論文 / 著書情報
Article / Book Information

題目(和文)	真核生物遺伝子構造アノテーション手法の研究開発と実ゲノムデータへの適用
Title(English)	
著者(和文)	谷口丈晃
Author(English)	Takeaki Taniguchi
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12617号, 授与年月日:2023年12月31日, 学位の種別:課程博士, 審査員:伊藤 武彦,本郷 裕一,立花 和則,二階堂 雅人,山田 拓司
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12617号, Conferred date:2023/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

東 京 工 業 大 学

博 士 学 位 論 文

真核生物遺伝子構造アノテーション手法の研究開発と実ゲノムデータへの適用

2023 年 9 月

谷口 丈晃

はじめに

真核生物種のゲノム配列決定が比較的容易になった 2023 年においても、精度が十分な遺伝子構造アノテーションを得ることは容易ではない。本研究では、既存手法に対して精度が勝ること、導入と実行が容易であることを目的として、真核生物種ゲノム用の新規の遺伝子構造予測手法を研究・開発した。既存ツールを含めた比較ベンチマークテストにより、本研究において開発した新規ツールである GINGER は、既存ツールに勝る精度を有していることを確認できた。GINGER では、最終結果出力時の閾値自動設定機能やアイソフォーム提示機能なども備わっている。また、ソースツリー、Conda パッケージ、Docker イメージを整備・公開することにより、研究コミュニティに対して広く発信することができている。さらに本研究では、GINGER の有効性を確認するためのデモンストレーションとして、新規に決定したアカムツのゲノム配列に対して GINGER を適用し、遺伝子構造予測結果の網羅性が非常に高いものであることを確認することができた。

本論文は、GINGER に関する研究・開発と、アカムツゲノムに対してのデモンストレーションの二部構成をなっている。前半では、GINGER の基本設計、新規に追加した機能、比較ベンチマークテスト、GINGER に組み込まれた手法が精度に及ぼす影響についての分析、ソフトウェアとしてのパッケージングについて述べる。後半では、アカムツ個体の採取、ゲノム配列決定、GINGER の適用、相同遺伝子同定と機能解析について述べる。

本研究は、東京工業大学の伊藤研究室にて研究・開発が引き継がれてきた真核生物種ゲノム用遺伝子構造予測手法のプロトタイプがベースとなっている。その開発には伊藤研究室に所属する・所属していた多くの方々が関わっている。特に、その中心人物であった篠田氏・小林氏・伊藤教授には心から感謝している。アカムツゲノムを再構築した大内氏にも感謝している。また、本活動に理解を示してくれた前職の上司、時間を割くことを許容してくれた現職の同僚にも感謝したい。

2023 年 9 月

目次

図目次

表目次

1. 序論

ある生物種のゲノム配列が決定された際に、そこにコードされている遺伝子を同定することは、より詳細な解析の出発点であり、その生物種をどの程度正しく理解することができるか、起点となる遺伝子の同定にかかっていると言っても過言ではない。遺伝子を同定する解析工程は、ゲノム配列上にコードされている領域を特定し、ゲノム配列データにその情報を付与する「遺伝子構造アノテーション」と、遺伝子の機能を特定し個々の遺伝子構造に付与する「遺伝子機能アノテーション」に分けることができる。いうまでもなく、遺伝子構造アノテーションの精度が、その後に続く遺伝子機能アノテーションの精度にも影響を与えることになる。更に、遺伝子の発現量を定量することや、ゲノム配列上にあるプロモーター領域を同定するといった後続の解析にも、遺伝子構造アノテーションの精度が影響を及ぼす。

以前から、精度の高い遺伝子構造アノテーションを目指して各所で手法の研究開発が行われている。一言で遺伝子構造と言っても、遺伝子にはタンパク質配列をコードするものや機能性 RNA をコードするものが存在し、タンパク質をコードする遺伝子においては、原核生物種と真核生物種でその構造は大きく異なり、前者と比べて後者の遺伝子構造は非常に複雑であると言える。最も顕著な違いとしては、前者はほとんどの生物種において単一の連続した領域、つまり、単一のエクソンにタンパク質配列がコードされているのに対して、後者では複数のエクソンに分割される形でコードされていることが一般的である点が挙げられる。このようなエクソン・イントロン構造を持つことや、スプライシングの際に mRNA を構成するエクソンが取捨選択されて、つまり選択的スプライシングを受けて、mRNA およびその翻訳産物であるタンパク質に多様性が生じるといった機構があることから、真核生物を対象として、ゲノム配列上のエクソン・イントロン構造までを正確に同定することは非常に難しいテーマとして今なお研究開発が続いている。

東京工業大学の伊藤研究室でも、真核生物種に対するより精度が高い遺伝子構造アノテーション手法を得るべく研究・開発を行ってきており、プロトタイプとして受け継いできている。本研究では、そのプロトタイプを基盤としてアルゴリズムの全体を精査、全面的に改良を施した上で、配布可能なソフトウェアパッケージとして整備することで、既存手法による遺伝子予測よりも精度高く、より簡便な方法で真核生物ゲノムからの遺伝子構造予測を実現することを目標に開発を行うこととした。また、幅広いモデル生物種に対するベンチマークテストを通じて、既存手法に比べての予測精度の高さを示すとともに、非モデル生物種においても本研究で開発する遺伝子予測手法を適用することで、有効性を示すことをもう一つの目標に研究を実施することとした。

本論文の主題は遺伝子構造アノテーションに用いる予測手法の開発についてであるが、上述のように研究内容が手法開発と実データへの適用に分けることができるため、構成を大きく 2 つの章に分けることとした。手法開発に関しては「2. 遺伝子構造アノテーション

手法の開発」に、実データへの適用に関しては「3. 実ゲノムデータへの適用」で述べることとする。それぞれにおいて、背景と目的、研究開発の内容、考察を述べることとするが、最後の「4. 総論」に全体をまとめることとする。

2. 遺伝子構造アノテーション手法の開発

2.1. 背景と目的

DNA シーケンサーとゲノム配列再構築手法の進歩により、完全またはほぼ完全なドラフトゲノム配列データの取得が大幅に容易となった。ゲノム配列データは、生物種やその生物種が持つタンパク質といった構成要素の機能や、進化上の位置付けを解き明かすために利用可能であり、さらに遺伝子発現や代謝に関するオミクスデータを統合することで、より高次の生命現象を探るために利用可能である。もはや、ドラフトゲノム配列の決定は日常的になりつつあり、タンパク質をコードする遺伝子構造の予測は、上述のような二次解析に入っていくための第一段階として避けることのできない工程となっている。これまで、線虫 [1]、ハエ [2,3]、ヒト [4,5] など、第一世代に当たる配列決定技術に基づいた大規模ゲノムプロジェクトでは、ゲノム配列再構築と遺伝子構造や遺伝子機能のアノテーションに多大な人数と膨大な時間を要した。近年になって、ゲノムプロジェクトの金銭的・人的・時間的規模は縮小し、遺伝子構造アノテーションのため遺伝子構造予測は、十分なデータが不足した中で、限られたリソースで実施することが多くなっており、遺伝子構造予測の工程が研究プロセスの大きなボトルネックとなっている場合も多い。通常、これらの予測された遺伝子構造は、追加的な研究から得られる機能情報を付与した形で、一つまり、遺伝子機能情報を付与した形で一公共データベースに登録することになる。しかし、公共データベースに登録されている遺伝子構造・遺伝子機能情報は精度の面から十分とは言えず [6,7]、二次解析によってさまざまな生物種の特徴を正確に理解していくためには、より正確で効率的なアノテーション技術が不足している状況にあると言える。

現在までに、さまざまな遺伝子構造予測法が開発されており、*ab initio* 予測に基づく手法（以下、“*ab initio*-based method”と呼ぶ。）、RNA-Seq データのマッピングや mRNA 配列の再構築とアライメントに基づく手法（以下、“RNA-Seq-based method”と呼ぶ。）、タンパク質配列のアライメントに基づく手法（以下、“Homology-based method”と呼ぶ。）、そしてそれらの結果を統合して最終的な予測を導き出す手法（以下、「統合手法」と呼ぶ。）の 4 つのカテゴリに大別することができる。*ab initio*-based method は、過去の研究において得られている信頼性の高い遺伝子構造情報に基づいてトレーニングされた統計モデルを使用して遺伝子構造を予測するものであり、AUGUSTUS[8]、SNAP[9]、GeneMark[10] といったツールで実現されている。RNA-Seq-based method は、mRNA を配列決定することによって得られる断片配列データを用いるものであり、StringTie[11] や HISAT2[12]、STAR[13] といった、断片配列をゲノム配列に対してマッピングした結果に基づいて遺伝子構造を再構築する手法（以下、“Genome-guided assembly-based method”と呼ぶ。）と、Trans-ABySS[14]、SOAPdenovo-Trans[15]、Trinity[16]、Oases[17] といった、断片配列から mRNA 配列を再構築した後にゲノム配列にアライメントする手法（以下、“*de novo* assembly-based method”と呼

ぶ。）に分類できる。前者はマッピング結果の位置情報を使用して遺伝子構造を構築する。後者は、エクソン・イントロン構造におけるスプライスサイトで保存されている配列パターンをルールとした上で最適なアライメントを探索する手法を用いており、アライメント結果が遺伝子構造予測結果となる。このアライメント手法はスプライスドアライメントと呼ばれている。RNA-Seq-based method は、遺伝子構造予測対象種の mRNA データを用いるものであることから、それを用いて得られる予測結果の信頼性は高くなるであろうことが想定されるが、標的部位あるいは組織で発現していない遺伝子の構造を予測することは原理的に不可能であるといった感度面での限界がある。Homology-based method は、近縁種のタンパク質配列をゲノム配列に対してスプライスドアラインメントするものであり、AAT[18]、GeneWise[19]、Exonerate[20]、genBlastG[21]、Spaln[22] 等が存在している。近縁種のタンパク質配列が豊富に利用可能な場合、RNA-Seq-based method や *ab initio*-based method よりも感度が高くなるが、そうではない場合には大幅に感度が下がることになり、加えて、当然ながら他生物種の遺伝子配列パターンとは似つかない遺伝子―つまり、進化的に遠く離れた独自の遺伝子―については遺伝子構造を予測することはできない。

統合手法は、これらの欠点を互いに補うべく、個別予測手法の出力を統合して遺伝子構造を再構築し、新しい予測結果として出力するものであり、MAKER[23,24]、EVidenceModeler[25]（以降、“EVM”と呼ぶ。）、Finder[26]、EvidentialGene[27] 等のツールが発表されている。文献 25 では、統合手法が、単一の方法では予測できない遺伝子構造を含む、より包括的で正確な遺伝子構造を予測できることが示されている。このアプローチは近年のゲノム研究で選択されることが多くなってきているが、入力データが大量に必要であるため、データセットを構築することが障壁であることと、データの多さや質の精査に人的コストが要求されることから逆に大量に間違った予測を含んでしまうこと―つまり偽陽性を多く含んでしまうこと―が問題として残っている。遺伝子構造予測の成果を評価する観点として、エクソン領域に含まれる塩基をどの程度正しく予測できているかに関する「塩基レベル」の精度、エクソン領域の始点および終点をどの程度正しく予測できているかに関する「エクソンレベル」の精度、遺伝子を構成するエクソン群の始点および終点をどの程度正しく予測できているかに関する「遺伝子レベル」の精度がある。MAKER や EVM 等は多くの研究プロジェクトで用いられているが、それらは塩基レベルとエクソンレベルでは高い予測精度を備えてはいるものの、遺伝子レベルでそうではないことが分かっている。要因の 1 つには、*ab initio*-based method への依存度が高いことである。これまでの研究では、これらの方法では複数の別個の遺伝子が結合される現象（ミスライゲーション）か、単一の遺伝子が複数の遺伝子として分断されてしまう現象（ミスセグメンテーション）の傾向があることが示されている[28]。MAKER および EVM は共に 2008 年の論文で発表されたものであり、開発当時では RNA-Seq で得られる mRNA の断片配列データ（以下、「RNA-Seq データ」と呼ぶ。）やタンパク質配列データが現在ほど豊富ではなかったことから、上述のように *ab initio*-based method に依存度を高めるセッティングあるいはアルゴリズムになっていると想像すること

ができる。

近年における技術進歩により、大量の RNA-Seq データが低コストで入手できるようになり、研究コミュニティの継続的な尽力により、タンパク質配列データをはじめとして公共レポジトリに登録されるデータが大幅に増加している [29]。このような背景から、伊藤研究室では、より正確な遺伝子構造アノテーションを得るために、*ab initio*-based method、RNA-Seq-based method および Homology-based method が出力する予測結果を統一的な方法で統合することで上述の問題を克服できる可能性を模索しつつ、独自のアルゴリズムによる統合手法の研究開発を続けてきている。これらは、2018 年度篠田氏修士課程研究、2019 年度小林氏修士課程研究などとして、都度プロトタイプ化されてきた。しかしそれらプロトタイプは、配布可能にパッケージ化されたものではなく、予測精度の面からも生物種によってバラツキが大きく安定した利用が難しく、広く公開できる状態にまでは仕立てられていなかった。本研究では、これまでのプロトタイプ開発による成果を踏まえ、開発コンセプトは維持したままアルゴリズムと実装を全面的に見直し、アルゴリズムの大幅な改良、機能追加を行うとともに、精度確認のためのベンチマークテストと、研究コミュニティでこの統合手法を広く利用可能とするためのパッケージングを実施した。以降では、まずこの統合手法のアルゴリズムについて述べる。その上で、ベンチマークの方法と結果を述べ、最後にソフトウェアのパッケージングについて述べる。本研究内では、このソフトウェアパッケージおよびそこに実装したアルゴリズムを“GINGER”というコードネームで読んできた。本論文でも同様に GINGER と呼ぶことにする。

2.2. アルゴリズムの開発

2.2.1. 基本コンセプト・処理フローの全体像

GINGER の基本コンセプトは、前述のとおり、*ab initio*-based method、RNA-Seq-based method および Homology-based method の個別手法が出力する予測結果を統合することであり、塩基レベル・エクソンレベル・遺伝子レベルにおいて高い感度を保持したまま、特に遺伝子レベルの偽陽性を削減し、陽性的中率を向上させることを狙って設計されている。

GINGER では、3 通りの個別手法で遺伝子構造予測を行う段階を“Preparation phase”、それらの出力結果を統合して最終的な予測を行う段階を“Merge phase”として分けている。全体を通した流れを図 1 に示す。

Preparation phase では、個別手法から得られる各エクソンに対して独自にスコアを与えることによりできるだけ高い感度を維持しつつ、また陽性的中率も保持することを狙っている。さらに RNA-Seq-based method では、予測結果をそのまま用いるのではなく、最終的な予測結果を出力する直前の段階で TransDecoder によるコード領域予測を挟むことでさらなる精度向上を図っている。「付録 B. TransDecoder の影響」に後述するとおり、ベンチマークにおける比較対象である EVM に対して TransDecoder の結果を与えた場合に遺伝

子レベルでの陽性的中率が向上していることから、本機能が効果的に働いていることが確認できる。

Merge phase では、複数のエクソンから構成される遺伝子（以降、「マルチエクソン遺伝子」と呼ぶ。）と単一のエクソンから構成される遺伝子（以降、「シングルエクソン遺伝子」と呼ぶ。）を異なる手順・基準に従って予測する。まず、マルチエクソン遺伝子の遺伝子構造予測を以下の流れで行う（詳細については後述）。

1. Preparation phase における個別手法が出力したエクソンに対してのスコアの付与
2. エクソン候補のグループ化
   （以下、グループ化されたそれぞれを「エクソングループ」と呼ぶ。）
3. 信頼性の低いイントロン領域でのエクソングループの分割
4. オーバーラップするエクソン、イントロンおよび遺伝子間領域の集約とスコアの付与
5. 各エクソングループにおいて存在する可能性がある遺伝子構造群からの最適解の取得（以下、この最適解の取得を「再構築」と呼ぶ。）

この再構築結果を、マルチエクソン遺伝子の候補として取り扱う。これらにシングルエクソン遺伝子の候補を合わせたうえで、それらに付与されているスコア（以下、「遺伝子構造スコア」と呼ぶ。）に対して閾値を適用することで、最終的な予測結果を絞り込み出力する。

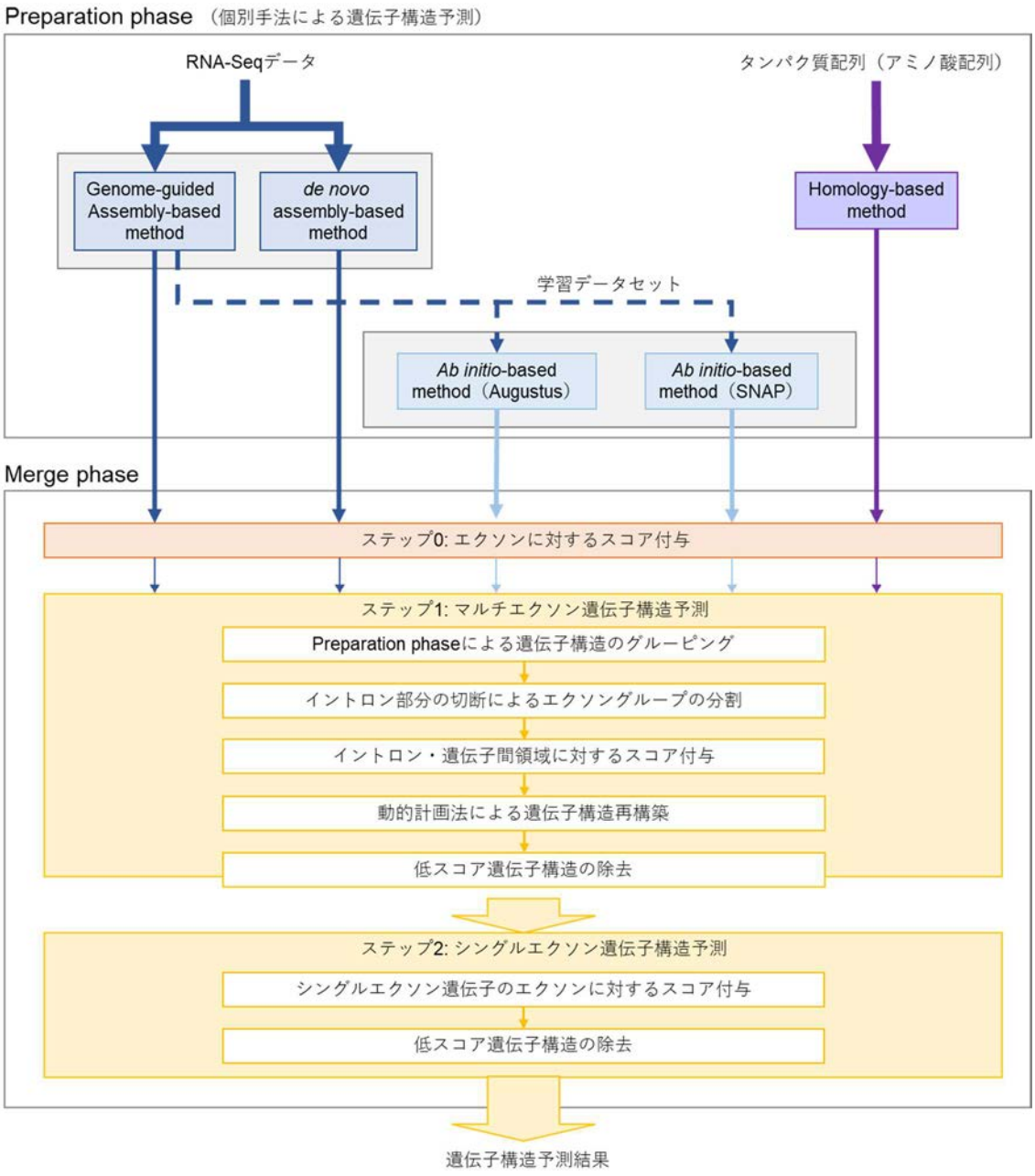


図 1　統合手法（GINGER）の全体像

2.2.2. Preparation phase における処理フロー

Preparation phase では、RNA-Seq-based method、*ab initio*-based method および Homology-based method の 3 通りの個別手法を用いて遺伝子構造予測を行う。

【RNA-Seq-based method】

RNA-Seq-based method の処理内容と実行パラメータを表 1 に示す。「2.1. 背景と目的」で述べたとおり、RNA-Seq-based method は、さらに Genome-guided assembly-based method と *de novo* assembly-based method の 2 種類の手法から構成され、それらによる予測結果を合わせたものが RNA-Seq-based method による最終的な予測結果となる。

Genome-guided assembly-based method では、まず、HISAT2 を使用して RNA-Seq データをゲノム配列にマッピングし、次に、StringTie を使用してマッピング結果から遺伝子構造を予測する。遺伝子の向き情報が与えられなかったシングルエクソン遺伝子については、順鎖・逆鎖を含めた 6 フレームをスキャンし、最も長いものをオープンリーディングフレーム（ORF）として採用する。マルチエクソン遺伝子については StringTie で得たスプライスサイト情報に基づいた遺伝子の向き情報を採用する。

de novo assembly-based method では、マッピングで使用したものと同一の RNA-Seq データに対して Trinity と Oases を適用して mRNA 配列をアセンブリングする。その結果から 300 塩基以上のコンティグを抽出し、CD-HIT[30] でクラスタリングすることにより重複する配列を除去する。これにより得た配列を、GMAP[31] を用いてゲノム配列に対してアラインメントし、95%以上の塩基一致率を有する遺伝子構造を抽出する。

最終的には、Genome-guided assembly-based method と *de novo* assembly-based method の出力結果に対して TransDecoder[32] を適用し、90 アミノ酸以上の ORF を併合した上で最終的な予測結果として出力する。なお、Genome-guided assembly-based method の遺伝子構造予測結果は、後述の *ab initio*-based method の学習データセットとしても用いる。

表 1　RNA-Seq-based method の処理フロー

処理内容および実行パラメータ	
Genome-guided assembly-based method	1. **HISAT2** を用いた RNA-Seq 断片配列のゲノム配列へのマッピング **オプション:--dta, --no-discordant, --no-mixed** 2. マッピング結果を入力とした **StringTie** による遺伝子構造予測 **オプション:デフォルト** 3. 側鎖の決定 • シングルエクソン遺伝子：順鎖・逆鎖を含めた 6 フレームをスキャンし、最長の ORF を参照して遺伝子の向きを決定 • マルチエクソン遺伝子：**StringTie** を用いることで、GT-AG ルールに則って遺伝子の向きを決定 4. **TransDecoder** による ORF 予測（90 アミノ酸以上のものを抽出）
de novo assembly-based method	1. 同一の RNA-Seq データに対して、**Trinity** と **Oases** を別個に適用してアセンブリングし、300 塩基以上のコンティグを抽出 オプション：**デフォルト** 2. **CD-HIT** でクラスタリングし、塩基一致率が 100%のコンティグ群を 1 本に集約 **オプション:-c 1.0** 3. **GMAP** を用いてゲノム配列に対してアライメントし、95%以上の塩基一致率でアライメントされたものを抽出 **オプション：-S -n 1** 4. **TransDecoder** による ORF 予測（90 アミノ酸以上のものを抽出）

【*ab initio*-based method】

ab initio-based method の処理内容と実行パラメータを表 2 に示す。この手法では、まず、Genome-guided assembly-based method が出力した遺伝子構造予測結果から、1,000 個をランダムにサンプリングし学習データセットを構築する。RepeatMasker[33] をゲノム配列に適用することによって検出した反復配列と、学習データセット内のエクソンの位置を照らし合わせて、反復配列とオーバーラップするエクソンを持つ遺伝子構造は学習データセットから排除する。なお、この際に反復配列として取り扱うものには、低複雑度配列は

含まないこととする。この手順により構築した学習データセットを用いて、Augustus と SNAP を使用して個別に統計モデルの学習を行う。得られた学習済み統計モデルをゲノム配列に適用し、Augustus による予測結果と、SNAP による予測結果を個別に Merge phase に受け渡す。

表 2　*ab initio*-based method の処理フロー

処理内容および実行パラメータ
【学習データセットの構築】 1. Genome-guided assembly-based method による遺伝子構造予測結果からシングルエクソン遺伝子を除去 ↓ 2. 300 塩基以上の ORF を抽出 （ある遺伝子から複数の ORF が抽出される場合には最長のものを選択） ↓ 3. **CD-HIT** によるクラスタリングによって重複した遺伝子構造を除去 **オプション:-c 1.0** ↓ 4. **RepeatMasker** をゲノム配列に適用することによって検出した反復配列とオーバーラップするエクソンを持つ遺伝子構造を除去（マイクロサテライトや、プリン塩基で構成される単純配列、ピリミジン塩基で構成される単純配列は対象外） ↓ 5. 遺伝子構造から 1,000 個だけをランダムにサンプリング **【学習および予測】** 1. 同じ学習データセットを **Augustus** と **SNAP** に与えて、統計モデルを学習 ↓ 2. 学習済み統計モデルをゲノム配列に適用して遺伝子構造予測結果を取得（**Augustus** で得た統計モデルは **Augustus** を用いて遺伝子構造を予測；**SNAP** についても同様）

【Homology-based method】

Homology-based method の処理内容と実行パラメータを表 3 に示す。この手法では、複数の近縁生物種についてタンパク質配列のデータセットを用意し、それぞれに対して Spaln を適用することでスプライスドアライメントを得る。ORF の途中に終止コドンが存在する遺伝子構造を除去して最終的な遺伝子構造予測結果とする。

表 3　Homology-based method の処理フロー

処理内容および実行パラメータ
1. タンパク質配列（アミノ酸配列）を問い合わせとして、ゲノム配列に対して **Spaln** を適用してスプライスドアライメントを取得 **オプション:-Q7 -O4 -M1 -T**
⇩
2. アミノ酸一致率の計算
⇩
3. ORF の途中に終止コドンがある遺伝子構造を除去

【エクソンポテンシャルの付与】

個別手法によって得られるエクソンにはエクソンらしさを数値化するためにエクソンポテンシャル（p_{exon}）を付与する。表 4 に p_{exon} の計算方法をまとめる。表に示すとおり、p_{exon} は(0,1]の範囲に収まる値となる。

表 4　エクソンポテンシャルの計算方法

個別手法		エクソンポテンシャル（p_{exon}）
RNA-Seq-based method	Genome-guided assembly-based method	RNA-Seq 断片配列がエクソン領域をカバーしている程度（以下、「カバー率」と呼ぶ。）を用いて p_{exon} を得る。カバー率には StringTie が出力する値を用いる。 カバー率が 100 以上である場合 $p_{exon} = 1$ カバー率が 1 未満である場合 $p_{exon} = 0.7$ その他の場合 $p_{exon} = \log_{10} coverage + 0.7$

個別手法		エクソンポテンシャル（p_{exon}）
	de novo assembly-based method	RNA-Seq 断片配列をアセンブリングして得られたコンティグをゲノム配列に対して GMAP でアライメントして得られる塩基一致率を用いる。 塩基一致率が 100%の場合 $p_{exon} = 1$ その他の場合 $p_{exon} = 0.67$
ab initio-based method	Augustus	Augustus が出力したスコアを用い、以下の式に従って p_{exon} を得る。 $p_{exon} = 10^{Augustus\ score}/10$
	SNAP	SNAP によるスコアの範囲は予め定まっていないことから、SNAP が出力したエクソンをスコア順にソートし、その順序に従ってエクソン数が同等になるように 10 グループに分け、その順序に従って p_{exon} を得る。 $p_{exon} = 10^{\text{何番目のグループであるか}/10}/10$
Homology-based method		Spaln によって得られるアミノ酸一致率を用いる。 アミノ酸一致率 $= \frac{\min(l_{aln}) - l_{mis}}{\max(l_{aln})} \times 100,$ l_{aln}はアライメントに含まれるアミノ酸数、l_{mis}はミスマッチのアミノ酸数 アミノ酸一致率が 30%以上の場合 $p_{exon} =$ アミノ酸一致率$(\%) \times 0.01$ その他の場合 $p_{exon} = 0$

2.2.3. Merge phase における処理フロー

Merge phase では、Preparation phase において個別手法が出力した遺伝子構造予測結果を受け取り、それらをエクソンレベルで組み合わせ、尤もらしい遺伝子候補を導き出すことで最終的な遺伝子構造予測結果を得る。図 2 はその全体像をまとめたものである。概要としては、まず Preparation phase の個別手法が出力したエクソンに対してスコアを付与し、ゲノム配列上でオーバーラップするエクソン同士をエクソングループとしてまとめ上げ、エクソンスコアを活用して定められている基準に従ってエクソングループの分割を行う（図中 A)。続いて、各手法が出力したエクソン・イントロン・遺伝子間領域の内でオーバーラップするものを集約し、それらに対して改めてスコアを付与する（図中 B)。最後に、動的計画法によってスコアの合計が最も高くなる遺伝子構造を抽出する（図中 C)。

