jh210018-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した研究

長﨑 正朗（京都大学学際融合教育研究推進センター）

概要

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、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、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過程で出てくる課題を整理しつつ、円滑に上の一部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。

昨年度は、有償で追加利用を行った東京大学および京都大学の計算資源も含め、約 5,000 検体の全ゲノムリファレンスパネルの構築を進めた。その中で、各拠点での計算機資源の特徴を考慮し、解析パイプラインの各ステップを試行錯誤しつつ実行を進めた。本年度は、そこで得られた知見に基づき、有償で追加資源を確保することを並行することで、2021 年度に当初目的としてた総計約 10,000 検体の全ゲノムリファレンスパネルの構築を達成した。

1. 共同研究に関する情報

(1) 共同研究を実施した拠点名

東京大学　京都大学　九州大学

(2) 共同研究分野

超大容量ネットワーク技術分野

(3) 参加研究者の役割分担

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

長﨑 正朗（代表者）ハイブリッドクラウド構築統括
山田 亮（副代表者） ゲノムデータ解析助言
松田 文彦　ゲノムデータ解析助言
Gervais Olivier　ゲノムデータ解析助言
山口 泉　ゲノムデータ解析環境構築設計
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
浅倉 章宏　　ゲノムデータ解析環境基盤構築
橋本 洋希　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
男澤 良子　ゲノムデータ解析支援
Wang Yen Yen（王 妍雁）　ゲノムデータ解析支援
寺岡 凌　ゲノムデータ解析支援

京都大学・学術情報メディアセンター

深沢 圭一郎 各大学間連携・データ伝送（京都大学拠点）

東京大学・東京大学情報基盤センター

関谷 勇司　クラウド運用および設計アドバイス
塙 敏博　各大学間連携・データ伝送（東京大学拠点）

九州大学・生体防御医学研究所

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

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、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、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は、2020 年度には約 5,000 検体、2021 年度には１万検体以上の全ゲノムシークエンスが集積されており、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、最新のシークエンサによって取得された情報の転送、データシェアリング等によって利用可能になった国内外のシークエンス配列との統合解析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、円滑に上の一部の情報について試験的に解析を行うことを目的として研究を進めた。

2020 年度は、有償で追加利用を行った東京大学および京都大学の計算資源も含め、約 5,000 検体の全ゲノムリファレンスパネルの構築を進めた。その中で、各拠点での計算機資源の特徴を考慮し、解析パイプラインの各ステップを試行錯誤しつつ実行を進めることで知見を得てきた。

そこで、本年度は、2020 年度に得られた知見に基づき、有償で想定される追加資源を確保することを並行しつつ、2021 年度に利用可能となる合計約 10,000 検体に対する全ゲノムリファレンスパネルの構築を目標に進めた。

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得られた解析結果は、疾患の原因やマーカーを含むことが十分に考えられ意義があると考えている。また、数年後の全ゲノム情報の大規模解析のための設計を視野に入れて課題を洗い出すことは意義があると考える。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。

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に運用に関連した研究」の課題を設定具体的に以下の２つの副課題を設定し解決することを目的とする。

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

3. 当拠点公募型研究として実施した意義

申請時には、当センターには約 5,000 検体の全ゲノム情報が集積されており、今後は海外の事例にあるように数万以上の規模での集積が想定されていた。

実際、当初想定していた通り 2021 年度半ばで 10,000 検体以上の全ゲノム情報が当センターに集積される状況となった。規模感が大きくなるにつれ、特に全体を統合する解析において各バッチ単位の計算時間がかかること、メモリサイズの設定の変更、拠点間の転送の時間の増加などさまざまな課題がでてきた。

これらの各課題については、昨年度に得られていた知見に基づき、ハイブリッドクラウドとして複数拠点を安全に接続しつつ、拠点毎の特徴を活かした解析を柔軟にできる仕組みを活用し、複数の解析ステップから構成される１万人の全ゲノム解析を期間内に達成することができた。

当該研究期間終了後も、さらに継続して人数規模が拡大し、数万以上の規模となることが想定されており、今回は１万検体であったがさらに情報量が拡大した場合に向けた解析環境の原案を作成できた点についても大変意義がある。

