UMI-Gen: a new UMI-based read simulator

for variant calling evaluation in paired-end sequencing

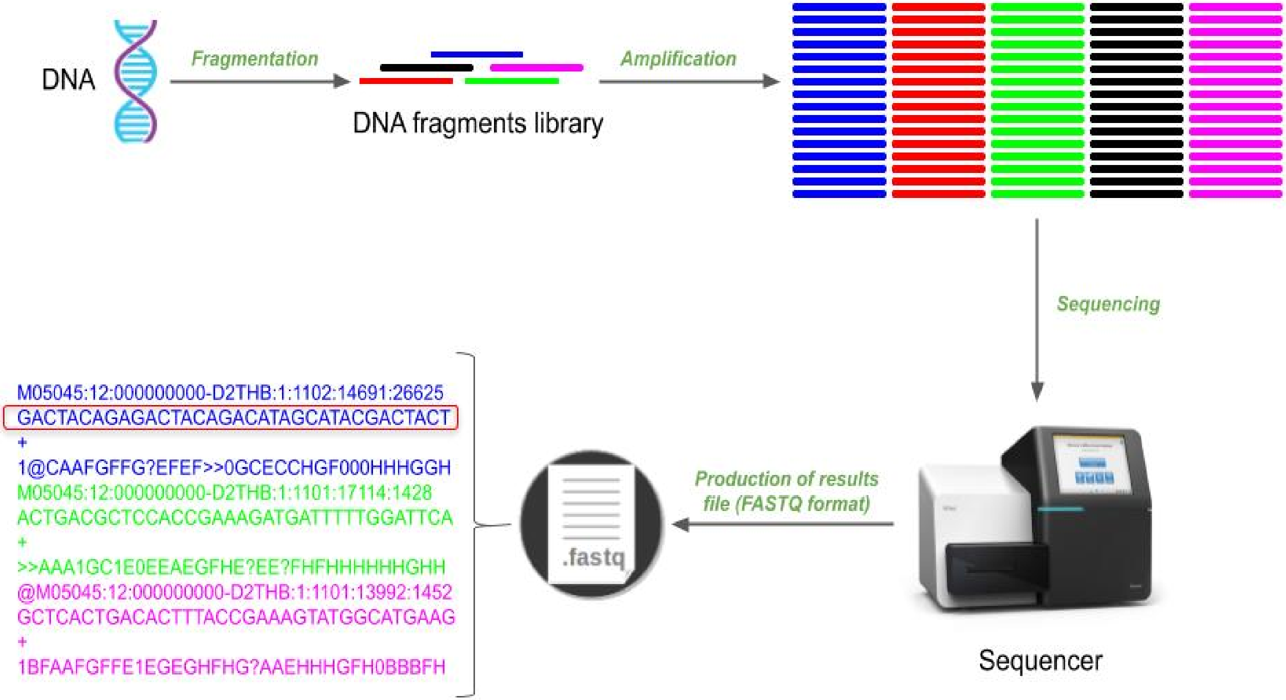
Vincent SATER

Thierry LECROQ, Pierre-Julien VIAILLY, Elise PRIEUR-GASTON,
Philippe RUMINY, Caroline BÉRARD and Fabrice JARDIN.

SeqBIM 2020

23 November 2020

Next Generation Sequencing (NGS)

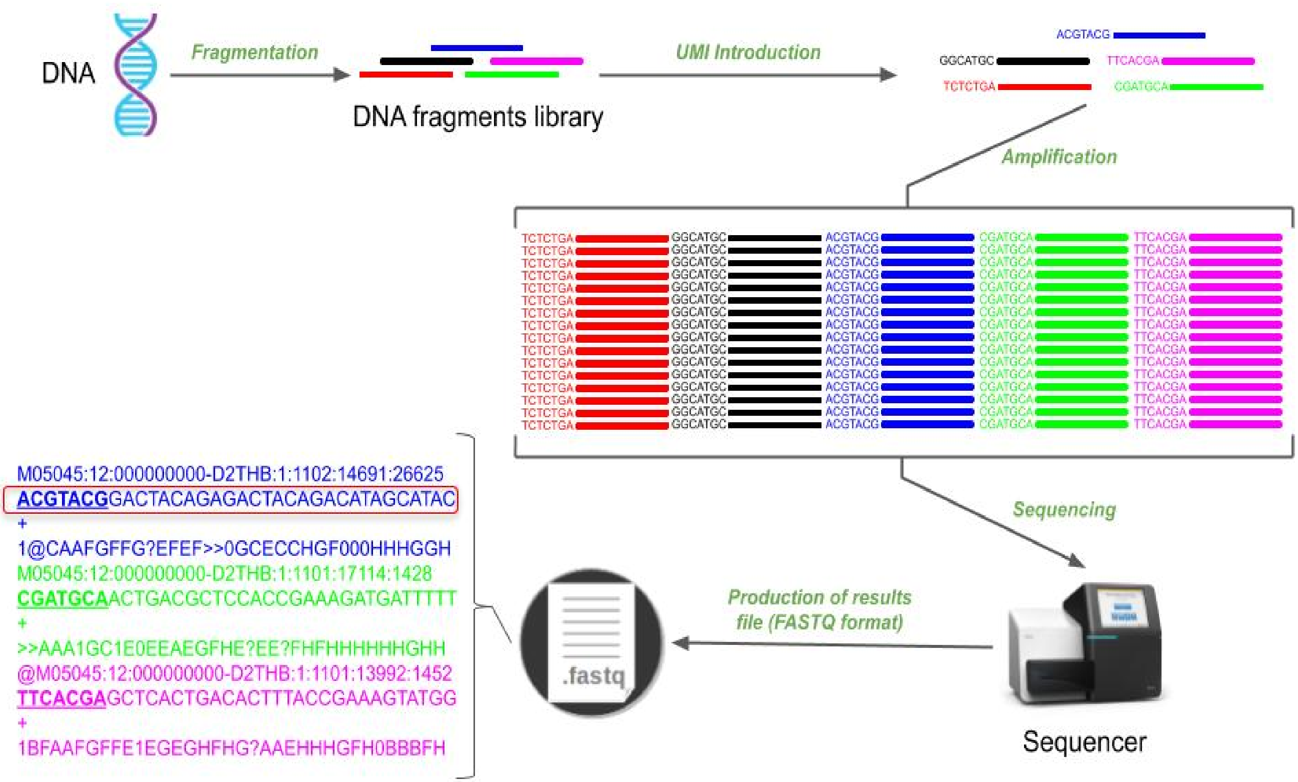


```
M05045:12:000000000-D2THB:1:1102:14691:26625
GACTACAGAGACTACAGACATAGCATACGACTACT
+
1@CAAFGFFG?EFEF>>0GCECCHGF000HHHGGH
M05045:12:000000000-D2THB:1:1101:17114:1428
ACTGACGCTCCACCGAAAGATGATTTTTGGATTCA
+
>>AAA1GC1E0EEAEGFHE?EE?FHFHHHHHHGHH
@M05045:12:000000000-D2THB:1:1101:13992:1452
GCTCACTGACACTTTACCGAAAGTATGGCATGAAG
+
1BFAAFGFFE1EGEGHFHG?AAEHHHGFH0BBBFH
```

What are UMIs ?

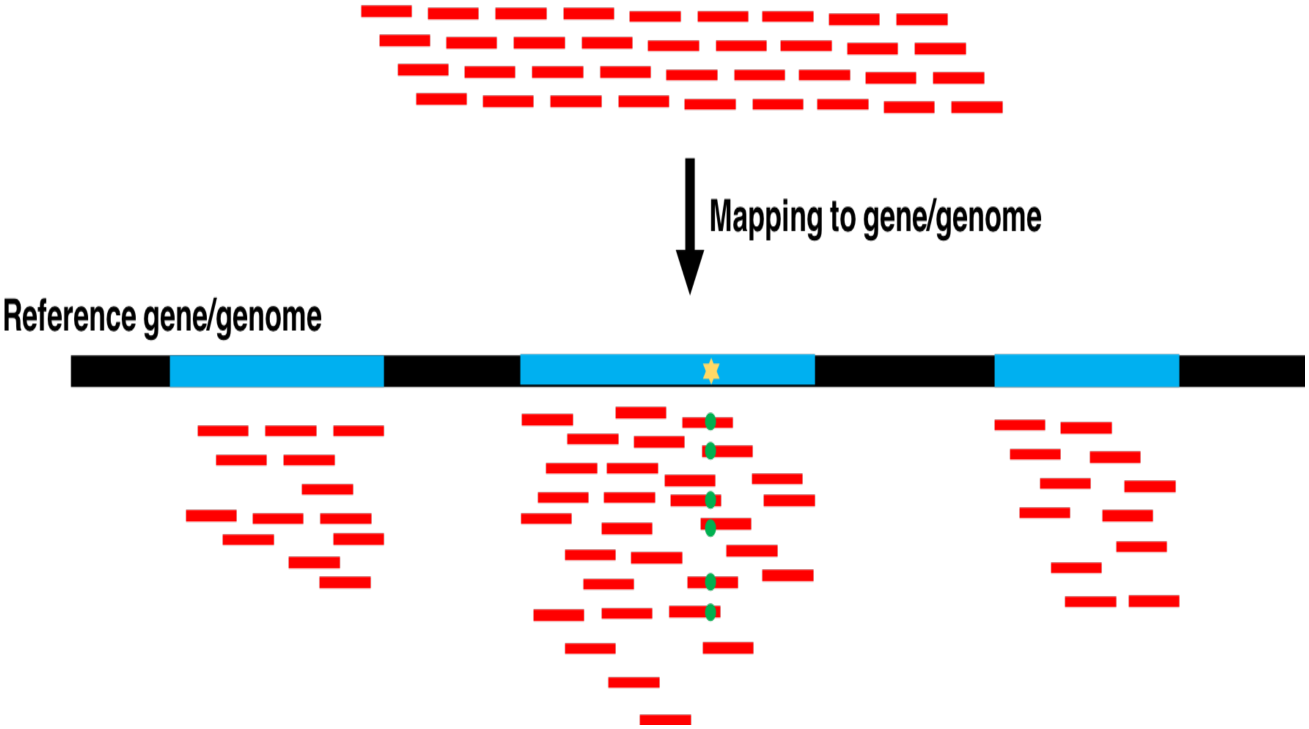
- Unique Molecular Identifier
- Short arbitrary nucleotide sequences
- Increased usage in Next Generation Sequencing since 2015

UMIs in NGS

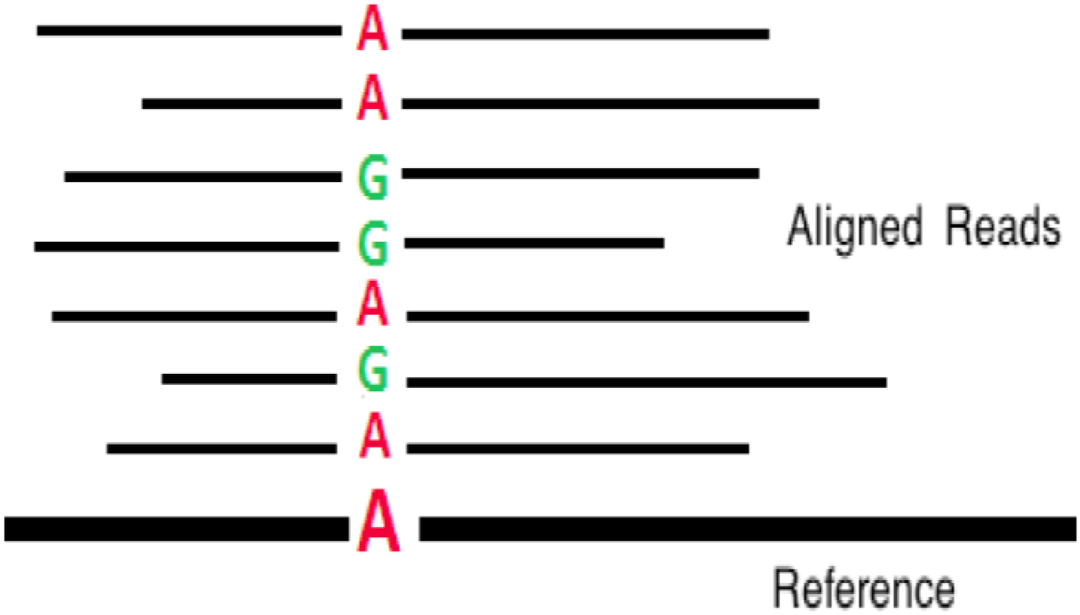


```
M05045:12:000000000-D2THB:1:1102:14691:26625
ACGTACGGACTACAGAGACTACAGACATAGCATAC
+
1@CAAFGFFG?EFEF>>0GCECCHGF000HHHGGH
M05045:12:000000000-D2THB:1:1101:17114:1428
CGATGCAACTGACGCTCCACCGAAAGATGATTTTT
+
>>AAA1GC1E0EEAEGFHE?EE?FHFHHHHHHGHH
@M05045:12:000000000-D2THB:1:1101:13992:1452
TTCACGAGCTCACTGACACTTTACCGAAAGTATGG
+
1BFAAFGFFE1EGEGHFHG?AAEHHHGFH0BBBFH
```

