Bioinformatics, 31(18), 2015, 2963–2971
doi: 10.1093/bioinformatics/btv309
Advance Access Publication Date: 18 May 2015
Original Paper

Sequence analysis

IMSEQ—a fast and error aware approach to immunogenetic sequence analysis

Leon Kuchenbecker[1,2,3,*], Mikalai Nienen[1], Jochen Hecht[1,3], Avidan U. Neumann[1,4], Nina Babel[1,5], Knut Reinert[2,3] and Peter N. Robinson[1,2,3,6,*]

[1]Berlin-Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin, Berlin, [2]Department of Computer Science, Freie Universität, Berlin, [3]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany, [4]Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, [5]Marien Hospital Herne, Ruhr University Bochum, Bochum and [6]Institute of Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on February 20, 2015; revised on May 8, 2015; accepted on May 11, 2015

Abstract

Motivation: Recombined T- and B-cell receptor repertoires are increasingly being studied using next generation sequencing (NGS) in order to interrogate the repertoire composition as well as changes in the distribution of receptor clones under different physiological and disease states. This type of analysis requires efficient and unambiguous clonotype assignment to a large number of NGS read sequences, including the identification of the incorporated V and J gene segments and the CDR3 sequence. Current tools have deficits with respect to performance, accuracy and documentation of their underlying algorithms and usage.

Results: We present IMSEQ, a method to derive clonotype repertoires from NGS data with sophisticated routines for handling errors stemming from PCR and sequencing artefacts. The application can handle different kinds of input data originating from single- or paired-end sequencing in different configurations and is generic regarding the species and gene of interest. We have carefully evaluated our method with simulated and real world data and show that IMSEQ is superior to other tools with respect to its clonotyping as well as standalone error correction and runtime performance.

Availability and implementation: IMSEQ was implemented in C++ using the SeqAn library for efficient sequence analysis. It is freely available under the GPLv2 open source license and can be downloaded at www.imtools.org.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: lkuchenb@inf.fu-berlin.de or peter.robinson@charite.de

1 Introduction

T- and B-cells are lymphocytes that form part of the adaptive or acquired immune system in vertebrates. These cells express an antigen receptor on their surface, the *T-cell receptor* (TCR) or *B-cell receptor/Immunoglobulin* (IG), which is responsible for the recognition of antigens. Due to the enormous number of potential antigens, a large diversity among these receptors is required. The key behind that diversity is that in the germline genome there is no functional TCR or IG gene. Instead, it is generated by recombination of germline encoded gene segments as well as random base additions and deletions upon differentiation of T- and B-cell precursors in the thymus and bone marrow. This somatic recombination process is

referred to as *V(D)J-recombination* and explained in more detail in the following section. The cells transmit their recombined receptor gene to daughter cells upon proliferation, allowing for a targeted immune response against antigens.

Recently, methods have been developed to enrich these recombined genes in order to capture the repertoires of unique gene sequences using next generation sequencing (NGS). The derived data, i.e. populations of such unique gene sequences, referred to as *clonotypes*, can be used to better understand how the adaptive immune system works. The underlying samples for this analysis can be full T or B cell populations from blood, urine, biopsies or other sources or antigen specific subpopulations that have been exposed to antigen *in vitro* and subsequently selected for activation markers using methods such as flow cytometry. Also, specialized cells (e.g. regulatory, memory or effector T cells) can be sorted with immunophenotyping protocols.

NGS based T- and B-cell clonotype analysis thus opens up new possibilities of qualitative and quantitative repertoire assessment. Recently, such methods were applied for TCR and IG repertoire analysis in healthy and diseased conditions. A number of data characteristics can be measured and are medically relevant, including repertoire size and diversity under different conditions or for different subpopulations as well as shared clonotypes across different individuals and the tracking of particular clonotype subgroups. Applications in medical and clinical research include the identification and tracking of clonotypes with known antigen specificities in vaccination studies (Jackson *et al.*, 2014; Vollmers *et al.*, 2013), differential diagnosis in transplant patients (Dziubianau *et al.*, 2013) or the usage of clonotypes as biomarkers in patients with lymphoid malignancies (Wu *et al.*, 2012).

Here, we present IMSEQ, a method to derive clonotypes from NGS sequence data. It can recognize and correct likely sequencing- or PCR-errors, can quickly process large datasets and was extensively tested on simulated and experimental data. We also show that it outperforms currently available software tools. IMSEQ is open source and available online (http://www.imtools.org).

1.1 Preliminaries

The generation of a functional TCR or IG gene follows the same principle and will be briefly explained using the TCR as an example. In humans, the TCR gene locus consists of an α and a β chain. The TCRβ locus on chromosome 7 has a cluster of over 50 variable (V) segments located at some distance to other clusters, two diversity (D) segments and 13 joining (J) segments. During differentiation of T lymphocytes in the thymus, somatic recombination in the thymus leads to the production of a functional TCRβ gene with one V, D and J segment each as well as a constant, conserved (C) segment. Additional diversity is added through random nucleotide removal and addition at the junction sites between the segments. A homologous process governs recombination of both the TCRs and the IGs, which are protein complexes that comprise two subunits, the beta (TCRB) or heavy chain (IGH), as well as the alpha (TCRA) or light chain (IGL). All four genes undergo an analogous somatic recombination process, however, the TCRA and IGL genes comprise only a V and J segment without an intervening D segment.

The structure of a recombined TCRB locus is shown in Figure 1. Given the number of functional gene segments in the human genome as well as the average number of nucleotides randomly added or removed at the junction sites, it has been estimated that the size of the theoretical space of clonotypes that can be generated is on the order of 10^{18} (Janeway *et al.*, 1999). *In vivo*, about 10^6 unique TCR recombinations (clonotypes) among 10^{12} cells were detected in the peripheral blood of humans (Arstila *et al.*, 1999).

