1.5 有意差があるというのはどういうこと

統計学の基本であるが、有意差があるということについて考えてみる。有意差検定とは仮説に対してその仮説の正しさを検定しているといっていい。その仮説であるが、ある多型頻度が患者と対照で差があるということを仮説として設定することは簡単ではない。通常、差がない（帰無仮説）ことを仮説とし、帰無仮説が棄却できるかどうかを検定する。有意差があるということは、ややっこしいが「差がないことはない」である。統計学的にはあくまでも差があるという結論ではないことは注意いただきたい。

帰無仮説を棄却する過誤がいわゆる P 値である。帰無仮説が正しい場合の観察確率といってもいいだろう。偽陽性といった方がなじみ深いかもしれないが、偽陽性のおこる確率である。ある有意水準をあらかじめもうけ、それより小さい場合に有意であると判定する。有意水準は便宜的なものであるが、0.05 がよく使われる。再度、仮説検定についてのべる。研究者は通常なにか仮説を立ててそれを検証する。アソシエーション・スタディの場合、興味ある遺伝子多型が病気と関連する、が仮説であろう。関連することを証明したいわけだ。しかしながら統計学では通常、差がないを仮説におく（帰無仮説）。関連するということは差があるということながら、どの程度の差かわからない、一方、差がないということは差は 0 である、と仮定を設定しやすいからと考えていただいても結構である。繰り返すが、得られた P 値は帰無仮説の正しい観察確率である。

有意差があるということは差がないこと（帰無仮説）の可能性が低いということであり、関連の強さを示すものではない。

1.6 得られた有意差をどのように評価したらいい？

ここで非常に低い *P* 値を得たとしよう、つまりある遺伝子多型が疾患と強い有意差を得た。このことはこの遺伝子多型が疾患と強く関連することを示すものだろうか。なんども説明したように、帰無仮説を検定しているのであり、強い有意差とはほぼ間違いなく帰無仮説を棄却できるというだけである。実際に χ^2 値は検体数に比例すると述べた。検体数を増やすことで（というより増やす必要がある）、強い有意差を得ることができる。すなわち、χ^2 検定で有意差検定できても、どのくらい強く関連しているかはわからない。関与の強さの指標として odds ratio が用いられる。アレルカウントを用いた 2×2 の分割表は不適切なところがあるが、簡単のため 2×2 で説明したい。odds ratio は表 1 の分割表から単純に以下で与えられる。

$$\text{odds ratio} = \frac{NaNd}{NbNc} \qquad \text{(式 4)}$$

アレル a を有すると有しない人に比べ、例えば 2 倍病気にかかりやすい。という結論が通常導かれる。実際にはアレル分布は 0、1、2 であるので個々人の遺伝型に従い、劣性モデルなら aa の方とそうでない人との比較、優性モデルなら Aa, aa と AA を有する人との比較で odds ratio は検討すべきだろう。

近年ゲノム全域アソシエーション・スタディの結果が洪水のようにおしよせており、疾患遺伝子が続々と同定されている。ひとつの例として乳がんの解析例を示そう。16,423 例の乳がん患者、17,109 例の対照でアソシエーション・スタディをおこなった結果、CASP8 で有意差（$5X10^{-7}$）を得ている。しかしながら、odds ratio はアレルにおいて 0.88 であり、この多型を有すると 0.88 倍乳がんに罹りやすいという結論となる(Cox A et al. 2007)。有意差は強くてもこのようにごく小さい関与（この場合抵抗性に働く）を有する遺伝子多型の生物学的意義は不明である。現時点では、強い有意差を得たことをよしとして、その後の思考を停止させているところがある。odds ratio がほとんど 1 と変わらない場合、現在のどのような実験手法であれ、機能的変化を証明することは困難であろう。さらに相互作用の解析手法やテクノロジーの進歩を待つ必要があるかもしれない。

Odds ratio は危険度の強さの指標となる。

1.7 Hardy-Weinberg 平衡と乖離した場合の統計遺伝解析

対照を一般集団においた場合、通常アレル分布は Hardy-Weinberg 平衡（HWE）に従うことが期待される。対立遺伝子を A, a とし、その頻度を p, q とする。p+q=1 である。HWE に従うと遺伝子型　AA, Aa, aa の頻度はそれぞれ p^2, 2pq, q^2 となる。HWE が成りたつための条件として 1）集団が大きく、検体数が十分であること、2）任意交配であること、3）移住、自然選択がないこと、があげられる。上記条件が満たされない場合、HWE からの乖離が観察されることがある。HWE からの乖離は自由度 1 のカイ検定でおこなう。

アソシエーション・スタディにおいて問題となる HWE からの乖離について考えてみよう。まず、乖離が観察される要因を上記に示したが。現実的には原因の多くはタイピングエラーである。少ない数の SNP サイトでは再度タイピングの精度を確認した方がいい。一方、ゲノム全域アソシエーション・スタディのように 30–50 万 SNP を解析する場合には、全体の精度確認など、品質管理に留意しなくてはならない。対照を非罹患者とするとある病気に罹りにくい対象を収集しているので、ある遺伝子多型が防御的に関与する可能性があり、HWE からずれることもあるかもしれない。ある SNP が疾患と関連していると、対照で HWE にはあるが、患者群でずれがある場合が想定される。それでも、遺伝モデル（優性か劣性か）を決定できることは稀であり、古典的なアソシエーション・スタディでは、それでも通常通りカイ検定をおこなうといい。ゲノム全域解析の場合、P 値の低い順に候補を絞り込む必要があるので、HWE から乖離している場合、なんらかの遺伝モデルに従った解析が必要となる。そこで 2x3 でありながら検体数（2N）を減らすことなく解析できる Cockran-Armitage trend test が採用されることが多い。

