CoLoRMap: Correcting Long Reads by Mapping short reads

Supplementary Material

Ehsan Haghshenas [1,2], Faraz Hach [1,3], S. Cenk Sahinalp [1,3,4] and Cedric Chauve [5]

[1]School of Computing Sciences, Simon Fraser University, Burnaby (BC), Canada, V5A 1S6
[2]MADD-Gen Graduate Program, Simon Fraser University, Burnaby (BC), Canada, V5A 1S6
[3]Vancouver Prostate Centre, Vancouver, BC, Canada, V6H 3Z6
[4]School of Informatics and Computing, Indiana University, Bloomington, IN, USA, 47405
[5]Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada, V5A 1S6

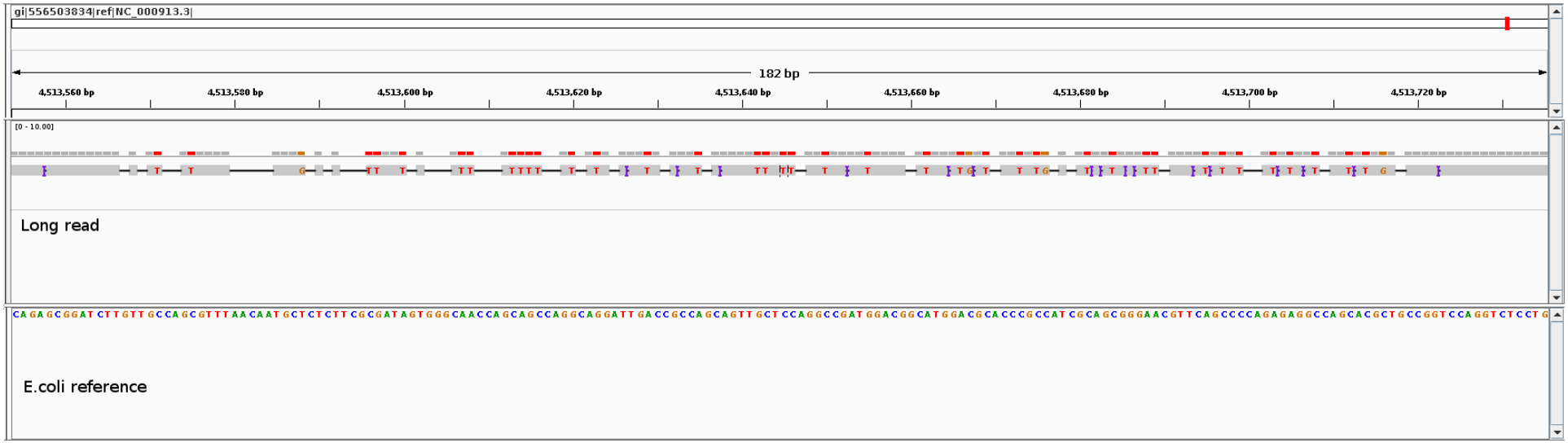

Figure S1: An example of a gap (region uncovered by short reads) on long read, exported from IGV software. There are so many sequencing errors that mapping short read in this region is very challenging. In the region shown here, the maximum exact match between long read and the reference genome is 4 bp long, in a region of size ≈ 150 bp.

Table S1: Data

	Bacteria	Yeast	Fruit fly
Reference organism			
Name	E. coli	S. cerevisiae	D. melanogaster
Strain	K-12 substr. MG1655	S288C	iso-1
Reference sequence	NC_000913	NC_0011{33-48} NC_001224	NT_0337{77-79} NC_0043{53-54} NC_0245{11-12} NT_037436
Genome size	4.6 Mbp	12.2 Mbp	140 Mbp
Pacbio data			
Accession ID	DevNet[1]	DevNet[2]	Bergman Lab[3]
Number of reads	33,360	231,604	901,564
Avg read length	2,938	6,055	1,505
Max read length	14,494	30,164	13,885
Number of bases	98 Mbp	1,402 Mbp	1,358 Mbp
Coverage	21x	114x	9.7x
Illumina data			
Accession ID	ERR022075[4]	SRR567755	ERX645969[4]
Number of reads	2,316,614	4,503,422	70,000,000
Read length	100 & 102	101	101
Coverage	50x	37x	50x
Insert size	504 ± 27	190 ± 80	240 ± 54

[1] Obtained from `https://github.com/PacificBiosciences/DevNet/wiki/E coli K12 MG1655 Hybrid Assembly`. Reads shorter than 100bp were filtered out.

