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 22 年度　総括･分担研究年度終了報告書

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 22 年度　総括･分担研究年度終了報告書

研究代表者　木下　茂

平成 23（2011）年　3 月

目　　次

［ Ⅰ ］

総括・分担研究報告

厚生労働科学研究費補助金（感覚器障害研究事業）
総括・分担研究報告書

「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

緑内障は適切な治療により進行を遅延させることができる疾患である一方で、９割の例で本人が気づいていないことが問題点である。本研究の最終目標は、簡便な血液検査で疾患感受性リスク判定健診を行い、緑内障発症が高リスクな例については緑内障専門医による精密検査を行い、その後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。血液検査の具体的内容としては、緑内障診断 SNP チップと変形プロテオミクスを統合して診断精度の向上を図るため、独自アルゴリズムを構築しての判定を試みている。その実現のために、以下の研究を実施した。

１．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

原発開放隅角緑内障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最適なデータ形式で扱う方法、各データに適用可能な解析手法の選定、および各データや解析を効果的に組み合わせる方法の検討を行い、独自の統合診断アルゴリズムの構築を試みた。

２．落屑緑内障の発症マーカーSNP を同定する研究

他疾患に対する統合診断アルゴリズムの応用を視野に入れ、落屑緑内障の研究を進め、同疾患の発症マーカーSNP を同定する目的で約 200 症例を収集し、アフィメトリクス社 1000K チップ実験により基盤となるデータを得た。

以上の結果より、統合的診断アルゴリズムを用いた原発開放隅角緑内障発症リスク判定の試験的な運用を行った結果、7 割以上の診断が可能な方法を得ることができたので、特許を出願するに至った。

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. 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

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チップを用いた全ゲノム関連解析（Genome Wide Association Study、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事とした。

（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んだ。

（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。

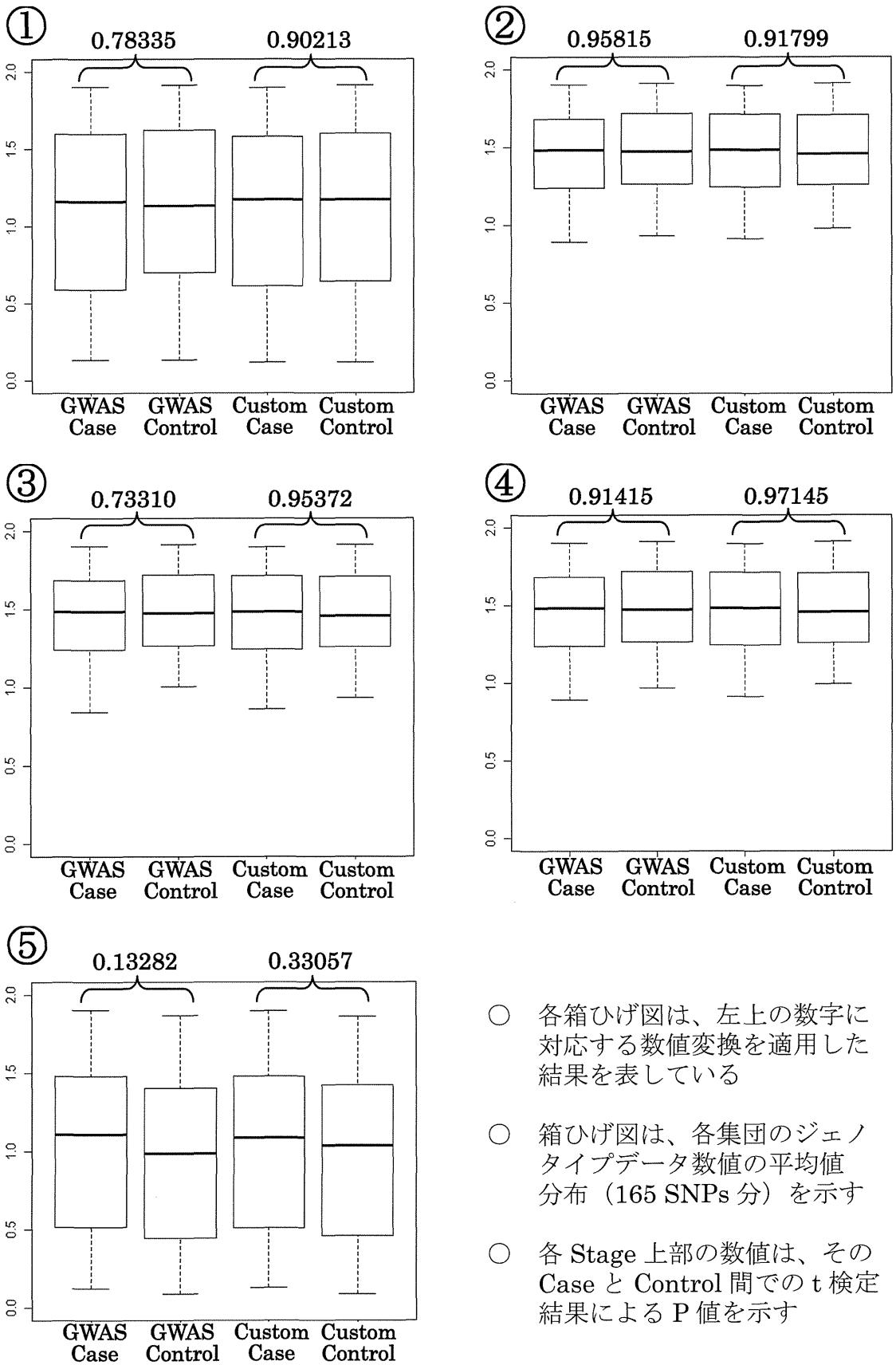
次に、GWAS と Custom 各 Stage の Case 集団と Control 集団毎に、数値変換されたジェノタイプデータの平均値を SNP 毎に計算した。そして、それら４集団分の 165 SNPs の平均値について、各集団の分布傾向を見るために箱ひげ図を作成した（図.1）。また、各 Stage で Case と Control 間の Student の t 検定も合わせて実施した。

