Repeat and haplotype aware error correction in nanopore sequencing reads with DeChat – Supplementary Material

Yuansheng Liu[1,†] Yichen Li[1,†], Enlian Chen[2], Jialu Xu[2], Wenhai Zhang[2], Xiangxiang Zeng[1], Xiao Luo[2,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2] College of Biology, Hunan University, Changsha, China
[†]These authors contributed equally to the work.
[*]To whom correspondence should be addressed.
Email: xluo@hnu.edu.cn

Supplementary Tables and Figures

Method	#Reads	Haplotype coverage (%)	Mismatch per 10 kbp	Non-Hp Ins per 10 kbp	Non-Hp Del per 10 kbp	Hp-Ins per 10 kbp	Hp-Del per 10 kbp	Errors per 10 kbp
coverage=10X								
DeChat	19548	**100.0**	**0.356**	**0.108**	**0.083**	**0.134**	**0.076**	**0.757**
VeChat	19118	99.8	0.983	0.307	1.287	0.234	1.131	3.942
Racon	19546	96.2	41.396	14.566	10.855	0.716	9.837	77.371
Herro	2263	85.6	8.600	1.763	4.661	3.668	4.004	22.696
Canu	19454	99.0	45.125	8.018	19.860	1.632	17.163	91.798
Daccord	**19598**	97.3	59.220	14.338	37.390	0.431	32.127	143.507
hifiasm	19548	**100.0**	29.756	25.197	21.148	33.298	15.528	124.928
Raw	19548	100.0	50.972	43.910	39.329	63.306	29.396	226.913
coverage=20X								
DeChat	**38628**	**100.0**	**0.149**	0.096	**0.049**	0.065	**0.036**	**0.395**
VeChat	37799	**100.0**	0.295	**0.073**	0.808	**0.056**	0.698	1.930
Canu	37819	99.9	43.355	7.466	16.487	2.062	14.238	83.608
Herro	4392	99.1	1.940	0.580	1.150	1.633	0.994	6.296
Racon	38561	99.5	41.300	12.500	10.381	0.672	8.988	73.841
Daccord	38623	99.5	56.444	15.169	43.145	0.315	37.312	152.385
hifiasm	**38628**	**100.0**	24.438	20.772	17.047	26.517	12.454	101.228
Raw	38628	100.0	50.508	43.761	39.155	62.950	29.246	225.619
coverage=30X								
DeChat	**57908**	**100.0**	0.231	0.116	**0.073**	0.101	**0.059**	**0.580**
VeChat	56657	**100.0**	**0.226**	**0.060**	0.838	**0.045**	0.712	1.882
Herro	6554	99.9	1.167	0.415	0.814	1.111	0.689	4.197
Canu	38410	**100.0**	36.169	5.419	10.962	0.388	9.604	62.543
Racon	57730	99.5	38.163	10.703	10.097	0.644	8.915	68.522
Daccord	50746	99.8	47.660	10.475	30.055	0.245	26.690	115.126
hifiasm	**57908**	**100.0**	23.286	19.761	16.091	25.012	11.800	95.951
Raw	57908	100.0	50.826	43.895	39.196	63.023	29.402	226.342
coverage=40X								
DeChat	**77484**	**100.0**	0.288	0.147	**0.095**	0.134	**0.078**	**0.741**
VeChat	75754	**100.0**	**0.133**	**0.028**	0.994	**0.027**	0.853	2.035
Herro	8670	99.9	0.734	0.269	0.519	0.738	0.442	2.702
Canu	47111	**100.0**	27.488	1.538	6.096	0.294	5.333	40.749
Racon	77244	99.6	40.193	10.116	12.175	0.751	10.456	73.692
Daccord	50961	99.5	53.289	11.405	36.977	0.246	32.531	134.447
hifiasm	**77484**	**100.0**	22.328	18.887	15.433	23.943	11.238	91.828
Raw	77484	100.0	50.613	43.712	39.120	62.862	29.237	225.544
coverage=50X								
DeChat	**96350**	**100.0**	0.354	0.172	**0.112**	0.159	**0.091**	**0.889**
VeChat	94204	**100.0**	**0.120**	**0.029**	1.023	**0.027**	0.859	2.059
Herro	10841	99.9	0.696	0.264	0.504	0.724	0.441	2.629
Canu	51748	**100.0**	16.413	1.005	1.518	0.208	1.364	20.508
Racon	**96035**	99.6	38.435	11.299	10.841	0.781	9.354	70.710
Daccord	49684	98.0	51.242	18.176	47.063	0.274	41.561	158.317
hifiasm	**96350**	**100.0**	22.037	18.636	15.109	23.317	11.025	90.125
Raw	96350	100.0	51.070	44.097	39.394	63.224	29.475	227.259

Supplementary Table 1. Error correction benchmarking results for simulated ONT simplex reads of various sequencing coverages. The average sequencing coverage per haplotype is set as 10x, 20x, 30x, 40x, and 50x, respectively. The ploidy is two and sequencing error rate is approximately 2%. The genomes used in these coverage experiments are the same as those in the "sim-diploid-ecoli" dataset.

