DEVOUR: Deleterious Variants on Uncovered Regions in Whole-Exome Sequencing

Erdem Türk[1,2], Akif Ayaz[3], Ayhan Yüksek[1] and Barış E. Süzek[1,2]

[1] Department of Computer Engineering, Muğla Sıtkı Koçman University, Muğla, Turkey
[2] Bioinformatics Graduate Program, Muğla Sıtkı Koçman University, Muğla, Turkey
[3] Department of Medical Genetics, School of Medicine, İstanbul Medipol University, İstanbul, Turkey

ABSTRACT

The discovery of low-coverage (i.e. uncovered) regions containing clinically significant variants, especially when they are related to the patient's clinical phenotype, is critical for whole-exome sequencing (WES) based clinical diagnosis. Therefore, it is essential to develop tools to identify the existence of clinically important variants in low-coverage regions. Here, we introduce a desktop application, namely DEVOUR (DEleterious Variants On Uncovered Regions), that analyzes read alignments for WES experiments, identifies genomic regions with no or low-coverage (read depth < 5) and then annotates known variants in the low-coverage regions using clinical variant annotation databases. As a proof of concept, DEVOUR was used to analyze a total of 28 samples from a publicly available Hirschsprung disease-related WES project (NCBI Bioproject: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB19327), revealing the potential existence of 98 disease-associated variants in low-coverage regions. DEVOUR is available from https://github.com/projectDevour/DEVOUR under the MIT license.

Subjects Bioinformatics, Computational Biology, Computational Science
Keywords Next-generation sequence (NGS) analysis, Whole-exome sequencing (WES) analysis, Medical genetics, Genetic disposition to disease, Genetic diseases, Genetic variants, Clinical NGS informatics

Submitted 9 May 2023
Accepted 13 August 2023
Published 15 September 2023

Corresponding author
Erdem Türk, erdemturk@mu.edu.tr

Academic editor
Lasse Folkersen

Additional Information and Declarations can be found on page 12

DOI 10.7717/peerj.16026

INTRODUCTION

Whole-exome sequencing (WES) is a technique for sequencing the protein-coding (*i.e.,* exonic) regions of genes in a genome. WES is typically utilized to identify genetic variants that are associated with clinical phenotypes such as observable traits of a disease. As a result, WES has been widely utilized in both academic research and clinical diagnosis. Although WES is a cost-effective and convenient approach that has enabled the detection of clinically significant (*e.g.,* deleterious) variants in almost all coding regions of the genome (*Tetreault et al., 2015*), it has some limitations in analysis processes (*Bergant et al., 2018*).

In the context of WES analysis, the presence of variants in low-coverage regions poses a challenge for accurate variant detection and interpretation. Low-coverage regions are areas of the genome with no or low sequencing coverage, resulting in a lack of read depth to confidently identify genetic variants. Read depth refers to the number of sequencing reads that align to a specific genomic region. Regions with insufficient coverage have a significantly lower number of reads compared to adequately covered regions. By analyzing the read depth across the exome, it is possible to identify regions with low-coverage.

In the literature, several studies have reported that a substantial proportion of the genome may remain low-coverage in WES experiments. For example, *Bick et al. (2012)* observed that approximately 20% of the exome was not adequately covered in their study cohort. Similarly, *Ku et al. (2012)* reported that around 15% of the exome was poorly covered in their analysis. These findings suggest that a significant portion of the exome may be susceptible to being classified as low-coverage regions. It is important to note that the number of variants expected in low-coverage regions can vary depending on the sample population, sequencing platform, exome capture kit used, and other technical factors.

The presence of low-coverage regions in whole-exome sequencing (WES) experiments can be attributed to various factors. One primary factor is the design limitations of exome capture kits (*Kong et al., 2018*). These kits are designed to target specific genomic regions of interest, typically focusing on protein-coding exons. However, certain genomic regions, such as those with high GC content or segmental duplications, may be challenging to capture effectively, leading to reduced coverage and potentially missed variants (*Choi et al., 2009*; *Clark et al., 2011*; *Sheppard et al., 2018*).

In a study conducted by *Pengelly et al. (2020)*, the performances of two commercial exome capture kits were compared for clinical diagnostics. The researchers reported achieving 76% and 91% coverage for 100X experiments using these kits. Additionally, at 20X resolution, different exome capture kits were capable of covering 98% of the targeted genomic regions, and at 30X resolution, the coverage dropped to 97% of the genome. These findings emphasize the impact of coverage depth on the ability to capture genomic regions of interest and underscore the need to carefully consider the trade-offs between coverage depth and cost in WES experiments.

Although these exome capture kits achieved high coverage levels for most regions, the remaining uncovered regions warrant careful consideration, especially in clinical diagnostics, to ensure comprehensive variant detection and accurate genetic analysis. Low-coverage regions may contain clinically relevant variants, and their accurate identification is crucial for making informed clinical decisions.

In addition to the capture kit limitations, the lack of coverage in specific regions can also be influenced by technical factors and library preparation protocols. Sequencing biases, such as preferential amplification of certain genomic regions or biases introduced during library construction, can contribute to uneven coverage and subsequently affect variant detection (*Gnirke et al., 2009*; *Ross et al., 2013*).

Furthermore, the presence of low-coverage regions may also arise from biological factors. For example, segmental duplications and repetitive elements in the genome can hinder accurate mapping of reads, resulting in zero or low-coverage in those regions (*Kiezun et al., 2012*).

Therefore, the detection of low-coverage regions containing clinically significant variants, especially when they are relevant to the clinical phenotype of the patient, is of utmost importance and calls for further validation protocols such as deep sequencing.

Hence, there is a need for tools to alert clinicians regarding the risks associated with low-coverage regions that are directly related to the patient's health.

