病原性微生物ゲノムの分子進化についての研究

病原性微生物ゲノムの分子進化についての研究

A study on the molecular evolution of pathogenic microbial genomes

2020 年

岐阜大学大学院連合創薬医療情報研究科

医療情報学専攻

吉崎　純夫

目次

Ⅰ．序論

ヒトを含めたあらゆる生物は遺伝情報という設計図に基づいて構築されており、各生物の特徴に応じて遺伝情報は大きく異なっている。この情報は DNA を構造基盤とするゲノム（genome）とよばれ、ゲノムが生物の有する様々な特徴を決めている。各種生物のゲノムは、その進化の過程で大きく変化してきており、これが現存生物の多様性を生み出している。一般に、真核生物ゲノムが各染色体に分断された直鎖状 DNA であるのに対して、細菌などの原核生物ゲノムは環状構造の DNA である。現在、これら各種生物のゲノムを解読、すなわち全塩基配列を明らかにする研究が急速に進行しつつある。NCBI データベースによれば、ヒトを含む真核生物では 12,000 を越える生物種のゲノムが完全解読されている。これに対して､原核生物では完全解読されたゲノムがすでに 250,000 生物種を上回っている。真核生物に存在するイントロンが原核生物ゲノムにはほとんどないために、原核生物ゲノムでは遺伝子の同定や推定が容易となっている。こうした多くの生物種で大量に得られているゲノムデータを解析することにより、様々なことが分かるようになってきた。

微生物の中には病原性を示すものが多く知られているが、この病原性は遺伝子の変化を伴う進化を経て獲得されたものである。したがって、病原性微生物が進化する過程で積極的に変化してきた遺伝子を同定することができれば、病原性発現のしくみを明らかにできる。こうした遺伝子の進化的特性を調べるためには、現存生物のゲノムを用いた比較ゲノム解析が大きな威力を発揮する。すなわち、進化の原動力はダーウィンの提唱した自然選択であるため、比較ゲノム解析によって正の自然選択（positive selection）を受けた遺伝子を検出する。

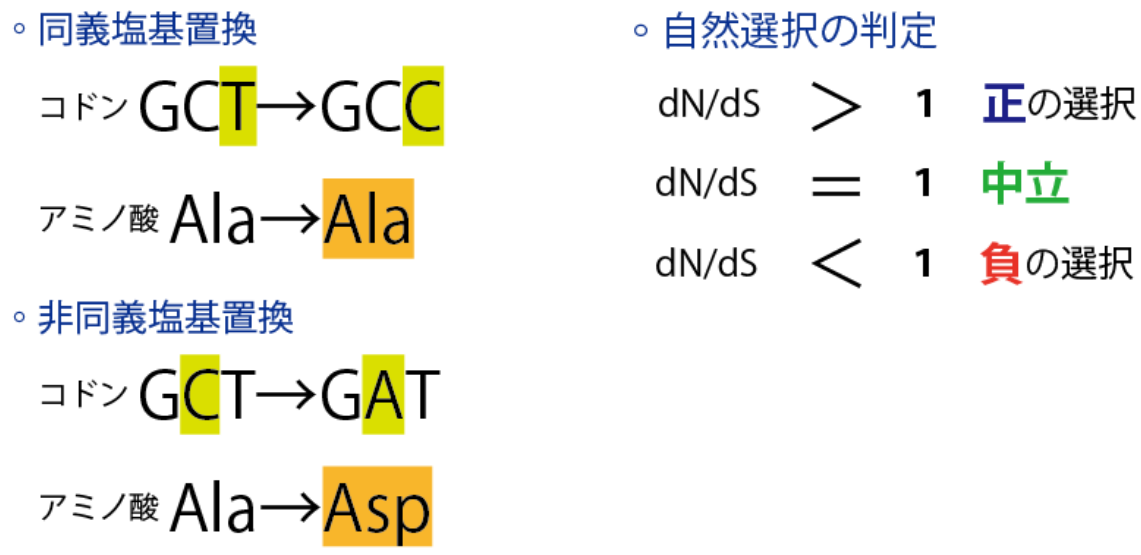

図１　塩基配列変化と進化

塩基配列の一部が変化した時、翻訳されるアミノ酸が変化しない同義塩基置換（dS）と変化する非同義置換（dN）がある。この dN と dS の比によって、正の自然選択を検出する。

この手法では、進化の過程で変化した塩基配列に着目するが、塩基配列が変化しても翻訳されるアミノ酸が変わる場合（非同義置換：dN）と変わらない場合（同義置換：dS）がある。アミノ酸が変わる非同義置換が高頻度で起きた遺伝子は、アミノ酸が大きな速度で変化してきていることを意味しており、正の自然選択（正の選択）が働いたことを示唆する（図１）。この原理に従って、dN と dS の比（dN/dS）を基準にして正の選択を受けた遺伝子を検出することができる。そして、正の選択が働いているアミノ酸配列部分は、速く進化してきた部分であると同時に、その遺伝子にとって新規機能の獲得に重要な部分であると推定できる（図２）。したがって、病原性を示す微生物のみに特徴的な正の自然選択は、その微生物の病原因子やそれにかかわる遺伝子の推定に利用できることになる。また、組み換え（recombination）も微生物が進化する上で重要な原動力となっている。組み換えとは、塩基配列が徐々に変化して起こる進化とは異なり、他生物に由来する遺伝子の一部分が丸ごと入れ替わるように変化する現象である。そのために、近い属種の遺伝子と比較した場合には、その生物だけに異常に大きな配列変化が認められる。

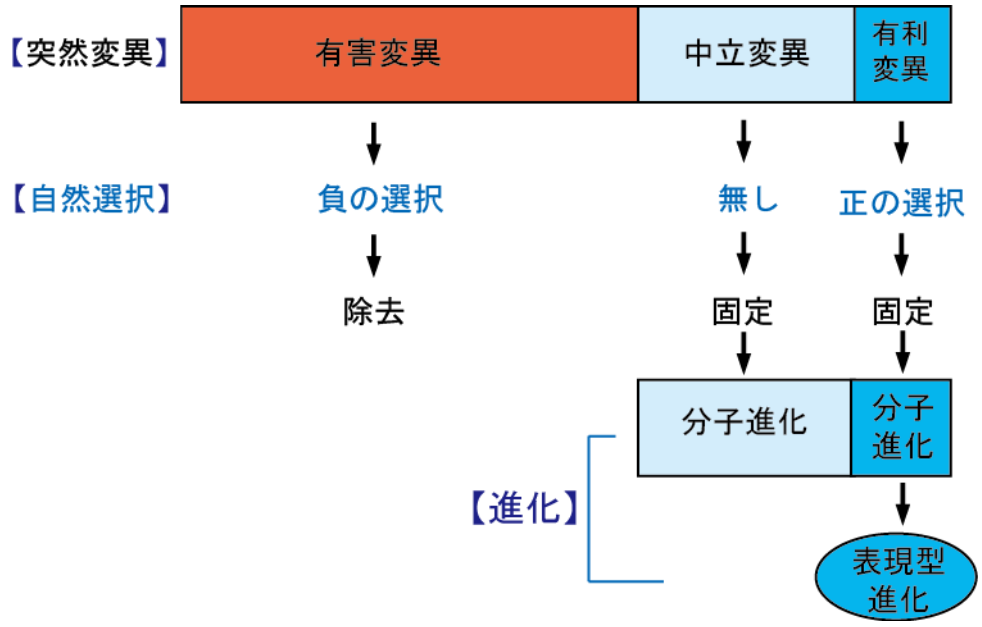

図２　自然選択と進化の関係

遺伝子に起きた変化が有害変異（負の自然選択）であった場合、個体の生存が困難となり種集団から除去されていく。中立変異と有利変異は種集団に存続することが可能で、その中でも正の選択は集団中に急速に固定し、その種に特異的な表現型を進化させる。

実際に多くの微生物の進化において、正の選択と組み換えは進化の重要な推進力といえる[1, 2]。鈴木と Stanhope らの大腸菌研究[3]、Urwin ら[4]や Andrews ら[5]の髄膜炎菌研究、Smith らの *Pseudomonas aeruginosa* 研究[6]、および Stanhope らの *Streptococcus pneumoniae* 研究[7]のように、微生物病原体における病原性遺伝子の進化に、正の選択が

寄与することを示唆する報告が多い。さらに、微生物のゲノム全体で正の選択を調べる研究は、重要な病原体の進化に関する包括的な探求に貢献してきた[8-10]。その中には、連鎖球菌属[11]、*Salmonella* 血清型[12]、大腸菌[13]、*Mycobacterium tuberculosis* [14]、*Botrytis* spp*[15]*、*Trypanosoma cruzi[16]*、などの研究が含まれている。こうした多様な研究から、環境に適応するために有利な進化を固定する正の選択は、感染プロセスの最適化と宿主免疫応答からの防御という両方の観点において、微生物の病原性獲得の過程で最も重要な原動力といえる[17]。

本研究で注目した *Bacteroides fragilis* はグラム陰性嫌気性菌であり、ヒトの通常の腸内細菌叢の主要構成菌種である。糞便分離株の *B. fragilis* の生存細胞数は、他の腸内 *Bacteroides* 属の生存細胞数の 10～100 分の 1 程度であるが[18]、腹腔内感染、膿瘍、および血液から最も頻繁に分離される病原性嫌気性菌である[19, 20]。*B. fragilis* の病原性には、膜多糖[21, 22]、プロテアーゼ[23]および *B. fragilis* 毒素[24]などが関与しているとの研究がされている。また、腸管外感染症に寄与する酸化ストレスなどへの抵抗性因子も報告されている[25]。これらの要因は病原性にとって重要であると考えられてきたが、現時点ではそれらの相対的な寄与の程度は知られておらず、他の可能なメカニズムも考慮しなければならない。

トキソプラズマ原虫 *Toxoplasma gondii* は、アピコンプレクサ門に属する組織嚢胞形成コクシジウム寄生虫である。*T. gondii* は多くの野生動物および家畜に感染し、ヒトに人畜共通感染症を引き起こす[26]。免疫不全の宿主では脳炎を誘発し、新生児の死亡率および子供の先天異常の発生率を上昇させる[27]。そのユニークな細胞内毒性戦略は、組織嚢胞内で潜在的にゆっくりと成長するブラディゾイトとして生き残ることである。*T. gondii* は、中間宿主の組織に存在する無性ブラディゾイトおよびその最終宿主によって排泄される環境耐性オーシストの摂取後に水平伝播する可能性がある[28]。*T. gondii* 感染の生物学的メカニズムは集中的に研究されており、分泌キナーゼなどのさまざまな病原因子が侵入時に重要な役割を果たすことが示されている[29]。また、比較ゲノム解析により、分泌タンパク質が宿主の多様化に寄与していることも明らかにされている[30, 31]。これらの要因は、*T. gondii* の病原性にとって重要であると考えられているが、*T. gondii* の系統間での進化的多様化の程度は不明であり、近縁の非病原性原虫との差異についても考慮する必要がある。

本研究では、病原性微生物のなかで原核生物として腸内細菌の *B. fragilis*、また真核生物として *T. gondii* を選び、これらのゲノムを用いて正の自然選択を検出するための進化解析を行った。正の自然選択の検出には、各アミノ酸サイトや進化系統で個々に自然選択の検出が可能な最尤法による手法を用い、病原性の獲得に至る進化的変化を同定した。また、正の自然選択が検出されたアミノ酸サイトについて、タンパク質の立体構造上での配置を解析し、タンパク質機能との関連を考察した。

Ⅱ．材料と方法

Ⅱ－１．解析に用いたゲノムデータセット

Bacteroides 属８菌種のゲノム配列は、FASTA フォーマットで統合微生物ゲノム（IMG）データベース（http://img.jgi.doe.gov/cgi-bin/w/main.cgi）からダウンロードした。また、９系統の *T. gondii* と近縁の *Neospora caninum*、*Hammondia hammondi* のゲノム配列は、ToxoDB database (http://toxodb.org/toxo/)からダウンロードした。遺伝子オルソログの同定は、OrthoMCL（v1.4）[32]を使用して、BLAST E 値のカットオフが 10^{-5} および 1.5 のインフレーションパラメーターで実行した。早期終止コドンまたは 50 コドンより短い配列を持つ遺伝子は、その後の分析から除外した。ベン図は、Vennerable R パッケージ（http://r-forge.r-project.org/projects/vennerable ）で作成した。

Ⅱ－２．多重アラインメント、組み換え検出、および系統樹の作成

同じクラスターにグループ化されたコア遺伝子オルソログは、プログラム MUSCLE [33] あるいは、PRANK [34]をデフォルト設定で使用して、多重アラインメントを行った。各アラインメントはコドンレベルで実施した。また、アミノ酸配列から塩基配列への変換は、PAL2NAL ソフトウェアパッケージを使用した[35]。

組み換えを受けた遺伝子は本来の進化とは異質な配列パターンを示すので、正の選択検出や系統樹の推定に影響を与えることが示唆されている [36]。そのため、HyPhy プログラム [37, 38]を用いた Single break point（SBP）解析、あるいは GARD algorithm を用いて、組み換えの検出を行った。

対象微生物の進化系統関係を推定するために、コア遺伝子クラスターから組換え遺伝子クラスターを除外することにより作成されたオルソログクラスターのアラインメントを単一の配列に連結した。結果として得られた連結塩基配列を用いて、PhyML（Phylogenetic Estimation Using Maximum Likelihood）プログラム[39, 40]を GTR＋γ 塩基置換モデルの条件下で実行した。ブランチサポートは、PhyML プログラムに実装されたノンパラメトリック下平 - 長谷川様式（SH 様式）の近似尤度比検定（aLRT）を使用して計算した。

Ⅱ－３．正の自然選択の検出

最尤法を用いた PAML version 4.5 あるいは version 4.7 [41]の codeml プログラムを使用して、多重アラインメントから自然選択の検出を行った。推定に用いた塩基置換モデルは、site model と branch-site model である。site model は、系統全体についてアミノ酸サイトごとに正の選択の検出を行う方法である [42]。検出に際しては、中立進化モデル（M1a model）と正の選択モデル（M2a model）のそれぞれで計算を実行し、両者の尤度を比較して χ^2 検定を行う。branch-site model では、系統樹上のある 1 つの進化経路における自然選択の検出を可能にするため、中立進化モデル（A1 model）と正の選択モデル（A model）の尤度を比較して χ^2 検定を行う方法である [43]。

PAML では、このそれぞれのモデルについての対数尤度が算出されるので、これらを用いて尤度比検定を行い p 値を計算した。p 値の計算には、PAML に付属する chi2 プログラムを使用した。自然選択の程度（選択圧）は、非同義置換／同義置換（dN/dS=ω）で表され、ω が 1 より大きい時（ω>1）では正の選択を受けており、ω が 1 の時（ω=1）は中立進化、ω が 1 より小さい時（ω<1）は負の選択を受けていることを示す [44]。PAML では、この ω の数値もそれぞれのモデルに対応して推定される。実際の codeml プログラムや chi2 プログラムの実行では、全ての遺伝子の多重アラインメントを連続して処理する必要があるために、Perl でスクリプトプログラムを作成して、各プログラムを連続実行した。

多重検定補正は、Benjamini と Hochberg [45]によって報告された手順を使用し、q 値は R プログラムの QVALUE を用いて算出した。
（https://www.bioconductor.org/packages/release/bioc/html/qvalue.html）

Ⅱ－４．正の自然選択を受けたアミノ酸サイトの同定と解析

codeml プログラムで正の選択を可能にするモデルの場合、Bayes Empirical Bayes アプローチを使用して、各コドンが正の選択の下で進化した事後確率（PP）が計算される[46]。ω > 1 クラスに由来する各アミノ酸部位で PP が大きいアミノ酸は、正の自然選択を受けてきたと推定される。本研究では、PP> 0.95 のカットオフを使用して、これ以上の PP 値を示すアミノ酸を正の選択下のアミノ酸として特定した。

正の選択を受けたアミノ酸の物理化学的性状の変化は、TreeSAAP 3.2 プログラムを使用して推定した[47]。このプログラムは、進化経路中のアミノ酸変化に伴うタンパク質の物理化学的性状変化を推定し、幾つかの判定基準に従ってタンパク質全体への影響を明らかにする。判定基準としては、合計 31 の構造的および生化学的アミノ酸特性の変化が考慮され、1～8 のスコアが与えられる。スコア 8 が最も重要な変化で、本研究では、根本的な機能的または構造的変化（すなわち、スコア 6～8）を示すカテゴリーのアミノ酸変化のみを検討した。

Ⅱ－５．タンパク質の３次元立体構造解析

正の選択を示した遺伝子によってコードされたタンパク質の三次元（3D）構造モデルを構築するために、Phyre2（Protein Homology / analogY Recognition Engine）サーバー（http：//www.sbg .bio.ic.ac.uk / phyre2 / html / page.cgi？id = index）[48]を使用してホモロジーモデリングを行った。本研究では multi template で行う intensive mode の計算条件を用いた。この計算条件では、一部の構造は *ab initio* 法によって推定されるが、タンパク質の全構造を得ることができる。推定された立体構造は、一般公開されている分子表示プログラムである PyMOL (https://pymol.org/2/)によって表示した。

膜タンパク質については、その膜内外ドメインの推定を行った。このドメイン予測には β ストランドの推定が良好に行える PRED-TMBB を利用した

(http://bioinformatics.biol.uoa.gr/PRED-TMBB/)。得られたドメイン配置のデータは、TOPO2(http://www.sacs.ucsf.edu/TOPO2/)で、アミノ酸ごとに表示し post script ファイルに保存した。

Ⅱ－６．遺伝子オントロジーおよびエンリッチメント解析

正の選択が検出された遺伝子の生物学的機能を理解するために、遺伝子オントロジー（遺伝子注釈）とそのエンリッチメント解析を行った。解析に用いた遺伝子注釈データは、*Bacteroides* については、COG（Clusters of Orthologous Groups）機能分類による遺伝子注釈を IMG データベースから取得して用いた。各 COG に該当する遺伝子は、Perl script で検索・同定し、エンリッチメントの有意性は二項検定で推定した。*Toxoplasma* については、GO カテゴリーの BP（Biological Processes）のオントロジーデータについて、ToxoDB（http://toxodb.org/toxo/）が提供する遺伝子注釈ツールを使用して、データベース上で解析を実施した[49]。GO カテゴリーのネットワーク表示は、Cytoscape[50]プラグインの Enrichment Map[51]を使用して、GO カテゴリー間で共有される遺伝子の数に基づいて相互に関連するネットワークを構築した。ネットワークでは、GO カテゴリーはノードとして示され、ノードをリンクするエッジは、遺伝子共有によって定義された GO カテゴリーのクロストークを表す[52]。

Ⅱ－７．計算環境と解析

本研究で行った多くの計算処理は、VMware player 上に構築した Linux 環境である CentOS で実行した。全てのスクリプトプログラムは、Perl 言語で記述し、Linux 上で実行した。また、HyPhy や codeml などの計算に長時間を要するプログラムは、連合創薬医療情報研究科の並列計算サーバー（24 CPU）または名古屋大学情報基盤センター全国共同利用システムの分散並列型 Linux 演算サーバ（HX600）を利用した。これらの計算機では、可能な場合には MPI 並列化による計算を行った。付録には、本研究で用いた主な Perl スクリプトを掲載した（Appendix 1）。

Ⅲ．*Bacteroides fragilis* についての研究成果

Ⅲ－１．*Bacteroides* 属コアゲノムの比較分析

本研究で用いた *Bacteroides* 属８菌種のゲノムについて、表１に示した。*Bacteroides* 属ゲノムのサイズ（塩基数）は、原核生物の特性を反映して比較的小さく、タンパク質をコードする遺伝子数も 3,436～4,917 であった。

表１　研究に用いた *Bacteroides* 属８菌種のゲノム

Bacteroides strain	GenBank accession No.	No. of CDS	Genome size (Mbp)	GC%
***B. fragilis* 638R**	NC_016776	4,417	5.373	43.4
***B. fragilis* NCTC9343**	NC_003228	4,403	5.241	43.1
***B. fragilis* YCH46**	NC_006347	4,730	5.310	43.2
***B. helcogenes* P36-108**	NC_014933	3,436	3.998	44.7
***B. salanitronis* DSM18170**	NC_015164	3,838	4.308	46.5
***B. thetaiotaomicron* VPI-5482**	NC_004663	4,917	6.293	42.9
***B. vulgatus* ATCC8482**	NC_009614	4,195	5.163	42.2
***B. xylanisolvens* XB1A**	FP929033	4,466	5.976	41.9

CDS: Coding sequence;　　GC%: Guanine plus cytosine content.

