scientific reports

Check for updates

OPEN

Scalable long read self-correction and assembly polishing with multiple sequence alignment

Pierre Morisse[1]✉, Camille Marchet[2], Antoine Limasset[2], Thierry Lecroq[3] & Arnaud Lefebvre[3]

Third-generation sequencing technologies allow to sequence long reads of tens of kbp, that are expected to solve various problems. However, they display high error rates, currently capped around 10%. Self-correction is thus regularly used in long reads analysis projects. We introduce CONSENT, a new self-correction method that relies both on multiple sequence alignment and local de Bruijn graphs. To ensure scalability, multiple sequence alignment computation benefits from a new and efficient segmentation strategy, allowing a massive speedup. CONSENT compares well to the state-of-the-art, and performs better on real Oxford Nanopore data. Specifically, CONSENT is the only method that efficiently scales to ultra-long reads, and allows to process a full human dataset, containing reads reaching up to 1.5 Mbp, in 10 days. Moreover, our experiments show that error correction with CONSENT improves the quality of Flye assemblies. Additionally, CONSENT implements a polishing feature, allowing to correct raw assemblies. Our experiments show that CONSENT is 2-38x times faster than other polishing tools, while providing comparable results. Furthermore, we show that, on a human dataset, assembling the raw data and polishing the assembly is less resource consuming than correcting and then assembling the reads, while providing better results. CONSENT is available at https://github.com/morispi/CONSENT.

Third-generation sequencing technologies Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have become widely used since their inception in 2011. In contrast to second-generation technologies, producing reads reaching lengths of a few hundreds base pairs, they allow the sequencing of much longer reads (10 kbp on average[1], and up to $>$ 1 million bps[2]). These long reads are expected to solve various problems, such as contig and haplotype assembly[3,4], scaffolding[5], and structural variant calling[6]. However, they are very noisy. More precisely, basic ONT and non-CCS PacBio reads can reach error rates of 10 to 30%, whereas second-generation short reads usually display error rates of 1%. The error profiles of these long reads are also much more complex than those of the short reads. Indeed, they are mainly composed of insertions and deletions, whereas short reads mostly contain substitutions. As a result, error correction is often required, as the first step of projects dealing with long reads. As the error profiles and error rates of the long reads are much different from those of the short reads, correcting long reads requires specific algorithmic developments.

The error correction of long reads has thus been tackled by two main approaches. The first approach, hybrid correction, makes use of additional short reads data to perform correction. The second approach, self-correction, aims at correcting the long reads solely based on the information contained in their sequences.

Hybrid correction methods rely on different techniques such as: 1. Alignment of short reads to the long reads (CoLoRMAP[7], HECiL[8]) ; 2. Exploration of de Bruijn graphs, built from short reads k-mers (LoRDEC[9], Jabba[10], FMLRC[11]) ; 3. Alignment of the long reads to contigs built from the short reads (MiRCA[12], HALC[13]) ; 4. Hidden Markov Models, initialized from the long reads sequences and trained using the short reads (Hercules[14]) ; 5. Combination of different strategies (NaS[15] (1 + 3), HG-CoLoR[16] (1 + 2)).

Self-correction methods usually build around the alignment of the long reads against each other (PBDAGCon[17], PBcR[18]). We give further details on the state-of-the-art of self-correction in the "Related works ".

As first long reads sequencing experiments resulted in highly erroneous long reads (15–30% error rates on average), most methods relied on the additional use of short reads data. As a result, hybrid correction used to be much more widespread. Indeed, in 2014, for five hybrid correction tools, only two self-correction tools were available.

However, third-generation sequencing technologies evolve fast, and now manage to produce long reads reaching error rates of 10–12%. Moreover, the evolution of long-read sequencing technologies also allows to

[1]Univ Rennes, Inria, CNRS, IRISA, 35000 Rennes, France. [2]Univ. Lille, CNRS, UMR 9189 - CRIStAL, 59000 Lille, France. [3]Normandie Univ, UNIROUEN, LITIS, 76000 Rouen, France. ✉email: pierre.morisse@inria.fr

produce higher throughputs of data, at a reduced cost. Consequently, such data became more widely available. As a result, self-correction can now efficiently be used in place of hybrid correction in data analysis projects dealing with long reads.

Related works. Since third-generation sequencing technologies evolve fast, and now reach lower error rates, various efficient self-correction methods have recently been developed. Most of them share a common first step of computing overlaps between the long reads, which can be performed in two different ways. First, a mapping approach can be used, to simply provide the positions of similar regions of the long reads (Canu[19], MECAT[20], FLAS[21]). Second, an alignment approach can be used, to provide the positions of similar regions, but also their actual base to base correspondence in terms of matches, mismatches, insertions, and deletions (PBDAGCon, PBcR, Daccord[22]). A directed acyclic graph (DAG) is then usually built, in order to summarize the 1V1 alignments and compute consensus, after recomputing actual alignments of mapped regions, if necessary. Other methods rely on de Bruijn graphs, either built from small windows of the alignments (Daccord), or directly from the long reads sequences with no prior overlapping step (LoRMA[23]). These graphs are explored, using the solid *k*-mers (*i.e. k*-mers occurring more frequently than a given threshold) from the reads as anchor points, in order to correct low quality, weak *k*-mers regions.

However, methods relying on direct alignment of the long reads are prohibitively time and memory consuming, and current implementations thus do not scale to large genomes. Methods solely relying on de Bruijn graphs, and avoiding the overlapping step altogether, usually require deep long reads coverage, since the graphs are usually built from large, solid *k*-mers. As a result, methods relying on a mapping strategy constitute the core of the current state-of-the-art for long read self-correction.

Contribution. We present CONSENT, a new self-correction method that combines different efficient approaches from the state-of-the-art. CONSENT indeed starts by computing multiple sequence alignments (MSA) between overlapping regions of the long reads, in order to compute consensus sequences. These consensus sequences are then polished with local de Bruijn graphs, in order to correct remaining errors, and further reduce the final error rate. Moreover, unlike current state-of-the-art methods, CONSENT computes actual MSA, using a method based on partial order graphs[24]. We also introduce an efficient segmentation strategy based on *k*-mer chaining, which allows to reduce the time footprint of the MSA. This segmentation strategy thus allows to compute scalable MSA. In particular, it allows CONSENT to efficiently scale to ONT ultra-long reads.

Our experiments show that CONSENT compares well to the latest state-of-the-art self-correction methods, and outperforms them on real ONT datasets. In particular, they show that CONSENT is the only method able to efficiently scale to ONT ultra-long reads, allowing to perform correction of a full human dataset, containing reads reaching up to 1.5 Mbp in 10 days.

Additionally, CONSENT is also able to polish assemblies generated from raw long reads. Our experiment on a full human dataset shows that assembling the raw data and polishing the assembly is less resource consuming than correcting and then assembling the data, while offering better results. Our experiments also show that CONSENT outperforms state-of-the-art assembly polishing tools in terms of resource consumption, while providing comparable results.

Results

Impact of the segmentation strategy. Before comparing CONSENT to state-of-the-art self-correction tools, we first validate our segmentation strategy. To this aim, we simulated a 50x coverage of long reads from *E.coli*, with a 12% error rate, using SimLoRD[25]. The following parameters were used for the simulation: `--probability-threshold 0.3 --prob-ins 0.145 --prob-del 0.06`, and `--prob-sub 0.02`. We then ran the CONSENT pipeline, with, and without the segmentation strategy. Results of this experiment are given in Supplementary Table 1. We obtained these results using ELECTOR[26], a tool specifically designed to precisely measure correction accuracy on simulated data. In particular, ELECTOR reports metrics such as recall (number of modified bases among erroneous bases in the original data), precision (number of properly corrected bases among bases modified by the error-correction tool), and error rate before and after correction. These results show that, compared to the regular MSA implementation, our segmentation strategy allows a 47x speedup, reducing the runtime from 5 h 31 min to 7 min. Moreover, our segmentation strategy also allows to reach lower memory consumption and higher quality. In particular, the post-correction error rate is decreased by 1.77x, and the precision increases by almost 0.15%. This gain in quality can be explained by the fact that our segmentation strategy allows to get rid of spurious sequences and thus to compute more accurate alignments and consensus.

