

jh200047-NWH

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する研究

長崎 正朗（京都大学学際融合教育研究推進センター）

概要

ヒト全ゲノム情報についてシーケ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、それらの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、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ことができるヒト全ゲノム情報解析のためのシステム環境を、具体的な課題を整理しつつ構築することを目的とし研究を進めた。2021 年度末において、ハイブリッドクラウドおよび京都大学ゲノム医学センターのオンプレミス、東京大学、京都大学大規模電算機資源などの複数拠点から構成されるハイブリッドクラウドを構築するとともに、各拠点において情報解析の役割を整理することで、5000 人規模のヒト全ゲノム情報解析を行った。これらの成果を踏まえつつ今後は 1 万規模を超える全ゲノム解析に向けた情報解析基盤の整理を進めていく。

1. 共同研究に関する情報

構築

橋本 洋希 ゲノムデータ解析環境基盤構築

(1) 共同研究を実施した拠点名

構築

東京大学 京都大学 九州大学

宮本 敏 ゲノムデータ解析環境基盤構築

関谷 弥生 ゲノムデータ解析支援

(2) 共同研究分野共同研究分野

男澤 良子 ゲノムデータ解析支援

超大容量ネットワーク技術分野

Wang Yen Yen (王 妍雁) ゲノムデータ

解析支援

(3) 参加研究者の役割分担

京都大学・学術情報メディアセンター

京都大学・学際融合教育研究推進センタ

ーおよび医学系研究科

深沢 圭一郎 各大学間連携・データ伝送

(京都大学拠点)

長崎 正朗（代表者）ハイブリッドクラウド

構築統括

東京大学・東京大学情報基盤センター

山田 亮（副代表者）ゲノムデータ解析

助言

関谷 勇司 クラウド運用および設計アド

松田 文彦 ゲノムデータ解析助言

バイス

Gervais Olivier ゲノムデータ解析助言

塙 敏博 各大学間連携・データ伝送（東

山口 泉 ゲノムデータ解析環境構築設計

京大学拠点)

川口 喬久 ゲノムデータ解析環境構築設計

稲富 雄一 ゲノムデータ解析環境構築設計

九州大学・生体防御医学研究所

浅倉 章宏 ゲノムデータ解析環境基盤

大川 恭行 シークエンス情報関係アドバイス

前原 一満 データ転送（九州大学拠点）

九州大学・情報基盤研究開発センター

南里 豪志 大学間連携（九州大学拠点）

情報通信研究機構・総合テストベッド研究開発推進センター

村田 健史 拠点間データ転送ソフトウェア提供・設定アドバイス

の情報について試験的に解析を行うことを目的とし研究を進めた。

本計画で扱う情報は、疾患をもつゲノム情報を含んでおり、今回の設計で扱う情報からえられた解析結果は、疾患の原因やマーカーを含むことが十分に考えられ学術的・社会的な意義がある。さらに、数年後の全ゲノム情報の大規模解析のための設計を進め課題を洗い出すことについて、学術的な意義があると考え。特に、データ転送や複数拠点での効率的な解析は重要であると考え。

そのため、「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する研究」の課題を設定し以下の課題を解決することを目的とする。

2. 研究の目的と意義

ヒトゲノム情報についてシーケ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、それらの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。

1つの拠点では、上の目的を達成することが困難な状況となっており、オンプレミス、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。

一方、京都大学のゲノム医学センターにおいて5,000 検体の全ゲノムシーケンスが行われており、それらの情報のバックアップ（1 検体当たり100GB～200GB）、ヒトゲノムリファレンス配列の更新に伴う再解析（1 検体当たり、32・48Core 搭載 CPU で1 日から1 週間）、また、下流解析（計算時間は解析の内容によって異なる）が必要となっている。他に、最新のシーケンサによって取得された情報の転送のための実装が求められている。

そこで、当研究においては、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を整理しつつ、円滑に上の一部

課題1) 複数拠点間を効率的に運用できるハイブリッドクラウドシステムの設計と運用

課題2) シーケンサから取得された情報を他の拠点に効率良く展開するための設計検討

3. 当拠点公募型研究として実施した意義

現在、5,000 検体の規模の解析であるが、今後は海外の事例にあるように数万以上の規模が想定されている。そのような規模において提案されてくるバイオインフォマティクス手法などによる解析が必須である。また、それらの解析によって、シーケンスされた生データから新たなリスク要因が同定されることが想定されている。

