МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ, БИОИНФОРМАТИКА / MATHEMATICAL BIOLOGY, BIOINFORMATICS

DOI: https://doi.org/10.18454/jbg.2023.21.2

ALIGNMENT OF PSEUDOREADS OBTAINED FROM HOMOLOGOUS SEQUENCES IN IDENTIFYING POTENTIALLY TOLERATED GENOMIC VARIANTS

Research article

Bug D.S.[1, *], Narkevich A.N.[2], Petukhova N.V.[3]
[1] ORCID : 0000-0002-5849-1311;
[2] ORCID : 0000-0002-1489-5058;
[3] ORCID : 0000-0001-6397-824X;
[1, 3] Pavlov First Saint Petersburg Medical State University, Saint-Petersburg, Russian Federation
[2] South Ural State Medical University, Chelyabinsk, Russian Federation

* Corresponding author (bug.dmitrii[at]yandex.ru)

Abstract

The use of Next-Generation Sequencing (NGS) has proven to be clinically beneficial, but it has also revealed a significant number of variants that we are unable to accurately define and categorize in terms of pathogenicity. These variants are known as variants of uncertain significance (VUS) which are detected en masse in each NGS run. Unlike amino acid substitutions and splice site mutations, common variants in non-coding regions have not been extensively studied and are still mostly classified as VUS. In this paper, a new concept was proposed to identify potentially tolerated variants, including variants in non-coding regions, based on the Genetic Alignment of “Pseudoreads” from Homologs (GAPH) method. We have discovered a total of 5,859,205 variants, the majority of which have never been documented in the largest population database, GnomAD, and only 0.0015% (88 variants) were classified as pathogenic according to the ClinVar database. Overall, the results of this study demonstrate the efficacy of our new method to refine a variant tolerability, many aspects of which could be further adjusted to optimize the results.

Keywords: genetics, variants of uncertain significance, variant effect prediction, non-coding, benign variant refinement.

ВЫРАВНИВАНИЕ ПСЕВДОРИДОВ ГОМОЛОГИЧНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В ОПРЕДЕЛЕНИИ ПОТЕНЦИАЛЬНО ПЕРЕНОСИМЫХ ГЕНОМНЫХ ВАРИАНТОВ

Научная статья

Буг Д.С.[1, *], Наркевич А.Н.[2], Петухова Н.В.[3]
[1] ORCID : 0000-0002-5849-1311;
[2] ORCID : 0000-0002-1489-5058;
[3] ORCID : 0000-0001-6397-824X;
[1, 3] Первый Санкт-Петербургский Государственный Медицинский Университет им. акад. И.П. Павлова, Санкт-Петербург, Российская Федерация
[2] Южно-Уральский государственный медицинский университет, Челябинск, Российская Федерация

* Корреспондирующий автор (bug.dmitrii[at]yandex.ru)

Аннотация

Секвенирование следующего поколения (NGS) доказало свою пользу в клинической практике, однако, значительное количество генетических вариантов, которые обнаруживаются при помощи данного метода, невозможно точно определить и классифицировать с точки зрения их патогенности. Эти варианты известны как варианты неопределённого значения (VUS), и они массово обнаруживаются при каждом запуске NGS. В отличие от аминокислотных замен и мутаций сайтов сплайсинга, обычные варианты в некодирующих областях не изучались столь широко и по-прежнему в основном классифицируются как VUS. В данной статье для идентификации потенциально переносимых вариантов, включая варианты в некодирующих областях, предлагается новая концепция, основанная на методе генетического выравнивания “псевдоридов” гомологичных последовательностей (GAPH). Мы обнаружили в общей сложности 5 859 205 вариантов, большинство из которых не задокументированы в крупнейшей популяционной базе данных GnomAD, и только 0,0015% (88 вариантов) были классифицированы как патогенные в соответствии с базой данных ClinVar. В целом, результаты этого исследования демонстрируют эффективность нашего нового метода для улучшения оценки переносимости варианта, при этом многие аспекты метода могут быть дополнительно скорректированы для оптимизации результатов.

Ключевые слова: генетика, варианты неопределённого значения, предсказание эффекта вариантов, некодирующие варианты, оценка доброкачественности варианта.

