An Integrated Framework for Analysis and Prediction of Impact of Single Nucleotide Polymorphism Associated with Human Diseases

Evolutionary Bioinformatics
Volume 20: 1–14
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769343241249916
Sage

Syed Shah Muhammad[1], Muhammad Shoaib[1]
and Muhammad Tariq Pervez[2]

[1]Department of Computer Science, University of Engineering & Technology, Lahore, Punjab, Pakistan. [2]Department of Biological Sciences, Virtual University of Pakistan, Lahore, Punjab, Pakistan.

ABSTRACT: Single nucleotide polymorphisms are most common type of genetic variation in human genome. Analyzing genetic variants can help us better understand the genetic basis of diseases and develop predictive models which are useful to identify individuals who are at increased risk for certain diseases. Several SNP analysis tools have already been developed. For running these tools, the user needs to collect data from various databases. Secondly, often researchers have to use multiple variant analysis tools for cross validating their results and increase confidence in their findings. Extracting data from multiple databases and running multiple tools at a time, increases complexity and time required for analysis. There are some web-based tools that integrate multiple genetic variant databases and provide variant annotations for a few tools. These approaches have some limitations such as retrieving annotation information, filtering common pathogenic variants. The proposed web-based tool, namely IPSNP: An Integrated Platform for Predicting Impact of SNPs is written in Django which is a python-based framework. It uses RESTful API of MyVariant.info to extract annotation information of variants associated with a given gene, rsID, HGVS format variants specified in a VCF file for 29 tools. The results are in the form of a CSV file of predictions (1) derived from the consensus decision, (2) a file having annotations for the variants associated with the given gene, (3) a file showing variants declared as pathogenic commonly by the selected tools, and (4) a CSV file containing chromosome coordinates based on GRCh37 and GRCh38 genome assemblies, rsIDs and proteomic data, so that users may use tools of their choice and avoiding manual parameter collection for each tool. IPSNP is a valuable resource for researchers and clinicians and it can help to save time and effort in discovering the novel disease-associated variants and the development of personalized treatments.

KEYWORDS: SNP analysis, SNPs prediction, automated tools, prediction tools evaluation, polymorphism

RECEIVED: November 28, 2023. **ACCEPTED:** April 10, 2024.

TYPE: Original Research Article

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Muhammad Tariq Pervez, Department of Biological Sciences, Virtual University of Pakistan, Satellite Town, Saidpur Road, College Chowk, Rawalpinid. Email: m.tariq@vu.edu.pk

Introduction

Single nucleotide polymorphisms (SNPs) play an important role in human genetic diversity, disease susceptibility, and drug responses. Several SNP analysis tools have already been developed, each with its strengths and weaknesses. Some popular SNP analysis tools are SIFT (Sorting Intolerant from Tolerant),[1] PROVEAN (Protein Variation Effect Analyzer),[2] MVP (Multivariate Prediction),[3] PolyPhen-2 HVAR (Polymorphism Phenotyping v2-Highly-Confident Missense Variant),[4] PolyPhen-2, HDIV (HumDiv, PrimateAI, REVEL (Rare Exome Variant Ensemble Learner),[5] MPC (Missense Prediction and Classification). There are multiple research domains where more than one tool is used for SNP analysis such as (i) genome-wide association studies (GWAS)[6] which involves scanning the entire genome of individuals to identify common genetic variants that are associated with a particular trait or disease. (ii) whole genome sequencing (WGS)[7] which encompasses sequencing the entire genome of an individual, providing a complete picture of all genetic variants present in the genome. This approach can be used to identify rare or novel genetic variants that may be missed by other methods and can also provide insight into the functional consequences of genetic variation. (iii) Population genetic studies employ SNP analysis to identify patterns of genetic variation within and between populations, providing insight into the evolutionary history and migration patterns of various groups.

It is beneficial to analyze SNPs from several tools, as it helps to improve accuracy and reliability of the obtained results, highlight the error observations or biases in the results, and have a deep understanding to genetic variations under study.

PredictSNP[8] is an online platform that predicts the impact of nsSNPs on protein structure and function. This platform combines a machine learning model for prediction and some other SNP prediction tools for annotating the SNPs. It is an efficient tool with the limitation that input needs to be calculated manually for each annotation. This may be a time taking task. Secondly, the authors of PredictSNP were unable to integrate a reasonable number of tools on this platform. This tool lacks the filtering of all results into pathogenic or benign groups. Capriotti et al[9] developed a platform named MetaSNP which integrates 4 SNP analysis tools namely SIFT, PhDSNP (Predictor of human Deleterious Single Nucleotide Polymorphism),[10] SNAP (an integrated SNP annotation platform),[11] and PANTHER (Protein ANalysis THrough

Evolutionary Relationships).[12] This platform has the same problem of manual input preparation. Result filtering into common pathogen and benign variants was not available. The dbNSFP (Database of Human Non-synonymous SNPs and their functional predictions)[13] is an integrated platform which provides annotation of SNPs executed by 36 tools. However, the provided results are not user friendly and the user cannot obtain common pathogenic variants. SNPnexus[14] is a web server developed by Oscanoa (et al 2020). This web server provides annotation by 8 tools namely FitCons (fitness consequences),[15] DeepSEA,[16] CADD (Combined Annotation Dependent Depletion),[17] Eigen,[18] FATHMM-MKL[19] (Functional Analysis through Hidden Markov Models—multiple kernel learning), FunSeq2,[20] GWAVA (genome wide annotation of variants)[21] and ReMM (regulatory Mendelian mutation).[22] However, this also has the same limitations as described for other tools.

Keeping in view the limitations of available integrated platforms, the proposed web-based tool namely "IPSNP: An Integrated Platform for Predicting Impact of SNPs" was developed in Django which is a python-based framework. It uses RESTful API of NCBI and MyVariant.info for obtaining required data/annotations of all variants associated with a given gene. The input for IPSNP is GeneID, rsID, HGVS format variants or a VCF file for the selected tools out 29 tools. The results include: (i) a CSV file comprising of IPSNP predictions derived from the consensus decisions of the selected SNP analysis tools (ii) a CSV file for each of the selected tool having annotation information for the variants associated with the given gene (iii) a CSV file showing variants declared as pathogenic commonly by the selected tools (iv) a CSV file(s) containing chromosome coordinates based on GRCh37 and GRCh38 genome assemblies, rsIDs and proteomic data including UniProt id/name, protein sequence, RefSeq protein id and amino acid substitutions so that the users may use tools directly of their choice. Now the user does not need to collect the input parameters manually for each tool from other databases like NCBI and Uniprot.

Problem Statement

The major problem identified from the literature is that input calculation for the SNP analysis tools is performed manually. It takes a lot of time and effort. A user must perform a number of steps in order to prepare input for SNP analysis tools. These steps may involve selecting pathogenic or benign SNPs as reported on the ClinVar.[23] Different SNP predictions tools have different input parameters. A prediction server may require gene ID, rsID, chromosome number, position and orientation, wild type allele, mutated allele, amino acid positions, protein accession ID and protein sequence in FASTA (Fast All) format. All these parameters are not available on any single platform or a website. A user has to collect all the required input data for a selected SNP prediction tool from various websites and databases. A user has to visit ClinVar, National Center for Biotechnology Information (NCBI) database of SNPs (dbSNP). Uniprot for protein sequences or any other database, as required by the prediction tools.

For example, if a user wants to execute SNP analysis through SIFT prediction tool, the parameters required to perform the predictions are chromosome number, chromosome coordinates, chromosome orientation, wild type and mutant alleles. The input for SIFT looks like 1,41285565,1,G/A. Collecting the parameters for hundreds of SNPs is very cumbersome job.

Second, the output from different SNP analysis tools is in different formats. The terminology for each output is different, for example, some of the tools use the term pathogenic, some use as damaging and others as deleterious. Similarly for benign variations, the terms non-damaging or neutral are used. Some of the tools even calculate the score and the user must determine whether the given score lies in the pathogenic or benign based on the given benchmark. The problem is to filter out the result in a consistent format to make the decision. All this work is done manually and can create inconsistencies in the input data.

Third, no SNP analysis tool provides common pathogenic variants. The user has to perform this task manually which is difficult and time consuming task especially if the results have been compiled from several tools. IPSNP provides a comprehensive solution to these problems by providing an easy to use, user friendly and efficient tool to obtain variant's data, predictions of various tools and consensus result of IPSNP.