To address the challenge of identifying genomic regions with no or low-coverage in WES data, we introduce DEVOUR, a desktop application specifically designed for this purpose. Unlike traditional variant callers such as HaplotypeCaller (*McKenna et al., 2010*), Platypus (*Rimmer et al., 2014*), and Freebayes (*Garrison & Marth, 2012*), which focus on variant calling, DEVOUR takes a unique approach by prioritizing variant annotation and interpretation. DEVOUR is the first of its kind in this regard, aiming to identify low-coverage regions and annotate clinically important variants within them using clinical variant annotation databases.

By combining coverage analysis and comprehensive variant annotation, DEVOUR provides a powerful solution to uncover and analyze clinically significant variants, even in challenging low-coverage regions. This pioneering approach makes DEVOUR a valuable tool for researchers and clinicians in the pursuit of accurate clinical diagnosis.

In this study, we demonstrate DEVOUR's capabilities using publicly available Hirschsprung disease-related WES data provided by *Gui et al. (2017)* as an example. Hirschsprung disease is a congenital disorder caused by the absence of ganglion cells in the colon, and it is known to be caused by mutations in several genes, including RET, EDNRB, and SOX10 (*Tang et al., 2023*).

According to the ClinVar database (*Landrum et al., 2020*), there are total of 32 pathogenic and likely pathogenic variants related to Hirschsprung disease. These variants are mainly clustered on chromosomes 4 and 10, with additional variants present on chromosomes 1 and 22, as depicted in Fig. 1.

By employing DEVOUR's coverage analysis and variant annotation approach, we successfully identified low-coverage regions in Hirschsprung disease-related WES data. DEVOUR's comprehensive analysis allowed us to annotate clinically significant variants within these regions, providing valuable insights into potential disease-causing mutations. This capability enhances the accuracy of WES data analysis and demonstrates the significance of DEVOUR in uncovering crucial genetic information for clinical diagnosis.

MATERIALS & METHODS

DEVOUR works with a set of inputs; a read alignment file (SAM or BAM) for the WES experiment, a Browser Extensible Data (BED) file for the regions targeted by the WES capture kit, a read depth threshold to identify low-coverage regions, a human reference genome version (*i.e.,* hg19 or hg38), and a list of annotation resources for clinically significant (*e.g.*, deleterious) variants. DEVOUR works in three main steps as represented in Fig. 2.

The first step identifies low-coverage genomic regions in a WES experiment. To this end, a user-provided read alignment file (SAM or BAM) is sorted and indexed using Samtools (*Danecek et al., 2021*). Per-base depth calculation is performed using Mosdepth (*Pedersen & Quinlan, 2018*) which is a command-line tool for calculating the sequencing coverage. During the depth calculation process, an exome capture file is used to limit the calculation

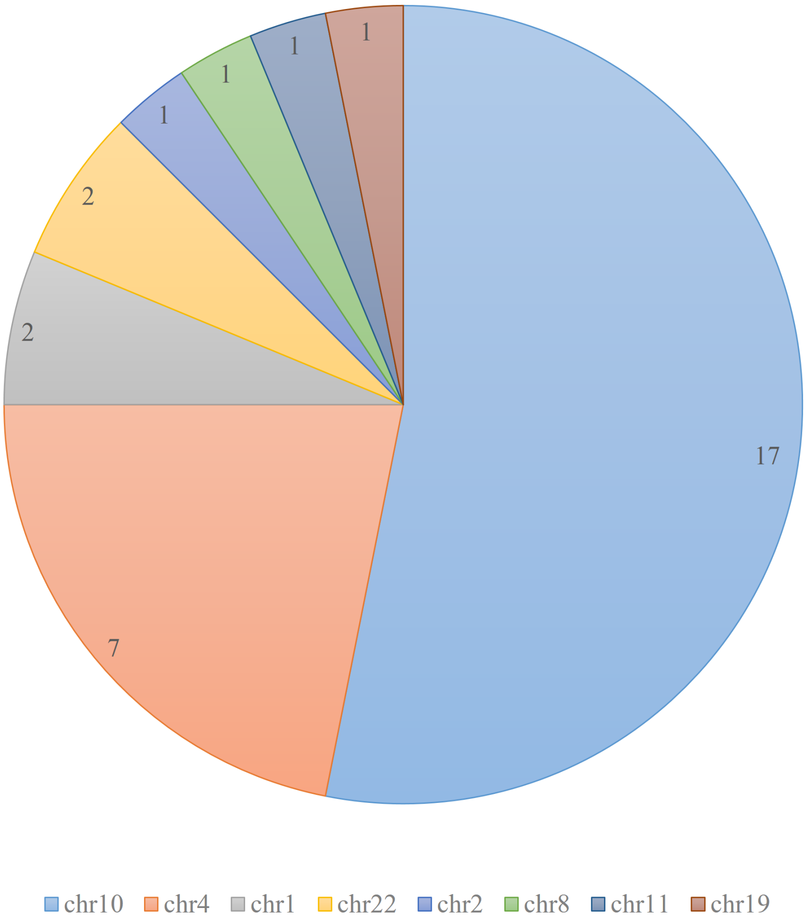


Figure 1 Distribution of pathogenic and likely pathogenic variants related to Hirschsprung disease in ClinVar database. The figure illustrates the distribution of 32 pathogenic and likely pathogenic variants associated with Hirschsprung disease in the ClinVar database. Each chromosome is represented by a different color. The majority of the pathogenic variants are clustered on chromosomes 4 and 10.

