OXFORD

Briefings in Bioinformatics, 2024, **25(3)**, bbae194
https://doi.org/10.1093/bib/bbae194
Problem Solving Protocol

DEGAP: Dynamic elongation of a genome assembly path

Yicheng Huang[1], Ziyuan Wang[2], Monica A. Schmidt[3], Handong Su[1,4], Lizhong Xiong[1], Jianwei Zhang[1,*]

[1]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Wuhan 430070, China
[2]Department of Pharmacy Practice & Science, College of Pharmacy, University of Arizona, Tucson, AZ 85721, USA
[3]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[4]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518000, China

*Corresponding author: Jianwei Zhang, National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, No. 1 Shizishan Street, Hongshan District, Wuhan 430070, China. Tel.: +86-27-8728-6166; E-mail: jzhang@mail.hzau.edu.cn

Abstract

Genome assembly remains to be a major task in genomic research. Despite the development over the past decades of different assembly software programs and algorithms, it is still a great challenge to assemble a complete genome without any gaps. With the latest DNA circular consensus sequencing (CCS) technology, several assembly programs can now build a genome from raw sequencing data to contigs; however, some complex sequence regions remain as unresolved gaps. Here, we present a novel gap-filling software, DEGAP (Dynamic Elongation of a Genome Assembly Path), that resolves gap regions by utilizing the dual advantages of accuracy and length of high-fidelity (HiFi) reads. DEGAP identifies differences between reads and provides 'GapFiller' or 'CtgLinker' modes to eliminate or shorten gaps in genomes. DEGAP adopts an iterative elongation strategy that automatically and dynamically adjusts parameters according to three complexity factors affecting the genome to determine the optimal extension path. DEGAP has already been successfully applied to decipher complex genomic regions in several projects and may be widely employed to generate more gap-free genomes.

Keywords: *gap closure; genome assembler; HiFi reads*

INTRODUCTION

DNA sequencing has had a long and inspiring history of development. Each generation of sequencing technologies has propelled the decryption of the genetic code and aided in the understanding of the biology and evolution of species. Complete and accurate genome assembly is an essential component for genomic analyses. Presently, genome assemblies are commonly accomplished by combining inaccurate long sequencing reads with accurate short reads [1–3]. Pacific Biosciences (PacBio) currently provides high-fidelity (HiFi) sequencing technology to enhance both the accuracy and length of sequence reads, routinely producing reads exceeding 10 kbp in length with 99.9% accuracy [4].

With the development of enhanced sequencing technologies, several software tools (e.g. Canu [5], MECAT [6], FALCON [7], Flye [8], hifiasm [9] and Wtdbg [10]) have become available for *de novo* assembling of genomes by using various raw sequencing data. However, gaps still remain in many assemblies due to either low sequencing depth or high sequence complexity. The processes must account for the individual differences among various genomes to achieve a high-quality assembly. While most genome assemblies can be handled by general whole-genome assemblers, intricate regions may necessitate additional efforts, such as manual editing of long tandem repeats [11]. Previous studies have shown that higher occurrences of repeats often lead to gaps in the assembly [12]. Commonly used assemblers, such as Canu and FALCON, typically group similar reads to rectify sequencing errors by utilizing the consensus of the majority of reads and produce contigs through sequence overlap. The presence of similar surrounding bases makes distinguishing certain variants from sequencing errors challenging, particularly in highly repetitive regions like the centromere. This can lead to misassembled contigs and result in gaps within the genomes. Hence, a high-resolution tool capable of easily distinguishing

Yicheng Huang is a graduate student at Huazhong Agricultural University. She specializes in the development of methods and pipelines for genomic studies.
Ziyuan Wang is a graduate student at the University of Arizona. He specializes in advancing methods for single-cell RNA and Nanopore sequencing technologies, contributing to cutting-edge research in genomics and transcriptomics.
Monica A. Schmidt received a PhD in Genetics from the University of British Columbia then after a few different post-doctoral positions she became a faculty member at the University of Arizona.
Handong Su received his PhD from Institute of Genetics and Developmental Biology, Chinese Academy of Sciences in 2019. He is a professor at Huazhong Agricultural University. His main research interest is to study the mechanisms of distant hybridization and polyploidization in plants.
Lizhong Xiong is a professor at Huazhong Agricultural University. His research interests focus on functional genomics and abiotic stress biology of rice.
Jianwei Zhang is a professor at Huazhong Agricultural University. His research interests focus on genomics and bioinformatics with high-throughput sequencing, computer science and other technical means.
Received: October 25, 2023. **Revised:** March 25, 2024. **Accepted:** April 11, 2024

low-frequency variants and accurately closing gaps is a necessary requirement to assemble complete genome sequences.

In this study, we developed an automated fine-tuning tool, named DEGAP (Dynamic Elongation of a Genome Assembly Path), to enhance the contiguity of genome assembly. DEGAP employs dynamic strategies to select HiFi reads and elongate gap edge sequences, thereby generating continuous assembly paths for closing or shortening gap regions. DEGAP can efficiently resolve gaps in chromosome-level assemblies and non-chromosome-level assemblies without the need to generate more sequencing data and cost-effectively reduce the overall number of gaps in an entire genome. DEGAP's ability to resolve gap regions will undoubtedly make it widely useful in the generation of more gap-free genomes.

