RACER: Rapid and Accurate Correction of Errors in Reads

Lucian Ilie and Michael Molnar

— Supplementary material —

1 Dataset description and mapping

Table 4 gives the full name of the organisms used in the paper and details of the read searching and BWA mapping procedures. The two procedures have very different purposes. Reads found by the exact search are considered error free. Exact search is used for evaluating the error-correcting performance of the algorithms by performing exact search before and after correction. For instance, only 45% of the *D.melanogaster* reads are error free before correction. BWA mapping however is used to alter the datasets by retaining only the reads that can be mapped. For instance, in the case of *D.melanogaster*, 83.91% of the reads could be mapped and hence the remaining 16.09% are removed.

Table 4: Full organism names and details of exact search and BWA mapping of the datasets.

Dataset	Exact search			BWA		
	reads found	base pairs	%	reads mapped	base pairs	%
Lactococcus lactis	3,625,393	130,514,165	82.96	4,274,880	153,895,680	97.82
Treponema pallidum	5,216,848	182,589,671	73.13	6,860,022	240,100,770	96.16
Escherichia coli	2,119,404	158,955,289	61.36	3,444,268	258,320,100	99.72
Bacillus subtilis	2,272,114	170,408,576	64.56	3,314,397	248,579,775	94.17
Escherichia coli	2,717,070	203,780,256	62.59	4,276,543	320,740,725	98.51
Pseudomonas aeruginosa	9,005,025	324,180,884	96.76	9,291,197	334,483,092	99.83
Escherichia coli	2,514,335	118,173,733	17.45	4,674,277	219,691,019	32.44
Leptospira interrogans serovar Lai	5,460,023	546,002,338	77.27	6510694	651,069,400	92.14
Leptospira interrogans serovar Copenhageni	5,782,338	578,233,793	81.13	6584317	658,431,700	92.38
Escherichia coli	17,649,121	635,368,369	84.78	20,462,872	736,663,392	98.30
Haemophilus influenzae	20,270,782	851,372,838	84.69	23,466,293	985,584,306	98.04
Staphylococcus aureus	6,700,708	509,253,774	26.22	15,336,233	1,165,553,708	60.02
Saccharomyces cerevisiae	29,961,488	2,277,073,060	57.55	46,405,503	3,526,818,228	89.14
Caenorhabditis elegans	47,684,047	4,768,404,660	70.52	63,663,585	6,366,358,500	94.15
Drosophila melanogaster	46,209,308	3,141,591,308	45.50	85,208,029	5,677,956,242	83.91

2 Actual time and space values for raw datasets

TIME (s)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	1,741	852	1,694	623	(b)	174	208	559	(b)	23
T.pallidum	7,325	2,636	4,309	2,266	5,373	780	534	1,028	548	71
E.coli 75a	6,330	3,074	842	2,248	6,672	1,366	512	658	787	99
B.subtilis	6,192	3,114	1,309	1,988	6,322	1,067	513	698	767	111
E.coli 75b	8,427	4,001	1,109	3,306	8,504	1,322	682	727	1,348	118
P.aeruginosa	3,663	916	1,348	4,744	5,652	409	424	1,000	658	63
E.coli 47	3,960	7,862	22,874	4,509	(b)	2,790	738	3,596	(b)	483
L.interrogans L	64,343	7,837	1,678	3,334	18,115	2,268	3,456	851	2,460	180
L.interrogans C	69,289	8,831	1,840	3,288	18,724	2,029	3,648	1,174	2,070	163
E.coli 36	19,133	9,610	10,123	4,295	13,681	1,515	1,497	2,228	1,608	202
H.influenzae	84,724	13,562	6,108	10,309	18,736	2,127	4,282	(d)	2,125	266
S.aureus	142,233	3,753	(d)	29,496	(a)	8,518	8,714	(d)	(a)	1,077
S.cerevisiae	359,097	8,081	8,915	29,174	100,425	13,732	18,894	4,812	20,212	1,294
C.elegans	(a)	(a)	19,975	104,010	(a)	34,165	(a)	6,906	(a)	2,618
D.melanogaster	(a)	(c)	69,747	128,981	(a)	46,476	(a)	24,352	(a)	6,242