Full-size DOI: 10.7717/peerj.16026/fig-1

of low-coverage regions to the ones targeted by the respective exome capture kit. Next, a selected read depth threshold (default = 5) is used to identify genomic regions with low-coverage. The output of this step is a list of low-coverage genomic regions in BED format. It is important to note that setting a read depth threshold of zero in DEVOUR leads to the identification of exclusive regions with absolutely no coverage. This threshold allows for the specific detection of regions that lack any sequencing reads, highlighting areas where no genetic information is captured. The human reference genome release name is specifically required in DEVOUR to accurately present users with the appropriate versions of the annotation sources. It is important to note that while the reference genome release name is necessary, users do not need to provide the actual reference genome FASTA file. DEVOUR utilizes the reference genome release name to ensure that the corresponding annotation sources aligned to the specific genome version are made available during the analysis process. This allows users to select the correct and relevant annotation sources for

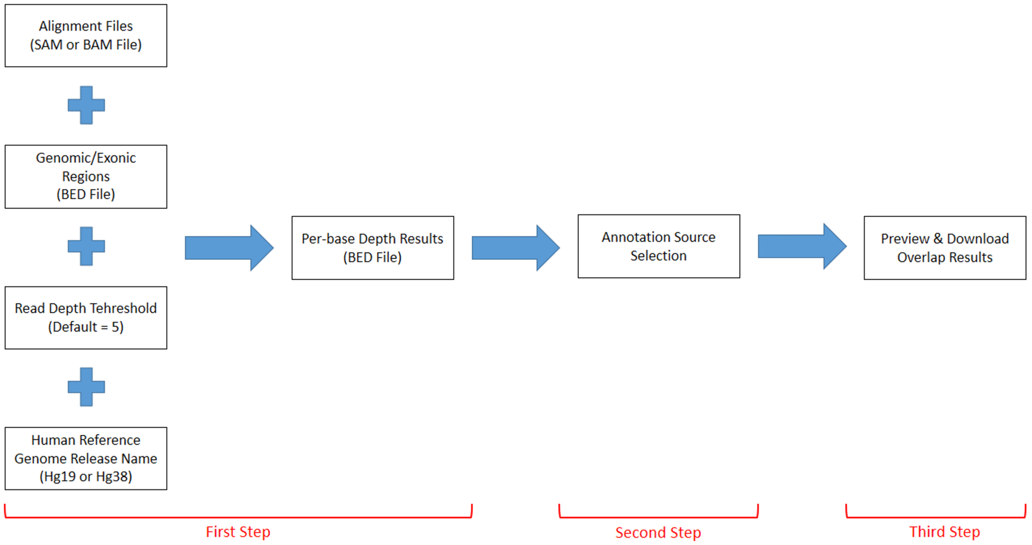


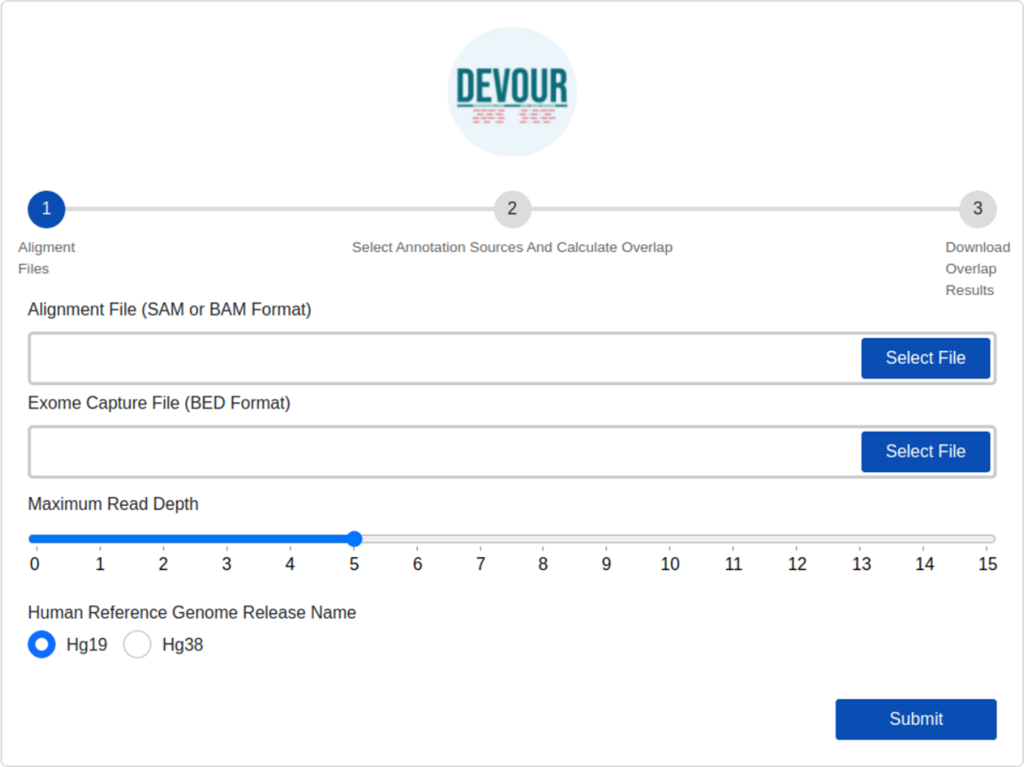
Figure 2 **The representation of DEVOUR's workflow.**

Full-size DOI: 10.7717/peerj.16026/fig-2

accurate variant annotation and interpretation. The user interface developed for the first step is shown in Fig. 3.

The second step aims to enrich the low-coverage regions with variant annotations as a step toward assessing the clinical significance of variants found in these regions. Variant annotations can either come from custom in-house annotation sources or clinically significant variant databases provided by ANNOVAR (*Wang, Li & Hakonarson, 2010*) such as ClinVar (*Landrum et al., 2020*). To ensure compatibility with DEVOUR, a custom annotation database must be provided as a BED-like file. This file format should consist of four tab-delimited columns, arranged in the specified order. The columns should include the chromosome name, start coordinate, end coordinate, and variant annotation in free text. DEVOUR has configuration screens to handle the incorporation of these custom annotation sources as shown in Fig. 4. Similarly, DEVOUR has implemented mechanisms to fetch disease-specific variant databases from ANNOVAR repositories and transform them into BED-like formats like custom annotation sources. Once the annotation sources are properly configured in DEVOUR, the analysis process presents the user with a comprehensive list of available databases. This list includes all the configured annotation sources, allowing the user to easily select the desired databases for the analysis. The availability and visibility of these databases during the analysis process enhance the user's ability to make informed decisions and effectively utilize the relevant annotation sources within DEVOUR.