Introduction

Molecular diagnostics has undergone a stage of rapid development and growth in the last decade, intensively increasing the amount of genomic data [1], [2]. However, the clinical significance of the majority of variants being detected is unknown. Variants in non-coding regions, which account for about 98% of the genome, are mostly classified as variants of uncertain significance (VUS) [3]. There is increasing evidence that non-coding variants in different diseases, including cancer, are becoming more recognized as contributors to the development of these diseases [4], [5], [6], [7].

Computational variant effect predictors leverage the vast amount of biological data currently available to infer the fitness effects of human variants. The American College of Medical Genetics and Genomics (ACMG) previously developed the guidance for interpreting sequence variants, which involves utilizing *in silico* predictors to classify neutral (benign, BP4 criterion) and pathogenic (pathogenic, PP3 criterion) variants [10].

The majority of widely used and well-established computational predictors were designed to analyze protein sequences or splicing sites and, therefore, cannot be applied to variants in non-coding regions. Nevertheless, there are multiple tools that predict the deleteriousness of non-coding region variants based on the evolutionary conservation of studied genomic fragments [11], [12], [13], [14]. Other variant effect predictors, which can analyze variants in non-coding regions, are based on the annotation data [15], [16], [17]. However, none of these existing tools are capable of taking into account the specific sequence changes observed during evolution.

For such a reason, we have developed a new method, Genetic Alignment of Pseudoreads from Homologs (GAPH), that identifies differences between human genes and genomic sequences of various species. We have also evaluated whether this approach is appropriate and relevant enough to use it for variant tolerance prediction.

Research methods and principles

263 cancer-related genes were obtained from the COSMIC database [18], and the longest protein isoform was derived for each gene from the NCBI RefSeq database [19]. These amino acid sequences were then queried against the RefSeq protein database using BLAST with an *E-value* threshold of 10 and a maximum number of resulting sequences set to 1000, and the corresponding nucleotide sequences were obtained.

The resulting gene sequences were fragmented into 70 nucleotides pseudoreads with a 35-nucleotide overlap. Each nucleotide was assigned a *Phred-score* of 30. The resulting pseudoreads were aligned to the originally queried human gene sequence using BWA-MEM with default parameters: according to the original BWA paper, only reads with no more than 4 mismatches or gaps were mapped to the reference sequence [20]. Finally, variants were called using VarScan v2.4.4 [21] with at least two reads support and a minimum of eight read-depth coverage required.

GAPH-resulted variant set was compared with known benign and pathogenic variants from ClinVar (accessed on July 14, 2022) [22], as well as variants from GnomAD v3.1.2 (accessed on June 3, 2023) [23].

Since benign variants tend to have a relatively high population allele frequency, we have used a subset of the GnomAD v3.1.2 database along with its full version, which included variants with population allele frequency > 0.01% only.

Main results

Using GAPH, we identified 5,859,205 nucleotide variants that are present in homologous sequences and therefore might be evolutionarily permissible. Of the 10,368,623 variants of the studied genes found in the GnomAD v3.1.2 database, 965,593 (9.3%) were in the resulting set. A subset of the GnomAD database was created, consisting of 569,998 variants with a minor allele frequency (MAF) greater than 0.1%, which accounted for 5.5% of all variants in GnomAD. Out of these, 72,778 variants were found in the GAPH-resulted set, making up 7.5% of GnomAD variants with MAF > 0.1% and 1.2% of the entire GAPH variant set (figure 1).

