Integrating Temporal Gene Expression Dynamics with GWAS Signals for Causal Gene Prioritization in Type 2 Diabetes (T2D)

March 3, 2025

1 Background and Introduction

Type 2 diabetes (T2D) is a global health crisis, affecting over 500 million individuals worldwide, and driving significant morbidity through its vascular and metabolic complications (Imamura & Maeda, 2024). T2D is a polygenic disease, where the interplay of genetic and environmental factors contributes to its onset and progression. Genome-wide association studies (GWAS) have identified hundreds of loci associated with T2D risk (Ghatan et al., 2024), with many of these loci residing in non-coding regions of the genome. These non-coding regions are often poorly understood in terms of their functional roles and the genes they regulate (Grotz, Gloyn, & Thomsen, 2017).

Current prioritization tools, such as MAGMA, primarily rely on static functional annotations (e.g., chromatin accessibility, tissue-specific gene expression) to link genetic variants to potential causal genes (Kolosov, Daly, & Artomov, 2021; Sarsani et al., 2024). However, these approaches fail to capture the dynamic, context-dependent nature of gene regulation, which is critical in diseases like T2D. For example, glucose-responsive enhancer activity in pancreatic beta cells or epigenetic changes in adipocytes during insulin resistance are aspects that traditional tools cannot account for (Sarsani et al., 2024). Furthermore, genes like *TCF7L2*(del Bosque-Plata, Martínez-Martínez, Espinoza-Camacho, & Gragnoli, 2021), a strong T2D-associated locus, are involved in insulin secretion specifically under hyperglycemic conditions highlighting the temporal nature of gene regulation in T2D (Grotz et al., 2017).

Additionally, a major limitation in T2D GWAS is the population bias. Over 90% of T2D GWAS data are derived from individuals of European ancestry, ignoring the genetic and regulatory diversity present in other populations, which is crucial for understanding the full spectrum of T2D risk variants (Ghatan et al., 2024). This limitation results in a lack of representation of high-risk populations, such as those of African descent.

Emerging machine learning (ML) techniques, such as positive-unlabeled learning (PU-learning) (Kolosov et al., 2021) and graph neural networks (GNNs) (Nicholls et al., 2020), offer potential solutions to address some of these challenges by integrating diverse multi-omics data (including single-cell epigenomics and metabolomics) and resolving context-specific gene-variant relationships. However, most tools have not incorporated temporal regulation, such as circadian rhythms or metabolic stress, or validated predictions in disease-relevant experimental systems. In this work, we explore how next-generation ML approaches, coupled with time-resolved multi-omics data and diverse population cohorts can bridge the gap between T2D-associated loci and their mechanistic roles. We aim to prioritize T2D risk variants and genes by accounting for cellular context, environmental triggers, and temporal regulation, thus accelerating the discovery of novel therapeutic targets and advancing precision medicine for T2D.

2 Literature Review

This review synthesizes advancements in genome-wide association studies (GWAS), machine learning (ML), and prioritization strategies, particularly in the context of diabetes.

2.1 Machine Learning Approaches in GWAS

Machine learning has significantly improved the ability to prioritize SNPs, predict disease risk, and annotate genetic variants. Some key studies in this field include:

- **Deep learning for noncoding variants**: (Zhou & Troyanskaya, 2015) developed *DeepSEA*, a convolutional neural network (CNN) that predicts the chromatin effects of noncoding SNPs (e.g., transcription factor binding) with high accuracy (AUC = 0.95). This tool prioritized T2D-associated SNPs near *TCF7L2*, though it lacked validation in pancreatic islets.

- **PU-learning for gene prioritization**: (Kolosov et al., 2021) introduced *GPrior*, an ensemble method that uses positive-unlabeled learning to prioritize genes for schizophrenia and IBD. While it has not been applied to T2D, its integration of tissue-specific features (e.g., GTEx) makes it promising for loci such as *TCF7L2*.

- **Clinical-genomic integration**: (Pedersen et al., 2016) used artificial neural networks (ANNs) to predict T2D remission after bariatric surgery by integrating clinical traits (e.g., HbA1c) with genomic data. The model prioritized *ABCA1* and *STXBP5L*, but was limited by a European-centric cohort.