Method	#Reads	Haplotype coverage (%)	Mismatch per 10 kbp	Non-Hp Ins per 10 kbp	Non-Hp Del per 10 kbp	Hp-Ins per 10 kbp	Hp-Del per 10 kbp	Errors per 10 kbp
coverage=10X								
DeChat	3047429	99.9	8.660	5.526	4.151	5.832	4.072	28.241
Raw	3047429	100.0	51.514	45.830	35.660	61.041	32.631	226.676
coverage=20X								
DeChat	6110105	100.0	8.054	5.562	4.267	5.883	4.164	27.929
Raw	6110105	100.0	51.697	45.977	35.784	61.265	32.749	227.472
coverage=30X								
DeChat	9117385	100.0	7.943	5.633	4.342	5.965	4.254	28.137
Raw	9117385	100.0	51.284	45.616	35.519	60.814	32.510	225.743
coverage=40X								
DeChat	12136092	100.0	8.187	5.888	4.604	6.284	4.508	29.470
Raw	12136092	100.0	51.151	45.497	35.400	60.660	32.405	225.114
coverage=50X								
DeChat	15179267	100.0	8.504	6.173	4.843	6.633	4.754	30.906
Raw	15179267	100.0	51.171	45.518	35.445	60.687	32.450	225.271

Supplementary Table 2. Error correction benchmarking results for simulated ONT simplex reads of various sequencing coverages(complex potato genome). The average sequencing coverage per haplotype is set as 10x, 20x, 30x, 40x, and 50x, respectively. The ploidy is two and sequencing error rate is approximately 2%. The genomes used in these coverage experiments were sim-diploid-potato.

Method	#Reads	Haplotype coverage (%)	Mismatch per 10 kbp	Non-Hp Ins per 10 kbp	Non-Hp Del per 10 kbp	Hp-Ins per 10 kbp	Hp-Del per 10 kbp	Errors per 10 kbp
ER=1%								
DeChat	10915	**100.0**	0.168	0.150	**0.054**	**0.060**	**0.037**	**0.468**
VeChat	10760	99.8	**0.158**	**0.132**	0.206	0.086	0.168	0.751
Canu	10870	99.9	1.253	0.621	0.725	0.572	0.518	3.689
Racon	10913	99.9	0.327	0.224	0.434	0.092	0.357	1.434
hifiasm	10915	99.9	6.641	5.695	4.045	6.031	2.887	25.299
Daccord	**11062**	99.7	8.519	0.533	0.517	0.119	0.394	10.082
Raw	10915	99.9	22.153	19.369	16.957	27.307	12.763	98.549
ER=2%								
DeChat	10624	**100.0**	0.901	0.778	**0.307**	0.333	**0.197**	**2.517**
VeChat	10308	99.9	**0.692**	**0.567**	0.896	0.449	0.700	3.304
Canu	10549	99.9	2.459	1.653	1.654	1.518	1.199	8.484
Racon	10622	99.9	0.942	0.731	0.968	0.442	0.765	3.848
hifiasm	10624	99.9	42.334	36.089	26.682	40.212	19.174	164.491
Daccord	**10810**	99.8	9.897	0.571	0.709	**0.162**	0.544	11.882
Raw	10624	99.9	66.081	56.678	45.230	70.156	33.135	271.281
ER=3%								
DeChat	9443	**99.9**	**1.046**	0.904	**0.360**	0.421	**0.241**	**2.972**
VeChat	9131	99.7	1.089	0.917	0.989	0.749	0.752	4.496
Canu	9370	**99.9**	3.070	2.102	2.149	1.901	1.560	10.782
Racon	9440	**99.9**	1.252	1.079	1.113	0.712	0.830	4.986
hifiasm	9443	**99.9**	71.484	61.142	47.119	72.050	34.369	286.164
Daccord	**9634**	99.7	8.787	**0.567**	0.440	**0.181**	0.327	10.303
Raw	9443	99.9	78.372	67.046	52.613	80.931	38.517	317.479
ER=4%								
DeChat	8433	**99.9**	**1.830**	1.615	0.658	0.752	0.450	**5.304**
VeChat	8116	99.8	2.832	2.451	2.364	2.000	1.815	11.462
Canu	8367	**99.9**	3.993	2.862	2.651	2.481	1.910	13.897
Racon	8432	**99.9**	2.668	2.343	1.707	1.633	1.277	9.628
hifiasm	8433	**99.9**	105.954	89.976	68.394	103.345	49.499	417.167
Daccord	**8691**	99.8	6.921	**0.391**	**0.516**	**0.189**	**0.412**	8.428
Raw	8433	99.9	106.954	90.853	69.252	104.759	50.165	421.984

