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、一台の 4U サーバーに統合しています。
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の GATK Best Practice と 99.99% 以上の互換性を保ったまま 100 倍以上高速に動作します。ZettaVEGA のゲノム解析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 Reference Genome, Fastq, VCF などの入出力といくつかのオプションを指定するだけで簡便に使用可能でき、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に加え、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

近年、次世代シーケンサー（NGS）の普及により、ゲノム解析の需要は研究・医療の両面で急速に拡大しています。しかし従来の解析手法では、1 検体あたり数十時間を要し、大規模コホート研究や臨床現場での即応性には大きな制約がありました。ZettaVEGA は、この課題を解決するために開発された高速ゲノム解析ソフトウェアです。PEZY-SC3 アクセラレータを搭載した ZettaScaler-3.0 システム上で動作し、GATK Best Practices と 99.99% 以上の互換性を保ちながら、従来比で最大 100 倍の速度で解析を実現します。これにより、1 日あたり最大 96 検体（33× カバレッジ）を処理可能となり、研究のスループットを飛躍的に高めるだけでなく、臨床現場での迅速な診断支援や個別化医療の実現にも大きく貢献します。本ホワイトペーパーでは、その技術的背景、性能評価、および導入における利点について詳しく紹介します。

アップデート内容

ZettaVEGA のバージョン 2.26.0 をリリースしました。アップデート内容は以下の通りです。

- GRCh38・CHM13 T2T などのリファレンスにおけるアライメント速度の向上
- GATK との一致度の向上
- HaplotyepCaller での Physical phasing の実装

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。

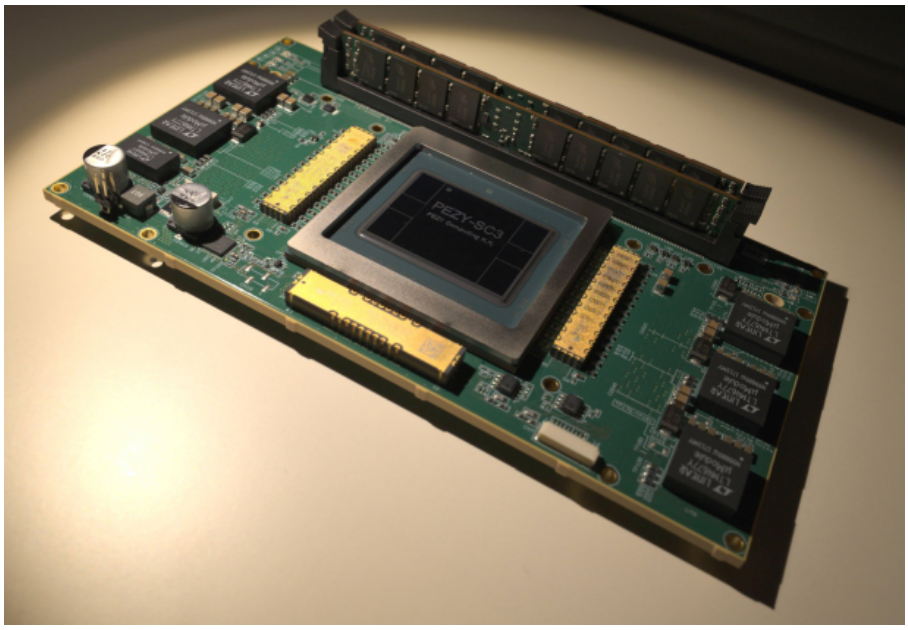
以下に ZettaScaler-3.0 の諸元を示します。

Table 1: ZettaScaler3.0 Spec

	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 が開発した並列計算に特化した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 19.66 TFlops、単精度で 39.32TFlops。半精度で 78.64 TFlops となります。また、メモリとして HBM2 を採用しており、1.2 TB/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 32 GB/s の帯域を持ちます。消費電力はモジュールあたり最大で 600W となっています。



