aws

SUMMIT ONLINE

JAPAN

京都大学
KYOTO UNIVERSITY

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

長﨑 正朗 (Masao Nagasaki)
特定教授　nagasaki@genome.med.kyoto-u.ac.jp

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

aws SUMMIT ONLINE

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

本セッションの概要

- ヒト全ゲノムのシークエンス技術と国内外の状況について
- 京都大学大学院医学系研究科附属ゲノム医学センターの取り組み
- スーパーコンピュータとクラウドとのハイブリッド運用について
  - → 要件およびAWSを選択した理由
  - → システム構成
  - → 解析の流れ
- まとめと今後の予定について

ヒト全ゲノムのシークエンス技術
と
国内外の状況について

ヒト全ゲノムのシークエンス技術
概要

① ヒトゲノム全長の情報

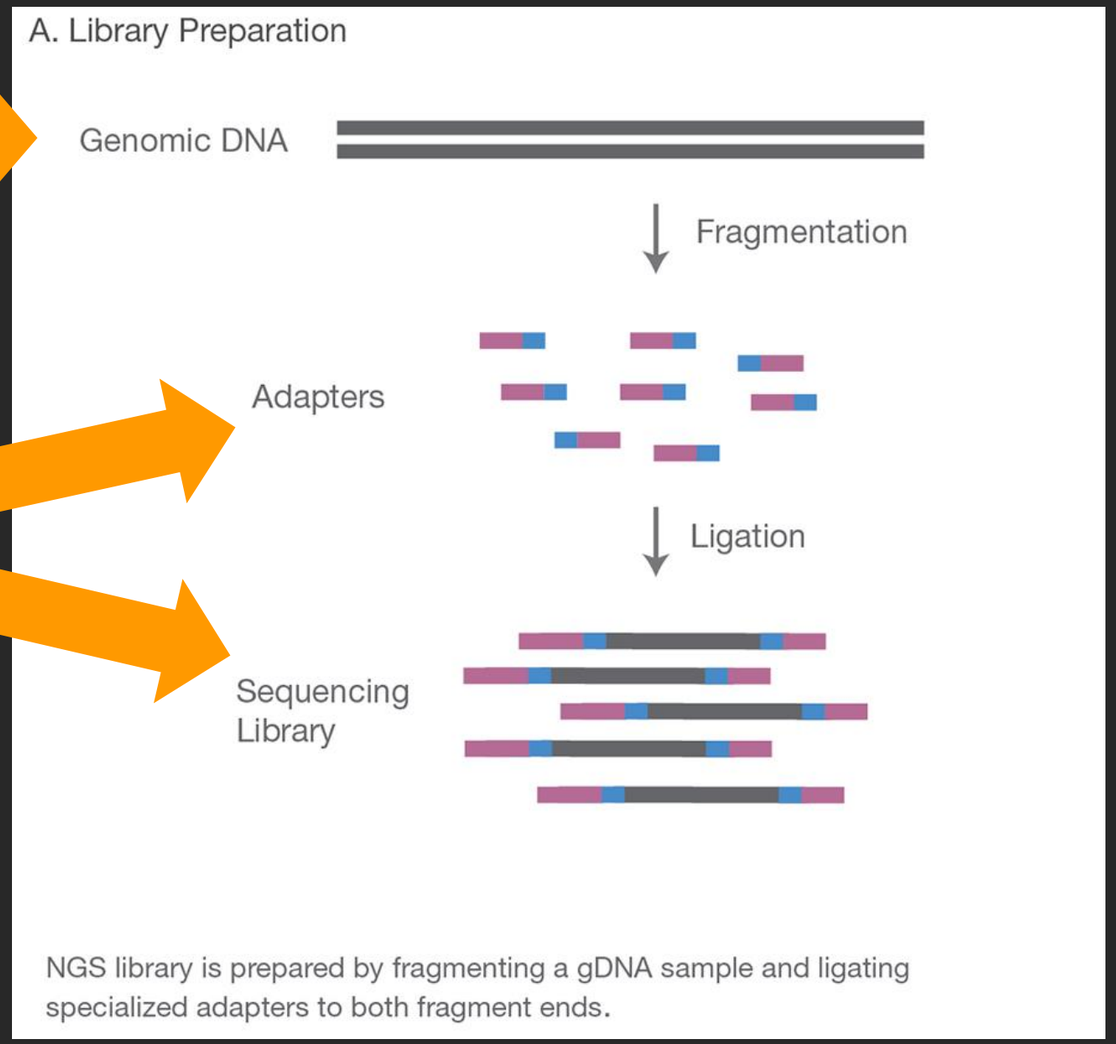
31億文字 (A/T/G/C) の塩基対情報からなる

② シークエンサで行うこと

約350-500塩基に断片化されたヒトゲノム情報の読み取り※1

ヒトゲノムの場合、大体ヒトゲノム全長を30回被覆できるぐらい読む

→ 一人当たり930億文字の情報



Courtesy of Illumina, Inc.