Materials & Methods

Integrated tools

The SNP prediction tools are available in 2 forms. First, SNP prediction tools are available as command base version such as dbNSFP, SIFT, PROVEAN, and Polyphen 2. Second the online web-servers such as PredictSNP, MetaSNP, SNPs&Go,[24] and Mutation Assessor.[25] In command base versions of SNP prediction tools, it is difficult to download and install/configure it on the local machine. This is not possible for an individual researcher as it involves high costs in terms of time and resources. The second option is to run online tools for the required predictions. In case 2, each tool has different input parameters and formats. Preparing input for several individual tools is a time taking job and may lead to inconsistent inputs.

The scope of tools selected (Table 1) for IPSNP platform is limited to the tools which can be accessed/run through APIs. Therefore, only those tools are embedded in this platform whose APIs are exposed for the accessibility from other servers.

Comparison IPSNP with existing prediction tools

The required genetic SNP variant data was downloaded from ClinVar and the Prtein IDs were collected from the Uniprot databases which are publicly available. We selected 100 benign

Table 1. List of SNP prediction tools integrated in IPSNP.

SR. NO	PREDICTION TOOL	COMPLETE NAME	LINK
1	SIFT	Sorting intolerant from tolerant	https://sift.bii.a-star.edu.sg/www/Extended_SIFT_chr_coords_submit.html
2	SIFT4G	Sorting intolerant from tolerant 4G	https://sift.bii.a-star.edu.sg/www/SIFT_dbSNP.html
3	PROVEAN	Protein variation effect analyzer	http://provean.jcvi.org/index.php
4	MVP	Missense variant pathogenicity prediction	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7820281/
5	Polyphen2_HVAR	Polymorphism phenotyping v2	http://genetics.bwh.harvard.edu/pph2/index.shtml
6	Polyphen2_HDIV	Polymorphism phenotyping v2	http://genetics.bwh.harvard.edu/pph2/index.shtml
7	PrimateAI	PrimateAI	https://www.nature.com/articles/s41588-018-0167-z
8	REVEL	Rare exome variant ensemble learner	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065685/
9	MPC	Missense badness, PolyPhen-2, and constrain	efaidnbmnnnibpcajpcglclefindmkaj/https://www.biorxiv.org/content/10.1101/148353v1.full.pdf
10	MutPred	MutPred	http://mutpred.mutdb.org/
11	MutationTaster	MutationTaster	https://www.mutationtaster.org/
12	MutationAssessor	MutationAssessor	http://mutationassessor.org/r3/
13	MetaRNN	MetaRNN	http://www.liulab.science/metarnn.html
14	MetaSVM	Meta-analytic support vector machine	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5270233/
15	M-CAP	Mendelian clinically applicable pathogenicity	http://bejerano.stanford.edu/mcap/
16	LIST-S2	LIST-S2	https://list-s2.msl.ubc.ca/;jsessionid=2F4AE6DF20C0D193843105AC54402693?session=2F4AE6DF20C0D193843105AC54402693
17	FATHMM	Functional analysis through hidden Markov model	https://fathmm.biocompute.org.uk/disease.html
18	FATHMM-XF	Functional analysis through hidden Markov model extra features	https://fathmm.biocompute.org.uk/fathmm-xf/
19	FATHMM-MKL	Functional analysis through hidden Markov model multiple kernel learning	https://fathmm.biocompute.org.uk/fathmmMKL.htm
20	PERCH BayesDel(addAF)	Polymorphism evaluation, ranking, and classification for a heritable trait	https://pubmed.ncbi.nlm.nih.gov/27995669/
21	BayesDel(noAF)	Polymorphism evaluation, ranking, and classification for a heritable trait	https://pubmed.ncbi.nlm.nih.gov/27995669/
22	VEST4	Variant effect scoring tool	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3665549/
23	DANN	Deleterious annotation of genetic variants using neural networks	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341060/
24	Eigen	Eigen score	https://pubmed.ncbi.nlm.nih.gov/26482676/
25	Eigen-PC	Eigen-PC	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4731313/
26	DEOGEN2	DEOGEN2	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570203/
27	GenoCanyon	GenoCanyon	https://www.nature.com/articles/srep10576
28	CADD	Combined annotation dependent depletion	https://academic.oup.com/nar/article/47/D1/D886/5146191
29	SnpEff	SNP effect	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/

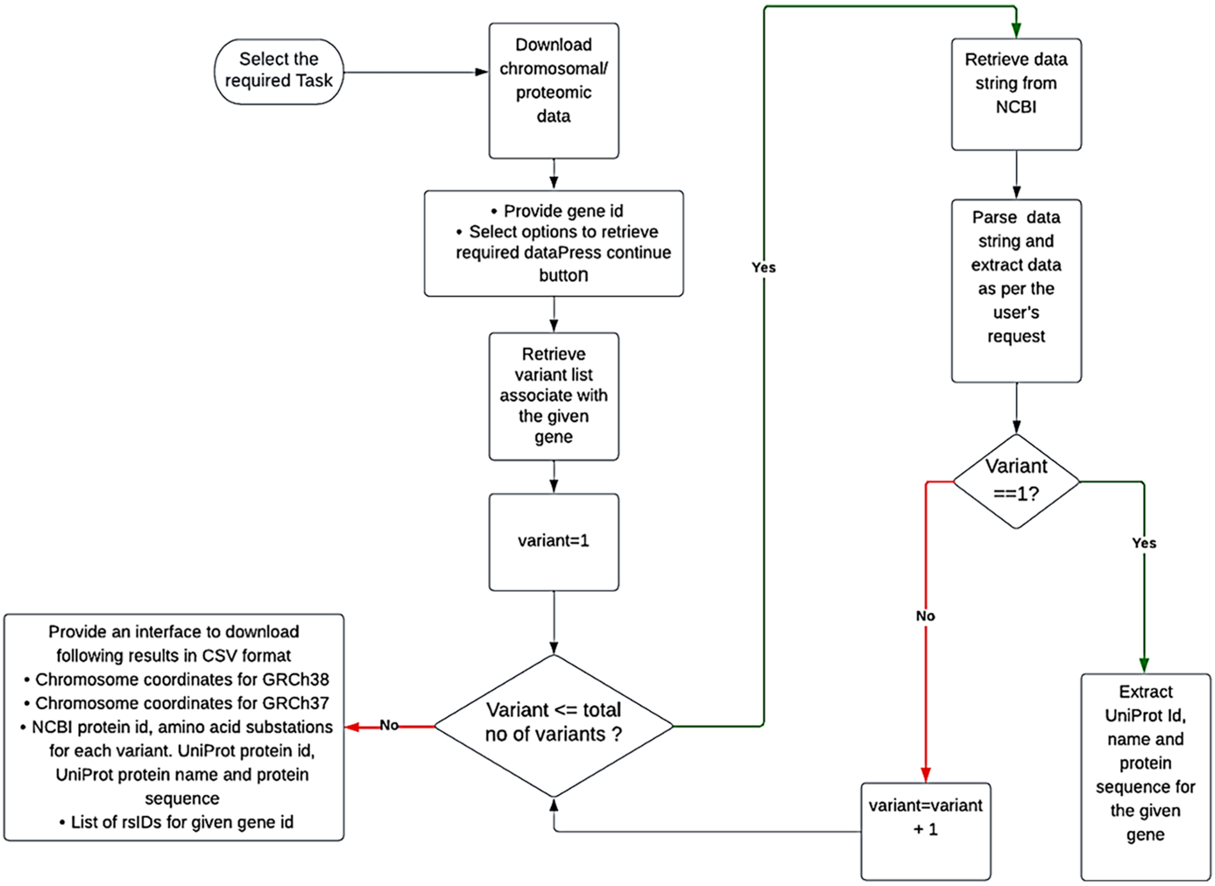

Select the required Task
Download chromosomal/ proteomic data
• Provide gene id
• Select options to retrieve required dataPress continue button
Retrieve variant list associate with the given gene
variant=1
Variant <= total no of variants ?
No
Provide an interface to download following results in CSV format
• Chromosome coordinates for GRCh38
• Chromosome coordinates for GRCh37
• NCBI protein id, amino acid substations for each variant. UniProt protein id, UniProt protein name and protein sequence
• List of rsIDs for given gene id
Yes
Retrieve data string from NCBI
Parse data string and extract data as per the user's request
Variant ==1?
No
Yes
variant=variant + 1
Extract UniProt Id, name and protein sequence for the given gene

