A journey to understanding cabbage

The genomic reveals

Zahraa Hasan

24/9/2020

Outline

- Introducing *Brassica*
- Genome paper
- Mesopolyploidy?
- General information about the genome
- Sequencing strategies
- Sequencing methods
- Assembly
- Genome outcome

Introducing *Brassica*

- 15 *Brassica* species,
  28 *Brassica* subspecies,
  8 *Brassica rapa* subspecies
- e.g. mustard, broccoli, cauliflower, and cabbages
- Cabbage cultivation started 6000 years ago in Shensi, China

Genome paper

Wang *et al*,. (2011). The genome of the mesopolyploid crop species Brassica rapa. Nature genetics, 43(10), 1035-1039.

nature genetics

The genome of the mesopolyploid crop species *Brassica rapa*

The *Brassica rapa* Genome Sequencing Project Consortium

We report the annotation and analysis of the draft genome sequence of *Brassica rapa* accession Chiifu-401-42, a Chinese cabbage. We modeled 41,174 protein coding genes in the *B. rapa* genome, which has undergone genome triplication. We used *Arabidopsis thaliana* as an outgroup for investigating the consequences of genome triplication, such as structural and functional evolution. The extent of gene loss (fractionation) among triplicated genome segments varies, with one of the three copies consistently retaining a disproportionately large fraction of the genes expected to have been present in its ancestor. Variation in the number of members of gene families present in the genome may contribute to the remarkable morphological plasticity of *Brassica* species. The *B. rapa* genome sequence provides an important resource for studying the evolution of polyploid genomes and underpins the genetic improvement of *Brassica* oil and vegetable crops.

Model species have provided valuable insights into angiosperm (flowering plant) genome structure, function and evolution. For example, *A. thaliana* has experienced two genome duplications since its divergence from *Carica*, with rapid DNA sequence divergence, extensive gene loss and fractionation of ancestral gene order eroding the resemblance of *A. thaliana* to ancestral Brassicales[1]. Compared with an ancestor at just a few million years ago, *A. thaliana* has undergone a ~30% reduction in genome size[2] and 9–10 chromosomal rearrangements[3,4] that differentiate it from its sister species *Arabidopsis lyrata*. Whole-genome duplication has been observed in all plant genomes sequenced to date. *A. thaliana* has undergone three paleo-polyploidy events[5]: a paleohexaploidy (γ) event shared with most dicots (asterids and rosids) and two paleotetraploidy events (β then α) shared with other members of the order Brassicales. *B. rapa* shares this complex history but with the addition of a whole-genome triplication (WGT) thought to have occurred between 13 and 17 million years ago (MYA)[6,7], making 'mesohexaploidy' a characteristic of the Brassiceae tribe of the Brassicaceae[8].

Brassica crops are used for human nutrition and provide opportunities for the study of genome evolution. These crops include important vegetables (*B. rapa* (Chinese cabbage, pak choi and turnip) and *Brassica oleracea* (broccoli, cabbage and cauliflower)) as well as oilseed crops (*Brassica napus*, *B. rapa*, *Brassica juncea* and *Brassica carinata*), which provide collectively 12% of the world's edible vegetable oil production[9]. The six widely cultivated *Brassica* species are also a classical example of the importance of polyploidy in botanical evolution, described by 'U's triangle'[10], with the three diploid species *B. rapa* (A genome), *Brassica nigra* (B genome) and *B. oleracea* (C genome) having formed the amphidiploid species *B. juncea* (A and B genomes), *B. napus* (A and C genomes) and *B. carinata* (B and C genomes) by hybridization. Comparative physical mapping studies have confirmed genome triplication in a common ancestor of *B. oleracea*[11] and *B. rapa*[12] since its divergence from the *A. thaliana* lineage at least 13–17 MYA[6,7,13].

Using 72× coverage of paired short read sequences generated by Illumina GA II technology and stringent assembly parameters, we assembled the genome of the *B. rapa* ssp. *pekinensis* line Chiifu-401-42 and analyzed the assembly (Online Methods and **Supplementary Note**). The final assembly statistics are summarized in **Table 1**. The assembled sequence of 283.8 Mb was estimated to cover >98% of the gene space (**Supplementary Table 1**) and is greater than the previous estimated size of the euchromatic space, 220 Mb[14]. The assembly showed excellent agreement with the previously reported chromosome A03 (ref. 15) and with 647 bacterial artificial chromosomes (BACs)[14] (Online Methods) sequenced by Sanger technology. Integration with 199,452 BAC-end sequences produced 159 super scaffolds representing 90% of the assembled sequences, with an N50 scaffold (N50 scaffold is a weighted median statistic indicating that 50% of the entire assembly is contained in scaffolds equal to or larger than this value) size of 1.97 Mb. Genetic mapping of 1,427 markers in *B. rapa* allowed us to produce ten pseudo chromosomes that included 90% of the assembly (**Supplementary Table 2**).