Comparison to the state-of-the-art. We compare CONSENT against state-of-the-art error correction methods. We include the following tools in the benchmark: Canu, Daccord, FLAS, and MECAT. We voluntarily exclude LoRMA from the comparison, as it tends to aggressively split the reads, and thus produce reads that are usually shorter than 900 bp. We however report LoRMA's result in Supplementary Tables S3 and S4. We also exclude hybrid error correction tools from the benchmark, as we believe it makes more sense to only compare self-correction tools. We performed experiments both on simulated and real data. We ran all tools with default or recommended parameters, and with a number of threads corresponding to the maximum number of cores available on the systems experiments were performed on. For CONSENT, we set the minimum support to define a window to 3, the window size to 500, the overlap size between consecutive windows to 50, the *k*-mer size used for chaining and polishing to 9, the solidity threshold for *k*-mers to 4, and the solidity threshold for the anchors chain computation to 8. Additionally, only windows for which at least two anchors could be found during the segmentation algorithm were processed.

Datasets. For our experiments, we used simulated PacBio long reads, as well as real PacBio and ONT long reads. PacBio simulated reads were generated with SimLoRD, using the previously mentioned parameters. We generated two datasets with a 12% error rate for *E. coli*, *S. cerevisiae* and *C. elegans*: one with a 30x coverage, and one with a 60x coverage, corresponding to typical sequencing depths in current long reads experiments. As for the real data, we used a 89x coverage dataset from *S. cerevisiae*, a 63x coverage dataset from *D. melanogaster*, and a 29x coverage from *H. sapiens* chr 1, containing ultra-long reads, reaching lengths up to 340 kbp. Human data were sequenced by[2]] and are publicly available with accession number PRJEB23027 (release 4). *D. melanogaster* data are publicly available with accession number SRX3676783. Real *S. cerevisiae* data are publicly available with accession number SRR9617898. Further details for all the datasets are given in Supplementary Table S1. We used the reference genomes K-12 substr. MG1655 for *E. coli*, S288C for *S. cerevisiae*, Bristol N2 for *C. elegans*, BDGP Release 6 for *D. melanogaster*, and GRCh38 for *H. sapiens*. Further details on the reference sequences are given in Supplementary Table S2.

Comparison on simulated data. To precisely assess the accuracy of the different correction methods, we first tested them on the simulated PacBio datasets. ELECTOR was used to evaluate the correction accuracy of each method. Correction statistics of all the aforementioned tools on the different datasets, along with their runtime and memory consumption, are given in Table 1. For methods having distinct, easily identifiable, steps for overlapping and correction (*i.e.* Daccord, MECAT and CONSENT), we additionally report runtime and memory consumption of these two processes apart. We ran all the correction experiments on a computer equipped with 16 2.39 GHz cores and 32 GB of RAM. All tools were thus run with 16 threads.

Daccord clearly performed the best in terms of number of bases and quality, outperforming all the other methods on the *E. coli* and the *S. cerevisiae* datasets. However, the overlapping step, relying on actual alignment of the long reads against each other, consumed high amounts of memory, 3x to 11x more than CONSENT or MECAT mapping strategies. As a result, Daccord could not scale to the *C. elegans* datasets, DALIGNER reporting an error upon start, even when run on a cluster node equipped with 128 GB of RAM. On the contrary, Canu displayed the highest error rates on all the datasets, except on the *C. elegans* dataset with a 30x coverage, but consumed relatively stable, low amounts of memory. In particular, on the two *C. elegans* datasets, it displayed the lowest memory consumption among all the other methods.

MECAT performed the best in terms of runtime, outperforming all the other tools on all the datasets. Its overlapping strategy was also highly efficient, and displayed the lowest memory consumption among all the other strategies, on all the datasets. However, compared to Minimap2 (the overlapping strategy adopted in CONSENT) MECAT's overlapping strategy displayed higher runtimes, although it remained faster than Daccord's DALIGNER. Minimap2's memory consumption, however, was larger than that of MECAT's overlapping strategy, on all the datasets. The memory consumption of Minimap2 can nonetheless easily be reduced, at the expense of a slightly larger runtime, by decreasing the size of the index used for computing the overlaps, which CONSENT sets to 1 Gbp by default.

Compared to both FLAS and CONSENT, MECAT displayed lower number of bases on all the datasets. As for FLAS, this can be explained by the fact that it is a MECAT wrapper, allowing to retrieve additional overlaps, and thus correct a greater number long reads. As a result, since it relies on MECAT's error correction strategy, FLAS displayed highly similar memory consumption. Runtime was however higher, due to the additional steps allowing to retrieve supplementary overlaps, and to the resulting higher number of reads to correct. Numbers of bases and error rates of FLAS and CONSENT were highly similar on all the datasets, varying by 0.12% at most, on the *S. cerevisiae* dataset with a 30x coverage. CONSENT was also faster than FLAS on the *E. coli* and *S. cerevisiae* datasets. However, on the *C. elegans* datasets, CONSENT displayed slightly higher runtimes.

As for the memory consumption of the error correction step, CONSENT was less efficient than MECAT on most datasets. This can be explained by the fact that CONSENT loads the correction jobs into a queue of default size 100,000. This queue thus loads the necessary data for 100,000 correction tasks into memory at once. We then iteratively allocate new tasks to threads as they become available. Reducing the size of this queue would allow CONSENT to consume less memory, at the expense of a slightly higher runtime, since the queue would have to be repopulated more often.

Comparison on real data. We then evaluated the different correction methods on a real PacBio *S. cerevisiae* dataset, as well as on larger, real ONT datasets from *D. melanogaster* and *H. sapiens* (chr 1). For these datasets, we not only evaluate how well the corrected long reads realign to the reference genome, but also how well they assemble. For the alignment assessment, we report how many reads were corrected, their total number of bases, their N50, the proportion of corrected reads that could be aligned, the average identity of the alignments, as well as the genome coverage, that is, the percentage of bases of the reference genome to which at least a nucleotide aligned. For the assembly assessment, we report the overall number of contigs, the number of contigs that could be aligned, the NGA50 and NGA75, and, once again, the genome coverage. We obtained alignment statistics using ELECTOR's second module, which performs alignment to the reference genome with Minimap2. We performed assemblies using Minimap2 and Miniasm[27]. Moreover, for the *H. sapiens* (chr 1) dataset, we performed additional assembly experiments using the more modern and sophisticated assembler Flye[28]. Assembly statistics were obtained with QUAST-LG[29]. Results are given in Table 2 for the alignment assessment, and in Table 3 for the assembly assessment. Runtimes and memory consumption of the different methods are also given in Table 2. As for the simulated data, we report runtime and memory consumption of the overlapping and correction steps apart, when possible. We ran all the correction experiments on a cluster node equipped with 28 2.39 GHz cores and 128 GB of RAM. All tools were thus run with 28 threads.

