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

Sakshar Chakravarty[1], Glennis Logsdon[2] and Stefano Lonardi[1*]

[1*] *Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA*

E–mail: *schak026@ucr.edu, stelo@cs.ucr.edu*

[2] *Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19103, USA*

E–mail: *glogsdon@pennmedicine.upenn.edu*

(Received: ; Accepted:)

Abstract

Repetitive regions in eukaryotic genomes often contain important functional or regulatory elements. Despite significant algorithmic and technological advancements in genome sequencing and assembly over the past three decades, modern *de novo* assemblers still struggle to accurately reconstruct highly repetitive regions.

In this work, we introduce RAmbler (**R**epeat **A**sse**mbler**), a reference-guided assembler specialized for the assembly of complex repetitive regions exclusively from PacBio HiFi reads. RAmbler (i) identifies repetitive regions by detecting unusually high coverage regions after mapping HiFi reads to the draft genome assembly, (ii) finds single-copy k-mers from the HiFi reads, (i.e., k-mers that are expected to occur only once in the genome), (iii) uses the relative location of single-copy k-mers to barcode each HiFi read, (iv) clusters HiFi reads based on their shared barcodes, (v) generates contigs by assembling the reads in each cluster, and (vi) generates a consensus assembly from the overlap graph of the assembled contigs.

Here we show that RAmbler can reconstruct human centromeres and other complex repeats to a quality comparable to the manually-curated telomere-to-telomere human genome assembly. Across over 250 synthetic datasets, RAmbler outperforms hifiasm, LJA, HiCANU, and Verkko across various parameters such as repeat lengths, number of repeats, heterozygosity rates and depth of sequencing.

Key words. DNA sequencing, reference-guided genome assembly, repetitive regions, segmental duplications, PacBio HiFi, tandem repeats, single-copy k-mers

Introduction

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

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

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

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

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

Results

Experimental Results on *Homo sapiens*

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

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

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

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

We carried out these five experiments on the centromeric regions of Chromosomes 8 and X, as well as three non-centromeric regions on Chromosome 19. We ran RAmbler, hifiasm, LJA, HiCANU, and Verkko on the subset of HiFi reads mapped to these regions. SDA was excluded because "*is no longer maintained and should not be used . . . assembly tools like Flye, HiCanu, and hifiasm outperform any results previously possible with SDA.*" (quote from `https://github.com/mrvollger/SDA`). CentroFlye was also excluded because we were unable to run it on our 2.8 GHz 32-core processor server with 512GB of RAM. We tried first to reconstruct the centromere of Chromosome X using the HiFi reads and the HORs provided by the authors, but CentroFlye failed during the error-correction step. When we tried to reconstruct the same centromere using the CHM13 ONT reads from (Bzikadze and Pevzner 2020), CentroFlye ran out of memory. The instructions claim that CentroFlye requires about 800GB of RAM to complete the assembly of Chromosome 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**	
									Total (bp)	4,678,725	4,678,719	4,663,887	4,669,944	N/A
									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.

