

به نام خدا

مسئله ۳

برای انتخاب ابزار شبیه سازی از مقاله "A broad survey of DNA sequence data simulation tools" که در سال ۲۰۱۹ چاپ شده است استقاده کردیم و با توجه به تقسیم بندی ای که در این مقاله به صورت شکل زیر بود، تصمیم گرفتیم که illumina را شبیه سازی کنیم و همینطور با جدولی که روش های state-of-the-art را مشخص کرده بود، تصمیم گرفتیم که اشتراک این دو مجموعه را بررسی کنیم. بنابراین ابزارها به ابزارهای شبیه سازی NGS و state-of-the-art کاهش یافتد. سپس با مقایسه این ابزارها در نهایت ابزار GemSim انتخاب شد.

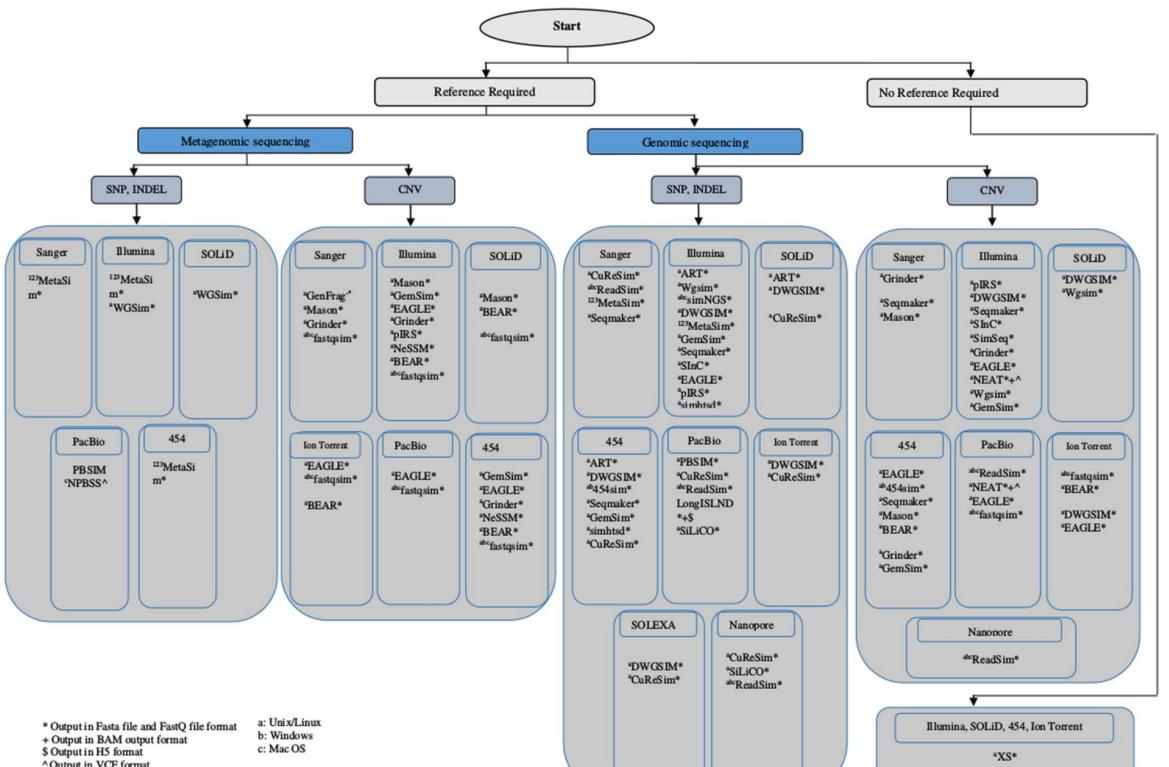


Table 1. Evaluating 20 state-of-art DNA sequence simulation tools that use reference genome and can generate read files in standard fastq files format based on sensitivity and precision in producing reads