[2] `https://github.com/PacificBiosciences/DevNet/wiki/Saccharomyces-cerevisiae-W303-Assembly-Contigs`

[3] `bergmanlab.ls.manchester.ac.uk/data/genomes/2057_PacBio.tgz`

[4] Only a subset of the data was used; the read file was truncated to 50x coverage.

Table S2: The effect of chunking on correction quality for CoLoRMap. CoLoRMap-W represents running of our software on the whole long read set.

			Aligned					
data set	Method	#Reads[a]	#Reads[b]	#Bases[c]	Size[d] (%)	Matched[e] (%)	Identity[f] (%)	Gen. cov.[g]
E.coli	CoLoRMap-W	33360	31247	83214224	89.49	86.59	99.02	100.00
(Full)	CoLoRMap	33360	31271	83344272	89.92	87.53	99.27	100.00
	CoLoRMap-W+OEA	33360	31165	82556432	89.07	86.63	99.20	100.00
	CoLoRMap+OEA	33360	31215	82915378	89.66	87.58	99.38	100.00
E.coli	CoLoRMap-W	30501	30302	76706135	95.98	93.44	99.23	100.00
(Trim)	CoLoRMap	30396	30190	76671240	96.26	94.24	99.46	100.00
	CoLoRMap-W+OEA	30501	30285	76338574	95.88	93.87	99.43	100.00
	CoLoRMap+OEA	30396	30183	76434210	96.21	94.56	99.58	100.00
E.coli	CoLoRMap-W	57458	57281	71449338	98.90	98.76	99.88	99.91
(Split)	CoLoRMap	48987	48840	73728458	99.11	98.99	99.90	99.91
	CoLoRMap-W+OEA	44037	43847	73062465	98.77	98.59	99.86	99.91
	CoLoRMap+OEA	40256	40101	74571341	98.99	98.84	99.89	99.91
Yeast	CoLoRMap-W	231594	223919	1211630012	88.07	83.12	96.69	99.85
(Full)	CoLoRMap	231594	223641	1207729568	88.60	85.62	98.30	99.83
	CoLoRMap-W+OEA	231594	223693	1207654403	88.02	83.61	97.10	99.85
	CoLoRMap+OEA	231594	223497	1205652269	88.55	85.72	98.40	99.83
Yeast	CoLoRMap-W	214765	211702	1004246265	93.35	88.61	96.94	99.85
(Trim)	CoLoRMap	211324	208188	1017551673	92.84	90.46	98.79	99.82
	CoLoRMap-W+OEA	214765	211710	1001174433	93.38	89.33	97.44	99.81
	CoLoRMap+OEA	211324	208310	1017391347	92.95	90.76	98.92	99.82
Yeast	CoLoRMap-W	1043237	1038397	631786029	96.65	96.14	99.44	99.68
(Split)	CoLoRMap	435140	432750	943502213	97.56	97.29	99.69	99.79
	CoLoRMap-W+OEA	676091	672731	707315725	97.36	96.60	99.28	99.77
	CoLoRMap+OEA	349998	347516	952997735	97.26	96.95	99.66	99.79

[a]the number of DNA sequences available after running the correction tool (may contain uncorrected sequences); in case of original data set, shows the total number of long reads. [b]the number of aligned sequences. [c]the number of bases aligned to the reference genome. [d]the percentage of aligned bases; that is column c / summed length of sequences in column a. [e]the percentage of matched bases; that is total number of matched bases / summed length of sequences in column a. [f]average identity; that is total number of matched bases / summed length of aligned regions in the reference genome. [g]percentage of the reference genome covered by the aligned sequences.