Fig. 1. The basic structure of a recombined TCRB or IGH gene. V, D and J denote conserved germline sequence, R_1 and R_2 denote random nucleotides at the junction sites. The highlighted area between the conserved Cys- and Phe-triplets denotes the *CDR3 region*

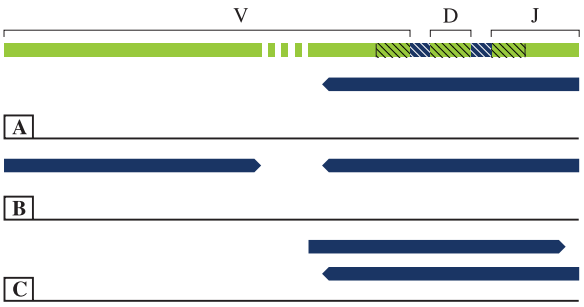

Fig. 2. Read positioning supported by IMSEQ. The V(D)J region must be covered by a single read. This also reflects the minimal configuration, as shown in **(A)**. A second read covering additional parts of the V region as shown in **(B)** is optional and improves the accuracy of the V segment identification. If a paired-end configuration where the reads overlap **(C)** is chosen, the reads must be pre-processed and merged before IMSEQ analysis

The recombined genes are commonly enriched by a multiplex PCR targeting the V and J segments. Different target sites result in different sequencing read positions. The configurations supported by IMSEQ are shown in Figure 2.

1.2 Technical artifacts in TCR and IG sequencing

PCR amplification and sequencing are error-prone processes, mainly due to *false base incorporations*. These errors cause the detection of additional unique clonotypes which are in fact not present in the sample, particularly when the errors lie within the hypervariable CDR3 region, since it cannot be compared with any reference and redundancy through coverage (Sims *et al.*, 2014) cannot be applied to handle these errors. Sequencing reads usually come with a quality score per base and it is common practice to define a quality threshold for reads to be discarded. False base incorporations do however not occur independent of the sequence, hence they can alter the derived clonotype frequency distributions if reads are selected for their quality scores. An extreme case of such a distribution change is shown in Figure 3. This analysis demonstrates that it is preferable to process the raw data in an attempt to correctly assign the clonotypes of low-quality reads rather than removing them.

Another technical artifact that can occur in TCR and IG gene sequence data are *primer induced base substitutions* due to multiplex PCR amplification. If a primer designed to amplify a fragment incorporating segment A binds to a fragment incorporating segment B due to similarity in the primer sequence, the replicated sequence will have false base incorporations making it more similar to segment A at the primer binding site. Generally, this only plays a role for J segment identification, since for single-end (SE) V(D)J reads the V primer site is usually not part of the read and for paired-end data there is sufficient V segment sequence present in the V read to safely

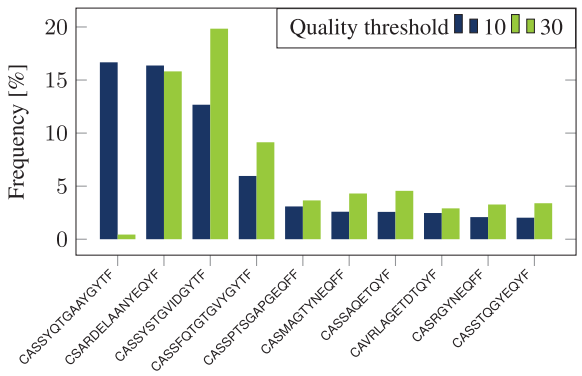

Fig. 3. A dataset from an antigen specific sample with frequencies for the top 10 clonotypes. The reads were initially filtered according to average quality scores, once with a threshold of 10 and once with a threshold of 30 [CMV specific (pp65+) T cells sequenced on an Illumina GAIIx sequencer, 150 bp SE]

identify the V segment. Therefore, immunogenetic sequence analysis tools should provide methods to quantify this error in the data and accurately identify the gene segments despite such base substitutions.

1.3 Available tools

Many resources supporting research on immunogenetic sequence data have been made available by the International Immunogenetics Information System (IMGT), including a database providing sequences and annotations of known gene segments (Giudicelli *et al.*, 2005) as well as Web-based tools for the analysis of recombined receptor gene sequences, *V-QUEST* (Brochet *et al.*, 2008) and *HighV-QUEST* (Li *et al.*, 2013). *IgBlast* (Ye *et al.*, 2013) provides a similar functionality as V-QUEST and is also available as a standalone application, as is the software *Decombinator* (Thomas *et al.*, 2013). These tools, however, do not attempt to characterize the TCR repertoires, i.e. they only perform clonotyping per read and do not incorporate further error corrections that take the entire repertoire into account. Tools that include error correction mechanisms are *MiTCR* (Bolotin *et al.*, 2013) and *MIGEC* (Shugay *et al.*, 2014); however, the latter requires additional experimental effort by using primers with random identifier sequences, so called *unique molecular identifier (UMI)* tags, in order to do so. Another tool, *TCRklass* (Yang *et al.*, 2015), although performing some error correction doesn't correct false bases inside the CDR3 region.

1.4 Scope of IMSEQ

IMSEQ is designed to identify the clonotypes of recombined TCR or IG genes from NGS data, to correct PCR and sequencing errors, and to count identical clonotypes to characterize the repertoire of clones in a population of cells. We will explain the algorithms and the use of the software in the context of the TCR beta gene (TCRB, Entrez Gene ID 6957 for *Homo sapiens*) in detail, but the TCR alpha chain or the IG genes, as well as the corresponding genes of other species, can be investigated simply by substituting FASTA files with the corresponding sequences. Our method requires every gene copy present in the sample to be represented by a read covering the CDR3 region of the gene as well as parts of the flanking V and J regions. Additionally, it can process a second, paired read covering only the V region. If the read pairs overlap, they must be pre-processed with a paired read overlapper such as PEAR (Zhang *et al.*, 2014) and presented to IMSEQ as single end data. The different configurations supported by IMSEQ are illustrated in Figure 2.

