Ultra-Efficient Short Read Sequencing of Immune Receptor Repertoires

Janelle M. Montagne[1], Xuwen Alice Zheng[1], Iago Pinal-Fernandez[2,3], Jose C. Milisenda[4], Lisa Christopher-Stine[5], Thomas E. Lloyd[3], Andrew L. Mammen[2,3], & H. Benjamin Larman[1,*]

[1]Division of Immunology, Pathology Department, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

[2]Muscle Disease Unit, Laboratory of Muscle Stem Cells and Gene Regulations, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD, USA.

[3]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

[4]Internal Medicine Department, Hospital Clinic, Universitat de Barcelona, Barcelona, Spain and Centro de Investigación Médica en Red Enfermedades Raras.

[5]Division of Rheumatology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

Abstract:

Immune receptor repertoire (IRR) sequencing is increasingly employed to characterize adaptive immune responses. However, current IRR sequencing methodologies are complex, expensive, or both, thereby limiting routine utilization. Here we present Framework Region 3 AmplifiKation sequencing ("FR3AK-seq"), a simplified multiplex PCR-based approach for the ultra-efficient analysis of IRRs. By using minimal primer sets that target a conserved region adjacent to the hypervariable VDJ sequence, undistorted amplicon can be analyzed via short read, single-end sequencing. We find that FR3AK-seq is sensitive and quantitative, with a per sample cost of ~50-fold below the current industry standard. Inference of V-allele usage from FR3AK-seq data is demonstrated using a novel algorithm: Inferring Sequences via Efficiency Projection and Primer Incorporation ("ISEPPI"). FR3AK-seq and ISEPPI were utilized to quickly and inexpensively characterize the T cell infiltrates of 146 muscle biopsies obtained from patients with idiopathic inflammatory myopathies (IIMs) and controls. A cluster of related T cell receptors were identified in samples from patients with sporadic inclusion body myositis, suggesting the presence of an

unknown shared antigen. The ease and cost of FR3AK-seq analysis removes the current barriers to routine, large-scale IRR analyses.

Introduction:

Immune receptor repertoire (IRR) analysis has emerged as a powerful tool for examining adaptive immune responses. IRRs, generated by the process of V(D)J recombination, encompass the T and B cell clones within a given individual, which can be classified by phenotype and tissue occupancy, and evolve over time.[1,2] Development of next generation sequencing (NGS)-based technologies within the last decade has enabled the unprecedented analysis of IRRs.[3–6]

T cell receptor (TCR) recognition requires antigen processing and presentation of peptide epitopes on autologous major histocompatibility complex (MHC) molecules. Analysis of TCR repertoires both within and between individuals can be indicative of disease status and antigen specificity.[7–13] Clonally expanded or tissue-infiltrating T cell clones can also reveal the nature of local immune responses; integrating TCR repertoire analyses with phenotypic measurements (e.g. single cell transcriptional profiling[14] or flow cytometric sorting of T cell subpopulations[15,16]) can further illuminate the nature of T cell responses (**Fig 1a**).

A variety of methods have been used to generate TCR libraries for NGS, including multiplex PCR, 5'-RACE, and target enrichment.[3–6,17,18] Generally, these technologies prioritize sequencing of the hypervariable complementarity determining region 3 (CDR3) of the TCR beta (TCRB) chain, which harbors the greatest amount of sequence diversity, primarily confers the antigen specificity to a T cell, and can be used to uniquely determine T cell clonal identity.[19–21] Despite their utility, however, current approaches remain complex, expensive, or both. These limitations have greatly hindered the utilization of TCR repertoire-based analyses, especially for studies involving large

numbers of samples. We have therefore developed a quantitative, multiplex PCR-based approach that prioritizes the ultra-efficient analysis of CDR3 sequences of the TCRB chain. Additionally, our strategy can readily be adapted to B cell receptor (BCR) analysis and to the characterization of immune receptor repertoires of any organism with sequenced V-genes.

The sequence diversity of the TCRB chain variable domain, which is encoded by ~128 unique TCRBV alleles, necessitates the use of complex primer pools for multiplex PCR amplification. Primer competition and differential amplification efficiencies distort clonal amplicon abundance, confounding quantification using high-throughput DNA sequencing.[22] Sophisticated techniques have been developed to computationally correct amplification bias, such as the use of an 'artificial immune system' or the incorporation of molecular barcodes.[23,24] Adaptive Biotechnologies has developed the former approach as a commercial product, the immunoSEQ Assay. At a charge of $500-$1000 per sample, however, the scale of affordable studies has been greatly constrained. Molecular barcoding requires multi-day protocols, complicated analytical pipelines, and deeper sequencing to obtain sufficient coverage of the distorted libraries.[18,25] By designing maximally compact primer sets and a streamlined workflow, we have essentially eliminated PCR amplification bias and dramatically reduced assay cost, enabling CDR3 analysis via single-end, short-read (100 nt) sequencing.

To benchmark the performance of our human TCRB FR3AK-seq assay, we directly compared the results against the industry standard for quantitative TCR repertoire sequencing, the hsTCRB immunoSEQ assay. We found that our sequencing data, which was analyzed using open-source software, was in excellent agreement with the hsTCRB immunoSEQ assay. Finally, we harnessed the cost effectiveness of TCRB FR3AK-seq to identify shared T cell responses within biopsied muscle tissues from a cohort of 146 patients with idiopathic inflammatory myopathies (IIMs) and controls.

Results:

Minimal primer sets targeting the TCRB framework region 3 (FR3) efficiently amplify CDR3 sequences

We hypothesized that minimizing the number of potentially competing V-gene primers, while also minimizing amplicon length, would increase the fidelity of multiplexed PCR amplification of the TCRB CDR3 region. To identify candidate primer binding sequences upstream and proximal to CDR3, we constructed a phylogenetic tree of all TCRB FR3 regions available in the Immunogenetics (IMGT) database.[26] This revealed a high degree of sequence conservation within the 3' terminus of FR3 (**Fig 1b**). This homology likely reflects a critical role for these sequences in the optimal presentation of the CDR3 loop. The 3' 20 nucleotides of all IMGT FR3s were therefore extracted and used as target primer binding sequences. Among the 128 functional TCRBV alleles reported, there are only 47 distinct 3'-terminal 20-mers. Beyond their sequence homology, additional advantages of targeting these sequences include minimal CDR3 amplicon length (and thus optimal amplification efficiency) and the ability to sequence using short (<100 nt) single end reads initiated proximal and upstream of CDR3. We refer to this approach to IRR CDR3 analysis as "FR3AK-seq" (FR3 AmplifiKation sequencing).