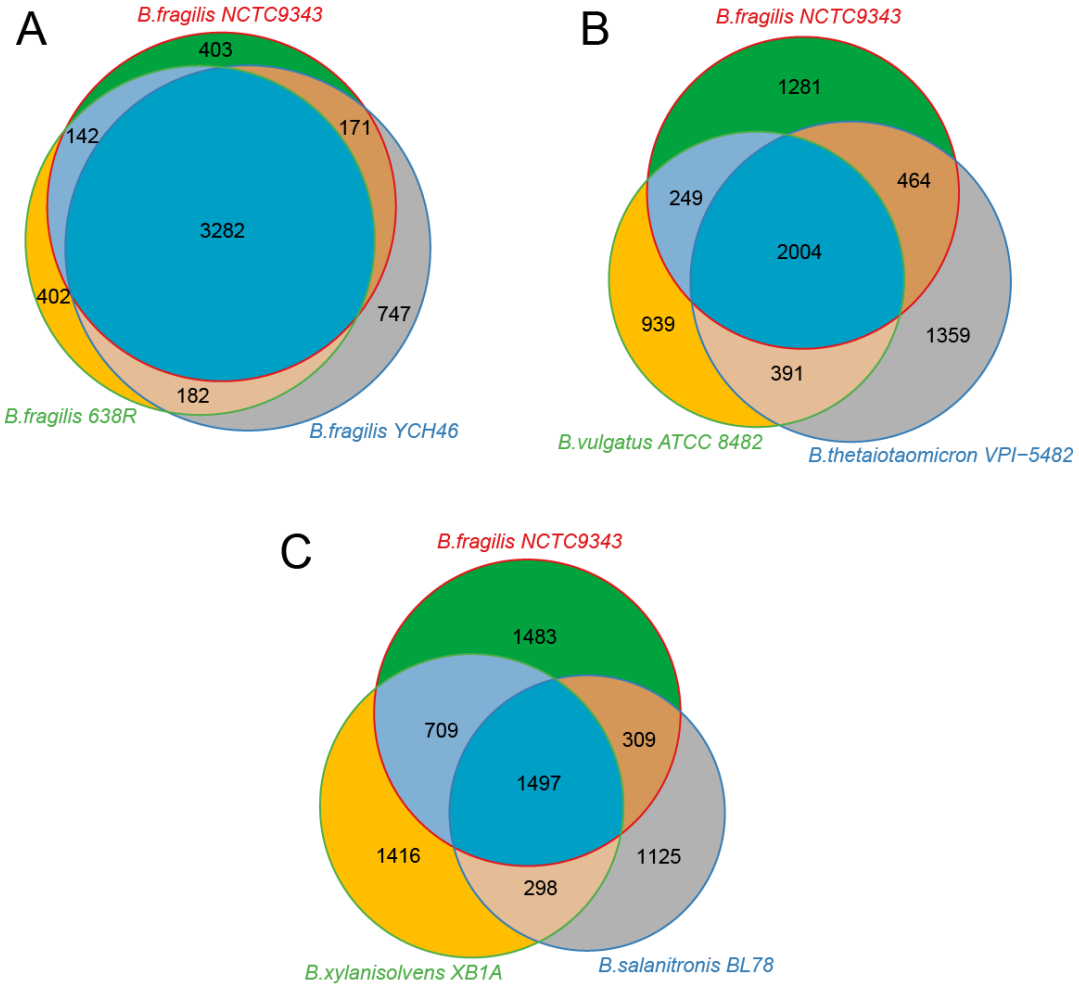

図３　*Bacteroides* 属間の遺伝子共有

OrthoMCL を使用して、これら 8 つの *Bacteroides* ゲノム全てに存在する 1,275 個のオルソログ遺伝子群を特定し、これらの遺伝子群をコアゲノムとした。次に、OrthoMCL 出力によって得られた遺伝子のコンテンツテーブルに基づいて、8 菌種の遺伝子組成の特徴を検討した。図 3 に示すように、*B. fragilis* の 3 つの同種の菌株間では、遺伝子の約 75%（3,282 個）を共有していたが、他菌種間の比較では共有遺伝子は相対的に少なかった。また、*B. fragilis* と密接に関連する病原性 *Bacteroides* 種である *B. vulgatus* および *B. thetaiotaomicron* [19]の 3 者間での比較では、2,004 個の共通遺伝子が見いだされた（図 3 B)。これに対し、*B. fragilis* と非病原性の *Bacteroides* 種である *B. xylanisolvens* 及び *B. salanitronis* 間の比較では、共通遺伝子は少なく 1,497 個であった（図 3 C)。こうした共有遺伝子数のパターンは、系統樹に描かれた進化距離を完全には反映していないことから（図 4）、進化過程での分岐後の時間以外の要因の関与が予想される。病原性を獲得した菌種間では、病原性に必要な遺伝子は共通の遺伝子が使われている可能性もある。

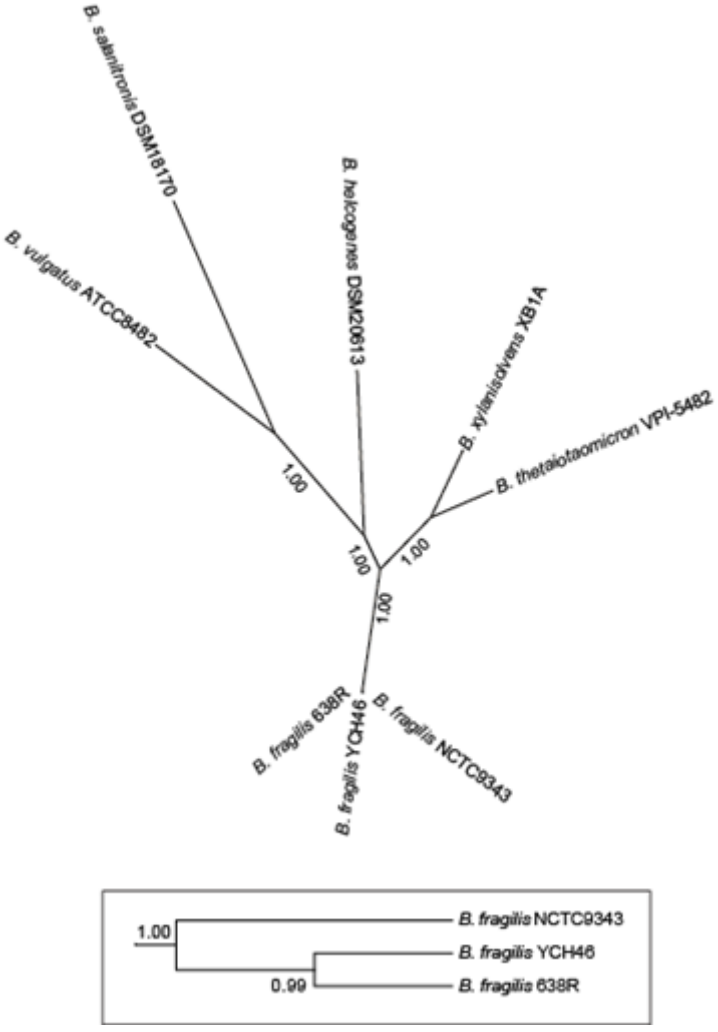

図 4　*Bacteroidesis* 属 8 菌種の系統樹

系統樹上に示した数字は分岐の正確性を表しており、1.00 はほぼ 100%の確率を示す。

図５は、８つの *Bacteroides* ゲノム間での遺伝子の分布を示している。ゲノム全遺伝子の46%は１つのゲノムのみに存在しており、これらは系統特異的遺伝子である。系統特異的遺伝子の次に多いのは、８つの *Bacteroides* ゲノム全てに共有されている遺伝子で、ゲノム全遺伝子の 13%を占める。これらが 1,275 個のコアゲノムである（図５）。コアゲノムは、*Bacteroides* 属で基本的で必須な機能に関わっていると推定される。他方、系統特異的遺伝子が多く見いだされることは、*Bacteroides* 属の菌種あるいは菌株間の機能分化が著しいことも示唆している。

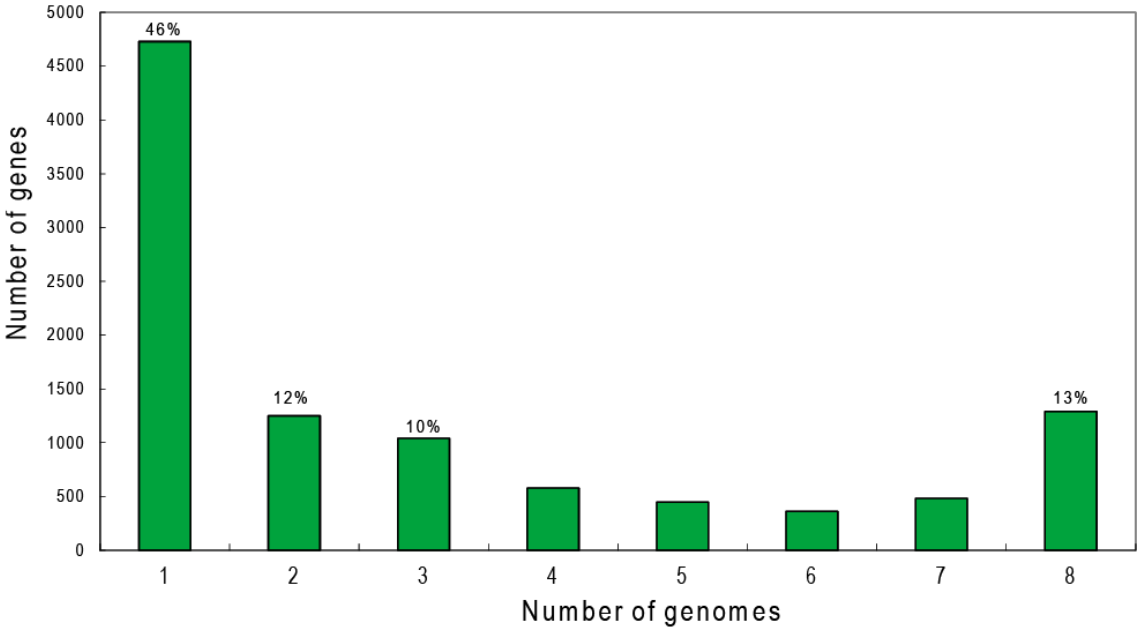

図５　*Bacteroides* 属間での遺伝子分布

Bacteroides ８菌種の遺伝子を比較し、共通する遺伝子をクラス分けした。一番左は１つの菌種のみに存在する遺伝子である。一番右に示した８菌種全てに存在した 1,275 遺伝子をコアゲノムとした。

コアゲノム遺伝子セット内の組換えまたは水平遺伝子伝達が系統樹推定および正の選択解析に影響する可能性を排除するために、HyPhy [38]で実装された SBP 解析および KH テストを使用して、*Bacteroides* 種間のコアゲノム遺伝子の組換えの有無について調べた。この方法では、単一の組換えブレークポイントを仮定する尤度モデルと、組換えを仮定しないモデルを比較する。HyPhy 分析では、1,275 個のコアゲノム遺伝子のうち 61 個のみで p 値<0.05 の有意な組換えブレークポイントが検出された。これらの組換え遺伝子は、通常の正の選択解析の遺伝子リストからは除外したが、組換えブレークポイントを境にして２つの遺伝子断片に分割し、別々に正の自然選択を評価した。

Ⅲ－２．*Bacteroides* の正の自然選択の検出

８つの *Bacteroides* 種間の進化的関係を知るために、組換え遺伝子を除く連結されたコアゲノム遺伝子に基づいて系統樹を推定した（図４）。この系統樹を使用して、PAML パッケージに実装された branch-site model で正の自然選択の解析を実行した[43]。全体として、正の選択は全ての *Bacteroides* 属の系統にわたって見られたが、系統間で著しいばらつきがあった（表２）。正の自然選択を受けた遺伝子が最も多く検出されたのは、それぞれ *B. salanitronis* と *B. vulgatus* へ至る進化系統であった。

表２　各 *Bacteroides* 種へ至る進化経路で検出された正の自然選択

Lineage	No. of genes with evidence for positive selection	Percentage in total core genome genes (%)
B. fragilis [a]	52[b]	3.8
***B. helcogenes* P36-108**	88	6.9
***B. salanitronis* DSM18170**	167	13.1
***B.thetaiotaomicron* VPI-5482**	10	0.8
***B. vulgatus* ATCC8482**	164	12.9
***B. xylanisolvens* XB1A**	7	0.5

a：*B. fragilis*３菌種へと至る進化経路

b：組み換え遺伝子を検出した遺伝子を含む

B. salanitronis への進化経路で正の選択が検出された 167 個の遺伝子は、分析に含まれるコアゲノム遺伝子の 13.1%に達してしていた。また、*B.xylanisolvens* と *B. thetaiotaomicron* 系統で、正の選択が検出された遺伝子数が最小であった。これらの結果は、系統ごとに正の選択下にある遺伝子の数は分岐からの距離にほぼ比例していることを示している。したがって、正の選択を受けた遺伝子の数が少ない *B. xylanisolvens* と *B. thetaiotaomicron* の系統は進化の期間が短いためであると考えられる。

次に、*B. fragilis* の病原性と自然選択の関連を明らかにするために、*B. fragilis* へと至る進化経路で正の選択が検出された遺伝子の詳細な解析を行った。branch-site model の尤度比検定（FDR <2%）に基づいて、合計 48 の遺伝子が正の選択下にあると特定された。さらに、組換え遺伝子をブレークポイントで分割して解析した結果から、4 つの組換え遺伝子の正の選択も検出され、最終的に 52 個の遺伝子に正の選択が検出された（表２）。

表３　*B. fragilis* への進化経路で正の選択を検出した 52 遺伝子

Cluster ID	COG category [a]	Gene annotation	$2\Delta lnL$ [b]	*q*-value
170	[S]	tetratricopeptide repeat protein	37.49512	1.16E-06
462	[I]	acyl-CoA dehydrogenase	33.69269	3.47E-06
697	[T]	two component system sensor histidine kinase	31.02918	9.65E-06
269	[M]	lipoprotein	29.99908	1.24E-05
471	[C]	ferredoxin oxidoreductase	24.33866	0.00015903
787	[C]	hypothetical protein	24.30007	0.00015903
1083	[P]	TonB dependent receptor	23.01494	0.00026584
169	[P]	TonB dependent receptor	22.20886	0.00035391
644	[S]	membrane protein	20.10175	0.00094479
1136	[P]	alkaline phosphatase	19.83070	0.00094577
268	[S]	hypothetical protein	19.71620	0.00094577
726	[T]	two-component system, sensor kinase	19.15439	0.00116340
745	[C]	Na^+translocating NADH-quinone reductase subunit F	17.71081	0.00229070
168	[L]	DNA topoisomerase I	17.14196	0.00286911
832	[E]	dipeptidase	16.41834	0.00381506
189	[J]	polyribonucleotide nucleotidyltransferase	16.28441	0.00381506
978	[G]	hypothetical protein	15.82599	0.00396907
535	[R]	amidohydrolase	15.75123	0.00396907
904	[P]	oxalate/formate antiporter	15.67711	0.00396907
1205	[F]	formyl transferase	15.66830	0.00396907
851	[I]	O-succinylbenzoate-CoA ligase	15.61942	0.00396907
973	[G]	glyceraldehyde 3-phosphate dehydrogenase	15.58608	0.00396907
340	[G]	phosphoglucomutase	15.48465	0.00401333
1059	[L]	DNA polymerase III alpha subunit	14.90654	0.00523281
465	[E]	aminopeptidase C	14.77003	0.00540919
770	[K]	RNA-binding protein	14.30354	0.00643397
1206	[Q]	acyl carrier protein	13.54749	0.00893195
596	[F]	amidophosphoribosyltransferase	13.52283	0.00893195
626	[S]	glucose-1-phosphate adenylyltransferase	13.49570	0.00893195
1117	[M]	outer membrane protein/Omp85	13.38540	0.00917682
163	[U]	protein-export transmembrane SecDF protein	13.23510	0.00964147
1114	[H]	riboflavin biosynthesis protein	13.17641	0.00965289
180	[E]	aminopeptidase	12.93830	0.01014399

1153	[O]	peptidyl-prolyl cis-trans isomerase	12.92560	0.01014399
361	[M]	hypothetical protein	12.64840	0.01145469
1245	[E]	dipeptidyl peptidase IV	12.57262	0.01162270
788	[P]	sulfate adenylyltransferase subunit 1	12.44627	0.01166327
702	[M]	penicillin-binding protein 1A	12.42766	0.01166327
399	[I]	cardiolipin synthetase	12.30724	0.01215106
868	[O]	META domain protein	12.12239	0.01311172
611	[E]	aminopeptidase	11.98312	0.01334565
311	[V]	ABC-2 type transporter	11.97787	0.01334565
816	[M]	transmembrane glycosyltransferase	11.96647	0.01334565
423	[M]	hypothetical protein	11.80581	0.01424485
1013	[C]	NADH-quinone oxidoreductase chain C/D	11.71577	0.01464575
506	[U]	signal recognition particle protein	11.61612	0.01514261
915	[U]	tetratricopeptide repeat protein	11.35812	0.01705632
946	[I]	YegS//BmrU family lipid kinase	11.09953	0.01922939
186 [c]	[R]	hypothetical protein	25.71775	0.00011435
176 [c]	[F]	phosphoribosyl aminoimidazole carboxylase	14.93504	0.00523281
1095 [c]	[V]	ABC transporter	13.62999	0.00859239
1192 [c]	[J]	translation initiation factor IF-2	13.64192	0.00859239

a：図 6 で示した COG カテゴリー分類

b：尤度比検定の統計値

c：組み換え遺伝子において正の選択を検出した遺伝子

表 3 に示すように、*B. fragilis* へと至る進化経路で正の選択が検出された 52 個の遺伝子は、多様な機能に関わっており、20 の COG 機能分類のいずれかに該当している。図 6 には、各 COG 中で正の選択を受けた遺伝子と全遺伝子が占める比率を示した。*B. fragilis* 系統で正の選択を受けた遺伝子は、「Lipid transport and metabolism」、「Intracellular trafficking, secretion and vesicular transport」、「Defense mechanisms」の 3 つの COG カテゴリーに多く含まれており、全遺伝子の比率と比較して統計的にも有意であった（片側二項検定 $p<0.05$）。一方、COG カテゴリー「Cell wall/ membrane / envelope biogenesis」では正の選択を受けた遺伝子が有意に多いとは言えないが、遺伝子注釈から他の COG カテゴリーに属する遺伝子でも表面/膜に局在するタンパク質をコードすることが示唆される遺伝子は多い（表 3）。表 3 の遺伝子注釈から、52 の正の選択が検出された遺伝子の中で 17 個（33%）の遺伝子が表面/膜構造に関連することが推定された。これらの知見は、正の選択を受けた多くのタンパク質が細胞表面に露出し又は表面膜に局在することを示した以前のゲノムワイド研究と一致している[13, 53, 54]。本研究で見いだした正の選択を受けた細

