ゲノムワイドな古いハプロタイプブロックによる疾患関連形質の多様性解析

小貫　律子

2018 年 5 月

要旨

疾患関連形質の集団間の多様性についての遺伝情報は、公衆衛生、個別化予防にとって重要であり、疾患関連形質の集団間の多様性に関連する座位をゲノムワイドに探索することが必要になっている。これまでの研究では、近年の正の選択を検出する研究が多くなされている。こうした研究では約 1 万年前以降の進化過程と疾患関連形質の集団間の多様性を関連させる。一方、最近の研究では、古く(10 万年以上前)から起こってきた病原体への適応がその後の集団の移動にともない集団間で異なる自然選択を受け、その違いが現在の自己免疫疾患や代謝疾患などの集団間の多様性に影響を与えている事例が示されている。しかし、古い(1 万年以上前の)自然選択を検出する研究はまだ少なく、実際にこうした長期的な進化プロセスを経て集団間の違いを生じたゲノム領域があるのかどうかあきらかでない。また、こうした長期の進化プロセスを経たゲノム領域を検出することにより、これまで見出しにくかった多様性情報を得ることができるのかどうかはあきらかではない。そこで本研究では、古い正の選択を受け、かつその後に多様化が起ったゲノム領域を検出する手法を独自に開発した。本研究では、まず新たに HHD というディプロタイプ間距離の推定法を開発した。さらに、HHD を利用することにより、上述のプロセスを経たゲノム領域を検出するためのパイプラインを構築した。このパイプラインではゲノムワイドな一塩基多型データのうち、連鎖不平衡にあるものの中から古い正の選択が起こったゲノム領域を検出する。そして、古い正の選択の候補領域について集団間で異なる近年の自然選択を受けてきた領域かどうかを評価し、近年の自然選択の候補領域の検出感度の向上を試みた。そして最後に疾患関連形質の多様性との関連を検討する。HapMap の遺伝子型データにこのパイプラインを適用したところ、古い正の選択かつ近年の自然選択による多様化の候補領域に含まれてい

た遺伝子の 75%は既存の近年の正の選択の研究では検出されていなかったもので、検出された多くの遺伝子は免疫関連の遺伝子であった。また、疾患関連形質の多様性との関連について検討したところ、実際に観測された多様性との整合性がみられる例がいくつか見つかった。例えば、セリアック病はアフリカの集団で発症率が高く、NOD2 認識パスウェイの活性化に関与する感受性遺伝子には細菌感染からの防御のための正の選択が近年アフリカで起こったことがすでに知られているが、本研究では、古い時期にもアフリカで正の選択が起こり、発症率の違いに関与していることが示唆された。また、C 型肝炎は集団間の発症率の違いが知られているが、本研究では C 型肝炎ウィルスの除去率に関与する T 細胞受容体シグナル伝達経路の遺伝子がアフリカで古い正の選択を受け、発症率の違いに関与している可能性が示唆された。このほか、アレルギー感作に関連する Jak-STAT シグナル伝達経路の受容体付近で機能する遺伝子がアフリカにおいて古い正の選択を受けたことにより、アレルギー感作の集団間の違いが生じている可能性が示唆された。今後、疾患関連形質の集団間の違いをもたらす座位、機構を明らかにしていくために、これらの集団間の違いが生じた背景について本研究の仮定では考慮しなかった要因についてもさらに検討し、本研究の結果を検証していく必要がある。

目次

要旨 2
第 1 章　序論 9
1.1 疾患関連形質の集団間の違い 9
1.2 ゲノムワイド関連解析による疾患関連形質座位の同定 10
1.2.1 一塩基多型と遺伝子型 10
1.2.2 連鎖不平衡 12
1.2.3 ありふれた疾患とゲノムワイド関連解析 15
1.3 進化的な観点から研究されている集団間の形質の違い 17
1.4 古い正の選択と疾患の関係 18
1.5 本論文の目的と構成 20
第 2 章　ディプロタイプ間距離の推定法の提案 26
2.1 ディプロタイプの分類のためのディプロタイプ間距離 26
2.2 先行研究 27
2.2.1 Allele Sharing Distance (ASD) 27
2.2.2 ハプロタイプ間距離 27
2.2.3 HIT アルゴリズム 28
2.3 ディプロタイプ間距離の定義 29
2.3.1 Population Model-based Distance (PMD) 30
2.3.2 HIT HMM-based Distance (HHD) 31
2.3.3 PMD と複数祖先仮説 32
2.4 連鎖平衡・不平衡による HHD の評価 38
2.4.1 データセット 38
2.4.2 ASD と HHD を使ったクラスタリングの精度の比較 40
2.4.3 HHD を使った機械学習の精度について 44
第 3 章　ディプロタイプ間距離 HHD を用いた古い正の選択の検出方法の開発 50
3.1 古い正の選択について 50

3.2 古い正の選択を検出するパイプライン 51
3.2.1 古いハプロタイプブロックの抽出 52
3.2.2 ディプロタイプ間距離 HHD の計算 54
3.2.3 集団間の違いを持つ古いハプロタイプブロックの抽出 55
3.2.4 古いハプロタイプブロックのクラスタリング 56
3.3 パイプラインにより抽出された古いハプロタイプブロックの機能アノテーション 57
3.3.1 遺伝子のエンリッチメント解析による機能アノテーション 57
3.3.2 SNP の GWAS カタログによる機能アノテーション(SNP のエンリッチメント解析による機能アノテーション) 57
3.4 パイプラインの実データへの適用 58
3.4.1 データセット 58
3.4.2 抽出された古いハプロタイプブロック 59
3.4.3 古いハプロタイプブロックのスコアリングとクラスタリング 61
3.4.4 上位 1%ブロックの新規性と機能アノテーション 71
3.5 考察 94
3.5.1 古いハプロタイプブロックの特徴 94
3.5.2 上位 1%ブロックと F_{st} 94
3.5.3 疾患関連形質の集団間の形質の違いと免疫システム関連パスウェイ 95
3.5.4 疾患関連形質の集団間の形質の違いと機能モジュール 97
第 4 章　総括 99
謝辞 102

図目次

図 1-1　ハプロタイプブロックの生成過程
図 1-2　古い自然選択と近年の自然選択
図 1-3　古い正の選択と近年の自然選択を受け多様化が起こったアレル
図 1-4　古い正の選択により生じた古いハプロタイプブロックと集団特異的な自然選択
図 2-1　HIT アルゴリズムの HMM
図 2-2　ASD と PMD の違いがみられるディプロタイプの例
図 2-3　図 2-2 のaとbに対する最適な祖先ハプロタイプ
図 2-4　図 2-2 のディプロタイプの変異数に基づいた$H = PMD_{M_1}$とPMD_{M_2}のクラスタリングの結果
図 2-5　HapMap の SNP についてのベン図
図 2-6　CER の平均値とh_Bの関係のプロット
図 2-7　h_Bと成功の割合の関係のプロット
図 2-8　Haploview を使った HapMap の SNP の NAT2 遺伝子領域における LD プロット
図 3-1　提案するパイプラインと機能アノテーションの概観図
図 3-2　古い正の選択の候補領域の例
図 3-3　古いハプロタイプブロックの例
図 3-4　各古いハプロタイプブロックで作成された 270x270 の HHD 行列
図 3-5　スコア分布
図 3-6　古いハプロタイプブロックの分類
図 3-7　クラスターごとのスコア分布
図 3-8　集団間のアレル頻度の大きな違いが見られた SNP のアレル分布
図 3-9　上位 1%ブロックの各クラスターに想定した歴史
図 3-10　上位 1%ブロックの SNP のアレル頻度と想定された歴史
図 3-11　得られた遺伝子がマップされた Jak-STAT シグナル経路

表目次

表 1-1　古い正の選択と近年の自然選択の起こり方の可能な組合せ
表 2-1　図 2-2 のディプロタイプ間の変異数
表 2-2　図 2-2 のディプロタイプ間の距離
表 2-3　ハプロタイプ頻度で与えられた集団モデル
表 2-4　図 2-2 のディプロタイプの候補ディプロタイプの表 2-3 で与えられた集団モデルの下での条件付確率
表 2-5　CEU データセットの 3-fold cross-validation の結果
表 2-6　YRI データセットの 3-fold cross-validation の結果
表 2-7　シミュレーションデータセットの 3-fold cross-validation の結果
表 2-8　ASN データセットの 3-fold cross-validation の結果
表 2-9　クローン病データセットの 3-fold cross-validation の結果
表 2-10　自己免疫疾患データセットの 3-fold cross-validation の結果
表 3-1 先行研究で検出された領域の平均長
表 3-2　上位 1%ブロックで遺伝子を含んでいるもののリスト
表 3-3　上位 1%ブロックの各クラスターに含まれる遺伝子
表 3-4　上位 1%ブロックの各クラスターに含まれる nsSNP とそれらを含む遺伝子数
表 3-5　上位 1%ブロックの各クラスターに含まれるミスセンス/ナンセンス SNP を含む遺伝子
表 3-6　クラスタリングに用いた t-統計量スコアプロファイル
表 3-7　スコアリングとクラスタリングの結果
表 3-8　上位 1%ブロックのクラスターごとの既に正の選択が報告されている遺伝子
表 3-9　エンリッチメント解析で得られたパスウェイとマップされた遺伝子
表 3-10　エンリッチメント解析で得られた免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイにマップされた遺伝子とその機能
表 3-11　上位 1%ブロックの ns SNP のうち免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイのコード領域に存在するもの
表 3-12　免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイにマップされた遺伝子を含むブロックを構成する SNP のうち nsSNP ではなく、また、イントロン以外に存在する SNP
表 3-13　各クラスターの遺伝子がマップされた関連パスウェイ
表 3-14　上位 1%ブロックのクラスターごとの既に何らかの形質との関連が

GWAS によって報告されている SNP
表 3-15　クラスター4 の遺伝子がマップされたパスウェイ
表 3-16　本研究で注目したパスウェイにマップされたイントロン以外の SNP もしくは GWAS カタログで報告されている SNP のうち、アレル頻度の集団間の大きな違い(F_{st} >0.15)が確認できたもの

第 1 章　序論

1.1 疾患関連形質の集団間の違い

ヒトの疾患のうち一般的に良く見られる発症頻度の高い疾患(以下、ありふれた疾患と呼ぶ)は由来の異なる集団により発症率の違いがある[1-4]。多くのありふれた疾患、たとえば、冠動脈疾患などの心疾患、高血圧・2 型糖尿病・肥満などの代謝疾患、アルツハイマー病などの神経変性疾患、リウマチ性関節炎・1 型糖尿病・アレルギー性疾患などの自己免疫疾患は、アフリカ・ヨーロッパ・アジアなど由来の異なる集団によって発症率の違いが見られる。また、これらの疾患の感受性アレルにおいては集団間によるアレル頻度の違いが見られることも分かっている[2, 5]。ここで感受性アレルとはゲノムワイド関連解析などの検出手法により検出された、疾患の発症に関連する可能性の高いアレルで、主に複数の遺伝子が原因の疾患に用いられる。なかでもよく報告されているのが 2 型糖尿病についてである[6]。アフリカの集団はアジアの集団より 2 型糖尿病の発症率が高いことが知られており、また、アフリカの集団でアジアの集団に比べて高いアレル頻度を示す感受性アレルが複数見つかっている。

また、 がんについても由来が異なる集団間で発症率の違いがあることが報告されている。アフリカ系アメリカ人ではそれ以外の集団に比べて前立腺がんの発症率が高いことが観測されている[7]。腎細胞がんや肺がんの発症率もアフリカ系の集団において高いことが分かっている[8, 9]。一方で、乳がん、子宮内膜がん、またメラノーマはヨーロッパ由来の集団が、そのほかの集団よりも高い発症率を持つことが観測されている[10-12]。これらの違いには環境要因や社会経済などの原因が示唆されているが、遺伝的要因についてはあまり研究されていない[13]。

その他、疾患関連の形質では、由来の異なる集団によって薬物応答の違いが見られることも知られており、特にシトクロム P450 についてはよく知られている[14]。CYP3A5 の第 3 イントロンに存在する SNP のアレルは CYP3A5 の発現や臨床的に重要な薬(免疫抑制タクロリムス・HIV プロテアーゼ阻害薬サキナビル)の代謝に関わっていることが示されており、一方でアレル頻度がアフリカ由来とヨーロッパ由来の集団で異なることから、異なる由来の集団によってこれらの薬物応答が異なると考えられている[15]。

1.2 ゲノムワイド関連解析による疾患関連形質座位の同定

1.2.1 一塩基多型と遺伝子型

同種の集団内において、異なる個体間で、ゲノム配列上の同じ場所が異なる場合、それを多型(polymorphism)といい、その違いが一塩基の場合、一塩基多型(single nucleotide polymorphism; SNP)と呼ぶ。SNP がコード領域に起きたとき、その変異は、同義置換、ミスセンス置換、ナンセンス置換の 3 種類に分類される。同義置換はアミノ酸配列に変化がない変異、ミスセンス置換はアミノ酸配列に変化がある変異、ナンセンス置換は終止コドンに変化する変異のことをいう。コード領域に 1 塩基、2 塩基単位、もしくは 3 の倍数以外の塩基単位の挿入/欠失が起こった場合は、それ以降コドンへの翻訳にずれが生じるので frame shift という。

ここで、本論文で用いる単語について説明する。一般に座位(locus)とは染色体上の遺伝子の位置、アレル(allele)とは遺伝子座位の片親由来の対立遺伝子をあらわすが、本研究では座位とは SNP が観測された一塩基の位置、アレルは SNP 座位で観測された相同染色体上の両親由来の二塩基のうち片親由来の塩基に対して用いる。SNP 座位で観測された両親由来のアレルのペアは遺伝子型

(genotype)と呼ぶ。同じ染色体上のアレルの並びをハプロタイプ(haplotype)という。相同染色体上のハプロタイプのペアをディプロタイプ(diplotype)という。遺伝子型を構成するアレルがそれぞれ相同染色体のうちどちら由来か分かっていないとき、つまり、ハプロタイプが分かっていない時、”相が不明”という。家系情報があれば相は明らかだが、親族を含まないデータの場合にも相を明らかにするために EM アルゴリズムなどを使った手法が開発されている。

次に、これらの表記法について説明する。今、同じ染色体上に並んでいるm個のSNPを、並んでいる順番に$1, \dots, m$と番号づける。各SNPのアレルは、$S = \{0, 1\}$の要素で 0 か 1 をとるとする。ある SNP でアレル頻度の低いアレルをマイナーアレル、高い方のアレルをメジャーアレルというが、ここでは 0 と 1 をそれぞれマイナーアレル、メジャーアレルとする。ハプロタイプはアレルの配列で、S^m $(e.g.,\ 10101 \in S^5)$の要素として表されるとする。また、遺伝子型は、$D = S \times S$ $(e.g., \{0,1\} \in D)$の要素とする。相が不明なディプロタイプは遺伝子型の配列で、D^m $(e.g., \{1,0\} - \{0,0\} - \{1,0\} - \{1,1\} - \{1,0\} \in D^5)$の要素として表す。相が明らかなハプロタイプのペア、ディプロタイプは、例えば、{10010,00111}のように表記する。

SNP は、ヒトゲノムにおいて最も一般的な多型である[16]。SNP はおよそ 1,000 塩基対(base pair; bp)に 1 回の頻度でヒトゲノム中に分布している。SNP は病因遺伝子の同定をするための遺伝子マーカーとして有用であると考えられ、大規模に SNP を同定しデータを収集する必要性が認められている。現在では、1,000 人ゲノムプロジェクトなどが大量のゲノムワイドな SNP データをインターネット上で一般公開し、それらのデータは、ゲノムワイド関連解析などの研究に用いられている[17]。

1.2.2 連鎖不平衡

ヒトゲノム中の連鎖不平衡(Linkage disequilibrium; LD)の構造を詳しく理解することも、疾患関連形質座位を同定する際に非常に重要であると考えられる。ゲノム配列に起こった変異の中でも他の変異に比べて早い速度(世代)で集団内に広まるものがあり、それらがその集団にとって有利なものであった場合、その変異は正の選択を受けたという。早い速度で広まった変異は、組み換えの影響をあまり受けずに周辺の領域が保存されたまま集団に広まり、ハプロタイプブロックが生じると考えられる(図 1-1)。そこで、集団内で保存された、変異を含む配列(ハプロタイプ)を探索することは、重要な機能を持つ遺伝子を検出するのに有用だと考えられている。その際重要なのが、変異を含むハプロタイプがどの程度集団内に広がっているかを測ることで、連鎖不平衡の尺度が用いられる。

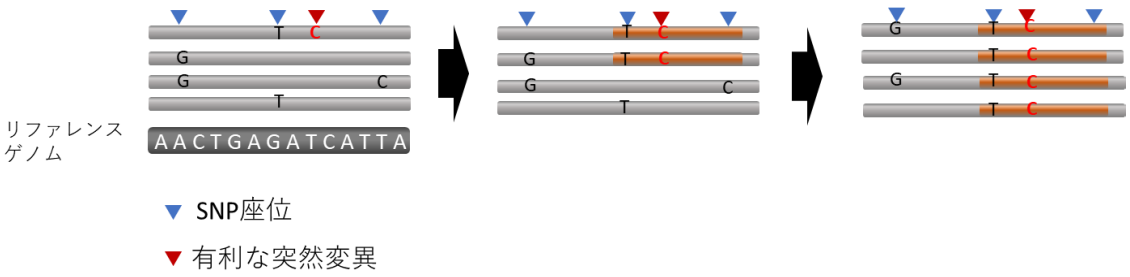


図 1-1 ハプロタイプブロックの生成過程

有利な変異が生じ、その変異が集団内で急速に広がり、有利な変異の周辺は組み換えの影響を受けずにハプロタイプブロック(オレンジ色の領域)が生じる。

連鎖平衡・連鎖不平衡とは、同じ染色体上に並んでいるアレル間の関係により SNP 間の関係性をあらわすものである[18]。より具体的には、同じ染色体上

のアレルが同じ祖先に由来するものなのか、つまり同じ染色体上の 2 つのアレル間に相関があるのかどうかで SNP 間の関係性をあらわす。2 つのアレル間に相関がなく互いに独立な関係であるとき、2 つの SNP は連鎖平衡、2 つのアレル間に相関があるとき、2 つの SNP は連鎖不平衡であるという。この 2 つのアレル間の相関の強さを数学的に表現しようと試みられ用いられるようになったのが、連鎖不平衡の尺度である。

今までに、様々な連鎖不平衡の尺度が提案されているが、それらは共通して、同じ染色体上の 2 つのアレルからなるハプロタイプの頻度の観測値と 2 つのアレルの相関がないと仮定したときのハプロタイプの頻度の期待値の差を表わすものである。今、隣り合う 2 つの SNP を考える。1 つ目の SNP(座位 1 とする)のアレルを A と a で表わし、2 つ目の SNP(座位 2 とする)のアレルを B と b で表わす。今、$\pi_A, \pi_B, \pi_a, \pi_b$ をアレル A、B、a、b の頻度、また、$\pi_{AB}, \pi_{Ab}, \pi_{aB}, \pi_{ab}$ をハプロタイプ AB、Ab、aB、ab の頻度とする。もし、座位 1 と座位 2 の相関がなく、互いに独立で、それらの関係が連鎖平衡(Linkage equilibrium; LE)であった場合には、例えばハプロタイプ AB の頻度π_{AB}はアレル A、B の頻度の積$\pi_A\pi_B$から計算される期待値として表される。しかし、連鎖不平衡がある場合には、ハプロタイプ頻度は、2 つのアレル頻度の積だけでは表されない。祖先ハプロタイプが世代を経ても保存され、アレル頻度から予測されるハプロタイプ頻度より観測されるハプロタイプ頻度がこれを上回るからである。このとき、

$$\pi_{AB} = \pi_A\pi_B + D$$

$$\pi_{Ab} = \pi_A\pi_b - D$$

$$\pi_{aB} = \pi_a\pi_B - D$$

$$\pi_{ab} = \pi_a\pi_b + D$$

と表し、Dを連鎖不平衡係数と呼ぶ。式変形によりDは、$D = \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}$と

なる。連鎖不平衡がない場合、つまり連鎖平衡のときは$D = 0$である。

連鎖不平衡は、染色体連鎖の概念との関連から説明することができる。染色体連鎖というのは、同一染色体上の近傍に位置するマーカーセットが家族の世代を超えても物理的に同一染色体上の近傍に乗った状態を保っていることをいう。この染色体上のブロックを分離するものは、世代が変わるときに起こる減数分裂の際の組み換えである。よって固定サイズの集団内で、random mating によって世代が変われば変わるほど連続した染色体のブロックは分断される。そして最終的には連鎖平衡ブロック、すなわち複数の祖先由来でそれぞれ関連性のない独立したブロックが細かく組み合わさり、関連があるマーカーが近傍に並ぶことがなくなる状態となる。この連鎖平衡の状態に比べてどれだけ元の状態に近いかを測る尺度が連鎖不平衡である。アフリカ系統は、もっとも古くから存在する系統だとされており、系統内での組み換えの蓄積により連鎖不平衡の領域は小さく、ヨーロッパ系統やアジア系統では、アフリカ系統よりも大きな連鎖不平衡の領域が存在する[19]。

Haploview はこの連鎖不平衡ブロックをハプロタイプブロックとして検出するツールである[20]。Haploview のハプロタイプブロックの検出法は、2 つの SNP 間の連鎖不平衡係数に基づいている[21]。2 つの SNP 間の連鎖不平衡係数の 95%信頼区間を観測データによる正確確率分布から計算する。そして、95%信頼区間の上限が>0.98 もしくは下限が>0.7 の時、連鎖が強い、逆に上限が<0.9 の時、組み換えがあったという強い証拠があったとする。対象の領域を一定値、例えば 50kb ずつに区切り、その中の SNP 間の連鎖不平衡係数を計算し、連鎖が強いと判定された区間を次々つなげていく。最終的に検出されたブロックは可視化され、三角形の太線に囲まれた領域として表現される。デフォルトでは Haploview は 500kb 以上離れている SNP ペアの比較は行わないことにしてい

る。

ゲノム中の SNP を解析する際に、すべての SNP を使うのではなく、ある領域の複数の SNP とおなじ情報を持つとみなす SNP を代表として選択して解析する場合も多く、選ばれた SNP を tagSNP(手法によっては haplotype-tagging SNP (htSNP))と呼ぶ[18]。tagSNP は、連鎖不平衡によって相関がみられる SNP を表現する最小の SNP 集合である。tagSNP として選ばれた SNP 集合は、集合内の各 SNP が、集合内の少なくとも 1 つの他の SNP と、ある閾値以上の連鎖不平衡係数を持つように選ばれる。連鎖不平衡のパターンは集団によって異なるので、ある集団で選ばれた tagSNP が、必ずしもそのほかの集団でも tagSNP になるかどうかは分からない。ほとんど tagSNP が存在しない領域は、SNP 間の相関が高く、少ない SNP で周辺領域の SNP を表現することができるため LD と解釈することができる。逆に、多くの tagSNP が存在する領域は、その周辺の SNP はお互いに相関がない LE と解釈することができる。このような連鎖不平衡を用いて選択される tagSNP は、遺伝子型決定する SNP が冗長性のある情報を持たないようにするために有効に使われている。HapMap プロジェクトによるデータの解析によると、遺伝子型決定の際に一般的に使われる SNP アレイで扱うおよそ 100 万 SNP の部分集合によって、ヨーロッパ由来の集団の一般的な SNP(マイナーアレル頻度が 0.05 以上の SNP)の 80%以上を扱うことができる[18]。

1.2.3 ありふれた疾患とゲノムワイド関連解析

ありふれた疾患共通変異仮説(common disease common variant hypothesis; CD/CV 仮説)とは、ありふれた疾患、つまり発症頻度の高い疾患で遺伝と関連している場合は、原因変異遺伝子が、家系が異なっていても共通のものが多い

であろうという仮説である[18]。きわめてまれな遺伝病の場合、特に優性の遺伝形式をとる場合、原因となる変異遺伝子は、次世代に伝わりにくく、家系によって異なることが多い。CD/CV 仮説のもとでは、多くの家系で見られるありふれた疾患は、共通の祖先突然変異に由来すると考え、家系によらず、ありふれた疾患を発症した集団で高い頻度が観測された変異遺伝子をありふれた疾患の感受性遺伝子とみなす。

