# GVC: A superfast and universal genomic variant caller

Zhongbo Zhang[1], Longhui Yin[1], Lihua Cao[1], Changling Liu[1], Lingtong Hao[1] and Shaoping Ling[1]

[1]Genowis Inc., Beijing, China

## Abstract

Germline and somatic variant detection from human and cancer whole-genome sequencing data is a challenge task for genome-wide association study and cancer genomics in precision medicine. Many confounding factors contribute the difficulties including complexity of variant, sequencing and alignment error, tumor clonality and sample purity etc. Current genomic variant callers are too time-consuming to meet the requirement of clinical application in precision medicine. We developed superfast and universal Genomic Variant Caller (GVC), which can simultaneously detect various genomic variants including SNV, sINDEL and SV from personal and normal-cancer paired whole-genome/exome sequencing data within fifteen minutes. What's more, it achieved higher sensitivity and precision than popular variant callers including GATK4, Mutect, NovoBreak in germline and somatic variant detection from NA12878 and ICGC-TCGA Dream Challenge Datasets respectuvely. It is worth mentioning that GVC achieved comparable performance in variant detection from NA12878 sequenced by three different high-throughput sequencing platforms including Illumina HiSeq2000, NovaSeq and BGISEQ-500.

## Results

GVC adopts multithread to read bam file and hash structure to save all statistics and improve processing speed. GVC version in iGenomeComputer (HiX) only costs 15 minutes for detecting all three types of germline and somatic variants from paired cancer whole-genome sequencing data (Normal: 30x vs Cancer: 30x).

As to somatic variant detection, the ICGC-TCGA DREAM Genomic Mutation Calling Challenge[3] set5 data is used to evaluate the performance of GVC. For somatic SNV, GVC outcompeted MuTect[6] with sensitivity of 99.46% vs 92.13%, precision of 92.26% vs 99.98% and F1-SCORE of 99.36% vs 95.89% respectively. For SV, GVC outcompeted NovoBreak[7] with sensitivity of 91.52% vs 92.28%, precision of 100% vs 82.56% and F1-SCORE of 95.57% vs 87.15%. For sINDEL, GVC outcompeted pindel with sensitivity of 62.47% vs 50.47%, precision of 71.71% vs 70.90% and F1-SCORE of 66.77%% vs 58.97% (Table 1).

For germline variant detection, we used the Genome in a Bottle Consortium (GIAB)[5] truth set derived from the well-studied human sample NA12878. Three NA12878 sequencing datasets from NovaSeq, HiSeq2000 and BGISEQ-500 and HiSeq2000 were used to evaluate the GVC performance of germline variant detection (Table 2).

For SNP,GVC achieved the F1-score of 99.90%, 99.74% and 99.82% on the NovaSeq, BGISEQ-500 and HiSeq2000 datasets respectively. It outcompeted GATK4, which achieved the F1-score of 98.99%, 97.44% and 97.12% on NovaSeq, BGISEQ-500 and HiSeq2000 datasets respectively (Table 3).

For INDEL,GVC achieved the F1-score of 97.04%,97.30% and 96.29% on the NovaSeq, BGISEQ-500 and HiSeq2000 datasets respectively. It outcompeted GATK4 which achieved the F1-score of 98.34%, 90.98% and 85.05% on NovaSeq, BGISEQ-500 and HiSeq2000 datasets respectively (Table 4).

In respect of sensitivity and specificity metrics, GVC and GATK4 both have high performance in SNP detection. By comparing AUC (Area Under Curve), GVC achieved AUC of 0.9370, 0.9447 and 0.9208 on the NovaSeq, BGISEQ-500 and HiSeq2000 datasets which are superior to GATK4 which achieved AUC of 0.8766, 0.9213 and 0.8908 on three datasets (Fig 1a). In sINDEl detection, GVC has much more advantage compared with GATK4 (Fig. 1b).

From the box plot in Figure 2, we can see the generalization ability of each chromosome. The F1-scores are higher than 0.99 in SNV and 0.96 in sINDEL. But in model built by chromosome 21 and 22, the boxes show a larger volatility. The reason maybe that chromosome 21 and 22 have less

information than other chromosomes.

