博士論文

機械学習を用いたゲノムワイド遺伝子多型情報に基づくうつ状態のリスク予測

東北大学大学院医学系研究科医科学専攻

神経・感覚器病態学講座精神神経学分野

高橋　雄太

目次

Ⅰ　要約

研究背景

ゲノムワイド遺伝子多型データを用いて、うつ状態の脆弱性を予測したこれまでの研究では、あまり高い予測精度が得られていない。その主な原因としては、一塩基多型の表現型に対する効果サイズが小さく、十分な検出力を得るのが難しいことがある。そして、表現型に効果をもたない null variants が多く予測モデルに含まれることにより、学習の段階では見かけ上、高い予測精度が得られるのにもかかわらず、テストデータで検証すると低い予測精度しかえられない、過剰適合という現象が起きてしまうためである。STMGP （Smooth-Threshold Multivariate Genetic Prediction）法は過剰適合を軽減させることで予測精度を向上させるために開発された、機械学習を用いたゲノムワイド遺伝子多型データに基づく予測モデルである。

研究目的

ゲノムワイド遺伝子多型データからのうつ状態脆弱性の予測に STMGP 法を用いることで、従来法と比較して予測精度が向上するかどうかを検討する。

研究方法

東北メディカル・メガバンク計画によって、宮城県でリクルートされた 3,685 人のゲノム情報を用いて予測モデルを学習させ、岩手県でリクルートされた 3,048 人のゲノム情報を用いて、予測モデルの予測精度を評価した。両方のコホートは東北メディカル・メガバンクプロジェクトで収集されたもので、ゲノタイピングは

HumanOmniExpressExome BeadChip Array を用いて行った。うつ症状は Center for Epidemiologic Studies-Depression Scale (CES-D) で評価した。STMGP 法による予測精度と過剰適合の程度は、遺伝子スコア法、GBLUP (genomic best linear unbiased prediction) 法, SBLUP (summary-data-based best linear unbiased prediction) 法, BayesR 法, Ridge 回帰法と比較した。

研究結果

STMGP 法による予測精度 (predictive correlation coefficients ± 標準誤差) は 0.0769 ± 0.0173 であり、遺伝子スコア法 (0.0332 ± 0.0178)、GBLUP 法 (0.0309 ± 0.0178)、SBLUP 法 (0.0164 ± 0.0178)、BayesR 法 (0.0100 ± 0.0185)、Ridge 回帰法 (0.0260 ± 0.0178) よりも高かった。また STMGP 法ではトレーニングデータでの見かけ上の予測精度は 0.3232 ± 0.0153 であり、遺伝子スコア法 (0.9027 ± 0.0076)、GBLUP 法 (0.9627 ± 0.0017)、SBLUP 法 (0.9554 ± 0.0019)、BayesR 法 (0.9633 ± 0.0015)、Ridge 回帰法 (0.9998 ± 0.000) よりも低く、過剰適合の程度が軽減されていた。

結論

STMGP 法は過剰適合を軽減することで、従来法よりもゲノムワイド遺伝子多型データからうつ症状を予測する際に、高い精度を示した。ポリジーンモデルに従う複雑な遺伝疾患の脆弱性の予測に STMGP が有効であることが示唆された。

Ⅱ　研究背景

うつ病は生涯有病率 8-12%とされ、その多くが就労年齢である若年で発症し、社会的な機能の低下、高い自殺率が問題となっている(1)。患者だけでなく、家庭、企業、国家の経済的損失も大きい疾患である(2)。うつ病の生物学的な発症機序は、現時点では不明とされているが、遺伝素因、環境素因、遺伝子環境相互作用が複雑にかかわる多因子疾患とされている(3)。現在、うつ病の診療では、DSM (Diagnostic and Statistical Manual of Mental Disorders) や ICD (International Statistical Classification of Diseases and Related Health Problems) の診断基準に従い、症状のみに基づいた診断が行われている。診断の一致率が限られていること、うつ病と診断された集団が異質であるために、診断と治療方針が一対一対応しないことが問題となっている(4)。これらの解決のために、症状のみではなく、客観的な生物学的指標を用いて、疾病の診断や分類をすることで、診断と治療方針が対応するような個別化医療の開発が期待されている(4)。生物学的指標としては、ゲノム、血中や髄液中の生化学的な指標、脳画像、生理学的な指標が研究されている(5, 6)。この中で、ゲノム情報は遺伝素因の機序を直接的に調べられ、遺伝子多型は検査装置や概日リズム、生活習慣、服薬状況などの環境因からの影響を受けないという長所がある(7)。今回、我々はうつ病やその状態像（うつ状態）への遺伝要因を探求し、その貢献度を推定することを目指し、遺伝子多型情報からうつ状態を予測する研究を行った。ゲノムからうつ状態を予測するスコアを作成することができれば、個別の患者に対して遺伝的要

因と環境的要因の関連の強さの割合を得ることができるので、想定される病態生理をもとに患者をサブタイプ分類することが可能となる。こうして、これまでは試行錯誤の末に決定されていた治療方針についても、サブタイプに基づいて決定できることが期待される。

双生児を対象としたメタアナリシスではうつ病の易罹病性に対する相加的な遺伝分散は37%と計算されている(8)。最近の大規模なGWAS（genome-wide association analysis；全ゲノム関連解析）では、数十～百前後の、うつ病と統計的に有意な関連を示すSNP（single nucleotide polymorphism；一塩基多型）が報告されている(8-10)。これらの研究の結果、うつ病へのそれぞれのSNPの効果サイズは、オッズ比で1.05程度と小さく、多くの小さい効果の遺伝子多型が発症にかかわる、ポリジーンモデルが想定されている(10)。したがって、単一のSNPではなく、複数のSNPの効果を効率的に組み合わせることが、うつ病の脆弱性を予測するためには必要である。しかし、ポリジーンモデルにおいて、小さいSNP効果を効果的に足し合わせるには、下記に示すような統計学的な問題がある。

ポリジーンモデルの表現型を扱う際に、GWASの結果に基づいて、精度の高い予測モデルを作成することは、容易ではない。なぜならば、真に（弱い）関連を示すSNPを一度のGWASに基づいて選択しようとする際に、現在入手できるサンプルサイズでは限られた検出力しか得られないためである。このことの概念図を図１Aに示した。例えば、先行報告(10)に基づいて、真にうつ症状に対して効果のあるSNPのオッズ比

が 1.05 だと設定し、その SNP の MAF（minor allele frequency；マイナーアレル頻度）が 50%で、症例 10,000 人と対照群 10,000 人の症例対象研究で有病率が 10%の疾患について検討することをシミュレーションすると、真に関連がある SNP の P 値が 1×10^{-7} 未満となる確率は 0.001%である。この確率は P 値の閾値を 0.5 に上げると 67%に上昇する(7)。この限られた検出力がゲノムからの予測モデルに過剰適合を引き起こす。過剰適合とは、トレーニングデータで予測モデルを作成する段階で算出される見かけ上の予測精度が高いのにも関わらず、独立したテストデータに基づいて計算された予測精度が低くなってしまう事象である。過剰適合が起きる主要な原因は対象とする表現型に対して全く効果をもたない SNP が予測モデルに混入されるためである。このような SNP を、この論文の中では、先行報告(11)にならって null variants と呼ぶこととする。

ポリジーンモデルに従う複雑な疾患をゲノムから予測する際に、頻繁に用いられている方法に遺伝子スコア法 (Polygenic Risk Score) と GFLUP (genomic best linear unbiased prediction) 法がある。遺伝子スコア法は、Purcell ら(12)により提案されたもので、まずトレーニングデータで GWAS を実行し、任意の P 値のカットオフを設定し、P 値がそのカットオフよりも小さい SNP の回帰係数を足し合わせるものである（図 1 B)。遺伝子スコア法はポリジーンモデルの予測に頻回に使用されるが、40 万人以上のサンプルを用いた解析でもうつ病の予測精度（R^2 値）は 0.02 と予測精度は限られている(10)。うつ病のゲノム研究の領域では、遺伝子スコア法による予測精度の

低さは主に検出力が限られていることに起因している(8, 10, 13)。他には、遺伝子スコア法ではClampingという操作により互いに独立なSNPのデータセットしか用いないので、連鎖不平衡により互いに相関関係をもつSNP群を効果的に予測モデルに含めることができないという問題がある。検出力不足の問題は、より多くの被験者数を集めることでしか解決されないが、SNP同士の相関関係を効果的に予測に利用することについては、重回帰モデルを作ることで解決される可能性がある。また、null variantsを過剰評価して過剰適合を引き起こしてしまうことについては、罰則付き回帰モデルを作成することで軽減される可能性がある。

ポリジーンモデルの予測に頻用されている、もう一つのモデルはGBLUP法であり、これはSNPの効果をランダム効果とみなして、線形混合モデルを作成するものである（図１C)。遺伝子スコア法と異なり、GBLUP法は互いに相関するSNP群を効果的に予測モデルに含めることができる。しかし、GBLUP法でも、うつ病またはうつ状態の予測についてはPCC（predictive correlation coefficient）が0.045と予測精度は限られている(14)。GBLUP法の限界の一つとして、選択せずに全てのSNPを予測モデルに含めるために、多くのnull variantsが予測モデルに含まれ、結果的に過剰適合を招き、遺伝子スコア法で観察されたように、予測精度を下げてしまうという問題がある。

過剰適合や連鎖不平衡関係のSNP情報の損失といった問題を解決するために、STMGP（smooth-threshold multivariate genetic prediction）法が植木と田宮によ

り提案された(15)（図１D)。遺伝子スコア法と同様に、STMGP 法は GWAS での P 値に閾値を設定することにより SNP を選択し､それをもとに予測モデルを作成する｡しかし、単純に SNP の回帰係数を足し合わせる遺伝子スコア法と異なり、STMGP 法は、トレーニングデータでの GWAS の統計量で重みづけをしたうえで、選択された SNP 群で罰則付き回帰（一般化 ridge 回帰）モデルを作成する。このことで、(i) null variants の過大評価による過剰適合を軽減し、(ii) 相関関係のある SNP 群を効果的に予測精度向上に利用することができる。

STMGP 法は罰則付き回帰の機械学習である Elastic net と類似したモデルである。Elastic net や Lasso などの罰則付き回帰モデルは、他の手法と比較して SNP に基づく予測モデルにおいて、高い予測精度が報告されている (14, 15)。しかし、これらの手法は､チューニングパラメータを設定するために､交差検証法を使用しているので、莫大なコンピューターコストがかかってしまい、大規模なゲノムワイドデータに実装できないという問題が指摘されている (16)。実際に、今回解析に使用した東北大学東北メディカル・メガバンク機構のスーパーコンピュータ環境でも、1000 人程度の集団の解析を行うにも、1 週間経過しても解析が終了せず、実現が難しいことを確認している。一方、STMGP 法は交差検証法を用いずに、Unbiased Cp-type model selection criterion を用いて、チューニングパラメータを決定するので、コンピューターコストを軽減させることができる。