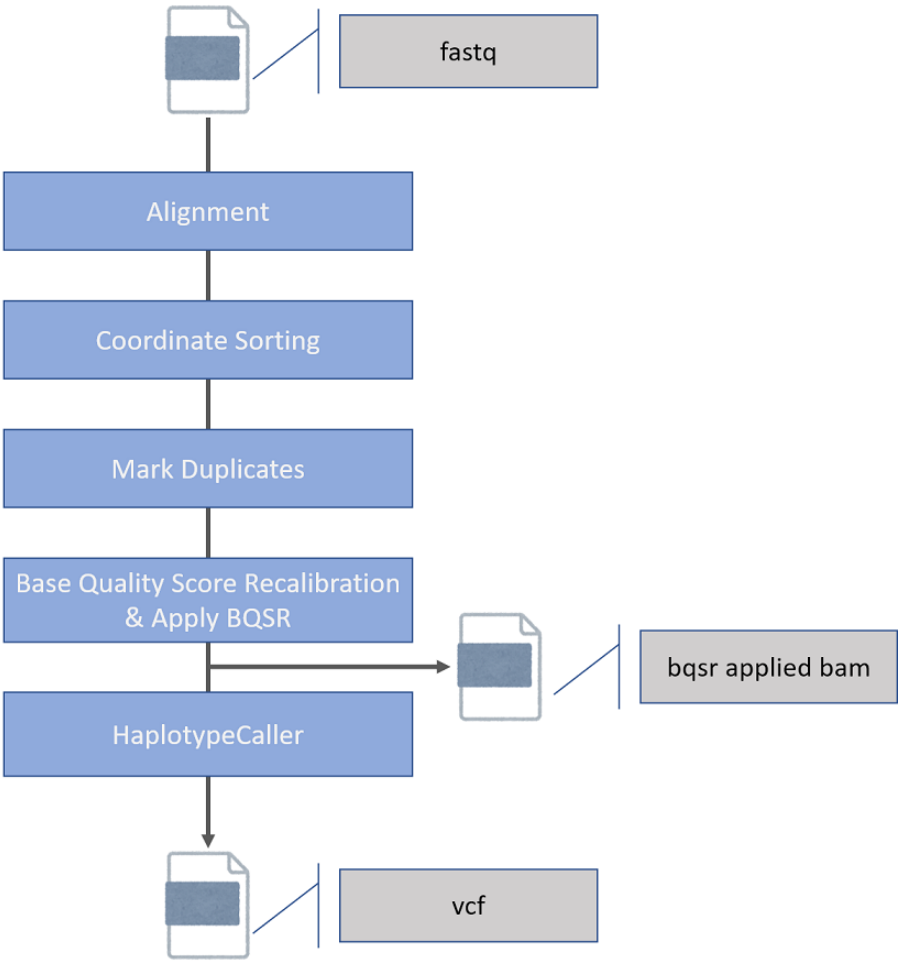
以下に PEZY-SC3 の諸元を示します。

Table 2: PEZY-SC3 Spec

	Spec
Number of PEs	4,096
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)	1,200
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。