Figure 1. Extraction of chromosomal and proteomic data.

variations and 100 pathogenic variations, as reported on ClinVar. The variations were formatted in the proper format as required by all the tools, separately. It contains Gene ID, Protein change, Chromosome number and its orientation, variation ID, genome assembly, dbSNP ID, wild type allele, and the mutant allele. Raw data can be in one of the following formats depending upon the requirements of selected SNP prediction tools:

GENE(S)	PROTEIN CHANGE	CHROMOSOME	LOCATION	DBSNP ID
USH2A	S5188G	1	215799170	rs58257972

NAME	GENE(S)	PROTEIN CHANGE	VARIATIONID	ALLELEID(S)	DBSNP
NM_000059.4 (BRCA2):c.3G>A (p.Met1Ile)	BRCA2	M1I	51579	66247	rs80358650

Design and implementation

IPSNP provides 3 interfaces; one is for downloading chromosomal and proteomic data (CPD) associated with all the variants of a given gene id. This interface was named "CPDEI" (Chromosomal and Proteomic Data Extraction Interface). It is pronounced as cp-de. Here, the user provides only the gene id and selects the desired data to be obtained and IPSNP provides the required data in CSV format. The complete working of the cpdei is shown in Figure 1.

The second interface, which was termed AEFI (Annotation Extraction and Filtering Interface) allows the user for getting annotation of all variants associated with the given gene id for selected tools. In addition, it also provides common pathogenic variants, the variants for which annotation is not found and an option for extracting CPD for these variants as shown in Figure 2. The third interface termed as IPCP (IPSNP consensus predictions) is for obtaining IPSNP's predictions based on the consensus results of the selected SNP analysis tools shown in Figure 5.

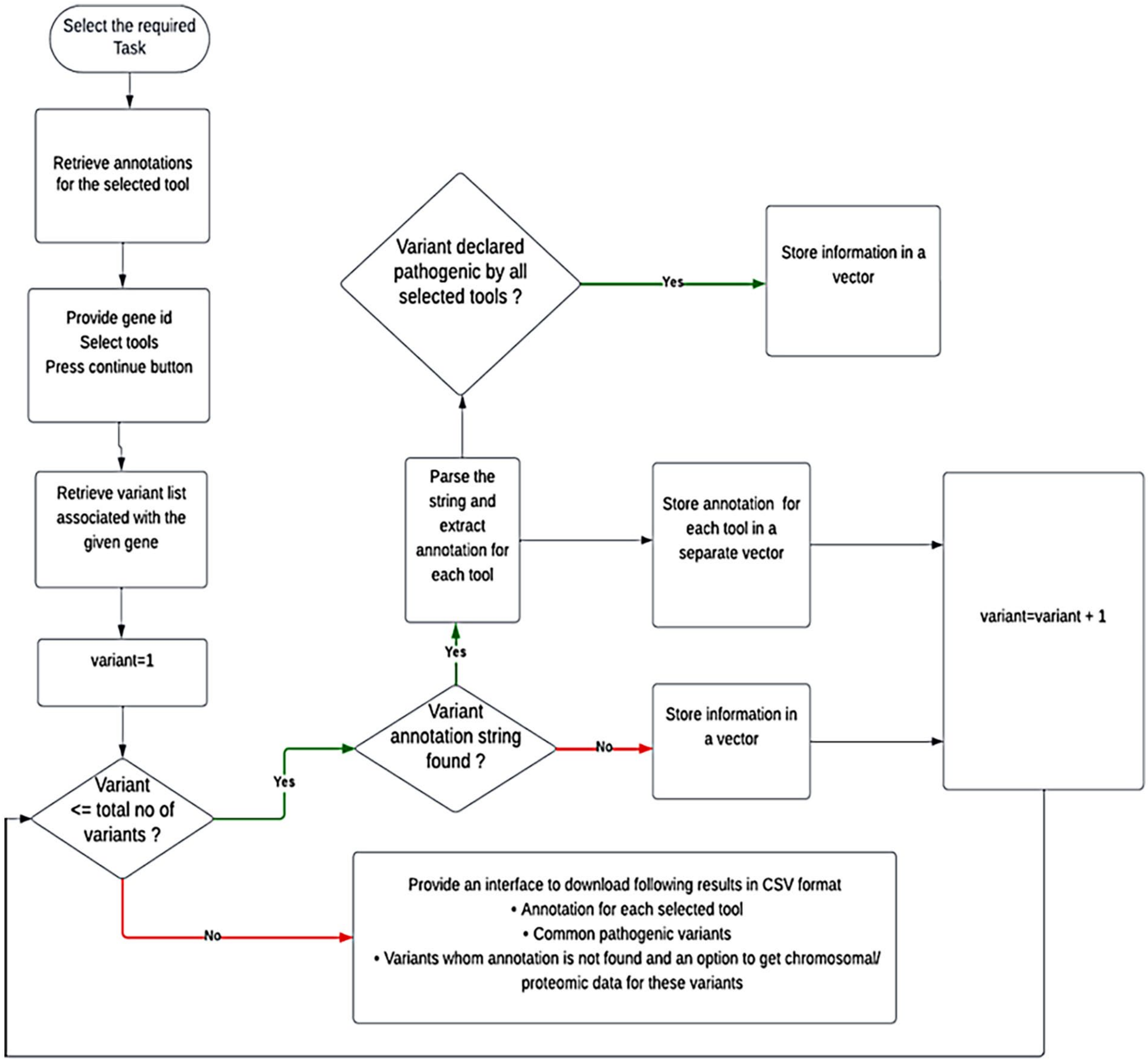

Select the required Task
Retrieve annotations for the selected tool
Provide gene id
Select tools
Press continue button
Retrieve variant list associated with the given gene
variant=1
Variant <= total no of variants ?
Yes
No
Variant annotation string found ?
Yes
No
Parse the string and extract annotation for each tool
Variant declared pathogenic by all selected tools ?
Yes
Store information in a vector
Store annotation for each tool in a separate vector
Store information in a vector
variant=variant + 1
Provide an interface to download following results in CSV format
• Annotation for each selected tool
• Common pathogenic variants
• Variants whom annotation is not found and an option to get chromosomal/ proteomic data for these variants

Figure 2. Retrieving SNP annotations from the selected prediction tools.

IPSNP had been developed using Django 4.1.3 and Python 3.10.5. Biopython 1.81 was used to obtain missense variants from dbSNP.[26] NCBI[27] API was used to download a string and own written python script was used to extract chromosome coordinates for GRCh37 and GRCh38 genome assemblies, and proteomic data including NCBI protein id and amino acid substations for each variant. UniProt API[28] was used to get UniProt protein id, UniProt protein name and protein sequence. Four CSV files are provided to the user: One file stores GRCh38 chromosome coordinates, second file has GRCh37 chromosome coordinates, third file provides proteomic data, and fourth file has rsIDs of all variants associated with the given gene. MyVariant.info API was used for obtaining a string comprising of annotations of variants associated with the given gene. Self-written python script was used to extract annotations for the selected tools and allow the user to download the results in CSV format.

The design of the proposed IPSNP framework is implemented as depicted in the following algorithm:

```
IPSNP (gene_id, SNP_analysis_tools)

   Variants ← fetch_variants_from_clinvar(gene_id)

   for j ← 1 to length[Variants]
      for k ← 1 to length[SNP_analysis_tools]
          tool_specific_variant_data[j, k] ← fetch_tool_specific_variant_data(gene_id,
          Variant[j], SNP_analysis_tools[k])
          tool_prediction[j, k] ← execute(gene_id, tool_specific_variant_data[j, k],
          SNP_analysis_tools[k])

  for j ← 1 to length[Variants]
      variant_prediction[j] ← consensus_prediction(tool_prediction[j])

  return variant_prediction


consensus_prediction (tool_prediction)

      total_count ← 0
      pathogenic_count ← 0

      for j ← 1 to length[tool_prediction]
           if tool_prediction[j] = "pathogenic"
                 then pathogenic_count ← pathogenic_count + 1
           total_count ← total_count + 1

      if (pathogenic_count / total_count) >= 0.5
           then return "pathogenic"
  else
           return "benign"Dataset and Preprocessing
```

To measure accuracy of IPSNP 100 benign and 100 pathogenic SNPs were collected from ClinVar public archive. These selected missense SNPs are reported by different research institutes and reviewed by an expert panel. The reported benign and pathogenic SNPs were downloaded and formatted as per the requirements of the tools chosen to compare them IPSNP.