Supplementary Table 3. Error correction benchmarking results for simulated ONT simplex reads of various sequencing error rate. The average sequencing error rate is approximately set as 1%,2%,3%,4%, respectively. The ploidy is two and sequencing coverage is 10X per haplotype. The genomes used in these coverage experiments are the same as those in the "sim-diploid-ecoli" dataset. ER = sequencing error rate.

Method	#Reads	Haplotype coverage (%)	Mismatch per 10 kbp	Non-Hp Ins per 10 kbp	Non-Hp Del per 10 kbp	Hp-Ins per 10 kbp	Hp-Del per 10 kbp	Errors per 10 kbp
Average length=1319bp								
DeChat	166096	**100.0**	**0.350**	**0.112**	**0.090**	0.141	**0.083**	**0.777**
VeChat	102546	99.9	2.434	0.123	10.170	**0.134**	9.098	21.959
LoRMA	165438	**100.0**	2.645	1.328	0.746	1.976	0.578	7.274
Canu	99053	99.1	21.321	1.605	3.871	0.275	3.342	30.414
Racon	161819	95.5	26.854	3.593	15.861	0.545	13.630	60.483
Daccord	105640	95.7	23.948	1.444	5.851	0.165	4.953	36.361
hifiasm	**166277**	**100.0**	16.879	14.287	11.722	18.272	8.575	69.735
Raw	166509	100.0	52.747	44.534	40.397	64.466	30.295	232.440
Average length=6955bp								
DeChat	**43513**	**100.0**	**0.230**	0.127	**0.103**	0.124	**0.087**	**0.671**
VeChat	43508	**100.0**	1.427	**0.111**	3.770	**0.065**	3.397	8.770
LoRMA	43250	**100.0**	1.801	1.280	0.893	1.979	0.687	6.641
Canu	43497	99.4	16.059	3.950	8.189	0.252	6.999	35.449
Racon	**43513**	93.6	26.851	10.705	13.164	0.555	11.273	62.547
Daccord	36835	95.5	32.153	6.017	36.451	0.183	31.008	105.812
hifiasm	**43513**	**100.0**	21.032	17.811	14.650	22.766	10.704	86.962
Raw	43513	100.0	51.441	44.555	39.791	64.057	29.818	229.662
Average length=19308bp								
DeChat	**15674**	**100.0**	**0.251**	0.153	**0.105**	0.170	**0.099**	**0.779**
VeChat	**15674**	**100.0**	0.442	**0.048**	0.816	**0.021**	0.717	2.043
Canu	15179	**100.0**	18.228	11.704	13.179	0.357	11.457	54.925
Racon	**15674**	99.3	20.012	20.036	11.317	0.497	9.734	61.596
Daccord	13352	99.8	41.853	17.603	47.797	0.264	41.810	149.326
hifiasm	**15674**	**100.0**	21.629	18.323	14.898	23.067	10.917	88.834
Raw	15674	100.0	49.041	42.623	37.956	61.202	28.537	219.359
Average length=60953bp								
DeChat	**4966**	**99.9**	0.360	0.261	**0.187**	0.324	0.179	1.311
VeChat	**4966**	**99.9**	**0.150**	**0.057**	0.215	**0.002**	**0.174**	**0.598**
Canu	4965	99.8	18.100	20.649	24.001	0.458	21.807	85.016
Racon	**4966**	99.2	8.769	14.075	5.658	0.321	4.883	33.706
Daccord	4297	**99.9**	46.600	24.430	43.982	0.304	38.896	154.214
hifiasm	**4966**	**99.9**	24.723	20.906	16.919	25.954	12.399	100.900
Raw	4966	99.9	51.434	44.653	39.442	63.382	29.624	228.535

Supplementary Table 4. Error correction benchmarking results for simulated ONT simplex reads of various read length. The average sequencing length per haplotype is set as 1319bp, 6955bp, 19308bp,and 60953bp, respectively. The ploidy is two and sequencing error rate is approximately 2%. The genomes used in these read length experiments are the same as those in the "sim-diploid-ecoli" dataset.

