jh230016

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（九州大学　生体防御医学研究所）

概要

ヒトゲノム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対し上の一部の情報についての試験的な解析を円滑に行うことを令和 2-3 年度 jh200047-NWH,jh210018-NWH の課題において進め、論文成果を報告した(Tanjo *et al.* 2021, Nagasaki *et al.* 2023(10:6))。さらに、昨年度の課題（jh220014）では、日本人の長鎖型情報を活用し、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。今年度は、前年度の成果を発展させ、独自に取得した約 100 検体の長鎖型情報を活用し、約 5,000 検体の短鎖型法の検体のさらなる構造多型カタログ構築とアノテーションを進めた。その一部は論文として成果を得ることができた（Hiyarasu *et al.* 2024)。

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

(2) 課題分野

データ科学・データ利活用課題分野

(3) 共同研究分野(HPCI 資源利用課題のみ)

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

九州大学の長﨑、松田のチーム（他、関谷弥生、男澤、寺岡、町田、松原）は、構造多型の解析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う。東京大学の塙、関谷は、東京大学の電算機資源(Wisteria)、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う。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る。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う。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う（他、浅倉、橋本)。研究課題２のシークエンサからの拠点間データ転送においては、九州大学の大川チーム（前原、南里）で得られたデータの他拠点転送と情報解析（長﨑、浅倉、橋本が担当）をシームレスに行うことで評価を進める。

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。シークエンス拠点との連携など、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。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対し上の一部の情報についての試験的な解析を円滑に行うことを「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)」において進めまた論文として成果を報告した(Tanjo *et al.* Journal of Human Genetics 2021, Nagasaki *et al.* Human Genome Ver 2023 (10:6))。

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。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)。また、海外の先行研究が行われている（図１）。そこで、昨年度の課題 (jh220014) では、申請者が進めている日本人の約 50 検体の長鎖型情報を活用し、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の約 100 遺伝子を中心に解析を行い構造多型のカタログの構築を進めた。今年度は、前年度の成果を発展させ、申請者が取得した約 100 検体の長鎖型情報を活用し、より多くの構造多型のカタログの構築とアノテーションを進める。

なお、Bioinformatics 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。また、一部は GPU 計算ノードを用いることで効率的に解析することが前年度の試験検証によって示された。そこで、GPU 計算ノードを含むハイブリッドクラウド上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。また、複数拠点間の L2VPN 等よりセキュリティに配慮した構成にできないかなども併せて検討を進める。

3. 当拠点の公募型共同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。また、得られた新規構造多型の機能を引き続き実験することで形質や疾患に関わるあらたな遺伝要因の探索につなげることができる。

4. 前年度までに得られた研究成果の概要

50 人の長鎖型シークエンサの情報を鋳型にして約 5,000 人の 100 遺伝子の日本人の srWGS のゲノム情報の解析を試験的に進めた。

昨年度の試験的な解析成果に基づき、発展的に今年度の 100 人の長鎖型シークエンサの情報を鋳型にした約 5,000 人の数千遺伝子の本格解析を進めた（図２）。

5. 今年度の研究成果の詳細

課題１）数千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、関谷、塙、深沢、村田、大川、松田）

昨年度に得られた知見に基づき、以下の改良をすることで日本人の構造多型のリファレンスパネルの構築を進めている。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おおよそ出来上がりつつあることから、2023 年度は同手法を用いた解析に切り替え、遺伝子を拡大することを目標に計算を進めた。

Step1 については、新たに得られた長鎖型のシークエンス配列を京都大学の学術情報メディアセンターの計算リソースを用いて計算を進めた。

また、Step2 については、解析実績のある bwa (https://github.com/lh3/bwa) を用いて、Step1 の鋳型のハプロタイプにアライメントをおこなうことで解析を進めている。128G (56Cores) で１検体あたり２日以内に計算が完了することができる。同計算は、九州大学情報基盤研究開発センターで解析を進めた。

