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。

1．大規模ゲノムコホートデータを用いた疾患リスク予測手法の検証

背景

主な事例	対象形質/疾患	人数	SNP数
Ueki & Tamiya (2016)[1]	アルツハイマー病	713	8,657,877
Takahashi et. al. (2020)[2]	抑うつ症状 CES-D値	4,974	615,386
今回	糖尿病 HbA1c、尿酸値	**53,365**	**11,657,214**

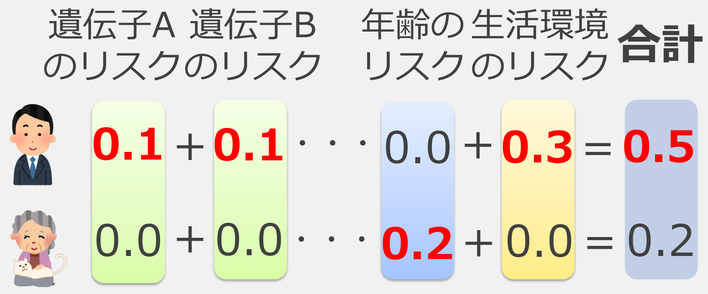
近年のゲノムデータを活用した論文により、疾患リスク予測の有用性が注目を集めている。我々研究チームが開発したSTMGP法は、アルツハイマー病や抑うつ症状で、高い予測精度を示した。我が国での疾患リスク予測の実用化に向けて、より多くの疾患を対象に大規模なデータでの検証が求められる。

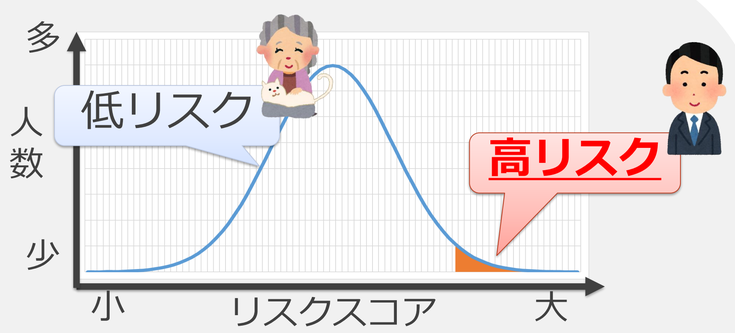
これまでの進捗

ToMMoの大規模データを解析し、疾患リスク予測手法の精度を検証した。規模が大きくなるほど、STMGP法が他手法よりも精度が高く、分割計算してもその精度を落とさずに使用メモリを半減できることを確認した。今後は検証範囲を広げ、10万人規模の解析や他疾患/形質への適用を目論む。

