日本統計学会誌
第38巻，第2号，2009年3月
189頁～198頁

全ゲノム関連解析（GWAS）の統計的手法

鎌 谷 直 之*

Statistical methods in genome-wide association study

Naoyuki Kamatani*

ライフサイエンスの分野で近年，膨大なデータが得られるようになった．2003年にヒトゲノムの全配列が発表され，SNP（single-nucleotide polymorphism）などの多型データの収集が始まった．一人当たり数十万のSNP遺伝子型データが数千人から収集されるようになり，それを用いた遺伝的関連解析が行われている（GWAS: genome-wide association study という）．GWASにはタイピングデータの品質管理，多重比較問題，集団の構造化問題など困難な問題も存在し，それらの解決のために統計的手法が必要である．遺伝的データの特殊性はそれが生み出された過程の確率の安定性にある．そのため，検定，推定や予測の結果が信頼できる．GWASで多くの遺伝的原因がわかったのちはそれらを用いた表現型予測が重要になると思われる．

In the field of life science, massive high throughput data are being accumulated. Polymorphism data such as SNP (single-nucleotide polymorphism) began to be collected, after human whole-genome sequences were released in 2003. Genome-wide association study (GWAS) has been conducted based on a few thousands of human samples with hundreds of thousands of SNP genotype data per person. GWAS has some difficulties, such as quality control of typing data, multiple testing problems, and population structure problems. In order to meet these problems, statistical methods are being required. Peculiarity of genetic data lies in the stability of probabilities in the process of producing these data. Therefore, statistical testing, inference and prediction can be reliable. Once GWAS detects the genetic causes, phenotype prediction based on them would become even more significant.

Key Words and Phrases: SNP, genome-wide association study

1 ライフサイエンスにおけるデータの洪水

ここ10年程の間にライフサイエンスの分野でデータの洪水が押し寄せて来た．更に，最近数年でデータの規模が加速的に増えている．きっかけは1990年に米国で立てられたヒトゲノム計画である．これは約30億個存在すると考えられたATCGの4つの文字で書かれたヒトゲノムの全配列をすべて解明するという壮大な計画であった．それにはゲノム配列をできるだけ速く読む機械（シークエンサー）の開発と，得られた配列データをコンピュータを用いて整理し固定していく技術（アセンブリー）の開発が重要であった．

ヒトゲノムのドラフトは2000年にクリントン大統領とブレア首相により解明が宣言され，ワトソン・クリックのDNAの二重らせん構造の発表から50年後の2003年にヒトゲノムの全解読が宣言された．その後，ヒト以外の動物，植物のゲノム配列解明がなされた．また，個人

* 理化学研究所ゲノム医科学研究センター：〒108-8639 東京都港区白金台4-6-1 東京大学医科学研究所内ヒトゲノム解析センター3階

個人のゲノム配列には大きな違いがあり，それが人種の違いや容貌，性格の違い，更には病気の原因になったり，薬に対する効果や副作用に関係していることがわかってきた．約30億個のゲノム配列の個人間の違い（本当は配偶子間の違いと考えるほうが正しい）を多型（polymorphism）という．多型の種類はいくつかあるが，中でも一つの塩基の置換が最も多く，これをSNP（single-nucleotide polymorphism）という．

全部で数百万から一千万個程度のSNPが存在することがわかったため，3人種（白人，黒人，東洋人）を対象としてSNPの頻度やハプロタイプの構造を解明するためのHapMapプロジェクトが始まった（日本人は45人のサンプルを提供）．これは3人種の基本的な多型を解明するための国際プロジェクトである．日本は一施設としては世界最大のタイピングデータ（全データの24.3%）を供給した（理化学研究所遺伝子多型研究センター，センター長，中村祐輔氏）．

HapMapプロジェクトの開始以前から極めて多数（数万-10数万）のSNPの遺伝子型を解明する技術が発達してきた．特に，理化学研究所遺伝子多型研究センターが改良したインベーダー法は高い精度を示し，初期の研究をリードする役割を果たした．この技術を用いて患者群と対照群の遺伝子型やアレルの頻度を比較し，疾患に関連するSNPを発見する遺伝的関連研究が始まった．この研究により理化学研究所遺伝子多型研究センターは2002年から心筋梗塞や関節リウマチに関連した遺伝子を発表してきたが，2007年になって米国のイルミナ社の開発したビーズ法，アフィメトリックス社のチップ法が世界中で商用化され，海外でも大規模な遺伝的関連解析の結果が発表されるようになった．このような研究は，全ゲノム関連解析（Genome-wide association study, GWAS）と呼ばれるようになった．2007年のScience誌が選んだScientific breakthrough of the yearは“Human genetic variation（ヒトの遺伝的多様性）”であった（iPS細胞は次点，2006年のbreakthroughはポアンカレ予想の証明）が，これはひとえにこのGWASの成功によるものである．病気の原因を解明するという作業は，医学の研究において最も重要な分野であるが，おそらく数年以内にGWASにより頻度の高い重要な病気の遺伝的原因の多くが明らかになるであろうと考えられている．GWASのためには一つの疾患で数千人単位のデータが得られるが，一人から得られるSNPの数は数十万である．このようなデータの洪水が今起きている．

