

Supplementary Material

Yuansheng Liu¹, Yicai Zhang¹, Yichen Li¹, Sisi Yuan^{2,*}, Xiao Luo^{3,*}

¹ College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

² College of Computer Science and Software Engineering, Shenzhen University, Shenzhen,
China

³ College of Biology, Hunan University, Hunan, China

*To whom correspondence should be addressed.

Email: syuan4@charlotte.edu

Email: xluo@hnu.edu.cn

Supplementary Tables and Figures

Method	Mismatches per 100kbp	Indels per 100kbp	Genome fraction (%)	Local misassemblies	Contigs
10x					
DADEC	386.04	48.98	99.298	2,461	83,869
FMLRC	1,248.38	2,680.50	99.63	6,388	83,864
F_HERO	2,461.06	1,253.96	99.348	66,525	83,865
Ratatosk	455.75	1,297.24	99.70	480	83,862
R_HERO	1,499.22	651.03	99.359	19,924	83,862
LoRDEC	853.87	1,852.19	99.557	687	83,862
L_HERO	998.14	502.03	98.831	18,107	83,863
CoLoRMap	2,553	9,219.62	99.73	268	83,896
proovread	648.04	205.63	99.062	191	83,859
20x					
DADEC	212.98	18.52	99.729	451	83,869
FMLRC	674.13	949.55	99.839	3,241	83,864
F_HERO	1,039.22	364.80	99.84	23,031	83,864
Ratatosk	245.75	429.09	99.827	322	83,861
R_HERO	566.72	183.54	99.832	6,134	83,861
LoRDEC	648.17	1,309.31	99.669	605	83,861
L_HERO	887.89	570.70	99.732	14,444	83,864
CoLoRMap	2,703.49	6,415.38	99.813	539	83,884
proovread	433.81	112.04	99.833	202	83,861
30x					
DADEC	130.69	10.12	99.674	199	83,880
FMLRC	328.49	375.53	99.717	2,663	83,864
F_HERO	529.93	154.46	99.834	9,497	83,868
Ratatosk	233.55	328.09	99.837	283	83,862
R_HERO	328.97	101.72	99.835	3,333	83,861
LoRDEC	393.82	586.41	99.718	555	83,862
L_HERO	504.33	249.18	99.828	6,864	83,861
CoLoRMap	2060.53	3,958.01	99.633	635	83,886
proovread	333.29	94.01	99.837	248	83,859
40x					
DADEC	231.25	14.89	99.844	87	83,889
FMLRC	311.63	259.02	99.827	2,140	83,864
F_HERO	429.84	134.88	99.83	5,889	83,865
Ratatosk	297.97	453.23	99.82	281	83,862
R_HERO	276.88	105.85	99.828	3,460	83,862
LoRDEC	325.30	326.63	99.833	518	83,862
L_HERO	317.17	144.79	99.827	3,071	83,863
CoLoRMap	2,065.83	3,228.97	99.82	672	83,896
proovread	257.44	114.07	99.832	596	83,859
50x					
DADEC	155.87	7.09	99.702	77	83,884
FMLRC	217.21	189.00	99.653	2,080	83,864
F_HERO	179.97	70.84	99.211	4,465	83,863
Ratatosk	329.75	620.9	99.667	241	83,862
R_HERO	216.78	106.42	99.182	4,082	83,861
LoRDEC	190.80	202.40	99.699	512	83,862
L_HERO	173.11	101.45	98.969	1,692	83,863
CoLoRMap	1,644.05	2,225.43	99.628	666	83,891
proovread	142.01	80.74	99.621	505	83,859

Supplementary Table 1. Error correction benchmarking results for 30strains of varioussequencing coverages. The average sequencing coverage is set as 10x, 20x, 30x, 40x, and 50x.

Method	Mismatches per 100kbp	Indels per 100kbp	Genome fraction (%)	Local misassemblies	Contigs
<i>Arabidopsis thaliana</i>					
DADEC	305.06	145.95	99.776	4627	222197
FMLRC	500.15	509.29	99.571	11873	222205
F_HERO	527.77	395.51	85.308	11599	222205
Ratatosk	494.51	1,495.31	99.901	883	222205
R_HERO	581.95	1,319.83	86.876	2975	222201
LoRDEC	495.77	398.83	99.718	1709	222205
L_HERO	552.79	454.12	89.48	2729	222203
CoLoRMap	555.31	893.22	99.888	557	222205
proovread	-	-	-	-	-

Supplementary Table 2. Benchmarking results for error correction on simulated PacBio datasets of *Arabidopsis thaliana* with 20x coverage and approximately 10% error rate . Bolded values indicate the optimal results.

