# High-throughput pairing of T cell receptor alpha and beta sequences

**Article** *in* Science translational medicine · August 2015

DOI: 10.1126/scitranslmed.aac5624 · Source: PubMed

IMMUNOLOGY

# High-throughput pairing of T cell receptor α and β sequences

Bryan Howie,[1]* Anna M. Sherwood,[1]* Ashley D. Berkebile,[1] Jan Berka,[1] Ryan O. Emerson,[1] David W. Williamson,[1] Ilan Kirsch,[1] Marissa Vignali,[1] Mark J. Rieder,[1] Christopher S. Carlson,[2] Harlan S. Robins[1,2†]

**The T cell receptor (TCR) protein is a heterodimer composed of an α chain and a β chain. TCR genes undergo somatic DNA rearrangements to generate the diversity of T cell binding specificities needed for effective immunity. Recently, high-throughput immunosequencing methods have been developed to profile the TCR α (TCRA) and TCR β (TCRB) repertoires. However, these methods cannot determine which TCRA and TCRB chains combine to form a specific TCR, which is essential for many functional and therapeutic applications. We describe and validate a method called pairSEQ, which can leverage the diversity of TCR sequences to accurately pair hundreds of thousands of TCRA and TCRB sequences in a single experiment. Our TCR pairing method uses standard laboratory consumables and equipment without the need for single-cell technologies. We show that pairSEQ can be applied to T cells from both blood and solid tissues, such as tumors.**

## INTRODUCTION

The αβ T cell receptor (TCR) protein, which determines the antigenic specificity of an αβ T cell, is a heterodimer composed of two peptides: a longer β chain (TCRB) and a shorter α chain (TCRA) (*1*–*6*). Recently, high-throughput sequencing assays have been developed to profile TCRA and TCRB chains, as well as immunoglobulin heavy and light chains, with multiple diagnostic applications, including clinical diagnostics for the detection of cancer clones and the measurement of minimal residual disease in lymphoid malignancies (*7*–*20*). However, these high-throughput methods can sequence only one chain of the TCR at a time. To reconstitute TCRs for functional analysis, therapeutic use, or modeling of receptor-antigen binding, the TCRA and TCRB chains from a complete TCR must be identified as a pair.

There have been multiple attempts to pair heavy and light chains in B and T cells using single-cell technology (*21*–*32*). One approach is to isolate individual lymphocytes and physically link the chains by bridge polymerase chain reaction (PCR) before sequencing; alternatively, the heavy and light chains can be barcoded at the single-cell level. Although single-cell methods have improved substantially, they are still technically challenging and limited in throughput. Herein, we present a technology to pair lymphocyte receptor sequences at high throughput without the need for isolated single-cell methods; our strategy uses combinatorics, rather than physical isolation, to match the pairs. As a first demonstration, we have applied this technology to pair hundreds of thousands of TCRA and TCRB gene sequences from the peripheral blood of two healthy donors, as well as thousands of sequences from tumor-infiltrating T cells in nine pairs of matched tumor and blood samples.

## RESULTS

### Experimental design

Our method, called "pairSEQ," relies on the observation that rearranged TCRA and TCRB nucleotide sequences are nearly unique for each clonal population of T cells. Distinctive TCR sequences arise through recombination of gene segments and stochastic deletion and insertion of nucleotides at the V-J, V-D, and D-J junctions in somatic cells during lymphocyte development (*33*). This extraordinary diversity means that mRNA sequences encoding the TCRA and TCRB chains of a given T cell clone will usually be unique to that clone. We leverage this diversity to identify pairs originating from particular clones by splitting a sample of T cells into multiple subsets and then reverse-transcribing and sequencing the TCRA and TCRB transcripts in each subset; the TCRA and TCRB sequences from a clone should be seen in the same subsets of T cells and only those subsets (Fig. 1). Designing an experiment to pair cognate TCR sequences then becomes a combinatorics problem: T cells must be split into enough subsets that each clone of interest is unlikely to occupy the same collection of subsets as another clone. The ability to accurately pair a clone's TCRA and TCRB sequences is a function of the frequency of that clone in the original sample, the number of T cells in each subset, and the number of subsets created. The last two parameters are part of the experimental design, and they can be tuned to choose the range of T cell clonal frequencies that will be accurately paired (fig. S1 and Supplementary Materials).

Figure 1 shows how these ideas are implemented in a pairSEQ experiment. First, T cells are collected and distributed among the wells on a microtiter plate to create 96 distinct subsets. Total cellular RNA is then extracted from the cells in each well and reverse-transcribed into complementary DNA (cDNA); the presence of multiple molecules of mRNA per cell for most genes makes cDNA preferable to genomic DNA (gDNA), allowing for some loss of material during extraction without compromising sequence detection. The TCRA and TCRB sequences are then PCR-amplified within each well, using primers specific for the V and C gene segments. Because T cells do not undergo somatic hypermutation, the VDJ-recombined complementarity-determining region 3 (CDR3) region is sufficient to characterize the entire TCR. Well-specific oligonucleotide DNA barcodes are then attached to the amplified receptor molecules, and the amplified molecules are pooled for high-throughput DNA sequencing, which reads both the receptor sequence and the barcode for each strand. The sequenced barcodes allow mapping of the receptor sequences to the wells of origin. The fact that a cell being present in a given well does not always guarantee observation of both

[1]Adaptive Biotechnologies, Seattle, WA 98102, USA. [2]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.
*These authors contributed equally to this work.
†Corresponding author. E-mail: hrobins@fredhutch.org

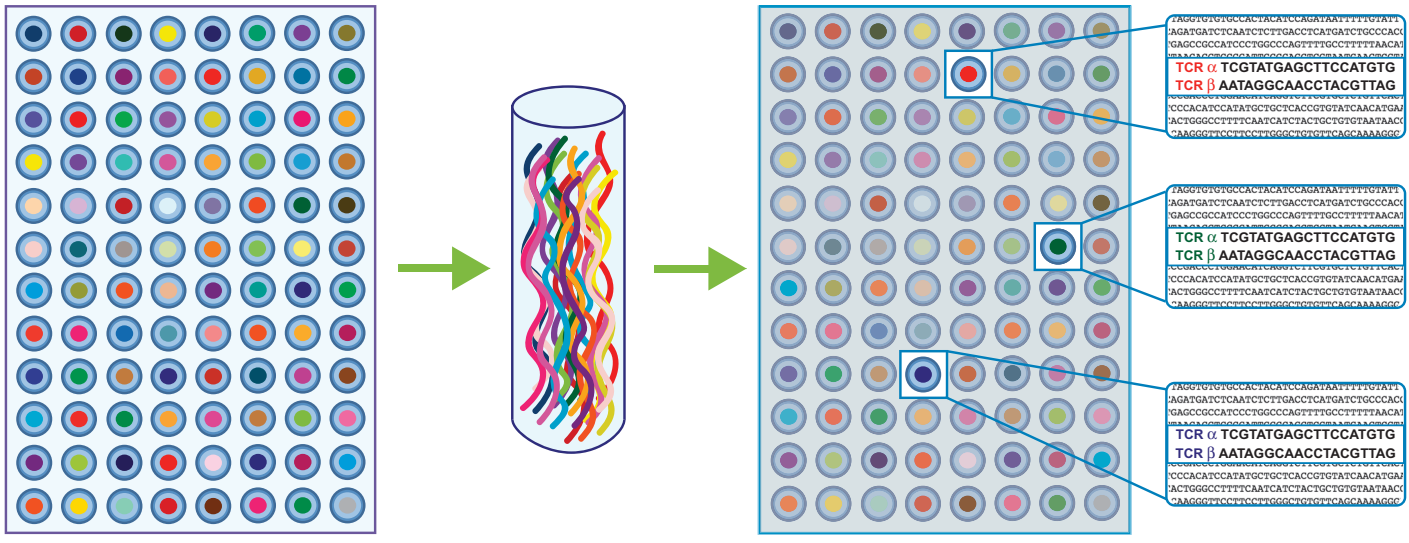



**Fig. 1. Schematic of the pairSEQ approach.** A fixed number of T cells are randomly allocated to each well on a 96-well plate, and their mRNA is extracted, converted to cDNA, and amplified by TCR-specific primers. Well-specific barcodes are attached, and the TCR molecules are pooled for sequencing, followed by computational demultiplexing to map each TCR sequence back to the wells in which it originated. The immune repertoire is highly diverse, and the probability that two clones will occupy exactly the same sets of wells is miniscule, so any pair of TCRA and TCRB sequences that uniquely share a set of wells can be inferred to have come from the same clone.

its TCRB and TCRA sequences (fig. S3 and Supplementary Materials) means that a statistical model is necessary to deal with imperfect overlap in the well occupancy patterns of TCRA/TCRB pairs. To address this issue, putative pairs are identified by comparing the well occupancy pattern of every TCRA sequence against that of every TCRB sequence; sequences that share more wells than expected by chance are marked as possible pairing partners. High-confidence pairs are then identified by generating a null distribution by permutation and identifying the pairs that satisfy a target false discovery rate (FDR; see the Analysis section of Materials and Methods for more details).

