Neural Computing and Applications (2021) 33:15669–15692
https://doi.org/10.1007/s00521-021-06188-z

ORIGINAL ARTICLE

New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies

Luigi Donato[1,2] · Concetta Scimone[1,2] · Carmela Rinaldi[1] · Rosalia D'Angelo[1] · Antonina Sidoti[1]

Received: 27 January 2020 / Accepted: 2 June 2021 / Published online: 16 June 2021

Abstract

During the last (15) years, improved omics sequencing technologies have expanded the scale and resolution of various biological applications, generating high-throughput datasets that require carefully chosen software tools to be processed. Therefore, following the sequencing development, bioinformatics researchers have been challenged to implement alignment algorithms for next-generation sequencing reads. However, nowadays selection of aligners based on genome characteristics is poorly studied, so our benchmarking study extended the "state of art" comparing 17 different aligners. The chosen tools were assessed on empirical human DNA- and RNA-Seq data, as well as on simulated datasets in human and mouse, evaluating a set of parameters previously not considered in such kind of benchmarks. As expected, we found that each tool was the best in specific conditions. For Ion Torrent single-end RNA-Seq samples, the most suitable aligners were CLC and BWA-MEM, which reached the best results in terms of efficiency, accuracy, duplication rate, saturation profile and running time. About Illumina paired-end osteomyelitis transcriptomics data, instead, the best performer algorithm, together with the already cited CLC, resulted Novoalign, which excelled in accuracy and saturation analyses. Segemehl and DNASTAR performed the best on both DNA-Seq data, with Segemehl particularly suitable for exome data. In conclusion, our study could guide users in the selection of a suitable aligner based on genome and transcriptome characteristics. However, several other aspects, emerged from our work, should be considered in the evolution of alignment research area, such as the involvement of artificial intelligence to support cloud computing and mapping to multiple genomes.

Keywords NGS · DNA-Seq · RNA-Seq · Mapping · Benchmark

1 Introduction

Starting from Sanger sequencing 40 years ago, more precise and rapid sequencing technologies expanded scale and resolution of various biological applications, including the detection of genome-wide single nucleotide polymorphisms (SNPs) and structural variants [1], quantitative analysis of transcriptome (RNA-Seq) [2], identification of protein binding sites (ChIP-Seq) [3], understanding methylation patterns in DNA [4], the assembly of new genomes or transcriptomes [5], determining species composition using metagenomic workflows. However, the huge amount of generated data explains almost nothing about the DNA without the appropriate analysis tools and algorithms. Therefore, bioinformatics researchers started to think about new ways to efficiently manage and analyze such enormous amount of data. The first crucial step in the analysis of next-generation sequencing (NGS) data, posterior to quality control and filtering steps, is alignment (mapping) of generated sequencing reads to the respective reference [6]. However, this step is biased by many errors due to the following reasons [7]: (1) a reference genome is generally long ($\sim$ billions) and presents complex regions such as repetitive elements (repetitive regions are usually masked because there is no consensus about how to deal with them,

✉ Rosalia D'Angelo
rdangelo@unime.it

[1] Division of Medical Biotechnologies and Preventive Medicine, Department of Biomedical and Dental Sciences and Morphofunctional Imaging, University of Messina, via C. Valeria 1, 98125 Messina, Italy

[2] Department of Biomolecular Strategies, Genetics, Cutting-edge Therapies, I.E.ME.S.T., 90139 Palermo, Italy

yet), (2) reads are short in length (typically, 50–150 bp), causing issues with efficiency and accuracy, aligning more likely in multiple locations rather than to unique positions in the reference genome, (3) the subject genome could inherently be different from the reference genome due to acquired alterations over time. In order to face the above challenges, many new alignment tools have been developed during the last years. Such tools exploit the specific advantages of each new sequencing technology, such as the short sequence length of Helicos (range of read length = 25–1000 bp), Illumina (range of read length = 36–300 bp) and SOLiD reads (range of read length = 35–75 bp), the high base quality toward the 5'-end of Illumina and 454 reads [8, 9], the di-base encoding of SOLiD reads, the low InDel error rate of Illumina reads (InDel rate = 0.000005%) [10], the low substitution error rate of Helicos reads (substitution rate = 0.2%) [11] and the Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) third generation sequencing technologies, superior in long-read assembly (PacBio Maximum Read Length = up to 40 Kb; ONT Maximum Read Length = 10 kb), with respect to accuracy and completeness [12]. The common alignment tools are based on (1) spaced-seed indexing or (2) Burrows–Wheeler transform [13, 14]. The first are slow, use more memory but correctly maps long gaps [15]. The second, instead, based on heuristic approaches, are fast, consume less memory and can be used to map short reads [16]. Spliced aligners such as GEM [17] and Tophat [18] are most frequently used for aligning transcripts. BWA [19], BLAT [20] and Bowtie2 [21] are frequently used for aligning DNA sequences. BWA and Bowtie2 are index-based aligners exploiting Burrows–Wheeler indexing algorithm. NovoAlign [22] uses dynamic programming, taking advantage of Needleman–Wunsch algorithm with affine gap penalties to score the alignment. Many other algorithms have been developed [23]. Selecting mapping tool, based on the characteristics of the organism, shall be a fundamental focus of bioinformatics, because the choice may affect downstream analysis. Several benchmarking analyses guided users in choosing aligners, but although several studies have been published for evaluating sequence mapping tools, the problem is still open and further perspectives were not faced [24–27]. Limitations of such works regard the focus on tool group classifications rather than evaluations of their own performance on selected settings [28], the use of small and unrealistic data sets (e.g., 500,000 reads) jointly with small reference genomes (e.g., 500 Mbps) [29], the use of simulated data only [30], as well as the mis-usage of the aligner options and algorithmic features. Therefore, selection of aligners based on genome characteristics is poorly studied, and a quantitative evaluation to systematically compare mapping tools in multiple aspects is still needed. We extend the "state of art" evaluations of mapping algorithms, comparing 17 different aligners (BBMap (sourceforge.net/projects/bbmap/) [31], Bowtie2, BWA, BWA-MEM [19, 32], Qiagen CLC Genomics Workbench (https://www.qiagenbioinformatics.com) [33–36], DNASTAR Lasergene Suite [37, 38], GEM [39], Hisat2 [40], Magic-BLAST [41], Minimap2 [42], Novoalign [43], YARA [44], RUM [45], Segemehl [46], STAR [47], Subread [48], TopHat2 [18, 49]) applied on empirical and simulated DNA- and RNA-Seq human and mouse datasets. Our study provides guidelines for the selection of a suitable aligner based on genome and transcriptome characteristics.

2 Materials and methods

2.1 Simulated data

In order to realize a complete evaluation of aligning tools, minimizing the possible bias coming from real sequenced data, we firstly produced simulated NGS reads, mapped them to the human and murine reference genomes and assessed read alignment accuracy using a tool evaluating if each individual read has been aligned correctly. In order to take in account the genomic features for following mapping analyses, we decided to simulate data from two organisms with two different read lengths (50 bp and 150 bp) using five different reads simulators. Each used read simulator introduced mutations within the *homo sapiens* GRCh38.p13 (RefSeq assembly accession: GCF_000001405.39, downloaded on January 4, 2021) or within the *mus musculus* Genome Reference Consortium Mouse Build 39 (GRCm39) (RefSeq assembly accession: GCF_000001635.27, downloaded on January 4, 2021) FASTA reference genomes and produced reads as genomic substrings with randomly added sequencing errors. Such simulations of artifacts and errors observed in real data were generated by different statistical models, mainly based on coverage, read sequencing errors and genomic mutations distributions, as well as CG-content. Finally, the origin of each read (generated for a given simulator) is encoded in a read group specific for each algorithm, and the reads were saved into FASTQ files. In order to enforce the usefulness of simulations, we used 5 different simulators, each one with own features: (1) ART [50]; (2) DWGSIM (http://github.com/nh13/dwgsim); (3) WGSIM (http://github.com/lh3/wgsim); (4) MASON [51]; (5) CURESIM [52]. The average Phred score over all simulated reads was of 29. Details of each algorithm, such as type of simulated data (Illumina or Ion Torrent), are available in Table 1.

Table 1 List of simulation tools with main features

Sim. Algorithm	Version	Technology	Run types	Read Length	Error Rate (%)	Processing	RAM (Gb)	Time (min)	Input
MASON	Mason, v.0.1.2	ILLUMINA	PE	50 bp and 150 bp (both Mouse and Human)	0.004 – 0.0285	P	30	43 (Mouse, 50 bp) 87 (Mouse, 150 bp) 47 (Human, 50 bp) 106 (Human, 150 bp)	FASTA (Ref. Gen.), VCF
DWGSIM	DwgSim, v.0.1.12	ION TORRENT	PE	50 bp and 150 bp (both Mouse and Human)	0.0037 – 0.034	P	30	17 (Mouse, 50 bp) 24 (Mouse, 150 bp) 32 (Human, 50 bp) 41 (Human, 150 bp)	FASTA (Ref. Gen.)
WGSIM	WgSim, v.1.0.2	ILLUMINA	PE	50 bp and 150 bp (both Mouse and Human)	0.0092 – 0.0336	P	30	19 (Mouse, 50 bp) 26 (Mouse, 150 bp) 34 (Human, 50 bp) 43 (Human, 150 bp)	FASTA (Ref. Gen
ART	ART v.2016–06-05	ILLUMINA	PE	50 bp and 150 bp (both Mouse and Human)	0.0095 – 0.06	P	30	21 (Mouse, 50 bp) 26 (Mouse, 150 bp) 29 (Human, 50 bp) 39 (Human, 150 bp)	FASTA (Ref. Gen.)
CURESIM	Customized Read Simulator, v.1.3	ION TORRENT	SE	50 bp and 150 bp (both Mouse and Human)	0.005 – 0.215	P	30	13 (Mouse, 50 bp) 17 (Mouse, 150 bp) 21 (Human, 50 bp) 29 (Human, 150 bp)	FASTA (Ref. Gen.)

Table 1 (continued)

Sim. Algorithm	Version	Technology	Output	Prog. Lang	Description	Docs	Bib
MASON	Mason, v.0.1.2	ILLUMINA	FASTQ, SAM	C + +	Starting from a genome, Mason can simulate variants and optionally also methylation levels. Optionally, it can simulate bisulfite treatment	Y	Holtgrewe (2010)
DWGSIM	DwgSim, v.0.1.12	ION TORRENT	FASTQ, VCF	C, Perl, Python	Based on WgSim originally released in the SAMtools software package (Danacek et al., 2021), it was modified to handle ABI SOLiD and Ion Torrent data	Y	Heng Li et al. (2011)
WGSIM	WgSim, v.1.0.2	ILLUMINA	FASTQ, VCF	C	WgSim is able to simulate diploid genomes with SNPs and insertion/deletion (InDel) polymorphisms, and simulate reads with uniform substitution sequencing errors	Y	Heng Li et al. (2011)
ART	ART, v.2016–06-05	ILLUMINA	FASTQ, ALN, MAP, SAM, BED	C ++ , Perl	ART consists of a set of tools supporting genome and amplicon sequencing simulation of SE, PE and MP reads of Illumina's Solexa, Roche's 454 and Applied Biosystems' SOLiD. ART	Y	50
CURESIM	Customized Read Simulator, v.1.3	ION TORRENT	FASTQ	Java	CuReSim supports read simulation for major letter-base sequencing platforms. Wrappers to integrate CuReSim in Galaxy are also available	Y	52

Each reads simulator showed different features, considering both mouse (*Mus musculus*, GRCm39) and human (*Homo sapiens*, GRCh38.p.13) reference genomes.

PE paired end, *SE* single end, *P* parallel processing (accepts multithreading). The error rate is the mean value of error rates of all samples for each simulator. Ref. Gen. = reference genome. Prog. Lang. = programming language. Docs = documentation. Bib. = bibliography.

