Open access • Posted Content • DOI:10.1101/2020.08.07.242461

Raven: a de novo genome assembler for long reads — Source link

Robert Vaser, Robert Vaser, Mile Šikić, Mile Šikić

Institutions: Genome Institute of Singapore, University of Zagreb

Published on: 10 Aug 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Genome, Sequence assembly and Human genome

Raven: a de novo genome assembler for long reads

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

We present new methods for the improvement of de novo genome assembly from erroneous long-reads incorporated into a straightforward tool called Raven (https://github.com/lbcb-sci/raven). Raven maintains similar performance for various genomes and has accuracy on par with other assemblers which support third-generation sequencing data. It is one of the fastest options while having the lowest memory consumption on the majority of benchmarked datasets.

Sequencing technologies have come a long way, from tiny fragments at their infancy to large chunks obtainable today. The relentless advances in both length and accuracy continue to alleviate the puzzle-like reconstruction problem of the sequenced genome, as more repetitive structures can be resolved naturally. Amidst the excess of available state-of-the-art options for de novo genome assembly[1–6], we present a fast, memory frugal, reliable, and easy to use tool called Raven. It is an overlap-layout-consensus based assembler which accelerates the overlap step, builds an assembly graph[4] from reads that were pre-processed with pile-o-grams[7], implements a novel and robust simplification method based on graph drawings, and polishes the unambiguous graph paths with Racon[8], all of which is compiled into a single executable.

Short substring matching is a conventional approach for similarity search in bioinformatics[9,10]. However, even with minimizers[4] the overlap step of de novo assembly can take a substantial amount of time when handling larger genomes. To tackle this problem we enhanced the minimap[4] algorithm following the MinHash approach[11], where we select a fixed number of lexicographically smallest minimizers as the sequence sketch. The combination of MinHash on top of minimizers was already explored within the sequence mapper MashMap[12], while a similar idea with hierarchical minimizers is the core of de novo assembler Peregrine[13]. Based on empirical evaluations, we opted for retaining $|read|/k$ minimizers per read, where k is the minimizer length. Without any other algorithmic modifications to minimap, we are able to identify contained reads and create pile-o-grams for read pre-processing in a fraction of time and with a small impact on sensitivity. Suffix-prefix overlaps needed for graph constructions are found with the unmodified minimap algorithm within the containment-free read set, which is usually smaller than the whole sequencing yield by almost an order of magnitude.

Raven loads the whole sequencing sample into memory in compressed form, and finds overlaps in fixed-size blocks to decrease the memory footprint. Found overlaps are immediately transformed into pile-o-grams and discarded, except the longest few per read which are used for containment removal. Chimeric reads are iteratively identified and chopped by detecting sharp declines of coverage in pile-o-grams using coverage medians inferred from the stored overlaps. As minimap ignores the most frequent minimizers, which are critical for good repeat annotations, we lower this threshold while overlapping all contained reads to the set of containment-free reads, and search the updated pile-o-grams for sharp coverage inclines followed by sharp declines, both above the coverage median. Afterwards, the containment-free read set is overlapped to itself and repeat annotations are used to remove false overlaps between reads containing repetitive regions. Once the assembly graph is created, it is simplified stepwise with transitive reduction, tip removal, and bubble popping. Eventually, we simplify the graph with a novel method which lays out the graph in a two-dimensional

Euclidean system, searches for edges that connect distant parts of the graph and removes them. Applying the force-directed placement algorithm[14], which draws tightly connected vertices together, we can distinguish undetected chimeric or repeat-induced edges which are elongated with respect to others due to their rareness (Figure 1). Collapsing unambiguous paths while leaving room near junction vertices, coupled with the hierarchical force-calculation algorithm[15], makes this drawing based simplification method feasible for even the largest assembly graphs. To finalize the assembly, contiguous paths of the graph are passed to two rounds of Racon.

