

46. タンパク質の三次構造を考慮した新規解析手法の開発と疾患原因遺伝子の同定

東京大学大学院医学系研究科国際保健学専攻人類遺伝学分野 教授 藤本 明洋

概要

ヒトの遺伝的多様性は疾患のリスクに影響することが知られている。疾患のメカニズムやリスク予測を行うため、ゲノム全域を対象として遺伝的多様性（バリエーション）と疾患の関連を調査するゲノムワイド関連解析（GWAS; Genome-Wide Association Study）が広く行われてきた。しかし、GWAS では遺伝率（疾患リスクの遺伝が説明する割合）の一部しか説明できないことが大きな問題となっている（Missing Heritability 問題）。このことは、疾患に関連する遺伝的多様性の多くが GWAS では見逃されていることを強く示唆している。

GWAS では主に頻度が高い遺伝的多様性（頻度 > 1% のコモンバリエーション）が解析されている。そのため、レアバリエーション（頻度 < 1%）を見逃している可能性が考えられ、レアバリエーションに注目した解析が行われてきた。代表研究者らは、タンパク質の立体構造を考慮して複数のレアバリエーションをまとめ疾患との関連を発見する新たな手法（3D Burden test）を考案した。この方法を用いて、世界最大のゲノムコホートである UKBB（UK biobank）のデータを用いて疾患に関連する多型を解析した。363 の疾患を含む表現型を解析したところ、424 個の遺伝子において有意な変異の集積が観察され、多くが新規発見であった。

背景および目的

ヒトの遺伝的多様性は疾患のリスクに影響することが知られている。疾患のメカニズムやリスク予測を行うため、ゲノム全域を対象として遺伝的多様性（バリエーション）と疾患の関連を調査するゲノムワイド関連解析（GWAS; Genome-Wide Association Study）が広く行われてきた。これまでに 5,000 を超える GWAS が行われ、70,000 を超える表現型とバリエーションとの関連が報告されている(1)。これらの研究は新たな疾患メカニズムの解明や疾患リスクの予測に貢献している。しかしながら、GWAS では遺伝率（疾患リスクの遺伝が説明する割合）の一部しか説明できないことが大きな問題となっている（Missing Heritability 問題）(1)。このことは、疾患に関連する遺伝的多様性の多くが GWAS では見逃されていることを強く示唆している。

GWAS では主に頻度が高い遺伝的多様性（頻度 > 1% のコモンバリエーション）が解析されている。そのため、レアバリエーション（頻度 < 1%）を見逃している可能性が考えられ、レアバリエーションに注目した解析が行われてきた(2)。レアバリエーションは頻度が低く、単一のバリエーションの解析では統計的な検出力が弱い。そのため、遺伝子内の複数のレアバリエーションをまとめて疾患との関連を解析する方法が用いられてきた（Gene Burden テストなど）(2)。しかし、遺伝子内には、異なる機能を持つドメインが存在することや、疾患に関連するバリエーションとそうでないバリエーションが混在していると予測され、このような単純な手法では、検出力が十分ではない可能性がある。

代表研究者は、タンパク質の立体構造を考慮して複数のレアバリエーションをまとめ疾患との関連を発見する新たな手法（3D Burden test）を考案した。この方法を用いて、世界最大のゲノムコホートである UKB（UK biobank）のデータを用いて疾患に関連する多型を解析した。

方法

遺伝的多様性、表現型およびタンパク質の3次構造

本研究では、世界最大のゲノムコホートである UKB(UK biobank)のデータを用いて疾患に関連する多型を解析した。UK Biobank(UKB)は、募集時の年齢が40歳から69歳までの約50万人の参加者を対象とした大規模な生物医学データベースである。参加者は2006年から2010年にかけて英国で募集され、継続的に追跡調査されている。参加者の募集時の平均年齢は56.5歳で、54%が女性である。参加者のデータには、環境、ライフスタイル、遺伝子に関する広範なデータが含まれている(3)。UKBでは、表現型(身長、体重、食習慣、検査値など)や疾患(がん、生活習慣病など)の情報を含む18,395種類のデータ項目と454,756例のゲノムデータが利用可能である。