解析手法

疾患リスク予測

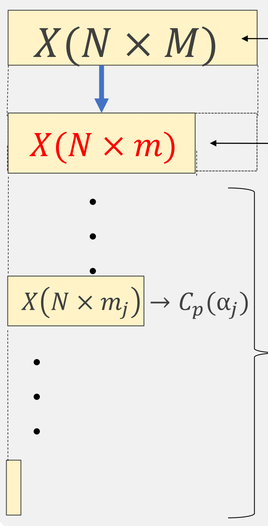




STMGP法

被験者数N
SNP数M

$\alpha_{min}, \alpha_{max}$：$P$値の閾値



1) GWASの実施
→ $p < \alpha_{max}$のSNP(m)を抽出

2) $[\alpha_{min}, \alpha_{max}]$の区間を50分割し各区間に含まれるSNPを特定
1区間を1モデルとする

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

3) 各モデルでC_p値を計算
最適モデル(C_pを最小化)を特定

4) 最適モデルにおける各SNPの効果量(β)の推定

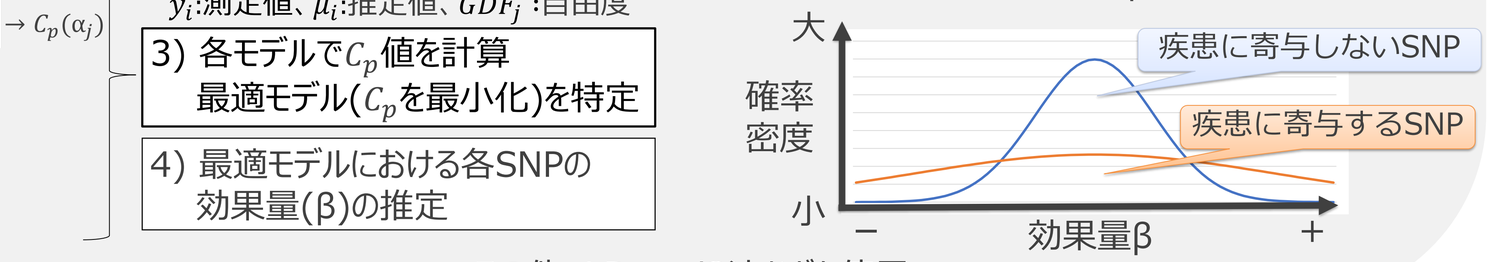
C+T法

P値とr2値でSNPを剪定

SNP	P値	r2値	
A	8.89E-08	0.6	
B	5.34E-08	0.8	
C	1.25E-08	0.2	採用

PRS-cs法

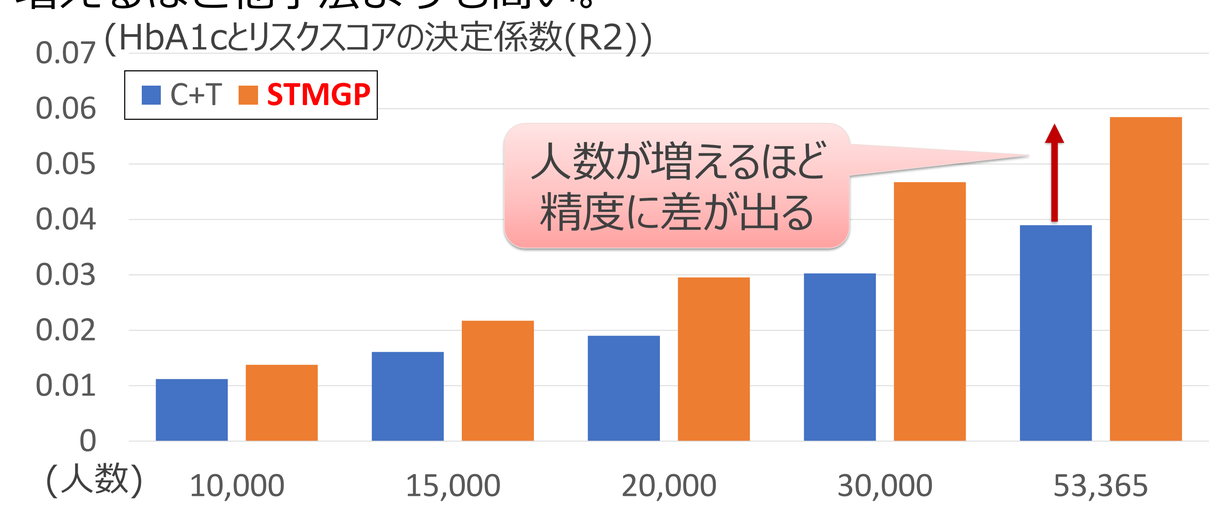
各SNPの疾患への効果量βの事前分布を仮定



その他、LDpred2法なども使用。

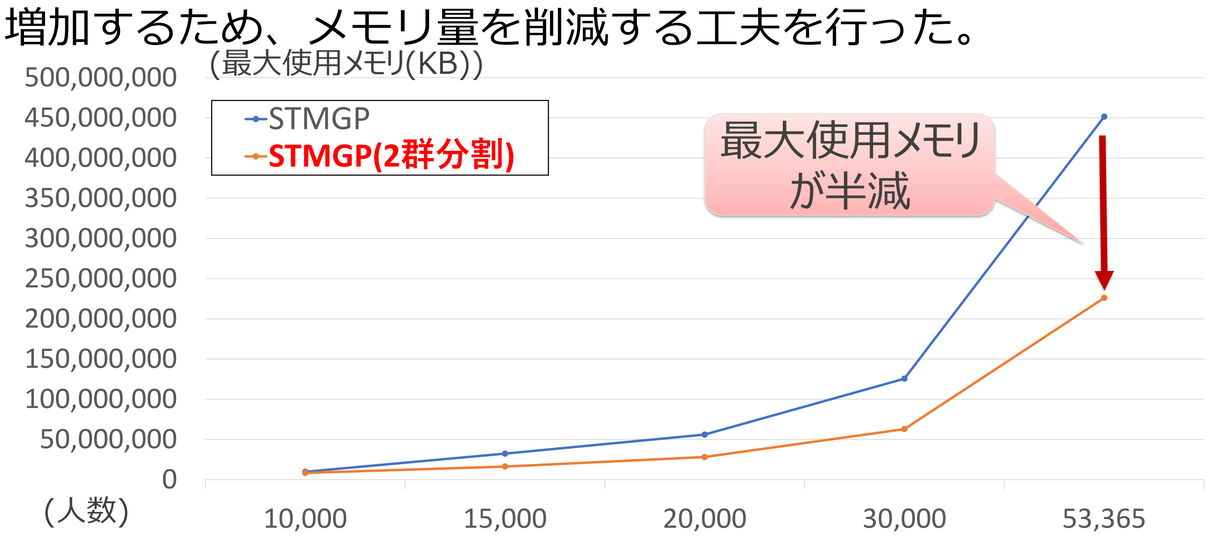
結果I.　ToMMoの大規模データ解析

STMGPが大規模データ解析に対応。その精度は、対象人数が増えるほど他手法よりも高い。