植木と田宮は、シミュレーションデータを用いた解析と、公開されている 713 人

のアルツハイマー病と健常人の全ゲノムシークエンスデータによって、STMGP 法が遺伝子スコア法や GBLUP 法よりも優れた予測精度を示すことを報告している(15)。植木と田宮の研究とは異なり、今回の研究では、STMGP 法を、より予測が困難な条件設定でその有効性を試している。第一に、今回の対象となる表現型のうつ状態は、アルツハイマー病と比較して、個別の SNP の効果サイズが小さい。第二に、今回の研究では、予測精度の評価を、トレーニングデータとは別にリクルートされた独立したテストデータセットで評価している。植木と田宮は同一データセット内の交差検証法で予測精度を評価したが、この方法では、我々の手法と比較して、予測精度が過大に評価されるリスクがある。

植木と田宮の報告(15)において、STGP 法が様々な分布の表現型のシミュレーションデータにおいて、遺伝子スコア法や GBLUP 法と比較して高い予測精度を示したことに基づき、我々は、STMGP 法は、実際の被験者のうつ症状をゲノムワイド SNP データから予測する際にも、従来の手法と比較して高い予測精度を示すことができると仮説を立てた。今回の研究では、ゲノムワイド SNP データからうつ状態への脆弱性を予測するために、3,685 人の宮城県で収集されたデータをもとに予測モデルを作成し、3,048 人のテストデータで予測モデルの精度を検証した。各予測モデルについては、予測精度と過剰適合の程度により評価した。

比較する対象の予測モデルとしては、現在頻用される遺伝子スコア法と GBLUP 法だけでなく、より公平に STMGP 法の精度を比較するために、3 モデルを用意した。一

つ目は、GBLUP 法の亜型であり、GWAS 統計量をもとに BLUP を行う SBLUP（summary-data-based best linear unbiased prediction）(16)法、2 つ目は、最近開発され、ベイズ階層モデルを作成することで、過剰適合を軽減させることに一定の効果があると報告されている BayesR 法(17)、3 つ目は、STMGP 法の一般化 Ridge 回帰のような GWAS 統計量による重みづけを行っていない、古典的な Ridge 回帰法である。

Ⅲ　研究目的

前述の通り、STMGP 法は、過剰適合を軽減することでポリジーンモデルに従う表現型の、ゲノムからの予測精度を向上させることを目的に開発された。今回は、ゲノムワイド SNP データからうつ症状を予測する際に、STMGP 法の予測精度と過剰適合の程度を、現時点で state-of-the-art と評されている各手法（遺伝子スコア法、GBLUP 法、SBLUP 法、BayesR 法、Ridge 回帰）と比較し、STMGP 法の性能が既存手法よりも優れていることを確認することを目的とする。

Ⅳ　研究方法

1　研究集団

東北メディカル･メガバンク計画の地域住民コホート調査(18)では 8 万人以上の、岩手県や宮城県在住の成人が、2013 年 3 月から 2016 年 3 月までの間にリクルートされた。東北メディカル・メガバンク計画のデザインやリクルート方法については、先

行報告に記載がある(18)。簡潔には、被験者は、20 歳から 75 歳までの成人で、社会疫学的因子・生活習慣・既往歴などの幅広い質問紙項目に回答している。血液検査はベースライン調査の際に実行され、血液サンプルはバイオバンクに保存されている。岩手県、宮城県に在住の被験者は、岩手医科大学、東北大学がそれぞれ別個にリクルートしている。この事業は岩手医科大学、東北大学の両方の倫理委員会から承認を得ており、今回のうつ症状の予測研究については東北大学の倫理委員会の承認（2018-1-851）を得ている。この研究はヘルシンキ宣言に則って行われており、研究参加時に、すべての参加者が、書面でインフォームド・コンセントを提出している。

２　ゲノタイピング

2013 年に登録された 9,966 人の被験者（20）について、Human OmniExpressExome BeadChip Array （Illumina Inc., San Diego, CA, USA）を用いて得られたシグナル強度の情報（.idat 形式）を Illumina 社の GenomeStudio ソフトウェアプログラムで分析し、アリルコールを行った。出力ファイルを PLINK ソフトウェアの BED 形式に変換し、以下の作業を行った。Human OmniExpressExome BeadChip Array は人類集団中で MAF が 5%以上の高頻度のタグ SNP についてはほぼ網羅されており、もっとも性能が高いタグ SNP セットの一つとなっている(19)。ただし、SNP セットを作成する際に参照したデータセットである 1000 Genomes Project と Exome Consortia においては、アジア人がそれぞれ 17%（186/1092)、2.7%（327/12031）しか含まれないため、日本

人サンプル集団では、搭載されているSNPの一部は分離されないとの報告がある(19)。東北大学によりリクルートされた宮城県在住の 4,974 人をトレーニングデータセット、岩手医科大学によりリクルートされた岩手県在住の 4,992 人をテストデータセットとした。Call rate が低い検体（0.98 未満；トレーニングデータで 2 人、テストデータで 3 人）は除外した。検体間の血縁関係について、PLINK ソフトウェアの PI_HAT 値が 0.09375 を上回るという基準で調べたところ、2,156 ペア（トレーニングデータで 620 ペア、テストデータで 1,536 ペア）が検出された。これらのペアはトレーニングデータ内、テストデータ内、トレーニングデータとテストデータ間のペアを含んでいる。それから、それぞれのペアにおいて、Call rate の低いほうの被検者を解析から除外した。さらに、Call rate が低い（0.99 未満）バリアント、Hardy-Weinberg equilibrium exact 検定で P 値が低い（1×10^{-4}未満）バリアント、そして、MAF が低い(0.01 未満)バリアントは除外した。表現型や共変量の情報が欠損している検体(トレーニングデータで 669 人、テストデータで 408 人）を除外した。今回のバリアントの品質管理操作において除外されたバリアントのうち 98%以上は、MAF ＞0.01 のフィルターで除外されており、日本人集団においては Human OmniExpressExome BeadChip アレイで分離されないバリアントが一定数含まれていたことによると考えらえる。最終的に、トレーニングデータの 3,685 人と、テストデータの 3,048 人で 615,386 バリアントが予測モデル作成の解析に使用された。我々は、結果の一貫性を調べるための追加の解析として、インピュテーションデータを用いた解析も行った。

３　インピュテーション

結果の一貫性の確認のために、インピュテーションされたSNPデータを用いた予測モデルの作成と予測精度の評価も行った。インピュテーション前の前処理として、HumanOmniExpressExome BeadChip Arrayでゲノタイピングされた9,966人のデータで、Call rateが低い（0.98未満）検体を除外し、重複するバリアント、完全に欠損しているバリアント、Hardy-weinberg equilibrium exact testのP値が低い（0.05未満）バリアント、MAFが低いバリアント（0.05未満）を除外した。その後、9,961人の490,981バリアントはSHAPEIT2（v2. r837）(20)を用いて、--duohmm -W 5 -thread 16 オプションを設定してフェージングした。インピュテーションは、フェージングされたゲノタイプに対して、IMPUTE2（version 2.3.2）(21)を用いて、2つのフェージングされたリファレンスパネルを用いて行った。一つは東北メディカル・メガバンク機構で2,049人の日本人ゲノムデータを用いて作成されたもの(22)で、もう一つは1000人ゲノムプロジェクトPhase 3のデータセットのうちEast Asiansのものである(23)。IMPUTE2では、-use_prephased_g, -Ne20000, -align_by_maf_g,-k_hap 4000のオプションを使用した。Oxford GEN formatの、インピュテーションされたゲノタイプデータはPLINK BEDフォーマットに、最も事後確率の高いゲノタイプを選択することで変換された。変換の際に、INFO値が0.9未満のバリアントは除外した。インピュテーションしていないゲノタイプデータセットと同様にPH_HAT値が0.09375を上

回る、2,156 の血縁者のペアについては、Call rate が低い検体を除外した。インピュテーションされたゲノタイプデータに含まれる 11,030,858 バリアントのうち、Call rate が低い（0.99 未満）バリアント、Hardy-Weinberg equilibrium exact test で P 値が低い（$1x10^{-4}$ 未満）バリアント、MAF が低い（0.01 未満）バリアントは除外し、5,949,462 バリアント（トレーニングデータセット 3,685 人、テストデータセット 3,048 人）を、インピュテーションデータを用いた解析に使用した。

４　アウトカム指標

うつ症状は Center for Epidemiological Studies-Depression scale （CES-D）(24)によって評価した。CES-D は各項目 0,1,2,3 の 4-point 尺度により計測され、20 項目の質問から構成される自記式の質問紙票である。最近 7 日間の、感情や行動に関する頻度についての質問が含まれ、スコアが高いほどうつ状態が重篤であることを示唆する。トレーニングデータセットとテストデータセットで、それぞれ 2.0%と 0.7%の欠損値を認め、欠損値のある被験者（トレーニングデータ 618 人、テストデータ 408 人）は解析から除外された。

今回使用したデータセットにおける、CES-D の分布を図２に示す。先行報告(25)と同じく、CES-D スコアの分布は正規分布ではなく、高スコア方向へ裾が伸びた分布となっている。我々は、今回の CES-D の分布が予測精度に与える影響を評価するために、CES-D の分布に編集を行った追加解析を行い、編集しない CES-D を予測した主解

析との結果の一貫性を確認した。まず、第一の追加解析では、CES-D の分布が正規分布とは異なっていることの影響を評価した。我々は Box-Cox 変換により CES-D を正規分布に近づけた表現型を作成し、それを予測した精度を調べた。Box-Cox 変換後の CES-D スコアは図２に示されており、正規分布を示している。第二の追加解析では、CES-D の外れ値の予測精度への影響を評価した。外れ値は boxplots adjusted for skewed distribution(26)を用いて検出し、0 点、または 33 点より大きい CES-D（トレーニングデータセットとテストデータセットでそれぞれ 3.5%と 4.7%）が外れ値と判定された。これらの外れ値は、この追加解析では除外した。

５　予測精度の評価指標

PCC を、先行研究に従って(14, 17)、予測精度の評価指標として採用した。PCC は予測値と真の値との間の相関係数であり、ゲノムデータからの表現型の予測では、頻回に使用される予測指標である(14, 15, 27, 28)。STMGP の PCC と他の予測モデルの PCC の差について、統計学的な有意差の検定をするために、我々は William’s test(29)を用いた。William’s test は二つの相関係数において一方の変数を共有している場合に、それらの相関係数の違いを検定するものであり、R の pysch package に実装されている。

６　STMGP のパッケージとパラメータ

STMGP のソースコードは、CRAN（R の公式パッケージアーカイブ）(30)で入手可能な STMGP v1.0 を使用し、パッケージ内の stmgplink 関数を用いて予測モデルの作成を行った。stmgplink 関数の入力として、SNP データ、表現型データ、共変量データを、トレーニングデータとテストデータのそれぞれについて用意し、τ と γ のパラメータを設定した。stmgplink 関数は連続的に、(A) GWAS の P 値を計算し、(B) 最適な P 値カットオフを Mallow's Cp criterion に基づいて決定し、(C) STMGP モデルに基づいて予測モデルを作成し（具体的には、GWAS の統計量と SNP 間の相関に基づいて、重みづけされた一般化 Ridge 回帰に基づいて、SNP の回帰係数を決定する)、(D) テストデータセットの表現型を予測する(15)。