For SV, due to the lack of the “ground-truth” set, it is not available for comparing the performance of GVC seekSV[8] and DELLY[2] directly. We set the truth set by the SV break point that supported by at least two software. GVC achieved the sensitivity of 90.93% and 93.93% on the HiSeq2000 and BGISEQ-500 datasets respectively, which outcompeted DELLY[2] with the sensitivity of 73.70 % and 76.60% on HiSeq2000 and BIGSEQ-500 datasets respectively, and seekSV[8] with the sensitivity of 85.69 % and 80.68% on HiSeq2000 and BIGSEQ-500 datasets respectively (Table 5).

For 30X tumor-normal sequence data somatic analysis, GVC can obtain the results of SNV, SV and Indel simultaneously in 15 minutes. However, it almost takes three days for GATK4 obtaining the INDEL and SNP results, and one day for DELLY[2] getting the SV results.
Meanwhile, for 50X sequence data germline analysis, GVC can obtain the results of SNP, SV and Indel simultaneously in 15 minutes. However, it almost takes one day for MuTect[6] obtaining the INDEL and SNP results, and eight hours for novoBreak[7] getting the SV results.
For tumor-normal sequence cancer sequence data, it is available for GVC to get the results of somatic and germline at the same time.

To verify that model built by xgboost algorithm has good generalization performance, we use each chromosome in NA12877 HiSeq dataset to test model based on the corresponding chromosome in NA12877 HiSeq dataset. From Figure 3, we can see that SNP and sINDEL in NA12877 HiSeq dataset are well detected by NA12878 model. But in model built by chromosome 21 and 22, the calling performances are not so good because that chromosome 21 and 22 contain less positions comparing with other chromosomes.

In summary, GVC can provide the most complete function of analysis of SNV, SV and Indel simultaneously in 15 minutes. Compared with other commonly used genome analysis software, such as GATK4[1], MuTect[6], novoBreak[7] and DELLY[2], GVC is much more powerful and can be applied to the prevalent sequencing platform, such as HiSeq2000, NovaSeq and BGISEQ, meeting the requirement of clinical application in precision oncology.

## Methods

In the first mapping step, GVC pipeline aligns paired-end FASTQ files to the human GRCh37 reference genome with BWA-MEM. In the second step, GVC extracts feature information from Binary Alignment Map (BAM) file to assemble feature vector space, which includes: 1) the quality statistics information, such as the consensus base, consensus quality, SNP and RMS quality, 2) some statistics, such as the split reads supporting rates; abnormal pair reads number; reads mismatch number; 3) the distance from variant point to read end; 4) Phred-scaled p-value and

somatic score. The dbSNP data set and validated somatic variants(positive or negative sites) are regarded as the labels. XGBoost4, an efficient machine-learning algorithm, is applied to learn different decision rule for different type of genomic variants, the candidate mutation set is obtained by using the well-trained machine learning model. Finally, the unsure mutation points can be validated, and the validated database can be updated by adding the newly assured mutation points (Fig.4).

Availability: GVC is available from http://gvc.0cancer.me

## References


1. Ga V D A, Carneiro M O, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline[J]. Current Protocols in Bioinformatics, 2013, 43(1110):11.10.1.
2. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis[J]. Bioinformatics, 2012, 28(18).
3. Ewing A D, Houlahan K E, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection[J]. Nature Methods, 2015, 12(7): 623-630.
4. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. knowledge discovery and data mining, 2016: 785-794.
5. Zook J M, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls[J]. Nature Biotechnology, 2014, 32(3): 246-251.
6. Cibulskis K, Lawrence M S, Carter S L, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. [J]. Nature Biotechnology, 2013, 31(3): 213-219.
7. Chong Z, Ruan J, Gao M, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. [J]. Nature Methods, 2017, 14(1):65.
8. Liang Y, Qiu K, Liao B, et al. seeksv: an accurate tool for somatic structural variation and virus integration detection[J]. Bioinformatics, 2016, 33(2):184.
9. Huang J, Liang X, Xuan Y, et al. A reference human genome dataset of the BGISEQ-500 sequencer[J]. Gigascience, 2017.

