厚生労働科学研究費補助金
感覚器障害研究事業

「緑内障診断 SNP チップと変形プロテオミクス クラスター解析による緑内障統合的診断法の開発」に関する研究

平成 20-22 年度　総合研究報告書

研究代表者　木下　茂

平成 23（2011）年　3 月

厚生労働科学研究費補助金
感覚器障害研究事業

「緑内障診断 SNP チップと変形プロテオミクス クラスター解析による緑内障統合的診断法の開発」に関する研究

平成 20－22 年度　総合研究報告書

研究代表者　木下　茂

平成 23（2011）年　3 月

目　　次

［ I ］

20－22 年度　総合研究報告

厚生労働科学研究費補助金（感覚器障害研究事業）
総合研究報告書

「緑内障診断 SNP チップと変形プロテオミクスクラスター解析による緑内障統合的診断法の開発」に関する研究

研究代表者　木下 茂

京都府立医科大学大学院医学研究科 視覚機能再生外科学 教授

研究要旨

緑内障は日本において中途失明を生じる疾患の第１位であるが、その９割で患者本人が無自覚なことが問題となっている。米国の大規模追跡調査により、早期治療で失明確率が低下可能なことは証明されているため、眼科スクリーニング検査等で発症予測診断が可能となれば、国民の健康維持に大きく貢献することになる。従って本研究の最終目標は、簡便な血液検査で緑内障感受性リスク判定健診を行い、発症リスクが高い例については緑内障専門医による追加の精密検査を実施し、その後は必要に応じて直ちに治療を開始するという一連の仕組みを構築することにある。この目標を実現すべく、以下の研究を実施した。

１．ゲノムワイド関連解析(GWAS)による緑内障マーカーSNP 同定

ゲノム情報を絡めた本疾患の発症・予後予測診断技術の開発を目的として、広義の原発開放隅角緑内障（以下 POAG）と健常対照群を含めた 1,575 例の血液検体について、DNA アレイを用いた２段階のゲノムワイド関連解析(GWAS)を行った。それらの結果、染色体上の 3 領域にわたる統計的に有意な 6 つ(①rs547984 、 ②rs540782 、 ③rs693421 、 ④rs2499601 、 ⑤rs7081455 、⑥rs7961953)の POAG 疾患マーカーSNPs が得られ、2009 年 7 月に Proceedings of National Academy of Sciences(PNAS: 2009;106(31):12838-12842)に報告を

行った。なお、①から④までの SNPs は同一の連鎖不平衡ブロックの上に存在している。

２．変形プロテオミクスによる緑内障マーカータンパク質同定

本研究では、POAG のリスク予測への応用を現実的なものとするために、マーカーSNPs による緑内障診断チップ作製と同時に、その診断支援・強化に役立つと考えられるデータを取得すべく、Cytometric Bead Array 法による変形プロテオミクス実験を分担して実行した。125 例について 29 種類のサイトカインの血漿中濃度を同時測定した結果、4 種類で POAG 群において有意差を認めた。

３．信頼性向上のための症例収集

上記の解析と並行して、本学緑内障専門外来で受診中の患者より POAG や落屑緑内障の症例、またボランティア健診事業受診者より正常対照例を募り、各種検査、血液サンプルの収集、問診票の蓄積、臨床データの解析を行った。いずれも十分なインフォームドコンセントとともに書面による同意を得て、種々の臨床機器を用いた緑内障精密検査を行い、緑内障性視神経障害の有無と程度に関する臨床データを合わせて、平成 23 年 3 月末時点で総数は 4,000 例を越えた。また正常対照例でも、視神経の形態によりランク分けを行った。

４．POAG 発症リスク判定に向けての統合的判定方法樹立

POAG 発症リスク判定に向けて、SNP チップにより得られたジェノタイピングデータと、変形プロテオミクス解析により得られた血中サイトカイン濃度データを統合して診断する方法を検討し、実際に開発・試験運用を行った。その結果、7 割以上の診断が可能な方法を得ることができたので、特許出願した。

５．GWAS による落屑緑内障マーカーSNP 同定

さらに落屑緑内障の発症マーカーSNP を同定するために、同疾患を約 200 症例収集して DNA アレイ実験を実施した。

分担研究者
田代 啓　（京都府立医科大学大学院医学研究科 ゲノム医科学 教授）
長崎 生光（京都府立医科大学大学院医学研究科 統計学 教授）
森 和彦　（京都府立医科大学大学院医学研究科 視覚機能再生外科学 講師）

＜1．ゲノムワイド関連解析(GWAS)による緑内障（POAG）マーカーSNP同定
2．変形プロテオミクスによる緑内障マーカータンパク質同定
3．信頼性向上のための症例収集について＞

