Supplementary Information of "PBSIM2: a simulator for long read sequencers with a novel generative model of quality scores"

Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada*

A Supplementary Tables

Table S1: Simulators for long reads

Simulator	Long read	Error model
PBSIM [1]	PacBio	nucleotide sequence-independent error model
DAZZ_DB/simulator [a]	PacBio	nucleotide sequence-independent error model
ReadSim [2]	PacBio, Nanopore	nucleotide sequence-independent error model
SimLoRD [3]	PacBio	nucleotide sequence-independent error model
SiLiCO [4]	PacBio, Nanopore	nucleotide sequence-independent error model
LongISLND [5]	PacBio, Nanopore	entended-kmer based error model
NanoSim [6]	Nanopore	alignment-based trained model, which does not use k-mer error bias
SNaReSim [7]	Nanopore	k-mer based error model
NPBSS [8]	PacBio	nucleotide sequence-independent error model
DeepSimulator [9]	Nanopore	pore model generates raw signal, and basecaller converts raw signal into fastq
DeepSimulator1.5 [10]	Nanopore	pore model generates raw signal, and basecaller converts raw signal into fastq
Naopore SimulatION [11]	Nanopore	pore model generates raw signal, and basecaller converts raw signal into fastq
PaSS [12]	PacBio	k-mer based error model
Badread [13]	PacBio, Nanopore	k-mer based error model

[a] https://github.com/thegenemyers/DAZZ_DB/blob/master/simulator.c

*To whom correspondence should be addressed. Department of Electrical Engineering and Bioscience Faculty of Science and Engineering, Waseda University 55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo 169-8555, Japan. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp

Table S2: Datasets for PacBio sequencers

Reference	Chemistry	ReadlLength mean	SD	Read accuracy mean	SD	URL
E-coli K12 MG1655	P4–C2	5,254	3,677	81%	5%	https://github.com/PacificBiosciences/DevNet/wiki/E.-coli-20kb-Size-Selected-Library-with-P4-C2
S.cerevisiae	P4–C2	5,856	4,422	82%	5%	https://github.com/PacificBiosciences/DevNet/wiki/Saccharomyces-cerevisiae-W303-Assembly-Contigs
N.creassa	P4–C3	5,581	3,838	81%	4%	https://github.com/PacificBiosciences/DevNet/wiki/Neurospora-Crassa-\%28Fungus\%29-Genome\%2C-Epigenome\%2C-and-Transcriptome
H.sapiens	P5–C3	6,383	5,562	83%	3%	https://github.com/PacificBiosciences/DevNet/wiki/H_sapiens_54x_release [a]
D.melanogaster	P5–C3	10,095	7,224	84%	2%	https://github.com/PacificBiosciences/DevNet/wiki/Drosophila-sequence-and-assembly [b]
E.coli K12 MG1655	P6–C4	8,582	6,953	85%	4%	https://github.com/PacificBiosciences/DevNet/wiki/E.-coli-Bacterial-Assembly
C.elegans	P6–C4	11,560	7,667	85%	3%	https://github.com/PacificBiosciences/DevNet/wiki/C.-elegans-data-set [c]

SD: standard deviation. Read accuracy was computed from quality scores.
[a] Only m130929_024849 and m130929_161837 were used.
[b] Only Dro1_24NOV2013_398 was used.
[c] Only m140928_184123, m140928_230547, m140928_033247 and m140928_075857 were used.

Table S3: Datasets for Oxford Nanopore sequencers

Reference	Chemistry	Base-caller	Read length mean	SD	Read accuracy mean	SD	URL
H.sapiens	R9.4	Guppy 3.6.0	16,975	38,159	80%	15%	https://github.com/nanopore-wgs-consortium/CHM13 [a]
C.elegans	R9.4	poretools	3,962	5,250	83%	6%	https://www.ncbi.nlm.nih.gov/sra/SRX2764157
E.coli O127:H6	R9.4	Guppy 3.1.5	9,231	14,187	90%	3%	https://www.ncbi.nlm.nih.gov/sra/SRX8094377
S.cerevisiae	R9.4	poretools	12,473	14,182	86%	10%	https://www.ncbi.nlm.nih.gov/sra/SRX2849976
D.melanogaster	R9.5	Albacore 2.1.0	6,699	6,703	92%	6%	https://www.ncbi.nlm.nih.gov/sra/SRX3676783, Run=SRR6821890
P.koreensis	R9.5	Albacore 2.1.3	25,740	15,090	84%	8%	https://www.ncbi.nlm.nih.gov/sra/SRX3923115
R.sphaeroides	R9.5	Guppy 3.0.3	5,650	7,247	83%	8%	https://www.ncbi.nlm.nih.gov/sra/SRX7341766
C.armoricus	R9.5	Albacore 2.3.4	9,966	9,004	90%	3%	https://www.ncbi.nlm.nih.gov/sra/SRX6887881
E.coli K12 MG1655	R10.3	Guppy 3.4.5	6,397	4,708	85%	3%	https://www.ncbi.nlm.nih.gov/sra/ERX3900444

SD: standard deviation. Read accuracy was computed from quality scores.
[a] 100,000 reads with a length of 100 bp or more were sampled from the rel5 dataset.

Table S4: Top30 6-mer with high rate of insertion

C.elegans P6–C4		*R.sphaeroides* R9.5		*E-coli* K12 R10.3	
6-mer	rate	6-mer	rate	6-mer	rate
GTATAG	7.1%	TAGAGT	10.4%	CCTAGA	8.9%
GCACGC	7.0%	TATTAA	9.7%	CTAGAT	7.8%
GTATAC	6.9%	TTAGCT	9.1%	CCAGGA	7.8%
CACGCG	6.9%	TAATTA	9.1%	CTAGTC	7.7%
TACCGA	6.7%	CACTAA	9.0%	GCCTGG	7.6%
TACGTC	6.7%	TAAGGA	8.9%	CCAGGC	7.5%
CGGTGT	6.7%	TGTACG	8.8%	TCCTGG	7.4%
TCACGC	6.6%	ACTAAG	8.6%	CCAGGT	7.3%
GGTGTA	6.6%	GGTACA	8.1%	CCTGGT	7.3%
TACGCG	6.5%	ACACTA	7.8%	CGATCG	7.2%
ACACCG	6.5%	TAACAT	7.7%	ACCAGG	7.2%
GGCGTA	6.5%	AATAAC	7.6%	GAGATC	7.2%
GTACGT	6.5%	GAGTCA	7.5%	GATCTA	7.2%
GCGCGA	6.5%	CTGTAC	7.5%	TCTAGT	7.2%
GTATAA	6.5%	GTACAC	7.4%	ACCTGG	7.0%
TGTAGC	6.4%	GTACAT	7.3%	ATCTAG	6.8%
ACCGCG	6.4%	CCTAGT	7.2%	CGATCT	6.8%
TAACGC	6.4%	CTAACA	7.1%	AGATCG	6.8%
CGCGCA	6.4%	ATACAG	6.9%	CCCAGG	6.7%
ACGTCG	6.4%	AGTACA	6.9%	TGATCG	6.6%
GCGGTA	6.4%	CGAGTC	6.9%	GATCTC	6.6%
GTACCG	6.4%	ACTTTA	6.8%	AGATCA	6.6%
ACCGCA	6.4%	TCTAGT	6.8%	GATCAT	6.5%
TGCGGT	6.4%	ACCTTA	6.7%	GATCGG	6.5%
CGCGCG	6.3%	CATGTA	6.7%	GATCGC	6.5%
GTGTAG	6.3%	CTAGTG	6.7%	TGATCA	6.5%
GTCGTA	6.3%	TAAGAG	6.7%	GTCTAA	6.4%
ACGCGT	6.2%	GTCATA	6.6%	GATCAG	6.4%
TTATAC	6.2%	ATCTAA	6.6%	GATCGA	6.4%
GTGCAG	6.2%	TATACA	6.6%	ACGATC	6.4%

Error rates were calculated for 6 bp long sequence on the reference sequence in the alignment of real reads, while shifting the 6 bp long sequence by 1 bp from end to end. Next, the averaged error rate of each 6-mer was calculated.

Table S5: Top 30 6-mer with high rate of deletion

C.elegans P6–C4		*R.sphaeroides* R9.5		*E-coli* K12 R10.3	
6-mer	rate	6-mer	rate	6-mer	rate
GGGCCC	9.6%	GGGGGG	22.9%	GGGGGG	27.7%
GGGGCC	8.9%	CCCCCC	22.7%	CCCCCC	27.3%
GCCCCC	8.6%	TTTTTT	18.9%	TCCCCC	20.7%
AGGGGG	8.6%	AGGGGG	18.3%	GGGGGA	20.0%
CCCGGG	8.4%	CCCCCT	18.3%	CCCCCT	19.7%
CCCCCG	8.3%	TCCCCC	17.5%	AGGGGG	19.1%
GGGGGT	8.2%	AAAAAA	17.3%	CCCCCA	18.9%
CCCCCA	8.2%	CTTTTT	17.2%	CCCCCG	18.2%
CCCCCT	8.0%	GGGGGA	16.7%	TGGGGG	18.1%
TGGGGG	8.0%	AAAAAG	14.9%	CGGGGG	17.7%
CCGGGG	8.0%	AAGGGG	14.7%	GGGGGC	17.6%
GGGGAC	8.0%	CCCCTT	14.7%	ACCCCC	17.3%
ACCCCC	8.0%	CCTTTT	14.5%	GCCCCC	17.2%
CCCCCC	7.9%	GGGGGT	14.4%	GGGGGT	17.2%
GCCGGG	7.9%	CGGGGG	14.4%	TTCCCC	14.6%
AGGGCC	7.8%	TTAAGC	14.3%	GGGGAA	14.6%
CCCCGG	7.8%	CTCCCC	14.2%	CCCCGA	14.5%
GGCCCC	7.8%	CCCCCG	14.2%	CAGGGG	14.0%
GGGGTT	7.8%	ACCCCC	14.1%	CCCCTC	14.0%
GGGGGA	7.7%	TAAAAA	13.9%	CCGGGG	13.8%
GGGGGC	7.7%	AGGGGA	13.9%	CCCCAG	13.7%
GAGGGG	7.7%	CTCTCT	13.8%	TCGGGG	13.6%
GGCCCT	7.6%	TAAGGG	13.8%	CGGGGA	13.5%
CCGGCC	7.5%	CTTAAG	13.7%	AAAAAA	13.5%
GGCCCA	7.5%	CCCCTC	13.7%	TGCCCC	13.4%
TCCCCC	7.4%	CAAAAA	13.7%	GGGGAG	13.4%
GCCCGG	7.4%	GCCCCC	13.6%	GGGGTA	13.4%
AGGGGC	7.4%	AAAAGG	13.6%	TCCCCG	13.2%
GGGGGG	7.3%	TCCCCT	13.5%	GGGGCG	13.2%
GGGCCT	7.3%	TAAAAG	13.5%	CCCCGG	13.2%

Error rates were calculated the same as in Table S4.

