FY2024/2024年度

Statistical Genetics Team
Gen Tamiya
遺伝統計学チーム　田宮 元

RIKEN Center for Advanced Intelligence Project

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を進めている。

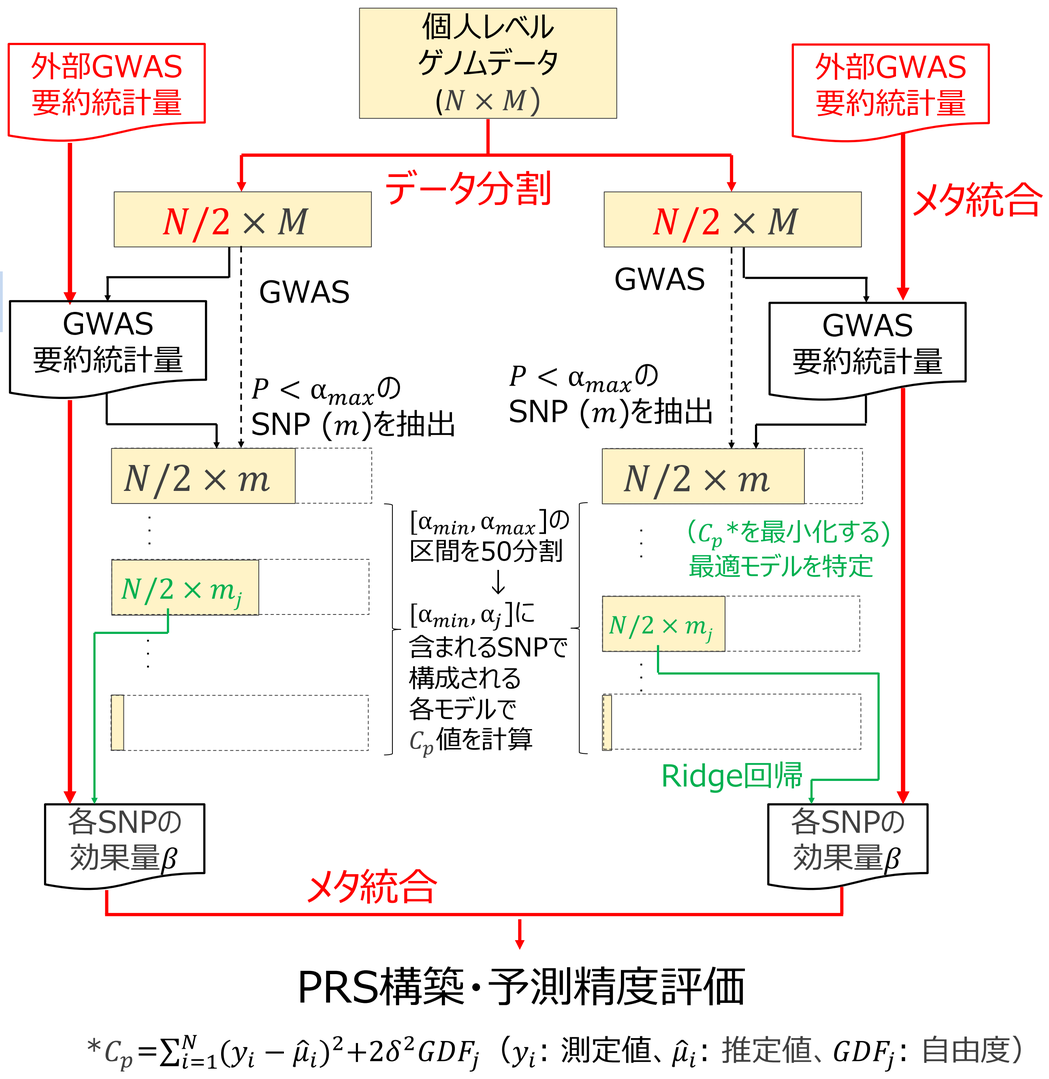
これまでの進捗

[2023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た解析により、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、分割計算により、予測精度を維持しながら計算負荷の軽減が可能であること、外部で公開されているGWAS（ゲノムワイド関連解析）の要約統計量を取り込むことで、予測精度が向上することを示した。

[2024年度の成果]
昨年度までの成果を踏まえ、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た。UKバイオバンクデータを用いた検証の結果、計算負荷の軽減と、予測精度の向上を同時に実現することができた。

提案手法

外部のGWAS要約統計量を利用したSTMGP法



*$C_p = \sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2 + 2\delta^2 GDF_j$（$y_i$: 測定値、$\hat{\mu}_i$: 推定値、$GDF_j$: 自由度）