過去およそ 10 年間のあいだに、CD/CV 仮説に基づいたゲノムワイド関連解析(Genome-Wide Association Study; GWAS)が様々なありふれた疾患に対して試みられ、ありふれたアレルがありふれた疾患の感受性アレルとして役割を果たしていることが報告されている。GWAS とは、興味のある形質を持つ個体集団の DNA サンプルに対して、ゲノム配列中の変異を捕える高密度の遺伝子マーカーのアレイを用いて遺伝子型を決定し、得られた遺伝子型と形質の関連を検出し、ある形質の感受性遺伝子を検出することである[18]。2003 年に約 30 億塩基からなるヒトゲノムの DNA 配列が解明されて以降、過去およそ 10 数年にわたり、SNP を遺伝子マーカーとしたヒトの疾患に関する GWAS が盛んに行われた。その際、連鎖不平衡の情報を用いることにより、非コード領域の SNP でもコード領域の重要な変異と連鎖不平衡にあることで、疾患と関連する重要なマーカーとなることが示された。また、GWAS の対象として tagSNP を用いることにより、より効率的にマーカーを検出することが可能となると考えられている[18]。

NHGRI GWAS カタログ(http://www.genome.gov/gwastudies)には、ありふれた疾患・形質との関連が検出された 3,600 以上の SNP がリスト化されている。ここでもありふれた疾患との関連性が検出された感受性アレルの 80%以上がコード領域以外に存在することが分かり、関連解析には、コード領域、非コード

領域どちらも重要であることが分かっている。

1.3 進化的な観点から研究されている集団間の形質の違い

集団特異的に見つかる有益な形質に関連する遺伝子を見つけるために、ヒトゲノム中で正の選択の影響を受けた領域を検出することが提案された[22]。なかでも近年の正の選択の影響を受けた領域についての研究が多く行われている。ここで近年とは、ヒトがアフリカ大陸から拡散した後の、今からおおよそ 10,000 年前以降のことをいう。10,000 年前というのは新石器時代の始まりの時期であり、植物の栽培化や動物の家畜化が行われ始めた時期である。この頃からヒトは人口密度を増やし、新しい気候や食習慣、感染症に直面し、環境への適応や病気への耐性などに関連するたくさんの重要な遺伝的な変異がヒトに広まりはじめたと考えられる[23]。

正の選択の影響を受けた領域をヒトゲノムから検出する手法は多く研究されており、それらは、要約統計量、LD に基づく統計量、比較ゲノム、中立モデルの検定という 4 つのグループに分けられる。乳糖耐性の集団間の違いに関連する LCT 遺伝子や皮膚色素の集団間の違いに関連する SLC24A5 遺伝子の近年の正の選択は Long-Range Haplotype, integrated Haplotype Score (iHS), extended haplotype homozygosity (EHH), Cross Population EHH(XP-EHH) などの LD に基づく統計量の検定[23-25]によって検出された。また、2 型糖尿病の感受性アレルの集団間の違いは、iHS, XP-EHH などの LD に基づく統計量によって明らかにされた[6, 26, 27]。これらの手法はアフリカで誕生した現代人の祖先がアフリカ大陸以外に拡散した後の正の選択を検出することを目的として応用されてきた[24, 25]。

1.4 古い正の選択と疾患の関係

ありふれた疾患の感受性アレルの中でも疾患によっては特に集団間の分化の程度が高いものがあり、それらは自然選択の影響を受けたと考えられている。最近の研究では、1 型糖尿病・セリアック病などの自己免疫疾患やマラリア、敗血病の感受性アレルは近年の正の選択を受けたことが示された[26, 28]。また、肥満、2 型糖尿病、高血圧などの代謝疾患の関連遺伝子についても近年の正の選択を受けたことが報告されている[29]。

さらに最近の研究では、自然選択のなかでも古くに起こったものと近年に起ったものとがあり、古い自然選択を受けかつ集団で異なる近年の自然選択を受けるという長期的な進化プロセスが、自己免疫疾患などの免疫を介したありふれた疾患や代謝疾患の発症率の地域差の一因となっていると考えられるようになってきた[30](図 1-2)。ここで、古い、とは、アフリカ大陸からヒトが拡散する前の 100,000 年以上前のことを指す。古くからアフリカ大陸に存在していた病原体に対する適応も、現在の疾患に関連があり重要であると考えられるようになっている。例えば、睡眠病の病原体への適応のためにアフリカで起こった古い正の選択が、最近のアフリカ系アメリカ人の腎疾患の発症率の高さの一因となっていることが報告されている[31]。また、古くから起こってきた病原体に対する適応が、セリアック病、1 型糖尿病、多発性硬化などの自己免疫疾患の感受性座位の地域差に影響を与えたことが報告されている[32]。

図 1-2 古い自然選択と近年の自然選択

古い自然選択はヒトがアフリカ大陸から移動する前、近年の自然選択はヒトがアフリカ大陸から移動した後に起こった自然選択を表す。

さらに具体的な例としては、マラリアの病原体 *Plasmodium* spp.への適応がある。*Plasmodium spp.* は 100,000 年以上前にアフリカで発生した。マラリア耐性に関する遺伝子(例えば, FY, HBB, G6PD, SLC4A1)の SNP のアレルは現在、鎌形赤血球症、α-地中海貧血症、グルコース-6-リン酸脱水素酸素欠損症、楕円赤血球症などの血液疾患と関連することが報告されており、マラリアの頻度が高い地域では、これらの疾患が高頻度で観測されている。例えば、鎌形赤血球症は、赤血球内のヘモグロビンを構成するβグロビン鎖の変異によって起こる疾患であるが、この変異はもともとマラリアへの適応のための有利な変異で、古くに自然選択を受けたと考えられる[30, 33]。βグロビン鎖の変異によってヘモグロビンはヘモグロビン S と呼ばれるタンパク質となり、赤血球内で互いに凝集しやすく、赤血球を鎌形に変形させる。ヒトがマラリアに感染すると、マラリア原虫が二酸化炭素を産生し血液中の pH が下がり、ボーア効果によりヘモグロビンが酸素を解離しやすくなる。この時ヘモグロビン同士が凝集しやすく

赤血球が鎌形の方がマラリアに対して有利だったためこの形質が自然選択を受けたと考えられている。

しかしどのようなメカニズムで鎌形の赤血球がヒトをマラリアから守っていたのか詳細はいまだに分かっていない。もっともらしい説として、鎌形の赤血球内ではマラリア原虫の成長が妨げられた、もしくは鎌形の赤血球は脾臓での排除が促進された、などが挙げられている。このように多くのマラリア感受性座位が見つかっている一方で、具体的な免疫メカニズムの詳細についてはほとんど分かっておらず、古くからのマラリアに対する有利な形質が、現在の免疫疾患、炎症疾患などのありふれた疾患にどのような影響を与えているかは分かっていない[33]。

1.5 本論文の目的と構成

疾患関連形質の集団間の多様性についての情報は、公衆衛生、個別化予防の際の事前情報として重要であり、疾患関連形質の集団間の多様性に関連する座位をゲノムワイドに探索することが必要になっている。近年の正の選択(10,000年前以降)についての研究は数多く行われ、集団間の多様性に関連する遺伝子も数多く検出されている。一方で、最近の研究では、近年の正の選択に加えて、古く(100,000 年以上前)から起こってきた自然選択についても、最近の自己免疫疾患などの免疫を介したありふれた疾患や代謝疾患などの集団間の多様性を理解するために重要であると考えられるようになってきた。

マラリアに対する有利な遺伝子は現在、鎌形赤血球症の原因となっており、マラリア原虫 *P. falciparum* が存在する地域では balancing selection を受けヘテロ接合体としてある一定の頻度に保たれ、マラリア原虫が存在しない地域ではその遺伝子は減少している。この他にも、古くに起こった有利な形質に関連

する遺伝子が、現在の疾患、特にありふれた疾患の原因となり、古くにアフリカで正の選択を受けた後、集団の移動にともない、各地域の地理的条件などの相違により集団間で異なる近年の自然選択を受け、集団間の多様性に関与している例があるだろうか。また、独自に手法を開発することにより、今まで検出されにくかった、自然選択や多様性情報が得られ、現在のありふれた疾患へ与える影響、特に免疫メカニズムへの影響などがわかるだろうか。今のところ古い正の選択を検出する研究は少なく、これらのことはあきらかではない。

集団間の多様性が生じる可能性があるのは近年の自然選択のみを受けた領域(表 1-1(1))と古い正の選択を受けかつ近年の自然選択を受けた領域(表 1-1(3))であると考えられる。本研究で‘古くから起こってきた’として着目するのは後者の、古い正の選択を受けかつ近年の自然選択を受けたゲノム領域(表 1-1(3))である。現在多くの研究が行われている、ヒトがアフリカ大陸から拡散した後の近年の正の選択は、その中でも古い正の選択を受けたものと受けなかったものとがある(表 1-1 (1) (3))と考えられ、本研究の対象領域である。しかし、すでに検出されている近年の正の選択を受けた領域が、古い正の選択を受けた領域かどうかを判別する研究はまだ行われていない。また、近年に受けた自然選択が、正の選択以外の、purifying selection や balancing selection などだった場合、近年の正の選択を検出する研究では、古い正の選択を受けかつ近年の自然選択を受けたゲノム領域(表 1-1(3))は、検出できない可能性があると考えられる。

		近年の自然選択	
		あり	なし
古い正の選択	なし	(1)	(2)
	あり	(3)	(4)

表 1-1 古い正の選択と近年の自然選択の起こり方の可能な組合せ

そこで、本研究では、新たに HHD という距離の推定法を開発して、HHD を利用することにより、古い正の選択を受け、かつその後に多様化が起ったゲノム領域(図 1-3)を検出する手法を独自に構築した。ここで古いとは近年の正の選択以前の 10,000 年以上前のことを指す。本パイプラインでは、ゲノムワイドな SNP のうち、連鎖不平衡にあるもののなかから古い正の選択を受けた領域を検出する。また、古い正の選択の候補領域について、集団間で異なる近年の自然選択を受けてきたかどうかを評価することにより、近年の自然選択の検出感度の向上を試みた。そしてなかでも古くから病原体に適応するために正の選択を受けてきた遺伝子を見つけ、それらがどのように現在のありふれた疾患、特に免疫を介したありふれた疾患や代謝疾患の集団間の多様性に関与するのかを検討した。

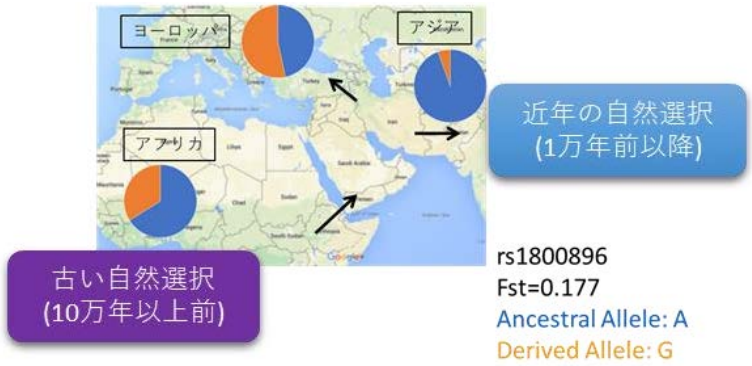


図 1-3 古い正の選択と近年の自然選択を受け多様化が起こったアレル

本研究ではこのような多様性を見せる感受性アレルを検出することを目的とする。

開発されたパイプラインでは、まず始めに古い正の選択の候補領域として古いハプロタイプブロックを抽出する。古い時期に正の選択を受けて生じたハプロタイプブロックのうち、ヒトの移住とともに広がった後もハプロタイプブロックが保存された領域を古いハプロタイプブロックと定義した(図 1-4(A)の点線で囲まれた領域)。パイプラインでは古いハプロタイプブロックを抽出した後、古いハプロタイプブロックが異なる近年の自然選択を受け多様化したかどうかを評価する。その際に集団間の関係性をネットワークで表現し、ネットワークの多様性により古いハプロタイプブロックの多様性を表す(図 1-4(B))。

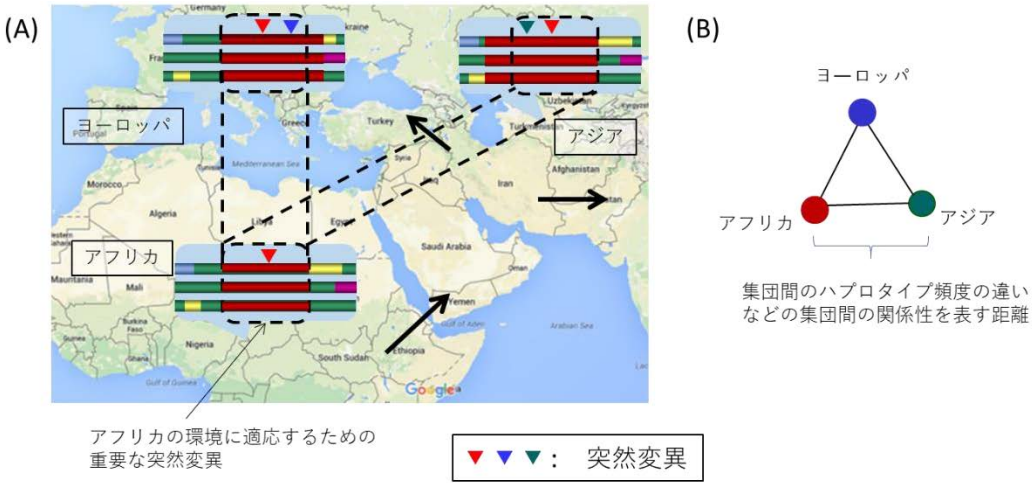


図 1-4 古い正の選択により生じた古いハプロタイプブロックと集団特異的な自然選択[34]

(A)古いハプロタイプブロックについて。古い時期にアフリカで環境に適応するための重要な変異が生じ(赤い三角)正の選択によりハプロタイプブロックが作られた。このハプロタイプブロックがヒトの移住とともに広がった。その後、古いハプロタイプブロックには、それぞれの環境で生じた新たな変異(青・緑の三角)などに対し、近年の自然選択が起こった。全集団のハプロタイプブロックのうち重複のある領域(点線で囲まれた)を古いハプロタイプブロックと定義する。

(B)古いハプロタイプブロックの多様性を表すために提案したネットワークモデル。それぞれのノードが各地域の集団を表す。赤、青、緑のノードがそれぞれアフリカ、ヨーロッパ、アジアの集団を表すことにする。エッジは集団間の関係性を表す。本研究では、集団間の関係性を、ハプロタイプ頻度の違いを使って表す t-統計量スコアを使って評価する。

次に、古くから集団によって異なる自然選択を受けてきたとされた古いハプロタイプブロックに含まれている遺伝子と SNP のアノテーションを行う。遺伝子のアノテーションは KEGG mapper を使って遺伝子を KEGG pathway にマップすることにより行われる。SNP のアノテーションは、まず NHGRI GWAS カタログを用いて、得られた SNP が何らかの形質との関連が報告されているかを探し、関連が見つかった SNP に関しては関連する遺伝子を用いて SNP を KEGG pathway にマップする。得られた遺伝子や SNP がすでに集団間の多様性が見られる疾患関連の経路、もしくは免疫を介したありふれた疾患との関連が想定される免疫システムや感染症関連の経路に含まれるかどうかを確認する。

本論文では、まず第 2 章で古い正の選択を検出するパイプラインで用いるディプロタイプ間距離の推定法について説明する。その際に、クラスタリングや機械学習への応用を通してその性質について述べる。第 3 章では第 2 章で推定法を提案したディプロタイプ間距離を利用して開発した古い正の選択を検出する手法について述べる。この手法をゲノムワイドな遺伝子型の公開データベースである HapMap に登録されている SNP データへ適用した結果についても併せて報告する。その際、本研究で得られた結果を既知の結果と関連させながら整合性について考察し、本研究の妥当性について検討する。最後に第 4 章において本研究で提案した手法について総括する。

第 2 章　ディプロタイプ間距離の推定法の提案

2.1 ディプロタイプの分類のためのディプロタイプ間距離

SNP データに基づくディプロタイプの分類は生物医学分野の研究において重要である[35, 36]。ここで、ディプロタイプとは、相が不明なディプロタイプ、あるいは相が明らかなディプロタイプ、どちらも含む。SNP データを用いてディプロタイプを分類するということは、そのディプロタイプを持つ集団内の個人を分類、グループ化することである。

一般的に、データを分類する手法はクラスタリングと呼ばれ、クラスタリング手法としては既存のものが多種存在している。それらのうち、例えばウォード法[37, 38]、k-medoids 法[39]、Density-based spatial clustering of application with noise (DBSCAN)[40]や、系統クラスタリング手法である近隣結合法[41]などは、クラスタリングする対象間の類似度を測る尺度が必要となる。正確な尺度を定義することはこれらの類似度に基づいたクラスタリングには必須である。

一般的に SNP アレイを使って検出される SNP の場合、ヘテロ接合体でかつ家系情報が得られない場合は相が不明で、相が不明なディプロタイプのクラスタリングに関する研究が数多く行われている。相が不明なディプロタイプのクラスタリングの中で距離に基づく手法の多くは、Allele Sharing Distance (ASD)[36, 42-44]と呼ばれる距離を用いた手法である。ASD は基本的には、距離として一番簡単だと考えられるハミング距離を拡張した、相が不明なディプロタイプ間の距離である。

遺伝解析では、遺伝的に異なる集団間では、集団の性質について考慮することが非常に重要である[44-46]。相が不明なディプロタイプ間の距離を設計する際にも同様のことが言える。しかし、すでに多く用いられている ASD は、距離

の計算の際、そのような集団の情報を用いていない。そこで第 2 章では、集団の情報を用いた、相が不明なディプロタイプ間の距離 Population model-based distance (PMD)を提案する。そして、実データへの適用により、PMD の計算の際、ハプロタイプ間の類似度をそのハプロタイプが属する集団における推定頻度で重み付けして使うことが、クラスタリングの精度にどのように影響を与えるかを ASD との比較により検証する。

2.2 先行研究

2.2.1 Allele Sharing Distance (ASD)

ASD は最もよく用いられている標準的なディプロタイプ間の距離である[36, 42, 44]。定義は以下に示す。

相が不明なディプロタイプ、$g, g' \in D^m$（mは SNP 数とする）に対して、ディプロタイプgとg'間の ASD は以下のように定義される:

$$D(g, g') = \frac{1}{2m}\sum_{l=1}^{m} d(g[l], g'[l]),$$

ここで$g[l]$は、相が不明なディプロタイプgのl番目の遺伝子型で、$d(g[l], g'[l])$は、gとg'のl番目の座位で共通でもっていないアレルの数とする。

2.2.2 ハプロタイプ間距離

ハミング距離[47-51]はハプロタイプのような DNA 配列間の最も一般的で単純な類似度である。ハプロタイプ$h \in S^m$（ここで、mはhの長さとする）に対して$h[k]$を、hのk番目の座位のアレルとする。2 つのハプロタイプhとh'間のハミング距離は、

$$s(h,h') = \sum_{k=1}^{m} I(h[k], h'[k]),$$

と定義される。ここで、

$$I(a,b) = \begin{cases} 0\ (a = b) \\ 1\ (a \neq b) \end{cases}$$

とする。ハミング距離は、ハプロタイプの長さに依存する距離なので、ハプロタイプの長さに依存しない距離とするため、以下のようなハプロタイプhとh'の間の距離、$A(h,h')$を定義する:

$$A(h,h') = \frac{s(h,h')}{m}.$$

2.2.3 HIT アルゴリズム

Haplotype Inference Technique (HIT)アルゴリズム[52]は隠れマルコフモデル(Hidden Markov Model;HMM)[53]に基づく、相が不明なディプロタイプの相を明らかにするためのアルゴリズムである。図 2-1 に、HIT で使われる HMM を示す。HIT アルゴリズムの HMM は、複数の祖先集合を仮定して設計されている。HMM は相が不明なディプロタイプの集合を用いて、EM アルゴリズム[54]を使った、教師なし学習法によりトレーニングされる。HIT アルゴリズムは、相が不明なディプロタイプから、ヒューリスティックに HMM からの出力確率が最も高いディプロタイプを探し、それを解とする。

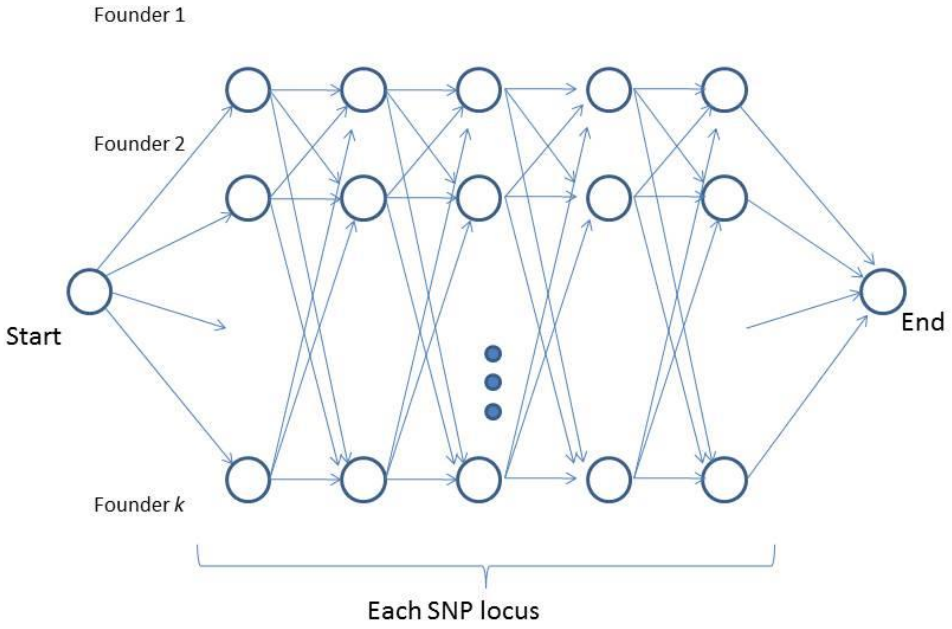


図 2-1　HIT アルゴリズムの HMM[55]
各行のノードはそれぞれ祖先を表わし、各列のノードは SNP を表わす。それぞれのノードからメジャーアレルとマイナーアレルのいずれかが出力される。スタートノードからエンドノードまでのパスはハプロタイプを表わす。祖先の数 k は、HIT を実行する際に指定する。本研究では開発者が推奨する k=7 として実行した。

2.3 ディプロタイプ間距離の定義

本節では、まず始めに、2.3.1 で、相が不明な 2 つのディプロタイプ間の距離、Population Model-based Distance(PMD)を新たに提案する[55]。PMD は集団の情報を集団モデルとして使い計算する距離の総称であり、その 1 つの実装として、HIT アルゴリズムで予測した HMM を集団モデルとして用いる距離、HIT HMM-based distance (HHD)を 2.3.2 で定義する。2.4 では HHD の性質につい

て議論する。

2.3.1 Population Model-based Distance (PMD)

新しく PMD を定義する前に、まず始めに、2.2.2 で説明したハプロタイプ間距離を以下のように拡張して、ディプロタイプ間の距離を扱うことができるようにする。

$a = \{h_1, h_2\}$ と $a' = \{h_1', h_2'\}$ を比較対象の 2 つのディプロタイプとする。ここで、$h_1, h_2, h_1', h_2' \in S^m$とする。このとき、ディプロタイプ$a$と$a'$間の距離を本研究では以下のように定義した。

$$H(a, a') = min\left\{\frac{A(h_1, h_1') + A(h_2, h_2')}{2}, \frac{A(h_1, h_2') + A(h_2, h_1')}{2}\right\},$$

ここで、Aは 2.2.2 で定義されたハプロタイプ間距離である。

ここで定義されたディプロタイプ間距離は、ディプロタイプの相が明らかな時にはそのディプロタイプに適用し、計算することができるが、相が不明なディプロタイプに対しては適用できない。そこで、このHを、相が不明なディプロタイプにも適用できるように、集団モデル(population model) Mを用いて拡張する。

任意の相が不明なディプロタイプに対して、可能なディプロタイプの候補を列挙することができる。例えば、相が不明なディプロタイプ$\{1,0\} - \{1,0\} - \{1,0\}$には、4 つのディプロタイプの候補がある。つまり、$\{111,000\}$、$\{110,001\}$、$\{101,010\}$、$\{011,011\}$である。この中から、何らかの集団の情報を用いて、最も可能性の高いディプロタイプを見つける。