4. 前年度までに得られた研究成果の概要

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 5,000 人規模の解析を進めた。また、この成果と得られた課題を元に 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クエ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こととした。

特に、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半自動的にレジューム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こととした。

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から、ハイブリッドクラウド上であ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こととした。

拠点間の転送について 100TiB 以上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こととした。

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こととした。

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

2021 年１月から、開発者の村田（分担研究者の一人）から提供される HCPTools 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 年後半に得られるシークエンス情報の転送を課題２においても並行して進めていくこととした。

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（長﨑 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、関谷 弥生、男澤 良子、Wang Yen Yen、寺岡 凌、深沢 圭一郎、関谷 勇司、塙 敏博）

概要は図１参照、成果は図２, ３参照

2022 年度は昨年度の 5,000 検体の解析実績に基づき、10,000 検体の解析を進めた。具体的には、個別の検体の解析については、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、専用ゲートウェイサーバを通じて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とオンプレに情報を書き戻すことで計算を進めた。

また、昨年度 singularity v3 を整備したことから、各電算機資源のバッチジョブシステムの違いのみを更新することで同一のパイプラインが実行できるように改良をして解析を進めた。なお、2020 年度後半に、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ることができたため、2021 年度は、拠点間のデータ転送を専用ゲートウェイ経由で行える体制としている。また、当初計画していた DRAGEN についてもオンプレ内で利用ができるように整備を進めた。

さらに、昨年度に試験的に構築をしていた SINET5 の L2VPN 経由でオンプレサーバ区画のみから商用クラウドに接続できるシステム構築を完了したことから、京大、東大拠点に加え、統合解析の一部を商用クラウド上においても同一の Singularity v3 とバッチジョブによるパイプラインを用いて並行して実施することで 2022 年 3 月末までに統合解析を完了できた（図２、３参照）。

2021 年度、5,000 検体の段階において全体解析の約 4,000 並列実行のジョブの３割において、計算が一部指定時間内に完了しなかったが、2022 年度は、特に１万人規模以上の検体となったことから実行上限の課題は特に顕著となり７割以上が時間内に１度のジョブでは計算完了しなかったが、昨年度に試験実装していたレジューム機能を改良することで運用を行った。最大、同システム上で 5 回レジュームを実行することで対象とした領域の計算を完了することができた。

上記ハイブリッドクラウドの設計や解析フローの内容の一部について、2021 年 10 月開催の日本人類遺伝学会年会、2021 年 11 月の医療情報学連合大会年会、2021 年 12 月の日本分子生物学会年会において発表を行い広く周知をおこなった。

なお、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、商用クラウドの計算資源を有償で契約・活用することで当センターに集積している 10,000 検体の全ゲノム情報解析を 2021 年度内に達成したことを補足する。

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（大川恭行、前原 一満、南里 豪志、深沢 圭一郎、村田 健史）

概要及び成果は図４参照

シークエンサから得られる情報、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を進めた。

特に本年度は、高速転送ソフトウェア HCPTools を重点的に拠点間通信で活用し rsync との特徴の差異、ファイル数が大量であるときに転送効率が良い点などを考慮し研究分担者の村田と連携しつつ評価を進めた。本年度は、138 万ファイル総量 487TiB を京大と東大間の拠点で転送を行った（別途 rsync で 7.7TiB を転送）

特に統合解析において、約 4,000 並列ジョブの１つ１つの解析ジョブが約１万個以上の細かく分割されたファイルを転送する必要がありこれらのファイルの効率的な転送が必要となる。特に HCPTools はこのように細かく分割されたファイルについて性能がでることが開発者の村田の試験結果から得られており、同ソフトウェアの恩恵を受けることができた。

さらに、海外（カナダ）との HCPTools を用いた転送を 3,781 ファイル、総量 30.1TiB の転送を２回に分けて 524 時間かけて実施し実際に海外との連携においても運用できることを示した。

5. 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

当初予定していた、10,000 検体の解析についていくつかの課題は出たものの、昨年度に経験していた課題点とその対応策を検討していたこと、例えば、統合解析において一時的に大量の並列計算が必要になることから複数拠点から構成されるハイブリッドクラウドの構成としていたことから、期間内に目標を達成することができた。なお、オンプレのみ、京大拠点のみ、東大拠点のみ等、各々、１拠点のみでの構成では期間内の統合解析の計算は困難であったと考えられる。

