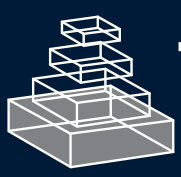
frontiers

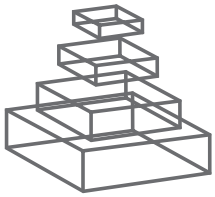
RESEARCH TOPICS

THE BRASSICA GENOME

Topic Editors
Xiaowu Wang and Michael Freeling



frontiers in
PLANT SCIENCE



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-136-9
DOI 10.3389/978-2-88919-136-9

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE BRASSICA GENOME

Topic Editors:
Xiaowu Wang, Institute of vegetables and flowers, CAAS, China
Michael Freeling, University of California, Berkeley, USA

Brassica rapa var Chiifu. These Chinese cabbages, being hardened during the winter of 2012, are the source of the reference Brassica genome. Freeling owns this image.

The genus Brassica is comprised of diploid and tetraploid species and includes many important crop plants. Several Brassica genomes have been sequenced are the subject of intensive investigation. The immediate impetus for a special Research Topic is the publication of genome sequence of B. rapa . B. rapa is of relatively recent paleopolyploid origin. Its triplicated genome is old enough such that the three genomes have diverged significantly, and young enough such that useful comparisons can be made using Arabidopsis thaliana as an out group, making the B. rapa genome an interesting model for comparative genomics and the analysis of genome evolution. Analysis of B. rapa is also informed by analyses of other Brassica genomes, and reciprocally, understanding of those genomes will be informed by comparisons with the B. rapa genome.

We welcome all types of articles on subjects including comparative genomics, genome evolution, and functional genomics, as well as analyses of specific gene families or genes in specific pathways and utilization of genomic data in molecular breeding of Brassica species.

Table of Contents

04 ***The Brassica Genome***
Michael Freeling and Xiaowu Wang

05 ***Syntenic Gene Analysis Between Brassica rapa and Other Brassicaceae Species***
Feng Cheng, Jian Wu, Lu Fang and Xiaowu Wang

11 ***A Phylogenetic Analysis of the Brassicales Clade Based on an Alignment-Free Sequence Comparison Method***
Klas Hatje and Martin Kollmar

23 ***Sequencing of Chloroplast Genome Using Whole Cellular DNA and Solexa Sequencing Technology***
Jian Wu, Bo Liu, Feng Cheng, Nirala Ramchiary, Su Ryun Choi, Yong Pyo Lim and Xiao-Wu Wang

30 ***Inferring the Brassica rapa Interactome Using Protein–Protein Interaction Data from Arabidopsis thaliana***
Jianhua Yang, Kim Osman, Mudassar Iqbal, Dov J. Stekel, Zewei Luo, Susan J. Armstrong and F. Chris H. Franklin

45 ***Unleashing the Genome of Brassica rapa***
Haibao Tang and Eric Lyons

57 ***Escape from Preferential Retention Following Repeated Whole Genome Duplications in Plants***
James C. Schnable, Xiaowu Wang, J. Chris Pires and Michael Freeling

65 ***The Impact of Genome Triplication on Tandem Gene Evolution in Brassica rapa***
Lu Fang, Feng Cheng, Jian Wu and Xiaowu Wang

72 ***Identification and Characterization of Orthologs of AtNHX5 and AtNHX6 in Brassica napus***
Brett A. Ford, Joanne R. Ernest and Anthony R. Gendall

84 ***A Candidate Gene-based Association Study of Tocopherol Content and Composition in Rapeseed (Brassica napus)***
Steffi Fritsche, Xingxing Wang, Jinquan Li, Benjamin Stich, Friedrich J. Kopisch-Obuch,Jessica Endrigkeit, Gunhild Leckband, Felix Dreyer, Wolfgang Friedt, Jinling Meng and Christian Jung

108 ***DNA-Based Genetic Markers for Rapid Cycling Brassica rapa (Fast Plants Type) Designed for the Teaching Laboratory***
Eryn E. Slankster, Jillian M. Chase, Lauren A. Jones and Douglas L. Wendell

116 ***Genetic Analysis of Morphological Traits in a New, Versatile, Rapid-cycling Brassica rapa Recombinant Inbred Line Population***
Hedayat Bagheri, Mohamed El-Soda, Inge van Oorschot, Corrie Hanhart, Guusje Bonnema,Tanja Jansen-van den Bosch, Rolf Mank, Joost J. B. Keurentjes, Lin Meng, Jian Wu, Maarten Koornneef and Mark G. M. Aarts

frontiers in
PLANT SCIENCE

EDITORIAL
published: 30 May 2013
doi: 10.3389/fpls.2013.00148

The Brassica genome

***Xiaowu Wang*[1]* *and Michael Freeling*[2]**

[1] *The Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China*
[2] *Plant and Microbial Biology, University of California, Berkeley, CA, USA*
**Correspondence: wangxw@mail.caas.net.cn*

Edited by:
Richard A. Jorgensen, University of Arizona, USA

Reviewed by:
Richard A. Jorgensen, University of Arizona, USA

BRASSICA GENOME RESEARCH TOPIC

Brassica species include important crops and provide unique materials for the study of genome evolution. These crops include six important vegetables and oilseed crops, which have been classically described by "U's triangle'. The three diploid species *B. rapa* (A genome), *B. nigra* (B genome), and *B. oleracea* (C genome) have 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. Moreover, the three diploid species themselves are evolved from a paleohexaploid, that is old enough to allowing the species being evolved into diploid species and young enough to maintain significant synteny with its ancestors. These make the *Brassica* species of the uniqueness of polyploidy in botanical evolution.

Brassicas are closely related to the model plant *Arabidopsis thaliana*, one of the most extensively studied species in the world. Sequencing of the genome of *Brassica rapa* provided a great opportunity to bridge the rich knowledge obtained from Arabidopsis to be transferred to a cultivated species. Yet, tools and resources need to be established to accomplish the knowledge transfer. The release of the *B. rapa* var. Chiifu genome is not only of importance for genome evolution research, but also facilitates the gene discovery and breeding of the Brassica crops. Conversely, using the duplicated *Brassica* species as "deletion machines" to better understand the cis/trans-relationships of ENCODE-like features that are accumulating all over the arabidopsis genome is an additional, fundamental reason for continued study of the *Brassica* species.

This Research Topic—The *Brassica* Genome—gathers contributions that report establishment of novel tools and methods, comparative genomics, gene discovery, molecular marker development, and genetic dissection of important traits. We hope this modest compendium marks the beginning of a vibrant future for *Brassica* comparative genome biology, and points the way toward how the Brassica lineage of crucifers, with arabidopsis as the outgroup, can and will revolutionize studies of eukaryotic gene and genome regulation and the phenotypes they specify.

Received: 30 April 2013; accepted: 01 May 2013; published online: 30 May 2013.
*Citation: Wang X and Freeling M (2013) The Brassica genome. Front. Plant Sci. **4**:148. doi: 10.3389/fpls.2013.00148*
This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 30 August 2012
doi: 10.3389/fpls.2012.00198

Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species

Feng Cheng, Jian Wu, Lu Fang and Xiaowu Wang*

Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China

Edited by:
Michael Freeling, University of California, Berkeley, USA

Reviewed by:
Michael Freeling, University of California, Berkeley, USA
Haibao Tang, J. Craig Venter Institute, USA

****Correspondence:***
Xiaowu Wang, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China.
e-mail: wangxw@mail.caas.net.cn

Chromosomal synteny analysis is important in genome comparison to reveal genomic evolution of related species. Shared synteny describes genomic fragments from different species that originated from an identical ancestor. Syntenic genes are orthologs located in these syntenic fragments, so they often share similar functions. Syntenic gene analysis is very important in Brassicaceae species to share gene annotations and investigate genome evolution. Here we designed and developed a direct and efficient tool, SynOrths, to identify pairwise syntenic genes between genomes of Brassicaceae species. SynOrths determines whether two genes are a conserved syntenic pair based not only on their sequence similarity, but also by the support of homologous flanking genes. Syntenic genes between *Arabidopsis thaliana* and *Brassica rapa*, *Arabidopsis lyrata* and *B. rapa*, and *Thellungiella parvula* and *B. rapa* were then identified using SynOrths. The occurrence of genome triplication in *B. rapa* was clearly observed, many genes that were evenly distributed in the genomes of *A. thaliana*, *A. lyrata*, and *T. parvula* had three syntenic copies in *B. rapa*. Additionally, there were many *B. rapa* genes that had no syntenic orthologs in *A. thaliana*, but some of these had syntenic orthologs in *A. lyrata* or *T. parvula*. Only 5,851 genes in *B. rapa* had no syntenic counterparts in any of the other three species. These 5,851 genes could have originated after *B. rapa* diverged from these species. A tool for syntenic gene analysis between species of Brassicaceae was developed, SynOrths, which could be used to accurately identify syntenic genes in differentiated but closely-related genomes. With this tool, we identified syntenic gene sets between *B. rapa* and each of *A. thaliana*, *A. lyrata*, *T. parvula*. Syntenic gene analysis is important for not only the gene annotation of newly sequenced Brassicaceae genomes by bridging them to model plant *A. thaliana*, but also the study of genome evolution in these species.

Keywords: synteny, ortholog, *Brassica rapa, Arabidopsis thaliana, Arabidopsis lyrata, Thellugiella parvula,* Brassicaceae

INTRODUCTION

The genomes of species from Brassicaceae are composed of 24 basic genomic blocks, A–X (24 GBs, also called the ancestral karyotypes, AK) (Schranz et al., 2006), as detected by studies using comparative chromosomal painting (CCP) (Mandakova and Lysak, 2008), and also directly observed in *Arabidopsis thaliana* (Initiative, 2000), and the newly sequenced genomes of *Arabidopsis lyrata* (Hu et al., 2011), *Brassica rapa* (Wang et al., 2011), and *Thellungiella parvula* (Dassanayake et al., 2011). Brassicaceae genomes have experienced different levels of genomic reshuffling, fragmentation, deletion, segmental or whole genome duplication (Schranz et al., 2006; Mandakova and Lysak, 2008). Along with drastic genome changes, gene contents have also rapidly evolved (Cheng et al., 2012; Tang et al., 2012). However, as they share a common AK, the chromosomal synteny relationship among these genomes is considered to be well preserved despite the long time for evolution following the divergence of these species (Tang et al., 2008; Cheng et al., 2012).