相が不明なディプロタイプ$g, g' \in D^m$に対して、$c_i = \{h_{i1}, h_{i2}\}\ (1 \le i \le M)$と$c'_j = \{h_{j1}', h_{j2}'\}\ (1 \le j \le M')$を$g$と$g'$それぞれの、$i$番目と$j$番目のディプロタイプの候補とする。$M$と$M'$は$g$と$g'$のディプロタイプの候補数とする。ここで何らか

の集団モデル Mが与えられたとすると、相が不明なデータgに対して、ディプロタイプの候補cが、与えられたgの、本当の候補である確からしさを、確率$Prob(c|g,M)$として計算することができる。$p_i = Prob(c_i|g,M)$ と $p_j' = Prob(c_j'|g,M)$ をモデルMと仮定したときの、ディプロタイプc_iとc_j'の条件付き確率とする。このとき、ディプロタイプgとg'間のPMD_Mを以下のように定義する:

$$PMD_M(g,g') = \sum_{i=1}^{M}\sum_{j=1}^{M'} H(c_i,c_j')\cdot q_i \cdot q_j',$$

ここで$q_i = p_i/\sum_{k=1}^{M} p_k$、$q_j' = p_i'/\sum_{k=1}^{M'} p_k'$である。$q_i$と$q_j'$はディプロタイプの候補$c_i$と$c_j'$の条件付き確率の推定値で、標準化された値である。このときPMDは、集団モデル Mの仮定のもとでのディプロタイプの候補間距離$H(c_i,c_j')$の期待値となっている。集団モデル Mとしてハプロタイプ頻度が与えられている場合、q_iとq_j'はハプロタイプ頻度から計算されるディプロタイプ頻度となる。

2.3.2 HIT HMM-based Distance (HHD)

2.3.1 において、PMD を計算する際、なんらかの適当な集団モデルが必要であった。以下に、その実装例として、HIT アルゴリズムで予測した HMM を集団モデルとして用いる距離、HIT HMM-based Distance (HHD) を定義する。

まず始めに与えられたすべての相が不明なディプロタイプを用いて HIT アルゴリズムにより HMM をトレーニングする。そして、集団モデルM^*を HIT アルゴリズムでトレーニングして予測した HMM として、HHD を以下のように定義する。

$$HHD(g,g') = PMD_{M^*}(g,g').$$

ここで、それぞれのディプロタイプ候補の確率は、予測された HMM の条件付き出力確率として、HMM の forward アルゴリズム[54]により計算される。

2.3.3 PMD と複数祖先仮説

ヒトゲノムの多くの領域、特に機能的に重要な遺伝子を含む領域では、集団中のハプロタイプの多様性が低くなり、少数のカテゴリーにしか分類されない（ハプロタイプブロックと呼ばれる）[56, 57]。このことは、これらの領域では、少数の祖先ハプロタイプが集団に広まったことを示唆している。この、少数のハプロタイプが祖先として存在するという仮説は、非常に有用な仮説で、様々な研究に用いられている。例えば、連鎖不平衡マッピング[58-60]や集団の進化解析[61, 62]などである。

PMD は祖先ハプロタイプの存在を反映することができる距離である。図 2-2 で例を示す。この例では 3 人のディプロタイプがあり、それぞれ、$a = \{1011,0110\}, b = \{1101,0110\}, c = \{1111,1000\}$とする。今、これらは相が不明なディプロタイプであるとする。つまり、$\{1,0\}-\{1,0\}-\{1,1\}-\{1,0\},\ \{1,0\}-\{1,1\}-\{1,0\}-\{1,0\}, \{1,1\}-\{1,0\}-\{1,0\}-\{1,0\}$ としか分かっていないとする。

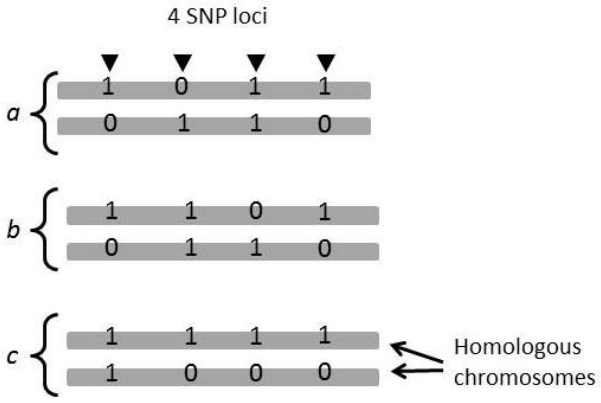


図 2-2　ASD と PMD の違いがみられるディプロタイプの例[55]

表 2-1 はa、b、cの祖先ディプロタイプを図 2-3 のように仮定した際のa、b、c間の変異数を示している。そもそも 2 つの配列間の距離は、それらの配列間の点変異の数に基づいて計算される。つまり、2 つの配列間で多くの変異があればあるほど配列間の距離が離れていくと考える。変異数は、複数の祖先ハプロタイプの存在を仮定して、定義することもできる。祖先ディプロタイプを介して、aとb間の変異数は 2 となる。祖先ディプロタイプとc間の変異数は 3 であり、祖先ディプロタイプを介したaとc間の変異数は 4 となる。同様に、bとc間の変異数も 4 となる。これらの変異数は、異なる 2 種類のハプロタイプからなる祖先ディプロタイプを仮定することにより、a、b、cは図 2-4 のようにクラスタリングすることができることを示している。クラスタリングされたaとbは、同じハプロタイプ0110を共有していて、この図のクラスタリングの結果は妥当であると考えられる。

表 2-1　図 2-2 のディプロタイプ間の変異数[55]

	a	b	c
a	0	2	4
b	-	0	4
c	-	-	0

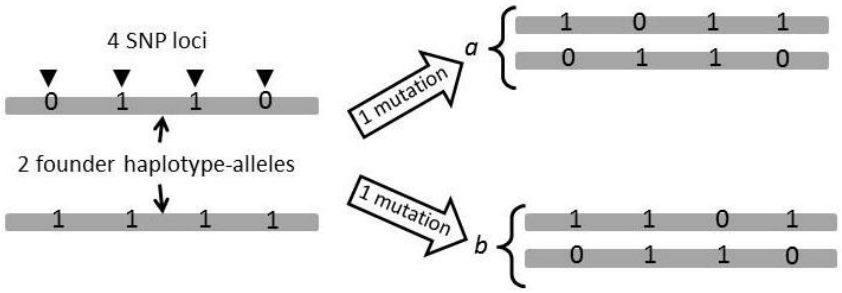


図 2-3　図 2-2 のaとbに対する最適な祖先ハプロタイプ[55]

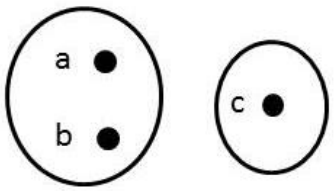


図 2-4　図 2-2 のディプロタイプの変異数に基づいた$H = PMD_{M_1}$とPMD_{M_2}のクラスタリングの結果[55]

一方で、ASD を用いることでは得られない。

ディプロタイプ間の距離Hは、表 2-1 の変異数を上手く表現できている。aとb間のHは 0.25、bとc間のHは 0.5 で(表 2-2(2))、図 2-4 のように 3 人をクラスタリングすることができる。一方で、これらa、b、cの 3 人からの任意の 2 人の間の ASD はすべて0.25であり(表 2-2(1))、祖先ディプロタイプを仮定した時に妥当だと考えられる図 2-4 のようなクラスタリングの結果を得ることが出来ない。少なくともこの例のような祖先ハプロタイプの存在を仮定すると、Hは ASD より適切であるということができる。

次に 2 つの集団モデル、M_1、M_2(表 2-3)が与えられたときの PMD を考える。集団モデルとして集団内に存在するハプロタイプの頻度がそれぞれ与えられている。モデルM_1の仮定のもとでは、100%の信頼度をもって相が不明なディプロタイプを相付けすることができ、結果としてPMD_{M_1}はHと一致する（表 2-4 と表 2-2(2)を参照)。モデルM_2では、それぞれの相が不明なディプロタイプに対して、複数のディプロタイプの候補が存在する（表 2-4 と表 2-2(3)を参照)。

3 人を$PMD_{M_1}(= H)$に基づいてクラスタリングすると、図 2-3 と同じクラスタリングの結果を得ることができる。さらに、PMD_{M_2}を用いても同じクラスタリングの結果を得ることができる。よって、もしある適当な集団モデルを与えることができれば、複数祖先仮説のもとでは、PMD は祖先ハプロタイプを反映するという意味で ASD より適切な距離ということができる。本研究では、PMD として、2.3.2 で定義した HHD を用いる。具体的には、集団モデルとして HIT アルゴリズムで予測された HMM からハプロタイプ頻度を抽出し使用する。

表 2-2 図 2-2 のディプロタイプ間の距離[55]

(1)ASD

	a	b	c
a	0	0.25	0.25
b	-	0	0.25
c	-	-	0

(2)$H = PMD_{M_1}$

	a	b	c
a	0	0.25	0.5
b	-	0	0.5
c	-	-	0

(3)PMD_{M_2}

	a	b	c
a	0	0.301	0.450
b	-	0	0.500
c	-	-	0

表 2-3　ハプロタイプ頻度で与えられた集団モデル[55]

	集団内のハプロタイプ頻度	
ハプロタイプ	(i) M_1	(ii) M_2
1111	0.40	0.20
1110	0.00	0.07
1101	0.20	0.08
1011	0.25	0.10
0011	0.00	0.05
0110	0.10	0.30
0101	0.00	0.05
1100	0.00	0.05
1000	0.05	0.10
その他	0.00	0.00

表 2-4　図 2-2 のディプロタイプの候補ディプロタイプの表 2-3 で与えられた集団モデルの下での条件付確率[55]

	相がわかっていない		条件付き確率	
個人	ディプロタイプ	ディプロタイプ候補	(i) M_1	(ii) M_2
a	{1,0}-{1,0}-{1,1}-{1,0}	{1011, 0110}	1.0000	0.8955
		{1110, 0011}	0.0000	0.1045
		その他	0.0000	0.0000
b	{1,0}-{1,1}-{1,0}-{1,0}	{1101, 0110}	1.0000	0.8727
		{1110, 0101}	0.0000	0.1273
		その他	0.0000	0.0000
c	{1,1}-{1,0}-{1,0}-{1,0}	{1111, 1000}	1.0000	0.8000
		{1011, 1100}	0.0000	0.2000
		その他	0.0000	0.0000

2.4 連鎖平衡・不平衡による HHD の評価

ここでは本研究で新たに推定法を提案したディプロタイプ間距離 HHD の性質を、HHD を使った手法の精度と連鎖平衡・不平衡との関連から検討する。また既存のディプロタイプ間距離 ASD との比較についても検討する。

2.4.1 データセット

HapMap データセット

HapMap release24[63]に登録されていたヒトの 22 本の常染色体の相が不明なディプロタイプデータを使った。このデータセットは、270 人の相が不明なディプロタイプデータからなり、その内訳は以下の通りである。 90 人のナイジェリア、イバダンのヨルバ族(YRI)、90 人の北ヨーロッパや西ヨーロッパ由来の米国ユタ州の住民(CEU)、45 人の、東京の日本人、45 人の、北京の漢民族系中国人(ASN)。本研究では YRI、CEU、ASN をそれぞれアフリカ、ヨーロッパ、アジアの集団とする。本研究で用いたのは全 3,976,554 SNP のうち 3 集団が共通に持っている 3,619,226 座位(図 2-5)で、さらにそのうちの 270 人全員のデータが存在する 879,657 座位を最終的な解析対象とした。これらの SNP は、連続した 100 SNP ごとの 8,930 ブロックに区切られ、ASD と HHD のクラスタリングの精度の比較に用いられた。

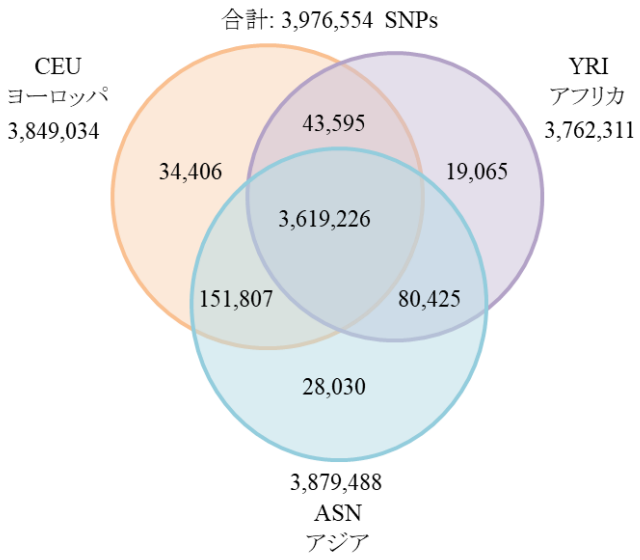


図 2-5　HapMap の SNP についてのベン図[34]
このベン図は本研究でダウンロードした HapMap の 22 本の常染色体上の SNP 数を表している。

NAT2 データセット

HapMap の 879,657 SNP から NAT2 遺伝子が存在する領域の SNP を抽出し、HHD を使った学習アルゴリズムの精度の検証を行った。NAT2 遺伝子は、薬物代謝の際のアセチル化の速度に関連する変異を複数持ち、変異を含むハプロタイプのパターンはヨーロッパとアジアでは異なる。NAT2 がアセチル化の速度に関わるリウマチの治療薬では、アセチル化が遅いと副作用がおこることが示されており[64]、薬の投与量の決定の際に NAT2 のハプロタイプの集団間の違いを考慮することの重要性が示されている。

疾患関連データセット

また、クローン病[65]と自己免疫疾患[66]に関連する領域の SNP を用い HHD を使った学習アルゴリズムの精度の検証を行った。クローン病関連領域のハプ

ロタイプブロックはSNPを使って高精度に検出された初めてのハプロタイプブロックで、ゲノム上の領域は5q31であると報告されている。用いたのは、129のトリオのデータである。すべての子供はケース集団に属し、すべての両親はコントロール集団に属する。ケース集団は144人、コントロール集団は計243人となる。

自己免疫疾患に関連する領域は、CD28、CTLA4、ICOSなどのタンパク質コーディング領域であると報告されている。これらの領域、330kbをシーケンシングし、384人のケース、652人のコントロールについて108SNPが得られている。

シミュレーションデータセット

GeneArtisan[67]はCoalescentモデルを使った遺伝子型シミュレーションデータ作成ツールである。GeneArtisanを用いてなんらかの疾患関連領域のSNPのシミュレーションデータを作成した。計1000人分のディプロタイプを作成し、そのうち500人をケース、500人をコントロールとした。これらのデータについても、HHDを使った学習アルゴリズムの検証に用いられた。

2.4.2 ASDとHHDを使ったクラスタリングの精度の比較

ここでは、HHDが集団モデルを仮定することにより、ASDと比べてどのようにクラスタリングの精度を上げたのかをHapMapデータセットを使ってみていく。クラスタリングの評価にはClassification Error Rate (CER)[68]を使った。CERは、ASN、CEU、YRI、それぞれ同じ集団に属する人同士がどれだけ同じクラスターに含まれているかを表す指標である。もともと同じ集団に属する人が同じクラスターに多く含まれていた時、それ以外の集団に属する人がそのク

ラスターにどれだけ含まれていたかを割合として表す。

2.4.1 で HapMap データセットから作成した連続した 100 SNP からなる 8,930 ブロックに対して 270 x 270 の距離行列を ASD と HHD を使って計算し、Ward 法によるクラスタリングを行った。Ward 法によるクラスタリングは統計ソフト R の hclust コマンドによって実行する。Ward 法は階層型クラスタリングであるので、結果として得たいクラスター数を指定して樹形図をクラスター分割する必要がある。樹形図の分割は統計ソフト R の cutree コマンドによって実行する。本研究ではクラスター数を 3 として Ward 法で作成した樹形図を分割し、3 つのクラスターを得た。cutree コマンドを実行するとそれぞれの要素がどのクラスターに属するかの情報を得ることが出来るので、その情報を用いて CER を計算し、ASD と HHD の結果の比較を行った(図 2-6、図 2-7)。

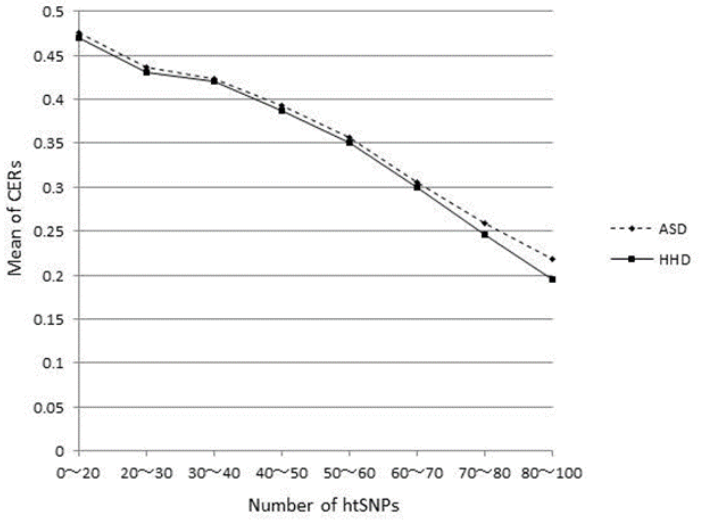


図 2-6　CER の平均値とh_Bの関係のプロット[55]

点線と直線はそれぞれ ASD と HHD の結果を表している。それぞれのブロック中の htSNP(tagSNP)の数をh_Bとする。x~yは$x \leq h_B < y$を意味する。HHD は ASD よりいつも優れた結果を出していることがわかる。

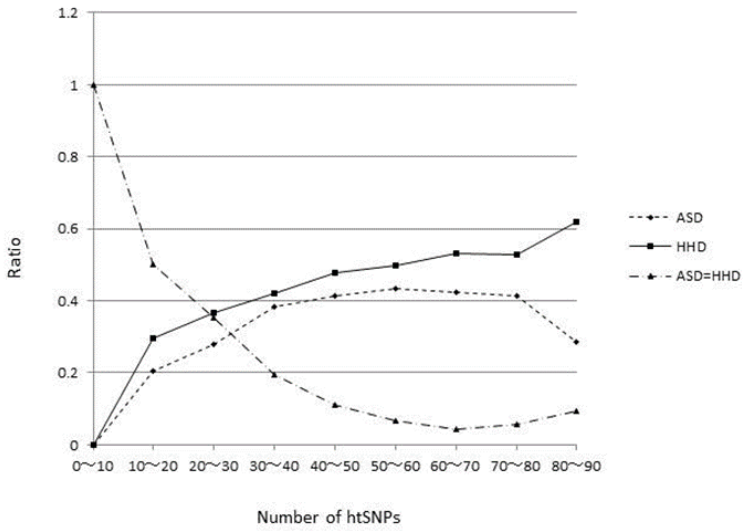


図 2-7　h_Bと成功の割合の関係のプロット[55]

それぞれのブロック中の htSNP(tagSNP)の数をh_Bとする。点線と直線はそれぞれ ASD と HHD の結果を表している。ASD=HHD を表す線は、ASD と HHD のパフォーマンスが全く一致していたことを示す。

まず始めに CER の平均値の比較を行った(図 2-6)。HHD を用いたクラスタリングの CER の平均値は 0.356、ASD を用いたクラスタリングの CER の平均値は 0.361 であり、これらの平均値の差が有意であるかどうか t 検定を行ったところ、p-値は 0.004 で、HHD を用いたクラスタリングは ASD を用いたクラスタリングよりも有意に精度が良いことが分かった。

また、符号検定による CER の比較も行った。ここで、ASD の CER が HHD の CER より低いとき、'ASD の成功'といい、HHD の CER が ASD の CER より低いとき、'HHD の成功'ということにする。まず、各ブロックで ASD、HHD それぞれによるクラスタリングの CER を比較した。すると、8,930 ブロック中、4,366 ブロックで ASD より HHD を用いた CER が低かった。また、3,969 ブロックについては HHD より ASD を用いた CER が低かった。残りの 868 ブロックでは ASD と HHD で CER は同じになった。これらの符号検定の p-値は$8.98 \cdot 10^{-14}$で、HHD が ASD より有意な頻度で良い結果を出していることが分かった。

それぞれのブロック中の htSNP(tagSNP)の数h_Bと ASD と HHD の違いの現れ方の関係についても比較を行ったところ、ASD、HHD いずれの距離に基づいた結果においても、h_Bが増加すると、CER が減少することが分かった(図 2-6)。また、h_Bが増加すると、ASD と HHD に基づいた距離の CER の差が広がっていく。h_Bが増加するとき、HHD の成功率が上がり、また ASD の成功率も上がる。しかし、h_Bが増加するとき、ASD と HHD の成功率の差も増加する。一方で、h_Bが増加するとき、ASD と HHD が同じ結果を得る場合は減少する。

図 2-7 では$80 \leq h_B < 90$のとき、HHD は ASD より精度が良い。この結果は、もし、情報が多ければ（すなわちゲノム領域に tagSNP が多いとき)、クラスタリングの精度が上がるということで、これは妥当な結果といえる。ASD と HHD の成功率の差も$80 \leq h_B < 90$のとき最大になる。このとき、13 ブロックで HHD

が成功し、残りの 18 ブロックのうち、6 ブロックでのみ ASD が成功していた。

2.4.3 HHD を使った機械学習の精度について

2.3.2 で定義した HHD を用いて、Support Vetor Machine (SVM) のカーネルを定義することができる。HHD の SVM のカーネルへの応用を通して HHD の精度と連鎖平衡・連鎖不平衡との関連を見て行く。

2.4.3.1 HHD を用いた SVM の学習アルゴリズム

1. HHD の計算

学習データとして選んだ相が不明なディプロタイプ間 HHD を計算する（詳細は 2.3.2 を参照)。

2. カーネルの計算

1.で計算した HHD に基づいて 2 通りの方法でカーネル行列を計算する。

2.1 $e^{-HHD(g,g')}$をカーネルとする。これを、自然対数 HHD と呼ぶことにする。

2.2 $1-HHD(g,g')$をカーネルとする。これを、線形 HHD と呼ぶことにする。

2.1 と 2.2 のカーネルを使い、270×270 のカーネル行列を作成する。大部分の場合、自然対数 HHD と線形 HHD を使ったカーネル行列は半正定値となる。半正定値にならない場合は、以下の処理を行う。

Kをカーネル行列とする。Kは、$K = PMP^{-1}$とかける。ここでMは三角行列で対角成分がKの固有値、Pは直行行列とする。本アルゴリズムでは、Mの負の対角成分を 0 で置き換え、M'とする。そして、$K' = PM'P^{-1}$をカーネル行列として用いる。

3. カーネル行列を使った SVM の学習

2.で計算したカーネル行列を使い、SVM[69]を学習させる。

Gist2.3(www.chibi.ubi.ca)を使うとカーネル行列を指定してSVMを学習させることが出来る。学習コマンドgist-train-svmを実行する際に、-matrixオプションをつけると、-trainで指定したファイルのカーネル行列を使って学習することができる。本研究では、2で自然対数HHDと線形HHDを使って作成したカーネル行列を新たなカーネル行列として提案し、学習の際に使用した。

4. SVMを使った予測

3で学習させたSVMを使ってテストデータを分類する。Gist2.3のgist-classifyコマンドを使用した。

2.4.3.2 NAT2・疾患関連・シミュレーションデータセットへの適用結果

手法の精度の検証では3-fold validationの結果を比較する。比較対象として、Brinzaの方法[70]、SVM-Fisher[71]、メジャーアレルを使った方法を選んで提案手法との比較を行った。SVM-FisherはHMMに基づいた手法であるが、ここではHHDの定義で用いたHITアルゴリズムのHMMを用いた。メジャーアレルを使った手法では、各SNPのメジャーアレルの数を数えてそれぞれのディプロタイプごとに特徴ベクトルを作成した。

連鎖不平衡が強いデータの結果

NAT2遺伝子領域におけるSNPに対する結果では、CEUとYRIのデータにおいて提案手法がその他の手法より精度が良かった(表2-5、表2-6)。またシミュレーションデータに対する結果でも、提案手法が他の手法より精度が高かった(表2-7)。これらのデータについて詳しく調べると、NAT2領域ではCEUとYRIのどちらの集団でも連鎖不平衡が強いことが分かった。また、シミュレーションでは、SNPがハプロタイプブロックを構築するようにデータが作成され

ている。連鎖不平衡が強く、ハプロタイプブロックの構造がはっきりする場合、ハプロタイプ推定の精度が上がり、その結果 HHD の計算精度も上がり、結果として新たに提案したカーネルを用いた予測手法の精度が上がるのではないかと考えられる。