※1 後述のIllumina Inc. MGI社の場合

ヒトゲノムの全ゲノムシークエンス
最新機種イルミナ社 NovaSeq 6000の場合

S4フローセル

2日のオペレーション

3000G 塩基 (bases)
の出力

計算機での解析へ

※2 x 150プロトコルの場合

ヒトゲノムの全ゲノムシークエンス
最新機種イルミナ社 NovaSeq 6000の場合

1機器年間 情報出力量	(3000 Giga bases x 365日)/ 2日 =**547.5 Tera bases** 1塩基はATGCの4文字のいずれか(つまり2bitで表現可能)、ただし１つ１つの塩基に信頼度や１つ１つのリードの名前などが付随する。
ヒト一人当たり に必要情報量	ヒト全ゲノム長 ≒ 3.1 億 塩基 (bases) =3.1 Giga bases 総被覆回数 30xなので **93 Giga bases**
1機器年間 出力人数	(32人 x 365日) / 2日 =**5,840人**

【参考】2000年あたり、10年かけて全世界の研究者が１つのヒト参照ゲノムを構築（数十人のゲノムを合成したもの。半分以上は1人に由来）

ヒトゲノムシークエンス情報の増加の様子

2013年～2018年

米国の国際シークエンス情報登録データベース (Sequence Read Archive: SRA) におけるゲノム情報の推移

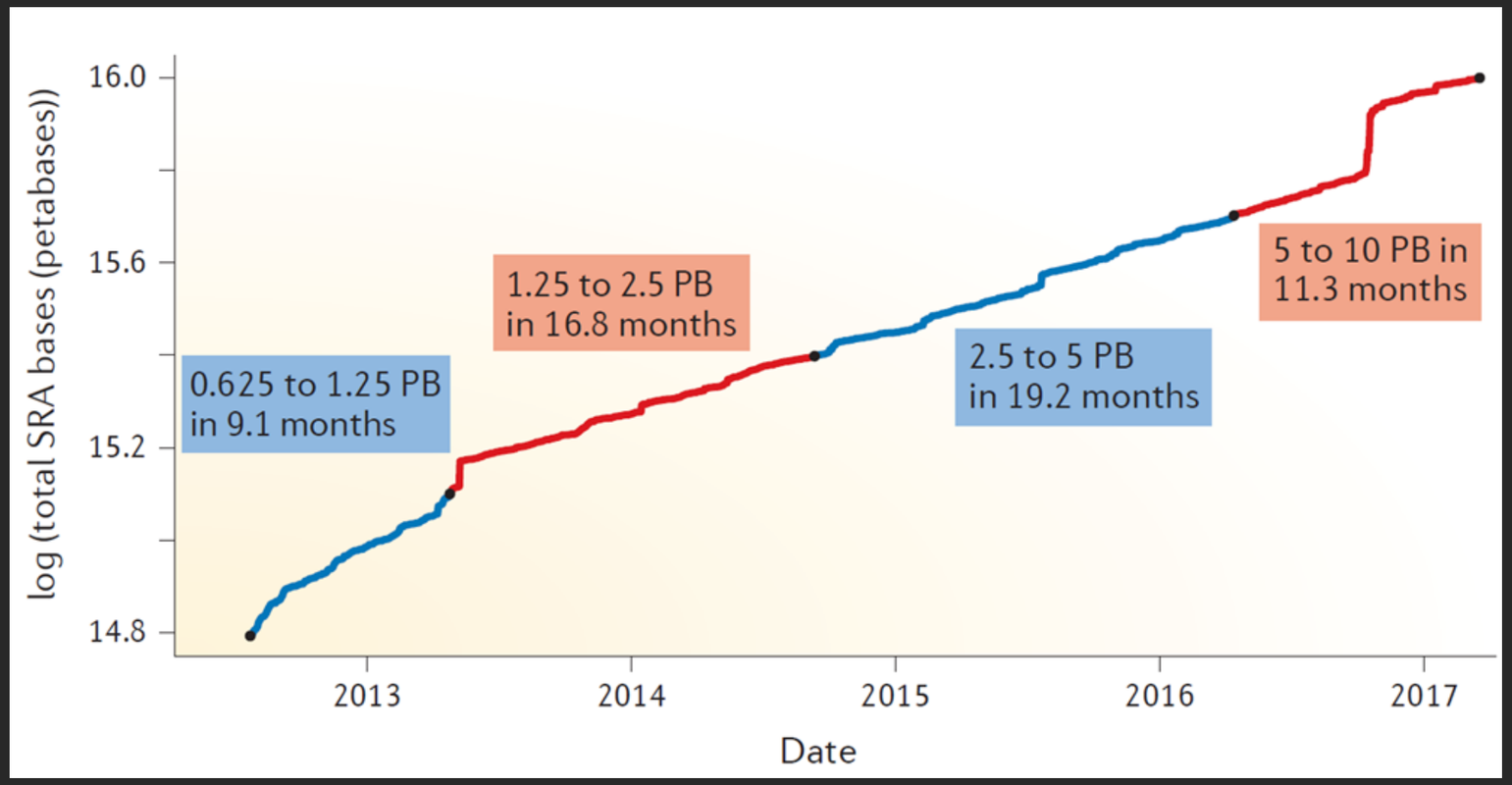


Figure 1 in Cloud computing for Genomic Data Analysis and Collaboration Ben *et al.* *Nature Rev Genet* 2018

2020年

12 Peta Byteに到達

ヒトゲノム研究における解析対象人数の増加

研究プロジェクト毎の解析対象人数 (下図での*N*) が大規模化するとともにまた研究数が増えていることがわかる（※）

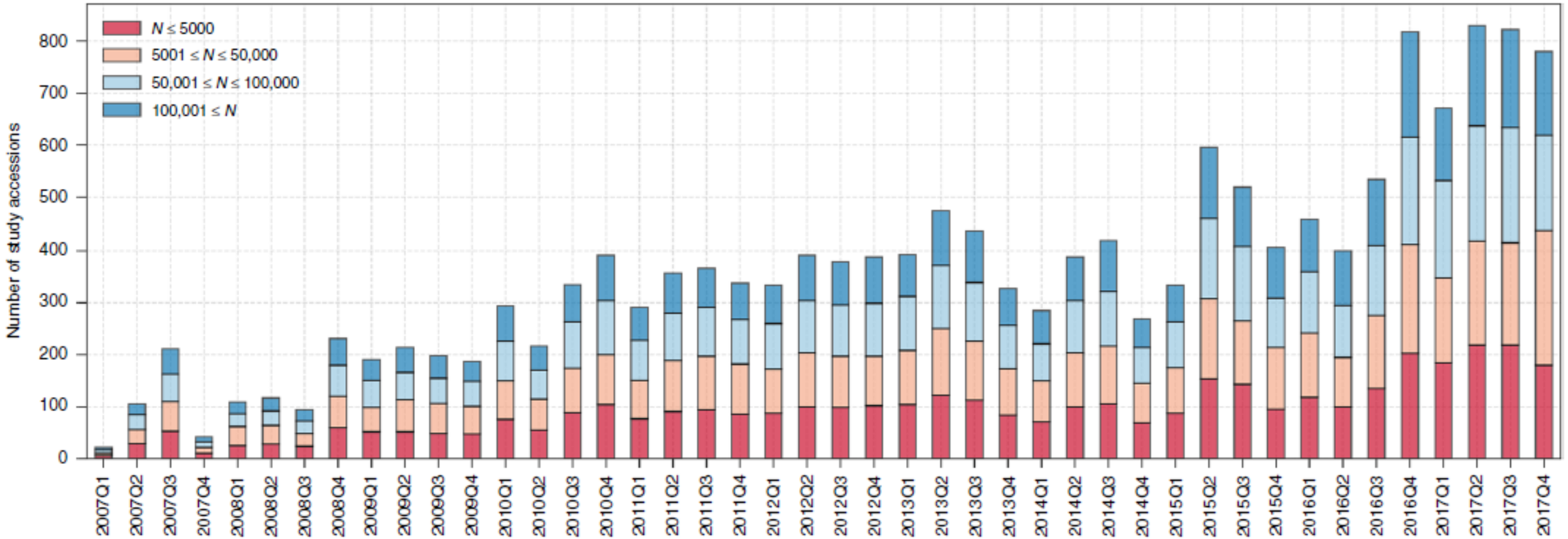


Figure 1 in a scientometric review of genome-wide association studies Melinda *et al Communication Biology* 2019

