أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

An Automated Pipeline for Functional Annotation and Phenotype Mapping

Eshraq Zakri, Hussein Mohsen, John A. Morris

eshraq.zakri@kaust.edu.sa, hussein.mohsen@utoronto.ca, john.morris@utoronto.ca

Abstract

Interpreting the functional impact of genetic variants remains a fundamental challenge in genomics. Functional annotation and phenotype mapping resources represent different channels of evidence on the potential biological impact of genomic variants. These resources are heterogeneous in structure and format, requiring access to multiple independent databases. In this project, we build a pipeline to annotate sets of single-nucleotide polymorphisms (SNPs) with a focus on associations resulting from genome-wide association studies (GWAS). The pipeline generates variant-centric reports capturing functional annotation and phenotype mapping. Functional annotations are generated using ANNOVAR, Ensembl VEP, and SnpEff, which characterize variants across multiple biological dimensions, including predicted molecular consequences, clinical significance, and regulatory effects across diverse ancestries. Phenotype mapping is retrieved from the GWAS Catalog, a curated repository of published GWAS findings, and OpenGWAS, a large collection of GWAS summary statistics enabling phenome-wide variant queries. We applied the pipeline to 222 genome-wide significant variants from a trans-population osteoporosis GWAS, mapping SNPs to 19 genes, including ones known to be associated with osteoporosis and BMD such as *ESR1*, *LRP5*, *SPP1*, *AXIN1*, and *COLEC10*. Phenotype mapping retrieved more than 200 variant–trait associations from the GWAS Catalog and over 1,000 from OpenGWAS, with strong intersection with previously reported osteoporosis and bone mineral density associations. The results demonstrate that the pipeline successfully annotates GWAS-significant variants and retrieves biologically meaningful associations. The pipeline is generalisable to any SNP set, consolidating functional and phenotype evidence into structured variant-level summaries that support variant prioritisation, cross-trait interpretation, and downstream genetic analyses.

Date: August 25, 2026

Contents

1 Introduction 2

2 Related Work 2

3 Methodology 2

3.1 Pipeline Overview . . . 3

3.2 Input Data . . . 3

3.3 Functional Annotation and Phenotype Mapping . . . 4

4 Results 4

5 Discussion 7

6 Limitations and Future Work 7

7 Conclusion 8

8 Acknowledgement 8

References 8

1 Introduction

Genome-wide association studies (GWAS) have identified millions of genetic variants associated with complex traits and diseases, yet the functional significance of most GWAS variants remains poorly understood [1, 2]. GWAS variant interpretation requires integrating evidence from multiple heterogeneous resources [2]. These resources include functional annotation databases that predict the biological effects of variants [3], clinical databases describing their medical significance [4], and variant–trait association databases reporting links to traits or diseases identified in genetic studies [5]. However, these resources differ in their structure, format, and the biological information they provide. Researchers often need to query multiple databases and combine their results to obtain a comprehensive interpretation of a variant.

Here, we develop a general-purpose pipeline for functional annotation and phenotype mapping of GWAS-identified single-nucleotide polymorphisms (SNPs). The pipeline was applied to significant variants from a trans-population osteoporosis GWAS (GWAS Catalog accession ID: GCST90480691). The pipeline is designed to be applicable to any SNP sets, and generates structured results based on evidence from multiple resources.

2 Related Work

Functional annotation of GWAS-associated variants is an important component of post-GWAS analysis. Tools such as Ensembl VEP [3] and SnpEff [6] have been widely adopted for predicting variant consequences at the gene, transcript, and protein levels. In addition, ANNOVAR enables annotation using a broad range of external databases [7]. Although widely adopted, these tools are typically applied independently, producing separate outputs across different formats. They may also produce different predictions for the same variant [8].

Thus, interpreting GWAS variants often requires integrating evidence beyond individual annotation tools. Zhao *et al*. developed an integrative framework that combined GWAS/PheWAS associations with diverse functional genomic annotations, including eQTLs, transcription factor binding sites, and enhancers [9]. The integration enabled the identification and characterisation of potentially regulatory variants, highlighting the value of combining multiple evidence sources for variant interpretation.

