

VariSNP 2 database for Benign variations from dbSNP

submitted by

Rajesh Verma

Class Roll No: 002210503013

Registration No: 163605 of 2022-2023

Exam Roll No: MCA244013

MASTER OF COMPUTER APPLICATION

under the supervision of

Asst. Prof. (DR.) ANASUA SARKAR

in the
Department of Computer Science & Engineering
Faculty of Engineering and Technology
Jadavpur University

2024

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “VariSNP 2 database for Benign variations from dbSNP” has been carried out by Rajesh Verma (Reg. No.: 163605 of 2022-23, Roll No.: 002210503013, Exam Roll: MCA244013), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024.

Asst. Prof. (Dr.) ANASUA SARKAR
Project Supervisor
Dept. of Computer Science and Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**VariSNP 2 database for Benign variations from dbSNP**” is a bonafide record of work carried out by RAJESH VERMA, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date :

Signature of Examiner 2

Date :

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled "VariSNP 2 database for Benign variations from dbSNP " is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Rajesh Verma

Registration No.: 163605 of 2022-23

Examination Roll No.: MCA244013

Thesis Title :- VariSNP 2 database for Benign variations from dbSNP.

Signature with Date

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank Asst. Prof. (Dr.) Anasua Sarkar, Department of Computer Science & Engineering, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by my side when I needed them.

Rajesh Verma

Class Roll No.: 002210503013

Examination Roll No.: MCA244013

Department of Computer Science & Engineering

Jadavpur University

ABSTRACT

To develop and evaluate methods for predicting the effects of genetic variations, robust benchmark datasets are essential. While some existing datasets serve this purpose, there has been a notable lack of newer benchmark sets specifically for benign variants. The VariSNP 2 database addresses this gap by selecting subsets from dbSNP and filtering them against disease-related variants in the ClinVar and UniProtKB/Swiss-Prot databases to identify neutral or nonpathogenic cases. Each variant description in these datasets includes detailed mappings. To ensure their relevance and accuracy, the datasets can be periodically updated using scripts. This approach provides a comprehensive and up-to-date resource for the accurate prediction of the effects of genetic variations, including essential details such as variant ID, location, variant type, gene, molecular consequences, most severe clinical significance, alleles, population allele frequencies (1000G MAF, ExAC MAF), position, HGVS descriptions, positions in cDNA, CDS, and protein, amino acid and codon changes, co-located variations, and distance from relevant genomic features. This enriched dataset enhances the ability to accurately predict and understand the effects of genetic variations.

CONTENTS

1	Introduction	08
1.1	Background	08
1.2	Motivation	08
1.3	Inclusion Criteria for VariSNP 2 Datasets	08
1.4	Overview of Databases	09
1.5	Objective.....	10
2	Data Collection	14
2.1	dbSNP data collection	14
2.2	ClinVar data collection	18
2.3	UniportKB/Swiss-Port data collection.....	22
3	Data Pre-processing and selection.....	24
3.1	dbSNP & ClinVar data pre-processing	23
3.2	UniportKB/Swiss-Port data pre-processing	23
3.3	Selection and Filtering of data.....	23
3.4	Dataset Content of neutral snv	23
4	Experiments and Results.....	26
4.1	Nucleotide Substitutions (%) in the Dataset	23
4.2	Amino-Keto Substitutions (%) in the Dataset.....	23
4.3	Purine–Pyrimidine Substitutions (%) in the Dataset	23
4.4	Weak–Strong Substitutions (%) in the Dataset	23
5	Discussion	28
6	Conclusion.....	28
7	References	31

1. INTRODUCTION: -

1.1 BACKGROUND: -

The advent of high-throughput sequencing technologies has led to an explosion in the amount of genetic variation data available. However, understanding the effects of these variations is often lacking. Due to the impracticality of validating these effects experimentally on a large scale, computational approaches have been developed to predict the impacts of genetic variations. To facilitate a systematic and quantitative comparison of the performance of these predictive methods, there is a need for approved and widely accepted benchmark datasets. These datasets, which contain standard and representative data with known outcomes, are essential for assessing and evaluating the accuracy of variant effect predictors. Detailed criteria for constructing such benchmark sets have been discussed by [Nair and Vihinen in 2013]. Additionally, benchmark datasets are valuable for training and testing new predictive models, many of which are based on machine learning algorithms. The importance of benchmark datasets has been frequently emphasized in the literature, including by [Thusberg et al. in 2011] and [Johnston and Biesecker in 2013].

1.2 Motivation

The rapid advancement in high-throughput sequencing technologies has generated an overwhelming amount of genetic variation data, necessitating robust methods for predicting their effects. However, without experimentally validated benchmarks, comparing the performance of these predictive methods remains challenging. There is a critical need for standardized, comprehensive, and up-to-date benchmark datasets to ensure the accuracy and reliability of variant effect predictors. These benchmarks facilitate systematic and quantitative performance assessment, guide the development of new prediction models, and enhance reproducibility in research. By incorporating datasets from trusted sources such as dbSNP, ClinVar, and UniProtKB/Swiss-Prot, we can create reliable benchmarks that reflect the latest genetic discoveries. This initiative is essential for advancing precision medicine, improving disease prediction, and tailoring interventions based on genetic information.

