Article

PhenGenVar: A User-Friendly Genetic Variant Detection and Visualization Tool for Precision Medicine

JaeMoon Shin [1,2,†], **Junbeom Jeon** [1,†], **Dawoon Jung** [1], **Kiyong Kim** [3], **Yun Joong Kim** [4,5], **Dong-Hoon Jeong** [6,*] and **JeeHee Yoon** [1,*]

1   Department of Computer Engineering, Hallym University, Chuncheon 24252, Korea; shin@dbcls.rois.ac.jp (J.S.); jjbblue12@hallym.ac.kr (J.J.); kozer@naver.com (D.J.)
2   Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Chiba-Ken, Kashiwa-Shi 277-0971, Japan
3   Department of Electronic Engineering, Kyonggi University, Suwon 16227, Korea; eye4eye@kyonggi.ac.kr
4   Department of Neurology, Yonsei University College of Medicine, Seoul 03722, Korea; yunjkim@yuhs.ac
5   Department of Neurology, Yongin Severance Hospital, Yonsei University Health System, Yongin 16995, Korea
6   Department of Life Science and Multidisciplinary Genome Institute, Hallym University, Chuncheon 24252, Korea
*   Correspondence: dhjeong73@hallym.ac.kr (D.-H.J.); jhyoon@hallym.ac.kr (J.Y.)
†   These authors contributed equally to this work.

Abstract: Precision medicine has been revolutionized by the advent of high-throughput next-generation sequencing (NGS) technology and development of various bioinformatic analysis tools for large-scale NGS big data. At the population level, biomedical studies have identified human diseases and phenotype-associated genetic variations using NGS technology, such as whole-genome sequencing, exome sequencing, and gene panel sequencing. Furthermore, patients' genetic variations related to a specific phenotype can also be identified by analyzing their genomic information. These breakthroughs paved the way for the clinical diagnosis and precise treatment of patients' diseases. Although many bioinformatics tools have been developed to analyze the genetic variations from the individual patient's NGS data, it is still challenging to develop user-friendly programs for clinical physicians who do not have bioinformatics programing skills to diagnose a patient's disease using the genomic data. In response to this demand, we developed a Phenotype to Genotype Variation program (PhenGenVar), which is a user-friendly interface for monitoring the variations in a gene of interest for molecular diagnosis. This allows for flexible filtering and browsing of variants of the disease and phenotype-associated genes. To test this program, we analyzed the whole-genome sequencing data of an anonymous person from the 1000 human genome project data. As a result, we were able to identify several genomic variations, including single-nucleotide polymorphism, insertions, and deletions in specific gene regions. Therefore, PhenGenVar can be used to diagnose a patient's disease. PhenGenVar is freely accessible and is available at our website.

Keywords: precision medicine; NGS; exome browser; genetic variations

1. Introduction

Precision medicine, also known as personalized medicine, is a new emerging field of medicine that combines the individual patient's clinical phenotypes, health records, and diverse omics data for tailored diagnosis, treatment, and prevention of human diseases [1]. With the completion of the human genome project and advances in next-generation DNA sequencing (NGS) technologies, it is now possible to identify the patient's genetic variations at single-nucleotide resolution and interpret the phenotypic consequences of them in human diseases [2–5].

Several NGS approaches have been used to identify complex human disease-associated genetic variations. These include whole-genome sequencing, whole-exome sequencing, and

targeted gene panel sequencing [6,7]. Recent large-scale consortia efforts have facilitated the meta-analysis of genome-wide association studies (GWAS) to identify genetic variants associated with human diseases, such as Alzheimer's disease, cancers, coronary artery disease, and type 2 diabetes mellitus [4,8–10]. For instance, several genes, such as *APOE*, *CD33*, *CLU*, *CR1*, *EPHA1*, *PICALM*, and *TREM2*, have been identified as major genetic risk factors for Alzheimer's disease [11–15]. *BRCA1* and *BRCA2* mutations account for approximately 5% of all breast cancer cases [16,17]. However, pathogenic mutations in *BRCA1* and *BRCA2* increase the risk of breast cancer by 65% and 45%, respectively [18]. This implies that despite the low total heritability explained by polymorphisms in these genes, monitoring genetic variations in *BRCA1* and *BRCA2* is advantageous for the diagnosis of breast cancer in high-risk patients [19]. Therefore, it is essential to obtain a catalog of genes associated with specific human diseases and phenotypes in order to facilitate genetic risk prediction. There are many clinical databases, such as Medical Subject Headings (MeSH), the National Cancer Institute's (NCI) Thesaurus, Online Mendelian inheritance in man (OMIN), SNOMEDCT, United Medical Language System (UMLS), and Human Phenotype Ontology (HPO) [20–25]. Among these, HPO provides a more comprehensive resource that contains semantic links for disease genes and ontologies for computational analysis of human phenotypes.

