様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　4 月　6 日現在

研究種目：基盤研究(B)
研究期間：2007 ～ 2008
課題番号：19300095
研究課題名（和文）　疾患関連遺伝子発見のための症例対照研究における統計学的問題とその解決策の検討
研究課題名（英文）　Statistical genetics for designing multistage genome-wide association studies to detect genetic risk factors

研究代表者：赤澤　宏平（AKAZAWA KOHEI）
新潟大学・医歯学総合病院・教授
研究者番号：10175771

研究成果の概要：本研究では、SNPs データを用いて疾患関連遺伝子を同定するための多段階デザイン症例-対照相関解析の統計学的性質を詳細に検討した。研究成果として以下が挙げられる。
１．R 言語を使い、多段階デザイン症例－対照研究の検出力を推定するプログラムを開発した。
２．3 段階法において、各ステージの症例割合を種々変化させたときの検出力とタイピング数を調べた．　これらの結果は、国際学術誌 Bioinformatics、Journal of Human Genetics に掲載された。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	4,900,000	1,470,000	6,370,000
2008 年度	4,100,000	1,230,000	5,330,000
年度			
年度			
年度			
総　計	9,000,000	2,700,000	11,700,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：医薬生物統計、疾患感受性遺伝子

１．研究開始当初の背景

1 塩基多型 (single nucleotide polymorphism, SNP) やマイクロサテライトマーカー等の DNA 多型データを用いた case-control 相関解析は、疾患感受性遺伝子を見出すツールとして頻繁に用いられている。たとえば、第 1 染色体上の transcription factor 7-like 2 (TCF7L2) 遺伝子上のイントロン 3 内に存在するマイクロサテライトマーカー・DG10S478 が、アイスランド人、デンマーク人、アメリカ人の 2 型糖尿病と相関することを発見したのも case-control 相関解析によるものである。さらに、DG10S478 のリスクアリルをヘテロ、ホモで持つ人の相対リスクは、持たない人のそれぞれ 1.45 倍、2.41 倍であり、また、2 型糖尿病全体の 21%に DG10S478 のリスクアリルが寄与していることが分かった。
SNP ベースの case-control 相関解析は、患者群と健常群とを用い、どの SNP が患者群と有意に関連しているのかを遺伝統計学的に調べ、最終的には有意な SNP の近傍に位置する疾患感受性遺伝子を同定する方法である。
SNP 解析法の向上に伴い、現在では、1 症例につき数千個から数十万個の SNP データが迅速に採取できるようになった。仮に、100,000

個の SNP 解析を 1,000 例（疾患群 500 例、対照群 500 例）で行なうとすると、研究全体で 1 億 SNP のデータ採取が必要となる。SNP 解析法が向上したとはいえ、10 万 SNP あたりにかかるコストは約 16 万円であり 1 億 SNP のデータ収集には１億 6 千万円のお金がかかることになる。疾患の病因の解明、診断精度の向上、治療方法の開発という大儀はあるにせよ、途方もないお金をつぎ込まない限り SNP 解析はできないことになる。
そこで、遺伝統計学の分野では、より少ない検体数で、最小の False negative rate を保ちつつ疾患関連遺伝子を検出する方法が議論されてきた。Two-stage association study もそのひとつで、コストから逆算される限られた症例数を First stage と Second stage に分け、First stage である程度有意な SNP をスクリーニングした後、Second stage で感受性遺伝子かどうかの確証的な検定を行なう。たとえば、利用できる全症例数が 1,000 例（疾患群 500 例、健常群 500 例）あり、First stage で 500 例（疾患群 250 例、健常群 250 例）を使い有意水準 0.1 の検定を行なうと、genotyping の総件数は one-stage association study の全数調査の 55%程度ですむことになる。しかしながら、two-stage association study は全症例数を分割して検定を行なうので、それぞれの検出力は低下することになり、真の疾患感受性遺伝子を取りこぼしてしまう確率も増加する。
分子生物学の専門誌、Nature Genetics, Human Molecular Genetics, American Journal of Human Genetics などの解析プロセスを調べてみると、各論文で異なる Two-stage association study design と検定処理が行なわれている。具体的には以下の相違点があり、これまでの研究では当該研究者の恣意によりこれらのパラメータは決められてきた。

