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

【課題番号】jh240015【研究代表者名】長﨑 正朗

九州大学生体防御医学研究所高深度オミクスサイエンスセンターバイオメディカル情報解析分野

【研究目的と成果】

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。

本研究課題の拠点とメンバ構成

【九州大学】

生体防御医学研究所

研究課題代表者	長﨑 正朗
副代表者	大川 恭行
他メンバ	関谷 弥生　浅倉 章宏　橋本 洋希　寺岡 凌 男澤 良子　町田宗聡　松原太一　前原 一満
情報基盤研究開発センター	南里 豪志

【東京大学】

情報基盤センター	塙 敏博
大学院情報理工学系研究科	関谷 勇司

【京都大学】

学術メディアセンター	深沢 圭一郎

【国立情報学研究所】

アーキテクチャ科学研究系	竹房　あつ子　合田　憲人
クラウド基盤研究開発センター	大江　和一
生命情報・DDBJ研究センター	丹生　智也

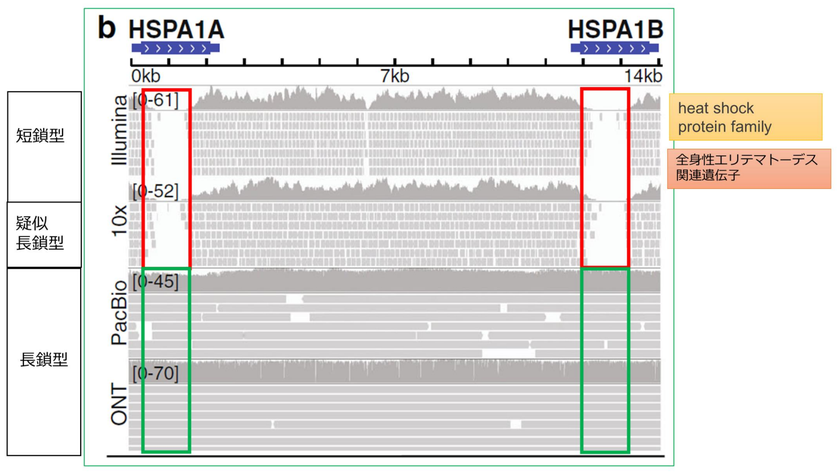
【情報通信研究機構】

総合テストベッド研究開発推進センター　村田 健史

敬称略

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

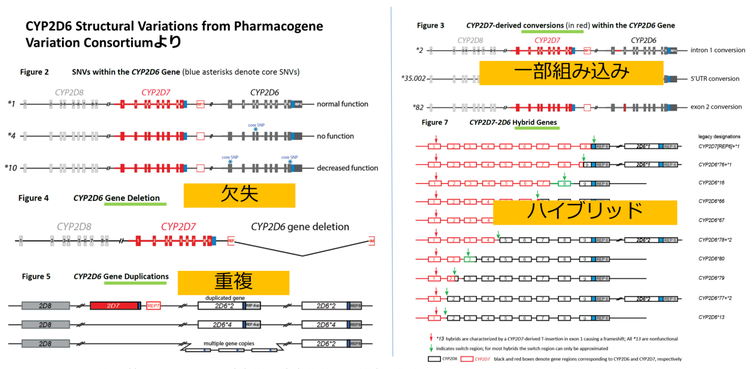
Genome Biology 2019 Ebbert et al

当拠点公募型共同研究として実施する必要性　CYP2D6

処方される薬剤の25%がCYP2D6遺伝子で代謝される。人毎にコピー数、また、１つ１つの遺伝子の代謝能力が異なる。
→ 高精度長鎖型シークエンサ※でようやく解読できる状況となってきた。

※ここで高精度とは１塩基の誤り精度が短鎖型と同等の性能を有するタイプの長鎖型シークエンサを指す

鋳型になる塩基配列のグラフ構造がわかれば、短鎖型のシークエンス情報でもある程度、どのような構造になっているかを推定することが可能



https://a.storyblok.com/f/70677/x/7073c64512/cyp2d6_structural-variation_v2-1.pdf

当拠点公募型共同研究として実施する必要性

海外の先行研究

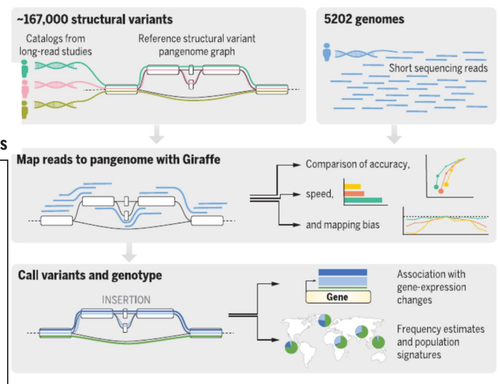
RESEARCH ARTICLE SUMMARY

GENOMICS

Pangenomics enables genotyping of known structural variants in 5202 diverse genomes