NGS data of individual patients contain diverse types of genetic variations, such as single nucleotide variants (SNVs), insertions and deletions (indels), and genetic aberrations, such as inversions and translocations. There are many computational tools for variant calling from NGS data, such as DeepVariant, DELLY, FermiKit, GATK HaplotyperCaller (GATK HC), Pindel, Platypus, Strelka2, and VarScan [26–33]. Using these variant detection tools, it is possible to detect the genetic variants in the disease-related genes for molecular diagnosis. The genetic variants identified from the NGS data are stored as variant call format (VCF), mutation annotation file (MAF), mutation file format (MUT), or other types of files. Among these, VCF is the most widely used community standard for storing mutation data [34]; however, the analysis of VCF files requires intensive bioinformatic expertise. Therefore, there is a need to develop tools for managing and visualizing VCF data. The currently available VCF processing programs include SNVerGUI, database.bio, DaMold, mirVAFC, GAVIN, and gNOME [35–40]. In addition, other tools, such as integrative genomic viewer (IGV), variant call miner (VCF-Miner), VCF.Filter, myVCF, BrowseVCF, and VCF-Server, were also developed to identify disease-associated genetic variants from NGS data [41–46]. Among these, IGV is the most widely used tool for visualizing genome variation data. However, there is still a need to develop an easy and comprehensive tool for non-bioinformatician researchers to analyze a patient's genomic data for diagnosis and clinical decision support.

In this study, we developed the Phenotype to Genotype Variation (PhenGenVar) browser application as an intuitive user interface. This tool was designed for physicians or researchers to monitor the patient's genetic variations from the set of selected genes that are associated with a specific disease or phenotype. With exome and genome browsers in a PhenGenVar application, it is possible to detect genetic variations at the gene or exon level, as well as at the single-nucleotide level.

2. Materials and Methods

2.1. Development of Visual Interface for PhenGenVar Browser

The PhenGenVar browser application was written in the C# programing language using Microsoft.Net framework. The application was developed as a graphic interface that can be performed on a Windows operating system. The human genome reference dataset, including the position, annotation, and amino acid composition, was obtained from the UCSC genome browser database [47]. To remove duplicated reads and minimize errors in sequence alignment, the flag and concise idiosyncratic gapped alignment report (CIGAR) string fields were utilized for error correction and realignment of the read data in the BAM file [48]. We adopted a data partitioning and indexing method for efficient

visualization of large-scale genomic data. Genomic data are divided into partitions that can be managed and accessed separately, using an efficient indexing scheme. To reduce the information loss caused by partitioning, each partitioned region included overlapping regions between the flanking partitions. In display processing, indexed partitions are loaded into memory, and the corresponding genomic regions are displayed on the screen according to the resolution. In the browser, the resolution of the sequence alignment was adjusted from a 100 bp to 100 kb window. To minimize the decline in speed to render many output images of multiresolution, Direct2D-based rendering was used.

2.2. Database Embedded in the PhenGenVar Application

Human reference genome data were downloaded and pre-installed in the PhenGenVar application as the hg19 and hg38 versions [47]. The SNP database (dbSNP) was downloaded from the National Center for Biotechnology Information [49]. The dbSNP versions available for the hg19 human genome are SNP138, SNP141, SNP142, SNP144, SNP146, SNP147, SNP150, and SNP151. The dbSNP versions available for hg38 are SNP141, SNP142, SNP144, SNP146, SNP147, SNP150, and SNP151. The gene sets were downloaded from the human phenotype ontology (HPO) to provide a list of genes associated with human disease and phenotype [24].

2.3. Sample Data Used in This Study

To test the PhenGenVar application for a patient's diagnosis, publicly available personal whole-genome sequence data were retrieved from the 1000 Genomes Project with sample accession number NA11995 [50].

2.4. Implementation

The PhenGenVar is available at http://dblab.hallym.ac.kr/PhenGenVar, and the copyright of the software was provided freely. PhenGenVar was implemented on a Windows 10 platform with an Intel Core i7 3.3GHz CPU, 32 GB main memory, and 1TB hard drive.

3. Results

3.1. Development of a PhenGenVar Browser Application

We developed a PhenGenVar browser application as an intuitive user interface to enable clinical physicians and biological researchers to explore genetic variations in a gene of interest for molecular diagnosis. It allows users to browse for genomic regions and variants associated with a specific phenotype. To protect the personal information of patients and enhance the speed of the analysis, the application was developed as a personal computer-based software program that runs locally instead of a web-based program.

PhenGenVar consists of two separate browsers: an exome browser and a genome browser (Figure 1). The exome browser of PhenGenVar is the main program designed for gene-level analysis by selecting a user-defined phenotype-related gene set and then monitoring the genetic variations of the specific gene. The exome browser output displayed the corresponding genomic regions with the reported variants and detailed information. The genome browser can subsequently be called from the exome browser and utilized to closely browse the variant calling results based on read alignment. Thus, the PhenGenVar application provides a convenient and intuitive user interface at various levels of genomic resolution.

