MUTATION REPORTER: FAST IDENTIFICATION OF MUTATIONS IN NGS FILES

A PREPRINT

Mikaela Teodoro[1,2], Rosana V. das Chagas[2], José A. Yunes[1,3], Natacha A. Migita[1], and João Meidanis[4]

[1]Boldrini Research Center, Boldrini Children's Hospital, Brazil
[2]Institute of Biology, University of Campinas, Brazil
[3]Medical Genetics Department, Faculty of Medical Sciences, University of Campinas, Brazil
[4]Institute of Computing, University of Campinas, Brazil

May 24, 2024

ABSTRACT

Summary: Next-generation sequencing (NGS) has accelerated precision medicine by enabling analysis of multiple genes at once and detection of low-frequency mutations. Nevertheless, there is a paucity of software applications enabling users not trained in bioinformatics to discern and adjust the quality parameters employed for analyzing sequencing data, which yield the output containing detected mutations alongside their corresponding variant allele frequency (VAF). Mutation Reporter was developed to identify point mutations, offering users the ability to adjust parameters including alignment e-value, minimum read size, minimum read depth, and minimum reported VAF.

Availability and implementation: Mutation Reporter is available to users through a free GNU software license and can be accessed on GitHub (https://github.com/meidanis-lab/mutation-reporter) and as a Code Ocean capsule (https://codeocean.com/capsule/4682823/tree/v1).

Keywords BLAST, NGS, mutations, pipeline

1 Introduction

Next-generation sequencing (NGS) has emerged as a pivotal tool in precision medicine in oncology by enabling analysis of multiple genes at once and facilitating the detection of low-frequency mutations in patients, which may be implicated in treatment resistance, thereby underscoring the clinical significance of NGS in therapeutic decision-making ([5, 6]). However, the scarcity of freely available open-source software tools capable of analyzing sequencing data to identify mutations pose a notable challenge for the biologist, not trained in bioinformatics. For instance, a recently described tool, the RNAMut software, swiftly detects mutations and displays their frequencies, but lacks functionalities that would allow users to comprehend and modify the parameters used to select high-quality reads for analysis, or to specify the minimum variant allele frequency (VAF) for reported mutations ([4]). Furthermore, the absence of disclosure regarding the reference sequence used for comparison is a critical limitation, particularly for subsequent functional investigations of identified mutations.

In pursuit of a more precise analysis of sequencing data, we have developed a user-friendly software tool designed to pinpoint single nucleotide substitutions. This software provides comprehensive information, including the gene harboring the mutation, its position, the altered amino acid, the number of mutated and non-mutated reads, and the mutation VAF. Notably, our software offers users the unique capability to comprehend and adjust the quality parameters governing read selection for analysis, such as minimum read depth, minimum quality and length of alignments, and minimum VAF thresholds for reporting. In addition, users have the ability to provide the protein sequences for comparison.

2 Installation and use

Our software was developed for Unix-like operating systems, including Linux. However, it can easily be installed on Mac computers, which run a Unix-like operating system underneath, and in Windows systems also, by means of the Windows Subsystem for Linux (WSL) platform, a Microsoft product freely available. We summarize here the main steps needed for installation and use. Detailed instructions are provided in the code repository.

Mutation Reporter requires the installation of the Basic Local Alignment Search Tool (BLAST) suite from the National Center for Biotechnology Information (NCBI) ([1]), which is also freely available, and the make command ([3]), another free tool. Both can be easily installed on most Unix-like platforms.

To run the software, one needs to define an **analysis**, which consists of a set of peptide sequences in a fasta file, and the parameters in a text file. We recommend that the description string of each protein identifies the gene it came from, because it will appear in Mutation Reporter's final output. Different protein isoforms with distinct description strings can be used as well. A command in Mutation Reporter builds a BLAST database using the peptide or protein sequences within the fasta file, generating the .pdb, .phr, .pin, .pog, .pos, .pot, .psq, .ptf, and .pto requisite files utilized in the alignment performed by BLAST.

