Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms

Dmitry A. Bolotin*, Ilgar Z. Mamedov*, Olga V. Britanova, Ivan V. Zvyagin, Dmitriy Shagin, Svetlana V. Ustyugova, Maria A. Turchaninova, Sergey Lukyanov, Yury B. Lebedev and Dmitriy M. Chudakov

1 Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia
2 Evrogen JSC, Moscow, Russia

The TCR repertoire is a mirror of the human immune system that reflects processes caused by infections, cancer, autoimmunity, and aging. Next generation sequencing (NGS) is becoming a powerful tool for deep TCR profiling; yet, questions abound regarding the methodological approaches for sample preparation and correct data interpretation. Accumulated PCR and sequencing errors along with library preparation bottlenecks and uneven PCR efficiencies lead to information loss, biased quantification, and generation of huge artificial TCR diversity. Here, we compare Illumina, 454, and Ion Torrent platforms for individual TCR profiling, evaluate the rate and character of errors, and propose advanced platform-specific algorithms to correct massive sequencing data. These developments are applicable to a wide variety of next generation sequencing applications. We demonstrate that advanced correction allows the removal of the majority of artificial TCR diversity with concomitant rescue of most of the sequencing information. Thus, this correction enhances the accuracy of clonotype identification and quantification as well as overall TCR diversity measurements.

Keywords: Bioinformatics · T cells · TCR

Introduction

The adaptive immune system drives the immune response via specific hypervariable molecules: B cell-generated immunoglobulins and immunoglobulin-like TCRs on T lymphocytes. These molecules are formed by genomic recombination in order to recognize foreign structures. The antigen specificity of each TCR is largely determined by the hypervariable CDR3 of the receptor beta chain, which is formed by recombination of V, D, and J gene segments with template-independent addition of nucleotides at the gene segment junctions.

Realistic analysis of the TCR CDR3 diversity is crucial for the understanding of the basic molecular mechanisms of adaptive immunity function in health and disease, including

*These authors contributed equally to this work.
consequences of aging, immunosuppressive and blood transplantation therapies, irradiation, autoimmunity, infections, and other conditions [1]. Moreover, high convergence of TCR beta CDR3 amino acid sequences [2], along with the expanding knowledge on their specificities [3], suggests the potential for diagnostics of multiple infections and pathological states via deep individual TCR profiling. These perspectives demand the development of powerful technological and software approaches for deep error-free TCR profiling.

Next generation sequencing (NGS) techniques have revolutionized this field by allowing massive analysis of TCR and antibody repertoires from individual blood samples [4–6] both before and after therapy [7, 8] and for different lymphocyte subsets [2, 9, 10]; however, performing deep, unbiased, and quantitative analysis of millions of CDR3 sequences that include highly homologous variants is quite challenging, due to numerous bottlenecks, accumulation of PCR and sequencing errors, and ratio bias. Altogether, these technical challenges lead to the loss of the original TCR repertoire of an analyzed T-cell sample, generation of false TCR diversity, and inability to interpret sequence information in a quantitative way, thus challenging intelligent analysis and cross-comparison of acquired datasets and generally complicating adaptive immunity studies.

The specific problem that distinguishes TCR repertoire analysis from sequencing analysis of a known genome or transcriptome resides in the fact that generating more sequencing reads results in the expansion of erroneous sequence variants with one, two, or more mismatches, and these changes raise the question of whether this sequence represents the same CDR3 with introduced errors or a different independently formed TCR. Here, we compare different NGS platforms for TCR profiling and propose advanced algorithms for efficient and valid correction of platform-specific errors with minimal loss of quantitative information originally contained in a lymphocyte sample.

Results

TCR library preparation and NGS

To provide reproducibility controls, we isolated RNA from three equal aliquots of peripheral blood mononuclear cells (PBMCs; hereafter called A, B, and C) that were obtained from the three parallel blood samples from the same donor. The first stage of library preparation was performed as described [8]. Briefly, cDNA synthesis was performed using a specific primer targeted to the constant region of TCR beta. A template switching effect was employed to generate a universal primer site at the 5′-end (Supporting Information Fig. 1). Next, a universal nested primer corresponding to the constant TCR beta region and a universal 5′-primer were used in the first amplification step. For the final pre-sequencing sample preparation, we employed different strategies for the 454, Ion Torrent, and Illumina platforms (see Materials and methods for details).

Importantly, the total number of PCR amplification cycles for pre-sequencing library preparation for each of the platforms did not exceed 22 cycles (taking into account the dilution of the first PCR reaction). Considering the average efficiency of PCR amplification as 1.8 (measured by real-time PCR, data not shown), the starting number of molecules that entered PCR amplification after cDNA synthesis was estimated to be approximately $1.2 \times 10^8$ for each sample. This bottleneck determined the possible depth of TCR profiling in our experimental system (see Discussion).

