DOI: 10.25205/978-5-4437-1843-9-3

IDENTIFICATION OF EVOLUTIONARILY TOLERATED GENETIC VARIANTS USING PSEUDOREADS OF ORTHOLOGOUS SEQUENCES

D. S. Bug, A. V. Tishkov, N. V. Petukhova

Pavlov First Saint Petersburg State Medical University

✉ bug.dmitrii@yandex.ru

Abstract

We present GAPO, a novel method for detecting evolutionarily tolerated genetic variants by analyzing orthologous sequences. Applied to 272 cancer genes, GAPO identified 9,054 rare variants (98 % absent in gnomAD), revealing changes preserved across species but infrequent in humans. Unlike population variation databases, GAPO pinpoints variants permitted by long-term evolution, offering insights into benign genomic variation.

Introduction

The rapid advancement of molecular diagnostics has created a growing demand for improved variant interpretation methods. Current tools primarily focus on predicting the pathogenicity of coding and splicing variants, with limited attention given to non-coding regions—until recently, with the introduction of AlphaGenome. It is important to note that AlphaGenome does not incorporate evolutionary conservation into its pathogenicity predictions. Furthermore, existing variant effect predictors do not take into account specific sequence changes observed throughout evolution.

Previously, we developed the Genetic Alignment of Pseudoreads for Homologs (GAPH) method, which identifies specific sequence changes between human genes and homologous genomic sequences across species. However, the homology search was insufficiently selective, which may lead to capturing paralogs—evolutionarily related but functionally distinct genes—alongside orthologs, resulting in noisy datasets.

Materials and methods

For this study, we analyzed 272 classic cancer genes from the COSMIC database [1]. Ortholog identification was performed using "Pavlov's COGs", our Linux-based tool for evolutionary reconstruction based on clusters of orthologous groups (COGs). This tool leverages a local BLAST database of high-quality genomes pre-selected using BUSCO completeness scores. Orthologs were defined as genes belonging to the largest maximal clique in the orthology graph.

Orthologous gene sequences were retrieved using the NCBI Datasets utility [2] and fragmented into 70-nucleotide pseudoreads with a 35-nucleotide overlap. Each base was assigned a Phred quality score of 30. Pseudoreads were aligned to the corresponding human reference gene sequences using BWA-MEM (default parameters) [3]. Variant calling was performed with VarScan v2 [4], applying a minimum threshold of two supporting reads and 8x read-depth coverage. Finally, variants were annotated using the Franklin VCF-annotation tool.

Results and discussion

We present GAPO (Genetic Alignment of Pseudoreads for Orthologs), an advanced method for identifying evolutionarily conserved variants in cancer-related genes. Applying GAPO to 272 COSMIC classic cancer genes, we detected 9,054 genetic variants, most of which were classified as likely benign or variants of unknown significance (VUS). Only five variants were predicted as likely pathogenic, and no pathogenic mutations were identified (see table).

Notably, only 2 % (178/9,054) of the discovered variants were previously documented in population allele frequency databases (gnomAD, 1000 Genomes, ExAC). This stark discrepancy underscores GAPO's unique capacity to identify rare variants that have been evolutionarily permitted across species—a critical distinction from population databases, which primarily reflect variants that have survived recent human selection pressures. While population databases capture variants that are tolerated in modern human populations, GAPO reveals a deeper evolutionary perspective by detecting variants that have persisted across long-term species divergence, despite their rarity in human populations.

Our prior approach, GAPH, identified over 5 million variants across homologous sequences [5]. In contrast, GAPO reduces this to a highly refined set of 9,054 variants by enforcing strict orthology criteria and high-quality sequence inclusion. While this conservative approach enhances precision, adjusting GAPO's parameters (e. g., relaxing orthology thresholds) could expand variant detection for specific applications.

In conclusion, GAPO represents a significant methodological advancement for identifying potentially benign variants in orthologous sequences, offering a robust framework for evolutionary and clinical genomics.

Distribution of variants by their clinical relevance

Clinical relevance category	Number of variants
Benign	13
Likely benign	6
Possibly benign variants of unknown significance	8581
Uncertain significance	413
Possibly pathogenic variants of unknown significance	36
Likely pathogenic	5
Pathogenic	0

References

1. Tate J. G., Bamford S., Jubb H. C. et al. COSMIC: The catalogue of somatic mutations in cancer // Nucleic Acids Research. 2019. Vol. 47. P. D941–D947.

2. O'Leary N. A., Cox E., Holmes J. B. et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets // Scientific Data. 2024. Vol. 11. P. 732.

3. Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform // Bioinformatics. 2009. Vol. 25. P. 1754–1760.

4. Koboldt D. C., Zhang Q., Larson D. E. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing // Genome Research. 2012. Vol. 22. P. 568–576.

5. Bug D., Narkevich A., Petukhova N. Alignment of pseudoreads obtained from homologous sequences in identifying potentially tolerated genomic variants // Journal of Bioinformatics and Genomics. 2023. Vol. 21. P. 7.