Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Life Sciences und
Facility Management

IAS Institut für
Angewandte Simulation

A study of dynamics of Indels using ProPIP, PRANK and MAFFT

Master's Thesis 2020 by Eldhose Poulose
Supervisors: Dr. Manuel Gil, Dr. Maria Anisimova, Dr. Massimo Maiolo

30.06.2020 09:00 CEST

Objective and Study Plan

- Can PIP adapt to long indels or not ?
- Comparative study using ProPIP, PRANK and MAFFT
- Dataset: Using long indel (INDELible, Real data) and single indel data (PIP)
- Study using supplementary features of ProPIP.

ProPIP

- Built under an explicit indel process, PIP (Poisson Indel Process)

  Basic PIP equations:

  Asymptotic expected sequence length, $E = \lambda/\mu$

  Indel Intensity, $I = \lambda \cdot \mu$

- Progressive dynamic programming approach.
- Marginal likelihood calculation at each internal node of the guide tree
- Alignment selection using Maximum Likelihood
- ProPIP requirements: input sequences, trees, rate matrix Q, λ and μ

ProPIP

1. Discrete Gamma distribution
    1. To tune the variation of indel rates across the sites
    2. Parameters: Number of classes (n) and shape parameter (α)

2. k-Factor
    - To tune the parameters λ (insertion rate) and μ (deletion rate)
    - Parameter: k

$$\lambda' = k.\lambda \text{ and } \mu' = k.\mu$$

$$E = \lambda'/\mu' = \lambda/\mu$$

$$I = \lambda' . \mu' = k^2 . \lambda . \mu$$

3. Discrete Gamma distribution and k-Factor

ProPIP

4. Stochastic Backtracking DP algorithm (SBDP)
    i. Ensemble of sub-optimal solutions at each internal node
    ii. Randomise the sub-optimal MSAs selected using SB algorithm
    iii. The distortion parameter, T

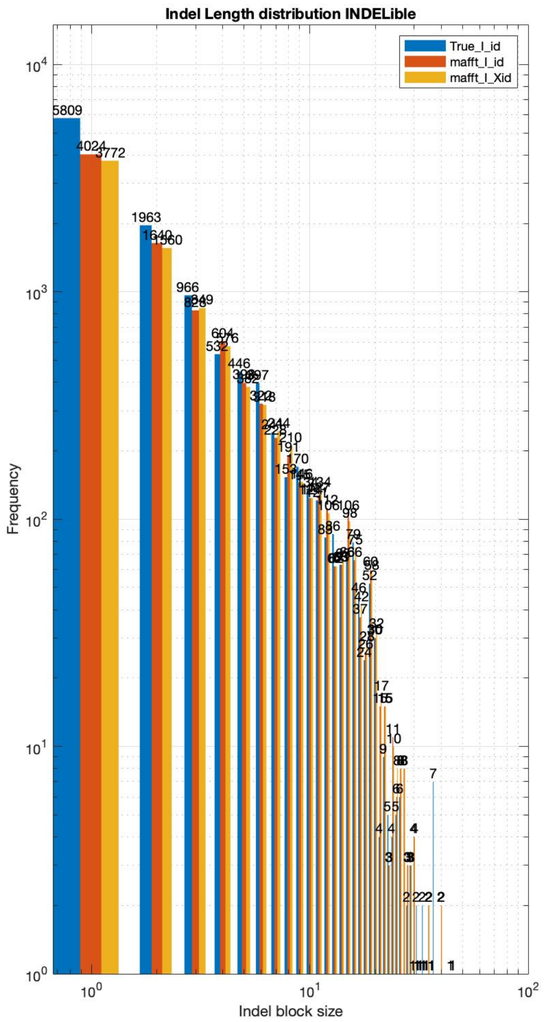
5. Short-Time Fourier Transform (STFT)
    i. Detection of homologous regions using STFT
    ii. For reducing computational complexity
    iii. Window functions and its size

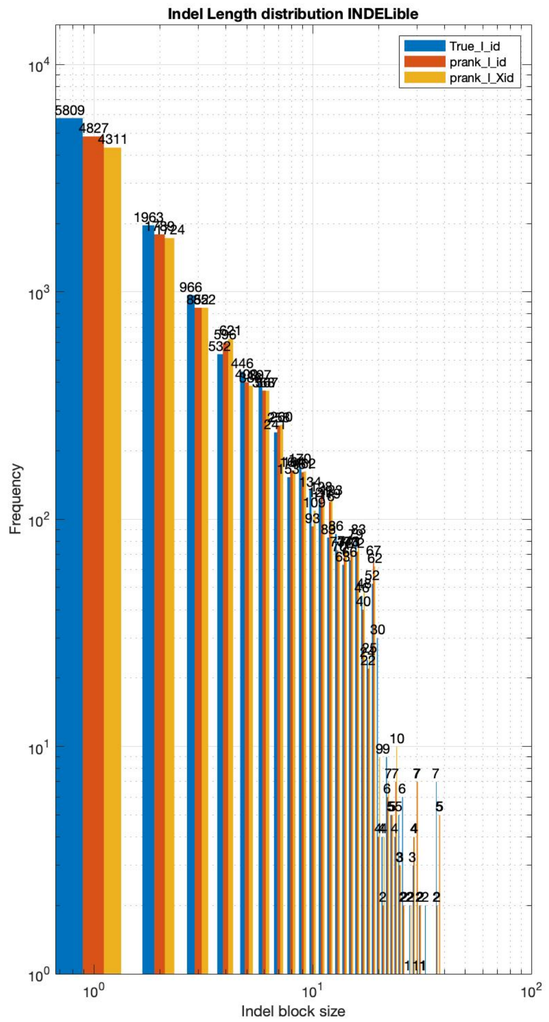
MSA Evaluation methods

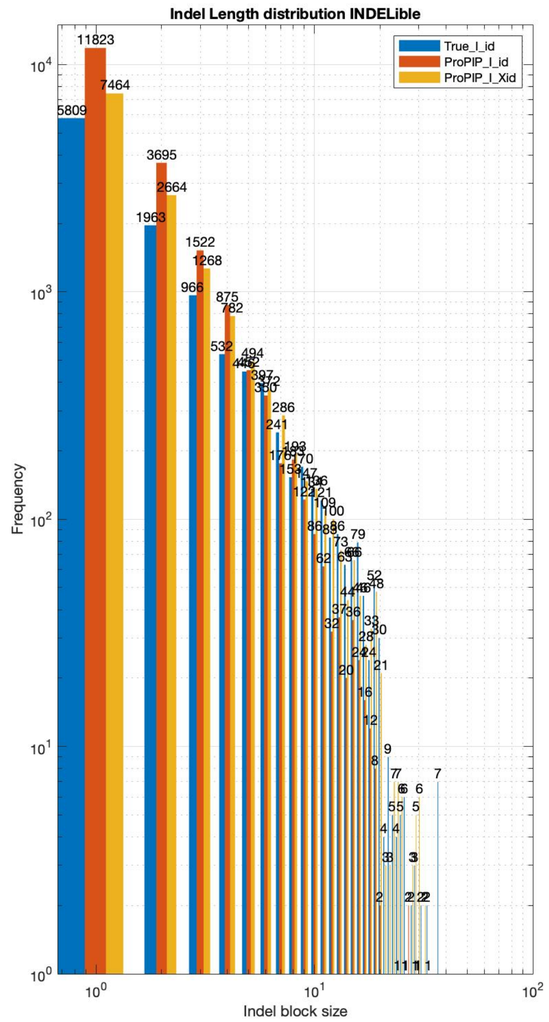
- Comparative study using two state of art aligners MAFFT and PRANK.
- Indel length and Indel block method
- Sequence −−AC−−T−AAG−−− the length of gaps [2,2,1,3] in Indel length method
- Same sequence under Indel block method [2,3,3]
- Test the additional features in our aligner to understand the indel placements in the Inferred MSAs.

- Log-Log plot and Summary statistics Table.
- Visualisation using the Pixel-plot.
- Quality analysis using qscore.

INDELible data analysis results



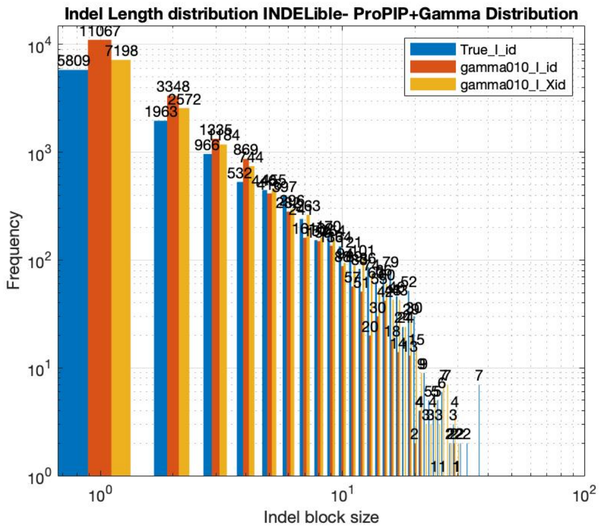


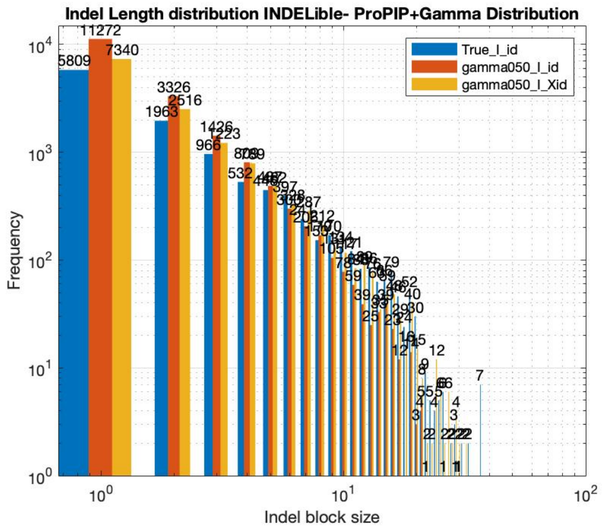


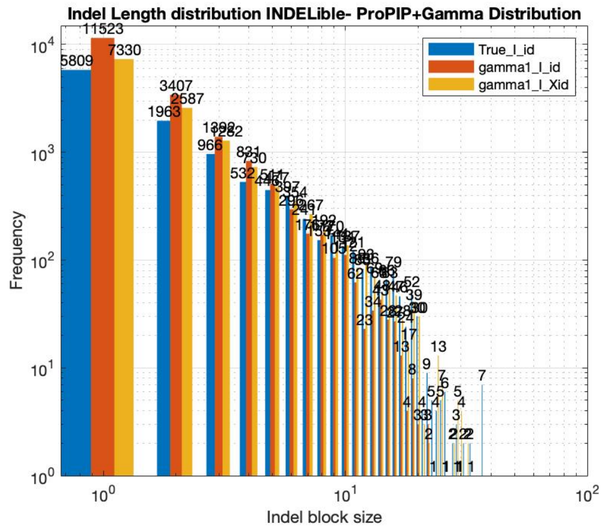
(100,8)	True(id)	MAFFT v7.453		PRANK v.170427		ProPIP	
		id	Xid	id	Xid	id	Xid
nIndels	11511	9279	8971	10226	9692	19539	14418
Max-IL	37	45	45	38	38	30	33
Mean	3.116	3.654	3.780	3.306	3.488	2.041	2.767
Median	1	2	2	2	2	1	1
SD	3.850	4.394	4.506	3.307	3.489	2.132	3.321

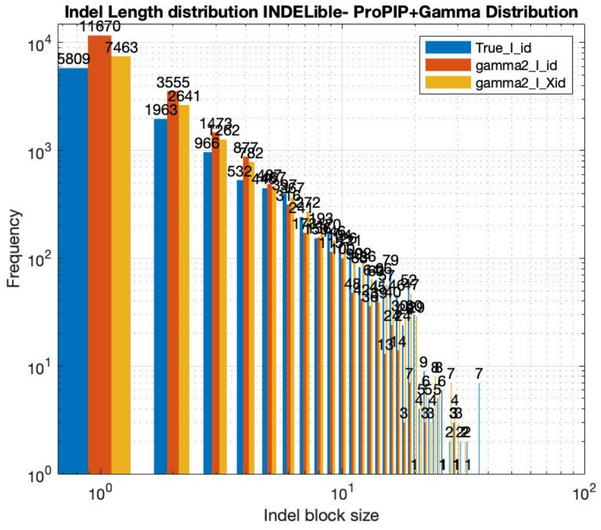
Table 6.1: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, and ProPIP. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution

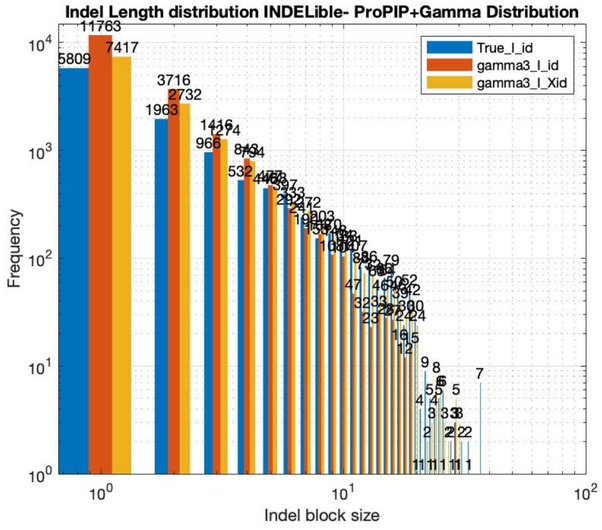
Figure 6.1: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.17042 (MSA 3), and ProPIP (MSA 4). Note: black pixel represents Characters and white pixel represents Indels.


Indel Length distribution INDELible- ProPIP+Gamma Distribution
True_I_id
gamma010_I_id
gamma010_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+Gamma Distribution
True_I_id
gamma050_I_id
gamma050_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+Gamma Distribution
True_I_id
gamma1_I_id
gamma1_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+Gamma Distribution
True_I_id
gamma2_I_id
gamma2_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+Gamma Distribution
True_I_id
gamma3_I_id
gamma3_I_Xid
Frequency
Indel block size

(100,8)	True (id)	G0.10		G0.50		G1		G2		G3	
		id	Xid	id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11511	18108	13700	18435	13964	18776	14060	19163	14283	19318	14346
Max-IL	37	30	30	30	32	30	33	30	33	30	33
Mean	3.116	2.045	2.703	2.050	2.707	2.043	2.728	2.023	2.723	2.026	2.728
Median	1	1	1	1	1	1	1	1	1	1	1
SD	3.850	2.1692	3.256	2.1604	3.228	2.148	3.269	2.105	3.290	2.142	3.255

Table 6.2: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under Gamma n=4, α=0.10, ProPIP with k=1 under Gamma n=4, α=0.50 , ProPIP with k=1 under Gamma n=4, α=1, ProPIP with k=1 under Gamma n=4, α=2, ProPIP with k=1 under Gamma n=4, α=3. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