We developed a heuristic algorithm to automate the design of three candidate sets of primers, each targeting the 47 distinct FR3 3' 20-mers (**Fig S1**): one primer set lacks any universal bases or mismatches ("Full" set), one primer set lacks universal bases but allows a single mismatch to occur ("1MM" set), and one primer set contains parsimonious incorporation of up to 3 universal (inosine) bases and allows one mismatch ("Compact" set). No mismatches were permitted within the five nucleotides of the 3' end of the primers, as a precaution to minimize interference with polymerase extension.[27–29] The Full set consists of 47 primers (equal to the number of unique FR3 3' 20-mers), the 1MM set consists of 34 primers, and the Compact set consists of 20 primers (**Table S1**). We calculated the expected efficiency for each primer to amplify each FR3 sequence;

these efficiencies are visualized in the form of a clustered heatmap (**Fig 1c**) and are used later for inference of V-gene utilization.

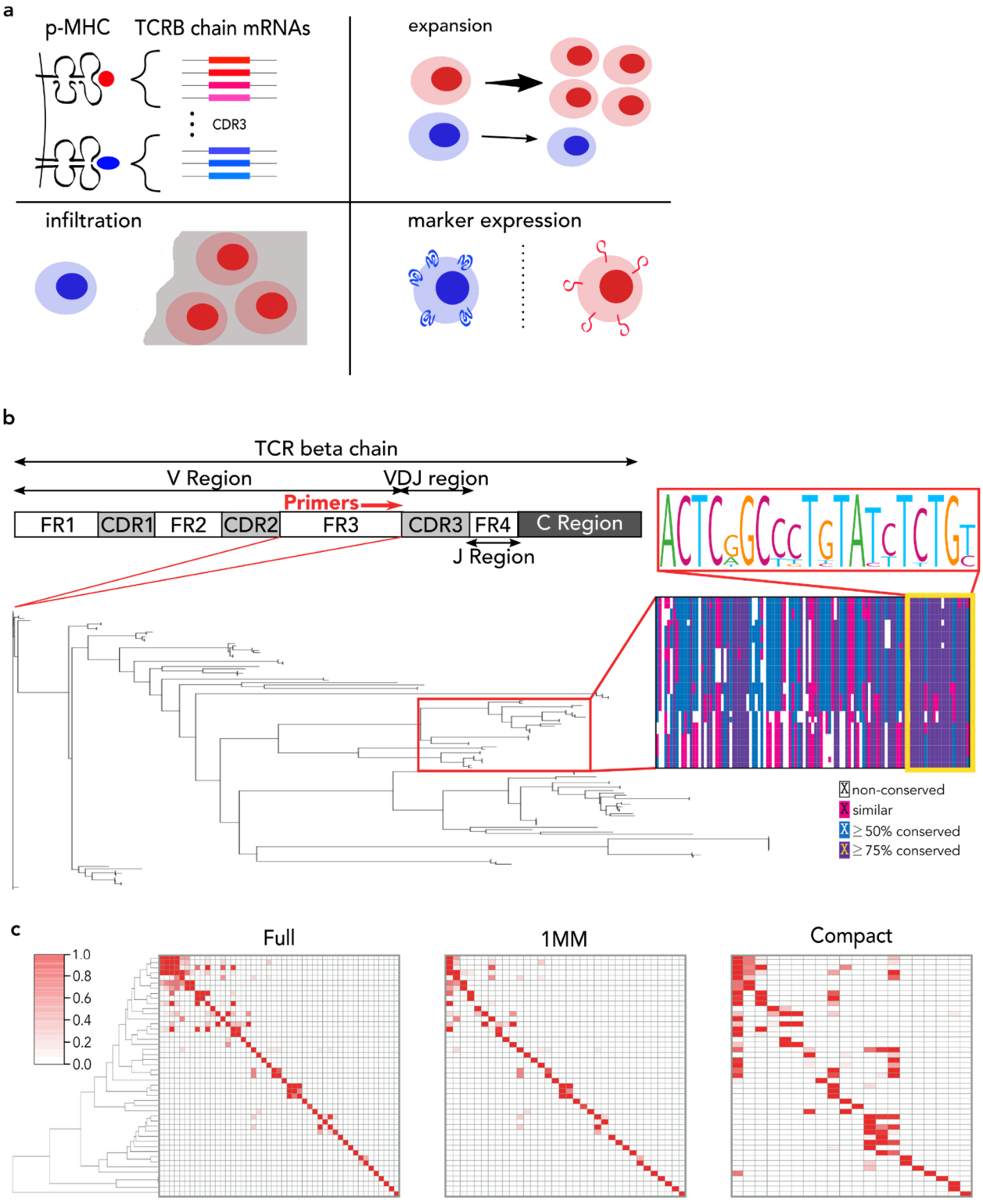

Figure 1: Utility of T cell receptor beta (TCRB) chain sequencing and FR3AK-seq multiplex PCR primer design. **a.** TCRB chain sequencing can be used to (i) identify related CDR3s that may share antigen specificity, (ii) detect expanded clones, (iii) identify tissue infiltrating clones, and (iv) link CDR3 sequences with T cell phenotypes. **b**. A schematic of the TCRB chain. An UPGMA (Unweighted Pair Group Method with Arithmetic Mean) phylogenetic tree of FR3s extracted from the 128 unique human TCRBV alleles in the IMGT database. Multiple sequence alignment from a dominant FR3 sequence cluster (red box on tree) identifies homology within the 3' ~20 nucleotides (yellow box). **c.** Clustered heatmaps showing primer amplification efficiencies for each primer from each set for each of the 47 unique 3' 20 nucleotides of the human TCRBV FR3 region. Inosines were considered exact matches for the Compact set. Y axis: 47 unique TCRBV FR3s. X axis: primers. The order of the primers in each heatmap can be found in **Table S1**. The order of unique TCRBV FR3s is identical to that of the full primer set.

TCRB FR3 primers amplify CDR3 sequences with negligible bias

Each primer was tested separately and within its set for the ability to produce amplicon from peripheral blood mononuclear cell (PBMC) cDNA. Agarose gel electrophoresis revealed a PCR product at the expected size of ~220 base pairs (**Fig 2a**). Importantly, incorporation of the inosine base did not prevent amplification by the KAPA Fast HotStart *Taq* DNA Polymerase. While lower PCR primer annealing temperature might increase off-target priming, it may also reduce unwanted bias in amplification from primers containing a single nucleotide mismatch.[30] We therefore utilized gradient PCR to assess the effect of annealing temperature on the number of unique CDR3 sequences detected using the Compact primer set. As expected, more unique CDR3 clones were recovered at lower annealing temperatures (**Fig 2b**), and a similar trend was observed for the 1MM primer set. The annealing temperature that produced the maximum number of unique clones was identified for each set as follows: Full: 52°C, 1MM: 47°C, Compact: 42°C. Use of a hemi-nested RT-PCR strategy (**Fig S2**) resulted in no apparent amplification of non-TCRB sequences, even at the lowest PCR annealing temperature.

