

## Supplementary Materials

### Supplementary Tables

**Table S1.** Data sets information

Data set	Accession	Basecaller	Coverage
Escherichia coli K12 - ONT	<a href="http://lab.loman.net/2016/07/30/nanopore-r9-data-release/">http://lab.loman.net/2016/07/30/nanopore-r9-data-release/</a>	Guppy	319X
Escherichia coli str. K-12 substr. MG1655 - Ground Truth	GenBank NC_0913		
Saccharomyces cerevisiae w303 - ONT	<a href="http://www.tgsbioinformatics.com/necat/">http://www.tgsbioinformatics.com/necat/</a>	Albacore	512X
Saccharomyces cerevisiae S288c - Ground Truth	GenBank GCA_00146045.2		
NA12878 Chromosome 21 - ONT	<a href="https://github.com/nanopore-wgs-consortium/NA12878/">https://github.com/nanopore-wgs-consortium/NA12878/</a>	Albacore	35X
GrCH38-with-decoys - Ground Truth	<a href="https://github.com/nanopore-wgs-consortium/NA12878/">https://github.com/nanopore-wgs-consortium/NA12878/</a>		
NA12878 Chromosome 20 - ONT	<a href="https://github.com/nanopore-wgs-consortium/NA12878/">https://github.com/nanopore-wgs-consortium/NA12878/</a>	Albacore	35X
GrCH38-with-decoys - Ground Truth	<a href="https://github.com/nanopore-wgs-consortium/NA12878/">https://github.com/nanopore-wgs-consortium/NA12878/</a>		
Arabidopsis thaliana - ONT	ENA PRJEB21270	Albacore	30X
Arabidopsis thaliana assembly TAIR10.1 - Ground Truth	GenBank GCA_00001735.2		

**Table S2.** Comparison of the error rates of polishing results by Racon, Medaka, MarginPolish, HELEN and NeuralPolish on the data assembled by Canu.

dataset	tools	mismatch(%)	deletion(%)	insertion(%)	total error rate(%)
<i>E. coli</i>	Canu	0.10	0.31	<b>0.02</b>	0.43
	Racon	0.11	0.18	0.13	0.42
	Medaka	0.14	<b>0.05</b>	0.09	0.28
	MarginPolish	0.12	0.11	0.13	0.36
	HELEN	0.13	0.21	0.11	0.45
	NeuralPolish	<b>0.06</b>	0.11	0.07	<b>0.24</b>
<i>S. cerevisiae</i>	Canu	<b>0.10</b>	1.02	<b>0.34</b>	1.46
	Racon	0.12	0.71	0.52	1.35
	Medaka	0.11	<b>0.51</b>	0.40	<b>1.02</b>
	MarginPolish	0.14	0.64	0.50	1.27
	HELEN	0.17	0.86	0.45	1.48
	NeuralPolish	0.12	0.62	0.64	1.38
<i>NA12878 chr21</i>	Canu	<b>0.27</b>	4.83	<b>0.17</b>	5.27
	Racon	0.30	2.91	0.29	3.50
	Medaka	0.36	2.55	0.26	3.17
	MarginPolish	0.32	2.87	0.27	3.46
	HELEN	1.55	9.14	0.83	11.52
	NeuralPolish	0.32	<b>1.61</b>	0.45	<b>2.38</b>
<i>NA12878 chr20</i>	Canu	0.30	5.17	<b>0.16</b>	5.63
	Racon	<b>0.29</b>	3.00	0.25	3.54
	Medaka	0.37	2.63	0.23	3.23
	MarginPolish	0.32	3.03	0.25	3.59
	HELEN	1.66	9.67	0.85	12.18
	NeuralPolish	0.34	<b>1.73</b>	0.37	<b>2.44</b>
<i>A. thaliana</i>	Canu	<b>0.62</b>	5.35	<b>1.65</b>	7.62
	Racon	0.65	3.79	1.86	6.30
	Medaka	0.66	3.47	1.88	6.01
	MarginPolish	0.68	4.00	1.79	6.47
	HELEN	1.06	5.21	1.89	8.16
	NeuralPolish	0.69	<b>3.09</b>	2.10	<b>5.88</b>