Tools for comparing with IPSNP

Five most popular SNP prediction tools namely SIFT, Provean, PolyPhen-2, Fathmm, and Mutation assessor were selected to compare their accuracy with the accuracy of IPSNP. Input data was prepared manually for the variants to run on the selected tools. Each tool has different input format and parameters. The input was given to the tools individually and results were calculated.

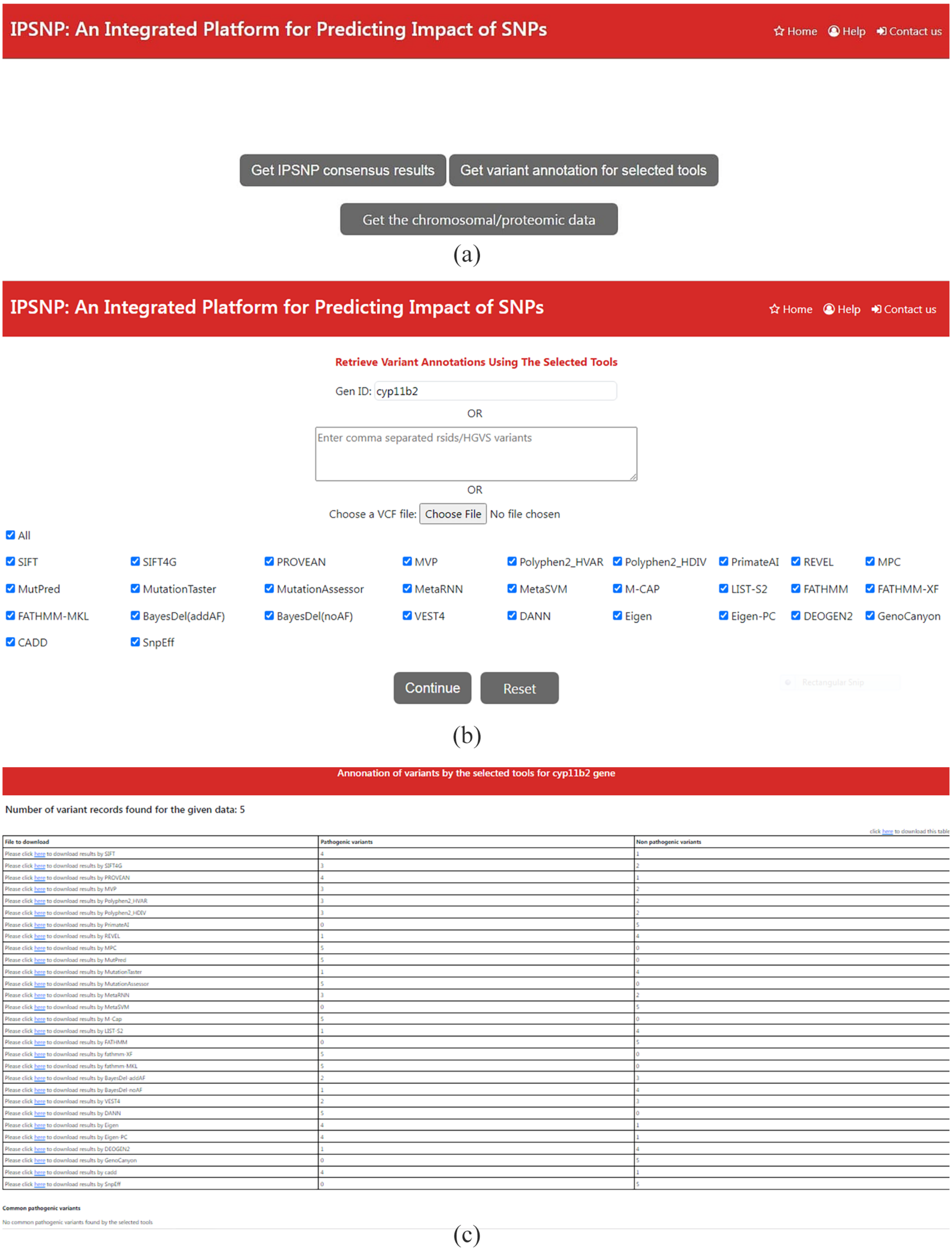
Working of IPSNP

Several SNP analysis tools are provided on IPSNP platform. The given input was executed on all tools. For a given input, IPSNP executes all prediction tools to get the output. Now this output is calculated on the pattern of random forest tree to get the consensus prediction from all these tools. It provides the output in the form of consensus prediction. If a variation is predicted as benign or pathogenic by 15 or more tools then it will be declared as benign pathogenic, accordingly.

Results

This study presented an easy-to-use web application for analyzing missense variants in the human genome and extracting CPD associated with a given gene. The user can select the required option from home page to get chromosomal/proteomic data, annotations for the selected tools or consensus prediction results (Figure 3a). In case of AEFI, (Figure 3b) the user can provide gene id, HGVS or a VCF file as per NCBI format such as BRCA1 (Breast Cancer type 1 susceptibility protein), CYP11B2 (Cytochrome P450 family 11 subfamily B member 2) and Hand1 (Cytochrome P450 family 11 subfamily B member 2) and select the desired tools for which annotations are required. The IPSNP then performs the whole job for the user that is, providing output of each selected tool in a separate CSV file, common pathogenic variants, the variants that are not found by MyVariant.info and an option to get CPD for these variants through CPDEI.

Home interface of IPSNP lets user to select the desired task (Figure 3a). The user can provide the input to get annotation results and can select all or desired tools to get SNP analysis results (Figure 3b). Interface shown by subfigure c provides The output of the AEFI on the next interface (Figure 3c). The tab allows the user to download results of each selected tool (column 1), shows variants declared pathogenic by each tool (column 2), the variants declared non-pathogenic by each tool (column 3). The link on the top right of table allows the user to put the table in CSV format. At the same interface, the user is provided common pathogenic variants by all selected tools.



IPSNP: An Integrated Platform for Predicting Impact of SNPs

☆ Home
Help
Contact us
Get IPSNP consensus results
Get variant annotation for selected tools
Get the chromosomal/proteomic data

(a)

IPSNP: An Integrated Platform for Predicting Impact of SNPs

Retrieve Variant Annotations Using The Selected Tools

Gen ID: cyp11b2

OR

Enter comma separated rsids/HGVS variants

OR

Choose a VCF file: Choose File No file chosen

☑ All

☑ SIFT ☑ SIFT4G ☑ PROVEAN ☑ MVP ☑ Polyphen2_HVAR ☑ Polyphen2_HDIV ☑ PrimateAI ☑ REVEL ☑ MPC

☑ MutPred ☑ MutationTaster ☑ MutationAssessor ☑ MetaRNN ☑ MetaSVM ☑ M-CAP ☑ LIST-S2 ☑ FATHMM ☑ FATHMM-XF

☑ FATHMM-MKL ☑ BayesDel(addAF) ☑ BayesDel(noAF) ☑ VEST4 ☑ DANN ☑ Eigen ☑ Eigen-PC ☑ DEOGEN2 ☑ GenoCanyon

☑ CADD ☑ SnpEff

(b)

Annonation of variants by the selected tools for cyp11b2 gene

Number of variant records found for the given data: 5

click here to download this table

File to download	Pathogenic variants	Non pathogenic variants
Please click here to download results by SIFT	4	1
Please click here to download results by SIFT4G	3	2
Please click here to download results by PROVEAN	4	1
Please click here to download results by MVP	3	2
Please click here to download results by Polyphen2_HVAR	3	2
Please click here to download results by Polyphen2_HDIV	3	2
Please click here to download results by PrimateAI	0	5
Please click here to download results by REVEL	1	4
Please click here to download results by MPC	5	0
Please click here to download results by MutPred	5	0
Please click here to download results by MutationTaster	1	4
Please click here to download results by MutationAssessor	5	0
Please click here to download results by MetaRNN	3	2
Please click here to download results by MetaSVM	0	5
Please click here to download results by M-Cap	5	0
Please click here to download results by LIST-S2	1	4
Please click here to download results by FATHMM	0	5
Please click here to download results by fathmm-XF	5	0
Please click here to download results by fathmm-MKL	5	0
Please click here to download results by BayesDel-addAF	2	3
Please click here to download results by BayesDel-noAF	1	4
Please click here to download results by VEST4	2	3
Please click here to download results by DANN	5	0
Please click here to download results by Eigen	4	1
Please click here to download results by Eigen-PC	4	1
Please click here to download results by DEOGEN2	1	4
Please click here to download results by GenoCanyon	0	5
Please click here to download results by cadd	4	1
Please click here to download results by SnpEff	0	5

