

本件の情報公開は、既に解禁されています。

2024年12月26日

## 国産コンピュータ・プロセッサで実現した ヒトゲノム解析の革新的な高速化と制約無しの完全公開

近年、ゲノムの読み取り技術は急速に進展を遂げていますが、ゲノム情報を希少疾患の原因究明やがん医療へ応用するには、数万人規模のヒトゲノム解析が不可欠です。これには数ペタバイト<sup>(1)</sup>に及ぶ大量のデータを処理する高度な計算技術が求められていました。

国立遺伝学研究所 生命情報・DDBJ センター(センター長 有田正規)と株式会社 PEZY Computing(代表取締役社長 高橋一夫)、株式会社 ゲノムアナリティクスジャパン(代表取締役 八谷剛史)は共同で、PEZY Computing社が開発した国産メニーコアプロセッサPEZY-SC3<sup>(2)</sup>を搭載した高速ヒトゲノム解析システム「ZettaVEGA」の実証実験を行いました。ZettaVEGAではヒトゲノム解析で最もよく使用されるソフトウェアGATKのベスト・プラクティス<sup>(3)</sup>をPEZY-SC3に移植し、メニーコアを最大限まで活かした超並列計算による高速化を実現しています。

まず、国際的によく利用される参照ゲノム配列GRCh38<sup>(4)</sup>に精度検証用のヒトゲノムデータ<sup>(5)</sup>を照合する解析をZettaVEGAで実施し、GATKベスト・プラクティスの実装と結果が99.9995%一致することを確認しました。GATKベスト・プラクティスの実装は、CPU16コアを用いても一人分のヒトゲノム解析に1日以上 of 計算時間を要します。一方で、ZettaVEGA では33分で計算が完了しました。これはNVIDIA社のGPUを用いた計算時間(H100 8枚を用いた場合に37分、V100 4枚では90分)よりも高速で、PEZY-SC3プロセッサの高い性能をよく表しています。

次に、GRCh38よりも新しく、完成度が高い参照ゲノム配列CHM13<sup>(4)</sup>に、4つの国際コンソーシアムが公開するデータセットを照合する解析を実施しました(表1)。CHM13との照合は更に大きな計算コストがかかりますが、DDBJ<sup>(6)</sup>とゲノムアナリティクスジャパンは照合結果をインターネットで制限を設けずに公開<sup>(7)</sup>し、誰もが利用できるようにしています。

表1 ZettaVEGA(PEZY-SC3)を用いた公開データセットの実行結果

| コンソーシアム         | サンプル数 | 合計計算時間<br>(一人あたり平均 [分]) | 合計データ量 [TB]<br>(一人あたり平均 [GB]) |
|-----------------|-------|-------------------------|-------------------------------|
| 1000 人ゲノムプロジェクト | 2,504 | 1,651 時間<br>(39.56 分)   | 123 TB<br>(49 GB)             |
| ヒトゲノム多様性プロジェクト  | 689   | 589 時間<br>(51.25 分)     | 54 TB<br>(78 GB)              |
| 合計              | 3,193 | 2,240 時間<br>(42.09 分)   | 177 TB<br>(55 GB)             |

## まとめ

国立遺伝学研究所スーパーコンピュータの既存システム（NVIDIA社のV100 GPU 4枚）と比べ、ZettaVEGAを用いた計算速度は2.8倍になりました。ZettaVEGA 1台あたり、年間で約1.2万人のゲノムデータを解析できるペースです。国立遺伝学研究所ではZettaVEGAを2台導入しており、成果を公開・共有することで国内に不足する計算資源の補完に貢献しています。

## ■ 用語解説

### (1) ペタバイト

ペタとは、ギガ(10億)の100万倍にあたります。ヒト全ゲノムの情報を1万人分管理するのに、おおよそ1ペタバイトのストレージを要します。

### (2) PEZY-SC3

PEZY-SC3プロセッサはHPC(High Performance Computing)用の高速演算コアを4,096個搭載した、超並列計算向けの高性能かつ省電力 IC チップです。

製品の解説 <https://www.pezy.co.jp/products/pezy-sc3/>

ZettaVEGAの解説 [https://www.pezy.co.jp/wp-content/uploads/2024/12/zettavega\\_whitepaper-20241202.pdf](https://www.pezy.co.jp/wp-content/uploads/2024/12/zettavega_whitepaper-20241202.pdf)

### (3) GATKベスト・プラクティス

GATK (Genome Analysis Tool Kit) は米国ブロード研究所が開発したヒトゲノム解析用のソフトウェア群です。複数の処理結果を比較可能にするために計算の処理過程を揃えたものをベスト・プラクティスと呼び、世界中で利用されています。

### (4) 参照ゲノム配列 GRCh38・CHM13

ゲノム解析の際に参考とするヒトゲノムの全染色体情報を参照ゲノム(リファレンスゲノム)といいます。基準ゲノム配列とよばれることもあります。GRCh38とは国際的な学術組織が公開する参照配列で、最新バージョン38は2013年より公開されています(特定個人の情報ではありません)。CHM13は胞状奇胎由来の1倍体完全ゲノムで、2022年に発表・公開されました。

### (5) 精度検証用のヒトゲノムデータ

米国食品医薬品局(FDA)が提供するヒトゲノムデータ PrecisionFDA HG001を、GATKベスト・プラクティスの実装(16コアのCPUを使用)とZettaVEGA(PEZY-SC3を使用)を用いて解析し、結果の一致度を評価しました。

### (6) DDBJ (DNA Data Bank of Japan)

国際的に主要な遺伝子情報のリポジトリ(投稿形式のデータベース)で、生命科学の研究者がゲノムデータを共有・公開するために利用しています。

センターウェブサイト <https://www.ddbj.nig.ac.jp/index.html>

今回の公開データサイト <https://ddbj.nig.ac.jp/public/public-human-genomes/>

### (7) ヒトゲノムデータの解析結果の公開

ヒトゲノムデータへのアクセス管理は「制限なし公開」と「制限あり公開」の2通りあります。今回の結果は、制限なし公開データセットを解析したものです。

## ■ 問い合わせ先

<事業活動に関すること>

- 情報・システム研究機構 国立遺伝学研究所 生命情報・DDBJセンター  
センター長 有田 正規（ありた まさのり）

<報道担当>

- 情報・システム研究機構 国立遺伝学研究所 広報室

配付先

文部科学記者会、科学記者会、三島記者クラブ