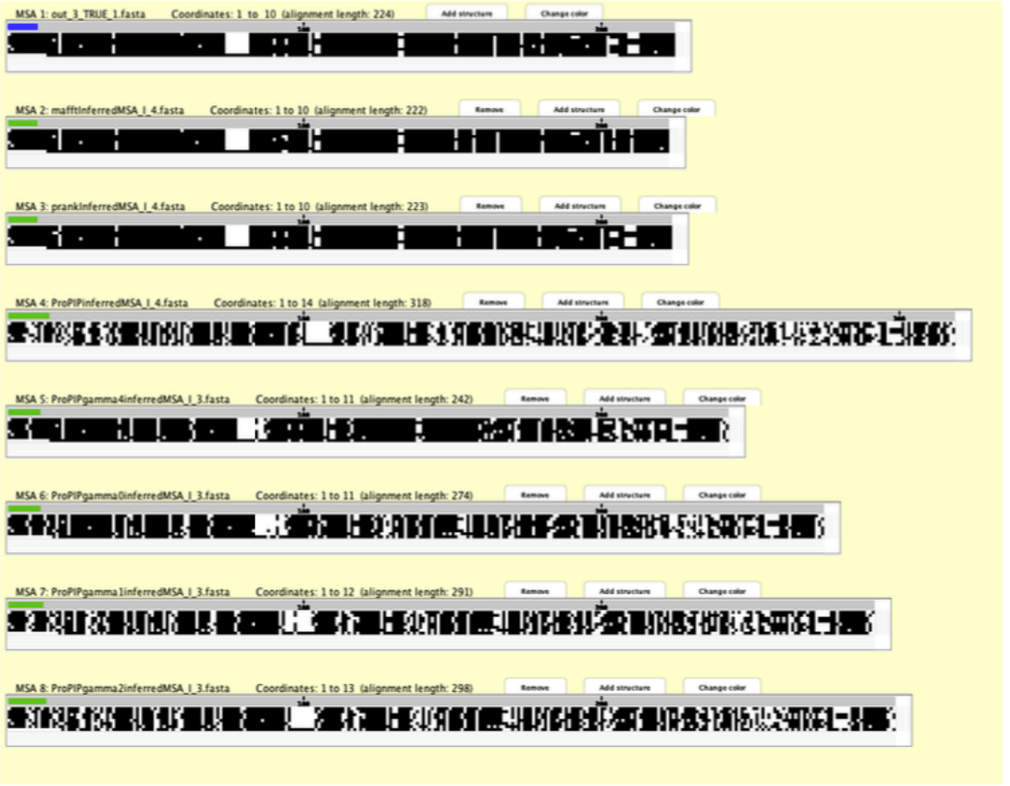
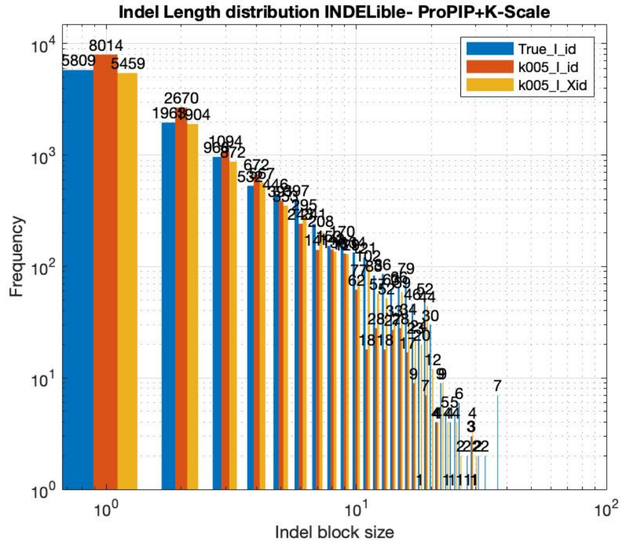
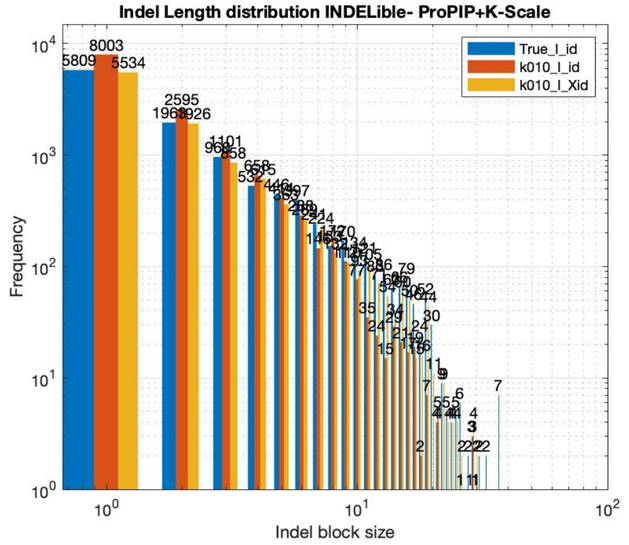
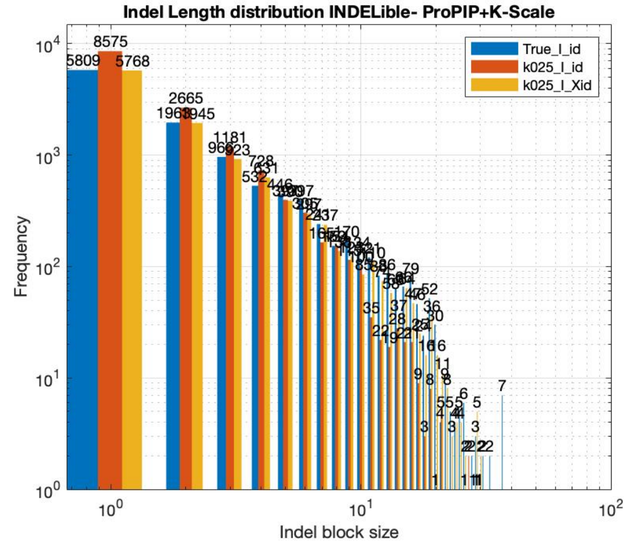
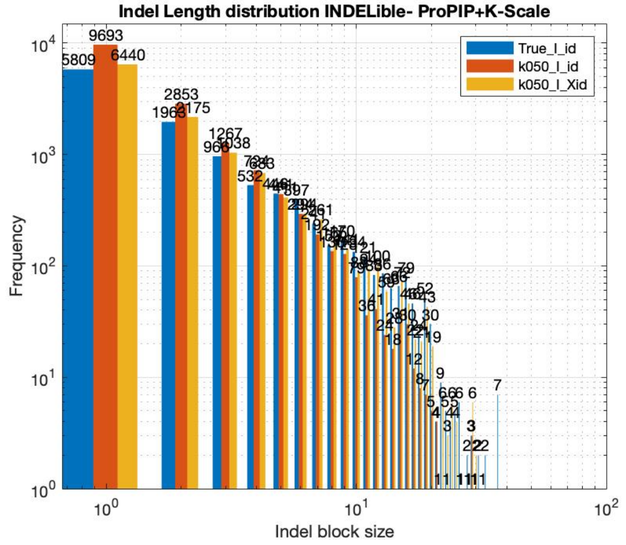


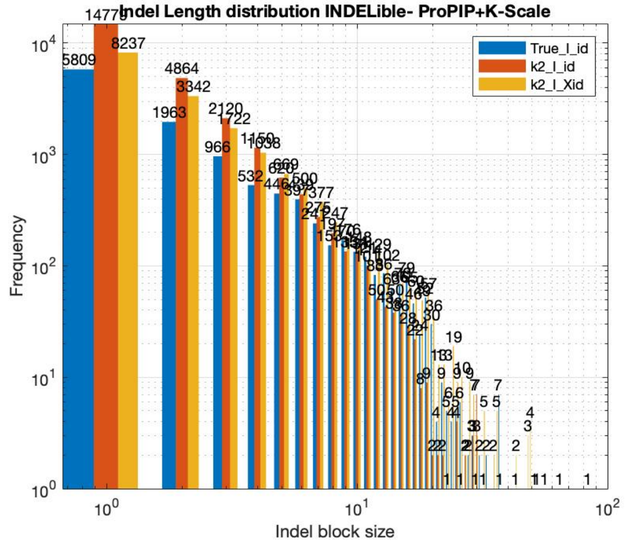
Figure 6.4: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.170427 (MSA 3), ProPIP with k=1 (MSA 4), ProPIP with k=1 under Gamma n=4, α=0.10 (MSA 5), ProPIP with k=1 under Gamma n=4, α=0.50 (MSA 6), ProPIP with k=1 under Gamma n=4, α=1 (MSA 7), and ProPIP with k=1 under Gamma n=4, α=2 (MSA 8). Note: black pixel represents Characters and white pixel represents Indels.

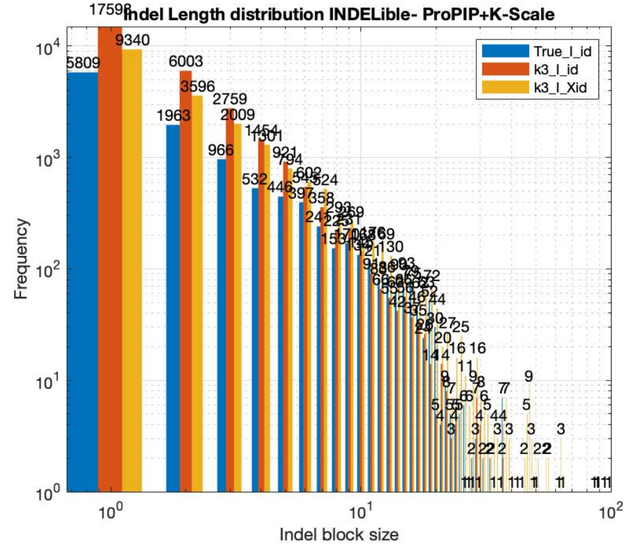

Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k005_I_id
k005_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k010_I_id
k010_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k025_I_id
k025_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k050_I_id
k050_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k2_I_id
k2_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K-Scale
True_I_id
k3_I_id
k3_I_Xid
Frequency
Indel block size

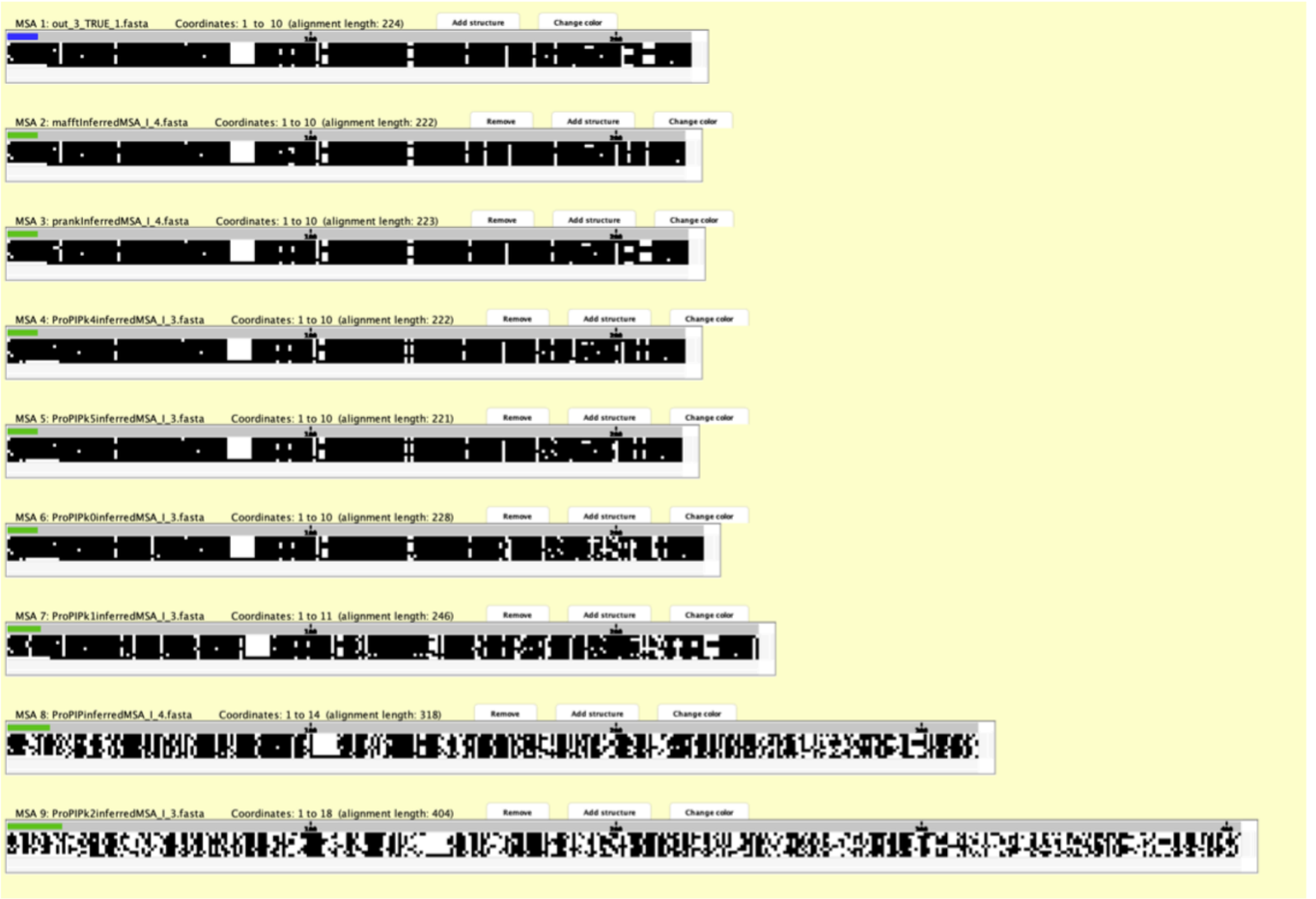
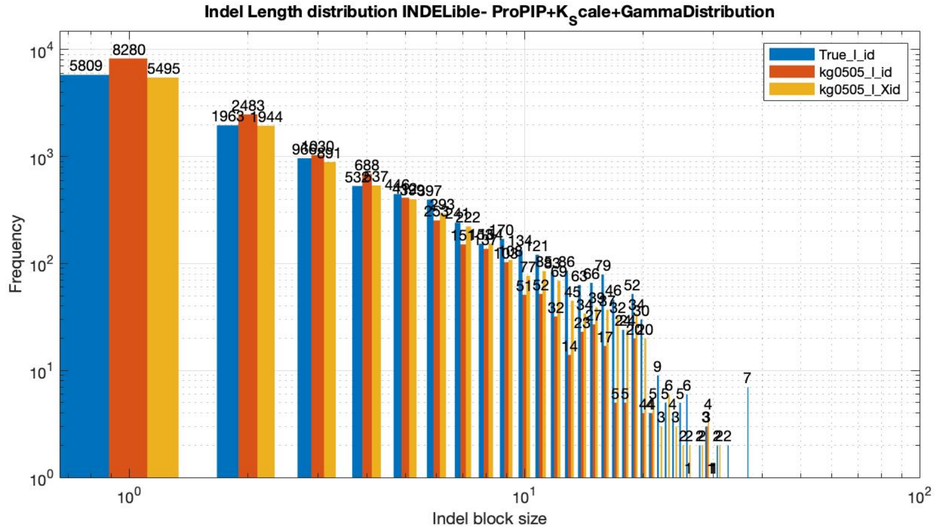
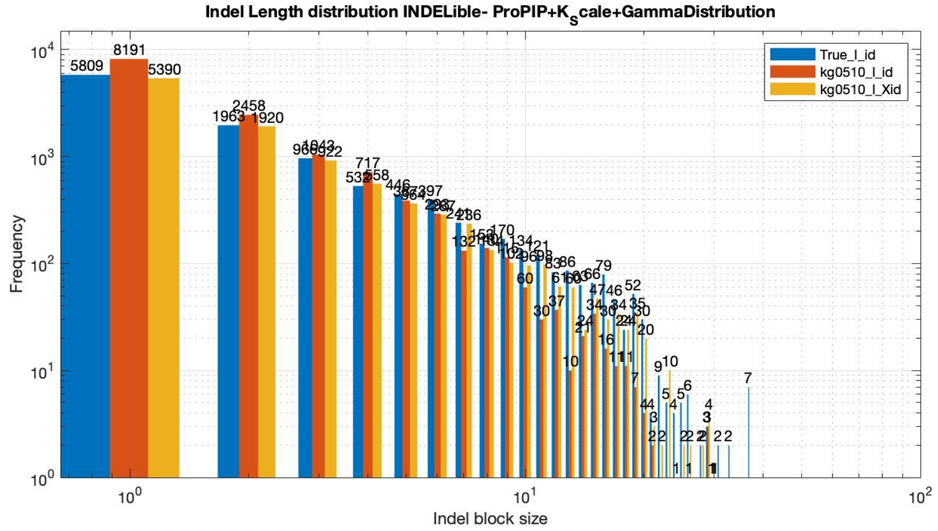


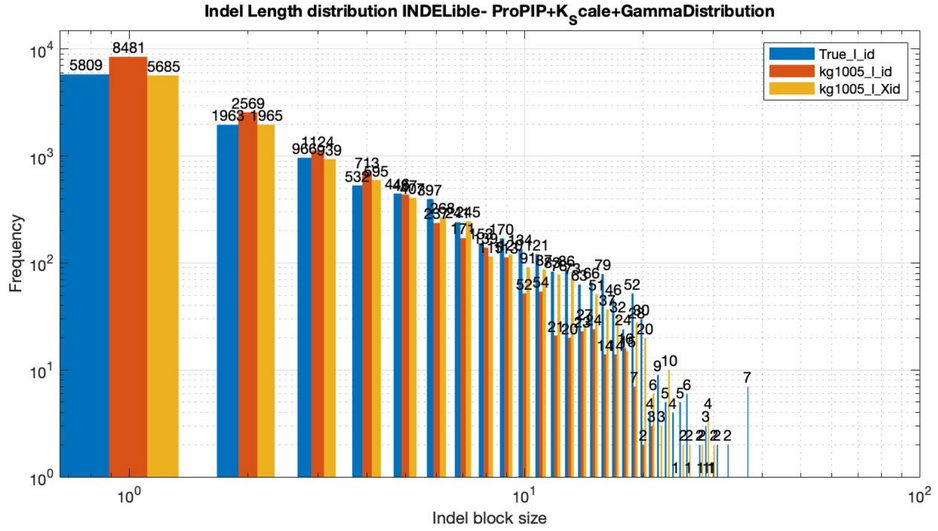
Figure 6.5: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.170427 (MSA 3), ProPIP with k=0.05 (MSA 4), ProPIP with k=0.10 (MSA 5), ProPIP with k=0.25 (MSA 6), ProPIP with k=0.50 (MSA 7), ProPIP with k=1 (MSA 8), and ProPIP with k=2 (MSA 9). Note: black pixel represents Characters and white pixel represents Indels.

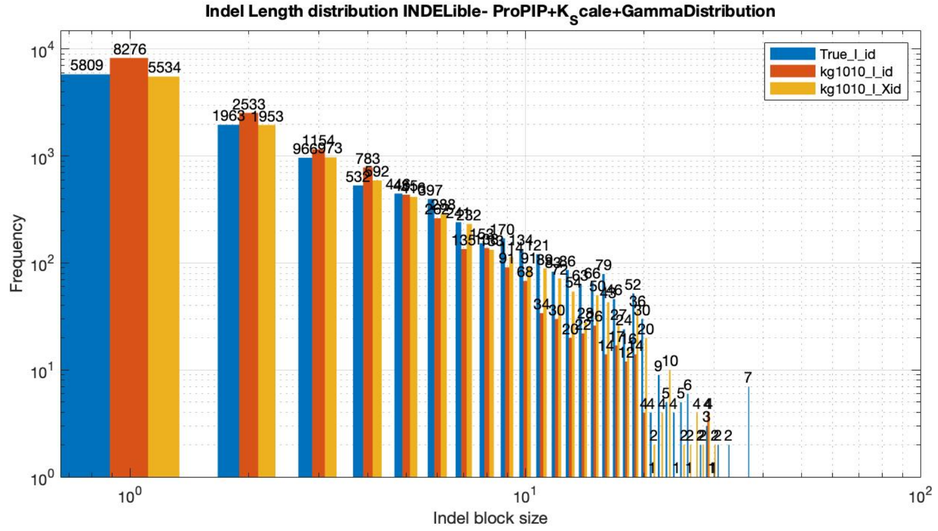
(100,8)	True (id)	k0.05		k0.10		k0.25		k0.50		k2		k3	
		id	Xid	id	Xid	id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11511	13725	10466	13745	10618	14562	11060	16030	12231	25071	17307	30713	20091
Max-IL	37	30	30	30	30	30	30	30	32	43	105	37	118
Mean	3.116	2.119	2.779	2.161	2.797	2.140	2.817	2.102	2.755	2.090	3.028	2.1663	3.312
Median	1	1	1	1	1	1	1	1	1	1	2	1	2
SD	3.850	2.207	3.316	2.261	3.352	2.226	3.376	2.255	3.294	2.230	3.985	2.347	4.740