RESULTS

Two gap-closure modes: GapFiller and CtgLinker

DEGAP provides two operational modes, GapFiller and CtgLinker, each tailored to distinct different scenarios: chromosome-level and non-chromosome-level genome assemblies. The HiFi reads provided for DEGAP remain consistent in both modes; however, these modes require different assembly sequences as additional input. In the GapFiller mode, the input sequences consist of two flanking sequences from both gap edges, whereas in CtgLinker mode, contig or scaffold sequences are used as input. The GapFiller mode focuses on filling known gaps, specifically in chromosome-level assemblies. DEGAP takes both left and right sequences of gap regions as seed sequences and uses raw HiFi reads to incrementally elongate the seed sequences to eventually reach the sequences on the other side of each gap. The CtgLinker mode is designed for handling non-chromosome-level assemblies with unknown gaps. In this mode, DEGAP first filters out the overly short contigs, then cuts off the edges of the contigs to reduce the mis-assembly caused by sequencing errors and pursue accurate elongation results. DEGAP elongates all contigs with supplied HiFi data and generates assembly paths for connecting well-matched contigs through overlapping sequences. The automated process persistently conducts elongation tasks until all gaps are filled, or until no extension sequences (or reads) are found while gaps remain.

Dynamic process of assembly path elongation

DEGAP adopts an iterative elongation strategy that automatically uses dynamic parameter thresholds to select sequences for extension according to the complexity of the actual situation, which can be divided into four steps: (i) extracting reads that are located on one side of the gap, (ii) processing the candidate extension reads or sequences, (iii) elongating the edge sequence using the optimal candidate in the previous step and (iv) determining if the elongated sequence reaches the other side of the gap (Figure 1A). In the fourth step, DEGAP checks if the elongated sequence has reached the other side of the gap to assess if the elongation process is complete (the gap was filled) or needs to continue using the newly elongated sequence as the input sequence for the next elongation round (Figure 1B).

In each round of elongation, the basic principle for DEGAP is to select the correct reads or extension sequence to elongate the edge sequences between gaps. The accuracy depends on the assessment of whether an extension read/sequence is eligible for elongation, and this depends on the alignments to an edge sequence. Considering the differences in the input edge sequence for each round, DEGAP will dynamically change the threshold values of four different parameters to select the alignment blocks: (i) the distance from the alignment area to the sequence boundary, (ii) alignment identity, (iii) alignment length and (iv) extension length of the reads (Figure 1c, Supplementary Figures S1 and S2).

During each round of elongation, extension reads or contigs are dynamically selected based on assembly error and qualified alignment results. Using minimap2 [13] to map all HiFi reads back to one edge of a gap, the dynamic read finding process in DEGAP finds a set of candidate extension reads that have a high sequence identity and high coverage alignment with the edge sequence for potentially extending the edge sequence (Supplementary Figure S1). Subsequently, the dynamic extension sequence finding process may *de novo* assemble the extension reads obtained in the previous step by using hifiasm [9] or find the common sequences between extension HiFi reads, and then select the best extension sequence to precisely elongate the edge sequence (Supplementary Figure S2).

DEGAP determines qualified alignment results by considering both maximum alignment identity and maximum alignment length against an edge sequence (Supplementary Figure S3). Meanwhile, DEGAP can also provide a fuzzy alignment in a second dynamic process when no qualified alignment results are found (i.e. repeat-caused multiple alignment results) (Supplementary Figure S3D and E).

By taking the edge sequence at one end of a gap as a seed, DEGAP generates a continuous sequence path through the iterative extension process. Therefore, there are three possible outcomes of the DEGAP process: (i) the gap is filled; (ii) no extension reads/sequences are found and (iii) no new extension reads are found, and the process is stuck in an infinite loop. The log file generated by DEGAP comprehensively documents the extension process, detailing mapping quality, alignments and the utilized reads. This design allows users to further examine the accuracy in gap filling (Supplementary Figure S4).

Two study cases of DEGAP

Rice centromere sequences resolved in the GapFiller mode

Centromeric DNA usually contains highly repetitive DNAs, such as satellite DNAs, which are among the most difficult to resolve in genome assemblies. In this study, we used the gap-free genome (MH63RS3) of rice Minghui 63 (MH63) assembled with multiple sequencing data as a reference, as it contains complete centromere sequences [14], including *CentO* satellite repeats [15]. We used the flanking sequences bordering the centromere as input for elongating all centromere sequences. By using DEGAP, we bidirectionally reproduced all centromere sequences in MH63RS3 and compared them with the corresponding sequences in MH63RS3 to validate the DEGAP results (Figure 2). The results were that DEGAP successfully filled the complete centromere sequences in Chr03, Chr06 and Chr10 from both directions, and Chr07 and Chr09 from the left side and Chr12 from the right side of the centromere (Figure 2A).