P 値のカットオフに加えて、STMGP 法には τ と γ のチューニングパラメータが存在する。τ は全体の罰則の程度を調整する。具体的には、損失関数$\|y - X_A\beta_A\|^2$を制御する。ここで、y、X_A、β_Aはそれぞれ表現型のベクトル、説明変数の行列（選択された遺伝子多型のアリル数を標準化したもの)、回帰係数のベクトルを表している。A は、設定された P 値のカットオフで選択された遺伝子多型群を表している。したがって、残差二乗和である損失関数は被験者数（N）に応じて増大するために、τ は N に依存して調整される必要がある。植木と田宮による先行報告(15)では、シミュレーション研究や実際のゲノムデータを用いた研究に基づき、τ を$N/\sqrt{\log\ (N)}$と設定することを推奨している。今回の研究では、τ としては、推奨されている$N/\sqrt{\log\ (N)}$を設定した主解析を行うとともに、結果の一貫性の確認のために、

$N/0.1$、 $N/1$、 $N/10$に設定した解析も行った。

γ は一般化 Ridge 回帰において、GWAS 統計量の入力の、遺伝子多型の重みづけへの影響の程度を制御している。遺伝子スコア法の P 値の閾値設定は、悉無律に従うハードなものであるのに対して、STMGP 法の P 値の閾値設定は連続的でスムースなものである。このパラメータは、先行報告(31)において、adaptive lasso(32)の γ と等価であることが示されており、adaptive lasso でよく設定されるように(33-36)、γ は１と設定した。

７　STMGP 法における共変量の調整方法

性別や年齢、それに、集団階層性を調整するための有意な主成分得点は、先行報告に従い(15)、主解析では共変量として予測モデルに加えられた。Eigensoft パッケージ(37, 38)で、トレーニングデータとテストデータを合わせたデータに主成分分析を行う際に、Tracy-Widom 分布に基づいて、各主成分の P 値を算出し、P 値が 0.05 未満の主成分を有意な主成分として共変量に利用した。トレーニングデータとテストデータを合わせたデータで主成分分析を行ったのは、トレーニングデータとテストデータで同じ数の主成分を共変量としてモデルに含めるためである。主成分分析の第１主成分と第２主成分の散布図を図３に示した。トレーニングデータでは全体で一つの円形クラスターを示しているが、テストデータではクラスターの形がややいびつである。トレーニングデータとテストデータを合わせたデータでは外れ値を示す被験者はお

らず、第１主成分、第２主成分共に、６標準偏差以内に全被験者が存在する。共変量を含めた予測モデルは、下記のようにトレーニングデータで作成された。

$$y_{train} = b_0 + b_1 \times AGE_{train} + b_2 \times SEX_{train} + b_3 \times PC1_{train} + b_4 \times PC2_{train} + \cdots + \boldsymbol{SNP}_{train}\boldsymbol{b}_{SNP} + e_{train} \quad (1)$$

回帰係数（b_0，b_1，b_2，b_3，b_4，… $\mathbf{b}_{\mathbf{SNP}}$）は、STMGP 法によって得られるもので、これらの値が、テストデータにおいて、下記のような式を作って予測を行うのに利用される。

$$\mu_{test} = b_0 + b_1 \times AGE_{test} + b_2 \times SEX_{test} + b_3 \times PC1_{test} + b_4 \times PC2_{test} + \cdots + \boldsymbol{SNP}_{test}\boldsymbol{b}_{SNP} \quad (2)$$

上記の式で、y、AGE、SEXは各被験者の表現型スコア（CES-D）、年齢、性別をそれぞれ表す。PC1、PC2…. は調整されるべき主成分得点、e_{train}は誤差項である。$\boldsymbol{SNP}$は選択された SNP のアリル数の標準化されたベクトルを表す。

共変量の調整方法は、予測精度に影響をもたらすため、我々は共変量の扱い方を複数行い、結果の一貫性についての確認をした。上記（１）（２）の数式で表されるアプローチ以外に、2 つのアプローチを行った。第一のアプローチは、予測モデルを作成する前に、表現型スコアを共変量で調整するもので、予測モデルには共変量を加えないものである(14, 39)。この方法では、表現型は事前に下記のように共変量で調整される。

$$y_{train} = b_0 + b_1 \times AGE_{train} + b_2 \times SEX_{train} + b_3 \times PC1_{train} + b_4 \times PC2_{train} + \cdots + e_{train} \quad (3)$$

$$y_{test} = b'_0 + b'_1 \times AGE_{test} + b'_2 \times SEX_{test} + b'_3 \times PC1_{test} + b'_4 \times PC2_{test} + \cdots + e_{test} \quad (4)$$

そして、結果の残差をr_{train}と r_{test}と表現すると、r_{train}と r_{test}を CES-D の代わりに表現型として扱うこととする。そして、予測モデルは下記のように学習される。

$$r_{train} = \boldsymbol{SNP}_{train}\boldsymbol{b}_{SNP} + e'_{train} \quad (5)$$

ここでe'_{train}は誤差項を表し、テストデータでの予測スコアは

$$\hat{r}_{test} = \boldsymbol{SNP}_{test}\boldsymbol{b}_{SNP} \quad (6)$$

と表される。

第二のアプローチは、東日本大震災での被害の程度を共変量として加える方法で、この方法を試す根拠は、うつ症状という表現型はこのイベントから影響を受ける可能性があるからである(40-42)。被害の程度は地方自治体により測定された家屋倒壊度に基づいており(40)、4 は全壊、3 は大規模半壊、2 は半壊、1 は一部損壊、0 は損壊がないか被災地に居住していないとなるように、コーディングをした。予測モデルは下記のようにあらわされる。

$$y_{train} = b_0 + b_1 \times AGE_{train} + b_2 \times SEX_{train} + b_3 \times Damage_{train} + b_4 \times PC1_{train} + b_5 \times PC2_{train} + \cdots + \boldsymbol{SNP}_{train}\boldsymbol{b}_{SNP} + e_{train} \quad (7)$$

$$\mu_{test} = b_0 + b_1 \times AGE_{test} + b_2 \times SEX_{test} + b_3 \times Damage_{test} + b_4 \times PC1_{test} + b_5 \times PC2_{test} + \cdots + \boldsymbol{SNP}_{test}\boldsymbol{b}_{SNP} \quad (8)$$

ここで、$Damage$は上記の、東日本大震災による被害の程度を表している。

８　STMGP 法以外の予測モデルのパッケージやパラメータ

遺伝子スコア法については、現在最も使用されているパッケージの一つである PRSice （v1.25）パッケージを使用した。Clumping とは、GWAS の統計量に基づき、情報量の削減を最小限にしつつ、SNP の数を減らしてデータサイズを小さくする手法である。具体的には、GWAS の P 値が小さい順に SNP を並べて、順番に SNP の前後一定の間隔以内で、定めた r^2 値よりも大きい連鎖不平衡が認められる SNP は削除される。今回、遺伝子スコア法では連鎖不平衡を調べる範囲を、パッケージの推奨に従い、対象 SNP の前後 250 キロ塩基対で $r^2 > 0.1$ の SNP を削除するように設定した。遺伝子スコア法におけるカットオフの P 値候補については、パッケージの推奨に従い、トレーニングデータで 10-fold の交差検証を行った。可能な限り多くのカットオフ候補から選択できるよう、5×10^{-4} から 0.5 まで、5×10^{-4} 間隔での P 値候補を使用した。GBLUP 法と SBLUP 法を行うために、現在最も使用されているパッケージの一つである Genome-wide complex trait analysis （GCTA） v1.26.0 パッケージを使用した。GBLUP 法は SNP データを用いて、Genetic relation matrix を作成し、Restricted maximum likelihood analysis により解析対象の全ての SNP から説明される分散を推

測し、BLUP 法を用いて、遺伝子多型によるランダム効果を予測する。SBLUP 法は GWAS による統計量を用いた BLUP 解析である。SBLUP 法において cojo-sblup は、パッケージの推奨に従い、未インピュテーションデータの解析では 1.14e7 に、インピュテーションデータの解析では 9.50e7 に設定した。cojo-wind と thread-num オプションはそれぞれパッケージの推奨に従い、1000 と 20 に設定した。BayesR 法ではベイズ階層モデルのもと、Markov chain Monte Carlo （MCMC）を使用し、関連する SNP の検出と SNP で説明される分散の推測と、表現型の遺伝的構造の描出と、遺伝子多型からの表現型の予測を同時に行う(17)。我々は先行報告(17)に従い、MCMC 鎖の長さを 50,000 に設定し、burnin steps を 20,000 に設定した。また、計算コストの削減のために SNP 効果の更新に関してはパッケージの推奨通り、“-msize 500 -mrep 5000” オプションを設定した。STMGP と異なり、GWAS の統計量で重みづけをしない、一般的な Ridge 回帰（罰則つき回帰）法では、glmnet ver2.0-13 パッケージを使用した。Ridge 回帰法でのチューニングパラメータの λ は、パッケージの推奨に従い、トレーニングデータセットにおける 10-fold の交差検証で決定した。

９　STMGP 法以外の予測モデルのためのゲノムデータの準備

GBLUP 法と BayesR 法、Ridge 回帰法は、STMGP と同様に SNP データを、それぞれのパッケージの入力として使った。遺伝子スコア法と SBLUP 法は GWAS の統計量が入力情報として必要なため、STMGP と同じ共変量を使用して GWAS を行い、結果の統計量

をパッケージの入力として用いた。また、インピュテーションをしていない SNP データの解析において Ridge 回帰法が、またインピュテーションデータを用いた解析では Ridge 回帰法、SBLUP 法、BayesR 法が、計算負荷が大きすぎるために解析を行うことができなかった。具体的には、解析環境で用意できないサイズのメモリを要求する、または計算時間が 1 週間を超える、のいずれかである。上記のように計算負荷がかかりすぎて実現が困難であった解析については、先行研究(43)に従い、SNP データは SNP 数が約 30,000 になるまで Clumping をおこなった。Clumping 操作ではある一定範囲内の塩基対のうち、もっとも GWAS の P 値が小さいものを残す操作であるため、SNP 数の削減に比して、情報量の削減は小さい手法であり、先行の予測研究でも頻用されている(43, 44)。ただし、Ridge 回帰法や SBLUP 法、BayesR 法のように、連鎖不平衡関係にある SNP も含めて予測に使用するモデルにおいては、Clumping による情報量の低下が予測精度に影響する可能性は考えられる。

１０　STMGP 法以外の予測モデルの共変量の調整について

遺伝子スコア法で使用した PRSice （v1.25）パッケージに実装されている共変量オプションは、年齢、性別、主成分といった共変量の効果を除去するためにのみ使用されており、予測精度を上げるためには使用されない。SBLUP 法や BayesR 法は共変量を使用するオプションが存在しない。したがって、遺伝子スコア法、SBLUP 法、BayesR 法を、STMGP と公平に比較するために、我々は、下記のようなモデルを作成し、共変

量を予測精度を向上するために使用した。予測モデルは下記のようなモデルによって、トレーニングデータセットで学習した。

$$y_{train} = b_0 + b_1 \times AGE_{train} + b_2 \times SEX_{train} + b_3 \times PC1_{train} + b_4 \times PC2_{train} + \cdots + b_{GS} \times GS_{train} + e_{train} \quad (9)$$

推測された回帰係数を用いて、テストデータでの予測スコアは下記のモデルで得られた。

$$\mu_{test} = b_0 + b_1 \times AGE_{test} + b_2 \times SEX_{test} + b_3 \times PC1_{test} + b_4 \times PC2_{test} + \cdots + b_{GS} \times GS_{test} \quad (10)$$