長﨑はソースコードから singularity ベースでのコンパイル、同プログラムでのヒト染色体の一部の領域での数検体での試験解析を行うことで、プログラムの試験稼働などの準備を行った後に解析を進めた。なお、Step2 の解析パイプラインの一部のアライメント処理について、2022 年度に小数検体で GPU 計算機を試験的に用いることで CPU マシンに対して、より高速にできる結果を得ており、同知見を活かし、2023 年度は CPU に加えて、Step1 で利用を行う GPU 計算機(mdx 等)を併用し高速化やそのほか改良を進めた。

なお、全体の解析フローは、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と mdx にデータを展開し前処理を行った後、これらの情報を、東京大学の電算システムの固定ノードに高速通信による転送をした後（Wisteria または大規模仮想環境 mdx）に展開し情報解析を行う予定であったが、九州大学に 4 月に異動したことから、九州大学のオンプレシステムを前半期間において構築し、これらのシステム経由で接続できるように組み換えを進め後半年度において解析を継続することで目標を達成した。

なお、昨年度は Oakbridge-CX を用いて Step2 相当の解析を進めてきたが、今年度停止することから Wisteria-O での環境への構築を新たに行う予定であったが、アーキテクチャが異なることから計算パフォーマンスが得られない（主なソフトウェアが動作しない）ことから後半年度は、Wisteria-A 電算資源にリソースの変更申請を行うことで解析を進めた。

それらの計算結果について、課題２と連携しつつ、九州大学の前半期間において構築を進めたオンプレサーバ（高速転送のために専用ノードが準備される場合にはそのノード）に転送をおこなうことで最終的なデータ統合解析を進めた。

その中の１つの成果として、次のことが挙

げられる。Leukocyte immunoglobulin (Ig)-like receptors (LILRs)の領域のうち特にLILRB3 と LILRA6 の間においてコピー数多型が存在していることからsrWGSのみの解析では確定することが困難であった。しかし、長崎が独自に得た、長鎖型に基づく日本人集団のシークエンス情報を鋳型（図３）として使うことで高精度に一般的なsrWGS解析(図４）では得られないコピー数のハプロタイプ構造を推定する新規ソフトウェアの開発を行うことができた (Nagasaki *et al.* In submission)。同ソフトウェアを用いることで図５のようなdiploidでのプロファイルの構築が可能となるとともに、haploid のペアを推定することが可能となった。この中で、いままで知られていなかった LILRB3 と LILRA6 の融合遺伝子を同定することができた（図６） (Hirayasu *et al.* Frontiers Immunol. 2024)。さらに、同融合遺伝子は、日本人集団に低頻度で存在することが確認され、今後このLILRB3-LILRA6の融合遺伝子の形質や疾患への影響をｹﾞﾉﾑｻｲｴﾝｽの研究コミュニティーを中心として研究をおこなえる基盤としての成果を得ることができた。

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、長崎、深沢、村田）

課題１で鋳型となる長鎖型シークエンサの情報に関連した不死化リンパ球を九州大学（大川）が管理を行っている。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のオンプレミス環境とハイブリッドクラウド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村田の課題と連携をして行うとともに、課題１のアノテーションなどの拡充に活用している。今年度全体を通じて約40 検体の長鎖型シークエンサによる情報の取得、および、同ハイブリッドクラウド内へのデータ転送と解析を進めた。

6. 進捗状況の自己評価と今後の展望

課題１

必要に応じて、SINET6 の L2VPN をもちいて、パブリッククラウドに展開を行いアノテーションなどの下流解析をおこなえるように構築を後半年度に進める予定であったが、九州大学構内のSINETと接続するためのネットワークの新規敷設に関連した機器の納入が年度内に完了しなかったことから 2024 年度以降の課題となった。

課題２

京都大学時に構築をしていたハイブリッドクラウド相当の実装を九州大学のオンプレ経由で実装するとともに、九州大学のオンプレシステムで得られたシークエンス情報等を同ネットワークを通じて各拠点に転送して解析できる安定的な系の構築を進めた。後半には、分担研究者の村田が開発を進めているhsyncの試験運用を行うことで来年度以降のより安定的な拠点間通信ができる体制を整えることができた。

7. 研究業績

(1) 学術論文 （査読あり）

[1] 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.

