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

【課題番号】　jh210018

【研究代表者名】長﨑 正朗

**京都大学学際融合教育研究推進センター スーパーグローバル
コース医学生命系ユニット
京都大学大学院医学研究科附属ゲノム医学センター**

本研究課題のメンバ全体

研究課題代表者・副代表者

【京都大学】

長﨑 正朗　山田 亮

申請課題の全共同研究者

【京都大学】

松田 文彦　山口 泉　川口 喬久　稲富 雄一

深沢 圭一郎　関谷 弥生　浅倉 章宏　寺岡 凌　男澤 良子

Olivier Gervais　Wang Yen Yen　橋本 洋希

【情報通信研究機構】

村田 健史

【東京大学】

関谷 勇司

【東京大学情報基盤センター】

塙 敏博

【九州大学】

大川 恭行　前原 一満　南里 豪志

【研究目的】

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。

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しつつ、R3に利用可能となる合計約10,000検体の全ゲノムリファレンスパネルの構築を目標に進めていく。

そこで、当研究においては、

１．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

２．仮想環境や大規模電算機資源上でゲノム情報の解析パイプラインの実装を行い

３．円滑に上の一部の情報（数百検体を予定）について試験的に拠点間転送と２のパイプラインを用いたデータ解析を行う

ことを目的とする。

特に、昨年度では十分に対応ができなかった拠点間のデータ転送について重点的に整備を進める。

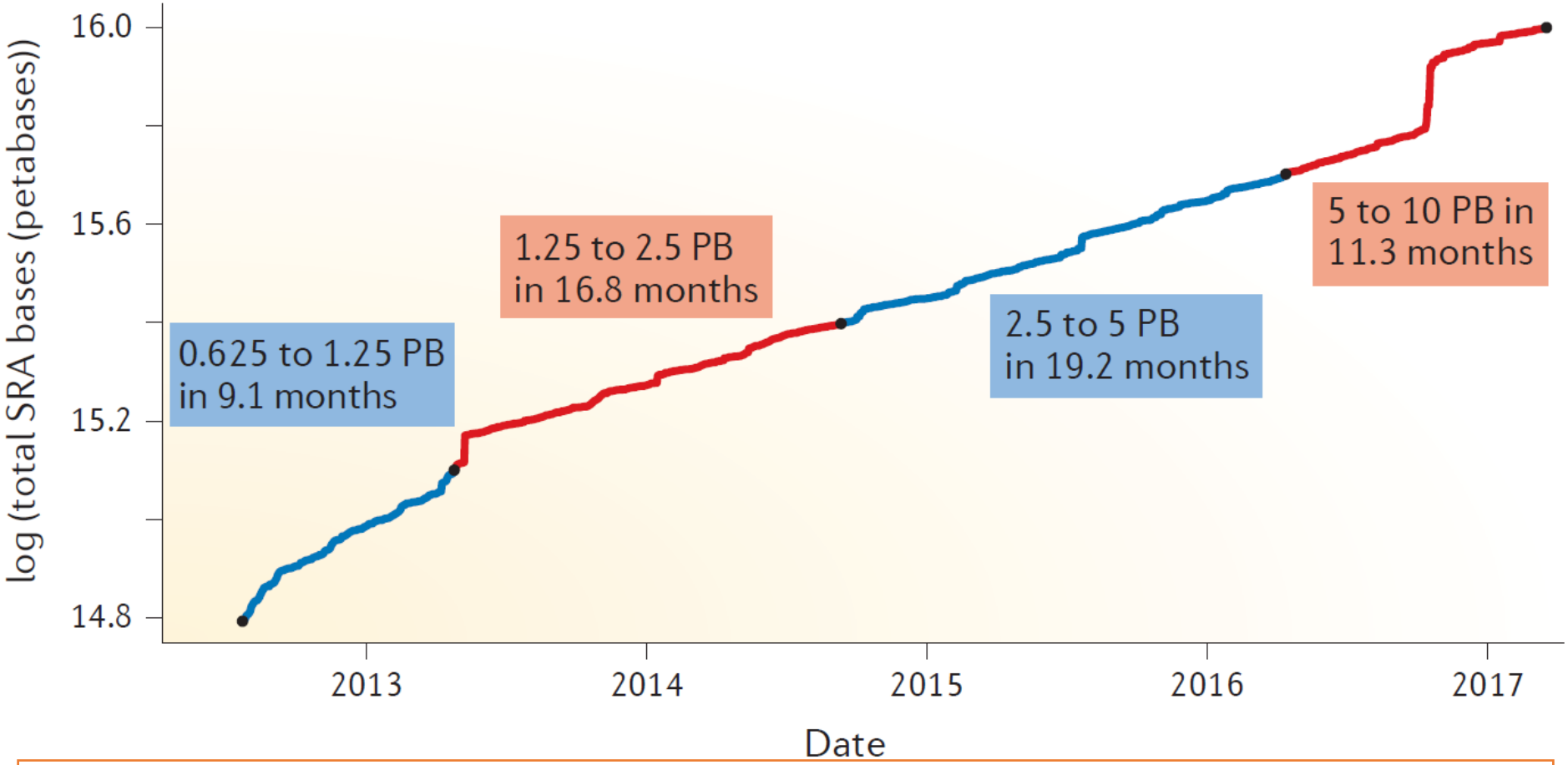
当拠点公募型共同研究として実施する必要性

昨年度5,000検体の規模の解析を行ったが、**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**（**次スライド参照**）。当センターにおいても１万人規模の検体が収集されている。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、シークエンスされた生データから新たな疾患の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。**

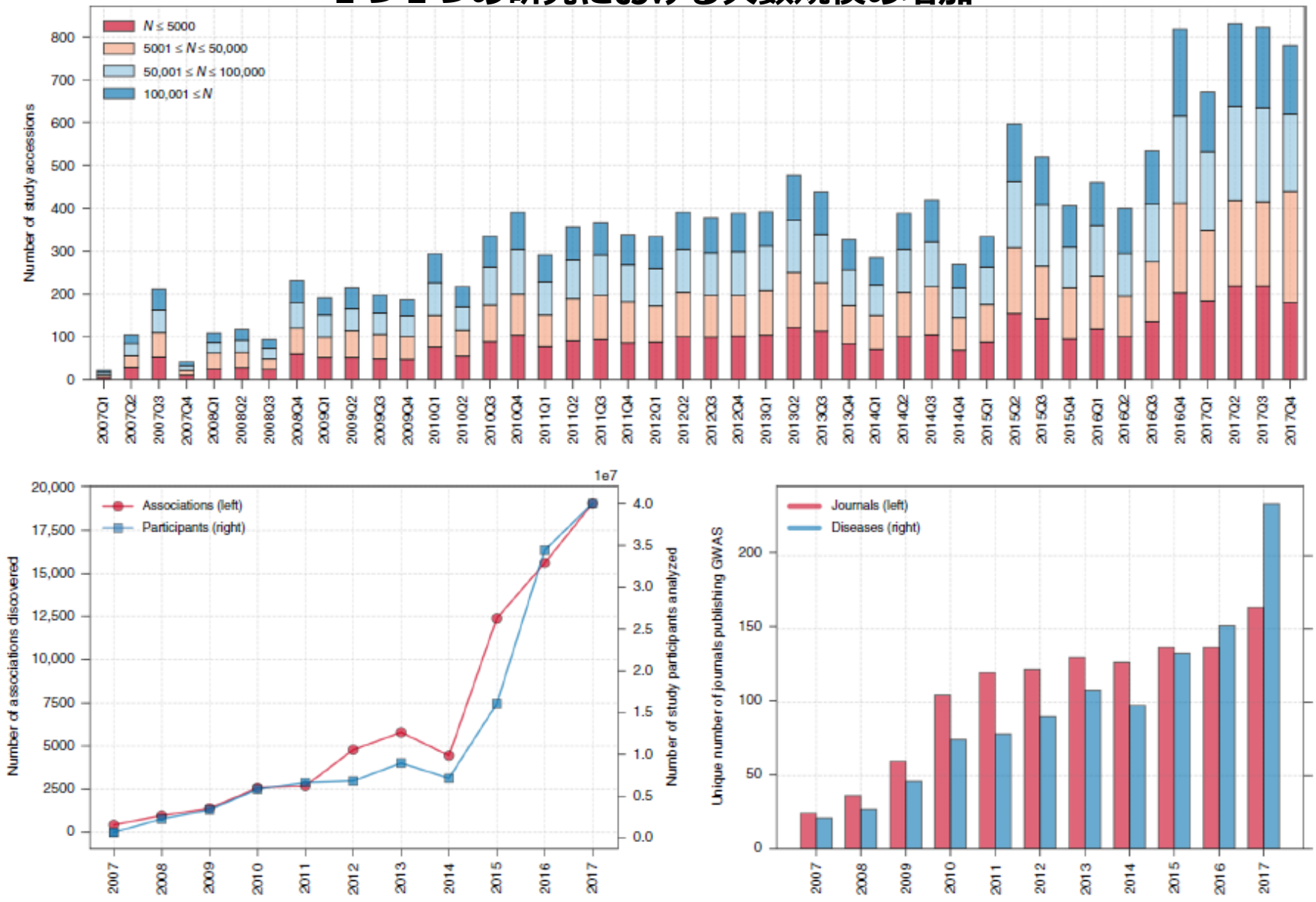
ヒトゲノムシークエンス情報の増加の様子



【図1】ヒトゲノムシークエンス情報の増加の様子
米国のヒトゲノムシークエンス情報の国際データベース (Sequence Read Archive: SRA) に登録されているゲノム情報の総データサイズ (テラバイト) の推移（2020年時点で12PB）
おおよそ1検体あたり30Gとして約40万人のゲノム情報が保存されていることになる。