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。

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

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

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

図.2 乱数を用いた欠損値補正の解析結果への影響

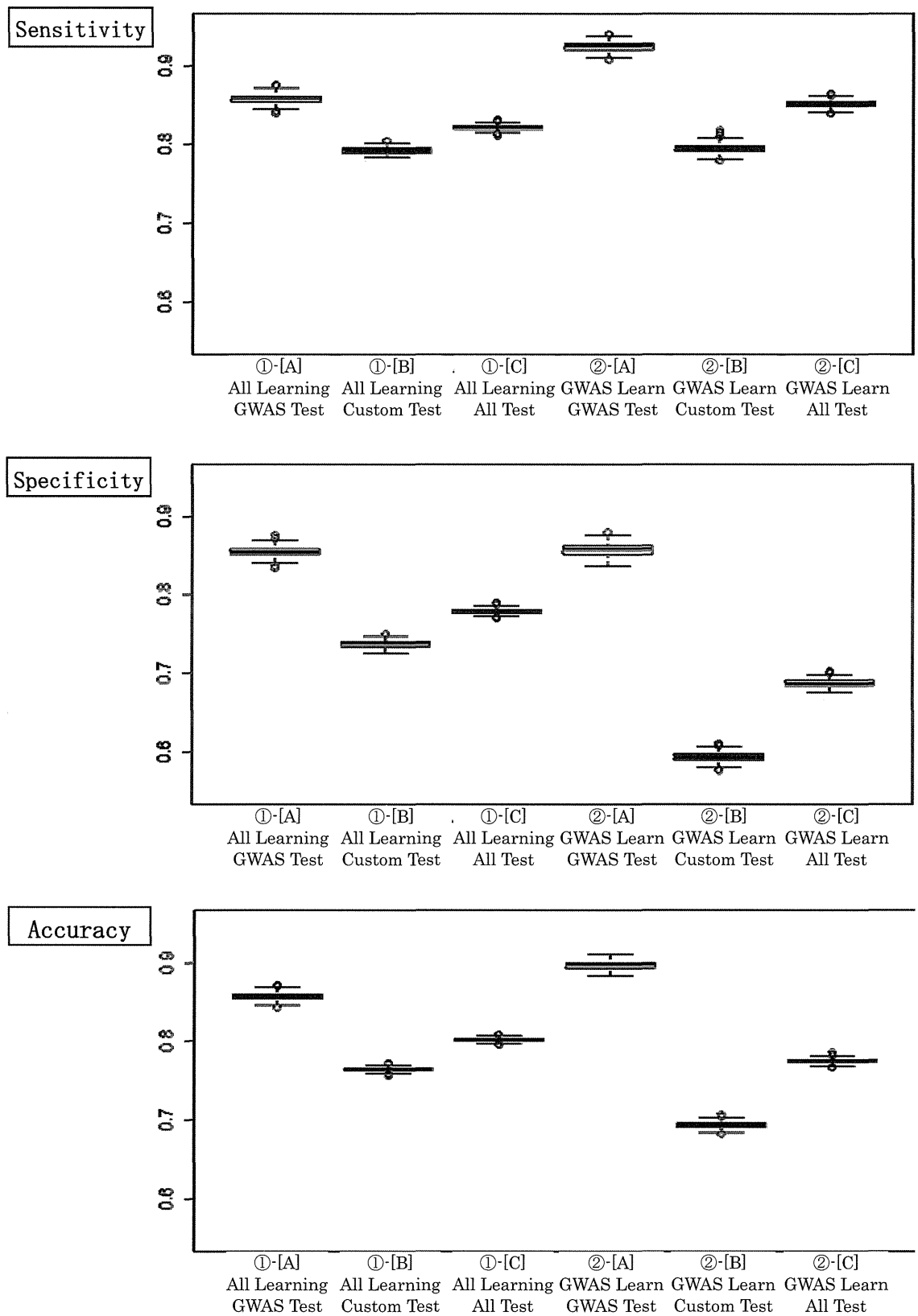


表.3 欠損値補正パターンの違いによる解析結果への影響

	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

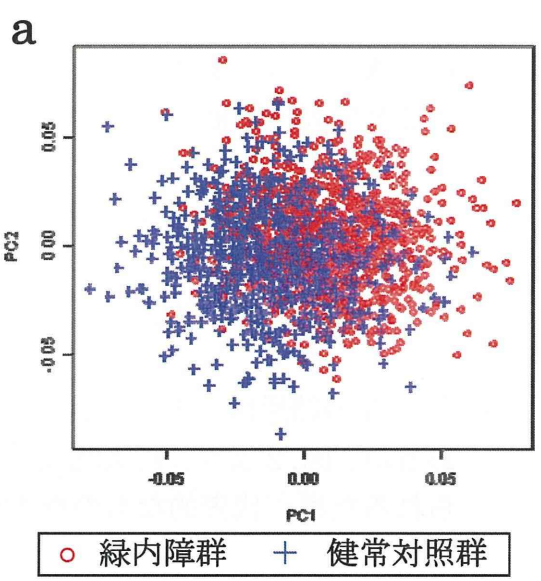