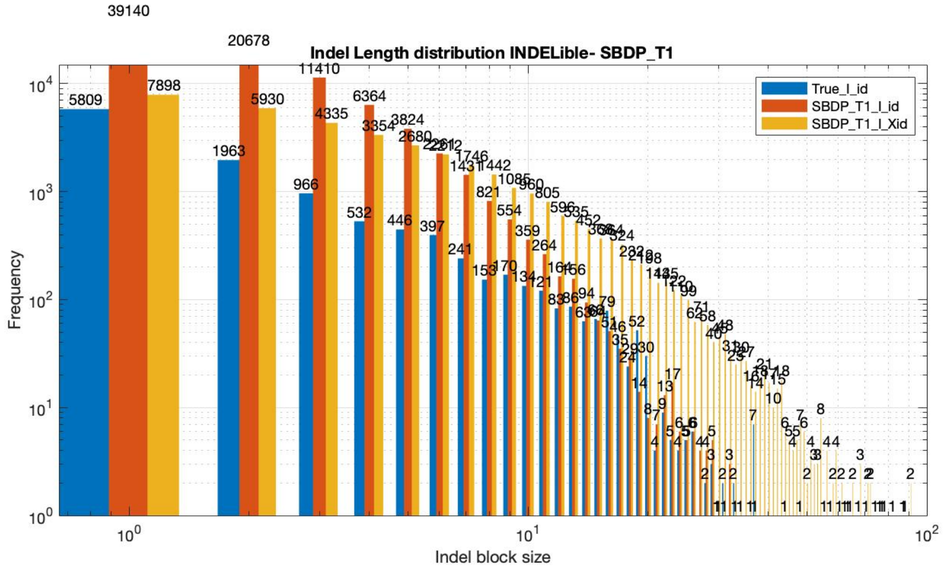
Table 6.3: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=0.05, ProPIP with k=0.10, ProPIP with k=0.25, ProPIP with k=0.50, ProPIP with k=2, ProPIP with k=3. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

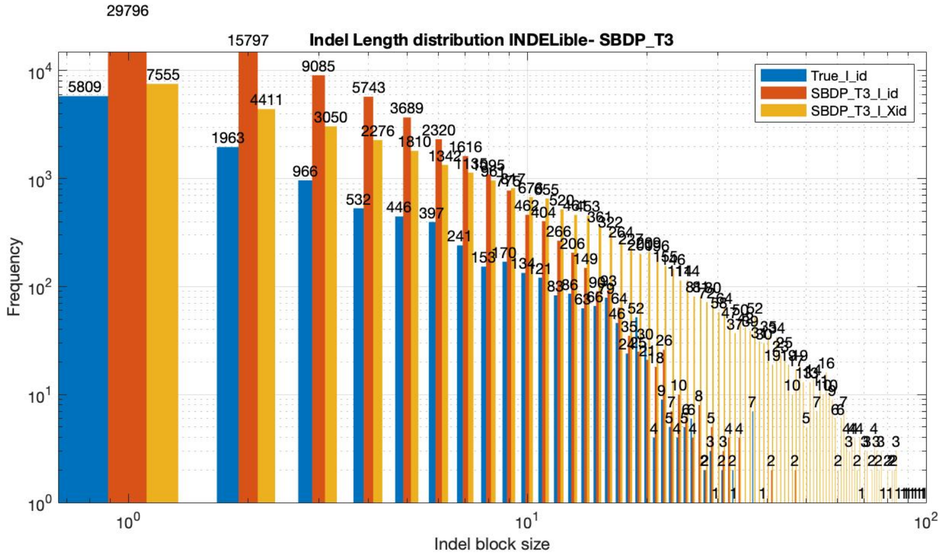

Indel Length distribution INDELible- ProPIP+K$_S$cale+GammaDistribution
True_I_id
kg0505_I_id
kg0505_I_Xid
Frequency
Indel block size

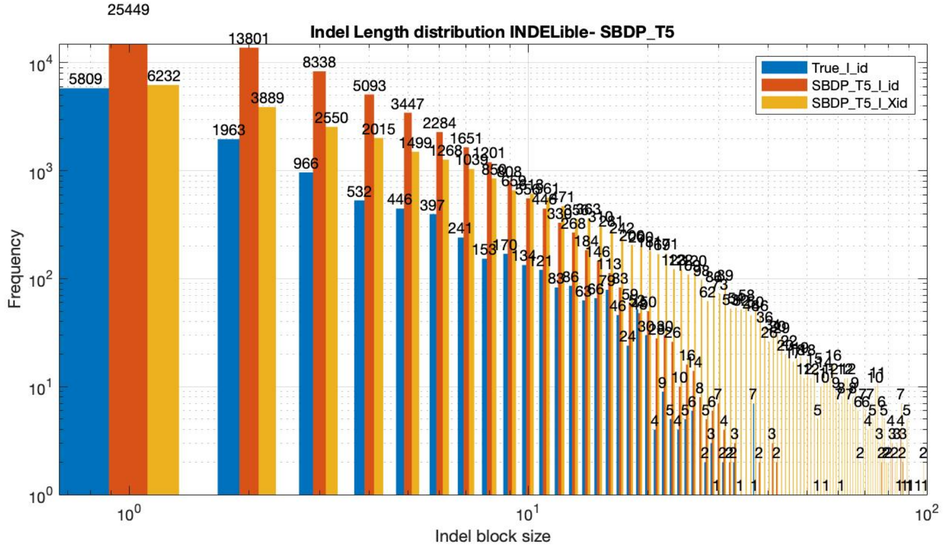

Indel Length distribution INDELible- ProPIP+K$_S$cale+GammaDistribution
True_I_id
kg0510_I_id
kg0510_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K$_S$cale+GammaDistribution
True_I_id
kg1005_I_id
kg1005_I_Xid
Frequency
Indel block size


Indel Length distribution INDELible- ProPIP+K$_S$cale+GammaDistribution
True_I_id
kg1010_I_id
kg1010_I_Xid
Frequency
Indel block size


39140
20678
Indel Length distribution INDELible- SBDP_T1
True_I_id
SBDP_T1_I_id
SBDP_T1_I_Xid
Frequency
Indel block size


29796
15797
Indel Length distribution INDELible- SBDP_T3
True_I_id
SBDP_T3_I_id
SBDP_T3_I_Xid
Frequency
Indel block size


25449
13801
Indel Length distribution INDELible- SBDP_T5
True_I_id
SBDP_T5_I_id
SBDP_T5_I_Xid
Frequency
Indel block size

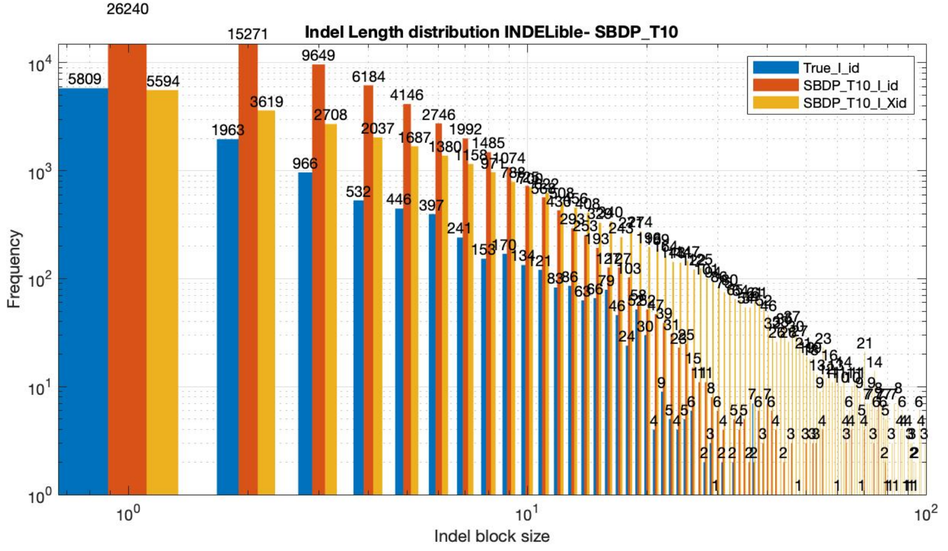

26240
15271
Indel Length distribution INDELible- SBDP_T10
True_I_id
SBDP_T10_I_id
SBDP_T10_I_Xid
Frequency
Indel block size

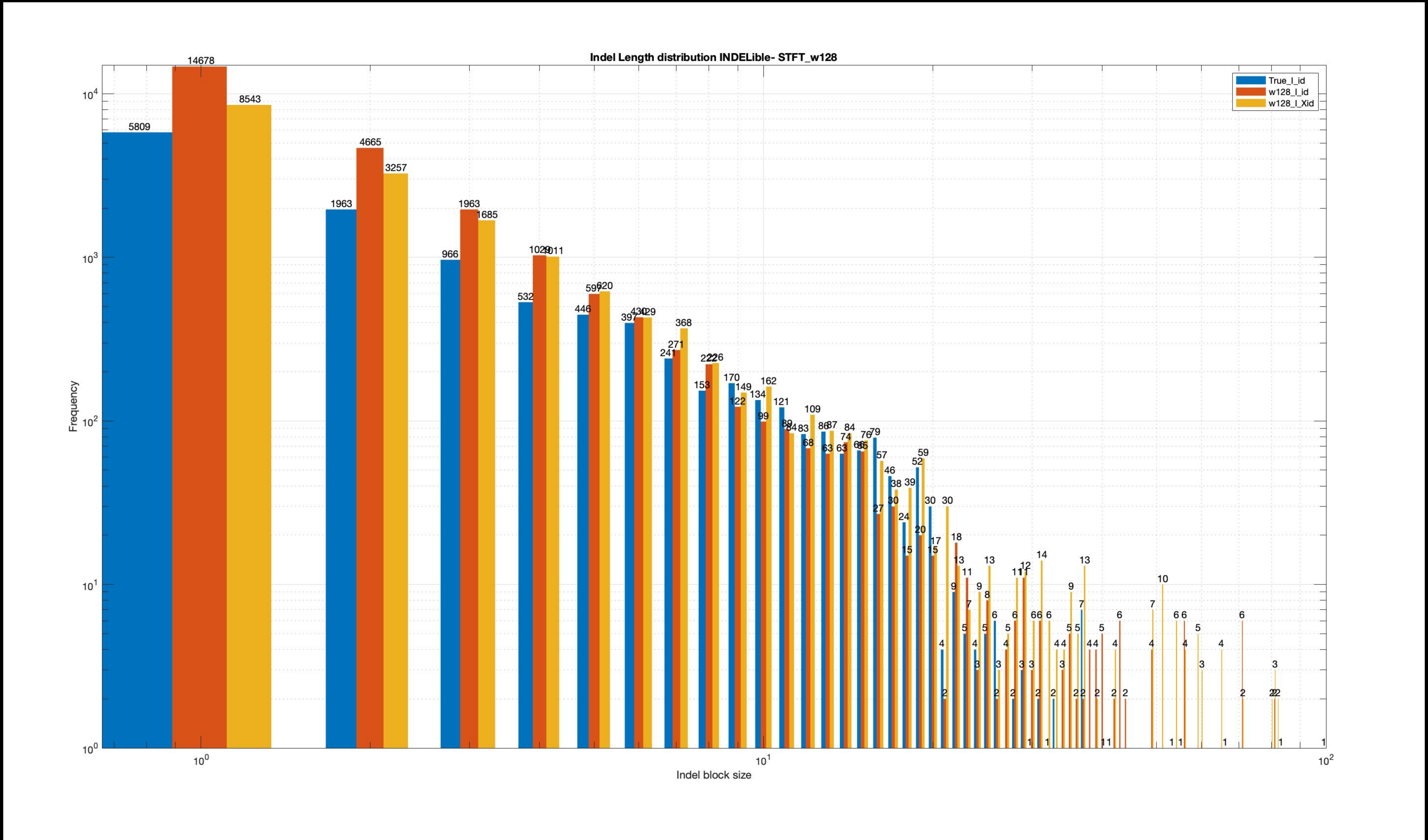
Figure 6.9: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.170427 (MSA 3), ProPIP with k=1 (MSA 4), ProPIP with k=1 under SBDP with T= 1 (MSA 5), ProPIP with k=1 under SBDP with T= 3 (MSA 6), ProPIP with k=1 under SBDP with T= 5 (MSA 7), and ProPIP with k=1 under SBDP with T= 10 (MSA 8). Note: black pixel represents Characters and white pixel represents Indels.

(100,8)	True (id)	T1		T3		T5		T10	
		id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11511	87799	37004	71834	29594	64539	25934	72090	26998
Max-IL	37	48	91	47	133	144	215	380	383
Mean	3.116	2.410	5.717	2.730	6.627	2.979	7.414	3.374	9.010
Median	1	2	4	2	3	2	4	2	4
SD	3.850	2.150	6.367	2.579	8.797	3.226	10.567	7.343	16.754

Table 6.6: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under SBDP with T= 1, ProPIP with k=1 under SBDP with T= 3, ProPIP with k=1 under SBDP with T= 5, ProPIP with k=1 under SBDP with T= 10. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

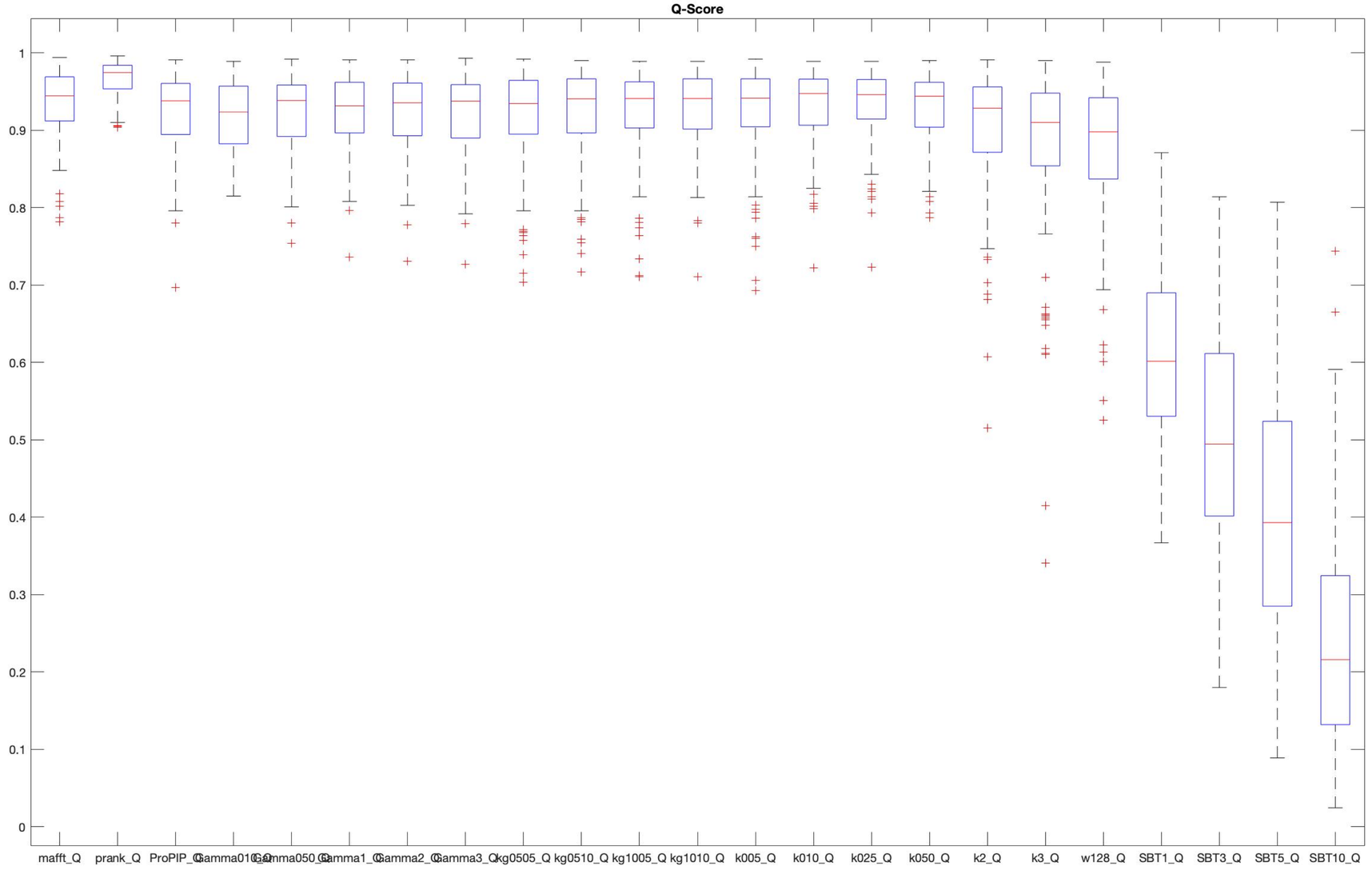
(100,8)	True (id)	w128	
		id	Xid
nIndels	11511	24670	17336
Max-IL	37	81	164
Mean	3.116	2.280	3.245
Median	1	1	2
SD	3.850	3.404	6.297

Table 6.7: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under STFT with fiter: welch and filter size: 128. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

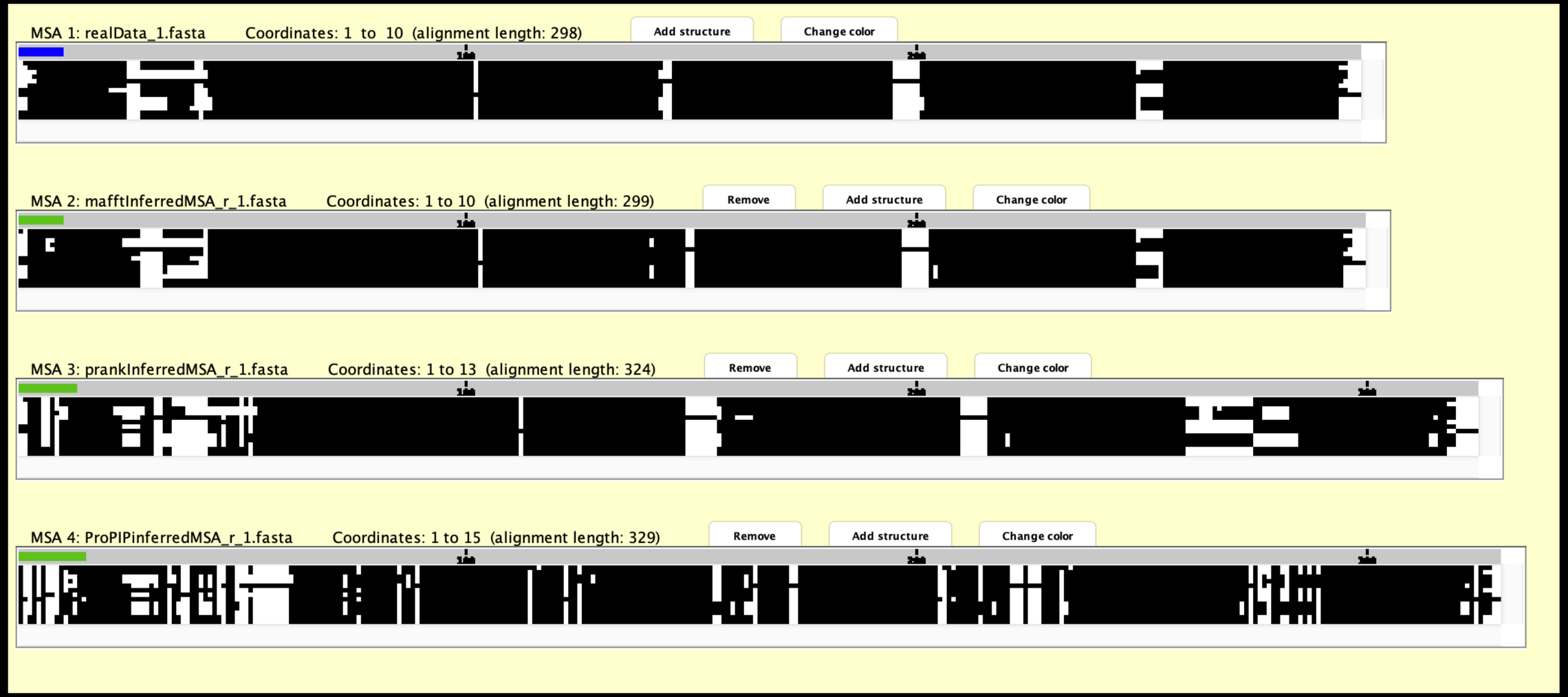

Indel Length distribution INDELible- STFT_w128
True_I_id
w128_I_id
w128_I_Xid
Frequency
Indel block size


