CaReAl: Capturing Read Alignments in a BAM file Rapidly and Conveniently

Yoomi Park
Seoul National University College of Medicine
Heewon Seo
Seoul National University College of Medicine
Kyunghun Yoo
Seoul National University College of Medicine
Ju Han Kim (juhan@snu.ac.kr)
Seoul National University Biomedical Informatics (SNUBI)

Research

Keywords: high-throughput sequencing, data visualization, variant evaluation

Posted Date: November 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-113235/v1

Version of Record: A version of this preprint was published on January 26th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00418-w.

Abstract

Some of the variants detected by high-throughput sequencing (HTS) is 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 ease 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 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 will facilitate the acquisition of highly curated data for obtaining reliable analytic results.

Background

The recent rapid evolution of high-throughput sequencing technology has resulted in the generation of huge volumes of data. Accordingly, various kinds of alignment and variant calling methods have been developed to obtain high-accuracy genome calls. However, there are variations in the sequence profiles obtained using different sequencing platforms [1], indicating that it is still challenging to clearly distinguish 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, 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 refined sequencing information is to manually review aligned sequence assemblies with the aid of visualization techniques [2]. Various visualization tools have been developed for investigating the read-alignment status, such as the Integrative Genomics Viewer (IGV) [3], GBrowse [4], Table [5], BamView [6], Savant [7], and Artemis [8], which support interactive explorations of multiple types of genome features. 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 it is too slow and it does not show inserted bases or the full depth of reads by dynamically adjusting track height depending on the regions of interest. In this study, we developed a rapid and convenient Capturing Read Alignments (CaReAl) tool for the efficient handling of heterogeneous genome sequence data sets. 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 in order 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 that BAM (Binary Alignment Map) and VCF (Variant Call Format) data formats. BAM files should be sorted into karyotypic order and indexed. Using SAMTools [9], CaReAl retrieves the reads/sequences in a specific region from a given BAM and a reference genome sequence. The window size is 81 bp (40 bp 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 3,000 bp, depending on the hardware specification. Tabix [10] 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. 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.

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 are 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 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 5,092 positions that exceed 30x at a given locus extracted from a whole-genome sequence generated by Illumina HiSeq2000. 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. 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.

Table 1
Comparing characteristics between CaReAl and IGV ‘snapshot’.

Features	**CaReAl**	**IGV ‘snapshot’**
Language	Python and R	Java
Time*	5.26 min / 100 imgs	39.47 min / 100 imgs
	(Avg. 3.16 sec / img)	(Avg. 23.68 sec / img)
Parallel computing	✓	
Display inserted base(s)	✓	
Support full-depth	✓	
* Randomly selected 100 targets in whole-genome sequencing with 40x.		

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 [2]. 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.

Discussion And Conclusions

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 for downstream analysis. The presence of error profiles specific to sequencing platforms [2, 11] 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 [12]. 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 [13]. Minikel et al. [14] 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 are required in the field of clinical genomics.

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

Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

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

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 of the authors had approved the final manuscript.

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.

Acknowledgements

Not applicable.

References

1. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. Beerenwinkel N, editor. PLoS Comput Biol. 2013;9:e1003031–18.
2. Seo H, Park Y, Min BJ, Seo ME, Kim JH. Evaluation of exome variants using the Ion Proton Platform to sequence error-prone regions. Bandapalli OR, editor. PLoS ONE. Public Library of Science; 2017;12:e0181304.
3. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. Oxford University Press; 2013;14:178–92.
4. Donlin MJ. Using the Generic Genome Browser (GBrowse). Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2009. pp. 28:9.9:9.9.1–9.9.25.
5. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet–next generation sequence assembly visualization. Bioinformatics. 2010;26:401–2.
6. 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;26:676–7.

7. Fiume M, Williams V, Brook A, Brudno M. Savant: genome browser for high-throughput sequencing data. Bioinformatics. 2010;26:1938–44.
8. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. Oxford University Press; 2000;16:944–5.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
10. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27:718–9.
11. 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 Research. 2015;43:e37–7.
12. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. Human Genomics; 2015;:1–14.
13. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. BioMed Central Ltd; 2013;14:R51.
14. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying prion disease penetrance using large population control cohorts. Science Translational Medicine. 2016;8:–322ra9.

