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　４ 年　 ６ 月 １５ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０３２０６

研究課題名（和文）NGSデータからの相同組換え位置特定アルゴリズムの開発および遺伝情報との関連解析

研究課題名（英文）Development of an algorithm for determining the location of homologous recombination sites from NGS data and analysis of the relationship with genetic information

研究代表者
伊藤　武彦（Takehiko, Itoh）

東京工業大学・生命理工学院・教授

研究者番号：９０５０１１０６

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：本研究では、各種真核生物において相同組換えの位置を網羅的に特定し、その結果を基盤とした相同組換えに関する新たな研究展開を図ることを目的としている。この目的実現のため、Illumina pair-end, mate-pairリードおよびPacBio HiFiを入力とした相同組換え位置特定プログラムの開発を実施した。またこのプログラムを用いて、マウスF1 (B6 x CAST)およびイトマキヒトデのpoolされた精子のゲノムシークエンスデータを解析し、ゲノムワイドな組換え候補位置の抽出を行い、その頻度・分布などの解析を実施した。

研究成果の学術的意義や社会的意義
本研究を通じて開発されたプログラムの使用により、bulkで取られた精子などのシークエンスデータに基づいて比較的容易にゲノムワイドな組換え位置、頻度情報が得られることが期待される。ここ1-2年で同様な解析がSingle-Cellベースで行われている事例が報告されているが、そのような方法と比べて圧倒的に簡便かつ網羅性が高いデータが得られる。今後は、相同組換えに関与するタンパク質の局在情報などと合わせて解析にしていくことで、減数分裂時の組換えの理解が一層進むことが期待される。

研究成果の概要（英文）：The objective of this research is to comprehensively identify the location of homologous recombination sites in various eukaryote genomes. To achieve this goal, we developed a novel program that detects homologous recombination sites using Illumina pair-end, mate-pair reads and PacBio HiFi as input. Using this program, we analyzed genome sequencing data of pooled sperm from mouse F1 (B6 x CAST) and from starfish, extracted genome-wide candidate positions for recombination, and analyzed their frequency and distribution.

研究分野：ゲノム情報解析

キーワード：ゲノム情報解析　相同組換え

１．研究開始当初の背景

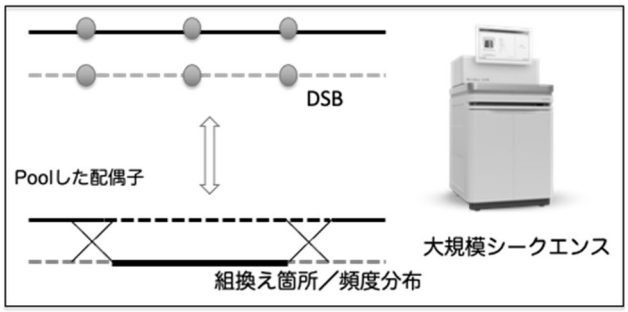
地球上の多くの生物種は、DNA を遺伝情報の担い手として生存し、そのゲノム DNA は世代間継承の過程において微小な変異を積み重ね、これが進化の原動力となり種々の生物が誕生したと考えられている。中でも真核生物においては、減数第一分裂時に引き起こされる相同組換えによって二個体のゲノムが交換され、DNA 配列の変動に大きな役割を果たす。このため、対立遺伝子を用いた連鎖・組換え頻度の研究により、遺伝地図がモーガンらによって 100 年以上も昔に作成されるなどその研究は極めて古い。この数十年では、分子生物学の発展に伴い、メカニズムを明らかにすべくどのようなタンパク質がどのように働いているのか、相同組換え時のゲノム DNA／タンパク質複合体の構造はどのようになっているのかといった様々な視点からの研究が盛んに行なわれている。これにより、Rad51, Dmc1 といったタンパク質の関与や二本鎖切断(DSB)とその修復メカニズムなどが次々と明らかになっている。

