jh250010

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 jh240015 の継続課題であり２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た。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７年度は 19,194 遺伝子に対する 370 万ハプロタイプ辞書として一般公開（JoGo1.0; https://jogo.csml.org/）と国際誌への掲載を完遂した。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）を効率的に運用することで本研究を達成することができた。

1. 共同研究に関する情報

(1) 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 I

(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、南、Huangfu、Chen、Owusu、Tang）は、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、および実行支援を行った。

東京大学の塙、関谷は、東京大学の電算機資源(Miyabi)、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、試験環境の整備を行った。

拠点間の高速データ転送については、情報通信研究機構 村田が開発を進めている実装を用いた。

また、京都大学の計算機資源におけるデータの効率的な保存については、京都大学の深沢らが整備を行った。シークエンサからの拠点間データ転送においては、九州大学の大川のチーム（前原、南里）で得られたデータの転送の担当を行った。

さらに、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については、SINET6 の設定、ソフトウェア設定と改修について、国立情報学研究所の竹房チーム（他、大江、丹生、合田）が支援を行った。

共同研究の推進にあたってゲノムサイエン

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した。

2. 研究の目的と意義

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。本研究課題では、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している (図１)。

長﨑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 (Nagasaki *et al Hum Immunol* 2025) 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。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た CPU と GPU 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。

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、関谷 (弥)、浅倉、男澤、寺岡、橋本、町田、松原 (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)

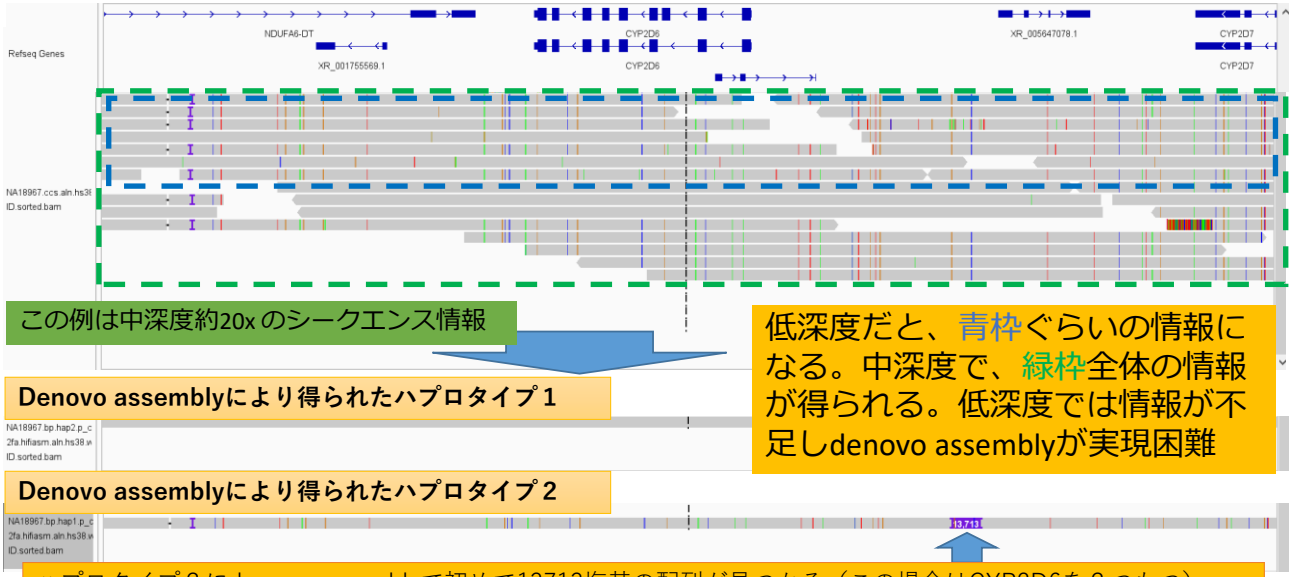
本研究提案の解析においては GPU を含む大規模な計算資源、また、効率的な各拠点での解析が必要となる。令和 6 年度の 258 検体の中・高深度の長鎖型ゲノムに基づくハプロタイプパネルの構築において、最低各ノード当たり 128G 程度の電算機資源、また、過去の実績から CPU の 8 ノード分の通年の電算資源と GPU の 1 ノードの電算資源、300TB のストレージが必要となった。令和 6 年度内に海外で約 200 検体のシークエンスが令和 6 年度に行われており、令和 7 年度はそれらに対する鋳型生成のための追加解析が必要となり昨年度と同等の計算資源が必要となる。