遺伝学的背景が判明している501,495人のUKB参加者のうち、472,124人が英国人であった。今回は、英国人(Data-Field 21000(Ethnic background)#1001(British))かつ他の参加者との遺伝的近親性がない(Data-Field 22021(Genetic kinship to other participants)#0(No kinship found))292,176人のエクソーム配列データを使用した。今回は一塩基多様体のみに着目し、挿入や欠失は解析から除外した。また、今回はアミノ酸の違いを引き起こす一塩基多様体であるミスセンス SNV(single nucleotide variant)のみを抽出して解析に使用した。

AlphaFold2(4)で予測されたタンパク質の三次元構造データを用い、三次元構造データが提供されている18,078遺伝子に対して解析を行った。

解析手法

3D Burden test は以下の手順でアミノ酸置換を引き起こすレアバリエントと疾患の関連を検出した(図1)。

- (1) タンパク質の三次構造において各アミノ酸残基を中心とする半径nÅの球をwindow(注目する領域(黒丸内))として定義し、球の中心を一次元のアミノ酸配列に沿って動かし、立体構造上でwindowを移動させ、立体構造を考慮した領域分割を行った。
- (2) 各window内に存在するアミノ酸置換に注目し、疾患患者と対照者(非患者)をレアバリエントを持つ個体と持たない個体に部類し、2X2の分割表を作成し、フィッシャーの正確確率検定で割合の差を検定した。
- (3) レアバリエントを有する個体が非常に少ない(全患者数の0.5%未満)windowを除き、検定に用いられたwindow数で多重検定補正(ボンフェローニ法)を行い、疾患に関連するwindowと遺伝子を同定した。

3D Burden test(多発性嚢胞腎のTBL3遺伝子の例)

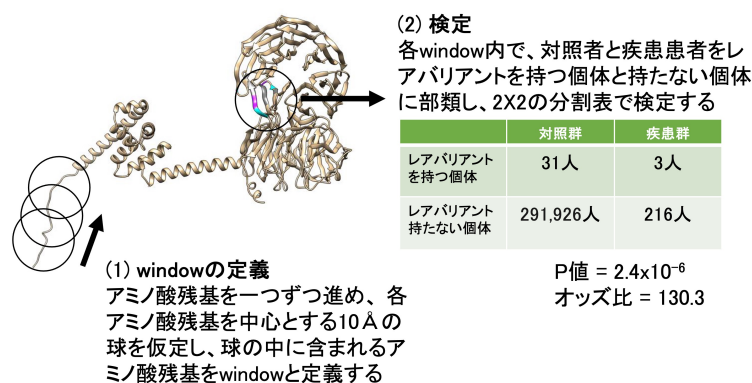


図1. 3D Burden test の方法

タンパク質立体構造を考慮してwindowを定義し、window内のレアバリエントを検定する。立体構造上で近いバリエントは(基質との結合のような)同じ機能に関連している可能性が高く、立体構造を考慮することでレアバリエントを適切にまとめることが可能になると考えられる。

結果および考察

統計解析の結果の信頼性を評価した。各形質で、p 値が一様分布すると仮定した場合の p 値の中央値と実際の p 値の中央値の比(lamdaGC)を求めた。ゲノム解析では、多くの遺伝子は形質に関係しないと考えられるため、lamdaGC は 1 に近くなることが予想される。本解析では、症例数(case のサンプル数)が少ない形質($n < 50$)において lamdaGC が高く、症例数が結果の信頼性に大きく影響することが示唆された。

UKB のデータを用いて先行研究(2)で重要な遺伝子が検出された多発性嚢胞腎、湿疹/皮膚炎を対象として 3D Burden test を実施した (図 2)。先行研究では、遺伝子ごとにレアバリエントをまとめた解析 (Gene Burden テスト) が行われ、多発性嚢胞腎では RDK1 と PDK2 遺伝子が、湿疹/皮膚炎では FLG 遺伝子が疾患遺伝子として同定されていた(2)。3D Burden test では、多発性嚢胞腎では RDK1 と PDK2 遺伝子に加えて ADRB3 遺伝子が同定された (図 2ab)。湿疹/皮膚炎では、HRNR 遺伝子が同定された。ADRB3 と HRNR は先行研究で当該疾患に関係することが報告(5,6)されており、3D Burden test が、UKB の解析が見逃している新たな遺伝子を同定することができる有用な方法であることが示唆された。