We next quantified PCR amplification bias inherent to the three primer sets. cDNA first underwent 20 cycles of PCR using each primer set separately. An aliquot of this product was diluted 2^{10}-fold and then subjected to 10 more cycles of PCR; CDR3 sequencing was then performed on both amplicons and the resulting data sets compared to each other. In this experiment, significant per-cycle amplification bias would manifest as discordance. However, amplification bias was undetectable for all three sets of primers, suggesting that these primer sets negligibly bias the CDR3 population during cycles of PCR amplification (**Fig 2c**).

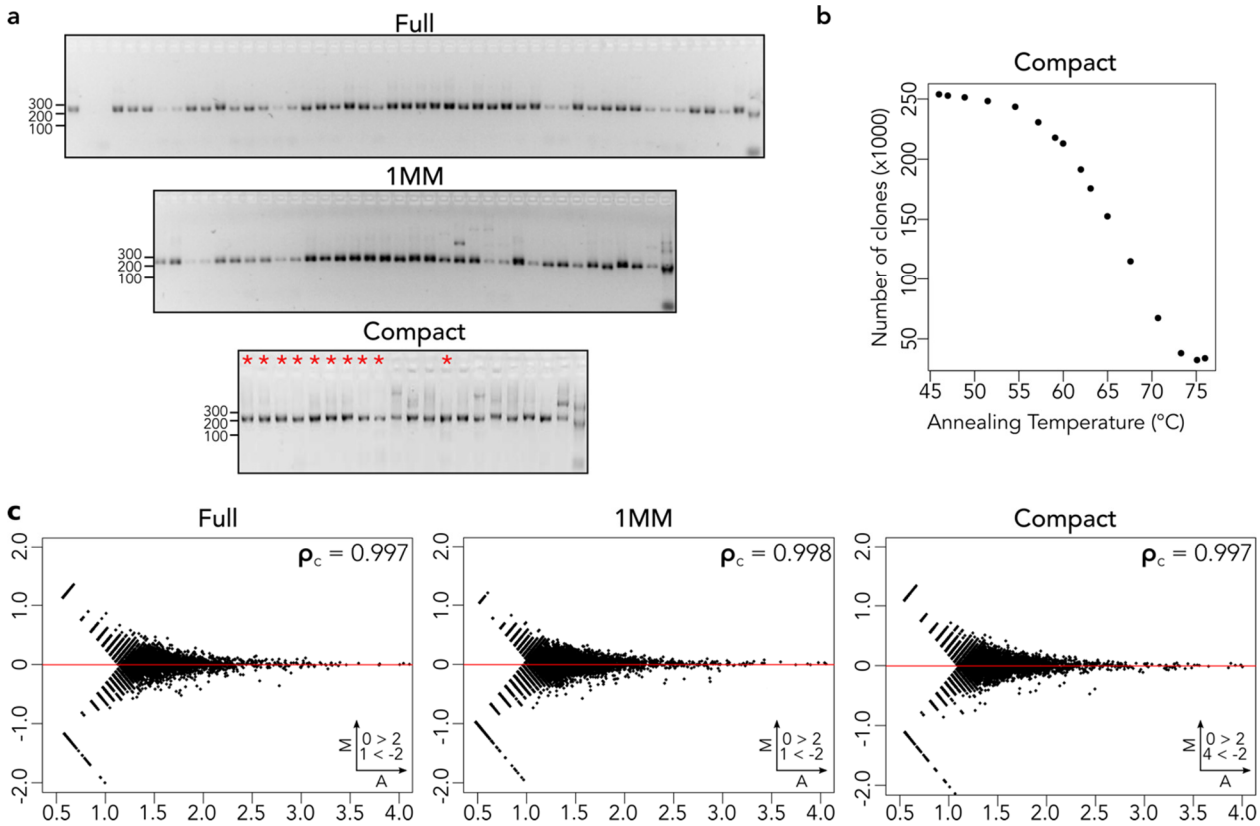

Figure 2: FR3AK-seq multiplex PCR performance. **a.** Individual primers from all three sets produce amplicons of the expected size from human PBMCs cDNA (~220 bp). Asterisks indicate inosine-containing primers. Order of primers can be found in **Table S1**. **b.** The number of clones detected versus annealing temperature using the Compact primer set is shown. A similar trend was observed for the 1MM primer set. To account for differences in number of reads between samples, FASTQ sequencing files were trimmed to identical line numbers before analysis. **c.** MA plots of 2^{10} dilution experiments for each primer set

demonstrate high linear concordance (Full: 0.997, 1MM: 0.999, Compact: 0.980), indicating negligible PCR bias. Clones with ≥10 counts in either dataset were included. Insets indicates number of clones above or below Y axis limits (set to 2 and –2 for visualization).

FR3AK-seq performs comparably to the current industry standard

PBMC RNA from two donors, A and B, as well as a third sample comprised of 90% donor A RNA and 10% donor B RNA (sample "C"), were used to generate cDNA for benchmarking studies against the current industry standard, the immunoSEQ hsTCRB "Deep Resolution" sequencing service offered by Adaptive Biotechnologies. Separate cDNA aliquots were also subjected to FR3AK-seq analysis using each of the three primer sets. The open source software MiXCR v2.1.11 was used to define and quantify CDR3 sequences from both assays for direct comparison.[31] The technical reproducibility of the immunoSEQ assay was quantified by comparing the abundances of donor A's unique CDR3 sequences against their corresponding abundances in sample C (concordance correlation coefficient, ρ = 0.989, left panel **Fig 3a**). The same analysis was applied to the FR3AK-seq data sets, which were generated using each of the three FR3 primer sets. Equally high measures of internal concordance were observed (ρ = 0.990 for the 1MM primer set, right panel **Fig 3a;** ρ = 0.990 for the Full primer set and ρ = 0.987 for the Compact primer set, **Fig S3a**). Importantly, the concordance between immunoSEQ and FR3AK-seq measurements of CDR3 abundance was also very high (ρ = 0.975, **Fig 3b**, **Fig S3b**). The vast majority of even moderately abundant immunoSEQ-defined CDR3 sequences were also detected using any of the three FR3 primer sets (**Fig 3c**). The 9:1 mixture composed of cells from donor A and donor B (sample C) was also used to determine quantification accuracy of the FR3 primer sets relative to the immunoSEQ assay (**Fig 3d**, **Fig S3c**). These data indicate that all three FR3 primer sets provide CDR3 quantification comparable to the current industry standard, with the 1MM set performing most accurately.