ここでは Armitage trend test での χ^2 値（$\chi^2_{AC\text{-}trend}$）と通常のカイ検定の比較のみをみてみよう。

$$\chi^2_{AC-trend} = \frac{N-1}{N} \times \frac{\chi^2_{Allelic}}{2 - \frac{Het_{obs}}{Het_{exp}}} \qquad (式5)$$

N: total number of individuals in cases and contols
Het_{obs}: observed number of heterozygotes in cases and controls
Het_{exp}: expected number of heterozygotes in cases and controls under HWE

アレル頻度からのカイ2乗値と、Cockran-Armitage trend 統計量との関係は、上の数式で示される。その数式から明らかなように、ケースコントロール合算ヘテロ接合数が、HWE のもとでの期待値と一致する場合は、両者の統計量はほぼ一致する。一方、ヘテロ接合超過の場合は、アレル頻度からの統計量が大きくなり、逆にホモ接合超過の場合には、トレンドテストからの統計量が大きくなる。多くのゲノム全域タイピング手法では、ヘテロ接合がタイピングされない傾向があり、結果的に「ホモ接合超過」のデータが得られやすく、トレンドテストからの P 値がより小さい値を示しやすい特徴を有する。この点は注意しておかなければならない。

Hardy-Weinberg 平衡は常に確認する。乖離が観察され、遺伝子タイピングに問題がない場合、アソシエーション・スタディの検定に Armitage trend test を採用する。

1.8 さいごに

common disease は複雑なネットワークが相互作用しつつ発症にいたると考えられる。遺伝要因はそのひとつに過ぎないし、関与する遺伝子もひとつではない。当然、SNP によるアソシエーション・スタディのみでなくいくつかの手法を組み合わせて検討する必要がある。本稿では、SNP 対連鎖不平衡、ハプロタイプによるアソシエーション・スタディにはまったく触れなかった。国際 HapMap 計画によりゲノム全域にわたる連鎖不平衡、ハプロタイプ構造が明らかになりデータベース化されている。本書の 2 章を参考にしていただければと思う。

ゲノム全域解析など多くの遺伝子多型を統計解析すると必ず有意差を有する多型がでる。多くは偽陽性であり、偽陽性を減らすためには検定数による補正が一般的である。このような多重検定補正法についても詳細は 5 章を読んでいただきたい。

複雑なネットワークが疾患原因のベースに存在する以上、遺伝子間、遺伝子-環境要因間の相互作用解析が重要な位置を占める。しかしながら解析法についてはまだ発展途上であり、本稿では示さなかった。今後重要になる解析なので、興味ある方は 6 章を読んでいただきたい。

2. SNP を用いた疾患感受性遺伝子同定

2.1 はじめに

疾患感受性遺伝子の同定とは、研究対象とする疾患に関わる遺伝子がゲノム上のどこに位置し、どのようなタイプの変異をもち、そして、その変異がどのようにして疾患に関わっているのかを明らかにすることである。メンデル型の遺伝様式を示す単一遺伝病については家系解析による疾患遺伝子マッピング法が確立しており、数多くの責任遺伝子が同定されてきている。一方、メンデル型の遺伝様式を示さない多因子疾患の遺伝子マッピングについては、様々な手法を用いて世界中で精力的に研究がなされているにもかかわらず、依然解決すべき課題が多く、定まった方法は確立されていない。本章では、SNP を遺伝マーカーとして用いた多因子疾患の感受性遺伝子マッピング法の概略について紹介する。その詳細に関しては、以降の各章を参照されたい。

2.2 SNP を用いた疾患感受性遺伝子マッピングの概略

2.2.1 研究対象とする「多因子疾患」の遺伝性について

疾患発症リスクが単一の遺伝要因からなる疾患（単一遺伝病）はまれであり、多くの疾患では、その発症に遺伝要因、環境要因（生活習慣、外的環境など）がともに関わっている（図 2.1）。これらの要因は、それぞれ単独で疾患発症に寄与するというよりも、相互に、かつ複雑に影響し合う（遺伝子−遺伝子相互作用、遺伝子−環境相互作用など）ことにより発症に関与していると考えられており、このことが多因子疾患の発症メカニズム解明を困難にしている原因のひとつとなっている。このような複雑さを内包する多因子疾患ではあるが、単一遺伝病と同様、遺伝要因は原因そのものであるため、疾患遺伝子同定が成因解明に直結し、さらにテーラーメイド医療と呼ばれる個人差に応じた医療実現のための基盤となることが期待されている。

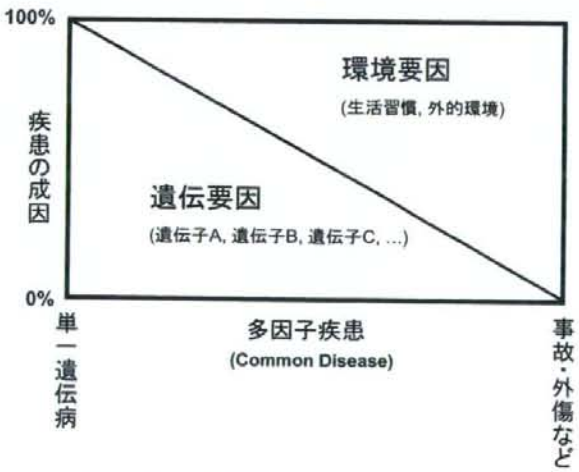


図 2.1　疾患発症要因についての概念図

メンデル型の遺伝様式を示す単一遺伝病は、単一の遺伝要因のみが疾患発症に寄与するのに対して、事故などによる外傷のほとんどは、外的環境要因のみが関与すると考えられる。多因子疾患は、遺伝要因、環境要因がともに疾患発症に関わっており、それらの比率は疾患ごとに異なる。

