jh220014

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

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、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「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)」において進めてきた。当研究課題では、さらに、申請者が進めている日本人の長鎖型情報を活用し、令和４年度において、先行研究が進めた 5,202 検体と同程度の約 5,000 検体の短鎖型法の検体を長鎖型の情報を鋳型に解析を行うことで 100 遺伝子の構造多型のカタログ構築を進めた。

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（該当するものを残す）

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（該当するものを残す）

データ科学・データ利活用課題分野

(3) 共同研究分野（HPCI 資源を利用している研究課題のみ, 該当するものを残す）

超大容量ネットワーク技術分野

(4) 参加研究者の役割分担

京都大学の長﨑、松田のチーム（他、関谷弥生、男澤、山口、川口、稲富、Olivier Gervais、王、寺岡）は、研究課題１のうち、特に、オンプレ、各電算資源間の効率的な解析パイプラインの構築について検討と実装を進めた。東京大学の関谷、塙は、クラウド実装におけるアドバイス、また、試験環境の整備を進めた。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試用した。また、京都大学の計算機資源への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進めている。他に、深沢は、京都大学の SINET6 を用いたパブリッククラウドへの VPN 接続管理においても支援を行った（他、浅倉、橋本）。研究課題２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、また必要に応じて、解析結果の返却（京大側は、長﨑、浅倉、橋本が担当）を主に行い評価を進めた。

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、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、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を中心として約 10,000 検体の短鎖型法(1 つの DNA 断片の読み取り長が 300-400 塩基)による全ゲノムシークエンスを行い、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、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「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)」において進めてきた。

一方、近年、長鎖型法(1 つの DNA 断片の読み取り長が 10,000 塩基以上)により全ゲノムデータの取得が進められ始めている。長鎖型は短鎖型に比べ構造多型の解析の精度が高いことが知られている（図１)。さらに、同情報によって得られた配列情報を鋳型とすることで、短鎖型法で得られたシークエンス情報を再解析することで海外においてヒトゲノムに含まれている遺伝的な形質に関連する構造多型が特定されてきている(Nature Comm 12(4250) 2021, Nature 374(1461) 2021)。そこで、当研究課題では、申請者が進めている日本人の長鎖型情報を鋳型として活用して、令和４年度において、先行研究が進めた 5,202 検体と同程度の約 5,000 検体の短鎖型法の検体を中心に解析を行うことで 100 か所の遺伝子に対する構造多型カタログの構築を行う。これにより将来的に疾患リスクに関連する遺伝子の詳細なハプロタイプ解析をおこなうことができる基盤構築技術の１つとして確立を進めていく。また、バイオインフォマティクス解析においては大規模なメモリを必要とする解析環境が解析ソフトウェアによって大きく異なる。そこで、いままで開発を進めてきたハイブリッドクラウドを改良を進めることで運用するとともに、本解析においては一部、大規模仮想環境を用いることで柔軟に情報解析をすすめることを試みる。

そのために、「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」の課題設定として、以下の２副課題を具体的に設定し研究を進める（図２)。

課題１）100 遺伝子の構造多型のリファレンスパネルの構築とそのための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情報基盤の設計と運用

課題２）長鎖型シークエンサから取得されたアノテーションに活用する情報を他の拠点に効率良く展開するための設計検討と実装

3. 当拠点の公募型研究として実施した意義

本研究提案の解析においては大規模な計算資源、また、効率的な各拠点での解析が必要となる。約 100 検体の長鎖型ゲノムに基づくグラフゲノムの構築において最低 256G 程度の電算機資源、また、過去の実績から個別の 5,000 検体の短鎖型のシークエンスデータの解析に 128G(48Cores)の 10 ノード分の年間資源が想定されている。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。

4. 前年度までに得られた研究成果の概要

新規課題のため該当なし

5. 今年度の研究成果の詳細

課題１）100 遺伝子の構造多型のリファレンスパネルの構築とそのための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情報基盤の設計と運用（長﨑、松田、関谷、塙、深沢、村田）（図３）