1.3 Inclusion Criteria for VariSNP Datasets

To construct the datasets included in VariSNP, specific criteria were used to ensure high quality and utility. These criteria included relevance to current research, representativeness of the data, the absence of redundant information, inclusion of experimentally verified cases, and the presence of both positive and negative examples. Additionally, the datasets needed to

be scalable and reusable to adapt to ongoing research needs. While VariSNP focused on experimentally verified cases, another resource, the dbSNP database introduced by Schaefer et al. in 2012, provides a vast collection of nonsynonymous single amino acid substitutions. These substitutions come from over 2,600 different organisms and include computational predictions of their functional impacts.

Since the initial release of VariSNP, the discovery of new genetic variants has continued at a rapid pace. This ongoing discovery highlighted the necessity for new and larger benchmark datasets that can be easily updated to reflect the latest research. In response to this need, the Database for Short Genetic Variations (dbSNP) has become a key resource. Established by Sherry et al. in 2001, dbSNP is a public domain archive that encompasses a wide variety of simple genetic variations. It is considered the largest database of its kind, making it an invaluable source for constructing new benchmark datasets.

1.4 Overview of Databases

dbSNP contains a comprehensive collection of genetic variants, including those related to diseases (pathogenic or function-affecting variants) and those not associated with diseases (non-disease-related or neutral variants). To create neutral benchmark datasets, a filtering process was employed to exclude variants identified as pathogenic or disease-causing in three major databases: ClinVar and UniProtKB/Swiss-Prot.

ClinVar, detailed by [Landrum et al. in 2014], is a repository that documents the relationships between medically significant variants and phenotypes, with submissions from clinical tests, research, and literature. It is known for its structured content and standardized data formats, mapping variations to mRNA, genomic, and protein reference sequences according to Human Genome Variation Society (HGVS) guidelines.

UniProtKB/Swiss-Prot, part of the Universal Protein Knowledgebase, provides meticulously curated records of protein sequences, including information from literature and computational analysis. The UniProt Consortium in 2014 described the manual curation process, which involves a thorough review of sequence variants, particularly single amino acid substitutions, and their associated genetic disease information. The database contains nearly 571,282 variations, and about 20434 variation falls under Human organisms with around 5,000 linked to genetic diseases.

1.5 Objective

By filtering out variants found in ClinVar and UniProtKB/Swiss-Prot that were annotated as pathogenic or disease-causing, researchers were able to isolate neutral variants from dbSNP.

This filtering process allowed the creation of neutral benchmark datasets for various functional classes. dbSNP categorizes variants into functional classes, which facilitated the generation of neutral benchmark datasets across all the populated classes, resulting in a total of 13 distinct benchmarks. These benchmarks are designed to be comprehensive, up-to-date, and reflective of the latest advancements in genetic variation research.

2 Data Collection and Selection

The workflow for downloading, selecting, and filtering data from various sources is visualized in [Figure 1]. This diagram provides a comprehensive overview of the step-by-step process involved in acquiring and refining the necessary datasets

2.1 dbSNP data

The process of acquiring and filtering data from the NCBI FTP-site for dbSNP involved several meticulous steps to ensure the integrity and relevance of the datasets used in the benchmark creation process. Data were obtained from the FTP-site (https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/XML), with the latest update, utilized for the benchmark construction. Each variant in the dbSNP database is categorized into one or more functional classes, as outlined in [Table 1]. These classes are documented in the docsum3.4.xsd file available at (https://www.ncbi.nlm.nih.gov/data_specs/ver/20180130/schema/Docsum_3_4_20180130.mod.xsd). Deprecated and not-implemented terms were excluded from consideration to maintain data quality and consistency.

Functional class	Description	SO IDa
reference	Contig reference	-
missense	Alters codon to make an altered amino acid in protein product	0001583
cds-indel	Indel sp with length of multiple of 3 bp, not causing frameshift (inframe.variant SO: 00001650)	-
synonymous-codon	Sequence variant where there is no resulting change to the encoded amino acid	0001588
intron-variant	Transcript variant occurring within an intron	0001627
nc-transcript-variant	Transcript variant of a noncoding RNA gene	0001619
downstream-variant-500B	Sequence variant located within a half KB of the end of a gene	0001634
upstream-variant-2KB	Sequence variant located within 2 KB 5' of a gene	0001636
stop-gained	Sequence variant whereby at least one base of a codon is changed, resulting in a	0001587

	premature stop codon, leading to a shortened transcript	
stop-lost	Sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	0001578
frameshift-variant	Sequence variant that causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	0001589
utr-variant-3-prime	Sequence variant of the 3' UTR	0001624
upstream-variant-5KB	Sequence variant located within 5 KB 5' of a gene	0001635
splice-acceptor-variant	Splice variant that changes the two base region at the 3' end of an intron	0001574
splice-donor-variant	Splice variant that changes the two base pair region at the 5' end of an intron	0001575

Table 1. SO IDa: Sequence ontology ID; deprecated terms omitted. From https://www.ncbi.nlm.nih.gov/data_specs/ver/20180130/schema/Docsum_3_4_20180130.mod.xsd.