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。

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)



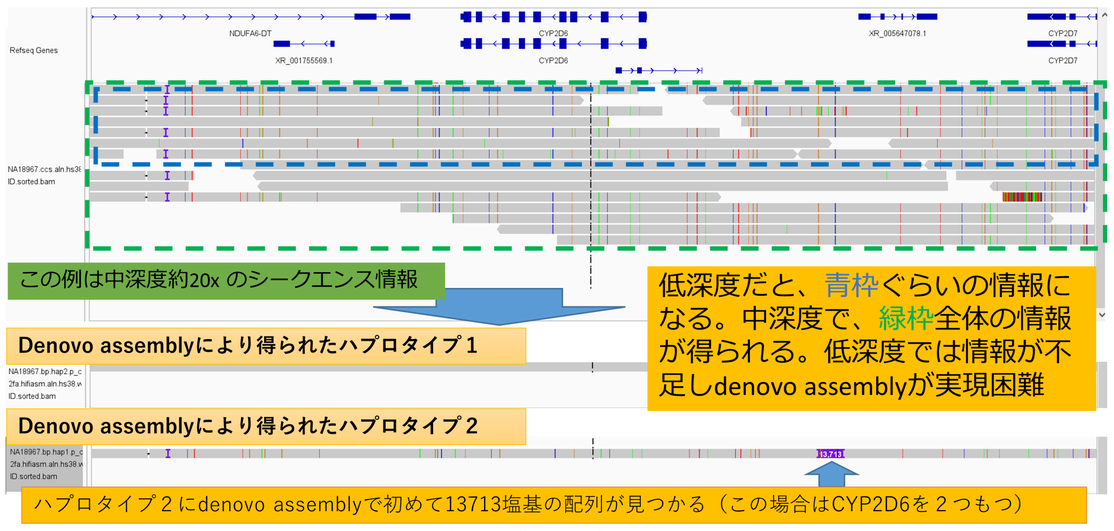
https://github.com/graph-genome/graph_summarization

構造多型のカタログ構築解析のためには、大規模ストレージへの情報集約と電算機資源による解析が必要

【研究目的】日本人のゲノム情報について申請者が取得した１００検体の低深度長鎖型シークエンス情報を中・高深度の長鎖型シークエンス情報へ拡充し、*denovo* assemblyなどの新たな情報解析を行い高精度なハプロタイプリファレンスパネルの構築すること

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)