Since an earlier version of Raven proved as one of the best performers in a comprehensive benchmark[16] at prokaryotic level, we evaluated several state-of-the-art assemblers alongside Raven on five model eukaryote datasets (Table 1), obtained by third-generation sequencing technologies, namely Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Emergence of PacBio's High-Fidelity sequencing protocol (HiFi), and novel assemblers[13,17,18] suitable for its highly accurate data, led us to evaluate the assembly reconstruction prospects of different sequencing approaches, that is ONT, Pacbio CLR (continuous long reads) and Pacbio HiFi, on three human samples (Table 2). Alongside default assembly quality metrics such as NGAx, genome fraction and accuracy, we evaluated gene completeness (single and multi-copy genes present both in the reference and the assembly), and where possible, the number of bacterial artificial chromosomes (BAC) resolved in an assembly. Details about computational cost can be found in the Supplementary (Table S1).

On erroneous data, Raven is one of the fastest assemblers, and uses the least amount of memory on all but two datasets, while having better or comparable contiguity and accuracy. It especially stands out in the number of contigs with similar genome reconstruction fractions, and in the number of retained multy-copy genes and resolved BACs on human datasets. On the other hand, Raven does not utilize the accuracy of HiFi reads, which results in longer running times and subpar assembly results on more accurate data. We believe that more carefully tweaked parameters for the overlap step will lead to performance improvements.

We also run Raven on a couple of ONT plant datasets from two scientific studies[19,20] and compared their results (Table 3). On datasets *B. oleracea*, *B. rapa* and *M. schizocarpa* Raven produces comparable assemblies to those obtained with Ra[21]. Furthermore, both *O. sativa* assemblies are more contiguous than the ones reported with Flye, but the BUSCO[22] scores are lower as we did not polish our assemblies with Illumina data.

Presented results indicate that PacBio HiFi assemblers achieve better overall reconstruction metrics, although ONT assemblies do not fall far off. ONT sequencing is still more approachable due to affordable consumables and portable devices, while requiring less gDNA than regular PacBio protocols. In addition, the length advantage of ONT reads and the recent increase in accuracy with the newest version of the Bonito basecaller (still in testing phase) justify the usage of assemblers which support this technology.

We showcased new algorithms for the overlap and layout phases of de novo genome assembly that reduce execution time and increase contiguity of the final assembly. We integrated them with an overlap module based on minimap, and the consensus module Racon, into a powerful standalone tool called Raven which is optimized for error-prone long reads. We argue that its performance coupled with the reduced cost per base of long-read sequencing technologies will enable assembly of large genomes even to laboratories with limited funding.

Methods

Raven starts the assembly by constructing pile-o-grams (one-dimensional structures storing per-base coverage) and removing contained reads with the minimap algorithm, using 15-mers, a sliding window of 5 bases and discarding 10^{-3} most frequent minimizers. The whole sequencing data set is loaded into memory, replacing nucleotides with two bits and merging 64 succeeding Phred quality scores with their average. Reads are overlapped to each other in 1Gbp vs 4Gbp chunks, and only the lexicographically smallest $\lfloor read \rfloor / 15$ minimizers are picked in both the index and the query (Supplementary Figure S1-S3; accuracy comparison in Supplementary Table S2). Once a block is processed, all overlaps are stacked into pile-o-grams which are decimated to every 16-th base. The longest 16 overlaps per read are stored for containment removal and connected component retrieval. When all pairwise overlaps are obtained, coverage medians are calculated for each pile-o-gram, reads are trimmed to the longest region covered with at least 4 other reads, and potential chimeric sites are detected by finding bases which have 1.82 times smaller coverage than their neighboring bases. Contained reads are dropped only if the containing read does not have a potential chimeric region. Decreasing the number of reads through containment removal enables faster verification of chimeric annotations. Given the stored suffix-prefix overlaps, Raven finds connected components and their coverage median, which approximates the sequencing depth. Each annotated coverage drop is used to chop problematic reads to their longest non-chimeric region, if the drop is consistent with the coverage median of the connected component the read belongs to. The whole process is done iteratively to capture different molecule copy numbers, because resolving chimeric reads tends to the forming of new connected components. Another containment check is carried out once chimeric sequences are resolved.

