# Written 03/01/2019 by Chunlei Yu.

Germline VariantCalling Tools

This note is to compare the performance of Germline variantcalling tools

###Tools:
DeepVariant
GATK4 CNN
GATK4 “hardfilter”
GATK4 VQSR
Strelka
Freebayes

###Data resources

##Testing BAM file :
`151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bam`

Downloaded from https://github.com/genome-in-a-bottle/giab_data_indexes/blob/master/AshkenazimTrio/alignment.index.AJtrio_OsloUniversityHospital_IlluminaExome_bwamem_GRCh37_11252015

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bam
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bai

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG003-EEogPU_v02-KIT-Av5_TCTTCACA_L008.posiSrt.markDup.bam
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG003-EEogPU_v02-KIT-Av5_TCTTCACA_L008.posiSrt.markDup.bai
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG004-EEogPU_v02-KIT-Av5_CCGAAGTA_L008.posiSrt.markDup.bam

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG004-EEogPU_v02-KIT-Av5_CCGAAGTA_L008.posiSrt.markDup.bai

##Reference FASTA

hs37d5.fa.gz
The original file came from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence. Because DeepVariant requires bgzip files, we had to unzip and bgzip it, and create corresponding index files.

wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz.fai

wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz.gzi

##Truth VCF and BED

HG002_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-22_v.3.3.2_highconf_* are from NIST, as part of the Genomes in a Bottle project. They are downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.3.2/GRCh37/

##Capture target BED file
According to the paper "Extensive sequencing of seven human genomes to characterize benchmark reference materials", the HG002 exome was generated with Agilent SureSelect. In this case study we'll use the SureSelect v5 BED (`agilent_sureselect_human_all_exon_v5_b37_targets.bed`) and intersect it with the GIAB confident regions for evaluation.

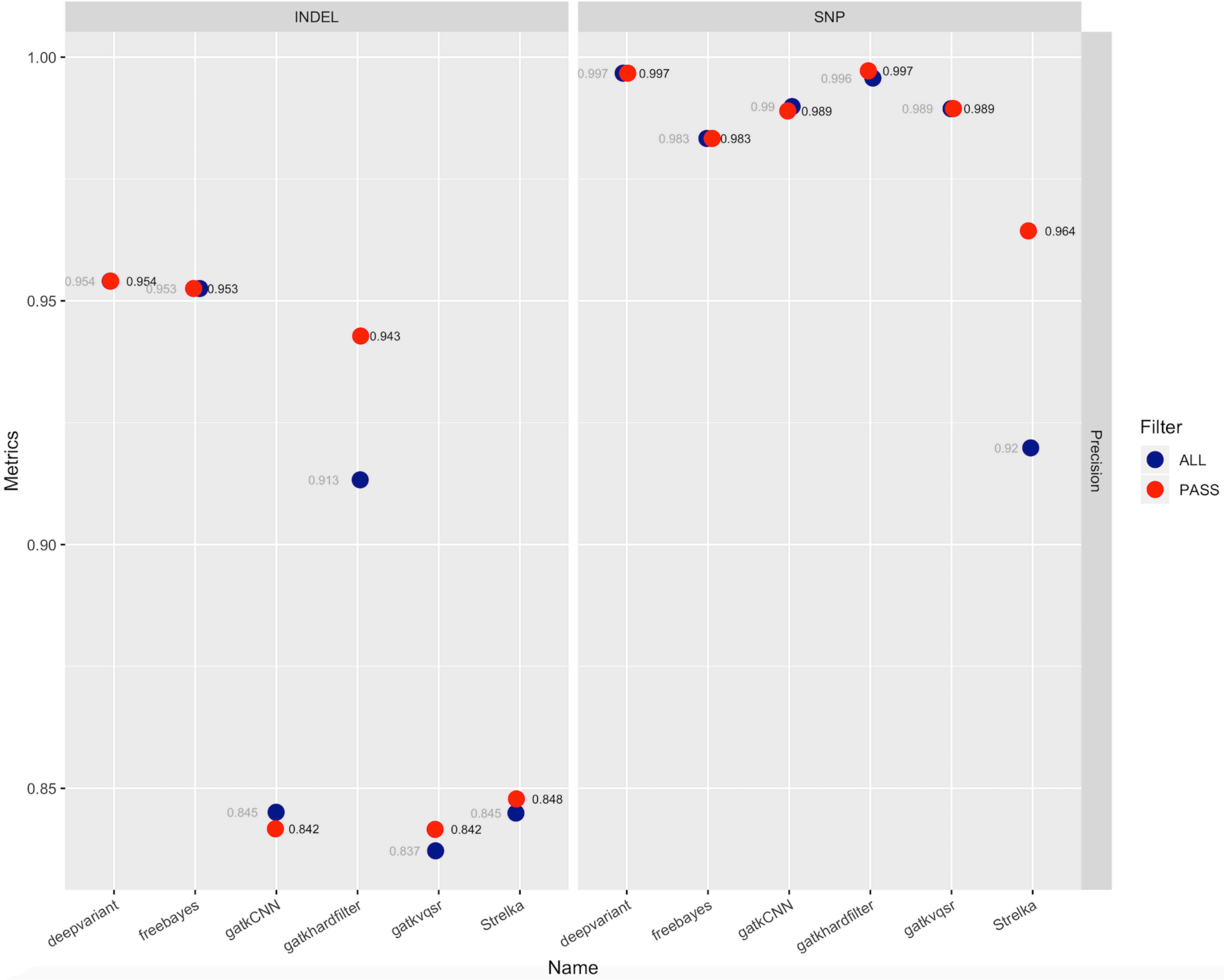
##Docker images
Docker resources:
DeepVariant: https://hub.docker.com/r/dajunluo/deepvariant
GATK suite: docker pull broadinstitute/gatk
Strelka: Dockerfile
FreeBayes: Dockerfile

###Performance:

Tools	Run time	CPU
DeepVariant	5hr11min, CPU:20	20
GATK-CNN	6hr56min, CPU:10	10
GATK-hardfilter	1hr, CPU:2	2
GATK-VQSR	2hr	2
Strelka	10min	2
FreeBayes	1hr19min	2

##Precision:

(Comparisons to the Genome in a Bottle truth set for this sample were performed using the hap.py software, available on GitHub at http://github.com/Illumina/hap.py, using the same version of the GIAB truth set (v3.2.2) used by pFDA.)


INDEL
SNP
Precision
Metrics
1.00
0.95
0.90
0.85
Name
deepvariant
freebayes
gatkCNN
gatkhardfilter
gatkvqsr
Strelka
deepvariant
freebayes
gatkCNN
gatkhardfilter
gatkvqsr
Strelka
0.954
0.954
0.953
0.953
0.943
0.913
0.845
0.842
0.842
0.837
0.845
0.848
0.997
0.997
0.983
0.983
0.99
0.989
0.996
0.997
0.989
0.989
0.964
0.92
Filter
● ALL
● PASS