Figures

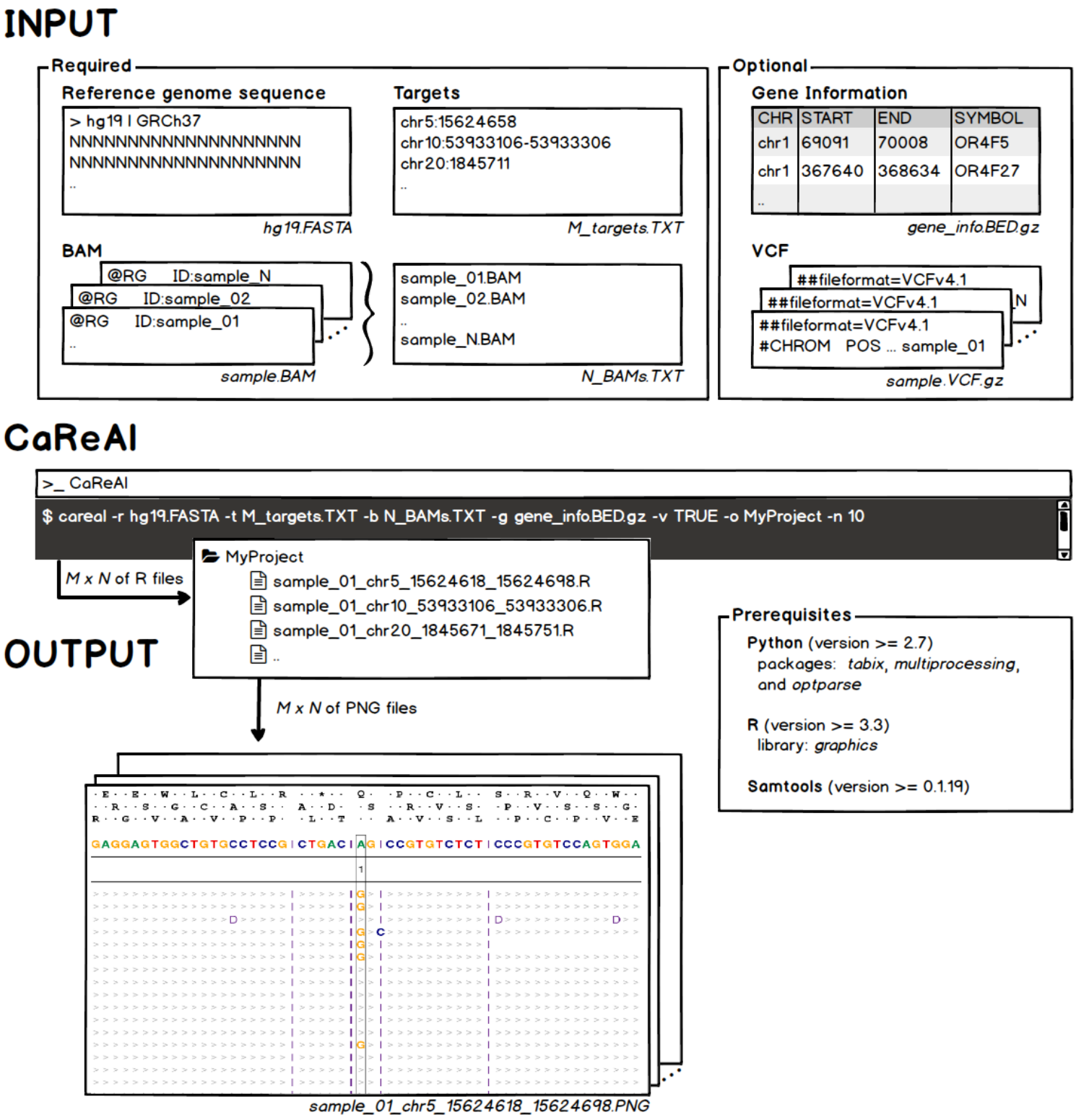

Figure 1

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