しかもデータの増大速度は加速するばかりである．数年後には個人の30億個のゲノム配列を極めて短い時間で読む技術が実用化される見込みである．そうなると1,000ドルで個人のゲノムすべてを読むことができるという．一人当たり30億のデータポイントを持つデータが数千人，数万人から供給されるという事態に我々は対処しなければならない．そうなると，最も重要になるのは情報学，統計学であることは議論の余地が無い．

データの洪水はゲノム以外のライフサイエンスの分野でも起きている．ゲノムということばは，もともと遺伝子（gene）と染色体（chromosome）を結合させた造語である（ウインクラー）．しかし，今では個人や種の遺伝的全データの一セットのことをさす．これを転用し，mRNA，蛋白質，生体内小分子の個人ごとの一セットを取り扱う学問分野をオミックス（-omics）研究というようになった．Transcriptome（mRNA），Proteome（蛋白質），Metabolome（生体内小分子），あるいはtranscriptomics，proteomics，metabolomicsという言葉も用いられる．これらの分野からも膨大なデータが供給されると予想されている．既にmRNAの発現を網羅的に解析するtranscriptomeの技術は実用化され，海外では病気の予後判定などにも用いられるようになった．

2 遺伝的データの特殊性

遺伝的データは特殊である．金融や気象，あるいは心理のデータと異なった取り扱いが必要である．それが，統計学者の遺伝統計学分野への参入の壁を高くしている．しかし，もともと統計学は遺伝学と歴史的に深い関係がある．Galton，Pearson，Fisher などの初期の統計学者はすべて遺伝学の大家でもあった．著者は，統計学が遺伝学とのかかわりの中でどのような歴史を経て英国で成立したかを知ることは，データをどのような哲学の下で解析するかを考える上でも重要であると考えている．例えば Galton，Pearson の記述統計学と Fisher の推計統計学の違いは，メンデルの法則を信じるか信じないかが大きな影響を与えている．

遺伝的データの特徴は，それを生み出す確率の安定性にある．そのような安定な確率を記載したものがメンデルの法則を初めとした遺伝継承法則（laws of inheritance）である．確率が安定とは，「大数の法則」「重複対数の法則」「中心極限定理」などが現実に成立するという意味で安定だということである．それに比較して金融や気象のデータを生み出すシステムにおいては確率が安定していない．統計的解析は，それらの法則が成り立つことを前提にしているので，一般に遺伝的データをもとにした統計的検定，推定，予測は信頼できる．生物種は確実な過程の下では存続できない．長い時間を考えると，環境は常に変化し続けるからである．不確実な過程の中で自己の種の存続を図らざるを得ない．そして，安定した確率を基礎にし，信頼できる予測が可能なシステムこそ長期存続の条件なのである．安定した確率は，生物が進化により獲得したものである．我々もその一員である生命の根源に，このような安定した確率が存在することは驚きでもある．

3 全ゲノムから疾患関連座位を抽出する手法

形質関連座位を全ゲノムの中から探索する手法の代表は連鎖解析である．Fisher が最初に考えた連鎖解析は，統計的に遺伝子の染色体上の位置を解明する方法である．連鎖解析では家系情報と表現型情報，およびゲノム多型情報を利用して，遺伝継承法則に当てはめ形質に連鎖する座位を探す．この方法により家系が得られる人間のほとんどの遺伝病の原因は解明された．これに対し，関連解析は主として家系情報がない場合に集団の個人の表現型とゲノム情報のみで同様の座位を探す．連鎖解析には家系が必要であるという制限の他に，マーカーの数が少なくてすむ（300-500 程度），陽性領域の範囲が広い（10 cM $\simeq$ 1 Mb 程度）などの特徴がある．

