複合遺伝性疾患の疾患関連遺伝子研究を進める上でのデータ解析補助資料集

理化学研究所 遺伝子多型研究センター

山田 亮　山本 一彦　中村 祐輔

監修・執筆・文責

理化学研究所 遺伝子多型研究センター

慢性関節リウマチ関連遺伝子研究チーム

山田 亮

分担執筆

東京大学医科学研究所 ヒトゲノム解析センター

ゲノムシークエンス解析分野

芳賀 久典

協力

理化学研究所 遺伝子多型研究センター 心筋梗塞関連遺伝子研究チーム

田中 敏博

大西 洋三

理化学研究所 遺伝子多型研究センター 遺伝子多型情報解析研究チーム

角田 達彦

平成 12 年度 東京大学医科学研究所/JST SNP ミーティング

構成員一同

序文

遺伝学の発展及び、ヒトゲノムプロジェクトの成果のおかげで、疾患関連遺伝子の研究は単一遺伝子疾患から、複合遺伝性疾患の遺伝子同定へとその目標を広げています。その複合遺伝性疾患の中でも、罹患者数が多い“遺伝性のある common disease ”の遺伝的背景の解明は、強く望まれてきた医学課題の一つです。しかしながら、common disease の関連遺伝子の同定が困難なことに異論を差しはさむ人はいませんが、その最善の戦略が何であるかもわからないままです。そんな中、我々理化学研究所 遺伝子多型研究センター(RIKEN SRC)では、SNP を用いたゲノムワイド ケース･コントロール関連解析をその手法として採用しています。この方法論に対する悲観的な意見が国内外に無いわけではありませんが、罹患同胞対法を用いた研究チームから、ローカスの絞込みをするには成功とは言いがたい LOD スコアの報告が続くなど、相対的な意味で、やはり関連解析の手法に頼らざるを得ないことが、理論面ではなく現実的な解析結果からも明らかになってきていると言えるのではないでしょうか。このような状況において、我々RIKEN SRC が common disease 関連遺伝子の同定に成功することの重要性がますます高まっています。

RIKEN SRC が発足して 1 年ほどが過ぎようとしていますが、疾患関連遺伝子研究の基盤となる日本人 SNP データベース(J-SNP)の整備が、東京大学医科学研究所/JST のゲノムワイド SNP ディスカバリープロジェクトにより順調に進み、また当センター タイピング研究･支援チームの SNP 大量高速タイピングの体制も完成が秒読みの段階になってきています。こうした中、疾患関連遺伝子を同定するために次に必要なのは、得られるデータの有効な活用です。これをテーマに平成 12 年度の 1 年間、様々な形で、思いつくままにデータの質や処理方法について考察して参りました。その整理をかねて、もしいくらかでも遺伝子多型研究センターの皆さんの役に立てば、との趣旨からまとめたのが本資料集です。

門外漢が、相当程度、我流で勉強しながらまとめたものの積み重ねなので、系統だっているとも言いかねますし、網羅的とも言えず、誤解や不十分な記載も多く残されているとは思います。しかしながら、手元に置いて自分だけのものとしておくよりは、多くの人の目に触れさせて、少しでも改善させることができれば良いし、参考にして頂けることがあれば望外、という心積もりで公開することに致しました。あえて宣伝しても良いと自負している点があるとすれば、本書の中にいくつも出てくる計算用・シミュレーション用のエクセルファイルです。残念ながら、PDF(プリントアウト)版ではすぐに使う訳にはいかない代物ですが、すべてのファイルは RIKEN SRC 内で公開されています。これらは使い方を理解するまでに若干の努力が必要かもしれませんが、わかってしまえば結構いろいろなことが分かり、面白い代物ですし、よそでは手に入りません。

最後に、このような考察をする機会を提供して下さり、また、適切なご意見を下さった、ヒトゲノム解析センター ゲノムシークエンス解析分野の中村祐輔先生を肇め、東京大学医科学研究所/JST SNP 棟ミーティングに参加された皆さんと、1 年間慢性関節リウマチの研究よりもこちらを優先させる自由を許して下さったチームリーダーの山本一彦先生、また資料等の執筆・編集・校正にあたり、非常に多くの時間を割いて下さった、東京大学医科学研究所 ヒトゲノム解析センター ゲノムシークエンス解析分野/小野薬品工業の芳賀久典さん、慢性関節リウマチ関連遺伝子研究チーム アシスタントの龍英理さんへの感謝をこの場を借りて申し上げます。

平成 13 年　春

山田 亮

目次

1　概論
1-1 SNP を用いて複合遺伝性疾患関連遺伝子解析を行うにあたっての基礎知識
1-2 SNP データベース
1-2-1 RIKEN SRC で用いる日本人一般集団の SNP
1-2-2 The International SNP Working Group による SNP 同定の現状
1-3 ゲノムワイド疾患関連 SNP スクリーニング方法の概略
1-4 参考 シミュレーションでよく用いる疾患相関 SNP の genotype 頻度の算出方法
2　遺伝性疾患であることの確認
2-1 遺伝性疾患であることの確認（λとγの推定）
2-2 $\lambda_{MZ/locus}$-γの関係(シミュレーション用エクセル)
2-3 $\lambda_{sib/locus}$-γの関係(シミュレーション用エクセル)
2-4 参考 どの程度の強さの遺伝子がいくつくらいあるか
3　関連解析
3-1 関連解析の種類
3-2 性染色体上多型の特殊性
3-2-1 性染色体上の SNP の検出、タイピング、タイピングデータに基づく疾患関連解析
3-2-2 X 染色体特異的遺伝子上の SNP 検定法
3-3 ケース・コントロール関連解析
3-3-1 ゲノムワイド ケース･コントロール関連解析を行うにあたって
3-3-2 関連解析の検定力に影響を与える因子
3-3-2-1 直接関連と間接関連
3-3-2-2 コントロール集団
3-3-2-2-1 ケース集団とコントロール集団の構成について
3-3-2-2-2 参考 性･年齢マッチ・コントロールは必要か？
3-3-2-2-3 一般集団をコントロール集団として用いることについて
3-3-2-2-4 コントロール群へのケース群の混入シミュレーション(シミュレーション用エクセル)
3-3-2-3 連鎖不平衡(Linkage Disequilibrium)
3-3-2-3-1 連鎖不平衡とは
3-3-2-3-2 D'計算(シミュレーション用エクセル)
3-3-2-3-3 真のローカスに近接する SNP の、真のローカスとの LD の強さとアレル頻度の乖離が検定力に及ぼす影響について
3-3-2-3-4 LD と χ^2 値の関係
3-3-2-3-5 参考 連鎖不平衡の実際及び間接関連検出用の SNP マーカーと真のローカスとの関係について
3-3-2-3-6 参考 いろいろな linkage disequilibrium 指標とその挙動
3-3-2-3-7 参考 組み換えと組み換え率
3-3-2-4 集団間の民族学的遺伝的差（階層化を含む、その問題点と存在の有無の検定）

3-3-3 関連解析に必要な検体数と第１種過誤(偽陽性率)と第２種過誤(偽陰性率)
3-3-3-1 第１種過誤(偽陽性率)と第２種過誤(偽陰性率)
3-3-3-2 必要検体数(シミュレーション用エクセル)
3-3-3-3 β過誤率(シミュレーション用エクセル)
3-3-4 1 SNP ケースコントロール関連解析用エクセル
(χ^2検定、HWE 検定、OR(95%CI)を一括出力)
3-4 ケース内解析
3-4-1 Hardy-Weinberg 平衡検定
3-4-1-1 Hardy-Weinberg 平衡(HWE)とは
3-4-1-2 Hardy-Weinberg 不平衡の程度の評価法とその原因解明の手順
3-4-1-3 参考 Hardy-Weinberg 平衡検定 p 値分布が、homozygote 過剰の場合に統計的有意 p 値を示しがちなことに関する考察
3-4-1-4 1 SNP data の Hardy-Weinberg 平衡評価 2 方法(エクセル)
3-5 複数点解析
3-5-1 マッピング(有意相関 SNP が検出された場合の linkage disequilibrium mapping、Hardy-Weinberg 不平衡 mapping の手順)
3-5-2 ハプロタイプ
3-5-2-1 ハプロタイプとは
3-5-2-2 参考 EM-algorithm による haplotype 頻度の推定
3-5-2-3 2 SNP Haplotype 頻度推定(エクセル)
3-5-2-4 3 SNP Haplotype 頻度推定(エクセル)
3-5-2-5 4 SNP Haplotype 頻度推定(エクセル)
3-5-3 連鎖していない複数 SNP の組合せ解析
4 タイピングデータの質の管理（有意差検定を用いて）

1
概論

1-1 SNP を用いて複合遺伝性疾患関連遺伝子解析を行うにあたっての基礎知識

＜複合遺伝性疾患(Complex Genetic Diseases)＞

その特徴

1. 多遺伝因子の関与
2. 多環境因子の関与
3. 個々の因子の関与が小さい
4. 複数の遺伝・環境両因子が相互に影響を及ぼし合いながら Phenotype の出現を決定する
5. 低い Penetrance
6. 高い Phenocopy
7. Quantitative Genetic Trait にみられるように Phenotype が連続形質であるものを含む

＜Phenocopy と Penetrance＞

SNP によるゲノムワイド研究の対象となる複合遺伝性疾患の原因遺伝因子は、その因子の有無が、疾患の存在(Phenotype)と 1 対 1 対応することは決してない。そうではなく、ある遺伝因子を保有することはその他の要因と相互作用を及ぼし合いながら、当該疾患の発症しやすさに影響を与えている。このことを簡単に言い換えると、罹患者の中に当該遺伝因子を持たないことがあり、また、当該遺伝因子を持つものの発症しないことがあるということである。前者の現象を“ Phenocopy が存在する ”、後者を“ 浸透率(Penetrance)が 1 に満たない ”と表現する。

＜疾患と遺伝因子との関連の強さを表す因子(再発危険率と相対危険度)＞

再発危険率：発端者とある遺伝関係にある個人を考える。その個人が発症する確率が、一般人の何倍であるかを表現した数値のこと。発端者との遺伝関係により一卵性双生児再発危険率、同胞再発危険率、子再発危険率などと呼ばれる。略号としては通常λ(ラムダ)を用いる。
λ_{MZ}、λ_{sib}、λ_{off}などと表記される。これらの値は、疫学研究によって推定される。

相対危険度：ある条件を保有する個人が発症する確率が、その条件を持たない個人の発症確率の何倍であるかを表現した数値のこと。ここでいう、ある条件に含まれるものには、“ ある遺伝因子を保有する ”というような因子の場合もあれば、環境因子の場合もある。
通常γ(ガンマ)をその略号として用いる。

Genotypic risk ratio：ある疾患ローカスの疾患アレルをヘテロで保有するときの相対危険度のことである。上記の遺伝子保有に関するγに相当する。最も一般的なモデルでは、この疾患アレルをホモで持つ場合の相対危険度はγ^2とされる。

Genotypic relative risk：疾患ローカスにおいては、genotype 別に相対危険度が存在する。2 アレルの SNP の場合には、genotype は 2 種類のホモと 1 種類のヘテロの 3 種類である。この場合、疾患アレルをヘテロで保有する場合と、疾患アレルをホモで保有する場合のそれぞれに相対危険度がある。genotypic risk ratio

で与えたモデルではホモ相対危険度＝γ^2、ヘテロ相対危険度＝γであった。
その他によく用いるモデルとして、優性遺伝形式の場合の
　　ホモ相対危険度＝γ^2、ヘテロ相対危険度＝γ
と、劣性遺伝形式の場合の
　　ホモ相対危険度＝γ^2、ヘテロ相対危険度＝1
とがある。このように、個々の genotype に定義された相対危険度のことを genotypic relative risk と呼ぶ。

＜SNP(Single Nucleotide Polymorphism; 1塩基多型)＞

ヒトゲノム全体に最も高密度に分布する多型であり、数百から千塩基対に1個程度の頻度で存在する。アレルの数は通常2つ(Biallelic, Diallelic)である。

○複合遺伝性疾患関連遺伝子同定研究において SNP と疾患の関連を検出するときには

(1) 多型そのものが Phenotype(疾患)の決定と因果関係にある場合(直接関連)
(2) Phenotype を決定する遺伝因子を検出するためのマーカーとしての役割を果たす場合(間接関連)

の2つの場合が想定される。これに関して、詳しくは“3-3-2-1 直接関連と間接関連”の項を参照のこと。

○SNP と他の多型(RFLP、VNTR、マイクロサテライト)との相違点

(1) SNP は他の多型に比べはるかに稠密に分布すること
(2) SNP の多型情報量は他の多型の情報量に比べて小さいこと
　多型情報量とは
　　(a)アレル数
　　(b)heterozygote の個体の占める割合
　の2ファクターによって決まる。
　SNP は Biallelic なので(a)、(b)ともに小さい
(3) SNP の genotyping assay 法は他の多型の assay 法に比べ簡便であること

○SNP を用いたゲノムワイド ケース・コントロール関連解析について

複合遺伝性疾患は、単一遺伝子疾患の原因遺伝子の同定と同じ方法論ではその原因遺伝子の決定が困難であると考えられている。その主な理由は、個々の遺伝因子の寄与分が小さいことによる。したがって、単一遺伝子疾患の原因遺伝子の同定に大きな役割を果たした、連鎖解析(Linkage analysis)とは異なる方法論として、ゲノムワイド ケース・コントロール関連解析が有力視されている。この方法論の理論的根拠は以下の通りである。

1. 疾患関連遺伝因子はケース・コントロール間でその分布に差が存在する。

2. 疾患関連遺伝因子の周辺には、連鎖不平衡が存在する。
3. 連鎖不平衡内に存在する遺伝子多型は疾患関連遺伝因子と連鎖している。
4. 連鎖不平衡内の遺伝子多型のアレルにはケース・コントロール間で分布差が存在する。
5. ケース・コントロール間で分布差のある SNP と疾患との関連は検定可能である。

1-2 SNP データベース

1-2-1 RIKEN SRC で用いる日本人一般集団の SNP

＜SNP 解析領域＞

RIKEN SRC でゲノムワイドスクリーニングに用いる日本人一般集団の SNP は、東京大学医科学研究所/JST SNP 同定プロジェクトにおいて同定した SNP である。このプロジェクトが対象としている解析領域は、ヒトゲノム全体に分布する遺伝子及び遺伝子制御領域である。すなわち、第１エクソンから最終エクソンまでの領域(イントロンも含む)とプロモーター領域である。エクソン、イントロン及びプロモーター領域の定義は、GenBank 配列注釈に構造情報の記載があれば、それをそのまま使用し、ない場合は GENSCAN による構造予測や mRNA/cDNA/Unigene レコードに対する相同性検索によって、遺伝子構造を推定した結果を用いている。

＜SNP 同定法＞

解析領域を PCR 増幅した後、ダイレクトシークエンスを行い、SNP を検出している。よって解析領域にプライマーを設計する必要がある。プライマー設計には、繰り返し配列をある基準以上含まないよう設計し、またプライマーの GC 含量、Tm 値などが至適条件になるよう設計している。よって、繰り返し配列や GC 含量等の影響により、設計できない領域が存在するため、ヒト全遺伝子(制御)領域を網羅している訳ではないが、現在報告されているヒトゲノム配列全てに対して設計を試みている。また、PCR 反応で増幅されなかった場合も、SNP 同定することができないし、シークエンスが読めていない場合も、その領域に対して SNP 同定することが出来ない。

また、常染色体と性染色体とを区別して解析する必要があるため、検体は XX(女性)のみを用いており、Y 染色体は SNP 同定の対象から外されている。性染色体に関する SNP 同定の詳細は“3-2 性染色体上多型の特殊性”参照のこと。

検体数は、日本人一般集団 24 人、すなわち 48 クロモソームである。PCR 増幅不良やシークエンス不良による読みこぼしを考慮しても、1-2-2 で解説している The SNP Consortium による解析クロモソーム数(2～5)や The International Human Genome Sequencing Consortium による解析クロモソーム数(2～)に比べて、圧倒的に多い。このことは、我々の方が、解析した領域中に存在する SNP を取りこぼす確率が低いことを意味している。

＜これまでに同定した SNP の概要＞

2001 年 3 月 13 日時点で、シークエンス解析済み 72,724,315 塩基中、

エクソン	16,000,766 塩基(22.0%)
イントロン	56,526,476 塩基(77.7%)
プロモーター領域	197,073 塩基(0.3%)

を解析した。そのうち、71,639,046 塩基について SNP 検出を終了し、73,549 SNP 同定した(欠失・挿入型多型を含む。その割合は約 6.9%である)。約 1,050 塩基に 1SNP(欠失・挿入を除いて計算)の割合であるが、解析塩基数にはプライマー領域やシークエンスが読めていない領域についても含んでいるため、日本人の SNP 頻度はもう少し高頻度で存在すると考えられる。

インターネット上で公開している(http：//snp.ims.u-tokyo.ac.jp)SNP 数は 34,064 SNP で、構造領域別 SNP 数と染色体別 SNP 数を以下の表 1、表 2 に掲げる。

表 1. 構造領域別 SNP 数

構造領域	SNP 数
プロモーター領域	426
エクソン	5,955
イントロン	27,683
合計	34,064

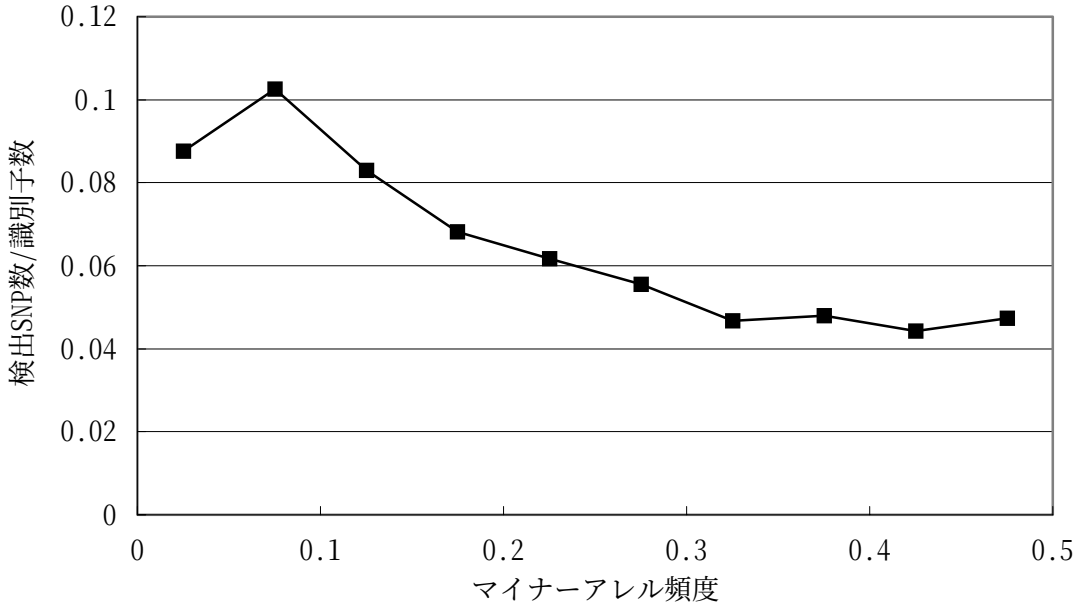
表 2. 染色体別 SNP 数

染色体番号	1	2	3	4	5	6	7	8	9	10	11	12	13
SNP 数	2372	1057	1159	766	898	2676	2780	640	625	555	1399	1704	228
染色体番号	14	15	16	17	18	19	20	21	22	X	Y	不明	合計
SNP 数	1265	569	1462	1037	326	1603	1439	1172	3191	1422	43	3676	34064

Finished sequence から先に解析を始めているため、染色体別 SNP 数にはばらつきがあるが、今後は遺伝子領域の長さに比例して SNP 数が増加するものと予想される。

最後に、2001 年 2 月～4 月期におけるマイナーアレル頻度別検出 SNP 数を図 1 に示す。
ヒトゲノム上の SNP に関して、低頻度のものほど、数多く存在していることが示唆されている(“1-2-2 The International SNP Working Group による SNP 同定の現状”を参照)が、この点と我々の結果は一致している。

図 1. マイナーアレル頻度別検出 SNP 数



検出 SNP 数をマイナーアレル頻度別に 10 段階(0～5/5～10/……/45～50%)で集計した。検出 SNP 数は 1 識別子当たり(＝約 500bp)に換算している。

なお、NCBI の dbSNP には、2,558,364 SNP の登録があるが、これらには、複数の塩基配列解読結果の照合結果によって抽出した SNP が大部分を占め、偽陽性がかなり含まれていると予想されている。また、あらゆる人種の SNP を含んでいるため、日本人には存在しない SNP も登録されている。SNP 同定プロジェクトの日本人一般集団 SNP と dbSNP との一致率は、1201/10131 件(11.9%)(2000/9/25 現在で JST 調査結果による)であった。

1-2-2 The International SNP Working Group による SNP 同定の現状

2001 年 2 月に Nature 誌にヒトゲノム解読結果が報告され、それと同時にヒトゲノム上の SNP に関しても 142 万個同定し、マッピングしたとの報告[1)]が The International SNP Working Group(TISWG)からなされた。

彼らは、以下の 3 つの方法により、候補 SNP の同定を行った。

1. The SNP Consortium(TSC)による Shotgun 法[#]を用いた民族学的に異なる 24 個人[##]ゲノム DNA を用いての SNP 同定(1,023,950 個)

   [#]適当な制限酵素でゲノム DNA を切断後、アガロースゲル上でサイズ分画し、プラスミドベクターにサブクローニング後、ベクタープライマーを用いてランダムシークエンスした。

   [##]領域ごとにシークエンスされたクロモソーム数は、ランダムに決まる方法であり、実際に 1 領域あたり解読したクロモソーム数は 1 領域あたり 2〜5 である。
2. The International Human Genome Sequencing Consortium(TIHGSC)による複数の BAC、PAC クローンのシークエンス結果から配列の違い[###]を見つけることによる同定(971,077 個)

   [###]すなわち、最低 2 クロモソーム以上の解析となる。1 領域あたりの解析クロモソームについての見解はないが、それほど多いとは予想されない。
3. EST 重複配列などから同定(候補 SNP の約 5%)

以上、総計 2,067,476 個の候補 SNP の内、1,433,393 個がゲノム上の 1 ローカスに、*in silico* でマッピングされた。BAC 領域の重複解析等により、同一の SNP が複数個カウントされていることを差し引くと、結局のところ、1,419,190 個が重複のない SNP として同定され、平均 1.91kb 間隔で SNP がヒトゲノム上にマッピングされた。

TSC では、同定した SNP の一部を用いて、SNP 同定に用いた 24 個人に SNP が存在するか、再確認した。ゲノム DNA を PCR で増幅後、ダイレクトシークエンスしたところ、95%が真であった。また、TSC と TIHGSC で同定した SNP の一部を、3 民族で SNP が存在するか再確認したところ、10%以上のマイナーアレル頻度を有する SNP を検出できる系で、82%が少なくとも 1 民族以上で真の SNP であった。

以上より、Kruglyak らは真の SNP 率は 82〜95%の間であろうと推測し、真の SNP 数は 1,419,190 個中、1,160,000〜1,350,000 個と見積もった[2)]。また、集団遺伝学の古典的中立仮説から 1%以上のアレル頻度で存在する SNP は、ヒトゲノム 32 億塩基中 1,100 万個と推定される(図 1、表 1)ので、TISWG による真の SNP 数 1,160,000〜1,350,000 個は、ヒトゲノム全 SNP 数の 11〜12%を同定したに過ぎないと考察した[2)](表 1)。

また、TISWG は RefSeq を用いた mRNA 上にマップされた SNP から、142 万個中、

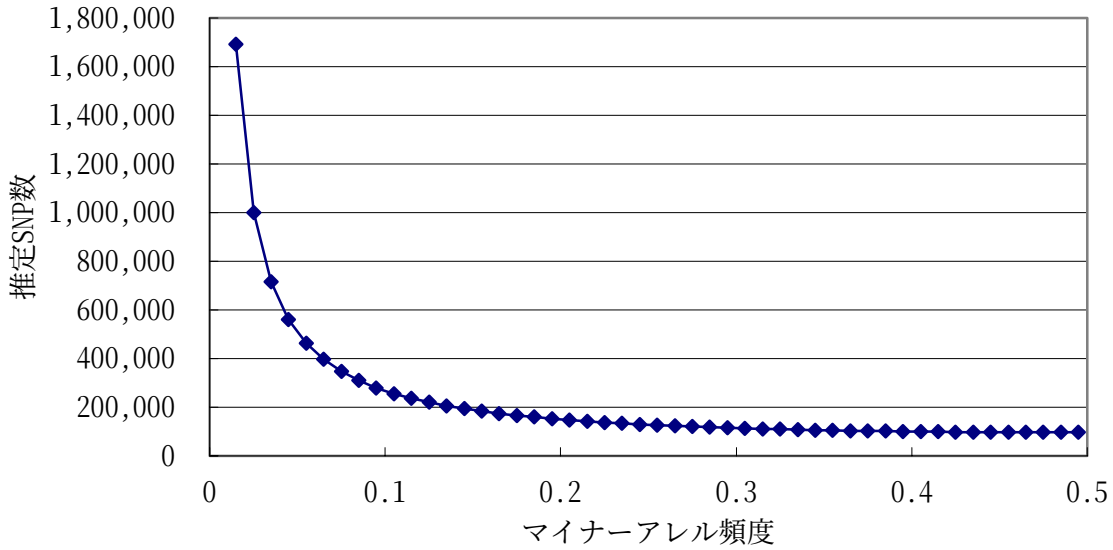
1. 60,000 SNP がエクソン領域内
2. 93%の遺伝子座が少なくとも 1 SNP 含む
3. 98%の遺伝子座は 5kb 以内に SNP を有する
4. 85%のエクソンは 5kb 以内に SNP を有する

と推定した[1)]。

なお、東京大学医科学研究所/JST SNP との大きな違いは、TISWG は遺伝子領域外に存在する gSNP が

データの大半を占めるが、東京大学医科学研究所/JST SNP は遺伝子領域内に存在する SNP のみであるという点である。

図 1.ヒトゲノム上のマイナーアレル頻度別推定 SNP 数



マイナーアレル頻度を f とすると、ヒトゲノム上に f 以上の頻度で存在する SNP 数 S_f は

$S_f = S_2 \times \ln((1-f)/f)$

で記述される。

2 本のハプロイドゲノムを解析した結果、1331 塩基に 1 SNP 同定されたこと[1)]より、

S_2＝32 億(ヒトゲノム総塩基対)/1331

を用いて、1%毎の推定 SNP 数を描いたグラフが図 1 である。

表 1. ヒトゲノム中の推定 SNP 数と TISWG によって同定された SNP 数の比較

最小ｱﾚﾙ頻度	推定 SNP 数	推定 SNP 頻度 (1SNP/Nbp の N で表示)	推定 SNP 数に対する TISWG が同定した SNP 数の占める割合
1%	11,000,000	290bp	11～12%
5%	7,100,000	450bp	15～17%
10%	5,300,000	600bp	18～20%
20%	3,300,000	960bp	21～25%
30%	2,000,000	1,570bp	23～27%
40%	970,000	3,280bp	24～28%

参考文献：

1) The International SNP Working Group, Nature 409: 928-33 (2001)

2) L Kruglyak et al, Nature Genetics 27: 234-6, (2001)

1-3　ゲノムワイド疾患関連 SNP スクリーニング方法の概略

1. 検体：心筋梗塞(MI)、慢性関節リウマチ(RA)、気管支喘息(AS)、アトピー(AT)、膝変形性関節症(KO)、糖尿病性腎症(DN)、糖尿病性網膜症(DR)、股関節変形性関節症(HO)
   の 8 疾患群の検体のデータが得られる。
2. 各疾患群のサンプル数は 94 である。
   8 疾患群合計のサンプル数は 8×94＝752 である。
3. それぞれの検体はそれぞれの疾患チーム・疾患担当者が責任を持って採集したサンプルであり、その疾患 phenotype については問題がないと考えられる。
4. 収集母集団は日本国内(おそらく日本人)という点で共通である。収集地域は疾患によりばらつきがあり、

MI	関西
RA	関東
AS	関西
AT	関西
KO	関東
DN	関東・関西
DR	関東・関西
HO	関東

   を中心としたサンプリングとなっている。
5. マーカーSNP：東京大学医科学研究所/JST による日本人一般集団検体を用いて同定された SNP を使用している。
6. タイピング法：Multiplex-PCR によって産物をテンプレートとした蛍光 2 色インベーダーアッセイ法。
   ABI9700 による multiplex PCR
   384-well カード式･恒温水槽インベーダー反応
   RAVEN による反応終了時蛍光強度測定
7. 1 SNP あたりのデータは各疾患群別にジェノタイプ別観測数として出力される。

これらの処理に関することがこの資料集の内容である。

＜1-4 参考 シミュレーションで用いる疾患相関 SNP の genotype 頻度の算出方法＞

この解析資料集の検定・シミュレーション・各種計算においては疾患と相関のある SNP のアレル頻度と genotype 別の相対危険度から一般集団・疾患集団・非疾患集団の genotpype 頻度を算出する機会が非常に多い。その基本方法を参考として以下に掲載する。

罹患率 d を入力する （0＜d＜1）
（例）d＝0.01
SNP の疾患アレル頻度を入力する （0＜p＜1）
（例）p＝0.3
疾患アレルの遺伝形式を入力する
g：疾患アレルホモ接合体の相対危険度 （g＞1）
ff：疾患アレルヘテロ接合体の相対危険度を決める因子 （0≦ff≦1）
ff＝0 のとき、ヘテロ接合体の相対危険度は 1、つまり危険度の上昇はなし
ff＝1 のとき、ヘテロ接合体の相対危険度は g に等しい、つまりホモ接合体と危険度は等しく、このアレルは優性形式
0＜ff＜1 のときは両者の中間、一般的な中間型モデルを調べたいときは 0.5 を入力する。

以下に計算式を示す。
s：集団全体の genotype 頻度
u：ケース集団の genotype 頻度
e：非ケース集団の genotype 頻度
DD：疾患アレルホモ
DN：ヘテロ
NN：非疾患アレルホモ

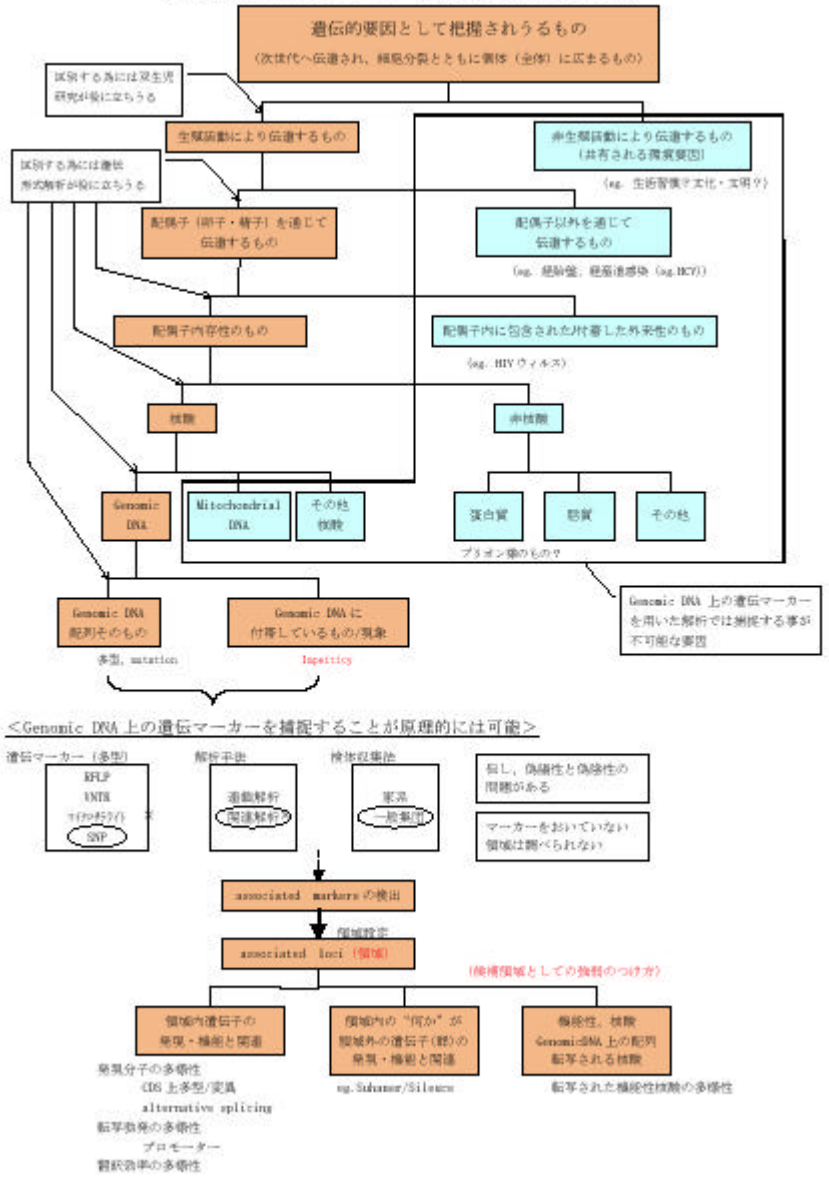
$s_{DD}=p^2$
$s_{DN}=2p(1-p)$
$s_{NN}=(1-p)^2$
$U=gp^2+g^{ff}\times 2p(1-p)+(1-p)^2$
$u_{DD}=gp^2/U$
$u_{DN}=g^{ff}\times 2p(1-p)/U$
$u_{NN}=(1-p)^2/U$
$e_{DD}=(s_{DD}-d\times u_{DD})/(1-d)$
$e_{DN}=(s_{DN}-d\times u_{DN})/(1-d)$
$e_{NN}=(s_{NN}-d\times u_{NN})/(1-d)$

1-5 網羅的遺伝的要因探索の落とし穴

ゲノムワイドに疾患関連遺伝子を解析していると、ついつい、全ての遺伝的要因について網羅的に解析を行っているかのような錯覚に陥る危険性があります。ゲノムワイド SNPs を用いたケース・コントロール関連解析の手法にも、当然のことながら様々な落とし穴があります。この項では、それらについて体系的に再確認することを目的としています。

遺伝性があるとしてとらえられ得る要因と、genomic DNA 上の遺伝マーカーを用いた解析で補足される遺伝的要因(genomic DNA 上の関連遺伝マーカーの捕捉)

図１．Genomic DNA 上のマーカーに関連する遺伝的要因の捕捉



2
遺伝性疾患であることの確認

2-1 遺伝性疾患であることの確認(λとγの推定)

＜遺伝性の根拠の確認(λの推定)＞

そもそも研究対象としている疾患が遺伝性であることが確認されなくては、研究対象とする必然性がない。遺伝性疾患であるということは、罹患者の血縁者が一般人よりも罹患しやすいということである。その場合、

λ_{MZ} = 一卵性双生児再発危険率

λ_{sib} = 同胞再発危険率

が、1 より大きくなる。これらの数値が求められない疾患の中にも、遺伝性のある疾患は存在すると思われるが、それを通常の解析で検出することは検出力の観点から、非常に困難であると予想される。

具体的には、以下の要領でλ_{MZ}、λ_{sib}を算出する。

疫学研究結果から以下の表の、空欄の情報を引き出す。

表 1

	遺伝的一致率	再発危険率・罹患率	λの算出
一卵性双生児研究	1	ア	λ_{MZ} = ア/ウ
同胞(二卵性双生児を含む)研究	0.25	イ	λ_{sib} = イ/ウ
一般集団研究	～0.001	ウ	N/A

以上により、対象疾患が遺伝性であることの根拠は得られたが、これだけでは対象疾患の原因遺伝子がどのくらいの強さで罹患しやすさ(susceptibility)を上昇させるかに関しての情報は得られない。それに関する情報を与えるのはgenotypic risk ratio(γ)と呼ばれる数値であったり、genotypic relative risk と呼ばれる数値であったりする。この genotypic risk ratio や genotypic relative risk がわかると、その遺伝子の検出の容易さ(困難さ)に関しても情報が得られ、解析に必要な検体数をはじめ、解析計画における様々な予測が可能となる。しかしながら、λからγを算出することは単純ではなく、疾患と疾患原因遺伝子との関係についていくつかの仮定をしてやらなければならない。

なお、γの数値に関しては、λから推定する以前に、ある程度の予想値がある。というのは、γの値がある程度以上であれば、罹患同胞対法などにより、比較的容易に原因遺伝子ローカスが見つかるはずであるという理論的根拠があるからである。従って、罹患同胞対法などが原因遺伝子ローカスの同定に失敗したような疾患では、それよりもγの値は小さいと予想されている。このようなγの値は、疾患アレル頻度が極端に小さくない場合には 4 程度と見積もられている(γ=4 は罹患同胞対 300 程度で十分に検出されるはずの強さである)。

次にこのλからγを算出する手順について説明する。

＜Genotypic risk ratio(γ)を推定せよ(λをγに変換せよ)＞

Genotypic risk ratio(γ)は個々の遺伝子・遺伝ローカスごとに決まる値である。一方、複数の遺伝子が関係している場合のλは複数のγの、作用の総体として決まる値である。従って、ある多遺伝子疾患のλを個々の遺伝子・遺伝ローカスのγに分解してやるためには、いくつの遺伝子・遺伝ローカスが相互にどのように作用しながら疾患 susceptibility を決定しているかを知らなくてはならない。しかしながら、原因遺伝子の数や相互関係について予め情報が得られているはずはない。

そこで、次のような手順を踏むことによって、おおまかな推定をする。まず、個々の遺伝子・遺伝ローカスのγとそのアレル頻度・genotypic relative risk の型(γ^{2}型、優性遺伝型・劣性遺伝型)を決め、その１遺伝子・遺伝ローカスが単一で及ぼすλの値を算出する。これを$\lambda_{MZ/locus}$、$\lambda_{sib/locus}$と表記する。

ここで genotypic relative risk の型について説明する。genotypic risk ratio は個々のローカスについて１つだけ与えられる。しかし、ローカスによって、homozygote の genotypic relative risk と heterozygote の genotypic relative risk との関係は異なる。その genotypic risk ratio と genotypic relative risk との関係を表すのが、genotypic relative risk の型である。

最もよく用いられる定義の仕方は以下のとおりであり、本資料集ではγ^{2}型と呼ぶこととする。

「今、ある 2 アレル ローカスの genotypic risk ratio がγであるとする。この場合、このローカスのヘテロ個体の罹患しやすさはこの疾患アレルを持たない場合のγ倍である。さらにこの疾患アレルをホモで持つ個体の罹患しやすさはγ^{2}である。」

その他の典型的な genotypic relative risk の型を優性遺伝型、劣性遺伝型と呼ぶこととし、それぞれの genotypic relative risk は以下の表 2 の通りとする。ただし、genotypic relative risk とは疾患アレルを持たない個体(疾患アレル非保有個体)に対する疾患アレル ホモおよびヘテロの個体の relative risk のことである。

表 2

genotypic relative risk の型	genotype		
	疾患アレル ホモ	疾患アレル ヘテロ	疾患アレル非保有
γ^{2}型	γ×γ	γ	1
優性遺伝型	γ	γ	1
劣性遺伝型	γ	1	1

$\lambda_{MZ/locus}$とγとの関係は、疾患アレル頻度・genotypic relative risk の型により決定される。その関係をグラフにしたのが図 1～図 4 である。