特にシークエンス情報については日々算出されていることから、個々の独立した検体の解析については、オンタイムで可能であるが、後段の全体の検体の情報を必要とする統合解析においては、一部の検体のもれがあるとすべて再解析をし直しになるために、検体セットの確定（データフリーズ）が遅延しがちである。実際、本研究課題においても、個別の要望がでたために、２月半ばまで完了をすることができなかった。

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

当初予定していた rsync 以外の高速転送ソフトウェアの評価を行うという目的を達成することができた。

特に、2022 年度は、開発者の村田から提供された HCPTools を用いて拠点間の転送を行

うことでrsyncよりも大量のファイルがある場合に高速に転送をすることができたことから、より効率的に拠点間の運用を行うことができた。今後、京都大学の基幹システムでHCPToolsが導入され、本格的に利用できるようになる予定であることから、得られた知見を提供するとともに、今後、当センターにおいてもより効率的な運用ができると考えている。

6. 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）

⑴ 学術論文 （査読あり）

[1] 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, 2021.

⑵ 国際会議プロシーディングス （査読あり）

該当なし

⑶ 国際会議発表 （査読なし）

[2] Accelerating the pace of research in Kyoto University, 長﨑 正朗, AWS Public Sector Summit Online, 2021/4/15-16.

⑷ 国内会議発表 （査読なし）

[3]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, 長﨑 正朗, AWS Summit Online Japan 2020 (2020/9/8-2020/9/30).

[4]ゲノム医科学における国内外のヒトゲノム解析の状況およびハイブリッドクラウド計算環境の構築と活用, 長﨑 正朗, 山口 泉, 川口 喬久, 寺岡 凌, 稲富 雄一, 深沢 圭一郎, 関谷 勇司, 塙 敏博, 大川 恭行, 王 妍雁, Gervais Olivier, Khor Seik-Soon, 植野 和子, 浅倉 章宏, 関谷 弥生, 人見 祐基, 小野 彰, 男澤 良子, 河合 洋介, 前原 一満, 南里 豪志, 村田 健史, 橋本 洋希, 丹生 智也, 小笠原 理, 山田 亮, 松田 文彦, 徳永 勝士, 第 44 回日本分子生物学会年会, 2021/12/1-3.

[5] クラウドサーバとオンプレミスサーバを組み合わせたハイブリッドシステムの構築と活用, 山口 泉, 川口 喬久, 寺岡 凌, 稲富 雄一, 深沢 圭一郎, 関谷 勇司, 塙 敏博, 大川 恭行, 王 妍雁, Olivier Gervais, Seik-Soon Khor, 植野 和子, 浅倉 章宏, 関谷 弥生, 人見 祐基, 小野 彰, 男澤 良子, 河合 洋介, 前原 一満, 南里 豪志, 村田 健史, 橋本 洋希, 徳永 勝士, 松田 文彦, 山田 亮, 長﨑 正朗, 日本人類遺伝学会第 66 回大会, 2021/10/13-16.

[6] クラウドサーバとオンプレミスサーバを組み合わせたハイブリッドシステムの構築と活用, 山口 泉, 川口 喬久, 寺岡 凌, 稲富 雄一, 深沢 圭一郎, 関谷 勇司, 塙 敏博, 大川 恭行, 前原 一満, 南里 豪志, 村田 健史, 橋本 洋希, 松田 文彦, 山田 亮, 長﨑 正朗, 第41回医療情報学連合大会, 2021/11/20.

⑸ 公開したライブラリなど

該当なし

⑹ その他（特許，プレスリリース，著書等）

長﨑 正朗, AWS SUMMIT ONLINE JAPAN Report https://special.nikkeibp.co.jp/atcl/NXT/20/aws1030_01/ ※記事内で共同研究者、関谷勇司の記載含む、また、本研究課題の謝辞含む