そのため、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ができるか、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ことが必要である。

4. 前年度までに得られた研究成果の概要

初年度のため該当なし

5. 今年度の研究成果の詳細

課題 1) 複数拠点間を効率的に運用できるハイブリッドクラウドシステムの設計と運用 (長崎 正朗、山田 亮、松田 文彦、Gervais Olivier、山口 泉、川口 喬久、稲富 雄一、浅倉 章宏、橋本 洋希、宮本 敏、関谷 弥生、男澤 良子、Wang Yen Yen、深沢 圭一郎、関谷 勇司、埴 敏博)

概要は図 1 参照

前半期間は、5,000 検体すべての情報を本研究課題の解析資源で賄うことが困難であることから、一部のデータ (数百検体程度) の情報での設計と運用を進めた。具体的には、京都大学のオンプレから、京都大学のメディアセンターのサーバにデータの展開を行った後、これらの情報を、東京大学の電算システムに展開し情報解析を行うことを進めた。また、それらの結果について、京都大学のメディアセンターのサーバに転送するまでのフローの構築を進めた。各電算機資源において共通した利用ができるように、singularity v3 のインストールを京都大学 Cray XC40 のシステムに依頼し、2020 年 9 月末にインストールが完了し準備を整えた。

さらに、SINET5 の L2VPN をもちいて、パブリッククラウドにも接続を行い同定された変異情報のアノテーションなどの下流解析を行える試験実装についても進めた (図 2、図 3 参照)。

また、上の全体の実装や解析フローの内容について、2020 年 9 月開催の AWS

Summit Online Japan 2020 で発表を行った。

さらに、現在までに検討を進めた大規模ゲノム解析における現実的な実装の経験を活かした内容に関連する国際論文の執筆を行い 2020 年 10 月に Journal of Human genetics 誌で掲載された。掲載後半年で 5,000 以上のオンラインアクセスがなされている (2021 年 5 月現在)。

2020 年度後半において、東京大学の電算機システムにおいてデータ転送用に固定ノードの提供を受け、2021 年度において拠点間のデータ転送を行える設定の方法などの準備を進めた。

なお、当該研究課題時に配分された電算機資源のみで解析が可能な数百検体を目標として課題の整理、試験実装を進めたが、上記検討を踏まえ、2020 年度途中から、京都大学ゲノム医学センターのオンプレミスに加え、同センターで別途契約をおこなった京都大学および東京大学の大規模電算機資源を活用することで希少疾患を含む 5,000 検体の全ゲノム情報解析を 2020 年度内に達成している。

課題 2) シークエンサから取得された情報を他の拠点に効率良く展開するための設計検討 (大川 恭行、前原 一満、南里 豪志、深沢 圭一郎、村田 健史)

概要は図 4 参照

シークエンサから得られる情報は、特に希少な疾患については、これらの検体については、DNA サンプルが少ないことも多々あり、得られた情報自身が貴重であることも多い。

また、すべての大学において最先端のシークエンサが導入されていることはまれである。そこで、本研究課題においては、最

新のシークエンサが導入されている九州大学と主に京都大学の間でシークエンサによって得られた情報を効率よく転送するための設計や性能試験のための準備を進めた。

6. 今年度の進捗状況と今後の展望

課題 1) 複数拠点間を効率的に運用できるハイブリッドクラウドシステムの設計と運用

図 2、図 3 で示すように各拠点の構成及び役割を整理するとともに 5000 人規模の解析を進めた。2021 年度においては、現在の配列情報も含め合計 1 万人規模の新たなシーケンス配列が取得されることから、1 万人程度に解析対象人数が増えた場合のハイブリッドクラウド上の運用の解析フローについて解析を進める。

2020 年度において、ヒト全ゲノムを複数拠点間の各計算リソース、ストレージを用いて解析を完了することを実際に試験的に行ったが、構成がヘテロな環境であることからいくつかの課題が得られた。特に、計算が一部指定時間内に完了しない (Oakbridge-CX は 48 時間が最長) ためにそのサイトでの実行は断念し他のサイトで実行を行うこととした (京都大学の Cray XC40 はジョブの実行時間の制限が 2 週間となっている)。

