

令和 2 年 9 月 11 日現在

機関番号：63904

研究種目：基盤研究(B) (一般)

研究期間：2017～2019

課題番号：17H03610

研究課題名(和文) 比較ゲノム解析に基づく水平伝搬による原核生物ゲノム進化の包括的解析

研究課題名(英文) Comprehensive analysis of the role of horizontal gene transfer in prokaryotic genome evolution based on comparative genome analysis

研究代表者

内山 郁夫 (Uchiyama, Ikuo)

基礎生物学研究所・ゲノム情報研究室・助教

研究者番号：90243089

交付決定額(研究期間全体)：(直接経費) 13,200,000円

研究成果の概要(和文)：微生物ゲノム間でオーソログを対応づけたデータベース(MBGD)に基づいて、原核生物における水平移動遺伝子の網羅的解析を行った。同一種内の株間で保存されていない「非コア遺伝子」を水平移動の候補とし、それらの並び順が保存された構造を水平移動の痕跡「ゲノミックアイランド」として抽出した。また、種内の遺伝子レパートリー全体(パンゲノム)を種間に対応づけることにより、非コア遺伝子が遠縁種にまたがって出現するケースを抽出した。これらの結果に、以前から用いられている水平移動の指標や既知の可動因子の知見を加えて、水平移動の全体像を推定したデータベースを構築した。

研究成果の学術的意義や社会的意義

原核生物(細菌、古細菌)は、自然環境や生物体内など地球上の至る所に存在し、その生態を理解することは環境や健康など身近な問題を解決する上で重要である。原核生物がコンパクトなゲノムを持ちつつ多様な環境に素早く適応できる理由として、ゲノム間で遺伝子群が「水平移動」することが挙げられる。本研究は、ゲノム比較により水平移動を推定する独自のアプローチを開発し、これを大規模なゲノムデータに適用することで、原核生物における水平移動の全体像を示すデータベースを構築した。本成果は、自然界で現在進行中の現象でもある原核生物のゲノム進化を解明する上で、重要な基盤データになると期待される。

研究成果の概要(英文)：Using a microbial comparative genome database (MBGD), we have conducted comprehensive analysis of horizontal gene transfer among prokaryotic genomes. We considered “non-core genes,” i.e. genes that are not conserved among the genomes of the same species, as candidates of horizontally transferred genes, and constructed a structure consisting of non-core genes with conserved gene order. Moreover, we identified orthologous relationships among the entire gene repertoire of species (pan-genome) and extracted non-core genes that appear in distantly related organisms. In addition to these results, we added conventional indices of horizontal gene transfer and knowledge of known mobile genetic elements and constructed a database of comprehensive prediction of horizontally transferred genes among prokaryotes.

研究分野：ゲノム情報科学

キーワード：水平遺伝子移動 微生物ゲノム ゲノム進化

## 1. 研究開始当初の背景

原核生物はコンパクトなゲノムを持ち、生体内や土壌、海洋など、極めて多様な環境に生息している。コンパクトなゲノムサイズは、素早く増殖して生存競争を勝ち抜く上で有利にはたらく一方、ゲノム内の遺伝子の多様性は限定される。遺伝子の水平伝搬(Horizontal Gene Transfer; HGT)はそうした点を克服し、種内や集団内における遺伝子の多様性を維持し、様々な環境変化に素早く適応することを可能にする仕組みと考えられ、原核生物の進化において重要である。

我々はこれまでに、種内でシンテニー(遺伝子の並び順)が共通に保存されている構造を「コアゲノム」として、これを種の共通祖先から垂直的に伝搬して維持されている構造として抽出することを提唱し、ゲノムコア構造抽出プログラム CoreAligner の開発を行ってきた(Uchiyama 2008)。そして、我々が独自に構築しているオーソログ解析に立脚した微生物比較ゲノムデータベース MBGD(Uchiyama 2003)において、様々な系統群ごとにコア構造を抽出し、そのデータベース化を行ってきた(Uchiyama 2013)。一方、シーケンス技術の進展によって、様々な種で多数の個別株ゲノムが決定されるようになり、多くの種においてコアゲノム以外の「非コア」遺伝子が大量に存在することから、原核生物における種内ゲノム多様性の大きさが改めて浮き彫りとなった。そこで本研究では、これらのデータベース中の非コア遺伝子に着目した網羅的な解析を行うことにより、原核生物における HGT によるゲノム進化の全体像に迫ることができるのではないかと考えた。