表 2-5　CEU データセットの 3-fold cross-validation の結果[72]

Method	Sensitivity	Specificity	Accuracy
Exp HHD	0.822	0.828	0.832
Linear HHD	0.822	0.811	0.815
Major allele	0.811	0.800	0.804
SVM-Fisher	0.678	0.728	0.711
Brinza's method	0.733	0.755	0.748

表 2-6　YRI データセットの 3-fold cross-validation の結果[72]

Method	Sensitivity	Specificity	Accuracy
Exp HHD	0.722	0.939	0.867
Linear HHD	0.722	0.939	0.867
Major allele	0.733	0.911	0.863
SVM-Fisher	0.489	0.861	0.737
Brinza's method	0.667	0.856	0.793

表 2-7　シミュレーションデータセットの 3-fold cross-validation の結果[72]

Method	Sensitivity	Specificity	Accuracy
Linear HHD	0.960	1.000	0.980
Major allele	0.970	0.920	0.945
SVM-Fisher	0.980	0.620	0.800

連鎖平衡に近いデータの結果

図 2-8 に見られるように、NAT2 領域の ASN のデータは、他の集団より連鎖平衡に近い。そのため、提案手法は、ASN のデータでは予測精度が他の手法に比べてそれほど良くない原因となっていると考えられる(表 2-8)。提案手法は、疾患関連データセットにおいても精度が良くなかった(表 2-9、表 2-10)。これについても、ゲノム上のクローン病の関連領域は長い領域に渡り、その中でハプロタイプブロックが 3 つの領域に分断されていて、全体として連鎖不平衡の度合いが低くなっているためと考えられる。

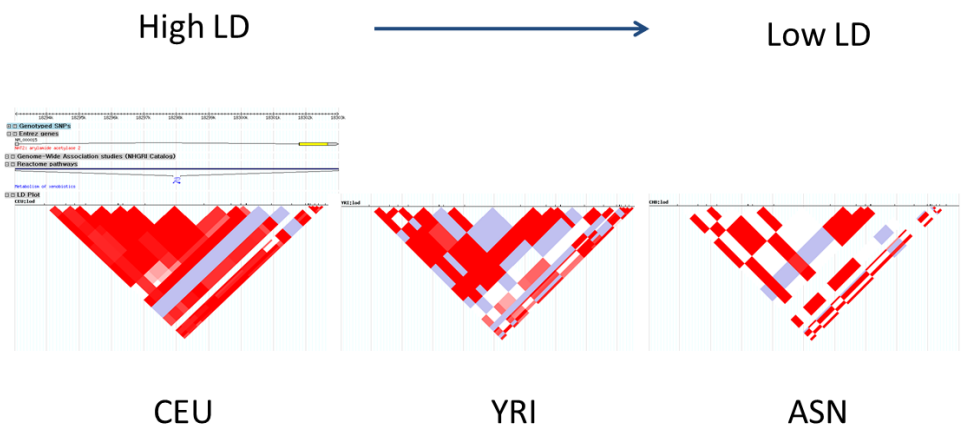


図 2-8　Haploview[20]を使った HapMap の SNP の NAT2 遺伝子領域における LD プロット[72]

赤色は 2 つの指標で強い LD が検出された、紫色は 1 つの指標で強い LD が検出された、白色は 2 つの指標で弱い LD が検出された、または LE の領域を示す。

表 2-8　ASN データセットの 3-fold cross-validation の結果[72]

手法	Sensitivity	Specificity	Accuracy
Exp HFD	0.911	0.800	0.832
Linear HFD	0.911	0.783	0.826
Major allele	0.922	0.744	0.804
SVM-Fisher	0.767	0.728	0.778
Brinza's method	0.911	0.800	0.837

表 2-9　クローン病データセットの 3-fold cross-validation の結果[72]

手法	Sensitivity	Specificity	Accuracy
Exp HFD	0.468	0.603	0.543
Linear HFD	0.561	0.536	0.545
Major allele	0.574	0.456	0.496
SVM-Fisher	0.668	0.352	0.447
Brinza's method	0.388	0.575	0.501

表 2-10　自己免疫疾患データセットの 3-fold cross-validation の結果[72]

手法	Sensitivity	Specificity	Accuracy
Exp HFD	0.530	0.521	0.525
Linear HFD	0.451	0.692	0.603
Major allele	0.425	0.684	0.587
SVM-Fisher	0.493	0.524	0.515
Brinza's method	0.601	0.570	0.580

第 3 章　ディプロタイプ間距離 HHD を用いた古い正の選択の検出方法の開発

3.1 古い正の選択について

ヒトは、古くはヒトがアフリカ大陸から拡散する前からマラリアなどの感染症からヒトを守るため、様々な病原体(ウィルス、細菌、菌類、原虫、昆虫、節足動物、寄生蠕虫など)に対し適応してきた。病原体は、年代や場所により異なり、現在のヒトは、どのような歴史を持つかによって病原体への耐性に関連するゲノム領域の変異は異なると考えられる。自然選択のなかでも古くに起こったものと近年に起ったものとがあり(表 1-1)、古い自然選択を受けかつ集団で異なる近年の自然選択を受けるという長期的な進化プロセスが、自己免疫疾患などの免疫を介したありふれた疾患や代謝疾患の集団間の形質の違い、例えば発症率の地域差の一因となっていると考えられるようになってきた[30]。しかし現在、近年の正の選択の検出手法は数多く研究され、集団間の形質の違いに関連する遺伝子も数多く検出されているものの、古い正の選択を検出する研究は少ない。

実際に、そのような進化プロセスを経て集団間の違いを持つようになったゲノム領域があるのだろうか、また、そのようなゲノム領域を検出する手法を開発することにより、これまで見出しにくかった集団間の多様性に関連する情報を得ることができるだろうか。このような観点から、本研究では、まず HHD という新たな距離の推定法を開発し、HHD を利用することによりパイプラインを構築した。本パイプラインでは、古い正の選択を受け、かつその後多様化が起こったゲノム領域を検出する。まず、ヒトがアフリカ大陸から拡散する前の古い時期に正の選択を受けた領域を、古いハプロタイプブロックとして抽出する。

そしてヒトがアフリカ大陸から移住する際に古いハプロタイプブロックが広がった後、近年の自然選択が起こったかどうかを評価する(図 1-4(A))。そして、古い正の選択の候補領域について集団間で異なる近年の自然選択を受けてきた領域かどうかを評価することで、近年の自然選択の候補領域の検出感度の向上を試みた。古い正の選択を受け、その後多様化が起こったとされた古いハプロタイプブロックに含まれる遺伝子や SNP は機能アノテーションされ、すでに集団間の多様性が知られている疾患の経路や、免疫を介したありふれた疾患との関連が想定される免疫システムや感染症の経路における機能について検討される。

3.2 古い正の選択を検出するパイプライン

本研究では、古い正の選択を検出するパイプライン(図 3-1(A))と機能アノテーション(図 3-1(B))を提案する。まず古い正の選択を検出するパイプラインについて 3.2.1 では古いハプロタイプブロックの抽出、3.2.2 ではディプロタイプ間距離 HHD の計算、3.2.3 では集団間の違いを持つ古いハプロタイプブロックの抽出、3.2.4 では古いハプロタイプブロックのクラスタリングについてそれぞれ手順ごとに説明する。

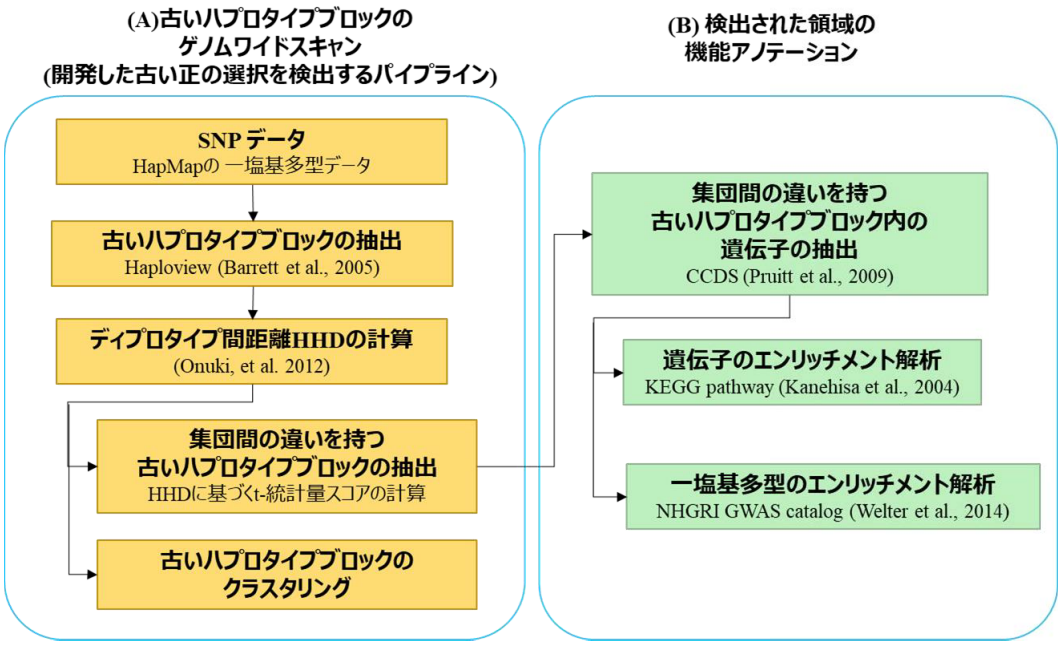


図 3-1　提案するパイプラインと機能アノテーションの概観図[34]
(A)HHD を使った古い正の選択を検出するパイプラインの手順。
(B)機能アノテーションの手順。それぞれの Box にはそれぞれの手順で使ったデータやツールが書かれている。

3.2.1 古いハプロタイプブロックの抽出

古い正の選択の候補領域を抽出するために、まず、Haplview4.2[20]を使ってそれぞれの集団ごとにハプロタイプブロックを検出する。次に、YRI 集団のハプロタイプブロックの中で CEU と ASN 両方のハプロタイプブロックと重複するものを抽出する。最初と最後の SNP のゲノム上の位置がi(bp)、j(bp)となるハプロタイプを$H[i..j]$とする。2 つのハプロタイプ、$H1[i..j]$ と $H2[k..l]$は、$i \le k \le j \le l$、 $k \le i \le l \le j$、 $i \le k \le l \le j$ または $k \le i \le j \le l$の時、重複があると考える。抽出された YRI 集団のハプロタイプブロックはヒトの移住とともに拡散した古い正の選択が起こった候補領域とみなす(図 3-2)。

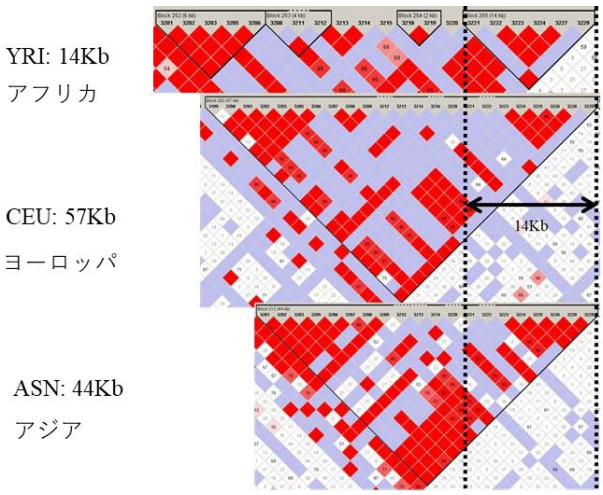


図 3-2 古い正の選択の候補領域の例[34]

黒い三角形で囲まれた領域は検出されたハプロタイプブロック、黒い点線で囲まれた領域は古いハプロタイプブロックを表す。

そして、この古い正の選択の候補領域のハプロタイプブロックを古いハプロタイプブロックと定義する。古いハプロタイプブロックを抽出するために、本パイプラインでは 3 集団共通のハプロタイプブロックを抽出する。本研究では、3 集団すべての遺伝子型データを使って Haploview を実行して検出されたハプロタイプブロックを共通のハプロタイプブロックとする。抽出された共通のハプロタイプブロックが、実際に古い正の選択の影響を受け、それぞれの集団に存在するかどうかを評価するために、さらに、共通のハプロタイプブロックが古い正の選択の候補領域と重複があるかどうかを確認し、重複があったものを最終的な古いハプロタイプブロックのセットとした(図 3-3)。

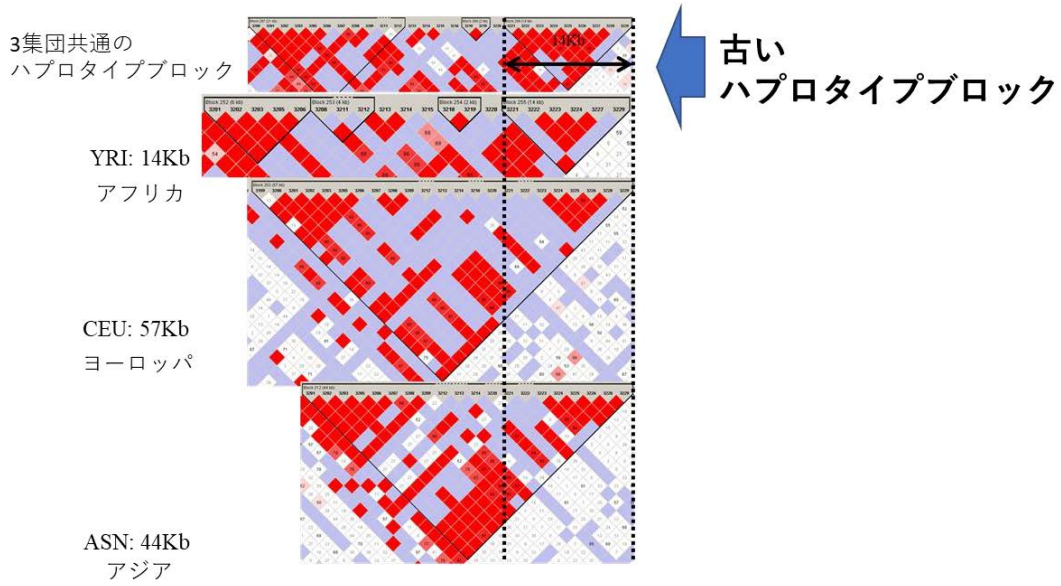


図 3-3 古いハプロタイプブロックの例[34]
3集団(YRI, CEU, ASN)それぞれから見つかったハプロタイプブロックと3集団共通のハプロタイプブロック。この3集団共通のハプロタイプブロックは古い正の選択の候補領域(図3-2)と重複があるので古いハプロタイプブロックとみなす。

3.2.2 ディプロタイプ間距離 HHD の計算

k番目の古いハプロタイプブロックについて、ディプロタイプiとjの間の HHD、d_{ijk} $(1 \leq i < j \leq 270)$を計算した。 具体的にはそれぞれの古いハプロタイプブロックに対して、3 集団のすべてのディプロタイプのペアについて HHD を計算し 270 × 270 HHD 行列を作成した(計算法については 2.3.2、図 3-4 を参照)。

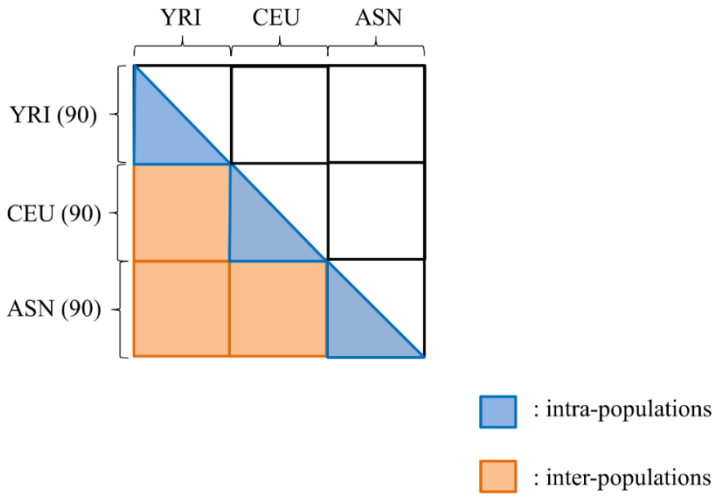


図 3-4 各古いハプロタイプブロックで作成された 270x270 の HHD 行列[34]
行列内の青で囲まれた距離は集団内距離、橙で囲まれた距離は集団間距離とする。

3.2.3 集団間の違いを持つ古いハプロタイプブロックの抽出

古い正の選択の候補である古いハプロタイプブロックのうち、集団間で多様化が起こったものがあるかどうかをみていく。共通の祖先ハプロタイプと集団特異的なハプロタイプが存在するような古いハプロタイプブロックを見つけるために、それぞれの古いハプロタイプブロックkについて集団間距離X_kと集団内距離Y_kに基づいた t-統計量スコアを計算する:

$$t_k = \frac{\overline{X_k} - \overline{Y_k}}{\sqrt{s_{XY_k}\left(\frac{1}{m}+\frac{1}{n}\right)}},$$

ここで

$$s_{XY_k} = \frac{(m-1)s_{X_k}+(n-1)s_{Y_k}}{m+n-2},$$

mは異なる集団に属する個人間の距離の総数、nは同じ集団に属する個人間の距離の総数を表す(図 3-4). $\overline{X_k}$ と $\overline{Y_k}$ は、集団間/集団内距離の標本平均、s_{X_k} と

s_{Y_k} は、集団間/集団内距離の不偏標本分散を表す。このスコアは異なる集団に属する個人間の HHD の平均と、同じ集団に属する個人間の HHD の平均が、どれだけ異なるかを測る。本研究では 3.2.2 で各古いハプロタイプブロックについて作成した HHD 行列を用いてこのスコアを計算し、古いハプロタイプブロックの順位付けをした。スコア分布の上位の裾、例えば上位 1%の古いハプロタイプブロックは、集団間と集団内のハプロタイプの違いが大きいことを表す。このとき、集団間の分化がみられるとし、古いハプロタイプブロックは、古い正の選択を受けた共通の祖先ハプロタイプと集団特異的なハプロタイプを含み、集団間の多様化が起こったとみなした。そして、上位 1%ブロックについては、さらに機能アノテーション等の検証を行った(詳しくは、3.5.2 を参照)。

3.2.4 古いハプロタイプブロックのクラスタリング

ネットワークのパターンにより古いハプロタイプブロックは分類される。ネットワークの 3 つのノードはそれぞれ集団 YRI, CEU, ASN、3 つのエッジは CEU と YRI, CEU と ASN, ASN と YRI という 2 集団間の t-統計量スコアで表す(図 1-4(B))。それぞれの古いハプロタイプブロックについて 3 つのエッジを t-統計量スコアで表した t-統計量スコアプロファイルを作成し、これに基づき古いハプロタイプブロックを分類した。ネットワークのエッジの長さが、長い・短い、の 2 種類があると考えると、ノードが 3 つのネットワークは 8 パターンをもつ。そこでクラスタリング手法の中でも k-means 法を$k = 8$として実行した。

3.3 パイプラインにより抽出された古いハプロタイプブロックの機能アノテーション

抽出された古いハプロタイプブロックの t 統計量スコア上位 1%に含まれる遺伝子・SNP について行われた機能アノテーションを 3.3.1、3.3.2 において説明する。

3.3.1 遺伝子のエンリッチメント解析による機能アノテーション

抽出された遺伝子と機能との関連付けはエンリッチメント解析によって行われた。エンリッチメント解析はなんらかの既知の機能カテゴリーを使って行われるが、本研究では、KEGG pathway を用いて行い、KEGG pathway のいずれのパスウェイマップに得られた遺伝子(つまり t 統計量スコアで上位 1%のブロックにある遺伝子)が存在する確率が高いかどうかをモンテカルロ検定により評価した。モンテカルロ検定では、古いハプロタイプブロック全体(30,966 個)を利用し､その総数の 1%にあたる 310 個の古いハプロタイプブロックを 10,000 回サンプリングした。サンプリングごとにサンプリングされた古いハプロタイプブロックに含まれる遺伝子リストを作成し、各パスウェイマップに含まれる遺伝子との重複を Jaccard index を使用して測った。パスウェイマップごとに 10,000 個の Jaccard index が計算され、その分布を作成できるが、この分布における、実データから得られた遺伝子リストとそのパスウェイの重複についての Jaccard index の p-値を計算した。そして、p-値が 0.05 より小さいとき、実データの遺伝子はそのパスウェイに多く含まれているとみなした。

3.3.2 SNP の GWAS カタログによる機能アノテーション(SNP のエンリッチメ

ント解析による機能アノテーション）

検出した領域に含まれる SNP については、まず、SNP と形質との関連がすでに報告されているかどうかを NHGRI GWAS カタログを使って調べた。NHGRI GWAS カタログは SNP とヒトの形質の関連性を集めたものである[73]。すでにヒトの形質との関連性が報告されている SNP は、関連が報告されている遺伝子によってパスウェイにマップされた。

3.4 パイプラインの実データへの適用

3.4.1 データセット

HapMap データ

相が不明な遺伝子型データは、2.4.1 の HapMap データを使用した。2.4.1 と同様に、HapMap データのうち 3 集団で共通の 3,619,226 SNP のうち(図 2-5)、270 人すべてのデータが揃っている、計 879,657 SNP を解析に用いる。

Entrez SNP データベース

Entrez SNP (https://www.ncbi.nlm.nih.gov/snp)を用いて、dbSNP build 132 の非同義 SNP(nonsynonymous SNP; nsSNP)を抽出した。本研究では 3 種類の nsSNP を用いた; 173,911 ミスセンス、6,838 ナンセンス そして 24,296 frame shift SNP である。これらのうち、4,361 nsSNP が、本研究で用いた HapMap データセットに含まれていた。また、CCDS[74] build 36.3 を用いて、それぞれの SNP がコード領域内に存在するかどうかを確認した。計 3,298 SNP が、22 本の常染色体上の 2,467 遺伝子上に存在していた。

KEGG PATHWAY データベース

KEGG PATHWAY データベースとは、知識ベースのデータベースで、分子相互作用ネットワーク（PATHWAY データベース）、遺伝子・タンパク質情報（GENES/SSDB/KO データベース）、そして生化学化合物とその反応（COMPOUND/GLYCAN/REACTION　データベース）を含む[75]。本研究は、KEGG PATHWAY を用いた。KEGG PATHWAY には、448 リファレンスパスウェイマップが存在している。そのうち、69 のマップがヒトの疾患関連のマップである。本研究で、すでに多様性がみられる疾患のパスウェイと免疫を介したありふれた疾患の多様性との関連が想定される免疫システムや感染症のパスウェイの機能カテゴリーには、それぞれ、”Cancers”は 21 パスウェイ、”Endocrine and metabolic diseases”は 4 パスウェイ、”Cardiovascular disease”は 4 パスウェイ、”Neurodegenerative diseases”は 5 パスウェイ”Immune system”は 16 パスウェイ、”Immune diseases”は 8 パスウェイ、”Infectious diseases”に 24 パスウェイが含まれていた。

KEGG Mapper は web ページ(http://www.genome.jp/kegg/tool/map_pathway2.html)から使用できるツールである。遺伝子リストを入力すると、リスト内の遺伝子が存在する KEGG パスウェイが全て出力される。機能アノテーションの際は、KEGG Mapper を用いてパイプラインで得られた遺伝子をパスウェイにマップし、パスウェイの機能を確認した。

3.4.2 抽出された古いハプロタイプブロック

Haploview は、YRI、CEU、ASN に対して、62,123、56,597、56,325 のハプロタイプブロックを 22 本の常染色体上に検出した。また、3 集団に共通のハ

プロタイプブロックは 76,119 個検出された。古いハプロタイプブロックの定義より、76,119 個のハプロタイプブロックのうち、39,228 個が古いハプロタイプブロックとみなされた。さらに古いハプロタイプブロックのうち 3SNP 以上で構成された 30,966 の古いハプロタイプブロックのみを本研究では使うことにした。抽出された古いハプロタイプブロックの大きさの、最大、最小、平均は、499,794 bp、42bp、24,584.36 bp であった。平均が 24,584.36bp というのは、long range haplotype test のような近年の正の選択を検出することを目的とした LD に基づく検定によって検出された領域と比べてかなり短い(表 3-1)。これは、本パイプラインは古い正の選択の候補領域を検出するために各集団のハプロタイプブロックの共通部分を抽出しているのでその結果である。ハプロタイプブロック領域内に存在する SNP 数は、3~97 個で、遺伝子数は、0~6 個だった。5,577 個の遺伝子、240,752SNP が、検出された古いハプロタイプブロック内に存在した。

