



Rambler resolves complex repeats in human Chromosomes 8, 19, and X

Sakshar Chakravarty, Glennis Logsdon and Stefano Lonardi

Genome Res. published online March 4, 2025

Access the most recent version at doi:[10.1101/gr.279308.124](https://doi.org/10.1101/gr.279308.124)

P<P	Published online March 4, 2025 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

16 Introduction

17 Given the broad biological impact of obtaining the genome for a new organism, *de novo* genome assembly is
18 one of the most critical problems in computational biology. Despite tremendous algorithmic progress, the
19 problem is not yet completely solved. The assembly problem remains challenging due to the high repetitive
20 content of eukaryotic genomes, short read length, uneven sequencing coverage, non-uniform sequencing errors
21 and chimeric reads. Repetitive regions (or segmental duplications) are the primary reasons for which *de novo*
22 genome assemblies are often fragmented and incomplete. A large fraction of eukaryotic genomes is made
23 of repetitive elements, including satellite DNA, mini-satellites, micro-satellites, and DNA/RNA transposons
24 (Mrázek et al. 2007; Jurka et al. 2007; Treangen et al. 2009; Bustos et al. 2022). For instance, Supplemental
25 Figure 1 illustrates the frequency of repeats present in seven plant species (Chan et al. 2015). Several
26 studies have shown that expansions or mutations of repetitive regions are linked to a variety of human
27 diseases, ranging from neurological diseases to cancers (see (X Liao et al. 2023) for a review). Although
28 many repeats were considered non-functional, they have been shown to impact gene expression, contributing
29 to genetic disorders (Hannan 2018; Ishiura et al. 2019; Shah et al. 2023).

30 The third generation of sequencing technology on the market, e.g., Pacific Biosciences (PacBio) (Eid
31 et al. 2008; Qin et al. 2012; Roberts et al. 2013; Huddleston et al. 2014; KE Kim et al. 2014) and Oxford
32 Nanopore (ONT) (Clarke et al. 2009; Quick et al. 2014; Ashton et al. 2015; Loose et al. 2016), offers longer
33 reads at a higher cost per base than the second generation, but the sequencing error rate is much higher. The
34 introduction of PacBio HiFi sequencing at the end of 2019 has been a “game-changer” in genome assembly,
35 because it can produce read lengths typically ranging 10-25 kb with accuracy greater than 99.8% (Wenger
36 et al. 2019). HiFi sequencing greatly improved human assemblies (Nurk, Walenz, et al. 2020; Porubsky et al.
37 2021; Shumate et al. 2020; Garg et al. 2021). The telomere-to-telomere human genome sequencing project
38 took advantage of PacBio HiFi and ultra-long ONT reads to close most of the repetitive gaps and achieved
39 99.9% completeness (Miga et al. 2020; Logsdon et al. 2021; Hoyt et al. 2022; Nurk, Koren, et al. 2022;
40 Rautiainen et al. 2023). In particular, the method developed to assemble human Chromosome 8 depended
41 on the use of single-copy *k*-mers (hereafter called *unikmers*, also known as SUNKs in (Logsdon et al. 2021)
42 or SUNs in (Sudmant et al. 2010)) to resolve repetitive regions.

43 The problem of reconstructing repetitive regions (segmental duplications) has been addressed several
44 times in the literature, e.g., (Chaisson et al. 2017; Vollger et al. 2019; Bzikadze and Pevzner 2020). The
45 Segmental Duplication Assembler (SDA) by (Vollger et al. 2019) is no longer maintained. More recently,
46 (Bzikadze and Pevzner 2020) proposed CentroFlye to address the problem of reconstructing human cen-
47 tromeres. CentroFlye is a specialized assembler that uses error-prone ONT or PacBio CLR reads. It also
48 requires additional information such as higher-order repeats (HORs) or monomers. This limits its applica-
49 bility to species for which HORs or monomers are known. In addition, CentroFlye is very demanding in
50 terms of computational resources. While SDA, CentroFlye and the assembler for human Chromosome 8 by
51 (Logsdon et al. 2021) are specialized tools for reconstructing segmental duplications or centromeres, they
52 were designed for error-prone long reads. To the best of our knowledge, there is no specialized assembler for
53 reconstructing repeats that uses PacBio HiFi reads exclusively.

54 In this work we introduce RAMbler, a reference-guided assembler that takes advantage of single-copy
55 *k*-mers to resolve complex repetitive regions. We show that RAMbler can resolve complex repeats in human
56 Chromosomes 8, 19, and X from HiFi data to a quality comparable to the telomere-to-telomere (T2T) human
57 genome assembly. Due to the lack of specialized HiFi assemblers for repeats, we compare RAMbler against

58 four general-purpose state-of-the-art assemblers, namely hifiasm (Cheng et al. 2021), LJA (Bankevich et al.
59 2022), HiCANU (Nurk, Walenz, et al. 2020) and Verkko (Rautiainen et al. 2023) on more than 250 synthetic
60 data sets and five real *H. sapiens* data sets.

61 Results

62 Experimental Results on *Homo sapiens*

63 We used three human genome assemblies as a reference, namely GRCh38.p13 (hereafter called HG38), T2T-
64 CHM13.v2.0 (T2T), and the maternal strand of HG002 (MAT002). We used PacBio HiFi reads from four
65 different cell lines, namely CHM13 (from the T2T project), HG002, HG00733, and HG01346 (these last three
66 from the human pan-genome project (WW Liao et al. 2023)). Supplemental Table 1 reports the accession
67 numbers and statistics for these human data sets, while Supplemental Table 2 summarizes accession numbers
68 and statistics for *S. cerevisiae* (used in our simulation studies).

69 We used RAmbl to assemble some of the repetitive regions in the reference assemblies HG38, T2T
70 and MAT002, using various sets of HiFi reads. We used the following naming convention to identify the
71 assemblies produced by RAmbl. Hereafter, a RAmbl assembly using reference assembly *G* and HiFi
72 reads *H* will be denoted by RA.*G.H*. For example, the RAmbl assembly using the HG38 assembly and
73 the HG00733 HiFi reads will be called RA.HG38.HG00733.

74 We focused on five complex repetitive regions within the human genome: the centromeres of Chromo-
75 some 8 and X, and three non-centromeric regions from Chromosome 19. While the selection of these regions
76 was somewhat arbitrary, it was motivated by a few factors: (1) these regions were over-collapsed in the HG38
77 assembly (Figure 1A), (2) Chromosomes 8, 19, and X have no unplaced contigs within the HG38 assembly,
78 and (3) Chromosome 19 contains a few unresolved non-centromeric repeats (Supplemental Figure 2). Our
79 intent was to show that using newer HiFi reads, RAmbl could improve some of the over-collapsed repetitive
80 regions, in particular in the HG38 assembly.

81 We carried out five experiments on human data sets. The first two experiments were aimed at demon-
82 strating RAmbl’s ability to resolve repeats without manual curation. For these two experiments, we used
83 as inputs (1) the T2T assembly with CHM13 HiFi reads, and (2) the MAT002 assembly with HG002 HiFi
84 reads. The other three experiments focused on RAmbl’s performance. For these three, we used as input
85 the HG38 assembly and HiFi reads from (3) HG002, (4) HG00733, and (5) HG01346.

86 We carried out these five experiments on the centromeric regions of Chromosomes 8 and X, as well as
87 three non-centromeric regions on Chromosome 19. We ran RAmbl, hifiasm, LJA, HiCANU, and Verkko
88 on the subset of HiFi reads mapped to these regions. SDA was excluded because “*is no longer maintained
89 and should not be used . . . assembly tools like Flye, HiCanu, and hifiasm outperform any results previously possible
90 with SDA.*” (quote from <https://github.com/mrvollger/SDA>). CentroFlye was also excluded because
91 we were unable to run it on our 2.8 GHz 32-core processor server with 512GB of RAM. We tried first to
92 reconstruct the centromere of Chromosome X using the HiFi reads and the HORs provided by the authors,
93 but CentroFlye failed during the error-correction step. When we tried to reconstruct the same centromere
94 using the CHM13 ONT reads from (Bzikadze and Pevzner 2020), CentroFlye ran out of memory. The
95 instructions claim that CentroFlye requires about 800GB of RAM to complete the assembly of Chromosome
96 X.

Chr	Input assembly	HiFi reads	Depth	Repetitive region		# Selected reads	Selected reads sum (Mb)	Expected size (Mb)		Assembler				
				start (Mb)	end (Mb)					Rambler	hifiasm	LJA	HiCANU	Verkko
8	HG38	HG002	67.5×	43.5	46.5	15,620	249.038	3.689	# contigs	2	3	39	-	66
									Total (bp)	3,885,290	3,632,973	7,573,022	-	8,186,622
									Longest contig	3,627,559	1,843,336	2,701,877	-	1,374,826
		HG00733	32.8×	43.5	46.5	7,241	99.356	3.029	# contigs	1	3	74	44	103
									Total (bp)	3,805,715	3,692,973	6,653,621	6,279,128	7,267,808
									Longest contig	3,805,715	3,483,042	1,384,931	1,002,430	1,710,467
	HG01346	26×	42.0	48.0	9,579	177.120	6.812	# contigs	3	5	91	73	134	
								Total (bp)	7,553,039	6,637,843	12,660,338	13,840,156	13,671,832	
								Longest contig	6,229,877	6,213,919	4,017,234	6,043,945	2,981,689	
	T2T	CHM13	32.4×	43.5	46.5	5,591	101.370	3.129	# contigs	1	1	1	10	1
									Total (bp)	3,040,965	3,040,965	3,040,948	3,343,237	3,032,370
									Longest contig	3,040,965	3,040,965	3,040,948	3,033,722	3,032,370
MAT002	HG002	66×	43.5	47.0	14,393	229.495	3.477	# contigs	1	3	26	-	51	
								Total (bp)	3,271,813	3,309,709	6,861,347	-	7,415,002	
								Longest contig	3,271,813	1,776,606	2,701,877	-	1,374,826	
X	HG38	HG002	34.2×	57.5	63.0	9,665	155.687	4.552	# contigs	1	1	1	1	N/A
									Total (bp)	4,678,725	4,678,719	4,663,887	4,669,944	-
									Longest contig	4,678,725	4,678,719	4,663,887	4,669,944	-
		HG00733	31.8×	57.25	64.0	15,333	209.268	6.581	# contigs	2	3	114	103	170
									Total (bp)	5,653,597	5,279,187	12,115,053	13,853,128	13,649,195
									Longest contig	3,813,195	3,408,909	1,675,295	2,970,002	3,984,187
	HG01346	25.5×	57.25	64.0	9,679	179.004	7.020	# contigs	1	5	74	92	140	
								Total (bp)	7,378,373	7,072,503	12,305,066	14,376,243	13,718,592	
								Longest contig	7,378,373	3,232,572	4,247,266	3,280,524	4,251,271	
	T2T	CHM13	34×	57.5	62.0	8,488	153.883	4.526	# contigs	1	2	3	22	10
									Total (bp)	4,468,109	4,533,307	4,536,587	5,081,155	4,575,146
									Longest contig	4,468,109	4,486,438	4,074,724	2,016,666	1,940,381
MAT002	HG002	33×	57.5	62.0	9,362	150.793	4.569	# contigs	2	2	3	-	4	
								Total (bp)	4,566,785	4,566,777	4,542,009	-	4,546,175	
								Longest contig	4,542,010	4,542,004	4,180,812	-	4,141,556	

Table 1: Comparison of the assemblies produced by Rambler, hifiasm, LJA, HiCANU and Verkko on the HiFi reads extracted around the centromeres of human Chromosomes 8 and X, for several choices of input assemblies and HiFi reads. Note: '-' means that the assembler timed out after 96 hours, 'N/A' means that the assembler finished execution but produced no output.