Data obtained with the three NGS platforms exhibited significant differences stemming from the sequencing technologies. Data from the 454 platform were less abundant, but the sequencing reads were the longest of the three platforms. The Illumina produced 100-fold more sequencing reads of shorter length, and the performance of the Ion Torrent was between the other two with respect to both parameters (Table 2 and Supporting Information Fig. 2).

Because reverse transcription, amplification, and sequencing each generate multiple errors [11], further analysis of NGS data required deep platform-specific processing to provide correct extraction and grouping of sequences with identical CDR3 regions that corresponded to distinct TCR clonotypes. Below, we provide optimized algorithms for such processing, which resulted in efficient correction of PCR and sequencing errors with minimal information loss.

Algorithms to eliminate PCR and sequencing errors for the Illumina platform

As was shown earlier, a significant number of PCR errors is accumulated during the course of pre-sequencing PCR and bridge amplification on the Illumina, and these errors can add to or even multiply the actual TCR repertoire diversity [6, 11]. To eliminate this artificial diversity, withdrawal of the low-abundance CDR3 variants that differ from the high-abundance variants by a single nucleotide mismatch [11] or removal of low-abundance sequence variants that comprise a total of 4% of all sequencing reads was suggested [6]. While these approaches appear reasonable, blind elimination of low-abundance TCR clonotypes can be disastrous for studies of converging clonotype variants. At the same time, complete elimination of artificial diversity is by far not achieved.

Another important issue is low-quality sequencing reads. Retaining low-quality sequences increases the number of erroneous clonotypes and thus, generates artificially increased diversity of the TCR repertoire. As sequence data are provided with a quality measure for each nucleotide, the low-quality sequences could potentially just be eliminated [6, 11]; however, we believe that this approach is erroneous for quantitative analysis for two reasons. First, strong filtering conditions result in loss of a significant portion of the data (up to 50% sequencing reads for the Illumina platform and even more for other platforms, see below). Second, NGS errors may occur in a sequence-dependent manner [12], leading to underestimation of certain TCR clonotypes.
Table 1. Typical examples of sequencing reads with PCR errors merged with the dominant “core clonotype”

| Number of sequencing reads | CDR3, nucleotide sequence | CDR3 length, nt. | CDR3, amino acid sequence | TRBV segment | TRBJ segment | TRBD segment |
|---------------------------|---------------------------|-----------------|---------------------------|--------------|--------------|--------------|
| 18732                     | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 74                        | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAVKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 2                         | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CVVKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 1                         | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 74                        | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 1                         | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 74                        | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 1                         | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |
| 74                        | tgtgccatcaagacagctacgaggttggatgtgtggtgagctggtttgct | 45              | CAIKTTSGIVDEQFF           | V10–3        | J2–1         | D2           |

*The dominant “core clonotype,” which determines the nucleotide sequence of the final clonotype, is shown in bold. V segments (orange), D segments (blue), J segments (green), and corrected errors (bold red) within the CDR3 nucleotide sequence are shown. Secondary PCR errors are highlighted in gray. Because of space limitations, a limited number of low-abundance subvariants is shown.*

Based on these considerations, we propose the following advanced error correction algorithm for processing Illumina TCR profiling data. This algorithm provides efficient and safe elimination of PCR and sequencing errors with minimal information loss.

i. CDR3 extraction. From each sequencing read, the CDR3 is extracted, if possible. This step is based on the alignment of each sequence to the set of genomic V, D, and J segments from the IMGT/GENE-DB database and identification of conservative Cys and Phe at the CDR3 boundaries. For the V, D, and J gene segment identification, low-quality nucleotides are treated as allowable mismatches (detailed in Materials and methods).

ii. Formation of the “core clonotypes.” Reads with identical CDR3 nucleotide sequences with high-quality sequence at each nucleotide position within CDR3 are grouped to form “core clonotypes”.

iii. Mapping of low-quality reads. Low-quality sequencing reads with up to three low-quality nucleotides within the CDR3 are merged with the latter “core clonotypes” (mismatches allowed in low-quality positions). Sequencing reads that contain more than three low-quality nucleotides within the CDR3 or fail to map to a “core clonotype” are removed.