Chromosomal synteny analysis is important in genome comparison to reveal the genomic evolution of related species (Tang et al., 2008). Shared synteny describes genomic fragments from different species that originated from a certain common ancestor (Lyons et al., 2008). Syntenic genes are orthologs located in these syntenic fragments, so they often share similar functions, and we can be highly confident when sharing their functional annotation information.

Determining syntenic genes and genomic regions among closely related species is important to both gene and genome studies (Lyons et al., 2008). On one hand, with a well-defined synteny relationship, we can share gene annotation information between well-studied genomes (such as the model species *A. thaliana*) and newly annotated genomes (such as *B. rapa*), and investigate the syntenic gene's differentiation after the species' divergence. On the other hand, syntenic fragments shared among genomes of different species offer us a chance to deduce the evolutionary processes of related species, or could even provide clues to the mechanisms behind these processes.

Three plant species, *B. rapa*, *B. juncea*, and *B. oleracea*, (also called genomes A, B, and C) and their allopolyploidies form the famous U's triangle (U, 1935). Many of these are important vegetable or oil crop species, and the sequencing of more genomes from these species is being planned. The A genome sequence

of *B. rapa* has been released (Wang et al., 2011), and compared to three other sequenced species from Brassicaceae (*A. thaliana*, *A. lyrata*, and *T. parvula*), *B. rapa* has experienced an extra genome triplication event after their divergence (Wang et al., 2011; Cheng et al., 2012).

To share gene annotations between *B. rapa* and the other annotated genomes specially for *A. thaliana* and dig for clues of the *B. rapa* genome evolution, we developed a tool, SynOrths, which is well suited for detecting syntenic orthologs between closely related species, and applied it to the synteny analysis between *B. rapa* and other sequenced genomes from Brassicaceae: the model plant *A. thaliana*, *A. lyrata*, and *T. parvula*. The datasets generated in this study will help the gene annotation and further genomic evolution analysis of these species.

RESULTS

DEVELOPMENT OF THE SYNORTHS TOOL TO IDENTIFY GENE PAIRS WITH SYNTENIC RELATIONSHIPS

A tool named SynOrths was developed to identify syntenic genes based on the protein sequences of *B. rapa* and other related species (http://brassicadb.org/brad/tools/SynOrths/). As shown in **Figure 1**, SynOrths determines two genes to be syntenic orthologs based on both their own sequence similarity and the homology of their flanking genes.

There are four main steps embedded in SynOrths: (1) finding orthologous gene pairs; (2) redundant tandem gene removal; (3) locating potential syntenic orthologs by the support of flanking genes; and (4) final syntenic gene pair determination. In the first step, SynOrths runs Blastp to get basic protein sequence homology information from pairwise genomes. Gene pairs that are the best hits or with Blastp e-values <1E-20 are selected for further analysis. For tandem duplicated genes, which would add complexity to syntenic gene finding, SynOrths keeps one gene from each tandem gene array as a representative. In the second step, we identify all tandem gene arrays across the genomes being compared. Each tandem array is composed of continuously distributed homologous genes (Blastp E-value <1E-20) and should not be interrupted by more than one non-homologous gene. After that, the genes of each tandem array are replaced by the first one in the corresponding tandem. The revised homologous gene pairs are then sent to step 3 to compute the supporting strength of the flanking genes. Here, we set a threshold to check if the gene pair in question is supported by their flanking genes and thus potentially syntenic. Genes located in both flanking regions of the two genes are selected and named as the flanking gene set. We then count the number of best hit genes between the pairwise genomes in the flanking gene set. If the ratio of the best hit genes is higher than the threshold, then

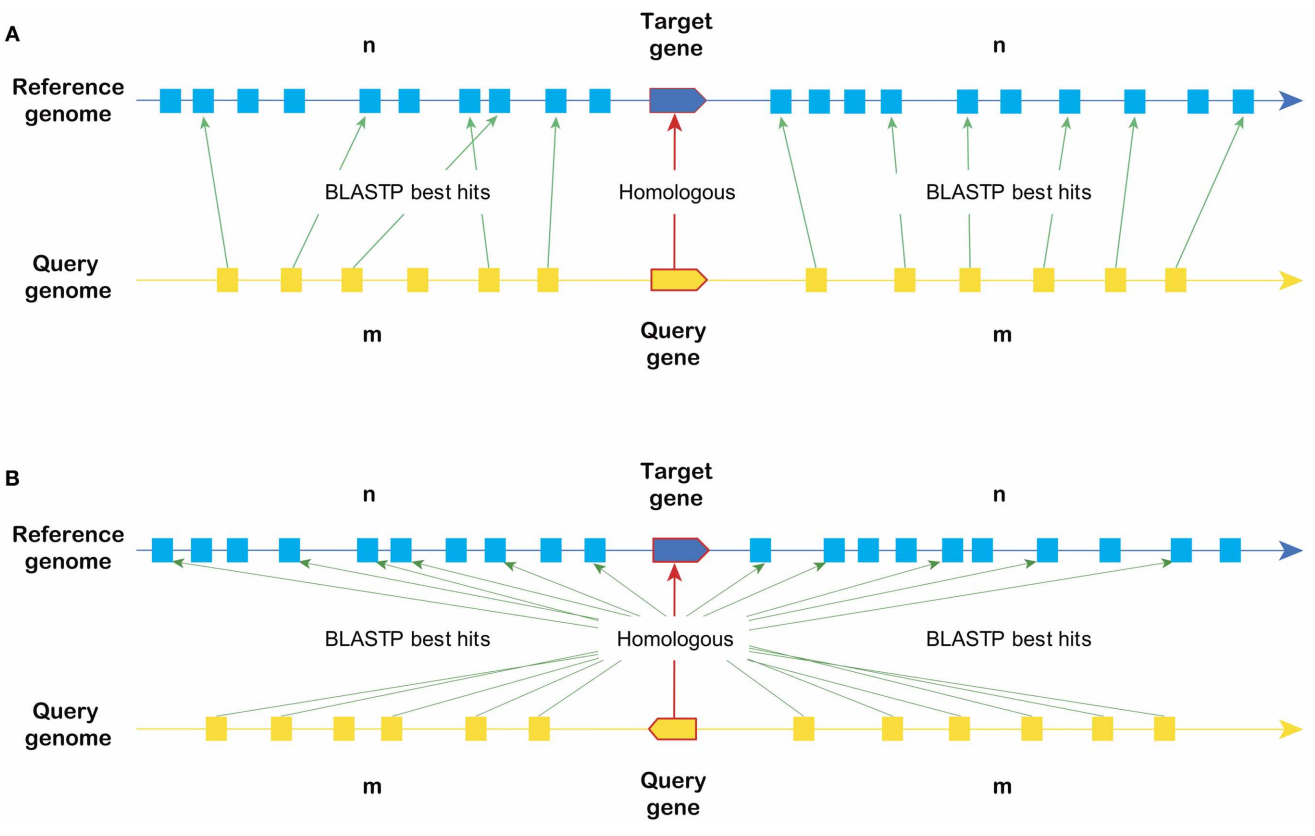


FIGURE 1 | The principles of syntenic gene identification in SynOrths. When determining whether two genes are under synteny, both the sequence homology of the two genes themselves and their flanking genes are considered. **(A)** Syntenic genes in the same direction in each genome. **(B)** Two syntenic genes located in inverted syntenic fragments.

the homologous gene pair is considered as potentially syntenic. In the fourth step, we further screen for the best syntenic gene pairs. After the first three steps, a certain gene might have more than one potential syntenic partner, so these candidate syntenic gene pairs are then compared based on the ratio of flanking gene support and their sequence homology. The gene pair with the highest supporting ratio of flanking genes and comparably higher sequence homology is finally determined as the best syntenic pair.

There are three main parameters that should be considered when using SynOrths. They are NumQ, the number of flanking genes on each side of the query gene; NumR, the number of flanking genes on each side of the reference gene; RatioQR, the ratio of best hit pairs among these flanking genes. Because *B. rapa* experienced whole genome triplication and subsequent intensive gene loss, the parameters should be carefully selected when using SynOrths to determine its syntenic genes with other species. Here, we chose three arrays for the three parameters (NumQ [5, 20, 60, 100], NumR [10, 40, 100, 150], and RatioQR [0.1, 0.2, 0.4, 0.8]) to perform SynOrths analysis between *B. rapa* and *A. thaliana*. As shown in **Figure 2**, the parameters NumQ = 20, NumR = 100, RatioQR = 0.2 gave a