Reference-guided repeat assembly with RAMbler

97 The comparative results for RAMbler, hifiasm, LJA, HiCANU, and Verkko on the centromeric regions of
98 Chromosomes 8 and X are summarized in Table 1. Columns 2, 3, and 4 indicate the input assembly, the HiFi
99 sample, and the chromosome-level sequencing depth, respectively. Columns 5 and 6 indicate the boundaries
100 of the regions to be re-assembled. Columns 7 and 8 indicate the number of HiFi reads mapped to the target
101 region and the total number of bases in these reads, respectively. In Column 9, we show RAMbler’s estimate
102 of the size of the repetitive region, obtained by computing the ratio between the total number of bases of
103 the selected HiFi reads and the average chromosome-level coverage depth. While it has been reported that
104 repetitive regions can affect HiFi sequencing coverage (Nurk, Koren, et al. 2022), we determined that this
105 coverage bias would have a small impact on RAMbler’s estimation of the length of the repetitive region
106 to be assembled. When we masked all repetitive regions from the human genome, the changes in HiFi
107 coverage was relatively small. Supplemental Table 3 shows that on Chromosomes 8, 19 and X, the change
108 in coverage after repeat masking was smaller than 4%. Overall, the variations in coverage were under 9% at
109 the chromosome-level and under 3% at the genome level. The largest change was observed for Chromosome
110 17 (8.57% variation in coverage using the HG002 HiFi reads). Such a coverage fluctuation would mean
111 that, in the worst case, a repetitive region with ten copies of the monomer could be underestimated by
112 one copy. Columns 11-15 show the assembly statistics for RAMbler, hifiasm, LJA, HiCANU, and Verkko.
113 Observe that RAMbler consistently produced the least fragmented assemblies, and in most cases the longest
114 contig produced by RAMbler was the longest among all assemblers. In many cases the total assembly size
115 generated by RAMbler was close to the expected assembly size, although hifiasm also produced assemblies
116 whose total size was also consistent with the expectation. Similar observations could be made for the three
117 non-centromeric regions on Chromosome 19 (Supplemental Table 4 for T2T and MAT002, Supplemental
118 Table 5 for HG38).

119 Unlike *de novo* assemblers which consistently produce the same outputs for the same set of input reads,
120 RAMbler’s assembly depends on the reference used as it relies on the input assembly to identify the reads
121 that belong to a repetitive region. Observe that in Table 1 we used the same set of HiFi reads (HG002) with
122 two reference assemblies, namely HG38 and MAT002. Despite the difference between HG38 and MAT002,
123 in particular in the centromeric regions, the analysis in Supplemental Table 6 shows that more than 85% of
124 the selected reads were shared. As a result, the RAMbler assemblies are relatively consistent. For instance,
125 on the centromere of Chromosome X, the length difference for the longest contigs was about 3%. On the
126 centromere of Chromosome 8, the length difference for the longest contigs was approximately 10%.

127 While the assembly contiguity statistics reported in Table 1, Supplemental Tables 4 and 5 are important,
128 they did not tell us whether these assemblies had high quality. In order to quantitatively assess the assembly
129 quality, we used a two-pronged approach. First, we used the manually-curated T2T assembly as the ground
130 truth and we measured the agreement between the RAMbler assemblies and the T2T assembly using QUASt
131 (Gurevich et al. 2013) and SyRI (Goel et al. 2019). This approach, however, can be problematic because the
132 HiFi reads used for the RAMbler assemblies did not originate from the cell line used to generate the T2T
133 assembly. Human centromeres are known to exhibit high variation across individuals (Altemose et al. 2022).
134 For instance, the NucFreq coverage plots in Supplemental Figure 2 show coverage spikes in the centromeric
135 regions of the T2T assembly when the HG01346 reads were mapped to it. As a consequence some level
136 of divergence would be expected in RAMbler’s centromeric assembly when compared to the T2T assembly.
137 In order to obtain both qualitative and quantitative assessments of RAMbler’s improvements, we measured
138 the assembly quality using CRAQ (K Li et al. 2023) and NucFreq (Vollger et al. 2019). CRAQ is a tool
139 that measures assembly quality without the need of a reference; in particular, it can detect local and global

Reference-guided repeat assembly with Rambler

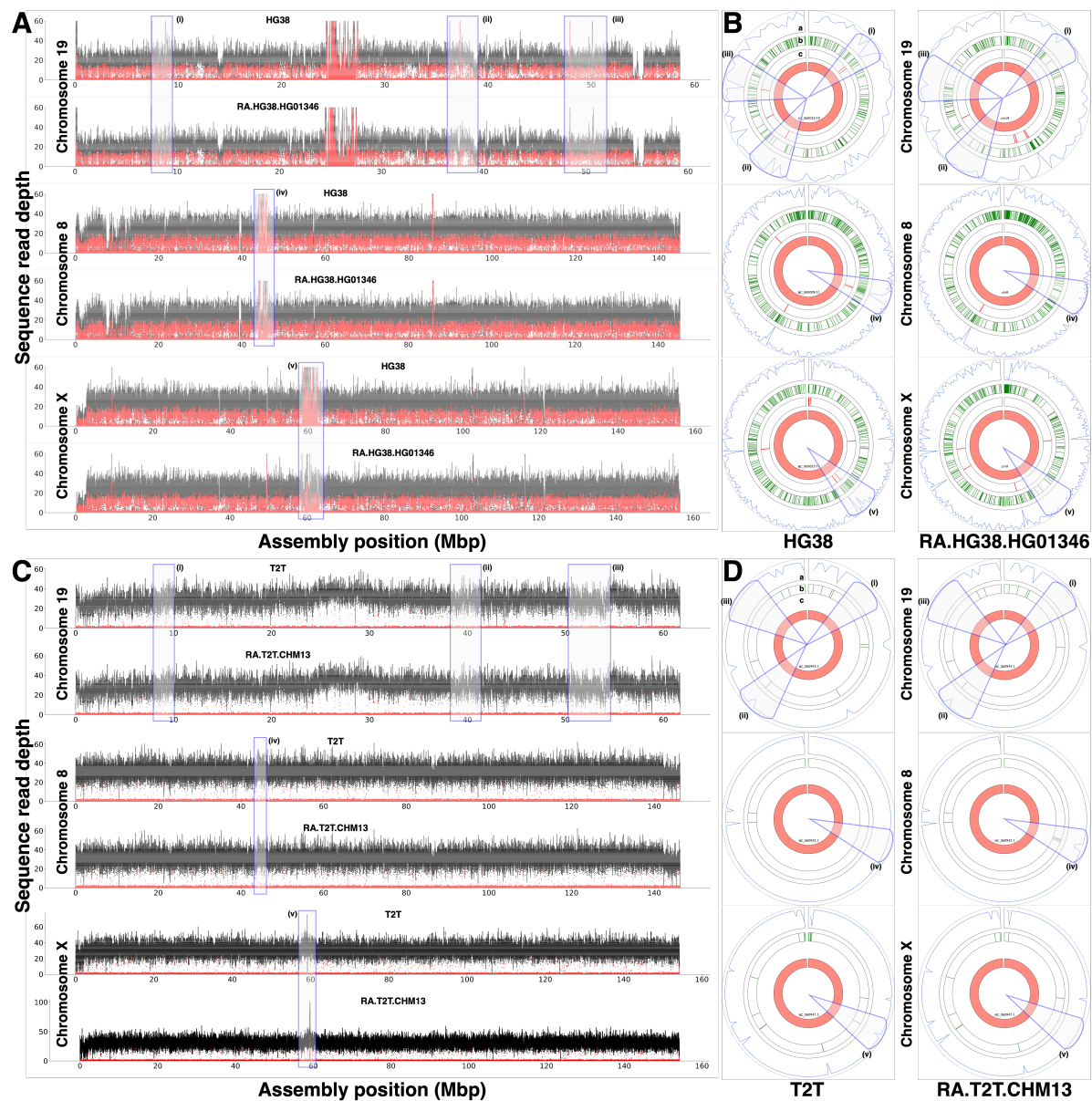


Figure 1: Comparing Rambler's assemblies (RA.HG38.HG01346 and RA.T2T.CHM13) against (A, B) the GRCh38.p13 assembly (HG38) and (C, D) the T2T-CHM13.v2.0 assembly (T2T) of human Chromosomes 8, 19, and X; in all plots, blocks (i), (ii), and (iii) are non-centromeric repeats in Chromosome 19; blocks (iv) and (v) are the centromeric regions of Chromosomes 8 and X, respectively; (A, C) NucFreq plots illustrating HiFi read mapping coverage (clipped at 60×); (B, D) CRAQ Circos plots illustrating (a) assembly quality index (AQI) score (higher is better), (b) base errors (fewer is better), and (c) misjoins (fewer is better)

Reference-guided repeat assembly with RAmbler

140 assembly errors based on the alignment information of mapped reads (long and short). CRAQ was given
141 in input Illumina reads (which were not used in the assembly) and HiFi reads (which were used for the
142 assembly). While we expected a reduction in coverage spikes in the repetitive regions of the NucFreq plots
143 because those HiFi reads were also used in the assembly, we were interested in measuring the extent of this
144 reduction. Additionally, we used NucFlag (an extension of NucFreq) to flag assembly errors or collapses
145 along with the coverage plots.

146 To establish RAmbler’s assembly quality, we carried out two sets of experiments. In the first set, we show
147 that RAmbler can drastically improve the HG38 assembly in the five repetitive regions to a QUASt-based
148 quality comparable to the T2T assembly using HG002, HG00733, or HG01346 HiFi reads. In the second set,
149 we show that if RAmbler was given the same set of HiFi reads used for the T2T or the MAT002 assemblies,
150 it would produce an assembly of the five repetitive regions that matches, if not exceeds, the quality of the
151 corresponding reference assemblies.

152 Figure 1A-B summarizes the results of one run from the first set of experiments using HG01346 HiFi
153 reads. Figure 1A shows the HiFi coverage. The coverage depth in Figure 1A is clipped at 60× for better
154 visualization, whereas Supplemental Figure 3 shows full-scale plots for HG38 and the RAmbler assembly
155 RA.HG38.HG01346. Observe that on Chromosome 19, RA.HG38.HG01346 has a more uniform coverage
156 across the three repetitive regions compared to HG38, in particular for regions (i) and (ii). Also observe
157 that (1) while the coverage on region (iv) of Chromosome 8 of HG38 is $\sim 600\times$, it reduces to $\sim 100\times$ in
158 RA.HG38.HG01346 indicating a partial resolution of the repeat; (2) while the coverage on region (v) of
159 Chromosome X of HG38 is $\sim 800\times$, it reduces to $\sim 30\times$ in RA.HG38.HG01346 indicating a full resolution of
160 the repeat. From a purely coverage viewpoint, these results indicate that RA.HG38.HG01346 is significantly
161 improved compared to HG38. Supplemental Figure 4A illustrates the NucFlag plots for the five regions
162 analyzed. Observe that NucFlag indicates much fewer errors in the RA.HG38.HG01346 assembly compared
163 to HG38.

