

遺伝統計学チームの研究ミッション

生命科学（医学や農学）のビッグデータを機械学習・人工知能ベースの遺伝統計学で解析を行い、ヒトの疾患や農業形質などの複雑な現象に寄与する要因を探り出す。今回、我々研究チームの取り組む課題のうち、疾患リスク予測を代表して紹介する。

I 大規模ゲノムコホートデータへの疾患リスク予測手法の適用

背景

近年、ゲノムデータを活用した疾患リスク予測の有用性が注目を集めている。当チームが開発したSTMGP法（smooth-threshold multivariate genetic prediction）^[1]は、他の手法と比較して高い予測精度を示しているが、疾患リスク予測システムの社会実装の実現に向けて、より多くの疾患、より大規模なデータでの検証が求められる。

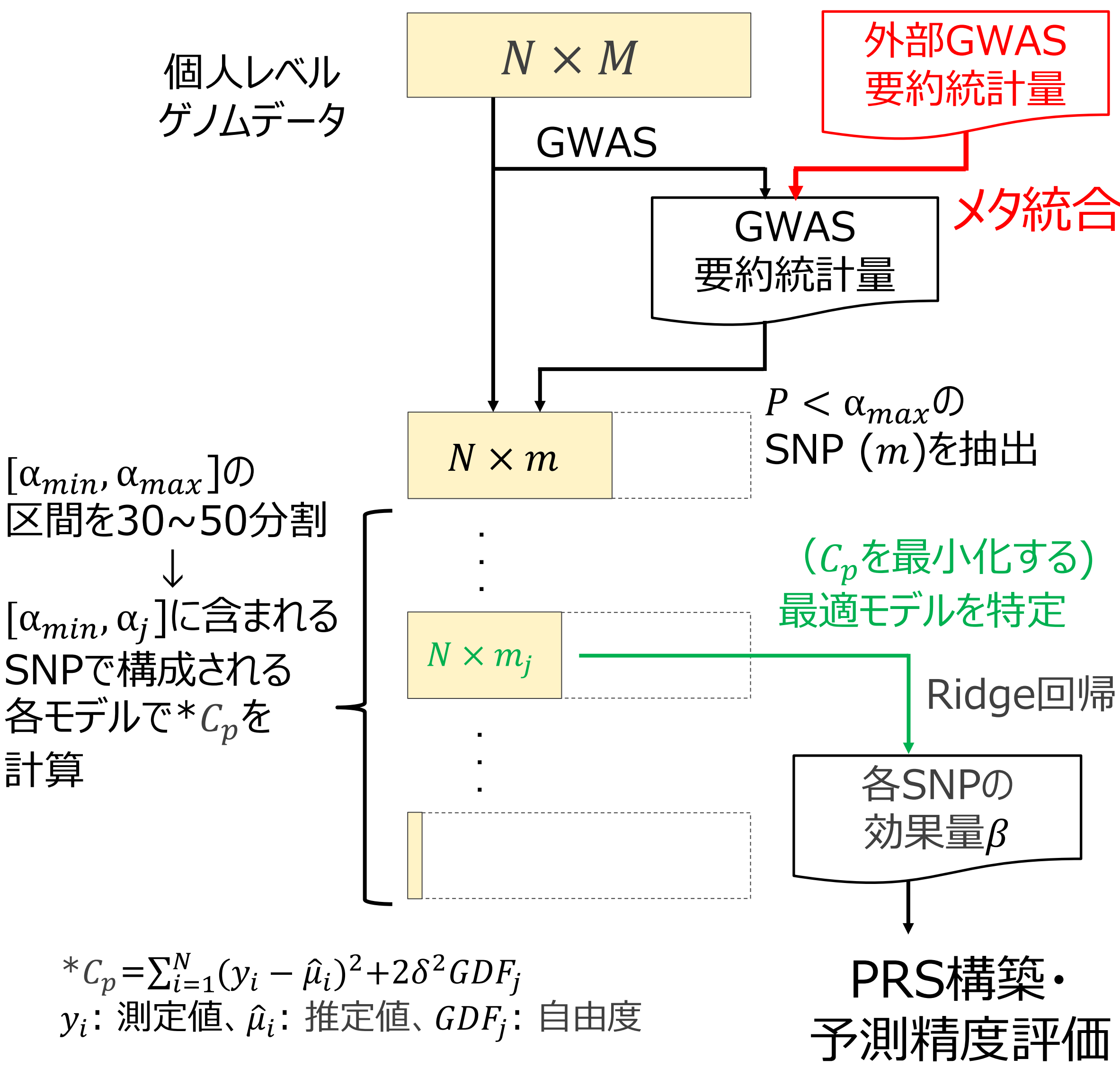
これまでの進捗

【2022年度までの成果】
UKバイオバンク（約14万人）および東北メディカル・メガバンク計画 地域住民コホート（約3万人）の大規模データを用いて検討した結果、STMGP法が他の手法よりも高い予測精度を示すことを確認した。
また、計算負荷の軽減のため、分割計算を可能とする機能を実装した。東北大学のデータによる検証では、データを二分割しても同等、もしくはそれ以上の予測精度が得られ、かつ使用メモリ量の削減が可能であった。
【2023年度の成果】
さらなる予測精度の向上を目指し、外部で公開されているGWAS（ゲノムワイド関連解析）の要約統計量をSTMGP法に取り込む機能を実装した。LDL-CのGWAS要約統計量を活用した検証の結果、予測精度の向上が見られた。

今後の予定

データ分割と、外部のGWAS要約統計量の取り込みを同時に行うアルゴリズムを実装し、計算負荷の増大を抑えつつ、予測精度のさらなる向上を目指す。同時に、STMGP法における尤度関数の最適化および行列分解アルゴリズムを改変し、計算の効率化を図る。