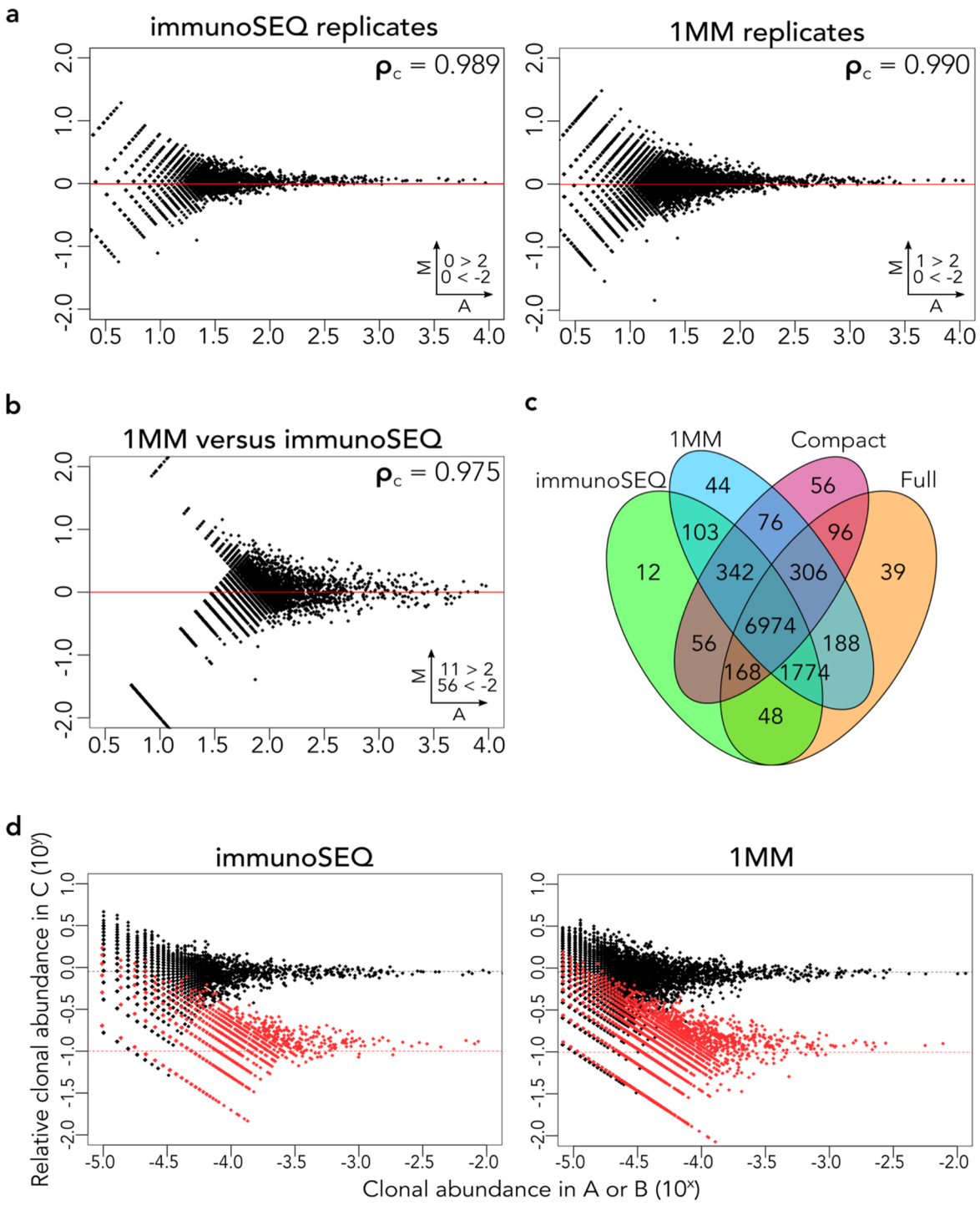

Figure 3: FR3AK-seq performs comparably to immunoSEQ. **a.** MA plots comparing technical replicates of the immunoSEQ and 1MM FR3AK-seq assays. **b.** MA plot comparing the 1MM FR3AK-seq assay against the immunoSEQ assay. Clones with ≥10 counts in either dataset were included. Y axis: log10(immunoSeq/1MM). **c.** A Venn diagram shows overlapping and non-overlapping CDR3 sequences

detected using immunoSEQ and all three FR3AK-seq primer sets. CDR3 sequences were included in this analysis if they had a clone count of at least 10 in any of the four data sets. **d.** Relative frequencies of Donor A (black dots) and Donor B (red dots) in sample C determined using the immunoSEQ assay or the 1MM FR3AK-seq assay. Insets in **a** and **b** indicate number of clones above or below Y axis limits (set to 2 and -2 for visualization).

V gene inference from FR3 primer usage

Of the 128 V alleles enumerated within the IMGT database, 56 can be uniquely distinguished via their complete FR3 domain sequences. Of these, 18 can be unambiguously identified by the 20 nucleotides we have targeted for amplification. We therefore hypothesized that specific patterns of FR3AK-seq primer utilization could be used to infer which of the underlying 47 unique 3' FR3 20-mers is associated with each CDR3. We first tabulated the usage of each primer (determined using the first 20 nt of each sequencing read) by each CDR3. Primer usage can be represented as a "usage vector" in N-dimensional space, where N is equal to the total number of primers contained in the primer set. Each primer usage vector can then be compared against 47 "efficiency vectors", which are defined by calculating the efficiency with which each primer will amplify each FR3 20-mer. We next assumed that for a given CDR3-linked FR3 20-mer sequence, the observed usage vector would be aligned with its corresponding efficiency vector in "primer space". Based upon this simple principle, we developed the Inferring Sequences via Efficiency Projection and Primer Incorporation ("ISEPPI") procedure to link CDR3s with their most likely associated FR3 20-mer (**Fig 4a**). ISEPPI was performed for donor A's CDR3-associated 1MM primer usage vectors and compared to the immunoSEQ-defined FR3 20-mer, which we considered ground truth (**Fig 4b,** Full primer set **Fig S4a**). This heatmap reveals (i) the high accuracy of ISEPPI, and (ii) that incorrect FR3 assignments most frequently involve assignments to closely related FR3 sequences. For moderately abundant CDR3 sequences (≥10 clone

counts), ISEPPI correctly assigns 58.4% of the CDR3 clone counts to the correct FR3 20-mer (**Fig 4c**). For clones with a higher abundance of ≥50 counts or ≥100 counts, 63.2% and 66.4%, respectively, are assigned to the exact FR3 20-mer observed in the immunoSEQ data set. As expected, the accuracy of ISEPPI is improved with use of the Full primer set: ≥10: 66.8%, ≥50: 72.2%, ≥100: 74% (**Fig S4b**).