164 Figure 1B shows the results of the CRAQ analysis (K Li et al. 2023). We used HG01346 HiFi reads and
165 HG01346 Illumina reads ($\sim 10\times$ coverage) to evaluate the quality of HG38 and RA.HG38.HG01346. The
166 Circos plots (Krzyszowski et al. 2009) in Figure 1B illustrate the (a) assembly quality indicators (AQI) score
167 (blue curve, higher is better, see Section “Performance Metrics for Real Data”, (b) base errors (green bands,
168 fewer is better), and (c) misjoins (red bands, fewer is better). Cones (i)-(v) highlight the repetitive regions.
169 On Chromosome 19, observe that (1) the assembly quality index (AQI) score for RA.HG38.HG01346 is higher
170 than the AQI score for HG38 on all three regions; (2) RA.HG38.HG01346’s base errors are significantly better
171 than HG38; (3) CRAQ reports four misjoins in HG38, and zero in RA.HG38.HG01346. Also observe that on
172 the centromeres of Chromosome 8 and Chromosome X, the AQI scores for RA.HG38.HG01346 are higher than
173 the AQI scores for HG38. On Chromosome 8, HG38 has two misjoins, and RA.HG38.HG01346 has none.
174 RA.HG38.HG01346 also has fewer base errors. On Chromosome X, RA.HG38.HG01346 has no misjoins
175 (HG38 has two) with fewer base errors than HG38. The NucFreq, the Circos/CRAQ and the NucFlag plots
176 for HiFi reads HG002 and HG00733 are shown in Supplemental Figures 5 and 6. The coverage plots and the
177 Circos plots from both of these runs demonstrate similar results as shown in Figure 1A-B. Table 2 reports all
178 the statistics produced by CRAQ, which clearly indicates that RA.HG38.HG002, RA.HG38.HG00733 and
179 RA.HG38.HG01346 are improved compared to HG38.

180 Supplemental Table 7 reports the QUASt metrics for HG38, RA.HG38.HG002, RA.HG38.HG00733 and
181 RA.HG38.HG01346, using the T2T assembly as the ground truth and for HG38 and RA.HG38.HG002,
182 using the MAT002 assembly as the ground truth. When T2T was used as the reference, (1) the number

Reference-guided repeat assembly with RAmbler

Chromosome	Metric	HG002		HG00733		HG01346	
		HG38	RA.HG38.HG002	HG38	RA.HG38.HG00733	HG38	RA.HG38.HG01346
19	Coverage (%)	90.57	94.25	89.95	94.16	90.32	94.05
	R-AQI	77.33	78.03	75.96	76.45	67.66	68.92
	S-AQI	81.29	82.12	92.69	89.81	89.29	91.41
8	Coverage (%)	92.96	98.19	92.47	97.34	92.41	97.72
	R-AQI	82.11	83.79	80.90	83.81	79.16	79.22
	S-AQI	89.48	92.60	94.21	96.54	98.52	98.60
X	Coverage (%)	94.02	98.31	93.91	97.95	94.06	97.71
	R-AQI	81.52	81.76	86.66	86.17	83.60	82.85
	S-AQI	92.15	93.66	94.69	94.86	96.32	97.42

Table 2: Comparing RAmbler’s assemblies (RA.HG38.HG002, RA.HG38.HG00733, and RA.HG38.HG01346) for selected regions of human Chromosomes 19, 8, and X against the GRCh38.p13 assembly (HG38) using CRAQ; CRAQ reports assembly coverage rate (i.e., the fraction of the genome assembled), Regional Assembly Quality Indicator (R-AQI) score (higher is better) and Structural Assembly Quality Indicator (S-AQI) score (higher is better); CRAQ was provided Illumina reads ($\sim 13.5\times$ coverage) and long PacBio HiFi reads ($\sim 66\times$ coverage) for HG002; Illumina reads ($\sim 12\times$ coverage) and long PacBio HiFi reads ($\sim 33\times$ coverage) for HG00733; Illumina reads ($\sim 10\times$ coverage) and long PacBio HiFi reads ($\sim 26\times$ coverage) for HG01346; numbers in bold indicate the best scores

183 of misassemblies on Chromosome 8 decreased from 371 in HG38 to 239-278 in the RAmbler assemblies; on
 184 Chromosome 19, they decreased from 101 in HG38 to 92-100; on Chromosome X, they decreased from 374 in
 185 HG38 to 205-250, (2) the genome fraction for Chromosome 19 increased from about 90.7% to about 91.5%;
 186 on Chromosome X, it increased from about 98.6% to about 99%; on Chromosome 8, it remained the same
 187 around 98%. When MAT002 was used as the reference, (1) the number of misassemblies on Chromosome 8
 188 decreased from 382 in HG38 to 205 in RA.HG38.HG002; on Chromosome 19, they reduced to 102 from 119;
 189 on Chromosome X, they decreased from 382 in HG38 to 169, (2) the genome fraction for Chromosome 8
 190 increased from 97.5% to 98.4%; on Chromosome 19, it improved from 91.2% to 92.3%; on Chromosome X,
 191 it increased from 98.6% to 99.6%. Supplemental Figure 7 shows the synteny analysis based on SyRI (Goel
 192 et al. 2019). Observe the much stronger synteny between the three RAmbler assemblies and T2T, compared
 193 to HG38.

194 Figure 1C-D summarizes the results of one of the runs from the second set of experiments obtained by
 195 RAmbler using the PacBio HiFi reads that were used for the T2T project ($\sim 32.4\times$ coverage). Hereafter,
 196 this RAmbler assembly is called RA.T2T.CHM13. Observe in Figure 1C that T2T and RA.T2T.CHM13
 197 have nearly identical HiFi coverage across all five regions, which is also reflected in the NucFlag analysis in
 198 Supplemental Figure 4B. Figure 1D illustrates the CRAQ assessments based on the alignment of CHM13
 199 HiFi reads and CHM13 Illumina reads. Observe that both T2T and RA.T2T.CHM13 have similar AQI scores
 200 across all three chromosomes with no misjoins, except for three base errors introduced in the centromere of
 201 Chromosome 8 and one base error corrected in the Chromosome X by RAmbler (CRAQ numerical scores are
 202 reported in Supplemental Table 8). We also used ultra-long ONT reads (longer than 100 kb, $\sim 17\times$ coverage)
 203 to further evaluate and compare these assemblies. The UL-ONT coverage and CRAQ plot in Supplemental
 204 Figure 8, and the CRAQ numerical scores in Supplemental Table 9 indicate that the RA.T2T.CHM13 has
 205 the same quality as T2T. The NucFreq plots, the Circos plots based on CRAQ (numerical scores are reported
 206 in Supplemental Table 10) and the NucFlag plots for the other run from this set of experiments are shown
 207 in Supplemental Figures 9 and 10. We used the MAT002 assembly as input assembly with the HG002
 208 HiFi reads. The coverage plots and the Circos plots from this run demonstrate similar results as shown in
 209 Figure 1C-D. However, the NucFlag analysis reveals that the RA.MAT002.HG002 assembly had fewer errors
 210 than MAT002 in three out of the five repetitive regions. Supplemental Figure 11A indicates a perfect synteny
 211 between RA.T2T.CHM13 and T2T across Chromosomes 8, 19, and X. Similarly, Supplemental Figure 11B

Reference-guided repeat assembly with RAMbler

212 illustrates a perfect synteny between RA.MAT002.HG002 and MAT002 across Chromosomes 8 and 19, and
 213 a near perfect synteny with a small inversion in Chromosome X.

214 Supplemental Table 11 shows the runtime and the memory consumption for RAMbler to resolve the
 215 centromeres of human Chromosomes 8 and X. Currently RAMbler takes longer than hifiasm, LJA and
 216 Verkko. RAMbler’s memory consumption is reasonable. The current implementation of RAMbler is not
 217 multi-threaded, so there is an opportunity to make it faster and more scalable.

218 **Experimental Results on Synthetic Data**

219 **Synthetic Data Generation and Parameter Optimization**

220 We generated synthetic repetitive regions by selecting a combination of (1) repeat unit size: 10 kb, 15 kb,
 221 20 kb and 25 kb; (2) 2, 5, and 10 copies of the repeat unit; (3) mutation rate in each copy of the repeat
 222 $p = \{1/100, 1/250, 1/500, 1/1000, 1/2000\}$. For each combination, we generated HiFi reads using PBSim
 223 (Ono et al. 2012) on the CCS model with read coverage of $10\times$, $20\times$, $30\times$, and $40\times$. PBSim requires other
 224 parameter values to be set before generating reads, which are provided in Supplemental Table 12.

225 RAMbler has five main parameters (summarized in Supplemental Table 13). Based on the analysis in
 226 Section “Analysis of k -mer Distribution”, we determined that $k = 21$ and $t = 3$. To find the optimal
 227 values for to and th , we conducted a grid search where $to = \{1, 5, 10, 15, 20\}$ and $th = \{5, 10, 15, \dots, 50\}$ (50
 228 combinations).

229 RAMbler was tested on 135 synthetic data sets, obtained from the combinations of different choices of the
 230 repeat unit size $\{10, 15, 20\}$ kb, repeat copies $\{2, 5, 10\}$, mutation rate $p = \{1/100, 1/250, 1/500, 1/1000, 1/2000\}$,
 231 and read coverage depth $\{20\times, 30\times, 40\times\}$. Supplemental Figure 12 shows the experimental results for dif-
 232 ferent metrics (namely number of contigs, number of misassembled contigs, effective genome fraction per
 233 contig and normalized NG50), with the best choices for to and th highlighted in colored rectangles.

234 When considering the number of contigs (Supplemental Figure 12A), observe that RAMbler achieved the
 235 best performance for $to = \{15, 20\}$ and $th = \{10, 15\}$. Regarding the number of misassembled contigs pro-
 236 duced by RAMbler (Supplemental Figure 12B), the best values were $to = \{15, 20\}$ and $th = \{10, 15, 20, 25\}$.
 237 In terms of effective genome fraction per contig (ζ) (Supplemental Figure 12C), RAMbler had better results
 238 with $to = \{10, 15, 20\}$ and $th = \{10, 15, 20\}$. When considering the normalized NG50 (η) (Supplemental
 239 Figure 12D), the best outcomes were obtained with $to = \{15, 20\}$ and $th = 15$. By combining all these
 240 metrics in the assembly score (defined later in Section “Assembly score”), we determined that the optimal
 241 values were $to = 15$ and $th = 15$ (Supplemental Figure 13).