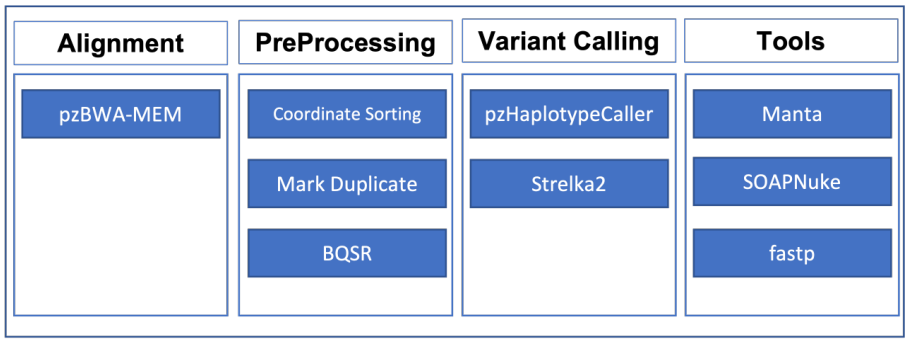


ゲノム解析パイプラインはシンプルなコマンドラインインターフェースを備えており、以下のように入力となるリファレンスゲノム・Fastq・出力先といくつかのオプションを指定するだけで複数のソフトウェアで構成されるパイプライン全体を実行可能です。

```
$ /opt/pezy/pzwgs/bin/pipeline.py
  --ref /data/ref/GRCh38/Homo_sapiens_assembly38.fasta \
  --known_sites /data/ref/GRCh38/Homo_sapiens_assembly38.dbsnp138.vcf \
  --known_sites /data/ref/GRCh38/Homo_sapiens_assembly38.known_indels.vcf \
  --known_sites /GRCh38/Mills_and_1000G_gold_standard.indels.hg38.vcf \
  --work_dir /scratch/local/tmp \
  --fq1 /data/fq/read1.fq \
  --fq2 /data/fq/read2.fq \
  --out_vcf /data/result/result.vcf \
  --preset_vc gatk-pcr-free
```

シンプルなインターフェースに加えて、個別のソフトウェアに様々なオプションを指定することができ柔軟性も兼ね備えております。
NCGM-genome/WGSpipeline 準拠の解析パイプラインなども備えており、これ以外のお客様の既存のパイプラインからの移行についても弊社が責任をもってサポートをいたします。

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前処理や統計情報取得を行うためのツールです。ZettaScaler3.0 の大容量メモリを活用してオンメモリにアライメントデータを置くことで高速なデータ処理を行います。以下のツールに相当する機能が使用可能です。

- Coordinate Sorting
- Picard MarkDuplicates
- GATK BQSR, Applying BQSR
- Save as bam or cram
- Picard CollectWgsMetrics

pzHaplotypeCaller

pzHaplotypeCaller は、GATK4.2.6.1 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。

- GATK4.2.6.0 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

GATK との互換性

ZettaVEGA は弊社があらかじめ設定している高精度なパラメータプリセットに加えて、GATK Best Practice と互換な結果を出力するためのパラメータプリセットを備えており、ユーザーのニーズに合わせて簡単に切り替えることが可能です。
本セクションでは　CPU で実行した GATK Best Practice (GATK version 4.2.6.0) の解析結果と、ZettaVEGA GATK Compatible モードでの解析結果を比較します。
比較対象とする GATK は Math.log 関数を使用している部分を StrictMath.log に修正してビルドしたものを使用しています。詳しくは、Appendix の GATK の StrictMath についての説明 に記載しております。出力形式は GVCF、リファレンスゲノムは GRCh38 を使用し、入力 Fastq は PrecisionFDA challenge V1 HG001: 53.6x, PrecisionFDA challenge V2 HG002: 41.8x を使用しました。

実行時間

Table 3: Processing time (GATK Compatible mode)

Input data	CPU mode	ZettaVEGA
Precision FDA challenge V1 HG001	> 30 h	2,998 s
Precision FDA challenge V2 HG002	> 30 h	2,167 s

一致度

出力された VCF を bcftools version 1.19 の bcftools isec コマンドをで比較した結果が以下の表です。すべてのコール結果が一致していることが示されています。

Table 4: Compatibility for GATK

input data	CPU GATK only	ZettaVEGA only	Call in both	Compatibility
Precision FDA challenge V1 HG001	0	0	148,966,181	100%
Precision FDA challenge V2 HG002	0	0	440,917,612	100%

加えて、以下のような diff コマンドを用いて vcf ファイルの比較を行ったところ、HG001, HG002 ともにヘッダ行以外での差は一行も検出されませんでした。
アップデートにより physical phasing に対応したことに加え、INFO や FORMAT の値についても徹底的な比較と修正を重ね、オリジナルの GATK との完全一致を実現できました。

```
diff -y –suppress-common-lines cpu.vcf zettavega.vcf > diff.vcf
```

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。
ZettaVEGA は GATK Best Practice と高い互換性を有するのみではなく、高精度な結果を得るためのパラメータプリセットを用意しており、ユーザーのニーズに合わせて簡単に切り替えることが可能です。
使用した ZettaVEGA バージョンは v2.26.0 です。
pzBWA と pzHaplotypeCaller には、弊社が用意している高精度なパラメータプリセットを指定して実行しております。GATK compatible モードの時の F-measure を表中の「F-measure (GATK)」として示しています。