3.2. PhenGenVar Exome Browser for Gene-Level Variation Analysis

The exome browser of PhenGenVar is an overview page for gene-level analysis of genetic variations from a patient's genomic data, such as exome sequencing or whole-genome sequencing. This browser page mainly consists of data uploading areas, gene and variant filtering panels, and an exon viewer panel (Figure 1A).
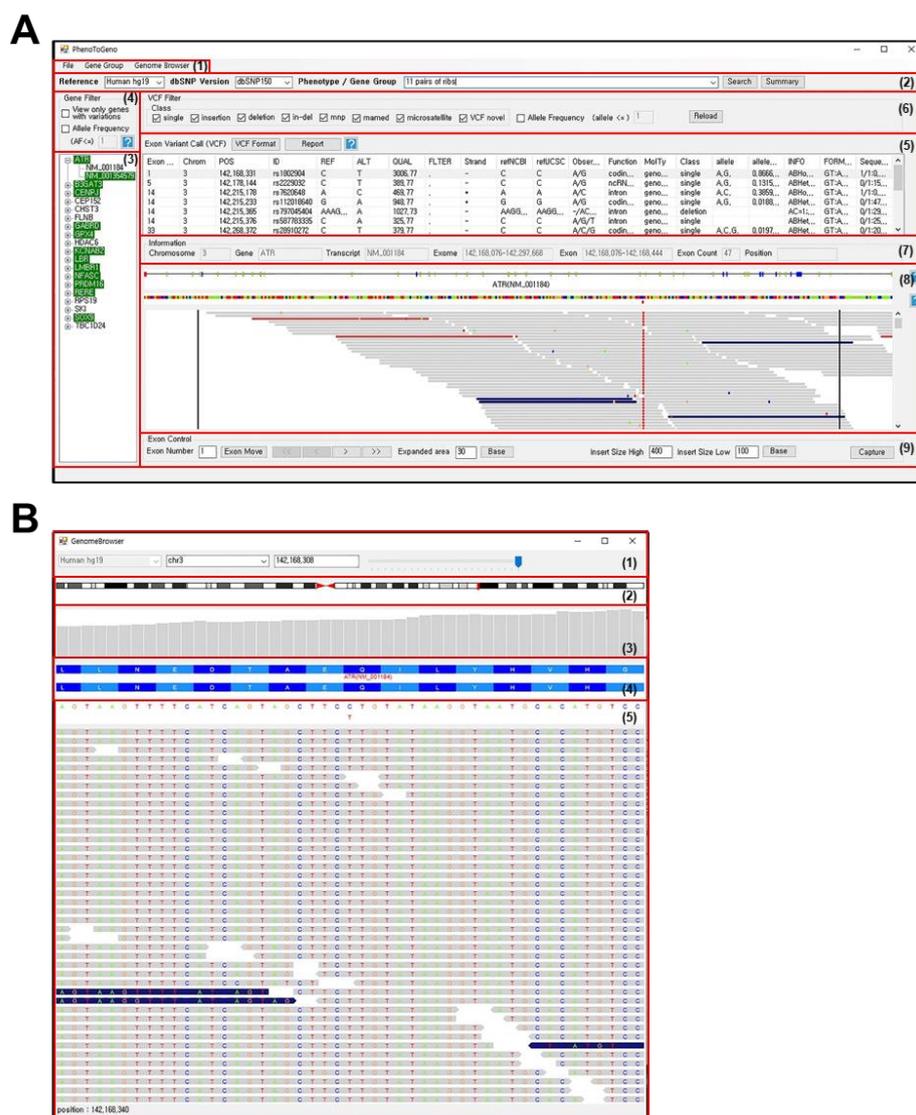
Figure 1. Main pages of PhenGenVar exome and genome browser. (**A**) Main page of PhenGenVar exome browser. The browser consists of menu bar (1), control panel (2), gene/transcript view panel (3), gene filter panel (4), exon variant call panel (5), VCF filter panel (6), information panel (7), main exon view panel (8), and exon view control panel (9). (**B**) Main page of PhenGenVar genome browser. The browser consists of control panel (1), cytoband panel (2), coverage graph panel (3), genetic structure panel (4), and main genome view panel (5).

To begin data analysis, it is necessary to upload the binary sequence alignment map (BAM) file and VCF file, both of which are generated from the patient's exome or whole-genome sequencing data. To explore the genetic variations of genes related to the patient's disease phenotype, a gene list can be created or typed individually as HUGO Gene Nomenclature Committee (HNGC) official gene symbol names [51]. In addition, the exome browser provides the gene groups registered in human phenotype ontology (HPO), which is a comprehensive resource for over 13,000 terms related to phenotypic abnormalities found in human diseases [24]. To analyze the genetic variations of each gene using the VCF/BAM file data, users can flexibly select a specific human reference genome sequence version and dbSNP version.

Once the patient's genome data and gene groups are uploaded, the analyzed output data are represented as a summary table, as shown in Figure 2A. This table shows the SNP data for each gene in the selected gene group. The SNP data include not only the SNPs registered in the dbSNP database [52] but also novel SNPs, which might be rare variations