Year	Tools	Sensitivity (%)	Precision (%)	Running time (h:min:s)
2009	WGSim* [26]	77.01%	78.6%	3:04:22
	DWGSim* [30]	76.37%	79.2%	3:14:09
2012	GemSim [15]	72.55%	73.1%	1:54:19
	EAGLE* [7]	66.28%	69.8%	1:40:01
2013	pIRS [34]	69.12%	74.4%	1:14:42
	Wessim [36]	86.41%	87.1%	6:23:03
2014	BEAR [38]	73.38%	74.8%	3:24:13
	CuReSim [39]	70.24%	73.4%	4:00:22
2016	FastQSim [16]	74.71%	76.1%	2:01:12
	ReadSim* [19]	76.91%	85.7%	5:41:43
2017	XS [40]	78.67%	82.7%	5:13:04
	SInC [13]	71.10%	76.9%	3:34:12
2016	LongISLND [43]	71.33%	73.3%	3:34:42
	NEAT [21]	85.62%	86.2%	5:52:01
2017	SILICO [18]	75.36%	77.6%	3:24:02
	CapSim [8]	72.81%	74.8%	1:44:07
2017	LRSim [46]	74.13%	74.53%	1:44:22
	Gargammel [47]	72.52%	74.2%	2:04:02
2017	NanoSim [48]	70.91%	78.1%	3:11:01
	Pysim-sv [51]	61.12%	72.2%	2:04:52

زمان اجرا، دقت و سال انتشار در تصویر بالا مشخص شده است.

نکات	دایکیومنت مناسب	تعداد ارجاع به مقاله	
زمان اجرای بالا، نسبتاً قدمی	ندارد	-	Wgsim
زمان اجرای بالا، پشتیبانی کوتاه مدت	دارد	-	DWGSIM
تولید فقط با رفرنس ژنوم	دارد (بسیار مناسب)	۱۷۲	GemSim
زمان اجرای بالا	ندارد	۵۶	SInC
دقت نسبتاً پایین	دارد	-	EAGLE
دقت مناسب و زمان اجرای پایین	دارد	۱۹۱	pIRS

با استفاده از ابزار GemSim و با استفاده از دستور زیر

```
python GemSIM_v1.6/GemReads.py -r Sample.fa -n 750002 -l 100 -u d -m
```

```
GemSIM_v1.6/models/ill100v5_p.gzip -q 64 -o sample_read -p
```

عملیات شبیه سازی را انجام می دهیم در اینجا برای داشتن پوشش برابر با 30x به ۲۵۰۰۰۲ و در مجموع ۱۵۰۰۰۴ read نیاز داریم. این عدد نیز به این صورت به دست آمدہ است که در ابزار دیگری که پارامتری برای coverage داشت، با پوشش 30x به تعداد گفتشده خوانش تولید کرد.

با توجه به اینکه ابزار GemSim با زبان پایتون ۲ توسعه داده شده است تنها کافی است که به طور مثال یک virtual

environment با پایتون ۲ بسازیم و از این ابزار در آن محیط استفاده کنیم.

