Published online 13 July 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3 **1**
https://doi.org/10.1093/nargab/lqac049

Benchmarking computational methods for B-cell receptor reconstruction from single-cell RNA-seq data

Tommaso Andreani [1,*], Linda M. Slot [2], Samuel Gabillard[3], Carsten Strübing[4], Claus Reimertz[4], Veeranagouda Yaligara[5], Aleida M. Bakker[2], Reza Olfati-Saber[6], René E.M. Toes[2], Hans U. Scherer [2], Franck Augé[7] and Deimantė Šimaitė[1]

[1]AI & Deep Analytics—Omics Data Science, Sanofi, Frankfurt am Main 65926, Germany, [2]Department of Rheumatology, Leiden University Medical Center, 2333 RC Leiden, The Netherlands, [3]Life & Soft, Le Plessis-Robinson, Paris 92260, France, [4]Immunology & Inflammation Research, Sanofi, Frankfurt am Main 65926, Germany, [5]Molecular Biology & Genomics, Translational Science Unit, Sanofi, Chilly-Mazarin 91385, France, [6]AI & Deep Analytics, Sanofi, Cambridge, MA 02142, USA and [7]AI & Deep Analytics—Omics Data Science, Sanofi, Paris 91385, France

Received January 25, 2022; Revised May 30, 2022; Editorial Decision June 07, 2022; Accepted June 21, 2022

ABSTRACT

Multiple methods have recently been developed to reconstruct full-length B-cell receptors (BCRs) from single-cell RNA sequencing (scRNA-seq) data. This need emerged from the expansion of scRNA-seq techniques, the increasing interest in antibody-based drug development and the importance of BCR repertoire changes in cancer and autoimmune disease progression. However, a comprehensive assessment of performance-influencing factors such as the sequencing depth, read length or number of somatic hypermutations (SHMs) as well as guidance regarding the choice of methodology is still lacking. In this work, we evaluated the ability of six available methods to reconstruct full-length BCRs using one simulated and three experimental SMART-seq datasets. In addition, we validated that the BCRs assembled *in silico* recognize their intended targets when expressed as monoclonal antibodies. We observed that methods such as BALDR, BASIC and BRACER showed the best overall performance across the tested datasets and conditions, whereas only BASIC demonstrated acceptable results on very short read libraries. Furthermore, the *de novo* assembly-based methods BRACER and BALDR were the most accurate in reconstructing BCRs harboring different degrees of SHMs in the variable domain, while TRUST4, MiXCR and BASIC were the fastest. Finally, we propose guidelines to select the best method based on the given data characteristics.

INTRODUCTION

The recent development of single-cell RNA sequencing (scRNA-seq) techniques has enabled the quantification of genes expressed in individual cells. This has contributed to the identification of different cell types characterizing tissues and to the discovery of previously unknown cell populations, in which the underlying gene expression programs were found to be critical for embryonic development, autoimmune disease pathogenesis and an in-depth understanding of the tumor microenvironment (1–4). Similarly, following the increasing evidence of the importance of B lymphocytes in health and disease, single-cell technologies have been applied to quantify the expression levels of genes, coding for heavy and light chains (HC and LC) of a B-cell receptor (BCR) and thus the BCRs/antibodies these cells produce.

Human BCRs consist of a pair of independent HC and LC that are interconnected by disulfide bonds, each of which contains both constant (C) and variable (V) regions, genetically encoded in three different loci. The immunoglobulin heavy chain locus (IGH) on chromosome 14 contains gene segments for the immunoglobulin HC, whereas LC genes are encoded by two loci: the immunoglobulin kappa (κ) chain locus (IGK) on chromosome 2 and the immunoglobulin lambda (λ) chain locus (IGL) on chromosome 22. During the HC somatic recombination, one of the diversity (D) gene segments is joined to one of the joining (J) gene segments in an event called the D–J recombination (5). Afterward, the D–J segment binds one of the variable segments and all constant regions are retained at the end of the mRNA to produce a functional HC. Since LC does not have a D segment, only the V–J recombination occurs. During these processes, random nucleotides are added into the V(–D–)J joining regions, resulting in a higher

*To whom correspondence should be addressed. Tel: +49 6930526381; Email: Tommaso.Andreani@sanofi.com

number of HC and LC than it would be possible by simply joining gene segments available at each HC and LC gene's locus. This process, in fact, can generate an immunoglobulin repertoire of $>5 \times 10^{13}$ different antigen specificities (5). The introduction of these new nucleotides is particularly challenging for BCR assembly algorithms because of the randomness of such introductions and their absence in the reference genome and in immunoglobulin annotation databases.

The characterization of BCR repertoires from scRNA-seq data has been instrumental in the investigation of groups of B cells sharing a common ancestor, called clonotypes. An overrepresentation of pathogenic clonotypes was noticed in different diseases such as breast cancer (6), multiple sclerosis (MS) (7) and acute myeloid leukemia (AML) (8). In these studies, the presence of repertoires with expanded clonotypes defined different tumor microenvironments in breast cancer, was associated with increased inflammation in MS and could be used to stratify AML patients. In addition, BCRs are known to acquire somatic hypermutations (SHMs) in their variable domain [defined as four framework regions (FWRs) and three complementarity-determining regions (CDRs)] after activation of the B cell by an antigen. In the process of affinity maturation, the affinity for a given antigen can be enhanced by introducing somatic mutations predominantly in the CDRs. For this reason, BCRs of memory B cells display SHMs that are not germline encoded when compared to naïve B cells. As memory B cells represent the cell population that has been triggered by a given antigen and hence are mostly connected to processes studied in e.g. autoimmunity or infectious diseases, these cells often gain most interest in studies investigating BCR composition. For example, an increasing number of SHMs in anti-citrullinated protein antibodies during rheumatoid arthritis (RA) development (9) have been described together with a high frequency of defined sequences called N-glycosylation sites that can potentially be used as a predictor of RA progression (10). Furthermore, BCRs of patients with diffuse large B-cell lymphoma harbor variable levels of SHMs in the variable regions of IGH and IGL/IGK genes. Here, high levels of clonal IGHV SHMs were associated with a prolonged overall survival of patients, whereas an increased CDR3 length of HC and the presence of IGHV ongoing SHM were associated with poor prognosis (11). Overall, these studies show the importance of patients' BCRs in the pathogenesis of multiple diseases and the potential usage of their characterization as a proxy in personalized medicine.

Previous sequencing techniques targeting the immunoglobulin genes, such as Ig-seq (12), allowed the quantification of the entire set of genes belonging to the HC and LC in the total population of B cells (also called BCR repertoire), giving a snapshot of the repertoire composition. However, the inability to obtain full-length BCRs consisting of the variable domain of an HC–LC pair set in an individual cell has been a limiting factor of such approaches. Nowadays, there are several different scRNA-seq technologies available. These can be plate based, which are generally low throughput but can be used to sequence full-length transcripts, or droplet based, allowing to sequence thousands of cells at the same time. Two plate-based approaches, namely SMART-seq (13,14) and a modification of it named SPEC-seq (15), have been instrumental in BCR sequencing and full-length BCR reconstruction due to their ability to obtain full-length transcripts of HC and LC genes allowing the reconstruction of the variable domain of an HC–LC pair set in a single cell. The droplet-based 10x Genomics Chromium Single Cell Immune Profiling Solution [CG000148_10x_Technical_Note(ctfassets.net)] also enables sequencing of the V–D–J genes of a B cell to obtain paired HC and LC. However, this comes with the cost of losing the full-length BCR information for a considerable number of single cells (16).

Although it is not feasible to determine a pair of HC and LC constituting a BCR of a cell using bulk RNA-seq techniques, several methods have been proposed to delineate BCR composition from complex datasets. Methods such as ImRep (17), V'DJer (18), TRUST (19–21) and Imonitor (22) are only suitable for the reconstruction of CDR3 of the variable domain of HC and LC. Therefore, BCRs, reconstructed using the above methods, miss CDR1 and CDR2 regions and the four FWRs. Nonetheless, these regions considerably contribute to the antigen recognition and binding and aid in maintaining the overall structure of an antibody (23). For this reason, the analysis of BCR repertoires from scRNA-seq data required the development of algorithms capable of dealing with highly mutated sequences, dissimilar to the reference genome in order to reconstruct the full variable domain of HC–LC pair of a B cell. Recently, several open-access methods have been proposed for preprocessing of raw SMART-seq and Chromium data to reconstruct BCRs (Table 1). The first developed method was MiXCR (24), which consists of a collection of algorithms based on a proprietary aligner that perform clustering to accomplish BCR reconstruction and annotation. BASIC, a semi *de novo* algorithm (25), was the second method that was made available with the advantage of being able to process libraries as short as 25 bp. Lately, several algorithms based on a *de novo* assembly but using different approaches to map reads and assign V–D–J genes such as BRACER (26), BALDR (27), VDJPuzzle (28) and TRUST4 (29) have emerged, increasing the choice but also the difficulty in selecting the best tool for a given dataset. Importantly, BASIC (25), MiXCR (24), VDJPuzzle (28) and TRUST4 (29) can be used for both BCR and T-cell receptor (TCR) assembly, making them suitable for more elaborate immune repertoire studies.

Given the methodological differences in the algorithms (see Supplementary Data), they can report different results under certain experimental setups. Conditions such as different sequencing technologies, read library properties and the number of SHMs expected within the variable domains of the BCRs can influence the output of the tools. Thus, it is essential to assess the performance of each algorithm and quantitatively understand how sensitive it is in reconstructing BCRs when compared to the 'ground truth' (defined as the original variable domain sequence obtained by classical Sanger sequencing). In addition, given varying numbers of SHMs within the variable domains of BCRs, and the observation of B cells with high SHM counts in several diseases such as follicular lymphoma (30), RA (9) and diffuse large B-cell lymphoma (11), as well as in anti-HIV antibod-

Table 1. Description of the computational methods for BCR reconstruction from scRNA-seq data evaluated in this benchmark

Method	Availability	Language	Type of algorithm	Annotation	Species	Sequencing	Publication
MiXCR	https://github.com/milaboratory/mixcr	Java	Proprietary aligner using *k*-mers and assembler	IMGT	Human, mouse	SMART-seq2	Bolotin *et al.* (24)
BASIC	https://github.com/akds/BASIC	Python	Semi *de novo* with anchors and *k*-mers	IMGT	Human, mouse	SMART-seq2	Canzar *et al.* (25)
BRACER	https://github.com/Teichlab/bracer	Python/R	*De novo*	IMGT, combinatorial recombinome	Human, mouse[a]	SMART-seq2	Lindeman *et al.* (26)
BALDR	https://github.com/BosingerLab/BALDR	Perl	*De novo*	IMGT, combinatorial recombinome	Human, rhesus macaque	SMART-seq2	Upadhyay *et al.* (27)
VDJPuzzle	https://github.com/simone-rizzetto/VDJPuzzle	Roff/Shell	*De novo*	IMGT	Human, mouse	SMART-seq2	Rizzetto *et al.* (28)
TRUST4	https://github.com/liulab-dfci/TRUST4	C++/Python	*De novo* with *k*-mers	IMGT	Human, mouse	SMART-seq2 + 10x	Song *et al.* (29)

[a]Possibility to obtain other species.

ies (31–33), it is crucial to assess the number of SHMs the algorithms can tolerate while still accurately reconstructing BCRs. Finally, the comparison of these methods using several datasets will reveal the stability of their performance across different library preparations.

