A quick view at the cauliflower genome

Presented by:

Nada Alhadidi

Intro to genomics (485)

Spring 2020

Outline

- Genome paper
- Interesting facts
- General information about the genome
- Sequencing strategy
- Sequencing method
- Assembly
- Annotation
- Questions

Genome paper

Horticulture Research

Article | Open Access | Published: 01 July 2019

Draft genome sequence of cauliflower (*Brassica oleracea* L. var. *botrytis*) provides new insights into the C genome in *Brassica* species

Deling Sun, Chunguo Wang, Xiaoli Zhang, Wenlin Zhang, Hanmin Jiang, Xingwei Yao, Lili Liu, Zhenghua Wen, Guobao Niu & Xiaozheng Shan

Interesting facts

- Cauliflower comes in 4 different colours (**creamy-white, purple, orange, green**)
- Sunlight can give it a yellow tint.
- The stems and leaves are edible.
- It can reduce the risk of developing certain types of cancers.
- It's high in fiber and can help in the digestion process.
- It Contains 10% Of Your Daily Vitamin C Need.

Interesting facts

Most common

presence anthocyanins
water-soluble pigment

contains beta-carotene
as the orange pigment

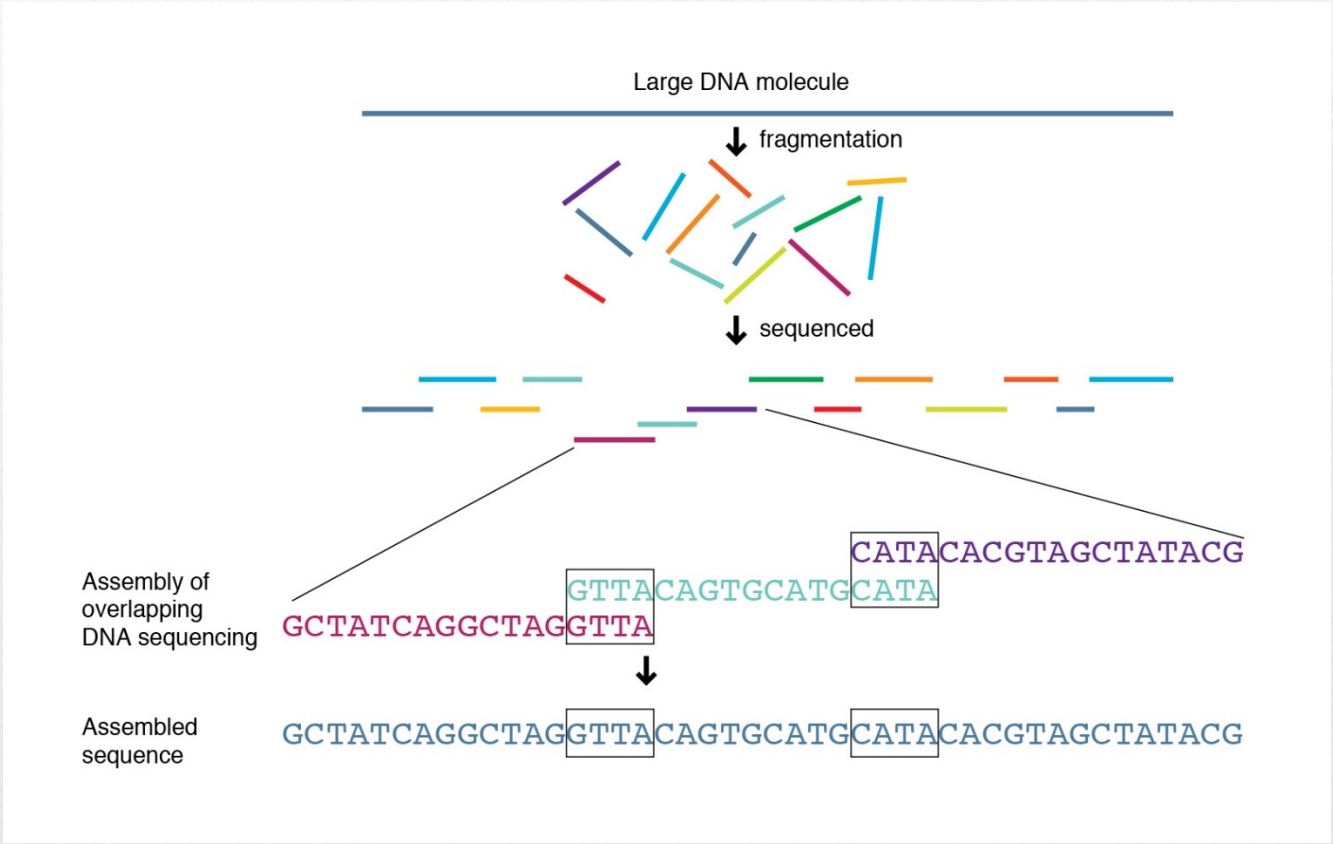
sometimes called
broccoflower

General information about the genome

- Scientific name is (Brassica oleracea L. var. Botrytis)
- Genus Brassica contains three basic genomes (A, B and C) that form three diploid species:-
  - Brassica rapa (AA genome)
  - Brassica nigra (BB genome)
  - Brassica oleracea (CC genome)
- It differs from most Brassica species in its formation of a specialized organ called the curd
- Curds are the primary edible organs of cauliflower
- Raw cauliflower is 92% water, 5% carbohydrates, 2% protein, and contains negligible fat

Sequencing strategy

Whole genome shotgun sequencing (WGS).



