counterr: Characterization of Context Dependent MinION Sequencing Errors

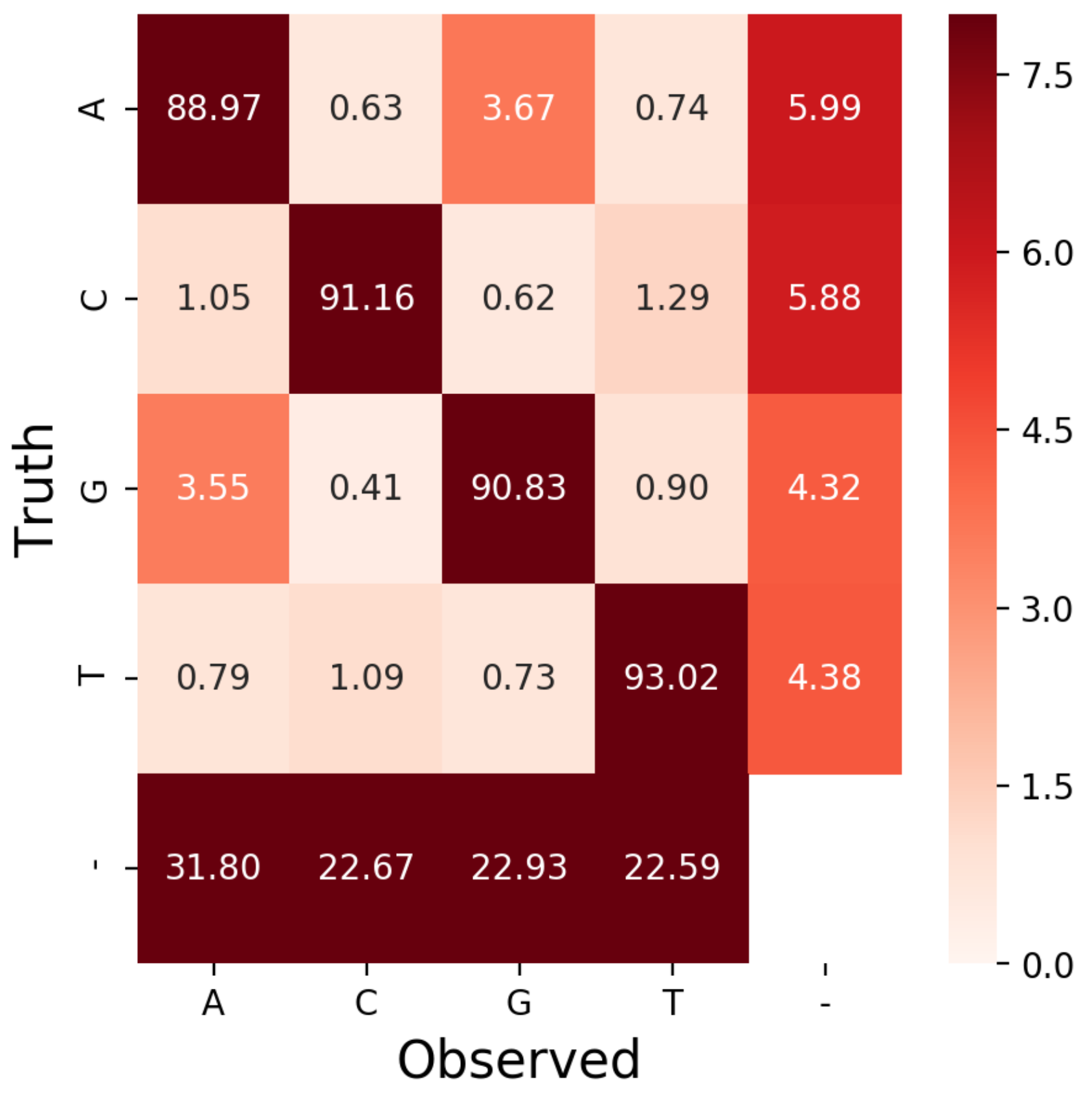
Jae Hyeon Lee*, Ian Herriott*, Elliott Bartsch*, Miriam H. Huntley*
* Day Zero Diagnostics, Inc.

Introduction

Errors in sequencing data depend on various factors such as library preparation, flowcell chemistry, and basecalling software. Moreover, sequencing errors can be context dependent, i.e. non-uniformly distributed. To better understand these errors, we developed ***counterr***, a lightweight command line tool that characterizes the context dependent error profile of sequencing reads using their alignment to a reference genome. We used *counterr* to characterize the error distributions in both amplified and native microbial ONT MinION sequencing data. Our results confirm a widely held belief that errors in MinION data strongly depend on sequence context. We hope that this improved error characterization can be useful for read error correction.