Phenotype mapping through phenome-wide association studies (PheWAS) has emerged as a complementary approach. Topaloudi *et al*. applied PheWAS to characterize the shared genetic architecture across autoimmune disorders, illustrating how querying variants across thousands of GWAS studies can reveal pleiotropic effects [10]. The OpenGWAS platform, which provides programmatic access to thousands of GWAS summary statistics, has become a key resource for PheWAS variant characterization [11]. Resources such as the GWAS Catalog further support this by providing curated variant–trait associations from published genetic studies, enabling the exploration of previously reported associations between variants and diverse traits [5].

Overall, previous work shows that researchers use multiple tools and sources of evidence to characterise GWAS-associated variants from different perspectives. Combining complementary sources can provide a more complete view of a variant than relying on a single source of evidence. This motivates a unified variant-centric workflow that combines functional consequence annotation and phenotype mapping evidence, providing a consistent framework for downstream variant interpretation.

3 Methodology

3.1 Pipeline Overview

The pipeline takes a set of GWAS-identified SNPs as input and generates variant-centric reports by running two parallel streams: functional annotation and phenotype mapping.

Three tools are used for functional annotation: SnpEff [6], Ensembl VEP [3], and ANNOVAR [7]. Two tools are used for phenotype mapping: the GWAS Catalog [5] and OpenGWAS [11] databases. The five tools operate independently and produce separate outputs, which together provide a multi-dimensional characterization of each variant.

The pipeline operates in the following steps:

1. **Input processing** — filter, standardize, and convert GWAS summary statistics into tool-compatible formats
2. **Functional annotation** — annotate variants using SnpEff, Ensembl VEP, and ANNOVAR
3. **Phenotype mapping** — query GWAS Catalog and OpenGWAS for known variant–trait associations
4. **Output** — structured reports per tool, ready for downstream analysis

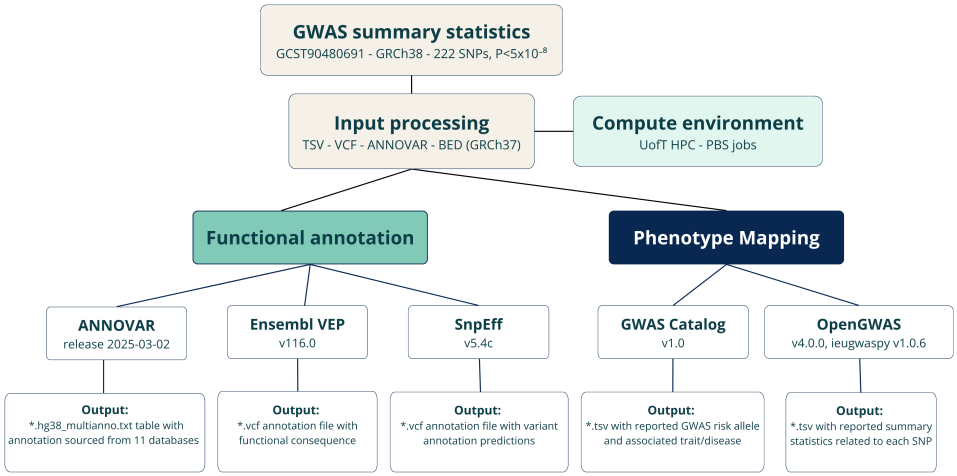


Figure 1 Pipeline workflow diagram.

3.2 Input Data

Summary statistics were obtained from a trans-population GWAS for osteoporosis available on the GWAS Catalog (accession GCST90480691). The original file contained 26,964,422 variant-level association records. Variants were filtered to retain significant associations ($P < 5 \times 10^{-8}$), yielding 222 SNPs.

Variants were standardized to single-nucleotide identifiers in CHROM:GENPOS:REF:ALT format. The processed TSV file was retained for the GWAS Catalog and converted into three

additional formats for downstream tool compatibility: VCF for SnpEff and Ensembl VEP, AVINPUT for ANNOVAR, and BED for OpenGWAS.

3.3 Functional Annotation and Phenotype Mapping

Functional annotation was performed using three annotation tools: SnpEff (v5.4c) [6], Ensembl Variant Effect Predictor (VEP; v116.0) [3], and ANNOVAR (release 2025-03-02) [7]. ANNOVAR annotation was performed using 11 databases, as detailed in Table 1. All three tools were executed locally.