（１）Stage1、Stage2 それぞれにおいて、SNP を候補因子として抽出するための有意水準の取り方が研究により異なる： Stage1 で厳しくし Stage2 でゆるめにするのか、Stage1 でできるだけ多くの候補因子を拾い上げ、Stage2 で候補因子を厳しく絞り込むのかの統計学的な評価がなされていない。

（２）Stage1、Stage2 に割り当てる症例数の割合がそれぞれ異なる： 研究全体で利用できる症例数は研究費に依存して決まるが、それを Stage1、Stage2 にどのように割り当てると最適かはよく知られていない。

（３）症例の収集方法がまちまちである：研究組織内で収集できた症例を 2 分割する研究もあれば、他の研究グループの症例（生データ、もしくは必要とする統計量）を利用することもある。
このような解析手法の選択や解析に用いるパラメータが各研究で異なることが、研究の再現性や信頼性の評価を難しくしており、疾患関連遺伝子の発見を遅らせる原因ともなっている。

２．研究の目的

Two-stage association study は、より少ない検体数で One-stage association study（Bonferroni の有意水準補正による疾患関連遺伝子の検出）での疾患関連遺伝子の検出と同程度の検出を可能にする方法とされている。最近では、Nature genetics, Human Molecular Genetics, American Journal of Human Genetics などの専門誌に、応用例、研究デザインの開発、それらの統計学的性質が掲載されている。
本研究では、発表者らが経験したアルツハイマー病の SNP 解析データを用いた疾患感受性遺伝子探索の過程から症例対照研究の問題点を抽出し、その解決策を考察する。特に、Two-stage association study での検定方法として、従来の方法(Replication-based analysis)と 2006 年６月に提唱された Joint analysis を取り上げ、症例数の配分割合、有意水準、対立遺伝子の割合、genotype の発生頻度モデル等を変化させたときの検出力を検討する。
主な研究目的を以下に列挙する。

（１）Two-stage design における解析方法の性能比較
① Replication-based analysis 、 Joint analysis、logistic analysis などの解析方法の検出力、サイズを解析的に求め比較する。その際のパラメータとして、症例群・対照群のサンプル数の割合、各 stage に割り付けられた症例数の割合、それぞれの stage での有意水準を用いる。
②疾患群、対照群の遺伝子型(AA, Aa, aa など)が人種や地域の違いでゆらいだとき、検出力の挙動をモンテカルロシミュレーションで推定する。
③アルツハイマー病の SNP データを使い、Replication-based analysis や Joint analysis、logistic analysis などの検出力を実際に計算する。

（２）健常群に疾患例が混入したときの解析結果への影響
①健常群への疾患例の混入は、アウトカムに計測誤差を含むデータととらえることができる。このような計測誤差が、症例-対照研

究での疾患関連遺伝子検索の検出力に与える影響をシミュレーションにより解明する。②先天的要因（SNP 情報)、後天的要因（生活習慣、食べ物、環境など）を含むロジスティック回帰分析での回帰係数の推定値は、計測誤差を含んでいるときに誤った推定値を算出する。本研究では、アウトカムに計測誤差を含む場合の回帰係数推定のバイアスを、特別な尤度関数を用いて推定する。

（３）研究対象 SNP の選択方法
SNP による疾患関連遺伝子の探索を行う際に、候補 SNP の選定は最も重要な作業となる。研究対象 SNP の最適な選定方法を、疾患群・対照群のハプロタイプの選定にベイズ理論を用いた手法により新規に開発する。