並行して、当課題に対応するため、制限時間内で解析された染色体上の領域から再実行できるようにするなどのレジューム機能を試験的に実装することで解析フローの一

部について解決を試みた。2021 年度は、同試験結果を取り込むことでより自動的に自動レジュームが実行できるように実装することや、対象の染色体上の領域をさらに分割することで対象時間内に完了できるようにするなどの改良を進める。

また、2021 年度に向けてオンプレミスに FPGA によって高速に定型解析を行うことができる DRAGEN を 2021 年 3 月に試験的に導入したこと。これらの機能についてもハイブリッドクラウド上で合わせて効率よく利用するための方策について検討を進めていく。

拠点間の転送について 100TB のデータ転送が定期的に発生することが考えられることから、これらの転送をより効率的にできるように高速転送のツールの評価と拠点間実装を優先して行う。

上記を総合的に進めることで、2020 年度に出た課題の確認、また、改善を進め 2021 年度は、1 万人規模の解析を達成していく。

課題 2) シークエンサから取得された情報を他の拠点に効率良く展開するための設計検討

2021 年 1 月から、開発者の村田から提供される HCPTtools を用いた課題 1 におけるデータ転送準備を進めてきた。同成果を活用して 2021 年後半には得られるシーケンス情報の転送を課題 2 においても進めていく。

課題 1) 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する研究 (長崎、山田、松田、関谷、深沢、村田、塙) システム全体構成と役割担当