※ここでの研究は全ゲノムを対象とした研究でシークエンスを対象としていないものも含まれる。

京都大学大学院医学系研究科
附属ゲノム医学センターの取り組み

京都大学
KYOTO UNIVERSITY

ゲノム医学センターの取り組み

ヒトゲノム情報などのライフサイエンス分野における超大規模ヘテロデータに対し、情報科学・数理統計学の知識を基に大規模計算資源を活用して解析を行い、新たな知識を発見するための研究を行っています。

特に、多因子型遺伝病や癌を解析対象とすることで、次世代の疾患解析モデルの構築、予防、治療へなど医学への還元を目指した研究を進めています。

関連して長浜0次コホート調査事業や難病プラットフォーム事業を行っています。

ながはま0次コホート事業

ながはま0次コホート事業

ながはま0次予防コホート事業では、長浜市民１万人から集めた様々な健康情報、血液や尿の成分、環境・生活習慣の情報などを統合して解析することで、病気の原因や老化のメカニズムを解明し、「医学の発展」と「市民の健康づくり」に貢献することを目指しています。
従来、病気の予防は１次予防（病気にならないように気をつける）、２次予防（病気がひどくなる前に見つける）、３次予防（かかってしまった病気の悪化を防ぐ）の３段階で考えられていました。しかし、科学技術の発展によって、ヒトの体質（ゲノム）やごく僅かな血液成分の変化も測定できるようになったことから、病気になる遙か手前で先手を打って予防することも夢ではなくなりつつあります。このような従来の１次予防に先んじる予防を「０次予防」と名付け、その実現を目指し、2007年に長浜市長と京都大学医学研究科長が「０次予防健康づくり推進事業の共同実施に関する協定書」を締結しました。京都大学大学院医学研究科と長浜市、そして、長浜市民とが、産官学民連携で進めています。

ゲノム情報を含むコホート調査事業

ながはま０次コホート事業は 長浜市民 10,000 人 が参加する
０次予防実現のための 最先端研究事業 です。

http://zeroji-cohort.com/project/index.html 研究代表　京都大学医学系研究科附属ゲノム医学センター長　松田文彦

難病プラットフォーム事業
Rare Disease Data Resigtory of Japan

RADDAR-J

難病プラットフォーム

Rare Disease Data Registry of Japan

- TOP
- 難病プラットフォームとは
- データシェアリングについて
- 各種支援
- 中央倫理審査
- 研究形態
- レジストリ・バイオレポジトリ一覧
- 研究体制
- 連携研究班一覧

統括的情報共有と連携。
難病の研究と治療方法の開発を推進します。

難病プラットフォームは、日本医療研究開発機構（AMED）および厚生労働省の難病研究班が収集した臨床情報や生体試料から得られた情報を集約する情報統合基盤です。集約したデータのシェアリングや二次利用を通じて、わが国の難病研究を推進させることを目指します。
難病プラットフォームは、AMEDの難治性疾患実用化研究事業の研究費により運営されています。

日本医療研究開発機構（AMED）の「難治性疾患実用化研究事業」のページはこちら

https://www.raddarj.org/　研究代表　京都大学医学系研究科附属ゲノム医学センター　センター長　松田文彦

難病プラットフォームによる
難病研究データのシェアリングの促進

※AMEDが指定する公的データベースとして認定
https://www.amed.go.jp/koubo/datasharing.html

難治性疾患政策研究事業（厚労省）

約88の研究班を支援
- 診療体制構築、疫学研究、普及啓発
- 診断基準・診療ガイドラインの作成・改訂
- 小児成人期移行医療の推進
- 関連研究やデータベース等との連携
- 指定難病医療費助成のための臨床情報（臨床調査個人票）を収集　等

難治性疾患実用化研究事業（AMED）

約200の研究班を支援
- 病態解明・バイオマーカーの開発
- エビデンスの創出
- ゲノム・オミックス解析
- 未診断疾患イニシアチブ（IRUD）
- 新規治療薬・医療機器等の開発
- 医師主導治験　等

