Masters Thesis on

Benchmarking Somatic Variant Callers

By R. S. Chintalapati

29th April 2021

Declaration

Abstract

Variant calling pipelines are of two types namely Germline and Somatic. In terms of benchmarking, many Germline variant calling pipelines were compared and documented by the Genome in a Bottle (GIAB) Consortium standards. For Somatic variants calling pipelines, only a few comparisons were achieved because the comparisons are challenging and less established. In the diploid human genome, a variant can be found either homozygously, heterozygously or not at all and these three levels complicate the comparison process. Tumors, on the other hand, are inhomogeneous cells that might carry variants possibly with rare mutations not found in the others making the Somatic variant calling pipelines benchmarking challenging. The goal of this thesis is to consider an artificial dataset pair from the GIAB project and benchmarking two Somatic variant callers based on a few factors to choose a variant caller for cancer research.

Chapter 1

Introduction

Chapter 2

Related Works

Chapter 3

Background

Chapter 4

Approach

Chapter 5

Experiments

To compare the Strelka, VarScan and Truth VCF files, the first step is calculating the number of positions, SNPs and Indels. This step helps in finding out the common positions, SNPs and Indels amongst the VCF files and even helps in normalising the variants in the second step.

In the second step, to know the bias of the variant callers, each variant type is counted and normalised. Through this step, information about a variant caller showing a specific bias in calling variant combinations quite often can be known in comparison to the truth data.

In the third step, to observe genetic diversity, allele frequencies between the truth data and the Strelka and VarScan VCF files are compared to learn which variant caller is comparably close to the truth data and which variant caller calls the reads with higher allele frequencies.

In the fourth step, to know the number of times a nucleotide has been read, the read depth of each VCF file is compared. This step reveals the coverage of a read in every variant caller while providing an idea as to which variant caller has the better coverage.

In the fifth and final step, ALT variants in each of the VCF files are compared for benchmarking. Through this step, the truth data is compared with the Strelka VCF file and VarScan VCF file individually to know the number of true positives, true negatives, false positives and false negatives.

5.1 Positions, SNPs & Indels

A single-nucleotide polymorphism is a substitution of a single nucleotide at a specific position in the genome that is present in a sufficiently large fraction of the population. Indel is a molecular biology term for an insertion or deletion of bases in the genome of an organism. In this section, Positions, SNPs and INDELs from different somatic variant callers are compared with the artifical truth data obtained from https://ftp-trace.ncbi.nlm.nih.gov/

Tumor Purity	**Positions**	**SNPs**	**Indels**
0.3	13,315	12,897	418
0.5	13,315	12,897	418
0.7	13,315	12,897	418

Table 5.1: Values from Strelka VCF files

Tumor Purity	**Positions**	**SNPs**	**Indels**
0.3	29,316	26,312	3,004
0.5	29,294	26,290	3,004
0.7	29,171	26,185	2,986

Table 5.2: Values from VarScan VCF files

Tumor Purity	**Positions**	**SNPs**	**Indels**
1.0	11,04,786	10,07,793	96,993

Table 5.3: Values from Truth Data VCF files

5.2 Variants

A variant is an alteration in the most common DNA sequence. In this section, variants from different somatic variant callers are compared with the artifical truth data obtained from https://ftp-trace.ncbi.nlm.nih.gov/

Type	Positions
Strelka	13,315
VarScan	29,316
Truth Data	11,04,786
Strelka & VarScan	3,104
VarScan & Truth Data	2,843
Strelka & Truth Data	3,791
Strelka, VarScan & Truth Data	1,052

Table 5.4: Positions comparison with Tumor Purity 0.3

Type	Positions
Strelka	13,315
VarScan	29,294
Truth Data	11,04,786
Strelka & VarScan	3,096
VarScan & Truth Data	2,843
Strelka & Truth Data	3,791
Strelka, VarScan & Truth Data	1,052

Table 5.5: Positions comparison with Tumor Purity 0.5

Type	Positions
Strelka	13,315
VarScan	29,171
Truth Data	11,04,786
Strelka & VarScan	3,051
VarScan & Truth Data	2,843
Strelka & Truth Data	3,791
Strelka, VarScan & Truth Data	1,052

Table 5.6: Positions comparison with Tumor Purity 0.7

5.3 Allele Frequencies

Allele frequency, or gene frequency is defined as the relative frequency of an allele at a particular locus in a population, expressed as a fraction or