先行研究として、*Helicobacter pylori* の非コア遺伝子について、比較ゲノム解析ツール RECOG(Uchiyama 2017)を用いて、遺伝子のあるなし(系統パターン)と遺伝子の近傍関係の保存性を組み合わせたクラスタリングによって、ゲノミックアイランドを抽出する研究を行っており(Uchiyama et al. 2016)、本研究ではこれをよりシステムティックに行うゲノミックアイランド抽出手法を開発して多様な生物種に適用し、これを核として研究を進めることを計画した。

## 2. 研究の目的

本研究では、種内の遺伝子レパートリー全体(パンゲノム)のうち、コアゲノムに属さない「非コア遺伝子」に着目し、HGT に関与している可能性を評価する。特に、従来あまり追求されてこなかった比較ゲノム解析的アプローチに基づく HGT の推定手法を開発し、これを従来用いられてきた HGT 指標と比較し、統合することを目指す。具体的には、以下の観点から研究を行う。(1) 非コア遺伝子の並び順を比較して、並び順が保存された構造をゲノミックアイランドの候補として抽出する手法を開発する。(2) 種内、属内、全体比較と、階層的にオーソログ解析を進めていく手続きによって、非コア遺伝子を含む種パンゲノムを、近縁および遠縁の種間で比較できるようにする。(3) 従来からある塩基組成ベースおよび配列類似性ベースの HGT 指標を計算し、非コア遺伝子に基づく本研究の開発手法と組み合わせて評価する。(4) 既知の可動遺伝因子(MGE)のデータベースや予測プログラムを用いて MGE を同定し、予測されたゲノミックアイランドとの対応を調べる。(5) これらの成果をまとめてデータベース化し、適切なインターフェイスを通して HGT 遺伝子の全体像を見られるようなシステムを構築する。

## 3. 研究の方法

本研究は、我々が開発し公開している微生物比較ゲノムデータベース MBGD を用いて行う。MBGD では微生物全体にわたるオーソロググループに加えて、系統群ごとのオーソロググループも定義しており、このうち本研究では種ごとのオーソロググループ(その全体がパンゲノムに相当する)と、それをもとに CoreAligner を用いて定義されたコアゲノムを出発点として用いる。

### (1) ゲノミックアイランド(GI)抽出手法の開発

種内オーソロググループの中からコアゲノムに含まれるものを除き、残った非コア遺伝子の間で、ゲノム上の近傍関係が複数の株で保存されている遺伝子セットを GI 候補として抽出する。アルゴリズムは、シンテニーの保存性に基づいてコアゲノムを定義する CoreAligner アルゴリズムを基に開発する。

### (2) 階層的オーソロググループの定義

種パンゲノム全体に対するオーソロググループを定義するため、種レベルのオーソロググループから始めて、階層的にオーソロググループを構成する手続きを作成する。

### (3) 既存 HGT 指標および MGE との対応による評価

開発した非コア遺伝子に基づく HGT 推定手法を評価するため、従来用いられている HGT 評価指標として、塩基組成( $k$ -mer 頻度)の異常性に基づく指標と、配列類似性の異常性に基づく指標を遺伝子ごとに計算して比較する。また、ファージや IS などの MGE を、既存データベースや検出プログラムを用いて同定し、比較する。

#### (4) データベースの構築

以上の結果をデータベースに格納した。MBGD を格納している MySQL 上にデータベースを構築し、MBGD と連携して検索できるようにする。

### 4. 研究成果

#### (1) ゲノミックアイランド(GI)抽出手法の開発

CoreAligner プログラムを拡張して、非コア遺伝子の並び順の保存性から GI 候補を抽出する FindIsland プログラムを作成した。CoreAligner は、オーソロググループ内の各遺伝子の近傍関係に基づいて近傍グラフを作成し、そこからなるべく長い直線的な構造を、コンセンサスの並びとして抽出する。FindIsland は、種内オーソロググループ全体（パンゲノム）からコア遺伝子を除いた非コア遺伝子についてこのアルゴリズムを適用し、並び順が保存された構造を抽出する。ただし、GI では遺伝子、およびその並びの保存性はコアゲノムほど高くないので、保存性の条件は緩く設定した。また、先行研究である FindMobile プログラム(Uchiyama et al. 2016)に準じて、抽出されたアイランドを、各ゲノムにおけるコアゲノムとの位置関係の保存性から、Mobile, Semi-Mobile, Stable の3つに分けてゲノム上の可動性を評価した。