Data: An amplified DNA dataset was created from 9 bacterial isolates (comprising *E. coli*, *P. aeruginosa*, *S. aureus*, and *K. pnuemoniae*) individually prepped with the PCR-based SQK-RPB004 and sequenced on a MinION (R9.4.1). Reads were basecalled with Albacore v2.3 and aligned with Minimap2 to matched short-read assemblies (generated with paired-end Illumina NextSeq data, assembled with SPAdes v3.8.1). Native genomic DNA was obtained from [1] (*K. pneumoniae* sequenced with MinION R9.4.1 and basecalled with Albacore v1.1.2).

Error Matrix



Truth \ Observed	A	C	G	T	-
A	88.97	0.63	3.67	0.74	5.99
C	1.05	91.16	0.62	1.29	5.88
G	3.55	0.41	90.83	0.90	4.32
T	0.79	1.09	0.73	93.02	4.38
-	31.80	22.67	22.93	22.59	

We show the percent of cases where a base will undergo a substitution or deletion (context independent). “A” and “G” are the most frequently confused pair of bases. The dominant mode of error is deletion for all bases.

Context Dependent Substitution/Deletion

We examined the context around errors to see if certain k-mers are more (or less) subject to substitution or deletion errors. We show the percent of time the central letter of the 5-mer (indicated in bold) is incorrectly called, along with observed errors for each base. In amplified DNA k-mers can have as little as 2% errors and as high as 30% errors. In the native genomic DNA, errors go from <12.5% to >50% for CC**T**GG and CC**A**GG due to *Dcm* methylation [2].

Amplified Microbial DNA

	Context	%Error	Counts	%A	%C	%G	%T	%-
most errors	GG**A**GT	30.31	2104958	69.69	0.40	3.48	0.67	25.76
	TC**C**AA	27.97	1589000	3.63	72.03	0.98	1.06	22.31
	GG**A**GC	27.79	3209245	72.21	0.70	3.47	0.48	23.15
	GC**T**TG	23.84	3185883	3.13	1.96	1.18	76.16	17.57
	GC**A**GT	22.91	4135843	77.09	0.89	15.88	0.81	5.33
fewest errors	AG**T**TC	2.43	3450757	0.50	0.30	0.52	97.57	1.12
	GA**T**CC	2.38	4423638	0.29	0.29	0.34	97.62	1.46
	AA**T**CA	2.33	5420979	0.18	0.59	0.28	97.67	1.27
	GA**T**CA	2.07	5454337	0.17	0.55	0.14	97.93	1.22
	GA**T**CG	2.05	5934583	0.18	0.44	0.13	97.95	1.30

Native Microbial DNA

	Context	%Error	Counts	%A	%C	%G	%T	%-
most errors	CC**A**GG	62.99	12192022	37.01	3.91	29.85	1.05	28.17
	CC**T**GG	50.30	12198776	4.51	25.87	1.44	49.70	18.48
	GG**A**GC	32.70	6056412	67.30	0.81	4.92	0.75	26.22
	CT**A**GG	32.68	237965	67.32	1.61	16.81	1.68	12.58
	CA**A**CT	32.54	3400929	67.46	0.81	25.05	1.18	5.52
fewest errors	AA**T**CG	1.76	6275281	0.16	0.40	0.18	98.24	1.02
	GG**A**TC	1.69	9237061	98.31	0.05	0.32	0.27	1.05
	AA**T**CA	1.62	7098031	0.15	0.43	0.21	98.38	0.84
	GA**T**CG	1.26	11948130	0.10	0.23	0.08	98.74	0.84
	GA**T**CA	1.25	9506225	0.09	0.26	0.09	98.75	0.82