Phenotype mapping was performed using two resources. The GWAS Catalog (v1.0) [5] was queried locally by matching input variants to the full association table by chromosome and position, retrieving the reported risk allele, associated trait, rsID, and allele match status for each hit. OpenGWAS (API v4.0.0, python wrapper ieugwaspy v1.0.6) [11] was queried against the `/phewas` API endpoint. Variants were submitted individually in CHR:POS format, with coordinates first lifted over to GRCh37/hg19 using UCSC LiftOver for compatibility. Results were filtered to $P \leq 1 \times 10^{-5}$.

Table 1 ANNOVAR databases used for functional annotation.

Database	Description
refGeneWithVer	Gene annotation
GTEx_v8_eQTL	Tissue-specific eQTL associations
ljb26_all, dbnsfp42c, revel	Functional and pathogenicity scores
gnomad41_genome, gnomad41_exome	Population allele frequencies
1000g2015aug, gme	Population-specific allele frequencies
clinvar_20250721	Clinical significance classifications
cosmic104	Somatic mutation overlap

All analyses were performed on the Morris Lab research computing cluster at the Donnelly Centre, University of Toronto.

4 Results

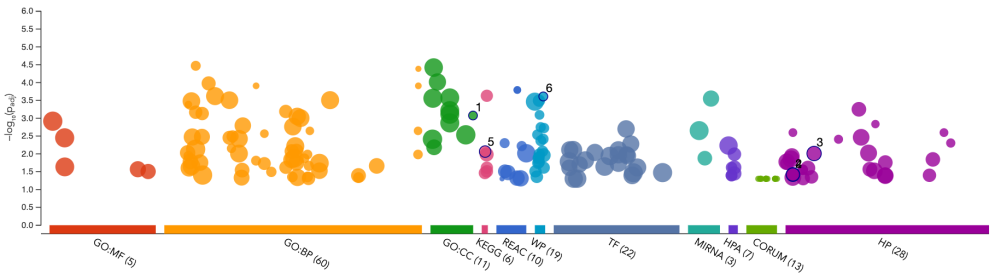
We test the pipeline on an osteoporosis GWAS that includes 222 significant SNPs ($p < 5 \times 10^{-8}$) out of 26,964,422 tested for association. The pipeline mapped the significant SNPs to 19 genes, including ones known to be associated with osteoporosis and BMD such as *ESR1*, *LRP5*, *SPP1*, *AXIN1*, and *COLEC10*, among others (Table 2).

Enrichment analysis using g:Profiler, which is a web-based tool for testing statistical associations with biological and phenotypic terms, returned a wide range of significantly associated terms with the 19 genes mapped by our pipeline. Amongst these are important functional and phenotypic terms related to osteoporosis (Figure 2). A cluster of terms also emerged in relation to the disease and its related biological pathways, including WNT and the Vitamin D receptor pathway. Other terms are also associated with bone structure, density, and morphology (Figure 3).

In terms of phenotypic mapping, GWAS Catalog and OpenGWAS results retrieved by our pipeline demonstrate a strong intersection with osteoporosis and bone density associations in previous studies. These include 241 associations in the GWAS Catalog and >1000 in OpenGWAS, further highlighting the validity of the pipeline results.

Table 2 Example genes mapped by the annotation pipeline from a GWAS study of osteoporosis, a bone disease. Mapped genes have known associations in the literature with the disease or bone mass density (BMD).

Gene	Trait	Citation
SPP1	Both	Chen *et al*. 2014
AXIN1	Both	Nassar *et al*. 2023
ESR1	Both	Li *et al*. 2010
LRP5	Both	Li *et al*. 2010
COLEC10	BMD	Xu *et al*. 2025
AKAP11	Both	Correa-Rodríguez *et al*. 2018, *Int J Med Sci*
CYP19A1	Both	Li *et al*. 2010, *Hum Genet*
MEPE	Both	Surakka *et al*. 2020, *Nat Commun*
CCDC170	Both	Liu *et al*. 2021, *Biol Open*



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:CC	GO:1990909	Wnt signalosome	8.405×10^{-4}
2	HP	HP:0000939	Osteoporosis	3.745×10^{-2}
3	HP	HP:0004348	Abnormality of bone mineral density	9.763×10^{-3}
4	HP	HP:0000938	Osteopenia	3.847×10^{-2}
5	KEGG	KEGG:04512	ECM-receptor interaction	8.616×10^{-3}
6	WP	WP:WP5368	Sulfatase and aromatase pathway	2.454×10^{-4}

version	e114_eg62_p19_27110d83
date	05/08/2026, 16:40:16
organism	hsapiens

g:Profiler

Figure 2 g:Profiler enrichment analysis results of the 19 genes mapped by the annotation pipeline.