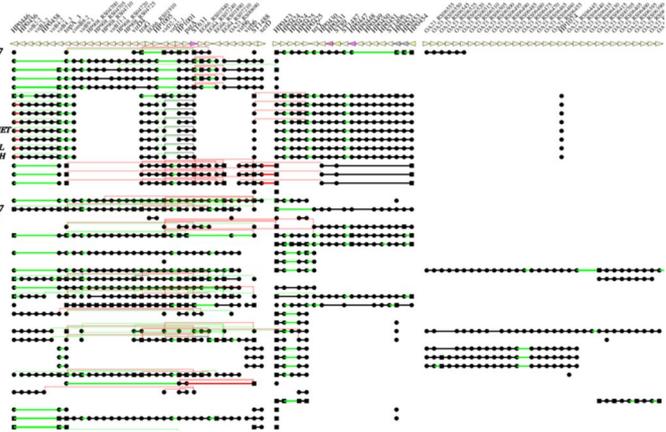


図1 抽出されたアイランドの可視化

本手法を *H. pylori* パンゲノムに適用して抽出されたアイランドの例を図1に示す。この図は、CoreAligner で用いられていると同様の可視化手法を用いて、縦軸方向に株、横軸方向にコンセンサスの並び順に並べられた遺伝子（オーソロググループ）を配置した図の一部（スクリーンショット）である。

本手法を、MBGD 2018 年版に登録された原核生物の種パンゲノムのうち、6 株以上を含んでいる 143 種(計 3077 株)に適用した。図2は、パンゲノム中におけるコア遺伝子 (core 青) アイランド遺伝子 (island 赤) それ以外の非コア遺伝子 (non-core 黄色) の割合を、パンゲノムサイズの大きさ上位 25 種、および全体(円グラフ)において調べた結果を表している。パンゲノム中のコア遺伝子やアイランド遺伝子の割合は種によって違いがあるが、平均的にはパンゲノム中のアイランドの割合は 16%程度であり、パンゲノムサイズが大きい場合には 20-30%に達していた。

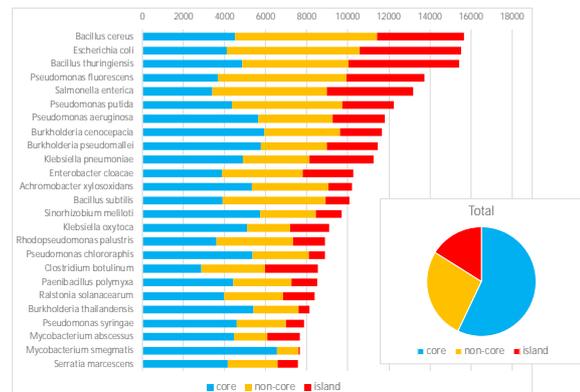


図2 パンゲノム中のコア/アイランドの割合

#### (2) 階層的オーソロググループの定義

MBGD のデータ更新プロトコルを修正して、種レベル、属レベル、全体と3段階で階層的にオーソログ解析を行うようにした(Uchiyama et al. 2019)。具体的に、種レベルのオーソロググループから代表配列を一つ抽出したものを「種パンゲノム」とし、これを入力として属レベルのオーソロググループを作成して、同様に「属パンゲノム」を作成し、これを入力として全体のオーソロググループを作成する、という手続きにより階層的にオーソロググループを作成するようにした(図3)。代表配列は、ドメイン融合や分裂が起きていないこと、他のメンバー配列と比べて類似性が顕著に低くないこと、などの条件によって選定した。