Type	SNPs
Strelka	12,897
VarScan	26,312
Truth Data	10,07,793
Strelka & VarScan	2,778
VarScan & Truth Data	2,484
Strelka & Truth Data	3,150
Strelka, VarScan & Truth Data	875

Table 5.7: SNPs comparison with Tumor Purity 0.3

Type	SNPs
Strelka	12,897
VarScan	26,290
Truth Data	10,07,793
Strelka & VarScan	2,770
VarScan & Truth Data	2,484
Strelka & Truth Data	3,150
Strelka, VarScan & Truth Data	875

Table 5.8: SNPs comparison with Tumor Purity 0.5

Type	SNPs
Strelka	12,897
VarScan	26,185
Truth Data	10,07,793
Strelka & VarScan	2,729
VarScan & Truth Data	2,484
Strelka & Truth Data	3,150
Strelka, VarScan & Truth Data	875

Table 5.9: SNPs comparison with Tumor Purity 0.7

percentage. In this section, allele frequencies from different somatic variant callers are compared with the artifical truth data obtained from https://ftp-trace.ncbi.nlm.nih.gov/

Type	Indels
Strelka	418
VarScan	3,004
Truth Data	96,993
Strelka & VarScan	110
VarScan & Truth Data	200
Strelka & Truth Data	127
Strelka, VarScan & Truth Data	29

Table 5.10: Indels comparison with Tumor Purity 0.3 & 0.5

Type	Indels
Strelka	418
VarScan	2,986
Truth Data	96,993
Strelka & VarScan	107
VarScan & Truth Data	200
Strelka & Truth Data	127
Strelka, VarScan & Truth Data	29

Table 5.11: Indels comparison with Tumor Purity 0.7

Figure 5.1 shows Strelka Allele Frequency Counts.
Figure 5.2 shows VarScan Allele Frequency Counts.

5.4 Read Depth

Read Depth describes the number of times that a given nucleotide in the genome has been read in an experiment. In this section, read depths from different somatic variant callers are compared with the artifical truth data obtained from https://ftp-trace.ncbi.nlm.nih.gov/

