# 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),<br>学位授与機関:東京工業大学,<br>報告番号:甲第12745号,<br>授与年月日:2024年3月26日,<br>学位の種別:課程博士,<br>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),<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第12745号,<br>Conferred date:2024/3/26,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Category(English) | Doctoral Thesis |
| 種別(和文) | 審査の要旨 |
| Type(English) | Exam Summary |

（博士課程）

# 論文審査の要旨及び審査員

| 報告番号 | 甲第　　号 | | 学位申請者氏名 | | 大内 俊 | |
|---|---|---|---|---|---|---|
| 論文審査<br>審査員 | | 氏　名 | 職　名 | | 氏　名 | 職　名 |
| | 主査 | 伊藤 武彦 | 教授 | 審査員 | 山田 拓司 | 准教授 |
| | 審査員 | 本郷 裕一 | 教授 | | | |
| | | 立花 和則 | 准教授 | | | |
| | | 二階堂 雅人 | 准教授 | | | |

論文審査の要旨（2000字程度）

本論文は「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（Development of a chromosome-level haplotype-resolved genome assembly tool using Hi-C）」と題し，二倍体高等真核生物ゲノム配列を対象に，Hi-C 法データを活用することによって効果的に Scaffolding，Phasing を実現する情報解析手法の新規開発および，開発手法のゲノムデータへの適用結果について述べたものであり，以下の４章より構成されている。

第一章「序論」では，まずゲノム配列決定の概要を述べ，その工程がシーケンサーで読み取った短いゲノム断片配列を計算機上で繋ぎ合わせ contig と呼ばれる配列を作成する段階と，contig を様々な配列情報に基づいてさらに繋ぐ Scaffolding と呼ばれる段階からなることを示している。また近年では，二倍体生物のゲノム決定において，父親と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)を区別し，両親由来の配列をそれぞれ構築する Phasing を行うことも求められていることが述べられている。続いて，Scaffolding および Phasing 時の問題点について示し，現存するツールの使用のみでは，連続性が高く高精度なハプロタイプ別ゲノム配列を構築することは難しいという課題を指摘し，染色体レベルのハプロタイプを区別したゲノム配列構築が可能な新規手法の開発が必要であると述べている。

第二章「Hi-C Scaffolding、Phasing 手法 GreenHill の開発」では，既存ゲノムアセンブラからの出力 contig を入力とし，染色体高次構造解析手法である Hi-C 法のデータを活用して，Scaffolding，Phasing を行う新規開発手法 GreenHill のアルゴリズムを説明している。GreenHill では，配列相同性と coverage 情報に基づき対立アレル由来の入力 contig 同士を対応づけた上で統合し，統合された contig の Scaffolding を行った後に，Phasing により両アレル配列を構築するアルゴリズムを採用していることを述べている。また，既存手法にはない GreenHill 独自の機能として，Hi-C データに加えてロングリード情報も Scaffolding 時に合わせて用いる機能や，Hi-C のコンタクトマップを用いてミス検出・修正を行う機能を導入することにより，染色体レベルのハプロタイプ別配列構築を実現していることも合わせて述べている。

第三章「ベンチマーク」では，様々な生物種のデータに対して GreenHill を適用することにより，有用性を示している。前半では，シミュレーションによる Hi-C データを用いることで，染色体立体構造の影響のない理想条件下での GreenHill の性能を既存ツールと比較しており，ゲノム既知の参照配列を用いて詳細な精度を評価している。後半では，GreenHill を様々な生物種の実データに適用することで有効性を検証しており，トリオデータ（両親と子供のデータ）を用いて Phasing 精度を評価している。具体的にはシミュレーションデータを用いたベンチマークとして，線虫の PacBio CLR リード，ショウジョウバエの PacBio HiFi リードを用いてテストを行い，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が高くかつ，高精度なハプロタイプ配列を構築できることを示している。実データを用いたベンチマークでは，ゲノムサイズが大きいウシ（約 3Gb），ヘテロ接合度が高いキンカチョウ（約 1.47%），さらには幅広い実サンプルとしてセキセイインコ，クロサイ，コチョウザメのデータを用いており，様々な生物種で高い連続性，高い精度のハプロタイプ配列構築ができることを示し，GreenHill の堅牢性の高さを述べている。

第四章「総括」では，上記３章のまとめと展望を述べており，本研究で開発された GreenHill の応用可能性と課題，今後の展望について論じている。

以上を要するに，本論文は Hi-C 法を用いて高精度に染色体レベルのハプロタイプ配列を構築することができる Hi-C 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を用いてその有用性・応用可能性を示したものであり，工学上ならびに工業上貢献するところが大きい。よって，本論文は博士（工学）の学位論文として十分な価値があるものと認められる。