TCR sample count, sequencing depth and read cutoff. TCRpower can be recalibrated with pilot data from alternative sequencing methods, beyond those considered in this study, and thereby provide laboratory protocol-specific predictions of TCR detection power for future applications.

Material and methods

Human subjects

To generate RNA from effector memory CD4+ T cells for the study, we obtained blood samples from two randomly selected donors. One donor was an anonymous blood donor at the blood bank of Oslo University Hospital (OUS), from whom we obtained a buffy coat made from full blood. We obtained a blood sample from another donor via the Gastroenterology unit at Oslo University Hospital-Rikshospitalet after receiving informed written consent.

Generation of TCR dataset with spike-in TCRs

Effector memory CD4+ T cells were isolated from peripheral blood samples by using the CD4+ Effector Memory T Cell Isolation Kit (Miltenyi, Germany) followed by total RNA extraction using the RNeasy Mini Kit (Qiagen, Germany) and cleanup using the RNeasy MinElute Cleanup Kit (Qiagen, Germany). In order to generate a panel of diverse spike-in TCRs, we selected 45 T-cell clones (TCCs) with unique known TCRs (Supplementary Table S1) and isolated total RNA using the RNeasy Mini Kit (Qiagen, Germany). The RNA from these 45 TCCs were mixed in titrated amounts, with nine different concentrations (0.001, 0.003, 0.01, 0.05, 0.3, 1, 3, 10 and 50 ng) containing 5 TCC each. This spike-in RNA mix (~320 ng) was combined

with RNA from CD4 Effector memory T cells (~680 ng), to generate a final RNA mix (~1000 ng) designed to mimic the broad range of biological TCR frequencies found in *in-vivo* TCR repertoires (Supplementary Table S1). Consequently, this final RNA mix contained RNA from 45 TCC with known TRA and TRB sequences present in nine different frequencies (1, 3, 10, 50, 300, 1000, 3000, 10 000 and 50 000 RNA per one million RNA molecules) where RNA from five TCC were present in each of these frequencies.

We prepared sequencing libraries from the final RNA mix under different conditions (Figure 1). In Set 1, the final RNA mix was split into three replicas prior to cDNA synthesis, whereas in Sets 2 and 3 the cDNA sample was split into six/three replicas prior to PCR amplification. In Set 2, the PCR amplification for TRA and TRB sequences were performed as separate reactions, whereas in Set 3 it was performed as one reaction. As controls, we also performed TCR sequencing on the RNA from spike-in RNA mix only (Control spike-in TCC mix) and RNA of the effector memory CD4 T cells only (Control CD4 TEM). All of these sets were generated in duplicates (a, b) with the only difference being the use of two slightly different Template-switch oligo in set a (TSO_a) and set b (TSO_b). The sequences of the oligos and primers used in cDNA synthesis and the PCR reactions are provided in Supplementary Table S2.

The RNA was reverse transcribed to generate cDNA in two steps using a protocol based on 5′ RACE [16, 20]. In the first step, RNA was mixed with 10 mM Tris–HCl pH 8 (Sigma Aldrich, USA), 0.2% Tween-20 (Sigma Aldrich, USA), 1 mM of deoxynucleotide (dNTP) (ThermoFisher Scientific, USA), 1 μM of oligo dT (Biomers.net, Germany), 1 U/μl RNase Inhibitor (New England Biolabs, USA) in a total reaction of 24.75 μl and subjected to 72°C for 3 min followed by 1 min on ice. In the second step, 1X FS buffer (Invitrogen, USA), 0.8 M Betaine (Sigma Aldrich, USA), 6 mM MgCl2 (Sigma Aldrich, USA), 2.5 mM DTT (Invitrogen, USA), 2 μM TSO_a (IBA Lifesciences, Germany) or 2 μM TSO_b (Biomers.net, Germany), 1.5 U/μl RNase Inhibitor (New England Biolabs, USA), 5 U/μl SuperScript II (Invitrogen, USA) were added in a total volume of 25.25 μl and the cDNA was synthesized at 42°C for 90 min followed by 72°C for 15 min.

Following cDNA synthesis, three rounds of PCR were carried out. In the first PCR, cDNA from each sample was divided into replicates for amplification of TRA and TRB genes (Figure 1). The first PCR was performed with 200/40 nM forward primer mix (STRT-fwd S/L; Biomers.net, Germany), 200 nM reverse primer (TRAC_rev1 or TRBC_rev1; Biomers.net, Germany), 200 μM of dNTP (ThermoFisher Scientific, USA) and Phusion High-Fidelity DNA Polymerase (ThermoFisher Scientific, USA), in total volume of 20 ul. The cycling conditions were: 1 min at 98°C followed by 5 cycles (10 s × 98°C, 60 s × 72°C), 5 cycles (10 s × 98°C, 30 s × 70°C, 40 s × 72°C), 8 cycles (10 s × 98°C, 30 s × 65°C, 40 s × 72°C) and a final elongation at 72°C for 4 min. The second PCR was performed with 200 nM indexed forward primers (R2_In; Biomers.net, Germany), 200 nM barcoded reverse primers (TRA_In or TRB_In; Biomers.net, Germany), 200 μM of dNTP (ThermoFisher Scientific, USA) and Phusion High-Fidelity DNA Polymerase (ThermoFisher Scientific, USA) in total volume of 10 ul. The primers used for barcoding different sets and replicates are provided in Supplementary Table S3. The cycling conditions were: 2 min at 98°C followed by 10 cycles (20 s × 98°C, 30 s × 60°C, 40 s × 72°C) with final elongation at 72°C for 5 min. A final PCR reaction was carried out with 200 nM forward primer (Illumina Seq Primer R2; Biomers.net, Germany), 200 nM reverse primer (Illumina Seq Primer R1; Biomers.net, Germany) and KAPA HiFi HotStart ReadyMix (Roche, South Africa) in a total reaction of 10 μl to prepare the sequencing library for the Illumina MiSeq platform. The cycling conditions were: 2 min at 95°C followed by 20 cycles (20 s × 98°C, 30 s × 60°C, 40 s × 72°C) with final elongation at 72°C for 5 min. The PCR products were pooled and cleaned using the Monarch PCR & DNA Cleanup Kit (New England Biolabs, USA) followed by gel extraction. The PCR product excised from the gel was cleaned with the Monarch DNA Gel Extraction Kit (New England Biolabs, USA) and the Monarch PCR & DNA Cleanup Kit (New England Biolabs, USA). The resulting amplicon library was sequenced using the HiSeq 3000 platform at the Norwegian Sequencing Centre, a core facility at the University of Oslo and Oslo University Hospital. The raw sequencing data have been deposited in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under the accession number PRJNA760684.

Data processing and software

MiXCR [21] was used to process the raw TCR sequences obtained from Illumina sequencing to obtain quantitated clonotypes. The nucleotide CDR3s of the MiXCR output (i.e. the clones list) were searched for the nucleotide CDR3s of the spike-in sequences. To convert the MiXCR-formatted CDR3s to IMGT format (used by the spike-in TCR sequences), three nucleotides were trimmed off the 5′ and 3′ ends of the CDR3s. Identical converted nucleotide CDR3s were assumed to signify identical TCRs; other information such as V-gene usage was not utilized. For each set, the two duplicates (a, b) were merged for downstream analysis. Python 3 with Jupyter [22], and the packages numpy [23], scipy [24] and statsmodels [25] were used for calculations, along with TCRpower, our custom built TCR detection power calculator. Data visualizations were created with the packages seaborn [26] and matplotlib [27]. Biorender was used to create Figure 1. TCRpower is publicly available via GitHub repository (https://github.com/GabrielBalabanResearch/TCRpower) and Zenodo (https://doi.org/10.5281/zenodo.5638319).