(2) 国際会議プロシーディングス （査読あり）

(3) 国際会議発表（査読なし）

(4) 国内会議発表（査読なし）

[2]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

[3]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

[4]長﨑 正朗, “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”, 学際大規模情報基盤共同利用・共同研究拠点 第 15 回シンポジウム, 2023/7/6

(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

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

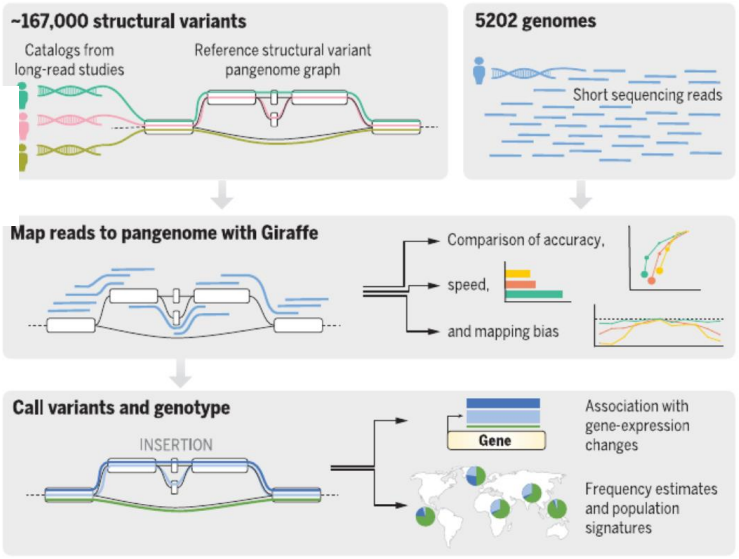
GENOMICS

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

図 2

【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**の長鎖型鋳型活用による情報解析による**数千遺伝子の構造多型**のカタログ同定と疾患解析での活用

【課題1】ハイブリッドクラウド情報基盤解析拠点の実装と運用
【課題2】シークエンス拠点構築とアノテーション基盤情報の取得と拠点間転送

2022年度 jh220014

- 約50人の日本人のIrWGSゲノム情報での解析
- 約5000人の100遺伝子日本人の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

2023年度 jh230016

- 約100人の日本人のIrWGSゲノム情報での解析
- 約5000人の数千遺伝子日本人の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

※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。IrWGSを用いることで徐々に解明されつつある。

図３ 鋳型として使うlrWSGに基づくテンプレートの塩基配列 （構築を行った内容）

長鎖型シークエンスを用いて確定したLILRB3付近の配列パターン
常染色体のため１検体あたり２本存在する。紫の四角が構造多型の領域 おおよそ19kb/38kbの挿入が19kbの欠失がヒト毎に異なることがわかる。

図４ srWSGの一般的な解析結果

短鎖型シークエンスの同一領域のアライメントの結果　lrWGSでみられた19kb/38kbの挿入が19kbの欠失が全く得られないこと、赤枠の部分にアライメントができていないことがわかる。