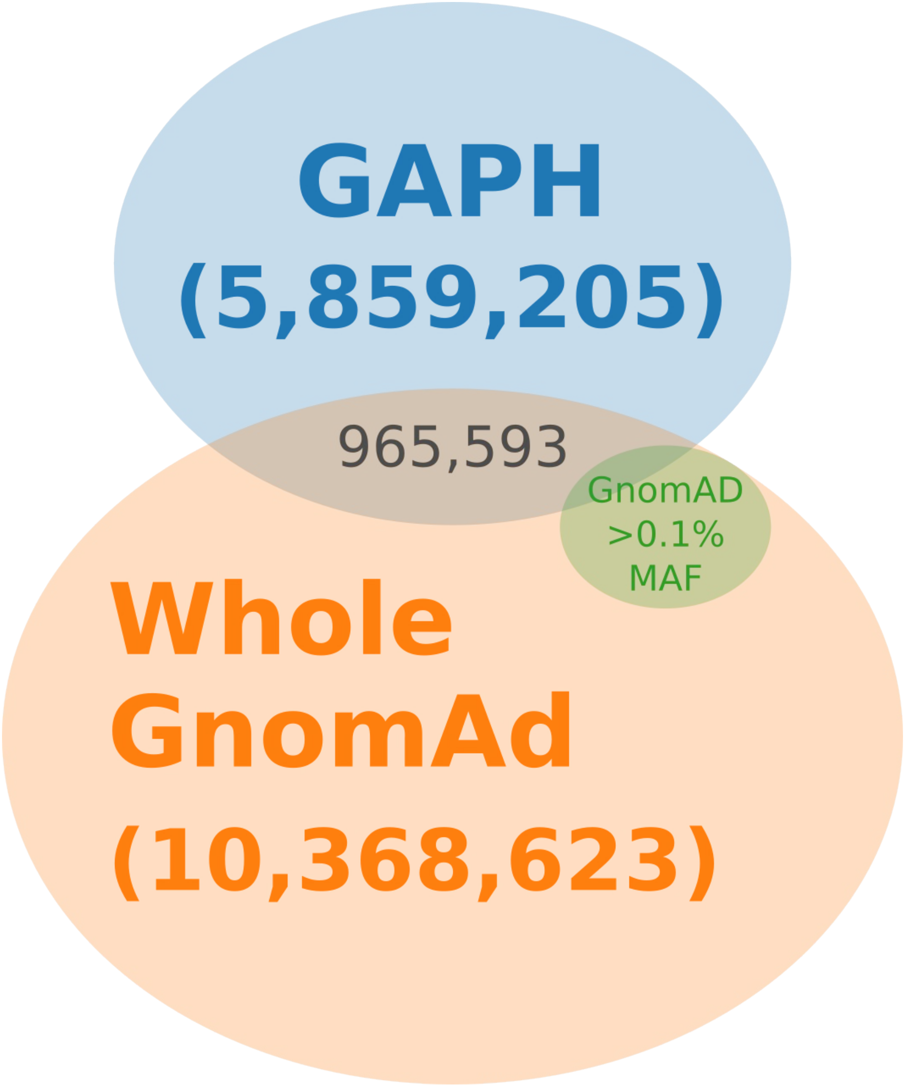

Figure 1 - Venn diagram illustrating the overlap between GAPH-resulted set, variants found in GnomAD v3.1.2, and subset of GnomAD variants with minor allele frequency (MAF) > 0.1%.
DOI: https://doi.org/10.18454/jbg.2023.21.2.1

Among the 18,358 benign variants from the ClinVar database, 6,460 (35%) were in the GAPH-resulted set, and 16,434 (90%) were present among the variants from GnomAD. On the other hand, out of the 27,092 pathogenic variants, only 88 (less than 1%) were present in the obtained set, while 1,180 (4%) were found in the GnomAD database. There were 11,050 benign variants in the studied subset of the GnomAD database, which only included variants with MAF > 0.1%, while no pathogenic variants were found in this group (figure 2).