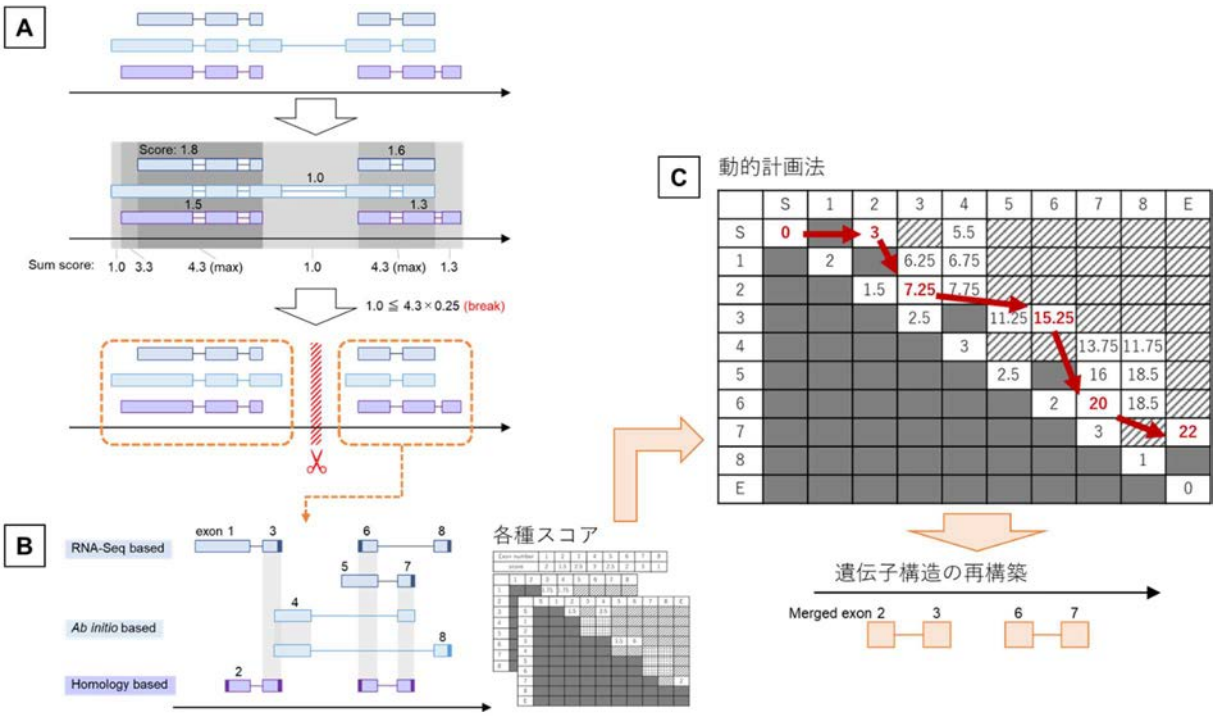


図 2　Merge phase の全体像

【個々のエクソンに対するスコアの算出】（図 1 ステップ 0）

まず、個別手法が出力した遺伝子構造に含まれる個々のエクソンに対してスコアの付与を行う。個々のエクソンについてのスコア（s_{exon}）は、次式のとおり決定する。

$$s_{exon} = p_{exon} \times w_{exon} \qquad (1)$$

p_{exon}はエクソンポテンシャル、w_{exon}は各手法に対して定めた重みである（p_{exon}については、表 4 を参照)。

【マルチエクソン遺伝子の再構築】（図 1 ステップ 1）

このステップでは、Preparation phase において各手法が出力したエクソンをグループ化し、それらのエクソングループにおいて尤もらしい単一の遺伝子構造を再構築する。ただし、以下に示す 2 つの基準に従って分割すべきエクソングループを検出し、分割する工程を挟んだうえで遺伝子構造の再構築に入る。

1 つ目は、あるエクソングループが存在する領域の塩基に対して、エクソングループ領域の繋がり具合を数値化したスコア（s_{base}）を付与し、そのスコアが閾値を下回る場合にその箇所でグループを分割する工程である。s_{base}は以下の式で得る。

$$s_{base} = \sum s_{gene} \text{ and } s_{gene} = \sum(s_{exon} \times l_{exon}) / \sum l_{exon} \quad (2)$$

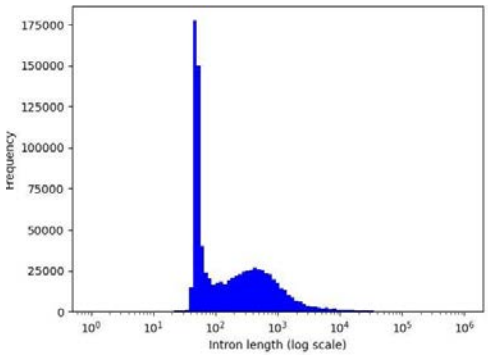
s_{gene}及び s_{exon}はステップ 0 で計算したスコアであり、l_{exon}はエクソンの長さである。閾値t_{group}は次の式で定まる。

$$t_{group} = s_{base\_max} \times 0.25 \quad (3)$$

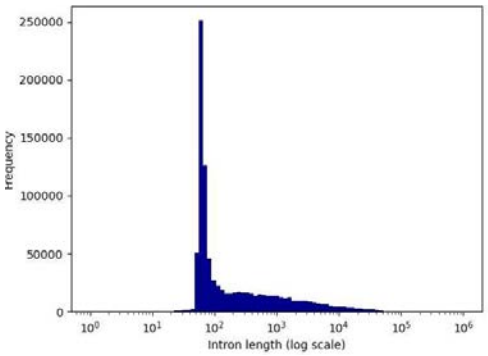
s_{base_max}は、遺伝子構造再構築対象領域（図 2A の灰色領域）における s_{base}の最大値である。

2 つ目は、非常に長いイントロンで分割する工程である。著しく長いイントロンは遺伝子構造予測手法に固有のアーチファクトであることが多く、そのようなアーチファクトを排除することによって、予測精度の向上と計算時間の短縮を実現することができる。アーチファクトの原因は、例えば Homology-based method においては、タンパク質配列をゲノム配列に対してスプライスドアライメントする際に、タンパク質配列の末端がゲノム配列上には存在しないにも関わらず、非常に長いイントロン領域を挟むことにより、末端がアライメントできる領域を無理やりに探し出した場合などに起きる。ある生物種が持つイントロン長の分布は 1〜3 つの極値または対数正規分布の重ね合わせで表現できるという研究報告 [34] があることも参考にして、GINGER では Preparation phase で得られた遺伝子構造予測結果から得られるイントロン長の分布に基づいてイントロン長の上限を決定することとしている。図 3 は、線虫、ハエ、イネ、ゼブラフィッシュ、ヒトのゲノム配列に対して Preparation phase を適用して得たイントロン長の分布であり、上述の研究報告どおりの分布を確認できる。

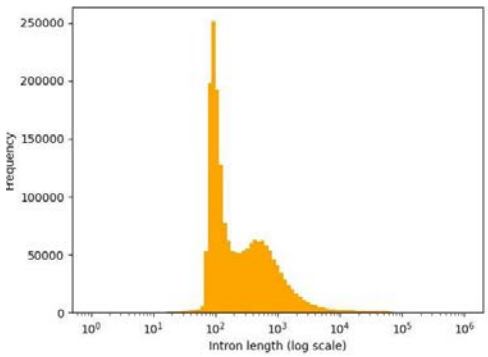
線虫



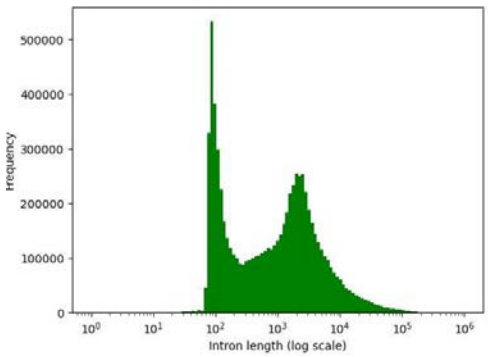
ハエ



イネ



ゼブラフィッシュ



ヒト

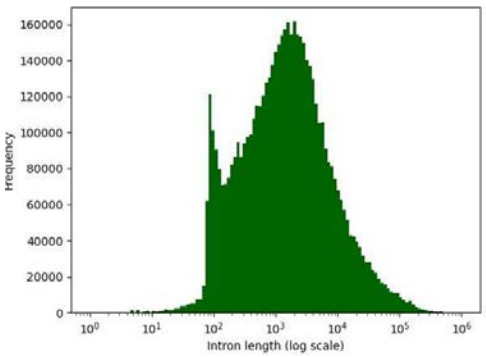


図 3　Preparation phase の遺伝子構造予測結果に基づくイントロン長の分布

本工程では、事前に取得したイントロン長の分布を参照し、イントロン長が長い方の峰に対して正規分布をフィッティングし、その平均値から標準偏差の 3 倍の点を閾値として設定する（図 4)。その閾値を用いて、閾値以上に長いイントロンでエクソングループを分割する。

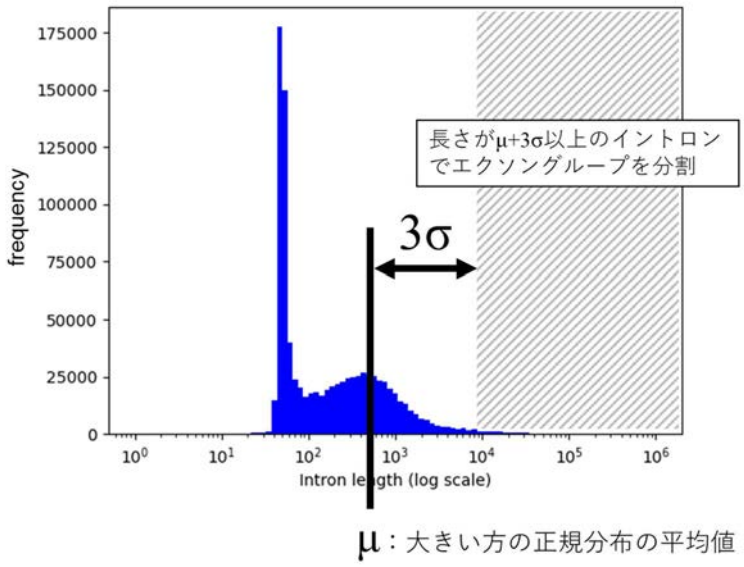


図 4　イントロン長の閾値

エクソングループの分割後は、個別手法が出力したエクソン・イントロン・遺伝子間領域のそれぞれの領域についてオーバーラップするものを集約し、集約した個々の領域に対してスコアを付与する（図 5)。集約したエクソンのスコア（s_{merged_exon}）の値は、個別手法によるエクソンのスコアとその長さの積に基づくものである。

$$s_{merged\_exon} = \sum\left(s_{exon\_max} \times l_{merged\_exon}\right) \tag{4}$$

S_{exon_max} は個別手法に由来するエクソンが複数存在する場合に、その中での最大スコアに該当するものであり、l_{merged_exon} はエクソンの長さである。s_{exon_max} はエクソンスコアを$(0, 1)$の範囲に正規化することで与えられる。イントロンおよび遺伝子間領域の信頼性はエクソンの信頼性に依存すると考えることができるため、イントロンスコアおよび遺伝子間領域スコアは、それらの両端に隣接するエクソンスコアを参照して計算する。具体的には、イントロンスコア（s_{intron}）は、隣接するエクソンの最小スコアによって決定する。遺伝子間領域スコア（$s_{intergenic}$）は、隣接するエクソンが開始コドンで始まるか終止コドンで終わる場合には $s_{intergenic}$=s_{merged_exon} とし、隣接するエクソンが部分的な遺伝子のものである場合には $s_{intergenic}$=1-s_{merged_exon} とする。集約したイントロンスコア（s_{merged_intron}）と遺伝子間領域スコア（$s_{merged\_intergenic}$）は次式のとおりとなる。

$$s_{merged\_intron} = \sum\left(\underset{both\ ends}{\text{average}}(s_{intron}) \times l_{intron}\right) \tag{5}$$

$$s_{merged\_intergenic} = \sum \left(\underset{both\ ends}{\text{average}}(s_{intergenic}) \times l_{intergenic} \right) \quad (6)$$

ここで l_{intron} および $l_{intergenic}$ は、イントロンおよび遺伝子間領域の長さである。

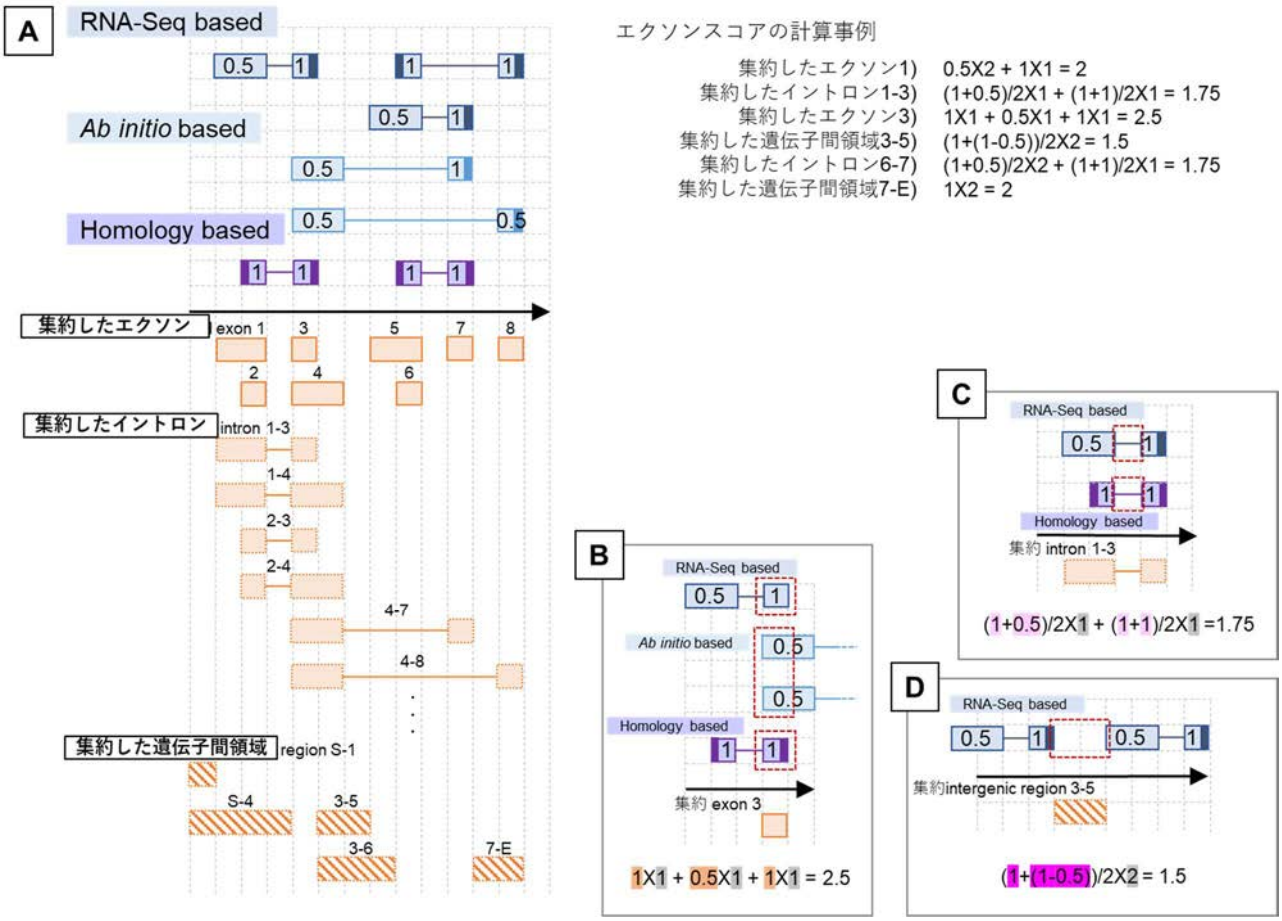


図 5　オーバーラップするエクソン・イントロン・遺伝子間領域の集約

最終的には、エクソングループに含まれるエクソン、イントロンおよび遺伝子間領域を組み合わせることで遺伝子構造を再構築する。これらのゲノム配列上の位置関係と、コドンの読み枠を崩さないという制約を守りながら、付随するスコアの総和が最大となるものを探索する。それらの組み合わせは「経路」として捉えることができ、最もスコアの高い経路を探索する最適化問題と見なすことができる。以下、この経路に付随するスコアを、パススコア（s_{path}）と呼ぶ。GINGER では動的計画法によって最適解を求める（図 6）。s_{path} の値は以下の式で得られる。

$_p$番目のエクソンが遺伝子構造中の末端のものである場合、

$$s_{path,f-c} = max\left(\max\left(s_{path,f-p}\right) + s_{merged\_exon,c} + s_{merged\_intergenic,p-c}\right) \quad (7)$$

そうではない場合、

$$s_{path,f-c} = max\left(\max\left(s_{path,f-p}\right) + s_{merged\_exon,c} + s_{merged\_intron,p-c}\right) \quad (8)$$

$_f$、$_p$および$_c$は、それぞれ、経路中における最初、一つ前および現在のエクソンを指す添え字である。エクソンの途中に終止コドンがあり、その後に開始コドンが続く場合、その位置で別の遺伝子構造が始まるものとして取り扱う。遺伝子構造の最初と最後に開始コドンあるいは終止コドンがあることは必須とはしておらず、部分的な遺伝子構造として取り扱う。動的計画法により得た遺伝子構造のスコア（s_{gene}）が閾値以上である場合にその遺伝子構造を最終結果に含める。s_{gene}は次式に従って決定する。l_{group}は、その遺伝子構造が存在する領域の塩基長である。

$$s_{gene} = s_{path}/l_{group}, \quad (9)$$

s_{gene}に対しての閾値の決定方法については「2.2.5. 遺伝子構造スコアに対する閾値」で述べる。

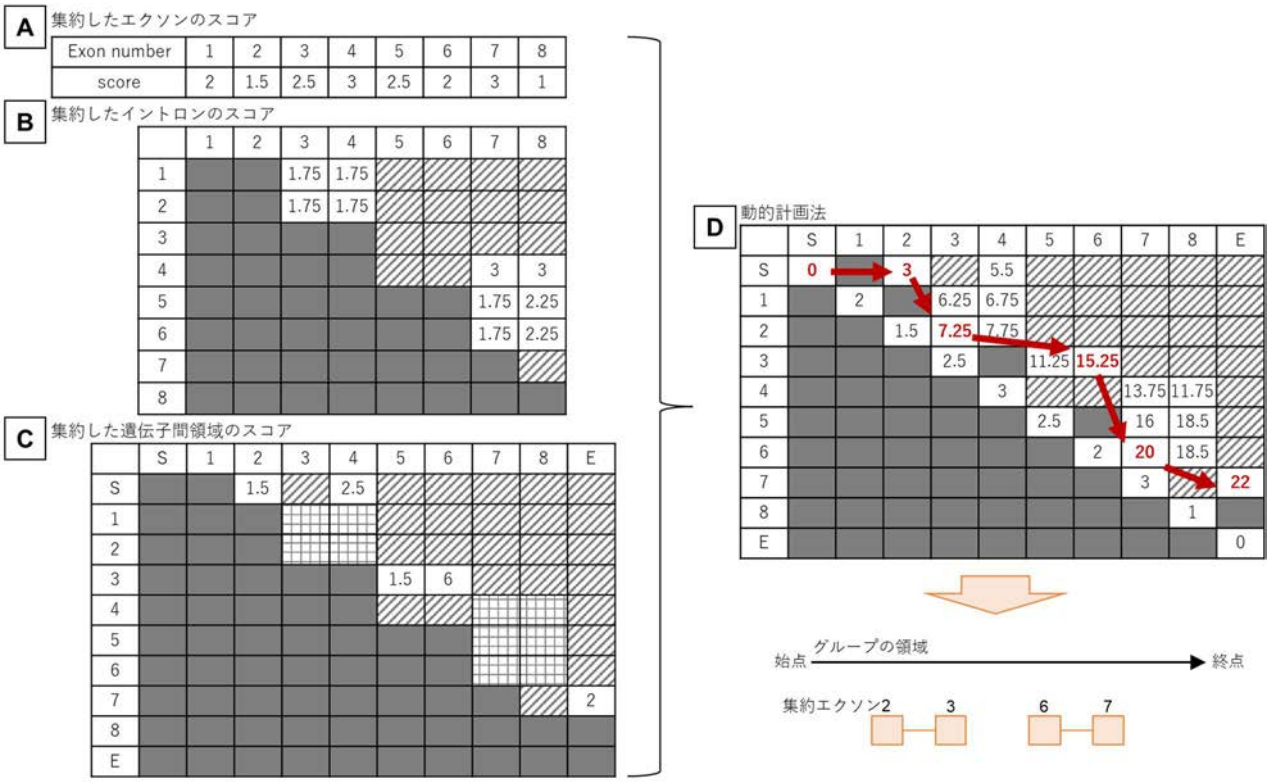


図 6　動的計画法による遺伝子構造の最適解探索

【シングルエクソン遺伝子の抽出】（ステップ 2）

シングルエクソン遺伝子の予測においては、偽遺伝子を予測結果に含めてしまうことが起きてしまいがちであり、マルチエクソン遺伝子の場合と比較して予測精度を向上させることが困難である。そこで GINGER では、少なくとも 2 つの個別手法が支持するシングルエクソンの s_{gene} を求め、閾値以上である場合に最終結果として抽出する。ここで用いる閾値は、マルチエクソン遺伝子の場合に用いる値と同じものである。

2.2.4. エクソンポテンシャルに対する重み

GINGER は、式(1)にあるとおり、エクソンポテンシャル（p_{exon}）と重み係数（w_{exon}）を掛け合わせることでエクソンスコア（s_{exon}）を得る。GINGER の実装では、個別手法に対する重み係数をユーザが設定可能なオプションとして持っているが、デフォルト値も設定されている。デフォルト値の設定は、Genome-guided assembly-based method に対しては 1.6、*de novo* assembly-based method に対しては 1.4、Augustus に対しては 1.8、SNAP に対しては 1.2、Homology-based method に対しては 2.0 であり、これらの値は、遺伝的アルゴリズムによる最適化の結果を踏まえて決定したものである。まず、3 つのモデル生物種（線虫、イネおよびゼブラフィッシュ）それぞれにおいて、エクソンレベルの感度が最大となるように w_{exon} を最適化した。その手順を図 7A に示す。遺伝的アルゴリズムにおける各個体は、重み係数のベクトルであり、各世代 32 個体と設定した。初期値としてランダムに重み係数の値を設定し、20 回の反復でより良い個体を選択した。評価指標にはエクソンレベルでの感度を用いた。母集団から 4 個体ずつ、8 つのグループを作成し、そのグループ内で最も優れた個体をトーナメントセレクション形式で選択し、最終的に 8 個体を抽出した。図 7B は交叉に影響を及ぼすパラメータを説明した図である。交叉には BLX-α 法を用いた。個体 1 と個体 2 の差分（図中の Δx および Δy）とマージンサイズ（$\alpha\Delta x$ および $\alpha\Delta y$）によって次世代個体を選択する範囲を定めた（この場合の α は 0.3）。変異の導入には Wright 法を用いており、そのマージンサイズも 0.3 とした。表 5 は、3 つのモデル生物種に対して遺伝的アルゴリズムにより得られた重み係数の結果と、それらを踏まえて決定した GINGER のデフォルト値である。

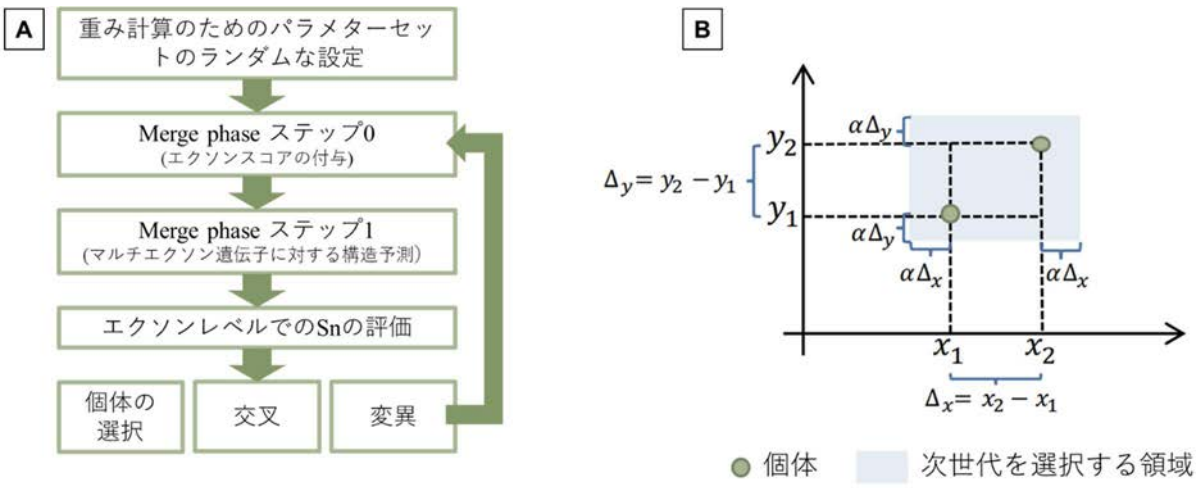


図 7　遺伝的アルゴリズムによる重み係数の最適化

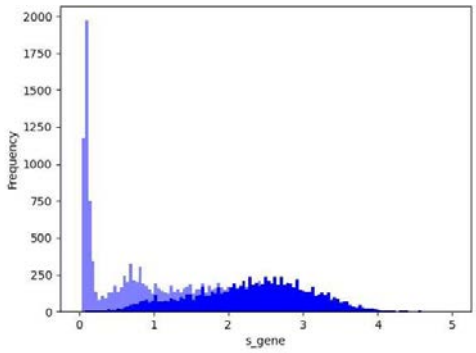
表 5　モデル生物種に対して遺伝的アルゴリズムを適用して得られた重み係数

生物種	RNA-Seq-based method		Homology-based method	*ab initio*-based method	
	Genome-guided assembly-based method	*de novo* assembly-based method		Augustus	SNAP
線虫	1.52	0.98	2.03	1.23	2.24
イネ	1.46	1.24	2.32	1.83	1.15
ゼブラフィッシュ	1.76	1.46	1.82	1.81	1.15
GINGER のデフォルト値	1.6	1.4	2.0	1.8	1.2

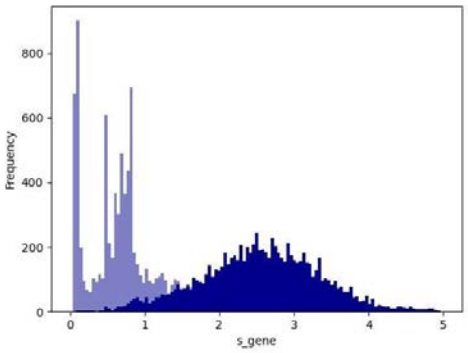
2.2.5.　遺伝子構造スコアに対する閾値の決定方法

Merge phase の最終段階において、遺伝子構造候補の中から閾値以上の遺伝子構造スコア（s_{gene}）を有するものを抽出して最終的な結果に含める。この閾値を s_{gene} の分布から自動決定する機能を以下の手順に基づき開発した。まず、モデル生物種に対して予測された全遺伝子候補に付随する遺伝子構造スコアの分布を図 8 に示す。横軸は s_{gene}、縦軸は予測された遺伝子候補の頻度である。分布中の濃い色の部分は、NCBI が提供する信頼性の高い遺伝子構造アノテーションと一致するものの頻度である。このプロットからもわかるとおり、全体として分布は 2〜3 の多峰性を示し、明確に、NCBI が提供する遺伝子構造と一致する「正しいと考えられる遺伝子構造」が大きい方の峰を形成する傾向が見られた。しかし同時に「正しいもの」と「正しくないもの」が同スコアを持つ領域が存在し、あるスコアを単純に閾値と設定することで明確に両者を二分することが困難であることも確認できた。

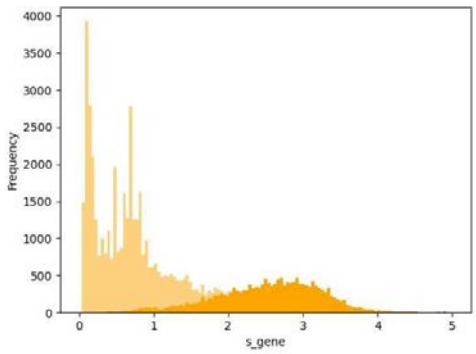
線虫



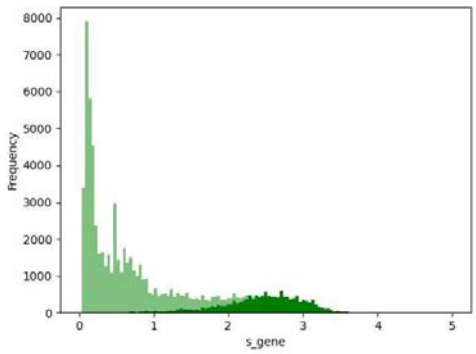
ハエ



イネ



ゼブラフィッシュ



ヒト

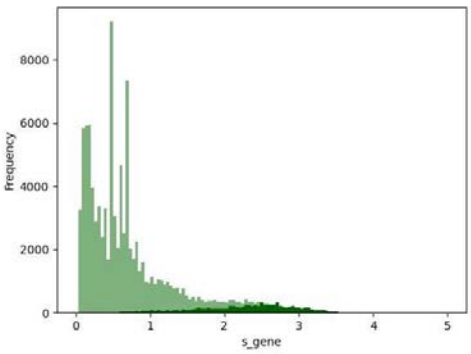


図 8　遺伝子構造スコアの分布

※ 色の濃い部分は NCBI が提供する遺伝子構造アノテーションと遺伝子レベルで一致するものの頻度、薄い部分はそれ以外の頻度

この問題を解決するため、より詳細に s_{gene} の分布を分析した。まず、エクソン数に着目した分析結果を示す。図 9 は、ヒトゲノムから GINGER により予測された遺伝子構造のエクソン数とスコアを 2 軸にとって分布を可視化した結果である。A）は、シングルエクソン遺伝子とマルチエクソン遺伝子を区別せずに取得した分布であるが、スコアが 1.0 の辺りを中心にエクソン数が 1 の遺伝子構造―つまりシングルエクソン遺伝子―が多数存在していることがわかる。一方 B)は、シングルエクソン遺伝子を除去して得られた分布であるが、スコアが 1.0 の左右に二つの峰が形成されることが確認できる。

