Journal of Big Data

RESEARCH Open Access

CaReAl: capturing read alignments in a BAM file rapidly and conveniently

Yoomi Park[1†], Heewon Seo[1,2†], Kyunghun Yoo[1] and Ju Han Kim[1,3*]

*Correspondence: juhan@snu.ac.kr
[†]Yoomi Park and Heewon Seo contributed equally to this work
[1] Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul National University Biomedical Informatics (SNUBI), Seoul, Korea
Full list of author information is available at the end of the article

Abstract

Some of the variants detected by high-throughput sequencing (HTS) are often not reproducible. To minimize the technical-induced artifacts, secondary experimental validation is required but this step is unnecessarily slow and expensive. Thus, developing a rapid and easy to use visualization tool is necessary to systematically review the statuses of sequence read alignments. Here, we developed a high-performance alignment capturing tool, CaReAl, for visualizing the read-alignment status of nucleotide sequences and associated genome features. CaReAl is optimized for the systematic exploration of regions of interest by visualizing full-depth read-alignment statuses in a set of PNG files. CaReAl was 7.5 times faster than IGV 'snapshot', the only stand-alone tool which provides an automated snapshot of sequence reads. This rapid user-programmable capturing tool is useful for obtaining read-level data for evaluating variant calls and detecting technical biases. The multithreading and sequential wide-genome-range-capturing functionalities of CaReAl aid the efficient manual review and evaluation of genome sequence alignments and variant calls. CaReAl is a rapid and convenient tool for capturing aligned reads in BAM. CaReAl facilitates the acquisition of highly curated data for obtaining reliable analytic results.

Keywords: High-throughput sequencing, Data visualization, Variant evaluation

Introduction

The recent rapid evolution of high-throughput sequencing technology has resulted in the generation of huge volumes of data [1, 2]. From the nucleotide sequences revealed by this technology, the order of approximately 3 billion base pairs, variant calls are obtained via the variant calling process by which we identify genomic variants from sequence data [3]. In the field of genomics, a series of DNA 'bases' (adenine, guanine, cytosine, and thymine) make up the nucleic acid sequence. A 'read' means a sequence of base pairs, and we perform genome assembly by taking these small fragments of reads and merging them into a longer DNA sequence. The number of unique sequence reads at a target position determines the depth of coverage. A genomic call is the conclusion of a nucleotide difference from a reference sequence at a given position, typically categorized as substitutions, insertions, and deletions (indels), etc. that describe different combinations of DNA gains, losses, or rearrangements. To obtain high-accuracy genome calls, various kinds of alignment and variant calling methods have been developed. The goal of