本年度の解析を進めるために、約 5 万塩基から 10 万塩基の範囲の 100 か所の遺伝子に対して、長鎖型法で得られた配列情報を鋳型の作成、また、作成した鋳型に対して、5,000 検体の短鎖型のシークエンス情報をアライメントすることで再解析することを行い全体の解析を行うための解析パイプラインの実装を東京大学の電算機資源および mdx を主に利用することで確立することができた（図３に解析概要を、また、図４に１遺伝子領域のハプロタイプを視覚化した結果を示す）。なお、5 に記載を行ったが GPU におけるアラメントソフトウェアが学術機関に無償公開されたことから、mdx の利用を想定し、より高速に解析処理を行える可能性を評価するために、同ソフトウェアの小規模なオンプレシステム上での実行パイプラインの試験実装を行った。

また、ネットワークの構築として、SINET6 の L2VPN を使ったネットワーク構成を京都大学・東京大学のスパコン・MDX 間を実現するための準備を進めた。

他に拠点間の高速データ転送のために、村田が新たに開発を行った rsync の代替高速転送ソフトウェア hsync を試験導入しパフォーマンス評価を進めた。

課題２）シークエンサから取得されたアノテーションに活用する情報を他の拠点に効率良く展開するための設計検討と実装（大川、長﨑、深沢、塙、関谷、村田、大川、前原、南里）（図６）

本研究では、鋳型の情報として長鎖型のシークエンサの情報が必要となる。2022 年 10 月の九州大学の大川が導入した長鎖型のシークエンサから出力される一部の情報について、京都大学やデータ活用社会創成プラットフォーム（MDX）拠点での解析が必要となった。そこで、本研究課題においては、前半期間において、九州大学と京都大学の間の当該シークエンサとの接続のための 10G 接続のネットワークを九州大学のシークエンサルームに新たに導入することで拠点間でスムーズにデータ転送ができるように構築を進めた。また、10 月末には、京都大学のチーム（長﨑、浅倉、橋本）が訪問し機器の立ち上げとネットワーク構築を九州大学の大川、南里、前原

と進めた。これにより、新たにシークエンサが導入された九州大学と、主に京都大学の拠点間において、効率よく転送して情報解析を効率良く進めていくための準備を前半期間において整えた。

さらに、後半期間においては、そのシステムを用いて実際にシークエンスされたデータについて mdx や東大システムとも hcptools や rsync を用いて転送を行い前半期間で構築を進めた情報解析パイプラインを用いて解析を進めることができた。また、一部の解析については、GPU ベースの解析が必要であり、後半に提供された GPU システムを用いて mdx 上で情報解析を進めた(DeepConsensus 他)。

6. 進捗状況の自己評価と今後の展望

課題１

約 5 万塩基から 10 万塩基の範囲の 100 か所の遺伝子に対して、長鎖型法で得られた配列情報を鋳型の作成、また、作成した鋳型に対して、5,000 検体の短鎖型のシークエンス情報をアライメントすることで再解析することを行い全体の解析を行うための解析パイプラインの実装を東京大学の電算機資源および mdx を主に利用することで確立することができた。

2022 年度申請時は、グラフゲノムの構築(Step1)については vg (https://github.com/vgteam/vg) の Giraffe を用いて行う計画を進めていたが、年度半ばに申請者の保有する長鎖型のシークエンサをもちいたリファレンスハプロタイプ群の同定手法の開発と実装。また、Step2 において、Step1 で得られた鋳型を用いた、短鎖型シークエンサに対するハプロタイプ推定手法の実装、構造多型の推定手法のプロトタイプ実装が完了したことから、2023 年度は同手法を用いた解析に切り替えて 5,000 遺伝子を目標に計算を進める準備を整えることができた。