胞表面/膜に関連した遺伝子は、*B. fragilis* がヒトの腸管内で他の細菌と競合するために獲得した戦略や、宿主の免疫システムへの適応などに寄与している可能性が大きい[55, 56]。

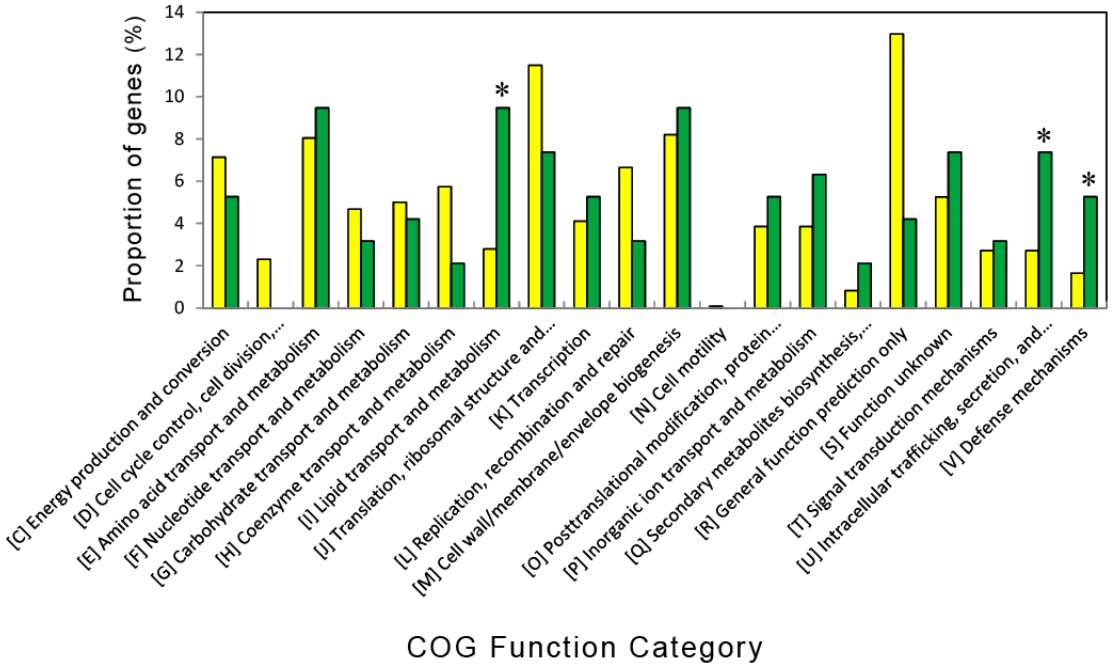

図６　ＣＯＧ機能分布

黄色は遺伝子全体のカテゴリー分布で、緑色は正の選択を検出した 52 遺伝子のカテゴリー分布を表す。

＊：遺伝子全体の分布と比較して、52 遺伝子で正の選択を有意に検出したカテゴリー分類

Ⅲ－３．*Bacteroides* タンパク質の３次元立体構造解析

正の選択が果たす役割についてより多くの洞察を得るために、タンパク質の 3D モデルで正の選択を受けたアミノ酸部位をマッピングした。細胞表面/膜に関わる遺伝子の多くが正の選択を示したので、*B. fragilis* の外膜に局在する 2 つの代表的なタンパク質に焦点を当てた。図 7 には、Phyre2 サーバーを使用したホモロジーモデリングに基づいた、TonB dependent receptor （表 3 の Cluster ID 1083）および Outer membrane protein/Omp85（表 3 の Cluster ID 1117）の 3D 構造を示した。いずれのタンパク質もβバレル構造をもつことが特徴であり、その小さい q 値から強い正の選択を受けていることが解る（表 3）。Bayes Empirical Bayes アプローチを使用した解析から、TonB dependent receptor の 9 つのサイト 212L、267P、391M、424R、 473S、537Q、561R、575P、610L は、branch-site model の下で高い事後確率（PP> 0.95）で $\omega > 1$ の正の選択を受けたことが推定された。これら 9 つのアミノ酸サイトを TonB dependent receptor の 3D 構造上にマッピングすると、7 つのアミノ酸が細胞外ループに位置していた（図 7 A)。膜内のβ鎖境界に面する部分には、2 つの正の選択を受けたアミノ酸残基のみが位置していた。図 8 には、TonB dependent

receptor の膜内外ドメインと正の選択を受けたアミノ酸を模式的に示している。Omp85 では、単一のアミノ酸部位 829A のみで正の選択が検出された。Omp85 の正の選択部位も細胞外ループに位置していた（図 7 B)。

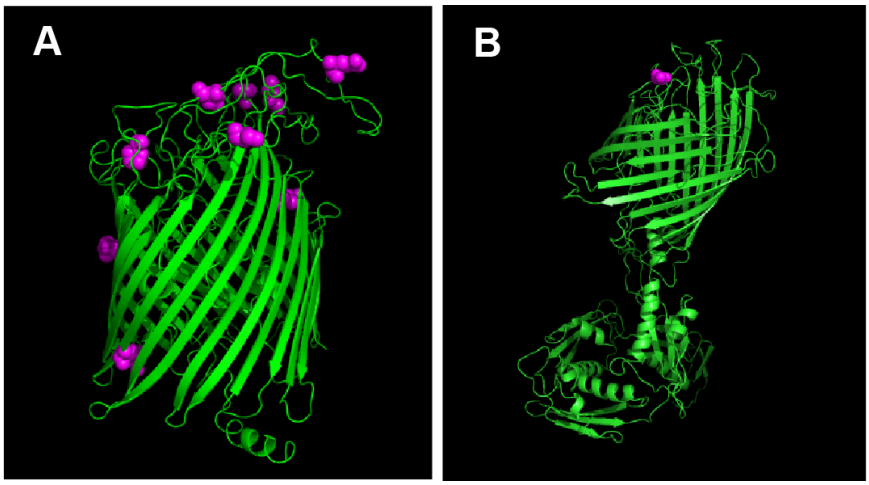

図 7　TonB dependent receptor（A）と Outer membrane protein/Omp85（B）の 3D 立体構造
正の選択を検出したアミノ酸を紫色で 3D 立体構造上に表示した。

現在までに、細菌の TonB dependent receptor の重要な役割として、Fe やビタミン B_{12} など様々な栄養素の生理的な取り込みに関わることが明らかにされている[57]。病原性細菌の宿主中での生存は、宿主と競合する鉄などの栄養素を獲得する能力に依存するため[58]、TonB dependent receptor は病原性にとっても重要である可能性がある[59]。別の研究では、*B. fragilis* の TonB dependent receptor が血漿フィブロネクチンと結合することが示されており、宿主組織との接着分子として機能している可能性が示唆されている[60]。本研究で示した TonB dependent receptor の正の選択は、細菌が外界と接する細胞外ループで主に起きていた。したがって、正の選択を受けたアミノ酸は、細菌が侵入、感染する際の効率的な栄養素の識別や宿主組織への接着に重要な役割を果たすことも充分に予想される。さらに、表面抗原としての Omp85 は、宿主免疫系と相互作用し、大腸菌で正の選択が検出されることが報告されている。[61]。

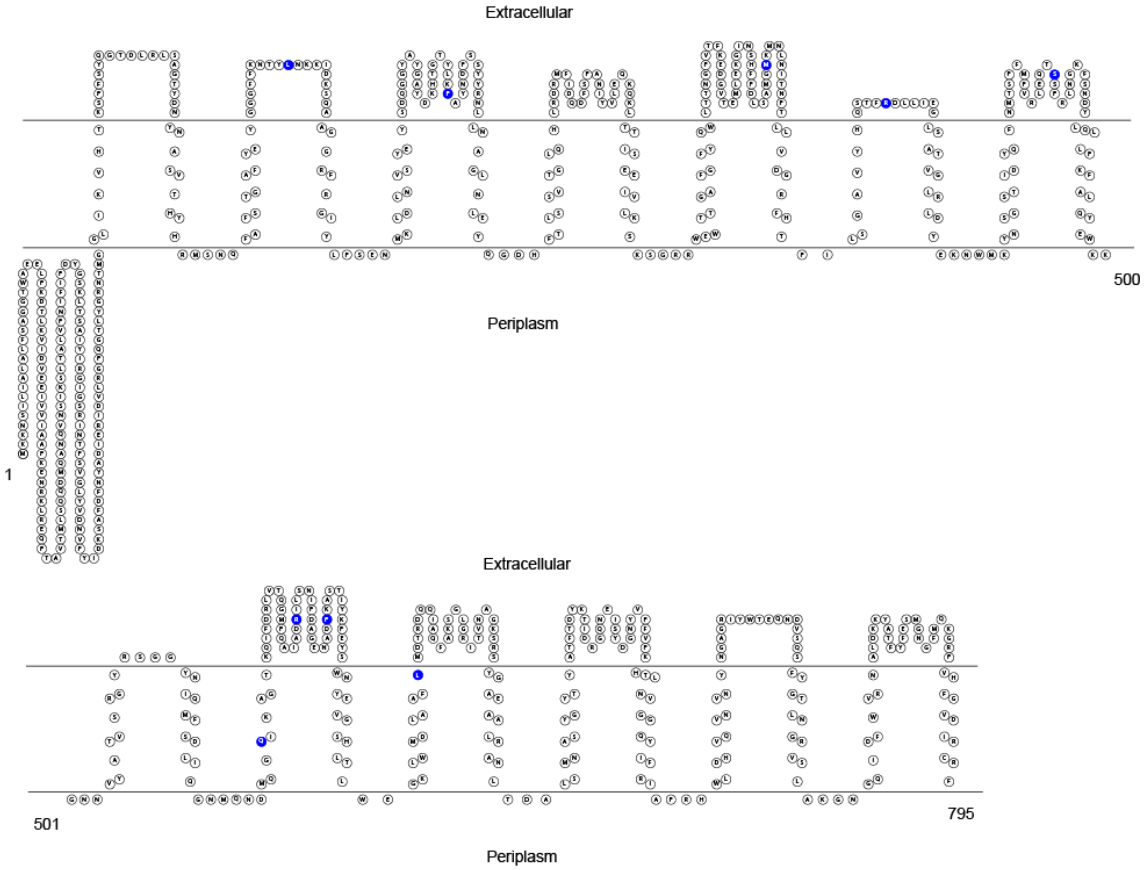

図８　TonB dependent receptor の膜内外ドメイン図

TonB dependent receptor（Cluster ID 1083）のアミノ酸を並べて、膜貫通の有無を模式化した。青い球は正の選択を検出したアミノ酸を示す。平行線の上方が細胞外で、平行線の下方がペリプラズム、平行線の中部分が外膜内を示す。

Ⅳ．*Toxoplasma gondii* についての研究成果

Ⅳ－１．*T. gondii* と近縁種コアゲノムの比較解析

本研究で用いた９系統の *T. gondii* と２近縁種のゲノムについて、表４に示した。*T. gondii* 及び近縁種のゲノムサイズ（塩基数）は、原核生物の *Bacteroides* ゲノムに比べて 10 倍程度大きく、タンパク質をコードする遺伝子数も 7,122～10,122 であった。

以前の研究で[31, 62, 63]、*Neospora caninum* と *Hammondia hammondi* は、*T. gondii* に密接に関連した近縁種であり、両方ともにヒトに対して非病原性であることが明らかにされている。そこで、３つの *Toxoplasma* 系統である *T. gondii* ME49、*T. gondii* GT1、*T. gondii* VEG に *Neospora caninum* と *Hammondia hammondi* を加えて、合計５種類の原虫ゲノムを用いて病原性 *T. gondii* 系統への進化経路で正の選択を受けた遺伝子の検出を行った。最初に OrthoMCL を使用して、5 つのゲノム全てに存在する 5,788 個のオルソログ遺伝子を同定した。さらに、これらコアゲノム遺伝子セット内での組換えが正の選択分析

に影響を与える可能性を排除するために、GARD algorithm を用いて組み換えの検出を行った。5,788 個のコアゲノム遺伝子のうち 1 つだけで p 値<0.05 の有意な組換えブレークポイントが検出された。この組換え遺伝子は、コード配列全体を使用した正の選択解析には含まれていない。また、組換え遺伝子を除く連結されたコアゲノム遺伝子に基づいて、5 つの原虫の系統樹を推定した（図９）。

表４　研究に用いた *Toxoplasma* と近縁種の 11 ゲノム

Species	Protein-coding genes	Genome size (Mbp)
***Neospora caninum* Liverpool**	7,122	59.10
***Hammondia hammondi* H.H.34**	8,003	67.70
***Toxoplasma gondii* GT1**	8,460	63.95
***Toxoplasma gondii* ME49**	8,322	65.67
***Toxoplasma gondii* VEG**	8,410	64.52
***Toxoplasma gondii* ARI**	9,958	64.69
***Toxoplasma gondii* FOU**	10,117	64.53
***Toxoplasma gondii* p89**	9,701	64.16
***Toxoplasma gondii* RUB**	10,027	64.96
***Toxoplasma gondii* TgCatPRC2**	10,121	64.19
***Toxoplasma gondii* VAND**	9,255	64.27

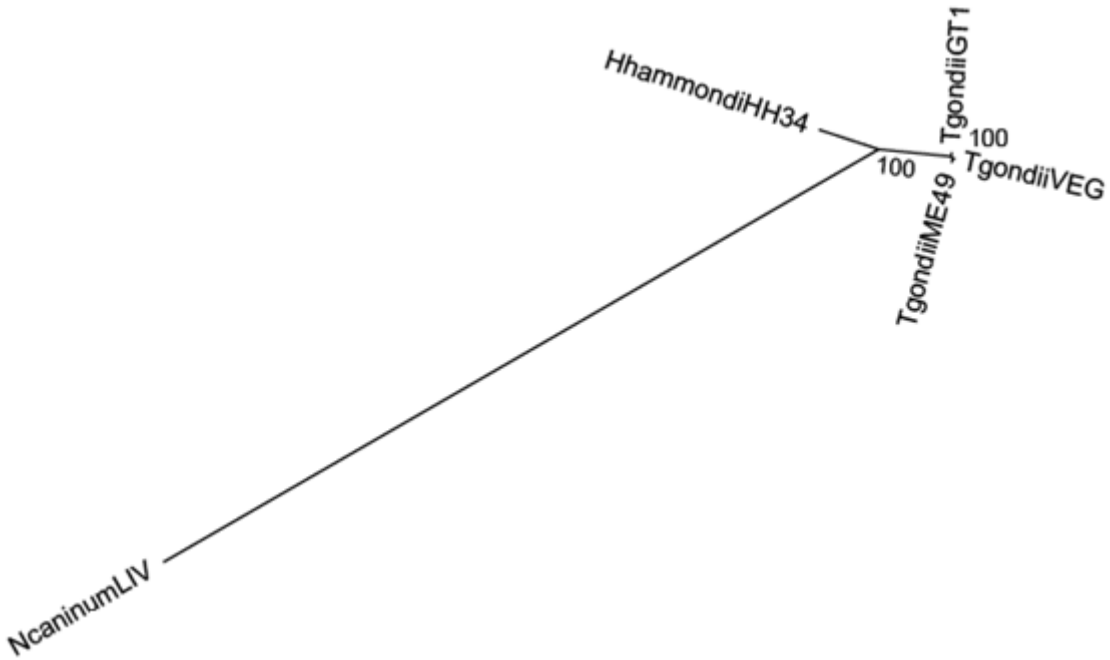

図９　５原虫の系統図（放射状）

系統樹上に示した 100 という数字は、分岐の正確性を表している。

この系統樹を使用して、PAML パッケージに実装された site model と branch-site model に従って、正の選択の解析を行った。Site model では、q 値の 0.05 を基準として判定した結果、正の選択を受けた遺伝子として 261 の遺伝子が同定された。261 の遺伝子の全ては、Appendix 2 に示した。これら正の選択が検出された遺伝子は、5,788 のコアゲノム遺伝子データセットの 4.5%に相当した。対照的に、branch-site model では正の選択が検出された遺伝子は site model ほど多くなかったが、系統間で大きく変動していた（図 10）。正の選択下にある遺伝子の数は、*H. hammondi*（図 10、Hh）および *T. gondii*（図 10、N - T）に至る進化経路に沿って最大であった。両方の系統で正の選択下にある 32 個の遺伝子（Hh）と 31 個の遺伝子（N - T）は、コアゲノム遺伝子の 1.1%に相当する（Appendix 3 に遺伝子リストを示した）。少数ではあるが、*T. gondii* 系統内の進化経路でも正の選択を受けた遺伝子が見いだされた。これらの結果は、系統ごとの正の選択下にある遺伝子の数は進化経路の長さにほぼ関連しており（図 9）、*T. gondii* 系統内で正の選択遺伝子の検出数が少ないことは分岐後の進化時間が短いためであることを示している。

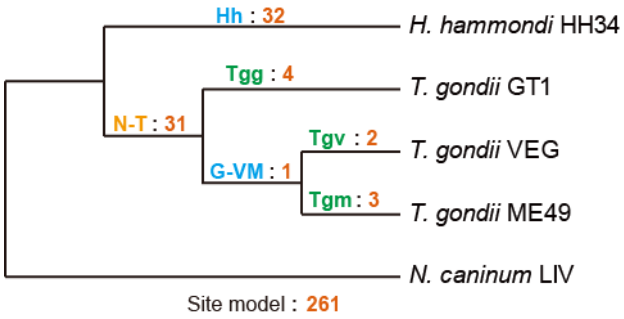

図 10　5 原虫の系統樹と各進化経路での正の選択検出数

各進化経路の数字は、正の選択を検出した遺伝子数である。この系統樹は、進化時間を反映していない。

Ⅳ－２．正の選択下での遺伝子の機能的分析

正の選択を受けた遺伝子と *Toxoplasma* 病原性との関わりを明らかにするために、正の選択を示す遺伝子リストの遺伝子オントロジーとエンリッチメント解析を行った。Site model では、正の選択を受けた遺伝子は、36 の生物学的プロセス、43 の分子機能、2 つの細胞成分を含む合計 81 の GO カテゴリーに有意に関連していることが解った（Appendix 4）。生物学的プロセスの上位 5 つの GO カテゴリーは、regulation of metabolic process（GO : 0019222）、macromolecule modification（GO : 0043412）、regulation of catalytic activity（GO : 0050790）、regulation of molecular function（GO : 0065009）および regulation of macromolecule metabolic process（GO : 0060255）であった。分子機能の最も重要な GO カテゴリーは、transferase activity、transferring phosphorus-containing groups、catalytic activity、phosphotransferase activity、alcohol group as acceptor、DNA-directed DNA polymerase activity および nucleic acid binding transcription factor activity であった

（Appendix 4）。Branch-site model の *T. gondii* へ至る N - T 進化経路（31 遺伝子）の場合では、8 つの生物学的プロセスと 17 の分子機能を含む合計 25 の GO カテゴリーで、正に選択された遺伝子が有意に関連していることが解った（Appendix 5）。

