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　５ 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：挑戦的研究（萌芽）
研究期間：2021～2022
課題番号：２１Ｋ１９２１１

研究課題名（和文）メタゲノム解析のためのk-merベースGWAS手法の開発

研究課題名（英文）Development of a k-mer-based GWAS method for metagenomic analysis

研究代表者
伊藤　武彦（Itoh, Takehiko）

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）　　5,000,000円

研究成果の概要（和文）：本研究では、ある二群の環境サンプルからそれぞれメタゲノムシークエンスを行い、比較解析することで、環境特異的なゲノム領域を効率的に抽出するための手法開発を目指した。当初は、両環境から取得したショートリードから特異的なk-merを抽出、アセンブルすることで実現予定であった。しかし、エラー由来のk-merとの頻度差による区別が困難などの理由から、両環境全体のpanメタゲノムアセンブルを実現後、特異的なゲノム領域をアセンブルグラフのバブル構造から抽出する方法へと方針を変更し、開発を行った。また実データへの適用により、同一菌種内で環境により多様性を持つ領域の抽出に成功し、本手法の有効性を示した。

研究成果の学術的意義や社会的意義
近年、ある環境に生息する細菌群集からゲノムDNAを直接回収・解析するメタゲノム手法は注目されている研究分野の一つである。しかし、アセンブル・ビンニングにより各菌のゲノムを再構成させ、さらには表現型との因果関係まで繋げようとすると解析は極めて困難であると言わざるを得ない。このような問題に対して、二群間で異なる多様性を持つゲノム領域を効率的に抽出することを可能にする本成果の利用は、研究遂行を容易にする効果が期待される。残念ながらデータ量の不足などから、両群の表現型とのリンクまでを直接的に見出すことは、現時点で困難であるが、候補箇所の効率的な抽出による候補の絞り込みが可能となり、その意義は大きい。

研究成果の概要（英文）：In this study, we aimed to develop a method to efficiently extract environment-specific genomic regions from two groups of environmental samples by metagenomic sequencing and comparative analysis, respectively. Initially, we planned to achieve this region by extracting and assembling specific k-mer from short reads obtained from both environments. However, due to the difficulty of distinguishing k-mer from error-derived k-mer by frequency difference, we changed the plan and developed a method to extract the genomic regions specific to each environment from the “bubble structure” of the assembly graph after pan- metagenome assembly of the whole sequence data from both environments. By applying the method to real data, we succeeded in extracting regions of diversity within the same bacterial species depending on the environment, demonstrating the effectiveness of our method.

研究分野：ゲノム情報解析

キーワード：メタゲノム　アセンブル

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

次世代シークエンサ（NGS）の登場に伴い、膨大な塩基配列が安価に産出される時代となり、NGS データを活用した研究は多岐に渡っている。適用分野の一つとして、ある環境に生息する細菌群集からゲノム DNA を直接回収・解析するメタゲノム手法も注目されている。多くの難培養性細菌情報を網羅的に解析できるメタゲノム解析は、様々な環境へ適用されているが、多くは 16S を用いた環境を構成する菌種組成の推定や変動を見るなどに止まっている。16S データからは菌種情報以上は基本的に得られず、ある環境でその菌種が果たしている役割は不明である。このため、ある環境中の全菌種の網羅的なゲノム情報を得ようと全ゲノムショットガン法によるフルメタゲノム解析研究も行われるようになってきている。しかし得られた断片配列から、アセンブル・ビンニングにより各菌のゲノムを再構成させる必要があり実現は容易ではない。現在でもメタゲノムから数多くの個別菌ゲノムを完成させることで、大きな研究成果とみなされる状況にあり、さらに一歩踏み込んだ表現型との因果関係まで繋げることは、極めて実現のハードルが高いと言わざるを得ない。

一方、ヒトなどの個別ゲノム解析では、疾患の原因をアレイ・SNP を用いた GWAS 解析により同定する研究が広く行われている。NGS の活用により、ヒトのみならずイネ・トマトなどの植物やバクテリアなどでも広く行われ、GWAS に基づき表現型との関連を調べる研究は一つのトレンドを形成している。GWAS のメタゲノムへの応用も当然見られ、ここ数年成果報告が続いている。しかし GWAS 解析は、SNP 頻度情報などと質的・量的形質との関連を統計的に調べる方法であるため、基盤となるゲノム情報が必要であり、その意味からメタゲノムへの応用では、構成する菌種のゲノム情報が充実しているヒト腸内細菌研究に限られているのが状況であった。

