# ZettaVEGA: ZettaScaler Verifying Environment for Genome Analysis

Accelerating genomic data analysis with PEZY-SC3

## 概要

ZettaVEGAは、PEZY-SC3を四基搭載したZettaScaler-3.0システム上で動作する、PEZY Computing社製の高速なゲノム解析ソフトウェアです。ZettaScaler-3.0は、ストレージ、ネットワーク、演算性能といったゲノム解析に必要なITリソースを、コンパクトでスケーラブルなソリューションに統合しています。
ZettaVEGAは、四基のPEZY-SC3を使用することで、33xカバレッジのヒト全ゲノムシークエンスのデータを最大96検体/日処理することができます。ZettaVEGAは、デファクトスタンダードとなっているGATK Best Practiceパイプラインと比較して100倍以上高速に動作しますが結果の互換性は99.99%以上保証しています。ZettaVEGAが使用しているソフトウェアはGATK Bset Practiceパイプラインを構成するソフトウェアと互換になるように設計・開発されています。ZettaVEGAのFastqを入力しVCFを得るためのパイプラインはシンプルなインターフェースを備えており、ユーザーはZettaVEGAの各ソフトウェアを従来のソフトウェアと同様の感覚で使用することが可能なため、従来のゲノム解析パイプラインからZettaVEGAへのスムーズな移行が可能です。またZettaVEGAでは、GATK4.2で導入されたFRDやBQDなどの高精度な確率モデルや、アライメントにおけるALT-contigからのLiftover機能や、チューニングされたパラメータのプリセットを利用することができ、高感度と低い偽陽性を両立した結果を得ることが可能になります。

本ドキュメントでは、ZettaVEGAのアーキテクチャとパフォーマンス・精度について解説します。

## 目次

# イントロダクション

2003年にヒトゲノムの解読が完了して以降、ヒトゲノムを解析しようとする試みは20年にわたり行われてきました。近年ではゲノムシーケンシングのコストがムーアの法則を凌ぐ勢いで下がり、個人の全ヒトゲノムシーケンシングにかかるコストが$1000以下となってきています。それに伴い、プレシジョン・メディシンという、遺伝情報、個々人の生活環境やライフスタイルの差異を考慮し、疾病予防や治療を行う、新しい医療の考え方も現実的なものとなってきました。また、次世代シーケンサによるゲノムシーケンシングの速度も高速化の一途を辿っており、高速なゲノム解析エンジンの重要性は非常に高まっています。

そこで、PEZY Computingでは、PEZY-SC3アクセラレータを使用し、高速なゲノム解析エンジンであるZettaVEGAを開発しました。

## ZettaVEGA

ZettaVEGAは、ZettaScaler-3.0システム上で動作する、高速なゲノム解析ソフトウェアです。

### ZettaScaler-3.0

ZettaScaler-3.0は、PEZY Computingが開発したサーバーです。最大の特徴として、独自開発した汎用アクセラレータ、PEZY-SC3を四基搭載していることが挙げられます。PEZY-SC3の理論ピーク性能は倍精度で19.66TFlopsであり、システム全体としては78.64TFlopsとなります。また、ホストCPUにはEPYC-7713Pを採用しており、PEZY-SC3で高速化されていない従来のツールを使用する場合にも高速に処理を行うことができます。外部ネットワークインタフェースは10G EthernetまたはInfiniband EDRをサポートし、ネットワーク上のファイルサーバーなどとの高速なネットワーク通信も可能です。
ZettaScaler-3.0では、ゲノム解析に必要な、リファレンスデータファイルやfastqファイル、またvcfデータなどの大容量フ