Dataset	Corrector	Number of bases (Mbp)	Error rate (%)	Recall (%)	Precision (%)	Overlapping		Correction		Total	
						Runtime	Memory (MB)	Runtime	Memory (MB)	Runtime	Memory (MB)
E. coli 30x	Original	140	12.2862	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	130	0.4156	99.7647	99.5887	_	_	_	_	19 min	4613
	Daccord	**131**	**0.0248**	**99,9965**	**99,9757**	1 min	6813	13 min	639	14 min	6813
	FLAS	130	0.2720	99.9291	99.7385	_	_	_	_	12 min	1639
	MECAT	107	0.2569	99.9302	99.7533	25 s	**1600**	**1 min 14 s**	1083	**1 min 39 s**	**1600**
	CONSENT	130	0.3111	99.9328	99.6934	**18 s**	2345	7 min 16 s	**532**	7 min 34 s	2345
E. coli 60x	Original	279	12.2788	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	219	0.7404	99.4781	99.2658	_	_	_	_	24 min	3674
	Daccord	**261**	**0.0214**	**99.9971**	**99.9790**	3 min	18,450	51 min	1191	54 min	18,450
	FLAS	260	0.1547	99.9546	99.8526	_	_	_	_	38 min	2428
	MECAT	233	0.1714	99.9547	99.8362	**1 min**	**2387**	**4 min**	**1553**	**5 min**	**2387**
	CONSENT	259	0.1833	99.9771	99.8196	**1 min**	4693	25 min	1757	26 min	4693
S. cerevisiae 30x	Original	371	12.283	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	226	1.1052	99.1766	98.9036	_	_	_	_	29 min	3681
	Daccord	**348**	**0.1259**	**99.9874**	**99.8762**	7 min	31,798	1 h 12 min	3487	1 h 19 min	31,798
	FLAS	344	0.3272	99.9131	99.6843	_	_	_	_	29 min	2935
	MECAT	285	0.3040	99.9160	99.7072	**1 min**	**2907**	**4 min**	1612	**5 min**	**2907**
	CONSENT	344	0.4102	99.9192	99.5956	**1 min**	5519	21 min	**1503**	22 min	5519
S. cerevisiae 60x	Original	742	12.2886	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	599	0.7919	99.4488	99.2148	_	_	_	_	1 h 11 min	**3710**
	Daccord	**695**	**0.0400**	**99.9928**	**99.9606**	10 min	32,190	2 h 16 min	1160	2 h 26 min	32,190
	FLAS	689	0.2034	99.9418	99.8049	_	_	_	_	1 h 30 min	4984
	MECAT	616	0.2088	99.9428	99.7996	4 min	**4954**	**12 min**	**2412**	**16 min**	4954
	CONSENT	688	0.2897	99.9532	99.7145	**2 min**	11,378	1 h 11 min	4754	1 h 13 min	11,378
C. elegans 30x	Original	3006	12.2806	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	2773	0.5008	99.7103	99.5040	_	_	_	_	9 h 09 min	**6921**
	Daccord	_	_	_	_	_	_	_	_	_	_
	FLAS	2729	0.7613	99.8613	99.2541	_	_	_	_	3 h 07 min	10,565
	MECAT	2084	**0.3908**	**99.8903**	**99.6212**	27 min	**10,535**	**21 min**	**2603**	**48 min**	10,535
	CONSENT	**2789**	0.6495	99.8846	99.3596	**16 min**	16,711	3 h 40 min	5338	3 h 56 min	16,711
C. elegans 60x	Original	6024	12.2825	_	_	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	5112	0.7934	99.4573	99.2131	_	_	_	_	9 h 30 min	**7050**
	Daccord	_	_	_	_	_	_	_	_	_	_
	FLAS	5584	0.3997	99.9175	99.6104	_	_	_	_	10 h 45 min	13,682
	MECAT	4938	**0.2675**	**99.9258**	**99.7415**	1 h 28 min	**10,563**	**1 h 15 min**	**3775**	**2 h 43 min**	10,563
	CONSENT	**5587**	0.3858	99.9428	99.6201	**56 min**	15,732	12 h 50 min	7921	13 h 46 min	15,732

Table 1. Metrics output by ELECTOR on the simulated PacBio datasets. Daccord results are missing for the two *C. elegans* datasets, as DALIGNER failed to perform alignment, reporting an error upon start, even when ran on a cluster node with 28 2.4 GHz cores and 128 GB of RAM. Recall and precision are not reported for original reads, since they cannot be computed from uncorrected reads. Best results for each metric is highlighted in bold.

On these three datasets, Daccord failed to run, as DALIGNER could not perform alignment, for the same reason as for the simulated *C. elegans* datasets. For all the datasets, CONSENT corrected the largest number of reads, output the largest number of bases, and reached the largest genome coverage. On the *S. cerevisiae* dataset, CONSENT reached slightly lower alignment identity than the other tools. However, on the two more complex ONT datasets, CONSENT displayed the highest alignment identity. Additionally, the N50 of CONSENT was the highest on the *S. cerevisiae* dataset, and was higher than that of all the others methods, except Canu, on the two ONT datasets.

When it comes to runtime and memory consumption, MECAT outperformed all the other methods, as in the experiments on simulated data. However, it is worth noting that the runtime was comparable to other methods, and that CONSENT was the second fasted tool after MECAT on the two ONT datasets. Finally, MECAT reached the highest proportion of aligned reads, on all datasets, although CONSENT was close, since only 0.24–1.21% fewer reads could be aligned.

Moreover, on the *H. sapiens* (chr 1) dataset, CONSENT and Canu were the only tools able to deal with ultra-long reads. Indeed, other methods reported errors when attempting to correct the original dataset. As a result,

Dataset	Corrector	Number of reads	Number of bases (Mbp)	N50 (bp)	Aligned reads (%)	Alignment identity (%)	Genome coverage (%)	Overlapping		Correction		Total	
								Runtime	Memory (MB)	Runtime	Memory (MB)	Runtime	Memory (MB)
S. cerevisiae	Original	121,640	1083	12,048	96.27	84.63	99.65	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	74,658	644	11,171	99.67	96.67	99.18	_	_	_	_	1 h 24 min	3870
	Daccord[2]	_	_	_	_	_	_	_	_	_	_	_	_
	FLAS	80,087	665	10,815	99.78	97.65	99.16	_	_	_	_	1 h 02 min	6195
	MECAT	46,614	477	11,123	**99.84**	**97.91**	98.21	**2 min**	**7176**	**4 min**	**3023**	**6 min**	**7176**
	CONSENT	**116,391**	**965**	**11,331**	99.60	95.47	**99.52**	4 min	19,828	1 h 13 min	3108	1 h 17 min	19,828
D. melanogaster	Original	1,327,569	9064	11,853	85.52	85.43	98.47	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	829,965	6993	**12,694**	98.05	95.20	97.89	_	_	_	_	14 h 04 min	**10,295**
	Daccord[2]	_	_	_	_	_	_	_	_	_	_	_	_
	FLAS	855,275	7866	11,742	95.65	94.99	98.09	_	_	_	_	10 h 18 min	18,820
	MECAT	827,490	7288	11,676	**99.87**	96.52	97.34	**28 min**	**13,443**	**1 h 26 min**	7724	**1 h 54 min**	13,443
	CONSENT	**1,096,046**	**8308**	12,181	98.66	**97.17**	**98.20**	1 h 07 min	31,282	8 h 53 min	**5639**	10 h 00 min	31,282
H. sapiens (chr 1)	Original	1,075,867	7256	10,568	88.24	82.40	92.46	N/A	N/A	N/A	N/A	N/A	N/A
	Canu	717,436	5605	**11,002**	97.60	90.40	92.33	_	_	_	_	22 h 06 min	12,802
	Daccord[1,2]	_	_	_	_	_	_	_	_	_	_	_	_
	FLAS[1]	670,708	5695	10,198	99.06	91.00	92.37	_	_	_	_	4 h 57 min	14,957
	MECAT[1]	655,314	5479	10,343	**99.95**	91.69	91.44	26 min	**11,075**	**1 h 27 min**	**4591**	**1 h 53 min**	**11,075**
	CONSENT	**893,738**	**6502**	10,870	98.89	**93.15**	**92.38**	**23 min**	17,290	2 h 47 min	4645	3 h 10 min	17,290

Table 2. Statistics of the real long reads, before and after correction with the different methods. Best results for each metric is highlighted in bold. [1]Reads longer than 50 kbp were filtered out, as ultra-long reads caused the programs to stop with an error. There were 1824 such reads in the original dataset, accounting for a total number of 135,364,312 bp. [2]Daccord could not be run on these two datasets, due to errors reported by DALIGNER.

in order to allow these methods to perform correction, we had to manually remove the reads longer than 50 kbp. There were 1,824 such reads, accounting for a total number of 135,364,312 bp. However, even if it managed to scale to the correction of ultra-long reads, Canu was almost seven times slower than CONSENT, making CONSENT the only tool to efficiently scale to ultra-long reads.