MSA 1: out_3_TRUE_1.fasta Coordinates: 1 to 10 (alignment length: 224)
Add structure
Change color
MSA 2: mafftInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 222)
Remove
Add structure
Change color
MSA 3: prankInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 223)
Remove
Add structure
Change color
MSA 4: ProPIPinferredMSA_I_4.fasta Coordinates: 1 to 14 (alignment length: 318)
Remove
Add structure
Change color
MSA 5: ProPIP_stft_w_128_MSA_I_4.fasta Coordinates: 1 to 18 (alignment length: 361)
Remove
Add structure
Change color

Q score


Q-Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
mafft_Q
prank_Q
ProPIP_Q
kg0505_Q
kg0510_Q
kg1005_Q
kg1010_Q
k005_Q
k010_Q
k025_Q
k050_Q
k2_Q
k3_Q
w128_Q
SBT1_Q
SBT3_Q
SBT5_Q
SBT10_Q

Real data analysis results



(4,13)	True(id)	MAFFT v7.453		PRANK v.170427		ProPIP	
		id	Xid	id	Xid	id	Xid
nIndels	314	343	338	532	475	1309	866
Max-IL	43	48	48	63	81	34	59
Mean	8.930	8.481	8.610	8.325	9.324	2.828	4.275
Median	4	3	3	5	6	1	2
SD	11.517	11.883	11.952	10.170	10.725	3.368	6.267

Table 6.15: The summary statistics of the indel length distribution of protein alignments (True(id)) used in the study is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, ProPIP. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

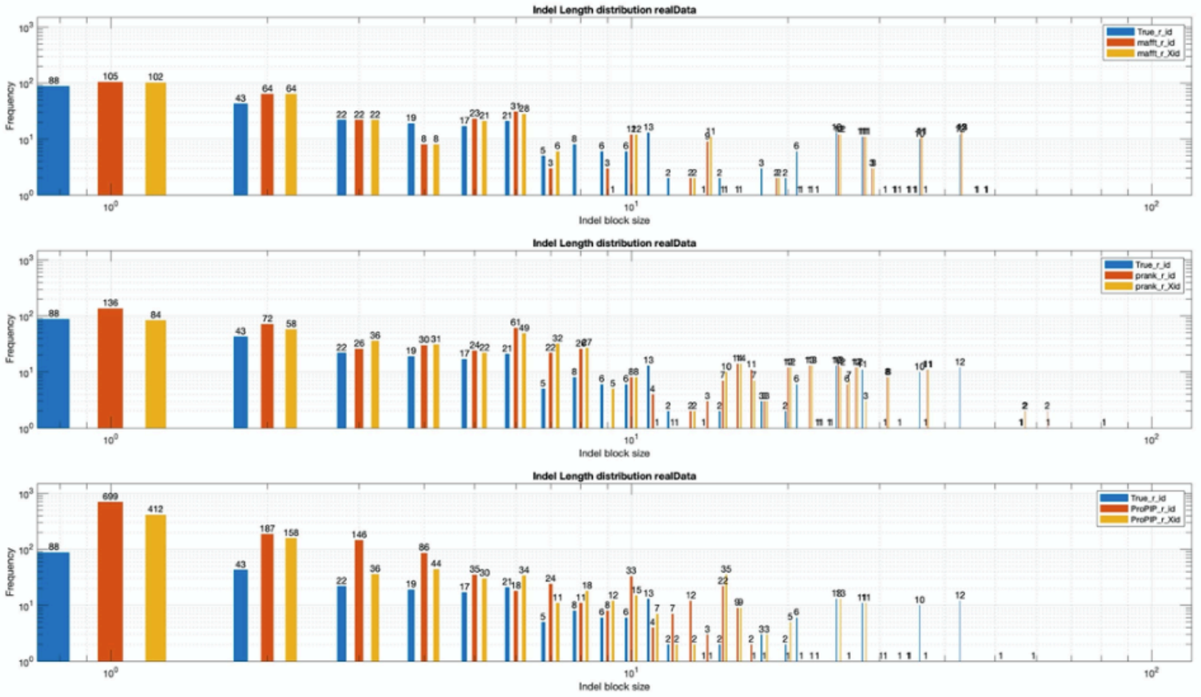
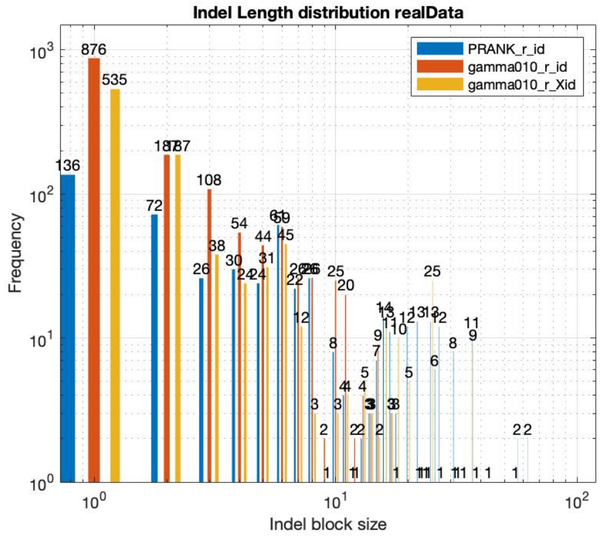
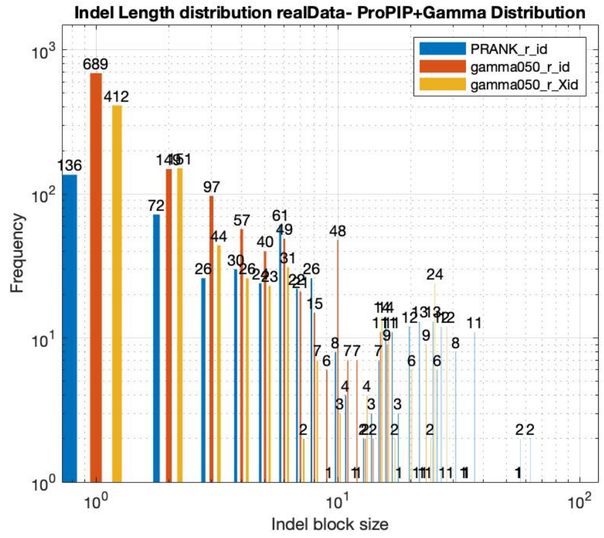
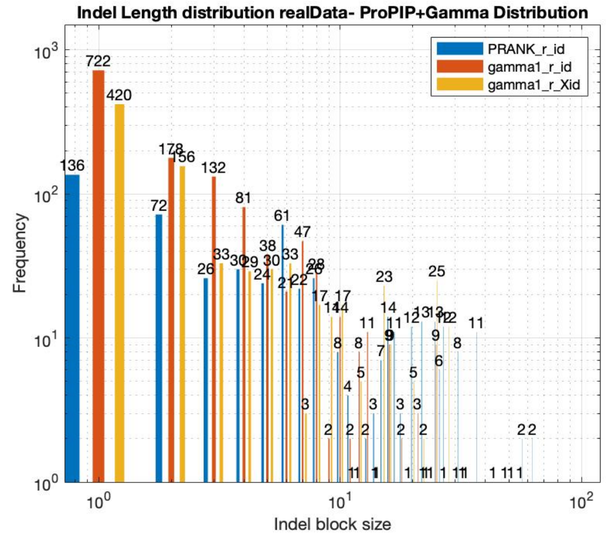
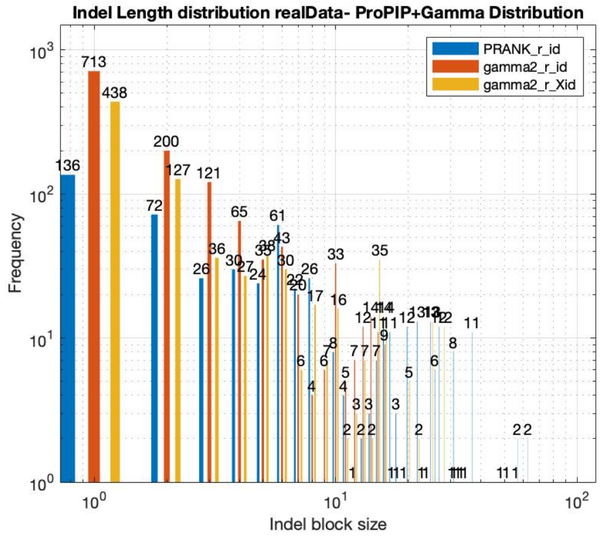


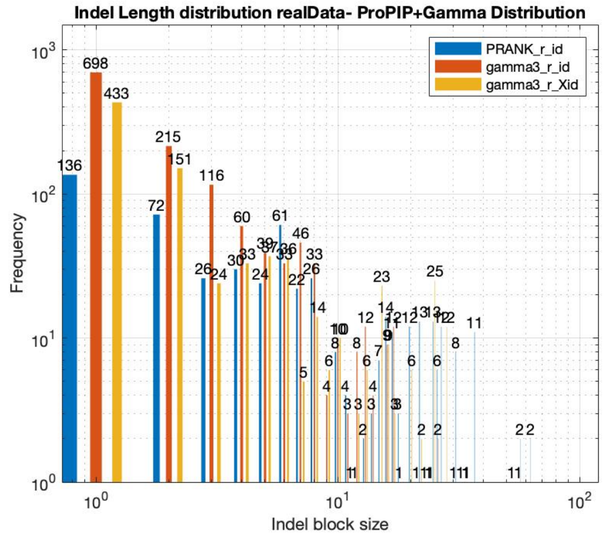
Figure 6.30: The Log-Log Plot. The indel length distribution of protein alignments (simulated using CLUSTALW1.8 (Ref: [13])) used in the study is compared with Indel length and Indel block distribution (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, and ProPIP. Note: The legend blue represents protein alignment indel length distribution, red represents inferred indel length distribution and yellow represents inferred indel block distribution.


Indel Length distribution realData
PRANK_r_id
gamma010_r_id
gamma010_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+Gamma Distribution
PRANK_r_id
gamma050_r_id
gamma050_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+Gamma Distribution
PRANK_r_id
gamma1_r_id
gamma1_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+Gamma Distribution
PRANK_r_id
gamma2_r_id
gamma2_r_Xid
Frequency
Indel block size

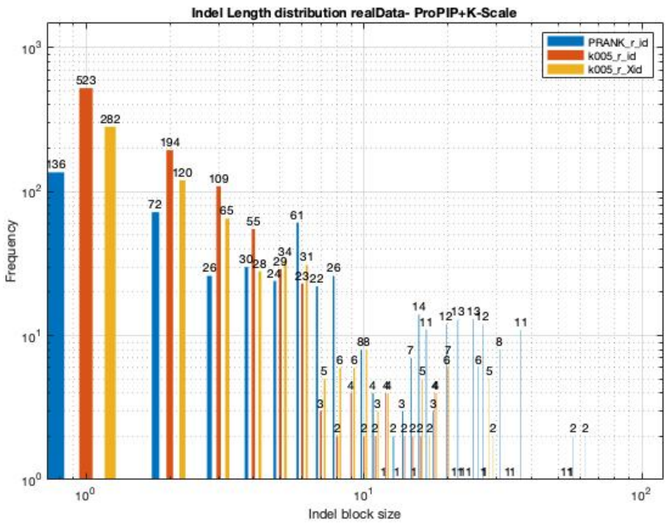

Indel Length distribution realData- ProPIP+Gamma Distribution
PRANK_r_id
gamma3_r_id
gamma3_r_Xid
Frequency
Indel block size

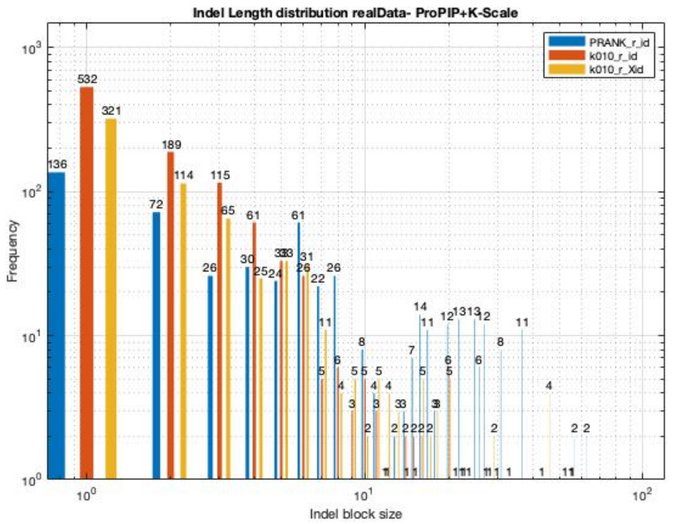
1. Discrete Gamma distribution

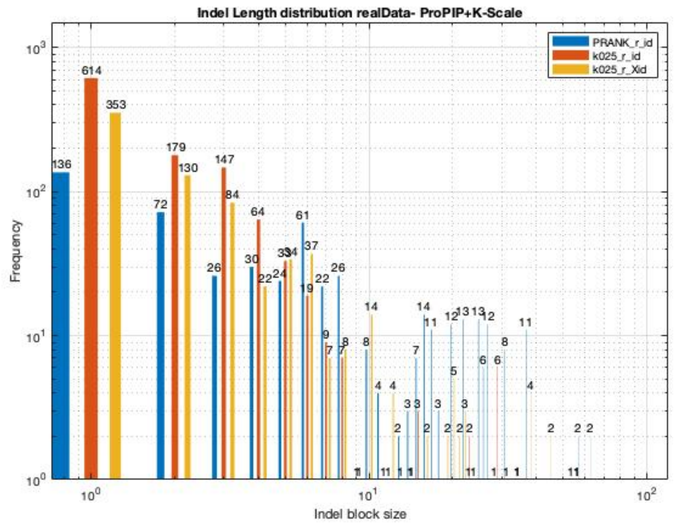
Figure 6.31: The Pixel Plot[1] (Section 5.5). A protein alignment reconstructed by PRANK v.170427 (MSA 1) is compared with MSA's generated by MAFFT v7.453 (MSA 2), ProPIP with k=1 (MSA 3), ProPIP with k=1 under Gamma n=4, α=0.10 (MSA 4), ProPIP with k=1 under Gamma n=4, α=0.50 (MSA 5), ProPIP with k=1 under Gamma n=4, α=1 (MSA 6), ProPIP with k=1 under Gamma n=4, α=2 (MSA 7), and ProPIP with k=1 under Gamma n=4, α=3 (MSA 8). Note: black pixel represents Characters and white pixel represents Indels.