メンデル型遺伝病に対しては連鎖解析は極めて有効な方法であるが，複雑な形質（多因子形質）については有効性が限られていることが知られている．その理由は検出力が低いことと，陽性領域が広すぎて更に絞り込むことがしばしば困難であることである．

関連解析はそれらの欠点の一部を補完する．関連解析にも研究デザインなどの違いによりコホート研究，症例・対照研究などがあるが（鎌谷（2007）），本稿では症例・対照研究についてのみ解説する．これは症例の集団と，対照（コントロール）の集団を別々に集め，その二つの集団の間でアレル頻度，あるいは遺伝子型頻度が異なるかどうかを検討する方法である．それにより，そのアレル，あるいは遺伝子が形質と関連するかどうか，および関係あるとすると関連の強さはどの程度か（効果サイズ）を検討できる．

4 連鎖不平衡構造とタグ SNP

関連解析には候補遺伝子アプローチと網羅的アプローチがある．網羅的アプローチでは直接の原因座位を標的にすることは困難である．SNP の数は 1 千万個以上あると考えられるので，限られた数の SNP の検査により直接の原因を探索することは不可能と考えられる．しかし，

直接形質と関連している座位だけではなく，染色体上でその近傍にあるマーカー座位においても形質との関連が存在することが多い．このような考えにより理研で世界最初に開始された網羅的アプローチによる関連解析は（Ozaki et al. (2002), Suzuki et al. (2003)），今や，大規模タイピングシステムの商品化により世界的に広がっている（Wellcome Trust Case Control Consortium (2007)）．直接関連する座位の近傍のマーカー座位も疾患に関連する理由は，直接関連する座位とマーカー座位の間で，連鎖不平衡が存在するからである．

連鎖不平衡とは集団において（正確には配偶子について）2つの連鎖する座位のアレルの存在が独立ではないことを言う（鎌谷（2007））．2つの座位のアレルの存在が独立である場合は連鎖平衡という．一つの座位が形質と直接と関係し，連鎖するもう一つのマーカー座位は直接その形質とは関連しないとする．集団における特定の疾患の原因変異を起こした突然変異が別々に起きたものであれば，マーカー座位は疾患と関連しないはずである．しかし，特定の集団においてその疾患の原因突然変異が同じ突然変異に由来するものならば，マーカー座位も疾患と関連する可能性がある．疾患に直接関連する座位（遺伝子の変異が直接の原因で疾患を来たす場合）とマーカー座位の間で，アレルの存在の非独立性（連鎖不平衡）が存在するからである．多くの集団において，ありふれた疾患の原因突然変異は共通の起源を持つという仮説が，common disease common variant 仮説である（鎌谷（2007））．これまで得られた証拠によると，確かにこの仮説が成り立つ場合は比較的多いようである．

Common disease common variant 仮説が正しければ，網羅的関連解析により疾患原因座位が特定できる可能性が高い．即ち，用いるマーカー座位が疾患と直接関係しなくても，直接関係する座位と連鎖不平衡の関係にあれば，関連解析でそれを検出できる．

問題はどのようなマーカーのセットを用いれば効率的に全ゲノムをカバーできるかである．一般にゲノム上には連鎖不平衡の強い領域と比較的弱い領域がある．強い領域は数 kb–数 10kb（場合によっては 100kb 以上）の範囲で連続しており，これを連鎖不平衡ブロック（あるいはハプロタイプブロック）という．全ゲノムは 60–70%の連鎖不平衡ブロックとそれ以外の領域に分かれる．連鎖不平衡ブロックの領域は組み換え割合が低く，それ以外は高いことがこのような連鎖不平衡ブロック構造ができる理由のようである．連鎖不平衡ブロック構造は SNP ペア間の連鎖不平衡の指標（r^2，または D'）などをもとに統計的に決められる（鎌谷（2007））．連鎖不平衡ブロック内の SNP は比較的少数の SNP により代表され，これをタグ SNP（tagging SNP）と呼ぶ（Clark et al. (1998)）．このタグ SNP を選択する手法，それらのタグ SNP で全染色体の何%をカバーできるかの推定などに統計的手法が用いられる．