## Figure Legends

Figure1. Germline calling performance: receiver operating characteristic (ROC) curves comparing the performance of GVC to GATK4 in three NA12878 sequencing datasets

Figure2.The performance of each chromosome's training model by predicting all other chromosomes

Figure 3. SNP and sINDEL calling performance on HiSeq NA12877 (model based on NA12878 )

Figure 4. The comparison between GVC, Delly and Seeksv.

Figure5. The comparison between GVC, Delly and Seeksv.

# Tables

| | Ture Positive | False Positive | Sensitivity | Precision | F1-SCORE | Time |
|---|---|---|---|---|---|---|
| GVC | 45138 | 337 | 99.46% | 99.26% | 99.36% | 15min |
| Mutect | 41811 | 8 | 92.13% | 99.98% | 95.89% | 50h30min |
| GVC | 723 | 0 | 91.52% | 100.00% | 95.57% | 15min |
| novoBreak | 729 | 154 | 92.28% | 82.56% | 87.15% | 2h35min |
| GVC | 10437 | 4118 | 62.47% | 71.71% | 66.77% | 15min |
| pindel | 8433 | 3461 | 50.47% | 70.90% | 58.97% | 19hour |

Note: All timings reported herein were performed on iGenomeComputer HIX.

**Table 1. Somatic mutation calling performance**

| Sequencing platform | Study accession | Reads Length | Mapping Rate | Proper pair Rate | Depth |
|---|---|---|---|---|---|
| NovaSeq | NA12878-REP1 | PE151 | 99.54% | 95.79% | 58 |
| BGISEQ-500 | ERP017158 | PE100 | 99.88% | 98.16% | 36 |
| HiSeq2000 | ERP001960 | PE101 | 99.35% | 97.17% | 50 |

Note : NovaSeq data download from https://basespace.illumina.com/sample/42318284/NA12878-rep1

**Table 2. NA12878 Data information**

| | Sequencing platform | True Postive | | Sensitivity | | Precision | | F1-SCORE | | Time | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 |
| SNP | NovaSeq | 3188373 | 3132935 | 99.90% | 98.16% | 99.91% | 99.83% | 99.90% | 98.99% | 15min | 19hour23min |
| | BGISEQ-500 | 3178160 | 3035849 | 99.58% | 95.12% | 99.90% | 99.88% | 99.74% | 97.44% | 15min | 15hour8min |
| | HiSeq2000 | 3182948 | 3015606 | 99.73% | 94.49% | 99.91% | 99.90% | 99.82% | 97.12% | 15min | 20h13min |

Note: All timings reported herein were performed on iGenomeComputer HIX.

**Table 3. SNP Calling Performance on NA12878**

| | Sequencing platform | True Postive | | Sensitivity | | Precision | | F1-SCORE | | Time | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 | GVC | GATK4 |
| sINDEL | NovaSeq | 367109 | 361917 | 98.44% | 97.20% | 99.37% | 99.52% | 98.90% | 98.34% | 15min | 19hour23min |
| | BGISEQ-500 | 355143 | 312639 | 95.38% | 83.96% | 99.30% | 99.28% | 97.30% | 90.98% | 15min | 15hour8min |
| | HiSeq2000 | 347845 | 276158 | 93.42% | 74.17% | 99.34% | 99.69% | 96.29% | 85.05% | 15min | 20h13min |