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

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

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

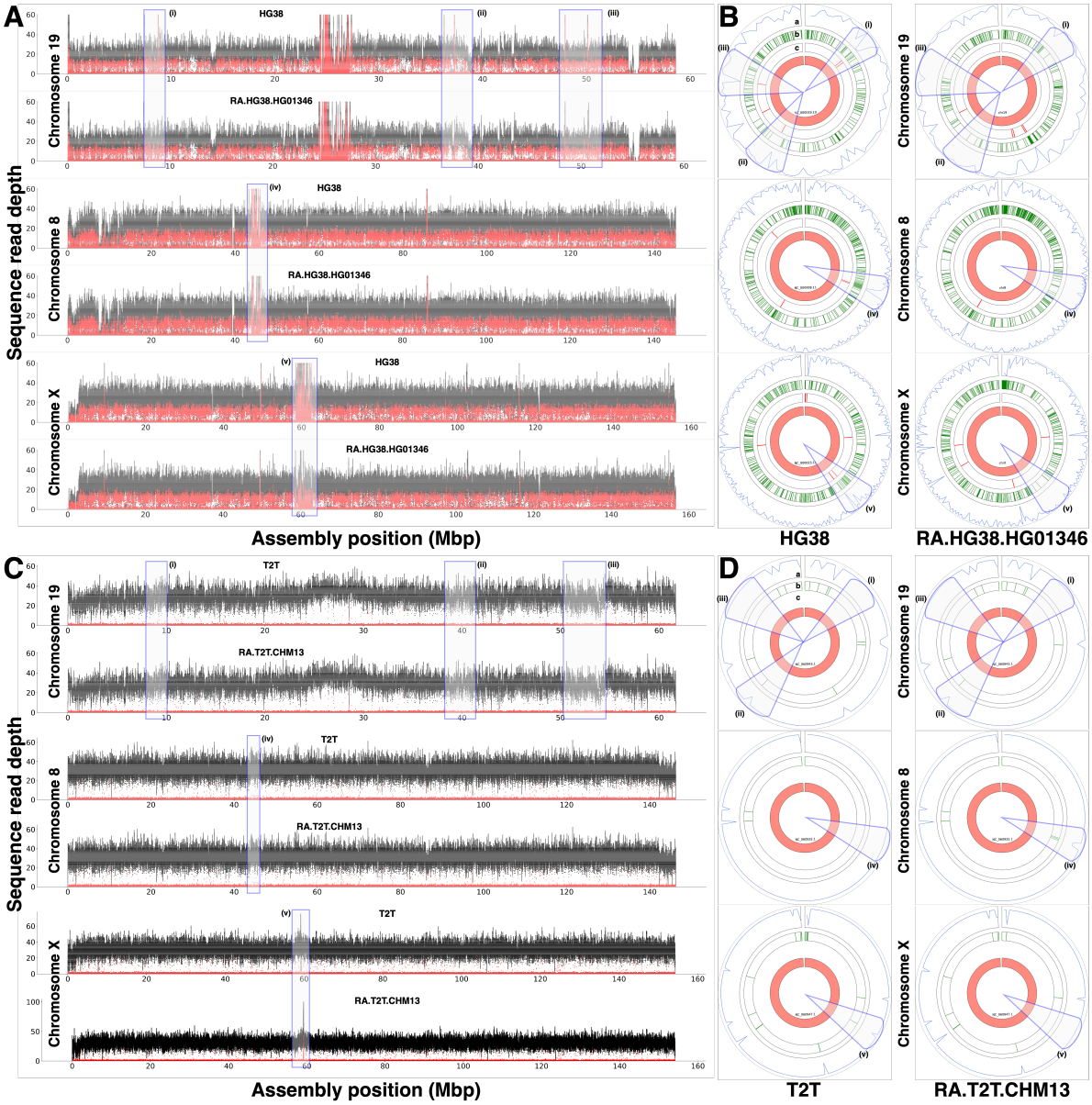

Figure 1: Comparing RAmbler's assemblies (RA.HG38.HG01346 and RA.T2T.CHM13) against **(A, B)** the GRCh38.p13 assembly (HG38) and **(C, D)** the T2T-CHM13.v2.0 assembly (T2T) of human Chromosomes 8, 19, and X; in all plots, blocks (i), (ii), and (iii) are non-centromeric repeats in Chromosome 19; blocks (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)

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

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

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

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

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

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 (∼13.5× coverage) and long PacBio HiFi reads (∼66× coverage) for HG002; Illumina reads (∼12× coverage) and long PacBio HiFi reads (∼33× coverage) for HG00733; Illumina reads (∼10× coverage) and long PacBio HiFi reads (∼26× coverage) for HG01346; numbers in bold indicate the best scores