5 多型データの品質管理（QC）

遺伝的関連解析の基本データは個人の遺伝子型データである．従って，遺伝子型が正しく判定されることは極めて重要である．しかし膨大な数の遺伝子型タイピングを行えば，ある小さな割合のタイピングエラーは必ず出る．従って，単一の座位のみを検討する場合はもちろん，多数の座位の検討では膨大なデータが得られるため品質管理（QC: quality control）が欠かせない．実際のところ，全ゲノム関連解析とはタイピングエラーとの格闘である．更に，これはタイピングエラーとは言えないが，欠失アレルや CNV（copy number variation）があると，この座位については通常の SNP 座位に関する法則を当てはめることはできない．従って，欠失アレルや CNV の存在もできるだけ検出して，通常の SNP を用いた関連解析から除外することが望ましい．一連の手法を用いて品質の高い SNP のみを採用するが，その選択に用いられる一連の手法を品質管理フィルターという．

遺伝子型測定のデータ管理については特有の手法がある．それは，遺伝的データについては

遺伝継承法則が適用されるからである．

同一集団から得られた遺伝子型データについては，濃厚な近親婚，分集団化，最近の複数の分集団の混合など，特殊な場合を除き Hardy-Weinberg の法則が適用される．Hardy-Weinberg 法則に従わない遺伝子型データはタイピングエラーの結果である可能性がある．この手法はタイピングエラーの検出法として極めて有効であるが，また欠失や CNV を検出する方法としても有効である（これらを更に敏感に検出するソフトウェアとして PennCNV がある）(Wang et al. (2007))．多数の検体の遺伝子型データが Hardy-Weinberg 法則に従っているかどうかの検定法には正確法，適合度検定，尤度比検定などがあるが具体的な手法は（鎌谷（2007），Wigginton et al. (2005)）を参照してほしい．

関連解析全体の問題点を把握するために Hardy-Weinberg 平衡テストの QQ-plot（quantile-quantile plot）が有用である．これは期待される統計量（観察された統計量の順位から計算できる）と観察された統計量の関係を示す，すべての SNP の検定結果のプロットである．これが $y=x$ の関係からずれた場合はそれを説明する何らかの要因を考える必要がある．タイピングエラーが多い場合，集団に構造化が大きい場合，多くの関連 SNP がある場合などに QQ-plot がずれる．

6 集団の構造化（層別化）とその修正

例えば，米国のように各人種が混合した集団で関連解析を行う時には注意が必要である．もともと異なった人種間のアレル頻度は異なるので，患者群と対照群が異なった割合で各人種を含むと問題が起きる．日本人集団についてはどれだけ均一とみなせるであろうか．一般に日本人は比較的均一であることは想像できるが，その程度と関連解析に与える影響，さらには問題がある場合の解決法については知っておく必要がある．

集団の構造化の解析と，構造化を前提とした関連解析の手法はいくつか発表されている (Pritchard et al. (2000), Nakamura et al. (2005))．しかし，極めて多数の SNP と個体を対象とするには適当ではない．全ゲノム関連解析において極めて有効な人口の構造化解析の手法は genomic control の概念（Devlin and Roeder (1999)）を用いる方法，及び主成分分析と Multidimensional scaling（MDS）法である．ここでは主成分分析を用いる方法を紹介する．MDS 法は PLINK というプログラムで実行できる．PLINK は MDS 法以外にも，GWAS のデータ解析のさまざまな手法をサポートしている（Purcell et al. (2007)）．

多数の個体から得られた 1 万以上の SNP 遺伝子型データを用いて主成分分析や MDS 法によりクラスタリングを行うと，驚くほど明確なクラスタリングの結果が得られることが多い．遺伝以外の領域ではこれほど明確な結果が得られない．その理由は，前述のような確率の安定性にある．遺伝的データは他の分野と異なって確率が極めて安定しているため明確な結果が得られるのであろる．主成分分析により集団の構造化を分析する方法として EIGENSTRAT というソフトウェアを用いる方法がある（Price et al. (2006)）．目的は個人のクラスタリングである．

主成分分析は多次元空間で，その直線への射影の分散が最大となるように直線を引く手法である．そのためには，多次元空間の多数の点の間の分散共分散行列を計算する必要がある．そして，その行列の固有ベクトルを計算すればよい．例えば，1000 人から 100,000 個の SNP の遺伝子型データが得られたとする．1000 人の個体をクラスタリングするために，それぞれの SNP を次元とし 100,000 次元空間における 1000 個の点を考える．しかし，そのためには 100,000×100,000 の行列の共分散行列を計算する必要がある．これはそれぞれが 1000 個の点を含む 50 億個の共分散の計算が必要なことを意味し，現在の計算機では計算が困難である．