**Table S3.** Quast evaluation of polishing results by Racon, Medaka, MarginPolish, HELEN and NeuralPolish on the data assembled by Flye.

dataset	tools	N50 (Mb)	mismatches per 100kbp	indels per 100kbp	number of indels>5bp	genome fraction(%)	Identity (%)
<i>E. coli</i>	Flye	4.6	182.31	130.32	<b>0</b>	<b>99.998</b>	99.69
	Racon	4.6	109.86	204.66	49	99.994	99.62
	Medaka	4.6	135.11	<b>111.84</b>	1	99.994	99.76
	MarginPolish	4.6	106.36	147.44	5	99.959	99.73
	HELEN	4.6	93.82	158.56	16	99.959	99.71
	NeuralPolish	4.6	<b>67.56</b>	126.08	2	99.965	<b>99.79</b>
<i>S. cerevisiae</i>	Flye	0.8	95.68	274.38	313	93.045	99.65
	Racon	0.8	141.06	348.03	503	92.923	99.49
	Medaka	0.8	99.12	134.25	<b>279</b>	93.135	99.83
	MarginPolish	0.8	<b>87.31</b>	<b>124.59</b>	368	<b>94.361</b>	<b>99.84</b>
	HELEN	0.8	87.45	178.4	1403	94.344	99.67
	NeuralPolish	0.8	134.82	355.15	305	94.148	99.54
<i>NAI2878</i> <i>chr21</i>	Flye	3.7	428.42	1890.74	13425	87.289	97.12
	Racon	3.7	372.63	1852.24	13057	87.751	97.15
	Medaka	3.7	<b>343.77</b>	1565.97	12770	<b>88.378</b>	97.49
	MarginPolish*	3.7	293.86	1328.66	12995	1.149	97.77
	HELEN*	3.1	885.46	707.79	120	0.032	97.77
	NeuralPolish	3.7	359.56	<b>1004.22</b>	<b>10102</b>	88.139	<b>98.31</b>
<i>NAI2878</i> <i>chr20</i>	Flye	8.7	322.12	1995.33	24798	92.866	96.96
	Racon	8.7	<b>295.16</b>	1912.53	23895	92.993	97.02
	Medaka	8.8	359.29	1636.96	23803	93.88	97.36
	MarginPolish*	8.8	275.56	1388.23	24753	1.979	97.68
	HELEN*	7.1	917.36	1938.98	99	0.008	96.18
	NeuralPolish	8.9	301.35	<b>1033.24</b>	<b>18685</b>	<b>94.387</b>	<b>98.23</b>
<i>A. thaliana</i>	Flye	13.2	1332.57	1448.7	45508	88.633	97.18
	Racon	13.2	1249.91	1287.21	45548	89.158	97.33
	Medaka	13.2	1113.78	1070.79	45243	90.173	97.62
	MarginPolish	13.2	<b>1062.12</b>	1096.03	46309	89.784	97.68
	HELEN	13	2576.23	1492.56	53361	78.047	96.8
	NeuralPolish	13.3	1052.42	<b>930.5</b>	<b>39247</b>	<b>90.408</b>	<b>97.92</b>

The results marked with \* have no comparative meaning. Because the indicators are calculated according to the contig sequences which align to the reference, and there are very few contig sequences aligning to the reference in these results.

**Table S4.** Quast evaluation of polishing results by Racon, Medaka, MarginPolish, HELEN and NeuralPolish on the data assembled by Wtdbg2.