３．研究の方法

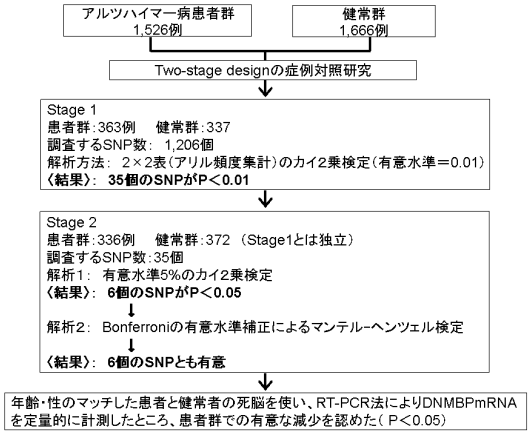
１．疾患関連遺伝子探索のための SNP を用いた症例対照研究の問題点の発掘

（１）文献調査に基づく問題点の発掘
①疾患関連遺伝子発見を目的とした１塩基多型（single nucleotide polymorphism, SNP）を用いた症例対照研究の論文を、Nature Genetics, Human Molecular Genetics, American Journal of Human Genetics などの学術専門誌から 100 編程度選択する。各論文を精査して統計学的な問題点を探り出す。
・研究デザイン(One stage design、Two-stage design、その他)
・Multi-stage design の場合のパラメータ(症例群・対照群のサンプル数の割合、各 stage に割り付けられた症例数の割合、それぞれの stage での有意水準)
・統計手法（カイ２乗検定、ロジスティック解析、その他）
・有意水準の補正方法
・サンプル数
・解析結果（発見遺伝子とその再現性、オッズ比等）
②Biometrika, Biometrics, Statistics in Medicine, Journal of American Statistical Association 等のバイオ統計学の専門雑誌から、SNP 解析の症例対照研究の問題点とその解決策を扱った論文を２０編程度選択し抄読する。これまでの解決方法、特に、Two-stage design の統計学的性質､解析方法、連鎖不平衡がある場合の有意水準の補正方法、などをレヴューする。

（２）アルツハイマー病の DNA データに基づく問題点の発掘
　分担研究者の桑野・宮下らのミレニアム研究プロジェクト（特定領域研究　平成１２～16 年度)、および、ポストミレニアム研究（特定領域研究　平成１７～１８年度）で収集された５，８００例の DNA データを再解析することにより、SNP による症例対照研究の実用上の問題点を発掘する。アルツハイマー病関連遺伝子探索の手順等は Human Molecular Genetics, 2006, Vol.15(13), 2170-82（Kuwano R, Miyashita A, Toyabe S, Akazawa K, et.al.)に記述されている。桑野論文の研究目的は､APOE-ε3*3 遺伝子型を有する LOAD の遺伝的なリスク因子を同定する、即ち、従来からアルツハイマー病のリスク遺伝子とされている APOE-ε4 遺伝子型以外の遺伝子を探索することである。アルツハイマー病の候補染色体 10q の 60-107Mb 領域(NCBI build 35.1）を解析した。本基盤研究(B)での解析計画を説明する際に必要となるので桑野論文の概要を次頁に図示する（図１)。

桑野論文での一連の統計解析から明らかとなった問題点
①桑野論文では Two-stage design データの解析を、査読者の修正要求に従い独自の解析方法により行なった。一方、他の論文でも２つの Stage の検定統計量のまとめ方として Replication-based analysis や Joint analysis など多種多様な方法がとられている。即ち、国際的に見て、標準的な解析方法が確立されていない。

②桑野論文での有意水準の補正は適切とは言えない。染色体全体での有意水準を 0.05、Stage1 における有意水準を 0.01 として 1,206SNP を検定する場合、Stage2 での有意水準は Bonferroni の補正を用いるならば 0.00414 としなければならない。この有意水準で解析すると､Stage2 で有意となる SNP は 1 個となる。結果的には桑野論文で発見された遺伝子が抽出された。