3 Real data samples

In order to cover popular sequencing platforms in biomedical science, a heterogeneous group of four samples (datasets unpublished) was chosen to perform the benchmark analysis. One RNA-Seq and one whole-genome sequencing (WGS) outputs, resulting from separate single-end experiments on Ion Torrent platform, came from the whole transcriptome analysis of retinal pigmented epithelial (RPE) cells and from the whole-genome sequencing of a patient affected by an orphan form of retinitis pigmentosa, respectively. Then, another one RNA-Seq and one whole-exome sequencing (WES) outputs, resulting from separate paired-end experiments on Illumina platform, came from the whole transcriptome analysis of exudate of a patient affected by osteomyelitis and from the whole-exome sequencing of a patient affected by an atypical form of retinitis pigmentosa, respectively. Produced raw data were quality checked by FastQC (v.0.11.7) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and 5′ end-trimmed according to Phred score threshold of 30. Residual adaptor sequences have been removed (min read length for a read to be kept = 50 bp). Detailed features of analyzed samples, characterized by read length ranging from 100 to 200 bp after trimming, and by total read number ranging from about 22 million to 88 million, are available in Table 2.

4 Aligner selection

The 17 benchmarked aligners were selected to represent different algorithms, many of them which are indexing-based (e.g., Bowtie2, BWA), hashing based (e.g., NovoAlign) and exploiting suffix array approaches (e.g., STAR). Main elements we considered to choose aligners were the number of citations (the highest cited, well-exploited aligners. and the lowest, already fully explored ones) and the continuously updating of algorithm code. All alignments were realized using the GRCh38.p13 and the GRCm39 reference genomes, for human and mouse, respectively. The list of all chosen mappers, along with their salient features, is highlighted in Table 3. Aligner parameters used during evaluations are listed in Table 4. About real data, YARA alignment was only completed on single-end RPE cells transcriptome, probably due to computational hardware limitations. Additionally, RUM mapping on WGS and simulated data outputted errors, probably due to conflicts between algorithmic specific features and qualities of real sample data.

Table 2 Description of datasets analyzed in our benchmarking study

Datasets	NGS Platform	Sequencing Strategy	Read Length (bp)	Expected Coverage	Number of Raw Reads
RPE Cell Transcriptome	Ion Torrent Proton	SE	200	20X	22,266,648
RP WGS	Ion Torrent Proton	SE	150	15X	44,755,937
Osteomyelitic Exudate Transcriptome	Illumina HiSeq 2500	PE	100 × 2	40X	36,851,618
RP WES	Illumina HiSeq 2500	PE	100 × 2	75X	88,068,730

Table presents description of the four datasets used during benchmark analyses. M reads = million reads. RPE = retinal pigment epithelial. RP = retinitis pigmentosa. PE = paired end. SE = single end.

5 Benchmarking workflow design

The mapping is a string-matching problem. It needs to integrate the known properties of the DNA sequences with the sequencing technologies, increasing complexity to the mapping procedure. To speed up the alignment process, most tools (and especially all the benchmarked tools we assess in this paper) pre-build an index on the reference genome to query faster the origin of each read. Several algorithms build index on the reads, but the most recent ones create index on the reference genome [53]. The latter method depends on the fact that the same index once built on a reference genome can be used repeatedly for aligning different read sets. After realized alignments, we used the Picard command line tool suite (https://broadinstitute.github.io/picard/) for processing and analyzing obtained data, especially focusing on internal control metrics, alignment summary metrics, GC bias metrics, quality by cycle, quality distribution, duplication metrics, insert size and deletion metrics. Additionally, clipped reads were also analyzed by the specific *ExtractSVReads* module of Genome Rearrangement IDentification Software Suite (GRIDSS) [54]. As regards the simulated datasets, where we have a gold standard for the origin of each read in the reference genome, we used the evaluation tool Alfred [55] to assess if any read has been affected to the correct location in the reference genome (and, possibly, with the appropriate edit operations). Final statistics strongly depended on the definition of a correctly mapped read, mapping qualities and multimapped reads.

6 Estimation of aligners' sensitivity and efficiency

Mapping task can represent a bottleneck in the NGS analysis pipeline due to the ever-increasing volume of the sequencing data, giving above approaches the need to adopt a trade-off between accuracy and speed, e.g., based on gaps allowed. Consequently, it is important to assess the performance of the aligners on both accuracy and computational efficiency of read alignment, particularly because mapping accuracy directly influences the results of many downstream tasks and the running time could potentially be a computational burden. Sensitivity (*S*) is determined as the ratio of reads mapped correctly to the reads mapped incorrectly at a particular threshold (*S* = number of reads mapped correctly/number of reads mapped incorrectly) [27]. We take the advantage of having simulated reads with known biases (considered as gold standard) to assess the performance of the 17 aligners and then test them on the empirical datasets. We decided to compare alignment performance firstly by stratifying against all reported MAPQ and Alignment Scores (AS), found in SAM format [56], then analyzing mapping efficiency (*E*), as accounted for other authors [57], and bearing in mind that the MAPQ assumes specific ranges for each alignment algorithm. AS is a score describing sequence similarity between a query and a reference. AS increases with the number of matches and decreases with the number of mismatches and gaps (rewards and penalties for matches and mismatches depend on the used scoring matrix). MAPQ is a metric affected by position. It equals 10 log$_{10}$ of probability that mapping position is wrong, rounded to the nearest integer. Based on quantiles from AS and MAPQ distributions (Table S1), we considered high AS and low MAPQ if reads aligned perfectly at multiple positions ("non-uniquely mapped reads"), while low AS and high MAPQ if reads aligned with mismatches but the reported position is still much more probable than any other ("uniquely mapped reads"). In order to make comparable MAPQ and AS across aligners, we evaluated the percentages of reads beyond the two parameters thresholds, mainly focusing on MAPQ. The mapping efficiency, corresponding to the total number of reads that align, generally depends on various factors such as (1) the read length, (2) quality of the reads, (3) the absence of contaminants (such as other species or adapter contamination), (4) the mapping software used and (5) the

Table 3 List of alignment tools with salient features

ALIGNER	Version	Designed for (Type of Data)	Prog. Lang	Algorithm	Input	Output	Mode	Clip	Mismatch Evaluation	Max InDels	Gaps
BBMAP	38.90	DNA, RNA	Java	SW	FASTA, FASTQ	BAM	Global	Soft	Simple Score (Levenshtein distance model)	8	Y
BOWTIE2	2.4.2	DNA, RNA	C++	BWT	FASTA, FASTQ	SAM	Local	Soft	Substitutional Matrices Score	8	Y
BWA	0.7.17	DNA	C++	BWT	FASTA, FASTQ	SAM	Local	Soft and Hard	Simple Score (Levenshtein distance model)	8	Y
BWA-MEM	0.7.17	DNA	C++	BWT	FASTA, FASTQ	SAM	Local	Soft and Hard	Simple Score (Levenshtein distance model)	8	Y
CLC GENOMICS WORKBENCH	21.0.3	DNA, RNA	Java	BWT	FASTA, FASTQ	BAM	Global, Local	Soft and Hard	Substitutional Matrices Score	8	Y
DNASTAR LASERGENE SUITE	17.2.1	DNA, RNA	Java	NW	FASTA, FASTQ	BAM	Global, Local	Soft and Hard	Simple Score (Levenshtein distance model)	8	Y
GEM	3.6	DNA, RNA	Python	BWT	FASTA, FASTQ	SAM	Global, Local	Soft	Substitutional Matrices Score	8	Y
HISAT2	2.2.1	DNA, RNA	Python	BWT	FASTA, FASTQ	SAM	Local	Soft	Substitutional Matrices Score	8	Y
MAGICBLAST	1.5.0	DNA, RNA	C++	BWT	FASTA, FASTQ	SAM	Local	Soft	Substitutional Matrices Score	8	Y
MINIMAP2	2.16	DNA, RNA	C, Python	BWT	FASTA, FASTQ	SAM	Global	Soft	Substitutional Matrices Score	8	Y
NOVOALIGN	4.03.02	DNA, RNA	C++	NW	FASTA, FASTQ	BAM	Global, Local	Soft and Hard	Substitutional Matrices Score	8	Y
YARA	1.0.3	DNA	C++	BWT	FASTA, FASTQ	BAM	Global	Soft	Substitutional Matrices Score	8	Y
RUM	2.0.5	DNA, RNA	Perl	BLAT	FASTA, FASTQ	BAM	Local	Soft	Substitutional Matrices Score	8	Y
SEGEMEHL	0.3.4	DNA, RNA	C	BWT	FASTA, FASTQ	SAM	Local	Soft and Hard	Substitutional Matrices Score	8	Y
STAR	2.7.0f	RNA	C++	BWT	FASTA, FASTQ	SAM	Local	Soft	Substitutional Matrices Score	8	Y
SUBREAD	2.0.1	DNA, RNA	C	BWT	FASTA, FASTQ	BAM	Local	Soft	Substitutional Matrices Score	8	Y
TOPHAT2	2.1.1	RNA	C++	BWT	FASTA, FASTQ	BAM	Global	Soft	Substitutional Matrices Score	8	N

ALIGNER	Version	Designed for (Type of Data)	MAPQ Cut-off	AS Cut-off	Index/ Hash	Salient Features	Bib	N° Cit
BBMAP	38.90	DNA, RNA	40	0.76 (Min. Aligner Identity)	Hashing	Rapidly indexes genome using short kmers, without size or scaffold count limit. Higher sensitivity and specificity than Burrows–Wheeler aligners, with comparable or increased speed. Could operate on Sanger, 454, PacBio, Illumina and Ion Torrent data. Splice-aware	Bushnell et al. (2017)	218
BOWTIE2	2.4.2	DNA, RNA	42	51 (50 bp), 60 (150 bp), 62 (200 bp)*	FM index	Modified Ferragina and Manzini matching algorithm, quality aware backtracking. Its sensitivity is high for reads > 50 bp when compared with Bowtie	Langmead et al. (2009)	25,200

Table 3 (continued)

ALIGNER	Version	Designed for (Type of Data)	MAPQ Cut-off	AS Cut-off	Index/ Hash	Salient Features	Bib	N° Cit
BWA	0.7.17	DNA	37	51 (50 bp), 60 (150 bp), 62 (200 bp)*	FM index	BWT-based aligner that uses the Ferragina and Manzini matching algorithm to find seeds, followed by a backtracking algorithm which searches for matches between a substring of the reference genome and the query within a specific defined distance. Generally used for mapping less divergent sequence	19	27,425
BWA-MEM	0.7.17	DNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	FM index	Conceptually based on BWA, but able to handle long reads and considered to be faster and more accurate	19	27,425
CLC GENOMICS WORKBENCH	21.0.3	DNA, RNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Suffix Array	Based on CLC Assembly Cell 5.0 (CLC5), employs a BWT of the reference genome rather than using a memory-intensive Suffix Array. In contrast to BWA algorithm, CLC5 parallelizes the transformation of individual chromosomes	QIAGEN (2012)	36,500
DNASTAR LASERGENE SUITE	17.2.1	DNA, RNA	70	160	Hashing	Performed by DNASTAR's SeqMan NGen® software, it aligns reads against a database of genomic templates, performing reference-guided assemblies and de novo assemblies of up to 30 million sequence reads (genome sizes up to 50 megabases)	DNASTAR (2012)	26,300
GEM	3.6	DNA, RNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	FM index	More accurate and efficient than Bowtie2 or BWA, it is characterized by high-quality alignment engine (exhaustive mapping with substitutions and INDELs). Various standalone biological applications (mappability, mapper and other) provided	39	495
HISAT2	2.2.1	DNA, RNA	42	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Hierarchical Graph FM index (HGFM)	Based on an extension of BWT for graphs, the HGFM implements a global graph FM index (GFM), that represents a population of genomes, along with a large set of small GFM indexes (local indexes) that collectively cover the whole genome	40	6245
MAGICBLAST	1.5.0	DNA, RNA	42	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Hashing	Each read alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-Seq, locating the candidate introns and adding up the score of all exons. It performs hit extensions by local walk and jump, and recursive clipping of mismatches near 5'- and 3'-ends	Boratyn et al. (2019)	50
MINIMAP2	2.16	DNA, RNA	60	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Hashing	It is faster and more accurate on simulated long reads and produces. For > 100 bp Illumina short reads, it is several times as fast as BWA-MEM and Bowtie2, and as accurate on simulated data	42	1757
NOVOALIGN	4.03.02	DNA, RNA	67	90	Hashing	Alignment quality scores uses posterior alignment probability (Allison et Wallace, 1994). It could report multiple alignments per read	NOVOCRAFT (2014)	3880