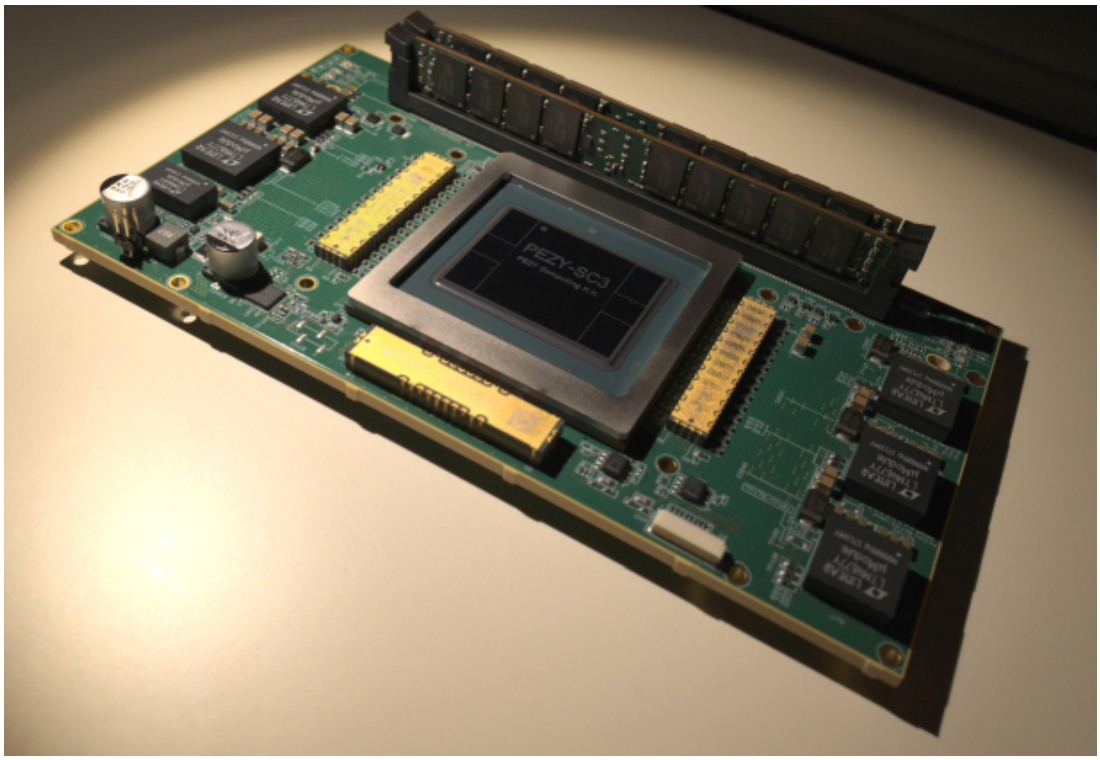
ァイルを高速に読み書きするために、四基のPCI Express接続のNVMeストレージを搭載し、これらをRAID0またはRAID01で構成しています。これにより、ゲノム解析で必要になる大容量データを高速に読み書きすることが可能です。

以下にZettaScaler-3.0の諸元を示します。

| | **Spec** |
|---|---|
| CPU | EPYC-7713P |
| RAM | 1TB(or 2TB) |
| SSD | 256GB(OS, SATA) |
| | 8TB(Data, NVMe) |
| Network | 10G Ether |
| | Infiniband EDR(100Gbps) |
| Accelerator | PEZY-SC3 * 4 |

## PEZY-SC3

PEZY-SC3はPEZY Computingが開発した汎用アクセラレータカードです。PEZY-SCシリーズとしては第三世代の製品となります。高いピーク性能と面積効率・電力効率を両立するようにデザインされており、非常に高い演算性能と非常に高い電力効率を誇ります。理論ピーク性能は倍精度で19.66TFlops, 単精度で39.32TFlops, 半精度で78.64TFlopsとなります。また、メモリとしてHBM2を採用しており、1.2TB/sと非常に高速なメモリ帯域を持ちます。ホストとのインタフェースはPCIExpress Gen4 x16を採用し、双方向32GB/sの帯域を持ちます。消費電力はモジュールあたり600Wとなっています。