外部のGWAS要約統計量を利用したSTMGP法



対象データ
形質: LDL-C
個人レベルデータ:
東北メディカル・メガバンク計画 地域住民コホート29,135人/
112,860,112座位
外部GWAS要約統計量:
GLGC (Global Lipids Genetics Consortium) 国際メタGWAS^[2]
約165万人/49,757,577座位

メタ統合
個人レベルデータと外部GWAS要約統計量で共通する18,246,488座位の効果量 β をMETAL^[3]による固定効果モデルで統合

外部GWAS要約統計量	m	Adjusted R^2 (95%CI)
なし	10,766	0.041 (0.022-0.065)
あり	109,661	0.044 (0.025-0.070)

外部GWAS要約統計量を取り込むことで、より多くの座位が予測モデルに用いられ、予測精度の向上につながった。

[1] Ueki M & Tamiya G. Genet Epidemiol 40:233-243, 2016.
[2] Graham SE et al. Nature 600:675-679, 2021.
[3] Willer CJ et al. Bioinformatics 26:2190-2191, 2010.

II 疾患パスウェイの混合を考慮した変数選択法による糖尿病臨床的サブクラス予測

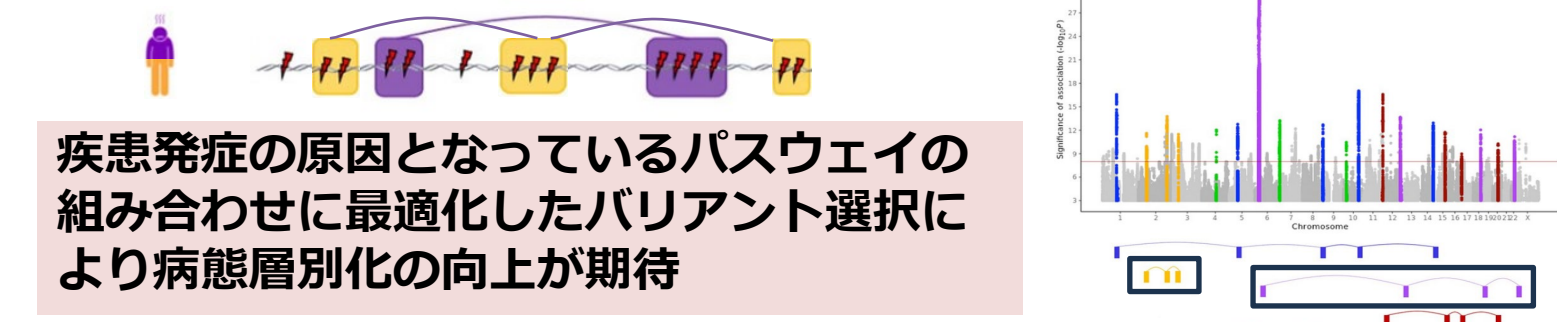
背景

糖尿病をはじめとする複雑かつ多様な疾患では、リスク予測に最適なバリエーションセットをただ一つ選択することは容易ではない。そこで、Udlerらによって発見された疾患パスウェイ^[1]と、Ahlgqvistらによって提唱された5つの臨床的サブクラス^[2]に注目した。両者の研究から糖尿病は様々な疾患パスウェイが濃淡を持ちながら組み合わせたり、いずれかの臨床的サブクラスを規定することでより病態に最適化した予測モデルの構築に繋がると考えられる。