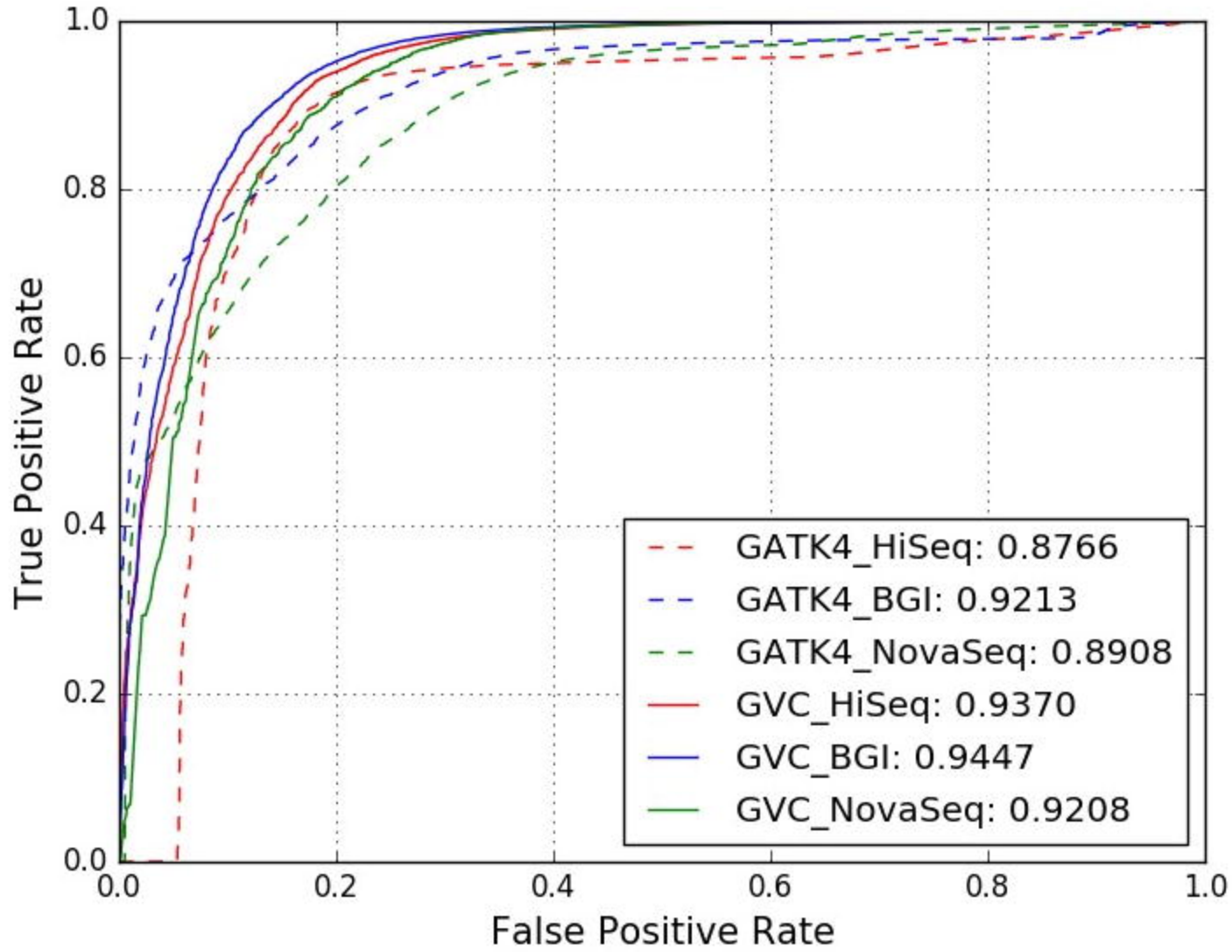
Note: All timings reported herein were performed on iGenomeComputer HIX.