２．研究の目的

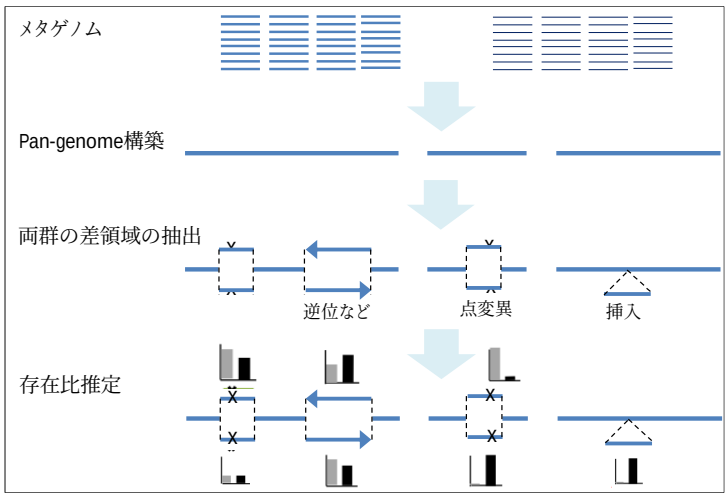
本研究では、現状を打破すべくメタゲノム解析本来の目的に立ち返り、ある二群の環境サンプルからそれぞれ複数のメタゲノムデータを取得、それらを直接的に比較し、GWAS 的な発想を組み込むことにより二群の違いを引き起こす要因となるゲノムの相違箇所を統計的に抽出し、その差分箇所のみのアセンブルを実施、明らかにするという手法の開発を目指す。本手法の実現によりアセンブルに成功した一部のみの比較とならざるを得ない現状を回避するとともに、膨大なデータを相手にすることなく、導出された機能との連関が疑われる箇所の配列のみを集中的に下流解析することが可能となることが期待される。

３．研究の方法

計画初期においては、以下の手順に従い手法の開発を試みた。アウトラインを以下に示す。

・各環境を構成している複数サンプルからの各 Illumina ショートリードデータの取得。
・各群において得られたシークエンスデータから、部分配列である k-mer の頻度情報を取得。
・両群間で有意に多く存在する k-mer のみの抽出。
・抽出された k-mer をアセンブルすることにより連続した配列を作成

この一連の解析により、二群間の表現型の違いと統計的に関連があるとみなされるゲノム領域のみに絞り込むことで、該当箇所の網羅的な抽出が可能となることを期待したものである。しかしながら、成果にも記載するように、本手法では芳しい結果が得られなかったため、以下の通り方向転換を図り、研究を実施した。



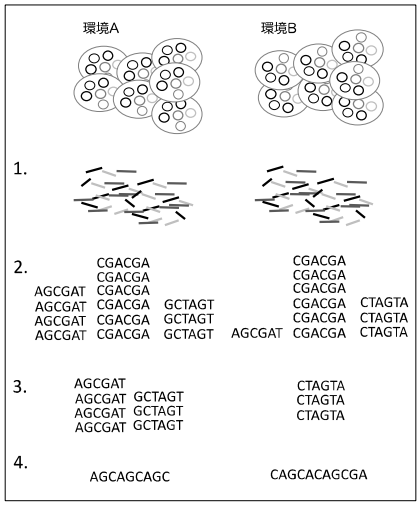
集団内で多様性を持つ箇所を抽出する目的で、まず全データを用い、より長くアセンブル結果を得るためのメタゲノムアセンブラの開発を実施し、その中で多様性を持つ領域を「バブル構造」として検出、さらには各環境サンプルをマッピングすることで、さらにマイナーな多様性を持つ領域の検出を実施する。最後に、多様性を持つ領域の存在比を推定し、群特異的領域を決定することで、当初目的の実現を目指した。具体的に開発した方法は、研究成果で示す。

