nature portfolio

https://doi.org/10.1038/s43588-021-00073-4

Supplementary information

Time- and memory-efficient genome assembly with Raven

In the format provided by the authors and unedited

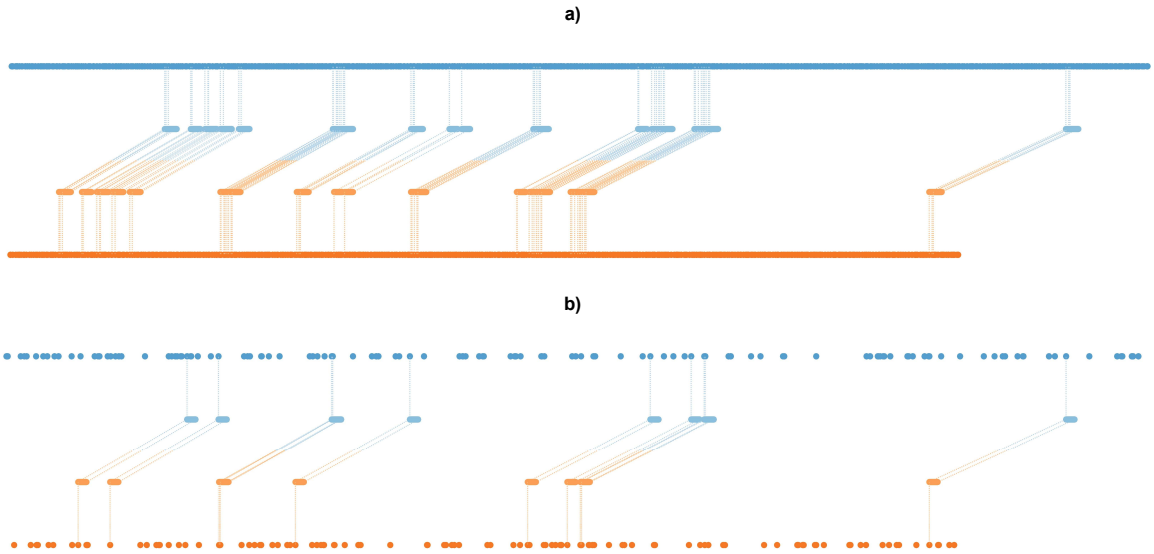
Time and memory efficient genome assembly with Raven

Robert Vaser[1,2] and Mile Šikić[1,2,*]

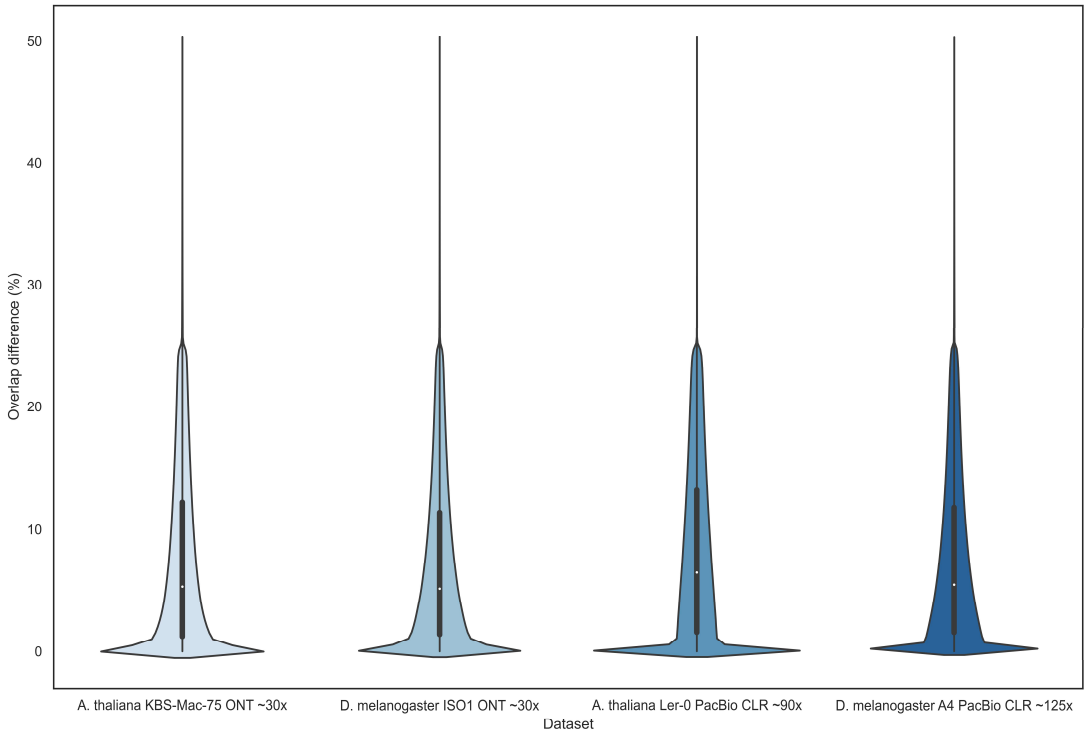
[1] Laboratory for Bioinformatics and Computational Biology, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia

[2] Laboratory of AI in Genomics, Genome Institute of Singapore, A*STAR, Singapore

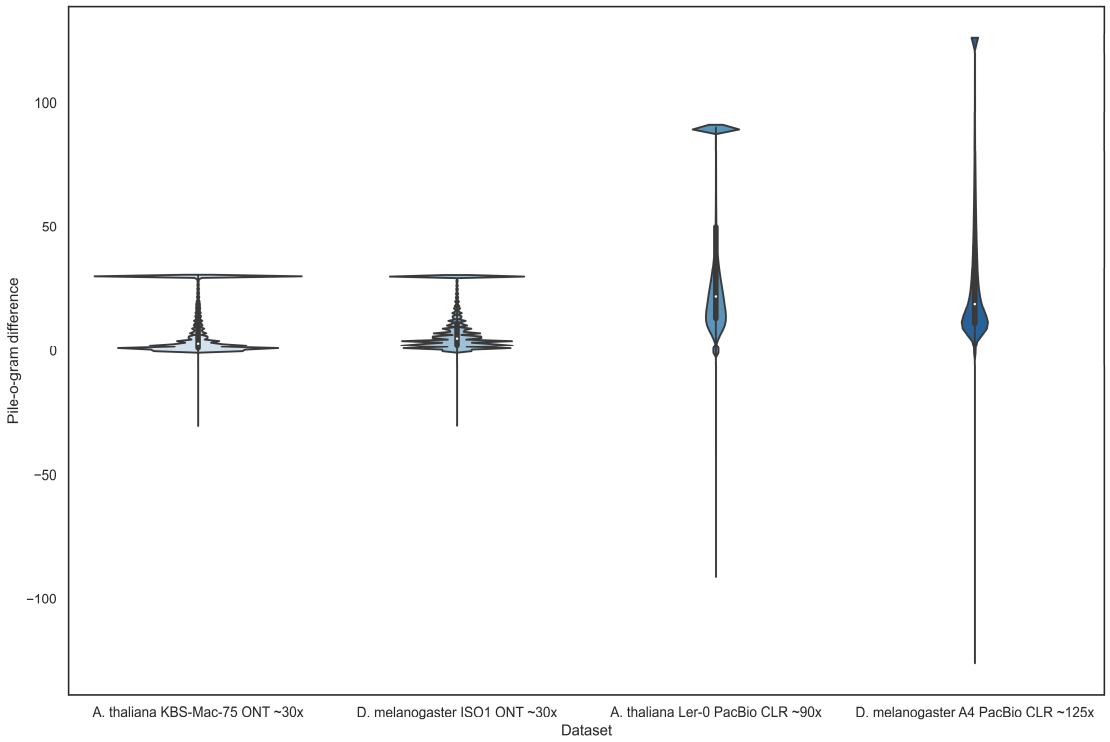
Supplementary information



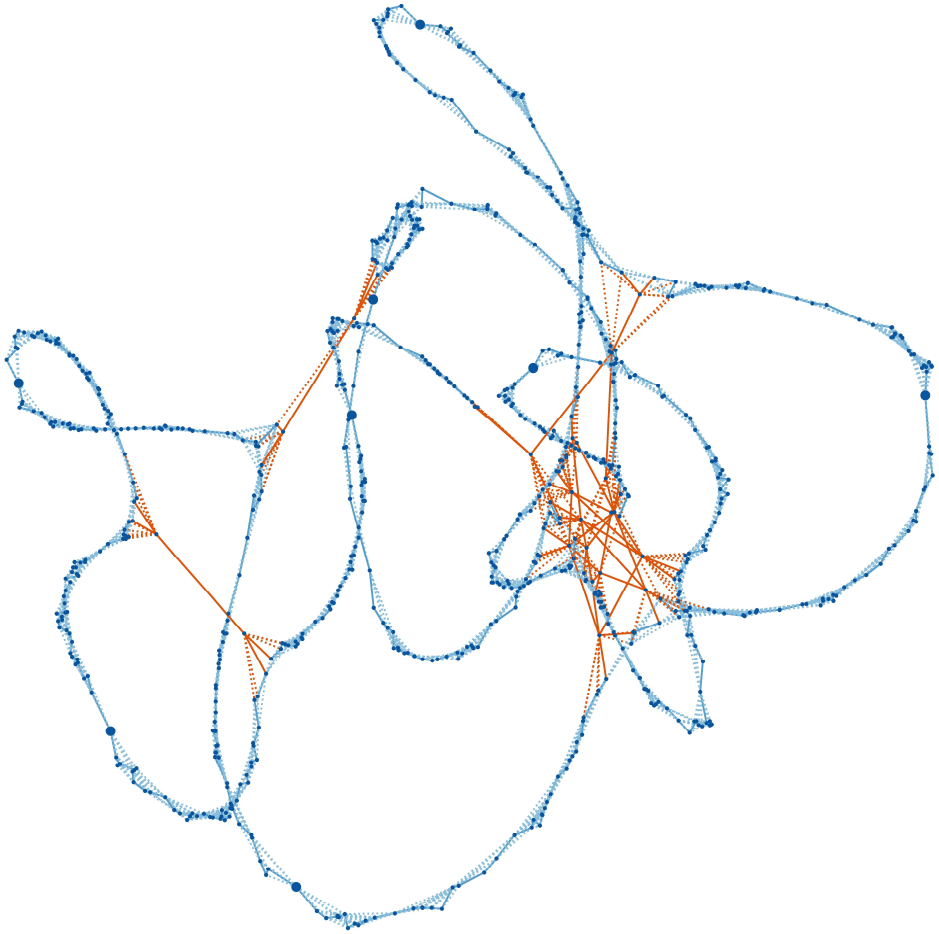
Supplementary Figure 1. Overlap between two erroneous reads based on minimizer matches. Raven uses the minimap algorithm to find pairwise overlaps, in which lexicography smallest k-mers in sliding windows (minimizers) of both reads are collected (blue and orange) and a linear chain of matches is found. a) While collecting all minimizers from a small sliding window ensures the retrieval of most overlaps between similar reads, b) a decent amount of overlaps can be retained by picking only a portion of the smallest minimizers. Shrinking the minimizer search space, without any other modifications, greatly accelerates the algorithm, and justifies the impact on sensitivity for containment removal and pile-o-gram creation.



Supplementary Figure 2. Overlap difference distributions when employing the MinHash paradigm on top of minimizers. Raven reduces the number of minimizers used in the minimap algorithm by choosing only a portion of smallest values per read, which affects the beginning and ending position of pairwise overlaps between reads but enables faster containment removal with a small sensitivity degradation. Depicted values represent the absolute difference between old and new coordinates divided by the old overlap length (overlaps which cover less than 75% of the old region on either of the reads are discarded).