Method	Assembler	#Contigs	Haplotype coverage (%)	N50 (bp)	NGA50 (bp)	Error rate (%)	Mismatch (%)	Indel (%)	# Misassemblies
sim-diploid-ecoli									
DeChat	hifiasm	40	99.1	537398	537398	0.008	0.004	0.004	0
Raw	hifiasm	111	92.4	139149	127606	0.291	0.087	0.204	8
sim-triploid-ecoli									
DeChat	hifiasm	68	69.1	1946151	1610698	0.008	0.004	0.004	1
Raw	hifiasm	86	66.0	277865	128913	0.406	0.110	0.296	1
sim-tetraploid-ecoli									
DeChat	hifiasm	117	81.8	546201	546200	0.006	0.003	0.003	0
Raw	hifiasm	143	74.9	261373	194410	0.286	0.082	0.204	4

Supplementary Table 5. Assembly benchmarking results for *E.coli* genomes of various ploidy (ploidy=2,3,4) using hifiasm. The average sequencing coverage per haplotype is 10x, and the sequencing error rate is approximately 2%. The error rate is equal to the sum of mismatch and indel rate. The results are sorted by the error rate in ascending order.

Method	Assembler	#Contigs	Haplotype coverage (%)	N50 (bp)	NGA50 (bp)	Error rate (%)	Mismatch (%)	Indel (%)	# Misassemblies
sim-tetraploid-potato									
DeChat	hifiasm	17283	73.4	417853	276851	0.098	0.038	0.060	2304
Raw	hifiasm	21307	63.8	238033	136318	0.230	0.065	0.165	1971
sim-haploid-potato									
DeChat	hifiasm	5528	96.8	401207	570351	0.072	0.031	0.042	332
Raw	hifiasm	9366	84.9	123936	123543	0.329	0.086	0.243	365

Supplementary Table 6. Assembly benchmarking results for simulated ONT reads of haploid and tetraploid potato genomes using hifiasm.The average sequencing coverage per haplotype for potato is 10x with a sequencing error rate of approximately 2%. The error rate is equal to the sum of mismatch and indel rate. The results are sorted by the error rate in ascending order. Note that N50 is lower than NGA50 in some hifiasm(-meta)'s assemblies, which means the total contig length is longer than the reference haplotype length. This is likely because that hifiasm(-meta) tends to produce larger assemblies containing duplicated sequences when assembling error-corrected simplex ONT reads.

Method	Assembler	#Contigs	Haplotype coverage (%)	N50 (bp)	NGA50 (bp)	Error rate (%)	Mismatch (%)	Indel (%)	# Misassemblies
sim-meta-low									
DeChat	hifiasmmeta	1767	59.2	274359	71431	0.173	0.109	0.064	124
Raw	hifiasmmeta	1603	39.2	127206	-	0.366	0.108	0.258	64
sim-meta-high									
DeChat	hifiasmmeta	29792	37.6	72700	-	0.451	0.266	0.185	3155
Raw	hifiasmmeta	17517	17.9	53692	-	0.900	0.257	0.643	1084

Supplementary Table 7. Assembly benchmarking results for sim-meta-low (60 strains) and sim-meta-high (1000 strains) using hifiasm-meta. The error rate is equal to the sum of mismatch and indel rate. The results are sorted by the error rate in ascending order. Note that NGA50 is missing if the "Haplotype coverage" is less than 50% (the length of all aligned blocks does not even reach to a half of the reference size).