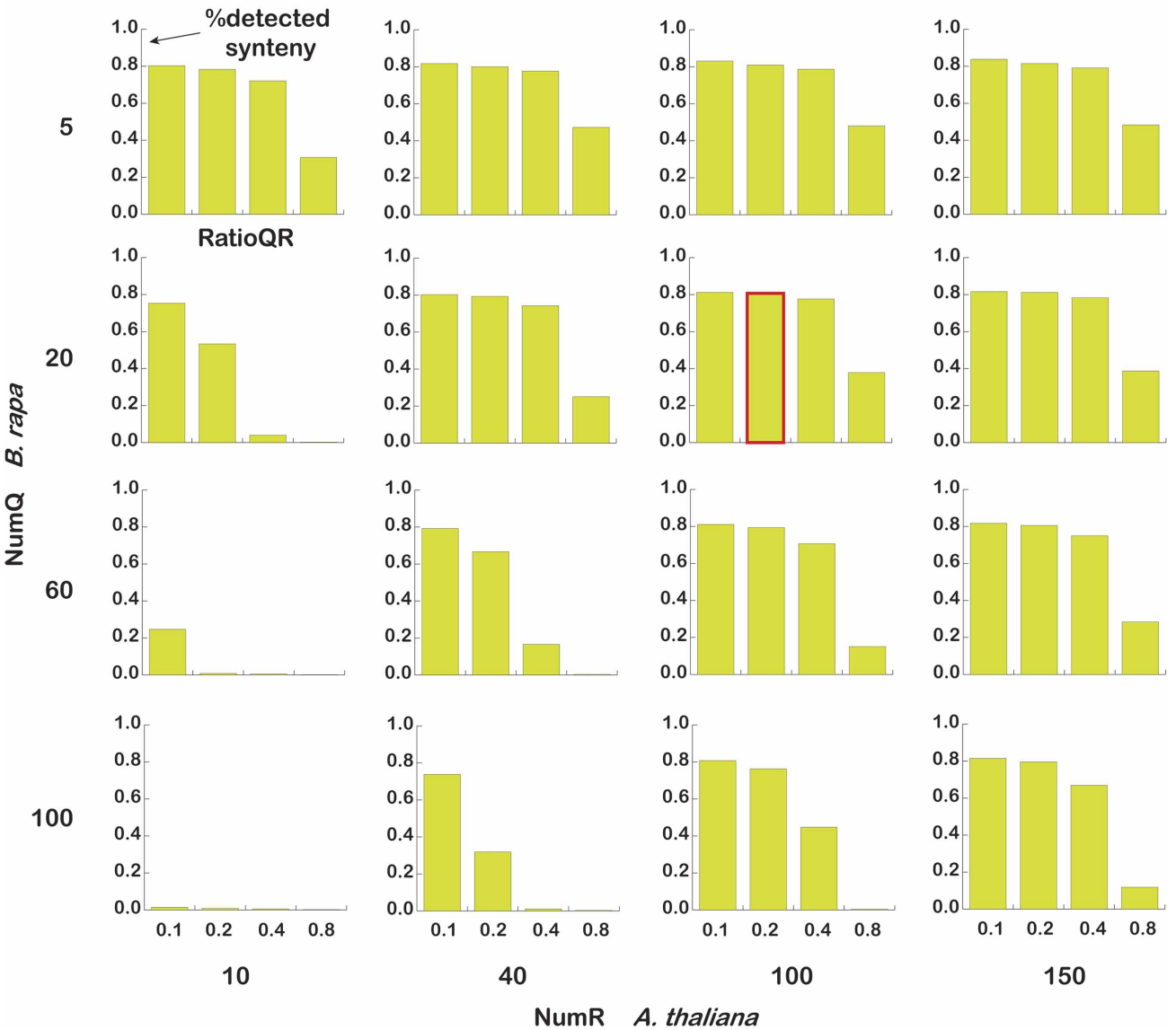


FIGURE 2 | Parameter estimation in SynOrths. The number of query (*B. rapa*) flanking genes [5, 20, 60, 100], the number of reference (*A. thaliana*) flanking genes [10, 40, 100, 150], and the threshold of the flanking genes' support ratio [0.1, 0.2, 0.4, 0.8] were set to run SynOrths. The bars indicate the proportions of syntenic genes identified out of 38,161 *B. rapa* genes, "%detected synteny" means percent of identified syntenic genes to the 38,161 *B. rapa* genes. The bar with a red border is the run with parameters NumQ = 20, NumR = 100, and RatioQR = 0.2; SynOrths returned stable and relatively more syntenic genes under these parameters.

considerably better result. This parameter set was then chosen to identify syntenic gene pairs between *B. rapa* and *A. thaliana*, *A. lyrata*, and *T. parvula*.

SYNTENIC GENE DETERMINATION BETWEEN *B. rapa* AND EACH OF *A. thaliana*, *A. lyrata*, AND *T. parvula*

Syntenic gene pairs between *B. rapa* and *A. thaliana*, *B. rapa* and *A. lyrata*, and *B. rapa* and *T. parvula* were identified using SynOrths. There were a total of 41,174, 27,379, 33,410, and 28,910 annotated proteins for *B. rapa*, *A. thaliana*, *A. lyrata*, and *T. parvula*, respectively. After removing the redundancy of duplicated tandem genes (keeping one gene from each tandem array), 38,161, 24,939, 30,773, and 27,344 genes were left for syntenic gene determination (**Table 1**). *B. rapa* returned 30,615 genes syntenic to 18,410 genes in *A. thaliana*; 30,250 genes syntenic to 18,125 *A. lyrata* genes; and 29,473 genes syntenic to 17,303 *T. parvula* genes. *A. thaliana* had the highest syntenic gene ratio (80.1%) compared to *A. lyrata* (79.6%) and *T. parvula* (77.5%).

The genome triplication event in *B. rapa* was well supported (**Figure 3**), because many genes that were evenly distributed in genomes of *A. thaliana* (14.3%), *A. lyrata* (14.6%), and *T. parvula* (16.1%) had three syntenic copies in *B. rapa*. Additionally, among the 7,546 *B. rapa* genes that had no syntenic orthologs in *A. thaliana*, 849 were syntenic to genes in *A. lyrata*, and 1,416 syntenic to *T. parvulla* genes. In total, there were 32,310 *B. rapa* genes with at least one syntenic ortholog in either *A. thaliana*, *A. lyrata*, or *T. pavula*, and only 5,851 *B. rapa* genes that had no syntenic counterparts in any of the other species.

For these non-syntenic genes between *B. rapa* and the other species, we considered them non-syntenic orthologs if their similarity satisfied sequence identity >70%, and coverage for each of the two genes >60%. *B. rapa* returned 1,391, 1,226, and 1,605 non-syntenic orthologs to 2,561 genes in *A. thaliana*, 1,877 in *A. lyrata*, 3,909 in *T. parvula*, respectively. However, for the 5,851 genes that had no syntenic orthologs in all three species, only

Table 1 | The homologous relationships of genes between *B. rapa* and *A. thaliana*, *A. lyrata*, or *T. parvulla*.

	Total genes	Tandem redundancy removed	Syntenic orthologs to *B. rapa*[a]	Non-syntenic orthologs to *B. rapa*[b]	Non-orthologs
B. rapa	41,174	38,161	N	N	N
A. thaliana	27,379	24,939	18,410/30,615	2,561/1,391	3,968/6,155
A. lyrata	33,410	30,773	18,125/30,250	1,877/1,226	10,771/6,685
T. parvulla	28,901	27,344	17,303/29,473	3,909/1,605	6,132/7,083
At + *Al* + *Tp*	N	N	N/32,310	N/808	N/5,043

[a] *Numbers left of the '/' indicate gene numbers in At, Al, or Tp; numbers right of the '/' represent the gene numbers in Br.*

[b] *Non-syntenic orthologs were defined as gene pairs with sequence identity >70% and coverage >60%.*

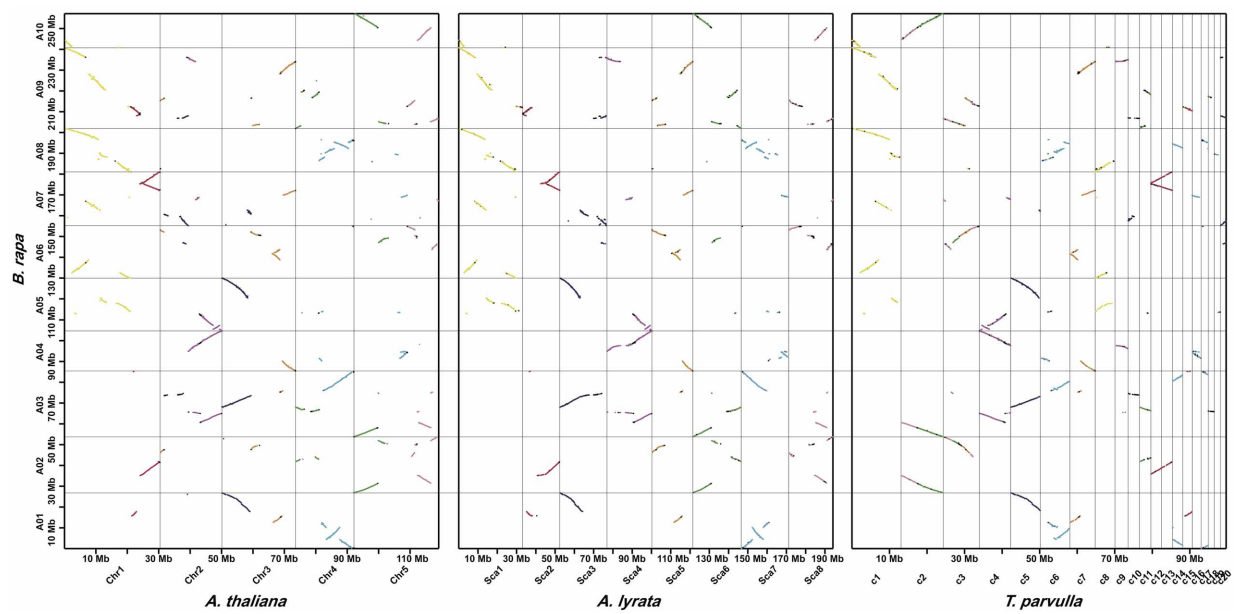


FIGURE 3 | Syntenic genes identified by SynOrths between *B. rapa* and *A. thaliana*, *A. lyrata*, or *T. parvula*. For each segment in *A. thaliana*, *A. lyrata*, or *T. parvula*, there were three syntenic copies observed in *B. rapa*, which clearly reflected the genome triplication experienced by *B. rapa*. Colors of the dots represent for the 24 ancestral blocks of Brassicaceae species, which has been defined previously (Schranz et al., 2006).

Table 2 | Syntenic tandem genes between *B. rapa* and *A. thaliana*, *A. lyrata*, or *T. parvulla*.

	Tandem (#arrays|#genes)	Syntenic tandem to *B. rapa*[a]
B. rapa	2,137|5,150	N
A. thaliana	1,569|4,009	1,223|3,157/1,649|4,033
A. lyrata	1,751|4,388	1,204|3,098/1,751|4,267
T. parvulla	1,135|2,692	857|2,071/1,689|4,140
At + Al + Tp	N|N	N|N/1,864|4,542

[a] *Numbers left of the '/' indicate tandems in At, Al, or Tp; numbers right of the '/' represent tandem numbers in Br.*

808 genes were non-syntenic orthologs of at least one gene in the other three species. These 808 genes could have been generated by gene transposition in *B. rapa* after its divergence from *A. thaliana*, *A. lyrata*, and *T. parvula*.

