jh240015

# 長鎖型シークエンスに基づくハプロタイプカタログ構築と異なるクラウド拠点間での横断的バッチジョブシステム試験実装

長﨑正朗（九州大学　生体防御医学研究所）

概要

本研究課題は２年間のプロジェクトの１年目にあたる。独自に取得した高深度の日本人の長鎖型シークエンス情報と海外の長鎖型シークエンス情報とを統合し、この情報を鋳型として用いることで、国内外の集団における遺伝子全長の配列をより高精度で取得整備することを目的とし研究を進めている。これにより日本人の遺伝子の集団としての特性、また、疾患研究に資する遺伝子のより精密なハプロタイプパターン理解が可能となる。令和６年度は約 19,000 遺伝子に対する 360 万の鋳型のドラフト版の構築を完了した。さらに、難読化領域に対する解析技術を活用することでソフトウェア開発をおこなうことができた。また、これらの情報は公開可能なヒトゲノム情報であることから、いままで構築をすすめてきている CPU と GPU 電算資源双方を必要とするハイブリッドクラウド(jh220014, jh230016)について課題であった複数のパブリッククラウドを横断的に電算機資源としてシームレスに利用することを試験的に進めている。

## 1. 共同研究に関する情報

### ⑴ 共同利用・共同研究を実施している拠点名

東京大学 情報基盤センター

京都大学 学術情報メディアセンター

九州大学 情報基盤研究開発センター

mdx

### (2) 課題分野

データ科学・データ利活用課題分野

### (3) 参加研究者一覧と役割分担

九州大学の長﨑のチーム（他、関谷弥生、男澤、寺岡、町田、松原、浅倉、橋本）：ハプロタイプパネル構築に関連したソフトウェア調査、電算機資源の実行スクリプトの作成、および実行支援

東京大学（塙、関谷）：東京大学の電算機資源（Wisteria）、および、大規模仮想環境（mdx）での最適利用に関連したアドバイス、また、試験環境の整備

情報通信研究機構（村田）：拠点間の高速データ転送プログラムの提供

京都大学（深沢）：京都大学の計算機資源におけるデータの効率的な保存

九州大学の大川のチーム（前原、南里）：シークエンサからの拠点間データ転送

国立情報学研究所の竹房チーム（他、大江、丹生、合田）：学認クラウドオンデマンド構築サービスのソフトウェア群を用いた複数クラウド間でのシームレスなジョブ実行のための SINET6 の設定、ソフトウェア設定と改修

本研究課題は、ゲノムサイエンス（長・短鎖ゲノムシークエンス取得（大川）、ゲノム情報解析（長﨑（九大）、松田（京大））、クラウド管理（mdx（塙、関谷）、NII（合田））、ネットワーク管理（南里（九大）、深沢（京大）、

関谷（東大)）、大規模計算資源管理（塙、関谷、南里、深沢)、ネットワーク転送（村田)、クラウド統合（竹房、大江）の専門性が異なる多数の異分野融合によるチームで構成をしており、拠点公募型共同研究として初めて研究を推進を進めている。

## 2. 研究の目的と意義

本研究課題は２年間のプロジェクトの１年目にあたる。

独自に取得した高深度の日本人の長鎖型シークエンス情報と海外の長鎖型シークエンス情報を統合して鋳型として用いることで、国内外の集団における遺伝子全長の配列をより高精度で取得整備することを目的とする。

また、これらの情報は公開可能なヒトゲノム情報であることから、いままで構築をすすめてきている CPU と GPU 電算資源双方を必要とするハイブリッドクラウドについて課題であった複数のパブリッククラウドを横断的に電算機資源としてシームレスに利用することを目指す。そのために、本研究で必要となる一部の計算について、国立情報学研究所が進めている学認クラウドオンデマンド構築サービスで提供されているソフトウェア群を活用することで mdx を含めて試験的に環境整備と実行を行う。

これにより日本人の遺伝子の集団としての特性、また、疾患研究に資する遺伝子のより精密なハプロタイプパターン理解、さらに、ゲノムサイエンスにおける解析環境構築のリファレンス実装を進める。

## 3. 当拠点の公募型共同研究として実施した意義

（課題の学際性）共同研究の推進にあたって構成拠点において研究グループや研究者の協力が必要な項目に記載したとおり、ゲノムサイエンス（長・短鎖ゲノムシークエンス取得（大川)、ゲノム情報解析（長﨑（九大)、松田（京大)))、クラウド管理（mdx (塙、関谷)、NII (合田))、ネットワーク管理（南里（九大)、深沢（京大)、関谷（東大))、大規模計算資源管理（塙、関谷、南里、深沢)、ネットワーク転送（村田)、クラウド統合（竹房、大江）の専門性が異なる多数の異分野融合によるチームで構成をしており、拠点公募型共同研究として初めて研究を推進できる。