Common pathogenic variants

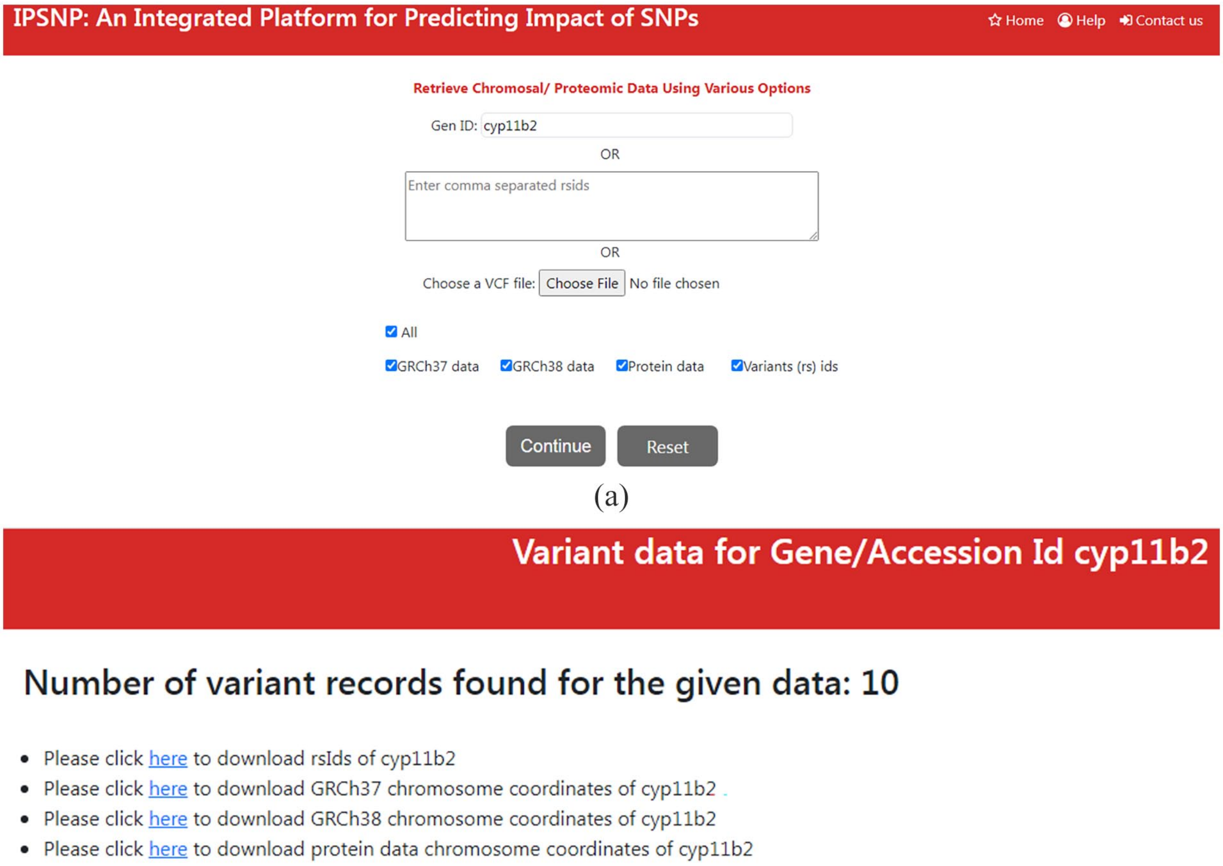
No common pathogenic variants found by the selected tools

(c)

Figure 3. (a) Interface of IPSNP for predicting impact of variants for a given gene (Home page). (b) Interface of IPSNP for predicting impact of variants for a given gene (Tools selection). (c) Interface of IPSNP for predicting impact of variants for a given gene (Output results).

The variants that were not found by MyVarint.info API are also shown on the output page.

The second interface that is, CPDEI has made it very convenient to retrieve chromosome coordinates based on GRCh37 and GRCh38, proteomic data and rsIDs. This information is very useful for running SIFT, PolyPhen and other tools, through their own web interfaces as shown in Figure 4. The CPDEI main interface allows the user to select desired task (Figure 4a). The CPDEI results interface provides results in the form of separate files for each selected tool to be



IPSNP: An Integrated Platform for Predicting Impact of SNPs

Retrieve Chromosal/ Proteomic Data Using Various Options

Gen ID: cyp11b2

OR

Enter comma separated rsids

OR

Choose a VCF file: Choose File No file chosen

☑ All

☑GRCh37 data ☑GRCh38 data ☑Protein data ☑Variants (rs) ids

(a)

Variant data for Gene/Accession Id cyp11b2

Number of variant records found for the given data: 10

- Please click here to download rsIds of cyp11b2
- Please click here to download GRCh37 chromosome coordinates of cyp11b2
- Please click here to download GRCh38 chromosome coordinates of cyp11b2
- Please click here to download protein data chromosome coordinates of cyp11b2

(b)

Figure 4. (a) The interface to provide gene id, choices to be selected and to download required information. (b) The interface to provide gene id, choices to be selected and to download required information.

downloaded by the user (Figure 4b). The user can download rsIDs, GRCh37 chromosome coordinates, GRCh38 chromosome coordinates and proteomic data in CSV format so that it can be easily used in other studies.

The third interface, that is, IPCP analyzes predictions provided by tools and declares a variant as benign or pathogenic based on the predictions by majority of the selected tools. For example, if a user selects 10 tools, IPCP declared a variant as pathogenic if more than 5 tools declare it as a pathogenic variant. This utility of IPSNP is the most accurate as shown in the section entitled "Performance of IPSNP" (Figure 5a). The user can input a gene ID, HGVS or a VCF file to get consensus results (Figure 5b). The output of consensus results can be downloaded in a CSV file (Figure 5c). The details architecture of IPSNP is presented in Figure 6.

Performance of IPSNP

The proposed IPSNP model has novel features that were not available in all existing platforms.

Genomic data provision

Genomic data is provided to user in a CSV file which includes chromosome coordinates based on both GRCh 37 and GRCh 38 genome assemblies. A user is able to download a separate file of proteomic data which contains the variants along with required parameters used by different SNP prediction tools. Data collection from NCBI and Uniprot databases has been automated. This data enables the user to choose a SNP prediction tool without the burden of collecting manual input parameters.