2 Methods

Processing the input read data happens in three stages. At first, a **pre-processing** step filters reads based on quality constraints. The user can specify a minimum average quality score, reads that fall below that value are discarded. As discussed in Section 1.2, this threshold should not be set too high, however, some threshold, e.g. 10, should be set in order to remove entirely erroneous reads. Furthermore, **clonotyping** assigns each read to a clonotype determined according to its V and J segment identity and the sequence of the CDR3 region. Finally, in the **repertoire generation** step, identical clonotypes are consolidated and counted. If desired, these clonotype clusters can be merged further to correct for sequencing and PCR errors. IMSEQ outputs both the repertoire in form of clonotype counts as well as detailed information containing the clonotype or reason for rejection for every single input read. The read clonotyping and repertoire generation steps are described in more detail in the following sections.

2.1 Read clonotyping

The initial step of the read clonotyping is the identification of the V and J gene segments. For that purpose, the best matching segments are identified by pairwise alignments between the *read sequence* and the germline V and J *segment sequences*. Although other tools implement alignment free segment identification approaches (Thomas *et al.*, 2013; Yang *et al.*, 2015) in order to improve the clonotyping runtime, only alignment based approaches can actually output detailed per read information regarding exact error positions in the V and J segment alignments. This can be valuable information for the assessment of the data quality or base substitutions induced by unspecific primer binding, as discussed in Section 1.2. An example for the analysis of such effects from the IMSEQ output is shown in Section 5.3. To accelerate the alignment based identification process, the segment sequences are pre-processed and a fast alignment filter is used to exclude sequences that cannot yield a good alignment with the read and reduce the required computations for those that can.

2.1.1 Segment core fragment matching

To identify the V and J segments that yield the best scoring overlap alignments against the read sequence in an efficient way, IMSEQ initially matches a set of short segment substrings, denoted as *segment core fragments (SCFs)*, against each read. The SCFs are built in a pre-processing step from the V and J segment sequences and have a fixed length and location relative to the Phe/Cys triplet within the segments and are not necessarily unique. They are chosen to be small enough to enable a semi-global alignment against the read sequences, which can be computed efficiently using filtering and alignment score computation methods as described below. Only the SCFs alignments with at most $\delta_{\max}^{\mathrm{V-SCF}}/\delta_{\max}^{\mathrm{J-SCF}}$ errors are then extended to full segment overlap alignments, which drastically reduces the number of full overlap alignments that have to be computed. The pre-computed SCF alignments form the innermost part of the final overlap alignments, i.e. for V segments the SCF alignment is extended in the 5′ direction and for J segments in the 3′ direction.

For every read, the set of SCFs is reduced to those that can actually yield a semi-global alignment with the read sequence within the specified error margins using the SWIFT filter (Rasmussen *et al.*,

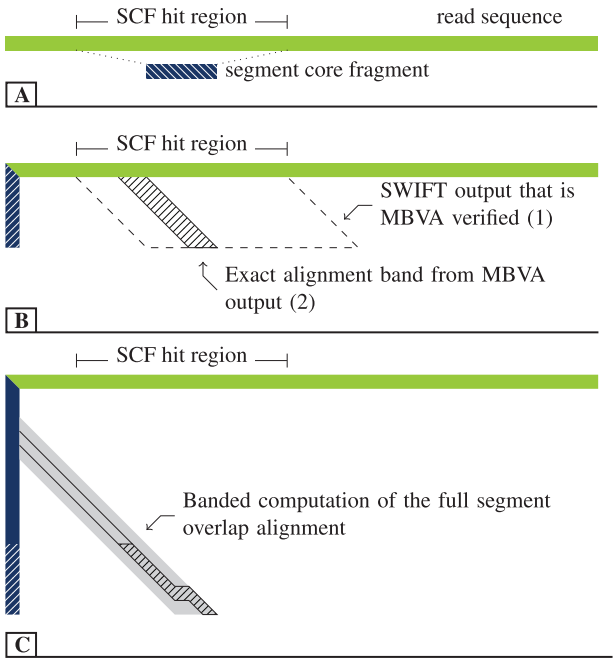

Fig. 4. The three stages of V segment identification (J is analogous except for the positioning of the J segment within the read). **(A)** Given the read, the SWIFT filter returns a hit region for one or more SCFs. **(B)** Each SCF with a hit region is verified using Myers bitvector algorithm. A full alignment is computed for an SCF only it has a match within the user specified error rate. **(C)** The full overlap alignment between the read and the segment sequence is computed within a band derived from the exact SCF match location determined by MBVA and the desired maximum error rate. The best overall alignment is reported as the matching segment

2006). It is fully sensitive, i.e. it guarantees to report every SCF that might yield a match. The SWIFT filter returns not only a set of candidate SCFs but also the corresponding potential *hit region* within the read, i.e. it reduces the search space within the read sequence as illustrated in Figure 4A. Whether or not there is an occurrence of the SCF in the hit region within the error threshold is verified using Myers bit vector algorithm (MBVA) (Myers, 1999) (Fig. 4B). Its asymptotic runtime is linear in the length of the hit region if the SCF length is chosen such that it does not exceed the machine word size of the underlying architecture in bits. It does not compute an alignment, but can determine the score as well as the begin and end position of the best alignment.

Because the SCF alignments later form the innermost part of the segment overlap alignments, the position of the SCFs defines if and how far the alignment reaches into the CDR3 regions. Because bases inside the CDR3 region undergo modification during the recombination, they shouldn't be considered for segment identification. However, since the J segment region of the read often suffers from base substitutions originating from unspecific PCR primer binding, it can make sense to include a few bases from the CDR3 region to increase the segment identification precision. Therefore, the V alignment is configured to end with the Cys triplet while the J alignment starts 6 bp before the Phe triplet. In turn, the default values for allowed SCF errors are set to $\delta_{\max}^{\mathrm{V-SCF}} = 1$ and $\delta_{\max}^{\mathrm{J-SCF}} = 2$. The J SCF length is set to 12 bp and the V SCF length is set based on the shortest read in the input data by default. For reads with a length of at least 120 bp it is set to 20 bp, otherwise it is reduced down to 10 bp if reads of length 100 bp or shorter occur in the input data. All parameters can be modified by the user.