Afterwards, Raven searches for suffix-prefix overlaps between the remaining reads enforcing the use of all minimizers. In addition, all contained reads are overlapped to the containment-free read set in order to increase the coverage of repetitive regions, again employing the MinHash approach. Decreasing the minimizer frequency filter to 10^{-5} enables proper repeat annotation in which sought bases need to have coverage at least 1.42 times larger than the component coverage median. Repetitive regions at either end of a read are used to iteratively remove false overlaps, i.e. overlaps that connect different copies of bridged repeats (repetitive genomic regions that are entirely contained in at least one read).

Once the overlap set is cleaned, the assembly graph is built and simplified stepwise with standard layout algorithms such as transitive reduction, tipping, and bubble popping. Information about transitive connections is kept for the last simplification step, which plots the assembly graph in a two-dimensional space, in order to increase the connections between neighboring vertices. Raven searches for edges connecting remote parts of the graph, which are usually present due to leftover sequencing artefacts or unresolved repeats. The force-directed placement algorithm enlarges most of such edges due to their rareness. Given the quadratic time complexity $O(|V|^2)$[14] and an approximate of 100 iterations until convergence, we shrink the graph by creating unitigs (paths in the graph consisting of vertices with only one ingoing and one outgoing edge) that are 42 vertices away from any junction vertex (vertices with more than one outgoing or ingoing edge). Furthermore, approximating the forces of distant vertices by replacing them with their centre of mass enables linearithmic time complexity $O(|V| \log |V|)$[15], and the use of this method on larger genomes. Depending on vertex distances in a finished drawing, Raven removes outgoing edges that are at least twice as long as any other outgoing edge of that junction vertex. As the drawing heavily depends on an initial layout, which is random but with a fixed seed, the whole procedure is restarted 16 times. It should be noted that if there exist a lot of false connections in a single area of the graph (usually induced by repeats), the drawing algorithm will not be able to sufficiently enlarge all of these edges for removal (Supplementary Figure S4).

Finally, paths of the assembly graph without external branches are polished with a library version of Racon, using small windows of 500bp and partial order alignment with linear gaps, in a total of two iterations. All constant values used in various Raven stages were empirically determined based on a large set of real datasets of various sizes.

Because of resource limitations we chose the best performing genome assemblers for erroneous third-generation data from recent scientific papers[3,6,16]. The assemblers are Raven (v1.3.0), Canu (v2.0), Flye (v2.8.1), miniasm (v0.3-r179) coupled with minimap (v0.2-r123) and polished with two iterations of Racon (v1.4.13), Ra (v0.2.1), Shasta (v0.7.0) and Wtdbg2 (v2.5). Raven was run without any additional parameters on ONT and PacBio CLR datasets. On PacBio HiFi datasets, we increase k-mer length from 15 to 29, and window length from 5 to 9, in order to decrease the number of found pairwise overlaps (comparison with default parameters can be found in Supplementary Table S3). We use options '-pacbio' or '-nanopore' for Canu, '-pacbio-raw' or '-nano-raw' for Flye, '-x ont' or '-x pb' for Ra, '-x sq', '-x rs' or '-x ont' for Wtdbg2, and configuration files Nanopore-Dec2019, Nanopore-Sep2020 or PacBio-CLR-Dec2019 for Shasta. For ONT runs we modified the Shasta consensus caller to better match the basecaller used to obtain the corresponding dataset, while we decreased the minimal read length to 5000 for non-human datasets, except PacBio CLR *D. melanogaster* dataset for which Shasta produced a decent assembly. Canu and Wtdbg2 require approximate genomes sizes which were 120 Mb, 144 Mb, and 3 Gb for *A. thaliana*, *D. melanogaster* and *H. sapiens* datasets, respectively. All assemblers were run with 64 threads on a server with 1 TB RAM and two AMD EPYC™ 7702 64-core processors. Due to high memory requirements, the ONT CHM13 dataset was benchmarked with 48 threads on a server with two Intel® Xeon® Platinum 8260L 24-core processors and 1.5 TB of Optane™ Persistent Memory. Shasta was unable to assemble the PacBio CLR HG00733 dataset on the first machine due to memory requirements, so it was run on the second machine. Also, it was not able to assemble the ONT CHM13 dataset on either machine, so we found the assembly in its publication. Canu was not run on human datasets due to its long running time, but we found assemblies in other publications[5,23] (NA12878 assembly was polished with Illumina data so it was excluded from accuracy comparison). We omitted Ra from the human dataset benchmark due to its complexity on larger genomes. Hifiasm human assemblies were found in its publication[18].

