様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１４６２４

研究課題名（和文）ハプロタイプ配列群に基づくアレル特異的ゲノム動態解析手法の開発

研究課題名（英文）Development of an analysis pipeline for allele-specific genome dynamics based on haplotype sequences

研究代表者
梶谷　嶺（Kajitani, Rei）

東京工業大学・生命理工学院・助教

研究者番号：４０７５６７０６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：多倍体生物のゲノムにおいて、片方のアレルからのみ遺伝子の発現が観察される現象については、X染色体不活性化や哺乳類におけるゲノムインプリンティング等が長く研究対象となってきており、疾患との関連も報告され生命活動の維持に重要であることが示されている。本研究の具体的な目的は、汎用的なアレル特異的現象のin silico解析パイプラインを開発することであり、単一の代表ゲノム配列に依存せず、各ハプロタイプ配列を別々に構築し、それぞれに特異的な遺伝子構造や発現等の動態を解析するツールを開発した。

研究成果の学術的意義や社会的意義
各ハプロタイプ配列を出力するツールは研究開始時点でも存在していたが、これらツールの配列を直接入力可能な既存解析ツールは存在していなかった。この状況の要因としては、各ハプロタイプの構築手法は１分子DNAシークエンサー等の新技術を活用することで近年にようやく実用性を帯びてきた事情が挙げられる。近年はオミクス解析の主流が、単一の代表ゲノム配列に依存したプロトコルから、ハプロタイプ配列の総体を「ゲノム」として扱うものへと移りつつあり、本研究の成果は関連研究の推進の一助になると期待される。

研究成果の概要（英文）：Phenomena such as X-chromosome inactivation and imprinting in polyploid genomes have been studied. Gene expression is sometimes observed only from one allele, and have been reported to be associated with disease and shown to be important for phenotypic traits. In this study, I developed an in-silico analysis pipeline for allele-specific genomic dynamics, which is expected to be effective to analyze the gene structures and gene expressions specific to each haplotype, without relying on a single representative reference genome.

研究分野：バイオインフォマティクス

キーワード：ゲノム　ハプロタイプ　遺伝子構造アノテーション

１．研究開始当初の背景

多倍体生物のゲノムにおいて、片方のアレルからのみ遺伝子の発現が観察される現象については、X 染色体不活性化や哺乳類におけるゲノムインプリンティング等が長く研究対象となってきており、疾患との関連も報告され生命活動の維持に重要であることが示されている（総説 Reik & Walter, Nat. Review Genetics 2001 等）。これらが初期の研究対象となった要因としては、顕微鏡下での染色体形態の大幅な変化や、片方のアレルの発現量がほぼゼロとなる性質等の観察が容易な条件が伴っていたことが挙げられる。近年は超並列 DNA シークエンサーの普及により網羅的に遺伝子発現やエピジェティックな現象を観察することが可能となってきており、ヒトやマウスで数多くのアレル特異的な遺伝子発現（Andergassen et al, eLIFE 2017 等)、ゲノム DNA メチル化（Shoemaker et al, Genome Res. 2010 等）、ヒストン修飾(Mikkelesen et al. Nature 2010 等）が報告されている。1000 Genomes Project （1000 Genomes Project Consortium, Nature 2015）をはじめとする大規模な集団ゲノムデータ収集計画によって変異（variants）のカタログ化も進められ、アレル特異的な現象の全容解明に向けては好条件が整いつつある。しかしながら、変異データを精査することによってそれまでは認知されていなかった問題も明らかになっている。ヒトのシークエンシングデータを *de novo* アセンブリ(断片配列からの新規配列構築手法)を用いて再調査したところ、種内のハプロタイプの多様性はそれまでの推定値より大きく、僅か 25%しか標準的な変異解析手法では検出できていないという報告がなされており（Weisenfeld et al, Nat. Genetics. 2014）、その後も *de novo* アセンブリによって多くの変異を新たに発見したという報告が複数存在している（Pendleton et al, Nat. Methods 2015; Seo et al, Nature 2016; Maretty et al, Nature 2017）。ここでの「標準的な変異解析手法」とは単一の参照ゲノム配列に DNA シークエンサーの出力する断片配列（リード）をマップする手法を指しているが、重要な問題点となるのは、トランスクリプトーム解析（RNA-seq）やエピジェネティクス解析（BS-seq, ChIP-seq）の手法のほぼ全てが同様に単一の参照ゲノム配列のみを活用している状況であり、真の網羅的解析には複数のハプロタイプ配列を活用する新規の解析手法を開発することが不可欠である。