Variants located in the low-coverage regions from the previous step are annotated using the selected variant annotation sources. For this annotation, DEVOUR utilizes an overlap computation algorithm leveraging the interval trees; a data structure that allows for efficient computation of intervals that overlap with a query interval. In this step, the genomic coordinates for clinical annotations are stored in an interval tree data structure



Alignment File (SAM or BAM Format)

Exome Capture File (BED Format)

Maximum Read Depth: 5

Human Reference Genome Release Name

● Hg19 ○ Hg38

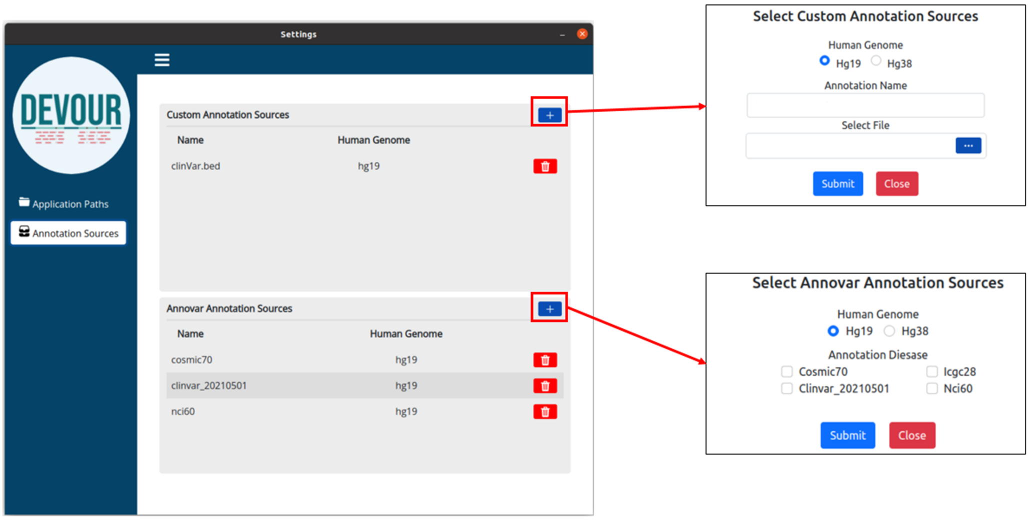
Figure 3 The illustration of the user interface for providing the parameters: input files (an alignment in SAM or BAM format and an exome capture file in BED format), depth threshold and human genome reference version. The initial stage in DEVOUR's analysis pipeline is to identify potentially uncovered or low-coverage genomic regions. A list of low-coverage genomic regions in BED format is the result of this stage.

Full-size DOI: 10.7717/peerj.16026/fig-3

per chromosome and queried with the low-coverage regions from the previous step. The output of this step is a set of tables where each table contains seven tab-delimited columns; the chromosomal location for the low-coverage regions (chromosome name, start/end coordinates), depth of the region, the chromosomal location for the annotated variant (start/end coordinates), and detailed annotation retrieved from respective source. The user interface developed for the first step is shown in Fig. 5.

The final step provides files to assist clinical diagnosis highlighting the variants in low-coverage genomic regions with clinical significance (*e.g.*, pathogenicity). DEVOUR helps users to preview and export each annotation-based table from the previous step in TSV or Excel format for inspection as shown in Fig. 6.

DEVOUR is developed using the Electron framework (http://www.electronjs.org). In addition, DEVOUR has some prerequisites that must be installed; Samtools, Mosdepth, ANNOVAR, and some Python libraries; intervaltree (https://pypi.org/project/intervaltree/), pandas (*Reback et al., 2022*), and openpxyl (https://openpyxl.readthedocs.io/en/stable/). As part of DEVOUR installation process, the paths for an application working directory and the prerequisite applications need to be configured.



Custom Annotation Sources

Name	Human Genome
clinVar.bed	hg19

Annovar Annotation Sources

Name	Human Genome
cosmic70	hg19
clinvar_20210501	hg19
nci60	hg19

Select Custom Annotation Sources

Human Genome
● Hg19 ○ Hg38

Annotation Name

Select File

Select Annovar Annotation Sources

Human Genome
● Hg19 ○ Hg38

Annotation Diesase
□ Cosmic70 □ Icgc28
□ Clinvar_20210501 □ Nci60

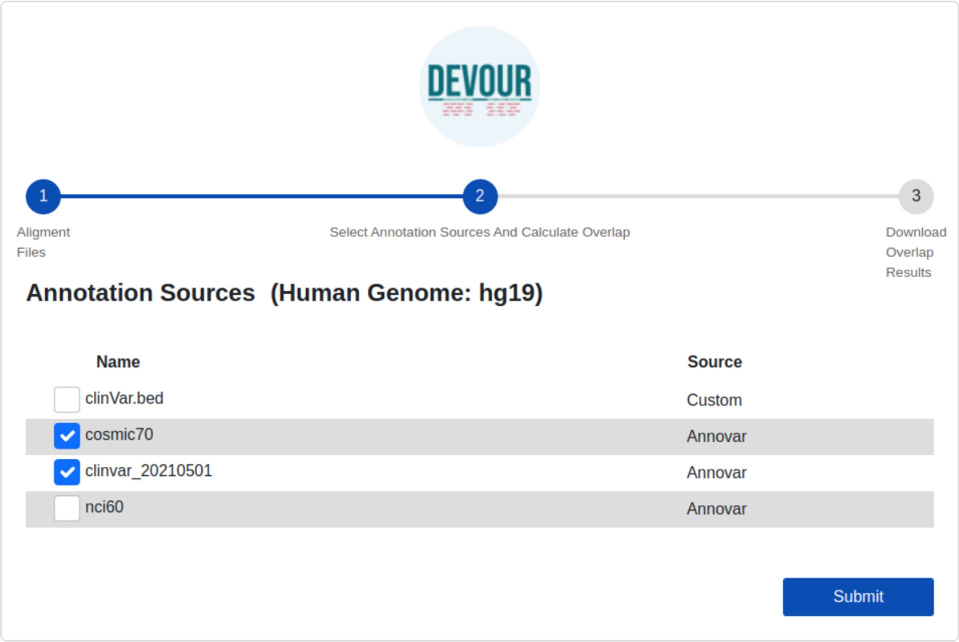
Figure 4 The illustration of user interfaces developed for users to create an annotation library. Either custom annotation sources or ANNOVAR's disease-associated variant databases, like ClinVar, can be used to annotate variants. A BED-like file with four tab-delimited columns holding the chromosomal name, start coordinate, end coordinate, and the variant annotation in free text, in that order, can serve as the custom annotation source. To handle the inclusion of various unique annotation sources, DEVOUR offers settings panels. Similar to custom annotation sources, DEVOUR has created methods to retrieve disease-associated variant databases from ANNOVAR repositories and convert them into BED-like formats.