A）シングルエクソン遺伝子とマルチエクソン遺伝子から得られる分布

B）マルチエクソン遺伝子のみから得られる分布

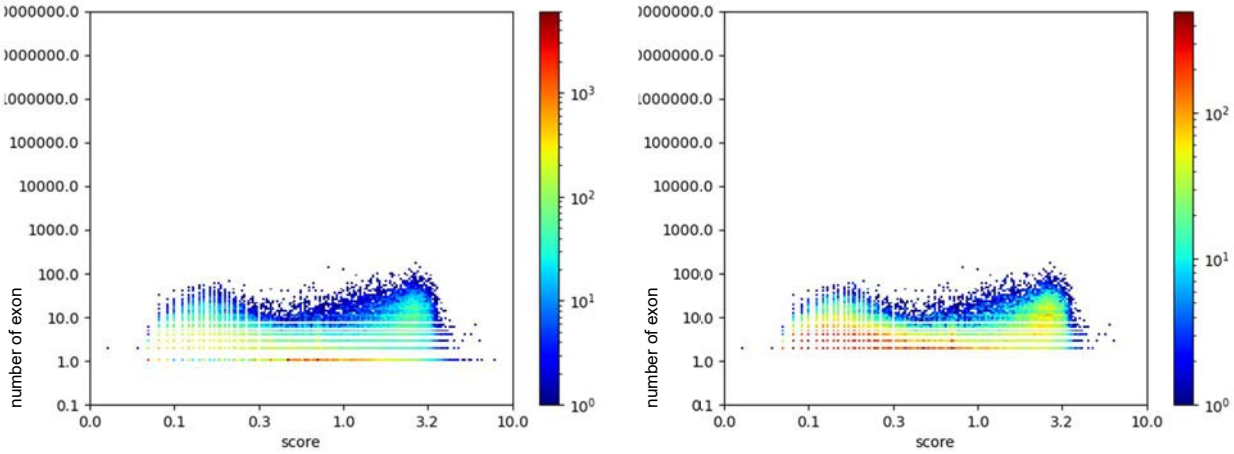


図 9　遺伝子構造のエクソン数とスコアの分布

※ 縦軸がエクソン数、横軸が遺伝子構造スコアであり、色が頻度を表す（赤い方が、頻度が高いことを示す）

続いて図 10 に、遺伝子構造の ORF 長とスコアを 2 軸にとって分布を取得した結果を示す。A）は、シングルエクソン遺伝子とマルチエクソン遺伝子を区別せずに取得した分布であるが、スコアが 3.1 の辺りと 0.1 の辺りのそれぞれに ORF 長が 100～10 万塩基くらいまで広がりがある峰が存在していることと、ORF 長が 1000 塩基未満にスコアが 0.1 から 3.1 までの間に多数の遺伝子構造が存在していることが見て取れる。B)はシングルエクソン遺伝子を除去して得られた分布であり、スコアが 1.0 の左右に二つの峰をより明確に見て取ることができる。

A）シングルエクソン遺伝子とマルチエクソン遺伝子から得られる分布

B）マルチエクソン遺伝子のみから得られる分布

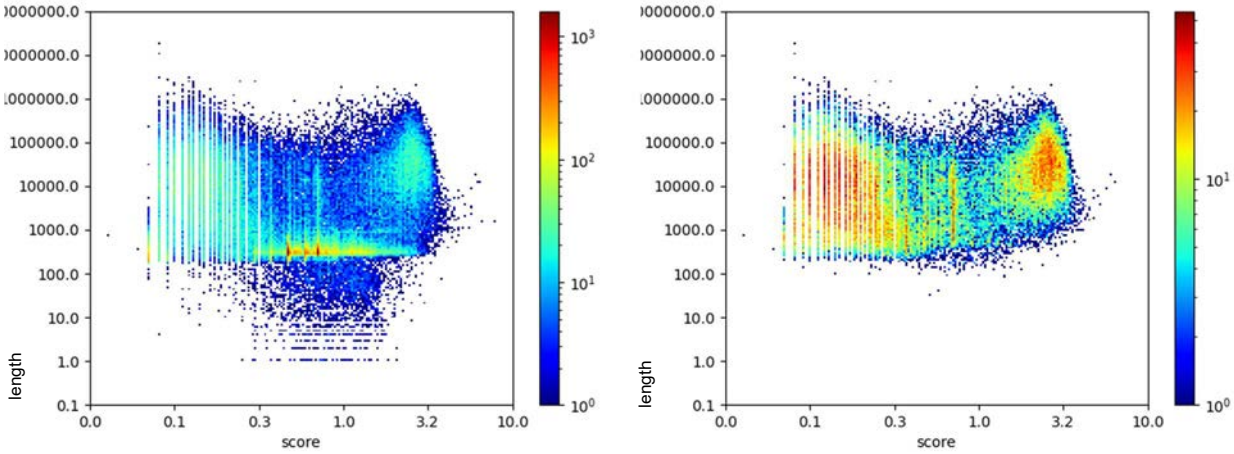


図 10　遺伝子構造の ORF 長とスコアの分布

※ 縦軸が ORF 長、横軸が遺伝子構造スコアであり、色が頻度を表す（赤い方が、頻度が高いことを示す）

これらの分析から、マルチエクソン遺伝子のみのスコア分布から閾値を決定する方が、シングルエクソン遺伝子も含めて得られたスコア分布に基づいた閾値決定よりも、正解・不正解を高い精度で分離する可能性が高いことが示唆された。図 11 には、A）シングルエクソン遺伝子のみのスコア分布、B）マルチエクソン遺伝子のみのスコア分布を示す。確かに B）は二峰性に近い形になった一方で、A）はところどころに突出して多い bin が存在する形となっている。上述したように、そもそも、シングルエクソン遺伝子の偽陽性を低減させることが困難であったことからも、B)のみを参照して閾値を決定することに妥当性はあると考えられた。

A）シングルエクソン遺伝子

B）マルチエクソン遺伝子

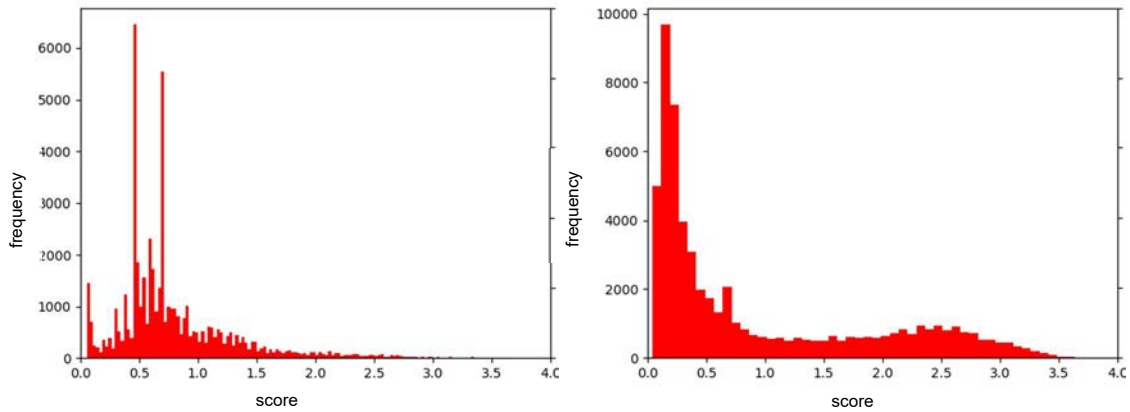
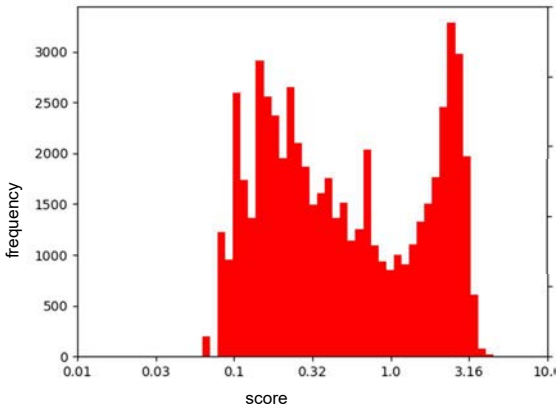


図 11　遺伝子構造スコアの分布

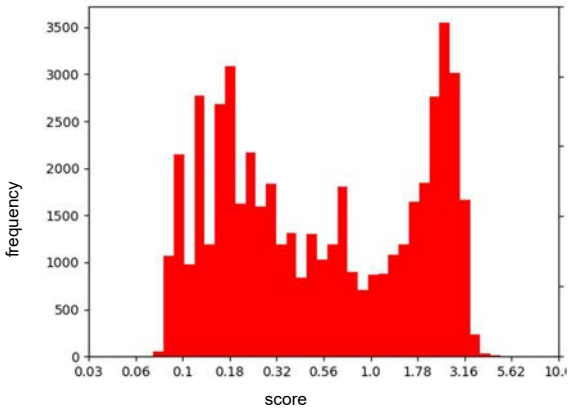
※ 縦軸は頻度

図 12 は、マルチエクソン遺伝子のみを対象に、遺伝子スコアの対数値で分布を取得した結果である。スコアが 1.0 辺りを極小としてそれよりもスコアが大きい方に明確な一つの峰が形成された。一方 ORF 長との関連性に関しては、大きな差を見て取ることはできなかった。

A）ORF 長が 100 以上



B）ORF 長が 1,000 以上



C）ORF 長が 10,000 以上

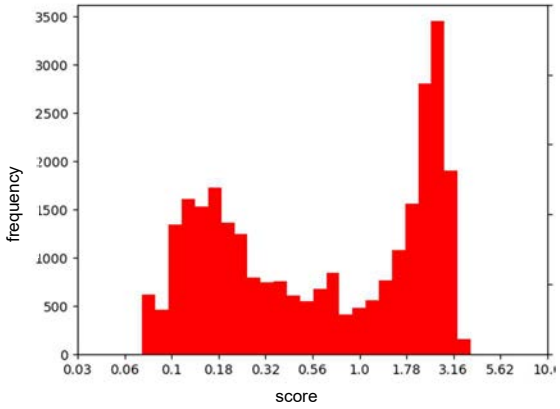


図 12　マルチエクソン遺伝子における遺伝子構造スコアの対数分布

※ 縦軸は頻度、横軸は底を 10 として得られるスコアの対数値

以上の結果より、本研究開発では、s_{gene} に対する閾値を以下に示す手順で自動決定するようにした。

1. ORF 長が 1,000 以上のマルチエクソン遺伝子のみを対象にして、底を 10 として得られる s_{gene} の対数値を取得
2. bin サイズが 0.05 の分布を取得
3. 着目点を中心とした 0.5 の範囲でスムージング
4. スコアが大きい方からスキャンして最初の極大値の後に出現する極小値を閾値とする

この自動決定法を用いることによって、ヒトに関しても「2.3.4. 統合手法間の精度比較」に

示すように他手法に勝る精度を得ることができるようになった。

2.2.6. アイソフォーム候補の提示

前節までに示したアルゴリズムに基づく GINGER のデフォルト出力は、動的計画法によって得られた遺伝子構造スコアが最大となる最適解のみであるが、その準最適解にはアイソフォームが含まれている可能性は十分にある。様々な生命現象において、アイソフォームが重要な役割を果たすことが知られていることから準最適解を出力するオプションも合わせて開発した。「2.2.3. Merge phase における処理フロー」で述べた動的計画法において、準最適解を得るにあたって全ての経路を保持していくことも可能ではあるが、大量のメモリと計算時間を消費してしまう。そこで、エクソンを経路に追加していくにあたって、複数の経路候補を評価することとなるが、その際に、それらのスコア（s_{path_cand}）を参照し、最大のものから一定の範囲内にあるもののみを保持していくものとすることで、準最適解を効果的に出力できるように開発した。この範囲は d_{drop} としてオプション指定できるようになっている（図 13）。

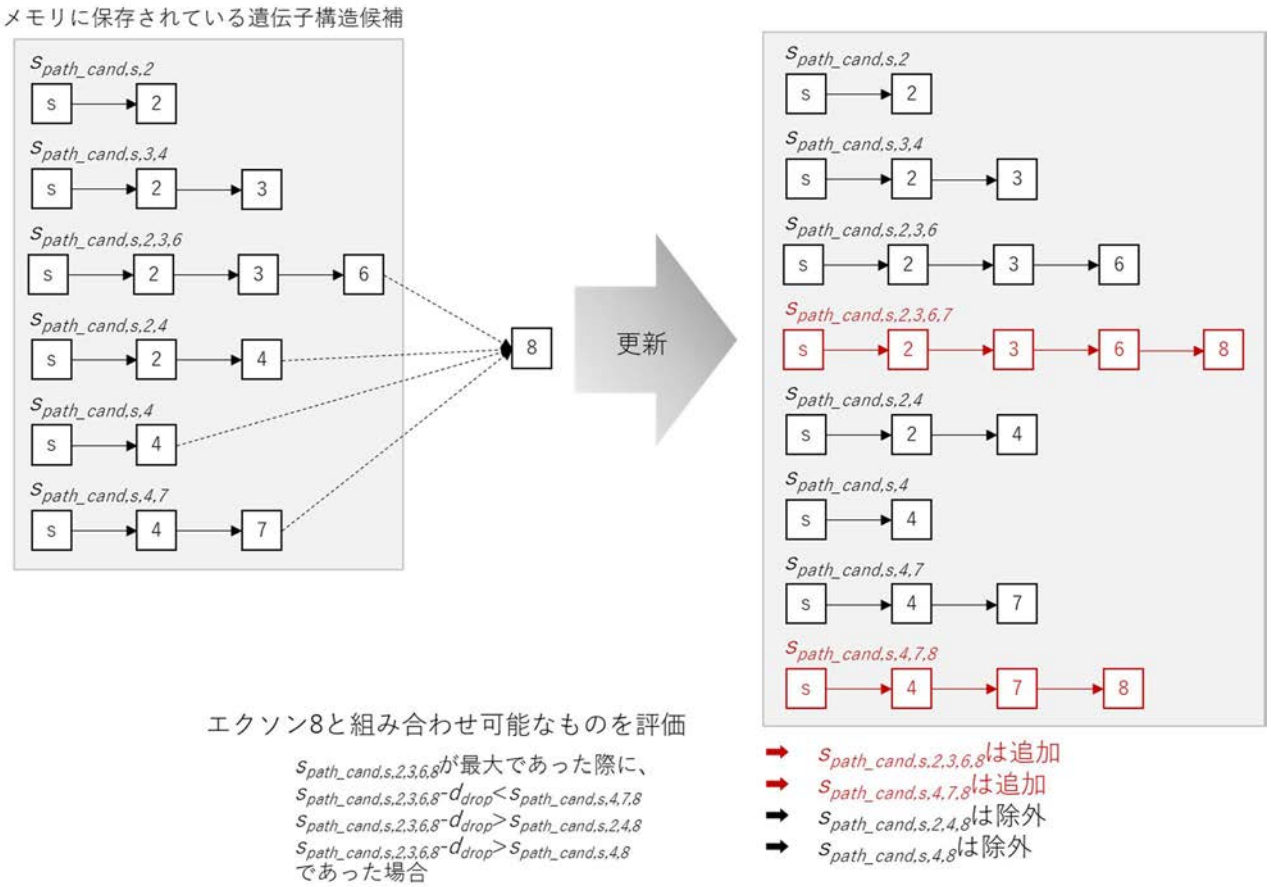


図 13　遺伝子構造の準最適解を保持するロジック

経路探索の過程において、あるエクソンを追加する際の s_{path_cand} を観察すると、様々なパターンが見られた。図 14 は、その際の s_{path_cand} を降順にソートし、順位と値の大きさをプロットしたものである。例示したように、最大スコアと同じ大きさ、あるいはほぼ同じ大きさの経路候補が複数存在する場合が多々見受けられた。一方で例 4 のようにスコアが徐々に下がっていく傾向を見せるものもあった。

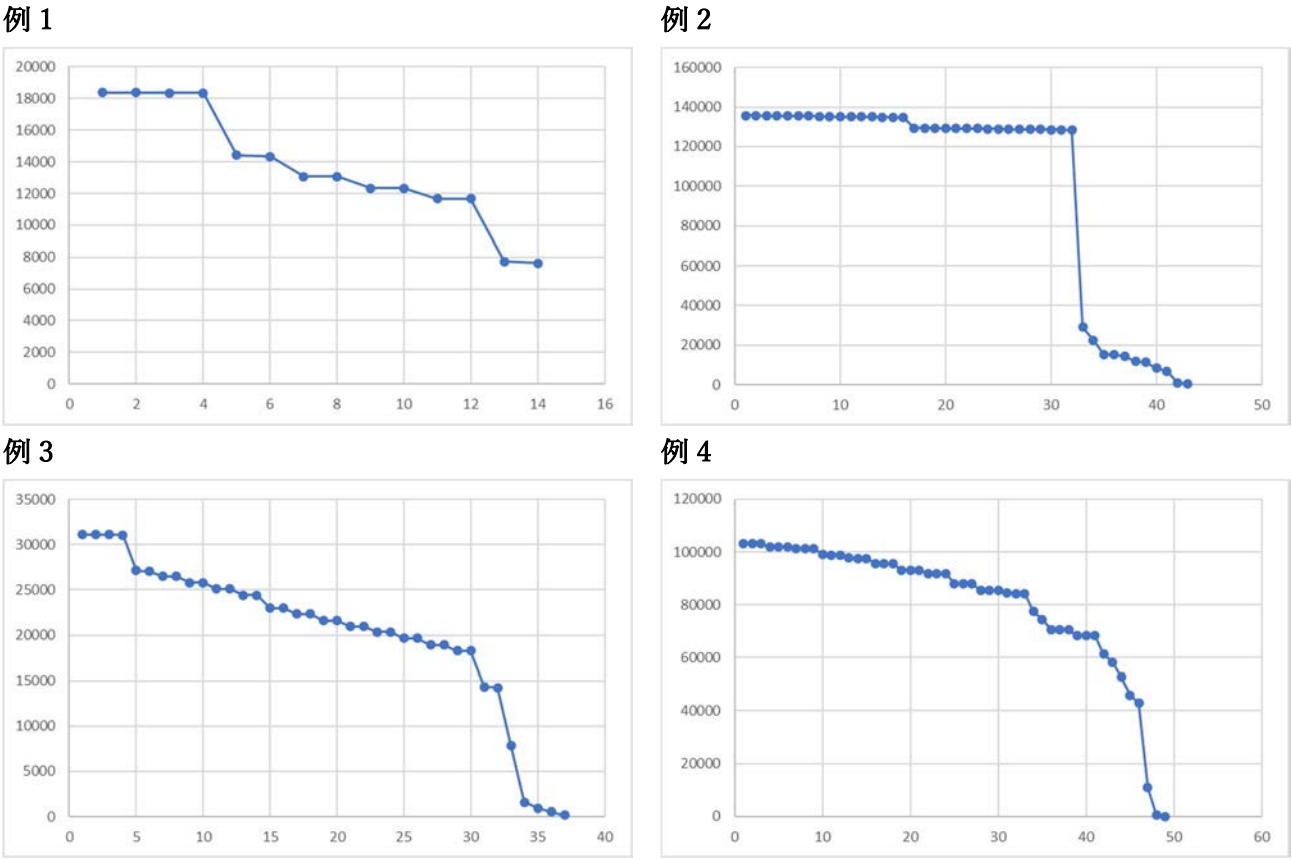
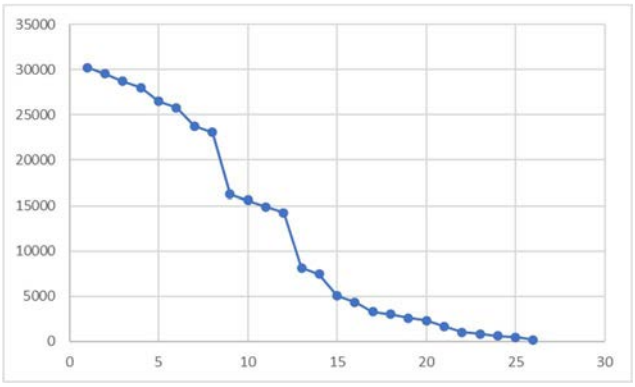


図 14　パススコア s_{path_cand} の大きさと順位の関係性

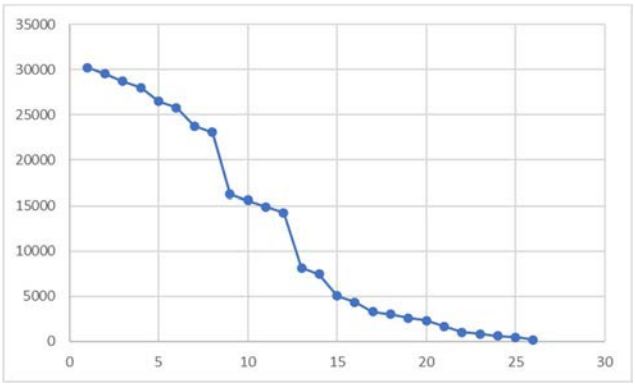
※ 横軸：スコアの順位、縦軸：スコアの値

分析を踏まえ、あるグループにおける遺伝子構造候補群を得た後に、付随するスコア s_{gene_cand} を参照し、最適解のスコア（つまり、最大の s_{gene_cand}）から一定の範囲内にあるもののみを準最適解として出力することとした。この範囲は、d_{gene} としてオプション指定できるようにした。図 15 は、あるグループについて s_{gene_cand} を降順にソートし、順位と値の大きさをプロットしたものである。s_{path_cand} の場合と異なり滑らかに下降していくことが多いが、例 4 のように下降が穏やかな場合も存在する。

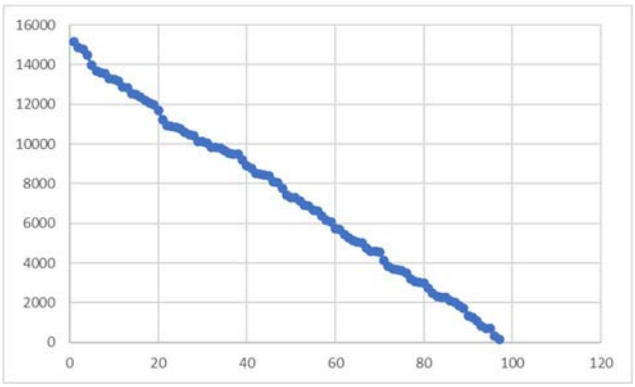
例 1



例 2



例 3



例 4

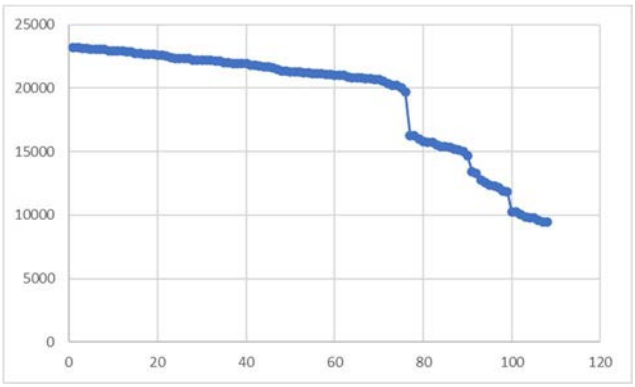


図 15　遺伝候補スコア s_{gene_cand} の大きさと順位の関係性

横軸：スコアの順位、縦軸：スコアの値

以下に、本機能を線虫ゲノムに対して適用して得られた結果の予測精度評価結果を示す。この評価には「2.3. ベンチマーク」で示したものと同じ線虫のデータを用いた。表 6 に最適解の評価と、準最適解の評価をまとめて示す。この評価では、d_{drop} は設定せず（つまり、全ての遺伝子構造候補を保持した上で）、d_{gene} を様々な値に振って感度と陽性的中率を調べた。表中にも示したとおり、これ以降、感度を“Sn”、陽性的中率を“Pr”と表記する。これらの精度指標の取得には GffCompare（「2.3.3. 評価ツール」を参照）を用いた。この評価では、各ゲノムコンソーシアムが維持する遺伝子構造アノテーションを正しいものとして参照した（表中の「参照データ」および「参照遺伝子構造」）。アイソフォームを含む参照データ（28,407 個）に対して GINGER で予測した最適解を照らし合わせると、予測遺伝子構造が 19,675 個であり、その中で参照遺伝子構造と一致したものが 12,658 個となり、遺伝子レベルにおける Sn が 44.6%、Pr が 64.3%となった。それに対して、d_{gene} を 10 に設定した場合には、予測遺伝子構造が 24,482 個、参照遺伝子構造と一致したものの数が 12,814 個、Sn が 45.1%、Pr が 52.3%であった。d_{gene} を 50 に設定した場合には、予測遺伝子構造が 31,765 個と参照遺伝子構造の数を大きく超えてしまった。図 16 に、

d_{gene} の変化に伴う Sn と Pr の関係性を示している。この結果から d_{gene} を大きくすると Sn は微増するが、Pr が大幅に低減することが見て取れる。

表 6　アイソフォーム候補を含む遺伝子構造予測

参照データ	GINGER の設定	参照遺伝子構造（#）	予測遺伝子構造（#）	参照と予測の完全一致（#）	塩基レベル		エクソンレベル		遺伝子レベル	
					Sn	Pr	Sn	Pr	Sn	Pr
アイソフォーム候補無し	最適解のみ	19,984	19,675	11,696	92.2	93.2	85.8	86.9	58.5	59.4
アイソフォーム候補あり	最適解のみ	28,407	19,675	12,658	91.8	93.6	81.9	88.2	44.6	64.3
	d_{gene}:10	28,407	24,482	12,814	91.9	93.5	82.2	87.5	45.1	52.3
	d_{gene}:50	28,407	31,765	13,126	92.0	93.4	82.6	85.9	46.2	41.3
	d_{gene}:100	28,407	37,670	13,354	92.2	93.2	82.9	84.4	47.0	35.4
	d_{gene}:500	28,407	73,533	13,832	92.6	92.4	83.8	78.3	48.7	18.8
	d_{gene}:1000	28,407	,108,166	14,054	92.9	91.9	84.3	75.0	49.5	13.0
	d_{gene}:5000	28,407	301,328	14,549	93.4	90.1	85.2	66.3	51.2	4.8
	d_{gene}:10000	28,407	490,564	14,706	93.6	89.3	85.5	63.3	51.8	3.0
	d_{gene}:50000	28,407	1,180,815	14,881	93.9	88.2	85.8	60.3	52.4	1.3
	d_{gene}:1000000	28,407	1,266,525	14,886	93.9	88.2	85.8	60.1	52.4	1.2
	閾値無し	28,407	1,270,869	14,887	93.9	88.2	85.8	60.1	52.4	1.2

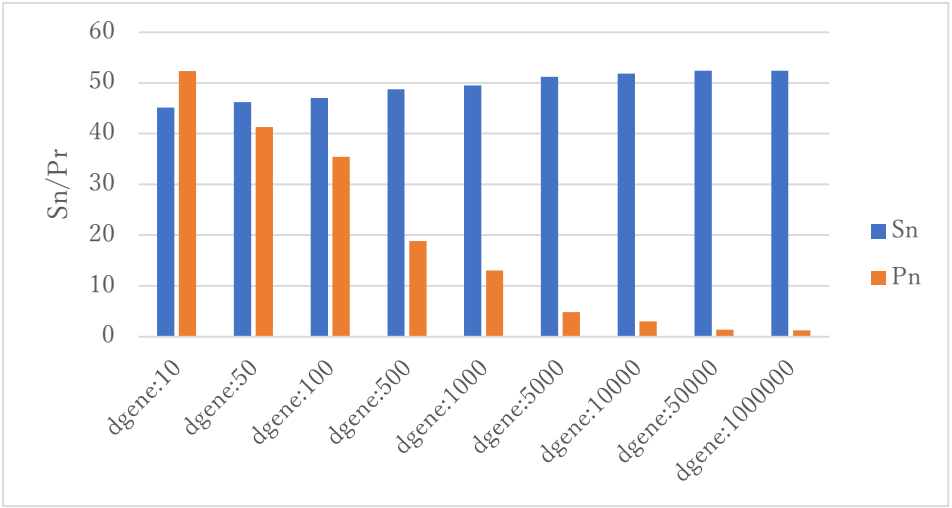


図 16　アイソフォーム候補を含む遺伝子構造予測での Sn および Pr

以上に示した遺伝子レベルでの評価では、予測遺伝子構造を構成する全エクソンの両端位置が完全に参照遺伝子構造と一致しているもののみを集計対象としたが、GffCompare により不完全なものも集計した結果を表 7 および図 17 に示す。これらの表・図にある“=”は完全一致した遺伝子、“c”は参照遺伝子構造と比較して余分なエクソンは持っておらずイントロンは一致した遺伝子、“k”は余分なエクソンを持っているがイントロンは一致しした遺伝子、“m”はイントロンの境界が一致しした（正解ではエクソン・イントロン・エクソンという構造になっている部分を一つのエクソンとして予測してしまっている場合を含む）遺伝子、“n”は一部のイントロン境界を予測できていないが予測できているイントロン境界が存在する遺伝子を意味する。d_{gene} を大きくしても完全一致する遺伝子構造数は微増するに過ぎないが、”c”, “k”など一致の条件を緩めると相対的に増加数が大きくなることが見て取れる。

