THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

[Next Generation Genomics]
[ICA]

Student Exam Number: …B196466……

The file name of your uploaded document **must** include your exam number

Quality assessment of Trachidermus fasciatus Genome

The rough-skinned cuckoo fish Trachidermus fasciatus can be found in East Asian beaches. The populations of this species are under peril due to environmental degradation and overfishing. It is critical to sequence the genome of this endangered species in order to safeguard it. However, there is no Trachidermus fasciatus reference genome, which makes studying this species extremely difficult. We employed a mix of nanopore long-read long sequencing, illumina short-read long sequencing, and Hi-C techniques to fully reconstruct the Trachidermus fasciatus genome. According to the findings, the estimated genome size is 542.6 Mbp (2 n = 40) and the scaffold N50 is 24.9 Mbp.[1]

Illumina short-read long sequencing

Illumina sequencing technology, which can produce hundreds of G or even T of sequencing data in tens of hours, is more than capable of achieving high-throughput sequencing's throughput needs. Its sequencing precision is likewise 100% assured. Because the 3'hydroxyl end of dNTP has a chemically cleavable part, it only allows the incorporation of a single base in each cycle, the surface of the reaction plate is scanned with a laser at this time, and the fluorescence of dNTP is used to read the nucleotides polymerized in each round of reaction for each temperature. The fluorescence of dNTP is used to read the nucleotide species of each template sequence at this point, and the base order of the target segment is finally determined through the "synthesis-wash-photography" cycle. The result shows that more than 350 million genomic and Hi-C sequencing reads were collected, totaling more than 50 billion nucleotide bases.

They chose this technique because not only can illumina short-read sequencing obtain a large amount of short sequence data at once and the cost of a single sequence is very low, but it can also generate chromosome-scale assemblies by combining short-read DNA sequencing and Hi-C chromatin interaction mate-pair sequencing. Integrating Hi-C and short-read data resulted in a scaffold spatial orienting accuracy of 99 percent.[1]

However, my observations are that the sequence read length of this technique is short, with a maximum of 250-300bp on the Illumina platform, and that because PCR is used to enrich sequences in library construction, some sequences with less content may not be amplified in large quantities, resulting in some information loss, and that there is a chance of mismatched bases being introduced during the PCR process. Higher coverage of sequencing is necessary to produce accurate and longer length splicing results, resulting in more mistakes and increasing cost of results.

Nanopore long-read long Sequencing

Nanopore sequencing is a type of sequencing that uses nanopores to provide highly sensitive single-molecule sequencing. When a molecule passes through a nanopore of equal size to its own, it has a considerable effect on the ion flow within the nanopore channel, which is subsequently turned into a detectable and analyzable electrochemical signal, similar to the Coulter principle of cell counters. Relevant information about the detected object can be obtained by detecting the state change of the nanopore, and its highly sensitive electrical signal can detect and analyse the state and structure of molecules passing through the pore in real time, allowing for single-molecule detection with high detection sensitivity. Approximately 4 million quality-controlled reads containing over 87 billion nucleotide bases were obtained after

nanopore sequencing. With a N50 of around 30 kbp, the longest readings were over 240 kbp.[1] They picked this technique because it can be utilised for genome assembly in a variety of species and can also considerably reduce the cost of current gene sequence analysis while also increasing the speed of analysis. Nanopore sequencing allows single molecules of DNA or RNA to be sequenced without the requirement for PCR amplification or chemical tagging. During the course of any previously created sequencing approach, at least one of these aforementioned phases is required. Nanopore sequencing has the potential to provide relatively low-cost genotyping, high mobility for testing, and the ability to rapidly process samples that can display results in real time. Solid-state nanopores, on the other hand, have numerous hurdles in manufacture, sequencing, and integration as a new sequencing method. Simultaneously, this approach employs hydrolytic sequencing, which cannot be repeated and so fails to attain adequate sequencing accuracy.