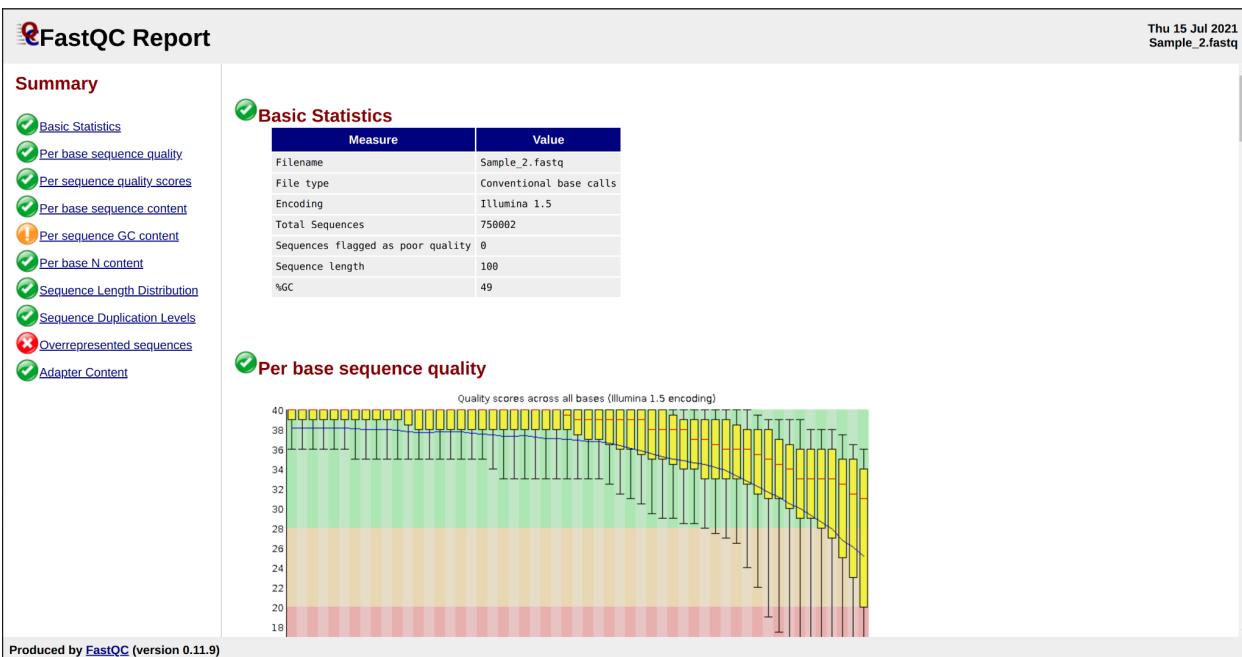
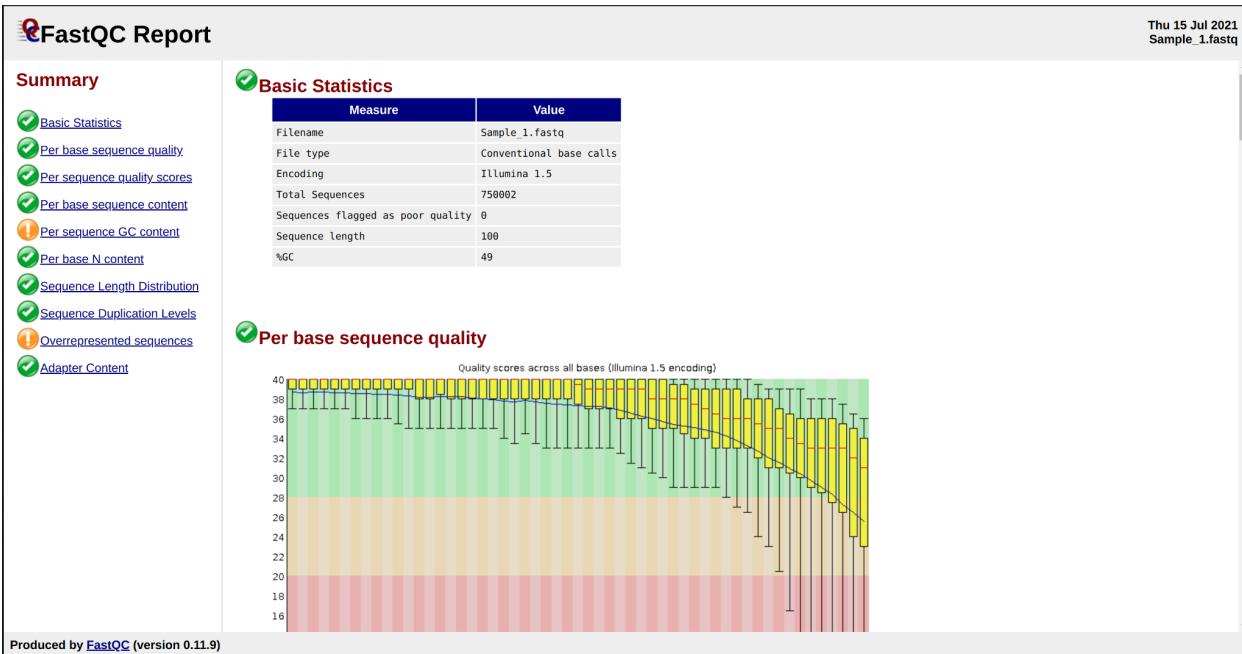
سپس دو فایل با فرمت fastq به صورت زیر در اختیار ما قرار می دهد.