To address these questions, we designed a comprehensive analysis framework (Figure 1). First, we selected two publicly available BCR sequencing datasets of plasmablast origin with the available ground truth (25,27) and generated an additional dataset with the ground truth, named 'Leiden', using sorted isotype-switched memory B cells that either recognize tetanus toxoid (TT) or have an unknown antigen specificity. Second, using random sampling, we created multiple libraries with different levels of coverage and read length using the above-mentioned experimental datasets. Third, we simulated a fully synthetic dataset (see the 'Materials and Methods' section) in which HC and LC harbor different amounts of SHMs. Subsequently, we used these datasets to test six available BCR reconstruction methods. We evaluated their abilities to obtain productive HC and LC, defined by the absence of stop codons or out-of-frame V–J junctions (see the 'Materials and Methods' section). This was carried out using the sensitivity as a metric for the experimental data and accuracy as a metric for the synthetic data (see the 'Materials and Methods' section). In addition, we experimentally validated the benchmarked methods by determining that antibodies, generated using the productive sequences, can recognize their intended antigens. Moreover, we measured the time needed to run these algorithms to provide information on time and scalability. Finally, we propose a final performance score for each method by aggregating the results of all the experiments and provide recommendations for the selection of the method that best suits a given dataset and scientific question.

MATERIALS AND METHODS

Experimental datasets

Leiden dataset (SMART-seq2).

Cell sorting, cDNA synthesis, ARTISAN PCR and Sanger sequencing. TT-specific B cells and B cells with unknown specificity were isolated as described before (9). In short, peripheral blood mononuclear cells were isolated by Ficoll-Paque gradient centrifugation and stained with Fixable Violet (405 nm) Dead Cell Stain Kit (Thermo Fisher), CD3 Pacific Blue (clone UCHT1, BD Pharmingen), CD14 Pacific Blue (clone M5E2, BD Pharmingen), CD19 APC-Cy7 (clone SJ25C1, BD Pharmingen), CD20 AlexaFluor 700 (clone 2H7, BD Pharmingen), CD27 PE-Cy7 (clone M-T271, BD Pharmingen), IgG BV510 (clone G18-145, BD Horizon), IgD FITC (clone IA6-2, BD Pharmingen) and APC- and PE-labeled TT. $CD19^+CD20^+CD27^+IgG^+IgD^-$ B cells were considered TT-specific if they stained double positive for fluorescently labeled TT with two different fluorochromes: TT-APC and TT-PE. Cells negative for two labeled antigens were considered as 'cells with unknown specificity'. Cells were single cell sorted on a FACS ARIA sorter and mRNA was lysed directly in lysis mix: 0.2% Triton X-100 (Sigma) in ddH_2O, RNase inhibitor (25 U, TaKaRa), oligo-dT30VN (10 pmol, IDT) and dNTPs (10 nmol, Thermo Fisher). cDNA synthesis and subsequent preamplification and purification were performed according to the SMART-seq2 protocol (14). Anchoring Reverse Transcription of Immunoglobulin Sequences and Amplification by Nested (ARTISAN) PCR was performed using purified cDNA.

The ground truth ARTISAN Sanger sequencing of each single cell was performed on an Applied Biosystems 96-capillary (ABI3730xl) sequencer. After counting the HC and the LC of the sequenced 72 single cells, 56 HC, 56 kappa light chains (LcK) and 27 lambda light chains (LcL) were defined as *assembled*. These were HC and LC that did not contain a stop codon in variable and constant regions as defined by IgBLAST and Change-O (34). Furthermore, 38 HC, 48 LcK, 8 LcL and 27 paired HC + LC (K or L) were classified as *productive*. IgBLAST and Change-O (34) define a chain *productive* if it is an *assembled* chain with in-frame V–J junctions. These 27 single cells with paired HC + LC were used as the ground truth to compute the sensitivity of each method for this dataset (Supplementary Figure S1A and Supplementary Table S3).

Canzar dataset (SPEC-seq). We retrieved the dataset included in the publication of BASIC (25) from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116500). Out of 295 single plasmablast cells published in this dataset, we selected 190 that were sequenced using a 50-bp paired-end mode. We further filtered for 113 samples that had at least 1.25 million reads. Besides the fastq files, we obtained fasta sequences generated using Sanger sequencing, which we used as the ground truth

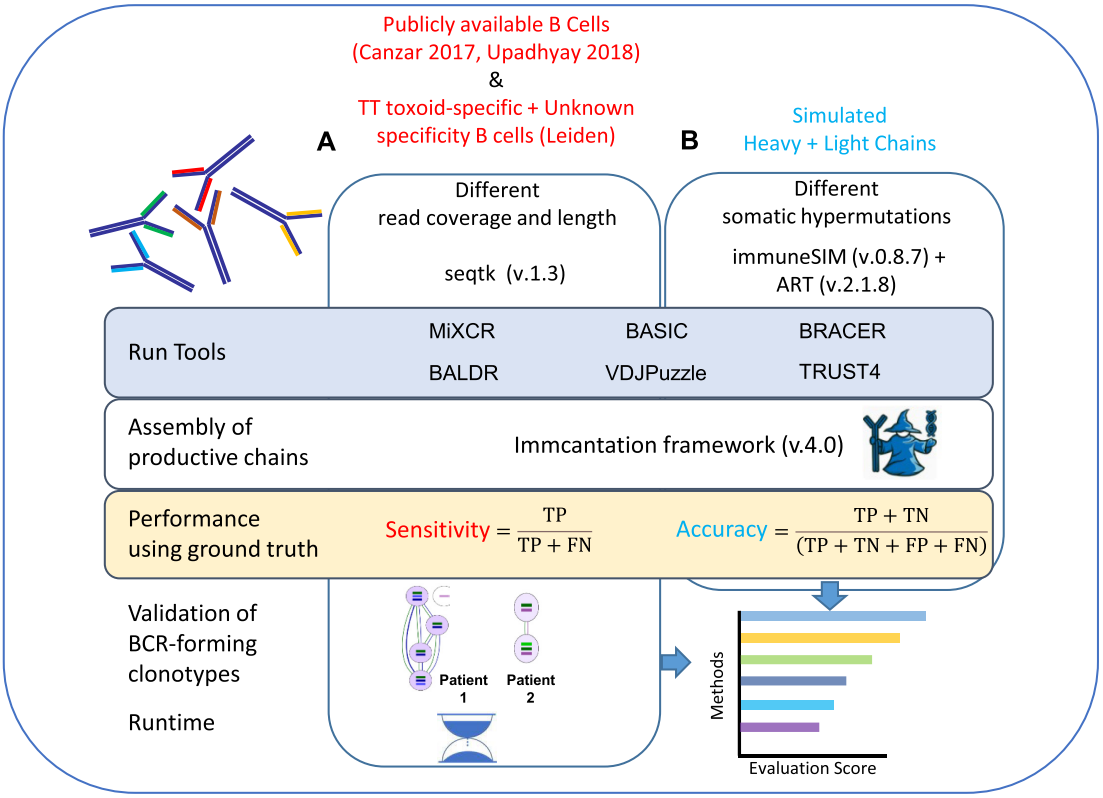

Figure 1. Benchmark framework. (**A**) Available datasets with ground truth (25,27) consisting of plasmablasts with unknown antigen specificity were obtained from the corresponding publications. In addition, an scRNA-seq dataset including TT-specific B cells and B cells with unknown antigen specificity was generated in this work (Leiden). All the datasets were downsampled to achieve different read coverages and lengths. (**B**) An additional dataset was simulated to investigate the effects of different levels of SHMs in the variable domains of BCRs on the performance of each method. Sensitivity and accuracy were used as metrics to evaluate each method depending on the type of used data. Antibodies were produced using a subset of clonotype-forming patient-specific BCRs and their specificity was experimentally validated. Finally, the execution time was investigated, and a final score was calculated to give a final recommendation on the method choice.

for this dataset. Afterward, we ran the Immcantation framework v4.0 (34) to assess the number of *assembled* and *productive* HC and LC (K or L) in each single cell and to determine germline genes in each sequence. As a result, we assembled 113 HC, 62 LcK and 57 LcL, including 102 cells having complete HC + LC pairs. After the assembly, 68 HC, 48 LcK and 46 LcL, including 45 paired HC + LC (complete BCRs), were labeled as *productive*. Hence, we used these 45 single cells with full-length paired and productive HC and LC to compute the sensitivity of each method for this dataset (Supplementary Figure S1B).

Upadhyay dataset (SMART-seq2). We retrieved the human AW1 plasmablast dataset included in the BALDR publication (27) from SRA (https://www.ncbi.nlm.nih.gov/sra/?term=SRP126429). This dataset, consisting of 51 single cells, was first investigated for the read quality using fastQC v0.11.9 (Babraham Bioinformatics; FastQC: a quality control tool for high-throughput sequence data). During this quality control step, we observed that libraries contained duplicated reads, overrepresented sequences and *k*-mers despite the quality of the reads being adequate at 3′ and 5′ ends (Supplementary Data S1 and Supplementary Figure S3). After downloading fastq files and fasta sequences of the ground truth that were generated using the Sanger method, we ran the Immcantation framework v4.0 (34) to assess the number of *assembled* and *productive* HC and LC in each single cell and to determine germline genes in each sequence. As a result, we reconstructed 34 HC, 19 LcK and 22 LcL, including 23 paired HC + LC that were termed as *assembled*. Out of these, 34 HC, 19 LcK and 21 LcL were labeled *productive*, including 23 HC + LC pairs (BCRs). Consequently, we used these 23 single cells with full-length paired and productive HC and LC to compute the sensitivity of each method for this dataset (Supplementary Figure S1C).

Simulation of experimental datasets with different read lengths and coverages

We used seqtk [lh3/seqtk: toolkit for processing sequences in FASTA/Q formats (github.com)] version 1.3-r106 with the option '-s100' to perform random sampling without reintroduction of additional reads to generate libraries with different levels (from 50 000 up to 1.25 million reads) of coverage. Afterward, the sampled reads were trimmed using seqtk with the option 'trimfq' to obtain final libraries with read lengths ranging from 25 up to 50 nucleotides for the Canzar dataset (25), and from 25 up to 100 nucleotides for the Upadhyay dataset (27) and for the Leiden dataset.

In total, we simulated 70 distinct libraries covering different read lengths, coverages, types of B cells (memory and plasmablasts) and plate-based techniques (SMART-seq2 versus SPEC-seq).

Simulation of synthetic chains with different levels of SHMs

The process of generating SHMs has been described as stochastic in nature for many decades. However, it has recently been demonstrated that intrinsic biases are present *in vivo* mostly due to the activity of the activation-induced cytidine deaminase (35). Nevertheless, we wanted to test whether a mutation accruing at any position of the CDR in the variable domain could generalize an effect on the performance of the methods used in this benchmark. For this, we used immuneSIM (36) version 0.8.7 to simulate four libraries containing 100 HC, 100 LcL and 100 LcK that harbored 15, 30, 45 and 60 SHMs. We used the option 'shm.mode=data', which focuses on mutation events in the CDRs (based on IMGT) occurring at any position during the process of SHM. First, this tool computes the frequencies of the V–D–J germline genes, insertions and deletions in the V and J junctions using repertoires present in different studies (37–40). Second, it recreates the sequences of HC and LC trying to maintain the statistical patterns and ratios of the number of germline genes. In case an introduction of SHMs is required in a simulation, the tool considers only chains without stop codons. In case a stop codon occurs, it resamples genes and other features until the requested number of simulated sequences without stop codons is reached. We noticed, though, that even if the resulting simulated sequences of HC or LC had no stop codons, they could still have out-of-frame V–J junctions that eventually resulted in a nonproductive chain. The output fasta file of each of the simulated HC, LcL and LcK was used as a reference to create synthetic reads and to obtain paired-end Illumina libraries containing 500 000 75-bp-long reads, using ART (41) version 2.1.8 (parameters *-l* 75, *-f* 500 000). Finally, we annotated each of the simulated sequences using the Immcantation framework v4.0 (34) to identify the germline genes and the number of *assembled* and *productive* HC and LC (Supplementary Figure S2A–D and Supplementary Table S4), which we used as the ground truth to compute the accuracy of each tool.