集約（制限共有）

臨床情報 | 生体試料情報 | ゲノム情報

難病プラットフォーム※

集約した情報の二次利用やレジストリ研究の促進
（制限共有、制限公開、非制限公開）

数千人の全ゲノム情報

数万人の全ゲノム情報

国内のゲノム情報の取り巻く状況

コホート調査事業

すでに数千人規模の全ゲノム情報

↓

１万人規模の全ゲノム情報

難病研究事業

数千人の全ゲノム情報

↓

数万人の全ゲノム情報

- UK/USですでに十万人以上　　UKなどですでに数万人

https://www.genomicsengland.co.uk/ TOPMed https://www.nhlbiwgs.org/

スーパーコンピュータとクラウドとのハイブリッド運用について

→ 要件およびAWSを選択した理由

→ システム構成

→ 解析の流れ

ハイブリッドクラウド構築における要件

1. オンプレ、スパコン、クラウド各拠点のそれぞれの優位性の活用
2. ゲノム情報の大規模情報化における解析インフラ環境の拡張性
3. 解析環境の新規導入時の水平展開の容易性
4. 多様な種類のデータベースや解析ツールの導入から利用までのコスト（人的資源・費用・期間）の低減
5. 国際共同学位のための海外の学生と教員が利用できる環境の用意
6. ヒト全ゲノム情報の研究グループ間の高度な情報共有システム構築

AWSを選択した理由

1. 学術情報ネットワーク SINET5 への接続および学術での運用実績があること。
2. クラウドの老舗ならではの豊富な技術情報、ホワイトペーパー等、開発や運用のノウハウの得やすさがあること。
3. AWSの豊富なセキュリティ機能や第三者認証（ISO/IEC 27017等）と多数の利用実績があること。
4. AWS活用による品質向上（可用性、運用保守性、性能拡張性等）、コスト低減、納期短縮の将来的な発展性があること。
5. AWSプロフェッショナルサービス活用による技術要件の時間内の解決能力を持つこと。

スーパーコンピュータとクラウドとのハイブリッド運用について

→ 要件およびAWSを選択した理由

→ システム構成

→ 解析の流れ

システム構成概要

AWS

CGM security Amazon VPC

SINET5

L2VPN (Direct Connect) 10G

京都大学

ゲノム医学センター (CGM)　オンプレ
高速ストレージ:　数ペタ (GPFS)
計算ノード　512 Gb memory /72～
120 cores x 数ノード

情報環境機構　スパコン
高速ストレージ: DDN ExaScaler
- 24ペタバイト
計算ノード: Cray CS400 2820XT
- 36 cores x 850 ノード (Xeon Broadwell / 128 Gb memory)
の一部を利用

SINET5

東京大学

情報基盤センター
仮想環境: Fennel
を利用

情報基盤センター　スパコン
高速ストレージ: DDN ExaScaler
- 12ペタバイト
計算ノード: PRIMERGY CX2550 M5
- 56 cores x 1240ノード (Xeon Platium 8280 / 192 Gb memory)
の一部を利用

ハイブリッドクラウドシステムの大枠の役割（１）

AWS　CGM security Amazon VPC

京都大

ゲノム医

【環境１】
（役割）
ヒトゲノム情報解析で**より汎用的な解析**が求められる解析パイプランの実行
（長所）
・OSや内部のバージョンの依存関係が複雑なため【環境2】では実現困難な解析パイプラインを容易に実施できる
・【環境2】が混んでいる場合に一時的な大規模な解析を実行できる
・すでに商用サービスとしてイメージが準備されておりすぐ使える

情報環境機構

情報基盤センター

【環境２】
（役割）
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ンの実行
（長所）
・計算リソース内で実現可能な定型解析を高速に実行できる

ハイブリッドクラウドシステムの大枠の役割（２）

AWS　CGM security Amazon VPC

【環境１】
（役割）ヒトゲノム情報解析で**より汎用的な解析**が求められる解析パイプランの実行
（短所）
・計算ノード / ストレージのコストが【環境２】に比べて高い

京都大

ゲノム医学センター (CGM)オンプレ

情報基盤センター

SINET5

情報環境機構

情報基盤センター

【環境２】
（役割）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ンの実行
（短所）
・東大の場合は実行時間の上限が2日、ノード毎のジョブ実行なので計算効率がソフトウェアによってはパフォーマンスがでない
・160Gメモリ上限があり一部のソフトウェアが実行できない
・混んでいるとジョブがすぐには入らない