図 1. $\lambda_{MZ/locus}$と疾患アレル頻度 p 及びγとの関係（γ^2型）

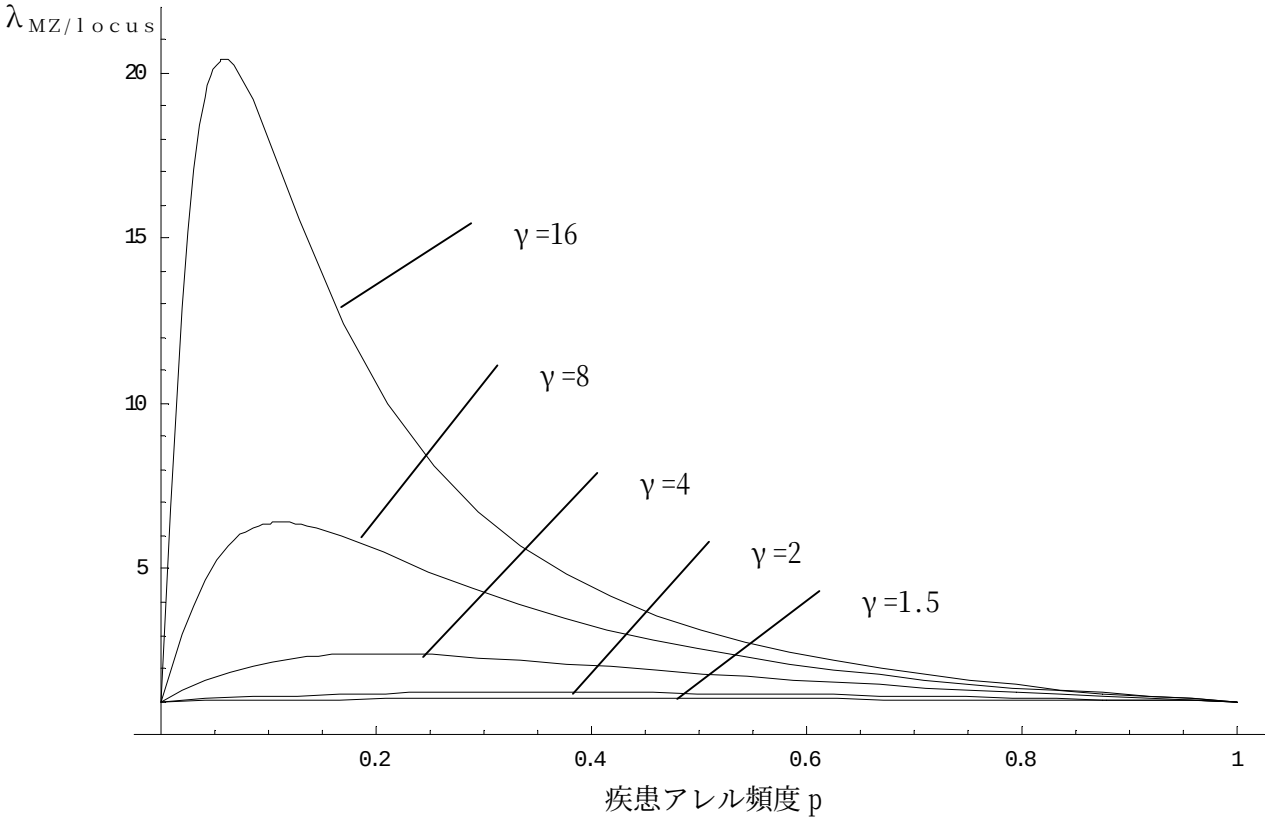


図 2. $\lambda_{MZ/locus}$と疾患アレル頻度 p 及びγとの関係（優性遺伝型）

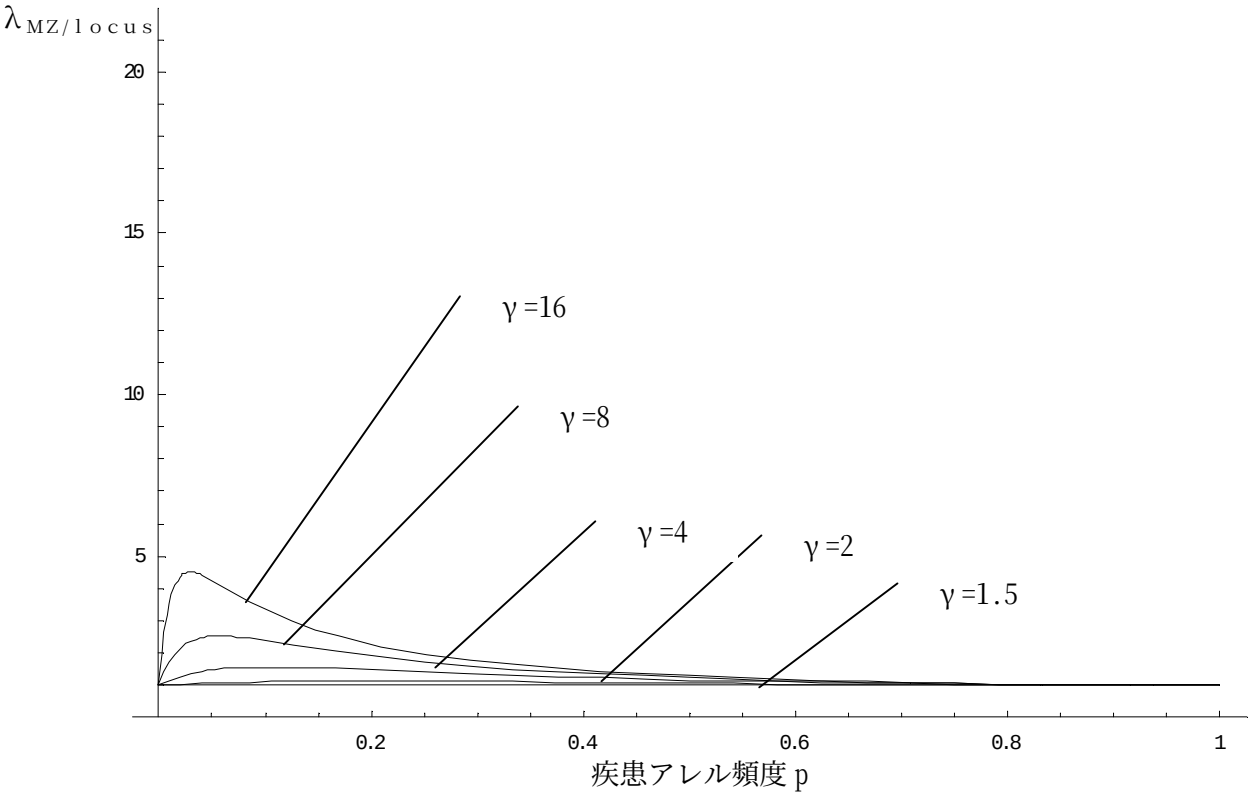


図 3. $\lambda_{MZ/locus}$と疾患アレル頻度 p 及びγとの関係（劣性遺伝型）

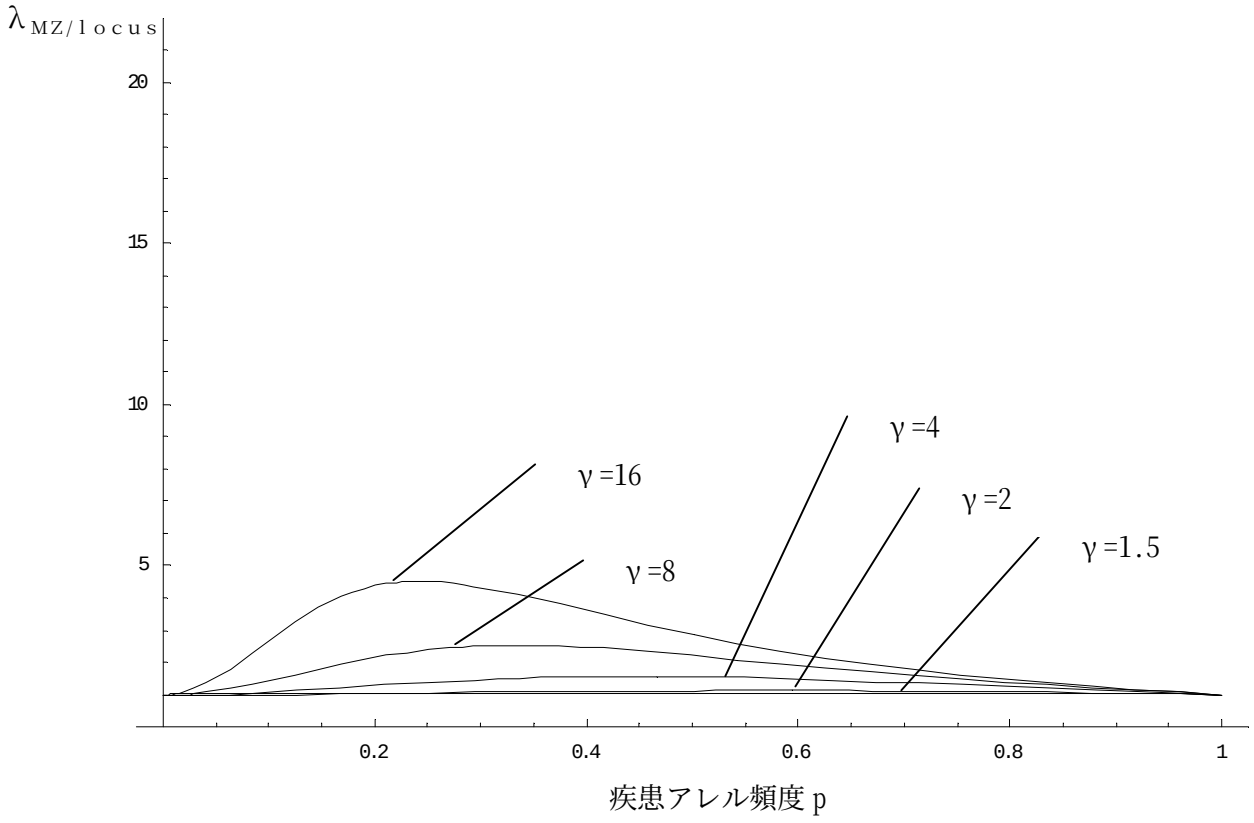
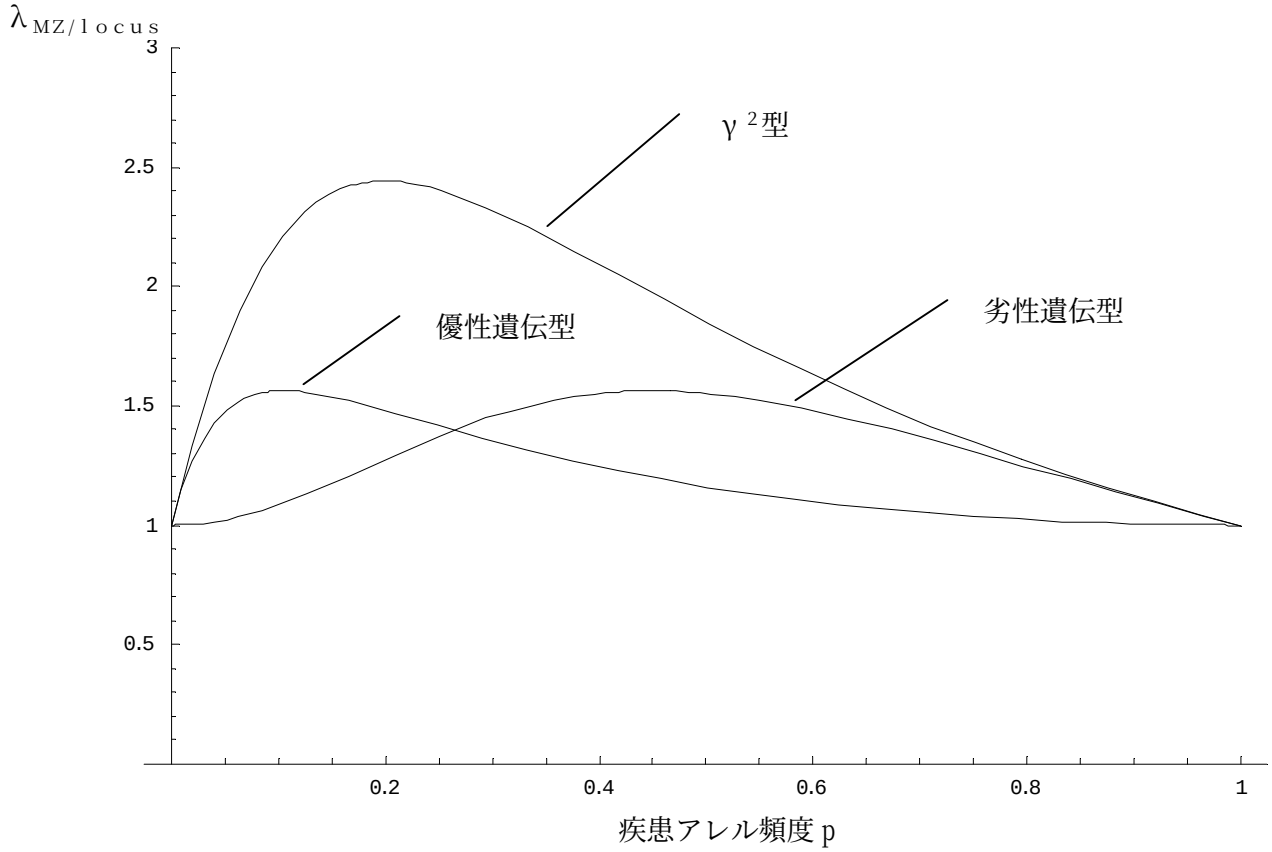


図 4. $\lambda_{MZ/locus}$の genotypic relative risk 型による違い（γ =4）



このグラフを描く基となる数式は以下の通りである。

疾患アレル頻度 p

q=1-p

とおいて、その関係を数式で表すと、

γ^2型

$$\lambda_{MZ/locus} = (\gamma^4 p^4 + 2\gamma^4 pq + q^4)/(\gamma^2 p^2 + 2\gamma pq + q^2)^2$$

優性遺伝型

$$\lambda_{MZ/locus} = (\gamma^2 p^2 + 2\gamma^2 pq + q^2)/(\gamma p^2 + 2\gamma pq + q^2)^2$$

劣性遺伝型

$$\lambda_{MZ/locus} = (\gamma^2 p^2 + 2pq + q^2)/(\gamma p^2 + 2pq + q^2)^2$$

さらに、シミュレーションを目的としてエクセルファイル “ $\lambda_{MZ/locus}$-γ ” を使用することができる。

このファイルは、任意のγとpに対して$\lambda_{MZ/locus}$が得られる。

$\lambda_{sib/locus}$とγとの関係も同様に、以下の数式から求められる(図5〜図8)。

エクセルファイル“ $\lambda_{sib/locus}$-γ ”を用いて計算することもできる。

図5. $\lambda_{sib/locus}$と疾患ｱﾚﾙ頻度p及びγとの関係(γ^2型)

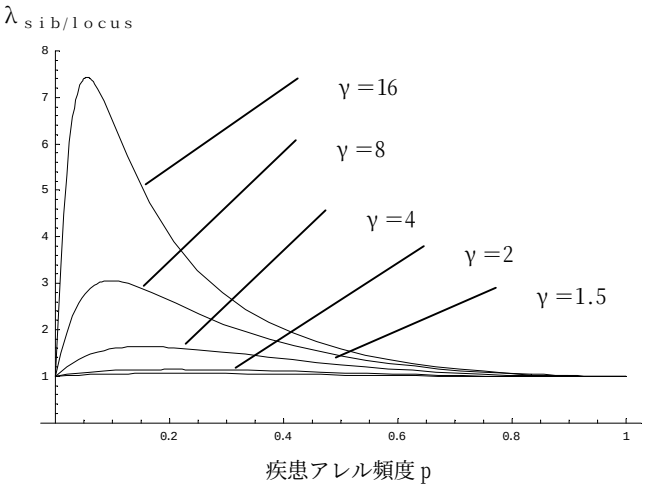


図6. $\lambda_{sib/locus}$と疾患ｱﾚﾙ頻度p及びγとの関係(優性遺伝型)

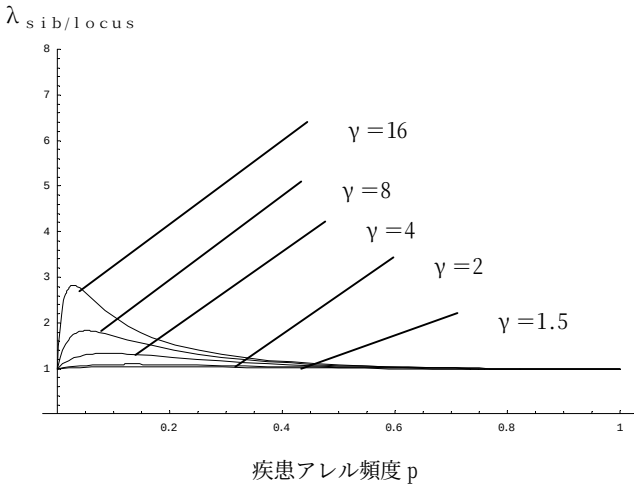


図 7. $\lambda_{sib/locus}$と疾患ｱﾚﾙ頻度 p 及びγとの関係(劣性遺伝型)

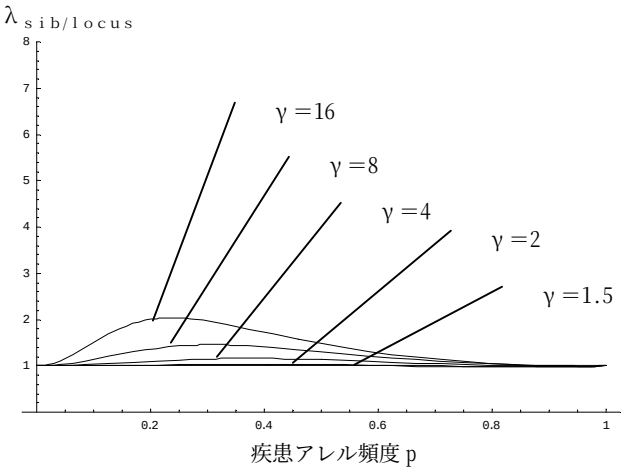
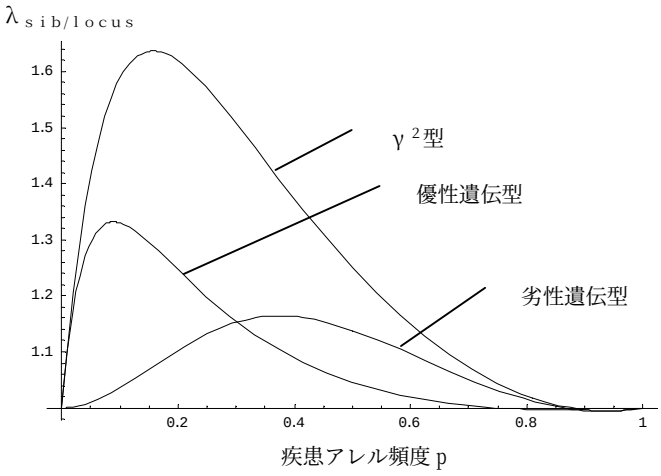


図 8. $\lambda_{sib/locus}$の genotypic relative risk 型による違い(γ=4)



ローカスごとのλの値はかなり小さいことがわかる。言い換えれば、疫学調査から得られるλ_{MZ}、λ_{sib}の値が小さくとも、原因遺伝子の数が少なければ、個々のローカスの、γの値は相当大きいわけである。

以上より、γと$\lambda_{sib/locus}$の関係がわかったが、では、あるλ_{MZ}、λ_{sib}が知られている場合に、1ローカスごとのλはどのくらいと考えるべきであろうか？ 言い換えれば、どのくらいのγのローカスがいくつくらいあると想定すればよいのであろうか？ これについては、単純化したモデルによるシミュレーションの結果を“2-4 参考 どのくらいの強さの遺伝子がいくつくらいあるか”に記した。

γ^2型

$$\lambda_{sib/locus} = \frac{8p^8+36p^7q+24\varepsilon p^7q+8\varepsilon^2p^7q+70p^6q^2+100\varepsilon p^6q^2+66\varepsilon^2p^6q^2+8\varepsilon^3p^6q^2+2\varepsilon^4p^6q^2+75p^5q^3+170\varepsilon p^5q^3+185\varepsilon^2p^5q^3+54\varepsilon^3p^5q^3+15\varepsilon^4p^5q^3+46p^4q^4+139\varepsilon p^4q^4+256\varepsilon^2p^4q^4+139\varepsilon^3p^4q^4+46\varepsilon^4p^4q^4+15p^3q^5+54\varepsilon p^3q^5+185\varepsilon^2p^3q^5+170\varepsilon^3p^3q^5+75\varepsilon^4p^3q^5+2p^2q^6+8\varepsilon p^2q^6+66\varepsilon^2p^2q^6+100\varepsilon^3p^2q^5+70\varepsilon^4p^2q^6+8\varepsilon^2pq^7+24\varepsilon^3pq^7+36\varepsilon^4pq^7+8\varepsilon^4q^8}{(4p^2+4pq+q^2)(2p^2+3pq+2q^2)(p^2+4pq+4q^2)(p^2+2\varepsilon pq+\varepsilon^2q^2)^2}$$

優性遺伝型

$$\lambda_{sib/locus} = \frac{8p^8+36p^7q+24\varepsilon p^7q+8\varepsilon^2p^7q+70p^6q^2+110\varepsilon p^6q^2+66\varepsilon^2p^6q^2+75p^5q^3+201\varepsilon p^5q^3+223\varepsilon^2p^5q^3+46p^4q^4+183\varepsilon p^4q^4+397\varepsilon^2p^4q^4+15p^3q^5+85\varepsilon p^3q^5+399\varepsilon^2p^3q^5+2p^2q^6+18\varepsilon p^2q^6+226\varepsilon^2p^2q^6+68\varepsilon^2pq^7+8\varepsilon^2q^8}{(4p^2+4pq+q^2)(2p^2+3pq+2q^2)(p^2+4pq+4q^2)(p^2+2\varepsilon pq+\varepsilon q^2)^2}$$

劣性遺伝型

$$\lambda_{sib/locus} = \frac{8p^8+68p^7q+226p^6q^2+18\varepsilon p^6q^2+2\varepsilon^2p^6q^2+399p^5q^3+85\varepsilon p^5q^3+15\varepsilon^2p^5q^3+397p^4q^4+183\varepsilon p^4q^4+46\varepsilon^2p^4q^4+223p^3q^5+201\varepsilon p^3q^5+75\varepsilon^2p^3q^5+66p^2q^6+110\varepsilon p^2q^6+70\varepsilon^2p^2q^6+8pq^7+24\varepsilon pq^7+36\varepsilon^2pq^7+8\varepsilon^2q^8}{(4p^2+4pq+q^2)(2p^2+3pq+2q^2)(p^2+4pq+4q^2)(p^2+2pq+\varepsilon q^2)^2}$$

＜多因子遺伝であることの確認＞

以下の内容は、「医科遺伝学(改定第2版)松田一郎監修(南江堂)」からの抜粋でありその理論的根拠について筆者(山田)は確認をとっていないので、参考として挙げるが、責任はとりかねる。

遺伝性疾患は単因子遺伝性疾患と多因子遺伝性疾患があります。RIKEN SRC では多因子遺伝性疾患を扱う。

多因子遺伝か単因子遺伝かの区別は以下のような方法で推定することがある程度可能である。

(1) 一卵性双生児相対危険度が同胞相対危険度の 4 倍以上であれば($\lambda_{MZ} \geqq 4 \times \lambda_{sib}$)、多因子遺伝と考える。4 倍未満の場合は単因子遺伝の可能性がある。

$\lambda_{MZ}/\lambda_{sib}$＝＿＿＿＿(試算して下さい)

(2)Common diseases の場合、多因子遺伝性疾患の一般集団罹患率(ウ)と同胞再発危険率(イ)との間には、

イ　ウ^2

の関係であることが知られている。

イ/ウ^2 ＝＿＿＿＿(試算して下さい)

遺伝性疾患であり、多因子遺伝であることが確認されたら、次に 1 遺伝子当たりの$\lambda_{MZ/locus}$推定する。

個々の疾患原因遺伝子が持つ相対危険度$\lambda_{MZ/locus}$は、

$\lambda_{MZ/locus} \leqq \lambda_{MZ}$

という制限がある。この不等式で等号が成立するのは単因子遺伝性疾患の場合であるから、多因子遺伝である場合には$\lambda_{MZ/locus}$としてλ_{MZ}よりも相当小さな数値を想定する必要がある。数遺伝子が遺伝性の主要部分を占めていると仮定すると$\lambda_{MZ/locus}$はλ_{MZ}の数分の１程度の値をとりうると仮定して研究計画を立てることができる。

今、$\lambda_{MZ/locus}$とλ_{MZ}との関係を述べたが、疾患によってはλ_{MZ}は不明でλ_{sib}に関する情報のみが得られることがある。λ_{sib}に関する情報が疾患遺伝子解析上に一定の役割を持つ、罹患同胞対解析という手法に関して必要なであるためである。$\lambda_{sib/locus}$とλ_{sib}との関係は$\lambda_{MZ/locus}$とλ_{MZ}との関係と同様である。

2-4 参考　どの程度の強さの疾患関連遺伝子がいくつくらいあるか

遺伝性の要因の存在が予想され、かつ複数の遺伝子の関与が予想されている疾患が我々の研究の対象である。しかも個々の遺伝子の、寄与の強さが、我々の研究の成否を相当程度規定していることは“2 遺伝性疾患であることの確認”及び“3-3-2 関連解析の検定力に影響を与える因子”を参照するとわかる。

ではλ_{MZ}やλ_{sib}が知られている場合に、一体どのくらいの数の遺伝子がどのくらいの genotypic risk ratio を持っていると考えればよいのであろうか？

一般に、common diseases と呼ばれる疾患の関連遺伝子が有する genotypic risk ratio はせいぜい高くても 4 くらいまでであろうと考えられている(“2-1 遺伝性疾患であることの確認”参照)。これは、これまでの疾患関連遺伝子検出の歴史から推定されている。というのは、相当数の罹患同胞対を用いた linkage analysis では、これより高い genotypic risk ratio を持つ主要ローカスはすでに検出されていてしかるべきであるのに、検出されていないということは、それよりも genotypic risk ratio が低いローカスしか存在しないということを示唆していると考えるべきだからである。

では、genotypic risk ratio がそれほど高くない遺伝子がいくつくらい集まると、われわれが疫学的に知っているλの値が得られるのであろうか？

以下のグラフは、簡略化のために、アレル頻度も genotypic risk ratio も同一の遺伝子が、複数個、相互に独立に疾患 susceptibility を上昇させていると仮定した上で、その場合にλ_{MZ}がいくつくらいになるかを表している。

実際には異なるアレル頻度の遺伝子がそれぞれ異なる genotypic risk ratio を持ち、さらに関連遺伝子同士が独立ではなく、組合せによって複雑に genotypic risk ratio を変化させることが予想されるので、このシミュレーションから関連遺伝子の数やそのアレル頻度・genotypic risk ratio を予見することはできないが、genotypic risk ratio と遺伝子の数について大雑把なイメージをつかむ役には立つと思われる。

グラフの他に、UNIX 上のプログラム(Nlamgam-input.pl)では、任意のローカス数(13 程度まで)、任意のアレル頻度、任意の genotypic risk ratio、任意の遺伝形式を指定して、その場合のλ_{MZ}が算出できる。

図 1 は、genotypic risk ratio＝2 で、genotypic relative risk の型がγ^2型のローカス(genotypic relative risk が疾患アレルホモ 4、ヘテロ 2)が複数存在する場合をシミュレートしている。ローカス数は 1 から 10 個とし、お互いに独立に疾患 susceptibility を上昇させている場合である。全てのローカスの疾患アレル頻度も同一であると仮定している。

例えば疾患アレル頻度が 0.3 のローカスが 10 個合わさっているような疾患ではλ_{MZ}が 10 を少し越えるくらいになることがわかる。

図 2 は、ある共通の疾患アレル頻度と genotypic risk ratio を持つγ^2型のローカスが 8 個集まって疾患 susceptibility を決めているという仮定をしたときに、その genotypic risk がいくつくらいであると、ある与えられたλ_{MZ}に到達するかをシミュレートしている。

p＝0.6(ピンクの▲)のグラフを例にとる。genotypic risk ratio が 1.5 や 2 の遺伝子が 8 個集まってもλ_{MZ}は 2 や 4 に過ぎない。しかしながら genotypic risk ratio が 2.5、3 と上昇するとλ_{MZ}急速に上昇し、それぞれ 9 超、18 と高くなる。このレベルは遺伝性が確認されている common disease に珍しくない数値である。

図 1

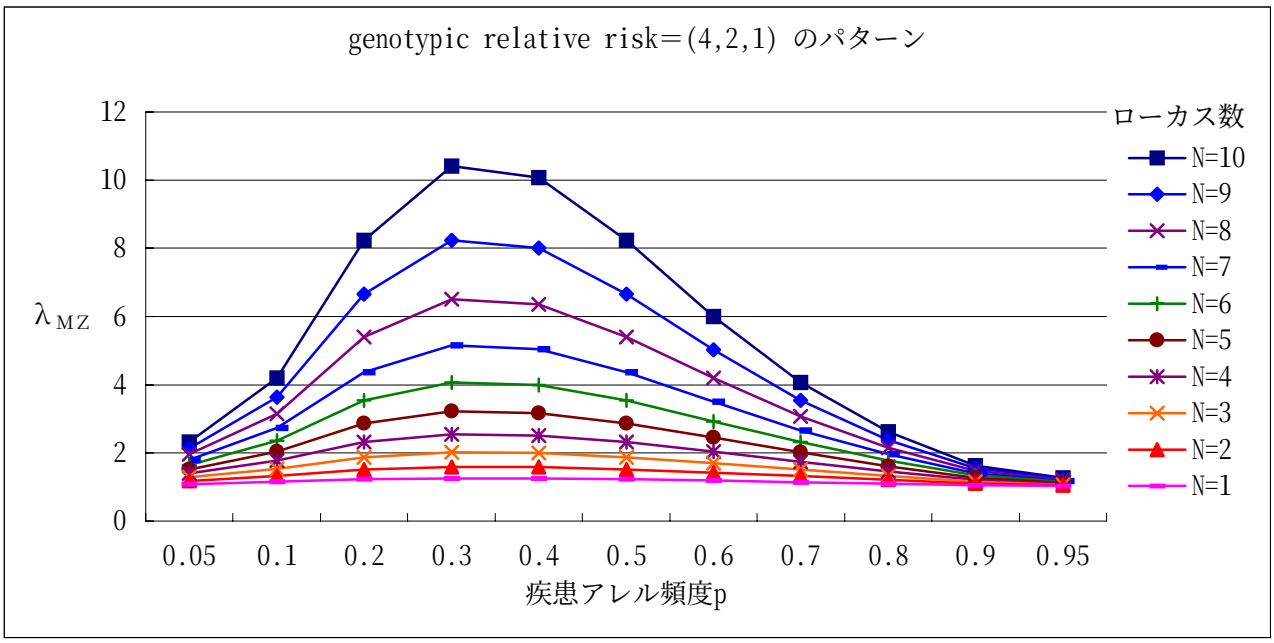
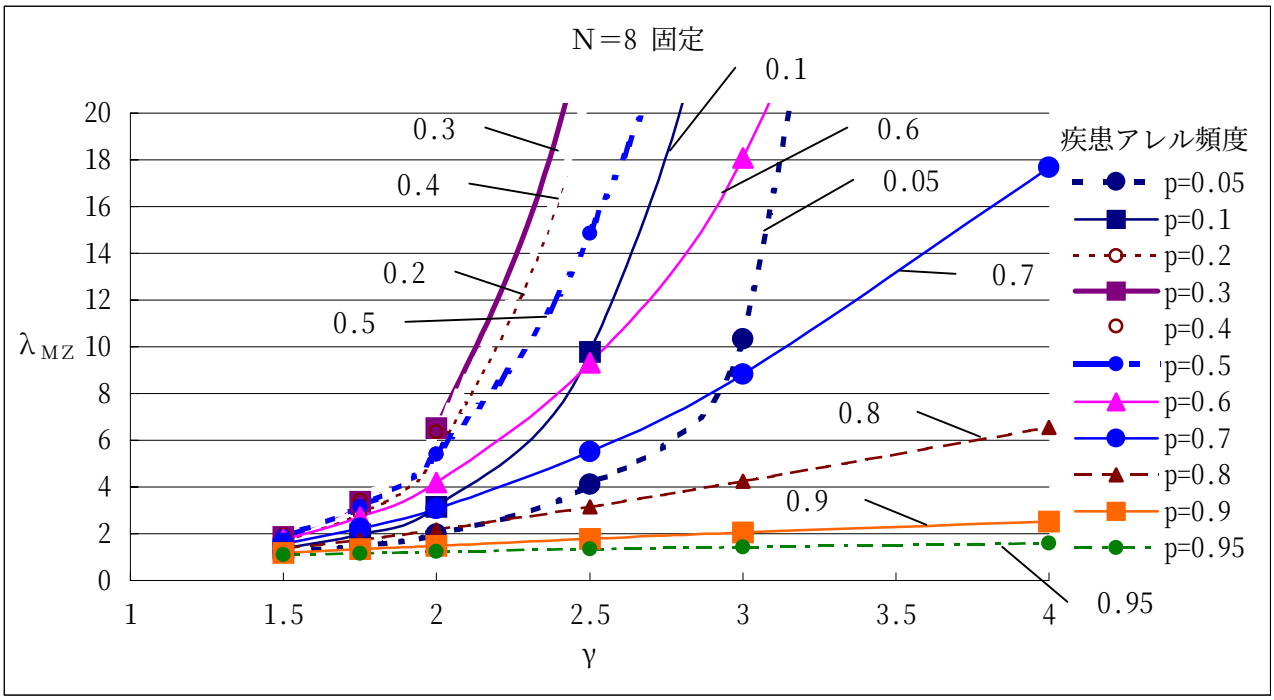

genotypic relative risk=(4,2,1) のパターン
λMZ
12
10
8
6
4
2
0
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
疾患アレル頻度p
ローカス数
N=10
N=9
N=8
N=7
N=6
N=5
N=4
N=3
N=2
N=1

図 2


N=8 固定
λMZ
20
18
16
14
12
10
8
6
4
2
0
1
1.5
2
2.5
3
3.5
4
γ
0.1
0.3
0.4
0.2
0.5
0.6
0.05
0.7
0.8
0.9
0.95
疾患アレル頻度
p=0.05
p=0.1
p=0.2
p=0.3
p=0.4
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
p=0.95

3
関連解析

3-1　関連解析の種類

疾患と SNP との関連を検出する方法はいくつかの基準で分類できる。

分類の基準

1. 使用する SNP の数は１つか複数か
2. 関連を示唆する“差”や“ずれ”を解析するときにケース群のみを用いるか、またはケースとコントロールの２群を比較するか
3. 統計的推測の方法としてモーメント法を用いるか、または最尤法を用いるか（この推定法の差に関しての説明はこの資料集の範囲を越えるので詳しくは統計学の成書を参照のこと）

3 の統計的推測の方法は解析を行う基となる統計的推測量を求める方法であるから、1 と 2 の基準による組合せでできる全ての解析が、モーメント法による推定量を基に行われうるし、最尤法による推定量によっても行われうる。

1 と 2 が作る解析の種類は、以下の表１の通りである。

表１

<table>
<tr><th></th><th colspan="2">単一 SNP</th><th colspan="2">複数 SNP</th></tr>
<tr><td rowspan="3">ケース・
コントロール
比較</td><td rowspan="2">分割表検定
(χ^2, Fisher's exact probability test)</td><td rowspan="2">モーメント法</td><td>ハプロタイプ解析
(近接複数 SNP ﾃﾞｰﾀの統合)</td><td>モーメント法
最尤法</td></tr>
<tr><td>マッピング法
(単一 SNP 解析、
ﾊﾌﾟﾛﾀｲﾌﾟ解析ﾃﾞｰﾀ適用可)</td><td rowspan="2">モーメント法
最尤法</td></tr>
<tr><td>分布比の尤度比検定</td><td>最尤法</td><td>機能的相互関連遺伝
ローカス組み合せ
(遠隔複数 SNP ﾃﾞｰﾀの統合)</td><td>最尤法</td></tr>
<tr><td>ケースデータ
のみの使用</td><td>Hardy-Weinberg 平衡検定</td><td>モーメント法
最尤法</td><td>マッピング法
(Hardy-Weinberg 平衡)</td><td>モーメント法
最尤法</td></tr>
</table>

この RIKEN SRC で関連 SNP スクリーニングに用いるのはケース・コントロール関連解析（分割表の χ^2 検定）である。また、生データの質をチェックする目的に主に用いられるのが Hardy-Weinberg 平衡検定であり、それを補助するのが分割表の χ^2 検定である。

スクリーニングにより有意とされた SNP に対しては検体数を増やした上で、単一 SNP 解析を適用してその統計的有意差の強さを確認し、関連が確認された場合に、その近傍 SNP と組み合わせてハプロタイプ解析やマッピング法を適用する。

3 の統計的推測の方法については、原則としてモーメント法と最尤法との間に大差はなく、モーメント法を用いていれば問題はない、と理解していればよい。計算も圧倒的にモーメント法が簡略である。

ここでモーメント法がどういうものかに頭を悩ませる必要はない。なぜならば、もし、あなたが“最

尤法 ” を知らないならば、あなたが考えるデータの処理方法は必ず “ モーメント法 ” と呼ばれる方法であるからである。

最尤法がモーメント法に対してより有用になるのは次のような限られた場合になると思われる。分割表分析が比較的不得手とするタイプのデータである、大サイズの分割表データの処理(数多くのハプロタイプに対して解析を行う場合)と、モデルを導入しなければならない場合(遺伝ローカス同士の相互作用をモデル化したり、コントロールサンプルを一般集団から採ったために、罹患率のケース混入の分を割り引いて解析を行ったりする場合)などである。

以下この資料集では単一の SNP に関するデータを “ 3-3 ケース・コントロール関連解析 ” と “ 3-4 ケース内解析 ” に 2 分して説明していく。さらに、複数の SNP のデータを組み合わせて解析する方法についても “ 3-5 複数点解析 ” で述べる。また、解析について話を進めるにあたってもう一点、区別して進めるべきことがある。それは解析する SNP が常染色体上にあるか、性染色体上にあるかという違いである。この差は SNP の genotype 構成を変えるので、解析手法は同じでも別に扱う必要がある。この件に関しては “ 3-2 性染色体上多型の特殊性 ” で述べる。

3-2　性染色体上多型の特殊性

3-2-1 性染色体上の SNP の検出、タイピング、タイピングデータに基づく疾患関連解析

＜序＞

ヒト X 染色体は全長 163Mb で、約 1000 個の遺伝子が存在すると考えられているが、ヒト Y 染色体は機能を持った遺伝子は 100 個程しかなく、また全長 51Mb 中、27Mb が転写不活性なヘテロクロマチンである。

＜性染色体上の相同領域＞

性染色体上の相同領域を理解する上で、X 染色体の不活性化を理解する必要があるので、X 染色体の不活性化について、まず解説する。男女間の遺伝子量の差を補正するため、女性では 2 本の X 染色体のうちどちらか一方を、発生初期に細胞ごとにランダムに不活性化する。その不活性化のパターンは、娘細胞に引き継がれる。結果、父方 X 染色体が不活性化している細胞系と、母方 X 染色体が不活性化している細胞系の混合体(モザイク)となっている。ただし、この X 染色体不活性化の法則に当てはまらない場合も存在する。なぜなら X 染色体不活性化は、Y 染色体上にその相同領域を持たない X 染色体上の遺伝子(ほとんど全部)を量的に補正する機構である。よって X 連鎖遺伝子のうち、Y 染色体上に機能を持った相同領域がある場合には不活性化を免れる(この場合 XX の両方が発現する)。ただし、例外遺伝子もあり、UBE1 や SB1.8 遺伝子のように不活性化されていないが、Y 染色体上に相同遺伝子がないものや、Y 染色体上に相同遺伝子を持っても偽遺伝子の場合がある。この場合、XX の方が発現量は 2 倍多いが、淘汰圧に耐えるようである。

X と Y 染色体上の遺伝子や部分配列を比較すると、大部分は非相同領域である。しかしながら、X と Y 染色体は古代の同型対染色体から進化してきたという点や重複転位現象によって、以下の 3 種類の部分的に相同な領域が知られている。

A) 主擬似常染色体領域(Major Pseudoautosomal Region : PAR1、XY 短腕最先端 2.6Mb)

この領域は、雄の減数分裂中に交差を必ず起こす。また、この領域内の遺伝子は X 不活性化と無関係に発現される。すなわち、減数分裂上及び遺伝子発現上は、常染色体と同じである。

B) 副擬似常染色体領域(PAR2、XY 長腕最先端 0.32Mb)

主擬似常染色体領域と異なり、X-Y 間での交差はそれほど頻繁に起こらない。PAR2 上には 4 遺伝子が知られているが、そのうち 2 遺伝子が X 不活性化を起こしている。減数分裂上は、常染色体と同じである。

C) その他の相同領域

その他の相同領域として、Xp-Yq 間/ Xq-Yp 間/ Xp-Yp 間/ Xq-Yq 間相同領域がある。

X 染色体不活性化による発現量(haploid または diploid 発現)と相同、非相同領域の対応関係を以下の表 1 にまとめる。男女とも haploid 発現の場合は、領域として「X 染色体上の非相同領域の大部分」と「副擬似常染色体領域の一部」が該当する。例には遺伝子名を示し、さらに X 染色体不活性化を受けている場合には「不活性」、そうでない場合には「活性」と記載している。

表１．X 染色体不活性化による発現量(haploid または diploid 発現)と相同、非相同領域の対応関係

発現量	領域	例	X 不活性化の状態	備考
男女とも haploid 発現	X 染色体上の非相同領域の大部分		不活性	
	副擬似常染色体領域の一部	SYBL1	不活性	Y 上不活性#
		HSPRY3	不活性	Y 上不活性#
男女とも diploid 発現	主擬似常染色体領域		活性、不活性	
	副擬似常染色体領域の一部	IL9R	活性	
		CXYorf1	活性	
	その他の相同領域の大部分		活性	
女 diploid 男 haploid 発現	X 染色体上の非相同領域の一部	UBE1	活性	
		SB1.8	活性	
	その他の相同領域の一部	KAL1	活性	Y 上偽遺伝子
		STS	活性	Y 上偽遺伝子

#偽遺伝子のため発現しないのではない。なぜ Y 上の遺伝子が不活性であるかは現在のところ不明。

＜性染色体上の SNP 同定＞

まず、性染色体上の遺伝子分類とその配列の X、Y 染色体上の有無、X 染色体と Y 染色体で配列に差があるかどうかのまとめは以下の表の通りである。

表 1

遺伝子分類	対応配列 X 上	対応配列上	X と Y の差
① Y 染色体特異的遺伝子	なし	あり	あり得ない
② X 染色体特異的遺伝子	あり	なし	あり得ない
③ X/Y 染色体上の相同遺伝子※	あり	あり	あり
④ 高頻度組み換え領域の遺伝子(擬似常染色体領域)	あり	あり	なし

※③には Y 上が偽遺伝子である場合も含まれる。

次に、性染色体上の SNP 同定で検体間に差が認められた時に、SNP と判断するか、SNP ではなく X と Y との相同配列の差を検出しているのかを解釈する際に、使用検体を女性(XX)のみか男女混合(XX/XY)では、解釈が異なる(表 2)。

“○”は X または Y 上の SNP として一意に決まる場合、“×”は決まらない場合を示す。例えば、検体 XX のみの場合、②、③、④において一意に決まるが、XX/XY 混合の場合、②しか一意に決まらない。また、XX/XY 混合の③では、検体間の差が、X 上の SNP であるか Y 上の SNP であるか X と Y の差であるかを区別出来ない。

表 2

検体 XX のみ	検体 XX/XY 混合	プライマー作成の由来
①	○ (Y 上の SNP)	Y
② ○ (X 上の SNP)	○ (X 上の SNP)	X
③ ○ (X 上の SNP)	× (X 上の SNP) (Y 上の SNP) (X と Y の差)	X, Y
④ ○ (X 上の SNP)	× (X 上の SNP) (Y 上の SNP)	X, Y

SNP として一意に決まるものは SNP 同定の解析対象とするので、遺伝子分類別の方針としては、

① 検体 XY でのみ実施可能。Y 染色体上の SNP として登録。

② 検体 XX 検体 XY どちらでも実施可能。X 染色体上の SNP として登録。

③ X 染色体上でプライマー設計し、検体 XX でのみ実施可能。X 染色体上の SNP として登録。

④ X 染色体上でプライマー設計し、検体 XX で実施するのが望ましい。X 染色体上の SNP として登録。

となる。

以上よりまとめると、

①に関しては、Y 染色体上でプライマー設計、検体は XY のみ、登録は Y 染色体上で行う。ただし、プロジェクトの方針として①は SNP 同定を行わない。また Y 上の③,④は SNP 同定不可能である。

よって、性染色体上の SNP 同定は、X 染色体上でプライマー設計、検体は XX のみ、登録は X 染色体上で行う。

＜性染色体上タイピングの目的と手順及び結果の関連解析への利用法＞

タイピングの目的は、タイピング結果から“個体レベルでの発現活性遺伝子型”を推定し、その活性遺伝子型と疾患との関連を見い出すことである。その手順は以下の通りである。

↓ タイピング結果（観測型）を得る (1/12/2)
　　11/12/22 ではないことに注意

↓ 観測型から DNA 上の遺伝子型(DNA 型)を推定する
　　常染色体と性染色体では、推定の仕方が異なる
　　(観測型 1/12/2→DNA 型 11/12/22 ではない)

↓ DNA 型から“細胞レベルでの活性遺伝子型”を推定する
　　性染色体では X 染色体不活化を考慮しなければならない
　　(DNA 型＝活性遺伝子型ではない)

↓ “細胞レベルでの活性遺伝子型”から“個体レベルでの活性遺伝子型”を推定する
　　個体レベルでは X 染色体不活化モザイクになっているのでその点を考慮しなければならない

↓ “個体レベル活性遺伝子型”と疾患との関連を検定する
　　観測型から活性遺伝子型が推定できれば、性染色体上 SNP でも関連検定可能である

※常染色体の場合は観測型→“個体レベルでの活性遺伝子型”が単純に決定できるため、その解釈と利用法が簡便である。

タイピング結果(観測型)から DNA 上の遺伝子型の推定と、X 染色体不活性化を考慮した発現活性レベルでの遺伝子型(活性遺伝子型)の推定は、以下の通りである。

	女性検体 XX			男性検体 XY		
	観測型	DNA 型	細胞ﾚﾍﾞﾙでの活性遺伝子型	観測型	DNA 型	細胞ﾚﾍﾞﾙでの活性遺伝子型
①	0	00	00	1/2	10/20	10/20
②	1/12/2	11/12/22	10/<u>10</u>,<u>20</u>/20	1/2	10/20	10/20
③-1	1/12/2	11/12/22	11/12/22	1/12	$11_y/21_y$	$11_y/21_y$
③-2	1	11	11	1/12	$11_y/12_y$	$11_y/12_y$
③-3	1/12/2	11/12/22	11/12/22	1/12/2	$11_y/\underline{21}_y$, $\underline{12}_y/22_y$	$11_y/\underline{21}_y$,$\underline{12}_y/22_y$
④	1/12/2	11/12/22	11/12/22	1/12/2	11/12/22	11/12/22