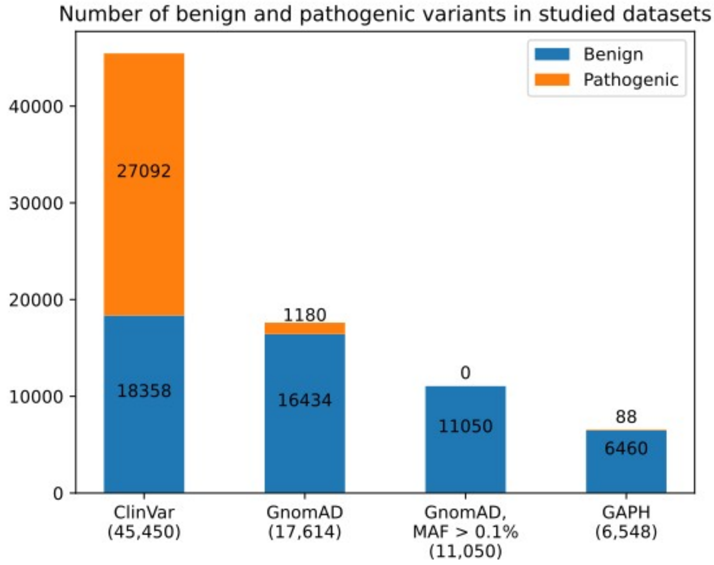

Figure 2 - Distribution of benign and pathogenic variants from ClinVar, GnomAD v3.1.2, GnomAD subset with MAF of variants > 0.1%, and the GAPH-resulted variant set
DOI: https://doi.org/10.18454/jbg.2023.21.2.2

Note: the total number of benign and pathogenic variants is indicated in parentheses

Discussion

In this paper, we have demonstrated GAPH, a method for obtaining a set of variants that are present in homologous sequences from a nucleotide sequence alignment. The core element of this approach is a pseudoreads usage, which is a concept previously applied in transcriptomic studies [24], [25] and genome assembly software [26]. To our knowledge, this is the first study utilizing pseudoreads for variant effect prediction, and that is why this report is focused on presentation of the method concept and its general applicability with no intention to conduct an exhaustive study. Many additional aspects of it might be adjusted further to achieve a desirable level of accuracy, combined with validation and verifications studies.

We used amino acid sequences as queries for the search of homologous sequences, which is a common approach [27], [28]. Despite our relatively high *E-value* threshold of 10, we further implemented the BWA-MEM algorithm, which prevents mapping of 70 nucleotide-long pseudoreads with more than 4 mismatches. However, our method, as well as the current variant effect prediction tools, does not take into account orthologous/paralogous relationships, which reduces the overall performance of categorizing disease-causing and benign variants [29].

For example, a comprehensive phylogenetic study of the *BRCA2*, *TP53*, *DICER1*, *PIK3CA*, and *BRCA1* genes decreased the number of variants obtained through GAPH in these genes from 77,154 to 72,505. It also reduced the number of matches with GnomAD from 12,345 to 11,484, which might aid to decrease the total number of variants erroneously classified as potentially tolerated (unpublished). The precise identification of orthologs for major disease-causing genes might be further introduced in our method.

There are also several widely used nucleotide sequence alignment tools, *e.g.* Bowtie [30] and BWA-SW [31]. While BWA-MEM is the most modern algorithm in the BWA package, which is utilized by popular biological pipelines like GATK [32] and SPAdes [33], other alignment tools exist that could potentially enhance the performance of the method presented here. The operational length of a pseudoread was chosen based on the minimum allowable read length for alignment using BWA-MEM [20] but it also might be adjusted for the better performance.

The main flaw of this method is that identification of damaging variants is not available, similarly to the GnomAD population database [34]. Therefore, our approach is intended to be used similarly to GnomAD in order to identify only potentially tolerated variants. However, there is a significant difference between the GnomAD database and GAPH-resulted variant set: only less than 10% of variants are present in both sets (figure 1). We propose that such a discrepancy is likely due to the disparity in the genomic diversity data used from various sources: while GnomAD is based on the study of the human genome variation, GAPH specifically focuses on sequences of homologous genes in various organisms.

We understand that not all GnomAD variants should be treated as benign, and there is a discrepancy regarding the population allele frequency threshold that should be applied to minimize the probability of a variant pathogenicity [35]. We used a threshold of 0.1% MAF to evaluate the predictive capability of this GnomAD subset and compare it to our dataset.

Overall, the current paper demonstrates the potential of our novel method to generate a database of potentially tolerated variants, which utilizes phylogenetic information unlike all human population databases.

Conclusion

This paper presents a concept of new methodology for obtaining a set of potentially tolerated variants from an alignment of pseudoreads generated from sequences of homologous genes. Our proposed approach has enabled us to identify numerous new variants in the homologs of the genes under investigation. These variants can be used in a similar manner to those found in population databases such as GnomAD. However, while GnomAD relies on genetic variation only in human genes, our method utilizes sequences of homologous genes in different organisms. Since GAPH-resulted set is based on a different principle, it can be used together with GnomAD for variant tolerance prediction.