Full-size DOI: 10.7717/peerj.16026/fig-4

Our DEVOUR tool was benchmarked using a publicly available WES project conducted by Gui and colleagues (*Gui et al., 2017*), focusing on Hirschsprung disease, a congenital abnormality characterized by the absence of nerves in portions of the intestine. This WES project (NCBI Bioproject: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB19327) contains a total of 72 samples obtained from two sequencing platforms (Illumina and ABI SOLiD) from the Sequence Read Archive (*Leinonen et al., 2011*). For testing purposes, we limited our analysis to 28 paired-end samples obtained from the Illumina platform, as we did not have access to the exome capture kit information for the samples obtained from the ABI SOLiD platform. The samples analyzed on the Illumina platform had a mean coverage of 27.9X, indicating that, on average, each base in the target region was sequenced about 28 times. Moreover, approximately 74% of the bases in the target region had a sequencing coverage greater than 10 times (*Gui et al., 2017*).

To conduct our analysis, we acquired the FASTA files for each sample and performed sequence alignment against the reference human genome (build hg19), using HISAT2 alignment tool (*Kim et al., 2019*), to obtain the corresponding BAM files. Subsequently, we processed these BAM files individually using DEVOUR. The first supplementary file provides a detailed description of the analysis process for one of the samples, specifically NCBI SRA: ERR1840777 (see Supplemental File 1). This example serves to showcase the usage of DEVOUR and provides step-by-step instructions for our analysis.

To identify clinically significant annotations in low-coverage regions (read depth < 5) and high-coverage regions (read depth >= 5) of these samples, we fed the BAM files to



Annotation Sources (Human Genome: hg19)

Name	Source
□ clinVar.bed	Custom
☑ cosmic70	Annovar
☑ clinvar_20210501	Annovar
□ nci60	Annovar

Figure 5 **The illustration of the user interface for selecting the desired annotation source(s).** At this stage, the annotation resources in the DEVOUR library, which were prepared according to the human reference genome version selected in the previous stage, are listed.

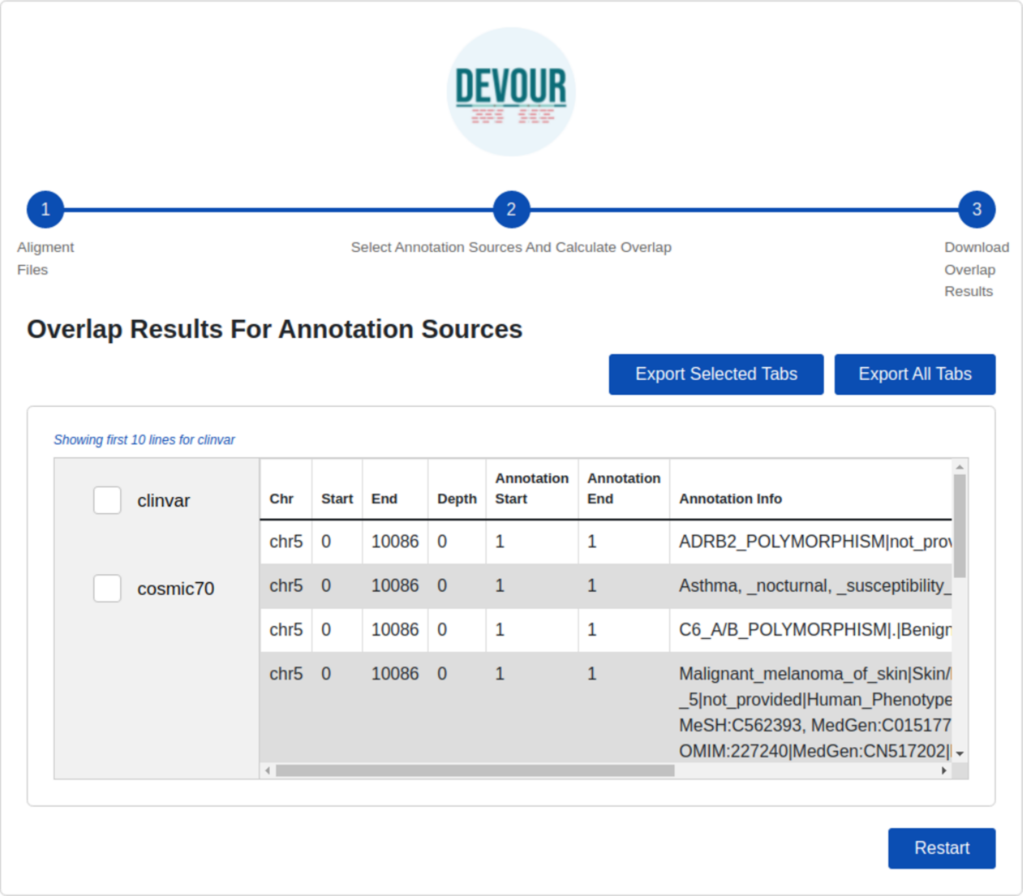
Full-size DOI: 10.7717/peerj.16026/fig-5

DEVOUR along with the exome capture kit used by the Illumina sequencing platform (Illumina, San Diego, CA, USA).