3 E.coli	all			insert			delete			mismatch		
	OC	UC	CC	OC	UC	CC	OC	UC	CC	OC	UC	CC
raw	-	16,513,543	-	-	7,194,851	-	-	4,989,735	-	-	4,328,957	-
DADEC	47,498	44,410	16,335,011	567	2,777	16,376,644	644	2,269	16,377,152	46,296	39,468	16,339,953
FMLRC	214940	414,993	15,964,428	53,208	148,343	16,231,078	40,009	100,571	16,278,850	124,270	171,966	16,207,455
F_HERO	179,514	292,150	16,087,271	38,762	102,845	16,276,576	27,763	69,972	16,309,449	115,084	123,535	16,255,886
Ratatosk	571,922	1,462,430	14,916,991	153,277	570,100	15,809,321	106,799	386,166	15,993,255	315,563	524,649	15,854,772
R_HERO	301,346	986,423	15,392,998	40,495	370,616	16,008,805	27,092	250,311	16,129,110	235,467	379,688	15,999,733
LoRDEC	263,412	1,746,432	14,632,989	44,702	684,510	15,694,911	32,234	466,271	15,913,150	188,255	617,958	15,761,463
L_HERO	296,566	1,159,616	15,219,805	51,603	443,015	15,936,406	33,594	300,405	16,079,016	213,607	432,254	15,947,167
CoLoRMap	657,303	369,571	16,009,850	92,519	125,457	16,253,964	65,703	86,382	16,293,039	502,004	161,959	16,217,462
proovread	458,374	402,486	15,976,935	36,432	131,285	16,248,136	41,267	97,866	16,281,555	382,290	178,577	16,200,844

Supplementary Table 3. The statistics of over-correction (oc), under-correction (uc), and correct-correction (cc) counts for various error correction methods on the 3E.coli dataset, with separate breakdowns for three genomic error types: insert errors, delete errors, and mismatches.

<i>Arabidopsis thaliana</i>	all			insert			delete			mismatch		
	OC	UC	CC	OC	UC	CC	OC	UC	CC	OC	UC	CC
DADEC	7,494,452	3,972,527	186,020,993	929,742	1,465,167	188,528,353	711,404	1,057,123	188,936,397	5,894,894	1,471,890	188,521,630
FMLRC	19,611,713	12,415,598	177,585,945	4,242,028	4,747,680	185,253,863	4,265,770	3,532,934	186,468,609	11,440,521	4,267,148	185,734,395
F_HERO	126,895,297	15,580,404	174,420,788	38,183,439	5,108,523	184,892,669	17,008,837	2,947,980	187,053,212	78,832,920	8,166,307	181,834,885
Ratatosk	14,711,611	41,342,673	148,658,390	5,806,008	18,388,263	171,612,800	4,138,748	12,803,695	177,197,368	4,855,883	10,429,959	179,571,104
R_HERO	114,032,442	33,432,453	156,566,398	37,108,693	13,602,750	176,396,101	16,132,488	8,749,767	181,249,084	67,105,768	11,778,314	178,220,537
LoRDEC	12,629,135	17,361,844	172,640,314	1,540,266	6,833,995	183,168,163	1,241,045	4,757,606	185,244,552	9,917,001	5,934,313	184,067,845
L_HERO	116,236,714	20,916,609	169,085,079	35,502,349	7,487,552	182,514,136	14,255,410	4,457,488	185,544,200	73,192,473	9,641,365	180,360,323
CoLoRMap	19,784,038	14,119,111	175,883,047	6,263,805	5,995,242	184,006,916	4,580,826	4,235,888	185,766,270	9,055,539	3,990,088	186,012,070

Supplementary Table 4. the statistics of over-correction (oc), under-correction (uc), and correct-correction (cc) counts for various error correction methods on the *Arabidopsis thaliana* dataset, with separate breakdowns for three genomic error types: insert errors, delete errors, and mismatches.

Method	FDR	FNR	read level			base level		
			OC	UC	CC	OC	UC	CC
raw	-	-	-	-	-	-	16513543	-
DADEC	0.00289	0.00271	40	57	57	47,498	44,410	16,335,011
	0.00289	0.00271				47,427	44,398	16,315,045
FMLRC	0.01328	0.02534	108	61	53	214,940	414,993	15,964,428
	0.01314	0.25378				211,473	413,568	15,882,500
F_HERO	0.01103	0.01784	232	65	49	179,514	292,150	16,087,271
	0.01039	0.01777				166,798	287,484	15,888,308
Ratatosk	0.03692	0.08928	245	81	33	571,922	1462,430	14,916,991
	0.03644	0.09005				556,922	1457,509	14,727,517
R_HERO	0.01920	0.06022	375	78	36	301,346	986,423	15,392,998
	0.01858	0.06099				286,185	980,226	15,114,768
LoRDEC	0.01768	0.10662	91	81	33	263,412	1746,432	14,632,989
	0.01740	0.10702				257,916	1745,437	14,563,717
L_HERO	0.01911	0.07079	232	65	49	296,566	1,159,616	15,219,805
	0.01038	0.01778				166,798	287,484	15,888,308
CoLoRMap	0.03944	0.02256	276	49	65	657,303	369,571	16,009,850
	0.03911	0.02257				646,236	366,575	15,877,582
proofread	0.02789	0.02457	382	38	76	458,374	402,486	15,976,935
	0.02759	0.02465				448,821	399,910	15,818,850