242 The value of mo (minimum overlap for the overlap graph) was set to 1,000 bp in all our experiments
 243 (both on synthetic and real data). In order to ensure that $mo = 1,000$ was sufficiently stringent to avoid
 244 spurious overlaps in the human genome, we carried out the following experiment. (i) We collected the
 245 centromeric regions for all human chromosomes in the T2T assembly (excluding the rDNA regions). (ii)
 246 We extracted 1 kb sequences every 100 bp in these centromeric regions. (iii) We computed an alignment
 247 between all pairs of sequences extracted in Step (ii) using minimap2 (H Li 2018). Out of 2,860,045 sequences
 248 generated in Step (ii), only three had an identical full match to another sequence in the set. There was
 249 (i) a 1 kb sequence on Chromosome 16 starting at position 33,712,533 matching a 1 kb sequence starting
 250 at position 35,108,033, (ii) a 1 kb sequence on Chromosome 16 starting at position 33,713,733 matching
 251 a 1 kb sequence starting at position 35,109,233, and (iii) a 1 kb sequence on Chromosome 3 starting at
 252 position 95,760,872 matching a 1 kb sequence starting at position 95,843,772. This analysis demonstrates

Reference-guided repeat assembly with RAMbler

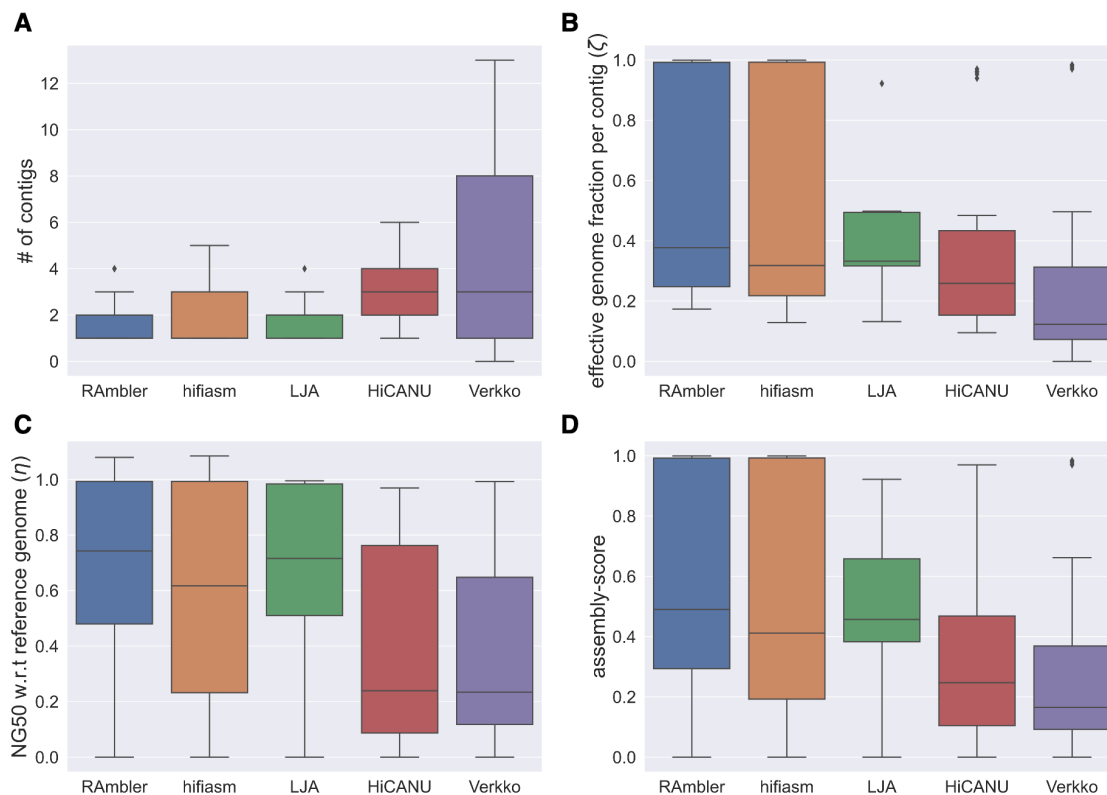


Figure 2: Assembly statistics for RAMbler, hifiasm, LJA, HiCANU, and Verkko for 36 different combinations of synthetic data with repetitive regions having repeat sizes $\{15, 20\}$ kb, number of copies $\{5, 10\}$, mutation rate $p = \{1/250, 1/500, 1/1000\}$ and coverage depth $\{20\times, 30\times, 40\times\}$; **(A)** number of contigs, **(B)** effective genome fraction per contig ζ , **(C)** normalized NG50 η , and **(D)** assembly-score

253 that spurious overlaps that are 1 kb or longer are extremely rare in the human genome, even in repetitive
 254 regions like the human centromeres. We used windows every 100 bp due to the computational cost aligning
 255 pairwise all these sequences. The minimap2 alignment file for 2.86 M sequences was 1.43 TB. The number
 256 of 1 kb sequence every 10 bp would be about 28.6 M, the number of 1 kb sequence every bp would be about
 257 286 M. Since the alignment file grows quadratically with the number of sequences, it would quickly become
 258 infeasible to process. We believe that the conclusions would still hold with $10\times$ or $100\times$ more data.

259 Comparing RAMbler, hifiasm, LJA, HiCANU, and Verkko on fixed-length repeats

260 We conducted an extensive performance comparison of RAMbler with other state-of-the-art HiFi assemblers,
 261 namely hifiasm, LJA, HiCANU, and Verkko, using 36 synthetic data sets. SDA is no longer maintained and
 262 CentroFlye requires additional auxiliary information (HORs and monomers), thus they were both excluded
 263 from this comparison. We generated repetitive regions with two choices of the repeat unit size $\{15, 20\}$ kb,
 264 two choices for the number of copies $\{5, 10\}$, and three values of mutation rate $p = \{1/250, 1/500, 1/1000\}$.
 265 Synthetic HiFi reads were generated using PBSim with coverage depths of $\{20\times, 30\times, 40\times\}$ (see Section
 266 “Synthetic Data Generation and Parameter Optimization” for details).

267 Figure 2 summarizes the results in terms of four different performance metrics. In Figure 2A, we compare
 268 the number of contigs produced by the different assemblers, where a single contig would be the ideal assem-
 269 bly. Observe that RAMbler consistently produced the lowest number of contigs among all the assemblers.

Reference-guided repeat assembly with RAMbler

Synthetic data set			Assembly score				
Repeat unit size	Number of copies	p	RAMbler	hifiasm	LJA	HiCANU	Verkko
15 kb	5	1/250	0.493	0.357	0.327	0.239	0.251
		1/500	0.993	0.545	0.658	0.165	0.204
		1/1000	0.454	0.426	0.346	0.547	0.408
20 kb		1/250	0.422	0.599	0.418	0.441	0.276
		1/500	0.797	0.309	0.618	0.312	0.246
		1/1000	0.896	0.213	0.395	0.206	0.294
25 kb		1/250	0.637	0.486	0.636	0.227	0.354
		1/500	0.773	0.173	0.970	0.254	0.471
		1/1000	0.465	0.231	0.409	0.278	0.345

Table 3: Assembly score results for nine different data sets of synthetic HiFi reads with coverage $30\times$, based on a repetitive region with a variable-length (up to $\pm 5\%$ per copy) repeat unit of length $\{15, 20, 25\}$ kb, and $p = \{1/250, 1/500, 1/1000\}$; numbers in bold indicate the best assembly score on each row

270 Figures 2B, C, and D show the results in terms of effective genome fraction per contig (ζ), normalized NG50
 271 (η), and assembly score (numerical scores are in Supplemental Table 14), respectively. For all these metrics,
 272 the best assembly is the one that gets closer to 1.0. Observe again that RAMbler achieved higher values on
 273 all these metrics compared to the other assemblers.

274 Comparing RAMbler, hifiasm, LJA, HiCANU, and Verkko on variable-length repeats

275 To evaluate RAMbler’s ability to resolve tandem repeats in the presence of variable-length repeat units, we
 276 created synthetic repetitive regions in which each repeat copy can vary up to $\pm 5\%$ of the length of a repeat
 277 unit. We used five copies of repeat units of $\{15, 20, 25\}$ kb with mutation rates $p = \{1/250, 1/500, 1/1000\}$.
 278 Synthetic HiFi reads with a coverage depth of $30\times$ were generated with PBSim. Table 3 summarizes assembly
 279 score results for these nine data sets. RAMbler outperformed the other assemblers in six out of nine runs.
 280 Supplemental Figure 14 shows that RAMbler, hifiasm and LJA produce contigs with either zero or one
 281 misassembled contig. It also shows that RAMbler, hifiasm and LJA produce more contiguous assemblies
 282 than HiCANU and Verkko. In general, while HiCANU and Verkko rarely introduce misassemblies, they
 283 produce more fragmented assemblies than RAMbler, hifiasm and LJA. This is reflected by the RAMbler’s
 284 best assembly score in Supplemental Figure 15.

285 Comparing RAMbler, hifiasm, LJA, HiCANU, and Verkko on repetitive regions with copy number variation

286 Note that RAMbler was not designed to produce a haplotype-resolved assembly. An important question is
 287 what assembly would RAMbler produce in case there are copy number variations between the two haplotypes.
 288 To address this question we carried out several experiments on synthetic diploid genomes, as follows. (i)
 289 We created a synthetic repetitive region where the primary haplotype (hap1) contained either 5 or 10
 290 copies of a repeat unit. Each repeat unit was $\{5, 10, 15, 20, 50, 100\}$ kb long, with a mutation probability of
 291 $\{1/250, 1/500, 1/1000\}$. (ii) We produced the secondary haplotype (hap2) as follows. When the hap1 had 5
 292 copies, hap2 had $\{3, 4, 5\}$ copies. If hap1 had 10 copies, hap2 had $\{6, 8, 10\}$ copies. (iii) We added a 50 kb
 293 sequence upstream and downstream of the repetitive region on both hap1 and hap2. (iv) We used PBSim
 294 to generate $30\times$ -coverage HiFi reads from these $2 \times 6 \times 3 \times 3 = 108$ synthetic diploid genomes (hap1+hap2).
 295 (v) We assembled the synthetic HiFi reads using RAMbler, hifiasm, LJA, HiCANU, and Verkko.

296 The goal of these experiments was to evaluate the quality of the assemblies produced by RAMbler, hifiasm,
 297 LJA, HiCANU, and Verkko on these synthetic diploid genomes using various metrics. We recorded the total

Reference-guided repeat assembly with RAmbler

number of contigs, the overall assembly size, the number of resolved repeat copies, and the number of contigs with haplotype switching. On these synthetic HiFi datasets with minimal divergence between haplotypes, LJA, HiCANU, and Verkko are expected to produce phased assemblies, but not fully haplotype-resolved assemblies due to the limited length of HiFi reads. They typically require either long-range sequencing data (e.g., ultra-long ONT or Hi-C reads) or phasing data (e.g., maternal and paternal reads) for full haplotype resolution. In HiFi-only mode, LJA, HiCANU, and Verkko generate a single FASTA file containing both haplotypes. hifiasm instead, attempts to resolve haplotypes even in HiFi-only mode by joining phased assembly blocks to achieve greater contiguity. It produces separate FASTA files for primary and alternate assemblies. Thus, the assemblies produced by LJA, HiCANU, and Verkko are expected to have a smaller number of haplotype switches, but more fragmented than hifiasm. Since LJA, HiCANU, and Verkko are expected to resolve repeat copies from both haplotypes, their total assembly size is expected to be close to the sum of the two haplotype lengths. Instead, since hifiasm assigns phased blocks to either the primary or the alternate assembly, it is expected to incur more haplotype switch errors, and generate pseudo-haplotypes that capture hap1 as primary and hap2 as alternate (or vice versa). RAmbler also prioritizes long contig generation and outputs a single FASTA file representing the primary pseudo-haplotype. This assembly is expected to capture hap1, which contains the larger repeat copy count, possibly including some haplotype switches.