**Table 4. sINDEL Calling Performance on NA12878**

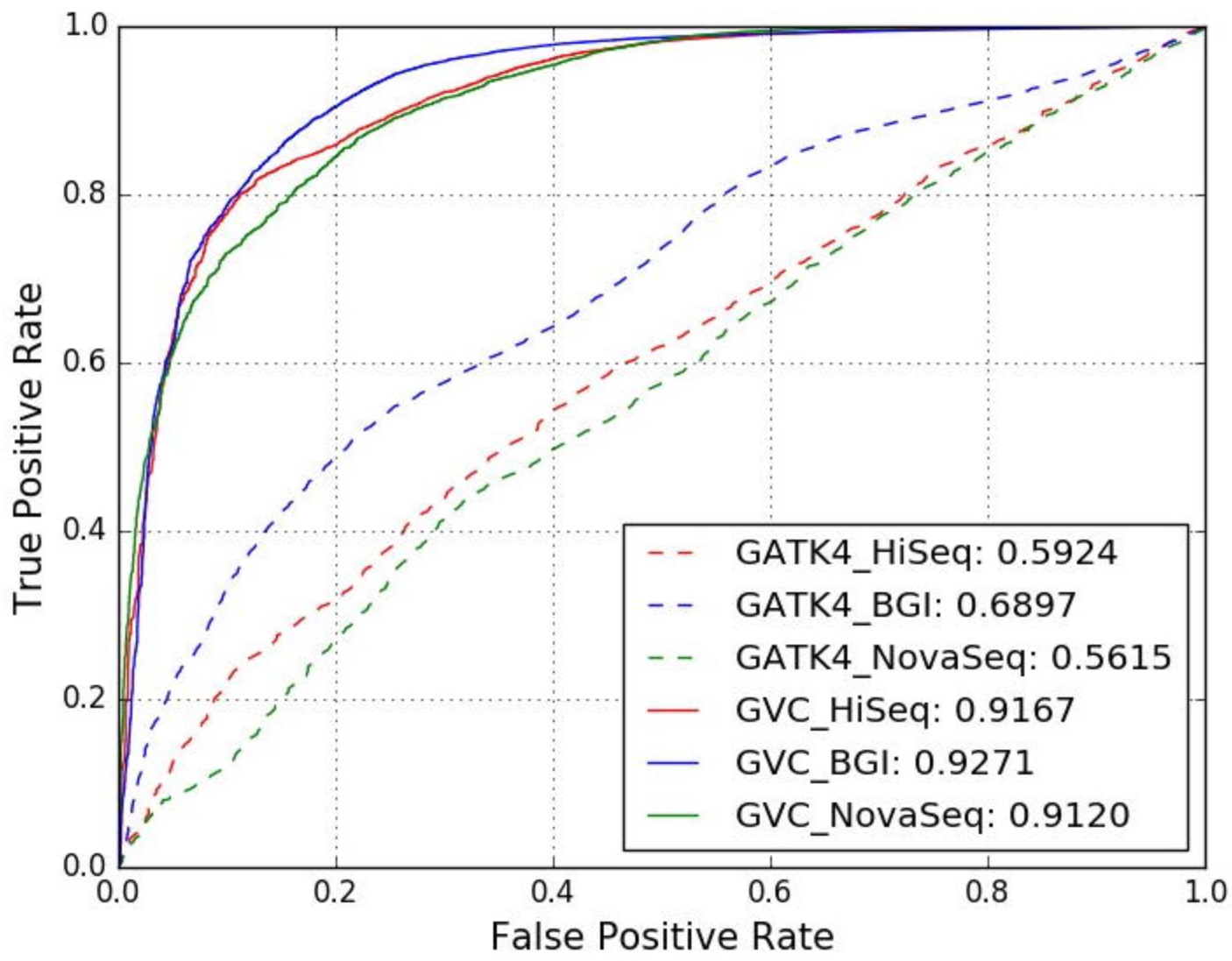
| Sequencing platform | break point that supported by at least two software | DELLY | SEEKSV | GVC |
|---|---|---|---|---|
| NovaSeq | 5765 | 79.46% | 82.48% | 85.10% |
| BGISEQ-500 | 4466 | 76.60% | 80.68% | 93.93% |
| HiSeq2000 | 4940 | 73.70% | 85.69% | 90.93% |