図１

【研究目的】令和6年度の国内外の258検体に加え、新規の海外の200検体の中・高深度の長鎖型シークエンス情報の情報解析を行うことで昨年度に構築を進めた情報解析を適用することで高精度なハプロタイプリファレンスパネルの拡充を進めること

denovo assemblyの実例

ヒト遺伝子領域 CYP2D6 (薬の代謝に関係する遺伝子)



※父親、母親から１本づつ引き継ぐためハプロタイプ１とハプロタイプ２がある

4. 前年度までに得られた研究成果の概要

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（Nagasaki *et al Hum Immunol* 2025）をおこなうことができた。

5. 今年度の研究成果の詳細

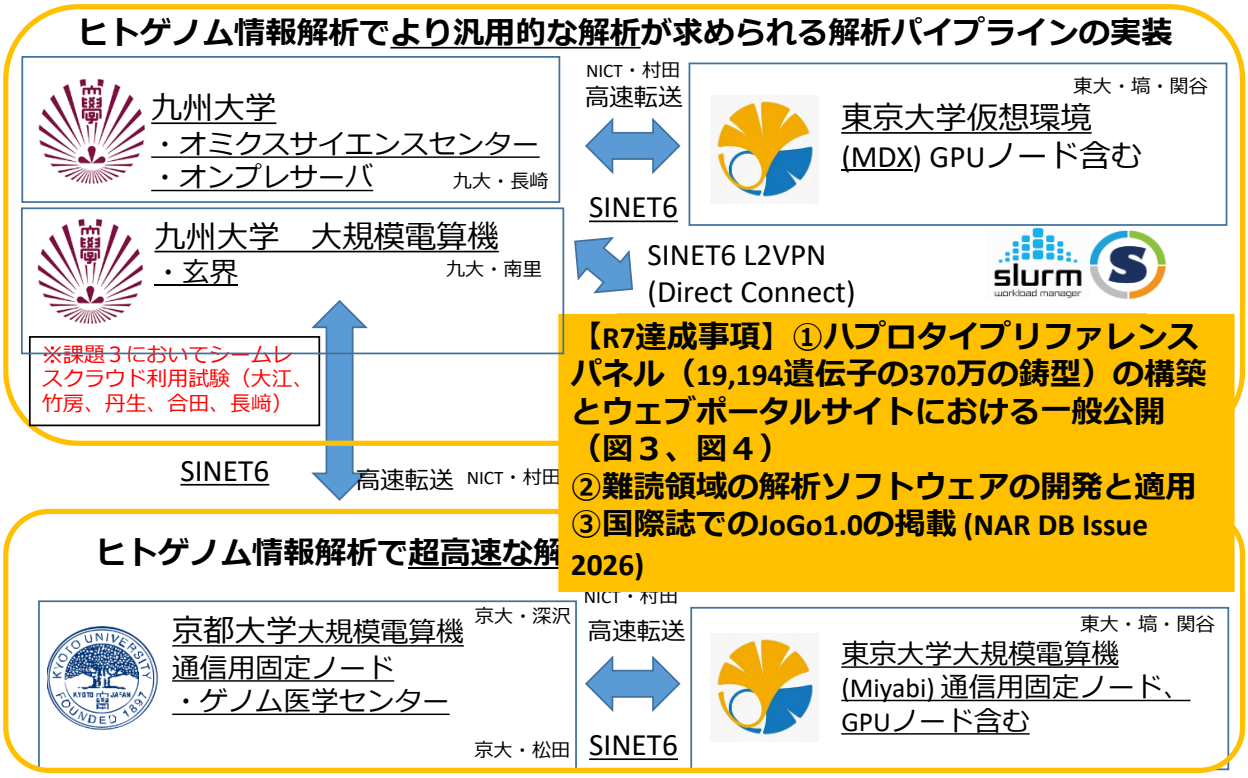
課題１）中高深度長鎖シークエンス情報に基づくハプロタイプリファレンスパネルの構築とそのための複数拠点間のハイブリッドクラウド情報基盤の運用（長﨑、関谷、塙、深沢、大川、松田）（図 2 に概念図と成果概要を示す）