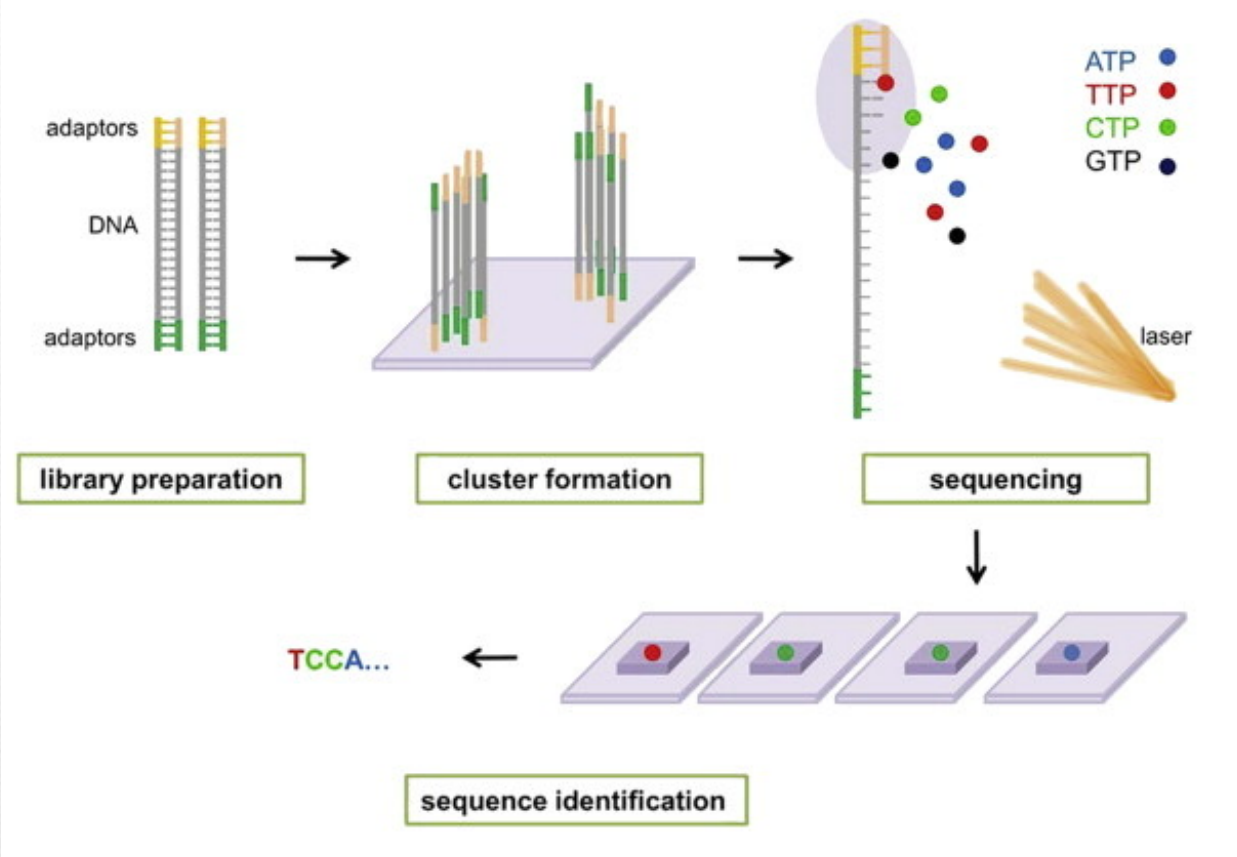
Sequencing method: illumina

- Sequencing by synthesis (SBS) is a widely adopted next-generation sequencing (NGS)
- Responsible for generating more than 90% of the world's sequencing data.
- Supports both single-read and paired-end libraries.
- A combination of short inserts and longer reads that increases the ability to fully characterize any genome
- Using a proprietary method that detects single bases as they are incorporated into growing DNA strands.

Sequencing method: illumina

- A fluorescently labeled reversible terminator is imaged as each dNTP is added
- Then cleaved to allow incorporation of the next base
- Since all 4 reversible terminator bound, dNTPs are present during each sequencing cycle
- The natural competition minimizes incorporation bias
- The end result is true base-by-base sequencing that enables accurate data