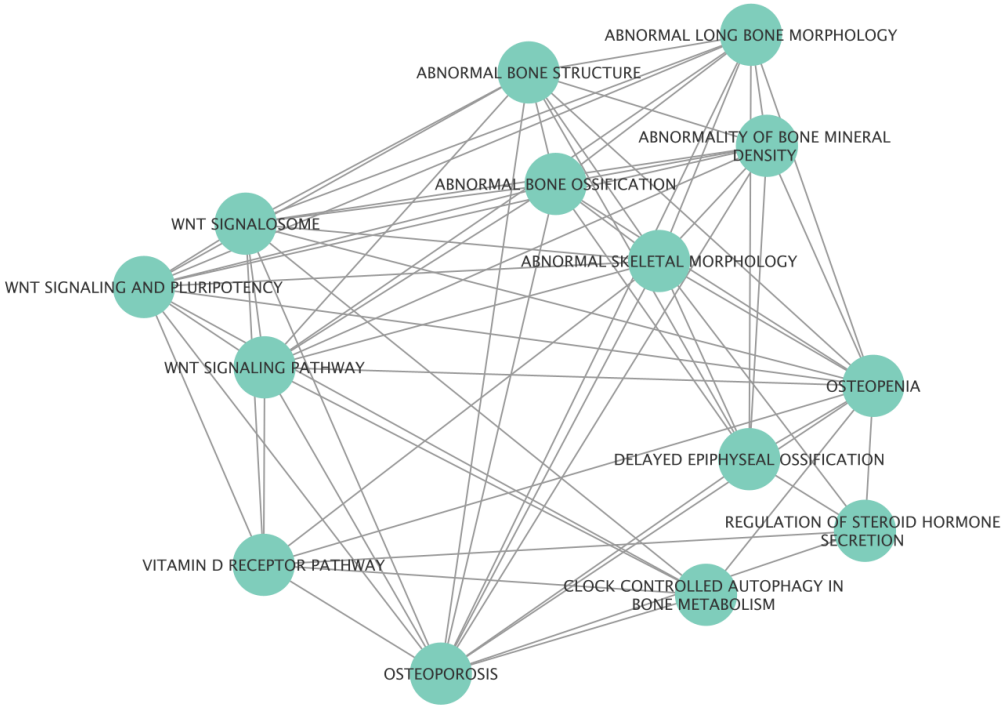


Figure 3 Cluster of significant enrichment results related to osteoporosis, its pathways, and bone properties. Edges represent shared genes driving the annotations.

5 Discussion

The results demonstrate that the pipeline can provide complementary evidence for characterising GWAS-associated variants from both functional and phenotype perspectives. Applied to 222 genome-wide significant SNPs from a trans-population osteoporosis GWAS, the pipeline mapped variants to 19 genes, including *ESR1*, *LRP5*, *SPP1*, *AXIN1*, and *COLEC10*, which have previously been associated with osteoporosis or BMD. The enrichment results further showed that the mapped genes were associated with terms related to osteoporosis, bone structure, density, and morphology, as well as pathways including WNT and the vitamin D receptor pathway. These findings are consistent with the biological context of the input GWAS and indicate that the annotations retrieved by the pipeline can provide biologically relevant context for the associated variants.