dataset	tools	N50 (Mb)	mismatches per 100kbp	indels per 100kbp	number of indels>5bp	genome fraction(%)	Identity (%)
<i>E. coli</i>	Wtdbg2	4.6	66.85	296.06	1	99.996	99.62
	Racon	4.6	104.95	213.86	31	99.995	99.61
	Medaka	4.6	132.33	<b>112.16</b>	<b>0</b>	99.995	99.75
	MarginPolish	4.6	101.56	149.73	6	<b>99.998</b>	99.73
	HELEN	4.6	90.90	158.89	13	<b>99.998</b>	99.71
	NeuralPolish	4.6	<b>57.29</b>	132.24	10	99.997	<b>99.78</b>
<i>S. cerevisiae</i>	Wtdbg2	0.8	115.17	514.81	397	97.325	99.34
	Racon	0.8	149.64	351.62	616	97.167	99.49
	Medaka	0.8	102.46	134.6	<b>330</b>	97.409	<b>99.83</b>
	MarginPolish	0.8	<b>91.28</b>	<b>128.05</b>	435	97.383	<b>99.83</b>
	HELEN	0.8	94.45	184.28	1491	<b>97.448</b>	99.66
	NeuralPolish	0.8	114.33	360.79	353	97.405	99.55
<i>NAI2878</i> <i>chr21</i>	Wtdbg2	4.3	<b>200.30</b>	2942.30	7759	52.351	96.42
	Racon	4.3	907.78	1706.41	13973	85.830	97.14
	Medaka	4.4	782.87	1469.07	13574	86.347	97.43
	MarginPolish	4.3	928.22	1408.48	15511	84.998	97.34
	HELEN*	4.1	864.27	1658.23	250	6.286	96.67
	NeuralPolish	4.4	552.09	<b>961.26</b>	<b>10570</b>	<b>86.798</b>	<b>98.23</b>
<i>NAI2878</i> <i>chr20</i>	Wtdbg2	0.5	700.11	2317.86	11208	55.510	96.37
	Racon	0.5	<b>318.53</b>	1853.13	23915	92.477	97.04
	Medaka	14	356.22	1679.48	23723	92.566	97.31
	MarginPolish	0.5	953.89	1479.41	28622	90.585	97.15
	HELEN*	0.5	768.34	1689.86	239	5.440	96.52
	NeuralPolish	0.5	556.65	<b>1024.25</b>	<b>20308</b>	<b>93.574</b>	<b>98.12</b>
<i>A. thaliana</i>	Wtdbg2	9.8	1545.41	1965.26	29724	55.628	96.61
	Racon	10.0	1094.28	1200.97	35032	72.762	97.47
	Medaka	10.0	<b>1020.58</b>	979.02	34228	73.754	97.75
	MarginPolish	10.0	1263.31	1253.93	38301	68.692	97.21
	HELEN	9.7	731	1598.33	18190	68.692	96.27
	NeuralPolish	10.0	1105.95	<b>849.23</b>	<b>29575</b>	<b>74.009</b>	<b>97.91</b>

The results marked with \* have no comparative meaning. Because the indicators are calculated according to the contig sequences which align to the reference, and there are very few contig sequences aligning to the reference in these results.

**Table S5.** Quast evaluation of polishing results by Racon, Medaka, MarginPolish, HELEN and NeuralPolish on the data assembled by Canu.

dataset	tools	N50 (Mb)	mismatches per 100kbp	indels per 100kbp	number of indels>5bp	genome fraction(%)	Identity (%)
<i>E. coli</i>	Canu	4.7	99.84	280.25	4	99.998	99.60
	Racon	4.7	110.35	213.51	57	99.998	99.62
	Medaka	4.7	136.68	<b>116</b>	<b>3</b>	99.998	99.75
	MarginPolish	4.7	112.83	169.19	20	99.998	99.72
	HELEN	4.7	112.66	191.14	24	99.976	99.72
	NeuralPolish	4.7	<b>64.76</b>	133.86	7	99.998	<b>99.78</b>
<i>S. cerevisiae</i>	Canu	0.8	116.93	504.53	732	99.153	99.39
	Racon	0.8	164.12	386.36	717	99.044	99.49
	Medaka	0.8	119.26	156.87	<b>435</b>	99.133	<b>99.83</b>
	MarginPolish	0.8	122.6	<b>153.93</b>	564	98.914	<b>99.83</b>
	HELEN	0.8	<b>113.06</b>	200.81	1536	98.166	99.67
	NeuralPolish	0.8	147.55	404.23	496	<b>99.160</b>	99.55
<i>NAI2878</i> <i>chr21</i>	Canu	1.1	969.55	2149.83	14304	65.860	96.18
	Racon	1.1	<b>288.32</b>	1882.49	12441	85.361	97.15
	Medaka	1.1	338.33	1608.22	12271	85.906	97.45
	MarginPolish	1.1	360.27	1515.52	15659	84.932	97.31
	HELEN*	1.0	1022.92	264.87	169	6.419	96.71
	NeuralPolish	1.1	327.17	<b>1056.99</b>	<b>9995</b>	<b>86.252</b>	<b>98.22</b>
<i>NAI2878</i> <i>chr20</i>	Canu	0.6	954.41	1973.19	17922	57.831	96.10
	Racon	0.6	<b>265.66</b>	1951.28	21889	88.686	97.06
	Medaka	0.6	322.76	1676.24	21530	89.399	97.35
	MarginPolish	0.6	364.57	1589.62	27824	87.557	97.14
	HELEN*	0.6	997.66	32.98	41	5.159	96.42
	NeuralPolish	0.6	304.71	<b>1092.18</b>	<b>17467</b>	<b>89.527</b>	<b>98.15</b>
<i>A. thaliana</i>	Canu	4.0	1367.31	1787.24	43584	63.020	96.48
	Racon	4.0	1099.75	1204.75	34693	71.429	97.48
	Medaka	4.1	<b>1018.12</b>	979.14	33814	72.332	97.75
	MarginPolish	4.0	1082.91	1113.64	39469	70.450	97.41
	HELEN*	4.0	1511.36	1406.54	37874	56.094	96.64
	NeuralPolish	4.1	1079.65	<b>854.46</b>	<b>29608</b>	<b>72.910</b>	<b>97.91</b>