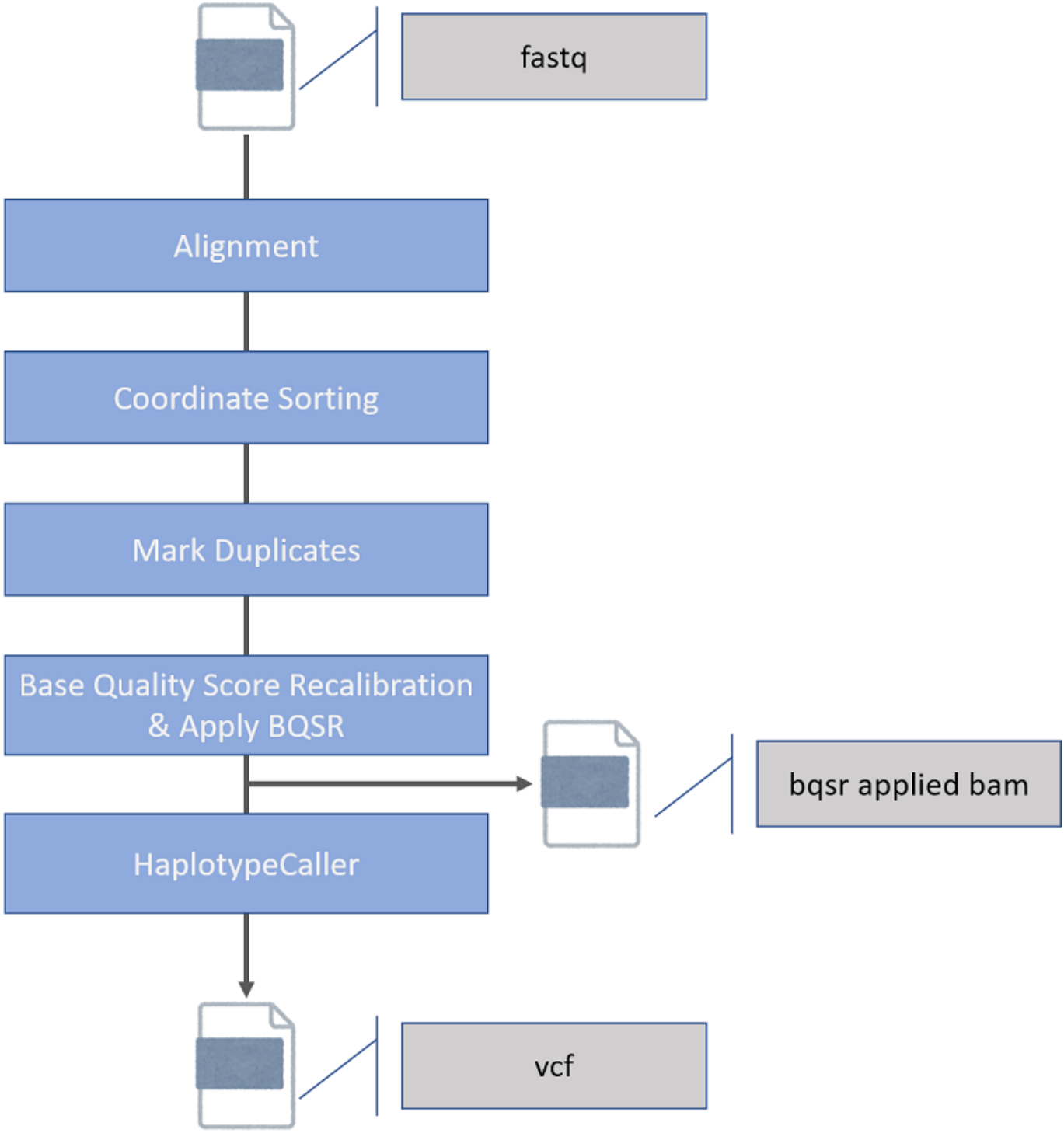



以下にPEZY-SC3の諸元を示します。

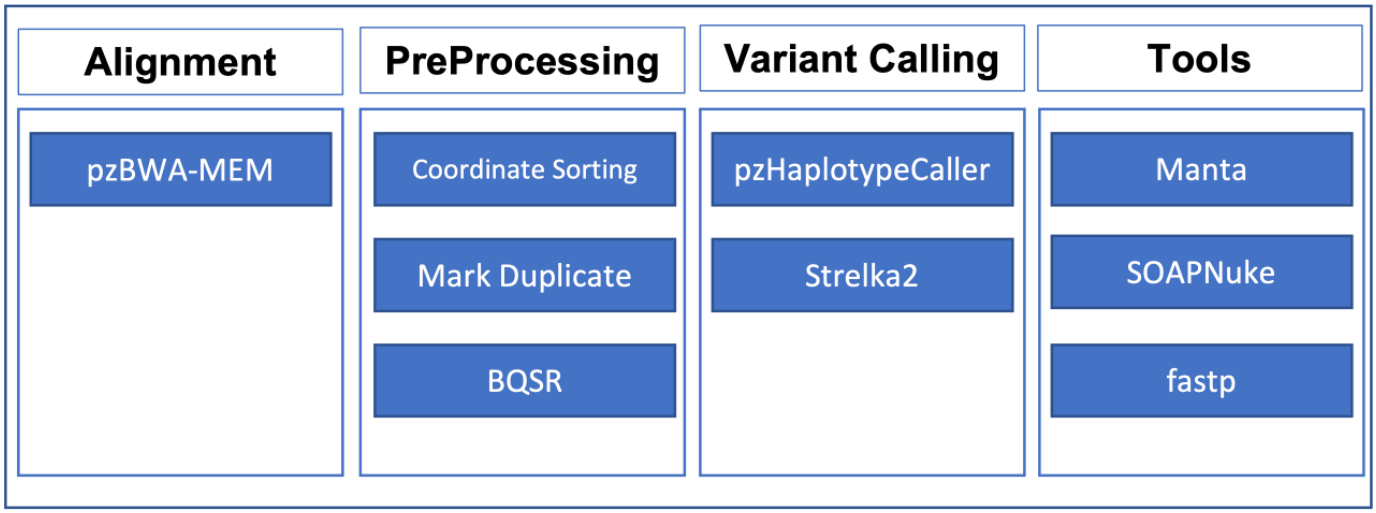
| | **PEZY-SC3** |
|---|---|
| Number of PEs | 4096 |
| Clock(PE / GHz) | 1.2 |
| Clock(BUS / GHz) | 1 |
| Peak FP64 Flops(TFlops) | 19.66 |
| Peak FP32 Flops(TFlops) | 39.32 |
| Peak FP16 Flops (TFlops) | 78.64 |
| Memory Type | HBM2 |
| Memory Size(GB) | 32 |
| Peak Memory bandwidth(GB/s) | 1200 |
| Power Consumption | 600W |

# ZettaVEGA ソフトウェア概要

ZettaVEGAでは、以下のヒトゲノム解析パイプラインを提供しています。



また、ZettaVEGAを構成しているソフトウェアは以下の通りです。




以下に、主要なソフトウェアを解説します。

### pzBWA-MEM

pzBWA-MEM は、BWA-MEM version 0.7.17 (r1198) をベースに、PEZY Computingによる改良を加えた、高速なアライメントソフトウェアです。改良点は以下のとおりです。

- PEZY-SC3 によるアライメント処理の高速化
- パイプライン段数とパイプライン構造の最適化による処理の高速化
- Fastq 読み込みの最適化による高速なクエリデータの読み込み
- スコアを調整するためのオプション等の追加
- Alternate contig へのアライメントのリフトオーバー機能
- ヒト集団に対するロングリードでのバリアントコール結果とリフトオーバー機能を活用したアライメント感度の向上

### reshz

reshz は、pzBWA-memの出力したアライメントデータに対して、Coordinate Sorting, Picard MarkDuplicates, GATK BQSR, Applying BQSRを行うツールです。 ZettaScaler3.0の大容量メモリを活用することで高速なデータ処理を行います。

### pzHaplotypeCaller

pzHaplotypeCallerは、GATK4.2.0.0のHaplotypeCallerをベースに、PEZY Computingで高速化・高精度化を行ったゲノム変異解析用ソフトウェアです。 改良点は以下の通りです。

- GATK4.2.0.0 HaplotypeCallerをもとにC++でフルスクラッチ
- PEZY-SC3を使用したSmith-WatermanアライメントとPairHMM処理の高速化
- CPU処理の最適化による高速化
- GATK4.2 で実装されたFRD, BQDなどの確率モデルがオプションで使用可能