Sequencing method: illumina



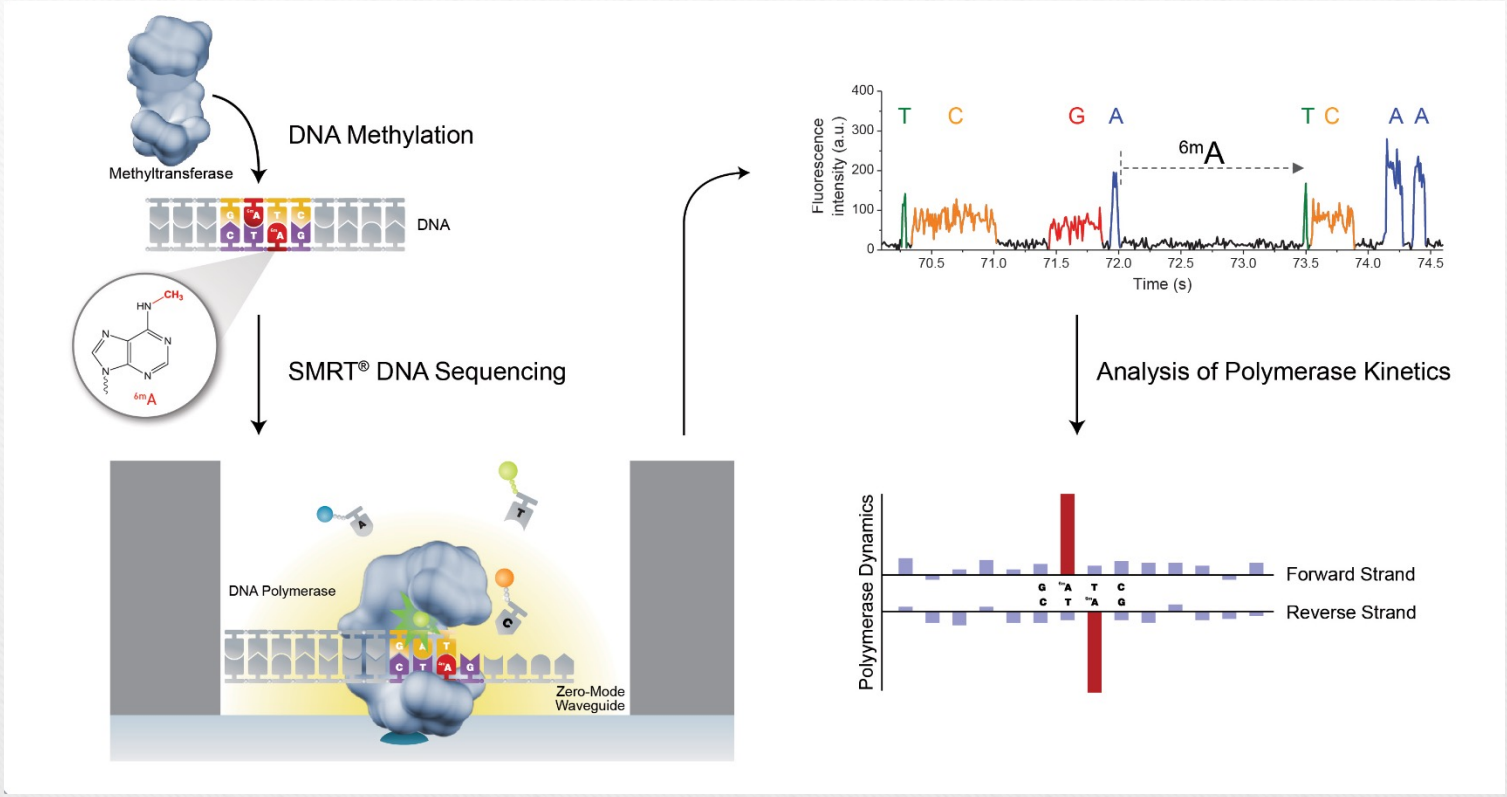
Sequencing method: PacBio

- It's a method for real time sequencing and doesn't require pause between steps.
- Offers longer read lengths, making it well-suited for unsolved problems in genome.
- With longer reads, we can sequence through extended repetitive regions and detect mutations.
- Although it offers long read lengths, it has a higher error rate.
- Sequence full-length transcripts or fragments with significant lengths.
- Provides information that is useful for the direct detection of base modifications, such as methylation.

Sequencing method: PacBio

- A template called a SMRTbell created by ligating hairpin adaptors to both ends of (dsDNA)
- SMRTbell diffuses into a sequencing unit called a zero-mode waveguide (ZMW)
- Four fluorescent-labeled nucleotides are added to the SMRT cell
- SMRT uses the innovation of (ZMW) to distinguish between ideal and strong fluorescent signal
- A light pulse is produced and the pulses corresponding to each (ZMW)
- The dye-linker-pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW, ending the fluorescence pulse