We are still at the early stages of exploring GAPH, and the potential for adjusting various aspects to enhance its performance is under ongoing study. For example, a comprehensive deep phylogenetic analysis could be applied additionally, a different alignment algorithm could be tested, and the pseudoread length could be further modified to improve performance and applicability.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Cook C. E. The European Bioinformatics Institute in 2016: Data growth and integration / C. E. Cook, M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney, R. Apweiler // Nucleic Acids Research. — 2016. — 44. — p. D20-D26. — URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1352 (accessed: 31.07.2023) DOI: 10.1093/nar/gkv1352.

2. Gagan J. Next-generation Sequencing to Guide Cancer Therapy / J. Gagan, E. M. Van Allen // Genome Medicine. — 2015. — 7. — p. 80. — URL: http://genomemedicine.com/content/7/1/80 (accessed: 31.07.2023) DOI: 10.1186/s13073-015-0203-x.

3. Federici G. Variants of Uncertain Significance in the Era of High-Throughput Genome Sequencing: a Lesson from Breast and Ovary Cancers / G. Federici, S. Soddu // Journal of Experimental & Clinical Cancer Research. — 2020. — 39. — p. 46. — URL: https://jeccr.biomedcentral.com/articles/10.1186/s13046-020-01554-6 (accessed: 31.07.2023) DOI: 10.1186/s13046-020-01554-6.

4. Whiffin N. Characterising the Loss-of-Function Impact of 5’ Untranslated Region Variants in 15,708 Individuals / N. Whiffin, K. J. Karczewski, X. Zhang, S. Chothani, M. J. Smith, D. G. Evans, A. M. Roberts, N. M. Quaife, S. Schafer, O. Rackham, J. Alföldi, A. H. O’Donnell-Luria, L. C. Francioli // Nature Communications. — 2020. — 11. — p. 2523. — URL: https://www.nature.com/articles/s41467-019-10717-9 (accessed: 31.07.2023) DOI: 10.1038/s41467-019-10717-9.

5. Borck G. Father-to-Daughter Transmission of Cornelia de Lange Syndrome Caused by a Mutation in the 5′ Untranslated Region of theNIPBL Gene / G. Borck, M. Zarhrate, C. Cluzeau, E. Bal // Human Mutation. — 2006. — 27. — p. 731-735. — URL: https://onlinelibrary.wiley.com/doi/10.1002/humu.20380 (accessed: 31.07.2023) DOI: 10.1002/humu.20380.

6. Johnston J. J. NAA10 Polyadenylation Signal Variants Cause Syndromic Microphthalmia / J. J. Johnston, K. A. Williamson, C. M. Chou, J. C. Sapp // Journal of Medical Genetics. — 2019. — 56. — p. 444-452. — URL: https://jmg.bmj.com/lookup/doi/10.1136/jmedgenet-2018-105836 (accessed: 31.07.2023) DOI: 10.1136/jmedgenet-2018-105836.

7. Weinhold N. Genome-Wide Analysis of Noncoding Regulatory Mutations in Cancer / N. Weinhold, A. Jacobsen, N. Schultz, C. Sander // Nature Genetics. — 2014. — 46. — p. 1160-1165. — URL: https://www.nature.com/articles/ng.3101 (accessed: 31.07.2023) DOI: 10.1038/ng.3101.

8. Fredriksson N. J. Systematic Analysis of Noncoding Somatic Mutations and Gene Expression Alterations across 14 Tumor Types / N. J. Fredriksson, L. Ny, J. A. Nilsson, E. Larsson // Nature Genetics. — 2014. — 46. — p. 1258-1263. — URL: https://www.nature.com/articles/ng.3141 (accessed: 31.07.2023) DOI: 10.1038/ng.3141.

9. Evans D. G. R. A Dominantly Inherited 5′ UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer / D. G. R. Evans, E. M. Van Veen, H. J. Byers, A. J. Wallace // The American Journal of Human Genetics. — 2016. — 103. — p. 213-220. — URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929718302295 (accessed: 31.07.2023) DOI: 10.1016/j.ajhg.2018.07.002.