CGMハイブリッドクラウドシステムの実装（1）

バッチジョブシステム

全世界のTOP500 supercomputers の 半数以上で利用

コンテナの利用による再現性・再利用性・信頼性の担保

Singularity v3を用いたパイプライン構築

CGMハイブリッドクラウドシステムの実装（2）

AWS

CGM security Amazon VPC

SINET5

【環境１】一度作成した解析パイプラインはどこでも実行可能

Amazon Parallel Cluster

Amazon FSx for Lustre

での実現

京都大学

ゲノム医学センター (CGM)オンプレ

東京大学

情報基盤センタ

SINET5

情報環境機構

情報基盤センター

【環境２】計算ジョブの実装コストはかかるがパイプライン内の各実行コンポーネントは【環境１】をそのまま利用可能

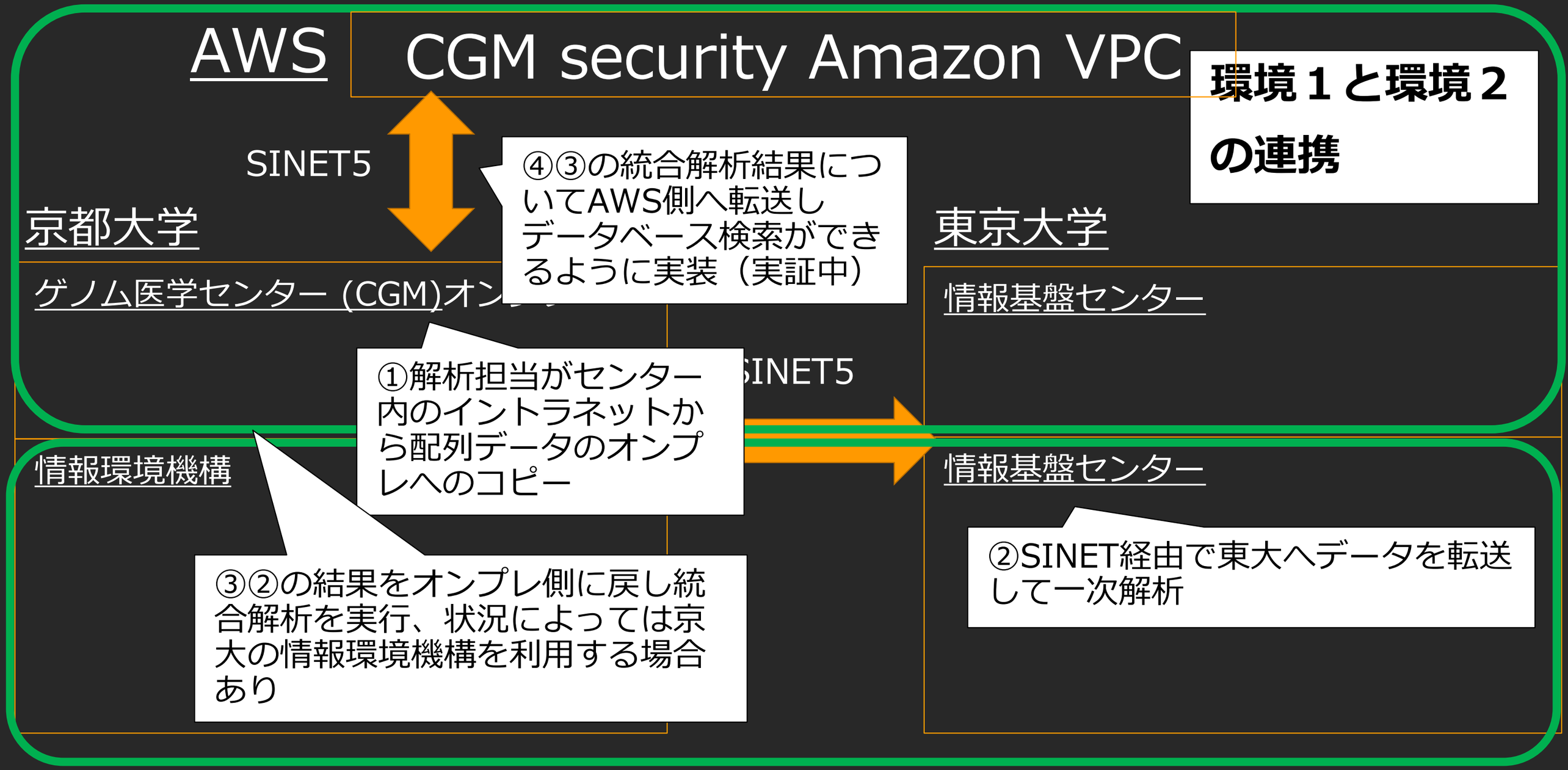
スーパーコンピュータとクラウドとのハイブリッド運用について

→ 要件およびAWSを選択した理由

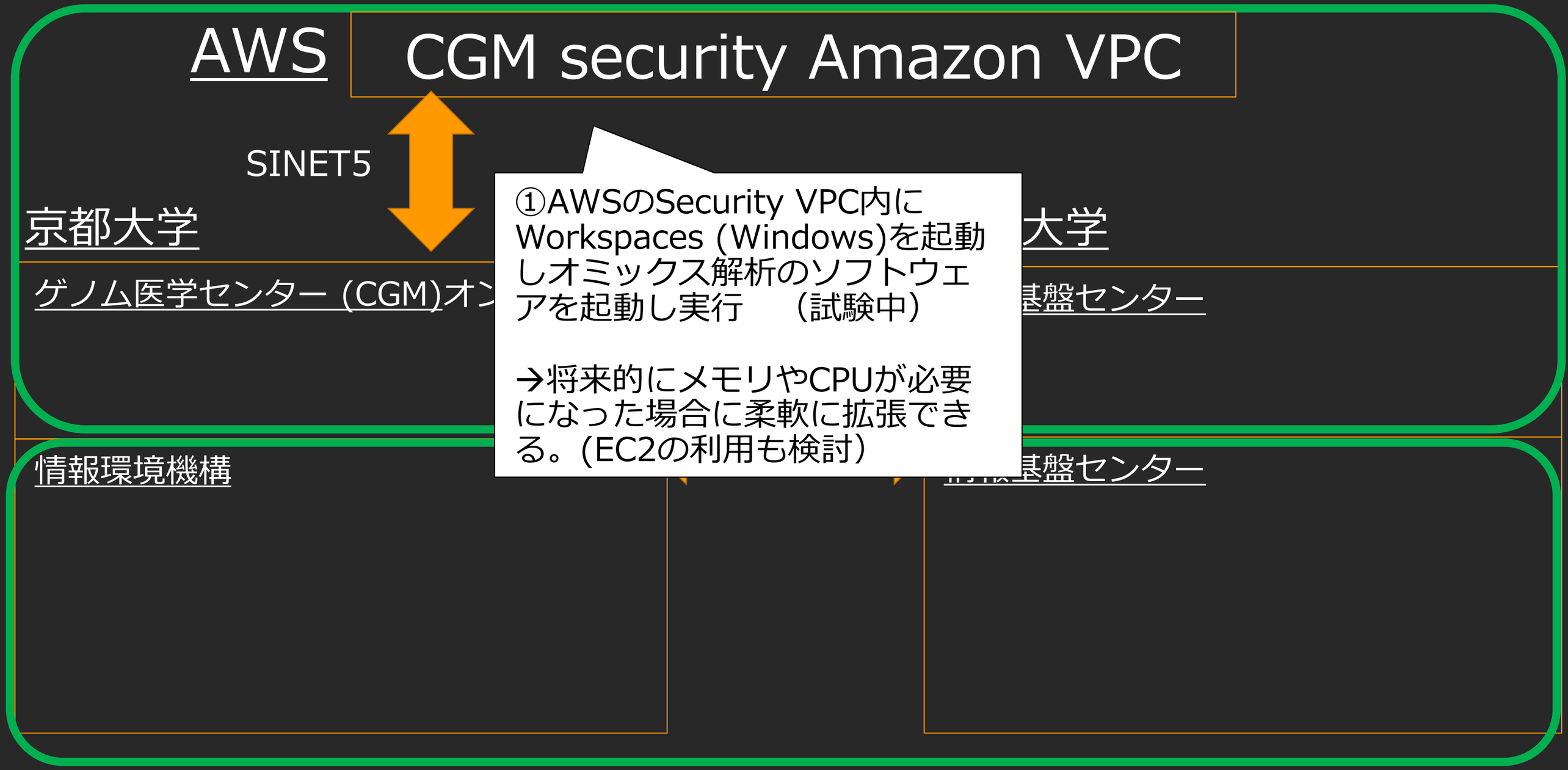
→ システム構成