Tool-specific annotations

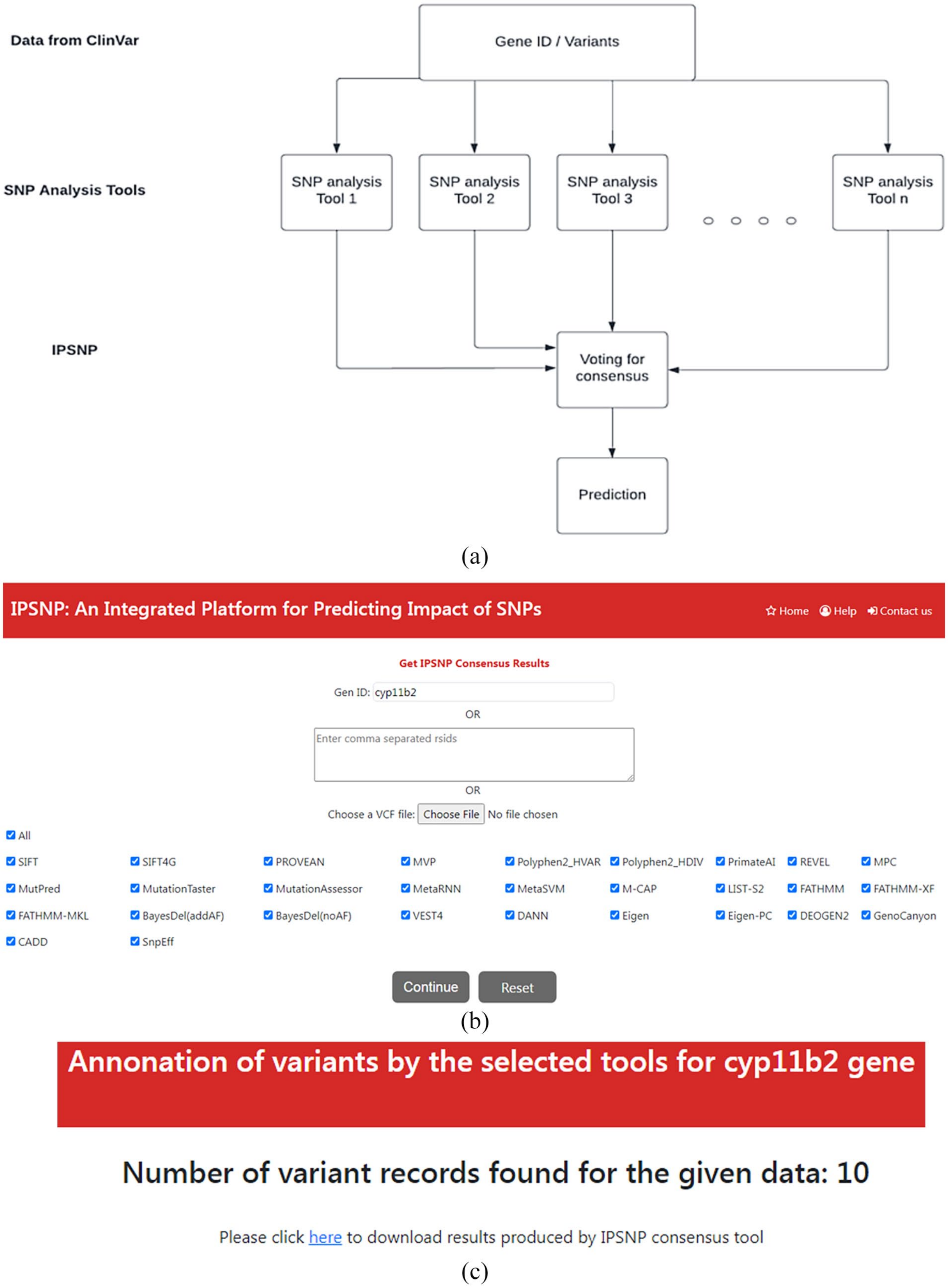
IPSNP provides the segregated annotation results for benign and pathogenic variants in a CSV format for all the variations in a selected gene. In this way, the user is able to study and analyze the annotation results that are provided by each SNP prediction tool separately.

Common pathogenic variants

Our proposed model provides the users with the common pathogenic or benign variants in a CSV file for a given gene. The users are able analyze the significance of these variations using this feature.

Consensus predictions

Finally, our model provides the facility to find out the consensus results in CSV file. The consensus results provide the ease



(a)

IPSNP: An Integrated Platform for Predicting Impact of SNPs

Get IPSNP Consensus Results

Gen ID: cyp11b2

OR

Enter comma separated rsids

OR

Choose a VCF file: Choose File No file chosen

☑ All

☑ SIFT ☑ SIFT4G ☑ PROVEAN ☑ MVP ☑ Polyphen2_HVAR ☑ Polyphen2_HDIV ☑ PrimateAI ☑ REVEL ☑ MPC

☑ MutPred ☑ MutationTaster ☑ MutationAssessor ☑ MetaRNN ☑ MetaSVM ☑ M-CAP ☑ LIST-S2 ☑ FATHMM ☑ FATHMM-XF

☑ FATHMM-MKL ☑ BayesDel(addAF) ☑ BayesDel(noAF) ☑ VEST4 ☑ DANN ☑ Eigen ☑ Eigen-PC ☑ DEOGEN2 ☑ GenoCanyon

☑ CADD ☑ SnpEff

(b)

Annonation of variants by the selected tools for cyp11b2 gene

Number of variant records found for the given data: 10

Please click here to download results produced by IPSNP consensus tool

(c)

Figure 5. (a) IPSNP Consensus Predictions (IPCP). (b) Interface for providing input to get consensus results. (c) Interface to download consensus results.

to see the output results in the form of pathogenic or benign results which are declared as pathogenic or benign by the majority of selected tools.

Different tools apply a separate criteria for the analyzing the mutations. Some of the tools provide distinct categorization of the results such as benign or pathogenic. The other tools

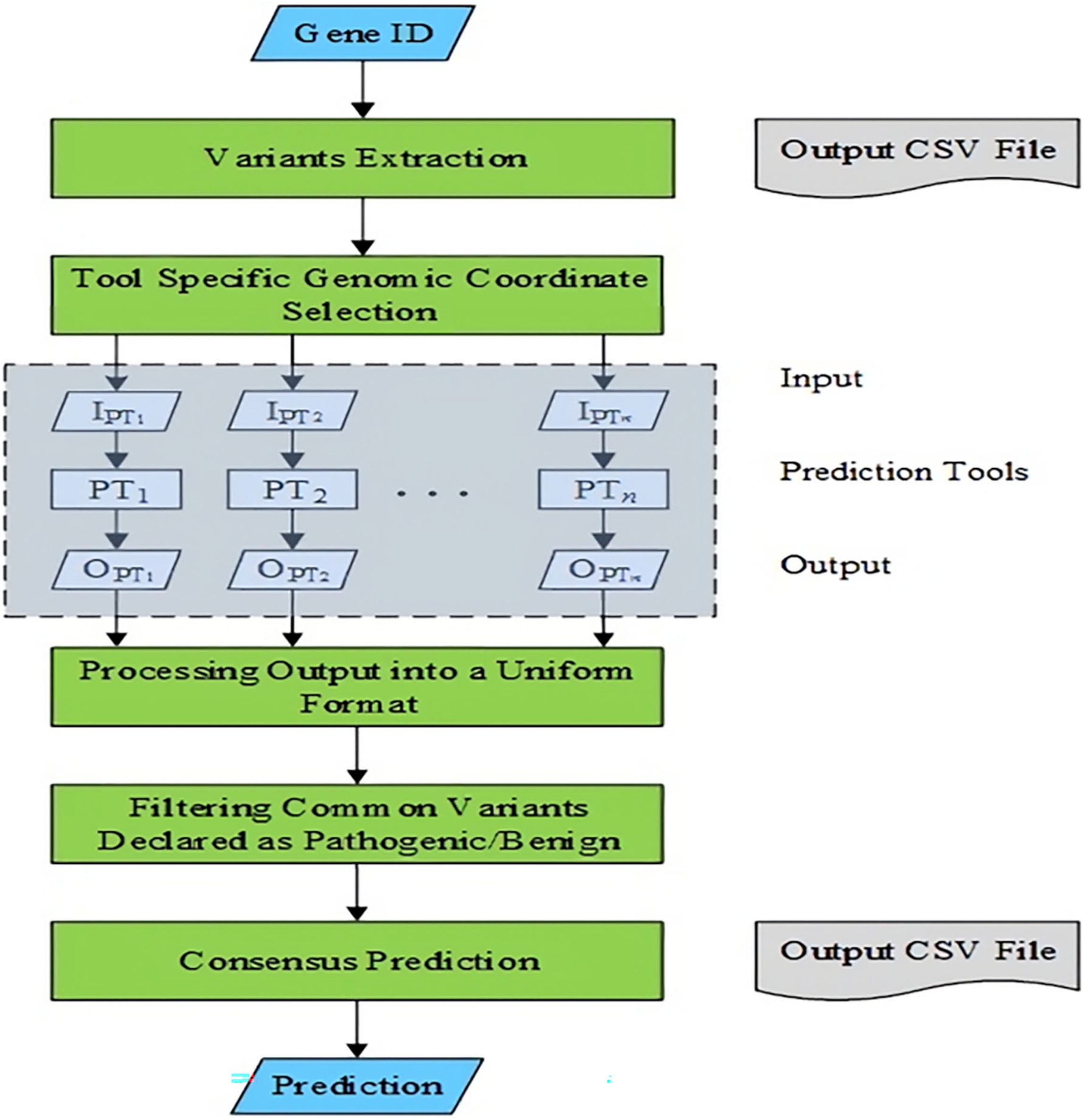


Figure 6. Proposed model of IPSNP.

calculate prediction score which may be a continuous value from 0 to 1. The threshold is set as benign when score is 0 to 0.5 and as damaging from 0.5 to the score as 1. IPSNP takes the adopted the same approach and provides the consensus result on the basis of maximum results as benign or pathogenic. Then the consensus result from all tools is provided.