We used BLAST (McGinnis and Madden 2004) to align all the repeat units in hap1 and hap2 to each target assembly. We defined a repeat unit R to be *resolved* by an assembly A if the BLAST output indicated a perfect identity (i.e., 99.99% or higher) of R in A . Any alignment with less than perfect identity was disregarded. We also recorded whether any contig of a target assembly contained a mix of repeat units from both haplotypes, which indicated switching errors.

The experimental results are summarized in Supplemental Tables 15–18. We omitted Verkko from the tables because it failed to produce any output in most cases. Verkko failed 92 runs out of 108 for unknown reasons (the logs were uninformative). Supplemental Tables 15 and 16 report the number of contigs and the total assembly size. Supplemental Tables 17 and 18 report the number of resolved repeat copies and the number of contigs containing haplotype switching errors. In Supplemental Tables 17 and 18, red numbers indicate incorrect repeat copy counts (neither matching either haplotype copy counts nor their sum), blue numbers indicate copy counts that match the sum of copies for both haplotypes (hap1+hap2), and grey cells indicate assemblies with haplotype switches.

Supplemental Tables 15 and 17 show the experimental results when hap1 had 5 copies and hap2 had 3 – 5 copies. For these data sets, RAmbler produced an assembly containing 5 copies of the repeat unit in 47 cases out of 54 (87%) with 8 haplotype switches. hifiasm produced the correct number of copies for hap1 in 48 cases out of 54 (89%), and the correct number of copies for hap2 in 46 cases out of 54 (85%) with 7 and 1 haplotype switches, respectively. RAmbler however, always produced a single contig, while hifiasm produced 2-3 contigs in some cases (the average number of contigs was 1.09 for both primary and alternate assembly). LJA resolved hap1+hap2 repeat copies in 19 cases out of 54 (35%), and sometimes only hap1 repeat copies in 2 cases out of 54 (4%). LJA produced more fragmented assemblies (the average number of contigs was 3.65), with a total of 7 haplotype switching errors. HiCANU resolved hap1+hap2 repeat copies in 10 cases out of 54 (19%), and sometimes only hap1 repeat copies in 9 cases out of 54 (17%). HiCANU produced more fragmented assemblies (the average number of contigs was 7.49), with a total of 12 haplotype switching errors. Overall, LJA and HiCANU performed worse than RAmbler and hifiasm on these datasets, producing assemblies that were more fragmented and contained less repeat copies.

Reference-guided repeat assembly with RAMbler

341 Supplemental Tables 16 and 18 summarize the cases where hap1 had 10 copies and hap2 had 6, 8 or
342 10 copies. RAMbler resolved 10 copies of the repeat unit in 32 cases out of 54 (59%) with 14 haplotype
343 switching errors. The average number of contigs over all RAMbler’s assemblies was 1.26. hifiasm resolved
344 10 repeat copies in 40 cases out of 54 (74%) with 17 haplotype switching errors. The average number of
345 contigs over all hifiasm’s assemblies was 1.06 for the primary, and 1.07 for the alternate. LJA’s assemblies
346 matched either hap1 or hap1+hap2 in 19 cases out of 54 (35%) with 18 haplotype switching errors. The
347 average number of contigs over all LJA’s assemblies was 7.52. HiCANU’s assemblies matched either hap1 or
348 hap1+hap2 in 16 cases out of 54 (30%) with 21 haplotype switching errors. The average number of contigs
349 over all HiCANU’s assemblies was 13.81. LJA and HiCANU, again, performed worse than RAMbler and
350 hifiasm on these datasets.

351 In summary, RAMbler captured the larger repeat count between the two haplotypes in most cases. It
352 frequently produced a single contig with a low number of haplotype switching errors, relying exclusively on
353 HiFi reads.

354 Discussion

355 We introduced RAMbler, a reference-guided genome assembler aimed at resolving complex repetitive regions.
356 To the best of our knowledge, there is no other specialized assembler for resolving complex repeats that
357 uses HiFi reads exclusively. Both SDA and CentroFlye expect as input error-prone PacBio CLR or ONT
358 long reads. SDA is no longer maintained and CentroFlye requires very high computational resources and
359 additional information about the centromeres of interest.

360 RAMbler leverages *unikmers* to detect overlaps and locally assemble the HiFi reads. By taking advantage
361 of shared *unikmers*, RAMbler can select safe and informative overlaps that are difficult to identify by tradi-
362 tional assemblers. Traditional HiFi assemblers rely on highly accurate (but not necessarily perfect) overlaps
363 to build a string/overlap graph, which is a method that works very well for non-repetitive regions of the
364 genome. For instance, hifiasm carries out an all-pairs sequence alignment of the HiFi reads before building
365 the string graph. Highly repetitive regions generate an overwhelming number of prefix-suffix overlaps, which
366 are difficult to process. In contrast, the use of *unikmers* allows RAMbler to detect informative overlaps
367 without the need of sifting through a very large number of sequence alignments. The use of *unikmers* also
368 allows RAMbler to eliminate the need of the error correction step to compensate for rare sequencing errors
369 in HiFi reads, as it is done in hifiasm.

370 The extensive set of experiments on human Chromosomes 8, 19, and X across centromeric and non-
371 centromeric complex repetitive regions demonstrated RAMbler’s ability to achieve T2T-level assembly qual-
372 ity using PacBio HiFi reads exclusively, without manual intervention. Our comparative experimental results
373 on more than 250 synthetic data sets, and on real data for *H. sapiens*, indicated that RAMbler outperformed
374 hifiasm, LJA, HiCANU, and Verkko on reconstructing these repetitive regions in the majority of cases. RAM-
375 bler generated assemblies with the fewest contigs, achieving higher completeness, contiguity, and accuracy
376 in them. Our analysis of synthetic diploid genomes with haplotype-dependent copy number illustrated that
377 RAMbler can produce less fragmented assemblies with fewer haplotype switching errors compared to LJA,
378 HiCANU, and Verkko by relying exclusively on HiFi reads. hifiasm performed comparably to RAMbler on
379 synthetic data sets with fixed-length repeat units and copy number variation between the haplotypes, but
380 its performance declined with variable-length repeats and further deteriorated on real datasets. LJA also
381 performed well, producing assemblies with a low number of contigs, second only to RAMbler on synthetic

Reference-guided repeat assembly with RAMbler

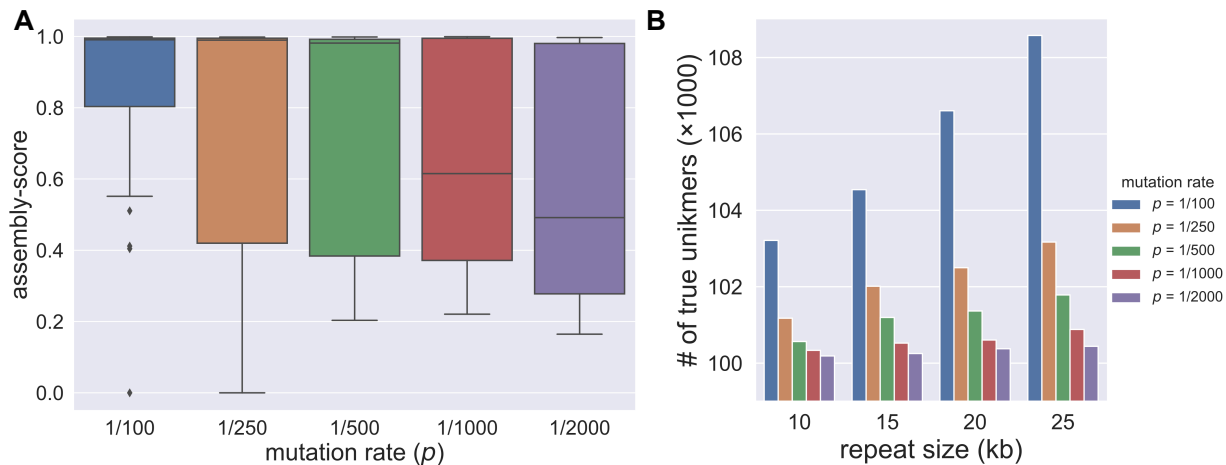


Figure 3: Relationship among RAMbler’s performance, mutation rate p , and number of true unikmers; (A) Performance of RAMbler for several choices of the mutation rate p over 27 different combinations of repetitive regions with repeat sizes $\{10, 15, 20\}$ kb, number of copies $\{2, 5, 10\}$, coverage depth $= \{20\times, 30\times, 40\times\}$, (B) Number of true unikmers as a function of mutation rate and repeat unit’s sizes (five copies)

382 data. However, LJA performed poorly on contiguity when the region to be assembled was longer than 1 Mb
 383 on real data. HiCANU and Verkko generated assemblies with a large number of contigs. Although HiCANU
 384 and Verkko rarely produced misassembled contigs, they suffered from poor contiguity, creating many small
 385 contigs and inflating the total assembly size. Additionally, there were instances where HiCANU failed to
 386 complete the assembly on real datasets. Similarly, Verkko failed to generate an assembly in two separate
 387 instances of *H. sapiens* dataset and majority cases of synthetic diploid genomes.

388 RAMbler still has some limitations. In simulations, its performance drops significantly when the mutation
 389 rate p is smaller than 1/1000 (see Figure 3A). In this case, the individual copies inside the repetitive
 390 regions are almost identical, thus there are not enough unikmers to resolve them. This observation is
 391 supported by Figure 3B, which illustrates that the number of unikmers decreases with p . However, the
 392 mutation rate in eukaryotic genomes is generally higher than one SNP over a thousand base pairs (Risch
 393 2000; International Human Genome Sequencing Consortium 2001; Orr and Chanock 2008). The current
 394 implementation of RAMbler is not multi-threaded which penalizes its runtime. This is an opportunity
 395 to improve RAMbler’s runtime. In addition, RAMbler generates pseudo-haplotype assemblies from HiFi
 396 data, while LJA, HiCANU, and Verkko produce phased assemblies and hifiasm always attempts to produce
 397 haplotype-resolved assemblies. HiCANU, hifiasm and Verkko support trio-based enhanced phasing. hifiasm
 398 and Verkko can use Hi-C reads to further improve haplotype resolution. Currently our tool lacks these
 399 features, which we plan to add in future versions of RAMbler.