Method	Assembler	#Contigs	Haplotype coverage (%)	N50 (bp)	NGA50 (bp)	Error rate (%)	Mismatch (%)	Indel (%)	# Misassemblies
real-D.melanogaster									
Raw	Flye	188	93.2	5321187	2304871	0.056	0.028	0.028	605
Dechat	Flye	196	92.9	6612384	2595112	0.061	0.031	0.029	680
Dechat	hifiasm	7484	94.6	392352	1020875	0.065	0.033	0.031	5175
Raw	hifiasm	800	49.8	163092	35758	0.468	0.300	0.168	2237
real-A.thaliana									
Herro	Flye	35	69.8	11350327	8585563	0.015	0.007	0.008	27
Dechat	Flye	70	88.8	11296224	9312360	0.019	0.007	0.012	24
Raw	Flye	131	89.1	11296772	8852727	0.023	0.011	0.012	33
Herro	hifiasm	48	78.1	16588563	14631567	0.017	0.010	0.007	76
Dechat	hifiasm	6686	95.1	558664	1886302	0.038	0.017	0.020	552
Raw	hifiasm	6355	96.5	410177	1178747	0.063	0.033	0.030	685
real-HG002									
Raw	Flye	669	47.9	41718591	-	0.087	0.056	0.032	1430
Dechat	Flye	692	47.9	45484342	-	0.089	0.055	0.033	1367
real-meta-Zymo									
DeChat	hifiasmmeta	1178	65.9	97312	104661	0.047	0.027	0.020	1770
Raw	hifiasmmeta	816	51.3	91651	43326	0.330	0.169	0.161	971

Supplementary Table 8. Assembly benchmarking results for real datasets (real-Dmelanogaster, real-A.thaliana, real-HG002 and real-meta-Zymo) using Flye or hifiasm(-meta). The error rate is equal to the sum of mismatch and indel rate. The results are sorted by the error rate in ascending order. Note that N50 is lower than NGA50 in some hifiasm(-meta)'s assemblies, which means the total contig length is longer than the reference haplotype length. This is likely because that hifiasm(-meta) tends to produce larger assemblies containing duplicated sequences when assembling error-corrected simplex ONT reads. Note that NGA50 is missing if the "Haplotype coverage" is less than 50% (the length of all aligned blocks does not even reach to a half of the reference size). Note that in the real-A.thaliana dataset, DeChat uses filtered reads(follow Herro's protocol) for assembly. For real-HG002 dataset, we downloaded ONT reads of HG002 from the file `12_1_22_R1041_ULCIR_HG002_1_dorado0.4.0_sup4.1.0_5mCG_5hmCG.bam` deposited at HPRC `https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/5b73fa0e-658a-4248-b2b8-cd16155bc157--UCSC_GIAB_R1041_nanopore/HG002_R1041_UL/dorado/v0.4.0_wMods/`) . The average sequencing coverage of real-HG002 is about 20x.

Method	Classifier	Precision	Recall	F1 score
sim-meta-low				
Raw	Centrifuge	0.170	0.933	0.288
DeChat	Centrifuge	0.299	0.933	0.453
sim-meta-high				
Raw	Centrifuge	0.159	0.787	0.264
DeChat	Centrifuge	0.169	0.787	0.279

