論文 / 著書情報
Article / Book Information

題目(和文)	Hi-C法を活用した染色体レベルのハプロタイプゲノム構築手法の開発
Title(English)	Development of a chromosome-level haplotype-resolved genome assembly tool using Hi-C
著者(和文)	大内俊
Author(English)	Shun Ouchi
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12745号, 授与年月日:2024年3月26日, 学位の種別:課程博士, 審査員:伊藤 武彦,本郷 裕一,立花 和則,二階堂 雅人,山田 拓司
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12745号, Conferred date:2024/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	生命理工学 系 生命理工学 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 工学 ）
学生氏名： Student's Name	大内 俊	審査員主査： Chief Examiner	伊藤 武彦

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

分子生物学研究をする上で，対象生物のゲノム配列情報は不可欠であり，様々な生物のゲノム配列が決定されてきている。未知のゲノム配列を新規に決定することは新規ゲノム配列構築(*de novo* アセンブリ)と呼ばれ，一般的に以下の流れで行われる。まず対象生物のゲノム DNA をシーケンサーで読み取り，リードと呼ばれる短いゲノム配列情報を得た後，オーバラップ情報を元に計算機上で繋げ合わせ contig と呼ばれる配列を作成する。その後，Mate-Pair などの追加情報を元に scaffolding と呼ばれる contig をさらに繋ぎ合わせ長い配列を得る工程が行われる。より長く繋がったゲノム配列を構築するために，Mate-Pair やロングリードを使う方法など様々な方法が開発されてきたが，セントロメアに代表される長い繰り返し配列の問題により，染色体レベルで繋がったゲノムを構築するのは難しいのが現状である。また，ヒトなど二倍体生物のゲノムは，父親，母親からそれぞれ引き継いだ相同染色体の配列(ハプロタイプ)で構成されており，ハプロタイプ間の違いを考慮し両親由来のハプロタイプを分けて構築する phasing を行うことが理想であるが，染色体レベルで phasing を実現することは困難である。そのため近年 Hi-C 法という空間的に近接しているゲノム領域の情報を網羅的に取得できる方法が scaffolding，phasing に応用されているが，既存の Hi-C データを用いた scaffolding，phasing ツールでは，参照配列の使用なしに連続性の高く高精度なハプロタイプを構築するのは難しいという課題がある。そこで，本研究では Hi-C 法を用いて染色体レベルの高精度なハプロタイプを構築するツールの開発を目的とした。

本研究では，Hi-C データを用いて scaffolding，phasing を行い染色体レベルで繋がったハプロタイプを構築するツール GreenHill を新規に開発した。GreenHill では，入力として既存ツールで作成した contig を用いており，はじめに contig 内の両親由来の配列から対応する配列を検出し一つにまとめた後，scaffolding を行い，ロングリードや Hi-C データを用いて染色体レベルに contig を繋げる。最後に phasing を行い，ロングリードや Hi-C を用いて母親由来，父親由来の配列に分けることで染色体レベルで繋がったハプロタイプを構築する。GreenHill には，既存ツールにはない独自の機能として，Hi-C データだけでなくロングリードの情報も scaffolding 時に用いる機能や Hi-C のコンタクトマップを用いたミス検出・修正機能があり，連続性が高くかつ高精度にハプロタイプを構築できるようになっている。

様々な生物種のデータを用いてベンチマークを行い，GreenHill の性能を既存ツールと比較した。はじめに Hi-C のシミュレーションデータを用いてテストを行い，染色体の立体構造の影響のない理想的な条件下で，GreenHill の基本性能を確認した。第一のベンチマークでは線虫の PacBio CLR リード，第二のベンチマークではショウジョウバエの PacBio HiFi リードを用いてベンチマークを行い，CLR と HiFi リードのどちらが入力の場合でも，GreenHill が既存ツールよりも連続性高く高精度なハプロタイプ配列を構築できることを示した。次に様々な生物種の実データに対して GreenHill を適用し，実データでの有効性を検証した。第三のベンチマークではゲノムサイズが約 3 Gb と大きいウシの実データを用いており，ゲノムサイズが大きい実データでも GreenHill が高い性能を発揮することを確認した。第四のベンチマークではキンカチョウの実データを用いて，ヘテロ接合度が 1.47 %と高くハプロタイプ間の違いが大きいサンプルでも GreenHill が染色体レベルのハプロタイプを構築できることを示した。第五のベンチマークでは，セキセイインコ，クロサイ，コチョウザメのデータを用いて様々な生物種で高い連続性，精度のハプロタイプ配列を構築できることを示し，GreenHill の堅牢性が高いことを示した。

本論文では，Hi-C 法を用いて高精度で染色体レベルのハプロタイプを構築することができる scaffolding，phasing ツール GreenHill を新規に開発し，様々な生物種のデータを用いたベンチマークでその有用性を示した。ゲノム構築に GreenHill を活用することで，多くの生物種の染色体レベルのハプロタイプ配列がより簡便に構築できるようになり，ハプロタイプ間の構造変異の解析や遺伝子発現量の解析など様々な下流解析を容易にすることが期待される。

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	生命理工学 系 生命理工学 コース	申請学位（専攻分野）： Academic Degree Requested	博士 (工学) Doctor of
学生氏名： Student's Name	大内 俊	審査員主査： Chief Examiner	伊藤 武彦

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

Most higher eukaryotic organisms, including humans, are diploid and possess two copies of homologous chromosomes. A nucleotide sequence from one chromosome is called a haplotype, and high-quality haplotype-resolved genomes, constructed by separating homologous chromosomes, are an important resource in molecular biology. However, most standard de novo assemblers ignore differences between haplotypes and generate mosaic haplotypes that do not exist, thereby misleading biological interpretation in downstream analysis. Tools for haplotype-resolved genome assembly, which reconstruct each haplotype from sequencing data, are developed, however they require parental data or reference genomes and often fail to provide chromosome-level haplotypes.

In this study, I developed Greenhill, a novel scaffolding and phasing tool using Hi-C to generate chromosome-level haplotypes. GreenHill receives contigs constructed from reads by other assemblers as input. First, the merge haplotype step is performed to detect contig pairs that consist of the same loci from homologous chromosomes and merge each contig pair into a single contig as a consensus contig. Second, the consensus scaffolding step is performed to connect and order contigs at the chromosome-level using long reads and Hi-C. Finally, the phasing step is performed to construct a chromosome-level haplotype by separating maternal and paternal haplotypes using long reads and Hi-C. Unique functions of GreenHill include new error correction based on Hi-C contacts and the simultaneous use of Hi-C and long reads, enabling to build haplotypes with high continuity and accuracy.

The performance of GreenHill on de novo assemblies was evaluated using both simulation data and actual data from a variety of species. Benchmarks revealed that GreenHill outperforms other existing tools in contiguity and phasing accuracy. GreenHill, which can construct high-quality chromosome-level haplotypes, is expected to facilitate a large variety of downstream analyses, such as structural variation analysis or gene analysis between haplotypes.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).