RESULTS

DEVOUR was used to identify ClinVar annotations overlapping with the identified low and high-coverage regions. In our analysis of 27 samples using DEVOUR, we detected at least one Hirschsprung-related pathogenic variant in high-coverage regions, as shown in Fig. 7 (see Supplementary File 2). On average, we identified approximately 17 pathogenic variants per sample, with a standard deviation of ±6.

Furthermore, in 18 of the samples, we found at least one Hirschsprung-related pathogenic variant located in low-coverage regions (see Supplementary File 3). The average number of such pathogenic variants identified per sample was approximately 8, with a standard deviation of ±6. In one sample (NCBI SRA: ERR1840777), no Hirschsprung-related pathogenic variants were identified in high-coverage regions. However, using DEVOUR, we detected a total of 27 Hirschsprung-related pathogenic variants in low-coverage regions within this sample. Notably, 25 out of these variants were located in regions with no coverage, indicated by a depth value of 0 (see Supplementary File 4). To validate our findings, we utilized the NCBI Sequence Viewer (*Rangwala et al., 2021*) to map the sequence reads obtained from sample ERR1840777 to the reference genome, allowing us to assess the coverage. Figure 8 illustrates the sequence coverage on chromosome 10 between coordinates 43,000,000 and 44,000,000 for this sample. The regions without bars



Overlap Results For Annotation Sources

Showing first 10 lines for clinvar

□ clinvar

□ cosmic70

Chr	Start	End	Depth	Annotation Start	Annotation End	Annotation Info
chr5	0	10086	0	1	1	ADRB2_POLYMORPHISM|not_prov
chr5	0	10086	0	1	1	Asthma, _nocturnal, _susceptibility_
chr5	0	10086	0	1	1	C6_A/B_POLYMORPHISM|.|Benign
chr5	0	10086	0	1	1	Malignant_melanoma_of_skin|Skin/ _5|not_provided|Human_Phenotype MeSH:C562393, MedGen:C015177 OMIM:227240|MedGen:CN517202|

Figure 6 The illustration of the user interface for reviewing the results. This stage seeks to generate result files to aid clinical diagnosis by indicating mutations in genomic regions with low-coverage that have clinical significance. DEVOUR enables users to evaluate each annotation-based table from the previous stage by previewing and exporting it in TSV or Excel format.

Full-size DOI: 10.7717/peerj.16026/fig-6

indicate low or no read coverage, which further supports the presence of low-coverage regions in the dataset. This visual representation enhances the comprehensibility of our identified low-coverage regions and bolsters our findings.

To gain further insights from DEVOUR's analysis, we conducted an evaluation of the total length of low-coverage regions per chromosome for each sample (see Supplementary File 5). Our analysis revealed that samples with Hirschsprung-related variants tend to exhibit larger proportions of low-coverage regions. This observation suggests the need for repeating WES experiments for these specific patients/samples to enhance coverage and improve the accuracy of variant detection.

An illustrative example is sample ERR1840777, wherein Hirschsprung-related variants were identified exclusively in low-coverage regions. In this scenario, DEVOUR's analysis serves as a valuable guide for clinicians, directing them to focus on these specific chromosome regions through more in-depth approaches such as deep sequencing. By conducting targeted experiments, clinicians can ascertain the presence of genuine variants

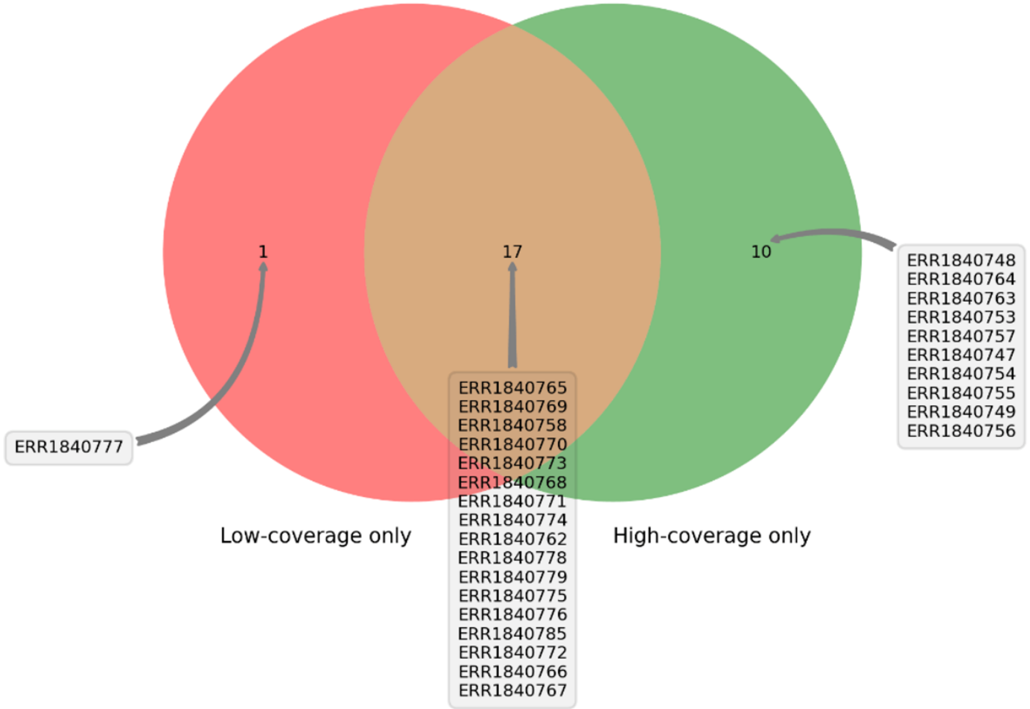


Figure 7 The distribution of samples containing Hirschsprung-related pathogenic variants on low- and high-coverage regions. DEVOUR analysis (with the default read depth threshold) revealed at least one Hirschsprung-related pathogenic variant in low-coverage regions (read depth < 5) and high-coverage regions (read depth ≥ 5) for 18 and 27 out of 28 samples, respectively. Sample ERR1840777 is distinctive as the only sample containing Hirschsprung-related pathogenic variants exclusively in low-coverage regions.