On the *S. cerevisiae* dataset, the assembly yielded from FLAS corrected reads outperformed all the other assemblies in terms of number of contigs. Oppositely, the assembly yielded from CONSENT corrected reads displayed the largest number of contigs. It also displayed more errors per 100 kbp and a slightly higher number of misassemblies compared to other assemblies. Canu corrected reads yielded the assembly covering the largest proportion of the reference genome, whereas MECAT corrected reads yielded the assembly displaying the lowest number of misassemblies and errors per 100 kbp. However, all assemblies were comparable in terms of NGA50 and NGA75, and CONSENT even outperformed MECAT in terms of NGA75.

On the *D. melanogaster* dataset, the assembly yielded from CONSENT corrected reads outperformed all the other assemblies in terms of number of contigs, NGA50, NGA75, as well as error rate per 100 kbp. In particular, the NGA50 of the CONSENT assembly was 2.1-6.8 times larger than that of other assemblies. The assembly yielded from Canu corrected reads outperformed all the other assemblies in terms of genome coverage, but was composed of the highest number of contigs, after the assembly obtained from the raw reads. However, the genome coverage of the CONSENT assembly was slightly larger than that of FLAS and MECAT. Finally, in terms of misassemblies, the assembly generated from MECAT corrected reads outperformed the other assemblies, although they all remained comparable on this metric.

On the *H. sapiens* (chr 1) dataset, the assembly obtained from CONSENT corrected reads once again outperformed all the other assemblies in terms of number of contigs, NGA50, and NGA75. In particular, the NGA50 of the CONSENT assembly was almost 1.6-4.3 times larger than that of other assemblies. However, 12 contigs of the CONSENT assembly could not be aligned to the reference. As a result, compared to the assemblies obtained from FLAS and MECAT corrected reads, the assembly yielded from the CONSENT corrected reads covered 2.6% less of the reference sequence, and displayed a higher error rate per 100 kbp. These unaligned contigs and differences could likely be reduced by further adapting both CONSENT and Miniasm parameters. Moreover, the assembly yielded from CONSENT corrected reads displayed a slightly higher number of misassemblies, although it remained comparable to that of other assemblies. On this metric, the assembly generated from MECAT corrected reads once again slightly outperformed other assemblies.

In addition, as we previously mentioned, we also performed additional assembly experiments using the more modern and sophisticated assembler Flye on the *H. sapiens* (chr 1) dataset. Here, the assembly yielded from CONSENT corrected reads displayed the smallest number of contigs, and the lowest number of errors per 100

Dataset	Corrector	Contigs	Aligned contigs (%)	NGA50 (bp)	NGA75 (bp)	Genome coverage (%)	Errors / 100 kbp	Misassemblies
S. cerevisiae	Original	29	93.10	408,751	179,653	84.67	10,514	45
	Canu	27	**100.00**	549,622	426,490	**96.64**	1291	53
	Daccord[1]	_	_	_	_	_	_	_
	FLAS	**21**	**100.00**	542,516	**447,884**	95.63	1423	43
	MECAT	22	**100.00**	**550,249**	305,358	95.81	**1154**	**38**
	CONSENT	37	97.30	524,568	419,018	94.68	1548	60
D. melanogaster	Original	423	96.45	864,011	159,590	83.22	10,690	810
	Canu	410	92.93	2,757,690	822,577	**92.95**	1896	845
	Daccord[1]	_	_	_	_	_	_	_
	FLAS	407	98.53	1,123,346	363,017	92.16	2736	838
	MECAT	310	**99.68**	1,414,076	480,297	92.02	1731	**554**
	CONSENT	**287**	98.61	**5,906,563**	**1,143,682**	92.26	**1502**	804
H. sapiens (chr 1)	Original[2]	201	93.53	1,008,692	_	77.52	11,318	98
	Canu	361	98.61	946,029	245,015	94.85	4689	49
	Daccord[1]	_	_	_	_	_	_	_
	FLAS	259	**100.00**	1,378,242	287,113	94.89	4413	50
	MECAT	237	100.00	1,698,601	289,968	**94.97**	**4404**	**44**
	CONSENT	**154**	92.21	**2,777,701**	**736,664**	92.30	4486	78
H. sapiens (chr 1) (Flye)	Original	319	97.81	11,231,592	2,893,011	**96.24**	2204	44
	Canu	181	98.90	3,022,928	1,237,577	95.69	2521	**27**
	Daccord[1]	_	_	_	_	_	_	_
	FLAS	169	**99.41**	7,733,334	2,298,510	95.73	2677	33
	MECAT	170	98.82	7,625,451	1,475,937	95.61	2732	36
	CONSENT	**153**	98.69	**12,088,173**	**3,089,752**	95.71	**2057**	28

Table 3. Statistics of the assemblies generated from the raw and corrected real long reads. Best results for each metric is highlighted in bold. [1]As previously mentioned, Daccord results on the three datasets are absent, since it could not be run. [2]For the assembly of the original reads on the *H. sapiens* (chr 1) dataset, QUAST-LG did not provide a metric for the NGA75.

kbp. Its number of misassemblies was also smaller than that of all other assemblies, except the one generated from Canu corrected reads. Morever, the proportion of aligned contigs and the genome coverage were highly similar for all assemblies. Interestingly, only the assembly generated from CONSENT corrected reads allowed to reach larger NGA50 and NGA75 than the assembly generated from the raw reads. Compared to other correction tools, CONSENT is the only method allowing to enhance the quality of the Flye assembly in such a way. Indeed, all the other assemblies, despite displaying a smaller number of contigs than the assembly generated from raw reads, reached smaller NGA50 and NGA75. These results thus underline the fact that Flye can benefit from CONSENT correction to generate higher quality assemblies.

Assembly polishing. As an additional feature, CONSENT also allows to perform assembly polishing. The process is pretty straightforward. Indeed, instead of computing overlaps between the long reads, as presented in the previous sections, overlaps are simply computed between the assembled contigs and the long reads used for the assembly. The rest of the pipeline remains the same. We present assembly polishing results on the simulated *E. coli*, *S. cerevisiae*, and *C. elegans* datasets with a 60x coverage, as well as on the real *S. cerevisiae*, *D. melanogaster* and *H. sapiens* (chr 1) datasets. We compare CONSENT to RACON[30], a state-of-the-art assembly polishing method. Results are presented in Table 4. We ran all the polishing experiments on a cluster node equipped with 28 2.39 GHz cores and 128 GB of RAM. All tools were thus run with 28 threads. Moreover, we only compare the runtimes of the actual polishing steps of the tools. Thus, the runtime of Minimap2, and the runtime of CONSENT pre-processing steps (sorting and reformatting the overlaps file) are not taken into account in the comparisons. Although these pre-processing steps can be slow for large datasets and large overlaps files, and are not required by RACON, we would like to underline that they allow to avoid the burden of loading the full overlaps file into memory, as required by RACON. The benefits of these pre-processing steps can be confirmed by comparing the memory consumption of CONSENT and RACON, as commented below. Moreover, to further emphasize the interest of these pre-processing steps, additional experiments on another human dataset (accession number NA12878, release 6), which are not presented here, have shown RACON could require more than 2 TB of RAM, while CONSENT displayed a peak of 39 GB.