表 3-1 先行研究で検出された領域の平均長[34]

先行研究	平均長 (bp)
LRH, iHS[76]	310,049.59
LRH, iHS, XP-EHH[77]	151,579.03
EHHS[28]	336,811.55
CMS[78]	86,178.84
XP-CLR[6]	1,280,084.33
HaploPS[79]	449,043.75
本研究の古いハプロタイプブロック	24,584.36
上位 1%ブロック	35,803.89

3.4.3 古いハプロタイプブロックのスコアリングとクラスタリング

古いハプロタイプブロックの評価

3 集団間の違いを示す古いハプロタイプブロックを見つけるために、3.2.2 で定義した t-統計量スコアを各古いハプロタイプブロックについて計算した。そのスコア分布を図 3-5 に示す。このスコア分布は極値分布とみなすことができる。スコアが大きければ大きいほど集団間、集団内、の違いが大きいことを表す。スコアが大きいブロックを抽出することにより、特に集団間のハプロタイプの違いがはっきりとみられる少数のブロックを抽出することが出来ると考えられる。スコアが大きい順にならべた古いハプロタイプブロックの上位 1%には、310 の古いハプロタイプブロックがあり、130 遺伝子(表 3-2、表 3-3)、2,803 SNP が含まれていた。310 の古いハプロタイプブロックの平均長は 35,803.89 bp だった(表 3-1)。310 の古いハプロタイプブロックが存在するゲノム領域中に含まれる nsSNP は、246 ミスセンス SNP、10 ナンセンス SNP あった(表 3-4)。これらの nsSNP は、それぞれ 72、8 遺伝子に含まれていた(表 3-5)。また、上位 5%、1%のブロック内の SNP で、F_{st} が 0.2 より大きいものはそれぞれ 35%、49%存在した。F_{st} の平均値は、上位 5%、1%のブロックの SNP についてそれぞれ 0.162 と 0.187 で、それら差は Welch の t 検定で有意(p-値<0.05)であることが分かった。

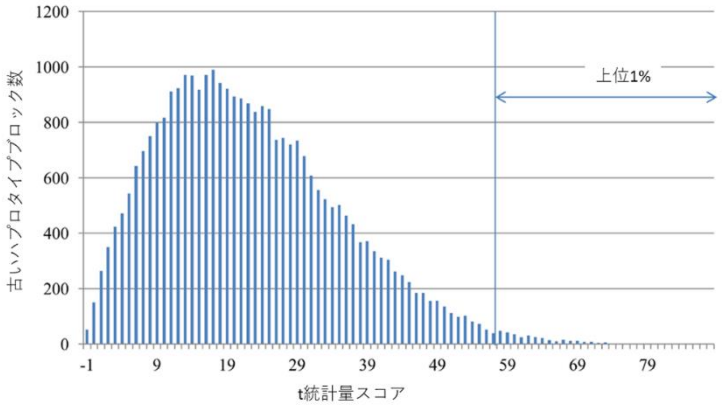


図 3-5 スコア分布[34]

x 軸は t-統計量スコア、y 軸は古いハプロタイプブロックの数を表す。

表 3-2 上位 1%ブロックで遺伝子を含んでいるもののリスト[34]

スコア	染色体	開始位置(bp)	終了位置(bp)	遺伝子
86.4	9	123944551	123953921	NDUFA8
83.28	15	41410208	41604696	ADAL,ZSCAN29,TUBGCP4,TP53BP1,MAP1A
80.32	15	32026096	32055908	AVEN
79.75	5	112189368	112219971	APC
79.44	18	19655922	19713486	LAMA3
74.55	1	65069828	65083850	JAK1
74.08	5	1850735	1879259	NDUFS6
73.78	6	102245541	102268491	GRIK2
73.31	19	38242901	38294696	RHPN2,GPATCH1
73.06	7	79671526	79691383	GNAI1
72.84	3	121019507	121275299	GSK3B
72.28	3	50417804	50499562	CACNA2D2
72.01	15	47985162	48007314	ATP8B4
71.82	9	15933931	15992275	C9orf93
71.13	10	25656389	25672837	GPR158
71.09	14	63645327	63688587	SYNE2
70.38	16	68979732	69181984	ST3GAL2,FUK,COG4,SF3B3
69.4	8	25190279	25215802	DOCK5
69.24	5	131772473	131785379	LOC441108
68.86	6	143632155	143637580	AIG1
68.83	3	52829953	52871771	ITIH4,TMEM110
68.5	11	26680597	26699178	SLC5A12
68.23	1	231831166	231833646	KCNK1
68.01	7	2106516	2127625	MAD1L1
67.56	12	6337707	6338466	SCNN1A
66.8	6	116671081	116706358	NT5DC1
66.78	6	97690663	97750368	KLHL32,C6orf167
66.7	5	145800605	145851267	TCERG1
66.53	5	460230	465694	AHRR
66.29	20	47958234	48006194	SPATA2,RNF114
65.96	5	118834319	118840649	HSD17B4
65.52	14	56728943	56819453	MUDENG
65.03	15	29636202	29704566	OTUD7A
64.89	1	205011575	205013520	IL10
64.82	14	69950782	70000315	SYNJ2BP
64.75	15	78518870	78521229	ARNT2
64.43	4	155435821	155464121	DCHS2
64.39	22	45649881	45665319	TBC1D22A

64.26	5	112238314	112262448	REEP5
64.23	1	218752203	218861725	MARK1
64.14	17	9713768	9736460	GLP2R
63.92	1	168672987	168819281	GORAB
63.8	7	7538454	7571438	COL28A1
63.77	10	14093666	14098918	FRMD4A
63.7	22	22501924	22509132	SMARCB1
63.51	15	80230095	80293845	EFTUD1
63.41	19	40405882	40412616	FAM187B
63.37	4	106671248	106698641	FLJ20184
63.37	17	64552655	64568764	ABCA9
63.33	8	14561614	14600425	SGCZ
63.26	14	35847635	35901442	MBIP
63.24	20	19232480	19286182	SLC24A3
62.95	11	15999048	16133464	SOX6
62.95	21	30185479	30215280	GRIK1
62.88	9	70730434	70736264	PIP5K1B
62.79	2	47143532	47148792	TTC7A
62.77	10	34983712	35012850	PARD3
62.71	20	19286645	19299929	SLC24A3
62.63	6	167064009	167067731	RPS6KA2
62.51	8	26692502	26726712	ADRA1A
62.46	4	94471706	94570269	GRID2
62.45	3	141764774	141768194	CLSTN2
62.36	12	110341870	110506735	SH2B3,ATXN2
62.35	14	80341951	80433377	C14orf145
62.15	15	71136686	71223975	NEO1
62.13	9	96083443	96096004	ZNF169
62.08	3	124822524	124875823	MYLK
61.99	13	90931942	90933438	GPC5
61.93	3	121596818	121611630	FSTL1
61.76	1	11456640	11484524	PTCHD2
61.68	11	48071665	48114445	PTPRJ
61.38	19	8418940	8458273	HNRNPM
61.17	11	34925688	34988152	PDHX
61.08	7	94459605	94571457	PPP1R9A
61.06	12	100400555	100417446	SPIC
61.03	7	140014377	140260681	BRAF
60.91	7	127714910	127753862	RBM28
60.79	20	9449645	9492722	C20orf103,PAK7
60.71	3	30793000	30806248	GADL1
60.7	1	203336801	203424382	RBBP5,DSTYK
60.43	9	15911782	15927653	C9orf93
60.41	1	159279214	159303096	USF1,ARHGAP30
60.25	15	64047948	64053170	MEGF11
60.23	17	25156608	25318074	SSH2
60.14	11	83019720	83033768	DLG2
60.05	11	62930260	62970259	SLC22A9
59.96	15	76716427	76733688	CHRNB4
59.9	1	203964468	204028762	NUCKS1,SLC41A1
59.77	20	41214300	41222873	PTPRT
59.75	18	48922319	48948096	DCC
59.66	11	16136851	16177326	SOX6
59.64	1	244312399	244340166	SMYD3
59.55	1	75880476	76009893	ACADM
59.48	13	66276153	66328988	PCDH9
59.4	14	60830345	60869673	PRKCH
59.35	22	41523585	41641840	ARFGAP3,PACSIN2
59.35	14	68442503	68462017	ACTN1
59.25	4	72332316	72342980	SLC4A4
59.22	11	12176480	12184257	MICAL2
59.21	11	20663510	20687919	NELL1
59.2	18	48961290	49003933	DCC
59.19	1	64312721	64326387	ROR1
58.99	21	25947970	25962999	JAM2
58.93	11	83432571	83446491	DLG2
58.93	6	38651355	38740681	BTBD9
58.92	6	152328398	152340128	ESR1
58.85	2	42326398	42356444	EML4
58.81	1	116084319	116101491	CASQ2
58.79	12	10122660	10142287	CLEC1A
58.77	11	30427273	30444885	MPPED2
58.76	12	79791042	79812822	LIN7A

58.66	16	56344212	56392233	KATNB1,KIFC3
58.48	12	130977857	131110596	EP400
58.44	14	63884064	63940963	MTHFD1
58.37	3	85888887	85932752	CADM2
58.36	6	170459223	170529596	FAM120B
58.35	22	42372045	42383885	EFCAB6

表 3-3 上位 1%ブロックの各クラスターに含まれる遺伝子[34]

クラスター	遺伝子
2	ADAL, ARHGAP30, BTBD9, CLEC1A, DOCK5, ESR1, FAM187B, GADL1, GLP2R, GORAB, GPATCH1, GPR158, GSK3B, HSD17B4, LAMA3, MAP1A, MBIP, MICAL2, MPPED2, MYLK, NDUFS6, NUCKS1, OTUD7A, PTCHD2, PTPRT, RHPN2, SCNN1A, SLC41A1, SLC4A4, SYNJ2BP, TP53BP1, TUBGCP4, USF1, ZSCAN29
3	ACTN1, ADRA1A, ARFGAP3, ATXN2, CHRNB4, CLSTN2, FSTL1, IL10, JAM2, MEGF11, PACSIN2, SH2B3, SLC22A9, SOX6, SPIC, TCERG1, TTC7A
4	ACADM, ATP8B4, C6orf167, EML4, FAM120B, KLHL32, NELL1, PPP1R9A, SMYD3
5	ABCA9, AHRR, AIG1, APC, ARNT2, AVEN, BRAF, C14orf145, C20orf103, C9orf93, CACNA2D2, CADM2, CASQ2, COG4, COL28A1, DCC, DCHS2, DLG2, DSTYK, EFCAB6, EFTUD1, EP400, FLJ20184, FRMD4A, FUK, GNAI1, GPC5, GRID2, GRIK1, GRIK2, HNRNPM, ITIH4, JAK1, KATNB1, KCNK1, KIFC3, LIN7A, LOC441108, MAD1L1, MARK1, MTHFD1, MUDENG, NDUFA8, NEO1, NT5DC1, PAK7, PARD3, PCDH9, PDHX, PIP5K1B, PRKCH, PTPRJ, RBBP5, RBM28, REEP5, RNF114, ROR1, RPS6KA2, SF3B3, SGCZ, SLC24A3, SLC5A12, SMARCB1, SPATA2, SSH2, ST3GAL2, SYNE2, TBC1D22A, TMEM110, ZNF169

表 3-4 上位 1%ブロックの各クラスターに含まれる nsSNP とそれらを含む遺伝子数

クラスター	missense		nonsense	
	#SNP	#遺伝子	#SNP	#遺伝子
2	82	21	5	4
3	32	8	2	1
4	24	5	0	0
5	108	38	3	3
Total	246	72	10	8

表 3-5 上位 1%ブロックの各クラスターに含まれるミスセンス/ナンセンス SNP を含む遺伝子

クラスター	ミスセンス	ナンセンス
2	ADAL, (ADAM21)*, ARHGAP30, DOCK5, FAM187B, GLP2R, GORAB, GPATCH1, HSD17B4, LAMA3, MAP1A, MBIP, MICAL2, MYLK, (OTUD1), PTCHD2, (TM4SF4), TP53BP1, USF1, (WDR87), ZSCAN29	ARHGAP30, FAM187B, GPATCH1, GSK3B
3	ARFGAP3, ATXN2, (LOC100287812), PACSIN2, SH2B3, SLC22A9, (SLC28A2), SPIC	(SLC28A2)*
4	(ADAMTS20), ATP8B4, C6orf167, FAM120B, (NRG1)	-
5	ABCA9, (ADCK2), APC, (ARHGEF38), C14orf145, CACNA2D2, COG4, DCHS2, (DERL3), DSTYK, (EFCAB5), EFTUD1, EP400, (EXOC5), FUK, GRID2, ITIH4, JAK1, KATNB1, KIFC3, (LOC100288867, LOC441687, LOC644620), MUDENG, (MUSTN1), PAK7, PDHX, PTPRJ, RBM28, REEP5, RNF114, SF3B3, SLC5A12, SYNE2, (TSPYL4, USO1, XCR1), ZNF169	EFTUD1, JAK1, (TSPYL4)

*本研究で解析対象とした HapMap release24 には登録されていないが dbSNP build 132 には登録されている nsSNP のうち、古いハプロタイプブロックのゲノム領域内に含まれるとされたものを含む遺伝子は括弧で囲んだ。それらの遺伝子は表 3-3 には含まれていない。

古いハプロタイプブロックのクラスタリング

クラスタリングは、t-統計量スコアプロファイルに基づいて行われた(表 3-6)。結果として、すべてのエッジが長いクラスターは得られず、かわりにクラスター5 と類似のエッジのパターンを持つクラスター5'を得た(図 3-6)。クラスター5 に比べてクラスター5'の YRI とつながるエッジはクラスター5 にくらべてかなり短い。30%もの古いハプロタイプブロックがすべてのエッジが短いクラスタ

ー1 に分類された(図 3-6、表 3-7)。いずれか 1 つのエッジが長いクラスター2、3、4、YRI とつながっているエッジが長いクラスター5 には、ほぼ同じ数の古いハプロタイプブロックが分類され、全体的にエッジが短いクラスター6、7 には、クラスター2、3、4、5 のほぼ 2 倍の古いハプロタイプブロックが分類された。それぞれのクラスターに異なる自然選択の歴史を想定し、結果を検討した(3.4.4 を参照)。

表 3-6　クラスタリングに用いた t-統計量スコアプロファイル[34]

2、3、4 列目には 2 集団間の t-統計量スコア、5 列目にはクラスタリングで付与されたクラスター番号、6 列目には 3 集団間の t-統計量スコア、7 列目にはそれぞれの古いハプロタイプブロック内の遺伝子が示されている。古いハプロタイプブロックの ID は古いハプロタイプブロックを構成する SNP のうち最初の SNP の rs 番号である。上位 1%ブロックで遺伝子を含むものについてのみここでは表記する。

ブロック ID	CEU-YRI	CEU-ASN	ASN-YRI	クラスター	3 集団間の比較	遺伝子
rs1556217	76.63	27.53	84.68	5	86.4	NDUFA8
rs3742971	113.74	36.36	39.19	2	83.28	ADAL ZSCAN29 TUBGCP4 TP53BP1 MAP1A
rs525243	95.82	19.75	54.5	5	80.32	AVEN
rs2546108	72.73	21.97	90.95	5	79.75	APC
rs12957168	84.2	37.03	47.74	2	79.44	LAMA3
rs1048007	93.99	20.1	52.95	5	74.55	JAK1
rs2276986	73.09	63.37	24.57	2	74.08	NDUFS6
rs2786254	70.67	10.58	72.66	5	73.78	GRIK2
rs1559088	95.99	19.91	43.59	2	73.31	RHPN2 GPATCH1
rs4731273	61.46	12.54	81.58	5	73.06	GNAI1
rs6438550	77.74	34.37	49.02	2	72.84	GSK3B
rs2282755	66.65	30.7	54.87	5	72.28	CACNA2D2
rs12440932	36.97	52.31	73.41	4	72.01	ATP8B4
rs10810488	64.37	5.46	84.15	5	71.82	C9orf93
rs16925561	70.41	54.91	21.9	2	71.13	GPR158
rs17101669	73.52	8.71	63.05	5	71.09	SYNE2
rs4985526	84.86	10.92	58.33	5	70.38	ST3GAL2 FUK COG4 SF3B3
rs2666156	85.61	21.97	35.09	2	69.4	DOCK5
rs1016988	47.69	31.36	67.79	5	69.24	LOC441108
rs7765576	51.19	34.56	60.22	5	68.86	AIG1
rs2535646	68.7	9.46	65.34	5	68.83	ITIH4 TMEM110
rs12289000	97.41	6.71	62.89	5	68.5	SLC5A12
rs12407694	89.06	15.61	63.87	5	68.23	KCNK1
rs3778969	50.63	10.18	89.36	5	68.01	MAD1L1
rs10849445	85.65	29.28	26.13	2	67.56	SCNN1A
rs1204826	38.19	24.77	86.7	5	66.8	NT5DC1
rs1206144	34.09	28.22	77.19	4	66.78	KLHL32 C6orf167
rs6889741	46.49	70.84	23.5	3	66.7	TCERG1
rs732111	66.8	24.74	47.43	5	66.53	AHRR
rs2769982	43.17	24.94	74.02	5	66.29	SPATA2 RNF114
rs257973	59.93	46.77	40.43	2	65.96	HSD17B4
rs8011432	67.55	5.3	59.15	5	65.52	MUDENG
rs11630385	77.83	26.12	35.62	2	65.03	OTUD7A
rs3021094	8.34	69.05	59.41	3	64.89	IL10
rs4902815	67.79	50.29	20.32	2	64.82	SYNJ2BP
rs1037124	51.37	2.88	77.92	5	64.75	ARNT2
rs10003606	74.21	2.64	55.34	5	64.43	DCHS2
rs131864	75.2	3.94	50.04	5	64.39	TBC1D22A
rs419155	50.38	17.66	75.61	5	64.26	REEP5
rs17007868	58.76	17.52	64.08	5	64.23	MARK1
rs9674995	59.22	47.23	36.03	2	64.14	GLP2R
rs12128140	51.31	43.89	46.39	2	63.92	GORAB
rs2108065	69.91	0.56	65.31	5	63.8	COL28A1
rs12571559	44.22	30.12	60.69	5	63.77	FRMD4A
rs6003899	69.95	4.24	62.95	5	63.7	SMARCB1
rs16973420	62.16	6.13	74.49	5	63.51	EFTUD1
rs10416023	59.41	64.64	12.33	2	63.41	FAM187B
rs3960769	43.15	23.5	71.14	5	63.37	FLJ20184
rs10512523	53.59	19.14	60.77	5	63.37	ABCA9
rs11203648	75.01	19.41	46.39	5	63.33	SGCZ
rs17175276	58.32	45.7	35.31	2	63.26	MBIP
rs6035283	63.29	14.81	59.29	5	63.24	SLC24A3
rs16932469	32.57	60.5	43.04	3	62.95	SOX6
rs2832478	79.03	6.41	47.28	5	62.95	GRIK1
rs10781282	66.27	21.93	44.54	5	62.88	PIP5K1B
rs4953454	25.4	66.65	36.93	3	62.79	TTC7A
rs12241700	66.87	10.92	70.43	5	62.77	PARD3

rs7264626	62.89	4.88	58.72	5	62.71	SLC24A3
rs7773974	61.51	2.32	68.67	5	62.63	RPS6KA2
rs4732646	37.55	66.83	25.36	3	62.51	ADRA1A
rs10002068	67.49	2.71	68.05	5	62.46	GRID2
rs4683509	35.81	72.91	21.27	3	62.45	CLSTN2
rs2078863	24.79	77.32	32.77	3	62.36	SH2B3 ATXN2
rs10141192	44.39	24.11	69.52	5	62.35	C14orf145
rs11632639	63.77	22.04	47.9	5	62.15	NEO1
rs7874441	43.43	17.65	74.9	5	62.13	ZNF169
rs3845915	78.79	54.77	10.89	2	62.08	MYLK
rs9560814	37.09	16.1	81.71	5	61.99	GPC5
rs13709	21.34	89.78	23.78	3	61.93	FSTL1
rs10159458	65.28	27.24	34.58	2	61.76	PTCHD2
rs3942852	61.68	0.36	69.11	5	61.68	PTPRJ
rs17160491	67.63	0.8	70.88	5	61.38	HNRNPM
rs11539202	70.2	11.69	46.91	5	61.17	PDHX
rs854745	30.3	46.37	57.41	4	61.08	PPP1R9A
rs11110820	8.03	61.89	59.32	3	61.06	SPIC
rs7792746	64.34	7.44	60.1	5	61.03	BRAF
rs1545444	36.9	9.08	82.1	5	60.91	RBM28
rs2283635	60.36	17.82	54.88	5	60.79	C20orf103 PAK7
rs9882060	87.18	32.16	13.51	2	60.71	GADL1
rs11578293	35.49	20.09	77.5	5	60.7	RBBP5 DSTYK
rs1536685	47.28	6.96	81.14	5	60.43	C9orf93
rs2073655	77.89	38.82	12.28	2	60.41	USF1 ARHGAP30
rs333556	38.23	74.34	17.69	3	60.25	MEGF11
rs8081325	44.26	15.71	73.26	5	60.23	SSH2
rs6592134	37.47	16.43	77.14	5	60.14	DLG2
rs7101446	7.58	65.63	48.53	3	60.05	SLC22A9
rs12441998	45.24	69.64	12.35	3	59.96	CHRNB4
rs1620334	69.55	30.48	31.47	2	59.9	NUCKS1 SLC41A1
rs2000135	49.4	45.82	28.02	2	59.77	PTPRT
rs9956477	41.38	12.77	76.11	5	59.75	DCC
rs16932650	24.78	60.87	40.82	3	59.66	SOX6
rs2788019	32.57	40.34	57.11	4	59.64	SMYD3
rs12732722	31.51	34.19	66.17	4	59.55	ACADM
rs9317615	67.98	2.28	61.3	5	59.48	PCDH9
rs7158527	74.08	1.85	56	5	59.4	PRKCH
rs4393836	29.48	69.3	23.71	3	59.35	ARFGAP3 PACSIN2
rs4899269	12.35	82.1	30.02	3	59.35	ACTN1
rs2045012	73.65	44.31	10.42	2	59.25	SLC4A4
rs17477991	71.92	48.97	6.75	2	59.22	MICAL2
rs1793000	19.43	41.14	64.28	4	59.21	NELL1
rs8083850	38.6	23.44	68.53	5	59.2	DCC
rs1772618	29.5	14.78	93.22	5	59.19	ROR1
rs9974110	38.12	65.83	17.98	3	58.99	JAM2
rs1384751	64.07	1.17	59.8	5	58.93	DLG2
rs17543178	45.42	42.22	40.79	2	58.93	BTBD9
rs985695	55.1	36.32	37.11	2	58.92	ESR1
rs17671409	29	23.1	75.28	4	58.85	EML4
rs7544242	70.7	18.18	40.73	5	58.81	CASQ2
rs11053573	50.87	44.86	29.8	2	58.79	CLEC1A
rs11031094	66.32	19.34	36.72	2	58.77	MPPED2
rs1716546	79.83	6.89	41.27	5	58.76	LIN7A
rs2965798	42.84	36.26	46.74	5	58.66	KATNB1 KIFC3
rs7978708	61.36	8.19	53.46	5	58.48	EP400
rs1256112	44.05	16.12	68.97	5	58.44	MTHFD1
rs12714637	48.45	23.71	64.32	5	58.37	CADM2
rs9356637	18.43	53.84	59.19	4	58.36	FAM120B
rs137744	29.77	15.82	75.38	5	58.35	EFCAB6