Human genome sequences improved with the CtgLinker mode

The first near-complete human reference genome (GRCh38.p13) was released in 2013 [16] by the Genome Reference Consortium. This genome assembly contained 349 unfinished repetitive and polymorphic regions as gaps. A Telomere-to-Telomere (T2T) version (CHM13) of the reference human genome was assembled in 2022 by using PacBio HiFi and Oxford Nanopore ultralong-read sequencing technologies [1]. The T2T-CHM13 reference genome (CHM13v1.1) was used to test the DEGAP performance,

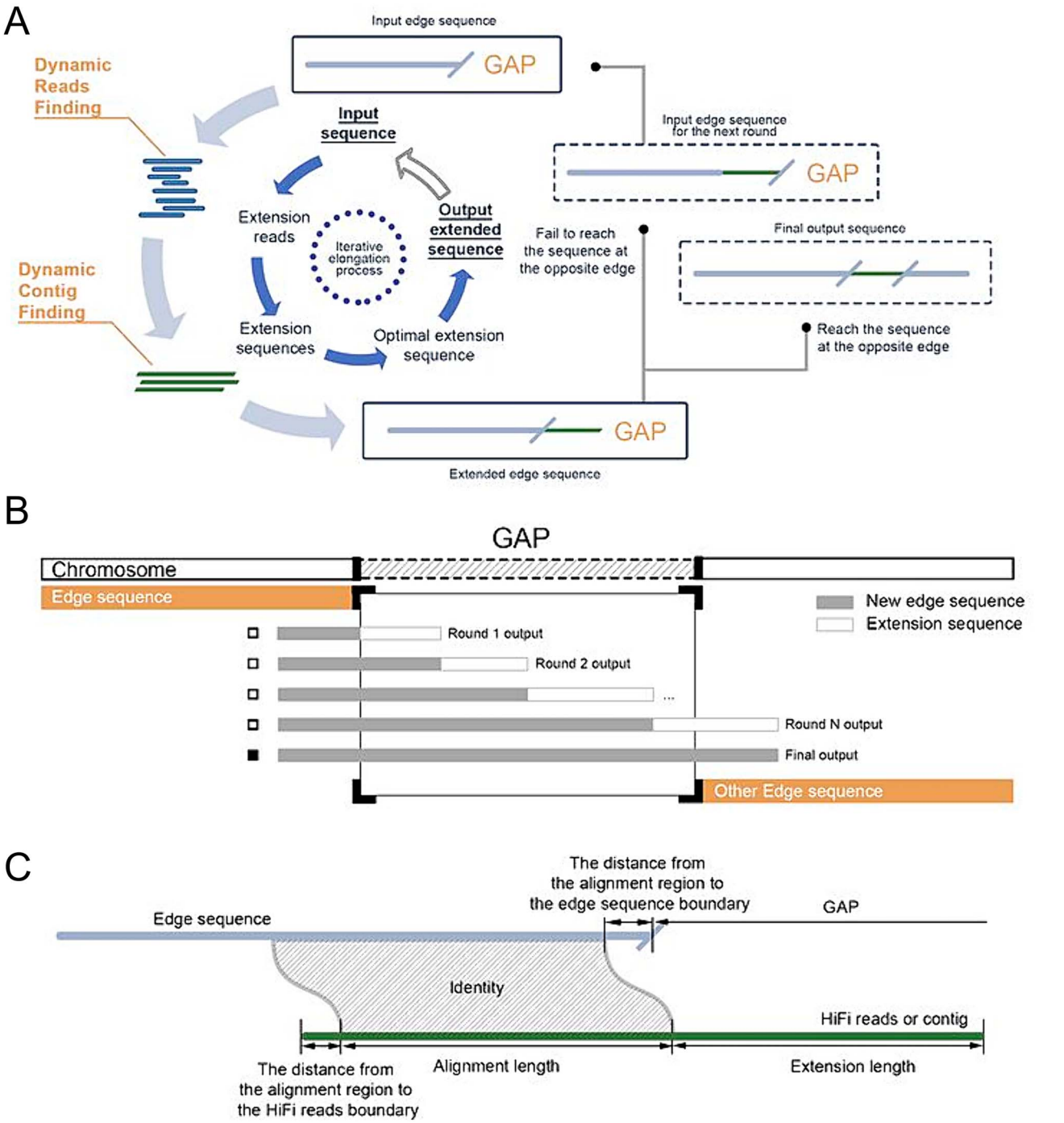

Figure 1. DEGAP pipeline. (**A**) Iterative elongation strategy of DEGAP. (**B**) Complete elongation process when DEGAP filled the gaps. (**C**) Parameters to evaluate whether HiFi reads/sequences are eligible to extend the edge sequence.

and the former non-T2T assembly (CHM13v0.7) was used as the input dataset to test if DEGAP could resolve a non-chromosome-level assembly. As an illustrative example, we used the Chr19 sequence (NCBI AC No. WNKB01000034.1) from the previous version of the human reference genome (CHM13v0.7) as the input (Supplementary Figure S5) and endeavored to bridge gaps with HiFi reads. Using DEGAP CtgLinker mode, the input sequences were automatically split into smaller contigs containing no gaps, then elongated by extending the edge sequences. As a result, two contiguous sequences were generated from nine fragmented contigs, and the gap number was reduced from eight to one (Supplementary Figure S5). This comparison result validates the high collinearity between DEGAP-resolved sequences and the gap-free CHM13v1.1 reference genome (Supplementary Figure S5). In addition to Chr19, DEGAP demonstrates its capability across other chromosomes. For instance, it effectively closed three gaps in Chr01, two in Chr02 and two in Chr03 (Supplementary Figure S5B–D).