実データへの適用

以下のデータに適用
形質: LDL-C
個人レベルデータ:
　UKバイオバンク[2]136,364人
外部GWAS要約統計量:
　GLGC（Global Lipids Genetics Consortium）国際メタGWAS[3]
　約165万人

メタ統合:
METAL[4]による固定効果モデル

データ分割	外部GWAS要約統計量	Adjusted R^2（統合後）	最大使用メモリ量（GB）
なし	なし	0.0927	189
	あり	0.0966	182
あり	なし	0.0838 0.0851 (0.0836)	64 64
	あり	0.0924 0.0910 (0.0962)	97 99

データ分割により、予測精度が一定程度低下したものの、使用メモリ量を大幅に削減できた。同時に、外部GWAS要約統計量の統合により、分割なしの場合と同様に、予測精度の向上を実現できた。

[1] Ueki M & Tamiya G. Genet Epidemiol 40:233-243, 2016.
[2] Bycroft C et al. Nature 562:203-209, 2018.
[3] Graham SE et al. Nature 600:675-679, 2021.
[4] Willer CJ et al. Bioinformatics 26:2190-2191, 2010.

今後の予定

関連多型の数や効果、遺伝率など、遺伝的アーキテクチャの異なる他の疾患/形質に適用し、予測精度を評価する。
また、構築された予測モデルを祖先性の異なる集団に適用し、その汎用性について検証する。

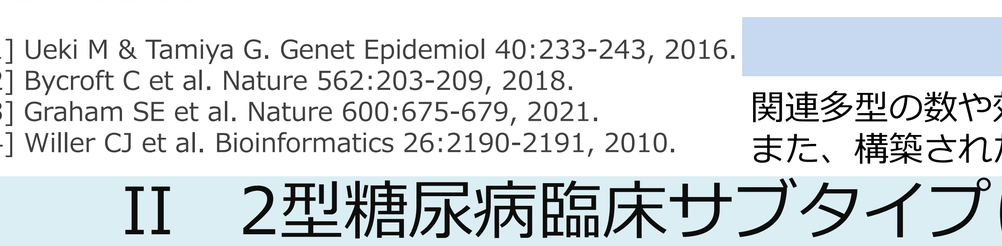
Ⅱ　2型糖尿病臨床サブタイプに対するポリジェニックリスク予測手法の開発

背景・目的

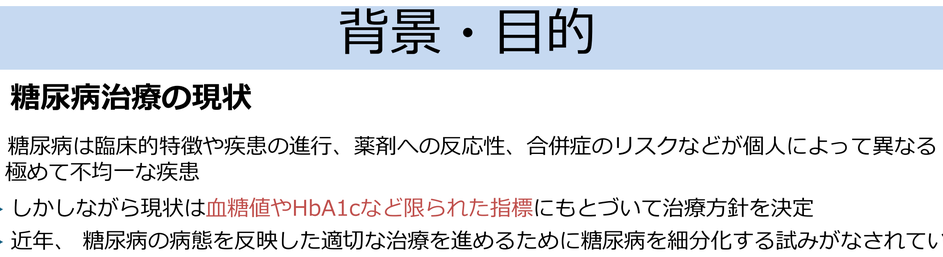
■ **糖尿病治療の現状**

- 糖尿病は臨床的特徴や疾患の進行、薬剤への反応性、合併症のリスクなどが個人によって異なる極めて不均一な疾患
  - ▶ しかしながら現状は血糖値やHbA1cなど限られた指標にもとづいて治療方針を決定
  - ▶ 近年、 糖尿病の病態を反映した適切な治療を進めるために糖尿病を細分化する試みがなされている

■ **K-meansクラスタリングによる糖尿病サブタイプ分類** [1][2]