found in the patient's genome. A set of genes with genetic variation can be selected using the gene filter panel of the exome browser, as shown in Figure 2B. In this filter panel, allele frequency values can also be adjusted to narrow down genes with more significant genetic variations. In addition to gene filtering, a more detailed filtering option is available using the VCF filter panel (Figure 2C, upper panel). Once a specific gene is selected from the gene filter panel, all variant information of the corresponding gene is represented in the exon variant call panel of the exome browser (Figure 2C, lower panel). This panel provides detailed variant information, such as position, dbSNP reference number (rsID), sequence alterations compared to the reference genome, and type of variations. In the above VCF filter panel, additional filtering was performed by clicking the type of sequence variations and adjusting the allele frequency.
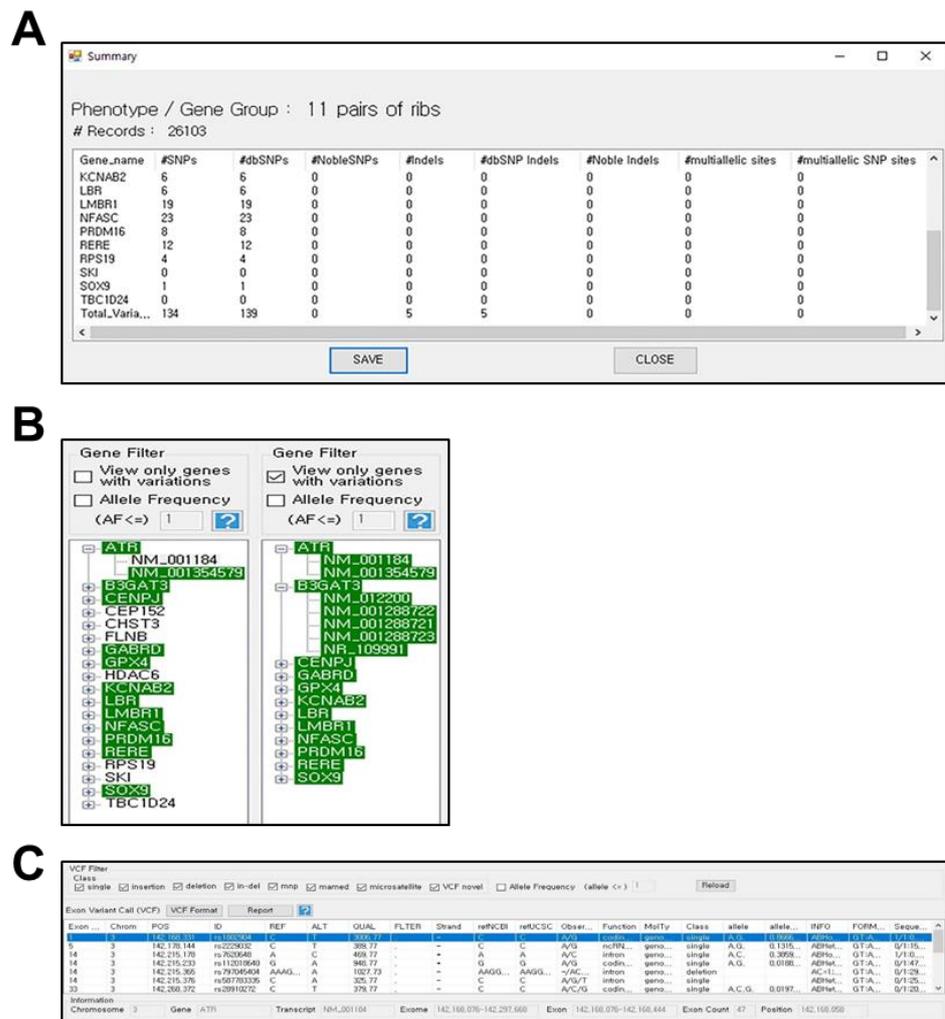

Figure 2. Output data of the exon variant information. (**A**) Statistical summary of the variant information from the VCF file. (**B**) Example of the gene filter panel and gene/transcript view panel. The genes and transcript with variants are highlighted with green color. (**C**) Example of the variant output in the VCF filter panel and the exon variant call panel.

The final output of the exome browser is shown in the main exon view panel, in which the gene structure, variants along the reference sequence, and read alignment are represented (Figure 3). To test how the PhenGenVar application can be utilized to detect genetic variation, we uploaded a publicly available personal whole-genome sequence data from 1000 human genome project and monitored it to identify various genetic variations [50]. In the gene structure area, exons with or without sequence variations are represented in blue or green boxes, respectively. The read alignment viewer showed various types

of sequence variations, such as SNPs (Figure 3A), deletions (Figure 3B), and insertions (Figure 3C). Therefore, we were able to identify specific sequence variations in the read alignment viewer. The sequence variation regions can also be represented by selecting the specific variation from the exon variant call panel (Figure 3C). Since the PhenGenVar application was designed to show the read alignment in the 30 bp upstream and downstream of the exon as a default setting, the exome browser also provides sequence variation in the intron. The scope of intron area output was adjustable in the Viewer Control Panel. We demonstrated that the exome browser of the PhenGenVar application can be used to identify genetic variations, which might be related to the patient's disease phenotype.
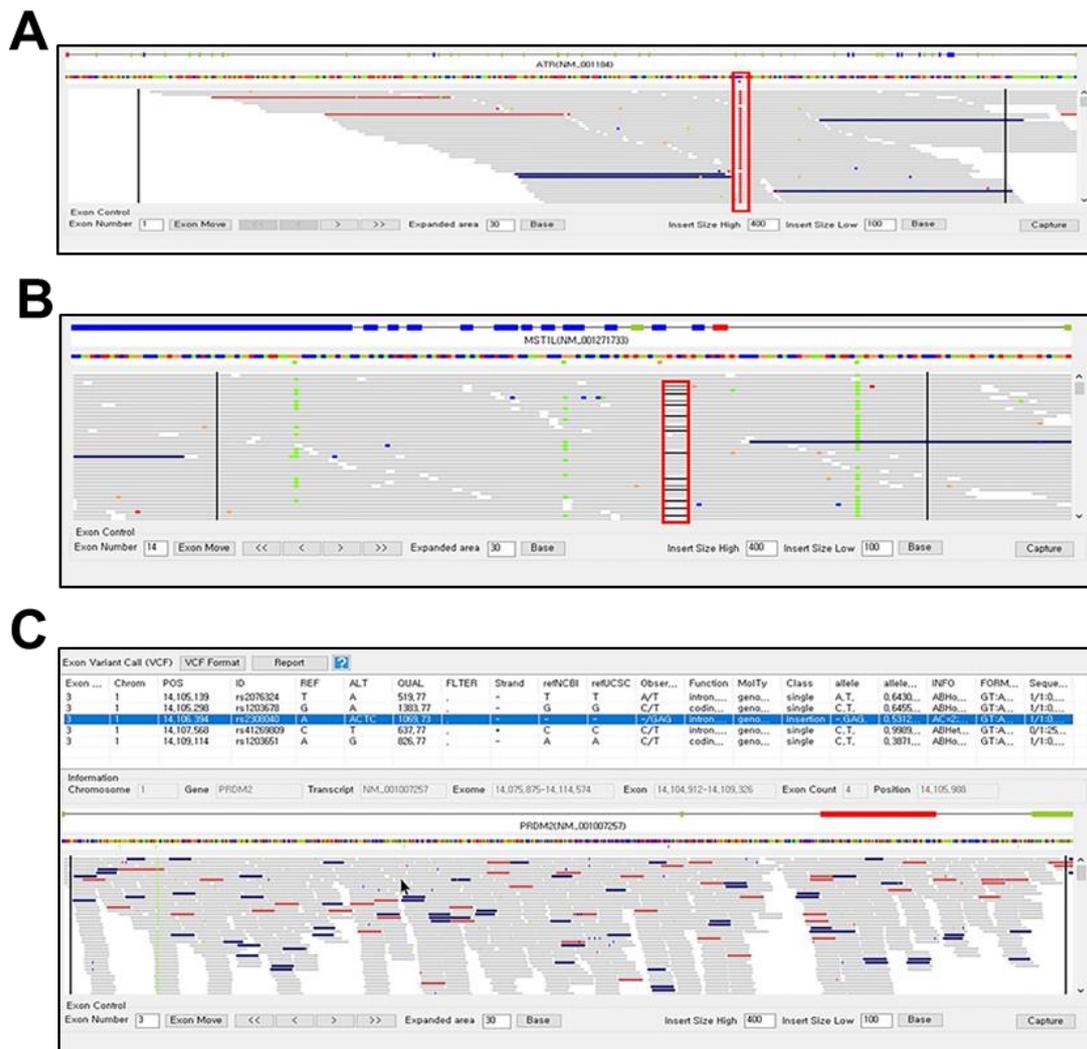