Most of the tandem arrays in *B. rapa* showed a syntenic relationship to *A. thaliana*, *A. lyrata*, or *T. parvula* (**Table 2**). For all 2,137 tandem arrays in *B. rapa*, 1,649 (77.16%) were syntenic to *A. thaliana*; 1,751 (81.94%) syntenic to *A. lyrata*, and 1,689 (79.04%) syntenic to *T. parvula*. In total, 1,864 (87.23%) tandem arrays in *B. rapa* had syntenic counterparts in at least one of the other species.

The dataset of genes' syntenic relationship among above four species had been integrated into BRAD (Brassica Database, http://brassicadb.org/brad/searchBrMultiSynteny.php) (Cheng et al., 2011). This resource built bridges between model plant *A. thaliana* and other Brassicaceae species, so the information of genes' function studies in *A. thaliana* were linked to the newly sequenced and annotated Brassicaceae genomes. For crop species such as *B. rapa, B. oleracea*, and *B. napus*, with the resource of syntenic relationships we can rapidly transfer knowledges from *A. thaliana* to the breeding research and apllication, and further production of the crops.

DISCUSSION

Because SynOrths determines syntenic orthologs and genomic regions based on protein sequences, it cannot be applied to genomes without protein annotation. Coding regions are one of the most conserved elements in the genome, as most active genes are functionally conserved to defend against non-synonymous mutations. Even after long periods of divergent evolution, proteins with the same function still show high levels of sequence homology. Thus, proteins that preserve ancestral traits provide good data for synteny analysis. After the determination of syntenic orthologs, syntenic gene pairs could be used as anchors to identify genomic regions that are in synteny. Conversely, most non-coding genomic regions (except conserved non-coding DNA) are under almost neutral selection. These regions can be rapidly changed in the evolutionary process, even in true syntenic regions non-coding sequences almost always show complete non-homology. SynOrths is not suited for synteny analysis of genomes without protein sets, but methods that include a step to search for conserved non-coding elements for synteny analysis might handle this problem.

The performance of SynOrths is impacted by gene annotation; the more accurate the gene annotation sets, the more accurately syntenic orthologs would be identified. However, the impact is acceptable and can even be very limited when we tune down the parameters to deny gene pairs to be in synteny. Because the annotation of different genomes is completed by different software with different parameters, the robustness of synteny analysis is very important. Furthermore, where the annotation of different genomes is unbalanced, we cannot draw genetic relationships based on the number or ratio of identified syntenic genes. Syntenic gene sets could be used to address whole genome genetic variation patterns, but for research issues focusing on individual genes, more detailed work should be done to confirm the results, or the parameters should be turned up to generate more stringent results.

Synteny analysis is one of the most important fields in comparative genome analysis as it is the basis of evolutionary studies at both the gene and genome levels. Most practically, it helps improve the gene annotation of newly sequenced genomes. In this study, we used synteny analysis to link bulk gene information from the model plant *A. thaliana* to the newly sequenced Brassicaceae genomes *B. rapa*, *A. lyrata*, and *T. parvula*.

We designed and developed a direct and efficient tool, SynOrths, to identify pairwise syntenic genes between *B. rapa* and other sequenced genomes of Brassicaceae species. With this tool we conducted syntenic gene analysis between *B. rapa* and *A. thaliana*, *A. lyrata*, and *T. parvula*, and generated valuable datasets for further analysis. Whole genome synteny analysis and the datasets generated are important for both gene annotation of the newly sequenced Brassicaceae genomes and the study of genomic and chromosomal evolution in Brassicaceae species.

MATERIALS AND METHODS

Gene annotation and protein sequences of *B. rapa* version 1.2 were downloaded from BRAD (http://brassicadb.org) (Cheng et al., 2011). The *A. thaliana* dataset was retrieved from TAIR (http://www.arabidopsis.org/index.jsp), and version TAIR9 was used. Genome and gene annotation files of version 6 for *A. lyrata* were downloaded from the JGI database (http://genome.jgi-psf.org/Araly1/Araly1.home.html) (Hu et al., 2011). *T. parvula* datasets version 7 were obtained from the *T. parvula* genome sequencing group (Dassanayake et al., 2011).

Blastp (NCBI-BLAST package, ftp://ftp.ncbi.nih.gov/blast/) was used for protein sequence alignment in SynOrths. SynOrths now is freely available at http://brassicadb.org/brad/tools/SynOrths/. The data generated in this study can be searched through http://brassicadb.org/brad/searchBrMultiSynteny.php or by contacting Feng Cheng.

AUTHOR CONTRIBUTIONS

Xiaowu Wang and Feng Cheng conceived the study. Feng Cheng developed SynOrths and performed the synteny analysis. Feng Cheng prepared the manuscript, Xiaowu Wang and Jian Wu improved the manuscript. Jian Wu and Lu Fang tested SynOrths

and provided feedback. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Dong-Ha Oh for his help in retrieving the *T. parvula* dataset. The work was funded by National Program on Key Basic Research Projects (The 973 Program: 2012CB113900), and the National High Technology R&D Program of China (2012AA100201) to Xiaowu Wang. The work was done in the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, P. R. China.

REFERENCES

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., and Wang, X. (2011). BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biol.* 11, 136.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of *Brassica rapa*. *PLoS ONE* 7:e36442. doi: 10.1371/journal.pone.0036442

Dassanayake, M., Oh, D. H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D. J., Bressan, R. A., Zhu, J. K., Bohnert, H. J., and Cheeseman, J. M. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* 43, 913–918.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y. L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., and Freeling, M. (2008). Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 148, 1772–1781.

Mandakova, T., and Lysak, M. A. (2008). Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (Brassicaceae). *Plant Cell* 20, 2559–2570.

Schranz, M. E., Lysak, M. A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* 11, 535–542.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488.

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., Wang, X., Freeling, M., and Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* 190, 1563–1574.

U. N. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of B. napus and peculiar mode of fertilication. *Jpn. J. Bot.* 7, 389–452.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J. H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Yu, J., Meng, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Z., Li, Z., Xiong, Z., and Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 July 2012; paper pending published: 07 August 2012; accepted: 08 August 2012; published online: 30 August 2012.

*Citation: Cheng F, Wu J, Fang L and Wang X (2012) Syntenic gene analysis between Brassica rapa and other Brassicaceae species. Front. Plant Sci. **3**:198. doi: 10.3389/fpls.2012.00198*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 29 August 2012
doi: 10.3389/fpls.2012.00192

A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method

Klas Hatje and Martin Kollmar*

Abteilung NMR-Basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Göttingen, Germany

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Ruiqiang Li, Peking University, China
Bernhard Haubold, Max-Planck-Society, Germany

****Correspondence:***
Martin Kollmar, Abteilung NMR-Basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany.
e-mail: mako@nmr.mpibpc.mpg.de

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes. The alignment of whole genome sequences of higher eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments. To overcome these limitations, we here used an alignment-free method to compare genomes of the Brassicales clade. For each nucleotide sequence a Chaos Game Representation (CGR) can be computed, which represents each nucleotide of the sequence as a point in a square defined by the four nucleotides as vertices. Each CGR is therefore a unique fingerprint of the underlying sequence. If the CGRs are divided by grid lines each grid square denotes the occurrence of oligonucleotides of a specific length in the sequence (Frequency Chaos Game Representation, FCGR). Here, we used distance measures between FCGRs to infer phylogenetic trees of Brassicales species. Three types of data were analyzed because of their different characteristics: (A) Whole genome assemblies as far as available for species belonging to the Malvidae taxon. (B) EST data of species of the Brassicales clade. (C) Mitochondrial genomes of the Rosids branch, a supergroup of the Malvidae. The trees reconstructed based on the Euclidean distance method are in general agreement with single gene trees. The Fitch–Margoliash and Neighbor joining algorithms resulted in similar to identical trees. Here, for the first time we have applied the bootstrap re-sampling concept to trees based on FCGRs to determine the support of the branchings. FCGRs have the advantage that they are fast to calculate, and can be used as additional information to alignment based data and morphological characteristics to improve the phylogenetic classification of species in ambiguous cases.

Keywords: Chaos game representation, Brassicales, *Brassica rapa*, phylogenetic tree, bootstrap re-sampling, frequency Chaos game representation

INTRODUCTION

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes, which assume the conservation and contiguity over the total sample length between homologous sequences (Blair and Murphy, 2011). The alignment of whole genome sequences of eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments (Dewey, 2012). In particular, genetic recombination and shuffling during species evolution complicate whole genome alignments limiting species genome versus single gene, multiple gene, or transcriptome comparisons. However, it would be beneficial for the significance of the species trees, if also whole genome assembly data were taken into account. In the past two decades several methods have been suggested for alignment-free sequence analyses that mainly group into word (oligomer) frequency methods and methods that do not resolve the fixed word-length distance measures and are thus absolutely independent from the assumption of conservation and contiguity (reviewed in Vinga and Almeida, 2003). The latter category includes the Chaos Theory (Jeffrey, 1990) and the theoretical concept of Kolmogorov complexity (Li et al., 2001). More recent methods include the alignment-free estimation of the number of substitutions per site (Domazet-Loso and Haubold, 2009) and feature frequency profiles (Sims et al., 2009).