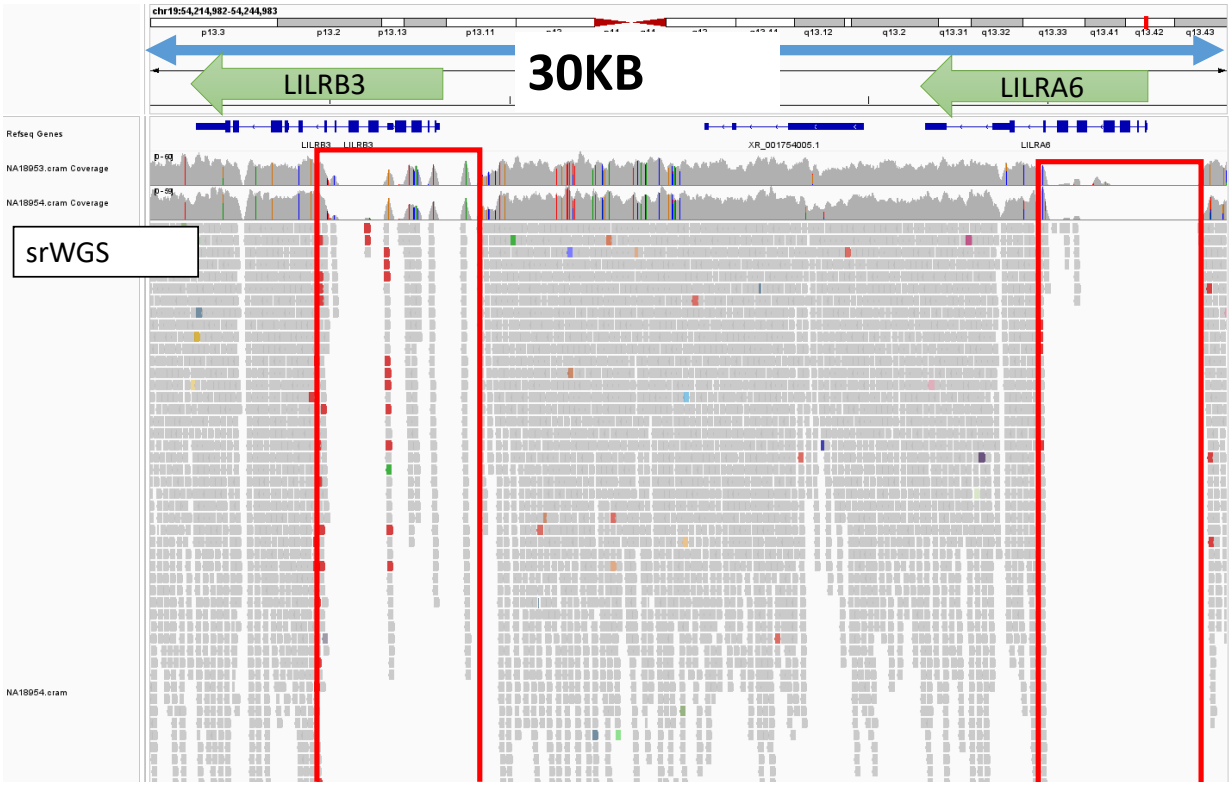
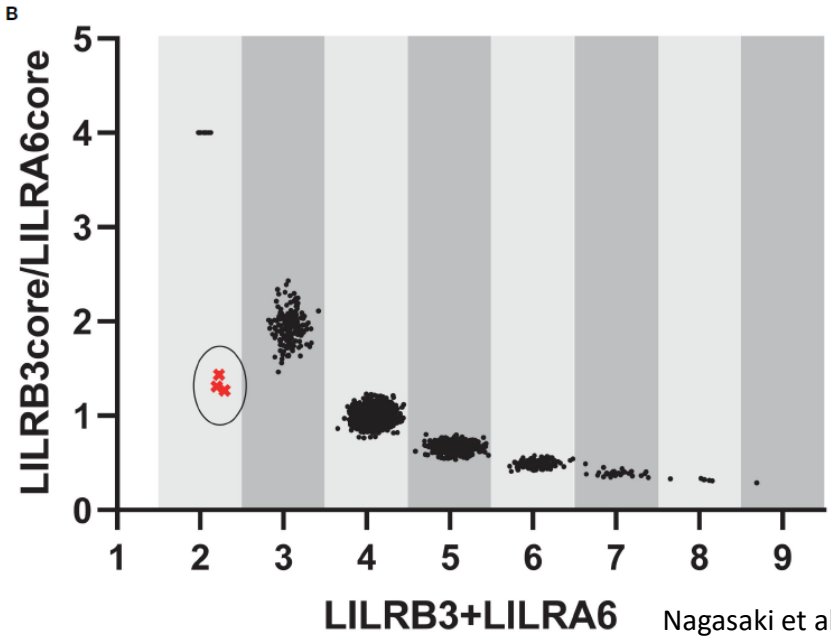