Supplementary Table 5. This table presents the False Negative Rate (FNR) and False Discovery Rate (FDR) of various error correction methods on the 3E.ecoli dataset, along with the counts of over-correction (oc), under-correction (uc), and correct-correction (cc) at both the base level and reads level. It also includes the base-level counts obtained after filtering out reads classified as over-correction (oc) at the reads level.

Method	FDR	FNR	read level			base level		
			OC	UC	CC	OC	UC	CC
DADEC	0.03873	0.02091	4,023	2,182	589	7,494,452	3,972,527	186,020,993
	0.03673	0.02088				6,970,734	182,795,600	3,897,361
FMLRC	0.09945	0.06535	9,916	2,206	565	19,611,713	12,415,598	177,585,945
	0.09707	0.06531				18,092,887	11,760,307	168,295,093
F_HERO	0.42114	0.082	19,435	2,221	550	126,895,297	15,580,404	174,420,788
	0.41436	0.0807				111,978,739	13,894,024	158,265,968
Ratatosk	0.09005	0.21759	2,288	2,352	419	14,711,611	41,342,673	148,658,390
	0.08835	0.21872				14,254,630	41,177,600	147,086,337
R_HERO	0.42141	0.17596	11,198	2,312	459	114,032,442	33,432,453	156,566,398
	0.41346	0.41346				17,310,090	32,448,434	148,131,997
LoRDEC	0.06817	0.09138	2,678	2,304	467	12,629,135	17,361,844	172,640,314
	0.06740	0.09210				12,331,504	17,310,090	170,638,498
L_HERO	0.40739	0.11009	11,769	2,308	463	116,236,714	20,916,609	169,085,079
	0.39782	0.11090				105,777,017	19,971,641	160,108,805
CoLoRMap	0.10111	0.74310	2,299	2,280	491	19,784,038	14,119,111	175,883,047
	0.09928	0.07433				19,210,694	13,994,979	174,281,481

Supplementary Table 6. This table presents the False Negative Rate (FNR) and False Discovery Rate (FDR) of various error correction methods on the simulated PacBio datasets of *Arabidopsis thaliana*, along with the counts of over-correction (oc), under-correction (uc), and correct-correction (cc) at both the base level and reads level. It also includes the base-level counts obtained after filtering out reads classified as over-correction (oc) at the reads level.

Method	Mismatches per 100kbp	Indels per 100kbp	Genome fraction (%)	NGA50 (kbp)	Contigs
<i>Drosophila melanogaster</i>					
(canu)					
DADEC	206.98	78.07	95.22	454,875	984
FMLRC	258.41	219.28	58.11	40,768	1,584
F_HERO	334.08	319.79	7.36	-	495
Ratatosk	152.00	152.49	59.87	39,906	1,504
R_HERO	360.13	379.96	21.96	-	941
LoRDEC	207.10	146.26	75.00	95,559	1,485
L_HERO	351.07	363.85	11.53	-	588
CoLoRMap	198.45	134.91	80.99	132,340	1,202
proovread	235.12	92.17	91.17	502,254	809
NWC-PacBio					
(hifiasm)					
DADEC	170.66	14.19	65.243	9,261	9,261
FMLRC	190.87	44.1	61.734	5,463	5,463
F_HERO	324.26	65.13	64.922	4,226	4,226
Ratatosk	344.29	236.49	47.452	1,042	1,042
R_HERO	374.76	118.31	56.24	917	917
LoRDEC	1015.1	399.38	37.458	238	238
L_HERO	1245.11	209.13	44.28	-	186
CoLoRMap	1543.19	1847.43	38.97	-	224
proovread	-	-	-	-	-
NWC-PacBio					
(canu)					
DADEC	2,698.12	225.16	66.484	57,049	248
FMLRC	3,075.98	430.39	66.406	59,986	248
F_HERO	3,331.52	374.1	67.081	40,423	351
Ratatosk	902.21	620.86	50.834	55,903	191
R_HERO	2,468.41	569.32	65.593	35,426	414
LoRDEC	2,968.33	363.27	64.451	51,408	256
L_HERO	3,306.37	424.79	64.876	40,524	330
CoLoRMap	2,066.33	1,124.47	61.088	45,032	274
proovread	1,854.84	1,120.28	62.285	45,765	248

Supplementary Table 7. Quality metrics of the *Drosophila melanogaster* and NWC assemblies generated by hifiasm and canu on datasets corrected with different tools. Bolded values indicate optimal results.