Supplementary Figure 3. Pile-o-gram difference distribution when employing the MinHash paradigm on top of minimizers. Raven reduces the number of minimizers used in the minimap algorithm by choosing only a portion of smallest values per read, which affects per-base coverage in pile-o-grams, but it is negligible for chimeric and repeat annotations. Depicted values represent per-base differences between old and new pile-o-grams, which are saturated to the sequencing depth in absolute value (the biggest differences in coverage are in repetitive regions due to different minimizer sampling).



Supplementary Figure 4. Bacterial assembly graph constructed without read pre-processing and drawn with the force-directed placement algorithm. The graph simplification method based on vertex distances in two-dimensional Euclidean system underperforms when there are many repeat induced edges (orange) which connect distant genomic regions. Compared to Figure 1, overlap removal based on repeat annotations is skipped, which yields a more tangled graph at the end of the layout phase. Although, the simplification method is still able to correctly resolve all false connections for this case, but in a total of two iterations. Number of edges removed constitutes a small portion of edges in the assembly graph and depends on many intertwined factors (for *A. thaliana* and *D. melanogaster* datasets this value ranges between 1% and 11% with respect to number of remaining edges at the end of the layout phase). To fully utilize this method, read pre-processing should be applied beforehand.

Supplementary Table 1. Assembly cost estimates. Values with an asterisk were found in publications of corresponding assemblers, while values with a tilde are approximations when more threads are invoked. Canu assemblies were omitted due to long running times (according to (Shafin et al. 2020), its estimate cost is around $19000 for all three ONT datasets).

Dataset	Assembler	Number of threads	Real Time (h)	Memory (Gb)	AWS instance	AWS cost / hour ($)	AWS cost ($)
H. sapiens CHM13 ONT ~130x	Raven	48	123.12	251	m5a.16xlarge	2.75	~253.94
	Flye	48	141.78	873	x1.16xlarge	6.67	~709.25
	Shasta	128*	5.28*		x1.32xlarge	13.34	70.43
	Wtdbg2	48	128.58	423	r5a.16xlarge	3.62	~349.09
H. sapiens HG002 ONT ~60x	Raven	64	25.21	105	c5a.16xlarge	2.46	62.02
	Flye	64	49.86	951	x1.16xlarge	6.67	332.57
	Shasta	64	3.09	771	x1.16xlarge	6.67	20.61
	Wtdbg2	64	36.59	352	r5a.16xlarge	3.62	132.46
H. sapiens HG00733 ONT ~80x	Raven	64	26.88	131	m5a.16xlarge	2.75	73.92
	Flye	64	57.14	546	x1.16xlarge	6.67	381.12
	Shasta	64	2.83	870	x1.16xlarge	6.67	18.88
	Wtdbg2	64	31.46	345	r5a.16xlarge	3.62	113.89
H. sapiens CHM13 PacBio CLR ~50x	Raven	64	10.41	98	c5a.16xlarge	2.46	25.61
	Flye	64	35.29	407	r5a.16xlarge	3.62	127.75
	Shasta	64	1.54	547	x1.16xlarge	6.67	10.27
	Wtdbg2	64	8.06	180	m5a.16xlarge	2.75	24.64
H. sapiens HG002 PacBio CLR ~80x	Raven	64	23.78	129	m5a.16xlarge	2.75	65.40
	Flye	64	63.50	562	x1.16xlarge	6.67	423.55
	Shasta	64	1.50	567	x1.16xlarge	6.67	10.01
	Wtdbg2	64	9.29	207	m5a.16xlarge	2.75	25.55
H. sapiens HG00733 PacBio CLR ~95x	Raven	64	30.92	138	m5a.16xlarge	2.75	85.03
	Flye	64	110.94	663	x1.16xlarge	6.67	800.00
	Shasta	48	1.28	1012	x1.32xlarge	13.34	~6.40
	Wtdbg2	64	25.12	340	r5a.16xlarge	3.62	90.93
H. sapiens PacBio HiFi ~35x	hifiasm	48*	9*	150*	m5a.12xlarge	2.06	18.54