表 7　アイソフォーム候補を含む遺伝子構造予測の評価結果

参照データ	GINGER の設定	コードが付与された予測	コード					合計 (=,c,k,m,n)
			=	c	k	m	n	
アイソフォーム候補無し	最適解のみ	17,686	11,696	980	408	138	42	13,260
アイソフォーム候補あり	最適解のみ	17,875	12,658	781	555	142	38	14,171
	d_{gene}:10	17,914	12,814	802	572	174	47	14,354
	d_{gene}:50	18,119	13,126	889	646	347	64	14,730
	d_{gene}:100	18,361	13,354	1,086	753	629	102	15,066
	d_{gene}:500	19,193	13,832	3,604	1,440	980	368	16,110
	d_{gene}:1000	19,736	14,054	5,367	1,801	1,116	542	16,711
	d_{gene}:5000	21,595	14,549	9,783	2,514	1,528	959	18,442
	d_{gene}:10000	22,375	14,706	11,562	2,668	1,648	1,142	19,131
	d_{gene}:50000	23,588	14,881	14,039	2,830	1,796	1,325	20,133
	d_{gene}:1000000	23,730	14,886	14,262	2,843	1,805	1,331	20,219
	閾値無し	23,748	14,887	14,291	2,844	1,806	1,331	20,227

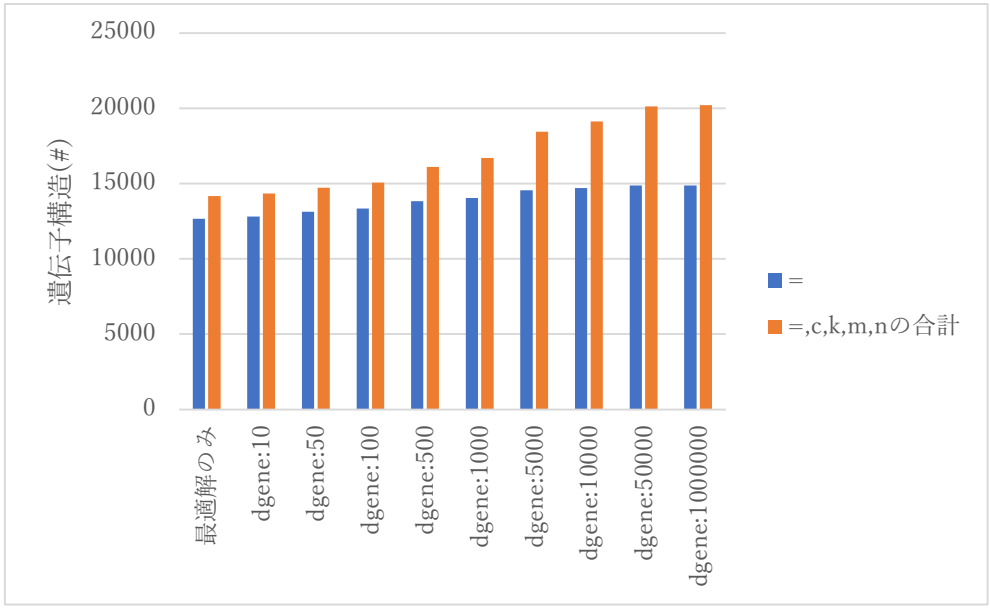


図 17　アイソフォーム候補を含む遺伝子構造予測での完全一致と不完全一致

2.3.　ベンチマーク

本節では、2.2. 節で開発内容について述べた GINGER の性能を確かめるため、既存の統合手法の中で多くの論文で使用されている MAKER および EVM との比較ベンチマークテストを行った結果を示す。ベンチマークには、信頼性の高い遺伝子情報が提供されているモデル生物である、線虫（*Caenorhabditis elegans*）、ハエ（*Drosophila melanogaster*）、イネ（*Oryza sativa*）、ゼブラフィッシュ（*Danio rerio*）、ヒト（*Homo Sapiens*）について全ゲノム配列からの遺伝子構造予測を行い、それぞれのゲノムコンソーシアムから提供されている遺伝子情報と比較することで、精度を検証した。

2.3.1.　使用ツールのバージョン

GINGER は Preparation phase において多数の既存ツールを用いる。ベンチマークのために使用した既存ツールのバージョンを表 8 にまとめる。

表 8　Preparation phase で使用した既存ツールおよびそのバージョン情報

既存ツール	バージョン
Augustus	3.3.2
CD-HIT	4.8.1
Gffread	0.11.7
GMAP	2018-07-04
HISAT2	2.2.0
Oases	0.2.09
SAMtools	1.5
SNAP	2017-03-01
Spaln	2.3.3
StringTie	1.3.4
TransDecoder	5.5.0
Trinity	2.8.4

表 9 にベンチマーク対象とした統合手法のバージョン情報をまとめる。

表 9　ベンチマーク対象とした統合手法とそのバージョン情報

統合手法	バージョン
GINGER	1.0.1
EVM	2.1.0
MAKER	3.01.04

2.3.2. ベンチマークテストに用いたデータ

本ベンチマークでは、線虫、ハエ、イネ[35]、ゼブラフィッシュ[36]、ヒトのゲノム配列および RNA-Seq データと、それらの近縁生物種が保有するタンパク質配列を用いた（表 10）。これらのデータは NCBI の FTP サイトから取得した。

表 10　ベンチマークに使用したデータ

生物種	バージョン 上段　：　ゲノム配列 中断　：　RNA-Seq データ 下段　：　タンパク質配列
線虫 Nematode (*Caenorhabditis elegans*) 100.3 Mb	GCF_000002985.6 in WBcel235
	SRR5849933, SRR5849934, SRR5849935, SRR5849936, SRR5849937, SRR5849938, SRR6266292
	GCF_000004555.2 in CB4 (*Caenorhabditis briggsae*) GCA_000180635.4 in EI_Paco_v._4 (*Pristionchus pacificus*), GCF_000001215.4 in Release_6_plus_ISO1_MT (*Drosophila melanogaster*), GCF_000001405.40 in GRCh38.p14 (*Homo sapiens*)
ハエ Fruit fly (*Drosophila melanogaster*) 143.7 Mb	GCF_000001215.4 in Release_6_plus_ISO1_MT
	SRR1573961, SRR3018852, SRR3018855, SRR3018862
	GCF_018153725.1 in ASM1815372v1 (*Drosophila mojavensis*), GCF_014905235.1 in Bmori_2016v1.0 (*Bombyx mori*), GCF_000004555.2 in CB4 (*Caenorhabditis briggsae*), GCF_000001405.40 in GRCh38.p14 (*Homo sapiens*)
イネ Rice (*Oryza sativa*) 374.4 Mb	GCF_001433935.1 in IRGSP-1.0
	SRX1030320, SRX4321411, SRX4321421, SRX4321425
	GCF_000231095.2 in ObraRS2 (*Oryza brachyantha*), GCF_002211085.1 in PHallii_v3.1 (*Panicum hallii*), GCF_902167145.1 in Zm-B73-REFERENCE-NAM-5.0 (*Zea mays*), GCF_000001405.40 in GRCh38.p14 (*Homo sapiens*)
ゼブラフィッシュ Zebrafish (*Danio rerio*) 1,373.5 Mb	GCF_000002035.6 in GRCz11
	SRX3632952, SRX4156978, SRX4156979
	GCF_018340385.1 in ASM1834038v1 (*Cyprinus carpio*), GCF_002234675.1 in ASM223467v1 (*Oryzias latipes*), GCF_901000725.2 in fTakRub1.2 (*Takifugu rubripes*), GCF_000001405.40 in GRCh38.p14 (*Homo sapiens*)
ヒト Human (*Homo sapiens*) 3,099.4 Mb	GCF_000001405.40 in GRCh38.p14
	SRR1693851, SRR3340902, SRR6333712, SRR7349910
	GCF_000001635.27 in GRCm39 (*Mus musculus*), GCF_002863925.1 in EquCab3.0 (*Equus caballus*), GCF_000002285.5 in Dog10K_Boxer_Tasha (*Canis_lupus_familiaris*), GCF_000003025.6 in Sscrofa11.1 (*Sus scrofa*)

2.3.3. 評価ツール

本ベンチマークでは、遺伝子構造を比較するツールである GffCompare[37] を使用して精度比較を行った。用いた指標としては、Sn（感度）と Pr（陽性的中率）およびこれらから計算される F-score（$2\,(Sn \cdot Pr)/(Sn + Pr)$）である。

2.3.4. 統合手法間の精度比較

近年のゲノム研究では、MAKER と EVM が広く使用されていることからこの 2 つのツールをベンチマークの対象とした（以下、統合手法が実装されたツールを「統合ツール」と呼ぶ。）。統合ツールとしての公正な評価のために、GINGER の Preparation phase から得たデータを MAKER および EVM の入力データとしても用いた。具体的には、MAKER の入力データは、Trinity の出力結果（RNA-Seq データのアセンブリング結果）、Augustus および SNAP によって学習したそれぞれの統計モデル、Homology-based method で使用したものと同じタンパク質配列で構成される。また、EVM の入力データは、Genome-guided assembly-based method、*de novo* assembly-based method、*ab initio*-based method（Augustus および SNAP）、そして Homology-based method の出力結果で構成される。GINGER に関しては、設定可能な実行パラメータは全てデフォルトのものとした（重み係数である w_{exon} はデフォルト、s_{gene} に対する閾値は自動決定とした）。EVM に関しては、事前検討の結果に基づき、GINGER の重みを参考にした設定（Genome-guided assembly-based method：10.0、*de novo* assembly-based method：4.0、Augustus：9.0、SNAP：4.0、Homology-based method：8.0）を採用した（事前検討については「付録 A. EVM の重み係数」を参照）。

このベンチマークでは、各ゲノムコンソーシアムが維持する遺伝子構造アノテーションを正しいものとして参照し（以下、これらのデータを「参照遺伝子構造」と呼ぶ。）、塩基レベル、エクソンレベル、遺伝子レベルでの予測精度を評価した。アイソフォームが存在する遺伝子については、最も長いアイソフォームを正解として使用した。評価指標としては上述のとおり GffCompare の使用で求められる Sn、Pr 並びに F-score を用いた。

表 11 に 3 つの統合ツールに対しての評価結果をまとめて示す。Sn、Pr ともに塩基レベル、エクソンレベル、遺伝子レベルの順に値が減少していることが確認できる。この傾向は、ゲノムサイズが大きくなるにしたがって顕著になる。「2.3.5. GINGER における個別手法間の精度比較」には Preparation phase における個別手法単独での精度を示しているが、それらと比較しても、EVM と GINGER は、全ての生物種においてより高い F-score を達成しており統合の有効性を示している。また、GINGER はエクソンレベルおよび遺伝子レベルにおいて全ての精度指標で他ツールより優れた結果を示した（塩基レベルでは、イネおよびヒトの Sn が EVM より低い値となった）。特に Pr と F-score はエクソンレベルや遺伝子レベルで他のツールとの差が大きく、GINGER の陽性的中率が他に比べ相対的に高いことが明らかとなった。この差はゲノムサイズが大きくなるにつれて拡大することも確認で

きた。例えば、線虫で GINGER に次いで 2 番目に優れた性能を示す EVM との遺伝子レベルでの F-score の差は約 6 ポイントであったが、ゼブラフィッシュとヒトでは 14〜17 ポイントと広がっていることが確認できた。さらに、GINGER は遺伝子レベルでも非常に高い Sn を有していた。重み係数の決定には、線虫、イネ、ゼブラフィッシュを用いているが、ハエとヒトに関しても遺伝子構造全体および完全なエクソンの予測において優秀な成績を収めていることが確認できた。

表 11　統合ツールの評価結果

	参照遺伝子構造(#)	予測遺伝子構造(#)	塩基レベル			エクソンレベル			遺伝子レベル		
			Sn	Pr	F-score	Sn	Pr	F-score	Sn	Pr	F-score
線虫 (*Caenorhabditis elegans*) 100.3Mb											
GINGER		19,675	92.2	93.2	92.7	85.8	86.9	86.3	58.5	59.4	58.9
EVM	19,984	17,694	88.8	92.2	90.5	81.6	84.2	82.9	49.6	56.1	52.7
MAKER		18,056	80.9	89.6	85.0	62.1	54.4	58.0	24.3	26.9	25.5
ハエ (*Drosophila melanogaster*) 143.7 Mb											
GINGER		12,347	90.2	98.1	94.0	84.5	91.4	87.8	64.6	73.1	68.6
EVM	13,963	13,042	88.1	94.5	91.2	81.1	83.3	82.2	58.9	63.1	60.9
MAKER		12,640	84.8	95.9	90.0	69.9	72.3	71.1	40.0	44.2	42.0
イネ (*Oryza sativa*) 374.4 Mb											
GINGER		28,021	87.8	93.2	90.4	86.9	89.0	87.9	66.2	67.8	67.0
EVM	28,721	38,168	92.5	83.1	87.5	87.2	77.6	82.1	64.7	48.7	55.6
MAKER		33,487	81.2	81.9	81.5	69.0	59.4	63.8	21.8	18.7	20.1
ゼブラフィッシュ (*Danio rerio*) 1,373.5 Mb											
GINGER		25,579	90.4	93.0	91.7	90.3	90.3	90.3	58.4	58.3	58.3
EVM	25,536	33,239	89.4	84.6	86.9	88.6	79.1	83.6	51.0	39.2	44.3
MAKER		30,967	72.5	84.2	77.9	54.2	58.8	56.4	11.7	9.6	10.5
ヒト (*Homo sapiens*) 3,099.4 Mb											
GINGER		20,811	84.8	94.5	89.4	83.2	91.1	87.0	45.3	47.2	46.2
EVM	21,658	36,275	88.7	77.7	82.8	85.4	77.1	81.0	39.6	23.6	29.6
MAKER											

各統合ツールの傾向をより詳細に把握するために、ゼブラフィッシュのベンチマーク結果について各統合ツールによる予測遺伝子構造と参照遺伝子構造との重複を、塩基レベル、エクソンレベル、遺伝子レベルで比較した(表 12)。塩基レベルとエクソンレベルで

は、どの統合ツールでも高い割合で正しい予測を得た（塩基レベルで 67.0%、エクソンレベルで 50.3%)のに対し、遺伝子レベルでは非常に悪い結果となった。どの統合ツールによっても正解を予測できなかった参照遺伝子構造の数は、塩基レベルとエクソンレベルでは非常に少なかったのに対して、遺伝子レベルでは 36.4%にまで達していた。中でも GINGER と EVM が正解するケースは比較的重なっており、塩基レベルとエクソンレベルでは両者ともに正解とするケースが 85%を超えた。しかしながら遺伝子レベルでは両者ともに正解となるケースは 46.9%にまで落ち込んだ。さらに GINGER のみが正しく遺伝子レベルで予測したケースが 10.1%であるのに対して EVM のみが正しく予測したケースは 3.7%のみであり、遺伝子レベルでは両者の予測傾向が異なることが見て取れた。

GINGER のみで正しく予測できた遺伝子を詳細に解析したところ、他の統合ツールでは最初のエクソンが正しく予測されない場合が多いことが判明した。GINGER のみで予測できた 2,590 個の遺伝子のうち、EVM は 1,682 個について最初のエクソンを正しく予測できていなかった（MAKER では 1,871 個が最初のエクソンを間違っていた)。図 18〜図 20 に、GINGER は正しく予測できていたのに対して EVM および MAKER では予測できていなかった遺伝子の事例を示す。正しくは別々の複数の遺伝子として予測すべきところを誤って繋げてしまった事例や余分なエクソンを含んで予測した事例であり、GINGER 固有の機能である、入力となるエクソン群を適切にグループ化することの重要性を確認できた。

表 12　統合ツールによる予測結果と正解との関係性

統合ツール			塩基レベル			エクソンレベル			遺伝子レベル		
GINGER	EVM	MAKER	塩基数 (#)	(%)		エクソン数 (#)	(%)		遺伝子構造数 (#)	(%)	
■	■	■	29,803,416	67.0%	*	123,516	50.3%	*	2,235	8.8%	
■	■		8,869,898	19.9%		88,082	35.9%	*	9,732	38.1%	*
■		■	1,037,428	2.3%		4,444	1.8%		374	1.5%	
	■	■	748,740	1.7%		2,244	0.9%		115	0.5%	
■			720,635	1.6%		5,916	2.4%		2,590	10.1%	*
	■		358,221	0.8%		3,232	1.3%		948	3.7%	
		■	661,240	1.5%		2,507	1.0%		256	1.0%	
			2,274,382	5.1%		15,502	6.3%		9,286	36.4%	*
		Total	44,473,960	100.0%		245,443	100.0%		25,536	100.0%	

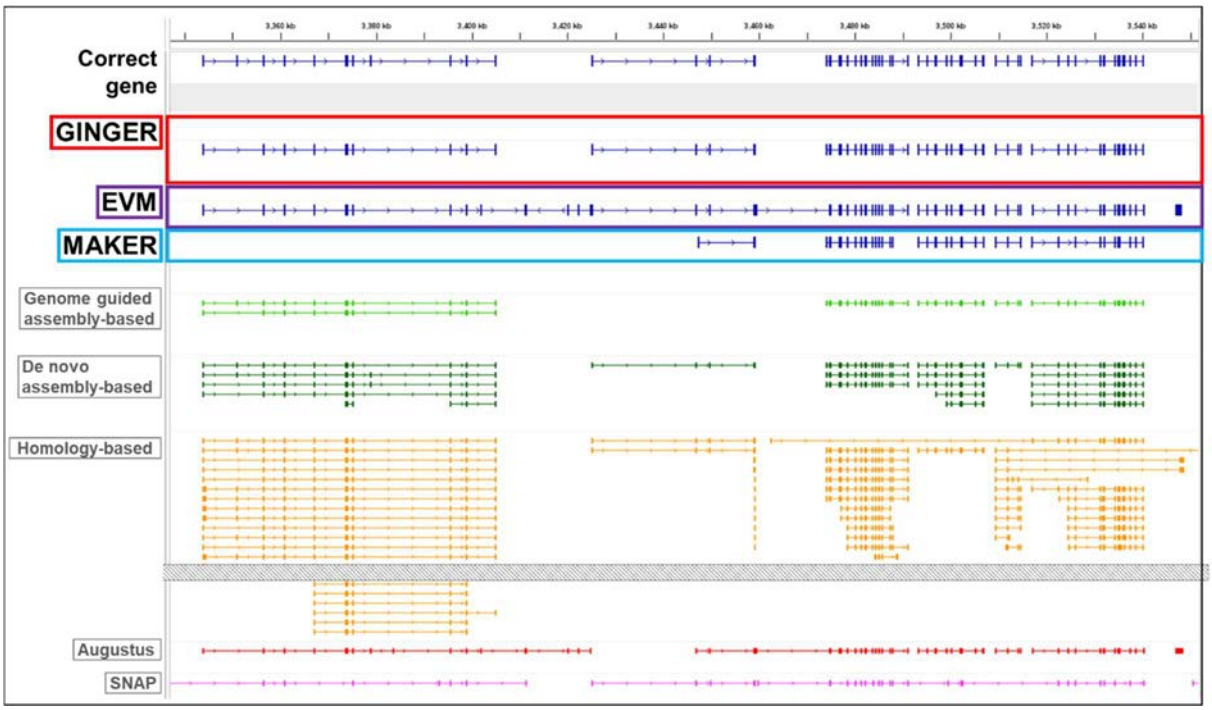


図 18　GINGER のみが正解した例（1）

※ 可視化には IGV[38] を使用

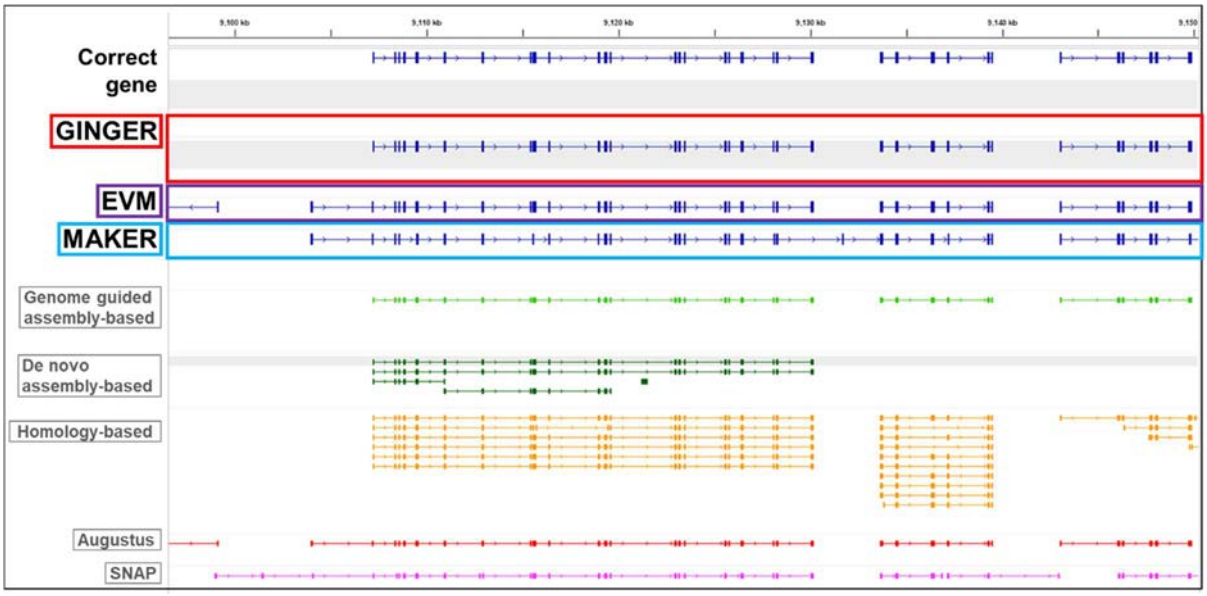


図 19　GINGER のみが正解した例（2）

※ 可視化には IGV[38] を使用

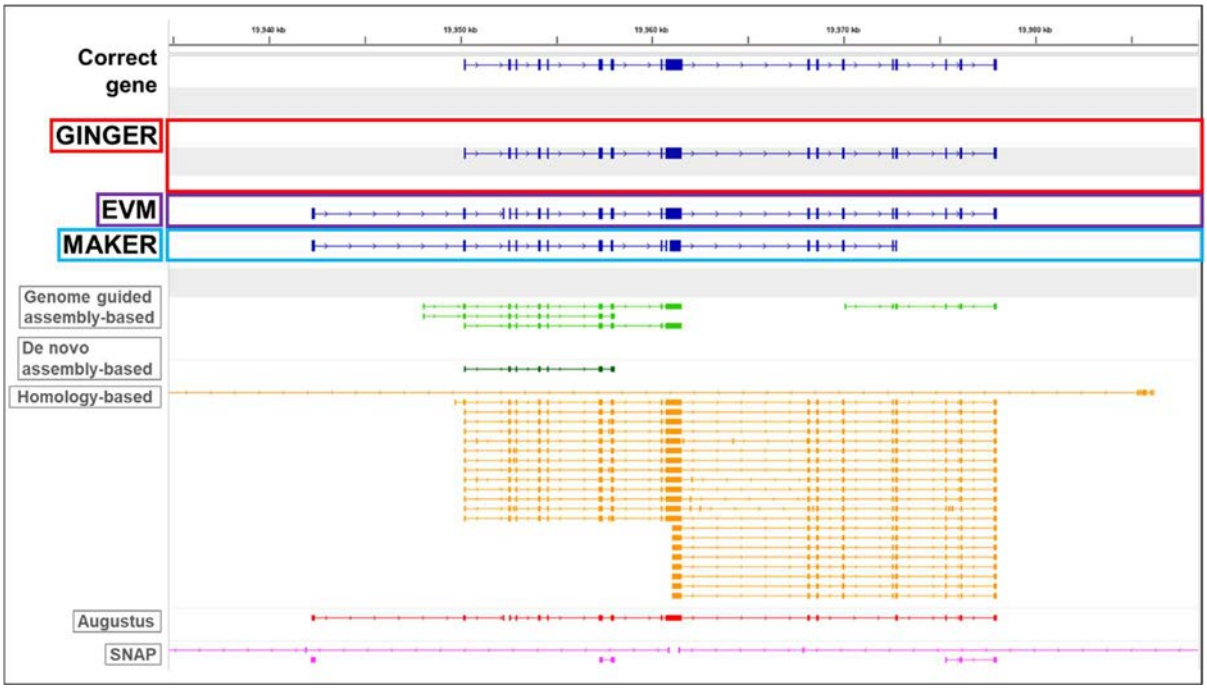


図 20　GINGER のみが正解した例（3）

※ 可視化には IGV[38] を使用

2.3.5. GINGER における個別手法間の精度比較

Preparation phase の個別手法による予測結果を表 13 および図 21～図 25 に示す。基本的に、Sn と Pr の両方において GINGER による統合結果と比べて、塩基レベル、エクソンレベルおよび遺伝子レベルのいずれにおいても低いことが見て取れる。それらの中では、*ab initio*-based method の Augustus および Genome-guided assembly-based method が、どの生物種の場合であっても最も高い F-score を得る結果となったが、統合後の傾向と同様に、F-score はゲノムサイズの増加に伴って減少していた。この傾向は Augustus の予測結果に対しての遺伝子レベルでの評価において顕著に認められた。エクソンレベルではゲノムサイズが増加することによって F-score が約 10 ポイント低下するのみであったが、遺伝子レベルでは最大 40 ポイント近く低下していた。また、*ab initio*-based method は、どの場合においても Sn と Pr のバランスがとれているが、*de novo* assembly-based method や Homology-based method では、Sn に対して Pr が大幅に低い場合が認められた。*de novo* assembly-based method では、RNA-Seq データをミスアセンブルしたコンティグ配列が含まれていると、それらをゲノム配列に対してアライメントしても正しい遺伝子構造を得られないため Pr が下がることになると考えられる。Homology-based method では、データセット中に多数のリモートホモログを含んでしまうと正しいアライメントが得られずに Pr が下がることになると考えられる。これらのことから、入力データの吟味が重要であることが改めて確認できた。

表 13　Preparation phase の個別手法に対する評価結果

	参照遺伝子構造(#)	正解遺伝子構造(#)	塩基レベル			エクソンレベル			遺伝子レベル		
			Sn	Pr	F-score	Sn	Pr	F-score	Sn	Pr	F-score
線虫 (*Caenorhabditis elegans*) 100.3 Mb											
Genome-guided assembly		22,509	78.35	62.16	69.32	70.23	57.98	63.52	33.88	30.08	31.87
de novo assembly		31,703	53.20	30.92	39.11	37.37	26.61	31.09	13.03	8.21	10.07
ab initio (Augustus)	19,984	17,154	87.93	91.42	89.64	77.32	81.71	79.46	37.53	43.72	40.39
ab initio (SNAP)		25,769	79.72	81.22	80.46	61.53	50.65	55.57	13.09	10.15	11.44
Homology		70,706	69.49	18.63	29.38	60.07	14.64	23.54	31.74	8.97	13.99
ハエ (*Drosophila melanogaster*) 143.7 Mb											
Genome-guided assembly		20,962	77.68	54.79	64.26	72.47	49.22	58.63	50.38	33.53	40.26
de novo assembly		48,232	75.38	20.47	32.20	56.75	16.78	25.90	36.09	10.44	16.19
ab initio (Augustus)	13,950	13,449	90.82	94.58	92.66	76.68	78.38	77.52	51.18	53.09	52.12
ab initio (SNAP)		15,359	78.37	86.23	82.11	53.31	42.80	47.48	18.08	16.42	17.21
Homology		96,975	83.72	11.25	19.83	73.50	5.75	10.67	43.87	6.31	11.03
イネ (*Oryza sativa*) 374.4 Mb											
Genome-guided assembly		50,579	76.76	56.21	64.90	78.99	53.22	63.59	47.62	26.88	34.37
de novo assembly		113,258	73.14	23.93	36.06	57.84	18.92	28.51	30.28	7.64	12.20
ab initio (Augustus)	28,556	40,532	90.09	78.47	83.88	74.87	66.45	70.41	41.41	29.17	34.23
ab initio (SNAP)		55,269	48.31	54.54	51.24	27.46	16.17	20.35	0.44	0.23	0.30
Homology		147,671	84.77	16.85	28.12	83.13	11.49	20.19	58.97	11.40	19.11
ゼブラフィッシュ (*Danio rerio*) 1,373.5 Mb											
Genome-guided assembly		42,422	83.72	53.79	65.50	84.04	52.90	64.93	53.11	31.96	39.90
de novo assembly		114,035	79.40	22.48	35.04	76.53	21.01	32.96	44.57	9.98	16.30
ab initio (Augustus)	25,536	38,987	86.28	78.00	81.93	77.17	69.25	73.00	21.60	14.14	17.09
ab initio (SNAP)		50,230	39.68	57.96	47.11	18.46	17.90	18.18	0.83	0.42	0.56
Homology		358,530	89.34	5.87	11.01	87.17	4.32	8.23	52.39	3.73	6.96
ヒト (*Homo sapiens*) 3,099.4 Mb											
Genome-guided assembly		67,749	67.94	36.44	47.43	67.88	38.03	48.74	31.69	10.14	15.37
de novo assembly		193,511	70.89	14.54	24.13	67.44	14.05	23.26	23.76	2.66	4.79
ab initio (Augustus)	21,670	33,031	81.13	73.47	77.11	68.32	64.86	66.55	15.31	10.04	12.13
ab initio (SNAP)		91,453	39.11	19.53	26.05	14.35	5.41	7.86	1.03	0.24	0.39
Homology		317,873	90.28	5.61	10.57	88.20	4.84	9.18	55.79	3.80	7.12

