（報告様式４）

【16gm6010003h0001】

平成２９年 ５月 ３１日

# 平 成 ２ ８ 年 度　委 託 研 究 開 発 成 果 報 告 書

## I. 基本情報

事　業　名：（日本語）革新的先端研究開発支援事業
（英　語）Advanced Research and Development Programs for Medical Innovation

研究開発課題名：（日本語）高完成度ドラフトゲノム構築による種内変異レベル解像度のメタゲノミクス
（英　語）High-resolution metagenomics for intra-species variations based on assembly of the comprehensive draft genomes

研究開発担当者　（日本語）国立大学法人東京工業大学　生命理工学院　助教　梶谷嶺
所属 役職 氏名：（英　語）Rei Kajitani, Assistant Professor, School of Life Science and Technology, Tokyo Institute of Technology

実 施 期 間：平成２８年１０月１日 ～ 平成２９年３月３１日

分担研究開発課題名：（日本語）
（英　語）

研究開発分担者　（日本語）
所属 役職 氏名：（英　語）

## II. 成果の概要（総括研究報告）

・ 研究開発代表者による報告の場合

・ 研究開発分担者による報告の場合

和文

当該期間の前半では、真核生物用 *de novo* アセンブラ Platanus (Kajitani et al. 2014 *Genome Res.*) にメタゲノム用の機能を追加した MetaPlatanus（β 版）をベースに、種内変異レベルのゲノム配列の差異を統合する機能を追加し、結果配列の長さの点でより高性能な *de novo* アセンブリツールの開発を進めた。ツールのベンチマークでは、paired-end および mate-pair 法でライブラリ調整された、Illumina シークエンサーの全メタゲノムショットガンデータを対象としている。公開されている牛の反芻胃サンプル（Hess et al. 2011 *Science*）を最初は使用していたが、insert-size（シークエンシング対象の DNA 断片長）が 5 kbp までの mate-pair ライブラリのみしか存在しておらず、DNA シークエンサのバージョンも古いという問題もあり、研究開発計画書に記した「N50 長が 100 kbp 以上」という目標性能を達成するには至らなかった。ここで、N50 長は結果配列の長さの指標であり、平均長より外れ値の影響を軽減できる利点から、*de novo* アセンブリ結果の評価に一般的に用いられる。前述の目標達成の条件を探るため、九州大学 後藤恭宏 博士より、シークエンシング済の牛の蹄のメタゲノムデータを提供して頂き、開発されたツールを適用した。当データは insert-size が 10 kbp までの複数の mate-pair ライブラリを含み、DNA シークエンサのバージョンもより新しい。その結果、混入した牛のゲノム配列を除いたアセンブリ結果の N50 長は約 170 kbp、合計長は約 360 Mbp となり、目標性能を達成した（未発表、成果発表は次年度以降に予定）。様々な既存のメタゲノムアセンブリ結果をまとめた論文（Kuleshov et al. 2016 *Nat. Biotech.*）を参照すると、先行研究では N50 長は 4.1–49 kbp（合計長は 300–656 Mbp）に留まっており、本研究で達成した N50 長はそれらを大きく引き離す値である。本手法により、培養過程を省略して環境 DNA サンプルから微生物ゲノムを構築し、解析が大幅に効率化されることが期待される。

当該期間の後半では、サンプルに含まれる菌株の全ハプロタイプを構築するアプローチのアルゴリズムを開発し、より解像度の高いメタゲノム解析の基盤構築を図った。手順としては、含まれるハプロタイプが 2 種類であることが明らかな 2 倍体生物種のデータへのアルゴリズム開発から開始した。このような対象データは、含まれるハプロタイプ数が不明かつ各株のシークエンシング量が不明なメタゲノムデータと比較すると、アルゴリズム開発は容易であると考えられる。結果として、既存の類似用途のツールより高精度なハプロタイプ配列を構築するツールを開発することができたため、本来の課題であるメタゲノム用のツール開発に移行した。多ハプロタイプの混合も想定した、*de novo* アセンブリ時のグラフ上の経路探索機能などを実装し、公開データであるヒト腸内メタゲノムサンプルに適用して、2 倍体生物種と同様に良好な結果を得たため、第 11 会日本ゲノム微生物学会年会にて成果を発表した（演題名：メタゲノムショットガンデータからの株ハプロタイプ配列構築手法の開発）。更に、10x Genomics 社の Linked reads と呼ばれる、数十 kbp の各 DNA 断片由来のリードにバーコードと呼ばれるタグ配列を付加する新手法のデータへの対応も開始した。九州大学 小椋義俊 博士、後藤恭宏 博士より、牛の腸内および蹄のメタゲノムサンプルを提供頂き、それを当該手法の受託サービスにて調整し、結果のデータを得ることができた。精度評価方法は検討の余地があるものの、バーコード情報を用