→ 解析の流れ

全ゲノム解析の場合のデータフロー



オミックス解析の場合のデータフローの一例

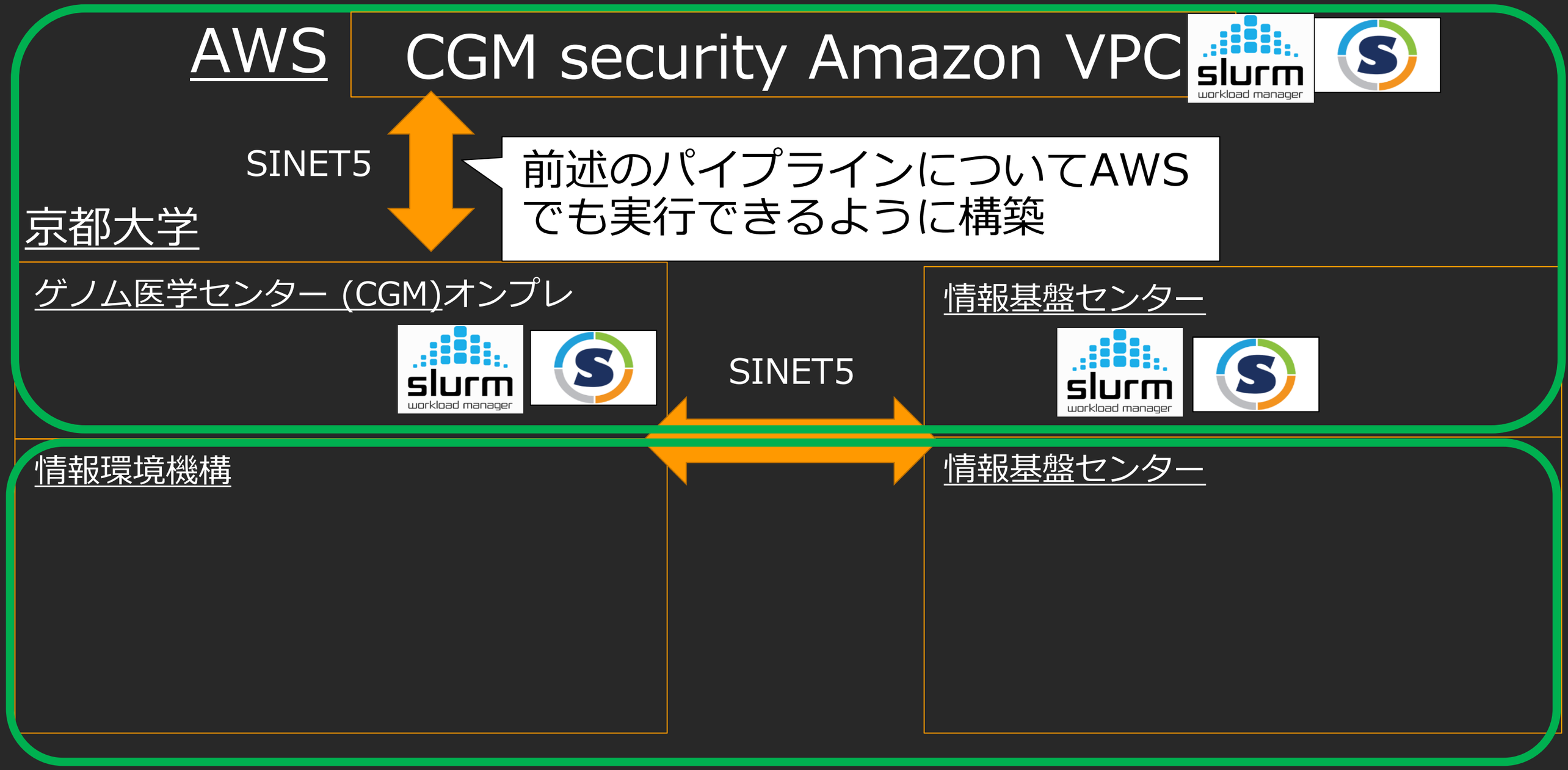


ヒトゲノム情報解析で<u>より汎用的な解析</u>が求められる解析パイプライン

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） 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によっててで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

SNPアレイ解析の場合のデータフロー



利用中の主なサービス群

Storage

Amazon S3
Amazon S3 Glacier
Amazon FSx for Lustre

Compute resource

Amazon EC2
(含むGPU)

Database

Amazon Elasticsearch Service
MarkLogic

Batch System

AWS ParallelCluster
AWS Batch

Network

AWS Direct Connect

Log & Security

Amazon CloudWatch
AWS CloudTrail

GUI Frontend

Amazon Workspaces
X2Go

ドラフト実装後の最終運用においてプロフェッショナルサービスの多大な協力

まとめと今後の予定

1. ヒトゲノム情報の機器のイノベーションによりデータ量が膨大となっている。
2. 各研究において対象とする人数規模も増加している。
3. ゲノム医学センターにおいても大規模な解析が必要となっている。
4. ハイブリッドクラウドの構築によりそれぞれの長所を活かした解析を行うことを目標としている。
5. できるだけ同一環境とするためのいくつかのキーテクノロジーを考慮した実装としている。
6. AWSにより、オンプレやスパコンのみでは実現できない柔軟さを得ることができた。
7. すでに一部の実践の解析においてその柔軟性や機動性が役立ち始めている。
8. AWSの導入によるハイブリッドクラウドの優位性を活かした最適配置。
9. AWS ParallelCluster/ Amazon FSx for Lustre を活かした最小限のコストでのオンデマンドHPC環境の積極的な活用。