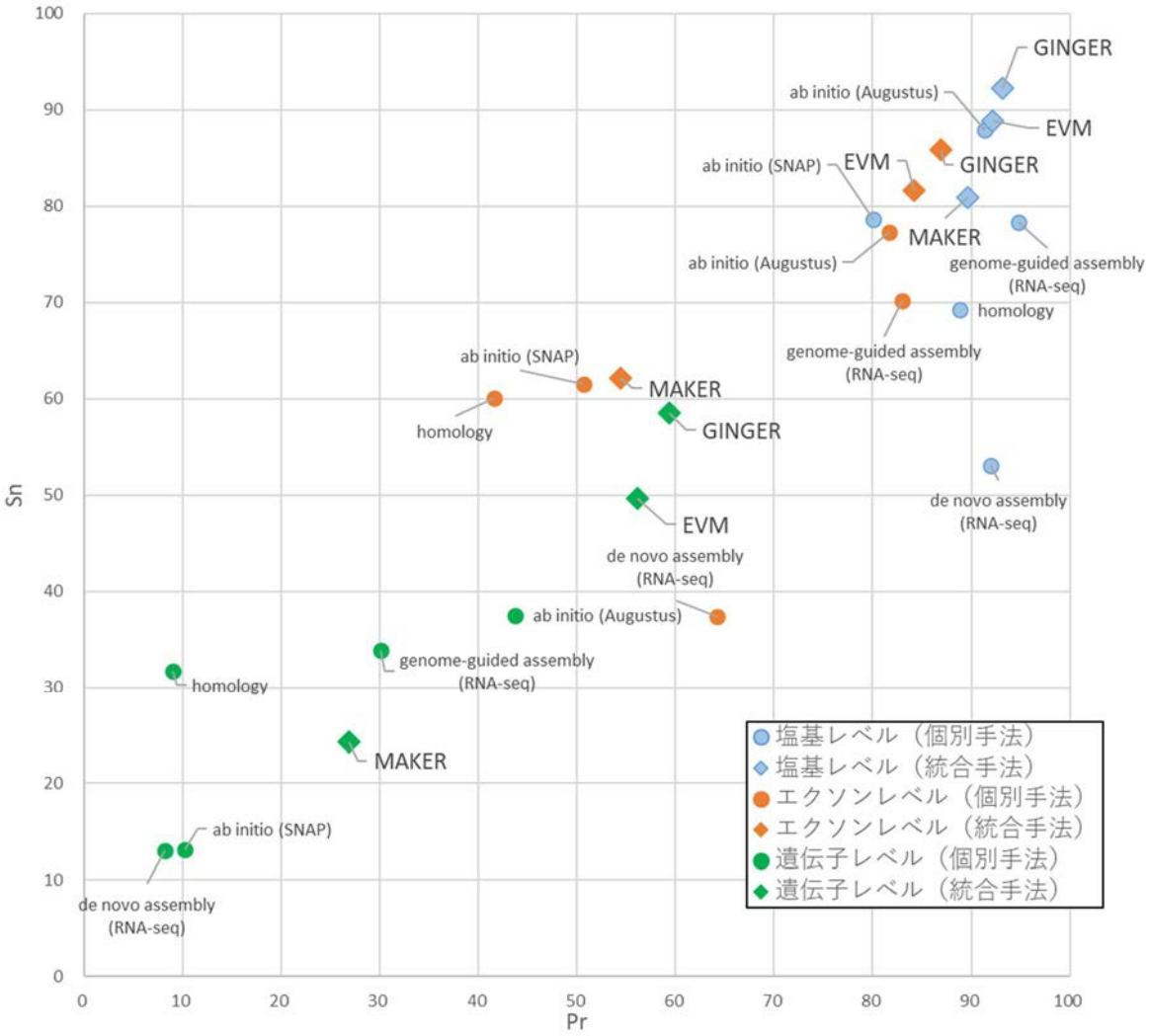


図 21　線虫ゲノムに対しての予測精度の比較

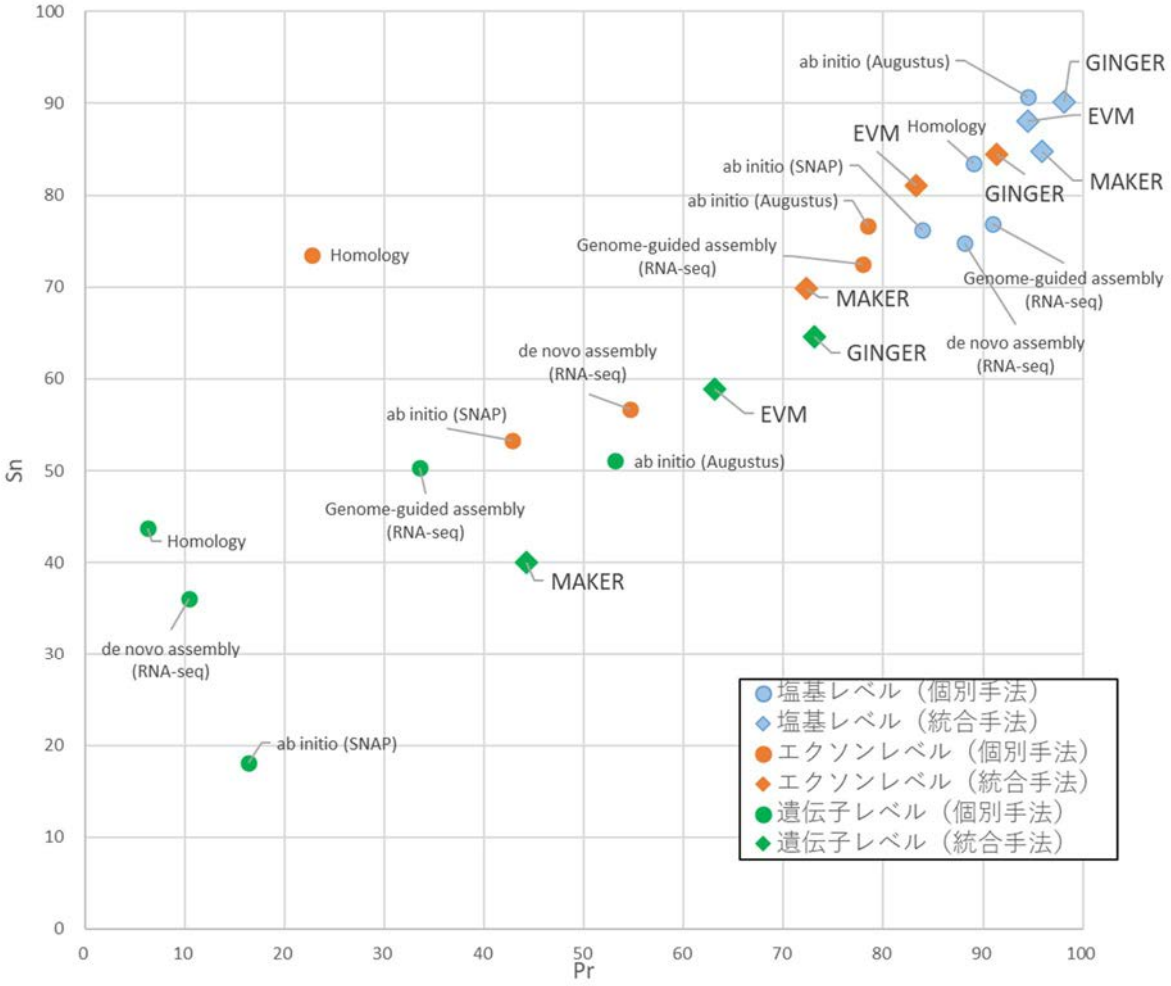


図 22　ハエゲノムに対しての予測精度の比較

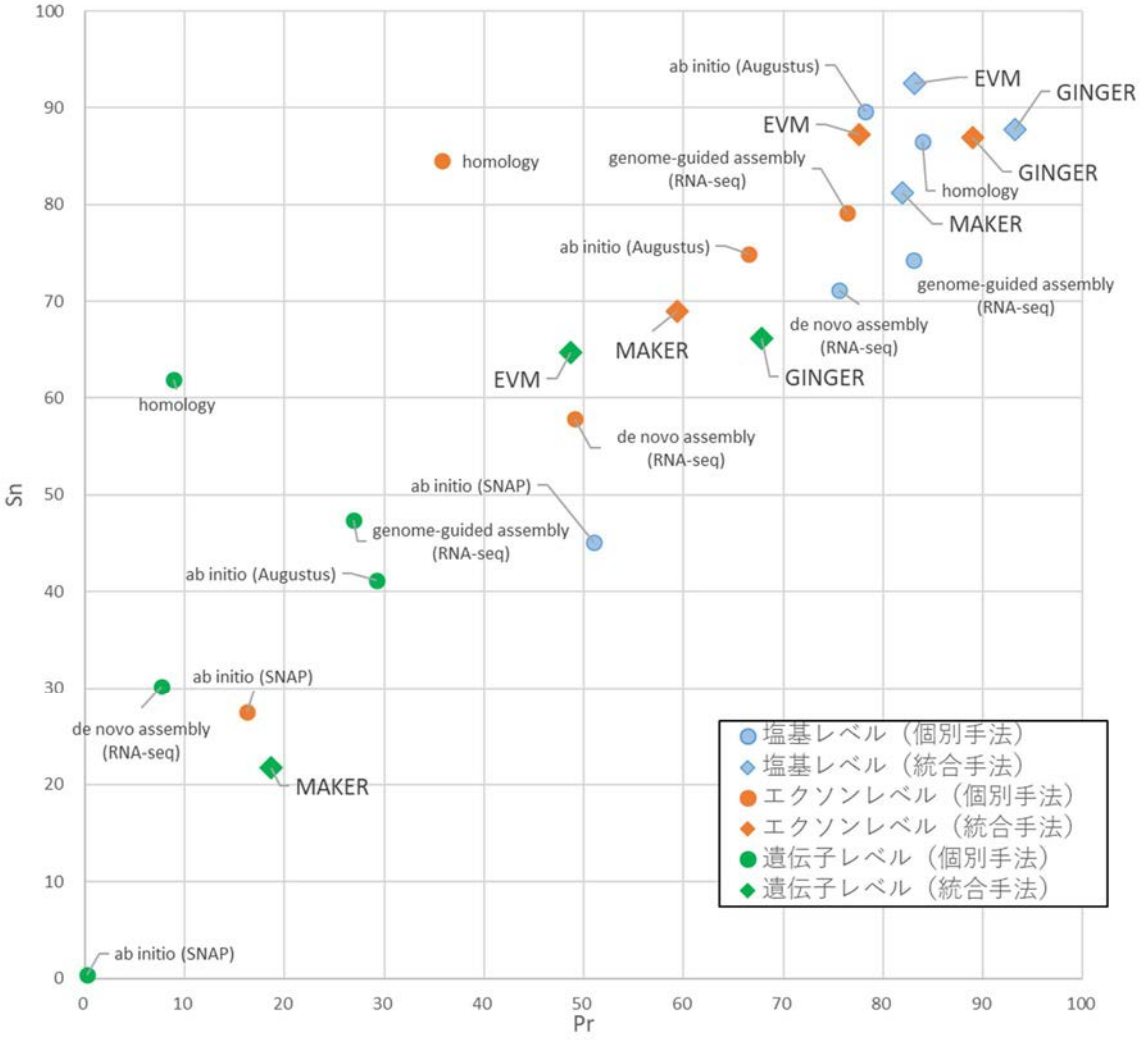


図 23　イネゲノムに対しての予測精度の比較

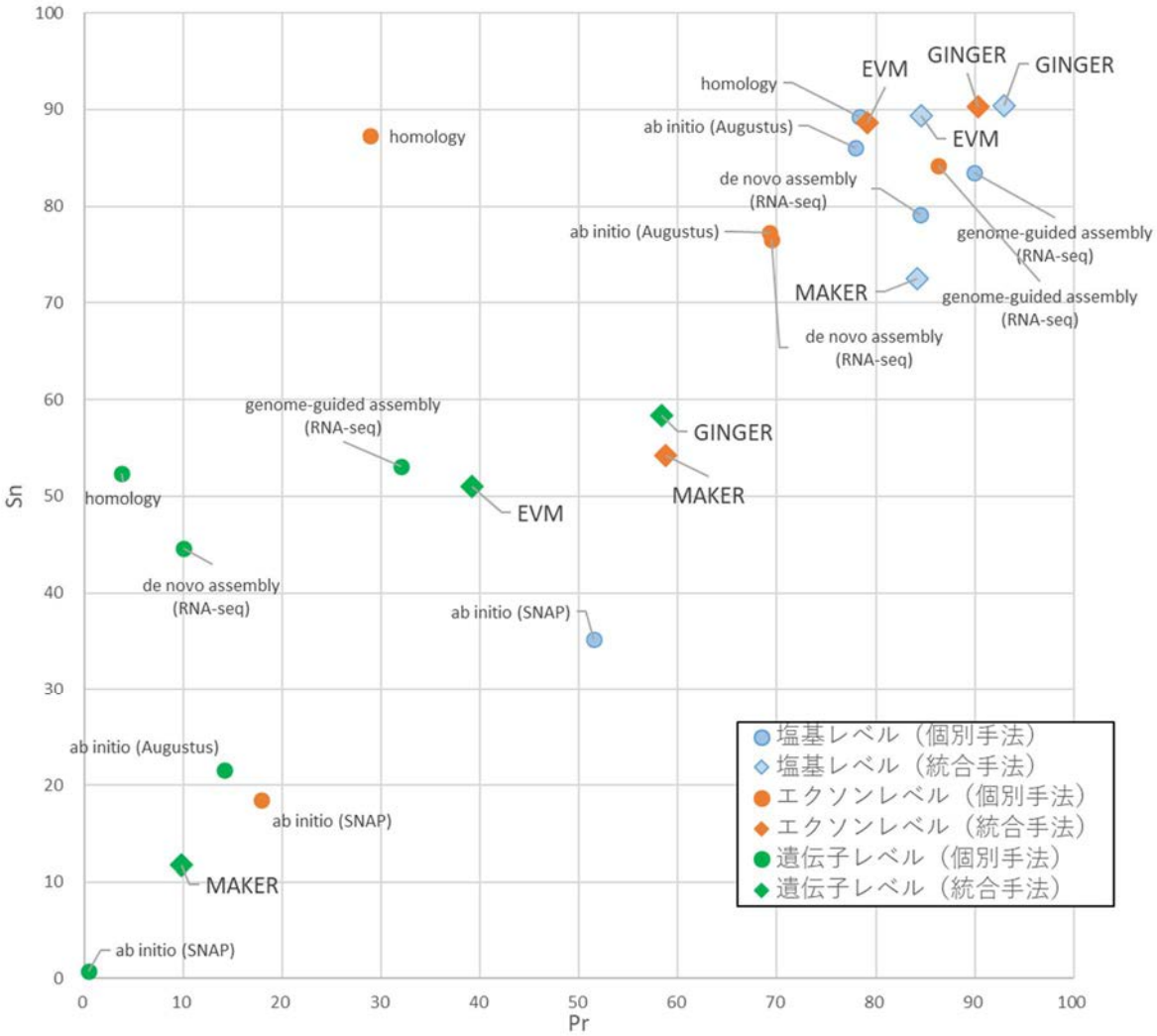


図 24　ゼブラフィッシュゲノムに対しての予測精度の比較

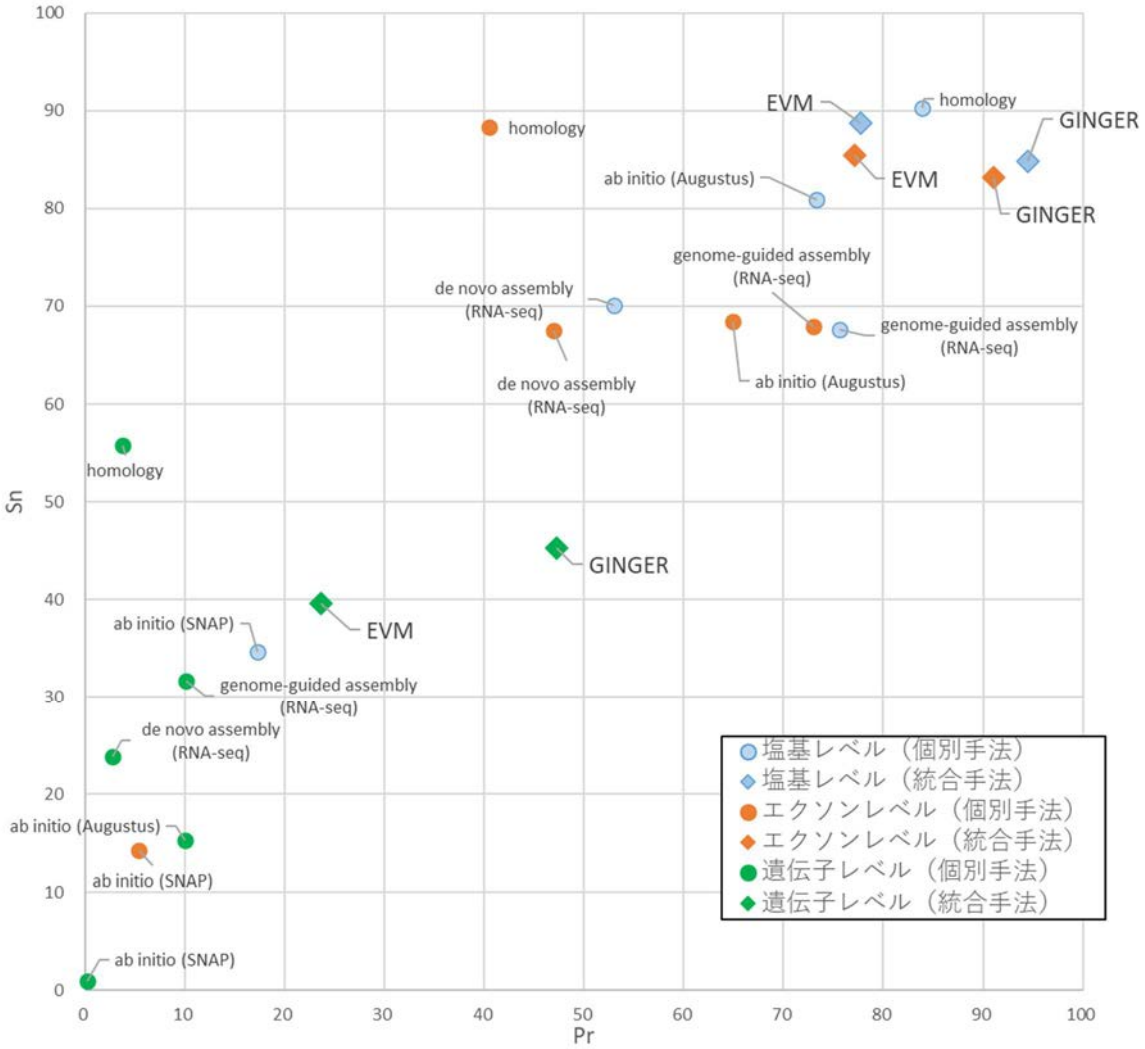


図 25　ヒトゲノムに対しての予測精度の比較

2.3.6.　エクソンポテンシャルの効果の検討

GINGER の予測精度が他の統合手法に勝る理由の分析の一環として、GINGER 独自の機能である Preparation phase におけるエクソンポテンシャルが及ぼす影響を検討した。具体的には、GINGER を用いて通常どおりにエクソンポテンシャルを取得して遺伝子構造予測までを行った場合と、エクソンポテンシャルを 1.0 に固定した場合の予測精度を比較した。結果を以下の表 14 に示す。通常の GINGER と比較したところ、全ての評価指標において値が低下していることが確認できた（エクソンレベルの F-score で 2.2～3.6 ポイント、遺伝子レベルの F-score で 3.2～8.5 ポイント）。この結果より、Preparation phase の個別手法におけるエクソンポテンシャルの定式化が、GINGER の性能に寄与していることが示唆される。

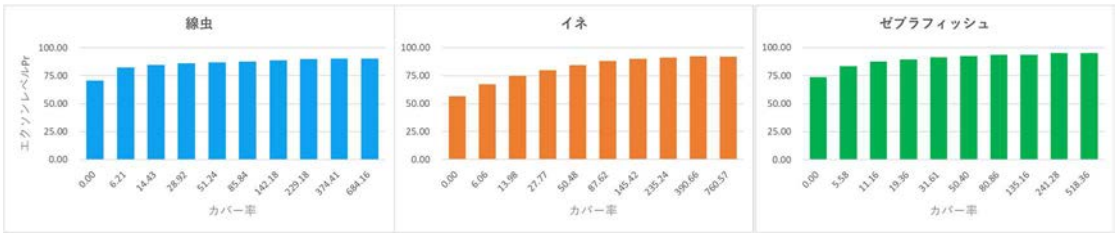
表 14　エクソンポテンシャルを 1.0 に固定した場合の予測精度

	参照遺伝子構造(#)	予測遺伝子構造(#)	塩基レベル			エクソンレベル			遺伝子レベル		
			Sn	Pr	F-score	Sn	Pr	F-score	Sn	Pr	F-score
線虫 (*Caenorhabditis elegans*) 100.3Mb *											
GINGER (default)	19,984	19,675	92.2	93.2	92.7	85.8	86.9	86.3	58.5	59.4	58.9
GINGER (エクソンポテンシャル = 1.0)		17,849	89.9	92.7	91.3	82.1	83.4	82.7	47.7	53.4	50.4
ハエ (*Drosophila melanogaster*) 143.7 Mb											
GINGER (default)	13,963	12,347	90.2	98.1	94.0	84.5	91.4	87.8	64.6	73.1	68.6
GINGER (エクソンポテンシャル = 1.0)		12,710	91.4	96.6	93.9	84.3	86.9	85.6	61.7	67.8	64.6
イネ (*Oryza sativa*) 374.4 Mb *											
GINGER (default)	28,721	28,021	87.8	93.2	90.4	86.9	89.0	87.9	66.2	67.8	67.0
GINGER (エクソンポテンシャル = 1.0)		29,220	88.3	91.1	89.7	86.0	83.9	84.9	62.2	61.1	61.6
ゼブラフィッシュ (*Danio rerio*) 1,373.5 Mb *											
GINGER (default)	25,536	25,579	90.4	93.0	91.7	90.3	90.3	90.3	58.4	58.3	58.3
GINGER (エクソンポテンシャル = 1.0)		26,094	90.3	91.6	90.9	89.8	86.5	88.1	55.7	54.6	55.1
ヒト (*Homo sapiens*) 3,099.4 Mb											
GINGER (default)	21,658	20,811	84.8	94.5	89.4	83.2	91.1	87.0	45.3	47.2	46.2
GINGER (エクソンポテンシャル = 1.0)		23,226	85.6	90.0	87.7	83.0	84.6	83.8	42.4	39.5	40.9

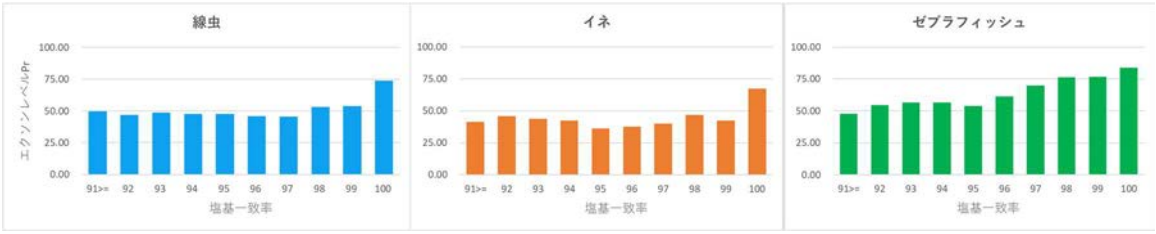
図 26 に、各個別手法で予測されたエクソンについて、エクソンポテンシャルの計算に用いられた統計指標とエクソンの Pr との関係性を示す。Augustus、Homology-based method では、指標と Pr の間に明確な相関が認められたのに対し、RNA-Seq-based method では二つの手法双方において、明確な相関は認められなかった。また、各生物種のベンチマークテストで得られたエクソンポテンシャルおよびエクソンスコアとエクソンの Sn および Pr との関係性を図 27～図 31 に示す。A と B のグラフは、エクソンポテンシャルと Sn および Pr の関係を示し、C と D はエクソンポテンシャルに重み係数を適用して算出したエクソンスコアと Sn および Pr との関係を示す。重み係数を掛け合わせているだけであるため、基本的な傾向は変わらないが、正しいエクソンと、エクソンポテンシャ

ルとの間に高い相関関係を見いだすことができるツールに大きな重み係数がかかっており、この重み係数が Sn と Pr を向上させるために正しく機能していることがわかる。一方で SNAP および RNA-Seq-based method には低い係数が適用されており、これは上述のとおり正しいエクソンと、エクソンポテンシャルを決定する指標との間の相関が低いこと、言い換えると、これらの手法ではエクソンポテンシャルに基づく正解と不正解の判別性能が高くはないことに起因している。

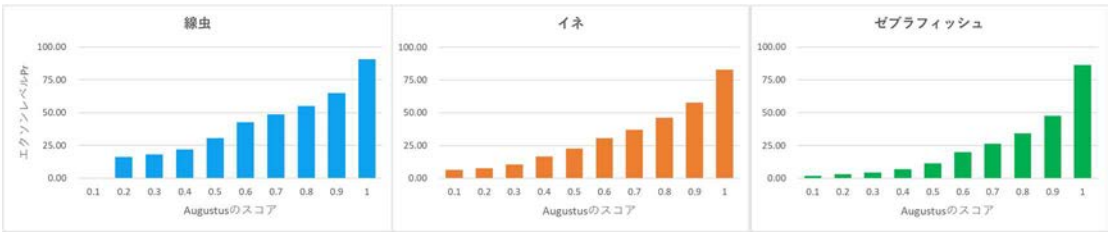
Genome-guided assembly-based method（エクソンに対するカバー率）