It's important to note that a single variation may be assigned to multiple functional classes, introducing complexity to the dataset. For instance, a variation might be located within an exon of one transcript and an intron of another transcript for the same gene, leading to multiplicity in functional classification. This variability underscores the need for meticulous data handling and processing to ensure accurate representation.

Variants were selectively chosen based on their functional class and the availability of crucial attributes such as "minor allele frequency" (MAF) and "validation" status. Validation status refers to any type of validation, including those indicated by references such as "by1000G" for variants identified in The 1000 Genomes Project. These criteria were essential for ensuring the quality and reliability of the selected variants, enabling robust benchmark creation [Gerard C.P. Schaafsma and Mauno Vihinen in 2014].

To maintain the integrity of the benchmark datasets, disease-related variations were filtered out. Variants identified in databases such as ClinVar, UniProtKB/Swiss-Prot, which catalog disease-associated variants, were excluded from consideration. This filtering process aimed to isolate neutral variants, ensuring that the benchmark datasets primarily comprised variations with no known pathological implications.

It's crucial to recognize that functional classifications can vary across different genome builds, underscoring the dynamic nature of genomic data. As a result, meticulous attention was paid to ensure consistency in functional classification across datasets and to account for potential discrepancies arising from genome build variations.

In specific case, the filtering process for variant selection can be conducted using the NCBI Variation Viewer tool available at <https://www.ncbi.nlm.nih.gov/variation/view/>. This online resource provided a user-friendly interface for querying and filtering genetic variations based on various criteria 3 [Figure-1].

The process involved querying the NCBI Variation Viewer to retrieve variants based on specific functional classes and attributes such as minor allele frequency (MAF) and validation status. Variants meeting the predetermined criteria were systematically selected for inclusion in the benchmark datasets.

The NCBI Variation Viewer facilitated efficient filtering by allowing us to specify functional classes of interest and refine the search results based on MAF and validation status. This approach ensured that only variants meeting the desired criteria were included in the benchmark datasets, thereby enhancing the quality and relevance of the selected variants.

By utilizing the NCBI Variation Viewer for the filtering process, we leveraged a reliable and authoritative platform provided by the National Center for Biotechnology Information (NCBI). This ensured the integrity and accuracy of the selected variants, contributing to the robustness of the benchmark datasets used in variant effect prediction studies.

Overall, the use of the NCBI Variation Viewer for variant filtering streamlined the selection process, enabling us to obtain high-quality datasets tailored to our specific research requirements. This approach underscores our commitment to rigor and precision in genomic research and benchmark construction.