Full-size DOI: 10.7717/peerj.16026/fig-7

and achieve precise clinical interpretations. Unfortunately, NCBI SRA did not provide any clinical phenotype for this sample, but regardless, our finding asks for further investigation to improve or potentially correct this sample's Hirschsprung disease genotype.

The computation for each sample took eight minutes on average (min: 4 min, max: 12 min) on an Intel i7-5820K based virtual machine with 10 GB memory. These results show the importance of inspecting low-coverage regions using alternative methodologies such as deep sequencing as these regions may contain variants potentially critical for the clinical diagnosis. Without DEVOUR, the potential existence of such variants would not have been possible to identify and may result in missing potential diagnoses.

With DEVOUR's ability to incorporate these extended databases seamlessly, it becomes even more valuable in facilitating precise and comprehensive variant analysis, aiding in the identification of pathogenic variants, and contributing to improved diagnostic outcomes for patients. As the number of clinically important variants available in both public and private variant annotation resources continues to increase over time, we anticipate that DEVOUR will become increasingly beneficial, particularly for undiagnosed patients. The expanding databases of clinically significant variants provide a valuable resource for accurate variant interpretation and diagnosis.

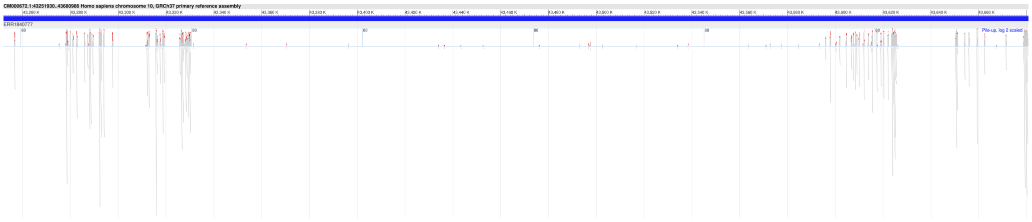


Figure 8 Sequence coverage on chromosome 10 (Coordinates: 43,000,000 - 44,000,000) for sample ERR1840777. The figure displays the sequence coverage on chromosome 10 in the genomic region spanning coordinates 43,000,000 to 44,000,000 for sample ERR1840777. Regions without bars on the graph indicate low or no read coverage, highlighting the presence of low-coverage regions in the dataset. The visual representation provides valuable insights into the distribution of sequence coverage and confirms the identification of low-coverage regions in the sample, supporting the findings of this study. NCBI Sequence Viewer Link: https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=CM000672.1&tracks=[key:sequence_track,name:T378820,display_name:Sequence,id:T378820,dbname:GenBank,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:1][key:alignment_track,name:ERR1840777,display_name:ERR1840777,id:STD2123359385,dbname:SRA,setting_group:cSRA,annots:ERR1840777,Layout:Adaptive,StatDisplay:15,Color:ShowDifferences,UnalignedTailsMode:glyph,HideSraAlignments:none,sort_by:,LinkMatePairAligns:false,ShowAlnStat:true,AlignedSeqFeats:false,Label:false,IdenticalBases:false,shown:true,order:7]&srz=ERR1840777&assm_context=GCA_000001405.3&mk=42833500|42833500|blue|9&v=43251931:43680986&c=FFFFFF&select=null&slim=0..

Full-size DOI: 10.7717/peerj.16026/fig-8

CONCLUSIONS

In the context of whole-exome sequencing (WES) analysis, it is crucial to identify and notify clinicians about clinically significant variants located in low-coverage regions to avoid missing potential diagnoses. Low-coverage regions can contain important genetic variants that contribute to the observed clinical phenotype but may go undetected due to insufficient sequencing coverage. By identifying these variants, clinicians can gain valuable insights into the underlying genetic basis of the patient's condition and make more informed diagnostic and treatment decisions. Therefore, it is essential to develop tools and strategies that effectively address the challenge of detecting and interpreting variants in low-coverage regions to maximize the diagnostic yield of WES analysis.

To address this need, we have developed DEVOUR, a desktop application that facilitates the analysis of WES experiments and identifies clinically significant variants in regions with low or no coverage. DEVOUR serves as a valuable addition to WES analysis pipelines, particularly those focused on detecting and annotating variants in covered genomic regions.

Looking towards the future, we envision DEVOUR's expansion to encompass not only ClinVar but also central protein databases such as Uniprot, as well as established mutation databases like HGMD. This evolution is expected to significantly augment DEVOUR's versatility, extending its applicability to both germline studies and somatic analyses. The anticipated outcome is enhanced robustness, particularly in scenarios where pinpointing precise target points is of utmost importance, such as in somatic investigations. Moreover, our vision includes the potential extension of DEVOUR's capabilities to encompass whole-genome sequencing (WGS) experiments, thereby broadening its scope to cater to an even broader array of genomic analyses. The pipeline integrated into DEVOUR is designed to accommodate WGS data seamlessly. Users are advised to provide a BED file that encompasses the entire genomic coordinate instead of solely exonic regions when working with WGS data. This BED file, generated manually by including the desired genomic regions for analysis, ensures effective utilization of DEVOUR's comprehensive variant annotation and interpretation capabilities in WGS data analysis.

By leveraging DEVOUR, researchers and clinicians can enhance their understanding of genomic variants, enabling more accurate and informed decision-making in clinical settings. With its versatility in handling both WES and potentially WGS data, DEVOUR is a valuable tool for comprehensive variant analysis and interpretation in a clinical research setting.