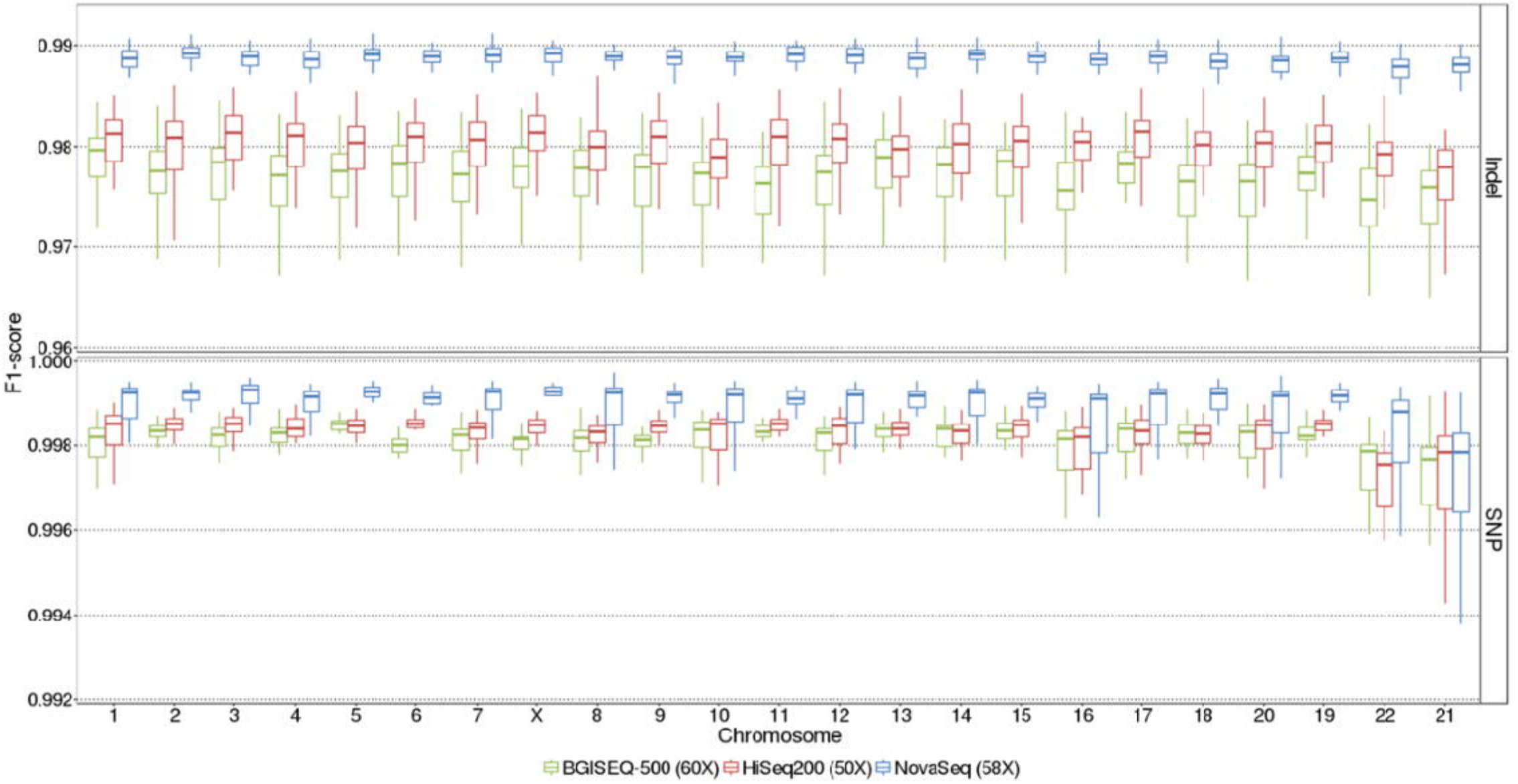
**Table 5. Somatic SV Calling Performance**

a


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
GATK4_HiSeq: 0.8766
GATK4_BGI: 0.9213
GATK4_NovaSeq: 0.8908
GVC_HiSeq: 0.9370
GVC_BGI: 0.9447
GVC_NovaSeq: 0.9208