iv. Correction of PCR errors. Mismatches within V, D, or J segments of CDR3 do not arise naturally, because TCRs do not undergo somatic hypermutation. Therefore, nucleotide mismatches within the segments can only arise from PCR and sequencing errors and may be safely corrected. At this step, low-abundant “core clonotypes” are merged with the more abundant (at least fold more abundant) “core clonotype” that differs by no more than three nucleotides within the V, D, or J segments of CDR3 (since percentage of sequencing reads carrying more than three errors is negligible, see below). Of these mismatches, no more than two are allowed within the V gene segment (excluding the last two identified nucleotides), no more than two are allowed within the J gene segment (excluding the first two and the last two identified nucleotides), and no more than one is allowed within the D gene segment (excluding the first two identified nucleotides). This additional constraint was introduced to avoid clusterization of clones with homologous but actually different segments. Such correction is safe with respect to the potential risk of losing natural TCR diversity, since no mismatches within the nucleotides added or trimmed during TCR rearrangement are used for correction (Table 1). In contrast to the blind cutoff of 4% sequencing reads of low-abundance CDR3 variants proposed by Warren et al. [6], this approach allows rational elimination of PCR errors while maintaining the natural TCR diversity.

Each of the final clonotypes is characterized by its unique CDR3 nucleotide sequence, identified V, D, and J gene segments (all determined by the dominant “core clonotype”), and the total merged reads number that corresponds to the relative abundance of a T-cell clone within the initial sample. According to the
proposed algorithm, these reads include identical high-quality CDR3 sequencing reads, homologous low-quality reads, and homologous high-quality reads of lower abundance (transcription, reverse transcription, PCR, and sequencing errors). Importantly, none of the low-quality reads forms a final clonotype. Instead, low-quality reads are rescued by mapping to the final clonotypes formed by high-quality sequencing reads.

Advanced correction resulted in a dramatic reduction of the estimated TCR diversity (twofold reduction of extracted clonotypes number) in the Illumina data obtained for the complex mix of blood-derived T-cell sample (Table 2, advanced correction). At the same time, 90% of the quantitative sequencing information (i.e., total number of evaluable sequencing reads) was preserved, resulting in a nearly twofold increase in reads per clonotype, a parameter that generally determines quantification power of the whole method (see below). In contrast, removal of the low-quality reads resulted in a substantial loss of quantitative information, although the artificial diversity was not eliminated efficiently (Table 2).

The percentages of the sequencing reads that mapped to the 1000 major CDR3 clonotypes with single, double, and triple nucleotide mismatches within the CDR3 were 3.59%, 0.12%, and 0.01%, respectively, confirming that percentage of sequencing reads with >3 errors is negligible. Notably, 100% of the mapped sequences with double and triple nucleotide mismatches represented minor descendants of the mapped sequences with single nucleotide mismatches (such as shown in gray in Table 1), indicating these mismatches originated from secondary errors during later PCR cycles.

Advanced correction eliminated the major portion of PCR errors. To illustrate this result, the number of CDR3 sequences that were found only once in a dataset that mostly comprised reads with sequencing or PCR errors was greatly reduced after advanced correction (compare Fig. 1A and C), whereas straightforward removal of the low-quality reads was much less efficient in this respect (Fig. 1B). The number of erroneous but high-quality CDR3 sequence variants that were successfully mapped to a single large clonotype often exceeded 100 (Supporting Information Dataset 1), providing a vivid example of how accumulated PCR errors can multiply real TCR diversity.

By definition, proposed algorithm cannot eliminate errors within added nucleotides or within the edges of the V, D, and J gene segments. Assuming uniform distribution of error positions within CDR3, this limitation leads to undercorrection of approximately 15–35% of PCR errors, depending on the nature of the analyzed TCR library. By excluding these limitations, remaining errors can be efficiently corrected as well (data not shown) but the significant risk appears to lose natural TCR diversity, which remains minimal with the cautious correction algorithm proposed above.

Analysis algorithms for 454 and Ion Torrent TCR data

One problem that is specific to the 454 and Ion Torrent platforms is imperfect interpretation of homopolymeric stretches, as this difficulty leads to erroneous nucleotide deletions and insertions. To
Figure 1. Abundance distribution of clonotypes sequenced 1–30 times (Illumina data) (A) before any correction, (B) after removal of the low-quality sequencing reads, and (C) after correction of PCR errors and mapping of low-quality sequencing reads. Data shown are representative of three independent experiments.

take the uncertainty in nucleotide lengths into account for further analysis, we initially rebuild sequences from the raw flowgrams, where we interpret ambiguous signals by rounding up with the last position in a stretch designated as “uncertain”. For example, raw flowgram values of 3.8–4.2 are interpreted as a 4-nt stretch, whereas values of 4.21–4.79 are interpreted as a 5-nt stretch with the fifth position designated “uncertain”. Values of 4.8–5.2 are interpreted as certain 5-nt stretch. The “uncertain” positions are treated in further analysis as possible insertions, thus, allowing us to correct for this type of sequencing error as described below.

