

Software

SNPraefentia: a toolkit to prioritize microbial genome variants linked to health and disease

Nadeem Khan^{1,†}, Muhammad Muneeb Nasir^{1,†}, Ammar Mushtaq²,
Masood Ur Rehman Kayani^{1,*} 

¹Metagenomics Discovery Lab, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST), Islamabad 44000, Pakistan

²Computational Aeronautics Lab, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST), Islamabad 44000, Pakistan

*Corresponding author. Metagenomics Discovery Lab, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST), Sector H-12, Islamabad 44000, Pakistan. E-mail: masood.kayani@gmail.com

[†] = equal contribution.

Associate Editor: Alia Benkhla

Abstract

Motivation: Analysis of genomic variation in microbial genomes is crucial for understanding how microbes adapt, interact with their hosts, and influence health and disease. In metagenomic studies, where genetic material from entire microbial communities is sequenced, thousands of single-nucleotide polymorphisms can be detected across species and samples. However, identifying which of these variations has biologically or functionally relevant impacts remains a significant challenge.

Results: To address this, we present SNPraefentia, a Python-based toolkit for prioritizing microbial SNPs based on their predicted functional relevance. The tool integrates multiple biologically meaningful parameters, including sequencing depth, physicochemical impact of amino acid substitutions, and the structural and functional context of mutations within annotated protein domains. SNPraefentia extracts variation depth and amino acid changes, annotates protein domains using UniProt, and computes individual impact scores. These are then integrated into a composite prioritization score that reflects the potential biological importance of each variant. Overall, SNPraefentia provides researchers with a systematic and reproducible approach to filter and rank microbial variants for downstream functional analysis or experimental validation.

Availability and implementation: The toolkit and test data are freely available at <https://github.com/muneebdev7/SNPraefentia>.

1 Introduction

Single-nucleotide polymorphisms (SNPs) in microbial genomes are key drivers of evolutionary change, phenotypic diversity, and functional adaptation. In clinical and ecological contexts, specific SNPs can influence antimicrobial resistance, virulence, host interaction, and metabolic capabilities. Accurately identifying and prioritizing these SNPs (further referred to as variants) is essential for understanding microbial behavior and guiding downstream experimental analyses (Riccardi and Resta 2020, Li *et al.* 2021, Miller and Chiu 2021, Zhang and Knight 2023, Bhunia *et al.* 2024, Linz *et al.* 2024). Several tools, such as SnpEff and Snippy, are currently deployed for detecting and annotating variants in microbial genomes (Seemann 2015). However, with the increasing complexity of microbial genomic datasets, owing to the advances in genome-resolved metagenomics and strain-resolved analyses, there is an urgent need for more advanced methods that go beyond identification and annotation and enable prioritization of variants relevant to health and disease.

To address this gap, we present SNPraefentia, a tool developed for prioritizing variants in microbial genomes. Unlike currently available variant prioritization tools, designed for

human datasets, SNPraefentia is tailored for microbial systems (Gong *et al.* 2015, Yaury *et al.* 2018). It incorporates multiple biologically significant features, including variant read depth, changes in amino acid physicochemical properties (e.g. size, charge, hydrophobicity, polarity), and the presence of mutations in known functional protein domains. The integration of these features yields a composite prioritization score that reflects the cumulative biological importance of each variant. This framework enables users to rank SNPs based on their potential to influence microbial function and clinical relevance. By automating this process, SNPraefentia empowers researchers to focus on high-impact variants, facilitating more targeted studies across diverse microbial genomes, including those identified or reconstructed from metagenomic samples.

2 Tool description

SNPraefentia is implemented in Python and developed as a command-line tool for transparency, modularity, and ease of use. The tool is packaged and publicly available through the Python Package Index (PyPI) and can be installed using the

following command: `pip install snpraefentia`. The source code is publicly available at <https://github.com/muneebdev7/SNPræfentia>. The overall architecture of SNPræfentia is shown in Fig. 1, outlining the sequential flow from parsing the input to the generation of a prioritized variant list.