Sensitivity as a performance metric for experimental data

In this study, ground truth was considered as a set of HC and LC sequences coming from different single cells and obtained by Sanger sequencing. In principle, every cell should have a pair of HC and LC in the ground truth. However, we realized that only ∼40% of single cells in each of three analyzed datasets had paired HC and LC that were also defined as *assembled* and *productive* using the criteria mentioned earlier (Supplementary Figure S1A–C). Therefore, the ground truth information was available only for a subset of single cells (and thus productive BCRs) that could be used as a ground truth for all the single cells in each dataset. A metric such as specificity could not be used to evaluate the performance of the tested algorithms for BCR reconstruction due to the nature of the data (every antigen-specific B cell must have an HC and an LC), which prevents calculation of the false positive values. In other words, if Sanger sequencing misses one HC or one LC, assigning a false positive to a productive HC and LC assembled by a tool would be a wrong assumption since the absence of the chain would reflect a technical problem with Sanger sequencing and not the true biology. Therefore, we used sensitivity as a metric to compare different tools. First, we identified genes of productive HC and LC (K or L) of the ground truth for each cell in each dataset as described earlier. Second, we used different tested tools to assemble and annotate HC and LC (K or L) in different experimental libraries that we generated as described earlier. Then, we compared genes of each productive chain of the ground truth to the corresponding productive ones obtained by a tested method. As IgBLAST and tested tools can sometimes output more than one of each V–D–J germline gene, an HC was defined as a true positive (TP), if at least one of each of the V–D–J genes of a productive HC in the ground truth matched one of each of the corresponding V–D–J genes obtained by the computational method. In case an HC was not reported as assembled and productive by the tool, or when at least one of the V–D–J genes in the ground truth did not match any of the corresponding V–D–J genes obtained by a computational method, it was considered a false negative (FN). The same approach was used for the LC, but only V–J genes were matched. Knowing that a B cell can have two LC and [a process named allelic inclusion (42)] in case two LC were assembled in the Sanger sequencing, we compared only the productive one to the one obtained by the given method. Sensitivity was then calculated using the following formula:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Accuracy as a performance metric for simulated synthetic BCRs

In the simulated synthetic dataset described earlier, we counted both productive and nonproductive HC and LC (Supplementary Figure S2A–D). In humans, mature antigen-specific B cells have one productive HC and LC after SHM, as cells that fail to display functional BCRs are negatively selected and undergo apoptosis. However, the introduction of SHMs by immuneSIM returned several HC and LC with out-of-frame V–J junctions, which we interpreted as a proxy of the SHMs in the variable regions, resulting in nonproductive chains that would be negatively selected in the real world. Having assumed this, we counted true positive (TP), false negative (FN), false positive (FP) and true negative (TN) chains for each tool to compute the accuracy. Here, TP and FN were calculated the same way as for the sensitivity. If assembled and nonproductive chains in the ground truth were marked as assembled and productive by the tools, such chains were called FP. In case highly mutated and nonproductive HC and LC in the synthetic ground truth were reported as nonproductive by the tool, they were TN. To evaluate this, we computed the accuracy:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Calculation of gene overlap among the methods

To sustain the sensitivity as a metric for performance evaluation of the tested methods, we overlapped the V–(D)–J genes, mapped during the BCR reconstruction by each tool, in a pairwise manner separately for HC and LC. To do this, we counted all the HC that were reconstructed in each dataset by at least one method, and we used this number as denominator of our calculation. This was necessary because some methods were incapable of reconstructing the entire number of HC in each dataset. Afterward, we overlapped lists of separate V–(D)–J genes reconstructed by the methods in a pairwise manner and divided this number by the denominator. The same procedure was also performed for the LC (Supplementary Table S6).

Assessment of the execution time of each method

We considered the execution time as time needed to reconstruct a BCR from fastq files. For each tool, samples were run as parallelized jobs, in which a sample was run in parallel to all the others at the same time. We first created a multi-node high-performance computing cluster employing SLURM (version 18.08.5-2) as a job scheduler with 10 nodes using the Amazon Elastic Compute Cloud (Amazon EC2) and the r5.2xlarge instance that uses the Intel Xeon® Platinum 8000 processor, 3.1 GHz with eight vCPUs and 64 GB of RAM. All tools were then run on this virtual cluster requiring at least two CPUs and 6 GB of RAM for each job using the following parameters: '-n=2' and '–mem-per-cpu=6G'.

Final score to evaluate the methods

For each experimental dataset, we aggregated the sensitivity scores of HC and LC separately. For the HC, we summed the sensitivity values across different coverages and read lengths for all the libraries and divided this value by the number of libraries tested (where a library is a dataset with a defined read length and coverage). Given that BASIC was the only method capable of reconstructing HC in libraries of 25 bp length and that BRACER could not reconstruct HC when the read length was 50 bp in most of the libraries, we divided the sum of the sensitivity values across the different libraries of each dataset by the number of libraries in which the reconstruction was successful. The same was performed for the LC. Afterward, the two mean values obtained for the HC and the LC were added up and divided by 2. This score represents the evaluation of each tool for a given experimental dataset.

For the simulated SHM dataset, we first summed up the accuracy values of the HC reconstruction across the four SHM libraries (where each library contained HC with a specific amount of SHMs) and divided this number by the total number of libraries. We then performed the same procedure for the LC. These two values were added up and divided by 2 to obtain a score for this dataset.

Finally, we computed a cumulative average score by averaging the scores obtained from the four datasets and used it as a final evaluation metric of each method.

Creation of the *scBCR* docker image and data usage

The *scBCR* docker image that was used to install and run the methods tested in this study was built using Debian 10.6 docker. It embeds various tools that share dependences of different versions that can potentially be incompatible. To avoid this, different Anaconda environments were used to separate different versions of the same software. The native image has BRACER commit 131c2b9, MiXCR 3.0.13 and TRUST4 1.0.4. The conda *omics* environment contains BALDR commit 461d9b0. The conda *vdjpuzzle* environment contains VDJPuzzle 3.0 and BASIC 1.5.0. Along with the files needed to run each tool, the image also contains human genome and annotation version GRCh38, human transcripts and Bowtie index version GRCh38. The IGMT annotation is also provided to run the tools (43). All the tools were run using standard parameters for all the datasets, if not stated differently. For VDJPuzzle, we had to modify the first line of each fastq read when running samples from the Upadhyay dataset. Each pair of reads was modified from e.g. '@SRR6471013.1 1' to '@ SRR6471013_1/1' for read 1, and from '@SRR6471013.1 2' to '@ SRR6471013_1/2' for read 2. In case of BALDR, we used the output from the 'Ig_mapped & Ig_mapped + Unmapped' model to evaluate the reconstructed HC and LC since it was reported as the best option by the authors (27). In case of BASIC and BRACER, we took the sequences of a given chain and used Change-O (34) within the Immcantation framework v4.0 (34) to obtain the germline genes, productivity and in-frame V–J junctions. This information was already present in the output folders of TRUST4, VDJPuzzle, BALDR and MiXCR.

Validation of reconstructed TT-specific BCRs

We selected three pairs of reconstructed HC and LC belonging to three different BCRs for experimental validation. These TT-specific (TT+) BCRs were part of clonotypes identified in our dataset containing (TT+) memory B cells. We obtained these clonotypes by running BRACER function '*bracer summarise*' (Supplementary Figure S4). We synthesized and cloned variable domains of the assembled HC and LC sequences to produce and test the resulting monoclonal antibodies (mAbs). First, sequences were codon-optimized via GeneArt Gene Synthesis (Life Technologies) and the HC/LC variable genes together with 5′-BamHI and 3′-XhoI restriction sites, the Kozak sequence and the respective leader sequence were ordered from GeneArt (Life Technologies). The constructs were then ligated into a pcDNA3.1(+) expression vector (Invitrogen) carrying the IGHG1/4 or the IGLC1/IGKC constant regions (UniProt), respectively, flanking a 3′-XhoI site. The recombinant mAbs were produced in Freestyle™ 293-F cells (Gibco) as previously stated (44). Supernatants were harvested 5–6 days post-transfection. IgG antibodies were purified using a 1-ml HiTrap® Protein G HP affinity column (GE Healthcare) followed by a direct buffer exchange using a 53-ml HiPrep™ 26/10 Desalting column (GE Healthcare) according to the manufacturer's instructions. An IgG enzyme-linked immunosorbent assay (ELISA) was used to determine IgG concentrations of the mAbs according to the

manufacturer's protocol (Bethyl Laboratories). Anti-TT reactivity was assessed by a TT ELISA. TT (NIBSC) was directly coated onto C96 Maxisorp NuncImmuno plates (Thermo Fisher Scientific). mAbs were tested in multiple concentrations. Bound anti-TT IgG was detected by polyclonal rabbit anti-human IgG horseradish peroxidase (Dako), ABTS and H_2O_2 (Sigma-Aldrich).

RESULTS

Effect of coverage and read length on BCR assembly, productivity and sensitivity

To evaluate the experimental parameters affecting the performance of the BCR reconstruction tools in different B-cell types, we generated datasets of distinct library lengths and read coverages. To comprehensively test this, we selected three datasets with different characteristics. The first dataset named 'Leiden' consisted of reads obtained from TT-specific memory B cells and memory B cells with unknown antigen specificity. It was generated as part of this study using the SMART-seq2 protocol and had read length of 100 bp. The Canzar (SPEC-seq) (25) and Upadhyay (SMART-seq2) (27) datasets were obtained by sequencing plasmablast cells, yielding read lengths of 50 and 100 bp, respectively. We downsampled all three datasets to create libraries with variable coverage levels (from 50 000 to 1 250 000). Similarly, read trimming was done to simulate libraries with different read lengths (from 25 to 100 bp). Finally, we assessed whether a set of productive HC and LC (LcK or LcL) could be assembled by each tool for each single cell in the total set of 70 different libraries (see the 'Materials and Methods' section) and used the ground truth information (see Supplementary Figure S1A–C) to compute the sensitivity (see the 'Materials and Methods' section).

Leiden

We observed different outcomes for the different methods in terms of the percentage of cells with assembled and productive HC and LC (Figure 2A) using libraries generated from this dataset of 72 cells. Specifically, the number of assembled and productive HC was dependent on the number of reads for all the tools and increased with the rising coverage of the libraries. Moreover, this effect was strikingly pronounced for BASIC, which was able to assemble only <40% of the HC in libraries with 100 000 reads. Furthermore, BRACER showed difficulties in reconstructing HC in a scenario of 50-bp libraries with a very pronounced effect for coverages below 500 000 reads (Supplementary Figure S5A). BALDR, BRACER and TRUST4 displayed a similar high performance, assembling HC in up to 90% of the single cells. Finally, the average sensitivity for the HC remained relatively stable across the different coverage levels with an average value of 85% for BRACER, followed by BALDR with 77% and TRUST4 with 66% (Figure 2A and Supplementary Table S1).