A. 研究目的

緑内障は日本における中途失明を生じる第1位の疾患であり、眼科スクリーニング検査等で発症予測診断が可能になれば、早期治療により失明を予防することが出来るため、国民の健康維持に大きく貢献することになる。日本人集団を対象に、ゲノムワイド関連解析（Genome Wide Association Study、GWAS）とその確認試験による広義の原発開放隅角緑内障（Primary Open Angle Glaucoma、POAG）マーカーとなる一塩基多型（Single Nucleotide Polymorphisms、SNPs）の同定、及び変形プロテオミクスデータ取得による緑内障血中マーカータンパク質同定を目的とした研究を行った。さらに信頼性向上のために症例を集め、データを継続的に追加収集した。

B. 研究方法

（1）ジェノタイピングデータ

本研究開始前の718例によるGWAS結果に基づいて設計したイルミナ社製カスタムチップのハイブリダイゼーション実験を行った。21年4月以降に分担研究者の森らが新たに検査と採血を行って収集した300例以上のPOAG症例と健常対照例について、細胞株化とゲノムDNA抽出を行うと共に、GWASとは別集団の約857例のゲノムDNAについてカスタムチップハイブリダイゼーション実験を行った。その後、Mantel-Haenszel法にてGWAS結果との統合解析を行った。

（2）サイトカインデータ

変形プロテオミクスの手法では、125 例について 29 種類のサイトカインの血漿中濃度同時測定を、ビーズとフローサイトメトリー装置を用いる Cytometric Bead Array 法にて POAG 血中マーカータンパク質として同定した。

（3）症例収集

本学倫理委員会の承認のもとで、倫理基準を遵守して症例収集を実施した。京都府立医科大学附属病院眼科・緑内障専門外来に通院中の POAG 患者の中から、十分なインフォームドコンセントを行った後に書面による同意を得ることができた患者を本研究に組み入れた。緑内障病型の診断基準は、日本緑内障学会による緑内障診療ガイドライン、日本緑内障学会が主体となり岐阜県多治見市で行われた緑内障疫学調査（多治見スタディ）、およびヨーロッパ緑内障学会の緑内障判定基準に準拠して行った。また、正常コントロール例については当院における緑内障正常ボランティア健診事業受診者の中から選択し、緑内障専門外来におけるものと同等の緑内障精密検査を行い、状態に応じたランク分けも行った。正常ボランティア健診事業における緑内障精密検査内容は、視野検査として FDT スクリーナー（マトリックス、カールツァイスメディテック社)、ハンフリー視野計プログラム SITA fast（カールツァイスメディテック社；視野異常出現時)、レフケラトノンコンタクトトノメーター（ニデック社)、視神経乳頭形状解析検査として HRT-II（ハイデルベルグエンジニアリング社)、GDxVCC（カールツァイスメディテック社)、無散瞳眼底写真（トプコン社）を行った。また2年目より新たに追加検査として前房隅角解析装置の Visante（カールツァイスメディテック社)、眼軸測定として IOL マスター（カールツァイスメディテック社)、網膜神経線維層厚解析としてフーリエドメインの 3 DOCT（トプコン社)の 3 つを検査に組み入れた。全対象者に対して細隙灯顕微鏡による前後眼部検査を施行した。視野異常検出症例ならびに眼圧が 21 mmHg を超えた症例に対しては、2 次検査として緑内障専門医がゴールドマン圧平眼圧計での眼圧測定なら

びに隅角検査を行った。これらの全ての検査結果をもとに、複数の緑内障専門医が独立して緑内障の有無を判定した。さらに視神経乳頭の形状を基に、独自に作成した診断基準により健常対照例のランク分けを行った。さらに本研究に参加した健常対照例およびPOAG患者に対し、臨床的背景因子を探る目的で問診表を記入してもらい、緑内障家族歴、眼既往症、全身合併症、睡眠時間、飲酒・喫煙を含む生活習慣などを調査した。

C. 研究結果

（1）ジェノタイピングデータ

POAG および対照例あわせて 857 例を用いてカスタムチップ実験を行った結果を、GWAS の結果と Mantel-Haenszel 法により統合解析したところ、染色体上の 3 領域にわたる 6 個の POAG マーカーSNPs (①rs547984 、 ②rs540782 、 ③rs693421 、 ④rs2499601 、 ⑤rs7081455、⑥rs7961953)を同定した。これらの結果は PNAS. 2009; 106 (31):12838-12842.に報告した。