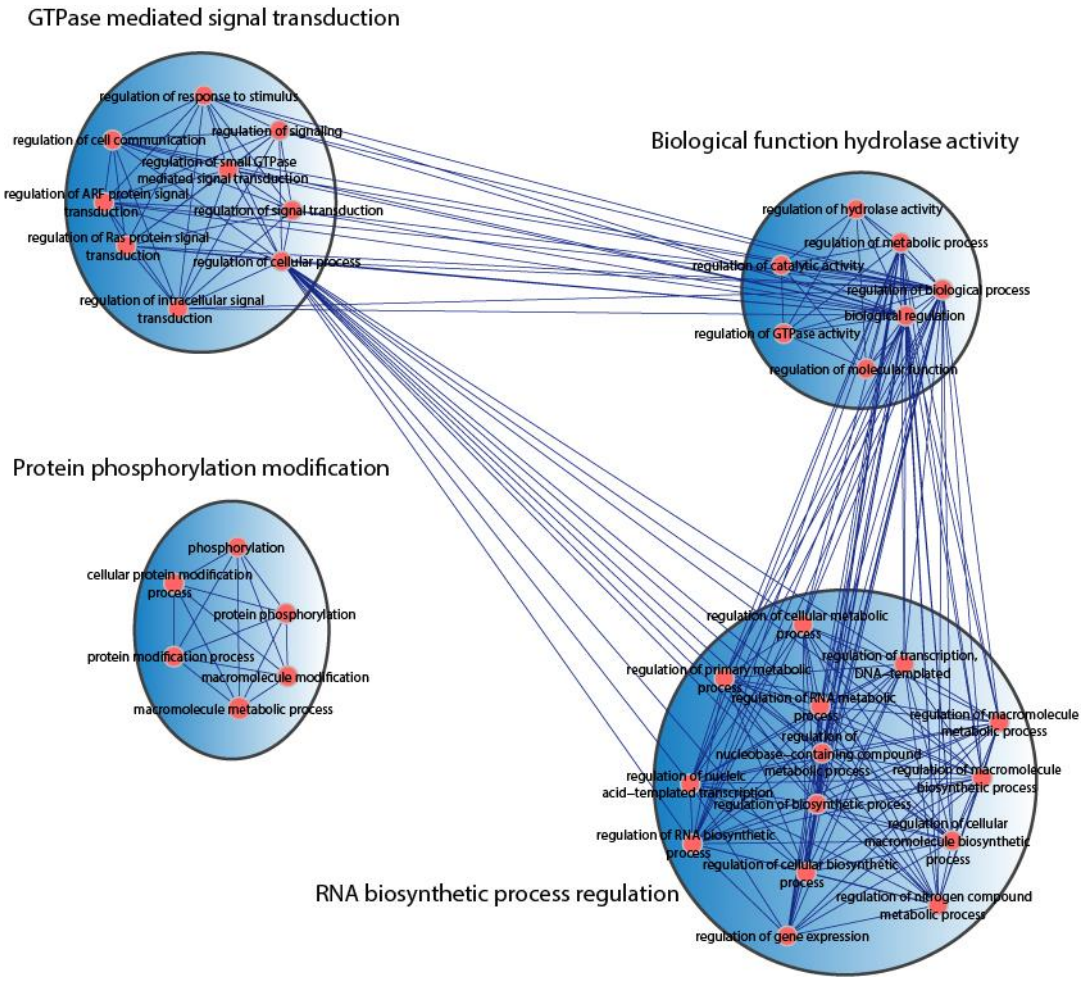

図 11　Site model での GO カテゴリーネットワーク図

Site model で正の選択を検出した遺伝子の GO カテゴリーがどのように関連しているかについて、ネットワーク図で表示した。

正に選択された遺伝子と生物学的機能との関連をより簡潔に示すために、Enrichment Map を用いて GO カテゴリー間の関わりを表すネットワーク図を構築した[51]。Site model で検出された遺伝子について、図 11 にカテゴリー間の関係を示している。4 つの大きなクラスターが形成され、各 GO カテゴリー間で遺伝子にかなりの重複が観察された。最大のクラスターには 14 の GO カテゴリーが含まれ、それらの多くは RNA 生合成調節に関連していた。2 番目のクラスターには 9 つの GO カテゴリーが含まれ、GTPase を介したシグナル伝達に最も多く関与していた。3 番目と 4 番目のクラスターには、それぞれ加水分解酵素

活性とタンパク質リン酸化修飾に関わる GO カテゴリーが含まれた。一方、branch-site model の *T. gondii* へ至る進化経路では（N-T）、タンパク質のリン酸化修飾と制御分子活性を表す 2 つの小さなクラスターが生成された（図 12）。これらのクラスターの中では、タンパク質リン酸化や低分子量 GTPase などによって媒介されるシグナル伝達に関連する GO カテゴリーが顕著であることが解る。興味深いことに、こうしたシグナル伝達機能を有するタンパク質の 1 つであるタンパク質キナーゼが *T. gondii* が宿主に侵入する際に重要な役割を果たすことが報告されている[64]。したがって、正の選択によって急速に進化してきた遺伝子群は、*T. gondii* のシグナル伝達に関わっているとともに、それらのいくつかが宿主-原虫相互作用システムの重要な要因である可能性が示唆される。

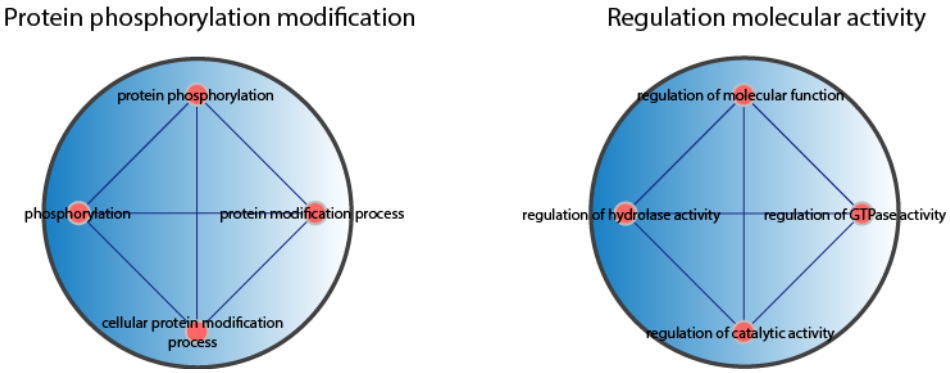

図 12　Branch-site model での GO カテゴリーネットワーク図

Branch-site model で正の選択を検出した遺伝子の GO カテゴリーがどのように関連しているのかについて、ネットワーク図で表示した。

Ⅳ－３．*T. gondii* コアゲノムでの正の自然選択の検出

T. gondii 系統内で正の自然選択をさらに詳細に検出するために、9 つの *T. gondii* 原虫系統のゲノムを選択し（表４）、9 ゲノム全てに存在する 5364 個のオルソログ遺伝子を特定した。前述の GARD algorithm と KH テストを使用して、9 つの *T. gondii* ゲノム間で遺伝子内の組換えについて解析した。組換えがあったと推定される遺伝子は、5364 コアゲノム遺伝子中の 86 個で、p 値<0.05 で有意な組換えブレークポイントを検出した。これらの組換え遺伝子を除外した後、branch-site model を用いて 9 つの *T. gondii* 系統に至る各進化経路に沿った正の選択の解析を実施した。使用した系統樹は、コアゲノムから組換え遺伝子を除いた塩基配列から推定し、注目している進化経路に＊を付した（図 13）。

今回の解析では、q 値<0.05 の条件下で、各 *T. gondii* 系統に至る進化経路で正の選択を受けたと推定される 2～20 の遺伝子が特定された（表 5）。正の選択を検出した遺伝子によってコードされるタンパク質産物には、secretory pathogenesis determinants（SPD）[30]とよばれる、以前から *T. gondii* の病原性に関ることが知られている多くの分泌タンパク質または表面タンパク質が含まれた[26]。*T. gondii* の 9 系統に至る全ての進化経路におい

て、少なくとも 1 つの SPD 遺伝子が正の選択下にあることが判明した（表 5）。その他の正の選択を受けた遺伝子としては、宿主侵入[65]や原虫の伝播[66]で何らかの役割を果たすことが知られている plant-like AP2 transcription factor や oocyst wall protein（表 5）も高頻度に見いだされた。これらの知見は、正の選択がさまざまな微生物の病原性因子の進化に重要な役割を果たし、宿主-病原体動態の重要なメカニズムであることを示した以前のゲノムワイド研究と一致している[67-69]。分泌タンパク質または表面タンパク質をコードする遺伝子の正の選択と急速な進化は、*T. gondii* が宿主であるヒトの腸管を含むさまざまな急速に変化する環境に適応し侵入するために重要であると考えられる[26]。

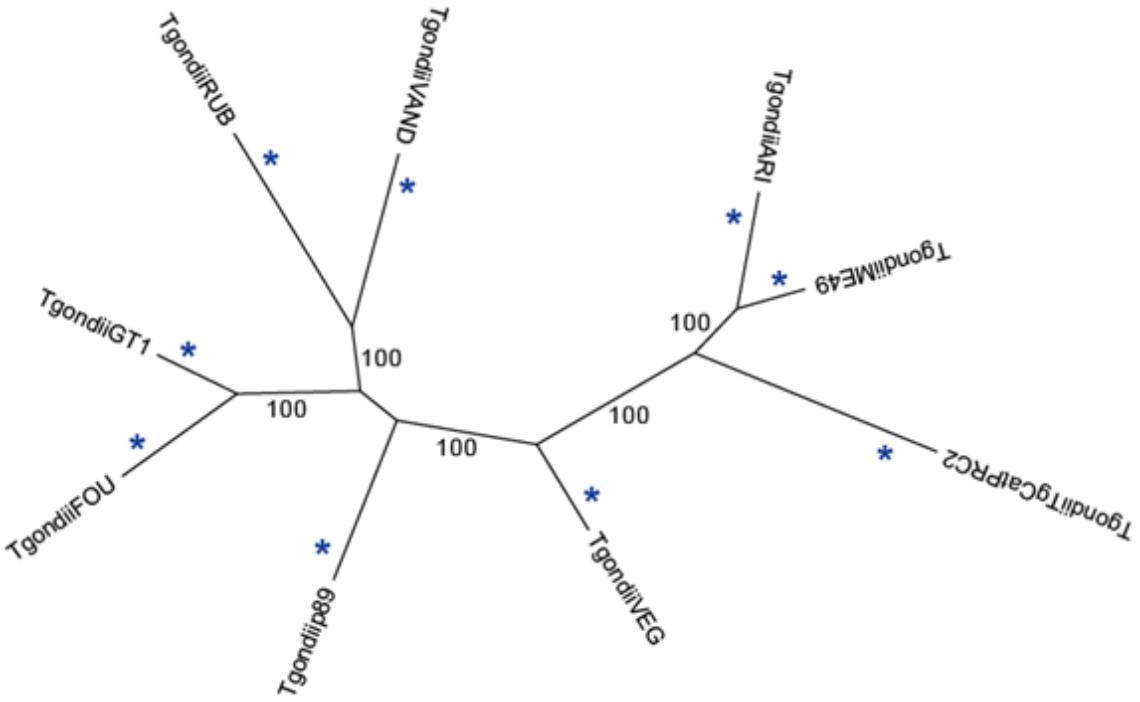

図 13　9 種類の *T. gondii* 原虫についての系統樹

＊：正の選択検出に用いた 9 つの進化経路

系統樹上に示した 100 という数字は、分岐の正確性を表している。

表 5　*T. gondii* の 9 系統で正の選択を受けた遺伝子

Lineage	No. of genes with evidence for positive selection		No. of SPD** genes	Positively Selected SPD and Pathogenicity related Genes						
	q -value < 0.2*	*p* -value < 0.01	*p* < 0.01							
Toxoplasma gondii GT1	8	24	3	rhoptry protein ROP10	SAG-related sequence SRS54	SAG-related sequence SRS16C				
Toxoplasma gondii ME49	14	32	4	MIC2-associated protein M2AP	SAG-related sequence SRS38A	SAG-related sequence SRS59J	SAG-related sequence SRS53D	AP2 domain transcription factor AP2VIII-5		
Toxoplasma gondii VEG	11	32	2	Toxoplasma gondii family A protein	SAG-related sequence SRS59J	AP2 domain-containing protein	AP2 domain transcription factor AP2X-2			
Toxoplasma gondii ARI	3	42	5	Toxoplasma gondii family A protein	microneme protein, putative	SAG-related sequence SRS35B	rhoptry kinase family protein ROP37	SAG-related sequence SRS53D	oocyst wall protein	AP2 domain transcription factor AP2X-9
Toxoplasma gondii FOU	2	24	3	SAG-related sequence SRS30C	SAG-related sequence SRS53B	SAG-related sequence SRS54	oocyst wall protein	AP2 domain-containing protein	AP2 domain transcription factor AP2IX-6	
Toxoplasma gondii p89	20	54	2	toxofilin	rhoptry kinase family protein ROP39	AP2 domain transcription factor AP2XI-2				
Toxoplasma gondii RUB	14	53	1	SAG-related sequence SRS30C	AP2 domain transcription factor AP2XII-2					
Toxoplasma gondii TgCatPRC2	5	43	1	SAG-related sequence SRS57	oocyst wall protein	AP2 domain transcription factor AP2VIIa-3				
Toxoplasma gondii VAND	11	49	2	rhoptry protein ROP18	SAG-related sequence SRS16C	oocyst wall protein				

＊：Benjamini＆Hochberg の補正値

＊＊SPD：分泌型病原性の決定因子（表中の水色）

oocyst wall protein を橙色、plant-like AP2 transcription factor を緑色で示した。

Ⅳ－４．*Toxoplasma* タンパク質の３次元立体構造解析

正の選択がタンパク質の立体構造に及ぼす影響を明らかにするため、9 つの *T. gondii* 系統で検出された正の選択を受けた遺伝子のなかの Toxofilin に注目した。Toxofilin は *T. gondii p89* 系統で正の選択が検出されており、SPD のグループに属する。また、Lee らの構造研究[70]により、３次元立体構造が決定されており、5 つの連続したαヘリックスが比較的独立したアクチン結合部位を形成することが実証されている。本研究での Bayes Empirical Bayes アプローチを使用した解析によれば､Toxofilin の 4 つのアミノ酸サイト、43A、115 V、166I、179F は、branch-site model の下で高い事後確率（PP> 0.90）で $\omega > 1$ の正の選択を受けたことが推定された。（表 6 、図 14A)。Toxofilin の結晶構造はアクチンと共結晶化されたアミノ酸領域 69-196 についてのみ解明されているため[70]、Toxofilin の 3D 構造上のこのアミノ酸領域に含まれる 2 つの正の選択を受けたアミノ酸サイトをマッピ

ングした（図 14）。図 14B から明らかなように、2 つの正に選択されたアミノ酸サイト、115V および 166I は、3 番目および 5 番目のαヘリックスに位置していた。次に、正の選択が検出された部位の機能的意義についての洞察を得るために、アミノ酸置換の物理化学的特性の変化の影響を推定する TreeSAAP プログラムを使用した[47]。注目すべきことに、PAML によって特定された Toxofilin の全ての正の選択を受けたアミノ酸サイトは、TreeSAAP によって物理化学的特性の大きな変化の下にあることが検出された（表 6）。すなわち、*T. gondii p89* 系統に至る進化過程で起きるアミノ酸置換に対して、これらのアミノ酸サイトはスコア 6〜8 で大きな物理化学的特性変化を示すことが明らかになった。

表 6　Toxofilin で正の選択を受けたアミノ酸サイトと物理化学的特性変化

ω values (Proportion)[a]	**Positively selected sites[b] (Posterior probability)**	**TreeSAAP properties**	
		Amino acid changes	Radical changes in physicochemical properties[c]
$\omega_0 = 0$, $\omega_1 = 1.0$, $\omega_2 = 999.0$	43A (0.998)	S → A	*Pα, Pc, Pt*
($p_0 = 0.68488$, $p_1 = 0.29001$, $p_2 = 0.02511$)	115V (0.914)	I → V	*pK′*
	166I (0.974)	L → I	*pK′*
	179F (0.976)	Y → F	*F*

a：p_0〜p_2 は ω_0〜ω_2 の比率を示す

b：正の選択を検出したアミノ酸サイトとその事後確率

c：TreeSAAP によって検出されたアミノ酸特性の変化（6〜8 カテゴリー）

Pα, alpha-helical tendencies; *Pc*, coil tendency; *Pt*, turn tendencies; *pK′*, equilibrium constant (ionisation COOH); *F*, mean r.m.s. fluctuational displacement.

Toxofilin は *T. gondii* の rhoptry organelle に含まれており、宿主細胞への侵入時に分泌される[71]。Toxofilin は宿主細胞のアクチンと結合し、Actin フィラメントのターンオーバーを促進・加速することが知られている[72]。また、哺乳動物細胞での Toxofilin の過剰発現は、マイクロフィラメントと Actin ストレスファイバーの数を減らすことが報告されている[72]。これらの知見は、Toxofilin が宿主の Actin 細胞骨格を破壊し、感染中の原虫の侵入を促進できることを示唆している[73]。Lee らの研究によれば[70]、Toxofilin のαヘリックス 3〜5 は、Actin 結合ドメインを構成している。本研究で行われた進化解析により、Toxofilin のヘリックス 3 およびヘリックス 5 に正の選択を受けた 2 つのアミノ酸サイトが存在することが示された（図 14A）。さらに、ヘリックス 3 の正に選択されたアミノ酸 115V は、Actin の Toxofilin 結合溝と緊密に接触していた（図 14B）。これらの結果は、正の選択

を受けたアミノ酸が Toxofilin と宿主アクチンとの効率的な相互作用に関係している可能性を強く示唆している。したがって、正の選択は Toxofilin-Actin-結合を微調整することによって *Toxoplsasma* の侵入プロセスを高度に進化させたことも充分に考えられる。

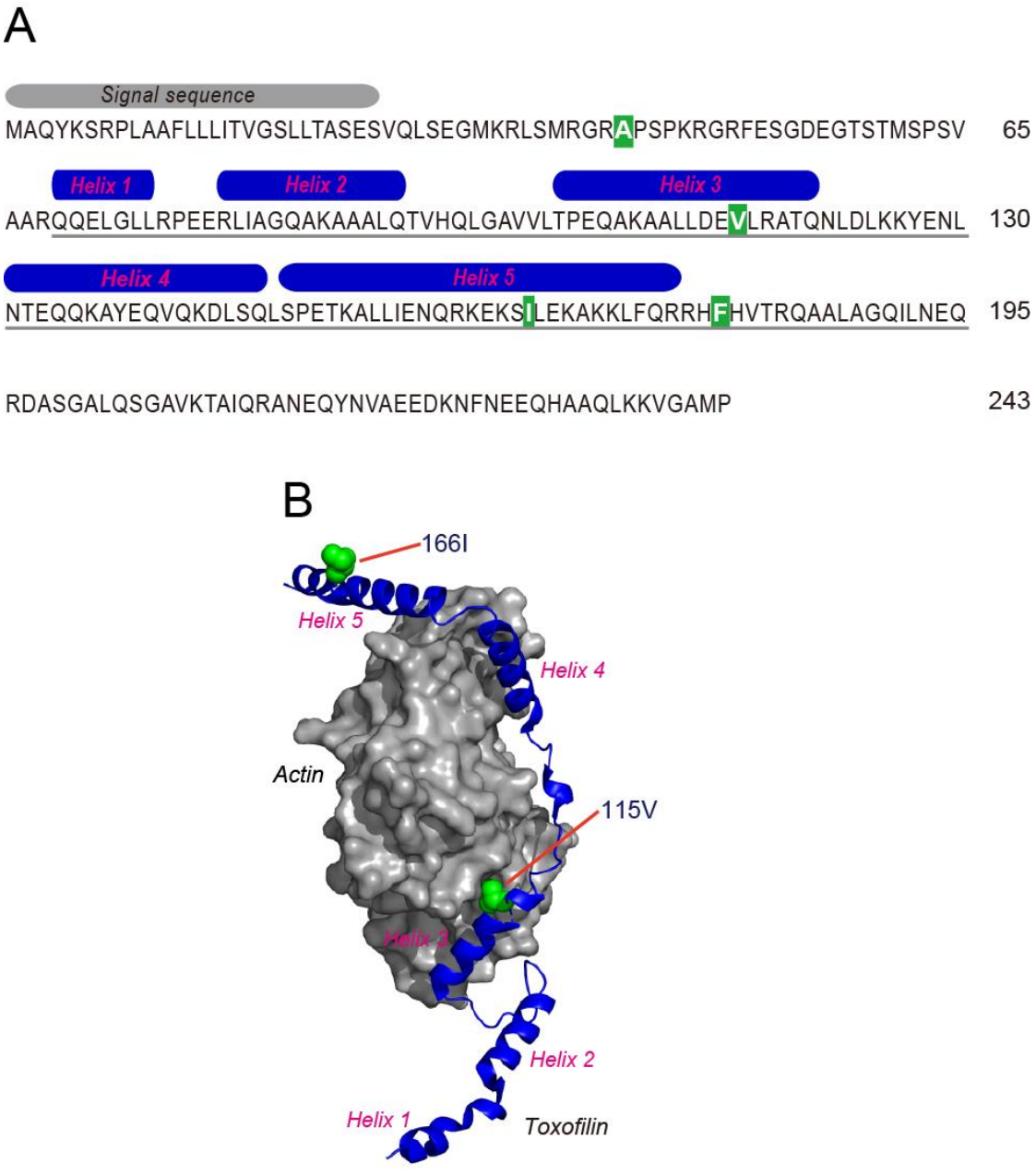

図 14　Toxofilin と Actin の 3D 立体構造

A：Toxofilin のアミノ酸配列上で、ヘリックス構造と正の選択を受けたアミノ酸の位置を表示。

B：Toxofilin と Actin の結合状態における 3D 立体構造。Toxofilin はリボンモデルで、Actin は分子表面モデルで表示した。正の選択を検出したアミノ酸（115V、166 I）を緑色で示した。