Table S6: Top 30 6-mer with high rate of substitution

C.elegans P6–C4		*R.sphaeroides* R9.5		*E-coli* K12 R10.3	
6-mer	rate	6-mer	rate	6-mer	rate
GTCTGG	2.7%	CGAATC	10.8%	GATCTA	14.2%
GGGGGG	2.3%	GATTCG	10.4%	TAGATC	14.1%
AGTCTG	2.1%	ACTAGG	9.0%	CTACTA	12.8%
ATTGGG	2.1%	GTCTAG	8.5%	CAGATC	12.8%
ACGGCC	2.0%	TGTCTA	8.1%	TCGATC	12.5%
CGGGGG	2.0%	TACTAG	7.7%	GATCGA	12.3%
GGGGGT	1.9%	GACTAG	7.6%	CTAGAC	12.3%
ACCCCC	1.9%	CTAGGC	7.5%	TCTAGT	12.2%
CCCCCC	1.9%	GCCTAG	7.4%	TCTAGG	12.2%
TTGGGA	1.8%	CTAGAT	7.3%	GATCAA	12.2%
CCCCCT	1.7%	GAATCG	7.3%	CTAGTA	12.2%
GTGGGG	1.7%	CTAGTA	7.3%	CCGATC	12.2%
TGGGGG	1.7%	CGATTC	6.9%	CTAGTG	12.1%
CCCCCA	1.7%	GAATCC	6.9%	CGATCA	12.0%
CACCCC	1.6%	GGATTC	6.8%	GATCTG	11.9%
GGTACC	1.6%	AAGTTA	6.7%	TAGTAT	11.9%
AGGGGT	1.6%	AACTAG	6.7%	TACTAG	11.9%
CAGGGG	1.6%	GAATCT	6.7%	CGATCG	11.8%
GGGGTG	1.6%	GCTAGC	6.7%	GTCTAG	11.6%
GGGGTT	1.5%	TCTAGG	6.6%	GATCGG	11.6%
AGGGAC	1.5%	CTAAGT	6.5%	TGATCA	11.6%
CCCCTA	1.5%	AGATTC	6.5%	TGATCG	11.5%
CTAGGG	1.5%	CTAGCC	6.5%	TTGATC	11.5%
TGGGAA	1.5%	GTCTAC	6.4%	TAGTAG	11.3%
GGCCAG	1.5%	ACTGTA	6.3%	ACTAGA	11.2%
CCCCTG	1.5%	CTAATG	6.3%	TATCTA	11.2%
CGGCCA	1.5%	CTGTAG	6.3%	AGATCA	11.1%
GGGGTA	1.5%	GTCTAA	6.3%	ACTAGT	11.1%
ACCCCT	1.5%	CCTAGG	6.2%	CGATCT	11.0%
CCGGTC	1.4%	CAGACT	6.2%	AGATCG	11.0%

Error rates were calculated the same as in Table S4.

B Supplementary Figures

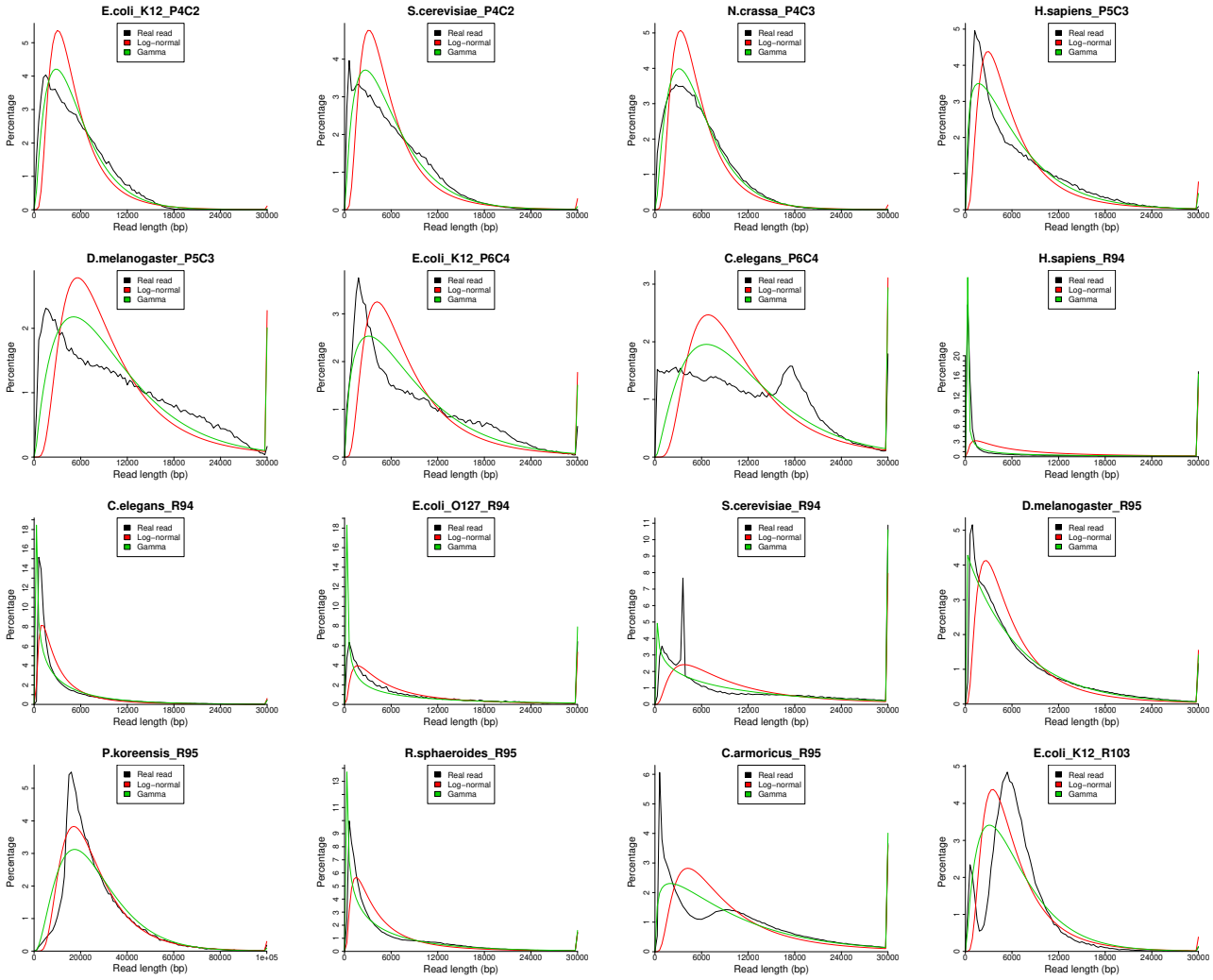


Figure S1: Read length distribution for each of the datasets in Tables S2 and S3. Dataset name is species (e.g., *E. coli*_K12) + chemistry (e.g., P4C2). Each graph shows distribution of real read length, as well as log-normal and gamma distributions with parameters derived from real reads.

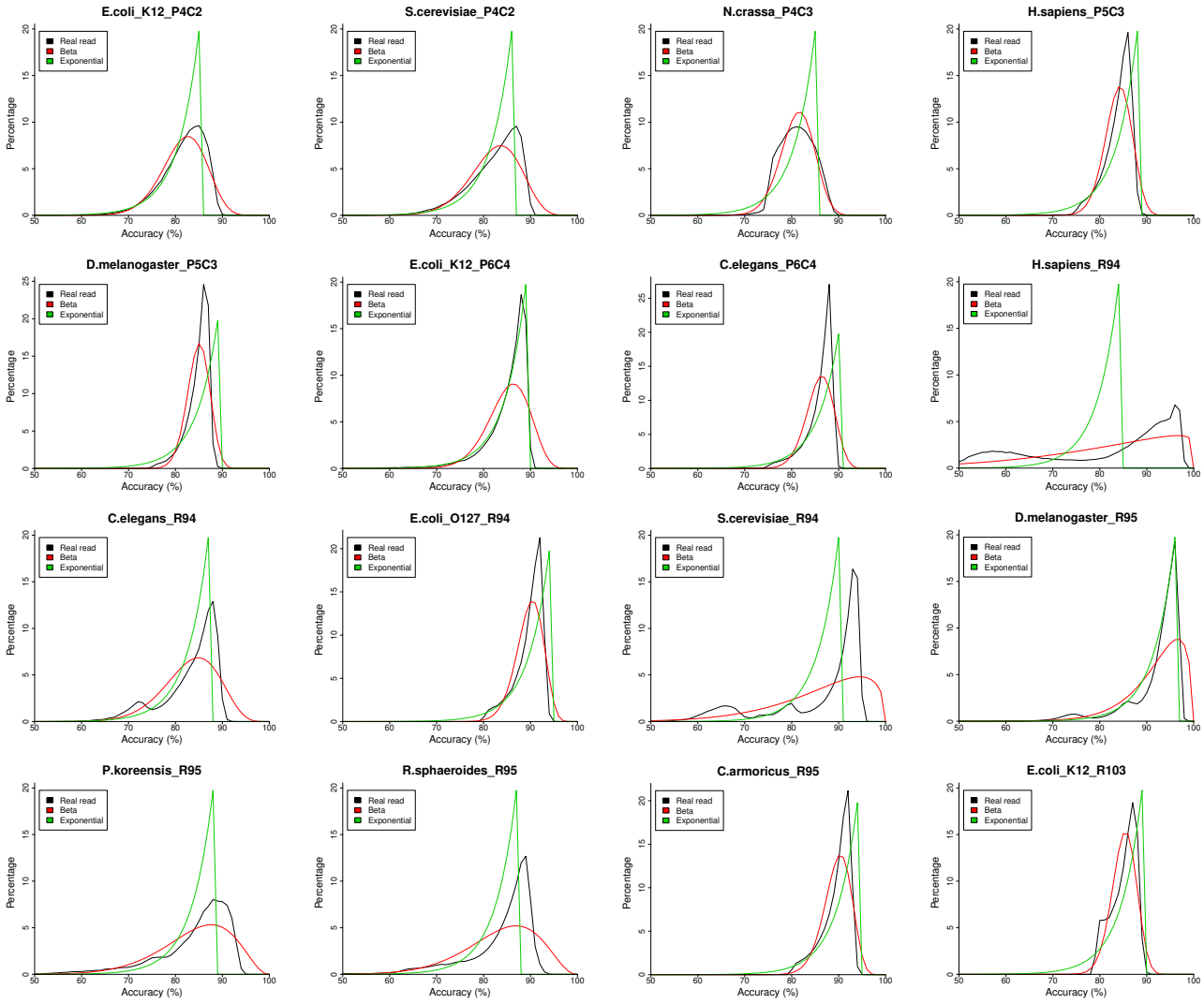


Figure S2: Read accuracy distribution for each of the datasets in Tables S2 and S3. Dataset name is species (e.g., *E. coli*_K12) + chemistry (e.g., P4C2). Each graph shows distribution of real read length, as well as beta and exponential distributions with parameters derived from real reads. Read accuracy was computed from quality scores.