The results marked with \* have no comparative meaning. Because the indicators are calculated according to the contig sequences which align to the reference, and there are very few contig sequences aligning to the reference in these results.

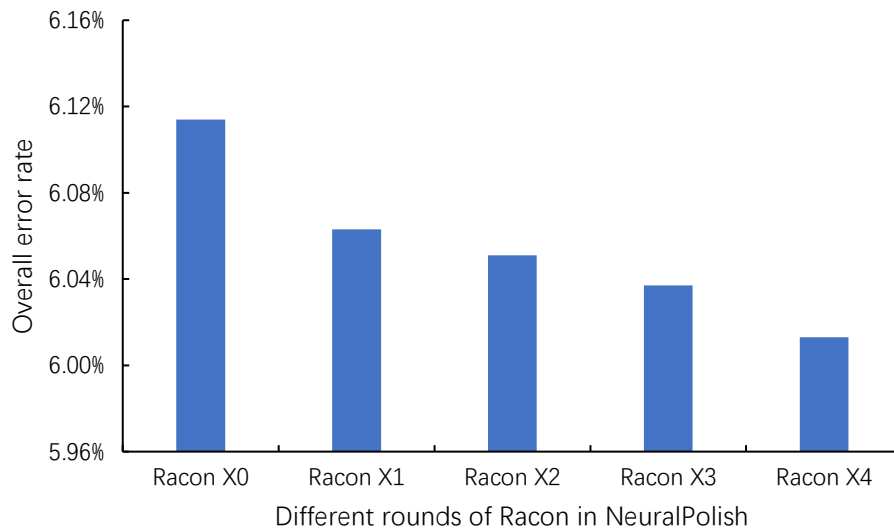
**Table S6.** Iteration polishing of NeuralPolish on the draft assembly from *E. coli*, *S. cerevisiae*, chromosome 21 of NA12878, chromosome 20 of NA12878 and *A. thaliana* dataset.

Dataset	Rounds	Mismatch (%)	Deletion (%)	Insertion (%)	overall error rate (%)
<i>E. coli</i>	NeuralPolish x1	0.07%	0.12%	0.07%	0.26%
	NeuralPolish x2	0.06%	0.12%	0.07%	<b>0.25%</b>
	NeuralPolish x3	0.06%	0.12%	0.07%	0.25%
	NeuralPolish x4	0.06%	0.13%	0.07%	0.26%
<i>S. cerevisiae</i>	NeuralPolish x1	0.11%	0.71%	0.56%	<b>1.37%</b>
	NeuralPolish x2	0.11%	0.71%	0.58%	1.40%
	NeuralPolish x3	0.11%	0.71%	0.57%	1.38%
	NeuralPolish x4	0.11%	0.71%	0.57%	1.38%
NA12878 chr21	NeuralPolish x1	0.44%	1.59%	0.57%	2.60%
	NeuralPolish x2	0.42%	1.55%	0.59%	2.55%
	NeuralPolish x3	0.42%	1.53%	0.58%	<b>2.53%</b>
	NeuralPolish x4	0.42%	1.54%	0.58%	2.54%
NA12878 chr20	NeuralPolish x1	0.33%	1.60%	0.51%	2.43%
	NeuralPolish x2	0.33%	1.57%	0.53%	2.43%
	NeuralPolish x3	0.33%	1.56%	0.54%	2.43%
	NeuralPolish x4	0.32%	1.55%	0.53%	<b>2.41%</b>
<i>A. thaliana</i>	NeuralPolish x1	0.73%	2.98%	2.36%	6.06%
	NeuralPolish x2	0.72%	2.93%	2.36%	6.01%
	NeuralPolish x3	0.72%	2.96%	2.35%	6.03%
	NeuralPolish x4	0.72%	2.90%	2.34%	<b>5.95%</b>