Supplementary Table 2. Evaluation of long-read assemblers on older sequencing datasets. Values in columns that are missing CPU time and memory were obtained with assemblies from other publications. Canu assembly of dataset NA12878 was polished with short accurate reads and is excluded from accuracy comparison. Ra was not run on the NA12878 dataset due to its complexity on larger genomes. NG50 metric is defined as the length of the contig which coupled with longer contigs covers 50% of the reference genome. NGA50 is calculated the same way but on top of alignments between contigs and the reference. Quality value denotes the Phred base error rate of the assembly obtained by comparing k-mers between short accurate reads and the assembly. Multi-copy genes are those that occur multiple times both in the reference and the assembly. 99.5% of bases of a BAC need to be present in the assembly for it to be resolved. Bolded values represent the best metric scores.

Dataset	Metric	Raven	Canu	Flye	miniasm	Ra	Shasta	Wtdbg2
A. thaliana KBS-Mac-74 ONT ~30x	Genome fraction (%)	99.28	95.39	**99.88**	99.51	99.74	76.32	97.50
	No. of contigs	**25**	448	118	62	57	1382	353
	NG50 (Mb)	11.1	2.6	**13.3**	11.2	7.4	0.3	9.8
	NGA50 (Mb)	5.6	2.2	**9.2**	7.0	5.7	0.3	3.1
	NGA75 (Mb)	3.3	0.4	**4.9**	3.3	2.5	-	1.0
	No. of misassemblies	261	368	653	256	420	**41**	500
	Mismatch fraction (%)	0.30	**0.16**	0.30	0.18	0.33	0.51	0.36
	Indel fraction (%)	1.73	2.25	1.59	**1.41**	1.42	2.57	3.00
	Single-copy genes (%)	75.91	38.23	81.40	**84.25**	83.47	11.92	25.83
	Duplicated genes (%)	0.01	0.01	0.04	0.01	0.01	**0.0**	**0.0**
	Multi-copy genes (%)	0.0	0.0	**2.08**	0.0	**2.08**	0.0	0.0
	CPU time (h)	4.5	1157.5	22.4	6.0	9.5	**0.6**	19.8
	Memory (GB)	**9.6**	10.6	87.9	21.7	30.5	21.6	15.8
A. thaliana Ler-0 PacBio CLR ~90x	Genome fraction (%)	99.60	99.07	**99.69**	99.30	99.62	22.48	99.28
	No. of contigs	**74**	591	174	155	112	1508	280
	NG50 (Mb)	10.8	0.7	**14.0**	8.7	6.8	-	12.2
	NGA50 (Mb)	6.1	0.7	**6.7**	6.2	6.4	-	6.1
	NGA75 (Mb)	3.1	0.3	**4.5**	1.8	2.3	-	2.7
	No. of misassemblies	792	1189	798	611	833	**22**	728