Table 3 (continued)

ALIGNER	Version	Designed for (Type of Data)	MAPQ Cut-off	AS Cut-off	Index/ Hash	Salient Features	Bib	N° Cit
YARA	1.0.3	DNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Hashing	Exhaustive enumeration of sub-optimal end-to-end alignments under the edit distance; excellent speed, memory footprint and accuracy; support for reference genomes consisting of millions of contigs	44	22
RUM	2.0.5	DNA, RNA	42	20	FM index	It exploits the advantages of both genome and transcriptome mapping as well as combining the speed of Bowtie with the sensitivity and flexibility of Blat (Kent, 2002). It also has a strand specific mode	45	359
SEGEMEHL	0.3.4	DNA, RNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Enhanced suffix array	It is able to detect not only mismatches but also insertions and deletions. It is not limited to a specific read length and is able to detect primer- or polyadenylation contaminated reads correctly	Hoffmann et al. (2009)	522
STAR	2.7.0f	RNA	240	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Uncompressed suffix array	Designed to align the non-contiguous sequences, originally due to intron–exon boundaries, directly to the reference genome, it consists of two major steps: seed searching step and clustering/ stitching/scoring step	47	15,763
SUBREAD	2.0.1	DNA, RNA	40	51 (50 bp), 60 (150 bp), 62 (200 bp)*	Hashing	Superfast and accurate read aligner, employs a novel mapping paradigm named seed-and-vote	Liao et al. (2013)	1535
TOPHAT2	2.1.1	RNA	50	51 (50 bp), 60 (150 bp), 62 (200 bp)*	FM index	It aligns RNA-Seq reads to mammalian-sized genomes using the ultra-high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons	18	9819

In this table are shown all tools exploited for alignment comparison with their own specific features. Prog. Lang. = programming language. Clip = clipping. Bib. = bibliography. SW = Smith–Waterman. BWT = Burrows–Wheeler transform. NW = Needleman–Wunsch. Max InDels = maximum n° allowed InDels for read (default). Gaps = gapping alignment. MAPQ = mapping quality score (its main value and range differ across mappers). AS = alignment score. N° Cit. = number of citations by January 2021. * This score is calculated on the basis of "Local Alignment = 20 + 8.0 * ln(L)," where "L" is the read length.

quality of the reference genome assembly. Additionally, the type of library may impact the aligner efficiency: for instance, some library protocols [58] tend to enrich for repeated sequences, possibly due to a high number of PCR cycles and/or starting from a small amount of material, and would display lower mapping efficiency than others. Additionally, regardless of the limitations related to both reads and the quality of the reference, several aligners are able to reach high mapping efficiencies by performing "clipping" of the reads. With this procedure, the portions of the read that do not align to the reference on either side of the read are ignored, though label in the CIGAR string. This process usually comes along with a small penalty for each clipped base within the read, but aims to maximize the alignment score and thus amounts to a significantly smaller alignment penalty than mismatched bases. Mapped reads clipping profile, representing the distribution of clipped nucleotides across reads were compared for all algorithms.

7 Read distribution

Different aligned data coming from different aligners were analyzed to calculate how mapped reads were distributed over genome feature (like CDS exon, 5'UTR exon, 3′ UTR exon, Intron, Intergenic regions), using the Human GENECODE annotation v.36 (www.gencodegenes.org). When genome features were overlapped (e.g., a region could be

Table 4 Aligner parameters used in the evaluation

ALIGNER	Command line used
BBMAP	bbmap.sh ref = genome.fa -Xmx20g
	bbmap.sh in1 = sequence.R1.fq in2 = sequence.R2.fq out = alignment.bam -Xmx20g
BOWTIE2	bowtie2-build -f reference.fa reference
	bowtie2 -x reference.fa -U sequence.fastq -S alignment.sam
BWA	bwa index reference.fa
	bwa aln -t 4 reference.fa sequence.fastq | bwa samse reference.fa - sequence.fastq > alignment.sam
BWA-MEM	bwa index reference.fa
	bwa mem -t 4 reference.fa sequence.fastq > alignment.sam
CLC GENOMICS WORKBENCH	Index automatically produced within software
	No Command Line, but GUI, with the following parameters: Match score = 1; Mismatch cost = 2; Cost of insertion and deletions = Linear gap cost; Insertion cost = 3; Deletion cost = 3; Length fraction = 0.5; Similarity fraction = 0.8; Global alignment = no; Auto-detect paired distances = yes; Non-specific match handling = Map randomly
DNASTAR LASERGENE SUITE	Index automatically produced within software
	No Command Line, but GUI, with the following parameters: Minimum aligned length = 35; Maximum gap size = 20; Minimum match percentage = 93; Match score = 10; Mismatch penalty = 15; Gap penalty = 40; Gap extension penalty = 5; Alignment cutoff = 160
GEM	gem-indexer -i reference.fa -o reference
	gem-mapper -I reference4.gem -1 sequence.R1.fq -2 sequence.R2.fq -o alignment.sam -t 4 –report-file = alignment
HISAT2	hisat2-build -f reference.fa reference
	hisat2 -f -x reference-1 sequence_R1.fq -2 sequence_R2.fq -S alignment.sam
MAGICBLAST	makeblastdb -in reference.fa -dbtype nucl -parse_seqids -out reference
	magicblast -query sequence.R1.fq -query_mate sequence.R2.fq -db reference -infmt fastq -out alignment.sam -num_threads 4
MINIMAP2	minimap2 -d reference.mmi reference.fa
	minimap2 -a reference.mmi sequence.fq > alignment.sam
NOVOALIGN	novoindex reference.nix reference.fa
	novoalign -d reference.nix -f sequence.R1.fq sequence.R2.fq -o BAM > alignment.bam
YARA	yara_indexer reference.fa -o reference.index
	yara_mapper reference.index sequence.R1.fq sequence.R2.fq -o alignment.bam
RUM	perl create_indexes_from_ucsc.pl NAME_genome.txt NAME_refseq_ucsc
	rum_runner align
	–index $RUM_INDEXES/REFERENCE \
	–output data/Lane1\
	–name Lane1\
	–chunks 1\
	data/Lane1/forwardreads.txt data/Lane1/reversereads.txt
SEGEMEHL	segemehl.x-x reference.idx-d reference.fa
	segemehl.x-i reference.idx-d reference.fa -q sequence_R1 -p sequence_R2 > alignment.sam
STAR	STAR–runThreadN 4 –runMode genomeGenerate –genomeDir Genome_data/star\
	–genomeFastaFiles Genome_data/reference.tar.gz
	STAR --readFilesIn sequence.fastq\--alignIntronMax 1\
	--genomeLoad LoadAndKeep\--genomeDir /path/to/genomeFasta/\--runThreadN 4 \
	--outStd SAM > alignment.sam
SUBREAD	subread-buildindex-o reference.fa
	subread-align -d 50 -D 600 -t 1 -T 4 -i mm10 -r sequence.R1.fq -R sequence.R2.fq -o alignment.bam

Table 4 (continued)

ALIGNER	Command line used
TOPHAT2	bowtie2-build -f reference.fa reference tophat -o alignment reference sequence.R1.fq sequence.R2.fq

The first point refers to genome indexing creation, while the second to the read mapping task. All other arguments are set to default if not specified. reference.fa = reference genome in FASTA format. sequence_R1.fq and sequence_R2.fq = input paired-end reads

annotated as both exon and intron by two different transcripts), they were prioritized as: CDS exons > UTR exons > Introns > Intergenic regions. Tags assigned to "TSS_up_1kb" were also assigned to "TSS_up_5kb" and "TSS_up_10kb," tags assigned to "TSS_up_5kb" were also assigned to "TSS_up_10kb." Therefore, "Total Assigned Tags" = CDS_Exons + 5'UTR_Exons + 3'UTR_Exons + Introns + TSS_up_10kb + TES_down_10kb. When tags were assigned to genome features, each one was represented by its middle point. Reads were unassigned if: (1) hit to regions covered by both 5′- and 3′ UTR, when two head-to-tail transcripts are overlapped in UTR regions; (2) hit to intergenic regions that beyond region starting from TSS upstream 10Kb to TES downstream 10Kb; (3) hit to regions covered by both TSS upstream 10Kb and TES downstream 10Kb.

8 Deletion and insertion profiles

Sequencing reads covering InDels typically map with more difficulties since their correct alignment involves complex gapped alignment. Software tools needed during high-throughput sequencing focus their improvements on InDel detection analysis. Several studies described the effects of different alignment tools on detection's efficiency [59]. These works recommend the use of gap-aware aligners. Nevertheless, further knowledge on the effects of these tools is still required. Distributions of deletions and inserted nucleotides across reads were calculated for all considered algorithms, distinguishing the effects of gapped alignments.

9 GC content distribution of reads

The GC content distribution was evaluated by FastQC [60]. Among the challenges, GC bias in NGS data is known to aggravate genome assembly/alignment [61–66]. However, it is not clear to what extent GC bias affects genome assembly or mapping in general. The average GC% along the reads indicates, for example, if considered reads are properly trimmed, in order to avoid residual adapter sequences. The whole distribution of GC content is shown as a function of the position in each read. In an unbiased library, the mean GC per read distribution must be Gaussian and centered on expected GC% (the GC% for humans is known and is 41%), if no contamination/technical bias is present. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset, possible emerging from incorrect alignment. Warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads.

10 Duplication profiles

Duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation (e.g., library construction using PCR) or result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument (optical duplicates). They were detected with Picard MarkDuplicates tool [67], which compares the anchors of mapped reads from a SAM/BAM file. After duplicate reads are collected, the tool differentiates the primary and duplicate reads using an algorithm that ranks reads by the sums of their base-quality scores. It is still unclear whether removing read duplicates computationally improves accuracy and precision by reducing PCR bias and noise or whether it decreases accuracy and precision by removing genuine information. Generally, we mark duplicates (e.g., do not remove them) only for data from WGS/exome experiments or from analyses where amplification artifacts might be a problem (ChIP-Seq for example), while should never be removed in any quantitative experiment, such as RNA-seq, because they may be part of small highly expressed transcripts [68, 69]. Two strategies were used to determine read duplication rate: (1) sequence based: reads with identical sequence are considered duplicated reads. (2) mapping based: reads mapped to the exactly same genomic location are evaluated as duplicated reads. For spliced reads, reads mapped to the same starting position and splice the same way are regarded as duplicated reads.

11 Saturation profiles

Restricted to RNA-Seq data, saturation profiles highlight how the precision of any summary statistics (Reads Per Kilobase Million, RPKM) is affected by sample size (sequencing depth). Such profiles are obtained by taking random sub-samples (repeated 20 times) of a given number of reads every 5th percentile up to the total library size and then by calculating the RPKM value of each subsample. Such curve allows to check if the RPKM estimates plateaus, allowing to adjust for differences in sample sizes. In this way, it is possible to check if the current sequencing depth was saturated or not (or if the RPKM values were stable or not) in terms of gene expression estimation. If sequencing depth was saturated, the estimated RPKM value will be stationary or reproducible. For each transcript, we recorded at which bin the saturation is reached and used this percentage to classify the transcripts. In details, twenty RPKM values (using 5%, 10%,…, 95%,100% of total reads) were calculated for each transcript, and the "percent relative error," used to measure how the RPKM estimated from subset of reads (i.e., $\mathrm{RPKM}_{\mathrm{obs}}$) deviates from real expression level (i.e., $\mathrm{RPKM}_{\mathrm{real}}$), was derived and plotted. As a proxy, the RPKM estimated from total reads was used to approximate $\mathrm{RPKM}_{\mathrm{real}}$.

$$\text{Percent relative error} = \left(|\mathrm{RPKM}_{\mathrm{obs}} - \mathrm{RPKM}_{\mathrm{real}}| \,/\, \mathrm{RPKM}_{\mathrm{real}}\right) \times 100$$

Then all transcripts were sorted in ascending order according to expression level (RPKM) and finally divided into 4 groups:

- Q1 (0–25%): transcripts with expression level ranked below 25 percentile.
- Q2 (25–50%): transcripts with expression level ranked between 25 and 50 percentile.
- Q3 (50–75%): transcripts with expression level ranked between 50 and 75 percentile.
- Q4 (75–100%): transcripts with expression level ranked above 75 percentile.