<u>注釈</u>

タイピングは2色の蛍光標識で検出する系を用いているので、観測型としては1/12/2の3通りである。

原則として不活化している②では“細胞レベルでの活性遺伝子型”と“個体レベルでの活性遺伝子型”が1対1に対応しないが、その他のグループの遺伝子は原則として“細胞レベルの活性遺伝子型”と“個体レベルの活性遺伝子型”は1対1対応する。

③の相同遺伝子では、X 上の1と Y 上の1を区別するため、Y 上の遺伝子型に$_y$を付与してある。

③-1 は、X 上に SNP がある場合、③-2 は、Y 上に SNP がある場合、③-3 は、X/Y 上両方に SNP がある場合を表す。

活性遺伝子型は、haploid/diploid の差を 10/11(もしくは 20/22)として表現している。

表の灰色部分は、対応が1対1でない部分である。

表より、

①、③-1、③-2、④は、観測型から DNA 型、DNA 型から活性遺伝子型が一意に決まる。

②は、検体 XX の DNA 型から“細胞レベルでの活性遺伝子型”が一意に決まらない。

③-3 は、検体 XY の観測型から DNA 型が一意に決まらない。

よって、

①、③-1、③-2、④　“個体レベルでの活性遺伝子型”は常染色体のそれとは異なるが、phenotype と組み合わせて分割表を作成して解析可能。

②　“X 染色体不活化モザイクの総体としての個体”の活性遺伝子型が主に影響する疾患（多くの全身性疾患など)では DNA 型がヘテロの個体の活性遺伝子型はモザイク化の影響を受ける。このモザイク化の影響を確率論で補正することによってヘテロ個体を検定に組み入れることは可能かもしれないが、その妥当性は明確ではない。

③-3 観測型 12 から DNA 型 21_y、12_yを一意に決定できないので、現状の実験系では検定不可。ただし、X と Y 上の相同遺伝子を PCR で分けるなどして区別した上で、注目している多型をタイピングするという手順を踏めば、観測型から DNA 型への対応は1対1となり、分割表化して解析可能となる。

その他、②の場合で細胞レベルでの活性遺伝子型が主要な影響を与える疾患(癌など)では当該細胞の活性遺伝子型を知ることが必要になる。

大規模なスクリーニングの場合、現実的には同定の確実性と数の多さから、②と④の SNP のみがタイピングするのが適切と考えられる。

②を DNA 型まで推定し、“ 個体レベル ” では

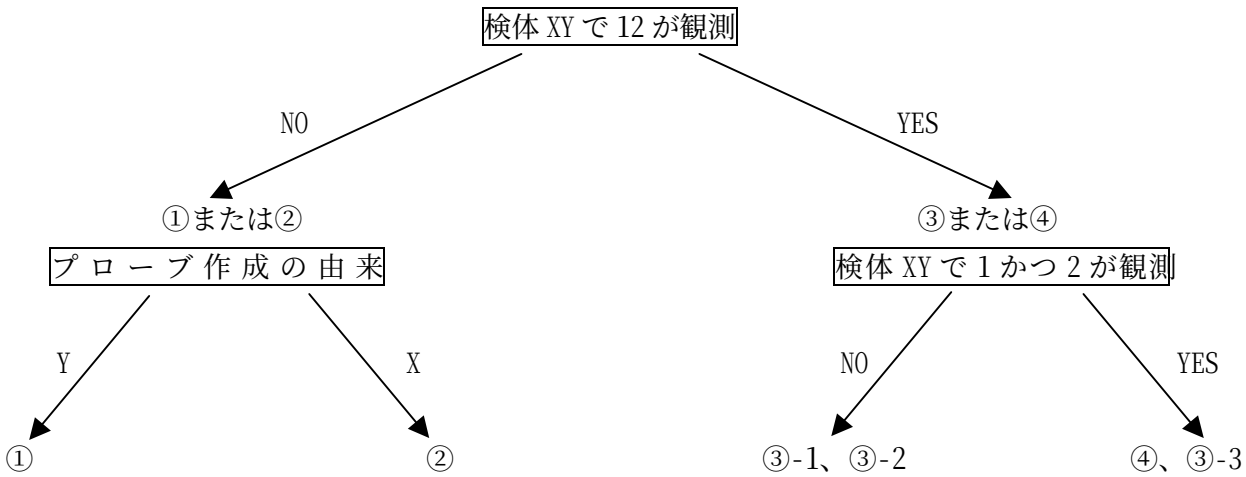
女性ホモ個体と男性個体のみで解析する　もしくは

女性ヘテロ個体を中間的な発現レベルを有すると解釈して何らかの形で解析に組み込む

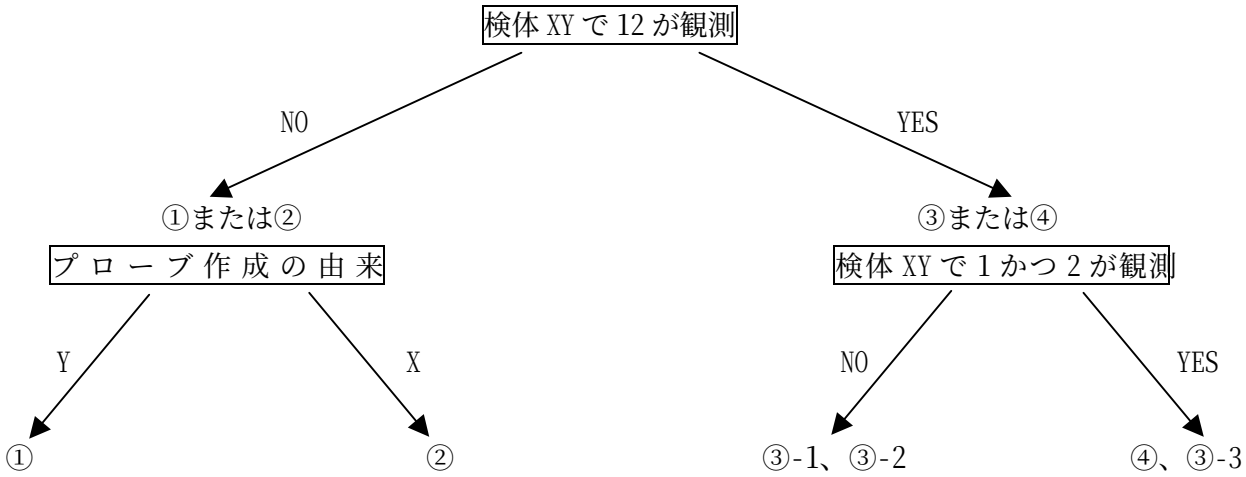
④を常染色体上遺伝子と同形式で解析する

＜観測型からの遺伝子タイプ別分類の推定＞

ある性染色体上の SNP が①、②、③または④のどれに属するかの判断は、個々の遺伝子がどこに属するかの情報が得られればそれを用いてもよいし、男女別の観測型から逆に類推してもよい。観測型の結果から、以下のように分類することができる。



実際に、タイピング領域が X 上または不明のものは、XY 染色体上に相同配列があるかどうか、多数の男女別のタイピングデータを基に高い精度で推定可能であり、RIKEN SRC のデータもこの観点からチェックを受けている。具体的には男性でヘテロと判定された観測数が、女性の結果と比べて、偏っているかどうかを判断するプログラムにかける。ある程度の観測誤差による偏りと相同配列による偏りは、最尤法で区別がほぼ確実であると思われる。



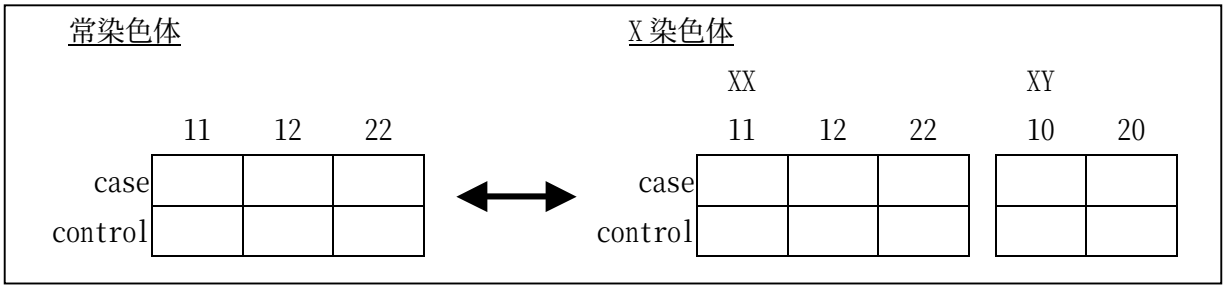
実際に、タイピング領域が X 上または不明のものは、XY 染色体上に相同配列があるかどうか、多数の男女別のタイピングデータを基に高い精度で推定可能であり、RIKEN SRC のデータもこの観点からチェックを受けている。具体的には男性でヘテロと判定された観測数が、女性の結果と比べて、偏っているかどうかを判断するプログラムにかける。ある程度の観測誤差による偏りと相同配列による偏りは、最尤法で区別がほぼ確実であると思われる。

3-2-2 X 染色体特異的遺伝子上 SNP 検定法

以下の説明に該当するのは、“3-2-1 性染色体上の SNP の検出、タイピング、タイピングデータに基づく疾患関連解析”の“② X 染色体特異的遺伝子”の場合の検定法についてである。

1. genotype を用いた分割表検定において、常染色体では 2×3 分割表を用いるが、X 染色体の場合、2×5 分割表を用いる(以下の図 1 参照)。

図 1



また、genotype を優性か劣性遺伝形式かによって 2 群に分ける方法も異なる。常染色体では 2 パターンしか存在しないが、X 染色体では 4 パターン存在する。

(1) リスクの上昇に allele 1 homozygosity が必要
11 vs (12+22)　　11 vs (12+22+10+20)

(2) リスクの上昇に allele 2 の不在が必要
11 vs (12+22)　　(11+10) vs (12+22+20)

(3) リスクの上昇に allele 1 が必要
(11+12) vs 22　　(11+12+10) vs (22+20)

(4) リスクの下降に allele 2 homozygosity が必要
(11+12) vs 22　　(11+12+10+20) vs 22

2. アレル頻度を用いた分割表検定において、アレル頻度の計算方法が、常染色体と X 染色体とでは異なる。
1をもつアレル頻度は、
常染色体：(11 観測数×2+12 観測数)/全観測数×2
X 染色体：(11 観測数×2+12 観測数+10 観測数)/(XX 観測数×2+XY 観測数)
で計算される。

3. Hardy-Weinberg 平衡検定の検定方法が、常染色体と X 染色体とでは異なる。
常染色体では、2.のように 1 と 2 のアレル頻度を計算した後、そのアレル頻度から期待値 11/12/22 を算出する。X 染色体では、2.のように 1 と 2 のアレル頻度を計算した後、そのアレル頻度から期待値 11/12/22/10/20 を算出する。

3-3
ケース・コントロール
関連解析

3-3-1 ゲノムワイド ケース･コントロール関連解析を行うにあたって

ケース･コントロール関連解析は群間比較の一番初歩的な方法であり、改めてその方法について述べるほどのものではないが、検出できるかできないかのぎりぎりの線上で、ゲノムワイドに行うという事情に関連して、いくつか確認しておいたほうがよい点があると思われるので、以下の順序で説明を加える。

マーカーを使った解析の検定力が弱くなる理由がいくつかある。それらを

3-3-2 関連解析の検定力に影響を与える因子

3-3-2-1 直接関連と間接関連

3-3-2-2 コントロール集団

3-3-2-3 連鎖不平衡

3-3-2-4 集団間民族学的遺伝的差(階層化を含む、その問題点と存在の有無の検定)

の 4 項目で、それぞれ説明する。

また、関連解析における第 1 種過誤、第 2 種過誤についてと、それと関連する解析に必要な検体数に関する話と、さらに、RIKEN SRC のプロジェクトによって得られるデータの統計的解釈について

3-3-3 関連解析に必要な検体数と第 1 種過誤(偽陽性率)･第 2 種過誤(偽陰性率)

3-3-3-1 第 1 種過誤(偽陽性率)･第 2 種過誤(偽陰性率)

3-3-3-2 関連解析の必要検体数

で解説する。

3-3-2　関連解析の検定力に影響を与える因子

3-3-2-1 直接関連と間接関連

ある SNP とある疾患との間に関連が見出される場合、その関連は直接の因果関係に基づく“ 直接関連 ” と、直接の因果関係はないものの “ 直接関連 ” のある要因を介して間接的に関連を認める “ 間接関連 ” とがある。“ 直接関連 ” を認めた SNP と疾患の場合の、相互の関係の解釈は単純である。

一方、“ 間接関連 ” を認めた場合には介在する要因のタイプによってその解釈は分かれる。

間接関連：その介在要因のタイプと対象疾患・対象 SNP との関係

介在要因	介在要因と 当該疾患との関連	介在要因と 当該 SNP との関係
①当該 SNP 近傍の遺伝子多型	真の直接関連あり	連鎖不平衡あり
②当該疾患以外の phenotype	真の間接関連あり	直接関連あり、もしくは 連鎖不平衡による 間接関連あり
③比較集団間の 遺伝的背景の差異	真の関連なし	SNP が集団間の差異の 構成要素である

間接関連を有する SNP を検出した場合の対処法
①の場合：
ゲノムワイドに関連遺伝子を探し、SNP をそのマーカーとして用いる場合にはこのような間接関連 SNP の検出を第 1 の目標にしているわけである。この間接関連 SNP を手がかりにその SNP と連鎖不平衡内にある多型･変異について疾患と直接関連のあるものを同定していく。連鎖不平衡に関しては “ 3-3-2-3-1 連鎖不平衡とは ” 参照のこと。
連鎖不平衡と連鎖不平衡が関連検定に及ぼす影響については “ 3-3-2-3-3 真のローカスに近接する SNP の真のローカスとの LD の強さとアレル頻度の乖離が検定力に及ぼす影響について ” を参照のこと。
②の場合：
当該疾患の直接関連遺伝子を同定するという目標から外れるものの、当該疾患と関連する phenotype を規定する遺伝子の同定に役立つので利用価値は高い。しかしながら、当該疾患の直接関連遺伝子が関連 phenotype の直接関連遺伝子よりも遺伝的寄与が小さくなってしまうと、関連 phenotype の関連遺伝子ばかりが同定される危険がある。そのような場合には、サンプルを関連 phenotype について層別化して解析するなどの対応が必要である。
③の場合：
ケース・コントロール サンプリングにおいてはケース集団とコントロール集団との間には疾患の有無という phenotype 以外の差が無いことを前提としているが、何らかの原因でケース集団とコントロール集団との間に遺伝的な背景に差がある場合がある。当該 SNP がその集団間の遺伝的な背景の差に含ま

れる場合には、疾患という phenotype とは無関係であっても、関連があるものとして見出される。ケース・コントロール関連解析の根幹に関わる問題であり、両集団間に遺伝的背景の差が存在する場合に、ゲノムワイドに関連マーカーの検出を行うと、相当数のマーカーが疾患と関連あるものとして認められてしまう。ケース・コントロール両集団間に遺伝的背景の差の無いことをあらかじめ確認しておくことが必要である。

集団間の遺伝的背景の差(階層化がその代表例である)とその存在の有無を検定する方法、また、階層化が存在する場合のケース･コントロール関連解析の手法については“ 3-3-2-4 集団間の民族学的遺伝差(階層化を含む、その問題点と存在の有無の検定) ”参照のこと。

3-3-2-2　コントロール集団

3-3-2-2-1 ケース集団とコントロール集団の構成について

ケース・コントロール関連解析を行うときの前提は、

1. ケース個体の phenotype が定義されており、その phenotype を持つことを基準にサンプリングされることである。
2. コントロール個体としての最低条件は、ケース集団をサンプリングする母集団と遺伝的に均質の集団からサンプリングされることである。

この 2 大条件を満たせばケース群とコントロール群の間の差を検出することは理論的に可能である。しかしながら、実際に疾患関連遺伝子を同定しようとする場合、この差が僅かであるために、その検出が非常に難しい。

この困難を乗り越えるために、以下のような様々な工夫がなされる。

(1)ケースの phenotype を均質化する
　①ケースを亜分類する
　など
(2)ケース集団とコントロール集団の遺伝的背景をなるべくそろえる
　①限られた、厳密な均一母集団から、ケース・コントロールサンプリングを行う
　②ケースの血縁者（内部コントロール）をコントロールとして採用する
　など
(1)ケース集団とコントロール集団との phenotype 上の差を大きくする
　①コントロール群に混入するケース個体を減らす
　　a) ケースでないことの確認の取れた個体をコントロールとして採用する
　　b) 一生涯ケースとしての phenotype をとらなかったコントロール(super-control)を採用する
　など

しかしながら、実際に Common diseases の疾患遺伝子解析においてサンプルを集める段階には、ケース集団とコントロール集団に関して制約が存在する。

(1)疾患自体に根ざした要因
　①ケース集団が複数の遺伝的要因の総和からなるため、phenotype は同一でも遺伝的に不均一な集団であること(disease heterogeneity)
　②疾患の浸透率が低いこと
(2)サンプリングを実行に移す段階の要因
　①コントロールのサンプリングはそもそも難しい
　②コントロールに条件をつけて(内部コントロール、super-control、sex-age-matched コントロールなどを用いて)サンプリングすることはさらに難しい

実際に RIKEN SRC プロジェクトでは、コントロールのサンプリング条件を厳しくし、解析に使用する検体数を少なく済ませるのではなく、コントロールのサンプリング条件は緩くし、検体数を集めやすくする、「検定力の低下を検体数の増加で補う」方針を選択している。

具体的には、ケースのサンプリングは、疾患によって異なるが、新規発症者をサンプリング対象としている場合と、罹病者集団からしている場合とがある。いずれの方法にしろ検体の年齢構成・性別分布は、サンプリングの段階で操作することはせず、その疾患のサンプリング方法と発症年齢と罹病期間とによって規定されてくる。

一方、コントロール集団はケースの疾患とは異なる、その他の複数の common diseases のケースとしてサンプリングされたものを混合した検体群を模擬一般集団とみなして、ケース集団対模擬一般集団の比較を行っている。

この模擬一般集団を一般集団としてみなしてよいかが問題になるが、

1. Common diseases のケースであって、稀な遺伝性疾患のケースはその対象ではない
2. 複数の common diseases を混合したものである
3. それぞれの疾患集団間に遺伝的背景の差が認められないことを、多数の SNP タイピング結果を用いて検証する

という手続きを踏むことで、乗り越えようとしている。

また、疾患関連 SNP の同定は、タイピング結果に基づく統計的検定のみで行われるものではなく、あくまでも疾患 susceptibility に影響を与える遺伝子の機能的裏付けを必要とするものであり、大規模関連解析においては一般集団をコントロール集団として用いることが適当であるとみなしている。さらに、コントロール集団として一般集団を用いることは、関連を見い出しにくくするという弱点はあるものの、この 2 群間で検出された関連は、コントロール群としてより条件の厳しいコントロールサンプリングを行った場合にはもっと強い関連を生ずるはずであり、検定としてはより保守的であり、データの信用性が低くなることは無いと言える。

3-3-2-2-2-1　参考 性・年齢マッチ-コントロールは必要か？

ケース群とコントロール群とのサンプリングの工夫の１方法として、性・年齢マッチ-コントロールを使用する方法がある。

ケースとコントロールの性・年齢をマッチさせることは、医学・生物学分野の研究でしばしば要求される。医学・生物学分野で数多く行われる interventional study と呼ばれる研究手法では、特に性･年齢マッチが強く要求される。interventional study の代表格は、ある治療の効果を加療群とプラセボ群との間の差を見い出す研究である。このようなタイプの研究では、intervention(治療するかしないかなど。研究者がコントロールすることが前提となっている)がもたらす結果を計測するし、結果が加療群とプラセボ群とで差があるかどうかを調べる。このような場合、intervention を与える対象は、与える時の状態ができる限り均一であることが望ましい。そのような時に均一にするべき条件の中に、性と年齢は当然のことながら含まれる。例えば、ある死亡率を計測する場合、intervention を与える時点で、高齢者であるか、若年者であるかでその結果が異なることは当然であることからも容易に理解できる。このようなタイプの研究方法は intervention を与えてから変化の出現を追っていくので前向き研究(prospective study)と呼ばれる。

では、大規模なケース・コントロール関連解析で関連遺伝子・関連マーカーを研究する場合にあてはめて考えるとどうなるだろうか？

発症と病態に性差がある疾患は数多く、これらを別個に扱うことは適切である。また、年齢により発症や病態に違いがある疾患も多く、これを別個に扱うことも適切である。

しかし、interventional study とケース・コントロール関連研究との間には大きな違いがある。それは、前者が前向き研究(prospective study)であるのに対し、後者が後ろ向き研究(retrospective study)であることである。後ろ向き研究とは、何らかの因子(遺伝因子など)の影響の結果、すでにある事態(発病など)が起きてしまっているところから、過去の因子の役割を研究するからである。言い換えると、前者は、intervention(研究者がコントロールすることが前提となっている)がもたらす結果を計測するのに対し、後者は、複雑な遺伝因子と環境因子の複合の結果である疾患 phenotype(interventional study の“結果”に相当する)を基にサンプリングを行い、そのサンプルが所有する遺伝因子の差(interventional study の“intervention”に相当する)を検出しようとしている訳である。従って、interventional study と同様の意味でケース・コントロール関連解析を計画するということは、未発症の集団から、性・年齢をマッチさせた遺伝因子陽性サンプルと遺伝因子陰性サンプルとを確保し、prospective にある一定期間追跡した結果、その因子を保有する群と保有しない群とで発症のパターンに差があることを明らかにする必要がある（いわゆるコホート研究）。このような研究計画は未知の遺伝ローカスを同定する目的にはそぐわない。なぜならば、比較の対象となるあらゆる危険因子について揃えてから追跡を始める必要があるからである。

しかしながら、遺伝因子という“intervention”は都合の良いことに、どの個人も受精卵のとき以来持ち続けるという特徴がある。これは、遺伝因子を観測する結果は、観測されるサンプルの年齢によらずいつでも同じということである。また、民族学的に変化の無い集団においては性･年齢による分布差が無いということである。（もしあるとすればそのローカスが連鎖している遺伝子は個体の生存に大きな影響を与えると予想しなくてはならない。そのような遺伝子多型は数多くはない。ただし、性染色体上の遺伝子は性差に関してこの限りではない。）

この 2 点を考慮した場合、一般集団をコントロール集団とする場合には、あえてケースとコントロールの性･年齢構成をマッチさせる必然性は消失する。

では、疾患が持つ性･年齢による特性という情報は使用することが出来ないのかというと、そうではない。その情報はケース集団を亜分類する目的に使用すればよい。そうすることによって性･年齢に特徴的な遺伝因子が凝縮されたケース亜集団を得ることができる。そのような性･年齢によって亜分類されたケース集団に対して、どのようなコントロール集団を用いるかは別の問題である。

発症が性・年齢の影響を受ける疾患の関連解析において、一般集団をコントロールとして用いることは、性･年齢マッチ-コントロールをコントロールとして用いる場合に比べ、検定力が落ちる可能性はある。しかし、この検定力の低下は本項で述べてきた通り、コントロール集団にどのくらいの割合でケースが混入するかによって決まるものであり、場合によっては性･年齢をマッチさせない集団をコントロールとして採用するほうがケース集団とコントロール集団との差を大きくし、検定力が強くなる可能性もある。その顕著な例は高齢の罹患歴を持たない個体をコントロールとする場合(super-control)である。

以上のことからわかるように、コントロール集団として性･年齢マッチ-コントロールにこだわることで得られる解析力の向上と、研究の質の向上は必ずしも必要なものではないと考えられる。

3-3-2-2-2-2-1　参考 年齢情報のいろいろ(「ケース」対「非ケース」関連解析の場合について)

年齢情報というときに、解析時年齢、採血時年齢、臨床情報取得時年齢、発病時年齢の４種類が考えられる。この４種類の年齢情報は、ケース･コントロール関連解析において取得できるか否かを示した表は以下のようになる。

	解析時年齢	採血時年齢	臨床情報取得時年齢	発病時年齢
ケース	○	○	○	○
コントロール	○	○	○	×

○：取得可能、×：取得不可能

１　解析時年齢の意味と、それが必要な場合

「解析時年齢を考慮する」とは、言い換えれば「世代コホート別の genotype 分布の差を考慮する」ということである。日本人を扱う場合､通常、これを考慮する必要はない。なぜなら、ある研究が遂行される期間は比較的短時間であり､その短時間の間に検体提供者となり得る「日本人」はその期間内に生存する世代（3-4 世代程度）に関わらず、均質であると考えてよいからである。もし仮に、現存の人の genotype 分布が、世代別に異なるとすれば、その集団には何らかの民族学的な移動や選択などが働いたことになる。

なお、解析時年齢には、生年月日を基に算出しうるので、データの確実性が高いという特徴がある。

２　採血時年齢の意味と、それが必要な場合

採血期間が短期間に限定されているとき、「採血時年齢」と「解析時年齢」はほぼ同じか､採血-解析間のタイムラグの分だけ並行移動した情報となるだけで、本質的に変わるところはなく、通常、考慮する必要はない。

しかしながら、考慮する必要がある可能性があるのは、寿命に影響を与える genotype が存在する場合である｡言い換えれば､ある種の genotype の所有者は、比較的早死にをしたり、比較的長寿であったりすると、その genotype がある高齢世代に占める割合はそれぞれ、低く、もしくは高くなる。このような genotype が存在し､かつその genotype が寿命のほかに解析対象である疾患と関連している場合には、このことを考慮した解析が必要になる。(Ref Application of log-linear model of inference on disease susceptibility gene effects under independence of genotype and age. (Poster presentation by N. Tanaka on ASHG Annual Meeting 2001)

このように、採血時年齢は「採血のときに少なくとも生存していた」という「情報」が付随した年齢情報である｡この付帯情報を考慮するときには､たとえ､採血期間が短期間に限局されているとしても、「解析時年齢」と「採血時年齢」とを同等に扱うことはできない。

３　臨床情報取得時年齢の意味と、それが必要な場合

人の形質(phenotype)は常に変化する。例えば、あるときある疾患を発症していないとしても､5 年後

には発症しているかもしれないし、1 日後に発症しているかもしれない。10 年後には間違いなく発症するはずだったにも関わらず、不慮の事故で発症前に死亡するといったこともありえる。そういう意味では、ケース･コントロール関連解析に用いる「非ケース」としてのコントロールは、その臨床情報を取得した時点においてしか「非ケース」であるとして扱うことはできない。従って「非ケース」の臨床情報取得時年齢は「少なくともその年齢まで発病していない」という意味で用いるべき情報である。他方、これに対応する「ケース」の年齢情報は次の２通りがある。１つは臨床情報取得時年齢であり、これは「その年齢に達するまでに発病に至った」という意味で「非ケース」の臨床情報取得時年齢と対比して解析に用いるべきである。もう１つは、後で述べる「発病時年齢」である。こちらは「その年齢で発病した」という意味において、「非ケース」の臨床情報取得時年齢と対比して解析に用いるべきものである。

４　発病時年齢の意味と、それが必要な場合

ケースの発症に年齢が影響する場合には、どうしても避けて通れない情報である。実際にケースの発症年齢を解析に組み込むときにコントロールの年齢情報をどのようにして組み込むかは非常に重要である。そもそも、コントロール側についてある疾患にかかっているか否かの情報が得られない場合には「コントロール」は「非ケース」ではなく、「日本人集団(一部発症者を含む)」として扱う以外に方法はない。一方、「コントロール」として「非ケース」であるという臨床情報が得られた場合には、それと対比する形で情報を解析に利用する必要がある。

※　年齢が発病に関与している場合の年齢情報の組み込みについては「3-3-2-2-4 年齢が発病に影響を与える場合の情報組み込みモデル」を参照

3-3-2-2-2-2-2 年齢が発病に影響を与える場合の年齢情報組み込みモデル

条件設定

① Genotype　G1 と G2 が存在する Gi 群の人口を Pi で表す。Pi は t の関数である。

② Genotype Gi 群は疾患 X を発病した群 Gdi と、発病しない群 Ghi に分けられる。それぞれの群の人口は Pki(k=d or h)で表す。

③ Gki(k=d or h)群の死亡率を年齢 t の関数として Dki(t)と表す。

④ 新規発症は非罹患群 Ghi からのみおきるとして、その発症率は年齢の関数として Rhi(t)と表す。ケース vs 非ケースのケース・コントロールスタディを行うとして、それぞれの群から得られる年齢情報 Ta(解析時年齢)、Tb(採血時年齢)、Tc(臨床情報取得時年齢)、Td(発病年齢)

$$\mathrm{Ph}_1(t)+\mathrm{Ph}_2(t)+\mathrm{Pd}_1(t)+\mathrm{Pd}_2(t)=1$$
$$\mathrm{Phi}(T)=\mathrm{Phi}(o)-\int_0^T \mathrm{Phi}(t)\times \mathrm{Dhi}(t)dt-\int_0^T \mathrm{Phi}(t)\times \mathrm{Rhi}(t)dt$$
$$\mathrm{Pdi}(T)=\mathrm{Pdi}(o)-\int_0^T \mathrm{Pdi}(t)\times \mathrm{Ddi}(t)dt+\int_0^T \mathrm{Phi}(t)\times \mathrm{Rhi}(t)dt$$

ある年齢Tにおける人口がTの関数で与えられたケースを、このような集団からサンプリングすると、Genotype Ｉの占める率

$$\mathrm{Fdi}(t)=\frac{\mathrm{Pdi}(t)}{\mathrm{Pd}_1(t)+\mathrm{Pd}_2(t)}$$

コントロールをこのような集団からサンプリングすると、

$$\mathrm{Fhi}(t)=\frac{\mathrm{Phi}(t)}{\mathrm{Ph}_1(t)+\mathrm{Ph}_2(t)}$$

但し、このときの T は、臨床情報取得時年齢であるので、Tc を用いる。

以上が一般的なケース・非ケース集団からのサンプリングを行った場合である。

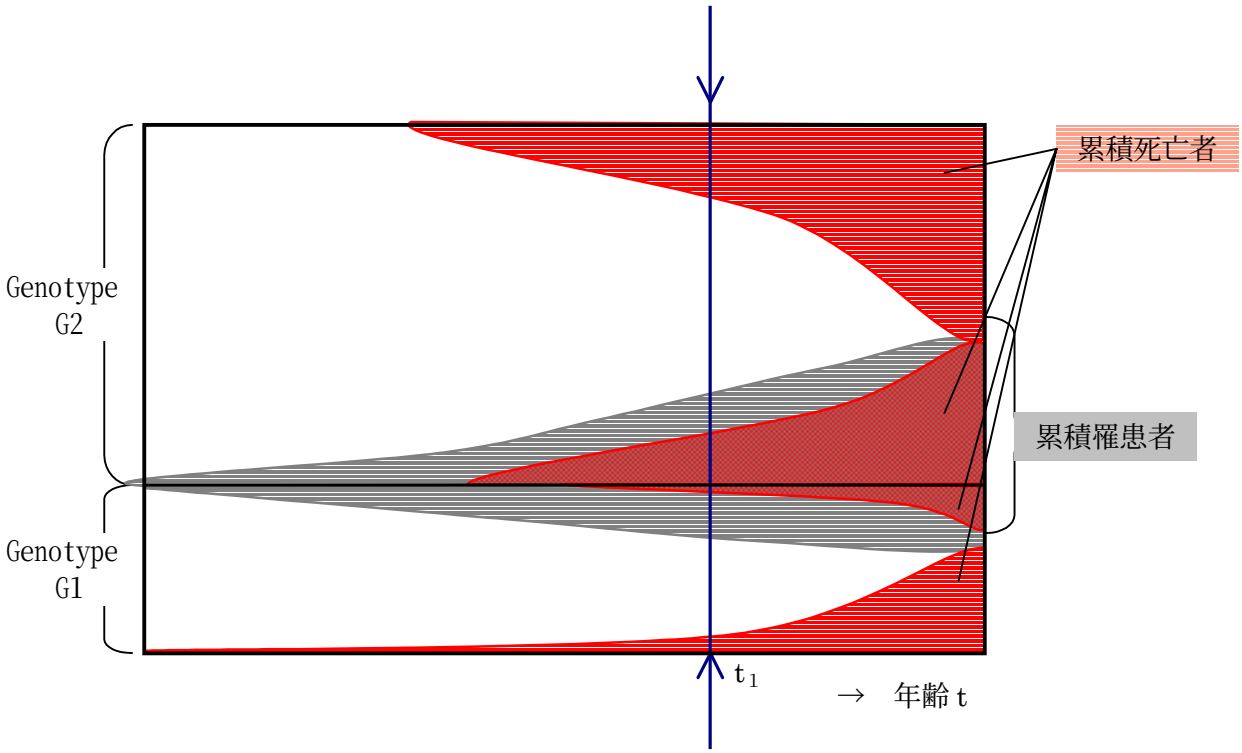
上記のモデルについて、実際のデータを解析することを考える。

ケースデータが $Tc=t_j$ につき Ndi 人、非ケースのデータが同様に Nhi 人　観測されたとすると、このような観測データが得られる。

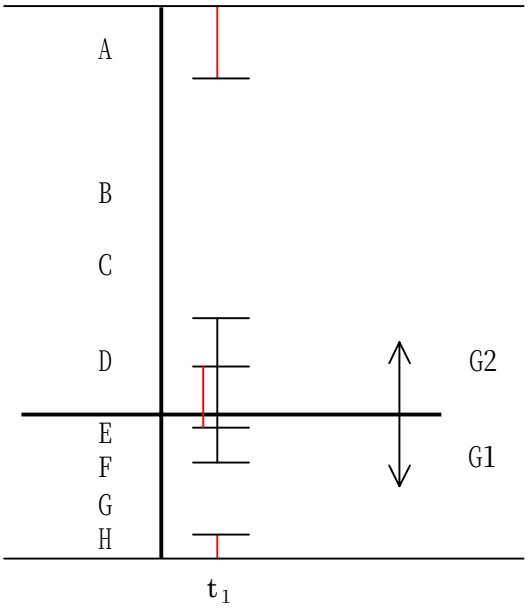
対数尤度 L は

$$L=\sum_T \mathrm{Nd}_1(T)\ln \mathrm{Fd}_1(T)+\sum_T \mathrm{Nd}_2(T)\ln \mathrm{Fd}_2(T)+\sum_T \mathrm{Nh}_1(T)\ln \mathrm{Fh}_1(T)+\sum_T \mathrm{Nh}_2(T)\ln \mathrm{Fh}_2(T)$$

ある観測データを基に最尤推定を行い、T の関数 $\mathrm{Rh}_1(T)$と $\mathrm{Rh}_2(T)$とに有意な差が認められれば、Genotype1/2 は発症に関連していると検定される。



年齢別 t_1 でのサンプリング



A：Genotype G2 で年齢 t_1 に到るまで、ある疾患に罹患せず死亡した画分
B：Genotype G2 で年齢 t_1 にて、ある疾患に罹患せず存在している画分
C：Genotype G2 で年齢 t_1 までに、ある疾患に罹患し生存している画分
D：Genotype G2 で年齢 t_1 までに、ある疾患に罹患し、かつ死亡している画分
E：Genotype G1 にて G2 の D に相当する画分
F：Genotype G1 にて G2 の C に相当する画分
G：Genotype G1 にて G2 の B に相当する画分
H：Genotype G1 にて G2 の A に相当する画分

簡略化したモデルを用いてみる。

Genotype1/2、疾患の有無 d/h は死亡率に無関係であるり、60 歳まで死亡率は 0、以後加齢比例して死亡率は上昇する。

$$Dd_1(t) = Dd_2(t) = Dh_1(t) = Dh_2(t)$$

$$\begin{cases} 0 \leqq t \leqq 60 & D^{**}(t) = 0 \\ t < 60 & D^{**}(t) = d(t - 60) \end{cases}$$

また、20 歳以前の発病はなし、以後加齢に比例して発症率は上昇し、Genotype G2 の発症率は G1 のそれの r 倍と固定されているとする。

$$Rh_2(t) = cRh_1(t)$$

$$Rh_1(t) = r(t - 20) \qquad (0 \leqq c \leqq 1)$$

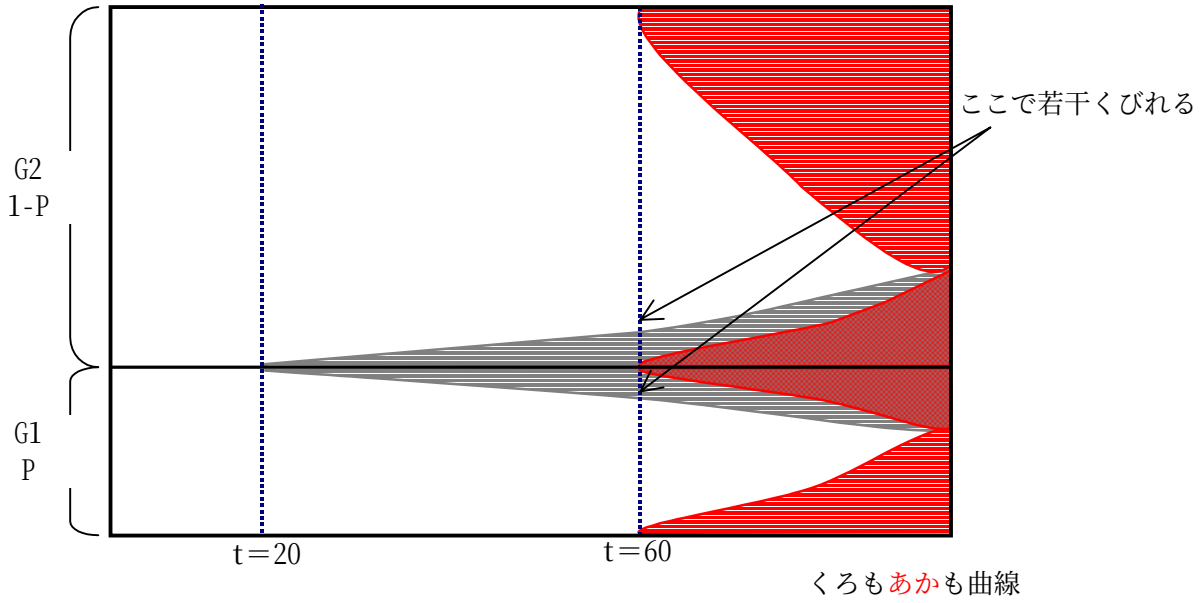
$$Pd_1(0) = 0$$

$$Pd_2(0) = 0$$

$$Ph_1(0) = P$$

$$Ph_2(0) = 1 - P$$

簡略化モデルの概略図



3-3-2-2-3　一般集団をコントロール集団として用いることについて

ケースは疾患という phenotype を所有する集団からランダムにサンプリングされている。コントロールは phenotype に関して通常とられる方法として 2 つのサンプリング方法がある。1 つは疾患 phenotype を所有していないことが確認された集団（非ケース集団）からのサンプリング、もう一つは疾患 phenotype について不明な集団（一般集団）からのサンプリングである。この外に家系内コントロールを採用する方法や、高齢になっても当該疾患を発病していない集団を採用する方法などもあるが、RIKEN SRC プロジェクトでは少なくともこの方法が実現することはないと思われる。

上記の特別なコントロール集団の場合を除けば、ケース集団とコントロール集団との間の差が大きいのは非ケース集団をコントロール集団として用いる場合である。しかしながら RIKEN SRC のケース・コントロール比較はケース集団 対 一般集団となっている。しかも一般集団は真の一般集団ではなく、ある別の疾患 phenotype を持った集団の混成である(擬似一般集団)。

コントロール集団として非ケース集団を用いずに擬似一般集団を用いることによる検定力の低下を決める要素は以下の通りである。

1. コントロール集団にケースもしくは将来的にケースになってしまう個体が多いほど検定力は低下する
   (1) 罹患率（高ければ高いほど検定力は低下する）
   (2) 若年者の割合（高ければ高いほど検定力は低下する。特に加齢とともに罹患率の上昇する疾患の場合）
   (3) 擬似一般集団を構成するその他の疾患 phenotype と当該疾患 phenotype とのオーバーラップの高さ

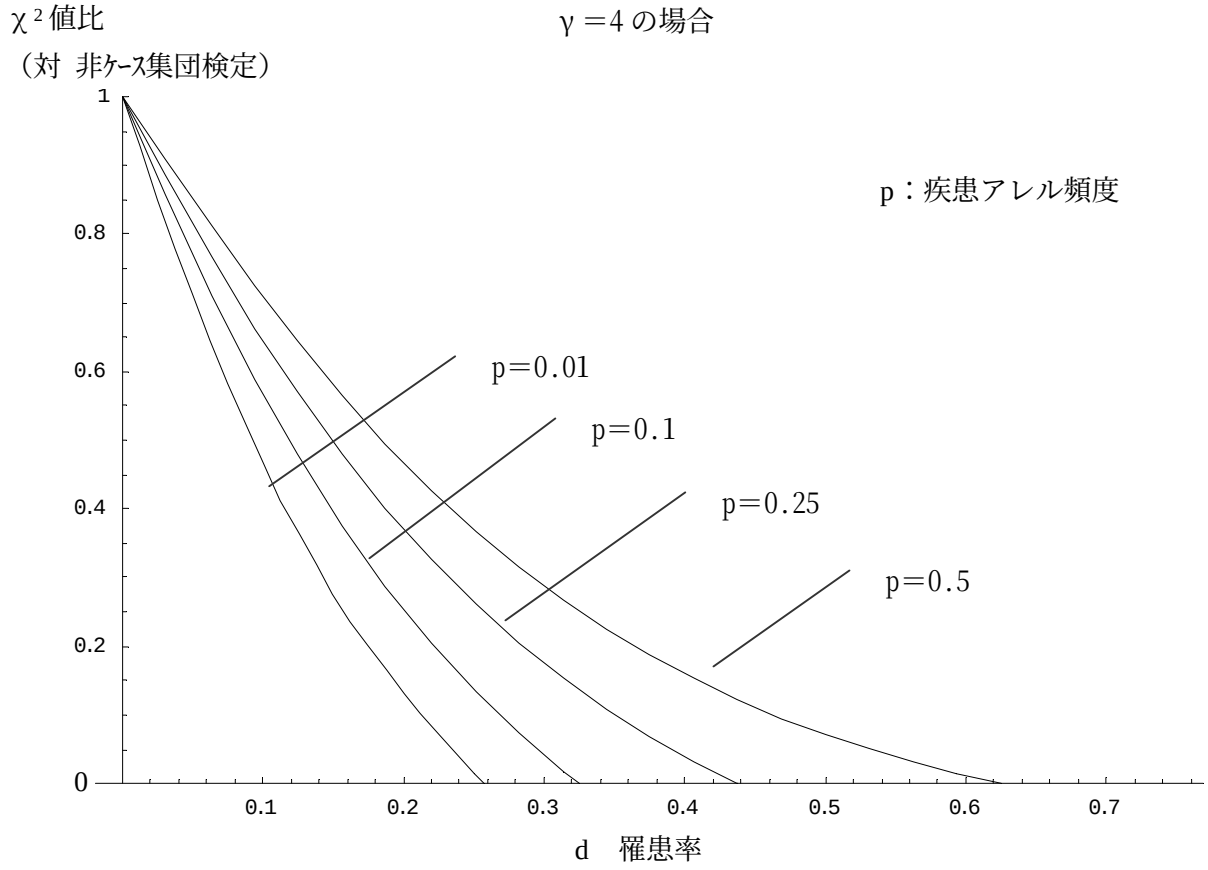
この検定力の低下をケース・コントロール アレル頻度 χ^2検定でシミュレートする。
非ケース集団をコントロール集団とした場合に得られるχ^2値に対する、コントロール集団にケースが混入している場合に得られるχ^2値の比をグラフ化したものである(図 1～図 4)。