The whole analysis algorithm is generally similar to the algorithm we proposed for the Illumina:

i. Rebuilding sequences from raw data. Ambiguous flowgram values are rounded up, with the last position in the resulting nucleotide stretch designated as “uncertain”.

ii. CDR3 extraction. Similar to the algorithm for the Illumina, for identification of the V, D, and J gene segments, “uncertain” nucleotides are treated as possible insertions.

iii. Formation of “core clonotypes.” As for the Illumina, high-quality reads without ambiguous flowgram values within the CDR3 are merged to form “core clonotypes”.

iv. Mapping of low-quality reads. Similar to the algorithm for the Illumina, up to three “uncertain” nucleotides within the CDR3 are treated as possible insertions and corresponding sequencing reads are merged with the “core clonotypes”.

v. Correction of PCR errors. Correction is identical to the algorithm for the Illumina platform.

Additional modification of the algorithm is possible by employing asymmetric interpretation of raw flowgram data that have an initial systemic shift in signal readout. For example, flowgram values from 3.9 to 4.3 nt can be interpreted with certainty as a 4-nt stretch, leading to a −0.1-nt shift in the average lengths of CDR3 sequences.

As shown in Table 2, advanced error correction works efficiently for the 454 and Ion Torrent data and results in a dramatic reduction of interpreted TCR diversity, minimal loss of quantitative information, and up to a fourfold increase in the read-per-clonotype parameter.

To evaluate the error rate of deletions and insertions, we performed in silico spectratyping analysis of sequence datasets, in which functional “in-frame” clonotypes have CDR3 nucleotide sequence lengths in multiples of 3. Stretch errors result in the generation of artificial shorter or longer out-of-frame sequences (Fig. 2). Without correction, as much as 29.3% of the 454 sequences and 25.0% of the Ion Torrent sequences fall to the out-of-frame columns compared to only 1% of the Illumina sequences. In the latter platform data, these sequences may result from the naturally presented nonfunctional out-of-frame TCR beta mRNA molecules that have escaped nonsense-mediated decay. The percentage of out-of-frame CDR3 sequencing reads was essentially decreased after advanced correction. In fact, for the 454 data, these sequences were decreased to 22.2%, and for the Ion Torrent data, these sequences were decreased to 7.2%. Overall, advanced correction allowed the conversion of Ion Torrent data to Illumina-like results that are intrinsically free of deletion/insertion errors (Table 2 and Fig 2D). With other more abundant 454 datasets of higher quality, correction for the out-of-frame reads was also quite efficient (data not shown).

Still, relatively low number of high quality CDR3-containing sequencing reads obtained from Ion Torrent and 454 compared to Illumina create additional bottleneck at the sequencing level that leads to essential loss of low abundant clonotypes and decreases efficiency of error correction algorithms.

Error rates

The overall rate of corrected PCR errors for each of the three NGS platform datasets was analyzed by counting the number of mismatch-containing high-quality sequencing reads mapped to the high-quality “core clonotypes”. The percentage of such mismatch-containing sequencing reads was notably higher for the Illumina dataset than for the 454 and Ion Torrent datasets (3.2%, 1.4%, and 1.2% of sequencing reads, respectively), corresponding to recently
reported data that demonstrated a high rate of PCR errors during Illumina bridge amplification, leading to up to 3.19% of high-quality Illumina reads with PCR errors [11].

As expected, the rate of mismatched nucleotide incorporation was not evenly distributed, and transitions were favored over transversions (Supporting Information Table 2). Interestingly, we obtained a significantly higher rate for A-to-G substitutions as compared to T-to-C substitutions for all platforms (38–41% versus 13–20%). As PCR substitutions should be symmetrical (i.e., G-to-A substitutions on the antisense strand will result in a T-to-C substitution on the sense DNA strand, yielding equivalent rates for G-to-A and C-to-T substitutions), we suggest that these errors were generated during the course of one of the asymmetrical processes, which include natural transcription errors, cDNA synthesis, or sequencing. As this type of bias was observed in all datasets generated by each of the different sequencing technologies and the natural transcription error rate is generally low [13], these errors were most likely generated by reverse transcription. Importantly, due to their early origin and high abundance within the final NGS data, these frequent errors cannot be corrected by previously proposed algorithms [6, 11]; however, our described algorithms efficiently capture such errors.