All tools demonstrated a consistently high number of assembled and productive LC (K or L) across the different coverage and read length levels, with BALDR, BRACER and TRUST4 assembling LC in >90% of single cells (Figure 2A). Likewise, BRACER, BALDR and TRUST4 reached LC assembly sensitivities of 99%, 94% and 93%, respectively, being the highest among the tested tools (Figure 2A and Supplementary Table S1). In conclusion, BRACER, BALDR and TRUST4 were the best performing tools, assembling HC and LC with the highest sensitivity in this dataset.

Canzar 2017

In contrast to the other two datasets, we could only test the tools using 25- and 50-bp read length scenarios due to the nature of the original libraries of 113 plasmablasts in the Canzar dataset. As in the Leiden dataset, we observed varying effects of the tested parameters on the different tools (Figure 2B). Surprisingly, BRACER failed to assemble HC in all the tested libraries, seemingly due to the shortness of reads. On the other hand, all the other tested tools successfully assembled >75% of the productive HC. Moreover, BASIC reached up to 95% sensitivity across different coverage levels for the HC, followed by BALDR and TRUST4 with a sensitivity of 63% and 54%, respectively (Figure 2B and Supplementary Table S1). Although MiXCR and VDJPuzzle assembled comparably high percentages of the HC, they were less sensitive than others.

Furthermore, all the tested tools successfully assembled LC in >80% of the single cells. Additionally, changing coverage did not affect the assembly and productivity rates, except for VDJPuzzle, where we observed a drop to 75% of the assembled LC in the 1 250 000 read libraries. Moreover, BASIC and BRACER showed the highest sensitivity of 88% that remained stable across the different coverage levels. Other methods were less sensitive, with BALDR, TRUST4, MiXCR and VDJPuzzle reaching the sensitivities of 82%, 79%, 78% and 59%, respectively (Figure 2B and Supplementary Table S1). Notably, BASIC was the only method capable of assembling productive HC and LC from 25-bp libraries, displaying 54% sensitivity in reconstructing HC and 88% sensitivity for LC, with the average of 71% for both chains (Supplementary Figure S6 and Supplementary Table S1). Altogether, BASIC, BALDR and TRUST4 were the methods capable of assembling the highest number of productive HC and LC with the highest sensitivity values in the Canzar dataset.

Upadhyay 2018

Comparably to the Canzar dataset, the Upadhyay dataset was originally generated by sequencing 51 single plasmablasts in a paired-end mode, using 100-bp libraries instead of 50 bp. This could potentially explain some of the differences we observed (Figure 2C). The number of assembled and productive HC across the different coverage levels was ≥85% for all tools analyzed except for MiXCR, which assembled <60% productive chains. Consistently with the Canzar dataset, BRACER was not able to assemble HC when the length of reads dropped to 50 bp and below (Supplementary Figure S7A). Similarly, BASIC achieved the highest average sensitivity value (90%) across the different coverage levels for the HC. It was followed by BALDR and TRUST4 with 64% and 52% of sensitivity, respectively. Finally, VDJPuzzle and BRACER showed an in-

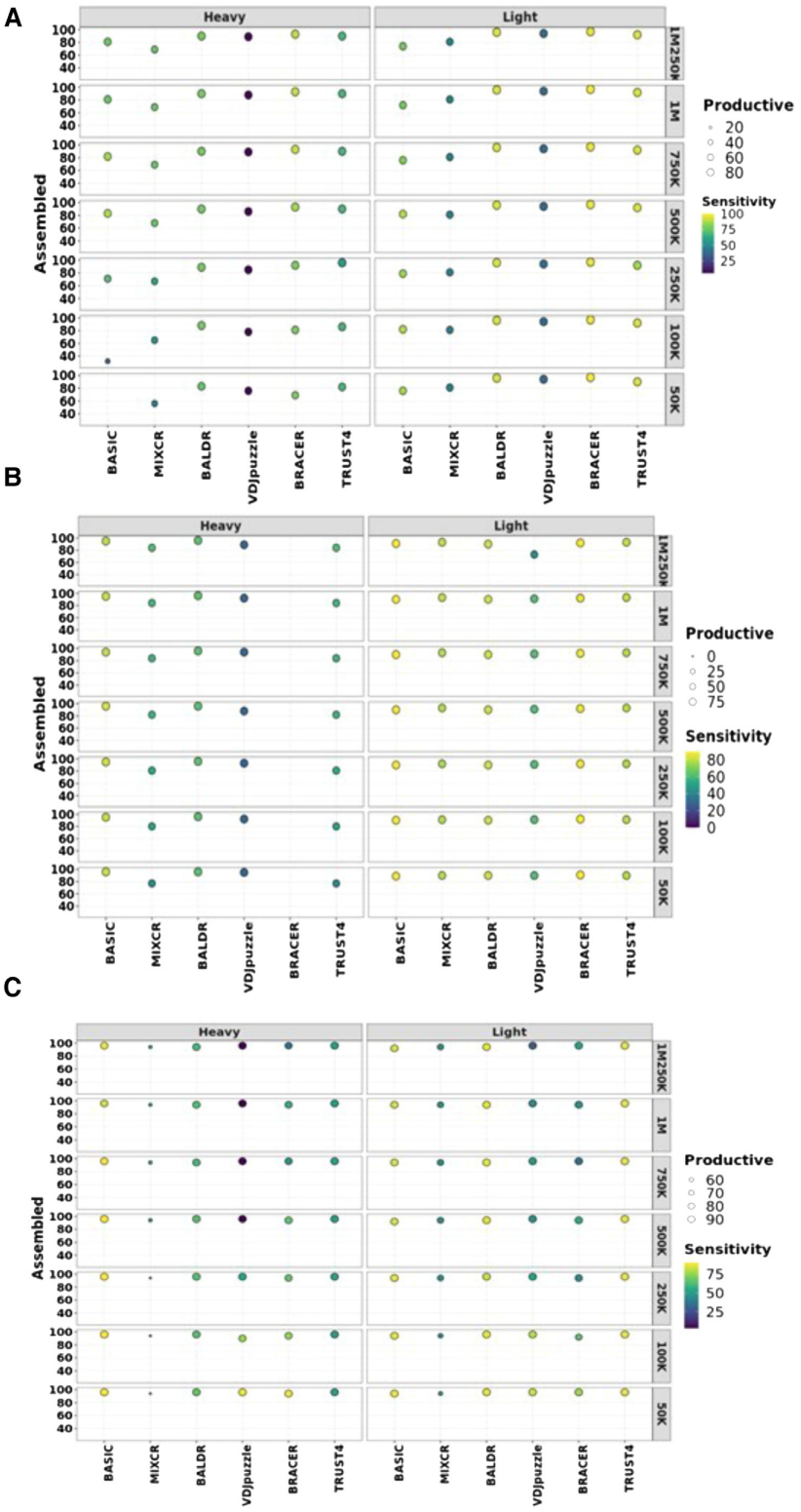

Figure 2. Performance of each method on different datasets. Effect of sequencing depth on the assembly, productivity and sensitivity of each tool in (**A**) the Leiden dataset created in this work, (**B**) the Canzar dataset and (**C**) the Upadhyay dataset, consisting of 72, 51 and 113 paired-end single-cell libraries. Original libraries were generated using either SPEC-seq or SMART-seq2 technology and had 100-, 50- and 100-bp-long reads, respectively. These datasets were downsampled to different coverages as displayed in the figures to test the effect of coverage on the assembly, productivity and sensitivity of BCR heavy and light chains (LcK or LcL). *Assembled* are heavy and light chains without stop codons. *Productive* are assembled heavy and light chains with in-frame V–J junctions. Left *y*-axis depicts % of assembled chains over the total number of single cells in each dataset. Right *y*-axis corresponds to coverage. The size of the circles is proportional to the % of the productive chains. Higher intensity of the yellow color of the circles corresponds to the higher sensitivity.

verse sensitivity–coverage relationship, with both reaching the sensitivity of 88% in libraries made up of 50 000 reads.

Only MiXCR failed to achieve >90% of assembled and productive LC across the different coverage levels. Among the remaining tools, TRUST4, BALDR and BASIC were the most sensitive, reaching average sensitivity levels equal to 87%, 87% and 86%, respectively (Supplementary Table S1). As for the HC, the sensitivity of VDJPuzzle and BRACER increased up to 83% and 77% with the simultaneous decrease of the number of reads to 50 000. An investigation of the read quality of all samples (Supplementary Data S1, Supplementary Figure S3 and Supplementary Table S5) revealed that the Upadhyay libraries harbored an overrepresentation of *k*-mers and duplicated reads, which might explain the inverse sensitivity–coverage relationship observed for BRACER and VDJPuzzle. Thus, BASIC was the method capable of assembling the highest number of productive HC and LC with the highest sensitivity value followed by BALDR in the Upadhyay dataset.

Method sensitivity is reflected by the V–(D)–J gene overlap

Next, we asked whether the performance of the tested methods was driven by the mere availability of a subset of HC and LC and the corresponding ground truth sequences in each dataset (see Supplementary Figure S1A–C) or by the capability of each method to detect the same genes and, therefore, alleles during the mapping procedure before reconstruction. To comprehensively test this, we did a pairwise comparison of separate V–(D)–J genes that were reconstructed in HC and LC by different methods. We selected libraries of 100 bp with 1.25 million reads from the Leiden and Upadhyay datasets and 50 bp with 1.25 million reads from the Canzar dataset for this analysis (Figure 3). Starting with the Leiden dataset (Figure 3A), we observed that BRACER and BALDR had the highest overlap of annotated HC V–(D)–J genes (0.88, 0.72 and 0.84). Moreover, although MiXCR and BRACER had a high overlap of the D genes, a poor overlap of the V and J genes was reflected in a general lower sensitivity of this tool. Furthermore, BALDR and BRACER showed the highest overlap of the V and J genes of LC (0.9 and 0.9). These results were in line with the overall better performance of these two tools on the Leiden dataset when the ground truth was used. Besides, we observed that BASIC and BALDR had a consistently good overlap of annotated HC V–(D)–J genes (0.71, 0.51 and 0.76) and the best overlap of annotated LC V–J genes (0.86 and 0.8) when compared to the match among other tools while using the Canzar dataset (Figure 3B). It is important to note that BRACER did not show any overlap of HC genes with any of the tools due to the inability of BRACER to reconstruct HC using 50-bp read length libraries. Finally, BASIC and BALDR had consistently the best overlap of V–(D)–J genes of the HC (0.86, 0.59 and 0.86) and LC (0.9 and 0.92) in the Upadhyay dataset (Figure 3C). Again, this agreed with the best performance of these two tools on this dataset. Overall, this analysis, which did not rely on the ground truth, reflected the performance of each method, which was evaluated when using the ground truth and the sensitivity as a metric.