Plain Text ▾ Tab Width: 8 ▾ Ln 311554, Col 99 ▾ INS

Plain Text ▾ Tab Width: 8 ▾

به دلیل حجم بالای این دو فایل، آن‌ها را آیلود نخواهیم کرد.

پس از تولید دو فایل fastq، با ابزار fastqc، خوانش‌های paired-end را کنترل کیفیت می‌کنیم. خروجی fastqc در زیر آمده است.



در گزارش‌های تولیده شده، مقدار GC-Content برابر با ۴۹ درصد گزارش شده است که نزدیک به ۵۰ درصد است و این مقدار قابل قبولی است.

(۳)

حال مانند ویدیو کارگاه نیاز است که در ابتدا از ژنوم مرجع، index و dictionary ساخته شود که همانطور که کد آن‌ها در فایل، آمده است. این دو را ساختیم با استفاده از دو ابزار bwa و picard. سپس با ابزار bwa خوانش‌ها را به ژنوم مرجع align می‌کنیم و فایل alignment.sam را تولید می‌کنیم. سپس نیاز است که این فایل مرتب بشود که با ابزار picard این کار انجام می‌شود. سپس از ابزار mark duplicate، picard، mark duplicate انجام می‌دهیم. که بخش‌هایی که تکراری هستند علامت میخورند ولی حذف نمی‌شوند. سپس این فایل به دست آمده را با picard گروه‌بندی می‌کنیم. سپس از فایل bam به دست آمده، با استفاده از ابزار picard اقدام به تولید index می‌کنیم. سپس با ۳ ابزار گفته شده عملیات

و ساخت فایل‌های vcf را انجام می‌دهیم که در اینجا فایل samtools variant calling حجم بسیار بالای دارد و به همین دلیل آن را نیز آپلود نمی‌کنیم. کدهای استفاده شده به صورت زیر هستند.

```

1 #####create index from reference genome#####
2 bwa index chr19.fa
3 #####create dictionary from reference genome#####
4 java -jar ../../../../../../SetupFiles/picard.jar CreateSequenceDictionary REFERENCE=chr19.fa OUTPUT=chr19.dict
5 #####align reads to the reference genome#####
6 bwa mem -t 2 -a -M chr19.fa ./Part2/Sample_1.fastq ./Part2/Sample_2.fastq > alignment.sam
7 #####sort the sam file#####
8 java -jar ../../../../../../SetupFiles/picard.jar SortSam I=alignment.sam O=alignment.bam SORT_ORDER=coordinate
9 #####mark duplicates#####
10 java -jar ../../../../../../SetupFiles/picard.jar MarkDuplicates INPUT=alignment.bam OUTPUT=mdup_alignment.bam METRICS_FILE=alignment.metrics
11 #####group reads in the mark duplicated bam file#####
12 java -jar ../../../../../../SetupFiles/picard.jar AddOrReplaceReadGroups INPUT=mdup_alignment.bam OUTPUT=group_mdup_alignment.bam RGID="alignment" RGLB="Exome"
13 #####create index from group duplicated bam file#####
14 java -jar ../../../../../../SetupFiles/picard.jar BuildBamIndex I=group_mdup_alignment.bam
15 #####create index from reference genome#####
16 samtools faidx chr19.fa
17 #####gatk4 variant calling#####
18 java -jar ../../../../../../SetupFiles/gatk-4.2.0.0-gatk-package-4.2.0.0-local.jar HaplotypeCaller -R chr19.fa -I group_mdup_alignment.bam -O gatk4.vcf
19 #####samtools variant calling#####
20 samtools mpileup -uf chr19.fa group_mdup_alignment.bam > samtools.bcf
21 bcftools view samtools.bcf > samtools.vcf
22 #####freebayes variant calling#####
23 freebayes -f chr19.fa group_mdup_alignment.bam > freebayes.vcf
24
25
26 #source=HaplotypeCaller

```

و همینطور فایل‌های vcf به دست آمده نیز به صورت زیر هستند.