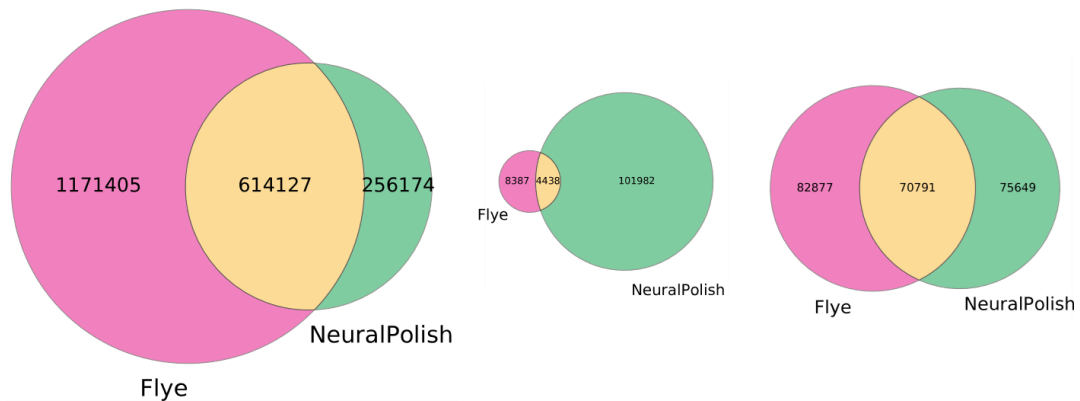
**Table S7.** The runtime of each polishing methods on the test data sets.

Datasets	Polishing tools	CPU runtimes(sec)	GPU runtimes(sec)	Total runtimes(sec)
<i>E. coli</i>	Racon	898	0	898
	Medaka	898	326	1224
	MarginPolish	954	0	954
	HELEN	954	108	1062
	NeuralPolish	574	213	<b>787</b>
<i>S. cerevisiae</i>	Racon	4121	0	4121
	Medaka	4121	692	4813
	MarginPolish	4595	0	4595
	HELEN	4595	233	4828
	NeuralPolish	2031	500	<b>2531</b>
<i>NA12878 chr21</i>	Racon	326	0	<b>326</b>
	Medaka	326	579	905
	MarginPolish	1408	0	1408
	HELEN	1408	182	1590
	NeuralPolish	420	1538	1958
<i>NA12878 chr20</i>	Racon	702	0	<b>702</b>
	Medaka	702	903	1605
	MarginPolish	2333	0	2333
	HELEN	2333	230	2563
	NeuralPolish	984	2560	3544
<i>A. thaliana</i>	Racon	1235	0	<b>1235</b>
	Medaka	1235	1905	3140
	MarginPolish	12066	0	12066
	HELEN	12066	1194	13260
	NeuralPolish	1777	5757	7534