400 Methods

401 Problem Formulation

402 We assume that (1) the genome G contains n repetitive regions $\{R_1, \dots, R_i, \dots, R_n\}$, (2) each repetitive
 403 region R_i is composed of t_i tandem copies of a string α_i , (3) each tandem copy has sufficient variations that
 404 allows it to be distinguished from another copy; we assume that each copy contains SNPs with probability
 405 p (e.g., $p = 1/100$), and its length can increase or decrease by at most $L\%$. Given a set T of HiFi reads and

Reference-guided repeat assembly with RAMbler

406 the draft genome G , the objective is to produce a set $\{F_1, \dots, F_i, \dots, F_n\}$ of n assemblies, where each F_i
 407 is as “similar as possible” to R_i . In particular, if the assembly F_i contains t'_i copies of the repeat unit, we
 408 want t'_i as close as t_i as possible (see Supplemental Figure 16). For synthetic data, we measure the quality
 409 of the assembly F_i by comparing it to R_i using QUAST (Gurevich et al. 2013), i.e., we report the fraction of
 410 R_i covered by F_i (ideally 100%), the number of mis-assembled contigs in F_i (ideally zero), and the number
 411 of contigs in F_i (ideally one). When the ground truth is unavailable (i.e., for real data sets), a qualitative
 412 assessment of the assembly’s accuracy can be obtained by aligning F_i with the corresponding repeat unit α_i .
 413 The alignment, visualized as a dot plot, can provide a qualitative measure on how well the repeat units are
 414 assembled.

415 **Repeat Identification**

416 To identify the repetitive regions $\{R_1, \dots, R_n\}$, we map the HiFi reads T against the draft genome assembly
 417 G . Since unresolved tandem repeats are collapsed in the draft assembly, they can be identified by a spike
 418 in mapping coverage. For instance, Figure 4 shows the mapping coverage of an unresolved tandem repeat
 419 in Chromosome XII of *Saccharomyces cerevisiae* which is known to contain ~ 150 tandemly repeated copies
 420 of a 9.1 kb rDNA unit (Johnston et al. 1997; YH Kim et al. 2006). This region is the only unresolved
 421 non-telomeric gap in the current *S. cerevisiae* assembly.

422 **Analysis of k -mer Distribution**

423 Recall that we assume that the copies in the repetitive region are not identical to each other. If all the copies
 424 were identical, the problem of resolving repeats would be impossible, unless one can produce reads so long
 425 that they span the entire repetitive region. We rely on the presence of the SNPs to distinguish and partition
 426 the HiFi reads that belong to different repeat copies within a repetitive region. When distinct SNPs are
 427 present among the different copies, we expect those copies to have their own SNP signatures.

428 Each SNP is likely to induce a unique (or single-copy) k -mer, i.e., a k -mer that occurs a number of times
 429 approximately equal to the expected sequencing coverage. We call these k -mers, *unikmers*. Unikmers (called
 430 SUNKs in (Logsdon et al. 2021) or SUNs in (Sudmant et al. 2010)) were crucial to resolve the assembly of
 431 human Chromosome 8, but to the best of our knowledge they have not been used in any other assembly
 432 method. Please note that unikmers are NOT k -mers that appear only once in the reads: those k -mers
 433 correspond to sequencing errors.

434 One of the contributions of our study is to provide a method to identify unikmers from the reads, and
 435 analyze its accuracy and precision. RAMbler finds unikmers by selecting all k -mers in the HiFi reads that
 436 have a number of occurrences within the interval $[\mu - t\sigma, \mu + t\sigma]$, where μ is the average sequencing depth of
 437 the HiFi reads, σ is the standard deviation of the sequencing depth, and k and t are user-defined parameters.

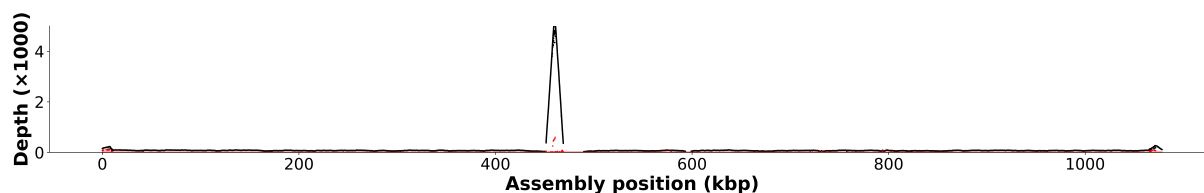


Figure 4: PacBio HiFi mapping coverage for Chromosome XII in *Saccharomyces cerevisiae* illustrated using NucFreq; the coverage spike indicates the presence of a repetitive region which is known to contain ~ 150 tandemly repeated copies of a 9.1 kb rDNA unit

Reference-guided repeat assembly with Rambler

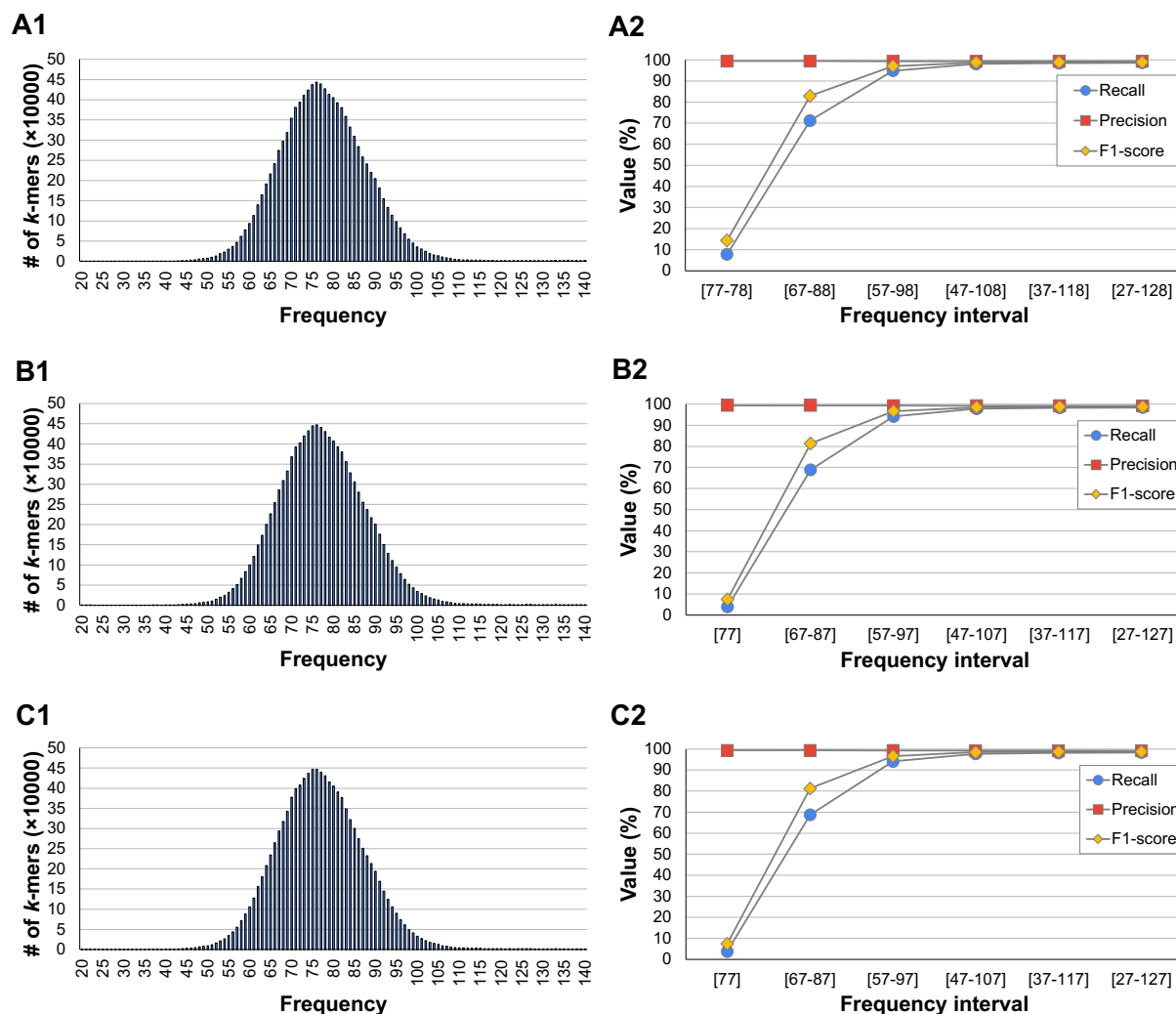


Figure 5: PacBio HiFi k -mer distribution for $k = 17$ (A1), $k = 21$ (B1) and $k = 25$ (C1) for *Saccharomyces cerevisiae*; precision, recall, and F1-score for unikmers when $k = 17$ (A2), $k = 21$ (B2) and $k = 25$ (C2) for longer and longer intervals centered at the average sequencing coverage

438 We investigate how to choose k and t in the following analysis. (1) We determine the set of true unikmers
 439 in the *Saccharomyces cerevisiae* genome to serve as the ground truth. (2) We compute the k -mer distribution
 440 for a set of real HiFi reads (SRA accession SRR13577847) and ONT reads (SRA accession SRR18365585)
 441 for *S. cerevisiae* using Jellyfish (Marçais and Kingsford 2011). The left column of Figure 5 shows the k -mer
 442 distribution for HiFi reads for $k = 17$ (Figure 5-A1), $k = 21$ (Figure 5-B1) and $k = 25$ (Figure 5-C1). Odd
 443 integers in the range $[17, 25]$ are typical choices for k to estimate genome size (see, e.g., (Vurture et al. 2017;
 444 Ranallo-Benavidez et al. 2020)) or the construction of the de Bruijn graphs for eukaryotic genomes (see,
 445 e.g., (Zerbino and Birney 2008)). Observe that the distributions are almost identical, which indicate that
 446 any of these k -mer choices would be appropriate. (3) We compute the average sequencing depth μ and the
 447 standard deviation of the sequencing depth σ from the k -mer distribution. (4) The k -mers in the HiFi reads
 448 that have a number of occurrences within the interval $[\mu - t\sigma, \mu + t\sigma]$ for $t = 0, 1, 2, 3, 4, 5$ are compared
 449 against the true unikmers: true positive, false positive, true negative and false negative are recorded.

450 The results of this analysis (precision, recall and F1-score) for HiFi reads are shown in the right column

Reference-guided repeat assembly with RAMbler

of Figure 5. The x-axis represents the choice of t , i.e., longer and longer intervals centered around the mean (the first interval is for $t = 0$, the second is for $t = 1$, etc.). Figure 5-A2 shows the results for $k = 17$, Figure 5-B2 illustrates the results for $k = 21$ and Figure 5-C2 shows the results for $k = 25$. Observe that in all cases precision and recall are very close to 100% as soon as $t = 3$. For instance, there are 11,137,337 21-mers that occur [47 – 107] times in the HiFi reads, i.e., at most $t = 3$ standard deviations away from the average coverage. Of those, 11,058,290 are truly *unikmers* which correspond to a precision of 99.29%; only 79,047 are false positives (0.71% of the total). For $t = 3$, this method recalls 97.84% of the *unikmers* in the genome. Almost identical results can be obtained from $k = 17$ or $k = 25$. This analysis indicates that selecting k -mers that have a number of occurrences in the interval $\mu \pm 3\sigma$ in HiFi reads can recover almost 98% of the true single-copy k -mers in the genome with a false positive rate less than 1%. The same analyses carried out on ONT reads show that precision, recall, and F1-score for ONT reads are slightly lower than those obtained from PacBio HiFi reads, likely due to the higher rate of sequencing errors in ONT reads (see Supplemental Figure 17).