発想を逆にし，人を次元に，点をSNPにして考える．1000次元の空間で100,000個の点を想定し，1000×1000の共分散行列を計算すればよく，計算が可能ととなる．この場合は，通常の主成分分析のように，固有ベクトルを含む直線への射影によりクラスタリングを行うのではなく（それだとSNPのクラスタリングになる），固有ベクトルの因子によりクラスタリングを行うのである（これにより個人がクラスタリングされる）．

また，これは構造化とは少し異なるが，サンプル中に近縁関係の個体，あるいは極端な場合同一個体からのサンプルが含まれている場合は，それを平均IBS（identity by state）（Rioux et al.（2007）），またはIBD（identity by descent）（Milligan（2003））の推定により検出し，一方を除外することが望ましい．

7 関連解析の手法

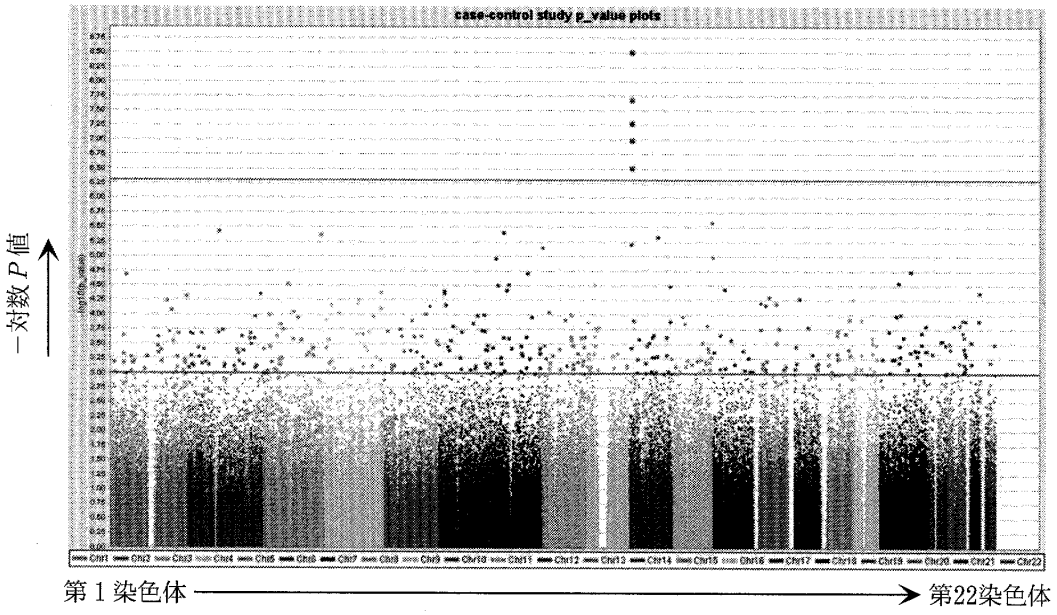
各SNPを単点とした質的形質とSNP座位との関連の検定の為には三つの遺伝子型と二つの表現型（例えば疾患あり，なし）の分類による3×2の偶現表（contingency table）が基礎になる．この表を用いた検定はCochrane-Armitage検定（傾向検定）（Armitage（1955）），あるいは自由度2のχ^2分布を用いたPearsonの独立性の検定がしばしば用いられる．さらにこの表を基にし，2×2の表を作成し自由度1のχ^2分布を用いた，アレル頻度，優性モード，劣性モードによる検定もしばしば行われる（鎌谷（2007））．

一つの研究で数十万もの検定を行うので，その結果の表示は大変である．P値の結果を第一染色体から第22染色体（場合によってはX染色体も含め）まで，短腕から長腕方向に並べて表示する．縦軸は$-\log_{10}P$の値の点として表示することが多い（HaploViewというプログラムで効率的に表示できる）（Barret et al.（2005））．このようなプロットに加え，QQ-plotの図を作成することが望ましい（鎌谷（2007），Weir et al.（2004））（図1）．

Genomic control（gc）の概念は集団の構造化などによる統計量の増大を調べるために重要である．多くのSNPの情報を用いて（ほとんどのSNPは表現型と関連が無いはずである）例えばCochrane-Armitage検定により統計量を計算する．すべてのSNPに関する統計量の分布から帰無仮説の下での分布（Cochrane-Armitage検定の場合は自由度1のχ^2分布）と比較し，どの程度統計量が膨張するかの比率（variance inflation factor）をλで表す（Devlin and Roeder（1999））．