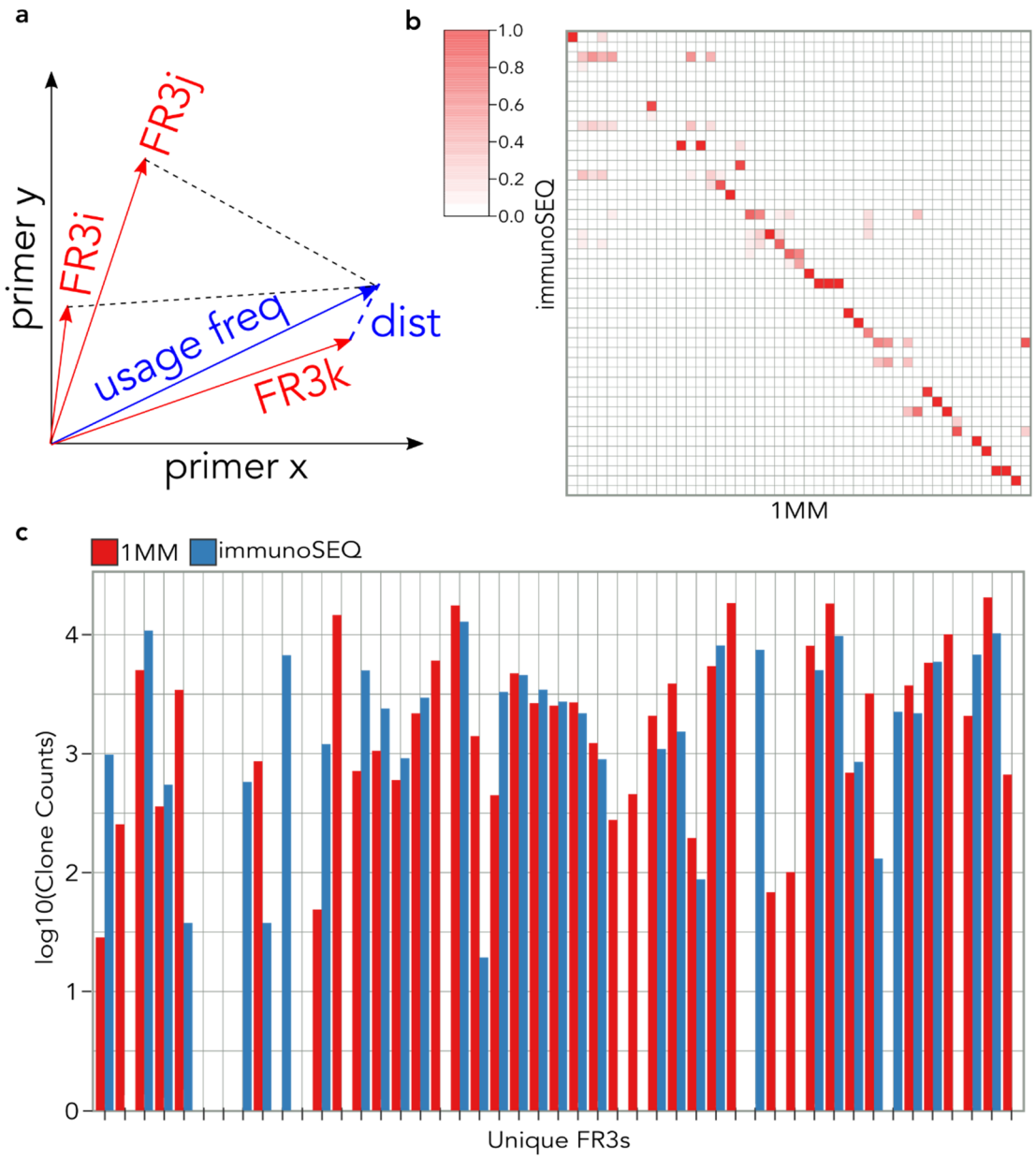

Figure 4: Inferring Sequences via Efficiency Projection and Primer Incorporation (ISEPPI). **a.** Each primer's amplification efficiency is calculated for each of the 47 unique TCRB FR3 20-mer sequences, thereby forming each FR3 amplification efficiency vector (red vectors, FR3i-k). A primer usage vector for each CDR3 is also determined (blue vector). Each CDR3 is assigned to a FR3 sequence by finding the nearest FR3 amplification efficiency vector (Euclidean 'dist', blue dashed line). **b.** Heatmap comparing FR3 assignments to CDR3s using ISEPPI analysis of 1MM FR3AK-seq data (X axis) versus immunoSEQ data (Y axis, ground truth). CDR3s were included in the analysis if they had a clone count of at least 10 in both datasets. The order of FR3s is the same as in **Fig 1c**; heatmap columns are normalized to 1. **c.** ISEPPI correctly assigned 58.4% of clone counts (for CDR3s with a clone count of at least 10). The order of FR3s is the same as in **Fig 1c**.

FR3AK-seq enables inexpensive cohort-scale repertoire studies

146 muscle biopsies from patients with idiopathic inflammatory myopathies (IIMs, 124) as well as healthy controls (HC, 9) and non-myositis controls (NM, 13) were analyzed using the 1MM primer set at a total cost of ~$20 per sample (**Table S2**). IIM patients included those with dermatomyositis (DM, 40), immune-mediated necrotizing myopathy (IMNM, 49), inclusion body myositis (IBM, 14), and anti-synthetase syndrome (ASyS, 21). For absolute T cell number estimation, 1,000 cell equivalents of RNA derived from a clonal T cell line[32] was spiked into each sample after RNA purification. Aggregated CD4 and CD8 sequencing reads from RNA-seq analysis of the same samples tightly correlated with a FR3AK-seq based estimate of cellular equivalents (**Fig 5a**). In further analysis, CDR3 sequences present at levels below 1 cell equivalent per biopsy were considered 'bystander' T cell clones, unlikely to be involved in the disease process. By this metric, the number of non-bystander T cells were elevated in each disease subgroup versus the controls (**Fig. 5b**). In agreement with previous immunohistochemical analyses, IBM biopsies contained high levels of T cell infiltrates.[33–35] We determined this to be the case both in terms of absolute T cell number and clonal diversity (**Fig S5**).