## pairSEQ validation

To validate this experimental design and statistical framework, we performed an experiment that directly assesses the credibility of putative pairs identified by our method (Table 1, experiment 1, and Fig. 2). We used the immunoSEQ assay (*14*, *34*) to sequence the TCRA and TCRB gDNA repertoires of 14 million peripheral blood mononuclear cells (PBMCs) from each of two healthy adult subjects, labeled "X" and "Y." This produced 4,522,509 TCRA and 4,655,290 TCRB sequences that were unique to subject X, and 3,721,491 TCRA and 3,900,602 TCRB sequences that were unique to subject Y. The sequencing of the repertoire of subjects X and Y provided two useful pieces of information. First, determining the frequencies of TCRBs allowed us to design a pairSEQ experiment that would capture the majority of clones present at a frequency of at least 1/10,000 T cells in peripheral blood; for each subject, our study design algorithm (see the Supplementary Materials) determined that ~4500 PBMCs should be allocated to each of the 96 wells to achieve ~2000 T cells per well (Table 1). Second, knowing the unique clone sequences from each subject allowed us to mix their cells on a single plate, run a standard pairSEQ analysis, and then assign the pairs

**Table 1. Summary of experiments and pairing results.**

| Experiment | Subjects | Input T cells per well | Pairs called at FDR = 1% |
|---|---|---|---|
| 1 | X and Y | 2,000 per subject | 4,143 |
| 2 | X and Y | 80,000 per subject | 155,805 |
| 3 | X | 160,000 | 212,651 |
| 4 | Nine tumors | Tailored to each sample* | 6,172 |
| 5 | Nine matched PBMCs | Tailored to each sample* | 14,123 |

*For each tumor or blood sample, TCRB repertoire frequencies from the relevant tissue were used to choose a number of input T cells for each of the nine samples such that common clones were likely to be paired.

to the samples of origin post facto using the data from the sequencing of the individual repertoires. About half of the false TCRA/TCRB pairs should comprise a TCRA from one subject and a TCRB from the other, so counting these "cross-subject pairs" provides an empirical measurement of the false pairing rate independent of our statistical framework.

Figure 3 shows the FDR analysis of experiment 1. We divided the pairs into three categories: (i) pairs in which both TCR sequences were exclusive to subject X ("X/X", blue); (ii) pairs in which both TCR sequences were exclusive to subject Y ("Y/Y", orange); and (iii) pairs in which one sequence was exclusive to subject X and the other to subject Y ("X/Y", gray). At a predicted FDR of 1%, we observed 4143 pairs (Table 1), of which 1621 corresponded to X/X pairs, 1616 to Y/Y pairs, and 16 to X/Y pairs (Fig. 3A). An additional 890 called pairs could not be assigned to these categories, typically because the TCRA sequence was observed in the repertoires of both individuals. If the cross-subject

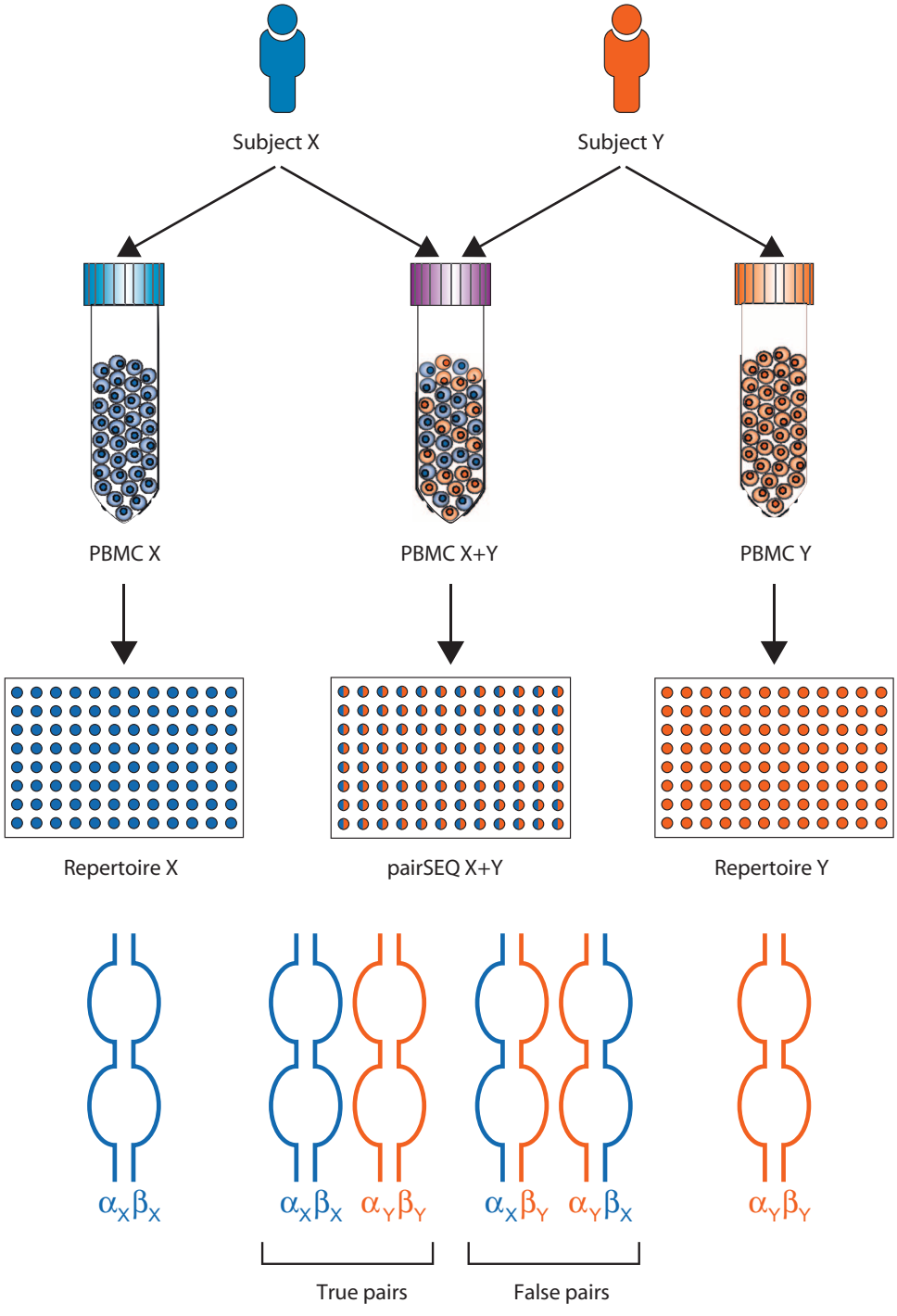



**Fig. 2. Schematic of FDR validation experiment (experiment 1).** Peripheral blood was collected from two subjects, X and Y, and deep-sequenced by immunoSEQ to characterize the TCRA and TCRB repertoire of each subject. PBMCs from the two subjects were then mixed, and the resulting mix was used to perform a pairSEQ experiment. True-positive pairs must include a TCRA and a TCRB from the same subject, whereas about half of false-positive results will be cross-subject TCRA/TCRB pairs.

pairs represented half of the false pairings, as expected, the empirical FDR at this pairing threshold would be 0.98%, which shows excellent agreement with the nominal FDR value of 1%. Figure 3B extends this false-pairing analysis to predicted FDR values ranging from 0.1 to 20%. The nominal FDR values closely track the empirical values across this entire range, showing that our analysis framework provides valid FDR estimates that can be tuned to balance specificity and sensitivity for different applications.

Experiment 1 also included a positive control, in which we spiked in cells from a clonal sample with known TCRA and TCRB sequences [E6-1 T-ALL cell line (*35*)]. We added a solution of Jurkat cells at a concentration of 1/30,000 PBMCs to the bulk mixture of 900,000 PBMCs before allocating cells to the plate, for a target spike-in of 30 Jurkat cells. After running the experiment, we observed Jurkat sequences in 28 of 96 wells, of which 24 contained both the TCRA and TCRB Jurkat sequences. These sequences were confidently paired in our statistical analysis (FDR << 1%), and neither the TCRA nor the TCRB Jurkat sequence showed a significant match to any non-Jurkat sequence, further confirming our method's ability to identify biologically paired sequences.

## High-throughput pairing of targeted clone frequencies

To demonstrate how our method can be extended to capture rarer clones at high throughput, we conducted two additional pairSEQ experiments. The first one contained 80,000 T cells per well from each of subjects X and Y for 160,000 total T cells per well (Table 1, experiment 2), whereas the second one contained 160,000 T cells per well from subject X only (Table 1, experiment 3). Experiment 2 was designed to target clones with frequencies of 1/10,000 or less in both subjects, to complement the common clones paired in experiment 1. Experiment 3 was designed to capture as many rarer clones as possible to further test the ability of the method to pair large numbers of TCRs on a single 96-well plate.