なお、$m(j)$、$p(j)$を計算する際には、欠損値を無視した総数で計算を行う。また、正規化後の欠損値は一律で「0.0」とし、これは欠損値を平均値と同値をみなす事を意味している。この欠損値補正法では、乱数の取り方により値が変化する事が無いため、安定的に運用できる。しかし、検体数が少ない場合、算出されるアレル頻度の精度が低くなり、悪影響を及ぼし得る。この欠損値補正法を用いた場合のジェノタイプデータを検証するために、LDA を用いた解析を行った（表.3)。ここでは、リスクアレル重視の数値変換方法⑤（参照 B 研究方法(1)）で数値化した後、乱数による欠損値補正を行ったデータと解析結果を比較した。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



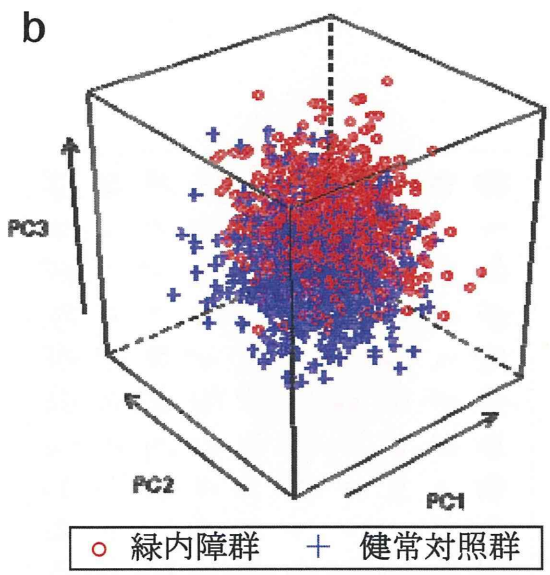


図.3 a は、第 1 主成分（PC1）と第 2 主成分（PC2）を 2 次元にプロットした図であり、同 b はそれに第 3 主成分（PC3）を加えた 3 次元プロットの図である。

