メンデル遺伝疾患のExome解析における候補変異数

西野 穣　　数理・推論研究系　特任研究員

1. 背景

近年, 次世代シーケンサーと総称されるDNA配列決定技術の発展に伴い, 個人ゲノム配列決定のコストダウン化が進んでいる. 疾患原因変異探索の観点からは, ゲノムの中でもタンパク質をコードする “Exome“ に特に注意が払われている. 何故なら, Exomeはゲノムの1-2%と小さい割合でありながら疾患原因が存在する可能性が大きく, この領域に限定して原因変異を探す方法は, 費用・技術コスト両面において効率的であるからである.

メンデル遺伝疾患に対するExome解析の有効性は,常染色体優性疾患であるフリーマン・シェルドン症候群の既知の原因遺伝子がExomeデータのみを用いて同定できることが示されたことで実証された(Ng et al. 2009). 解析手法としては, 4名の患者Exomeにおいて遺伝子単位で特異的に存在する変異を8名の健常者Exomeと公共データベースを用いて抽出という分かりやすいものであった. 現在, Exome解析により100以上のメンデル遺伝性疾患の原因変異が新規に同定されている.

2. 目的

これまでのExome解析は, 不完全浸透やフェノコピーのほとんど無い明瞭なメンデル性遺伝パターンに従う疾患が主な対象だった. 今後はそれほどメンデル遺伝パターンが明白ではない疾患 (さらには複合疾患) に対するExomeデータの利用も重要な課題となる. そこで, 本研究ではメンデル遺伝疾患のExome解析は, 不完全浸透およびフェノコピーの存在下で実際に有効か？ という問に対して, 遺伝的異質性が無い(=疾患の原因は唯一の変異による) という条件の下, "候補変異数"という観点からこの問いに答える. なお, 本研究は間野修平准教授との共同研究である.

3. 方法と結果

以下の図は, Exome解析の模式図である (1名のフェノコピーを許容している).候補変異数は, サンプル中の変異の頻度が大きく関わっている事が予想される.

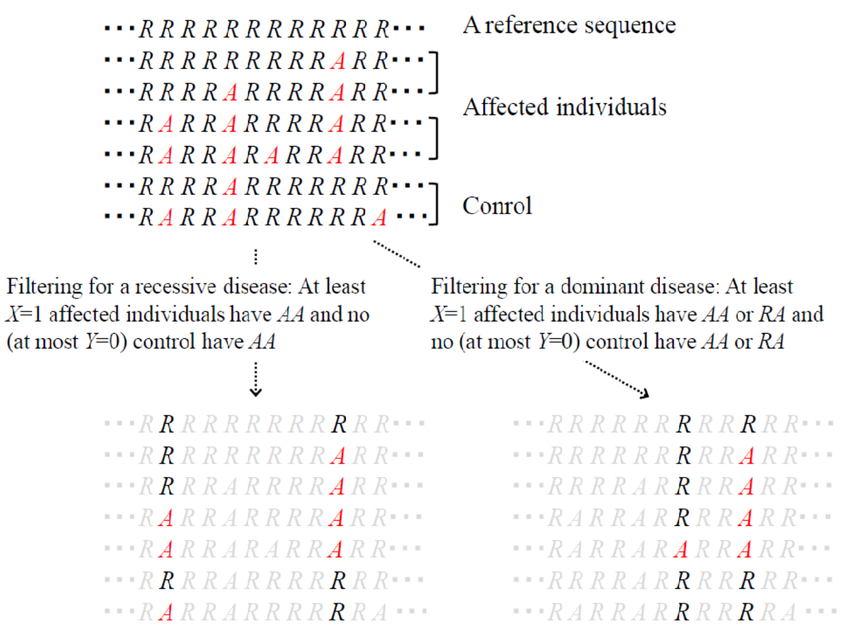

図 1. Exome 解析におけるフィルタリングの模式図

3.1 SNVの頻度スペクトラム

ある生物の遺伝的多様性は, メンデル遺伝を基礎とする確率過程の産物である. 特に重要なWright-Fisher modelは,「集団サイズ一定, 自然選択がない」というモデルであり, 近似的に遺伝的多様性をよく説明するケースも多い. このモデルの下では, n 個のExomeサンプル中でi 個の突然変異型と$(n\text{-}i)$ 個の祖先型を持つSNV (single nucleotide variant) の数M_i (Site frequency spectrum)の期待値は,

$$E[M_i] = \frac{\theta}{i}, \qquad 1 \le i \le n-1, \qquad (1)$$

となる (Fu 1995). ここでθ =4×突然変異率×集団サイズである. なお, この式は, 集団サイズの変動が起こったか, 又はあるクラスのDNAサイトで自然選択が働いているか調べる際に頻繁に用いられる.