2.1.2 Alignment computation and segment assignment

Last, based on the exact alignment location of the SCF and the maximum segment match error rates $e_{\max}^{\mathrm{V}}$ and $e_{\max}^{\mathrm{J}}$, IMSEQ computes a full overlap alignment between the read and the segment sequence in a banded fashion (Chao *et al.*, 1992). The maximum error rates can be specified by the user and should be chosen generously according to the expected error rate of the underlying PCR protocols and sequencing platform. The default values are $e_{\max}^{\mathrm{V}} = 0.05$ and $e_{\max}^{\mathrm{J}} = 0.15$, since the J-segment area of the read often suffers from primer-induced base substitutions. The alignment computation step is illustrated in Figure 4C.

If an SCF originates from more than one gene segment, the alignment has to be computed for all of these segments. In the end, the best scoring overlap alignment is used to assign the gene segment to the clonotype. If no alignment within the user-specified error rate boundary can be found, the read is rejected. If there is no unique best matching segment, multiple segments are assigned. For the V segment, this can be due to insufficient read lengths, i.e. the read does not cover enough of the gene segment to uniquely identify the incorporated segment. This can be addressed by non-overlapping paired-end sequencing, which can resolve most ambiguity even at the same overall read length, since the forward read can be used to cover upstream regions with a higher degree of sequence divergence (CDR1 and CDR2) (Tonegawa, 1983) and thus identifiability. IMSEQ supports this approach and also implements another strategy to handle V ambiguity, as described below.

2.1.3 Non overlapping paired-end segment alignment

If each VDJ-covering read has an additional paired-end read covering only the V region of the read (Fig. 2B), the SCF matching and alignment computation as described earlier is omitted for the V segment. Instead, a combination of the SWIFT filter and MBVA is used to find the best match of the V read against the set of reference V segments, and then to compute an overlap alignment with the VDJ-read.

2.1.4 CDR3 identification

After the V and the J segment have been identified, the CDR3 region is determined based on the Cys_{104} and Phe_{118} triplet positions known from the V and J segment sequences. If these triplets are out of frame or if a stop codon occurs inside the CDR3 region, the read is rejected as non-functional. Each rejected read as well as the reasons for the rejection are recorded in a log file.

2.2 Repertoire generation

After all input reads have been clonotyped, the clonotype repertoire is derived from the counts of all uniquely occurring clonotypes. The number of unique clonotypes identified is generally over-estimated due to PCR and sequencing errors as well as V or J segment identification ambiguities. To resolve this, the repertoire generation involves several correction steps that cluster clonotypes in order to correct these errors. IMSEQ offers three clustering methods, two of which are designed to cluster clonotypes with different CDR3 sequences (*simple clustering* and *quality clustering*) and one that corrects for ambiguously identified V or J segments, the *segment ambiguity clustering*.

2.2.1 Segment ambiguity clustering

As described in Section 2.1.2, for some reads the V- or J-segment cannot be uniquely identified and multiple segments are assigned to that read. If sufficiently many other reads that originated from the

same clonotype can be unambiguously assigned, then it is possible to correct the ambiguous reads under the assumption that relatively few clonotypes share the same CDR3 region and that this is unlikely to coincide with overlapping V- and J-segment assignments. Thus, in order to correct for additional clonotypes induced by such segment matching ambiguity, we partition the set of unique clonotypes $\mathcal{C}$ into sets of clonotypes that share the same CDR3 sequence, $\mathbf{C} = \{\mathcal{C}_1^*, \cdots, \mathcal{C}_n^*\}$. For every ordered pair of clonotypes (c_1, c_2) where $c_1, c_2 \in \mathcal{C}_i^*, c_1 \neq c_2$ IMSEQ checks whether the set of identified V and J segments of c_1 are a subset of or equal to the corresponding sets of c_2, i.e. $V(c_1) \subseteq V(c_2) \wedge J(c_1) \subseteq J(c_2)$. If that is the case, the pair (c_1, c_2) is added to the set of correctable clonotype pairs $\mathcal{P}$. After all pairs in one clonotype subset have been checked, the count of every clonotype c_e is distributed among all $c^* \in \{c_i \mid (c_i, c_e) \in \mathcal{P}\}$, maintaining the frequency ratios among these clonotypes. This procedure is repeated for all clonotype subsets contained in $\mathbf{C}$.

2.2.2 Quality clustering

Erroneous base calls inside the CDR3 region generate new clonotypes, i.e. increase the repertoire size, and falsify the clonotype frequencies, as they affect some clonotypes more than others due to the sequence dependency of sequencing errors (Dohm *et al.*, 2008). To keep this bias as low as possible, reads with low quality bases should also be incorporated into the analysis. This requirement, however, increases the demand for error correction inside the CDR3 sequence. Therefore, we perform a post-processing step that checks for every identified clonotype cluster whether it is likely to be erroneously derived from another clonotype cluster. Let $q_k(c)$ denote the mean sequencing quality of the kth base within the CDR3 region among all reads that were clonotyped as c and $\overline{q}(c)$ and $\sigma_{q(c)}$ denote the mean and SD of these mean qualities. In order to consider clonotype cluster c_e to be an erroneous version of c_i, we demand that the same V and J segments were identified in c_e and c_i, that their CDR3 sequences are of the same length and differ only in e^q_{max} positions and that

$$q_k(c_e) \leq \overline{q}(c_e) - s_{min} \cdot \sigma_{q(c_e)}$$

for all $k \in \mathscr{E}(c_i, c_e)$, where s_{min} and e^q_{max} are user-defined parameters (see Section 2.2.4) and $\mathscr{E}(c_i, c_e)$ is the set of positions at which c_i and c_e differ. Clonotype pairs that fulfill these conditions are stored and the counts of erroneous clonotype clusters redistributed as described in Section 2.2.1.