Using multiple functional annotation tools provides information from different annotation frameworks and databases. SnpEff and Ensembl VEP provide predicted molecular and transcript-level consequences, while ANNOVAR extends the annotation using multiple external databases, including eQTL, functional and pathogenicity scores, clinical and machine learning classifications, and other genomic resources. The outputs of the different annotation tools may provide overlapping or complementary information, depending on the type of annotation and database used. This is consistent with previous work showing that annotation tools may produce different predictions and that interpretation of GWAS variants benefits from considering multiple sources of evidence [8]. The phenotype mapping component provides a complementary perspective by identifying previously reported associations between variants and traits. The strong intersection of the retrieved GWAS Catalog and OpenGWAS associations with osteoporosis and bone density provides additional context for interpreting the variants identified in the input GWAS.

The results therefore illustrate the benefit of considering functional annotation and phenotype mapping. Functional annotation can indicate the potential biological consequences or genomic context of a variant, whereas phenotype mapping provides evidence about previously observed relationships between variants and phenotypes. However, the present pipeline does not establish that an annotated variant is causal or functionally responsible for a phenotype. Rather, the different evidence sources provide information that can be used to support subsequent investigation and prioritisation of candidate variants.

6 Limitations and Future Work

A limitation of the current pipeline is that the annotation and phenotype-mapping tools operate independently and generate separate output files. Although these outputs collectively provide multiple dimensions of evidence, they do not currently produce a single unified record containing all available evidence for each variant. This means that further processing is required when comparing or integrating annotations across tools. This limitation is particularly relevant because the tools also require different input formats and provide outputs with different structures and types of information.

Future work can focus on extending the pipeline toward fine-mapping of GWAS-associated loci. The current pipeline provides functional and phenotype-level evidence that can be used as additional information when prioritising variants within associated regions.

7 Conclusion

This project developed a pipeline for functional annotation and phenotype mapping of GWAS-identified SNPs using SnpEff, Ensembl VEP, ANNOVAR, the GWAS Catalog, and OpenGWAS. Applied to 222 genome-wide significant variants from a trans-population osteoporosis GWAS, the pipeline identified genes and phenotype associations that were consistent with the biological context of osteoporosis and bone mineral density. The results demonstrate the value of examining functional and phenotype evidence from multiple complementary resources when characterising GWAS-associated variants.

The completed pipeline provides a foundation for downstream variant prioritisation and fine-mapping. Future integration of statistical fine-mapping with functional annotation can further narrow candidate variants and support the investigation of potentially causal variants.

8 Acknowledgement

I would like to sincerely thank the **Morris Lab at the Donnelly Centre, University of Toronto**, for providing the research environment and computational resources that supported this project.

I am grateful to **Hussein Mohsen** and **John A. Morris** for their supervision, guidance, and valuable feedback throughout the project.

I would also like to thank the **KAUST Academy**, **Data Science Institute**, and **RZM Investment** for their support and for providing the opportunity to undertake this research project.

I also acknowledge the support of the **Alagil Foundation** and the institutions involved in supporting this research opportunity.

References

[1] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[2] Mark I McCarthy et al. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature Reviews Genetics* 9.5 (2008), pp. 356–369.

[3] William McLaren et al. “The Ensembl variant effect predictor”. In: *Genome Biology* 17.1 (2016), p. 122.

[4] Melissa J Landrum et al. “ClinVar: public archive of interpretations of clinically relevant variants”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D862–D868.

[5] Elliott Sollis et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D977–D985.

[6] Pablo Cingolani et al. “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff”. In: *Fly* 6.2 (2012), pp. 80–92.

[7] Kai Wang, Mingyao Li, and Hakon Hakonarson. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. In: *Nucleic Acids Research* 38.16 (2010), e164.

[8] Eleftherios Pilalis et al. “Genome-wide functional annotation of variants: a systematic review of state-of-the-art tools, techniques and resources”. In: *Frontiers in Pharmacology* 16 (2025), p. 1474026.

[9] Junfei Zhao et al. “An integrative functional genomics framework for effective identification of novel regulatory variants in genome-phenome studies”. In: *Genome Medicine* 10 (2018), p. 47.

[10] Anastasia Topaloudi et al. “PheWAS and cross-disorder analysis reveal genetic architecture, pleiotropic loci and phenotypic correlations across 11 autoimmune disorders”. In: *Frontiers in Immunology* 14 (2023), p. 1147573.

[11] Benjamin Elsworth et al. “The MRC IEU OpenGWAS data infrastructure”. In: *bioRxiv* (2020).