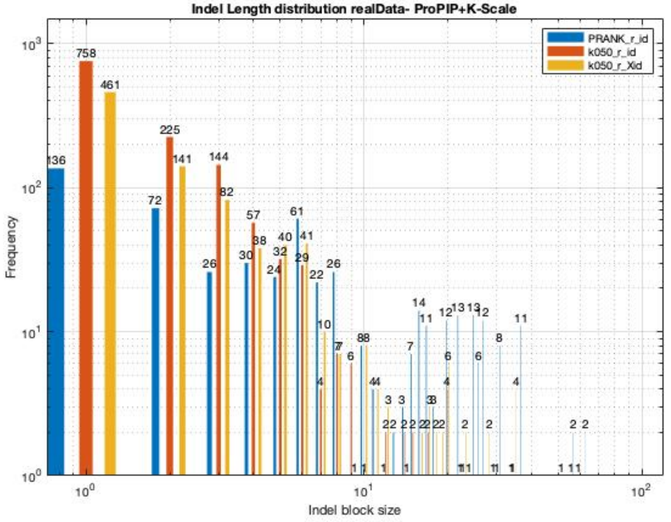
(100,8)	PRANK (id)	G0.10		G0.50		G1		G2		G3	
		id	Xid	id	Xid	id	Xid	id	Xid	id	Xid
nIndels	532	1450	966	1211	792	1311	841	1302	843	1306	845
Max-IL	63	34	43	34	56	33	51	33	51	34	52
Mean	8.325	2.4813	3.725	2.788	4.264	2.784	4.340	2.783	4.299	2.775	4.289
Median	5	1	1	1	1	1	2	1	1	1	1
SD	10.170	2.841	6.367	3.219	6.932	3.578	6.684	3.376	6.379	3.310	6.549

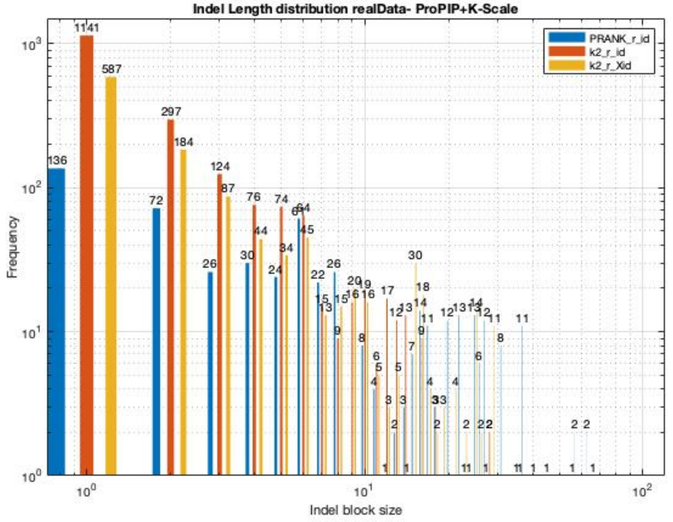
Table 6.16: The summary statistics of the indel length distribution of PRANK alignments (PRANK(id)) used in the study is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under Gamma n=4, α=0.10, ProPIP with k=1 under Gamma n=4, α=0.50 , ProPIP with k=1 under Gamma n=4, α=1, ProPIP with k=1 under Gamma n=4, α=2, ProPIP with k=1 under Gamma n=4, α=3. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

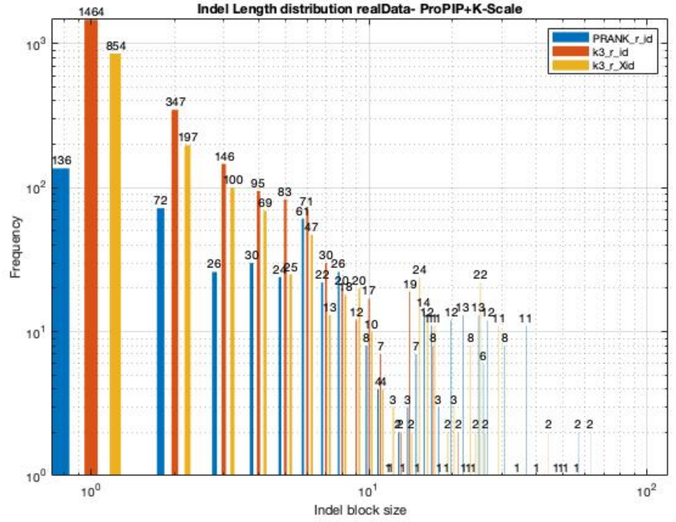

Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k005_r_id
k005_r_Xid
Frequency
Indel block size

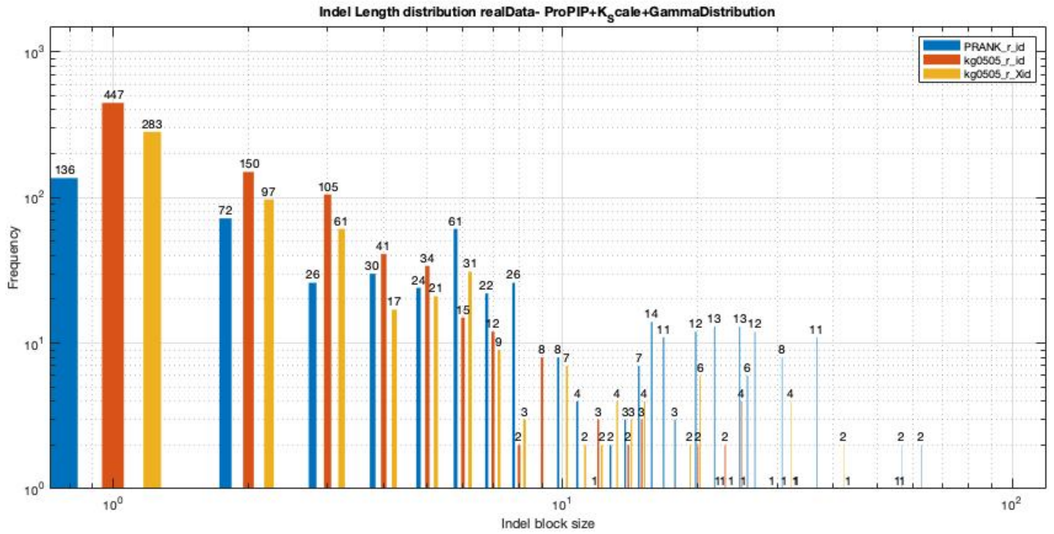

Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k010_r_id
k010_r_Xid
Frequency
Indel block size

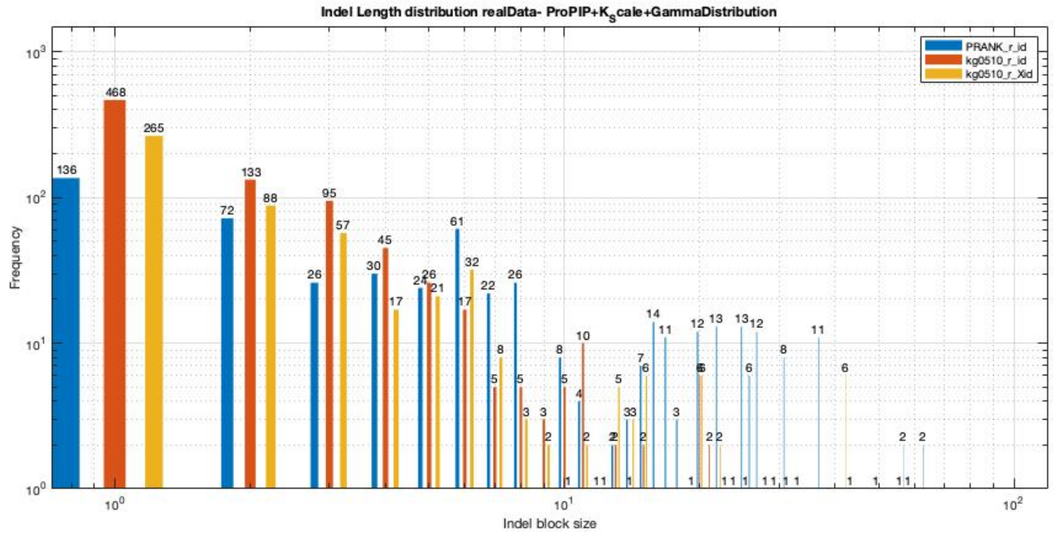

Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k025_r_id
k025_r_Xid
Frequency
Indel block size

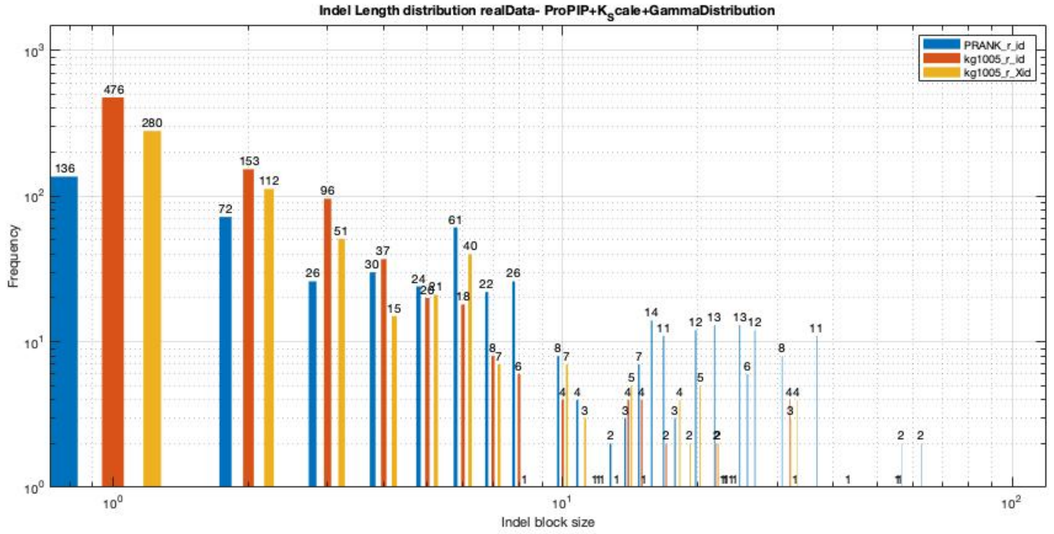

Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k050_r_id
k050_r_Xid
Frequency
Indel block size

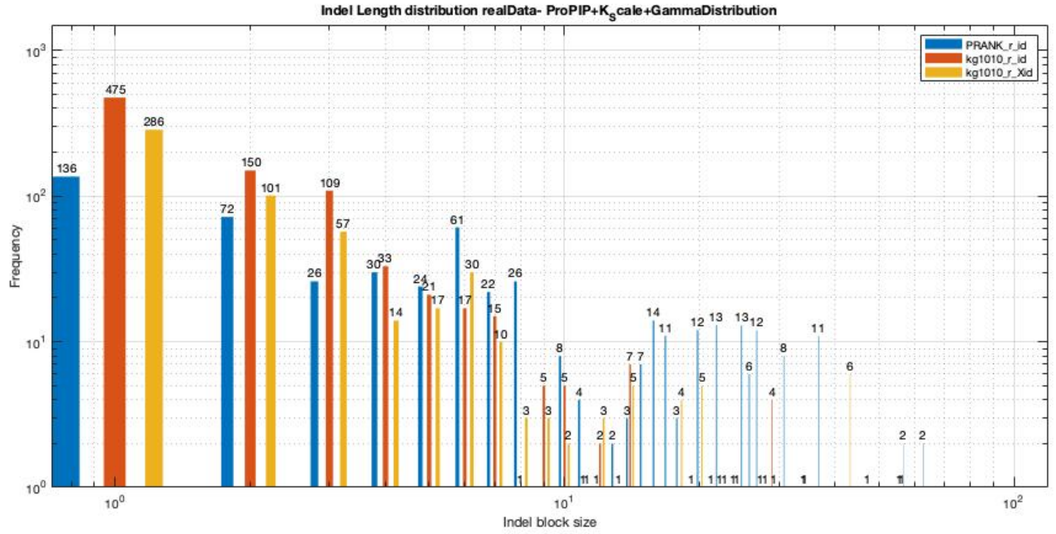

Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k2_r_id
k2_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+K-Scale
PRANK_r_id
k3_r_id
k3_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+K$_S$cale+GammaDistribution
PRANK_r_id
kg0505_r_id
kg0505_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+K$_S$cale+GammaDistribution
PRANK_r_id
kg0510_r_id
kg0510_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+K$_S$cale+GammaDistribution
PRANK_r_id
kg1005_r_id
kg1005_r_Xid
Frequency
Indel block size


Indel Length distribution realData- ProPIP+K$_S$cale+GammaDistribution
PRANK_r_id
kg1010_r_id
kg1010_r_Xid
Frequency
Indel block size

(100,8)	PRANK (id)	kg0505		kg0510		kg1005		kg1010	
		id	Xid	id	Xid	id	Xid	id	Xid
nIndels	532	831	565	827	532	837	569	847	557
Max-IL	63	33	57	33	58	33	56	34	56
Mean	8.325	2.389	3.515	2.402	3.733	2.342	3.445	2.314	3.519
Median	5	1	1	1	2	1	2	1	1
SD	10.170	3.026	5.958	3.058	6.727	3.347	5.650	3.070	6.458

Table 6.18: The summary statistics of the indel length distribution of PRANK alignments (PRANK(id)) used in the study is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=0.05 under Gamma n=4, α=0.05, ProPIP with k=0.05 under Gamma n=4, α=0.10, ProPIP with k=0.10 under Gamma n=4, α=0.05, ProPIP with k=0.10 under Gamma n=4, α=0.10. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

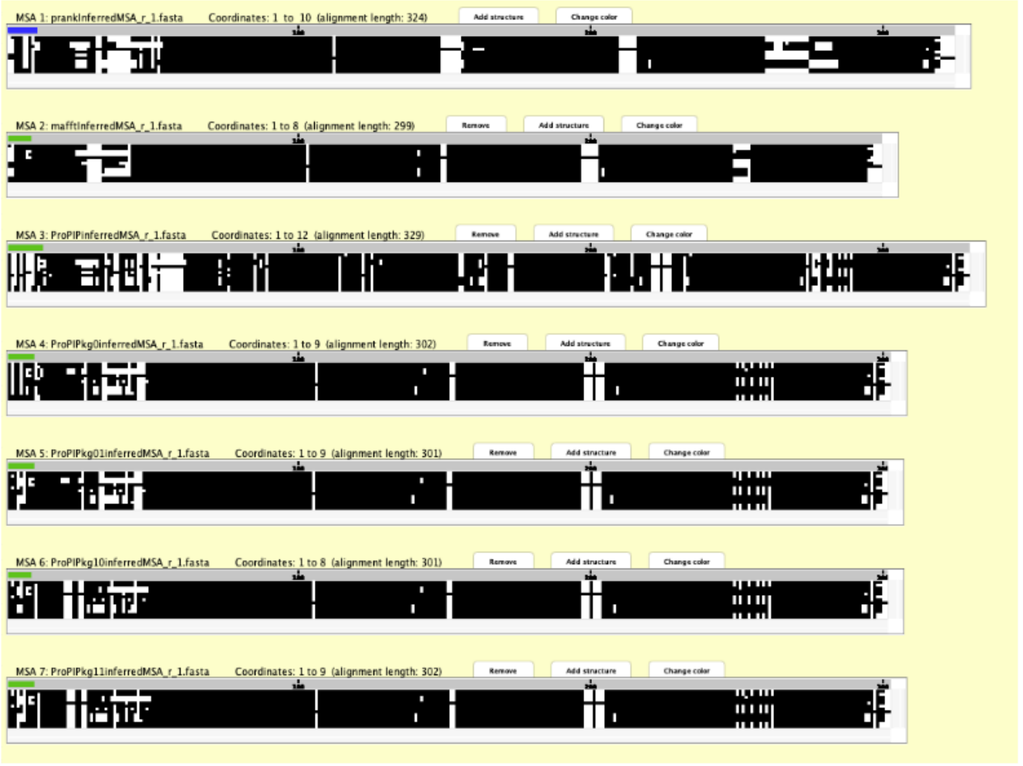
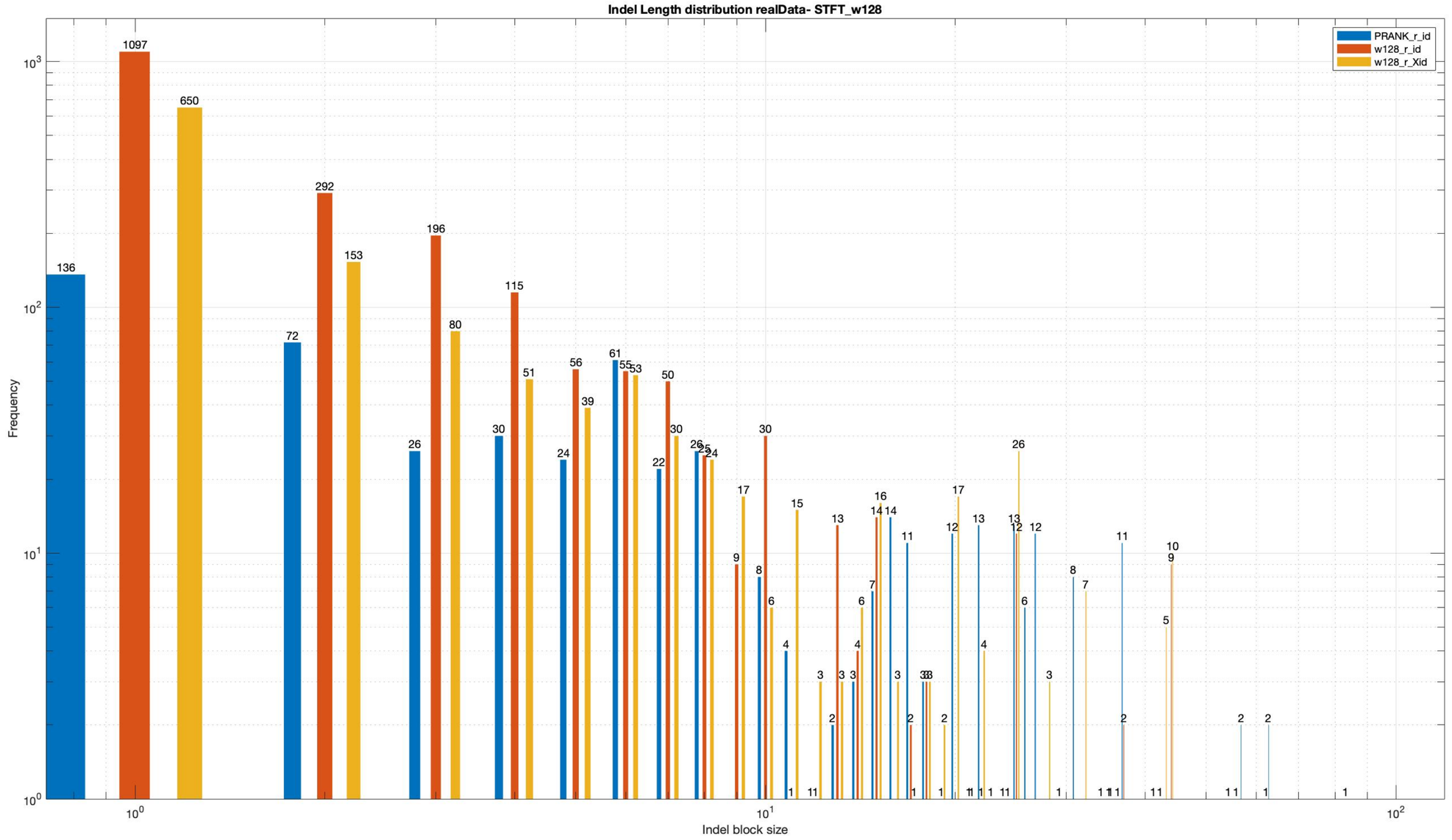


Figure 6.36: The Pixel Plot[1] (Section 5.5). A protein alignment reconstructed by PRANK v.170427 (MSA 1) is compared with MSA's generated by MAFFT v7.453 (MSA 2), ProPIP with k=1 (MSA 3), ProPIP with k=0.05 under Gamma n=4, α=0.05 (MSA 4), ProPIP with k=0.05 under Gamma n=4, α=0.10 (MSA 5), ProPIP with k=0.10 under Gamma n=4, α=0.05 (MSA 6), and ProPIP with k=0.10 under Gamma n=4, α=0.10 (MSA 7), Note: black pixel represents Characters and white pixel represents Indels.