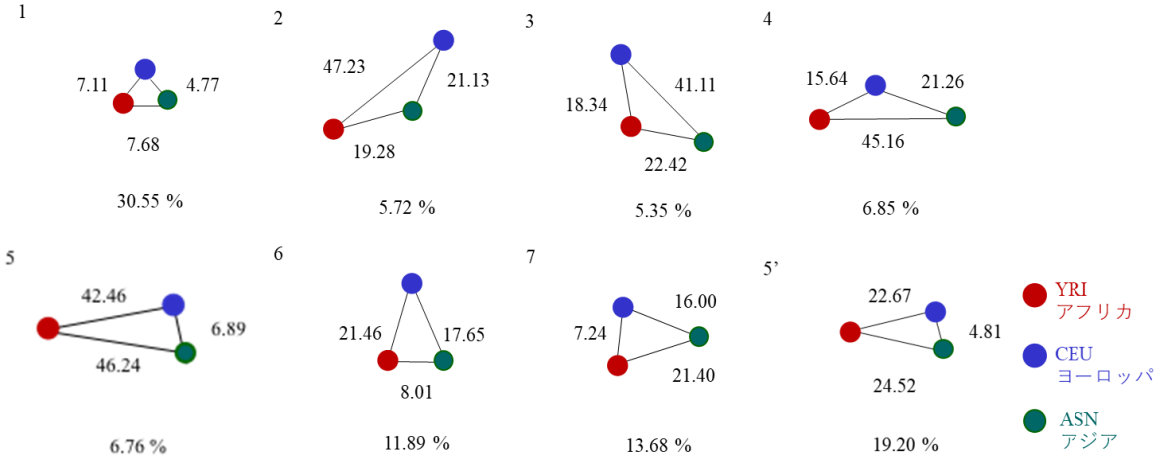


図 3-6 古いハプロタイプブロックの分類[34]

t-統計量スコアプロファイルを使ってネットワークに基づいて行われたクラスタリングにより得られた 8 つのクラスター。それぞれのエッジに書かれた数字は t-統計量スコアの平均値。スコアが小さくなるとエッジも短くなる。

表 3-7 スコアリングとクラスタリングの結果[34]

クラスター	1	2	3	4	5	6	7	5'	合計
上位 1%	0	76	39	35	160	0	0	0	310
	(0%)	(24.52%)	(12.58%)	(11.29%)	(51.61%)	(0%)	(0%)	(0%)	
合計	9,459	1,772	1,657	2,121	2,094	3,682	4,237	5,944	30,966
	(30.55%)	(5.72%)	(5.35%)	(6.85%)	(6.76%)	(11.89%)	(13.68%)	(19.20%)	

表のそれぞれの要素は得られた古いハプロタイプブロックの数。括弧の中は全体の古いハプロタイプブロック数に対する割合を表す。

スコア分布をクラスターごとに見てみると、クラスター1 からなるグループ I、クラスター2、3、4、5 からなるグループ II、クラスター6、7、5'からなるグループ III の 3 グループに分類される(図 3-7)。大部分の古いハプロタイプブロックはグループ I に分類され、これらのスコアは 18 より小さく、集団間の違いがみられないものである。グループ III とグループ II のスコアは 11 から 39、23 から 86 の範囲に広がっており、集間の大きな違いが見られる古いハプロタイプブロックはグループ II にのみ含まれていることが分かった。

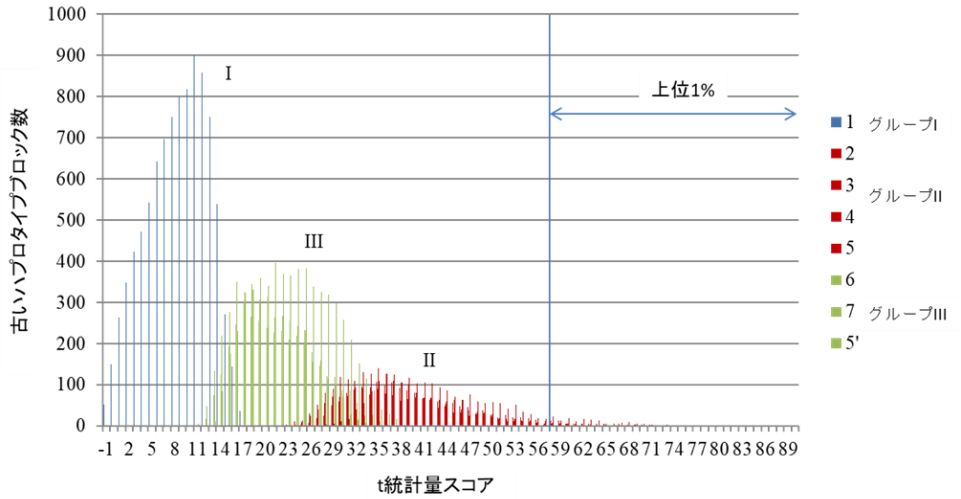


図 3-7 クラスターごとのスコア分布[37]

古いハプロタイプブロックのスコア分布がクラスターごとに示されている。クラスターは I、II、III の 3 つのグループに分類される。グループ I はクラスター1 からなる(青)。グループ II はクラスター2、3、4、5 からなる(赤)。グループ III はクラスター6、7、5'からなる(緑)。

集団間の大きな違いがみられる上位 1%ブロックにはグループ II のみが含まれ、中でも全体のブロックに対して有意に多くのクラスター2 と 5 のブロックが含まれていた。クラスター5 はアフリカの集団とその他の集団との違いが大きい

ブロックで、アフリカの集団は他の集団との遺伝的な距離が大きいという先行研究の結果と一致する。また、アフリカとヨーロッパの集団の大きな違いが見られるクラスター2 が多く含まれているというのは先行研究とは異なる結果となった。

3.4.4 上位 1%ブロックの新規性と機能アノテーション

本研究で新たに検出された自然選択の候補領域について

すでに近年の正の選択を検出する研究は多く行われており、本研究で調査した 6 つの論文[6, 28, 76-79]で近年の正の選択の候補として報告されていた遺伝子は合計 828 遺伝子だった。本研究で抽出された t 統計量スコア上位 1%の古いハプロタイプブロックに含まれていたのは 130 遺伝子で、そのうち既知の 828 遺伝子との重複は 33 遺伝子だった(表 3-8)。つまり、本研究で検出した古い正の選択を受けかつ近年の自然選択によって多様化された候補領域に含まれていた遺伝子の 75%が既存の近年の正の選択を検出する研究では報告されていないことが分かった。これらの遺伝子は、近年の自然選択が正の選択ではなかった場合か、もしくは古い正の選択を受けたことを条件にしたことにより新たに候補の上位となった近年の正の選択の可能性がある。

表 3-8 上位 1%ブロックのクラスターごとの既に近年の正の選択が報告されている遺伝子[34]

遺伝子は正の選択が検出された集団ごとに分類されている。

クラスター	遺伝子数	すでに報告されている遺伝子数	すでに報告されている遺伝子		
			アフリカ	ヨーロッパ	アジア
2	34	9	ARHGAP30 MYLK USF1	ADAL MAP1A MYLK SYNJ2BP TP53BP1 TUBGCP4 ZSCAN29	ADAL MAP1A TP53BP1 TUBGCP4 ZSCAN29
3	17	8	ACTN1	CLSTN2 SPIC TCERG1	ACTN1 ATXN2 FSTL1 MEGF11 SH2B3
4	9	2	-	-	ACADM EML4
5	70	14	EFTUD1 PRKCH	APC DCC DLG2 FLJ20184 NDUFA8 PDHK PRKCH SLC24A3	APC DCC DLG2 FLJ20184 GPC5 GRID2 REEP5 ROR1 TBC1D22A

上位 1%ブロックのエンリッチメンチ解析による機能アノテーション

上位 1%のブロック(310 ブロック)に含まれていた遺伝子についてエンリッチメント解析を行ったところ、130 遺伝子は、機能カテゴリーの“Metabolism”、“Genetic Information Processing”、“Cellular Processes”、“Organismal Systems”、そして“Human Diseases”に含まれる 22 パスウェイに多く含まれていることが分かった(表 3-9)。本研究では、すでに形質の多様性が見られる疾患のパスウェイとして、”Cancers”、”Endocrine and metabolic diseases”、”Cardiovascular disease”、”Neurodegenerative diseases”、また、免疫を介したありふれた疾患の多様性に関連するパスウェイとして、”Immune system”、”Immune

diseases”、”Infectious diseases”に注目するが、エンリッチメント解析で得られたパスウェイには、C 型肝炎、非アルコール性脂肪肝疾患、がん、などすでに集団間の発症率の違いがみられる疾患のパスウェイが見つかったのでこれらを中心に結果を見て行く。

表 3-9 エンリッチメント解析で得られたパスウェイとマップされた遺伝子[34]

機能カテゴリー	パスウェイ	マップされた遺伝子 (クラスターごとに示す)				p-値
		2	3	4	5	
Organismal Systems						
Immune system	T cell receptor signaling pathway	GSK3B	IL10		PAK7	0.029
Nervous system	Neurotrophin signaling pathway	GSK3B	SH2B3		BRAF, RPS6KA2	0.007
Endocrine system	Progesterone-mediated oocyte maturation				BRAF, GNAI1, MAD1L1, RPS6KA2	0.005
Metabolism						
Metabolism of other amino acids	beta-Alanine metabolism	GADL1		ACADM		0.016
Genetic Information Processing						
Translation	Ribosome biogenesis in eukaryotes				EFTUD1, RBM28	0.039
Environmental Information Processing						
Signaling molecules and interaction	Neuroactive ligand receptor interaction	GLP2R	ADRA1A, CHRNB4		PARD3, GRID2, GRIK1, GRIK2	0.012
Signal transduction	Hippo signaling pathway	GSK3B			APC, DLG2, PARD3	0.048
Cellular Processes						
Cellular community	Focal adhesion	GSK3B, LAMA3, MYLK	ACTN1		BRAF, PAK7	0.018
	Signaling pathways regulating pluripotency of stem cells	GSK3B			APC, JAK1	0.019
	Tight junction		ACTN1, JAM2		GNAI1, PARD3, PRKCH	0.038
Cell motility	Regulation of actin cytoskeleton	MYLK	ACTN1		APC, BRAF, PAK7, PIP5K1B, SSH2	0.001
Human Diseases						
Infectious diseases	Toxoplasmosis	LAMA3	IL10		GNAI1, JAK1	0.003
	Hepatitis C	GSK3B			BRAF,	0.022

				JAK1	
	Pertussis		IL10	GNAI1	0.023
	Leishmaniasis		IL10,	JAK1	0.025
Cancers	Colorectal cancer	GSK3B		APC, BRAF, DCC	0.001
	Renal cell carcinoma			ARNT2, BRAF, PAK7	0.008
	Endometrial cancer	GSK3B		APC, BRAF	0.018
	Basal cell carcinoma	GSK3B		APC	0.023
	Viral carcinogenesis		ACTN1	JAK1, MAD1L1	0.046
Endocrine and metabolic diseases	Non-alcoholic fatty liver disease (NAFLD)	GSK3B, NDUFS6		NDUFA8	0.016
Neurodegenerative diseases	Parkinson's disease	NDUFS6		GNAI1, NDUFA8	0.013

C 型肝炎は、発症率や症状、治療応答に関して集団間に多様性が見られる[80]。例えばアフリカ由来の集団では C 型肝炎ウィルスの除去率が低く、そのため、慢性 C 型肝炎の発症率が高い。パイプラインから得られた 130 遺伝子の中では、BRAF (クラスター5), GSK3B (クラスター2), JAK1 (クラスター5) が“Hepatitis C”にマップされた。マップされた遺伝子のうち、近年の正の選択がすでに見つかっているのが GSK3B で、アメリカ合衆国、ロサンゼンルスのメキシコ系住民で近年の正の選択が報告されている[79]。GSK3B は”Wnt signaling” モジュールに含まれる遺伝子である (表 3-10)。

また、T 細胞応答は C 型肝炎ウィルスの除去に重要な役割を果たすことが示されている[80]。本研究のエンリッチメント解析では“T cell receptor signaling pathway” が得られており、IL10 (クラスター3)、GSK3B (クラスター2)、PAK7 (クラスター5)がこのパスウェイにマップされている。マップされた IL10 を含む古いハプロタイプブロックは rs3021094、rs3024491、rs1800896 で構成されており、これらの SNP のうち rs1800896 はすでに C 型肝炎ウィルス感染応答との関連性が良く知られている SNP である[81]。

内分泌・代謝疾患である NAFLD は、集団間で病態生理学的に違いが見られることが示されている[82]。ラテン系が一番高い肝臓脂肪症の発症率を示し、アフリカ系が一番低い発症率を示し、ヨーロッパ由来の集団は中程度の発症率を示した。 本研究では NDUFS6 と GSK3B (クラスター2)、NDUFA8 (クラスター5)が“Non-alcoholic fatty liver disease (NAFLD)”にマップされた。マップされた遺伝子のうち、NDUFA8 ではすでにヨーロッパ由来の集団で近年の正の選択が見つかっている[28]。また、NDUFA8 と NDUFS6 はミトコンドリア内に存在する(ミトコンドリア電子伝達系 NAD 生成酵素)NADH 脱水素酵素のサブユニットで、それぞれ”NADH dehydrogenase (ubiquinone) 1 alpha subcomplex” モジュール、“NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria” モジュールを構成する遺伝子であり(表 3-9)、これらのモジュールは” Oxidative phosphorylation”パスウェイに含まれるものである。

がんについては、腎細胞がんはアフリカ系の男性で多く見られる[9]。子宮内膜がんはヨーロッパ由来の女性で他の集団より発症率が高い[10, 11]。基底細胞がんは肌の色が明るいヒトでよく見られる[83]。パスウェイマッピングの結果では、ARNT2 , BRAF, PAK7 (クラスター5) は “Renal cell carcinoma”に、APC, BRAF (クラスター5), GSK3B (クラスター2) は“Endometrial cancer”に、APC (クラスター5)と GSK3B (クラスター2) は“Basal cell carcinoma”にマップされた。マップされた遺伝子のうち、APC ではすでにヨーロッパとアジアの集団で近年の正の選択が報告されている[78, 79]。がん関連のパスウェイにマップされた遺伝子では、GSK3B の”Wnt signaling” モジュール、APC の”APC/C complex” モジュールが見つかった(表 3-10)。

表 3-10 エンリッチメント解析で得られた免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイにマップされた遺伝子とその機能

KEGG パスウェイ	遺伝子	KEGG での定義	モジュール
Immune System			
T cell receptor signaling pathway	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling
	IL10	interleukin 10	-
	PAK7	p21-activated kinase 7 [EC:2.7.11.1]	-
Infectious disease			
Toxoplasmosis	LAMA3	laminin, alpha 3/5	-
	IL10	interleukin 10	-
	GNAI1	guanine nucleotide-binding protein G(i) subunit alpha	cAMP signaling
	JAK1	Janus kinase 1 [EC:2.7.10.2]	JAK-STAT signaling
HepatitisC	BRAF	B-Raf proto-oncogene serine/threonine-protein kinase[EC:2.7.11.1]	-
	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling
	JAK1	Janus kinase 1 [EC:2.7.10.2]	JAK-STAT signaling
Pertussis	IL10	interleukin 10	-
	GNAI1	guanine nucleotide-binding protein G(i) subunit alpha	cAMP signaling
Leishmaniasis	IL10	interleukin 10	-
	JAK1	Janus kinase 1 [EC:2.7.10.2]	JAK-STAT signaling
Endocrine and metabolic disease			
NAFLD	NDUFA8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex subunit 8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex
	NDUFS6	NADH dehydrogenase (ubiquinone) Fe-S protein 6	NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria
	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling
Cancer			
Renal cell carcinoma	ARNT2	aryl hydrocarbon receptor nuclear translocator 2	-
	BRAF	B-Raf proto-oncogene serine/threonine-protein kinase [EC:2.7.11.1]	-
	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling
Endometrial cancer	APC	adenomatosis polyposis coli protein	APC/C complex
	BRAF	B-Raf proto-oncogene serine/threonine-protein kinase [EC:2.7.11.1]	-
	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling
Basal cell carcinoma	APC	adenomatosis polyposis coli protein	APC/C complex
	GSK3B	glycogen synthase kinase 3 beta [EC:2.7.11.26]	Wnt signaling

エンリッチメント解析で得られた、C 型肝炎、”T cell receptor signaling pathway”、がん、NAFLD のパスウェイにマップされた遺伝子の SNP うち、イントロン以外に存在する SNP は、GSK3B、IL10、PAK7、NDUFS6、APC、JAK1 に対し機能をもつもので(表 3-11、表 3-12)、中でも IL10 の上流に位置する rs1800896、PAK7 のミスセンス SNP である rs2297345、NDUFS6 の上流に位置する rs972890 の F_{st} は特に大きい値($F_{st} > 0.15$)を示している(図 3-8)。

表 3-11 上位 1%ブロックの ns SNP のうち免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイのコード領域に存在するもの

クラスター	rs 番号	遺伝子	機能
2	rs118027151	GSK3B	nonsense
	rs17202961	LAMA3	missense
	rs117387365	LAMA3	missense
	rs12457323	LAMA3	missense
	rs17187262	LAMA3	missense
	rs820463	MYLK	missense
	rs41366751	MYLK	missense
3	-	-	-
4	-	-	-
5	rs73220015	APC	missense
	rs459552	APC	missemse
	rs79617548	APC	missense
	rs2229995	APC	missense
	rs33974176	APC	missense
	rs41291734	CACNA2D2	missense
	rs117679986	JAK1	missense
	rs78190282	JAK1	nonsense
	rs2297345	PAK7	missense

表 3-12 免疫システム・感染症、もしくは形質の多様性が報告されている疾患のパスウェイにマップされた遺伝子を含むブロックを構成する SNP のうち nsSNP ではなく、また、イントロン以外に存在する SNP

クラスター	遺伝子	ブロック ID	SNP	機能の重要性	F_{st}
2	GSK3B	rs6438550	rs2037547	utr variant 3 prime	0.088
	NDUFS6	rs2276986	rs972890	upstream variant 2KB	0.282
3	IL10	rs3021094	rs1800896	upstream variant 2KB	0.177
5	PAK7*	rs2283635	rs16996034	downstream variant 500B	0.080
	GNAI1	rs4731273	rs12721454	synonymous codon	0.313
	JAK1	rs1048007	rs2230587	synonymous codon	0.024

* PAK7 は、解析対象の SNP ではなかったため、アレル頻度は dbSNP の web ページから確認した。

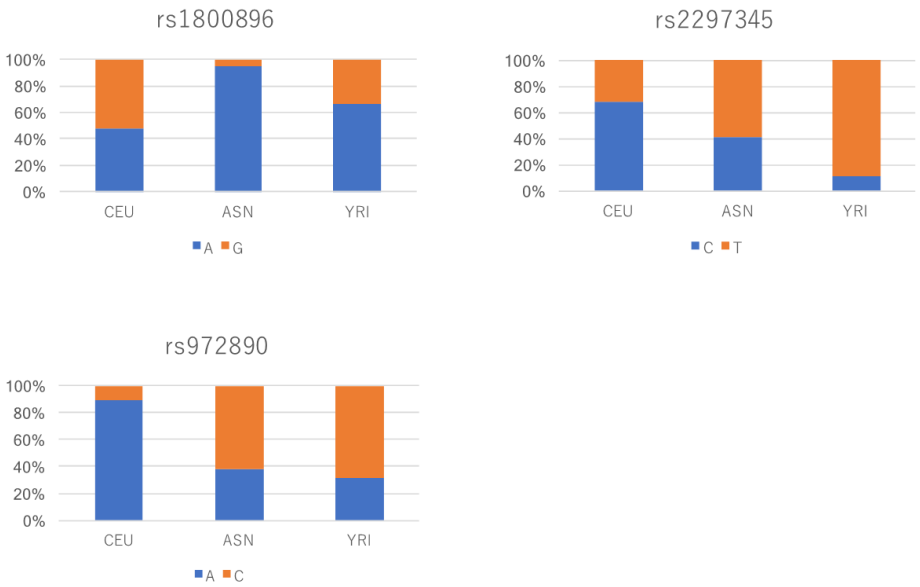


図 3-8　集団間のアレル頻度の大きな違いが見られた SNP のアレル分布

クラスターごとの遺伝子や SNP の機能アノテーション

上位 1%のブロックに含まれていた遺伝子はエンリッチメント解析により機能アノテーションを行ったが、エンリッチメント解析で得られなかったパスウェイについても、遺伝子がマップされているものがあるかどうか、つまり、上位 1%のブロックに含まれた遺伝子が高頻度ではないもののマップされているパスウェイがあるかどうかも詳細に見ていく。その際、上位 1%のブロックはクラスター2、3、4、5 のみに属するので、それぞれのクラスターごとに確認していく。具体的には、それぞれのクラスターごとに、各クラスターが含む遺伝子や SNP を、パスウェイや GWAS カタログにマップし、マップされた遺伝子について、すでに知られている多様性情報と整合性が見られるかどうかを確認する。そして、実際に集団間の違いが見られるかどうか F_{st} を使って確認する。パスウェイに遺伝子をマップする際には、疾患の発症率の違いに関連あるパスウェイとして、特に、”Cancers”、”Endocrine and metabolic

diseases”、”Cardiovasular diseases”、”Neurodegenerative diseases”、また、免疫を介したありふれた疾患の多様性に関連するパスウェイとして、”Immune system”、”Immune diseases”、”Infectious diseases”に含まれるパスウェイに注目する。

ネットワークの多様性と自然選択の歴史

クラスターごとに結果を見ていく際、上位 1%ブロック内の各クラスターを表すネットワークが(図 3-6)、各クラスターに含まれる領域が受けてきた古くからの自然選択の歴史を表現していると考え、既知の多様性情報との整合性を確認する(図 3-9)。

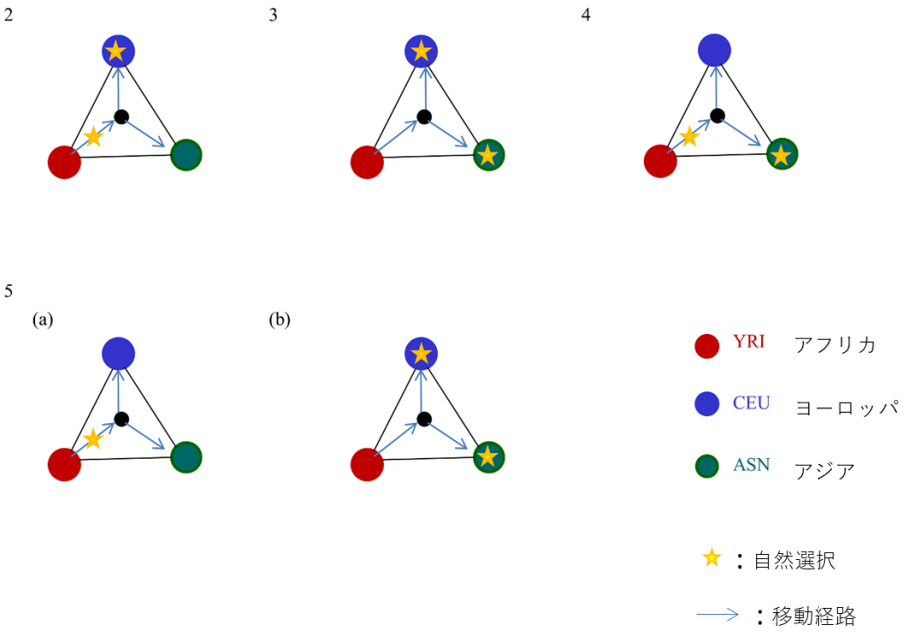


図 3-9 上位 1%ブロックの各クラスターに想定した歴史[34]
それぞれのノードは集団、エッジは 2 集団間の関係性を表す。赤、青、緑のノードはそれぞれ YRI、CEU、ASN を表し、矢印はヒトのアフリカからの移動経路、星印は自然選択を表す。

クラスター2 では、アフリカとヨーロッパの集団間の違いが大きいことから、アフリカでの古い正の選択が起こった後、ヨーロッパで近年の自然選択が起こったと考えた。同様に、クラスター3 では、ヨーロッパとアジアの集団間の違いが大きいことから、アフリカでの古い正の選択に加えてヨーロッパとアジアでも近年の自然選択が、クラスター4 では、アジアとアフリカの集団間の違いが大きいことから、アフリカの古い正の選択に加えてアジアで近年の自然選択が起こったと考えた。クラスター5 では、アフリカの集団が他の集団と大きな違いがあり、アフリカで古い正の選択が起こった後、(a)アフリカからヒトが拡散する

