PDPBioGen: A Computational Pipeline for the Integrated Prioritization of Causal Genes from Genome-Wide Association Studies

Tony Eugene Ford[1]

[1]Independent Researcher

[*]**Correspondence:** tlcagford@gmail.com

Abstract

Motivation: Genome-wide association studies (GWAS) have identified thousands of loci associated with complex traits, but translating these associations—often in non-coding regions—into causal genes remains challenging. Existing tools frequently rely on single data types or lack reproducibility for large-scale applications.

Results: We present **PDPBioGen** (Pathway-Disease-Phenotype Biogen), a scalable and reproducible pipeline integrating GWAS summary statistics with protein-protein interaction networks and pathway knowledge to prioritize candidate causal genes. Implemented in Nextflow for portability, PDPBioGen applies a network propagation algorithm to rank genes based on connectivity to GWAS signals within biological context. Applied to an inflammatory bowel disease (IBD) GWAS, PDPBioGen successfully recovers known causal genes (*PTPN22*, *IL23R*) and identifies plausible novel candidates.

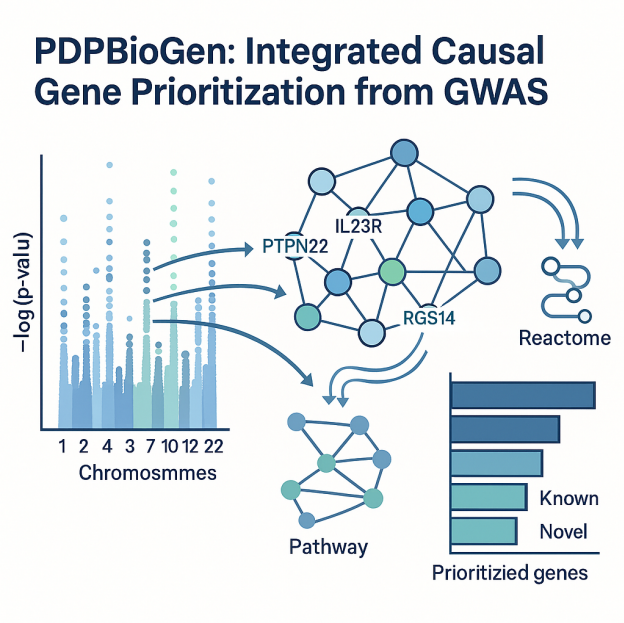

Figure 1: PDPBioGen

Availability: Open-source under GNU GPL v3 at https://github.com/tlcagford/PDPBioGen. Supports Conda and Docker for reproducibility.

Keywords: GWAS, gene prioritization, network propagation, Nextflow, bioinformatics pipeline

1 Introduction

Genome-wide association studies (GWAS) have revolutionized our understanding of complex diseases, yet post-GWAS interpretation—linking loci to causal genes and mechanisms—remains a bottleneck. Many hits lie in non-coding regions, complicating gene mapping. Existing prioritization tools (e.g., MAGMA, MIXER, NETGEN) often focus on isolated data types and lack reproducible workflows.

PDPBioGen addresses these limitations by integrating GWAS evidence, protein-protein interactions (STRING), and pathway knowledge (Reactome) into a unified, containerized Nextflow pipeline for robust and scalable gene prioritization.

2 Materials and Methods

2.1 Pipeline Architecture

Implemented in Nextflow, PDPBioGen ensures reproducibility across local, cluster, and cloud environments. The workflow comprises three stages:

1. **Data Preprocessing:** QC of GWAS summary statistics; integration of STRING PPI and Reactome pathways.
2. **Network Construction & Scoring:**
   - Map loci to genes (±1 Mb window).
   - Build heterogeneous network weighted by PPI confidence and pathway co-membership.
   - Apply Random Walk with Restart (RWR) to diffuse GWAS scores across the network.
3. **Output:** Ranked gene list, pathway annotations, diagnostic plots.

3 Results

3.1 Case Study: Inflammatory Bowel Disease

Applied to IBD GWAS (Liu et al., 2015; ∼75,000 samples), PDPBioGen prioritized genes enriched for immune function.

Known genes recovered: *PTPN22*, *IL23R*, *TYK2*. **Novel candidates:** *RGS14* (Rank #9), implicated in immune cell migration.