These results show that CONSENT outperformed RACON in terms of quality of the results, especially dealing better with errors, and thus greatly reducing the error rate per 100 kbp, on the *E. coli*, *S. cerevisiae*, and *C. elegans* datasets. RACON outperformed CONSENT in terms of NGA50, NGA75, genome coverage, and number of misassemblies, but the two methods were highly comparable on these three datasets. Oppositely, RACON

Dataset	Method	Contigs	Aligned contigs (%)	NGA50 (bp)	NGA75 (bp)	Genome coverage (%)	Errors / 100 kbp	Misassemblies	Runtime	Memory (MB)
E. coli 60x	Original	**1**	**100.00**	**4,939,014**	**4,939,014**	**99.91**	10,721	**0**	N/A	N/A
	RACON	**1**	**100.00**	4,663,914	4,663,914	99.90	499	**0**	5 min 55 sec	**643**
	CONSENT	**1**	**100.00**	4,638,842	4,638,842	**99.91**	**117**	**0**	**30 sec**	786
S. cerevisiae 60x	Original	**29**	**100.00**	**579,247**	**456,470**	**96.14**	10,694	**5**	N/A	N/A
	RACON	**29**	**100.00**	539,472	346,116	96.09	637	**5**	15 min 47 sec	1703
	CONSENT	**29**	**100.00**	532,189	332,977	96.05	**217**	7	**1 min 49 sec**	**1052**
C. elegans 60x	Original	**47**	**100.00**	5,201,998	2,511,520	**99.78**	10,974	5	N/A	N/A
	RACON	**47**	97.87	**6,405,523**	**2,726,529**	99.74	819	**2**	2 h 24 min	14,288
	CONSENT	**47**	**100.00**	6,340,451	2,699,930	99.73	**375**	3	**11 min**	**3648**
S. cerevisiae real	Original	**29**	93.10	408,751	179,653	84.67	10,514	**45**	N/A	N/A
	RACON	**29**	**100.00**	518,943	330,455	93.74	**1193**	52	1 h 17 min	3708
	CONSENT	**29**	**100.00**	**522,799**	**411,537**	**94.23**	1400	**50**	**2 min**	**1667**
D. melanogaster	Original	**423**	96.45	864,011	159,590	83.20	10,690	**810**	N/A	N/A
	RACON	422	98.34	**1,446,703**	**552,532**	**93.03**	**961**	1013	3 h 29 min	19,508
	CONSENT	422	**98.82**	1,235,999	465,133	92.00	2213	1024	**1 h 14 min**	**5358**
H. sapiens (chr 1)	Original[1]	**201**	93.53	1,008,692	_	77.52	11,318	98	N/A	N/A
	RACON	**201**	97.01	**3,481,900**	**1,282,763**	**95.69**	**2393**	**57**	2 h 30 min	16,202
	CONSENT	**201**	**97.51**	3,295,244	924,899	94.16	4727	65	**31 min**	**5399**

Table 4. Statistics of the assemblies, before and after polishing with RACON and CONSENT. The missing contig for the CONSENT and RACON polishings on the *D. melanogaster* dataset is 428 bp long, and could not be polished, due to the window size of the two methods being larger (500). Best results for each metric is highlighted in bold. [1]For the assembly of the original reads on the *H. sapiens* (chr 1) dataset, QUAST-LG did not provide a metric for the NGA75.

Corrector	Number of reads	Number of bases (Mbp)	N50 (bp)	Aligned reads (%)	Alignment identity (%)	Genome coverage (%)	Overlapping		Correction		Total	
							Runtime	Memory (MB)	Runtime	Memory (MB)	Runtime	Memory (MB)
Original	**15,243,243**	**112,970**	12,196	80.57	82.74	**93.56**	N/A	N/A	N/A	N/A	N/A	N/A
CONSENT	11,913,704	102,543	**12,880**	**98.44**	**93.86**	93.35	3 days 21 h	98,332	6 days 11 h	45,296	10 days 8 h	98,332

Table 5. Statistics of the full *H. sapiens* dataset, before and after correction with CONSENT. Best results for each metric is highlighted in bold.

outperformed CONSENT in terms of errors per 100 kbp on the *S. cerevisiae* real dataset, but CONSENT outperformed RACON in terms of number of misassemblies NGA50, NGA75, and genome coverage.

For the larger, eukaryotic *D. melanogaster* dataset, RACON outperformed CONSENT in terms of number of errors per 100 kbp, and genome coverage, but the NGA50, NGA75 of the two methods remained comparable. On the *H. sapiens* (chr 1) dataset, RACON once again outperformed CONSENT in terms of error rate and genome coverage, and also displayed larger NGA50 and NGA75. However, polishing the assembly with CONSENT allowed to align a greater proportion of contigs, compared to both the raw and the RACON polished assembly. On these two datasets, RACON slightly outperformed CONSENT in terms of number of misassemblies, although the two methods remained highly comparable. 2Additionally, on all the datasets, the polishing step of CONSENT was 3x to 38x faster than RACON, and also consumed up to four times less memory, thanks to its pre-processing steps.

Results on a full human dataset. To further validate the scalability of CONSENT, we present results on a full ONT human dataset. This dataset is composed of 113 Gbp, displays an error rate of 17%, and contains ultra-long reads reaching lengths up to 1.5 Mbp. Data were sequenced by[2]] and are publicly available with accession number PRJEB23027. Further details are given in Supplementary Table S1.

In this experiment, we not only evaluate how CONSENT behaves on such a large dataset, but also study the impact of the correction / assembly order on the quality of the results. We thus correct the raw data with CONSENT, and then assemble the corrected long reads, but also assemble the raw long reads first, and then polish the assembly with CONSENT. Alignment statistics of the raw and corrected long reads are presented in Table 5, while statistics of the different assemblies are presented in Table 6. We only report CONSENT results, since other tools could not scale to this dataset. Indeed, Daccord crashed upon start due to memory limitations, while FLAS and MECAT reported errors during correction, owing to the presence of ultra-long reads, as reported in previous experiments. Canu, on the other hand, did not crash, but required an unreasonable amount of time to

Assembly	Contigs	Aligned contigs (%)	NGA50 (bp)	NGA75 (bp)	Genome coverage (%)	Errors / 100 kbp	Misassemblies	Runtime	Memory (MB)
Raw	750	95.47	534,347	_	69.83	11,175	3,414	**2 days 20 h**	**382,191**
Corrected	780	**98.08**	2,103,452	_	78.76	5,652	**1,919**	17 days 9 h	1,097,627
Polished	**749**	97.86	**2,964,053**	**232,884**	**83.83**	**4,591**	2,290	7 days 12 h	**382,191**

Table 6. Statistics of the different assemblies for the full *H. sapiens* dataset. Raw corresponds to the assembly generated from raw reads. Corrected corresponds to the assembly generated from corrected reads. Polished corresponds to the assembly generated from raw reads, and polished with CONSENT. Runtime and memory consumption are reported for the whole correction + assembly or assembly + polishing pipelines. QUAST-LG did not provide a metric for the NGA75 of the assembly generated from corrected reads. Best results for each metric is highlighted in bold.

run. We ran all the experiments on a cluster node equipped with 28 2.39 GHz cores and 128 GB of RAM. Tools were thus run with 28 threads.

Alignment statistics of Table 5 show that CONSENT managed to process the whole dataset in 10 days, and required less than 100 GB of RAM. More precisely, the more computationally expensive step, in terms of memory consumption, was actually the overlaps computation, and not the error correction itself. The corrected reads displayed a higher N50 than the raw reads, the longest read reaching 929 kbp. Moreover, almost 99% of the reads could be realigned to the reference genome. The average identity of the alignments reached more than 93.5%, which is slightly higher, but consistent with the results on chr 1, presented in Table 2. Moreover, CONSENT managed to correct a large number of reads, and thus barely reduced the genome coverage of the original dataset.