また、GPU におけるアラメントソフトウェア(Parabricks)が 2022 年度途中で学術機関に無償公開されたことから、2023 年度での本格運用の試験検証を行い CPU を用いた計算に比べて約 10 倍程度高速に処理できる知見を得ることができた。今後、GPU と CPU をどのように併用すればよいか検討が可能な知見とすることができた。

課題２

本研究では、鋳型の情報として長鎖型のシークエンサの情報が必要となる。2022 年度の前半において、九州大学と京都大学の間の当該シークエンサとの接続のための 10G 接続のネットワークを九州大学のシークエンサルームに新たに導入した。

また、10 月末には、京都大学のチーム（長﨑、浅倉、橋本）が訪問し機器の立ち上げとネットワーク構築を九州大学の大川、南里、前原と進めた。これにより、新たにシークエンサが導入される九州大学と他拠点との間に京都大学の間でシークエンサ情報の解析を効率よく転送して情報解析を進めていくための準備を整えた。さらに、2022 年度後半においては、25 検体の鋳型のシークエンサの情報を取得することで、2023 年度の日本人の約 100 検体の鋳型に基づいた解析を行える体制を整えられたと考えている。

さらに、拠点間の高速データ転送のために、村田が新たに開発を行った rsync の代替高速転送ソフトウェア hsync(beta 版)を一部のシステムに導入し、2023 年度の本格利用を視野にいれて試験利用を進めた。全体としては、1PB 以上の拠点間データ転送をおこなったが、hcptools のように想定されたパフォーマンスが hsync において、同環境では得られなかったことから、今後さらにハイブリッドクラウドの環境で性能がでるように調査を進めていく予定である。

7. 研究業績

(1) 学術論文 （査読あり）

[1] M. Nagasaki, Y. Sekiya, A, Asakura, R. Teraoka, R. Otokozawa, H. Hashimoto, T. Kawaguchi, K. Fukazawa, Y. Inadomi, K. T. Murata, Y. Ohkawa, I. Yamaguchi, T. Mizuhara, K. Tokunaga, Y. Sekiya, T. Hanawa, R. Yamada, F. Matsuda. Design and implementation of hybrid cloud system for large-scale human genomic research, *Hum Genome Var*, 10: 6, 2023.

[2] M. Fujiwara, H. Hashimoto, K. Doi, M. Kujiraoka, Y. Tanizawa, Y. Ishida, M. Sasaki, M. Nagasaki. Secure secondary utilization system of genomic data using quantum secure cloud. *Sci Rep*, 12, 18530, 2022.

(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）

[3] 長﨑 正朗, “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”, 学際大規模情報基盤共同利用・共同研究拠点 第 14 回シンポジウム, 2022/7/7