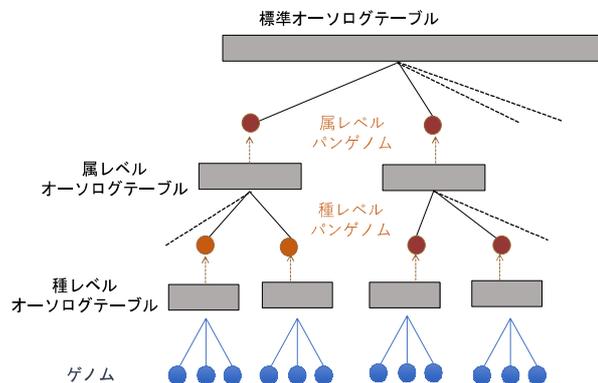


図3 階層的オーソロググループ

本データは、MBGD の 2018 年版としてすでに公開されている。従来の属あたり 1 つの代表

ゲノムを選定してオーソログ解析を行う方法と比べて、パンゲノムを対象にすることによってトップレベルのオーソログ解析には 1.24 倍の遺伝子が解析に取り込まれ、結果としてできるオーソロググループ数は 1.56 倍になった。

### (3) 既存 HGT 指標および MGE との対応による評価

組成ベースの指標としては、個々の遺伝子の  $k$ -mer ( $k=4$ ) 頻度に対して、あらかじめ計算した各ゲノム (パンゲノム) 全体の  $k$ -mer 頻度を比較して、自身のゲノムが最も近いゲノムであるかどうかで評価した。計算は、PSTk プログラム (Dalevi et al. 2006) を本目的に改変したものをを用いて行った。配列類似性ベースの指標としては、配列類似性検索結果から、同一種に対するヒットを除いた上位のヒットの系統群が、自身の系統群と比べてタキシソミーツリー上で離れているかどうかで評価した。これらの結果を、上述の core, island, non-core の 3 つのクラスの遺伝子ごとに整理した結果を図 4 に示す。期待通り、core 遺伝子は  $k$ -mer 頻度でも類似性検索でも HGT と推定されないケースが多く、特に両方の指標で HGT と推定されるケースは非常に少なかった。逆に非コア遺伝子については、いずれかの指標で HGT と推定されるケースが多く、両方で HGT と推定されるケースも多く見られた。その程度は、アイランドを形成しているものとそうでないものとを比べると、後者の方がより HGT と推定される遺伝子の割合が大きかった。

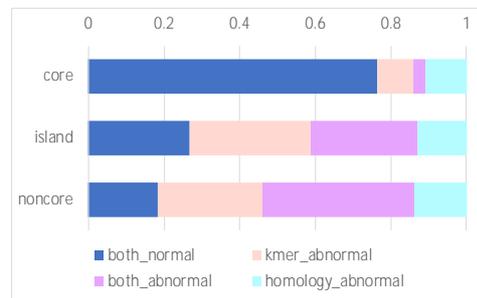


図 4 既存 HGT 指標との対応

一方、既知 MGE として、ファージ (VirSorter; Roux et al. 2015)、IS (ISESearch; Xie et al. 2017)、ICE (Iceberg; Liu, et al. 2018) に加えて、防御関連のアイランドに見られる CRISPR-Cas (CRISPRCasFinder; Couvin et al. 2018) と制限修飾系遺伝子 (REBASE)、さらに塩基組成などに基づく一般的な GI 予測 (IslandViewer; Bertell et al. 2017) を行い (以上、カッコ内は推定に用いたプログラムまたはデータベース)、推定結果を遺伝子ごとにアサインして、本研究で抽出されたアイランドと比較した。結果を図 5 に示す。抽出されたアイランドは、3 つの可動性クラス Mobile, Semi-mobile, Stable に分類し、さらに長さ (遺伝子数) によって分けて評価した。従来法で推定された GI との比較で見ると、長さが 10 を超える場合はよく一致するが、それ以下のものについては一定程度違いが見られた。Mobile, Semi-mobile については半数以上が一致したが、Stable についてはより大きな違いが見られた。既知 MGE との対応で見ると、ファージの割合が最も大きく、短いものでは IS も一定の割合を占めた。