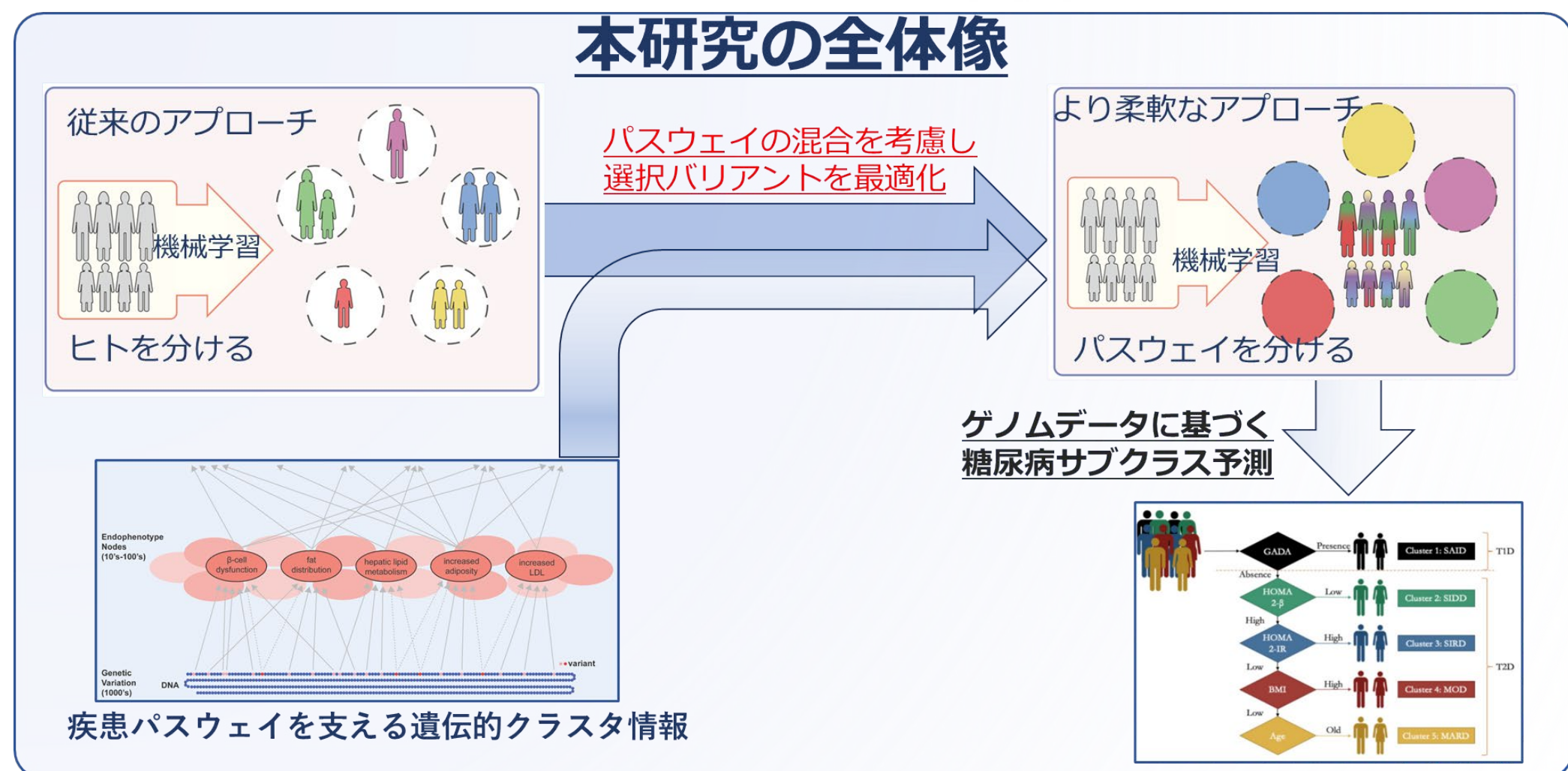
目的

パスウェイを支える遺伝的クラスター情報を考慮した機械学習により、5つの臨床的サブクラスを高精度に予測することを目的とする。

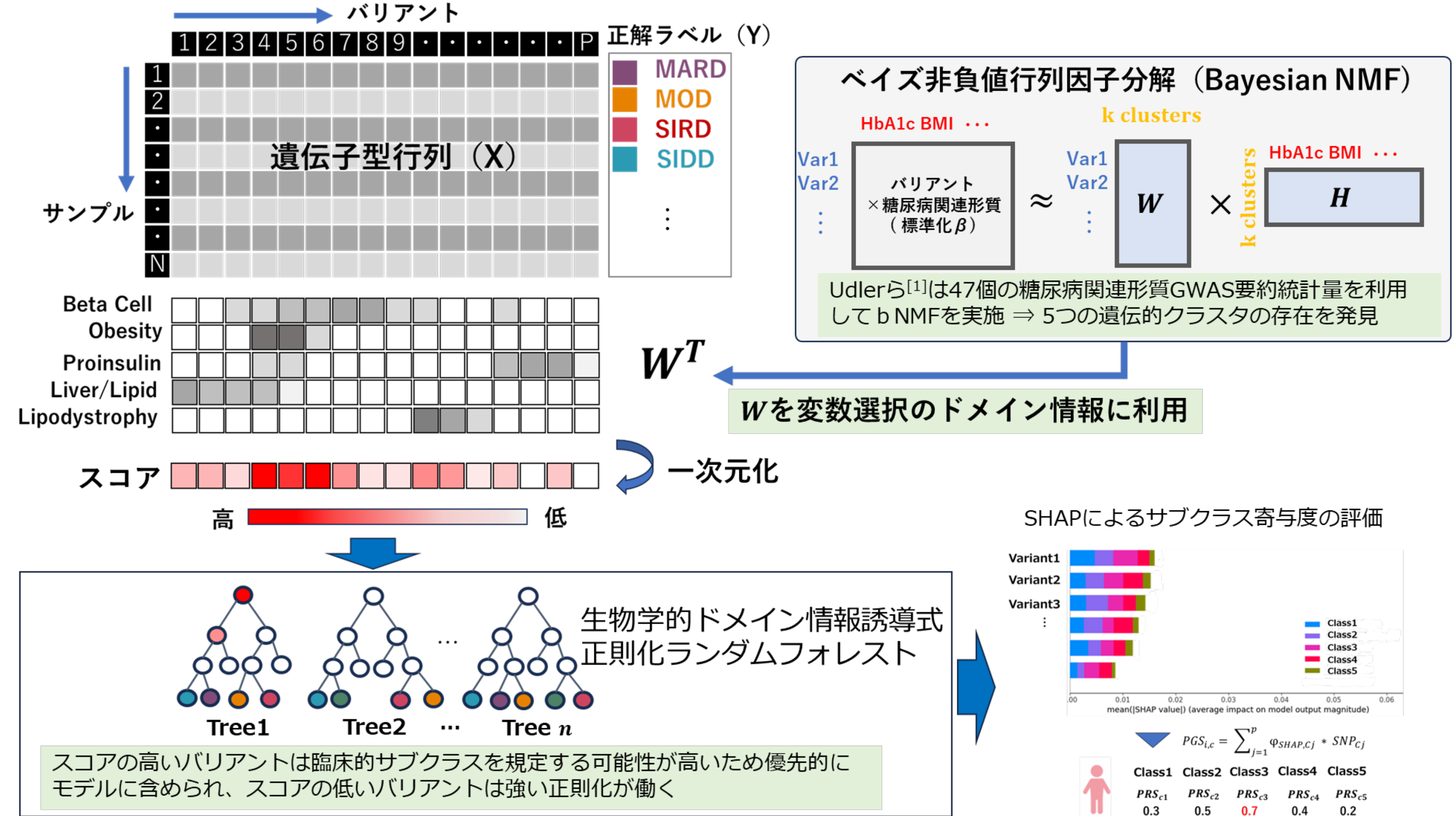
■ パスウェイの混合を考慮した変数選択の意義



本研究の全体像



解析方法



これまでの進捗

【2023年度の成果】
疾患パスウェイを支える遺伝的クラスターの混合を捉えるための変数選択法の考案および手法の確立を行った。東北メディカル・メガバンク計画 地域住民コホートの糖尿病患者の臨床的サブクラスのラベル付けを実施した。

参考文献

[1] Udler MS et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018
[2] Ahlgqvist E et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 2018
[3] Choi SW et al. PRSet: Pathway-based polygenic risk score analyses and software. PLoS Genet. 2023
[4] Udler MS. Type 2 Diabetes: Multiple Genes, Multiple Diseases. Curr Diab Rep. 2019
[5] Guan X et al. Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery. BMC Bioinformatics. 2020