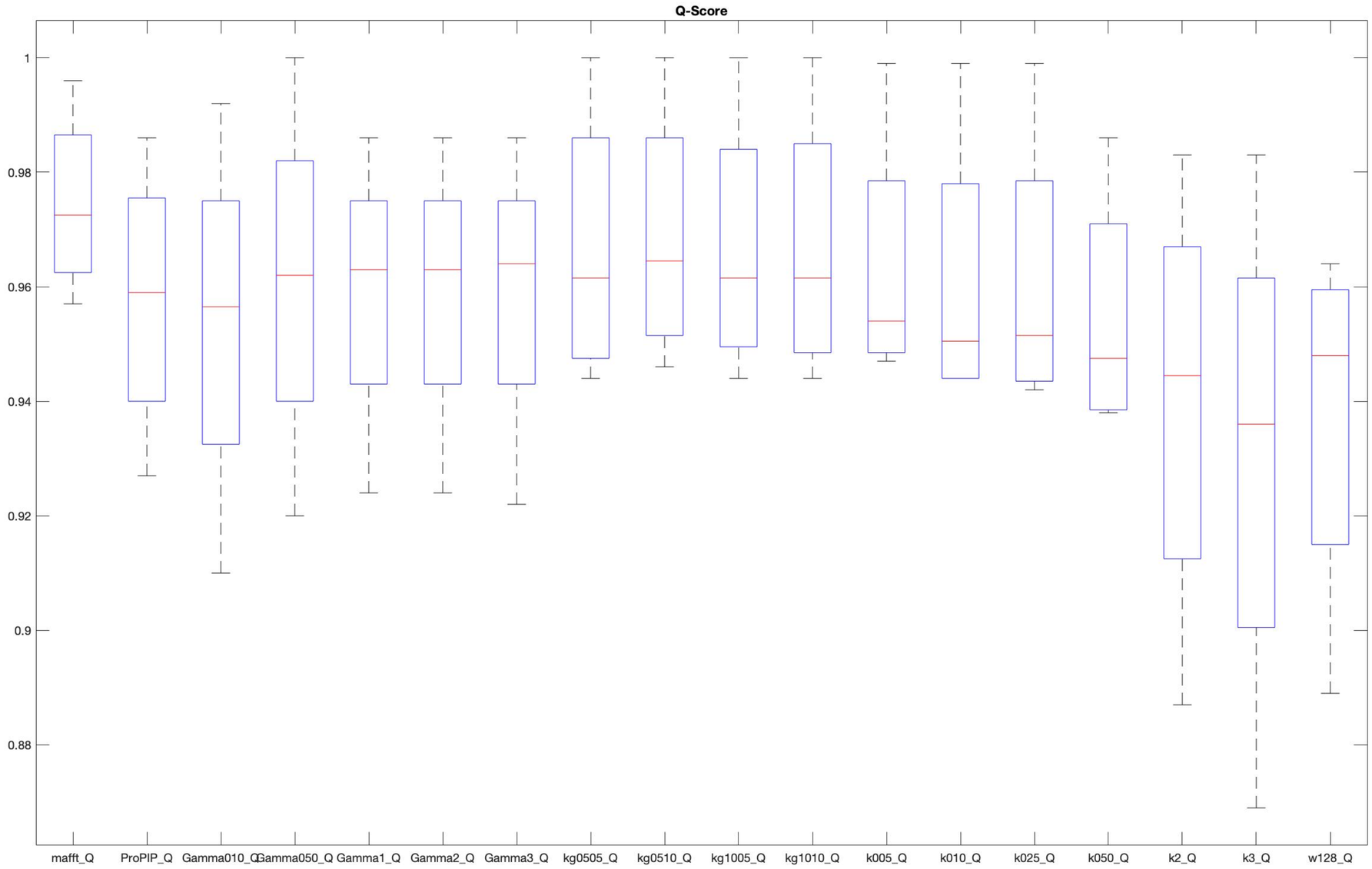

Indel Length distribution realData- STFT_w128
PRANK_r_id
w128_r_id
w128_r_Xid
Frequency
Indel block size
10^0
10^1
10^2
10^3

(4,13)	PRANK (id)	w128	
		id	Xid
nIndels	532	1991	1237
Max-IL	63	44	83
Mean	8.325	2.871	4.622
Median	5	1	1
SD	10.170	4.477	8.000

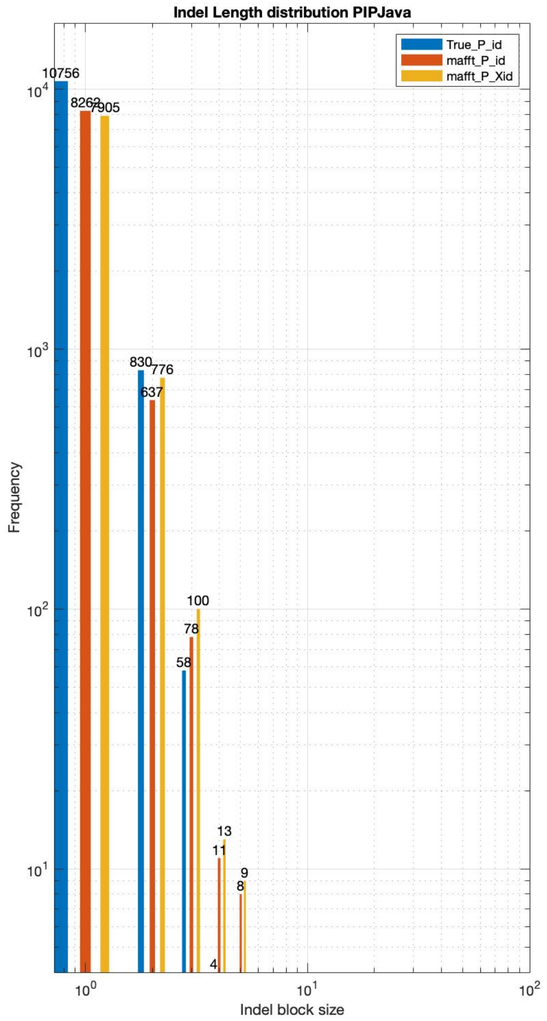
Table 6.19: The summary statistics of the indel length distribution of PRANK alignments (PRANK(id)) used in the study is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under STFT with fiter: welch and filter size: 128. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

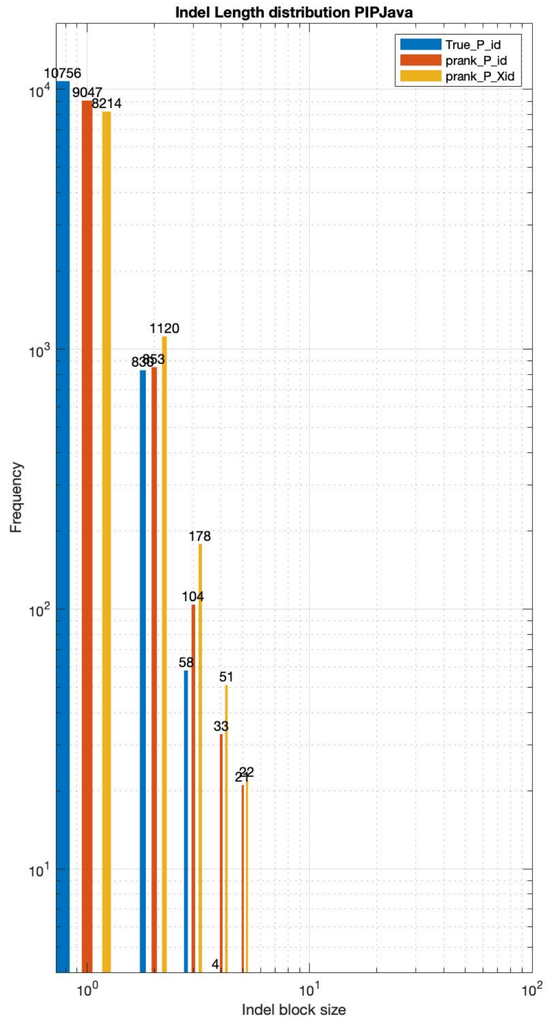
Figure 6.37: The Pixel Plot[1] (Section 5.5). A protein alignment reconstructed by PRANK v.170427 (MSA 1) is compared with MSA's generated by MAFFT v7.453 (MSA 2), ProPIP with k=1 (MSA 3), and ProPIP with k=1 under STFT with fiter: welch and filter size: 128 (MSA 4). Note: black pixel represents Characters and white pixel represents Indels.

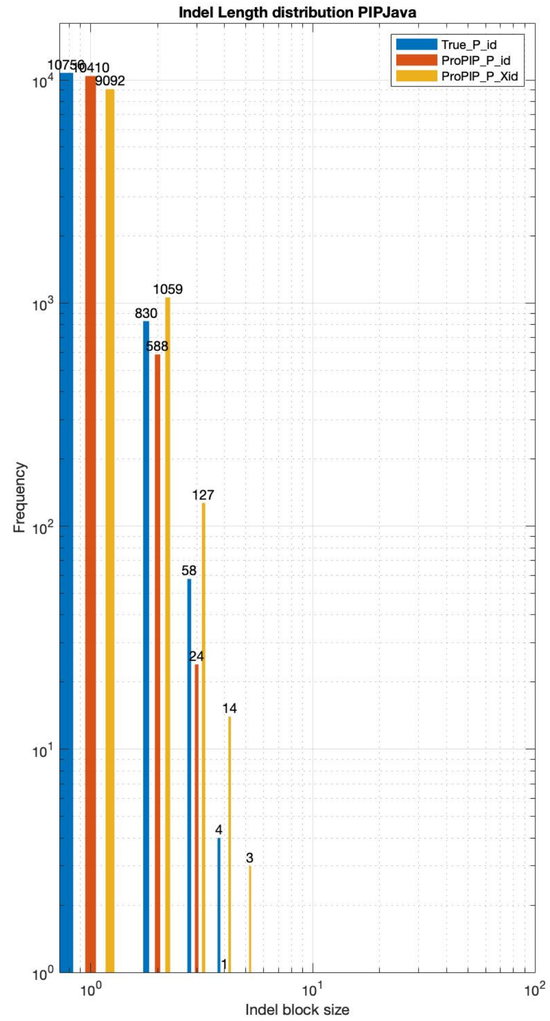
Q Score


Q-Score
1
0.98
0.96
0.94
0.92
0.9
0.88
mafft_Q
ProPIP_Q
Gamma010_Q
Gamma050_Q
Gamma1_Q
Gamma2_Q
Gamma3_Q
kg0505_Q
kg0510_Q
kg1005_Q
kg1010_Q
k005_Q
k010_Q
k025_Q
k050_Q
k2_Q
k3_Q
w128_Q

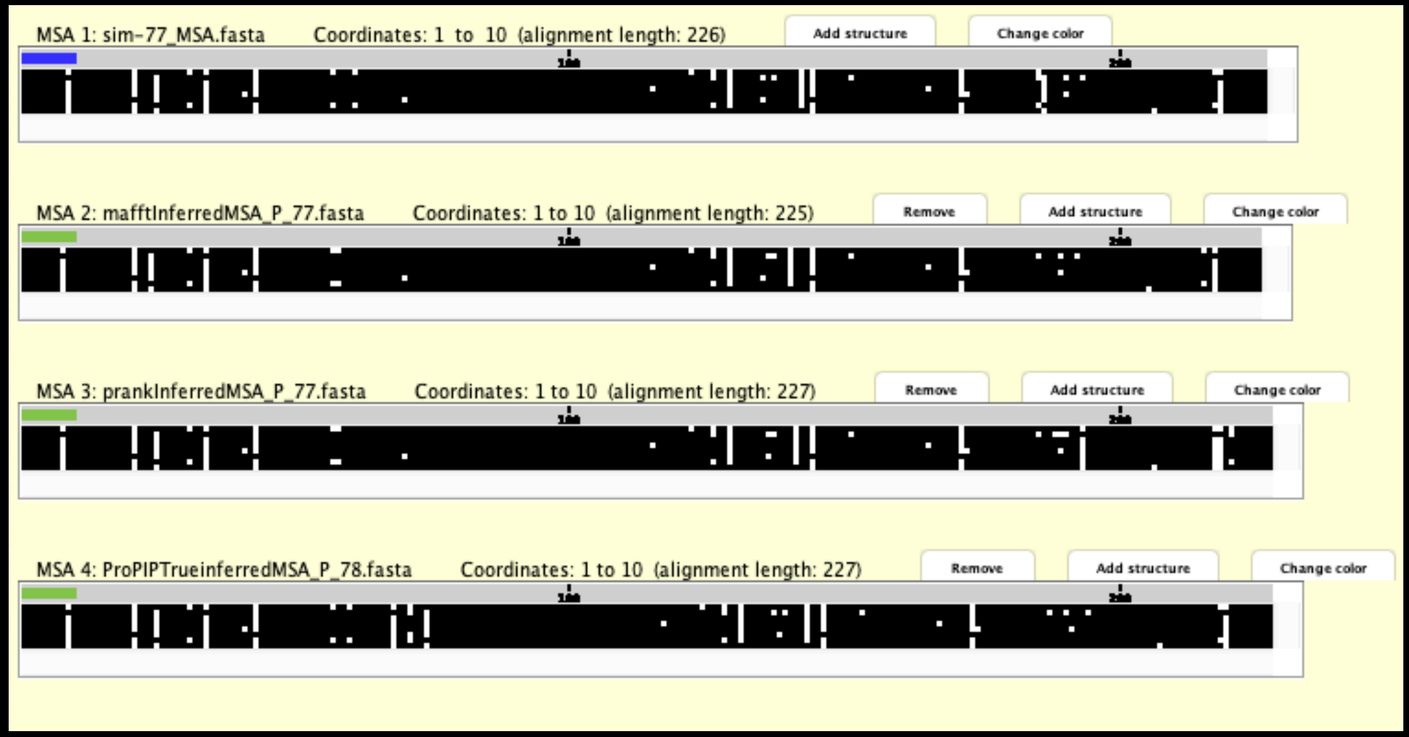
PIP data analysis results







1. Discrete Gamma distribution

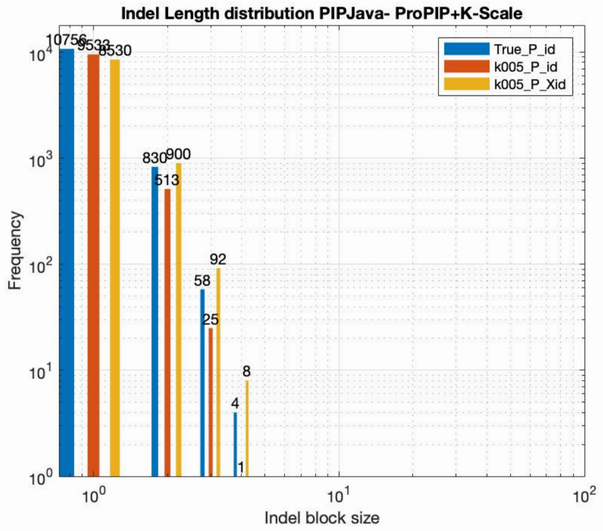


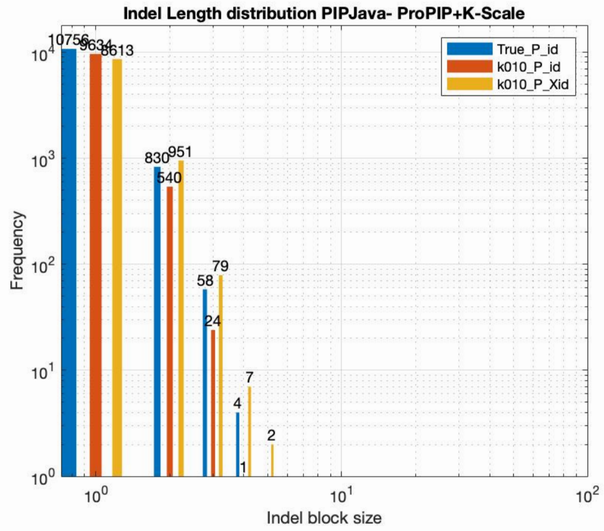
(100,8)	True (id)	MAFFT v7.453		PRANK v.170427		ProPIP	
		id	Xid	id	Xid	id	Xid
nIndels	11648	8996	8803	10058	9585	11023	10295
Max-IL	4	5	5	5	5	4	5
Mean	1.082	1.095	1.119	1.124	1.179	1.058	1.133
Median	1	1	1	1	1	1	1
SD	0.296	0.348	0.386	0.417	0.494	0.244	0.389

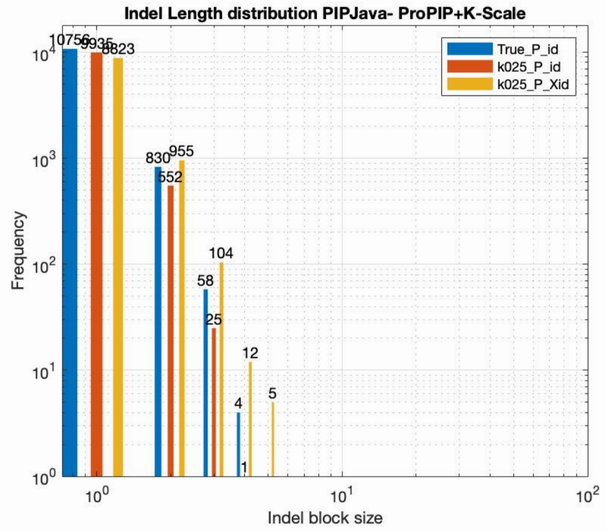
Table 6.8: The summary statistics of the 'true' Indel length distribution of PIP data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, ProPIP. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

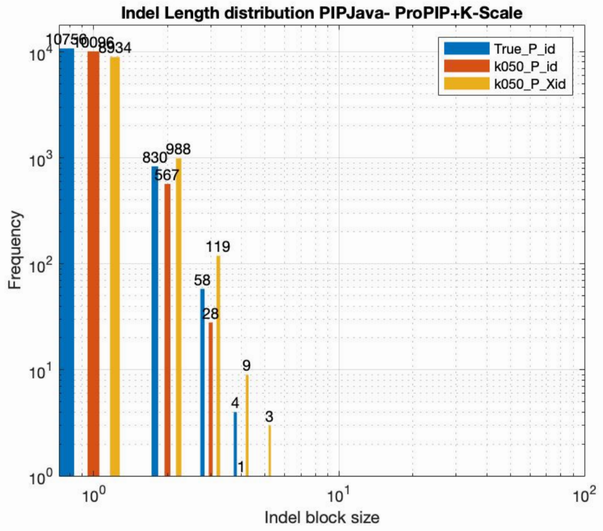
Indel-length	True (count)	default ProPIP	Discrete Gamma		k-Factor	
			α=0.10	α=3	k=0.05	k=3
1	10756	10410	10713	10391	9533	11260
2	830	588	616	595	513	695
3	58	24	43	23	25	44
4	4	1	3	1	1	2

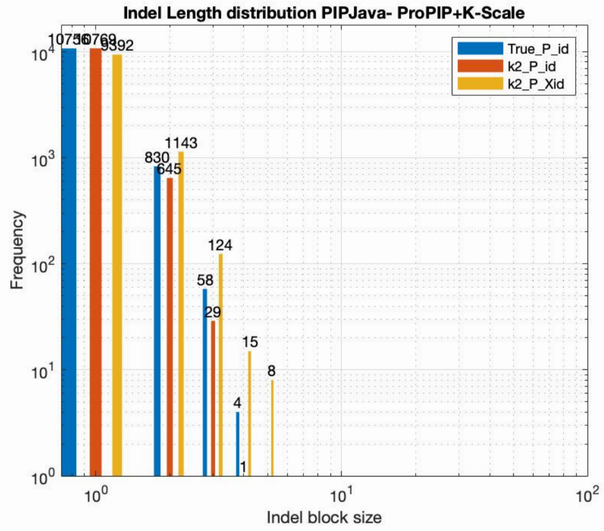
Table 6.9: Indel length frequency variation in 'true' PIP data, default ProPIP, ProPIP under Discrete Gamma distribution and ProPIP under k-Factor.