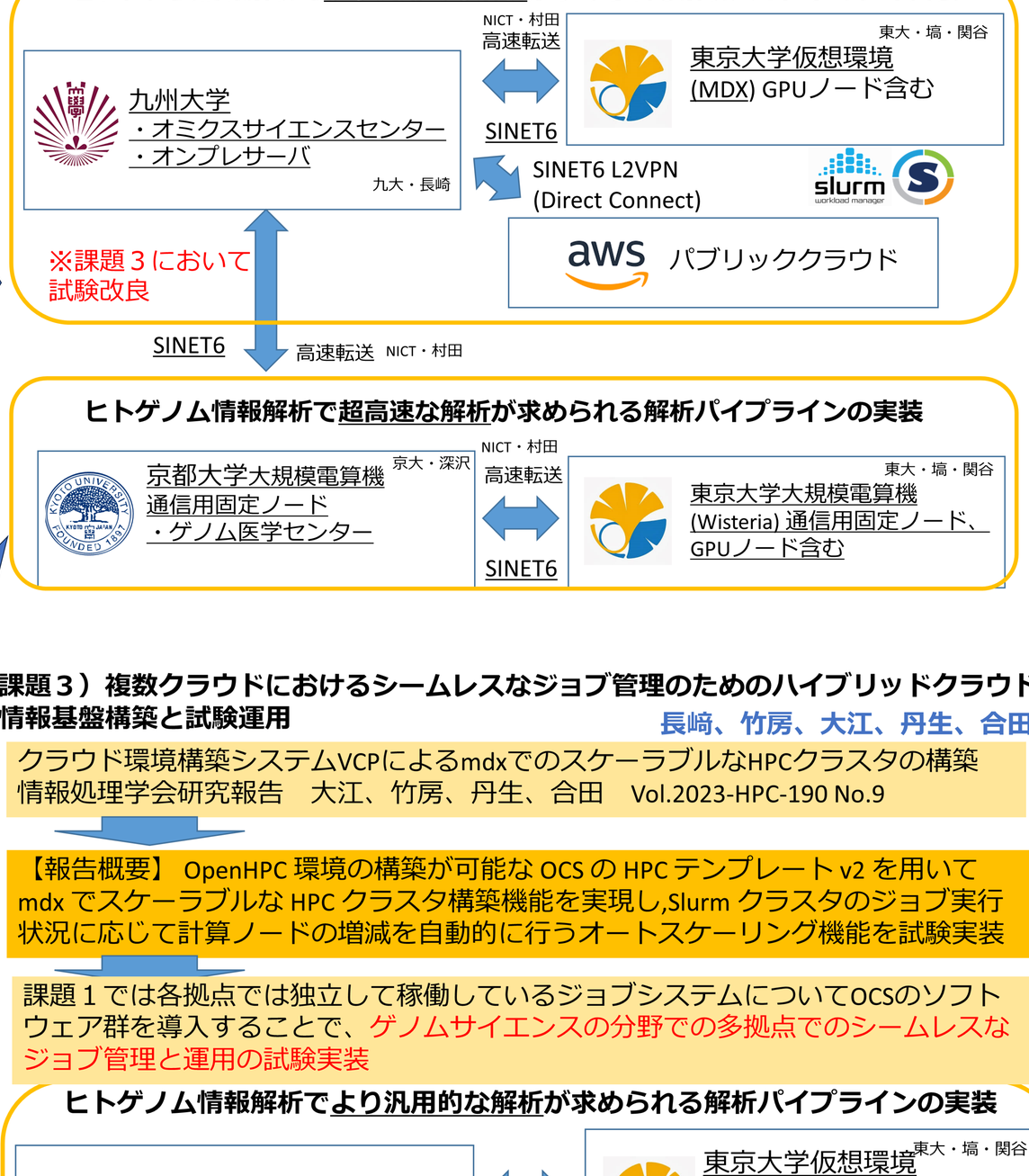


低深度だと、青枠ぐらいの情報になる。中深度で、緑枠全体の情報が得られる。低深度では情報が不足しdenovo assemblyが実現困難

※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

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、大川他

システム全体構成と役割担当



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

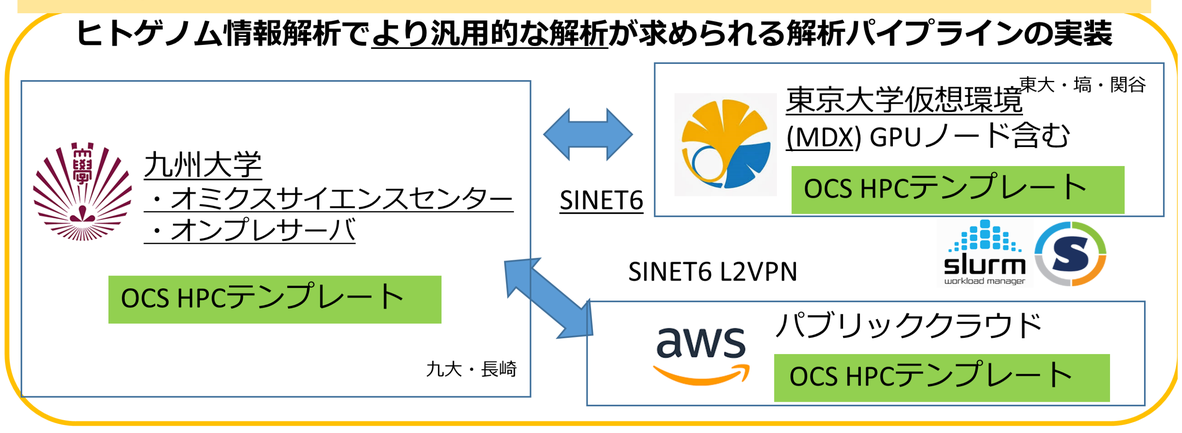
課題３）複数クラウドにおけるシームレスなジョブ管理のためのハイブリッドクラウド情報基盤構築と試験運用

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



【本研究課題に関連する論文】

K. Hirayasu, S.S. Khor, Y. Kawai, M. Shimada, Y. Omae, G. Hasegawa, Y. Hashikawa, H. Tanimoto, J. Ohashi, K. Hosomichi, A. Tajima, H. Nakamura, M. Nakamura, K. Tokunaga, R. Hanayama, M. Nagasaki. 'Identification of the hybrid gene LILRB5-3 by long-read sequencing and implication of its novel signaling function', *Front Immunol* (15), 1398935, 2024.

【国内会議発表】

長﨑 正朗, 未来医療へのヒト情報解析基盤構築と実装", 未来社会デザイン統括本部＆データ駆動イノベーション推進本部　合同シンポジウム2023, 2022/9/4

長﨑 正朗, ヒトゲノム情報や臨床情報のセキュリティと情報解析の取り組みについて, αxSC2023Q セキュリティとスーパーコンピュータシンポジウム, 2023/7/31

長﨑 正朗, ヒトゲノムと臨床情報の統合解析に向けたハイブリッドクラウド基盤構築とパブリッククラウドの活用, 教育と研究のDXフォーラム, 2023/7/27

https://nagalab.csml.org/