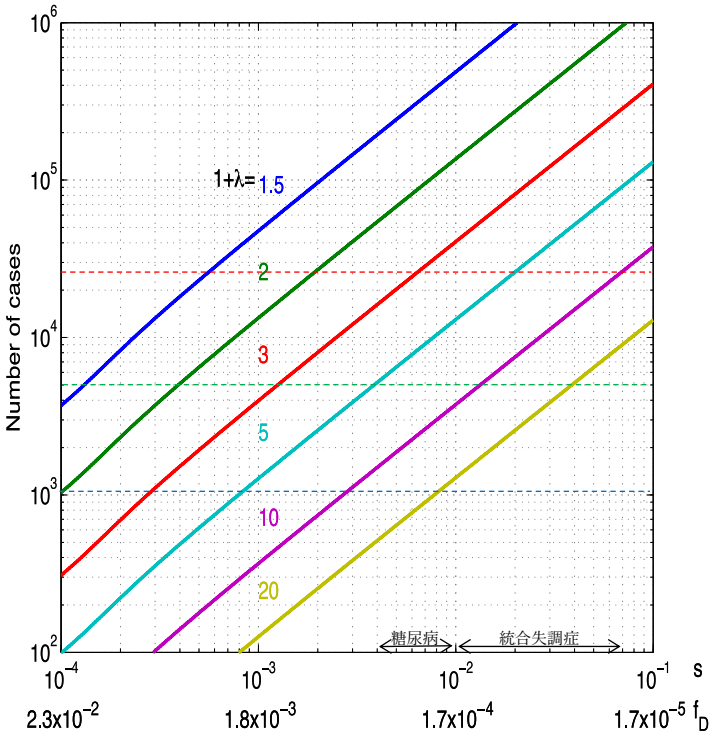
資料 2

レアバリアント関連解析の検出力について　2019 年 11 月 18 日

東京大学大学院新領域創成科学研究科　複雑形質ゲノム解析分野　鎌谷洋一郎

突然変異と自然選択を考慮した集団遺伝学モデルを用いた Zuk らの論文 [1] によると、次世代シーケンサー（NGS）を用いた解析により複雑性疾患（多因子疾患）を引き起こすレアバリアントをレアバリアント関連検定（RVAS、特に burden test[1]）により検出するためには、症例の人数として 25,000 人が必要であると報告されている。この人数は、(1) 概ねの疾患で 10 倍以上、(2) 累積アレル頻度が1.7×10^{-4}前後より大きいなら 5 倍以上、(3) 累積アレル頻度が1.8×10^{-3}前後より大きいなら 2 倍以上、などの相対リスクを示す遺伝子を検出力 90%で検出するサンプルサイズである。

その根拠は下図（論文 supplementary Fig 9）で、横の赤点線が 25000 人を示す。横点線より下にある斜め実線のリスク効果の遺伝子は、その遺伝子の累積アレル頻度（f_D）において検出可能である。確かに $f_D > 1.8 \times 10^{-3}$ の範囲において、リスク比 2, 3, 5, 10, 20 の直線は、25000 人を示す赤の横点線の下にあることを確認できる。



Zuk らの論文によれば集団遺伝学パラメータsは統合失調症で$s = 10^{-1.3} \sim 10^{-2}$、2 型糖尿病で$10^{-2} \sim 10^{-2.7}$とあり、それに対応した累積アレル頻度$f_D$を見ることで必要サンプルサイズを検討できる。

左図より、5000 人（緑点線）であっても、$f_D = 1.7 \times 10^{-4}$だとすると 10 倍以上のリスクの遺伝子を検出できる。

また、1000 人（青点線）の場合、$f_D = 1.7 \times 10^{-4}$だとすると 20 倍のリスクの遺伝子ならば検出力 90%で検出可能である。

すでに心筋梗塞について LDLR 遺伝子のリスクは 18.1 倍であると報告されており、また 2019 年のアメリカ人類遺伝学会においては、統合失調症に対して 20 倍以上のリスク効果を持つバリアントが報告されている。20 倍のリスクという数値は想像上の数字ではない。

---

[1] RVAS は主に遺伝子レベルでレアバリアントを集約的に解析するもので、burdern test、adaptive burden test、variance-component test、combined test、EC test に分類できる [3]。これらによる検出力を推定するためには遺伝子レベルの集約的なアレル頻度（累積アレル頻度）の検討が必要である。

論点１：検定の種類

前項の図は、機能欠失型バリアントの遺伝子ベース検定を行った場合の必要サンプルサイズを示している。実際の全ゲノムシークエンス（WGS）を用いたレアバリアント解析には以下の可能性がある[2]。

1. シングルバリアント関連検定
2. 機能欠失型バリアントの遺伝子ベース検定
3. MAF や機能予測スコアでフィルタしたミスセンスバリアントの遺伝子ベース検定
4. 遺伝子セット検定
5. 非コード領域の解析

このうち、前項の図は、2 番の解析のサンプルサイズのみを示している。

論点２： sについて

Zuk らの論文では、突然変異・自然選択・集団サイズ可変の Wright-Fisher モデルを構築し、集団遺伝学的な推定を行っている（移住と集団混合の影響を無視している）。

この時、集団遺伝学パラメータであるsとは、対象とするバリアントアレルの子世代での相対的減少を表し、t世代から$t+1$世代へのレアバリアントアレル頻度の変化について$f_{t+1} \approx (1-s)f_t$として得られるが、この算出は容易ではない。同論文内では$s = (\lambda\pi)s_D$を用いて統合失調症、2 型糖尿病、アルツハイマー型認知症でのバリアントタイプごとの選択係数を計算している。s_Dは decreased reproductive fitness、πは有病率、λはバリアントタイプごとの超過リスク（相対リスクとしては$1+\lambda$）を表す。

これは明確に決定できるものではなく、sは、このくらいの範囲になると把握するものだと考えられる。

＜小括＞

本稿では、複雑性疾患のレアバリアント関連解析に係る Zuk らの論文（(以下「論文」という。）において、25000 例程度を目安としていることについて、現在の我が国の難病の全ゲノム解析等の数値目標を検討するにあたって適切かについての検討をおこなった。

その結果、論文においては、疾患の特性に応じて定められる累積アレル頻度が低い疾患であっても、25000 例を行うことにより、疾病に関連する有意なレアバリアントを見つけられる可能性が高いということを示しており、翻って解釈すると疾患によっては 1000 例程度であっても、リスクの高い遺伝子を見つけることできるともいえる。

難病は、多岐にわたり、また疾病毎に累積アレル頻度を推定することも困難といえることから、第 1 回の検討会で辻参考人が指摘した 1000 例、5000 例と段階的に検体を収集していく戦略は、一定の遺伝統計学的妥当性を有すると考える。

1. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455-64 (2014).
2. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0177-4
3. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).