12 Estimation of computational time and mean RAM usage for alignment

We performed all analyses on a high-end MacBook Pro with Intel® Core$^{\mathrm{TM}}$ (Intel Core i7-8750H CPU @ 2.2 GHz with Turbo Boost to 4.1 GHz six-core) processor and a maximum memory of 32 GB of RAM with AMD Radeon Pro 555X graphics and MacOS Big Sur 11.2.0/Ubuntu 20.04 LTS OS for alignment jobs. The scalability of the mapping tools may be different under different parallel settings. Many tools support multithreading, which is expected to yield linearly increasing speedup with an incremented number of CPU cores. However, using multiprocessing is more general and may improve the throughput even for tools that do not support multithreading, where multiprocessing refers to using more than one process in a distributed memory fashion while communicating through a message passing interface. All our used tools support multithreading/multiprocessing, except YARA and RUM.

13 Evaluation of simulated data statistics

Alfred (https://www.gear-genomics.com/alfred) allowed a complete integration of high-throughput settings, enabling the monitoring of sequence data quality and characteristics across samples. It parsed the aligned BAM files only once and pre-allocated data structures for counting primary, secondary, supplementary and spliced alignments. Sequencing error rates were elaborated separately for mismatch, insertion and deletion errors. InDel size distribution was estimated and potential homopolymer sequence regions and a fragment-based GC bias curve were estimated from the reference context. Evaluated parameters of Alfred alignment metric are available in Table 5.

14 Results

14.1 Simulated data highlight 7 potential best-performer aligners

Alignment of simulated data by all 17 considered algorithms showed that the best mapping was reached by BBMAP, BWA-MEM, Novoalign, DNASTAR, YARA, Segemehl and TopHat2 for all paired-end simulated data, while the worst one was Subread. Only Mason data showed a high-quality alignment for BWA. A different situation resulted from mapping of CuReSim data, due to the ability of tool to generate single-end reads only. Thus, this time the best outputs emerged from BBMAP, YARA and DNASTAR, while non-optimal alignments were performed by Bowtie2, BWA, Hisat2, Minimap2, CLC and STAR. Distributions of mapped and unmapped reads by selected aligners on simulated data are shown in Fig. 1, while detailed statistics of each simulated data alignment is available in Table S2.

15 Accuracy and efficiency of mapping evaluation on real data

We evaluated the tools on four types of data sets, namely two DNA-Seq samples (paired-end WES and single-end WGS). During an evaluation procedure, it is essential to choose the right data set type to improve the applicability of the tools. We concatenated our 4 datasets in a single file and added a read group to find back the origin of each read. The uniquely mapped reads were separated from multiple mapped (the percentage of reads mapped to more than one location with the same number of mismatches, highlighting that these reads could fall in repetitive regions) and unmapped ones by Alfred. Furthermore, we reported the distribution of mapped reads over genomic features, as number of mapped tags per kb (tags/kb) of reference genome. Indexing and matching times for selected tools were split (see further). Figure 2 reports the percentage of mapped (uniquely and multiple) and unmapped reads of various sizes aligned using all chosen 17 different aligners. Figure 3 highlights distribution of reads assigned to specific genome features for each selected aligner in all samples. CLC, BWA-MEM, GEM and Magic-BLAST mapped the highest number of reads in all samples (% of mapped reads $> 90\%$), while the smallest fraction of aligned reads was realized by TopHat2 in three samples of four (only single-end RNA-Seq data saw BWA and Bowtie2 worse than it). Interestingly, paired-end samples showed a high mapping percentage ($> 99\%$) for DNASTAR, BBMap and RUM, while a lower percentage was evidenced by STAR ($< 45\%$ in osteomyelitis sample). About accuracy, it was globally high for both DNA-Seq samples and more variable through different aligners in RNA-Seq ones. Novoalign and DNASTAR showed the highest accuracy in all alignments, followed by CLC algorithm. Segemehl reached the highest result for exon mapping (591 tags/kb), while BBMap obtained the worst result in all samples (12 tags/kb represented its peak, in WES sample). Additionally, only for WGS sample, Novoalign got the highest value of total read mapping parameter (160 tags/kb). Considering the read distribution, the best aligner varies with the analyzed sample: RPE transcriptome showed that Segemehl, BWA-MEM, Magic-BLAST and Minimap2 reached the best results aligning exon sequences; the same results were achieved by Hisat2, STAR, TopHat2 and RUM in osteomyelitis sample (here RUM also highlighted a peak for intron aligned sequences). RUM itself, with Segemehl, reached the best mapping score for uniquely aligned coding sequences in WES analysis, similarly for Novoalign in WGS sample. The mapping tool which performed the worst results in exon alignment was BWA.

16 Clipped reads

Commonly used aligners perform “clipping,” excluding the unalignable portion of a read which was not mapped in its full length. With soft-clipping, highlighted in the CIGAR string with the letter “S,” the clipped sequence bases will not be used by variant callers or other downstream tools. As expected, Bowtie2, Segemehl and TopHat2 showed no clipped reads during alignment in all samples, as YARA for RPE transcriptome and BBMap for retinitis pigmentosa WES. Very interestingly, all other mappers reached a very high percentage of non-clipped reads in both DNA-Seq data ($> 98\%$ in WES and $> 92\%$ in WGS analyses, respectively). An increased number of clipped reads are, instead, revealed by both RNA-Seq data, especially by RPE cell transcriptome. About the latter, BWA-MEM, Magic-BLAST and CLC did not pass the value of 82% non-clipped reads until read position 35 for the first two, and 60 for the third, respectively. A similar trend was shown by Minimap2, Subread and Novoalign that reached the same value after 35 bp across the mapped reads, with an additional decrease to 75% on non-clipped reads around 150 bp. Regarding osteomyelitis sample, the only noticeable variation was highlighted by BWA-MEM, in which the percentage of 92% was obtained only in read length range of 20-80 bp. Detailed clipping profiles are represented in Figs. S1–S4.

17 Insertions and deletions

The “highest” insertion percentage was detected by GEM in RPE transcriptome sample, with a value of about 2%, while in the osteomyelitis data all tools reached a very little percentage (about 0.02%). In DNA-Seq samples, the insertion percentage remained very low, ranging from the 0.002% of Magic-BLAST, BWA, CLC, Minimap2 and TopHat2 to 0.010% of GEM in WES data, and ranging from 0.15% of Bowtie2, BWA and TopHat2 to 0.5% of GEM and Segemehl in WGS one. About distribution of deletions across the reads, the compared analysis was more complex, because alternative transcripts could create InDels. The only tool which outputted common results for all samples was BBMap, which did not detect any deletion. RPE cell transcriptome data showed a huge growing trend for the most of aligners, with a little decrease only for Subread, and lower peaks (value about 20) for Bowtie2, BWA and TopHat2, the last of whom also presented a very irregular deletion distribution along the read. About Illumina aligned data, both showed a stable trend around a specific deletion number for the most of mappers, mainly distributed around 30 bp of read length. The highest peak

Table 5 Alignment metrics used in the evaluation

Alignment Metric	DNA-Seq WGS	DNA-Seq WES	RNA-Seq
Mapping Statistics Duplicate	✔	✔	✔
Statistics Sequencing Error Rates	✔	✔	✔
Base Content Distribution Read	✔	✔	✔
Length Distribution Base Quality	✔	✔	✔
Distribution Coverage Histogram	✔	✔	✔
Insert Size Distribution InDel Size	✔	✔	✔
Distribution InDel Context	✔	✔	✔
GC Content	✔	✔	✔
On-Target Rate		✔	
Target Coverage Distribution		✔	
Spliced Alignments			✔
Feature Counting	✔	✔	✔
Feature Annotation	✔	✔	✔

The absence of tick indicates that the specific metric is not available for that aligner

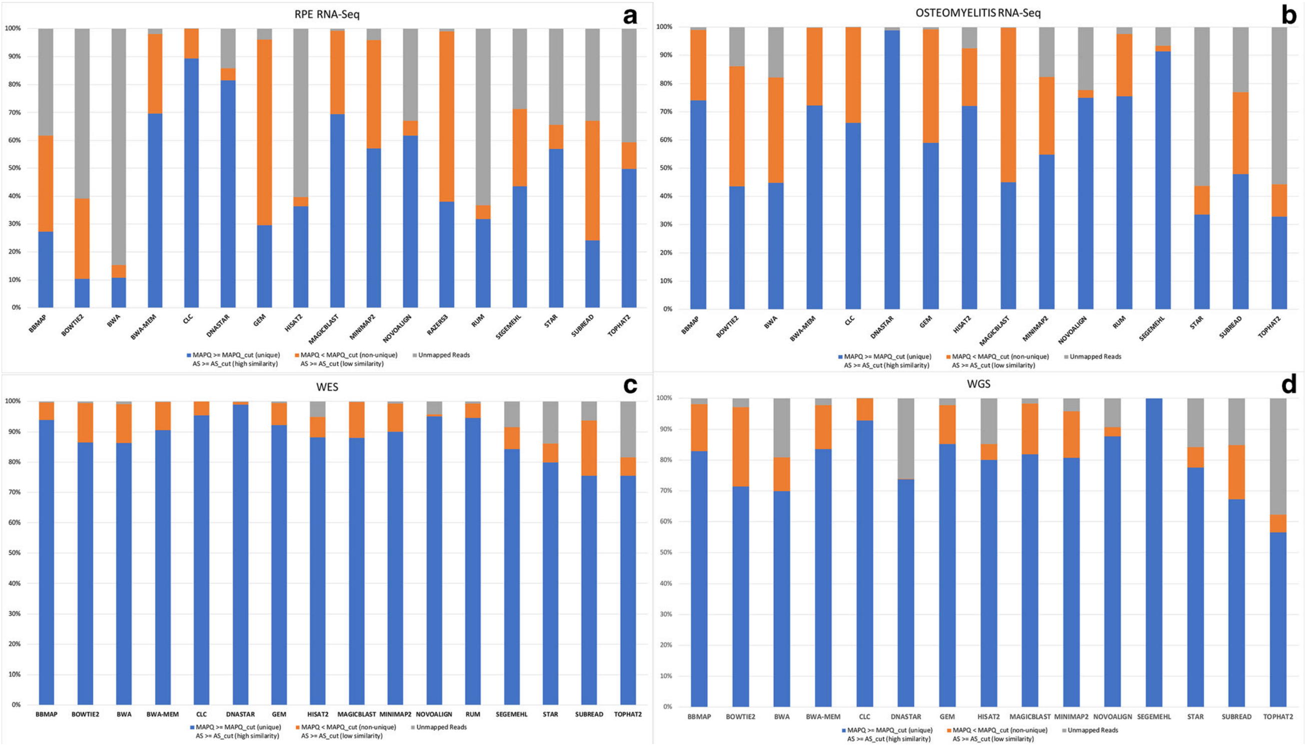

Fig. 1 Distribution of mapped and unmapped reads by selected aligners on simulated data. Normalized bar plots indicate the partitioning of unmapped and mapped (uni- vs multimapped) reads, based on MAPQ and AS scores, for all selected aligning algorithms in five simulated datasets. A = Mason. B = DwgSim. C = WgSim. D = Art. E = CuReSim. Refer to Table 3 for MAPQ and AS thresholds for each aligner

was reached by CLC (5500 in osteomyelitis and RPE samples, at 30 bp and 70 bp, respectively) and the lowest by Subread (0 in osteomyelitis data and 1800 in WES), while an irregular "up and down" distribution was highlighted by GEM (between 15,000 at 25 bp and 1,500 at 38-70 bp in osteomyelitis sample; 4,100 at 18, 30 and 50 bp, 4,500 at 65 bp and 5,500 at 85 bp). Finally, whole-genome sequencing data presented an increasing deletion distribution throughout the entire read length for all mapping tools, with the highest value ranging from 600 for BWA-MEM to 3200 for Segemehl, both at 120 bp. The exception was represented by Subread and TopHat2, which exhibited a very particular behavior, with many and irregular peaks across the reads. Results of insertion and deletion analyses are plotted in Figs S5–S12.