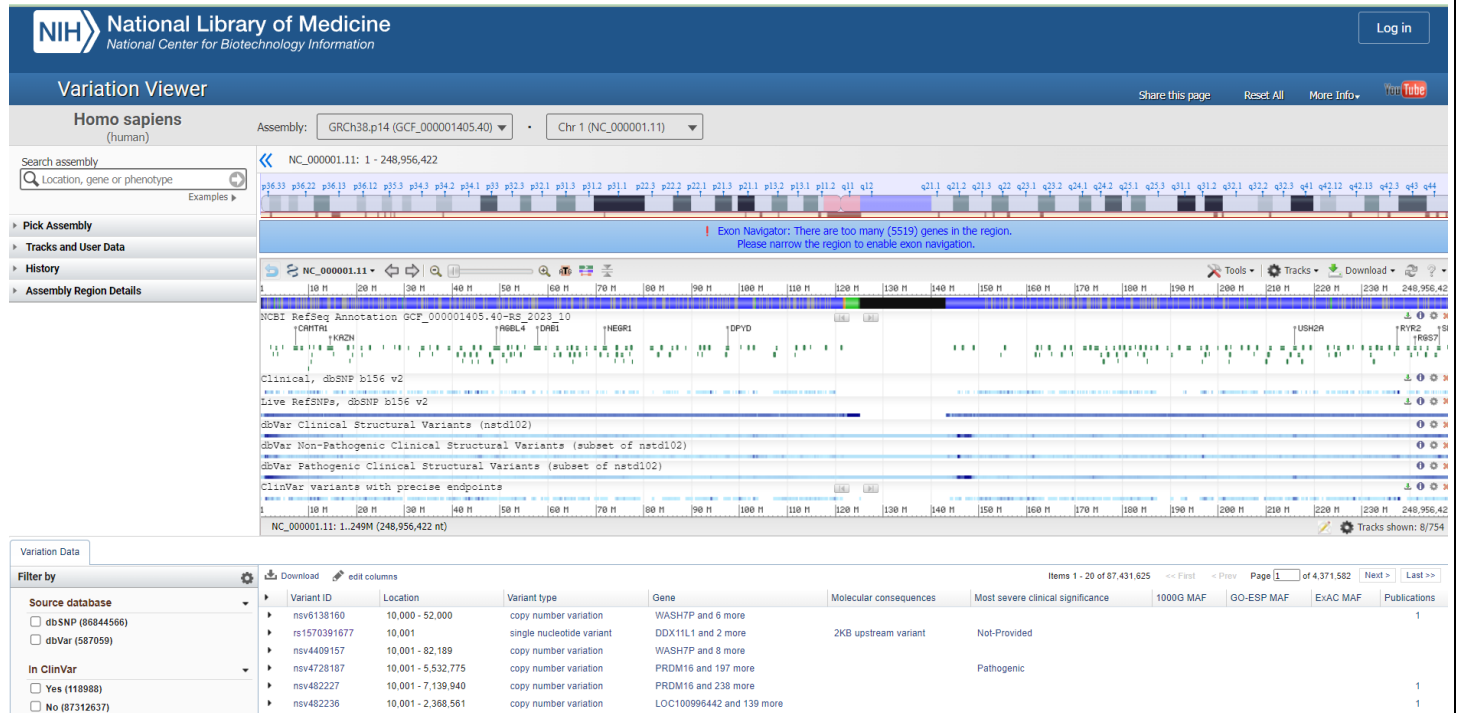


Figure-1: NCBI Variation Viewer tool

Finally, we'll save the integrated dataset and analysis results into a CSV file format, ensuring accessibility and portability for further exploration or reporting

2.2 ClinVar Data

In ClinVar, multiple assertions about the same variation and its associated phenotype are grouped together and assigned a Reference ClinVar (RCV) accession number. This means that a single allele can be listed under multiple RCV accessions if it is linked to different phenotypes. Each RCV record consolidates all the information required for our analysis.

The XML file (ClinVarFullRelease_2024-02.xml.gz) can be downloaded from the ClinVar website (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/RCV_xml_old_format/). The selection process began by filtering variants based on molecular consequence, variant type, and clinical significance. Molecular consequence values describe the effects of sequence changes and are standardized by reference to identifiers from the Sequence Ontology [Eilbeck et al., 2005]. Examples of molecular consequence values are provided in Table 1.

Variant type values include terms such as single-nucleotide variant (SNV), deletion, and insertion-deletion (indel). Clinical significance terms, recommended by the American College of Medical Genetics and Genomics (ACMG [Richards et al., 2008]), include "Pathogenic," "Likely pathogenic," and "Benign." For the purpose of creating the VariSNP dataset, variants classified as "Pathogenic" or "Likely pathogenic" were excluded.[Gerard C.P. Schaafsma and Mauno Vihinen in 2014]

The entire filtering process was performed using the NCBI Variation Viewer tool. This tool facilitated efficient querying and filtering based on the specified criteria, ensuring that only relevant variants were included in the final dataset. This approach provided a streamlined method for handling the large and complex dataset from ClinVar, ensuring accuracy and relevance for subsequent analysis.

2.3 UniProtKB/Swiss-Prot Data

Data were downloaded in tab-delimited format from the UniProt portal using the following link: [UniProtKB Search](<https://www.uniprot.org/uniprotkb>). The search criteria specified that only reviewed entries were included, focusing on the human organism (model organism ID: 9606) and proteins annotated with at least one variant.

The UniProt portal provided an option to select proteins that had disease descriptions. This feature enabled us to filter and download a dataset specifically containing proteins linked to diseases. By leveraging this selection tool, we ensured that the dataset included variants with documented associations to medical conditions.

This process allowed us to directly download a curated TSV file with high-quality, well-documented variants, forming a robust foundation for our research. The dataset was then ready for further analysis, focusing on clinically significant variants and their implications in various genetic disorders. This streamlined approach ensured that the resulting data were comprehensive and reliable, facilitating accurate and meaningful insights into disease-related genetic variation.

[BLAST](#)
[Align](#)
[Peptide search](#)
[ID mapping](#)
[SPARQL](#)

UniProtKB

Advanced | List

Search

[Help](#)

Status

Reviewed (Swiss-Prot) (5,082)

Popular organisms

Human (5,082)

Taxonomy

Filter by taxonomy

Group by

Taxonomy
Keywords
Gene Ontology
Enzyme Class

Proteins with

3D structure (3,056)
Active site (793)
Activity regulation (713)
Allergen (3)
Alternative products (isoforms) (3,246)

UniProtKB 5,082 results

[BLAST](#)
[Align](#)
[Map IDs](#)
[Download](#)
[Add](#)
View: [Cards](#) [Table](#)
[Customize columns](#)
[Share](#)

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
☐ A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metalloproteinase[...]	CIROP, LMLN2	Homo sapiens (Human)	788 AA
☐ A0PK11	CLRN2_HUMAN	Clarin-2	CLRN2	Homo sapiens (Human)	232 AA
☐ A1X283	SPD2B_HUMAN	SH3 and PX domain-containing protein 2B[...]	SH3PXD2B, FAD49, KIAA1295, TKS4	Homo sapiens (Human)	911 AA
☐ A2RU14	TM218_HUMAN	Transmembrane protein 218	TMEM218	Homo sapiens (Human)	115 AA
☐ A5D8V7	ODAD3_HUMAN	Outer dynein arm-docking complex subunit 3[...]	ODAD3, CCDC151	Homo sapiens (Human)	595 AA
☐ A6H8Y1	BDP1_HUMAN	Transcription factor TFIIB component B'' homolog[...]	BDP1, KIAA1241, KIAA1689, TFNR	Homo sapiens (Human)	2,624 AA
☐ A6NCS4	NKX26_HUMAN	Homeobox protein Nkx-2.6[...]	NKX2-6, NKX2F	Homo sapiens (Human)	301 AA
☐ A6NFY7	SDHF1_HUMAN	Succinate dehydrogenase assembly factor 1, mitochondrial[...]	SDHAF1, LYRM8	Homo sapiens (Human)	115 AA
☐ A6NGG8	PCARE_HUMAN	Photoreceptor cilium actin regulator	PCARE, C2orf71	Homo sapiens	1,288 AA