We used QUAST-LG[24] (v5.0.2) for assembly evaluation and ran it with minimal identity of 80%. For *H. sapiens* datasets we used the T2T (telomere-to-telomere) reconstruction of CHM13 (and options '--large' in QUAST), while for *A. thaliana* and *D. melanogaster* datasets we used appropriate NCBI assemblies or references depending on the strain. The assembly quality value (QV) was obtained with yak (v0.1), which is available at https://github.com/lh3/yak, by comparing 31-mers found in short accurate reads and the assembly for datasets NA12878, HG002 and HG00733. Gene completeness was evaluated with paftools (v2.17-r982) asmgene function, found inside the minimap2[25] package. We mapped annotated Ensembl cDNA sequences (v102 for *D. melanogaster* and *H. sapiens*, and v49 for *A. thaliana*) to the references and the assemblies. Identity of 97% was used to find single-copy and duplicated single-copy genes, while 99% identity was used for multi-copy genes. We validated BAC resolution with a pipeline available at https://github.com/skoren/bacValidation (commit 4f3e463), where 99.5% of bases of a BAC need to be present in the assembly for it to be resolved. We used VMRC53 (237 BACs), VMRC59 (647 BACs) and VMRC62 (190 BACs) clones for NA12878, CHM13 and HG00733, respectively. BUSCO (v4.1.4) scores for the five plant datasets were found with the *embryophyta* database, although the current version contains more orthologs (1614 in total).

ONT dataset for *A. thaliana* is available under the accession number ERR2173373, for *D. melanogaster* under SRR6702603, for *H. sapiens* NA12878 here (release 6), for *H. sapiens* CHM13 here (release 6), for *H. sapiens* HG002 here, and for *H. sapiens* HG00733 here.

PacBio CLR dataset for *A. thaliana* is available here, for *D. melanogaster* under accession number SRR5439404, for *H. sapiens* CHM13 here (extracted from draft v1.0 bam), for *H. sapiens* HG002 here, and for *H. sapiens* HG0073 under SRR7615963.

PacBio HiFi dataset for *H. sapiens* CHM13 is available from accession number SRR11292120 to SRR11292123, for *H. sapiens* HG002 under SRR10382244, SRR10382245, SRR10382248 and SRR10382249, and for *H. sapiens* HG00733 under ERX3831682.

Illumina reads for yak evaluation are available from accession number SRX1049768 to SRX1049782 for *H. sapiens* NA12878, here (extracted from 60x bam) for *H. sapiens* HG002, and under accession number SRR7782677 for *H. sapiens* HG00733.

Accession numbers of the plant datasets used for separate Raven evaluation can be found in corresponding publications.

All generated assemblies in this research can be found at Zenodo under DOI 10.5281/zenodo.4443062.

Acknowledgments

This work has been supported in part by the Croatian Science Foundation under the project Single genome and metagenome assembly (IP-2018-01-5886), by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS), and by the A*STAR Computational Resource Centre through the use of their high-performance computing facilities. R.V. and M.Š. have been partially supported by funding from A*STAR, Singapore. We acknowledge Intel® Corporation for allowing us to test with Intel® Optane™ Persistent Memory server and providing us with high-quality technical support. Finally, we thank Goran Žužić from Carnegie Mellon University for useful discussions in the field of graph drawings.