Effect of SHMs on BCR assembly, production and accuracy

After assessing the effect of the read coverage and length on BCR *assembly* and *production*, we next investigated the consequences of different levels of SHMs in the CDRs of the variable domains of the BCRs on the performance of each tool. To address this, we first simulated datasets that consisted of 100 HC, 100 LcK and 100 LcL and contained distinct levels of SHMs (ranging from 15 to 60) in the CDRs of their variable domains. Second, we created synthetic Illumina libraries and tested the capability of each method to reconstruct HC and LC harboring various levels of SHMs. Finally, we annotated the simulated HC and LC and used them as the ground truth to compute the accuracy of each tool (see the 'Materials and Methods' section and Supplementary Figure S2A–D).

The first noticeable outcome of this experiment was a pronounced decrease in the percentage of assembled HC and LC (LcK and LcL) with the increasing number of introduced SHMs in the CDRs of the variable domains for MiXCR, VDJPuzzle and TRUST4 (Figure 4). Interestingly, we observed a stronger effect of SHMs on the assembly of LC than HC for MiXCR and BALDR compared to other tools. In contrast, BASIC, TRUST4, VDJPuzzle and BRACER did not show large differences in the assembly of HC and LC. Importantly, not all simulated HC and LC were productive in all the datasets (Supplementary Figure S2A–D). In fact, the number of nonproductive chains increased with the higher number of SHMs. Consequently, using the tested tools to assemble BCRs from such libraries would result in nonproductive HC and LC. Using the proposed metric of accuracy, we found that BRACER returned the highest average values of 95% and 97% across different SHM levels for HC and LC, respectively (Figure 4 and Supplementary Table S1). This was reflected by the capability of this method to correctly assemble a chain and assign its (non)productivity as in the ground truth. Other tools, such as BASIC, BALDR and TRUST4, were also stable across the different SHM levels in HC and LC but demonstrated on average lower accuracy of 66%, 77% and 83% for the HC and 92%, 87% and 79% for the LC, respectively (Figure 4 and Supplementary Table S1). In conclusion, BRACER was the most accurate tool across different SHM levels in correctly assembling HC and LC for this dataset, followed by BALDR.

Validation of the specificity of assembled BCRs

To validate the specificity of the BCR sequences assembled by the different algorithms, we produced three mAbs using productive BCR sequences that were assigned to three different clonotypes by BRACER using the SMART-seq2 Leiden dataset obtained in this study. Two mAbs were based on BCR sequences of single TT-specific B cells collected from patient 1 (cells B10 and D8) and a third mAb was based on a sequence isolated from patient 2 (cell G1) (Figure 5A and Supplementary Figure S4). First, the BCR sequences of these three cells assembled by all algorithms were compared to the ground truth that was obtained using ARTISAN PCR followed by Sanger sequencing. The assembled sequences of 1-B10 LC, 1-D8 LC and 2-G1 HC were identical among all different algorithms and the ground truth.

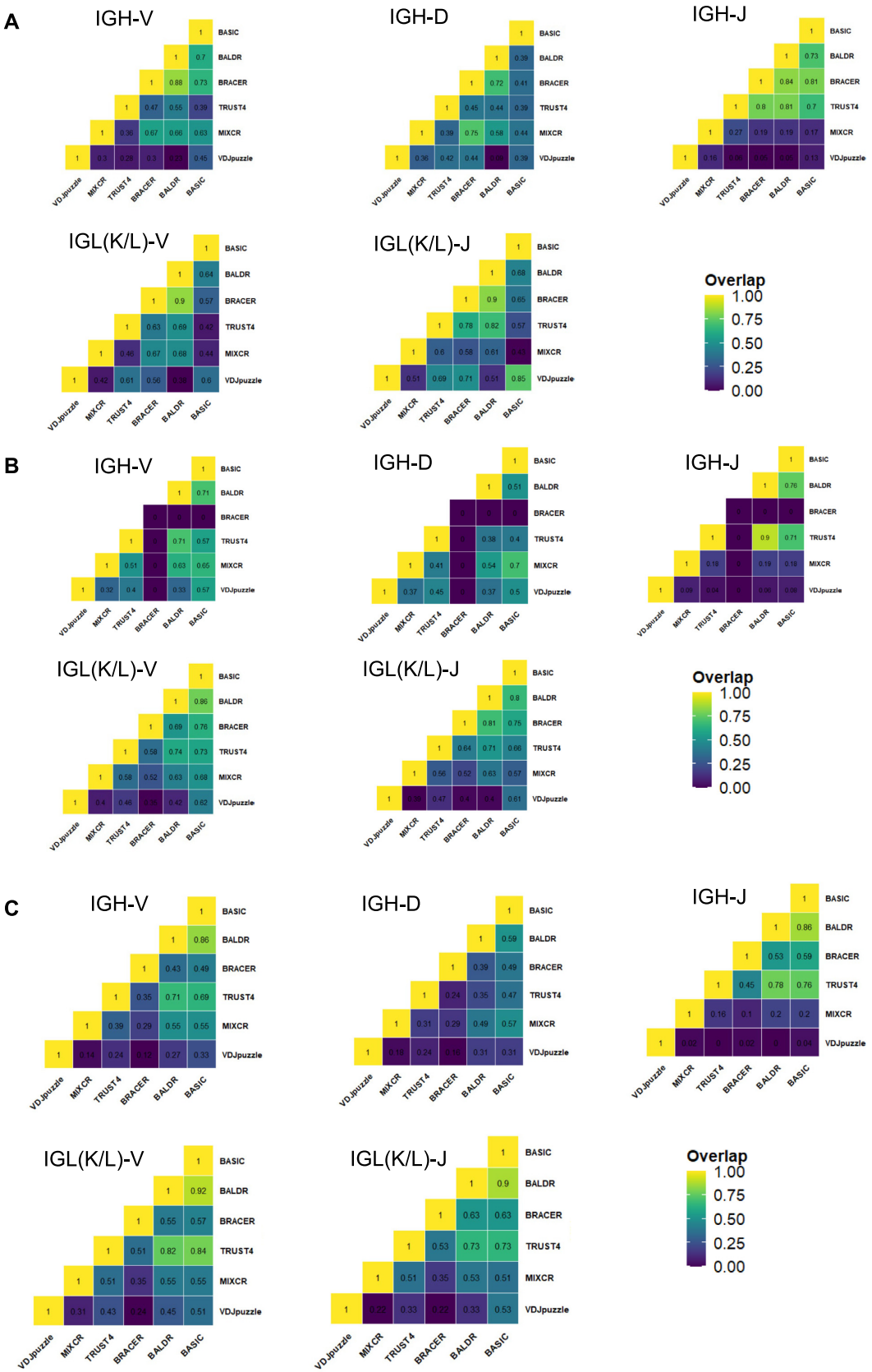

Figure 3. Pairwise overlap of V–(D)–J genes. Proportions of overlapping separate V–(D)–J genes among different computational methods for heavy chains (IGH) and light chains [IGL(K/L)] in the (**A**) Leiden, (**B**) Canzar and (**C**) Upadhyay datasets. Higher intensity of the yellow color in heatmaps corresponds to the higher overlap.

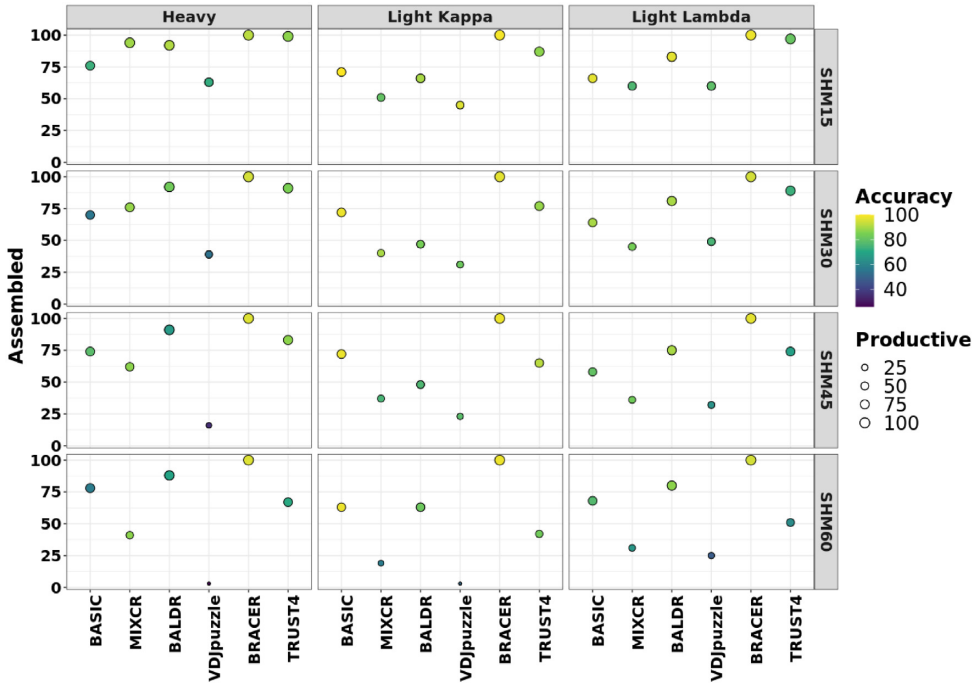

Figure 4. Effect of the number of SHMs in the variable domain on the performance of each method. One hundred HC, 100 LcK and 100 LcL were simulated with immuneSIM while introducing different amounts of SHMs in the CDRs of the variable domain (SHMs from 15 to 60). Using these sequences as a reference, Illumina libraries were created with ART (41) tool and all the methods were tested using those libraries. The obtained HC, LcL and LcK were compared to the initially simulated sequences to assess the assembly, productivity and accuracy rates of each method. *Assembled* are HC and LcK/LcL without stop codons. *Productive* are assembled HC or LcL/LcK with in-frame V–J junctions. Left *y*-axis corresponds to the % of assembled chains. Right *y*-axis shows the number of SHMs in each simulated library. The size of the circles is proportional to the % of the productive chains. Higher intensity of the yellow color of the circles corresponds to the higher accuracy.

However, MiXCR assembled two differences when compared to all other algorithms and the ground truth in the sequence of 1-B10 HC. In the 1-D8 HC and 2-G1 LC, several differences between the ground truth and the output of all algorithms were found (Figure 5B). We decided to produce the mAbs based on the sequences assembled by BRACER because two of the three selected sequences showed clonal relationship with BCR sequences determined by this tool in the dataset of TT-specific B cells (Figure 5C). After cloning and expression of selected mAbs, an IgG ELISA was first performed to stratify concentrations up to 1.5 μg/ml followed by a TT ELISA to test the TT binding of these antibodies. We confirmed that an unrelated mAb with a proven specificity for citrullinated antigens was negative in the TT ELISA, while a validated TT mAb was positive. All three tested mAbs showed a clear TT binding, with 1-D8 displaying the strongest signal (Figure 5D). In conclusion, except for sequence regions in which all the tools reported a different nucleotide composition when compared to the ground truth among themselves, only MiXCR reported individual amino acid differences in the CDR2 region. Nevertheless, the experimentally validated clonotype-forming BCRs showed high antigen specificity.

Evaluation of the execution time

If an experiment includes thousands or millions of cells to be assembled, the runtime during which tools complete BCR assembly might become a determining factor. Therefore, we evaluated the execution time of each tested method using a standard virtual cluster (see the 'Materials and Methods' section and Supplementary Table S2). As a result, we observed that runtimes increased with the increase of coverage and read length for all methods, except BRACER that demonstrated stable execution times across all coverages, starting from 50 bp (Figure 6 and Supplementary Figure S8A–D). Finally, TRUST4 was the tool that processed the highest number of reads per second, followed by MiXCR, BASIC, BALDR, BRACER and VDJPuzzle.