Criteria for pathogenicity prediction

The criteria to predict pathogenicity of variants are tool specific. Each tool has its own criteria for prediction. As the proposed web server IPSNP is getting predictions from 29 different tools. So, there is no common criterion for finding the pathogenicity. The result obtained by different prediction tools have their own methods for it. IPSNP collects results from each prediction tools and get consensus results on voting basis like the random forest algorithm.

Feature Based Comparison

A few platforms are available that provide impact of SNPs from multiple tools. Some examples include: PredictSNP predicts the impact of nsSNPs on protein function, using a combination of 8 bioinformatics tools. However, the major problem in this integrated platform is that the user has to collect the input data manually, which is one of the most time-consuming tasks. Second, they integrated only 9 tools on a platform. Third, it does not provide facility for filtering variants that are declared pathogenic by all the selected tools. MetaSNP includes predictions from 4 tools namely SIFT, PhDSNP, SNAP, and PANTHER. However, the user has no choice to select the tools and must prepare the input for them manually. It does not filter common pathogenic variants. The dbNSFP provides annotation of SNPs of 38 tools on a single platform. However, this server does not provide a facility for retrieving data automatically. To get results using dbNSFP, the user must provide input in a specific format which is not an easy task especially when multiple SNPs have to be analyzed from the multiple variant annotation tools. SNPnexus provides annotation by 8 tools namely FitCons, DeepSEA, CADD, Eigen, FATHMM-MKL, FunSeq2, GWAVA, and ReMM. However, this tool does not have the mechanisms to get the input automatically. Recently, Hassan et al[29] developed a platform ISTJRip. They integrated 5 snp prediction tools; mutation assessor, iFish, Provean, SIFT and FATHMM. They found it was better to use the combined prediction tools rather than individual one. The major limitation in all these tools/databases is that they do not provide an approach for getting input data automatically. IPSNP outperformed all these tools (Table 2) by accepting only gene id. There is no need to provide protein sequence, chromosome coordinates, mutations, or genetic variant ids (rsIDs). All other platforms require input data to be entered by the user manually which leads to inconsistent input. Secondly, IPSNP provides an interface for retrieving CPD from NCBI dbSNP and UniProt based only on the gene id which is very useful for the user who wants to run SNP analysis tools directly by visiting their web servers or on their own local machines for analyzing a long list of genetic variants.

IPSNP provides results for almost all inputs provided to it. However, if there is some error reported for missing values, then the user is provided with facility to download the required input coordinates so that a user can get results for the same input directly from respective SNP prediction tool. Although there are many prediction tools, but in this research, 5 tools were selected for the comparison with IPSNP. The reason behind selecting these tools was that these tools are available online and easy to operate. Some other tools are also available, but they need to be downloaded and run on the local machine. It is required to download database for and configuring it is very difficult. The hardware requirements for running these tools may be very high in the form of cost and resources. Therefore, the scope of comparing this approach was focused to these 5 tools.

Accuracy Based Comparison

To compare IPSNP with other tools, 100 benign and 100 pathogenic variants were obtained from ClinVar. All tools supported by IPSNP were selected to obtain the prediction results. Other SNP analysis tools namely SIFT, Provean, PolyPhen-2, Fathmm, and Mutation assessor were run with provided variants. The prediction was performed from the selected tools. The results of all tools were taken and documented. The same dataset of benign and pathogenic variations was analyzed on our proposed model IPSNP. The results were transformed in the form of True Negative (TN), False Positive (FP), True positive (TP), and False Negative (FN). The results of all prediction tools were analyzed using the evaluation parameters accuracy, precision, recall/sensitivity, specificity, negative predictive value (NPV), Mathew correlation coefficient (MCC),

Table 2. Comparison of IPSNP with other similar tools/approaches. The main parameters which made IPSNP unique include provision of (i) results of each selected tool, (ii) common pathogenic variants, (iii) retrieving CPD for the given gene id in CSV format. It does not require any list of mutations, chromosome coordinates or any other complex form of data.

TOOL	NO. OF TOOLS INTEGRATED	COMMAND BASE VERSION (CBV)	DATABASE REQUIRED FOR CBV	COMMON PATHOGENIC FILTERING	EXTRACTION OF CHROMOSOME COORDINATES (GRCH38)	EXTRACTION OF CHROMOSOMAL COORDINATES (GRCH37)	EXTRACTION OF PROTEIN SEQUENCE AND MUTATIONS	EXTRACTION OF RSIDS	INPUT	OUTPUT FORMAT
PredictSNP	8	No	NA	No	No	No	No	No	Protein sequence and list of mutations	CSV
MetaSNP	4	No	N/A	No	No	No	No	No	Protein sequence and list of mutations	Text
Condel	5	Yes	Yes	No	No	No	No	No	Protein sequence and list of mutations	WAS
dbNSFP v4	38	Yes	Yes	No	No	No	No	No	Chromosome coordinates	Text
SNPnexus	5	No	N/A	No	No	No	No	No	Chromosome coordinates or dbSNP ids (rsIDS)	Text, VCF,and TSV
VannoPortal[30]	11	No	No	Yes	Yes	No	No	No	dbSNP ID, VCF, rsID, Genomic coordinates, HGVS	Json, VCF
IPSNP	29	Yes	No	Yes	Yes	Yes	Yes	Yes	NCBI Gene Id	CSV

Table 3. Accuracy evaluation of IPSNP.

TOOS	SIFT		PROVEAN		POLYPHEN-2		FATHMM		MUTATION ASSESSOR		IPSNP	
PARAMETERS	SCORE	%AGE	SCORE	%AGE	SCORE	%AGE	SCORE	%AGE	SCORE	%AGE	SCORE	%AGE
Accuracy	0.67	66.5	0.83	82.9	0.72	72	0.76	75.84	0.72	72.16	0.88	87.5
Precision	0.61	60.93	0.78	78.15	0.64	64.04	0.67	66.67	0.67	67	0.82	81.51
Specificity	0.41	41	0.72	72.04	0.56	56.38	0.55	54.84	0.65	64.52	0.78	78
Sensitivity/ Recall	0.92	92	0.93	93	0.9	90.12	0.99	98.82	0.81	80.72	0.97	97
NPV	0.84	83.67	0.91	90.54	0.87	86.89	0.98	98.08	0.79	78.95	0.96	96.3
MCC	0.38	38.36	0.67	66.84	0.49	48.66	0.59	58.94	0.46	45.59	0.76	76.39
F-score	0.73	73.31	0.85	84.93	0.75	74.87	0.8	79.62	0.73	73.22	0.89	88.58

and F-score. All these parameters were calculated based on above mentioned 4 values. The accuracy of our proposed tool is the highest of all these selected tools. The accuracy of IPSNP was 87.5% which is more than all the tools executed for predicting the variants as benign or pathogenic. High accuracy means that IPSNP has a high proportion of correct predictions. Precision of our proposed tool IPSNP is 81.5, which is greater than all selected tools. High precision means the result is correct as compared to other prediction tools. Then specificity was calculated for all tools. It was 78% for IPSNP. It was from 41% to 72% for the other prediction tools. Sensitivity / recall for IPSNP are 97% which was among the best performing prediction algorithms. NPV for IPSNP is 96.3% which is the best of all selected prediction tools. MCC is 76.4 that mean IPSNP performs balanced results between TP, FN, TN, and FP. F-Score is 88.58 that mean the precision and recall is balanced. Overall IPSNP performed better than all selected tools based upon the given metrics. The results of accuracy and other parameters are shown in Table 3. The scope of comparison of IPSNP results was focused to the selected 5 tools. The rationale behind selecting these tools is that they are available online. A user does not have to put time and effort to download and install these tools on the local machine. Downloading and installing the SNP prediction tools involves much time and raise the cost for hardware resources.