一方、次世代シークエンサ(NGS)とりわけ HiSeq の登場により、100-250bp 程度の短い配列ではあるものの、一度の稼働で数百 Gb にも及ぶ塩基配列が安価に産出される時代になり、変異解析、RNA-seq さらには様々な生物種の新規ゲノム配列決定にも適用されている。また、目的タンパク質に特異的な抗体を利用することで、タンパク質のゲノム上での局在を特定する ChIP-seq 法なども開発・普及し、上述した相同組替えに関わるタンパク質のゲノム上での動態も明らかになりつつある。さらには DSB による非相同末端結合(NHEJ)時に、二本鎖オリゴデオキシヌクレオチドを取り込ませて、その取り込む位置を検出する Guide-seq と呼ばれる手法により、DSB の位置をゲノムワイドに網羅的に明らかにする手法なども開発されている。

このような分子生物学の発展に伴い、相同組換えに関与するタンパク質、DSB などの知見が明らかになりゲノム上での局在情報などは知ることができるようになってきている、また、遺伝情報そのものであるゲノム DNA 配列も解読されているが、ゲノム中のどこで・どれくらいの頻度で相同組換えを起こしているのか、塩基レベルで網羅的に明らかにすることはほとんどできていない状況であった。

２．研究の目的

以上の背景を踏まえ本研究では、各種真核生物において相同組換えの位置をシークエンスデータから網羅的に特定する手法を新規に開発することが最大の目的となる。この目的が達成できれば、その結果を基盤とした相同組換えに関する新たな研究展開を図ることが大いに期待できる。具体的には、第一にゲノム上で相同組換えの位置を塩基レベルで網羅的に特定する新規解析手法の開発、第二に開発した手法を用いたモデル／非モデル生物種における相同組換え位置の網羅的な特定を進める。これらの目標を達成することで、有性生殖により父方・母方それぞれの遺伝情報がどのように子孫に引き継がれていくのかを網羅的に解析し、遺伝情報の継承に関する全ゲノムにわたる普遍性と組換え抑制によるゲノム領域に依存した特異性とを明らかにすることが今後の研究で期待される。



本研究の最大の特徴は、配偶子のゲノムをプールした状態でシークエンサにより解読し、減数分裂時に生じた組換え位置を直接網羅的に決定することを目指していることである。従前の研究では、ヒトなど特定な種において、膨大な個体数の対立遺伝子を調べることで連鎖状況から組換え箇所／頻度を推定したり、組換え時に起きる DSB 箇所から推定したりといった間接的な手法しか取り得なかった。

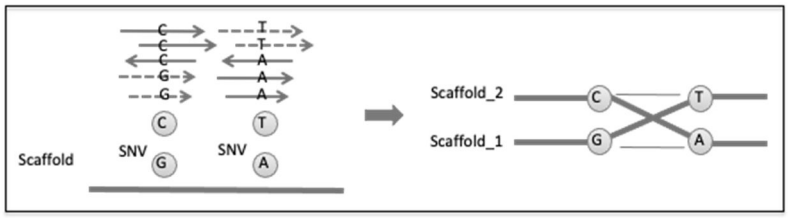
しかし本提案は、全ゲノムショットガンデータから相同染色体を区別したアセンブル結果と共に、組換え箇所を網羅的に出力する解析プログラムを開発し、Illumina シークエンサの精度の高さと相同染色体間の差異を有効に活用することで、塩基レベルでの精度で網羅的に組換え位置を特定することを可能にすることを目指す、従来の研究とは一線を画すものである。

網羅的な塩基レベルでの組換え位置の情報は、組換えに関与するタンパク質、DSB の局在情報などと比較することで、局在箇所と実組換え位置の関連性などの研究が初めて実現できる。同様に子孫のゲノム情報と比較することで、網羅的な致死アレルの検出なども可能になる。また組換え抑制が起きているゲノム領域の特定も可能になり、これらの領域は変異が蓄積されるため種分化を産み出す原動力となり得ることから、種の進化研究への発展も大いに期待できる。