The open source software GLIPH (grouping of lymphocyte interactions by paratope hotspots)[11] was next used to search for clusters of related CDR3 sequences within the aggregated set of non-bystander CDR3s. 143 clusters with sequences contributed by at least 3 individuals were identified. We first examined clusters that had a disproportionate number of sequences contributed by single individuals. **Fig 5c** provides an example of a cluster dominated by related CDR3 sequences contributed by one ASyS patient. Compared to CDR3 sequences from the same patient which were not in any cluster, the clonal abundances of these CDR3s were significantly increased, suggestive of shared antigen-driven expansion. These results suggest that GLIPH is likely able to cluster functionally related CDR3 sequences that have been detected and quantified via the FR3AK-seq assay. We next determined whether any of the GLIPH clusters were composed of sequences contributed disproportionately by individuals from one of the disease subgroups. Indeed, one GLIPH cluster was composed exclusively of sequences contributed by IBM patients. Specifically, 5 of the 12 IBM patients with non-bystander T cells, contributed at least one sequence to this cluster ($p = 1.4x10^{-6}$, **Fig 5d**). ISEPPI analysis revealed focused V-gene usage within this cluster, as visualized in the sequence logo. Four of the 12 (33.3%) assigned FR3s from 3 different patients are tightly clustered (FR3s 1, 2, and 3 from IBM82, IBM27, and IBM13), while 6 of 12 (50%) ISEPPI assigned FR3s from one patient are identical (FR3 30, IBM82). Remarkably, all CDR3 sequences in this IBM cluster are linked to a single J-chain (TRBJ2-2*01), providing additional support for shared antigen recognition**.** Taken together, this study demonstrates the biomedical utility of FR3AK-seq and the ISEPPI procedure in probing patients' T cell responses at cohort scale.

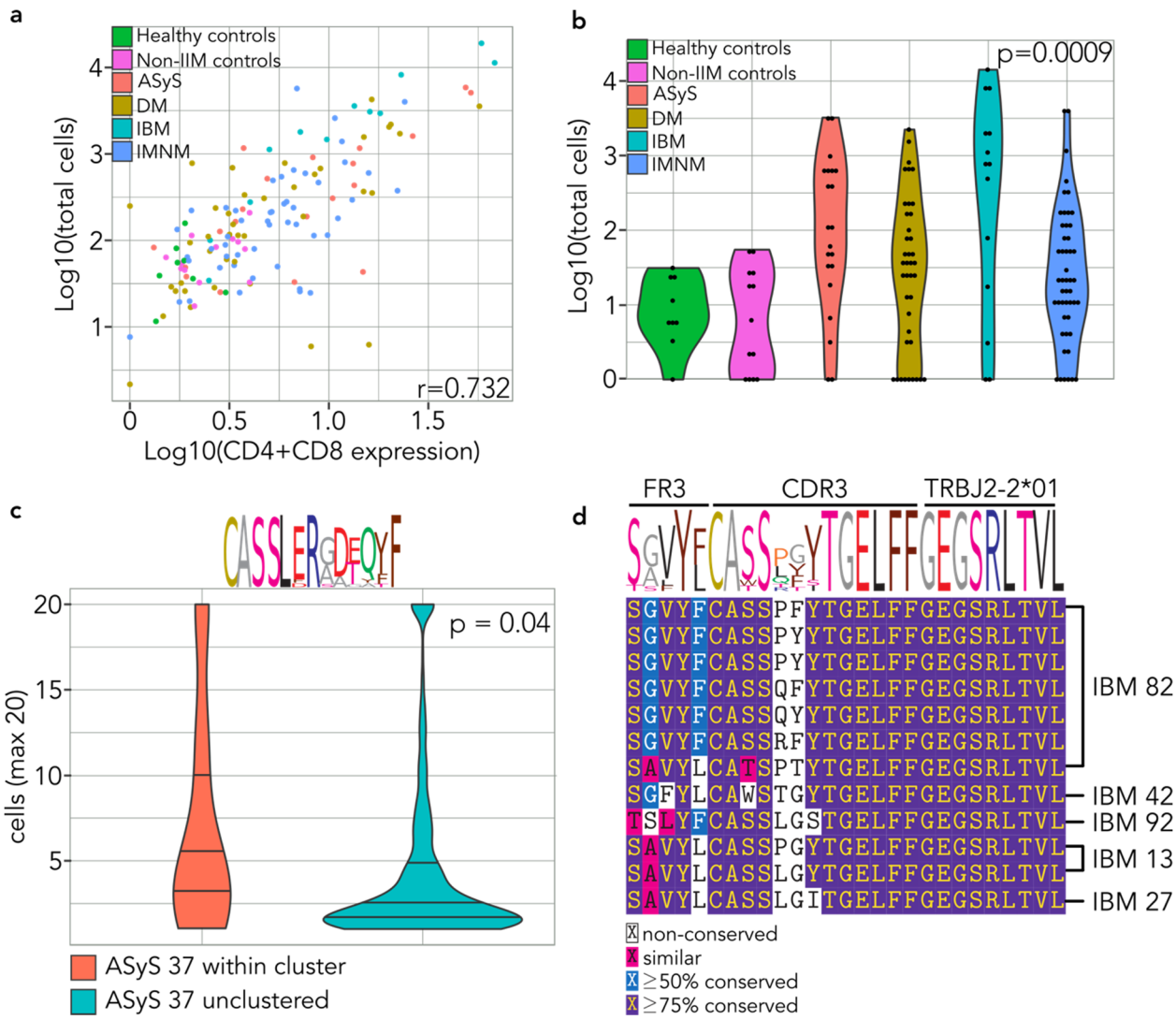

Figure 5: FR3AK-seq detects biologically meaningful T cell responses in patients with idiopathic inflammatory myopathies. **a.** Scatterplot of cell counts per patient versus matched CD4+CD8 expression levels obtained from total RNA sequencing. Each patient data point is colored to represent IIM subgroup. Pearson correlation coefficient (r) is provided. **b.** Distributions of cell counts per patient for each IIM subgroup and controls. Kruskal-Wallis p-value is provided. **c.** A CDR3 cluster enriched for sequences from a single patient with ASyS was identified using GLIPH software ($p = 1.4x10^{-7}$). The amino acid sequence logo of the CDR3 region is shown above the violin plot. The clones within this cluster are more expanded than corresponding unclustered clones from the same patient. Mann-Whitney p = 0.04. Cell counts were set to a maximum of 20 for visualization. **d.** An IBM-exclusive cluster was also identified using GLIPH, encompassing 5 of 12 IBM patients who had non-bystander T cell infiltrates (41.7%, $p=1.4x10^{-6}$). Multiple

sequence alignment (using MUSCLE) is shown for the corresponding ISEPPI-defined FR3s, MiXCR-defined CDR3s, and the sequenced J chain allele (TRBJ2-2*01).