References

1. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

2. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

3. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

4. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

5. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0503-6.

6. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).

7. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res.* **27**, 747–756 (2017).

8. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search

tool. *J. Mol. Biol.* **215**, 403–410 (1990).

10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

11. Broder, A. Z. On the resemblance and containment of documents. in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* 21–29 (1997). doi:10.1109/SEQUEN.1997.666900.

12. Jain, C., Dilthey, A., Koren, S., Aluru, S. & Phillippy, A. M. A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. in *Research in Computational Molecular Biology* (ed. Sahinalp, S. C.) 66–81 (Springer International Publishing, 2017).

13. Chin, C.-S. & Khalak, A. Human Genome Assembly in 100 Minutes. *bioRxiv* (2019) doi:10.1101/705616.

14. Fruchterman, T. M. J. & Reingold, E. M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129–1164 (1991).

15. Barnes, J. & Hut, P. A hierarchical O(N log N) force-calculation algorithm. *Nature* **324**, 446–449 (1986).

16. Wick, R. R. & Holt, K. E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 2; peer review: 4 approved]. *F1000Research* **8**, (2020).

17. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).

18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

19. Belser, C. *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).

20. Choi, J. Y. *et al.* Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* **21**, 21 (2020).

21. Vaser, R. & Šikić, M. Yet another de novo genome assembler. in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)* 147–151 (2019). doi:10.1109/ISPA.2019.8868909.

22. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

23. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

24. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

25. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Figure 1 Bacterial assembly graph drawn with the force-directed placement algorithm. Raven uses vertex distances in two-dimensional Euclidean system to find elongated edges (red) that connect junction vertices and removes the longest ones. Those represent false connections which occur either due to sequencing errors or repetitive genomic regions. Without unitig creation (large circles) and the hierarchical force calculation, the drawing algorithm would partake an extensive amount of time on larger genomes. In addition, transitive edges (dotted green) are reinstated to increase the connectivity of neighboring vertices.

Table 1 Evaluation of long-read assemblers.