Quantification power of NGS-based TCR profiling

Quantitative NGS-based analysis of relative T-cell clone abundances within an initial lymphocyte sample is highly desirable for the individual and for cohort comparative studies of adaptive immunity [8]. In order to compare the relative quantification accuracy for the three platforms, we plotted the mean percentages of each J beta gene, V beta gene, and the 50 most abundant individual CDR3 clonotypes for the Illumina dataset and compared these to the 454 and Ion Torrent datasets (Fig. 3). We did not observe significant bias between the 454 and Illumina datasets with respect to the abundance of J beta and V beta gene segments as well as with respect to the major CDR3 clonotypes (Fig. 3A–C). We did, however, observe a significantly weaker correlation between the Illumina and Ion Torrent data (Fig. 3D–F) with a more than 100-fold difference for particular CDR3 clonotypes. Comparison of NGS data with flow cytometry data obtained from the blood cells of the same patient using TCR V beta family-specific antibodies demonstrated that, while the 454 and Illumina data correlated well with the real abundance of corresponding T-cell subsets, the Ion Torrent results were skewed with respect to the relative TCR V beta gene abundances (Fig. 4).

The main bias in TCR quantification may arise from multiplex PCR, which is used on the second amplification step during Illumina and Ion Torrent sample preparation, since multiplex PCR is known to skew the abundance of template molecules [14, 15]. Therefore, we attribute the dramatic skew in Ion Torrent data to multiplex amplification with a complex mix of TCR V beta-specific primers that had to be employed for the Ion Torrent library preparation. These data indicate that multiplex PCR for pre-sequencing library preparation should be avoided or very thoroughly corrected to provide quantitative measurement of TCR clonotype...
concentrations by NGS profiling. Apparently, multiplex amplification with a simple mix of 13 highly homologous J genes-specific primers at the second step of Illumina sample preparation did not significantly skew the natural abundance of TCR molecules. However, this multiplexing can be avoided as well by using primers specific to the constant region of TCR genes and 150 bp Illumina read length.

The overall accuracy of quantification of individual clonotypes using Illumina sequencing was estimated via comparison of the datasets obtained for the three independent blood samples (A, B, and C). These data were corrected using our advanced algorithm (see above) or by removal of low-quality reads. As expected from the increased reads-per-run parameter and the number of rescued low-quality reads, advanced correction significantly ($p = 3.45E-7$ by the Kruskal–Wallis test) increased the accuracy of quantification for T-cell clonotypes with concentrations greater than 0.01% (Fig. 5). As a result, a quantification accuracy of 50% (i.e., 95% CI for concentration value is ($C/1.5$:$C \times 1.5$), where $C$ is the measured relative concentration of a clonotype) was achieved for the clonotypes that constitute at least 0.1% of all T cells.

**Discussion**

Application of NGS for deep TCR repertoire characterization raised issues concerning the accuracy of sequencing and the quantification of individual TCR clonotypes. Several methods of probe preparation and data processing for NGS TCR repertoire analysis have been proposed [4–6, 8], but the issue of artificial TCR diversity generation has not been fully resolved. Here, we propose correction algorithms that allow the elimination of most accumulated errors without the risk of losing natural independent clonotypes. Insertion/deletion errors characteristic of the 454 and Ion Torrent technologies can also be eliminated quite efficiently after
reinterpretation of raw flowgram data. These algorithms, thus, allow use of these NGS platforms in applications where longer reads (precise TCR V beta/alpha gene identification or antibody sequencing) or faster or cheaper (per run but not per read) results are required.

Accurate quantification of certain T-cell clones in the analyzed sample is an important goal in TCR repertoire studies that aim to track changes made by therapeutic procedures or immune responses. We demonstrated that rational mapping of low-quality reads instead of blind filtering allows to minimize loss of quantitative information and to significantly increase the resulting accuracy of clonotype quantification.

The depth of TCR profiling is limited by bottlenecks that may appear at any stage, including the initial T-cell count, number of molecules that successfully entered PCR amplification, and number of output sequencing reads. Thus, uniform amplification starting from a sufficient number of molecules is crucial for the quantitative and unbiased analysis of complex TCR gene libraries. Albeit with several pros and cons, genomic DNA or mRNA can be used as a starting material for TCR profiling. We believe the latter is preferable in most cases due to the following reasons:

i. Each T-cell contains multiple copies of RNA molecules that encode the TCR beta and alpha chains, and these copies expand the potential bottleneck between sampled T cells and the final TCR amplicon.

ii. When starting from genomic DNA, the entire sample isolated from certain PBMC aliquots must be amplified to gain a comprehensive representation of the TCR repertoire. This inclusion may be technically challenging when large T-cell populations are studied. For example, a starting sample of 10 million T cells requires amplification of approximately 100 μg of PBMC-derived genomic DNA. In contrast, a reasonable aliquot (5–10 μg) of an RNA sample obtained from the same cell population is sufficient to sample the diversity.

iii. At the DNA level, each T-cell carries two rearranged TCR beta genes, and one of them is nonfunctional. In contrast, out-of-frame mRNA molecules are efficiently degraded by the nonsense-mediated decay mechanism [16, 17], and thus, these nonfunctional molecules are not sampled.

iv. TCR beta cDNA can be synthesized and amplified using universal primers specific to the constant region of TCR beta genes [5, 8]. In addition, the template switching effect can be used in reverse transcription [18], providing universal priming at the 5′ end of mRNA. This technique allows us to avoid use of the complex multiplex primer sets required to amplify all TCR V beta and J beta gene variants and thus, decreases PCR bias.