Ⅴ．総括

原核生物のモデルとして、*Bacteroides* のゲノムワイド解析により、3 つの *B. fragilis* 病原性株を含む *Bacteroides* 属のコアゲノム遺伝子を特定した。コドンモデル最尤法に基づいて、系統特異的な正の選択をコアゲノムから同定し、*Bacteroides* の広範囲の遺伝子が正の自然選択を受けていることを示した。とりわけ、表面/膜に関連するタンパク質をコードする遺伝子で正の選択が顕著であり、例えば TonB dependent receptor および Outer membrane protein/Omp85 などが正の選択の主要な標的である。これらの正の選択を受けた遺伝子の適応変化は、宿主の免疫および防御システムによって引き起こされる動的な相互作用に関連している可能性がある。

真核生物のモデルとして、*Neospora caninum* と *Hammondia hammondi* の 2 つの非病原性種を含む、病原性 *Toxoplasma* のコアゲノム遺伝子を定義した。コドンモデル最尤法にに基づいて、これらの種の広範囲の遺伝子でアミノ酸サイトおよび系統特異的な正の選択を特定し、正の選択がこれら微生物の進化に寄与したことを示した。特に、secretory pathogenesis determinants（SPD）をコードする遺伝子などの病原性に関わる遺伝子の正の選択は、*T. gondii* の系統内で多く見いだされた。SPD の代表的遺伝子の 1 つである Toxofilin では、立体構造レベルでの適応進化が推定された。これらの正の選択を受けた遺伝子の適応変化は、*Toxoplasma* の感染および病原性発現プロセスと密接に関わっており、宿主-病原体相互作用の進化に大きく影響してきたことが示唆された。

原核生物と真核生物について、それぞれの正の選択を受けた遺伝子は様々であったが、両者において病原性に関わる遺伝子の多くが正の選択を受けていた。正の選択を顕著に受けている遺伝子の中には、宿主への侵入、感染に際して機能するタンパク質をコードしているものも多い。こうした研究成果は、微生物の病原性解明へのアプローチの一つとして、比較ゲノム解析による正の自然選択の検出が有効であることを示している。今後、こうした分子進化解析の結果に基づいて実験検証を含めた研究をさらに進めることにより、微生物の病原性についてより詳細な分子機構の解明が期待できる。

Ⅵ. 文献

1. Aguileta, G., et al., *Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists.* Infect Genet Evol, 2009. **9**(4): p. 656-70.
2. Perfeito, L., et al., *Adaptive mutations in bacteria: high rate and small effects.* Science, 2007. **317**(5839): p. 813-5.
3. Suzuki, H. and M.J. Stanhope, *Functional bias of positively selected genes in Streptococcus genomes.* Infect Genet Evol, 2012. **12**(2): p. 274-7.
4. Urwin, R., et al., *Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB.* Mol Biol Evol, 2002. **19**(10): p. 1686-94.
5. Andrews, T.D. and T. Gojobori, *Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis.* Genetics, 2004. **166**(1): p. 25-32.
6. Smith, E.E., et al., *Evidence for diversifying selection at the pyoverdine locus of Pseudomonas aeruginosa.* J Bacteriol, 2005. **187**(6): p. 2138-47.
7. Stanhope, M.J., et al., *Positive selection in penicillin-binding proteins 1a, 2b, and 2x from Streptococcus pneumoniae and its correlation with amoxicillin resistance development.* Infect Genet Evol, 2008. **8**(3): p. 331-9.
8. Cao, P., et al., *Genome-Wide Analyses Reveal Genes Subject to Positive Selection in Pasteurella multocida.* Front Microbiol, 2017. **8**: p. 961.
9. Rasigade, J.P., F. Hollandt, and T. Wirth, *Genes under positive selection in the core genome of pathogenic Bacillus cereus group members.* Infect Genet Evol, 2018. **65**: p. 55-64.
10. Yu, D., et al., *A genome-wide identification of genes undergoing recombination and positive selection in Neisseria.* Biomed Res Int, 2014. **2014**: p. 815672.
11. Lefebure, T. and M.J. Stanhope, *Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition.* Genome Biol, 2007. **8**(5): p. R71.
12. Soyer, Y., et al., *Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected Salmonella serotypes.* BMC Evol Biol, 2009. **9**: p. 264.
13. Petersen, L., et al., *Genes under positive selection in Escherichia coli.* Genome Res, 2007. **17**(9): p. 1336-43.
14. Zhang, Y., et al., *Genes under positive selection in Mycobacterium tuberculosis.* Comput Biol Chem, 2011. **35**(5): p. 319-22.

15. Aguileta, G., et al., *Genes under positive selection in a model plant pathogenic fungus, Botrytis.* Infect Genet Evol, 2012. **12**(5): p. 987-96.
16. Flores-Lopez, C.A. and C.A. Machado, *Differences in inferred genome-wide signals of positive selection during the evolution of Trypanosoma cruzi and Leishmania spp. lineages: A result of disparities in host and tissue infection ranges?* Infect Genet Evol, 2015. **33**: p. 37-46.
17. Toft, C. and S.G. Andersson, *Evolutionary microbial genomics: insights into bacterial host adaptation.* Nat Rev Genet, 2010. **11**(7): p. 465-75.
18. Salyers, A., *Bacteroides of the human lower intestinal tract.* Annual Reviews in Microbiology, 1984. **38**(1): p. 293-313.
19. Finegold, S., *Anaerobic infections in humans*. 1989: Access Online via Elsevier.
20. Snydman, D.R., et al., *National survey on the susceptibility of Bacteroides fragilis group: report and analysis of trends in the United States from 1997 to 2004.* Antimicrob Agents Chemother, 2007. **51**(5): p. 1649-55.
21. Kasper, D.L., A.B. Onderdonk, and J.G. Bartlett, *Quantitative determination of the antibody response to the capsular polysaccharide of Bacteroides fragilis in an animal model of intraabdominal abscess formation.* J Infect Dis, 1977. **136**(6): p. 789-95.
22. Onderdonk, A.B., et al., *The capsular polysaccharide of Bacteroides fragilis as a virulence factor: comparison of the pathogenic potential of encapsulated and unencapsulated strains.* J Infect Dis, 1977. **136**(1): p. 82-9.
23. Duerden, B.I., *Virulence factors in anaerobes.* Clin Infect Dis, 1994. **18 Suppl 4**: p. S253-9.
24. Sears, C.L., et al., *The C-terminal region of Bacteroides fragilis toxin is essential to its biological activity.* Infect Immun, 2006. **74**(10): p. 5595-601.
25. Sund, C.J., et al., *The Bacteroides fragilis transcriptome response to oxygen and H2O2: the role of OxyR and its effect on survival and virulence.* Mol Microbiol, 2008. **67**(1): p. 129-42.
26. Hunter, C.A. and L.D. Sibley, *Modulation of innate immunity by Toxoplasma gondii virulence effectors.* Nat Rev Microbiol, 2012. **10**(11): p. 766-78.
27. Pappas, G., N. Roussos, and M.E. Falagas, *Toxoplasmosis snapshots: global status of Toxoplasma gondii seroprevalence and implications for pregnancy and congenital toxoplasmosis.* Int J Parasitol, 2009. **39**(12): p. 1385-94.
28. Blader, I.J. and J.P. Saeij, *Communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion, and virulence.* Apmis, 2009. **117**(5 - 6): p. 458-476.
29. Melo, M.B., K.D. Jensen, and J.P. Saeij, *Toxoplasma gondii effectors are master*

regulators of the inflammatory response. Trends Parasitol, 2011. **27**(11): p. 487-95.

30. Lorenzi, H., et al., *Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes.* Nat Commun, 2016. **7**: p. 10147.
31. Reid, A.J., et al., *Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy.* PLoS Pathog, 2012. **8**(3): p. e1002567.
32. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.
33. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
34. Loytynoja, A. and N. Goldman, *Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis.* Science, 2008. **320**(5883): p. 1632-5.
35. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W609-12.
36. Anisimova, M., R. Nielsen, and Z. Yang, *Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.* Genetics, 2003. **164**(3): p. 1229-36.
37. Kosakovsky Pond, S.L., et al., *GARD: a genetic algorithm for recombination detection.* Bioinformatics, 2006. **22**(24): p. 3096-8.
38. Pond, S.L., S.D. Frost, and S.V. Muse, *HyPhy: hypothesis testing using phylogenies.* Bioinformatics, 2005. **21**(5): p. 676-9.
39. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Syst Biol, 2010. **59**(3): p. 307-21.
40. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.
41. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
42. Nielsen, R. and Z. Yang, *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.* Genetics, 1998. **148**(3): p. 929-36.
43. Zhang, J., R. Nielsen, and Z. Yang, *Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.* Mol Biol Evol, 2005. **22**(12): p. 2472-9.

44. Yang, Z. and J.P. Bielawski, *Statistical methods for detecting molecular adaptation.* Trends Ecol Evol, 2000. **15**(12): p. 496-503.
45. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995: p. 289-300.
46. Yang, Z., W.S. Wong, and R. Nielsen, *Bayes empirical bayes inference of amino acid sites under positive selection.* Mol Biol Evol, 2005. **22**(4): p. 1107-18.
47. Woolley, S., et al., *TreeSAAP: selection on amino acid properties using phylogenetic trees.* Bioinformatics, 2003. **19**(5): p. 671-2.
48. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. **4**(3): p. 363-71.
49. Gajria, B., et al., *ToxoDB: an integrated Toxoplasma gondii database resource.* Nucleic Acids Res, 2008. **36**(Database issue): p. D553-6.
50. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
51. Merico, D., et al., *Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.* PLoS One, 2010. **5**(11): p. e13984.
52. Isserlin, R., et al., *Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results.* F1000Res, 2014. **3**: p. 141.
53. Orsi, R.H., Q. Sun, and M. Wiedmann, *Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes.* BMC Evol Biol, 2008. **8**: p. 233.
54. Xu, Z., H. Chen, and R. Zhou, *Genome-wide evidence for positive selection and recombination in Actinobacillus pleuropneumoniae.* BMC Evol Biol, 2011. **11**: p. 203.
55. Duerkop, B.A., S. Vaishnava, and L.V. Hooper, *Immune responses to the microbiota at the intestinal mucosal surface.* Immunity, 2009. **31**(3): p. 368-76.
56. Massari, P., et al., *The role of porins in neisserial pathogenesis and immunity.* Trends Microbiol, 2003. **11**(2): p. 87-93.
57. Schauer, K., D.A. Rodionov, and H. de Reuse, *New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'?* Trends Biochem Sci, 2008. **33**(7): p. 330-8.
58. Miethke, M. and M.A. Marahiel, *Siderophore-based iron acquisition and pathogen control.* Microbiol Mol Biol Rev, 2007. **71**(3): p. 413-51.
59. Koebnik, R., *TonB-dependent trans-envelope signalling: the exception or the rule?* Trends Microbiol, 2005. **13**(8): p. 343-7.
60. Pauer, H., et al., *A TonB-dependent outer membrane protein as a Bacteroides fragilis*

fibronectin-binding molecule. FEMS Immunol Med Microbiol, 2009. **55**(3): p. 388-95.

61. Fitzpatrick, D.A. and J.O. McInerney, *Evidence of positive Darwinian selection in Omp85, a highly conserved bacterial outer membrane protein essential for cell viability.* J Mol Evol, 2005. **60**(2): p. 268-73.
62. Su, C., et al., *Recent expansion of Toxoplasma through enhanced oral transmission.* Science, 2003. **299**(5605): p. 414-6.
63. Walzer, K.A., et al., *Hammondia hammondi, an avirulent relative of Toxoplasma gondii, has functional orthologs of known T. gondii virulence genes.* Proc Natl Acad Sci U S A, 2013. **110**(18): p. 7446-51.
64. Lourido, S., et al., *Calcium-dependent protein kinase 1 is an essential regulator of exocytosis in Toxoplasma.* Nature, 2010. **465**(7296): p. 359-62.
65. Wang, J., et al., *Lysine acetyltransferase GCN5b interacts with AP2 factors and is required for Toxoplasma gondii proliferation.* PLoS Pathog, 2014. **10**(1): p. e1003830.
66. Fritz, H.M., et al., *Proteomic analysis of fractionated Toxoplasma oocysts reveals clues to their environmental resistance.* PLoS One, 2012. **7**(1): p. e29955.
67. Chen, S.L., et al., *Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding.* Proc Natl Acad Sci U S A, 2009. **106**(52): p. 22439-44.
68. Matute, D.R., et al., *Evidence for positive selection in putative virulence factors within the Paracoccidioides brasiliensis species complex.* PLoS Negl Trop Dis, 2008. **2**(9): p. e296.
69. Rojas, T.C.G., et al., *Genome-Wide Survey of Genes Under Positive Selection in Avian Pathogenic Escherichia coli Strains.* Foodborne Pathog Dis, 2017. **14**(5): p. 245-252.
70. Lee, S.H., et al., *Toxofilin from Toxoplasma gondii forms a ternary complex with an antiparallel actin dimer.* Proc Natl Acad Sci U S A, 2007. **104**(41): p. 16122-7.
71. Bradley, P.J., et al., *Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in Toxoplasma gondii.* J Biol Chem, 2005. **280**(40): p. 34245-58.
72. Poupel, O., et al., *Toxofilin, a novel actin-binding protein from Toxoplasma gondii, sequesters actin monomers and caps actin filaments.* Mol Biol Cell, 2000. **11**(1): p. 355-68.
73. Delorme-Walker, V., et al., *Toxofilin upregulates the host cortical actin cytoskeleton dynamics, facilitating Toxoplasma invasion.* J Cell Sci, 2012. **125**(Pt 18): p. 4333-42.

謝辞

本研究の遂行及び論文の作成にあたり、多大なご指導を賜りました岐阜大学大学院連合創薬医療情報研究科 武藤吉徳 教授に心より感謝いたします。

また、本論文をご精読いただき、多くのご助言を賜りました岐阜大学大学院連合創薬医療情報研究科 田中香お里 教授、 大橋憲太郎 准教授に深謝いたします。

研究を進めるにあたり、家族や同じ大学院生の仲間の励ましが原動力となり、博士論文をまとめることが出来ました。特に赤堀洋通さんには、発案段階での議論を通して本研究の理解を深めることができたことを感謝いたします。また、支えていただいた全ての皆様に心より感謝申し上げます。

Appendix

Appendix 1　研究のために作成、使用した Perl Script.

1―1　pret_rev_2.pl　FASTA ファイルの前処理プログラム

```
#!/usr/bin/perl -w
#
#配列の前処理
#カレントディレクトリにある ori_pep 内のファイルを処理　→結果は sp_pep に保存
# 種名=ファイル名（拡張子は除いたもの）
#
@bn　　=　　qw(B_fragilis_YCH46　　B_thetaiotaomicron_VPI-5482 B_vulgatus_ATCC8482);

#５０アミノ酸以下の配列を除外してファイルに保存
$n = 1;
print "¥n","***Delete short sequences: < 50 aa***","¥n";
foreach $name (@bn) {
　　open(IN, "<", "ori_pep/".$name.".fas") || die "cannot open :$!¥n";
　　open(OUT1, ">", "tmp_pep/".$name.".fas");
　　open(OUT2, ">", $name.".del");
　　print $n, ") ", $name, "¥n";
　　$sc = "";
　　while(<IN>) {
　　　　if(/^>/) {
　　　　　　if($sc ne "") {
　　　　　　　　$scd = $sc;　#ここから５０アミノ酸以上のチェック処理
　　　　　　　　$scd =~ s/¥s//g;　#スペースの除去
　　　　　　　　if(length($scd)<50) {
　　　　　　　　　　print OUT2 $sn;
　　　　　　　　}
　　　　　　　　else {
　　　　　　　　　　print OUT1 $sn, $sc;
```

```
                        }
                }
                $sn = $_;
                $sc = "";
        }
        else {
                $sc .= $_;
        }
}

#最後の１配列の処理
$scd = $sc;   #ここから５０アミノ酸以上のチェック処理
$scd =~ s/¥s//g;  #スペースの除去
        if(length($scd)<50) {
                print OUT2 $sn;  # ５０アミノ酸以下の配列を書き出
す
        }
        else {
                print OUT1 $sn, $sc;
        }

close(OUT1);
close(OUT2);
close(IN);
++$n;
}


#STOP コドンのある配列を除外する
$n = 1;
print "¥n","***Delete sequences with multiple STOP codons***","¥n";
foreach $name (@bn) {
        open(IN, "<", "ori_nuc/".$name.".fas") || die "cannot open :$!¥n";
        open(OUT1, ">", "tmp_nuc/".$name.".fas");
        open(OUT2, ">", $name.".stop");
        print $n, ") ", $name, "¥n";
```

```
	$sc = "";
	$sn = "";
	while(<IN>) {
		if(/^>/) {
			&fstp($sn, $sc) if($sc ne "");
			$sn = $_;
			$sc = "";
		 }
		else {
			$sc .= $_;
		}
	}

	#最後の１配列の処理
	&fstp($sn, $sc);

	close(OUT1);
	close(OUT2);
	close(IN);
	++$n;
}

sub fstp { # ストップコドン検索
	my $snd = shift @_;
	my $scd = shift @_;
	my $bf = $scd;
	$bf =~ s/¥s//g; # 改行除去
	$bf =~ s/([atgc]{3})$//i; # 最終コドン除去

	while($bf=~s/^([atgc]{3})//i) { # 頭から１コドンずつ取り出し
		next unless($1=~/(taa|tag|tga)/i); # ストップコドン以外無
視
		print OUT2 $snd;  # STOP コドンのある配列を書き出す
		return 1;
	}
```

```
        print OUT1 $snd, $scd;
        return 0;
}


# 塩基配列ファイルに存在するタンパク質配列のみを残して、タンパク質ファイルとして保存
# これはすべてのタンパク質配列が該当するはず
$n = 1;
print "¥n","***Extraction of common sequences***","¥n";
print "---Protein sequences***","¥n";
foreach $name (@bn) {
        open(IN1, "<", "tmp_pep/".$name.".fas") || die "cannot open :$!¥n";
        open(IN2, "<", "tmp_nuc/".$name.".fas") || die "cannot open :$!¥n";
        open(OUT1, ">", "sp_pep/".$name.".fas");
        open(OUT2, ">", $name."_pep.diff");
        print $n, ") ", $name, "¥n";
        # 塩基配列読み込み
        undef %dp;
        while(<IN2>) {
                if(/^>([^¥s]+)/) {
                        $dp{$1}=1; # 配列名を控える
                }
        }
        close(IN2);

        # タンパク配列読み込みと書き出し
        $of = 0;
        while(<IN1>) {
                if(/^>([^¥s]+)/) {
                        if(exists $dp{$1}) {
                                $of = 1; # 対応する配列があれば出力オン
                                delete $dp{$1}; # 1回出力したら重複は許さな
い
                        }
                        else {
```

```
                        $of = 0;
                    }
                }
                print OUT1 $_ if($of>0);
        }
        close(IN1);
        close(OUT1);

        # タンパク配列中に対応するものがなかった塩基配列
        #  t RNA 等の遺伝子が該当する
        print OUT2 "***Nucleotide sequences not present in Protein sequence
file***", "¥n", "¥n";
        foreach $key (keys %dp) {
                print OUT2 $key, "¥n";
        }
        close(OUT2);
        ++$n;
}


#タンパク質ファイルに存在する塩基配列のみを残して、塩基配列ファイルとし
て保存
$n = 1;
print "¥n","***Extraction of common sequences***","¥n";
print "---Nucleotide sequences***","¥n";
foreach $name (@bn) {
        open(IN1, "<", "tmp_nuc/".$name.".fas") || die "cannot open :$!¥n";
        open(IN2, "<", "sp_pep/".$name.".fas") || die "cannot open :$!¥n";
        open(OUT1, ">", "sp_nuc/".$name.".fas");
        open(OUT2, ">", $name."_nuc.diff");
        print $n, ") ", $name, "¥n";
        # アミノ酸配列読み込み
        undef %dp;
        while(<IN2>) {
                if(/^>([^¥s]+)/) {
                        $dp{$1}=1; # 配列名を控える
```

```
		}
	}
	close(IN2);

	# 塩基配列読み込みと書き出し
	$of = 0;
	while(<IN1>) {
		if(/^>([^¥s]+)/) {
			if(exists $dp{$1}) {
				$of = 1; # 対応する配列があれば出力オン
				delete $dp{$1}; # 1 回出力したら重複は許さな
い
			}
			else {
				$of = 0;
			}
		}
		print OUT1 $_ if($of>0);
	}
	close(IN1);
	close(OUT1);

	# 塩基配列中に対応するものがなかったタンパク質配列
	# これは無しのはず
	print OUT2 "***Protein sequences not present in Nucleotide sequence
file***", "¥n", "¥n";
	foreach $key (keys %dp) {
		print OUT2 $key, "¥n";
	}

	close(OUT2);
	++$n;
}

exit;
die "The Unhappy End";
```