令和 5 年度に取得した約 100 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約 160 検体の超高深度シークエンス情報を用いることで、日本人集団に留まらない、世界集団から構成される 258 検体からなる国際パネルを構築することができた。

特に、シークエンス情報の積み増しを行うことで、新たに取得された長鎖型シークエンス情報（課題２に関連）に対する実験毎の新規の情報解析を必要とするステップ（ステップ１）、さらに、すでに取得済みの検体の情報解析結果と新規に取得された情報を合わせて解析するステップ（ステップ２）が情報解析のために必要となりその解析を継続的に進めた。

技術的には、昨年度に構築を行ったドラフト版において得られた課題（ホモポリマーにおいてエラー率が向上すること）について、アミノ酸のフレームシフトとして扱われて

図２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　　システム全体構成と役割担当



しまうパターンが観測され、複数のハプロタイプにおいてサポートされるか、また、その前後の塩基パターンを確認することで同エラーを補正できるかどうかを検討し、その検討に基づき改良した計算ステップを実装することで本年度の公開版を構築する目途が立った。

計算資源については、ステップ１については、特に、九州大学の玄界、京都大学の大規模電算機を並行して活用することで、19, 194遺伝子全数の領域の計算を完遂できた。

また、ステップ２については、統合解析であり、九州大学のオンプレミスのシステムに各拠点で計算した内容を転送すること（課題

図３　**JoGo 1.0ポータルサイト**

https://jogo.csml.org/

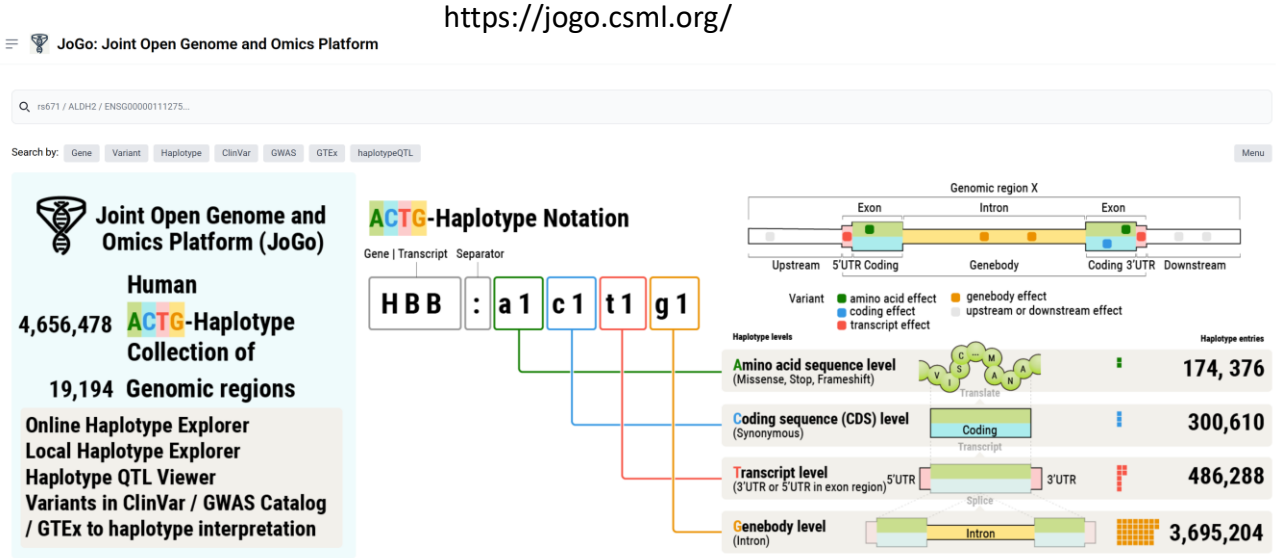
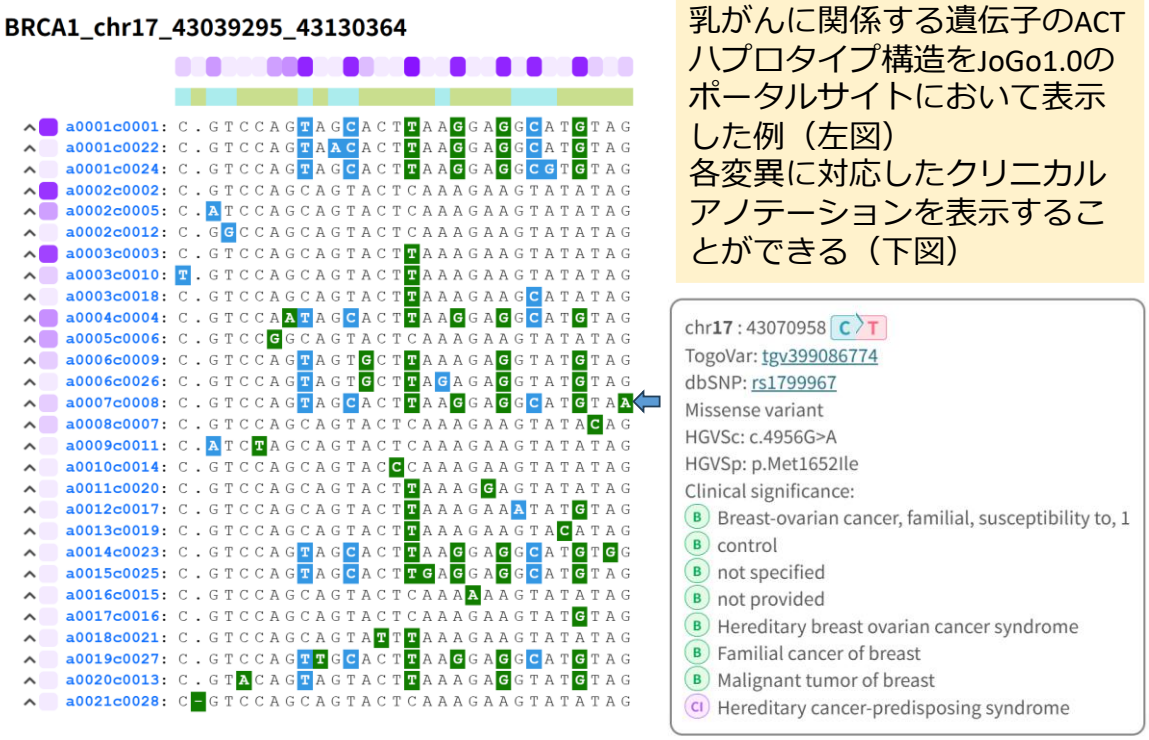