2.2.3 Simple clustering

Errors in the multiplex PCR or library preparation steps are not associated with reduced sequencing quality scores, but usually display a low edit distance to a genuine clonotype. Additionally to the quality dependent clonotype clustering approach described in Section 2.2.2, we therefore also implemented the same clustering approach without the quality score requirement, i.e. simply based on a threshold e^s_{max} for the maximum number of errors inside the CDR3 region. In contrast to quality clustering, this step can therefore correct PCR errors. However, there is a risk that this clustering method might erase true positive clonotypes that are highly similar to other clonotypes.

2.2.4 Clustering parameters

The two clustering methods previously described are performed in a combined fashion. That is, given two clonotypes that differ in given number of positions inside the CDR3 region, if only some of them can be explained by a low-quality score clustering will still be performed if the number of remaining errors is less or equal than that allowed in the simple clustering step. The default parameters for the number of allowed errors are $e^q_{max} = 4$ and $s_{min} = 1$ for quality based and $e^s_{max} = 1$ for simple clustering. The parameters can be adjusted by the user. Optionally a maximum ratio $n(c_e)/n(c_i) \leq r^s_{max}$ can be configured, where $n(c_e)$ denotes the count of the erroneous clonotype and $n(c_i)$ denotes the count of the target clonotype for clustering.

2.3 Implementation

IMSEQ was implemented in C++ using SeqAn (Döring *et al.*, 2008) and is available for download at www.imtools.org. In the most basic setup it requires the reference V and J segment sequences in FASTA format with annotations that indicate the positions of the Cys_{104} and Phe_{118} triplets in the segments as well as at least one input file in FASTA or FASTQ format. GZIP compression is supported for all input files. The program can write different kinds of output files. The most verbose output file is a tab separated file containing detailed per-read information such as the identified V and J segment(s), the begin and end position of the CDR3 region, the number of V and J bases aligned outside the CDR3 region as well as the error positions among those bases. IMSEQ can also write aggregated files with (potentially corrected) clonotype counts, using either nucleotide or amino acid-based clonotype identities. Furthermore, a file containing the IDs of all rejected reads associated with the rejection reason such as low sequence quality, out of frame CDR3 boundaries or stop codons inside the CDR3 can be written.

3 Clonotyping evaluation

We compared IMSEQ to five other tools: HighV-QUEST (Li *et al.*, 2013), Decombinator (Thomas *et al.*, 2013), MITCR (Bolotin *et al.*, 2013), MIGEC-CdrBlast (Shugay *et al.*, 2014) and TCRklass (Yang *et al.*, 2015). TCR beta chain genes were simulated as described in the following section, and Illumina reads were simulated using Mason 2 (Holtgrewe, 2010). Additionally, data from a real sequencing experiment was analyzed.

3.1 Data simulation

We simulated VDJ-recombined human TCRB genes by implementing a simple model: The V, D and J segments are chosen uniformly from the set of functional gene segments as defined by the IMGT (Giudicelli *et al.*, 2005). For every junction and involved segment, n_d nucleotides are removed from the segment end, n_p nucleotides are added as a complementary palindrome and n_n nucleotides are added randomly. For every junction, an overlap of n_o nucleotides is created between the involved segments and nucleotide mismatches are resolved randomly. The distributions for the parameters n_d, n_p, n_n and n_o were manually tuned such that the CDR3 length distribution matches the one observed in five deeply sequenced (average 7.5 mio. reads), unspecific TCRB repertoires from different donors. To incorporate PCR errors, a simple replication model was used where in every cycle every sequence has a probability of P_r to be replicated and every base incorporation has a probability of P_e to to result in a random false base. Insertions and deletions were not considered in the PCR model.

3.2 Dataset description

We generated two human TCR-β gene datasets, one with 5×10^5 unique clonotypes and one with 150 unique clonotypes. To simulate

the data as realistically as possible, the reference segments passed to the simulator were truncated according to the enrichment PCR primer sites used by Dziubianau *et al.* (2013). As some of the tools come with their own gene segment reference data, in order to exclude effects of different reference segment datasets, only segments denoted as 'functional' in the IMGT reference database were used during the simulation, as this can be considered the minimal set of gene segments that should be supported by all tools. The first dataset was kept *error-free* and used to validate the basic clonotyping functionality of each tool tested. The dataset was also tested against IMGT HighV-QUEST in order to ensure that the clonotype simulation generated valid sequences. The second dataset was in-silico PCR-amplified to a total size of 1.5×10^5 sequences using a per cycle amplification probability of $P_a = 0.8$ and a per-base substitution probability of $P_e = 10^{-4}$, referred to as *PCR*. Finally, another dataset was generated, introducing sequencing errors into each of the PCR amplified fragments and reducing the fragment length to 150 bp in order to establish conditions that correspond to real sequencing as closely as possible. This step was performed using Mason 2 (Holtgrewe, 2010), using its default error profile options. This dataset is referred to as *PCR + Seq*. Although the simulator provides paired-end data, only the VDJ-read was used for the analysis since not all of the tools in our comparison can handle paired-end data. A separate evaluation showing the improvement of split paired-end (SPE) workflows is presented in Section 4.

Additionally, data originating from a real experiment was analyzed to show the runtime performance in real life applications as well as differences in the estimated number of clonotypes. The dataset originates from antigen unspecific $CD4^+$ T cells which were prepared according to the protocol described by Dziubianau *et al.* (2013) and sequenced on an Illumina HiSeq sequencing system. The dataset is referred to as *real data* and contains 3 504 203 reads.

3.3 Evaluation criteria

To assess the clonotyping and repertoire generation performance, two different evaluations were performed. One assesses the performance only on a clonotype level, i.e. discarding the frequencies of the clonotypes, and the other evaluation assesses the per-read clonotyping performance.