γ^2型
優性遺伝型
劣性遺伝型

の 3 つの型について描図する。

(擬似)一般集団内に混入するケース個体の割合がもたらす検定力の低下の程度を独自にシミュレーションしたい場合には、エクセル“コントロール群へのケースの混入シミュレーション”を用いてその程度についてイメージをつかむことができる。このエクセルファイルは罹患率によるケースの混入のみをシミュレートするが、その他の原因による混入も割合の点のみが問題なので適宜読み換えてシミュレートすることができる。

図 1. 一般集団をコントロールとして用いるときの罹患率による χ^2 値の減衰曲線
(γ^2 型ローカスの場合)



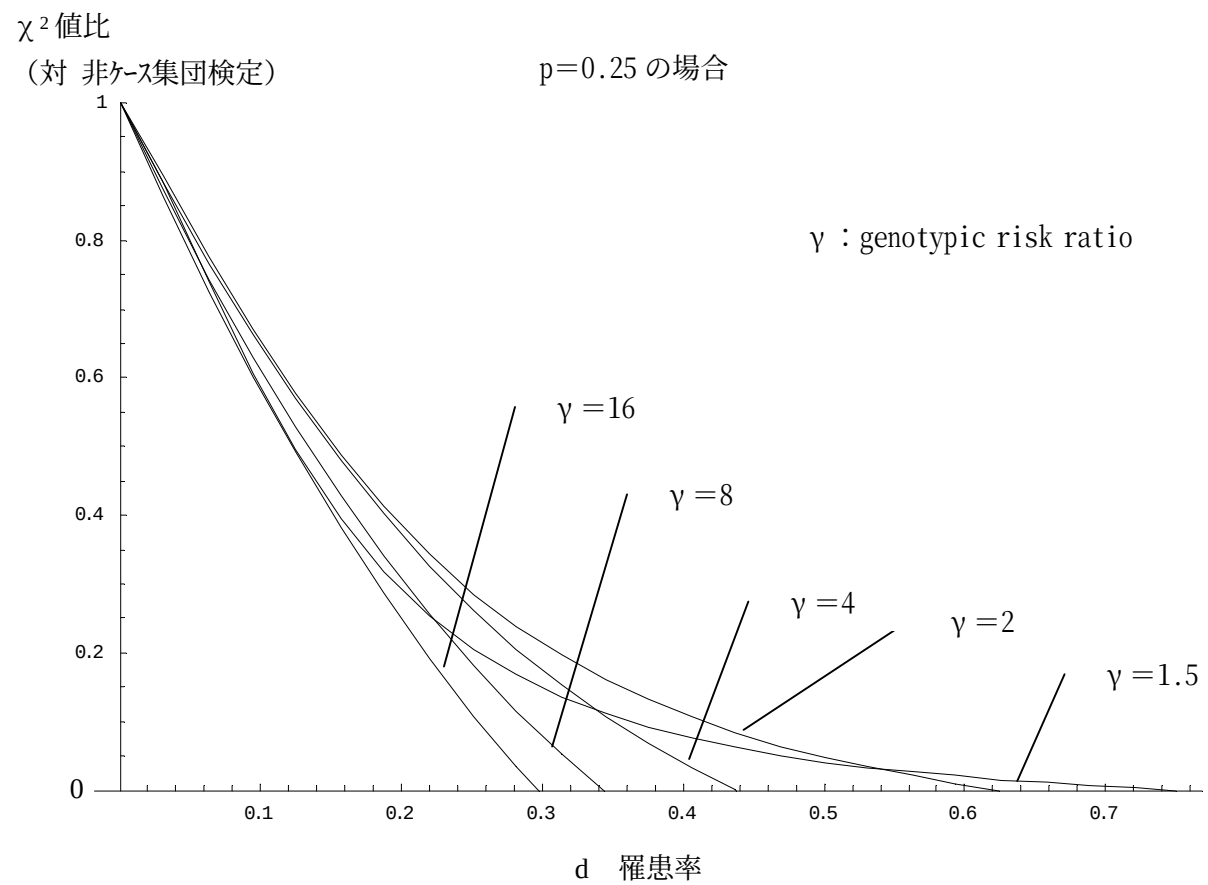
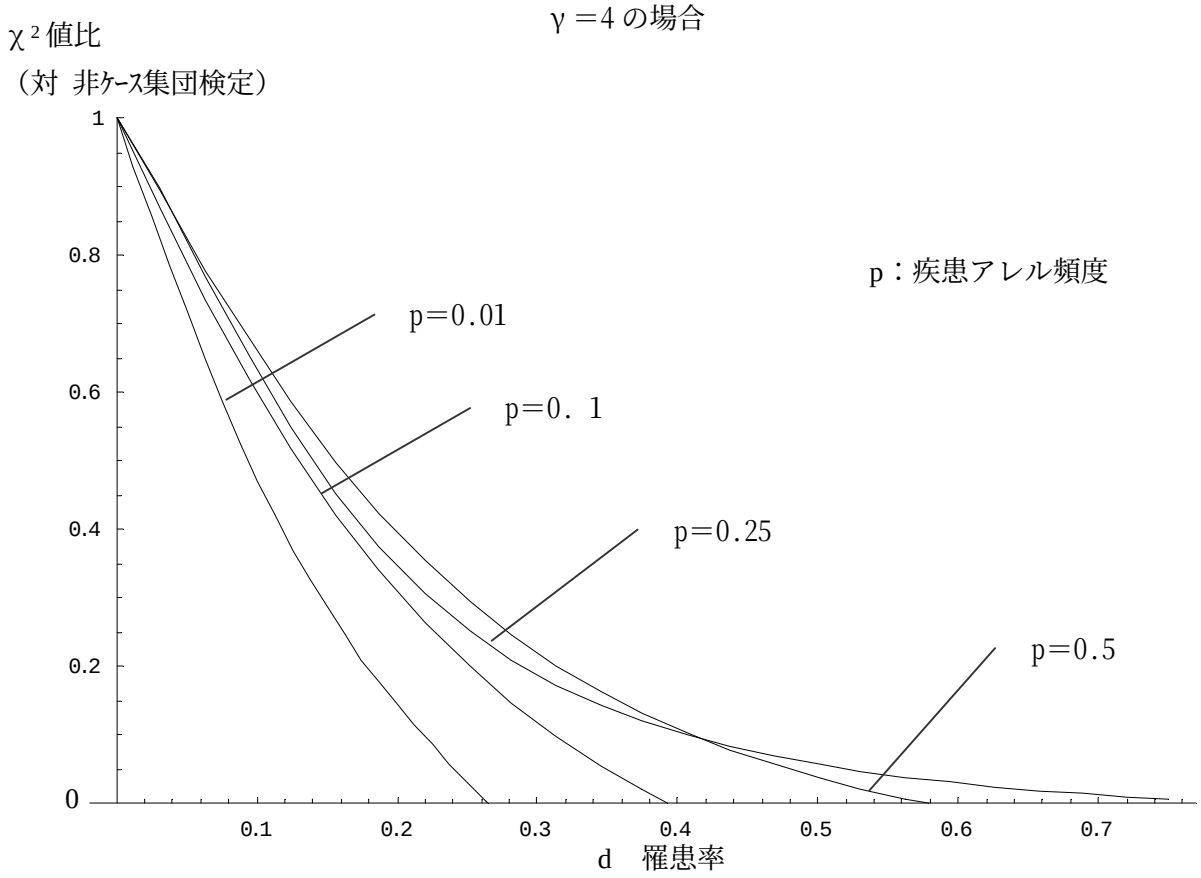


図 2. 一般集団をコントロールとして用いるときの罹患率による χ^2 値比の減衰曲線
(優性遺伝型ローカスの場合)



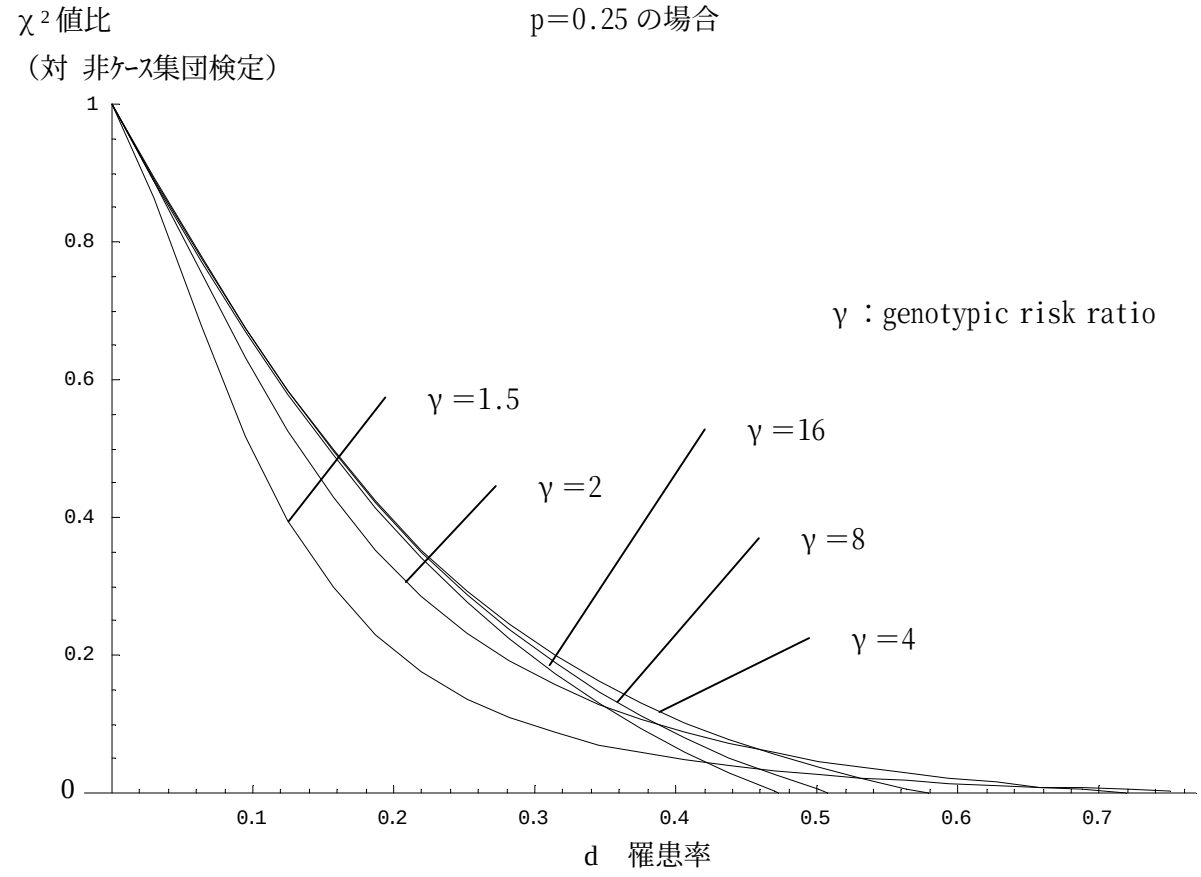
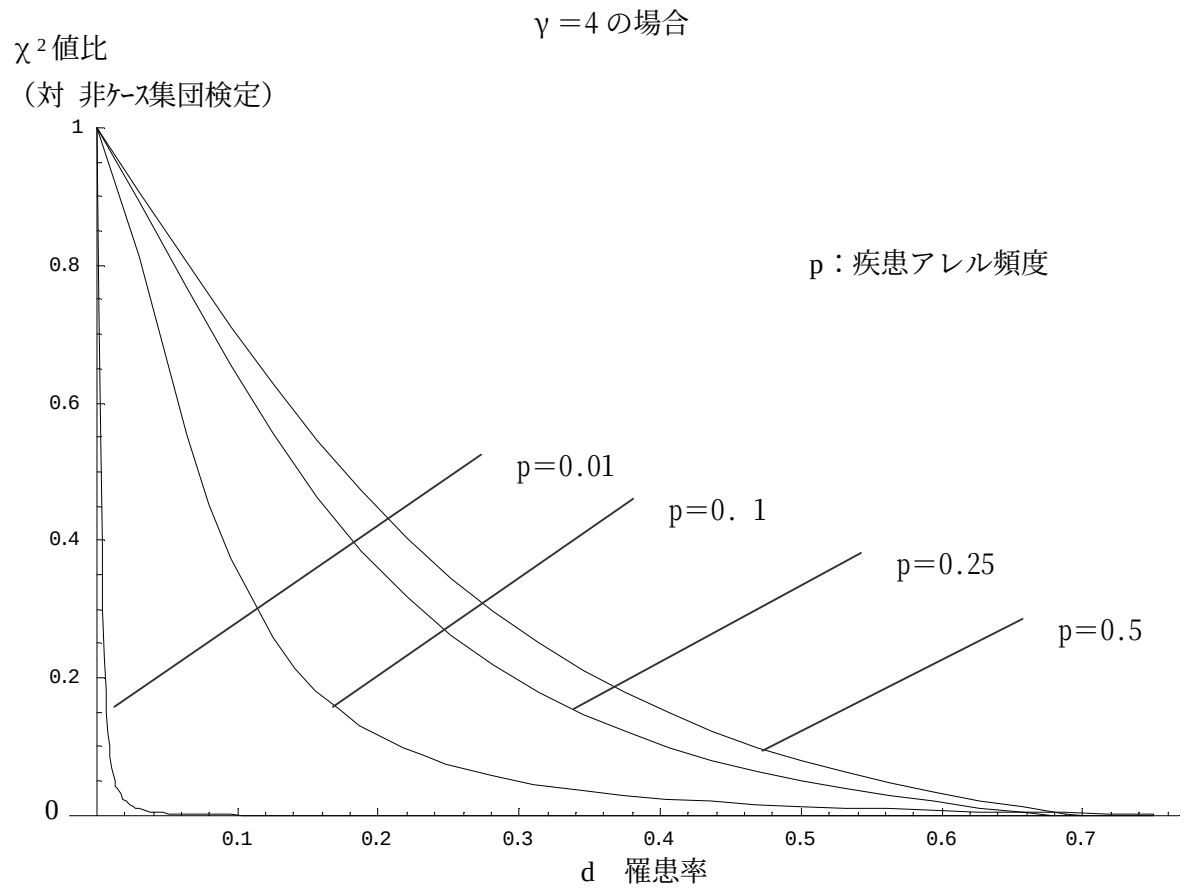


図 3. 一般集団をコントロールとして用いるときの罹患率による χ^2 値比の減衰曲線
(劣性遺伝型ローカスの場合)



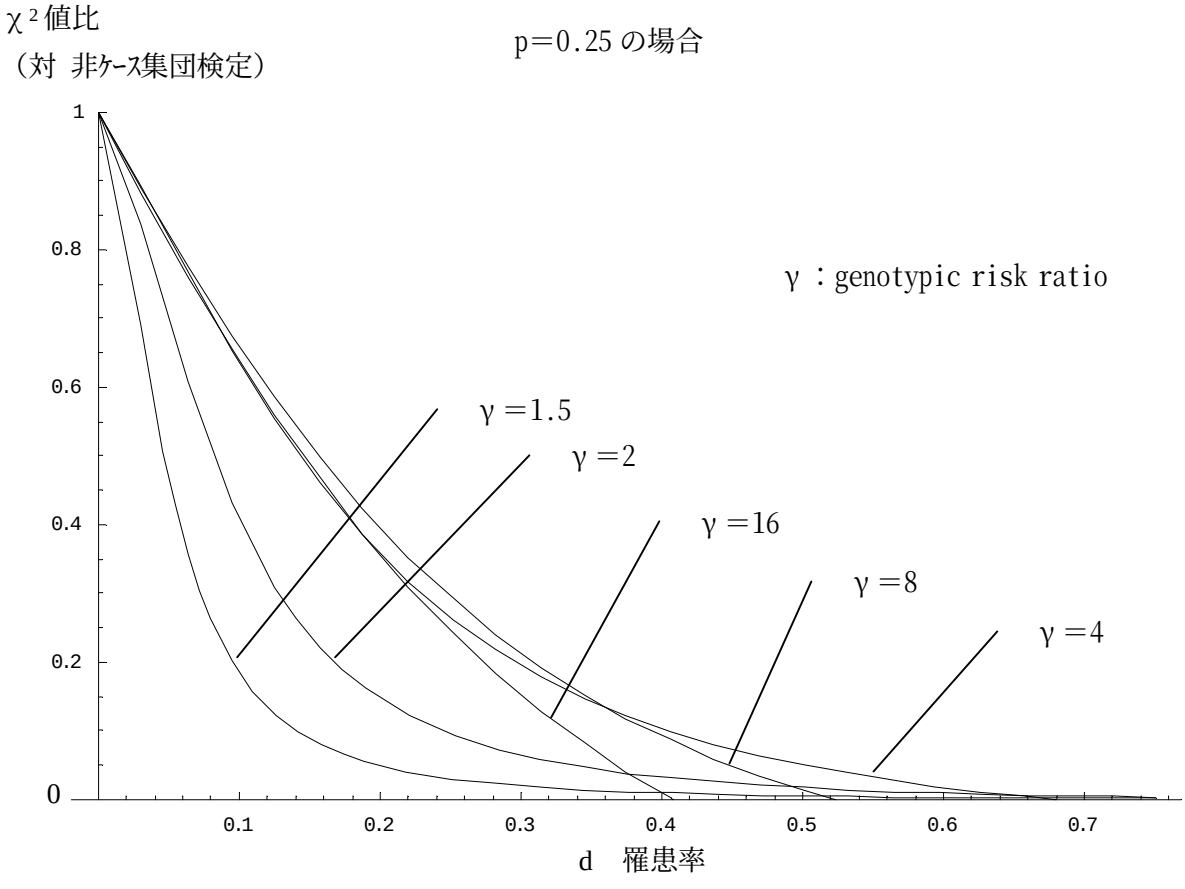
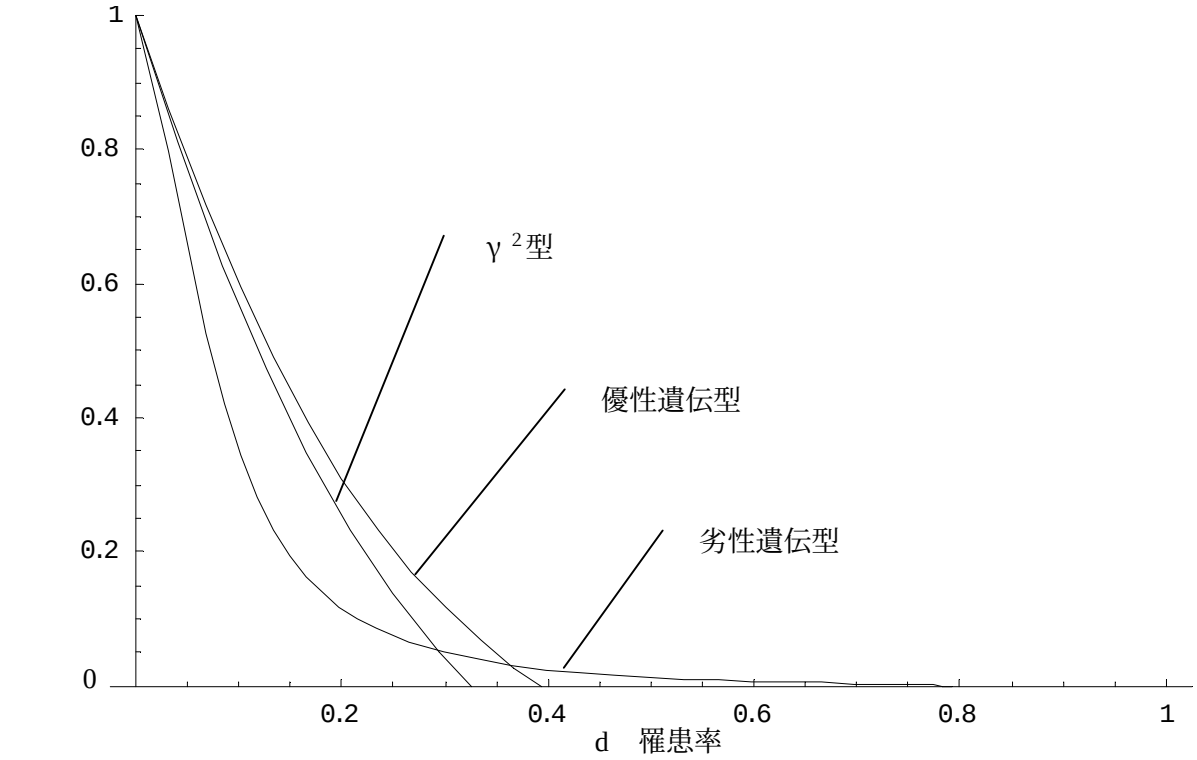
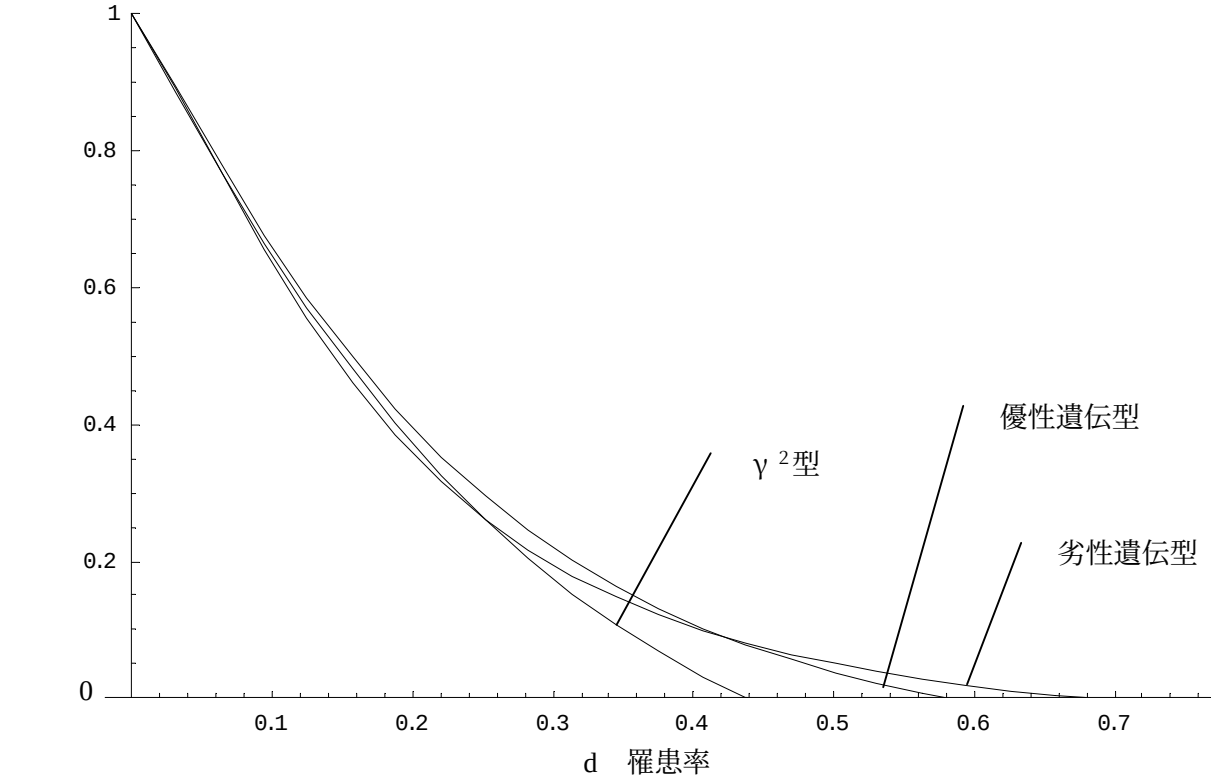


図 4. 一般集団をコントロールとして用いるときの罹患率による χ^2 値比の減衰曲線

genotypic relative risk 3 型の比較　p=0.1, γ =4 の場合



genotypic relative risk 3 型の比較　p=0.25, γ =4



グラフ描写のための数式を以下に示す。

γ^2型

$$\chi^2\text{値比}= \frac{\left(-d+\varepsilon p^2+pq+\varepsilon pq+q^2\right)\left(-d\varepsilon+\varepsilon p^2+pq+\varepsilon pq+q^2\right)\left(\varepsilon^2 p-2\varepsilon^2 p^3+\varepsilon^2 p^5+q-4\varepsilon p^2 q-2\varepsilon^2 p^2 q+2\varepsilon p^4 q+3\varepsilon^2 p^4 q-2pq^2-4\varepsilon pq^2+p^3 q^2+6\varepsilon p^3 q^2+3\varepsilon^2 p^3 q^2-2q^3+3p^2 q^3+6\varepsilon p^2 q^3+\varepsilon^2 p^2 q^3+3pq^4+2\varepsilon pq^4+q^5\right)}{(p+q)(\varepsilon p+q)^2\left(-d\varepsilon-2d^2\varepsilon-d^3\varepsilon+2d\varepsilon p^2+2d^2\varepsilon p^2+\varepsilon^2 p^2+2d\varepsilon^2 p^2+d^2\varepsilon^2 p^2-d\varepsilon p^4-2\varepsilon^2 p^4-2d\varepsilon^2 p^4+\varepsilon^2 p^6+pq+2dpq+d^2 pq+4d\varepsilon pq+4d^2\varepsilon pq+\varepsilon^2 pq+2d\varepsilon^2 pq+d^2\varepsilon^2 pq-dp^3 q-4\varepsilon p^3 q-6d\varepsilon p^3 q-4\varepsilon^2 p^3 q-5d\varepsilon^2 p^3 q+2\varepsilon p^5 q+4\varepsilon^2 p^5 q+q^2+2dq^2+d^2 q^2+2d\varepsilon q^2+2d^2\varepsilon q^2-2p^2 q^2-4dp^2 q^2-8\varepsilon p^2 q^2-10d\varepsilon p^2 q^2-2\varepsilon^2 p^2 q^2-4d\varepsilon^2 p^2 q^2+p^4 q^2+8\varepsilon p^4 q^2+6\varepsilon^2 p^4 q^2-4pq^3-5dpq^3-4\varepsilon pq^3-6d\varepsilon pq^3-d\varepsilon^2 pq^3+4p^3 q^3+12\varepsilon p^3 q^3+4\varepsilon^2 p^3 q^3-2q^4-2dq^4-d\varepsilon q^4+6p^2 q^4+8\varepsilon p^2 q^4+\varepsilon^2 p^2 q^4+4pq^5+2\varepsilon pq^5+q^6\right)}$$

優性遺伝型

$$\chi^2\text{値比}= \frac{\left(-d\varepsilon+\varepsilon p^2+2\varepsilon pq+q^2\right)\left(-d\varepsilon p+\varepsilon p^3-dq+3\varepsilon p^2 q+pq^2+2\varepsilon pq^2+q^3\right)\left(\varepsilon^2 p^3-2\varepsilon^2 p^5+\varepsilon^2 p^7+3\varepsilon^2 p^2 q-10\varepsilon^2 p^4 q+7\varepsilon^2 p^6 q+2\varepsilon pq^2+\varepsilon^2 pq^2-4\varepsilon p^3 q^2-16\varepsilon^2 p^3 q^2+2\varepsilon p^5 q^2+19\varepsilon^2 p^5 q^2+q^3-12\varepsilon p^2 q^3-8\varepsilon^2 p^2 q^3+10\varepsilon p^4 q^3+25\varepsilon^2 p^4 q^3-2pq^4-8\varepsilon pq^4+p^3 q^4+18\varepsilon p^3 q^4+16\varepsilon^2 p^3 q^4-2q^5+3p^2 q^5+14\varepsilon p^2 q^5+4\varepsilon^2 p^2 q^5+3pq^6+4\varepsilon pq^6+q^7\right)}{(p+q)\left(\varepsilon p^2+2\varepsilon pq+q^2\right)^2\left(-d\varepsilon^2 p-2d^2\varepsilon^2 p-d^3\varepsilon^2 p+\varepsilon^2 p^3+4d\varepsilon^2 p^3+3d^2\varepsilon^2 p^3-2\varepsilon^2 p^5-3d\varepsilon^2 p^5+\varepsilon^2 p^7-d\varepsilon q-2d^2\varepsilon q-d^3\varepsilon q+2d\varepsilon p^2 q+2d^2\varepsilon p^2 q+3\varepsilon^2 p^2 q+10d\varepsilon^2 p^2 q+7d^2\varepsilon^2 p^2 q-d\varepsilon p^4 q-10\varepsilon^2 p^4 q-14d\varepsilon^2 p^4 q+7\varepsilon^2 p^6 q+2\varepsilon pq^2+8d\varepsilon pq^2+6d^2\varepsilon pq^2+\varepsilon^2 pq^2+4d\varepsilon^2 pq^2+3d^2\varepsilon^2 pq^2-4\varepsilon p^3 q^2-8d\varepsilon p^3 q^2-16\varepsilon^2 p^3 q^2-22d\varepsilon^2 p^3 q^2+2\varepsilon p^5 q^2+19\varepsilon^2 p^5 q^2+q^3+2dq^3+d^2 q^3+2d\varepsilon q^3+2d^2\varepsilon q^3-dp^2 q^3-12\varepsilon p^2 q^3-16d\varepsilon p^2 q^3-8\varepsilon^2 p^2 q^3-13d\varepsilon^2 p^2 q^3+10\varepsilon p^4 q^3+25\varepsilon^2 p^4 q^3-2pq^4-3dpq^4-8\varepsilon pq^4-10d\varepsilon pq^4-2d\varepsilon^2 pq^4+p^3 q^4+18\varepsilon p^3 q^4+16\varepsilon^2 p^3 q^4-2q^5-2dq^5-d\varepsilon q^5+3p^2 q^5+14\varepsilon p^2 q^5+4\varepsilon^2 p^2 q^5+3pq^6+4\varepsilon pq^6+q^7\right)}$$

劣性遺伝型

$$\chi^2\text{値比}= \frac{\left(-d+\varepsilon p^2+2pq+q^2\right)\left(-d\varepsilon p+\varepsilon p^3-dq+2p^2 q+\varepsilon p^2 q+3pq^2+q^3\right)\left(\varepsilon^2 p^3-2\varepsilon^2 p^5+\varepsilon^2 p^7+p^2 q+2\varepsilon p^2 q-8\varepsilon p^4 q-2\varepsilon^2 p^4 q+4\varepsilon p^6 q+3\varepsilon^2 p^6 q+3pq^2-8p^3 q^2-12\varepsilon p^3 q^2+4p^5 q^2+14\varepsilon p^5 q^2+3\varepsilon^2 p^5 q^2+q^3-16p^2 q^3-4\varepsilon p^2 q^3+16p^4 q^3+18\varepsilon p^4 q^3+\varepsilon^2 p^4 q^3-10pq^4+25p^3 q^4+10\varepsilon p^3 q^4-2q^5+19p^2 q^5+2\varepsilon p^2 q^5+7pq^6+q^7\right)}{(p+q)\left(\varepsilon p^2+2pq+q^2\right)^2\left(-d\varepsilon p-2d^2\varepsilon p-d^3\varepsilon p+2d\varepsilon p^3+2d^2\varepsilon p^3+\varepsilon^2 p^3+2d\varepsilon^2 p^5+d^2\varepsilon^2 p^5-d\varepsilon p^5-2\varepsilon^2 p^5-2d\varepsilon^2 p^5+\varepsilon^2 p^7-dq-2d^2 q-d^3 q+p^2 q+4dp^2 q+3d^2 p^2 q+2\varepsilon p^2 q+8d\varepsilon p^2 q+6d^2\varepsilon p^2 q+3\varepsilon^2 p^2 q+10d\varepsilon^2 p^2 q+7d^2\varepsilon^2 p^2 q-2dp^4 q-8\varepsilon p^4 q-10d\varepsilon p^4 q-2\varepsilon^2 p^4 q-3d\varepsilon^2 p^4 q+4\varepsilon p^6 q+3\varepsilon^2 p^6 q+3pq^2+10dpq^2+7d^2 pq^2+2d\varepsilon pq^2+2d^2\varepsilon pq^2-8p^3 q^2-13dp^3 q^2-12\varepsilon p^3 q^2-16d\varepsilon p^3 q^2-d\varepsilon^2 p^3 q^2+4p^5 q^2+14\varepsilon p^5 q^2+3\varepsilon^2 p^5 q^2+q^3+4dq^3+3d^2 q^3-16p^2 q^3-22dp^2 q^3-4\varepsilon p^2 q^3-8d\varepsilon p^2 q^3+16p^4 q^3+18\varepsilon p^4 q^3+\varepsilon^2 p^4 q^3-10pq^4-14dpq^4-d\varepsilon pq^4+25p^3 q^4+10\varepsilon p^3 q^4-2q^5-3dq^5+19p^2 q^5+2\varepsilon p^2 q^5+7pq^6+q^7\right)}$$

3-3-2-3　連鎖不平衡

3-3-2-3-1 連鎖不平衡とは

連鎖不平衡(Linkage Disequilibrium(LD))は、allelic association とも呼ばれる。その説明を以下に行う。

ある集団において、2 つのローカスに多型が存在するとする。片方の多型がどのアレルを持つかということと、もう一方の多型がどのアレルを持つかということが、完全にランダム（独立）な場合、その 2 つの多型の間には連鎖不平衡は無いという。一方、片方の多型のあるアレルを持つときに、もう一方の多型の特定のアレルを持つ傾向がある場合に、この 2 つのアレルは連鎖不平衡にあるという。この連鎖不平衡は世代を経るごとに失われていくのだが、その理由は減数分裂のときに発生する組み換えのためである。この組み換えの結果生じた親と異なるアレルの組み合わせを持つ染色体を組み換え体(recombinant)と呼ぶ。従って、多型間の組み換えが起こりやすいほど速やかに失われる。組み換えの頻度はゲノム上の場所によって異なるが、ある特定の領域に注目する限り距離が遠いほど組み換え頻度は高くなる。

今、ある集団（例えば日本人）の中で、互いに遺伝的に無関係な 2 人の 2 多型を比べたとする。本当に無関係であれば、その 2 多型はどんなに近接していようとも連鎖していない。しかしながら、ヒトはその祖先において、ごく限られた個体数であったために、互いに親族関係が無いと考えられている者同士であっても、共通の祖先を有するとみなせる。そうすると、染色体のある限られた範囲に関してはその共通の祖先が持っていたアレルの組み合わせを持っていることがわかっている。この特定のアレルの組み合わせが、ある集団内で多いことを連鎖不平衡が存在すると呼ぶ。

この連鎖不平衡の程度を説明するのは Linkage disequilibrium (LD) indices である。この複数の指標はそれぞれ異なる数式にて表されるが、それらが表現しようとしている内容は同一である。その内容とは、世代が十分若い時期には完全に連鎖していた多型同士が、世代を経るに従ってその間に発生した組み換えの結果、組み換え体(recombinant)を生じる。この組み換え体が占める程度が LD の程度である。LD の程度を表す指標として複数の定義式が提案されており、それぞれに特徴があるが、ここでは比較的論文などによく登場する D'の定義式をあげ、その他の指標に関しては“参考 いろいろな Linkage Disequilibrium 指標とその挙動”を参照してほしい。

D'の定義式

今、SNP A と SNP B とがあり、それぞれのアレルを A,a,B,b とする。SNP A と SNP B とが作る 4 haplotypes を AB, Ab, aB, ab とし、それぞれの haplotype 頻度を P_{AB}, P_{Ab}, P_{aB}, P_{ab} とすると、

$$D' = (P_{AB} \times P_{ab} - P_{Ab} \times P_{aB}) / \mathrm{Minimum}(((P_{AB}+P_{aB}) \times (P_{aB}+P_{ab})),((P_{AB}+P_{Ab}) \times (P_{Ab}+P_{ab})))$$

但し、$\mathrm{Minimum}(((P_{AB}+P_{aB}) \times (P_{aB}+P_{ab})),((P_{AB}+P_{Ab}) \times (P_{Ab}+P_{ab})))$は、
$(P_{AB}+P_{aB}) \times (P_{aB}+P_{ab})$と$(P_{AB}+P_{Ab}) \times (P_{Ab}+P_{ab})$との内、値の小さい方をとることを意味する。

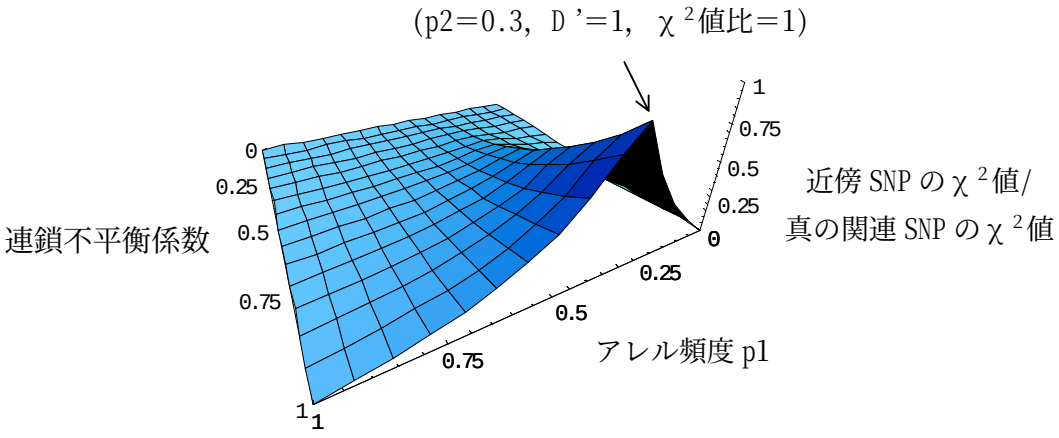
2 SNP が作る 9 genotype 別の観測データから、2 SNP が作る 4 haplotype 頻度を計算することはエクセルファイル“2SNPsHaplotype 頻度推定”により可能である。また、4 haplotype 頻度のデータから D'を算出することは、上の式を利用してもよいが、エクセルファイル“D'計算”でも可能である。

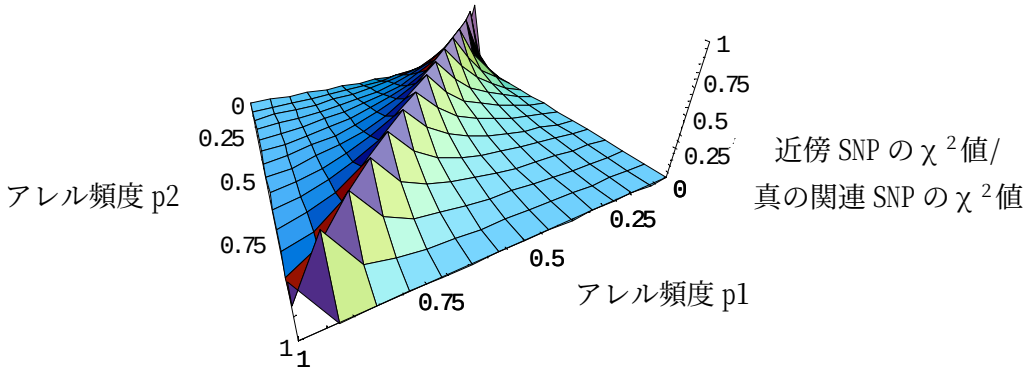
3-3-2-3-3 真のローカスに近接するSNPの､LDの強さとアレル頻度の乖離が検定力に及ぼす影響について

真の疾患関連 SNP における χ^2 値とその SNP と連鎖不平衡関係にある SNP における χ^2 値との関係をグラフで示す。

第 1 のグラフは、真の関連 SNP の疾患アレル頻度 p2 を 0.3 とした時に、近傍 SNP のアレル頻度 p1 と 2 SNP 間の連鎖不平衡係数とによって近傍 SNP の χ^2 値がどのように減衰するかを示したグラフである。2 SNP のアレル頻度が等しく、完全連鎖をしている場合に相当する点にピークがあり、そこから全方向に減衰している様子がわかる。

第 2 のグラフは、固定した連鎖不平衡係数の下で、2 SNP のアレル頻度が変化する場合、どのように減衰するかを示したグラフである。真の関連 SNP の疾患アレル頻度と近傍 SNP の連鎖しているアレル頻度とが一致している時に χ^2 値のピークが認められる。また､同時に､同じ連鎖不平衡係数の場合で 2 SNP のアレル頻度が等しい場合でも、疾患アレル頻度が中程度の場合をピークに、得られる χ^2 値が減少することも読み取れる。さらに、疾患アレル頻度が中程度の場合は、2 SNP 間のアレル頻度の差が開いても、χ^2 値の減衰の程度が緩やかであることも読み取れる。





3-3-2-3-5 参考 連鎖不均衡(LD)の実際及び間接関連検出用の SNP マーカーと真のローカスとの関係について

以下のグラフはヒトゲノム上のある領域における連鎖不平衡の広がりを実際の genotype data から haplotype 頻度を EM-algorithm で推定してプロットしたものである。
このグラフが示していることは以下のようにまとめられる。