SPACE (MB)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	1,996	2,979	493	579	(b)	515	52,544	1,619	(b)	722
T.pallidum	3,271	4,738	1,297	768	35,178	569	53,819	1,838	99,492	798
E.coli 75a	3,810	4,895	1,230	1,765	35,987	1,437	54,355	1,944	99,394	1,773
B.subtilis	4,243	4,981	1,332	1,595	36,444	1,438	54,789	1,945	99,600	1,773
E.coli 75b	4,910	6,064	1,732	3,598	36,454	1,477	55,427	2,180	99,606	1,803
P.aeruginosa	3,579	6,340	1,851	790	35,865	1,167	54,189	2,575	99,525	1,429
E.coli 47	11,934	12,755	3,364	1,914	(b)	6,433	62,453	4,045	(b)	6,944
L.interrogans L	8,210	7,253	2,673	2,431	38,030	966	58,727	3,226	99,628	1,383
L.interrogans C	7,982	7,343	2,800	2,322	37,961	964	58,499	3,139	99,591	1,252
E.coli 36	8,235	14,178	4,090	1,070	37,872	1,411	58,816	5,008	99,515	1,949
H.influenzae	10,231	19,132	4,253	1,060	38,142	1,278	60,811	(d)	99,637	1,669
S.aureus	43,981	36,117	(d)	8,837	(a)	4,561	94,499	(d)	(a)	5,109
S.cerevisiae	41,278	77,893	15,056	4,421	69,125	5,628	91,858	15,581	100,002	6,267
C.elegans	(a)	(a)	32,001	10,406	(a)	17,803	(a)	32,688	(a)	18,263
D.melanogaster	(a)	(c)	36,374	21,069	(a)	41,206	(a)	36,868	(a)	42,229

3 Actual time and space values for mapped datasets

TIME (s)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	1,667	823	1,261	1,760	2,626	157	183	551	317	20
T.pallidum	7,366	1,669	2,170	2,565	3,964	462	518	906	448	49
E.coli 75a	6,017	3,069	709	2,370	6,442	951	485	640	831	92
B.subtilis	6,239	2,926	1,101	1,829	5,272	881	480	657	624	80
E.coli 75b	8,534	4,028	865	2,666	9,033	1,187	702	787	1,656	186
P.aeruginosa	3,573	867	1,136	4,189	5,726	498	386	835	642	43
E.coli 47	1,705	2,375	9,053	3,205	5,232	819	252	1,747	1,041	166
L.interrogans L	61,714	7,084	1,538	4,381	16,883	1,665	3,210	932	1,983	155
L.interrogans C	65,612	8,062	1,777	3,006	16,459	1,318	3,419	848	2,650	155
E.coli 36	18,691	5,108	4,693	6,008	12,872	942	1,392	1,588	1,443	101
H.influenzae	26,876	7,404	4,573	8,993	17,428	1,269	4,419	(d)	1,876	160
S.aureus	9,214	3,928	5,567	9,528	31,471	4,901	6,388	2,857	8,101	81
S.cerevisiae	357,529	7,074	7,799	34,476	84,381	11,899	19,434	3,924	14,040	840
C.elegans	262,581	(a)	15,831	104,010	(a)	33,072	(a)	5,773	(a)	1,932
D.melanogaster	152,651	(c)	46,248	148,164	(a)	30,117	(a)	10,827	(a)	2,883

SPACE (MB)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	1,962	2,914	233	543	34,238	467	52,573	1,553	99,514	730
T.pallidum	3,067	4,557	449	689	34,829	478	53,616	1,616	99,538	768
E.coli 75a	3,786	4,927	371	1,724	35,907	1,393	54,330	1,533	99,615	1,730
B.subtilis	3,511	4,694	424	1,680	34,837	1,389	54,056	1,713	99,648	1,727
E.coli 75b	4,747	5,974	399	1,718	38,430	1,420	55,268	1,734	99,548	1,747
P.aeruginosa	3,571	6,330	311	781	35,418	1,079	54,118	1,507	99,528	1,277
E.coli 47	3,892	4,209	1,184	1,043	36,399	1,530	54,409	1,782	99,598	1,852
L.interrogans L	7,375	6,683	611	2,268	36,964	859	57,891	2,413	99,578	1,217
L.interrogans C	7,279	6,759	472	2,712	36,675	859	57,798	2,409	99,647	1,217
E.coli 36	7,772	13,937	819	878	36,996	1,141	58,354	2,882	99,718	1,683
H.influenzae	9,628	18,757	629	1,016	36,930	843	60,146	(d)	99,592	1,351
S.aureus	18,267	22,941	3,083	3,086	53,599	2,163	68,782	3,591	99,948	2,535
S.cerevisiae	34,491	65,590	3,041	3,411	50,815	3,233	85,069	10,828	99,789	3,937
C.elegans	75,096	(a)	6,685	10,406	(a)	16,765	(a)	24,747	(a)	17,271
D.melanogaster	83,163	(c)	21,516	17,802	(a)	39,482	(a)	24,339	(a)	40,272

4 Comparison on the mapped datasets

A comparison has been performed also on the BWA-mapped datasets and the results are presented in Table 5. The values have been computed in the same way with those in Table 2 of the paper. In general, similar performance is shown by the programs, except the percentage of errors corrected are higher and there are fewer dataset that could not be run. Unexpectedly, the error correction of Reptile decreased very much, a thing we could not correct even by varying the parameters.