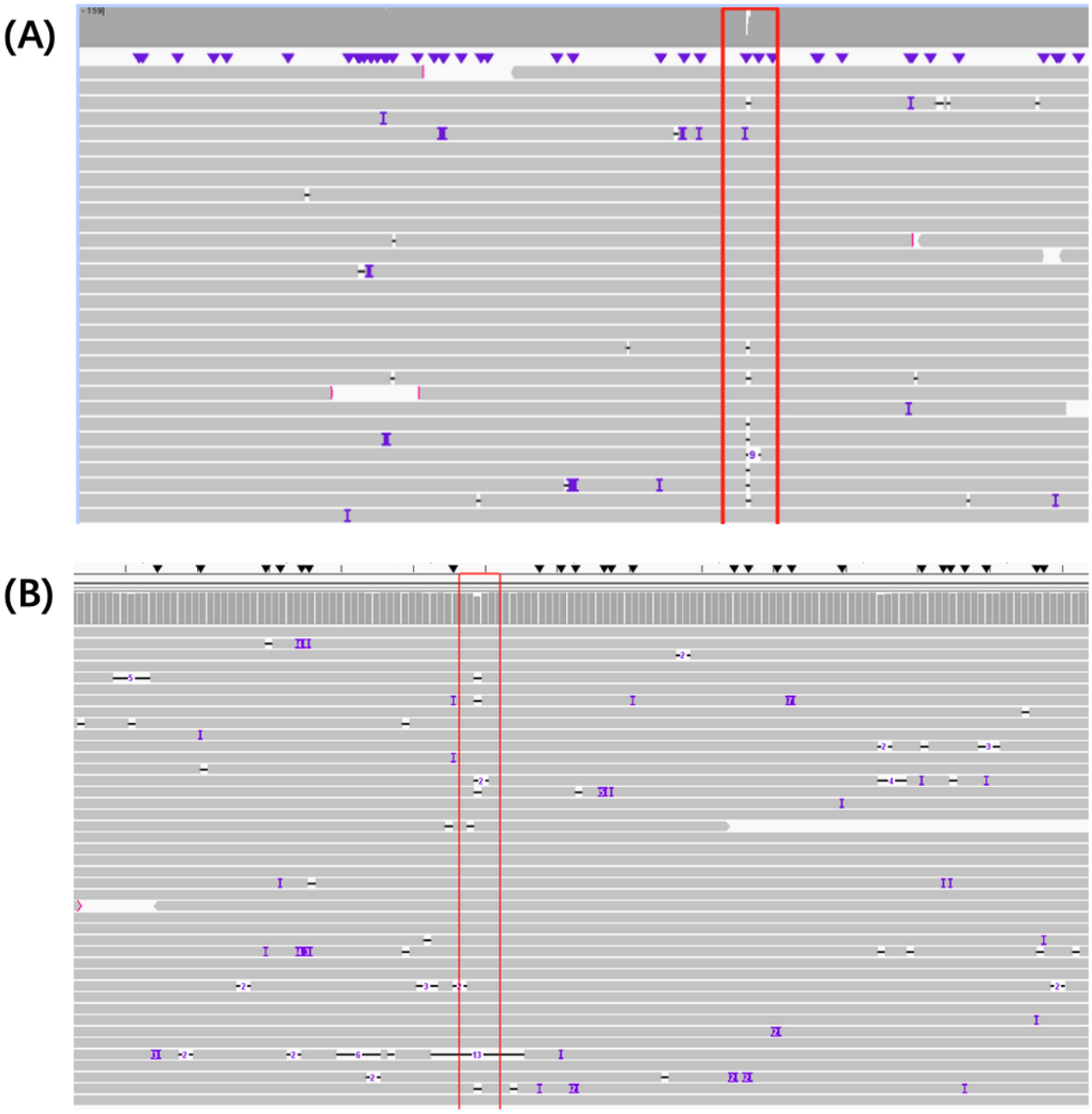
Supplementary Table 9. Benchmarking results of strain-level taxonomic classification. At the level of taxa (strains), Precision measures the fraction of correctly identified taxa among all taxa that were identified. Recall, at the level of taxa, evaluates the proportion of correctly identified taxa compared to all taxa present in the sample. The F1 score, the harmonic mean of precision and recall, is frequently employed as a balanced metric for evaluating classifiers. We retrieved high-quality, complete (in particular gap-free) bacterial genomes from NCBI's RefSeq (Release 219). This establishes a comprehensive reference database, consisting of 34,697 complete genomes. We selected a representative genome for each species as the reference genome, resulting in 8778 species (13404 strains) and their corresponding reference genomes.

Method	DeChat	VeChat	LoRMA	CONSENT	Racon	Canu	Daccord	hifiasm
sim-diploid-ecoli								
CPU(h)	0.33	1.48	4.82	1.76	3.75	1.32	0.63	0.15
Wall Time(h)	0.025	0.062	0.29	4.89	0.269	0.145	0.033	0.014
Memory(GB)	16.71	8.79	56.46	1.78	4.18	3.29	11.59	16.55
sim-haploid-potato								
CPU(h)	165.62	>2000	-	-	>2000	300.01	-	12.93
Wall Time(h)	12.21	>100	-	-	>100	6.60	-	1.87
Memory(GB)	26.27	690.34	-	-	711.38	16.60	-	24.97
real-D.melanogaster								
CPU(h)	233.58	>2000	-	-	2165.97	360.63	-	213.87
Wall Time(h)	10.70	>100	-	-	75.13	35.72	-	12.46
Memory(GB)	71.10	678.06	-	-	678.46	13.82	-	59.22
real-A.thaliana								
CPU(h)	544.63	>2000	-	-	>2000	1105.01	-	512.36
Wall Time(h)	17.66	>100	-	-	>100	46.85	-	19.32
Memory(GB)	121.43	341.36	-	-	1198.78	12.97	-	107.43
sim-meta-high								
CPU(h)	121.76	346.35	-	-	127.84	180.93	80.04	12.68
Wall Time(h)	6.79	7.94	-	-	6.34	4.25	24.03	2.17
Memory(GB)	40.98	89.78	-	-	92.01	41.40	23.36	40.96
real-meta-zymo								
CPU(h)	8.89	49.50	77.86	-	15.99	31.78	14.50	3.17
Wall Time(h)	14.87	2.05	3.21	-	47.80	39.50	41.81	14.11
Memory(GB)	17.49	28.06	143.10	-	18.34	7.05	50.27	17.15

Supplementary Table 10. The CPU time (h), wall time (h), and memory (G) of various tools across different datasets. The ">" symbol indicates that the tool was terminated due to excessive runtime. The CPU time reflects the total runtime of the tool. The tools were run with approximately 24-36 threads.