ここで GS はそれぞれの予測モデルでゲノム情報をもと予測されたスコアを表している。

GBLUP 法では、GCTA の共変量オプションを使用するのに加えて、GCTA 開発者によって推奨されているように、共変量による固定効果を予測モデルに含めた。Ridge 回帰法では共変量と SNP 情報の両方が予測モデルに含まれ、予測モデルの数式は STMGP と同様に記載できる（(１）と（２))。

Ⅴ　研究結果

トレーニングデータセット、テストデータセットそれぞれの人口統計情報を表１に示した。トレーニングデータとテストデータが別々にリクルートされていることもあり、女性の割合、年齢、最終学歴、東日本大震災による家屋倒壊度、被災と CES-D

評価の間の時間差の項目は、データセット間で有意な違いがあった。こういったトレーニングデータセットとテストデータセット間の差異は、予測を難しくする可能性がある。

今回の主解析における、各予測モデルでの精度を表２に示した。独立したテストデータセットでの予測精度（PCC）を計算すると、STMGP 法の予測精度は、他の予測モデルと比較して有意に高かった。一方、予測モデルを作成するために使用したトレーニングデータセットで評価した、見かけ上の PCC を調べると、STMGP 法は他の手法よりも値が低く、トレーニング過程における過剰評価の程度が低かった。トレーニングデータでの過剰評価と、テストデータでの予測精度の差によってあらわされる、過剰適合の程度(45)は、STMGP 法において軽減させることに成功した。今回使用したゲノムデータにおける GWAS のマンハッタンプロットと QQ プロットを図４に示す。今回のデータでは CES-D とゲノムワイド有意に関連を認める SNP は存在しなかった。QQ プロットにおいては、体系的な P 値のインフレーションは認めなかった。STMGP 法で予測に使用された SNP とその回帰係数は表３に示す。STMGP 法の計算において、計算時間と、使用されたメモリ（ピーク時）はそれぞれ、107 分と 13GB であり、多くのコンピューターサーバーで扱える計算負担であった。

STMGP 法によって使用された GWAS の P 値のカットオフは 2.7×10^{-4} であり、これは、遺伝子スコア法によって使用された P 値のカットオフである 0.022 よりも低い値であった。遺伝子スコア法において、STMGP 法と同じ P 値カットオフを用いた修正遺

伝子スコア法の解析では予測精度（PCC）は 0.0285 であり、P 値カットオフ設定に交差検証を用いたもともとの遺伝子スコア法の予測精度（0.0332）よりも低かった。

STMGP 法において、異なる τ パラメータを用いた予測精度は表４に示している。今回設定したどの τ パラメータにおいても、競合する予測モデルよりも高い予測精度を示し、$\tau = N/10$と設定したものが最も高い PCC （0.1201）を示した。また、結果の一貫性を調べるために、表現型（CES-D）に Box-Cox 変換を行ったものや、外れ値を除外したもの、さらに、共変量の調整方法を変えて行ったものの結果は表５、６にそれぞれ示している。全ての方法において、STMGP 法のテストデータセットにおける予測精度は他のモデルよりも高く、STMGP 法の過剰適合の程度（トレーニングデータでの見かけ上の予測精度とテストデータでの予測精度の差）は他モデルと比較して小さかった。このことは、STMGP 法は過剰適合を抑えることで最良の結果を出すことを示している。

インピュテーションを行ったゲノムデータを使用した、各予測モデルの精度は表７に示している。インピュテーションデータをもとに予測した STMGP 法の PCC(0.0817)は、元の SNP データをもとに予測した STMGP 法の PCC（0.0769）よりも統計学的に有意に高くはなかった（$P=9.54\times10^{-2}$）。しかし、STMGP 法は依然として、過剰適合を軽減することで、他の予測モデルよりも高い精度を示した。

Ⅵ　考察

今回の研究は、STMGP 法をはじめて実際に収集されたゲノムデータに用いて、その有用性を確認したものである。STMGP 法は従来法と比較して、高い予測精度を示した。比較的少ないサンプルサイズでトレーニングをおこなった予測モデルでも、テストデータで有意な予測精度を示し、SNP 群が有意にうつ状態リスクに貢献していることを示した。STMGP 法と他の予測モデルの精度の違いの一因として、null variants の過剰評価を避けることで過剰適合を軽減させていることが考えられた。

今回の研究で STMGP が他の予測モデルよりも高い予測精度を示したことは、植木と田宮らの先行報告(15)の結果を、より困難な条件設定で再現したこととなる。今回のデータセットにおいては、うつ状態との関連が最も大きい SNP の標準化回帰係数が 0.057 であったが、これは、植木らが使用したアルツハイマー病のデータベースの 0.238 よりも小さい。STMGP 法以外の予測モデルにおいて深刻な過剰適合が起きた原因としては、効果サイズの小さい真の感受性バリアントと null variants の区別が困難であったことが考えられる。SNP のうつ症状への効果サイズの小ささは先行研究と一貫しており、うつ症状の予測の難しさを説明している。そのような困難な条件でも、STMGP 法は過剰適合を軽減して、有意な予測精度を示すことに成功した。このことは、STMGP 法が幅広い、複雑な遺伝疾患の予測へ応用ができる可能性を示している。

今回の研究では、単一の SNP とうつ症状との関連を示すには不十分なサンプルサイズの研究でも、STGMP 法を用いた方法では、表現型と有意な関連を示す遺伝的な脆

弱性スコアを作成することに成功した。今回のトレーニングデータでの GWAS ではゲノムワイド有意な SNP は一つも認めず、P 値が 5×10^{-5} を下回る SNP は 11 個あり、この上位 11SNP ではテストデータセットの表現型の 3.6×10^{-3}%の分散しか説明しなかった。GWAS の結果は、それぞれの SNP の表現型との関連を議論するには不十分な検出力しか得られていないことを示し、GWAS の P 値は、ランダムノイズの影響を大きく受けていることを示唆している。STMGP 法と遺伝子スコア法は両方とも GWAS の P 値に基づいて SNP を選択するが、STMGP 法は遺伝子スコア法よりも高い予測精度を示した。この結果は、STMGP 法の戦略である、選択の確実性を反映するように遺伝子多型を重みづけすることが、比較的小さなサンプルサイズでも成功していることを示している。反対に、遺伝子スコア法のように単純に遺伝子多型を合計していく方法では、検出力不足に大きく影響される可能性がある。

今回の研究では、ゲノム情報に注目しており、他の生物学的指標（生化学的な指標、脳画像、生理学的な指標）を用いた研究と異なり、環境因やうつ症状に起因する変化を排した、遺伝によって説明される CES-D の割合を計算することができる。トレーニングデータセットにおいて、GREML （genome-wide restricted maximum likelihood）によって計算した SNP 遺伝率は今回の研究では 0.05（標準誤差 0.07）であった。この結果は、これまでに CES-D を表現型として行われた最大のメタアナリシスの一つで 70,017 人を対象に行われたゲノムワイド関連メタアナリシス(46)の結果である、SNP 遺伝率が 0.04（標準誤差 0.01）という結果と矛盾しないものであっ

た。今回の研究で標準誤差が大きく算出されたのは、サンプルサイズの小ささによるものと考えられる。

今回の研究で、STMGP 法以外の手法では、統計学的に有意な予測精度が得られなかった。遺伝子スコア法、GBLUP 法、SBLUP 法、BayesR 法、Ridge 回帰法に共通した問題点としては、トレーニングデータで見かけ上の予測精度が高く、深刻な過剰適合が生じていた。過剰適合が起きる原因として、それぞれのモデルにおいて、Null variant を過大評価していることが考えらえる。また、BayesR 法はこういった過剰適合を抑制するために開発されたモデルではあるが、今回の解析では深刻な過剰適合が観察された。BayesR 法では設定しているパラメータの数が多く、今回のようにサンプルサイズが小さい場合には過剰適合の強い影響を受ける可能性が示唆された。また、SBLUP 法、BayesR 法、Ridge 回帰法は共変量のオプションが設定されておらず、共変量を予測に効果的に利用できていない可能性がある。これらのモデルでは、方法の（９）（１０）で記した数式を用いて共変量を含めて回帰モデルを作成したが、この方法が、もともと共変量のオプションがある STMGP 法、遺伝子スコア法、GBLUP 法と比較すると、共変量の情報を効果的に予測に利用できていない可能性がある。

今回のうつ症状を予測した予測精度と、これまでのゲノムに基づく先行予測研究の予測精度とを比較することは興味深いものの、異なる研究の予測精度を比較するには、多くの項目が一致している必要がある。予測精度は、予測モデルだけでなく、サンプルサイズ、表現型の遺伝的構造、表現型の評価方法（二値変数か量的変数か）、共

変量の調整方法にも依存している(11, 27)。したがって、今回計算された各モデルのPCC の値は、今回の研究内での比較には有用であるが、他の研究で計算された他の予測モデルの予測精度との比較はできない。今回の研究のデータセット、設定した条件下では、STMGP 法は他の手法よりも高い精度で CES-D を予測することに成功した。

STMGP 法は従来法よりも今回の研究で高い予測精度を示したが、その PCC は 0.1を下回っており、このままでは臨床に役立てるのは難しい。今後、STMGP 法を用いてうつ病のリスク予測を向上させるうえで、考えられる手段がいくつかある。最も直截的なアプローチは、トレーニングデータセットにより大きなサンプルサイズを用意することである(10, 11, 27)。Lasso や Elastic net などで、パラメータを決定する際に交差検証法を用いるような機械学習の手法と比較して、STMGP 法は交差検証法を用いないために、計算機の負担を抑え、より大きなサンプルサイズのデータに使用できる可能性がある。他には、全ゲノムシークエンスデータのように、より多くの遺伝子多型を含むデータセットを予測に用いるということも、これまでのレアバリアントがうつ病発症に関わるという報告を考慮すると、予測精度を上げる可能性がある。しかし、全ゲノムシークエンスデータは計算にコンピューターコストがかかるだけでなく、変数の数が増えるために過剰適合を引き起こしやすいという問題がある。STMGP 法はすでに全ゲノムシークエンスデータでも他の手法よりも優れた予測精度を示している(15)。STMGP 法を用いて、よりサンプル数、遺伝子多型数が多いデータによるトレーニングを行うことが、ゲノムデータからうつ状態を予測する精度をあげるために、

実現可能なアプローチと考える。さらに、ゲノム以外のオミクス情報をモデルに入れることや遺伝子環境相互作用をモデルに含めることも予測精度を上げていく可能性がある。

今回の解析では、STMGP 法では予測精度が他の手法と比較して高かったことに加えて、Clumping をせずにインピュテーションされたゲノムデータの解析が可能であり、他の手法に比べて計算負担が低いことが確認された。ゲノム情報や生化学的な情報、脳画像情報といった情報を入力として、それらから重要な変数を抽出したり、疾患を予測したりする機械学習が個別化医療の実現のために注目されている。しかし、機械学習の手法は計算負担が大きいことが課題である(47)。STMGP 法では機械学習において多大な計算負担を要する交差検証を用いないことで、計算負担を下げている。STMGP 法は医療現場での実装に関して、計算負担面で実現可能性が高い手法と考えられる。