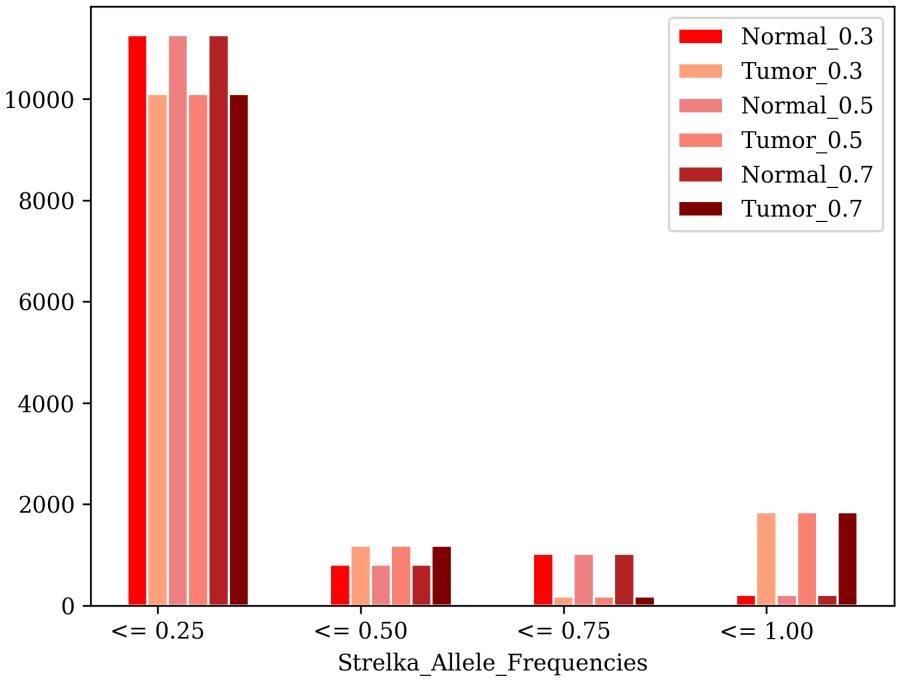

Figure 5.1: Strelka Allele Frequency Counts

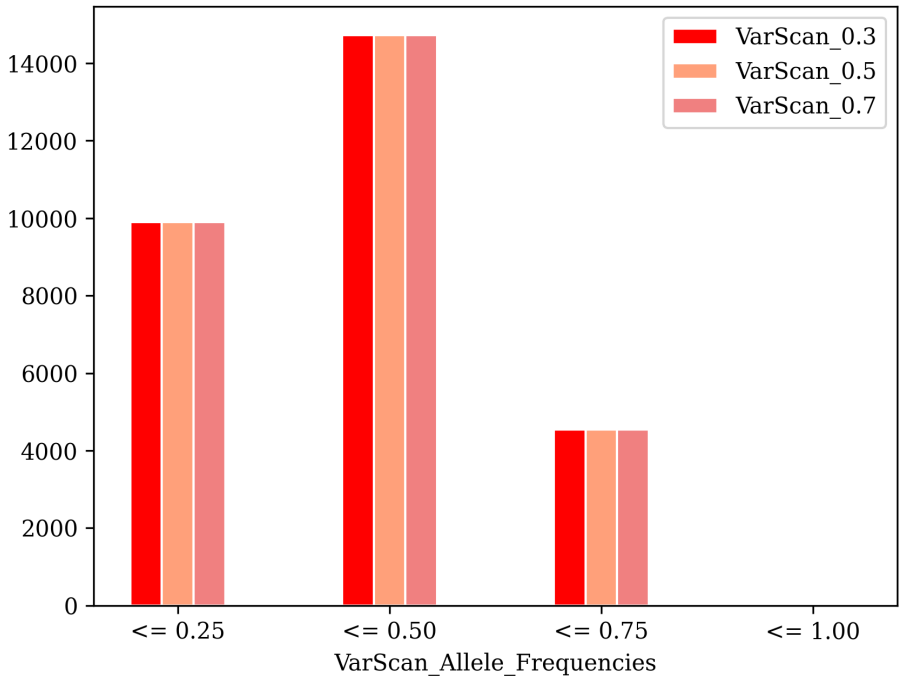

Figure 5.2: VarScan Allele Frequency Counts

Combination	**Tumor Purity 0.3**	**Tumor Purity 0.5**	**Tumor Purity 0.7**
AA	0	0	0
AT	393	393	393
AG	1,196	1,196	1,196
AC	1,942	1,942	1,942
TT	0	0	0
TA	647	647	647
TG	1,491	1,491	1,491
TC	1,245	1,245	1,245
GG	0	0	0
GA	1,700	1,700	1,700
GT	882	882	882
GC	478	478	478
CC	0	0	0
CA	1,072	1,072	1,072
CT	1,441	1,441	1,441
CG	409	409	409

Table 5.12: Variant counts in Strelka variant caller

Combination	**Tumor Purity 0.3**	**Tumor Purity 0.5**	**Tumor Purity 0.7**
AA	0	0	0
AT	662	662	662
AG	4,354	4,354	4,350
AC	962	958	947
TT	0	0	0
TA	722	720	715
TG	990	989	976
TC	4,343	4,342	4,336
GG	0	0	0
GA	4,960	4,959	4,950
GT	969	966	949
GC	1,212	1,212	1,210
CC	0	0	0
CA	1,059	1,054	1,033
CT	4,843	4,840	4,829
CG	1,235	1,233	1,231

Table 5.13: Variant counts in VarScan variant caller

Combination	**Tumor Purity 1.0**
AA	0
AT	32,639
AG	1,52,935
AC	38,384
TT	0
TA	32,817
TG	37,707
TC	1,53,474
GG	0
GA	1,91,603
GT	44,538
GC	43,851
CC	0
CA	44,253
CT	1,91,442
CG	44,049

Table 5.14: Variant counts in Truth Data

Combination	**Strelka**	**VarScan**	**Truth Data**
AA	0	0	0
AT	3.04	2.51	3.23
AG	9.27	16.54	15.17
AC	15.05	3.65	3.80
TT	0	0	0
TA	5.01	2.74	3.25
TG	11.56	3.76	3.74
TC	9.65	16.50	15.22
GG	0	0	0
GA	13.18	18.85	19.01
GT	6.83	3.68	4.41
GC	3.70	4.60	4.35
CC	0	0	0
CA	8.31	4.02	4.39
CT	11.17	18.40	18.99
CG	3.17	4.69	4.37