The Chaos Game Representation (CGR) denotes an algorithm, which produces fractal pictures and can be adapted to reveal patterns in DNA (Li et al., 2001) and even protein sequences (Basu et al., 1997; Pleissner et al., 1997). These CGR pictures exhibit the fractal property that the overall pattern of the CGR picture is repeated in smaller parts of the picture. It has been shown that this self-similarity even holds for whole genome sequences and its sub-sequences, like single chromosomes, contigs, or genes (Deschavanne et al., 1999; Almeida et al., 2001; Joseph and Sasikumar, 2006). Commonly, the pictures of DNA sequences are generated as squares such that the lower (A + T) and the upper (C + G) halves indicate the base composition and the diagonals the purine/pyrimidine composition. CGRs are unique descriptions of each DNA sequence and, in the case of whole genome sequences, can therefore be regarded as genomic fingerprints. However, the CGRs are not directly comparable. If the CGR pictures are divided into smaller squares by grid lines, each grid square

represents the frequencies of the respective oligonucleotides as found in the whole sequence (Deschavanne et al., 1999; Almeida et al., 2001). These frequencies can be represented in Frequency Chaos Game Representation (FCGR) pictures with a gray scale to express the number of points within each grid square and with pictures for each length k oligonucleotide (with $k = 1, 2, 3\ldots$). FCGRs are numerical matrices and can be used to infer phylogenetic trees based on distance methods (Wang et al., 2005). So far this approach has only been applied to reconstruct the phylogeny of 20 birds using nuclear genome data (Edwards et al., 2002), to analyze the mitochondrial genomes of 26 sample eukaryotes (Wang et al., 2005), and to sub-typing of HIV-I (Pandit and Sinha, 2010). One of the advantages of using FCGRs for phylogenetic reconstructions is that sequence, which cannot be aligned, can be used.

Here, we performed phylogenetic analyses based on three different types of data. Firstly we used the whole genomic sequence assemblies of all so far sequenced species in the taxon Malvidae, including that of *Brassica rapa*. Because a reference tree including all these species was not available we assembled and annotated all actin capping (CP) protein sequences (Cooper and Sept, 2008) and the sequences of the actin-related proteins Arp2 and Arp3 (Goley and Welch, 2006). These proteins are present in all eukaryotes and as single copies in the *Arabidopsis thaliana* genome. Thus they were not expected to exist in duplicates in the other analyzed species avoiding the ortholog-paralog problem. To infer the phylogeny of the different *Brassica* species, for which whole genome assemblies have not yet been produced, we used EST and mitochondrial genome DNA. The quality of the phylogenetic analyses depends on the resolution of the FCGRs (length of k) and thus on the length of the nucleotide sequences. Thus we only included those species for which a considerable number of EST clones were available. To estimate the support for the branchings, here, we apply the concept of bootstrap re-sampling to the comparison of FCGRs for the first time.

MATERIALS AND METHODS

DATA ACQUISITION

The genome files were retrieved from diArk[1] (Hammesfahr et al., 2011), and the mitochondrial genomes and EST reads from the NCBI database, each in FASTA format (**Table 1**). For the generation of the CGRs the contigs and reads of each dataset were concatenated. The whole genome assemblies as available from the sequencing centers contain both the nuclear and mitochondrial genomes, and potentially still some contaminations from other species' DNA. However, given the sizes of the whole genome datasets, the contributions of the mitochondrial genomes and contaminating DNA to the FCGRs are negligible. The FCGRs of the whole genome data can thus be regarded as identical to the FCGRs of the nuclear genomes.

IMPLEMENTATION OF THE ALGORITHM

The algorithm to calculate CGRs and FCGRs was implemented in C/C++. CGR positions were generated as lists in plain text and plotted for graphical presentations in the Scalable Vector Graphics (SVG) format[2]. Based on the CGR position values, FCGRs were calculated for each k in 1, …, 8. Distance calculations were implemented in Ruby[3].

GENERATING CHAOS GAME REPRESENTATIONS

Chaos game representations of the nucleotide sequences were generated by the following algorithm. A 1×1 square is drawn and each vertex labeled by a nucleotide. In agreement with other analyses we placed C in the upper left, G in the upper right, A in the lower left, and T in the lower right vertex. The starting point is defined as the geometric center of the square at position (0.5, 0.5). The respective nucleotide sequences are then plotted sequentially. For the first nucleotide a point is plotted on half the distance between the starting point (0.5, 0.5) and the vertex corresponding to this nucleotide. Subsequently for each following nucleotide a point is placed as mid-point between the previously plotted point and the vertex corresponding to the nucleotide (**Figure 1A**).

The algorithm can be expressed by the following equations:

$$\mathrm{CGR}_0 = (0.5, 0.5) \tag{1}$$

$$\mathrm{CGR}_i = \begin{cases} \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (0.0, 0.0)) & \text{if } \mathrm{seq}_i = \text{'C'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (1.0, 0.0)) & \text{if } \mathrm{seq}_i = \text{'G'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (0.0, 1.0)) & \text{if } \mathrm{seq}_i = \text{'A'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (1.0, 1.0)) & \text{if } \mathrm{seq}_i = \text{'T'} \end{cases} \tag{2}$$

The resulting plot is unique for each sequence. The overall pattern of points is repeated in each sub-square of the plot (**Figure 1B**). In addition, each plot based on a sub-sequence of the whole sequence has a similar appearance. Thus similar sequences result in similar CGR plots. **Figure 1B** shows the CGR of the first 1,000,000 nt of the *B. rapa* genome sequence.

The calculation of the frequencies of points within each sub-square results in an FCGR. Thus each FCGR represents the occurrence of oligonucleotides in the whole sequence. For dinucleotides ($k = 2$) the binary square is divided into a 4×4 grid, for trinucleotides ($k = 3$) into an 8×8 grid, and in general into a $2^k \times 2^k$ grid. **Figure 1C** shows an FCGR ($k = 3$) of the whole *B. rapa* genome sequence.

If the nucleotide sequences differ in length, the resulting FCGRs will also differ in there overall frequencies. To overcome this sequence length bias each FCGR was standardized (Wang et al., 2005). If the FCGR is represented as for example a $2^k \times 2^k$ matrix, the matrix $A = (a)_{2^k \times 2^k}$ is transformed to a standardized FCGR as follows:

$$\bar{A} = \frac{4^k}{\sum_{i=1}^{k} \sum_{j=1}^{k} a_{i,j}} A \tag{3}$$

The nucleotide sequences of each data file (whole genome, EST, or mitochondrial genome data) were concatenated and the reverse

[1] http://www.diark.org

[2] http://www.w3.org/Graphics/SVG

[3] http://ruby-lang.org

Table 1 | List of the species used in the analysis.

Species	Whole genome			EST		Mitochondrial genome		
	Contigs	Nucleotides	Accession numbers	Reads	Nucleotides	Contigs	Nucleotides	Accession numbers
Arabidopsis lyrata	695	206667935	GL348713–GL349407					
Arabidopsis thaliana	5	119145879	NC_003070–NC_003071, NC_003074–NC_003076	1529700	400512451	1	366924	NC_001284
Brassica rapa	51658	273071614	AENI01000001–AENI01051658	213605	122970377	1	219747	NC_016125
Capsella rubella	853	134834574						
Carica papaya	3207	331271729	DS981520–DS984726	77393	54789864			
Citrus clementina	1128	295550349						
Citrus sinensis	12574	319231331						
Eucalyptus camaldulensis	274001	654922307	DF097775–DF126446					
Eucalyptus grandis	4952	691297852						
Eutrema halophilum	639	243117811		38022	20080214			
Eutrema parvulum	7	114396853	CM001187–CM001193					
Gossypium raimondii	1448	763818933						
Theobroma cacao	1782	351351221	FR720657–FR725448					
Vitis vinifera	33	486265422	FN597015–FN597047	446643	284204927	1	773279	NC_012119
Brassica napus				643437	381399492	1	221853	NC_008285
Brassica oleracea				179150	125257248	1	360271	NC_016118
Limnanthes alba				15331	8582959			
Raphanus raphanistrum				164119	104536170			
Raphanus sativus				150680	97973638			
Tropaeolum majus				10507	6436290			
Brassica carinata						1	232241	NC_016120
Brassica juncea						1	219766	NC_016123
Lotus japonicus						1	380861	NC_016743
Millettia pinnata						1	425718	NC_016742
Ricinus communis						1	502773	NC_015141

The number of contigs/reads and the number of nucleotides for the whole genome, EST, and mitochondrial genome data files are given. In addition, for whole genome and mitochondrial genome data the NCBI accession numbers are given if available.

complement of the concatenated sequence was appended. Characters other than "C," "G," "A," or "T" were ignored. Some example FCGRs generated with $k = 8$ are shown in **Figures 1D–L**. Already by visual inspection it is obvious, that whole genome, EST, and mitochondrial genome FCGRs have distinct patterns (**Figures 1D–F**), while the FCGRs generated from the same data type of closely related species are very similar (**Figures 1G–L**). EST data disproportionately contain poly-A sequences, resulting in unusually high frequency values in the FCGRs. These subsequently dominate the distance matrix calculation for higher order FCGRs ($k > 5$) and misdirect the calculation of the phylogenetic trees (data not shown). Therefore, in the case of EST data, the two entries in each FCGR that contain poly-A and poly-T stretches were set to zero.

DISTANCES

In order to reveal the phylogenetic relation between the analyzed species we calculated pair-wise distances between the FCGRs. In general all distances that are applicable to matrices could be used. The following distances have already been described for comparing FCGRs: The Hamming distance (Campbell et al., 1999; Wang et al., 2005), the Euclidean distance (Edwards et al., 2002; Vinga and Almeida, 2003; Wang et al., 2005; Pandit and Sinha, 2010), the Image distance defined in Wang et al. (2005), and the Pearson distance (Almeida et al., 2001; Vinga and Almeida, 2003; Wang et al., 2005). Here, we chose the Pearson distance as a statistical distance and the Euclidean distance as a geometrical distance, which performed best in a comparison of difference distance methods (Wang et al., 2005). The Euclidean distance between two points in two-dimensional space is defined as the length of the line segment between these two points and can be calculated using the Pythagorean equation. This concept can be adapted to calculate the distance between two FCGRs. The Euclidean distance between two standardized FCGRs $A = (a)_{2^k \times 2^k}$ and $B = (b)_{2^k \times 2^k}$ is defined as follows:

$$d_{\text{Euclidean}}\left(\bar{A}, \bar{B}\right) = \frac{\sqrt{2^k}}{4^k} \sqrt{\sum_{i=1}^{2^k} \sum_{j=1}^{2^k} \left(a_{i,j} - b_{i,j}\right)^2} \tag{4}$$

The Pearson distance is based on a weighted Pearson correlation coefficient (Almeida et al., 2001; Wang et al., 2005). To calculate the Pearson distance, the FCGRs are represented as lists of the