今回の研究で STMGP 法において予測に使用された SNP は、結果に記したようにゲノムワイド有意に CES-D と関連するものは認められなかった。したがって、個別の SNP と CES-D の関連については今回の研究結果に基づいて議論することは難しい。その一方で、STMGP 法により有意な予測精度が示されたことより、STMGP 法で予測に使用された SNP102 個の中に、宮城県データと岩手県データに共通して、うつ状態に寄与している SNP が含まれている可能性がある。STMGP 法で使用された SNP102 個が位置する遺伝子の機能解析をするために、Ingenuity Pathway Analysis (QIAGEN

bioinformatics, Germantown, MD, USA)でパスウェイ解析を行った。SNP102 個のうち、60 個がデータベース上の多型情報にマップされた。そして、今回予測に使用された rs7312048 (Intron)、rs7138222 (Intron)、rs6581940 (Intron)が位置する KCNMB4 遺伝子、rs449998(Intron)、rs2837657(Intron)が位置する DSCAM 遺伝子、rs17029241(Intron)、rs17029245(Intron)が位置する CMTM8 遺伝子、rs788159(Intron)が位置する METAP1D 遺伝子、rs12806(Non Coding Transcript Variant)、rs17112705(Intron)が位置する ERLIN 遺伝子、rs2230804(Missense Variant)、rs12570957(Intron)が位置する CHUK 遺伝子が、ドパミン代謝系とグルタミン代謝系で構成される同一ネットワーク内に集積して存在していた。ドパミン代謝系やグルタミン代謝系はすでに抗うつ薬のターゲットとして注目されているネットワークであり(48, 49)、うつ状態に関連する多型としても矛盾はしない。

今回の解析で使用したトレーニングデータとテストデータの間では、男女比、被災状況 (家屋倒壊度)、最終学歴において分布に違いがあった。特に、被災状況についてはトレーニングデータでは損壊なしが 28.2%であったのに対し、テストデータでは損壊なしが 71.3%と、トレーニングデータのほうが、被災程度が重い集団であった。実際に臨床場面で予測モデルを使う際には、予測を行うテスト患者と、完全に同じ背景のトレーニングデータは用意できないので、臨床場面の状況に近いデザインでの予測実験ができたといえる。その一方、データを一つにまとめて交差検証法を用いたほうが、よりテストデータに近縁な被験者がトレーニングデータに含まれ、予測精度が

向上すると報告されている(47)。我々は、今回の研究においてなるべく一般性の高い予測モデルを作成するために、被災程度が重度から軽度まで含まれ、都市在住者と地方在住者が両方含まれるような宮城県をトレーニングデータとして採用した。

今回の試験の限界としては、表現型としてCES-Dを用いており、うつ病の診断を表現型とした場合とは、異なるSNP群が選択されている可能性がある。うつ病を予測するためにSTGMPモデルを使用する場合は、同じ表現型を用いたトレーニングデータセットで学習を行う必要があるかもしれない。また、サンプルサイズが比較的小さいので、予測に用いられたSNP群と表現型との関連性の議論は慎重に行う必要がある。さらに、うつ病への介入に関する情報が欠如していることも限界の一つである。内服薬剤に関する情報は、解析段階では、トレーニングデータセットの情報でのみ得られた。トレーニングデータセットでは、56人（1.5%）が抗うつ薬を内服していた。表1に示すように、うつ病の既往を有する被験者の頻度分布は、トレーニングとテストデータの間でほぼ同等であったため、トレーニングデータセットと同等の割合の被験者が、テストデータにおいても抗うつ薬を内服していた可能性がある。

現在のSTMGP法の実装面では二つの限界がある。STMGPアルゴリズムは重みづけされたL2罰則をもつ、本質的には一般化線形モデルの多型であるので、P値でカットオフされたSNPを説明変数とした線形/ロジスティック回帰と、スケーラビリティは同等である。しかしながら、現在の実装で計算可能なデータサイズは、多くても数千人規模のSNPデータ・インピュテーションされたゲノムデータ、全ゲノムシークエン

スデータであり、数万人規模は難しい。さらに、STMGP は現時点では、SNP 同士の相関の計算のために、個人単位の SNP データを必要としており、GWAS の統計量のみでは計算ができない。我々は、スケーラビリティの問題や統計量のみを用いた実装が可能となるように現在、新たなソフトウェアを開発中である。

Ⅶ　結論

今回の研究は、STMGP 法を初めて実際の大規模ゲノムデータに実装し、その予測精度の高さと計算負担面における実現可能性の高さを示したものである。ゲノムワイド SNP データからのうつ状態の予測において、STMGP 法が従来法よりも高い予測精度を示した要因として、過剰適合を軽減していることが示された。加えて、STMGP 法は過剰適合を抑える機械学習アルゴリズムでありながら、インピュテーションされたゲノムデータにも実装可能な程度の低い計算負担であった。したがって、STMGP 法は精神疾患のようにポリジーンモデルを呈する疾患の予測に、現時点で最適な選択の一つであることが示唆された。生物学的な情報を用いてより精緻なうつ病のサブタイプ分類をすることで個別化医療を目指すにあたり、高性能で実現可能性の高い STMGP 法が貢献する役割は大きいと考える。

Ⅷ　謝辞

本研究は、著者が東北大学東北メディカル・メガバンク機構の多くの先生方から

ご指導をいただき、執筆したものです。まず、研究の機会を与えていただき、精神医学の側面に関して丁寧にご指導をいただきました富田博秋先生に感謝の意を表します。また、遺伝統計学に関して細部に至るまでご指導をいただきました植木優夫先生、田宮元先生に深く感謝いたします。また、基礎論文について査読いただき、様々な視点より有益なご助言をくださった荻島 創一教授、木下賢吾教授、寶澤 篤教授、峯岸直子教授、長神 風二教授、福本 健太郎先生、大塚耕太郎教授、丹野 高三 教授、坂田 清美 教授、清水 厚志 教授、佐々木 真理教授、祖父江憲治教授、呉 繁夫 教授、山本 雅之 教授に深く感謝いたします。最後に、東北大学東北メディカル・メガバンク機構といわて東北メディカル・メガバンク機構の全ての部門の先生方、スタッフの方々、そして東北メディカル・メガバンクプロジェクトに参加された被験者の方々に深く感謝申し上げます。

Ⅸ 文献

1. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet (London, England). 2006;367(9524):1747-57.

2. Wittchen HU, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jonsson B, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. 2011;21(9):655-79.

3. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. The American journal of psychiatry. 2000;157(10):1552-62.

4. Health WFfM. DEPRESSION: A Global Crisis. 2012.

5. Lopresti AL, Maker GL, Hood SD, Drummond PD. A review of peripheral biomarkers in major depression: the potential of inflammatory and oxidative stress biomarkers. Prog Neuropsychopharmacol Biol Psychiatry. 2014;48:102-11.

6. Zheng H, Zheng P, Zhao LC, Jia JM, Tang SL, Xu PT, et al. Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. Clinica Chimica Acta. 2017;464:223-7.

7. Flint J, Kendler KS. The genetics of major depression. Neuron. 2014;81(3):484-503.

8. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nature neuroscience. 2019;22(3):343-52.

9. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nature genetics. 2016;48(9):1031-6.

10. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature genetics. 2018;50(5):668-81.

11. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS genetics. 2013;9(3):e1003348.

12. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.

13. Colodro-Conde L, Couvy-Duchesne B, Zhu G, Coventry WL, Byrne EM, Gordon S, et al. A direct test of the diathesis-stress model for depression. Molecular psychiatry. 2018;23(7):1590.

14. Maier R, Moser G, Chen G-B, Ripke S, Absher D, Agartz I, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. The American Journal of Human Genetics. 2015;96(2):283-94.

15. Ueki M, Tamiya G, Alzheimer's Disease Neuroimaging I. Smooth-Threshold Multivariate Genetic Prediction with Unbiased Model Selection. Genet Epidemiol. 2016;40(3):233-43.
16. Robinson MR, Kleinman A, Graff M, Vinkhuyzen AA, Couper D, Miller MB, et al. Genetic evidence of assortative mating in humans. Nature Human Behaviour. 2017;1(1):0016.
17. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS genetics. 2015;11(4):e1004969.
18. Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, Osumi N, et al. The Tohoku Medical Megabank Project: Design and Mission. Journal of epidemiology. 2016;26(9):493-511.
19. Zhang W, Ng HW, Shu M, Luo H, Su Z, Ge W, et al. Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. Journal of genetics. 2015;94(4):731-40.
20. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. American journal of human genetics. 2013;93(4):687-96.
21. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5(6):e1000529.
22. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nature communications. 2015;6:8018.
23. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
24. Radloff L. The CES-D scale: A self-report depression scale for research in the general population. Applied psychological measurement. 1977;1(3):385-401.
25. Wada K, Tanaka K, Theriault G, Satoh T, Mimura M, Miyaoka H, et al. Validity of the Center for Epidemiologic Studies Depression Scale as a screening instrument of major depressive disorder among Japanese workers. American journal of industrial medicine. 2007;50(1):8-12.
26. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Computational statistics & data analysis. 2008;52(12):5186-201.
27. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature genetics. 2013;45(4):400.
28. Momen M, Mehrgardi AA, Sheikhy A, Esmailizadeh A, Fozi MA, Kranis A, et al. A predictive assessment of genetic correlations between traits in chickens using markers. Genetics, selection, evolution : GSE. 2017;49(1):16.
29. Williams EJ. Regression analysis: wiley; 1959.
30. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2017.
31. Ueki M. A note on automatic variable selection using smooth-threshold estimating

equations. Biometrika. 2009:1005-11.

32. Zou H. The adaptive lasso and its oracle properties. Journal of the American statistical association. 2006;101(476):1418-29.

33. Bühlmann P, Van De Geer S. Statistics for high-dimensional data: methods, theory and applications: Springer Science & Business Media; 2011.

34. Gregory KB, Wang D, McMahan CS. Adaptive elastic net for group testing. Biometrics. 2018.

35. Huang J, Ma SG, Zhang CH. Adaptive Lasso for Sparse High-Dimensional Regression Models. Statistica Sinica. 2008;18(4):1603-18.

36. van de Geer S, Bühlmann P, Zhou S. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). Electronic Journal of Statistics. 2011;5:688-749.

37. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics. 2006;2(12):e190.

38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9.

39. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS genetics. 2013;9(7):e1003608.

40. Nakaya N, Nakamura T, Tsuchiya N, Tsuji I, Hozawa A, Tomita H. The Association Between Medical Treatment of Physical Diseases and Psychological Distress After the Great East Japan Earthquake: The Shichigahama Health Promotion Project. Disaster Med Public. 2015;9(4):374-81.

41. Tsuchiya N, Nakaya N, Nakamura T, Narita A, Kogure M, Aida J, et al. Impact of social capital on psychological distress and interaction with house destruction and displacement after the Great East Japan Earthquake of 2011. Psychiatry Clin Neurosci. 2017;71(1):52-60.

42. Yoshida H, Kobayashi N, Honda N, Matsuoka H, Yamaguchi T, Homma H, et al. Post - traumatic growth of children affected by the Great East Japan Earthquake and their attitudes to memorial services and media coverage. Psychiatry and clinical neurosciences. 2016;70(5):193-201.

43. Cherlin S, Plant D, Taylor JC, Colombo M, Spiliopoulou A, Tzanis E, et al. Prediction of treatment response in rheumatoid arthritis patients using genome-wide SNP data. Genet Epidemiol. 2018;42(8):754-71.

44. Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome - wide association studies. Genetic epidemiology. 2010;34(7):643-52.

45. Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? Contemp Clin Trials. 2013;36(2):636-41.

46. Direk N, Williams S, Smith JA, Ripke S, Air T, Amare AT, et al. An analysis of two genome-wide association meta-analyses identifies a new locus for broad depression phenotype. Biological psychiatry. 2017;82(5):322-9.

47. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, et al. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry. 2012;17(10):956-9.

48. Dailly E, Chenu F, Renard CE, Bourin M. Dopamine, depression and antidepressants. Fundam Clin Pharmacol. 2004;18(6):601-7.

49. Krystal JH, Sanacora G, Blumberg H, Anand A, Charney DS, Marek G, et al. Glutamate and GABA systems as targets for novel antidepressant and mood-stabilizing treatments. Molecular Psychiatry. 2002;7(S1):S71-S80.

X　表

	トレーニングデータ	テストデータ	P 値[a]
被験者数	3,685	3,048	
女性の割合	70.1%	65.3%	3.31×10^{-5}
CES-D の平均（標準偏差）	13.6 （7.2）	13.4 （6.9）	0.226
年齢の平均 （標準偏差）	58.5 （12.1）	62.0 （10.1）	$1{,}35\times10^{-38}$
最終学歴			6.54×10^{-37}
小中学校	640 （17.4%）	946 （31.0%）	
高等学校	1852 （50.3%）	1260 （41.3%）	
短期大学	903 （24.5%）	649 （21.3%）	
大学	279 （7.6%）	187 （6.1%）	
大学院	11 （0.3%）	6 （0.2%）	
東日本大震災による被害状況（家屋倒壊度）			1.09×10^{-278}
全壊	561 （15.2%）	218 （7.2%）	
大規模半壊	248 （6.7%）	61 （2.0%）	
半壊	302 （8.2%）	75 （2.5%）	
一部損壊	1534 （41.6%）	522 （17.1%）	
損壊なし	1040 （28.2%）	2172 （71.3%）	
精神疾患の既往歴			
うつ病	104 （2.8%）	81 （2.7%）	0.708
双極性障害	9 （0.2%）	6 （0.2%）	0.798
家族歴[b]			
うつ病	203 （5.5%）	167 （5.5%）	1.00
双極性障害	27 （0.7%）	26 （0.9%）	0.583
東日本大震災と CES-D 回答までの平均月数 （標準偏差）	28.5 （2.0）	30.8 （1.3）	9.88×10^{-324}
県	宮城	岩手	

表１：トレーニング・テストデータセットの人口統計情報

略語

CES-D（Center for Epidemiologic Studies - Depression Scale）

[a] P 値は CES-D、年齢、東日本大震災と CES-D 回答までの月数については Student’s t-tests を、女性の割合、最終学歴、東日本大震災による被害状況、精神科的な既往歴、家族歴については Fisher’s exact tests を使用して算出した。

[b] 家族歴は一親等の血縁者（両親、同胞、子）の精神科的な既往歴を示す。

	独立したテストデータでの PCC（標準誤差）	トレーニングデータでの PCC（標準誤差）	PCC の P 値	STMGP 法と他のモデルの PCC の違いについての P 値	予測モデルに使用されている SNP の数	SNP を選択する際の P 値のカットオフ
STMGP 法	0.0769 (0.0173)	0.3232 (0.0153)	2.114×10^{-5}		102	1.8×10^{-4}
遺伝子スコア法	0.0332 (0.0178)	0.9027 (0.0076)	0.06719	0.04992	13,421	2.2×10^{-2}
GBLUP 法	0.0309 (0.0178)	0.9627 (0.0017)	0.08806	0.04092	601,239	NA
SBLUP 法	0.0164 (0.0178)	0.9554 (0.0019)	0.3663	0.0103	599,149	NA
BayesR 法	0.0100 (0.0185)	0.9633 (0.0015)	0.5801	7.614×10^{-4}	615,386	NA
Ridge 回帰法	0.0260 (0.0178)	0.9998 (0.0000)	0.1511	0.02702	30,333	NA

表２：うつ状態の予測精度（PCC）

略語

PCC（predictive correlation coefficient）

STMGP 法（Smooth-Threshold Multivariate Genetic Prediction）

GBLUP 法（genomic best linear unbiased prediction）

SBLUP 法（summary-data based best linear unbiased prediction）

染色体	SNP ID	マイナーアレル	メジャーアレル	回帰係数
1	rs6696013	A	C	-0.000316822
1	rs2174596	A	G	-0.109634579
1	rs12136807	A	G	0.015725808
1	rs7531107	T	C	0.002871295
1	rs1539098	C	A	0.067651135
1	rs7540470	T	C	0.135646346
2	rs7576288	A	G	-0.076398888
2	rs6725242	G	A	-0.129266112
2	rs12623956	C	T	-0.081596897
2	rs788159	T	G	0.112805298
2	rs12478389	T	C	0.041326254
2	rs13424509	G	A	0.024054805
3	rs12497785	T	C	-0.000551287
3	rs17029241	G	A	0.093923526
3	rs17029245	G	A	0.01913246
3	rs16827675	G	T	-0.068114816
4	rs4690340	T	G	0.028884966
4	rs4690339	C	T	0.045099471
4	rs2687410	A	G	-0.112224539
4	rs17578337	T	C	0.024765976
5	rs10071484	C	T	-0.024929414
5	rs4400166	A	G	-0.04582722
5	rs10512781	C	T	-0.117021989
5	rs4957383	G	T	-0.185772013
5	rs4703712	C	A	0.059099728
5	rs414536	A	G	0.018826112
5	rs919769	C	A	-0.021608124
5	rs2678196	C	T	0.040359976
5	rs10477276	A	G	-0.040681735
5	rs11167904	T	C	-0.137892569
5	rs10463375	T	C	-0.00615862
5	rs7718967	A	G	-0.094092227
5	rs12657244	C	T	-0.007361145
5	rs17720191	G	A	-0.25733267
6	rs6901079	C	T	0.061791106
6	rs3130215	A	G	-0.036552542
6	rs2180346	C	T	0.109381191
7	rs6945486	G	A	-0.084014422
7	rs4731659	G	A	-0.051558373
7	rs4731660	A	G	-0.089169563
8	rs1355303	A	G	-0.00653999
8	rs10088391	T	C	-0.006422325
8	rs9643782	C	T	0.126589795
8	rs7000275	A	G	0.013606826
8	rs444282	C	T	0.034255137
8	rs2008398	G	A	0.095856347
8	exm2270975	T	C	0.070950047

染色体	SNP ID	マイナーアレル	メジャーアレル	回帰係数
8	rs9792192	T	C	0.070950047
9	rs4977974	C	T	-0.006805982
10	rs12778518	T	G	0.001469451
10	rs1112787	T	C	0.030488841
10	rs2488630	T	G	-0.096981862
10	rs975752	A	G	-0.032713161
10	rs12806	G	A	-0.021838503
10	rs17112705	C	T	0.14681868
10	exm849056	T	C	-0.038529499
10	rs2230804	T	C	-0.031072571
10	rs12570957	A	C	0.016234331
10	rs2270961	A	G	-0.072031089
10	rs10510138	T	C	0.141613533
10	rs2306702	C	A	0.125280557
11	rs4754140	A	G	0.014297365
11	rs10160360	C	T	0.014297365
11	rs1940791	C	T	0.013819073
11	rs2186671	A	G	0.0597174
11	rs7946281	C	T	-0.113669556
11	rs2082273	T	G	-0.137663404
11	rs7927195	A	G	-0.008985451
11	rs88493	T	C	-0.011338025
11	rs356240	A	C	-0.152995318
11	rs7108112	A	G	0.050707967
12	rs7964690	T	G	0.001774714
12	rs1910186	G	T	0.034079656
12	rs7312048	G	A	0.072823532
12	rs7138222	A	C	0.12254391
12	rs6581940	T	C	0.13685854
12	rs10777815	C	T	-0.056280781
12	rs1035203	A	G	-0.075077133
12	rs11108730	C	A	-0.063547139
12	rs11108735	G	A	-0.072923432
13	rs9581268	A	C	0.11632914
13	rs876505	A	G	0.037787236
14	rs8003430	G	A	0.141577636
14	rs6573779	C	T	0.047878185
14	rs10147970	G	A	0.218914031
14	rs17094008	A	G	-0.027513632
16	rs1477390	C	T	0.173393608
16	rs1424035	A	C	0.11422743
16	rs30882	C	T	0.056621262
16	rs153669	T	C	0.065869633
16	rs30893	G	A	0.031689929
16	rs648929	G	A	0.052856206
16	rs1774414	T	C	-0.033701265
18	rs11659603	A	G	-0.018735175

染色体	SNP ID	マイナーアレル	メジャーアレル	回帰係数
18	rs6506816	C	A	-0.04203481
20	rs8118592	C	T	-0.142498042
20	rs16981118	C	A	0.010909118
20	rs6112580	C	T	0.017373518
20	rs13045057	G	A	0.004369891
21	rs449998	A	G	-0.03332973
21	rs2837657	C	A	0.024467292
22	rs136947	T	C	0.227809713

表３：STMGP 法で予測に用いられた SNP

略語

SNP (single nucleotide polymorphism)

STMGP 法 (Smooth-Threshold Multivariate Genetic Prediction)

	独立したテストデータでの PCC （標準誤差）	トレーニングデータでの PCC （標準誤差）
STMGP 法 （$\tau = N/\sqrt{\log\ (N)}$）	0.0769 （0.0173）	0.3232 （0.0153）
STMGP 法 （$\tau = N/0.1$）	0.0531 （0.0175）	0.3739 （0.0150）
STMGP 法 （$\tau = N/1$）	0.0953 （0.0172）	0.3088 （0.0154）
STMGP 法 （$\tau = N/10$）	0.1201 （0.0176）	0.1690 （0.0159）

表４：異なる τ パラメータを設定した際の STMGP 法の予測精度

略語

PCC (predictive correlation coefficient)

STMGP 法 (Smooth-Threshold Multivariate Genetic Prediction)

	Box-Cox 変換		外れ値除外	
	独立したテストデータでの PCC （標準誤差）	トレーニングデータでの PCC （標準誤差）	独立したテストデータでの PCC （標準誤差）	トレーニングデータでの PCC （標準誤差）
STMGP 法	0.0684 (0.0175)	0.2849 (0.0154)	0.0485 (0.0183)	0.2839 (0.0148)
遺伝子スコア法	0.0150 (0.0180)	0.6829 (0.0108)	-0.0195 (0.0192)	0.5984 (0.0113)
GBLUP 法	0.0283 (0.0184)	0.9680 (0.0014)	0.0064 (0.0181)	0.9601 (0.0016)
SBLUP 法	0.0133 (0.0184)	0.9565 (0.0016)	-0.0110 (0.0182)	0.9579 (0.0014)
BayesR 法	0.0122 (0.0186)	0.9567 (0.0016)	-0.0229 (0.0181)	0.9589 (0.0014)
Ridge 回帰法	0.0282 (0.0180)	0.9998 (0.0000)	0.0091 (0.0187)	0.9998 (0.0000)