入力 FASTQ とリファレンスゲノムの詳細については、Appendix の 入力 FASTQ hg19+decoy GRCh38 に記載しております。簡単のため入力 FASTQ はすべて 100 Gbasepair に down sample してから入力しております。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 での精度と速度

hg19 + decoy リファレンスを使用した場合の実行時間と精度を以下の表に示します。

Table 5: Processing time (hg19)

Input Data	BWA MEM (s)	Preprocess (s)	HaplotypeCaller (s)	Whole time (s)
CNR0028192	318.11	164.75	245.26	728.16
CNR0028194	319.13	173.42	216.77	709.38
Precision FDA challenge V1 HG001	518.87	168.71	214.33	901.97
Precision FDA challenge V2 HG002	442.07	161.44	202.90	806.46

Table 6: SNP (hg19)

Input data	Precision	Recall	F-measure	F-measure (GATK)
CNR0028192	99.52%	99.42%	99.47%	99.40%
CNR0028194	99.77%	99.26%	99.51%	99.40%
Precision FDA challenge V1 HG001	99.03%	99.39%	99.21%	99.46%
Precision FDA challenge V2 HG002	99.77%	99.17%	99.47%	99.33%

Table 7: Indel (hg19)

Input data	Precision	Recall	F-measure	F-measure (GATK)
CNR0028192	98.80%	98.10%	98.45%	95.00%
CNR0028194	99.67%	99.29%	99.48%	99.35%
Precision FDA challenge V1 HG001	98.78%	98.77%	98.77%	98.56%
Precision FDA challenge V2 HG002	99.26%	98.97%	99.12%	98.85%

1 サンプルあたり 900 秒（15 分）から 710 秒ほどで解析が完了します。1 サンプルあたり 900 秒で解析できた場合、1 日に 96 検体の解析を完了することが可能です。

GRCh38 での精度と速度

GATK bundle の GRCh38 リファレンスを使用した場合の実行時間と精度を以下に示します。

Table 8: Processing time (GRCh38)

Input Data	BWA MEM (s)	Preprocess (s)	HaplotypeCaller (s)	Whole time (s)
CNR0028192	560.80	174.41	301.99	1037.23
CNR0028194	580.74	178.45	237.08	996.33
Precision FDA challenge V1 HG001	871.58	183.64	259.85	1315.13
Precision FDA challenge V2 HG002	704.07	175.19	246.72	1126.03

Table 9: SNP (GRCh38)

Input data	Precision	Recall	F-measure	F-measure (GATK)
CNR0028192	99.33%	99.35%	99.34%	99.35%
CNR0028194	99.65%	99.19%	99.42%	99.34%
Precision FDA challenge V1 HG001	98.99%	99.29%	99.14%	99.41%
Precision FDA challenge V2 HG002	99.63%	98.97%	99.30%	99.20%

Table 10: Indel (GRCh38)

Input data	Precision	Recall	F-measure	F-measure (GATK)
CNR0028192	98.78%	98.04%	98.41%	94.98%
CNR0028194	99.61%	99.22%	99.41%	99.32%
Precision FDA challenge V1 HG001	98.76%	98.66%	98.71%	98.53%
Precision FDA challenge V2 HG002	99.23%	98.81%	99.02%	98.75%

GRCh38+population-contig での精度と速度

精度向上のため、GRCh38 リファレンスに対し、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の情報を index のビルド時に付与しています。GRCh38 + population-contig リファレンスを使用した場合の実行時間と精度を以下の表に示します。

Table 11: Processing time (GRCh38 + population-contig)

Input Data	BWA MEM (s)	Preprocess (s)	HaplotypeCaller (s)	Whole time (s)
CNR0028192	729.81	181.22	327.43	1238.54
CNR0028194	752.68	177.44	248.47	1178.64
Precision FDA challenge V1 HG001	999.95	189.39	280.97	1470.35
Precision FDA challenge V2 HG002	820.99	177.67	250.62	1249.33