③Bonferroni の補正は検定する対象が独立であることを仮定しているが、SNP の場合は連鎖不平衡があり SNP 間は独立ではない。実

際に、アルツハイマー病の疾患関連遺伝子は4〜7 個程度とされ、有意水準１％の Stage1 解析で有意となる期待度数は、偽陽性 SNP と合わせても高々20 個程度であると推測される。ところが、実データの解析では 35 個の SNP が有意となった。このことは、ある DNA 領域の複数の SNP が同時に有意差ありとされた可能性を示唆している。従って、①の Stage2 の有意水準の補正は正しいとは言えず、SNP 解析にあった有意水準の補正方法を考案すべきである。

④Stage１で抽出された 35 個の SNP について、Stage1 と Stage2 のオッズ比を比較した。本来であれば、オッズ比の符号は２つの Stage で一致していなければならない。ところが正負逆転している SNP が 12 個もあることがわかった。この再現性の不確実性については統計学的に見ていくつかの原因があるものと考えられる。分担研究者で分子生物学者の桑野・宮下の話では、優性のアリル頻度が 0.5 に近い場合には、このような報告例が Am. J. Hum. Genet.などで報告されているという。このような理由で文献的な調査が必要であると判断した。

４．研究成果

1 塩基多型（single nucleotide polymorphism, SNP）やマイクロサテライトマーカー等の DNA 多型データを用いた症例-対照関連解析は、疾患感受性遺伝子を見出す解析法として頻繁に用いられている。症例-対照関連解析の中でも疾患感受性遺伝子を効率的に検出するための研究デザインとして、多段階関連解析法があり、その代表的な手法として Replication-based analysis（RBA）と Joint analysis（JA）のふたつが知られている。多段階関連解析で用いられる研究デザインの多くは 2 段階法であるが、最近、極めて多数の候補遺伝子を対象として、3 段階法を用いた疾患感受性遺伝子同定の研究が報告され始めた。しかし 3 段階法による疾患感受性遺伝子の統計学的検出力、陽性反応適中度（PPV）およびタイピング数に関する研究は殆ど行われていない。
本研究の目的は、多段階関連解析法の特性を明らかにすることである。具体的には、実際の SNPs 解析で起こりうるさまざまな条件(例えば、症例数、候補アリル数、実験全体の有意水準、各 k 段階(k は１以上の整数)への症例数の配分割合(πs,k)、各 k 段階での候補アリルの選択割合(πm,k)、等）の下で、RBA と JA の統計学的検出力と PPV ならびにタイピング数の特性を調べた。本研究目的のため、著者はプログラム言語として Mathematica ならびに R を用いて、RBA ならびに JA による検出力と PPV ならびにタイピング数を算出するプログラムを開発した。
本研究から得られた結果は以下のとおりである。
3 段階法において、各ステージの症例割合を種々変化させたときの検出力とタイピング数を調べた。その結果、いずれの場合においても JA が RBA よりも高い検出力を示し、例えばπs,1＝0.4、πs,2＝πs,3＝0.3 の時、1 段階法では、85.2%の検出力であるのに対して、3 段階法の RBA、JA の検出力はそれぞれ 49.8%、72.3%であり、JA の方が検出力の減少率は小さいことがわかった。その時のタイピング数は 1 段階法が 1,000,000,000 回であるのに対して、3 段階法では 403,030,000 回であり、約 60%の削減が可能であった。
さらに最終ステージに残るアリルの個数が等しくなるように 2 段階法では、πm,1＝0.0001、3 段階法では、πm,1＝0.01、πm,2＝0.01（すなわちπm,1×πm,2＝0.0001 となる）と設定して 2 段階法と 3 段階法とを比較した。その結果、例えばπs,1＝0.2、πs,2＝πs,3＝0.4 のとき、2 段階法の RBA と JA の検出力はそれぞれ 0.148、0.149、3 段階法の RBA と JA の検出力はそれぞれ 0.466、0.548 となり、同様に 0.1≦πs,1≦0.5 において、RBA ならびに JA による 3 段階法の検出力ならびに PPV は 2 段階法よりも高い値を示す傾向が認められた。特に、πs,2 とπs,3 が同じ程、また RBA よりも JA の方でその傾向が強くみられた。一方、0.6≦πs,1≦0.9 の場合は、RBA ならびに JA による検出力と PPV は、3 段階法よりも 2 段階法の方が常に高い値を示していた。
また、タイピング数を一定に設定した条件下において、各多段階デザインによる検出力ならびに PPV を比較したところ、3 段階法と 2 段階法との検出力の差が小さくなっていたものの、最終段階で選択されるアリルの数を一定とした条件の場合と同様、πs,1 が小さい範囲においては、RBA ならびに JA による 3 段階法の検出力ならびに PPV は 2 段階法よりも高い値を示す傾向が認められた。例えばπs,1＝0.2、πs,2＝0.3、πs,3＝0.5 のとき、2 段階法の RBA と JA の検出力はそれぞれ 0.459、0.465、3 段階法の RBA と JA の検出力はそれぞれ 0.460、0.490 であった。
さらに、症例数を増加させることにより各多段階デザインの検出力ならびにタイピング数も上昇したが、各デザインにおける上述のような変化のパターンの傾向は変わらなかった。なお JA において、各症例数毎に多段階法の検出力とタイピング数を比較すると、いずれの症例数においてもπs,1＝0.3 における 3 段階法の検出力は、πs,1＝0.5 における 2 段階法の検出力とほぼ同程度の検出力を