Assemblies statistics of Table 6 are particularly interesting. Indeed, they show that, in addition to being extremely more computationally expensive, correcting the reads before assembling them produces less satisfying results than assembling the raw reads first, and then polishing the assembly. Indeed, the correction + assembly pipeline required more than 17 days and 1 TB of RAM, while the assembly + polishing pipeline ran in 7.5 days, and consumed less than 400 GB of RAM. In addition, the polished assembly displayed better metrics than the assembly generated from corrected reads, reaching higher NGA50, NGA75, and genome coverage, and lower error rate per 100 kbp. These results underline the fact that, for large datasets and complex genomes, assembling the raw data first, and then polishing the assembly is much more efficient than correcting the reads and then performing assembly.

Discussion

Experimental results on the human datasets are particularly promising. Indeed, they show that CONSENT is the only method able to efficiently scale to the ultra-long reads they contain. More precisely, on the human chr 1 dataset, CONSENT is almost four times faster than Canu, the only other method able to scale to the correction of ultra-long reads. Moreover, it also produces more accurate results, and thus allows to yield a more contiguous assembly. As such reads are expected to become more widely available in the future, being able to deal with them will soon become a necessity. In addition, results on the complete human dataset show that CONSENT manages to efficiently process such large datasets in 10 days, using less than 100 GB of RAM. Moreover, this memory consumption could easily be reduced by adapting the parameters of Minimap2, and reducing the size of the jobs queue used during the actual correction step. At the expense of an increased runtime, CONSENT could thus process a full human dataset on a simple laptop. Using 8 threads, setting Minimap2 parameter `-I 900M`, and reducing the size of the jobs queue to 50,000, we indeed estimate CONSENT would run in a month, consume at most 16 GB of RAM, and require 5 TB of disk space to process a full human dataset. The 5 TB disk space requirement comes from the overlaps file, which can easily be stored on a external hard drive. Further experiments should therefore focus on the correction of larger and more complex organisms. However, the runtime of CONSENT's correction step tends to be higher than that of other state-of-the-art methods. We discuss how to further reduce these computational costs below.

Our experiments show that the runtime of the correction step tends to rise according to the complexity of the genome. This can be explained by the highest proportion of repeated regions in more complex genomes. Such repeated regions indeed impact the alignment piles coverages, and could therefore lead to the processing of piles having very deep coverages. For such piles, our strategy of only selecting the N highest identity overlaps might prove inefficient, especially when the length of the repeated regions grows longer. To further refine the overlaps selection, we could use a validation strategy similar to that of HALC. Such a strategy would allow us to only consider sequences from the pile that actually come from the same genomic region as the long read we are attempting to correct. This would, in turn, allow us to ensure the selected sequences display low divergence, which would speed up the MSA computation, while allowing to produce higher quality consensus.

Moreover, further optimization of the parameters shall also be considered. In particular, the window size and the minimum number of anchors to allow the processing of a window significantly impact the runtime. Running various experiments with different sets of parameters could therefore allow us to find a satisfying compromise between runtime and quality of the results. The fact that the CONSENT assembly covers a smaller proportion of the reference sequence also gives us further room for improvement. In particular, looking to the unaligned contigs more into details could help us further improve the mechanisms and principles of CONSENT. Another possible improvement would be to consider multiple k-mer size for the k-mer chaining strategy. By selecting

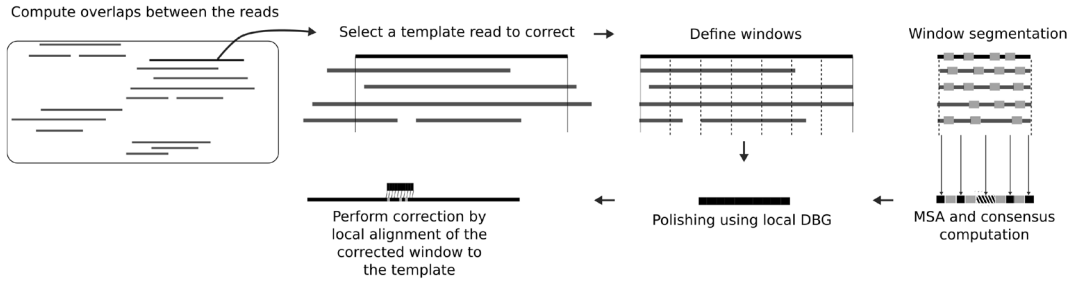

Figure 1. Overview of CONSENT's workflow for long read error correction.

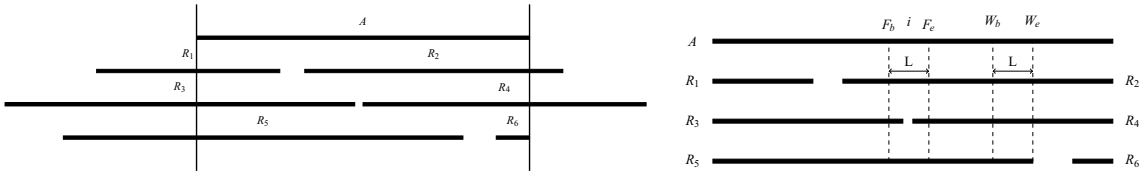

Figure 2. *Left:* An alignment pile for a template read A. The pile is delimited by vertical lines at the extremities of A. Prefixes and suffixes of reads overlapping A outside of the pile are not considered during the next steps, as the data they contain will not be useful for correcting A. *Right:* When fixing the length to L and the minimum coverage threshold to 3, the window (W_b, W_e) will be processed by CONSENT. With these same parameters, the window (F_b, F_e) will not be processed by CONSENT, as $A[i]$ is not supported by at least 3 reads $\forall F_b \leq i \leq F_e$.

the best possible chaining according to the coverage or the repetitive elements of a given window, the method could be more robust and more efficient by computing smaller MSA.

Finally, it is essential to note that CONSENT uses Minimap2 as its default overlapper, but does not depend on this tool. As a result, CONSENT will benefit from the progress of future overlapping strategies, and will therefore allow to propose better correction quality as the overlapping methods evolve.

Material and methods

Overview. CONSENT takes as input a FASTA file of long reads, and returns a set of corrected long reads, reporting corrected bases in uppercase, and uncorrected bases in lowercase. Like most efficient methods, CONSENT starts by computing overlaps between the long reads using a mapping approach. These overlaps are computed using an external program, and not by CONSENT itself. This way, only matched regions need to be further aligned in order to compute consensus. These matched regions are then divided into smaller windows, that are aligned independently. The alignment of these windows is performed via a MSA strategy based on partial order graphs. This MSA is computed by iteratively constructing and adding sequences to a DAG. It also benefits from an efficient heuristic, based on k-mer chaining, allowing to reduce the time footprint of computing MSA between noisy sequences. The DAG is then used to compute the consensus of the window it originates from. Once the consensus is computed, a second correction step makes use of a local de Bruijn graph. This allows to correct weakly supported regions, that are, regions containing weak k-mers, and thus reduce the final error rate of the consensus. Finally, the consensus is realigned to the read, and correction is performed for each window. CONSENT's workflow is summarized in Fig. 1.

Definitions. Before presenting the CONSENT pipeline, we recall the notions of alignment piles and windows on such piles, as proposed in Daccord, since we rely on them throughout the rest of the paper.

Alignment piles. An alignment pile represents a set of reads that overlap with a given read A. More formally, it can be defined as follows. For any given read A, an alignment pile for A is a set of alignment tuples $(A, R, Ab, Ae, Rb, Re, S)$ where R is a long read id, Ab and Ae represent respectively the start and the end positions of the alignment on A, Rb and Re represent respectively the start and the end positions of the alignment on R, and S indicates whether R aligns forward ($S = 0$) or reverse-complement ($S = 1$) to A. One can remark that this definition is slightly different from that of Daccord. In particular, Daccord adds an edit script to each tuple, representing the sequence of edit operations needed to transform $A[Ab..Ae]$ into $R[Rb..Re]$ if $S = 0$, or into $\overline{R}[Rb..Re]$ if $S = 1$ (where $\overline{R}$ represents the reverse-complement of read R). This edit script can easily be retrieved by Daccord, as it relies on DALIGNER[31] to compute actual alignments between the reads. However, as CONSENT relies on a mapping strategy, it does not have access to such information. We thus chose to exclude the edit script from our definition of a tuple. In its alignment pile, we call A the *template* read. The alignment pile of a given template read A thus contains all the information needed for its correction. An example of an alignment pile is given in Fig. 2 (left).