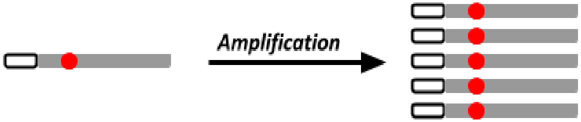
Aligning sequences to the reference genome



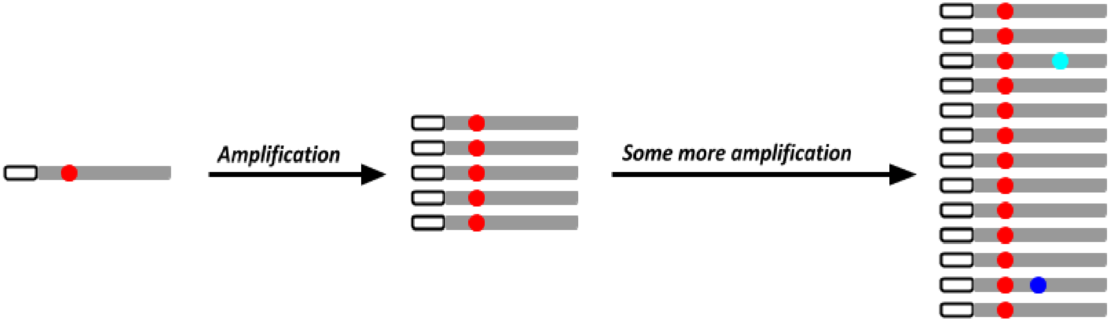
Variant Calling



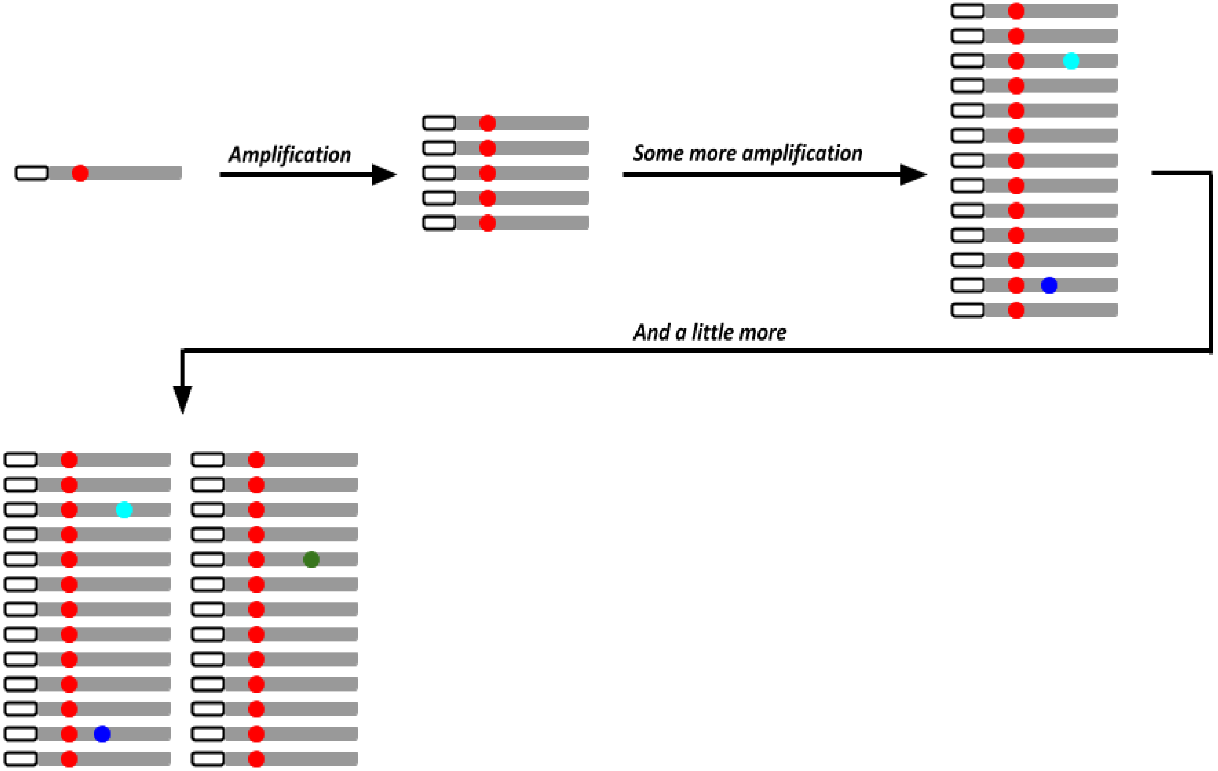
The theory



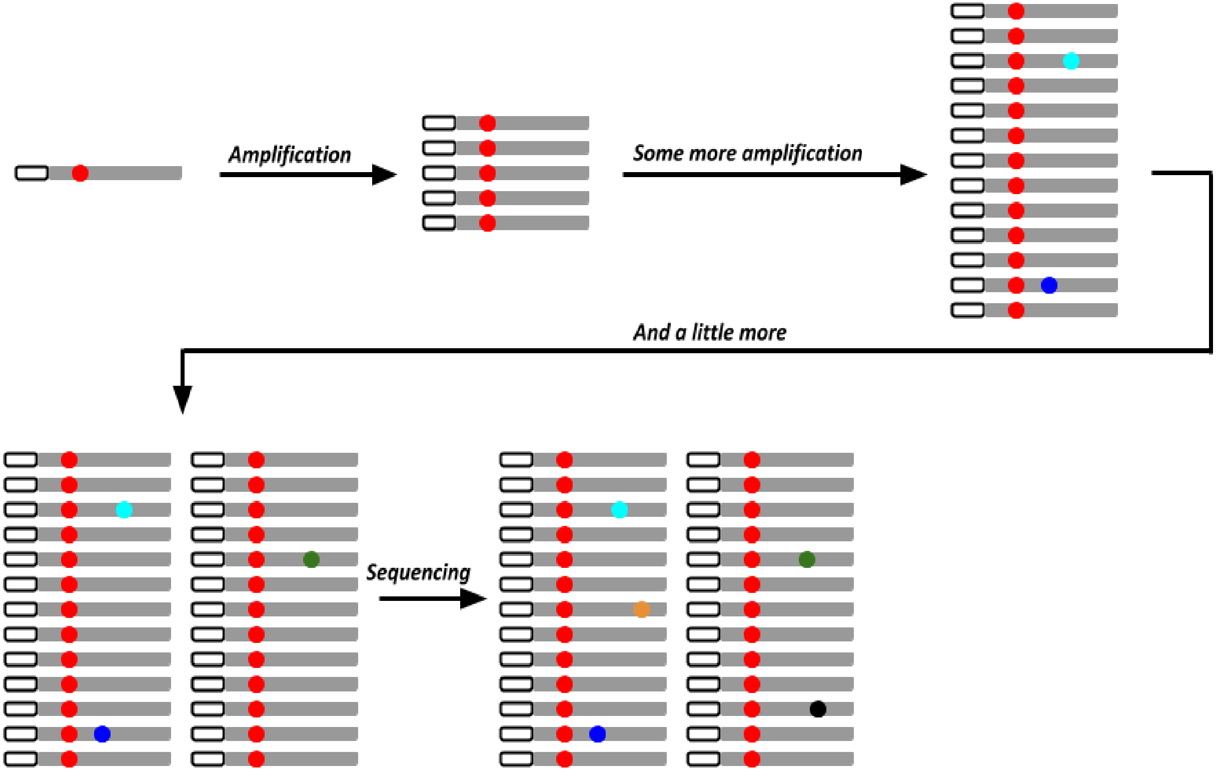
The theory



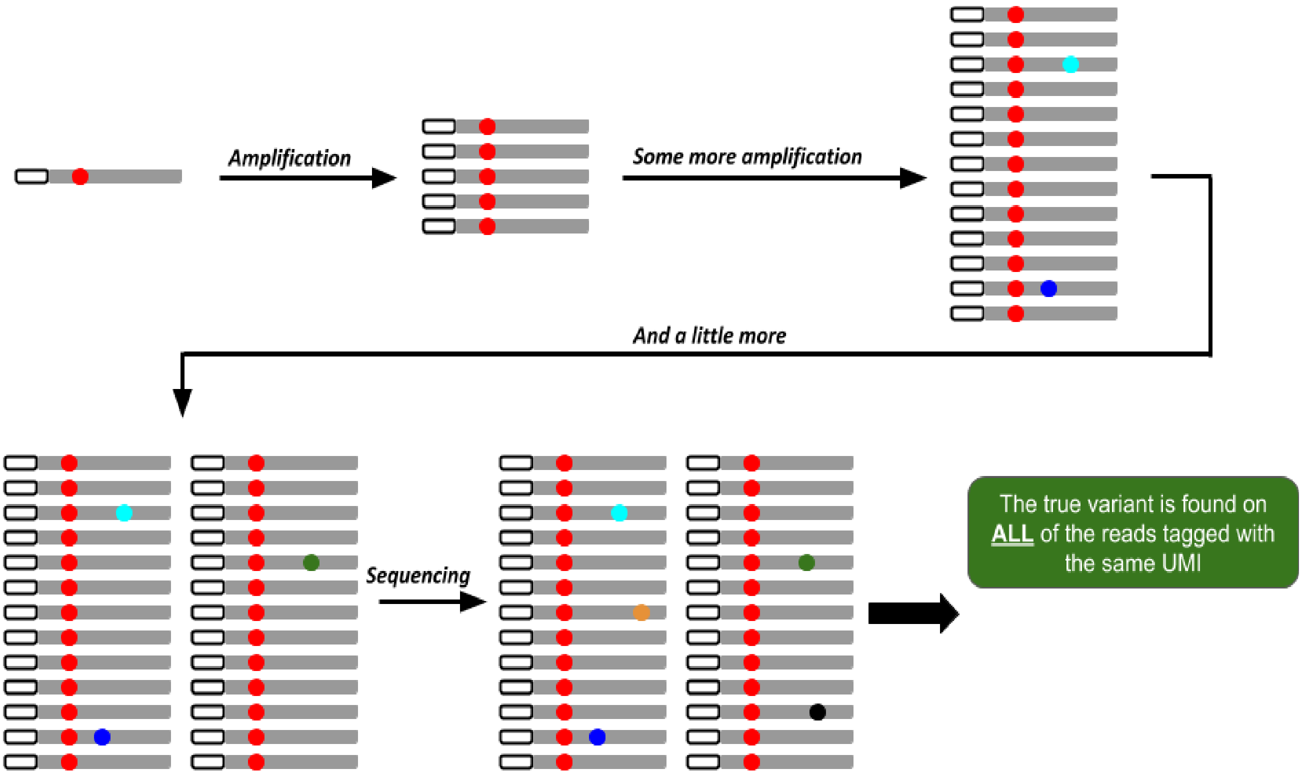
The theory



The theory

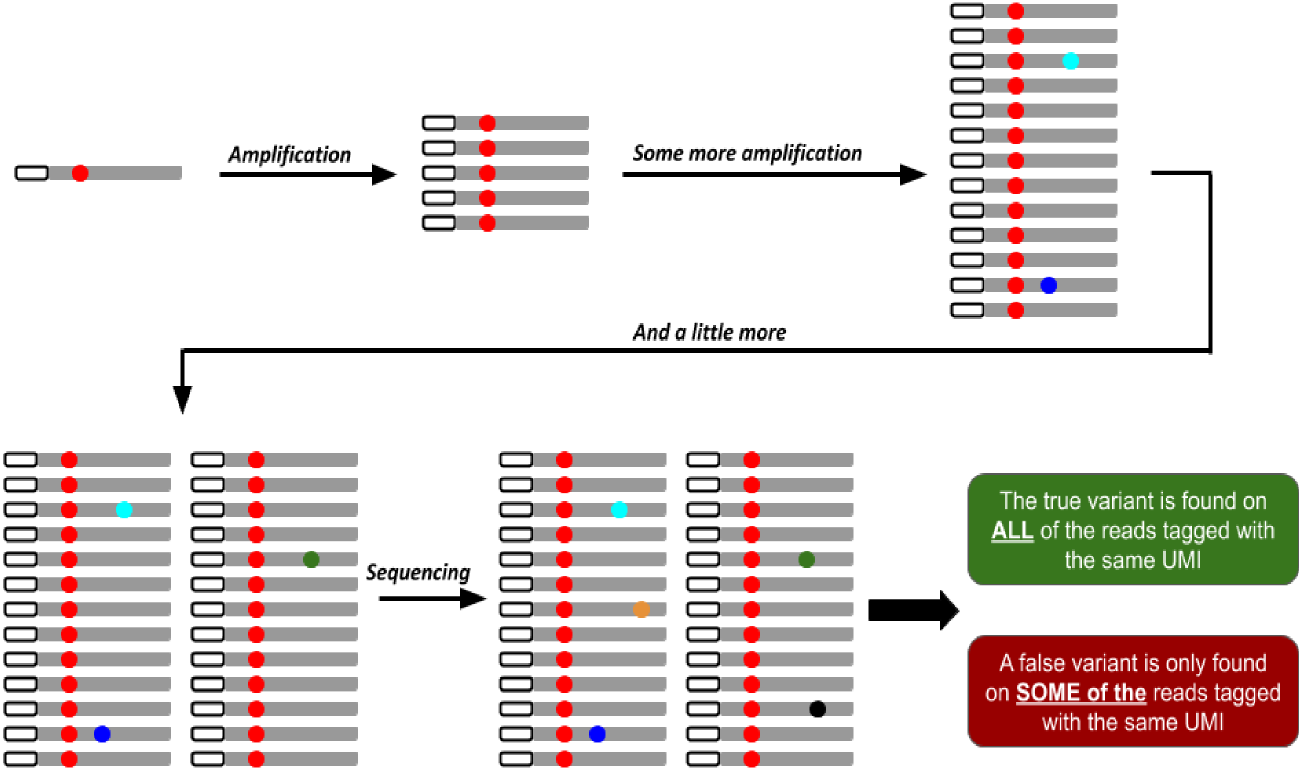

Amplification
Some more amplification
And a little more
Sequencing

The theory



The true variant is found on **ALL** of the reads tagged with the same UMI

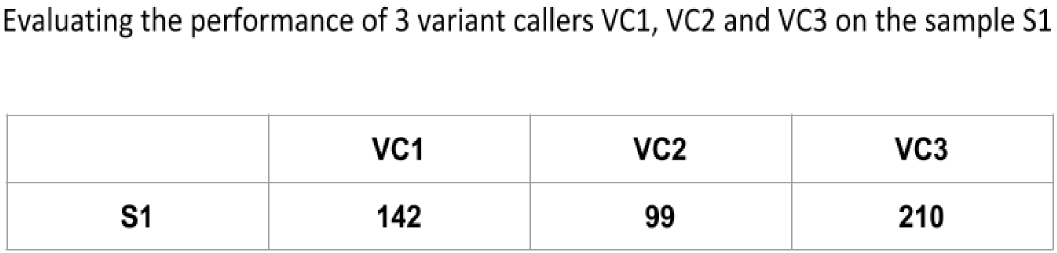
The theory



The true variant is found on **ALL** of the reads tagged with the same UMI

A false variant is only found on **SOME of the** reads tagged with the same UMI

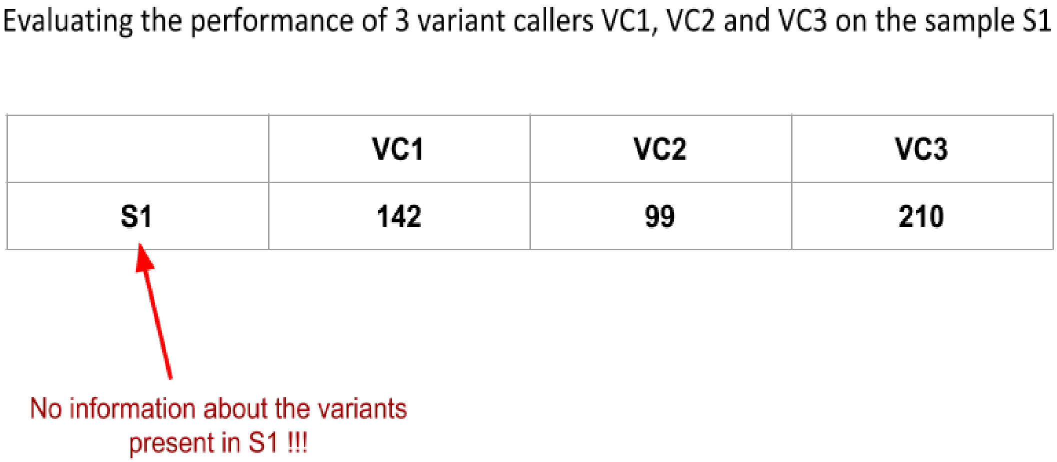
The problem



Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

	VC1	VC2	VC3
S1	142	99	210

The problem

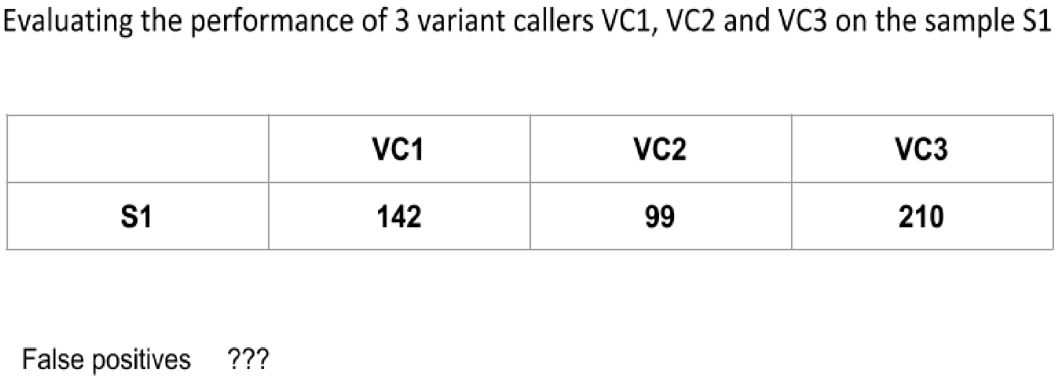


Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

	VC1	VC2	VC3
S1	142	99	210

No information about the variants present in S1 !!!

The problem

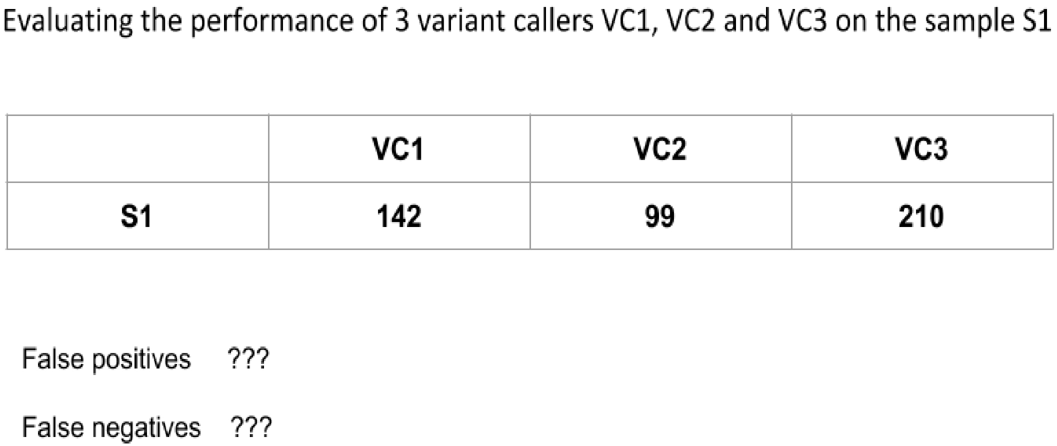


Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

	VC1	VC2	VC3
S1	142	99	210

False positives ???

The problem



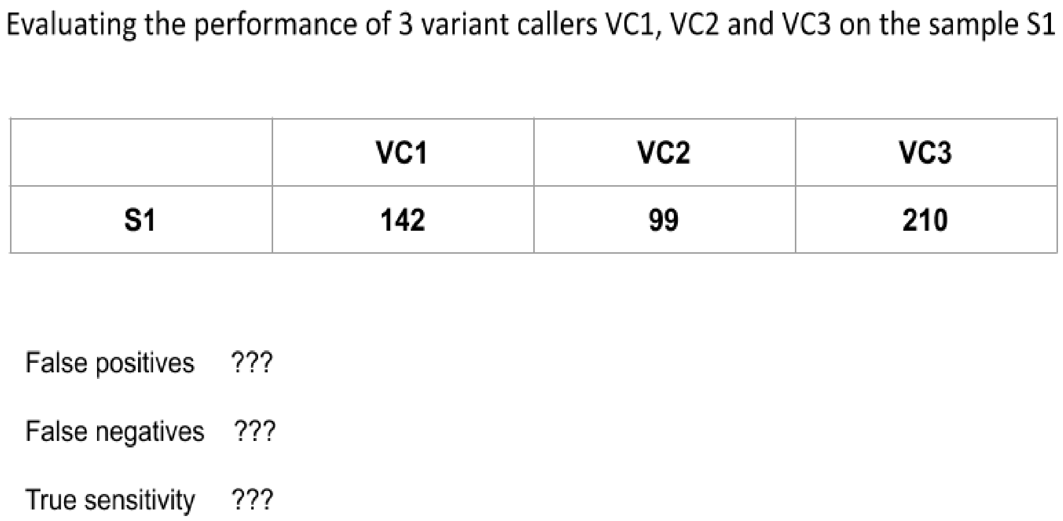
Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

	VC1	VC2	VC3
S1	142	99	210

False positives ???

False negatives ???

The problem



Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

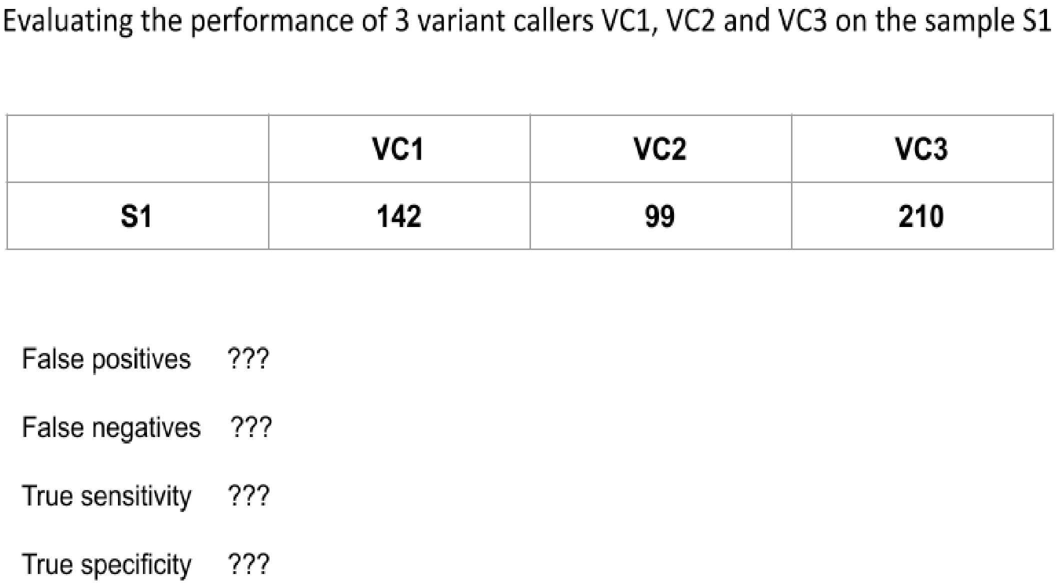
	VC1	VC2	VC3
S1	142	99	210

False positives ???

False negatives ???

True sensitivity ???

The problem



Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

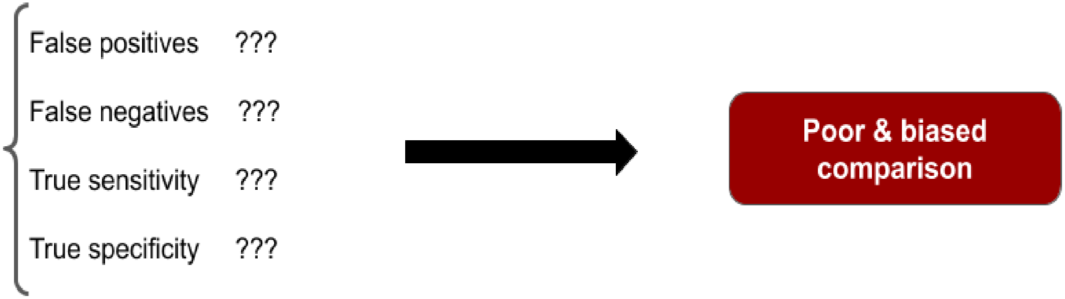
	VC1	VC2	VC3
S1	142	99	210

False positives	???
False negatives	???
True sensitivity	???
True specificity	???