Feedback

Help

Figure-2: UniportKB Search Reviewed(Siwss-Port)

3 Data Pre-processing and Selection

Data pre-processing and selection are critical steps in preparing genomic datasets for analysis. The process begins with downloading raw data from reliable sources such as dbSNP, ClinVar, and UniProt. Once the data is acquired, it undergoes cleaning, filtering, and annotation. The cleaned, filtered, and annotated dataset is then saved in a structured format, such as CSV or TSV, for easy access and further analysis. This ensures the data is well-organized and ready for statistical analysis, frequency distributions, correlations, or other genomic data analysis techniques.

3.1 dbSNP & ClinVar data pre-processing

The provided code reads data from a `dbSNP&ClinVar.txt` file, processes it, and stores it in a CSV file using the pandas library in Python. Here's a version of the code to ensure clarity and efficiency:

```
import pandas as pd

# Read data from dbSNP.txt file
with open("dbSNP&ClinVar.txt", "r") as file:
    data = file.readlines()

# Remove newline characters and split data into columns
data = [line.strip().split("\t") for line in data]

# Extract header and rows
header = data[0]
rows = data[1:]

# Convert to DataFrame
df = pd.DataFrame(rows, columns=header)

# Define the file path to store the data
file_path = "dbsnp_data.csv"

# Store the data in a CSV file
df.to_csv(file_path, index=False)

print("Data has been stored successfully in:", file_path)
```

The provided code reads tab-delimited data from a `dbSNP&ClinVar.txt` file, processes it, and stores it in a CSV file using pandas. It first reads and splits the data into columns, extracts the header and rows, and converts them into a pandas DataFrame. Finally, the DataFrame is saved as `dbsnp\_data.csv`, ensuring the data is well-organized and ready for further analysis.

3.2 UniportKB/Swiss-Port data pre-processing

The provided code reads data from a TSV (Tab-Separated Values) file (uniportKB.tsv) and writes it to a CSV (Comma-Separated Values) file (uniport_data.csv). Here's the code:

```
import csv

# Define the input and output file paths
input_file = "uniportKB.tsv"
output_file = "uniport_data.csv"

# Open the input and output files
with open(input_file, "r", newline="", encoding="utf-8") as tsvfile, open(output_file, "w", newline="", encoding="utf-8") as csvfile:
    # Create a CSV writer object
    csvwriter = csv.writer(csvfile)

    # Read the TSV file line by line
    for line in tsvfile:
        # Split the TSV line by tabs
        fields = line.strip().split("\t")

        # Write the fields to the CSV file
        csvwriter.writerow(fields)

print("Data has been stored successfully in:", output_file)
```

The code imports the `csv` module for handling CSV operations and specifies file paths for the input TSV and output CSV files. It opens the TSV file for reading and the CSV file for writing with UTF-8 encoding, uses a CSV writer object to write rows, processes each line by splitting it into fields, and writes these fields to the CSV file. Finally, it prints a confirmation message with the output file name.