図5 前述の長鎖型シークエンスを鋳型としてsrWGSでコピー数構造を推定した結果

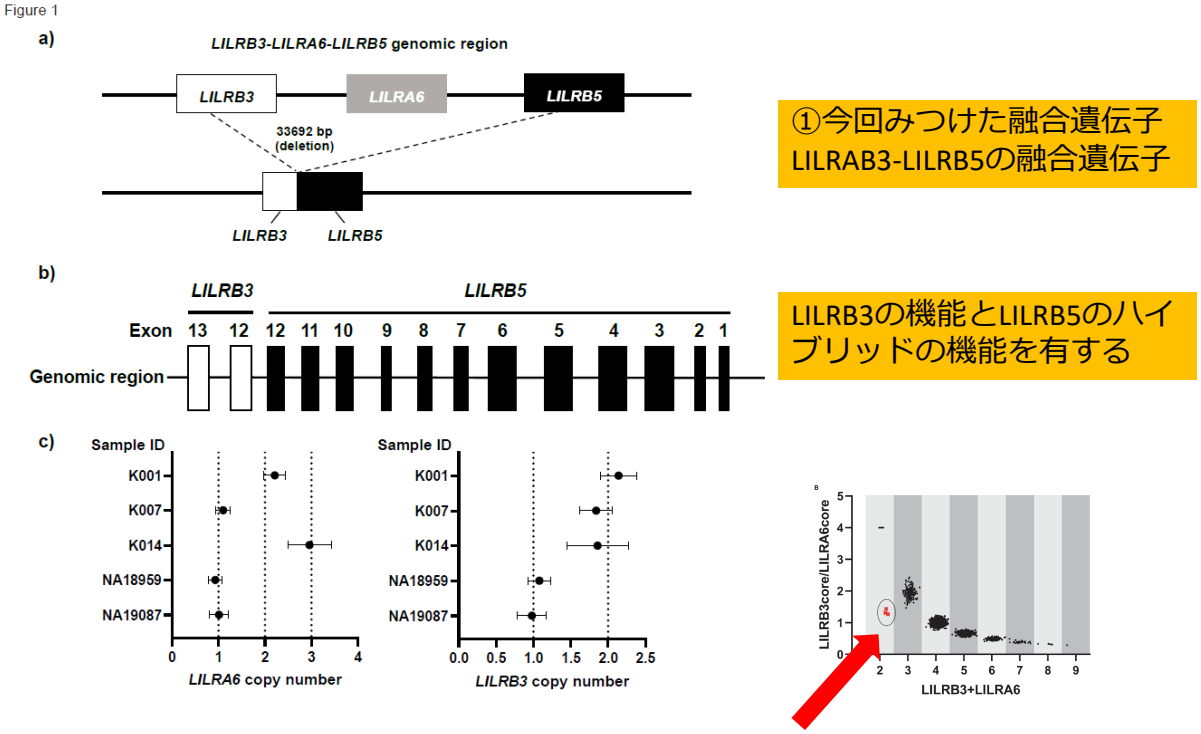
長鎖型シークエンスで得られた結果をテンプレートとしてsrWGSの情報のみ用いて検体のもつコピー数を推定した結果
LILRB3とLILRA6のパターンを分離することが可能となった。



Nagasaki et al In submission

Kyushu University Hirayasu *et al. Frontiers Immunol.* 2024

図6 新たに見つかった融合遺伝子



①今回みつけた融合遺伝子
LILRAB3-LILRB5の融合遺伝子

LILRB3の機能とLILRB5のハイブリッドの機能を有する

Hirayasu *et al. Frontiers Immunol.* 2024

Kyushu University 6