- **Ensemble methods**: (Nicholls et al., 2020) showed that ensemble models (e.g., Random Forests, Gradient Boosting) outperformed single algorithms in prioritizing causal loci (AUROC = 0.75–0.96). Challenges include class imbalance and lack of interpretability.

- **Systematic review of ML in GWAS**: (Enoma, Bishung, Abiodun, Ogunlana, & Osamor, 2022) highlighted ML's superiority in detecting epistasis and addressing the underrepresentation of African genomes, which is critical for T2D research.

2.2 Prioritization Strategies for Complex Diseases

Recent prioritization frameworks have focused on integrating multi-omics data and addressing locus heterogeneity. Some notable contributions include:

- **Cross-ancestry fine-mapping**: (Sarsani et al., 2024) developed *GPScore*, a tool that prioritizes adiponectin-associated genes (e.g., *CYP2R1*) using cross-ancestry GWAS data. However, European LD reference bias limited its trans-ethnic resolution.

- **Functional genomics**: (Grotz et al., 2017) combined fine-mapping (SuSiE), CRISPR screens, and chromatin conformation to validate T2D genes like *STARD10* and *PDX1*.

- **PRS clustering**: (Ghatan et al., 2024) stratified T2D loci into five pathways using metabolic trait colocalization. Lipodystrophic PRS was linked to lower BMI, though non-European validation is lacking.

2.3 Diabetes-Focused Applications

Machine learning and prioritization strategies have been applied to T2D research, with some notable contributions including:

- **Translational review**: (Imamura & Maeda, 2024) discussed the poor performance of European-derived polygenic risk scores (PRS) in diverse populations, advocating for more diverse cohort representation.

- **Multi-omics integration**: (Pedersen et al., 2016) and (Sarsani et al., 2024) showed improved accuracy by integrating genomic, clinical, and metabolomic data.

- **Therapeutic target discovery**: (Grotz et al., 2017) validated *STARD10* via CRISPR, while (Zhou & Troyanskaya, 2015) prioritized *TCF7L2* regulatory SNPs for experimental testing.

2.4 Cross-Cutting Challenges

Some of the major challenges in the field include:

- **Population diversity**: The over-representation of European cohorts in GWAS data (Ghatan et al., 2024) can be mitigated by federated learning across diverse biobanks (e.g., H3Africa).

- **Dynamic modeling**: Static annotations (Kolosov et al., 2021) fail to capture context-specific regulation. Future work should integrate single-cell epigenomics.

- **Validation gap**: Genes like *STXBP5L* lack CRISPR validation in disease-relevant tissues.
- **Interpretability**: Black-box models (Zhou & Troyanskaya, 2015) obscure mechanistic understanding, and explainable AI (XAI) could be a solution.

Machine learning-driven tools (e.g., (Kolosov et al., 2021), (Zhou & Troyanskaya, 2015)) and multi-omics integration have advanced our understanding of T2D genetics, but several gaps remain. Future work should focus on addressing population bias, improving dynamic modeling, and validating predictions in diverse experimental systems.

3 Problem Statement

Despite significant advances in identifying T2D-associated genetic variants and various methods for causal gene prioritization, a critical gap remains in the integration of temporal gene expression dynamics with genetic association data. Current approaches largely focus on static genetic associations and tissue-specific expression patterns, but fail to capture the dynamic nature of gene regulation during disease progression. This limitation is particularly important because T2D development involves complex temporal changes in multiple tissues. Existing methods don't adequately account for the temporal sequence of metabolic dysregulation leading to T2D, potentially missing important causal genes that may be active only during specific stages of disease development.

4 Proposed Method

We propose a hybrid machine learning framework that integrates temporal transcriptomic data with GWAS signals to prioritize causal genes associated with Type 2 Diabetes (T2D). The framework will consist of the following steps:

1. **Data Integration**: Integrate GWAS summary statistics from the DIAGRAM Consortium with time-series RNA-seq data from pancreatic islets (HPAP) and adipocytes, capturing dynamic gene expression patterns during disease progression.
2. **Dynamic Network Analysis**: Utilize graph neural networks (GNNs) to model temporal gene co-expression modules and their relationships to key disease processes, such as insulin resistance and beta-cell dysfunction.
3. **PU-Learning Prioritization**: Apply ensemble models based on positive-unlabeled learning (PU-learning) like GPrior to rank candidate genes based on both static GWAS signals and dynamic expression patterns.
4. **Validation**: Validate the top-ranked genes through CRISPR screens in iPSC-derived beta cells to confirm their functional relevance in T2D.