8 何十万もの多重比較にどう対処するか

多重比較の問題はGWASで大きな問題である．即ち，数10万ものSNPを用いてそれぞれのSNPについて検定を行えば，低いP値を示すSNPが多数出現する．すべてのSNPが疾患と関連無ければ（帰無仮説）P値は[0, 1]の一様分布に従う．従って，例えば50万個のSNPについて検定を行えば，$P\leq 0.05$となるSNPは平均$0.05\times 500{,}000=25{,}000$個あると考えられる．これが多重検定の問題である．それを避けるために最も厳しい補正法はBonferroniの補正法である．これは全体の有意水準（例えば，$\alpha=0.05$）を検定の数で除する方法である．即ち，50万個のSNPで単点解析を行う場合は，一つのSNPの有意水準を$0.05/500{,}000=10^{-7}$とする方法である．しかし，この補正法は個々のSNPの検定の棄却域がすべて互いに排他の時のみ正しい補正法である．すべての検定が独立であり，しかも個々の検定の有意水準が小さいときは近似的に正しい補正法である（鎌谷（2007））．もし，通常の関連解析で用いられるSNPの様に一部に連鎖不平衡が存在する場合は検定が独立ではなく，Bonferroniの補正は過剰に保守的である（有意であるものを落とす傾向がある）．このような問題にいくつかの対策が取られている．



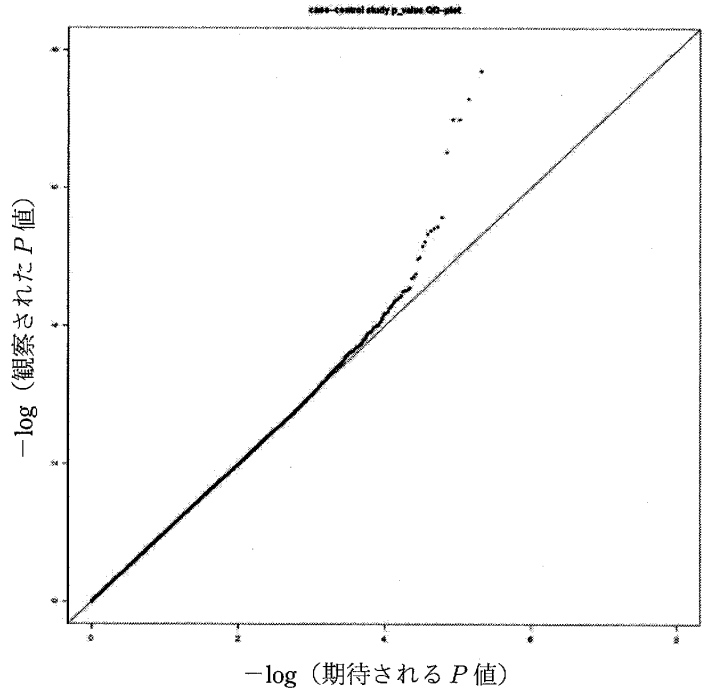


図 1　関連解析の結果を示す P 値の分布（上）と期待される P 値の $-\log$（又は統計量）と観察された値の関連を示す QQ-plot（下）.

Benjamini と Hochberg による false discovery rate（FDR）の概念を用いる手法もしばしば取り入れられる．これは個々の検定の偽陽性の確率を有意水準で調整するという従来の方法ではなく，陽性と報告された検定の中の誤って陽性と報告される SNP の割合を一定に（例えば 0.05, 0.1 など）調製する手法である．詳細については他の文献を参照してほしい（鎌谷（2007），Benjamini and Yekutieli (2005)）.

また Permutation 法もしばしば用いられる．これは症例・対照の表現型をランダムに入れ換え，そのように作成されたサンプルについて多くの SNP について検定を行う方法である．表現型をランダムに入れ換えることにより表現型と遺伝的多様性が関連の無い（帰無仮説）状態

が得られる．このような状態で得られた最小P値，あるいは最大統計量（全SNPの中で）の分布から帰無仮説の下でのP値，あるいは統計量の分布を調べる．その上で，観察データから得られたP値，あるいは統計量が，有意水準（例えば$\alpha=0.05$）に照らして有意かどうかにより検定を行うのである（鎌谷（2007），Sladek et al.（2007））．しかし，帰無仮説における信頼できる分布を得るための，シミュレーションにより生成されるべきサンプルの数が膨大になり，計算に時間がかかることである．これを補うため，importance sampling法を用いたRAT（rapid association test）法が発表されている（鎌谷（2007），Kimmel and Shamir（2006））．

さらに多重比較の補正法としてBayes的手法もある．WTCCCのGWASではBayes法を用い，50万SNPについて，数千のサンプルサイズで検定を行った場合，$\alpha=5\times10^{-7}$に設定している（Wellcome Trust Case Control Consortium（2007））．