GATK vcf

Description="RMS Mapping Quality">', and '#INFO=Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities"'. The FORMAT column shows AC, AF, and other quality scores. The file is 901 lines long."/>

```

gatk4.vcf /home/mahdi/Work/code/MBDA/HW1/Q3/Part3 - Geany
File Edit Search Document Project Build Tools Help
File Edit Search Document Project Build Tools Help
Symbols freebayes.vcf gatk4.vcf
No symbols found
18 #INFO=Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the sample."
19 #INFO=Description="RMS Mapping Quality">
20 #INFO=Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities"">
21 #INFO=Description="Variant Confidence/Quality Depth"
22 #INFO=Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias"
23 #INFO=Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias"">
24 #contig=

23:01:56: This is Geany 1.36.  
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/freebayes.vcf opened (1).  
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/gatk4.vcf opened (2).



Status


```

samtools vcf

```

samtools.vcf - /home/mahdi/Work/code/MBDA/HW1/Q3/Part3 - Geany
File Edit Search View Document Project Build Tools Help
Symbols freebayes.vcf gatk4.vcf samtools.vcf
No symbols found
9 ##INFO<=ID=IV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
10 ##INFO<=ID=IM,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
11 ##INFO<=ID=DP,Number=1,Type=Integer,Description="Read depth">
12 ##INFO<=ID=SPB,Number=1,Type=Float,Description="Strand Position Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
13 ##INFO<=ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
14 ##INFO<=ID=MOB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
15 ##INFO<=ID=BQ,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
16 ##INFO<=ID=MOSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
17 ##INFO<=ID=MQF,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
18 ##INFO<=ID=I16,Number=16,Type=Float,Description="Auxiliary tag used for calling, see description of bcf_callret1_t in bam2bcf.h">
20 ##INFO<=ID=OS,Number=R,Type=Float,Description="Auxiliary tag used for calling">
22 ##FORMAT<=ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
23 ##FORMAT<=ID=Comma,Number=1,Type=String,Description="Comma-delimited list of variants for each sample">
Date:Sun May 30 22:06:32 2021
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Hiseq4000
chr19 10839289 . T <=> 0 . DP=1;ID=1,0,0,0,36,1296,0,0,66,3660,0,0,0,0,0,0;OS=1,0;MQF=0 PL 0,3,36
chr19 10839290 . A <=> 0 . DP=1;ID=1,0,0,0,36,2704,0,0,66,3660,0,0,1,1,0,0;OS=1,0;MQF=0 PL 0,3,52
chr19 10839291 . A <=> 0 . DP=1;ID=1,0,0,0,64,4096,0,0,66,3660,0,0,2,4,0,0;OS=1,0;MQF=0 PL 0,3,68
chr19 10839292 . G <=> 0 . DP=1;ID=1,0,0,0,64,3660,0,0,66,3660,0,0,3,6,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839293 . T <=> 0 . DP=1;ID=1,0,0,0,64,5941,0,0,66,3660,0,0,4,16,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839294 . T <=> 0 . DP=1;ID=1,0,0,0,64,4996,0,0,66,3660,0,0,5,25,0,0;OS=1,0;MQF=0 PL 0,3,68
chr19 10839295 . T <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,6,36,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839296 . C <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,49,0,0;OS=1,0;MQF=0 PL 0,3,68
chr19 10839297 . T <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,56,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839298 . G <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,9,81,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839299 . T <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,18,100,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839300 . G <=> 0 . DP=1;ID=1,0,0,0,70,4900,0,0,66,3660,0,0,11,21,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839301 . C <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,12,144,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839302 . T <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,14,16,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839303 . C <=> 0 . DP=1;ID=1,0,0,0,71,5941,0,0,66,3660,0,0,15,225,0,0;OS=1,0;MQF=0 PL 0,3,60
chr19 10839304 . A <=> 0 . DP=2;ID=2,0,0,0,107,6337,0,0,120,7200,0,0,16,256,0,0;OS=1,0;MQF=0 PL 0,6,90
chr19 10839305 . A <=> 0 . DP=2;ID=2,0,0,0,107,6337,0,0,120,7200,0,0,18,256,0,0;OS=1,0;MQF=0 PL 0,6,103
chr19 10839306 . G <=> 0 . DP=2;ID=2,0,0,0,123,7745,0,0,120,7200,0,0,18,256,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839307 . T <=> 0 . DP=2;ID=2,0,0,0,123,7745,0,0,120,7200,0,0,18,300,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839308 . T <=> 0 . DP=2;ID=2,0,0,0,126,7956,0,0,120,7200,0,0,22,270,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839309 . T <=> 0 . DP=2;ID=2,0,0,0,131,8621,0,0,120,7200,0,0,24,416,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839310 . G <=> 0 . DP=2;ID=2,0,0,0,132,8762,0,0,120,7200,0,0,26,466,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839311 . G <=> 0 . DP=2;ID=2,0,0,0,136,8482,0,0,120,7200,0,0,28,520,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839312 . G <=> 0 . DP=2;ID=2,0,0,0,137,8762,0,0,120,7200,0,0,30,578,0,0;OS=1,0;MQF=0 PL 0,6,110
chr19 10839313 . G <=> 0 . DP=2;ID=2,0,0,0,132,8762,0,0,120,7200,0,0,32,640,0,0;OS=1,0;MQF=0 PL 0,6,110
23:01:56: This is Geany 1.36.
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/freebayes.vcf opened (1).
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/gatk4.vcf opened (2).
23:03:15: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/samtools.vcf opened (3).
Status
line: 22 / 5005125 col: 45 set: 1 INS TAB mode: LF encoding: UTF-8 filetype: None scope: unknown