## Supplementary Figures

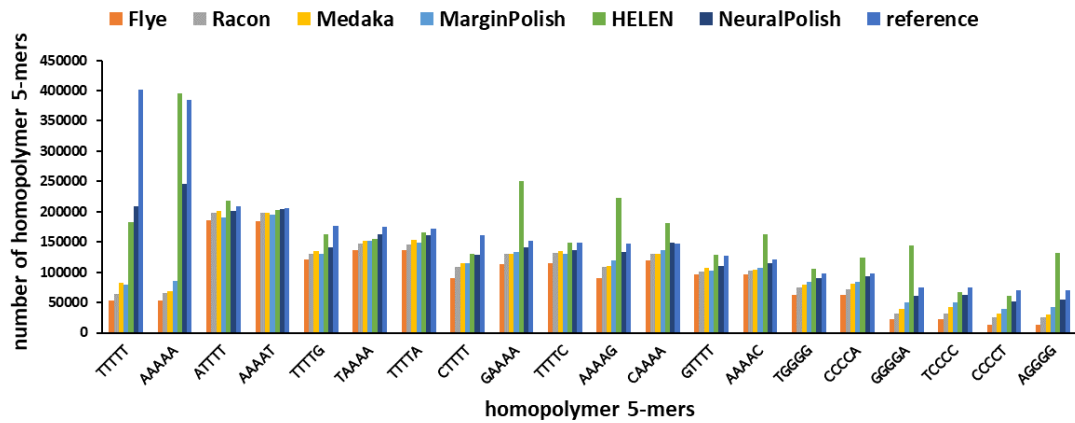


**Fig. S1.** Evaluation of different rounds of Racon in NeuralPolish. The draft assembly is assembled from *A. thaliana* by Flye. We perform different rounds of Racon to polish the draft assembly initially and then obtain the final polished result by NeuralPolish. The NeuralPolish model is trained on the dataset assembled by Wtdbg2. NeuralPolish with one round of Racon can significantly reduce the error rate of the polishing result than NeuralPolish without Racon. More rounds of Racon in NeuralPolish can further reduce the error rate, but they also bring more computational expense.

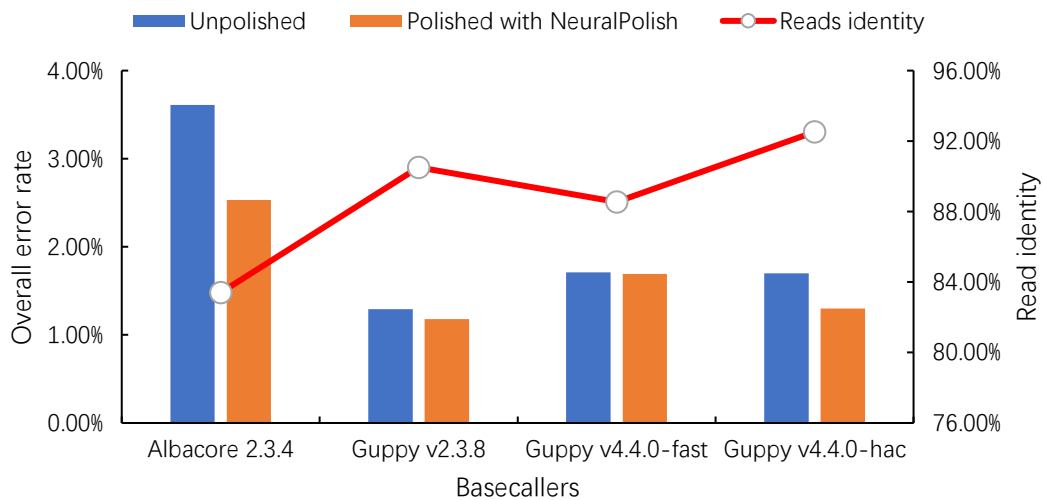


**Fig. S2.** Comparing the draft assembly and polished assembly of chromosome 20 of NA12878 in terms of the error sites. The left one presents the deletion errors in the draft assembly and polished assembly. The middle one presents the insertion errors and the right one shows the substitution errors.





**Fig. S3.** Comparison of the homopolymer 5-mers counts in the polished genome assembly and reference genome on the dataset of chromosome 20 of NA12878. The draft assembly is produced by assembler Flye and then polished by Racon, Medaka, MarginPolish, HELEN and NeuralPolish, respectively. The homopolymer 5-mers consists of a single base (*e.g.* AAAAA) or 4-base stretch of a single base (*e.g.* AAAAT) are picked. The 5-mers are sorted by the counts in the reference genome.



**Fig. S4.** The error rates of polished assemblies by NeuralPolish on the datasets with varying degree of errors. The read sets were basecalled by several basecallers and the assemblies were assembled using Flye. The red polyline shows the identities of reads sets.