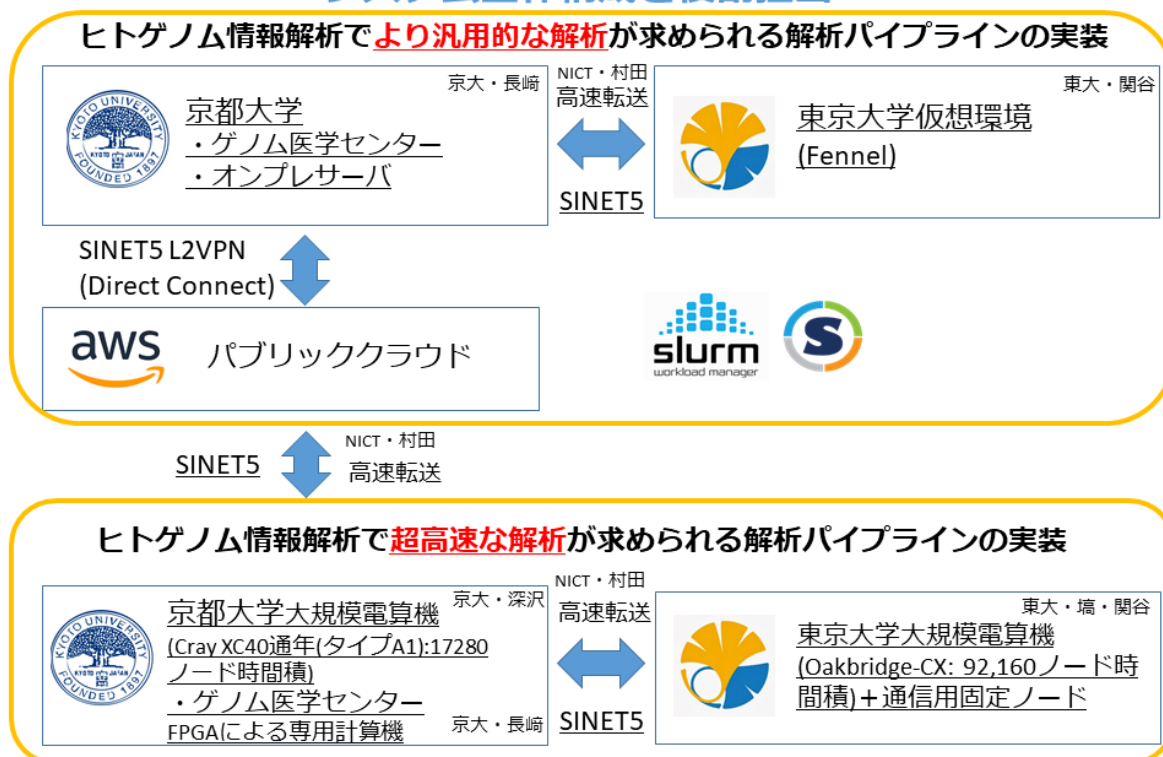


図 1 課題 1 の概要とチーム構成

ハイブリッドクラウドシステムに向けたインフラの設計と実装 ーシステム構成概要ー

成果1

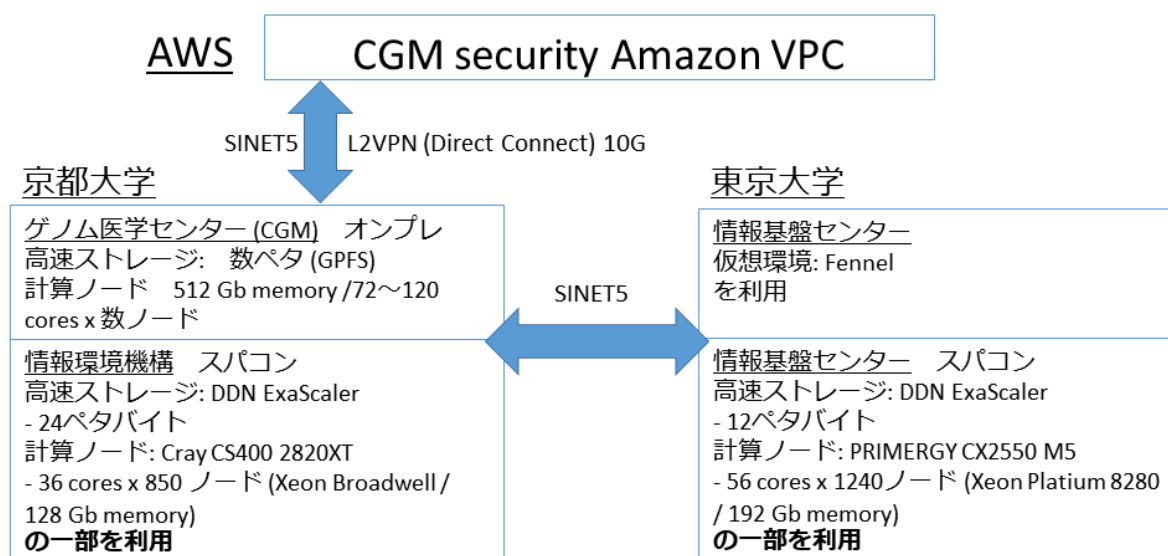


図 2 成果 1 システム構成概要

ハイブリッドクラウドシステムに向けたインフラの設計と実装 —解析パイプラインの対象技術の適用—

成果2

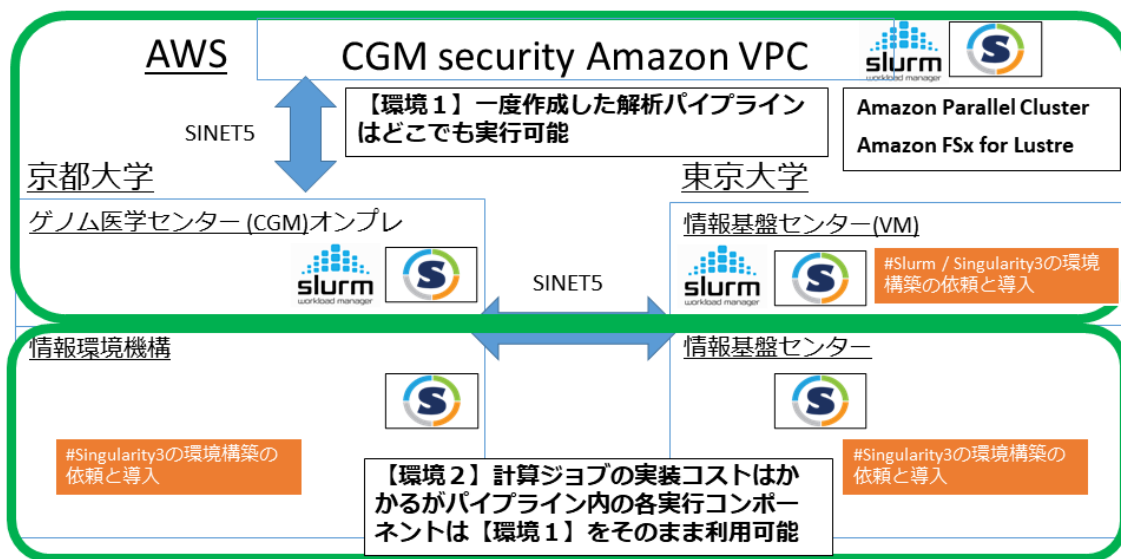


図3 成果2 各解析拠点のゲノム情報解析における役割

課題2) シークエンサから取得された情報および解析パイプラインの各ステップの入出力データの拠点間通信の最適化検討 (大川、前原、長崎、深沢、村田、塙、関谷)