We found the difference in the physical sizes of the *A. thaliana* and *B. rapa* genomes to be largely because of transposable elements (**Supplementary Table 3**). Although widely dispersed throughout the genome, as shown in **Figure 1**, the transposon-related sequences were most abundant in the vicinity of the centromeres. We estimated that transposon-related sequences occupy 39.5% of the genome, with the proportions of retrotransposons (with long terminal repeats), DNA transposons and long interspersed elements being 27.1%, 3.2% and 2.8%, respectively (**Supplementary Tables 4** and **5**).

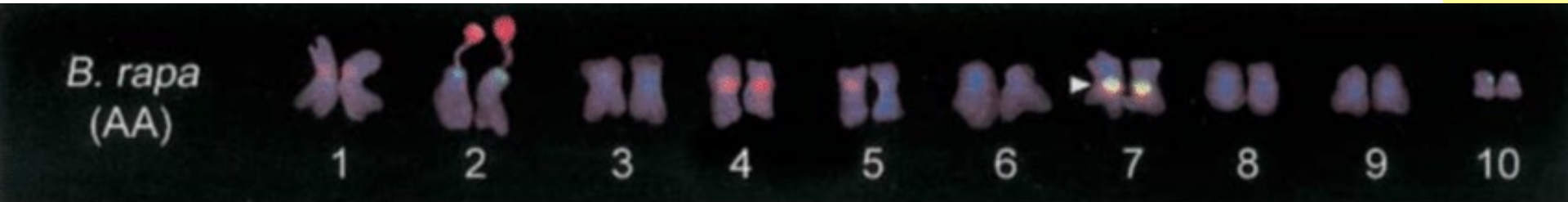
We modeled and analyzed protein coding genes (described in the Online Methods and the **Supplementary Note**). We identified 41,174 protein coding genes, distributed as shown in **Figure 1**. The gene models have an average transcript length of 2,015 bp, a coding length of 1,172 bp and a mean of 5.03 exons per gene, both similar to that observed in *A. thaliana*[16]. A total of 95.8% of gene models have a match in at least one of the public protein databases and 99.3% are represented among the public EST collections or *de novo* Illumina mRNA-Seq data. Among the total 16,917 *B. rapa* gene families, only 1,003 (5.9%) appear to be lineage specific, with 15,725 (93.0%) shared with *A. thaliana*[16] and 9,909 (58.6%) also shared by *Carica papaya*[17] and *Vitis vinifera*[18] (**Fig. 2**).

A full list of members appears at the end of the paper.

Received 7 March 2011; accepted 3 August 2011; published online 28 August 2011; doi:10.1038/ng.919

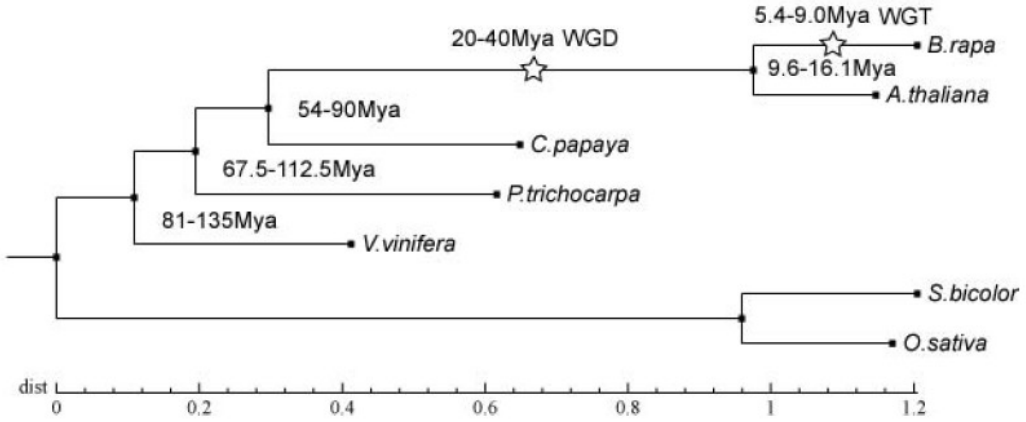
Mesopolyploid?

- A species that undergone whole genome multiplication in several million years ago
- Whole-genome duplication has been observed in all plant genomes sequenced to date
- *B. rapa* genome changed from diploid to tetraploid, then from tetraploid to hexaploid
- This generated genome triplication



Mesopolyploid?