Table 12: SNP (GRCh38 + population-contig)

Input data	Precision	Recall	F-measure
CNR0028192	99.62%	99.71%	99.67%
CNR0028194	99.79%	99.65%	99.72%
Precision FDA challenge V1 HG001	99.14%	99.70%	99.42%
Precision FDA challenge V2 HG002	99.76%	99.44%	99.60%

Table 13: Indel (GRCh38 + population-contig)

Input data	Precision	Recall	F-measure
CNR0028192	98.81%	98.25%	98.53%
CNR0028194	99.69%	99.43%	99.56%
Precision FDA challenge V1 HG001	98.85%	98.86%	98.85%
Precision FDA challenge V2 HG002	99.31%	99.01%	99.16%

通常の GRCh38 リファレンスと比較すると、100 ~ 200 秒実行時間の増加がみられるものの、すべてのサンプルにおいて精度の向上がみられました。

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 検体を 1 日に処理することが可能です。GATK best practice に含まれるソフトウェア群を実装しており、オリジナルのワークフローと完全に一致する結果と、高精度な結果の両方をユーザーに提供可能です。
今後、ロングリードのアプリケーションや Somatic Call への対応を予定しております。また、次世代プロセッサである PEZY-SC4s をリリース予定で、PEZY-SC4s を用いた製品では ZettaVEGA の約二倍の高速化や AI を用いたアプリケーションの実装が実現可能となる予定です。

Appendix

入力 FASTQ

入力データには以下表の FASTQ を使用しました。

Table 14: Input fastq

Run Name	Coverage	Total Gbp	Type	URL	sub-sampling factor
CNR0028192 (HG001)	46.1x	138.329	PCR	https://db.cngb.org/search/run/CNR0028192/	0.722929
CNR0028194 (HG001)	42.1x	126.174	PCR-Free	https://db.cngb.org/search/run/CNR0028194/	0.79257
PrecisionFDA challenge V1 HG001	53.6x	160.700	PCR-Free	https://console.cloud.google.com/storage/browser/genomics-public-data/precision-fda/input	0.62229
PrecisionFDA challenge V2 HG002	41.8x	125.356	PCR-Free	https://precision.fda.gov/challenges/10	0.79774

GATK との互換性を確認するセクションにおいては、PrecisionFDA challenge V1 HG001, PrecisionFDA challenge V2 HG002、のデータをそのまま入力として用いました。これらのデータを seqkit の sample コマンドを用いて 100 Gbp 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
```

hg19+decoy

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
```

BQSR 用の vcf ファイルは以下の三つを用いました。

- dbsnp_138.b37.vcf
- 1000G_phase1.indels.b37.vcf
- Mills_and_1000G_gold_standard.indels.b37.vcf

GRCh38

以下の GATK bundle のリンクより、リファレンスの FASTA と BQSR 用の VCF をダウンロードし解析に使用しました。
https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/
GRCh38 + population-congting の場合は、ヒト集団の Long read を用いたコール結果を用いた In house のデータセットを作成してインデックスをビルドしました。

BQSR 用の VCF ファイルは以下の三つを用いました。

- Homo_sapiens_assembly38.dbsnp138.vcf
- Homo_sapiens_assembly38.known_indels.vcf
- Mills_and_1000G_gold_standard.indels.hg38.vcf

GATK の StrictMath についての説明

Java の Math クラスは精度を犠牲にしてプロセッサに応じた高速な命令を使用する実装となっており、その挙動は CPU や JVM によって異なります。
GATK は数学関数に Java の Math クラスを使用しており、特に log 関数の誤差が原因で、C++ ベースで実装されている pzHaplotypeCaller と計算結果を一致させることが困難です。
Java に用意されている StrictMath クラスは IEEE 754 の Recommended Operations に準拠した C 言語のライブラリを使用しているため、GATK の Math クラスを使用している箇所を StrictMath に置き換えることで GATK の計算誤差を修正し、pzHaplotypeCaller と正しく比較することが可能となります。
StrictMath を使用するかどうかは、GVCF モードでのバリアントコールのときゲノム全体で十数〜数十か所のバリアント（NON_REF を含む）のコール結果に影響します。