Results

We developed a computational and experimental framework for quantifying the statistical power of nucleic acid

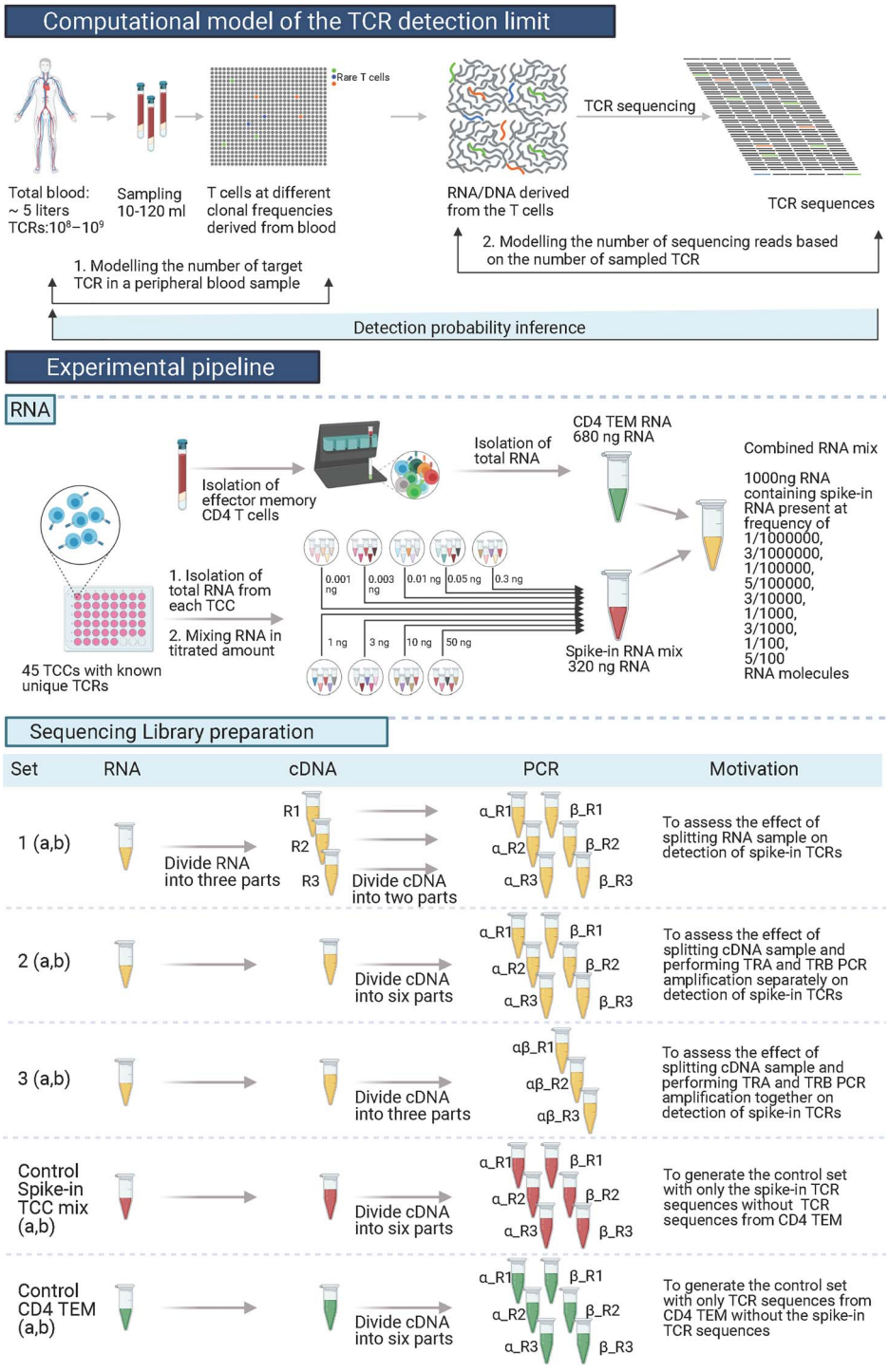

Figure 1. Study design. Our study presents a detection power calculator based on a computational model of TCR RNA read count in bulk sequencing data to enable efficient TCR sampling and RNA sequencing. Our model has two components [1] modeling the number of target TCR in a peripheral blood sample and [2] modeling the number of target TCR RNA sequencing reads based on the number of sampled target TCR. To calibrate our model, we mixed RNA from T cells with known (spike-in) concentration, together with RNA from CD4 effector memory T cells with unknown TCRs, and sequenced TCR from these mixtures using a 5′ RACE protocol. To investigate how library preparation choices affect detection power, we created three sequencing sets with different library preparation approaches. As controls, we performed TCR sequencing on the spike-in RNA mix only (Control spike-in TCC mix) and RNA of the effector memory CD4 T cells only (Control CD4 TEM). Created with Biorender.com.

sequencing methods to detect a target TCR in a peripheral blood sample. This framework is summarized in Figure 1, and includes a TCR detection power calculator and an experimental procedure to generate spike-in TCR calibration data.

A computational framework for quantifying the sequencing read count of a target TCR in a peripheral blood sample

As part of our statistical power calculator, we developed a TCR sequencing read count model. This model contains

two components: [1] modeling the number of T-cells with the target TCR that are sampled from the body [2] modeling the number of TCR sequencing reads obtained from the blood sample.

Model Component 1: The number of T-cells with the target receptor present in a peripheral blood sample. In our first model component, we account for the effects of blood sampling. In practical diagnostic scenarios, only a small portion of a patient's blood will be sampled. Thus, the TCR in this blood sample represents a subsampling of a patient's total circulating TCR population. We assume that this subsampling process is random, and model the number of target TCR sampled from the patient, C_{samp}, with a Poisson distribution

$$C_{\text{samp}} \sim \textbf{Poisson}\left(f_{\text{body}} T_{\text{samp}}\right). \tag{1}$$

Here f_{body} is the frequency of the target TCR in the patient's body, and T_{samp} the total number of sampled TCR. The expected number of target TCR in the blood sample is therefore $f_{\text{body}} T_{\text{samp}}$, which is also the rate parameter of the Poisson distribution.

Model Component 2: The number of target TCR sequencing reads obtained from the peripheral blood sample. In our second model component, we consider C_{read}, the sequencing read count of the target TCR in the blood sample. In particular, we model C_{read} with a negative binomial distribution

$$C_{\text{read}} \sim \textbf{negbin}\left(\mu, \sigma^2\right). \tag{2}$$

Here μ is the mean read count, and σ^2 the read variance. The negative binomial distribution allows for the variance of C_{read} to be greater than that expected by random subsampling, thereby taking into account technical factors associated with library preparation and sequencing, which can influence the read count (e.g. differences in primer binding and PCR amplification rates). We further parameterize the negative binomial mean and variance parameters to allow for flexible models that can account for the effects of various laboratory protocols and TCR sample frequencies

$$\mu = f_{\text{samp}} r_e T_{\text{read}}, \quad \sigma^2 = \mu + \eta \mu^{\lambda}. \tag{3}$$

Here f_{samp} is the frequency of the target TCR sequence within the sample, and T_{read} the total number of sequencing reads. The expected value μ of C_{read} is related to $T_{\text{read}}, f_{\text{samp}}$, and a sequencing method dependent read efficiency $r_e \in [0, 1]$. The variance σ^2 of C_{read} is controlled by the scaling parameters η, λ. If $\eta = 0$ then $\sigma^2 = \mu$ and the variance of C_{read} corresponds to a perfectly even subsampling (i.e. Poisson distribution). If $\eta > 0$, then the variance of C_{read} is increased beyond random subsampling, with the parameters η, λ controlling the shape of the mean–variance relationship.