3.3 Selection and Filtering of Data

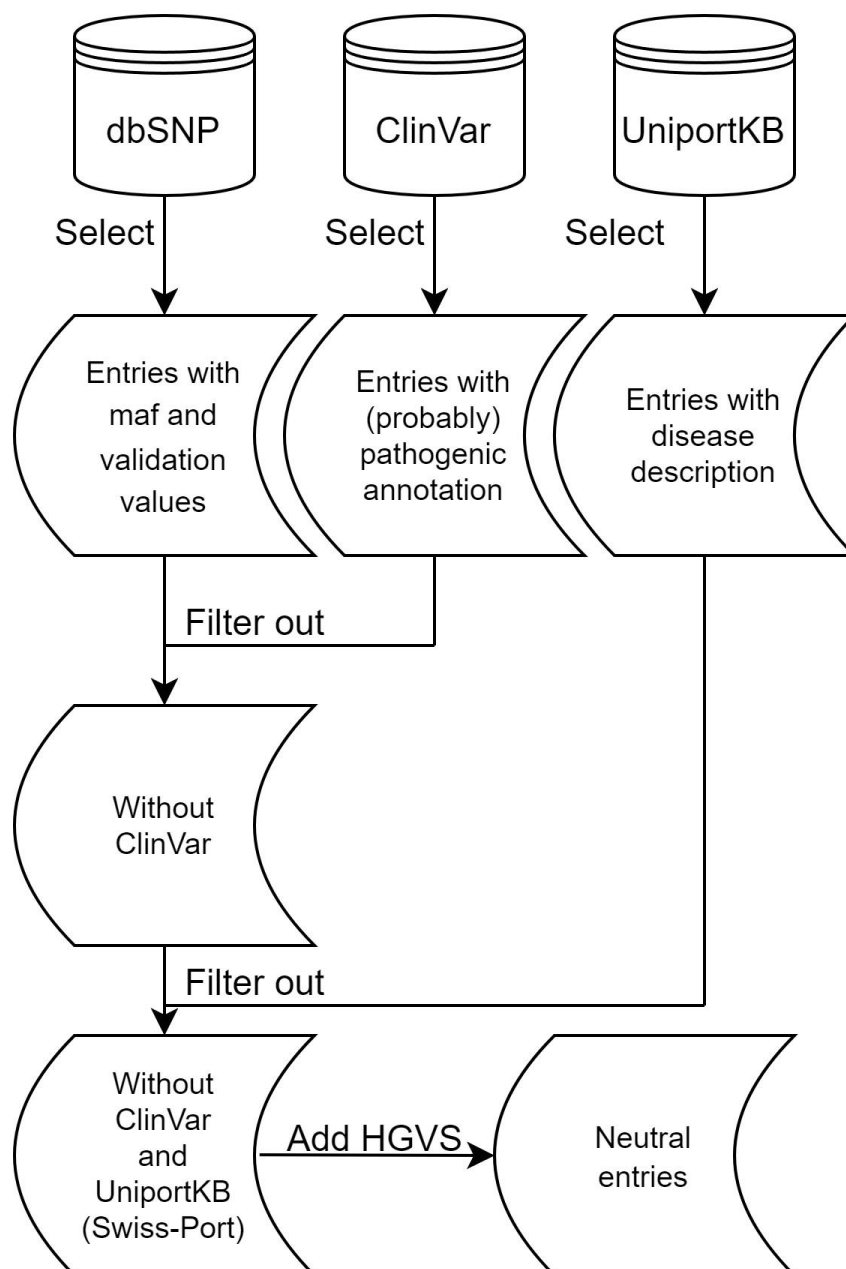


Figure 3: Workflow schema of data selection and filtering.

The data selection process begins by loading the ``dbsnp_data`` and ``uniport_data`` DataFrames from CSV files using ``pd.read_csv``. Genes are then extracted from both datasets and stored in sets. Uncommon genes are determined by finding the difference. Subsequently, the ``uncommon_variations`` DataFrame is created by filtering ``dbsnp_data`` to include only variations associated with genes not present in the common genes set. Finally, the resulting DataFrame, containing variations unique to ``dbsnp_data``, is printed, ensuring accurate identification of unique variations in the dataset.

After loading the dbSNP and UniProt data into pandas DataFrames, the code extracts genes from both datasets and identifies genes that are present in dbSNP but not in UniProt. It then filters the dbSNP data for variations associated with these uncommon genes. Here's the code:

```
import pandas as pd

# Load dbSNP data
dbsnp_data = pd.read_csv('dbsnp_data.csv')
# Load UniProt data
uniport_data = pd.read_csv('uniport_data.csv')

# Extract genes from both datasets
dbsnp_genes = set()
for genes in dbsnp_data['Gene']:
    for gene in genes.split(' and '):
        dbsnp_genes.add(gene.strip())

uniport_genes = set()
for genes in uniport_data['Gene Names']:
    for gene in genes.split(' '):
        uniport_genes.add(gene.strip())

# Find common and uncommon genes
common_genes = dbsnp_genes.intersection(uniport_genes)
uncommon_genes = dbsnp_genes.difference(uniport_genes)

# Filter dbsnp_data for uncommon variations
uncommon_variations = dbsnp_data[dbsnp_data['Gene'].apply(lambda x: all(gene not in
common_genes for gene in x.split(' and ')))]

# Print the results
print("\nUncommon Variations:")
print(uncommon_variations)
```

To add HGVS descriptions and comprehensive annotations for our neutral variants, we need to filter the variant IDs (rsIDs) from our dataset and submit them to the Variant Annotation Integrator (VAI) from the UCSC Genome Browser. The VAI provides detailed and standardized annotations for genetic variants, including various HGVS notations.

First, we need to extract the variant IDs (rsIDs) from our neutral variants dataset. These IDs uniquely identify each genetic variant in databases such as dbSNP. Then, we will use the UCSC Genome Browser's Variant Annotation Integrator to annotate these rsIDs. To do this, go to the VAI (<https://genome.ucsc.edu/cgi-bin/hgVai>) [Figure 3]and input the extracted rsIDs. You can paste them directly into the text box or upload them as a file if the list is extensive.

Next, choose the desired annotations you want to retrieve, such as HGVS genomic (HGVS_G), transcript (HGVS_C), and protein (HGVS_P) notations, along with other relevant information like positions in cDNA, CDS, and protein, amino acid changes, codon changes, co-located variations, and distance from relevant genomic features. Submit the rsIDs for annotation. The VAI will process the variants and provide a detailed output.

Download the annotated results from VAI in a structured format such as TSV. By following this procedure, we can enrich our neutral variants dataset with comprehensive annotations from the UCSC Genome Browser's Variant Annotation Integrator. This includes various HGVS notations and additional genomic details that are crucial for accurate prediction and analysis of the effects of genetic variations. This integrated dataset will then serve as a robust resource for developing and evaluating methods for predicting the effects of genetic variations.