3.3.1 Clonotype level evaluation

In many studies it is crucial to know whether or not a clonotype is present in a sample, while its exact frequency is less important. That includes different measures to assess the diversity of a sample, which are affected by over- or underestimation of the number of clonotypes or comparative methods that take into account whether or not two samples share a clonotype. Hence, if clonotypes are falsely removed from a repertoire, i.e. due to over-clustering or biased clonotyping, or if insufficient error handling enlarges the number of clonotypes in the repertoire, the performance of these methods will suffer. Therefore, one of the evaluations we performed is a binary classification based on the presence and absence of clonotypes in the repertoires generated by the different methods.

3.3.2 Per read clonotyping performance

Because the competing tools do not provide per-read result information, the performance was estimated based on the clonotype frequencies. That is, given the simulated count of a clonotype n_c and the detected count of that clonotype n_c', we define the number of true positive, false positive and false negative read to clonotype assignments for a particular clonotype c as

$$\mathrm{FP}_c = \max(0, n_c' - n_c)$$
$$\mathrm{FN}_c = \max(0, n_c - n_c')$$
$$\mathrm{TP}_c = \min(n_c, n_c').$$

The overall number of events is computed as the sum of events over all simulated and detected clonotypes.

3.4 Tool parameters

The tools DECOMBINATOR, HighV-QUEST and MIGEC-CdrBlast were run with their default parameters. None of these tools support error correction and therefore could not be parameterized in that respect. Each tool comes with their own set of reference sequences for human TCRB gene segments. To assess only the core clonotyping capability, the correction-capable tools were instructed to skip clustering (MITCR maxClusterizationRatio = 0, IMSEQ -qcme 0 -scme 0). For the erroneous datasets PCR and PCR + SEQ and the real data, MITCR was invoked with default parameters and the IMSEQ parameters were set as indicated.

4 Paired-end evaluation

To assess the impact of an SPE protocol versus an SE protocol under real world conditions, we analyzed 14 different (8 antigen specific and 6 polyclonal) samples on an Illumina MiSeq machine in paired-end mode with 100 bp forward reads and 200 bp reverse reads. The reads were then processed using IMSEQ, once in a 100/100 bp paired-end setting and once in a 200 bp SE setting. For each setting, the analysis was performed with and without segment ambiguity clustering (Section 2.2.1). The results depicted in Figure 5 show the proportion of V segment ambiguity removed relative to the 200 bp SE analysis without segment ambiguity clustering.

5 Results

5.1 Clonotyping and error correction

The results of the clonotyping performance for all tools are shown in Table 1. The evaluation of the *error-free* dataset analysed by IMGT HighV-QUEST shows that the simulated TCRB gene sequences are valid and correctly labelled according to the IMGT naming scheme. IMSEQ can correctly clonotype all provided

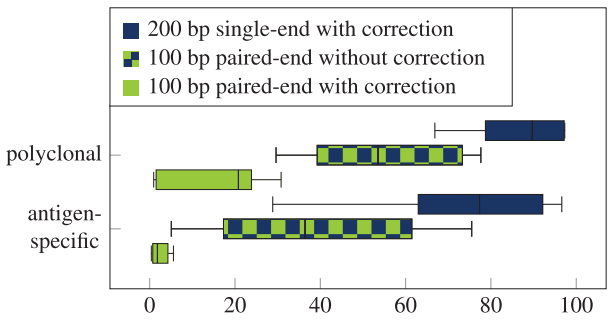

Fig. 5. Impact of paired-end sequencing on V segment ambiguity. Datasets from eight influenza specific and six polyclonal samples were analyzed in four ways: 200 bp SE and 100 bp paired-end, each with and without V ambiguity correction. For each sample type the percentage of remaining V-ambiguous clonotypes compared with the 200 bp SE data with no correction is shown

sequences in absence of any errors and also TCRklass reports only correct clonotype assignments, missing a small fraction of the reads resulting in a recall of 99.9%. MIGEC-CdrBlast also derives almost all clonotypes correctly with a very low number of false positive and false negative calls. MITCR fails to clonotype some reads correctly, resulting in about 94% precision and recall for both reads and clonotypes. Decombinator clonotypes only 57% of the reads correctly, generally having difficulties to define the CDR3 region correctly, which confirms previous results by Bolotin *et al.* (2013). Decombinator was therefore excluded from the subsequent evaluations. IMGT HighV/QUEST was only included to verify the validity of the simulated data, but was not included in the subsequent tests because it is of no practical use for the analysis of real world NGS data—it is limited to 500 000 reads, performs only per read clonotyping without repertoire generation and error correction, there is no standalone software and the online analysis of a dataset usually takes days.

For sequences with PCR errors, IMSEQ has the highest read level recall and precision even without any posterior clustering (99% each). It recovers all clonotypes, however, generates additional clonotypes which aren't corrected for without clustering (CT precision 79%). MIGEC-CdrBlast, which doesn't support error clustering, performs similarly. It is unclear whether TCRklass performs any kind of error correction inside the CDR3 region, however, it's performance is only slightly better than that of MIGEC-CdrBlast, indicating that it doesn't. With simple clustering enabled, IMSEQ outputs only correct clonotype assignments and covers 99% of the input reads. MITCR again performs worse than IMSEQ with about 95% precision and recall for both reads and clonotypes.

The evaluation with 150 bp sequencing reads simulated from the dataset containing PCR errors firstly illustrates the impact of posterior error correction. MIGEC-CdrBlast and IMSEQ without clustering detect about 10 000 unique clonotypes, whereas only 150 are actually present in the sample. Again, TCRklass performs only slightly better than these tools, overestimating the total number of clonotypes to be 8666. With both clustering methods enabled, IMSEQ detects 185 clonotypes with a 100% recall and 81% precision. On the read level, precision and recall reach 99%. MITCR performs again about 5% worse than IMSEQ in read precision as well as read and clonotype recall, the clonotype precision drops to 14% due to the many false positive clonotypes called. Overall, MITCR detects 1026 unique clonotypes.