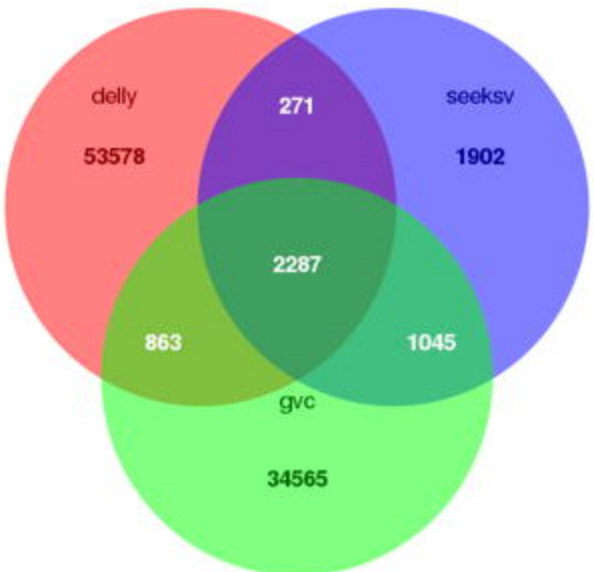
b


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
GATK4_HiSeq: 0.5924
GATK4_BGI: 0.6897
GATK4_NovaSeq: 0.5615
GVC_HiSeq: 0.9167
GVC_BGI: 0.9271
GVC_NovaSeq: 0.9120

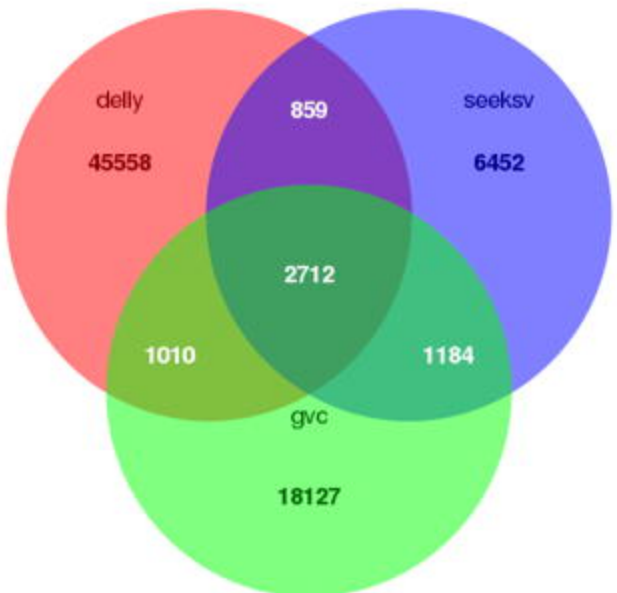

F1-score
0.99
0.98
0.97
0.96
Indel
1.000
0.998
0.996
0.994
0.992
SNP
1
2
3
4
5
6
7
X
8
9
10
11
12
13
14
15
16
17
18
20
19
22
21
Chromosome
BGISEQ-500 (60X)
HiSeq200 (50X)
NovaSeq (58X)

SNP and sINDEL calling performance on Hiseq NA12877
( model based on NA12878 )
100%
99%
98%
97%
96%
95%
94%
93%
92%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
InDel
SNP