Three complexity factors affecting the genome assembly path

For a genome assembly, it is easy to measure its quality from the sequence length (such as N50), but rarely from the data complexity. Here, in consideration of multiple variable factors,

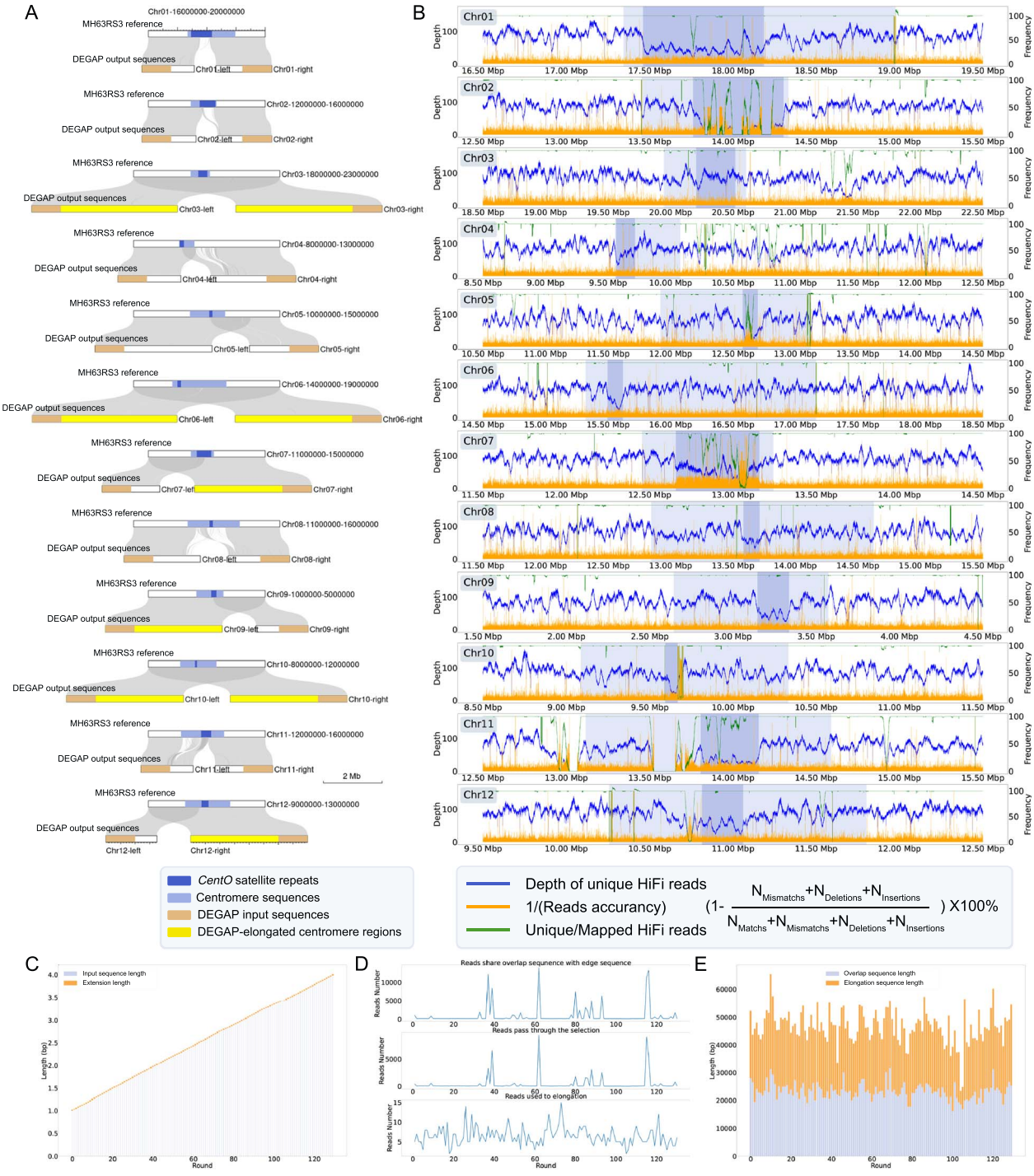

Figure 2. DEGAP-resolved results of MH63 centromere sequences in GapFiller mode. (**A**) Alignment of DEGAP results with the MH63 reference genome. Chr03, Chr06, Chr07, Chr09, Chr10 and Chr12 of MH63 had complete centromeric sequences elongated by DEGAP in GapFiller mode. (**B**) Sequences complexity in the MH63RS3 centromere sequences. The closer to the centromere region, the lower the depth of unique HiFi reads, the higher the inaccuracy of reads and the more difficult it is for DEGAP to obtain correct reads. The definition that we choose is the proportion of base mismatch, deletion and insertion in the matching columns in the alignment of the reads to the reference sequence. The ratio of unique HiFi reads among all mapped HiFi reads in the same area is used to assess the difficulty for DEGAP to obtain correct reads. (**C**) The input sequences length and extension length in each round of Chr03 left side DEGAP elongation. (**D**) The number of reads that matched edge sequences, reads passed through selection and reads used to elongation in each round of Chr03 left side DEGAP elongation. (**E**) The length of overlap and elongated sequences in elongation sequences in each round of Chr03 left side of DEGAP elongation.

we adopt three major characteristics of a sequence dataset for evaluating sequence complexity and ease of assembly from a global perspective for the entire genome: the depth value of unique reads, reads accuracy and the ratio of unique reads among all mapped reads (Figure 2).