Cloud computing for Genomic Data Analysis and Collaboration Ben *et al.* Nature Rev Genet 2018

１つ１つの研究における人数規模の増加



【図2】 １つ１つの研究における人数規模の増加の様子　研究プロジェクト毎の解析対象人数が大規模化するとともにまた数が増えていることがわかる。

A scientometric review of genome-wide association studies Melinda *et al* Communication Biology 2019

【研究計画】

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、山田、松田、関谷、深沢、村田）

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、長崎、深沢、村田、塙、関谷）

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

各拠点担当者とその役割

（拠点１）京都大学　ゲノム医学センター
【担当者】長﨑正朗　全体統括　（分担）山田亮、松田文彦　他
【役割】全ゲノム情報のハイブリッドクラウドにおける効率的な解析の設計
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（後スライド参照）

（拠点２）京都大学　メディアセンター
【担当者】深沢圭一郎
【役割１】京都大学と他拠点計算機資源との効率的なデータ分散及び拠点間転送支援
【役割２】京都大学と他拠点とのSINET5を用いたパブリッククラウドへのVPN接続管理

（拠点３）東京大学
【担当者】関谷勇司
【役割】クラウド実装におけるアドバイス、また、試験環境の整備

（拠点４）情報通信研究機構
【担当者】村田健史
【役割】拠点間の高速データ転送技術提供と評価

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

各拠点担当者とその役割

（拠点１）京都大学　ゲノム医学センター
【担当者】長﨑正朗　全体統括
【役割】拠点間の高速データ転送基盤整備と評価（拠点１担当）および全体評価

（拠点２）京都大学　メディアセンター
【担当者】深沢圭一郎
【役割】　拠点間の高速データ転送基盤整備と評価（拠点２担当）

（拠点３）東京大学
【担当者】塙 敏博・関谷勇司
【役割】　拠点間の高速データ転送基盤整備と評価（拠点３担当）

（拠点４）情報通信研究機構
【担当者】村田健史
【役割】拠点間の高速データ転送技術提供および運用支援

（拠点５）九州大学　生体防御医学研究所
【担当者】大川恭行　前原一満
【役割１】全ゲノムシークエンスデータの読み取りとその情報の拠点１への転送
【役割２】解析結果の拠点1からの受け取りと結果評価

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

システム全体構成と役割担当

ヒトゲノム情報解析で<u>より汎用的な解析</u>が求められる解析パイプラインの実装

京大・長崎

京都大学
・ゲノム医学センター
・オンプレサーバ
・DRAGEN

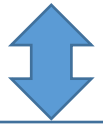
NICT・村田
高速転送

SINET5

東大・関谷

東京大学仮想環境
(Fennel)

SINET5 L2VPN
(Direct Connect)



パブリッククラウド

（重点目標）
- 拠点間の解析フローの運用の安定化
- FPGA機器の運用フローの統合
- １万人規模の全ゲノム解析の達成

SINET5

NICT・村田
高速転送

ヒトゲノム情報解析で<u>超高速な解析</u>が求められる解析パイプラインの実装

京大・深沢

京都大学大規模電算機
(Cray XC40通年(タイプA1):34,560
ノード時間積)

・ゲノム医学センター
FPGAによる専用計算機

京大・長崎

NICT・村田
高速転送

SINET5

京大・長崎

東京大学大規模電算機
(Oakbridge-CX: 146,880
ノード時間積)

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

京大・長崎

全世界のTOP500 supercomputers 60%で利用

➡ 京大オンプレ＆AWS＆東大仮想環境 (Fennel) での共通実装

Singularity v3を用いたパイプライン構築

➡ コンテナの利用による再現性・再利用性・信頼性の担保

解析パイプラインをさらに実装予定

ヒトゲノム情報解析で<u>より汎用的な解析</u>が求められる解析パイプライン

パイプラインはPython / R / C++ / Javaなどさまざまなバイオインフォマティクス解析ソフトウェアのワークフローで構成

処理名称	パイプラインの概要	入力ファイル	出力ファイル
Genotyping	SNPアレイ（Japonica Array (CEL)ファイル）から約66万か所の遺伝型をクラスタリングによって決定するためのパイプライン	CEL	VCF/BED
Imputation	国際1000人ゲノムやそのほかの全ゲノムリファレンスパネルを用いることでSNPアレイでタイピングされた約66万か所のSNPから数千万のSNP情報を復元するパイプライン	VCF/BED	VCF/BED
GWAS	インピュテーション（1KGP / GRIFFIN Panelなど）によって復元された数千万か所の変異情報について指定された条件でフィルタリングを行った後に疾患群と健常群などのcase/controlまたは検査情報などの連続量についての各SNPの偏りを統計手法により検定を行うパイプライン	VCF/BED	TXT
Annotation	GWASによってでてきた結果についてアノテーションを行うパイプライン	VCF/BED	TXT