（2）サイトカインデータ

125 例について 29 種類のサイトカイン血漿中濃度を変形プロテオミクスの手法により同時測定を試みたところ、11 種類のサイトカインの同時測定に成功し、そのうち 4 種類のサイトカインで POAG に有意差を認め、POAG 血中マーカータンパク質として同定した。

（3）症例収集

23 年 3 月末の時点で、POAG 1439 例、正常コントロール 1713 例、落屑緑内障症約 200 例を含むその他緑内障病型症も含め、合計 4079 例のゲノムサンプルならびに臨床データを収集することができた。

D. 考察

単一施設の統一診断基準で収集されたGWAS対応のケースコントロール集団としては、本邦最大規模の症例数となった。この症例数を元として、POAGの疾患マーカーとなる6 SNPsが同定できた。これらを含めて、カイ2乗検定で比較的高い有意差を示した複数の上位SNPについて、次項に示す統合的アルゴリズムでその有用性を検討することは有意義だと考えられ

る。また、変形プロテオミクスで同定されたマーカータンパク質についても、統合的アルゴリズムでその有用性を別角度から検討することは有意義だと考えられる。

E.　結論

血液検査による緑内障リスク判定に向けた基礎的研究が進展した。

F. 危険情報

当該なし

＜４．　原発開放隅角緑内障発症リスク判定に向けての統合的判定方法樹立について＞

A.　研究目的

ジェノタイピングデータと血中サイトカイン濃度データのそれぞれについて、解析に最適なデータ形式を検討し、統計学的・情報工学的な各種解析を行う事で有用な手法のスクリーニングを実施し、統合的診断アルゴリズム開発の基礎を固める事を、研究の第一の目的とした。また、それらの結果を元にして、最終的に診断アルゴリズムを構築する事にも取り組んだ。

これらの流れを具体的に説明する。まず本研究で扱う２種類のデータのうち、遺伝的情報である 1 塩基多型（Single Nucleotide Polymorphisms、SNPs）データは、各々は 3 種類のジェノタイプ（AA,AB,BB）と欠損値の計４つの値からなる離散値である。本研究のように DNA チップを用いた GWAS 結果を用いる場合、最終的に扱う SNPs 数は数百から数千の規模になる。また解析に使用する症例数も同様の規模となるため、最終的に扱うデータは SNP 数と症例数を乗じた数万から数十万の要素数を持つ離散値行列となり、複雑な特徴空間に分布する事が予想された。

これに対して血中サイトカイン濃度データは、一定の範囲内で連続値の形をとる点、および変形プロテオミクスの手法の性質上一度に計測可能な項目数が数十種類程度、症例数も数百例程度と、データの規模が一回り小さい点でジェノタイピングデータとは大きく異なる。従って、これら２種類のデータを単純に合わせる事は、データの性質、及びサイズの違いから容易

ではない。また、各データを個別に扱う際にも種々の問題が存在するため、両データに何らかの統一的な数値変換手法を施すことで、新たな数値空間に同時に展開するといった手法も困難であると考えられる。

本研究では、まず各データ固有の問題を個々に解消する最適なデータ形式を検討する事、次に既存の統計学的・情報工学的な各種解析手法を多数適用して、各々で一定の判別を可能とする手法の選定を行った。また、ジェノタイプデータに関しては、全データから冗長なものを取り除き、解析に使用するデータサイズを削減する試みも合わせて行った。

この様にして行われた各解析結果を利用することで、別の角度から2種類のデータを総合的に考慮し、最終的に診断精度を向上させるアルゴリズムのプロトタイプを構築、試験的な運用を行った。

B. 研究方法

（1）ジェノタイピングデータ変換

まず、A、T、C、G等の塩基を示す文字情報として表現されるジェノタイピングデータを、一般的な解析に適用しやすく、かつ診断精度の向上へ寄与するような数値情報に変換する方法を検討した。この検討に際しては、解析当時でクオリティチェックをクリアした最大検体数である表.1 の集団数を用いた。なお、「GWAS（Affy500k）」はアフィメトリクス社製DNAチップ「Affy500k」による全ゲノム解析の事であり、「Custom（iSelect）」はイルミナ社製カスタムDNAチップ「iSelect」を用いたGWASの再現性確認解析を意味する。各解析間で、使用した検体に重複は無い。

表.1 検討解析の使用検体数

Stage	緑内障群 (Case)	健常対照群 (Control)	計
GWAS (Affy500k)	411	289	700
Custom (iSelect)	521	519	1040
	932	808	1740

具体的な検討方法としては、次の5種類の数値変換方法を試した。

① 全てのジェノタイプデータを、SNP 毎に単純にアルファベットの若い順に数値化。（例えば

ATの場合、AA → 0、AT → 1、TT → 2として変換する）

② 各SNPについて、全ジェノタイプデータを用いたアレル頻度を計算し、それに従って変換。（例えば、Major Homoは2、Heteroは1、Minor Homoは0）