At an FDR of 1%, we were able to identify 155,805 TCR pairs in experiment 2 and 212,651 in experiment 3 (Table 1). Although both experiments received the same number of input T cells, more pairs were called in experiment 3 because the number of "pairable" clones increases rapidly with deeper sampling from a single repertoire. As in experiment 1, a cross-pairing analysis of experiment 2 showed that our algorithm controls false pairings across a wide range of FDR thresholds (fig. S2).

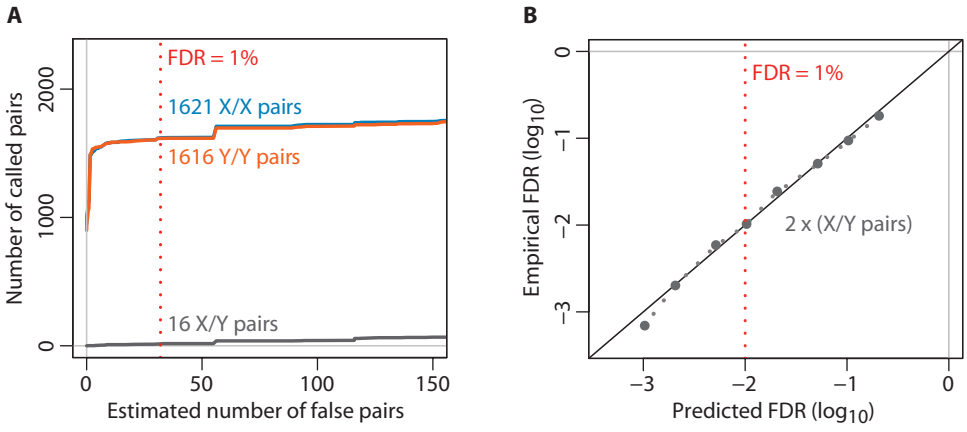



**Fig. 3. Validation of pairSEQ FDR using mixed PBMCs from two subjects (experiment 1).** (**A**) FDR curves for an experiment in which PBMCs from two subjects (X and Y) were mixed, with 2000 T cells per subject, and dispensed into each of 96 wells. Separately, ~6 million cells from each subject were immunosequenced to identify TCRB and TCRA sequences unique to either subject. Pairs were split into groups named "X/X" (both sequences of a pair seen only in subject X; blue), "Y/Y" (both sequences of a pair seen only in subject Y; orange), and "X/Y" (one sequence of a pair seen only in subject X and one seen only in subject Y; gray). The red dotted line shows the cutoff for an estimated FDR of 1%; below this threshold, there are 1621 X/X pairs, 1616 Y/Y pairs, and 16 X/Y pairs. (**B**) Predicted versus empirical $\log_{10}$ FDR in experiment 1. Predicted FDR values were provided by pairSEQ statistics, whereas empirical FDR values were computed as twice the number of X/Y pairs divided by the total number of called pairs, under the assumption that X/Y pairs represent half the total number of errors. As in (A), the red dotted line shows the cutoff for an estimated FDR of 1%.

Notably, the cross-pairing evaluation at a target FDR of 1% included 26,667 TCRB sequences with repertoire frequencies of 1 in 1 million or lower. Among these low-frequency clones, the cross-pairing rate was 0.17%, which implies a pairing error rate of 0.34%. Hence, this experiment validated our method on thousands of rare clones in parallel. The clones validated in our experiment were drawn from a native repertoire. We note that this is a rigorous validation: instead of testing the accuracy of pairing using a handful of spiked-in clones, this test assesses thousands of native clones. Additionally, we find spiked-in clones misleadingly easy to pair because clonal T cell lines can show unusually high mRNA expression.

The combined pairing results from experiments 1, 2, and 3 are shown in Fig. 4A. At an FDR of 1% (dotted red line), we discovered 362,528 unique pairs of TCRA and TCRB sequences, a value that is orders of magnitude higher than that single-cell pairing methods have achieved to date (*21*–*32*). The FDR curve is L-shaped, implying that most of the TCRA/TCRB pairs stood out clearly from the background noise and were called with extremely high confidence. Of the unique TCRB sequences identified, 9.9% were paired to two different productive TCRA sequences. Some of these TCRB sequences paired to two TCRA sequences likely represent true T cell clones that express one β chain and two α chains; the observed abundance is consistent with previous findings on the abundance of T cells that express two α chains (*23*, *36*). Similarly, 2.8% of TCRA sequences paired with two productive TCRB sequences. The probability that each of these pairings represents two T cell lineages with the same α or β chain by convergent recombination is vanishingly small. In principle, it should be possible to distinguish these two cases for each observed TCR rearrangement (that is, two cell lineages that recombined the same TCR sequence versus a single-cell lineage expressing two receptors) because the expected occupancy pattern is different: a cell expressing two TCRA or TCRB receptors should include both TCRs in the same set of wells, whereas receptor sequences from lineages with convergently rearranged TCRA or TCRB receptors should occupy distinct subsets of wells. Finally, 1.9% of TCRB sequences were paired only with a nonproductive TCRA, implying that in these cases, either the correct TCRA was not observed (for example, because of low expression) or it was not paired correctly.

Figure 4 (B and C) shows the pairing results for subject X as a function of TCRB repertoire frequency. Most of the common TCRBs were successfully paired: experiments 1 to 3 captured ~80% of the 10,000 most frequent TCRB sequences (Fig. 4B), which have repertoire frequencies of 1/80,000 or greater. It is possible that the other 20% of TCRB sequences remained unpaired due to technical issues with TCRA sequence amplification and detection, to the stringency of our FDR threshold, or to the very low expression of TCRA mRNA in certain clones (*23*). The pairing efficiency was lower among the top 100,000 clones (~40%; Fig. 4B) because many of these clones fell below our target frequency threshold for complete pairing.

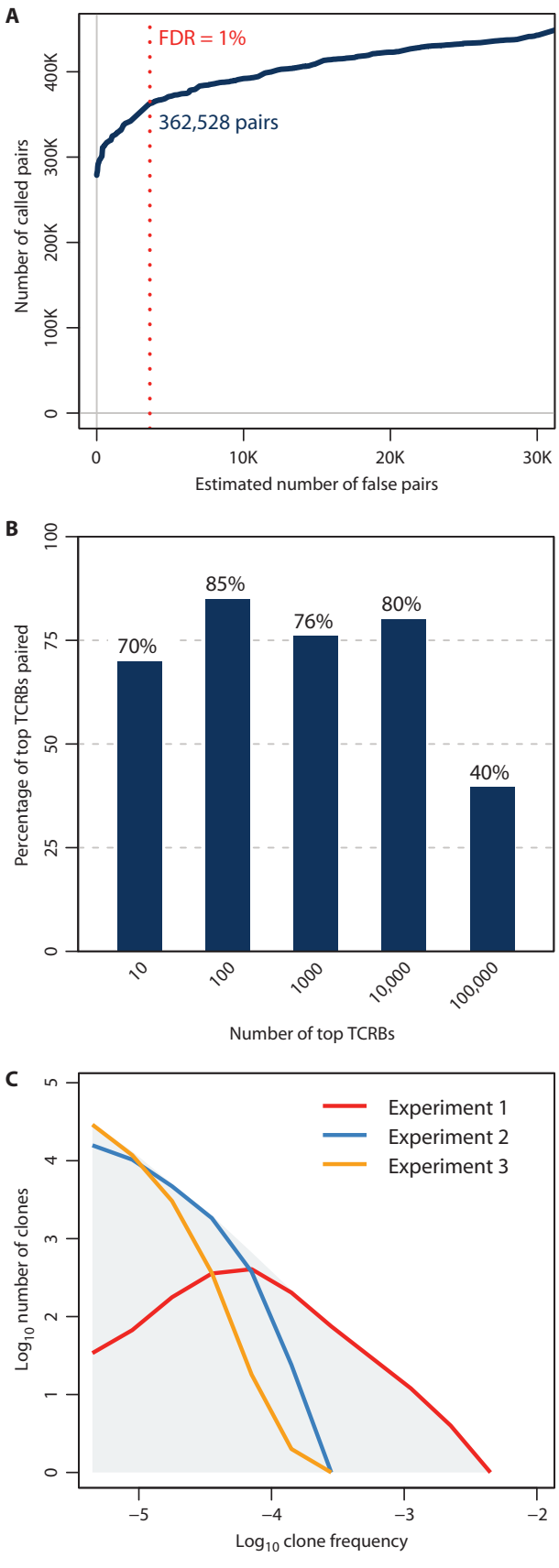
Figure 4B omits rare T cell clones for which our previous immunosequencing could not provide reliable, independent estimates of clonal abundance. However, the rarest clones we paired were observed in six wells (and thus almost certainly represent six cells) in experiment 3, so they are likely to have a repertoire abundance of roughly 1 cell in 2.6 million. This illustrates that our method can pair rare clones as long as they are represented by multiple cells in the input sample. Five or six copies are typically sufficient to prevent collisions between well occupancy patterns, although the 80% pairing efficiency in Fig. 4B requires 10 or more cells from a given clone.