The depth value of unique reads, a basic characteristic, indicates the actual number of sequencing reads in a certain region, and it is also the expected value of extended reads to be obtained in this region. This characteristic can be regarded as the true number of reads obtained at this location during the genome sequencing process, which is the optimally aligned reads in the entire genome alignment. In the case study of rebuilding MH63 centromere sequences, it was found that the closer to the centromere and the *CentO* regions, the fewer number of unique reads, which translates to a smaller number of extension reads generated by DEGAP. Results indicate that the closer to a centromere region, the lower depth value of unique reads (Supplementary Figure S6, Supplementary Table S1).

The reads accuracy, a quality characteristic, represented by the subtraction ratio of mismatches, deletions and insertions when mapping all reads back to the reference genome, is another data characteristic to indicate the assembly potential. The closer to the centromere region, the lower the reads accuracy, which elevates the difficulty to build consensus sequences. The results showed significant differences between non-centromere regions and centromere regions (non-*CentO* regions) and between non-centromere regions and *CentO* regions in all chromosomes. For example, the Chr02 *CentO* region was determined to have Cohen's $d > 0.8$, and the Chr11 centromere region (non-*CentO* region) had Cohen's $d > 0.5$ when compared to non-centromere regions, and also these gaps were not closed by DEGAP (Figure 2A, Supplementary Figure S6, Supplementary Table S1). These results indicate that the closer to centromere regions, the lower the reads accuracy.

The ratio of unique reads among all mapped reads, a differentiation characteristic, indicates how unique the locus is across the genome. As centromeric regions contain a large number of satellite repeats with high sequence similarity across chromosomes, the proportion of unique reads represents the level of difficulty for DEGAP to properly distinguish corresponding reads belonging to different chromosomes. The results showed that there are significant differences between non-centromere regions and centromere regions (non-*CentO* regions) and between non-centromere regions and *CentO* regions in all chromosomes. For example, Chr02, Chr05 and Chr07 *CentO* regions have Cohen's $d > 0.8$, and Chr11 centromere regions (non-*CentO* regions) have Cohen's $d > 0.5$ when compared to non-centromere regions (Supplementary Figure S6, Supplementary Table S1). These results indicate that the closer to a centromere region, the lower the ratio of unique reads.

Quality assessment with three complexity factors in genome assembly

We used the Guassian mixture model (GMM), or skewed distribution, to evaluate the overall quality of the rice MH63RS3, human CHM13v1.1 and human CHM13v2.0 assemblies by classifying genomic locus form into three groups using the above three characteristics (Supplementary Figure S7). Three groups from high to low represent the difficulty of genome assembly. For the rice genome analysis, the results showed that Group I is most prevalent in non-centromere regions, Group II is most prevalent in centromere regions (non-*CentO* regions) and Group III is most abundant in *CentO* regions. These results are consistent with the finding that the closer to centromere regions, the more difficult the regions are to assemble (Supplementary Figure S7).

The result shows that 99.78% of MH63RS3 assembly have been properly assembed and 94.73% in CHM13v1.1, which indicates that the MH63 genome was easier to assemble. Moreover, CHM13v2.0 has more sites in Group III compared to CHM13v1.1, which indicates CHM13v2.0 assembly resolved more complex regions when compared to CHM13v1.1 (Supplementary Table S2).

The skewed distribution was used to group MH63RS3 instead of GMM, due to Group I and Group II exhibiting a long-tail class distribution (Supplementary Table S2). We also used the same HiFi reads to *de novo* assemble the MH63 genome by Canu (version 2.0) and hifiasm (version 0.16.1-r375). The GMM was subsequently used to combine these two assemblies. The result was that neither of these two assemblies had better performance than MH63RS3. We found Canu assembled the more difficult parts of the genome better than hifiasm (Group III in Canu is 30.8% and in hifiasm is 9.47%), such as in more repetitive sequence regions, and these regions are also highly prone to assembly errors. Overall, the assembly results using hifiasm gave a better performance when compared to the assembly using Canu. Hifiasm assemblies consistently had higher numbers of Group II characteristics.

Software comparison

To evaluate the performance of DEGAP, we compared it with four other gap-filling software programs: TGS-GapCloser [17], FGAP [18], LR_Gapcloser [19] and PBJelly [20], which can take PacBio long reads as input. The published 20 pseudochromosomes (with 23 gaps) of *Perilla frutescens*, broken down by gaps, were used to compare the various gap-filling programs [21]. The genome was assembled using HiFi reads by hifiasm [9] (v0.16), and it was also combined with processed Omni-C reads for Hi-C integrated assembly. The gaps in nearly completed genomes have complex individual reasons and can serve as a dataset for evaluating the ability of gap-filling tools to improve assembly continuity within the limited existing data without generating additional data. DEGAP, TGS-GapCloser and FGAP filled 10, 8 and 3 gaps, respectively (Supplementary Tables S3 and S4). However, LR_Gapcloser and PBJelly did not fill any gaps in the *P. frutescens* assembly. These results illustrate that DEGAP performs better in filling gaps or extending sequences to build an assembly of greater completeness and higher contiguity than other existing tools.