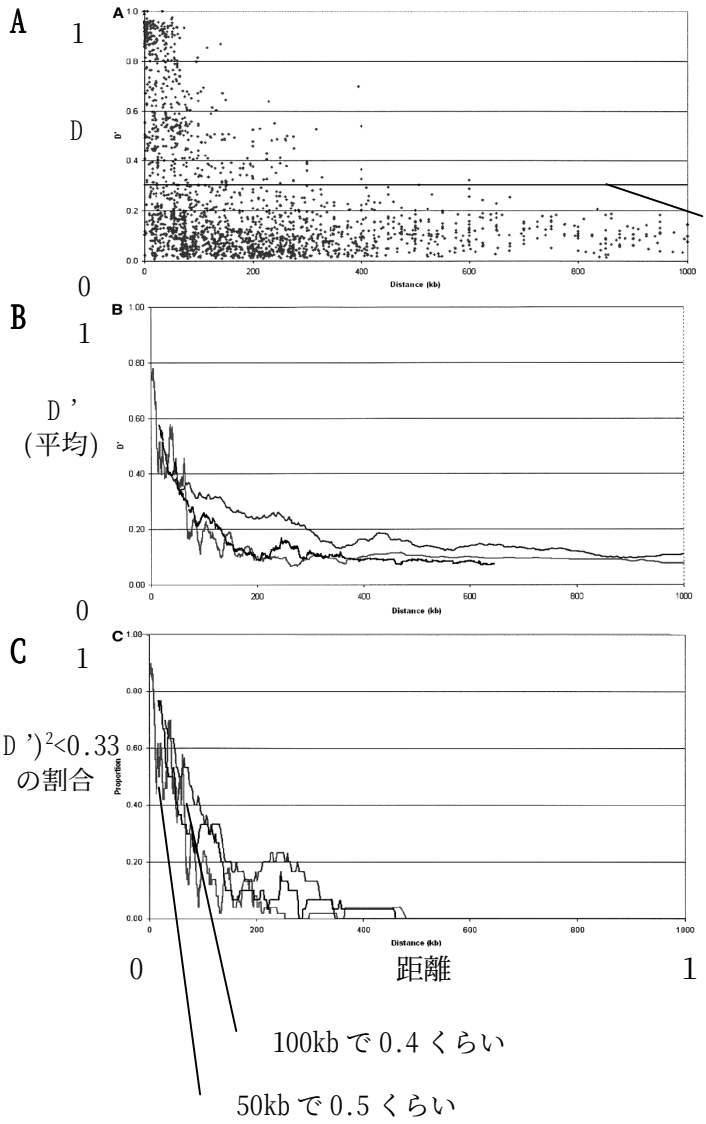
(1) 平均してみると LD の強さ(ここでは LD の指標として D'を使用している)は距離に応じて指数関数的に減少していく
(2) 個々のローカス間の LD はばらつきが大きく非常に近くても LD をほとんど認めない場合もあれば相当程度遠くてもある程度の LD を維持している場合もある
(3) SNP をマーカーとして利用し、間接関連を検出して、その近傍の真の関連ローカスに到達しようとする場合、真のローカスとマーカーローカスとの間にはある程度の LD がなければ現実的な検体数で関連を認めることは難しい。複合遺伝性疾患の関連遺伝子を同定するような場合でそもそも真のローカスとの直接関連を見出すのに必要な検体数が大きい場合には、LD として$(D')^2$に換算して$(D')^2>0.33$程度が必要であるとされている。
(4) ある真のローカスから前後 50kb の範囲にある SNP のうち約半数は$(D')^2>0.33$の条件を満たし、前後 100kb の範囲にある SNP のうち約 4 割は$(D')^2>0.33$の条件を満たしていた。
(5) このことから真のローカスから 50kb 程度の距離にある SNP３つを用いれば、そのうちの少なくとも１つが真のローカスと$(D')^2>0.33$の条件を満たす確率は

$1-(0.5)^3=0.875$ であり、

100kb 程度の距離にある SNP3 つを用いれば、そのうちの少なくとも１つが真のローカスと$(D')^2>0.33$の条件を満たす確率は

$1-(0.6)^3=0.784$ である。

ヒトゲノム上の 3 領域、127 マーカーのデータから 2108 マーカーペアの LD を求めた結果
－GR Abecasis e al, Am J Hum Genet 68：191-197, (2001)



A：2 点間の D’と距離との関係

$(D')^2>0.33$
のレベル

B：A の図はばらつきがあってやや解りにくいので 30 ペアごとの D’と距離との関係の平均をとることで集約させた図

C：LD の存在に依存した association study では、association が$(D')^2$に比例して減衰するので$(D')^2>0.33$ が解析に有効な LD の広がりとみなし、$(D')^2>0.33$ となるマーカーペアの割合が距離によってどのように減少するかを示した図

3-3-2-3-6-1 参考 いろいろな Linkage Disequilibrium 指標とその挙動

Reference：Devlin B and Risch N, A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29：311-322, (1995)

連鎖不平衡の程度を表現する指標として複数のものが提案されており、同一の 2 SNP データ間の関係を表現する場合にも、用いる指標により異なる値が得られる。主なものとしては、以下のように 5 つの LD の指標(Δ、D'、δ、d、Q)及び、d^2を加えた計 6 つの指標が用いられる。
それぞれの指標の特徴をつかむために挙動を視覚化する。
要点としては、各指標の値が、
(1)完全連鎖のときにいつも 1 をとるか、否かの点、
(2)Allele frequency の影響をどのように受けるかという点、
(3)ローカス間の対称性があるか、否かの点、
の３点が主な相違点となる。

Haplotype frequency を以下のように表記する

Haplotype	Frequency
AB	P_{AB}
Ab	P_{Ab}
aB	P_{aB}
ab	P_{ab}

P_{xx}の条件としては

$$P_{AB}+P_{ab}+P_{aB}+P_{ab}=1$$

$$Min(P_{xx})=P_{aB}$$

とすることで、あらゆるパターンを網羅することとする。

全ての指標の計算式の分子(N：Numerator)は共通しており、それは

$$N=P_{AB}\times P_{ab}-P_{Ab}\times P_{aB}$$

と書ける
この N を用いてそれぞれの指標を書き下すと

$$\triangle = N/SQRT((P_{AB}+P_{Ab})\times(P_{aB}+P_{ab})\times(P_{AB}+P_{aB})\times(P_{Ab}+P_{ab}))$$

$$D' = N/Minimum((P_{AB}+P_{aB})\times(P_{aB}+P_{ab}),((P_{AB}+P_{Ab})\times(P_{Ab}+P_{ab}))$$

$$\delta = N/((P_{AB}+P_{aB})\times P_{ab})$$

$$d = N/((P_{AB}+P_{aB})\times(P_{Ab}+P_{ab}))$$

$$Q = N/(P_{AB}\times P_{ab}+P_{Ab}\times P_{aB})$$

$$d^2 = (N/((P_{AB}+P_{aB})\times(P_{Ab}+P_{ab})))^2$$

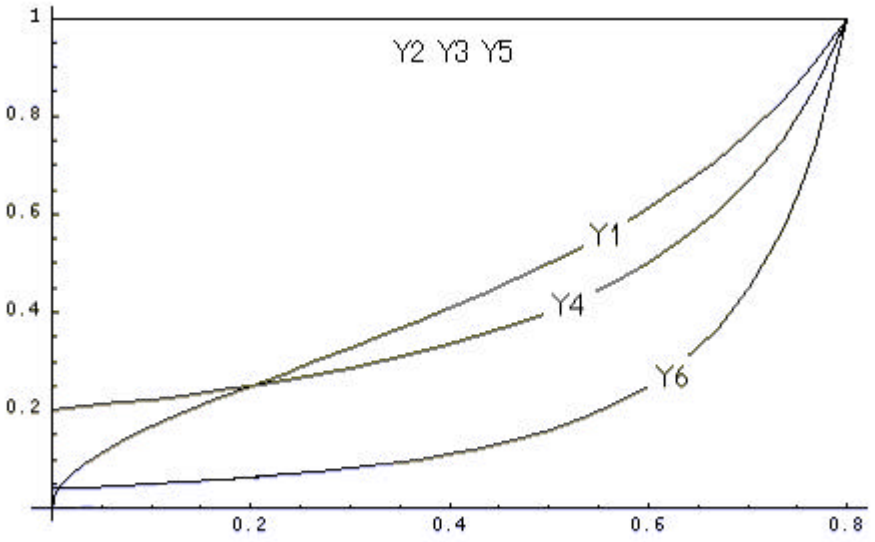
となる。

また、組換え率θとの関係で、世代数nを用いておのおのの指標を書き表すと

$$\triangle\times(1/P_{ab})\times SQRT((P_{Ax}/P_{xB})\times(1-P_{Ax})\times(1-P_{xB}))=(1-\theta)^n$$
$$D'\times(1+P_{ab}/P_{AB})\times(1+P_{aB}/P_{ab})/(1+P_{aB}/P_{AB})=(1-\theta)^n$$
$$\delta=(1-\theta)^n$$
$$d\times P_{xb}/P_{ab}=(1-\theta)^n$$
$$Q\times(1+(P_{ab}/P_{ab}-1)/(P_{AB}/P_{aB}+1))=(1-\theta)^n$$

という関係にあり、
δのみが、アレル頻度に関係なくθのみの関数となっている。

挙動描図1
完全連鎖の場合の挙動(1)
条件：$P_{aB}=0$，$P_{AB}+P_{Ab}=0.8$，$P_{ab}=0.2$，$0<P_{Ab}$，$P_{AB}<0.8$
として、$P_{AB}=w$**を変動範囲[0,0.8]としてLD指標をwの関数としてグラフ化すると以下のようになる。**



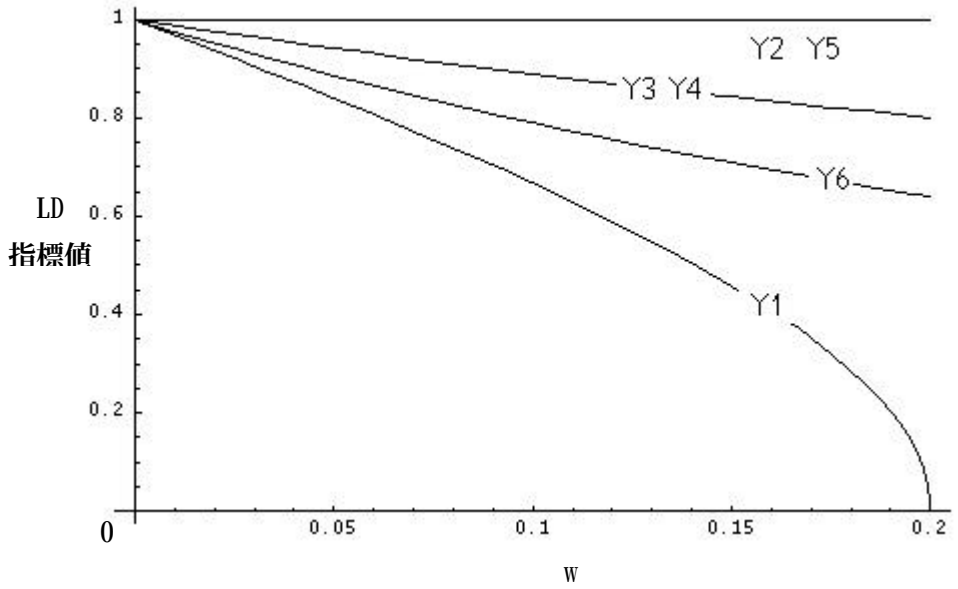
Y1＝Δ
Y2＝D’
Y3＝δ
Y4＝d
Y5＝Q
Y6＝d^2

挙動描図 2

完全連鎖の場合の挙動(2)

条件：$P_{Ab}=0$，$P_{AB}=0.8$，$P_{aB}+P_{ab}=0.2$，$0<P_{aB}<0.2$

として、$P_{aB}=w$**を変動範囲[0,0.2]として LD 指標を** w **の関数としてグラフ化すると以下のようになる。**



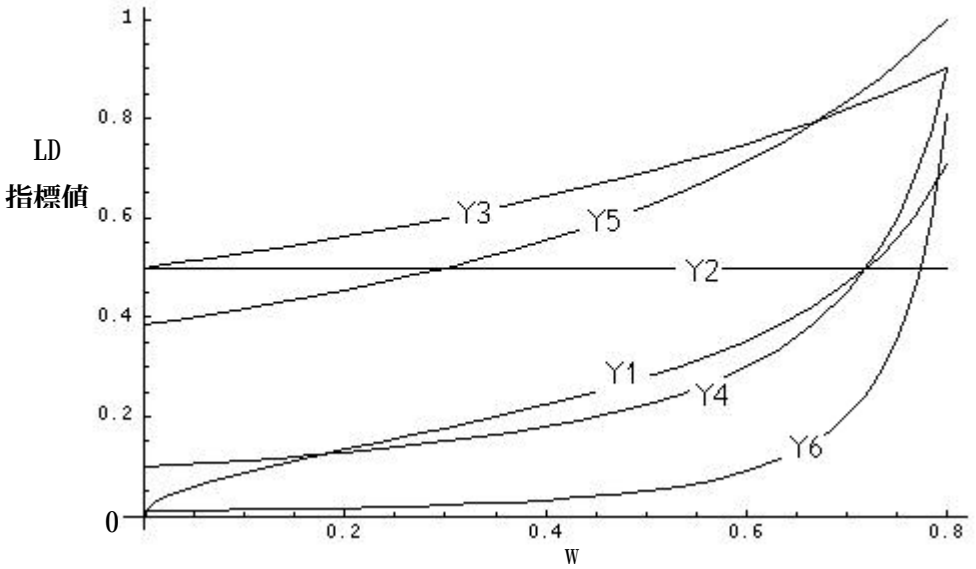
Y1＝Δ
Y2＝D’
Y3＝δ
Y4＝d
Y5＝Q
Y6＝d^2

挙動描図 3

D’＝0.5 で固定している場合のその他の指標の挙動

条件：$P_{aB}=P_{AB}/9$，$P_{Ax}=0.8$，D’=0.5

として、$P_{AB}=w$**を変動範囲[0,0.8]として LD 指標を** w **の関数としてグラフ化すると以下のようになる。**



Y1＝Δ
Y2＝D’
Y3＝δ
Y4＝d
Y5＝Q
Y6＝d^2

3-3-2-3-6-2 参考　連鎖不平衡係数D'の信頼区間

Reference : Zapata C, Alvarez G and Carollo C, Approximate Variance of the Standardized Measure of Gametic Disequilibrium D'. Am J Hum Genet 61:774-778, 1997 (Letters to the Editor)

D'の信頼区間を求めるのに、その分散を近似式により算出する方法をとる。
分散V(D')は複雑な挙動をとるが、簡単にまとめると、
D'の値が小さいほど、分散は大きくなる傾向がある。
アレル頻度が0.5から遠いほど、D'の変化につれ、大きく変化する。

詳しいことはReferenceに譲るが、

Haplotype を AB Ab aB ab とし、それぞれの頻度をP_{AB} P_{Ab} P_{aB} P_{ab}とする

簡便化のために、3-3-2-3-6-1と同様に
$\mathrm{Min}\,(P_{xx)}) = P_{aB}$
とし、
$N = P_{AB} \times P_{ab} - P_{Ab} \times P_{aB}$
とする。

さらに、簡便のため$N \geq 0$であるように設定する。
今、$D_{MAX} = \mathrm{Min}\,((P_{AB} + P_{aB}) \times (P_{aB} + P_{ab}), (P_{AB} + P_{Ab}) \times (P_{Ab} + P_{ab}))$
と置くと、
$D' = N / D_{MAX}$

$D_{MAX} = (P_{AB} + P_{aB}) \times (P_{aB} + P_{ab})$　の場合には、
$x_i = P_{Ab}$
$D_{MAX} = (P_{AB} + P_{Ab}) \times (P_{Ab} + P_{ab})$　の場合には、
$x_i = P_{aB}$
とすると、D'の分散は次のような式で近似されることがわかっている（但し、観測染色体数が多い場合である）

$$V(\hat{D}') \approx \left[\frac{1}{n(D_{MAX})^2}\right]\left\{(1-|D'|)\left[nV(D) - |D'|D_{MAX}(P_{AB}+P_{Ab})\times(P_{AB}+P_{aB}) + (P_{aB}+P_{ab})\times(P_{Ab}+P_{ab}) - 2|D|\right)\right] + |D'|x_i(1-x_i)\right\}$$

標準偏差　$s = \sqrt{V(\hat{D}')}$

95%CIはさらに、正規分布近似により

計算で得たD'±1.96 x sで近似する

3-3-2-3-7 参考 組み換えと組み換え率

＜用語の説明＞

組み換えとは：

減数分裂の際に発生する相同染色体間の DNA の交換現象。

組み換え率とは：

ある 2 ローカスについて考える。減数分裂後に生じた染色体がその 2 ローカスのうち、1 ローカスに父由来のアレルを持ち、もう 1 つのローカスに母由来のアレルを持つとき、その染色体はこの 2 ローカスについて組み換え体(recombinant)であるという。この組み換え体の占める割合を組み換え率(θ)と呼ぶ。

遺伝的距離 単位 Morgan(M, モルガン)：

ある距離において、1 回の減数分裂を経るごとに平均 1 回の組み換えが起こるとき、その距離を 1 Morgan と定義する。この単位によって規定される長さを Genetic map distance(遺伝的距離)と呼ぶ。

実際の減数分裂のとき、ヒトの染色体は 2 倍体なので、1M の距離にあるローカスの間では、2 回の組み換えが起きている(これは単純化すると、父由来の染色体上で 2 回の組み換えが起きている場合と母由来の染色体上で同様のことが起きている場合と、父由来染色体上で 1 回、母由来染色体で 1 回起きている場合が含まれる)。ちなみに、ヒト 1 番染色体(最も長い)のモルガン数は 2 近くであり、21 番染色体のそれは 0.5 を越える程度である。ヒト男性の減数分裂では、染色体全体において平均 53 キアズマ(chiasmata)が生じるので 26.5M(26.5＝53/2)に相当する。ヒト女性染色体全体のモルガン数は 39 である。遺伝的距離と物理的距離(塩基対数に相当)との対応関係は、組み換えの起こりやすい領域とそうでない領域があるため、厳密に対応しないが、1cM ＝1,000,000 塩基対といわれている。但し 100cM＝1M である。

＜(遺伝的)距離と組み換え率の関係(Map functions)＞

距離と組み換え率とのあいだには、近ければ組み換え率が低く、遠ければ高いという関係があることは容易に想像できる。これは距離が短い範囲では組み換えの回数が少なく、距離が長いと組み換えの回数が多くなるからである。つまり、距離と組み換え率との関係を決めるためにはまず、距離と組み換えの回数の関係を明らかにし、ついで組み換えの回数と組み換え率との関係を明らかにすることで距離と組み換え率との関係がわかる。ここで、自明のこととして次のことが言える。距離が０であれば組み換え率は０であり、距離が十分に遠ければ組み換え率は 1/2 に収束することも想像に難くない。また、組み換え回数と組み換え率との間の関係は次のようになっている。

ある距離において 2 倍体で組み換えが 1 回も起きていなければ、組み換え率は０である。ある 2 倍体において、ある距離で組み換えが奇数回生じたとすると、2 倍体のうちの、片方の 1 倍体の、両端のマーカーに注目すると、そこでは組み換えが起きており、もう片方の、1 倍体の両端のマーカーでは組み換えは起きていない。これは組み換え率が 1/2 であることと同義である。一方、ある 2 倍体において、ある距離で組み換えが偶数回生じたとすると、2 倍体のうちの、両方の 1 倍体で奇数回ずつ組み換えが起きている場合と、両方で偶数回ずつ組み換えが起きている場合の 2 通りが同一確率で想定される。

両端のマーカーのみに着目すると、前者の組み換え率は１、後者のそれは０である。これは組み換え率がやはり 1/2 であることと同義である(1/2×1+1/2×0＝1/2)。したがって組み換え回数が０回の時を除いて、組み換えの起きる回数によらず組み換え率は 1/2 である。今、組み換えが１度も起きない確率を P_0とすると、組み換えが１回以上起きている確率は $1-P_0$である。奇数回と偶数回の両方を考えあわせて、組み換え体が生じる確率は 1/2 となっている。
組み換えの回数を０回から無限大回まで通算すると、

$$0\times P_0+1/2\times(1-P_0)=1/2\times(1-P_0)$$

この関係を表す式を
Mather 's formula と呼ぶ。

$$\theta=1/2\times(1-P_0)$$

他方、距離(m)と組み換え回数の間の関係は自明ではなく、モデルが提唱されている。それを以下に示す。

モデル 1(Morgan map function)

$\theta=1/2\times(1-P_0)=m$,

仮定：組み換えは設定２点間で最大１回しか起こらないとする。(短距離間での近似である。)

$0 \leqq m \leqq 1/2$ において、
組み換えの起こる確率$(1-P_0)$は m に比例して上昇すると仮定し、
$P_0=1-2m$
と置くと、上の Function が得られる。

モデル 2(Haldane map function)

$\theta=(1-Po)/2=(1-e^{(-2m)})/2$,

$m=-1/2\times\ln(1-2\theta)$

仮定：組み換えはあらゆる場所で相互に独立にランダムに起こるとする。
このことは言いかえると、ある２点間で組み換えの起こらない確率は m に関する
Poisson distribution に従うと言える。その関係は、
$P_0=e^{(-2m)}$
と書き表される。

この Function は次のような考え方でも導かれる。
今３ loci、A、B、C がこの順序で並んでいるとする。Loci 間の距離を m(XY)で表すと、

$m(AC)=m(AB)+m(BC)$

一方、Loci 間の組み換え率を θ(XY)で表すと、

$$\theta(AC)=\theta(AB)\times(1-\theta(BC))+(1-\theta(AB))\times\theta(BC)$$
$$=\theta(AB)+\theta(BC)-2\theta(AB)\theta(BC) \quad ——— \quad (\%)$$

ここで式変換をすると、

$1-2\theta(AC)=(1-2\theta(AB))\times(1-2\theta(BC))$

式の両辺の対数をとると、

$\ln(1-2\theta(AC))=\ln(1-2\theta(AB))+\ln(1-2\theta(BC))$ ――― (#)

この式より、

$m(XY)=C\times\ln(1-2\theta(XY))$ ――― (!)

と置いてやると(#)を満足させることがわかる。

今、C=1/2 とすると(!)の導関数(-1)/(1-2θ)のθ値の小さい範囲での値は約１となり、これは近距離間の組み換え頻度はその距離に比例するという実情(Morgan map function)をよく説明するので、

$m=-1/2\ln(1-2\theta)$

$\theta=(1-e^{-2m})/2$

を得ることができ、これは Haldane map function である。

しかしながら、現実の組み換えは近距離では相互に干渉しあって起こりにくく、Poisson distribution には従わないことが知られている。近距離間では(%)の式は

$\theta(AC)=\theta(AB)+\theta(BC)$ (%')

が成り立つ。

モデル 3((1)Kosanbi map function、(2)Cartes-Falconer map function、(3)Felsenstein map function)

仮定：十分に近距離では

$\theta(AC)=\theta(AB)+\theta(BC)$ (%')

十分に遠距離では

$\theta(AC)=\theta(AB)+\theta(BC)-2\theta(AB)\theta(BC)$ (%)

が成り立ち、その中間領域では

$\theta(AC)=\theta(AB)+\theta(BC)-2c\theta(AB)\theta(BC)$ (%")

と表せる。ここで c は距離の関数である。

ここで、c を、距離 0 の時(θ=0)は 0 もしくはそれに近い値、距離無限大の時(θ=1/2)は１である、という条件を満たすθの関数として設定したモデルがいくつか示されている。

(1) Kosambi map function

$c=2\theta$

$\theta=1/2\times((e^{4m}-1)/(e^{4m}+1))$

$m=1/4\times\ln((1+2\theta)/(1-2\theta))$

(2) Carter-Falconer map function

$c=(2\theta)^3$

(3) Felsenstein map function

$c=K-2\theta(K-1)$

モデル 4(Sturt map function)

上記の 3 モデルは組み換えの起こる法則を推測して導いた関係であった。

今、経験的にある規則が知られており、それに基づいた関係式があり、Sturt map function と呼ばれる。

減数分裂時の組み換えには以下のような規則があることが観測されている。

染色体の減数分裂時には動原体を中心として広がる 4 分体(末端動原体(acrocentromea)を保持する染色体(13‐15、21‐22)を除く)では、ある長さ当たり必ず 1 回の組み換えが起こり(必然的組み換え)、さらに Poisson 分布に従って無作為に組み換えが起こっている(無作為分組み換え)と考えるのが最も妥当な発生分布であることが観察されている。

今、必ず 1 回の組み換えが起こる単位距離を L とすると、

距離 m の範囲に必然的組み換えが起こる確率は、

m/L

距離 m の範囲に偶然的組み換えが起こる確率は、

$1-e^{-m(2L-1)/L}$

である。

従って、距離 m において必然的組み換えも偶然的組み換えも起こらない確率は

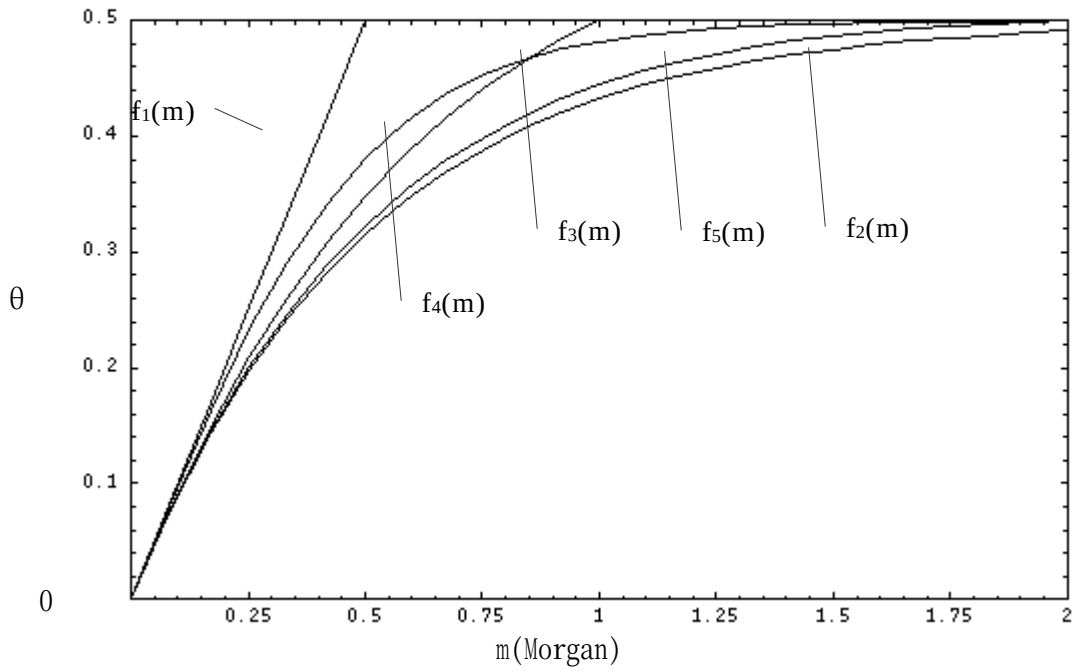
$P_0=(1-m/L)e^{-m(2L-1)/L}$

$\theta=(1-P_0)/2=1/2(1-(1-m/L)e^{-m(2L-1)/L})$

である。

図
縦軸　θ
横軸　m(Morgan)

Morgan map function, Sturt map function(L=0.5)
$\theta = m = f_1(m)$
Haldane map function
$\theta = (1-e^{-2m})/2 = f_2(m)$
Kosambi map function
$\theta = ((e^{4m}-1)/(e^{4m}+1))/2 = f_3(m)$
Sturt map function(L=1)
$\theta = (1-(1-m)e^{-m})/2 = f_4(m)$
Sturt mp function (L=2)
$\theta = (1-(1-m/2)e^{-3m/2})/2 = f_5(m)$



Morgan

$\theta = 2cm$

Haldane

$\theta = (1-e^{-2cm})$

Kosambi

$\theta = (e^{4cm}-1)/(e^{4cm}+1)$

Sturt(L=1)

$\theta = 1-(1-cm)e^{-cm}$

Sturt(L=2)

$\theta = 1-(1-cm/2)e^{-3cm/2}$

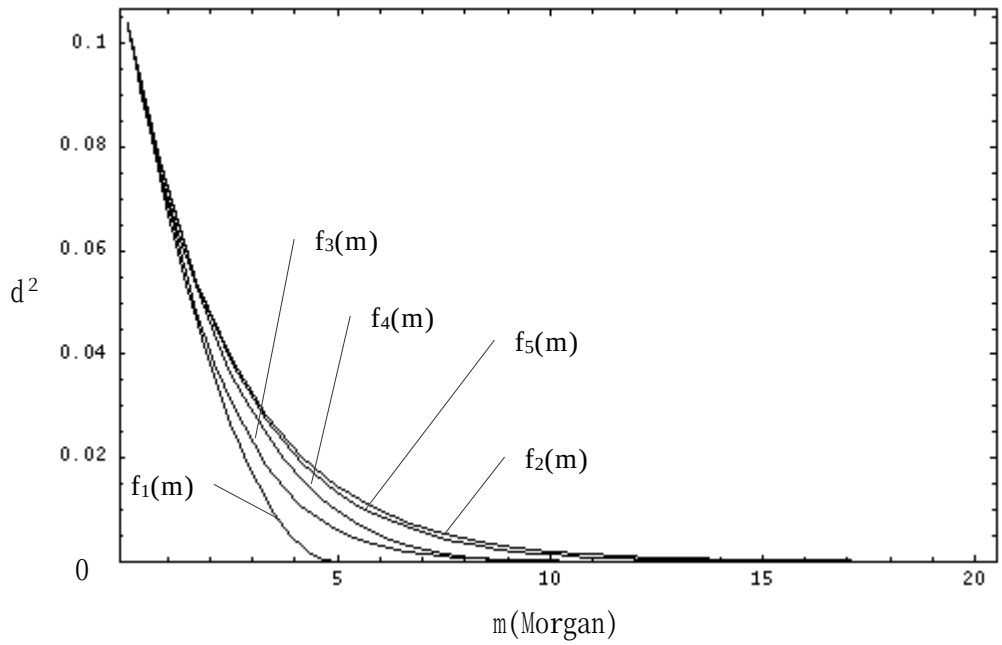
$P_{AB} = (1-\theta)\times P_{0AB} + \theta \times (P_{0AB}+P_{0Ab})\times(P_{0AB}+P_{0aB})$

$P_{Ab} = (1-\theta)\times P_{0Ab} + \theta \times (P_{0AB}+P_{0Ab})\times(P_{0Ab}+P_{0ab})$

$P_{aB} = (1-\theta)\times P_{0aB} + \theta \times (P_{0aB}+P_{0ab})\times(P_{0AB}+P_{0aB})$

$P_{ab} = (1-\theta)\times P_{0ab} + \theta \times (P_{0aB}+P_{0ab})\times(P_{0Ab}+P_{0ab})$

$d^2 = (P_{AB}\times P_{ab} - P_{Ab}\times P_{aB})^2 / ((P_{AB}+P_{aB})^2 \times (P_{Ab}+P_{ab})^2)$



3-3-2-3-8 最もよく使われる連鎖不平衡係数D'とΔ²の特徴と関係について

Δ²はR²とも表記されることがある

連鎖平衡･連鎖不平衡の最も特徴的な状態は以下のように３つである。

１　連鎖平衡にて　０

２つのbiallelic markersが作る「完全連鎖」には大きく２種類ある。

存在し得る４つのhaplotypesのうち、

２　２つのhaplotypesしか存在しない場合と、

３　３つのhaplotypesが存在する場合

前者をabsolute disequilibrium、後者をcomplete disequilibriumと分けて表現することもある

	Haplotype AB	Haplotype Ab	Haplotype aB	Haplotype ab
連鎖平衡	P(A)xP(B)	P(A)x(1-P(B))	(1-P(A))xP(B)	(1-P(A)x(1-P(B))
Absolute disequilibrium	P(A)	0	0	1-P(A)
Complete disequilibrium	P(A)	0	P(B)-P(A)	1-P(B)

但しP(A)はSNP Aのアレル頻度、P(B)はSNP Bのアレル頻度、AB、Ab、aB、abはそれぞれhaplotypeを表し、P(A), P(B)を用いた数式で表されたセル内の式は、連鎖の状態別のハプロタイプ頻度を表す。

この特徴的な３状態におけるD'とΔ²の値は

	D'	Δ²
連鎖平衡	0	0
Absolute disequilibrium	1	1
Complete disequilibrium	1	0より大、1未満

このことからわかるように、Absolute disequilibriumとComplete disequilibriumとの連鎖不平衡係数を区別したいときにはΔ²が適切であり、どちらも「完全連鎖」しているとして一まとめにして扱いたい場合にはD'が適当である。

一般的にΔ²の減衰はD'のそれよりも急峻であることが知られている。参考となる文献は
Linkage Disequilibrium in Humans: Models and Data. JK Prichard and M Przeworski, Am J Hum Genet. 69:1-14,2001

その他の事情として次のような関係が知られている。
多型の発生と消長を決定しているのは、変異の発生率(μ)と組替え率(θ)、及びrandom driftであるとするのが、一般的であるが、そのような立場にたつとき、Δ²はμ及びθと以下のような関係にある。
$E(\Delta^2)=(4N(\mu+\theta))^{-1}$
ここで$E(\Delta^2)$はΔ^2の期待値である。

3-3-2-4　集団間の民族学的遺伝的差(階層化を含む、その問題点と存在の有無の検定)

3-3-2-4-1 集団間の民族学的遺伝的差について

初めに、“集団間の民族学的遺伝的差”について説明する。ある集団とある集団が民族学的遺伝的に等しいという場合、その両集団は十分な期間、互いに交流して無作為な交配が行われ続けてきたことが必要である。このことを裏返せば、どんなに長い期間、距離的に近い関係にあったとしても無作為交配が行われてこなかった集団の間には何らかの民族学的遺伝的差が残ったり、蓄積したりする。このような集団間の民族学的遺伝的差は、アレル頻度・genotype 頻度の異なる多型がゲノムワイドに多数存在することで確認される。もし、比較する集団間の民族学的遺伝的差が小さい場合には、アレル頻度・genotype 頻度が異なる多型の数が少なく、また、アレル頻度・genotype 頻度の差が小さくなる。

ここで、話をケース・コントロール関連解析に戻す。統計的有意差を持つ疾患関連マーカーが検出されるということは、ケース群とコントロール群のマーカーのアレル頻度・genotype 頻度が異なり、その差が検出されるということである。しかしながら、このようなマーカーの genotype 頻度・アレル頻度の差は必ずしも疾患と関連しているとは限らない。このように疾患関連マーカーを検出しようとして、疾患と関連のないマーカーが誤って疾患と関連があるように検出された場合、その検出結果を偽陽性と呼ぶ。

偽陽性の原因は簡単に大きく分けて 2 つある。

偽陽性 A

ケース検体とコントロール検体をそれぞれの母集団から抽出したときに、母集団を代表する検体が選ばれなかったための偽陽性である。これはサンプリングバイアスによる偽陽性と呼ばれ、偶然に支配されているため、回避できない。

偽陽性 B

ケース母集団とコントロール母集団との間に民族学的遺伝的差が存在したために検出されてしまったという場合である。こちらはケース母集団とコントロール母集団の選び方を理想化すれば、回避可能であるし、本来回避するべき状況である。

しかしながら、厳密にケース母集団とコントロール母集団を同一の一般母集団の中に設定し、偽陽性 B を完全に回避することは難しい。その理由も大きく分けて 2 つある。１つは、対象疾患に罹りやすい集団と罹りにくい集団が混在している集団において、ケース・コントロール検体抽出を行う場合である。この場合はサンプル抽出の結果として得られるケース検体に罹患率の高い集団由来の検体が多くなり、逆にコントロール検体に罹患率の低い集団由来の検体が多くなる。このようにある疾患の罹患率について民族学的にムラがあるとき、その疾患についてこの集団は階層化しているという。従って、罹患率の異なる 2(複数)集団が存在していることが知られており、検体抽出時にその 2(複数)集団を区別することができれば、その 2(複数)集団を区別して検体抽出を行うことが必須である。しかしながら 2(複数)集団が区別できない場合には、この混合母集団を用いる限り、階層化の影響を排除することは不可能であり、その結果として生じる偽陽性の検出も避け得ない。

もう１つは、検体抽出の方法が不完全であるために、ケース母集団とコントロール母集団とを厳密に共通な一般母集団内に設定できない場合である。これは、検体抽出の現実的方法に由来する問題である。

ケース検体は通常、ある医療機関において治療を受ける罹患者集団から選ばれる。一方、コントロール検体はそれとは異なる理由で選ばれる(研究のために募られたボランティア、健康診断などのために集まった集団からの抽出、ケース抽出医療機関における他疾患罹患者集団からの抽出など、その方法は様々である)ため、ケース母集団と同一の母集団からのコントロール検体の抽出というのは難しい。RIKEN SRC では、大都会を中心とした複数地域から、罹患者を複数の疾患について個別に収集し、その混合を日本人の一般集団とみなしてコントロールとしている。この設定の背景には日本人は均一であるとの仮定、もしくは都会は多くの地域から流入した人々の混合であるので日本国内の民族学的遺伝的差が消失しているという仮定がある。

このようにケース・コントロール関連解析には常に両群の民族学的遺伝的差が存在するための偽陽性 B の可能性を疑う必要がある。ゲノムワイド関連解析の場合には、回避不可能なサンプリングバイアスによる偽陽性 A も多いことから、陽性結果の解釈は難しい。従ってケース・コントロール集団間の民族学的遺伝的差による偽陽性 B はできるだけ避けたいものである。

実は、複数の集団の間に民族学的遺伝的差があるか否かは多数の遺伝的マーカーについてのタイピング結果を統合して検討することにより、偽陽性 B が混入しているか否かを検定できる。これは、次のような論理に基づく。以下は、単純化のために 2 群間に民族学的遺伝的差があるか否かの解析について述べてある。

ゲノムワイドに多数のマーカーについて集団間有意差検定を行った場合、2 群間に民族学的遺伝的差が無いならば、有意差を認めるマーカーは疾患と関連するマーカーであるか、サンプリングバイアスに由来する偽陽性 A であるはずである。疾患と真に関連する遺伝子の数は限られており、真に関連するマーカーの数は無視できるほど少ない。従って、有意差を認めるマーカーはサンプリングバイアスに由来するもので全てが占められると考えればよい。このような場合、確かに相当数の“偽陽性”が検出されるが、その検出頻度は確率論的に予想された値に近くなるはずである。このことを利用して、多数のマーカーについての群間有意差検定値を集計してその集計値がサンプリングバイアスで説明されるべきか、それ以外の原因(両群間の民族学的遺伝的差)に由来するものと判断されるべきかの検定が可能である。簡単に言うと、有意差のあるマーカーの数が、偶然生じる個数程度であれば民族学的遺伝的差は無いが、有意差のあるマーカーの個数が偶然では説明できないほど多ければ民族学的遺伝的差があると判断される。

具体的には、全てのマーカーに関して得られる χ^2 検定値の総和を、そのマーカーの数の自由度の χ^2 分布に照らして両群間に民族学的遺伝的差が無いという帰無仮説が真である確率を求めることで検定する。RIKEN SRC のデータは定期的にこの検定を行い、民族学的遺伝的差が示唆される検体群を検出する予定である。

3-3-2-4-2 集団の遺伝的構造の解析

背景

遺伝的に均質とは：

ランダムな交配が継続してきた結果、中立的な多型の全てが相互に偏ることなく確率的に分布しているような集団は、遺伝的に均質であるという。ただし、同一の Ancestral haplotype に由来する連鎖不平衡の関係にある多型同士の分布の偏りはこの原則から除外する。

遺伝的に均質であるか否かを知ることの重要性：

Common Diseases 関連遺伝子解析において、対象集団が遺伝的に均質であるかいなかの重要性は、以下のとおりである。

疾患 Phenotype と Genotype との関連を解析するとき、対象集団が遺伝的に均質でないと、疾患 Phenotype と直接関連のない遺伝マーカー・遺伝ローカスが、偽陽性を呈することによって解析を妨害する。

方法

1. 複数の遺伝子多型マーカーについて、比較するべき集団間で分布の差があるかどうかを検定し、その検定結果が同一集団からのランダムサンプリングから得られた結果であるか否かを検定する(複数マーカー個別検定・累計法(筆者の(勝手な命名))。(ref Jonathan K. Prichard and Noah A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 65: 220-228,1999)

2. 個々人の複数の遺伝子多型マーカーの genotype data をもとに、クラスター別の多型マーカー頻度分布及び個々人の所属クラスターをマルコフ連鎖モンテカルロシミュレーションによって推定し、推定所属クラスターの分布が比較するべき集団間で異なるか否かを検定する(複数マーカーデータによる所属クラスター推定法(筆者の(勝手な命名)) 。(ref Jonathan K. Prichard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959, 2000)

3. 複数の遺伝子多型マーカーの genotype data を用いて、異なる染色体上の SNP s が作る複合 genotype の分布が、均一集団からのランダムサンプリングから得られた結果(正規分布)であるか否かを検定する(偽連鎖不平衡係数の正規性検定(筆者の(勝手な命名))。(ref なし)

3-3-2-4-3 マルコフ連鎖モンテカルロ法による民族学的差の解析について

（ref Jonathan K. Prichard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959, 2000）

1. 全体像
   (1) 扱うパラメタは以下のとおり。
      ① 既知パラメタ
         a)個人別・遺伝マーカー別 genotype
      ② 未知(推定)パラメタ
         a)クラスタ数
         b)クラスタ別・遺伝マーカー別頻度
         c)個人別所属クラスタ
   (2) 解決しようとしているのは
      ① 既知パラメタから、未知パラメタの期待値(その他信頼区間など)を求める。
   (3) 上記の問題は以下のような理由により、算術的に得られないことがしばしばである。
      ① 複雑な積分を用いる必要がある。
      ② 高次元の解である。
      ③ そもそも解析解が得られえない分布である。
   (4) したがって、算術的方法以外の解法が必要である。
2. 算術的方法以外の方法として採用される「マルコフ連鎖モンテカルロ法(MCMC)」の概説
   (1) モンテカルロ法とマルコフ連鎖を組み合わせた推定法のことである。集団の遺伝的構造解析のように、多数の未知パラメタの推定をおこなう場合には、MCMC 法を実践するにあたって、さらに工夫が必要で、Metropolis-Hastings(M-H)アルゴリズムを用いて定常分布(推定解)に近づく必要があり、さらに、M-H アルゴリズムの実行時に使用する関数(分布)には、Gibbs サンプラと呼ばれる条件付き分布(関数)を用いるのが通例となっている。この Gibbs サンプラからの乱数発生には Adaptive Rejection sampling という方法を併用する必要がある。
   (2) モンテカルロ法とは
      ① 統計学の実験的手法の１つである。乱数を発生させてシミュレーションを行い、確率的に解を得る方法である。適切な解を得るためには、乱数の発生に条件を設定する必要がある。その条件設定方法の一つがマルコフ連鎖である。
   (3) マルコフ連鎖とは
      ① ある変数を順次、発生させるときに、現世代のパラメタの値(のセット)のみをもとに次世代のパラメタの値を発生させる方法のことである。
   (4) Metropolis-Hastings アルゴリズムとは
      ① マルコフ連鎖を用いて、世代を順次進めていくときに、次世代のパラメタの値を得るためには、次世代のパラメタ値をある分布(関数)(サンプラと呼ぶ)から作り出し、その作られた値を次世代のパラメタの値として妥当かどうかを判断する、という手続きを踏むと未知パラメタの値(のセット)が定常状態(推定値)に向かうことが知られている。このような新世代のパラメタ値の作り方を M-H アルゴリズムと呼ぶ。

Ryo Yamada [1] Hiroto Kawakami [2] Masao Yamaguchi [3]
Eri Tatsu [1] Akihiro Sekine [4] Kazuhiko Yamamoto [1]
Yusuke Nakamura [4] Tatsuhiko Tsunoda [3]

[1] Laboratory for Rheumatic Diseases, SNP Research Center, RIKEN, Tokyo, JAPAN

[2] Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, JAPAN

[3] Laboratory for Medical Informatics, SNP Research Center, RIKEN, Tokyo, JAPAN

[4] Laboratory for Genotyping, SNP Research Center, RIKEN, Tokyo, JAPAN

Correspondence to Ryo Yamada, ryamada-tky@umin.ac.jp

Acknowlegments

We Thank Hisanori Haga [2] for discussing on presentation materials in detail.