Depending on the questions posed, the required depth of TCR repertoire analysis can vary across a wide range. Cell count intervals that are calculated according to the Poisson model suggest that a comprehensive analysis of a population containing a diversity of $10^7$ (estimated individual TCR beta diversity [4]) requires analysis of a sample of at least $10^8$ cells. The corresponding number of T cells is contained in approximately 50–70 mL of blood from a healthy individual. At any further stages of library preparation, starting from DNA or RNA purification and up to the NGS output, the number of molecules/sequencing reads should be well above 100 million. These conditions are potentially feasible (albeit laborious and expensive) prerequisites for the nearly comprehensive analysis of an individual's TCR beta diversity, which includes not only effector and memory T cells but also a considerable portion of the naïve T-cell repertoire. Much smaller bottleneck limits can be estimated for the analysis of a narrower T-cell population, such as a fraction of effector or memory T cells or a presorted subpopulation of specific TCR V beta family T cells, Treg cells, MHC-tetramer sorted cells, etc. [2, 19]. Still, a clear understanding of parameters that determine the sample preparation bottlenecks and their minimal required size is important for adequate experimental design.

In conclusion, rational analysis of the TCR repertoire by NGS demands intelligent design of the whole experimental pipeline, including blood sampling, DNA/RNA purification, cDNA synthesis, PCR amplification, pre-sequencing preparation, sequencing depth, platform choice, and finally intelligent interpretation of the NGS output with regard to all potential errors accumulated above. We believe that the algorithms proposed here for the efficient correction of PCR and sequencing errors and for rescue of the quantitative information contained in low-quality reads should contribute
to rational analysis of the TCR repertoire and fuel adaptive immunity studies.

Materials and methods

Accession number for all raw sequencing data of this study: NCBI Sequence Read Archive (SRA): SRP012846.

Sample preparation and sequencing

Three aliquots of peripheral blood from 43-year-old man were obtained. Informed consent was obtained from the patient for these procedures. PBMCs (3–4 × 10⁶) for each of three replicates (A, B, and C) were isolated by Ficoll-Paque (Paneco, Russia) density gradient centrifugation. RNA was isolated using the Trizol reagent (Invitrogen, USA) according to the manufacturer’s protocol. First-strand cDNA was synthesized for 2 h using the Mint cDNA synthesis kit (Evrogen, Russia) and the primer BC1R (5′-CAGTATCTGGGAGTCATTGA-3′), which is specific to both variants of the TCR beta constant regions. PlugoLigo (AAGGCAATGTATCAGGAGAAGGAGG-G-P, Evrogen) was added after 30 min of synthesis. The PCR amplification protocol was as follows for 18 cycles: 94 °C for 20 s, 65 °C for 20 s, and 72 °C for 50 s. The PCR mixture (15 μL) contained 1× Encyclo polymerase buffer (Evrogen), 0.125 mM of each dNTP, 10 pmol of universal primers: M1 (AAGGCAATGTATCAGGAGAAGGAGG-G-P, Evrogen) and BC2R (5′-CTTTCTCTGGTACCTAACAAC-3′), 0.3 μL of Encyclo polymerase mix, and 1 μL of undiluted first-strand cDNA. After the first-strand cDNA synthesis, the first PCR was performed in 15 tubes in order to amplify all TCR molecules for each of the independent sample replicates (A, B, and C). Then, the PCR products from all 15 tubes were mixed and diluted 1:10. Then, 1 μL of the diluted PCR product was used in each of the ten second-step PCR reactions (25 μL each), which were performed with conditions specific for each sample type.

For the 454 sample

We used nested 3′ primers with barcodes that distinguished the A, B, and C replicates in combination with a universal 5′ primer to generate a 500–550-bp amplicon readily flanked by 454 sequencing primers. Thus, 454 reads directionally started from the constant TCR beta region and covered the J gene segment, CDR3, and a part of the V gene segment that was sufficient for its unambiguous identification (Supporting Information Fig. 1). This technique permits amplification of the TCR beta genes with the same pair of primers for all TCR V beta gene variants, and this property decreases potential PCR bias. Each PCR reaction contained 1 μL of the first PCR (diluted), 1× PCR buffer, 0.125 mM of each dNTP, 10 pmol of the primer B-M1 (5′-GGCT TGCCAGCCCCTGATGGATGATCAAGCCAGGT-3′), and 10 pmols of one of the reverse primers, which contained specific barcodes for each of the three replicates (Supporting Information Table 1). The amplification protocol was as follows for 12 cycles: 94 °C for 20 s, 68 °C for 20 s, and 72 °C for 50 s. After the second PCR the amplicons were purified by agarose gel electrophoresis using a DNA gel extraction kit (Cytokin, Russia). The purified amplicons were re-amplified for two additional cycles and finally purified with the QiAquick PCR purifi-
cation kit (Qiagen, USA). Sequencing was performed using Genome Sequencer FLX, GS Em PCR Kit II (Roche Applied Science).