The problem

Evaluating the performance of 3 variant callers VC1, VC2 and VC3 on the sample S1

	VC1	VC2	VC3
S1	142	99	210



False positives ???

False negatives ???

True sensitivity ???

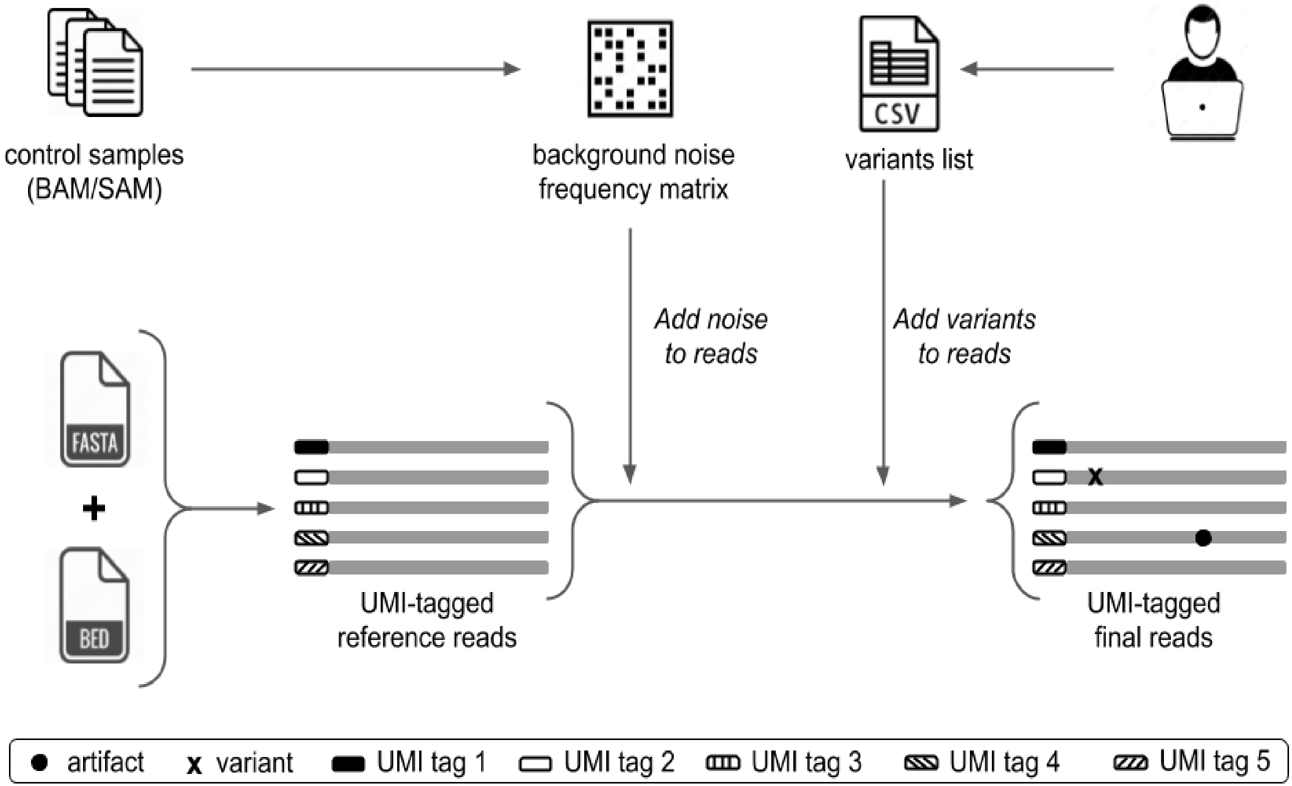
True specificity ???

Poor & biased comparison

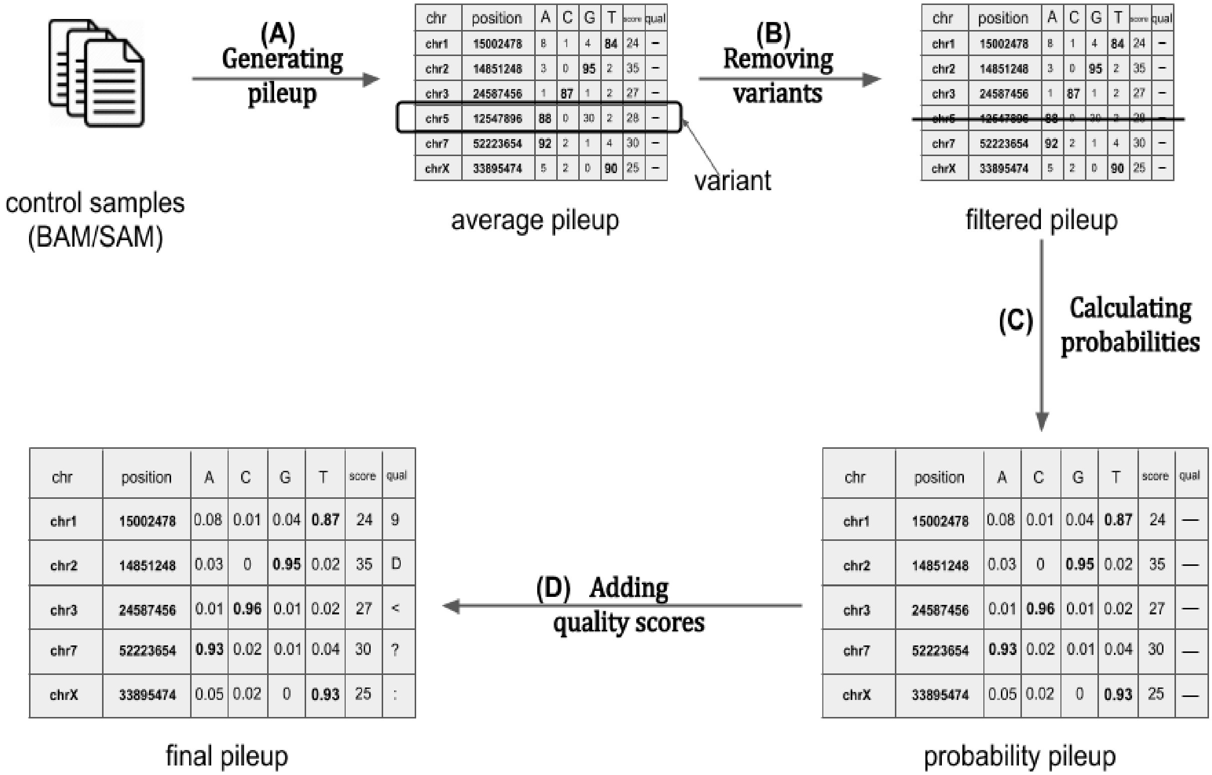
The solution

- Using a read simulator that mimic real tumor samples
- Real variants in the produced samples must be known
- No UMI-based read simulators available → we developed UMI-Gen

UMI-Gen's workflow



Background noise estimation



control samples (BAM/SAM)
(A) Generating pileup
(B) Removing variants
variant
(C) Calculating probabilities
(D) Adding quality scores

average pileup

chr	position	A	C	G	T	score	qual
chr1	15002478	8	1	4	**84**	24	–
chr2	14851248	3	0	**95**	2	35	–
chr3	24587456	1	**87**	1	2	27	–
chr5	12547896	**88**	0	30	2	28	–
chr7	52223654	**92**	2	1	4	30	–
chrX	33895474	5	2	0	**90**	25	–

filtered pileup

chr	position	A	C	G	T	score	qual
chr1	15002478	8	1	4	**84**	24	–
chr2	14851248	3	0	**95**	2	35	–
chr3	24587456	1	**87**	1	2	27	–
~~chr5~~	~~12547896~~	~~88~~	~~0~~	~~30~~	~~2~~	~~28~~	~~–~~
chr7	52223654	**92**	2	1	4	30	–
chrX	33895474	5	2	0	**90**	25	–

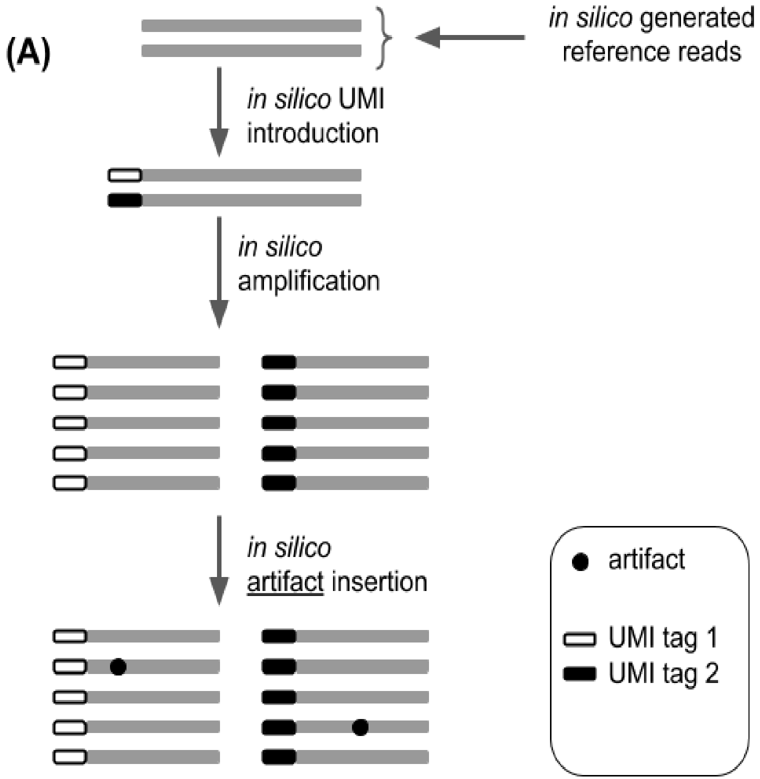
probability pileup

chr	position	A	C	G	T	score	qual
chr1	15002478	0.08	0.01	0.04	**0.87**	24	—
chr2	14851248	0.03	0	**0.95**	0.02	35	—
chr3	24587456	0.01	**0.96**	0.01	0.02	27	—
chr7	52223654	**0.93**	0.02	0.01	0.04	30	—
chrX	33895474	0.05	0.02	0	**0.93**	25	—

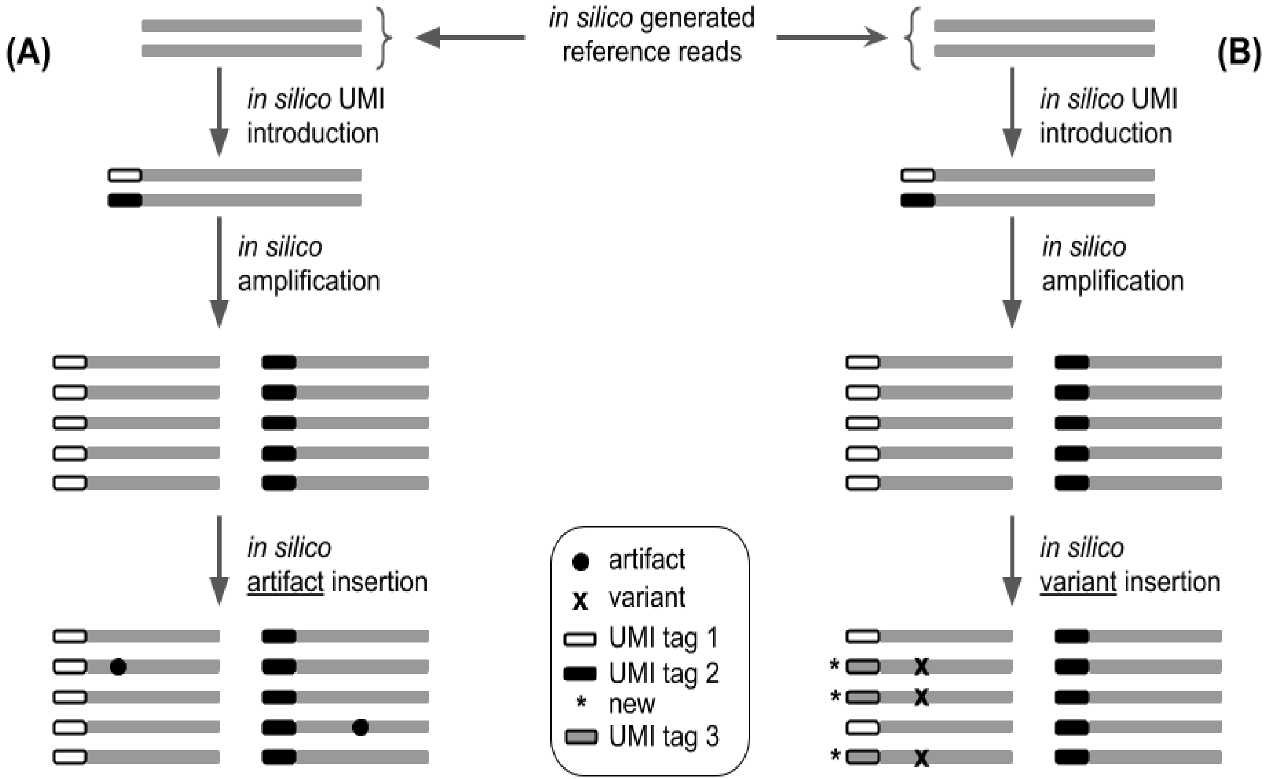
final pileup

chr	position	A	C	G	T	score	qual
chr1	15002478	0.08	0.01	0.04	**0.87**	24	9
chr2	14851248	0.03	0	**0.95**	0.02	35	D
chr3	24587456	0.01	**0.96**	0.01	0.02	27	<
chr7	52223654	**0.93**	0.02	0.01	0.04	30	?
chrX	33895474	0.05	0.02	0	**0.93**	25	:

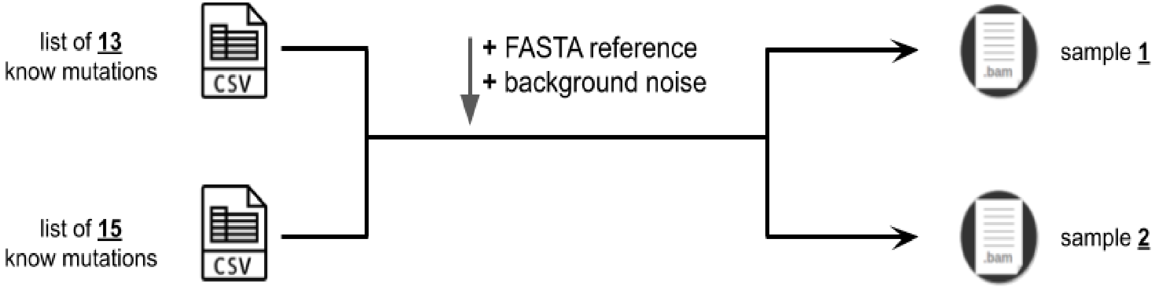
Adding the variants

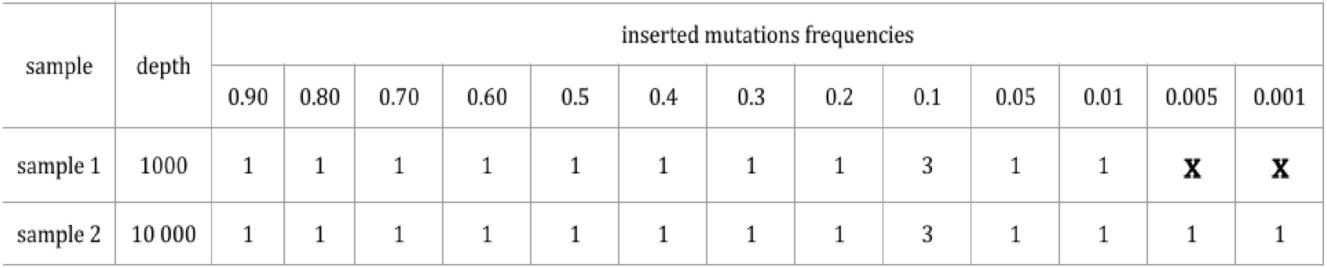

(A)
in silico generated reference reads
in silico UMI introduction
in silico amplification
in silico artifact insertion
● artifact
UMI tag 1
UMI tag 2

Adding the variants


(A)
in silico generated reference reads
(B)
in silico UMI introduction
in silico UMI introduction
in silico amplification
in silico amplification
in silico artifact insertion
in silico variant insertion
● artifact
x variant
UMI tag 1
UMI tag 2
* new
UMI tag 3

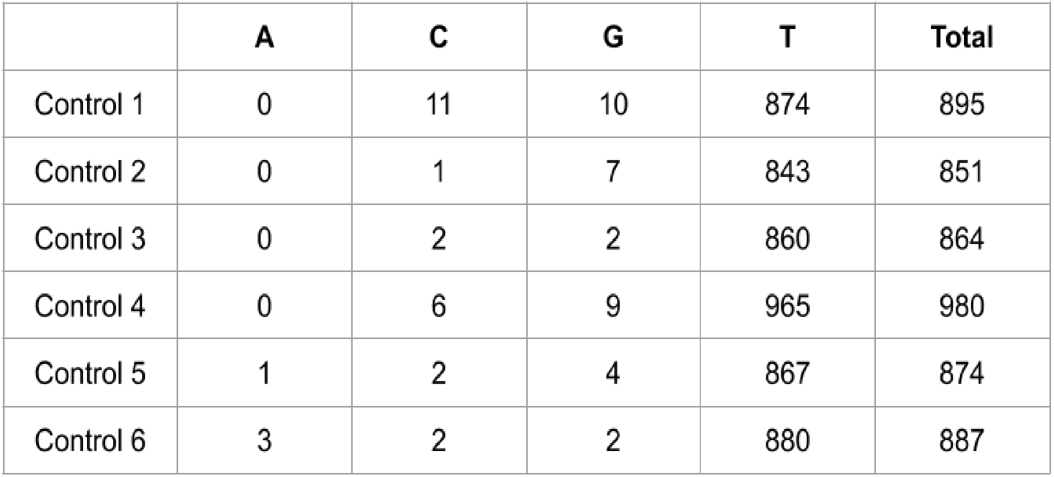
Sample Production





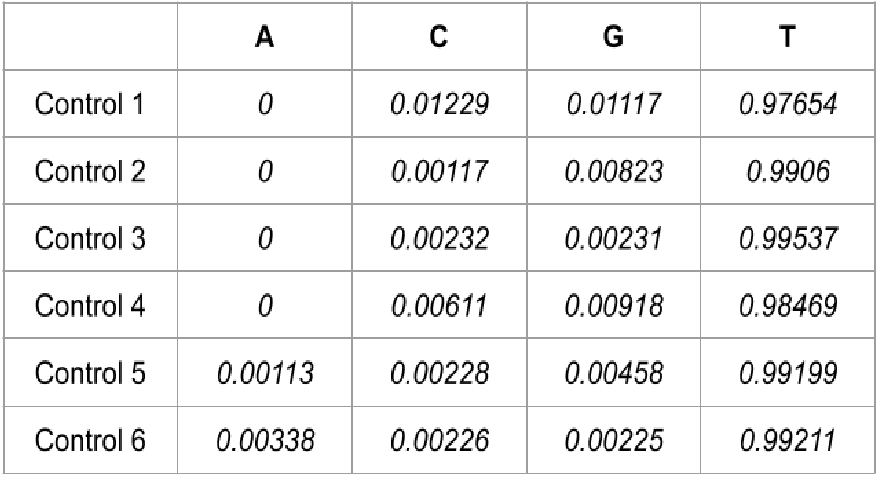
sample	depth	inserted mutations frequencies												
		0.90	0.80	0.70	0.60	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.005	0.001
sample 1	1000	1	1	1	1	1	1	1	1	3	1	1	**X**	**X**
sample 2	10 000	1	1	1	1	1	1	1	1	3	1	1	1	1

Noise reproduction from control samples



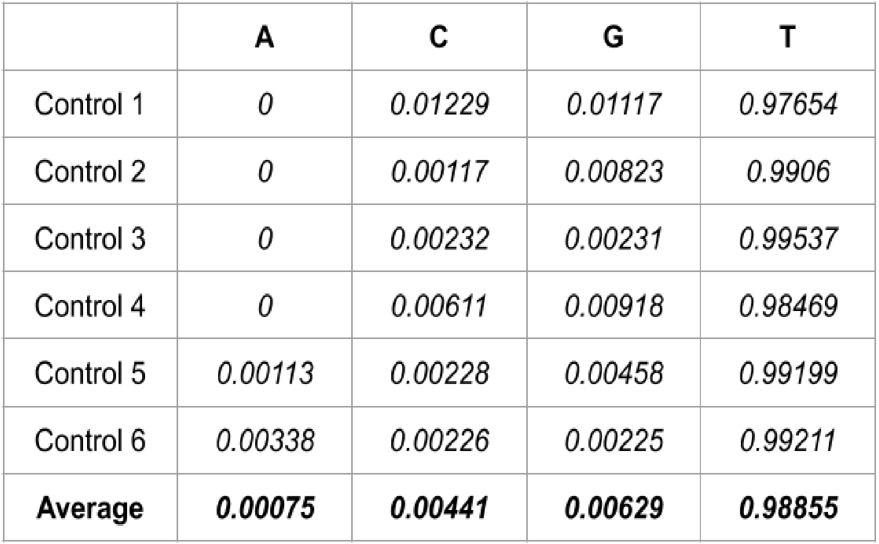
	A	C	G	T	Total
Control 1	0	11	10	874	895
Control 2	0	1	7	843	851
Control 3	0	2	2	860	864
Control 4	0	6	9	965	980
Control 5	1	2	4	867	874
Control 6	3	2	2	880	887

Noise reproduction from control samples



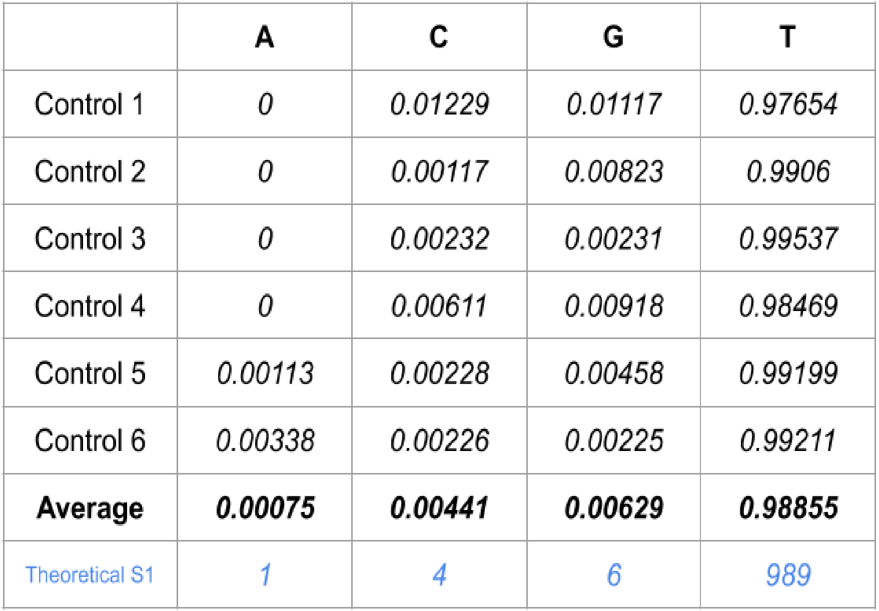
	A	C	G	T
Control 1	*0*	*0.01229*	*0.01117*	*0.97654*
Control 2	*0*	*0.00117*	*0.00823*	*0.9906*
Control 3	*0*	*0.00232*	*0.00231*	*0.99537*
Control 4	*0*	*0.00611*	*0.00918*	*0.98469*
Control 5	*0.00113*	*0.00228*	*0.00458*	*0.99199*
Control 6	*0.00338*	*0.00226*	*0.00225*	*0.99211*