本研究の学術的な問いとしては「アレル特異的な現象はどの程度厳密に制御され、どのような利益を個体にもたらすのか」という疑問が挙げられる。多倍体という性質を獲得したことは生物の進化史においては重大なイベントであるが、それによって得た利点は、集団内でゲノムの多様性を得ることと個体内でゲノムのバックアップを確保することだけであろうか。代表者は、多倍体化によって個体内でのハプロタイプの「分業体制」が実現され、より多様な制御が可能になった結果、多倍体を含む系統での複雑な形態が実現された、という仮説を立てている。仮説の検証は、多く報告されているヘテロ接合体の優位性（heterozygote advantage） （総説 Hedrick, Trends Ecol. Evol. 2012）の理解の一助になることも期待される。

２．研究の目的

本研究の具体的な目的は、汎用的なアレル特異的現象の *in silico* 解析パイプラインを開発することである。入力データは参照配列としての複数ハプロタイプ配列および超並列シークエンサーのリードであり、発現量やエピジェネティクス情報等の統計量とハプロタイプ配列の多様性の情報を同時に扱うことで、アレル特異的ゲノム動態の新知見獲得を目指す。

純粋な各ハプロタイプ配列を出力するツールは研究開始時点でも存在しているが（Chin et al, Nat. Methods 2016; Weisenfeld et al, Genome Res. 2017）、これらツールの配列を直接入力可能な既存解析ツールは存在していないと考えられる。この状況の要因としては、各ハプロタイプの構築手法は１分子 DNA シークエンサー等の新技術を活用することで近年にようやく実用性を帯びてきた事情が挙げられる。オミクス解析の主流が単一の代表ゲノム配列に依存したプロトコルが、ハプロタイプ配列の総体を「ゲノム」として扱うものへと移りつつあると代表者は推察している。本研究では多倍体ゲノムにおけるアレル（ハプロタイプ）特異的事象の解析のためのソフトウェアの開発が主目的となっており、その実行に際してはハプロタイプ配列を個別に決定して構築した参照配列と塩基配列シークエンサーのデータ（RNA-seq 等）の入力を想定している。また、各ハプロタイプの遺伝子発現量等の値と分子メカニズムの知見を結びつけるため、参照配列のアノテーション（遺伝子予測を含む注釈付け）の手法の開発も行う。

３．研究の方法

初年度はテスト用データを取得するための体制の構築と参照ゲノム配列アノテーションのソフトウェア開発を主に行った。本研究の前段階として各ハプロタイプ配列の決定が必要となり、当機能を担うツール：Platanus-allee は代表者を含むグループが開発と論文発表（Kajitani et al. 2019）を行ったが、これは別研究課題の対象である点は注記する。データ取得体制の構築としては、頭索動物や棘皮動物の豊富なデータを持つグループと共同研究の機会を得た。なお、開

発全般は所属研究室の大学院生の協力を得て行われた。

次年度では引き続き、２倍体ゲノム中のハプロタイプを分けて構築された配列セットを用いて、アレル特異的現象を検出するためのパイプラインの開発を行った。方法としては、ハプロタイプ配列セットを参照配列として、転写産物を対象とした RNA-seq や染色体の立体構造情報を捉えるための Hi-C といった技術の結果をマップすることが基本戦略となる。ハプロタイプ配列セットの構築には、前年度に代表者を含むグループで開発されたソフトウェアである Platanus-allee（Kajitani et al. 2019; 別研究課題の対象）を用いる方針とした。本年度は新たに Hi-C のデータを取得し、所属研究室の大学院生の協力を得て対応するソフトウェア開発を行った。このデータを用いると、本来の用途である染色体立体構造の推定のみならず、より長いハプロタイプ配列の再構築も同時に行うことができると期待される。

最終年度は、前年度までに開発された遺伝子構造アノテーションツールを活用しつつ、RNA-seq やゲノム DNA の立体構造を捉える手法である Hi-C 法のデータのマップによるアレル特異的現象の検出パイプラインの開発を行った。解析の障害となる配列ギャップに対しては、アノテーションツールによる対処の他に、前段階でのロングリードの活用によるギャップの削減も同時に検討した。また、Hi-C データを、ハプロタイプ配列構築時に使われるグラフ構造の単純化や経路探索に用いるとこにより長いハプロタイプ配列を得るソフトウェアの開発も行う予定である。テストデータとしては、共同研究の機会を得ている、頭索動物や棘皮動物を高ヘテロ接合性サンプルとして用いた。