2.1 Input file

SNPræfentia accepts variant annotations in tab-separated (tsv) or comma-separated (csv) formats (Supplementary Table 1, available as supplementary data at *Bioinformatics Advances* online) as input, typically generated from a variant calling pipeline (e.g. Snippy). The input file must contain the following four columns: (i) Gene, that indicates the microbial gene in which the variation has been identified; (ii) Amino_Acid_Position, representing the genomic variant position in the format x/y where x is the variation site and y is the total number of amino acids in the protein; (iii) Effect, describing the substitution at amino acid level (e.g. Val170Leu); and (iv) Evidence, indicating the variant depth.

2.2 Normalization of variant read depth

Variant read depth is one of the key metrics in prioritizing variants, since high-depth variants are more likely to represent true biological signals rather than sequencing noise (Jee *et al.* 2016, Farheen *et al.* 2017). To make depth values comparable across samples or genes with varying coverage, SNPræfentia applies min-max normalization (also referred to as min-max scaling) to transform variant read depth values into a standard range of [0,1]. This approach was selected purely for scaling purposes, as all input variants had already met the quality and depth thresholds defined by Snippy, ensuring the absence of false variant calls. The resulting scaled

values enable consistent integration into the final prioritization score without distorting their intrinsic relationships. The normalized depth ($D_{\text{normalized}}$) is calculated as shown in (1).

$$D_{\text{normalized}} = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}} \quad (1)$$

In this equation, D_i is the observed variant read depth for the i th SNP; $D_{\min}$ is the minimum depth across all SNPs in the dataset; $D_{\max}$ is the maximum depth across all SNPs in the dataset; and $D_{\text{normalized}}$ is the scaled depth value between 0 and 1.

This normalization ensures that all depth values contribute proportionally when integrated into the final prioritization score.

2.3 Calculation of amino acid impact score

Amino acid impact score is calculated by integrating four biologically meaningful properties of amino acids: (i) molecular weight, (ii) hydrophobicity, (iii) polarity, and (iv) charge. The differences in molecular weight (W_{diff}) and hydrophobicity (H_{diff}) are normalized, according to (2) and (3), using the maximum observed ranges of these properties across natural amino acids: 204 Da (tryptophan) to 75 Da (glycine) for weight (range ≈ 130), and -4.5 to $+4.5$ for hydrophobicity (range = 9). Polarity and charge differences are treated as binary indicators (0, 1). These four features are integrated (4) into a unified value (amino acid impact score), reflecting the degree of physicochemical disruption introduced by the variation.