（当拠点資源利用の必要性、研究の意義）本研究提案の解析においては GPU を含む大規模な計算資源、また、効率的な各拠点での解析が必要となる。約 260 検体の中・高深度の長鎖型ゲノムに基づくハプロタイプパネルの構築において最低各ノード当たり 192G 程度の電算機資源、また、過去の実績から CPU の 8 ノード分の通年の電算資源と GPU の 1 ノードの電算資源、200TB のストレージが想定されている。そこで、今回の申請において、各拠点でどのような解析を行うことで効率的に運用、セキュリティを担保した運用、また、将来的な情報量の増加に対応するか実際に設計・運用を行うことで検討を進める。それらの解決のために、各解析拠点のネットワーク、大規模解析、バイオインフォマティクス専門の研究者の融合した知識が必要である。また、申請者は日本人の長鎖型シークエンサについて、100 検体についてさらにシークエンス情報を積み増すことで拡充し、より正確なハプロタイプ情報を整理できる状況にあること、いくつかの遺伝子について生物学的に有用な成果（Hirayasu *et al* Front Immunol, Nagasaki *et al* Human Immunol.）を得ていることから計算科学だけでなくゲノムサイエンスでも貢献できると考えている。

## 4. 前年度までに得られた研究成果の概要

**該当なし**

**今年度の研究成果の詳細**

ヒトゲノム情報についてシークエンス技術の開発により爆発的に出力される情報が増えてきている。これらの情報について、情報量の増加とともに適切な計算環境において計算を行うこと、また、大規模演算により得られた計算結果を複数拠点にバックアップを持つなどの運用が必要となる。そこで、オンプレ、国内のスーパーコンピュータシステム、また、商用のクラウド環境の各々において、転送のコスト、費用、セキュリティなど総合的に勘案をして運用を行う必要がある。そこで、申請者は複数拠点間にわたる計算資源、ストレージを効率的に運用するにおいて出てくる課題に対し上の一部の情報についての試験的な解析を円滑に行うことを「ハイブリッドクラウド構築とゲノム情報解析の効率的な運用に関した研究（令和2−3年度 jh200047-NWH, jh210018-NWH）」において進めまた論文として成果を報告した（Tanjo *et al* Journal of Human Genetics 2021, Nagasaki *et al* Human Genome Ver 2023）。

一方、近年、長鎖型法（1つの DNA 断片の読み取り長が 15,000 塩基以上）により全ゲノムデータの取得が進められている。申請者も令和 4 年度においては 50 検体、令和 5 年度においては 100 検体の長鎖型シークエンサの情報（低深度）を取得し、これらの情報を鋳型として用いることで、ハイブリッドクラウド内において、短鎖型法（1つの DNA 断片の読み取り長が約 300 塩基）で取得された約 5,000 人の全ゲノム情報との統合解析を進め構造多型のパネルの構築を進めてきている（「ハイブリッドクラウドを用いたゲノム情報に基づく構造多型パネルの構築とアノテーション（jh220014, jh230016）」）（Hirayasu *et al* Front Immunol）。