Figure 3. Identification of genetic variants in an exome browser. (**A**) Single nucleotide polymorphism (SNP) in an ATR gene. The red box indicates the position of SNP. (**B**) Deletion in a MST1L gene. The deleted area is shown in a red box. (**C**) Insertion of the sequences in a PRDM2 gene. The position of the insertion is indicated by the arrow.

3.3. PhenGenVar Genome Browser for Single Base-Resolution Analysis

The PhenGenVar genome browser was designed to present detailed read alignment information at single-nucleotide resolution (Figure 1B). The genome browser is popped up from the exome browser by selecting variants from the exon variant call panel or by double clicking the variant region from the main view panel. In addition, read alignment can be performed in a specific region of the genome by typing the chromosome position or gene name in the control panel of the genome browser. The resolution of the read alignment panel can be adjusted by changing the levels in the trackbar of the control panel from 59 bp

to 121,704 bp. Thus, the structure of the corresponding gene and its neighboring genes can be shown at low resolution, while the nucleotide sequences and coded amino acid sequences can be shown at higher resolution.

To test whether the PhenGenVar genome browser can be utilized to identify specific genetic variations at single-nucleotide resolution, we monitored the various types of sequence variations using the personal genome sequencing data that were used above. As shown in Figure 4, we identified SNP variant, deletion, and insertion in specific genes of personal genome. The single nucleotide change from C to T in an *ATR* gene, in which the synonymous mutation does not change the coding amino acid, was identified from all the sequence reads (Figure 4A). In an *MST1L* gene, deletion of five nucleotide was confirmed in half of the sequence reads (Figure 4B). This implies that one of two paired chromosomes harbors a mutated MST1L gene with a frameshift. We also identified an insertion of additional CTC sequence in a *PRDM2* gene (Figure 4C). Since about half of the sequence reads contained this insertion, we estimate that one of two paired chromosomes has a mutated *PRDM2* gene with an additional CTC insertion. As a result, we showed that the visual inspection with a PhenGenVar genome browser is an effective and powerful tool for variant call validation, reducing the number of false positives and assisting in the confirmation of true genetic variant discoveries.
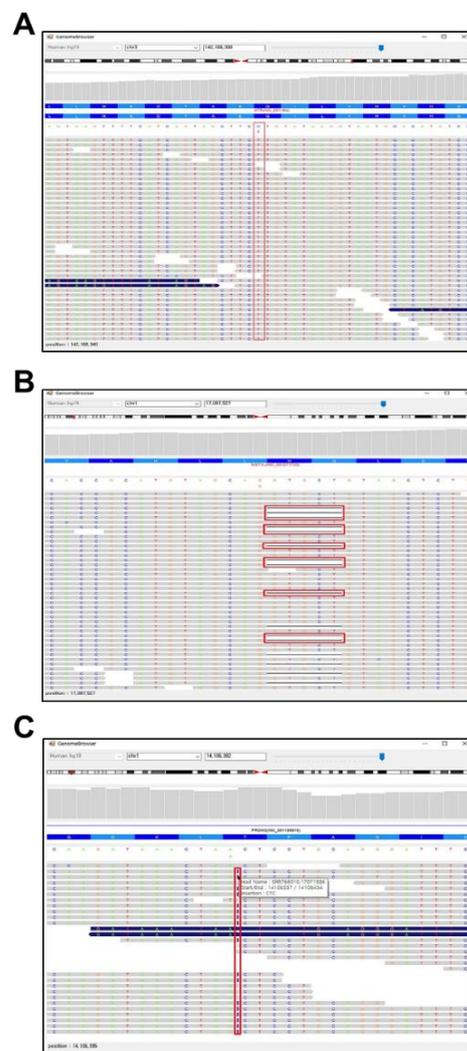

Figure 4. Identification of genetic variants in a genome browser with a single-nucleotide resolution. (**A**) Single nucleotide polymorphism (SNP) in an *ATR* gene. (**B**) Deletion in a *MST1L* gene. (**C**) Insertion of the sequences in a *PRDM2* gene. The positions of genetic variants are indicated with red boxes.

