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

【課題番号】　jh220014

【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

JHPCN第14回シンポジウム 2022採択課題研究紹介 2022/7/7

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・松田 文彦

申請課題の全共同研究者

【京都大学】

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】

シークエンス技術の進展によりヒト全ゲノム配列情報とそれに付随して必要となる解析資源が爆発的に増えている。

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上で出てくる課題を解決することで

「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）」を進めてきた。

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活用し、令和４年度において、国外において先行研究が行った5,202検体と同程度の約5,000検体の短鎖型法の検体の解析を目標に構造多型のカタログの構築（図１）を行う。

図 1 【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

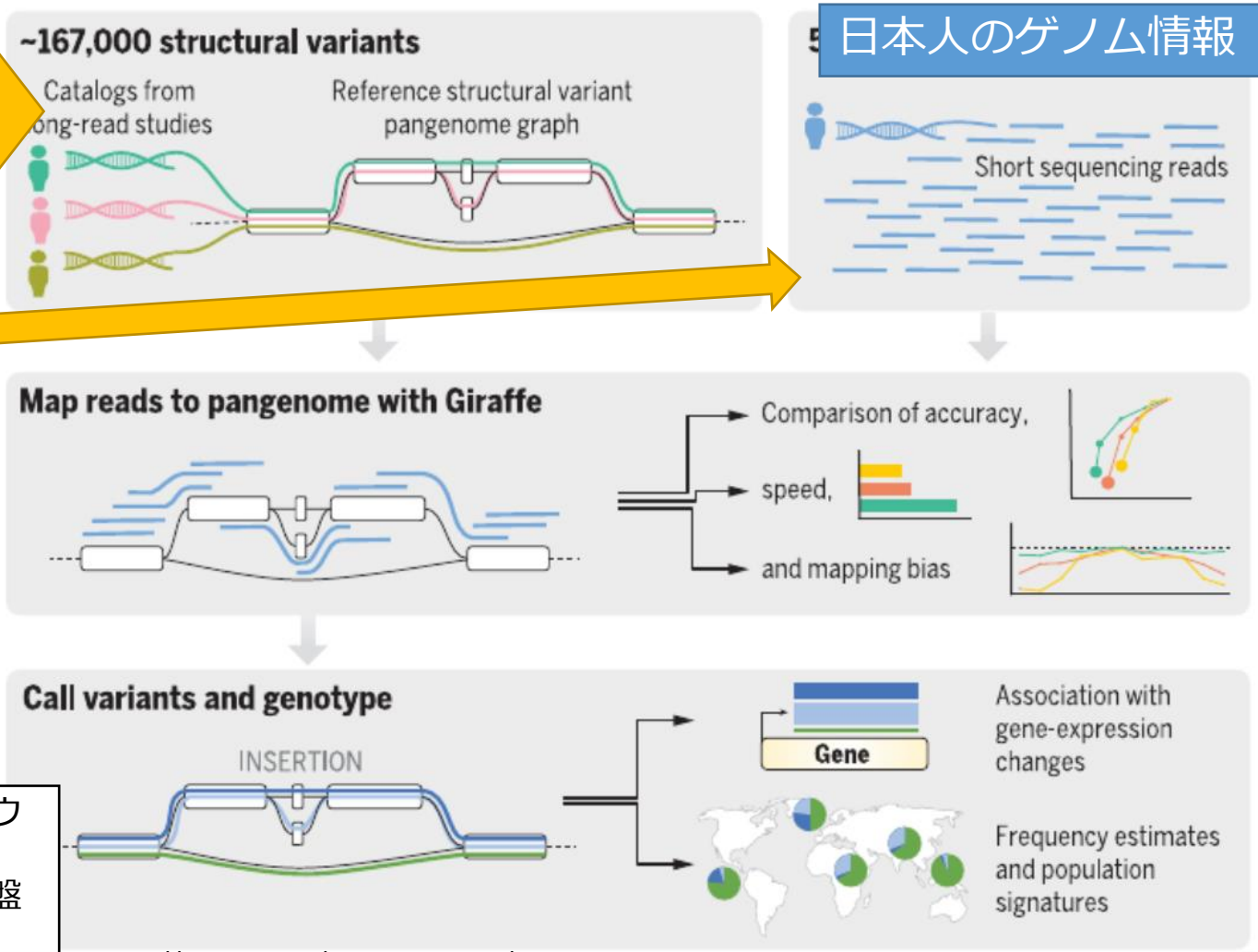
海外の先行研究

RESEARCH ARTICLE SUMMARY

GENOMICS

Pangenomics enables genotyping of known structural variants in 5202 diverse genomes

日本人のゲノム情報



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

Siren *et al* *Nature* 374(1461) 2021

【研究計画】

課題１）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（図２参照）
（長﨑、松田、関谷、塙、深沢、村田）

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（図３参照）（大川、長﨑、深沢、村田）

課題１）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

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）松田文彦　他

【役割】構造多型のリファレンスパネルの構築の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（後スライド図２参照）

（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SINET6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・塙 敏博