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

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

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

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

Experimental Results on Synthetic Data

Synthetic Data Generation and Parameter Optimization

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

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

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

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

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

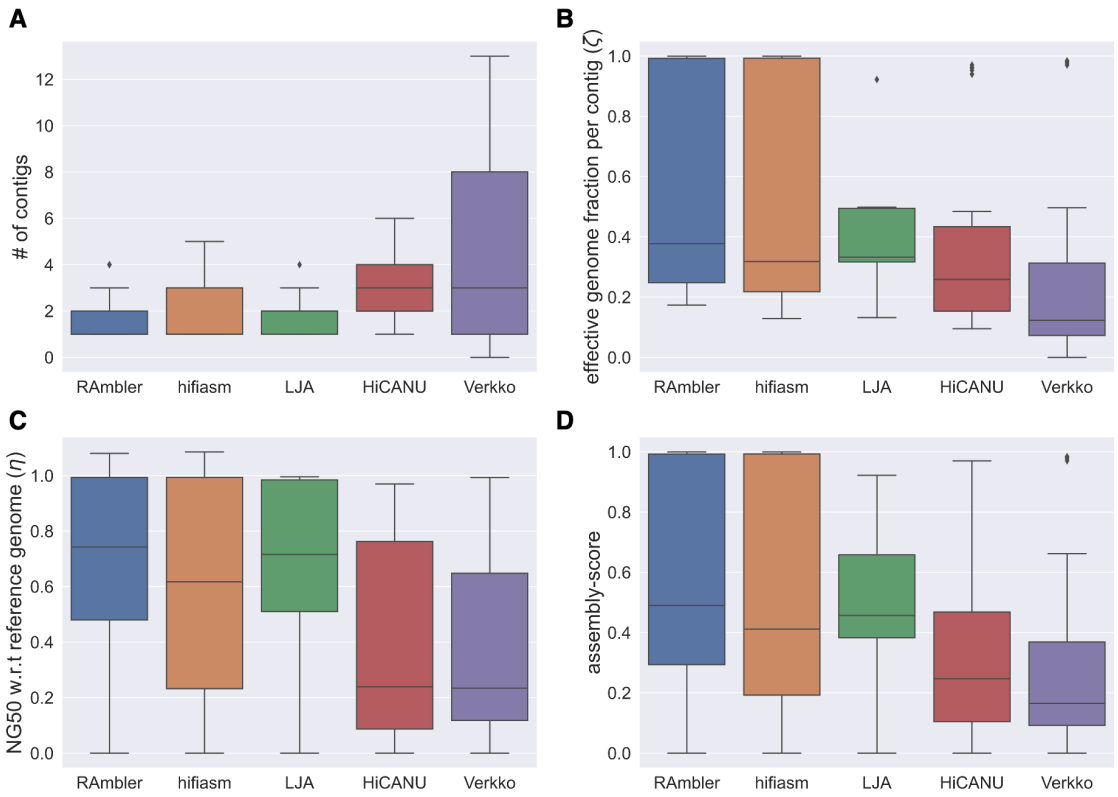

Figure 2: Assembly statistics for RAmbler, hifiasm, LJA, HiCANU, and Verkko for 36 different combinations of synthetic data with repetitive regions having repeat sizes $\{15, 20\}$ kb, number of copies $\{5, 10\}$, mutation rate $p = \{1/250, 1/500, 1/1000\}$ and coverage depth $\{20\times, 30\times,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

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

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

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

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

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×, based on a repetitive region with a variable-length (up to ±5% per copy) repeat unit of length $\{15, 20, 25\}$ kb, and $p = \{1/250, 1/500, 1/1000\}$; numbers in bolt indicate the best assembly score on each row

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

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

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

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

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

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

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.