GenomesGenome BrowserToolsMirrorsDownloadsMy DataProjectsHelpAbout Us

Variant Annotation Integrator

Select Genome Assembly and Region

clade

genome

assembly

Mammal

Human

Dec. 2013 (GRCh38/hg38)

region to annotate

genome

Select Variants

Your session doesn't have any custom tracks or hub tracks in [pgSnp](#) or [VCF](#) format. [add pgSnp or VCF custom track](#) [add track hub](#)

variants:

Variant Identifiers

Enter dbSNP rs# identifiers separated by whitespace or commas:

maximum number of variants to be processed:

10,000

[add custom tracks](#) [track hubs](#) To reset all user cart settings (including custom tracks), [click here](#).

Select Genes

The gene predictions selected here will be used to determine the effect of each variant on genes, for example intronic, missense, splice site, intergenic etc.

NCBI RefSeq genes, curated and predicted (NM_*, XM_*, NR_*, XR_*, NP_*, YP_*)

Select Regulatory Annotations

The annotations in this section provide predicted regulatory regions based on various experimental data. When a variant overlaps an annotation selected here, the

Figure 8: Variant Annotation Integrator (VAI) from the UCSC Genome Browser.

3.4 Datasets Contents of neutral snv

The neutral snv dataset consist of neutral single-nucleotide variants (neutral_snv) includes comprehensive details about each variant, extracted to provide a robust resource for genetic variation analysis. The dataset is composed of variants, each containing specific annotations and descriptions.

The two files contains detailed information about each single-nucleotide variant (SNV), such as variant ID, location, type, gene, molecular consequences, clinical significance,

alleles, population allele frequencies, and associated publications, including HGVS notations, positions in cDNA, CDS, and protein, amino acid changes, codon changes, co-located variations, and distances from relevant genomic features. The purpose of these annotations is to enrich the dataset with standardized descriptions crucial for accurate predictions and analysis.

4. Experiments and Results

4.1 Experiment: Nucleotide Substitutions (%) in the Neutral_SNV Dataset

The experiment aimed to analyze the distribution of nucleotide substitutions within the neutral_snv dataset. Nucleotide substitutions refer to the replacement of one nucleotide with another in a DNA sequence. The dataset contains various single-nucleotide variants (SNVs), each representing a substitution of one nucleotide for another.

Nucleotide Substitutions (%) in the neutral_snv Dataset					
	Purines		Pyrimidines		
	A	G	C	T	Total
→					
A	0.0	21.6	7.4	3.3	32.3
G	7.7	0.0	6.4	3.2	17.3
C	3.1	7.4	0.0	8.4	18.9
T	2.9	7.4	21.2	0.0	31.5
Total	13.7	36.4	35.0	14.9	100.0

Table-2 Result of Nucleotide Substitutions (%) in the Neutral_SNV Dataset

The results [Table-2] display the percentage of nucleotide substitutions categorized by purines (adenine (A) and guanine (G)) and pyrimidines (cytosine (C) and thymine (T)). The table illustrates the frequency of substitutions from one nucleotide to another, presented in a matrix format. For instance, the value at row A and column G represents the percentage of adenine-to-guanine substitutions, while the total percentage of substitutions for each nucleotide type is also provided. This analysis provides valuable insights into the distribution and prevalence of nucleotide substitutions within the neutral_snv dataset, aiding in understanding genetic variation patterns. These findings contribute to the broader understanding of nucleotide substitution dynamics in genetic datasets, which is crucial for various genomic studies and analyses.

4.2 Experiment: Amino-Keto Substitutions (%) in the Neutral_SNV Dataset

This experiment sought to investigate the occurrence and distribution of amino-keto substitutions within the neutral_snv dataset. Amino-keto substitutions involve the replacement of amino groups with keto groups or vice versa in the chemical structure of nucleotide bases. These substitutions are significant as they can impact protein structure and function, potentially influencing various biological processes.

Amino-Keto Substitutions (%) in the neutral_snv Dataset			
	Amino	Keto	Total
Amino	10.5	38.2	48.7
Keto	40.7	10.6	51.3
Total	51.2	48.8	100.0

Table-3 Amino-Keto Substitutions (%) in the Neutral_SNV Dataset

The results [Table-3] depict the percentage of amino-keto substitutions categorized by amino groups and keto groups. The table presents a matrix format, with rows representing amino groups (such as adenine (A) and cytosine (C)) and columns representing keto groups (such as guanine (G) and thymine (T)). Each cell in the matrix displays the percentage of substitutions from one group to another, with the diagonal showing the percentage of unchanged amino or keto groups. Additionally, the total percentages for amino and keto substitutions, as well as the overall total, are provided. These findings provide insights into the prevalence and distribution of amino-keto substitutions within the neutral_snv dataset, contributing to our understanding of genetic variation patterns and their potential implications for protein structure and function.