３．研究の方法

本研究ではまず、pool された配偶子のゲノムを対象とした NGS 全ゲノムショットガンデータから、網羅的な組換え位置検出プログラムの開発を実施する。今まで開発してきているアセンブラ(Platanus, Platanus_allee)では、Illumina シークエンサの高いシークエンス精度を活用し、diploid

ゲノムからハプロタイプ別のゲノム配列構築に成功している。これらのプログラムでは、①k-mer と呼ばれる部分文字列をノードとした graph 構造を解き、②リードの pair リンク情報を用いて連鎖を解決することによってハプロタイプごとのゲノム構築アセンブルを実現している。この際に、プログラム内部では、diploid を仮定しているため、シークエンスカバレッジが一定頻度よりも低い k-mer, pair リンクはシークエンスエラーに起因するとし、アセンブルには用いられていない。しかし、pool された配偶子のゲノムをシークエンスした場合、アセンブルアルゴリズム時には用いられていないデータとして、組換えにより生じたパスが存在する筈である。この頻度は極めて小さいことが予想されるが、あらかじめ（組換えを生じていないであろう）大多数のデータにより対立アレルを検出しておくことで、組換えにより生じることが期待されるリード配列を予測することが可能となるため、組換え由来のリードとシークエンスエラー由来のリードとの区別が可能となる。この原理を組み込むことで、減数分裂により生じる頻度の低い組換え由来のリードを抽出し、そのデータに基づき、組換え位置、頻度の特定を行うアルゴリズムを開発する。



また、頻度高く組換わるホットスポットなどが存在した場合には、pool された配偶子のゲノムのみからは、組換わったパスか親が持つパスなのかを明らかにすることはできない。このため、親の体細胞由来のゲノムが得られる場合には、比較することで、配偶子形成時の減数分裂で生じる相同組み換え位置、および頻度を網羅的に出力するプログラムを開発する。最後に上記で開発したプログラムを実サンプルに適用することで、網羅的な相同組換え情報の取得を実施する。

４．研究成果

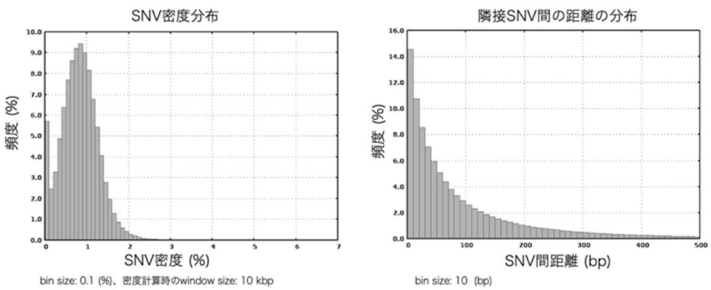
本研究では、上記方法にて記載した内容に基づき組換え位置同定プログラムを開発し、そのプログラムを pool した配偶子からのシークエンスデータに適用することで、全ゲノム中での組換え頻度や組換え位置の分布などを導出した。開発した組換え位置同定プログラムでは、pool した配偶子より得られた DNA からシークエンスされた Illumina pair-end, mate-pair に加え、PacBio HiFi リードに対応している。また、親個体のゲノムもしくは親個体同士の SNV 情報が得られている場合には、その情報も加えることで可能となっている。これにより、偽陽性が下げられることが期待される。本章では、①F1 マウス(B6 x CAST)の精子、②野生イトマキヒトデの精子から得られた Illumina pair-end データを用いて解析した結果を示す。なお①は、Hinch et al. 2019 にて公開されているデータであり、論文中では Single-Cell からの DNA 増幅法を開発し、217 細胞へ適用することで細胞ごとの組換え位置の特定を試みている。一方、本研究では同 F1 マウスの精子を pool して Illumina HiSeq2500 でシークエンスしたデータを用いている。②に関しては、野生イトマキヒトデ 1 個体から採取した精子を pool して Illumina HiSeq2500 でシークエンスしたデータを用いている。