Method	#Reads	Haplotype coverage (%)	Mismatch per 10 kbp	Non-Hp Ins per 10 kbp	Non-Hp Del per 10 kbp	Hp-Ins per 10 kbp	Hp-Del per 10 kbp	Errors per 10 kbp
sim-diploid-ecoli (NanoSim)								
Dechat	49596	100.0	**0.480**	4.257	**1.302**	0.049	1.133	**7.222**
Vechat	35144	100.0	0.568	**1.195**	3.592	**0.048**	3.041	8.444
LoRMA	29257	100.0	1.854	1.998	3.585	3.585	**0.499**	11.522
Canu	32221	100.0	28.345	7.314	8.244	0.233	7.019	51.155
Racon	36325	99.1	34.756	23.393	11.143	0.801	9.828	79.921
Daccord	28434	97.2	56.846	16.701	57.431	0.364	50.002	181.345
CONSENT	-	-	-	-	-	-	-	-
hifiasm	49596	100.0	1.796	39.934	24.546	9.291	19.596	145.161
Raw	49596	100.0	58.897	43.418	27.612	10.615	22.035	162.577
sim-diploid-potato (NanoSim)								
Dechat	992011	98.2	**14.174**	17.856	**7.109**	**1.451**	**7.639**	**48.229**
Vechat	-	-	-	-	-	-	-	-
LoRMA	267897	62.4	18.846	**15.103**	8.767	2.849	9.459	55.024
Canu	916465	97.1	54.969	26.430	22.821	4.561	24.227	133.007
Racon	674446	76.5	210.422	153.281	138.615	15.637	145.005	662.960
Daccord	-	-	-	-	-	-	-	-
CONSENT	-	-	-	-	-	-	-	-
hifiasm	992011	98.8	-	-	-	-	-	-
Raw	992011	98.8	58.469	42.243	26.523	9.560	26.329	163.124

Supplementary Table 11. Error correction benchmarking results for ONT simplex reads simulated by NanoSim. We used the real ONT simplex reads of *A. thaliana* for training during NanoSim's simulation process. The execution of VeChat was halted due to its excessively prolonged runtime, for example, on sim-diploid-potato (NanoSim), the wall time: 63.18 hours, CPU time: 1947.96 hours, RAM: 155 GB, with 32 threads. Note that CONSENT and Daccord failed to run due to segmentation fault.



Supplementary Figure 1. IGV screenshots for ONT read alignments. (A) real ONT simplex reads (A. thaliana). (B) simulated ONT simplex reads using Nanosim (Yang *et al.*, 2017) (trained with real reads of A. thaliana). Note that the red boxes denote the regions involving recurrent errors.

Supplementary Methods

Metrics for evaluating genome assembly

The genome assembly performance was evaluated by means of several commonly used metrics, routinely reported by QUAST V5.1.0 (Mikheenko *et al.*, 2018), as a prominent assembly evaluation tool. See below for specific explanations. In particular, we ran the `quast.py` program with the option `--unique-mapping` appropriately taking into account that our data sets reflect mixed samples (such as polyploid genome or metagenome).

Haplotype coverage (HC). Haplotype coverage is the percentage of aligned bases in the ground truth haplotypes covered by contigs, which is used to measure the completeness of the given sequence data.

N50 and NGA50. N50 is defined as the length for which the collection of all contigs of that length or longer covers at least half the given sequences. NGA50 is similar to N50 but can only be calculated when the reference genome is provided. NGA50 only considers the aligned blocks (after breaking reads/contigs at misassembly events and trimming all unaligned nucleotides), which is defined as the length for which the overall size of all aligned blocks of this length or longer equals at least half of the reference haplotypes. Both N50 and NGA50 are used to measure the contiguity of the assemblies.

Error rate (ER). The error rate is equal to the sum of mismatch rate and indel rate when mapping the obtained contigs to the reference haplotype sequences.