また、開発したアセンブラのベンチマークを、Mock として用意したゲノム配列既知のバクテリアを Mix したデータおよび本研究で取得した実データで実施した。

４．研究成果

A) 初期計画に基づいた、両群間で差のある領域のみのアセンブル手法の開発とテスト

まず期初の計画に基づき、二群を構成しているサンプルから取得した Illumina Pair-end データから k-mer(21-32 をテスト)頻度分布を作成し、両者を比較することにより群間で差のある k-mer の抽出を行うプログラムを開発した。続いて開発したプログラムを腸内メタゲノムサンプルなどのテストテータに適用し、その先のアセンブルを含めた工程を今後開発することにより、目的を達成できるかその見極めを試みた。しかしながら、パラメータなどをチューニングしても、得られる k-mer の種類が想定していたものよりも遥かに多くなってしまった。その原因を調べたところ、出現頻度の低い k-mer が非常に多く出力されており、これは一般的なゲノム解析では、出現回数の分布からエラーに起因した k-mer と実在する k-mer とを区別することが可能であるのに対し、メタゲノムサンプルではマイナーな菌種由来の k-mer との区別が困難であることが原因であると考えられた。



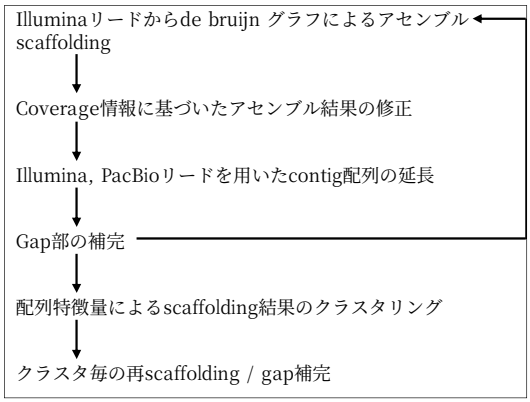
また二群間の比較において、当初想定していたものよりも二群間で重なる k-mer の少ないことも確認された。これもメタゲノム特有の菌種毎の頻度差に起因するものと考えられ、一般的なゲノム解析に必要となる coverage40 程度が確保できている菌種およびゲノム領域は限られており、多くの菌種・領域では十分な coverage が得られず、多くの特異的 k-mer を出力することになっていることが確認された。

さらに、得られた特異的 k-mer をアセンブルすると非常に短い断片配列が多く産出されることも確認された。この原因は頻度を稼ぐために k のサイズを小さくせざるを得ないため、特異的領域にもそれ以外の共通領域と共通に存在する k-mer の種類が多く含まれてしまうことで、そのような箇所でアセンブルが分断されることによることに起因すると考えられる。期初にはこのようなケースにおいても、精度が高い Illumina Mate-pair リードの活用による scaffolding で断片配列の結合を行うことを想定していたが、Mate-pair ライブラリ作成キットの終売に伴い、実現が困難となった。

A)で示したような結果が途中得られたため、方針を１年目の後半から転換し、まずは解析対象とする全てのシークエンスデータをアセンブルすることにより、pan-genome 状態のゲノム配列をできるだけ長く得ることを目的としたアセンブラを開発することを目指した。そしてアセンブラの中に群間で多様性を持つ領域をバブル構造として抽出する機能、さらには各群由来の配列をマッピングすることでより小規模な多様性を抽出する機能を開発、抽出された領域と各群との相関を見ることにより、研究の目的を達成する方法を目指した。

B) メタゲノム用アセンブラの開発

上記転換した方針に従い、まずはメタゲノム用アセンブラの開発を実施した。構築したアセンブラのフローを右図に示す。入力データとしては Illumina Pair-end と PacBio などのロングリードを想定している。



アルゴリズムの概要は以下の通りである。まず Illumina ショートリードを de bruijn グラフに基づき、接続して Contig を構築する。この際には、メタゲノムでは菌種間でリード量もばらつくことを考慮した設計としている。Illumina リードに基づいて Contig を構築することで、長さは十分でないものの正確性と網羅性に優れた Contig の構築が可能となる。続いて、Contig ペアを架橋する Illumina のペアおよびロングリードを検出し、その組を接続することで Scaffolding を行う。さらに得られた scaffold 両端の延長を試みる。ここまでの工程を複数回繰り返した後に、得られた scaffold を配列特徴量に基づき、クラスタリングを行う。さらに得られたクラスタ毎に scaffolding と gap 埋めの工程を行って、結果配列を出力する。