SNP を用いた疾患感受性遺伝子マッピングとは、本章では、多因子疾患の遺伝要因を見つけ出すことを意味する。従って、当然のことではあるが、遺伝要因の

関与が示唆される多因子疾患のみが研究対象なる。一般に、双生児（一卵性、二卵性）における発症一致率や、家系内集積性（ある疾患の発端者を含む家系内に同一疾患に罹患する血縁者が見つかりやすい傾向）といった近親度の高い人びとからの遺伝疫学データから、疾患に対する遺伝要因の関与の程度を推し量ることができる。さらに、各疾患の再発リスク比（recurrence risk ratio [λ_R]：R は血縁関係を表す）と呼ばれる数値も、疾患の遺伝性を考える上で重要な指標である。これは、別名、相対危険率と呼ばれ、疾患罹患者と特定の近親度にある血縁者（例えば、罹患者の兄弟姉妹といった同胞など）における疾患罹患率を、非血縁者からなる一般集団（例えば、日本人集団など）における疾患罹患率で除した値である。同胞再発リスク比は「同胞 Sibling」の頭文字 S をとってλ_Sと表記するが、例えば、多因子疾患のひとつである遅発性アルツハイマー病のλ_S値は 5 であることが報告されている（Scott et al. 1997）。再発リスク比の算出方法から明らかなように、λ_S=5 とは、罹患者同胞での疾患罹患率が、一般集団のそれの 5 倍高いことを意味するが、一般に、この値が 1 より有意に大きい場合には、対象疾患の遺伝性を考慮することが妥当であるといえる。ただし、同胞は食事など同一の環境のもとに育っていることが多く、この値が高めに推定されてしまうことは留意すべきである。

さて、研究対象としたい多因子疾患に遺伝性が確認されたとして、その発症にいくつの遺伝子が関与しているのか、それらの遺伝子の関与の強さはどの程度か、ということが次に注目すべきところとなる。実際には、疾患感受性遺伝子同定研究を始める前にこのようなことを知ることは不可能であるが、対象とする多因子疾患の遺伝様式について、あらかじめ情報収集、あるいは推定しておくことは重要である。例えば、同じ多因子疾患といっても、「ポリジーン遺伝モデル」では、非常に作用の弱い、多数の遺伝子が疾患発症に関わると考える、一方、「オリゴジーン遺伝モデル」では、罹病性に関わる多数の遺伝子の中に、相対的な寄与率の大きい遺伝子（これをオリゴジーンと呼ぶ）が幾つか存在すると考える、というような様式による違いがある。一般に、効果の強い因子の方が見つけやすいことは容易に想像されると思うが、効果の大きさがどの程度の遺伝因子を検出することを目的とするかは、後述するように、疾患感受性遺伝子同定のための研究デザイン（遺伝解析方法や解析に必要なサンプル数など）に大きく影響する。

疾患の遺伝性や関連遺伝子座、感受性遺伝子などの最新情報をカタログ化した公開データベースは複数存在するが、NCBI (National Center for Biotechnology Information、米国国立バイオテクノロジー情報センター)管理下の OMIM (Online Mendelian Inheritance in Man: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) は、他のデータベース（PubMed など）とのリンクも充実しており、研究対象とする疾患についての研究成果の概要をとらえるには有用である。データベースのタイトルとは異なり、メンデル遺伝型の単一遺伝病に加えて、非メンデル遺伝型の多因子疾患も収録されている。一方、既存の情報が存在しない場合でも、家系内集積性などの遺伝疫学データに基づく遺伝様式の推定（Risch 1990a; Risch 1990b; Risch 1990c も参照）や、血圧や体重といった量的形質に対する集団遺伝学的研究から、例えば、高血圧症のような量的疾患の感受性遺伝子の数や、その関与の程度の推定を行う（Rudan et al. 2003; Wright et al. 2003 など）、といった試みも積極的になされている。

2.2.2 疾患感受性遺伝子マッピングのための研究デザイン

遺伝性の確認された多因子疾患の感受性遺伝子を同定するために、次に考慮すべきことは、どのような人びとから DNA サンプルを収集して、どの SNP を調べれば良いか、ということである。すなわち、研究デザインの問題である。感受性遺伝子を同定する手法はいくつもあるが、SNP を用いた多因子疾患の感受性遺伝子マッピング法として現在頻用されているのは、アソシエーション・スタディと罹患同胞対連鎖解析である(図 2.2; Risch 2000 など)。アソシエーション・スタディ（関連解析）は、その研究対象とするサンプルの違いから、患者-対照関連解析（Population-based association study）と家族内関連解析（Family-based association study）とに大別できる。患者-対照関連解析では、多因子疾患の罹患者集団（患者群）と、それら罹患者とは血縁関係にない非罹患者集団（対照群）とを比較研究の対象とするのに対して、家族内関連解析では、主として、疾患に罹患した子供とその両親（両親は非罹患でも可）からなる家系サンプルを収集し、研究に供する。他方、罹患同胞対連鎖解析を行うには、同一家系内で疾患に罹患した同胞（ペア、トリオなど）からのサンプルを数多く集める必要がある（図 2.2)。いず

れにおいても、疾患感受性変異の存在しているゲノム領域（遺伝子など）は、患者間では、その共通祖先から同じものを受け継いでいる、という仮定に基づいている。すなわち、アソシエーション・スタディでは、感受性変異の集団における共有について、一方、罹患同胞対連鎖解析では、罹患同胞内での共有について統計学的に分析している、といえる。