Combined two-step model: We combine Components 1 and 2 to model the probability of observing C_{read} sequencing reads of a target TCR in a blood sample, whose frequency in the body is f_{body}. The joint probability of C_{read} and C_{samp} can then be written as

$$P\left(C_{\text{read}}, C_{\text{samp}}\right) = P_2\left(C_{\text{read}} | C_{\text{samp}}\right) P_1\left(C_{\text{samp}}\right), \tag{4}$$

where the probability P_1 is calculated by the Poisson distribution [1], and the probability P_2 is calculated from the negative binomial distribution [2]. Further details regarding the calculation of the probabilities P_1, P_2 are given in Appendix 1. In general, we do not know the value of C_{samp}, and we therefore marginalize over this variable to get the probability of obtaining a particular read count C_{read}, without knowledge of C_{samp}

$$P\left(C_{\text{read}}\right) = \sum_{c_{\text{samp}}=1}^{T_{\text{samp}}} P_2\left(C_{\text{read}} | C_{\text{samp}} = c_{\text{samp}}\right) \times P_1\left(C_{\text{samp}} = c_{\text{samp}}\right). \tag{5}$$

To fully specify the model, we need to provide the values $f_{\text{body}}, T_{\text{samp}}$ for the probability P_1, given by Equation (1), and the values $T_{\text{read}}, r_e, \eta, \lambda, f_{\text{samp}}$ for probability P_2 given by Equation (2). However, once the value C_{samp} is specified, we can deduce the value of f_{samp} by $f_{\text{samp}} = \frac{C_{\text{samp}}}{T_{\text{samp}}}$, which means that we can eliminate f_{samp} in the combined model. This gives us the fully parameterized formula for the probability of obtaining C_{read} target TCR sequencing reads in a blood sample,

$$P\left(C_{\text{read}}\right) = \sum_{c_{\text{samp}}=1}^{T_{\text{samp}}} P_2(C_{\text{read}} | C_{\text{samp}} = c_{\text{samp}}, T_{\text{read}}, r_e, \eta, \lambda, T_{\text{samp}}) P_1(C_{\text{samp}} = c_{\text{samp}} | f_{\text{blood}}, T_{\text{samp}}). \tag{6}$$

We assume that the parameters r_e, η, are specific to the library preparation and sequencing method, and therefore estimate them via maximum-likelihood using pilot read count data of spike-in sequences with known sample frequencies f_{samp}. We note that P_1 does not involve any sequencing method specific parameters, so that the maximum likelihood estimation of r_e, η, need only consider P_2. Further details regarding this maximum likelihood problem are given in Appendix 1. Once the parameters r_e, η, are estimated, we can use the combined Equation (6) as the basis for a TCR detection power calculator.

TCR detection power calculator

Our TCR detection power calculator is based on the read count model [6], while also accounting for read thresholds. Read thresholds are often used in TCR sequencing scenarios to reduce the chance of falsely detected sequences (i.e. false positives), which may be caused by sequencing errors [9]. Setting the read threshold at c_{thresh},

the probability of detecting a TCC by receptor sequencing a blood sample is then

$$P\left(C_{\text{read}} > c_{\text{thresh}}\right) = 1 - \sum_{i=0}^{c_{\text{thresh}}} P\left(C_{\text{read}} = i\right)$$

$$= 1 - \sum_{i=0}^{c_{\text{thresh}}} \sum_{c_{\text{samp}}=1}^{T_{\text{samp}}} P_2(C_{\text{read}} = i | C_{\text{samp}} = c_{\text{samp}}, T_{\text{read}}, r_e, \eta, \lambda, T_{\text{samp}}) P_1(C_{\text{samp}} = c_{\text{samp}} | f_{\text{body}}, T_{\text{samp}}) \quad (7)$$

where the second equation comes from the combined model [6]. The detection power calculation [7] is implemented in our Python-based power calculator, TCRpower, which is publicly available at https://github.com/GabrielBalabanResearch/TCRpower . We note that, in practice, a term of the double sum [7] only needs to be computed when P_1 is above machine precision. This means that for efficiency, we can precompute P_1, and discard the terms below machine precision before computing [7].

In addition to carrying out power calculations, TCRpower also contains functions for estimating the parameters r_e, η, from TCR sequencing data, thereby allowing TCRpower to be calibrated using pilot sequencing data with known TCR frequencies (i.e. spike-in sequences). Once calibrated in this way, TCRpower can be used to optimize further TCR sequencing scenarios. In particular, the effects of T_{read} and T_{samp} are often of interest, as these parameters are directly related to the financial cost of the sequencing, and the patient blood sample size, respectively. If we know f_{body}, and are interested in obtaining a certain target TCR detection probability, we can evaluate [7] directly. Alternatively, we can obtain the minimal TCR frequency that can be detected with a given confidence level α. In this case, we solve Equation (7) numerically for f_{body}, with the left hand side equal to α.

Accuracy and variability of spike-in TCR frequency measurements for the combined spike-in CD4-TEM experiments

We analyzed the read counts and relative read frequencies of the TRA and TRB sequences of the spike-in TCRs for the sequencing Sets 1–3, where we used the combined RNA mix. Of the 45 spike-in TCRs, 8 contained a TRA or TRB sequence that was undetected in all experimental sets (Supplementary Figure S1). These undetected TRA/TRB sequences were potentially lost during sample preparation, and their corresponding TCR were therefore removed from downstream analysis, leaving 37 spike-in TCRs in the analysis. For the remaining spike-in TCRs, we noted a linear relationship (Figure 2A) between the ground truth and measured TCR frequencies in all three Sets 1–3, with high coefficients of determination ($R^2 = 0.86$, 0.9, 0.92 for TRA Sets 1–3; $R^2 = 0.92$, 0.93, 0.92 for TRB Sets 1–3). This indicated a consistent linear relationship between the input TCR spike-in amount and output read count of our 5′ RACE library preparation and sequencing methods, for the 37 consistently detected TCR. This relationship is reflected in the TCRpower model by the linear relation between the read count, C_{read} and the TCR sample frequency, f_{samp} in Equation (3).

We quantified the variability in measured spike-in TCR frequency using the index of dispersion (std/mean). This allowed us to make a relative comparison of measured TCR frequency variability across our entire range of experimental TCR frequencies (Figure 2B). For both TRA and TRB, Sets 2 and 3 tended to have lower dispersion for the more frequent TCR (≥300 per million RNA) as compared with Set 1. We note that the index of dispersion tended to decrease with increasing spike-in frequency up to around 300 per million RNA before flattening out (Figure 2B). This indicated that it was relatively more difficult to accurately measure the frequency of the lower frequency TCR (< 300 per million RNA). This phenomenon may be partially explained by PCR chemistry and the central limit theorem of statistics. With increasing TCR RNA input, the random PCR doubling was most likely averaged over more input molecules, leading to the observed lower relative read count variability among the high frequency TCR.

Model calibration results and detection limit estimation for the combined spike-in CD4-TEM experiments

We sought to determine the detection limit of rare TCRs for our experimental Sets 1–3, in order to directly compare the efficacy of the underlying sequencing library preparation methods. To accomplish this, we estimated the minimal TCR frequency that could be detected with a standard 95% probability (f_{samp95}) for each experimental set and receptor type (TRA and TRB), and assuming a normalized sequencing depth of $T_{\text{read}} = 10^6$ reads. The calculation of each f_{samp95} value was performed using the negative binomial Model Component 2, calibrated separately to each experiment Set and TRA/TRB combination (Figure 3A). Figure 3A shows the 95% prediction interval of the calibrated TCR read count models, as compared with the measured read counts. These results show a good model to data match. In particular, our models were able to account for the experimentally observed, spike-in frequency dependent read count variability (Figure 2B). This variability is taken into account by our model derived detection limits f_{samp95}.

