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

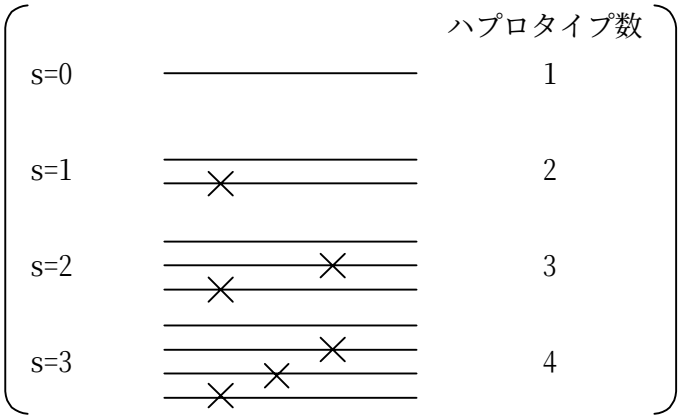
を考慮すべきである

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