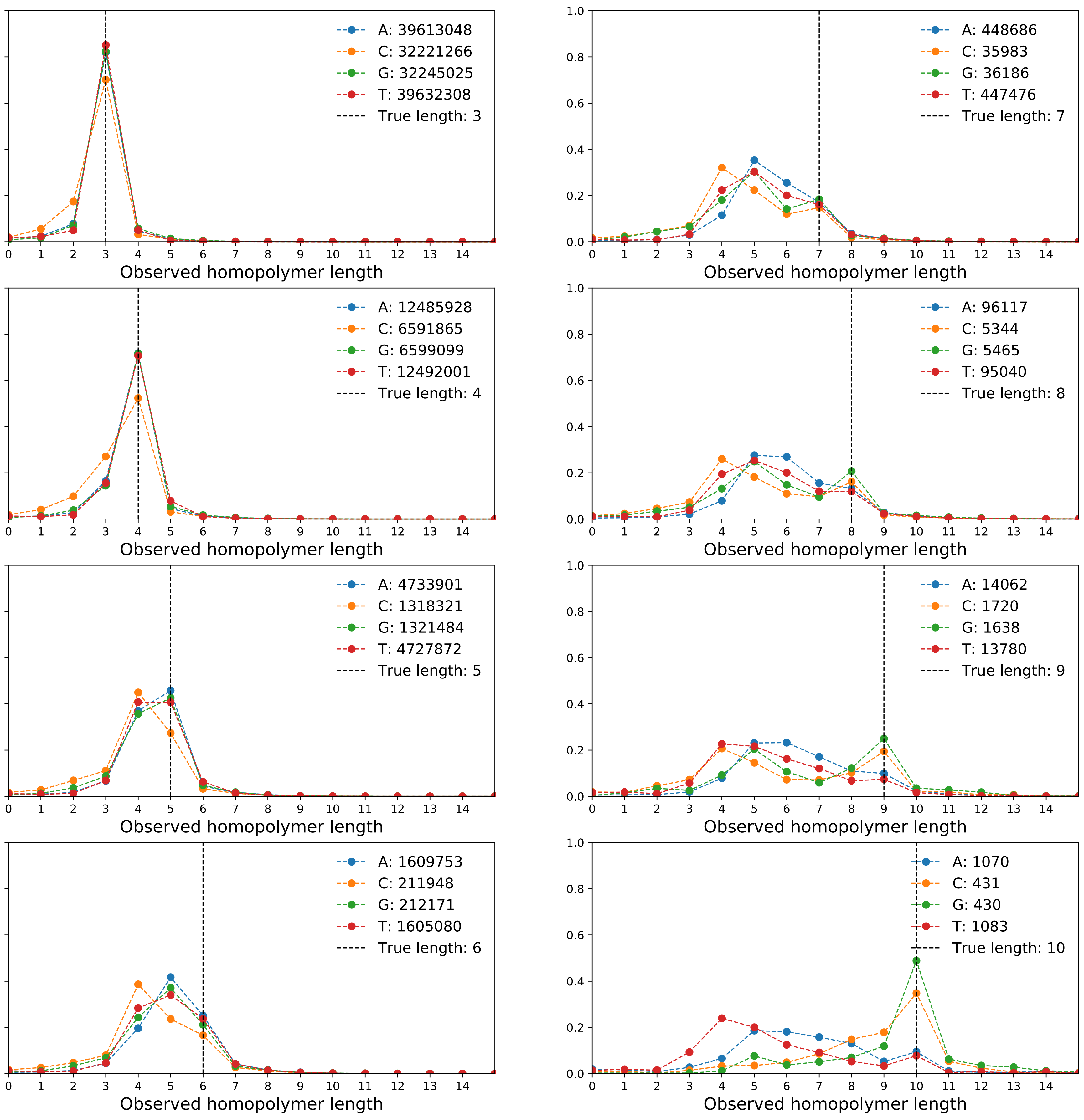
Context Dependent Insertion

We similarly examined which kmers are most subject to insertion errors. We show the percent of times there is an erroneous insertion between the two central letters (indicated in bold) within a 6-mer. The rate of insertion of each base is indicated by the corresponding column. Most insertions are single base insertions (not shown). The insertion rate ranges from 0.10% to 11.5%. The least erroneous cases have a T-dimer in the middle, showing insertions between T’s is rare.

	Context	%Error	Counts	%A	%C	%G	%T
most errors	AC**GT**GC	11.52	760651	69.18	9.33	10.72	10.77
	GT**CC**TT	10.39	497552	24.63	8.67	63.62	3.08
	AT**CC**TT	10.06	756856	27.71	9.04	59.64	3.61
	CT**CA**GT	9.86	474684	19.25	41.90	33.21	5.63
	GG**CA**AG	9.59	1138626	46.36	6.27	30.11	17.25
fewest errors	AG**TT**CC	0.12	752195	27.92	19.24	31.98	20.86
	CG**TT**CA	0.12	1245050	25.76	23.41	26.06	24.77
	AG**TT**CA	0.12	1009754	24.35	18.03	33.55	24.07
	GG**TT**CC	0.12	647142	19.61	20.28	36.05	24.06
	CG**TT**CC	0.11	897279	24.91	21.25	26.22	27.62

Homopolymer Length Errors

We examined the lengths of homopolymers in MinION data and found that their distribution has high variance and a biased mode relative to the true homopolymer length. The mode and mean of the distribution are smaller than the true length, leading to systematic under-estimation of the true homopolymer length when consensus correction is used. Interestingly, the length distributions for the four bases differ in shape.



Try our error profiler *counterr*!

Figures and tables were generated by our open source package *counterr* available at https://github.com/dayzerodx/counterr:

```
counterr align.bam ref.fasta --outdir results
```

Acknowledgements

We thank Jared Simpson and Keith Robison for helpful discussions and suggestions.

References

[1] Wick R, Judd L, Gorrie C, Holt K. *Completing bacterial genome assemblies with multiplex MinION sequencing* 14/09/2017. M Gen 3(10): doi:10.1099/mgen.0.000132
[2] Gomez-Eichelmann MC, Levy-Mustri A, Ramirez-Santos J. *Presence of 5-methylcytosine in CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria*. J Bacteriol. 1991;173(23):7692-4.