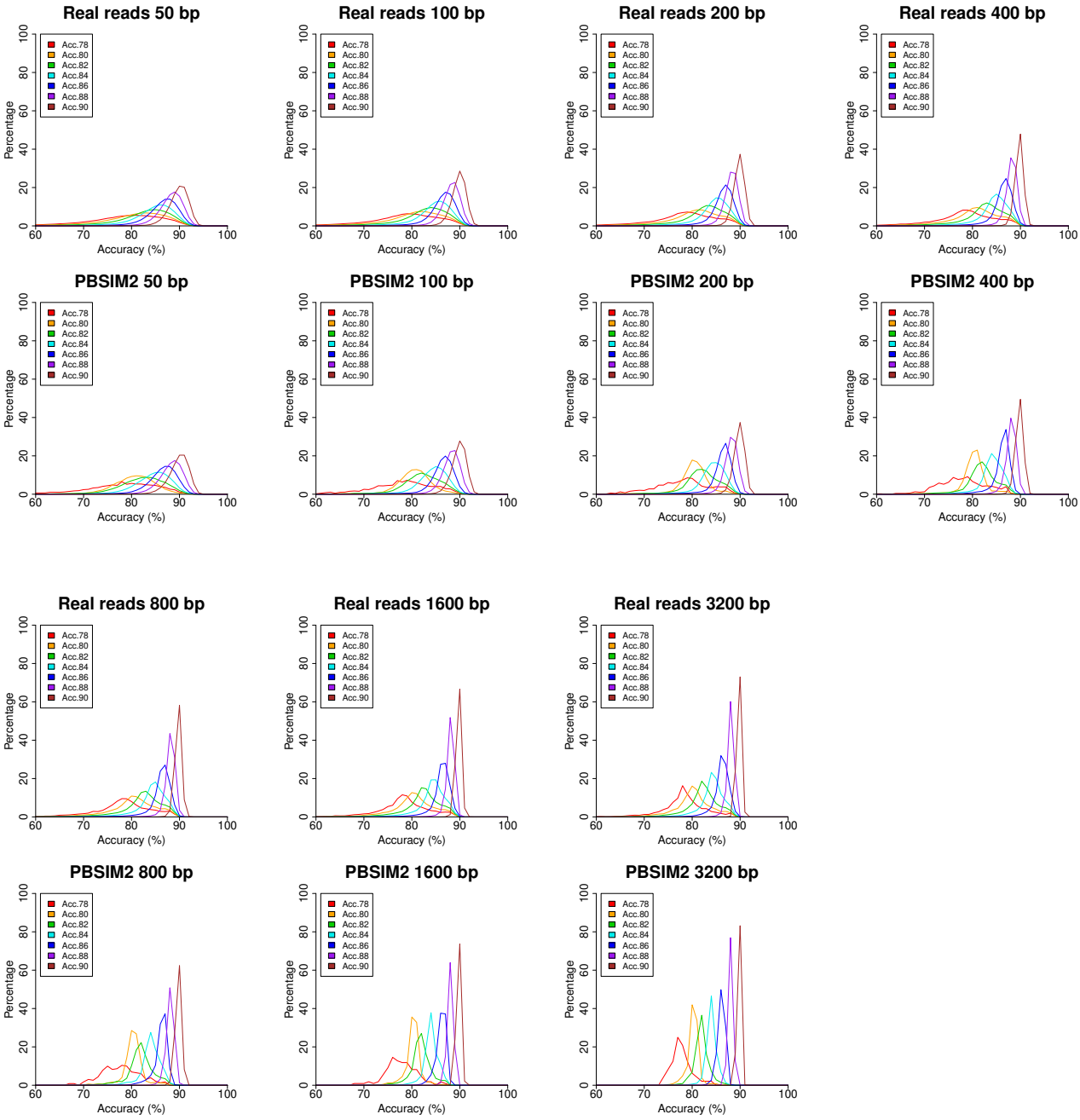


Figure S3: Non-uniformity of quality scores for real and simulated reads of PacBio P6–C4 for *C. elegans*. After grouping reads by their accuracy, the reads were segmented into fixed size (50, 100, 200, 400, 800, 1600, and 3200 bp) disjoint intervals, and the accuracy of each interval was computed from the quality scores. Each graph shows the distribution of averaged accuracy of each interval, where the colors of the plotted lines represent read groups (e.g., "Acc.78" is a read group whose accuracy is 77.5%-78.4%). Read groups (e.g., Acc.78) with insufficient data are not shown in the graphs.

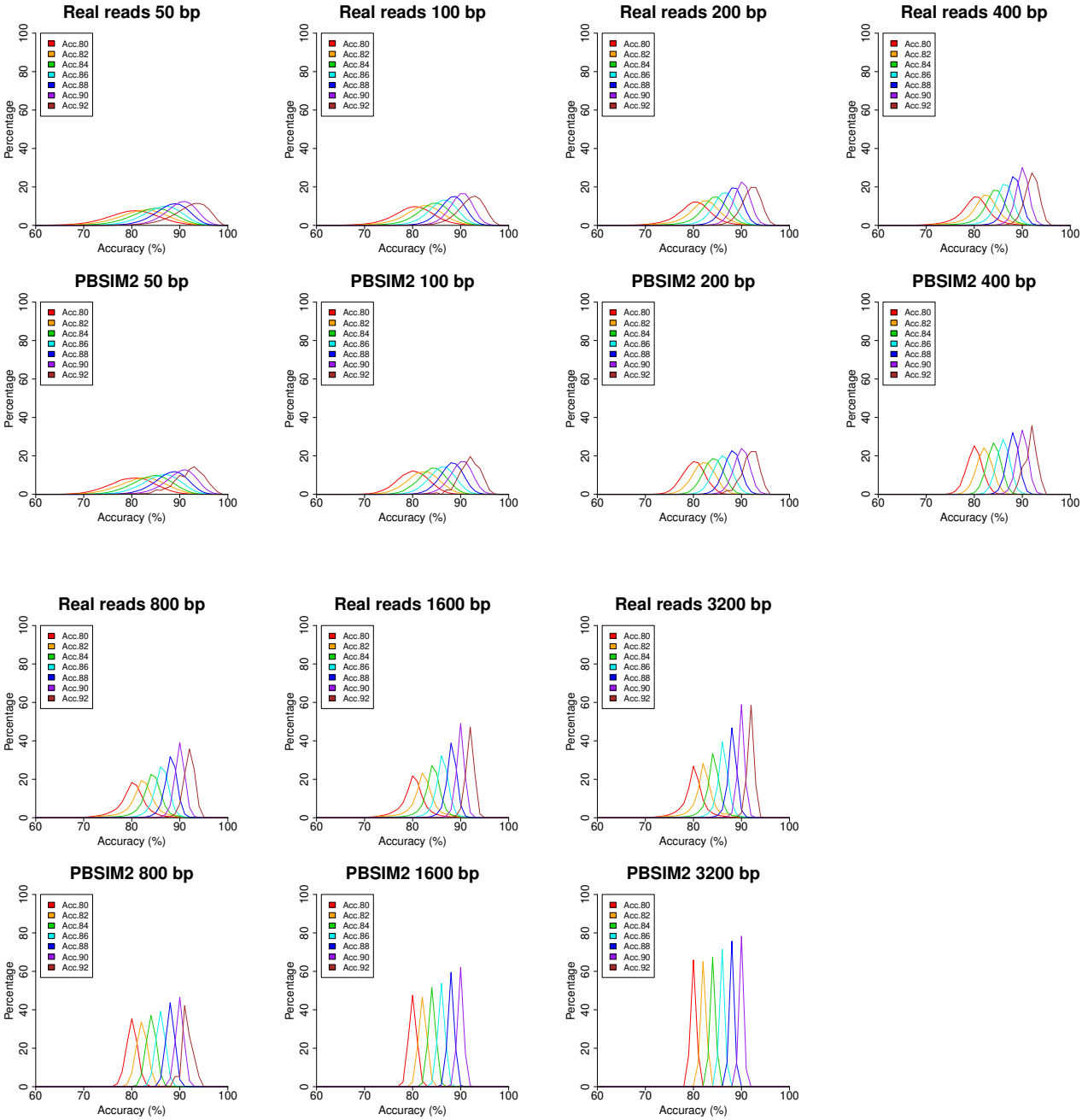
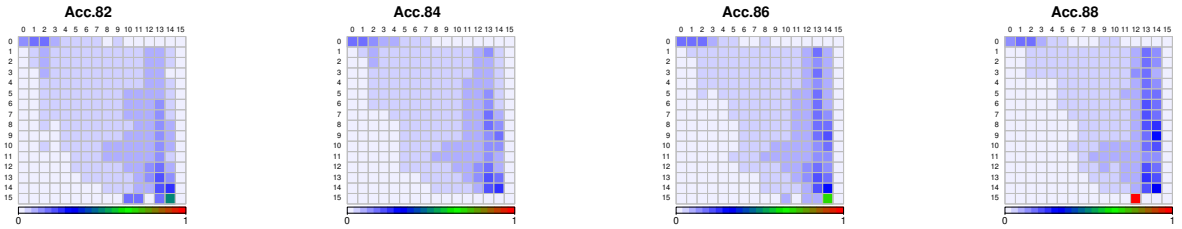


Figure S4: Non-uniformity of quality scores for real and simulated reads of Nanopore R9.5 for *R. sphaeroides*. Each graph shows the distribution of averaged accuracy of each interval in the same way as Figure S3.

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R.sphaeroides*

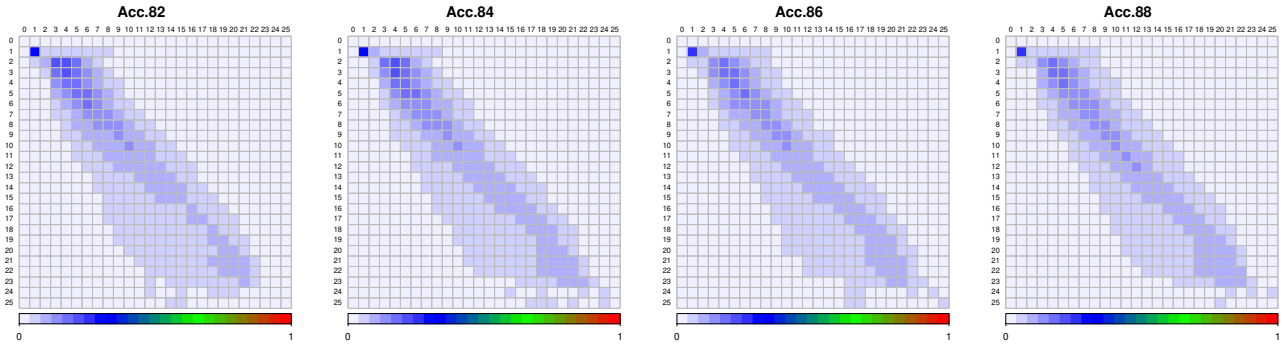
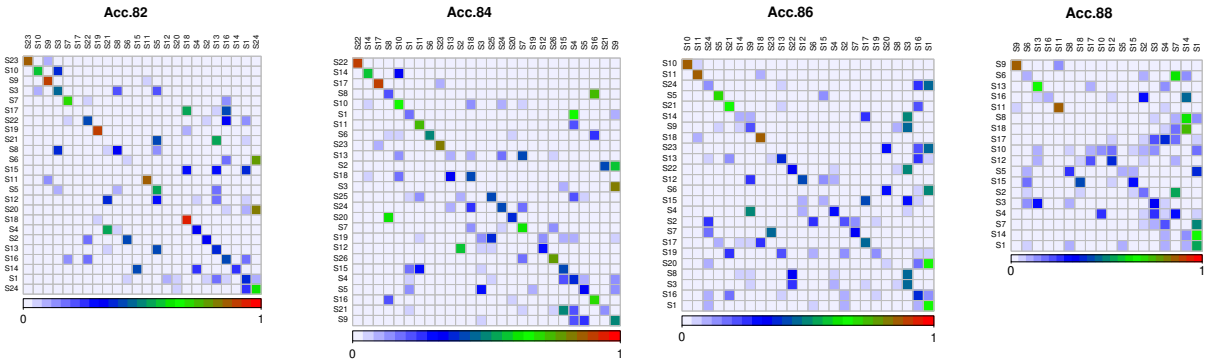


Figure S5: Transition probability matrices of quality scores of real reads. The vertical and horizontal axes are PHRED33 quality scores defined in terms of the estimated error probability (e.g., quality scores 4, 7, and 10 represent error probabilities of 40%, 20%, and 10%, respectively) [14]. Quality scores on the vertical axis transition to scores on the horizontal axis. The sum of transition probabilities on each quality score of the vertical axis is 100%. These are matrices of 'Acc.82'-'Acc.88' (e.g., 'Acc.84' refers to a read group with an accuracy of 83.5%-84.4%). In the Nanopore matrix, quality scores above 25 are not displayed.