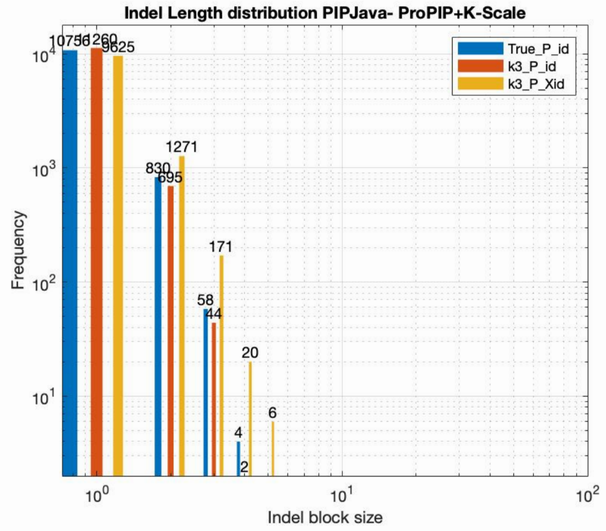

Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k005_P_id
k005_P_Xid
Frequency
Indel block size


Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k010_P_id
k010_P_Xid
Frequency
Indel block size


Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k025_P_id
k025_P_Xid
Frequency
Indel block size


Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k050_P_id
k050_P_Xid
Frequency
Indel block size

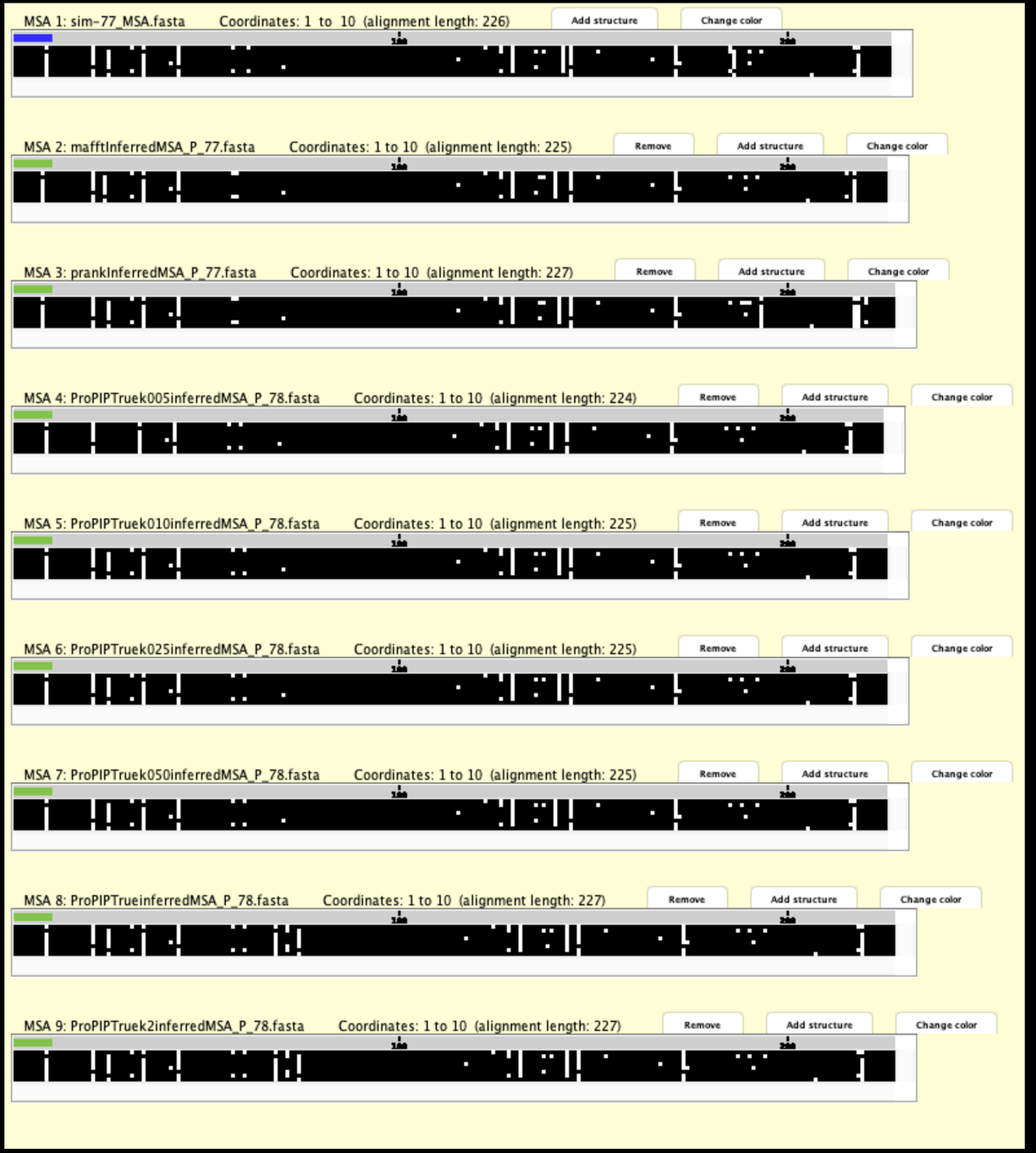

Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k2_P_id
k2_P_Xid
Frequency
Indel block size

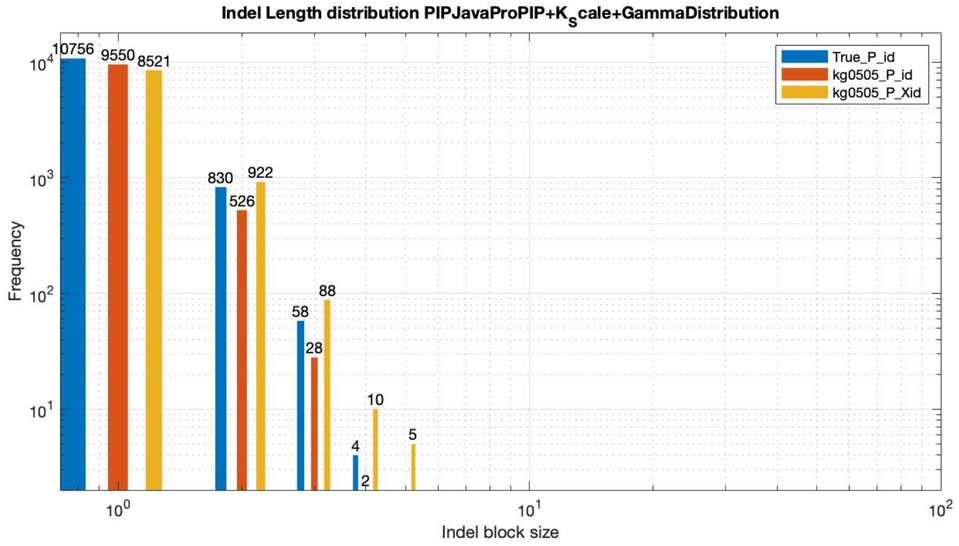

Indel Length distribution PIPJava- ProPIP+K-Scale
True_P_id
k3_P_id
k3_P_Xid
Frequency
Indel block size

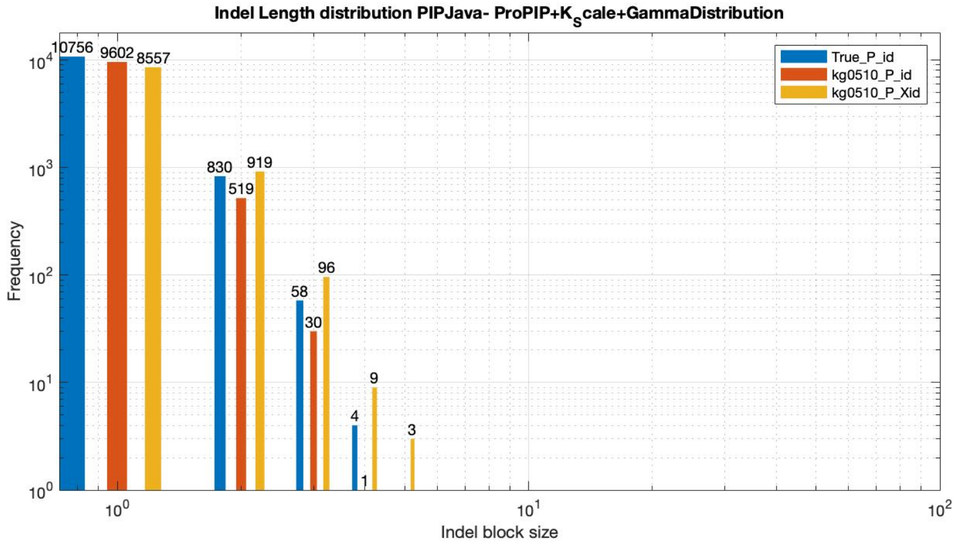
(100,8)	True (id)	k0.05		k0.10		k0.25		k0.50		k2		k3	
		id	Xid	id	Xid	id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11648	10072	9530	10199	9652	10513	9899	10692	100053	11444	10682	12001	11093
Max-IL	4	4	4	4	5	4	5	4	5	4	5	4	5
Mean	1.082	1.056	1.116	1.058	1.118	1.058	1.123	1.058	1.126	1.062	1.137	1.066	1.153
Median	1	1	1	1	1	1	1	1	1	1	1	1	1
SD	0.296	0.242	0.357	0.245	0.357	0.244	0.377	0.247	0.378	0.252	0.400	0.264	0.422

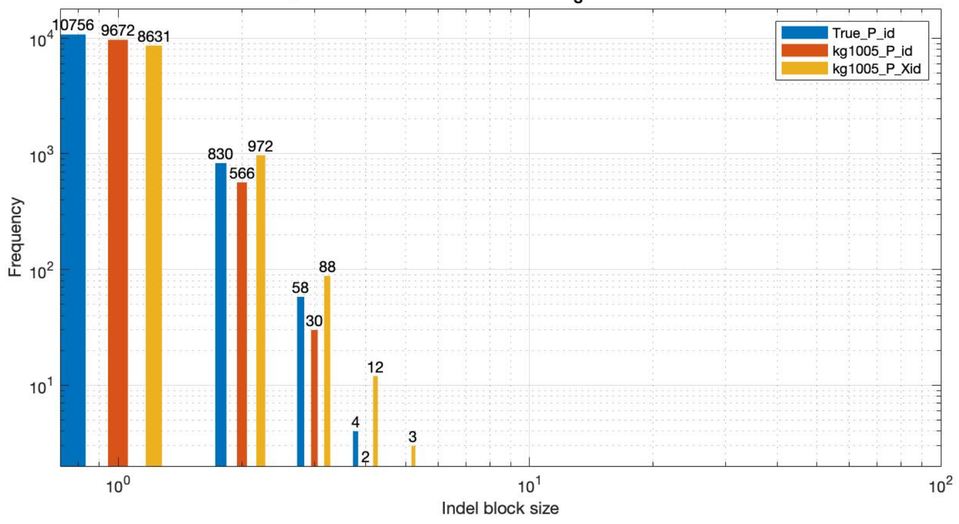
Table 6.11: The summary statistics of the 'true' Indel length distribution of PIP data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=0.05, ProPIP with k=0.10, ProPIP with k=0.25, ProPIP with k=0.50, ProPIP with k=2, ProPIP with k=3. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

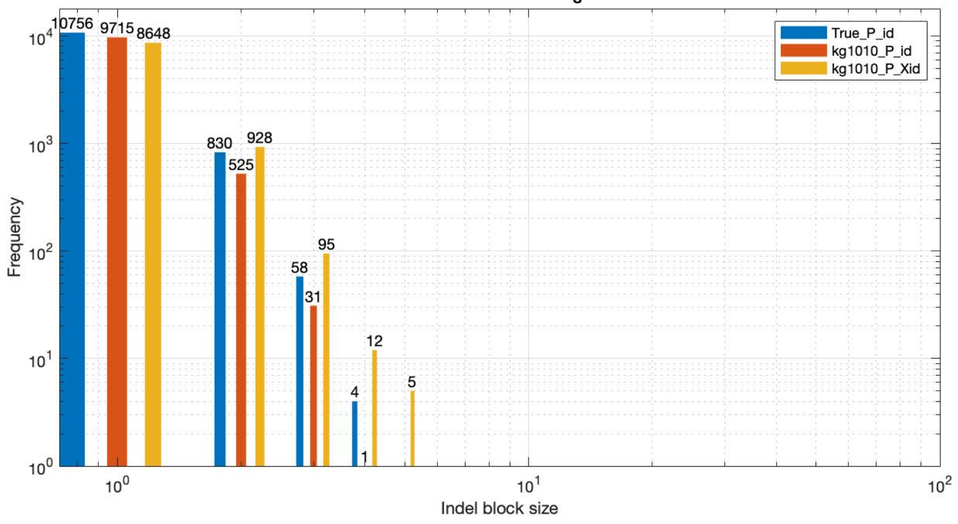
2. k-Factor

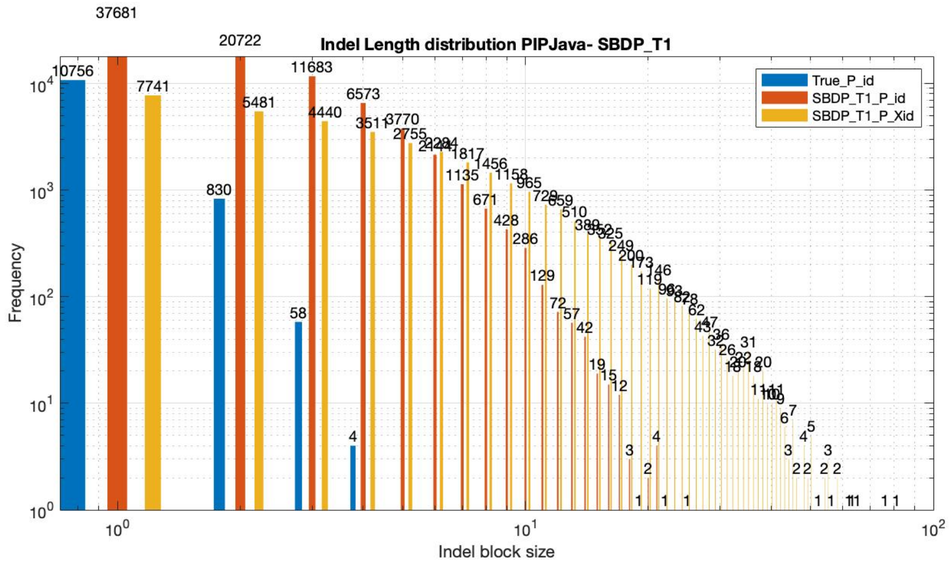

MSA 1: sim-77_MSA.fasta Coordinates: 1 to 10 (alignment length: 226)
Add structure
Change color
MSA 2: mafftInferredMSA_P_77.fasta Coordinates: 1 to 10 (alignment length: 225)
Remove
Add structure
Change color
MSA 3: prankInferredMSA_P_77.fasta Coordinates: 1 to 10 (alignment length: 227)
Remove
Add structure
Change color
MSA 4: ProPIPTruek005inferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 224)
Remove
Add structure
Change color
MSA 5: ProPIPTruek010inferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 225)
Remove
Add structure
Change color
MSA 6: ProPIPTruek025inferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 225)
Remove
Add structure
Change color
MSA 7: ProPIPTruek050inferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 225)
Remove
Add structure
Change color
MSA 8: ProPIPTrueinferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 227)
Remove
Add structure
Change color
MSA 9: ProPIPTruek2inferredMSA_P_78.fasta Coordinates: 1 to 10 (alignment length: 227)
Remove
Add structure
Change color

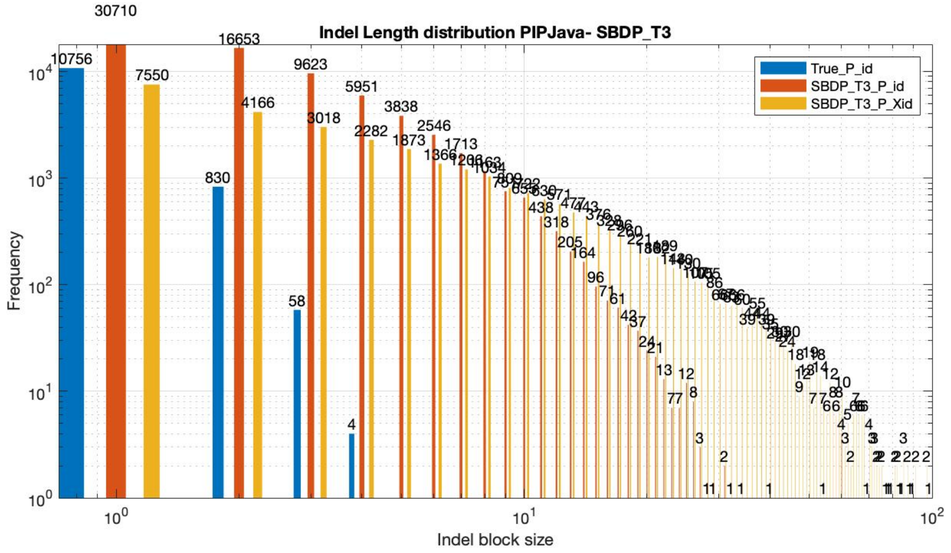

Indel Length distribution PIPJavaProPIP+K_Scale+GammaDistribution
True_P_id
kg0505_P_id
kg0505_P_Xid
Frequency
Indel block size
10756
9550
8521
830
526
922
58
28
88
4
2
10
5