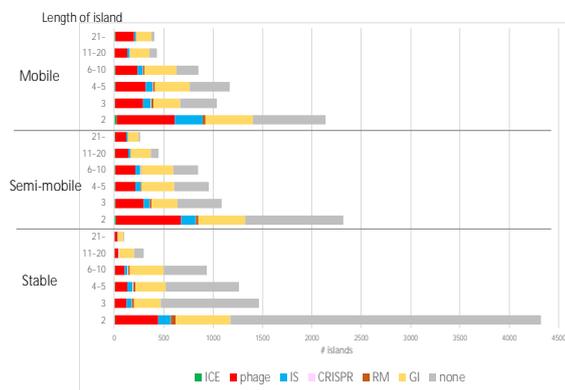


図 5 既知 MGE 等との対応

### (4) データベースの構築

以上の結果を MySQL データベースに登録して SQL を用いて検索できるようにした。パンゲノム中の各遺伝子は、MBGD の taxon specific ortholog table からオーソロググループあたり 1 つずつ選ばれており、taxonomy ID と cluster ID の組で表現されている。このそれぞれについて、コア、非コア、アイランドの別 (システニブロックへの帰属) や、 $k$ -mer および配列類似性に基づく HGT 指標、既知 MGE の推定結果などの情報を格納し、さらに階層的オーソロググループの情報をを使って、MBGD に含まれる個々の遺伝子情報や、上位のオーソログテーブルの情報とも合わせて検索できるようにした (図 6)。上述の結果は、本データベースの検索から導かれた結果の一部である。今後は、本データベースを MBGD と統合して行った解析結果を、Web を介して検索、表示できるようにシステムを拡張して、公開する予定である。

taxid	clusterid	class	mobility	length	kmer_rank	top_kmer_tax	normal_sim	taxonomy of tophits	rank	lca_rank
1063	1	noncore		141	58	mox	no	Bacteria	superkingdom	superkingdom
1063	2	core	CORE	2859	1	tax1063	yes	Rhodobacteraceae	family	family
1063	3	core	CORE	633	1	tax1063	yes	Rhodobacteraceae	family	family
1063	4	noncore		300	1	tax1063	no	Alphaproteobacteria	class	class
1063	5	noncore		432	18	sme	yes	Rhodobacteraceae	family	family
1063	6	core	CORE	622	1	tax1063	yes	Rhodobacteraceae	family	family
1063	7	core	CORE	1584	55	hne	no	Alphaproteobacteria	class	class
1063	8	core	CORE	1848	1	tax1063	yes	Rhodobacteraceae	family	family
1063	4993	noncore		306	31	sth	no	Proteobacteria	phylum	phylum
1063	4994	noncore		462	32	sti	no	Proteobacteria	phylum	phylum
1063	4995	noncore		183	85	hne	no			
1063	4996	island	MOBILE	252	1	tax1063	no	Chelativorans	genus	class
1063	4997	noncore		177	156	gms03554	no			
1063	4998	island	SEMI STABLE	1218	1	tax1063	no	Proteobacteria	phylum	phylum
1063	4999	island	SEMI STABLE	198	387	cgl	no			
1063	5000	island	SEMI STABLE	186	120	bja	no			

図 6 MBGD-HGT テーブル検索例

### < 引用文献 >

Bertell, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing

Group, Lau, B. Y., Hoad, G., Winsor, G. L., Brinkman, F. S. L. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets, 45, W30-S35, 2017.

Couvin, D., Bernhelm, A., Toffano, Nioche, C., Touchon, M., Michalik, J., Neron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., Pourcel, C. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids, Res., 46, W246-W251, 2018.

Dalevi, D., Dubhashi, D., Hermansson, M. Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. Bioinformatics, 22, 517-522, 2006.

Roux, S, Enault, F., Hurwitz, B. L., Sullivan, M. B., VirSorter: mining viral signal from microbial genomic data. Peer J., 3, e985, 2015.

Uchiyama, I., MGD: Microbial Genome Database for Comparative Analysis, Nucleic Acids Res., 31, 58-62, 2003.

Uchiyama, I., Multiple genome alignment for identifying the core structure among moderately related microbial genomes, BMC Genomics, 9, 515, 2008.

Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., MGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucleic Acids Res., 41, D631-D636, 2013.

Uchiyama, I., Albritton, J., Fukuyo, M., Kojima, K. K., Yahara, K., Kobayashi, I. A novel approach to *Helicobacter pylori* pan-genome analysis for identification of genomic islands. PLoS One, 11, e0159419, 2016.