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R.sphaeroides*

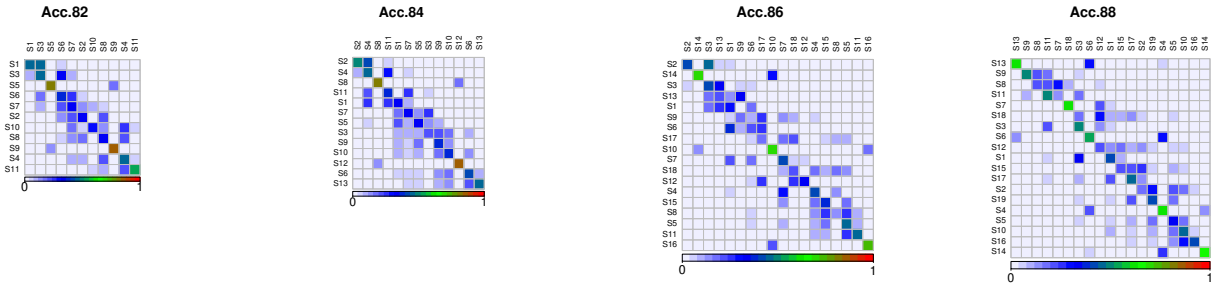
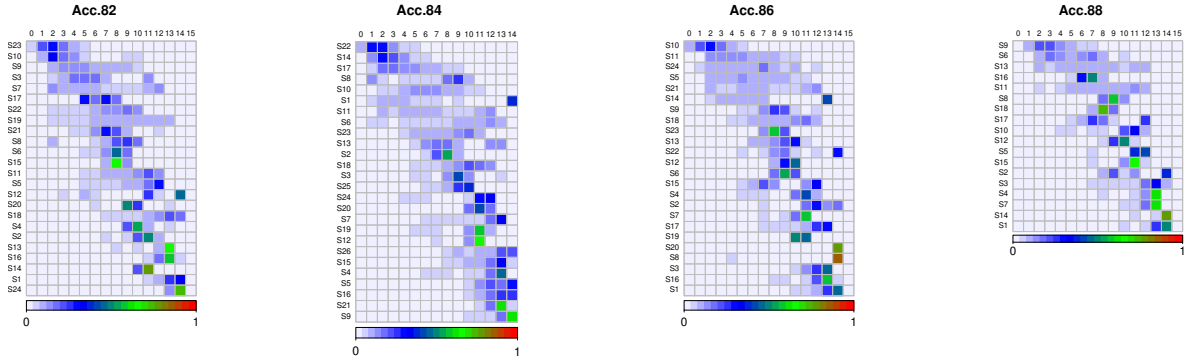


Figure S6: Transition probability matrices of states of FIC-HMM. The vertical and horizontal axes represent states of FIC-HMM, which are sorted in order of the increasing averaged quality score emitted by them. States on the vertical axis transition to states on the horizontal axis. The sum of transition probabilities on each state of the vertical axis is 100%. These are matrices of 'Acc.82'-'Acc.88' (e.g., 'Acc.84' refers to a read group with an accuracy of 83.5%-84.4%).

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R.sphaeroides*

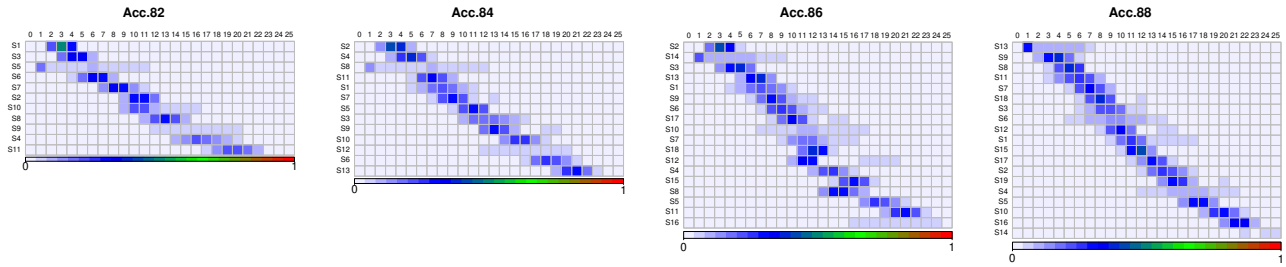
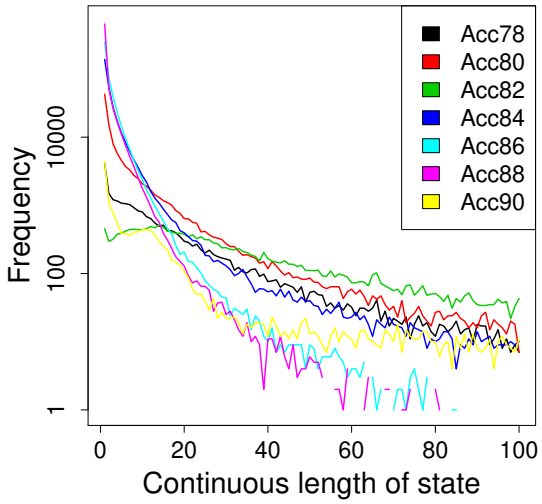


Figure S7: Emission probability matrices of states of FIC-HMM. The vertical axis represents states of FIC-HMM, which are sorted in the order of increasing averaged quality score emitted by them. The horizontal axis is PHRED33 quality score defined in terms of the estimated error probability (e.g., quality scores 4, 7, and 10 represent error probabilities of 40%, 20%, and 10%, respectively) [14]. States on the vertical axis emit quality scores on the horizontal axis. The sum of emission probabilities on each state of vertical axis is 100%. These are matrices of 'Acc.82'-'Acc.88' (e.g., 'Acc.84' refers to a read group with an accuracy of 83.5%-84.4%). In the matrix of Nanopore, quality scores above 25 are not displayed.

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R.sphaeroides*

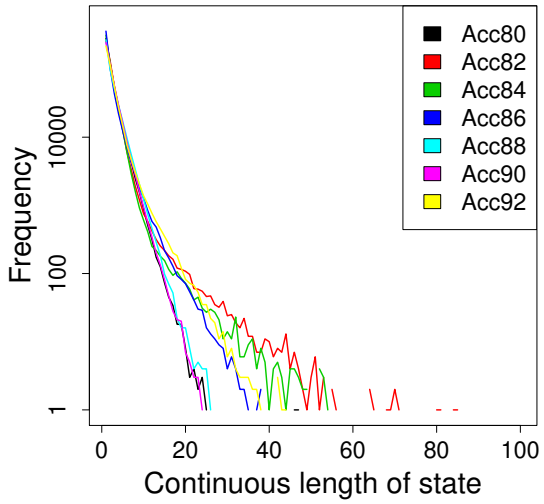


Figure S8: Distributions of continuous length of state. Training data for FIC-HMM was decoded into stats using the Viterbi algorithm. If the same state is lined up five times in a row, the continuous length is five. The vertical axis (log scale) is the frequency of each continuous length of state. The horizontal axis is the continuous length of state. Colors of plotted lines represent the read groups (e.g., 'Acc.78' refers to a read group with an accuracy of 77.5%-78.4%). Continuous lengths of state above 100 are not displayed.

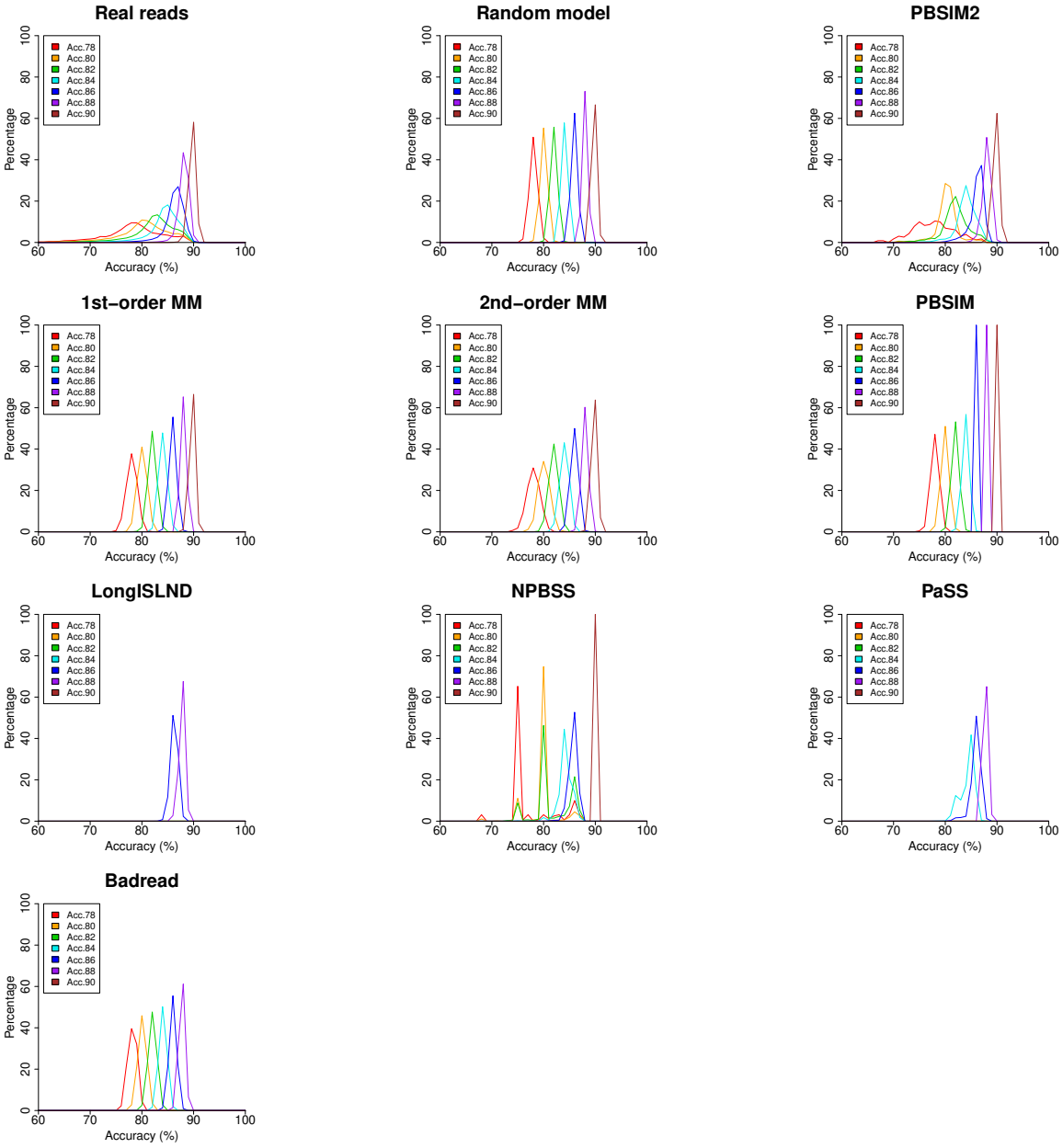


Figure S9: Non-uniformity of quality scores for real and simulated reads of PacBio P6–C4 for *C. elegans*. Each graph shows distributions of accuracy of 800 bp disjoint intervals in reads in the same way as Figure S3. PBSIM (i.e., previous version) has frequency tables of quality score for only Acc.60–85; thus, for Acc.86–90, 800 bp interval accuracy is constant.