本アセンブラの特徴として、菌種間のミスアセンブルの軽減策が挙げられる。具体的には、リードの coverage は基本的に菌種毎に一定であるというアイデアに基づき、Contig 内で coverage が大きく変化している箇所を検出し、その箇所で切断する機能を組み込んでいる。この機能により、結果として誤った contig 間の架橋を減らすことにもなり、途中で用いているグラフ構造の複雑性が低下し、結果的に長い配列を構築することにもつながっている。

C) 構築したアセンブラのベンチマークテスト

構築したアセンブラを用いたベンチマークテストの結果を示す。まずは GIS20 と呼ばれるゲノム既知の 20 菌種からなるシークエンスデータを用いた結果を以下の表に示す。データは Illumina pair-end, PacBio リードである。

ssembly	NG50 (bp)	NGA50 (bp)	# all-misassemblies	Mean sequence identity (%)	Genome fraction (%)	# inter-species misassemblies
本アルゴリズム	1 444 699	129 682	280	99.9390	69.95	0
metaSPAdes	374 634	112 139	219	99.8685	75.47	4
OPERA-MS	135 973	40 414	423	99.8566	71.51	13
metaFlye raw	235 065	11 717	242	99.7181	50.04	5
metaFlye polished	235 089	13 257	256	99.9221	50.32	5

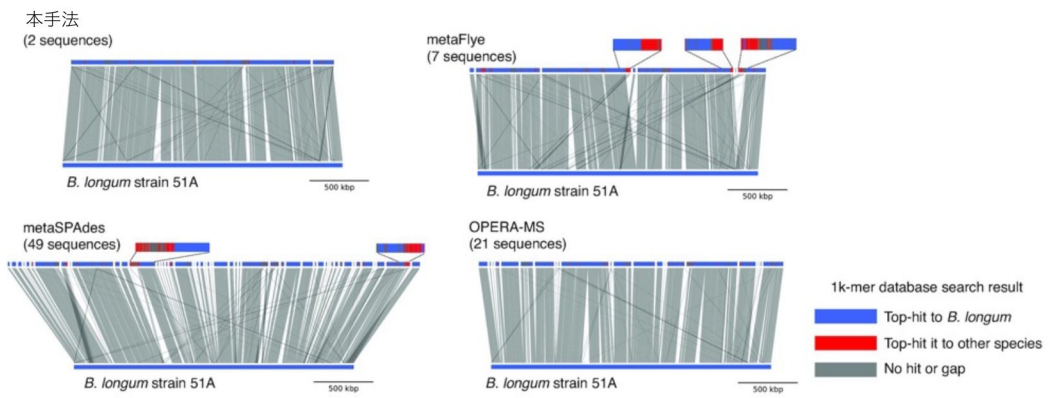
本アルゴリズムを用いて得られた結果から、繋がりの指標である NG50 の値が比較対象のアセンブラと比べて最も長く、また種間を間違えて繋いだミスアセンブルの数も 0 と最も少ないことが確認できる。

続いて、ヒトの腸内細菌データ(Bertrand D. et al. Nat. Biotech. 2019)を用いたベンチマークテストを実施した。scaffold N50 および、得られた scaffold から遺伝子を予測した中から SwissProt にヒットした遺伝子数を以下に示す。