- Chromosome number decreased by genome-wide chromosomes fusion around 6 to 9 million years ago
- Good source for genome evolution studies
- The understanding of *B. rapa* genome helped understanding other crops with triplicated genomes (e.g. bread wheat)

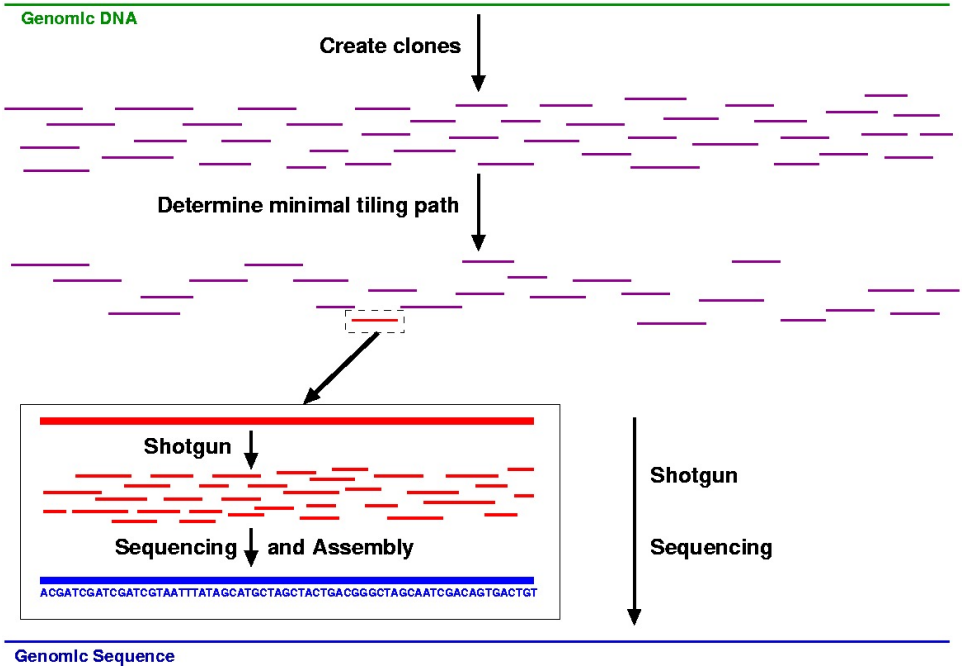


General information about the genome

Number of Chromosomes:	10 pairs, 2n=20
Genome Size:	283.8 Mb
# of coding genes:	41,174
Sequencing Strategies:	Hierarichal BAC clone sequencing and whole genome shotgun (WGS)
Sequencing Methods:	illumina and Sanger
shotgun coverage:	72
Contig N50:	27 kb
scaffold N50:	2 Mb
Assembly software:	SOAP de novo

Sequencing strategies

- Cabbage genome was shotguned
- 647 online published BAC sequences were compared with their equivalent WGS sequences



Sequencing methods

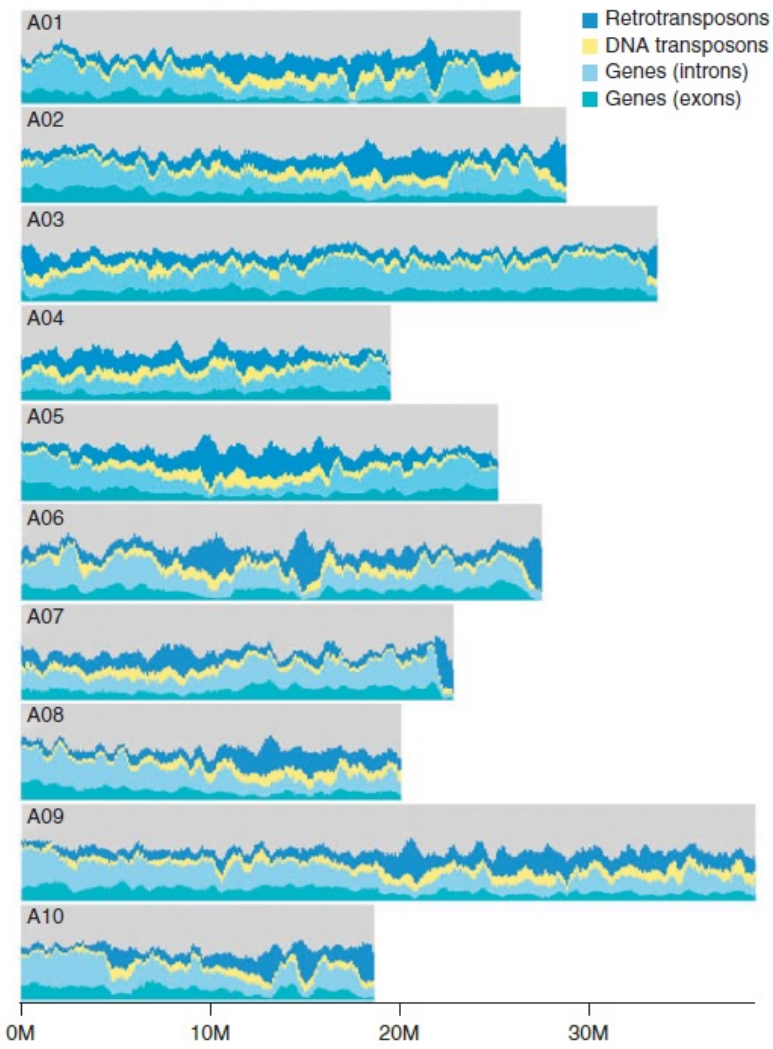
- In this project, DNA libraries were generated using illumina
- Illumina reads are short (~200 bp), medium (~500 bp), and long (~2 kb, 5 kb and 10 kb)
- Sanger DNA libraries were obtained from online sources
- Illumina sequences were aligned against Sanger sequences