18 GC content

In this work, we also conducted a systematic analysis on the effects of GC bias on genome mapping. In represented plots, the presence of double peaks in GC distribution could reflect some noise at the start of the per-base nucleotide distribution due to not-so-random hexamers. In the case of RNA samples, the second peak could also refer to an incomplete rRNA depletion during library preparation. Finally, the theoretical GC curve should be centered closer to the expected mean GC% value of the studied organism. If the read GC distribution is nonsignificantly different to the GC distribution of the reference genome, that implies no bias, assuming all of the reads originated from that reference genome. High and low GC tails, relative to the expected GC% of the organism, are more likely to represent repetitive regions and tandem repeats. So, using a poor mapper or a highly-repetitive organism determines that the seemingly higher coverage of extreme GC areas is actually due to the fact that they are collapsed repeats. Mapping with high error rates is not an issue, but there could be edge effects if you are mapping to short contigs. In RNA-Seq samples, BWA-MEM, CLC and Novoalign shown a normal distribution very close to the theoretical normal one. Moreover, the same result was obtained for Magic-BLAST, Minimap2 and Subread for RPE cell transcriptome data, and for BBMap and Hisat2 for osteomyelitis one. In the latter sample, interestingly, several tools (Bowtie2, BWA, GEM, Minimap2, STAR and Subread) highlighted a double peak, and two algorithms (Magic-BLAST and Segemehl) presented even irregular trend. About DNA-Seq data, Magic-BLAST and Segemehl show the most normal distribution of GC in WES sample, where all the other tools revealed a double peak in their trends, with the second higher than the first. In WGS data, instead, no algorithm could reach an approximal normal distribution, probably due to sequencing biases: all aligners shown a left-skewed distribution, with Magic-BLAST and Segemehl also characterized by a lower density of reads. GC distribution emerged from all alignments are represented in Figs. S13–S16.

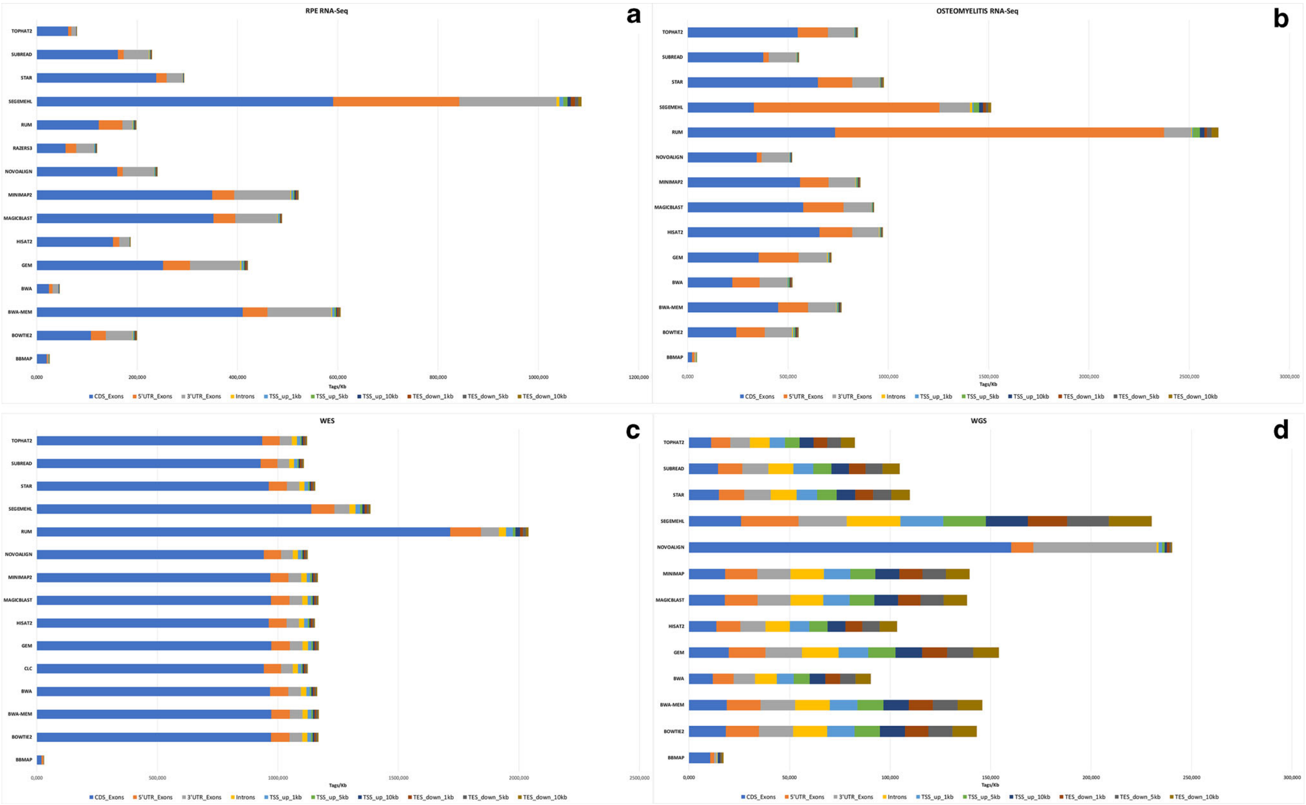

Fig. 2 Distribution of mapped and unmapped reads by selected aligners on empirical data. Normalized bar plots indicate the partitioning of unmapped and mapped (uni- vs multimapped) reads, based on MAPQ and AS scores, for all selected aligning algorithms in four analyzed samples (**a–d**). Refer to Table 3 for MAPQ and AS thresholds for each aligner. YARA results are only available for RPE RNA-Seq sample, probably due to computational hardware limitations

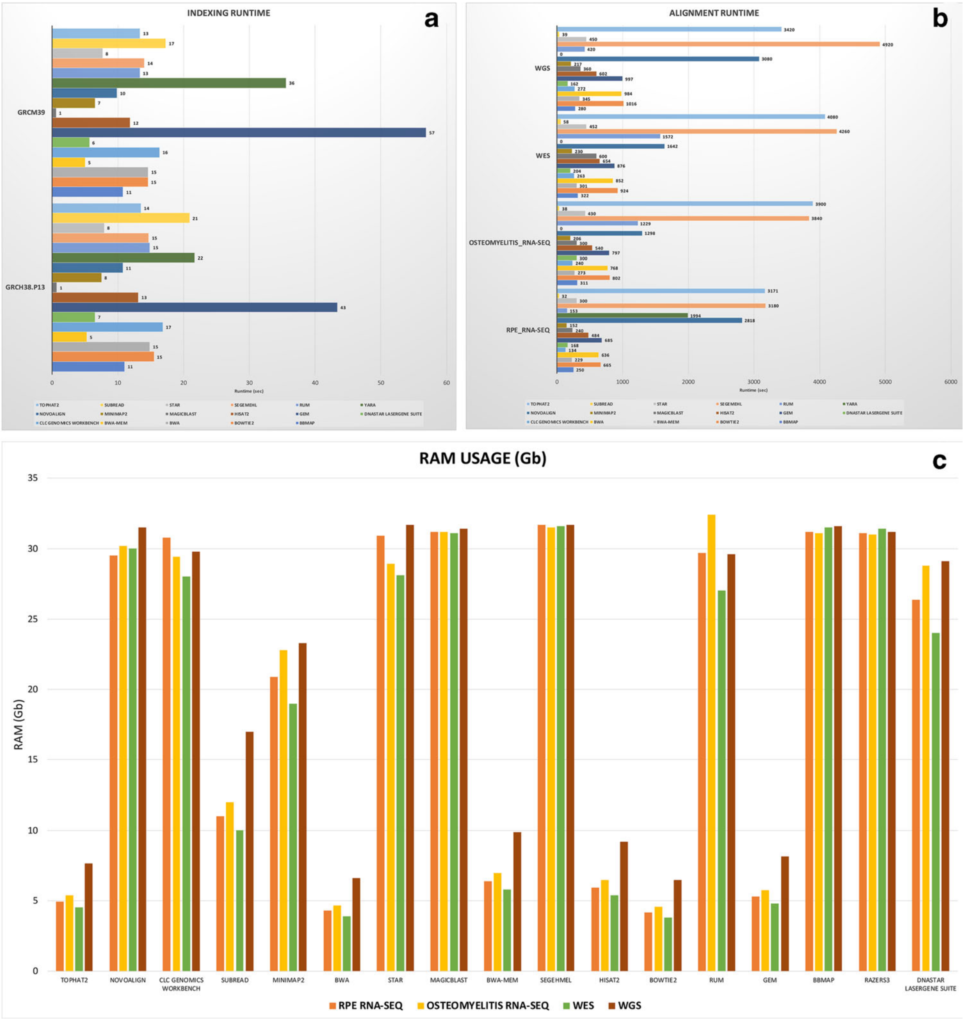

Fig. 3 Tag assigned to genome features by chosen mappers. Bar plot highlights distribution of reads assigned to specific genome features for considered aligners in all samples. TSS_up: upstream of the transcription start site. TSS_down: downstream from the transcription terminal site. Probably due to lack of sufficient computational resources or due to intrinsic algorithmic peculiarities, few mapping tools were unable to be evaluated for this parameter (CLC and DNASTAR in **a**; CLC, DNASTAR and YARA in **b**; DNASTAR and YARA in **c**; CLC, DNASTAR, YARA and RUM in **d**)

19 Duplication profiles

As expected, read duplication rates are very high in RNA-Seq samples and almost paltry in DNA-Seq ones, especially in WGS data. Read duplicates are not normally removed from the RNA-Seq data, because PCR duplicates are not distinguishable from the same fragments that are highly expressed. In RPE transcriptome, BBMap, BWA-MEM, Magic-BLAST and Segemehl detected the highest number of total duplications in the highest number of reads. Duplicate sequences detection resulted more divergent from duplicate mapped reads in BBMap, CLC and Magic-BLAST alignments. The same deviating trend is shown by Magic-BLAST and by Segemehl in the other three samples, with the latter aligner able to detect the most consistent number of duplicated mapped reads in WGS data. Figures S17–S20 highlight duplication rates from selected tools.