In Figure 3B we visualized the model derived detection limits, f_{samp95}, for Sets 1–3. We found that f_{samp95} for TRB was very similar among the sets (Set 1: $f_{\text{samp95}} = 2.08 \times 10^{-3}$, 95% CI = [2.01–2.14] $\times 10^{-3}$, Set 2: $f_{\text{samp95}} = 1.25 \times 10^{-3}$, 95% CI = [1.20–1.29] $\times 10^{-3}$ and Set 3: $f_{\text{samp95}} = 1.42 \times 10^{-3}$, 95% CI = [1.38–1.46] $\times 10^{-3}$). However, f_{samp95}

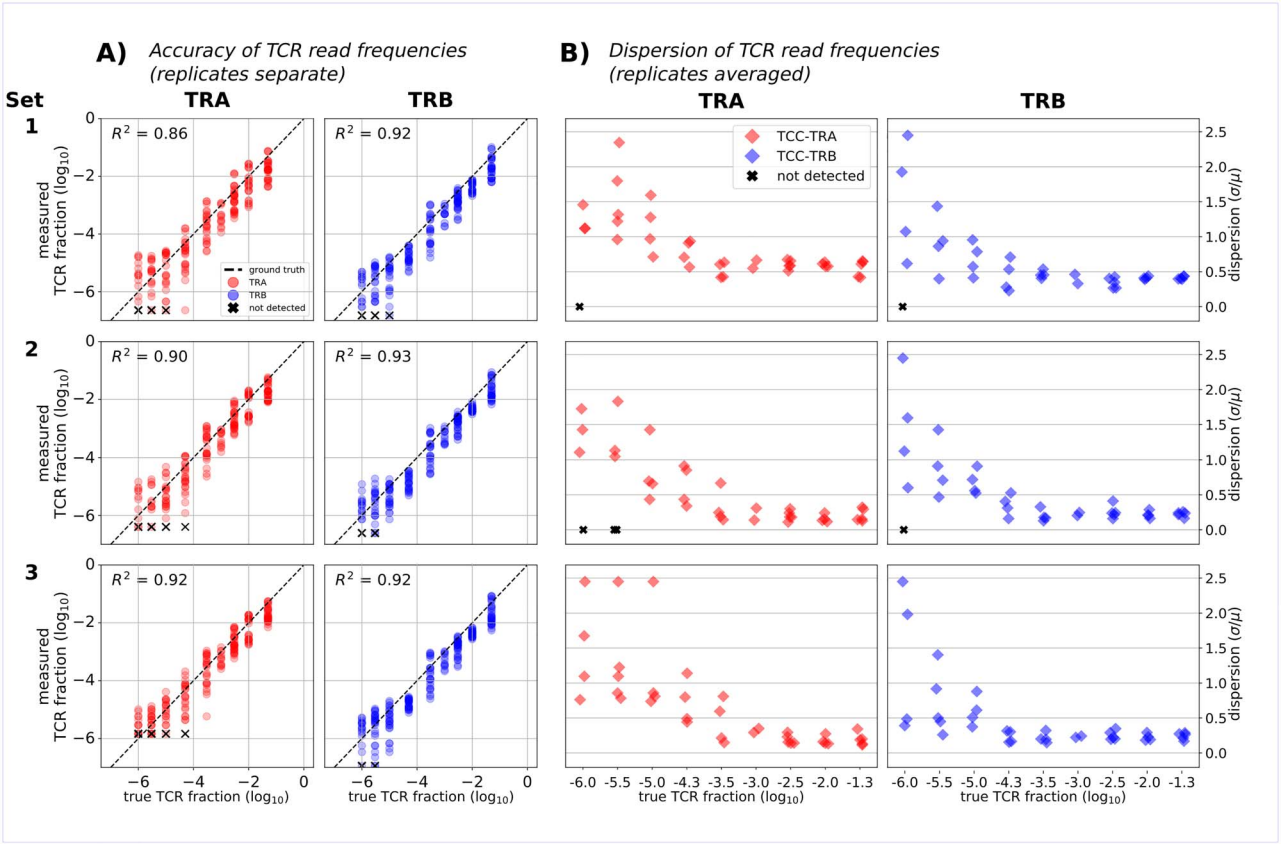

Figure 2. Accuracy and variability in TCR frequency measurement. (**A**) Ground truth versus measured TCR frequency of the spike-in TCR for experimental Sets 1–3 for all 6 replicates, showing consistent linear relationships. (**B**) TCR frequency dispersion index (std divided by mean) across 6 replicates of each TCR. The lower frequency TCRs have higher dispersion (R^2) than the higher frequency TCRs. Some low-frequency TCRs are undetected (marked by X) either for a certain replicate (panel A) or across all six replicates (panel B).

for TRA varied substantially more with the experimental setup. More specifically, f_{samp95} was lowest in Set 3 (f_{samp95} =1.23 × 10^{-3}, 95% CI = [1.19, 1.28] × 10^{-3}) and lower in Set 2 than in Set 1 (Set 2: f_{samp95} = 5.87 × 10^{-3}, 95% CI = [5.70, 6.05] × 10^{-3}, Set 1: f_{samp95} = 9.85 × 10^{-3}, 95% CI [9.51, 10.02] × 10^{-3}). We note the narrow size of the f_{samp95} confidence intervals, which were too small to be visualized in Figure 3, indicating a high model confidence in the detection limit values. Further details regarding the calculation of f_{samp95} and the corresponding confidence intervals are given in Appendix 2.

Taken together, these results indicate that our computational framework was able to account for our varying experimental conditions, and provide good model-data fits. With the calibrated models, we were then able to precisely estimate the detection limit of rare TCR (TRA and TRB sequences) for a given read count. In particular, we were able to detect many clonotypes down to a frequency of 10^{-6} and consistently detect clonotypes with frequency ≥ 10^{-4}. Our results also suggest that low-frequency TRB sequences have higher detection probabilities than low-frequency TRA sequences and combining amplification of TRA and TRB sequences improves the detection probabilities of TRA sequences.

Implementation of a read cutoff eliminated most false positive sequences

We investigated the potential for false positives results in our TCR sequencing, by examining the sequences in the Control spike-in TCC mix set, where we expected to find only the TCR sequences of the spike-in TCCs. All sequencing reads that did not match the TCR sequences of the spike-in TCCs were thus regarded as false positive sequences for the Control spike-in TCC mix set. We found that the majority of the sequences in the Control spike-in TCC mix set matched the spike-in TCRs, with a substantially lower false positive rate for TRB sequences (4.2%) as compared to TRA sequences (35.2%). We found that the relatively high error rate in TRA sequences was driven by three outlier TRAs with very high read counts, whose cause we were unable to identify. Upon removal of these outliers, the false positive rate for TRA sequences was reduced to 8.7%.

We categorized the false positive sequences into two groups, based on if they could be found in the Sets 1–3 of the sequencing library that contained TCR sequences from the CD4 TEM cells (Figure 4A). Based on this data, we noticed that a read cutoff of 18 reads could remove the majority of false positive sequences (TRA 99.0% and TRB 92.4%), including all false positive sequences that