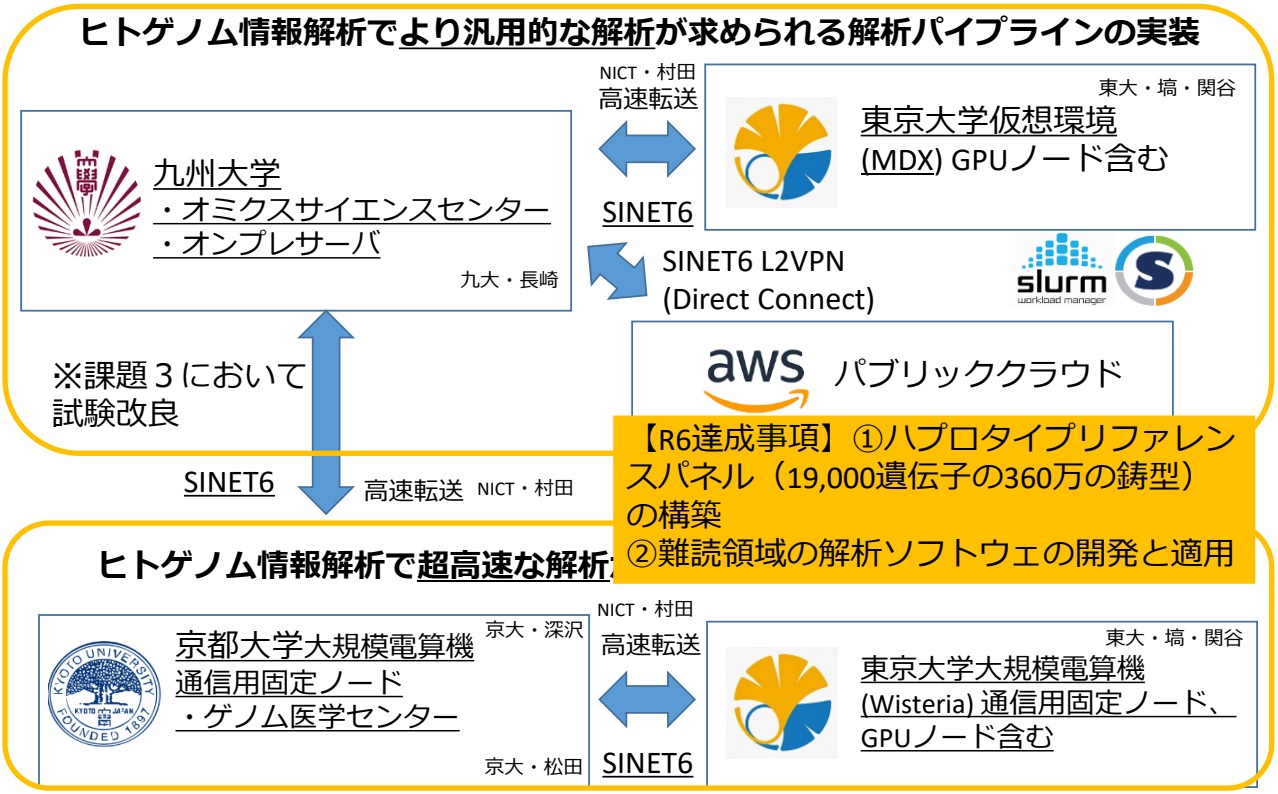
令和 4、5 年度においては、鋳型とする長鎖型シークエンスの情報取得コストが高額であることから、ヒトゲノム全長の被覆深度は平均して 10x 程度（低深度）であり、特に長い遺伝子等については扱いが難しいという課題があった。そこで令和 5 年度後半では、約 50 検体について追加で長鎖型シークエンスを行うことで被覆深度を 20-30 以上（中から高深度）にすることを進めてきた。また、海外において同一のシークエンサにおいて取得された欧米を中心とした 100 検体以上の情報（被覆深度 60x 以上（超高深度））が公開されている。

本研究課題では、2 年間のプロジェクトの 1 年目として、独自に取得した高深度の日本人の長鎖型シークエンス情報と海外の長鎖型シークエンス情報とを統合し、この情報を鋳型として用いることで、国内外の集団における遺伝子全長の配列をより高精度で取得整備することを目的とし、課題 1，2，3 を設定し研究を進めた。

課題 1）中高深度長鎖シークエンス情報に基づくハプロタイプリファレンスパネルの構築とそのための複数拠点間のハイブリッドクラウド情報基盤の運用（長﨑、関谷、塙、深沢、大川、松田）（図 1 に概念図と成果概要を示す）

令和 5 年度に取得した 105 検体の長鎖型シークエンス情報、および、令和 6 年度の前半に課題 2 において同一検体に対して積み増しシークエンス情報として新たに取得する長鎖型シークエンス情報、海外で取得されている 153 検体の超高深度シークエンス情報を用いることで、より高精度なハプロタイプリファレンスパネルの構築を進めている。現在のところ約 19,000 遺伝子に対す

図 1 課題１）中高深度長鎖シークエンス情報に基づくハプロタイプリファレンスパネルの構築とそのための複数拠点間のハイブリッドクラウド情報基盤の運用 長崎、関谷、塙、深沢、大川、松田　システム全体構成と役割担当