Windows on alignment piles. In addition to the notion of alignment piles, Daccord also underlined the interest of processing windows from these piles instead of processing them as a whole. A window from an alignment pile is defined as follows. Given an alignment pile for a template read A, a window of this pile is a couple (W_b, W_e), where W_b and W_e represent respectively the start and the end positions of the window on A, and are such as $0 \leq W_b \leq W_e < |A|$ (*i.e.* the start and end positions of the window define a factor of the template read A). We refer to this factor as the *window's template*. Additionally, in CONSENT, only windows having the two following properties are processed for correction:

- $W_e - W_b + 1 = L$ (*i.e.* windows have a fixed size);
- $\forall i, W_b \leq i \leq W_e, A[i]$ is supported by at least C reads of the pile, excluding A (*i.e.* windows have a minimum coverage threshold).

This second property allows to ensure that CONSENT has sufficient evidence to compute a reliable consensus for each window it processes. Examples of windows CONSENT does and does not process are given in Fig. 2 (right).

In the case of Daccord, this window strategy allows to build local de Bruijn graphs with small values of k, and overcome the high error rates of the long reads, which cause issues when using large values of k[32]. More generally, processing windows instead of whole alignment piles allows to divide the correction problem into smaller subproblems that can be solved faster. Specifically, in our case, as we seek to correct long reads by computing MSA, working with windows allows to save both time and memory, since the sequences that need to be aligned are significantly shorter.

Overlapping. To avoid prohibitive computation time and memory consuming full alignments, CONSENT starts by overlapping the long reads using a mapping approach. By default, this step is performed with the help of Minimap2[33]. However, CONSENT is not dependent on Minimap2, and the user can compute the overlaps with any other method, as long as the overlaps file follows the PAF format. We included Minimap2 as the default overlapper for CONSENT, since it offers good performances, and is thus able to scale to large organisms on reasonable setups.

Alignment piles and windows computation. The alignment piles are computed by parsing the PAF file generated by the overlapper during the previous step. Each line indeed contains all the necessary information to define a tuple from an alignment pile. It includes the identifiers of the two long reads, the start and the end positions of their overlap, as well as the orientation of the second read relatively to the first. Moreover, for each alignment pile, CONSENT only includes the N highest identity overlaps ($N = 150$ by default, although it can be user-specified), in order to reduce the time footprint, and avoid computing costly MSA of numerous sequences.

Given an alignment pile for a read A, we can then compute its set of windows. To this aim, we use an array of length $|A|$, which counts how many times each nucleotide of A is supported. We initialize the array with 0s at each position, and for each tuple $(A, R, Ab, Ae, Rb, Re, S)$, we increment values at positions i such as $Ab \leq i \leq Ae$. After processing all the tuples, we retrieve the positions of the piles by finding, in the array, sketches of length L of values $\geq C$. We search for such sketches because CONSENT only processes windows of fixed length and with a minimum coverage threshold. In practice, we extract overlapping windows instead of partitioning the pile into a set of non-overlapping windows. Indeed, since it is usually harder to exploit alignments located on sequences extremities, consensus sequence might be missing at the extremities of some windows. Such events would thus cause a lack of correction on the reads, and using overlapping windows allows to overcome the issue. Each window is then processed independently during the next steps. Moreover, the reads are loaded into memory to support random access and thus accelerate the correction process. Each base is encoded using 2 bits in order to reduce memory usage. The memory consumption is thus roughly 1/4 of the total size of the reads.

Window consensus. We process each window in two distinct steps. First, we align the sequences from the window using a MSA strategy based on partial order graphs, in order to compute consensus. This MSA strategy benefits from an efficient heuristic, based on k-mer chaining, allowing to decompose the global problem into smaller instances, thus reducing both time and memory consumption. Second, after computing the window's consensus, we further polish it with the help of a local de Bruijn graph, at the scale of the window, in order to get rid of the few errors that might remain.

Consensus computation. In order to compute the consensus of a window, CONSENT uses POAv2[24], an implementation of a MSA strategy based on partial order graphs. These DAGs, store all the information of the MSA. This way, at each step (*i.e.* at each alignment of a new sequence), the graph contains the current MSA result. To add a new sequence to the MSA, the sequence is aligned to the DAG, using a generalization of the Smith-Waterman algorithm.

Other methods usually compute 1V1 alignments between the read to be corrected and other reads overlapping with it, and then build a result DAG to summarize the alignments, and represent the MSA. In contrast, CONSENT's strategy allows to compute actual MSA, and to directly build the DAG, during the alignment computation. Indeed, the DAG is first initialized with the sequence of the window's template, and is then iteratively enriched by aligning the other sequences from the window, until it becomes the final, result graph. We then extract a matrix, representing the MSA, from the graph, and compute consensus by performing a majority vote. When a tie occurs, we chose the nucleotide from the window's template as the consensus base.

However, even on small windows, computing MSA on hundreds of bases from dozens of sequences is computationally expensive, especially when the divergence among sequences is high. To avoid the burden of building a consensus by computing full MSA, we search for collinear regions shared by these sequences, in order to split the global task into smaller instances. We thus build several consensus on regions delimited by anchors shared among the sequences, and reconstruct the global consensus from the distinct, smaller consensus sequences obtained. The rationale is to benefit from the knowledge that all the sequences come from the same genomic area. This way, on the one hand, we can compute MSA of shorter sequences, which greatly reduces the computational costs. On the other hand, we only use related sequences to build the consensus, and therefore exclude spurious sequences. This behavior allows a massive speedup along with an improvement in the global consensus quality.

To find such collinear regions, we first select k-mers that are non-repeated in their respective sequences, and shared by multiple sequences. We then rely on dynamic programming to compute the longest anchors chain $a_1, \ldots, a_n$ such as:

1. $\forall i, j$ such that $1 \leq i < j \leq n$, a_i appears before a_j in every sequence containing a_i and a_j;
2. $\forall i, 1 \leq i < n$, there are at least T reads containing a_i and a_{i+1} (with T a solidity threshold equal to 8 by default).

We therefore compute multiple, local consensus, using substrings bordered by consecutive anchors, in sequences that contain them, and are then able to reconstruct the global consensus of the window: $consensus(prefix) + a_1 + consensus(]a_1, a_2[) + a_2 + \cdots + consensus(]a_{n-1}, a_n[) + a_n + consensus(suffix)$. We illustrate this segmentation strategy in Supplementary Fig. S1 (longest anchors chain computation) and S2 (local consensus computation and global consensus reconstruction).

Consensus polishing. After processing a given window, a few erroneous bases might remain on the computed consensus. This might happen in cases where the coverage depth of the window is relatively low, and thus cannot yield a high-quality consensus. Consequently, we propose an additional, second correction phase, that aims at polishing the consensus obtained during the previous step. This allows CONSENT to further enhance its quality, by correcting weakly supported k-mers. This feature is related to Daccord's local de Bruijn graph correction strategy.

First, a local de Bruijn graph is built from the window's sequences, using only small, solid, k-mers. The rationale is that small k-mers allows CONSENT to overcome the classical issues encountered due to the high error rate of the long reads, when using large k values. CONSENT then searches for regions only composed of weak k-mers, flanked by sketches of n (usually, $n = 3$) solid k-mers. Afterwards, CONSENT attempts to find a path allowing to link a solid k-mer from the left flanking region to a solid k-mer from the right flanking region. We call these solid k-mers *anchors*. The graph is thus traversed, in order to find a path between two anchors, using backtracking if necessary. If a path between two anchors is found, the region containing the weak k-mers is replaced by the sequence dictated by this path. If none of the anchors pairs can be linked, the region is left unpolished. To polish sketches of weak k-mers located at the left (respectively right) extremity of the consensus, highest weighted edges of the graph are followed, until the length of the path reaches the length of the region to polish, or no edge can be followed out of the current node.