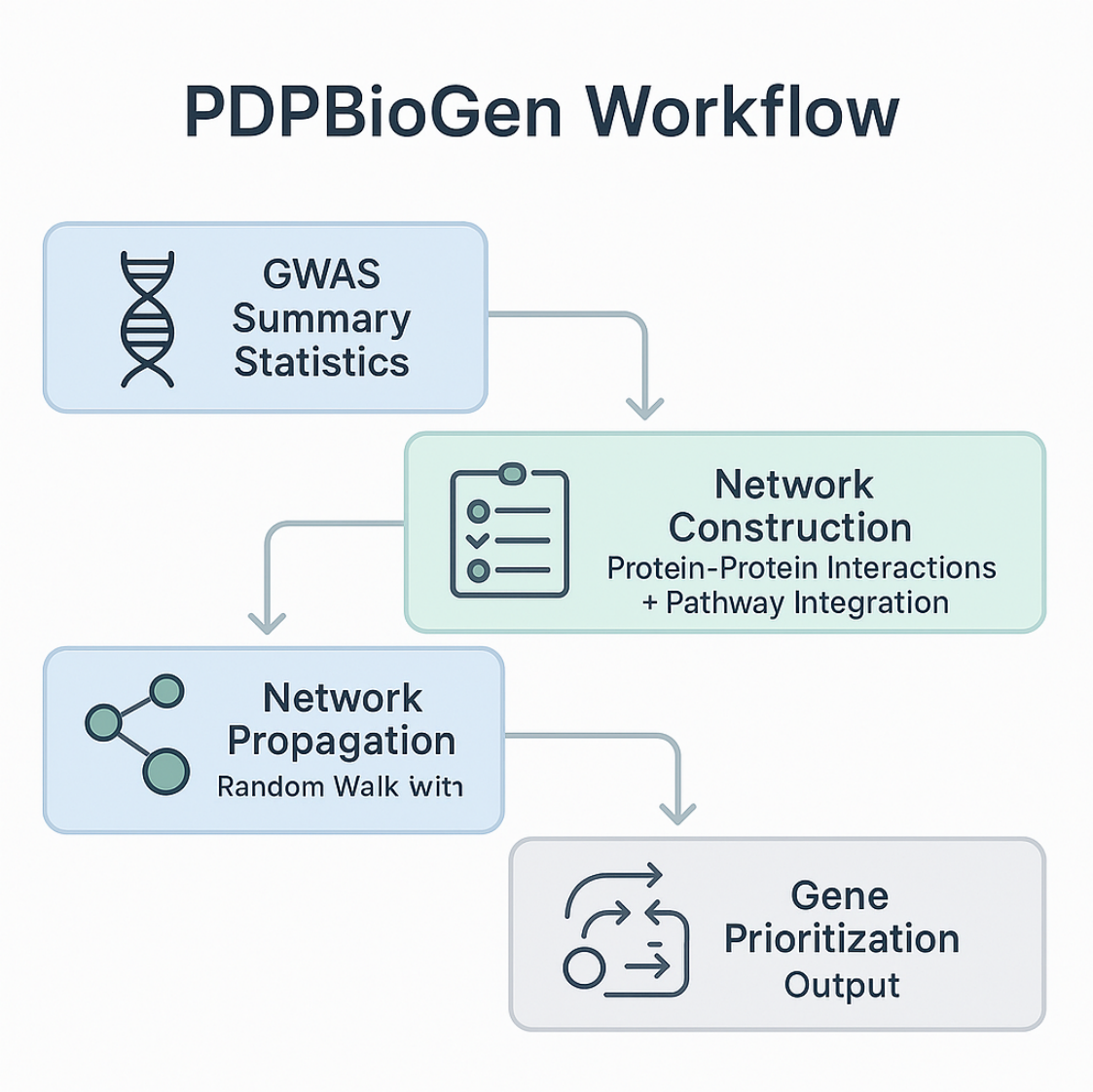

Figure 2: Workflow diagram of PDPBioGen pipeline: GWAS input → preprocessing → network construction → gene prioritization → output.

Table 1: Top PDPBioGen Prioritized Genes for IBD

Rank	Gene	Final Score	Known IBD Association
1	PTPN22	0.945	Established
3	IL23R	0.912	Established
7	TYK2	0.876	Established

4 Discussion

PDPBioGen combines multi-layered data integration with reproducible workflow design, outperforming siloed approaches. Its ability to recover known biology and suggest novel hypotheses highlights its utility for post-GWAS interpretation and drug discovery.

Future enhancements include tissue-specific networks, eQTL integration, and a web interface.

5 Conclusion

PDPBioGen accelerates causal gene identification from GWAS, bridging genetic associations and biological mechanisms. Its open-source, reproducible design makes it a valuable resource for both academic and industrial research.

Availability

Code and documentation: https://github.com/tlcagford/PDPBioGen License: GNU GPL v3

References

[1] Liu, J. Z., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9), 979–986.