Springer Open

© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

read-alignment is to map the vast quantities of short sequence fragments to a common reference genome to identify the correct genomic loci [4], where the reads are not long enough to sequence complete transcripts due to technical limitations. To overcome these problems, various strategies of different sequencing platforms are used, which creates variations in the sequence [5]. For example, the Ion Proton and Illumina devices, the two most widely known sequencing platforms, use different methods to capture the signal of variant calls: the mass of hydrogen ions and fluorescence intensity, respectively. These differences could create technology-induced artifacts potentially specific to the sequencing platform used [5–9]. One of the highest sequencing errors produced by Ion Proton is the homopolymer insertion/deletion, which derives from the nature of this technology to capture variant calls through changes in the pH of the solution. It is still challenging to clearly distinguish those technical errors from massive sequence data obtained using different sequencing platforms and experimental conditions. The most powerful way to overcome the challenges posed by technical biases is to experimentally measure the validity of variant calls. However, the experimental validation step is slow and impractically expensive, thus a comprehensive and systematic fashion of implementation has been required instead. One of the simplest ways to obtain high-quality sequencing information is to manually review aligned sequence assemblies with the aid of visualization techniques so that the pattern of technology-induced errors can be detected in advance [6]. Various visualization tools have been developed for investigating the read-alignment status, such as the Integrative Genomics Viewer (IGV) [10], GBrowse [11], Tablet [12], BamView [13], Savant [14], and Artemis [15], which support interactive explorations of multiple types of genome features, such as transcripts, exons, and genes. However, manually querying every single position in multiple regions is laborious. IGV is the only tool that provides an auxiliary tool for taking serial snapshots as a batch job, but this tool is not optimized for automated capturing since the java-based platform causes slow system performance and it does not show inserted bases or the full depth of reads by dynamically adjusting track height depending on the regions of interest. As some of the reads may have read-alignment patterns that are significantly different from the others, clearly visualizing the total depth of reads aligned at a given reference base position, including inserted bases, is important to detect platform-specific error patterns during the manual scanning process. In this study, we developed a rapid and convenient Capturing Read Alignments (CaReAl) tool for the efficient handling of heterogeneous genome sequence data sets. As a complementary tool to IGV 'snapshot', CaReAl supports the rapid and full-depth capture of wide-ranging genome locations in multiple samples, as well as displaying inserted bases. CaReAl focuses on the systematic exploration of the overall read-alignment status to minimize sequencing bias induced by technical errors, rather than on interactive searches for genome features. CaReAl-based research analyses of highly curated data facilitate the comprehension of the detailed alignment status of genome sequences.

Implementation

CaReAl is implemented in Python and R, and it supports the BAM (Binary Alignment Map) and VCF (Variant Call Format) data formats. BAM files should be sorted into karyotypic order and indexed. Using SAMTools [16], CaReAl retrieves the reads/sequences