Figure 4C shows the pairing results for subject X in each experiment. As anticipated, experiment 1 (red) captured common clones, experiment 2 (blue) captured a complementary set of rarer clones, and experiment 3 (orange) captured even rarer clones. These results confirm that the pairSEQ assay can be used to assay clones at different regimes of repertoire frequency simply by changing the number of input cells, which makes it a flexible tool for addressing both common and rare T cell populations.

### Analysis of tumor-infiltrating lymphocytes

TCR pairing will have applications in many sample types besides blood. T cells either selected or engineered to express tumor antigen–targeting receptors are proving to be powerful cancer therapies (*37*, *38*). The pairSEQ technology has the potential to rapidly identify sequence pairs of tumor-infiltrating lymphocytes (TILs), which can then be used to reconstruct TCR receptors within T cells engineered for cancer immunotherapy. To show that our technology can also work in this setting, we obtained samples from nine tumors (four kidney, four breast, and

**Fig. 4. Results of high-throughput pairSEQ experiments.** (**A**) FDR curve for the combined results of experiments 1 to 3. A total of 362,528 pairs of TCR sequences were called in subject X and subject Y at an FDR of 1% (red dotted line). (**B**) Percentage of paired TCRB sequences among the *N* most frequent in subject X, for *N* = 10, 100, 1000, 10,000, and 100,000. (**C**) Repertoire frequency distributions of paired TCRB sequences from subject X in experiments 1 to 3. Clone frequencies were estimated by the immunoSEQ assay using in-frame, expressed gDNA sequences. The frequency distribution of the full expressed repertoire (including unpaired TCRBs) is shown in gray.




one lung; Table 2) and matched blood samples. The clonal diversity of TILs is significantly reduced compared to blood: typically, only tens or hundreds of the most expanded TILs are of primary interest, so we leverage the high throughput of pairSEQ to run many samples in a single experiment. Experiments 1 and 2 demonstrated that samples can be multiplexed on a single plate for pairing; here, we extend that approach to pair the nine TIL and nine blood samples using one plate for each sample type.

We started by using the immunoSEQ assay to sequence the TCRA and TCRB repertoire of a portion of each sample to determine the T cell content, the TCRA/TCRB repertoire sequences, and the clonal frequencies. We then used our study design algorithm to determine how many cells should be collected from each sample to pair the TCR sequences of common clones. We dissociated the tumor cells and mixed the specified numbers of cells in two batches: one for tumor samples and one for PBMC samples. The tumor cell and PBMC mixtures were then distributed across separate 96-well plates (Table 1, experiments 4 and 5, respectively). Finally, after running the pairSEQ assay and identifying the cognate pairs on each plate, we assigned the pairs to samples of origin by finding exclusive matches between paired sequences and sample-specific repertoire sequences.

In total, we identified 6172 pairs from TILs in experiment 4 and 14,123 pairs from peripheral blood in experiment 5 (Table 2). Of these, 3284 pairs from experiment 4 and 7492 pairs from experiment 5 were unambiguously assigned to samples of origin. To ensure that multiplexing a larger number of samples did not inflate the FDR, we repeated the cross-pairing analysis used in experiment 1 and found that at a target FDR of 1%, the cross-pairing FDR was 0.7% in experiment 4 and 1.2% in experiment 5. We also assessed the consistency of pairing among 326 TCRB sequences that were paired in both tumor and matched PBMC. All but six of these sequences (1.8%) were paired with consistent TCRA sequences across sample types. This experiment confirms three important features of the pairSEQ experimental framework. First, when sampling depth in each subject is not of vital importance, for example, when profiling hundreds of TIL clones, the high-throughput nature of pairSEQ can be harnessed to efficiently profile many samples at lower comparative depth. Second, pairSEQ can perform well even with difficult sample types (such as solid tumor biopsies). Finally, concordant TCRA/TCRB pairs can be generated from different tissues. This last point is especially important when cell populations of interest are too small to generate a significant pair in a pairSEQ experiment (for example, TILs from a core needle biopsy) or when intact cells cannot be obtained (for example, formalin-fixed paraffin-embedded tumor tissue). In these cases, standard immunosequencing can be used to profile TCRB with near-perfect sensitivity, and much larger amounts of a different tissue (for example, peripheral blood) can be subjected to pairSEQ to identify the paired TCRA sequences.

**Table 2. Pairing results from nine tumors and matched PBMC samples.**

| Tumor sample | Tumor pairs | PBMC pairs | Pairs with same TCRB | Pairs with same TCRB and TCRA |
|---|---|---|---|---|
| Breast 1 | 6 | 13 | 0 | 0 |
| Breast 2 | 337 | 95 | 19 | 19 |
| Breast 3 | 188 | 1,782 | 67 | 66 |
| Breast 4 | 185 | 4,906 | 66 | 62 |
| Kidney 1 | 189 | 186 | 24 | 23 |
| Kidney 2 | 364 | 261 | 77 | 77 |
| Kidney 3 | 509 | 53 | 19 | 19 |
| Kidney 4 | 1,166 | 33 | 19 | 19 |
| Lung 1 | 340 | 163 | 35 | 35 |
| Total | 3,284* | 7,492† | 326 | 320 |

*A total of 2888 pairs called at FDR = 1% are not included in this count because they could not be unambiguously assigned to samples of origin. †A total of 6631 pairs called at FDR = 1% are not included in this count because they could not be unambiguously assigned to samples of origin.

## DISCUSSION

We have demonstrated the high accuracy and throughput of pairSEQ by identifying TCR pairs from a wide range of clonal frequencies in multiple sample types, including more than 200,000 pairs from a single 96-well plate. We anticipate that this technology will prove useful for many applications in immunology and translational medicine and in the development of T cell and antibody therapeutics. pairSEQ also yields massive volumes of data that can be used to investigate the connections between the primary sequence of adaptive immune receptors and their associated antigenic targets. With attention to experimental design and control of variables such as human leukocyte antigen type, pairSEQ could be a useful approach for establishing new criteria for responsiveness to routine or experimental vaccination and for epidemiological analyses of public exposures and shared responses. As is, pairSEQ can be used to identify the TCRs of TILs, and the resulting information can be used to engineer T cells to express tumor antigen–targeting receptors.

This technology is intentionally designed to be simple and scalable and thus accessible to a variety of laboratories; the pairSEQ assay can be run with standard laboratory supplies and equipment without the need for specialized expertise and is applicable to a broad potential range of starting sample types, including tumor samples, sorted cells, and cells in suspension. The main limitation of this approach is that it requires multiple cells from a given clone to recover its cognate pair: this technology is therefore not ideal for pairing clones that are rare in a particular sample, although this limitation can be overcome by taking a larger sample from the same person (such as a matched blood sample to accompany a small tumor biopsy, as discussed above in the section "Analysis of tumor-infiltrating lymphocytes").

In this work, we have used our analysis of TCRα/β as the proof of concept for the pairSEQ method. However, this approach should also work, with minimal changes to the experimental protocol, for TCRγ/δ. In addition, the analytical framework we have presented for combinatoric pairing should also work in principle for linking the immunoglobulin heavy and light chains of B cells. This will require the ability of high-throughput sequencers to generate adequate read lengths to support sequencing of the full variable regions (because the CDR3 region is not enough to fully characterize an antibody), as well as the design of experiments (for example, sorting or expansion protocols) to ensure that B cell clones of interest can be obtained at enough copies to efficiently pair. Given the practical interest in developing monoclonal antibodies, as well as the general importance of the humoral immune response, pairSEQ has the potential to become a foundational technology for biomedical discovery.

## MATERIALS AND METHODS

### Sample collection

Blood was collected in 10-ml EDTA blood tubes (Beckton Dickinson) from two healthy adults (X and Y) with written informed consent. PBMCs were isolated using SepMate tubes (STEMCELL Technologies), and red blood cells were lysed using 1× Red Blood Cell Lysis Solution (Miltenyi Biotec). The remaining cells were washed in phosphate-buffered saline, suspended in RNA*later* (Qiagen), and counted using a TC20 cell counter (Bio-Rad). Cell suspensions were adjusted relative to the experimental design using RNA*later* and then stored at 4°C until use.

As a positive control for pairing, a human T cell leukemia (T-ALL) cell line known to be positive for αβ TCR, Jurkat clone E6-1 (ATCC TIB-152) (*35*), was spiked into the cell suspension of experiment 1 at a ratio of 1 Jurkat cell per ~30,000 total cells. The TCRA and TCRB sequences of Jurkat cells were identified by deep sequencing (*14*, *34*).

### T cell counting for experimental design

To identify the optimal number of cells for each experiment, gDNA was extracted from cells using a DNeasy kit (Qiagen), followed by amplification and sequencing of the TCRB and TCRAD (TCR αδ) repertoire of each sample using the immunoSEQ assay (Adaptive Biotechnologies) and comparison to the IMGT (ImMunoGeneTics) database of immune genes (*39*). These data were used to estimate the abundance of individual clones and to estimate the proportion of T cells among total cells. Finally, these values were used to normalize the PBMCs to a constant concentration of T cell/μl to control for variation in the proportion of T cells in different biological samples.