1.) F1 マウス(B6 x CAST)の解析

Hinch et al. 2019 にて公開されている F1 マウス(B6 x CAST)の精子 bulk シークエンシングデータ (Illumina 150bp pair-end, 合計約 100Gb)を用いて解析を実施した。解析に当たっては参照ゲノムとして GRCm38 (mm10)を利用し、親個体のゲノム情報として MGI にて公開されている B6, CAST 間の SNV 情報も併せて利用した。

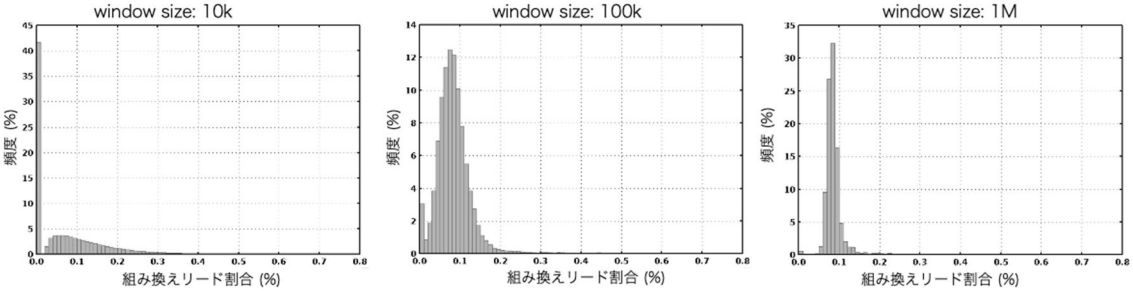
<u>B6 x CAST 間で認められる SNV</u>

まず、Illumina リードを bwa-mem を用いて参照ゲノムにマッピングし、identity, alignment coverage, 推定 insert size などの情報でフィルタ後、GATK HaplotypeCaller で SMV コールを行なった。この際にも QV や coverage depth による厳格なフィルタリングを実施した。さらにこのデータでは、親が持つ SNV 情報によるフィルタリングも実施した。その結果、マッピングのみで 19,106,423 箇所の SNV を、親情報によるフィルタリングで 18,013,608 箇所のより確実な SNV をコールすることができた。得られた SNV の密度および SNV 間距離の分布は右の通りである。この結果より、SNV 間の距離は Illumina pair-end データを用いても insert 距離内にほとんどが収まっていることが確認できた。



B6 x CAST 間で認められる組換え頻度

次に隣接する SNV 間で(頻度の低いパス)/(頻度の高いパス)を計算することにより、組換えリードの割合を算出し、ゲノムを 10kb, 100kb, 1Mb の window に分割した上で window 内の組換え頻度の平均、SD、さらに"hotspot"の数を求めた。"hotspot"は、組換えリードの割合が平均+ 2 x SD を超えている箇所として定義した。その結果、window 10kb においては、平均組換え頻度が 0.25%、hotspot window が 13,346 箇所、window 100kb においては、平均組換え頻度が 0.27%、hotspot window が 1,770 箇所、window 1Mb においては、平均組換え頻度が 0.25%、hotspot window が 213 箇所となった。各 window サイズにおける組換え頻度の分布は以下の通りである。

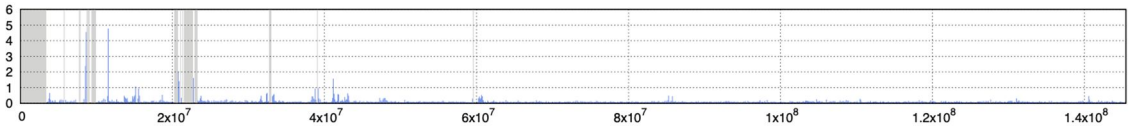


window サイズ 10kb においては組換えが認められない window が約半数であるのに対し、100kb とすると認められない window はほぼ存在せず、おおよそ 100kb に 1 箇所程度の組換えは起きていることが明らかとなった。