③ ②のアレル頻度算出を、GWASの健常対照群（Control）のみ用いる様に変更した方法で変換。（全ゲノム解析結果を重視した方針による数値化法）

④ ②のアレル頻度算出を、GWASとCustomの両Controlを用いる様に変更した方法で変換。（Control重視の数値化法）

⑤ ②のアレル頻度算出後、単純に頻度が高いアレルをMajorとしてMajor Homo等の判定を行うのではなく、CaseとControlのアレル頻度を比較して、Caseで高い方をリスクアレルとみなす。すなわち、Risk Allele Homoを2、Heteroを1、Non Risk Allele Homoを0となる様に数値変換を行う。

データ取得実験の過程で発生した欠損値の補正手法についても検討した。欠損値は、本研究で用いているSNPチップを含むDNAアレイチップ実験では、その性質上常に起こり得る問題である。つまり、解析結果への影響の少ない欠損値補正方法を確立する事は、将来的に実際の診断を行う上で克服すべき重要な課題である。具体的には、乱数によるランダム補正の影響調査、及び一定の規則性を与えて補正する方法を検討し、解析結果への影響が最小になるものを選ぶこととした。

（2）ジェノタイピング解析方法

（1）で検討した値を用いて、最も症例診断に有用な手法を検証すべく、各種解析法のスクリーニングを行った。基本的には、本研究プロジェクトが保有するデータを二分して、まず片方を学習データとして緑内障群と健常対照群の各特徴・特性を学習する。なおデータの二分は、GWAS、CustomのStage別に分ける方法を主に用いる。次いで、その学習結果を以って、残る一方をテストデータとして、各検

体に対しどちらの群に所属するのか（すなわち、緑内障に関して陽性もしくは陰性であるか）を予測・診断する。最終的に、各検体の本来属する群と診断結果がどの程度一致しているのかという精度を、感度・特異度・正診率（診断率）等の尺度を用いて評価する。

試行した解析手法は、サポートベクターマシンを詳細に試した他、最終的に主成分分析、自己組織化写像、線形判別分析、非線形判別分析であるMahalanobis距離、決定木、などを試した。また、上記の手法の幾つかに関しては、メタアナリシス手法を用いた追加の解析も行った。

これらの手法を用いて、単純にCase と Control の２つにパターン分類する、もしくはデータを 2 つのStage に分け、GWAS のパターンを機械学習し、Custom を Case と Control に判別する試みを行い、本研究で用いるデータの特性に合う手法の検討を行った。

（３）使用 SNPs の検討

（２）の検討と平行して、解析に使用するSNPsの組合せについての検討も行った。基本的には解析に使用するSNPs が多いほど既知検体に対する診断精度は上がる傾向があるが、過学習の結果、一方で未知検体に対する予測診断精度は下がる傾向にある。従って、ある程度の診断精度を確保しつつも、どの様な検体に対しても共通して診断に用いる事のできる最小限の SNPsセットを把握する事が重要と考えられる。

この様な検証に関しては、基礎的な統計手法である程度まで絞り込んだSNPs に対してそのまま解析を行う方法、及びその様な SNPs から更に組合せ最適化を施して絞り込む方法を行った。

（４）サイトカインデータ変換

本研究で用いる血中サイトカイン濃度データは、BD™ Cytometric Bead Array (CBA) Flex Set System を用いた変形プロテオミクス法により、緑内障患者と健常対照群それぞれで測定した。サイトカインデータには、表.２に示す様に、測定時期や項目、検体数などの違いで 2 つの Stage が存在する。なお、サイトカインの「1st Stage

(1st)」と「2nd Stage(2nd)」の各検体は、両方ともジェノタイプの GWAS 集団内から選んでおり、両 Stage 間に検体の重複は無い。これは、有意なサイトカインを得た場合、全ゲノム上のデータから、該当遺伝子情報を即座に見られる様にするためである。

表.2 サイトカイン解析使用検体数

Stage	緑内障群 (Case)	健常対照群 (Control)	計
1st Stage (29 項目)	42	42	84
2nd Stage (11 項目)	73	53	126
	115	95	210

1st では、有意なサイトカインのスクリーニングの目的で、CBA で同時測定可能な項目をなるべく多く設定し、29 項目を測定した。この 1st での結果を元に、測定でエラーが出たもの、正しく測定できなかったものなどを除き、2nd の実験を行った結果、11 項目の測定に成功した。