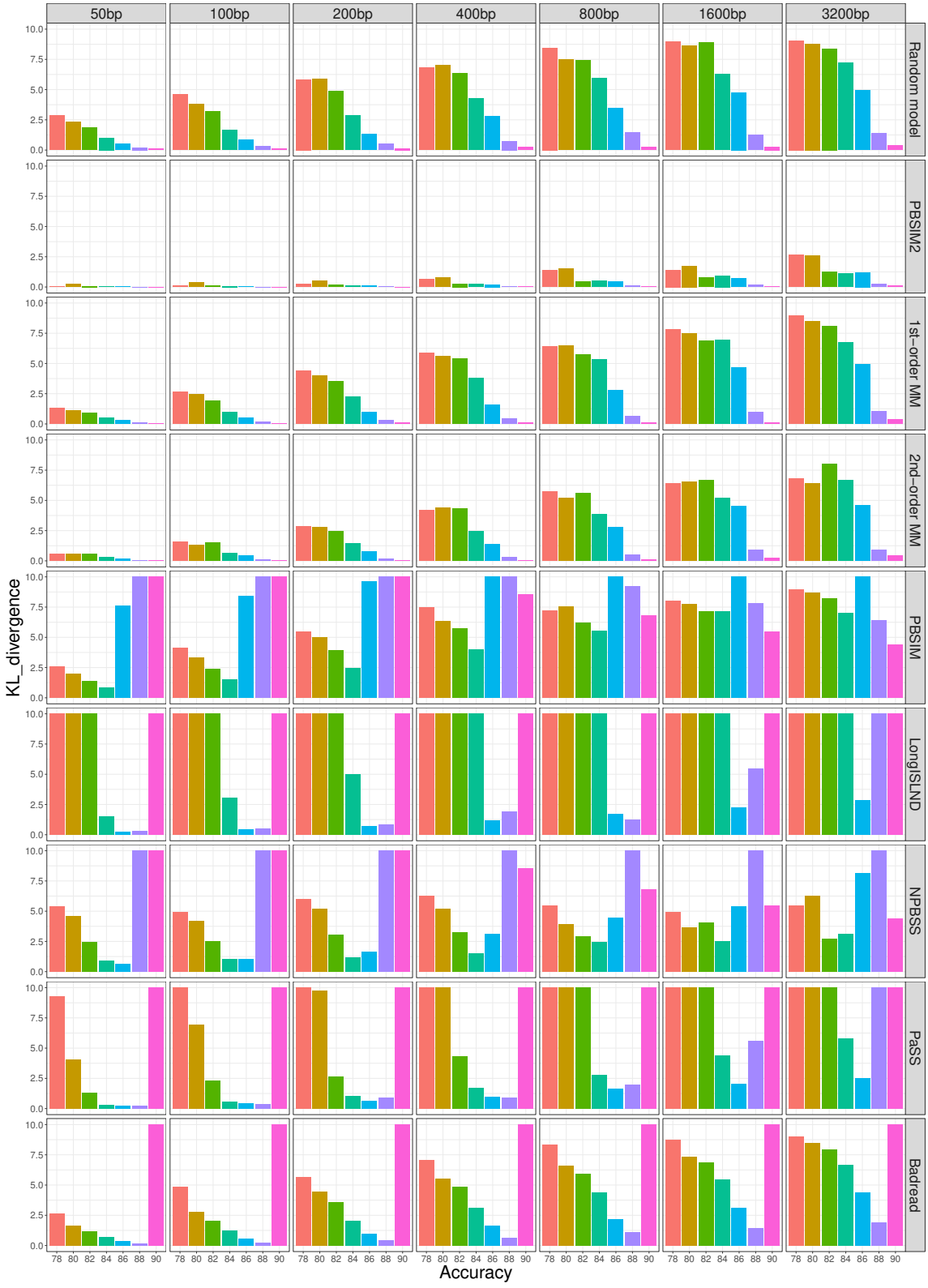


Figure S10: Kullback-Leibler (KL) divergence of distribution of accuracy of fixed size (50, 100, 200, 400, 800, 1600, and 3200 bp) intervals between real and simulated reads of PacBio P6–C4 for *C. elegans* in Figure S9. Upper-limit value of KL divergence is 10.

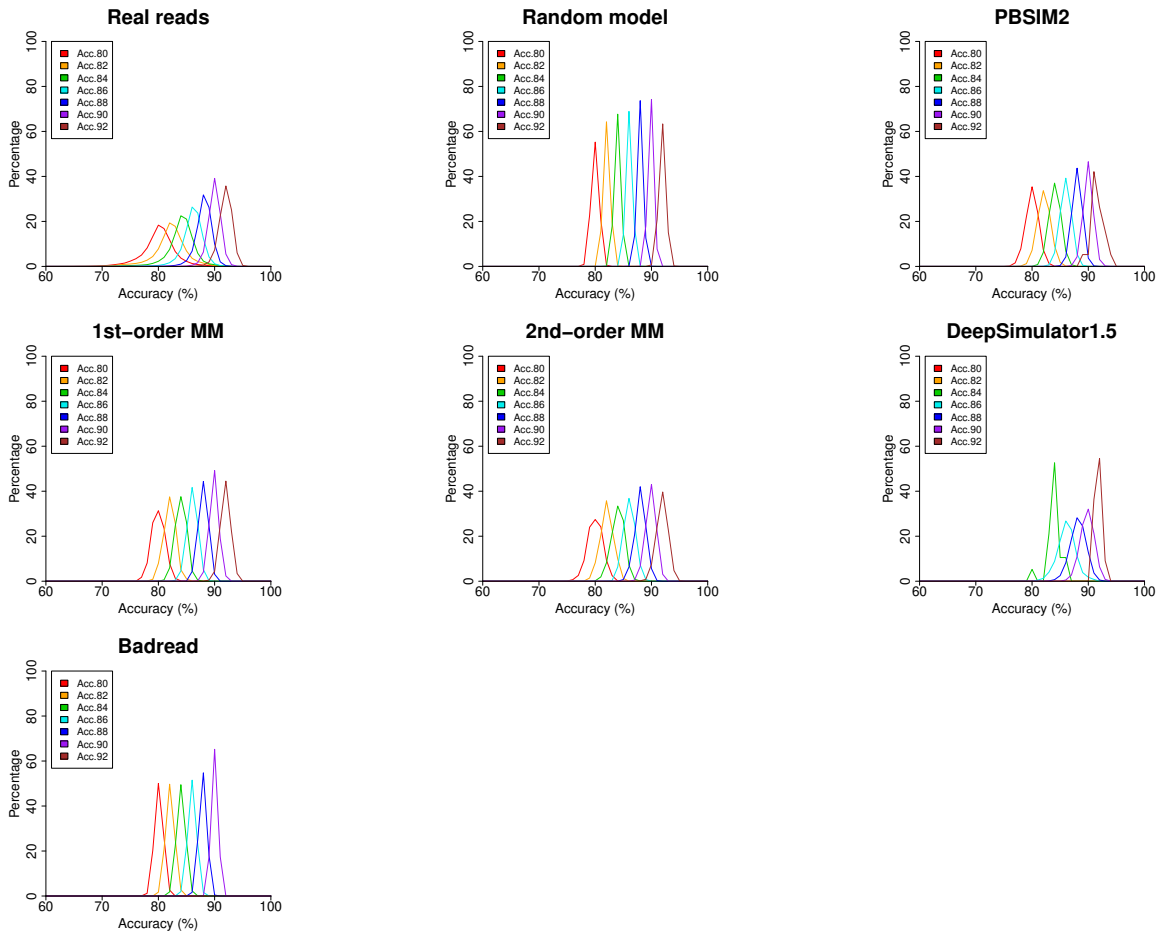


Figure S11: Non-uniformity of quality scores for real and simulated reads of Nanopore R9.5 for *R. sphaeroides*. Each graph shows distributions of accuracy of 800 bp intervals in reads in the same way as Figure S3. Read groups (e.g., Acc.78) with insufficient data are not shown in the graphs.