20 Saturation profiles

Saturation analyses of both RNA-Seq samples highlighted relevant differences between aligners in transcripts with expression level ranked below 25 percentile (Q1). In detail, RPE transcriptome data shown different RPKM saturations distributed in five tool groups: (1) Bowtie2, BWA and TopHat2, in which the median of percent relative error decreases from 100 to 75% and, soon after, under 50%, between 30 and 35 resampling percentage; (2) BWA-MEM, Magic-BLAST, Minimap2 and Segemehl, characterized by percent relative error median reaching 50% around 10–15% of resampling percentage; (3) GEM and Novoalign, whose median of percent relative error dropped below 50% only after 20% of resampling rate; (4) Hisat2, YARA and Subread, whose saturation plots reached the 50% of median of percent relative error at 25% of resampling percentage; (5) STAR that highlighted a trend very similar to group 4 aligners, but with first box-plots showing shorter quartiles than their counterpart in group 4. BBMap and CLC gave errors during saturation analysis, so they are not present in our results. Less intragroup differences were, then, shown by aligners in osteomyelitis sample, clustering tools into only 4 groups, one of which comprising the highest number of algorithms: (1) BBMap that highlighted a very strange trend, decreasing the median of percent relative error under 50% only after 65% of resampling, both in Q1 and Q2, and with incomplete box-plot representation in the same quartiles; (2) Bowtie2 and STAR, whose median of percent relative error fell below 50% after 20% of resampling rate; (3) BWA-MEM, BWA, GEM, Hisat2, Magic-BLAST, Minimap2, Segemehl, Subread and TopHat2 that reached the 50% of median of percent relative error at around 15 resampling percentage; (4) Novoalign, which showed a trend very similar to group four, but with the median of percentage error under 50% already at 10% of resampling. About osteomyelitis sample, data from CLC could not be analyzed. Sequencing saturation profiles are plotted in Figs. S21–S22.

21 Computational performance comparison of selected aligners

The computational time of read mapping of different aligners was directly proportional to the size of the reference genome. In our analyses, because all four samples were mapped against the same human reference genome, the alignment time mainly depended on the read size and number. A comparative analysis of alignment time showed Subread was significantly faster in all samples (~ 1936 s for single-end RPE transcriptome and ~ 2314 s for single-end WGS), followed by CLC and Minimap2 in RNA-Seq data, and by DNASTAR and Minimap2 in DNA-Seq ones. Segemehl, TopHat2 and Novoalign took the longest alignment time (~ 295,200 s for WGS sample). About memory consumption, all four samples showed a mean RAM usage with a similar trend for all aligners. TopHat2, BWA, BWA-MEM, Hisat2, Bowtie2 and GEM highlighted the lowest RAM consumption (mostly < 10 Gb, except Subread which has slightly exceeded 15 GB in the WGS sample). All other tools make maximum use of the available 32 Gb RAM. Indexing and alignment runtime, as well as mean RAM usage comparisons, are shown in Fig. 4.

22 Discussion

NGS technology has grown very rapidly during the last years, leading to huge data outputs of million sequences in a single run. However, such enormous quantity of generated data gives no really useful information about DNA [70] without the availability and the development of specific analysis algorithms and tools. For this reason, the bioinformatics research takes care to find new ways to efficiently manage and analyze such big data [71]. One of the most promising area that involves the most NGS bioinformaticians deals with the mapping of generated sequences [72]. Thus, selection of the correct aligner is fundamental. Mapper performance and specificity are determined by genome characteristics, so their evaluation depends on various criteria such as read distribution, properly paired, InDel and saturation profiles, duplications, and incorrectly mapped reads. We benchmarked 17 different aligners, starting from artificial reads obtained by 5 reads simulators, and then focused on different types of real data. Simulated data was obtained from both *homo sapiens* and *mus musculus*, with different read lengths, in order to widen the genomic features involvement into aligning analysis. In this way, we wanted to provide guidelines for the choice of the most suitable aligner [73–75]. The most unique feature of our study consisted of the use of all mappers on both RNA- and DNA-Sequencing data, even if several aligners were specifically developed for only one of them [76]. The idea of using an RNA-Seq mapping tool deals with the ability to align across intron–exon junctions. Whole genome/exome data consist of exon sequences (maybe plus a little intronic sequence in WES, depending on the probes), which are not supposed to be spliced together. Since it is not necessary to align across junctions, an RNA-designed aligner will not have an advantage over a DNA-designed aligner, but there is no evidence regarding why it could not be use either. Firstly, alignment performance by stratifying against all reported MAPQ scores and efficiency were evaluated. It is well known that they are both affected by genome size, read size and distribution of repeats [77]. Single-end long reads (150 bp and 200 bp) increased the number of unmapped reads, but decreased the number of multiple alignments, increasing the performance of alignment for all the mappers. Paired-end short reads, instead, showed the highest level of total mapped reads, increasing efficiency, and also reaching the highest percentage of uniquely mapped reads in WES data, probably the best compromise between mapping quality and efficiency. Furthermore, the trade-off between mapping quality and efficiency also depends on clipping process, that in our analyses evidenced the lowest number of clipped reads in DNA-Seq data, probably shifting balances in favor of align quality. Among different selected aligners, Magic-BLAST was the one able to map the highest number of reads in quite all samples, but at the expense of accuracy. Novoalign and DNASTAR, instead, were found to have the highest mapping quality with short and long reads, both in RNA- and DNA-Seq data. Furthermore, CLC showed similar pattern of align quality. The highest distribution of reads assigned to specific genome features was, instead,

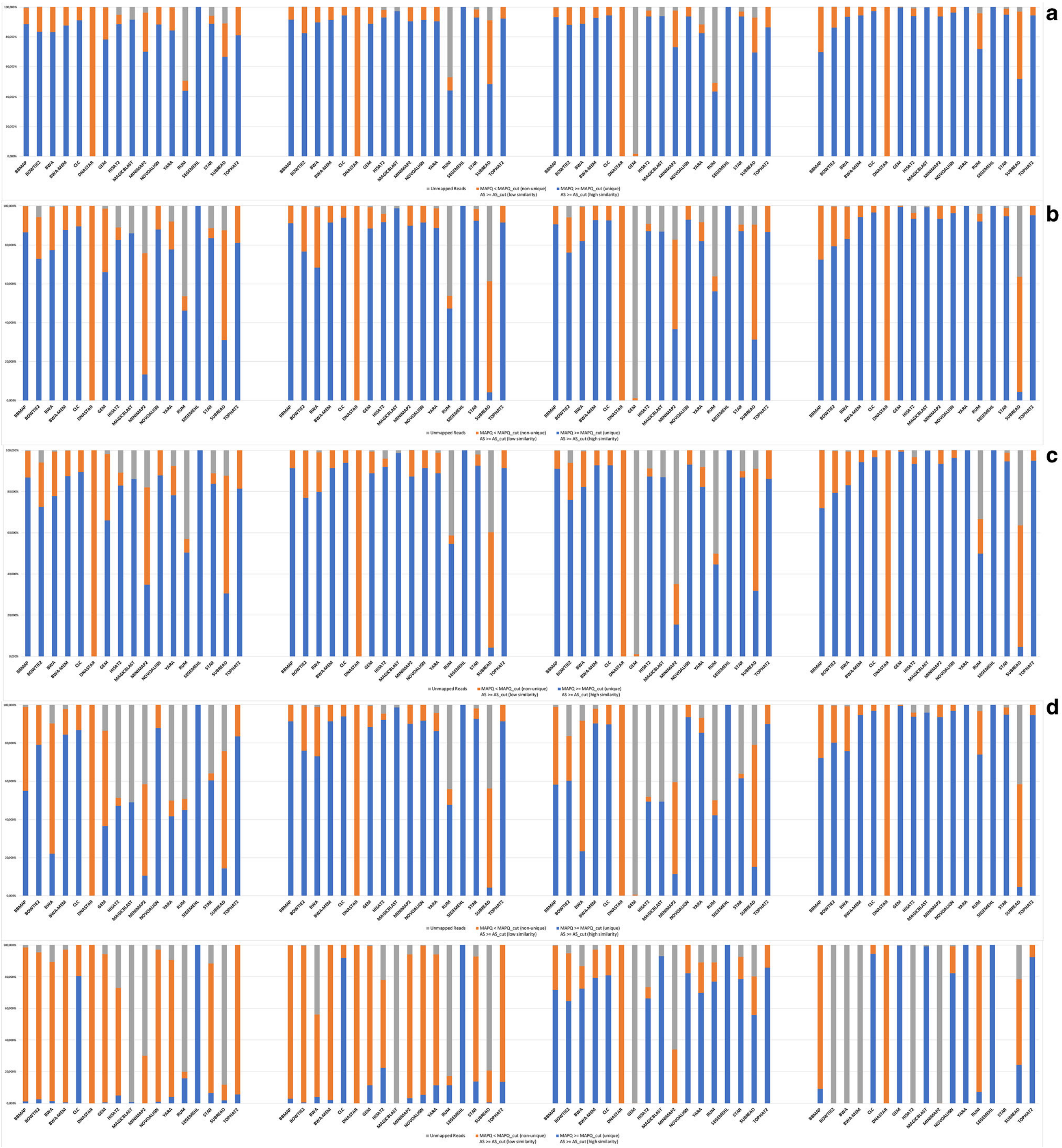

Fig. 4 Alignment computational performance comparison plots. Bar graphs show different runtime (in minutes) needed by analyzed mappers to index genomes (**A**), to complete the alignment of four considered samples (**B**), as well as RAM usage (in Gb, **C**)

reached by Segemehl and Novoalign, also able to map the huge number of exon sequences. The other aspect we focused on was the insertion/deletion profile, which usefully describe how downstream tools could correctly infer InDels [78]. Gapped alignment-based InDel detection algorithms require interpretation of the alignment results from a gapped aligner [79] such as BWA. The major obstacle of these methods is represented by the need that InDels have to be entirely contained within a read and correctly detected during the initial read mapping step (identified, in the CIGAR string, as ‘I’ for insertion and ‘D’ for deletion [80]). Furthermore, even if it is sufficient for small InDels detection, it becomes very difficult for InDels longer than 15% of the read length. In this case, supporting reads will frequently present too few bases able to match the reference genome or may contain only one end able to map correctly to the reference genome but the rest of the bases following the InDel get trimmed or soft-clipped by the NGS aligner [81]. Therefore, the previously clipping analysis resulted needed before interpreting data coming from insertion and deletion profiles. The highest insertion percentage was shown by GEM in RPE cell RNA-Seq sample, while the same parameter was really low in all other samples and aligners. Deletion percentage, instead, reached its highest peak in both Illumina samples, thanks to CLC algorithm, highlighting a basilar independency of this parameter from NGS kind of experiment. Another interesting side of alignment comparison deals with duplication rate. High coverage due to repeats has been known [82], and duplication in the genome plays a critical role in determining the quality of the aligners. Moreover, duplication should not be removed from RNA-Seq data, because they might be present due to high expression level [83]. BBMap, CLC, Magic-BLAST and, above all, Segemehl were able to detect the most elevated number of mapping-based duplications across all four samples, especially in WGS one. Furthermore, several aligners were more sensitive to several sequencing biases, like altered GC content and unbalanced nucleotide composition of reads [84]. Among them, BWA-MEM, CLC and Novoalign have suffered less from the effect of these errors in RNA-Seq samples, while Magic-BLAST and Segemehl did the same in WES data. Different situations appeared in WGS abnormal results for all aligners, probably due to low quality of the sample. Another interesting feature, this time only evaluable on RNA-Seq data, is represented by sequencing saturation, a measure of the fraction of library complexity [85]. The inverse of the sequencing saturation can be interpreted as the number of additional reads requested to detect a new transcript. Sequencing saturation is dependent on the library complexity and sequencing depth. Novoalign showed the best performance during evaluation of this parameter, highlighting the ability to detect new transcript easier than other tools. Relevant results in the same analysis were, also, reached by BWA-MEM, Magic-BLAST, Minimap2 and Segemehl. Finally, comparative analysis of computational performance showed Subread was significantly faster in all samples, followed by Minimap2. Interestingly, two of the commercial tools, CLC Genomics Workbench and DNASTAR Lasergene, resulted within the first three position in RNA-Seq and DNA-Seq alignment, respectively. Segemehl, TopHat2 and Novoalign, instead, evidenced the longest computational time of read mapping, probably due to their predisposition toward other parameters (e. g. accuracy). As already cited in results, simulated data alignment corroborated the mapping outputs of real data alignment. Data summarizing which aligner excels in relationship with input data (reads length and DNA- vs RNA-Seq) are reported in Table 6.

In this table are shown all tools exploited for alignment comparison with their own specific features. For each empirical dataset, each aligner is scored using the following criteria: Efficiency, Accuracy, Read Distribution, Deletion Profile, GC Count, Clipped Reads, Runtime, Duplication Profile and Saturation Profile. The reported number indicates the criteria an aligner performs best.