Key Words

- Population Structure
- Multiple Genetic Markers
- SNP
- Spurious Association
- Genomewide Association Study
- Markov Chain Monte Carlo Simulation
- Japanese

Abstract

Linkage disequilibrium (LD) mapping, using a dense map of single nucleotide polymorphisms (SNPs), has been advocated as the method of choice to locate loci of complex genetic diseases. However it has also been suggested that genetic population structure could lead to many spurious associations between genetic markers and a disease phenotype when LD mapping was performed, especially in case of studies using samples from ethnically structured populations.

Japanese population has been considered to be less structured, and, therefore, more appropriate for LD mapping than many other populations in the world, because of its ethnic isolation from both historical and geographical standpoints. However no analysis on their genetic homogeneity based on assayed genotype data has been reported. We evaluated homogeneity of two sub-populations, each of which were from one of the two largest Japanese metropolises by analyzing genotype data of 188 individuals for 303 unlinked SNPs on autosomal chromosomes. We adopted two analytical methods to evaluate structure. One method was to assess existence of difference in population structure between two sub-populations, by analyzing chi square statistics computed for multiple contingency tables constructed for allele distribution of each SNP between sampling sub-populations. The other was to infer population structure by model-based clustering method using Markov Chain Monte Carlo algorithm. We found that, if any, only subtle difference in population structure between two sub-populations existed. The possible allele difference between two sub-populations seemed small enough not to severely interfere case-control association studies for complex genetic traits.

Organization of This Presentation

- Six sets of simulated genotype data of two sub-populations with various degree of difference in allele frequencies of SNPs were analyzed with two methods for genetic structure.
- Real genotype data of two sub-populations in Japan were analyzed in the same way.
- The results of real data were compared with the results of simulated data.

Methods

Construction of genotype data for the structural analysis of simulated data

- Suppositions:
  - Two sub-populations were supposed, each of which had randomly mated and were homogeneous.
  - 180 SNPs were supposed, all of which were:
    - Biallelic.
    - In Hardy-Weinberg equilibrium.
    - Unlinked each other.
  - Genotypes of individuals, consisting of genotype data sets, were assigned as if individuals were randomly sampled from each sub-population.
  - Fifty individuals were sampled for each sub-population.
- Six data sets were constructed with variations in allele frequencies of a part of SNPs between two sub-populations.

Methods

Allele frequencies of SNPs in two sub-populations for 6 sets of simulated data (1)

Table 1 Number of SNPs and their allele frequencies in 2 sub-populations.

	No. SNPs	Sub-population 1	Sub-population 2	Difference of allele frequencies between sub-populations	Average allele frequency
	20	0.2	0.2	0	0.2
	20	0.3	0.3	0	0.3
	20	0.4	0.4	0	0.4
	10	0.2－Δa×0.5*	0.2＋Δa×0.5	-Δa	0.2
	10	0.2＋Δa×0.5	0.2－Δa×0.5	Δa	0.2
	10	0.3－Δa×0.5	0.3＋Δa×0.5	-Δa	0.3
	10	0.3＋Δa×0.5	0.3－Δa×0.5	Δa	0.3
	10	0.4－Δa×0.5	0.4＋Δa×0.5	-Δa	0.4
	10	0.4＋Δa×0.5	0.4－Δa×0.5	Δa	0.4
	10	0.2－Δb×0.5&	0.2＋Δb×0.5	-Δb	0.2
	10	0.2＋Δb×0.5	0.2－Δb×0.5	Δb	0.2
	10	0.3－Δb×0.5	0.3＋Δb×0.5	-Δb	0.3
	10	0.3＋Δb×0.5	0.3－Δb×0.5	Δb	0.3
	10	0.4－Δb×0.5	0.4＋Δb×0.5	-Δb	0.4
	10	0.4＋Δb×0.5	0.4－Δb×0.5	Δb	0.4
total	180	-	-		

Δa* and Δb & are parameters to make difference of allele frequency between sub-populations 1 and 2.
Δa and Δb are specified for each set of simulation on Table2

Allele frequencies of SNPs in two sub-populations for 6 sets of simulated data (2)

Table 2 Parameters to give allele frequency difference between sub-populations 1 and 2

Name of simulation set	Δa	Δb
1	0.025	0.025
2	0.025	0.05
3	0.05	0.1
4	0.1	0.1
5	0.05	0.15
6	0.1	0.2

Methods

Analytical Methods of Structure (1)

- Analytical Method 1

  Evaluation of sum of multiple chi statistics calculated for individual SNP[1]

- Analytical Method 2

  Inference of structure by Markov Chain Monte Carlo simulation method[2]

  – [1] Jonathan K. Prichard and Noah A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 65: 220-228,1999

  – [2] Jonathan K. Prichard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959, 2000

Methods

Analytical Method 1

- Principles of sum of multiple chi statistics calculated for individual SNP
  a. Null hypothesis and alternative hypothesis
     - Null Hypothesis: All the samples are from an identical population.
     - Alternative hypothesis: Samples are from two distinct sub-populations.
  b. Chi square value for each SNP genotype data
     - χ_i^2 (i=1,2,…N; N:Number of SNPs analysed) represents chi square value calculated for 2×2 contingency table of the observed number of alleles from two sub-populations.
  c. Sum of multiple chi square value(S) and degree of freedom(df)
     - $S = \sum_{i=1}^{N} \chi_i^2$
     - $df = N$
     - S is statistically evaluated as chi square value with degree of freedom being df.

Methods

Analytical Method 2

- Inference of structure by Markov Chain Monte Carlo (MCMC) simulation method*
  a. Parameter matrices were set as below:
     i. X: Genotypes of the sampled individuals
     ii. Z: Belonging sub-populations of the individuals
        – Number of sub-populations : Two
     iii. P: Allele frequencies of SNPs in all sub-populations
  b. Belonging sub-population was assigned to each individual at random at the beginning.
  c. MCMC algorithm was applied and converged result was obtained.
     i. Sample $P^{(m)}$ from $Pr(P|X,Z^{(m-1)})$#.
        • Beta distribution was used for allele frequency distribution of biallelic markers.
        • Gibbs sampler with adaptive rejection sampling was adopted.
     ii. Sample $Z^{(m)}$ from $Pr(Z|X,P^{(m)})$.
  d. Results of multiple runs were summed, considering phenomenon to converge into symmetrical modes.

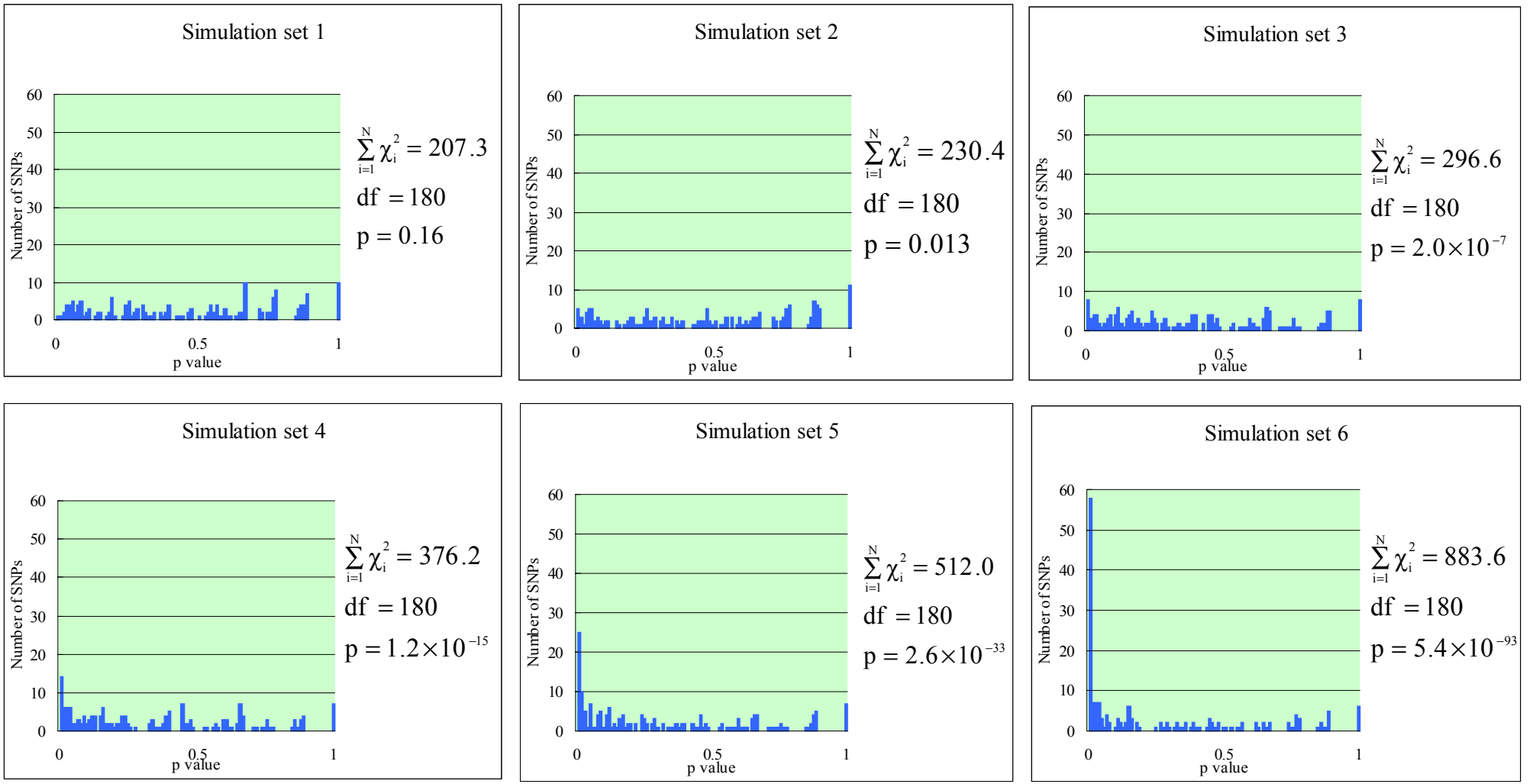
* Program source will be available in near future on request by the author.

m denotes number of iterations and Pr(Y|W) denotes conditional probability distribution of Y when W.

Results

Results of analysis 1(Sum of Chi Square Statistics) of simulated data

Fig1 p value distribution of individual SNPs and sum of chi square values and their corresponding p value



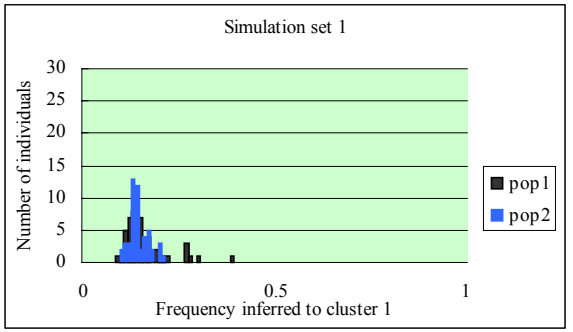
6 histograms represents result of 6 simulated data,respectively.

Null hypothesis in set 1 was not rejected. For sets 3, 4, 5 and 6 null hypothesis was strongly rejected.

Results

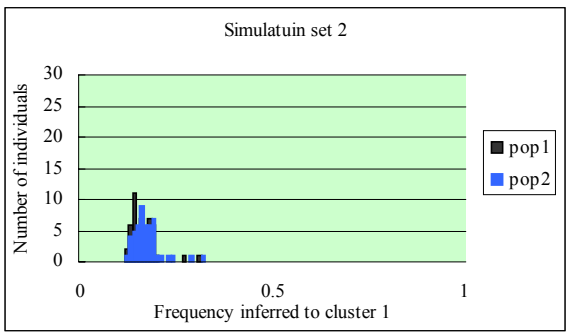
Results of analysis 2 (MCMC) of simulated data

Fig2 Histograms showing how often each individual was inferred to cluster 1



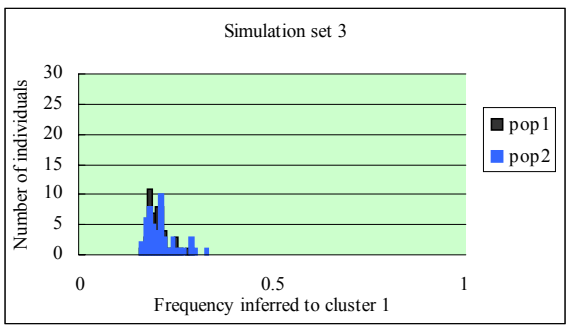
pop1
average 0.15
variance 0.011

pop2
average 0.14
variance 0.001



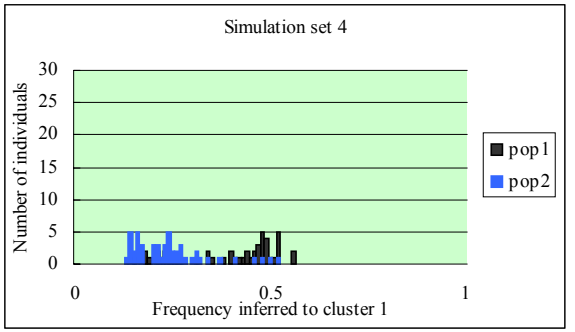
pop1
average 0.16
variance 0.001

pop2
average 0.17
variance 0.001



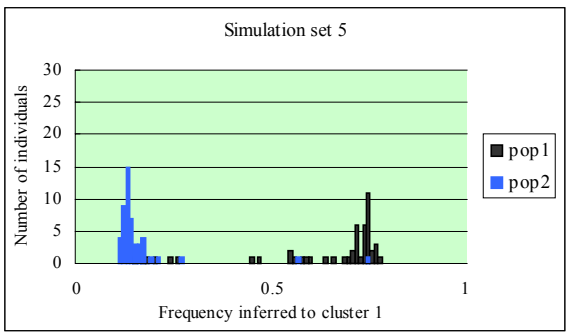
pop1
average 0.20
variance 0.001

pop2
average 0.20
variance 0.002



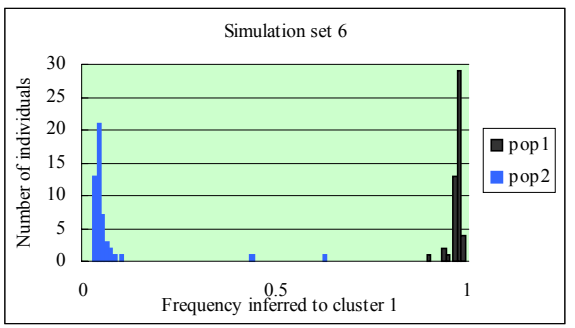
pop1
average 0.39
variance 0.014

pop2
average 0.24
variance 0.009



pop1
average 0.63
variance 0.033

pop2
average 0.16
variance 0.012



pop1
average 0.97
variance 0.000

pop2
average 0.06
variance 0.010

Distributions of sub-populations 1 and 2 of simulation sets 1, 2 and 3 appeared to be similar. Simulation sets 4, 5 and 6 segregated two sub-populations with appropriate progression of clarity of discrimination.

Methods

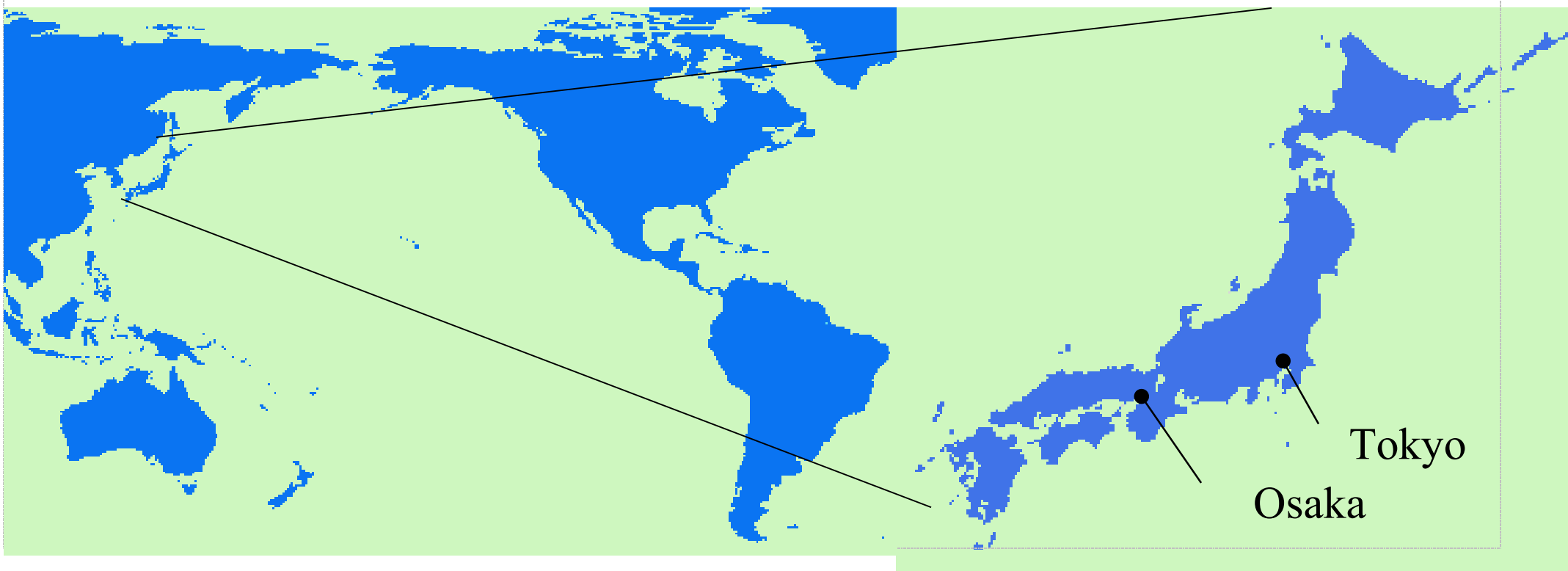
Samples(1)

Construction of genotype data for the structural analysis of real data

Tokyo: Sampling Metropolis 1

Osaka: Sampling Metropolis 2

Fig3 Location of Tokyo and Osaka



Methods

Samples(2)

Construction of genotype data for the structural analysis of real data

- Historical, geographical and ethnical information of sampling sub-populations

  The land of Japan consists of 4 major islands and many small islands, locating at the rim of northwestern Pacific ocean.

  Japanese ancestry is considered to be formed by migration of at least a few ethnic groups from the Eurasian continent before the end of the last glacial epoch, admixed by the Pacific islanders.

  No major migration or admixture happened for the last few thousands years, forming a strongly culturally distinct population with a limited degree of ethnic diversity.

  Tokyo and Osaka metropolitan areas are characterized by high-degree of influxes of people from diverse areas of Japan, followed by random mating, which is believed to have canceled genetic structure present in the past.

  Major differences between Tokyo and Osaka areas:

  - Population of Tokyo area has been formed by influx of people from almost all the areas of Japan for the last 4 hundreds years without a core ancestral population.
  - Population of Osaka area has been formed by admixture of many people from many areas of Japan, especially of south-western area, to the core ancestral population, which had been formed before early middle ages.

Methods

SNPs(1)

Construction of genotype data for the structural analysis of real data

- Source of SNPs:
  - SNPs on autosomal chromosomes from IMS-JST SNPs database, discovered by direct-sequencing method of multiple Japanese individuals'genomic DNAs targeting mainly on and around known and expected genes throughout the human genome. (http://snp.ims.u-tokyo.ac.jp/)
- Number of SNPs:
  - 4373 SNPs were genotyped initially.
  - 303 of 4373 SNPs were adopted for structure analyses by criteria described on the following sheet of paper.

Methods

SNPs(2)

Construction of genotype data for the structural analysis of real data

- SNPs were selected for analyses of population structure by the criteria as below:
  - Each SNP should:
    - Be assayed with appropriate call for more than 170 individuals.
    - Be biallelic.
    - Have minor allele frequency more than 0.1.
    - Be unlinked each other:
      - » Two SNPs on an identical chromosome were considered to be unlinked each other when linkage disequilibrium (LD) index D'* were less than 0.3

        *D': Index of LD was calculated as below:

        1. Haplotype frequencies were estimated by EM-algorithm
        2. $D' = |(P_{AB}^{**} \times P_{ab} - P_{ab} \times P_{aB})| / \text{Minimum}((P_{AB}+P_{aB}) \times (P_{aB}+P_{ab}), ((P_{AB}+P_{Ab}) \times (P_{Ab}+P_{ab}))$,

           AB, Ab, aB, and ab represent 4 haplotypes created by 2 SNPs.

           ** P_{cd} represents frequency of haplotype.
    - Be successfully assayed. Our internal criteria # was adopted for judgment of success of assays:

      # It based on the data in SRC, RIKEN, TOKYO, JAPAN (unpublished):

      - » Intensity of fluorescence signals should be more than a threshold.
      - » More than 90 % of samples should be called.
      - » Chi square statistic for Hardy-Weinberg Equilibrium should give p-value of larger than 1.0×10^{-3}.

Methods

Genotyping Assay

Construction of genotype data for the structural analysis of real data

- Invader assay with multiplex polymerase chain reaction (PCR), characterized as below:
  - Multiplex PCR was performed against 100 genomic regions simultaneously in a single reaction tube.
  - Each amplified fragment contained at least one SNP.
  - A specialized 384-well card system was adopted for the ultra high-throughput genotyping system.
  - Very small amount of genomic DNA was required for genotyping:
    - 40 ng of genomic DNA was used as a template for each multiplex PCR, meaning only 0.4 ng of genomic DNA per genotyping of one SNP.

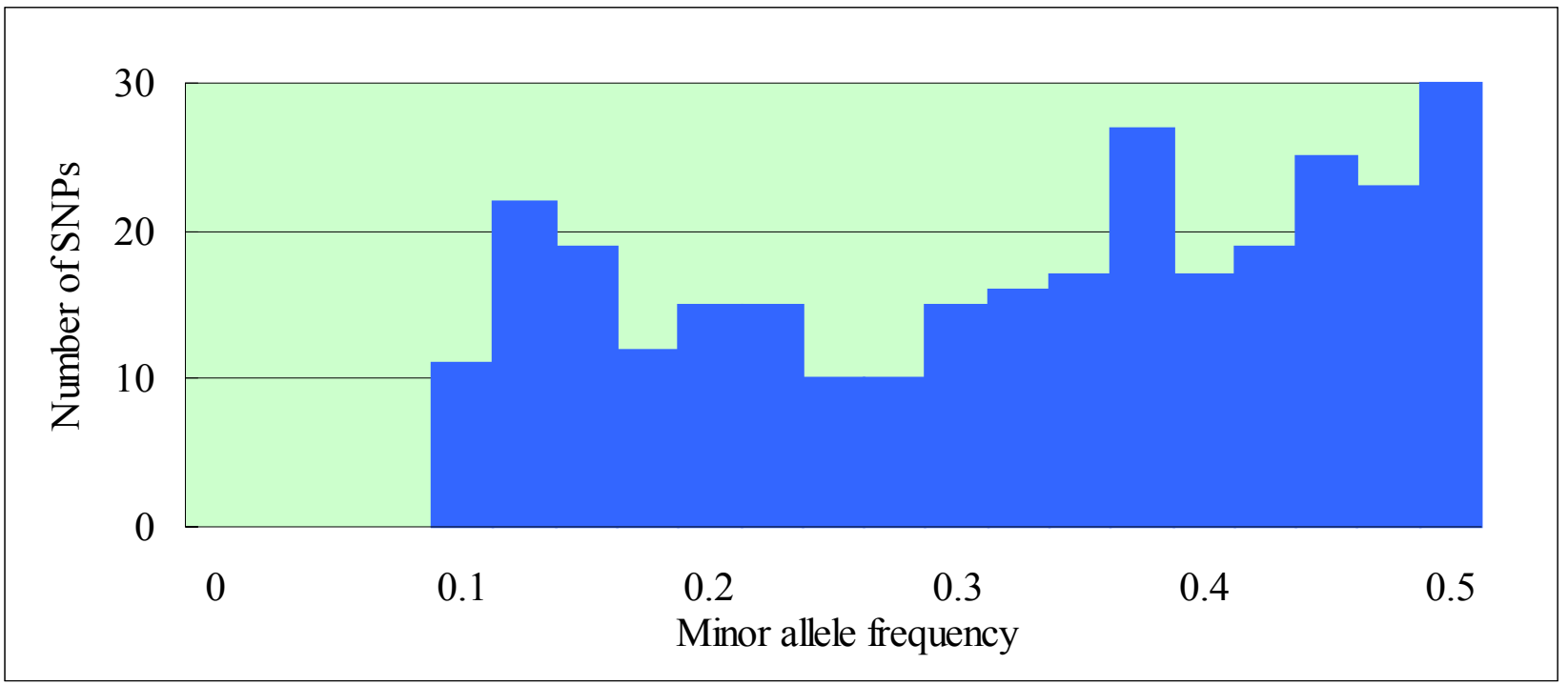
A high-throughput SNP typing system for genome-wide association studies.
J Hum Genet. 2001;46(8):471-7.
Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y.

Results

Characterization of selected SNPs of real data

- Characterization of SNPs included for analyses of population structure
  - 303 SNPs out of 4373 were selected.
  - Average rate of successful call of genotyping of 303 SNPs was 0.99
  - Minor allele frequency of each SNP was distributed as below:

Fig4 Distribution of minor allele frequency of 303 SNPs

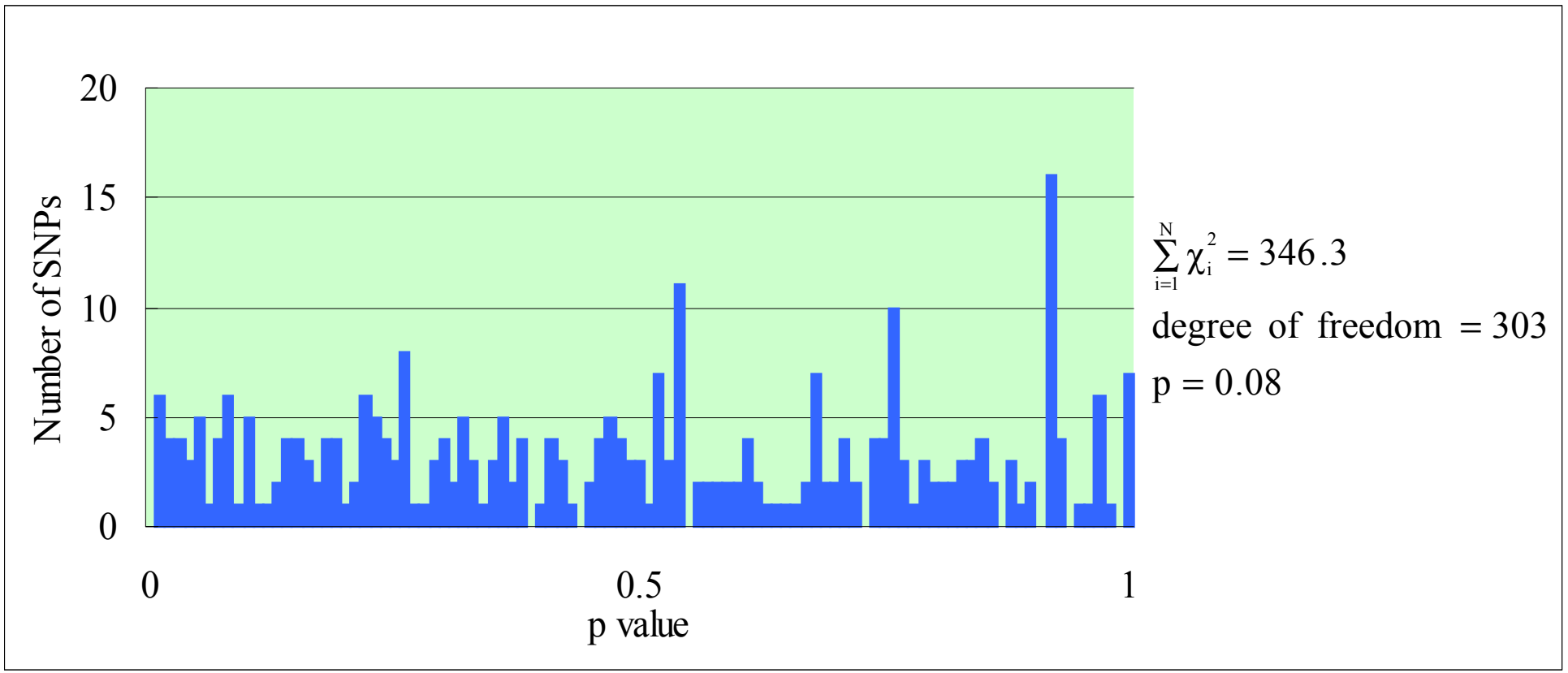


SNPs with higher minor allele frequency seemed to be included more frequently than the distribution in reality.

Results

Result of analysis 1(Sum of Chi Square Statistics) of real data

Fig5 p value distribution of 303 chi square tests and sum of chi square values and its corresponding p value with degree of freedom of 303

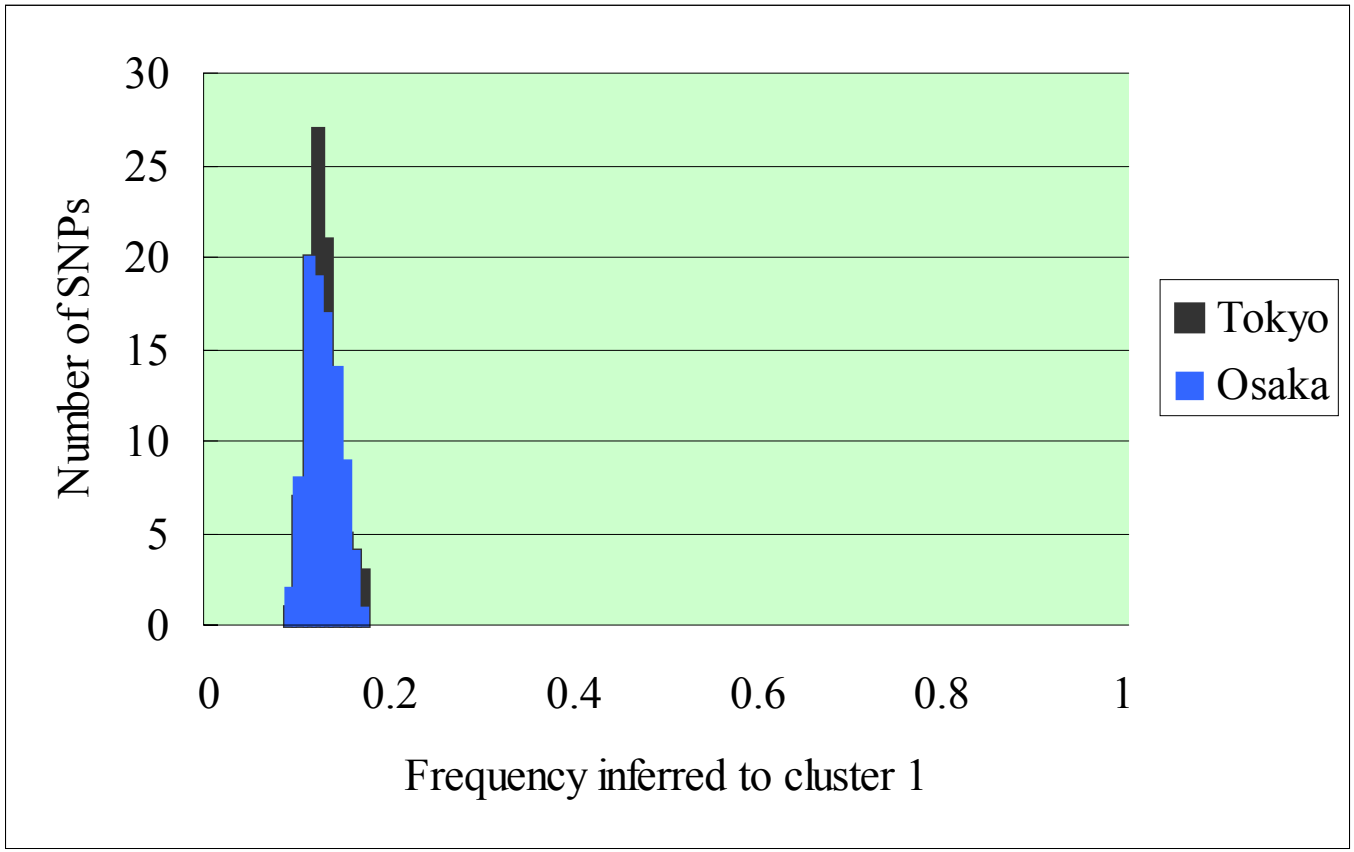


Null hypothesis that Tokyo and Osaka sub-populations were identical was not rejected with cut off p-value of 0.05.

Results

Result of analysis 2(MCMC) of real data

Fig6 Histogram of frequency that each individual was inferred to cluster 1



Tokyo		
	avarage	0.12
	variance	0.0003
Osaka		
	avarage	0.12
	variance	0.0003

Almost identical distributions were obtained for 2 sampling areas.

These distributions were similar to one of simulation sets 1, 2 and 3.

It suggested the difference between Tokyo and Osaka was as small as difference between two sub-populations in simulation sets 1, 2 or 3.

Discussion

The two analytical methods were applied to simulated genotype data and real data, both of which represented subtly heterogeneous two sub-populations. Result of analyses of simulated data revealed both analytical methods could dissect two populations when SNPs with allele frequency difference between populations more than 0.1 existed. Analysis 1 (sum-of-chi square values) showed statistical significance, even when SNPs with allele frequency difference of 0.05 dominated, if such SNPs existed frequently, although analysis 2 (MCMC) did not show visible discrimination in the distributions. In a situation, where population-based sampling association studies were carried out, difference of allele frequencies among subtly stratified populations would be negligible, as far as the difference between ethnic sub-populations was smaller than the difference between phenotypic sub-groups. The comparison between results of simulated data and real data gave an idea that the genetic structure between Tokyo and Osaka was somewhere between a simulated data with maximum allele frequency difference of 0.025 and another set with maximum allele frequency difference of 0.05. When complex genetic traits are analyzed with LD mapping, allele frequency difference between phenotypically distinct groups distributes most likely more than 0.05. From this standpoint, samples from Japanese metropolises should be treated as homogeneous.

(5) Gibbs サンプラとは

① M-H アルゴリズムを実行する(現世代パラメタ値(のセット)から次世代のパラメタ値(のセット)を得る)ときに用いるサンプラ(分布(関数))のタイプの一つである。このサンプラの特徴は、

「複数のパラメタがあり、その個々のパラメタを推定するときに、自身を除くその他の既知・未知パラメタの現時点での値によって推定することができる、言い換えれば、その推定作業には現時点の自身の値は影響を与えない」

というものである。このような分布のことを「フル条件付き分布」という。このような条件は、個々のパラメタの新世代の値を１つ１つ作ればよいのでアルゴリズムとして簡単である。しかしながら、アルゴリズムとして簡単であるがゆえに、乱数の発生が困難であることが多く、次の Adaptive Rejection sampling を導入することが必要である。

(6) Adaptive Rejection sampling とは

① フル条件付き分布から直接乱数を発生させることは容易ではないので(複雑な積分計算を伴うことが多いため)、乱数を発生させることが容易な代理の分布を作成し、その 2 分布の関係を用いて乱数を発生させるという方法である。

3. SNP を用いた MCMC 法による遺伝的構造解析の実際

(1) パラメタ

① 既知パラメタ

a) 個人別・遺伝マーカー別 genotype X

- Xは要素数＝(人数)×(SNP 数)の行列である

② 未知パラメタ

a) クラスタ数（固定し、クラスタ数別にシミュレーションする）

b) クラスタ別・遺伝マーカー別頻度 P

- Pは要素数＝(クラスタ数)×(ＳＮＰ数)の行列である

c) 個人別所属クラスタ Z

- Zは要素数＝(人数)の１次元行列である

(2) 推定の流れ

① Zの初期値設定(以下の２方法で行える) Z_0

② 以下、マルコフ連鎖による推定の繰り返し(m=1,2,3,…は繰り返し回数)

a) X及びZ_{m-1}より、P_mを推定する

b) X及びP_mよりZ_mを推定する

③ Zの初期設定の影響がなくなった後の推定P_m、Z_mの分布が得られる。これがある観測データXが与えられたときのP及びZの推定値の分布である

(3) 推定の各段階の詳細

① Zの初期値設定(以下の２方法で行える) Z_0

a) ランダムに設定(指定クラスタに均等な確率で割り当て)

b) サンプリング地域などにより個人別に特定

② マルコフ連鎖

a) X及びZ_{m-1}より、P_mを推定する

- クラスタ c・SNP s のアレル頻度が t(0≦t≦1)の確率分布 Pr(t)は

  $Pr(t)=t^{n1}x(1-t)^{n2}/(Constant)$

  で与えられる。但し Constant = $\int_0^1 Pr(t)dt$ を満たす。

  これはβ分布であり、Dirichlet 分布の要素が２の場合に補正係数を(1,1)とした場合(Uniform distribution を仮定している)である。つまり、Prichard らのプログラム “structure”と同様の設定となっている。
- クラスタ別・SNP 別にアレルの本数を数える。
- アレル１の本数を n1、アレル２の本数を n2 とすると

  n1=n2=0 の場合と

  n1=1 または n2=1 の場合は

  分布 Pr(t)から直接乱数が発生できるので、その値を p の次世代の値として採用する

  それ以外の場合は

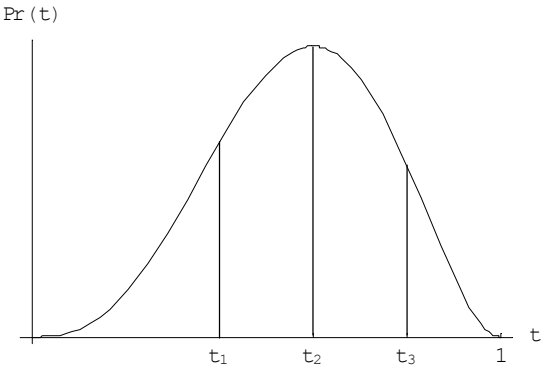
  Adaptive Rejection sampling を行って p の次世代の値を得る

  このとき必要な

  Π(t)≧Pr(t)を常に満たすような分布Π(t)は以下のようにして作成する。

  <Π(t)の作成>

  Pr(t)は下図のように描かれて、それは１次導関数、２次導関数の正負によって４区分に分けられる。



  区分１：0－t_1は下に凸の増加関数

  区分２：t_1－t_2は上に凸の増加関数

  区分３：t_2－t_3は上に凸の減少関数

  区分４：t_3－t_4は下に凸の減少関数

  例図は n1=3，n2=2 の場合である。

  但し、

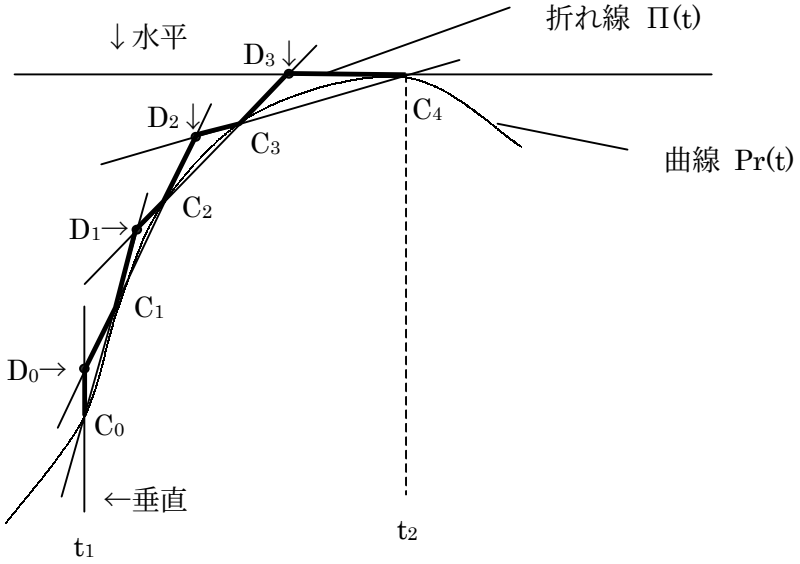
$$t_1 = \frac{n_1}{n_1 + n_2} - \frac{\sqrt{n_1 \times n_2 \times (n_1 + n_2 - 1)}}{(n_1 + n_2) \times (n_1 + n_2 - 1)}$$

$$t_2 = \frac{n_1}{n_1 + n_2}$$

$$t_3 = \frac{n_1}{n_1 + n_2} + \frac{\sqrt{n_1 \times n_2 \times (n_1 + n_2 - 1)}}{(n_1 + n_2) \times (n_1 + n_2 - 1)}$$

区分１と区分４とは下に凸なので、任意の点を結んだ直線がΠ(t)として採用できる。

一方、区分２と区分３とは、上に凸なので、Pr(t)上の 2 点(C_0,C_1…)を結ぶ直線は必ず Pr(t)そのものよりも小さい値をとる。したがって下図で→で指し示したように、Pr(t)上の隣り合う 2 点同士を結ぶ直線の交点として得られる点(D_0,D_1…) をΠ(t)の点として採用し、これと、Pr(t)上の点(C_0,C_1…)とを交互に結んだ折れ線をΠ(t)とする。



このようにして作成されたΠ(t)を用いて、乱数を発生させる。

Π(t)は有限個の点を結んだ折れ線なので、その積分は容易であり、したがって、このΠ(t)に比例する確率密度関数からは、乱数を発生させることは容易である。

今、Π(t)から発生させた乱数を r とする。

これとは別に[0-1]の一様乱数 q を発生させ、

q<Π(r)/Pr(r)

を満たした場合のみ r を採用すると

r は Pr(t)に比例する確率密度関数から発生させた乱数に一致する.