5 Available Datasets

- **DIAGRAM consortium T2D GWAS summary statistics** – Large-scale T2D genetic data for identifying risk loci.
- **UK Biobank T2D GWAS summary statistics** – Additional genetic data from a large cohort to complement DIAGRAM findings.
- **GTEx tissue-specific expression data** – Maps genetic variants to gene expression across tissues.
- **HPAP pancreatic islet time-series RNA-seq data** – Captures temporal gene expression in pancreatic islets for T2D understanding.
- **Adipocyte temporal expression data** – Provides insights into gene regulation in adipose tissue, crucial for insulin resistance.
- **Islet-specific eQTLs and chromatin accessibility data** – Links genetic variants to regulatory functions in islets.

- **Human islet single-cell RNA-seq datasets** – Enables analysis of cell-type-specific gene expression in T2D.
- **CRISPR validation datasets** – Validates predicted causal genes in experimental models.
- **AGEN (Asian Genetic Epidemiology Network) consortium** – Provides genetic data from Asian populations for cross-ancestry validation.
- **H3Africa (Human Heredity and Health in Africa) consortium** – Includes data from African populations to address underrepresentation in genetic research.

6 Conclusion

This project aims to address a critical gap in the temporal modeling of T2D genetics by integrating GWAS data with time-resolved transcriptomics. The proposed machine learning framework will prioritize context-dependent causal genes (e.g., *TCF7L2*) and tackle population bias by incorporating data from diverse cohorts, such as H3Africa. In the next phase, we will validate these predictions through CRISPR screens in relevant tissues and extend our analysis to include non-European populations to ensure the robustness and inclusivity of our findings.

References

del Bosque-Plata, L., Martínez-Martínez, E., Espinoza-Camacho, M. Á., & Gragnoli, C. (2021). The role of tcf7l2 in type 2 diabetes. *Diabetes*, *70*(6), 1220–1228.

Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., & Osamor, V. C. (2022). Machine learning approaches to genome-wide association studies. *Journal of King Saud University-Science*, *34*(4), 101847.

Ghatan, S., van Rooij, J., van Hoek, M., Boer, C. G., Felix, J. F., Kavousi, M., ... others (2024). Defining type 2 diabetes polygenic risk scores through colocalization and network-based clustering of metabolic trait genetic associations. *Genome Medicine*, *16*(1), 10.

Grotz, A. K., Gloyn, A. L., & Thomsen, S. K. (2017). Prioritising causal genes at type 2 diabetes risk loci. *Current diabetes reports*, *17*, 1–9.

Imamura, M., & Maeda, S. (2024). Perspectives on genetic studies of type 2 diabetes from the genome-wide association studies era to precision medicine. *Journal of Diabetes Investigation*, *15*(4), 410–422.

Kolosov, N., Daly, M. J., & Artomov, M. (2021). Prioritization of disease genes from gwas using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics*, *29*(10), 1527–1535.

Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., & Cabrera, C. P. (2020). Reaching the end-game for gwas: machine learning approaches for the prioritization of complex disease loci. *Frontiers in genetics*, *11*, 350.

Pedersen, H. K., Gudmundsdottir, V., Pedersen, M. K., Brorsson, C., Brunak, S., & Gupta, R. (2016). Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. *NPJ genomic medicine*, *1*(1), 1–8.

Sarsani, V., Brotman, S. M., Xianyong, Y., Silva, L. F., Laakso, M., & Spracklen, C. N. (2024). A cross-ancestry genome-wide meta-analysis, fine-mapping, and gene prioritization approach to characterize the genetic architecture of adiponectin. *Human Genetics and Genomics Advances*, *5*(1).

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, *12*(10), 931–934.

Group 17

KAJAANI B. 200279N
KOBINATH A. 200308F,
RAJEEVAN Y. 200501P
VIBULAN J. 200677H