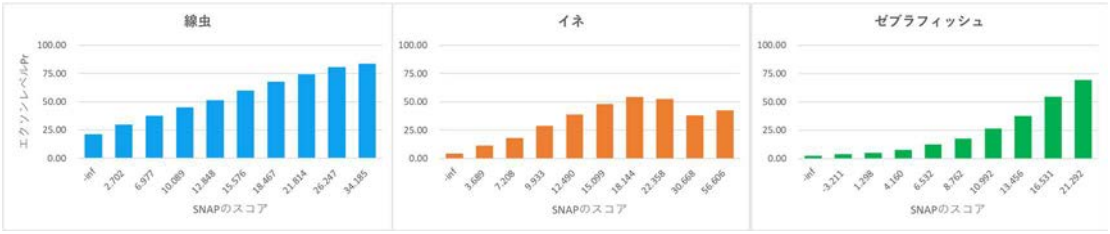
de novo assembly-based method（エクソン毎の塩基一致率）



ab initio-based method（Augustus が遺伝子構造に付与するスコア）



ab initio-based method（SNAP の出力に基づいて遺伝子構造に付与したスコア）



Homology-based method（エクソン毎のアミノ酸一致率）

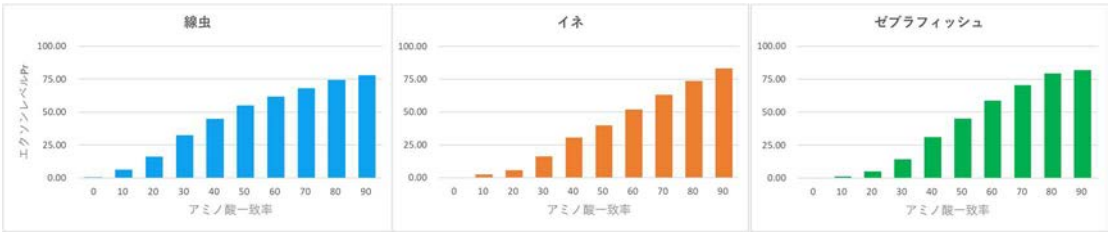


図 26　個別ツールから得られるエクソンに関する統計指標とエクソンの Pr との関係性

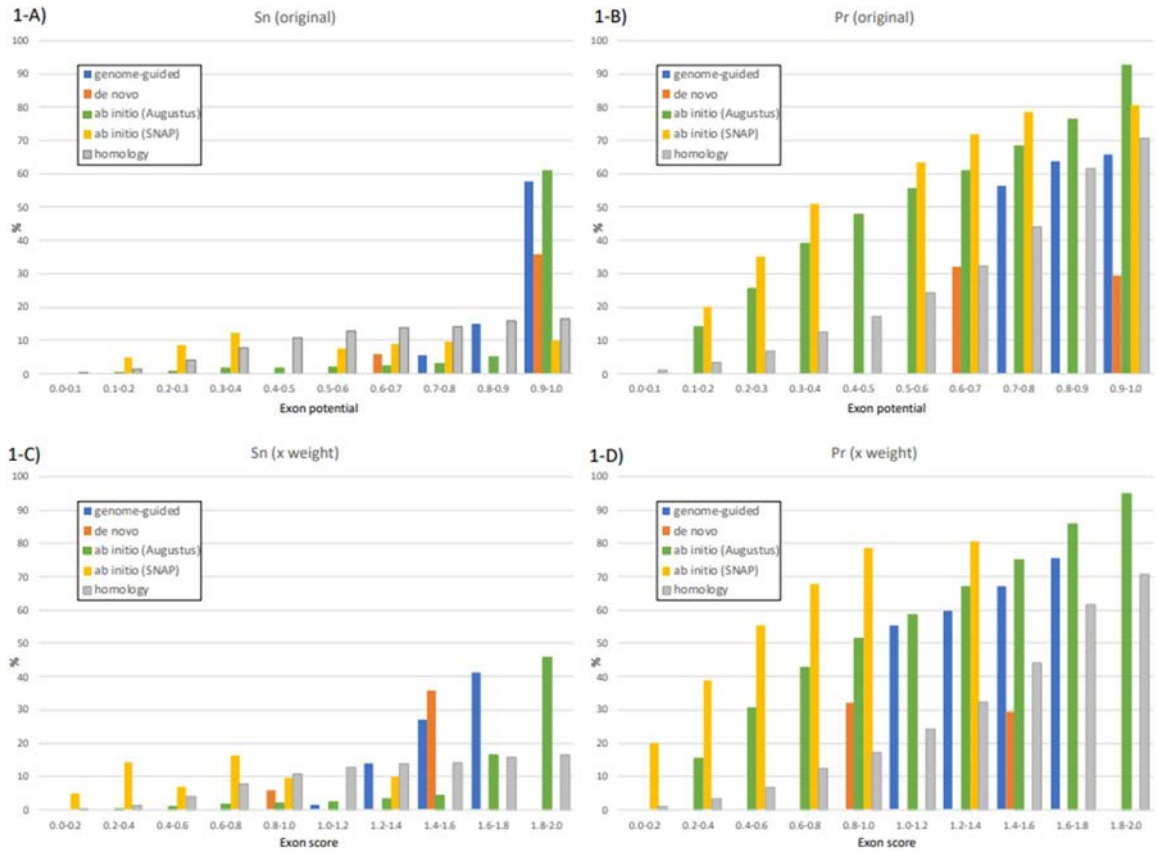


図 27　エクソンポテンシャル・エクソンスコアとエクソンの予測精度（線虫）

1-A)エクソンポテンシャルと Sn　1-B)エクソンポテンシャルと Pr

1-C)エクソンスコアと Sn　1-D)エクソンスコアと Pr

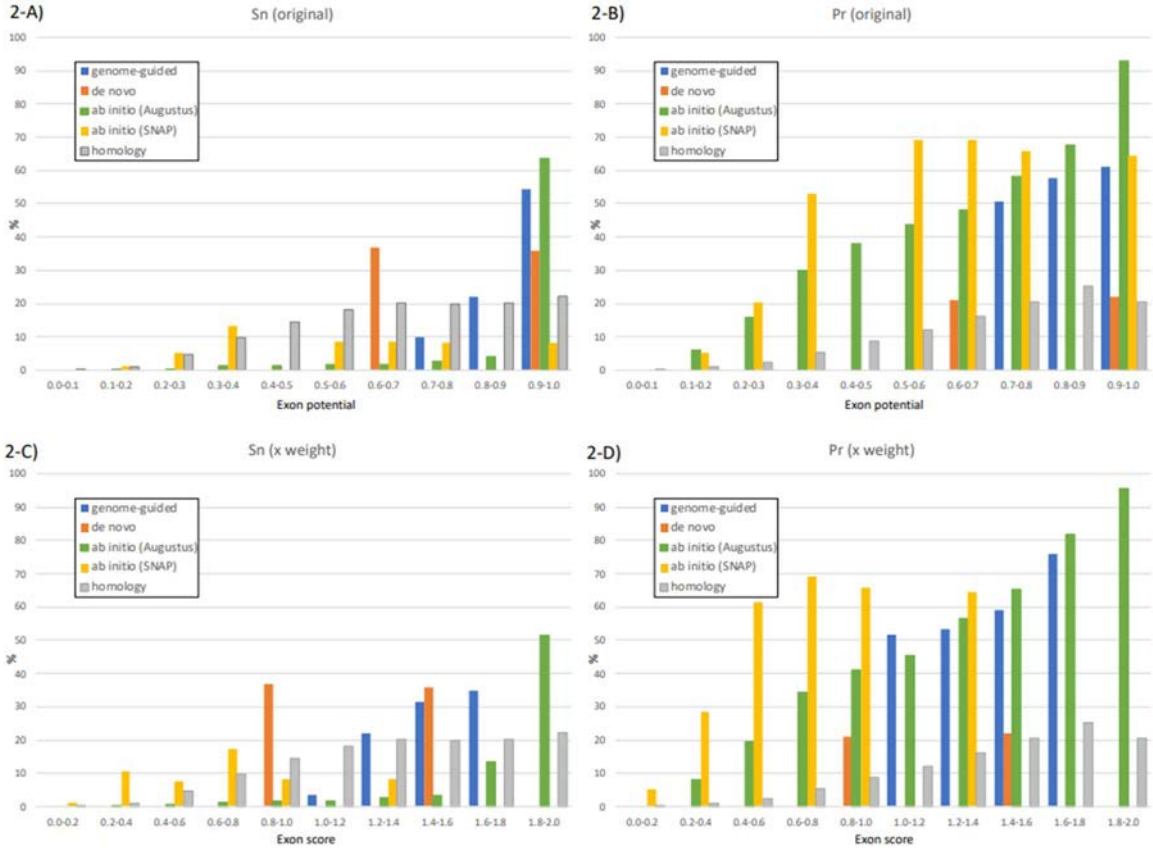


図 28　エクソンポテンシャル・エクソンスコアとエクソンの予測精度（ハエ）

2-A)エクソンポテンシャルと Sn　2-B)エクソンポテンシャルと Pr

2-C)エクソンスコアと Sn　2-D)エクソンスコアと Pr

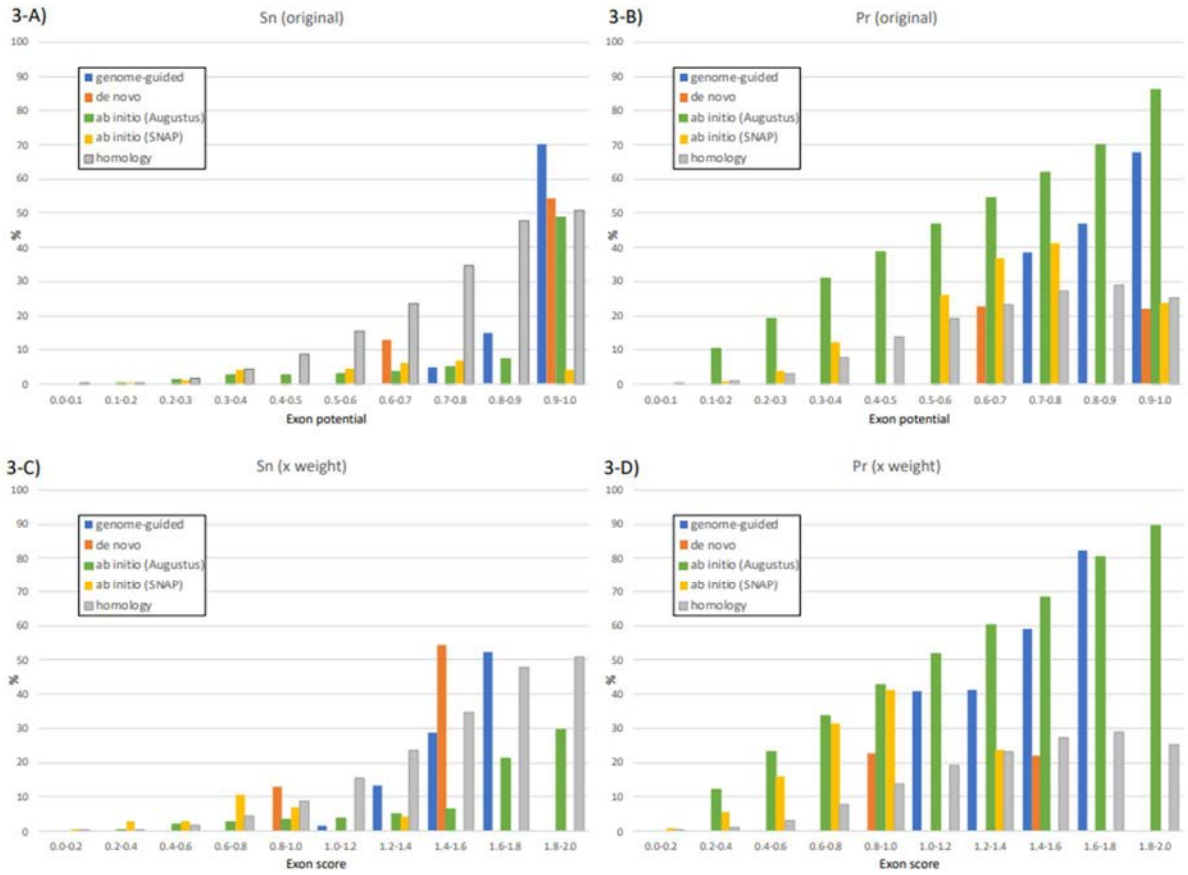


図 29　エクソンポテンシャル・エクソンスコアとエクソンの予測精度（イネ）

3-A)エクソンポテンシャルと Sn　3-B) エクソンポテンシャルと Pr

3-C)エクソンスコアと Sn　3-D) エクソンスコアと Pr

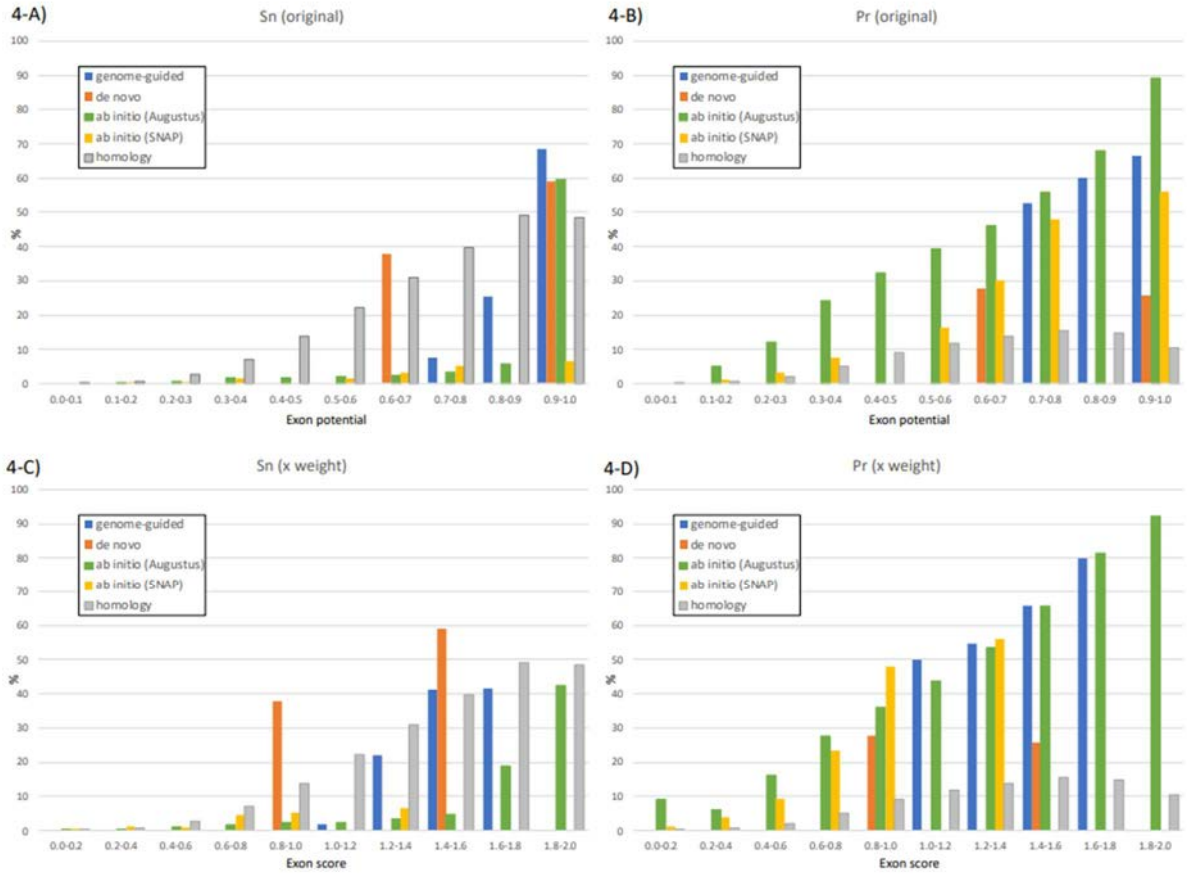


図 30　エクソンポテンシャル・エクソンスコアとエクソンの予測精度（ゼブラフィッシュ）

4-A)エクソンポテンシャルと Sn　　4-B) エクソンポテンシャルと Pr

4-C)エクソンスコアと Sn　　4-D) エクソンスコアと Pr

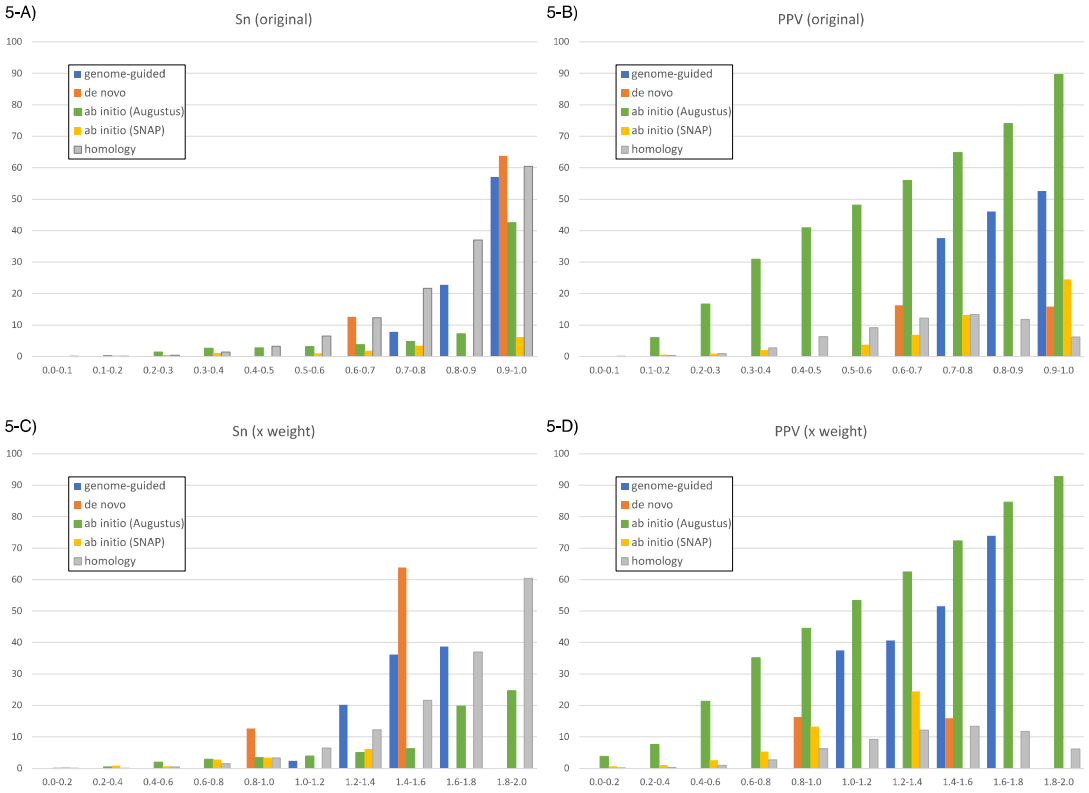


図 31　エクソンポテンシャル・エクソンスコアとエクソンの予測精度（ヒト）

5-A)エクソンポテンシャルと Sn　5-B) エクソンポテンシャルと Pr

5-C)エクソンスコアと Sn　5-D) エクソンスコアと Pr

2.4. ソフトウェアのパッケージング

2.4.1. Preparation phase における Nextflow の活用

GINGER は、Linux 系 OS で動作することを前提とした解析パイプラインであり、複数のツールを実行させていくことで解析目的を達成する。プロトタイプ開発時には、解析パイプラインは複数のシェルスクリプトで構成されており、ユーザはそれらのシェルスクリプト中に記載されている実行パラメータを編集し、利用可能な計算リソースを把握したうえで、順次実行させて行く必要があった。そのため、ユーザには、煩雑なファイル編集や、実行状況をこまめにモニタリングすることが求められていた。そこで、ユーザにかかる負荷を削減し、解析履歴の管理を容易とするために、実行パラメータを設定ファイルに集約し、あらかじめ指定した計算リソースの枠内で、予め定められた順番で計算を並列実行させていくことが可能となるようにパッケージングを行った。本パッケージングでは、その実現のために、ワークフローエンジンである Nextflow[39] を採用した。Nextflow は Java プラットフォーム上で動作する Groovy 言語をベースとして開発されたものである。以下、Nextflow の文法に従って記述されたロジックを「Nextflow 用ワークフロースクリプト」と呼ぶことにする。GINGER の解析パイプラインの中でも、入力ファイルや出力ファイルを適宜設定する必要がある部分や、実行の並列度を変化させたい部分を Nextflow 用ワークフロースクリプトに実装し直した（表 15）。

表 15　Preparation phase のための Nextflow 用ワークフロースクリプト

ファイル名	実行内容
`ginger.nf`	Preparation phase の全体
`mapping.nf`	Genome-guided assembly-based method の部分
`denovo.nf`	*de novo* assembly-based method の部分
`abinitio.nf`	*ab initio*-based method の部分
`homology.nf`	Homology-based method の部分

設定ファイルには、入力ファイルのパス、計算結果を出力するディレクトリのパス、スクラッチディスクのパス、使用する計算コア数、使用するメモリ量、実行させるコマンドのパス、実行パラメータ等を集約させた（表 16）。

表 16　Nextflow 用ワークフロースクリプトのための設定ファイル

変数名	設定内容
INPUT_GENOME	ゲノム配列データが格納されたファイルへの絶対パス
INPUT_MASKEDGENOME	リピート配列マスク済みのゲノム配列データが格納されたファイルへの絶対パス
INPUT_REPOUT	RepeatMasker によって出力されたリピート配列情報が格納されたファイル（拡張子が".out"）への絶対パス
INPUT_RNASEQR1	RNA-Seq のフォワード側断片配列データが格納されたファイルへの絶対パス
INPUT_RNASEQR2	RNA-Seq のリバース側断片配列データが格納されたファイルへの絶対パス
HOMOLOGY_DATA	近縁種のタンパク質配列データに関する情報を格納するための構造体であり、以下に示す PREFIX、PROTEIN、SPALNDB の配列で構成される
PREFIX	出力ファイル名に使われる接頭辞（任意の文字列であり、近縁種を区別するために使用）
PROTEIN	近縁種のタンパク質配列データが格納されたファイルへの絶対パス
SPALNDB	Spaln で使用される実行パラメータセット名
AUGUSTUS_SPEC	Augustus の統計モデルが配置されたディレクトリへの絶対パス
PDIR	GINGER の出力ファイルが配置されたディレクトリへの絶対パス
SCRATCH	一時的にデータが配置されるディレクトリへの絶対パス
N_THREAD	使用される最大スレッド数
MAX_MEMORY	使用される最大メモリ数
MIN0	TransDecoder の出力結果に対して適用する ORF 長の閾値（アミノ酸数）
MIN1	*ab initio*-based method の学習データセットが作成される際に用いられる ORF 長の閾値（塩基数）
SRA_FLAG	RNA-Seq データが SRA 形式であることを指定するフラグ（SRA 形式の場合には 1 を設定）
AUGUSTUS_TRAINING_SIZE	Augustus のトレーニングデータに含まれる遺伝子構造数
SNAP_TRAINING_SIZE	SNAP のトレーニングデータに含まれる遺伝子構造数
AUGUSTUS_TRAINING_DATA	Augustus のトレーニングデータが格納されたファイルへの絶対パス
SNAP_TRAINING_DATA	SNAP のトレーニングデータが格納されたファイルへの絶対パス
MAPPING_WEIGHT	Genome-guided assembly-based method によるエクソンポテンシャルに対する重み
DENOVO_WEIGHT	*de novo* assembly-based method によるエクソンポテンシャルに対する重み
HOMOLOGY_WEIGHT	Homology-based method によるエクソンポテンシャルに対する重み
AUGUSTUS_WEIGHT	Augustus によるエクソンポテンシャルに対する重み
SNAP_WEIGHT	SNAP によるエクソンにポテンシャル対する重み

2.4.2. Merge phase の実装

Merge phase は比較的に小さな計算リソースを要するのみであり、実行状況を小まめにモニタリングする必要がない。その一方で、ユーザは、遺伝子構造スコアに対する閾値を自動決定させない場合には、閾値を変更させながら何度も繰り返し実行することになる。そのため、Merge phase は、従来どおりにシェルスクリプトのままとし、記載内容の見通しをよくするためのリファクタリングは行ったが、機能的には上述の設定ファイルを読み込めるようにするに留めた（表 17）。

表 17　Merge phase のためのシェルスクリプト

ファイル名	実行内容
`phase0.sh`	Preparation phase における個別手法の予測結果を集約し、さらにエクソンスコアを付与する
`phase1.sh`	マルチエクソン遺伝子の遺伝子構造予測を行う（遺伝子構造スコアに対する閾値は自動決定）
`phase1_manual.sh`	マルチエクソン遺伝子の遺伝子構造予測を行う（遺伝子構造スコアに対する閾値はユーザが指定）
`phase2.sh`	シングルエクソン遺伝子の遺伝子構造予測を行う
`summary.sh`	予測した遺伝子の塩基配列・アミノ酸配列を生成

2.4.3. ソースツリーの整備

GINGER を広く配布するにあたって、ファイル構成の見通しが良くなるように再整理した（図 32 および表 18）。流通している多くのフリーソフトウェアと同様に、概要やインストール方法を記載したファイルも用意した。GINGER を動作させるためには、C++のソースコードをコンパイルする必要があることから、ソースツリーのトップディレクトリで make コマンドを実行するだけで必要なコンパイルが実行されるようにした。

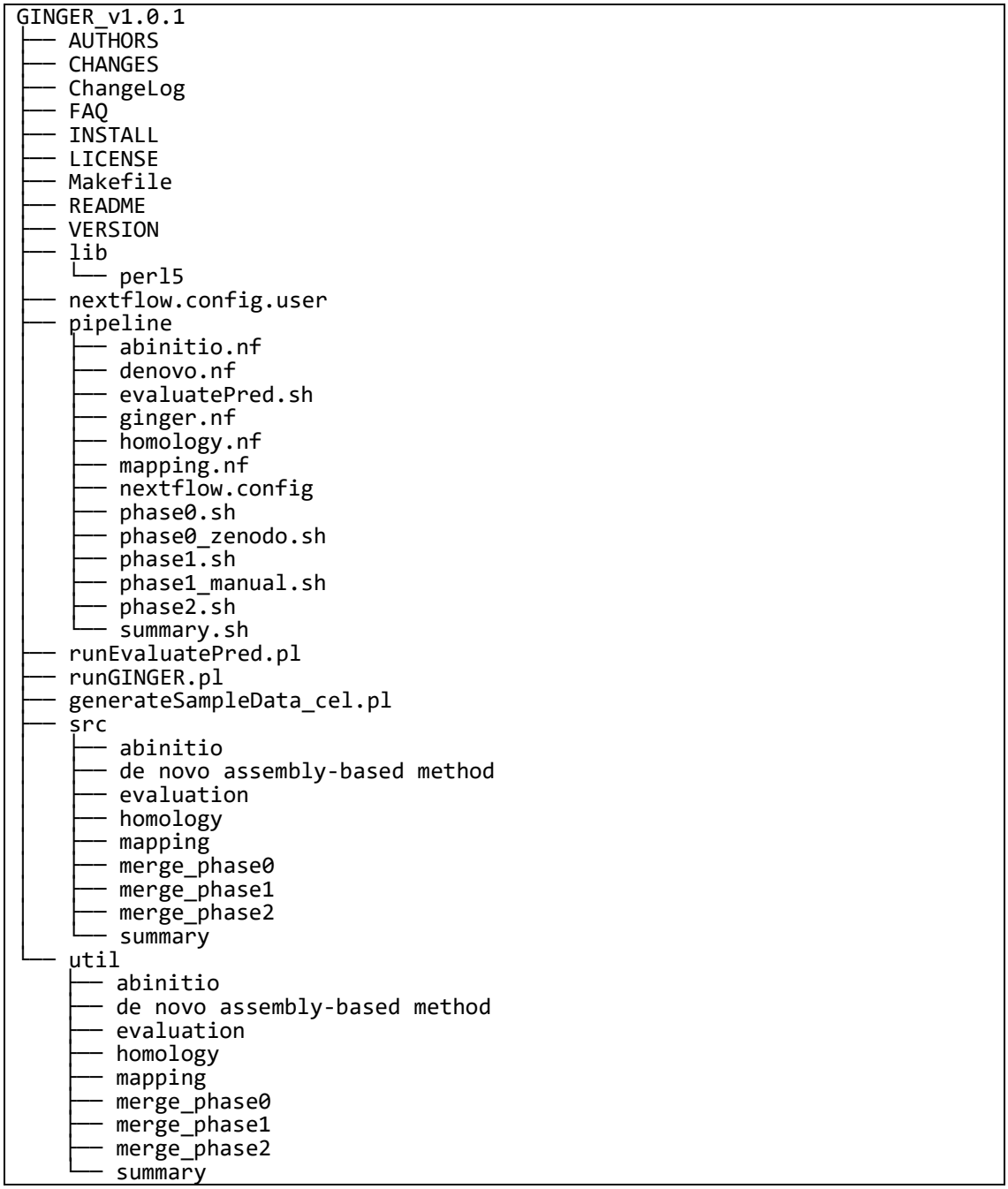


図 32　GINGER のソースツリー

表 18　ソースツリーを構成する主要なディレクトリ・ファイル

ディレクトリ・ファイル名	内容
README	概要、実行方法についての説明
INSTALL	インストール方法についての説明
Makefile	C++ソースコードをコンパイルし、所定のディレクトリに設置するための設定ファイル
nextflow.config.user	Docker 版 GINGER のための設定ファイル
runGINGER.pl	Docker 版 GINGER
generateSampleData_cel.pl	Docker 版 GINGER のためにサンプルデータをダウンロードするスクリプト
pipeline/	Preparation phase および Merge phase のパイプラインが配置されたディレクトリ
mapping.nf	Genome-guided assembly-based method のための Nextflow 用ワークフロースクリプト
denovo.nf	*de novo* assembly-based method のための Nextflow 用ワークフロースクリプト
homology.nf	Homology-based method のための Nextflow 用ワークフロースクリプト
abinitio.nf	*ab initio*-based method のための Nextflow 用ワークフロースクリプト
nextflow.config	Nextflow 用ワークフロースクリプトのための設定ファイル
phase0.sh	Merge phase ステップ 0 のためのシェルスクリプト
phase1.sh	Merge phase ステップ 1 のためのシェルスクリプト（遺伝子構造スコアに対する閾値を自動設定）
phase1_manual.sh	Merge phase ステップ 1 のためのシェルスクリプト（遺伝子構造スコアに対する閾値をユーザが指定）
phase2.sh	Merge phase ステップ 2 のためのシェルスクリプト
summary.sh	遺伝子構造予測結果の統計情報作成と ORF の塩基配列およびアミノ酸配列を生成するシェルスクリプト
src/	ソースコードが配置されたディレクトリ
util/	パイプラインから呼び出されるスクリプトおよびバイナリが配置されたディレクトリ

2.4.4. GitHub からのソースツリーの公開

本研究で整備したソースツリーは GitHub から公開している。URL は以下のとおりである。

https://github.com/i10labtitech/GINGER/

2.4.5. Conda パッケージ化と公開

本研究では、GINGER のインストール要件となっている外部ツール（Spaln 等）を含んだ Conda パッケージを作成し、ANACONDA のサイトから公開している。URL は以下のとおりである。

https://anaconda.org/i10labtitech/ginger

2.4.6. Docker 化と公開

本研究では、GINGER のインストール要件となっている外部ツール（Spaln 等）がインストールされ GINGER を実行可能な状態になっている Docker イメージを作成した。以下の URL から公開している。

https://hub.docker.com/r/i10labtitech/tools/tags

イメージ名は i10labtitech/tools:GINGER_v1.0.1 となっており、上述のソースツリーに含まれている runGINGER.pl、generateSampleData_cel.pl、runEvaluatePred.pl を用いることでイメージのダウンロードおよびその実行を行うことができるようになっている。

2.4.7. Zenodo からのベンチマークデータの公開

「2.3.4. 統合手法間の精度比較」で述べた精度比較に用いた入力データと実行コマンドは Zenodo から入手可能となっている。以下の URL から公開している。

https://zenodo.org/record/8126530

2.5. 考察

本研究では、複数の遺伝子構造予測手法から得られた結果を統合することでより精度の高い予測結果を得る真核生物用遺伝子予測ツール―GINGER―を新たに構築した。GINGER のアルゴリズムは、尤もらしさが反映されたスコアをエクソン候補に付与し、個々の遺伝子構造をコードすると考えられるゲノム領域毎にエクソン候補をグルーピングした