DISCUSSION

In this work, we extensively benchmarked six computational methods for BCR reconstruction using four different B-cell datasets, three experimental and one simulated. We focused our attention on the evaluation of methods capable of reconstructing full-length BCRs. The primary aim of this work was to provide guidance for the method of choice for different plate-based scRNA-seq datasets and scenarios (Figure 7C). In addition, we directed our attention to understanding the performance of each method on highly mutated BCRs that are common in autoimmune diseases (7,10), cancers (6,11) and in neutralizing anti-HIV antibodies (31–33).

Methods based on either 'semi *de novo*' (BASIC) or '*de novo*' (BALDR) assembly and using the IMGT annotation during the mapping procedure showed on average the highest performance when assembling both HC and LC

A

TT-IgG	IGHV	IGHD	IGHJ	IGH-CDR3 AA	IGH mutations	IGLV	IGLJ	IGL CDR3 AA	IGL mutations
1-B10	IGHV3-23	IGHD4-23	IGHJ4	CAKGPWDPPYHFDYW	26	IGLV2-23	IGLJ1	CCSYAGSSTYVF	21
1-D8	IGHV5-51	IGHD5-18	IGHJ4	CATHHVNTAMGNFASW	10	IGKV1-39	IGKJ1	CQQTYSTPPWTF	14
2-G1	IGHV3-23	IGHD3-16	IGHJ4	CARVPDGYDWGSYRYFDYW	24	IGKV3-11	IGKJ2	CQHRSNWPPYIF	12

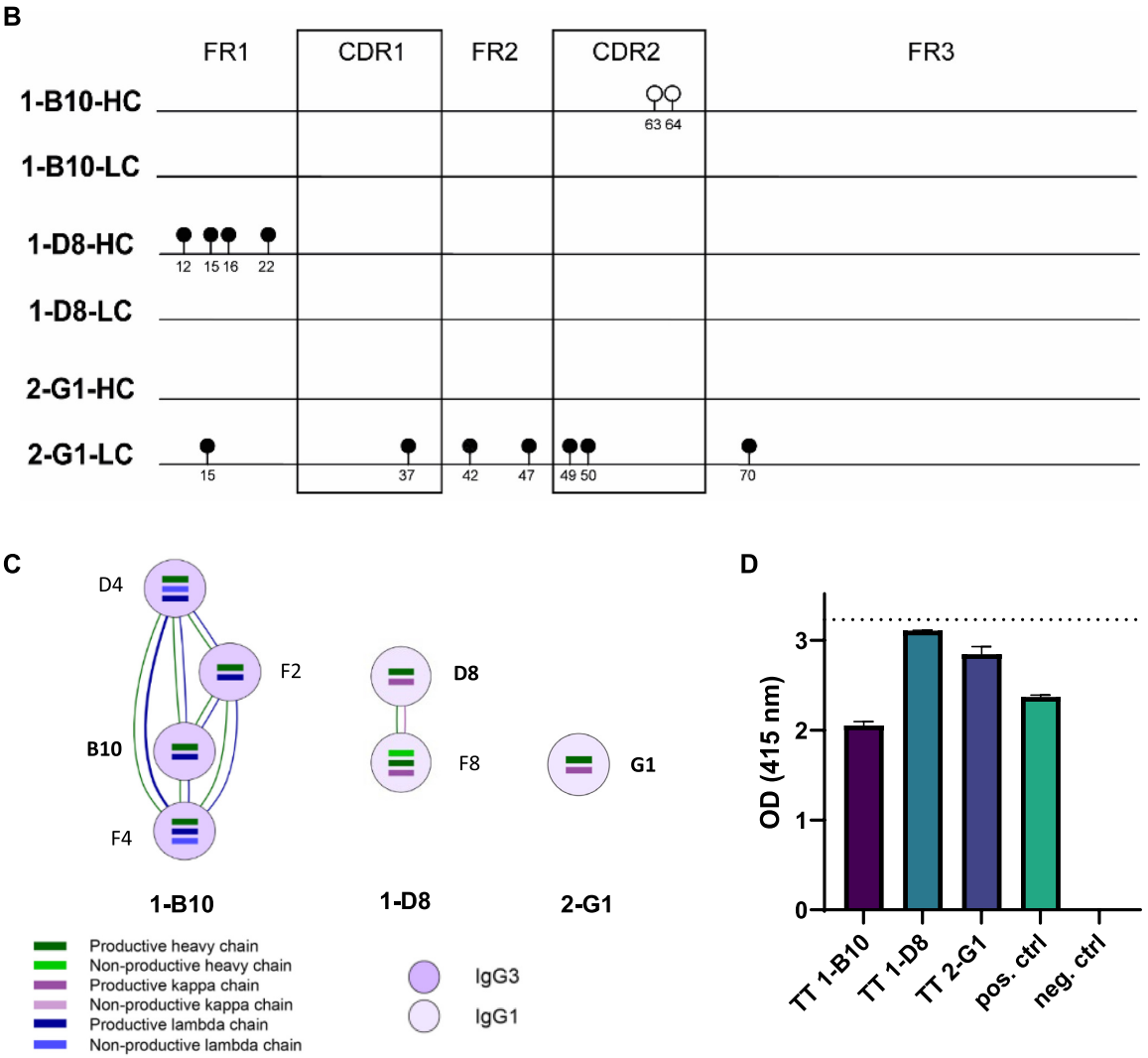

Figure 5. Validation of the TT specificity of the mAbs that were produced based on the assembled BCR sequences. (**A**) Characteristics of the TT IgG BCR variable region of the selected mAbs, based on the full-length BCR sequences obtained from the ground truth. V–D–J, CDR3 amino acid (AA) sequences and the numbers of nucleotide (nt) mutations compared to the germline sequence are depicted for the HC and LC of each TT IgG mAb. (**B**) Sequence alignment of the HC and LC of the sequences used for mAb production. The lollipops depict differences between the ground truth and the used tools. The positions of the lollipops are based on the IMGT amino acid numbering. Open circles depict differences of the MiXCR assembled sequence compared to the ground truth and the output of all other tools. Closed circles depict differences between the ground truth and all the evaluated tools. (**C**) The clonotype networks of the selected BCRs with other TT+ single cells obtained using BRACER (26). (**D**) Antibody validation using a TT ELISA for all antibodies. The ELISA was performed with a concentration of 1.5 μg/ml of the mAbs. Validated mAbs with (pos. ctrl) and without (neg. ctrl) tetanus specificity were used as controls. The optical density (OD) was measured at 415 nm 20 min after ABTS addition. The dashed line shows the upper detection limit of the TT ELISA.

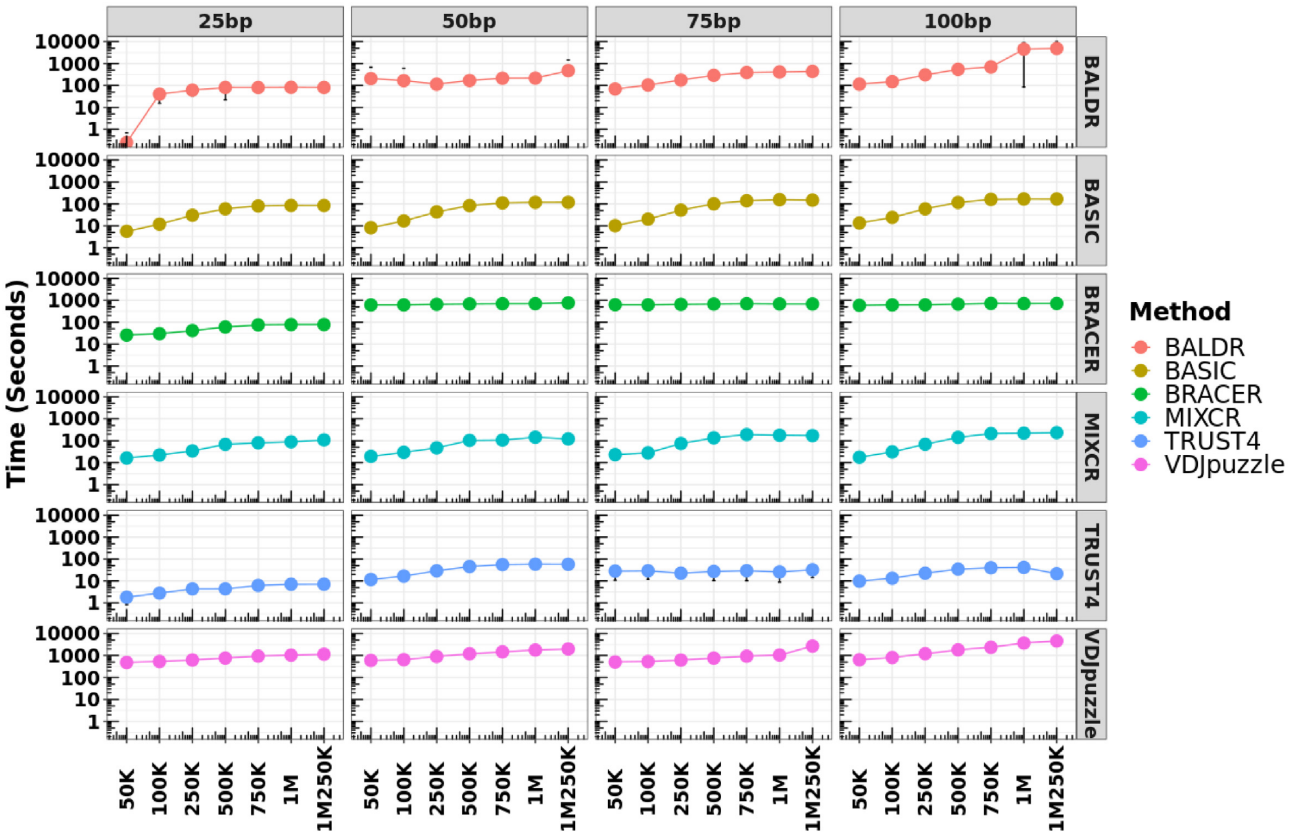

Figure 6. Execution time of each method using samples with different read lengths and coverages. Each method was tested using standard parameters and libraries with different read lengths and coverages. These libraries were simulated using the Leiden dataset. Each dot represents the median value (seconds) of the execution time calculated using all single cells from a particular library type. Top *x*-axis depicts the read length of the libraries. Bottom *x*-axis shows the coverage of the same libraries. Left *y*-axis corresponds to the runtime in seconds. Right *y*-axis shows the tested tool. The circle of each measurement corresponds to the median value of the time of execution of all the cells for a particular library type. The deviation from this value represents the standard deviation.