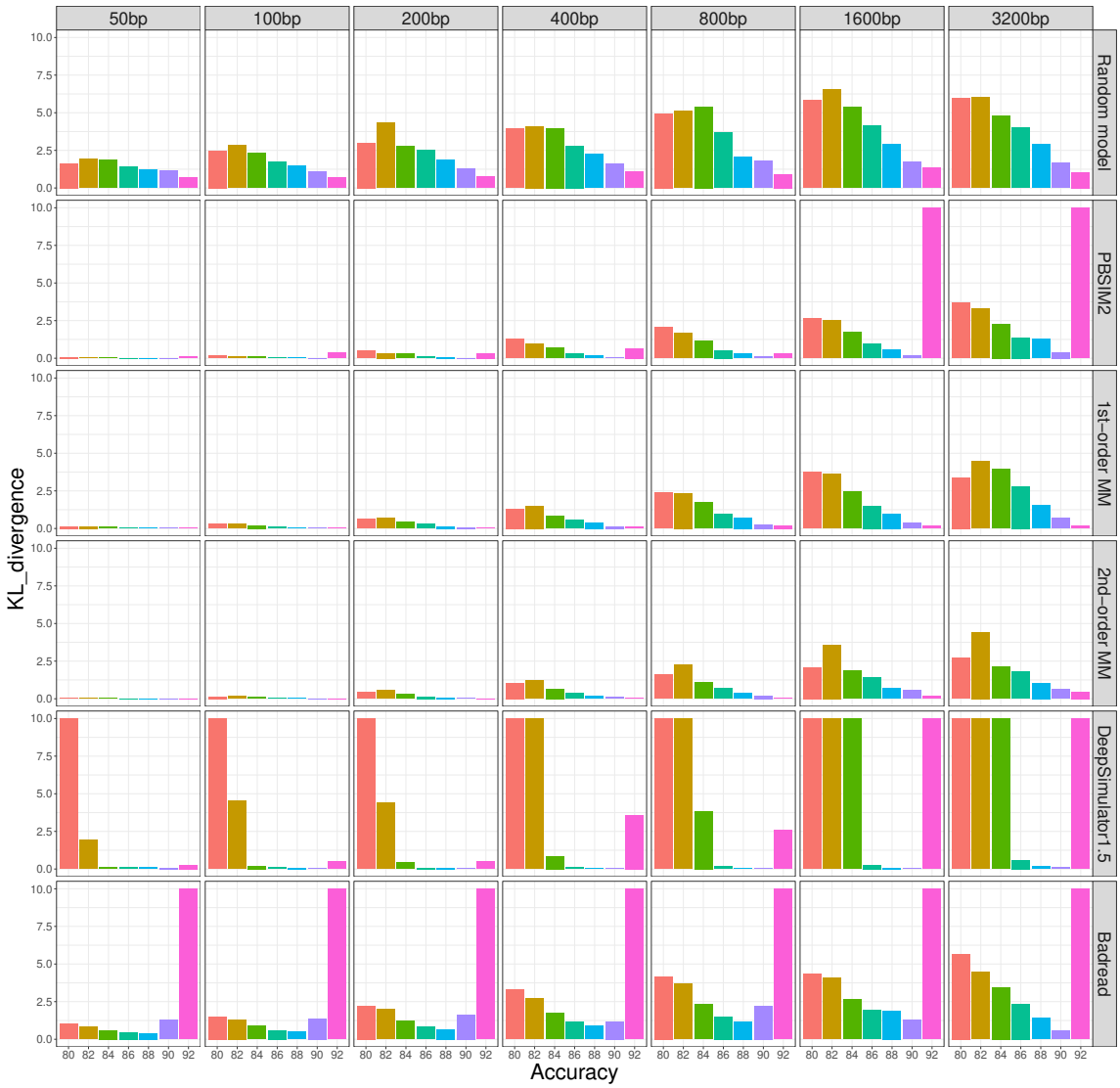


Figure S12: Kullback-Leibler (KL) divergence of distribution of averaged accuracy of fixed size (50, 100, 200, 400, 800, 1600, and 3200 bp) intervals between real and simulated reads of Nanopore R9.5 for *R. sphaeroides* in Figure S11. Upper-limit value of KL divergence is 10.

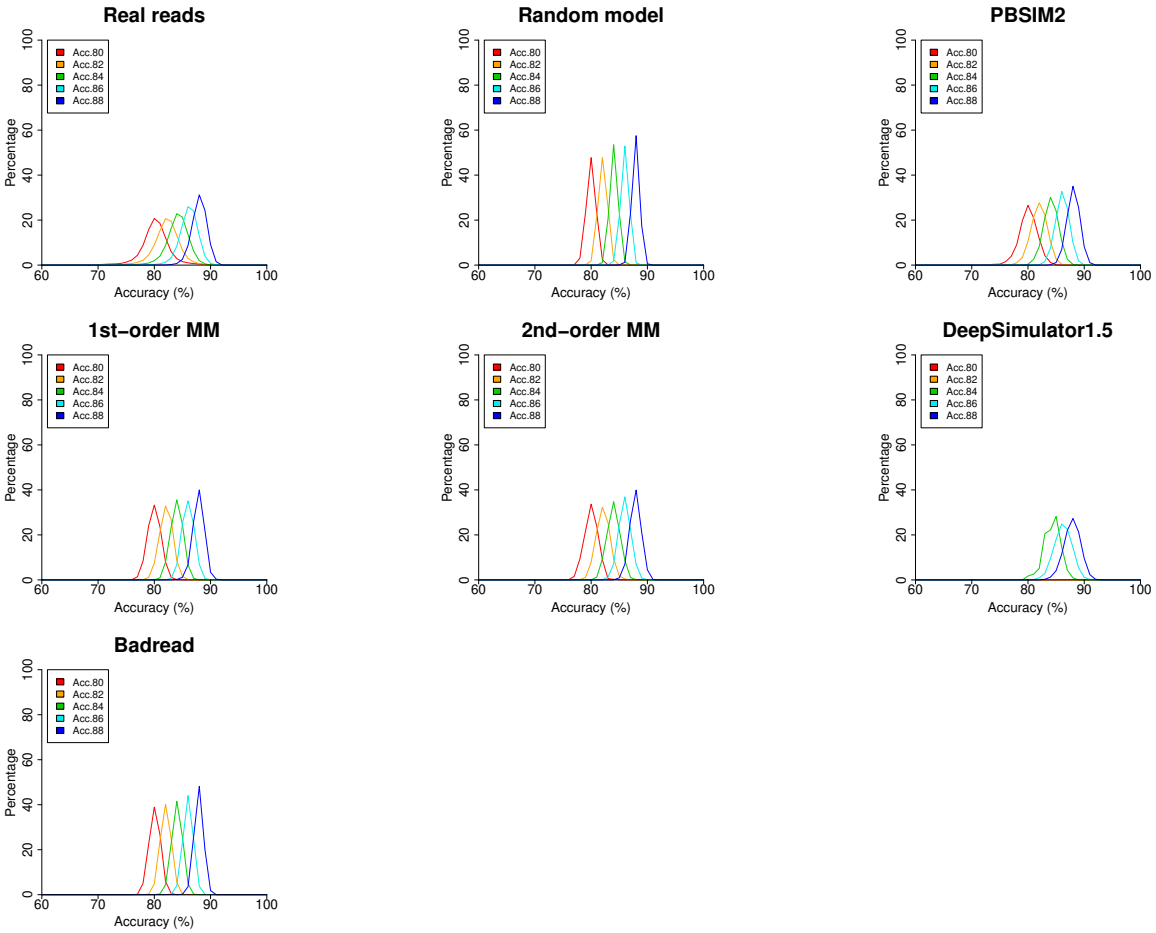
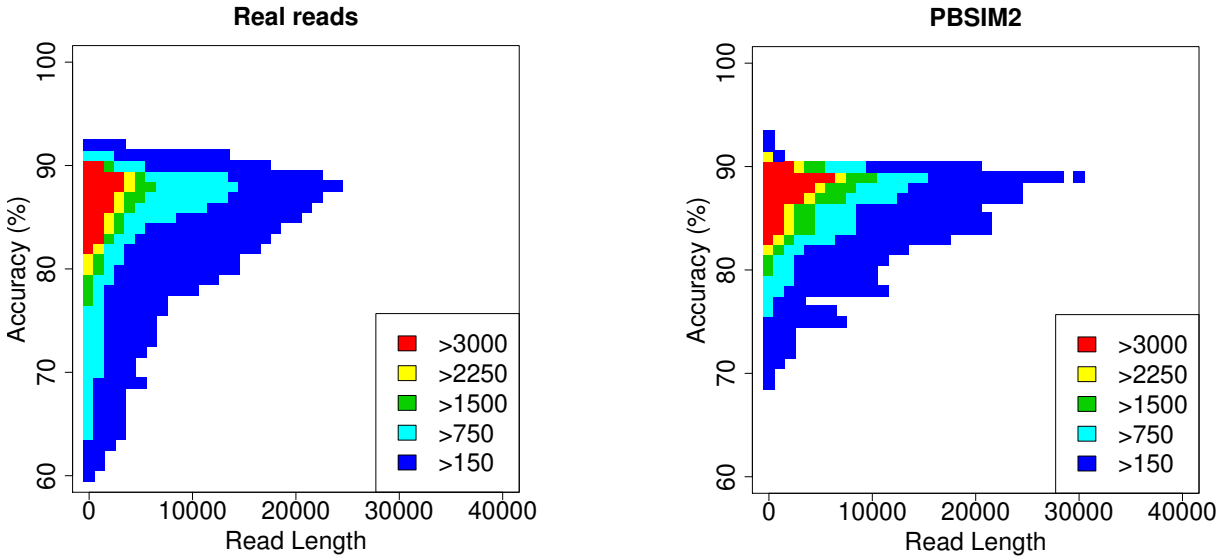


Figure S13: Non-uniformity of quality scores for real and simulated reads of Nanopore R10.3 for *E-coli* K12. Each graph shows distributions of accuracy of 800 bp intervals in read sequences, in the same way as Figure S9.

Figure S14: Kullback-Leibler (KL) divergence of distribution of averaged accuracy of fixed size (50, 100, 200, 400, 800, 1600, and 3200 bp) intervals between real and simulated reads of Nanopore R10.3 for *E. coli* K12 in Figure S13. Upper-limit value of KL divergence is 10.