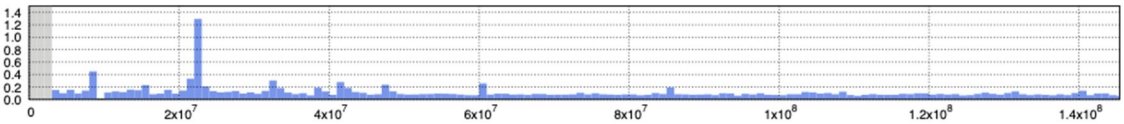
B6 x CAST 間で認められる組換え頻度の染色体内での分布

上記組換え頻度の偏りを、染色体スケールで分布を確認した。以下に例として 7 番染色体全体を Window サイズ 100kb, 1Mb で解析した例を示す。なお灰色で塗られている領域は十分量のデータが確保できていない箇所である。

・window 100kb



・window 1Mb



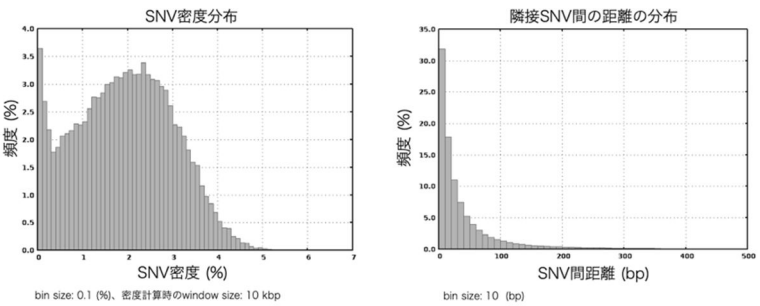
このように局所的に hotspot 様に組換え頻度の高い箇所が存在することが確認できる。また 7 番染色体の左側は、オス染色体にてインプリンティングが起きていることが知られている領域である。もしかするとインプリンティングと組換え頻度の高さは関連しているかもしれない。

2.) 野生イトマキヒトデの解析

新たに野生 1 個体の精巣から DNA を採取し、bulk でのシークエンス(Illumina pair-end, mate-pair,PacBio HiFi)を実施した。イトマキヒトデに関しては、参照ゲノム配列が公開されていないため、得られたシークエンスデータから Platanus-allee アセンブラにてアセンブルを行い、参照配列とした。この際相同染色体をできる限り「分けて」アセンブル(phasing)し、両者のゲノムを比較(minimap2 による比較)することで、ハプロタイプ間の SNV 検出の精度向上を試みた。なお、構築されたイトマキヒトデゲノムは総塩基長 968.8Mb、scaffold N50: 23.3Mb、Contig N50 825.9 kb である。