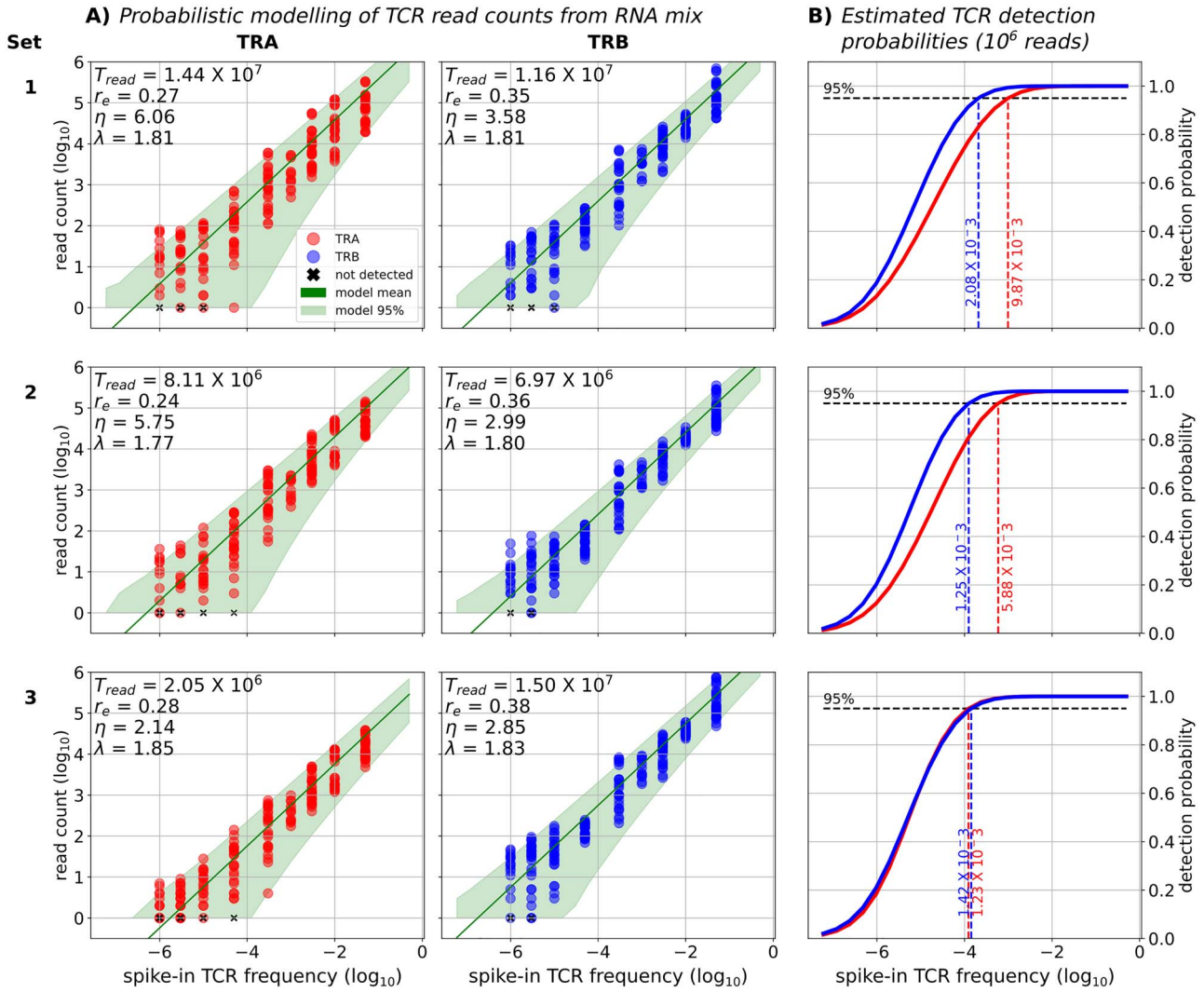

Figure 3. Negative binomial modeling of the spike-in TCR read counts and detection limit estimation. (**A**) TRA (red) and TRB (blue) read count versus spike-in TCR frequency under three experimental conditions (Sets 1–3). The dots represent measured read counts, whereas the green line and areas are the respective mean and 95% prediction interval of negative binomial models with read efficiency parameter r_e and mean-variance relationship parameters η, λ, fitted by maximum likelihood. T_{read} = the total TRA or TRB read count for the set. (**B**) Estimated detection probability (read count > 0) as a function of TCR frequency, along with the minimal fraction (dashed line) that can be detected with at least 95% probability. Note that for TRA, Set 3 stands out with the lowest η value (i.e. variance) and 95% detection probability.

were not found in the rest of the library, and the majority of the false positives that were also found in other sets (Figure 4A).

For the high read count sequences present in both control spike-in TCC mix and other sets (Figure 4B), we observed a linear trend, where the false positive sequences were present with 1–2 logs lower read counts than in the other sets. The only exception was a small cluster of four TRB sequences that did not follow the linear trend, as these false sequences had much substantially lower read counts in the Control spike-in TCC mix set as compared with the CD4 TEM Sets 1–3. Since we employed a single unique barcoding strategy, the observed trend is very likely an effect of the index-hopping phenomenon observed in the HiSeq 3000/4000 platforms [28–30]. Taken together, our observations indicate that implementing a read cutoff could potentially remove all false positive sequences not associated with index-hopping.

Example power calculations for TCR detection in patients, to optimize the number of sequencing reads and sampled T-cells

We used our power calculator TCRpower to perform detection power calculations for the experimental Sets 1–3. In particular, we estimated the minimum number of sampled TCR and sequencing reads required to achieve a 95% probability of detecting a target TCR with clonal frequency 10^{-4} in a patient. Based on the results of the previous section, we used an example detection read cutoff $c_{thresh} = 18$. For the model calibration parameters (r_e, η, λ), we used the previously estimated values shown in Figure 3A.

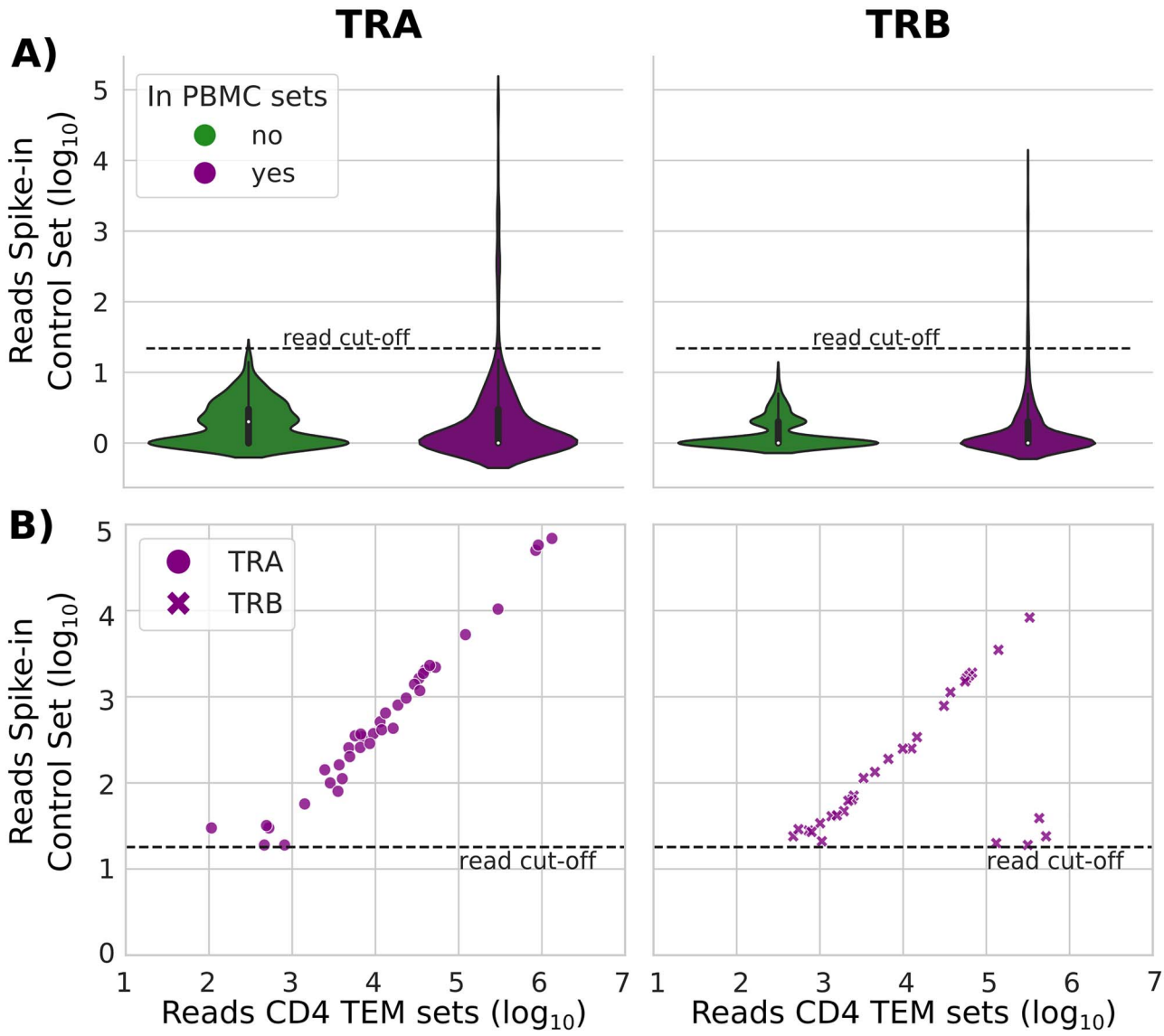

Figure 4. Falsely detected sequences in the Control Spike-in TCC set. (**A**) Count distributions of TRA and TRB reads that did not match the TCRs of the spike-in TCCs, but were nevertheless detected in the Control Spike-in TCC set. All of the falsely detected sequences were either present in other sets in the library (purple), or only exclusively found in the Control Spike-in TCC set (green). The dotted lines represent the read cutoff [18] (**B**) Read count total over all CD4 TEM containing sets versus read counts in the Control Spike-in TCC Mix for the falsely detected sequences with read count >18 in the Control Spike-in TCC Mix. Note the linear trend characteristic of the index-hopping phenomenon.

