統計数理(2012)
第 60 巻 第 2 号 317–325

特集「多様性と進化の統計解析」
［研究ノート］

Extended Haplotype Homozygosity (EHH) を用いる正の自然選択検出法の検出力比較

大橋 順†

(受付 2012 年 1 月 10 日；採択 3 月 1 日)

要　　旨

最近作用した自然選択の痕跡を検出する目的で，REHH 法や iHS 法など，extended haplotype homozygosity (EHH) を用いる検定手法が広く利用されている．しかし，REHH 法の検定統計量は，頻繁に正の無限大に発散するという欠点がある．そこで本研究では，EHH の定義を変更することで，検定統計量が発散することのない新たな REHH 法(改良 REHH 法)を提案する．合体(合祖)シミュレーションにより，自然選択強度やテストアリルの到達頻度が異なるいくつかのパラメタセットに対し，0.01 cM 間隔で連鎖する 101 個の SNP からなるハプロタイプデータを作製した．作製したハプロタイプデータに REHH 法，改良 REHH 法，iHS 法を適用し，それらの検出力を求めた．検出力を比較したところ，ほとんどのパラメタセットに対して，REHH 法の検出力がもっとも低かった．また，テストアリルの集団頻度が低い場合は iHS 法の検出力が最も高く，頻度が高い場合は改良 REHH 法の検出力が最も高いことがわかった．今回の結果は，テストアリルの集団頻度に応じて，改良 REHH 法と iHS 法を使い分ければ，検出力の増加が見込めることを示唆する．

キーワード：単塩基多型(SNP)，自然選択，extended haplotype homozygosity (EHH)，integrated EHH (iHH)，integrated haplotype score (iHS)，relative EHH (REHH)．

1. はじめに

ヒトゲノムの全塩基配列が解読されて以来，ヒトゲノム研究の興味はその多様性の理解へとうつった．国際 HapMap 計画により，一塩基多型(SNP)の詳細なハプロタイプ地図が整備され(The International HapMap Consortium, 2003; 2005)，ゲノムワイドに分布する数十万〜百万個の SNP をタイピングする技術も実用化された．高密度なゲノムワイド SNP データが利用可能になり，これまで困難であったヒト集団に作用した最近の自然選択の痕跡を検出できるようになった．

自然選択上有利な突然変異が誕生すると，突然変異が乗るハプロタイプの集団頻度が急速に増加する．そのため，ハプロタイプ内での多型性の程度は，同じ集団頻度のアリルが乗るハプロタイプ内の多型性の程度よりも相対的に低くなる．この多型性の度合いを Extended Haplotype Homozygosity (EHH)という指標であらわし，EHH に基づく検定統計量 REHH (relative EHH)を使用して自然選択の痕跡を検出する手法(REHH 法)が開発された(Sabeti et al., 2002)．REHH 法は，現在も多型的である遺伝子座(有利なアリルが固定していない遺伝子座)に作用してきた

† 筑波大学大学院 医学医療系分子遺伝疫学：〒305–8575 茨城県つくば市天王台 1–1–1

最近の正の自然選択の検出を目的としており，REHH 法を用いて正の自然選択を受けているアリルが次々と同定されている(たとえば，Sabeti et al., 2002; Fujimoto et al., 2008; Hirayasu et al., 2008)．

REHH 法は広く使用されているが，REHH 検定統計量は容易に発散する(理由は後述)という欠点を有している．本稿では，REHH が発散しないように EHH の定義を変更することで，REHH 法の欠点を克服する新たな手法を提案するとともに，EHH を用いる従来法との検出力の比較を行う．

2. EHH

EHH とは，着目するアリル(以下ではコアアリルという)が乗るハプロタイプのホモ接合度のことであり，サンプル中の 2 つの染色体コピーを取り出した際に，両者のハプロタイプが一致する確率と定義される(Sabeti et al., 2002)．コアアリルが存在する多型(以下ではコア多型という)から，上流もしくは下流のいずれか一方向に多型マーカーを順次含めながらハプロタイプを構成していくと，多型マーカー毎に EHH を計算することができる．コア多型における各コアアリルの EHH は 1 であり，考慮するマーカー数が増えるにつれ異なるハプロタイプが増えるため，EHH はコア多型からの遺伝距離に比例して減少(増加することのないという意味で，広義の単調減少)する．コアアリルの連鎖不平衡が遠方に及べば，多型マーカーが増えてもハプロタイプの種類が増えず，EHH は大きな値のまま維持される．多型マーカー毎に EHH を計算し，その値を直線で結んだものを EHH 曲線という．図 1 に EHH 曲線の例を示す．この図では，コアアリル 1 の連鎖不平衡はコアアリル 2 の連鎖不平衡よりも遠方にまで及んでいる．