(K or L) of BCRs (Figure 7A) in the two plasmablast datasets (Canzar and Upadhyay). Moreover, despite the Upadhyay dataset being biased toward the overrepresented *k*-mers, both BALDR and BASIC algorithms maintained a good overall performance as previously reported (27), with BASIC being independent of the type of immunoglobulin chain. Despite having a similar high performance when compared to the other methods in the Leiden dataset, BASIC was, however, sensitive to the decreasing coverage level when reconstructing HC, and was outperformed by the *de novo* assembly-based methods BALDR, BRACER and TRUST4 (Figure 7A). The main difference of the Leiden dataset compared to the previously published datasets was the type of cells used for library construction, memory B cells and plasmablasts, respectively. This could have had a potential impact on the results we observed as plasmablasts produce high amounts of antibodies and thus have higher amounts of BCR mRNA. However, an excess of reads that belong to the V–(D)–J genes in a sequencing library does not strictly implicate better performance of all the tools. This could be observed in Figure 7A for the two plasmablast datasets (Canzar and Upadhyay), in the which the final score of all the tools was lower than the one obtained for the Leiden dataset of memory B cells. This could be explained by an open problem in genomics, which describes a possibility of multiple reads, belonging to a short variable genomic region (in this case, V–D–J genes), to map to multiple locations. Moreover, in accordance to the previous study (45), MiXCR showed a generally low sensitivity in respect to the other tools, which was reflected by the inconsistency of genes annotated in the assembled HC and LC, which is probably due to the high number of gene mishits. Thus, our results suggest the adoption of BASIC, BALDR and BRACER for the investigation of B-cell repertoires where BCRs are carrying a particular antigen specificity to accelerate antibody-based drug design.

When assessing the capability of each method to reconstruct BCRs bearing variable levels of SHMs in their variable domains, we noticed the limited performance of VDJPuzzle. We propose that this was due to the intrinsic property of the tool to completely discard reads in case they do not map to any of the V–D–J genes and constant regions during the first assembly step (28). To support this notion, methods based on more sophisticated algorithms to reconstruct the variable portion of the HC and LC by overlapping the unmapped reads to those mapping to V–J junctions together with genes present in the constant regions of HC and LC showed to have adopted a better strategy. Therefore, BRACER, BALDR, BASIC and TRUST4 should be chosen when assembling highly mutated sequences.

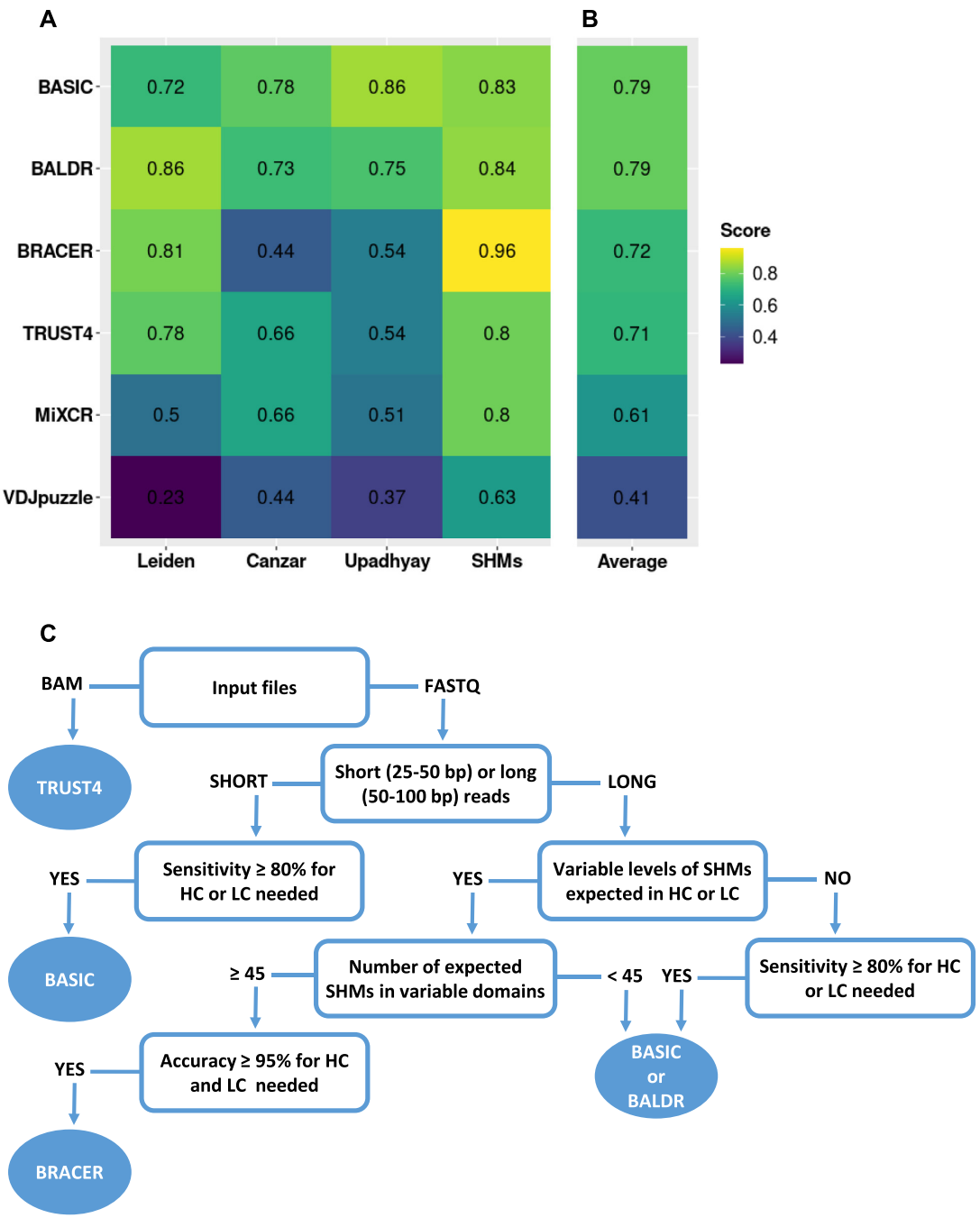

Figure 7. Average performance of benchmarked BCR methods on the different datasets. (**A**) For each method, a weighted average value for the sensitivity was calculated by averaging single sensitivity values that were obtained for HC and LC after running the tools on the different libraries of the three datasets: Leiden, Canzar and Upadhyay. A similar procedure was performed using the accuracy for the HC and LC obtained using the simulated SHM dataset. (**B**) A cumulative average score for each method across the different datasets was obtained. (**C**) Our recommendation tree for method selection, according to the research question and type of the dataset to be analyzed.

The experimental and simulated datasets used for this benchmarking study supplemented each other. On one hand, the experimental data reflected the real-world scenario of a limited number of available (antigen-specific) B cells and underlined the challenges in the assembly of their BCRs at different cell differentiation stages. On the other hand, larger sets of sequences with specific characteristics such as SHM load with absent sequencing challenges (e.g. duplicated reads, overrepresented *k*-mers, etc.) can be compared when using simulated datasets. However, in general, simulated datasets cannot completely reflect and thus replace the *in vivo* situation.

In terms of execution time, TRUST4, MiXCR and BASIC can process the highest number of reads per second.

They can run on a machine with as few as two CPUs Intel Xeon® Platinum 8000 with 3.1 GHz frequency each and 6 GB of RAM and process single-cell libraries of 1 million reads in ∼2 min, thus making them adoptable in every modern lab.

Importantly, some of the BCR assembly tools, namely BASIC (25), MiXCR (24), VDJPuzzle (28) and TRUST4 (29), can also be used to assemble TCRs from single-cell sequencing data. However, since BCR reconstruction is complicated by SHM, while this is not the case for TCRs, an independent benchmarking of the tools should be performed for TCR assembly.

With the goal to validate the specificity of the BCR sequences assembled by the different algorithms, three mAbs were produced. Remarkably, all tools showed consistent sequence (dis)similarities when compared to the ground truth, except for MiXCR, which assembled 1-B10 HC with two additional mutations. This is in line with the previous study (45) showing a high frequency of gene mishits during the alignment step of MiXCR and the resulting negative impact on the BCR sequence assembly. The disagreement of all the tools with the ground truth for the 1-B10 HC was in line with the low quality of the Sanger sequencing in the region where different nucleotides were called, suggesting that the computational methods can be superior to Sanger sequencing in such cases. However, the Sanger sequence of the 1-D8 LC was of high quality and may hint to either the amplification bias during the ARTISAN PCR or the issues during the BCR reconstruction step. Since all computational tools employ different methodologies for BCR assembly, the latter seems unlikely. Finally, the antigen affinity assay showed a clear TT binding by the cloned antibodies. This confirms that the adoption of these tools in the research and clinical setting would be beneficial for BCR and antibody reconstruction.

We would also like to emphasize that this benchmark study was explicitly done with datasets generated using plate-based scRNA-seq techniques, such as SMART-seq2 or a variant of it named SPEC-seq. TRUST4 was the only method, among the publicly available ones, capable of processing 10x BCR data (29). However, the unfeasibility to obtain the ground truth for 10x data and the scarcity of tools to process them made us exclude this type of data from our benchmark. Nevertheless, such comparisons should be addressed in the future studies. Moreover, although BRAPeS (46) was initially included in our evaluation, the extremely long runtimes to reconstruct just the CDR3 region resulted in a need to subset the reads (∼5000 as suggested by the developer). Since the scope of this work included understanding the effect of read coverage and length on tool performance, we excluded BRAPeS from further evaluations.

In conclusion, we provide clear guidance to select the best method (Figure 7C) according to the data type and research question the user has at the start of the BCR reconstruction experiment to facilitate the research. In our opinion, this work will help to improve the existing and develop new methods for BCR construction, especially adapting them to other sequencing technologies that are gaining increasing popularity, such as those using 10x, Oxford Nanopore and PacBio sequencing platforms.

DATA AVAILABILITY

Figures 2, 4 and 7 can be reproduced using data in Supplementary Table S1. Figure 3 can be reproduced using the data in Supplementary Table S6. Figure 5 can be reproduced using Supplementary Table S2. The Sanger sequencing results of the 'Leiden' dataset together with the annotation can be found in Supplementary Table S3. The fasta sequences of the simulated datasets with different levels of SHMs in the CDRs of the variable domains of HC and LC can be found in Supplementary Table S4.

We deposited the *scBCR* docker image together with the instructions to run all the tested methods in the following GitLab repository: https://gitlab.com/tAndreani/scBCR. In addition, we also created notebooks to compute the sensitivity and accuracy as well as to recreate the plots presented in this manuscript within the same directory.

ACCESSION NUMBERS

The single-cell RNA-seq data provided in this work are available at the SRA, BioProject ID PRJNA783770.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the Sanofi Global Postdoctoral Fellowship Program for financing the postdoc position of Tommaso Andreani. We thank the Flow Cytometry Core Facility (FCF) of Leiden University Medical Center (LUMC), The Netherlands, for the assistance with the cell sorting. We also thank Nima Nouri from Sanofi Precision Immunology group, and Franck Rapaport and Taylor Sorenson from Sanofi's AI & Deep Analytics, Omics Data Science and In Silico Drug Design groups, respectively, for reading the manuscript and providing feedback.

FUNDING

EU/EFPIA Innovative Medicines Initiative [777357].

Conflict of interest statement. T.A., C.S., C.R., R.O.-S., F.A., V.Y. and D.Š. are employees of Sanofi. D.Š., C.S., C.R. and V.Y. hold stocks of Sanofi.

REFERENCES

1. Stubbington,M.J.T., Rozenblatt-Rosen,O., Regev,A. and Teichmann,S.A. (2017) Single-cell transcriptomics to explore the immune system in health and disease. *Science*, **358**, 58–63.
2. Xiao,Z., Dai,Z. and Locasale,J.W. (2019) Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat. Commun.*, **10**, 3763.
3. Hong,X., Meng,S., Tang,D., Wang,T., Ding,L., Yu,H., Li,H., Liu,D., Dai,Y. and Yang,M. (2020) Single-cell RNA sequencing reveals the expansion of cytotoxic $CD4^+$ T lymphocytes and a landscape of immune cells in primary Sjogren's syndrome. *Front. Immunol.*, **11**, 594658.
4. Qian,J., Olbrecht,S., Boeckx,B., Vos,H., Laoui,D., Etlioglu,E., Wauters,E., Pomella,V., Verbandt,S., Busschaert,P. *et al.* (2020) A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.*, **30**, 745–762.