図４　**JoGo 1.0ポータルサイトのハプロタイプ情報の構築例 BRCA1**



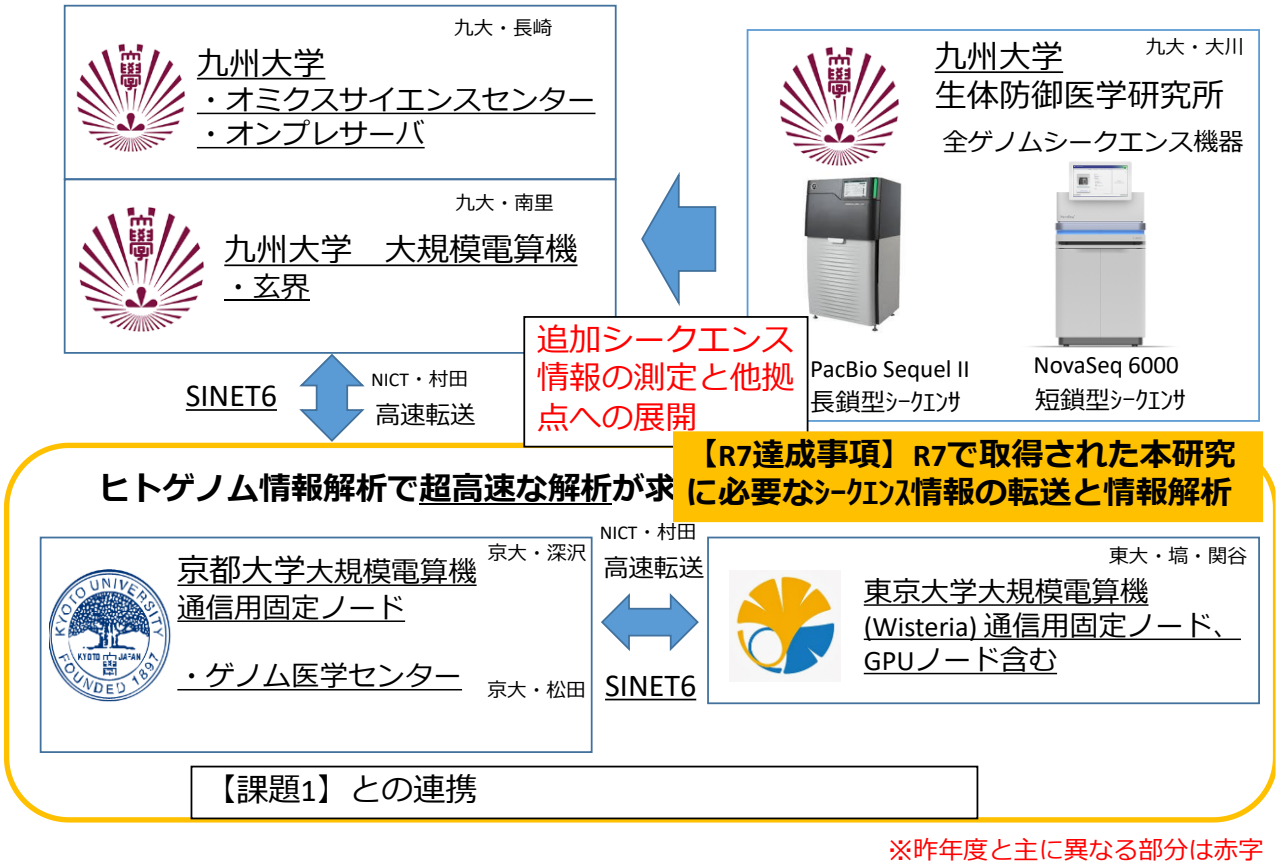
乳がんに関係する遺伝子のACTハプロタイプ構造をJoGo1.0のポータルサイトにおいて表示した例（左図）
各変異に対応したクリニカルアノテーションを表示することができる（下図）