### Amplification of cognate pairs

To initiate a pairSEQ process, cells were distributed into wells of 0.8-ml-deep 96-well plates (Axygen). RNA was isolated from each well using the QIAsymphony RNA kit and an automated protocol on the QIAsymphony laboratory robot (Qiagen) and eluted into another 96-well plate in 100 μl of ribonuclease (RNase)–free water and immediately stabilized with RNAstable (Biomatrica). RNA was concentrated using a vacuum centrifuge and re-eluted in 15 μl of RNase-free water. RNA was reverse-transcribed into single-stranded cDNA using the SuperScript VILO cDNA synthesis kit (Invitrogen) according to the manufacturer's protocol. Throughout the process, well identity was maintained relative to the original source plate.

The cDNA was used as a template for multiplex PCR using TCRA and TCRB gene-specific primers. The PCR uses multiple V gene–specific primers for TCRA and TCRB and a single C segment reverse primer (Integrated DNA Technologies) for each receptor (for the lists of PCR primers used, see tables S1 and S2), and Multiplex PCR Plus reagents (Qiagen) in a 50-μl reaction. Reactions were placed on Bio-Rad thermal

cyclers for 30 cycles of PCR amplification under the following conditions: 95°C denaturation for 300 s, followed by 30 cycles of denaturation at 95°C for 30 s, annealing at 59.3°C for 70 s, and extension at 72°C for 90 s.

To attach well-specific barcodes, PCR products were purified using SPRIselect magnetic beads (Beckman Coulter), and a 2-μl aliquot of the PCR reaction was used in a second PCR. The eight cycles of the second PCR incorporated, at each end of the amplicon, Illumina paired-end adapter sequences and an eight–base pair well-specific DNA sequence. This allows the samples to be sequenced on Illumina MiSeq or HiSeq (Illumina), and the DNA barcodes allow mapping of sequences to individual wells (*34*). Table S3 includes the complete list of the barcodes used. After PCR amplification, 5-μl aliquots were pooled for all 96 wells, and the molarity of the product was measured using the Caliper DNA 1K assay (PerkinElmer).

Sequencing libraries were loaded onto the Illumina HiSeq 2500 sequencer flow cell or in an Illumina MiSeq as per Illumina protocol, for on-board cluster generation using a Rapid Run format (Illumina). We sequenced in a single direction, using 15 cycles to capture the well-specific barcode and 150 cycles primed using the TCRA and TCRB C primers to sequence through the V(D)JC receptor domain. Depending on the number of cells per well, for some cases, multiple sequencing runs were necessary to achieve adequate sequencing depth.

To identify the lower range of cells per well on which we could efficiently perform RNA extraction, cDNA synthesis, and amplification, we conducted a series of control experiments. Naïve T cells were isolated from PBMC with a Naïve Pan T Cell Isolation Kit (Miltenyi Biotec). Cells were counted, normalized, and dispensed into wells across a dilution gradient ranging from 10,000, 2000, 400, 80, 16, 3.2, to 0 cells per well. Using the same protocol described above, RNA was isolated, and cDNA was synthesized. The unique TCRB and TCRA sequence count values were used as a proxy for cell capture. We validated our method down to 80 cells per well.

### FDR validation experiment (experiment 1)

To directly measure the FDR for pairing, we isolated PBMCs from the venous blood of two healthy adults using standard techniques, as described above. To assign TCRB and TCRA CDR3 sequences to each subject, gDNA was isolated from a subset of collected blood, and for each subject, 89.6 μg of DNA (~14 million genomes) was amplified using the TCRA and TCRB immunoSEQ assay (Adaptive Biotechnologies). After this test, the remaining cells from each subject were normalized to 5000 cells/μl.

The immunoSEQ assay provides accurate estimates of T cell abundance in a sample; for these subjects, T cells comprised ~45% of the PBMC samples. To empirically measure the FDR of our pairing method in an experiment with 2000 T cells per well per subject, the subjects' PBMCs were mixed at a 1:1 ratio, and ~900,000 total cells (~400,000 total T cells, ~200,000 T cells per subject) were equally distributed across 96 wells. Before distributing the cells, Jurkat cells were spiked in to the PBMC mixture at about 1 Jurkat cell per 30,000 total cells, with the goal of allocating ~30 Jurkat cells to the plate. Digital cell counting based on well occupancy on the plate provided post hoc confirmation that the number of allocated T cells was close to the target number. To identify TCR pairs, the pairSEQ process as described above was applied to the entire plate. Pairs in the combined plate were classified as belonging to subject 1, subject 2, or mixed based on the immunoSEQ data.

### High-throughput pairing experiments (experiments 2 and 3)

To demonstrate the ability of this method to identify both a large number and a large range of TCR pairs, we set up two additional experiments to identify the TCR pairs of rarer peripheral T cell clones. For both experiments, 16 million T cells in a suspension of PBMCs (36 million total cells) were distributed across a 96-well plate (160,000 T cells per well). For experiment 2, the cell suspension was mixed so that 80,000 T cells each from subject X and subject Y were added to each well, whereas for experiment 3, all T cells were from subject X. The number of cells per subject in each well was selected to identify clones that were present at a proportion of 1 cell in 10,000 or lower, as described in the section "Design principles for pairSEQ experiments." Each plate was processed using the pairSEQ process, as described above, and the pairs were identified using the method described below in the section "Design principles for pairSEQ experiments."

### Matched tumor and blood experiments (experiments 4 and 5)

To confirm feasibility in sample types other than blood, we identified receptor pairs from TILs in tumors and matched blood. Nine fresh tumors and matched blood were collected from consented patients through Conversant Bio. For each tumor, cells were disassociated using a gentleMACS Dissociator and a tumor dissociation kit (Miltenyi Biotec), PBMCs were separated from whole blood, and cell counts were obtained as described in the T cell counting for experimental design section.

To design an optimal pairSEQ experiment, as described in the "Design principles for pairSEQ experiments" section below, we first measured the abundance of T cell clones. Then, from each sample, we extracted gDNA from about 1 million cells and used the immunoSEQ TCRB and TCRAD assay (Adaptive Biotechnologies) to amplify and sequence the TCR repertoire from about 250,000 cells from each sample. These data were then used to design the corresponding pairSEQ experiment. Cells from the nine tumor samples and nine blood samples were mixed and then distributed across the 96-well plate. After this, the standard pairSEQ process was applied to the samples. Individual clones were then identified using the CDR3 chains from the immunoSEQ data and assigned to their sample.

### Analysis

**Statistics for TCR pairing.** For every possible TCRA/TCRB pair, we computed a *P* value for the observed number of shared wells. Consider a TCRA sequence that occupies $w^{(\alpha)}$ wells and a TCRB sequence that occupies $w^{(\beta)}$ wells. If these sequences share $w^{(\alpha\beta)}$ wells and each well contains the same number of T cells, the probability of seeing this amount of well sharing by chance is

$$P\left(w^{(\alpha\beta)} | w^{(\alpha)}, w^{(\beta)}, W\right) = \frac{\binom{W}{w^{(\alpha\beta)}}\binom{W-w^{(\alpha\beta)}}{w^{(\alpha)}-w^{(\alpha\beta)}}\binom{W-w^{(\alpha\beta)}-w^{(\alpha)}}{w^{(\beta)}-w^{(\alpha\beta)}}}{\binom{W}{w^{(\alpha)}}\binom{W}{w^{(\beta)}}} \quad (1)$$

conditional on the total number of wells $W$ and the marginal well counts $w^{(\alpha)}$ and $w^{(\beta)}$. A *P* value can be obtained for a putative TCR pair by summing the probabilities for all well configurations that have the same marginal counts and an equal or greater overlap of occupied wells.

Computing *P* values in this way serves two purposes: it accounts for the fact that TCR sequences in different numbers of wells have

different probabilities of overlapping by chance, and it captures departures from chance pairing without requiring perfect overlap between occupied wells. Imperfect well overlap is common among the members of a TCR pair because low numbers of mRNA transcripts per cell can lead to well dropouts (see section "Well dropout rates in cDNA subsamples" below), so it is essential to use a statistic that accounts for this feature of the data.

**Simulating null *P* value distributions.** Unlike many applications in biological data analysis, the null distribution of *P* values in a pairSEQ analysis is neither continuous nor uniform: discrete well occupancy patterns lead to discrete *P* values, and the *P* value for each TCR sequence is chosen as the smallest seen in many comparisons with possible pairing partners, which skews the null distribution toward smaller *P* values. TCR clones with different well occupancies also have different null distributions: more extreme *P* values can be observed in clones that occupy a moderate number of wells, but such wells also tend to be involved in relatively few comparisons that could generate extreme *P* values.