1―2 aablst_co.pl オルソログクラスタの構築プログラム

```
#!/usr/bin/perl -w
#
# クラスター化： Core gene で、１菌種１遺伝子のみのクラスタを得る
#
# 種名=ファイル名
@bn = qw(B_fragilis_NCTC B_fragilis_YCH46 B_helcogenes_P36-108
B_salanitronis_DSM18170 B_thetaiota_VPI-5482 B_vulgatus_ATCC8482);

# アミノ酸配列、配列名付け替え
undef @tn;
undef %dp;
$n = 1;
foreach $sn (@bn) {
        open(IN, "<", "sp_pep/".$sn.".fas") || die "cannot open :$!¥n";
        open(OUT, ">", $n.".fa");
        print $n, ") ", $sn, "¥n";
        while(<IN>) {
                chomp;
                if(s/^>//) {
                        $vl = $_;
                        s/¥s.+$//;
                        $ky = $sn . "@" . $_;
                        die if exists $dp{$ky}; # 重複チェック
                        $dp{$ky} = $vl;
                        print OUT ">" . $ky . "¥n";
                }
                else {
                        print OUT $_ . "¥n";
                }
        }
        close(OUT);
        close(IN);
        push(@tn, $n.".fa");
        ++$n;
}
```

```
# orthoMCL 実行
#       perl       orthomcl.pl       --mode       1       --fa_files
"B_fragilis_NCTC.fa,B_fragilis_YCH46.fa,B_helcogenes_P36-108.fa,B_sala
nitronis_DSM18170.fa,B_thetaiota_VPI-5482.fa,B_vulgatus_ATCC8482.fa"
print "¥n";
system("perl orthomcl.pl --mode 1 --fa_files ¥"".join(",",@tn)."¥"");

# クラスタ情報整形
open(IN, "<", "all_orthomcl.out") || die "cannot open :$!¥n";
open(OUT, ">", "oc.tsv"); # クラスタ情報ファイル
while(<IN>) {
       chomp;
       next unless s/^ORTHOMCL¥d+¥(¥d+ genes,¥d+ taxa¥):        //;
       @gm = split(/ /);
       undef %hs;
       undef @ed;
       $i = 0;
       foreach (@gm) {
              s/¥(.+¥)$//; # ファイル名削除=数字に限る?<実体変更
              @ed = split(/@/);
              next if exists $hs{$ed[0]};
              $hs{$ed[0]} = 1;
              $i++;
       }
       next unless($i>$#bn); # 全種類あり
       next unless($#gm==$#bn); # １菌種１遺伝子であること
       print OUT join("¥t",@gm), "¥n";
}
close(OUT);
close(IN);

exit;
die "The Unhappy End";
```

1―3　proc_1.pl　不完全なクラスタの除去プログラム

```
#!/usr/bin/perl -w
#
#カレントディレクトリに"oc.tsv"を置き、sp_pep 内にタンパク質ファイルを置く
#条件にあう cluster を"oc_mod.tsv"に出力
#条件に合わない cluster は"cluster.del"に出力
#
# 種名=ファイル名
@bn　=　qw(B_fragilis_NCTC　B_fragilis_YCH46　B_helcogenes_P36-108 B_salanitronis_DSM18170 B_thetaiota_VPI-5482 B_vulgatus_ATCC8482);

# 全アミノ酸配列読み込み　(全てのファイルについて)
undef %lp;
foreach $sn (@bn) {
	$vl = "";
	open(IN, "<", "sp_pep/".$sn.".fas") || die "cannot open :$!¥n";
	while(<IN>) {
		chomp;
		if(s/^>//) {
			$lp{$ky} = length($vl) if($vl); # 配列長
			s/¥s.+$//;
			$ky = $sn . "@" . $_;
			$vl = "";
		}
		else {
			$vl .= $_;
		}
	}
	close(IN);
	$lp{$ky} = length($vl);
}

# クラスタ処理
$bf = "";
open(IN, "<", "oc.tsv") || die "cannot open :$!¥n"; # クラスタ情報ファ
```

```
イル
open(OUT1, ">", "oc_mod.tsv");
open(OUT2, ">", "cluster.del");

while(<IN>) {
	chomp;
	@gm = split(/¥t/);
	$max = 0;
	$min = 0;
	foreach $ky (@gm) {
		die $ky unless exists $lp{$ky};
		$max = $lp{$ky} if($max<$lp{$ky});
		$min = $lp{$ky} if($min>$lp{$ky} || $min==0);
	}
	if($max/$min<1.5) { # 配列長が 1.5 倍未満
		print OUT1 $_, "¥n";
	}
	else { # 配列長が 1.5 倍以上なら使わない
		$bf .= $_ . "¥n";
	}
}
close(IN);
close(OUT1);

print OUT2 $bf;
close(OUT2);

exit 0;
die "The Unhappy End";
```

1―4 clst.pl　MUSCLE 多重配列アラインメントの作成

```
#!/usr/bin/perl -w
#
# クラスタ毎に塩基配列をアラインメントしてファイルに分ける
#
# 種名=ファイル名
@bn　=　qw(B_fragilis_NCTC　B_fragilis_YCH46　B_helcogenes_P36-108
B_salanitronis_DSM18170 B_thetaiota_VPI-5482 B_vulgatus_ATCC8482);

# 全塩基配列読み込み
undef %dn;
foreach $sn (@bn) {
	$vl = "";
	open(IN, "<", "sp_nuc/".$sn.".fas") || die "cannot open :$!¥n";
	while(<IN>) {
		chomp;
		if(s/^>//) {
			$dn{$ky} = $vl if($vl);
			s/¥s.+$//;
			$ky = $sn . "@" . $_;
			$vl = ">" . $ky . "¥n";
		}
		else {
			$vl .= $_ . "¥n";
		}
	}
	close(IN);
	$dn{$ky} = $vl;
}

# 全アミノ酸配列読み込み < 人間が用意した方が安心
undef %dp;
foreach $sn (@bn) {
	$vl = "";
	open(IN, "<", "sp_pep/".$sn.".fas") || die "cannot open :$!¥n";
	while(<IN>) {
```

```
                chomp;
                if(s/^>//) {
                        $dp{$ky} = $vl if($vl);
                        s/¥s.+$//;
                        $ky = $sn . "@" . $_;
                        $vl = ">" . $ky . "¥n";
                }
                else {
                        $vl .= $_ . "¥n";
                }
        }
        close(IN);
        $dp{$ky} = $vl;
}

# クラスタ処理
$n = 1;
open(IN, "<", "oc_mod.tsv") || die "cannot open :$!¥n"; # クラスタ情報
ファイル
while(<IN>) {
        chomp;
        print "¥n", $n, " ...¥n";
        @gm = split(/¥t/);

# アミノ酸配列書き出し
        open(OUT, ">", "tmp.fas") || die;
        foreach $ky (@gm) {
                die $ky unless exists $dp{$ky};
                print OUT $dp{$ky};
        }
        close(OUT);

# アミノ酸配列アラインメント
        system("./muscle -in tmp.fas -out tmp.aln");

# 塩基配列書き出し
```

```
        open(OUT, ">", "tmp.fas") || die;
        foreach $ky (@gm) {
                die $ky unless exists $dn{$ky};
                print OUT $dn{$ky};
        }
        close(OUT);

# アラインメント反映
        $ff = sprintf("oc_nuc/oc%04d.fas", $n);
        system("perl pal2nal.pl tmp.aln tmp.fas -output fasta > ".$ff); #
FASTA で書き出し

        ++$n;
}
close(IN);

print "Done all alignments¥n";

exit;
die "The Unhappy End";
```

1―5 clst_prank.pl Guidance-Prank 多重配列アラインメントの作成プログラム

```
#!/usr/bin/perl -w
#
# クラスタ毎に塩基配列をアラインメントしてファイルに分ける
# Guidance-PRANK によるアラインメント、Residue confidence → tempDir に保存
# Original Alignment → gp_nuc に保存
# maskLowScoreResidues.pl でマスキング処理
# masked alignment → mask_nuc ホルダに保存

#準備：sp_nuc_mod に塩基配列ファイル
#準備：oc.tsv にクラスタデータ
#上記ファイルの場所に CD 移動

# 下の、種名=ファイル名、は、読み込むファイル名に加えて、oc.tsv 中の種名と一致させる。

# 種名=ファイル名
@bn = qw(TgondiiME49_d TgondiiGT1_d TgondiiVEG_d TgondiiARI_d TgondiiFOU_d Tgondiip89_d TgondiiRUB_d TgondiiTgCatPRC2_d TgondiiVAND_d);

# ディレクトリの作成
if (!-d "./gp_nuc"){
   mkdir "./gp_nuc";
}
else{
   print "Directory already exists!¥n";
}

if (!-d "./mask_nuc"){
   mkdir "./mask_nuc";
}
else{
   print "Directory already exists!¥n";
```

```
}

# 全塩基配列読み込み
undef %dn;
foreach $sn (@bn) {
	$vl = "";
	open(IN, "<", "sp_nuc_mod/".$sn.".fas") || die "cannot open :$!¥n";
	while(<IN>) {
		chomp;
		if(s/^>//) {
			$dn{$ky} = $vl if($vl);
			s/¥s.+$//;
			$ky = $sn . "@" . $_;
			$vl = ">" . $ky . "¥n";
		}
		else {
			$vl .= $_ . "¥n";
		}
	}
	close(IN);
	$dn{$ky} = $vl;
}


# クラスタ処理
$n = 1;
open(IN, "<", "oc.tsv") || die "cannot open :$!¥n"; # クラスタ情報ファイル
while(<IN>) {
	chomp;
	print "¥n", $n, " ...¥n";
	@gm = split(/¥t/);

# 塩基配列書き出し
	open(OUT, ">", "tmp.fas") || die;
	foreach $ky (@gm) {
```

```
            die $ky unless exists $dn{$ky};
            print OUT $dn{$ky};
        }
        close(OUT);

# Guidance-PRANK 塩基配列アラインメント
        system('perl      ~/apli/guidance.v2.02/www/Guidance/guidance.pl
--program GUIDANCE --seqFile tmp.fas --msaProgram PRANK --MSA_Param ¥¥¥-F
¥¥¥-codon  --seqType  codon  --outDir  tempDir  --outOrder  as_input
--bootstraps 100');



# directory を含むファイル名の生成
        $ff = sprintf("gp_nuc/oc%04d.fas", $n);
        $mm = sprintf("mask_nuc/oc%04d.fas", $n);

# Original Alignment → gp_nuc に保存
        system('sh', '-c', "cp  tempDir/MSA.PRANK.aln.Sorted.With_Names
$ff");

# maskLowScoreResidues.pl でマスキング処理
        system("perl
~/apli/guidance.v2.02/www/Guidance/maskLowScoreResidues.pl
tempDir/MSA.PRANK.aln.Sorted.With_Names
tempDir/MSA.PRANK.Guidance_res_pair_res.scr ".$mm." 0.9 nuc");

        system('sh', '-c', "rm -rf tempDir");

        ++$n;
}
close(IN);

print "Done all alignments¥n";

exit;
die "The Unhappy End";
```

1―6 cont.pl PAML の連続実行 (Branch-Site model) プログラム

```
#!/usr/bin/perl -w
#
# 連続 PAML
# A-A1
#
$i = 1; # 開始
$n = $i-1+253; # 終了<588 個

# 出力ファイル
open(OUT2, ">", "res.tsv");

# ファイルループ
for(; $i<=$n; ++$i) {
	$ff = sprintf("oc_nuc/oc%04d.fas", $i);
	open(IN, "<", $ff) || die "cannot open :$!¥n";

# seq ファイル準備
	$ln = 0; # 配列長
	$ns = 0; # 配列数
	$sq = ""; # 配列
	$buf = ""; # FASTA 出力
	undef %hs;
	while(<IN>) {
		chomp;
		if(s/^>//) {
			die if($ln && $ln!=length($sq));
			$ln = length($sq);
			@ed = split(/@/);
			if(exists $hs{$ed[0]}) { # 1 種 1 配列以上は無視
				print $i, ") over spec!¥n";
				$ns = 0;
				last;
			}
			$hs{$ed[0]} = 1;
			$buf .= ">" . $ed[0] . "¥n"; # 種名で置換え
```

```
                $sq = "";
                ++$ns;
        }
        else {
                $buf .= $_ . "¥n";
                $sq .= $_;
        }
    }
    close(IN);
    next unless $ns>0;

# PAML 形式の seq ファイル書き出し
    open(OUT, ">", "A/bac.seq") || die;
    print OUT $ns, "¥t", $ln, "¥n¥n";
    print OUT $buf;
    close(OUT);
    system('sh', '-c', "cp A/bac.seq A1/");

# PAML 実行
    system('sh', '-c', "cd A; codeml");
    system('sh', '-c', "cd A1; codeml");

# 尤度拾い出し
    open(IN, "<", "A/mlc") || die "cannot open :$!¥n";  # 対立仮説
    while(<IN>) {
        next                                                    unless
/^lnL¥(ntime:¥s+¥d+¥s+np:¥s+(¥d+)¥):¥s+(¥-?¥d+¥.¥d+)/;
        $np1 = $1;
        $ll1 = $2;
    }
    close(IN);
    open(IN, "<", "A1/mlc") || die;  # 帰無仮説
    while(<IN>) {
        next                                                    unless
/^lnL¥(ntime:¥s+¥d+¥s+np:¥s+(¥d+)¥):¥s+(¥-?¥d+¥.¥d+)/;
        $np2 = $1;
```

```
                $ll2 = $2;
        }
        close(IN);
        print OUT2 $i, "¥t", $np1, "¥t", $np2, "¥t", $ll1, "¥t", $ll2, "¥n";
}
close(OUT2);

exit;
die "The Unhappy End";
```

1―7 cont_s.pl PAML の連続実行 (Site model) プログラム

```
#!/usr/bin/perl -w
#
# 連続 PAML
# M2a-M1a
#
$i = 1; # 開始
$n = $i-1+253; # 終了<588 個

# 出力ファイル
open(OUT2, ">", "res.tsv");

# ファイルループ
for(; $i<=$n; ++$i) {
        $ff = sprintf("oc_nuc/oc%04d.fas", $i);
        open(IN, "<", $ff) || die "cannot open :$!¥n";

# seq ファイル準備
        $ln = 0; # 配列長
        $ns = 0; # 配列数
        $sq = ""; # 配列
        $buf = ""; # FASTA 出力
```

```
	undef %hs;
	while(<IN>) {
		chomp;
		if(s/^>//) {
			die if($ln && $ln!=length($sq));
			$ln = length($sq);
			@ed = split(/@/);
			if(exists $hs{$ed[0]}) { # 1種1配列以上は無視
				print $i, ") over spec!¥n";
				$ns = 0;
				last;
			}
			$hs{$ed[0]} = 1;
			$buf .= ">" . $ed[0] . "¥n"; # 種名で置換え
			$sq = "";
			++$ns;
		}
		else {
			$buf .= $_ . "¥n";
			$sq .= $_;
		}
	}
	close(IN);
	next unless $ns>0;

# PAML形式のseqファイル書き出し
	open(OUT, ">", "M2a/bac.seq") || die;
	print OUT $ns, "¥t", $ln, "¥n¥n";
	print OUT $buf;
	close(OUT);
	system('sh', '-c', "cp M2a/bac.seq M1a/");

# PAML実行
	system('sh', '-c', "cd M2a; codeml");
	system('sh', '-c', "cd M1a; codeml");
```

```
# 尤度拾い出し
	open(IN, "<", "M2a/mlc") || die "cannot open :$!¥n";  # 対立仮説
	while(<IN>) {
		next unless
/^lnL¥(ntime:¥s+¥d+¥s+np:¥s+(¥d+)¥):¥s+(¥-?¥d+¥.¥d+)/;
		$np1 = $1;
		$ll1 = $2;
	}
	close(IN);
	open(IN, "<", "M1a/mlc") || die;  # 帰無仮説
	while(<IN>) {
		next unless
/^lnL¥(ntime:¥s+¥d+¥s+np:¥s+(¥d+)¥):¥s+(¥-?¥d+¥.¥d+)/;
		$np2 = $1;
		$ll2 = $2;
	}
	close(IN);
	print OUT2 $i, "¥t", $np1, "¥t", $np2, "¥t", $ll1, "¥t", $ll2, "¥n";
}
close(OUT2);

exit;
die "The Unhappy End";
```

Appendix 2　５原虫の系統で site model の下で正の選択が検出された遺伝子.