[4] 長﨑 正朗, “日本人の公開可能な長鎖型集団パネル構築とその意義について”, PacBio ユーザーグループミーティング 2022, 2022/5/17

(5) 公開したライブラリなど

該当なし

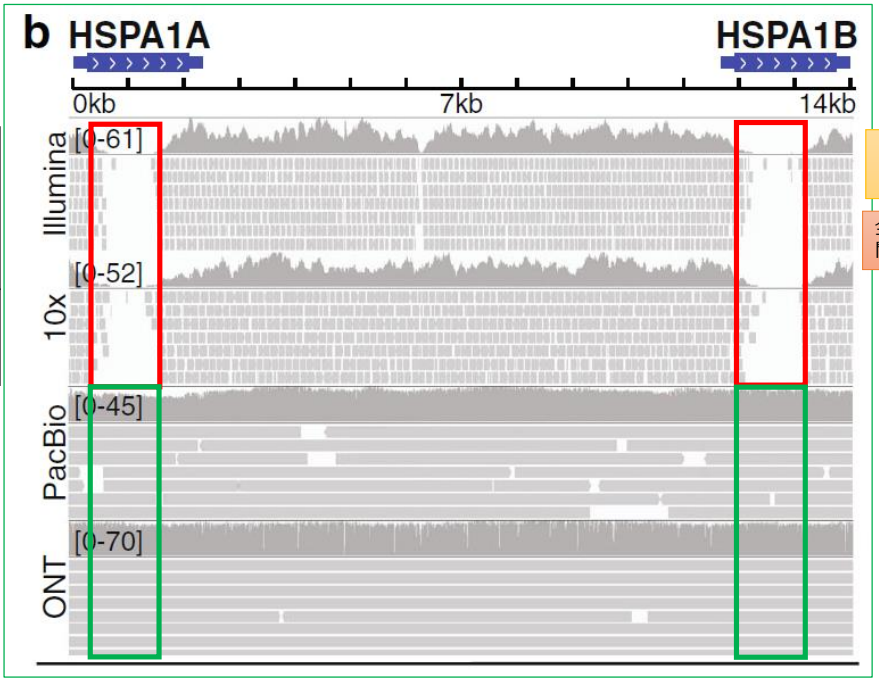
(6) その他（特許，プレスリリース，著書等）

該当なし

図 1

長鎖型シークエンスデータの解析の意義

構造多型領域の領域が未整備（下図：よく似た配列のためにsrWGSでは判別困難な遺伝子の領域（赤枠内））→ 長鎖型シークエンサで読み取ることで解決可能に



他の例は参考資料参照

20

Genome Biology 2019 Ebbert et al

図 2

【研究目的】日本人のゲノム情報についてグラフゲノムによる解析による**構造多型（※）**のカタログ同定と疾患解析での活用

構造多型：染色体上の複雑な配列の集団内の多様性を指す。SNVなどの１塩基の単純な集団内の多様性とは区別して記載する。lrWGSを用いることで徐々に解明されつつある。