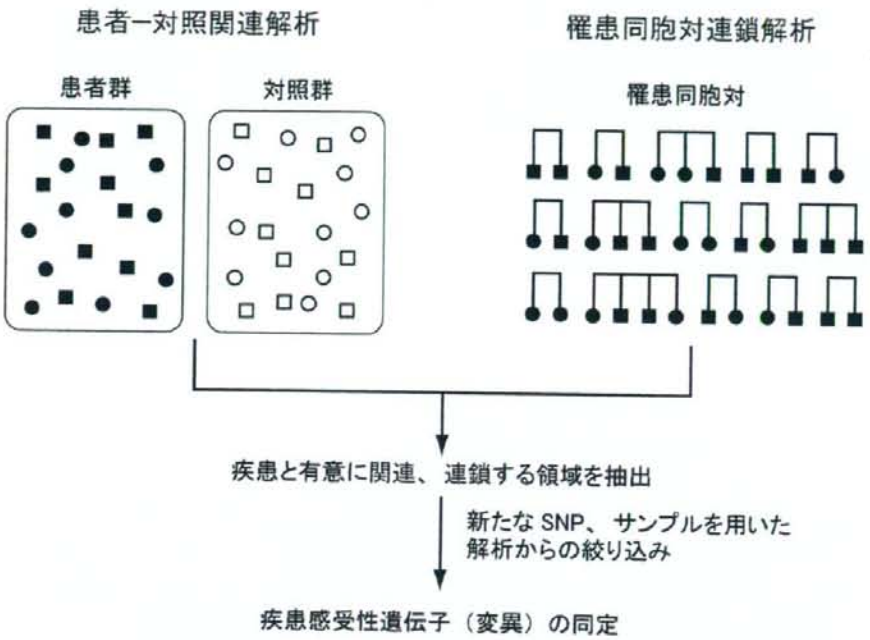


図 2.2　疾患感受性遺伝子マッピングの概略図

患者−対照関連解析および罹患同胞対連鎖解析において、疾患感受性遺伝子を同定するまでの大まかな流れを示している。それぞれの解析において、統計学的に疾患に有意に関連するマーカーSNP、あるいは疾患と連鎖する遺伝領域を見いだすことが最初の目標である。初期解析により抽出された遺伝マーカー周辺領域を遺伝学的に精査することに、疾患感受性遺伝子の同定にいたるが、その具体的な方法については他章などを参照されたい。

この 2 つの手法が頻用される理由のひとつは、多因子疾患という疾病の特徴に由来するサンプル収集上の容易さ・困難さに起因する。一般に、メンデル型の単一遺伝病は若年発症する場合が多いため、若年発端者からみて 3 世代以上を隔てた大家系の収集が可能であるのに対して、多くの多因子疾患は、子供の時には発症しない、いわゆる成人型疾患であるために、家系（特に複数世代にわたる大家系）の収集が非常に困難である。その一方、多因子疾患の疾患頻度は高いため、

同一の疾患表現型を持つ罹患者集団を集めることは比較的容易であり、また、少子化社会にある我が国においても、罹患同胞を集めることは十分可能である。

患者-対照関連解析と罹患同胞対連鎖解析の主な特徴を表 2.1 にまとめて示す。多因子疾患の遺伝研究に固有の問題としては、この疾病の遺伝的異質性、低浸透率、表現模写を挙げることができる。遺伝的異質性とは、数多くの遺伝子座が疾患を引き起こすことと関連して、ある集団や家系で疾患に関わることが明らかにされた遺伝子であっても、他の集団、家系では関与するとは限らないことを意味する。低浸透率とは、疾患に関わる遺伝子型を持っていても罹病しないことであり、一方、表現模写とは、疾患に関わる遺伝子型を持っていないのに（環境要因などにより）罹病することである。現在のところ、これらの問題への最適な対処法は存在せず、従って、どちらの解析手法を用いても、このような問題が感受性遺伝子マッピングの妨げとなり得る。表 2.1 に示したように、疾患への効果の大きさ（遺伝子型相対リスク[Genotype relative risk; GRR]を尺度とすることが多い）が同一の遺伝子座を見つけ出すための検出力（有意な結果を得る期待値）は関連解析の方が高く（Risch 2000 参照）、このことは、少ないサンプル数で感受性変異を同定することが可能であることを意味している。一方、連鎖解析は、どちらかといえば遺伝的異質性に対して頑健で、サンプル家系数を増やすことで、理論的には、どのような感受性遺伝子も同定できる可能性があるのに対して、関連解析の検出力は遺伝的異質性により低下する（Kruglyak 1997）。以上のような特徴から、これらの遺伝解析法はどちらかを選択すれば、もう片方は不要であるという関係というよりも、相互に補完し合うものであるといえよう。しかし、現実的には、サンプル収集のために必要なコスト、時間の制約から、両者を平行して研究を進めることは困難である場合が多い。対象とする多因子疾患が 1 個、ないしは 2 個程度の効果の大きい感受性遺伝子を持っていることが期待できる（オリゴジーン遺伝モデル）場合は、罹患同胞対連鎖解析からでも感受性遺伝子座の同定が十分可能である。一方、効果の大きな感受性遺伝子は期待できず、非常に効果の弱い多数の感受性遺伝子が疾患発症に関わっていることが想定される（ポリジーン遺伝モデル）場合は、患者-対照関連解析から感受性遺伝子に迫る、といったアプローチをとることが無難であろう。なお、家族内関連解析については、その実際の方法（伝達不平衡テスト[Transmission disequilibrium test; TDT]、ハプロタ

イプ相対危険率[Haplotype relative risk; HRR]テストなど）とともに、他章を参照されたい（Gauderman et al. 1999 なども参照）。