以下で，Sabeti et al. (2005)に従い EHH を数学的に定義する．n 本の染色体をサンプルしたとする．コア多型にコアアリルが m 種類存在し，その i 番目のコアアリルを含む染色体のサンプル数を c_i とする．ここで，$n=\sum_{i=1}^{m} c_i$ である．i 番目のコアアリルを含む染色体の中から，2 つの染色体を取り出す組合せ総数は $\binom{c_i}{2}$ である．i 番目のコアアリルを含む染色体の中で，異なるハプロタイプの種類数を t とし，j 番目のハプロタイプを有する染色体のサンプル数を e_j とする．ここで，$c_i=\sum_{j=1}^{t} e_j$ である．このとき，i 番目のコアアリルを含む染色体の中から，同一ハプロタイプを有する 2 つの染色体を取り出す組合せ総数は $\sum_{j=1}^{t}\binom{e_j}{2}$ である(ただし，$e_j=1$ であれば $\binom{e_j}{2}=0$)．以上より，i 番目のコアアリルの EHH は

$$\mathrm{EHH}_i=\frac{\sum_{j=1}^{t}\binom{e_j}{2}}{\binom{c_i}{2}} \tag{2.1}$$

と定義される(Sabeti et al., 2005)．ここで，$0\leq \mathrm{EHH}_i\leq 1$ である．

3. EHH を指標とする正の自然選択検出法

3.1 REHH 法(Sabeti et al., 2002)

あるアリルに正の自然選択が作用すれば，少ない世代数でその集団頻度が上昇する．経過世代数が少ないと，コア多型と近傍の多型マーカーとの間で十分な組換えが起こらないため，正の自然選択が作用するアリルの連鎖不平衡は遠方にまで及ぶことになる．すなわち，コア多型から遺伝距離がかなり離れた地点まで，正の自然選択が作用したアリルの EHH 値は大きいまま維持されると期待される．しかし，ゲノム中の組換え率は一定ではなく，組換えが起こりにくい領域(組換えのコールドスポットともよばれる)が存在するため，あるアリルの EHH 値が大きかったとしても，それが正の自然選択が作用したためなのか，単にコア多型近傍は組換えの起こりにくい領域であるためなのかを区別することができない．そこで，テストしたいアリル(テストアリル)の EHH と，コア多型の他のアリルの EHH との比をとった REHH を検定統

計量に用いる検定法が開発された(Sabeti et al., 2002)．組換えの起こりにくい領域であれば，コア多型の他のアリルの EHH も大きな値をとるため，REHH 法ではこれを組換えの内部コントロールとして利用する．一般的には，コア多型から 0.25 cM，もしくは 500 kb 離れた地点の多型マーカーの位置で REHH を計算する(Sabeti et al., 2002)．以下では，コア多型には 2 つのアリルが存在し($m=2$)，1 番目のアリルを検定したいテストアリル，2 番目のアリルが内部コントロールアリルとする．このとき，テストアリルの REHH は

$$\mathrm{REHH}=\frac{\mathrm{EHH}_1}{\mathrm{EHH}_2} \tag{3.1}$$

とあらわされる．テストアリルに正の自然選択が作用していれば $\mathrm{EHH}_1>\mathrm{EHH}_2$ となり，REHH は正の値をとる．

実データに適用する際は，まず解析対象サンプルのゲノムワイド SNP 遺伝子型データから，fastPHASE プログラム(Scheet and Stephens, 2006)などを用いて各染色体のハプロタイプを推定する．次に，テストアリルと同じサンプル頻度を有するアリル(以下では参照アリルという)を全て選出し，それらの REHH の分布(以下では経験分布という)を求める．REHH の計算に必用な SNP 間の遺伝距離は，HapMap データベースで公開されているデータを利用するか，LDhat プログラム(McVean et al., 2004)などを用いて SNP 遺伝子型データから推定すればよい．最後に，求めた経験分布とテストアリルの REHH と比較する．一般的には，テストアリルの REHH が経験分布の 95 パーセンタイル値より大きければ，テストアリルの EHH は有意に大きく，テストアリルに正の自然選択が作用した可能性があるといえる．なお，REHH の理論分布が不明である(パラメトリック検定ができない)ために経験分布を用いるが，経験分布を用いることで，対象集団の歴史(集団増加やボトルネックなどの集団サイズの変化)の影響を排除できるという利点がある．