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 成果1

システム全体構成と役割担当

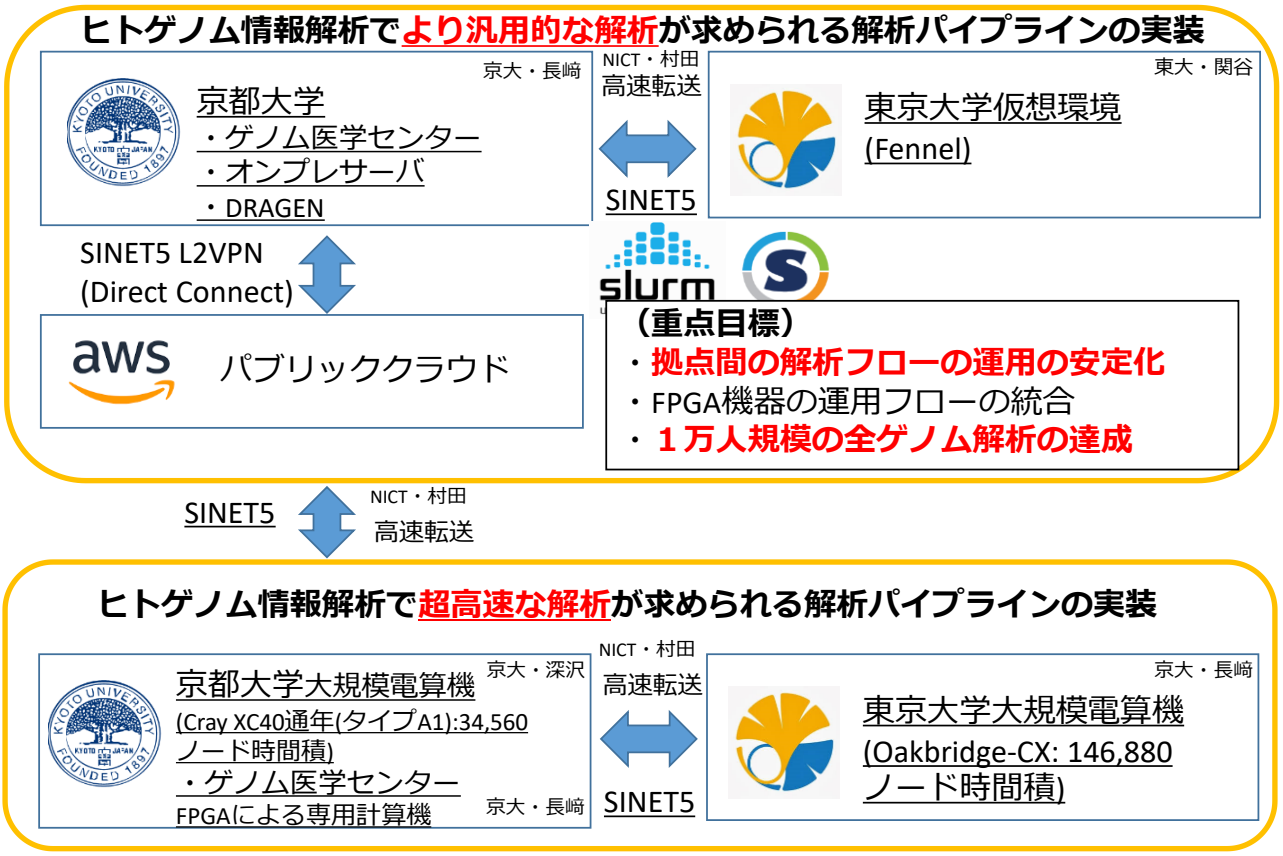

図１　課題１の概要とチーム構成と得られた成果１（特に赤字部分）

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2

