High quality reference genome assembly and annotation

Introduction

High-quality reference genomes, the keystone of modern genomics, are distinct from draft genomes due to their completeness, contiguity, and high percentage of bases assembled into pseudo-molecules[1]. Such high-resolution assembled genomes will enable the accurate identification of genomic components like genes, centromeres and intergenic regions.This is easily achieved for organisms with small genomes like viruses and prokaryotes, but the large size and complexity of eukaryotes, for example plants with large genome size and different ploidy levels, make it a difficult goal to accomplish. Notwithstanding the many rounds of sequencing and refinements, the current version of the rice reference genome, perhaps the most studied plant genome, still shows hundreds of gaps, mainly in complex regions like centromeres and telomeres.

Genome sequences provide essential resources for enabling research, development and application in the life sciences. High quality, accurately annotated reference genome sequences are key resources that can be employed for discovery and analysis of genetic variation and for linking genotypes to function. Genome sequences provide essential resources for enabling research, development and application in the life sciences. High quality, accurately annotated reference genome sequences are key resources that can be employed for discovery and analysis of genetic variation, for linking genotypes to function and for the development of panels of single nucleotide polymorphisms (SNPs). They help in the dissection of complex traits including responses to biotic and abiotic stresses. A wide range of comparative genomic and evolutionary studies can be performed, for example the identification of large structural variations, linkage blocks, genes, and alleles.

Technology platforms and combinations offered by AgriGenome

Short read sequencing by Illumina

Short read sequencing is excellent for generating high quality deep coverage for small to large size genomes. However, short read lengths have limitations in resolving complex regions, especially those with repetitive or heterozygous sequences. We use various Illumina sequencing platforms for short read sequencing such as HiSeqX, HiSeq4000, and HiSeq2500. They produce short reads of 50 to 150 bp.

Long read sequencing

We use PacBio[2] Sequel system and Oxford Nanopore Technologies[3,4] devices (MinION/GridION/PromethION) to generate long reads for Whole Genome Sequencing (WGS). PacBio Sequel system, based on Single Molecule Real-Time (SMRT) Sequencing technology, can generate up to 20 Gb per SMRT Cell with average read lengths up to 10 kb and high consensus accuracies (>99.999%). High-quality whole genomes can be generated within a short period using the PacBio Sequel system. Nanopore sequencers can produce long reads with median and maximum read lengths of 6 kb and 60 kb respectively.

Linked read sequencing using Chromium by 10x Genomics

Linked read technology using (Chromium) by 10x Genomics enables us to explore and characterize intracellular heterogeneity. It enables diploid genome assembly and phased genome assembly and resolves structural variation, nucleotide variants and reveals maps of the genomic landscape.

Technologies to improve genome contiguity

Hi-C

Hi-C[5] (Dovetail & Arima) method is used to study the 3-D architecture of genomes, which helps unveil the 3-D folding of chromosomes and arrangements of distant functional elements such as promoters and enhancers and ultimately, the relationship between chromosome organization and genome activity. Hi-C produces sequencing-ready libraries and high quality Hi-C data (on Illumina) that help to build chromosome-level assemblies.

Bionano optical mapping

Bionano uses optical maps generated by fluorescent-labeled ultralong DNA molecules of >150 kb in length. A *de novo* optical map assembly can be compared to a DNA sequence assembly and can be used for correction of mis-assemblies, stitching of scaffolds and correction of incorrect assembly of contigs in highly complex areas of the genome. Bionano's next-generation mapping combines proprietary NanoChannel arrays with optical mapping to image extremely long, high-molecular-weight DNA in its most native state[6]. Optical mapping also enables structural variation detection with high sensitivity and generates highly contiguous genome assemblies.