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

システム全体構成と役割担当

ヒトゲノム情報解析で超高速な解析が求められる解析パイプラインの実装

京都大学大規模電算機
(Cray XC40通年(タイプA1):34,560ノード時間積
・ゲノム医学センター
FPGAによる専用計算機

京大・深沢
京大・長崎

NICT・村田
高速転送
hcptools

SINET5

京大・長崎
東京大学大規模電算機
(Oakbridge-CX: 146,880ノード時間積)

【課題1】との連携

海外（Canada)との試験転送も予定

SINET5

NICT・村田
高速転送 hcptools

九大・大川/前原
九州大学大規模電算機
ITOサブシステムA:3,750ノード計算積
ストレージ 10TB

12
九州大学
生体防御研究所

Illumina社NovaSeq6000
（参考：次ページスペックシート）

全ゲノムシークエン機器

（重点目標）
- 拠点間の転送の安定化・高速化
- 海外拠点との転送試験

NovaSeq6000 スペックシート

Table 1: NovaSeq 6000 System flow cell specifications

Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
2 × 50 bp	65-80 Gb	134-167 Gb	333-417 Gb	N/A
2 × 100 bp	N/A	266-333 Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8 B	1.3-1.6 B	3.3-4.1 B	8-10 B
Paired-end reads CPF	1.3-1.6 B	2.6-3.2 B	6.6-8.2 B	16-20 B
Quality scores[c]				
2 × 50 bp	≥ 85% ≥ Q30			
2 × 100 bp	≥ 80% ≥ Q30			
2 × 150 bp	≥ 75% ≥ Q30			
2 × 250 bp	≥ 75% ≥ Q30			
Run time[d]				
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A
2 × 100 bp	N/A	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 × 250 bp	~38 hr	N/A	N/A	N/A

S4のフローセルを用いて

1台で年間 約8,000人のヒト全ゲノム情報を取得できる。

データ量としては
546000G塩基

1塩基はA/T/G/Cの文字列とASCIIコードで表現される1文字

1検体あたり圧縮して30Gbぐらい

8000検体で解析前のデータで
250TB

【参考関連論文】

1. N. Nariai, K. Kojima, S. Saito, T. Mimori, Y. Sato, Y. Kawai, Y. Yamaguchi-Kabata, J. Yasuda and M. Nagasaki, HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data, *BMC Genomics*, 16(2):S7, 2015.

2. M. Nagasaki *et al*, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals, *Nature Communications*, 6:8018, 2015

3. Y Kawai, T Mimori, K Kojima, N Nariai, I Danjoh, R Saito, J Yasuda, M Yamamoto and M. Nagasaki, Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals, *Journal of Human Genetics* 2015; 60: 581–587, 2015.

4. T. Hasegawa, K. Kojima, Y. Kawai, K. Misawa, T. Mimori, and M. Nagasaki, AP-SKAT: highly-efficient genome-wide rare variant association test, *BMC Genomics*, 17(1):745, 2016.

5. X. Jia, T. Horinouchi, Y. Hitomi, A. Shono, S.-S. Khor, Y. Omae, K. Kojima, Y. Kawai, M. Nagasaki, 17人略, K. Tokunaga, and K. Iijima, Strong Association of the HLA-DR/DQ Locus with Childhood Steroid-Sensitive Nephrotic Syndrome in the Japanese Population, *J. Am. Soc. Nephrol.*, vol. 29, no. 8, pp. 2189–2199, 2018.

6. Y.Y. Wang, T. Mimori, S. S. Khor, O. Gervais, Y. Kawai, Y. Hitomi, K. Tokunaga and M. Nagasaki, HLA-VBSeq v2: improved HLA calling accuracy with full-length Japanese class-I panel, *Hum Genome Var*, 6: 29, 2019.

7. O. Gervais, K. Ueno, Y. Kawai, Y. Hitomi, Y. Aiba, M. Ueta, M. Nakamura, K. Tokunaga and M. Nagasaki. Regional heritability mapping identifies several novel loci (STAT4, ULK4, and KCNH5) for primary biliary cholangitis in the Japanese population. European Journal of Human Genetics, 2021.

8. T. Tanjo, Y. Kawai, K. Tokunaga, O. Ogasawara and M. Nagasaki. Practical guide for managing large-scale human genome data in research. J Hum Genet, 66 (1) 39-52, 2021.

【参考発表】

1) 長﨑 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).

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