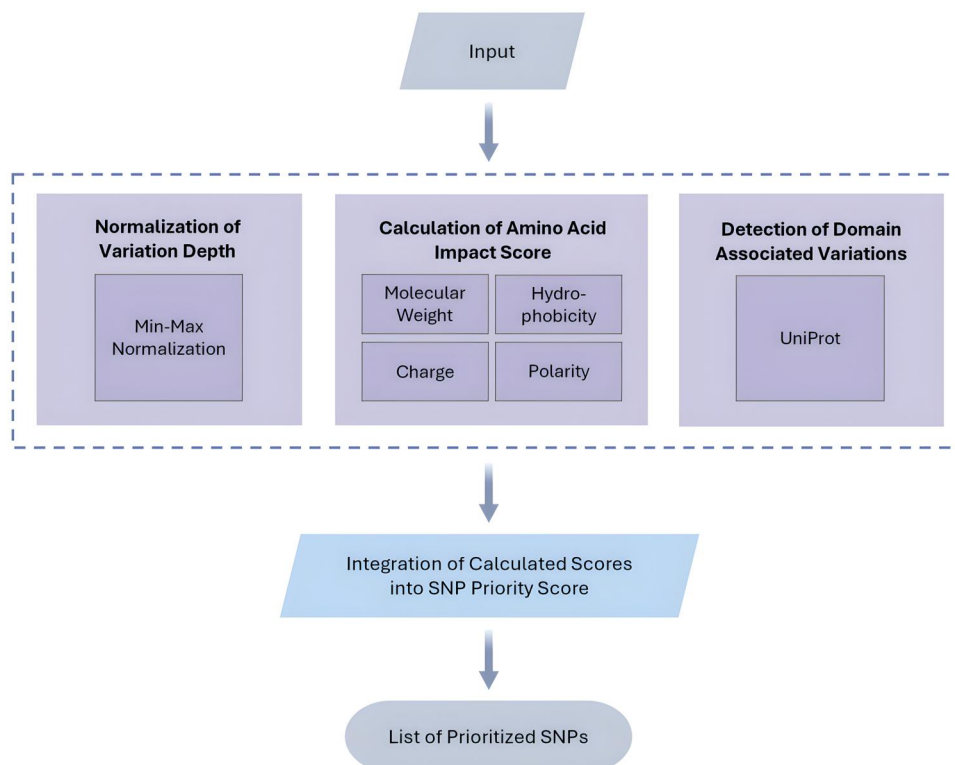


Figure 1. The workflow of SNPræfentia. The tool processes an input file of annotated variants. It separately calculates scores considering three key modules: variation depth, amino acid impact score, and presence of variation in the functional domain of the protein. These scores are consolidated to generate a final priority score for each variant.