Sequencing method: PacBio



Assembly

	Number	Size	Sequence coverage (X)	Percentage
Estimate of genome size		603.04 Mb		
PacBio reads		69.06 Gb	114.52	
Illumina reads		45.99 Gb	76.26	
Total reads		115.05 Gb	190.78	
Contigs	1,484	584.60 Mb		
Coverage of sequenced genome				96.94 %
N50 of contigs	82	2.11 Mb		
Longest contig		9.81 Mb		
GC content				36.76 %
Total repetitive sequences		331.20 Mb		56.65 %
Total protein-coding genes	47,772	108.40 Mb		18.54 %
Annotated protein-coding genes	46,628			97.60 %
Average length per gene (exon + intron)		2,035 bp		
Average exons per gene	4.78	242 bp		
Average length per intron		260 bp		
Noncoding RNAs	8,106	1.32 Mb		0.23 %

Annotation

- Combining between (de novo and homology)
- Total of 56.65% was composed of repetitive elements.
- LTRs were the most abundant → 32.71%
- TEs → 12.62%
- LINEs → 5.20%
- SINEs → 0.06%
- RNAs → 0.23%
- Tandem repeats → 9.34%

Questions

- What is the difference between cauliflower and some other species?

- cauliflower differ in its formation of a specialized organ called the curd

- What other parts of cauliflower can be eaten other than the head?

- Stems and leaves are edible