3.2　改良 REHH 法

前述の EHH の定義に従うと $0\leq\mathrm{EHH}_i\leq 1$ となる．そのため，$\mathrm{EHH}_2=0$ のときは式(3.1)で示される REHH の値は正の無限大に発散する(Sabeti et al., 2002)．先行研究ではこの点に注意を払っていないが，参照アリルの 5%以上で REHH が発散した場合は 95 パーセンタイル値が無限となり，テストアリルを評価することができなくなる．そこで本研究では，REHH を計算する際に限り，i 番目のコアアリルの EHH を

$$\mathrm{EHH}_i=\sum_{j=1}^{t}\left(\frac{e_j}{c_i}\right)^2 \tag{3.2}$$

と定義する手法(改良 REHH 法)を提案する．ここで，$\frac{e_j}{c_i}$ は i 番目のコアアリルを含む染色体サンプル中の j 番目のハプロタイプ頻度を示している．このように定義すれば，i 番目のコアアリルを含む染色体サンプル中に同一ハプロタイプが 2 本以上存在しない場合，すなわち，$t=c_i$ であっても $\mathrm{EHH}_i=0$ とはならず，EHH のとりうる値の範囲は

$$\frac{1}{c_i}\leq\mathrm{EHH}_i\leq 1 \tag{3.3}$$

となる．これにより，改良 REHH (rREHH)のとりうる値の範囲は

$$\frac{1}{c_1}\leq\mathrm{rREHH}\leq c_2 \tag{3.4}$$

となり，検定統計量が発散することはなくなる．

3.3　iHS 法(Voight et al., 2006)．

EHH の値はコア多型の地点では 1 であり，コア多型から離れるにつれて減少する(図 1)．各

アリルの EHH 曲線の下側の面積を，EHH の値が最初に 0.05 以下となる多型マーカーの地点まで足し合わせたもの(コア多型の上流側の面積と下流側の面積の和)を integrated EHH (iHH)とよぶ．テストアリルの iHH を iHH_1，コントロールアリルの iHH を iHH_2 とすると，integrated haplotype score (iHS)法で用いる検定統計量は

$$\text{unstandardized iHS} = ln\left(\frac{\mathrm{iHH}_2}{\mathrm{iHH}_1}\right) \tag{3.5}$$

であらわされる(Voight et al., 2006)．テストアリルに正の自然選択が作用していれば $\mathrm{iHH}_1 > \mathrm{iHH}_2$ となり，unstandardized iHS は負の値をとる．テストアリルの unstandardized iHS が経験分布の 5 パーセンタイル値より小さければ，テストアリルに正の自然選択が作用した可能性があるといえる．先行研究(Voight et al., 2006)では，各多型のアリル頻度に依存しない統一した検定統計量にすべく unstandardized iHS を標準化しているが，調べたいテストアリルが決まっている場合には，テストアリルと同一頻度の参照アリルの unstandardized iHS の経験分布を使用すればよい．

4. 検出力

各手法の検出力を求めるため，SelSim ソフトウェア(Spencer and Coop, 2004)を用いて合体(合租)理論に基づくコンピュータシミュレーションを行った．二倍体集団サイズ N は一定(5000 個体)，101 個の SNP が 0.01 cM 間隔で連鎖して並んでいると仮定し，51 番目の SNP (コア SNP)の派生アリル(突然変異により誕生したアリル)が正の自然選択の作用を受けると仮定した．残りの SNP は自然選択上中立とし，再起突然変異は仮定しなかった．本研究では，REHH 法と改良 REHH 法の検定において，コア多型から 0.25 cM 離れた上流(26 番目)と下流(76 番目)の SNP における REHH 値の和を検定統計量に用いる．ここで，0.01 cM に 1 個の SNP マーカーが存在すると仮定したが，この密度は，市販されているゲノムワイド SNP チップで解析できる SNP の密度をやや下回る程度である．

まず，自然選択が作用しない中立な状況(帰無仮説下)で期待される検定統計量の分布を求めた．コア SNP の派生アリルをテストアリルとし，その到達頻度が 0.15，0.3，0.6 となる時点で(厳密にいうと，初回到達頻度とは限らない)，集団から 120 本の染色体(60 個体)をサンプルした．120 本としたのは，国際 HapMap 計画で解析されているアフリカ系集団(YRI)とヨーロッ