Number of misassemblies (#MA). The misassembly event in assemblies indicates that left and right flanking sequences align to the true haplotypes with a gap or overlap of more than 1kbp, or align to different strands, or even align to different haplotypes or strains. Here, we report the total number of misassemblies in the given sequence data.

Metrics for evaluating read error correction

Since error-corrected long reads have very similar properties with assembled contigs and for the purpose of unified comparison, we also used QUAST for evaluating the performance(Haplotype coverage) of long-read error correction. Corrected reads and contigs with length less than 500bp were filtered before evaluation.

Haplotype coverage (HC). Haplotype coverage is the percentage of aligned bases in the ground truth haplotypes covered by corrected reads, which is used to measure the completeness of the given sequencing reads.

Commands and versions of tools used for comparison

- PBSIM2

```
haps=(JAIVGA01 JAIVGB01 JAIVGC01 JAIVGD01)
depths=(10 10 10 10)
fq_sample=sample_meta_ont.fq
for tag in "${haps[@]}"
do
  depth=${depths[$i]}
  prefix=$tag.pbsim.$depth"x"
  pbsim --prefix $prefix --depth $depth --sample-fastq $fq_sample $reference --length-min 1000
 done
```

- DeChat v1.0.1

```
dechat -i $raw_read -t $threads -o $corrected_read

#(real-meta-Zymo)
dechat -i $raw_read -t $threads -o -r1 3 -r 6 -e 0.04
```

```
#(sim-different-depth-diploid-potato)
dechat -i $raw_read -t $threads -o $corrected_read -r 6 -e 0.06
```

- VeChat v1.1.0

```
vechat -o $corrected_read --platform ont $raw_read
```

- LoRMA v0.5

```
#LoRMA can only process reads shorter than 40,000 bp, therefore long reads need to be trimmed

LoRMA -reads $raw_read -output $corrected_read -discarded $discarded_read -nb-cores $threads \
-k $kmer
```

- Racon v1.5.0

```
minimap2 -x ava-ont --dual=yes $raw_read $raw_read -t $threads > $overlap.tmp.paf
awk '$11>=500' $overlap.tmp.paf > $overlap.tmp.filter.paf
fpa drop --same-name --internalmatch $overlap.tmp.filter.paf > $overlap
racon -f -t $threads $raw_read $overlap $raw_read > $tmp.fa
Filter out reads shorter than 1000bp from $temp.fa to obtain $correct_read
```

- hifiasm v0.19.8-r603

```
hifiasm -o $corrected_read -t 32 $raw_read
```

- Canu v2.2

```
canu -p $prefix -d $path -correct useGrid=false genomeSize=$genomesize minInputCoverage=1 \
-nanopore $raw_read
```

- Daccord v0.0.10

```
seqkit seq -w 150 $raw_read > $raw_read_split
fasta2DAM $reads.dam $raw_read_split
DBsplit -s256 -x1000 $reads.dam
HPC.daligner $reads.dam -t 24 > $reads.las
daccord $reads.las $reads.dam > $corrected_read
```

- CONSENT v2.2.2

```
CONSENT-correct --in $raw_read --out $corrected_read --type ONT
```

- Quast v5.2.0

```
quast.py -r $ref --min-contig 500 -o out_uniq --ambiguity-usage one -t 32 $fa --fast
```

- Evaluation of reads using bamConcordance

```
minimap2 -a -Q --eqx --secondary=no -I9G -K8G -t $num_threads $ref $fa > mappings.sam
samtools view -Sb -@$num_threads mappings.sam > mappings.bam
bamConcordance $ref mappings.bam out.csv
```

- EDTA v2.2.1

```
EDTA.pl --genome $ref --species others --step all --sensitive 1 --anno 1 -t 32
```

- Flye 2.9.2-b1786

```
flye --nano-raw reads.fa -t 24 -o .
```

```
ln -fs assembly.fasta assembly.fa
```

- NanoSim 3.1.0
```
read_analysis.py genome -i $fq_sample -rg $reference_sample -o output_prefix -t 32
simulator.py genome -rg $reference -n 1000000 -c output_prefix -t 32 -max 20000 -min 5000
```

References

Mikheenko, A. et al (2018). Versatile genome assembly evaluation with quast-lg. *Bioinformatics*, **34**(13), i142–i150.

Yang, C. et al (2017). Nanosim: nanopore sequence read simulator based on statistical characterization. *GigaScience*, **6**(4), gix010.