Gene ID	Gene annotation	*2ΔlnL*	*p*-value	*q*-value
TGME49_306300	hypothetical protein	55.032796	1.12E-12	6.483E-09
TGME49_255180	ubiquitin carboxyl-terminal hydrolase	51.092634	8.04E-12	2.327E-08
TGME49_242415	histone lysine-specific demethylase	50.006526	1.38E-11	2.662E-08
TGME49_293820	calpain family cysteine protease domain-containing protein	42.690486	5.37E-10	7.77E-07
TGME49_211440	hypothetical protein	41.172498	1.15E-09	1.331E-06
TGME49_318770	aurora kinase(incomplete catalytic triad)	36.753874	1.05E-08	1.013E-05
TGME49_206450	autophagy-related cysteine peptidase atg4, putative	36.264678	1.33E-08	1.1E-05
TGME49_268370	non-specific serine/threonine protein kinase	35.812474	1.67E-08	1.208E-05
TGME49_257770	histone lysine methyltransferase SET2	35.000816	2.51E-08	1.574E-05
TGME49_318190	phosphoglycerate mutase family protein	34.842016	2.72E-08	1.574E-05
TGME49_210781	ubiquitin carboxyl-terminal hydrolase	34.435116	3.33E-08	1.752E-05
TGME49_304910	hypothetical protein	33.528072	5.24E-08	2.527E-05
TGME49_306020	hypothetical protein	33.268628	5.97E-08	2.567E-05
TGME49_219650	transporter, small conductance mechanosensitive ion channel (MscS) family protein	33.19015	6.21E-08	2.567E-05
TGME49_202880	carrier superfamily protein	31.963116	1.15E-07	4.437E-05
TGME49_203830	FHA domain-containing protein	30.96752	1.89E-07	6.837E-05
TGME49_304720	hypothetical protein	30.30081	2.63E-07	8.954E-05
TGME49_203520	hypothetical protein	29.477158	3.97E-07	0.0001163
TGME49_266830	Sec7 domain-containing protein	29.46186	0.0000004	0.0001163
TGME49_280800	SWI2/SNF2 SRCAP/Ino80	29.452414	4.02E-07	0.0001163
TGME49_206430	formin FRM1	29.181944	4.61E-07	0.0001271
TGME49_211350	CBS domain-containing protein	28.612416	6.12E-07	0.0001578
TGME49_232080	hypothetical protein	28.565984	6.27E-07	0.0001578
TGME49_223985	serine/threonine specific protein phosphatase	27.499772	1.07E-06	0.000258
TGME49_309910	hypothetical protein	27.270802	0.0000012	0.0002778
TGME49_216070	hypothetical protein	27.073618	1.32E-06	0.0002939
TGME49_250680	TBC domain-containing kinase (incomplete catalytic triad)	26.980674	1.38E-06	0.0002958

TGME49_294730	hypothetical protein	26.705648	1.59E-06	0.0003287
TGME49_203910	TBC domain-containing protein	26.461796	1.79E-06	0.0003573
TGME49_223970	translation elongation and release factors (gtpases), putative	26.207674	2.04E-06	0.0003936
TGME49_295710	HECT-domain (ubiquitin-transferase) domain-containing protein	25.729274	2.59E-06	0.0004739
TGME49_219640	hypothetical protein	25.703878	2.62E-06	0.0004739
TGME49_261850	helicase, putative	25.428536	3.01E-06	0.0005279
TGME49_275410	Proteasome/cyclosome repeat-containing protein	24.729786	4.27E-06	0.0006946
TGME49_318240	Tubulin-tyrosine ligase family protein	24.6702	0.0000044	0.0006946
TGME49_281910	hypothetical protein	24.649578	4.44E-06	0.0006946
TGME49_254135	hypothetical protein	24.608802	4.53E-06	0.0006946
TGME49_272550	hypothetical protein	24.59423	4.57E-06	0.0006946
TGME49_288940	hypothetical protein	24.545146	4.68E-06	0.0006946
TGME49_211340	hypothetical protein	24.374584	0.0000051	0.000738
TGME49_255220	AP2 domain transcription factor AP2VIIb-3	24.223076	0.0000055	0.0007764
TGME49_307860	hypothetical protein	24.015288	0.0000061	0.0008406
TGME49_253800	ribosomal protein L15, putative	23.880896	6.52E-06	0.0008695
TGME49_310950	AP2 domain transcription factor AP2XI-3	23.853416	6.61E-06	0.0008695
TGME49_253750	PLU-1 family protein	23.653626	7.31E-06	0.000921
TGME49_216500	tRNA synthetase, putative	23.65063	7.32E-06	0.000921
TGME49_229630	eIF2 kinase IF2K-A (incomplete catalytic triad)	23.44353	8.12E-06	0.001
TGME49_237830	DNA polymerase I domain-containing protein	23.2946	8.74E-06	0.0010477
TGME49_265190	Ulp1 protease family, C-terminal catalytic domain-containing protein	23.266874	8.87E-06	0.0010477
TGME49_310190	PIK3R4 kinase-related protein (incomplete catalytic triad)	23.083232	9.72E-06	0.0011252
TGME49_214960	AP2 domain transcription factor AP2X-8	22.833608	0.000011	0.0012355
TGME49_292330	hypothetical protein	22.815622	0.0000111	0.0012355
TGME49_202070	hypothetical protein	22.509676	0.0000129	0.0014088
TGME49_220910	HEAT repeat-containing protein	22.319972	0.0000142	0.001522
TGME49_282030	hypothetical protein	22.257006	0.0000147	0.0015435
TGME49_203665	hypothetical protein	22.219718	0.000015	0.0015435
TGME49_207180	indole-3-glycerol phosphate synthase	22.193006	0.0000152	0.0015435

	domain-containing protein			
TGME49_226890	hypothetical protein	22.09297	0.0000159	0.0015628
TGME49_227350	hypothetical protein	22.065898	0.0000162	0.0015628
TGME49_318480	SWI2/SNF2-containing protein RAD5	22.0583	0.0000162	0.0015628
TGME49_247440	hypothetical protein	21.740016	0.000019	0.0018028
TGME49_221230	hypothetical protein	21.687072	0.0000195	0.0018204
TGME49_227450	hydrolase, NUDIX family protein	21.647178	0.0000199	0.0018283
TGME49_289050	FIKK kinase, putative	21.597654	0.0000204	0.0018449
TGME49_224070	hypothetical protein	21.443862	0.0000221	0.0019679
TGME49_253580	CMGC kinase, CDK family	21.101438	0.0000262	0.0022977
TGME49_310910	WD domain, G-beta repeat-containing protein	21.01209	0.0000274	0.0023483
TGME49_264670	DNA polymerase family B protein	20.969106	0.000028	0.0023483
TGME49_273595	hypothetical protein	20.958778	0.0000281	0.0023483
TGME49_280390	HEAT repeat-containing protein	20.93809	0.0000284	0.0023483
TGME49_237410	protein phosphatase 2C domain-containing protein	20.751798	0.0000312	0.0025435
TGME49_209000	HECT-domain (ubiquitin-transferase) domain-containing protein	20.54183	0.0000346	0.0027751
TGME49_292300	DNA-directed RNA polymerase III RPC8	20.51978	0.000035	0.0027751
TGME49_268010	hypothetical protein	20.459946	0.0000361	0.0028236
TGME49_306410	hypothetical protein	20.423446	0.0000367	0.0028323
TGME49_211480	GTP-binding protein engA, putative	20.344698	0.0000382	0.0029092
TGME49_316150	ULK kinase	20.308644	0.0000389	0.0029241
TGME49_272040	WD domain, G-beta repeat-containing protein	20.240498	0.0000403	0.0029905
TGME49_312875	hypothetical protein	20.150082	0.0000421	0.0030845
TGME49_271030	AP2 domain transcription factor AP2VIII-6	20.03767	0.0000446	0.0032268
TGME49_281980	phosphatidate cytidylyltransferase	19.858826	0.0000487	0.0034658
TGME49_311230	hypothetical protein	19.823636	0.0000496	0.0034658
TGME49_264090	hypothetical protein	19.820296	0.0000497	0.0034658
TGME49_288440	NEK kinase	19.555284	0.0000567	0.0039069
TGME49_290990	HEAT repeat-containing protein	19.513842	0.0000579	0.0039103
TGME49_212735	hypothetical protein	19.507364	0.0000581	0.0039103
TGME49_209520	WD domain, G-beta repeat-containing protein	19.435598	0.0000602	0.004005
TGME49_224610	GYF domain-containing protein	19.41402	0.0000609	0.0040056
TGME49_247700	AP2 domain transcription factor AP2XII-4	19.198256	0.0000678	0.0044093

TGME49_305020	hypothetical protein	19.142436	0.0000697	0.0044825
TGME49_262900	hypothetical protein	19.086286	0.0000717	0.0045604
TGME49_247290	hypothetical protein	19.031074	0.0000737	0.0046367
TGME49_227000	hypothetical protein	18.922096	0.0000778	0.004842
TGME49_223880	zinc finger, C3HC4 type (RING finger) domain-containing protein	18.875822	0.0000797	0.0048802
TGME49_309890	hypothetical protein	18.86472	0.0000801	0.0048802
TGME49_264840	ATP-dependent DNA helicase, RecQ family protein	18.726036	0.0000858	0.005173
TGME49_206580	formin FRM2	18.634468	0.0000899	0.0053643
TGME49_203950	Myb family DNA-binding domain-containing protein	18.598382	0.0000915	0.0054041
TGME49_208550	hypothetical protein	18.542716	0.0000941	0.0055015
TGME49_223060	MORN repeat-containing protein	18.473266	0.0000974	0.0056375
TGME49_252210	pentatricopeptide repeat domain-containing protein	18.392652	0.000101	0.005788
TGME49_219450	WD domain, G-beta repeat-containing protein	18.238722	0.00011	0.0061776
TGME49_243610	C-5 cytosine-specific DNA methylase superfamily protein	18.23799	0.00011	0.0061776
TGME49_289710	AP2 domain transcription factor AP2IX-5	18.213928	0.000111	0.0061776
TGME49_233810	Sel1 repeat-containing protein	18.165896	0.000114	0.0062841
TGME49_315760	AP2 domain transcription factor AP2XI-4	18.101466	0.000117	0.0063289
TGME49_228660	Sec7 domain-containing protein	18.09988	0.000117	0.0063289
TGME49_267370	kinesin motor domain-containing protein	17.941938	0.000127	0.0068063
TGME49_202550	NLI interacting factor family phosphatase	17.827044	0.000135	0.0071686
TGME49_293280	cyclin protein	17.789962	0.000137	0.0072087
TGME49_248240	leucine rich repeat-containing protein	17.751436	0.00014	0.0072734
TGME49_318390	hypothetical protein	17.740294	0.000141	0.0072734
TGME49_295658	zinc finger in N-recognin protein	17.716932	0.000142	0.0072734
TGME49_277990	OTU family cysteine protease	17.636488	0.000148	0.0075142
TGME49_291940	hypothetical protein	17.548454	0.000155	0.0078012
TGME49_309000	hypothetical protein	17.395956	0.000167	0.0083327
TGME49_205200	hypothetical protein	17.374388	0.000169	0.0083604
TGME49_246060	DNA-dependent RNA polymerase	17.310236	0.000174	0.0085348
TGME49_232010	protein phosphatase 2C domain-containing	17.189814	0.000185	0.0089982

	protein			
TGME49_254950	RNA cap guanine-N2 methyltransferase	17.154498	0.000188	0.0090679
TGME49_229370	AP2 domain transcription factor AP2VIII-1	17.113554	0.000192	0.0091843
TGME49_225280	hypothetical protein	17.06572	0.000197	0.0093462
TGME49_218400	NEK kinase	17.015046	0.000202	0.0095055
TGME49_314890	ThiF family protein	16.958834	0.000208	0.0096775
TGME49_294860	hypothetical protein	16.943944	0.000209	0.0096775
TGME49_269290	hypothetical protein	16.91124	0.000213	0.0097845
TGME49_272270	radical SAM domain-containing protein	16.752458	0.00023	0.0104822
TGME49_306550	hypothetical protein	16.737208	0.000232	0.0104908
TGME49_213300	hypothetical protein	16.68723	0.000238	0.0105675
TGME49_245720	SWI2/SNF2-containing protein	16.6764	0.000239	0.0105675
TGME49_320150	elongation factor Tu GTP binding domain-containing protein	16.670452	0.00024	0.0105675
TGME49_275420	histone lysine-specific demethylase LSD1/BHC110/KDMA1A	16.657984	0.000241	0.0105675
TGME49_225230	hypothetical protein	16.5743	0.000252	0.0108043
TGME49_243590	endonuclease/exonuclease/phosphatase family protein	16.57355	0.000252	0.0108043
TGME49_219738	hypothetical protein	16.569342	0.000252	0.0108043
TGME49_273800	WD domain, G-beta repeat-containing protein	16.520564	0.000259	0.0110227
TGME49_311730	hypothetical protein	16.500274	0.000261	0.0110268
TGME49_321450	Myb family DNA-binding domain-containing protein	16.484996	0.000263	0.0110308
TGME49_221180	hypothetical protein	16.397978	0.000275	0.0113693
TGME49_241140	DEAD/DEAH box helicase domain-containing protein	16.395786	0.000275	0.0113693
TGME49_235490	hypothetical protein	16.364826	0.00028	0.0114939
TGME49_292950	hypothetical protein	16.31507	0.000287	0.0116974
TGME49_225105	hypothetical protein	16.296548	0.000289	0.0116974
TGME49_204360	subtilisin SUB4	16.190784	0.000305	0.0122593
TGME49_254860	hypothetical protein	16.131082	0.000314	0.012534
TGME49_316680	RNA pseudouridine synthase superfamily protein	16.06354	0.000325	0.0127966
TGME49_255300	hypothetical protein	16.061832	0.000325	0.0127966

TGME49_274000	hypothetical protein	15.973616	0.00034	0.0132352
TGME49_209590	hypothetical protein	15.96936	0.000341	0.0132352
TGME49_226810	histone lysine methyltransferase SET1	15.95396	0.000343	0.0132352
TGME49_316350	hypothetical protein	15.863078	0.000359	0.0137609
TGME49_253170	zinc carboxypeptidase, putative	15.844874	0.000363	0.0138227
TGME49_318275	peptidyl-prolyl cis-trans isomerase, FKBP-type domain-containing protein	15.797162	0.000371	0.014035
TGME49_311220	hypothetical protein	15.756952	0.000379	0.0141899
TGME49_229010	rhoptry neck protein RON4	15.751196	0.00038	0.0141899
TGME49_277550	UvrD/REP helicase domain-containing protein	15.682004	0.000393	0.0145813
TGME49_278730	guanine nucleotide-binding protein, putative	15.669034	0.000396	0.014599
TGME49_314340	Sodium:neurotransmitter symporter family protein	15.621116	0.000405	0.0148363
TGME49_232530	hypothetical protein	15.516402	0.000427	0.0155439
TGME49_212880	surface antigen repeat-containing protein	15.430038	0.000446	0.0161341
TGME49_288420	hypothetical protein	15.385668	0.000456	0.0163116
TGME49_271350	bifunctional protein FolC subfamily protein	15.382804	0.000457	0.0163116
TGME49_225720	hypothetical protein	15.363426	0.000461	0.0163116
TGME49_291930	RNA recognition motif-containing protein	15.352578	0.000464	0.0163116
TGME49_215300	hypothetical protein	15.346394	0.000465	0.0163116
TGME49_215100	PP-loop family protein	15.297162	0.000477	0.0166318
TGME49_208910	hypothetical protein	15.23773	0.000491	0.0170174
TGME49_226470	hypothetical protein	15.154428	0.000512	0.0176396
TGME49_236930	hypothetical protein	15.084096	0.00053	0.018079
TGME49_285530	ribosomal protein RPL35	15.080152	0.000531	0.018079
TGME49_244680	hypothetical protein	15.061358	0.000536	0.0181425
TGME49_228750	calcium dependent protein kinase CDPK7	15.037968	0.000543	0.0182004
TGME49_284620	hypothetical protein	15.03205	0.000544	0.0182004
TGME49_211600	hypothetical protein	14.990432	0.000556	0.018495
TGME49_262935	hypothetical protein	14.959112	0.000565	0.018687
TGME49_258450	hypothetical protein	14.874698	0.000589	0.0191748
TGME49_316430	target of rapamycin (TOR), putative	14.870874	0.00059	0.0191748
TGME49_285200	cleavage and polyadenylation specifity factor protein	14.866018	0.000591	0.0191748
TGME49_285540	DNA-directed DNA polymerase	14.86107	0.000593	0.0191748

TGME49_203780	hypothetical protein	14.838062	0.0006	0.0192933
TGME49_260840	hypothetical protein	14.80224	0.000611	0.0195385
TGME49_304955	serine/threonine specific protein phosphatase	14.786938	0.000615	0.0195584
TGME49_230990	hypothetical protein	14.73817	0.00063	0.0199259
TGME49_264120	Myb family DNA-binding domain-containing protein	14.726408	0.000634	0.0199295
TGME49_210700	hypothetical protein	14.717812	0.000637	0.0199295
TGME49_225960	STE kinase	14.696964	0.000644	0.0200402
TGME49_230140	vacuolar sorting protein 9 (vps9) domain-containing protein	14.672144	0.000652	0.0201806
TGME49_316620	WD domain, G-beta repeat-containing protein	14.658084	0.000656	0.0201964
TGME49_247270	hypothetical protein	14.558982	0.00069	0.0211308
TGME49_271270	hypothetical protein	14.48697	0.000715	0.0217812
TGME49_274010	hypothetical protein	14.418558	0.00074	0.0223984
TGME49_235950	subtilisin SUB8	14.410316	0.000743	0.0223984
TGME49_285895	AP2 domain transcription factor AP2V-2	14.371998	0.000757	0.0227022
TGME49_267540	AGC kinase	14.321354	0.000777	0.0231817
TGME49_233030	gliding-associated protein GAP70	14.310064	0.000781	0.0231817
TGME49_287230	inorganic anion transporter, sulfate permease (SulP) family protein	14.28988	0.000789	0.0232695
TGME49_305980	pyruvate dehydrogenase complex subunit PDH-E3I	14.281332	0.000792	0.0232695
TGME49_265090	hypothetical protein	14.26333	0.000799	0.0233566
TGME49_224870	hypothetical protein	14.21068	0.000821	0.0238791
TGME49_216430	TBC domain-containing protein	14.16552	0.00084	0.0242408
TGME49_288280	hypothetical protein	14.155348	0.000844	0.0242408
TGME49_273380	ion channel protein	14.150498	0.000846	0.0242408
TGME49_258990	bromodomain-containing protein	14.103186	0.000866	0.0246917
TGME49_298010	hypothetical protein	14.055322	0.000887	0.0251665
TGME49_287180	hypothetical protein	14.025788	0.0009	0.0254107
TGME49_306040	CHY zinc finger protein	14.010178	0.000907	0.0254841
TGME49_223920	rhoptry neck protein RON3	13.949306	0.000935	0.0261439
TGME49_251450	hypothetical protein	13.935686	0.000942	0.026213
TGME49_315190	CAM kinase, SNF1 family	13.714026	0.00105	0.0290785
TGME49_213392	surface antigen repeat-containing protein	13.707132	0.00106	0.0292156