これまでの検定ではすべて個々のSNPごとの単点解析を行った場合の方法を示した．それ以外にもハプロタイプを基礎にした色々の検定法が発表されているが，標準的な方法は定まっていない（鎌谷（2007），Epstein and Satten（2003），Furihata et al.（2006），Shibata et al.（2004））．

また，複数の遺伝子に存在するSNPの相互作用を仮定した検定法も発表されている（Hahn et al.（2003），Zhang and Lin（2007））．しかし，これについても標準的な方法は確定していない．

これらの複雑な解析法の標準化がこれから重要になるであろう．

9 再確認研究とメタ解析

初期の遺伝的関連研究では，統計解析により抽出された遺伝子の産物である蛋白質のレベルで，疾患と関連のある分子的証拠が発見されることが望ましいと考えられていた．しかし，慢性疾患の場合，何十年もかけて疾患が発生するのである．それを短時間の試験管内の実験で再現できるのは困難かもしれない．最近では，最初のサンプルとは独立のサンプルで統計的検定結果が再確認されることの方が重要であると考えられるようになった．今や，遺伝的関連研究の世界は統計学者の洗練された解析手法無くては不可能な時代に入ってきた．

また，複数の研究による統計的証拠を統合するメタ解析による検討も重要視されている．そのような再確認研究では必ずしも最初の研究の結果が再現されるとは限らない．しかし，注意深く十分のサイズのサンプルを用いて行われたGWASの結果は再現されることも多いことがわかってきた．人種を超えて，例えば，白人と日本人で同じ遺伝子の同じアレルがリスクとなっていることもあれば，日本人だけ，あるいは白人だけで関連が証明されることもある．

このように最初に得られた遺伝的多様性と表現型との関連を，独立のサンプルで再確認することは，それを臨床応用するための重要な条件である．このようなステップはclinical validity（臨床的妥当性）を確認する不可欠の過程として推奨されている．

10 表現型の予測システム

近年，GWASが大規模に行われるようになり，信頼できる結果がかなり蓄積されてきた．一般的に自己免疫疾患などではオッズ比のかなり高い，即ち効果サイズの大きな遺伝子が抽出されることが多い．また，別の自己免疫疾患の間で関連遺伝子を共有することも多い．しかし，糖尿病や高血圧症などの頻度の高い疾患では個々の関連遺伝子の効果サイズはそれほど高くないことが多い．以前から予測されていた通り，頻度の低い変異ほど効果サイズが大きいことが多く，頻度の高い変異ほど効果サイズが小さいことが多い．効果サイズが低いからといって重要でないわけではない．それらの遺伝子は頻度が高い傾向があり，集団全体に与える影響は高い傾向があるからである．その因子が集団全体の特定の疾患に寄与している割合をpopulation attributable risk（PAR）という．

多くの頻度の高い疾患について，GWASにより次々に遺伝的原因が解明されていくであろう．また薬の反応性についても関係する多型が発見されるであろう．それらを用いて表現型を予測するシステムを構築できるであろうか．

多くの予測は質的形質に対してのものである．例えば，糖尿病になるかどうか，特定の薬が効くかどうかの予測である．メンデル型の遺伝病では特定の遺伝子が表現型を決める．しかし，一般の複雑な形質では遺伝子は表現型を取る確率を変化させるに過ぎない．遺伝子や環境要因と質的表現型の関係についてはFisherの提唱した一般化線形モデルがしばしば用いられる．即ち，特定の座位の遺伝子型を数値化し遺伝子型値という変数にする．また環境要因も変数（環境値）にして，それらを線形結合した変数が疾患になりやすさ（liability）を決めると考える．Liabilityの値と，疾患になる確率との関係は，例えば正規分布の累積分布関数を考える．

最近では，liabilityの値と，疾患になる確率との関係をロジスティックモデルで解析することが多い．即ち，個人が疾患になるかならないかのオッズの対数（ロジット関数）が遺伝子型値や環境値（あるいは性別や年齢などの共変数）の線形結合で表されると考え，多くの患者や対照群から得たデータを用いてロジスティック回帰分析を行うのである．得られた推定係数を用いて，新たな個人の疾患になりやすさ（liability）を計算できる．Liabilityが特定の閾値を超えれば疾患，超えなければ疾患と予測する．この予測法は閾値を変化させることで変わる．閾値を変化させ感度，特異度の変化を二次元グラフで表したものがROC曲線である．ROC曲線の下の面積（AUC）が大きいほど良い診断法と結論できる．

