**百度一体机测评**

本次测评主要目的是测试一体机上WGS的HaplotypeCaller Flow 的**时间**和结果的**准确性**。

**测试一**

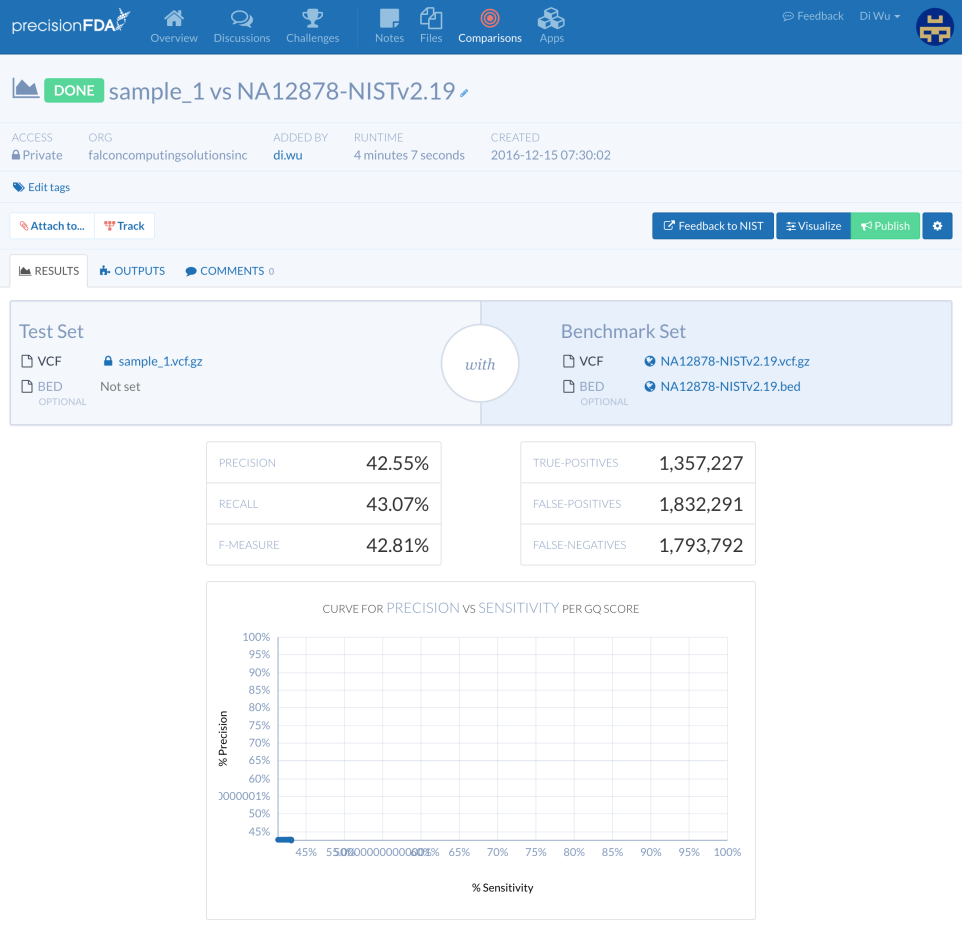
发现一体机上Flow只出gvcf结果，百度方debug后接着测试。

**测试二**

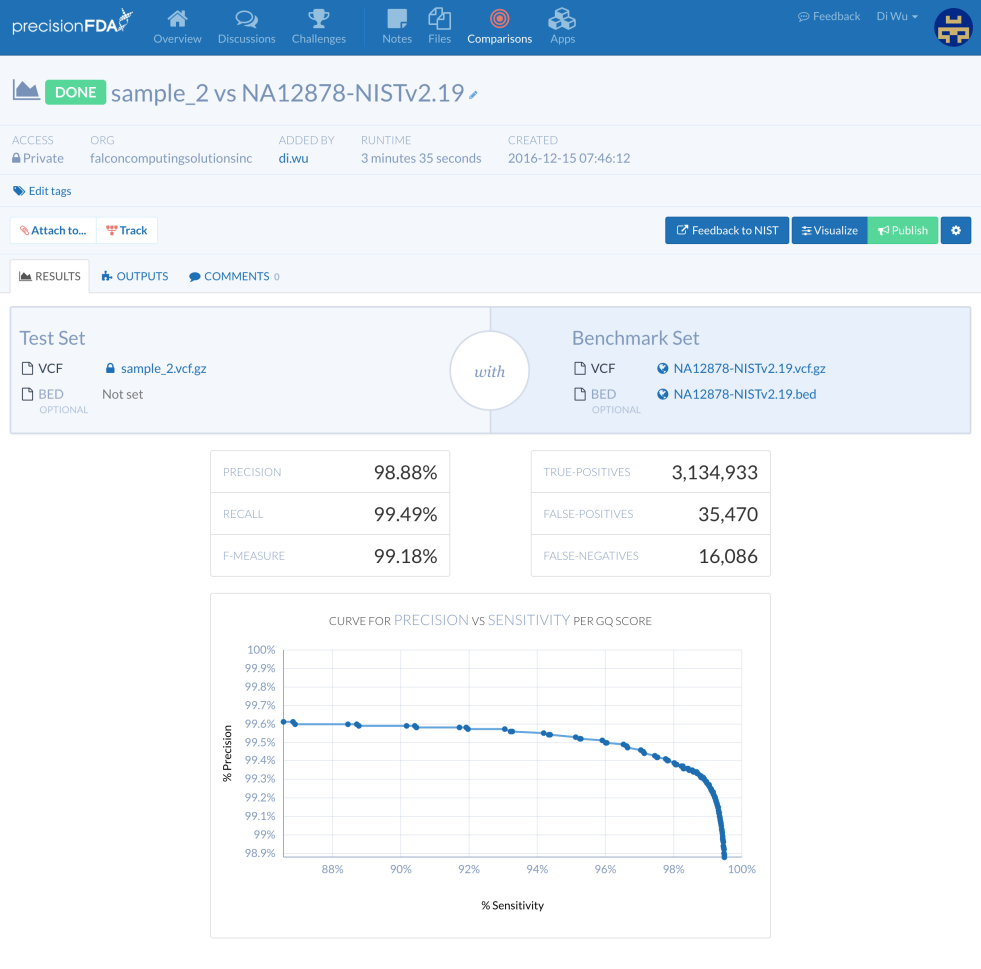
debug完后测试发现对单个样本call出来的结果仍未gvcf结果（虽然文件名为vcf）。推论：一体机上GATK Haplotype caller 接Genotypegvcf这个flow对单个样本不能得到vcf结果。

**测试三**

这次测试取了NA12877和NA12878两个样本一起进行Genotypegvcf。顺利得到vcf结果后，对每个样本vcf结果进行评估。PrecisionFDA结果如图。



图一 NA12877 VS GIAB2.19



图二 NA12878 VS GIAB2.19

TIME

Bwa -> Haplotype Caller Genotypegvcf

NA12877 13.69389h（期间有另外的pipeline在run） 20.8min

NA12878 8.448333h

Accuracy

NA12878的结果正常，NA12877结果异常。原因不知。

**Command**

**>> Script pipeline.ERR194146.sh**

sample\_id=$1

ref\_dir=/pool/local/ref

ref\_genome=$ref\_dir/human\_g1k\_v37.fasta

db138\_SNPs=$ref\_dir/dbsnp\_138.b37.vcf

g1000\_indels=$ref\_dir/1000G\_phase1.indels.b37.vcf

g1000\_gold\_standard\_indels=$ref\_dir/Mills\_and\_1000G\_gold\_standard.indels.b37.vcf