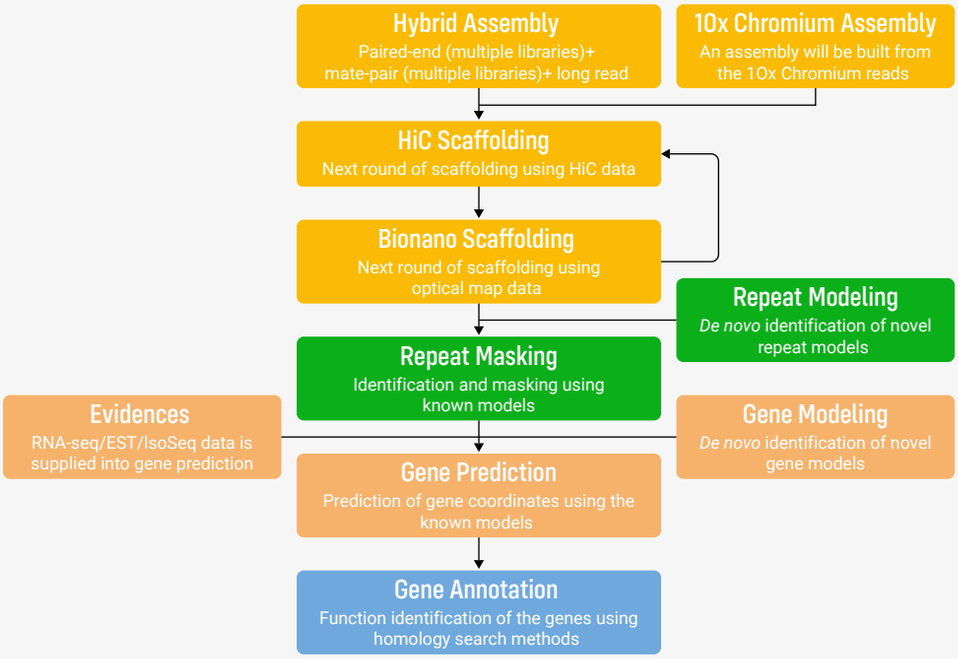
Deliverables

Item	Description
Raw Data	Raw data files in the respective formats like fastq, fasta, fast5 and bnx
Assembled genome	Nucleotide sequence file of the assembly in fasta format
Repeat information	The classes of repeats identified and the coordinates on the genome
Predicted gene coordinates	Genes and their coordinates in gff format
Functional annotations	Genes and their functional details in tabular format

Data analysis strategies at AgriGenome

Genome assembly is done at various levels depending on the sequencing platforms used and the specifications of the genome under study. The strategies are well explained in a research article from our groups on the reference genome of the Indian Cobra[7].

Analysis pipelines at AgriGenome employ a combination of various tools and databases to achieve the best quality genomes. Our in-house pipelines and tools to iteratively improve gene models and predict genes are capable of annotating all the genes in the genome and identify the best possible functional characterisation of them with optimum specificity and sensitivity. The checkpoints in the pipelines make it possible to evaluate each result and fine-tune the experimental settings to generate the best assembly and annotation outcomes.



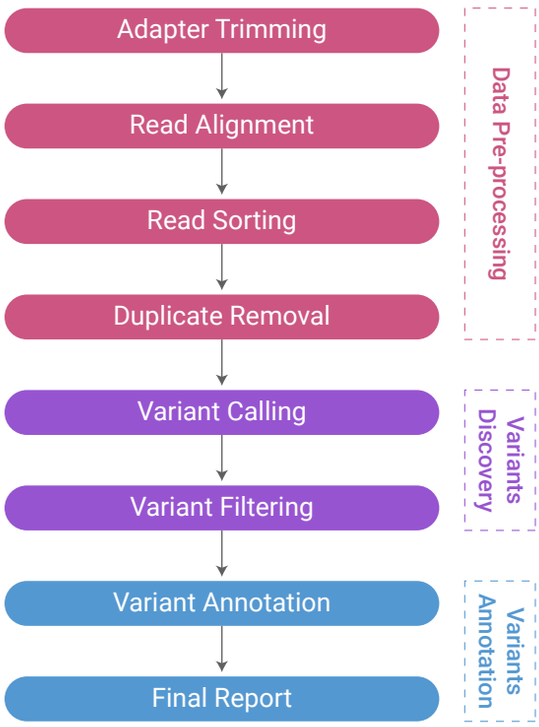
Whole genome re-sequencing

Applications

- Identifies nucleotide variants within species which help to study Genotype-Phenotype relation.
- Population screening, linkage mapping and QTL identification.
- Study genetic drift and evolutionary divergence between species.
- Variant annotation and functional elucidation of the pathways.

Bioinformatics Analysis

- Raw data quality control and length filter
- Reference-based mapping
- Variant calling information
- Variant annotation
- Comparison between samples (if required)



References

1. A reference standard for genome biology. Nat Biotechnol 36, 1121 (2018). https://doi.org/10.1038/nbt.4318
2. Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. Genomics, proteomics & bioinformatics, 13(5), 278-289.
3. Mikheyev, A. S., & Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. Molecular ecology resources, 14(6), 1097-1102.
4. Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., ... & Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nature communications, 9(1), 4844.
5. Van Berkum, Nynke L., et al. "Hi-C: a method to study the three-dimensional architecture of genomes." JoVE (Journal of Visualized Experiments) 39 (2010): e1869.
6. Reisner W, Larsen NB, Silahtaroglu A, Kristensen A, Tommerup N, Tegenfeldt JO, Flyvbjerg H. Single-molecule denaturation mapping of DNA in nanofluidic channels. Proc Natl Acad Sci U S A. 2010 Jul 27;107(30):13294-9. doi:10.1073/pnas.1007081107
7. Suryamohan, K., Krishnankutty, S.P., Guillory, J. et al. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nat Genet 52, 106–117 (2020) doi:10.1038/s41588-019-0559-8

FOR MORE INFORMATION

Kochi
5th Floor, SCK 01 Building "SmartCity Kochi",
Infopark Road, Kakkanad,
Kerala, 682 042, India

Hyderabad
DS-10, The Sustainability Innovation Center (SINC),
IKP Knowledge Park, Genome Valley, Koltur, Hyderabad,
Telengana, 500 078, India

Delhi
27 Ground Floor, TDI Centre,
Jasola District Centre,
New Delhi, 110 025, India

+91-484-4234-234 sales@aggenome.com