In the next step, users are required to provide the fastq.gz files obtained from sequencing. In the case of paired-end sequencing, the presence of two fastq.gz files with identical names, except for the inclusion of "R1" in the forward read file and "R2" in the reverse read file, is needed. Subsequently, both sets of reads are merged into a single file for further analysis.

3 Methodology and implementation

The first step in Mutation Reporter is to transform the fastq.gz files into fasta files for BLAST use. There is no need to trim or assess sequence quality at this point. Low quality sequences will be automatically discarded since they fail to produce good alignments. After that, BLAST is called to search every read against the user-defined reference database of proteins of interest. We opted to produce a XML file as output, which includes information on alignments, statistics, sequence identity, query, and database name. To process the XML output, we employed the Bio.Blast.Record module from the Biopython package's Bio.Blast module ([8],[2]). The 'parse' function retrieves a 'record' for each sequence in the query file. Each 'record' encompasses one or more alignments, representing matches with a reference sequence. Within each alignment, High-Scoring Segment Pairs (HSPs) denote regions where BLAST conducted local alignment. These HSPs contain crucial data, including identity metrics, e-values, scores, and the start and end positions of aligned amino acids, which are used to generate the mutation report.

First, for each HSP, the software verifies whether alignment length and percent identity are equal or higher than the values predetermined by the user. Upon meeting these criteria, the start and end residues of the hit are recorded, and the software proceeds to identify mutations (represented by ' ' or '+') within the alignment's middle or consensus sequence. For each detected mutation, the script stores the mutation gene, the amino acid in the reference sequence, the mutation's position, the corresponding amino acid in the query sequence and the number of times this mutation occurs.

To determine sequencing depth, the total number of reads at a specific position, necessary for calculating VAF, we use an indirect, but very fast, scheme. As sequencing depth varies along the peptide sequence, the best way to determine it for a given position is to count the number of valid hits that contain this position. However, a direct calculation would take too long. Instead, we record just the start and end residue positions for each valid hit. These numbers are transformed into start counts and end counts, and then into cumulative counts. Given that the total valid hits at position x can be computed as the difference between the number of reads initiating before x and the number of hits terminating before x, subtracting the start and end cumulative counts provides the total number of valid hits at position x. Employing a binary search, we quickly identify the correct start and end cumulative counts regarding x.

Finally, we report mutations with Variant Allele Frequency (VAF) greater than the threshold defined by the user.

To assess this pipeline, we analyzed NGS results from amplicon sequences corresponding to the Ligand Binding Domain (LBD) of RARa, and B1 and B2-box domains of PML within the PML-RARA fusion. These analyses were conducted utilizing bone marrow samples from 32 pediatric patients with Acute Promyelocytic Leukemia, who were treated at the Boldrini Children's Center, under approval by the ethics committee in research at the Boldrini Children's Center (CAAE 38962720.9.0000.5376). For comparative purposes, we subjected identical data files to analysis using RNAMut. As shown in Figure 1, it is evident that the majority of mutations were detected by both platforms, albeit with slight variations in VAFs. Notably, Mutation Reporter identified several mutations that were not reported by RNAMut, including one with a VAF of 99,45% in sample S32. The presence of this mutation and its VAF was further