Finally, the experimental data revealed a large variability regarding the number of clonotypes detected across the different tools. As expected, the tools that do not perform any correction of sequencing and PCR errors report the largest number of clonotypes (MIGEC-CdrBlast 34 006 and TCRklass 37 514 clonotypes). However, there is also a more than threefold difference between MITCR (6213 clonotypes) and IMSEQ (20 725). Manual inspection revealed that MITCR clusters all clonotypes with the same CDR3 sequence—even if they have clearly distinct V or J segments incorporated and additionally fails to detect some, even highly abundant clonotypes for no apparent reason. All tools were able to assign clonotypes to ~72% of the reads, except for MIGEC-CdrBlast, which detected clonotypes for only 65% of the input sequences.

The runtimes were measured on a dual 6-core Intel Xeon X5675 (3.06 GHz) Mac Pro system with 64 GB of RAM running Mac OS 10.9. This does not apply to IMGT/HighV-QUEST, which is a web-service hosted by the IMGT consortium. IMSEQ, MITCR and MIGEC-CdrBlast are multi-threading capable and were therefore run utilizing all 24 CPUs (including hyper-threading) presented by the operating system. Among the locally run tools regarding the 500 000 unique sequences dataset (*error-free*), IMSEQ processed the data in 37 s, followed by MITCR requiring 57 s. The other tools were significantly slower, i.e. the DECOMBINATOR analysis took 231 s, MIGEC-CdrBlast 393 s and TCRklass 341 s.

The PCR dataset was processed in 3 s by MITCR and IMSEQ, MIGEC-CdrBlast required 7 s and TCRklass 99 s. The PCR + Seq

Table 1. Evaluation results for four different datasets: 500 000 unique error-free clonotypes (first box), 150 unique clonotypes amplified up to 150 000 sequences, with and without simulation of errors induced by sequencing (second and third box) and a dataset originating from a real experiment (fourth box). The best values for each dataset are highlighted in bold

Tool	Options	Dataset	Clonotype level No. CTs	Precision	Recall	Read level Precision	Recall	Time
DECOMBINATOR	default	error-free	498 996	0.6456	0.5703	0.5757	0.5745	231 s
MITCR	no clust	error-free	498 415	0.9549	0.9418	0.9448	0.9418	57 s
MIGEC-CdrBlast	default	error-free	499 987	0.9987	0.9987	0.9987	0.9987	393 s
TCRklass	default	error-free	499 395	**1.0000**	0.9988	**1.0000**	0.9988	341 s
HighV-QUEST	no clust	error-free	**500 000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	>24 h
IMSEQ	no clust	error-free	**500 000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	37 s
MIGEC-CdrBlast	default	PCR	189	0.7884	0.9933	0.9887	0.9886	7 s
TCRklass	default	PCR	185	0.8054	0.9933	0.9970	0.9951	99 s
MITCR	default	PCR	151	0.9536	0.9600	0.9521	0.9519	3 s
IMSEQ	no clust	PCR	190	0.7895	**1.0000**	0.9969	0.9969	3 s
IMSEQ	$e^{s}_{max} = 1$	PCR	**150**	**1.0000**	**1.0000**	**1.0000**	**0.9999**	3 s
MIGEC-CdrBlast	default	PCR + Seq	10 413	0.0143	0.9933	0.8932	0.8859	40 s
TCRklass	default	PCR + Seq	8666	0.0172	0.9933	0.9154	0.8938	79 s
MITCR	default	PCR + Seq	1026	0.1404	0.9600	0.9464	0.9215	4 s
IMSEQ	no clust	PCR + Seq	9541	0.0157	**1.0000**	0.9073	0.8954	6 s
IMSEQ	$e^{s}_{max} = 1$	PCR + Seq	365	0.4110	**1.0000**	0.9985	0.9854	7 s
IMSEQ	$e^{q}_{max} = 4$, $e^{s}_{max} = 1$	PCR + Seq	**185**	**0.8108**	**1.0000**	**0.9991**	**0.9860**	10 s
MIGEC-CdrBlast		real data	34 006	—	—	—	—	228 s
TCRklass	default	real data	37 514	—	—	—	—	1310 s
MITCR	default	real data	6213	—	—	—	—	37 s
IMSEQ	$e^{q}_{max} = 4$, $e^{s}_{max} = 1$	real data	20 725	—	—	—	—	69 s

dataset analysis required 6, 7 and 10 s (no clustering, simple- and combined error correction) using IMSEQ, 4 s using MITCR, 40 s using MIGEC-CdrBlast and 79 s using TCRklass.

Last, the experimentally generated dataset was processed relatively quickly by MITCR (37 s) and IMSEQ (69 s), while MIGEC-CdrBlast required 228 s and the single threaded TCRklass took > 20 min to process it.

5.2 Split paired-end sequencing

The SPE data evaluation shows the impact of SPE sequencing (as depicted in Fig. 2B) on two types of real datasets, eight antigen-specific and six polyclonal repertoires. It shows that 100 bp SPE sequencing in conjunction with posterior segment ambiguity resolution can drastically reduce the V segment ambiguity compared with a SE-sequencing configuration with 200 bp, i.e. the same overall read length. The posterior clustering can correct a limited amount of ambiguity on the 200 bp SE data, i.e. 80–98% of the ambiguous clonotypes remain after correction compared with the 200 bp SE data without correction for the polyclonal data and 62–95% for the antigen-specific data. Running the analysis on the 100 bp SPE data instead reduces the number of ambiguous clonotypes down to 40–75% in the polyclonal and down to 18–61% in the antigen-specific case. In conjunction, i.e. when processing the 100 bp SPE dataset with posterior ambiguity clustering, almost all ambiguity can be resolved in the antigen specific samples (0–4% ambiguous clonotypes remain). In the polyclonal datasets the reduction leaves 20% or less of the ambiguous clonotypes.