(a) Nanopore R9.5 for *R. sphaeroides*



(b) Nanopore R10.3 for *E. coli* K12

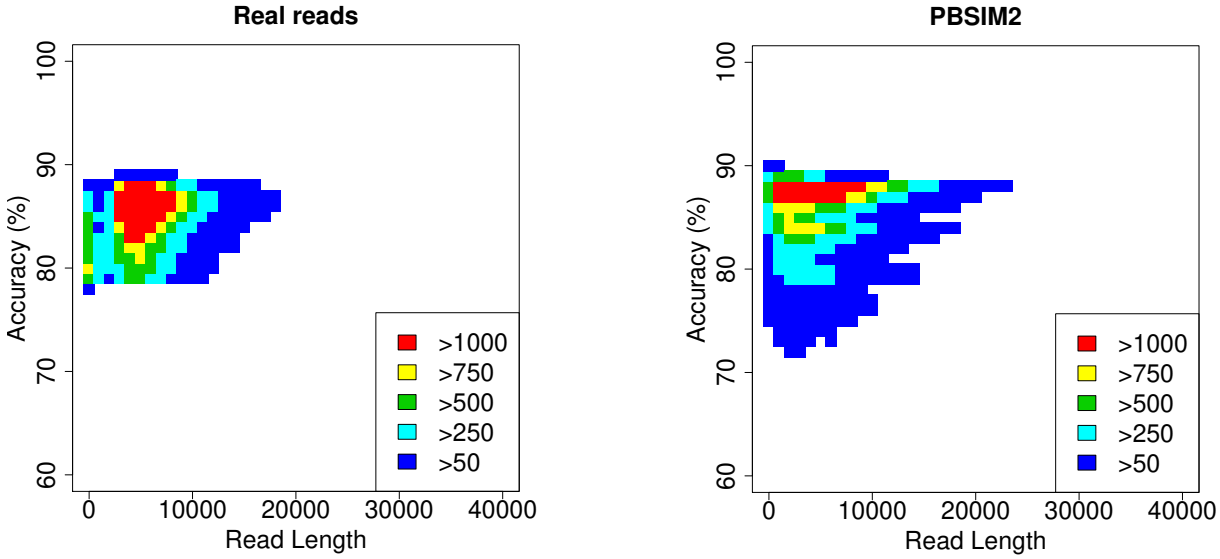
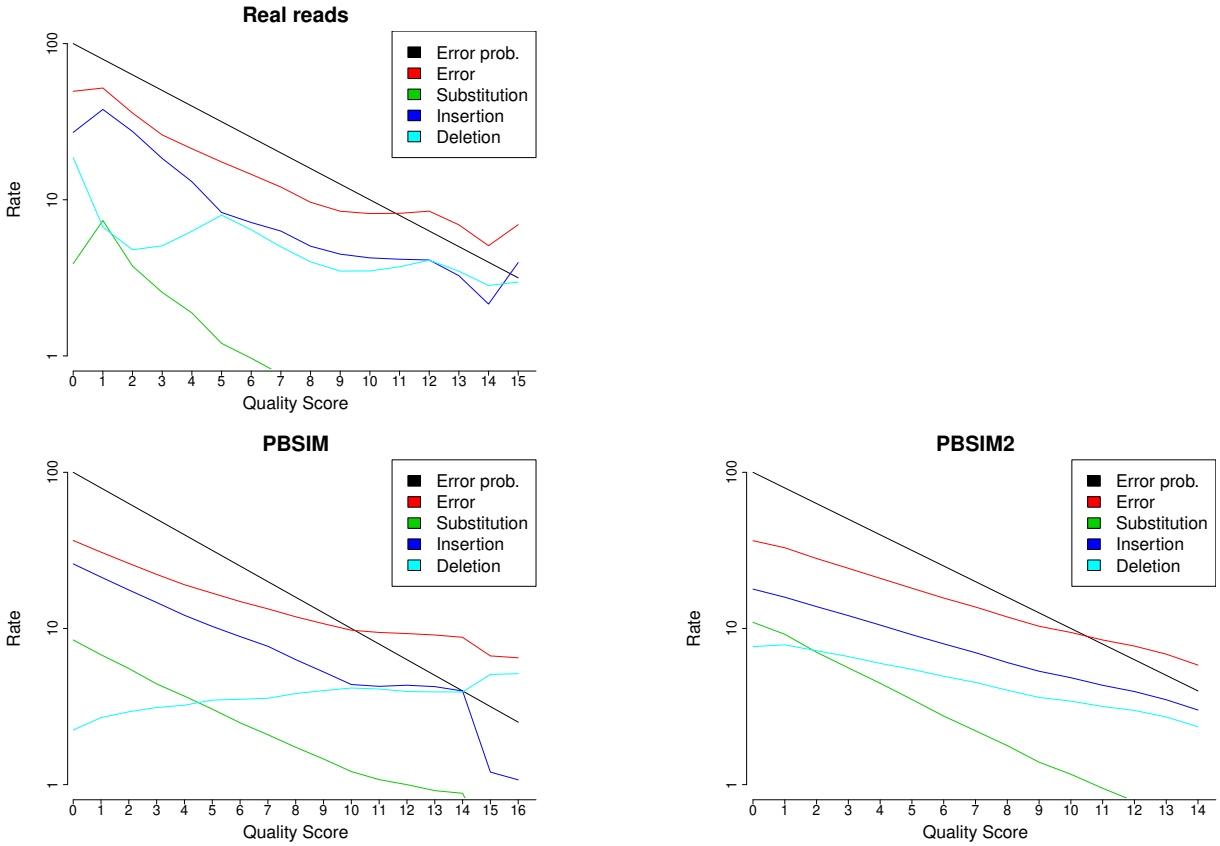


Figure S15: Correlation between read length and accuracy for each read. The accuracy of each read was calculated from the quality scores. PBSIM2 simulated reads with the same parameters (e.g., mean and standard deviation of read length and accuracy) as real reads. Colors indicate the different frequencies of each cell.

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R. sphaeroides*

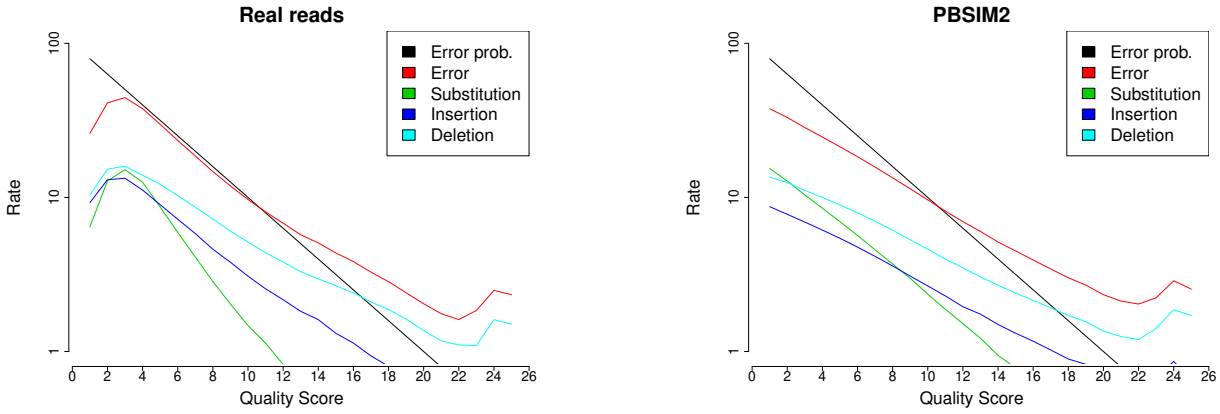


Figure S16: Relationship between the quality score and error rate for real reads and simulated reads. Each graph shows averaged error rate for each quality score. The horizontal axis is PHRED33, quality score defined in terms of the estimated error probability (e.g., quality scores 4, 7, and 10 represent error probabilities of 40, 20, and 10%, respectively) [14]. "Error" is the sum of the substitution, insertion, and deletion rates. Error rates were obtained from the alignment of the real and simulated reads to the reference sequences.

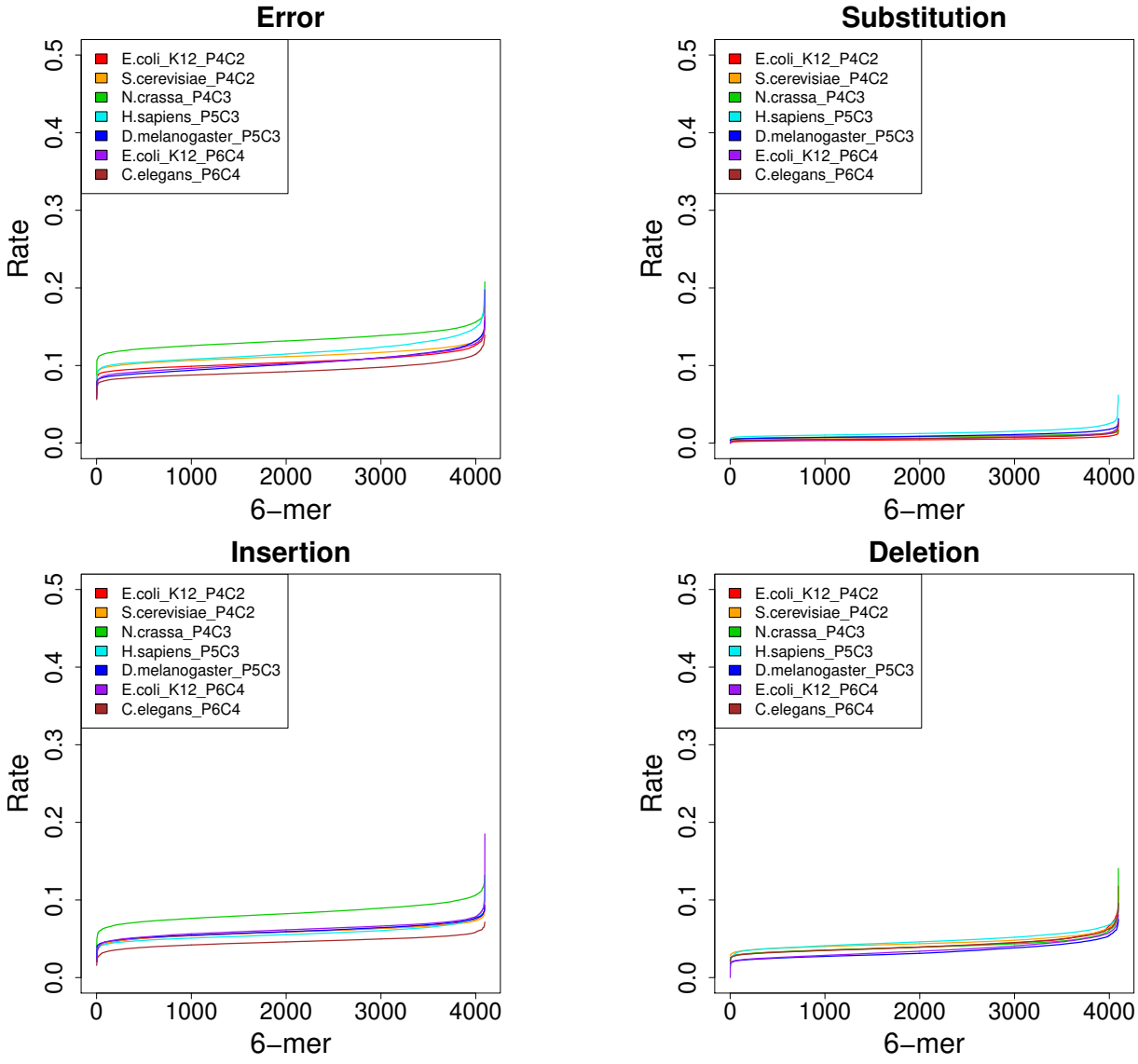


Figure S17: 6-mer error bias of PacBio. Several types of error rate were calculated for each 6-mer from the alignment of the real reads to the reference sequences. Colors of plotted lines represent each dataset in Tables S2. Dataset name is species (e.g., *E.coli*_K12) + chemistry (e.g., P4C2). The vertical axis represents the error rate, while the horizontal axis represents the k-mer index sorted by error rate. "Error" is the sum of the substitution, insertion, and deletion rates.