Based on this analysis, we used $k = 21$ and $t = 3$ for all the experiments (unless otherwise noted).

RAMbler’s Algorithm

The algorithm used in RAMbler is illustrated in Figure 6. It comprises of six major steps, the first two of which are data preprocessing.

A. Determine the reads corresponding to repetitive regions. As mentioned in Section “Repeat Identification”, RAMbler identifies repetitive regions by mapping all HiFi reads T against the draft genome. RAMbler generates the plot of the read coverage across the genome using NucFreq (Vollger et al. 2019). Unresolved repetitive regions produce distinctive peaks in the coverage plot as illustrated in Figure 4 and Supplemental Figure 2. Then, RAMbler selects the reads that map to the repetitive regions, as well as the reads extending 50 kb upstream and downstream of the peak, as shown in Figure 6, step A. We call the set of HiFi reads extracted in this step T_r , where r identifies the repetitive region.

B. Determine *unikmers*. RAMbler uses Jellyfish on the entire set of HiFi reads T to obtain the count distribution of 21-mers genome-wide. RAMbler calculates the mean μ and standard deviation σ of the distribution by excluding 21-mers that appear less than 5 times since these are most likely induced by sequencing errors. RAMbler then selects the 21-mers that fall within the interval $[\mu - 3\sigma, \mu + 3\sigma]$. According to our analysis in Section “Analysis of k -mer Distribution”, these 21-mers are true *unikmers* with high probability (Figure 6, step B).

C. Barcode reads. RAMbler uses the set of *unikmers* to barcode the HiFi reads T_r (Figure 6, step C). RAMbler searches for exact occurrences of the *unikmers* in the reads T_r or their reverse complement. The set of *unikmers* present in a read and their location is the *barcode* of that read. For each read, RAMbler stores pairs (u, j) , where u is a *unikmer* and j is the location within the read.

D. Cluster the barcoded reads. RAMbler compares the barcode of all pairs of reads to identify shared *unikmers*. This pairwise comparison allows RAMbler to determine which reads are overlapping. Two reads are overlapping if they share at least th *unikmers*, and the set of relative distances between the shared *unikmers* match within a tolerance up to to base pairs. Overlaps are stored in the *barcode graph*: each node

Reference-guided repeat assembly with RAMbler

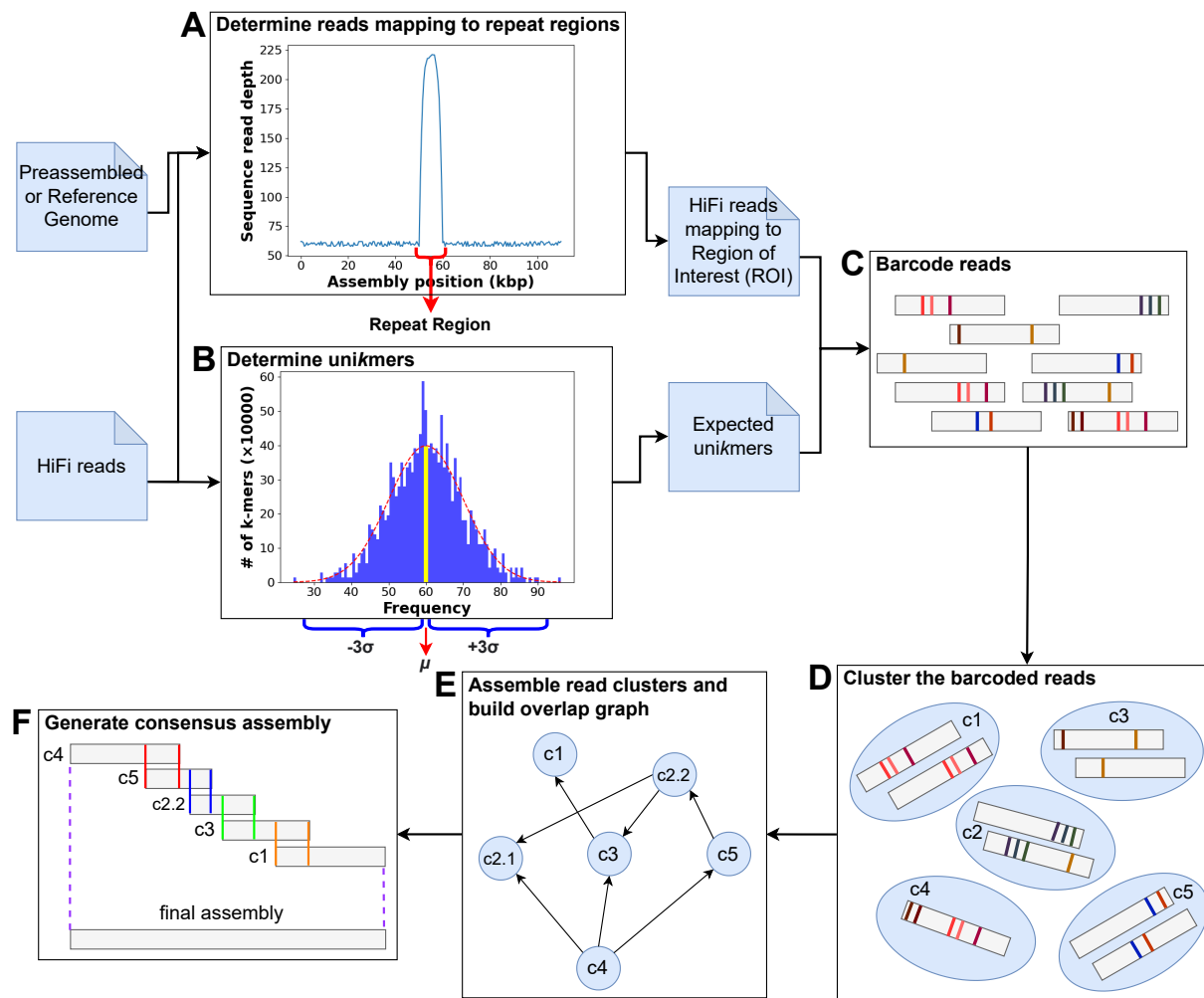


Figure 6: The algorithmic pipeline used in RAMbler

489 in the barcode graph represents a read; nodes in the graph are connected by an edge if the corresponding
 490 reads are overlapping according to the criteria described above. Once the graph is completed, RAMbler
 491 identifies clusters of reads by determining the connected components of the barcode graph (Figure 6, step
 492 D).

493 **E. Assemble read clusters and build overlap graph.** RAMbler carries out individual local assemblies for each
 494 set of clustered reads using a standard HiFi assembler (hifiasm in this case). Each read cluster is assembled
 495 in one or more contigs. RAMbler then uses minimap2 (H Li 2018) to align assembled contigs to each other.
 496 Any contig that is fully contained within another longer contig is removed. RAMbler constructs an overlap
 497 graph based on the overlap information provided by minimap2: each node in the overlap graph represents
 498 a contig; nodes are connected by edges if they have a suffix-prefix overlap (Figure 6, step E). It is worth
 499 noting there could be multiple suffix-prefix overlaps between a pair of contigs. RAMbler retains the overlap
 500 with the highest percentage of identity as long as the overlap is at least $mo = 1000$ bps. Furthermore, these
 501 suffix-prefix overlaps can occur between the positive or negative strands, resulting in three types of edges.
 502 Each edge is labeled by a pair (t, l) , where $t \in \{+, -, *\}$, and l is the length of the suffix-prefix overlap

Reference-guided repeat assembly with RAMbler

503 (Supplemental Figure 18). Given an edge (u, v) in the overlap graph, its type t is (1) “+” when there is an
 504 overlap between contig u and contig v , (2) “-” when there is an overlap between the reverse complement
 505 of contig u and contig v , (3) “*” when there is an overlap between contig u and the reverse complement of
 506 contig v .

507 **F. Generate consensus assembly.** At this stage, RAMbler needs an estimate of the size of the repetitive
 508 region. The size is obtained by computing the ratio between the total number of bases in the reads T_r
 509 (extracted in step A) and the average coverage depth.

510 To compute the final assembly, RAMbler first determines whether the overlap graph is acyclic. If it is
 511 acyclic, RAMbler enumerates all possible paths using DFS and generates a set of candidate assemblies. When
 512 computing the sequence consensus for suffix-prefix overlaps, if the suffix and the prefix do not match, RAM-
 513 bler arbitrarily picks the base either from the suffix or the prefix (Figure 6, step F). Among all assemblies,
 514 RAMbler selects the one that best matches the estimated length of the repetitive region.

515 When the overlap graph has cycles, RAMbler partitions the graph into three components: an acyclic pre-
 516 cycle subgraph, the cycle itself, and a post-cycle subgraph (which could be cyclic), as shown in Supplemental
 517 Figure 19. RAMbler repeats this process iteratively on the post-cycle subgraph until no cycles remain. Once
 518 the graph is completely decomposed in a set of acyclic subgraphs, RAMbler generates an assembly for each
 519 subgraph as described in the previous paragraph. RAMbler then enumerates all possible combinations of
 520 partial assemblies and selects the combination such that the sum of the individual assembly’s length best
 521 matches the estimated length of the repetitive region.

522 A summary of RAMbler’s main parameters k , t , th , to , and mo with their default values is shown
 523 in Supplemental Table 13. The optimization of these parameters is discussed in Section “Synthetic Data
 524 Generation and Parameter Optimization”.

525 Performance Metrics for Real Data

526 We used CRAQ (Clipping information for Revealing Assembly Quality) to assess the assemblies produced
 527 by RAMbler (K Li et al. 2023). CRAQ utilizes both NGS short reads and long reads for identifying and
 528 classifying errors in a given draft assembly. It reports assembly errors at different scales by transforming
 529 error counts into corresponding assembly quality indicators (AQIs) that reflect assembly quality at both
 530 regional and structural levels. CRAQ can distinguish between assembly errors and heterozygous loci based
 531 on the ratio of mapping coverage and the effective number of clipped reads; (1) Clip-based Regional Errors
 532 (CREs): If a region with clipped NGS reads is spanned by long reads with only SNP cluster features, and (2)
 533 Clip-based Structural Errors (CSEs): if the mapped long reads around a region with errors exhibit clipping
 534 features, i.e., the NGS reads simultaneously show clipping or no coverage.

535 Assembly quality index (AQI)

$$\text{AQI} = 100e^{-0.1N/L} \quad (1)$$

536 where N represents the cumulative normalized CRE or CSE count, and L indicates the total length of the
 537 assembly in the mega-base unit. Observe that a perfect assembly will yield an AQI score of 100. To avoid
 538 excessive impacts of specific regions enriched in errors (e.g, pericentromeric regions) on the overall AQI
 539 values, error counts were normalized within a sliding window of $0.0001 \times (\text{total assembly size})$.