るハプロタイプリファレンス（360 万の鋳型）の構築を進めた。

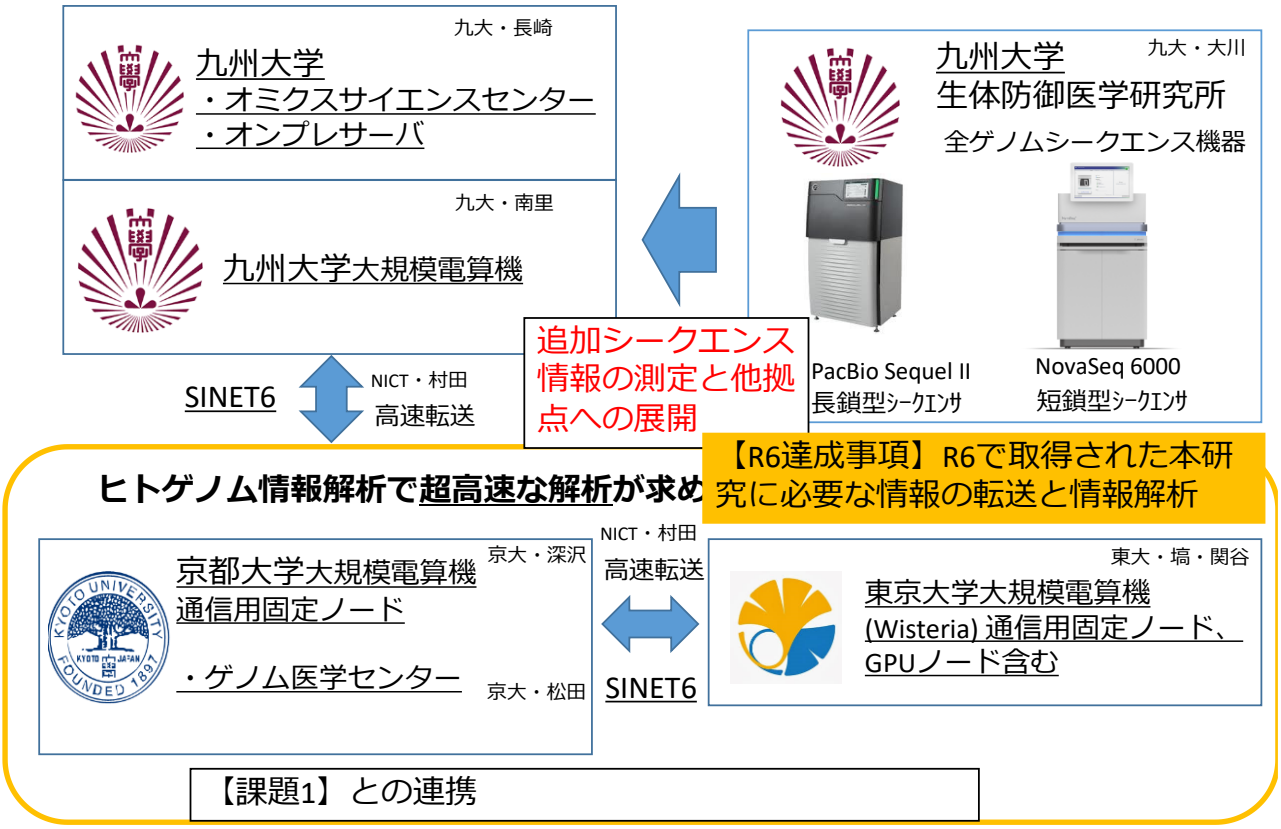
また、鋳型情報がそろうことで可能となった白血球免疫グロブリンスーパーファミリー受容体（LILR）が位置する難読領域の解析を進め、JoGo-LILR Caller のソフトウェアの公開、また、同ソフトウェアを用いることで、3,202 人の LILRB3 および LILRA6 のコピー数構造を同定することに成功した。

具体的には、LILR ファミリは、ヒト 19 番染色体 q13.4 領域にコードされており、遺伝的多様性に富む 11 種類の受容体から構成されており、中でも LILRB3 および LILRA6 は、リガンド結合に関与する免疫グロブリン様ドメインにおいて高い配列相同性を示し、コピー（CN）構造が個人毎に差があることが知られていたが、その正確な機能の解明は困難であった。

この課題に取り組むため、前述の「JoGo-LILR Caller」アルゴリズムおよびツールを開発した。本ツールにより、集団規模の全ゲノムショートリードシーケンシングデータから LILRB3 および LILRA6 のコピー数を同時に推定することが可能となった。また、JoGo-LILR Caller を、国際 HapMap 由来の 2,504 検体に適用することで、５大陸の CN 構造を構築した。さらに、このプロファイルは、Human Pangenome Reference Consortium（HPRC）が提供する長鎖型シークエンサ 40 検体のパンゲノムリファレンスアセンブリ由来の CN データと完全一致（100%の一致率）し、その妥当性を確認した。