DISCUSSIONS

As more and more genomes are sequenced and assembled, gap-free genomes are increasingly desired, if not expected. The genome assembly process initiates by obtaining long N50 reads and progresses to computer program operations capable of resolving complex repetitive regions. DEGAP proves to be a valuable new tool to assist in the generation of complete gap-free genomes using PacBio HiFi reads, and it is also a powerful tool to aid in the resolution of complex regions. The biggest strength of DEGAP is that it dynamically adjusts the elongation strategy according to the complexity of the gap region and accurately elongates sequences even by using sequences with low depths.

Accuracy and completeness are the most important requirements for genome assembly. Both metrics are substantially affected by the quality and complexity of the input raw data, especially in complex regions such as those found in centromeres. DEGAP can distinguish between reads having high sequence similarity, which allows it to resolve repetitive and complex regions in genomes. In consideration of the above

three characteristics of sequencing data, we designed DEGAP to dynamically recognize sequences in an elongation process.

In this study, we provide three complexity factors affecting the genome assembly path: depth value of unique reads, reads accuracy and the ratio of unique reads among all mapped reads. The higher the depth value of unique reads, the higher in the potential for a successful elongation; the higher the reads accuracy, the more precise assembly path for elongation; the higher the ratio of unique reads, the stronger the distinction of elongation assembly paths from similar sequences in other regions of the genome. Although all three characteristics of the sequence dataset observed in our study indicate the difficulty of assembling centromere sequences, in the MH63RS3 example, DEGAP was able to successfully rebuild 6 out of 12 centromere sequences with high identity to the reference genome (Figure 2A). For each round in DEGAP, sequences obtained from a previous round are used as input sequences for the seed sequence in the elongation process where, if possible, the edge sequences are extended. After the elongation process is complete for that sequence, DEGAP tests whether the extended output sequence can close the gap (Figure 2C). To exclude the influence of reads from similar sequences and enhance the likelihood of obtaining the correct unique reads for an extension, DEGAP uses two rounds of stringent dynamic selection process to filter out many abnormally mapped reads that may cause assembly errors (Figure 2D). However, this filtering process is not absolute, especially when edge sequences share identical sequences from other regions that exceed the reads length. Such difficulty can only be resolved by using longer sequencing technologies. To increase the extension quality, DEGAP follows an assembly path of higher accuracy rather than longer elongation, so for each process, it preferentially uses similar overlapping sequences in length as input sequences (Figure 2E).

These three characteristics affect not only the DEGAP extension accuracy, but also the correctness of an assembly in all existing assembly software. We used GMM or skewed distribution to group all sites in the assembly into three. Group I represents the easiest part of the genome to assemble, showing the highest values for all three characteristics. Group II represents parts of the genome that are less easy to assemble, showing high values for only one or two of all three characteristics. Group III represents the most difficult part of the genome to assemble, with lowest values in all three characteristics. For its simplicity, EM algorithm performs incredibly well in a range of scenarios. However, there are a number of potential limitations that we need to be aware of. One of the most concerning limitations is that its EM is slow for large datasets, such as an ultra-large genome. Skewed distribution is used to be an alternative method when Groups I and II (often occurs in high-quality assemblies like the gap-free genome) or Groups II and III (extremely high-heterozygosity assemblies in which Group III becomes a major part of the entire genome) exhibit a long-tail class distribution.

The issue of false contig connections arises when multiple highly similar long regions exist within a genome, making it difficult to discern HiFi reads originating from distinct regions. During the elongation processes, we utilize dynamic parameters tailored to specific situations at gap edges rather than employing unified ones. In DEGAP, there are two major dynamic selection processes to elongate the edge of gaps: HiFi reads selection and extension sequence selection. DEGAP dynamically adjusts the filtering thresholds, transitioning from stringent settings to prevent such occurrences and provides the entire process within each elongation. Users can manually check and track the results after running DEGAP if there are any suspicion regions. However, complete avoidance of these false positive connections remains elusive, particularly when assembling sequences from identical repeats shorter than the repeat region. The resolution of this issue hinges on advancements in sequencing technology, enabling the generation of longer reads capable of spanning highly similar regions.

Gaps within various genomes may stem from diverse and often intricate reasons. The primary objective of DEGAP is to enhance assembly contiguity using existing sequencing data without generating additional data. The origins of gaps in different genome assemblies can be multifaceted and may sometimes be resolved through iterative local assembly processes. It's worth noting that the *P. frutescens* genome used for comparison in this study was assembled using HiFi reads with Hifiasm (v0.16), supplemented by processed Omni-C reads for Hi-C [21]. DEGAP filled 10 out of 23 gaps in the published *P. frutescens* genome using their HiFi data, resulting in an improved contiguity of their genome without requiring extra data. Furthermore, we used the CtgLinker mode to address gaps in the hifiasm assembly of MH63, which initially consisted of 826 contigs with an N50 of 30 583 321 bp. As a result, a gap-free genome comprising 12 pseudochromosomes of 400 268 500 bp in length (Supplementary Table S5) was successfully regenerated, exhibiting high collinearity with the MH63RS3 reference sequences (Supplementary Figure S8).