示していたが、前者のタイピング数は後者のタイピング数の約 60%に抑えられていた。本研究により、2 段階法だけでなく、0.1≦πs,1≦0.5 において JA を用いた場合には 3 段階法の有用性も示唆された。逆に 0.6≦πs,1≦0.9 の場合は、RBA ならびに JA による検出力ならびに PPV は 3 段階法よりも 2 段階法の方が常に高い値を示していたことから、3 段階法は用いるべきではないと考えられた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

1. Kitamura N, Akazawa K, Miyashita A, Kuwano R, Toyabe SI, Nakamura J, Nakamura N, Sato T, Hoque MA:
Programs for calculating the statistical powers of detecting susceptibility genes in case-control studies based on multistage designs. (査読・有)
Bioinformatics, 25(2):272-273, 2009.

2. Toyabe S, Miyashita A, Kitamura N, Kuwano R, Akazawa K:
Prediction of Disease-associated Single Nucleotide Polymorphisms Using Virtual Genomes Constructed from a Public Haplotype Database. (査読・有)
Methods Inf Med. 47(6):522-528, 2008.

3. Kitamura N, Akazawa K, Toyabe SI, Miyashita A, Kuwano R, Nakamura J:
Sample-size properties of a case-control association analysis of multistage SNP studies for identifying disease susceptibility genes. (査読・有)
J Hum Genet 53(5):390-400, 2008.

4. Miyashita A, Arai H, Asada T, Imagawa M, Matsubara E, Shoji M, Higuchi S, Urakami K, Kakita A, Takahashi H, Toyabe S, Akazawa K, Kanazawa I, Ihara Y, Kuwano R:
Genetic association of CTNNA3 with late-onset Alzheimer's disease in females. (査読・有)
Human Molecular Genetics 16(23):2854-69, 2007.

〔学会発表〕（計１件）
1. 北村信隆、SNP データに基づく疾患感受性遺伝子同定のための多段階症例-対照研究デザインの特性、2008 年度統計関連学会連合大会、2008 年 9 月 9 日、慶應義塾大学

６．研究組織

(1)研究代表者
赤澤　宏平（AKAZAWA KOHEI）
新潟大学・医歯学総合病院・教授
研究者番号：10175771

(2)研究分担者
鳥谷部　真一（TOYABE SHIN-ICHI）
新潟大学・危機管理室・教授
研究者番号：20227648

桑野　良三（KUWANO RYOZO）
新潟大学・脳研究所・教授
研究者番号：20111734

宮下　哲典（MIYASHITA AKINORI）
新潟大学・脳研究所・助教
研究者番号：60323995

(3)連携研究者
（　　）

研究者番号：