【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

図２　　課題１）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

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・長﨑

<u>京都大学</u>
<u>・ゲノム医学センター</u>
<u>・オンプレサーバ</u>

NICT・村田
高速転送
⇔
<u>SINET6</u>

東大・関谷

<u>東京大学仮想環境</u>
<u>(MDX)</u>

SINET6 L2VPN
(Direct Connect)
⇕

aws　パブリッククラウド

slurm workload manager

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

<u>京都大学大規模電算機</u>
<u>通信用固定ノード</u>
<u>・ゲノム医学センター</u>
<u>FPGAによる専用計算機</u>

京大・長﨑

NICT・村田
高速転送
⇔
<u>SINET6</u>

東大・塙・関谷

<u>東京大学大規模電算機</u>
<u>通信用固定ノード</u>

<u>SINET6</u>
⇕
NICT・村田
高速転送

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

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

図３　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、村田）

九州大学
生体防御研究所

全ゲノムシークエンス機器

Illumina社NovaSeq6000
（参考：次ページスペックシート）

九大・大川/前原
九州大学大規模電算機

SINET6

NICT・村田
高速転送

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

京大・深沢
京都大学大規模電算機
通信用固定ノード

・ゲノム医学センター
FPGAによる専用計算機
京大・長﨑

NICT・村田
高速転送

SINET6

東大・塙・関谷
東京大学大規模電算機
通信用固定ノード

【課題1】との連携

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*, 29 (8):1282-1291, 2021.

8. T. Tanjo, Y. Kawai, K. Tokunaga, O. Ogasawara and M. Nagasaki. Practical guide for managing large-scale human genome data in research. *Journal of Human Genetics*, 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

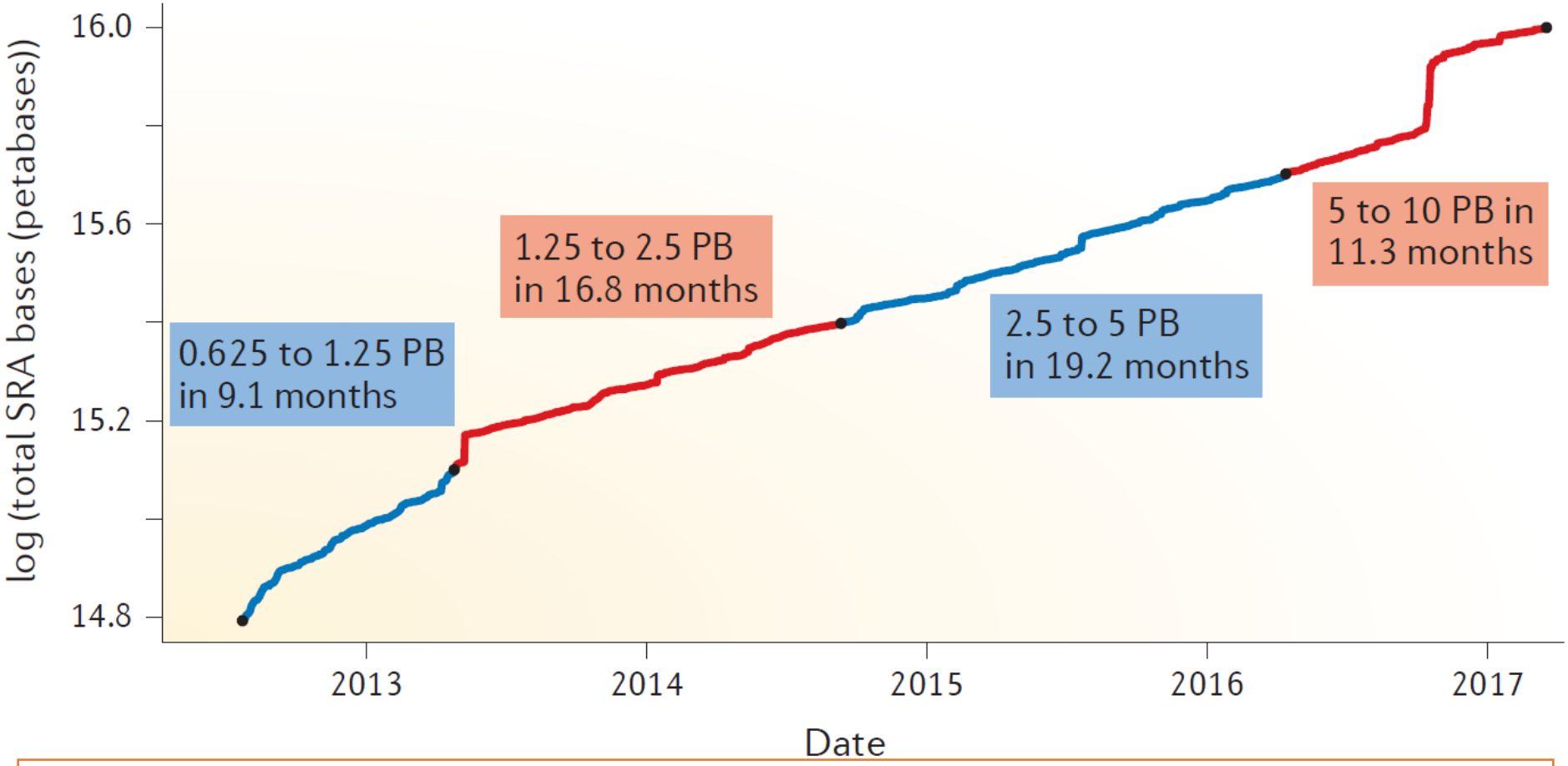
3. 長﨑 正朗, “Accelerating the pace of research in Kyoto University”, AWS Public Sector Summit Online, 2021/4/15-2021/4/16.