5.3 Segment alignment analysis

Because IMSEQ computes the full alignment between the gene segments and the input read, it can also reveal details about those alignments and the errors encountered in the input reads.

Supplementary Figure S1.1A shows the match/mismatch frequencies along the TRBJ 2-7 segment observed in a sample of unspecific CD8 + T cells generated from the detailed IMSEQ output. The genes were amplified using the multiplex PCR primers published by Dziubianau *et al.* (2013). In the underlying sample, >40 000 reads originated from clones that incorporate the TRBJ 2-7 segment. The distribution of alignment errors across the positions clearly shows that positions 17, 27 and 28 mismatch in about 50% of the reads, positions 23 and 26 even in about 75% of the reads. Supplementary Figure S1.1B shows the alignments of the TRBJ 2-7 gene (bases upstream the Phe triplet are omitted) against the primers designed to amplify genes incorporating the TRBJ segments 1-1 and 1-6. All the error positions are explained by unspecific binding of the TRBJ 1-1 primer, while positions 23 and 26 additionally suffer from unspecific binding of the TRBJ 1-6 primer.

An example for the alignment match/mismatch distribution in the V segment is shown in Supplementary Figure S1.1C. When the recombined gene is sequenced from the J segment, i.e. reverse, the coverage is uniform in the J segment while it varies towards the end inside the V segment, depending on the length of the J segment and the CDR3 region. The fewer V segment bases are covered by the read the higher is the chance that, in a SE experiment, the V segment cannot be uniquely identified.

Therefore, IMSEQ's detailed per-read alignment error diagnostics provide valuable information about potential issues regarding the gene enrichment and sequencing to the experimentalist. The data can be used to detect and quantify unspecific primer binding artifacts, analyze the V coverage distribution and also to detect errors in the reference database or new alleles.

6 Conclusion

We have presented IMSEQ, a fast and flexible tool for TCR/IG repertoire analysis from NGS data. Our method is capable of handling SE as well as overlapping and SPE data. Apart from per-read clonotyping it also features error correction methods to resolve false clonotype calls originating from PCR and sequencing errors as well as a method to correct for ambiguously assigned gene segments.

The results produced by IMSEQ were thoroughly validated using simulated data. The tools performance was compared with that of other state of the art standalone TCR/IG clonotypers available, namely DECOMBINATOR, TCRklass, MITCR and MIGEC-CdrBlast. Our evaluations show that the clonotyping capabilities of IMSEQ are superior to those of DECOMBINATOR and MITCR while being comparable to those of MIGEC-CdrBlast and TCRklass. However, IMSEQ substantially outperformed the latter tools with respect to runtime.

When it comes to the error correction performance, IMSEQ could only be compared with MITCR, and showed superior performance especially with regards to the precision of clonotype detection. MIGEC-CdrBlast has additional requirements for the experimental setup, namely the incorporation of UMIs, to be able to perform any kind of error correction. DECOMBINATOR does not support error correction at all and was also excluded from the evaluation due to its poor clonotyping performance. The TCRklass error correction does not seem to include the CDR3 region and therefore does not effectively reduce the number of false clonotype calls due to errors.

The evaluation of the impact of SPE sequencing showed a great improvement in V segment accuracy when compared with SE sequencing with the same overall read length.

To conclude, IMSEQ is a rapid, well tested, flexible and error-aware clonotyping and repertoire generation tool that was shown to be superior to available tools with regard to clonotyping and standalone error correction capabilities.

Funding

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the grants "Primage" (0315895A) to N.B. and "eKid" (01ZX1312A) to N.B. as well as by the Investitionsbank Berlin (IBB) within the "Profit" grant (10142562) to N.B.

Conflict of Interest: none declared.

References

Arstila,T.P. *et al.* (1999) A direct estimate of the human$\alpha\beta$ T cell receptor diversity. *Science*, **286**, 958–61.

Bolotin,D.A. *et al.* (2013) MiTCR: software for T-cell receptor sequencing data analysis. *Nature Methods*, **10**, 813–814.

Brochet,X. *et al.* (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized VJ and VDJ sequence analysis. *Nucleic Acids Res.*, **36**(Suppl. 2), W503–W508.

Chao,K.M. *et al.* (1992) Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.*, **8**, 481–487.

Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

Döring,A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Dziubianau,M. *et al.* (2013) TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell–related pathology. *Am. J. Transplant.*, **13**, 2842–2854.

Giudicelli,V. *et al.* (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**(Suppl. 1), D256–D261.

Holtgrewe,M. (2010) Mason–a read simulator for second generation sequencing data. *Technical report FU Berlin*. Institut für Mathematik und Informatik, Freie Universität Berlin

Jackson,K.J. *et al*. (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host and Microbe*, **16**, 105–114.

Janeway,C.A. *et al*. (1999) *Immunobiology: The Immune System in Health and Disease*. Current Biology Publications, New York, NY.

Li,S. *et al*. (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.*, **4**, 2333.

Myers,G. (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, **46**, 395–415.

Rasmussen,K.R. *et al*. (2006) Efficient q-gram filters for finding all ε-matches over a given length. *J. Comput. Biol.*, **13**, 296–308.

Shugay,M. *et al*. (2014) Towards error-free profiling of immune repertoires. *Nat. Methods*, **11**, 653–655.

Sims,D. *et al*. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.

Thomas,N. *et al*. (2013) Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, **29**, 542–550.

Tonegawa,S. (1983) Somatic generation of antibody diversity. *Nature*, **302**, 575–581.

Vollmers,C. *et al*. (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci.*, **110**, 13463–13468.

Wu,D. *et al*. (2012) High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Trans. Med.*, **4**, 134ra63.

Yang,X. *et al*. (2015) TCRklass: a new k-string–based algorithm for human and mouse TCR repertoire characterization. *J. Immunol.*, **194**, 446–454.

Ye,J. *et al*. (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, **41**, W34–W40.

Zhang,J. *et al*. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.