うえで、そのグループにおいて最適なエクソンの組み合わせを動的計画法で求めるアルゴリズムから成り立っている。このアルゴリズムは、従来から伊藤研究室でプロトタイプ的に用いられてきたものを全面的に見直したものであり、網羅的なベンチマークテストを通じ、MAKER、EVM といった既存の統合ツールに勝る性能を有することも合わせて確認した。また、アイソフォーム候補となりうる、準最適解を出力できることも GINGER が持つ一つの特徴である。そして、多くのユーザが利用できるように本ツールのソースコードとパッケージを整備し、GitHub、Conda および Docker を通じ一般公開した。また、ベンチマークテストでは、他の統合手法と予測精度を比較するのみでなく、本研究で導入したエクソンポテンシャルが精度向上に寄与していること、ならびにエクソンのグルーピングが他の統合ツールと比べた際の精度差に貢献していることを確認した。

エクソンに付与されるスコアは、個別の遺伝子構造予測手法から得られるエクソン候補に対して、それぞれの手法に応じた統計量に基づくエクソンポテンシャルを導入することと、手法の精度を反映した重み係数を導入することで、エクソンレベルの精度と相関するように設計されており、ベンチマークでもその成果を確認することができた。エクソン候補をグルーピングするだけでなく、エクソンに付与されたスコアを活用した分割基準や、アーチファクトと考えられる非常に長いエクソンでグループを分割することにより、ミスライゲーションを回避できていることもベンチマークにおいて確認できた。

また、シングルエクソン遺伝子とマルチエクソン遺伝子の遺伝子構造スコアを分けてスコアの分布を分析することで、遺伝子構造スコアに対する自動的な閾値設定を、精度を損なわないように安定化させるロジックが組み込まれていることも GINGER の利点の一つである。この機能により、ユーザが生物種に応じた閾値を設定することなく、自動で決定された閾値を利用し、精度の高い遺伝子予測結果を得ることが可能である。さらに、ワークフローエンジンである Nextflow を活用することで、実行パラメータを設定ファイルに集約させ、設定を容易にするだけでなく、設定および実行のログを容易に管理できるようにし、計算リソースと実行状況をモニタリングする必要から利用者を開放することで、全体的にオートメーション化を進めることができた。

GINGER パッケージの研究者コミュニティへの普及においても、GitHub、Conda、DockerHub から公開することができた。Docker イメージを用いることで、インストールの労力を大幅に削減させることができ、新規に決定されたゲノムに対して、Docker イメージを用いることで、容易に高精度な遺伝子構造予測を行なうことができるようになっている。ゲノム再構築や遺伝子構造予測に精通しない利用者にはぜひこの Docker イメージを活用してもらいたい。また、GINGER のソースコードも併せて GitHub から公開しているため、精度を究極まで突き詰めたい、特別な生物種や、特殊な遺伝子構造、シングルエクソン遺伝子に対応できるアルゴリズムを開発したいといった研究者がソースコードを活用し、改良して利用することも可能である。

3. 実ゲノムデータへの適用

3.1. 背景と目的

GINGER の有効性を確認するために、開発及びベンチマークに用いたモデル生物種以外の非モデル生物ゲノムに対して GINGER を適用し、遺伝子構造アノテーションを行った。適応対象はアカムツとした。学名は *Doederleinia berycoides* であり、スズキ目スズキ亜目ホタルジャコ科に属する魚類である。口腔内が黒いことから「ノドグロ」と呼ぶ地域もある。脂ののった白身魚であり、食味が良いことから近年では人気のある食用魚として取引されている。日本全国に分布しているが、比較的温暖な海域を好む魚種であり、北海道では釣りの対象にはなっていない。東京湾の水深 50m 付近でも釣れることはあるが、釣りの場合には 200〜300m 辺りを狙うことが多く、中深海に生息する魚として認識が広まっている。

食味の良いことから可食部の組成に興味が及ぶところであるが、その分析事例はほぼ見当たらず、唯一、韓国における研究事例が見られたのみである [40]。脂ののった魚であることからエイコサペンタエン酸（Eicosapentaenoic Acid：EPA）やドコサヘキサン酸（Docosahexaenoic Acid：DHA）といった健康に好影響を与えるとされる不飽和脂肪酸が豊富なことが期待されるが、上述の論文によると、EPA や DHA はタイリクスズキやホンニベよりも含有量は少ない一方で、イコセン酸（11-Eicosenoic acid）等の一価不飽和脂肪酸がそれらよりも比較的に多く含まれていることが示されている。脂肪酸（Fatty acid）の大くくりで見てもアカムツは他の魚種より多くはなく、一方で、脂質（Lipid）に関してはタイリクスズキやホンニベの 2 倍以上の含有量であることが示されている。

本研究では、新規に決定したアカムツゲノムに対して GINGER を適用し、アカムツゲノムにコードされている遺伝子構造をどの程度予測できるかを確認し、その上で、相同遺伝子の同定と、モデル生物種の機能アノテーションを参照することで食味に関係する可能性がある遺伝子を抽出した。

3.2. 材料

3.2.1. アカムツ個体

ゲノムデータの取得対象となるアカムツ個体は、筆者が茨城県沖で釣り上げたものである。以下に、その関連情報をまとめる。

個体採取場所　：　茨城県北茨城市平潟町沖
個体採取方法　：　ルアー釣り（ジギング；第十五隆栄丸に乗船）
個体採取日時　：　9:15 時頃
個体サイズ　　：　全長が 20cm 未満（図 33）

図 33　採取したアカムツ

撮影：2020/12/11 9:29

3.2.2.　ゲノムデータ

新鮮なアカムツ個体から鮮血を採取、100ul 1 x PBS/ 2.5ul 0.5M EDTA 入りチューブに加えて輸送し、遠心分離後ペレットをマイナス 80 度で保存した。保存したペレットを 1 x PBS で懸濁後、smart DNA Prep kit（analytikjena 社）の標準プロトコールに従い DNA 抽出を実施した。抽出された DNA サンプルを g-tube デバイスを用いて 30kb～100kb 程度の長さに断片化し、SMRTbell ライブラリを SMRTbell Express Template Prep Kit（PacBio 社）を用いて作成した。加えてコバリス社 M220 ウルトラソニケーターにて 500bp 平均になるように断片化し、TruSeq DNA PCR-Free Library Prep kit（Illumina 社）を用いて paired-end ライブラリを作成した。さらに鮮血から、Dovertail Omni-C Kit（Dovetail Genomics 社）を用い、標準プロトコルに従って Omni-C ライブラリを作成した。CLR SMRTbell ライブラリは PacBio Sequel II、pair-end ライブラリ、Omni-C ライブラリは NovaSeq6000 シーケンサーを用いてシークエンスし、以下の断片配列データを得た（表 19）。なお、ライブラリ調整およびシークエンスは国立遺伝学研究所 豊田特任教授に行っていただいた。

表 19　シークエンシングデータの統計値

	レイアウト	リード数	塩基数（Gbp）	カバー率の推計値
Illumina	PE (250bp×2)	385,556,242	96.4	133.1
Pacbio	SE	7,004,038	202.4	279.5
Omni-C	PE (150bp×2)	244,509,098	36.9	51.0

※カバー率は推定ゲノムサイズ 724Mb として算出

Illumina Pair-end データから GenomeScope2.0 を用いて推定されたゲノムサイズは約724Mbp、ヘテロ接合度は 0.81%であった（図 34）。

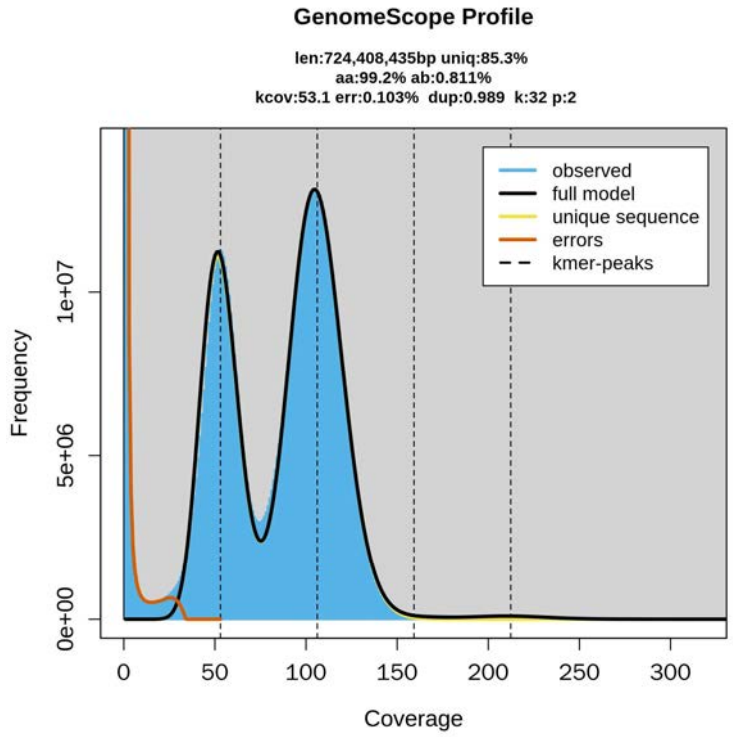


図 34　genomescope2.0 を用いたゲノムサイズの推定

続いて、PacBio CLR リードを用いてアセンブルを実施した。アセンブルには Canu v2.1.1[41]、Flye v2.9[42] および Nextdenovo v2.5.0[43] が用いられた。結果の統計値を表 20 に示す。

表 20　各種アセンブラーによるアセンブル結果の統計値

アセンブラー	Contig 数	合計長（bp）	最長 Contig 長（bp）	N50 (bp)
Canu v2.1.1	2,630	1,438,726,058	23,964,398	2,742,639
Canu v2.1.1 *	2,765	1,207,289,178	17,589,801	1,999,727
Flye v2.9	734	730,765,989	30,030,209	15,382,793
Nextdenovo v2.5.0	69	729,243,393	51,715,729	29,028,919

※ heteroOption: corOutCoverage=200 "batOptions=-dg 3 -db3 -dr1 -ca 500 -cp 50”

以上より、明らかに Nextdenovo v2.5.0 による結果の統計値がよいため、この結果を用いて下流の解析を進めることとした。得られたアセンブル結果に対し、PacBio CLR リードと Illumina リードを用いて、arrow を 2 回、Nextpolish を 3 回かけることでポリッシュを実施した。コンタミチェック後、Omni-C データを用いて 3D-DNA による Hi-C scaffolding を

実施後、マニュアルによるミスアセンブル箇所の切断などを行うことで、最終的に表 21 に示す結果を得た。中でも 1 Mb 以上の scaffold が 24 本、合計 720.8Mb（全 scaffold の 99.3%）であり、染色体 24 本のゲノム配列がほぼ構築できていることが確認できる。本研究ではこのゲノム配列を用いた。なお、ゲノムのアセンブルは伊藤研究室所属の大内氏が主に実施した。

表 21　最終的なアセンブリング結果

scaffold 数	合計長（bp）	scaffold 数（> 1 Mb）	合計長（bp）（> 1 Mb）	最長 Contig 長（bp）	N50（bp）
39	725,647,098	24	720,775,832	37,324,599	32,045,567

3.2.3.　RNA-Seq データ

DNA を採取したものと同一のアカムツ個体から、筋肉を船上でエッペンチューブにとりわけ、RNA Protect reagent に浸した状態で保存し持ち帰った。このサンプルをホモジナイザーを使用して組織の粉砕を行い、QIAGEN 社 RNeasy Protect Mini Kit を用い、標準プロトコルに従って RNA を抽出した。抽出した RNA は、Illumina Stranded mRNA Prep キットによりライブラリ調整され、DNBSEQ-T7 によりシークエンスされた。ライブラリ調整およびシークエンスは GeneWiz 社にて実施された。得られたシークエングデータは表 22 のとおりである。

表 22　RNA シークエンシングデータの統計値

レイアウト	リード数	塩基数（Gbp）
PE (150bp×2)	69,005,236	10.4

3.3.　方法

3.3.1.　使用するパッケージ

本適用では、「2.4. ソフトウェアのパッケージング」で述べた GINGER のパッケージ（バージョン：1.0.1）を用いた。

3.3.2.　入力データの用意

GINGER の入力データとなるゲノム配列データと RNA-Seq データは「3.2. 材料」の通りである。Homology-based method で用いる近縁種は以下の通り決定した。NCBI の Taxonomy において、アカムツはスズキ系真スズキ亜系（Eupercaria）と呼ばれる分岐群に属しており、300 種を超えるゲノム配列データが登録されている。NCBI RefSeq プロジェクトが作成している遺伝子構造アノテーションは信頼性が高いと考えられるが、上述の 300 種の中で 37 種について公開されていた。本解析では、その中からアカムツに比較的形

態が似ているキチヌ（*Acanthopagrus latus*）とストライプドバス（*Morone saxatilis*)、淡水魚ではあるがコクチバス（*Micropterus dolomieu*）とオオクチバス（*Micropterus salmoides*)、モデル生物種であるゼブラフィッシュを(*Danio rerio*)を用いることとした。用いた 5 種の配列情報を表 23 に示す。

表 23　アノテーションに用いた近縁種の配列情報

用いた生物種	アクセッション番号	Assembly 名
ゼブラフィッシュ（*Danio rerio*）[36]	GCF_000002035.6	GRCz11
キチヌ（*Acanthopagrus latus*）[44]	GCF_904848185.1	fAcaLat1.1
ストライプドバス（*Morone saxatilis*）[45]	GCF_004916995.1	NCSU_SB_2.0
コクチバス（*Micropterus dolomieu*）[46]	GCF_021292245.1	ASM2129224v1
オオクチバス（*Micropterus salmoides*）[47]	GCF_014851395.1	ASM1485139v1

3.3.1.　実行パラメータ設定

GINGER における全ての設定可能な実行パラメータはデフォルトのとおりとした。また、遺伝子構造スコアに対する閾値設定は自動決定とした。

3.4.　GINGER を用いた遺伝子構造予測結果

GIGNER を実行して得た遺伝子構造予測結果に関する基本的な統計値を表 24 に、BUSCO[48] v5.5.0（protein モード）による網羅性評価結果を表 25 に、それぞれ近縁種の情報と合わせて示した。BUSCO は、ある系統に含まれる生物種が単一コピーで保持していると考えられる遺伝子の何割を、評価対象であるゲノムあるいは予測遺伝子中に見つけることができるかにより、その網羅性を評価するツールである。評価に用いたデータベースは actinopterygii_odb10（3,640 遺伝子）である。遺伝子構造予測結果の統計値が近縁種とほぼ同じであり、BUSCO による評価でも Complete BUSCOs が 98.1%と極めて高い網羅性を示した。これらの結果より、GINGER による遺伝子構造予測がベンチマークに用いたモデル生物種のみではなく、非モデル生物のゲノムにも有効であることが確認できた。

表 24　アカムツに対する GINGER の予測結果と近縁種の遺伝子構造情報（RefSeq）

種	タンパク質コーディング遺伝子（#）	うちシングルエクソン遺伝子（#）	平均 CDS 長(bp)	平均エクソン長(bp)	平均イントロン長(bp)
アカムツ	25,814	1,152	1,681.8	172.8	1,301.7
ゼブラフィッシュ	25,526	2,052	1,744.9	181.2	3,009.6
キチヌ	23,706	1,666	1,854.6	175.5	1397.3
ストライプドバス	21,640	1,277	1,728.4	169.0	1273.7
コクチバス	24,663	1,580	1,730.0	173.2	1,836.3
オオクチバス	27,166	1,849	1,735.3	174.2	1,610.8

表 25　アカムツ予測遺伝子および近縁魚種 RefSeq 遺伝子に対する BUSCO を用いた網羅性評価

種	C(%)	S(%)	D(%)	F(%)	M(%)
アカムツ	98.1	97.3	0.8	0.8	1.1
ゼブラフィッシュ	96.9	95.5	1.4	1.3	1.8
キチヌ	98.7	98.0	0.7	0.3	1.0
ストライプドバス	91.5	90.5	1.0	2.7	5.8
コクチバス	97.7	96.8	0.9	1.5	0.8
オオクチバス	98.8	92.3	6.5	0.7	0.5

C: Complete BUSCOs…OrthoDB の遺伝子セット中で、ほぼ完全な形で同定できる遺伝子

S: Complete and single-copy BUSCOs…C のうち、単一コピーであるもの

D: Complete and duplicated BUSCOs…C のうち、重複しているもの

F: Fragmented BUSCOs…BUSCO の遺伝子セット中で、断片化した形で同定できる遺伝子

M: Missing BUSCOs…BUSCO の遺伝子セット中で、同定できなかった遺伝子

3.5.　近縁種との相同解析結果

近縁種との相同遺伝子を同定することで、アカムツがどのような機能の遺伝子を多く保有しているかを調査した。相同遺伝子の同定には SonicParanoid[49] を使用した。入力データセットには、アカムツ、ゼブラフィッシュ、ストライプドバス、キチヌ、オオクチバスおよびコクチバスのタンパク質配列を用いた。アカムツのタンパク質配列は GINGER による予測結果である。その他の 5 種のタンパク質配列は、RefSeq が提供するものであり、複数のアイソフォームを含む遺伝子に関しては最長のものを抽出してデータセットに含めた。SonicParanoid を実行した結果、21,049 の相同遺伝子グループが得られた。その中で最も多かったタイプは、各魚種からの 1 コピーの遺伝子から成るもの、つまり単一コピー遺伝子

から成るものであった。このような相同遺伝子グループは 17,314 個あり、その中でも、6 生物種の全てが単一コピーを有しているものは、12,366 個であった。相同遺伝子数が最大のグループは 228 個の遺伝子を含んでおり、100 個以上の遺伝子を含むものは 17 グループ、10 個以上かつ 100 個未満の遺伝子を含むグループは 1,111 グループであった。

上述の 12,366 個の相同遺伝子グループについてアミノ酸配列をチェックし、標準的な 20 種類のアミノ酸を示す文字のみで配列が構成されている 12,348 個を用いて 6 生物種の系統樹を作成した。各相同遺伝子グループでマルチプルアライメントを作成したのち、アライメントから gap 領域など余分な部分を除去し、最尤法で系統樹を作成した。使用したツールとオプションを表 26 に、ゼブラフィッシュ（*D.rerio*）をアウトグループとした場合の系統樹を図 35 に示す。この系統樹において、コクチバス（*M.salmoides*）とオオクチバス（*M.doloieu*）は特に近い関係にあること、ストライプドバス（*M.saxatilis*）、キチヌ（*A.latus*）、アカムツ（*D.berycoides*）の中でも、アカムツがコクチバス・オオクチバスに近い系統にあることが明らかとなった。

表 26　系統樹作成に使用したツールとオプション

ツール	計算内容	オプション
Mafft v7.520[50]	マルチプルアライメントの作成	--atuo （マルチプルアライメント構築戦略を自動決定）
Trimal v1.4.rev15[51]	マルチプルアライメントからの余分な部分の除去	なし
IQ-tree 2.2.3[52]	系統樹の作成	-m MFP-bb 10000 （ModelFinder Plus によって置換モデルを選択し、10,000 回のブートスタラップサンプリングを実行）
iTOL v6[53]	系統樹の可視化	ゼブラフィッシュの枝をルートに設定

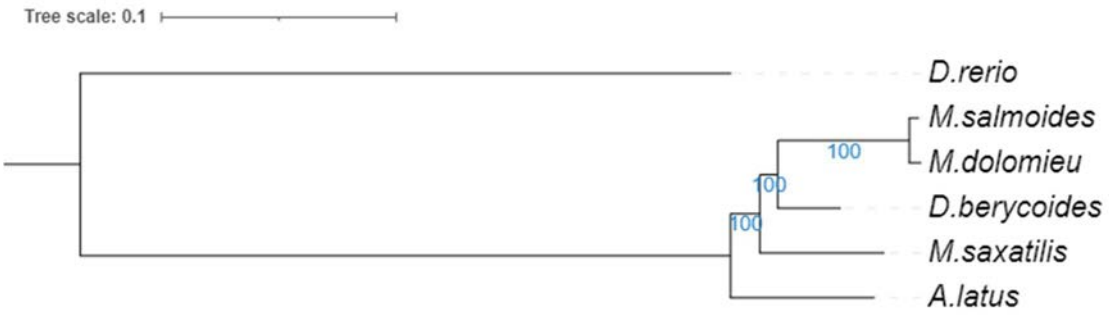


図 35　単一コピー遺伝子を用いて作成した系統樹

※青色の数字はブートストラップ値

続いて得られた相同遺伝子グループの中から、アカムツの遺伝子が他種に比べて多くなっているものに着目した。相同遺伝子グループに含まれる遺伝子の機能は、そのグループに含まれるゼブラフィッシュの遺伝子に付与されている機能アノテーションを参照した。その中でも食味に関連する可能性がある相同遺伝子グループを以下にまとめる。以下の表には、相同遺伝子グループに含まれる各魚種の遺伝子数と、ゼブラフィッシュの機能アノテーションを整理した。表頭における略称は以下の魚種に該当する。

［表頭の略称］
Dbe　：アカムツ
Ala　：キチヌ
Mdo　：ストライプドバス
Msal　：コクチバス
Msax　：オオクチバス
Dre　：ゼブラフィッシュ

【Asparagine synthetase を含む相同遺伝子グループ】

アスパラギンはうま味成分の一つとして認識されている。表 27 に列挙した相同遺伝子グループは、メンバーに含まれるゼブラフィッシュの遺伝子に付与されている機能アノテーションが Asparagine synthetase となっているものである。この中で、アカムツの遺伝子が多いものがあった（表 27 の No.1）。

表 27　Asparagine synthetase を含む相同遺伝子グループ

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
1	2	1	1	1	1	1	asparagine synthetase [glutamine-hydrolyzing] [asns]
2	1	1	1	1	1	1	asparagine synthetase domain-containing protein 1 [asnsd1]

【Glutathione synthetase を含む相同遺伝子グループ】

グルタチオンは、グルタミン、システインおよびグリシンの 3 つのアミノ酸が結合したペプチドであり、抗酸化機能を有している。酸化した脂肪酸は食味に悪影響を与える可能性があることからも、グルタチオンが食味に良い影響を与える可能性はあり得る。グルタチオンの合成を担う Glutathione synthetase の相同遺伝子グループにおいてアカムツの遺伝子が多かった（表 27）。

表 28　Glutathione synthetase を含む相同遺伝子グループ

No.1	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
1	3	2	2	1	1	1	glutathione synthetase isoform X1 [gss]

【Amino acid transporter を含む相同遺伝子グループ】

アミノ酸が輸送されることによってアミノ酸の含有量に変化が生じ、成長や、うまみ成分の合成に影響を及ぼすことが考えられる。アミノ酸の輸送を担う Amino acid transporter でもアカムツで遺伝子が多い相同遺伝子グループがあった（表 29 の No.1）。

表 29　Amino acid transporter を含む相同遺伝子グループ

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
1	9	6	3	3	3	3	sodium-dependent neutral amino acid transporter B(0)AT1 [slc6a19a.1]等
2	3	1	1	1	1	3	sodium- and chloride-dependent neutral and basic amino acid transporter B(0+) [slc6a14]等
3	4	2	3	3	3	2	sodium-coupled monocarboxylate transporter 1 [slc5a8]
4	2	2	1	1	2	1	sodium- and chloride-dependent GABA transporter ine isoform X1 [si:ch211-283g2.1]
5	2	2	2	1	2	2	sodium-dependent neutral amino acid transporter B(0)AT2 [slc6a15]等
6	2	2	2	2	2	1	vesicular inhibitory amino acid transporter [slc32a1]
7	1	1	1	1	1	1	large neutral amino acids transporter small subunit 2 [slc7a8a]
8	1	1	1	1	1	1	putative sodium-coupled neutral amino acid transporter 10 isoform X2 [slc38a10]
9	1	1	1	1	1	1	large neutral amino acids transporter small subunit 3 [slc43a1b]
10	1	1	1	1	1	1	vesicular inhibitory amino acid transporter [si:dkey-126h10.1]
12	1	1	1	1	1	1	solute carrier family 43 (amino acid system L transporter), member 1a [slc43a1a]
13	2	3	2	1	3	1	sodium-coupled neutral amino acid transporter 5 [slc38a5b]
14	1	1	1	1	1	1	sodium-coupled neutral amino acid transporter 3-like isoform X1 [slc38a3b]
15	1	1	1	1	1	1	Y+L amino acid transporter 1 [slc7a7]
16	1	1	1	1	1	1	sodium-coupled neutral amino acid transporter 4 isoform X1 [slc38a4]
17	1	1	1	1	1	1	B(0%2C+)-type amino acid transporter 1 [slc7a9]
18	1	1	1	1	1	1	cationic amino acid transporter 2 [slc7a2]
19	1	1	1	1	1	1	probable cationic amino acid transporter [slc7a14a]
20	1	1	1	1	1	1	putative sodium-coupled neutral amino acid transporter 7 [slc38a7]

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
21	1	1	1	1	1	1	high affinity cationic amino acid transporter 1 isoform X1 [slc7a1]
22	1	1	1	1	1	1	Y+L amino acid transporter 2 isoform X1 [slc7a6]
23	1	1	1	1	1	1	sodium-coupled neutral amino acid transporter 9 [slc38a9]
24	1	1	1	1	1	1	proton-coupled amino acid transporter 1 [slc36a1]
25	1	1	1	1	1	1	solute carrier family 3 (amino acid transporter heavy chain), member 2b [slc3a2b]
26	1	1	1	1	1	1	lysosomal amino acid transporter 1 homolog isoform X1 [pqlc2]
27	1	1	1	0	2	1	excitatory amino acid transporter 5 [slc1a8a]
28	1	1	1	1	1	1	large neutral amino acids transporter small subunit 4 isoform X1 [slc43a2b]
29	1	1	1	1	1	1	high affinity cationic amino acid transporter 1 [zgc:63694]
30	1	1	1	1	1	1	excitatory amino acid transporter 5 isoform 1 [slc1a7a]
31	1	1	1	1	1	1	b(0%2C+)-type amino acid transporter 1-like isoform X1 [zmp:0000001267]
32	1	1	1	1	1	1	cationic amino acid transporter 4 [slc7a4]
33	1	1	1	1	1	1	excitatory amino acid transporter 2 [slc1a2b]
34	1	1	1	1	1	1	putative sodium-coupled neutral amino acid transporter 11 isoform X1 [slc38a11]
35	1	1	1	1	1	2	sodium- and chloride-dependent taurine transporter [slc6a6a]等
36	1	1	1	1	1	2	large neutral amino acids transporter small subunit 1 [slc7a5] 等
37	1	1	1	1	2	1	excitatory amino acid transporter 3 [slc1a1]
38	1	1	1	1	1	2	probable sodium-coupled neutral amino acid transporter 6 [slc38a6] 等
39	1	2	1	1	1	1	asc-type amino acid transporter 1 [slc7a10a]
40	1	1	1	1	2	1	asc-type amino acid transporter 1 [slc7a10b]
41	1	2	1	1	1	2	uncharacterized protein LOC790938 [zgc:158423] 等
42	1	2	2	1	1	1	neutral amino acid transporter A isoform X1 [slc1a4]
43	1	1	2	2	1	1	excitatory amino acid transporter 4 isoform X1 [slc1a6]
44	1	1	1	2	2	2	sodium-coupled neutral amino acid transporter 2 [slc38a2] 等

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
45	1	1	2	2	1	2	putative sodium-coupled neutral amino acid transporter 8 [slc38a8a] 等

【Collagen を含む相同遺伝子グループ】

コラーゲンは、皮膚、軟骨、筋肉等で繊維構造を形成する。それらの硬さや口の中でのばらけ方といった食感に影響を及ぼす。相同遺伝子グループの中には、アカムツの遺伝子が多いものが存在した（表 30 の No.1）。

表 30　Collagen を含む相同遺伝子グループ

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
1	4	3	2	2	3	2	collagen alpha-6(VI) chain [LOC100535278], collagen alpha-6(VI) chain [col6a4a]
2	1	1	1	0	0	1	collagen alpha-1(I) chain isoform X1 [si:dkey-61l1.4]
3	1	1	0	0	2	1	collagen alpha-3(VI) chain-like isoform X1 [LOC108191316]
4	1	1	1	0	1	1	collagen alpha-1(XXVIII) chain isoform X1 [col28a1a]
5	1	1	1	0	1	1	collagen alpha-1(XII) chain isoform X1 [col12a1a]
6	2	2	2	2	2	1	collagen type IV alpha-3-binding protein isoform 1 [col4a3bpa]
7	1	1	1	0	1	1	collagen alpha-1(XXI) chain-like isoform X1 [si:dkey-225n22.4]
8	1	1	1	1	1	1	collagen alpha-1(XXIV) chain [col24a1]
9	1	1	1	1	1	1	collagen type XVIII%2C alpha 1 isoform X1 [col18a1a]
10	1	1	1	1	1	1	collagen alpha-1(VII) chain isoform X1 [col7a1]
11	1	1	1	1	1	1	procollagen C-endopeptidase enhancer precursor [pcolcea]
12	1	1	1	1	1	1	collagen alpha-1(I) chain-like isoform X1 [si:ch211-196i2.1]
13	1	1	1	1	1	1	collagen alpha-1(X) chain precursor [col10a1a]
14	1	1	1	1	1	1	collagen alpha-1(XI) chain isoform X1 [col11a1a]
15	1	1	1	1	1	1	collagen alpha-1(XXVII) chain B precursor [col27a1b]
16	1	1	1	1	1	1	collagen alpha-1(V) chain isoform X2 [col5a1]
17	1	1	1	1	1	1	collagen alpha-1(XVII) chain isoform X1 [col17a1b]
18	1	1	1	1	1	1	procollagen galactosyltransferase 2