To account for these features of the data, we developed a permutation algorithm that fully models the structure of a pairSEQ experiment. We designated one locus (typically TCRA) as the "query" locus and the other as the "target" locus. After counting the number of target sequences $T_i$ that occupy $i$ wells for $i = 1,\ldots,96$, a single permutation involves the following steps:

1) For each occupancy level $i$, sample $T_i$ random numbers in [0,1]. Denote the largest sampled number $\gamma_i$. In practice, this step can be performed quickly by sampling from an extreme value distribution.

2) For each occupancy level $j$, use $\gamma_i$ to determine the number of shared wells $N_{ij}$ in a cumulative distribution function for sequences that occupy $i$ and $j$ wells. This represents the largest number of wells that a sequence in $j$ wells would share with the best sequence in $i$ wells after sampling $T_i$ random configurations.

3) Use $N_{ij}$ to compute a *P* value $\delta_j$ for occupancy level $j$. If $\delta_j$ is smaller than the smallest *P* value seen so far at level $j$, store it.

Once these steps have been completed for every number of occupied target wells $i$, we will have a minimum *P* value $\delta_j$ for each possible number of query wells $j$. These values are stored as the outcome of one permutation, and they are directly analogous to the smallest *P* values seen when comparing a given query sequence against a collection of target sequences. All of the results in this manuscript are based on running 10,000 permutations for each pairSEQ experiment.

**FDR estimation.** The simulated null distributions described above can be used to estimate FDRs in a pairSEQ experiment. Clones at each well occupancy level have different null *P* value distributions, so they also require separate FDR estimates. Our approach follows that of Bancroft *et al.* (*40*), who developed a method for estimating FDRs with sequential permutation *P* values. Like the *P* values in a pairSEQ experiment, sequential permutation *P* values are discrete and nonuniform, so their method is directly applicable to our scenario. The main steps of the method are as follows:

1) For each well occupancy level $i = 1,\ldots,96$, construct as many bins as possible that have a probability of at least 0.05 under the null. This accounts for discontinuities in discrete distributions that can lead to overconservative estimates of the number of truly null hypotheses.

2) For each well occupancy level $i$, estimate the number of null hypotheses $m_{0i}$ among the observed *P* values via theorem 2 of Bancroft *et al.* (*40*), which is related to histogram-based estimators of $m_0$.

3) Compute the FDR at any rejection threshold $\alpha$ for *P* values at occupancy level $i$. To get experiment-wide FDR estimates like the ones shown in Figs. 2 and 4, sum the numbers of expected type I errors and total pairing calls across occupancy levels and divide the former sum by the latter.

Estimates of $m_0$ can be unstable when the number of observed *P* values is small, as can happen for occupancy levels with few TCR sequences. To get more reliable estimates, we combine occupancy levels that have small sequence counts and similar null distributions.

**Design principles for pairSEQ experiments.** The TCR clones of interest in a pairSEQ experiment will depend on the application. In some cases, the goal may be to pair sequences from the most common clones in a T cell repertoire, whereas in other cases, one may aim to pair a much larger set of low-frequency clones. The scope of a study may further be constrained by expenses or logistics. The pairSEQ assay can be tailored in multiple ways to address these variables.

The key design considerations, as illustrated by simulations based on a real TCR repertoire (from "subject X" in this study) are as follows:

1) The number of T cells allocated to each distinct subset (that is, well on a microtiter plate) determines the expected numbers of wells that will be occupied by clones of different repertoire frequencies (fig. S1A).

2) Clones that appear near the edge of the range of possible well occupancies—say, in just one well or in all available wells—are not "pairable" in the sense that other clones are likely to reside in exactly the same combinations of wells (fig. S1B). Pairing rates change slightly with the number of T cells per well (for example, 10,000 versus 50,000 cells per well; orange and dark blue, respectively), but the main point is that virtually every clone that occupies an intermediate number of wells can theoretically be paired. This is because the number of possible combinations increases rapidly with the number of occupied wells; for example, on a 96-well plate, there are $\binom{96}{3} \approx 143{,}000$ unique 3-well combinations, as compared to $\binom{96}{10} \approx 11.3$ trillion 10-well combinations.

3) For a fixed number of wells, the number of input T cells affects both the total number of pairable sequences and their frequency distribution (fig. S1C). Allocating more T cells to each well can greatly increase the number of called pairs; in this example of 10,000 (orange) versus 50,000 (dark blue) T cells per well, the latter experiment paired more than eight times as many TCRs. Conversely, experiments with fewer cells per well are better suited for pairing common clones; in this example, the experiment with 10,000 T cells per well captured 71 of the 100 most frequent clones, as compared to 2 of 100 for the experiment with a larger number of T cells. A hybrid experiment with different numbers of T cells allocated to different wells would yield an intermediate number of pairs across a broader frequency range than seen in either of these experiments alone.

4) Varying the total number of wells while adding a constant number of T cells to each well affects both the number of pairable sequences and their frequency range (fig. S1D). An experiment with a larger number of wells—96 (dark blue) versus 48 (light blue)—produced about four times as many paired TCRs in this example. Both experiments captured a similar fraction of common clones (2 and 1 of the 100 most frequent clones, respectively), but the 96-well design paired many more TCRs at lower frequencies. These trends were driven by three features of the experiment with a larger number of wells. First, increasing the total number of input T cells increased the number of clones available for pairing. Second, increasing the number of wells shifted rarer (and hence more numerous) clones into the "intermediate" well occupancy range, which has high pairing rates. Third, using

a larger total number of wells generated more possible combinations of each number of occupied wells, so that two clones seen in the same number of wells were less likely to exactly coincide on the plate.

The results and figures in this section are based on simulations from a real immune repertoire but without modeling the experimental conditions that can make it harder to pair TCR sequences in practice. Hence, these results should be treated as useful intuitions about how to design a pairSEQ experiment and not as quantitative predictions.

**Study design algorithm.** To design pairSEQ experiments that maximize the chances of pairing clones of interest, we developed an optimization algorithm that uses TCR repertoire data to generate custom pairSEQ design parameters (number of wells and number of T cells per well) for each sample. The main inputs to the algorithm are a set of repertoire clone frequencies (typically from TCRB) and the range of frequencies we want to capture. The algorithm assumes that the repertoire frequencies represent all clones in the target frequency range, and it aims to find parameters that will theoretically pair all of the clones in this range with high probability.

A key quantity in this algorithm is the probability of pairing a clone of any given frequency under a proposed study design with $W$ wells and $C$ input T cells ($C/W$ cells per well); we denote the probability of pairing clone $i$ as $\mathrm{P}(\mathrm{pair}_i \mid C, W)$. To compute this quantity, we begin by calculating the binomial probability of sampling $c_i$ cells from clone $i$ with repertoire frequency $f_i$:

$$P(c_i|f_i,C) = \binom{C}{c_i} f_i^{c_i}(1-f_i)^{C-c_i} \tag{2}$$

Some of the sampled cells may fail to produce any cDNA sequences because of experimental inefficiencies (for example, mRNA decay during sample preparation), so we model each clone as having a per-cell censoring probability drawn from a beta distribution with shape parameters $s_1$ and $s_2$. The beta parameters are fitted by maximum likelihood to data from an experiment that measures the attrition rates in our assay (see the section "Well dropout rates in mRNA subsamples" below). If we allow separate censoring parameters for each TCR locus $x$ in {α,β}, the probability that $c_i^{(x)}$ cells will produce a signal from locus $x$ can be computed using a beta binomial distribution:

$$P\left(c_i^{(x)}|c_i,s_1,s_2\right) = 1 - \binom{c_i}{c_i^{(x)}} \frac{\mathrm{B}(c_i^{(x)}+s_1, c_i-c_i^{(x)}+s_2)}{\mathrm{B}(s_1,s_2)} \tag{3}$$

where $\mathrm{B}(x,y)$ is the beta function.

Sampled cells are partitioned into a fixed number of wells in a pairSEQ experiment, and the probability that $c_i$ cells will occupy $w_i$ out of $W$ wells is

$$P(w_i|c_i,W) = \frac{W!\left\{ {c_i \atop w_i} \right\}}{(W-w_i)!W^{c_i}} \tag{4}$$

where $\left\{ {c_i \atop w_i} \right\}$ is a Stirling number of the second kind.