３に関連）で効率良く計算処理を行った。

これらの取り組みにより、ACTG ハプロタイプ辞書の構築と、2025 年内に JoGo ウェブポータルサイト（https://jogo.csml.org/）からリリース版(1.0 版)の公開を達成できた（図３）。

同 ACTG ハプロタイプ辞書は、米国の NCBI とヨーロッパの EMBL-EBI が共同で進めたヒト遺伝子の標準遺伝子とその標準転写物を統一する規格によって定義された遺伝子のうち 19,194 のヒト遺伝子を対象として、ハプロタイプ情報を収集し、合計 174,376 の A レベル、300,610 の C レベル、486,288 の T レベル、3,695,204 の G レベルハプロタイプ（総計 4,656,478）が最終的に収載している（図３）。図４に乳がんに関連することが知られている BRCA1 の JoGo1.0 における、ハプロタイプ情報を示す。

図５　**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**



※昨年度と主に異なる部分は赤字

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５に概念図と成果概要を示す）

前半期間において、課題１に関連したシークエンス情報が九大（大川）により取得された。同シークエンサが接続している、九大のオンプレミス環境とハイブリッドクラウド間で同シークエンス情報を村田が開発を行っている Archaea（旧 HCP Tools）などを使って転送を行うことで、課題１の計算を円滑に推進した。