De Novo Genome Assembly

All genome assemblies are sequenced by breaking the genome into small fragments with the help of whole genome birdshot method, and then these small fragments are reassembled to restore the genomic information. Genome de novo assembly is an assembly based entirely on sequenced read segments without reliance on a reference genome or other genomes. The reference genome sequence of each species is generated by sequencing to obtain sequenced read segments of the genome, followed by de novo splicing or assembly.Finally, the sequence of each chromosome of the sequenced species is restored, i.e. the order of the four bases of ATGC. First the genome is sequenced to produce reads, then the reads are assembled to produce long fragments Contigs, then the orientation and order of Contigs are determined, longer fragments Scaffolds are assembled, and finally Scaffolds are assembled and joined to obtain the complete chromosome sequence. NextDenovo is utilised in this article for genome assembly, which includes error repair, preliminary assembly, and genome polishing. next Raw read rectification and common sequence extraction are handled by the Correct module. next For preliminary assembly, the Graph module is utilised, and for genome polishing, the NextPolish module is used.
The advantage of a non-reference genome dependent genome de novo assembly strategy is that it allows better assembly of some species-specific genome fragments, and many tools and software have been developed that improve the automation of computations and reduce the computer skills required of researchers. And genomes perform better in terms of assembly of species-specific sequences. However, my evaluation of the technique is that it has several obvious drawbacks, which are that in highly complex regions of the genome, it is difficult to assemble them using non-initial assembly strategies. And the time to assemble from scratch can be long. De novo genome assembly remains challenging due to short read lengths, missing data, duplicated regions, polymorphisms, and sequencing errors. Therefore, for large genomes with high complexity and repetitive sequences, this strategy may perform poorly in terms of continuity of genome assembly, i.e. small N50 or N90, and may perform poorly in terms of accuracy, i.e. many misconnections in terms of overlapping clusters and scaffolds. In this paper, the combined coverage of 23 alleles reached 90% based on N90 information, indicating that the number of N90 alleles is close to the number of haploid chromosomes.[1]

Quality Evaluation of Genome Assembly

The final quality of the genome must be assessed after it has been assembled. Each allele in the generated allele file should be lengthy enough to contain the whole gene structure, and the assembled allele should have a low error rate and contain as much of the original sequence as possible. However, this is in fact contradictory. The higher the contiguousness, the more ambiguous nodes must be analysed, resulting in a higher total mistake rate. Contig will be quite fragmented to assure perfect accuracy. CEGMA and BUSCO were used to analyse the completeness of the assembled genome in this work, and the results showed a high percentage of completeness (98.39 percent and 96.95 percent , respectively).[1]

Hi-C Proximity-Guided Assembly of Chromosome-Level Scaffolds

Current second and third generation sequencing techniques can only assemble genomes to the Contig/Scaffold level, but not at the chromosomal level. Hi-C aided assembly technology, on the other hand, may mount Contig/Scaffold to various chromosomes to increase genome quality and play an important role in the publication of genomic publications. Because Hi-C aided assembly technology can help with genome assembly down to the chromosomal level, it is currently indispensable. Hi-C assisted assembly has become a standard strategy for assisting genome assembly, owing to the incomplete statistics of genome articles. In this research, 44 scaffolds were built utilising Hi-C chromatin interaction data to rearrange alleles based on the information.[1]

Hi-C stands for high-throughput chromosome conformation capture technology, which was developed to capture all intra- and inter-chromatin spatial interactions information within the entire genome. It has since been applied to the study of spatial regulation mechanisms of gene expression, the creation of chromosome-level reference genomes, and the creation of haplotype maps. Hi-C technology is derived from Chromosome Conformation Capture (3C) technology, which studies the spatial position of the entire chromatin DNA in the whole genome and obtains high-resolution chromatin three-dimensional structure information using high-throughput sequencing technology combined with bioinformatic analysis methods.

Hi-C-based genome assembly was used in this article because it provides higher coverage and specificity than genetic mapping and physical mapping, as well as avoiding lengthy population creation effort, a shorter experimental cycle time, and lower costs. However, I believe that Hi-C technology's resolution is limited, and that there are difficulties assembling repetitive sequences like mitogen and telomeres, and that the theory of Hi-C technology's genome assembly assistance is based on "proximity interaction over remote interaction," but that this rule does not always hold in some specific regions, etc.[2]