23 Limitations

One limit of the study regards the comparison of Illumina and Ion Torrent data, which could complicate the benchmark evaluation. A second problem was the need of sufficient localized computing resources to analyze data that, in our case, arrested the analysis of several resource-consuming algorithms (e.g., YARA). Another relevant issue might deal with scaling results to other datasets, element that should always keep in mind. Additionally, the choice of the aligners could be not totally exhaustive, due to practical reasons. Furthermore, we did not benchmark machine learning algorithms, such as DAVI [86], DeepFam [87], RLAlign [88] and DeepSF [89], namely neural networks and support vector machines, as well as the emerging developments in artificial intelligence, that represent the most emerging analytic field. Nevertheless, the major challenge is to deal with and interpret all the distinct output coming from each algorithm parameters.

24 Perspectives: improving the SARS-CoV-2 genome sequencing and variant discovery

Sequencing of genes and whole genomes has been recognized as a powerful technique to investigate viral pathogen genomes, understand outbreak transmission dynamics and spill-over events and screen for mutations that potentially play a pivotal role pathogenicity, transmissibility and/or countermeasures (e.g., diagnostics, antiviral drugs and vaccines). A standardized pipeline to characterize, name and report SARS-CoV-2 sequences has not been established yet, even if there are several methods available for sequencing SARS-CoV-2 from clinical samples [90]. About data analysis, methods based on reference mapping are most suitable for routine analysis, while minority variant determination or structural genomic variants detection are technically challenging and can often not be fully automated. Alignment of on-target reads to a canonical reference genome, such as the genome NCBI reference

Table 6 Scoring of aligners for each empirical dataset

ALIGNER	SCORING SYSTEM	DNA-SEQ			RNA-SEQ		
		WES	WGS	Total	RPE	Osteo	Total
BBMAP	Score	1	0	1	2	2	4
	Parameters	Efficiency	–		Deletion and Duplication Profiles	Efficiency, Clipped Reads	
BOWTIE2	Score	1	1	2	1	1	2
	Parameters	Clipped Reads	Clipped Reads		Clipped Reads	Clipped Reads	
BWA	Score	0	0	0	0	0	0
	Parameters	–	–		–	–	
BWA-MEM	Score	0	1	1	6	1	7
	Parameters	–	Efficiency		Efficiency, Read Distribution, Deletion Profile, GC Count, Duplication and Saturation Profiles	GC Count	
CLC GENOMICS WORKBENCH	Score	2	2	4	4	4	8
	Parameters	Accuracy, Deletion Profile	Efficiency, Accuracy		Efficiency, Accuracy, GC Count, Runtime	Accuracy, Deletion Profile, GC Count, Runtime	
DNASTAR LASERGENE SUITE	Score	3	2	5	2	2	4
	Parameters	Efficiency, Accuracy, Runtime	Accuracy, Runtime		Accuracy, Deletion Profile	Efficiency, Accuracy	
GEM	Score	0	1	1	3	0	3
	Parameters	–	Efficiency		Efficiency, Insertion and Deletion Profiles	–	
HISAT2	Score	0	0	0	1	1	2
	Parameters	–	–		Deletion Profile	Read Distribution	
MAGICBLAST	Score	2	2	4	5	1	6
	Parameters	GC Count, Duplication Profile	Efficiency, Duplication Profile		Efficiency, Read Distribution, Deletion, Duplication and Saturation Profiles	Duplication Profile	
MINIMAP2	Score	1	1	2	4	0	4
	Parameters	Runtime	Runtime		Read Distribution, Deletion and Saturation Profiles, Runtime	–	
NOVOALIGN	Score	1	2	3	3	3	6
	Parameters	Accuracy	Read Distribution, Accuracy		Accuracy, Deletion Profile, GC Count	Accuracy, GC Count, Saturation Profile	
YARA	Score	0	0	0	2	0	2
	Parameters	–	–		Clipped Reads, Deletion Profile	–	
RUM	Score	2	0	2	1	2	3
	Parameters	Efficiency, Read Distribution	–		Deletion Profile	Efficiency, Read Distribution	
SEGEMEHL	Score	4	3	7	5	2	7
	Parameters	Read Distribution, Clipped Reads, GC Count, Duplication Profile	Clipped Reads, Deletion and Duplication Profiles		Read Distribution, Clipped Reads, Deletion, Duplication and Saturation Profiles	Clipped Reads, Duplication Profile	

Table 6 (continued)

ALIGNER	SCORING SYSTEM	DNA-SEQ			RNA-SEQ		
		WES	WGS	Total	RPE	Osteo	Total
STAR	Score	0	0	0	1	1	2
	Parameters	–	–		Deletion Profile	Read Distribution	
SUBREAD	Score	1	1	2	2	1	3
	Parameters	Runtime	Runtime		Deletion Profile, Runtime	Read Distribution	
TOPHAT2	Score	1	1	2	1	2	3
	Parameters	Clipped Reads	Clipped Reads		Clipped Reads	Read Distribution, Clipped Reads	

sequence NC_045512, is generally realized by Bowtie2, Minimap2, BWA or BWA-MEM [91–94]. Regardless of the pipeline, nucleotide variants should not be called if the number of unique supporting reads at the site is lower than the required depth for confidence. Thus, in practice, such alignment-based methods are prone to bias ranging from false positives [95] to false negatives [96, 97]). In these scenarios, alignment-based methods may be not specific or sensitive enough. Such alignment-based methods are also computationally intensive and therefore not particularly fast or efficient. Thus, our study could shed new lights on more suitable alignment algorithms for SARS-CoV-2 analysis, permitting the bioinformaticians to evaluate the introduction of new mapping procedures inside their in-house pipelines, trying to improve the output of sequencing and reduce all types of just cited bias.

25 Conclusions

In conclusion, there is no *best* aligner among all of the analyzed ones; each tool was *the-best* in specific conditions. For Ion Torrent single-end RNA-Seq samples, the most suitable aligners resulted CLC and DNASTAR, while both DNA-Seq samples showed as “best performers” Segemehl and DNASTAR, with the first particularly performing well for WES data. Even if many studies have deeply evaluated and, then, improved actual algorithms, alignment still remains an active challenge. However, for the first time, we tried to analyze different NGS data (RNA-Seq and DNA-Seq ones) with 17 different mapping algorithms created for specific kind of experiment, showing possibilities of such attempt and limitations. We believe that also with machine learning algorithms support, the NGS technique will help scientists and clinicians to solve complex biological challenges, thus improving clinical diagnostics and opening new avenues for novel therapies development.

Authors contribution **LD** designed the study, performed analytical work, paper writing and drafting; **CS** participated in analytical work and paper drafting; **CR** participated in manuscript preparation; **RD** participated in paper drafting and assisted in study design; **AS** supervised all work and assisted in manuscript preparation. All authors read and approved the final manuscript.

Funding None.

Availability of data and material The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interests/competing interests.

Ethical approval This study was approved by the Ethics Committee of “Azienda Policlinico Universitario of Messina” and conformed to the tenets of the Declaration of Helsinki. All subjects had given written informed consent prior to participation in the study.

Consent for publication All authors give their consent for publications.

References

1. Zhao Y, Wang K, Wang WL, Yin TT, Dong WQ, Xu CJ (2019) A high-throughput SNP discovery strategy for RNA-seq data. BMC Genom 20(1):160. https://doi.org/10.1186/s12864-019-5533-4
2. Rodriguez-Garcia A, Sola-Landa A, Barreiro C (2017) RNA-Seq-Based comparative transcriptomics: RNA preparation and

bioinformatics. Methods Mol Biol 1645:59–72. https://doi.org/10.1007/978-1-4939-7183-1_5

3. Nakato R, Shirahige K (2017) Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform 18(2):279–290. https://doi.org/10.1093/bib/bbw023
4. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D (2017) DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Sci Int Genet 28:225–236. https://doi.org/10.1016/j.fsigen.2017.02.009
5. Sohn JI, Nam JW (2018) The present and future of de novo whole-genome assembly. Brief Bioinform 19(1):23–40. https://doi.org/10.1093/bib/bbw096
6. Al Kawam A, Khatri S, Datta A (2017) A survey of software and hardware approaches to performing read alignment in next generation sequencing. IEEE/ACM Trans Comput Biol Bioinform 14(6):1202–1213. https://doi.org/10.1109/TCBB.2016.2586070
7. Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1):36–46. https://doi.org/10.1038/nrg3117
8. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genom 12:245. https://doi.org/10.1186/1471-2164-12-245
9. Tan G, Opitz L, Schlapbach R, Rehrauer H (2019) Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep 9(1):2856. https://doi.org/10.1038/s41598-019-39076-7
10. Schirmer M, D’Amore R, Ijaz UZ, Hall N, Quince C (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinform 17:125. https://doi.org/10.1186/s12859-016-0976-y
11. Thompson JF, Steinmann KE (2010) Single molecule sequencing with a HeliScope genetic analysis system. Curr Protoc Mol Biol Chapter 7(Unit7):10. https://doi.org/10.1002/0471142727.mb0710s92
12. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Walker AS, Stoesser N, On Behalf Of The Rehab C (2019) Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom, 5(9). Doi:https://doi.org/10.1099/mgen.0.000294
13. Lindner R, Friedel CC (2012) A comprehensive evaluation of alignment algorithms in the context of RNA-seq. PLoS ONE 7(12):e52403. https://doi.org/10.1371/journal.pone.0052403
14. Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, Gibrat JF (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J Comput Biol 19(6):796–813. https://doi.org/10.1089/cmb.2012.0022
15. Girotto S, Comin M, Pizzi C (2018) Efficient computation of spaced seed hashing with block indexing. BMC Bioinform 19(Suppl 15):441. https://doi.org/10.1186/s12859-018-2415-8
16. Baichoo S, Ouzounis CA (2017) Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. Biosystems 156–157:72–85. https://doi.org/10.1016/j.biosystems.2017.03.003
17. Marco-Sola S, Ribeca P (2015) Efficient alignment of illumina-like high-throughput sequencing reads with the GEnomic Multi-tool (GEM) Mapper. Curr Protoc Bioinform 50:11–13. https://doi.org/10.1002/0471250953.bi1113s50
18. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36
19. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324
20. Bhagwat M, Young L, Robison RR (2012) Using BLAT to find sequence similarity in closely related genomes. Curr Protoc Bioinform Chapter 10(Unit10):18. https://doi.org/10.1002/0471250953.bi1008s37
21. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. https://doi.org/10.1038/nmeth.1923
22. Callari M, Sammut SJ, De Mattos-Arruda L, Bruna A, Rueda OM, Chin SF, Caldas C (2017) Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. Genome Med 9(1):35. https://doi.org/10.1186/s13073-017-0425-1
23. Kumar S, Agarwal S, Ranvijay (2019) Fast and memory efficient approach for mapping NGS reads to a reference genome. J Bioinform Comput Biol 17(2):1950008. https://doi.org/10.1142/S0219720019500082
24. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM (2020) Weighted minimizer sampling improves long read mapping. Bioinformatics 36(Supplement_1):i111–i118. https://doi.org/10.1093/bioinformatics/btaa435
25. Grytten I, Rand KD, Nederbragt AJ, Sandve GK (2020) Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. BMC Genom 21(1):282. https://doi.org/10.1186/s12864-020-6685-y
26. Schilbert HM, Rempel A, Pucker B (2020) Comparison of read mapping and variant calling tools for the analysis of plant NGS data. Plants (Basel). https://doi.org/10.3390/plants9040439
27. Thankaswamy-Kosalai S, Sen P, Nookaew I (2017) Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics 109(3–4):186–191. https://doi.org/10.1016/j.ygeno.2017.03.001
28. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5):473–483. https://doi.org/10.1093/bib/bbq015
29. Zhao Y, Wang X, Tang H (2018) A secure alignment algorithm for mapping short reads to human genome. J Comput Biol 25(6):529–540. https://doi.org/10.1089/cmb.2017.0094
30. Wilson-Sanchez D, Lup SD, Sarmiento-Manus R, Ponce MR, Micol JL (2019) Next-generation forward genetic screens: using simulated data to improve the design of mapping-by-sequencing experiments in Arabidopsis. Nucleic Acids Res 47(21):e140. https://doi.org/10.1093/nar/gkz806
31. Smith HE, Yun S (2017) Evaluating alignment and variant-calling software for mutation identification in C. elegans by whole-genome sequencing. PLoS ONE 12(3):e0174446. https://doi.org/10.1371/journal.pone.0174446
32. Houtgast EJ, Sima VM, Bertels K, Al-Ars Z (2018) Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths. Comput Biol Chem 75:54–64. https://doi.org/10.1016/j.compbiolchem.2018.03.024
33. Donato L, D’Angelo R, Alibrandi S, Rinaldi C, Sidoti A, Scimone C (2020) Effects of A2E-induced oxidative stress on retinal epithelial cells: new insights on differential gene response and retinal dystrophies. Antioxidants (Basel). https://doi.org/10.3390/antiox9040307
34. Donato L, Scimone C, Alibrandi S, Nicocia G, Rinaldi C, Sidoti A, D’Angelo R (2020) Discovery of GLO1 new related genes and pathways by RNA-Seq on A2E-stressed retinal epithelial cells could improve knowledge on retinitis pigmentosa. Antioxidants (Basel). https://doi.org/10.3390/antiox9050416