If the sampled cells from clone $i$ occupy $w_i$ wells, the censored sequences from locus $x$ must occupy a censored set of $w_i^{(x)}$ wells, with probability

$$P\left(w_i^{(x)}|c_i^{(x)},W\right) = \frac{W!\left\{ {c_i^{(x)} \atop w_i^{(x)}} \right\}}{(W-w_i^{(x)})!W^{c_i^{(x)}}} \tag{5}$$

Conditional on the censored well occupancies, the probability of pairing depends on the amount of overlap between wells that contain noncensored sequences. The probability that a clone's TCRA and TCRB chains will jointly occupy $w_i^{(\alpha\beta)}$ wells is

$$P\left(w_i^{(\alpha\beta)}|w_i^{(\alpha)},w_i^{(\beta)},w_i\right) = \frac{\binom{w_i}{w_i^{(\alpha\beta)}}\binom{w_i-w_i^{(\alpha\beta)}}{w_i^{(\alpha)}-w_i^{(\alpha\beta)}}\binom{w_i-w_i^{(\alpha\beta)}-w_i^{(\alpha)}}{w_i^{(\beta)}-w_i^{(\alpha\beta)}}}{\binom{w_i}{w_i^{(\alpha)}}\binom{w_i}{w_i^{(\beta)}}} \tag{6}$$

Finally, we must compute the probability that clone $i$ will be successfully paired at a given FDR. The FDR is a joint property of the collection of clones in a pairSEQ experiment—for example, the pairing probability for clone $i$ depends not only on the amount of overlap between the censored wells for that clone but also on the number of other clones that have similar enough occupancies to potentially create a false pairing. If we denote the full set of observed TCRA occupancies as $\Theta^{(\alpha)}$, the full set of TCRB occupancies as $\Theta^{(\beta)}$, and the event that clone $i$ satisfies a given FDR threshold as $\delta_i$, the probability that clone $i$ will be correctly paired can be written as $\mathrm{P}[\mathrm{pair}_i, \delta_i \mid w_i^{(\alpha\beta)}, w_i^{(\alpha)}, w_i^{(\beta)}, W, \Theta^{(\alpha)}, \Theta^{(\beta)}]$. This probability does not have a simple analytical form, but we can get a good approximation by simulating pairSEQ experiments with the same input parameters.

Putting all of these pieces together, we can compute the overall pairing probability for clone $i$ as follows:

$$\begin{aligned} P(\mathrm{pair}_i|C,W) = {} & \sum_{c_i=0}^{C} P(c_i|f_i,C) \sum_{w_i=0}^{\min(c_i,W)} P(w_i|c_i,W) \\ & \times \sum_{c_i^{(\alpha)}=0}^{c_i} P(c_i^{(\alpha)}|c_i,s_1^{(\alpha)},s_2^{(\alpha)}) \sum_{c_i^{(\beta)}=0}^{c_i} P(c_i^{(\beta)}|c_i,s_1^{(\beta)},s_2^{(\beta)}) \\ & \times \sum_{w_i^{(\alpha)}=0}^{w_i} P(w_i^{(\alpha)}|c_i^{(\alpha)},W) \sum_{w_i^{(\beta)}=0}^{w_i} P(w_i^{(\beta)}|c_i^{(\beta)},W) \\ & \times \sum_{w_i^{(\alpha\beta)}=\max(0,w_i^{(\alpha)}+w_i^{(\beta)}-w_i)}^{\min(w_i^{(\alpha)}w_i^{(\beta)})} P(w_i^{(\alpha\beta)}|w_i^{(\alpha)},w_i^{(\beta)},w_i) \end{aligned}$$

$$P(\mathrm{pair}_i,\delta_i|w_i^{(\alpha\beta)},w_i^{(\alpha)},w_i^{(\beta)},W,\Theta^{(\alpha)},\Theta^{(\beta)}) \tag{7}$$

This calculation can be made tractable by limiting the ranges of the summations to values with reasonable probabilities of occurring—for example, many possible values of the number of sampled cells $c_i$ have essentially zero binomial probability, so they can be skipped.

With the ability to calculate the probability of pairing any given clone in an experiment with $C$ total input cells and $W$ wells, we typically fix the number of wells to model a standard experimental plate ($W = 96$) and use a simple optimization to find a value $C$ that either (i) pairs all repertoire clones in the target frequency range with some minimum probability or (ii) pairs a specified expected number or fraction of clones. The optimal number of input cells is then used to design a real experiment whose results can be compared against the per-clone pairing probabilities from this model.

**Well dropout rates in cDNA subsamples.** The pairSEQ method depends on reliably detecting the TCRA and TCRB sequences from a T cell clone in each well it occupies. A typical T cell may carry just a few mRNA copies (possibly just 5 to 10) of each TCR rearrangement,

and biological variation or experimental inefficiencies can cause either locus to go unobserved in a subset of wells. To characterize this well dropout rate, we designed a pairSEQ experiment in which PBMCs were distributed across a 96-well plate for RNA extraction and cDNA synthesis (as usual) and then half of the volume was pipetted to another plate. The well configuration was maintained while transferring the solution from the first plate to the second, so each plate represented a technical pairSEQ replicate with half the usual amount of cDNA.

After running the rest of the pairSEQ assay on each plate, we measured the well dropout rate as follows: for each unique sequence, we treated the union of occupied wells across plates as the true set of wells that received that clone. We then determined the well dropout rate for each clone by comparing the single-plate occupancy with the union occupancy. TCR sequences that appear in less than 40 wells are likely to arise from one cell per well, so by focusing on union occupancies less than 40, we were able to estimate per-cell dropout rates.

The results of this experiment are shown in fig. S3; fig. S3A shows the per-cell dropout rates from the first pairSEQ plate, and fig. S3B shows the equivalent rates from the second plate. The dashed vertical lines show the median dropout rates (9 and 5%, respectively), and the black curves depict beta distributions fitted to the data. Most TCR sequences have less than a 10% chance of going unobserved, but a small fraction of sequences suffer significant attrition.

The cDNA attrition rates in fig. S3 were achieved after many rounds of assay optimization, and the pairSEQ method is robust enough to provide reliable results with this amount of noise. To account for this feature of the experiment as fully as possible, we have incorporated the fitted beta distributions into our algorithm for designing pairSEQ experiments, as discussed in the "Study design algorithm" section above.

**Assigning paired TCRs to samples of origin in multiplexed pairSEQ experiments.** In this work, we have presented a number of experiments in which multiple samples were combined on a single pairSEQ plate to increase efficiency. After calling the pairs on a given plate, we traced them back to samples of origin by computationally matching the paired sequences with sample-specific repertoire sequences. Regardless of the number of samples in a given experiment, we assigned pairs to samples according to the following rules:

1) If the TCRB sequence in a pair matches a sequence in exactly one repertoire, assign the pair to the corresponding sample.

2) If the TCRB sequence in a pair matches the repertoires of multiple samples or none at all, check the TCRA sequence for unique matches and assign the pair to a sample if any such match is found.

3) If neither of the above steps produces a match, declare that pair to have an "ambiguous" sample of origin. Pairs with ambiguous origins may be resolved by sequencing more cells from the sample repertoires; alternatively, the number of such pairs can be limited by targeting only common clones in multiplexed experiments.

## SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/7/301/301ra131/DC1
Fig. S1. Basic design parameters in a pairSEQ experiment.
Fig. S2. Predicted versus empirical $\log_{10}$ FDR in experiment 2.
Fig. S3. Per-cell dropout rates in mRNA subsamples.
Table S1. Sequences of TCRAD and TCRB V gene forward PCR primers.
Table S2. Sequences of TCRA and TCRB C gene reverse PCR primers.
Table S3. Sequences of DNA barcodes.