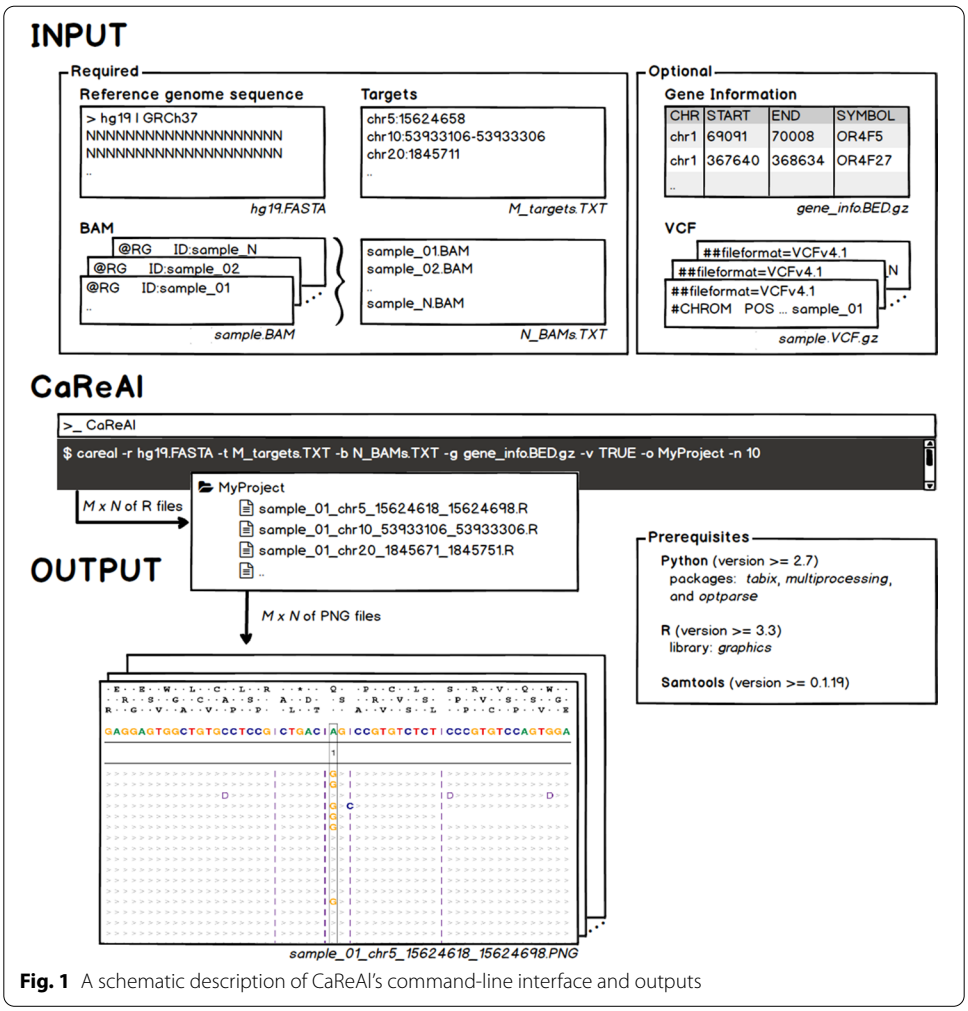

Fig. 1 A schematic description of CaReAl's command-line interface and outputs

in a specific region from a given BAM and a reference genome sequence. The window size is 81 bp (40 bp on either side of a single target variant) by default when a single position is submitted, but the range threshold can be flexibly resized based on the user's region of interest. The maximum supportable window size is 3000 bp, depending on the hardware specification. Tabix [17] is required to extract called variant information that overlaps specified regions using an indexing technique. Both image files in Portable Network Graphics (PNG) format with a resolution of 200 dpi and amendable R scripts are created in the results directory. CaReAl supports any type of next-generation sequencing data, such as whole-genome sequencing, whole-exome sequencing, or targeted sequencing. The time required for screen capture may vary depending on the coverage depth and the window size of the target area that users want to explore. An installation package that contains all of the software required to run CaReAl is provided. The recommended system requirements for CaReAl are as follows: OS (64-bit): CentOS, RAM: 16 gigabytes (GB). A detailed sketch of CaReAl's command line and utility with types of inputs is provided in Fig. 1.