Dataset	Metric	Raven	Canu	Flye	miniasm	Ra	Shasta	Wtdbg2
A. thaliana KBS-Mac-74 ONT ~30x	Genome fraction (%)	99.283	95.393	**99.883**	99.505	99.741	76.317	97.500
	No. of contigs	**25**	448	118	62	57	1382	353
	NG50 (Mb)	11.11	2.61	**13.26**	11.18	7.44	0.28	9.83
	NGA50 (Mb)	5.62	2.20	**9.21**	7.01	5.66	0.27	3.10
	NGA75 (Mb)	3.28	0.38	**4.91**	3.30	2.49	-	0.98
	No. of missaseblies	261	368	653	256	420	**41**	500
	Mismatch fraction (%)	0.298	**0.163**	0.299	0.179	0.325	0.509	0.363
	Indel fraction (%)	1.729	2.247	1.589	**1.414**	1.421	2.574	2.999
	Single-copy genes (%)	75.911	38.226	81.397	**84.252**	83.466	11.920	25.826
	Duplicated genes (%)	0.014	0.009	0.042	0.009	0.014	**0.0**	0.005
	Multi-copy genes (%)	0.0	0.0	**2.083**	0.0	**2.083**	0.0	0.0
	CPU time (h)	4.51	1157.51	22.41	5.99	9.47	**0.64**	19.79
	Memory (GB)	**9.64**	10.57	87.94	21.72	30.46	21.56	15.77
A. thaliana Ler-0 PacBio CLR ~90x	Genome fraction (%)	99.603	99.069	**99.692**	99.300	99.622	22.483	99.275
	No. of contigs	**74**	591	174	155	112	1508	280
	NG50 (Mb)	10.78	0.75	**13.98**	8.68	6.78	-	12.21
	NGA50 (Mb)	6.12	0.75	**6.68**	6.21	6.40	-	6.09
	NGA75 (Mb)	3.07	0.31	**4.55**	1.77	2.34	-	2.74
	No. of missaseblies	792	1189	798	611	833	**22**	728
	Mismatch fraction (%)	0.129	0.219	0.137	**0.107**	0.166	0.371	0.184
	Indel fraction (%)	0.252	0.077	**0.023**	0.231	0.577	2.118	0.279
	Single-copy genes (%)	98.659	98.752	**99.889**	98.632	96.581	8.544	99.174
	Duplicated genes (%)	0.070	0.088	0.028	0.116	0.074	**0.0**	0.023
	Multi-copy genes (%)	72.581	**93.548**	85.484	72.581	38.710	0.0	45.161
	CPU time (h)	22.86	238.86	62.18	25.62	29.06	**0.77**	43.44
	Memory (GB)	18.83	**12.22**	59.68	46.65	32.67	37.44	25.65
D. melanogaster ISO1 ONT ~30x	Genome fraction (%)	92.200	**94.326**	93.023	92.316	88.376	71.756	91.371
	No. of contigs	**148**	664	468	219	232	1852	635
	NG50 (Mb)	6.15	4.56	**19.65**	3.29	1.90	0.10	10.62
	NGA50 (Mb)	1.36	1.23	**1.70**	1.10	1.09	0.10	1.03
	NGA75 (Mb)	0.51	0.46	**0.56**	0.41	0.34	-	0.32
	No. of missaseblies	1230	3167	1316	1098	605	**342**	1974
	Mismatch fraction (%)	**0.163**	0.218	0.164	0.183	0.195	0.456	0.370
	Indel fraction (%)	0.713	0.935	**0.407**	0.737	0.727	1.800	1.556
	Single-copy genes (%)	98.573	98.059	**99.273**	98.219	97.864	63.432	96.083
	Duplicated genes (%)	0.071	0.284	0.035	0.151	0.071	**0.0**	0.027
	Multi-copy genes (%)	52.404	**57.212**	56.731	47.115	21.154	0.962	3.365
	CPU time (h)	5.05	520.75	25.64	7.91	13.71	**0.62**	26.90
	Memory (GB)	**12.86**	13.08	33.37	23.44	26.69	21.45	19.25
D. melanogaster A4 PacBio CLR ~125x	Genome fraction (%)	93.460	**95.967**	92.291	93.709	90.423	91.242	92.830
	No. of contigs	**121**	254	199	299	177	484	311
	NG50 (Mb)	12.83	13.80	15.63	6.54	4.27	3.46	**17.05**
	NGA50 (Mb)	3.92	**9.41**	8.28	3.20	2.55	2.68	4.54
	NGA75 (Mb)	1.21	1.99	**2.20**	1.34	0.77	0.91	1.43
	No. of missaseblies	771	774	609	791	**405**	416	761
	Mismatch fraction (%)	0.047	0.037	0.036	0.058	**0.033**	0.035	0.170
	Indel fraction (%)	0.118	0.041	**0.027**	0.121	0.125	0.135	0.285
	Single-copy genes (%)	99.533	99.023	**99.785**	99.177	99.159	99.196	99.551
	Duplicated genes (%)	0.140	0.897	0.075	0.495	0.196	**0.0**	0.037
	Multi-copy genes (%)	80.447	**92.737**	83.799	86.592	80.447	29.050	59.777
	CPU time (h)	25.54	389.18	75.83	37.87	61.39	**4.35**	20.54
	Memory (GB)	22.18	**19.08**	79.62	56.59	61.99	62.82	19.36
H. sapiens NA12878 ONT ~45x	Genome fraction (%)	92.267	92.037	**92.748**	90.611		91.491	87.356
	No. of contigs	**249**	1145	1264	502		2989	5147
	NG50 (Mb)	27.89	10.58	**31.82**	9.73		3.60	9.80
	NGA50 (Mb)	15.96	8.06	**19.40**	8.03		3.38	5.73
	NGA75 (Mb)	5.90	2.95	**8.49**	3.40		1.33	1.52
	Mismatch fraction (%)	0.135	0.152	**0.128**	0.140		0.151	0.242
	Indel fraction (%)	0.341	0.054	0.359	**0.248**		0.360	0.724
	Yak QV	25.659	35.063	25.479	**27.002**		25.209	22.450
	Single-copy genes (%)	90.285	94.045	90.021	**95.200**		70.849	58.881
	Duplicated genes (%)	0.198	0.252	0.299	0.525		**0.008**	0.016
	Multi-copy genes (%)	48.015	42.772	41.348	**49.139**		7.491	2.247
	Resolved BACs (%)	61.181	44.726	40.084	**63.713**		16.456	8.861
	CPU time (h)	470		1264	1373		**29**	1994
	Memory (GB)	**83**		730	401		391	279