```

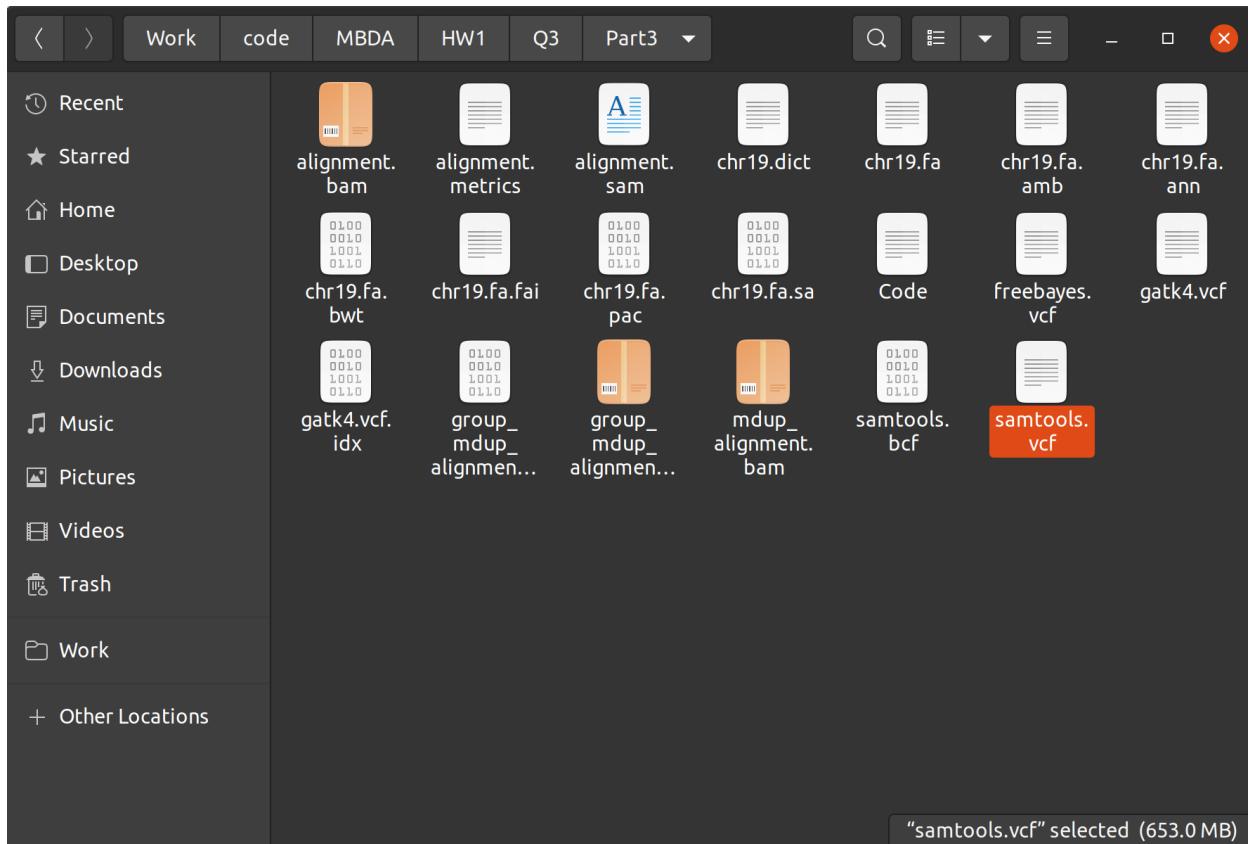
Freebayes vcf

```

freebayes.vcf - /home/mahdi/Work/code/MBDA/HW1/Q3/Part3 - Geany
File Edit Search View Document Project Build Tools Help
Symbols freebayes.vcf gatk4.vcf samtools.vcf
No symbols found
57 ##FORMAT<=ID=NR,Number=1,Type=Integer,Description="Reference allele observation count">
58 ##FORMAT<=ID=OR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
59 ##FORMAT<=ID=A0,Number=A,Type=Integer,Description="Alternate allele observation count">
60 ##FORMAT<=ID=AO,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">
61 ##FORMAT<=ID=DP,Number=1,Type=Intger,Description="Mean depth in gvcf output block.">
62 #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Hiseq4000
63 chr19 10839945 . T G 3.1160e-14 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=37;DPR=0;EPP=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
64 chr19 108404466 . A C 2.94216e-07 . AB=0;0.0740741;ABP=45,5551;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=27;DPR=0;EPP=3.0103;EPR=7.26639;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
65 chr19 108404472 . A C 1.65293e-08 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
66 chr19 108404486 . A C 7.64682e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
67 chr19 108404486 . A C 8.31026e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
68 chr19 108404487 . T G 3.48296e-14 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
69 chr19 108404502 . T G 6.10481e-08 . AB=0;0.089505;ABP=40,8993;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
70 chr19 108404505 . T G 6.10481e-08 . AB=0;0.089505;ABP=40,8993;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=22;DPR=22;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
71 chr19 108404477 . T G 3.21936e-12 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=7.35324;EPRR=3.0103;EPRR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
72 chr19 108404477 . T G 9.42393e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=7.35324;EPRR=3.0103;EPRR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
73 chr19 1084045682 . C T 1.45772e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=22;DPR=22;EPP=7.35324;EPRR=3.0103;EPRR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
74 chr19 1084047417 . A G 0 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=0;EPP=7.06999;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NUMALT=1;