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

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

Discussion

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

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

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

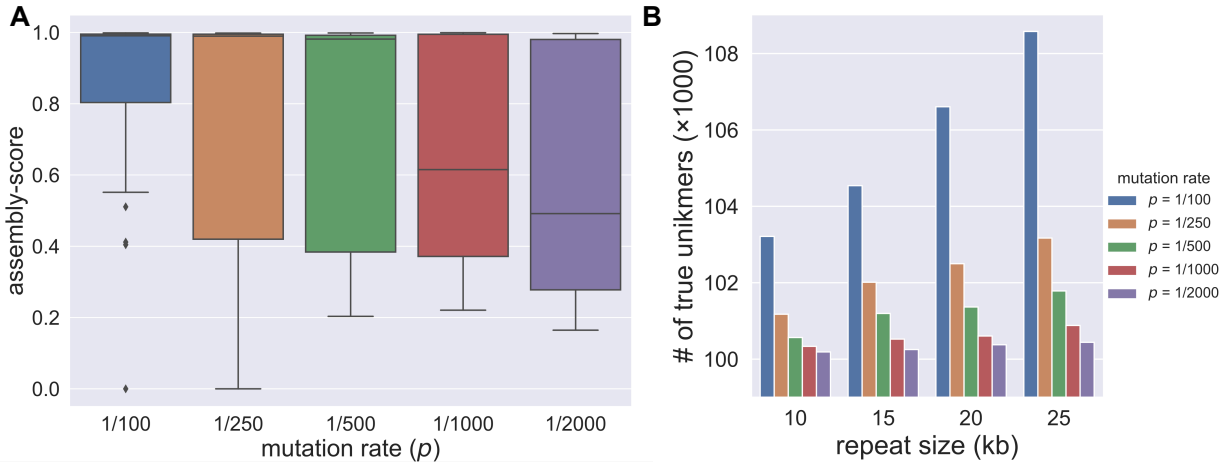

Figure 3: Relationship among RAmbler's performance, mutation rate p, and number of true uni$k$mers; **(A)** Performance of RAmbler for several choices of the mutation rate p over 27 different combinations of repetitive regions with repeat sizes $\{10, 15, 20\}$ kb, number of copies $\{2, 5, 10\}$, coverage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

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

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

Methods

Problem Formulation

We assume that (1) the genome G contains n repetitive regions $\{R_1, \ldots, R_i, \ldots, R_n\}$, (2) each repetitive region R_i is composed of t_i tandem copies of a string α_i, (3) each tandem copy has sufficient variations that allows it to be distinguished from another copy; we assume that each copy contains SNPs with probability 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

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

Repeat Identification

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

Analysis of k-mer Distribution

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

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

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

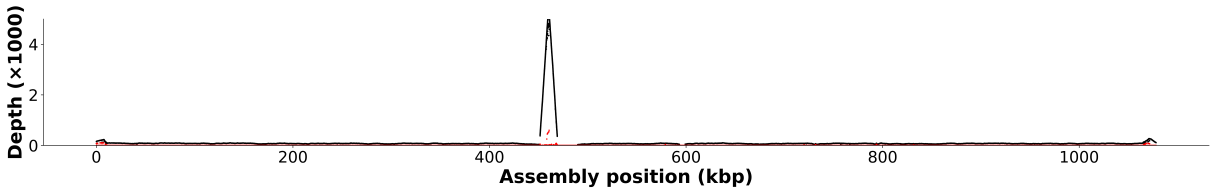

Figure 4: PacBio HiFi mapping coverage for Chromosome XII in *Saccharomyces cerevisiae* illustrated using NucFreq; the coverage spike indicates the presence of a repetitive region which is known to contain ∼150 tandemly repeated copies of a 9.1 kb rDNA unit

