# Written 03/01/2019 by Chunlei Yu.

Germline VariantCalling Tools

To identify clinical variants accurately and consistently, I have compared 6 germline variant calling pipelines. Each method was run according to the individual authors' best-practice recommendations. Method information used in germline variant calling pipelines are shown in Table.1.

Table.1 Methods used in variant calling pipelines.

Method	Version	Algorithm	References
DeepVariant	1.6	Deep neural network	Ryan Poplin, Pi-Chuan Chang. et al. (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36, 983–987. doi:10.1038/nbt.4235
Genome Analysis Toolkit (GATK) HardFilter	4.1.0.0	Filter variant calls based on INFO and/or FORMAT annotations	https://software.broadinstitute.org/gatk/
Genome Analysis Toolkit (GATK) CNNScoreVariants	4.1.2.0	Convolutional Neural Network (CNN)	https://software.broadinstitute.org/gatk/
Genome Analysis Toolkit (GATK) VQSR	4.1.2.0	Machine learning	https://software.broadinstitute.org/gatk/
FreeBayes	1.1.0	Bayesian genetic variant detector	Erik Garrison, Gabor Marth. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 (http://arxiv.org/abs/1207.3907)
Streka	2.9.7	Tiered haplotype-modeling strategy	Kim, S., Scheffler, K. et al. (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15, 591-594. doi:10.1038/s41592-018-0051-x

Case study for 6 germline variant calling pipelines.

Table.2 Data sources

Data	File Name
NA24385 BAM	151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bam 151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bai

Reference Genome	hs37d5.fa.gz hs37d5.fa.gz.fai hs37d5.fa.gz.gzi
Truth VCF	HG002_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-22_v.3.3.2_highconf_*
Truth BED	HG002_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_CHROM1-22_v3.2.1_highconf.bed
Capture target BED	Exome-Agilent_V6.bed.gz
dbSNP database	dbsnp_138.b37.vcf
Mills indel database used for BQSR	hapmap_3.3.b37.vcf
Known indels 1000G	1000G_phase1.indels.b37.vcf
HapMap genotypes and sites VCFs	hapmap_3.3.b37.vcf
OMNI 2.5 genotypes for 1000 Genomes samples	1000G_omni2.5.b37.vcf

1 GermlineVC_DeepVariant pipeline

Runtime

Step	Run time	CPU
make_examples	4hr48min	20
call_variants	1hr11min	20
postprocess_variants	25sec	20

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we applied Genome in a bottle truth dataset and evaluation methodology (hap.py) to GermlineVC_DeepVariant pipeline.

Type	Recall	Precision	F1_Score
InDel	0.969	0.992	0.980
SNP	0.99	0.997	0.993

2 VARIANTCALLING_HARDFILTER pipeline

Runtime

Step	Run time	CPU
BQSR	1hr12min	4
haplotypecaller	4hr24min	4
vf_indel	1min	4
vf_snp	1min	4
merge_snp_indel	27sec	4

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we appliedGenome in a bottle truth dataset and evaluation methodology (hap.py) to VARIANTCALLING_HARDFILTER_EN pipeline.

Type	Recall	Precision	F1_Score
InDel	0.967	0.982	0.974
SNP	0.970	0.997	0.983

3 VARIANTCALLING_GATKCNN pipeline

Runtime

Step	Run time	CPU
BQSR	1hr12min	4
RunHC4	3hr10min	10
CNNScoreVariants	41min	10
FilterVariantTranches	1min	4

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we appliedGenome in a bottle truth dataset and evaluation methodology (hap.py) to VARIANTCALLING_GATKCNN pipeline.

Type	Recall	Precision	F1_Score
InDel	0.925	0.973	0.948
SNP	0.839	0.989	0.908

4 VARIANTCALLING_VQSR pipeline

Runtime

Step	Run time	CPU
RunHC4	3hr10min	10
VariantRecalibratorINDEL	3min	4
ApplyRecalibrationINDEL	23sec	4
VariantRecalibratorSNP	11min	4
ApplyRecalibrationSNP	36sec	4

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we appliedGenome in a bottle truth dataset and evaluation methodology (hap.py) to VARIANTCALLING_VQSR pipeline.

Type	Recall	Precision	F1_Score
InDel	0.958	0.974	0.966
SNP	0.990	0.990	0.990

5 VARIANTCALLING_FreeBayes pipeline

Runtime

Step	Run time	CPU
Germlinecall	1hr17min	4

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we applied Genome in a bottle truth dataset and evaluation methodology (hap.py) to VARIANTCALLING_FreeBayes pipeline.