Exome解析では, DNAサイトが祖先型か突然変異型かは通常は問わず, リファレンス配列 (eg., hg19) との比較によって, 同じであればR (Reference Allele) 型, 異なればA (Alternative allele) 型かという区分をする. そこで, n 個のExomeサンプル中でn_A 個のA型と$(n\text{-}\ n_A)$ 個のR型を持つSNV (single nucleotide variant) の数M'_{nA} の期待値を求めると,

$$E\left[M'_{n_A}\right] = \frac{\theta}{n+1} + \frac{\theta}{n-n_R}\frac{n_R+1}{n+1} = \frac{\theta}{n-n_R} = \frac{\theta}{n_A}, \qquad 1 \le n_A \le n. \qquad (2)$$

となる. 興味深いことに, これはA 型を突然変異型とみなして(1)を使うことで得られる. 例えば, ヒトのθ は人種間で若干の変動はあるが, 13000程度である (個人の2つのExome の間で約13000個の異なるDNAサイトがある). (2)の式から, 4つのExome (n=4)を考えれば, 1個, 2個, 3個のA 型アリルを持つサイトは平均的に13000個, 6500個, 4333個と求めることができる.

3.2 特定の遺伝子型構成を持つSNVの数

次にN 人が非血縁者について考える. 遺伝子型がAA, RA, AAである個体がn_{RR}, n_{RA}, n_{AA} 人(合計N人)であるDNAサイトの数$K(n_{RR},n_{RA},n_{AA})$ の期待値は,

$$\begin{aligned} &E[K(n_{RR},n_{RA},n_{AA})] \\ &= E\left[M'_{n_A}\right] \times Prob(n_{RR},n_{RA},n_{AA}|n_A) \\ &= \frac{\theta}{n_A} \times \frac{2^{n_{RA}}N!}{n_{RR}!n_{RA}!n_{AA}!}\frac{n_R!n_A!}{(2N)!}. \end{aligned} \qquad (3)$$

となる. 例えば, N=2 (4つのExome) のとき, K(1,1,0), K(0,2,0), K(1,0,1), K(0,1,1), K(0,0,2) の期待値はそれぞれθ, $\theta/3$, $\theta/6$, $\theta/3$, $\theta/4$となり, θ=13000ならば, 13000, 4333, 2167, 4333 , 3250個となる.

3.3 N_aの疾患個体とN_cのコントロール個体

優性疾患を考える. 図1の右 では, 2名の疾患個体で少なくとも1名がAA, RA, コントロール1名ではAA,RAを持たないDNAサイトを抽出している. このようなフィルタリングで残るSNVの期待数は(3)と簡単な組合せ論から計算することができる.

図2に非血縁者Exomeを用いた優性疾患における候補SNV数の期待値をいくつかの条件で計算した. 疾患個体(N_a) とコントロール個体(N_c) が半々の時((b)の×)では, 不完全浸透やフェノコピーが20 %存在するときでも候補変異数は, 急速に小さい値になる, つまりExome解析は有効である事が分かる (N=50で0.2個).