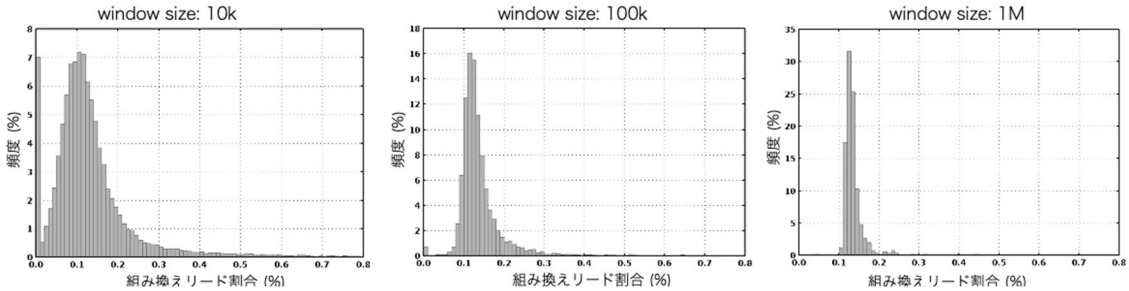
イトマキヒトデ相同染色体間で認められる SNV

まず、Illumina リードを bwa-mem を用いて参照ゲノムにマッピングし、identity, alignment coverage, 推定 insert size などの情報でフィルタ後、GATK HaplotypeCaller で SMV コールを行なった。この際にも QV や coverage depth による厳格なフィルタリングを実施した。さらにこのデータでは、ハプロタイプ別アセンブル結果を比較(minimap2 による)で得られた SNV 情報によるフィルタリングも実施した。その結果、マッピングの



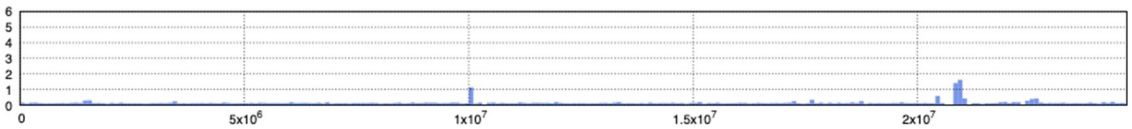
みで 5,498,609 箇所の SNV を、フィルタリングで 5,270,842 箇所のより確実な SNV をコールすることができた。得られた SNV の密度および SNV 間距離の分布は前ページの通りである。この結果より、SNV 間の距離は Illumina pair-end データを用いても insert 距離内にほとんどが収まっていることが確認できた。また、マウスの結果と比べて SNV 密度が全般的に高いとともに、密度の幅が広いことも確認された。(マウス:平均 0.82%, SD 0.44 に対しイトマキヒトデ:平均 1.94%, SD:1.07)
同様の解析を PacBio HiFi リードを用いても実施したが、coverage が低いこともあり検出された SNV は 2,831,795 箇所と Illumina で得られた SNV の約 54%と少なかったため、以下の結果は全て Illumina から得られた結果のみを示す。

<u>イトマキヒトデゲノムで認められる組換え頻度</u>
次に隣接する SNV 間で(頻度の低いパス)/(頻度の高いパス)を計算することにより、組換えリードの割合を算出し、ゲノムを 10kb, 100kb, 1Mb の window に分割した上で window 内の組換え頻度の平均、SD、さらに"hotspot"の数を求めた。その結果、window 10kb においては、平均組換え頻度が 0.16%、hotspot window が 949 箇所、window 100kb においては、平均組換え頻度が 0.15%、hotspot window が 72 箇所、window 1Mb においては、平均組換え頻度が 0.14%、hotspot window が 17 箇所となった。各 window サイズにおける組換え頻度の分布は以下の通りである。



平均組換え頻度はマウスの例と比べて低いが、window サイズ 10kb においてもゲノムの約 90%領域で組換えが確認された。これは hotspot の数がマウスと比べて大幅に低いこととも整合性の取れる結果となっている。

<u>イトマキヒトデゲノムで認められる組換え頻度の染色体内での分布</u>
最後に上記組換え頻度の偏りを、染色体スケールで分布を確認した。以下に例として 6 番染色体全体を Window サイズ 100kb で解析した例を示す。