	Scaffold N50 (bp)					**# CDSs hit to SwissProt.**				
Sample ID	**本手法**	metaSPAdes	OPERA-MS	metaFlye raw	metaFlye polished	**本手法**	metaSPAdes	OPERA-MS	metaFlye raw	metaFlye polished
V00-S-0509-S01	94,973	16,744	41,074	59,682	59,638	12,665	9,436	9,466	2,594	10,940
V00-S-0511-S01	16,478	18,918	11,175	15,204	15,207	4,406	5,443	4,506	0	0
V01-T-0506-S02	37,571	33,054	19,864	50,150	50,157	8,577	8,428	7,725	283	3,700
V02-T-0504-S03	52,447	39,185	31,962	37,453	37,455	6,223	6,301	5,545	663	1,804
V02-T-1664-S03	409,203	32,192	130,981	150,075	150,542	11,232	8,995	9,158	835	9,461
V02-T-1665-S03	181,937	35,610	69,496	321,411	322,103	3,321	2,954	2,787	569	3,390
V03-S-0457-S04	378,761	237,395	156,750	152,295	152,892	2,286	2,283	2,223	325	1,985
V03-S-1663-S04	50,103	43,925	27,766	32,736	32,736	2,415	2,643	2,349	66	662
V03-T-0504-S04	74,806	22,461	51,883	94,860	95,036	3,125	3,167	2,697	429	2,935
V03-T-0506-S04	51,084	45,393	33,659	57,798	57,890	12,165	10,112	10,267	929	6,443
V03-T-0508-S04	588,856	54,608	155,866	379,361	379,937	12,315	9,232	11,079	2,653	10,704
V04-S-0509-S04	24,955	22,382	14,976	28,988	29,036	8,286	8,260	7,374	236	2,042
V05-S-0512-S05	3,618,905	1,230,380	292,645	780,728	783,447	1,555	1,559	1,560	241	1,558
V05-T-0502-S06	17,679	19,238	13,412	100,961	101,044	5,713	5,940	5,771	0	0
V05-T-0513-S05	87,131	57,024	43,063	61,936	62,317	3,279	3,295	3,144	418	1,610
V06-T-0501-S07	33,927	11,254	25,630	31,806	31,833	1,240	1,197	1,072	263	1,038
V06-T-0502-S07	150,407	16,358	41,645	77,079	77,291	6,598	6,477	6,196	1,169	6,359
V07-T-0504-S08	29,531	16,156	18,649	101,642	100,518	4,752	5,096	4,572	137	1,347

結果より、本手法による結果は metaSPAdes, OPERA-MS といった Illumina をベースにアセンブルする他の手法よりも連続性が高いだけでなく、ロングリードベースのメタゲノムアセンブラである metaFlye よりも連続性の高いアセンブル結果が得られていることが確認できる。また、予測遺伝子数においても、多くのケースに置いて一番多い結果が得られている。このことから網羅性が高くアセンブルできていることも推察される。(なお、ロングロードベースの metaFlye においては、単なるアセンブル結果からは予測される遺伝子数が極めて少ない。これはロングリードが元々持っている高いエラーレートに起因している可能性が高い。本ベンチマークでは、アセンブル後に Illumina データでの polish を行うことで配列精度を向上した場合の結果も示している。Polish 後では多くの遺伝子が予測されていることが確認できる。)

最後に、各アセンブラの結果からゲノム既知の *B. longum* に相当する結果を抜き出し、アセンブル状況を調べた結果を以下に示す。結果より、本手法では 2 本に長くつながっている事に加え、他のアセンブラ結果で見られるような他のバクテリアゲノムと間違って繋がってしまったようなアセンブル領域が見られず、正しくアセンブルできていることが確認できる。

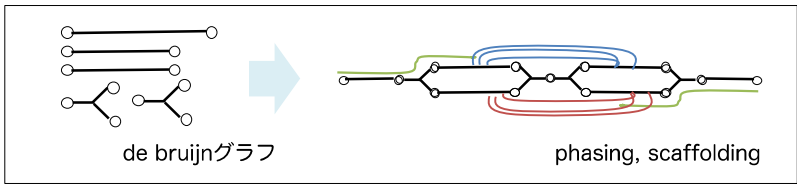


D)群間で多様性のある領域の抽出

続いて群間で多様性のある領域を抽出する機能開発を実施した。まず、両群のデータを用いてC)で開発したアセンブラにより pan-ゲノムを構築する。その pan-ゲノムを基に、両群間で配列に差異のある領域を以下の 2 手法で抽出する手法を開発した。

・アセンブラ内部で、二型をとる箇所をバブル構造として許容し、抽出する機能

この機能は、C)で開発したアセンブラに、de bruijn グラフの段階で二型に起因した枝分かれ構造を許してバブル構造として保持する機能を追加し、さらに pair-end 情報や long-read 情報、あるいは 10X などの link 情報によりバブル構造間の関係性を解く phasing の機能を追加することで実現した。基本的には二倍体ゲノム用アセンブラ platanus-allee で開発していた機能を基にしているが、二倍体の場合にはハプロタイプ間の頻度が等しいのに対し、メタゲノムのバブル間頻度は不均等であることなどに伴う細かい改良を施すことで実現を目指した。