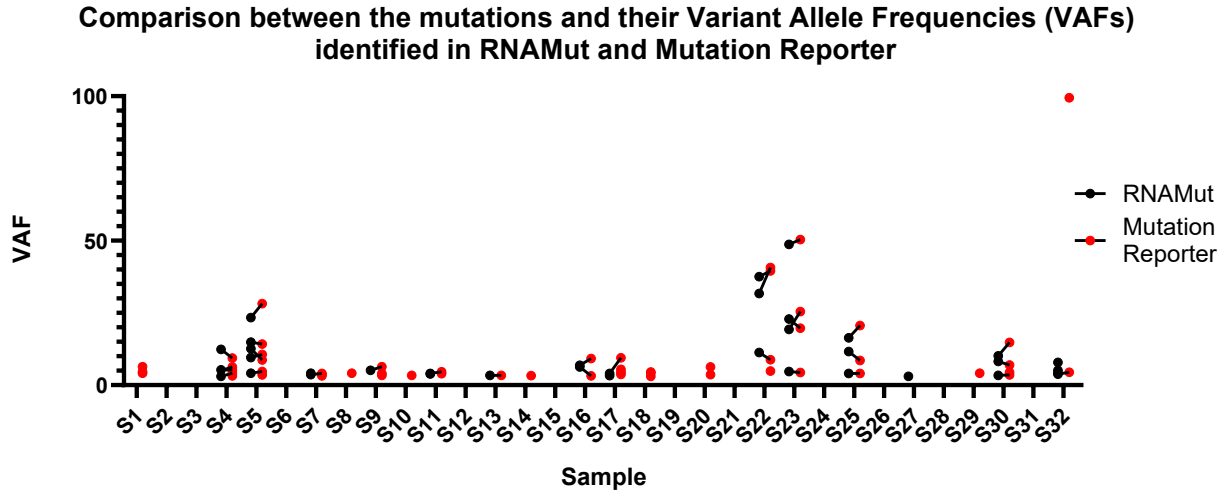

Figure 1: Comparison between mutations and their respective Variant Allele Frequencies (VAFs) identified in the RNAMut software (black) and the Mutation Reporter software (red) across 32 samples of patients diagnosed with Acute Promyelocytic Leukemia from the Boldrini Children's Center. Short straight lines join mutations found by both tools. Only mutations with a VAF higher than 3% are reported.

confirmed through visual inspection of sequence alignment using the Integrative Genomics Viewer (IGV) ([7]). The reason of why it was not detected by RNAMut is a mystery to us.

4 Conclusion

Mutation Reporter successfully achieved the identification of point mutations in NGS files, displaying the mutation, its VAF, and gene, providing users the control over the reference sequence utilized and the quality parameters employed for reporting mutations (e.g., e-value, alignment length, percent identity, sequencing depth, and minimum VAF). Our software demonstrated superior performance in comparison to RNAMut. Moreover, despite the absence of a Windows interface, our software proved to be swift, efficient, user-friendly, and conducive to practical implementation in clinical settings for rapid identification of mutations with potential prognostic and therapeutic implications.

5 Competing interests

No competing interest is declared.

6 Author contributions statement

MT implemented the final version of the software, wrote the manual, and conducted the experiments, RVC. implemented an initial version of the software, JAY. supervised the NGS of PML-RARA and reviewed the manuscript, NAM. conceived the idea for the software and reviewed the manuscript, JM supervised the implementations, helped with the writing of both the manual and the article, and organized the code repository.

7 Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by São Paulo Research Foundation [grant number 2022/06635-7]; and the National Council for Scientific and Technological Development [grant number 308399/2021-8 to J.A.Y].

References

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[3] S. I. Feldman. Make — a program for maintaining computer programs. *Software: Practice and Experience*, 9:255–265, 1978.

[4] Muxin Gu, Maximillian Zwiebel, Swee Hoe Ong, Nick Boughton, Josep Nomdedeu, Faisal Basheer, Yasuhito Nannya, Pedro M Quiros, Seishi Ogawa, Mario Cazzola, et al. RNAmut: robust identification of somatic mutations in acute myeloid leukemia using RNA-sequencing. *Haematologica*, 105(6):e290, 2020.

[5] Margaret Morash, Hannah Mitchell, Himisha Beltran, Olivier Elemento, and Jyotishman Pathak. The role of next-generation sequencing in precision medicine: A review of outcomes in oncology. *Journal of personalized medicine*, 8(3):30, 2018.

[6] Michael W. Schmitt, Scott R. Kennedy, Jesse J. Salk, Edward J. Fox, Joseph B. Hiatt, and Lawrence A. Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 109(36):14508–14513, 2012.

[7] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2012.

[8] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.