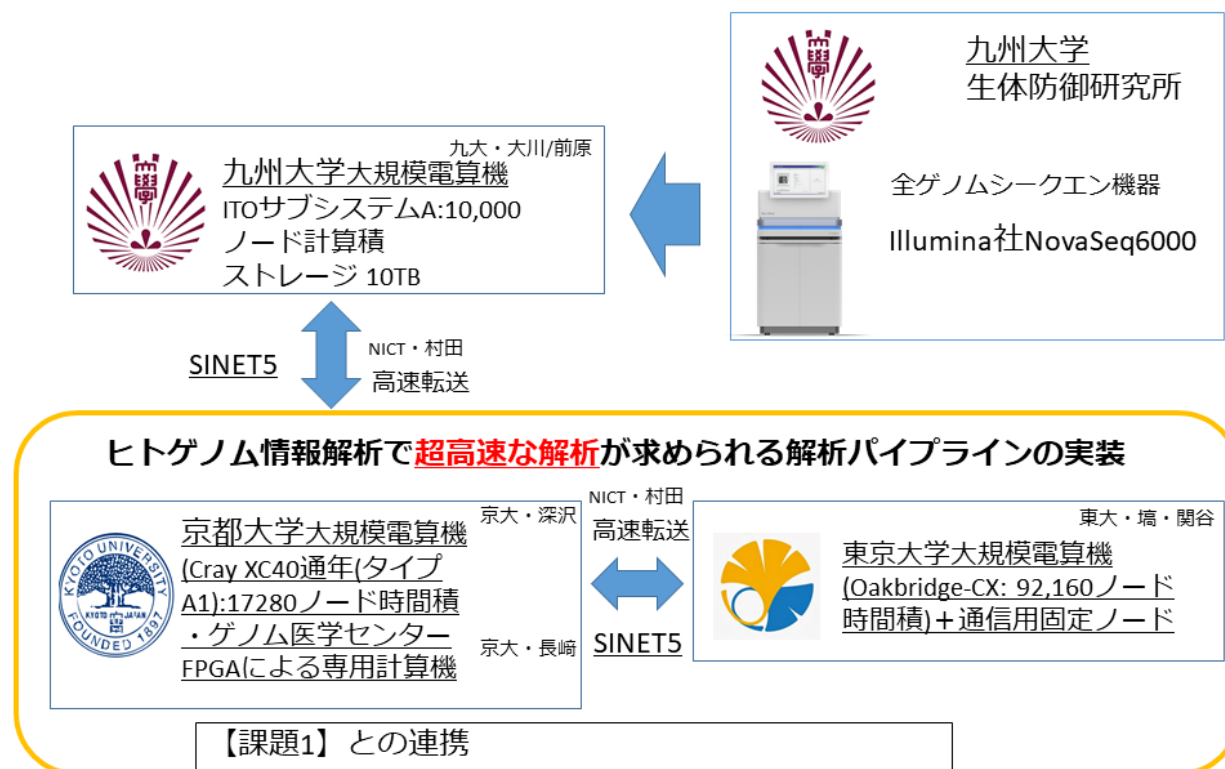


図4 課題2の概要とチーム構成

7. 研究業績一覧（発表予定も含む。投稿中・投稿予定は含まない）

(1) 学術論文（査読あり）

T. Tanjo, Y. Kawai, K. Tokunaga, O. Ogasawara, M. Nagasaki, 'Practical guide for managing large-scale human genome data in research', Journal of Human Genetics (66), pp. 39 – 52, 2021

(2) 国際会議プロシーディングス（査読あり）

該当なし

(3) 国際会議発表（査読なし）

該当なし

(4) 国内会議発表（査読なし）

長崎 正朗, 'ヒトゲノム情報統合解析に向けた京都大学ゲノム医学センターのハイブリッドクラウドシステム構築について', AWS Summit Online Japan 2020 (2020/9/8 - 2020/9/30)

(5) 公開したライブラリなど
該当なし

(6) その他（特許, プレスリリース, 著書等）

長崎 正朗, AWS SUMMIT ONLINE JAPAN Report
https://special.nikkeibp.co.jp/atcl/NXT/20/aws1030_01