途中で近年の自然選択が起こったとする場合、(b)ヨーロッパ、アジアでそれぞれ近年の自然選択が起こったとする場合の 2 パターンを考えた。

クラスター2

上位 1%ブロック内のクラスター2 には 34 の遺伝子が含まれ(表 3-3)、そのうち 5 遺伝子が本研究で注目したパスウェイにマップされた(表 3-13)。すでに多様性が知られている疾患のパスウェイである”Hapatitis C”、”Basal cell carcinoma”、”Endometrial cancer”、”Prostate cancer”、”Small cell lung cancer”、”NAFLD”にマップされたのは、GSK3B、MYLK、LAMA3、NDUFS6 で、免疫システム関連の”Immune sysstem”パスウェイにもマップされていたのは、GSK3B だった。GSK3B は、”B cell receptor signaling pathway”、”Chemokine signaling pathway”、”T cell receptor signaling aptwhay”にマップされていた。最近の研究では、ケモカインと NAFLD の関連性が報告されている[84]。GSK3B は NDUFS6 とともに、エンリッチメント解析でも得られた、NAFLD にもマップされており、特に、NDUFS6 の上流に位置する rs972890 は大きい F_{st} を示している(図 3-8)。NAFLD はアフリカとヨーロッパで発症頻度の多様性が見られることから、GSK3B と NDUFS6 がアフリカで古い正の選択を、ヨーロッパで近年の自然選択を受けたことが、NAFLD のアフリカとヨーロッパの集団間の多様性に関与している可能性が示唆された(図 3-10)。

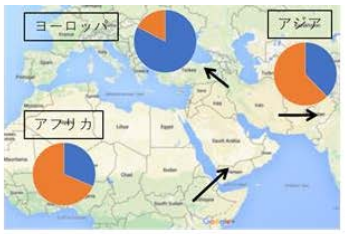
また GWAS カタログでは、クラスター2 の 76 の古いハプロタイプブロックに含まれる SNP のうち、5 SNP はすでに形質との関連が報告されており、骨ミネラル濃度、前立腺特異的な抗原レベル、髪の形態、乳がんに関連するものであった(表 3-14)。これらのうち乳がんに関連する rs9383951 のみが ESR1 を介

して”Estrogen signaling pathway”にマップされた。乳がんはヨーロッパで高頻度に観測される疾患で、ESR1 が受けたアフリカにおける古い正の選択とヨーロッパにおける近年の自然選択が、乳がんの多様性に関与している可能性が示唆された。

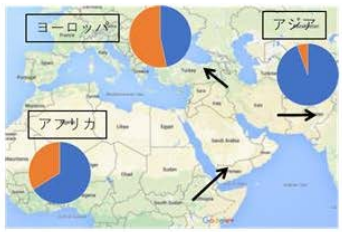
表 3-13　各クラスターの遺伝子がマップされた関連パスウェイ

機能カテゴリー	パスウェイ	遺伝子 (クラスターごとに示す)			
		2	3	4	5
Organismal Systems					
Immune system					
	B cell receptor signaling pathway	GSK3B	-	-	-
	Chemokine signaling pathway	GSK3B	-	-	BRAF, GNAI1, PARD3
	Fc gamma R-mediated phagocytosis	-	-	-	PIP5K1B
	Intestinal immune network for IgA production	-	IL10	-	-
	Leukocyte transendothelial migration	-	ACTN1, JAM2	-	GNAI1
	Natural killer cell mediated cytotoxicity	-	-	-	BRAF
	Platelet activation	MYLK	-	-	GNAI1
	T cell receptor signaling pathway	GSK3B	IL10	-	PAK7
Human Diseases					
Immune diseases					
	Allograft rejection	-	IL10	-	-
	Asthma	-	IL10	-	-
	Autoimmune thyroid disease	-	IL10	-	-
	Inflammatory bowel disease (IBD)	-	IL10	-	-
	Systemic lupus erythematosus	-	ACTN1, IL10	-	-
Infectious diseases					
	African trypanosomiasis	-	IL10	-	-
	Ameobiasis	LAMA3	ACTN1, IL10	-	-
	Chagas disease (American trypanosomiasis)	-	IL10	-	GNAI1
	Epithelial cell signaling in Helicobacter pylori infection	-	JAM2	-	
	Epstein-Barr virus infection	GSK3B	IL10	-	JAK1
	HTLV-I infection	GSK3B	-	-	APC, JAK1
	Hepatitis B	GSK3B	-	-	JAK1
	Hepatitis C	GSK3B	-	-	BRAF, JAK1
	Herpes simplex infection	-	-	-	JAK1
	Influenza A	-	-	-	JAK1
	Leishmaniasis	-	IL10	-	JAK1
	Malaria	-	IL10	-	-
	Measles	GSK3B	-	-	JAK1
	Pertussis	-	IL10	-	GNAI1
	Staphylococcus aureus infection	-	IL10	-	-
	Toxoplasmosis	LAMA3	IL10	-	GNAI1, JAK1
	Tuberculosis	-	IL10	-	JAK1
Cancers					
	Acute myeloid leukemia	-	-	-	BRAF
	Basal cell carcinoma	GSK3B	-	-	APC
	Bladder cancer	-	-	-	BRAF
	Choline metabolism in cancer	-	-	-	PIP5K1B
	Chronic myeloid leukemia	-	-	-	BRAF
	Colorectal cancer	GSK3B	-	-	APC, BRAF, DCC

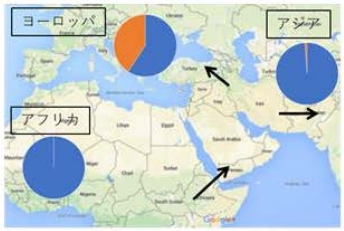
	Endometrial cancer	GSK3B	-	-	APC, BRAF
	Glioma	-	-	-	BRAF
	Melanoma	-	-	-	BRAF
	MicroRNA in cancer	-	-	-	APC
	Non-small cell lung cancer	-	-	EML4	BRAF
	Pancreatic cancer	-	-	-	BRAF, JAK1
	Prostate cancer	GSK3B	-	-	BRAF
	Proteoglycans in cancer	ESR1	-	-	BRAF
	Renal cell carcinoma	-	-	-	ARNT2 BRAF PAK7
	Small cell lung cancer	LAMA3	-	-	-
	Thyroid cancer	-	-	-	BRAF
	Transcriptional misregulation in cancers	-	-	-	ARNT2
	Viral carcinogenesis	-	ACTN1	-	JAK1, MAD1L1
Cardiovascular diseases					
	Arrythmogenic right ventricular cardiomyopathy (ARVC)	-	ACTN1	-	CACNA2D2
	Dilated cardiomyopathy (DCM)	-	-	-	CACNA2D2
	Hypertrophic cardiomyopathy (HCM)	-	-	-	CACNA2D2
Endocrine and metabolic diseases	Non-alcoholic fatty liver disease (NAFLD)	GSK3B, NDUFS6	-	-	NDUFA8
Neurodegenerative diseases	Alzheimer's disease	GSK3B, NDUFS6	-	-	NDUFA8
	Huntington's disease	NDUFS6	-	-	NDUFA8
	Parkinson's disease	NDUFS6	-	-	GNAI1, NDUFA8



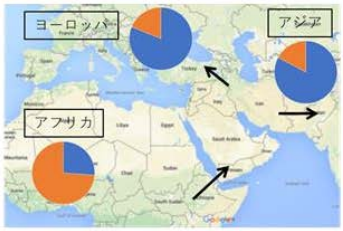
rs972890
Fst=0.282
Ancestral Allele: A
Derived Allele: C



rs1800896
Fst=0.177
Ancestral Allele: A
Derived Allele: G



rs3184504
Ancestral Allele: C
Derived Allele: T



rs10056340
Fst=0.299
Ancestral Allele: T
Derived Allele: G

図 3-10 上位 1%ブロックの SNP のアレル頻度と想定される歴史
rs3184504 は検出された古いハプロタイプブロックに含まれていたが、本研究の解析対象の SNP ではなかったため、アレル頻度は dbSNP の web サイトから確認した。そのため F_{st} は計算できなかった。

表 3-14 上位 1%ブロックのクラスターごとの既に何らかの形質との関連が GWAS によって報告されている SNP[34]

上位 1%ブロックに含まれる SNP で既に GWAS で形質との関連が報告されているものをクラスターごとに示す。3 列目は、各 SNP が存在するかまたは連鎖する遺伝子、5 列目はその遺伝子を介してマップされるパスウェイ、6 列目はそのほか得られた遺伝子でそのパスウェイにマップされたものを示している。

クラスター	SNP	関連遺伝子	形質/疾患	KEGGパスウェイマップ	マップされた遺伝子
2	rs10416218	GPATCH1	Bone mineral density	-	-
	rs823123	NUCKS1, RAB7L1	Prostate-specific antigen levels	-	-
	rs6679073	SLC41A1	Prostate-specific antigen levels	-	-
	rs7586898	AC113608.1	Hair morphology	-	-
	rs9383951	ESR1	Breast cancer	04915 Estrogen signaling pathway	ESR1, GNAI1
				04917 Prolactin signaling pathway	ESR1, GSK3B
				04919 Thyroid hormone signaling pathway	ESR1, GSK3B
				04961 Endocrine and other factor-regulated calcium reabsorption	ESR1
				05205 Proteoglycans in cancer	BRAF, ESR1
3	rs11949289	intergenic	Response to anti-depressant treatment in major depressive disorder	-	-
	rs7101446	SLC22A9	Economic and political preferences	-	-
	rs3184504	SH2B3	Eosinophil count	04722 Neurotrophin signaling pathway	SH2B3
4	rs4949874	ACADM	Blood metabolite ratios	00640 Propanoate metabolism	ACADM
				00071 Fatty acid degradation	ACADM
				00280 Valine, leucine and isoleucine degradation	ACADM
				00410 beta-Alanine metabolism	ACADM, GADL1
				03320 PPAR signaling pathway	ACADM
5	rs3818638	NDUFA8	Obesity-related traits	00190 Oxidative phosphorylation	NDUFA8, NDUFS6
				05010 Alzheimer's disease	GSK3B, NDUFS8, NDUFS6
				05012 Parkinson's disease	GNAI1, NDUFA8, NDUFS6
				05016 Huntington's disease	NDUFA8, NDUFS6
				04932 Non-alcoholic fatty liver disease (NAFLD)	GSK3B, NDUFA8, NDUFS6
	rs6475606	CDKN2B-AS1	Intracranial aneurysm	04350 TGF-beta signaling pathway	-
				04068 FoxO signaling pathway	BRAF, IL10
				04110 Cell cycle	-
				05203 Viral carcinogenesis	ACTN1, JAK1, MAD1L1
				05222 Small cell lung cancer	LAMA3
				05166 HTLV-I infection	APC, GSK3B, JAK1
	rs10965235	CDKN2BAS	Endometriosis	-	-

			(IRF1)	
			04917 Prolactin signaling pathway	ESR1, GSK3B
rs1016988	IRF1, SLC22A4, SLC22A5	Fibrinogen	05133 Pertussis	GNAI1, IL10
			05160 Hepatitis C	BRAF, GSK3B, JAK1
			(SLC22A4, SLC22A5)	
			05231 Choline metabolism in cancer	PIP5K1B
rs11242111	IRF1, C5orf56	Fibrinogen	-	-
rs11118346	LYPLAL1	Height	-	-
rs2820446	LYPLAL1	Type 2 diabetes	-	-
			(CAMK4)	
			04020 Calcium signaling pathway	ADRA1A, MYLK
			04024 cAMP signaling pathway	BRAF, GNAI1
			04921 Oxytocin signaling pathway	CACNA2D2, GNAI1, MYLK
			04725 Cholinergic synapse	CHRNB4, GNAI1
			04720 Long-term potentiation	BRAF, RPS6KA2
			04722 Neurotrophin signaling pathway	BRAF, GSK3B, RPS6KA2, SH2B3
rs10056340	CAMK4, TSLP, WDR36, SLC25A46	Allergic sensitization	04380 Osteoclast differentiation	JAK1
			05031 Amphetamine addiction	-
			05034 Alcoholism	BRAF, GNAI1
			(TSLP)	
			04630 Jak-STAT signaling pathway	IL10, JAK1
			04060 Cytokine-cytokine receptor interaction	IL10
			(WDR36)	
			03008 Ribosome biogenesis in eukaryotes	EFTUD1, RBM28
rs7620363	NR	Non-substance related behavioral disinhibition	-	-
rs3008706	intergenic	Bilirubin levels	-	-
rs11203649	SGCZ	Obesity-related traits	-	-
rs3942852	PTPRJ	Acute lymphoblastic leukemia (childhood)	04520 Adherens junction	ACTN1, PARD3, PTPRJ

クラスター3

上位 1%ブロック内のクラスター3 には 17 遺伝子が含まれ(表 3-3)、そのうち 3 遺伝子は本研究で注目したパスウェイにマップされた(表 3-13)。免疫疾患の “Asthma”、 “Inflammatory bowel disease (IBD)”、“Systemic lupus erythematosus”には IL10 がマップされた。IL10 は免疫システムの” T cell receptor signaling pathway”、古くからアフリカに存在した病原体による感染症の”African trypanosomiasis”、“Malaria”、“Tuberculosis”にマップされ、また IL10 の上流の SNP rs1800896 の F_{st} は大きな集団間差を示しており(表 3-12、図 3-8)、これは IL10 が古い正の選択の候補として検出されたことと整合性があると考えられる。エンリッチメント解析で rs1800896 はすでに C 型肝炎ウィルス感染応答との関連性が良く知られていることがわかっており、IL10 がアフリカにおける古い正の選択とヨーロッパ・アジアにおける近年の自然選択を受けたことが、C 型肝炎をはじめ、マップされた疾患の多様性に関与している可能性がある(図 3-10)。

また、セリアック病は、アフリカ、ヨーロッパ、アジアでの発症頻度の多様性が見られる疾患であるが、セリアック病関連遺伝子 SH2B3 が top1%ブロックのクラスター3 に含まれていた。SH2B3 のミスセンス SNP rs3184504 は GWAS カタログで好酸球数との関連が報告されており(表 3-14)、またこの SNP は、NOD2 認識パスウェイの活性化への関与もすでに報告されている[85]。SH2B3 はアフリカ、アジア、ヨーロッパで近年の正の選択が見られることがすでに報告されているが[85, 79]、本研究では SH2B3 にはアフリカですでに古い正の選択が起こっていたことが示唆され、SH2B3 のアフリカでの古くからの正の選択と、ヨーロッパ・アジアでの近年の自然選択が、セリアック病の多様性に関与している可能性が示唆された(図 3-10)。

クラスター4

上位 1%ブロック内のクラスター4 は 9 遺伝子を含んでいたが(表 3-3)、本研究で注目した免疫関連パスウェイにはマップされなかった(表 3-13)。その代わりに、ACADM と EML4 は注目したパスウェイ以外の 6 パスウェイにマップされ、そのうち 4 パスウェイは代謝パスウェイだった(表 3-15)。

表 3-15 クラスター4 の遺伝子がマップされたパスウェイ[34]

機能カテゴリー		パスウェイ	マップされた遺伝子
Metabolism	Amino acid metabolism	Valin, leucine and isoleucine degradation	ACADM
	Carbohydrate metabolism	Propanoate metabolism	ACADM
	Lipid metabolism	Fatty acid degradation	ACADM
	Metabolism of other amino acids	beta-Alanine metabolism	ACADM
Organismal Systems	Endocrine system	PPAR signaling pathway	ACADM
Human Diseases	Cancers	Non-small cell lung cancer	EML4

GWAS カタログでは、クラスター4 の 35 の古いハプロタイプブロックに含まれる SNP のうち、rs4949874 が血液中の代謝産物の割合に関連し、ACADM 遺伝子を介して、代謝関連の 4 つのパスウェイにマップされた(表 3-14)。

クラスター5

上位 1%ブロック内のクラスター5 には 70 遺伝子が含まれ(表 3-3)、そのうち 12 遺伝子が今回注目したパスウェイにマップされた(表 3-13)。すでに多様性が

知られている、エンリッチメント解析でも得られた、”Hapatitis C”、”Renal cell carcinoma”、”Endometrial cancer”、”Basal cell carcinoma”、”NAFLD”や、”Melanoma”、” Prostate cancer”、”Small cell lung cancer”にマップされたのは、APC、ARNT2、BRAF、JAK1、NDUFA8、PAK7 で、免疫システム関連の”Immune system”パスウェイにもマップされたのは、BRAF、PAK7 だった。BRAF、PAK7 は、”Chemokine signaling pathway”、”Natural killer cell mediated cytotoxicity”、”T cell receptor signaling pathway”にマップされた。これら免疫関連のパスウェイにマップされた遺伝子が受けた古い正の選択と近年の自然選択の影響が、パスウェイが得られた C 型肝炎やがんなどの疾患のアフリカとそれ以外の集団の多様性に関与している可能性がある。

GWAS カタログでは、クラスター5 の 160 の古いハプロタイプブロックに含まれる SNP のうち、12 SNP がすでに形質との関連が報告されており、そのうち、rs10056340 はヨーロッパ系集団でアレルギー感作(アレルギーが起こりやすい状態になること)に関連する(表 3-14)。rs10056340 は大きい F_{st} 値を示す(図 3-10)。rs10056340 は、rs10056340 を含む LD 領域とその近傍の 4 遺伝子、SLC25A46、TSLP、WDR36、CAMK4 のうち、TSLP によって、免疫関連パスウェイ“Jak-STAT signaling pathway”、“Cytokine–cytokine receptor interaction”にマップされた。TSLP は、すでに多くの炎症性疾患、ぜんそく、アレルギー性炎症、慢性閉塞性肺疾患などに関連していることが報告されており、これらの疾患の治療のターゲットとして考えられている。

そのほか、クラスター5 の遺伝子で “Jak-STAT signaling pathway”にマップされたのは JAK1 で、クラスター3 の IL10 もマップされていた。IL10 は Jak-STAT シグナル伝達経路の制御に関わっているサイトカインで、IL10 が利用するヤヌスキナーゼは JAK1 か Tyk2 であるが、本パイプラインからは JAK1

が得られており、いずれも“Jak-STAT signaling pathway”のレセプター付近にマップされた(図 3-11)。TSLP、JAK1 は、アフリカでの古い正の選択とその後の近年の自然選択を受け、IL10 とともにアレルギー感作のアフリカとそれ以外の集団間の多様性に関与している可能性があると考えられる。

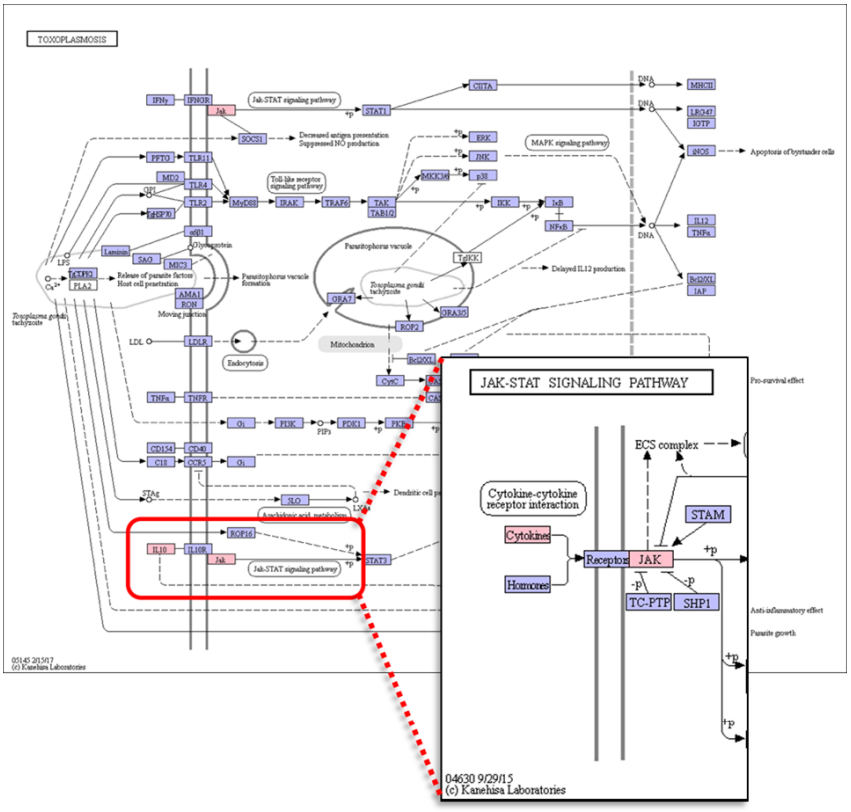


図 3-11　得られた遺伝子がマップされた Jak-STAT シグナル経路

パイプラインから検出された IL10、JAK1 は”Jak-STAT signaling pathway”のレセプター付近にマップされた。同様に、”Toxoplasmosis”に Jak-STAT singaling pathway の要素としてマップされた(赤線内部)。

3.5 考察

3.5.1 古いハプロタイプブロックの特徴

本研究で抽出された古いハプロタイプブロックは、その 75.32%がグループ I か III に属していた(図 3-6、図 3-7)。つまり、その大部分で、集団間で大きな差異がみられなかった。また、大きな集団間の違いを示す古いハプロタイプブロックの多くはクラスター5 に含まれ、YRI 特異的なハプロタイプを持つことが示された。すでに報告されている系統解析で、YRI がその他の集団から一番距離が離れていることが示されているが、本研究の結果はこのことと一致する[86, 87]。

しかし、本研究では、集団間の大きな違いが見られるクラスターの中では、クラスター2 の方がクラスター4 より大きな割合を占めることが示され、YRI と CEU の違いが、YRI と ASN の違いより大きく離れていることも示した。これは、既に報告されている系統解析の結果[86, 87]とは異なり、古いハプロタイプブロック特異的な結果だと考えられる。

3.5.2 上位 1%ブロックと F_{st}

F_{st} を使って、パイプラインが正の選択を検出したかどうか確かめた。本研究のパイプラインで計算されるスコアは、ハプロタイプ頻度に基づき集団の分化を測り、F_{st} はアレル頻度に基づきハプロタイプは考慮せずに集団の分化を測る指標であるが、正の選択は、アレル頻度の差異を大きくし、正の選択を受けた SNP は F_{st} の分布の上位にくるはずである[1]。

今 H_T を全集団のヘテロ接合度の期待値、H_S を各集団におけるヘテロ接合度の期待値の平均とする。全集団に対して、各集団でのヘテロ接合度の減少を、$F_{ST} = (H_T - H_S)/H_T$ で表す。集団の分化により生じた近交(inbreeding)でヘテロ

接合度が減少すると H_S が減少し F_{st} は大きくなる。$0 < F_{st} < 1$で、値が大きいほど遺伝的分化が大きいとみなされる。実際の計算は以下のように行った。ある SNP 座位における一方のアレルの頻度pを各集団(CEU、YRI、ASN)で計算し、その平均値をmとする。またその標本分散をVとする。この時、F_{st} は$F_{st} = V/m(1-m)$で計算することができる[88]。

本研究の結果では、上位 1%ブロックに含まれる F_{st} の平均は上位 5%ブロックに含まれる F_{st} の平均に比べて有意に大きいことが示された(p-値<0.05)。このことは、本手法が自然選択を受けた領域を抽出することを目的としていることとも整合性がある。その一方で、上位 0.5%ブロックに含まれる SNP のうち F_{st} が 0.2 より大きいものの割合は 54%、上位 1%ブロックに含まれる SNP のうち F_{st} が 0.2 より大きいものの割合は 49%で、上位 0.5%と上位 1%ブロックではあまり差がみられなかった。よって、パイプラインでは、上位 1%ブロックを集団間の違いが見られる古いハプロタイプブロックとして抽出することにした。

3.5.3 疾患関連形質の集団間の形質の違いと免疫システム関連パスウェイ

抽出された古いハプロタイプブロックのうち、集団間の違いがあるとみなされた上位 1%ブロックに含まれていた遺伝子や SNP は、免疫システム関連パスウェイや免疫システム関連疾患パスウェイに含まれる遺伝子である頻度が統計的に有意に高かった(p-値<0.05)。例えば、免疫システムパスウェイの“T cell receptor signaling pathway”、”Jak-STAT signaling pathway”、”Cytokine-cytokine receptor signaling pathway”の遺伝子の頻度が高かった。そこで、免疫システム関連のパスウェイにマップされた遺伝子が疾患関連形質の集団間の違いに関連する可能性について考察していく。

C 型肝炎は、発症率に関してアフリカ由来とヨーロッパ由来の集団間に違いが見られる疾患である[80]。T 細胞応答が C 型肝炎ウィルスの除去に重要な役割を果たすことが示されていることから、“Hepatitis C”パスウェイにマップされた遺伝子 GSK3B、BRAF、JAK1 以外にも、”T cell receptor signaling pathway”にマップされた IL10、GSK3B、PAK7 と C 型肝炎の集団間の違いとの関連を調べることも重要である。本研究で IL10 を構成する古いハプロタイプブロックの SNP rs1800896 はすでに C 型肝炎ウィルス感染応答との関連性が良く知られている[81]。