4. Discussion

In this study, we developed a knowledge-based PhenGenVar browser for disease-gene variation visualization for clinical diagnosis. Using personal whole-genome sequencing data, we proved that the PhenGenVar browser is useful for identifying genetic variations, such as SNPs, insertions, and deletions of specific genes from the user-defined gene set. Compared to other genome browsers, such as IGV, which were designed to monitor genetic variations of all genomic regions from large-scale datasets, our program was specialized to quickly detect the genetic variations of genes of interest from the patient's exome or whole-genome sequencing data. Moreover, because the PhenGenVar can be installed on a personal computer, the security of the patient's personal information can be better guaranteed.

The advent of high-throughput NGS technology has accelerated the discovery of common and rare disease-associated genetic variants in large population genomic data [53,54]. Based on this information, personal exome or whole-genome sequencing has been widely used to identify genetic variations and their associations with various human diseases [54,55]. For instance, *BRCA1* and *BRCA2* are partially responsible for breast and ovarian cancers, and their SNVs are widely used for cancer diagnosis [56]. Since our knowledge of the genes related to a specific disease phenotype has greatly increased, it is necessary to diagnose a patient's disease using a precise and fast method. The PhenGenVar application can be used as a tool for this purpose. Compared to previous tools, one advantage of the PhenGenVar program is that the exome browser is specifically designed to quickly monitor the genetic variations in each exon and its exon–intron junction area, which is critical for the deleterious mutations of disease-responsible genes. To prove this, we showed that the application could easily detect genetic variations in specific genes using randomly selected whole-genome sequencing data from an anonymous person. Thus, we believe that the PhenGenVar can be widely used for clinical diagnosis utilizing the relatively cost-effective exome sequencing data. Moreover, this method can be applied to custom-targeted gene panel sequencing data [57].

The main objective of PhenGenVar application is to quickly identify genetic variations from a knowledge-based gene set. Thus, the precise diagnosis using this application relies on the quality of the gene set associated with a specific disease. The gene sets from the HPO were pre-installed in the current version of PhenGenVar. We will update more disease gene sets upon user request in the next version.

5. Conclusions

PhenGenVar is a freely available program developed with the aim of supporting clinical diagnosis using the patient's NGS data. It is designed to monitor the genetic variations of selected gene lists that are associated with specific disease and phenotype. This user-friendly program is developed for researchers or physicians to analyze the patient's genetic data with no programing knowledge requirement. Because the PhenGenVar browsers provide the comprehensive GUI-based VCF visualization, it allows users to manage, filter, query, and export the variant results in a fast and effective way (Supplementary Figure S1). We expect that this tool can be widely used by the medical physicians to diagnose the patient's diseases from their genomic data.

Author Contributions: Conceptualization, Y.J.K., D.-H.J. and J.Y.; methodology, J.S., J.J., D.J. and K.K.; software, J.S., J.J., D.J. and K.K.; writing—original draft preparation, J.S., D.-H.J. and J.Y.; writing—review and editing, J.S., D.-H.J. and J.Y.; visualization, J.S., J.J., D.-H.J. and J.Y.; supervision, D.-H.J. and J.Y.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