ACKNOWLEDGEMENTS

We would like to acknowledge that this study is part of Erdem Türk's Ph.D. thesis at Muğla Sıtkı Koçman University Graduate School of Natural and Applied Sciences.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Scientific and Technological Research Council of Turkey (No. 120E522). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Scientific and Technological Research Council of Turkey: 120E522.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Erdem Türk conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Akif Ayaz analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Ayhan Yüksek performed the experiments, prepared figures and/or tables, contributed to the development of the tool, and approved the final draft.
- Barış E. Süzek conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

DEVOUR is available at GitHub: https://github.com/projectDevour/DEVOUR.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.16026#supplemental-information.

REFERENCES

Bergant G, Maver A, Lovrecic L, Cuturilo G, Hodzic A, Peterlin B. 2018. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. *Genetics in Medicine* **20**:303–312.

Bick AG, Flannick J, Ito K, Cheng S, Vasan RS, Parfenov MG, Herman DS, De Palma SR, Gupta N, Gabriel SB, Funke BH, Rehm HL, Benjamin EJ, Aragam J, Taylor J, Herman A. Fox ER, Newton-Cheh C, Kathiresan S, O'Donnell CJ, Wilson JG, Altshuler DM, Hirschhorn JN, Seidman JG, Seidman C. 2012. Burden of rare sarcomere gene variants in the Framingham and Jackson heart study Cohorts. *American Journal of Human Genetics* **91**:513 DOI 10.1016/j.ajhg.2012.07.017.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Özen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**:19096 DOI 10.1073/pnas.0910672106.

Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. 2011. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology* **29**:908–914 DOI 10.1038/nbt.1975.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**:giab008 DOI 10.1093/gigascience/giab008.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing DOI 10.48550/arXiv.1207.3907.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27**:182 DOI 10.1038/nbt.1523.

Gui H, Schriemer D, Cheng WW, Chauhan RK, Antinolo G, Berrios C, Bleda M, Brooks AS, Brouwer RW, Burns AJ, Cherny SS, Dopazo J, Eggen BJ, Griseri P, Jalloh B, Le TL, Lui VC, Luzon-Toro B, Matera I, Ngan ES, Pelet A, Ruiz-Ferrer M, Sham PC, Shepherd IT, So MT, Sribudiani Y, Tang CS, van denHout MC, vander Linde HC, van Ham TJ, van I WF, Verheij JB, Amiel J, Borrego S, Ceccherini I, Chakravarti A, Lyonnet S, Tam PK, Garcia-Barcelo MM, Hofstra RM. 2017. Whole exome sequencing coupled with unbiased functional analysis reveals new Hirschsprung disease genes. *Genome Biology* **18**:48 DOI 10.1186/s13059-017-1174-6.

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PIW, Purcell SM, Sunyaev SR. 2012. Exome sequencing and the genetic basis of complex traits. *Nature Genetics* **44**:623–630 DOI 10.1038/ng.2303.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**:907–915 DOI 10.1038/s41587-019-0201-4.

Kong SW, Lee I-H, Liu X, Hirschhorn JN, Mandl KD. 2018. Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genetics in Medicine* **20**:1617–1626.

Ku C-S, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. 2012. Exome sequencing: dual role as a discovery and diagnostic tool. *Annals of Neurology* **71**:5–14 DOI 10.1002/ana.22647.

Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, Lyoshin V, Maddipatla Z, Maiti R, Mitchell J, O'Leary N, Riley GR, Shi W, Zhou G, Schneider V, Maglott D, Holmes JB, Kattman BL. 2020. ClinVar: improvements to accessing data. *Nucleic Acids Research* **48**:D835–D844 DOI 10.1093/nar/gkz972.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. *Nucleic Acids Research* **39**:D19–D21 DOI 10.1093/nar/gkq1019.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, De Pristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303 DOI 10.1101/gr.107524.110.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**:867–868 DOI 10.1093/bioinformatics/btx699.

Pengelly RJ, Ward D, Hunt D, Mattocks C, Ennis S. 2020. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Scientific Reports* **10**:1–7 DOI 10.1038/s41598-020-60215-y.

Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, Joukov V, Lotov V, Pannu R, Rudnev D, Shkeda A, Weitz EM, Schneider VA. 2021. Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). *Genome Research* **31**:159–169 DOI 10.1101/gr.266932.120.

Reback J, Jbrockmendel , McKinney W, Van den Bossche J, Augspurger T, Roeschke M, Hawkins S, Cloud P, Young GF, Sinhrks , Hoefler P, Klein A, Petersen T, Tratner J, She C, Ayd W, Naveh S, Darbyshire J, Garcia M, Shadrach R, Schendel J, Hayden A, Saxton D, Gorelli ME, Li F, Zeitlin M, Jancauskas V, McMaster A, Wörtwein T, Battiston P. 2022. pandas-dev/pandas: Pandas 1.4.2. *Zenodo* DOI 10.5281/zenodo.6408044.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. 2014. Integrating mapping-assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**:912–918 DOI 10.1038/ng.3036.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biology* **14**:R51 DOI 10.1186/gb-2013-14-5-r51.

Sheppard S, Biswas S, Li MH, Jayaraman V, Slack I, Romasko EJ, Sasson A, Brunton J, Rajagopalan R, Sarmady M, Abrudan JL, Jairam S, De Chene ET, Ying X, Choi J, Wilkens A, Raible SE, Scarano MI, Santani A, Pennington JW, Luo M, Conlin LK, Devkota B, Dulik MC, Spinner NB, Krantz ID. 2018. Utility and limitations of exome sequencing as a genetic diagnostic tool for children with hearing loss. *Genetics in Medicine* **20**:1663–1676.

Tang CS, Karim A, Zhong Y, Chung PH, Tam PK. 2023. Genetics of Hirschsprung's disease. *Pediatric Surgery International* **39**:104 DOI 10.1007/s00383-022-05358-x.

Tetreault M, Bareke E, Nadaf J, Alirezaie N, Majewski J. 2015. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert Review of Molecular Diagnostics* **15**:749–760 DOI 10.1586/14737159.2015.1039516.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**:e164 DOI 10.1093/nar/gkq603.