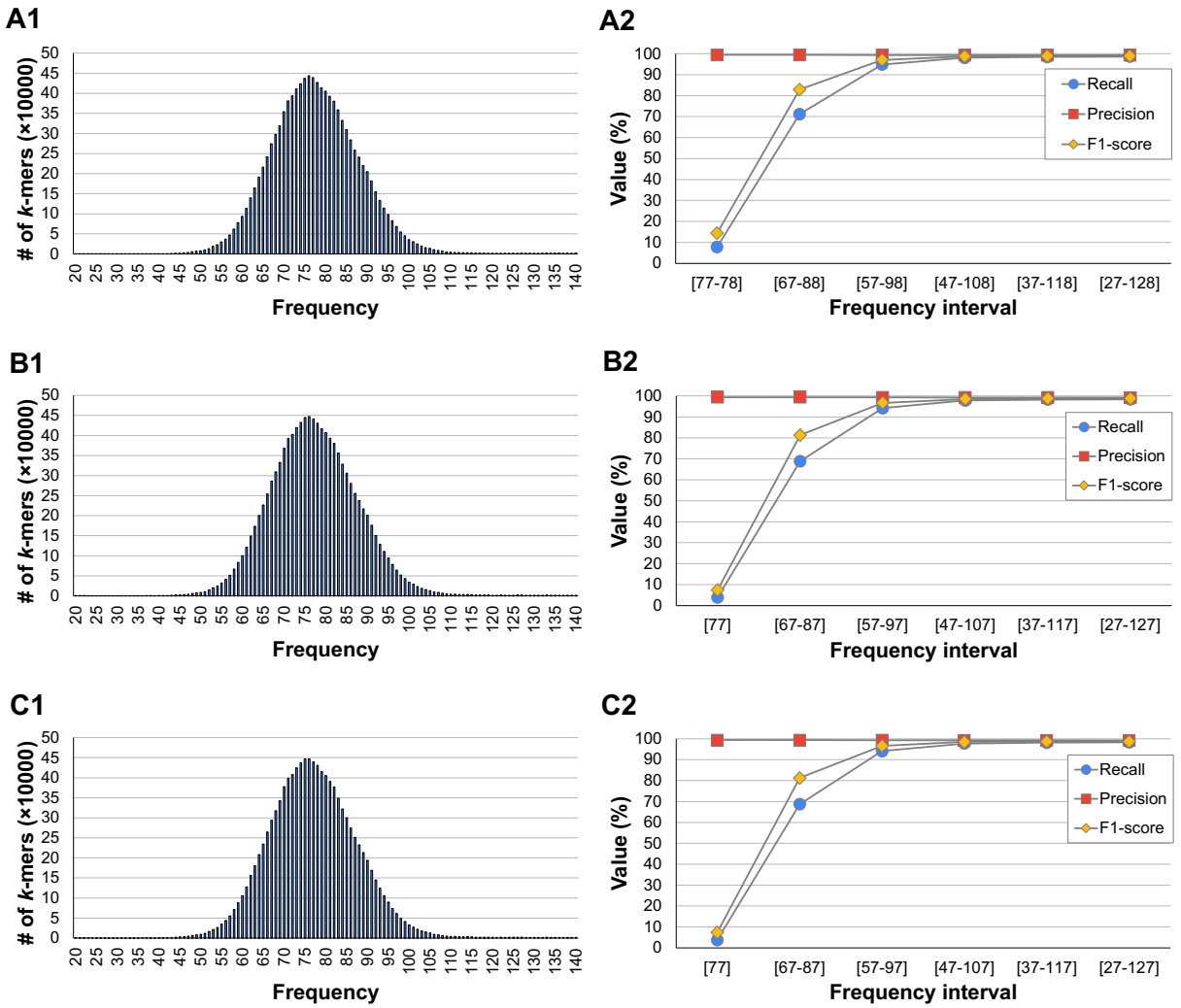

Figure 5: PacBio HiFi k-mer distribution for $k = 17$ **(A1)**, $k = 21$ **(B1)** and $k = 25$ **(C1)** for *Saccharomyces cerevisiae*; precision, recall, and F1-score for uni$k$mers when $k = 17$ **(A2)**, $k = 21$ **(B2)** and $k = 25$ **(C2)** for longer and longer intervals centered at the average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

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

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

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 uni$k$mers 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 uni$k$mers 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 uni$k$mers. 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 uni$k$mers with high probability (Figure 6, step B).