35. Donato L, Scimone C, Alibrandi S, Rinaldi C, Sidoti A, D'Angelo R (2020) Transcriptome analyses of lncRNAs in A2E-stressed retinal epithelial cells unveil advanced links between metabolic impairments related to oxidative stress and retinitis pigmentosa. Antioxidants (Basel). https://doi.org/10.3390/antiox9040318
36. Donato L, Scimone C, Alibrandi S, Abdalla EM, Nabil KM, D'Angelo R, Sidoti A (2020) New omics-derived perspectives on retinal dystrophies: could ion channels-encoding or related genes act as modifier of pathological phenotype? Int J Mol Sci. https://doi.org/10.3390/ijms22010070
37. Mo L, Shi J, Guo X, Zeng Z, Hu N, Sun J, Wu M, Zhou H, Hu Y (2018) Molecular characterization and phylogenetic analysis of a dengue virus serotype 3 isolated from a Chinese traveler returned from Laos. Virol J 15(1):113. https://doi.org/10.1186/s12985-018-1016-5
38. Donato L, Scimone C, Alibrandi S, Pitruzzella A, Scalia F, D'Angelo R, Sidoti A (2020) Possible A2E Mutagenic Effects on RPE Mitochondrial DNA from Innovative RNA-Seq Bioinformatics Pipeline. Antioxidants (Basel). https://doi.org/10.3390/antiox9111158
39. Marco-Sola S, Sammeth M, Guigo R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9(12):1185–1188. https://doi.org/10.1038/nmeth.2221
40. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360. https://doi.org/10.1038/nmeth.3317
41. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL (2018) Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. BioRxiv. https://doi.org/10.1101/390013
42. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191
43. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA, Crook DW, Walker AS (2020) Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience. https://doi.org/10.1093/gigascience/giaa007
44. Siragusa E (2015) Approximate string matching for high-throughput sequencing. Free University of Berlin. https://doi.org/10.17169/refubium-15562
45. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27(18):2518–2528. https://doi.org/10.1093/bioinformatics/btr427
46. Otto C, Stadler PF, Hoffmann S (2014) Lacking alignments? The next-generation sequencing mapper segemehl revisited. Bioinformatics 30(13):1837–1843. https://doi.org/10.1093/bioinformatics/btu146
47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635
48. Liao Y, Smyth GK, Shi W (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz114
49. Scimone C, Alibrandi S, Scalinci SZ, Trovato Battagliola E, D'Angelo R, Sidoti A, Donato L (2020) Expression of pro-angiogenic markers is enhanced by blue light in human RPE cells. Antioxidants (Basel). https://doi.org/10.3390/antiox9111154
50. Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28(4):593–594. https://doi.org/10.1093/bioinformatics/btr708
51. Holtgrewe M (2019) Mason—a read simulator for second generation sequencing data. Institut für Mathematik und Informatik, Freie Universität Berlin
52. Caboche S, Audebert C, Lemoine Y, Hot D (2014) Comparison of mapping algorithms used in high-throughput sequencing: application to Ion torrent data. BMC Genom 15:264. https://doi.org/10.1186/1471-2164-15-264
53. Hatem A, Bozdag D, Toland AE, Catalyurek UV (2013) Benchmarking short sequence mapping tools. BMC Bioinform 14:184. https://doi.org/10.1186/1471-2105-14-184
54. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27(12):2050–2060. https://doi.org/10.1101/gr.222109.117
55. Rausch T, Hsi-Yang Fritz M, Korbel JO, Benes V (2019) Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics 35(14):2489–2491. https://doi.org/10.1093/bioinformatics/bty1007
56. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18(11):1851–1858. https://doi.org/10.1101/gr.078212.108
57. Lim JQ, Tennakoon C, Guan P, Sung WK (2015) BatAlign: an incremental method for accurate alignment of sequencing reads. Nucleic Acids Res 43(16):e107. https://doi.org/10.1093/nar/gkv533
58. Bronner IF, Quail MA (2019) Best practices for illumina library preparation. Curr Protoc Hum Genet 102(1):e86. https://doi.org/10.1002/cphg.86
59. Pervez MT, Babar ME, Nadeem A, Aslam M, Awan AR, Aslam N, Hussain T, Naveed N, Qadri S, Waheed U, Shoaib M (2014) Evaluating the accuracy and efficiency of multiple sequence alignment methods. Evol Bioinform Online 10:205–217. https://doi.org/10.4137/EBO.S19199
60. Brown J, Pirrung M, McCue LA (2017) FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics 33(19):3137–3139. https://doi.org/10.1093/bioinformatics/btx373
61. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O,

Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59. https://doi.org/10.1038/nature07517

62. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A, Hansen LH (2020) GC bias affects genomic and metagenomic reconstructions underrepresenting GC-poor organisms. Gigascience. https://doi.org/10.1093/gigascience/giaa008
63. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16):e105. https://doi.org/10.1093/nar/gkn425
64. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5(2):183–188. https://doi.org/10.1038/nmeth.1179
65. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6(4):291–295. https://doi.org/10.1038/nmeth.1311
66. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5(12):1005–1010. https://doi.org/10.1038/nmeth.1270
67. Institute B Picard Tools. http://broadinstitute.github.io/picard/. Accessed 25 February 2021 2021
68. Fu Y, Wu PH, Beane T, Zamore PD, Weng Z (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genom 19(1):531. https://doi.org/10.1186/s12864-018-4933-1
69. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I (2016) The impact of amplification on differential expression analyses by RNA-seq. Sci Rep 6:25533. https://doi.org/10.1038/srep25533
70. Schroeder CM, Hilke FJ, Loffler MW, Bitzer M, Lenz F, Sturm M (2017) A comprehensive quality control workflow for paired tumor-normal NGS experiments. Bioinformatics 33(11):1721–1722. https://doi.org/10.1093/bioinformatics/btx032
71. Wordsworth S, Doble B, Payne K, Buchanan J, Marshall DA, McCabe C, Regier DA (2018) Using "Big Data" in the cost-effectiveness analysis of next-generation sequencing technologies: challenges and potential solutions. Value Health 21(9):1048–1053. https://doi.org/10.1016/j.jval.2018.06.016
72. Canzar S, Salzberg SL (2017) Short read mapping: an algorithmic tour. Proc IEEE Inst Electr Electron Eng 105(3):436–458. https://doi.org/10.1109/JPROC.2015.2455551
73. Krizanovic K, Echchiki A, Roux J, Sikic M (2018) Evaluation of tools for long read RNA-seq splice-aware alignment. Bioinformatics 34(5):748–754. https://doi.org/10.1093/bioinformatics/btx668
74. Giese SH, Zickmann F, Renard BY (2014) Specificity control for read alignments using an artificial reference genome-guided false discovery rate. Bioinformatics 30(1):9–16. https://doi.org/10.1093/bioinformatics/btt255
75. Holtgrewe M, Emde AK, Weese D, Reinert K (2011) A novel and well-defined benchmarking method for second generation read mapping. BMC Bioinformatics 12:210. https://doi.org/10.1186/1471-2105-12-210
76. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14(2):135–139. https://doi.org/10.1038/nmeth.4106
77. Xin H, Nahar S, Zhu R, Emmons J, Pekhimenko G, Kingsford C, Alkan C, Mutlu O (2016) Optimal seed solver: optimizing seed selection in read mapping. Bioinformatics 32(11):1632–1642. https://doi.org/10.1093/bioinformatics/btv670
78. Steglich M, Nubel U (2017) The challenge of detecting indels in bacterial genomes from short-read sequencing data. J Biotechnol 250:11–15. https://doi.org/10.1016/j.jbiotec.2017.02.026
79. Lee D, Hormozdiari F, Xin H, Hach F, Mutlu O, Alkan C (2015) Fast and accurate mapping of Complete Genomics reads. Methods 79–80:3–10. https://doi.org/10.1016/j.ymeth.2014.10.012
80. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352
81. Landman SR, Hwang TH, Silverstein KA, Li Y, Dehm SM, Steinbach M, Kumar V (2014) SHEAR: sample heterogeneity estimation and assembly by reference. BMC Genomics 15:84. https://doi.org/10.1186/1471-2164-15-84
82. Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B (2015) High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet 16:38–47. https://doi.org/10.1016/j.fsigen.2014.11.022
83. Bansal V (2017) A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC Bioinform 18(Suppl 3):43. https://doi.org/10.1186/s12859-017-1471-9
84. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40(10):e72. https://doi.org/10.1093/nar/gks001
85. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505(7484):495–501. https://doi.org/10.1038/nature12912
86. Gupta G, Saini S (2020) DAVI: Deep learning-based tool for alignment and single nucleotide variant identification. Mach Learn: Sci Technol 1(2):025013. https://doi.org/10.1088/2632-2153/ab7e19
87. Seo S, Oh M, Park Y, Kim S (2018) DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics 34(13):i254–i262. https://doi.org/10.1093/bioinformatics/bty275
88. Ramakrishnan RK, Singh J, Blanchette M (2018) RLALIGN: A reinforcement learning approach for multiple sequence alignment. In: 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 29–31 Oct. 2018. pp. 61–66. doi:https://doi.org/10.1109/BIBE.2018.00019
89. Hou J, Adhikari B, Cheng J (2018) DeepSF: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics 34(8):1295–1303. https://doi.org/10.1093/bioinformatics/btx780
90. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, Morton L, Jarman RG (2020) Next generation sequencing and bioinformatics methodologies

for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity. J Infect Dis 221(Suppl 3):S292–S307. https://doi.org/10.1093/infdis/jiz286

91. Charre C, Ginevra C, Sabatier M, Regue H, Destras G, Brun S, Burfin G, Scholtes C, Morfin F, Valette M, Lina B, Bal A, Josset L (2020) Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. Virus Evol 6(2):veaa075. https://doi.org/10.1093/ve/veaa075
92. Chen S, He C, Li Y, Li Z, Melancon CE (2020) A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data. Brief Bioinform. https://doi.org/10.1093/bib/bbaa231
93. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang Z (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol 92(6):667–674. https://doi.org/10.1002/jmv.25762
94. Control ECfDPa (2021) Sequencing of SARS-CoV-2: first update. ECDC, Stockholm
95. Zhang YZ, Holmes EC (2020) A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell 181(2):223–227. https://doi.org/10.1016/j.cell.2020.03.035
96. Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, Abel HJ, Pfeifer JD (2011) Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn 13(3):325–333. https://doi.org/10.1016/j.jmoldx.2011.01.006
97. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. Nat Methods 10(10):999–1002. https://doi.org/10.1038/nmeth.2634

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.