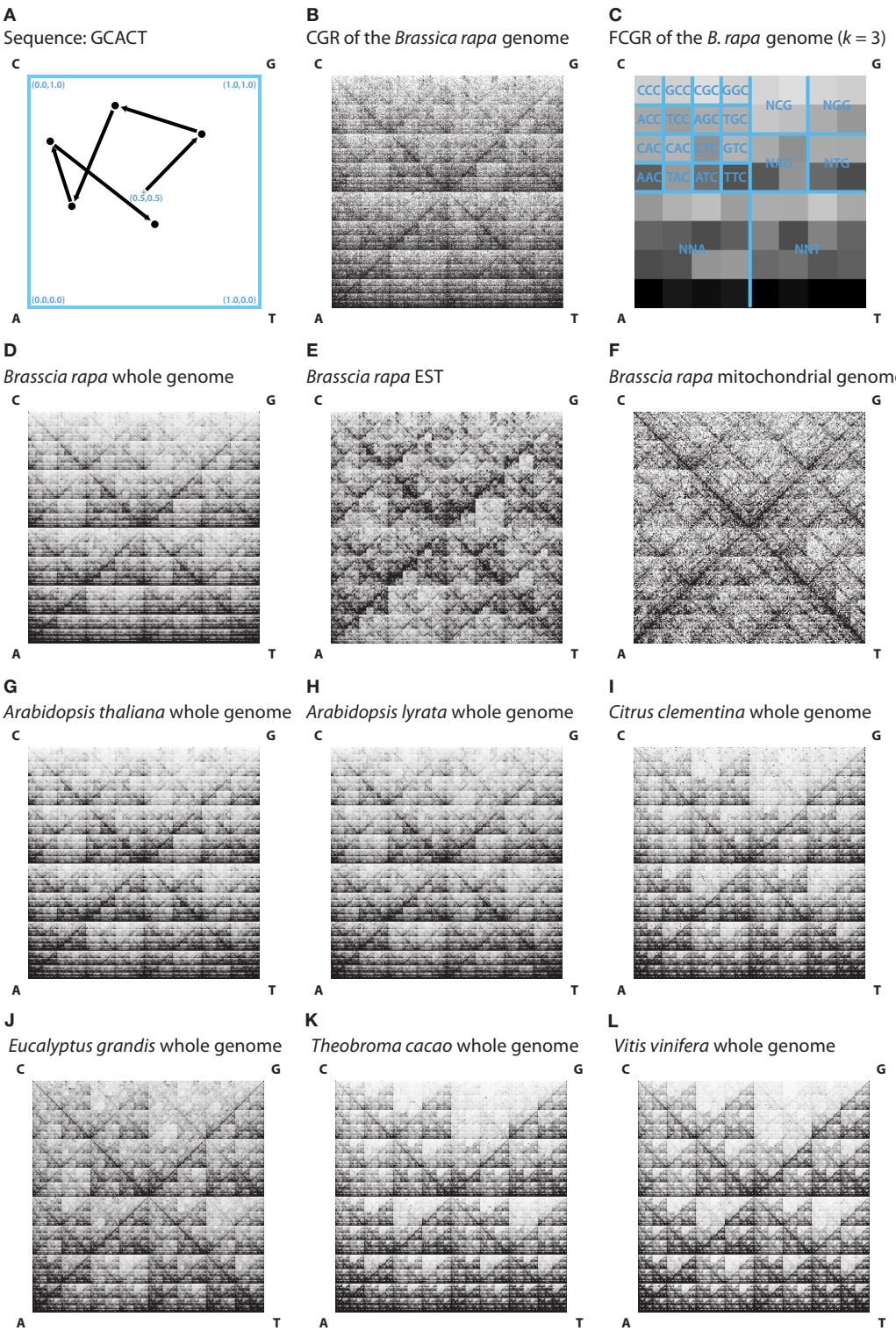


FIGURE 1 | (A) A Chaos game representation (CGR) image is generated by drawing a unit square and, starting at the center (0.5, 0.5), plotting for each nucleotide of the sequence a point on half the distance to the corresponding vertex. In this example the CGR for the sequence "GCACT" was drawn. **(B)** The image shows the CGR of the first 1,000,000 nt of the *Brassica rapa* genome. **(C)** The figure shows an FCGR ($k = 3$) of the whole *Brassica rapa* genome illustrating the frequencies of points in the CGR in an 8×8 grid. The squares of the grid represent the occurrence of specific trinucleotides, which are labeled in the figure. In **(D–L)** the FCGRs ($k = 8$) of the whole genome **(D)**, EST **(E)** and mitochondrial genome sequences **(F)** of *Brassica rapa* and the FCGRs ($k = 8$) of the whole genome sequences of some representatives of the different clades **(G–L)** are shown for visual comparison.

frequencies with $n = 4^k$ values. The Pearson distance between the non-standardized FCGRs $A = (x_1, \ldots, x_n)$ and $B = (y_1, \ldots, y_n)$ is defined as follows:

$$nw = \sum_{i=1}^{n} x_i \cdot y_i$$

$$\bar{x}w = \frac{\sum_{i=1}^{n} x_i^2 \cdot y_i}{nw}, \quad \bar{y}w = \frac{\sum_{i=1}^{n} y_i^2 \cdot x_i}{nw},$$

$$sx = \frac{\sum_{i=1}^{n} (x_i - \bar{x}w)^2 \cdot x_i \cdot y_i}{nw}, \quad sy = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}w\right)^2 \cdot x_i \cdot y_i}{nw}$$

$$d_{\text{Pearson}} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i - \bar{x}w}{\sqrt{sx}} \cdot \frac{y_i - \bar{y}w}{\sqrt{sy}} \cdot x_i \cdot y_i}{nw} \quad (5)$$

GENERATING PHYLOGENETIC TREES

To generate the phylogenetic trees, pair-wise distance matrices were calculated for each k in $1, \ldots, 0.8$ with the Euclidean distance method as defined in Eq. 4 and the Pearson distance as defined in Eq. 5. The distance matrices were subjected to the Neighbor joining (NJ) and Fitch–Margoliash algorithms as implemented in the Phylip package[4]. Statistical support for branchings was obtained by applying the bootstrap re-sampling method. For each FCGR, 500 datasets were generated by random sampling with replacement. Based on these re-sampled FCGRs 500 phylogenetic trees were reconstructed for each k in $1, \ldots, 0.8$. The trees of each dataset were summarized to consensus trees using the *consense* program of the Phylip package. The topologies of the consensus trees were fixed and the branch lengths calculated with the Fitch–Margoliash algorithm. In the case of the NJ trees, a bootstrapped tree was chosen that had the same topology as the consensus tree and the bootstrap values were plotted onto this tree. The bootstrap values represent the percentage each interior branch has the same partition as the consensus tree.

GENERATION OF THE REFERENCE TREE FOR THE WHOLE GENOME ANALYSIS

For the reference tree of those species for which whole genome assemblies are available we identified, assembled, and annotated the sequences of the heterodimeric actin capping protein (CAP), α- and β-CAP, and the sequences of the actin-related proteins Arp2 and Arp3. The *B. rapa* and *Gossypium raimondii* genomes contain duplicates of these genes due to species-specific duplications. Therefore, only one of the duplicates had been used for the phylogenetic tree reconstructions. The CAP and Arp sequences were aligned, concatenated, and phylogenetic trees reconstructed using the NJ and the Maximum likelihood (ML) method. The NJ tree was unrooted and generated using ClustalW (Chenna et al., 2003) with standard settings and the Bootstrap (1,000 replicates) method. The ML tree was calculated using the JTT (Jones et al., 1992) substitution model as suggested by ProtTest (Darriba et al., 2011) with estimated proportion of invariable sites and bootstrapping (1,000 replicates) using RAxML (Stamatakis et al., 2008).

[4]http://evolution.genetics.washington.edu/phylip.html

RESULTS

Phylogenetic trees based on whole genome, mitochondrial genome, and EST data were generated using the Euclidean or Pearson distance methods in combination with the NJ or the Fitch–Margoliash tree reconstruction algorithms. In order to reveal the influence of the lengths of the oligonucleotides we report trees of FCGRs generated with $k = 3$ (trinucleotides, 64 data points) and $k = 8$ (octanucleotides, 65,536 data points).

INFLUENCE OF SEQUENCE LENGTHS ON THE PHYLOGENETIC TREES

First we tested whether different sequence lengths have an influence on the results (**Figure 2**). For the whole genome assemblies and the EST datasets, sub-sections of the sequences were generated with lengths of 10^6, 10^7, and 10^8 nt. For that purpose the contigs or EST entries of each organism were shuffled, concatenated, and subsequently the sub-sequences generated by cutting the sequences at the respective positions. In the case of the whole genome data (**Figure 2A**), the FCGRs of the whole genome assemblies and the sub-sequences of each organism grouped together forming clusters. The only exceptions were the shortest 10^6-nt sequences of *Citrus sinensis*, *Citrus clementina*, *Arabidopsis lyrata*, and *A. thaliana*, which group to different species. The FCGRs of the EST data group together for each species independently of the lengths of the sequences (**Figure 2B**). For the mitochondrial genomes datasets with shorter sequences of 10^4 and 10^5 nt were generated. Here the FCGRs of the 10^4 nt sequences do not cluster together with those of the longer sequences of the corresponding species. The FCGRs of the mitochondrial sequences have been calculated based on hexanucleotides ($k = 64{,}096$ data points). Here, $k = 6$ was chosen, because in the case of higher k values ($k = 7$ or $k = 8$), the sequence length of the shortest sequences (10^4 nt) would be less than the number of data points in the FCGRs. In the shortest sequences (10^4 nt) many of the hexanucleotides are not covered at all resulting in many zero values for frequency positions, which lead to the unusual grouping of these FCGRs.