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
							precursor [colgalt2]
19	1	1	1	1	1	1	procollagen C-endopeptidase enhancer b precursor [pcolceb]
20	1	1	1	1	1	1	collagen alpha-1(VI) chain [col6a1]
21	1	1	1	1	1	1	collagen alpha-1(XX) chain-like isoform X1 [LOC101884357]
22	1	1	1	1	1	1	collagen type XXVIII alpha 1 a precursor [col28a2a]
23	1	1	1	1	1	1	procollagen galactosyltransferase 1-like [colgalt1]
24	1	1	1	1	1	1	collagen alpha-1(XXVI) chain-like [LOC101885935]
25	1	1	1	0	2	1	procollagen-lysine%2C2-oxoglutarate 5-dioxygenase 1 precursor [plod1a]
26	1	1	1	1	1	1	collagen alpha-2(VI) chain isoform X1 [col6a2]
27	1	1	1	1	1	1	collagenase 3 [mmp13b]
28	1	1	1	1	1	1	collagen alpha-1(X) chain [col10a1b]
29	1	1	1	1	1	1	collagen alpha-2(VIII) chain precursor [col8a2]
30	1	1	1	1	1	1	collagen triple helix repeat-containing protein 1 [cthrc1b]
31	1	1	1	1	1	1	collagen alpha-2(IX) chain precursor [col9a2]
32	1	1	1	1	1	1	collagen alpha-3(V) chain precursor [col5a3a]
33	1	1	1	1	1	1	collagen alpha-1(VIII) chain precursor [col8a1a]
34	1	1	1	1	1	1	acetylcholinesterase collagenic tail peptide isoform X1 [colq]
35	1	1	0	1	2	1	collagen alpha-1(VII) chain [col7a1l]
36	1	1	1	1	1	1	collagen alpha-1(XIV) chain isoform X1 [col14a1a]
37	1	1	1	1	1	1	collagen alpha-1(XII) chain isoform X1 [col12a1b]
38	1	1	1	1	1	1	collagen alpha-1(XXI) chain-like isoform X1 [col21a1]
39	1	1	1	1	1	1	collagen alpha-2(XI) chain isoform X1 [col11a2]
40	1	1	1	1	1	1	collagen alpha-3(IX) chain precursor [col9a3]
41	1	1	1	1	1	1	collagen alpha-1(VIII) chain [col8a1b]
42	1	1	1	1	1	1	collagen%2C type V%2C alpha 3b isoform X1 [col5a3b]
43	1	1	1	1	1	1	collagen alpha-1(XXIII) chain isoform X1 [si:dkeyp-44a8.4]
44	1	1	1	1	1	1	procollagen galactosyltransferase 1 precursor [si:ch211-114l13.7]

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
45	1	1	1	1	1	1	72 kDa type IV collagenase precursor [mmp2]
46	1	1	1	1	1	1	procollagen C-endopeptidase enhancer 2 precursor [pcolce2b]
47	1	1	1	1	1	1	collagen%2C type XVII%2C alpha 1a isoform X1 [col17a1a]
48	1	1	1	1	1	1	collagen alpha-1(XV) chain isoform X1 [col15a1b]
49	1	1	1	1	1	1	collagen alpha-1(XI) chain [col11a1b]
50	1	1	1	1	1	1	collagen alpha-1(XXII) chain-like [zmp:0000000760]
51	1	0	1	2	2	1	collagen alpha-1(XXIV) chain-like [LOC100332483]
52	1	1	1	1	2	1	collagen and calcium-binding EGF domain-containing protein 1 precursor [ccbe1]
53	0	0	1	0	0	1	collagen type IX alpha l precursor [col9a1b]
54	0	1	1	0	1	1	collagen alpha-1(XXV) chain isoform X1 [col25a1]

【Stearoyl-CoA desaturase を含む相同遺伝子グループ】

Stearoyl-CoA desaturase は短鎖不飽和脂肪酸であるオレイン酸の合成に関与している。オレイン酸を多く含む牛肉や豚肉はさっぱりとした食感を提供するとされている。相同遺伝子グループの中には、アカムツの遺伝子が多いものがあった（表 31 の No.1）。

表 31　Stearoyl-CoA desaturase を含む相同遺伝子グループ

No	Dbe	Ala	Mdo	Msal	Msax	Dre	ゼブラフィッシュの機能アノテーション
1	3	1	1	0	1	1	stearoyl-CoA desaturase b isoform X1 [scdb]
2	1	1	1	1	1	1	stearoyl-CoA desaturase 5 [scd]

3.6. GINGER による予測結果の精度に関する追加解析

シーケンサーやアセンブラーの高度化が進み、ゲノム配列決定が加速する一方で、ゲノム配列決定後の解析、特に遺伝子構造アノテーションに関しては、ゲノム配列決定ほどルーチン化はされておらず、様々な論文にて試行錯誤が行われており、高精度とは言えない結果が掲載されている場合も多い。アカムツが含まれる Eupercaria に属し、本年度（2023 年度）に論文にてゲノム配列および遺伝子構造アノテーションデータが公開されている魚種について、各種統計値と論文公開データに基づいた BUSCO による評価値を調べた。以下の表 32 に、そのまとめた結果と、本研究で対象としたアカムツのデータとを合わせて示す。いずれに対する評価も BUSCO v5.5.0、データベースは actinopterygii_odb10 (3,640 遺伝子)を用いている。

表 32　2023 年度に論文公開されたゲノム・遺伝子データに対する評価結果

種名	ゲノム統計値		BUSCO（genome モード）				予測遺伝子数（#）	予測遺伝子 BUSCO（protein モード）			
	合計長	N50	C	S	D	M		C	S	D	M
ホンソウワケベラ	726.4Mb	33.6Mb	98.6	97.5	1.1	1.2	28,138	94.9	91.3	3.6	3.4
ダーター	788.7Mb	30.8Mb	96.7	95.7	1.0	2.3	28,034	81.0	79.3	1.7	11.5
フォークランドアイナメ	606.3Mb	26.7Mb	98.0	97.0	1.0	1.6	27,321	94.3	89.7	4.6	3.6
スナウナギ	808.5Mb	33.7Mb	97.0	95.9	1.1	2.7	22,274	94.3	61.5	32.9	4.8
ケツギョ	716.2Mb	30.0Mb	97.9	97.3	0.6	1.7	28,905	92.1	58.7	33.4	6.7
コウライケツギョ	740.4Mb	30.5Mb	98.7	98.0	0.7	0.8	29,497	91.7	59.3	32.4	6.3
アカムツ	725.6Mb	32.0Mb	98.9	98.0	0.9	0.8	25,814	98.1	97.3	0.8	1.1

[種名] ホンソウワケベラ　：*Labroides dimidiatus*[54]　スナウナギ　：*Hyperoplus lanceolatus*[57]
ダーター　：*Etheostoma perlongum*[55]　ケツギョ　：*Siniperca chuatsi*[58]
フォークランドアイナメ　：*Eleginops maclovinus*[56]　コウライケツギョ　：*Siniperca scherzeri*[58]

［BUSCO 分類記号］C:“Complete BUSCOs” OrthoDB の遺伝子セット中で、ほぼ完全な形で同定できる遺伝子
S :“Complete and single-copy BUSCOs” C のうち、単一コピーであるもの
D:“Complete and duplicated BUSCOs” C のうち、重複しているもの
M:“Missing BUSCOs” BUSCO の遺伝子セット中で、同定できなかった遺伝子

まず、いずれの魚種においても、アセンブル結果の合計長と N50 は共に似た値となっており、染色体レベルでのゲノム配列が構築されていることを確認できる。また、ゲノム配列に対する BUSCO による評価においても、Complete BUSCOs（以下、“C”と呼ぶ。）が 96.7〜98.7%といずれも非常に高く、Complete and duplicated BUSCOs（以下、“D”と呼ぶ。）も 0.6〜1.1%と少ないことから、高い網羅性のハプロイドゲノムが構築できていることがわかる（評価指標の分類記号については表 32 の欄外を参照）。

一方、遺伝子構造予測結果では、遺伝子構造数だけで見ても 22,274〜29,497 個と幅が出ている。protein モードでの BUSCO による評価は、C が 81.0〜94.9%と低い値にとどまっている種もあり、D においては 30%を超えている結果が散見された。genome モードでの評価では D が低いことから考えて、ゲノム重複が起こっているとは考えづらく、アノテーションの精度が低いと判断することが妥当であると考えられる。genome モードと protein モードとで、得られた C と D の値を比較すると、protein モードでの C が低いことや D が高いことから、ゲノム上には存在しているものの、遺伝子構造予測では一部を予測から漏らした上に、余分なものを予測してしまっている状態にあることが見てとれる。一方で、アカムツでは genome モードでの C が 98.9%、protein モードでのそれが 98.1%と僅かに一部の予測漏れは想定されるものの、D のパーセンテージはほぼ同じであり、非常に高精度に遺伝子を予測できていることが推察される。他魚種の遺伝子構想予測は、MAKER、BRAKER、EVM あるいは研究プロジェクトに独自のツール等を用いて行われているが、いずれもアカムツに対して GINGER を用いて予測した結果よりも劣っており、大きく劣後するケースも散見された。

さらに、アカムツの遺伝子予測に用いた近縁種に目を向けると、たとえばオオクチバスのゲノム決定論文では、著者らによる独自の遺伝子構造アノテーション結果が示されている [44]（遺伝子構造アノテーションには GLEAN[59] が用いられている）。この論文において報告されている protein モードによる BUSCO の評価結果は、C: 89.2% (S: 82.9%, D 6.3%), F: 7.1%, M: 3.7%というものであった。一方で、公開されているオオクチバスのゲノム配列に対して BUSCO の genome モードを適用したところ、C: 97.9% (S: 91.9%, D: 6.0%), F: 0.7%, M: 1.4%と網羅性が高い結果が得られた。また、NCBI RefSeq から提供されているタンパク質配列に対して protein モードを適用したところ、C: 98.8% (S: 92.3%, D: 6.5%), F: 0.7%, M: 0.5%と更に網羅性が高い結果となった。このことからも、研究者自身が精度の高い遺伝子構造アノテーション情報を得ることがいかに困難であるかがわかる。

3.7. 考察

GINGER をアカムツのゲノム配列データに適用することによって、BUSCO による評価において網羅性の高いことが示された遺伝子セットを構築することができた。GINGER の遺伝子構造予測結果に基づくオーソログ解析から、アミノ酸やペプチドの合成、アミノ酸の輸送、コラーゲン、短鎖不飽和脂肪酸の合成に関する遺伝子がアカムツで増えた可能性が示唆された。もちろん、遺伝子が増えていることを確認するにはより精緻な解析が必要であり、遺伝子の増減だけでアカムツの食味を説明しきれるわけではないが、より詳細な解析、あるいは他魚種も含めた二次解析のための基本的なデータセットを GINGER によって用意できたと考えられる。

今回の解析で使用した、キチヌ、ストライプドバス、コクチバス、オオクチバスはゼブラフィッシュと比較すればアカムツに近い種ではあるものの、魚種全体でみればアカムツからは遠い系統のものであると考えられる。アカムツにおける遺伝子重複や消失を同定していくためには、より近縁種のデータセットが必要であると考えられる。食味が良くなるためのメカニズムを理解するには、生息環境に類似点がありかつ食味の良い近縁のシロムツや、分類学上の科がアカムツとは異なるムツ科に属するものではあるが、やはり生息環境に類似点があり、また同様に食味が良いクロムツのデータセットが望まれるところである。

上述の通り、アカムツゲノムに対しては網羅的な遺伝子構造予測結果を得られたという評価を得ることができた。その一方で、近年においてゲノム配列が決定された近縁種に対する調査から、精度の高い遺伝子構造アノテーション作業が間に合っていないという現状が明らかとなった。本研究では、2023 年になって論文にてゲノム配列および遺伝子構造アノテーションデータが公開された魚種を対象に、実際に、BUSUCO による網羅性評価の結果を活用して遺伝子構造アノテーションの精度を分析した。それらの遺伝子構造アノテーションデータの精度は、GINGER を用いたアカムツゲノムに対する遺伝子構造予測結果に対して大幅に劣後するものであった。一方で、NCBI における RefSeq プロジェクトでは網羅性の高い遺伝子構造アノテーションが公開されている。RefSeq における遺伝子構造アノテーションはその流れは公開されているものの、煩雑な手段を踏んでいる上に用いているツール群も公開されていない。このため、一般の研究者が容易に同程度の精度の遺伝子構造アノテーションを行うのは困難な現状がある。手法開発に関する考察（「2.5. 考察」）において強調した通り、GINGER は高い精度を有する遺伝子構造予測ツールであるだけでなく、導入と実行が非常に容易なツールである。上述の現状を踏まえると、より多くのゲノムプロジェクトで使ってもらえるように普及活動も必要であると考えられる。

4. 総論

従来から、ゲノム配列決定のコスト下落、計測時間の短縮は加速度的に進んでいることは周知のとおりであるが、ここ数年、PacBio HiFi などに代表されるロングリードシークエンス技術の急速な進展・普及により、一般的な 2 倍体ゲノムを持つ生物種についてはゲノム解読の難易度が大幅に低下しており、さらに Hi-C 法などと組み合わせることで染色体レベルのゲノム決定はもはや珍しくはなくなってきている。そのようなゲノム配列の再構築を実現するためのゲノムアセンブラも多く開発され、計算機リソースに関しても研究室レベルで保有可能なワークステーションで対応可能となった。そのため、高分子量の DNA が潤沢に採取可能な 1Gb 程度のゲノムサイズを持つ 2 倍体生物種が対象であれば、ゲノム配列決定のみについてはルーチンワークになりつつある。実際、染色体レベルのゲノム配列決定について数多くの論文が 2020 年代以降になって公開されている。

その一方で、真核生物種に対する遺伝子構造アノテーションは、いまだ解析工程全体の中でのボトルネックとなっている。近年になっても、論文上や著者から公共データベースに登録された遺伝子構造アノテーションにおいて、精度の低いものが散見される状況にある。

本研究では、真核生物種に対して適用可能な遺伝子構造予測手法を開発し、GINGER として解析パッケージを公開することができた。GINGER は、*ab initio*-based method や RNA-Seq-based method、Homology-based method といった個別の予測手法による遺伝子構造予測結果を統合し、より精度の高い結果を得る設計となっている。このような手法は「統合手法」というカテゴリに属するものであり、EVM や Maker といったツールが存在している。それらは比較的に精度が高いことから多くのゲノムプロジェクトで用いられているが十分なレベルには至っておらず、より高精度なツールが求められていた。本研究では、GINGER に EMV と Maker を加える形で比較ベンチマークテストを行ったところ、GINGER は他者を大幅に上回る精度を有していることが確認できた。加えて、GINGER は自動実行可能な形でシステム化されており、容易に実行可能なことも大きな利点である。自動的に設定される閾値を利用することで、最終的な遺伝子構造の出力までを安定的に高い精度で行うことができる。また準最適解の出力により、アイソフォームの遺伝子構造候補を提示することも可能である。さらに、GINGER は、ソースツリーはもちろん、Conda パッケージや Docker イメージの形でも配布されており、導入と実行が容易であり、高い精度での遺伝子構造予測を提供できることから幅広く様々なゲノムプロジェクトでの利用が期待される。

最後に、本研究では、デモンストレーションとして、新規に決定したアカムツのゲノム配列に対して GINGER による遺伝子構造予測を実施

した。この予測結果は、BUSCO を用いた評価により、2023 年度に論文公開された近縁種の遺伝子構造アノテーション結果との比較の中で、最も高い網羅性を持つことが示された。他のアノテーションは、それらの近縁種の研究に携わる研究者が自分たちのプロジェク

トにおいて最適と判断したプロトコルで実行したものである。その結果と比べて、GINGER を自動パラメータで実行した遺伝子構造予測結果が上回る網羅性を示したことは、GINGER のアルゴリズムが非モデル生物を対象としても機能し、GINGER の利用により、より多くの研究者が容易に高精度の遺伝子予測が可能であることを示している。また、このデモンストレーションによって遺伝子構造予測結果および比較解析結果が得られたことは、アカムツを理解するためのより詳細な解析への入り口までたどり着いたに過ぎないため、アカムツとより近縁な魚種や、食味が類似している魚種のゲノム配列を決定し、それらに対して GINGER を適用し詳細な比較解析を行うことで、アカムツの素晴らしい食味が明らかになることも合わせて期待したい。

参考文献

1. Berks, M. 1995, The C. elegans genome sequencing project. C. elegans Genome Mapping and Sequencing Consortium, Genome Res., 5, 99-104.
2. Adams, M. D., Celniker, S. E., Holt, R. A. et al. 2000, The genome sequence of Drosophila melanogaster, Science, 287, 2185-95.
3. Celniker, S. E., Wheeler, D. A., Kronmiller, B. et al. 2002, Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence, Genome Biol., 3, RESEARCH0079.
4. Venter, J.C. 2001, The sequence of the human genome, Science, 291, 1304-51.
5. International Human Genome Sequencing Consortium, 2004, Finishing the euchromatic sequence of the human genome, Nature, 431, 931-45.
6. Brenner, S.E. 1999, Errors in genome annotation, Trends Genet., 15, 132-3.
7. Schnoes, A.M., Brown, S.D., Dodevski, I. and Babbitt, P.C. 2009, Annotation error in public databases: misannotation of molecular function in enzyme superfamilies, PLoS Comput. Biol., 5, e1000605.
8. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, Nucleic Acids Res., 32, W309–12.
9. Korf, I. 2004, Gene finding in novel genomes, BMC Bioinformatics, 5, 59.
10. Besemer, J. and Borodovsky, M. 2005, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res., 33, W451–4.
11. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol., 33, 290-5.
12. Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. 2019, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. biotechnol., 37, 907-915.
13. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15-21.
14. Robertson, G., Schein, J., Chiu, R. et al. 2010, *De novo assembly-based method* assembly and analysis of RNA-seq data, Nat. Methods, **7**, 909–12.
15. Xie, Y., Wu, G., Tang, J. et al. 2014, SOAPde novo assembly-based method-Trans: de novo assembly-based method transcriptome assembly with short RNA-Seq reads, Bioinformatics, 30. 1660–6.
16. Grabherr, M. G., Haas, B. J., Yassour, M. et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol., 29, 644–52.

17. Schulz, M. H., Zerbino, D. R., Vingron, M. et al. 2012, *Oases*: robust *de novo assembly-based method* RNA-seq assembly across the dynamic range of expression levels, Bioinformatics, 28, 1086–92.
18. Huang, X., Adams, M. D., Zhou, H., & Kerlavage, A. R. 1997, A tool for analyzing and annotating genomic sequences. Genomics, 46, 37-45.
19. Birney, E., Clamp, M. and Durbin, R. 2004. GeneWise and Genomewise, Genome Res., 14, 988–95.
20. Slater, G. S. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, BMC Bioinformatics, 6, 31.
21. She, R., Chu, J. S., Uyar, B., Wang, J., Wang, K., & Chen, N. 2011, genBlastG: using BLAST searches to build homologous gene models, Bioinformatics, 27, 2141–3.
22. Gotoh O. 2008. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence, Nucleic Acids Res., 36, 2630–8.
23. Cantarel, B. L., Korf, I., Robb, S. M. et al. 2008, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Res., 18, 188–96.
24. Holt, C., and Yandell, M. 2011, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC Bioinformatics, 12, 491.
25. Haas, B. J., Salzberg, S. L., Zhu, W. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, Genome Biol., 9, R7.
26. Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., & Andorf, C. M. 2021, FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences, BMC Bioinformatics., 22, 205.
27. Gilbert, D. 2013, Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame.
28. Reese, M. G., Kulp, D., Tammana, H. and Haussler, D. 2000, Genie--gene finding in Drosophila melanogaster. Genome Res., 10, 529–38.
29. NCBI Resource Coordinators 2018, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 46, D8–13.
30. Li, W. and Godzik, A. 2006, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, 22, 1658–9.
31. Wu, T. D. and Watanabe, C. K. 2005, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics, 21, 1859–75.
32. Haas, B. J., TransDecoder at https://github.com/TransDecoder/TransDecoder
33. Smit, A. F. A., Hubley, R. and Green, P., RepeatMasker at http://repeatmasker.org/

34. Gotoh O. 2018, Modeling one thousand intron length distributions with fitild, Bioinformatics, 34, 3258–64.
35. Kawahara Y, de la Bastide M, Hamilton JP, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y). 2013;6(1):4. Published 2013 Feb 6.
36. Broughton, RE., Milam, JE., Roe, BA. The complete sequence of the zebrafish (Danio rerio) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome Res. 2001;11(11):1958-1967.
37. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:ISCB Comm J-304. Published 2020 Apr 28.
38. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W. et al. 2011, Integrative genomics viewer. Nat. Biotechnol., 29, 24–6.
39. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. 2017, Nextflow enables reproducible computational workflows, Nat. Biotechnol., 35, 316–9.
40. Lee, Du-Seok, Yoon, Ho Dong, KIM, Yeon Kye, Yoon, Na Young, Moon, Soo-Kyung, Kim, In-Soo, & Jeong, Bo-Young. (2011). Proximate and Fatty Acid Compositions of 14 Species of Coastal and Offshore Fishes in Korea. Korean Journal of Fisheries and Aquatic Sciences, 44(6), 569–576.
41. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722-736.
42. Freire B, Ladra S, Parama JR. Memory-Efficient Assembly Using Flye. IEEE/ACM Trans Comput Biol Bioinform. 2022;19(6):3564-3577.
43. Hu, J. et al. An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv 2023.03.09.531669 (2023)
44. Rhie, A., McCarthy SA, Fedrigo O, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737-746.
45. Baltzegar, DA., et al. Striped Bass Genome Project. Direct submission. 2019.4.26, https://www.ncbi.nlm.nih.gov/nuccore/SSXF00000000.1
46. Zarri, LJ., Reference genome of smallmouth bass (Micropterus dolomieu), assembled using PCR-free Illumina HiSeq 50x 150pe and anchored using largemouth bass (Micropterus salmoides) genome Direct submission. 2022.1.11, https://www.ncbi.nlm.nih.gov/nuccore/JAJNER000000000.1
47. Sun C, Li J, Dong J, et al. Chromosome-level genome assembly for the largemouth bass Micropterus salmoides provides insights into adaptation to fresh and brackish water. Mol Ecol Resour. 2021;21(1):301-315.

48. Simão, FA., Waterhouse, RM., Ioannidis, P., Kriventseva, EV., Zdobnov, EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-3212.
49. Cosentino, S., & Iwasaki, W. (2019). SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics (Oxford, England), 35(1), 149–151.
50. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772–780.
51. Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England), 25(15), 1972–1973.
52. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular biology and evolution, 37(5), 1530–1534.
53. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127-128.
Kang, J., Ramirez-Calero, S., Paula, J. R., Chen, Y., & Schunter, C. (2023). Gene losses, parallel evolution and heightened expression confer adaptations to dedicated cleaning behaviour. BMC biology, 21(1), 180.
54. Kang, J., Ramirez-Calero, S., Paula, J. R., Chen, Y., & Schunter, C. (2023). Gene losses, parallel evolution and heightened expression confer adaptations to dedicated cleaning behaviour. BMC biology, 21(1), 180.
55. MacGuigan, D. J., Krabbenhoft, T. J., Harrington, R. C., Wainwright, D. K., Backenstose, N. J. C., & Near, T. J. (2023). Lacustrine speciation associated with chromosomal inversion in a lineage of riverine fishes. Evolution; international journal of organic evolution, 77(7), 1505–1521.
56. Cheng, C. C., Rivera-Colón, A. G., Minhas, B. F., Wilson, L., Rayamajhi, N., Vargas-Chacoff, L., & Catchen, J. M. (2023). Chromosome-Level Genome Assembly and Circadian Gene Repertoire of the Patagonia Blennie Eleginops maclovinus-The Closest Ancestral Proxy of Antarctic Cryonotothenioids. Genes, 14(6), 1196.
57. Winter, S., de Raad, J., Wolf, M., Coimbra, R. T. F., de Jong, M. J., Schöneberg, Y., Christoph, M., von Klopotek, H., Bach, K., Pashm Foroush, B., Hanack, W., Kauffeldt, A. H., Milz, T., Ngetich, E. K., Wenz, C., Sonnewald, M., Nilsson, M. A., & Janke, A. (2023). A chromosome-scale reference genome assembly of the great sand eel, Hyperoplus lanceolatus. The Journal of heredity, 114(2), 189–194.

58. Tu, G. X., Zhang, X. S., Jiang, R. R., Zhang, L., Lai, C. J., Yan, Z. Y., Lv, Y. R., Weng, S. P., Zhang, L., He, J. G., & Wang, M. (2023). Long-read genome assemblies reveal a cis-regulatory landscape associated with phenotypic divergence in two sister Siniperca fish species. Zoological research, 44(2), 287–302.
59. Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., & Weinstock, G. M. (2007). Creating a honey bee consensus gene set. Genome biology, 8(1), R13.

付録

A　EVM の重み係数に関する検討

「2.3. ベンチマーク」のベンチマークにおいて、EVM による遺伝子構造予測結果を得るにあたって、以下に示す重みで得られる Sn と Pr を事前に取得した（表 33)。

（ア) EVM についての論文で推奨されている値（表 33 の表側「論文推奨値」)

Genome-guided assembly-based method　：　1.0
de novo assembly-based method　：　10.0
Augustus　：　0.3
SNAP　：　0.3
Homology-based method　：　1.0

（イ) GINGER のデフォルト値を参考にした重み（表 33 の表側「GINGER デフォルト値参照」)

Genome-guided assembly-based method　：　10.0
de novo assembly-based method　：　4.0
Augustus　：　9.0
SNAP　：　4.0
Homology-based method　：　8.0

(ア)の方が、塩基レベル、エクソンレベルおよび遺伝子レベルの Sn が高く、さらに遺伝子レベルの Pr が高かったため、(ア)を本ベンチマークに使用した。

表 33　EVM に対するベンチマークに用いた重み係数

	参照遺伝子構造(#)	予測遺伝子構造(#)	塩基レベル			エクソンレベル			遺伝子レベル		
			Sn	Pr	F-score	Sn	Pr	F-score	Sn	Pr	F-score
線虫 (*Caenorhabditis elegans*) 100.3Mb											
論文推奨値	19,984	17,292	87.01	92.22	89.54	79.42	82.34	80.85	41.71	48.20	44.72
GINGER デフォルト値参照		16,803	88.89	92.12	90.48	80.49	83.88	82.15	42.58	50.64	46.26
ハエ (*Drosophila melanogaster*) 143.7 Mb											
論文推奨値	13,950	11,753	85.74	95.90	90.53	75.05	78.91	76.93	48.90	58.04	53.08
GINGER デフォルト値参照		11,843	85.75	95.65	90.43	74.97	78.54	76.72	49.12	57.87	53.10

B　TransDecoder の影響の検討

EVM には転写産物をゲノム配列に対してアライメントした結果を入力とすることは可能であるが、通常の読み込ませ方では、アライメント領域をエクソン候補領域として（つまり、エクソン・イントロン構造に関する情報を損失した形で）取り扱うことになる。一方、GINGER の RNA-Seq-based method では、各ツールの予測結果に対して TransDecoder を適用し、予測された ORF 情報に基づき遺伝子構造予測結果を入力としている。この ORF 予測結果に含まれているエクソン・イントロン構造に関する情報を残したまま EVM に読み込ませることで精度が向上するかを確認した。具体的には、EVM に読み込ませる GFF の source 列を“OTHER_PREDICTION”とし、attributes 列で同じ遺伝子に紐づけられるエクソンを指定した上で読み込ませた。その結果、Se、Pr、F-score の全ての値において向上が見受けられた（表 34）。このことから、GINGER の標準プロトコルである TransDecoder を活用した RNA-Seq-based method の扱いが予測精度向上に寄与していると考えられる。

表 34　TransDecoder によるエクソン・イントロン構造の EVM に対する影響

	参照遺伝子構造(#)	予測遺伝子構造(#)	塩基レベル			エクソンレベル			遺伝子レベル		
			Sn	Pr	F-score	Sn	Pr	F-score	Sn	Pr	F-score
線虫 (*Caenorhabditis elegans*) 100.3Mb											
EVM	19,984	16,803	88.89	92.12	90.48	80.49	83.88	82.15	42.58	50.64	46.26
EVM (TransDecoder)		14,418	80.46	92.52	86.07	68.99	79.22	73.75	29.25	40.55	33.99
ハエ (*Drosophila melanogaster*) 143.7 Mb											
EVM	13,950	11,843	85.75	95.65	90.43	74.97	78.54	76.72	49.12	57.87	53.10
EVM (TransDecoder)		11,571	85.51	96.85	90.83	73.10	76.48	74.75	46.25	55.78	50.57
イネ (*Oryza sativa*) 374.4 Mb											
EVM	28,556	39,980	91.45	79.27	84.93	81.05	69.73	74.96	49.18	35.33	41.12
EVM (TransDecoder)		36,302	91.47	84.18	87.67	83.51	76.42	79.81	57.91	45.82	51.16
ゼブラフィッシュ (*Danio rerio*) 1,373.5 Mb											
EVM	25,526	31,200	87.14	83.40	85.23	82.69	75.55	78.96	30.05	24.60	27.06
EVM (TransDecoder)		28,930	84.93	85.24	85.09	79.61	73.57	76.47	26.19	23.12	24.56
ヒト (*Homo sapiens*) 3,099.4 Mb											
EVM	21,670	25,686	84.88	81.63	83.22	78.39	76.02	77.19	22.66	19.10	20.73
EVM (TransDecoder)		22,480	81.61	84.48	83.02	76.59	76.93	76.76	19.10	18.39	18.74