表 2.1　患者-対照関連解析および罹患同胞対連鎖解析の主な特徴

	患者-対照関連解析	罹患同胞対連鎖解析
戦略	ゲノム全域スキャン 候補遺伝子アソシエーション	ゲノム全域スキャン
検出力	高い	低い
遺伝的異質性	検出力低下	比較的頑健

それでは、どのような SNP を調べればよいのであろうか。患者-対照関連解析は、もともと、他の遺伝解析などから疾患候補ゲノム領域や、候補遺伝子群が絞り込まれている場合の感受性遺伝子同定法として利用され、現在もなお利用されている。この場合は、その候補ゲノム領域や候補遺伝子内に位置する SNP を調べれば良いこととなる（具体的な SNP 選定方法は後述する）。このような手法は着実な成果が期待できる一方、その研究スピードにはおのずと限界がある。そこで、先験的な知見を必要としない、ゲノム全域での体系的アソシエーション・スタディ（Genome-wide association study; GWAS）により感受性遺伝子を探索する方が直接的、効率的であるという議論がある（詳細については、8, 9 章を参照されたい）。2002 年に、約 10 万 SNP を用いた GWAS により、心筋梗塞感受性遺伝子の同定がなされた（Ozaki et al. 2002）ことを契機として、その後、様々な疾患で GWAS が試みられ、新規感受性遺伝子が報告されてきている（Maraganore et al. 2005; Klein et al. 2005 など）。GWAS の場合は、研究者自らが探索する SNP を選別するというよりも、DNA マイクロアレイと呼ばれる技術を用いて、カタログ化された SNP 群（数万～数十万）を網羅的、体系的に調べることになる。

罹患同胞対連鎖解析は、ゲノム全体を探索し、感受性遺伝子の存在する場所（遺伝子座）と、その相対的重要性を推定することを目的としているため、従来、SNP よりも多型情報量が大きく、少ない遺伝マーカー数（約 400～800 マーカー）でゲノム全域を網羅できるマイクロサテライト多型が用いられてきた。1994 年のイン

スリン依存性糖尿病の感受性遺伝子座同定報告（Davies et al. 1994）以降も、マイクロサテライト多型をマーカーとして、多因子疾患の感受性遺伝子座が相次いで報告されている（Onda et al. 2001; Tanaka et al. 2003 など）。その一方、上述したように、DNA チップを用いた SNP タイピング技術が格段に向上するに伴い、ゲノム全域高密度 SNP（1 万程度）タイピングによる罹患同胞対連鎖解析の可能性も議論されている。実際、高密度 SNP データの方が感受性遺伝子座を見いだすための検出力を高めることが明らかになり（Sawcer et al. 2004 など）、近年では、高密度 SNP マーカーを用いた罹患同胞対連鎖解析からの報告例が増えてきている（Arinami et al. 2005 など）。従って、SNP を用いた罹患同胞対連鎖解析を行う場合は、上記 GWAS 同様、既にカタログ化された高密度 SNP 群を調べることになる。4 章で SNP を用いた連鎖不平衡解析についてまとめている。

患者−対照関連解析や罹患同胞対連鎖解析において高密度 SNP タイピングが行われていること、特に、罹患同胞対連鎖解析においては高密度 SNP マーカーを使用する必要があることを上述してきたが、高密度 SNP タイピングに必要なコストは、その解析設備費を含め、依然高価であることは留意すべきである（実験設備が整っていたとしても、タイピング費用は、少なめに見積もって 1,000 万円以上必要(200 例タイピングとして)）。また、言うまでもないことではあるが、実際の研究を行うに際しては、任意の効果（遺伝子型相対リスク）を持った感受性遺伝子を少なくとも 80%の検出力で見いだすために必要なサンプル数について、患者−対照関連解析（Risch 2000; Botstein and Risch 2003; Ambrosius et al. 2004 など）、罹患同胞対連鎖解析（Risch 1990b など）のいずれにおいてもあらかじめ推定しておき、限りある研究費、時間内で、どの程度の遺伝子型決定（=(タイピングする SNP 数)×(タイピングするサンプル数)）が可能であるかを見積もることも研究デザインを決定する上で重要である。図 2.3 には、患者−対照関連解析において、有意水準 10^{-3}（図 2.3 A：50 個程度の少数 SNP のみを調べる場合）および 10^{-7}（図 2.3 B：GWAS により数十万 SNP を調べる場合）で両側検定を行った場合、検出力 80%を達成するために必要なサンプル数の一例を示している。疾患有病率 10%、感受性 SNP が相乗的に疾患発症に関わると仮定し、QUANTO ソフトウェア（Gauderman 2002a; 2002b; http://hydra.usc.edu/GxE/）を用いて計算した結果であるが、いずれの図においても曲線上側にある疾患感受性 SNP であれば検出力が 80%

以上となる。ひとつの患者-対照集団セットで統計学的に有意な成果を得るためには、患者、対照群ともに 1000 サンプル以上（図 2.3 A）、あるいは 2000 サンプル以上（図 2.3 B）あれば、例えば、感受性 SNP アレル頻度が 0.1、遺伝子型相対リスク GRR が 1.5 程度の疾患感受性遺伝子を 80%の検出力で同定することが可能であることが分かる。対象とする多因子疾患の感受性遺伝子の同定に成功するかどうかは、ここまでの準備段階が重要な役割を担っていることを強調しておきたい。