## 速度と精度評価

本セクションでは、ZettaVEGAの速度と精度を評価します。
使用したZettaVEGAバージョン: v2.20.6

使用したデータは以下の通りです。 これらのデータをseqkitのsampleコマンドを用いて100Gbpとなるようにsub-samplingを行い速度と精度の評価に用いました。

```
seqkit sample -p factor 入力FASTQファイル > 出力FASTQファイル
```

| Run Name | Coverage | Total Gbp | Type | URL | sub-sampling factor |
|---|---|---|---|---|---|
| CNR0028192 (HG001) | x46.1 | 138.329 | PCR | https://db.cngb.org/search/run/CNR0028192/ | 0.722929 |
| CNR0028194 (HG001) | x42.1 | 126.174 | PCR-Free | https://db.cngb.org/search/run/CNR0028194/ | 0.79257 |
| PrecisionFDA challenge V1 HG001 | x53.6 | 160.700 | PCR-Free | https://console.cloud.google.com/storage/browser/genomics-public-data/precision-fda/input | 0.62229 |
| PrecisionFDA challenge V2 HG002 | x41.8 | 125.356 | PCR-Free | https://precision.fda.gov/challenges/10 | 0.79774 |

リファレンスは以下を用いました。

- hg19 + decoy
  - https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
  - リファレンス作成方法はAppendixを参照してください。
- hg38 gatk bundle
  - https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/ + population-contig

精度評価には、正解データとして v4.2.1 benchmarkを、精度評価ソフトウェアとしてrtg tools v3.12.1のvcfevalコマンドを使用しています。

### hg19 + decoy

hg19 + decoyリファレンスを使用した場合の実行時間(単位:秒)を以下の表に示します。

PZWGS v2.2.3 (2023/05) AlmaLinux 9.1

| Run Name | BWA time | SHZ time | HaplotypeCaller time | ALL time |
|---|---|---|---|---|
| CNR0028192 | 532.73 | 123.28 | 242.06 | 781.65 |
| CNR0028194 | 421.81 | 114.92 | 211.55 | 748.30 |
| Precision FDA challenge V1 HG001 | 537.42 | 123.60 | 213.73 | 874.79 |
| Precision FDA challenge V2 HG002 | 530.95 | 108.43 | 217.44 | 856.84 |

PZWGS v2.20.6 (2024/11) Ubuntu 24.04.1

| Run Name | BWA time | SHZ time | HaplotypeCaller time | ALL time |
|---|---|---|---|---|
| CNR0028192 | 352.72 | 61.15 | 227.16 | 641.03 |
| CNR0028194 | 359.02 | 61.40 | 186.23 | 606.65 |
| Precision FDA challenge V1 HG001 | 485.66 | 60.97 | 192.91 | 739.54 |
| Precision FDA challenge V2 HG002 | 478.37 | 59.54 | 188.42 | 726.33 |

また、このときのそれぞれの精度は以下のようになります。

Summary PZWGS v2.2.3 (2023/05) AlmaLinux 9.1

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9937 | 0.9923 | 0.9930 |
| CNR0028194 | 0.9971 | 0.9924 | 0.9947 |
| Precision FDA challenge V1 HG001 | 0.9897 | 0.9926 | 0.9911 |
| Precision FDA challenge V2 HG002 | 0.9964 | 0.9910 | 0.9937 |

PZWGS v2.20.6 (2024/11) Ubuntu 24.04.1

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9943 | 0.9926 | 0.9934 |
| CNR0028194 | 0.9976 | 0.9926 | 0.9951 |
| Precision FDA challenge V1 HG001 | 0.9900 | 0.9931 | 0.9916 |
| Precision FDA challenge V2 HG002 | 0.9971 | 0.9913 | 0.9942 |

SNP PZWGS v2.2.3 (2023/05) AlmaLinux 9.1

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9945 | 0.9939 | 0.9942 |
| CNR0028194 | 0.9972 | 0.9924 | 0.9948 |
| Precision FDA challenge V1 HG001 | 0.9900 | 0.9934 | 0.9917 |

| Name | precision | recall | f-measure |
|---|---|---|---|
| Precision FDA challenge V2 HG002 | 0.9971 | 0.9913 | 0.9942 |