この採用された乱数が p の次世代の値である。

もし、

$q<\Pi(r)/Pr(r)$

が満たされなまった場合には、$\Pi(t)$の再作成をして、乱数を発生しなおす。そのときに、$Pr(t)$上の点$(t,Pr(t))$を新たに加えて、より細かい折れ線$\Pi(t)$を作成する。この新しい$\Pi(t)$は必ず、それよりも前の段階の$\Pi(t)$よりも$Pr(t)$に近く、乱数が採用される確率は高くなる。

b) X及びP_mよりZ_mを推定する

- 個人 i の genotype data から 1,…nc のそれぞれのクラスタの事後確率を求め、その比率に応じて、i の出身クラスタを推定する。
- 前の段階において、クラスタ別・SNP 別のアレル頻度は得られているからそれを $p_{(c,s)}$と表すと、個人 i がクラスタ c の出身である尤度 Pr(i from c)は

$$Pi(i \text{ from } c) = \prod_{s}^{\text{全てのSNP}} g_{(c,s)}$$ とする。

ただし、g(c,s)はクラスタ別・SNP 別の genotype 頻度であり、

個人 i の SNP s の genotype がアレル 1 のホモの場合は

$g_{(c,s)} = p_{(c,s)}^2$

genotype がヘテロの場合は

$g_{(c,s)} = 2 \times p_{(c,s)} \times (1 - p_{(c,s)})$

genotype がアレル 2 のホモの場合は

$g(c,s)=(1-p(c,s))^2$

で与えられる。

上記アルゴリズムを実行する上での詳細は、プログラムソース内の説明文を参照のこと。

3-3-3　解析関連に必要な検体数と第１種過誤(偽陽性率)と第２種過誤(偽陰性率)

3-3-3-1 第１種過誤(偽陽性率)と第２種過誤(偽陰性率)

　第１種過誤(偽陽性率)と第２種過誤(偽陰性率)とは、検定において用いられる用語であるので初めに統計学的検定について簡単に説明する。

　関連解析において統計的に検定を行うということは、
　あるデータを基に次のことを推定するということである。

今、ある検体数で、ある事柄が真実であるかどうかを検定しようとしている。帰無仮説と対立仮説を立て、それらのどちらが真実に近いとみなすのが妥当であるかを、確率の考え方に基づいて判定する。
この作業を統計学的検定といい、検定の結果得られる確率の値として、いわゆるｐ値を求め、その値を基に仮説の妥当性を判定する。
　観測データを統計学的に有意とするか否かはこのｐ値がある基準より大きいか小さいかで判断される。
　この基準をαとする。

　ｐ値がαより小さいとき、この観測データから言えることは、

“ある帰無仮説が真である確率 ｐ はαより小さいから、帰無仮説が真でないと考えることは、‘基準αにおいて正しい’と判断してよい。したがって対立仮説が真であると判断することも基準αにおいて正しい。しかし、100%の確率で対立仮説が真であると言えることは決してなく、確率α(偽陽性率)で対立仮説は真でない。”
関連解析の場合に、この文は
“関連がない確率ｐはαより小さいから、関連があると考えることは 有意水準αで正しい。しかし、関連があると判断することはαの確率で誤りである。”
と言い換えられる。

また、以下のことが知られている。
“ある検体数で、ある差を、ある有意水準で見い出そうという検定の場合、その差を見い出し損ねる確率(偽陰性)は統計的に算出される。”
このような差を見い出し損ねる確率をβとする。

ｐ値がαより大きいとき、このことを基に、上のようなデータを解釈すると、

“ｐ値は有意水準αに達していない。しかしながら、本当は差があるのにその差を見い出し損ねただけかもしれない。もし、見い出し損ねた確率を知りたいならば、それは計算可能である。この検定に用いた検体数と設定した有意水準αは定まっているので、見い出し損ねた確率βは、見い出そうとした差に依存して決まる。”

関連解析の場合に、この文は
“用いた検体数と、設定した有意水準はわかっている。今、見い出そうとしているSNPの遺伝的寄与の程度(Genotypic risk ratioもしくはGenotypic relative riskで規定される)を想定してやれば、本当は差があるのに差を見い出し損ねた確率βは計算可能である。”
と言い換えられる。

上の説明文のαを第１種過誤(偽陽性率)と言い、βを第２種過誤(偽陰性率)と言う。
実際のRIKEN SRCのゲノムワイド関連SNPスクリーニングでこの第１種過誤と第２種過誤とがどのような役割を果たしているかを以下に述べる。

＜スクリーニングの段階＞
そもそもスクリーニングの目的は、真の関連SNPとして可能性のあるSNPを絞り込むことであるから、“スクリーニングをする”とは次のように言い換えられる。

“スクリーニングをすることによって、本来は疾患と関連のないSNPが次の段階に進むことは、スクリーニングの本来の目的から言って全く不都合ではない。その逆に、本当は関連があるSNPを取りこぼすことはなるべく避けたいが、100%避けるということは、いかなる基準を設けても不可能である。従って、解析を進めるにあたっては、スクリーニングによってどのようなSNPがどのくらいの割合で取りこぼされているのかを理解しておくことが必要である。このスクリーニングで取りこぼす割合(偽陰性率)は遺伝的寄与の程度によって変わる。”

スクリーニングによる絞込みの基準をスクリーニング有意差検定で得られるp値に対して設定するとすると、遺伝的寄与の強いSNPを取りこぼす確率はより低く、寄与の弱いSNPを取りこぼす確率はより高くなる。
実際に、RIKEN SRCのスクリーニングは、スクリーニングデータに対して有意差検定を行い、有意水準αよりも有意であるSNPを次の段階に進めて検体数を増やしており、αの値として0.01を初期値として設定した。このスクリーニングに用いている検体数は一定なので、その条件下でのαとβと遺伝的寄与の強さとの関係を知れば、スクリーニングによってどの程度の遺伝的寄与を持つSNPがどのくらいの割合で取りこぼされているかがわかる。

＜検体数増加後の関連本検定の段階＞
この段階で求められるのは、第１種過誤(偽陽性率)の低さである。偽陽性率がある基準(有意水準)より低ければ、それだけ真に関連のある確率が高いという結論が引き出せるからである。
この段階で有意水準に達していないデータは、“関連がないとは言えない”と解釈される。これは、もう少し言葉を足せば、“真の関連はあるかも知れないが、ある有意水準からすると、関連があるとは言い切れない”と言うことである。
これまでに数多くの単ローカス関連解析が行われ、多くの“ネガティブ”データが公表されているが、それらの多くのものは、ある水準以上の有意差を見い出せなかったと言う段階にとどまり、関連がないことを結論付けたものではない。つまり、関連があるとも言えないし、関連がないとも言えないという

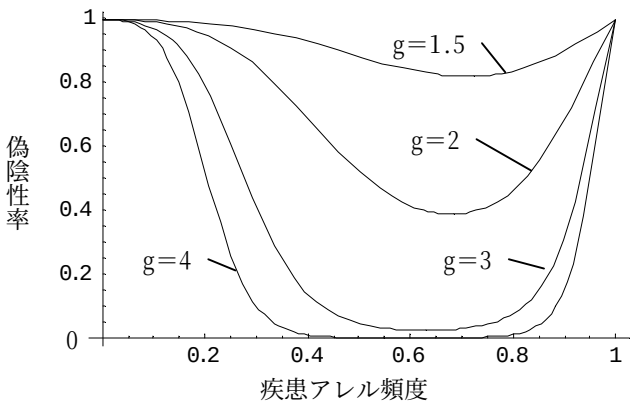
段階で放り出されてしまっているわけである。このようなデータの中途半端な結果になった理由は検体数が少な過ぎたということに尽きる。しかしながら、検体数が少ない“ネガティブ”な検定結果ももう少し使い道があるはずである。つまり、検体数が少なくとも、どのくらいの強さの遺伝的寄与を想定すれば偽陰性率がどのくらいになるかという情報は付加できるので、“この、結果としてパッとしなかったネガティブデータから、○○より強い遺伝的寄与は確率××で否定的である” という結論は引き出せる。

以下に検体数増加後の関連検定によって得られる p 値が表している内容を記す。

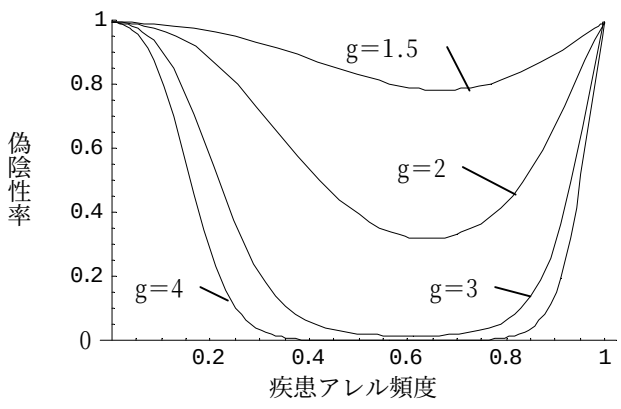
“関連がない確率は p である。”
“ p 値が有意水準αより小さい場合、それだけ関連がない可能性が低いのだから、関連があると判断してもよい。”
“ p 値が有意水準αより大きい場合、その有意水準では関連があるとは言えないが、関連がない、と結論付けられたわけではない。”
“ p 値が有意水準αより大きい場合、本当は関連がある SNP がどのくらいの確率で有意水準に達しそこねて偽陰性になっているかは、その SNP の遺伝的寄与の程度と検体数に依存する。偽陰性率は計算可能である。”

以下のグラフは RIKEN SRC のスクリーニングを模してケース 94 人対コントロール 94×7＝658 人を想定してβを算出している。偽陰性率は検定の際の、データの扱い方にもよるので、アレル頻度の差の検定を行う場合と、劣性･優性それぞれの遺伝形式を想定した上での genotype 頻度の差の検定を行う場合のそれぞれのシミュレーション結果を表現している。ケースとコントロールの人数が等しくないときの補正なども考慮している。以下のグラフの g は genotypic risk ratio を表し、RR は genotypic relative risk を表現している。優性遺伝形式の場合の RR はホモ・ヘテロで等しく g、劣性遺伝形式の RR はホモで g、ヘテロで 1 である。genotypic risk ratio, genotypic relative risk に関しては“2-1 遺伝性であることの確認”を参照のこと。

劣性遺伝形式のﾛｰｶｽを劣性遺伝形式の genotype 分布で検定
（α＝0.01、94VS658 でスクリーニング）

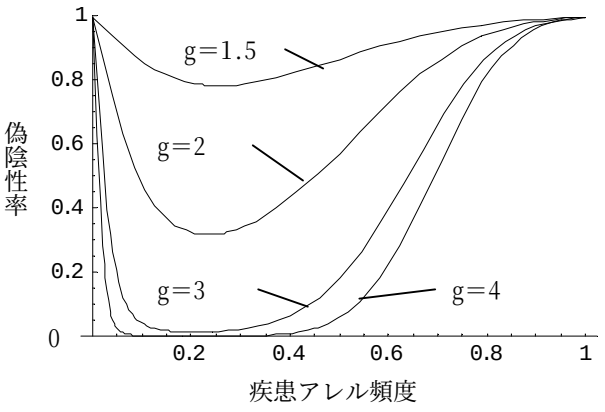


劣性遺伝形式のﾛｰｶｽを allele 頻度で検定
（α＝0.01、94VS658 でスクリーニング）



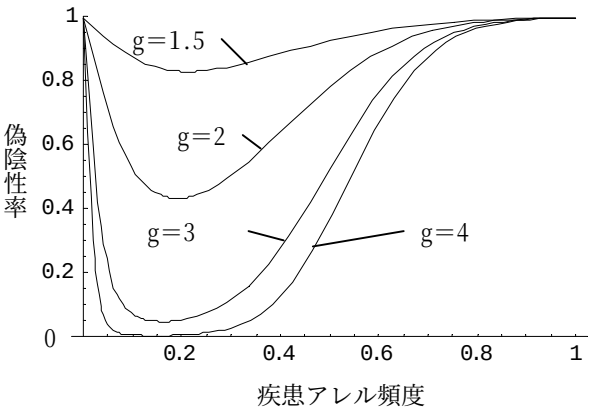
優性遺伝形式のﾛｰｶｽを優性遺伝形式の genotype 分布で検定

（α＝0.01、94VS658 でスクリーニング）



優性遺伝形式のﾛｰｶｽをllele 頻度で検定

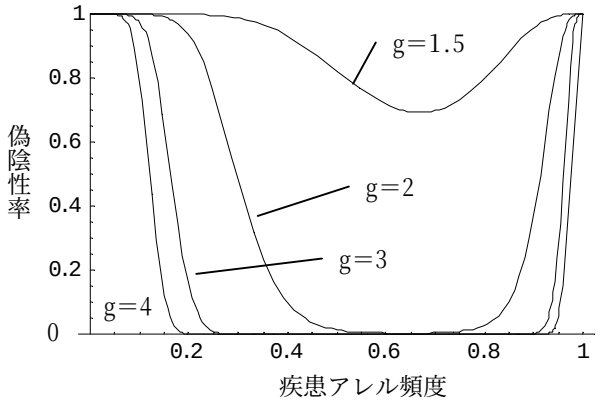
（α＝0.01、94VS658 でスクリーニング）



同様のシミュレーションを検体数を 1000 人対 1000 人に増やしたと仮定し、有意水準を 0.00001 で関連の有無を判定する場合の偽陰性率を、疾患アレル頻度と遺伝的寄与の程度を変化させてグラフ化したのが以下の図である。

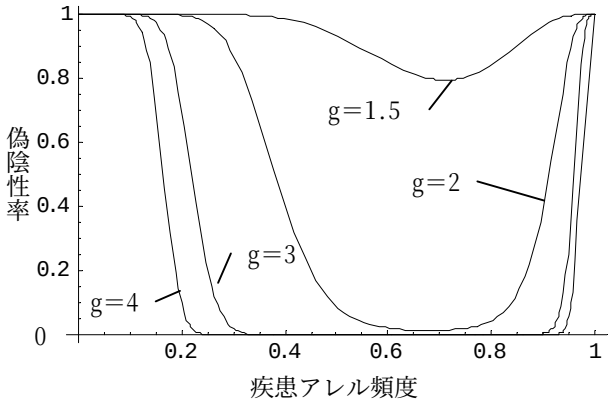
劣性遺伝形式のﾛｰｶｽを劣性遺伝形式の genotype 分布で検定

（α＝0.00001、1000 人 vs1000 人で検定）



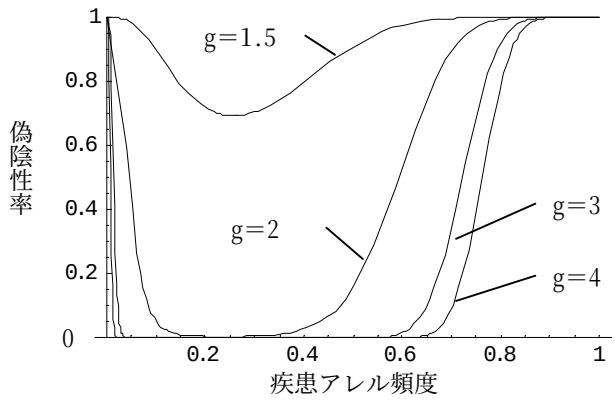
劣性遺伝形式のﾛｰｶｽをallele 頻度で検定

（α＝0.00001、1000 人 vs1000 人で検定）



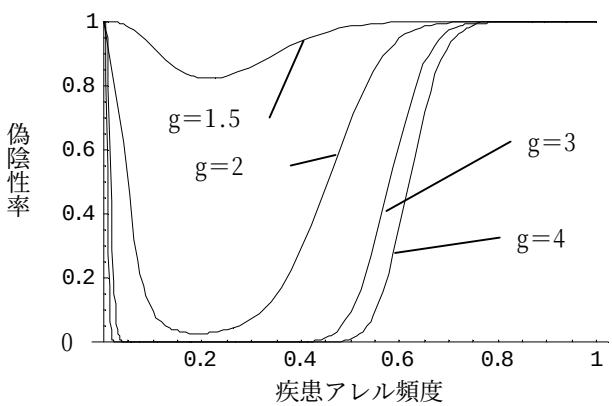
優性遺伝形式のﾛｰｶｽを優性遺伝形式の genotype 分布で検定

（α＝0.00001、1000 人 vs1000 人で検定）



優性遺伝形式のﾛｰｶｽをallele 頻度で検定

（α＝0.00001、1000 人 vs1000 人で検定）

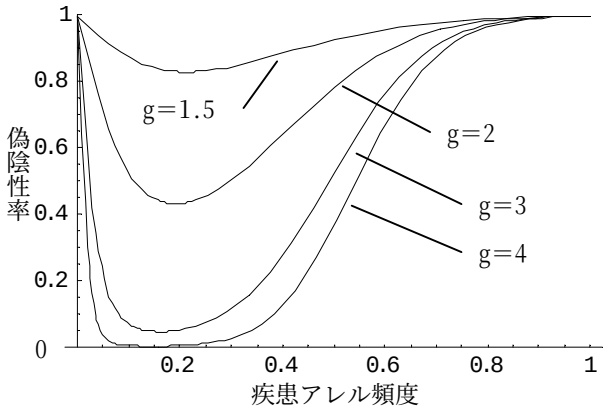


ここまでは、RIKEN SRC のスクリーニングの現状に合わせてのシミュレーションであったが、以下に、そのスクリーニングの方法を変更した場合にどうなるかを示し、今後のスクリーニングの方法の選択の判断基準を示す。

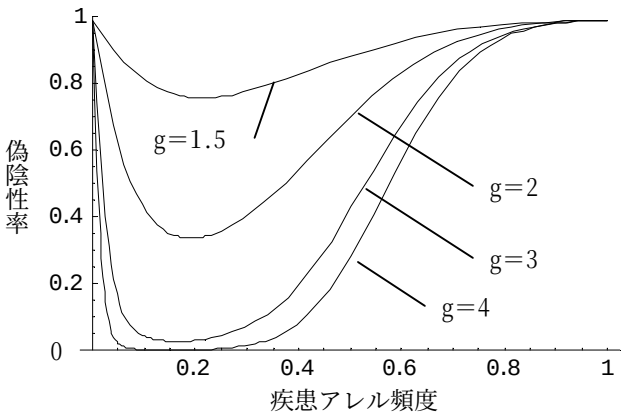
スクリーニングの方法を変更する場合に、閾値(p=0.01 から、p=0.02)を 2 倍に引き上げる場合と、検体数を 2 倍に増やす場合、閾値を引き上げ、かつ、検体数を増やした場合に、どのくらい偽陰性率が変化するかを以下に示す。

当然のことながら、閾値を 2 倍にし、検体数も 2 倍にするのがもっとも偽陰性率を下げる。閾値を 2 倍に引き上げるだけの場合と、検体数を 2 倍にするだけの場合との比較では、検体数を 2 倍にするほうが偽陰性率を低く抑えられることがグラフから読み取れる。但し、以下の例は、優性遺伝形式・アレル頻度比較の場合である。その他の遺伝形式の場合、グラフは示さないがそれらの変化も同様である。

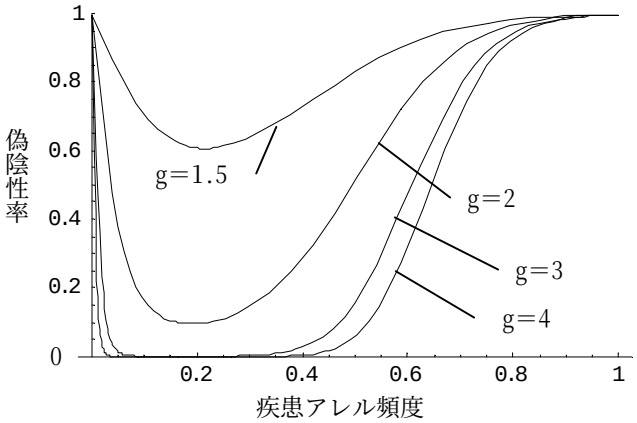
現行　ｹｰｽ 94 対ｺﾝﾄﾛｰﾙ 658　閾値　0.01



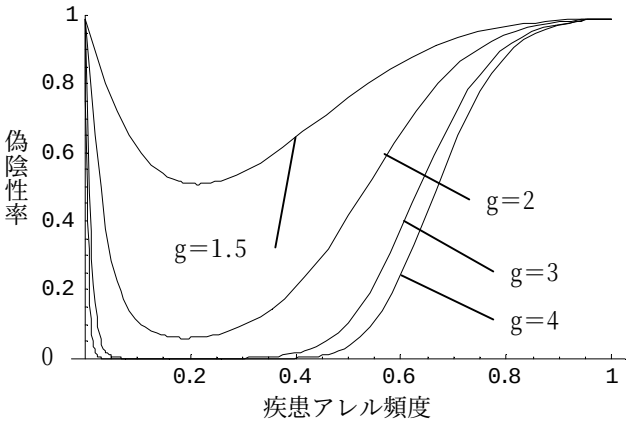
閾値 2 倍　ｹｰｽ 94 対ｺﾝﾄﾛｰﾙ 658 閾値 0.02



検体数 2 倍 ｹｰｽ 188 対ｺﾝﾄﾛｰﾙ 1,316 閾値 0.01



閾値･検体数 2 倍 ｹｰｽ 188 対ｺﾝﾄﾛｰﾙ 1,316 閾値 0.02



ここまでは、第 1 種過誤と第 2 種過誤と RIKEN SRC プロジェクトにおける関係を中心に述べてきた。しかしながら、統計学のテキストなどで扱われる場合、第 1 種過誤・第 2 種過誤と検体数のことを問題にするのは、研究計画をデザインする段階である。それに準じて話を進めると以下のようになる。

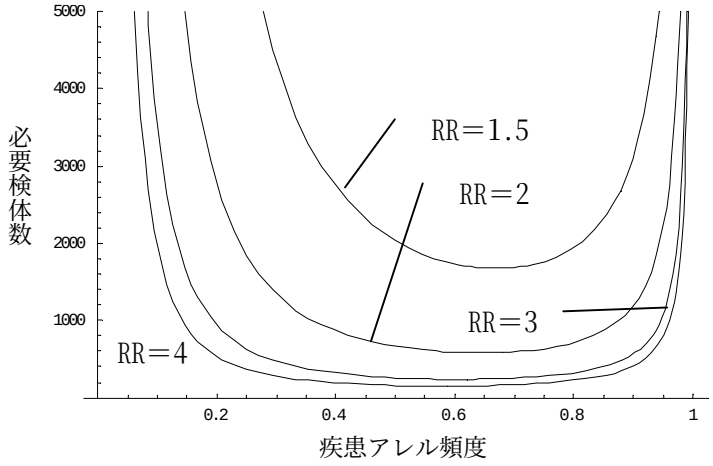
どのくらいの遺伝的寄与のある SNP を
どのくらいの有意水準(α)で
偽陽性率(β)はどのくらい容認するか

この 3 点を決めてやり、その結果得られる必要検体数を求めるという手順を踏む。

以下のグラフは、ある SNP が劣性遺伝形式または優性遺伝形式で genotypic relative risk を決めているという前提で、ケースとコントロールを、ケース集団と非ケース集団から同数ずつサンプリングする場合、指定の第１種過誤と第２種過誤との条件をクリアする検体数を、ケース・コントロールそれぞれに必要な検体数を縦軸にして描いてある。必要検体数は検定の際の、データの扱い方にもよるので、アレル頻度の差の検定を行う場合と、劣性･優性それぞれの遺伝形式を想定した上での genotype 頻度の差の検定を行う場合のそれぞれのシミュレーション結果を描いている。

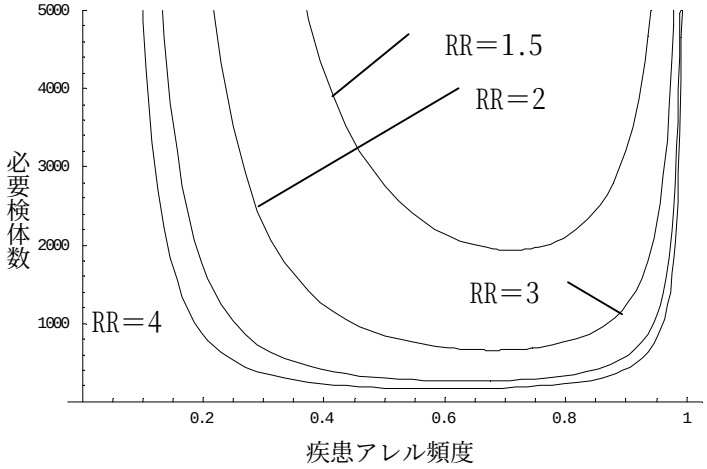
劣性遺伝形式のﾛｰｶｽを劣性遺伝形式の genotype 分布で検定

α＝0.00001、β＝0.2



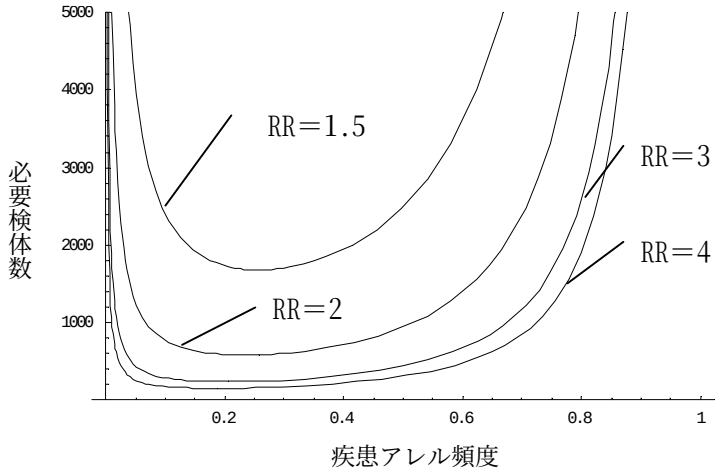
劣性遺伝形式のﾛｰｶｽの allele 頻度で検定

α＝0.00001、β＝0.2



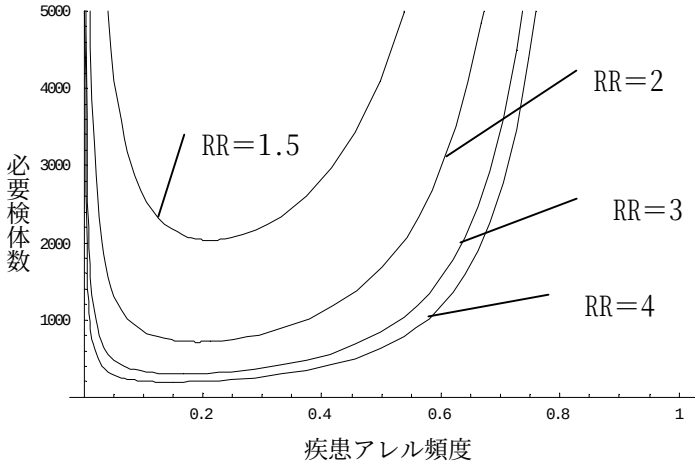
優性遺伝形式のﾛｰｶｽを優性遺伝形式の genotype 分布で検定

α＝0.00001、β＝0.2



優性遺伝形式のﾛｰｶｽの allele 頻度で検定

α＝0.00001、β＝0.2



以上の、α、β、遺伝的寄与(genotypic relative risk)、検体数との関係を知りたいときは、エクセルファイル “ β過誤率 ” および、“ 必要検体数 ” を用いてシミュレート可能である。

3-4-1 Hardy-Weinberg 平衡検定

3-4-1-1 Hardy-Weinberg 平衡（HWE）とは

ヒトを含む多くの生物は 2 倍体である(2 本の相同染色体を持つ)。したがってある多型において、遺伝子型(genotype)はホモとヘテロとからなる。2 アレル多型の SNP ではホモが 2 種類、ヘテロが 1 種類の計 3 種類の genotype が作られる。

今、ある集団内で、このアレルおよび genotype が生物学的に中立であり、この集団内でランダムな交配が行われているとする。このような場合、この集団における genotype の頻度分布はそのアレルの頻度依存性に一定となる。このような状態を Hardy-Weinberg 平衡に達しているという。

しかしながら、何らかの原因でこの平衡状態に達していないことがある。考えられる原因には以下のようなものが挙げられる。

1　SNP は中立でランダムな交配も行われているが、平衡状態に達するだけの世代が経過していない。
2　SNP は中立だが、ランダムな交配が行われていない。
3　SNP が中立でなく、特定の genotype もしくはアレルの存在が生存・生殖などに有利・不利に働く。
4　SNP は生存・生殖などに関しては中立で、適切な集団であれば平衡に達しているが、その内部のある集団(疾患罹患者集団など)において特定の genotype が高頻度もしくは低頻度になる要因がある。(例えば minor allele homozygote が、ある疾患の罹患リスクである場合には、集団全体では HWE が満たされていても、罹患亜集団においては minor allele homozygote が高率に占める。ただし、このように疾患などの特定の条件と genotype とが関連していても、必ずしも HWE から外れるわけではない。疾患と関連している SNP の HWE が保持されるのは、heterozygote の相対危険度が homozygote の相対危険度の平方根の場合である。このような場合、罹患者集団のアレル頻度と母集団のアレル頻度とは異なるが、各々の集団のなかでは HWE に達している。)

例えば 2 アレルの SNP でマイナーアレル頻度が p のときを考える。

HWE に達している集団では

minor allele homozygote frequency：p^2
heterozygote frequency：$2p(1-p)$
major allele homozygote frequency：$(1-p)^2$

となっている。

3-4-1-2 Hardy-Weinberg 不平衡の程度の評価法とその原因解明の手順

原則として関連解析に用いられる集団からの検体の genotype データは Hardy-Weinberg 平衡に達していると考えるのが適当である。しかし、何らかの原因で平衡から外れた値が得られたときに、まず、その外れの程度を評価し、その外れの原因を見極める必要がある。

＜Hardy-Weinberg 不平衡の程度の評価法＞

ある SNP に関して以下のような genotype 分布が観測されたとする。

観測値(1, 2 はそれぞれ minor allele , major allele を表す)

Genotype	11	12	22	sum
観測度数(N_{oij})	4	20	76	100
観測頻度(P_{oij})	0.04	0.2	0.76	

今、minor allele frequency = (4+20/2)/100 = 0.14
であるから、
期待値（HWE 成立）

Genotype	11	12	22	sum
期待度数(N_{eij})	2	24	74	100
期待頻度 (P_{eij})	0.02	0.24	0.74	

上の 2 つの genotype 分布の乖離の程度を評価する方法が 3 つある。そのうちの 2 つ（1, 2）を RIKEN SRC のプロジェクトでは常用する。

1. χ^2検定
2. Homozygote excess
3. 尤度比検定

1. χ^2 検定 (Ref. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. by DM Nielsen, MG Ehm and BS Weir, Am J Hum Genet 63: 1531-1540, 1990)

観測度数と期待度数とから Pearson 's χ^2を求める。この分布の自由度は 1 である。この方法は検定であるから、Hardy-Weinberg 平衡が成り立っているという仮説を帰無仮説としてその対立仮説の確からしさを判断する材料になる。

$$\chi^2=(N_{o11}-N_{e11})^2/N_{e11}+(N_{o12}-N_{e12})^2/N_{e12}+(N_{o22}-N_{e22})^2/N_{e22}$$

2. Homozygote excess

この指標は、homozygotes の観測度数が、homozygotes の期待観測度数に対してどのくらい過剰かま

たは不足しているかを、heterozygotes の期待度数に対する比として表したものである。こちらは１のχ^2検定と異なり、Hardy-Weinberg 平衡からのずれの程度を確率的に定量化することはできないが、Hardy-Weinberg 平衡からのずれが、ホモ個体が増える方向に向かっているのか、その逆にヘテロ個体が増える方向に向かっているのかを教えてくれる。この情報は Hardy-Weinberg 不平衡を示すデータが得られた原因を分類する上で役に立つ。(詳しくは、本章次項の Hardy-Weinberg 不平衡の原因の究明を参照)(Ref. A novel MHC class 1-like gene is mutated in patients with hereditary haemochromatosis. by JN Feder et al. Nat Genet 13: 399-408, 1996Statistics in human genetics by Pak Sham Arnold, pp39-44)

3. 尤度比検定(Ref. Testing heterozygote excess and deficiency. by F Rousset and M Raymond, Genetics 140: 1413-1419(1995))

この方法では、観測頻度が集団の genotype 頻度であるという仮定と期待頻度がそれであるという仮定から、観測度数が得られる確率を計算し、両者の差の程度から、観測度数が HWE を満たす集団を代表している確率を計算する。

観測頻度が集団の genotype 頻度であるという仮定において観測度数を得る確率 Lo
期待頻度が集団の genotype 頻度であるという仮定において観測度数を得る確率 Le とする。

$$L_o = N_{o11} \times \ln(P_{o11}) + N_{o12} \times \ln(P_{o12}) + N_{o22} \times \ln(P_{o22})$$

$$L_e = N_{o11} \times \ln(P_{e11}) + N_{o12} \times \ln(P_{e12}) + N_{o22} \times \ln(P_{e22})$$

$$\Delta L = Lo - Le$$

$$= N_{o11} \times \ln(P_{o11}/P_{e11}) + N_{o12} \times \ln(P_{o12}/P_{e12}) + N_{o22} \times \ln(P_{o22}/P_{e22})$$

$2 \times \Delta L$ は自由度１のχ^2分布に近似されるので、
その値から観測度数が HWE を満たす集団を代表している確率がわかる。

個別のデータを、χ^2検定および Homozygote excess-deficiency 法により、Hardy-Weinberg 平衡評価をするためには、エクセルファイル“3-4-1-4 1 SNP data の Hardy-Weinberg 平衡評価 2 方法”を用いる。
また、エクセルファイルによる個々の SNP genotype data の処理ファイルである、“3-3-4 1 SNP data-analysis-set”を用いても、ケース群・コントロール群各々の Hardy-Weinberg 平衡 χ^2検定結果が、ケースコントロール関連解析・相対危険度計算の結果とともに得られる。

＜Hardy-Weinberg 不平衡の原因解明の手順＞

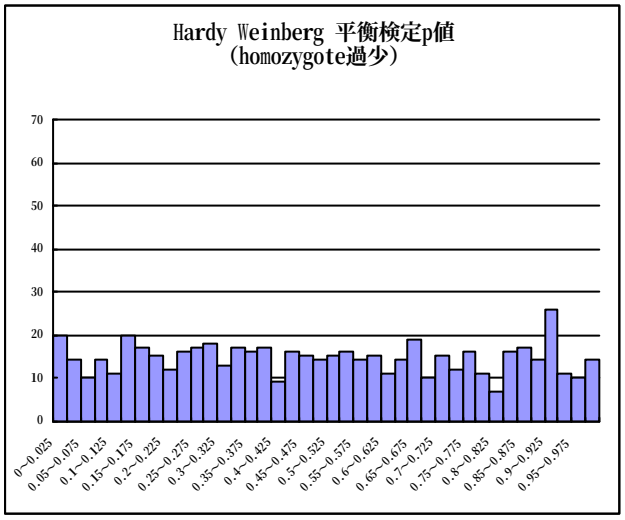
Hardy-Weinberg 不平衡の原因には大きく分けて２つある。1 つは実際の集団の Genotype の分布が Hardy-Weinberg 平衡から外れている場合であり、もう 1 つは実際の集団の Genotype 分布は Hardy-Weinberg 平衡を満たしているのにサンプリングの段階もしくはタイピングの段階で平衡から外れている場合である。列挙すると以下の通りである。このテーマは“4 データの質の管理のための検定”にて扱われているのでその項も参照のこと。

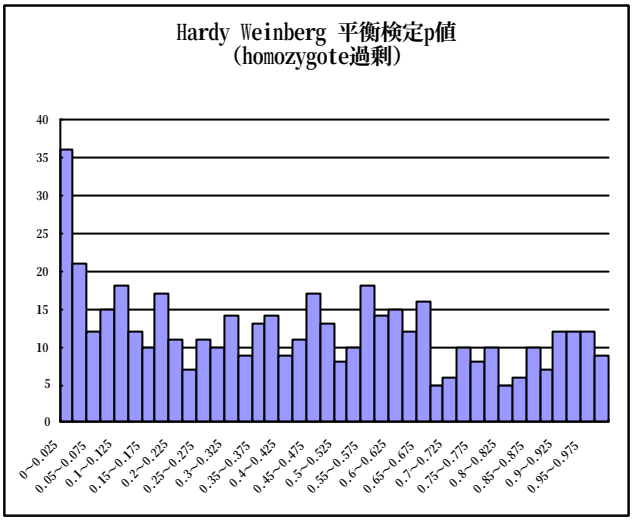
1．集団の genotype 分布が Hardy-Weinberg 不平衡にある
(1) SNP は中立でランダムな交配も行われているが、平衡状態に達するだけの世代が経過していない場合（水とお湯を混ぜたとき、すぐには全体が一定の温度には達しないのと同様の減少）
(2) SNP は中立だが、ランダムな交配が行われていない場合（集団に階層化がある）
(3) SNP が中立でなく、特定の Genotype もしくはアレルの存在が生存・生殖などに有利･不利に働く場合
(4) ある phenotype の集団で、その phenotype を生じさせる頻度が genotype によって異なる場合（疾患関連 SNP の場合）
2．集団の genotype 分布が Hardy-Weinberg 平衡にある
(1) サンプリングによる不平衡の出現
① ランダムサンプリングだが偶然に不平衡結果を得る場合
② サンプリングが当該 SNP に関して非ランダムに行われる場合
a) 異なるアレル頻度で Hardy-Weinberg 平衡に達している複数の集団からサンプリングした結果を 1 集団からのサンプリングとして扱う場合
(2) タイピングによる不平衡の出現
① SNP 自体の問題
a) SNP ではない
ｱ) 多型が認められない場合
ｲ) 相同配列間の差を多型とみなしてアッセイを行う場合
② タイピングに用いるテンプレートの問題
a) タイピングアッセイに PCR を用いる場合には、テンプレートである PCR 産物量にアレル特異的なアンバランスが生じうる。（連鎖している SNP が PCR プライマー上にある場合に、アレル特異的に PCR 効率に差が生じ、産物量がアンバランスとなる）
③ アッセイに関する問題
a) アレル特異的反応がアッセイ対象 SNP のアレルと 1 対 1 対応していない
ｱ)アレル特異的反応のアンバランス
○対象 SNP のアレル依存性にタイピング反応アンバランスがある
○対象 SNP の近傍配列依存性にタイピング反応アンバランスがある
b) アレル数の設定ミス(3 アレル SNP を 2 アレルとしてアッセイする、など)
④ ミスコール

RIKEN SRC 初期データの検討結果とその数式的解釈のためには、“3-4-1-3 参考 Hardy-Weinberg 平衡検定 p 値分布が、homozygote 過剰の場合に統計的有意 p 値を示しがちなことに関する考察 ”を参照のこと。

3-4-1-3 参考 Hardy-Weinberg 平衡検定 p 値分布が、homozygote 過剰の場合に統計的有意 p 値を示しがちなことに関する考察

RIKEN SRC SNP genotyping 初期データの Hardy-Weinberg 平衡検定の結果をまとめたものである。Homozygote が過剰になる場合と過少になる場合とに二分してその平衡からのずれの程度を χ^2 検定し、その p 値の分布をとると、homozygote 過剰群で有意 p が期待されるよりも多く認められるのに対し、homozygote 過少群では p 値の分布は期待通り(一様分布に似る)に分布する。この理由を明らかにするために一般論として Hardy-Weinberg 不平衡の原因とそれが homozygote 過少・過剰のどちらにずれるかを確認した。





Hardy-Weinberg 不平衡の原因	不平衡の方向(ホモ過剰・過少)
1．不適切な genotype 分類 例 1　3 allele SNP を 2 allele SNP としてタイピング 例 2　X 染色体上 SNP を常染色体上 SNP としてタイピング	Homozygote 過剰
2．非任意交配/繁殖(Non-random mating/breeding) (1)　　近親交配(inbreeding) 集団規模でこれがおこると階層化(stratification)が生じる (2)　　同類交配(assrotative mating/breeding)	Homozygote 過剰
3．Genotype が生存・生殖に影響を与える場合	場合による
4．選択的サンプリング(例 case を sampling)	場合による

上記テーブルが示すように、homozygote 過剰になる場合が多いことが確認された。

p 値が 0.001 を切る SNP に関してその SNP に関する情報、タイピングの生データなどを検討した結果、多型でない塩基を SNP として誤登録していた例、相同配列の差を SNP として登録していた例、対象 SNP の近傍にもう 1 つの SNP があったために PCR 産物の産生量にアンバランスの認められた例、同じくタイピングプローブの反応効率に差が生じていた例、ミスコールによる例、常染色体上 SNP と性染色体上 SNP との扱いを混同していたために異常とされた例が認められた。

これらに該当しない例で $p<0.00001$ の有意差を出す例はまだ認められていない(平成 13 年 2 月末現在)。次に上に挙げた例の homozygote 過剰となるか、過少となるか、どちらになるか不定であるかを数式的に検討した結果を添付する。

＜HW 不平衡の方向に関する数式的確認＞

1-例 1　不適切な genotype 分類 3 allele SNP を 2 allele SNP としてタイピング

ある SNP が 3 アレル A、B、C を持つとする。ある assay 系にてタイピングを行うと A、B は決定できるが、C は結果が出ないものとすると、実際の genotype とこの assay 系でのタイピング結果とは以下の表のような関係となる。

真の genotype	アッセイ上の genotype
AA	AA
AB	AB
AC	AA
BB	BB
BC	BB
CC	-

今、アレル頻度を a、b、c(但し a+b+c＝1)とし、HWE が保たれているものとする。

このような assay で得られた genotype の結果は、

アッセイ上の genotype	アッセイ上の genotype 頻度
AA	a^2+2ac
AB	2ab
BB	b^2+2bc
No Call	c^2

ここで homozygote の過剰を表す関数 H を次のように定義する。

$$H=Q(AA)\times Q(BB)-(1/2\times Q(AB))^2$$

H の値は HWE の時 0

ホモ過剰で HWD の時正、ホモ過少で HWD の時負 となる。

なぜならば、HWE の時、

$$Q(AA)=x^2 \qquad Q(AB)=2x(1-x) \qquad Q(BB)=(1-x)^2$$

であり

$$Q(AA)\times Q(BB)=1/4Q(AB)^2$$

が成り立っているからである。

今

$$\begin{aligned} H&=(a^2+2ac)\times(b^2+2bc)-(ab)^2 \\ &=2abc(a+b+2c) \\ &=2abc(1+c)>0 \end{aligned}$$

より、homozygote 過剰となる。

1－例 2　X 染色体上 SNP を常染色体上 SNP としてタイピング

ある X 染色体上の SNP がアレル A、B で構成されているとする。それとは知らずに常染色体上の SNP としてタイピング結果を取り扱ったとすると、

	真の genotype	常染色体上 SNP として扱った場合の genotype
女	AA	AA
	AB	AB
	BB	BB
男	AO	AA
	BO	BB

今、アレル頻度を a、b（a+b=1）とし、男女比 50：50 でサンプリングしているとすると、

アッセイ上の genotype	アッセイ上の genotype 頻度
AA	$1/2\times a^2+1/2\times a$
AB	$1/2\times 2ab=ab$
BB	$1/2\times b^2+1/2\times b$

今、$H=(1/2\times a^2+1/2\times a)\times(1/2\times b^2+1/2\times b)-(1/2\times ab)^2$

$=1/4\times ab(a+b+1)$

$=1/2\times ab>0$

2－(1)　階層化の例

今、ある混合集団 P には階層化が存在し、下位集団 p１、p２にて構成されているとする。

p１、p２の中では HWE が成立している。

この階層化とは独立にサンプリングを行い、p_1、p_2それぞれから同数のサンプルが得られたとする。

ある SNP が 2 アレル A、B からなり、p_1、p_2でのアレル頻度がそれぞれ a_1、b_1、a_2、b_2であるとする。

（但し、$a_1+b_1=1$、$a_2+b_2=1$）

	genotype	genotype 頻度
p_1	AA	a_1^2
	AB	$2a_1b_1$
	BB	b_1^2
p_2	AA	a_2^2
	AB	$2a_2b_2$
	BB	b_2^2
P	AA	$1/2a_1^2+1/2a_2^2$
	AB	$1/2\times 2a_1b_1+1/2\times 2a_2b_2$ $=a_1b_1+a_2b_2$
	BB	$1/2b_1^2+1/2b_2^2$