## REFERENCES AND NOTES


1. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
2. N. R. Gascoigne, Y.-H. Chien, D. M. Becker, J. Kavaler, M. M. Davis, Genomic organization and sequence of T-cell receptor β-chain constant-and joining-region genes. *Nature* **310**, 387–391 (1984).
3. J. Nikolich-Žugich, M. K. Slifka, I. Messaoudi, The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* **4**, 123–132 (2004).
4. G. Siu, S. P. Clark, Y. Yoshikai, M. Malissen, Y. Yanagi, E. Strauss, T. W. Mak, L. Hood, The human T cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene. *Cell* **37**, 393–401 (1984).
5. B. Toyonaga, Y. Yoshikai, V. Vadasz, B. Chin, T. W. Mak, Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor β chain. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 8624–8628 (1985).
6. Y. Yoshikai, S. P. Clark, S. Taylor, U. Sohn, B. I. Wilson, M. D. Minden, T. W. Mak, Organization and sequences of the variable, joining and constant region genes of the human T-cell receptor α-chain. *Nature* **316**, 837–840 (1985).
7. S. D. Boyd, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, A. Z. Fire, Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
8. M. L. Davila, I. Riviere, X. Wang, S. Bartido, J. Park, K. Curran, S. S. Chung, J. Stefanski, O. Borquez-Ojeda, M. Olszewska, J. Qu, T. Wasielewska, Q. He, M. Fink, H. Shinglot, M. Youssif, M. Satter, Y. Wang, J. Hosey, H. Quintanilla, E. Halton, Y. Bernal, D. C. Bouhassira, M. E. Arcila, M. Gonen, G. J. Roboz, P. Maslak, D. Douer, M. G. Frattini, S. Giralt, M. Sadelain, R. Brentjens, Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia. *Sci. Transl. Med.* **6**, 224ra225 (2014).
9. M. Faham, J. Zheng, M. Moorhead, V. E. Carlton, P. Stow, E. Coustan-Smith, C. H. Pui, D. Campana, Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* **120**, 5173–5180 (2012).
10. J. Glanville, W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. Day, D. Cox, A. Rajpal, J. Pons, Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 20216–20221 (2009).
11. S. A. Grupp, M. Kalos, D. Barrett, R. Aplenc, D. L. Porter, S. R. Rheingold, D. T. Teachey, A. Chew, B. Hauck, J. F. Wright, M. C. Milone, B. L. Levine, C. H. June, Chimeric antigen receptor–modified T cells for acute lymphoid leukemia. *N. Engl. J. Med.* **368**, 1509–1518 (2013).
12. K. Larimore, M. W. McCormick, H. S. Robins, P. D. Greenberg, Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–3230 (2012).
13. H. Robins, C. Desmarais, J. Matthis, R. Livingston, J. Andriesen, H. Reijonen, C. Carlson, G. Nepom, C. Yee, K. Cerosaletti, Ultra-sensitive detection of rare T cell clones. *J. Immunol. Methods* **375**, 14–19 (2012).
14. H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, C. S. Carlson, Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).
15. H. S. Robins, S. K. Srivastava, P. V. Campregher, C. J. Turtle, J. Andriesen, S. R. Riddell, C. S. Carlson, E. H. Warren, Overlap and effective size of the human $CD8^+$ T cell receptor repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).
16. A. M. Sherwood, C. Desmarais, R. J. Livingston, J. Andriesen, M. Haussler, C. S. Carlson, H. Robins, Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after αβ and γδ T cell commitment. *Sci. Transl. Med.* **3**, 90ra61(2011).
17. R. L. Warren, J. D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb, R. A. Holt, 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.* **21**, 790–797 (2011).
18. R. L. Warren, B. H. Nelson, R. A. Holt, Profiling model T-cell metagenomes with short reads. *Bioinformatics* **25**, 458–464 (2009).
19. W. K. Weng, R. Armstrong, S. Arai, C. Desmarais, R. Hoppe, Y. H. Kim, Minimal residual disease monitoring with high-throughput sequencing of T cell receptors in cutaneous T cell lymphoma. *Sci. Transl. Med.* **5**, 214ra171 (2013).
20. D. Wu, A. Sherwood, J. R. Fromm, S. S. Winter, K. P. Dunsmore, M. L. Loh, H. A. Greisman, D. E. Sabath, B. L. Wood, H. Robins, High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* **4**, 134ra63 (2012).
21. C. E. Busse, I. Czogiel, P. Braun, P. F. Arndt, H. Wardemann, Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* **44**, 597–603 (2014).
22. N. Chapal, M. Bouanani, M. J. Embleton, I. Navarro-Teulon, M. Biard-Piechaczyk, B. Pau, S. Peraldi-Roux, In-cell assembly of scFv from human thyroid-infiltrating B cells. *Biotechniques* **23**, 518–524 (1997).

23. P. Dash, J. L. McClaren, T. H. Oguin III, W. Rothwell, B. Todd, M. Y. Morris, J. Becksfort, C. Reynolds, S. A. Brown, P. C. Doherty, P. G. Thomas, Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J. Clin. Invest.* **121**, 288–295 (2011).
24. B. J. DeKosky, G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, G. Georgiou, High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166–169 (2013).
25. B. J. DeKosky, T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, G. Georgiou, In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2015).
26. M. Embleton, G. Gorochov, P. T. Jones, G. Winter, In-cell PCR from mRNA: Amplifying and linking the rearranged immunoglobulin heavy and light chain V-genes within single cells. *Nucleic Acids Res.* **20**, 3831–3837 (1992).
27. A. Han, J. Glanville, L. Hansmann, M. M. Davis, Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
28. S. M. Kim, L. Bhonsle, P. Besgen, J. Nickel, A. Backes, K. Held, S. Vollmer, K. Dornmair, J. C. Prinz, Analysis of the paired TCR α- and β-chains of single human T cells. *PLOS One* **7**, e37338 (2012).
29. P. J. Meijer, P. S. Andersen, M. Haahr Hansen, V. Steinaa, A. Jensen, J. Lantto, M. B. Oleksiewicz, K. Tengbjerg, T. R. Poulsen, V. W. Coljee, S. Bregenholt, J. S. Haurum, L. S. Nielsen, Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *J. Mol. Biol.* **358**, 764–772 (2006).
30. X. Sun, M. Saito, Y. Sato, T. Chikata, T. Naruto, T. Ozawa, E. Kobayashi, H. Kishi, A. Muraguchi, M. Takiguchi, Unbiased analysis of TCRα/β chains at the single-cell level in human $CD8^+$ T-cell subsets. *PLOS One* **7**, e40386 (2012).
31. Y.-C. Tan, L. K. Scalfone, S. Kongpachith, C.-H. Ju, X. Cai, T. M. Lindstrom, J. Sokolove, W. H. Robinson, Sequencing antibody repertoires provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin. Immunol.* **151**, 55–65 (2014).
32. M. A. Turchaninova, O. V. Britanova, D. A. Bolotin, M. Shugay, E. V. Putintseva, D. B. Staroverov, G. Sharonov, D. Shcherbo, I. V. Zvyagin, I. Z. Mamedov, C. Linnemann, T. N. Schumacher, D. M. Chudakov, Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).
33. J. D. Ashwell, A. Weissman, in *Clinical Immunology: Principles and Practice*, R. R. Rich, T. A. Fleisher, W. T. Shearer, B. L. Kotzin, H. W. Schroeder Jr., Eds. (Mosby International Limited, London, 2001), chap. 5, pp. 5.1–5.19.
34. C. S. Carlson, R. O. Emerson, A. M. Sherwood, C. Desmarais, M. W. Chung, J. M. Parsons, M. S. Steen, M. A. LaMadrid-Herrmannsfeldt, D. W. Williamson, R. J. Livingston, D. Wu, B. L. Wood, M. J. Rieder, H. Robins, Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
35. Y. Sandberg, B. Verhaaf, E. J. van Gastel-Mol, I. L. Wolvers-Tettero, J. de Vos, R. A. Macleod, J. G. Noordzij, W. A. Dik, J. J. van Dongen, A. W. Langerak, Human T-cell lines with well-defined T-cell receptor gene rearrangements as controls for the BIOMED-2 multiplex polymerase chain reaction tubes. *Leukemia* **21**, 230–237 (2007).
36. E. Padovan, G. Casorati, P. Dellabona, S. Meyer, M. Brockhaus, A. Lanzavecchia, Expression of two T cell receptor α chains: Dual receptor T cells. *Science* **262**, 422–424 (1993).
37. J. H. Park, C. Sauter, R. Brentjens, Cellular therapies in acute lymphoblastic leukemia. *Hematol. Oncol. Clin. North Am.* **25**, 1281–1301 (2011).
38. S. A. Rosenberg, N. P. Restifo, Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62–68 (2015).
39. M. Yousfi Monod, V. Giudicelli, D. Chaume, M. P. Lefranc, IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics* **20** Suppl. 1, i379–i385 (2004).
40. T. Bancroft, C. Du, D. Nettleton, Estimation of false discovery rate using sequential permutation *p*-values. *Biometrics* **69**, 1–7 (2013).


**Acknowledgments:** We thank T. Snyder and W. DeWitt for providing critical comments on a previous version of this manuscript. **Author contributions:** H.S.R. and C.S.C. designed the pairSEQ concept; A.M.S., J.B., and A.D.B. designed and performed the experimental work; B.H., R.O.E., and D.W.W. developed statistical methods and analyzed the data; and J.B., B.H., A.M.S., R.O.E., I.K., M.V., M.J.R., C.S.C., and H.S.R. wrote the manuscript. All authors discussed the results and commented on the paper. **Competing interests:** H.S.R. and C.S.C. have consultancy, equity ownership, patents, and royalties with Adaptive Biotechnologies; J.B., B.H., D.W.W., A.D.B., A.M.S., R.O.E., I.K., M.V., and M.J.R. have employment and equity ownership with Adaptive Biotechnologies; and H.S.R. is an inventor on the filed patent no. WO/2013/188831; PCT/US2013/045994, titled "Uniquely tagged rearranged adaptive immune receptor genes in a complex gene set." **Data and materials availability:** Data can be downloaded from http://adaptivebiotech.com/link/howie-2015-pairseq.

Submitted 12 May 2015
Accepted 7 July 2015
Published 19 August 2015
10.1126/scitranslmed.aac5624

**Citation:** B. Howie, A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, H. S. Robins, High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* **7**, 301ra131 (2015).

Science Translational Medicine
AAAS

# High-throughput pairing of T cell receptor α and β sequences

Bryan Howie *et al.*

*Sci Transl Med* **7**, 301ra131 (2015);
DOI: 10.1126/scitranslmed.aac5624

Editor's Summary

**T cell receptor chains pair off**

High-throughput immunosequencing can take a snapshot of the repertoire of immune cells, providing a broad picture of the immune response at any given time and tracking how the immune response changes as a result of perturbations such as vaccines, infection, or cancer. However, this approach has been limited by the inability to determine which TCR α and TCR β chains combine to form specific T cell receptors in a given cell. Now, Howie *et al.* report and validate a high-throughput method to pair TCR α and β segments without the need for single-cell technologies. They confirm that their method can be used for T cells from both blood and solid tissues.



*Science Translational Medicine* (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 