Table S3: Quality of corrected long reads for E.coli dataset obtained with different methods. Assessment is based on alignments of long reads to the reference genome obtained with BWA-MEM.

			Aligned					
data set	Method	#Reads[a]	#Reads[b]	#Bases[c]	Size[d] (%)	Matched[e] (%)	Identity[f] (%)	Gen. cov.[g]
E.coli	Original	33360	30830	86694498	88.45	76.66	94.07	**100.00**
E.coli	LSC	25426	25403	77867023	**93.06**	86.46	97.20	**100.00**
(Full)	proovread	24722	24046	73292276	91.83	**90.89**	**99.69**	**100.00**
	LoRDEC	33360	31371	82332501	90.16	88.74	99.44	**100.00**
	CoLoRMap	33360	**31693**	**84690697**	91.37	89.34	99.20	**100.00**
	CoLoRMap+OEA	33360	**31693**	84514038	91.39	89.67	99.33	**100.00**
E.coli	LSC	25426	25402	72255582	95.02	89.47	97.68	**100.00**
(Trim)	LoRDEC	31733	**31320**	**80137781**	94.32	93.49	**99.69**	**100.00**
	CoLoRMap	30396	30392	76686059	96.28	94.77	99.45	**100.00**
	CoLoRMap+OEA	30396	30392	76498317	**96.29**	**95.17**	99.59	**100.00**
E.coli	PacBioToCA	100100	100006	69100959	99.80	**99.77**	99.95	99.81
(Split)	proovread	30479	30477	71518136	99.50	99.40	**99.97**	99.67
	LoRDEC	49018	41679	**80036317**	99.33	99.28	99.96	99.83
	CoLoRMap	48987	48965	74256645	**99.82**	99.70	99.91	**99.91**
	CoLoRMap+OEA	40256	40235	75174811	99.79	99.65	99.90	**99.91**

[a]the number of DNA sequences available after running the correction tool (may contain uncorrected sequences); in case of original data set, shows the total number of long reads. [b]the number of aligned sequences. [c]the number of bases aligned to the reference genome. [d]the percentage of aligned bases; that is column *c* / summed length of sequences in column *a*. [e]the percentage of matched bases; that is total number of matched bases / summed length of sequences in column *a*. [f]average identity; that is total number of matched bases / summed length of aligned regions in the reference genome. [g]percentage of the reference genome covered by the aligned sequences.

Table S4: Quality of corrected long reads for Yeast dataset obtained with different methods. Assessment is done using alignments obtained from BWA-MEM.

			Aligned					
data set	Method	#Reads	#Reads	#Bases	Size (%)	Matched (%)	Identity (%)	Gen. cov.
Yeast	Original	231594	136943	742471424	52.94	47.05	92.79	99.69
Yeast	proovread	229702	223719	1216554868	88.78	83.79	95.86	99.75
(Full)	LoRDEC	231594	226827	1223763923	89.96	87.50	98.21	99.71
	CoLoRMap	231594	**228484**	**1240416300**	91.00	88.42	98.14	**99.71**
	CoLoRMap+OEA	231594	**228484**	1239535613	**91.03**	**88.66**	**98.27**	99.70
Yeast	LoRDEC	228893	**226632**	**1206108701**	91.46	89.25	98.40	**99.71**
(Trim)	CoLoRMap	211324	211206	1029575460	93.94	92.17	98.77	**99.71**
	CoLoRMap+OEA	211324	211206	1028609695	**93.98**	**92.47**	**98.93**	99.70
Yeast	proovread	225878	225670	245184675	**99.82**	**99.66**	**99.82**	60.64
(Split)	LoRDEC	1460179	925878	**1133579321**	97.90	97.41	99.52	**99.72**
	CoLoRMap	435140	434418	961260796	99.40	99.14	99.74	99.71
	CoLoRMap+OEA	349998	349421	973637656	99.37	99.07	99.72	99.70

Note: Please see Table S3 for description about each column.