表５：表現型の分布に関して結果の一貫性を確認する調査結果（Box-Cox 変換により表現型を正規分布にしたものと外れ値を除外したもの）

略語

PCC（predictive correlation coefficient）

STMGP 法（Smooth-Threshold Multivariate Genetic Prediction）

GBLUP 法（genomic best linear unbiased prediction）

SBLUP 法（summary-data based best linear unbiased prediction）

	共変量を予測に使用するモデル 共変量：年齢、性別、主成分得点、家屋倒壊度		共変量は事前に調整し、予測には使用しないモデル 共変量：年齢、性別、主成分得点		共変量は事前に調整し、予測には使用しないモデル 共変量：年齢、性別、主成分得点、家屋倒壊度	
	独立したテストデータでの PCC（標準誤差）	トレーニングデータでの PCC（標準誤差）	独立したテストデータでの PCC（標準誤差）	トレーニングデータでの PCC（標準誤差）	独立したテストデータでの PCC（標準誤差）	トレーニングデータでの PCC（標準誤差）
STMGP 法	0.0817 (0.0177)	0.3784 (0.0148)	0.0155 (0.0182)	0.3928 (0.0153)	0.0192 (0.0182)	0.4638 (0.0145)
遺伝子スコア法	0.0493 (0.0175)	0.9324 (0.0076)	-0.0005 (0.0182)	0.8076 (0.0089)	0.0090 (0.0178)	0.8835 (0.0081)
GBLUP 法	0.04253 (0.0174)	0.9628 (0.0017)	-0.0003 (0.0178)	0.9610 (0.0018)	-0.0044 (0.0177)	0.9604 (0.0018)
SBLUP 法	0.0163 (0.0177)	0.9539 (0.0019)	-0.0011 (0.0178)	0.9564 (0.0018)	-0.0052 (0.0177)	0.9562 (0.0018)
BayesR 法	0.0110 (0.0183)	0.9634 (0.0014)	0.0080 (0.0185)	0.9604 (0.0016)	0.0117 (0.0187)	0.9445 (0.0022)
Ridge 回帰法	0.0175 (0.0177)	0.9998 (0.0000)	NA[a]	NA[a]	NA[a]	NA[a]

表６：共変量の調整に関して結果の一貫性を確認する調査結果（共変量として家屋倒壊度を使用した解析と、共変量を事前に調整して予測には使用しない解析）

[a] 予測された表現型のスコアが全サンプルにおいて同じ値となったため、PCC は計算していない

略語

PCC (predictive correlation coefficient)

STMGP 法 (Smooth-Threshold Multivariate Genetic Prediction)

GBLUP 法 (genomic best linear unbiased prediction)

SBLUP 法 (summary-data based best linear unbiased prediction)

	独立したテストデータでのPCC（標準誤差）	トレーニングデータでのPCC（標準誤差）	PCCのP値	STMGPと他のモデルのPCCの違いについてのP値	予測モデルに使用されているSNPの数	SNPを選択する際のP値のカットオフ
STMGP法	0.0888 (0.0173)	0.1793 (0.0153)	9.187×10^{-7}		72	1.7×10^{-5}
遺伝子スコア法	0.0205 (0.0176)	0.8311 (0.0082)	0.2580	3.209×10^{-3}	9,005	1.5×10^{-3}
GBLUP法	0.0347 (0.0181)	0.9521 (0.0022)	0.05565	1.699×10^{-2}	5,949,462	NA
SBLUP法	0.0168 (0.0178)	0.9535 (0.0013)	0.3539	2.392×10^{-3}	30,007[a]	NA
BayesR法	0.0092 (0.0182)	1.0000 (0.0000)	0.6118	1.462×10^{-3}	30,007[a]	NA
Ridge回帰法	0.0080 (0.0184)	0.9998 (0.0000)	0.6573	8.696×10^{-4}	33,538[a]	NA

表７：インピュテーションされたゲノムデータを用いた予測精度

[a] SBLUP法、BayesR法、Ridge回帰法では、計算コストが大きすぎて計算が不可能であったために、先行報告[1]に従って、事前に約３万の数にまでSNPの数をClumpingにより減らしている。

略語

PCC（predictive correlation coefficient）

STMGP法（Smooth-Threshold Multivariate Genetic Prediction）

GBLUP法（genomic best linear unbiased prediction）

SBLUP法（summary-data based best linear unbiased prediction）

SNP（single nucleotide polymorphism）

XI　図

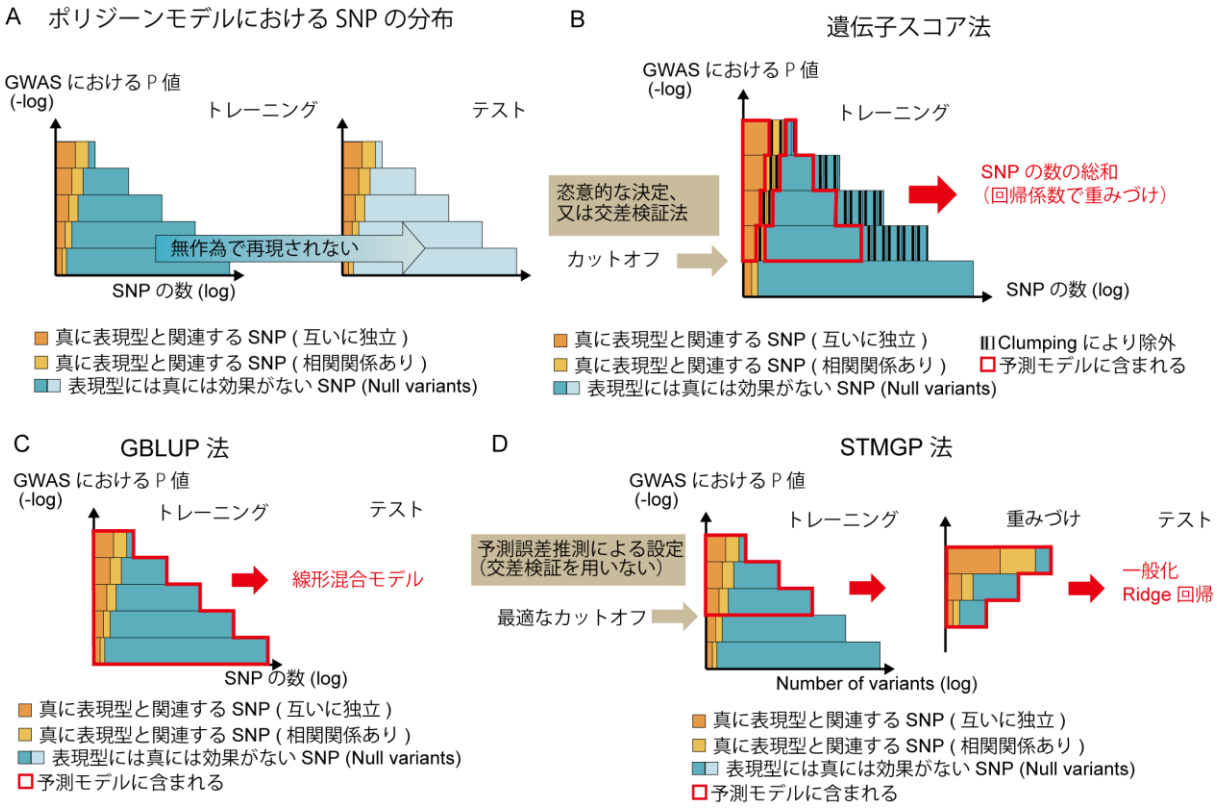


図１：ポリジーンモデルにおける遺伝的な構造と予測モデルに関する概念図。

（A）ポリジーンモデルに従う疾患におけるトレーニングデータセットとテストデータセットでの GWAS の P 値の分布。Y 軸は GWAS における P 値の負の対数を示し、X 軸は SNP の数の対数を示す。オレンジ色や黄色で示される、真に表現型と関連する SNP の P 値は小さい傾向があるが、検出力が十分でないために、一部は比較的大きな値の P 値が算出される。一方、青で示される、表現型には真には関連がない SNP （null variants）は P 値が大きくなる傾向があるが、検定の回数が多いために、一部は小さい P 値にもなりうる。

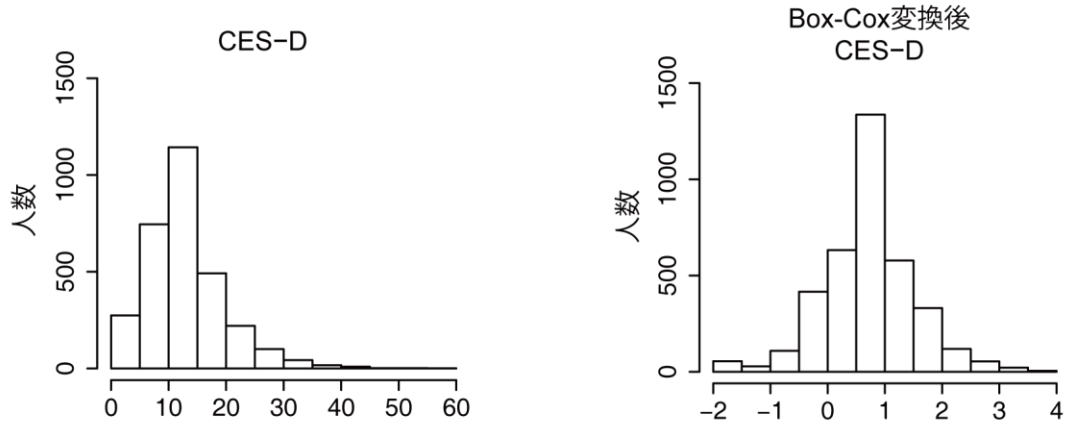
真に表現型と関連する SNP は互いに独立である群（オレンジ）と、それらと連鎖不平衡により相関関係にある群（黄色）に分けられる。真に関連する SNP（オレンジと黄色）は予測モデルに含まれた際にテストデータでの予測精度を向上させる。一方、検定回数が多いために偶然 P 値が低く出た null variants（青）が予測モデルに含まれてしまうと、トレーニングデータでの計算ではあたかも予測精度が高いように見えるが、テストデータでの計算では予測精度が低くなる。これを過剰適合と呼ぶ。現在手に入るサンプルサイズ（数十万人程度）では、うつ症状と真に関連がある SNP と null variants を、1 度の GWAS で区別することは困難である。

（B）遺伝子スコア法の概念図。遺伝子スコア法は真に関連する SNP を選択し、null variants からの影響を軽減させるために、GWAS の結果の P 値にカットオフを設けて SNP を選択する。しかしながら、ポリジーンモデルにおいては遺伝子スコア法で予測精度が十分に高くでない限界が 2 点ある。それは、(i) 真に関連する SNP を含めようとすると多くの null variants が含まれてしまい、それらを過大評価してしまうことと、(ii) Clumping を行い、互いに相関する、真に関連する

SNP で、予測モデルに含まれれば予測精度を高めるもの（黄）が除外されてしまうことである。
（C）GBLUP 法の概念図。GBLUP 法では真に関連する SNP のうち、互いに相関するものも含めて効果的に予測モデルに含めることができる。しかし、このモデルでは、遺伝子スコア法以上に多くの null variants を予測モデルに含めてしまい、過剰適合を引き起こし、予測精度が低くなってしまう可能性がある。
（D）STMGP の概念図。STMGP ではコンピューターコストのかかる交差検証法を使用せずに、予測精度向上に最適な P 値のカットオフを計算する。さらに、互いの相関関係と GWAS の統計量の両方を考慮した、罰則付き回帰（一般化 Ridge 回帰）によって SNP を重みづけすることで、相関関係にある SNP の効果的な利用と過剰適合の軽減を可能にする。