Table 5: Comparison on the mapped datasets. Average is computed for the same dataset as in Table 1 to facilitate comparison between raw and mapped datasets.

ERROR	serial						parallel (24 cores)			
CORRECTION (%)	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	75.22	92.16	81.67	0.11	84.87	92.02	75.21	81.82	84.74	92.02
T.pallidum	50.81	91.72	68.77	0.77	70.72	92.35	50.82	69.18	70.55	92.35
E.coli 75a	26.49	83.08	1.51	0.07	73.05	83.91	26.49	1.65	72.14	83.91
B.subtilis	73.93	92.64	63.90	32.25	47.93	93.55	73.93	63.92	47.46	93.55
E.coli 75b	11.65	78.22	1.42	0.07	40.35	77.80	11.65	1.16	39.56	77.80
P.aeruginosa	84.28	82.92	60.54	3.26	66.74	89.51	84.27	60.56	66.71	89.51
E.coli 47	3.78	77.00	45.79	0.02	56.44	82.67	3.78	45.97	55.53	82.67
L.interrogans L	75.18	91.58	76.07	0.12	85.44	90.81	75.20	76.11	84.22	90.81
L.interrogans C	75.35	90.05	75.33	1.61	80.31	89.27	75.34	75.47	79.44	89.27
E.coli 36	67.74	90.74	90.73	0.07	86.65	90.80	67.75	90.85	86.19	90.80
H.influenzae	48.65	80.04	69.60	8.32	60.66	84.33	48.57	(d)	60.45	84.33
S.aureus	0.25	0.43	32.75	0.07	40.49	47.00	0.25	32.83	39.35	47.00
S.cerevisiae	5.97	0.30	8.92	0.30	11.94	14.68	5.97	9.00	11.75	14.68
C.elegans	27.53	(a)	47.90	0.26	(a)	65.96	(a)	47.92	(a)	65.96
D.melanogaster	40.12	(c)	47.01	0.00	(a)	57.04	(a)	47.19	(a)	57.04
AVERAGE	58.18	87.62	54.78	4.78	68.90	88.50	58.18	54.86	68.28	88.50

TIME (s/MB)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	11.11	5.49	8.40	11.73	17.50	1.05	1.22	3.67	2.11	0.13
T.pallidum	30.94	7.01	9.11	10.77	16.65	1.94	2.18	3.80	1.88	0.21
E.coli 75a	24.36	12.42	2.87	9.59	26.08	3.85	1.96	2.59	3.36	0.37
B.subtilis	24.78	11.62	4.37	7.27	20.94	3.50	1.91	2.61	2.48	0.32
E.coli 75b	27.48	12.97	2.79	8.59	29.09	3.82	2.26	2.53	5.33	0.60
P.aeruginosa	11.18	2.71	3.56	13.11	17.92	1.56	1.21	2.61	2.01	0.13
E.coli 47	2.64	3.68	14.02	4.96	8.10	1.27	0.39	2.71	1.61	0.26
L.interrogans L	91.58	10.51	2.28	6.50	25.05	2.47	4.76	1.38	2.94	0.23
L.interrogans C	96.53	11.86	2.61	4.42	24.21	1.94	5.03	1.25	3.90	0.23
E.coli 36	26.15	7.15	6.57	8.41	18.01	1.32	1.95	2.22	2.02	0.14
H.influenzae	28.03	7.72	4.77	9.38	18.18	1.32	4.61	(d)	1.96	0.17
S.aureus	4.98	2.12	3.01	5.14	16.99	2.65	3.45	1.54	4.37	0.04
S.cerevisiae	94.75	1.87	2.07	9.14	22.36	3.15	5.15	1.04	3.72	0.22
C.elegans	40.72	(a)	2.46	16.13	(a)	5.13	(a)	0.90	(a)	0.30
D.melanogaster	23.18	(c)	7.02	22.50	(a)	4.57	(a)	1.64	(a)	0.44
AVERAGE	41.63	9.53	4.27	8.58	22.24	2.55	2.66	2.38	2.99	0.28