この様に 1st と 2nd の間には、使用した検体、測定項目の他、実験日などの複数の要因で異なる点がある。そして、本研究で用いた変形プロテオミクス法の特徴として、その様な測定条件や環境の違いにより、データの取り得る範囲や値の傾向に差が生じ得る事が有る。このため、複数回に分けて測定された血中サイトカイン濃度データを比較する際には注意が必要であるが、その様な場合に統一的にデータを扱うための一般的な手法自体が、未だ存在しない。そこでまず、測定条件の違いによる値の差を緩衝するために、本研究プロジェクト独自のデータ標準化方法を検討し、最も効率良く解析に利用でき、診断性能向上に寄与する手法を採用することにした。

（5）サイトカインデータ解析方法

検討の結果採用した方法で独自の標準化を行い、ジェノタイピングデータの時と同様の主成分分析、線形判別分析、サポートベクターマシン、自己組織化写像などの解析手法を試し、データの特性に合う Case と Control の判別に有用な手法の検討を行った。

またこの過程で、解析に使用するサイトカインの選別も同時に行った。ジェノタイプデータは、元々解析に使用可能な SNP 数自体が多かったために、その組合せについて慎重に配慮する

必要があった。これに対してサイトカインデータでは、元となる 1st・2nd に共通する最大項目数が 11 であるため、考慮すべき組合せ数が少ない。従って、作業の効率化と解析精度向上のため、解析手法のスクリーニングと並行して項目を絞り込む。

（6）統合的診断アルゴリズム

ジェノタイピングデータ、血中サイトカイン濃度データの各解析結果に対して一定の評価基準を設け、統合的に診断する手法を数種類検討した。

なお本研究の一連の解析には、統計ソフトの「R」、及び C 言語で作成したオリジナルプログラムを使用した。

C. 研究成果

（1）ジェノタイピングデータ変換

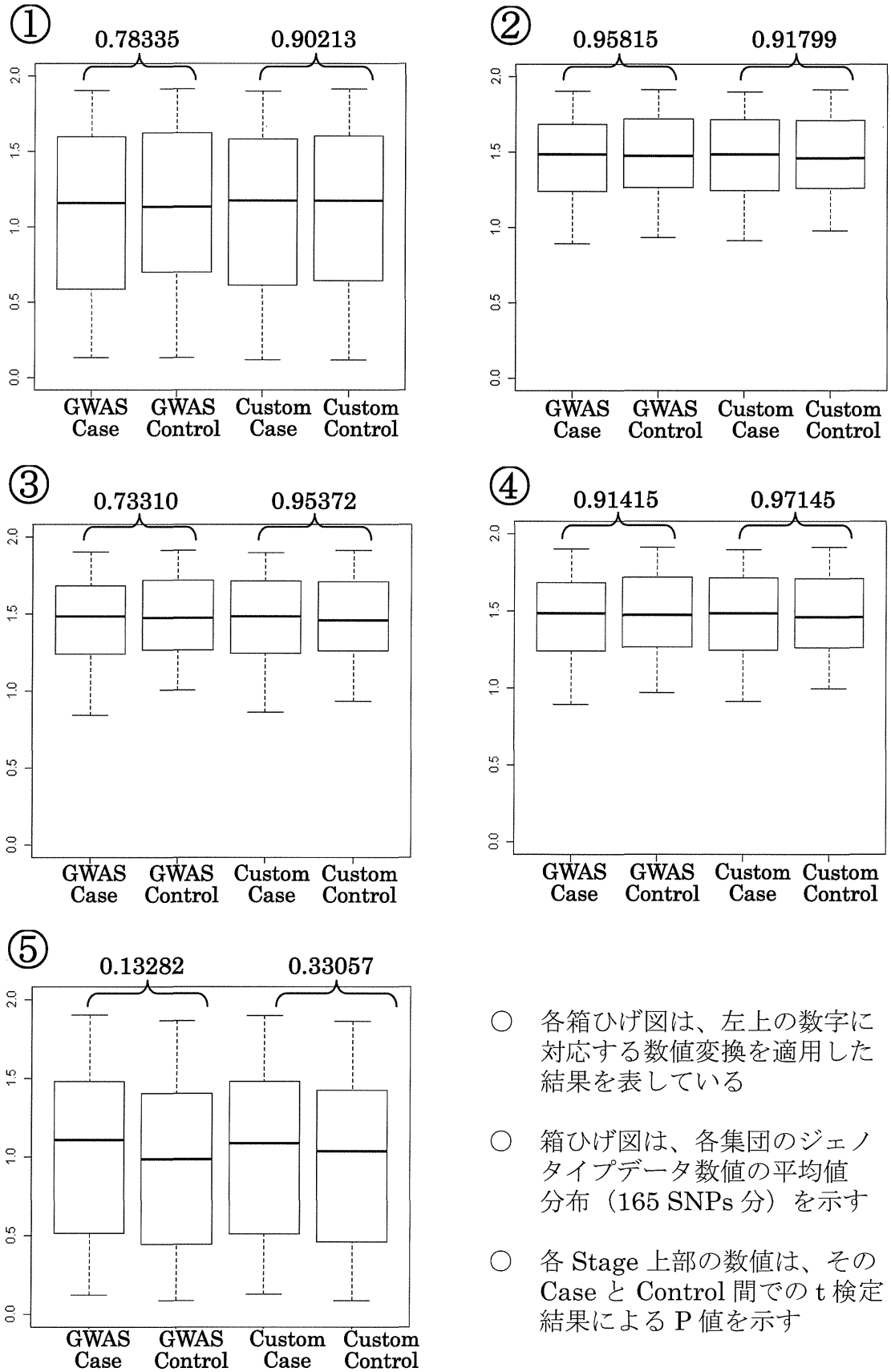
まず、GWAS と Custom に共通する全 SNPs の中から Filtering 等で一部の SNPs を選抜した。多数の選抜パターンを考慮した結果、まず一次的に 172 SNPs を抽出し、そこから更に選抜した 165 SNPs について、計 1,740 検体分のジェノタイピングデータに対し、前述の①～⑤の数値変換方法を実際に適用した（参照 B 研究方法-(1)）。この時点で全データが０、１、２のいずれかになっている。なお、これらの手法中で用いるアレル頻度の計算、及び変換対象のデータからは、欠損値を取り除いている。