他に、LILRB3-LILRA6 のコピー数ハプロタイプ構造（前述の CN 構造は 2 倍体における構造、この構造は 1 倍体における構造)の頻度を 5 つの大陸集団に対して推定し、グローバルなハプロタイプレベルの CN プロファイルを確立した。これにより、本

図２　**課題２）長鎖シークエンサから取得する情報を他拠点に効率良く展開するための設計検討と実装　大川、南里、長﨑、深沢、村田**




ツールは LILRB3-LILRA6 のハプロタイプ CN の 2 倍体におけるペアも推定可能となった。

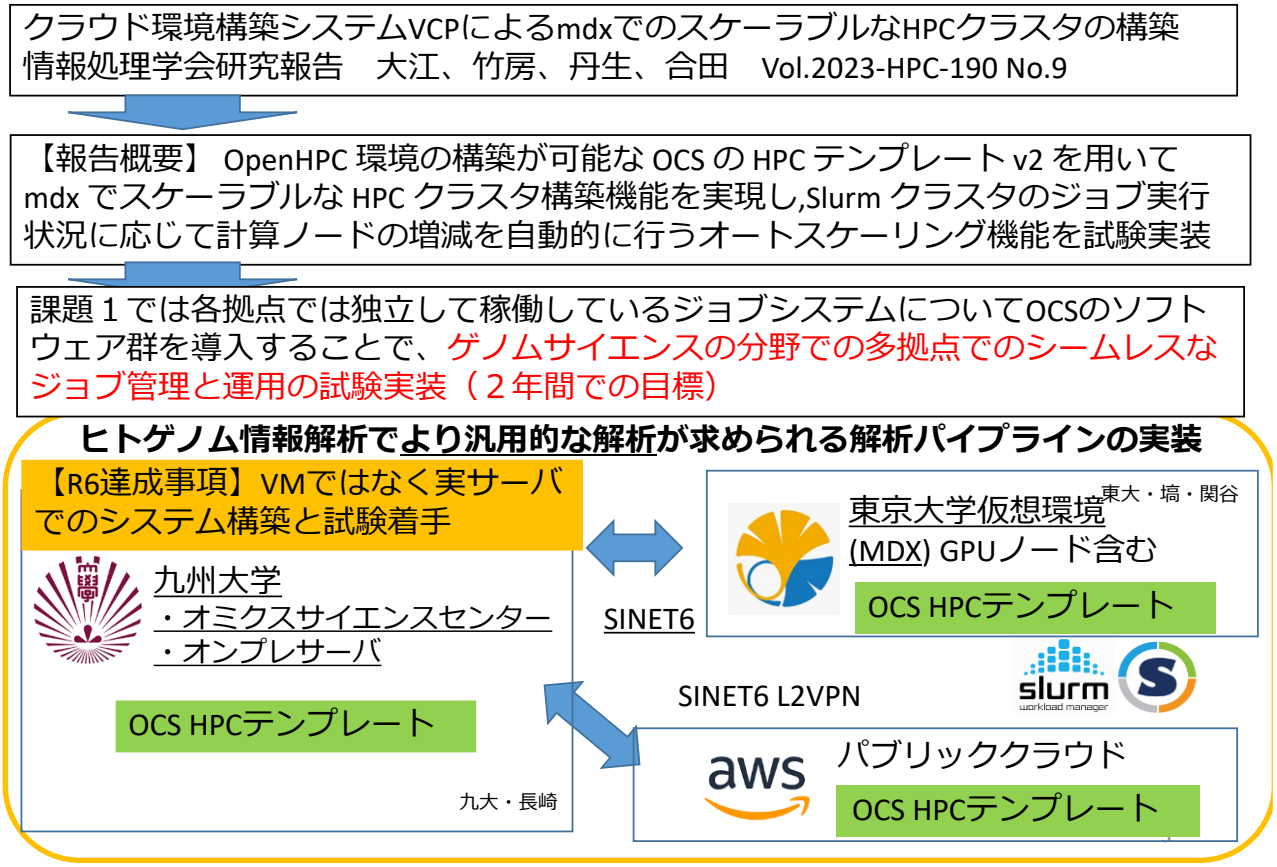
さらに、JoGo-LILR-trio は、親子三者データセットにおけるハプロタイプペアの予測精度を向上させ、40 例の子の検体において予測されたハプロイド CN 型のペアと、リファレンスアセンブリで確認された二倍体構造との間に 100%一致することを確認した。