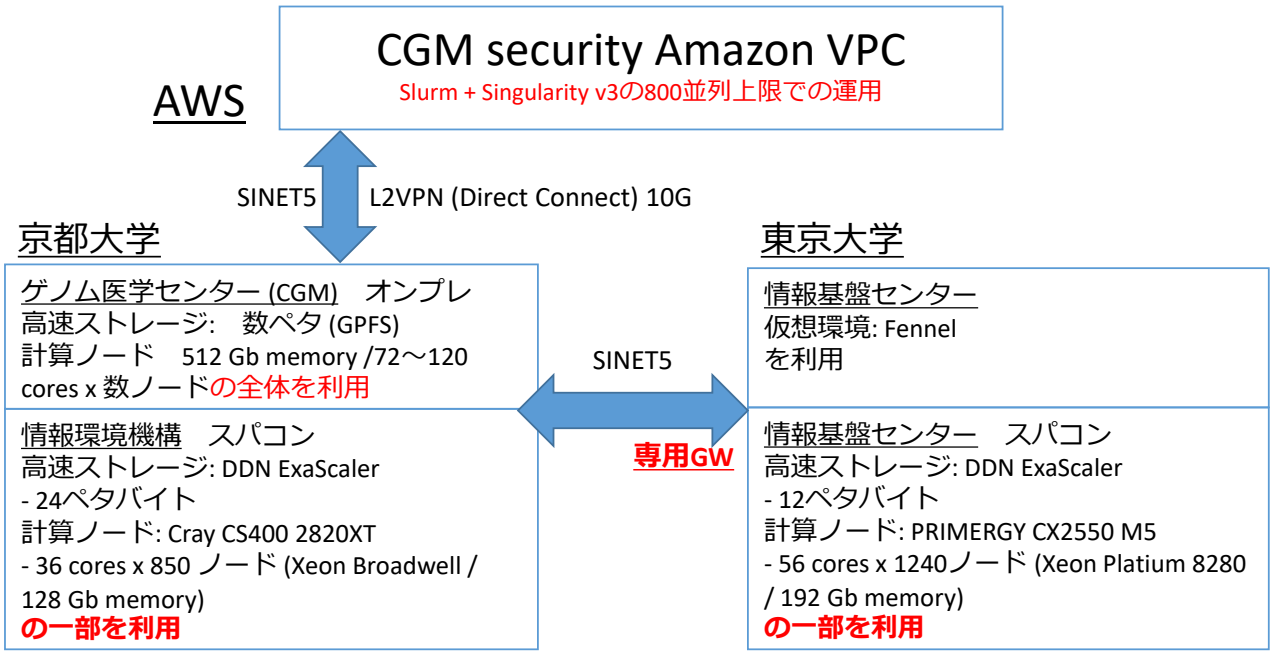

図２　システム構成概要と得られた成果２（特に赤字部分）

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
ー解析パイプラインの対象技術の適用ー

成果3

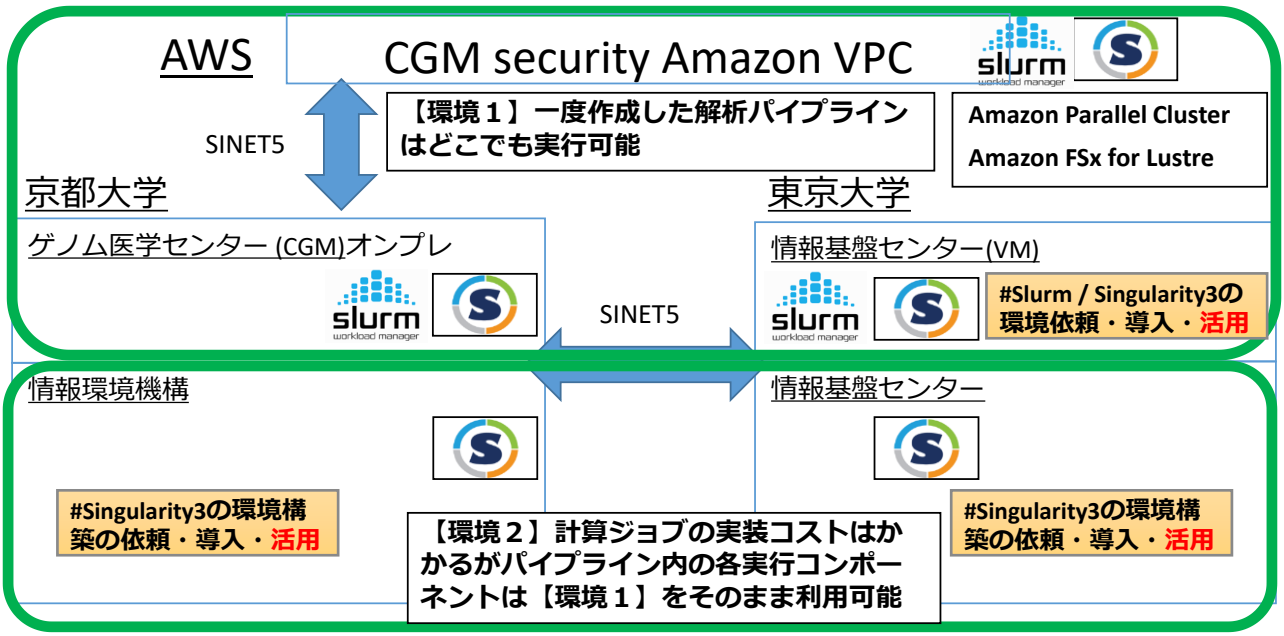