Table S5: Quality of Canu assemblies for E.coli data set corrected by different methods. The assessment is done using QUAST. All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted.

Assembly	Original	LoRDEC	proovread	CoLoRMap	CoLoRMap+OEA
# contigs (≥ 0 bp)	182	24	26	**19**	**19**
# contigs (≥ 1000 bp)	182	24	26	**19**	**19**
# contigs (≥ 5000 bp)	178	24	26	**19**	**19**
# contigs (≥ 10000 bp)	141	24	26	**19**	**19**
# contigs (≥ 50000 bp)	4	21	22	**19**	**19**
Total length (≥ 0 bp)	3508197	4623137	**4629719**	4624793	4627249
Total length (≥ 1000 bp)	3508197	4623137	**4629719**	4624793	4627249
Total length (≥ 5000 bp)	3492249	4623137	**4629719**	4624793	4627249
Total length (≥ 10000 bp)	3209268	4623137	**4629719**	4624793	4627249
Total length (≥ 25000 bp)	1710292	4623137	4616507	4624793	**4627249**
Total length (≥ 50000 bp)	228498	4495150	4492555	4624793	**4627249**
Largest contig	69266	920903	605792	1089140	**1089205**
Reference length	4641652	4641652	4641652	4641652	4641652
GC (%)	51.05	50.81	50.81	50.81	50.81
Reference GC (%)	50.79	50.79	50.79	50.79	50.79
N50	24663	226456	231774	**239066**	**239066**
NG50	17847	226456	231774	**239066**	**239066**
L50	48	6	7	**5**	**5**
LG50	76	6	7	**5**	**5**
# unaligned contigs	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Unaligned length	0	0	0	0	0
Genome fraction (%)	75.455	99.120	99.092	**99.244**	99.231
Duplication ratio	1.002	1.005	1.007	**1.004**	1.005
Largest alignment	69266	538466	398061	**698643**	**698643**
NA50	24663	202095	198530	**239066**	**239066**
NGA50	17847	202095	198530	**239066**	**239066**
LA50	48	8	9	**6**	**6**
LGA50	76	8	9	**6**	**6**
# misassemblies	0	6	7	**5**	6
# relocations	0	6	7	**5**	6
# translocations	0	0	0	0	0
# inversions	0	0	0	0	0
# misassembled contigs	0	4	**3**	**3**	4
Misassembled contigs length	0	1328532	**1076559**	1277904	1446651
# local misassemblies	1	2	3	**1**	**1**
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	8.17	15.63	18.00	**6.64**	7.36
# indels per 100 kbp	191.04	3.43	2.02	1.80	**1.74**
Indels length	7249	222	126	99	**98**

Table S6: Quality of Canu assemblies for Yeast data set corrected by different methods. The assessment is done using QUAST. All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted.