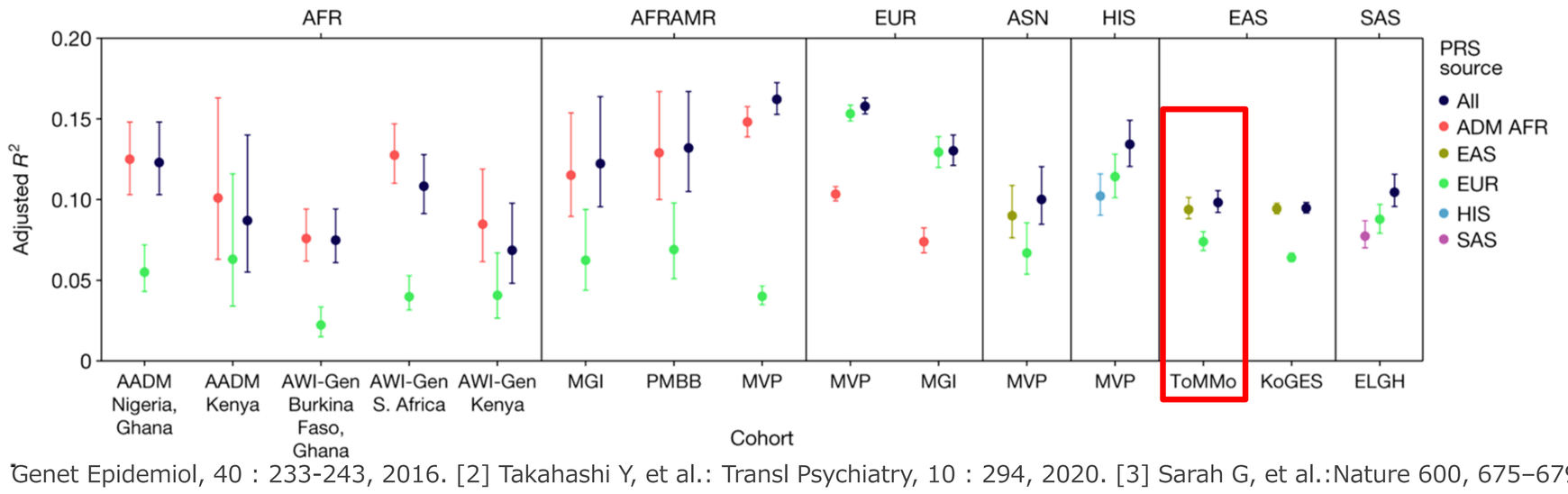
結果II.　STMGP法の分割計算の検証

対象人数が多くなると計算量が増え、使用メモリ量が指数的に増加するため、メモリ量を削減する工夫を行った。



2．多民族集団を対象とした大規模ゲノムワイド関連メタ解析と形質予測精度の検証

Global Lipids Genetics Consortium（GLGC）は様々な民族集団で構成される201コホート(165.5万人)を対象に、脂質関連5項目のゲノムワイド関連メタ解析を実施した[3]。我々の研究グループでは、メタ解析から得られたSNPと効果量βを用いて、地域住民コホート調査（宮城）の参加者28,000人のリスクスコアを算出し、LDL-Cの分布をどの程度説明できるかを決定係数で評価した。本研究の結果、各祖先グループごとよりも、集団全体のデータから特定されたSNPとその効果量βをもとに算出されたリスクスコアの方が、予測性能が高いことが示された。これまでの解析はヨーロッパ人集団を対象としたものが多いが、様々な民族集団を解析対象に含めることで、より信頼度の高いSNPの探索や疾患リスク予測が可能となると期待される。



[1] Ueki M & Tamiya G : Genet Epidemiol, 40 : 233-243, 2016. [2] Takahashi Y, et al.: Transl Psychiatry, 10 : 294, 2020. [3] Sarah G, et al.:Nature 600, 675–679, 2021.