PZWGS v2.20.6 (2024/11) Ubuntu 24.04.1

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9952 | 0.9942 | 0.9947 |
| CNR0028194 | 0.9977 | 0.9925 | 0.9951 |
| Precision FDA challenge V1 HG001 | 0.9903 | 0.9939 | 0.9921 |
| Precision FDA challenge V2 HG002 | 0.9977 | 0.9916 | 0.9946 |

non-SNP PZWGS v2.2.3 (2023/05) AlmaLinux 9.1

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9878 | 0.9806 | 0.9842 |
| CNR0028194 | 0.9965 | 0.9925 | 0.9945 |
| Precision FDA challenge V1 HG001 | 0.9874 | 0.9868 | 0.9871 |
| Precision FDA challenge V2 HG002 | 0.9920 | 0.9893 | 0.9906 |

PZWGS v2.20.6 (2024/11) Ubuntu 24.04.1
Indel

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9879 | 0.9810 | 0.9844 |
| CNR0028194 | 0.9967 | 0.9929 | 0.9948 |
| Precision FDA challenge V1 HG001 | 0.9878 | 0.9877 | 0.9877 |
| Precision FDA challenge V2 HG002 | 0.9930 | 0.9898 | 0.9914 |

GRCh38 + population-contig

hg38 リファレンスをダウンロードし、アライメント困難な領域におけるヒト集団のロングリードでの解析結果をindexのビルド時に付与しています。 hg38 + population-contig リファレンスを使用した場合の実行時間(単位:秒)を以下の表に示します。

PZWGS v2.2.3 (2023/05)

| Run Name | BWA time | SHZ time | HaplotypeCaller time | ALL time |
|---|---|---|---|---|
| CNR0028192 | 978.34 | 131.41 | 354.42 | 1464.25 |
| CNR0028194 | 1034.61 | 120.18 | 263.39 | 1418.22 |
| Precision FDA challenge V1 HG001 | 1147.73 | 114.66 | 296.67 | 1559.10 |
| Precision FDA challenge V2 HG002 | 1077.61 | 117.17 | 271.47 | 1466.28 |

PZWGS v2.20.6 (2024/11) Ubuntu 24.04.1

| Run Name | BWA time | SHZ time | HaplotypeCaller time | ALL time |
|---|---|---|---|---|
| CNR0028192 | 782.21 | 65.46 | 319.10 | 1166.77 |
| CNR0028194 | 791.59 | 66.34 | 237.62 | 1095.55 |
| Precision FDA challenge V1 HG001 | 940.48 | 66.04 | 255.87 | 1262.39 |
| Precision FDA challenge V2 HG002 | 854.58 | 64.03 | 236.15 | 1154.76 |

また、このときのそれぞれの精度は以下のようになります。

Summary
PZWGS v2.2.3 (2023/05)

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9952 | 0.9953 | 0.9952 |
| CNR0028194 | 0.9977 | 0.9963 | 0.9970 |
| Precision FDA challenge V1 HG001 | 0.9912 | 0.9960 | 0.9936 |
| Precision FDA challenge V2 HG002 | 0.9968 | 0.9946 | 0.9957 |

PZWGS v2.20.6 (2024/11)

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9952 | 0.9953 | 0.9952 |
| CNR0028194 | 0.9978 | 0.9962 | 0.9970 |
| Precision FDA challenge V1 HG001 | 0.9911 | 0.9959 | 0.9935 |
| Precision FDA challenge V2 HG002 | 0.9970 | 0.9937 | 0.9954 |

SNP PZWGS v2.2.3 (2023/05)

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9962 | 0.9972 | 0.9967 |

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028194 | 0.9979 | 0.9965 | 0.9972 |
| Precision FDA challenge V1 HG001 | 0.9915 | 0.9970 | 0.9943 |
| Precision FDA challenge V2 HG002 | 0.9975 | 0.9951 | 0.9963 |

PZWGS v2.20.6 (2024/11)

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9962 | 0.9971 | 0.9967 |
| CNR0028194 | 0.9979 | 0.9965 | 0.9972 |
| Precision FDA challenge V1 HG001 | 0.9915 | 0.9970 | 0.9942 |
| Precision FDA challenge V2 HG002 | 0.9976 | 0.9943 | 0.9959 |

non-SNP PZWGS v2.2.3 (2023/05)