Table 2 Evaluation of long-read assemblers across sequencing technologies.

		ONT					PacBio CLR				PacBio HiFi	
Dataset	Metric	Raven	Canu	Flye	Shasta	Wtdbg2	Raven	Flye	Shasta	Wtdbg2	Raven	hifiasm
H. sapiens CHM13 ONT ~130x PacBio CLR ~50x PacBio HiFi ~35x	Genome fraction (%)	93.392	**94.943**	93.444	92.552	88.668	91.825	**92.121**	91.442	91.783	92.551	99.778
	No. of contigs	**120**	558	548	1236	19029	**897**	2247	2937	3632	1755	470
	NG50 (Mb)	67.58	**79.50**	68.42	41.09	5.29	10.97	**20.83**	12.71	16.76	12.02	88.93
	NGA50 (Mb)	56.59	44.65	**56.77**	28.85	2.34	9.35	**17.45**	11.63	14.57	10.37	80.81
	NGA75 (Mb)	32.08	19.85	**32.20**	12.01	0.59	3.82	**6.04**	3.71	4.17	3.69	36.43
	No. of missaseblies	2847	3885	264	**126**	7046	869	316	**186**	954	2921	156
	Mismatch fraction (%)	0.073	0.117	**0.014**	0.039	0.285	0.036	**0.017**	0.034	0.072	0.059	0.002
	Indel fraction (%)	0.088	0.479	**0.085**	0.351	0.428	0.094	**0.020**	0.254	0.129	0.011	0.001
	Single-copy genes (%)	98.939	93.595	**99.275**	95.823	82.979	98.422	**98.472**	96.570	96.681	98.286	99.908
	Duplicated genes (%)	0.331	0.158	0.150	**0.014**	5.288	0.303	0.247	**0.017**	0.058	0.386	0.061
	Multi-copy genes (%)	**86.217**	49.513	62.547	14.607	34.831	**44.419**	30.262	5.393	6.592	44.644	99.700
	Resolved BACs (%)	**95.518**	88.717	72.798	43.895	30.294	**42.040**	36.785	33.076	35.858	39.104	96.600
	CPU time (h)	**4792**		4855		5978	498	1865	**36**	461	554	
	Memory (GB)	**251**		873		423	**98**	407	547	180	65	
H. sapiens HG002 ONT ~60x PacBio CLR ~80x PacBio HiFi ~35x	Genome fraction (%)	92.691	**94.034**	93.309	93.446	88.866	91.257	**91.707**	89.474	90.882	92.144	96.183
	No. of contigs	**192**	767	776	2039	10166	**2168**	2879	8425	4660	2375	383
	NG50 (Mb)	34.51	32.60	**50.42**	28.92	7.71	3.55	**11.56**	0.91	9.91	6.49	98.17
	NGA50 (Mb)	21.06	20.09	**26.84**	22.73	3.37	2.86	**8.99**	0.89	7.56	5.94	31.43
	NGA75 (Mb)	9.69	7.99	**12.61**	11.42	1.18	1.24	**3.08**	0.34	2.32	2.14	13.09
	Mismatch fraction (%)	0.159	0.222	**0.137**	0.153	0.380	0.143	**0.127**	0.143	0.192	0.185	0.239
	Indel fraction (%)	0.226	0.794	0.220	**0.182**	0.605	0.162	**0.048**	0.361	0.200	0.036	0.031
	Yak QV	28.032	21.887	28.221	**29.179**	24.307	29.455	**37.424**	25.647	29.464	42.265	48.675