10. Richards S. Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology / S. Richards, N. Aziz, S. Bale, D. Bick // Genetics in Medicine. — 2015. — 17. — p. 405-424. — URL: https://linkinghub.elsevier.com/retrieve/pii/S1098360021030318 (accessed: 31.07.2023) DOI: 10.1038/gim.2015.30.

11. Siepel A. Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes / A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs // Genome Research. — 2005. — 15. — p. 1034-1050. — URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.3715005 (accessed: 31.07.2023) DOI: 10.1101/gr.3715005.

12. Pollard K. S. Detection of Nonneutral Substitution Rates on Mammalian Phylogenies / K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel // Genome Research. — 2010. — 20. — p. 110-121. — URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.097857.109 (accessed: 31.07.2023) DOI: 10.1101/gr.097857.109.

13. Cooper G. M. Distribution and Intensity of Constraint in Mammalian Genomic Sequence / G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green // Genome Research. — 2005. — 15. — p. 901-913. — URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.3577405 (accessed: 31.07.2023) DOI: 10.1101/gr.3577405.

14. Garber M. Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns / M. Garber, M. Guttman, M. Clamp, M. C. Zody // Bioinformatics. — 2009. — 25. — p. 54-62. — URL:

https://academic.oup.com/bioinformatics/article/25/12/i54/187307 (accessed: 31.07.2023) DOI: 10.1093/bioinformatics/btp190.

15. Kircher M. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants / M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak // Nature Genetics. — 2014. — 46. — p. 310-315. — URL: https://www.nature.com/articles/ng.2892 (accessed: 31.07.2023) DOI: 10.1038/ng.2892.

16. Quang D. DANN: a Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants / D. Quang, Y. Chen, X. Xie // Bioinformatics. — 2015. — 31. — p. 761-763. — URL: https://academic.oup.com/bioinformatics/article/31/5/761/2748191 (accessed: 31.07.2023) DOI: 10.1093/bioinformatics/btu703.

17. Ritchie G. R. S. Functional Annotation of Noncoding Sequence Variants / G. R. S. Ritchie, I. Dunham, E. Zeggini, P. Flicek // Nature Methods. — 2014. — 11. — p. 294-296. — URL: https://www.nature.com/articles/nmeth.2832 (accessed: 31.07.2023) DOI: 10.1038/nmeth.2832.

18. Tate J. G. COSMIC: the Catalogue Of Somatic Mutations In Cancer / J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka // Nucleic Acids Research. — 2019. — 47. — p. D941-D947. — URL: https://academic.oup.com/nar/article/47/D1/D941/5146192 (accessed: 31.07.2023) DOI: 10.1093/nar/gky1015.

19. O'Leary N. A. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation / N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo // Nucleic Acids Research. — 2016. — 44. — p. D733-D745. — URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189 (accessed: 31.07.2023) DOI: 10.1093/nar/gkv1189.

20. Li H. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform / H. Li, R. Durbin // Bioinformatics. — 2009. — 25. — p. 1754-1760. — URL: https://academic.oup.com/bioinformatics/article/25/14/1754/225615 (accessed: 31.07.2023) DOI: 10.1093/bioinformatics/btp324.

21. Koboldt D. C. VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples / D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson // Bioinformatics. — 2009. — 25. — p. 2283-2285. — URL: https://academic.oup.com/bioinformatics/article/25/17/2283/210190 (accessed: 31.07.2023) DOI: 10.1093/bioinformatics/btp373.

22. Landrum M. J. ClinVar: Improvements to Accessing Data / M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen // Nucleic Acids Research. — 2020. — 48. — p. D835-D844. — URL: https://academic.oup.com/nar/article/48/D1/D835/5645007 (accessed: 31.07.2023) DOI: 10.1093/nar/gkz972.

23. Karczewski K. J. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans / K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings // Nature. — 2020. — 581. — p. 434-443. — URL: https://www.nature.com/articles/s41586-020-2308-7 (accessed: 31.07.2023) DOI: 10.1038/s41586-020-2308-7.

24. Cottier F. Advantages of Meta-Total RNA Sequencing (MeTRS) over Shotgun Metagenomics and Amplicon-Based Sequencing in the Profiling of Complex Microbial Communities / F. Cottier, K. G. Srinivasan, M. Yurieva, W. Liao // npj Biofilms and Microbiomes. — 2018. — 4. — p. 2. — URL: https://www.nature.com/articles/s41522-017-0046-x (accessed: 31.07.2023) DOI: 10.1038/s41522-017-0046-x.