いて *de novo* アセンブリ時のグラフ上の分岐構造を解決する機能の実装は行ない、前述の牛の蹄メタゲノムデータに対するテストではハプロタイプ配列の N50 長は 95 kbp となった。当該期間終了時点では、類似の用途の手法は Illumina TruSeq synthetic long read と呼ばれるライブラリ調整法に基づくものが唯一であったが、先行論文（Kuleshov et al. 2016 *Nat. Biotech.*）で報告された 19 kbp の N50 長を大きく上回る値を達成することができたと考えている。

英文

In the former half of this period, I developed a novel *de novo* assembler that could merge intra-species variations in metagenomic data and generate highly contiguous assembled sequences. The development was based on MetaPlatanus (beta version), which was the modified version of Platanus genome assembler (Kajitani et al. 2014 *Genome Res.*) In benchmarks, whole-metagenome shotgun data from Illumina sequencers were input as paired-end and mate-pair libraries. At first, I used the published bovine-rumen metagenomic data (Hess et al. 2011 *Science*). Due to the limited insert-sizes (mean size of targeted DNA fragments) that were smaller than 5 kbp and relatively old version of a DNA sequencer, a targeted performance (N50 length ≥100 kbp) was not achieved. Note that "N50 length" is an indicator of contiguity of assembled sequences and a summed length of sequences longer than N50 length correspond to 50% of entire total length. Compared to mean length, it can effectively exclude outliers. To investigate a condition to achieve the performance above, I was offered a bovine hoof metagenomic sample, which had been sequenced previously, from Dr. Yasuhiro Gotoh (Kyushu University). The sample had been prepared as mate-pairs of long insert-sizes (3–10 kbp) and sequenced by newer version of Illumina facilities. As a result, the developed assembler recorded N50 length of 170 kbp and total size of 360 Mbp (bovine genomic sequences were excluded), exceeding the targeted performance (under preparation for publication). In a reference of a comparison of existing methods (Kuleshov et al. 2016 *Nat. Biotech.*), N50 lengths remained 4.1–49 kbp (total sizes, 300–656 Mbp), and my method overwhelmed these ones. Applying my method to environmental DNA samples directly, it is expected that analysis of microbial genomes are substantially accelerated skipping culture processes.

In the latter half of the period, I developed an algorithm to construct whole haplotypes in metagenomic data, aiming at high-resolution analysis of intra-species variations. As the first step, data from diploid organisms, which were relatively simple in terms of the limited number of haplotypes, were targeted. Note that metagenomic data may contain more than two haplotypes (strains) for each species and relative abundances are unknown in advance. After development of a haplotype assembler for diploid organisms that performed high accuracy and contiguity compared to existing tools, I shifted to an original plan for metagenome analysis. Consequently, functions to search paths in graph structures used in *de novo* assembly of genomic data with many haplotypes were implemented, and a benchmark using published human gut data indicated contiguous resultant haplotypes. The method and results were reported in 11th annual conference of Society of Genome Microbiology (title, "Development of an assembly method to construct strain haplotypes from metagenomic shotgun data"). In addition, I started to deal with the data of 10x Genomics technology, called Linked reads. In this protocol, a short tag sequence ("barcode") is attached to reads from each

DNA fragment whose length ranges 10–200 kbp. I offered DNA samples of bovine hoof and gut from Dr. Yoshitoshi Ogura and Dr. Yasuhiro Gotoh (Kyushu University), and obtained Linked reads data of those metagenomic samples through a commercial service. Although there were not widely used metrics to measure accuracy of assembled haplotypes in metagenomes, I measured N50 length of haplotype sequences from the developed tool, resulting in 95 kbp (bovine hoof sample). At the time of the end of this period, the one and only competing method was based on the "Illumina TruSeq synthetic long read" protocol (Kuleshov et al. 2016 *Nat. Biotech.*). In that article, N50 length of haplotype sequences was reported as 19 kbp, and I supposed that my method significantly surpassed the existing one in terms of contiguity.

## III. 成果の外部への発表

（１）学会誌・雑誌等における論文一覧（国内誌　０　件、国際誌　０　件）

（２）学会・シンポジウム等における口頭・ポスター発表

1. メタゲノムショットガンデータからの株ハプロタイプ配列構築手法の開発, ポスター, 梶谷嶺, 小椋義俊，後藤恭宏，吉村大，奥野未来，林哲也，伊藤武彦，第 11 回日本ゲノム微生物学会年会，2017/3/4，国内.

（３）「国民との科学・技術対話社会」に対する取り組み

（４）特許出願