約100人の日本人のlrWGSゲノム情報での解析

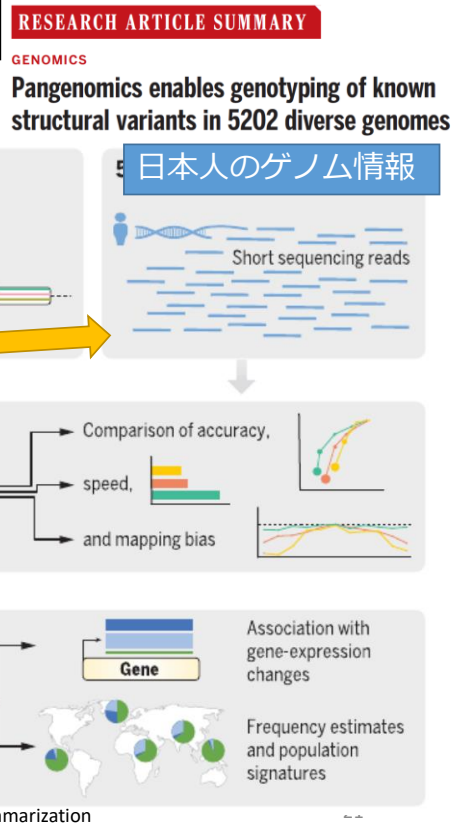
約5000人の日本人のsrWGSゲノム情報での解析

アノテーションと疾患解析での基盤としての活用

【課題1】ハイブリッドクラウド情報基盤での実装
【課題2】アノテーション基盤情報の取得と拠点間転送

Siren *et al Nature* 374(1461) 2021

海外の先行研究



https://github.com/graph-genome/graph_summarization

図３

システム全体構成と役割担当

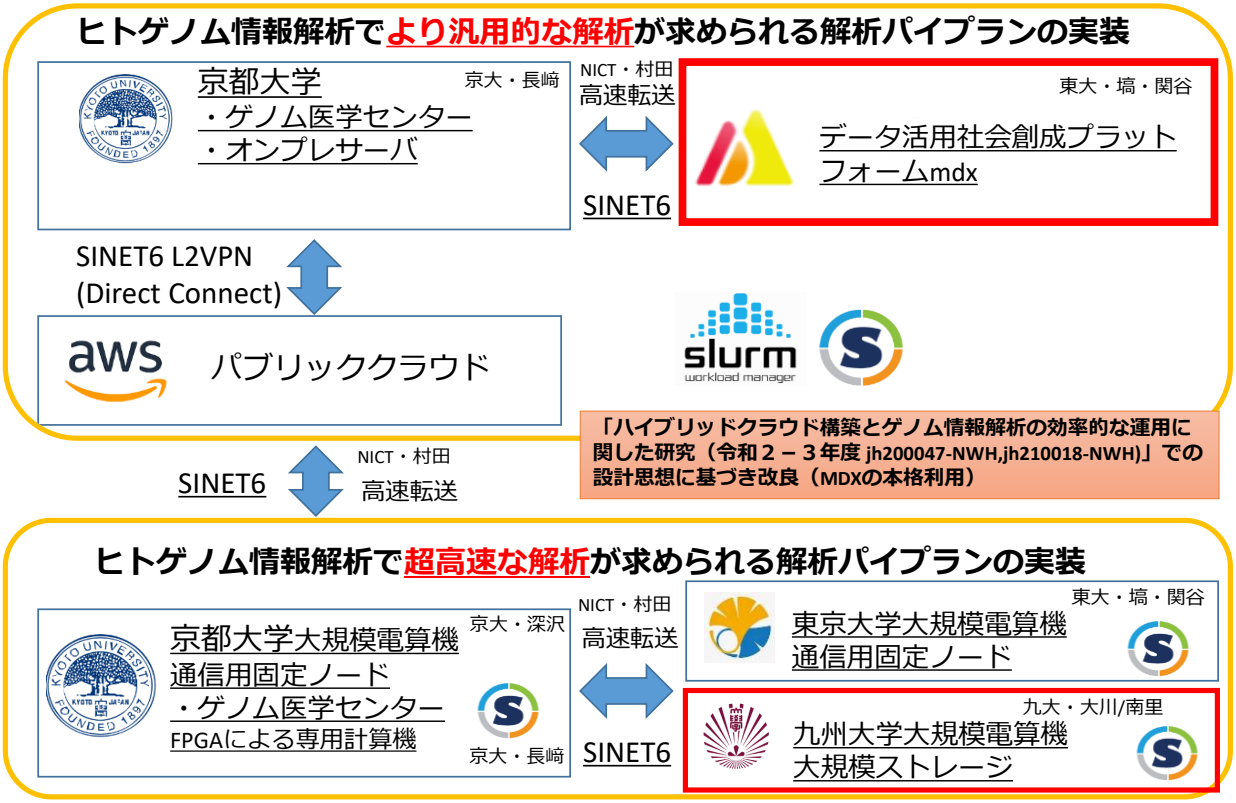


図４