略語
GWAS（genome-wide association study;全ゲノム関連解析）
GBLUP 法（genomic best linear unbiased prediction）
STMGP 法（Smooth-Threshold Multivariate Genetic Prediction）

トレーニングデータ



テストデータ

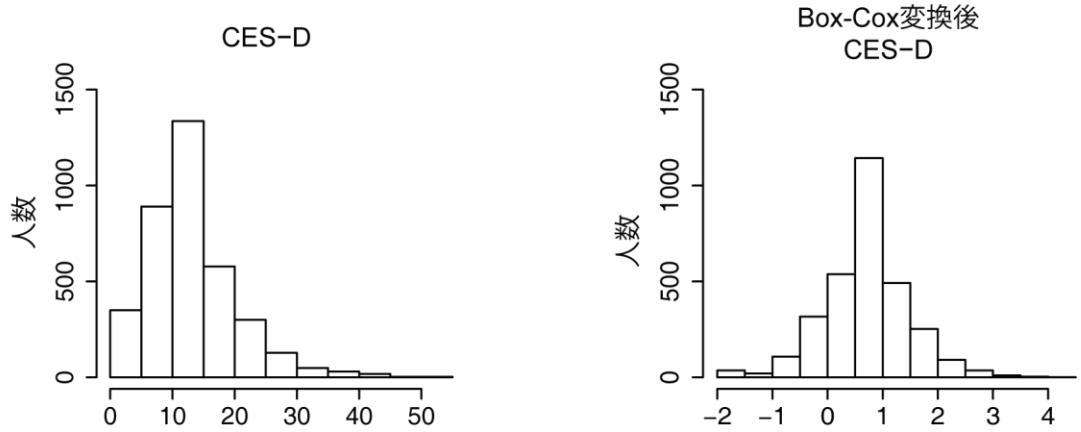


図２：CES-D スコアの度数分布図

略語

CES-D (Center for Epidemiologic Studies-Depression)

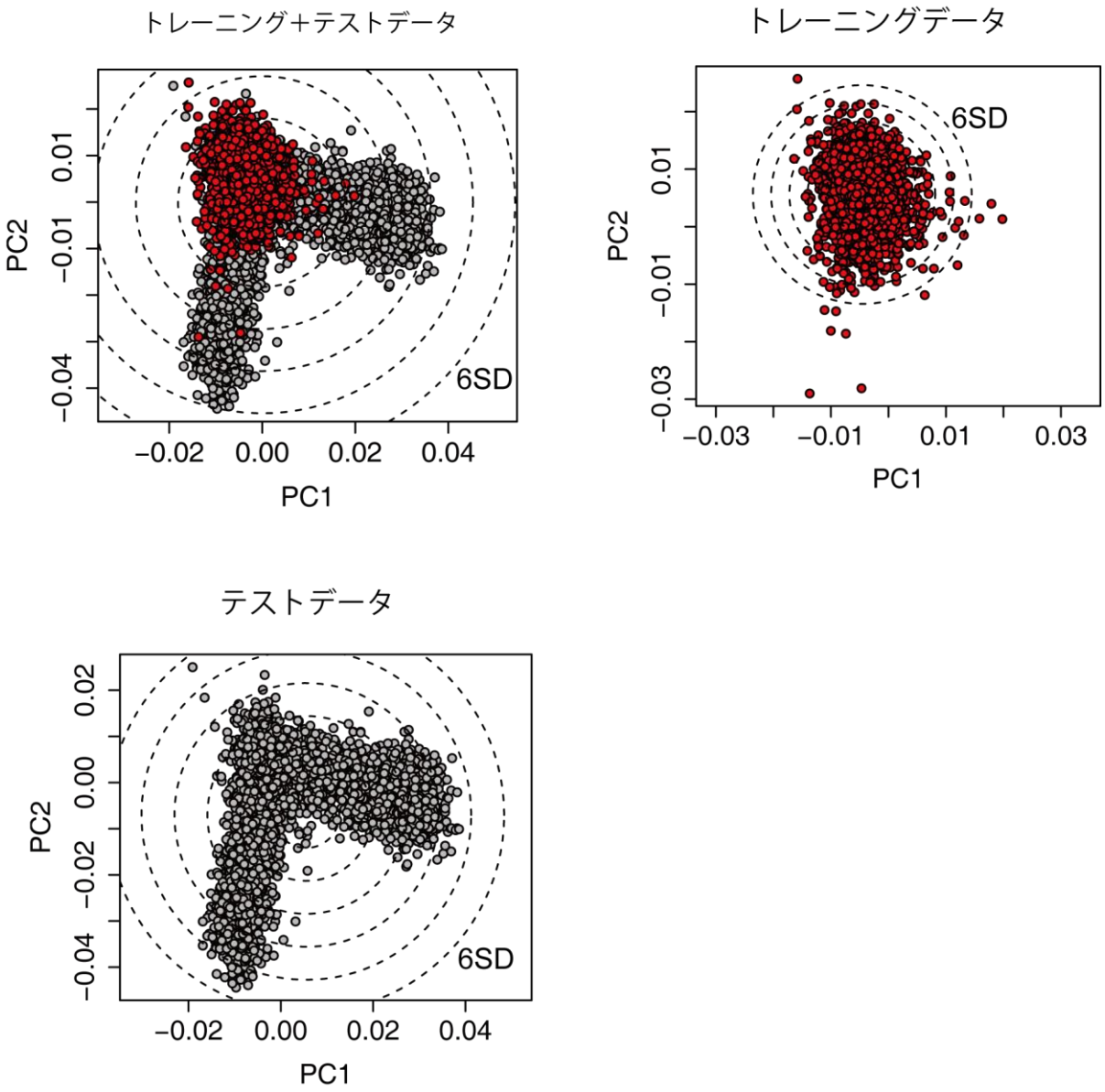


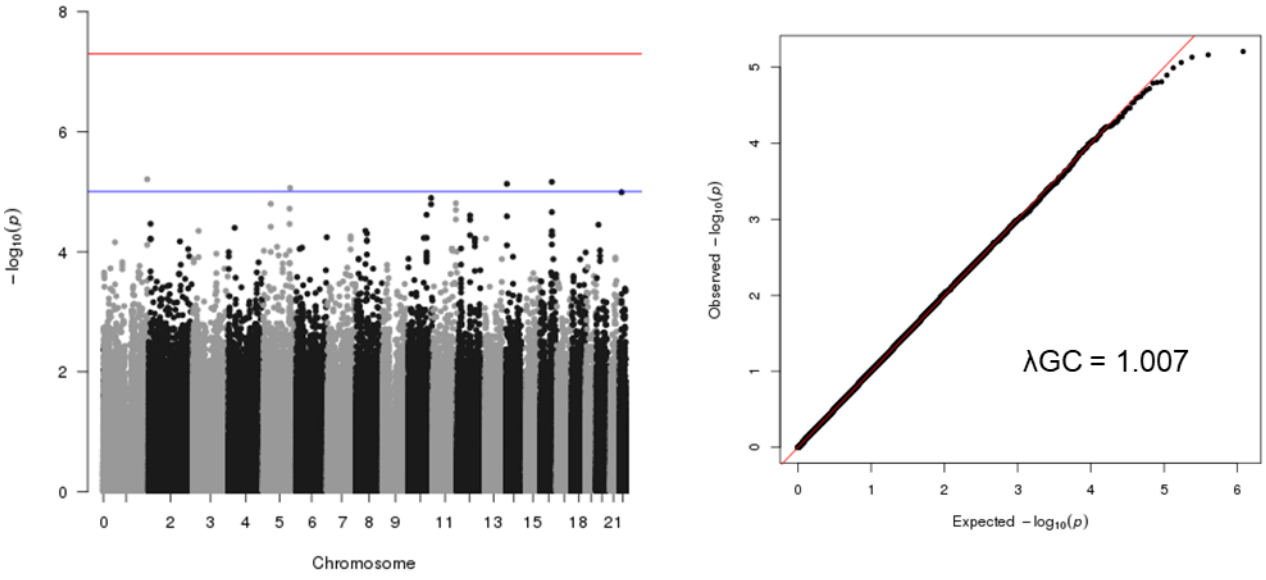
図３：使用したデータにおける集団階層性の検討（第１、第２主成分スコアの図示）。各図において、同心円はサンプル平均からの１－６SD を表している。

略語

PC（principal component；主成分）

SD（standard deviation；標準偏差）

トレーニングデータセット



トレーニング＋テストデータセット

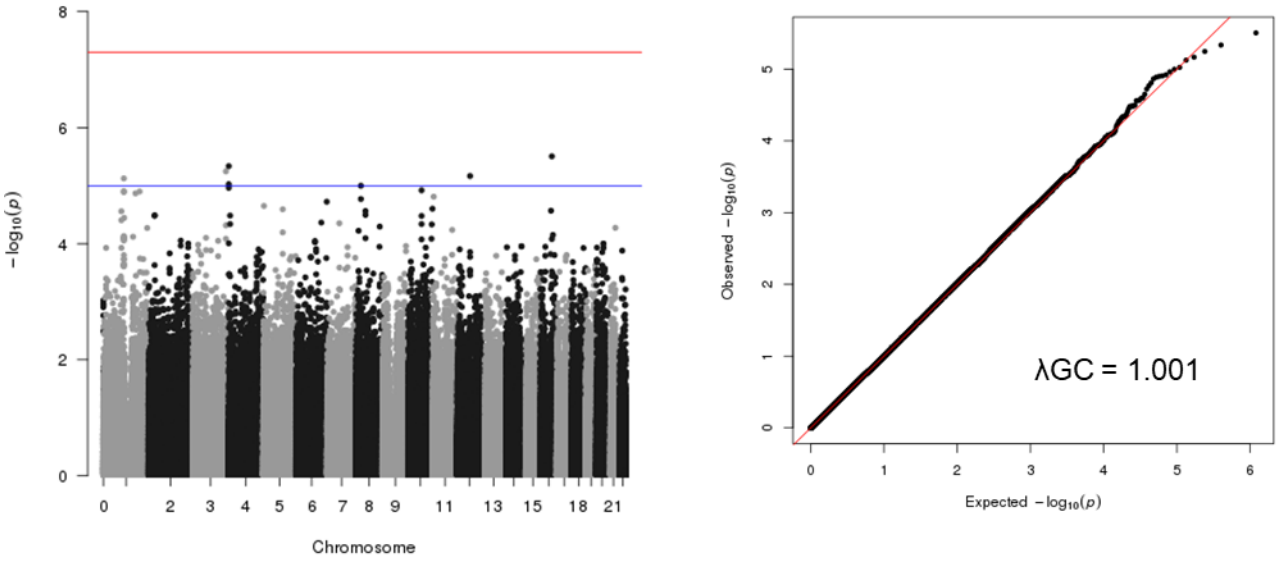


図４：解析集団における GWAS の結果（マンハッタンプロットと qq プロット）。
今回の解析集団では、表現型とゲノムワイド有意な関連を示す一塩基多型は認められなかった。
略語
GWAS（genome-wide association study;全ゲノム関連解析）