TGME49_264820	RbAp48	13.614396	0.00111	0.0304487
TGME49_200410	hypothetical protein	13.543116	0.00115	0.0312498
TGME49_260250	cyclin domain protein, cyclin H family protein	13.530528	0.00115	0.0312498
TGME49_250850	CMGC kinase, putative	13.512908	0.00116	0.0313742
TGME49_259920	Nitric-oxide synthase	13.489556	0.00118	0.0317667
TGME49_278030	hypothetical protein	13.462356	0.00119	0.0318876
TGME49_211310	hypothetical protein	13.403198	0.00123	0.0328076
TGME49_257110	hypothetical protein	13.372538	0.00125	0.0331881
TGME49_221330	DNA gyrase/topoisomerase IV, A subunit domain-containing protein	13.23519	0.00134	0.0351973
TGME49_236850	hypothetical protein	13.233778	0.00134	0.0351973
TGME49_270210	kinesin motor domain-containing protein	13.220844	0.00135	0.0351973
TGME49_201820	hypothetical protein	13.215406	0.00135	0.0351973
TGME49_288240	hypothetical protein	13.176536	0.00138	0.0356582
TGME49_228100	hypothetical protein	13.176024	0.00138	0.0356582
TGME49_258590	hypothetical protein	13.113012	0.00142	0.0363671
TGME49_206540	hypothetical protein	13.111046	0.00142	0.0363671
TGME49_282170	hypothetical protein	13.084234	0.00144	0.0367168
TGME49_308060	hypothetical protein	13.057304	0.00146	0.0370635
TGME49_248290	WD domain, G-beta repeat-containing protein	13.014482	0.00149	0.0376599
TGME49_308000	Gpi16 subunit, GPI transamidase component protein	12.995486	0.00151	0.0379995
TGME49_268320	hypothetical protein	12.949854	0.00154	0.0385867
TGME49_291120	trafficking protein mon1 subfamily protein	12.909308	0.00157	0.0391688
TGME49_271200	AP2 domain transcription factor AP2VIII-5	12.904782	0.00158	0.0392491
TGME49_239830	TBC domain-containing protein	12.828434	0.00164	0.0405655
TGME49_313630	hypothetical protein	12.817528	0.00165	0.0406392
TGME49_213310	hypothetical protein	12.804966	0.00166	0.0407122
TGME49_249560	DNA-directed RNA polymerase alpha chain rpoA	12.73349	0.00172	0.0420057
TGME49_230905	hydrolase, alpha/beta fold family protein	12.694178	0.00175	0.0425588
TGME49_250690	zinc finger (CCCH type) motif-containing protein	12.685414	0.00176	0.0426229
TGME49_299070	pyruvate kinase PyKII	12.658296	0.00178	0.0428741
TGME49_246600	ABC1 family protein	12.651246	0.00179	0.0428741

TGME49_232590	glutamate-cysteine ligase, catalytic subunit domain-containing protein	12.63773	0.0018	0.0428741
TGME49_289950	hypothetical protein	12.63667	0.0018	0.0428741
TGME49_269760	'chromo' (CHRromatin Organization MOdifier) domain-containing protein	12.61431	0.00182	0.0430571
TGME49_218570	Nin one binding (NOB1) Zn-ribbon family protein	12.610514	0.00183	0.0430571
TGME49_305470	hypothetical protein	12.603592	0.00183	0.0430571
TGME49_243290	hypothetical protein	12.598846	0.00184	0.0431171
TGME49_275430	hypothetical protein	12.574034	0.00186	0.04341
TGME49_244010	hypothetical protein	12.46046	0.00197	0.0454277
TGME49_314280	AAR2 protein	12.457496	0.00197	0.0454277
TGME49_264200	hypothetical protein	12.457174	0.00197	0.0454277
TGME49_265280	hypothetical protein	12.41392	0.00202	0.0462584
TGME49_259600	hypothetical protein	12.402188	0.00203	0.0462584
TGME49_223610	hypothetical protein	12.395154	0.00203	0.0462584
TGME49_268330	hypothetical protein	12.3409	0.00209	0.0472536
TGME49_245500	dipeptidyl peptidase iv (dpp iv) n-terminal region domain-containing protein	12.338838	0.00209	0.0472536
TGME49_312430	hypothetical protein	12.327032	0.00211	0.0475202
TGME49_292170	histone lysine methyltransferase, SET, putative	12.299858	0.00213	0.0477847
TGME49_320260	hypothetical protein	12.27919	0.00216	0.0482706
TGME49_271120	endonuclease/exonuclease/phosphatase family protein	12.218964	0.00222	0.0494206
TGME49_320680	AP2 domain transcription factor AP2IV-2	12.192712	0.00225	0.0498966

Appendix 3　Hh と NT 系統で Branch-site model の下で正の選択が検出された遺伝子.

Hh Lineage				
Gene ID	Gene annotation	*2ΔlnL*	*p*-value	*q*-value
TGME49_315860	EF hand domain-containing protein	74.28532	0	0
TGME49_255180	ubiquitin carboxyl-terminal hydrolase	34.790804	3.67E-09	1.0621E-05
TGME49_304910	hypothetical protein	28.298498	1.04E-07	0.00020065
TGME49_313630	hypothetical protein	23.163054	0.00000149	0.00215603
TGME49_306410	hypothetical protein	21.873524	0.00000291	0.00336862
TGME49_306300	hypothetical protein	21.453326	0.00000363	0.00350174
TGME49_224070	hypothetical protein	20.7454	0.00000525	0.004341
TGME49_213310	hypothetical protein	20.30014	0.00000662	0.00478957
TGME49_247440	hypothetical protein	19.643808	0.00000933	0.00600023
TGME49_294860	hypothetical protein	18.97734	0.0000132	0.00764016
TGME49_223060	MORN repeat-containing protein	18.152334	0.0000204	0.01073411
TGME49_311080	transporter, cation channel family protein	17.547356	0.000028	0.01349049
TGME49_309000	hypothetical protein	17.401996	0.0000303	0.01349049
TGME49_274000	hypothetical protein	17.16837	0.0000342	0.01413926
TGME49_239830	TBC domain-containing protein	16.581376	0.0000466	0.01798139
TGME49_290950	clathrin heavy chain, putative	16.43982	0.0000502	0.01815985
TGME49_257730	methionine aminopeptidase, type i, putative	16.08873	0.0000604	0.02056442
TGME49_315760	AP2 domain transcription factor AP2XI-4	15.73817	0.0000727	0.02337709
TGME49_203780	hypothetical protein	14.902872	0.000113	0.03032912
TGME49_223970	translation elongation and release factors (gtpases), putative	14.885202	0.000114	0.03032912
TGME49_224700	hypothetical protein	14.809528	0.000119	0.03032912
TGME49_259640	nucleoporin autopeptidase	14.790066	0.00012	0.03032912
TGME49_295658	zinc finger in N-recognin protein	14.757518	0.000122	0.03032912
TGME49_208430	serine proteinase inhibitor PI-2, putative	14.697516	0.000126	0.03032912
TGME49_214830	hypothetical protein	14.62579	0.000131	0.03032912
TGME49_233810	Sel1 repeat-containing protein	14.370484	0.00015	0.03253255
TGME49_247290	hypothetical protein	14.302866	0.000156	0.03253255
TGME49_254860	hypothetical protein	14.22778	0.000162	0.03253255
TGME49_304720	hypothetical protein	14.212738	0.000163	0.03253255
TGME49_267020	hypothetical protein	13.795738	0.000204	0.0393584

TGME49_292950	hypothetical protein	13.587528	0.000228	0.04256981
TGME49_227300	hypothetical protein	13.493502	0.000239	0.04322913
N-T Lineage				
Gene ID	Gene annotation	*2ΔlnL*	*p*-value	*q*-value
TGME49_255180	ubiquitin carboxyl-terminal hydrolase	41.700256	1.06E-10	6.1353E-07
TGME49_306300	hypothetical protein	24.68538	6.75E-07	0.00195345
TGME49_306410	hypothetical protein	22.183192	0.00000248	0.00478475
TGME49_213310	hypothetical protein	21.106606	0.00000434	0.00627998
TGME49_268370	non-specific serine/threonine protein kinase	20.195134	0.00000699	0.00677196
TGME49_224070	hypothetical protein	19.878192	0.00000825	0.00677196
TGME49_247440	hypothetical protein	19.643794	0.00000933	0.00677196
TGME49_292950	hypothetical protein	19.241104	0.0000115	0.00677196
TGME49_313630	hypothetical protein	19.217028	0.0000117	0.00677196
TGME49_239830	TBC domain-containing protein	19.21024	0.0000117	0.00677196
TGME49_304910	hypothetical protein	18.587966	0.0000162	0.00852415
TGME49_223060	MORN repeat-containing protein	17.53794	0.0000282	0.0136018
TGME49_220360	FAD binding domain-containing protein	17.253414	0.0000327	0.01455905
TGME49_235710	hypothetical protein	16.587596	0.0000465	0.01922443
TGME49_310300	hypothetical protein	16.429092	0.0000505	0.01948627
TGME49_314280	AAR2 protein	16.258274	0.0000553	0.01950896
TGME49_253580	CMGC kinase, CDK family	16.191166	0.0000573	0.01950896
TGME49_306550	hypothetical protein	15.780186	0.0000711	0.0228626
TGME49_275420	histone lysine-specific demethylase LSD1/BHC110/KDMA1A	15.440466	0.0000852	0.02595461
TGME49_236930	hypothetical protein	15.288214	0.0000923	0.02671162
TGME49_246170	ARID/BRIGHT DNA binding domain-containing protein	15.151752	0.0000992	0.02734141
TGME49_315760	AP2 domain transcription factor AP2XI-4	14.961742	0.00011	0.02894
TGME49_313640	hypothetical protein	14.811376	0.000119	0.02994661
TGME49_312370	RNA pseudouridine synthase superfamily protein	14.655614	0.000129	0.03032912
TGME49_295658	zinc finger in N-recognin protein	14.63252	0.000131	0.03032912
TGME49_294730	hypothetical protein	14.298044	0.000156	0.034728
TGME49_310190	PIK3R4 kinase-related protein (incomplete	14.112446	0.000172	0.0368717

	catalytic triad)			
TGME49_271350	bifunctional protein FolC subfamily protein	13.805836	0.000203	0.041963
TGME49_211310	hypothetical protein	13.571576	0.00023	0.04590483
TGME49_304720	hypothetical protein	13.332014	0.000261	0.04910465
TGME49_314920	hypothetical protein	13.320852	0.000263	0.04910465

Appendix 4　NT 系統で Branch-site model の下で正の選択が検出された遺伝子の GO カテゴリー.

Biological Process					
ID	Description	# Genes in GO	# Genes in Overlap	*p*-value	FDR *q*-value
GO:0019222	regulation of metabolic process	123	13	0.000143058	0.005150095
GO:0043412	macromolecule modification	294	19	0.001657531	0.020878166
GO:0050790	regulation of catalytic activity	27	5	0.00215551	0.020878166
GO:0065009	regulation of molecular function	28	5	0.002481643	0.020878166
GO:0060255	regulation of macromolecule metabolic process	102	9	0.004923199	0.020878166
GO:0050789	regulation of biological process	232	15	0.005277944	0.020878166
GO:0006464	cellular protein modification process	256	16	0.005409738	0.020878166
GO:0036211	protein modification process	256	16	0.005409738	0.020878166
GO:0043087	regulation of GTPase activity	21	4	0.005580684	0.020878166
GO:0006468	protein phosphorylation	167	12	0.005799491	0.020878166
GO:0051336	regulation of hydrolase activity	22	4	0.006447988	0.021102506
GO:0010468	regulation of gene expression	89	8	0.007171287	0.02151386
GO:0031323	regulation of cellular metabolic process	95	8	0.010198225	0.022671558
GO:0046578	regulation of Ras protein signal transduction	4	2	0.011921425	0.022671558
GO:1902531	regulation of intracellular signal transduction	4	2	0.011921425	0.022671558
GO:0032012	regulation of ARF protein signal transduction	4	2	0.011921425	0.022671558
GO:0051056	regulation of small GTPase mediated signal transduction	4	2	0.011921425	0.022671558
GO:0080090	regulation of primary metabolic process	98	8	0.012027602	0.022671558
GO:0065007	biological regulation	256	15	0.012141137	0.022671558
GO:2001141	regulation of RNA biosynthetic process	83	7	0.015857994	0.022671558
GO:0051252	regulation of RNA metabolic process	83	7	0.015857994	0.022671558
GO:1903506	regulation of nucleic acid-templated transcription	83	7	0.015857994	0.022671558

GO:0006355	regulation of transcription, DNA-templated	83	7	0.015857994	0.022671558
GO:0016310	phosphorylation	193	12	0.016285982	0.022671558
GO:0019219	regulation of nucleobase-containing compound metabolic process	85	7	0.017724577	0.022671558
GO:0009889	regulation of biosynthetic process	87	7	0.01974158	0.022671558
GO:0051171	regulation of nitrogen compound metabolic process	87	7	0.01974158	0.022671558
GO:0031326	regulation of cellular biosynthetic process	87	7	0.01974158	0.022671558
GO:0010556	regulation of macromolecule biosynthetic process	87	7	0.01974158	0.022671558
GO:2000112	regulation of cellular macromolecule biosynthetic process	87	7	0.01974158	0.022671558
GO:0010646	regulation of cell communication	6	2	0.021412027	0.022671558
GO:0048583	regulation of response to stimulus	6	2	0.021412027	0.022671558
GO:0023051	regulation of signaling	6	2	0.021412027	0.022671558
GO:0009966	regulation of signal transduction	6	2	0.021412027	0.022671558
GO:0043170	macromolecule metabolic process	937	37	0.045243969	0.046536654
GO:0050794	regulation of cellular process	204	11	0.049699442	0.049699442
Molecular Function					
ID	Description	# Genes in GO	# Genes in Overlap	*p*-value	FDR *q*-value
GO:0016772	transferase activity, transferring phosphorus-containing groups	354	22	0.001115347	0.023682619
GO:0003824	catalytic activity	1779	69	0.003330536	0.023682619
GO:0016773	phosphotransferase activity, alcohol group as acceptor	224	15	0.003872025	0.023682619
GO:0003887	DNA-directed DNA polymerase activity	19	4	0.004087329	0.023682619
GO:0001071	nucleic acid binding transcription factor activity	47	6	0.004229852	0.023682619
GO:0003700	transcription factor activity, sequence-specific DNA binding	47	6	0.004229852	0.023682619

GO:0034061	DNA polymerase activity	20	4	0.004795009	0.023682619
GO:0046592	polyamine oxidase activity	2	2	0.004956827	0.023682619
GO:0016647	oxidoreductase activity, acting on the CH-NH group of donors, oxygen as acceptor	2	2	0.004956827	0.023682619
GO:0005096	GTPase activator activity	21	4	0.005580684	0.02369415
GO:0030695	GTPase regulator activity	22	4	0.006447988	0.02369415
GO:0003674	molecular_function	3207	106	0.006612321	0.02369415
GO:0004672	protein kinase activity	173	12	0.007517676	0.024866159
GO:0008047	enzyme activator activity	24	4	0.008441231	0.025926639
GO:0060589	nucleoside-triphosphatase regulator activity	25	4	0.00957363	0.027444407
GO:0005086	ARF guanyl-nucleotide exchange factor activity	4	2	0.011921425	0.028570648
GO:0005085	guanyl-nucleotide exchange factor activity	4	2	0.011921425	0.028570648
GO:0003682	chromatin binding	14	3	0.01249343	0.028570648
GO:0098772	molecular function regulator	43	5	0.01262424	0.028570648
GO:0016301	kinase activity	239	14	0.015333314	0.032966624
GO:0003677	DNA binding	175	11	0.019758622	0.040458131
GO:0030554	adenyl nucleotide binding	514	24	0.021501104	0.04125678
GO:0000166	nucleotide binding	687	30	0.02302704	0.04125678
GO:1901265	nucleoside phosphate binding	687	30	0.02302704	0.04125678
GO:0017076	purine nucleotide binding	610	27	0.027071111	0.044048591
GO:0016740	transferase activity	671	29	0.028906383	0.044048591
GO:0036094	small molecule binding	703	30	0.030452177	0.044048591
GO:0005524	ATP binding	506	23	0.03198916	0.044048591
GO:0032559	adenyl ribonucleotide binding	507	23	0.03261623	0.044048591
GO:0044877	macromolecular complex binding	39	4	0.036366807	0.044048591
GO:0004721	phosphoprotein phosphatase activity	59	5	0.038873052	0.044048591
GO:0035639	purine ribonucleoside triphosphate binding	602	26	0.03911222	0.044048591
GO:0032550	purine ribonucleoside binding	603	26	0.039793711	0.044048591
GO:0001882	nucleoside binding	603	26	0.039793711	0.044048591
GO:0001883	purine nucleoside binding	603	26	0.039793711	0.044048591

GO:0032549	ribonucleoside binding	603	26	0.039793711	0.044048591
GO:0032555	purine ribonucleotide binding	603	26	0.039793711	0.044048591
GO:0032553	ribonucleotide binding	605	26	0.041183728	0.044048591
GO:0016779	nucleotidyltransferase activity	81	6	0.041690443	0.044048591
GO:0016881	acid-amino acid ligase activity	41	4	0.041999819	0.044048591
GO:0030234	enzyme regulator activity	41	4	0.041999819	0.044048591
GO:0016645	oxidoreductase activity, acting on the CH-NH group of donors	10	2	0.04675282	0.047865982
GO:0097367	carbohydrate derivative binding	614	26	0.047897903	0.047897903
Cellular Component					
ID	Description	# Genes in GO	# Genes in Overlap	*p*-value	FDR *q*-value
GO:0070258	inner membrane complex	37	4	0.031199998	0.031199998
GO:0020039	pellicle	37	4	0.031199998	0.031199998

Appendix 5　Site model の下で正の選択が検出された遺伝子の GO カテゴリー.

Biological Process					
ID	Description	# Genes in GO	# Genes in Overlap	*p*-value	FDR *q*-value
GO:0006468	protein phosphorylation	167	10	0.015069889	0.044639517
GO:0043087	regulation of GTPase activity	21	3	0.022873057	0.044639517
GO:0051336	regulation of hydrolase activity	22	3	0.02551487	0.044639517
GO:0016310	phosphorylation	193	10	0.035270037	0.044639517
GO:0050790	regulation of catalytic activity	27	3	0.041067925	0.044639517
GO:0036211	protein modification process	256	12	0.041661441	0.044639517
GO:0006464	cellular protein modification process	256	12	0.041661441	0.044639517
GO:0065009	regulation of molecular function	28	3	0.044639517	0.044639517
Molecular Function					
ID	Description	# Genes in GO	# Genes in Overlap	*p*-value	FDR *q*-value
GO:0008047	enzyme activator activity	24	4	0.005338645	0.044960493
GO:0060589	nucleoside-triphosphatase regulator activity	25	4	0.006071857	0.044960493
GO:0016773	phosphotransferase activity, alcohol group as acceptor	224	12	0.0173886	0.044960493
GO:0004672	protein kinase activity	173	10	0.018666113	0.044960493
GO:0005096	GTPase activator activity	21	3	0.022873057	0.044960493
GO:0030695	GTPase regulator activity	22	3	0.02551487	0.044960493
GO:0016772	transferase activity, transferring phosphorus-containing groups	354	16	0.026206337	0.044960493
GO:0030234	enzyme regulator activity	41	4	0.027838337	0.044960493
GO:0005509	calcium ion binding	85	6	0.029151037	0.044960493
GO:0030554	adenyl nucleotide binding	514	21	0.029726782	0.044960493
GO:0004709	MAP kinase kinase kinase activity	9	2	0.031254489	0.044960493
GO:0098772	molecular function regulator	43	4	0.032054273	0.044960493

GO:0004540	ribonuclease activity	25	3	0.034381553	0.044960493
GO:0004871	signal transducer activity	27	3	0.041067925	0.049031756
GO:0005524	ATP binding	506	20	0.045367304	0.049031756
GO:0032559	adenyl ribonucleotide binding	507	20	0.046147536	0.049031756
GO:0043167	ion binding	1052	36	0.04993126	0.04993126

発表論文

本学位論文は、学術研究雑誌に収載された次の原著論文を基にして作成し、岐阜大学大学院連合創薬医療情報研究科に提出したものである。

1. Yoshizaki, S., Umemura, T., Tanaka, K., Watanabe, K., Hayashi, M., and Muto, Y., Genome-wide evidence of positive selection in *Bacteroides fragilis*. *Computational Biology and Chemistry.* 52,43-50, (2014)
2. Yoshizaki, S., Akahori, H., Umemura, T., Terada, T., Takashima, Y. and Muto, Y., Genome-wide analyses reveal genes subject to positive selection in *Toxoplasma gondii*. *Gene*. 699, 73-79, (2019).