$H=(1/2a_1^2+1/2a_2^2)\times(1/2b_1^2+1/2b_2^2)-(1/2\times(a_1b_1+a_2b_2))^2$

$=1/4(a_1b_2-a_2b_1)^2$

$=1/4\{(a_1(1-a_2)-a_2(1-a_1))\}^2$

$=1/4((a_1-a_2)^2\geqq 0$

2－(2)　同類交配も想定上は近親交配と同様であり、結論とその数式上の確認過程は同じ。

3－例 1　genotype が生存・生殖に影響を与える場合

胎生致死劣性遺伝では、SNP アレル A、B、アレル頻度 a、b（a+b=1）のとき、

genotype	genotype 頻度
AA	$a^2/(a^2+2ab)$
AB	$2ab/(a^2+2ab)$
BB	0

$H=a^2\times 0-(a^2+2ab)=-a^2\times b^2<0$

3－例 2　genotype が生存・生殖に影響を与える場合

夭逝優性遺伝形式を有するサンプルのとき、夭逝 genotype の 1/2 が死亡しているとすると、同様の SNP にて

genotype	genotype 頻度
AA	a^2
AB	$1/2\times 2ab=ab$
BB	$1/2\times b^2$

$H=(a^2\times 1/2\times b^2)-(1/2\times ab)^2$

$=1/2(a^2\times b^2)-1/4(a^2\times b^2)$

$=1/4(a^2\times b^2)<0$

4　選択内サンプリング

遺伝性疾患のケースサンプリング genotype AA、AB の個体の genotype BB に対する相対危険度をそれぞれ

$g(g>1)$　$g^f(0\leqq f\leqq 1)$

とすると

genotype	genotype 頻度
AA	$g\times a^2/C$
AB	$g^f\times 2ab/C$
BB	b^2/C

但し、$C=g\times a^2+ g^f\times 2ab + b^2$

$H=(g\times a^2/C)\times(b^2/C)-(1/2\times(g^f\times 2ab)/C)^2$

$=1/C^2(ga^2b^2-g^{2f}a^2b^2)$

$=a^2b^2/ C^2(g-g^{2f})$

$f=1/2$ のとき $H=0$,

$f>1/2$ のとき $H<0$,

$f<1/2$ のとき $H>0$

Heterozygote の相対危険度が homozygotes のそれの平方根の場合(f=1/2 の場合)は SNP と疾患に相関はあるものの疾患集団内では HWE に達している。しかしながらこのとき、疾患集団でも一般集団でも HWE に達しているが、疾患集団の疾患アレル頻度は一般集団（こちらも HWE に達している）の疾患アレル頻度よりも高い点が異なる。

3-4-1-5 Hardy-Weinberg Disequilibrium が疾患 phenotype(s)と関連して認められた場合にどうするか

簡単に言えば：

1　2×3 分割表検定による分布差が認められれば、追及する価値がある。

2　HWE からの逸脱の理由を解明する→3-4-1-2、3-4-1-3 参照

その解説

HWE からの著しい逸脱は原則として、アッセイ系などの系統的な原因に起因し、疾患との関連を丁寧に追及することが適さないことが多いと予想され、本資料集でもそのような立場からアプローチすること(しないこと)を勧めている。

しかしながら、triallelic SNPs の場合や、SNP を含む領域に large deletion もしくは large insertion がある場合には、3 番目のアレル(insertion 及び deletion もアレルとみなす)はこの HWE からの逸脱をもたらしたアッセイ系によっては検出することが出来ず、それが理由で homozygote-excess を生じさせてしまうことは 3-4-1-2 及び 3-4-1-3 で述べた。このような第 3 のアレルが疾患 phenotype と関連している事例も知られていることから、その参照文献(Human Hypertension Caused by Mutations in WNK Kinases. FH Wilson et al. Science 293:1107-1112,2001)を提示するとともに、第 3 の検出不能アレルの存在を仮定した尤度比検定を行うためのエクセルファイル(3-4-1-6 検出不能な第 3 アレルを仮定した尤度比検定のエクセルファイル)を紹介する。

このエクセルはあくまでも隠れた第 3 のアレルの存在を予想するためのものであり、実在するかいなかは他の方法で確認する必要があることはもちろんである。

また、疾患 phenotype 間で genotype 分布に差が無ければ、追求することに意味が無いので、疾患関連 hidden 3rd allele の検討は、疾患 phenotype による分割表検定を行って有意であるかどうかを先に調べる必要がある。

このときに用いるべき分割表は 2×3 分割表である。なぜならば、それ以外の分割表(アレル別 2×2 分割表、優性・劣性遺伝形式 genotype 別 2×2 分割表)はそれぞれ、biallelic markers の場合に生物学的な意味を持つことから、集計して作成された表であり、biallelic でない場合には集計すること自体が無意味である。そのような場合には、生のデータである 2×3 分割表の観測度数分布が比較 phenotype 群間で差があるかどうかを検定することが先決である。

疾患 phenotype 関連 hidden 3rd allele の存在が示唆されたら、3rd allele の存在を確認し、さらに、それを含めたアレル検出系を用いて、新たな分割表を phenotype 別に作成して関連の有無を調べる必要がある。

3-5　複数点解析

3-5-1 マッピング(有意相関 SNP が検出された場合の Linkage Disequilibrium mapping、Hardy-Weinberg 不平衡 mapping の手順)

全体の流れ

1　前後 SNP 地図を作る
2　前後 SNP を分類する
3　検定有意水準マップを作る

ゲノムワイドスクリーニングで相関が有意となった single SNP を SNP X と呼ぶこととする。

各ステップ詳細

1．前後 SNP 地図を作る

考慮すべきことは、

(1)範囲の設定

mapping に有用な LD が認められる範囲

LD が認められる範囲の推定

一般論からの類推(default 範囲とする)

SNP X 周囲の既知の LD から推定

(2)SNP の出所

東京大学医科学研究所/JST SNP db

新規 SNP discovery を行う

以上を実現するための具体的な作業内容は以下のようになる。

Default 範囲(前後 50kb*ずつ)の既知 SNP を東京大学医科学研究所/JST SNP db から抽出する。

Default 範囲(前後 50kb ずつ)で、かつ東京大学医科学研究所/JST プロジェクトがスクリーニングしていない部分に関して、

SNPdiscovery を行う。Repeat-Masker を“Do not mask simple repeats and low complexity DNA”の設定にした上で、該当領域のスクリーニング対象を設定し、SNPdiscovery を行う。

Default 範囲の全 SNP について LD の概数が得られるだけのタイピングを行う。(但し、SNP X との LD の程度に関わらず、範囲内の全 SNP につき全サンプルのタイピングをするのであれば、ステップ 2 は無意味となる。)

当該領域固有の LD の広がりについての情報が得られるので、その情報を基にさらに広い範囲の SNP 地図の作成が必要であると判断されたら、その範囲について東京大学医科学研究所/JST SNP db を検索し、また新規 SNP discovery を行う。

2．前後SNPを分類する

上記1にてSNP X周囲のSNP地図とそれらのLDの程度についての情報がある。この情報により周囲SNPは以下の2群に大別される。今、前後SNPをSNP Yと呼ぶ事とする。

(1) SNP Xを解析上有意と判断される程度のLD(D'＞0.57：論拠は“3-3-2-3-5 参考 連鎖不平衡の実際及び間接関連検出用のSNPマーカーと真のローカスとの関係について”)をもつSNP Y
(2) SNP Xを解析上有意と判断される程度のLDをもたないSNP Y

(1) SNP Xを解析上有意と判断される程度のLDをもつSNP Y の説明

これは真の疾患ローカス(D)とSNP XとSNP Yとが1つのancestral haplotypeを形成していた場合に相当する

SNP X-SNP Y-D
SNP X-D-SNP Y
D-SNP X-SNP Y

の3通りの並び方が考えられる。

(2) SNP Xを解析上有意と判断される程度のLDをもたないSNP Yの説明

疾患ローカス(D)は複数の疾患ハプロタイプによって構成されていることが必要である。SNP X の疾患アレルと連鎖しているアレル(疾患アレル)が作るハプロタイプがその一つで、第2の疾患ハプロタイプに疾患ローカス・アレルとSNP Yの疾患アレルとが乗っていることになる。

この場合の疾患ローカス(D)は、ある遺伝子として定義されその中に 2 ヶ所の多型が存在することもあるし、おそらくそれよりはるかに低い確率で、ある遺伝子の1ヶ所の多型が2つの異なる疾患アレルを有することもあり得ない話ではなく、またその第2の例のバリエーションであるが異なる2つの疾患ancestral haplotypeがたまたま、疾患ローカス遺伝子の同一部位に同一の多型アレルを持っていたという可能性もやはり不可能ではない。2つの疾患ハプロタイプは、

SNP X-D
SNP Y-D

と、表記できる。

従って理論的にはSNP XとSNP Yのみでなく、連鎖していないSNP Yも多人数でタイピングすることが必要であるが、連鎖していないSNP Yの場合には検定で統計的有意差を得る可能性は相当程度低いことが容易に予想される。

しかしながら疾患 ancestral haplotype が非常に限定される“いわゆる遺伝病(発症率に民族差や地域差があることが普通で、疾患ローカス数も1つないしは極めてそれに近いような疾患)”においてはこのようなSNP Xと連鎖していないSNP Yが疾患ローカスと連鎖不平衡にある可能性がほとんどないことに比べれば、common diseaseの場合には無視できないレベルかもしれないと言える。

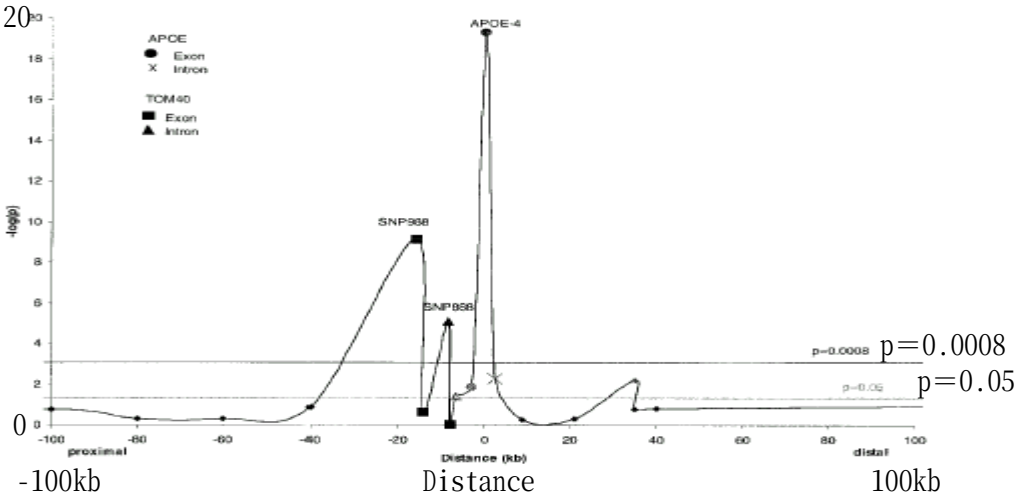
3．検定有意水準マップを作る(添付図参照)

大別して 2 種類の互いに独立した 2 検定

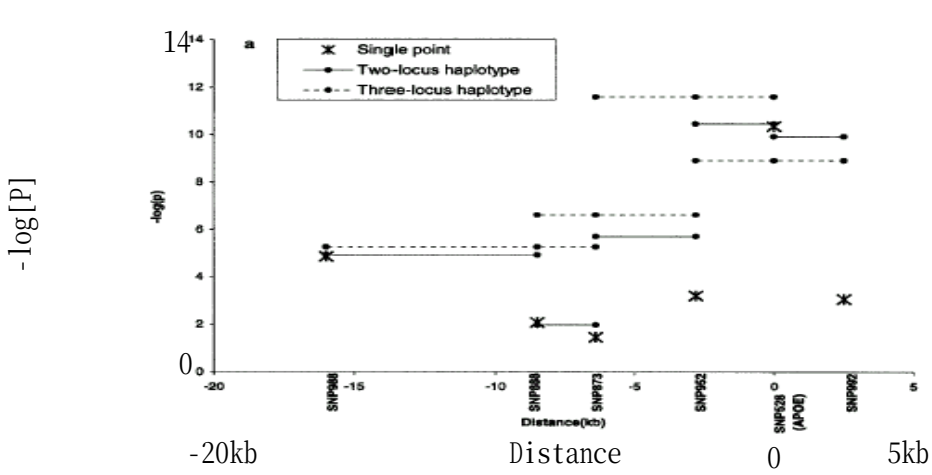
(1) ケース・コントロール相関検定

(ER Martin et al. Am J Hum Gent 67:383-394,(2000))

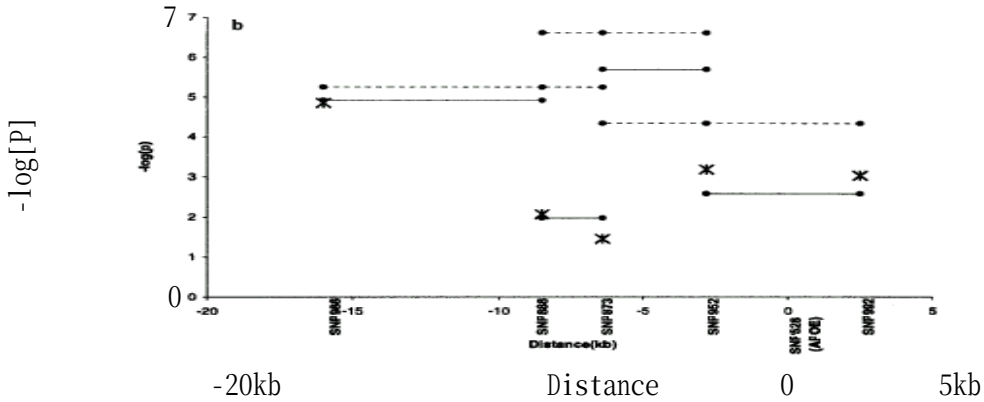
① 1 SNP ケース・コントロール相関



② Haplotype ケース・コントロール相関(真のローカスが解析に含まれる)

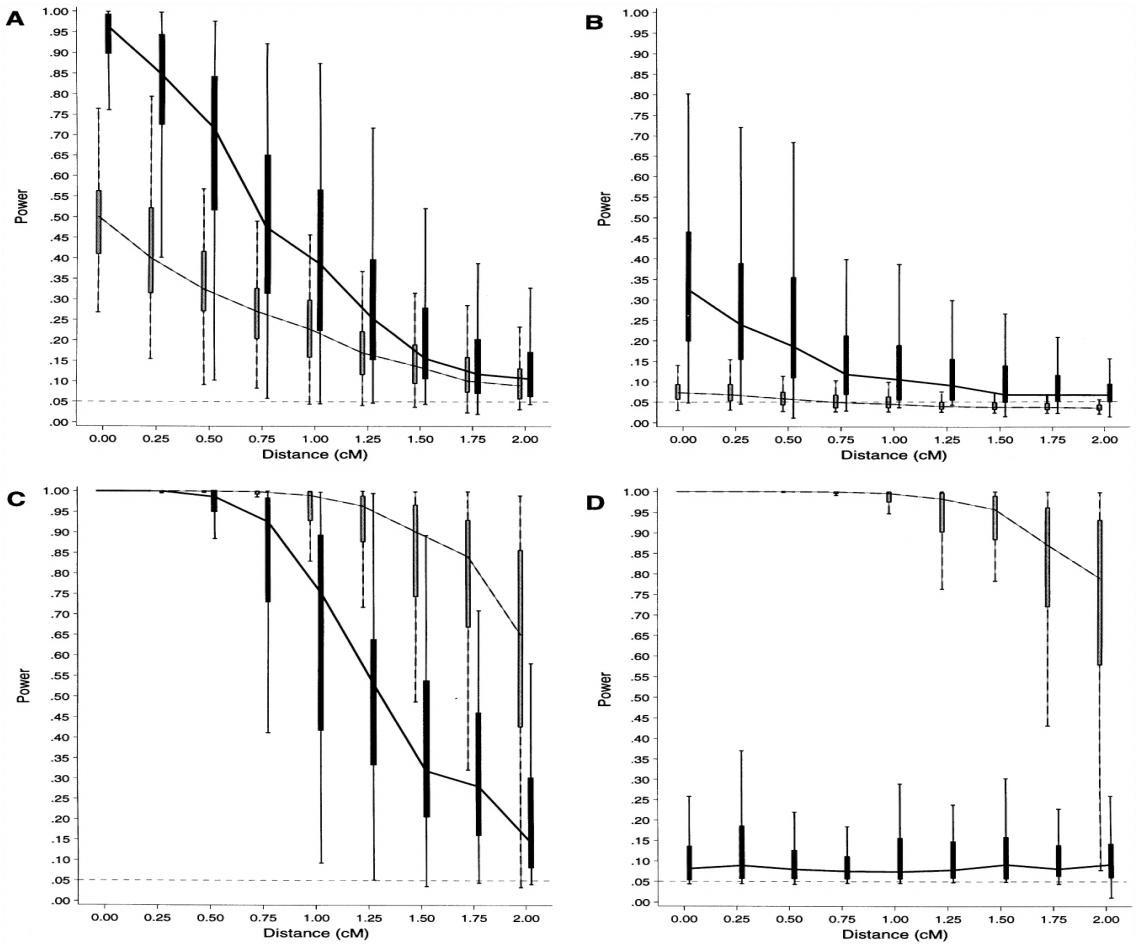


(真のローカスが解析に含まれない)



(2) Hardy-Weinberg 平衡検定

(DM Nielsen et al. Am J Hum Gent 63:1531-1540,(1999))



一般的に Hardy-Weinberg 平衡は LD よりも急峻に減衰することが知られている。また遺伝形式がはっきりしない場合(優性・劣性などではない場合)にはそもそも有意な不平衡が認められていない場合も多いことに留意する必要がある。

なお、(1)はケースとコントロールの両方のデータを用いた検定であり、(2)はケースのデータのみを用いた検定である。両者が揃って疾患ローカスの存在を当該 SNP の周辺に認めれば、独立の 2 解析がその仮説を支持することになり、有力である。

3-5-2　ハプロタイプ

3-5-2-1 ハプロタイプとは

ハプロタイプとは複数の遺伝子多型が 1 本の染色体上に形成するアレルの組合せのことである。例えば、2 つの SNP が近接しているとする。あるヒトのその 2 つの SNP の genotype がいずれもヘテロであったとする。1 つ目の SNP のアレルを A/a とし、もう 1 つの SNP のそれを B/b とする。このとき、このヒトのハプロタイプの可能性は、

{A-B, a-b}

{A-b, a-B}

の 2 通りある。この 2 通りのうち、どちらであるかを決めるためには染色体別に SNP のアレルを決めてやる必要がある。それは、1 つの SNP の genotype を決める作業に比べてはるかに煩雑な作業であり、2 つの SNP 間の距離によっては非常に難しいこともある。

このように、ある個人の複数の SNP によるハプロタイプを決定することは簡単ではないが、ハプロタイプを知ることが有益であることも多い。例えば、ある疾患関連ローカスがあるハプロタイプと連鎖している場合には、そのようなハプロタイプを同定することが有用である。また、複数の多型が及ぼす影響が cis エレメントとして機能する場合には、ハプロタイプ構成を明らかにしなければ複数の多型の総合的影響を判断することはできない。

以上のようにハプロタイプ構成を調べることの有用性は挙げられたが、実際問題として実験的にハプロタイプを決定することはやはり煩雑である。そこで個々の多型の genotype データも用いてそれらが作るハプロタイプの割合を推定する方法が提案されており、実際のデータと照らし合わせても悪くない結果が得られている。そのような方法は複数提案されているが、最も一般的なのが EM-algorithm と呼ばれる手順で推定する方法である。その原理と、2〜4 個の SNP のデータからハプロタイプ頻度を推定するためのエクセルファイルを添付する。

3-5-2-2 参考 EM-algorithm による haplotype 頻度の推定

2 SNP の観測データからは 9 genotype の観測値が得られる。そのような観測値を最も得やすい 4 haplotype 頻度の分布を求めるのに用いる標準的アルゴリズムが Expectation-Maximization algorithm (EM-algorithm)と呼ばれるものである。

2 SNP の 9 genotype 観測値から 4 haplotype 本数を推定するにあたり、genotype は 2 種類に分けられる。片方の種類に分類される genotype は特定の haplotype の本数を一意的に決定する(確定的 haplotype 数)。もう片方に分類される genotype は特定の haplotype の本数を確率的に決定する(非確定的 haplotype 数)。

EM-algorithm では非確定的 haplotype 数を 4 haplotype に均等に分配し、確定的 haplotype 数と合算し、その値を基に haplotype 頻度の初期推定値を算出する。ついで、その haplotype 頻度推定値を基に、非確定的 haplotype 数を分配し、新たな haplotype 頻度の推定値を得る。ついで、新たな haplotype 頻度推定値を基に非確定的 haplotype 数の再分配を行う。この作業を繰り返すと haplotype 頻度は収束することが知られ、しかもその収束推定値は現実のデータとの整合性が高いことが知られている。エクセルファイル“2 SNP Haplotype 頻度推定”でも EM-algorithm を使用している。

参考のため、その具体的な数式を以下に掲載する。

X/x、Y/y はそれぞれ１つの SNP の 2 つのアレルを表している。
a1…a9 は X/x、Y/y の作る 9 genotype の観測値を表している。
a1 は XXYY、a2 は XXYy、a3 は XXyy、… a8 は xxYy、a9 は xxyy に相当する。
また、p_{XYn}は haplotype XY の世代 n における推定頻度を表す。ただし、下付き数字 n は 0 の場合、haplotype 推定初期値に対応し、1 以上の場合は上記で説明した非確定的 haplotype 数の分配作業の回数に相当する。
S＝a1+a2+…+a9 である。

$$p_{XY0}=(a1+1/2\times a2+1/2\times a4+1/4\times a5)/S$$
$$p_{Xy0}=(a3+1/2\times a2+1/2\times a6+1/4\times a5)/S$$
$$p_{xY0}=(a7+1/2\times a4+1/2\times a8+1/4\times a5)/S$$
$$p_{xy0}=(a9+1/2\times a6+1/2\times a8+1/4\times a5)/S$$

$$p_{XYn}=(a1+1/2\times a2+1/2\times a4+(p_{XYn-1}\times p_{xyn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{Xyn}=(a3+1/2\times a2+1/2\times a6+(p_{Xyn-1}\times p_{xYn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{xYn}=(a7+1/2\times a4+1/2\times a8+(p_{Xyn-1}\times p_{xYn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{xyn}=(a9+1/2\times a6+1/2\times a8+(p_{XYn-1}\times p_{xyn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$

3-5-2-6 ハプロタイプ解析のバックグラウンド --- 多型・diversity の評価法・表記法を含めて
― 文献を読みながら理解しよう ―

Haplotype 解析

参考文献など

① DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene,Nat Genet 19 233-240,1998 D.A.Nickerson et al

② Haplotype Structure and Population Genetic Inferences from Nucleotide-Sequence Variation in Human Lipoprotein Lipase,Am J Hum Genet 63 595-612,1998 A.G.Clark et al

③ The Haplotype Map Meeting on July 18-19,2001,in Washington D.C. U.S.A.

<“多型学”に出現するいろいろな用語など>

（1） Heterozygosity

ある多型に着目したとき、その多型がヘテロである個体が集団に占める割合。

HWE 下では(SNP の場合) 2p(1-p)：但し p はアレル頻度 すなわち 12/(11+12+22)

（2） Heterozygosity のとりうる範囲

Diallelic 多型が HWE を満たしているとき

0≦Heterozygosity≦0.5

h=2p(1-p) を p について解くと、$p=\frac{1\pm\sqrt{1-2h}}{2}$

h=0：p=0 or 1

h=0.5：p=0.5

（3） HWE からのズレの表現

χ^2検定、Homozygosity excess “資料集 3-4-1-2 Hardy-Weinberg 不平衡の程度の評価法とその原因解明の手順”参照

(観測 Heterozygosity)-(HWE 期待 Heterozygosity)の差 ― 文献① Fig3 a

（4） Linkage disequilibrium に 2 つの表現(文献② p604)

Absolute disequilibrium(4 haplotypes のうち 2 haplotypes のみが存在する)

Complete disequilibrium(4 haplotypes のうち 3 haplotypes のみが存在する)

（5） 使用多型数を増やすと集団特有のハプロタイプの割合が増し、全ての個人を個別の集団の代表とみなせば、1(2)ハプロタイプということになる。(言われてみれば当然)

88 多型のつくる 88 ハプロタイプのうち 81 ハプロタイプは 3 群のいずれか 1 群にしか認められなかった。(文献 ②)

（6） Nucleotide Variation の指標(πとθ)

（ア）π：average heterozygosity per site

random に選んだ 2 本の相同染色体上のある塩基 site が相互に異なる塩基を持っている確率。多型がゲノム上に存在する密度とその多型の heterozygosity に依存する。

$$\pi = \frac{\sum_{j=1}^{s}(\text{Heterozygosity@多型j})}{N}$$

s：多型の個数、N：評価している範囲の塩基数

ある個人の塩基配列を K 塩基対にわたって調べたときに、P 塩基でヘテロであったとする。

$\pi = \left(\frac{P}{K}\right)$ の期待値　という関係がある。

(イ) θ：多型生起モデルによって仮定されるパラメーター

mutation の発生率/site、及び genotype drift の複合結果

集団中に発生し、維持される多型の量を規定する

$$\theta = 4N_e\mu \quad \begin{pmatrix} N_e : \text{effective population size} \\ \mu : \text{mutation rate / site} \end{pmatrix}$$

Infinite site model(後述)に基づいて表現されている。

今、n 本の染色体の塩基配列を調べて、S 箇所の segregating sites(多型箇所)が検出されたとする。

θ は測定不可能なパラメーターであるが、その推定値 $\hat{\theta}$ は

$$\hat{\theta} = \frac{S}{\sum_{i=1}^{n-1}\left(\frac{1}{i}\right)} \quad \text{で得られることが、モデルの設定から知られている。}$$

n 染色体を調べたときに、singleton(後述)が観測される塩基 site 数の期待値が θ となる。

（ウ）Infinite site model では、$\hat{\pi} = \hat{\theta}$　となる

（エ）観測データから π、θ を計算してみる。(文献② p599)

i）$\pi = \frac{\sum_{i=1}^{88}(\text{Heterozygosity@多型i})}{>900} = 19.4$　多型は 88 箇所

ii) $\theta = \frac{88}{\left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{141}\right)}$

$= \frac{88}{5.53}$

$= 15.9132\cdots$

以上はモデルが観測データの背後で成立したとしたときの π と θ の値

さらにモデルが成立していれば －。

i)’ $\hat{\pi}$ =（個人が持つヘテロ塩基 sites の数の平均値）　である。

実測データから求めると　17

ii)’ $\hat{\theta}$ =（singletons の箇所数）　である。

実測データから　10

（オ）Recombination を考慮に入れる(文献① Table1,p603)

$\theta = 4N_e\mu$　と同様に

$C = 4N_e c$　というパラメーターを導入する。

μ :mutation rate

c :recombination rate

2 多型間のアレルの組み合わせの偏りは、recombination が多く起きる方が、小さくなることを利用して計算する。

$\frac{\hat{C}}{\hat{\theta}}$　を求めることでμとc の比が求まる。→　μはcの$\frac{1}{2}$～2 （？）

“c is within a factor of 2 of mutation rate”

（7） Infinite site model(無限座位モデル)

（ア） 突然変異を起こすsite の総数が非常に大きく、突然変異率は非常に低く、変異可能site 数が無限大であると近似できる。

（イ） ある site に突然変異が起こるのは 1 回だけである。

（ウ） Recombination、conversion は発生しない。

ことを仮定している。

◎モデルで仮定された推定値同士の乖離を認めたとき、モデルの破綻を疑う。

破綻の主な理由

i） C = 4Neμの Ne(effective population size)が一定でない

bottleneck や急速な人口増加など

ii） recombination/conversion の効果

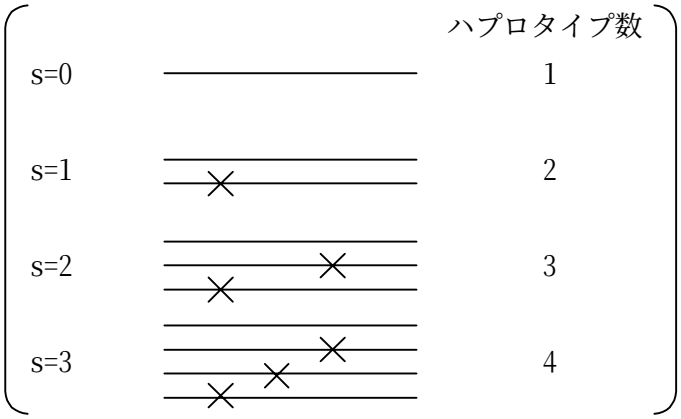
を考慮すべきである

逆に、モデル下推定値同士の乖離をもとに effective population size や recombination/conversion について検討を加えることが可能である。

（エ） Haplotype の増加パターン

多型 site 数　s ヶ所　→　haplotype 数　s＋1 種類

（recombination を許せば　→　2^s）



同じ site に 2 度目の mutation は発生しない recombination はおこらない

（8） Singleton、doubleton

複数の diploid 個体を調べているときに、稀な多型の場合、1 人だけヘテロで他全員がホモ

(Major allele のホモ)であることがある(1 人が逆ホモ、残りがホモでも可?)。
このような多型を、singleton と呼ぶ。同様に 2 人だけヘテロで全員が Major allele ホモのような多型を doubleton と呼ぶ。

（９） Haplotype 解析の実際

（ア）実験的決定

i) Haploid 化 template → serial typing(Moleculer cloning 法)

ii)Allele specific typing/amplification

iii)Pedgree analysis

（イ）算術的推定

i)EM algorithm － “3-5-2-2 参考 EM-algorithm による Hplotype 頻度の推定”

ii)その他の algorithm

iii)Clark 法(Mol.Biol.Erol 7(2),111-122(1990))

高頻度に存在する haplotype をホモで持つ個人は、相当高い確率で見つけられる

↓

このような haplotype を集めてきて、それらによって、すべての個人の haplotype 構成を説明してみる

↓

説明のつかない部分が Minor haplotype である

※文献② p597 では、Clark 法及び allele specific amplification を併用して、haplotype を推定している。

※一般論として、とりうるハプロタイプ数が十分大きい場合は Clark 法で、ハプロタイプを一部決定し、決まらないものに対して EM algorithm を適用するのが効果的である。もし、いきなり EM を適用すると、全てのとりうるハプロタイプに割当てられてしまう恐れがある。

Clark 法/EM 法併用の際、家系サンプルを用いれば、なお推定し易いであろう。(参考：Science 293,489-493(2001))

（１０） 集団間の多型分布の違いの評価(文献② p599)

（ア）ANOVA/分散分析 「論文が読める早わかり統計学」(Geoffrey R.Norman David L.Streiner) MEDSi p56 より

F：統計量=群間平均平方/郡内平均平方

95% CI の出し方

simulation-based method “Bootstrapping”

観測データから無作為抽出を繰り返して、その分布から信頼区間を得る方法

「統計モデル入門」(丹後俊郎) 朝倉書店 p20 より

（イ） permutation tests による null hypothesis の棄却の是非

「数学いらずの医科統計学」(津崎晃一) MEDSi p220 より

（１１） 参考文献①、②の図表の読み方

（ア）Nature Genetics ①

Fig3 a　実測 heterozygosity-期待 heterozygosity 差のプロット

Fig3 b　HWE ズレ検定の P 値分布

random なバラツキが観測されただけで HWE OK

Fig4　3 群別及び 3 群統合における、各多型の heterozygosity

“０”の多型は“多型ではない”

縦軸が >0.5 は、理論的におかしい

Fig5　Genotype 列の視覚的表現

（イ）Am J Hum Genet②

Table1　θ、π

Tajima's D^* ― excess heterozygosity($\hat{\pi} > \hat{\theta}$のこと)

↓

これが random な変動範囲か真にπとθに差があるためかを検証する。

(SD で割っているから、偏差値みたいなもの)

Infinite site model ではD^*の値は$\hat{D}^* = 0, SD = 1.0$の正規分布となるはず

$\hat{C}/\hat{\theta}$

Fig1 a　観測された多型のマイナーアレル頻度を高い順にプロット

Infinite site model での理論値±SE と対比 ― 適合性よい

Fig1 b　観測された haplotype の数(実測データより観測個人数(Bootstrap)をsimulate している)と Infinite site model での理論値を対比

― 明らかに実測値が大きい

↓

recombination、Gene conversion の存在が示唆される。

Fig2　横軸(多型)物理的並び順

縦軸(ハプロタイプ＋チンパンジーハプロタイプ) ― clustering 後

3 '側 半分の多型群が尾に clustering に効いている。

マイナーallele はマイナーallele 同士で並ぶべきだが、反転しているところがある。(70 番目の多型付近、80＋α番目の多型付近)

clustering 基礎の基礎

近似の度合をパラメーターを用いて“距離”に変換し、距離を図示したもの。

Fig3　“s 個の多型は s＋1 個の haplotype を形成する”という仮定からのズレを解析

範囲の全多型について tandem に評価

s=5 の場合 ― 理論上 haplotype 数　6

s=10 の場合 ― 理論上 haplotype 数　11

上に凸の領域は“何らかの原因”で haplotype 数が↑[20-30]

下に凸の領域は“何らかの原因”で haplotype 数が↓[45-60]

Fig4　A-B、A-b、a-B、a-b の 4 タイプがあれば黒 ― $D' \neq 1$、なければ白 ― $D' = 1$

Fig5　黒－LD⊕　(Fisher ’s exact probability test で P<0.01)
白＋ドット－P 不充分
Fig4-5 は白は白、黒は黒の基本的関係

Fig6　142 染色体がつくる 2 染色体ペア(142×141/2)について、何ヶ所の塩基が異なるかを数え上げ、ヒストグラム化した。
－度数総計が(142×141/2)になっていない・・・
2 峰性＝類似ペアと非類似ペアとの存在を意味する。
↓
2 群に分かれているという結論がどのくらい確からしいかを検定したい。
（文献② p605）
シミュレーションにより 2 群の染色体グループ(各群の染色体数は実測データの群分け結果に同じ)をつくる。
その 2 群について、群内シスマッチ指数と群間シスマッチ指数を計算し、その比を統計量とする。
その統計量の帰無仮説における分布が得られたので、実測データの“統計量”が帰無仮説のもとで得られる確率が得られる。

Table2　2×3Table の χ^2 検定

3-5-3 連鎖していない複数 SNP の組合せ解析

複数の、互いに連鎖していない遺伝子が相互に影響し合って疾患 susceptibility を決定しているという予想は、複合遺伝性疾患のモデルとしておそらく外れてはいないと思われる。しかしながら、そのような状況をどのようにして実際のデータに当てはめるかは未解決のままである。

Association between an SNP in the promoter of the human interleukin-3 gene and rheumatoid arthritis (RA) in Japanese patients, and maximum-likelihood estimation of combinatorial effect of two genetic loci on susceptibility to the disease.
Am J Hum Genet 68：674-685, (2001)

に提案した 2 SNP 間の関係をモデル化する試みも、ゲノムワイドスクリーニングにおいて候補遺伝子を効率よく拾い上げることを意図したものであり、まだまだ発展途上であると考えるべきである。しかし、その限界を知った上で利用する分には利用価値がないわけではないので、その限界と使用法を簡単に説明する。

プログラムの概要

1　連鎖していない 2 SNP に対して使用する。
2　2 つの SNP が相互に独立に疾患 susceptibility を上昇させている場合に、この 2 つの SNP は互いに独立であると考える(実際には 2 つの SNP は相乗効果を発揮している場合に相当する)。
3　2 つの SNP が相互に影響を与えながら疾患 susceptibility を決めているという仮説が、どのくらい確からしいかを推定する。この 2 つの SNP が相互に影響を与え合っているという状況は、2 SNP の疾患アレルの両方を持っていると、その相乗効果以上に疾患 susceptibility が上昇する場合、2 SNP の疾患アレルの片方を持っていれば、両方持っている場合と同程度に疾患 susceptibility が上昇する場合、片方の疾患アレルを持っていることがもう片方の疾患アレルの働きの前提条件となっている場合、などである。

詳細は、筆者山田まで問い合わせのこと。

4
タイピングデータの質の管理(有意差検定を用いて)

4 タイピングデータの質の管理(有意差検定を用いて)

タイピングのデータには必ず異常値が伴う。その異常値には偶然得られる場合と何らかの系統的な問題があったために必然的に生じる異常値とがある。ゲノムワイド多点解析においてはある一定の確率で異常値(＝滅多に得られるはずのないような観測数の分布)が得られるという前提がある。しかしながら全ての異常値を多点解析の結果であるとして容認することは不適切であるので、異常値をもたらす問題のタイプとそれがもたらす異常値のパターンとその検出方法を以下に説明する。

そもそも、タイピングデータを用いて関連解析をする場合の前提条件として、

1. タイピング対象の SNP が真の SNP である(SNP 自体の問題)
2. タイピングアッセイのコールが genotype を正確に反映している(アッセイに関する問題)
3. 実際の genotype の分布がサンプリング対象集団全体で均一かつ、無作為交配の原則に則っている(genotype 分布に関する問題)

の、3 つが挙げられる。この前提条件から外れたデータは系統的に異常値を出す。そしてそのような異常値はなるべく系統的異常値として検出し、関連遺伝子解析から外してやらなくてはならない。

これらを問題のタイプ別に具体例とともにまとめる。

1. SNP 自体の問題

(1) SNP ではない

① 1 つのタイプのホモのコールしか認められない

そもそも多型ではないのでコールの結果は 1 タイプのホモばかりとなる。

② ヘテロのコールしか認められない

ゲノム上に相同配列が存在するためにタイピング反応は複数のタイプに対応して進行するが、その比率に個人差はなく誰もがヘテロの genotype であるかのような結果を出す。

2. アッセイに関する問題

(1) アレル特異的反応がアッセイ対象 SNP のアレルと 1 対 1 対応していない

① テンプレート量のアンバランス

genomic DNA をテンプレートとする場合には問題ないが、PCR 産物を用いるときにはアレル依存性にその増幅効率が異なっている場合がある。そのためテンプレート量にアンバランスが生じ、その結果がコールに反映される。そのようなことが起きるのはタイピング対象 SNP そのものがその増幅反応の効率に影響している場合(可能性は低い)と、タイピング対象 SNP と連鎖している多型が PCR プライマー配列内にあるなどして、PCR 効率に影響している場合が考えられる。このような場合、ヘテロ個体が増幅効率の良いアレルのホモとしてコールされるため、Hardy-Weinberg 不平衡がホモ過剰側に偏る。

② アレル特異的反応のアンバランス

タイピング反応のテンプレート量にアンバランスはないが、タイピング反応にアレル特異的なアンバランスがある場合がある。

a) 対象 SNP のアレル依存性にタイピング反応アンバランスがある。

タイピングプローブの反応効率に著しい差がある場合には、全個体で 1 タイプのホモとなる。反応効率にかなりの差がある場合には 2 タイプのホモ個体のコールは正しくなるが、ヘテロ個体のコールが 1 タイプのホモに偏る場合がある。

b) 対象 SNP の近傍配列依存性にタイピング反応アンバランスがある

対象 SNP 近傍に多型があり、対象 SNP と連鎖している場合に近傍多形の影響でタイピング反応の効率が影響を受ける場合がある。この場合 Hardy-Weinberg 平衡からのずれはホモ過剰の方向である。その偏りの程度は対象 SNP と近傍多型とが作る 4 ハプロタイプの頻度によって決まる。

③ アレル数の設定ミス(3 アレル SNP を 2 アレルとしてアッセイする、など)

考慮に入れていないアレルが存在する場合にはそのアレルと考慮済みのアレルとの個体のコールは考慮済みアレルのホモとなるために、ホモ過剰方向の Hardy-Weinberg 不平衡が観測される。

(2) ミスコール

① 反応不十分による困難コール

テンプレートにもタイピングプローブにもアンバランスはないが、タイピング反応自体が不十分であると 3 genotypes の区別がつきにくくなりコールミスが生じる。このコールミスに一定の傾向が生じ、ある genotype が増える方向もしくは減る傾向にある場合、Hardy-Weinberg 不平衡が観測される。

② 単純な人的ミス

対象 SNP の取り違えたり、アレルプローブの対応を取り違えたりすることがある。

このような場合には、Hardy-Weinberg 不平衡が非常に強く出現する。

3. genotype 分布に関する問題

(1) 対象集団の均一性・無作為交配が守られていない

① Hardy-Weinberg 不平衡（詳細は“3-4-1 Hardy-Weinberg 平衡検定”参照）

Hardy-Weinberg 平衡からの著しい偏りはコールの信用性を落とすために、そもそもデータを検定に用いることを躊躇させるものである。しかしながらアッセイおよびコールに問題がなく Hardy-Weinberg 不平衡が観察された場合に考慮しなくてはならないのは、疾患 phenotype との関連があるために生じた Hardy-Weinberg 不平衡と、集団の無作為交配が行われていないために生じた Hardy-Weinberg 不平衡である。複合遺伝子疾患のような個々のローカスの遺伝的寄与が大きくない解析において強度の Hardy-Weinberg 不平衡が疾患との関連により認められることはあり得ないので、そのような SNP においては集団内の無作為交配に問題があると結論づけられる。つまり、サンプリング対象とする集団として不適切であるとみなされる。

② 集団の遺伝的背景の差(詳細は該当項目参照)

複数のサンプリング集団の遺伝的背景が異なると疾患と関連のない多数の SNP が擬陽性として検出される。従って集団間の遺伝的背景の差を検定して、関連解析に用いることが適切かどうかの判断を下す必要がある。

今まで述べてきた問題の検出方法は個々の SNP データについて検討することで判断を下してきたが、集団の遺伝的背景の差の有無に関して判断を下すためには、多数の SNP の関連解析の結果を統合する必要がある。

これらの異常は複数の疾患群間の分割表分析(χ^2検定)と Hardy-Weinberg 平衡検定とでチェックされる。そのうち、Hardy-Weinberg 平衡検定は genotype 観測値分布の異常を調べているので、特にその有用性が高く上記の問題のほとんど全てを網羅している。一方、χ^2検定での外れ値は、極度の異常の場合には SNP の取り違えやプローブの取り違えといった単純な人的ミスであることがほとんどである。

平成 13 年 3 月の時点で、χ^2検定、Hardy-Weinberg 平衡検定とも、異常の可能性のあるレベルは p 値として 0.01、明らかに異常であるとみなしたほうが良いレベルとして 0.00001 を設定して対処することとしている。

問題のタイプ	具体例
1.SNP 自体の問題	(1) SNP ではない ① 1 つのタイプのホモのコールしか認められない ② ヘテロのコールしか認められない ゲノム上に相同配列が存在する
2.アッセイに関する問題	(1) アレル特異的反応がアッセイ対象 SNP のアレルと 1 対 1 対応していない ① テンプレート量のアンバランス ② アレル特異的反応のアンバランス a)近傍に多型がない場合 b)近傍に多型がある場合 ③ アレル数の設定ミス (3 アレル SNP を 2 アレルとしてアッセイする、など) (2) ミスコール ① 反応不十分による困難コール ② 単純な人的ミス
3.genotype 分布に関する問題	(1) 対象集団の均一性・無作為交配が守られていない ① Hardy-Weinberg 不平衡 ② 集団の遺伝的背景の差