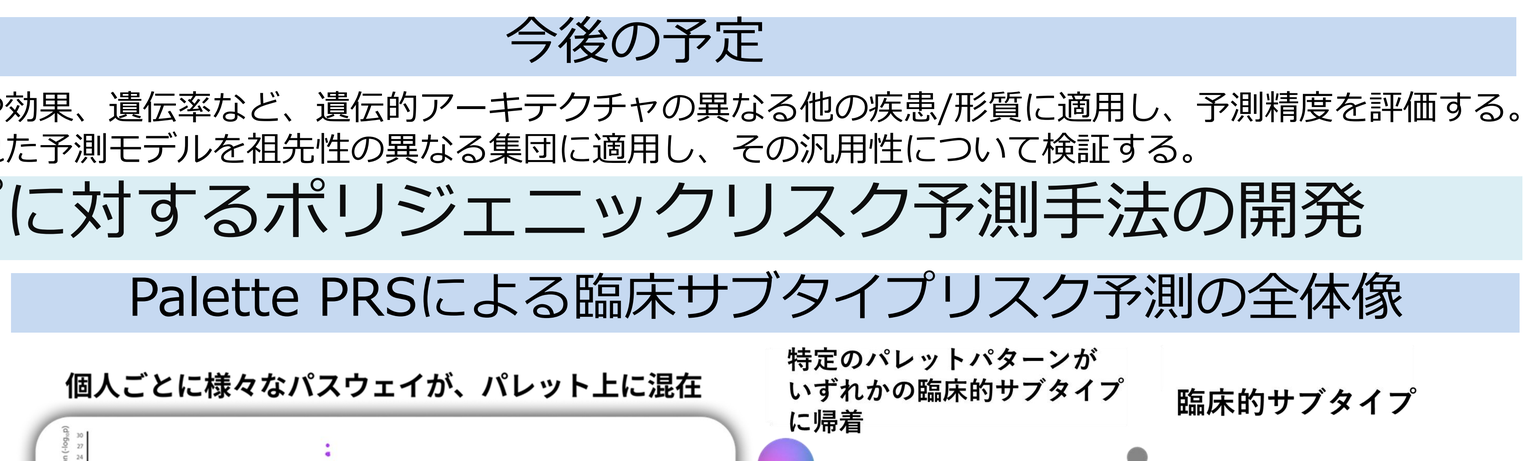
■ **サブタイプごとの糖尿病合併症発生率 (日本)** [1]



▶ 各臨床サブタイプの遺伝的リスクを予測するポリジェニックリスクスコア（polygenic risk score; PRS）を開発し、サブタイプに応じた予防的介入の実現を目指す

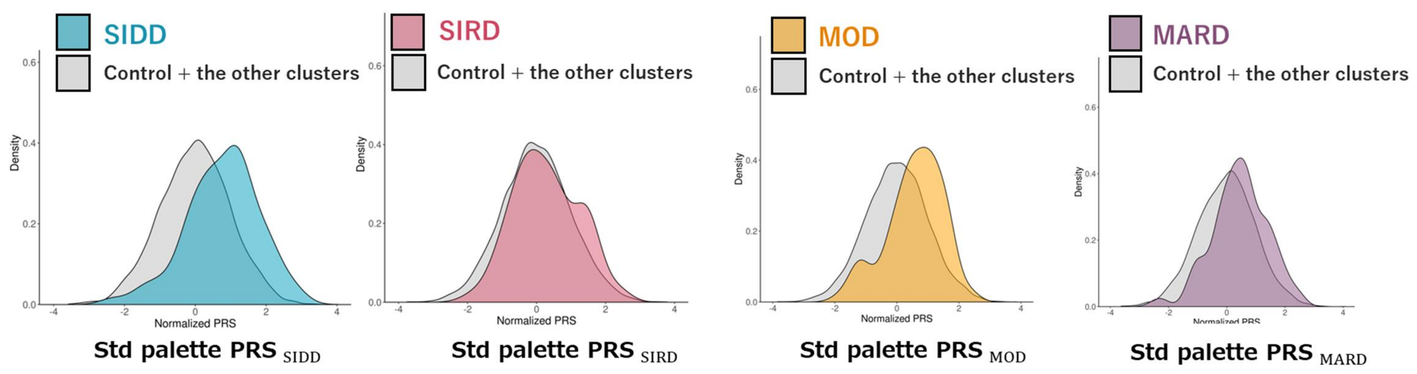
▶ベイジアン非負値行列因子分解（bNMF）により発見された12種類の糖尿病関連パスウェイ[3]を基に、各臨床サブタイプの病態をモデル化するPalette PRSを開発

Palette PRSによる臨床サブタイプリスク予測の全体像



実データへの適用

・**学習用データ**　東北メディカル・メガバンク計画 (T2D Case 3,800例 Control 30,000例)
　国立病院機構京都医療センター（T2D Case 350例）
・**検証用データ**　福島県立医科大学（T2D Case 430例）



重症サブタイプ（特にSIDD）の遺伝的高リスク群を高い精度で検出することに成功

[1] Tanabe H, et al. J Clin Med. 9:7, 2020.
[2] Ahlqvist E, et al. Lancet Diabetes Endocrinol. 6:361–369, 2018.
[3] Smith K, et al. Nat Med. 30:1065–1074, 2024.