次に、GWAS と Custom 各 Stage の Case 集団と Control 集団毎に、数値変換されたジェノタイプデータの平均値を SNP 毎に計算した。そして、それら４集団分の 165 SNPs の平均値について、各集団の分布傾向を見るために箱ひげ図を作成した（図.1）。また、各 Stage で Case と Control 間の Student の t 検定も合わせて実施した。

最終的に、t 検定の P 値が小さいほど、Case と Control 間の差を効果的に表現できる数値変換法であると考えられ、図.1 中に記した検定結果より、⑤の手法が最も P 値が小さかった。従って、これ以降の本研究で用いるジェノタイプデータの数値変換法は、全て⑤の方法を用いることとする。

また、⑤の数値化法を採用した上での欠損値補正方法について、乱数を用いた方法と、ジェノタイプデータを正規化する方法の検討を行った。

図.1 ジェノタイプデータに対する 5 つの数値化パターン結果の比較



○ 各箱ひげ図は、左上の数字に対応する数値変換を適用した結果を表している

○ 箱ひげ図は、各集団のジェノタイプデータ数値の平均値分布（165 SNPs 分）を示す

○ 各 Stage 上部の数値は、その Case と Control 間での t 検定結果による P 値を示す

まず、乱数による欠損値補正方法は、各 SNP のアレル頻度に応じた確率で、0、1、2の各値を割り当てるランダム補正による方法である。これにより補正されたジェノタイプデータは、各 SNP の持つ値の傾向を損なう可能性が低いものの、一方で乱数の取り方によって値が頻繁に変わってしまい得る。この影響を調査するため、1000回乱数を取り直した解析を行い、どの程度解析結果にばらつきが生じるのかを検証した。

解析手法には、与えられたデータから特徴を学習し、別のデータを安定的に 2 群へ判別可能な「線形判別分析（Linear Discriminant Analysis、LDA)」を用いた。学習には、全検体を対象とした場合（① All Learning)、及び GWAS 検体のみを対象とした場合（② GWAS Learn）の 2 種類を用いた。また、学習結果をテストするデータには、GWAS のみ（[A] GWAS Test)、Custom のみ（[B] Custom Test)、及び全検体（[C] All Test）の計 3 種類を用いた。なお、いずれも前述の 165 SNPs を用いた。