Type	Recall	Precision	F1_Score
InDel	0.922	0.968	0.944
SNP	0.989	0.988	0.988

6 VARIANTCALLING_Strelka pipeline

Runtime

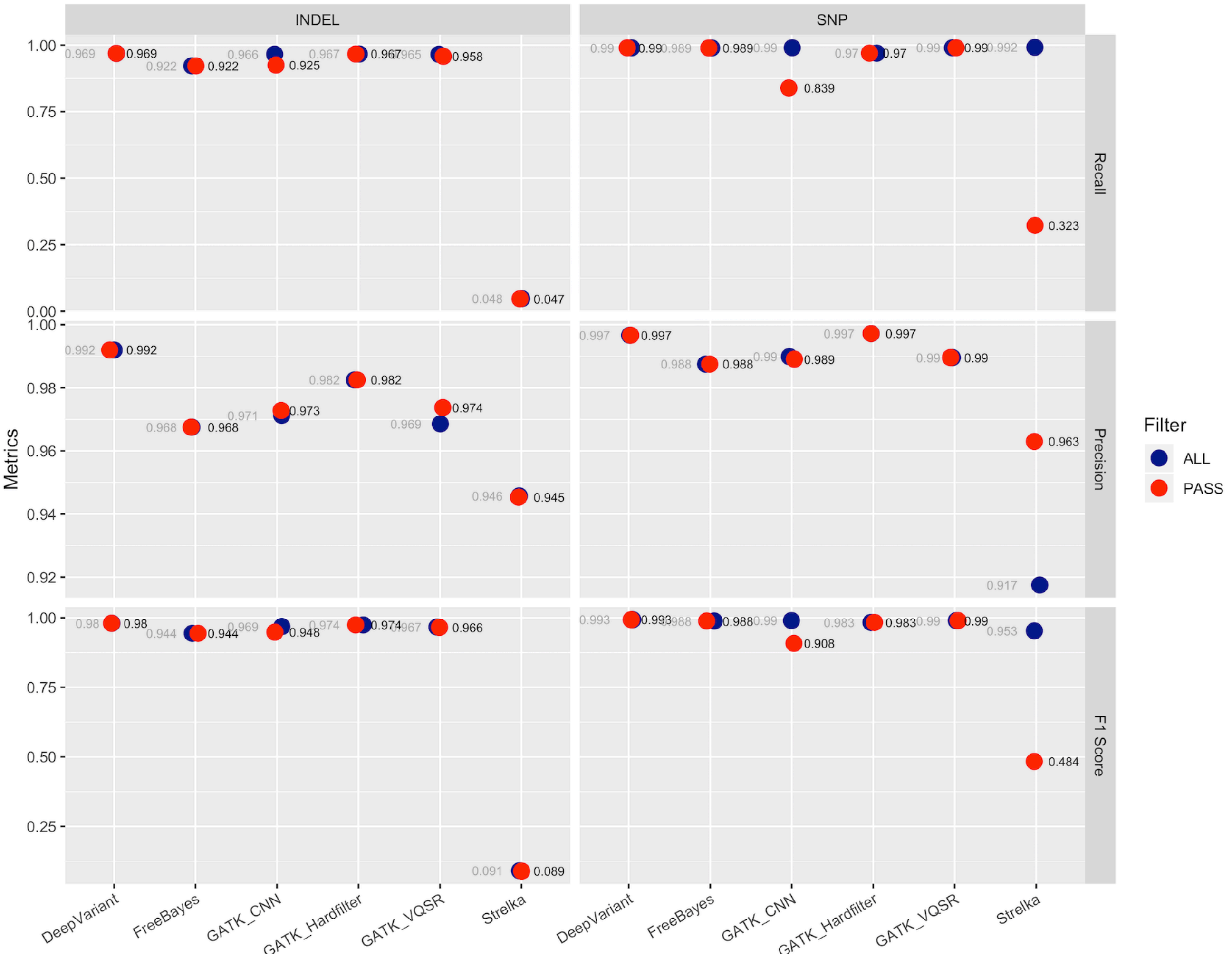
Step	Run time	CPU
configureStrelkaGermlineWorkflow	28min	4

Performance

To evaluate the sensitivity and precision for SNPs and InDels, we applied Genome in a bottle truth dataset and evaluation methodology (hap.py) to VARIANTCALLING_Strelka pipeline.

Type	Recall	Precision	F1_Score
InDel	0.047	0.945	0.089
SNP	0.323	0.963	0.484

Performance Summary:



Evaluation of 6 Germline variant calling methods on NA24385 BAM data mentioned in table.2.