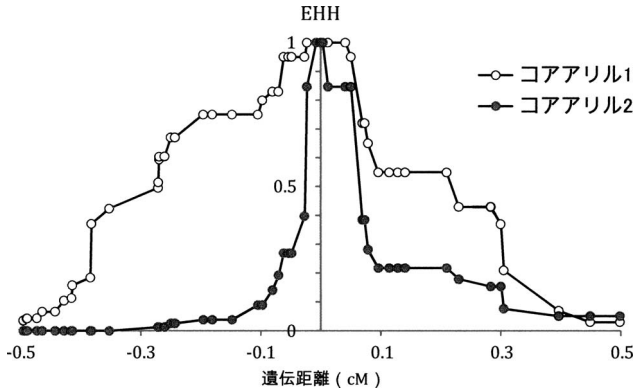

図 1. EHH 曲線の例．コア多型 (0 cM の地点) を中心に，各アリルの上流 (左側) と下流 (右側) の EHH 曲線を示す．

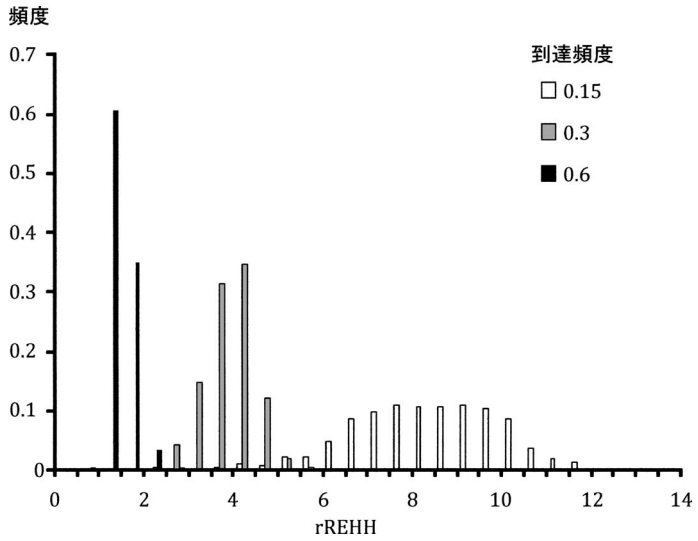

図 2.　中立な状況(帰無仮説下)での rREHH の分布．テストアリルの到達頻度は，0.15(白色棒)，0.3(灰色棒)，0.6(黒色棒)とした．

パ系集団(CEU)のサンプルサイズに合わせたためである．テストアリルを含む染色体サンプル数はその集団頻度に比例するとし，テストアリルの到達頻度が 0.15，0.3，0.6 である場合の，テストアリルを含む染色体サンプルを，それぞれ 18 本(コントロールアリルが 102 本)，36 本(コントロールアリルが 84 本)，72 本(コントロールアリルが 48 本)と設定した．図 2 に，各到達頻度に対し，1000 回の試行からえられた改良 REHH 法の検定統計量 rREHH の頻度分布を示す．到達頻度が低いほど，検定統計量の分散が大きかった．これは，REHH 法でも iHS 法でも観察される一般的な性質である(データは割愛)．

テストアリルの到達頻度が高い場合は，コントロールアリル(祖先アリル)の頻度が低くなるため，REHH 法では 0.25 cM 離れた地点で $EHH_2=0$ となりやすく，中立の下で REHH が正の無限大に発散するケースが多かった．到達頻度が 0.6 の場合は発散するケースが 9%以上もあり，頻度の高いテストアリルに対して REHH は適当な統計量ではないといえる．

次に，テストアリルに正の自然選択が作用した状況(対立仮説下)で期待される検定統計量の分布を求めた．自然選択モデルでは，テストアリルのホモ接合体，テストアリルとコントロールアリルのヘテロ接合体，コントロールアリルのホモ接合体の相対適応度を，それぞれ $1+s$，$1+0.5s$，1 とした(s は正の選択係数)．また，テストアリルの頻度は決定論的に変化するとした．自然選択下で 200 回試行し，REHH 法と改良 REHH 法では，検定統計量が中立な下で得られた検定統計量の分布の 95 パーセンタイル値より大きな値を示した割合を，iHS 法では 5 パーセンタイル値より小さい値を示した割合をそれぞれ検出力とした．図 3 に各手法の検出力を示す．ここでは，集団サイズと選択係数の積($2Ns$)で自然選択強度を示してある．改良 REHH 法と iHS 法では，自然選択強度が大きいほど，またテストアリルの到達頻度が高いほど，高い検出力が達成された．経験分布の 95 パーセンタイルが無限大となったため，到達頻度が 0.6 の場合の REHH 法の検出力は 0 とした．iHS 法の検出力は REHH 法の検出力よりも常に高かったが，到達頻度が 0.6 の場合は，改良 REHH 法の検出力が iHS 法の検出力を上回っていた．