To improve genome assembly, the first step is to improve the method. The present technology utilised in this paper is NextDenovo for genome assembly, but they can experiment with different algorithms, such as wtdbg, or build new algorithms to reduce computing time and enhance accuracy. In big genome assemblies, wtdbg is a new long read length sequencing assembly approach that is several times quicker and has little influence on assembly quality. Thus, to improve the continuity, completeness and accuracy of the assembly sequence. Secondly, Hi-C technique can collect genetic data at the chromosomal level, allowing scaffolds to swiftly locate chromosomes.[3]

Compared with Onychostoma macrolepis

Organism

Onychostoma macrolepis, a commercial carp, is a newcomer. The preliminary genome assembly size is 883.2 Mb, whereas the contig N50 size is 11.2 Mb.[5] There is no highly contiguous genome assembly of any carp family species equivalent to Trachidermus fasciatus based on long read length third generation sequencing. Although carps are not threatened, high-quality genome assembly would help researchers better understand the genomic structure and evolution of their polyploid relatives.[4]

sequencing approach

Trachidermus fasciatus uses illumina short-read sequencing and nanopore long-read sequencing in the same way. This demonstrates that the Nanopore sequencing platform, when combined with Illumina sequencing technology, may produce the best sequencing results. The Illumina TruSeq RNA Library Preparation Kit is also used to create RNA sequencing libraries. NextDenovo corrects sequencing errors, however WTDBG is used for genome assembly. After quality filtering 101.6 Gb of raw data, a total of 87.9 Gb was obtained via nanopore sequencing. The average read length was 18.16 kb, while the N50 read length was 23.51 kb. The longest read was 724.56 kb long. WTDBG's preliminary genome assembly size is around 876.5 megabytes, which is close to the expected genome size (928 Mb). The N50 of 914 overlapping clusters is 11.1 Mb. And these two outcomes are remarkably similar to the preceding species' nanopore sequencing result.[5]

assembly method

The greatest difference is that Bionano built the scaffold. The Nanochannel array-based technique developed by Bionano allows for high-throughput automated photo imaging of fluorescently labelled DNA molecules moving through the Nanochannel, allowing for the complete restoration of single-molecule structures. Genomic DNA is placed onto Saphyr chips after labelling and staining, and electrophoresed into massively parallel nanochannels to image fluorescently tagged DNA molecules, which are then tethered to a reference genome to form superscaffolds. The new method was developed by combining short read length DNA sequencing and Hi-C chromatin interaction paired sequencing, as opposed to the previous method, which used short read length DNA sequencing and Hi-C chromatin interaction paired sequencing.

Bionano optically mapped and created 969 corrected overlap clusters using Hi-C chromatin interaction data, resulting in a total of 44 scaffolds. All of these were put together to make 853 scaffolds. This indicates that Bionano may construct additional stands. The most striking similarity is that chromosomal assembly is likewise aided by Hi-C. Cohesive hierarchical clustering was used to anchor and target 881.3 MB of data from 526 scaffolds to 25 chromosomes. The process is the same as for the preceding species' assembly.[5]

Finally, the most prevalent methods of sequencing are illumina short-read long sequencing and nanopore long-read long sequencing. High percentages of repetitive sequences, high heterozygozity, excessive GC content, and the presence of genomes with difficult-to-eliminate heterologous DNA contamination can all be solved by combining the two. Due to the lack of a reference genome for these species, there is no way to compare sequenced sequences to the reference genome to facilitate assembly, hence de novo assembly is the only option. However, in terms of improving genome assembly, approaches such as Bionano and Hi-C can assist the genome in assembling in a more complete and reasonable manner.

References

[1] Xie, Gangcai, et al. "Nanopore Sequencing and Hi-C Based De Novo Assembly of Trachidermus fasciatus Genome." Genes 12.5 (2021): 692.

[2] Chen, Siyuan, et al. "HiC-LDNet: A general and robust deep learning framework for accurate chromatin loop detection in genome-wide contact maps." bioRxiv (2022).

[3] Ghurye, Jay. Genome Assembly and Variant Detection Using Emerging Sequencing Technologies and Graph Based Methods. Diss. University of Maryland, College Park, 2018.

[4] Wu, Jiayu, et al. "Age-Related Changes in the Composition of Intestinal Microbiota in Elderly Chinese Individuals." Gerontology (2022): 1-13.

[5] Sun, Lina, et al. "Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology." Molecular ecology resources 20.5 (2020): 1361-1371.