WHOLE GENOME ANALYSIS

In order to analyze the phylogenetic grouping of *B. rapa* in a whole genome context we searched for closely related plant species, for which whole genome assemblies are available. According to diArk (Hammesfahr et al., 2011), that comprises the most reliable and complete compilation of eukaryotic genome projects for which genome assemblies are available, the genomes of 13 different species (excluding different *A. thaliana* strains) of the taxon Malvidae have been sequenced and assembled: *A. lyrata* (Hu et al., 2011), *A. thaliana* (thale cress; Arabidopsis Genome Initiative, 2000), *B. rapa* subsp. *pekinensis* (Chinese cabbage; Wang et al., 2011), *Capsella rubella*, *Carica papaya* (Ming et al., 2008), *C. clementina*, *C. sinensis* (sweet orange), *Eucalyptus camaldulensis* (Murray red gum), *Eucalyptus grandis* (Flooded gum), *Eutrema halophilum* (salt cress), *Eutrema parvulum* (Dassanayake et al., 2011), *G. raimondii*, and *Theobroma cacao* (cacao plant; Argout et al., 2011). In addition, the genome of *Vitis vinifera* (grape vine; Jaillon et al., 2007; Velasco et al., 2007) was chosen as outgroup

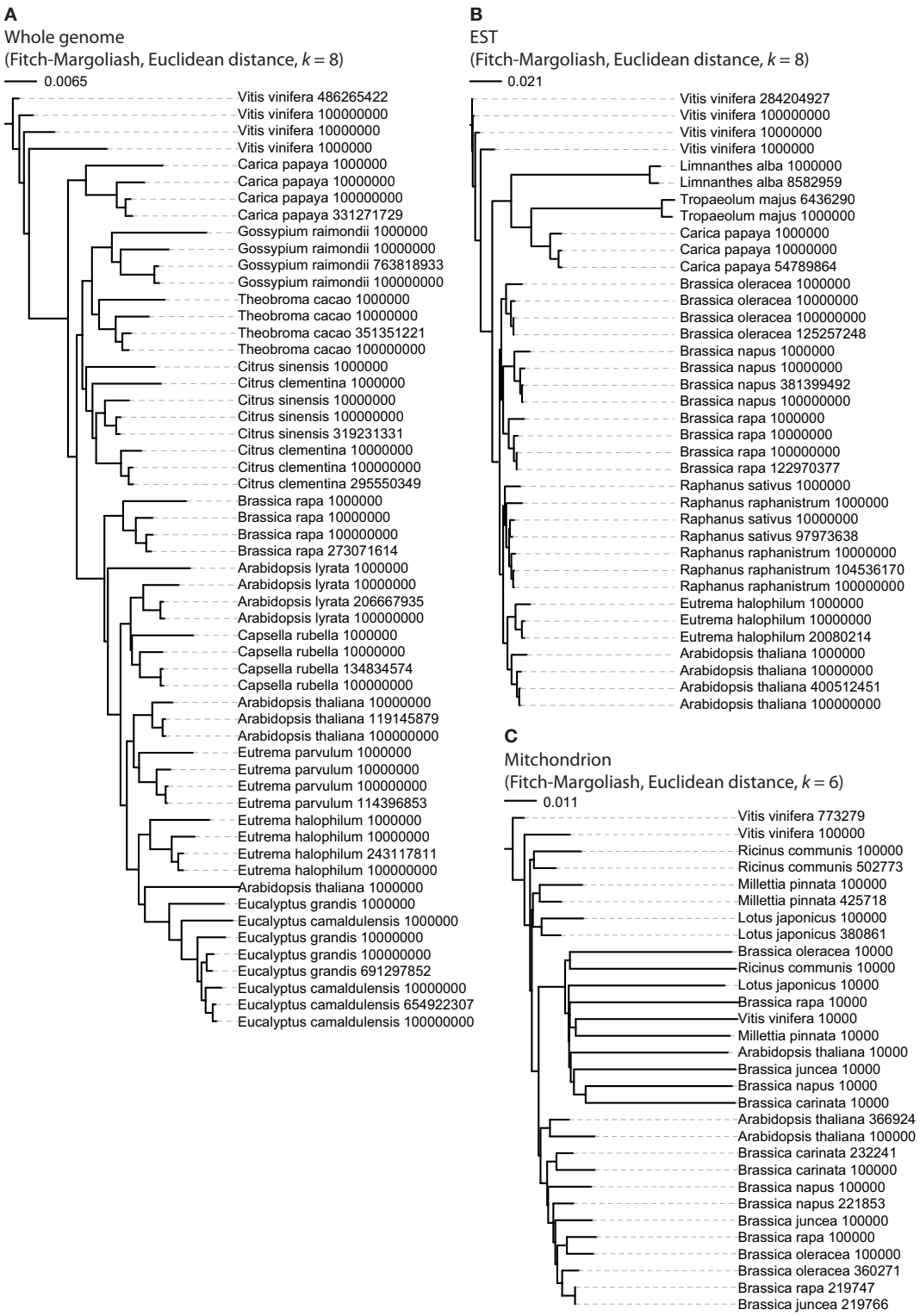


FIGURE 2 | Phylogenetic trees to reveal the potential influence of sequence length. For each dataset sub-sequences with defined lengths were generated and FCGRs calculated. The lengths of the sequences were supposed to be sufficient for reliable tree reconstructions if the datasets generated from the same species grouped together. For whole genome **(A)** and EST **(B)** data 10,000,000 nt should be sufficient while the full-length mitochondrial genomes **(C)** are needed for reliable tree reconstructions.

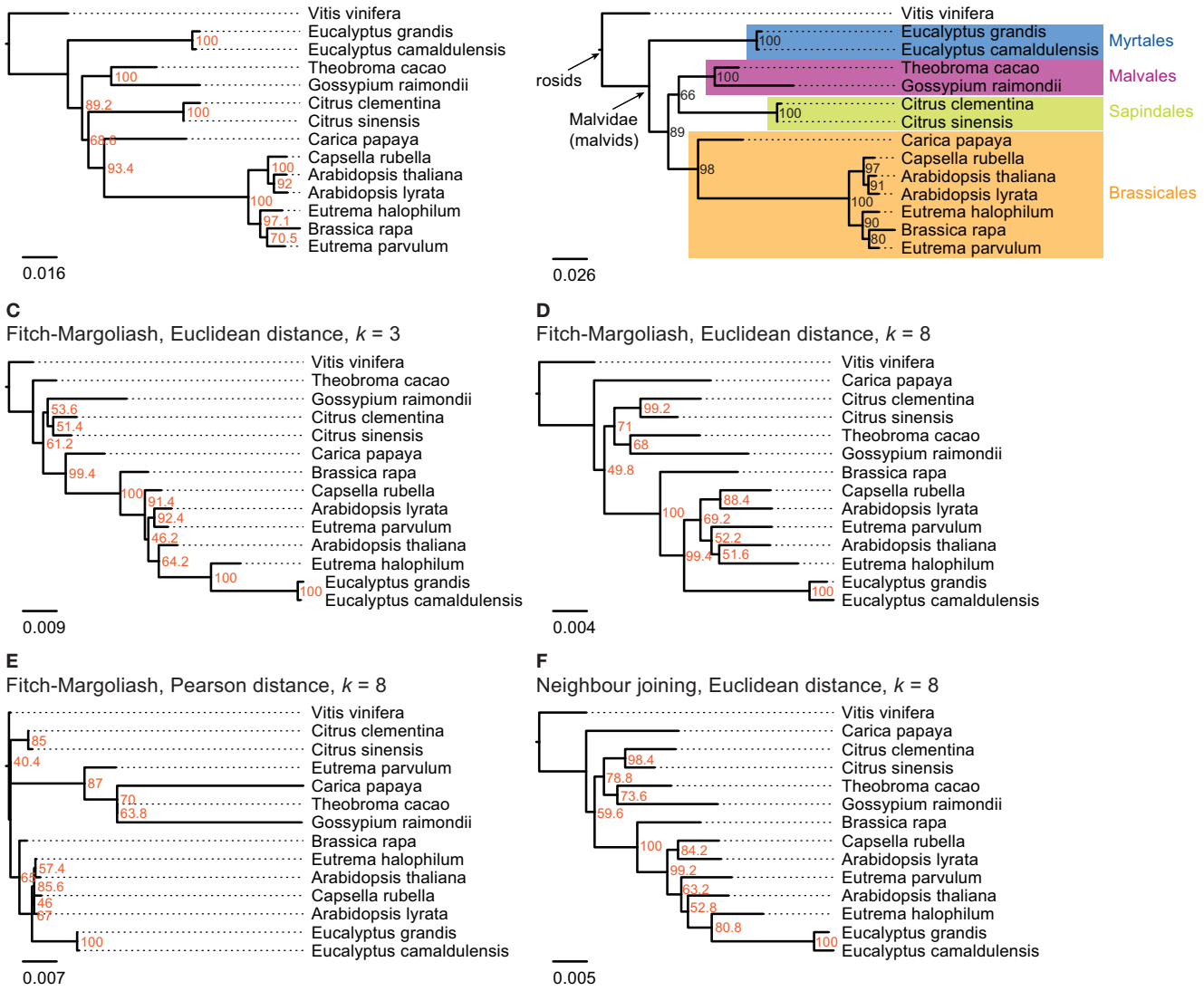


FIGURE 3 | The trees in (A,B) are based on a multiple sequence alignment of manually assembled CAP and Arp2/3 protein sequences. The trees were calculated using the Neighbor joining and the Maximum likelihood method, respectively, with 1,000 bootstraps for each tree. In **(C–F)** phylogenetic trees were generated applying different methods on FCGRs of whole genome sequence data of species of the taxon Malvidae. In **(C–E)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(F)** the Neighbor joining algorithm was used to calculate the tree.

to root the trees. A species tree including all these organisms is not available. For comparison we therefore reconstructed trees of these species based on the alignment of the concatenated protein sequences of the actin CAP, Arp2, and Arp3 proteins (**Figures 3A,B**). The trees based on the NJ and ML methods are almost identical and differ only in the grouping of the two *Citrus* species (Sapindales clade) as independent clade (NJ, **Figure 3A**) or as sister clade of the Malvales (ML, **Figure 3B**). While the bootstrap support for all branchings is high, the support for the grouping of the *Citrus* clade is low in both trees (68.6% in the NJ and 66% in the ML tree, respectively). Both trees are in general agreement with phylogenetic analyses of the mitochondrial matR proteins (Zhu et al., 2007) and 61 chloroplast protein-coding genes (Bausher et al., 2006), and the combined analysis of 10 plastid and 2 nuclear (18S and 26S rDNA) genes (Cantino et al., 2007) that also show different groupings of the Sapindales clade. All trees agree with the grouping of the Malvales, Sapindales, and Brassicales into one clade and the grouping of the Myrtales as a sister clade, *C. papaya* being the most divergent of the analyzed Brassicales species and *C. rubella* being the closest relative of the *Arabidopsis* species. Except for the grouping of the two *Citrus* species the topology of the tree based on the ubiquitous

cytoskeletal proteins CAP and Arp2/3 can thus be regarded as reference.