C. Barcode reads. RAmbler uses the set of uni$k$mers to barcode the HiFi reads T_r (Figure 6, step C). RAmbler searches for exact occurrences of the uni$k$mers in the reads T_r or their reverse complement. The set of uni$k$mers 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 uni$k$mer 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 uni$k$mers. This pairwise comparison allows RAmber to determine which reads are overlapping. Two reads are overlapping if they share at least th uni$k$mers, and the set of relative distances between the shared uni$k$mers match within a tolerance up to to base pairs. Overlaps are stored in the *barcode graph*: each node

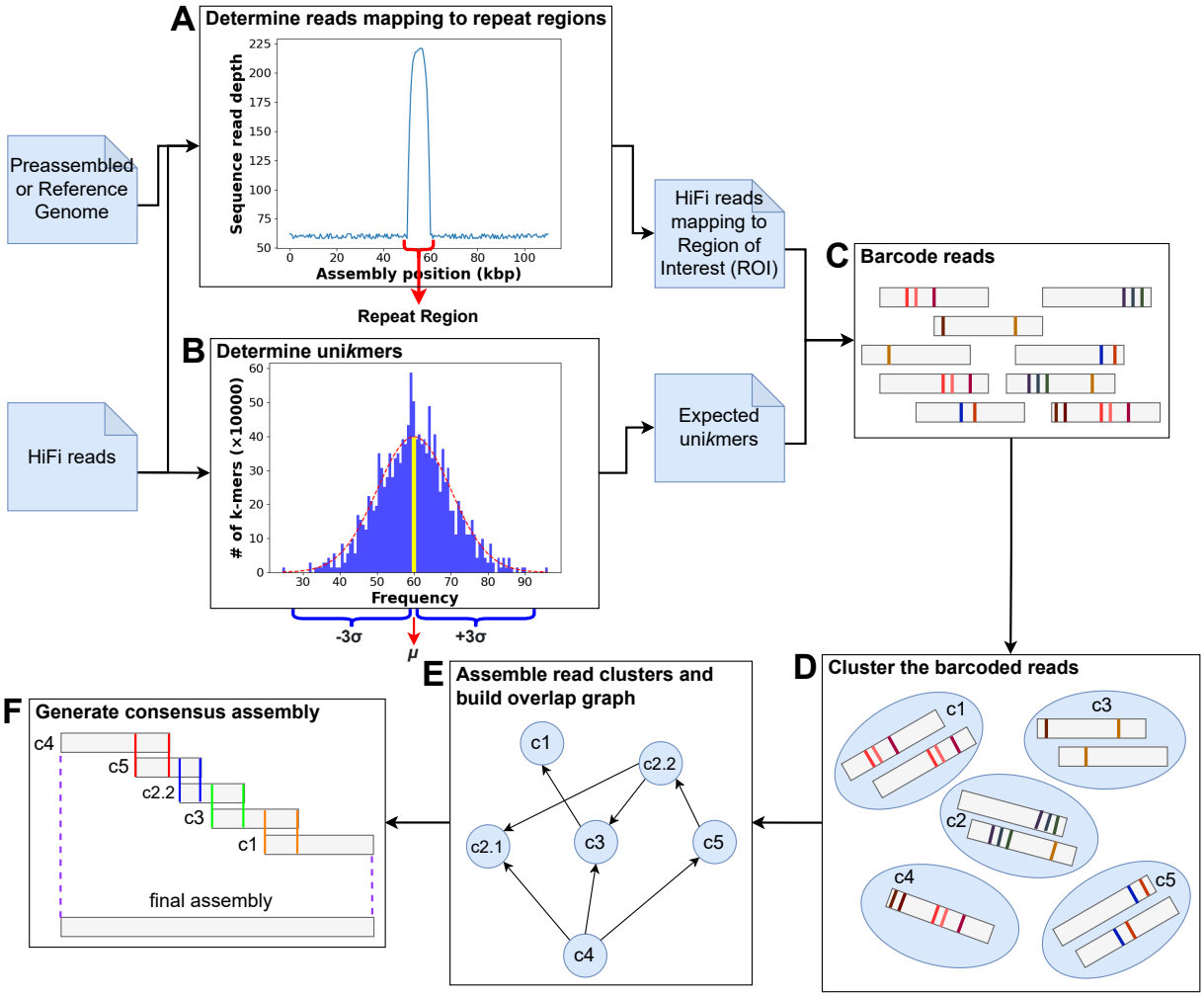

Figure 6: The algorithmic pipeline used in RAmbler

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

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

(Supplemental Figure 18). Given an edge (u, v) in the overlap graph, its type t is (1) "+" when there is an overlap between contig u and contig v, (2) "−" when there is an overlap between the reverse complement of contig u and contig v, (3) "$*$" when there is an overlap between contig u and the reverse complement of contig v.