Reference-guided repeat assembly with RAMbler

540 **Normalized error count in a window (N_w)**

$$N_w = \sum_{i=1}^m i^{-1} \quad (2)$$

541 where m is the actual number of CRE/CSEs in the block. R-AQI and S-AQI can be calculated separately
542 for CREs and CSEs.

543 **Performance Metrics for Synthetic Data**

544 Metrics such as NG50, genome fraction, number of contigs, and number of mis-assemblies have been tra-
545 ditionally employed to evaluate the quality of an assembly on synthetic data when the reference genome
546 is known. However, each of these metrics alone does not fully capture all the desired qualities of a “good
547 assembly”. To address this shortcoming we introduce here a new metric called the *assembly score* that sum-
548 marizes in one number the quality of an assembly in terms of accuracy, contiguity, and completeness. The
549 assembly score is based on two preliminary metrics, as explained next.

550 **Effective genome fraction per contig (ζ)**

551 Consider an assembly that consists of a single contig but contains one mis-assembly. To correct the mis-
552 assembly, the contig needs to be broken, resulting in the creation of an additional contig. Based on this
553 observation we define the *effective number of contigs* as the sum of the number of contigs in the assembly and
554 the number of mis-assemblies. We propose to calculate the effective genome fraction per contig as follows

$$\zeta = \frac{\text{genome_fraction}}{\#\text{contigs} + \#\text{mis-assembled contigs}} \quad (3)$$

555 As said, metric ζ takes into account the number of mis-assemblies and penalizes the score accordingly. An
556 assembly that covers 100% of the reference genome (without any mis-assemblies) would yield $\zeta = 1.0$. By
557 considering the effective genome fraction per contig, we can assess the assembly quality while accounting for
558 the presence of mis-assemblies, thereby providing a more comprehensive evaluation.

559 **Normalized NG50 (η)**

560 While NG50 is an essential metric for evaluating the contiguity of an assembly, it depends on the size of the
561 reference genome, making it challenging to use it to compare an assembler’s performance across genomes
562 of different lengths. To address this limitation, we normalize NG50 by the size of the reference genome,
563 yielding a metric called η defined as follows

$$\eta = \frac{\text{NG50}}{|\text{reference_genome}|} \quad (4)$$

564 By normalizing NG50 with respect to the reference genome size, η is constrained within the range of $[0, 1]$,
565 with a perfect assembly achieving $\eta = 1$. Observe that it is possible that η may exceed 1 when the assembly is
566 over-inflated, i.e., longer than the actual genome. In general, a higher value of η indicates a better assembly
567 quality, as long as it is smaller than 1.

Reference-guided repeat assembly with RAmbler

568 Assembly score

569 The assembly score is defined by computing the harmonic mean of ζ and η , as follows

$$\text{assembly_score} = \frac{2\zeta\eta}{\zeta + \eta} \quad (5)$$

570 Observe that while ζ is always within $[0, 1]$, η can exceed 1, which can result in an assembly score greater
571 than 1. Nevertheless an assembly score closer to 1 indicates a higher quality assembly. This score enables
572 a holistic assessment of the assembly's quality, taking into account accuracy, contiguity, and completeness,
573 which are all equally important.

574 Acknowledgments

575 This project was supported in part by NSF IIS #2444456, NSF CBET #2225878, and NIH 1-R01-AI169543-
576 01 to SL. The authors thank the anonymous reviewers what helped to improve the quality of this manuscript.

577 **Author contributions.** S.C., G.L. and S.L. conceptualized and designed the study.; S.C. wrote and tested
578 the code; G.L. provided guidance on the experimental design on the human genome; S.C. carried out the
579 experiments and generated figures and tables; S.C. and S.L. wrote the manuscript; all authors read and
580 approved the final manuscript.

581 Competing interest statement

582 The authors declare no competing interests.

583 Software availability

584 RAmbler is available from GitHub (<https://github.com/ucrbioinfo/rambler>) and as Supplemental
585 Code.

References

- 586
- 587 Altomose N et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science*. **376**: eabl4178.
588 doi:[10.1126/science.abl4178](https://doi.org/10.1126/science.abl4178).
- 589 Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, and O’Grady J. 2015. MinION nanopore
590 sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33**: 296–
591 300.
- 592 Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, and Pevzner PA. 2022. Multiplex de Bruijn graphs enable
593 genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* **40**: 1075–1081.
- 594 Bustos BI, Billingsley K, Blauwendraat C, Gibbs JR, Gan-Or Z, Krainc D, Singleton AB, Lubbe SJ, and (IPDGC)
595 IPDGC. 2022. Genome-wide contribution of common short-tandem repeats to Parkinson’s disease genetic risk.
596 *Brain*. **146**: 65–74.
- 597 Bzikadze AV and Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat.*
598 *Biotechnol.* **38**: 1309–1316.
- 599 Chaisson MJ, Mukherjee S, Kannan S, and Eichler EE. 2017. Resolving multicopy duplications de novo using poly-
600 ploid phasing. *Proceedings of RECOMB*. **10229**: 117–133.
- 601 Chan S, Wang W, Hallers B ten, Peters S, Gaiero P, Jong H de, Perez G, Hastie A, and Cao H. 2015. Detection,
602 Characterization, and Biological Analysis of Long Tandem Repeats Using Nanochannel Technology. In *poster at*
603 *Plant and Animal Genome conference*.
- 604 Cheng H, Concepcion GT, Feng X, Zhang H, and Li H. 2021. Haplotype-resolved de novo assembly using phased
605 assembly graphs with hifiasm. *Nat. Methods*. **18**: 170–175.
- 606 Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, and Bayley H. 2009. Continuous base identification for single-
607 molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**: 265–270.
- 608 Eid J et al. 2008. Real-time DNA sequencing from single polymerase molecules. *Science*. **323**: 133–138.
- 609 Garg S et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**: 309–
610 312.
- 611 Goel M, Sun H, Jiao WB, and Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence
612 differences from whole-genome assemblies. *Genome Biol.* **20**: 277.
- 613 Gurevich A, Saveliev V, Vyahhi N, and Tesler G. 2013. QUAST: quality assessment tool for genome assemblies.
614 *Bioinformatics*. **29**: 1072–1075.
- 615 Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**: 286–298.
- 616 Hoyt SJ et al. 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements.
617 *Science*. **376**: eabk3112. doi:[10.1126/science.abk3112](https://doi.org/10.1126/science.abk3112).
- 618 Huddleston J et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome*
619 *Res.* **24**: 688–696.
- 620 International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome.
621 *Nature*. **409**: 860–921.
- 622 Ishiura H et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyn-
623 godistal myopathy and an overlapping disease. *Nat. Genet.* **51**: 1222–1232.
- 624 Johnston M et al. 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature*. **387**: 87–90.
- 625 Jurka J, Kapitonov VV, Kohany O, and Jurka MV. 2007. Repetitive Sequences in Complex Genomes: Structure and
626 Evolution. *Annu. Rev. Genomics Hum. Genet.* **8**: 241–259.

Reference-guided repeat assembly with RAMbler

- 627 Kim KE et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data*. **1**:
628 140045.
- 629 Kim YH, Ishikawa D, Ha HP, Sugiyama M, Kaneko Y, and Harashima S. 2006. Chromosome XII context is important
630 for rDNA function in yeast. *Nucleic Acids Res.* **34**: 2914–2924.
- 631 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, and Marra MA. 2009. Circos: an
632 information aesthetic for comparative genomics. *Genome Res.* **19**: 1639–1645.
- 633 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**: 3094–3100.
- 634 Li K, Xu P, Wang J, Yi X, and Jiao Y. 2023. Identification of errors in draft genome assemblies at single-nucleotide
635 resolution for quality assessment and improvement. *Nat. Commun.* **14**: 6556.
- 636 Liao WW et al. 2023. A draft human pangenome reference. *Nature.* **617**: 312–324.
- 637 Liao X, Zhu W, Zhou J, Li H, Xu X, Zhang B, and Gao X. 2023. Repetitive DNA sequence detection and its role in
638 the human genome. *Commun. Biol.* **6**: 954.
- 639 Logsdon GA et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature.* **593**:
640 101–107.
- 641 Loose M, Malla S, and Stout M. 2016. Real-time selective sequencing using nanopore technology. *Nat. Methods.* **13**:
642 751–754.
- 643 Marçais G and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.
644 *Bioinformatics.* **27**: 764–770.
- 645 McGinnis S and Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools.
646 *Nucleic Acids Res.* **32**: W20–5.
- 647 Miga KH et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature.* **585**: 79–84.
- 648 Mrázek J, Guo X, and Shah A. 2007. Simple sequence repeats in prokaryotic genomes. *Proceedings of the National
649 Academy of Sciences.* **104**: 8472–8477.
- 650 Nurk S, Koren S, et al. 2022. The complete sequence of a human genome. *Science.* **376**: 44–53.
- 651 Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, and Koren S.
652 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long
653 reads. *Genome Res.* **30**: 1291–1305.
- 654 Ono Y, Asai K, and Hamada M. 2012. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinform-
655 matics.* **29**: 119–121.
- 656 Orr N and Chanock S. 2008. Chapter 1 Common Genetic Variation and Human Disease. In *Advances in Genetics*,
657 pp. 1–32. Academic Press.
- 658 Porubsky D et al. 2021. Fully phased human genome assembly without parental data using single-cell strand sequenc-
659 ing and long reads. *Nat. Biotechnol.* **39**: 302–308.
- 660 Qin J et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature.* **490**: 55–60.
- 661 Quick J, Quinlan AR, and Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION portable
662 single-molecule nanopore sequencer. *Gigascience.* **3**: 1–6.
- 663 Ranallo-Benavidez TR, Jaron KS, and Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profil-
664 ing of polyploid genomes. *Nat. Commun.* **11**: 1432.
- 665 Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, and Koren S. 2023.
666 Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**: 1474–1482.
- 667 Risch NJ. 2000. Searching for genetic determinants in the new millennium. *Nature.* **405**: 847–856.
- 668 Roberts RJ, Carneiro MO, and Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol.* **14**: 405.

Reference-guided repeat assembly with RAMbler

- 669 Shah NM et al. 2023. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat.*
670 *Genet.* **55**: 631–639.
- 671 Shumate A, Zimin AV, Sherman RM, Puiu D, Wagner JM, Olson ND, Pertea M, Salit ML, Zook JM, and Salzberg SL.
672 2020. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**: 129.
- 673 Sudmant PH et al. 2010. Diversity of human copy number variation and multicopy genes. *Science.* **330**: 641–646.
- 674 Treangen TJ, Abraham AL, Touchon M, and Rocha EP. 2009. Genesis, effects and fates of repeats in prokaryotic
675 genomes. *FEMS Microbiol. Rev.* **33**: 539–571.
- 676 Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson
677 MJP, and Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat. Methods.* **16**: 88–
678 94.
- 679 Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, and Schatz MC. 2017. GenomeScope:
680 fast reference-free genome profiling from short reads. *Bioinformatics.* **33**: 2202–2204.
- 681 Wenger AM et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly
682 of a human genome. *Nat. Biotechnol.* **37**: 1155–1162.
- 683 Zerbino DR and Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome*
684 *Res.* **18**: 821–829.