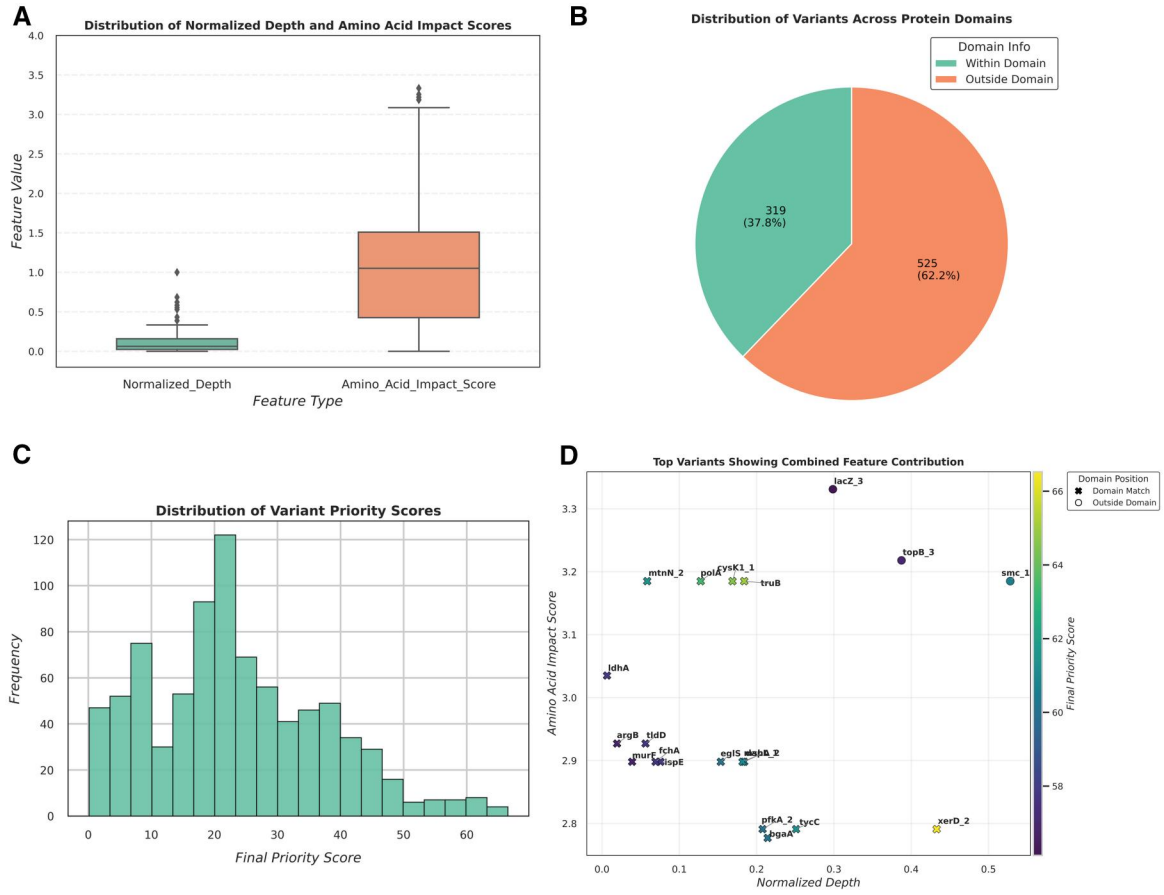


Figure 2. Graphical output generated by SNPraefentia. (A) Box plots showing the distribution of “Normalized_Depth” and “Amino_Acid_Impact_Score” across all variants. (B) Pie-chart illustrating the proportion of variants located within and outside conserved domains. (C) Histogram representing the distribution of “Final_Priority_Score” across the dataset. (D) Scatter plot integrating all feature contributions, displaying the relationship between “Normalized_Depth” (x-axis), “Amino_Acid_Impact_Score” (y-axis), “Domain_Position_Match” (point shape), and “Final_Priority_Score” (color intensity). Gene names are labeled for the top variants with the highest scores.

$$W_{\text{diff}} = \frac{|W_{\text{reference}} - W_{\text{mutated}}|}{130} \quad (2)$$

$$H_{\text{diff}} = \frac{|H_{\text{reference}} - H_{\text{mutated}}|}{9} \quad (3)$$

$$\text{Amino Acid Impact Score} = W_{\text{diff}} + H_{\text{diff}} + P_{\text{change}} + C_{\text{change}} \quad (4)$$

Here, $W_{\text{reference}}$ and W_{mutant} denote the molecular weights of the original and substituted amino acids. $H_{\text{reference}}$ and H_{mutant} represent respective hydrophobicity values. P_{change} is 1 if the polarity changes between the two residues, otherwise 0. C_{change} is 1 if the charge differs, otherwise 0.

2.4 Detection of domain-associated mutations

SNPraefentia determines whether a mutated amino acid resides within a known functional domain by querying domain annotations from the UniProt database. Gene, Taxonomic_ID, and Total_Protein_Length are used to identify the closest matching UniProt entry (UniProt_ID) via the UniProt REST API. If the variant residue falls within the functional domain region of the protein, the variant is assigned a score of 1; otherwise, 0. This binary feature, termed Domain_Position_Match, prioritizes variants that are likely to affect functionally relevant regions of microbial proteins.

To minimize potential biases from assembly errors or annotation inconsistencies, SNPraefentia, as mentioned, applies

a careful matching strategy that cross-references the bacterial species, gene identity, and total protein length to ensure accurate UniProt mapping. In cases where reliable domain information is unavailable, a neutral score of 0 is assigned instead of an inferred value, preventing artificial inflation of the prioritization score. Moreover, although domain mapping depends on UniProt annotations, the tool remains applicable to novel or less-characterized microbial species, like when domain data are missing, SNPraefentia integrates five additional biological features, each contributing to the final prioritization score. This design ensures stable and generalizable variant prioritization across both well-annotated and newly reconstructed microbial genomes.

2.5 Calculation of final priority score

The final priority score is a composite metric derived by integrating all relevant features into a single value (5). The scoring formula assigns double weight to the normalized depth, reflecting the increased confidence in high-frequency mutations, while the amino acid impact score and Domain Position Match are weighted equally. The total score is scaled by dividing by 7, the maximum possible sum of all weighted components, and multiplying it by 100, ensuring the output remains in the range of 0–100.

$$\text{Final Priority Score (\%)} = \frac{2 * D_{\text{normalized}} + \text{Amino Acid Impact Score} + \text{Domain Position Match}}{7} * 100 \quad (5)$$

Variants with higher scores are ranked as higher priority candidates for downstream characterization and experimental validation.

3 Working principle of the tool

For the demonstration, we processed an input dataset of 844 variants (Supplementary Table 1, available as supplementary data at *Bioinformatics Advances* online) using SNPraefentia. For successful execution of the analysis, users must provide an input file, the name of the target bacterial species (e.g. *Bacteroides uniformis*), and the output file. The `—input` or `-i` argument specifies the SNP file, while the `—specie` or `-s` argument is used for specifying the corresponding bacterial species. Additionally, the `—output` or `-o` argument defines the output file (Supplementary Table 2, available as supplementary data at *Bioinformatics Advances* online), which provides a comprehensive prioritization of input variants based on the mentioned parameters. The complete execution and progress of SNPraefentia during analysis is shown in Supplementary Fig. 1, available as supplementary data at *Bioinformatics Advances* online.