この結果からもマウスの例と比べて hotspot 様の箇所が少なく、ゲノム全体に渡って比較的満遍なく低い組換えが起きていることが確認できる。

以上、本研究を通じて開発したプログラムを用いて F1 マウス、イトマキヒトデの pool された精子のゲノム解析結果を示した。今回得られた結果には「正解」データが存在しないため、その精度や感度を求めることができていないことは今後の大きな課題である一方、pool された bulk の DNA を用いてゲノム全体における組換え箇所・頻度を解析できる手法が得られたことは、大きな成果であると考えられる。本研究課題がスタートして以降、マウス Hinch et al. (2019)、牛 Yang et al. (2022)それぞれについて、精子のゲノム解析を用いた組換え位置に関する研究が報告されている。しかしこれらの論文では single-cell 解析に基づいた解析が行われている。Single-cell 解析が実現できれば、その 1 細胞での組換え位置が求められるため非常に有益な情報が得られることは間違いない。しかし、single-cell ゲノム解析には技術的な困難さや、増幅時のバイアスが避けられないといった問題、ひいてはこれらの点に起因して多くの細胞での実施時の高コスト化につながり、様々な種で容易に行うことは現実的ではない。事実、上記論文では 200 細胞程度の実施でゲノムの 60%程度をカバーするに留まっており、hotspot とされている箇所も 10 細胞で見つかった領域を一箇所例示しているのみである。
一方本研究による手法は、高精度な相同染色体間に認められる SNV 検出に基づいた bulk での解析を実現可能にしたものであり、容易にデータが取得可能な方法に基づいていること、さらには数万レベルの細胞で起きている組換えイベントを網羅的に捉えることなどのメリットも多い。今後 DSB-seq などの知見と合わせて解析していくことで、更なる生殖細胞系列での組換えイベントの理解が進むことが期待される。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　4件）

１．著者名 Hideaki Yuasa, Rei Kajitani, Yuta Nakamura, Kazuki Takahashi, Miki Okuno, Fumiya Kobayashi, Takahiro Shinoda, Atsushi Toyoda, Yutaka Suzuki, Nalinee Thongtham, Zac Forsman, Omri Bronstein, Davide Seveso, Enrico Montalbetti, Coralie Taquet, Gal Eyal, Nina Yasuda, Takehiko Itoh	４．巻 28
２．論文標題 Elucidation of the speciation history of three sister species of crown-of-thorns starfish (Acanthaster spp.) based on genomic analysis	５．発行年 2021年
３．雑誌名 DNA Research	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/dnares/dsab012	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Rei Kajitani, Hideki Noguchi, Yasuhiro Gotoh, Yoshitoshi Ogura, Dai Yoshimura, Miki Okuno, Atsushi Toyoda, Tomomi Kuwahara, Tetsuya Hayashi, Takehiko Itoh	４．巻 49
２．論文標題 MetaPlatanus: a metagenome assembler that combines long-range sequence links and species-specific features	５．発行年 2021年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkab831	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

１．著者名 Miki Okuno, Shusei Mizushima, Asato Kuroiwa, Takehiko Itoh	４．巻 13
２．論文標題 Analysis of Sex Chromosome Evolution in the Clade Palaeognathae from Phased Genome Assembly	５．発行年 2021年
３．雑誌名 Genome Biology and Evolution	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/gbe/evab242	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

１．著者名 Yong-Woon Han, Rei Kajitani, Hiroya Morimoto, Maierdan Palihati, Yumiko Kurokawa, Rie Ryusui, Bilge Argunhan, Hideo Tsubouchi, Fumiyoshi Abe, Susumu Kajiwara, Hiroshi Iwasaki, Takehiko Itoh	４．巻 9
２．論文標題 Draft Genome Sequence of Naganishia liquefaciens Strain N6, Isolated from the Japan Trench	５．発行年 2020年
３．雑誌名 Microbiology Resource Announcements	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1128/MRA.00827-20	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 7.Miki Okuno, Rei Kajitani, Hiroyuki Tanaka, Shuhei Mizushima, Asato Kuroiwa, Takehiko Itoh
2．発表標題 Whole-genome sequencing and comparative analysis of Emu provides insights into sex chromosome evolution of the palaeognathae clade
3．学会等名 Plant and Animal Genome XXVIII Conference（国際学会）
4．発表年 2020年

1．発表者名 山部貴央，小林史弥，梶谷嶺，伊藤武彦
2．発表標題 アゲハチョウ属5種の比較ゲノム解析
3．学会等名 日本動物学会関東支部第72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