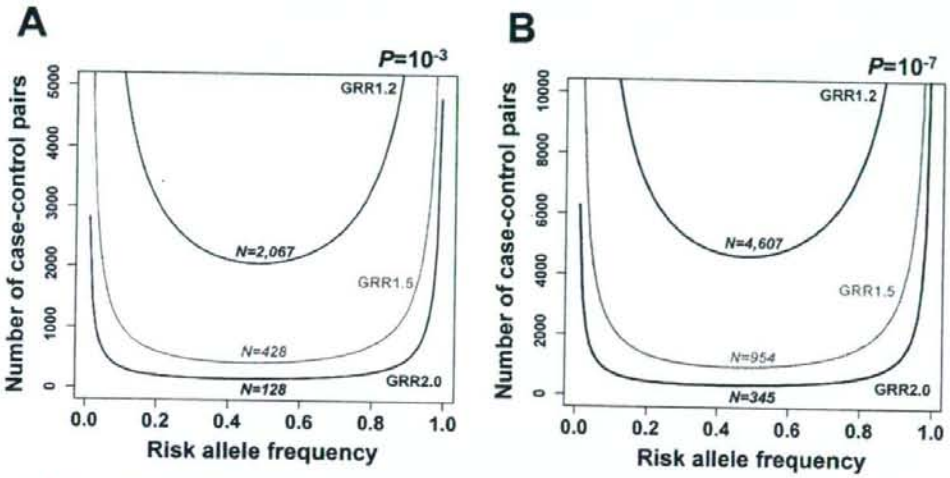


図 2.3　検出力 80%となる患者-対照関連解析を行うために必要なサンプル数

有意水準値を(A) 10^{-3}、(B) 10^{-7}に設定し、疾患有病率 10%、疾患感受性 SNP が相乗的に発症に関わると仮定し、検出力 80%を達成するために必要なサンプル数を QUANTO ソフトウェア（Gauderman 2002a; 2002b）により推定した結果を示す。図中に示した数値は、疾患感受性 SNP アレル頻度が 0.5、遺伝子型相対リスク GRR=1.2, 1.5 あるいは 2.0 の場合に必要な患者-対照サンプルペア数である。例えば、(A)の場合、アレル頻度 0.5、GRR=1.5 の感受性 SNP を検出力 80%で見いだすためには、患者、対照群ともに 428 サンプル必要となる。

2.3 タイピングするSNPについて

研究の正否を決めるもうひとつの重要なポイントは、タイピングするSNPの選択であるが、その具体的な方法に移る前に、「SNP」について要約しておこう。よくご存知のように、SNPは日本語では一塩基多型といわれ、その大多数が2種類の対立遺伝子(アレル)からなり、ゲノム上に多数網羅されている。前述したNCBIがサイト管理しているdbSNP (http://www.ncbi.nlm.nih.gov/SNP/)には、現在、約1200万のヒトSNPが登録されており(Build128)、より信頼性の高い検証済みSNP(多型性が確認されているSNPなど)の数も600万を超えている(図2.4)。

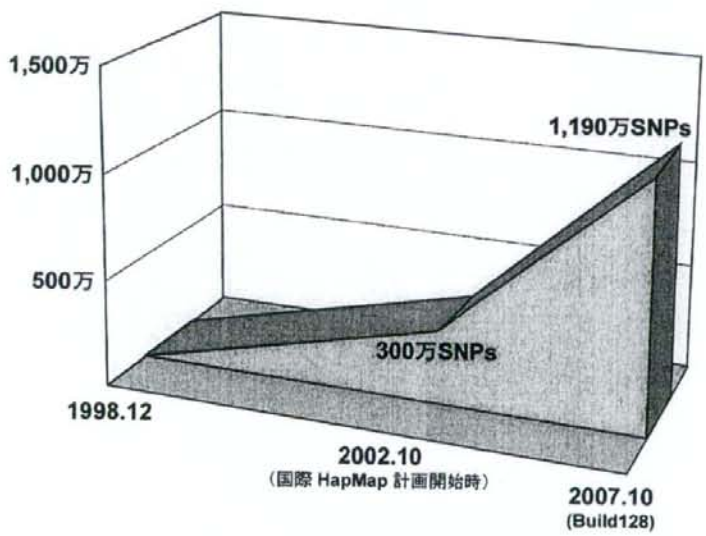


図2.4　SNPデータベースdbSNPにおける登録SNP数の変化

1998年12月に開設されたdbSNPデータベース(http://www.ncbi.nlm.nih.gov/SNP/)には、ヒトを含めた様々な生物のSNP情報が登録されている。ヒトSNP登録数の変遷を図示しているが、現在(データベースbulid128)、約1,200万種類のSNPが登録されている。

また、2002年10月に開始された国際HapMapプロジェクト(http://www.hapmap.org/)では、300万を超えるSNPデータを基にしたヒトゲノム中の詳細なハプロタイプ地図(同一染色体上の2つ以上のアレル間の連関の程度に関する詳細な地図であり、連関しているアレルの組み合わせのことをハプロタイプと呼ぶ)が作成され(The International HapMap Consortium 2003; 2005; 2007)、SNP間の遺伝的関係性

が明らかになってきている。このように、SNP データベースの充実や、SNP ハプロタイプについての知見の蓄積を背景として、SNP を用いた疾患感受性遺伝子同定研究が行われているのである。

これらの SNP の大多数は、疾患感受性に関わる実際の遺伝的変異とは連鎖不平衡（2 つの以上の変異が連関して存在している状態のこと）にあるなどの理由から、遺伝マーカーとしての役割を果たす（図 2.5）。従って、遺伝マーカーとして使用した SNP そのものが、疾患感受性 SNP である場合はあまり多くなく、図 2.2 に示したように、関連解析や連鎖解析においては、それぞれ疾患と有意に関連するマーカーSNP、あるいは疾患と連鎖して受け継がれている遺伝子座（SNP 群）を見いだすことが第一の目標となる。それぞれの解析の実際については他章で取り上げられているので、ここでは詳しくは触れないが、例えば、患者-対照関連解析では、各集団でのアレル、あるいは遺伝子型頻度の差が有意である SNP（このことを疾患に関連するマーカーSNP と呼ぶ）を同定することが目的であり（図 2.6 参照）、一方、罹患同胞対連鎖解析では、罹患同胞間で共有するアレル数が、その期待値（連鎖がないことを仮定すると共有するアレル数の期待値は 1 である）から有意に大きい遺伝子座（SNP 群）を連鎖領域として見いだすことが、その目標となる。