In Figure 5, we display the detection calculator results. As expected, the 95% detection regions have a rectangular shape with a rounded corner, meaning that there is a minimum number of sampled TCRs and sequencing reads needed to achieve 95% detection power. Within the rounded corners, sampled TCR and sequencing reads can be traded for one another while still maintaining the same detection power (Figure 5). For TRA, Set 3 has the best detection efficiency, requiring the least number of reads and sampled TCR to achieve 95% detection power. For TRB, Set 1 is slightly more efficient than Sets 2 or 3.

We note that the required number of sequencing reads for a 95% detection probability is several orders of magnitude more than the inverse of the desired target TCR frequency in all of our cases. This is due to the extra-Poisson variance in the number of sequencing reads that we attribute to library preparation and sequencing chemistry. We note that even in a perfect subsampling scenario (i.e. Poisson process), the number of sequencing reads would have to be substantially greater than the inverse of the target TCR frequency, to ensure a reasonable detection probability.

Discussion

We developed a combined computational and experimental pipeline for quantifying the detection power of bulk TCR sequencing based on 45 unique spike-in TCRs present at a wide range of clonal frequencies (5 per 100 to 1 per million). In particular, we demonstrated the possibility of consistently detecting TCRs with frequencies as low as 1 per million. We also observed that TRB sequences were more easily detected than TRA sequences present at the same low frequency.

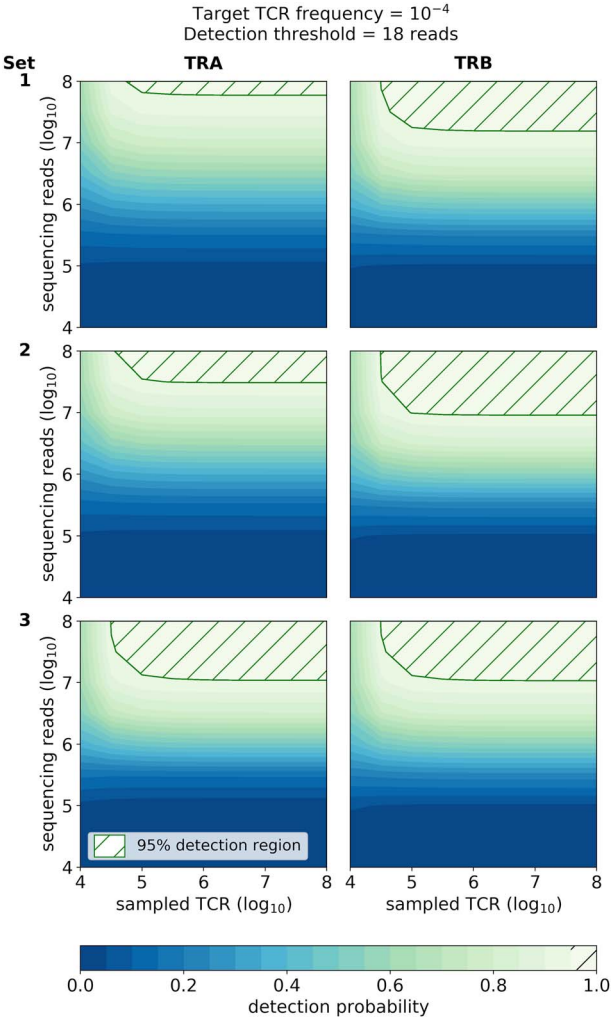

Figure 5. Detection power estimation. Example output from our power calculator TCRpower, showing the probability of detecting a TCR with frequency 10^{-4} and read count threshold 18, as a function of the number of sampled TCR and sequencing reads.

Furthermore, we developed the first computational model to estimate TCR detection power, and thereby enable reliable detection of disease-relevant TCR for diagnostic applications. Our model is implemented in a power calculator, which can be recalibrated with data from alternative TCR sequencing methods. Both the detection model and power calculator are available to the research community via the publicly available Python package TCRpower.

Importance of spike-in standards with multiple unique CDR3 sequences at variable frequencies

Spike-in standards with known sequences and frequencies provide a useful ground truth [9]. This ground truth is normally lacking in vast, diverse TCR sequencing repertoire data, derived from biological samples. Indeed, the Adaptive Immune Receptor Repertoire (AIRR) Community encourages the use of spike-in standards, to address a wide variety of technical issues (e.g. sensitivity, specificity, accuracy of clonotype quantification, reproducibility and false read removal) [9].

Our study highlights the importance of spike-in TCRs for understanding the impact of different library preparation choices on the detection of TCRs present at different frequencies. In particular, we investigated spike-in TCR (unique TRA and TRB) at a wider range of frequencies (5 per 100 to 1 per million) than had been previously considered. Previous studies using spike-in TCRs have either used synthetic DNA molecules in a multiplex PCR [19] or used a single spike-in TCR with a limited frequency range (1/10, 1/100, 1/1000) [11]. Furthermore, having a setup with unique spike-in TCRs at known concentrations allowed us to demonstrate for the first time the linearity/accuracy of clonotype frequency quantification for multiple unique TCRs. This is particularly important for TCR based diagnostics, as disease-relevant TCRs can be present in the body at a wide range of frequencies. The read counts of the spike-in TCRs enabled us to calibrate our power calculator, which we used to optimize the number of sampled TCR and sequencing depth. Furthermore, our calibrated power calculator can interpolate and extrapolate TCR detection probabilities to arbitrary TCR frequency ranges, an issue which has not been previously addressed. Finally, by sequencing the control sets with spike-in TCC mix only, we were able to examine the nature of falsely detected sequences and set an appropriate read cutoff.

Detection of rare TRA and TRB with 5′ RACE sequencing

We demonstrated that TCRs present at frequencies as low as 1 per million can be detected by a 5′ RACE sequencing method, one of the most widely used AIRR sequencing methods [9]. We observed increased replicate consistency with increased frequency of the TCR clonotypes, indicating that if the antigen-specific T cells of interest are found in relatively higher frequencies (above 50 per million in our study), RNA or cDNA replicates may be unnecessary. In our experiments, low-frequency TRB sequences were more consistently detected than low-frequency TRA sequences. More specifically, only TRAs present at frequency $\geq$ 50 per million and TRBs present at frequency $\geq$ 10 per million were detected consistently. This difference between TRB and TRA sequencing efficiency has also been described for several other TCR sequencing methods [11] and is most likely caused by a difference in transcript abundance, as TRB transcripts are two to three times more abundant than TRA transcripts [14]. Taken together, our results indicate that for abundant TCRs, both TRA and/or TRB sequencing provides reliable detection of TCR clonotypes. However, for the detection of rare clonotypes, TRB sequencing alone is sufficient and better suited, since including TRA will occupy sequencing depth without increasing detection power (Figure 3B). However, if one is also interested in TRA sequencing of a rare TCR clonotype using a 5′RACE based protocol, performing combined TRA and TRB PCR amplification could enable TRA detection without compromising the TRB detection.

TCR detection power calculator and read count model

We developed a model of the TCR detection limit, which can be calibrated with pilot TCR sequencing data with known TCR concentrations (i.e. spike-ins). This allowed our model to account for read count variability due to technical factors such as primer and PCR biases. For example, in the current study, we tested three different sample preparation setups (Figure 1), which had an effect on the measured read count variability, and thereby on the calculated detection probabilities via the estimated parameters (r_e, η, λ). Based on our read count model, we created our power calculator, TCRpower. To the best of our knowledge, TCRpower is the first power calculator specifically made for TCR sequencing.