４．研究成果

初年度に行った参照配列の遺伝子構造アノテーションパイプライン開発に関しては、当初は既存ツールの MAKER（Holt et al. 2011）等を組み込んでハプロタイプ毎のアノテーションを行うパイプラインの構築を予定していたが、容易であると考えられたハプロタイプのコンセンサス配列（疑似的な１倍体ゲノム配列）を対象としても、既存ツールの遺伝子領域予測精度が低いケースが多いという問題が浮上した。そのため、代表者を含むグループで新規ツールを開発し精度を向上させた（篠田ら、第 43 回日本分子生物学会年会、2018）。

構築したハプロタイプ配列のギャップ（配列不明領域）の影響の把握と対処に時間を費やしたことが挙げられる。Platnaus-allee はエラー率の低いショートリード（DNA シークエンサーの結果配列のうち、300 bp 以下の短いもの）を活用したときに高性能を発揮するが、構築された配列中に多くのギャップが入ることになる。それらギャップがパイプライン中の遺伝子構造アノテーション等に悪影響を及ぼすことが判明し、その対処法を代表者を含むグループで開発した（中村ら、第８回生命医薬情報学連合大会、2019）。

最終年度は、当初の目標であった各ハプロタイプに特異的な立体構造の特徴を明確に検出することはできなかったが、副次的な成果としてハプロタイプ配列を染色体スケールで決定する手法を所属研究室の大学院生と共同で開発することができた（大内ら、日本動物学会第 91 回大会）。ここでは、以前に開発されたハプロタイプ配列構築ツール: Platanus-allee の途中結果である scaffold グラフと呼ばれるデータ構造上に Hi-C データをマップし、同ハプロタイプ上の領域は空間的にも近接する頻度が大きいことを利用してその配列の決定（phasing）を行う機能を実装した。構築された配列のエラーを Hi-C コンタクトマップを利用して修正する機能も追加することで、大きく配列長を向上させることにも成功している。前年度までに開発された遺伝子構造アノテーションツールも、各ハプロタイプが分けて構築された配列セットに対してのテストを所属研究室の大学院生と共同で実施した。対象サンプルとしてはヘテロ接合性が極めて高い頭索動物を選択した。その際、最初に片方のハプロタイプ配列セットに遺伝子構造アノテーションを実施し、その結果のエクソン-イントロン構造をもう片方の配列セットにアライメントを行い、更にタンパク質コード領域の完成度（インフレームストップコドンの有無など）を確認した。アライメントは複数ツールの結果を統合することで性能を向上させている。結果として 75% 以上の遺伝子を両ハプロタイプで全長を構築することができた。このパイプラインは、両ハプロタイプが分けて構築された配列セットに対してのアノテーションツールとして新規性を有する。開発されたツールは、本研究課題の対象ではないものの軟体動物（Inoue et al, Sci. Rep. 2021）や棘皮動物（Yuasa, Kajitani et al, 査読中）の全ゲノム配列解析論文にも活用された。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 中村 優太、梶谷 嶺、小林 史弥、湯淺 英知、伊藤 武彦
２．発表標題 ドラフトゲノムに対応した遺伝子構造アノテーション自動構築パイプラインの開発
３．学会等名 第８回生命医薬情報学連合大会
４．発表年 2019年

１．発表者名 小林 史弥、梶谷 嶺、中村 優太、奥野 未来、湯淺 英知、伊藤 武彦
２．発表標題 ドラフトゲノムの不完全性が遺伝子構造アノテーションに及ぼす影響の解析
３．学会等名 第８回生命医薬情報学連合大会
４．発表年 2019年

１．発表者名 篠田 恭寛、梶谷 嶺、小林 史弥、髙橋 和希、中村 優太、湯淺 英知、伊藤 武彦
２．発表標題 網羅的遺伝子構造予測自動アノテーションパイプラインの開発
３．学会等名 第41回日本分子生物学会年会
４．発表年 2018年

１．発表者名 大内 俊、梶谷 嶺、伊藤 武彦
２．発表標題 Hi-C法による染色体レベルのscaffolding、ハプロタイプ構築手法の開発と高ヘテロ接合性サンプル（イトマキヒトデ）への応用
３．学会等名 日本動物学会第91回大会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関