マラリア関連遺伝子としても知られている IL10 は[89-92]、本研究でも”Malaria”パスウェイにマップされ(表 3-13)、病原体に対する防御のための古い正の選択が IL10 に起こったことが示唆された。また、過去のインターロイキン/インターロイキン受容体遺伝子の解析においては、地域による病原体の種類数の違いと IL10 のアレル頻度の差異の相関も示されており[93]、本研究では、IL10 が病原体への適応のために受けたアフリカにおける古い正の選択と、近年のヨーロッパとアジアで起こった自然選択が、現在の T 細胞受容体シグナル伝達経路における C 型肝炎ウィルスの除去率などの集団間の違いをもたらし、C 型肝炎の形質の多様性に関与している可能性が示唆された。

セリアック病は集団間の発症率の違いがみられ、特にアフリカの集団で発症率が高い。セリアック病関連遺伝子として知られる SH2B3 が本研究の古い正の選択の候補として含まれていた。SH2B3 は NOD2 認識パスウェイの活性化に関与することがすでに知られており、SH2B3 は細菌感染に対する防御のための強い近年の正の選択をヨーロッパやアフリカで受けたと考えられている[85]。本研究により SH2B3 はアフリカでさらに古い正の選択もすでに受けていたことが検出され、SH2B3 にアフリカで古くから起こってきた細菌感染に対する防

御のための正の選択とヨーロッパとアジアで起こった近年の正の選択が、現在のセリアック病の発症率が集団によって多様性を見せることに関与している可能性が示唆された。

機能アノテーションによりアレルギー感作との関連が示された rs10056340 は TSLP を介して、免疫関連パスウェイ“Jak-STAT signaling pathway”、“Cytokine–cytokine receptor interaction”にマップされた。TSLP は多くの炎症性疾患、ぜんそく、アレルギー性炎症、慢性閉塞性肺疾患などに関連しており、これらの疾患の治療のターゲットとして考えられている。このほか、“Cytokine-cytokine receptor interaction”にマップされたのは IL10(Cluster3)、“Jak-STAT signaling pathway”にマップされたのは IL10 (Cluster 3)、JAK1 (Cluster 5)で、TSLP、IL10、JAK1 はいずれも“Jak-STAT signaling pathway”のレセプター付近にマップされた。病原体との相互作用に関連する Jak-STAT シグナル伝達経路[32]のレセプター付近のこれらの遺伝子に、アフリカでの古い正の選択とその後の近年の自然選択が起こり、アレルギー感作の集団間の違いをもたらしている可能性が示唆された。

3.5.4 疾患関連形質の集団間の形質の違いと機能モジュール

疾患の発症率の違いに関連すると注目したパスウェイにマップされた遺伝子のうち、イントロン以外に存在する機能性の高い SNP で、F_{st} が大きいものを含む遺伝子は、セリン/スレオニンキナーゼ(EC 2.7.11.-)やサイトカイン、酸化的リン酸化に関連するもので、また、”Calcium signaling pathway”(MYLK)や、”T cell receptor signaling pathway”(PAK7、L10)、”Jak-STAT signaling pathway”(IL10、rs10056340)、など、シグナル伝達に関連するものが多いことも分かった(表 3-16)。また、NADH 脱水素酵素モジュールの遺伝子も 2 つ見つ

かった(表 3-16)。集団間の違いの関連する共通な機能モジュール等を探していくことも、今後集団間の違いが生じるメカニズムを理解する上で重要であると考えられる。

表 3-16　本研究で注目したパスウェイにマップされたイントロン以外の SNP もしくは GWAS カタログで報告されている SNP のうち、アレル頻度の集団間の大きな違い(F_{st} >0.15)が確認できたもの

nsSNP	機能	遺伝子	KEGG での定義	モジュール
rs820463	missense	MYLK	myosin-light-chain kinase [EC:2.7.11.18]	-
rs2297345	missense	PAK7	p21-activated kinase 7 [EC:2.7.11.1]	-
rs1800896	upstream variant 2KB	IL10	interleukin 10	-
rs972890	upstream variant 2KB	NDUFS6	NADH dehydrogenase (ubiquinone) Fe-S protein 6	NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria
rs3818638	GWAS (Obsetity-related trait)	NDUFA8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex subunit 8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex
rs10056340	GWAS (Allergy)	CAMK4	calcium/calmodulin-dependent protein kinase IV [EC:2.7.11.17]	-
		TSLP	thymic stroma; lymphopoietin	-
		WDR36	WD repeat domain 36	-
		SLC25A46	solute carrier family 25, member 46	-

第 4 章　総括

疾患関連形質の集団間の多様性についての遺伝情報は、公衆衛生、個別化予防にとって重要である。疾患関連形質の集団間の多様化が起こる原因は、環境要因と遺伝的要因が考えられるが、本研究では遺伝的要因に着目し、疾患関連形質の多様化に関与する可能性の高い遺伝的要因の探索を試みた。その際に、疾患関連形質の集団間の多様性に関連する座位を系統的に収集することが必要になる。最近では自然選択のなかでも古くに起こったものと近年に起ったものとがあり(表 1-1)、古い自然選択を受けかつ集団で異なる近年の自然選択を受けるという長期的な進化プロセスが、自己免疫疾患などの免疫を介したありふれた疾患や代謝疾患の集団間の形質の違い、例えば発症率の地域差の一因となっていると考えられるようになってきた[30]。そこで本研究では、今まであまり研究されていなかった古い正の選択に注目し、古い正の選択を受けかつその後の多様化が起こったゲノム領域をゲノムワイドな SNP から検出する手法を独自に開発することにより、実際にそのような進化プロセスを経て生じた疾患関連形質の多様性があるのかどうか検討した。

本研究では、まず新たに HHD というディプロタイプ間距離の推定法を提案し、HHD を利用することにより、古い正の選択が起こりその後に多様化が起こった領域を検出するパイプラインを構築した。開発したパイプラインでは、まず、ヒトゲノムの最も一般的な多型である SNP データから、古いハプロタイプブロックを抽出し、古い正の選択の候補領域とする。次に、古い正の選択の候補領域が、集団によって異なる近年の自然選択を受けてきたかどうかを評価し、近年の自然選択の検出感度の向上を試みた。

HapMap の遺伝子型データに本パイプラインを適用したところ、抽出された古い正の選択を受けかつ近年の自然選択による多様化された候補領域に含まれ

ていた遺伝子の 75%が既存の近年の正の選択を検出する研究で報告されていないものであった。これは、本研究で古い正の選択を考慮した結果、新たに検出された近年の正の選択の候補遺伝子である可能性がある。またパイプラインによって抽出された古いハプロタイプブロックに含まれる遺伝子やSNP に対して機能アノテーションを行った際には、得られた遺伝子は、免疫システム、感染症、集団間の多様性が知られているありふれた疾患関連パスウェイにマップされた遺伝子の頻度が高かった。そして、いくつかの疾患に関しては、免疫システム・感染症にマップされた遺伝子が、発症率の多様性に関与する可能性を考察することができた。さらに、それらの遺伝子のうち、機能性が高く F_{st} も大きいような SNP を含むものに関しては、シグナル伝達に関するものや、NADH 脱水素酵素モジュールのものが多かった。

今後、ありふれた疾患の発症率の違いに関連する共通の機能モジュールの探索や、検出された SNP のタンパク質の立体構造・タンパク質間相互作用への影響など、検出された遺伝子がどのように発症頻度の違いのメカニズムに作用するのか、さらに詳細に検討して行く必要がある。また、本研究で集団間のアレル頻度の違いが生じた背景として仮定した自然選択について、どのような淘汰が働いたのかを詳細に検討していきたい。

本研究で提案したパイプラインの改良点も多くの観点から考えられる。まず、本研究で立てた仮説の見直しである。本研究では集団間のアレル頻度が異なる原因として自然選択のみを仮定したが、実際にはこの他にも遺伝的浮動やネアンデルタールとの交雑など様々な原因が考えられる。これらの可能性を考慮した上で、新たな仮説を検討するなど、本研究の結果を検証していく必要がある。次に、使っている統計モデルの改良である。本研究では、スコア分布が正規分布に従うことを仮定して解析を進めたが、本研究で用いたスコア分布により近

い統計モデルを仮定することも試みていきたい。また、大量データに対応するために古いハプロタイプブロックの検出手法についても改良していく必要がある。機能アノテーションについては、アノテーションデータの充実についても検討していきたい。本研究では、遺伝子コード領域のみを使ったが、本研究を通して、非コード領域にも重要な形質と関連する SNP が多数存在することが分かり、今後の研究では、tRNA, rRNA, microRNA などの non-coding RNA の情報を使うことが必要であると考えられる。遺伝子コード領域の上流、例えば 1Mb 以内、にある SNP についても解析に追加していきたい。

進化的な観点からは、ヒトと相互作用する重要な環境要因の 1 つとして食物が考えられ、食物ゲノムの進化的な解析もヒトの進化の歴史を理解するのに重要であると考えられる。イネや小麦など食物として用いられてきた植物ゲノムの進化的な解析も、食習慣による疾患について理解を深めるために行っていきたい。また、本研究では、疾患の発症率の違いを説明する遺伝的要因のうち、集団間で共通してリスクとなるような感受性アレルに注目した。一方で、集団特異的な遺伝的要因も多く存在すると考えられ、集団特異的な感受性アレルを探す手法も今後必要であると考えられる。

謝辞

本研究を進めるにあたり、数々のご指導、ご助言により、研究を導いて下さった渋谷哲朗准教授、緒方博之教授、五斗進教授、金久實教授に深く感謝いたします。また、他研究室にもかかわらず、学生時に多くのご指導、ご助言をして下さった山口類准教授、山田亮教授に心より感謝いたします。最後に、学生時の金久研究室・渋谷研究室の皆様、緒方研究室の皆様、現在の職場の皆様に心より感謝いたします。

参考文献

[1] Myles S, Davison D, Barrett J, Stoneking M, Timpson N. Worldwide population differentiation at disease-associated SNPs. BMC Med Genomics. 2008;1:22. PubMed PMID: 18533027; PubMed Central PMCID: PMC2440747.

[2] Adeyemo A, Rotimi C. Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics. 2010;13(2):72-9. PubMed PMID: 19439916; PubMed Central PMCID: PMC2835382.

[3] Marigorta UM, Lao O, Casals F, Calafell F, Morcillo-Suárez C, Faria R, et al. Recent human evolution has shaped geographical differences in susceptibility to disease. BMC Genomics. 2011;12:55. PubMed PMID: 21261943; PubMed Central PMCID: PMC3039608.

[4] Luisi P, Alvarez-Ponce D, Dall'Olio GM, Sikora M, Bertranpetit J, Laayouni H. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Mol Biol Evol. 2012;29(5):1379-92. PubMed PMID: 22135191.

[5] Young JH, Chang YP, Kim JD, Chretien JP, Klag MJ, Levine MA, et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. PLoS Genet. 2005;1(6):e82. PubMed PMID: 16429165; PubMed Central PMCID: PMC1342636.

[6] Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20(3):393-402. PubMed PMID: 20086244; PubMed Central PMCID: PMC2840981.

[7] Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat Genet. 2011;43(6):570-3. PubMed PMID: 21602798.

[8] Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE, et al. Ethnic and racial differences in the smoking-related risk of lung cancer. N Engl J Med. 2006;354(4):333-42. PubMed PMID: 16436765.

[9] Stafford HS, Saltzstein SL, Shimasaki S, Sanders C, Downs TM, Sadler GR. Racial/ethnic and gender disparities in renal cell carcinoma incidence and survival. J Urol. 2008;179(5):1704-8. PubMed PMID: 18343443.

[10] Fejerman L, Romieu I, John EM, Lazcano-Ponce E, Huntsman S,

Beckman KB, et al. European ancestry is positively associated with breast cancer risk in Mexican women. Cancer Epidemiol Biomarkers Prev. 2010;19(4):1074-82. PubMed PMID: 20332279.
[11] Setiawan VW, Pike MC, Kolonel LN, Nomura AM, Goodman MT, Henderson BE. Racial/ethnic differences in endometrial cancer risk: the multiethnic cohort study. Am J Epidemiol. 2007;165(3):262-70. PubMed PMID: 17090617.
[12] Cormier JN, Xing Y, Ding M, Lee JE, Mansfield PF, Gershenwald JE, et al. Ethnic differences among patients with cutaneous melanoma. Arch Intern Med. 2006;166(17):1907-14. PubMed PMID: 17000949.
[13] Lohmueller KE, Mauney MM, Reich D, Braverman JM. Variants associated with common disease are not unusually differentiated in frequency across populations. Am J Hum Genet. 2006;78(1):130-6. PubMed PMID: 16385456.
[14] Bains RK. African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases. Evol Med Public Health. 2013;2013(1):118-34. PubMed PMID: 24481193; PubMed Central PMCID: PMC3868406.
[15] Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138(1):103-41. PubMed PMID: 23333322.
[16] Kim S, Misra A. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289-320. PubMed PMID: 17391067.
[17] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. PubMed PMID: 26432245; PubMed Central PMCID: PMC4750478.
[18] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822. PubMed PMID: 23300413; PubMed Central PMCID: PMC3531285.
[19] Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet. 2002;3(8):611-21. PubMed PMID: 12154384.
[20] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263-5.

PubMed PMID: 15297300.
[21] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225-9. PubMed PMID: 12029063.
[22] Kimura R, Fujimoto A, Tokunaga K, Ohashi J. A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS One. 2007;2(3):e286. PubMed PMID: 17356696.
[23] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832-7. PubMed PMID: 12397357.
[24] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004;74(6):1111-20. PubMed PMID: 15114531; PubMed Central PMCID: PMC1182075.
[25] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312(5780):1614-20. PubMed PMID: 16778047.
[26] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009;19(5):826-37. PubMed PMID: 19307593; PubMed Central PMCID: PMC2675971.
[27] Klimentidis YC, Abrams M, Wang J, Fernandez JR, Allison DB. Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. Hum Genet. 2011;129(4):407-18. PubMed PMID: 21188420; PubMed Central PMCID: PMC3113599.
[28] Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 2007;5(7):e171. PubMed PMID: 17579516; PubMed Central PMCID: PMC1892573.
[29] Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008;40(3):340-5. PubMed PMID: 18246066.
[30] Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014;15(6):379-93.

PubMed PMID: 24776769.
[31] Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010;329(5993):841-5. PubMed PMID: 20647424; PubMed Central PMCID: PMC2980843.
[32] Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7(11):e1002355. PubMed PMID: 22072984; PubMed Central PMCID: PMC3207877.
[33] Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 2005;77(2):171-92. PubMed PMID: 16001361; PubMed Central PMCID: PMC1224522.
[34] Onuki R, Yamaguchi R, Shibuya T, Kanehisa M, Goto S. Revealing phenotype-associated functional differences by genome-wide scan of ancient haplotype blocks. PLoS One. 2017;12(4):e0176530. Epub 2017/04/26. PubMed PMID: 28445522; PubMed Central PMCID: PMC5406033.
[35] Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006;38(11):1251-60. Epub 2006/10/22. PubMed PMID: 17057719.
[36] Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451(7181):998-1003. PubMed PMID: 18288195.
[37] RDC T. R: A language and environment for statistical computing R Foundation for Statistical Computing2007.
[38] Ward JH, Hook ME. Application of a hierarchical grouping procedure to a problem of grouping profiles. Educ. and Phsychol Mesurement. 1963;23.
[39] Kaufman L, Rousseeuw P. Finding groups in data: An introduction to cluster analysis: John Wiley and Sons, New York; 1990.
[40] Ester M, Kriegel HP, Sander J, Xu X. Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96): AAAI Press; 1996.
[41] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25. PubMed

PMID: 3447015.
[42] Gao X, Martin ER. Using allele sharing distance for detecting human population stratification. Hum Hered. 2009;68(3):182-91. PubMed PMID: 19521100.
[43] Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet. 2007;80(6):1171-8. Epub 2007/04/20. PubMed PMID: 17503334; PubMed Central PMCID: PMC1867104.
[44] Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, et al. Genetic similarities within and between human populations. Genetics. 2007;176(1):351-9. Epub 2007/03/04. PubMed PMID: 17339205; PubMed Central PMCID: PMC1893020.
[45] Beaty TH, Fallin MD, Hetmanski JB, McIntosh I, Chong SS, Ingersoll R, et al. Haplotype diversity in 11 candidate genes across four populations. Genetics. 2005;171(1):259-67. Epub 2005/06/18. PubMed PMID: 15965248; PubMed Central PMCID: PMC1456517.
[46] Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, et al. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res. 2001;11(1):143-51. PubMed PMID: 11156623; PubMed Central PMCID: PMC311030.
[47] Cover TM, Thoman JA. Elements of Information Theory: John Wiley & Sons, INC; 1991.
[48] Isaev A. Introduction to mathematical methods to bioinformatics: Springer; 2004.
[49] Lesk AM. Introduction to bioinformatics. Second ed: Oxford; 2005.
[50] Li J, Jiang T. Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics. 2005;21(24):4384-93. Epub 2005/10/25. PubMed PMID: 16249262.
[51] Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet. 2003;72(4):891-902. Epub 2003/02/27. PubMed PMID: 12610778; PubMed Central PMCID: PMC1180352.
[52] Rastas P, Koivisto PM, Mannila H, Ukkonen E. A hidden Morkov technique for haplotype reconstruction. Lecture Notes in Bioinformatics. 2005;3692.

[53] Rabiner LR, Juang BH. An introduction to hidden Markov models. IEEE ASSP Mag. 1986.
[54] Durbin R, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis: Cambridge Press, New York; 1998.
[55] Onuki R, Yamada R, Yamaguchi R, Kanehisa M, Shibuya T. Population model-based inter-diplotype similarity measure for accurate diplotype clustering. J Comput Biol. 2012;19(1):55-67. PubMed PMID: 22149683.
[56] Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol. 2010;6(10):e1000954. Epub 2010/10/14. PubMed PMID: 20976246; PubMed Central PMCID: PMC2954823.
[57] Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415-25. PubMed PMID: 20479773.
[58] Chung PY, Beyens G, Guañabens N, Boonen S, Papapoulos S, Karperien M, et al. Founder effect in different European countries for the recurrent P392L SQSTM1 mutation in Paget's Disease of Bone. Calcif Tissue Int. 2008;83(1):34-42. Epub 2008/06/10. PubMed PMID: 18543015.
[59] Gonzalez E, Bamshad M, Sato N, Mummidi S, Dhanda R, Catano G, et al. Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. Proc Natl Acad Sci U S A. 1999;96(21):12004-9. PubMed PMID: 10518566; PubMed Central PMCID: PMC18402.
[60] Haiman CA, Stram DO, Pike MC, Kolonel LN, Burtt NP, Altshuler D, et al. A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort. Hum Mol Genet. 2003;12(20):2679-92. Epub 2003/08/27. PubMed PMID: 12944421.
[61] Ahmad T, Neville M, Marshall SE, Armuzzi A, Mulcahy-Hawes K, Crawshaw J, et al. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Hum Mol Genet. 2003;12(6):647-56. PubMed PMID: 12620970.
[62] Gaudieri S, Leelayuwat C, Tay GK, Townend DC, Dawkins RL. The major histocompatability complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. J Mol Evol. 1997;45(1):17-23. PubMed PMID: 9211730.
[63] The International HapMap Consortium. A haplotype map of the human

genome. Nature. 2005;437(7063):1299-320. PubMed PMID: 16255080.
[64] Tanaka E, Taniguchi A, Urano W, Nakajima H, Matsuda Y, Kitamura Y, et al. Adverse effects of sulfasalazine in patients with rheumatoid arthritis are associated with diplotype configuration at the N-acetyltransferase 2 gene. J Rheumatol. 2002;29(12):2492-9. PubMed PMID: 12465141.
[65] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001;29(2):229-32. PubMed PMID: 11586305.
[66] Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature. 2003;423(6939):506-11. Epub 2003/04/30. PubMed PMID: 12724780.
[67] Wang Y, Rannala B. In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. Am J Hum Genet. 2005;76(6):1066-73. Epub 2005/04/07. PubMed PMID: 15818531; PubMed Central PMCID: PMC1196444.
[68] Gao X, Starmer J. Human population structure detection via multilocus genotype clustering. BMC Genet. 2007;8:34. PubMed PMID: 17592628.
[69] Vapnik V. Statistical Learning Theory: Wiley, NY; 1998.
[70] Brinza D, Zelikovsky A, editors. Discrete Methods for Association Search and Status Prediction in Genotyep Case-Control Studies. IEEE 7-th International symposium on Bioinformatics and Bioengineering; 2007.
[71] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Comput Biol. 2000;7(1-2):95-114. PubMed PMID: 10890390.
[72] Onuki R, Shibuya T, Kanehisa M. New kernel methods for phenotype prediction from genotype data. Genome Inform. 2010;22:132-41. PubMed PMID: 20238424.
[73] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(Database issue):D1001-6. PubMed PMID: 24316577; PubMed Central PMCID: PMC3965119.
[74] Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19(7):1316-23. PubMed PMID: 19498102.

[75] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277-80. PubMed PMID: 14681412.
[76] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61. PubMed PMID: 17943122; PubMed Central PMCID: PMC2689609.
[77] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-8. PubMed PMID: 17943131.
[78] Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 2010;327(5967):883-6. PubMed PMID: 20056855.
[79] Liu X, Ong RT, Pillai EN, Elzein AM, Small KS, Clark TG, et al. Detecting and characterizing genomic signatures of positive selection in global populations. Am J Hum Genet. 2013;92(6):866-81. PubMed PMID: 23731540; PubMed Central PMCID: PMC3675259.
[80] Pearlman BL. Hepatitis C virus infection in African Americans. Clin Infect Dis. 2006;42(1):82-91. PubMed PMID: 16323096.
[81] Oleksyk TK, Thio CL, Truelove AL, Goedert JJ, Donfield SM, Kirk GD, et al. Single nucleotide polymorphisms and haplotypes in the IL10 region associated with HCV clearance. Genes Immun. 2005;6(4):347-57. PubMed PMID: 15815689.
[82] Bambha K, Belt P, Abraham M, Wilson LA, Pabst M, Ferrell L, et al. Ethnicity and nonalcoholic fatty liver disease. Hepatology. 2012;55(3):769-80. PubMed PMID: 21987488; PubMed Central PMCID: PMC3278533.
[83] Kwasniak LA, Garcia-Zuazaga J. Basal cell carcinoma: evidence-based medicine and review of treatment modalities. Int J Dermatol. 2011;50(6):645-58. PubMed PMID: 21595656.
[84] Xu L, Kitade H, Ni Y, Ota T. Roles of Chemokines and Chemokine Receptors in Obesity-Associated Insulin Resistance and Nonalcoholic Fatty Liver Disease. Biomolecules. 2015;5(3):1563-79. Epub 2015/07/21. PubMed PMID: 26197341; PubMed Central PMCID: PMC4598764.
[85] Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, et al. Evolutionary and functional analysis of celiac risk loci reveals

SH2B3 as a protective factor against bacterial infection. Am J Hum Genet. 2010;86(6):970-7. PubMed PMID: 20560212; PubMed Central PMCID: PMC3032060.
[86] Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 1994;368(6470):455-7. PubMed PMID: 7510853.
[87] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4. PubMed PMID: 18292342.
[88] Weir BS, Cockerham CC. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution. 1984;38(6):1358-70. PubMed PMID: 28563791.
[89] Peyron F, Burdin N, Ringwald P, Vuillez JP, Rousset F, Banchereau J. High levels of circulating IL-10 in human malaria. Clin Exp Immunol. 1994;95(2):300-3. PubMed PMID: 8306505; PubMed Central PMCID: PMC1534910.
[90] Niikura M, Inoue S, Kobayashi F. Role of interleukin-10 in malaria: focusing on coinfection with lethal and nonlethal murine malaria parasites. J Biomed Biotechnol. 2011;2011:383962. PubMed PMID: 22190849; PubMed Central PMCID: PMC3228686.
[91] Huang BH, Liao PC. Tracing evolutionary relicts of positive selection on eight malaria-related immune genes in mammals. Innate Immun. 2015;21(5):463-76. PubMed PMID: 25201904.
[92] Pereira VA, Sánchez-Arcila JC, Teva A, Perce-da-Silva DS, Vasconcelos MP, Lima CA, et al. IL10A genotypic association with decreased IL-10 circulating levels in malaria infected individuals from endemic area of the Brazilian Amazon. Malar J. 2015;14:30. PubMed PMID: 25627396; PubMed Central PMCID: PMC4334410.
[93] Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, Clerici M, et al. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med. 2009;206(6):1395-408. Epub 2009/05/25. PubMed PMID: 19468064; PubMed Central PMCID: PMC2715056.