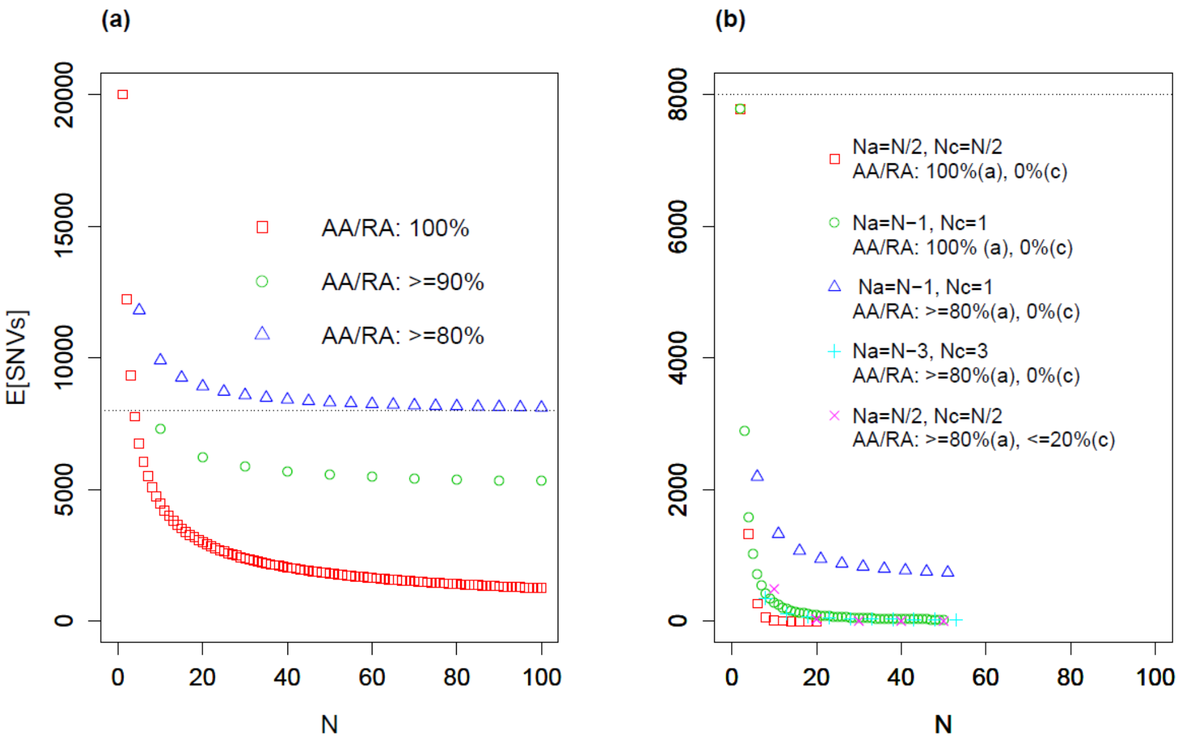

図 2. 非血縁者データを用いた優性疾患における候補SNV数

血縁者データを用いた優性疾患における候補SNV数を図3に示す. コントロールを用いた時(b)は非血縁者のときと近い値を示すことが分かる.

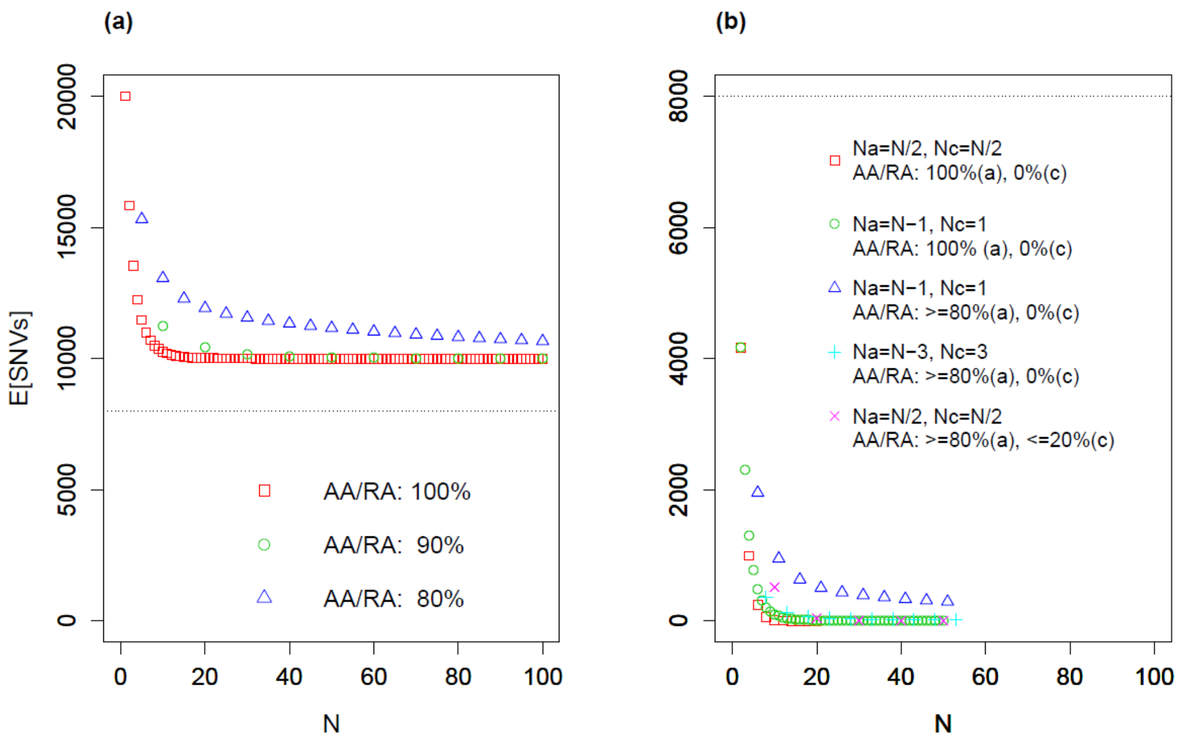

図 3. 血縁者データを用いた優性疾患における候補SNV数
(血縁者のモデルとして“sibs“ であることを仮定)

4. 実際の応用例

ある霊長類において, 特定の優性疾患に確実に罹患している3個体のExome は入手済みであった. このExome に対して, 何個体のコントロール個体のExomeを新しくシーケンシングすればどのくらいの候補SNVまで絞れるか?
⇒ まず, 既に読んでいる3個体のExomeからθ=35333 と推定した. (これはヒトの3倍弱である. ヒトは多様性が小さい事がよく知られている.) このθ の値と上述の議論を用いて以下のように推定した.

N_c	5個体	10個体	15個体	20個体
残る変異数(E[SNVs])	235個	43個	15個	6個