Discussion

Although the utility of IRR sequencing to characterize adaptive immune responses has been well established, the lack of a streamlined, quantitative, inexpensive, and non-proprietary assay has limited the scope and scale of feasible studies. By rationally designing minimal sets of primers that target the 3' terminus of FR3, we have developed a multiplex PCR-based approach for ultra-efficient library preparation and sequencing of CDR3 repertoires. By minimizing amplification bias (via reduction of primer number and amplicon length), the resulting sequencing libraries quantitatively capture clonal abundance distributions with an accuracy comparable to the industry standard. While the library preparation and sequencing strategy presented here already bring the per-sample cost to ~$20 (**Table S2**)**,** we expect that alternative approaches to library preparation, reduced sequencing depth, and declining sequencing costs, will likely enable another ~10-fold reduction in cost.

A key feature of our approach to IRR analysis is its generalizability to additional immune receptor types as well as to those of non-human species (see **Table S3** for *Mus musculus* TCRB primers). We have additionally used the same FR3 primer design and analysis principles to target human TCR alpha, gamma, and delta repertoires (**Table S4**), and for both the heavy and light chains of human B cell receptors (BCRs, antibodies, **Table S5**).

Analysis of V gene utilization is important for detecting shared specificities among T cell clones. The ISEPPI procedure, introduced herein, which is based on simple Euclidean distances measured in primer space, successfully assigns up to 74% of T cell clone counts to their exact FR3 20-mer sequence. Many incorrect assignments are made to FR3 sequences of high sequence similarity (**Fig 4a**). More careful consideration of sequence similarities may therefore

improve interpretation of ISEPPI results. Beyond a simple distance measure, ISEPPI improvements may utilize more sophisticated machine-learning classification strategies. In addition, it is likely that the ISEPPI principle may find applications beyond IRR studies, in settings where deconvolution of multiplex PCR primer usage is important.

Given the current price of a 100 nt single-end Illumina sequencing read (~$0.00001), the information-to-cost ratio is extremely favorable. Studies that identify public TCR clones and shared epitope recognition groups will further enhance the information content of IRR analyses by linking specific CDR3 sequences with relevant meta-data (e.g. disease association, antigenic target, etc.). In addition to reducing the overall cost of obtaining CDR3 sequences, we have developed the FR3AK-seq workflow to require minimal expertise, steps, and time. Quantitative sequencing libraries can easily be prepared within a single day. To demonstrate the use of FR3AK-seq for low cost cohort scale analysis of TCRB CDR3 repertoires, we characterized the muscle-infiltrating T cells present within 146 muscle biopsies obtained from inflammatory muscle disease patients and controls. CDR3 sequence clustering indicated both donor and disease specific antigen driven T cell responses. Future studies will seek the identity of specific antigens that drive the IBM-specific T cell responses we have identified. We anticipate that the FR3AK-seq assay will enable routine, large scale IRR studies like the one presented here.

Methods:

FR3 primer design

FR3 sequences from all human TCRB V alleles were downloaded from the IMGT/LIGM-DB.[26] Multiple sequence alignment (MSA) was performed using MUSCLE, and UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and neighbor-joining phylogenetic trees were generated via the R package msa.[36] The 3' 20 nucleotides of the TCRB chain FR3s were subsequently used

to design three sets of primers as described in the text. We automated design of the 1MM and the Compact primer sets with an algorithm outlined in **Fig S1**. Primer sequences can be found in **Table S1**.

PBMC preparation, T cell purification, RNA extraction, and cDNA synthesis

PBMCs from donors A and B were isolated by Ficoll-Paque (GE Healthcare) gradient centrifugation and cryopreserved. T cells were purified from thawed PBMCs using the EasySep Human T cell enrichment kit (STEMCELL Technologies). RNA was extracted from purified human T cells using the RNeasy Plus Minikit (Qiagen). 0.5 μg Donor B RNA was mixed with 4.5 μg Donor A RNA to create sample C for quantitation experiments. 4 μg each of donor A, donor B, and sample C RNA was reverse transcribed using a TCRB chain constant region reverse primer with Superscript III First-Strand Synthesis System (Invitrogen). cDNA was column purified with the Oligo Clean and Concentrator Kit (Zymo Research). 100 ng of purified cDNA from each sample was sent to Adaptive Biotechnologies for the immunoSEQ hsTCRb Deep sequencing service or used for FR3AK-seq PCR. All primer sequences are provided in **Table S1**.

Polymerase Chain Reaction (PCR) and sequencing library preparation

100 ng TCRB chain cDNA was used as template for PCR with KAPA2G Fast Multiplex Mix (Roche). Forward TCRB FR3 primers from the Full, 1MM, or Compact primer sets were used with a single nested TCR beta chain constant region reverse primer at 0.2 μM per primer. Samples underwent 30 cycles of PCR for visualization by agarose gel electrophoresis, or 20 cycles for NGS library preparation ("PCR1").

20 cycles of PCR2 were performed on PCR1 product (2 μL of PCR1 added to 18 μL of PCR2 master mix, which contained 0.25 uM each i5 and i7 sample barcoding primers) to incorporate sample barcodes and Illumina sequencing adaptors using Herculase II Fusion DNA Polymerase (Agilent). Equal volumes of barcoded PCR2 products (5 μL each) were pooled and PCR column

purified using QIAquick PCR Purification Kit (Qiagen). Libraries were quantified using KAPA Library Quantification Kit for Illumina Platforms (Roche).

Sequencing and CDR3 analysis

Sequencing was performed on either an Illumina NextSeq 500 (benchmarking comparisons) or HiSeq 2500 (IBM muscle biopsy analysis). CDR3s were identified and quantified using MiXCR v2.1.11 software[31] with default parameters except as noted in **Table S6.** Data obtained from Adaptive Biotechnologies were re-analyzed using MiXCR v2.1.11 and identical parameters for comparison to our own data. For reanalysis, full nucleotide sequences for each CDR3 from the Adaptive Biotechnologies ImmunoSEQ dataset were expanded to repeat as many times as indicated by the corresponding "count" column. This file, with each clonal nucleotide sequence represented as many times as its counts column, was converted to FASTA format for compatibility with MiXCR. Clone count cutoffs for each analysis are as described in the figure legends.