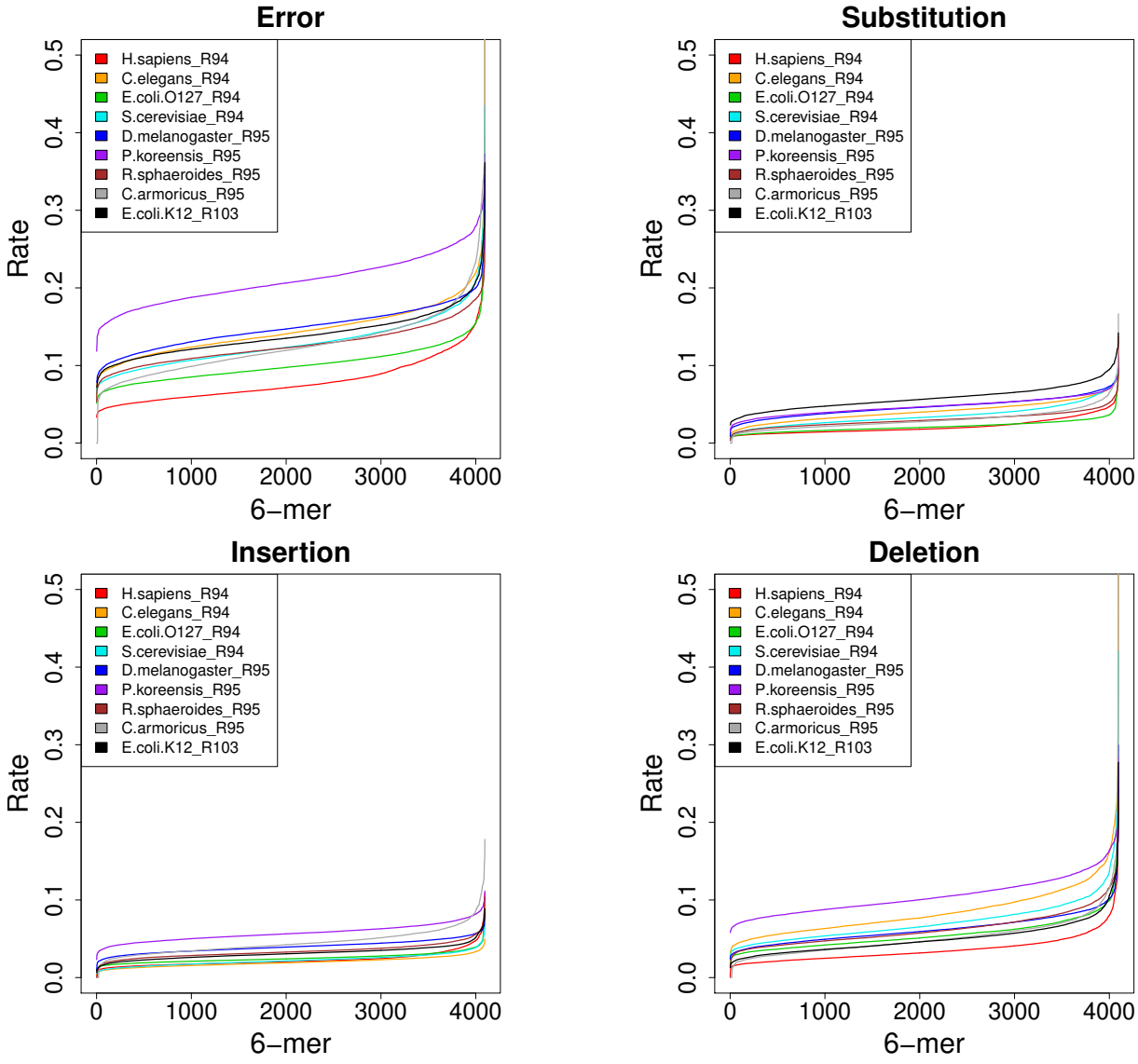
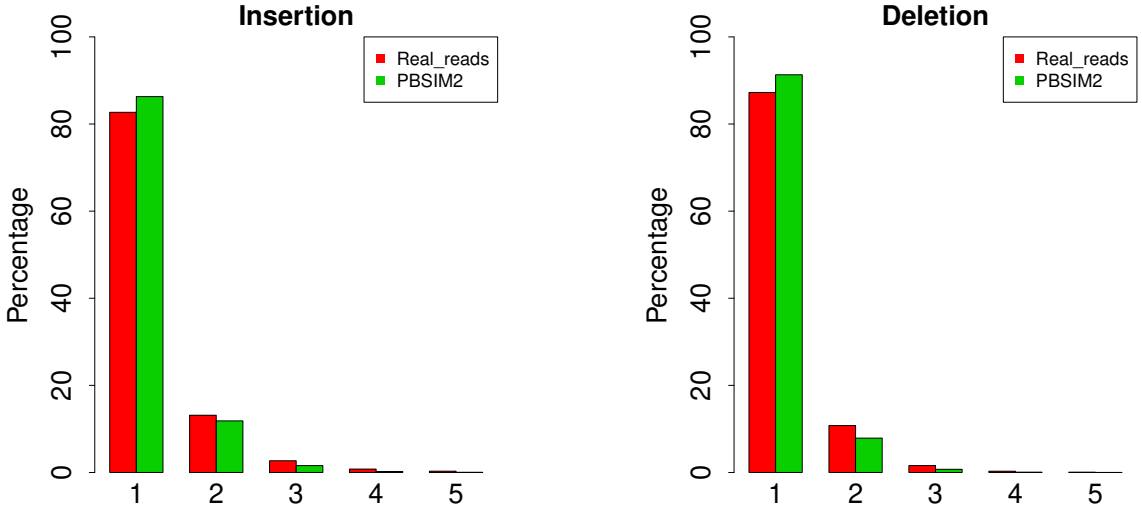
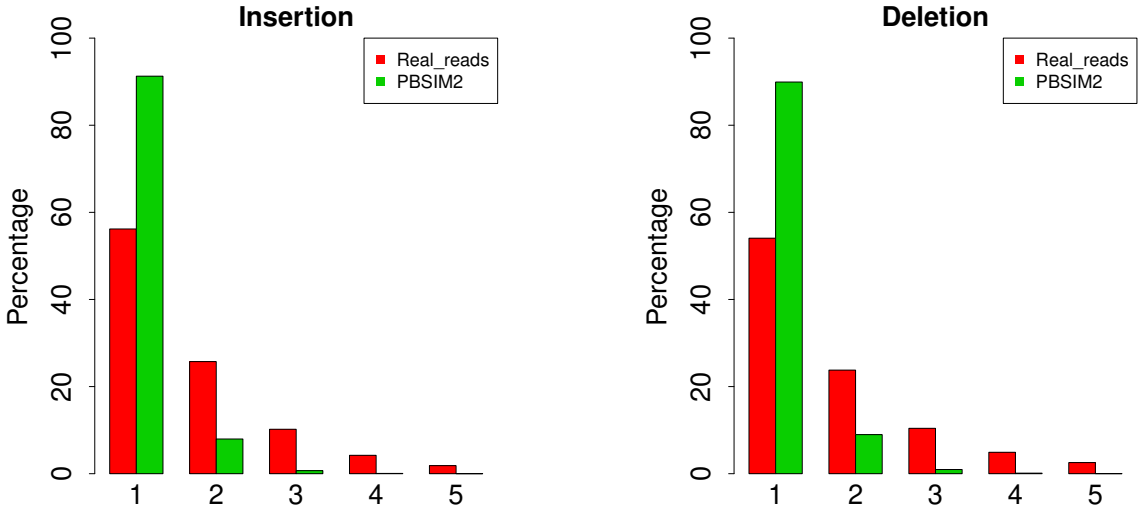


Figure S18: 6-mer error bias of Nanopore. Several types of error rates were calculated for each 6-mer from the alignment of the real reads to the reference sequences. Colors of plotted lines represent each dataset in Tables S3. Dataset name is species (e.g., *H.sapiens*) + chemistry (e.g., R94). The vertical axis represents the error rate, while the horizontal axis represents the k-mer index sorted by error rate. "Error" is the sum of the substitution, insertion, and deletion rates.

(a) PacBio P6–C4 for *C.elegans*



(b) Nanopore R9.5 for *R.sphaeroides*



(c) Nanopore R10.3 for *E-coli* K12

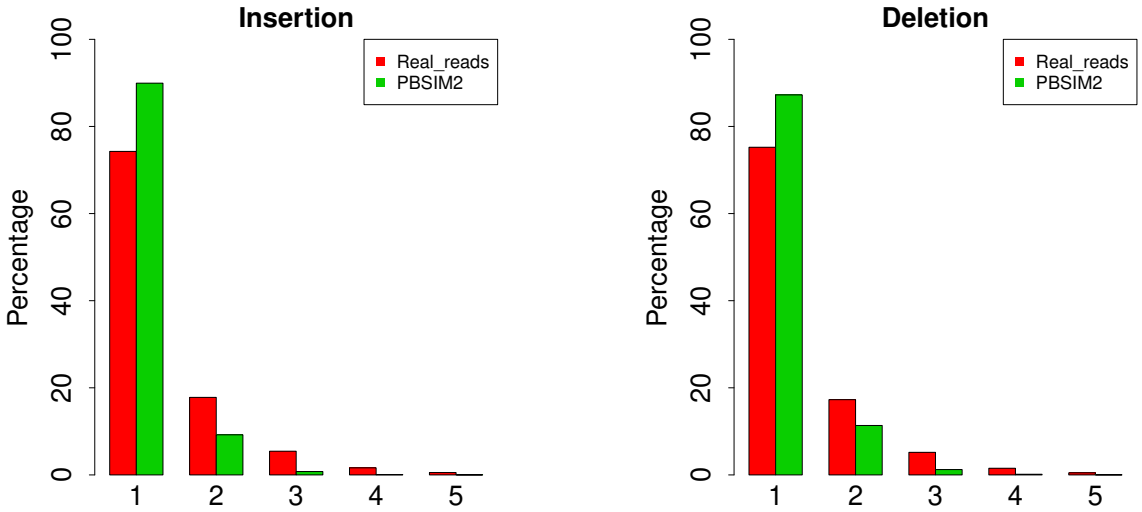


Figure S19: Distributions of insertion and deletion length (indel) for real reads and PBSIM2. The vertical axis represents the percentage, while the horizontal axis represents the indel length. Frequencies of indel length were obtained from the alignment of the real and simulated reads to the reference sequences.

References

[1] Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. Pbsim: Pacbio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2013.

[2] Hayan Lee, James Gurtowski, Shinjae Yoo, Shoshana Marcus, W Richard McCombie, and Michael Schatz. Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*, page 006395, 2014.

[3] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. Simlord: simulation of long read data. *Bioinformatics*, 32(17):2704–2706, 2016.

[4] Ethan Alexander Garcia Baker, Sara Goodwin, W Richard McCombie, and Olivia Mendivil Ramos. Silico: a simulator of long read sequencing in pacbio and oxford nanopore. *BioRxiv*, page 076901, 2016.

[5] Bayo Lau, Marghoob Mohiyuddin, John C Mu, Li Tai Fang, Narges Bani Asadi, Carolina Dallett, and Hugo YK Lam. Longislnd: in silico sequencing of lengthy and noisy datatypes. *Bioinformatics*, 32(24):3829–3832, 2016.

[6] Chen Yang, Justin Chu, René L Warren, and Inanç Birol. Nanosim: nanopore sequence read simulator based on statistical characterization. *GigaScience*, 6(4):gix010, 2017.

[7] Philippe C Faucon, Parithi Balachandran, and Sharon Crook. Snaresim: Synthetic nanopore read simulator. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 338–344. IEEE, 2017.

[8] Ze-Gang Wei and Shao-Wu Zhang. Npbss: a new pacbio sequencing simulator for generating the continuous long reads with an empirical model. *BMC bioinformatics*, 19(1):177, 2018.

[9] Yu Li, Renmin Han, Chongwei Bi, Mo Li, Sheng Wang, and Xin Gao. Deepsimulator: a deep simulator for nanopore sequencing. *Bioinformatics*, 34(17):2899–2908, 2018.

[10] Yu Li, Sheng Wang, Chongwei Bi, Zhaowen Qiu, Mo Li, and Xin Gao. Deepsimulator1. 5: a more powerful, quicker and lighter simulator for nanopore sequencing. *Bioinformatics*, 2020.

[11] Christian Rohrandt, Nadine Kraft, Pay Gießelmann, Björn Brändl, Bernhard M Schuldt, Ulrich Jetzek, and Franz-Josef Müller. Nanopore simulation–a raw data simulator for nanopore sequencing. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–8. IEEE, 2018.

[12] Wenmin Zhang, Ben Jia, and Chaochun Wei. Pass: a sequencing simulator for pacbio sequencing. *BMC bioinformatics*, 20(1):1–7, 2019.

[13] Ryan Wick. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, 4(36):1316, 2019.

[14] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.