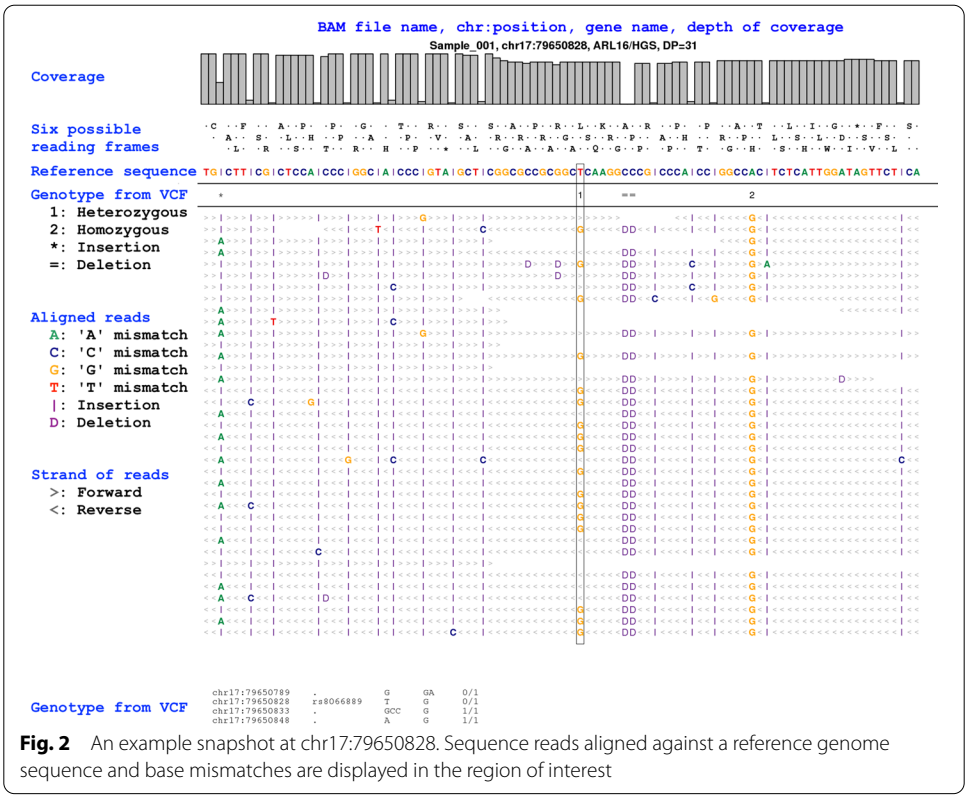

Fig. 2 An example snapshot at chr17:79650828. Sequence reads aligned against a reference genome sequence and base mismatches are displayed in the region of interest

Results

Features

CaReAl displays the full-depth read-alignment status (Fig. 2). An identifier that includes the BAM file name, the chromosomal position of interest, the gene symbol, and the depth of coverage is displayed at the top of the figure. The coverage histogram at the base of each genome position and six possible reading frames of the consensus sequence is provided for detailed background information. Reads and reference sequences overlapping specified regions are visualized in the middle. Bases are colored according to nucleotide type, and gray angle brackets displayed in the background of the sequences indicate the direction of each read strand: '>' for forward and '<' for reverse. The corresponding positions with insertions and deletions are indicated by a purple 'I' and 'D', respectively, and inserted nucleotides are displayed along a straight line. The center of the target position is indicated by the black box. To provide background information for comparing called variant information with displayed signals, called variant genotypes in VCF in the specified region are listed in the top panel as follows: '1' for heterozygous variants, '2' for homozygous variants, '*' for insertion calls, and '=' for deletion calls. Additional detailed variant information in VCF is listed at the bottom.

Performance

To compare the main characteristics of CaReAl with an IGV 'snapshot', the run-time performance was assessed by measuring the time taken to obtain 100 captures of different genome positions with a server equipped with an Intel Xeon 2.0HGz, 256 GB

Table 1 Comparing characteristics between CaReAl and IGV 'snapshot'

Features	CaReAl	IGV 'snapshot'
Language	Python and R	Java
Time[a]	5.26 min/100 imgs (Avg. 3.16 sec/img)	39.47 min/100 imgs (Avg. 23.68 sec/img)
Parallel computing	✓	✗
Display inserted base(s)	✓	✗
Support full-depth	✓	✗

[a] Randomly selected 100 targets in whole-genome sequencing with 40×

of RAM, and integrated graphics chipsets from Matrox Electronics Systems Ltd. (MGA G200e) under CentOS 6.9 (Table 1). We randomly selected 100 coding variants from 5092 positions that exceed 30x at a given locus extracted from a whole-genome sequence generated by Illumina HiSeq2000. As whole-genome sequences require an average of 30× coverage depth for accurate variant detection, targets with greater than 30x were considered to have a sufficient number of sequence reads to perform a test [18, 19]. CaReAl showed extremely good performance, taking 5.26 min without parallel computing (approximately 3 sec per image), compared to an IGV 'snapshot' taking 39.47 min (approximately 23 sec per image), indicating that CaReAl was approximately 7.5 times faster than IGV. By default, CaReAl uses four processor cores in parallel and has shown that this tool outperforms IGV 'snapshot' under a more conservative condition. This implies that this tool is optimized for rapidly capturing useful snapshots of the sequencing reads, while IGV 'snapshot' benefits from the dynamic display of various tracks for an alignment file. One powerful feature of CaReAl is that inserted bases are arranged linearly on images. With IGV 'snapshot', it is not possible to check how many bases and which nucleotide bases are inserted over the reads. Another unique and compelling feature of CaReAl is that captures are displayed with full-depth read alignments, which automatically import the maximum depth of coverage in a given region to adjust the PNG size. In contrast, IGV 'snapshot' displays aligned reads with a fixed coverage depth as specified by the user.