	Mismatch fraction (%)	0.13	0.22	0.14	**0.11**	0.17	0.37	0.18
	Indel fraction (%)	0.25	0.08	**0.02**	0.23	0.58	2.12	0.28
	Single-copy genes (%)	98.66	98.75	**99.89**	98.63	96.58	8.54	99.17
	Duplicated genes (%)	0.07	0.09	0.03	0.12	0.07	**0.0**	0.02
	Multi-copy genes (%)	72.58	**93.55**	85.48	72.58	38.71	0.0	45.16
	CPU time (h)	22.9	238.9	62.2	25.6	29.1	**0.8**	43.4
	Memory (GB)	18.8	**12.2**	59.7	46.7	32.7	37.4	25.7
	Genome fraction (%)	92.20	**94.33**	93.02	92.32	88.38	71.76	91.37
	No. of contigs	**148**	664	468	219	232	1852	635
	NG50 (Mb)	6.1	4.6	**19.6**	3.3	1.9	0.1	10.6
	NGA50 (Mb)	1.4	1.2	**1.7**	1.1	1.1	0.1	1.0
	NGA75 (Mb)	0.5	0.5	**0.6**	0.4	0.3	-	0.3
D. melanogaster	No. of misassemblies	1230	3167	1316	1098	605	**342**	1974
ISO1	Mismatch fraction (%)	**0.16**	0.22	**0.16**	0.18	0.19	0.46	0.37
ONT ~30x	Indel fraction (%)	0.71	0.93	**0.41**	0.74	0.73	1.80	1.56
	Single-copy genes (%)	98.57	98.06	**99.27**	98.22	97.86	63.43	96.08
	Duplicated genes (%)	0.07	0.28	0.04	0.15	0.07	**0.0**	0.03
	Multi-copy genes (%)	52.40	**57.21**	56.73	47.12	21.15	0.96	3.37
	CPU time (h)	5.1	520.8	25.6	7.9	13.7	**0.6**	26.9
	Memory (GB)	**12.9**	13.1	33.4	23.4	26.7	21.5	19.2
	Genome fraction (%)	93.46	**95.97**	92.29	93.71	90.42	91.24	92.83
	No. of contigs	**121**	254	199	299	177	484	311
	NG50 (Mb)	12.8	13.8	15.6	6.5	4.3	3.5	**17.0**
	NGA50 (Mb)	3.9	**9.4**	8.3	3.2	2.6	2.7	4.5
	NGA75 (Mb)	1.2	2.0	**2.2**	1.3	0.8	0.9	1.4
D. melanogaster	No. of misassemblies	771	774	609	791	**405**	416	761
A4	Mismatch fraction (%)	0.05	0.04	0.04	0.06	**0.03**	**0.03**	0.17
PacBio CLR ~125x	Indel fraction (%)	0.12	0.04	**0.03**	0.12	0.13	0.13	0.29
	Single-copy genes (%)	99.53	99.02	**99.78**	99.18	99.16	99.20	99.55
	Duplicated genes (%)	0.14	0.90	0.07	0.50	0.20	**0.0**	0.04
	Multi-copy genes (%)	80.45	**92.74**	83.80	86.59	80.45	29.05	59.78
	CPU time (h)	25.5	389.2	75.8	37.9	61.4	**4.3**	20.5
	Memory (GB)	22.2	**19.1**	79.6	56.6	62.0	62.8	19.4
	Genome fraction (%)	92.27	92.04	**92.75**	90.61		91.49	87.36
	No. of contigs	**249**	1145	1264	502		2989	5147
	NG50 (Mb)	27.9	10.6	**31.8**	9.7		3.6	9.8
	NGA50 (Mb)	16.0	8.1	**19.4**	8.0		3.4	5.7
	NGA75 (Mb)	5.9	2.9	**8.5**	3.4		1.3	1.5
	Mismatch fraction (%)	0.14	0.15	**0.13**	0.14		0.15	0.24
H. sapiens	Indel fraction (%)	0.34	0.05	0.36	**0.25**		0.36	0.72
NA12878	Quality value	25.66	35.06	25.48	**27.00**		25.21	22.45
ONT ~45x	Single-copy genes (%)	90.28	94.04	90.02	**95.20**		70.85	58.88
	Duplicated genes (%)	0.20	0.25	0.30	0.52		**0.01**	0.02
	Multi-copy genes (%)	48.01	42.77	41.35	**49.14**		7.49	2.25
	Resolved BACs (%)	61.18	44.73	40.08	**63.71**		16.46	8.86
	CPU time (h)	470		1264	1373		**29**	1994
	Memory (GB)	**83**		730	401		391	279

Supplementary Table 3. Raven plant assemblies. Values in brackets represent best assembly metrics in corresponding publications. *Oryza sativa* assemblies in the original publication were additionally polished with Illumina reads. N50 metric is defined as the length of the contig which coupled with longer contigs covers 50% of the assembly. Complete BUSCOs denote the percentage of single-copy and duplicated orthologs from the *embryophyta* database that are found in the assembly.