The primary output of SNPraefentia is the Final_Priority_Score, which serves as the core outcome of the tool, allowing users to quickly identify high-priority variants for downstream analysis. The score is derived from three supporting columns: Normalized_Depth, Amino_Acid_Impact_Score, and Domain_Position_Match, all of which are computed based on the input data and external annotations as described previously.

To provide a visual overview of these prioritization parameters, a comprehensive set of graphical summaries is generated by SNPraefentia (Fig. 2). The box plots display the overall distribution of Normalized_Depth and Amino_Acid_Impact_Score, enabling users to assess the variability and range of these continuous features. The pie chart depicts the proportion of variants occurring within and outside conserved domains, highlighting the structural relevance of prioritized mutations. The histogram illustrates the frequency distribution of variants across different Final_Priority_Score ranges, giving a clear view of the overall scoring landscape. Finally, the scatter plot integrates all feature contributions, visualizing how depth, impact score, and domain alignment collectively influence the final prioritization, with the top 20 variants and their corresponding genes annotated for user interpretation.

Acknowledgements

We thank the members of the Metagenomics Discovery Lab, SINES, NUST for their continuous feedback and support in this research.

Author contributions

Nadeem Khan (Data curation [equal], Formal analysis [equal], Methodology [equal], Software [equal], Writing—

original draft [equal]), Muhammad Muneeb Nasir (Data curation [equal], Formal analysis [equal], Methodology [equal], Software [equal]), Ammar Mushtaq (Supervision [equal], Writing—review & editing [equal]), and Masood ur Rehman Kayani (Project administration [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

All authors declare no competing interests.

Funding

No specific funding was obtained for this study.

Data availability

The toolkit and test data are freely available at <https://github.com/muneebdev7/SNPraefentia>.

References

- Bhunias PK, Verma D, Roy P *et al.* Mutation significance and assessment: mutation types and selection of mutants, importance in microbial technology. In: Verma P (ed.), *Industrial Microbiology and Biotechnology*. Singapore: Springer, 2024, 567–94. https://doi.org/10.1007/978-981-97-1912-9_23
- Farheen N, Sen N, Nair S *et al.* Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations. *Prog Biophys Mol Biol* 2017;128:14–23.
- Gong J, Liu C, Liu W *et al.* An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database* 2015;2015:bav029.
- Jee J, Rasouly A, Shamovsky I *et al.* Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* 2016; 534:693–6.
- Li H, Zhao L, Zhang M. Gut microbial SNPs induced by high-fiber diet dominate nutrition metabolism and environmental adaption of *Faecalibacterium prausnitzii* in obese children. *Front Microbiol* 2021;12:683714.
- Linz B, Sticht H, Tegtmeyer N *et al.* Cancer-associated SNPs in bacteria: lessons from *Helicobacter pylori*. *Trends Microbiol* 2024; 32:847–57.
- Miller S, Chiu C. The role of metagenomics and next-generation sequencing in infectious disease diagnosis. *Clin Chem* 2021; 68:115–24.
- Riccardi B, Resta S. The importance of the environment in the study of metagenomics and microbiome. *Pharmacophore* 2020;11:128–34.
- Seemann T. Snippy: rapid haploid variant calling and core SNP phylogeny. GitHub 2015. github.com/tseemann/snippy.
- Yauy K, Baux D, Pegeot H *et al.* MoBiDiC prioritization algorithm, a free, accessible, and efficient pipeline for single-nucleotide variant annotation and prioritization for next-generation sequencing routine molecular diagnosis. *J Mol Diagn* 2018;20:465–73.
- Zhang J, Knight R. Genomic mutations within the host microbiome: adaptive evolution or purifying selection. *Engineering* 2023; 20:96–102.

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics Advances, 2025, 00, 1–4

<https://doi.org/10.1093/bioadv/vbaf297>

Application Note