Application

We visualized a variant call identified in *ABI1* using CaReAl (Fig. 3). Since the alternative bases were only identified from PCR duplicates of one unique read, we flagged this variant as being a probable artifact induced from the polymerase process during amplification. To evaluate the accuracy of this call, Fluidigm SNV genotyping assays were carried out, and it turned out to be a false positive. Furthermore, four platform-specific error patterns of variant calls were previously reported by systematically visualizing sequence reads [6]. This highlights the CaReAl's capability which enables the systematic review of the quality of sequence alignment as a pre-evaluation step before the experimental functional assay.

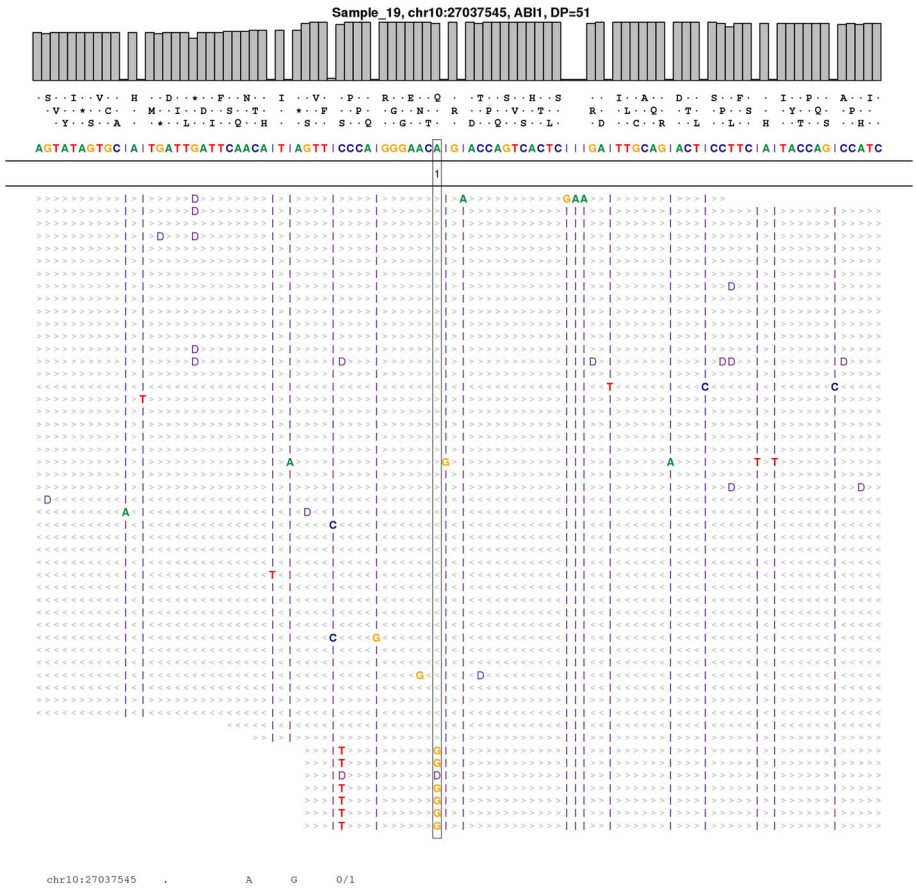

Fig. 3 An example snapshot of probable sequencing artifact at chr10:27037545

Discussion

CaReAl is a high-performance read-alignment capturing tool for systematically displaying genome features. It makes it possible to capture exome variants in a timely manner to detect systematic errors efficiently and to accurately inspect the full depth of the capture. The enormous increase in the amount of sequence data makes an in-depth understanding of the alignment status a prerequisite to obtaining reliable variant calls with high accuracy and validity for downstream analysis. The presence of error profiles specific to sequencing platforms described in the works of Seo et al. and Schirmer et al. [6, 20] makes it necessary to explore read patterns over the full depth of sequencing assemblies. In particular, exploring sequence assemblies not only for calls due to misalignment but also calls derived from a low coverage depth or incorrect enzymatic reactions are also important, because technical artifacts can be eliminated by the systematic scanning of aligned read patterns. Even though various variant calling tools have been developed, detecting reliable indel calls is still challenging [21]. For example, insertion calls in regions of low complexity, such as homopolymer/GC-rich regions, should be double-checked due to the high probability of introducing erroneous signals [22, 23]. Minikel et al. [24] applied this approach to real-world data analysis by manually reviewing the sequence alignment status of every single variant. This process yielded high-confidence

genotype calls on the risk of prion disease, which is required in the field of clinical genomics.

Conclusions

CaReAl is an optimized tool that systematically and rapidly displays temporal summaries of genomic sequences using python and R. This tool has an advantage over IGV 'snapshot' in that it rapidly visualizes the true sequence diversity across the entire depth of coverage, including the insertion bases. This helps to generate high-confidence variant calls in the downstream analysis without the cost and time burden.

Abbreviations

CaReAl: Capturing Read Alignments; IGV: Integrative Genomics Viewer; HTS: High-Throughput Sequencing; PNG: Portable Network Graphics; BAM: Binary-sequence Alignment Format; VCF: Variant Call Format; PCR: Polymerase Chain Reaction.

Acknowledgements

Not applicable.

Authors' contributions