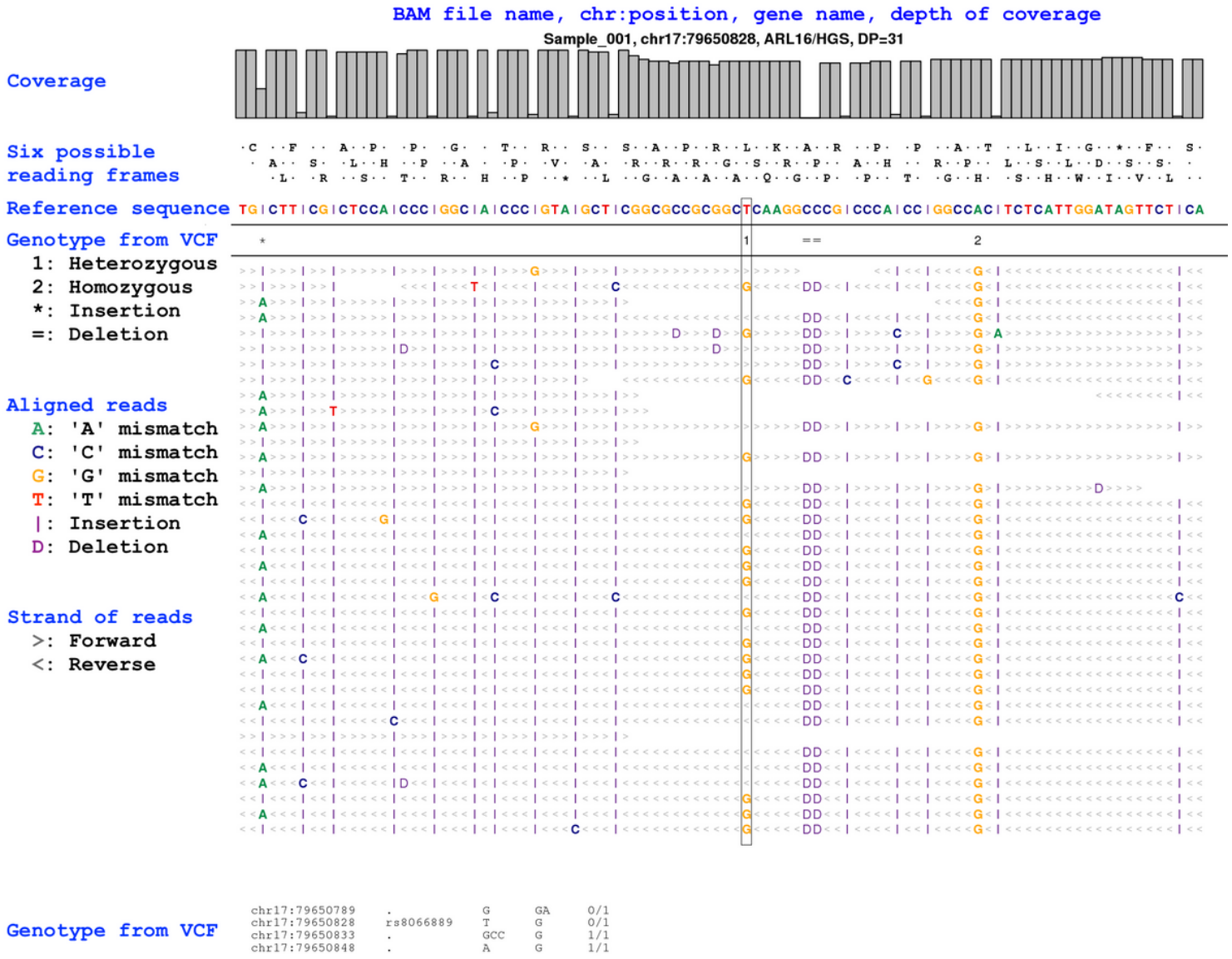

Figure 2

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