Noise reproduction from control samples



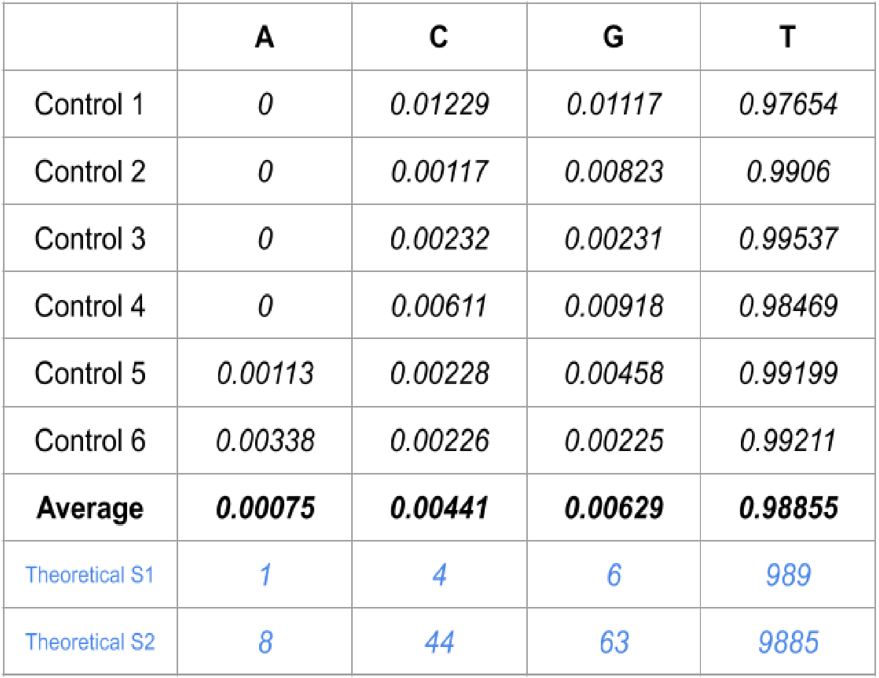
	A	C	G	T
Control 1	*0*	*0.01229*	*0.01117*	*0.97654*
Control 2	*0*	*0.00117*	*0.00823*	*0.9906*
Control 3	*0*	*0.00232*	*0.00231*	*0.99537*
Control 4	*0*	*0.00611*	*0.00918*	*0.98469*
Control 5	*0.00113*	*0.00228*	*0.00458*	*0.99199*
Control 6	*0.00338*	*0.00226*	*0.00225*	*0.99211*
Average	***0.00075***	***0.00441***	***0.00629***	***0.98855***

Noise reproduction from control samples



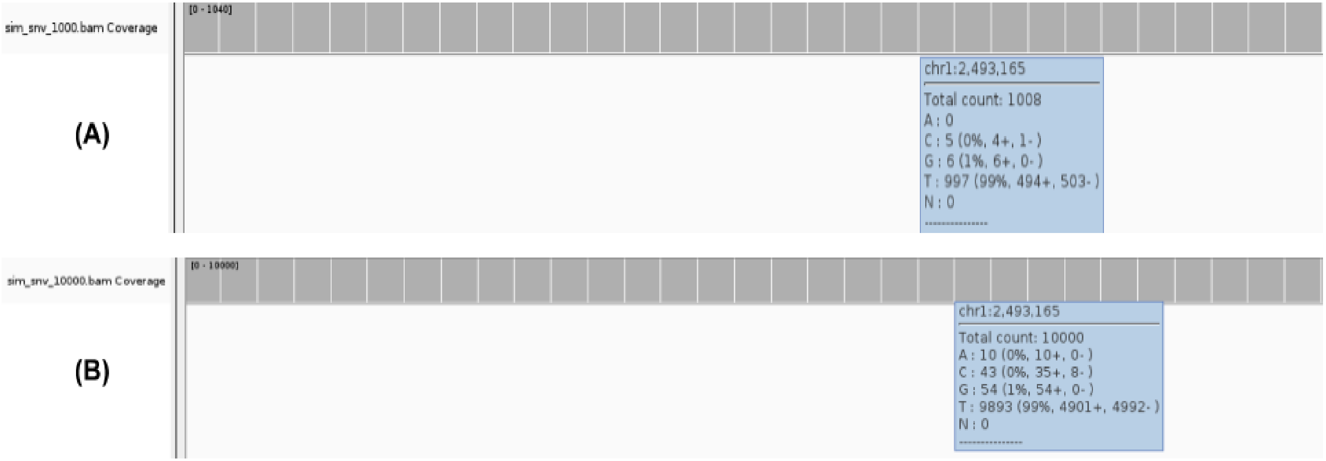
	A	C	G	T
Control 1	*0*	*0.01229*	*0.01117*	*0.97654*
Control 2	*0*	*0.00117*	*0.00823*	*0.9906*
Control 3	*0*	*0.00232*	*0.00231*	*0.99537*
Control 4	*0*	*0.00611*	*0.00918*	*0.98469*
Control 5	*0.00113*	*0.00228*	*0.00458*	*0.99199*
Control 6	*0.00338*	*0.00226*	*0.00225*	*0.99211*
Average	***0.00075***	***0.00441***	***0.00629***	***0.98855***
Theoretical S1	*1*	*4*	*6*	*989*

Noise reproduction from control samples



	A	C	G	T
Control 1	*0*	*0.01229*	*0.01117*	*0.97654*
Control 2	*0*	*0.00117*	*0.00823*	*0.9906*
Control 3	*0*	*0.00232*	*0.00231*	*0.99537*
Control 4	*0*	*0.00611*	*0.00918*	*0.98469*
Control 5	*0.00113*	*0.00228*	*0.00458*	*0.99199*
Control 6	*0.00338*	*0.00226*	*0.00225*	*0.99211*
Average	***0.00075***	***0.00441***	***0.00629***	***0.98855***
Theoretical S1	*1*	*4*	*6*	*989*
Theoretical S2	*8*	*44*	*63*	*9885*

Noise in the produced samples



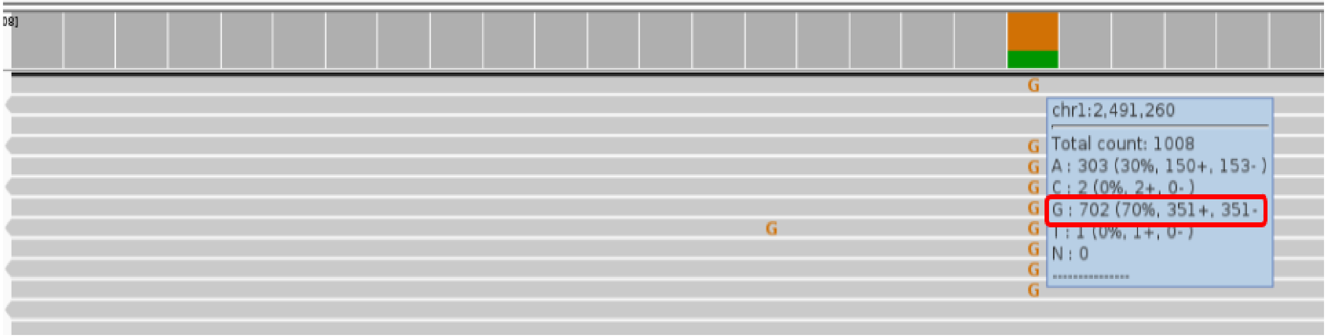
(A)

```
chr1:2,493,165
Total count: 1008
A : 0
C : 5 (0%, 4+, 1- )
G : 6 (1%, 6+, 0- )
T : 997 (99%, 494+, 503- )
N : 0
```

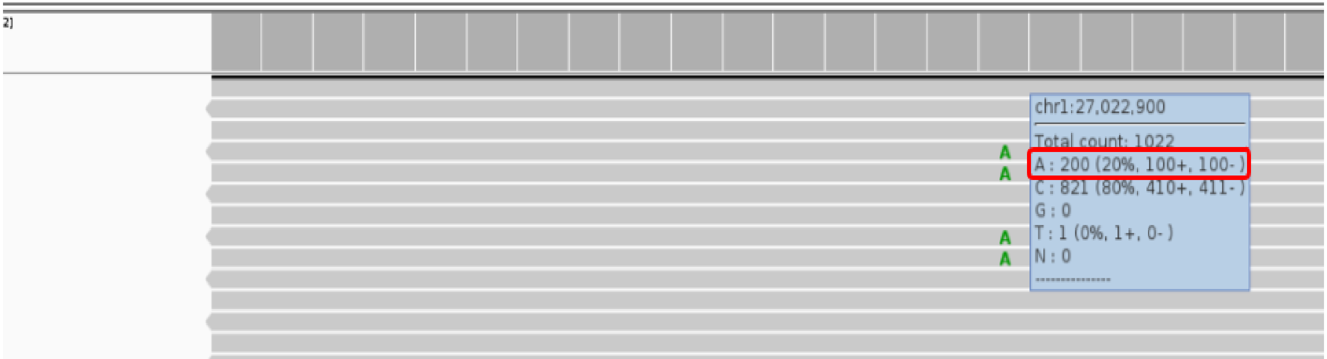
(B)

```
chr1:2,493,165
Total count: 10000
A : 10 (0%, 10+, 0- )
C : 43 (0%, 35+, 8- )
G : 54 (1%, 54+, 0- )
T : 9893 (99%, 4901+, 4992- )
N : 0
```

Real variants in the produced samples

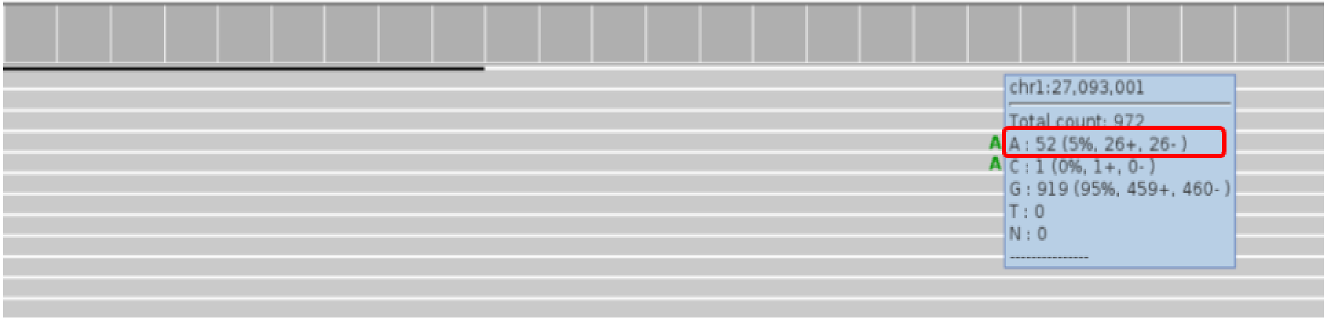


```
chr1:2,491,260
Total count: 1008
A : 303 (30%, 150+, 153- )
C : 2 (0%, 2+, 0- )
G : 702 (70%, 351+, 351-
T : 1 (0%, 1+, 0- )
N : 0
---------------
```

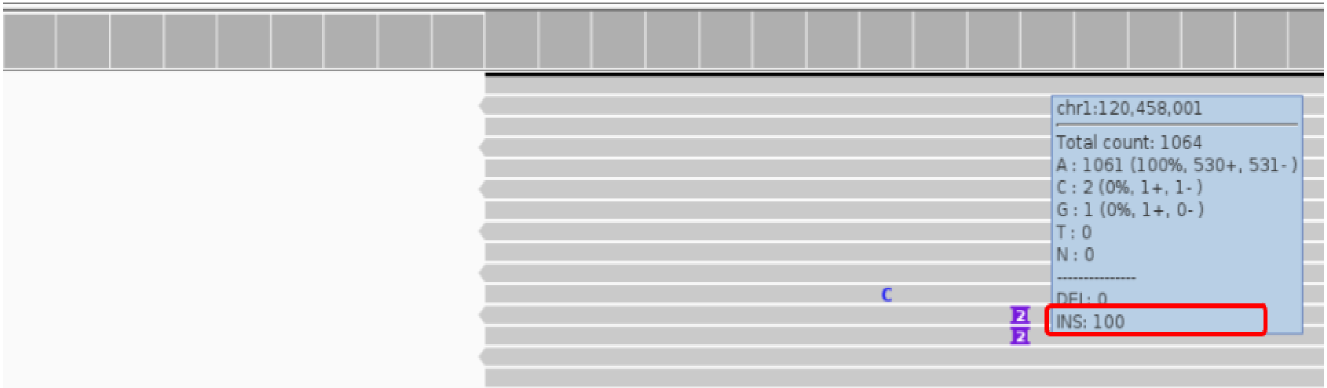


```
chr1:27,022,900
Total count: 1022
A : 200 (20%, 100+, 100- )
C : 821 (80%, 410+, 411- )
G : 0
T : 1 (0%, 1+, 0- )
N : 0
---------------
```