YP and HS conceived of the idea; YP, HS, and KY developed the analytic software; JHK supervised the whole project. All authors contributed to manuscript development. All authors read and approved the final manuscript.

Funding

This research was supported by a Grant (16183MFDS541) from the Ministry of Food and Drug Safety in 2019.

Availability of data and materials

CaReAl is implemented in Python and R and freely available for download (for Linux) from http://www.snubi.org/software/careal/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul National University Biomedical Informatics (SNUBI), Seoul, Korea. [2] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [3] Center for Precision Medicine, Seoul National University Hospital, Seoul, Korea.

Received: 9 November 2020 Accepted: 16 January 2021
Published online: 26 January 2021

References

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97.
2. Churko JM, Mantalas GL, Snyder MP, Wu JC. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circ Res. 2013;112(12):1613–23.
3. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016;107(1):1–8.
4. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol. 2009;27(5):455–7.
5. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol. 2013;9(4):e1003031.
6. Seo H, Park Y, Min BJ, Seo ME, Kim JH. Evaluation of exome variants using the Ion Proton Platform to sequence error-prone regions. PLoS One. 2017;12(7):e0181304.
7. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125.
8. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39(13):e90.
9. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019;20(1):50.
10. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
11. Donlin MJ. Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics. 2009; Chap. 9:Unit 9.

12. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet–next generation sequence assembly visualization. Bioinformatics. 2010;26(3):401–2.
13. Carver T, Bohme U, Otto TD, Parkhill J, Berriman M. BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics. 2010;26(5):676–7.
14. Fiume M, Williams V, Brook A, Brudno M. Savant: genome browser for high-throughput sequencing data. Bioinformatics. 2010;26(16):1938–44.
15. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944–5.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
17. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27(5):718–9.
18. Mauger F, Horgues C, Pierre-Jean M, Oussada N, Mesrob L, Deleuze JF. Comparison of commercially available whole-genome sequencing kits for variant detection in circulating cell-free DNA. Sci Rep. 2020;10(1):6190.
19. Yao RA, Akinrinade O, Chaix M, Mital S. Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients. BMC Med Genomics. 2020;13(1):11.
20. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43(6):e37.
21. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. Hum Genomics. 2015;9:20.
22. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.
23. Luo R, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10(1):998.
24. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med. 2016;8(322):322ra9.

Publisher's note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com