Discussion

A few databases and tools have been developed to analyze missense variants by using more than one tool simultaneously but each of these approaches has its own limitations such as some tools allow using limited number of tools, for example MetaSNP combined only 4 tools and SNPnexus supports results by 5 tools. Most of the tools lack the provision of command base tool. Other more important feature that lack in almost all these tools is that they don't provide facility to filter common pathogenic variants that may be needed by several studies including GWAS and analyzing SNPs associated with a gene which is one of the most focused studies. Deng et al[31] conducted research was confined to only APOC2 and APOA5 genes. They used only 8 SNPs of APOC2 genes and 17 APOC5 genes. The dataset used was very limited. They used 8 tools for predictions. In their research they did not consider the nod coding variations. In our proposed model the number of tools is sufficient. Also non-synonymous SNPs are considered for prediction and analysis. Prakasam et al (2023) studied the effects of SNPs on the structure and function of TL4 gene.[32] There was no integration of computational prediction tools. Shah et al[33] focused on UTR gene for studying functional importance of SNPs and their impact on the human diseases. They did not integrated and obtained any consensus results. Joshi et al[34] used Support Vector Machine, Artificial Neural Network and Random forest for implementation but their scope was limited to selected genes only. In contrast to all above mentioned tools, most important functionality provided by IPSNP is that now the user does not have to collect CPD manually by visiting different web pages. This task is required when the user intends to analyze the SNPs associated with a gene available in a repository such as dbSNP. For example, a user will have to retrieve chromosome coordinates based on GRCh37 or GRCh38 and protein mutations by visiting NCBI dbSNP, UniProt Id, protein sequence from UNiProt database of all variants one by one to investigate them using SIFT, Polyphen2, MetaSNP, SNPnexus, or other similar tools. This is a very cumbersome and laborious job especially when the user must analyze tens of hundreds of variants by using more than one annotation tools.

There were 2 main objectives of developing IPSNP. First objective was to provide an integrated framework that enables a user to execute predictions of more than one tool simultaneously and provision of list of variants that have been declared pathogenic by all selected tools. The second target was the provision of an application that allow the user to extract chromosome coordinates and protein mutations to get predictions of variants by running tools by visiting their web interfaces. First objective was achieved by developing AEFI and second interface that is, CPDEI allow the user to extract CPD by single click. The proposed study is very useful for getting predictions of more than one tool. The user provides only gene id and selects the desired tools. No need to provide chromosome coordinates, protein mutations or rsIDs that are the requirement of almost all SNP analysis tools. Secondly, if a user wants to analyze SNPs by running tools directly visiting their web interfaces (e.g. SIFT,[35] Polyphen-2,[36] PredictSNP, MetaSNP etc), then there is no need to get parameters needed by these web servers manually. CPDEI of the IPSNP will extract all required data and save a lot of time.

IPSNP outperforms in comparison with other discussed tools in the scope of this research. This platform provides flexibility to user in selection of prediction tools rather than hard bounded facility of tools. A user can select one or all the tools as per requirements. Secondly, the input methods provided by IPSNP are versatile. A user has liberty to provide either rsID, gene ID, or HGVS Id as per ease. Third, the user has been provided the facility to download the results in CSV format. Fourth, the user gets the common pathogenic results which are not available in other tools, as per our best of knowledge. Five, if a user does not want to run the prediction tools using the IPSNP platform then it is provided the facility to download the input coordinates based on GRCH37 and GRCH28 assemblies as required by the user some specific SNP prediction tools. Six, IPSNP provides the tools wise results. This unique feature differentiates the IPSNP from all other tool. Lastly, IPSNP provides the consensus results which enable the user to get confident results based on the fact that it has been declared as benign or pathogenic by most of the SNP prediction tools. Although IPSNP is an efficient platform, however it has some limitations. We are using data from NCBI and Uniprot; and API of MyVariant.info. If any of these servers is down then IPSNP platform will not be able to execute and give the required results. If any of selected SNP prediction tools is down, the user will not be able to run the selected tools to get results.

Conclusions

IPSNP is a powerful tool that brings together various software tools to assist the interpretation of genetic variants. The platform allows researchers and clinicians to analyze genetic variants in a more comprehensive manner and make more efficient diagnosis, treatment, and disease management. It has revolutionized the way genetic data is analyzed and interpreted and has led to significant advancements in personalized medicine. This study facilitates to obtain predictions of variants associated with a gene by integration of 29 SNP annotation tools with a single click, filtering common pathogenic variants and retrieving CPD to execute various SNP analysis tools by visiting their web interfaces. Overall, the IPSNP platform has the potential to greatly improve patient outcomes by providing clinicians with more information to make better decisions regarding treatment and management of genetic diseases.

Acknowledgements

We are grateful to Virtual University of Pakistan and higher education commission for their support during this research work.

Author's Contributions

Syed Shah Muhammad conceived the idea conducted literature survey, designed and performed experiments, analyzed the data, drafted the research article, prepared the figures and/or tables, revised it critically for important content. Muhammad Tariq Pervez contributed in code writing, design of Platform and in drafting the research article. Muhammad Shoaib contributed in designing experiments, verified the experiments and results, writing the research article and proof reading of the document.

Availability

The link of web application is: http://ipsnp.vu.edu.pk:8000/

The complete code of Web application and command line version is available at: https://github.com/usmanathar2023/snp/tree/snp_pline_cmd

DOI: 10.5281/zenodo.8319444

REFERENCES

1. Sim NL, Kumar P, Hu J, et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40:W452-W457.
2. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31:2745-2747.
3. Hartley SW, Monti S, Liu CT, Steinberg MH, Sebastiani P. Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. *Front Genet*. 2012;3:176.
4. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using polyphen-2. *Curr Protoc Hum Genet*. 2013;76:7.20.1-7.20.41.
5. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. 2016;99:877-885.
6. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Primers*. 2021;1:59.
7. Van El C, Cornel M, Borry P, et al. Whole-genome sequencing in health care. *Eur J Hum Genet*. 2013;21:580-584.
8. Bendl J, Stourac J, Salanda O, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol*. 2014;10:e1003440.
9. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*. 2013;14 Suppl 3:S2-S9.
10. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*. 2005;21:ii54-ii58.
11. Li S, Ma L, Li H, et al. Snap: an integrated SNP annotation platform. *Nucleic Acids Res*. 2007;35:D707-D710.
12. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. 2021;49:D394-D403.
13. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*. 2020;12:103-108.
14. Oscanoa J, Sivapalan L, Gadaleta E, et al. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res*. 2020;48:W185-W192.

15. Gulko B, Gronau I, Hubisz M, Siepel A. Probabilities of fitness consequences for point mutations across the human genome. *Nat Genet*. Posted online September 11, 2014. bioRxiv 006825. doi:10.1101/006825
16. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015;12:931-934.
17. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47:D886-D894.
18. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*. 2016;48:214-220.
19. Shihab HA, Gough J, Mort M, et al. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*. 2014;8:11-16.
20. Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*. 2014;15:480.
21. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods*. 2014;11:294-296.
22. Smedley D, Schubach M, Jacobsen JOB, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am J Hum Genet*. 2016;99:595-606.
23. Landrum MJ, Chitipiralla S, Brown GR, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48:D835-D844.
24. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat*. 2009;30:1237-1244.
25. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 2011;39:e118-e118.
26. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308-311.
27. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*. 2020;2020:baaa062.
28. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43:D204-D212.
29. Hassan MS, Shaalan AA, Khamis S, Barakat A, Dessouky MI. Integrated rules classifier for predicting pathogenic non-synonymous single nucleotide variants in human. *Gene Rep*. 2024;34:101887.
30. Huang D, Zhou Y, Yi X, et al. VannoPortal: multiscale functional annotation of human genetic variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids Res*. 2022;50:D1408-D1416.
31. Deng H, Li J, Shah AA, Ge L, Ouyang W. Comprehensive in-silico analysis of deleterious SNPs in APOC2 and APOA5 and their differential expression in cancer and cardiovascular diseases conditions. *Genomics*. 2023;115: 110567.
32. Prakasam P, Abdul Salam AA, Basheer Ahamed SI. The pathogenic effect of SNPs on structure and function of human TLR4 using a computational approach. *J Biomol Struct Dyn*. 2023;41:12387-12400.
33. Shah H, Khan K, Badshah Y, et al. Investigation of UTR variants by computational approaches reveal their functional significance in PRKCI gene regulation. *Genes*. 2023;14:247.
34. Joshi I, Bhrdwaj A, Khandelwal R, et al. Artificial intelligence, big data and machine learning approaches in genome-wide SNP-based prediction for precision medicine and drug discovery. In: Basak SC, Vračko M, eds. *Big Data Analytics in Chemoinformatics and Bioinformatics*. Elsevier; 2023; 333-357.
35. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11:863-874.
36. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248-249.