4. 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, 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, 丹生 智也, 小笠原 理, 山田 亮, 松田 文彦, 徳永 勝士, “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”, 第44回日本分子生物学会年会, 2021/12/1-2021/12/3.

5. 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, “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”, 日本人類遺伝学会第66回大会, 2021/10/13-2021/10/16.

6. 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, “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”, 第41回医療情報学連合大会, 2021/11/18-2021/11/21.

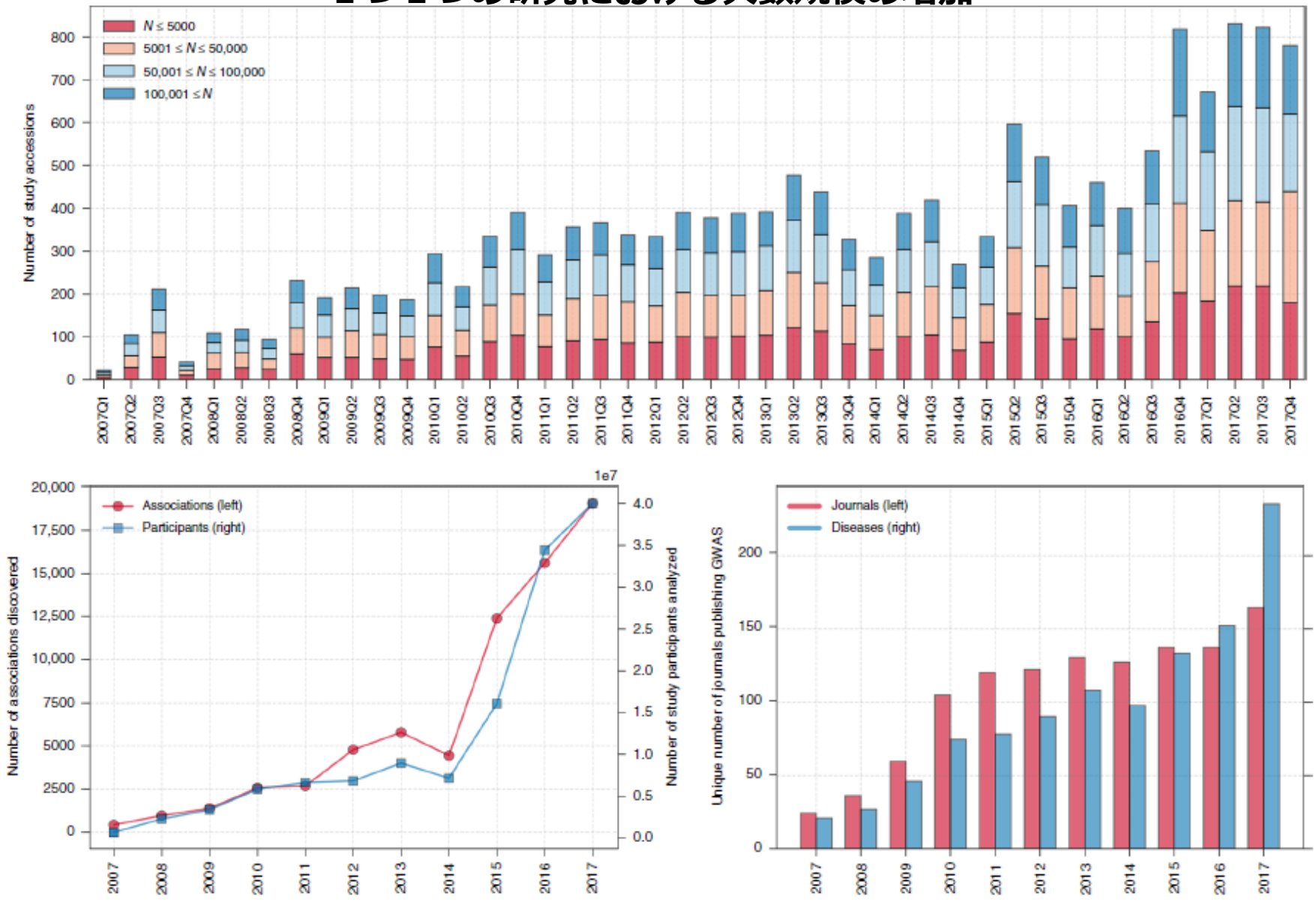
ヒトゲノムシークエンス情報の増加の様子



【図4】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5】 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