For the Illumina sample

Due to the relatively short sequencing reads produced by the Illumina platform, pre-sequencing preparation requires nested PCR amplification of the TCR beta library to achieve sequencing of the CDR3 region of interest. To this end, we used a set of multiplex J beta-specific primers containing an Mmel restriction site. Mmel treatment resulted in PCR fragments that ended with the triplet coding for the last conservative phenylalanine of CDR3 and were, thus, suitable for the 72-bp Illumina sequencing (Supporting Information Fig. 1). At the 5′ end, the universal primer that annealed to the template formed by the switching effect was used, resulting in a PCR library of fragments with approximately 450–500 bp. Each PCR reaction contained 1 μL of the first PCR (diluted), 1× PCR buffer, 0.125 mM of each dNTP, 5 pmol of primer MmelStep1 (5′-CATCCTATCCGACACGTTAG-3′), 0.5 pmol MmeSmart20 (5′-CATCCTATCCGACACGTTAGTATCAAC-3′), 3 pmol of each of the 13 MmeJ primers (Supporting Information Table 1), and 0.5 μL of the Encyclo polymerase mix (Evrogen). The amplification protocol was as follows for 12 cycles: 94 °C for 20 s, 68 °C for 20 s, and 72 °C for 50 s. After the second PCR step, the amplicons were purified with the QiAquick PCR purification kit (Qiagen, USA) and digested by Mmel (New England Biolabs, USA) according to the manufacturer’s protocol. The efficiency of Mmel digestion was estimated to be about 80% according to a control PCR with primers that annealed to the fragment to be removed by theendonuclease (data not shown). After Mmel digestion, the DNA samples were purified by the QiAquick PCR purification kit. After end polishing, the digested amplicons were ligated to barcoded adaptors to distinguish the A, B, and C replicates, processed using the standard Illumina sequencing protocol, and sequenced on the Genome Analyzer IIX (Illumina).

For the Ion Torrent sample

The Ion Torrent generates more that 10 Mbp data with sequencing reads of about 100 bp (314 chip). One of the points that distinguishes this platform from the 454 and Illumina is the length restriction of the DNA fragments (150–200 bp) that can bind to the sequencing chip. Therefore, to prepare a library suitable for Ion Torrent sequencing, multiplex PCR was implemented on both
sides of the TCR molecules at the second PCR stage (Supporting Information Fig. 1). The diluted first PCR reactions for the replicates A, B, and C were mixed together prior to the second PCR. Each of the second PCR reactions contained 1 μL of the first PCR (diluted), 1× PCR buffer, 0.125 mM of each dNTP, 3 pmol of each of the 13 BJ primers, 40 pmol of mixed bvs primers (Supporting Information Table 1), and 0.5 μL of the Encyclo polymerase mix (Evrogen). The amplification protocol was as follows for 17 cycles: 94°C for 20 s, 62°C for 20 s, and 72°C for 50 s. The PCR product was precipitated with ethanol and purified by agarose gel electrophoresis using a DNA gel extraction kit (Promega, USA). The obtained amplicon was ligated to adapters and sequenced on the Ion Torrent PGM (Life Technologies) machine.

**CDR3 extraction algorithm**

We implemented the following algorithm for the initial CDR3 extraction. The J gene segment is localized based on the identity of the short 6-nt motif to one of the 13 J genes from the IMGT database. Then, the region of identity is expanded into the CDR3 until the last matched nucleotide is encountered. The list of possible J segments for each read is generated to include the longest identical one(s) and the one(s) that are 2-nt shorter (if any). The V gene segment is then localized based on the identity of the short 7-nt motif that precedes the conservative Cys to one of the functional V genes from the IMGT database. Then, the region of identity is expanded in both directions until the last matched nucleotide is encountered. The list of possible V segments for each read is generated to include the longest identical one(s) and the one(s) not more than 2-nt shorter (if any). The D gene segment is localized based on the identity of at least 6 nt between the identified V and J gene segments. The CDR3 is extracted for each read as the nucleotide sequence between and including triplets coding for the conservative Cys located in V gene segment and for the conservative Phe located in J gene segment. The extracted CDR3 sequences are then further processed as described in the main text.