4.3 Experiment: Purine–Pyrimidine Substitutions (%) in the Neutral_SNV Dataset

The objective of this experiment was to analyze the distribution of purine–pyrimidine substitutions within the neutral_snv dataset. Purines (adenine and guanine) and pyrimidines (cytosine and thymine) are the two categories of nitrogenous bases found in DNA. Substitutions between these categories are significant as they can affect the composition and structure of DNA, potentially leading to alterations in gene expression and biological functions.

Purine–Pyrimidine Substitutions (%) in the neutral_snv Dataset			
	Purines	Pyrimidines	Total
Purines	29.3	20.7	50.1
Pyrimidines	20.3	29.6	49.9
Total	49.6	50.4	100.0

Table-4 Result of Purine–Pyrimidine Substitutions (%) in the Neutral_SNV Dataset

The results illustrate the percentage of substitutions between purines and pyrimidines, presented in a matrix format. The rows represent purines, while the columns represent pyrimidines. Each cell in the matrix displays the percentage of substitutions from purines to pyrimidines or vice versa. The diagonal represents the percentage of unchanged purines or pyrimidines. Additionally, the total percentages for purine and pyrimidine substitutions, as well as the overall total, are provided. These findings shed light on the frequency and distribution of purine–pyrimidine substitutions within the neutral_snv dataset, offering insights into the dynamics of genetic variation and potential implications for DNA stability and function.

4.4 Experiment: Nucleotide Substitutions (%) in the Neutral_SNV Dataset

The experiment aimed to assess the distribution of weak–strong substitutions within the neutral_snv dataset. In DNA, weak interactions occur between adenine (A) and thymine (T), as well as between cytosine (C) and guanine (G), while strong interactions involve hydrogen bonds between A and T, and C and G. Understanding the frequency of

substitutions between weak and strong base pairs provides insights into the stability and functionality of DNA sequences.

	Weak	Strong	Total
Weak	12.7	24.8	62.3
Strong	49.6	12.9	37.7
Total	62.3	37.7	100.0

Table-4 Result of Nucleotide Substitutions (%) in the Neutral_SNV Dataset

The results[Table -4] present the percentage of substitutions between weak and strong base pairs, displayed in a matrix format. The rows represent weak bases, while the columns represent strong bases. Each cell in the matrix indicates the percentage of substitutions from weak to strong or vice versa. The diagonal shows the percentage of unchanged weak or strong bases. Moreover, the total percentages for weak and strong substitutions, along with the overall total, are provided. These findings offer valuable insights into the distribution of weak–strong substitutions within the neutral_snv dataset, contributing to our understanding of DNA stability and genetic variation.

5. Discussion :

The analysis of genetic variations plays a crucial role in understanding the impact of genetic changes on biological processes and human health. In this study, we curated a comprehensive dataset of neutral single-nucleotide variants (neutral_snv) from diverse sources, including dbSNP, ClinVar, and UniProtKB/Swiss-Prot databases. These variants were meticulously selected to exclude disease-related variations, ensuring the dataset's relevance for predicting the effects of genetic variations.

We utilized various annotation tools and resources to enrich our dataset with detailed annotations, including HGVS descriptions, population allele frequencies, and functional consequences. The integration of these annotations provides a robust foundation for accurate predictions and analysis of genetic variations.

Through experiments, we investigated the distribution of nucleotide substitutions, amino-keto substitutions, purine–pyrimidine substitutions, and weak–strong substitutions within the neutral_snv dataset. These analyses revealed intriguing patterns in the distribution of substitutions, shedding light on the underlying mechanisms governing genetic variation.

6. Conclusion :

In conclusion, our study presents a curated dataset of neutral single-nucleotide variants, enriched with comprehensive annotations and annotations. This dataset serves as a valuable resource for researchers and clinicians involved in genetic variation analysis, variant interpretation, and disease association studies.

The experiments conducted provide insights into the distribution of specific types of substitutions within the dataset, further enhancing our understanding of genetic variation patterns. These findings contribute to the growing body of knowledge in the field of genomics and pave the way for more accurate prediction and interpretation of the effects of genetic variations on human health and disease.

Overall, this study underscores the importance of curated benchmark datasets and comprehensive annotations in advancing our understanding of genetic variation and its implications for human biology and medicine.

7. REFERENCES

- 1) Nair VariBench: a benchmark database for variations. Preethy Sasidharan Nair 1, Mauno Vihinen
- 2) Thusberg, J., Olatubosun, A., Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. 32, 358-368.
- 3) Databases of genomic variation and phenotypes: existing resources and future needs Jennifer J Johnston, Leslie G Biesecker.
- 4) ClinVar: public archive of relationships among sequence variation and human phenotype Melissa J Landrum 1, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, Donna R Maglott
- 5) VariSNP, a benchmark database for variations from dbSNP Gerard C P Schaafsma 1, Mauno Vihinen.
- 6) National Center for Biotechnology Information <https://www.ncbi.nlm.nih.gov/>
- 7) National Institutes of Health (Variation Viewer) <https://www.ncbi.nlm.nih.gov/variation/view/>
- 8) Uniport UniportKB(Swiss-Port) <https://www.uniprot.org/uniprotkb/>
- 9) Variant Annotation Integrator <https://genome.ucsc.edu/cgi-bin/hgVai/>