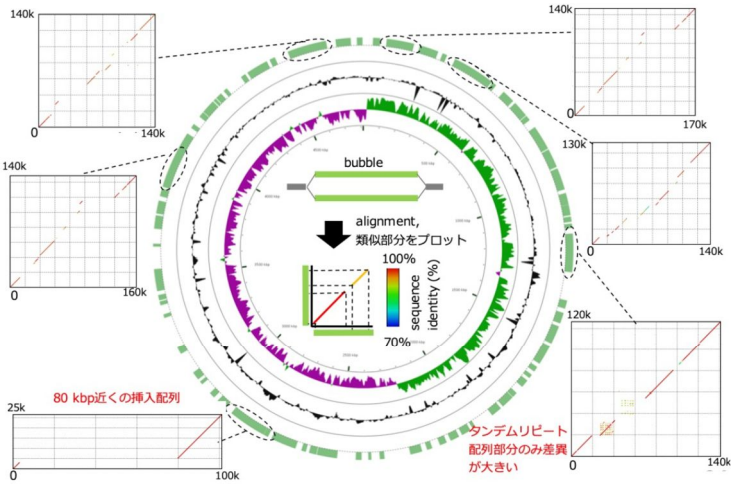
・アセンブル結果へのマッピングによる、細かい変異箇所の抽出機能

上記アセンブラによる方法で、群間の違いを網羅的に捉えられれば良いが、実際には SNP など点変異のような細かい差を抽出することは困難である。配列間にわずかに差が生じるごとにバブル構造とした場合には、シークエンスエラー部も枝分かれ構造となってしまい、グラフが複雑すぎ解決できなくなる。二倍体ゲノムなどの場合には、このような問題点を、シークエンスカバレッジで解決しているが、メタゲノムの場合には、マイナーな菌種由来のリードとエラー由来のリードとを頻度差で区別することが極めて困難なため、一定頻度以下のパスは除去することにより対応せざるを得ない。

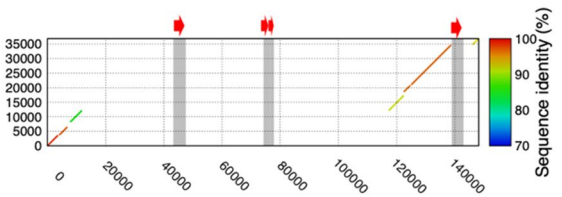
そのような箇所の差異を検出する補助的な機能として、アセンブル結果にリードをマッピングし、変異箇所の抽出機能も開発した。

E) 群間で多様性のある領域の抽出事例

D)の有効性を確かめる目的で、ヒト腸内メタゲノムデータ(Bishara et al. Nat. Biotech. 2018)に当てはめた例を示す。このサンプルは、10X と pair-end の 2 種類のライブラリから 36Gb ずつの配列が産出されている。まず、アセンブルした結果、near-complete クオリティの scaffold が 4 つ、intermediate-quality の scaffold が 8 つと論文に記載されていた、それぞれ 2 つ、4 つよりもより完成度の高いアセンブル結果が得られた。続いてアセンブル結果に D)で開発した機能を適用したところ、右図のようにコンセンサスとしてアセンブルされたゲノムに対し、二群間で配列の違う箇所が複数箇所得られた。



右図で示した 6 箇所においては、いずれもその周辺領域は 99%以上の高い相同性を示すのに対し、多様性を示す領域では、全く相同性を示さない二型を取っていることが確認された。このような領域では、両サンプル間の相同性が異なり過ぎるため、マッピングの解析では見出すことのできなかった領域である。同様に左下の領域は、片方の環境のみで 110kb 程度の挿入があることが明らかとなった。拡大図を右下に示す。赤矢印で示している箇所には薬剤耐性関連データベース CARD への配列検索でヒットを示した遺伝子が座乗している。このように、ある環境からのサンプルのみで検出されるゲノム領域の抽出ができている例を確認することができ、本研究で開発した手法の有効性を示すことができた。



5．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

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 e130
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 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 Shun Ouchi, Rei Kajitani, Takehiko Itoh
2．発表標題 A De Novo Chromosome-Level Scaffolding and Phasing Tool Using Hi-C
3．学会等名 Plant and Animal Genome 30 Conference（国際学会）
4．発表年 2023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