ここで，複数の遺伝子型のデータを変数として取り扱う場合に注意が必要である．ハーディー・ワインバーク平衡を仮定すれば，異なった染色体上の多型は必ず独立であり，同じ染色体上の多型は連鎖不平衡の程度に応じて関連する．解析上，優先するのはそれらの遺伝法則であり，統計的モデルによる結論ではない．このあたりも遺伝的データを取り扱うためには重要な要素である．

これらの表現型予測モデルについては，予測精度を示す指標を考慮する必要がある．例えば，感度，特異度，陽性的中率，陰性的中率を精度良く推定したり，ROC曲線を描いてAUC（area under the curve）を計算する必要がある．そのような解析により，このモデルを臨床に用いることの是非を判断する必要がある．このステップはclinical utility（臨床的有用性）の判断として重要なステップと考えられている．

参 考 文 献

Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics*, **11**, 375–386.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, **21**, 263–265.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait Loci analysis using the false discovery rate, *Genetics*, **171**, 783–790.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C.F. (1998). Haplotype structure and population genetic inferences from nucleotide-sequencevariation in human lipoprotein lipase, *Am J Hum Genet*, **63**, 595–612.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies, *Biometrics*, **55**, 997–1004.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotypedata, *Am J Hum Genet*, **73**, 1316–1329.

Furihata, S., Ito, T. and Kamatani, N. (2006). Test of association between haplotypes and phenotypes in case-control studies:examination of validity of the application of an algorithm for samples fromcohort or clinical trials to case-control samples using simulated and real data, *Genetics*, **174**, 1505–16.

Hahn, L. W., Ritchie, M. D. and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene andgene-environment interactions, *Bioinformatics*, **19**, 376–382.

鎌谷直之 (2007). 『遺伝統計学入門』. 岩波書店.

Kimmel, G. and Shamir, R. (2006). A fast method for computing high-significance disease association in largepopulation-based studies, *Am J Hum Genet*, **79**, 481–492.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness, *Genetics*, **163**, 1153–1167.

Nakamura, T., Shoji, A., Fujisawa, H. and Kamatani, N. (2005). Cluster analysis and association study of structured multilocus genotype data, *J Hum Genet*, **50**, 53–61.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated withsusceptibility to myocardial infarction, *Nat Genet*, **32**, 650–654.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wideassociation studies, *Nat Genet*, **38**, 904–909.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkageanalyses, *Am J Hum Genet*, **81**, 559–575.

Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W., Shugart, Y. Y., Griffiths, A. M., Targan, S. R., Ippoliti, A. F., Bernard, E. J., Mei, L., Nicolae, D. L., Regueiro, M., Schumm, L. P., Steinhart, A. H., Rotter, J. I., Duerr, R. H., Cho, J. H., Daly, M. J. and Brant, S. R. (2007). Genome-wide association study identifies new susceptibility loci for Crohndisease and implicates autophagy in disease pathogenesis, *Nat Genet*, **39**, 596–604.

Shibata, K., Ito, T., Kitamura, Y., Iwasaki, N., Tanaka, H. and Kamatani, N. (2004). Simultaneous estimation of haplotype frequencies and quantitative traitparameters: applications to the test of association between phenotype and diplotype configuration, *Genetics*, **168**, 525–539.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genomewide association study identifies novel risk loci for type 2 diabetes, *Nature*, **445**, 881–885.

Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T., Suzuki, M., Nagasaki, M., Nakayama-Hamada, M., Kawaida, R., Ono, M., Ohtsuki, M., Furukawa, H., Yoshino, S., Yukioka, M., Tohma, S., Matsubara, T., Wakitani, S., Teshima, R., Nishioka, Y., Sekine, A., Iida, A., Takahashi, A., Tsunoda, T., Nakamura, Y. and Yamamoto, K. (2003). Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylargininedeiminase 4, are associated with rheumatoid arthritis, *Nat Genet*, **34**, 395–402.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H. and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copynumber variation detection in whole-genome SNP genotyping data, *Genome Res*, **17**, 1665–1674.

Weir, B. S., Hill, W. G., Cardon, L. R.; SNP Consortium. (2004). Allelic association patterns for a dense SNP map, *Genet. Epidemiol*, **27**, 442–450.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, **447**, 661–678.

Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium, *Am J Hum Genet*, **76**, 887–893.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies, *Nat Genet*, **39**, 1167–1173.