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

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

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

A summary of RAmbler's main parameters k, t, th, to, and mo with their default values is shown in Supplemental Table 13. The optimization of these parameters is discussed in Section "Synthetic Data Generation and Parameter Optimization".

Performance Metrics for Real Data

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

Assembly quality index (AQI)

$$\mathrm{AQI} = 100e^{-0.1N/L} \tag{1}$$

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

Normalized error count in a window (N_w)

$$N_w = \sum_{i=1}^{m} i^{-1} \tag{2}$$

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

Performance Metrics for Synthetic Data

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

Effective genome fraction per contig (ζ)

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

$$\zeta = \frac{\text{genome\_fraction}}{\#\text{contigs} + \#\text{mis-assembled contigs}} \tag{3}$$

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

Normalized NG50 (η)

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

$$\eta = \frac{\text{NG50}}{|\text{reference\_genome}|} \tag{4}$$

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

Assembly score

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

$$\text{assembly\_score} = \frac{2\zeta\eta}{\zeta + \eta} \tag{5}$$

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

Acknowledgments

This project was supported in part by NSF IIS #2444456, NSF CBET #2225878, and NIH 1-R01-AI169543-01 to SL. The authors thank the anonymous reviewers what helped to improve the quality of this manuscript.
Author contributions. S.C., G.L. and S.L. conceptualized and designed the study.; S.C. wrote and tested the code; G.L. provided guidance on the experimental design on the human genome; S.C. carried out the experiments and generated figures and tables; S.C. and S.L. wrote the manuscript; all authors read and approved the final manuscript.

Competing interest statement

The authors declare no competing interests.

Software availability

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

References

Altemose N et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science*. **376**: eabl4178. doi:10.1126/science.abl4178.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, and O'Grady J. 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33**: 296–300.

Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, and Pevzner PA. 2022. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* **40**: 1075–1081.

Bustos BI, Billingsley K, Blauwendraat C, Gibbs JR, Gan-Or Z, Krainc D, Singleton AB, Lubbe SJ, and (IPDGC) IPDGC. 2022. Genome-wide contribution of common short-tandem repeats to Parkinson's disease genetic risk. *Brain*. **146**: 65–74.

Bzikadze AV and Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**: 1309–1316.

Chaisson MJ, Mukherjee S, Kannan S, and Eichler EE. 2017. Resolving multicopy duplications de novo using polyploid phasing. *Proceedings of RECOMB*. **10229**: 117–133.

Chan S, Wang W, Hallers B ten, Peters S, Gaiero P, Jong H de, Perez G, Hastie A, and Cao H. 2015. Detection, Characterization, and Biological Analysis of Long Tandem Repeats Using Nanochannel Technology. In *poster at Plant and Animal Genome conference*.

Cheng H, Concepcion GT, Feng X, Zhang H, and Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*. **18**: 170–175.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, and Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**: 265–270.

Eid J et al. 2008. Real-time DNA sequencing from single polymerase molecules. *Science*. **323**: 133–138.

Garg S et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**: 309–312.

Goel M, Sun H, Jiao WB, and Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**: 277.

Gurevich A, Saveliev V, Vyahhi N, and Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. **29**: 1072–1075.

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**: 286–298.

Hoyt SJ et al. 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*. **376**: eabk3112. doi:10.1126/science.abk3112.

Huddleston J et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**: 688–696.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*. **409**: 860–921.

Ishiura H et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat. Genet.* **51**: 1222–1232.

Johnston M et al. 1997. The nucleotide sequence of Saccharomyces cerevisiae chromosome XII. *Nature*. **387**: 87–90.

Jurka J, Kapitonov VV, Kohany O, and Jurka MV. 2007. Repetitive Sequences in Complex Genomes: Structure and Evolution. *Annu. Rev. Genomics Hum. Genet.* **8**: 241–259.

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. *Bioinfor-*
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.

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.

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