Several power calculators have been previously developed for RNA-sequencing and RNA microarray experiments [31–35], typically focusing on detecting differentially expressed genes via log-fold changes in RNA read counts. Unlike in these gene expression scenarios, TCR sequencing has to account for a much greater diversity of read sequences, many of which may not be present in a particular individual. Consequently, the potential presence or absence of a TCR is of particular importance for TCR sequencing diagnostics, which motivates our detection power calculator. Furthermore, biological sample size issues (e.g. volume of blood or size of biopsies) affect TCR-sequencing applications to a much greater degree than in typical gene expression studies. This is because each T-cell expresses only one single receptor sampled from an enormously large TCR sequence space, which necessitates the inclusion of the number of sampled TCR in our power calculations as an important parameter.

Due to the complexity of TCR repertoires in the body, sophisticated statistical and machine learning approaches have been developed for immune status classification based on TCR sequencing [2, 36–39]. These approaches typically infer a 'negative' diagnosis from the absence of disease related TCRs, which naturally leads to questions about the detection power of the underlying TCR sequencing methods [40]. In particular, knowledge of TCR detection power could help when transferring machine learning models to data generated by a sequencing method that differs from that used for the training data. In this scenario, optimizing the number of reads and the TCR sample size with our power calculator could help to ensure that the TCRs, which infer a 'positive' diagnosis can still be detected with the new sequencing method.

In the future, our work could be extended to consider family-wise or false discovery error rates, as has been done for RNA sequencing [31, 35]. Such an extension would allow for the quantification of detection power to entire sets of TCRs, which is especially relevant when considering 'public' TCR sets that are shared across many individuals [6, 16, 41]. Finally, our read count model could also be used to generate synthetic TCR sequencing repertoires. This could be useful to assess the diagnostic power of TCR diagnostic tests based on machine learning models that analyze entire TCR sets or repertoires.

A limitation of our power calculator is that it requires ground truth data for calibration. We also assume a perfectly even subsampling of TCRs from the body, which is reflected in the TCR sampling component of the 2-step model. This assumption is reasonable for globally prevalent TCR harvested from homogeneously mixed biological samples, such as peripheral blood. However, our model may need to be modified for T cells derived from nonhomogenous mix (e.g. tissue biopsies).

False positive sequences and read count thresholds

As the field of antigen-specific TCRs used for disease monitoring and diagnosis continues to grow, it is crucial to understand the nature of falsely detected sequences (i.e. sequencing reads that do not match any RNA sequences that were present in the original biological sample - here denoted as false positives) to develop appropriate bioinformatic pipelines and robust diagnostics. When we analyzed the control spike-in TCC mix set, where we expected to find only TCR sequences of the spike-in TCCs, we found false positive sequences. Most importantly, we found that an appropriate read cutoff could eliminate the majority of the false positives, including all the false positive sequences that were not found in the rest of the library. This demonstrates the importance of implementing a read cut off for improving the specificity of TCR diagnostic tests, which has also been highlighted for Ig sequencing [42].

We note that implementing a read count threshold can potentially remove true positive TCRs of interest, as a side effect of removing the false positives. This effectively creates a trade-off, where a higher count threshold increases the specificity of TCR detection, at the cost of sensitivity. In such a scenario, it could be desirable to maintain a sufficient detection probability by optimizing the number of sequencing reads. For this reason, the read cutoff is accounted for in our power calculator, and is available as a user-specified parameter.

We found that all of our false positives with high read count (> 18) were also present in other sets that contained TCRs from effector memory CD4 T cells. Since we have employed a single unique barcoding strategy, we suspect that these false positives were a result of the index-hopping phenomenon observed in Illumina sequencers employing patterned flow cells with Exclusion amplification chemistry (HiSeqX, HiSeq3000/4000 and NovaSeq) [28–30]. The use of nonredundant double indexing has been recommended to overcome the index-hopping phenomenon [43, 44]. Although we do not provide a remedy for index-hopping in our study, we show that index-hopping can give rise to false positive sequences and should be controlled for.

Concluding remarks and recommendations

We present the first statistical power calculator for detecting the presence of a T-cell clonotype by TCR sequencing, as well as a novel experimental procedure to generate spike-in receptor sequences for model calibration. Furthermore, the results of our sequencing experiments can be used to inform future TCR sequencing experimental designs. In particular, we confirm that TCRs as rare as 1 per million can be detected with a 5′RACE based TCR sequencing method, and that TRB sequences of rare TCRs are detected more consistently than TRA sequences. This suggests that TRB sequencing is optimal for efficiently detecting rare TCR clonotypes, whereas both TRA and TRB sequencing are sufficient for detecting TCR clonotypes that are relatively abundant.

For future TCR sequencing experiments, we recommend conducting pilot experiments with spike-in TCRs to identify the needed sequencing depth and number of cells with our calculator. When this is infeasible, including a small panel of low-frequency spike-in TCRs could help quantify TCR detection power without taking up significant sequencing depth space. This is especially crucial if the TCRs of interest are present in rare cells. We also recommend sequencing a panel of spike-in TCRs in the same sequencing library as a control set, to enable the identification of a read cutoff to reduce false positives. Taken together, we conclude that multiple unique spike-in TCRs at varying frequencies can assist both experimental and computational protocol development, which can in turn improve the reliability of TCR sequencing methods. For future studies, it would be interesting to further investigate read count threshold optimization, which we have touched upon but not fully addressed. We also encourage further use of our power calculator with alternative sequencing methods and in prospective studies, to further validate or extend our methodology.

Availability

TCRpower is publicly available in the GitHub repository (https://github.com/GabrielBalabanResearch/TCRpower) and via Zenodo (https://doi.org/10.5281/zenodo.5638319).

Accession numbers

The raw sequencing data have been deposited in the Sequence Read Archive under accession number PRJNA760684.

Key Points

- A novel statistical method (TCRpower) for calculating the detection power of T-cell receptor sequencing methods.
- Experimental procedure for generating spike-in TCR sequences for model calibration.
- TCR sequencing method optimization to efficiently detect a target T-cell clone in a patient using a blood sample.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgement

We thank all the donors who provided blood samples for this study. We thank staff at the Gastroenterology unit at Oslo University Hospital-Rikshospitalet for collecting blood samples. We also thank the blood bank at Oslo University Hospital for supplying blood samples. We are very grateful to M.K. Johannesen for laboratory-related assistance, and to Eric de Muinck for critical review of the manuscript.

Funding

This work was supported by Stiftelsen KG Jebsen (project SKGJ-MED-017); and the Norwegian Research Council via the ProCardio Center for Innovation (project 32481); and the IKTPLUSS project (#311341 to V.G. and G.K.S.). Funding for open access charge: Research Council of Norway IKTPLUSS project (#311341 to V.G. and G.K.S.).

References

1. Robins HS, Campregher PV, Srivastava SK, *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* 2009;**114**(19):4099–107.
2. Warren RL, Freeman JD, Zeng T, *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* 2011;**21**(5):790–7.
3. Dupic T, Marcou Q, Walczak AM, *et al.* Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput Biol* 2019;**15**(3):e1006874.
4. Manojlović LM. Photometry-based estimation of the total number of stars in the Universe. *Appl Optics* 2015;**54**(21):6589–91.
5. 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**(8):1070–8.
6. Yao Y, Zia A, Neumann RS, *et al.* T cell receptor repertoire as a potential diagnostic marker for celiac disease. *Clin Immunol* 2021;**222**:108621.
7. Ostmeyer J, Lucas E, Christley S, *et al.* Biophysicochemical motifs in T cell receptor sequences as a potential biomarker for high-grade serous ovarian carcinoma. *PLoS One* 2020;**15**(3):e0229569.
8. 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**(5):659–65.
9. Trück J, Eugster A, Barennes P, *et al.* Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *Elife* 2021;**10**:e66274.
10. Rosati E, Dowds CM, Liaskou E, *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol* 2017;**17**(1):61.
11. Barennes P, Quiniou V, Shugay M, *et al.* Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol* 2020;**39**:236–45.