DEGAP gradually elongates gap edges in a sequential manner, meaning a new round of elongation will not start until the previous one concludes. Consequently, filling extremely long gaps with DEGAP is a time-consuming process. It's important to recognize that DEGAP serves as a complementary fine-tuning tool for genome assembly rather than a replacement for comprehensive whole-genome assembly methods. The elongation assembly path generated in both modes in DEGAP is highly dependend on the input sequences. The sequences proved in GapFiller mode are the left and right edge sequences between gaps as input, whereas CtgLinker mode takes scaffolds with no specific order as input. The alignment with opposite edge sequences also serves as a termination signal. If the termination alignment regions with opposite edges are inaccurate, the elongation process will continue until no extension sequences are found, resulting in the retention of both the gaps and elongated input sequences.

The accuracy of the genome is often the key affecting subsequent research, such as gene annotation or genome-wide SNP analysis. For future genome assemblies, we need to shift our focus from pursuing more complete genomes to complete and precise genomes. In addition to *de novo* genome assemblies, DEGAP can also be used to verify the quality and correctness of regions in released genomes by comparing the differences between DEGAP extension results and other assembly results in difficult-to-be-assembled regions. DEGAP, as a new 'post' genome assembly software tool, will play a powerful role in building better-quality genome sequences with longer contiguity.

METHODS

Dynamic HiFi reads finding

DEGAP uses minimap2 [13] to generate the sequence alignment file in SAM/BAM format, sets the distance from the boundary threshold from 10 bp to 500 bp (the default is 500 bp and can be customized with the parameter '–edge'), MAPQ from 20 to 0, alignment length from 3000 to 500 bp and extension length from 1000 to 10 bp. After selecting all aligned HiFi reads and getting the extension HiFi reads in a loose threshold, DEGAP adaptively

adjusts the filtering thresholds from a stringent one for the most suitable HiFi reads (Supplementary Figure S1). To reduce the screening time, DEGAP selects the entire HiFi reads with parameters as NM (edit distance to the reference, including ambiguous bases but excluding clipping)/alignment length <0.1, MAPQ ≥0, alignment length ≥500, extension length ≥10 and distance from the boundary ≤500. If no extension reads are found, DEGAP stops the entire elongation process and exports the result. Alignment files were processed by SAMtools [22] and Pysam Python package.

Dynamic extension sequence finding

DEGAP assembles the extension reads by hifiasm [9] and aligns the assembled sequences to edge sequences by MUMmer [23]. The minimum alignment length may be set from 1000 to 500 bp, the minimum alignment identity from 99.0 to 95.0 and the distance from the boundary from 10 bp to maximum extension length. If there are no sequences assembled by hifiasm that pass the selection process, DEGAP uses the common sequences between any two HiFi reads (DEGAP uses the entire sequence if only one extension read is found). With the common sequences, DEGAP runs the dynamic selection process again to find the best extension sequence. The goal of this process is to find the best extension sequence by selecting the alignment blocks from the hifiasm assembly or common sequences. The result may contain multiple sequences (Supplementary Figure S3A).

Fuzzy alignment generation

DEGAP uses MUMmer to align the assembly or common sequences with the edge sequence to find the best alignment result, which can elongate the edge sequence (Supplementary Figure S3A and B). DEGAP uses the length of the alignment area, alignment identity and distance from the boundary to select the alignment result, which is similar to the extension reads selection. If the alignment is reversed, DEGAP takes the reverse complement of the extension sequence (Supplementary Figure S3B). If there are multiple alignment results, DEGAP chooses the most solid alignment block (Supplementary Figure S3C). Moreover, DEGAP can also consider some special circumstances where hifiasm's assembly cannot elongate the edge sequence. For example, if two sequences are greatly different or contain tandem repeats, the result does not always show a single alignment block at the edge (Supplementary Figure S3D and E). When the results generate more than one alignment block but none of them can pass the dynamic selection, DEGAP generates a fuzzy alignment result to test if the extension sequence can be utilized (Supplementary Figure S3D and E).

Three complexity factors affecting the genome assembly path

t-tests were conducted to compare complex regions, like centromere and euchromatin regions, on different chromosomes. Bonferroni correction was applied to adjust for multiple testing. Cohen's d was determined to indicate whether the effect is large enough to be meaningful in a real application and to estimate practical significance (Cohen's $d > 0.8$). The results showed significant differences between non-centromere regions and centromere regions (non-*CentO* regions) and between non-centromere regions and *CentO* regions in all chromosomes. All *CentO* regions, except on Chr03, were found to have significant differences with non-centromere regions in the depth value of unique HiFi reads, as indicated by Cohen's $d > 0.8$.