GATK Best Practice のワークフロー

比較対象とした GATK Best Practice のワークフローのコマンド例を以下に示します。

```
sample=HG001
benchmark_dir=$(pwd)
result_dir=${benchmark_dir}/result_gatk_best_practice_${sample}
mkdir -p ${result_dir}
cd ${result_dir}
echo run GATK best practice. The results will be in ${result_dir}

# input files
reference_dir=/data/ref/GRCh38/
```

```
ref=${reference_dir}/Homo_sapiens_assembly38.fasta
bwa_ref=${reference_dir}/Homo_sapiens_assembly38.fasta
known_sites1=${reference_dir}/Homo_sapiens_assembly38.known_indels.vcf
known_sites2=${reference_dir}/Homo_sapiens_assembly38.dbsnp138.vcf
known_sites3=${reference_dir}/Mills_and_1000G_gold_standard.indels.hg38.vcf

fastq_dir=/data/pFDA_V1/fastq/
fq1=${fastq_dir}/HG001-NA12878-pFDA_S1_L001_R1_001.fastq.gz
fq2=${fastq_dir}/HG001-NA12878-pFDA_S1_L001_R2_001.fastq.gz
read_group="@RG\tID:NA12878\tPL:TEST\tPU:TEST\tLB:TEST\tSM:NA24385\tCN:TEST"

echo "Input␣files:␣${fq1}␣,␣${fq2}"
# output file prefix
prefix=${sample}.GRCh38.cpu
vcf=${prefix}.vcf

# path to tools
bwa=${benchmark_dir}/tools/bwa-0.7.17/bwa
samtools=${benchmark_dir}/tools/bin/samtools
java=${benchmark_dir}/tools/jdk8u402-b06/bin/java
gatk=${benchmark_dir}/tools//gatk-4.2.6.1/build/libs/gatk-package-4.2.6.1-42-g8c348aa-SNAPSHOT-local
    .jar
n_threads=128

#BWA MEM alignment, conversion to bam, sort
${bwa} mem -t $n_threads -Y -M -K 160000000 $ref $fq1 $fq2 -R $read_group \
   | ${samtools} view -buhS \
   | ${samtools} sort -@ $n_threads -m 2G > ${prefix}.bam
${samtools} index -@ $n_threads ${prefix}.bam

#mark duplicates
${java} -jar -Xmx30g ${gatk} MarkDuplicates -I ${prefix}.bam -M ${prefix}.markdup.metrics.txt -O ${
    prefix}.markdup.bam --REMOVE_DUPLICATES false
${samtools} index -@ $n_threads ${prefix}.markdup.bam

#BQSR
${java} -jar -Xmx30g ${gatk} BaseRecalibrator -R $ref -I ${prefix}.markdup.bam -O bqsr.table \
   --known-sites $known_sites1 \
   --known-sites $known_sites2 \
   --known-sites $known_sites3 \

${java} -jar -Xmx30g ${gatk} ApplyBQSR -R $ref -I ${prefix}.markdup.bam --bqsr-recal-file bqsr.table
    -O ${prefix}.markdup.bqsr.bam
${samtools} index -@ $n_threads ${prefix}.markdup.bqsr.bam

#HaplotypeCaller
${java} -jar -Xmx30g ${gatk} HaplotypeCaller \
      --reference $ref \
      --input ${prefix}.markdup.bqsr.bam \
      --output $vcf \
      --pcr-indel-model NONE \
      --emit-ref-confidence GVCF \
      -pairHMM FASTEST_AVAILABLE \
      --smith-waterman FASTEST_AVAILABLE \
```

```
    --native-pair-hmm-threads $n_threads \
    --native-pair-hmm-use-double-precision true \
    --max-reads-per-alignment-start 0 \
    --max-prob-propagation-distance 51 \
    --minimum-mapping-quality 20
```