start\_ts=$(date +%s)

echo start time : $start\_ts

#set -x

fcs-genome align \

-r $ref\_genome \

-1 /pool/storage/fastq/${sample\_id}\_1.fastq.gz \

-2 /pool/storage/fastq/${sample\_id}\_2.fastq.gz \

-o ${sample\_id}.bam \

-R $sample\_id -S $sample\_id -L $sample\_id -P illumina

end\_ts1=$(date +%s)

echo "align finishes in $end\_ts1 seconds"

fcs-genome bqsr \

-r $ref\_genome \

-i ${sample\_id}.bam \

-o ${sample\_id}.recal.bam \

-K $db138\_SNPs \

-K $g1000\_indels \

-K $g1000\_gold\_standard\_indels

end\_ts2=$(date +%s)

echo "bqsr finishes in $end\_ts2 seconds"

fcs-genome htc \

-r $ref\_genome \

-i ${sample\_id}.recal.bam \

-o ${sample\_id}.gvcf

end\_ts3=$(date +%s)

echo "htc finishes in $end\_ts3 seconds"

echo end time : $end\_ts3

**>> work command**

nohup time sh pipeline.ERR194146.sh ERR194146 > pipeline.ERR194146.sh.o.log 2> pipeline.ERR194146.sh.e.log &