本ツールは今後、SNP アレイによるジェノタイピングデータから LILRB3-LILRA6 の CN 型をインピュート（推定）するソフトウェア開発や、炎症性腸疾患や高安動脈炎など、多様な表現型・疾患との関連解析を可能にする基盤としての活用を予定している。

課題２）長鎖及び短鎖シークエンサから取得する情報を他拠点に効率良く展開するための設計検討と実装（大川、南里、長﨑、深沢、村田）（図２に概念図と成果概要を示す）

課題１でハプロタイプ構築を行う長鎖型シークエンサの情報に関連した DNA 取得用の不死化リンパ球を九州大学（大川）が管理を行っている。本研究課題においては、長鎖型シークエンサが導入されている九州大学のオンプレミス環境とハイブリッドクラウド間で長鎖型シークエンサによって得られた低深度から高深度情報とするために取得したシークエンス情報を効率よく他拠点へ転送するための設計や性能試験を継続して進めている。後半期間は Hi-C などのデータを取得し、これらのデータの転送を重点的に進めた。

図３　**課題３）複数クラウドにおけるシームレスなジョブ管理のためのハイブリッドクラウド情報基盤構築と試験運用**　　**長﨑、竹房、大江、丹生、合田**




課題３）複数クラウドにおけるシームレスなジョブ管理のためのハイブリッドクラウド情報基盤構築と試験運用（長﨑、竹房、大江、丹生、合田）（図３に概念図と成果概要を示す）

課題１で構築を進めるハイブリッドクラウドは各拠点単体で slurm ノードやジョブ管理システムが稼働しており複数拠点における統合したジョブ管理を行う仕組みとはなっていない。そのため、一時的な情報解析のために、課題２などで準備されるシークエンス情報を異なる拠点に高速転送し、その後、ジョブを投入するというマニュアルなステップが必要となることが課題であった。

本課題では、学認クラウドオンデマンド構築サービスで開発が進められている HPC 向けのオンプレミスを VM の管理サーバとすることで、IPSec もしくは L2VPN を用いたネットワーク上で複数のクラウドの電算資源をシームレスにジョブ管理することができるシステムである。本研究課題においては、同環境をカスタマイズしつつ試験的に mdx とオンプレ、パブリッククラウド間で試験構築し、課題１で行う解析の一部を試験的に実行することで課題の洗い出しと個別に解決することを目的として進めようとしている。これにより将来的によりシームレスなハイブリッド統合環境構築の実装を進める。

当初は、VMWare を入れることで、必要とされるシステム構築を進める予定であったが、買収などの影響でライセンスを維持することが困難であると判断したため、本年度は、九州大学で本プロジェクト専用のサーバを設置する代替案を進め、年度内に構築を完了することができた。

## 5.　進捗状況の自己評価と今後の展望

課題１）JHPCN で割り当てられた CPU/GPU の電算資源を活用して 19,000 遺

伝子に対し、国内外の 258 名の検体から約 360 万個の鋳型のドラフト版の生成を進めるとともに、鋳型生成の過程で得られた技術を活用することで関連成果（Nagasaki *et al.* Human Immunol. 2024, Hirayasu *et al.* Front Immunol. 2024）を得ることができた。当初目標は十分に到達していると考える。

令和6年度内に海外でさらなる検体のシークエンスが行われており、令和 7 年度はそれらに対する鋳型生成のための追加解析を継続して進めていく。ストレージや電算資源が不足する場合には適宜追加契約を行って当初目標を達成していく。

課題２）令和６年度に、本プロジェクトに関連する長鎖型シークシークエンサ、および短鎖型シークエンサの情報は、九州大学のオンプレミスに格納された後、九州大学 情報基盤研究開発センター、京都大学、東京大学のシステム間のデータ転送を進めることで遅滞なく情報解析を進めることができた。

来年度においても、継続して同情報解析網を通じて解析を進めていく。

課題３）令和６年度後半に JHPCN でアサインされているチケットの範囲で mdx における学認クラウドオンデマンド設定と試験実行を行うための準備を進めてきた。本年度は、九州大学で本プロジェクト専用のサーバの設置と構築を完了したことから、来年度は、同サーバとパブリッククラウドや mdx との接続試験を課題１の一部の実行ができるように試験を進めていく予定である。当初予定としては、オンプレミスと１つのクラウド拠点との間でのジョブ実行を予定していたので、来年度に少し注力して進める必要があると考えている。

**※7. 研究業績はウェブ入力です**