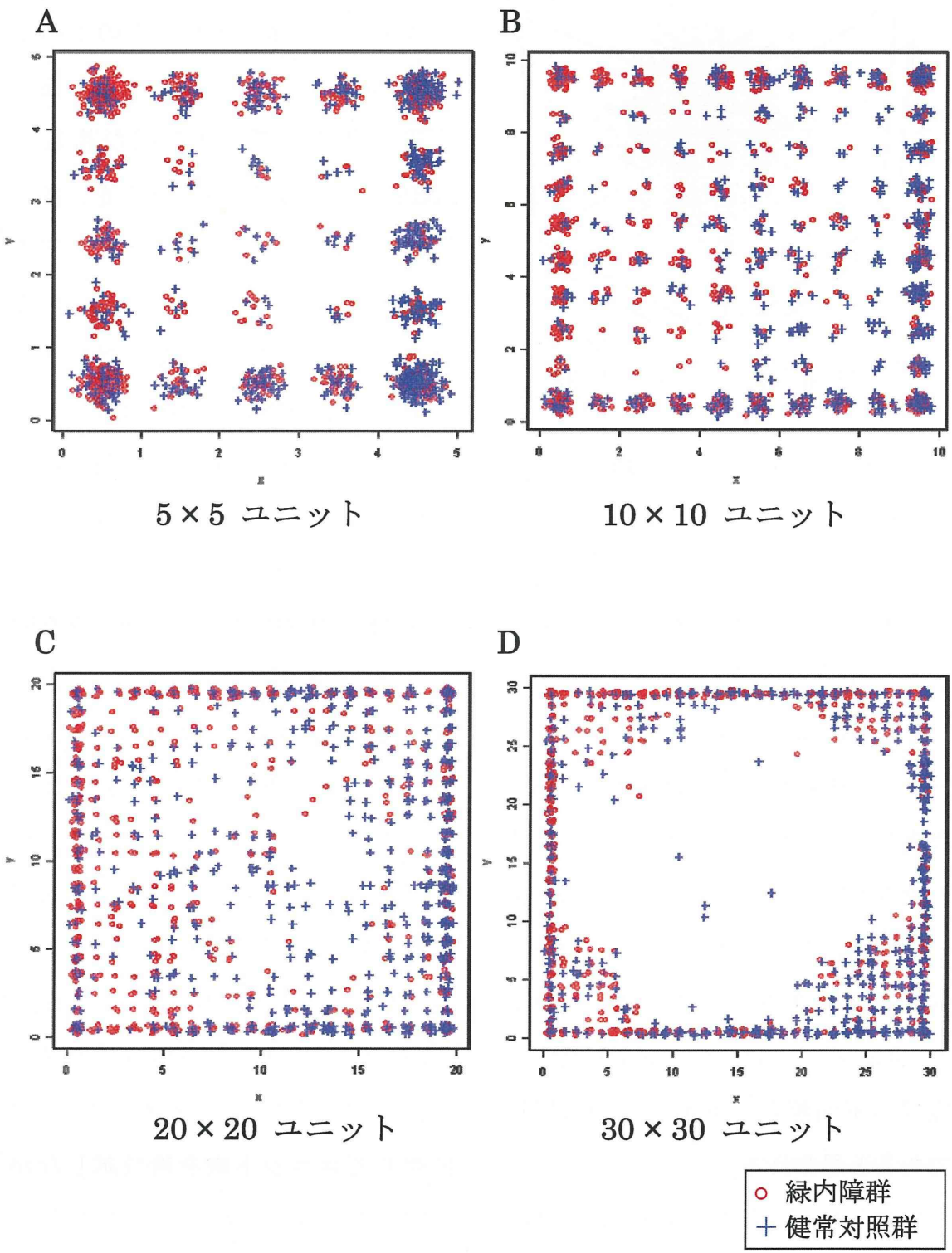
これらの結果より、ジェノタイプデータの数値特性としては、2 群はある程度分かれて分布しているものの、その境界は曖昧であり、単一の集団と化していると考えられる。また、数値特性を表す尺度である主成分についても、用いる主成分の数を増やした所で境界の不明瞭さを解消するのは困難であると思われる。

次に、「自己組織化写像（Self Organization Map、SOM）」という手法を試した。これは、入力されたデータ自らの特徴と互いの相互作用のみによって、任意の次元へと写像させる人工ニューラルネットワークの 1 つである。例えば、本プロジェクトで扱うジェノタイプデータの様な複数の SNPs からなる多次元データを、各検体の特徴を反映した形で 2 次元空間に写像する事が可能である。つまりこの手法を用いれば、PCA では捉えきれなかった各 SNPs 間の相互作用を反映させつつ、緑内障群と健常対照群を効果的かつ簡易的に分類する事が期待される。