**Acknowledgments:** This work was supported by the Molecular and Cell Biology program of the Russian Academy of Sciences; Russian Science Support Foundation, Russian Foundation for Basic Research (10-04-01771-a, 11-04-12042-off-m), and a grant from the President of the Russian Federation (MK-575.2011.4). We thank Alexey Kurnosov for the help in manuscript preparation.

**Conflict of interest:** The authors declare no financial or commercial conflict of interest.

References

1 Kedzierska, K., La Gruta, N. L., Stambas, J., Turner, S. J. and Doherty, P. C., Tracking phenotypically and functionally distinct T-cell subsets via T-cell repertoire diversity. Mol. Immunol. 2008. 45: 607–618.

2 Venturi, V., Quigley, M. F., Greenaway, H. Y., Ng, P. C., Ende, Z. S., McIntosh, T., Asher, T. E. et al., A mechanism for TCR sharing between T-cell subsets and individuals revealed by pyrosequencing. J. Immunol. 2011. 186: 4285–4294.

3 Venturi, V., Chin, H. Y., Asher, T. E., Ladell, K., Scheinberg, P., Bornstein, E., van Bockel, D. et al., TCR beta-chain sharing in human CD8+ T-cell responses to cytomegalovirus and EBV. J. Immunol. 2008. 181: 7853–7862.

4 Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Khalsai, O., Riddell, S. R. et al., Comprehensive assessment of T-cell receptor beta-chain diversity in alpha beta T cells. Blood. 2009. 114: 4099–4107.

5 Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. and Holt, R. A., Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009. 19: 1817–1824.

6 Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R. et al., Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 2011. 21: 790–797.

7 Britanova, O. V., Bochкова, A. G., Staroverov, D. B., Fedorenko, D. A., Bolotin, D. A., Mamedov, I. Z., Turchaninova, M. A. et al., First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. Bone Marrow Transplant. 2012. DOI: 10.1038/bmt.2012.44.

8 Mamedov, I. Z., Britanova, O. V., Bolotin, D. A., Chikalina, A. V., Staroverov, D. B., Zvyagin, I. V., Kotlobay, A. A. et al., Quantitative tracking of T-cell clones after haematopoietic stem cell transplantation. EMBO Mol. Med. 2013. 1: 201–207.

9 Robins, H. S., Srivastava, S. K., Campregher, P. V., Turtle, C. J., Andriesen, J., Riddell, S. R., Carlson, C. S. et al., Overlap and effective size of the human CD8+ T-cell receptor repertoire. Sci. Transl. Med. 2012. 2: 47ra64.

10 Klarenbeek, P. L., Tak, P. P., van Schaik, B. D., Zwinderman, A. H., Jakobs, M. E., Zhang, Z., van Kampen, A. H. et al., Human T-cell memory consists mainly of unexpanded clones. ImmunoL Lett. 2013. 133: 42–48.

11 Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C. and Geiger, T. L., Identification of errors introduced during high throughput sequencing of the T-cell receptor repertoire. BMC Genom. 2011. 12: 106.

12 Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J. et al., Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009. 10: R32.

13 Sydow, J. F. and Cramer, P., RNA polymerase fidelity and transcriptional proofreading. Curr. Opin. Struct. Biol. 2009. 19: 732–739.

14 Elnifro, E. M., Ashshi, A. M., Cooper, R. J. and Klapper, P. E., Multiplex PCR: optimization and application in diagnostic virology. Clin. Microbiol. Rev. 2000. 13: 559–570.

15 Markoulatos, P., Siafakas, N. and Moncany, M., Multiplex polymerase chain reaction: a practical approach. J. Clin. Lab. Anal. 2002. 16: 47–51.

16 Bhalla, A. D., Gudikote, J. P., Wang, J., Chan, W. K., Chang, Y. F., Olivas, O. R. and Wilkinson, M. F., Nonsense codons trigger an RNA partitioning shift. J. Biol. Chem. 2009. 284: 4062–4072.

17 Wang, J., Vock, V. M., Li, S., Olivas, O. R. and Wilkinson, M. F., A quality control pathway that down-regulates aberrant T-cell receptor (TCR)
transcripts by a mechanism requiring UPF2 and translation. J. Biol. Chem. 2002. 277: 18489–18493.

18 Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L. and Chenchik, A., Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res. 1999. 27: 1558–1560.

19 Wang, C., Sanders, C. M., Yang, Q., Schroeder, H. W., Jr., Wang, E., Babrzadeh, F., Gharizadeh, B. et al., High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T-cell subsets. Proc. Natl. Acad. Sci. USA 2010. 107: 1518–1523.

Abbreviation: NGS: next generation sequencing