Uchiyama, I., Ortholog identification and comparative analysis of microbial genome using MGD and RECOG. In "Protein Function Prediction: Methods and Protocols" (Daisuke Kihara et.), pp.147-168, Humana Press, 2017.

Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., Kato, M. MGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. Nucleic Acids Res., 47, D382-D389, 2019.

Xie, Z., Tang, H., ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics, 33, 3340-3347, 2017.

5. 主な発表論文等

〔雑誌論文〕 計2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 Uchiyama Ikuo, Mihara Motohiro, Nishide Hiroyo, Chiba Hirokazu, Kato Masaki	4. 巻 47
2. 論文標題 MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons	5. 発行年 2018年
3. 雑誌名 Nucleic Acids Research	6. 最初と最後の頁 D382 ~ D389
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/nar/gky1054	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 Uchiyama Ikuo	4. 巻 1611
2. 論文標題 Ortholog Identification and Comparative Analysis of Microbial Genomes Using MBGD and RECOG	5. 発行年 2017年
3. 雑誌名 Methods in Molecular Biology	6. 最初と最後の頁 147 ~ 168
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-1-4939-7015-5_12	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

〔学会発表〕 計11件（うち招待講演 1件/うち国際学会 2件）

1. 発表者名 Ikuo Uchiyama, Motohiro Mihara, Hiroyo Nishide
2. 発表標題 Orthology Relationships among Pan-Genomes: the Microbial Genome Database (MBGD) for Exploring Microbial Diversity
3. 学会等名 The 67th NIBB Conference "Quest for Orthologs" (招待講演) (国際学会)
4. 発表年 2019年

1. 発表者名 内山郁夫
2. 発表標題 微生物比較ゲノムデータベースを用いた水平移動遺伝子動態の網羅的解析
3. 学会等名 遺伝研研究集会 環境中のDNA循環
4. 発表年 2019年

1. 発表者名 内山郁夫
2. 発表標題 微生物比較ゲノムデータベースMBGDを用いた水平移動遺伝子動態の網羅的解析
3. 学会等名 第42回日本分子生物学会年会
4. 発表年 2019年

1. 発表者名 内山郁夫
2. 発表標題 微生物比較ゲノムデータベースMBGDに基づく水平移動遺伝子動態の網羅的解析
3. 学会等名 第14回日本ゲノム微生物学会年会
4. 発表年 2020年

1. 発表者名 内山郁夫
2. 発表標題 非コア遺伝子に着目した水平移動遺伝子の検出
3. 学会等名 遺伝研研究集会「DNA水平伝搬」
4. 発表年 2018年

1. 発表者名 内山郁夫
2. 発表標題 オーソログを探索する: Quest for Orthologs会議について
3. 学会等名 日本進化学会第20回大会
4. 発表年 2018年

1. 発表者名 内山郁夫
2. 発表標題 非コア遺伝子の並び順の保存性に基づくゲノミックアイランドの抽出
3. 学会等名 第13回日本ゲノム微生物学会年会
4. 発表年 2019年

1. 発表者名 Ikuo Uchiyama
2. 発表標題 A new protocol for constructing the ortholog table in Microbial Genome Database for Comparative Analysis
3. 学会等名 NIG International Symposium 2017 "Life, environment, and evolution revealed by genomes" (国際学会)
4. 発表年 2017年

1. 発表者名 内山郁夫
2. 発表標題 ゲノムコア構造から見た遺伝情報の水平移動
3. 学会等名 遺伝研研究集会「生物種間における遺伝情報の水平移動」
4. 発表年 2017年

1. 発表者名 内山郁夫
2. 発表標題 バクテリアのゲノムコア構造と遺伝子の水平伝搬
3. 学会等名 ConBio2017 フォーラム「遺伝子の水平伝搬とは何か」
4. 発表年 2017年

1. 発表者名 内山郁夫、三原基広、西出浩世、千葉啓和
2. 発表標題 公開ゲノムの全遺伝子レパートリーをカバーする微生物オーソログテーブルの構築
3. 学会等名 第12回ゲノム微生物学会年会
4. 発表年 2018年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

微生物比較ゲノムデータベースMBGD <a href="http://mbgd.genome.ad.jp/">http://mbgd.genome.ad.jp/</a>
---

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