Real variants in the produced samples

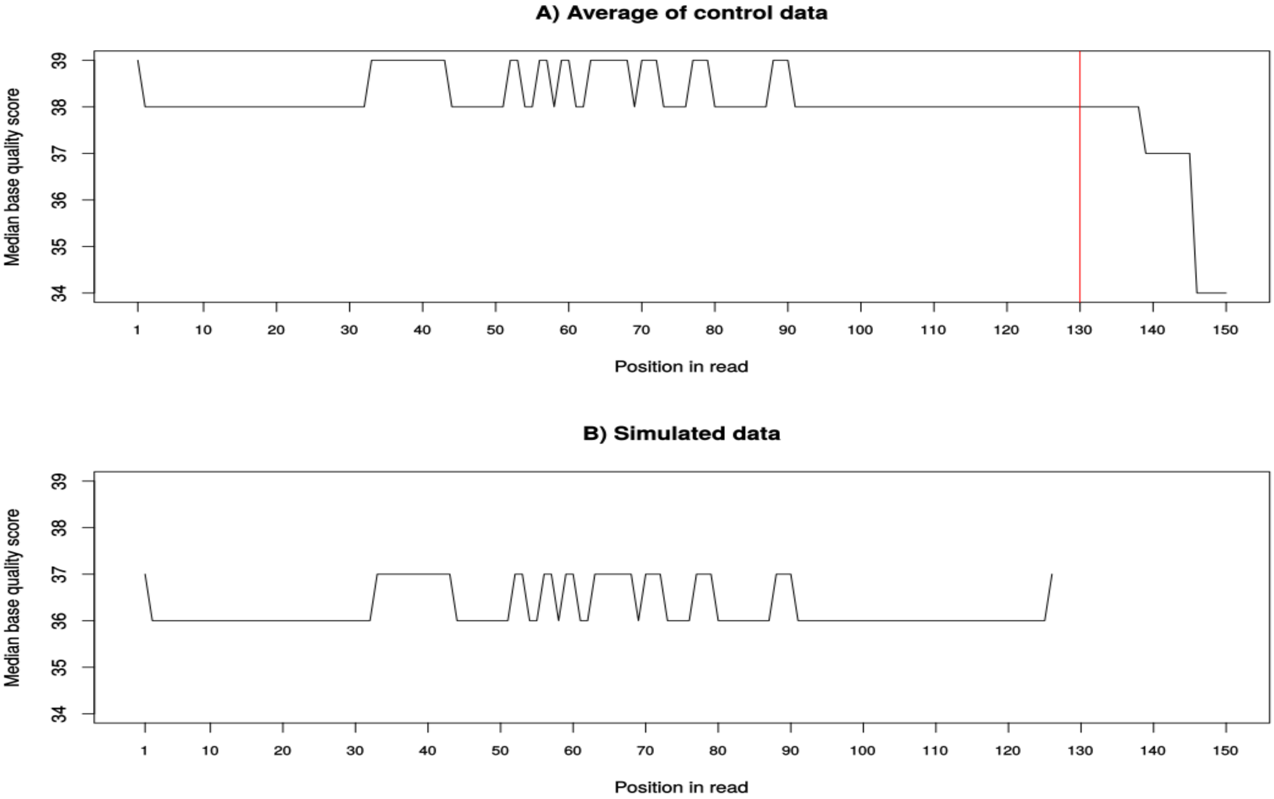


```
chr1:27,093,001
Total count: 972
A : 52 (5%, 26+, 26- )
C : 1 (0%, 1+, 0- )
G : 919 (95%, 459+, 460- )
T : 0
N : 0
---------------
```

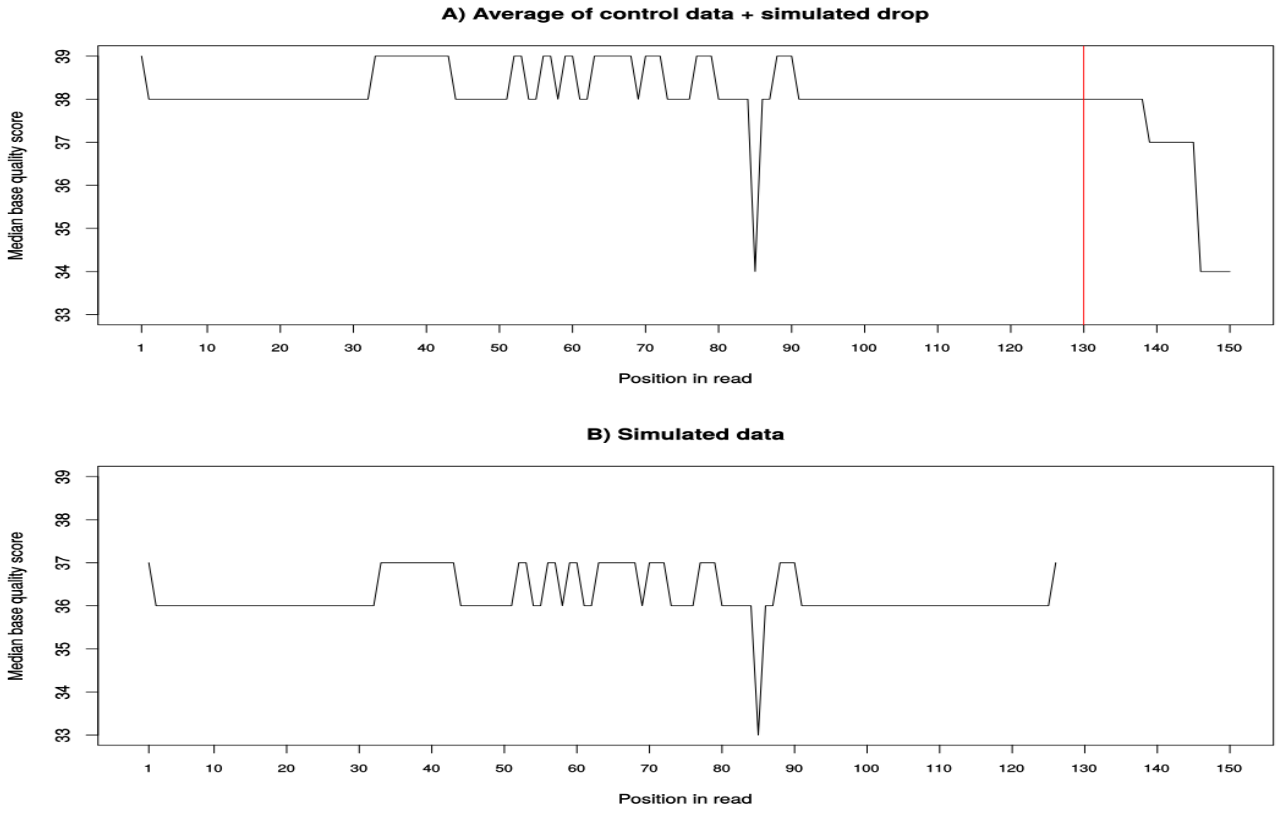


```
chr1:120,458,001
Total count: 1064
A : 1061 (100%, 530+, 531- )
C : 2 (0%, 1+, 1- )
G : 1 (0%, 1+, 0- )
T : 0
N : 0
---------------
DEL: 0
INS: 100
```

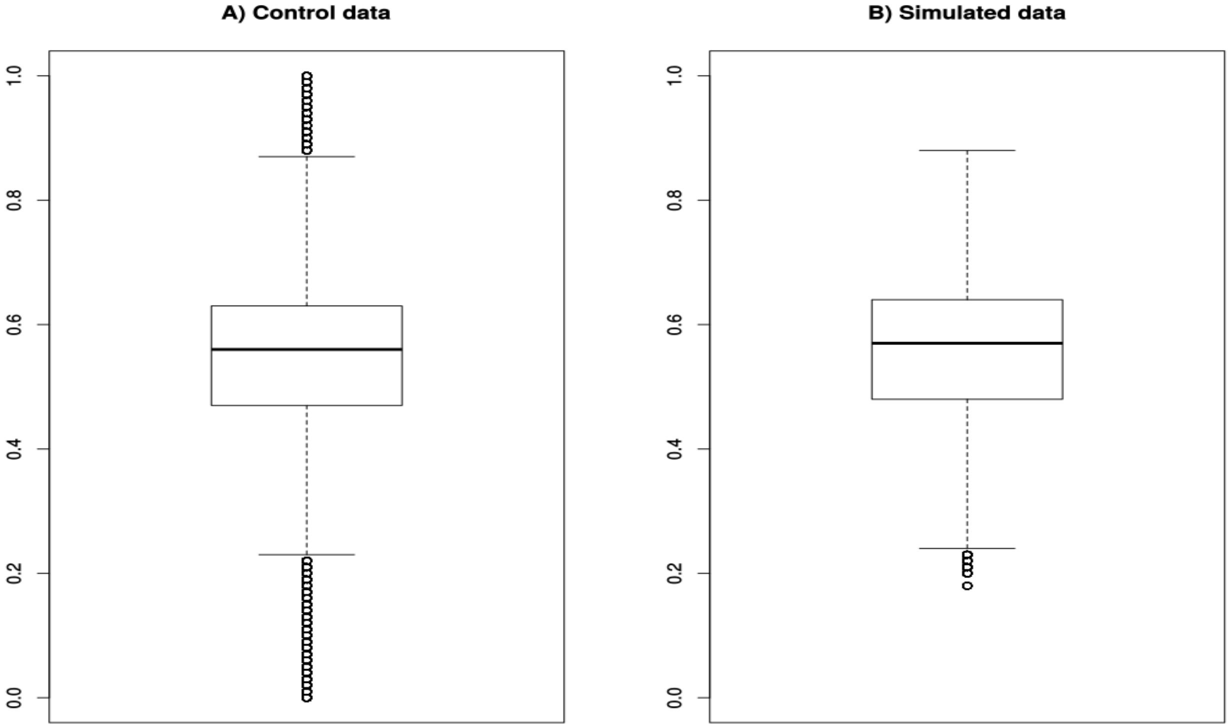
Quality score reproduction



Quality score reproduction



%GC comparison

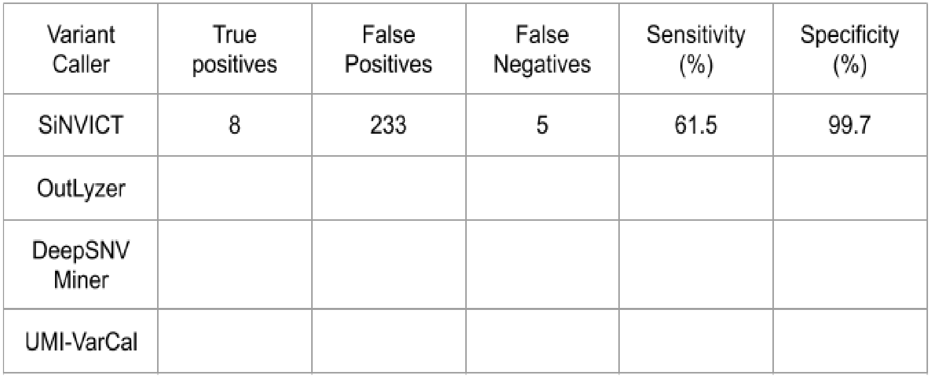


Evaluation on 4 variant callers

- 2 raw-read-based variant callers: SiNVICT & OutLyzer
- 2 UMI-based variant callers: UMI-VarCal & DeepSNVMiner
- Sample 1 has a depth of 1000x with 13 variants
- Sample 2 has a depth of 10,000x with 15 variants

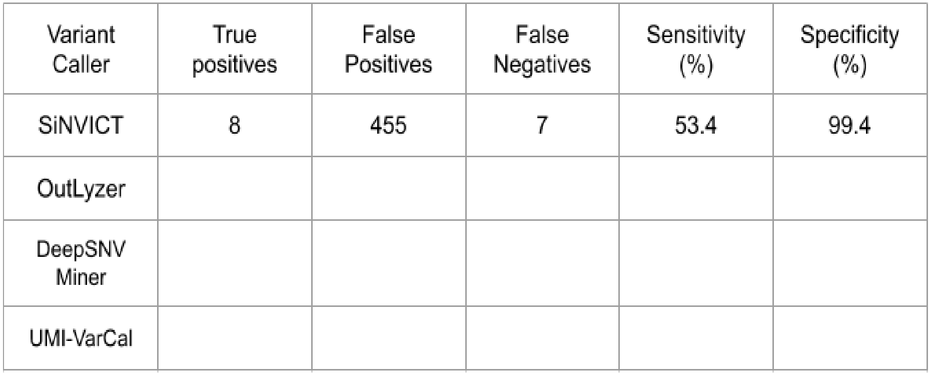
Evaluation on 4 variant callers

sample 1



Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	233	5	61.5	99.7
OutLyzer					
DeepSNV Miner					
UMI-VarCal					