Inferring Sequences via Efficiency Projection and Primer Incorporation (ISEPPI)

Primer amplification efficiencies for ISEPPI analysis were calculated for the 47 unique FR3 20-mers for both the 1MM and Full primer sets using the R package DECIPHER and its CalculateEfficiencyPCR function.[29] For primer usage vectors, we searched for each relevant CDR3 nucleotide sequence obtained from MiXCR within the FASTQ sequencing file and extracted the first 20 nucleotides from each matching read. These sequences were then compared to the corresponding primer sequences and use of each primer was tabulated to generate a vector of length N, where N equals the number of primers used for amplification. Sequences that were not an exact match to a primer sequence were presumed sequencing errors and excluded from analysis. All vectors were converted to unit vectors prior to distance measurements. Each CDR3 was assigned to the FR3 that had a unit efficiency vector nearest (minimum Euclidean distance) to the primer usage unit vector. For assessment of FR3-

assignment accuracy using the immunoSEQ data as ground truth, we extracted all sequences upstream of CDR3, including the first 3 nucleotides of CDR3 (which is included as FR3 in our primer design), from the immunoSEQ dataset. We then trimmed this FR3 sequence to the 3' 20 nucleotides and assigned its corresponding IMGT reported FR3 20-mer as ground truth. For this analysis, we only considered CDR3s with clone counts of ≥10 in both data sets.

Repertoire analysis using muscle biopsies from patients with idiopathic inflammatory myopathies (IIMs)

RNA isolated from 146 muscle biopsies taken from IIM patients or controls was reverse transcribed, PCR amplified, sequenced, and analyzed as described above. We added 1,000 cell equivalents of RNA from a clonal T cell line as a spike-in to each sample.[32] Sequencing of 6 spike-in only samples provided the CDR3 sequences to remove from downstream analyses of samples containing patient RNA. The spike-in counts were used for cell number quantification within patient samples, which distinguished 'bystander' versus 'non-bystander' T cell clones.

CDR3 sequences obtained from these samples were analyzed for disease-specific motifs using the GLIPH 1.0 group discovery algorithm (see **Table S6** for parameters).[11] To determine the statistical significance of each cluster, we performed a chi square analysis on all clusters with at least three contributing individuals, followed by Benjamini-Hochberg multiple comparison correction. This analysis was performed for both the number of sequences contributed by each disease subgroup to each cluster, as well as the number of patients in each disease group represented within each cluster. P-values ≤ 0.05 after multiple test correction were considered statistically significant.

Acknowledgements:

The authors would like to thank Dr. Erik Wright for his advice on primer design, Dr. Suzanne L. Topalian for the TIL1558 cell line, and Dr. Maximilian F. Konig for his review of the manuscript. We would also like to thank the Johns Hopkins Medical Institute Deep Sequencing and Microarray Core Facility and the Johns Hopkins Institute of Genetic Medicine Genetics Resources Core Facility. This work was supported by NIH grant U24AI118633 and a Prostate Cancer Foundation Young Investigator Award.

References:

1. Attaf, M., Huseby, E. & Sewell, A. K. αβ T cell receptors as predictors of health and disease. *Cell. Mol. Immunol.* **12**, 391–399 (2015).
2. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).
3. Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. & Holt, R. A. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
4. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099–4107 (2009).
5. Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
6. Boyd, S. D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
7. Muraro, P. A. *et al.* T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *J. Clin. Invest.* **124**, 1168–1172 (2014).

8. Gros, A. *et al.* PD-1 identifies the patient-specific CD8$^+$ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124**, 2246–2259 (2014).
9. Beausang, J. F. *et al.* T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the T cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10409–E10417 (2017).
10. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
11. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
12. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
13. DeWitt, W. S. *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* **7**, (2018).
14. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
15. Morris, H. *et al.* Tracking donor-reactive T cells: Evidence for clonal deletion in tolerant kidney transplant patients. *Sci. Transl. Med.* **7**, 272ra10 (2015).
16. Theil, A. *et al.* T cell receptor repertoires after adoptive transfer of expanded allogeneic regulatory T cells. *Clin. Exp. Immunol.* **187**, 316–324 (2017).
17. Mamedov, I. Z. *et al.* Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.* **3**, 201–207 (2011).
18. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).
19. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).

20. Turner, S. J., Doherty, P. C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6**, 883–894 (2006).

21. Miles, J. J., Douek, D. C. & Price, D. A. Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375–387 (2011).

22. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143 (2015).

23. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).

24. Peng, Q., Vijaya Satya, R., Lewis, M., Randad, P. & Wang, Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics* **16**, 589 (2015).

25. Egorov, E. S. *et al.* Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol. Baltim. Md 1950* **194**, 6155–6163 (2015).

26. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* **33**, D256-261 (2005).

27. Simsek, M. & Adnan, H. Effect of single mismatches at 3'-end of primers on polymerase chain reaction. *J. Sci. Res. Med. Sci.* **2**, 11–14 (2000).

28. Wu, J.-H., Hong, P.-Y. & Liu, W.-T. Quantitative effects of position and type of single mismatch on single base primer extension. *J. Microbiol. Methods* **77**, 267–275 (2009).

29. Wright, E. S. *et al.* Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. *Environ. Microbiol.* **16**, 1354–1365 (2014).

30. Ishii, K. & Fukui, M. Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Appl. Environ. Microbiol.* **67**, 3753–3755 (2001).

31. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

32. Deng, L. *et al.* Structural basis for the recognition of mutant self by a tumor-specific, MHC class II-restricted T cell receptor. *Nat. Immunol.* **8**, 398–408 (2007).

33. Arahata, K. & Engel, A. G. Monoclonal antibody analysis of mononuclear cells in myopathies. I: Quantitation of subsets according to diagnosis and sites of accumulation and demonstration and counts of muscle fibers invaded by T cells. *Ann. Neurol.* **16**, 193–208 (1984).

34. Fyhr, I. M., Moslemi, A. R., Lindberg, C. & Oldfors, A. T cell receptor beta-chain repertoire in inclusion body myositis. *J. Neuroimmunol.* **91**, 129–134 (1998).

35. Cavazzana, I., Fredi, M., Selmi, C., Tincani, A. & Franceschini, F. The Clinical and Histological Spectrum of Idiopathic Inflammatory Myopathies. *Clin. Rev. Allergy Immunol.* **52**, 88–98 (2017).

36. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinforma. Oxf. Engl.* **31**, 3997–3999 (2015).