Metric \ Dataset	*Brassica oleracea*	*Brassica rapa*	*Musa schizocarpa*	*Oryza sativa basmati 334*	*Oryza sativa dom sufid*
Total length (Mb)	535.9 (546.4)	351.7 (375.3)	534.4 (522.0)	382.4 (386.6)	380.5 (383.6)
N50 (Mb)	6.4 (7.3)	5.5 (3.8)	2.5 (2.13)	8.1 (6.3)	11.9 (10.5)
No. of contigs	252 (244)	410 (544)	546 (615)	116 (188)	107 (116)
Complete BUSCOs (%)	74.78 (74.30)	85.94 (79.70)	47.15 (53.80)	92.50 (97.60)	92.19 (97.00)
CPU time (h)	41 (261)	59 (316)	95 (246)	44 (N/A)	34 (N/A)

Supplementary Table 4. Impact on sensitivity and execution time when employing MinHash on top of minimizers. For overlap Jaccard score calculation new overlaps are declared valid if they cover at least 75% of the old regions on both reads.

	A. thaliana KBS-Mac-75 ONT ~30x	*D. melanogaster* ISO1 ONT ~30x	*A. thaliana* Ler-0 PacBio CLR ~90x	*D. melanogaster* A4 PacBio CLR ~125x
No. of overlaps - minimizers	33639084	38522454	250456231	376494493
No. of overlaps - minimizers and MinHash	8724876	15282409	26442472	153956450
Overlap Jaccard score	0.208	0.238	0.057	0.266
Containment Jaccard score	0.800	0.835	0.698	0.864
Speedup	4.42	3.62	4.21	4.01

Supplementary Table 5. Impact of overlap parameters on Raven's performance on PacBio HiFi data. The value pairs in brackets represent the k-mer length and the sampling window length. Number of misassemblies for HG002 and HG00733 datasets were removed due to differences between the samples and the reference genome. Dataset HG002 does not have available BAC clones. NG50 metric is defined as the length of the contig which coupled with longer contigs covers 50% of the reference genome. NGA50 is an extension which is calculated on top of alignments between contigs and the reference. Quality value denotes the Phred base error rate of the assembly obtained by comparing k-mers between short accurate reads and the assembly. Multi-copy genes are those that occur multiple times both in the reference and the assembly. 99.5% of bases of a BAC need to be present in the assembly for it to be resolved.

	H. sapiens CHM13 PacBio HiFi ~35x		*H. sapiens* HG002 PacBio HiFi ~35x		*H. sapiens* HG00733 PacBio HiFi ~35x	
Metric	Raven (15,5)	Raven (29,9)	Raven (15,5)	Raven (29,9)	Raven (15,5)	Raven (29,9)
Genome fraction (%)	91.49	92.55	91.20	92.14	91.14	91.96
No. of contigs	5689	1755	7282	2375	7615	2176
NG50 (Mb)	1.1	12.0	0.8	6.5	0.8	7.1
NGA50 (Mb)	0.8	10.4	0.6	5.9	0.6	6.1
NGA75 (Mb)	0.3	3.7	0.2	2.1	0.2	2.2
No. of misassemblies	3803	2921				
Mismatch fraction (%)	0.04	0.06	0.15	0.18	0.14	0.16
Indel fraction (%)	0.01	0.01	0.03	0.04	0.03	0.03
Quality value	43.52	43.51	41.94	42.27	39.66	40.06
Single-copy genes (%)	95.01	98.29	93.97	97.57	94.17	97.58
Duplicated genes (%)	1.91	0.39	1.82	0.48	1.60	0.49
Multi-copy genes (%)	42.10	44.64	36.93	38.73	33.26	37.15
Resolved BACs (%)	28.59	39.10			18.42	22.11
CPU time (h)	1313	554	3449	527	1300	486
Memory (GB)	87	65	91	67	97	70