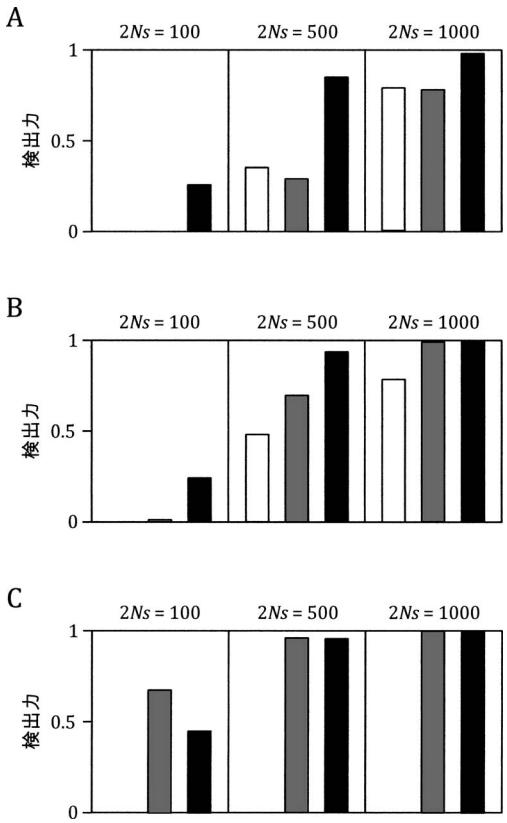

図 3. 検出力の比較．REHH 法(白色棒)，改良 REHH 法(灰色棒)，iHS 法(黒色棒)の検出力を示す．自然選択の強度は，集団サイズと選択係数の積($2Ns$)で示してある．テストアリルの到達頻度は，(A) 0.15，(B) 0.3，(C) 0.6 とした．(C) では，REHH 経験分布の 95 パーセンタイル値が発散したため，REHH 法の検出力を 0 とした．

5. まとめ

コンピュータシミュレーションによって作製したデータを用いて，REHH 法(Sabeti et al., 2002)，改良 REHH 法(本研究)，iHS 法(Voight et al., 2006)の検出力を比較したところ，テストアリルの到達頻度が低い場合は iHS 法の検出力が最も高く，到達頻度が高い場合は改良 REHH 法の検出力が最も高いことがわかった．このことは，テストアリルのサンプル頻度に応じて，改良 REHH 法と iHS 法を使い分ければ検出力の増加が見込める可能性を示唆している．検出力はさまざまな未知パラメタ(集団サイズ，選択係数など)の影響を受けるため，テストアリルの集団頻度に応じた検定法選択が実際に有効か否か，さらなる検討が望まれる．

テストアリルの到達頻度が 0.6 の場合には，設定条件下において REHH 分布の 95 パーセンタイル値が無限大となった．REHH の発散を防ぐためには，0.25 cM よりもコア多型に近い位置で REHH を計算すべきであるが，コア多型に近づき過ぎれば検出力が下がることは明白である．そのため，いくつかの距離に対して経験分布を求め，少なくとも 95 パーセンタイル値が発散せずに，なるべく距離が離れた地点で計算した分布を選ぶ必要がある．しかし，事後的にコア多型からの距離を決める(事後的に使用する経験分布を決める)ことは，検定の客観性を

担保する上では好ましくないであろう.