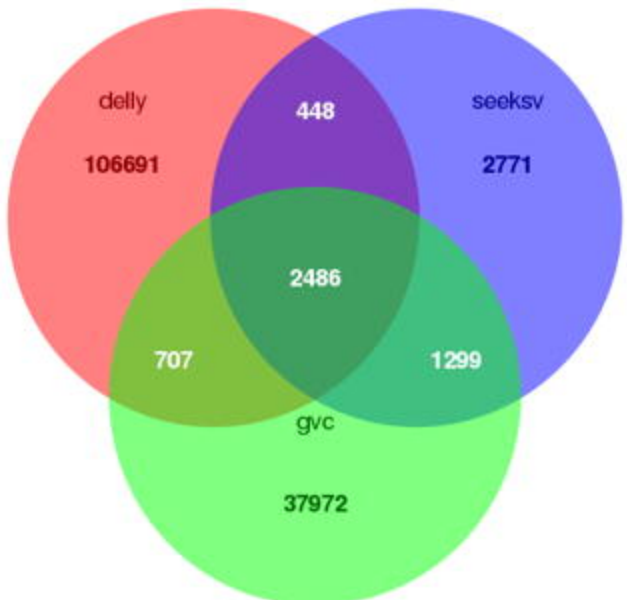
## BGI

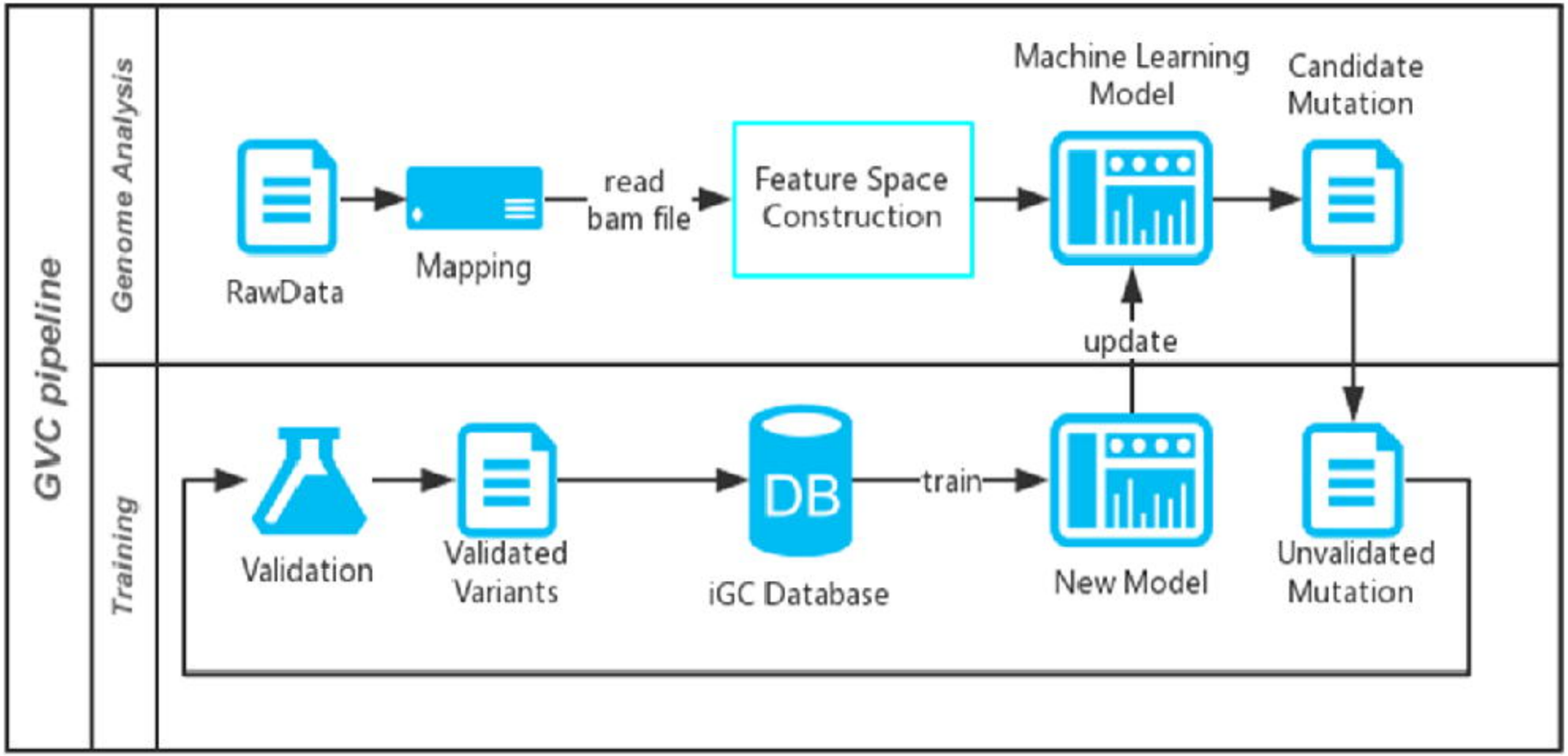



## NovaSeq




## Hiseq




GVC pipeline
Genome Analysis
Training
RawData
Mapping
read
bam file
Feature Space
Construction
Machine Learning
Model
Candidate
Mutation
update
Validation
Validated
Variants
DB
iGC Database
train
New Model
Unvalidated
Mutation