ASSEMBLY EVALUATION

To evaluate the performance of the assembly data, a quantified method was used. For each point in the genome, a distance toward the best sequencing point in the genome (peak point) based on three features was calculated as follow:

$$d_j = \|\mathbf{X_j} - P\|_2$$

where $X_j = (x_1, x_2, x_3)$ denotes the sequencing features of position j in the genome and x_1, x_2, x_3 denote the depth value of unique reads, reads accuracy and ration of unique reads, respectively. Before calculating the distance, each feature needs to be normalized using min-max normalization as below:

$$\hat{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \ (i = 1, 2, 3).$$

For a set of distances $\{d_1, d_2, \ldots, d_N\}$, we used GMM to model to classify genomic locus forms into three groups. The classification used Python SciKit-Learn package (version 1.3.0.)

$$p(d_j) = \sum_{k=0}^{K} \alpha_k \mathcal{N}(d_j|\mu_k, \sigma_k) \ (K = 1, 2, 3, j = 1, \ldots, N).$$

μ_k, σ_k means the mean distance and the standard deviation toward the best point in regions that can be properly ($k = 1$), partially ($k = 2$) and hardly ($k = 3$) *de novo* assembled. α_k means the proportion of each region in the genome.

Kolmogorov–Smirnov (KS) test was used to verify the groups classified by GMM were correct. If no significant difference (the *P*-value >0.05) showed in two adjacent groups, the skewed distribution was used as the alternative method to GMM. The random sample from two adjacent groups was used for KS test using Python SciPy (version 1.11.2).

An alternative method is based on skewed distributions using Python SciPy. After labeling outliers as Group III ($k = 3$), whose definition is data that are more than 1.5 times the inter-quartile range before first quartile (Q1) or after third quartile (Q3), the combination of Group I and Group II is considered as a skewed distribution. One-tailed *P*-value was used to determine the decision boundary $\hat{d}$ of Group I and Group II as below:

$$P(d < \hat{d}) = p \text{ if skewness} < 0$$

$$P(d > \hat{d}) = 1 - p \text{ if } skewness > 0.$$

Key Points

- DEGAP offers 'GapFiller' or 'CtgLinker' modes to bridge or reduce gaps in genomes.
- DEGAP autonomously adapts parameters based on three complexity factors affecting genome assembly, optimizing the extension path.
- DEGAP serves as a versatile tool for producing gap-free genomes, making it applicable across various contexts.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

The Science and Technology Innovation 2030 (2023ZD04062); the Major Project of Hubei Hongshan Laboratory (2022HSZD031); Huazhong Agricultural University (HZAU) Start-up Fund to J.Z.; National Natural Science Foundation of China (31821005) to L.X.; HZAU Special Funds for Interdisciplinary Scientific Research (SZYJY2022011) to H.S.

DATA AVAILABILITY

All described datasets are publicly available. The MH63 raw sequencing data and assembly MH63RS3 used for this project are previously archived at NCBI under accessions SRR10238608, SRR10188372, CP054676–CP054688 or at the National Genomics Data Center under BioProject no. PRJCA005549. Human reference genome GRCh38.p13 sequence data and genome were downloaded from https://github.com/marbl/CHM13.

CODE AVAILABILITY

DEGAP was developed as an open-source tool on a Linux platform with the Python programming language and could be downloaded at https://github.com/Jianwei-Zhang/DEGAP.

REFERENCES

1. Nurk S, Koren S, Rhie A, *et al.* The complete sequence of a human genome. *Science* 2022;**376**(6588):44–53.
2. Chen J, Wang Z, Tan K, *et al.* A complete telomere-to-telomere assembly of the maize genome. *Nat Genet* 2023;**55**(7):1221–31.
3. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. *Mol Ecol* 2017;**26**(20):5369–406.
4. Wenger AM, Peluso P, Rowell WJ, *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**(10):1155–62.
5. Koren S, Walenz BP, Berlin K, *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**(5):722–36.
6. Xiao CL, Chen Y, Xie SQ, *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* 2017;**14**(11):1072–4.
7. Chin CS, Peluso P, Sedlazeck FJ, *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;**13**(12):1050–4.
8. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**(5):540–6.
9. Cheng H, Concepcion GT, Feng X, *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**(2):170–5.
10. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**(2):155–8.
11. Huang Y, Koo DH, Mao Y, *et al.* A complete reference genome for the soybean cv. Jack. *Plant Commun* 2023;**5**(2):100765.
12. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;**13**(1):36–46.
13. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**(18):3094–100.
14. Song JM, Xie WZ, Wang S, *et al.* Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol Plant* 2021;**14**(10):1757–67.
15. Cheng Z, Dong F, Langdon T, *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* 2002;**14**(8):1691–704.
16. Schneider VA, Graves-Lindsay T, Howe K, *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;**27**(5):849–64.
17. Xu M, Guo L, Gu S, *et al.* TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* 2020;**9**(9):giaa094.
18. Piro VC, Faoro H, Weiss VA, *et al.* FGAP: an automated gap closing tool. *BMC Res Notes* 2014;**7**(1):371.
19. Xu GC, Xu TJ, Zhu R, *et al.* LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* 2019;**8**(1):8.
20. English AC, Richards S, Han Y, *et al.* Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. *PloS One* 2012;**7**(11):e47768.
21. Tamura K, Sakamoto M, Tanizawa Y, *et al.* A highly contiguous genome assembly of red perilla (*Perilla frutescens*) domesticated in Japan. *DNA Res* 2023;**30**(1):dsac044.
22. Li H, Handsaker B, Wysoker A, *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
23. Kurtz S, Phillippy A, Delcher AL, *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**(2):R12.