25. Tripp H. J. Misannotations of rRNA Can Now Generate 90% False Positive Protein Matches in Metatranscriptomic Studies / H. J. Tripp, I. Hewson, S. Boyarsky, J. M. Stuart // Nucleic Acids Research. — 2011. — 39. — p. 8792-8802. — URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr576 (accessed: 31.07.2023) DOI: 10.1093/nar/gkr576.

26. Buza K. RECORD: Reference-Assisted Genome Assembly for Closely Related Genomes / K. Buza, B. Wilczynski, N. Dojer // International Journal of Genomics. — 2015. — 2015. — p. 1-10. — URL: http://www.hindawi.com/journals/ijg/2015/563482/ (accessed: 31.07.2023) DOI: 10.1155/2015/563482.

27. Abascal F. TranslatorX: Multiple Alignment of Nucleotide Sequences Guided by Amino Acid Translations / F. Abascal, R. Zardoya, M. J. Telford // Nucleic Acids Research. — 2010. — 38. — p. W7-W13. — URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq291 (accessed: 31.07.2023) DOI: 10.1093/nar/gkq291.

28. Rokas A. Phylogenetic Analysis of Protein Sequence Data Using the Randomized Axelerated Maximum Likelihood (RAXML) Program / A. Rokas // Current Protocols in Molecular Biology. — 2011. — 96. — URL: https://onlinelibrary.wiley.com/doi/10.1002/0471142727.mb1911s96 (accessed: 31.07.2023) DOI: 10.1002/0471142727.mb1911s96.

29. Adebali O. Establishing the Precise Evolutionary History of a Gene Improves Prediction of Disease-Causing Missense Mutations / O. Adebali, A. O. Reznik, D. S. Ory, I. B. Zhulin // Genetics in Medicine. — 2015. — 18. — p. 1029-1036. — URL: https://linkinghub.elsevier.com/retrieve/pii/S1098360021044543 (accessed: 31.07.2023) DOI: 10.1038/gim.2015.208.

30. Langmead B. Fast gapped-read alignment with Bowtie 2 / B. Langmead, S. L. Salzberg // Nature Methods. — 2012. — 9. — p. 357-359. — URL: https://www.nature.com/articles/nmeth.1923 (accessed: 31.07.2023) DOI: 10.1038/nmeth.1923.

31. Li H. Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform / H. Li, R. Durbin // Bioinformatics. — 2010. — 26. — p. 589-595. — URL: https://academic.oup.com/bioinformatics/article/26/5/589/211735 (accessed: 31.07.2023) DOI: 10.1093/bioinformatics/btp698.

32. Van Der Auwera G. A. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline / G. A. Van Der Auwera, M. O. Carneiro, C. Hartl, R. Poplin // Current Protocols in Bioinformatics. — 2013. — 43. — URL: https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1110s43 (accessed: 31.07.2023) DOI: 10.1002/0471250953.bi1110s43.

33. Bankevich A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing / A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich // Journal of Computational Biology. — 2011. — 19. — p. 455-477. — URL: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021 (accessed: 31.07.2023) DOI: 10.1089/cmb.2012.0021.

34. Gudmundsson S. Variant Interpretation Using Population Databases: Lessons from gnomAD / S. Gudmundsson, M. Singer-Berk, N. A. Watts, W. Phu // Human Mutation. — 2022. — 43. — p. 1012-1030. — URL: https://onlinelibrary.wiley.com/doi/10.1002/humu.24309 (accessed: 31.07.2023) DOI: 10.1002/humu.24309.

35. Ghosh R. Updated Recommendation for the Benign Stand-Alone ACMG/AMP Criterion / R. Ghosh, S. M. Harrison, H. L. Rehm, S. E. Plon // Human Mutation. — 2018. — 39. — p. 1525-1530. — URL: https://onlinelibrary.wiley.com/doi/10.1002/humu.23642 (accessed: 31.07.2023) DOI: 10.1002/humu.23642.