References

1. Collins, F.S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795. [CrossRef]
2. Human Genome Sequencing Consortium, International. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*, 931–945. [CrossRef] [PubMed]
3. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef] [PubMed]
4. Easton, D.F.; Pharoah, P.D.; Antoniou, A.C.; Tischkowitz, M.; Tavtigian, S.V.; Nathanson, K.L.; Devilee, P.; Meindl, A.; Couch, F.J.; Southey, M.; et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **2015**, *372*, 2243–2257. [CrossRef]
5. Smedley, D.; Jacobsen, J.O.; Jager, M.; Kohler, S.; Holtgrewe, M.; Schubach, M.; Siragusa, E.; Zemojtel, T.; Buske, O.J.; Washington, N.L.; et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **2015**, *10*, 2004–2015. [CrossRef] [PubMed]
6. Choi, M.; Scholl, U.I.; Ji, W.; Liu, T.; Tikhonova, I.R.; Zumbo, P.; Nayir, A.; Bakkaloglu, A.; Ozen, S.; Sanjad, S.; et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19096–19101. [CrossRef]
7. 1000 Genomes Project Consortium; Abecasis, G.R.; Altshuler, D.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Gibbs, R.A.; Hurles, M.E.; McVean, G.A. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073. [CrossRef]
8. Lambert, J.C.; Ibrahim-Verbaas, C.A.; Harold, D.; Naj, A.C.; Sims, R.; Bellenguez, C.; DeStafano, A.L.; Bis, J.C.; Beecham, G.W.; Grenier-Boley, B.; et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **2013**, *45*, 1452–1458. [CrossRef]
9. Lu, X.; Wang, L.; Chen, S.; He, L.; Yang, X.; Shi, Y.; Cheng, J.; Zhang, L.; Gu, C.C.; Huang, J.; et al. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nat. Genet.* **2012**, *44*, 890–894. [CrossRef]
10. Sladek, R.; Rocheleau, G.; Rung, J.; Dina, C.; Shen, L.; Serre, D.; Boutin, P.; Vincent, D.; Belisle, A.; Hadjadj, S.; et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **2007**, *445*, 881–885. [CrossRef]
11. Corder, E.H.; Saunders, A.M.; Strittmatter, W.J.; Schmechel, D.E.; Gaskell, P.C.; Small, G.W.; Roses, A.D.; Haines, J.L.; Pericak-Vance, M.A. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **1993**, *261*, 921–923. [CrossRef] [PubMed]
12. Guerreiro, R.; Wojtas, A.; Bras, J.; Carrasquillo, M.; Rogaeva, E.; Majounie, E.; Cruchaga, C.; Sassi, C.; Kauwe, J.S.; Younkin, S.; et al. TREM2 variants in Alzheimer's disease. *N. Engl. J. Med.* **2013**, *368*, 117–127. [CrossRef] [PubMed]
13. Jonsson, T.; Stefansson, H.; Steinberg, S.; Jonsdottir, I.; Jonsson, P.V.; Snaedal, J.; Bjornsson, S.; Huttenlocher, J.; Levey, A.I.; Lah, J.J.; et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **2013**, *368*, 107–116. [CrossRef] [PubMed]
14. Harold, D.; Abraham, R.; Hollingworth, P.; Sims, R.; Gerrish, A.; Hamshere, M.L.; Pahwa, J.S.; Moskvina, V.; Dowzell, K.; Williams, A.; et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **2009**, *41*, 1088–1093. [CrossRef]
15. Hollingworth, P.; Harold, D.; Sims, R.; Gerrish, A.; Lambert, J.C.; Carrasquillo, M.M.; Abraham, R.; Hamshere, M.L.; Pahwa, J.S.; Moskvina, V.; et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.* **2011**, *43*, 429–435. [CrossRef]
16. Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **2000**, *83*, 1301–1308. [CrossRef]
17. Peto, J.; Collins, N.; Barfoot, R.; Seal, S.; Warren, W.; Rahman, N.; Easton, D.F.; Evans, C.; Deacon, J.; Stratton, M.R. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J. Natl. Cancer Inst.* **1999**, *91*, 943–949. [CrossRef]
18. Antoniou, A.; Pharoah, P.D.; Narod, S.; Risch, H.A.; Eyfjord, J.E.; Hopper, J.L.; Loman, N.; Olsson, H.; Johannsson, O.; Borg, A.; et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* **2003**, *72*, 1117–1130. [CrossRef]
19. Torkamani, A.; Wineinger, N.E.; Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **2018**, *19*, 581–590. [CrossRef]

20. Marc, D.T.; Khairat, S.S. Medical Subject Headings (MeSH) for indexing and retrieving open-source healthcare data. *Stud. Health Technol. Inform.* **2014**, *202*, 157–160.

21. Noy, N.F.; de Coronado, S.; Solbrig, H.; Fragoso, G.; Hartel, F.W.; Musen, M.A. Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages. *Appl. Ontol.* **2008**, *3*, 173–190. [CrossRef] [PubMed]

22. Spackman, K. SNOMED RT and SNOMEDCT. Promise of an international clinical terminology. *MD Comput.* **2000**, *17*, 29. [PubMed]

23. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef] [PubMed]

24. Kohler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef] [PubMed]

25. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2005**, *33*, D514–D517. [CrossRef]

26. Poplin, R.; Chang, P.C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987. [CrossRef]

27. Rausch, T.; Zichner, T.; Schlattl, A.; Stutz, A.M.; Benes, V.; Korbel, J.O. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, i333–i339. [CrossRef]

28. Li, H. FermiKit: Assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **2015**, *31*, 3694–3696. [CrossRef]

29. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]

30. Escaramis, G.; Docampo, E.; Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genom.* **2015**, *14*, 305–314. [CrossRef]

31. Rimmer, A.; Phan, H.; Mathieson, I.; Iqbal, Z.; Twigg, S.R.F.; Consortium, W.G.S.; Wilkie, A.O.M.; McVean, G.; Lunter, G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **2014**, *46*, 912–918. [CrossRef] [PubMed]

32. Kim, S.; Scheffler, K.; Halpern, A.L.; Bekritsky, M.A.; Noh, E.; Kallberg, M.; Chen, X.; Kim, Y.; Beyter, D.; Krusche, P.; et al. Strelka2: Fast and accurate calling of germline and somatic variants. *Nat. Methods* **2018**, *15*, 591–594. [CrossRef] [PubMed]

33. Koboldt, D.C.; Chen, K.; Wylie, T.; Larson, D.E.; McLellan, M.D.; Mardis, E.R.; Weinstock, G.M.; Wilson, R.K.; Ding, L. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **2009**, *25*, 2283–2285. [CrossRef] [PubMed]

34. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]

35. Wang, W.; Hu, W.; Hou, F.; Hu, P.; Wei, Z. SNVerGUI: A desktop tool for variant analysis of next-generation sequencing data. *J. Med. Genet.* **2012**, *49*, 753–755. [CrossRef]