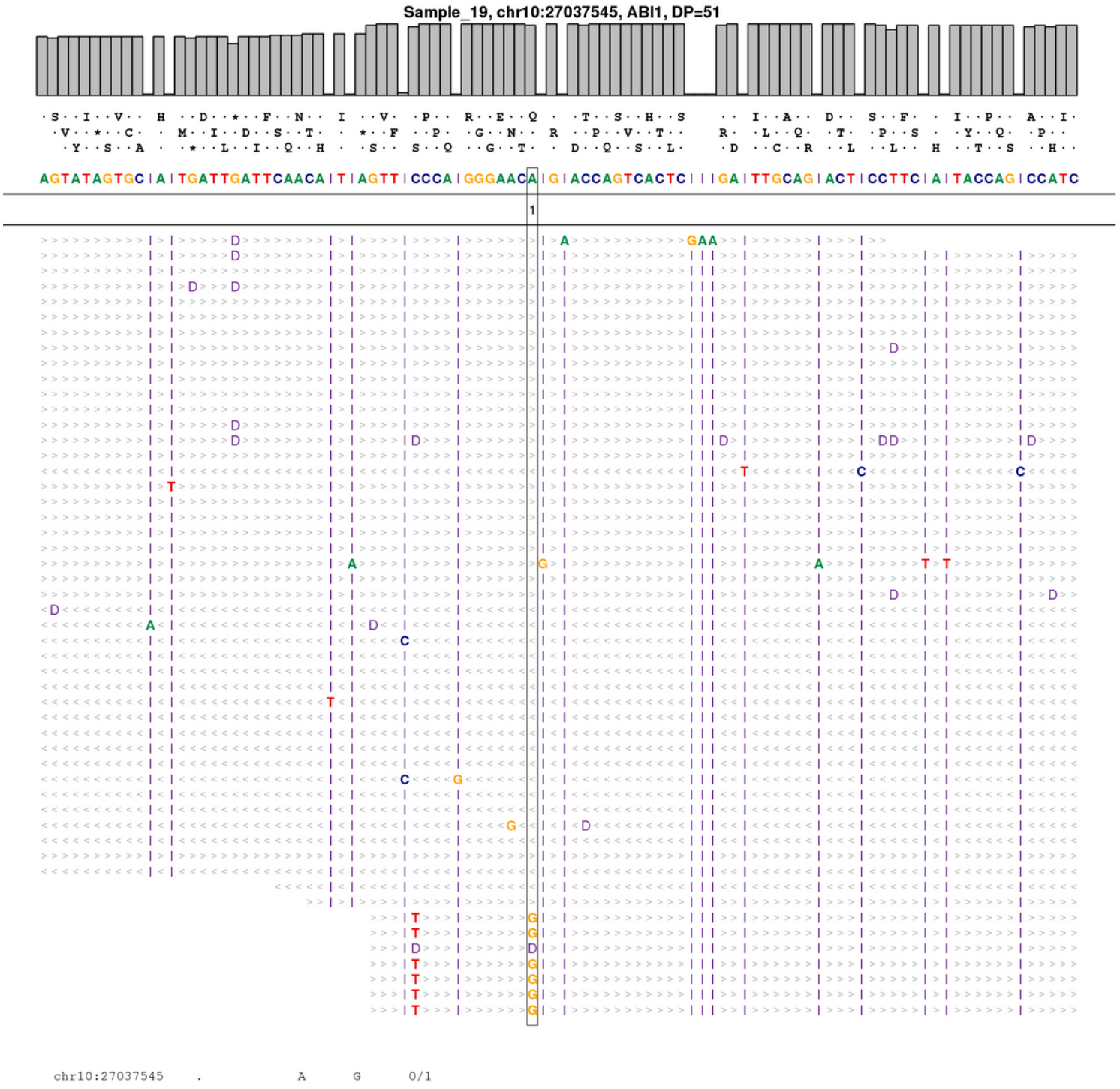

Figure 3

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