SOM の出力は離散的なデータ空間となり、主に格子状に“ユニット”と呼ばれる点を配置し、各ユニットが入力データの特徴を学習する人工ニューラルネットワークと直結された形で実装される。つまり、ユニット数が学習の特性に影響するために注意する必要があり、複数のパターンを試して総合的に検証した方が良いと考えられる。実際に、ジェノタイプデータに対してユニット数を複数試しながら解析を行った結果例を図. 4 に示す。

これら 4 つの結果は、ユニットを正方格子状に配置して、一辺の数を増加

図. 4　SOM を用いたジェノタイプデータの解析結果



させながら解析結果がどう変化するかという様子を見たものである。全体的にユニット数が増加しても、左側に赤で示された緑内障群、右側に青で示された健常対照群が寄る傾向があるものの、互いに上手く分離できていない検体が多い事も解る。また、SOMでは上手く分離できるデータの場合、基本的に似た特性のデータ同士が集まる島状の固まりが複数できるが、このデータでは全体的に外縁周辺に連なって分布している。これは、SOMの解析能力では、緑内障群と健常対照群を明確に分離できていないために起きたものと考えられる。ただし、SOM は分類精度が良くは無いものの、PCA の様に比較的単一に近いまとまりにはなっておらず、両群間の差に関する何らかの特徴を捉えつつある事が示唆される。

これらPCA・SOMによる結果を踏まえ、単純な統計解析手法による2群の判別よりも、情報工学的・人工知能的手法により、積極的に2群を分類する手法を用いた方が良いと考えられる。特に、先のデータ変換法の検討時におけるLDA結果（表.3）より、GWAS集団のみの学習によるCustom集団の診断結果が7割近くに達している事から、これを上回る結果を模索する事を目標とする。

また、GWAS 集団と Custom 集団は、実験に用いた SNP チップの製造元や形式が異なっている。つまり、この両者間の学習と診断で一定以上の効果を上げる事が出来れば、その様な差異によるバイアスに対してある程度堅固な性能を持つ手法と考えられる。これは、より複雑な環境要因の影響が懸念される医療現場で実際に運用される場合、各測定・検査の過程で様々なバイアスが混入し得る事を想定すると、重要視すべきことである。従って、これ以降の検討の基本方針は、GWAS 集団の特徴を学習した後、Custom集団を診断した場合の精度を重点的に評価することとする。

先に述べた様に、表.3にてまとめたLDA の結果では、予測診断精度は 7割程度であった。これは、対象となるデータが、線形の判別面によって2分される事を前提とした手法である。しかし、先のSOM等の結果を見る限りは、線形よりも非線形の判別面の方が

ジェノタイプデータの解析手法に沿っている可能性がある。そこで、非線形判別分析手法を試してみた。

具体的には、非線形判別分析の代表的な手法である Mahalanobis 距離を用いた解析を行った。これは、判別したい各群の群内分散と、それらの間の群間分散の両方を加味して各群の中心を規定し、各検体のデータについて中心からの距離を算出する方法である。この距離の事を Mahalanobis 距離という。算出方法は、各データの行列変換から求められ、本プロジェクトの場合は165 SNPsのジェノタイプデータがこれに相当する。つまり、緑内障群と健常対照群の各特徴を最も表していると考えられる部分から、各検体がどの程度離れているのかを算出し、近い方の群をその検体の所属する群と考える。この時、群間分散・郡内分散の算出に用いるデータを GWAS 集団のみに限れば、Custom 集団の各検体の距離を算出できるので、結果的に学習と予測診断という流れになる。

本プロジェクトのジェノタイプデータに対して、Mahalanobis 距離を算出した結果を図.5に示す。まず、図中Aより、全検体を用いた学習とテスト結果では、GWAS・Custom 共にほぼ100%の予測診断率を達成している。特に図を見れば、赤で示した緑内障群と青で示した健常対照群とが、ほぼ対角線上で二分されている事が解る。この結果より、ジェノタイプデータの特徴は Mahalanobis 距離で判別可能なものである事が解る。しかし一方で、現実的な診断を模した予測診断実験結果であるBでは、学習検体の判別は100%出来ているのに対して、テストデータではSpecificityがほぼ0%に近くなっている。これは、学習が非常に緑内障群の特徴を捉える方向で進んでしまったため、本来健常対照群である検体のほとんどを陽性と誤判定してしまった事による。以上をまとめると、Aより非線形判別分析による精度向上の可能性は高まったものの、一方で Mahalanobis 距離を用いた方法はジェノタイプデータの解析に合致しない事が解った。

次に別の非線形判別的なアプローチを可能とする「サポートベクターマシン（Support Vector Machine、SVM)」を試した。この手法は、ジェ