75 chr19 1084047647 . A C 2.75661e-12 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=0;EPP=7.06999;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NUMALT=1;
76 chr19 1084047652 . C A 2.95252e-09 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=0;EPP=7.06999;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
77 chr19 1084047659 . C A 9.05399e-08 . AB=0;0.0714286;ABP=7.6806;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=28;DPR=28;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
78 chr19 1084048338 . A C 1.96858e-14 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=28;DPR=28;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
79 chr19 1084048897 . A G 6.26888e-12 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=33;DPR=33;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
80 chr19 1084049596 . A G 1.42133e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=31;DPR=31;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
81 chr19 1084049606 . A G 9.42393e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=31;DPR=31;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
82 chr19 1084049924 . A G 9.03339e-11 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=39;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
83 chr19 1084050275 . T G 9.78689e-15 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=38;DPR=38;EPP=7.35324;EPRR=3.07234;GTI=0;LEN=1;MEANALT=2;MOM=60;MOM=60;NS=1;
84 chr19 1084050564 . T G 5.52344e-15 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=36;DPR=36;EPP=3.0103;EPR=7.06999;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NU=1;
85 chr19 1084050887 . A C 3.45581e-06 . AB=0;1.11111;ABP=0;AC=0;AF=0;CIGAR=1X;DP=27;DPR=27;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NU=1;
86 chr19 10851234 . A C 4.16163e-09 . AB=0;0.0740741;ABP=45,5551;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=27;DPR=27;EPP=3.0103;EPR=7.26639;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
87 chr19 108524732 . A C 2.46989e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=24;DPR=24;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
88 chr19 108524735 . C A 2.46989e-13 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=24;DPR=24;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
89 chr19 108524743 . A C 2.72428e-10 . AB=0;.0625;ABP=56,2114;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=32;DPR=32;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