Assembly	original	lordec	proovread	CoLoRMap	CoLoRMap+OEA
# contigs (≥ 0 bp)	26	28	32	**24**	29
# contigs (≥ 1000 bp)	26	28	32	**24**	29
# contigs (≥ 5000 bp)	26	28	31	**22**	28
# contigs (≥ 10000 bp)	26	27	30	**21**	28
# contigs (≥ 50000 bp)	22	**19**	24	**19**	20
Total length (≥ 0 bp)	12341981	**12497078**	12485995	12315869	12450479
Total length (≥ 1000 bp)	12341981	**12497078**	12485995	12315869	12450479
Total length (≥ 5000 bp)	12341981	**12497078**	12484209	12308283	12445656
Total length (≥ 10000 bp)	12341981	**12490996**	12474494	12302229	12445656
Total length (≥ 25000 bp)	12341981	12444116	**12456794**	12302229	12385648
Total length (≥ 50000 bp)	12218401	12257688	**12279045**	12239085	12217774
Largest contig	1543990	1552711	1537979	**1555857**	1538508
Reference length	12157105	12157105	12157105	12157105	12157105
GC (%)	38.18	38.21	38.22	38.17	38.20
Reference GC (%)	38.15	38.15	38.15	38.15	38.15
N50	777602	818962	777713	815158	**932935**
NG50	777602	818962	777713	815158	**932935**
L50	6	6	6	6	6
LG50	6	6	6	6	6
# unaligned contigs	1 + 1 part	**1 + 0 part**	**1 + 0 part**	**1 + 0 part**	**1 + 0 part**
Unaligned length	27953	**27982**	42350	34077	29118
Genome fraction (%)	98.638	98.791	98.687	98.716	**98.881**
Duplication ratio	1.027	1.038	1.037	**1.023**	1.033
Largest alignment	1084893	1073237	**1090741**	1073302	1085688
NA50	354598	377095	350112	**377108**	377106
NGA50	354598	377095	350112	**377108**	377106
LA50	11	11	11	11	11
LGA50	11	11	11	11	11
# misassemblies	107	124	108	**102**	112
# relocations	26	42	**29**	30	31
# translocations	79	82	79	**72**	80
# inversions	2	**0**	**0**	**0**	1
# misassembled contigs	21	25	24	**19**	24
Misassembled contigs length	10513374	12191557	**10639582**	11996690	10856637
# local misassemblies	31	**11**	14	**11**	12
# N's per 100 kbp	0.00	**0.00**	**0.00**	0.14	**0.00**
# mismatches per 100 kbp	75.76	89.07	96.59	87.75	**84.37**
# indels per 100 kbp	25.83	19.92	21.04	**13.64**	13.66
Indels length	6573	5899	6112	4901	**4627**

Table S7: Quality of Canu assemblies for D.melanogaster data set corrected by different methods. The assessment is done using QUAST. All statistics are based on contigs of size $\geq$ 500 bp, unless otherwise noted.

Assembly	original	lordec	CoLoRMap
# contigs ($\geq$ 0 bp)	217	**224**	260
# contigs ($\geq$ 1000 bp)	217	**224**	260
# contigs ($\geq$ 5000 bp)	159	**144**	161
# contigs ($\geq$ 10000 bp)	47	**33**	42
# contigs ($\geq$ 50000 bp)	0	**0**	2
Total length ($\geq$ 0 bp)	1768221	1730606	**2106055**
Total length ($\geq$ 1000 bp)	1768221	1730606	**2106055**
Total length ($\geq$ 5000 bp)	1543023	1410633	**1726065**
Total length ($\geq$ 10000 bp)	735933	653134	**910341**
Total length ($\geq$ 25000 bp)	58943	286003	**488439**
Total length ($\geq$ 50000 bp)	0	0	**142690**
Largest contig	30023	42661	**75766**
Reference length	137567484	137567484	137567484
GC (%)	38.17	37.92	38.22
Reference GC (%)	42.08	42.08	42.08
N50	8620	7664	**8485**
L50	64	**58**	**58**
# unaligned contigs	69 + 8 part	67 + 14 part	**61 + 17 part**
Unaligned length	770395	**861325**	986102
Genome fraction (%)	0.649	0.573	**0.764**
Duplication ratio	1.117	1.104	**1.066**
Largest alignment	16190	13571	**17993**
NA50	1442	-	**955**
NGA50	-	-	-
LA50	177	**-**	238
# misassemblies	175	**122**	138
# relocations	117	**73**	83
# translocations	58	**49**	54
# inversions	0	**0**	1
# misassembled contigs	67	**54**	73
Misassembled contigs length	562340	**358099**	478259
# local misassemblies	55	32	**21**
# N's per 100 kbp	0.00	0.00	0.00
# mismatches per 100 kbp	679.35	704.35	**583.04**
# indels per 100 kbp	401.99	273.08	**191.33**
Indels length	9235	**6132**	7931