12. Mora T, Walczak AM, Bialek W, *et al.* Maximum entropy models for antibody diversity. *Proc Natl Acad Sci* 2010;**107**(12):5405–10.
13. Greiff V, Miho E, Menzel U, *et al.* Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol* 2015;**36**(11):738–49.
14. Oakes T, Heather JM, Best K, *et al.* 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* 2017;**8**:1267.
15. Christophersen A, Raki M, Bergseng E, *et al.* Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. *United European Gastroenterol J* 2014;**2**(4):268–78.
16. Risnes LF, Christophersen A, Dahal-Koirala S, *et al.* Disease-driving CD4+ T cell clonotypes persist for decades in celiac disease. *J Clin Invest* 2018;**128**(6):2642–50.
17. Khan TA, Friedensohn S, Vries ARG, *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science. Advances* 2016;**2**(3):e1501371.
18. Friedensohn S, Lindner JM, Cornacchione V, *et al.* Synthetic standards combined with error and bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naïve and memory B cells. *Front Immunol* 2018;**9**.
19. Carlson CS, Emerson RO, Sherwood AM, *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* 2013;**4**(1):2680.
20. Quigley MF, Almeida JR, Price DA, *et al.* Unbiased molecular analysis of T cell receptor expression using template-switch anchored RT-PCR. *Curr Protoc Immunol* 2011; Chapter 10:Unit10.33.
21. Bolotin DA, Poslavsky S, Mitrophanov I, *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 2015;**12**(5):380–1.
22. Kluyver T, Ragan-Kelley B, Rez F, *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* IOS Press 2016;87–90.
23. Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020;**585**(7825):357–62.
24. Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**(3):261–72.
25. Seabold S, and Perktold J. Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*. 2010:92–6.
26. Waskom ML. Seaborn: statistical data visualization. *J Open Source Software* 2021;**6**(60):3021.
27. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;**9**(3):90–5.
28. Illumina I. *Effects of Index Misassignment on Multiplexing and Downstream Analysis*, Illumina 2017. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf
29. Sinha R, Stanley G, Gulati GS, *et al.* Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv* 2017;125724.
30. Yao Y, Zia A, Wyrożemski Ł, *et al.* Exploiting antigen receptor information to quantify index switching in single-cell transcriptome sequencing experiments. *PLoS One* 2018;**13**(12):e0208484.
31. Busby MA, Stewart C, Miller CA, *et al.* Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 2013;**29**(5):656–7.
32. Hart SN, Therneau TM, Zhang Y, *et al.* Calculating sample size estimates for RNA sequencing data. *J Comput Biol* 2013;**20**(12):970–8.
33. van Iterson M, van de Wiel MA, Boer JM, *et al.* General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol* 2013;**12**(4):449–67.
34. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* 2014;**20**(11):1684–96.
35. Wu H, Wang C, Wu Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* 2015;**31**(2):233–41.
36. Kanduri C, Pavlović M, Scheffer L, *et al.* Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. *bioRxiv* 2021;2021.05.23.445346.
37. Pavlović M, Scheffer L, Motwani K, *et al.* immuneML: an ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell* 2021;**3**:936–44.
38. Pertseva M, Gao B, Neumeier D, *et al.* Applications of machine and deep learning in adaptive immunity. *Annu Rev Chem Biomol Eng* 2021;**12**(1):39–62.
39. Widrich M, Schäfl B, Pavlović M, *et al.* Modern Hopfield networks and attention for immune repertoire classification. *Adva Neural Inf Process Syst* 2020;**33**:18832–45.
40. 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.
41. Dahal-Koirala S, Risnes LF, Neumann RS, *et al.* Comprehensive analysis of CDR3 sequences in gluten-specific T-cell receptors reveals a dominant R-motif and several new minor motifs. *Front Immunol* 2021;**12**:639672.
42. Greiff V, Menzel U, Haessler U, *et al.* Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol* 2014;**15**(1):1–14.
43. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 2012;**40**(1):e3.
44. Costello M, Fleharty M, Abreu J, *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 2018;**19**(1):332.
45. CFJ W. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* 1986;**14**(4):1261, 35–95.

APPENDIX 1: Probability calculations

In this appendix we give further details underlying the probability calculations used in the detection power calculator TCRpower. The probability P_1, corresponding to Model Component 1, is given by the Poisson function

$$P_1\left(C_{\text{samp}} \mid f_{\text{body}}, T_{\text{samp}}\right) = \frac{\left(f_{\text{body}} T_{\text{samp}}\right)^{C_{\text{samp}}} e^{-\left(f_{\text{body}} T_{\text{samp}}\right)}}{C_{\text{samp}}!}. \tag{8}$$

Probability P_2, corresponding to Model Component 2, is given by the negative binomial function

$$P_2\left(C_{\text{read}}, r, p\right) = \frac{\Gamma\left(C_{\text{read}} + r\right)}{\Gamma\left(C_{\text{read}} + 1\right)\Gamma(r)}(1-p)^{C_{\text{read}}} p^r, \tag{9}$$

where r, p are the standard negative binomial parameters, and Γ is the standard Gamma function. We can relate r, p to the parameters $T_{\text{read}}, r_e, \eta, \lambda, f_{\text{samp}}$ by

$$\begin{aligned} r &= \frac{\mu^2}{\sigma^2 - \mu} = \frac{\mu^{(2-\lambda)}}{\eta} = \frac{\left(f_{\text{samp}} r_e T_{\text{read}}\right)^{(2-\lambda)}}{\eta} \\ p &= \frac{\sigma^2 - \mu}{\sigma^2} = \frac{1}{1 + \eta\mu^{(\lambda-1)}} = \frac{1}{1 + \eta\left(f_{\text{samp}} r_e T_{\text{read}}\right)^{(\lambda-1)}}. \end{aligned} \tag{10}$$

Calibrating TCRpower consists of solving the following maximum likelihood problem

$$\max_{r_e, \eta, \lambda} \sum_{i=1}^{N} \log P_2\left(c^i_{\text{read}}, r_e, \eta, \lambda, f^i_{\text{samp}}\right), \tag{11}$$

where $c^i_{\text{read}}, f^i_{\text{samp}}$ are the respective sequencing read count and ground truth frequency of the ith spike-in TCR receptor, and N is the total number of spike-in TCR receptor types. The maximum likelihood estimation [11] is accomplished in TCRpower via a nested set of three likelihood problems, each of which is solved by Newton's method. First, the r_e parameter is estimated for a Poisson model with data c^i_{read} and rate parameter vector $f^i_{\text{samp}} N_{\text{reads}}$, where $N_{\text{reads}} = \sum_{i=1}^{N} c^i_{\text{read}}$. Second, the estimated r_e, along with $\eta = 0.001$ are used as an initial guess to solve the negative binomial likelihood problem [11], with the λ parameter fixed to $\lambda = 2$, giving us an updated estimate of r_e, η. Finally, the previously estimated r_e, η, along with $\lambda = 2$ are used to initialize the full negative binomial maximum likelihood problem [11]. In practice, this procedure gave us reliable estimates of r_e, η, λ without any numerical convergence difficulties.

APPENDIX 2: Detection limit and detection confidence interval calculation for experimental sets 1–3

We calculated the smallest TCR frequency (f_{samp95}) that could be detected with 95% probability for each experimental set and TCR type (TRA or TRB), assuming 10^6 reads. We calculated f_{samp95} by numerically solving the equation $P_2(C_{\text{read}} > c_{\text{thresh}} | f_{\text{samp95}}, \eta, \lambda, r_e) = 0.95$ for f_{samp95}, using the 'brentq' solver of the scipy package. In this equation, the probability P_2 comes from the negative binomial Model Component 2, calibrated to the experimental data via Equation (11).

We calculated the 95% confidence intervals of f_{samp95} with a jackknife resampling method [45]. More specifically, we created leave-one out datasets corresponding to each read count datapoint, with a different datapoint missing in each set. We re-estimated the r_e, η, parameters and recalculated f_{samp95} for each such dataset. This gave us a resampling distribution of f_{samp95} values. We then used the mean and standard deviation of the f_{samp95} resampling distribution to determine the confidence interval.