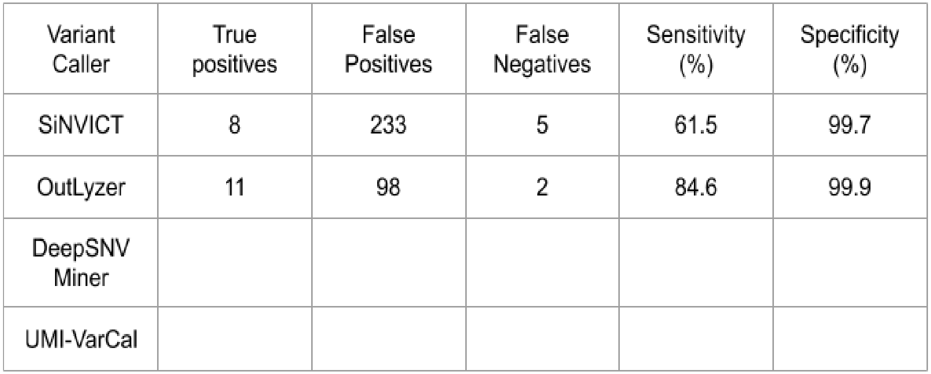
sample 2



Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	455	7	53.4	99.4
OutLyzer					
DeepSNV Miner					
UMI-VarCal					

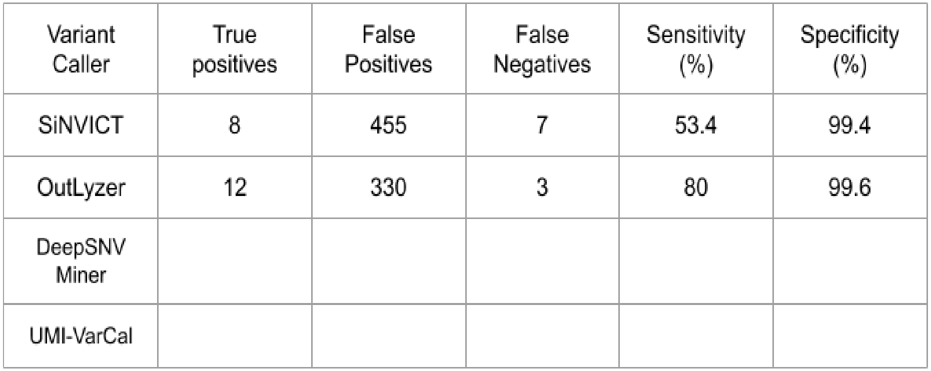
Evaluation on 4 variant callers

sample 1



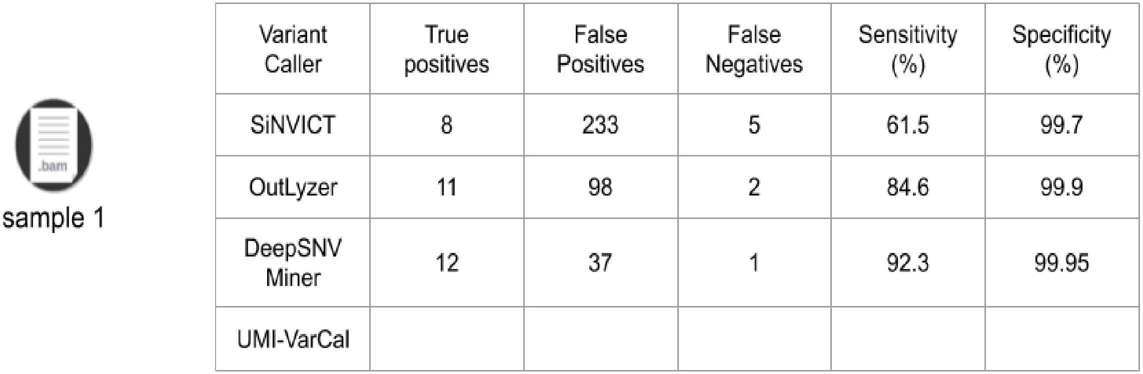
Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	233	5	61.5	99.7
OutLyzer	11	98	2	84.6	99.9
DeepSNV Miner					
UMI-VarCal					

sample 2



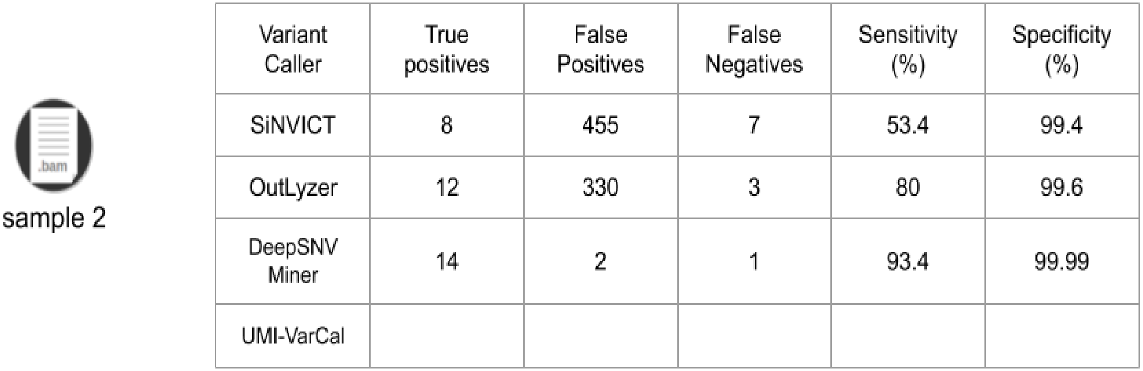
Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	455	7	53.4	99.4
OutLyzer	12	330	3	80	99.6
DeepSNV Miner					
UMI-VarCal					

Evaluation on 4 variant callers



sample 1

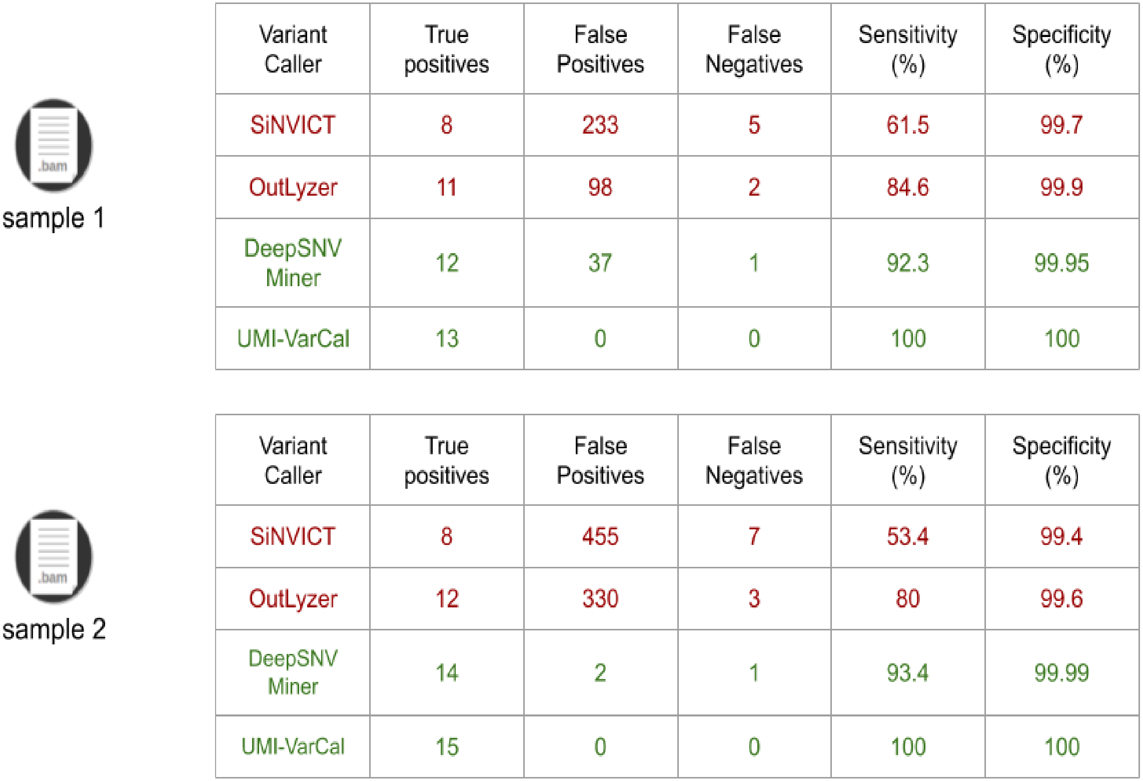
Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	233	5	61.5	99.7
OutLyzer	11	98	2	84.6	99.9
DeepSNV Miner	12	37	1	92.3	99.95
UMI-VarCal					



sample 2

Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	455	7	53.4	99.4
OutLyzer	12	330	3	80	99.6
DeepSNV Miner	14	2	1	93.4	99.99
UMI-VarCal					

Evaluation on 4 variant callers



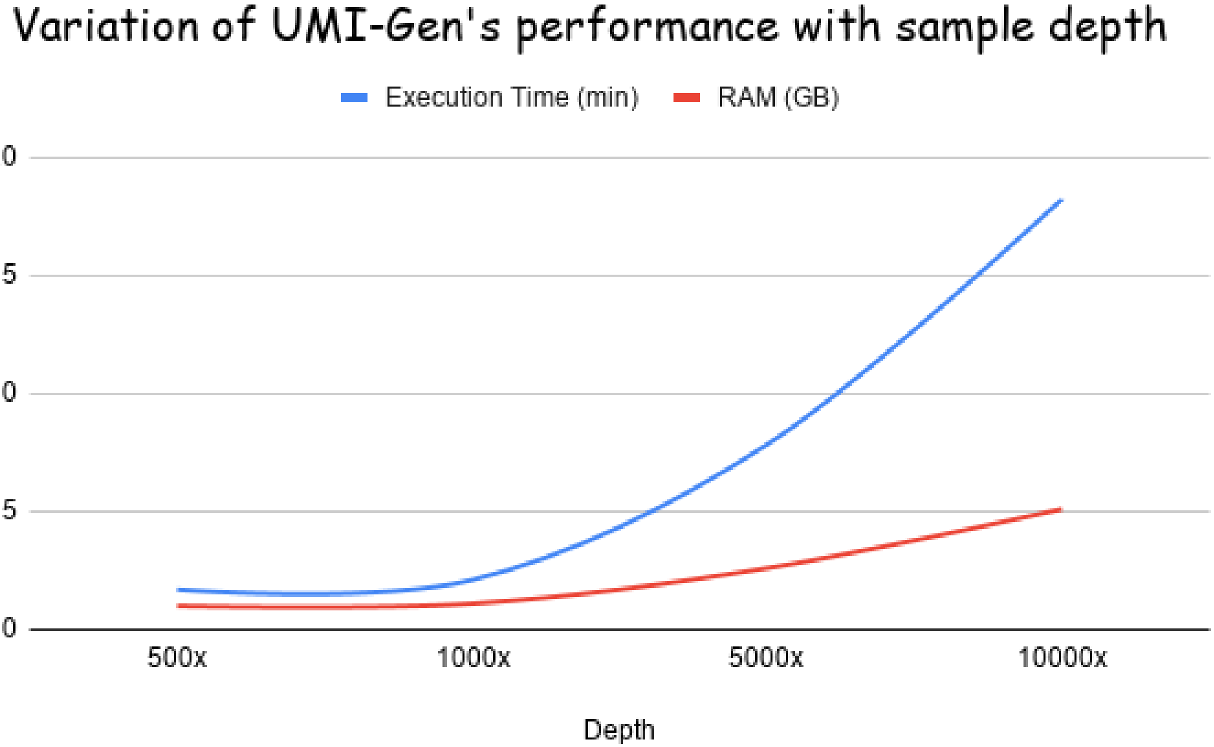
sample 1

Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	233	5	61.5	99.7
OutLyzer	11	98	2	84.6	99.9
DeepSNV Miner	12	37	1	92.3	99.95
UMI-VarCal	13	0	0	100	100

sample 2

Variant Caller	True positives	False Positives	False Negatives	Sensitivity (%)	Specificity (%)
SiNVICT	8	455	7	53.4	99.4
OutLyzer	12	330	3	80	99.6
DeepSNV Miner	14	2	1	93.4	99.99
UMI-VarCal	15	0	0	100	100

Performance



Conclusion

- UMI-based read simulator that is capable of producing artificial samples in which real variants are known
- Most of the parameters are customizable which allows total control over the simulation
- Generates FASTQ, BAM and SAM files
- Published in the Computational and Structural Biotecnology Journal and available for Python 3 on GitLab: https://gitlab.com/vincent-sater/umigen

Thank you for the attention