最後に，本研究で扱った自然選択検出法に共通する問題点を二つ指摘する．一つめは，SNP マーカー間の遺伝距離の推定に誤りがあると，検定統計量が大きな影響を受けることである．SNP 間の遺伝距離は連鎖不平衡データから推定したものであり，HapMap データベースで公開されている推定値も含めてそこには相当な誤差が含まれることに注意すべきである．二つめは，ゲノムワイド SNP データから網羅的に自然選択の痕跡を検出する場合に，検定統計量の理論分布が不明なため有意水準の補正ができない(多重検定の問題を解決できない)ことである．そのため，ゲノムワイドに多型をスクリーニングする際は，EHH のような連鎖不平衡の指標に加え，F_{st} のような集団分化の指標も考慮して，結果の蓋然性を高める努力をするべきであろう(たとえば，Kimura et al., 2007; Sabeti et al., 2007)．Kimura et al. (2007)は，上述の問題を解決すべく，遺伝距離情報やハプロタイプ情報を使用せずに(遺伝距離推定やハプロタイプ推定での誤りに影響を受けない)，連鎖不平衡と集団分化の程度から地域特異的に作用した最近の自然選択を検出する有力な手法を提案している．次世代シークエンサーの使用により全塩基配列情報が取得できるようになれば，多型頻度スペクトラムなどを用いてさらに多くの指標が利用可能となる．今後，より強力な検定手法が開発され，自然選択が作用したことが確かな全てのヒトゲノム領域が明らかになることに期待したい.

謝　辞

今回提案した手法は，木村亮介氏(琉球大学亜熱帯島嶼科学超域研究推進機構)との議論に基づいており，同氏から多くのご助言をいただいたことをここに記して感謝申し上げる．なお，この研究は統計数理研究所 共同研究課題(H23-J-4312)に基づくとともに，文部科学省科学研究費補助金・新学術領域研究(研究課題番号：23133502)からの助成による.

参 考 文 献

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., Morishita, Y., Furusawa, T., Nakazawa, M., Ohtsuka, R. and Tokunaga, K. (2008). A scan for genetic determinants of human hair morphology: *EDAR* is associated with Asian hair thickness, *Human Molecular Genetics*, **17**, 835–843.

Hirayasu, K., Ohashi, J., Tanaka, H., Kashiwase, K., Ogawa, A., Takanashi, M., Satake, M., Jia, G. J., Chimge, N. O., Sideltseva, E. W., Tokunaga, K. and Yabe, T. (2008). Evidence for natural selection on leukocyte immunoglobulin-like receptors for *HLA* class I in Northeast Asians, *American Journal of Human Genetics*, **82**, 1075–1083.

Kimura, R., Fujimoto, F., Tokunaga, K. and Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation, *PLoS One*, **2**, e286.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome, *Science*, **304**, 581–584.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**, 832–837.

Sabeti, Pardis C., Walsh, Emily, Schaffner, Steve F., Varilly, Patrick, Fry, Ben, Hutcheson, Holli B.,

Cullen, Mike, Mikkelsen, Tarjei S., Roy, Jessica, Patterson, Nick, Cooper, Richard, Reich, David , Altshuler, David, O'Brien, Stephen and Lander, Eric S. (2005). The Case for Selection at CCR5-Δ32, *PLoS Biology*, **3**(11), e378.doi:10.1371/journal.pbio.0030378

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S. and The International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations, *Nature*, **449**, 913–918.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase, *American Journal of Human Genetics*, **78**, 629–644.

Spencer, C. and Coop, G. (2004). SelSim: A program to simulate population genetic data with natural selection and recombination, *Bioinformatics*, **20**, 3673–3675.

The International HapMap Consortium (2003). The International HapMap Project, *Nature*, **426**, 789–796.

The International HapMap Consortium (2005). A haplotype map of the human genome, *Nature*, **437**, 1299–1320.

Voight B. F., Kudaravalli S., Wen X. and Pritchard J. K. (2006). A map of recent positive selection in the human genome, *PLoS Biology*, **4**, e72.

Proceedings of the Institute of Statistical Mathematics Vol. 60, No. 2, 317–325 (2012) 325

A Comparison of Statistical Power of EHH-based Methods for Detecting a Signature of Recent Positive Selection

Jun Ohashi

Molecular and Genetic Epidemiology, Faculty of Medicine, University of Tsukuba

Extended haplotype homozygosity (EHH)-based methods such as relative EHH (REHH) and integrated haplotype score (iHS) tests have been used for detecting a signature of recent positive selection in human populations. In this study, with a slight modification of definition of EHH, a revised REHH (rREHH) test is proposed in which the divergence of test statistic can be avoided. A comparison of the statistical power of three EHH-based methods for haplotype data obtained by coalescent simulation revealed that iHS test achieved the highest power for a selected allele with low frequency, while rREHH test showed the highest power for one with high frequency. For most parameters, REHH test showed the lowest power. The present results suggest that, to efficiently detect recent positive selection, rREHH and iHS methods should be used properly based on the population frequency of core allele to be tested.

Key words: SNP, positive selection, extended haplotype homozygosity (EHH), integrated EHH (iHH), integrated haplotype score (iHS), relative EHH (REHH).