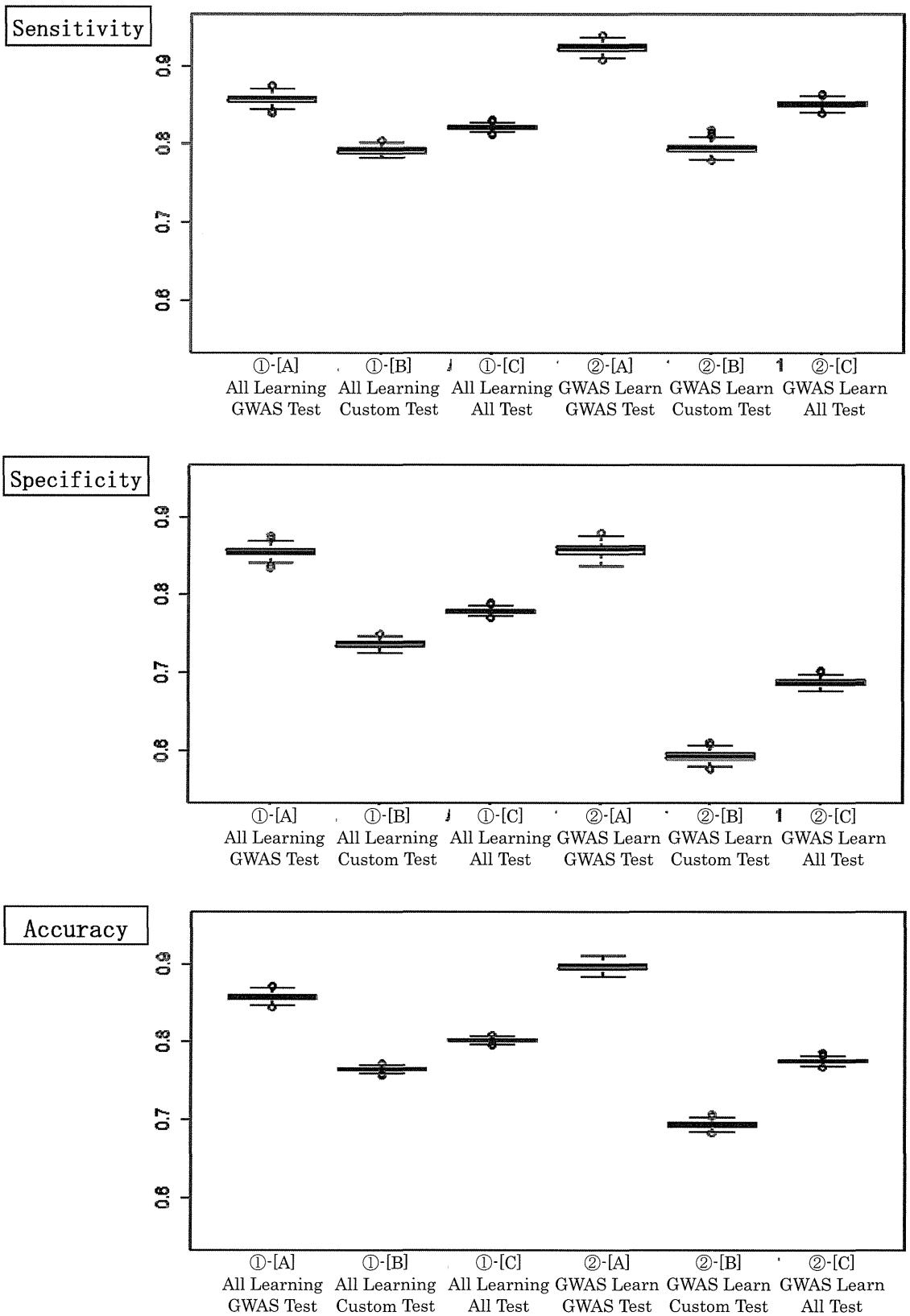
それら合計 6 種類のテストを、乱数を 1000 回変えて繰り返し、結果を感度(Sensitivity)、特異度(Specificity)、診断率（Accuracy）毎に分けて、箱ひげ図に表した（図.2)。この結果より、6種類の何れのテストにおいても、各感度・特異度・診断率の大半が非常に狭い範囲内に分布しており、各最大・最小値間の差は 5%未満に留まっている。従って、飛び抜けた改善や改悪も無かった事から、乱数による欠損値補正が解析に大きく影響を与える事は無いと考えられる。

次に、ジェノタイプデータを正規化する方法の検討を行った。これは、GWAS の分野で人種・集団間の遺伝差異を見る手法として多用されている、「EIGENSTRAT」というソフトウェア中で用いられている方法である。具体的には、式①を用いて、数値化されたジェノタイプデータを正規化して変換する手法である。

$$M(i,j)=\frac{C(i,j)-m(j)}{\sqrt{p(j)(1-p(j))}} \quad (式①)$$

- i は検体、 j は SNP の番号
- $C(i,j)$ は各ジェノタイプデータの数値（0, 1 or 2）

図.2 乱数を用いた欠損値補正の解析結果への影響



・ $m(j)$ はSNP j のジェノタイプデータの平均値
・ $p(j)$ はSNP j のアレル頻度
・ $M(i,j)$ は正規化後の値

なお、$m(j)$、$p(j)$を計算する際には、欠損値を無視した総数で計算を行う。また、正規化後の欠損値は一律で「0.0」とし、これは欠損値を平均値と同値をみなす事を意味している。この欠損値補正法では、乱数の取り方により値が変化する事が無いため、安定的に運用できる。しかし、検体数が少ない場合、算出されるアレル頻度の精度が低くなり、悪影響を及ぼし得る。