	Single-copy genes (%)	97.833	88.954	98.225	**98.522**	85.743	96.725	**97.645**	90.573	93.367	97.570	99.244
	Duplicated genes (%)	0.603	0.547	0.481	**0.147**	2.730	0.397	0.322	**0.022**	0.042	0.481	0.297
	Multi-copy genes (%)	**70.936**	28.390	56.704	27.865	15.655	**26.742**	18.652	4.045	5.169	38.727	85.243
	CPU time (h)	1157		1962	**128**	2191	987	3586	**34**	544	527	
	Memory (GB)	**105**		951	771	352	**129**	562	567	207	67	
H. sapiens HG00733 ONT ~80x PacBio CLR ~95x PacBio HiFi ~35x	Genome fraction (%)	92.511	**94.043**	92.716	92.904	89.176	**92.341**	92.334	92.071	90.796	91.960	96.089
	No. of contigs	**262**	778	1028	1953	4848	**559**	1589	2281	2863	2176	657
	NG50 (Mb)	33.32	**40.63**	37.74	18.43	13.95	22.45	26.53	14.03	**29.05**	7.12	68.31
	NGA50 (Mb)	18.32	22.51	**23.87**	13.47	8.23	17.27	18.00	12.21	**19.38**	6.09	29.94
	NGA75 (Mb)	8.29	**9.49**	9.35	5.36	2.40	**7.29**	6.84	4.28	6.43	2.16	12.83
	Mismatch fraction (%)	**0.128**	0.205	0.129	0.131	0.272	0.131	**0.110**	0.175	0.165	0.157	0.221
	Indel fraction (%)	0.347	0.677	0.405	**0.211**	0.715	0.142	**0.041**	0.381	0.233	0.033	0.031
	Yak QV	25.705	22.635	24.978	**27.979**	22.772	29.758	**37.310**	25.143	28.345	40.056	42.390
	Single-copy genes (%)	97.145	91.406	96.814	**97.650**	88.484	98.411	**98.672**	96.045	96.442	97.584	99.333
	Duplicated genes (%)	0.417	0.714	0.278	**0.069**	0.253	0.503	0.261	**0.039**	0.053	0.489	0.367
	Multi-copy genes (%)	**53.783**	34.307	41.798	14.457	4.719	**56.929**	36.479	5.543	11.461	37.154	88.689
	Resolved BACs (%)	**71.053**	67.895	42.632	26.842	15.790	**48.421**	34.211	22.105	29.474	22.105	80.000
	CPU time (h)	1234		2871	**98**	1895	1522	6473	**115**	1491	486	
	Memory (GB)	**131**		546	870	345	**138**	663	1012	340	70	

Table 3 Raven plant assemblies. Values in brackets represent assembly metrics in corresponding publications. Oryza genomes in the original publication were additionally polished with Illumina reads.

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)	6.35 (7.28)	5.52 (3.80)	2.48 (2.13)	8.14 (6.32)	11.86 (10.53)
No. of contigs	252 (244)	410 (544)	546 (615)	116 (188)	107 (116)
% complete BUSCOs	74.783 (74.300)	85.936 (79.700)	47.150 (53.800)	92.503 (97.600)	92.193 (97.000)
CPU time (h)	40.52 (261.40)	58.55 (315.70)	94.95 (245.60)	43.59 (N/A)	33.89 (N/A)