90 chr19 10853174 . A C 7.85595e-09 . AB=0;0.0714286;ABP=48,6806;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=28;DPR=28;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
91 chr19 10854192 . A G 0 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=39;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NU=1;
92 chr19 10854194 . A G 0 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=39;DPR=39;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;NU=1;
93 chr19 10854689 . A C 2.86001e-12 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=32;DPR=32;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
94 chr19 10854896 . A C 6.225594e-14 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=32;DPR=32;EPP=3.0103;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
95 chr19 10854997 . A C 0 .000398939 . AB=0;1.42857;ABP=26,2761;AC=1;AF=0.5;AN=2;AO=3;CIGAR=1X;DP=21;DPR=21;EPP=3.07342;EPR=7.35324;GTI=0;LEN=1;MEANALT=1;MOM=60;NS=1;
96 chr19 10855133 . T C 1.40545e-10 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=26;DPR=26;EPP=3.07324;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=1;MOM=60;MOM=60;NS=1;
97 chr19 10856540 . A C 3.51875e-10 . AB=0;ABP=0;AC=0;AF=0;CIGAR=1X;DP=36;DPR=36;EPP=3.07342;EPR=7.35324;EPRR=3.56868;GTI=0;LEN=1;MEANALT=2;MOM=60;MOM=60;NS=1;
23:01:56: This is Geany 1.36.
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/freebayes.vcf opened (1).
23:01:56: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/gatk4.vcf opened (2).
23:03:15: File /home/mahdi/Work/code/MBDA/HW1/Q3/Part3/samtools.vcf opened (3).
Status
line: 1 / 13576 col: 0 sel: 0 INS TAB mode: LF encoding: UTF-8 filetype: None scope: unknown

```

فایلهای ساخته شده در این مرحله به صورت زیر هستند که به دلیل حجم بالای برخی از آنها، آنها را آپلود نمی‌کنیم. به طور مثال همانطور که در تصویر دیده می‌شود، حجم فایل ساخته شده با samtools vcf حدود ۶۰۰ مگابایت است.



در نهایت با مقایسه ۳ فایل **vcf** ساخته شده متوجه می‌شویم که هر سه ابزار کروموزوم را به درستی شناسایی کردند و ابزار **gatk** تغییرات بسیار بیشتری را گزارش می‌دهد نسبت به دو روش دیگر و شاید به این دلیل باشد که به طور مثال ابزار **gatk** در ابتدا خوانش‌های با کیفیت پایین را دور میریزد ولی ابزار **samtools** به صورت پیش فرض تمامی خوانش‌ها را بررسی می‌کند. ابزار **gatk** برای ژنوم انسان نتایج بهتری را تولید می‌کند نسبت به **samtools**.