解析フロー概要

② ①と短鎖型シークエンス情報の活用

①のパネルの活用による短鎖型シークエンス情報から対象遺伝子領域の変異および完全phase情報を構築する手法

①のlrWGSのphasingパネルを用いsrWGSから日本人の50検体の完全なphasing情報を推定・復元した例（右図）

同様に他の遺伝子領域のハプロタイプを計算

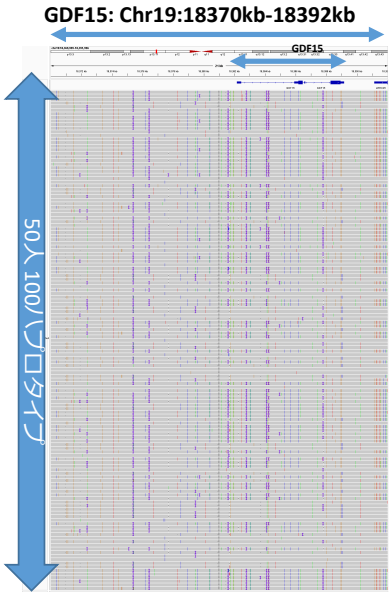


図５

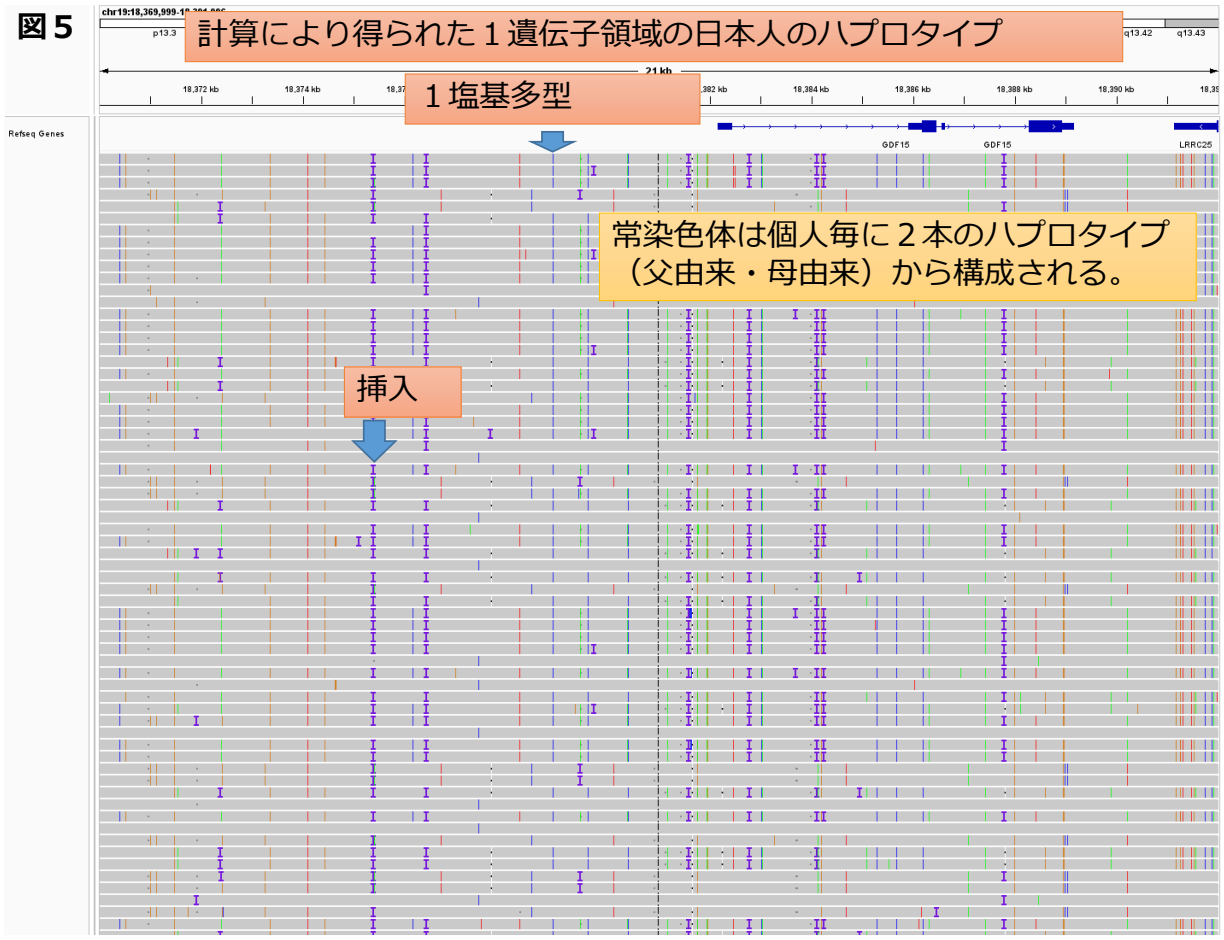


図６　課題２）シークエンサから取得されたアノテーションに活用する情報を他の拠点に効率良く展開するための設計検討と実装（長﨑, 大川, 深沢, 南里, 関谷, 塙, 村田）