Table 5.15: Normalised variant counts with Tumor Purity 0.3

Combination	**Strelka**	**VarScan**	**Truth Data**
AA	0	0	0
AT	3.04	2.51	3.23
AG	9.27	16.56	15.17
AC	15.05	3.64	3.80
TT	0	0	0
TA	5.01	2.73	3.25
TG	11.56	3.76	3.74
TC	9.65	16.51	15.22
GG	0	0	0
GA	13.18	18.86	19.01
GT	6.83	3.67	4.41
GC	3.70	4.61	4.35
CC	0	0	0
CA	8.31	4.00	4.39
CT	11.17	18.41	18.99
CG	3.17	4.69	4.37

Table 5.16: Normalised variant counts with Tumor Purity 0.5

Combination	**Strelka**	**VarScan**	**Truth Data**
AA	0	0	0
AT	3.04	2.51	3.23
AG	9.27	16.61	15.17
AC	15.05	3.61	3.80
TT	0	0	0
TA	5.01	2.73	3.25
TG	11.56	3.72	3.74
TC	9.65	16.55	15.22
GG	0	0	0
GA	13.18	18.90	19.01
GT	6.83	3.62	4.41
GC	3.70	4.62	4.35
CC	0	0	0
CA	8.31	3.94	4.39
CT	11.17	18.44	18.99
CG	3.17	4.70	4.37

Table 5.17: Normalised variant counts with Tumor Purity 0.7

Format	**Purity**	**$\leq$ 0.25**	**0.25 $>$ & $\leq$ 0.50**	**0.50 $>$ & $\leq$ 0.75**	**$>$ 0.75**
Normal	0.3	11,266	809	1,020	212
Tumor	0.3	10,095	1,185	177	1,845
Normal	0.5	11,266	809	1,020	212
Tumor	0.5	10,095	1,185	177	1,845
Normal	0.7	11,266	809	1,020	212
Tumor	0.7	10,095	1,185	177	1,845

Table 5.18: Allele Frequencies count from Strelka variant caller VCF files

Purity	**$\leq$ 0.25**	**0.25 $>$ & $\leq$ 0.50**	**0.50 $>$ & $\leq$ 0.75**	**$>$ 0.75**
0.3	9,893	14,732	4,546	0
0.5	9,893	14,732	4,546	0
0.7	9,893	14,732	4,546	0

Table 5.19: Allele Frequencies count from VarScan variant caller VCF files

$\leq$ **0.25**	**0.25** $>$ & $\leq$ **0.50**	**0.50** $>$ & $\leq$ **0.75**	$>$ **0.75**
11,266	809	1,020	212

Table 5.20: Allele Frequencies count from Truth Data VCF file

Type	$\leq$ **0.25**	**0.25** $>$ & $\leq$ **0.50**	**0.50** $>$ & $\leq$ **0.75**	$>$ **0.75**
Strelka Normal	469	561	2	0
Strelka Tumor	1,032	0	0	0
VarScan	0	0	1,032	0
Truth Data	0	1,032	0	0

Table 5.21: Allele Frequencies count with Tumor Purity of 0.3, 0.5 & 0.7

Format	**Purity**	**Minimum**	**Maximum**	**Mean**	**Median**	**Mode**
Normal	0.3	1	299	58.32	53	0 43
Normal	0.3	0	114	20.71	19	0 17
Normal	0.5	1	299	58.32	53	0 43
Tumor	0.5	0	114	20.71	19	0 17
Normal	0.7	1	299	58.32	53	0 43
Tumor	0.7	0	114	20.71	19	0 17

Table 5.22: Read Depth statistics from Strelka variant caller VCF files

Format	**Purity**	**Minimum**	**Maximum**	**Mean**	**Median**	**Mode**
Normal	0.3	10	99	∞	55	0 38
Normal	0.3	10	98	∞	19	0 17
Normal	0.5	10	99	∞	55	0 38
Tumor	0.5	10	98	∞	19	0 17
Normal	0.7	10	99	∞	55	0 38
Tumor	0.7	10	98	∞	19	0 17

Table 5.23: Read Depth statistics from VarScan variant caller VCF files

Minimum	Maximum	Mean	Median	Mode
100	999	∞	647	0 640

Table 5.24: Read Depth statistics from Truth Data VCF file

Type	Minimum	Maximum	Mean	Median	Mode
Strelka Normal	20	139	70.38	64	0 59
Strelka Tumor	14	61	26.61	25	0 19
VarScan Normal	16	117	55.11	50	0 41
VarScan Tumor	9	52	21.18	19	0 16
Truth Data	274	1,104	662.83	660	0 730

Table 5.25: Read Depth statistics from Tumor Purity 0.3, 0.5, & 0.7

Chapter 6

Conclusions

Chapter 7

Acknowledgments