Read correction via window consensus alignment. Once the consensus of a window has been computed and polished, we need to realign it to the template, in order to actually perform correction. To this aim, we use an optimized library of the Smith-Waterman algorithm[34]. To avoid time-costly alignment, we locally align the consensus around the positions of the window it originates from. This way, given a window (W_b, W_e) of the alignment pile of the read A, its consensus will be aligned to $A[W_b - O..W_e + O]$, where O represents the length of the overlap between consecutive windows processed by CONSENT ($O = 50$ by default, although it can be user-specified). Aligning the consensus outside of the original window's extremities as such allows to take into account the error profile of the long reads. Indeed, as insertions and deletions are predominant in long reads, it is likely that a consensus could be longer than the window it originates from, thus spanning outside of this window's extremities.

In the case alignment positions of the consensus from the ith window overlap with alignment positions of the consensus from the $(i + 1)$th window, we compute the overlapping sequences of the two consensus. The one containing the largest number of solid k-mers (where the k-mer frequencies of each sequence are computed from the window their consensus originate from) is chosen and kept as the correction. In the case of a tie, we arbitrarily chose the sequence from the $(i + 1)$th consensus as the correction. We then correct the aligned factor of the long read by replacing it with the aligned factor of the consensus.

Conclusion

We presented CONSENT, a new self-correction method for long reads that combines different efficient strategies from the state-of-the-art. CONSENT starts by computing overlaps between the long reads to correct. It then divides the overlapping regions into smaller windows, in order to compute MSA, and consensus sequences of each window independently. These MSA are performed using a method based on partial order graphs, allowing to perform actual MSA. This method is combined to an efficient k-mer chaining strategy, which allows to further divide the MSA into smaller instances, and thus significantly reduce computation times. After computing the consensus of a given window, it is further polished with the help of a local de Bruijn graph, at the scale of the

window, in order to further reduce the final error rate. Finally, the polished consensus is locally realigned to the read, in order to correct it.

Our experiments show that CONSENT compares well to, or even outperforms, other state-of-the-art self-correction methods in terms of quality of the results. In particular, CONSENT is the only method able to efficiently scale to the correction of ONT ultra-long reads, and is able to process a full human dataset containing reads reaching lengths up to 1.5 Mbp in 10 days. Although very recent, such reads are expected to further develop, and thus become more widely available in the near future. Being able to deal with them will thus soon become a necessity. CONSENT could therefore be the first self-correction method able to be applied to such ultra-long reads on a greater scale.

CONSENT's assembly polishing feature also offers promising results. In particular, our experiment on a full human dataset shows that assembling the raw reads and then polishing the assembly allows to greatly reduce the computational costs, but also provides better results than correcting and then assembling the reads. This conclusion raises the question of the interest of long-read error correction in assembly projects. Moreover, as the processes of long read correction and assembly polishing are not much different from one another, one can also wonder why more error correction tools do not offer such a feature. It indeed seems to be affordable at the expense of minimal additional work, while providing satisfying results. We believe that CONSENT could open the doors to more error correction tools offering such a feature in the future. Finally, it would also be interesting to evaluate already published correction tools on their ability to polish assemblies, at the expense of minimal modifications to their workflows.

The segmentation strategy introduced in CONSENT also shows that actual MSA techniques are applicable to long, noisy sequences. In addition to being useful for error correction, this could also be applied to various other problems. For instance, it could be used during the consensus steps of assembly tools, for haplotyping, and for quantification problems. The literature about MSA is vast, but lacks application on noisy sequences. We believe that CONSENT could be a first work in that direction.

Data availability

CONSENT is implemented in C++, wrapped in Python and Bash scripts, open source, supported on Linux platforms and available at https://github.com/morispi/CONSENT.

Received: 26 August 2020; Accepted: 22 December 2020
Published online: 12 January 2021

References

1. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **39**, 329–346 (2018).
2. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338 (2018).
3. Patterson, M. *et al.* Whatshap: Weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
4. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. & David, N. T. Hinge: long-read assembly achieves optimal repeat resolution. *Genome Res.* **27**, 747–756 (2017).
5. Cao, M. D. *et al.* Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* **8**, 14515 (2017).
6. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
7. Haghshenas, E., Hach, F., Sahinalp, S. C. & Chauve, C. CoLoRMap: Correcting long reads by mapping short reads. *Bioinformatics* **32**, i545–i551 (2016).
8. Choudhury, O., Chakrabarty, A. & Emrich, S. J. HECIL: A hybrid error correction algorithm for long reads with iterative learning. *Sci. Rep.* **8**, 1–9 (2018).
9. Salmela, L. & Rivals, E. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics* **30**, 3506–3514 (2014).
10. Miclotte, G. *et al.* Jabba: hybrid error correction for long sequencing reads. *Algorithms Mol. Biol.* **11**, 10 (2016).
11. Wang, J. R., Holt, J., McMillan, L. & Jones, C. D. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinform.* **19**, 1–11 (2018).
12. Kchouk, M. & Elloumi, M. An error correction and DeNovo assembly approach for nanopore reads using short reads. *Curr. Bioinform.* **13**, 241–252 (2018).
13. Bao, E. & Lan, L. HALC: High throughput algorithm for long read error correction. *BMC Bioinform.* **18**, 204 (2017).
14. Firtina, C., Bar-joseph, Z., Alkan, C. & Cicek, A. E. Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Res.* **46**, e125 (2018).
15. Madoui, M.-A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16**, 327 (2015).
16. Morisse, P., Lecroq, T. & Lefebvre, A. Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. *Bioinformatics* **34**, 4213–4222 (2018).
17. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
18. Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101 (2013).
19. Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
20. Xiao, C. L. *et al.* MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
21. Bao, E., Xie, F., Song, C. & Dandan, S. HALS: Fast and high throughput algorithm for PacBio long read self-correction. *RECOMB-SEQ* **35**, 3953–3960 (2019).
22. Tischler, G. & Myers, E. W. Non hybrid long read consensus using local de Bruijn graph assembly. *bioRxiv* (2017).
23. Salmela, L., Walve, R., Rivals, E. & Ukkonen, E. Accurate selfcorrection of errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799–806 (2017).
24. Lee, C., Grasso, C. & Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464 (2002).
25. Stöcker, B. K., Köster, J. & Rahmann, S. SimLoRD: Simulation of long read data. *Bioinformatics* **32**, 2704–2706 (2016).

26. Marchet, C. *et al.* ELECTOR: evaluator for long reads correction methods. *NAR Genom. Bioinform.* **2**, lqz015 (2019).
27. Li, H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *arXiv* **25**, 1–7 (2015).
28. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
29. Mikheenko, A., Prjibelski, A., Antipov, D., Saveliev, V. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
30. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 727–736 (2017).
31. Myers, G. Efficient local alignment discovery amongst noisy long reads. In *Algorithms in Bioinformatics* (eds Brown, D. & Morgenstern, B.) 52–67 (Springer, Berlin, Heidelberg, 2014).
32. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform.* **13**, 238 (2012).
33. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
34. Zhao, M., Lee, W. P., Garrison, E. P. & Marth, G. T. SSW library: An SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS ONE* **8**, 1–7 (2013).

Acknowledgements

Part of this work was performed using computing resources of CRIANN (Normandy, France), project 2017020. The authors would like to thank Pierre Marijon for his help with the Bioconda release. P.M. would also like to thank Pierre Marijon for his helpful advice with assemblers parameters.

Author contributions

P.M., C.M. and A.Li. designed and carried out research and wrote the manuscript. A.Le. and T.L. carried out research. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-80757-5.

Correspondence and requests for materials should be addressed to P.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.