SPACE (MB/MB)	serial						parallel (24 cores)			
	Coral	HiTEC	Quake	Reptile	SHREC	RACER	Coral	Quake	SHREC	RACER
L.lactis	13.08	19.42	1.55	3.62	228.20	3.11	350.41	10.35	663.28	4.87
T.pallidum	12.88	19.14	1.89	2.89	146.27	2.01	225.17	6.79	418.03	3.23
E.coli 75a	15.32	19.94	1.50	6.98	145.34	5.64	219.91	6.21	403.21	7.00
B.subtilis	13.95	18.65	1.68	6.67	138.39	5.52	214.73	6.80	395.85	6.86
E.coli 75b	15.29	19.24	1.29	5.53	123.77	4.57	178.00	5.58	320.61	5.63
P.aeruginosa	11.18	19.81	0.97	2.44	110.85	3.38	169.38	4.72	311.50	4.00
E.coli 47	6.03	6.52	1.83	1.61	56.36	2.37	84.25	2.76	154.22	2.87
L.interrogans L	10.94	9.92	0.91	3.37	54.85	1.27	85.91	3.58	147.77	1.81
L.interrogans C	10.71	9.94	0.69	3.99	53.96	1.26	85.03	3.54	146.60	1.79
E.coli 36	10.87	19.50	1.15	1.23	51.77	1.60	81.65	4.03	139.53	2.35
H.influenzae	10.04	19.56	0.66	1.06	38.52	0.88	62.74	(d)	103.88	1.41
S.aureus	9.86	12.39	1.66	1.67	28.94	1.17	37.14	1.94	104.25	1.37
S.cerevisiae	9.14	17.38	0.81	0.90	13.47	0.86	22.54	2.87	26.45	1.04
C.elegans	11.65	(a)	1.04	1.61	(a)	2.60	(a)	3.84	(a)	2.68
D.melanogaster	12.63	(c)	3.27	2.70	(a)	6.00	(a)	3.70	(a)	6.12
AVERAGE	12.64	17.02	1.26	4.14	103.15	3.16	157.47	5.16	285.39	4.08

5 Details on running the programs

Coral

The flag `-illumina` was needed since the default assumes the reads are 454. The flag `-fq` was needed for FASTQ format, the default assumes the files are in FASTA format. The command used for all data sets was:

`Coral -fq` [input file] [output file] `-illumina -p` [number of threads]

HiTEC

The genome size was set to the size given in Table 1. The per base error rate was set to 1% for all genomes, except for the survey data sets. The survey data sets per base error rates was set to 2% as indicated in the survey paper of Yang *et al.*. The command used for all data sets was:

`HiTEC` [input file] [output file] [genome size] [per base error rate]

Quake

The k-mer size was set according to the formula provided in the Quake paper and the website FAQ section. The k-mer size was set to 14 for *T.pallidum*, *L.lactis*, and *H.influenzae*, 16 for *S.cerevisiae*, and 17 for *D.melanogaster*. All the others had k-mer size set to 15. The command used for all data sets was:

`quake.py -r` [input file] `-k` [k-mer size] `-p` [number of threads]

Reptile

Reptile was run according to the instructions in the `readme` file. The first step in running Reptile is the conversion of the FASTQ file to a FATSA file with a separate quality file. The conversion script provided was used. The command used with all data sets in the input directory was:

`fastq-converter-v2.0.pl` [input directory] [output directory] `0 100 100 A`

The next two steps are parameter tuning used to find the proper settings for each data set. The `readme` file was used to find the proper settings. All data sets were run with the k-mer length set to 24, which is the default, and the parameters `KmerLen` and `Step` were set to 12 according to the instructions in the `readme` file. The data sets from the survey paper were set according to those in the paper. The error correcting program was then run using the command:

`reptile` [configuration file]

This program creates a file with the correction information. The final step creates the corrected data set using the original FASTA file and the correction information file. The command used for this step was:

`reptile_merger` [FASTA file] [correction information file] [output name]

The times provided for Reptile are the sum of the time required for each step. The memory is the maximum memory used from each of the steps.

SHREC

The memory given to the JVM was set to be 2 GB less than the memory available to the process (98 GB), according to the `readme` file for SHREC. The command used for all data sets was:

`java -Xmx96g SHREC -p` [number of threads] [input file] [output file]

RACER

The genome size was set to the size given in Table 1. The command used for all data sets was:

`RACER` [input file] [output file] [genome size]