引き続き、363 の疾患を含む表現型を解析したところ 424 個の遺伝子において有意な変異の集積が観察された。症例数が 50 以上の形質に絞ると 349 遺伝子に有意な関連が観察された。これらの遺伝子の詳細な機能は現在解析中であるが、多数の興味深い遺伝子が発見されている。例えば、高コレステロールと関連する遺伝子として ABCG5、APOB、FGB、LDLR、PCSK9 の 5 遺伝子が同定された。このうち、PCSK9 は先行研究においてコレステロール関連遺伝子として同定され、創薬につながった遺伝子である。LDLR と APOB は UKB の研究で関連が発見されている。FGB と ABCG5 については、UKB の先行研究では関連が見つからなかったが、ABCG5 に関しては先行研究で高コレステロールと関連することが報告されており(7)、本手法の信頼性が高いことが示唆された。

以上より、3D Burden test が有用な方法であり新規疾患関連遺伝子の発見が可能になることが強く示唆された。しかし、以下の課題もある。lamdaGC を用いた解析から、サンプル数が少ない場合に信頼性が低下することが示唆された。また、遺伝子の発見した後に変異の集積の意義を解析する必要があるため、ドメイン予測などと組み合わせる解析が必要であると考えられる。さらに、細胞等を用いた機能解析も必要であると考えられる。

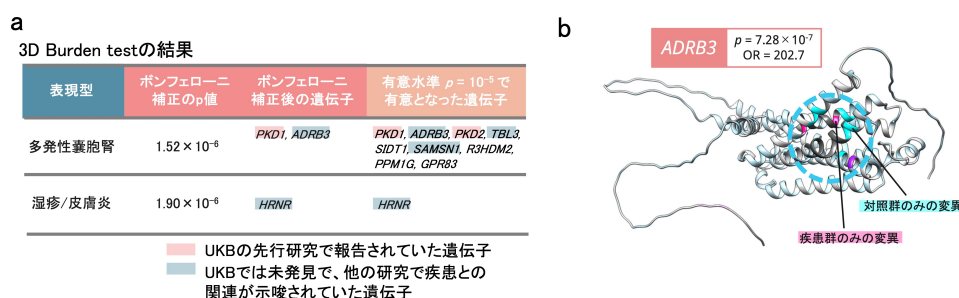


図 2. 3D Burden test の結果

a UKB による先行研究で解析されていた多発性嚢胞腎と湿疹/皮膚炎のサンプルを 3D Burden test を用いて再解析した結果

b 多発性嚢胞腎において、3D Burden test で有意な結果を得た遺伝子の例。疾患群において著しく変異が集積している window が検出された。

(完)

発表論文

- 1) Worakitchanon et al. Comprehensive analysis of *Mycobacterium tuberculosis* genomes reveals genetic variations in bacterial virulence. *Cell Host & Microbe*, Volume 32, Issue 11, 1972 - 1987.e6 (2024)
- 2) Yoshida-Tanaka, K., Ikemoto, K., Kuribayashi, R. et al. Long-read sequencing reveals the complex structure of extra dic(21;21) chromosome and its biological effects. *Hum. Genet.* 142, 1375–1384 (2023).
- 3) Ikemoto, K., Fujimoto, H. & Fujimoto, A. Localized assembly for long reads enables genome-wide analysis of repetitive regions at single-base resolution in human genomes. *Hum Genomics* 17, 21 (2023).

引用文献

- 1) Tam, V., Patel, N., Turcotte, M. et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 20, 467–484 (2019).
- 2) Wang, Q., Dhindsa, R.S., Carss, K. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597, 527–532 (2021).
- 3) Conroy, M.C., Lacey, B., Bešević, J. et al. UK Biobank: a globally important resource for cancer research. *Br J Cancer* 128, 519–527 (2023).
- 4) Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
- 5) Procino, G., Carmosino, M., Milano, S., Monte, D. M., Schena, G., Mastrodonato, M., Gerbino, A., Bagnoli, P., Svelto, M. β 3 adrenergic receptor in the kidney may be a new player in sympathetic regulation of renal function. *Kidney International*, 90(3), 555–567 (2016).
- 6) Smieszek, S.P., Welsh, S., Xiao, C., Wang, J., Polymeropoulos, C., Birznieks, G., & Polymeropoulos, M. H. Correlation of age-of-onset of Atopic Dermatitis with Filaggrin loss-of-function variant status. *Scientific Reports*, 10(2721). (2020).
- 7) Lee, MH., Lu, K., Hazard, S. et al. Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nat Genet* 27, 79–83 (2001).