5. Burger,J.A. and Wiestner,A. (2018) Targeting B cell receptor signalling in cancer: preclinical and clinical advances. *Nat. Rev. Cancer*, **18**, 148–167.
6. Hu,Q., Hong,Y., Qi,P., Lu,G., Mai,X., Xu,S., He,X., Guo,Y., Gao,L., Jing,Z. *et al.* (2021) Atlas of breast cancer infiltrated B-lymphocytes revealed by paired single-cell RNA-sequencing and antigen receptor profiling. *Nat. Commun.*, **12**, 2186.
7. Ramesh,A., Schubert,R.D., Greenfield,A.L., Dandekar,R., Loudermilk,R., Sabatino,J.J. Jr, Koelzer,M.T., Tran,E.B., Koshal,K., Kim,K *et al.* (2020) A pathogenic and clonally expanded B cell transcriptome in active multiple sclerosis. *Proc. Natl Acad. Sci. U.S.A.*, **117**, 22932–22943.
8. Zhang,J., Hu,X., Wang,J., Sahu,A.D., Cohen,D., Song,L., Ouyang,Z., Fan,J., Wang,B., Fu,J. *et al.* (2019) Immune receptor repertoires in pediatric and adult acute myeloid leukemia. *Genome Med.*, **11**, 73.
9. Vergroesen,R.D., Slot,L.M., Hafkenscheid,L., Koning,M.T., van der Voort,E.I.H., Grooff,C.A., Zervakis,G., Veelken,H., Huizinga,T.W.J., Rispens,T. *et al.* (2018) B-cell receptor sequencing of anti-citrullinated protein antibody (ACPA) IgG-expressing B cells indicates a selective advantage for the introduction of N-glycosylation sites during somatic hypermutation. *Ann. Rheum. Dis.*, **77**, 956–958.
10. Scherer,H.U., Huizinga,T.W.J., Kronke,G., Schett,G. and Toes,R.E.M. (2018) The B cell response to citrullinated antigens in the development of rheumatoid arthritis. *Nat. Rev. Rheumatol.*, **14**, 157–169.
11. Xu-Monette,Z.Y., Li,J., Xia,Y., Crossley,B., Bremel,R.D., Miao,Y., Xiao,M., Snyder,T., Manyam,G.C., Tan,X. *et al.* (2019) Immunoglobulin somatic hypermutation has clinical impact in DLBCL and potential implications for immune checkpoint blockade and neoantigen-based immunotherapies. *J. Immunother. Cancer*, **7**, 272.
12. Lopez-Santibanez-Jacome,L., Avendano-Vazquez,S.E. and Flores-Jasso,C.F. (2019) The pipeline repertoire for Ig-seq analysis. *Front. Immunol.*, **10**, 899.
13. Picelli,S., Bjorklund,A.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–1098.
14. Picelli,S., Faridani,O.R., Bjorklund,A.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
15. Neu,K.E., Guthmiller,J.J., Huang,M., La,J., Vieira,M.C., Kim,K., Zheng,N.Y., Cortese,M., Tepora,M.E., Hamel,N.J. *et al.* (2019) Spec-seq unveils transcriptional subpopulations of antibody-secreting cells following influenza vaccination. *J. Clin. Invest.*, **129**, 93–105.
16. Goldstein,L.D., Chen,Y.J., Wu,J., Chaudhuri,S., Hsiao,Y.C., Schneider,K., Hoi,K.H., Lin,Z., Guerrero,S., Jaiswal,B.S. *et al.* (2019) Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol.*, **2**, 304.
17. Mandric,I., Rotman,J., Yang,H.T., Strauli,N., Montoya,D.J., Van Der Wey,W., Ronas,J.R., Statz,B., Yao,D., Petrova,V. *et al.* (2020) Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.*, **11**, 3126.
18. Mose,L.E., Selitsky,S.R., Bixby,L.M., Marron,D.L., Iglesia,M.D., Serody,J.S., Perou,C.M., Vincent,B.G. and Parker,J.S. (2016) Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics*, **32**, 3729–3734.
19. Li,B., Li,T., Pignon,J.C., Wang,B., Wang,J., Shukla,S.A., Dou,R., Chen,Q., Hodi,F.S., Choueiri,T.K. *et al.* (2016) Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.*, **48**, 725–732.
20. Li,B., Li,T., Wang,B., Dou,R., Zhang,J., Liu,J.S. and Liu,X.S. (2017) Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data. *Nat. Genet.*, **49**, 482–483.
21. Hu,X., Zhang,J., Wang,J., Fu,J., Li,T., Zheng,X., Wang,B., Gu,S., Jiang,P., Fan,J. *et al.* (2019) Landscape of B cell immunity and related immune evasion in human cancers. *Nat. Genet.*, **51**, 560–567.
22. Zhang,W., Du,Y., Su,Z., Wang,C., Zeng,X., Zhang,R., Hong,X., Nie,C., Wu,J., Cao,H. *et al.* (2015) IMonitor: a robust pipeline for TCR and BCR repertoire analysis. *Genetics*, **201**, 459–472.
23. Sela-Culang,I., Kunik,V. and Ofran,Y. (2013) The structural basis of antibody–antigen recognition. *Front. Immunol.*, **4**, 302.
24. Bolotin,D.A., Poslavsky,S., Mitrophanov,I., Shugay,M., Mamedov,I.Z., Putintseva,E.V. and Chudakov,D.M. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, **12**, 380–381.
25. Canzar,S., Neu,K.E., Tang,Q., Wilson,P.C. and Khan,A.A. (2017) BASIC: BCR assembly from single cells. *Bioinformatics*, **33**, 425–427.
26. Lindeman,I., Emerton,G., Mamanova,L., Snir,O., Polanski,K., Qiao,S.W., Sollid,L.M., Teichmann,S.A. and Stubbington,M.J.T. (2018) BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods*, **15**, 563–565.
27. Upadhyay,A.A., Kauffman,R.C., Wolabaugh,A.N., Cho,A., Patel,N.B., Reiss,S.M., Havenar-Daughton,C., Dawoud,R.A., Tharp,G.K., Sanz,I. *et al.* (2018) BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Med.*, **10**, 20.
28. Rizzetto,S., Koppstein,D.N.P., Samir,J., Singh,M., Reed,J.H., Cai,C.H., Lloyd,A.R., Eltahla,A.A., Goodnow,C.C. and Luciani,F. (2018) B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics*, **34**, 2846–2847.
29. Song,L., Cohen,D., Ouyang,Z., Cao,Y., Hu,X. and Liu,X.S. (2021) TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods*, **18**, 627–630.
30. Gagyi,E., Balogh,Z., Bodor,C., Timar,B., Reiniger,L., Deak,L., Csomor,J., Csernus,B., Szepesi,A. and Matolcsy,A. (2008) Somatic hypermutation of IGVH genes and aberrant somatic hypermutation in follicular lymphoma without BCL-2 gene rearrangement and expression. *Haematologica*, **93**, 1822–1828.
31. Lorenzi,J.C.C., Mendoza,P., Cohen,Y.Z., Nogueira,L., Lavine,C., Sapiente,J., Wiatr,M., Mugo,N.R., Mujugira,A., Delany,S. *et al.* (2020) Neutralizing activity of broadly neutralizing anti-HIV-1 antibodies against primary African isolates. *J. Virol.*, **95**, 01909–01920.
32. Bournazos,S., Klein,F., Pietzsch,J., Seaman,M.S., Nussenzweig,M.C. and Ravetch,J.V. (2014) Broadly neutralizing anti-HIV-1 antibodies require Fc effector functions for *in vivo* activity. *Cell*, **158**, 1243–1253.
33. Scheid,J.F., Mouquet,H., Feldhahn,N., Seaman,M.S., Velinzon,K., Pietzsch,J., Ott,R.G., Anthony,R.M., Zebroski,H., Hurley,A. *et al.* (2009) Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, **458**, 636–640.
34. Gupta,N.T., Vander Heiden,J.A., Uduman,M., Gadala-Maria,D., Yaari,G. and Kleinstein,S.H. (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.
35. Yaari,G., Vander Heiden,J.A., Uduman,M., Gadala-Maria,D., Gupta,N., Stern,J.N., O'Connor,K.C., Hafler,D.A., Laserson,U., Vigneault,F. *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.*, **4**, 358.
36. Weber,C.R., Akbar,R., Yermanos,A., Pavlovic,M., Snapkov,I., Sandve,G.K., Reddy,S.T. and Greiff,V. (2020) immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics*, **36**, 3594–3596.
37. Greiff,V., Menzel,U., Miho,E., Weber,C., Riedel,R., Cook,S., Valai,A., Lopes,T., Radbruch,A., Winkler,T.H. *et al.* (2017) Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.*, **19**, 1467–1478.
38. Madi,A., Poran,A., Shifrut,E., Reich-Zeliger,S., Greenstein,E., Zaretsky,I., Arnon,T., Laethem,F.V., Singer,A., Lu,J. *et al.* (2017) T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife*, **6**, e22057.
39. DeWitt,W.S., Lindau,P., Snyder,T.M., Sherwood,A.M., Vignali,M., Carlson,C.S., Greenberg,P.D., Duerkopp,N., Emerson,R.O. and Robins,H.S. (2016) A public database of memory and naïve B-cell receptor sequences. *PLoS One*, **11**, e0160853.
40. Emerson,R.O., DeWitt,W.S., Vignali,M., Gravley,J., Hu,J.K., Osborne,E.J., Desmarais,C., Klinger,M., Carlson,C.S., Hansen,J.A. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, **49**, 659–665.

41. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
42. Pelanda,R. (2014) Dual immunoglobulin light chain B cells: trojan horses of autoimmunity?*Curr. Opin. Immunol.*, **27**, 53–59.
43. Lefranc,M.P., Giudicelli,V., Duroux,P., Jabado-Michaloud,J., Folch,G., Aouinti,S., Carillon,E., Duvergey,H., Houles,A., Paysan-Lafosse,T. *et al.* (2015) IMGT®, the international immunogenetics information system® 25 years on. *Nucleic Acids Res.*, **43**, D413–D422.
44. Kissel,T., van Wesemael,T.J., Lundquist,A., Kokkonen,H., Kawakami,A., Tamai,M., van Schaardenburg,D., Wuhrer,M., Huizinga,T.W., Scherer,H.U. *et al.* (2021) Genetic predisposition (HLA-SE) is associated with ACPA-IgG variable domain glycosylation in the predisease phase of RA. *Ann. Rheum. Dis.*, **81**, 141–143.
45. Smakaj,E., Babrak,L., Ohlin,M., Shugay,M., Briney,B., Tosoni,D., Galli,C., Grobelsek,V., D'Angelo,I., Olson,B. *et al.* (2020) Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. *Bioinformatics*, **36**, 1731–1739.
46. Afik,S., Raulet,G. and Yosef,N. (2019) Reconstructing B-cell receptor sequences from short-read single-cell RNA sequencing with BRAPeS. *Life Sci. Alliance*, **2**, 4.