Method	presion	recall	F1
raw	0.7039	0.7059	0.7049
DADEC	0.7143	0.7162	0.7153
FMLRC	0.7051	0.7095	0.7073
F_HERO	0.6718	0.6757	0.6737
Ratatosk	0.6935	0.6966	0.6950
R_HERO	0.6703	0.6731	0.6717
LoRDEC	0.7166	0.7155	0.7160
L_HERO	0.6791	0.6831	0.6811
CoLoRMap	0.6714	0.6689	0.6701
proovread	0.657	0.6353	0.6460

Supplementary Table 8. Benchmarking results of strain-level taxonomic classification.

Method	CPU(h)	wall time(h)	memory (G)
DADEC	0.462	0.03	2.168
FMLRC	0.529	0.017	0.581
F_HERO	5.418	0.186	2.835
Ratatosk	1.426	0.054	0.606
R_HERO	5.317	0.182	2.916
LoRDEC	4.074	0.271	1.971
L_HERO	5.435	0.178	3.236
CoLoRMap	3.529	0.433	10.135
proovread	6.752	0.356	7.369

Supplementary Table 9. The CPU time (h), wall time (h), and memory (G) for correcting simulated PacBio CLR reads of 3E.coli.

Method	CPU(h)	wall time(h)	memory (G)
DADEC	6.405	0.303	3.879
FMLRC	4.546	0.144	1.342
F_HERO	339.75	10.611	35.346
Ratatosk	6.294	0.209	1.075
R_HERO	361.033	11.37	33.276
LoRDEC	18.486	0.607	3.872
L_HERO	279.017	8.852	29.408
CoLoRMap	20.738	2.971	39.015
proovread	279.861	10.78	53.417

Supplementary Table 10. The CPU time (h), wall time (h), and memory (G) for correcting real PacBio CLR reads of the NWC.



Supplementary Figure 1. Error correction Genome fraction (%) for 30 strains of various sequencing coverages. The average sequencing coverage is set as 10x, 20x, 30x, 40x, and 50x.

Commands and versions of tools used for comparison

- pbsim2:

```
pbsim --prefix ecoli --depth 10 --hmm_model P6C4.model $ref --accuracy-mean 0.85 --length-min 1000
```

- art

```
art -ss HS25 -i $ref -o short_reads -l 150 -f 20 -p -m 500 -s 30 -sam
```

- NanoSim

```
simulator.py genome -dna_type linear -rg $ref -o n1.nanosim -n 4000 -med 9000 -sd 1.05 -max 100000 -min 1000 -c human_NA12878_DNA_FAB49712_albacore/training
```

- DADEC v1.0.1

```
DADEC -s $sreads -l $lreads -r 0.08
```

- FMLRC

```
cat $sreads | awk 'NR % 2 == 0' | sort | tr NT TN | ropebwt2 -LR | tr NT TN | fmlrc2-convert short.npy fmlrc2 short.npy $lreads fmlrc1.fa
```

- Lordec

```
lordec-correct -k 31 -s 5 -i $lreads -2 $sreads -o lordec1.fa
```

- Ratatosk

```
Ratatosk correct -v -s $sreads -l $lreads -o ratatosk1
```

- HERO

```
python HERO.py -r $sreads -lc fmlrc3.fa -p -o F_HERO.fa -i 1 -s 10  
python HERO.py -r $sreads -lc lordec3.fa -p -o L_HERO.fa -i 1 -s 10  
python HERO.py -r $sreads -lc ratatosk3.fa -p -o F_HERO.fa -i 1 -s 10
```

- ColorMap

```
sh runCorr.sh $lreads $sreads ./map_cor colormap
```

- prooveared

```
proovread -l $lreads -s $sreads --overwrite -p proovread
```

- hifiasm

```
hifiasm -o ./asm corrected.fa
```

- canu

```
canu -p canu -corrected -assemble useGrid=false genomeSize=$size minInputCoverage=1 -pacbio corrected.fa
```

- Quast

```
Python quast.py -r $ref corrected.fa -o quastEvaluate
```

- Centrifuge

```
centrifuge-build --conversion-table ref.conv \  
    --taxonomy-tree nodes.dmp \  
    --name-table names.dmp \  
    ref.fa ref  
centrifuge -f -x ref -U corrected.fa -S centrifuge.txt
```