Assembly

- Scaffolds were *de novo* assembled
- 60,521 contigs had a total size of 264 Mb
- 40,549 scaffolds had a total size of 283 Mb

Genome outcome

- Many genes near telomeres
- Many transposons near the centromeres



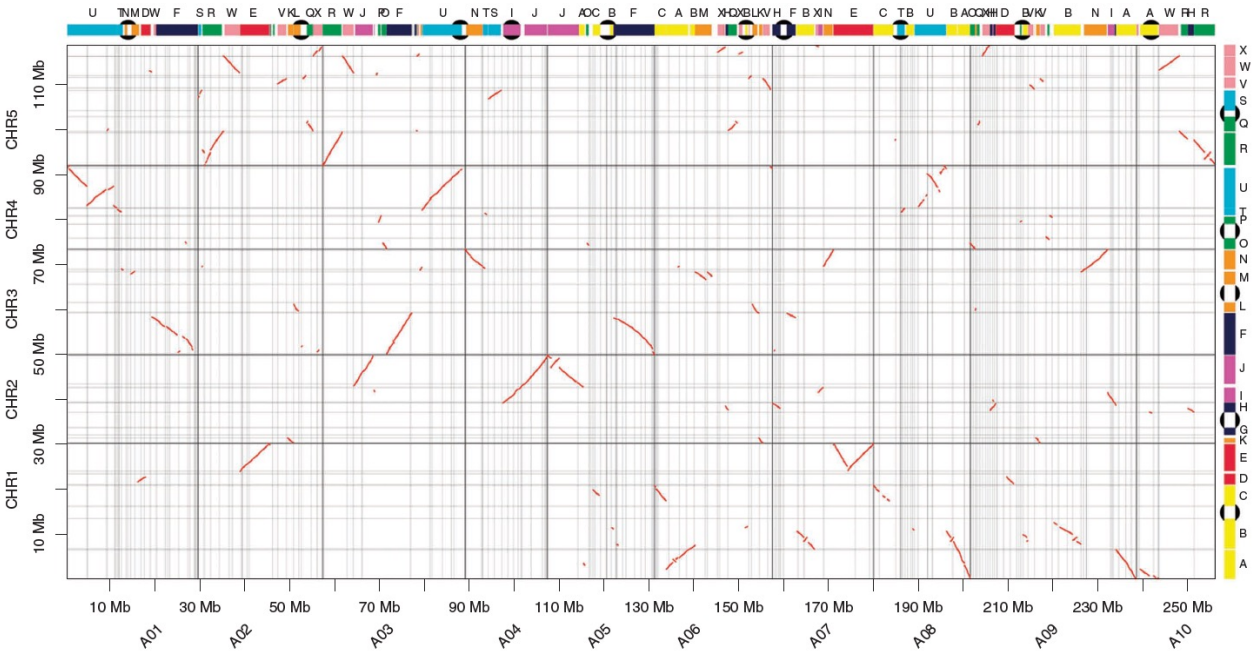
Interspersed repeats of *B. rapa*

39.5% of the genome is transposable elements

27.1% are retrotransposons, 3.2% are DNA transposons, and 2.8 % are LINEs

Cabbage and *Arabidopsis thaliana*

- *A. thaliana* experienced two genome duplications
- Cabbage experienced whole genome triplication
- Genomes were partitioned and visualized in this figure
- Cabbages & *A. thaliana*. have many shared orthologous regions



References

1. Mandakova, T., Joly, S., Krzywinski, M., Mummenhoff, K., & Lysak, M. A. (2010). Fast diploidization in close mesopolyploid relatives of Arabidopsis. The Plant Cell, 22(7), 2277-2290.
2. Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., ... & Denoeud, F. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome biology, 15(6), 1-18.
3. Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., ... & Huang, S. (2011). The genome of the mesopolyploid crop species Brassica rapa. Nature genetics, 43(10), 1035-1039.
4. http://brassicagenome.net/

Figures: [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]