36. Ou, M.; Ma, R.; Cheung, J.; Lo, K.; Yee, P.; Luo, T.; Chan, T.L.; Au, C.H.; Kwong, A.; Luo, R.; et al. database.bio: A web application for interpreting human variations. *Bioinformatics* **2015**, *31*, 4035–4037. [CrossRef]

37. Pandey, R.V.; Pabinger, S.; Kriegner, A.; Weinhausel, A. DaMold: A data-mining platform for variant annotation and visualization in molecular diagnostics research. *Hum. Mutat.* **2017**, *38*, 778–787. [CrossRef]

38. Li, Z.; Liu, Z.; Jiang, Y.; Chen, D.; Ran, X.; Sun, Z.S.; Wu, J. mirVAFC: A Web Server for Prioritizations of Pathogenic Sequence Variants from Exome Sequencing Data via Classifications. *Hum. Mutat.* **2017**, *38*, 25–33. [CrossRef]

39. Van der Velde, K.J.; de Boer, E.N.; van Diemen, C.C.; Sikkema-Raddatz, B.; Abbott, K.M.; Knopperts, A.; Franke, L.; Sijmons, R.H.; de Koning, T.J.; Wijmenga, C.; et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol.* **2017**, *18*, 6. [CrossRef]

40. Lee, I.H.; Lee, K.; Hsing, M.; Choe, Y.; Park, J.H.; Kim, S.H.; Bohn, J.M.; Neu, M.B.; Hwang, K.B.; Green, R.C.; et al. Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. *Hum. Mutat.* **2014**, *35*, 537–547. [CrossRef]

41. Hart, S.N.; Duffy, P.; Quest, D.J.; Hossain, A.; Meiners, M.A.; Kocher, J.P. VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. *Brief. Bioinform.* **2016**, *17*, 346–351. [CrossRef] [PubMed]

42. Muller, H.; Jimenez-Heredia, R.; Krolo, A.; Hirschmugl, T.; Dmytrus, J.; Boztug, K.; Bock, C. VCF.Filter: Interactive prioritization of disease-linked genetic variants from sequencing data. *Nucleic Acids Res.* **2017**, *45*, W567–W572. [CrossRef] [PubMed]

43. Pietrelli, A.; Valenti, L. myVCF: A desktop application for high-throughput mutations data management. *Bioinformatics* **2017**, *33*, 3676–3678. [CrossRef] [PubMed]

44. Salatino, S.; Ramraj, V. BrowseVCF: A web-based application and workflow to quickly prioritize disease-causative variants in VCF files. *Brief. Bioinform.* **2017**, *18*, 774–779. [CrossRef]

45. Jiang, J.; Gu, J.; Zhao, T.; Lu, H. VCF-Server: A web-based visualization tool for high-throughput variant data mining and management. *Mol. Genet. Genom. Med.* **2019**, *7*, e00641. [CrossRef]

46. Robinson, J.T.; Thorvaldsdottir, H.; Wenger, A.M.; Zehir, A.; Mesirov, J.P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **2017**, *77*, e31–e34. [CrossRef]

47. Navarro Gonzalez, J.; Zweig, A.S.; Speir, M.L.; Schmelter, D.; Rosenbloom, K.R.; Raney, B.J.; Powell, C.C.; Nassar, L.R.; Maulding, N.D.; Lee, C.M.; et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **2021**, *49*, D1046–D1057. [CrossRef]

48. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]

49. Sayers, E.W.; Bolton, E.E.; Brister, J.R.; Canese, K.; Chan, J.; Comeau, D.C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2022**, *50*, D20–D26. [CrossRef]

50. Genomes Project, C.; Abecasis, G.R.; Auton, A.; Brooks, L.D.; DePristo, M.A.; Durbin, R.M.; Handsaker, R.E.; Kang, H.M.; Marth, G.T.; McVean, G.A. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *491*, 56–65. [CrossRef]

51. Povey, S.; Lovering, R.; Bruford, E.; Wright, M.; Lush, M.; Wain, H. THE HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.* **2001**, *109*, 678–680. [CrossRef] [PubMed]

52. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]

53. Boycott, K.M.; Vanstone, M.R.; Bulman, D.E.; MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* **2013**, *14*, 681–691. [CrossRef]

54. Boycott, K.M.; Hartley, T.; Biesecker, L.G.; Gibbs, R.A.; Innes, A.M.; Riess, O.; Belmont, J.; Dunwoodie, S.L.; Jojic, N.; Lassmann, T.; et al. A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell* **2019**, *177*, 32–37. [CrossRef]

55. Goodrich, J.K.; Singer-Berk, M.; Son, R.; Sveden, A.; Wood, J.; England, E.; Cole, J.B.; Weisburd, B.; Watts, N.; Caulkins, L.; et al. Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat. Commun.* **2021**, *12*, 3505. [CrossRef] [PubMed]

56. Abul-Husn, N.S.; Soper, E.R.; Odgis, J.A.; Cullina, S.; Bobo, D.; Moscati, A.; Rodriguez, J.E.; Team, C.G.; Regeneron Genetics, C.; Loos, R.J.F.; et al. Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. *Genome Med.* **2019**, *12*, 2. [CrossRef]

57. Ogiso-Tanaka, E.; Shimizu, T.; Hajika, M.; Kaga, A.; Ishimoto, M. Highly multiplexed AmpliSeq technology identifies novel variation of flowering time-related genes in soybean (Glycine max). *DNA Res.* **2019**, *26*, 243–260. [CrossRef] [PubMed]