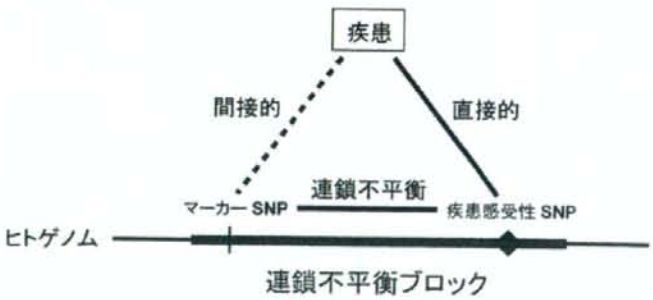
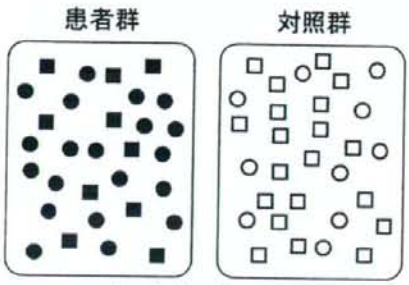


図 2.5　疾患マーカーSNP と疾患感受性 SNP との遺伝学的関係性

多くの場合において、感受性遺伝子マッピングの初期解析では疾患発症に間接的に関わるマーカーSNP が統計学的に見いだされ、そのマーカーSNP が位置する連鎖不平衡ブロック内を精査することにより、発症に直接的に関わる真の感受性 SNP が同定される。

患者−対照関連解析



	患者群	対照群
アレル		
A	20	10
B	10	20

P<0.01（χ^2 検定）

図 2.6　SNP を用いた患者−対照関連解析の一例

この例では、SNP の 2 つのアレル（A アレル、B アレル）のうち、A アレル（丸印）が患者群で統計学的に有意に高頻度出現し、疾患発症のリスクアレルである。

では、どのような SNP をタイピングすればよいのであろうか。前述したように、ゲノム全域の高密度 SNP マーカーをタイピングする場合、メーカーごとにカタログ化された SNP 群を探索することになる。GeneChip (Affymetrix 社)や Genotyping BeadChip (Illumina 社)などと呼ばれる SNP タイピング用の製品が複数のメーカーから販売されているが、搭載されているカタログ SNP のリストはメーカー間で異なることは留意すべきである（SNP タイピング用チップの特徴などは 8, 9 章を参照されたい）。他方、疾患候補ゲノム領域や、あるいは、疾患との関わりが機能面から予想される候補遺伝子群が既に明らかになっている場合、そこからの絞り込み段階（図 2.2 参照）で用いる SNP は、研究者が自ら選択しなければならない。ゲノム全域を網羅した既知 SNP についての優れた公開データベース（dbSNP など）が存在していることから、この段階で、研究対象とするゲノム領域に存在する SNP を自身で探索することは現在ではほとんど行われていない。すなわち、データベースに登録されている SNP の中から、研究の目的に適う SNP を選定し、実際にタイピングを行うことになる。「マーカーSNP」と称していることから理解できるよ

うに、対象集団であまりにも低頻度（例えば 1%程度）にしか観察されない SNP は、その目的には適わない。多くの研究では、マイナーアレル頻度が 5%以上の SNP が選択されている。また、候補ゲノム領域や候補遺伝子には多数の SNP がデータベース登録されているが、その全てを調べる必要はない。国際 HapMap プロジェクトにより明らかにされたヒトゲノム中の連鎖不平衡ブロックについての情報に基づき、特定のゲノム領域、遺伝子を代表する「タグ SNP」と呼ばれる比較的少数の SNP をマーカーとして選択し、疾患との関わりを精査すればよいこととなる。

この項の最後として、SNP のタイピング方法についても簡単に紹介しておこう。ご存知のように、DNA 塩基配列の直接決定法による SNP タイピングを含め、様々なタイピング方法が知られているが、そのほとんど全てが、核酸の塩基配列相補結合性（A-T, G-C 間のみで相補的な水素結合が生じること）をタイピングに利用し、また、アレル特異的な蛍光物質を用いることで、簡便なアレル識別を達成している。TaqMan プローブ（Applied Biosystems 社）と呼ばれるアレル特異的蛍光プローブと PCR 法とを組み合わせた SNP タイピング法（図 2.7）は、遺伝子型決定に蛍光検出器が必要であるものの、450 万以上の既知 SNP をタイピングするためのプローブ・PCR プライマーが、メーカーよりキット販売（TaqMan SNP Genotyping Assay）されており、選択した SNP を比較的容易にタイピングできることから、多くの研究で用いられている。ヒト遺伝子データベースのひとつである GeneCards (http://genecards.org/index.shtml)では、調べたい遺伝子上にカタログ化されている TaqMan SNP Genotyping Assay リストを見ることができる。図 2.7 には、TaqMan PCR タイピング例を模式的に示している。ある SNP において、T アレル（アレル 1）と C アレル（アレル 2）が存在するとする。それぞれのアレルに相補的な配列をもつ TaqMan プローブを、2 種類の異なる蛍光色素（FAM, VIC）で標識する（図 2.7 A）。ゲノム DNA と相補結合している状態（あるいは DNA と結合していない状態）では、消光物質（Q）が FAM, VIC の蛍光エネルギーを吸収するために、蛍光は検出されない（図 2.7 B1）。この SNP を含む DNA 領域を増幅するように設計された PCR プライマーと Taq DNA ポリメラーゼとを用いて PCR を行うと（図 2.7 B2）、Taq DNA ポリメラーゼの 5'ヌクレアーゼ活性により、PCR プライマーからの伸長反応に伴い、TaqMan プローブの標識蛍光色素が切断される。その結果、消光物質による蛍光吸収効果が働かなくなり、蛍光検出できるよ