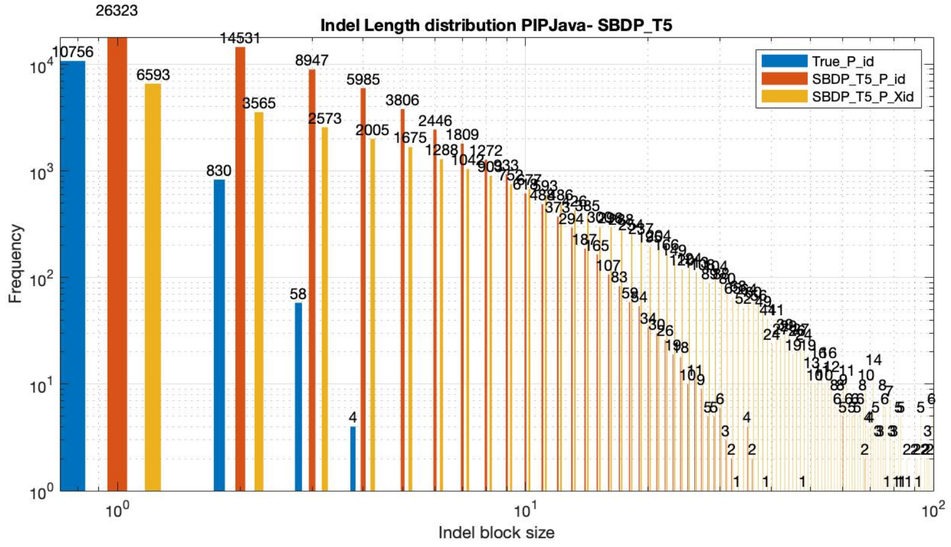

Indel Length distribution PIPJava- ProPIP+K_Scale+GammaDistribution
True_P_id
kg0510_P_id
kg0510_P_Xid
Frequency
Indel block size
10756
9602
8557
830
519
919
58
30
96
4
1
9
3

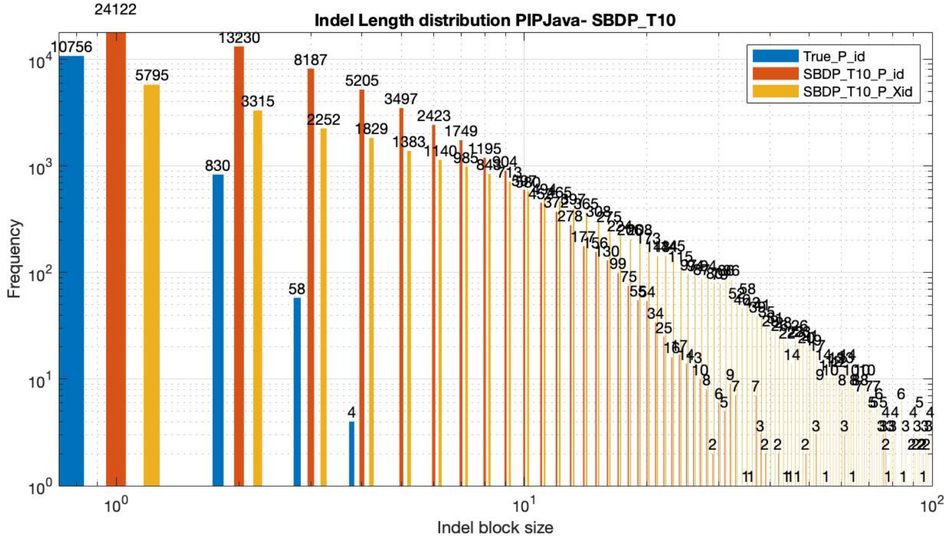

Indel Length distribution PIPJava- ProPIP+K_Scale+GammaDistribution
True_P_id
kg1005_P_id
kg1005_P_Xid
Frequency
Indel block size
10756
9672
8631
830
566
972
58
30
88
4
2
12
3


Indel Length distribution PIPJava- ProPIP+K_Scale+GammaDistribution
True_P_id
kg1010_P_id
kg1010_P_Xid
Frequency
Indel block size
10756
9715
8648
830
525
928
58
31
95
4
1
12
5


Indel Length distribution PIPJava- SBDP_T1
True_P_id
SBDP_T1_P_id
SBDP_T1_P_Xid
Frequency
Indel block size

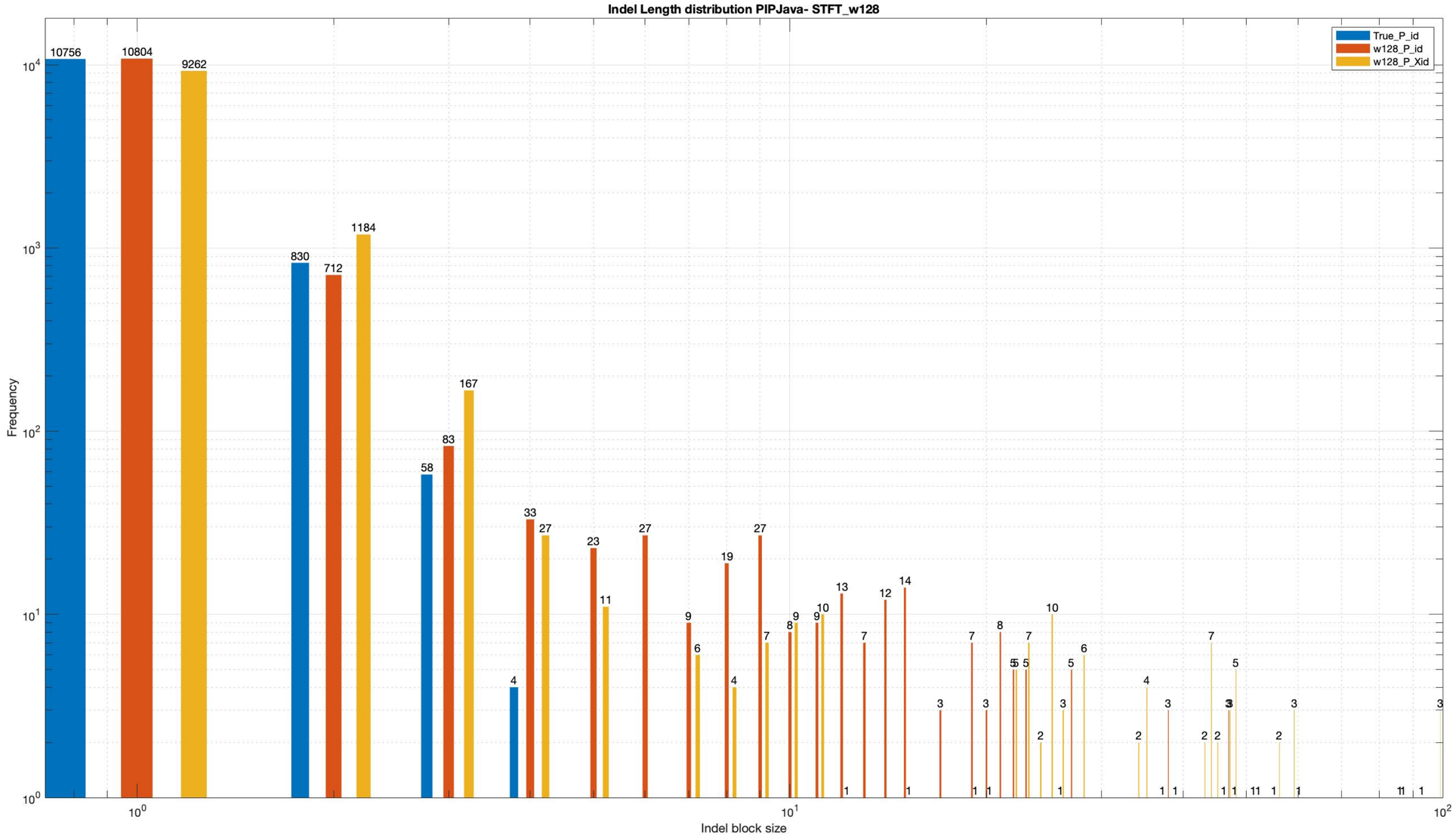

Indel Length distribution PIPJava- SBDP_T3
True_P_id
SBDP_T3_P_id
SBDP_T3_P_Xid
Frequency
Indel block size


Indel Length distribution PIPJava- SBDP_T5
True_P_id
SBDP_T5_P_id
SBDP_T5_P_Xid
Frequency
Indel block size


Indel Length distribution PIPJava- SBDP_T10
True_P_id
SBDP_T10_P_id
SBDP_T10_P_Xid
Frequency
Indel block size

(100,8)	True (id)	T1		T3		T5		T10	
		id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11648	85451	36277	75142	29963	68682	26905	63185	24162
Max-IL	4	25	81	60	113	168	168	370	410
Mean	1.082	2.303	5.425	2.756	6.911	3.016	7.700	3.153	8.245
Median	1	2	4	2	4	2	4	2	4
SD	0.296	1.782	5.600	2.588	8.975	3.213	10.843	4.785	14.121

Table 6.13: The summary statistics of the 'true' Indel length distribution of PIP data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under SBDP with T= 1, ProPIP with k=1 under SBDP with T= 3, ProPIP with k=1 under SBDP with T= 5, ProPIP with k=1 under SBDP with T= 10. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.


Indel Length distribution PIPJava- STFT_w128
True_P_id
w128_P_id
w128_P_Xid
Frequency
Indel block size
10^0
10^1
10^2
10^3
10^4
10756
10804
9262
830
712
1184
58
83
167
4
33
27
23
11
27
9
6
19
4
27
7
9
8
9
10
13
1
7
12
14
1
3
7
1
1
3
8
5
5
7
1
10
3
5
6
2
4
3
1
1
7
2
2
3
1
1
5
2
3
1
1
3

(100,8)	True (id)	w128	
		id	Xid
nIndels	11648	11853	10768
Max-IL	4	111	131
Mean	1.082	1.359	1.495
Median	1	1	1
SD	0.296	3.246	4.361

Table 6.14: The summary statistics of the 'true' Indel length distribution of PIP data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=1 under STFT with fiter: welch and filter size: 128. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

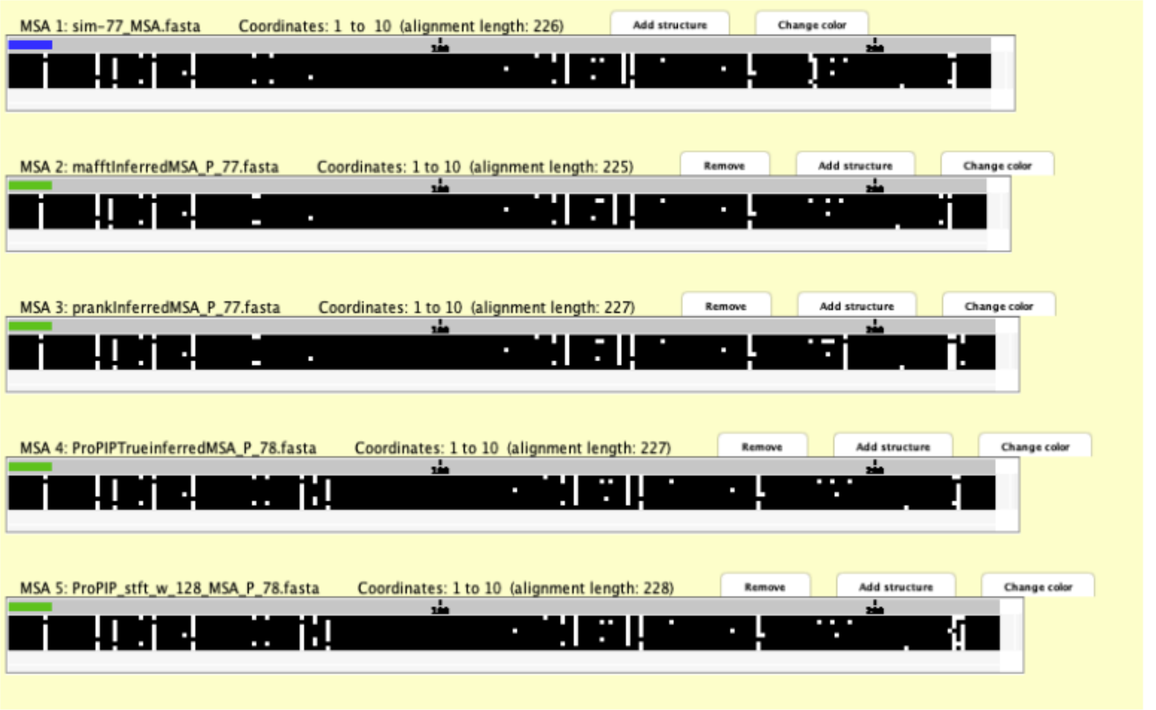
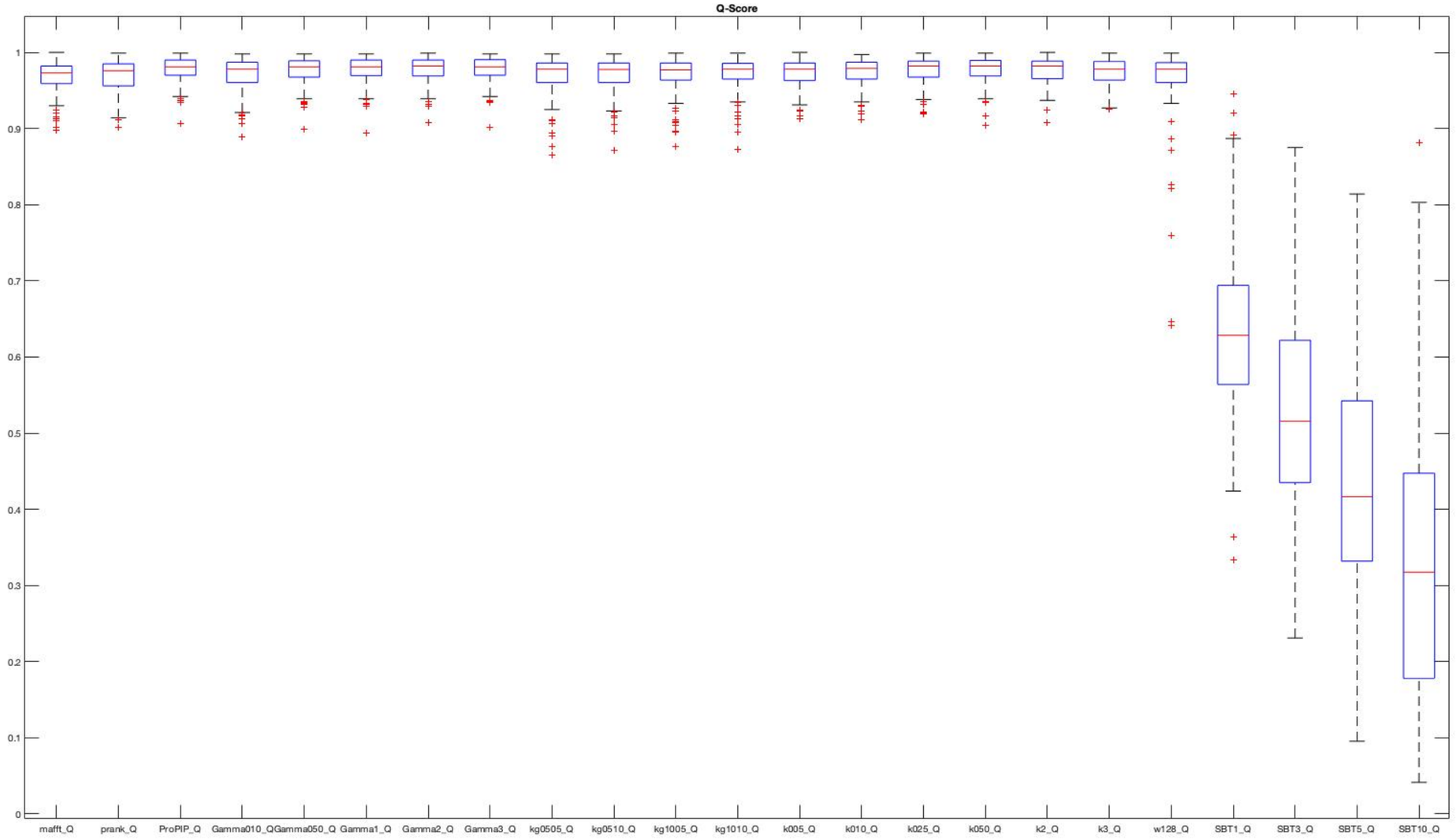


Figure 6.25: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using PIPJava is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.170427 (MSA 3), ProPIP with k=1 (MSA 4), and ProPIP with k=1 under STFT with fiter: welch and filter size: 128 (MSA 5). Note: black pixel represents Characters and white pixel represents Indels.

Q score


Q-Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
mafft_Q
prank_Q
ProPIP_Q
Gamma010_Q
Gamma050_Q
Gamma1_Q
Gamma2_Q
Gamma3_Q
kg0505_Q
kg0510_Q
kg1005_Q
kg1010_Q
k005_Q
k010_Q
k025_Q
k050_Q
k2_Q
k3_Q
w128_Q
SBT1_Q
SBT3_Q
SBT5_Q
SBT10_Q

Conclusions and Future works

THE CONCLUSION

- MAFFT and PRANK outperformed ProPIP in INDELible and real data (long indel data).
- ProPIP best fits the PIP data and performs better than other traditional aligners, MAFFT and PRANK.
- The ProPIP performance under long indel data can be improved using its additional features
- Parameter α is not suitable for INDELible data however combined with k can improve the alignment quality.
- When k= 0.05 and α= 0.05 we witnessed improvement. For relatively lower α the nIndels Increased.
- For relatively lower k — nIndels decreased Q score Increased for all data types same pattern.
- But for PIP, same pattern as above but fit obtained at k= 2
- SBDP- for relatively lower T the value of nIndels Increased — max indel length decreased
- STFT

THE FUTURE

- We observed the possibility of tuning PIP model in order to adapt long indel data
- K depends on the prior knowledge on MSA length
- K smaller or larger
- K independent or together with α. The improved performance of k and Gamma still depends on k ?
- SBDP and STFT need more tests to verify their poor performances.

THANK YOU !