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9881 | 0.9825 | 0.9853 |
| CNR0028194 | 0.9969 | 0.9943 | 0.9956 |
| Precision FDA challenge V1 HG001 | 0.9884 | 0.9884 | 0.9884 |
| Precision FDA challenge V2 HG002 | 0.9925 | 0.9908 | 0.9917 |

PZWGS v2.20.6 (2024/11)
Indel

| Name | precision | recall | f-measure |
|---|---|---|---|
| CNR0028192 | 0.9881 | 0.9824 | 0.9852 |
| CNR0028194 | 0.9969 | 0.9943 | 0.9956 |
| Precision FDA challenge V1 HG001 | 0.9885 | 0.9886 | 0.9885 |
| Precision FDA challenge V2 HG002 | 0.9933 | 0.9902 | 0.9917 |

# GATKとの互換性

CPUで実行したGATK 4.2.6.0の解析結果と、ZettaVEGA GATKモードの解析結果を比較した結果は以下の通りです。 入力データはオリジナルのファイルを使用しました。(HG001, HG002共にPrecision FDA challenge V1) PZWGS v2.20.6(2024年11月リリース)で実行時間，互換性共に大幅に改善されていることが分かります。

## 実行時間

実行環境: v2.2.3-AlmaLinux9.1, v2.20.6-Ubuntu24.04.1

| **Run Name** | **CPU mode** | **PZWGS v2.2.3(2023/5リリース)** | **PZWGS v2.20.6(2024/11リリース)** |
|---|---|---|---|
| Precision FDA challenge V1 HG001 | > 30hrs | 1618s | 1500s |
| Precision FDA challenge V1 HG002 | > 30hrs | 1725s | 1589s |

## 互換レベル

CPU実行結果との差異は以下の通りです。

- Precision FDA challenge V1 HG001

| | **PZWGS v2.2.3(2023/5リリース)** | **PZWGS v2.20.6(2024/11リリース)** |
|---|---|---|
| CPUのみの変異数 | 14463 | 5 |
| ZettaVEGAのみの変異数 | 1060 | 10 |
| 合致率(%) | 99.69527 | 99.99970 |

- Precision FDA challenge V2 HG002

| | **PZWGS v2.2.3(2023/5リリース)** | **PZWGS v2.20.6(2024/11リリース)** |
|---|---|---|
| CPUのみの変異数 | 15100 | 17 |
| ZettaVEGAのみの変異数 | 1303 | 4 |
| 合致率(%) | 99.67900 | 99.99958 |

# まとめ

本ドキュメントでは、高速なゲノム解析ソフトウェア ZettaVEGAを解説しました。PEZY-SC3を搭載したZettaScaler-3.0サーバーでは、33xのヒトゲノム解析を最大で100検体以上1日に処理することが可能です。また、GATK best practice guideに準拠したソフトウェア群を提供しているため、精度はGATKに準ずるかまたは向上しており、ユーザーのワークロードも、GATK best practice guideからZettaVEGAにシームレスに移行が可能です。 また、PZWGS v2.20.6(2024年11月リリース)で実行時間，精度が大幅に改善されました。

# Appendix

## hg19 + decoyの作成方法

hg19(hs37d5)リファレンスを以下のURLよりダウンロードします。

- https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

以下のコマンドでは1-22、X、Y、MTの主要染色体のデータとそれ以外のデータを一旦分離し、インデックス作成時に分離したデータをあらためてデコイ染色体として追加しています。これはZettaVEGA独自のインデックス作成方法です。デコイとして追加された染色体にマップされたリードは以後の処理ではマップされなかったものとして扱われます。

```
seqkit grep -n -r -p '^(X|Y|MT|[1-9])' hs37d5.fa -o hs37d5_pzbwa.fasta
seqkit grep -n -r -v -p '^(X|Y|MT|[1-9])' hs37d5.fa -o hs37d5_decoy.fasta
/opt/pezy/pzwgs/bin/bwa-index -i 8 -d hs37d5_decoy.fasta hs37d5_pzbwa.fasta
samtools faidx hs37d5_pzbwa.fasta
```