図３　各解析拠点のゲノム情報解析のおける役割と得られた成果３（特に赤字部分）

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

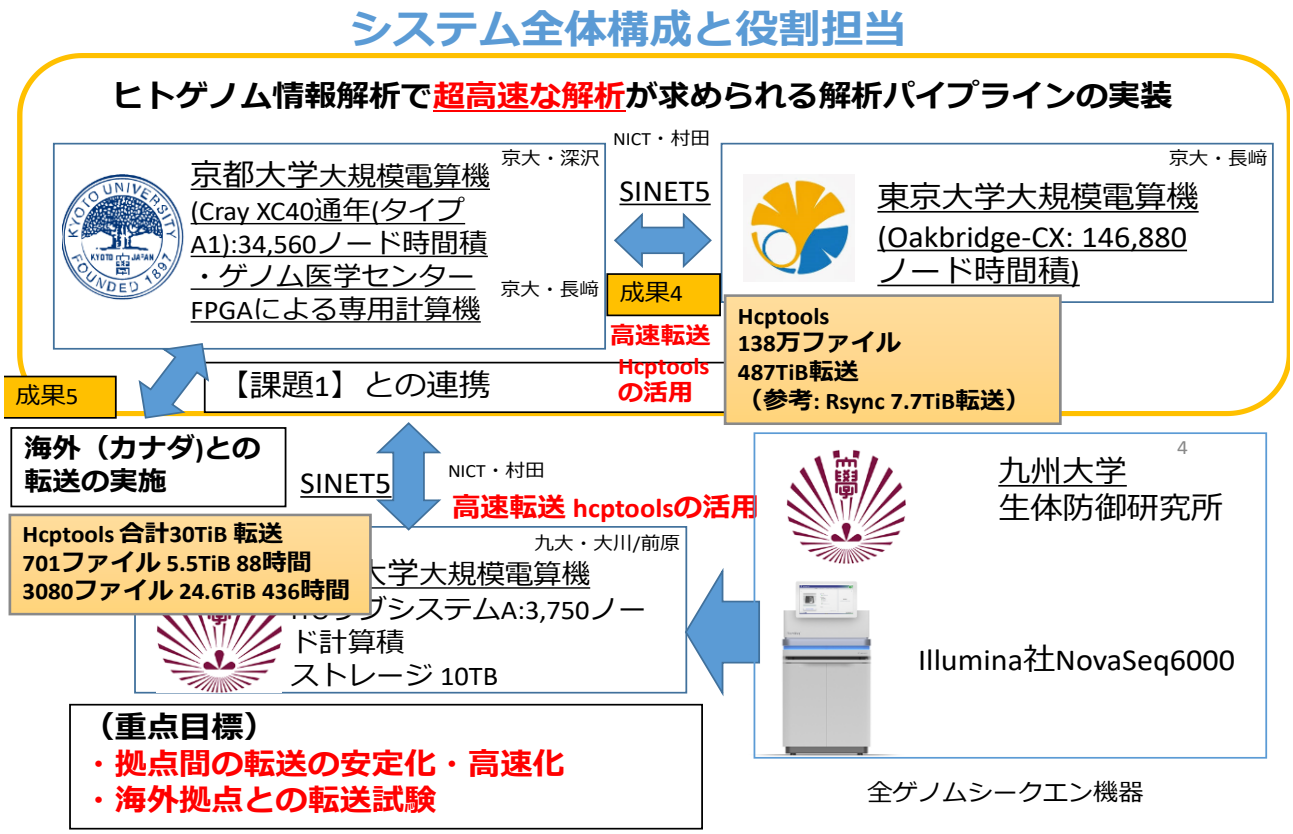

図４　課題２の概要、チームメンバ、および得られた成果４と成果５（特に赤字部分）