この欠損値補正法を用いた場合のジェノタイプデータを検証するために、LDAを用いた解析を行った(表.3)。ここでは、リスクアレル重視の数値変換方法⑤（参照 B研究方法-(1)）で数値化した後、乱数による欠損値補正を行ったデータと解析結果を比較した。この結果より、両手法間には大差が無いと考えられるが、正規化法を用いた方が安定的に欠損値を補正できるので、本研究では正規化法を用いる事とする。

（2）ジェノタイピング解析方法

（1）で決定した手法を用いて、数値化・正規化・欠損値補正を行ったジェノタイプデータに対して、各種解析法のスクリーニングを行った。

まず、複数の変数に対して、その全体の特性を求める手法である「主成分分析（Principal Component Analysis、PCA)」を用いて解析を行った。PCAのジェノタイプデータを用いた応用例としては、遺伝学の分野における集団構造化の評価などで用いられている。ここでは165 SNPsを用いた場合、GWAS・Custom 両 Stageの緑内障群と健常対照群との分布がどの程度異なっているのかを視覚的に表した。（図.3）

図.3 ジェノタイプデータのPCA結果

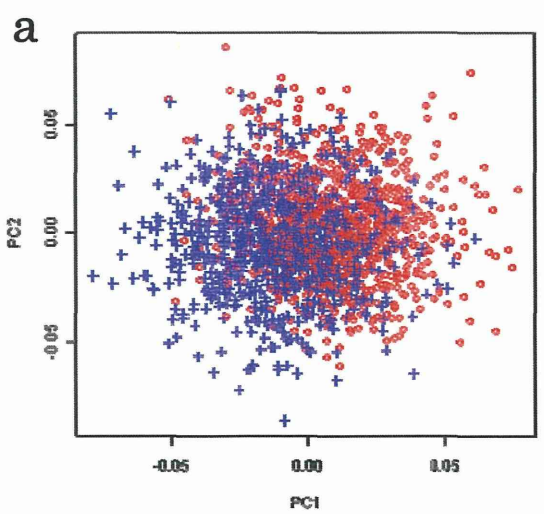


表.3 欠損値補正パターンの違いによる解析結果への影響

	Stage	集団	全検体による学習結果				GWAS集団のみによる学習結果			
			+	−	Sensitivity / Specificity	Accuracy	+	−	Sensitivity / Specificity	Accuracy
A	GWAS	Case	352	59	85.64%	86.00%	382	29	92.94%	89.71%
		Control	39	250	86.51%		43	246	85.12%	
	Custom	Case	416	105	79.85%	76.83%	412	109	79.08%	69.52%
		Control	136	383	73.80%		208	311	59.92%	
	All	Case	768	164	82.40%	80.52%	794	138	85.19%	77.64%
		Control	175	633	78.34%		251	557	68.94%	
B	GWAS	Case	356	55	86.62%	86.29%	378	33	91.97%	89.29%
		Control	41	248	85.81%		42	247	85.47%	
	Custom	Case	410	111	78.69%	76.35%	413	108	79.27%	69.04%
		Control	135	384	73.99%		214	305	58.77%	
	All	Case	766	166	82.19%	80.34%	791	141	84.87%	77.18%
		Control	176	632	78.22%		256	552	68.32%	

- いずれも、選抜した 165 SNPs を用いた場合の解析結果である
- A はリスクアレルを考慮したジェノタイプデータ数値化法（変換方法⑤）を適用した後、乱数による欠損値補正した場合の LDA による解析結果
- B は A と同様に数値化したジェノタイプデータを、式①による正規化を施すことで欠損値補正を行った場合の LDA による解析結果
- 各 Stage・集団の検体に対して、LDA による解析の結果、陽性（緑内障）と判断されたものを「＋」、陰性（健常者）と判断されたものを「－」として、それぞれの合計値を表に記載
- なおAの結果は、ある１つの乱数パターンで欠損値を補正した時のものであるが、図.２より、複数乱数パターンを変えても結果に大差が無いと考えられるため、代表的なものを１つ掲載した
- ただし、例え結果に及ぼす影響が少ないとしても、取り得る乱数パターンにより差が生じるのであれば、欠損値補正法としては不安定であるので、この点では常に安定して一意的な補正を可能とするBの方が優れている