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６に概念図と成果概要を示す）

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。

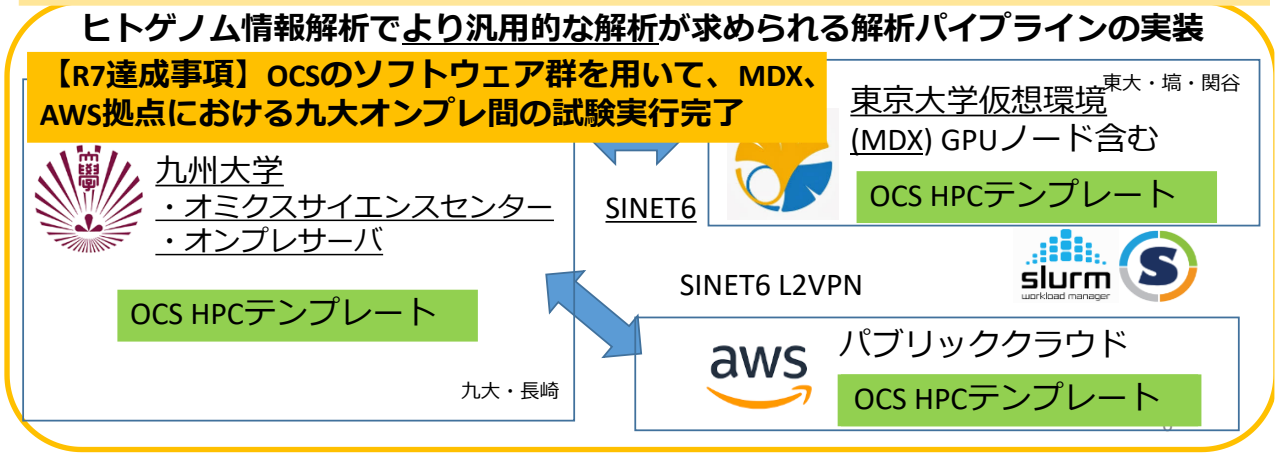
前半期間において、定期的（月１回、Slack を併用）、共同研究者間でミーティングを行うことで、9 月末に九州大学のオンプレシステム内に構築した学認クラウドオンデマンド構築のマスター環境経由から、パブリッククラウド AWS における疎通確認と簡単なジョブ実行を行うことに成功した。

図６　**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**

クラウド環境構築システムVCPによるmdxでのスケーラブルなHPCクラスタの構築
情報処理学会研究報告　大江、竹房、丹生、合田　Vol.2023-HPC-190 No.9

【報告概要】 OpenHPC 環境の構築が可能な OCS の HPC テンプレート v2 を用いて mdx でスケーラブルな HPC クラスタ構築機能を実現し,Slurm クラスタのジョブ実行状況に応じて計算ノードの増減を自動的に行うオートスケーリング機能を試験実装

課題１では各拠点では独立して稼働しているジョブシステムについてOCSのソフトウェア群を導入することで、ゲノムサイエンスの分野での多拠点でのシームレスなジョブ管理と運用の試験実装



後半期間においても、１カ月に１回定期ミーティングを行うとともに、Slack で OCS の技術サポートを継続的に受けることで、MDX での試験実行についても完遂することができた。

今年度の最終目標を達成することで、来年度以降、パブリッククラウド AWS と MDX に対して、九州大学のオンプレミス環境からシームレスにバッチジョブを投入するための環境構築の準備を整えることができた。

6.　進捗状況の自己評価と今後の展望

（自己評価）

課題１については、当初目標であるヒトの 19,194 遺伝子のハプロタイプ辞書の構築と一般公開を達成することができた。

本研究成果は、分子生物学分野で高い評価を受ける Nucleic Acids Research (NAR) の Database Issue に掲載され、特集号編集者から「世界初のヒト遺伝子を網羅的に収載するハプロタイプデータベース」である点を高く評価された。その結果、全掲載論文の上位 3%に付与される Breakthrough Article に選定され、本紙のウェブサイト及び論文 PDF に明記された。今後も情報拡充により研究コミュニティや医療基盤としての普及が期待される。

課題２についても、当初想定通り、本年度得られたシークエンス情報を課題１で必要となる解析を期間内に達成するために円滑に拠点間で展開、計算結果を回収することを目的としていた。担当メンバが一体となって、データ転送と各分散拠点での解析を村田の提供するソフトウェアを使うことなどで円滑に継続的に実施できた。

課題３についても、当初目標通り、複数拠点

での OCS 運用のために簡易試験に取り組み、前半期間においては、パブリッククラウドの AWS と、後半期間においては、MDX と九州大学のオンプレミスの間で疎通確認と試験実行を行うことができており、当初目標を達成できたと考えている。

（今後の展望）

課題１については、対象サンプルの拡充と並行した、遺伝子領域以外の非コード領域のハプロタイプ辞書の構築がバイオメディカル領域については必要となっている。

課題２については、第三世代シークエンサとして高出力の Oxford Nanopore 社のシークエンサーも九州大学で導入されたことから、既存の機器に加えて、新規機器の情報解析の円滑な運用が求められる。特にベースコールについて GPU デマンドであることからいままでの経験を活かした分散計算が求められる。

課題３については、各拠点間と九州大学のオンプレミスの間での試験実行に加えて、複数拠点を九州大学のオンプレミス経由で円滑に運用することでユーザビリティの向上を進めることが必要であると考えている。

※7. 研究業績はウェブ入力です