うになる（図 2.7 B3）。FAM, VIC 蛍光検出後の SNP タイピング結果の一例を図 2.8 に示す。対象サンプルがアレル 1 のホモ接合体であるならば、FAM 蛍光のみを認め、一方、アレル 2 のホモ接合体では、VIC 蛍光のみが検出される。ヘテロ接合体では両者の蛍光が検出される。複数サンプルからの蛍光データを同時解析すると、図 2.8 に示すような 3 つのクラスターが観察され、これに基づき、各サンプルの遺伝子型が決定される。

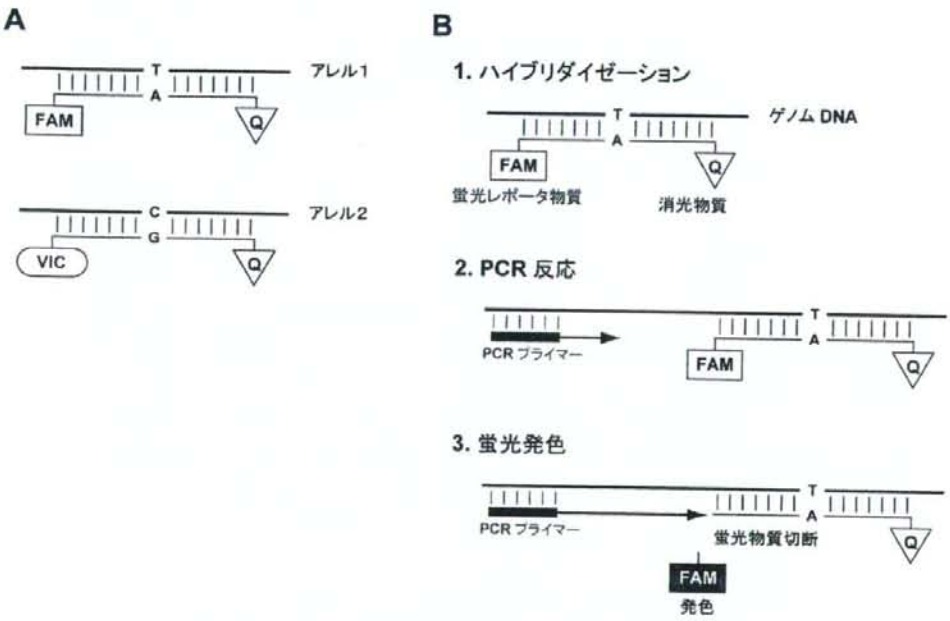


図 2.7　TaqMan PCR 法による SNP タイピング

(A) SNP の 2 つのアレル（アレル 1、アレル 2）に相補的な配列をもつ TaqMan プローブは、それぞれ FAM, VIC とよぶ蛍光物質で 3'末端標識されている。プローブ 5'末端の消光物質(Q)が FAM あるいは VIC の蛍光エネルギーを吸収するため、通常、蛍光は検出されない。(B) 2 種類の TaqMan プローブはそれぞれ完全相補であるゲノム中の標的配列にのみ結合し、PCR プライマーの伸長により分解され、蛍光を発する。

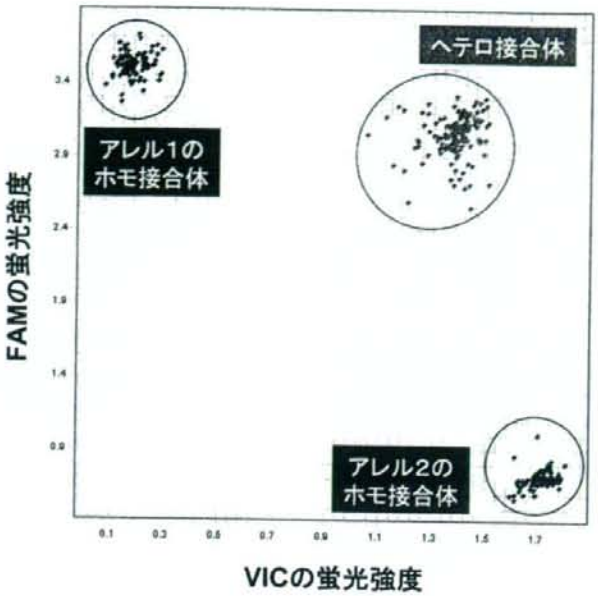


図 2.8　TaqMan PCR 法による SNP タイピングの実際

TaqMan PCR 実施後、蛍光検出器により FAM および VIC 蛍光強度を測定した時の典型例を示している。図の縦軸、横軸はそれぞれ、FAM, VIC 蛍光強度を ROX とよぶデータ補正用蛍光物質からの強度で除した値である。SNP の遺伝子型に対応した 3 つの異なるクラスターが認められ、各サンプルの遺伝子型が決定される。

TaqMan SNP タイピング法を一例として SNP タイピングの実際を示したが、どのようなタイピング方法を選択した場合でも、調べた SNP における各個人の遺伝子型情報（場合によっては、集団におけるアレル頻度情報）を取得することができる。適切な研究デザイン、SNP が選択されている場合は、得られた遺伝子型情報を基にした統計遺伝学的解析から、対象とする多因子疾患に関わるマーカー SNP（SNP 群）をこの段階で見いだすことができる。