The resulting phylogenetic trees of the FCGRs differ as a function of data and methods used (**Figures 3C–F**). We reconstructed two trees based on the Euclidean distance and the Fitch–Margoliash algorithm but based on FCGRs with different resolution ($k = 3$ and $k = 8$ in **Figures 3C,D**, respectively), a tree using a different method for the distance calculation, the Pearson distance (**Figure 3E**), and a tree by applying a different method for the tree reconstruction, the NJ method (**Figure 3F**). In general, the trees agree with the reference tree except for the *Eucalyptus* species, which are either placed as sister group to *E. halophilum* (**Figures 3C,F**) or at the base of the Brassicales (**Figures 3D,E**) and thus far from their position according to the reference tree. In addition, *T. cacao* in **Figure 2C**, *C. papaya* in **Figures 3D–F**, and *E. parvulum* in **Figure 2E** are in wrong positions. None of the combinations of methods and data resulted in a correct resolution of the very closely related *Arabidopsis*, *Eutrema*, and *Capsella* species.

The tree based on the Pearson distance method (**Figure 3E**) contains the most deviations from the reference tree and this method therefore seems to be the least appropriate for reconstructing phylogenetic trees of whole genome sequences. This observation is in accordance with Wang et al. (2005). In addition, the bootstrap values do not provide reasonable support for most of the branchings except for the monophyly of the *Citrus* and the *Eucalyptus* clades. The trees based on high-resolution FCGRs ($k = 8$) using the Euclidean distance method (**Figures 3D,F**) have identical topologies except for the *Eucalyptus* outliers. In both trees *C. papaya* is placed as closest species to *V. vinifera* and not at the base of the Brassicales, *A. thaliana* grouped to the *Eutrema* species instead to its closest relative *A. lyrata*, and *B. rapa* is found at the base of the Brassicales instead of grouping to the *Eutrema* species. However, the misplacement of *Carica* and *A. thaliana* is not well supported (bootstrap values of 50–60%). Thus, the considerably faster NJ algorithm is a good alternative to the Fitch–Margoliash algorithm if run time is important. In contrast, the phylogenetic tree based on the low-resolution FCGRs ($k = 3$) contains more differences compared to the reference tree (**Figure 3C**).

EST DATA ANALYSIS

For this analysis related species of *B. rapa* were chosen, for which more than 1,000 EST entries are available in the EST database of NCBI. There are ten species that belong to the Brassicales taxon and match this criteria: *A. thaliana*, *Brassica napus*, *Brassica oleracea*, *B. rapa*, *C. papaya*, *E. halophilum*, *Limnanthes alba*, *Raphanus raphanistrum*, *Raphanus sativus*, and *Tropaeolum majus* (**Table 1**). Again, *V. vinifera* was included as outgroup. The trees reconstructed from the FCGRs of the EST datasets are shown in **Figure 4**. The tree based on the Pearson distance and calculated with the Fitch–Margoliash algorithm (**Figure 4C**) shows many deviations from the known relationships of the species but also low support for the branchings. Like for the whole genome analysis, the Pearson distance concept is not appropriate for the reconstruction of reliable phylogenetic trees based on FCGRs. The trees based on the Euclidean distance (**Figures 4A–D**) have almost identical (low-resolution $k = 3$ compared to high-resolution data $k = 8$) to identical topologies (Fitch–Margoliash compared to NJ algorithm). Especially the species of the Brassicaceae clade are well resolved and their topology is highly supported in all trees. The Limnanthaceae, Tropaeolaceae, and Caricaceae are sister groups of the Brassicaceae. To our knowledge there is no highly resolved tree of these groups available that we could use as reference. Based on our experience with the whole genome data we suppose that the trees based on high-resolution data represent the more reliable topologies.

MITOCHONDRIAL GENOME ANALYSIS

For this analysis close relatives of *B. rapa* were chosen, for which sequenced mitochondria are available from NCBI. There were nine species in the Rosids taxon, whose mitochondrial genome sequences were available: *A. thaliana*, *Brassica carinata*, *Brassica juncea*, *B. napus*, *B. oleracea*, *B. rapa*, *Lotus japonicus*, *Millettia pinnata*, and *Ricinus communis* (**Table 1**). The mitochondrial genome of *V. vinifera* was used as outgroup. In contrast to the analyses of the other datasets, the trees based on the FCGRs of the mitochondrial genomes were very similar for the four different methods (**Figure 5**). Especially the sub-branches containing the five closely related *Brassica* species show exactly the same topology supported by high bootstrap values. While the topology of the Brassicales subfamily tree is well resolved the grouping of the Fabales *L. japonicus* and *M. pinnata* and the Malpighiales *R. communis*, which all belong to the fabids, is different in the four trees. Here, the trees based on the Euclidean distance with high-resolution FCGRs ($k = 8$) have the same well supported topology grouping the Fabales together (**Figures 5B,D**) independently which method has been used for the tree reconstruction. This is in agreement with the results from the whole genome and EST analysis that the use of FCGRs with high-resolution results in more reasonable trees, and that the Euclidean method for the calculations of the distances is more appropriate than the Pearson method.

COMPUTATIONAL RESOURCE COMPARISON

The algorithm to calculate the CGRs and FCGRs has linear time complexity $O(L)$ and space constant complexity $O(1)$, where L is the length of the nucleotide sequence. In the case of whole genomes, the calculation of the CGRs and FCGRs took about 7,600 s for each genome, for EST data 2,800 s for each species, and 140 s for each mitochondrial genome. The time the algorithm needs to calculate the phylogenetic trees mainly depends on the distance matrix calculated for each species against each other species. This calculation has time complexity $O(4^k s^2)$ and space complexity $O(s^2)$, where s is the number of species and k is the length of the oligonucleotide. The reconstructions of the phylogenetic trees took 98 s for $k = 8$ and the whole genome datasets ($k = 7$: 41 s, $k = 6$: 10 s, $k = 3$: 4 s), 86 s with $k = 8$ for the EST datasets ($k = 7$: 22 s, $k = 3$: 2 s) and 58 s for $k = 8$ and the mitochondrial genome datasets ($k = 7$: 13 s, $k = 3$: 1 s). These values refer to one round of bootstrapping. For comparison, one of the fastest whole genome alignment tools, called Mugsy, needs 45,000 s (ca. 12 h) to align the human and the mouse genomes (Angiuoli and Salzberg, 2011). However, whole genomes can only be aligned if they are from closely related species and, to our knowledge, phylogenies of multiple sequence alignments of the whole genomes from different eukaryotes have not been reconstructed yet.

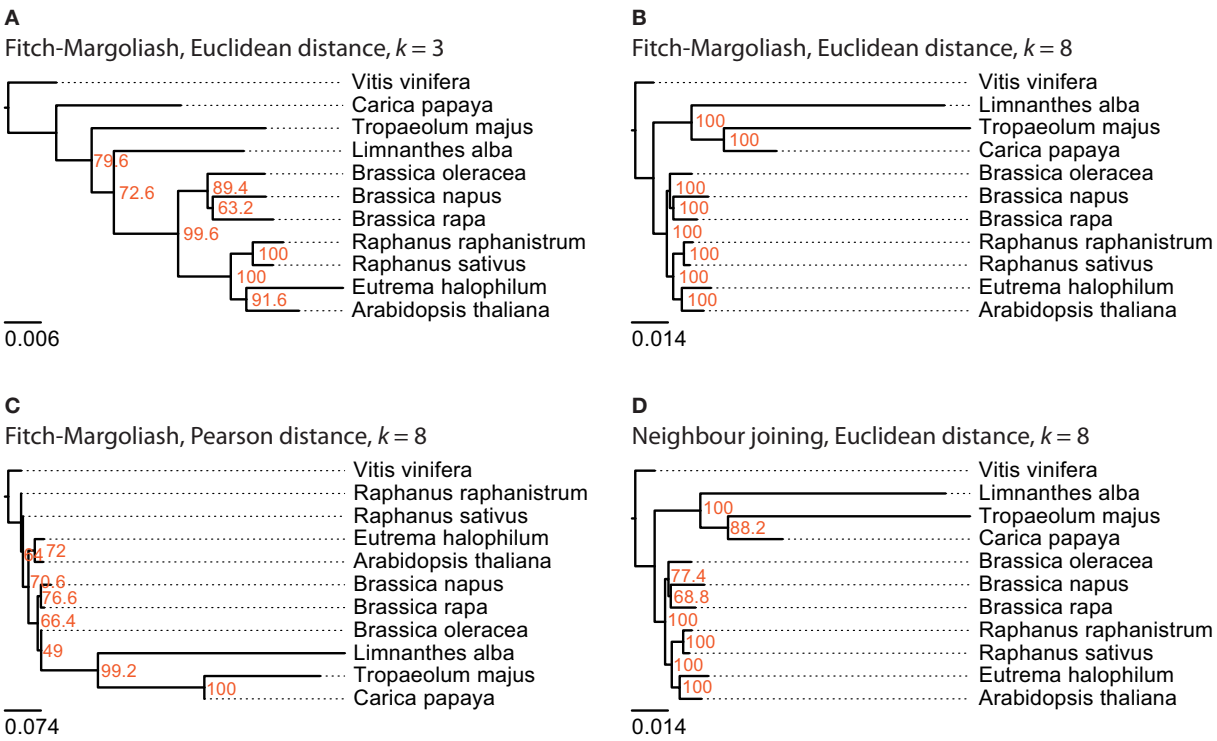


FIGURE 4 | The phylogenetic trees were generated applying different methods on FCGRs of public available EST data of the Brassicales taxon. In **(A–C)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The methods used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(D)** the Neighbor joining algorithm was used to calculate the tree.

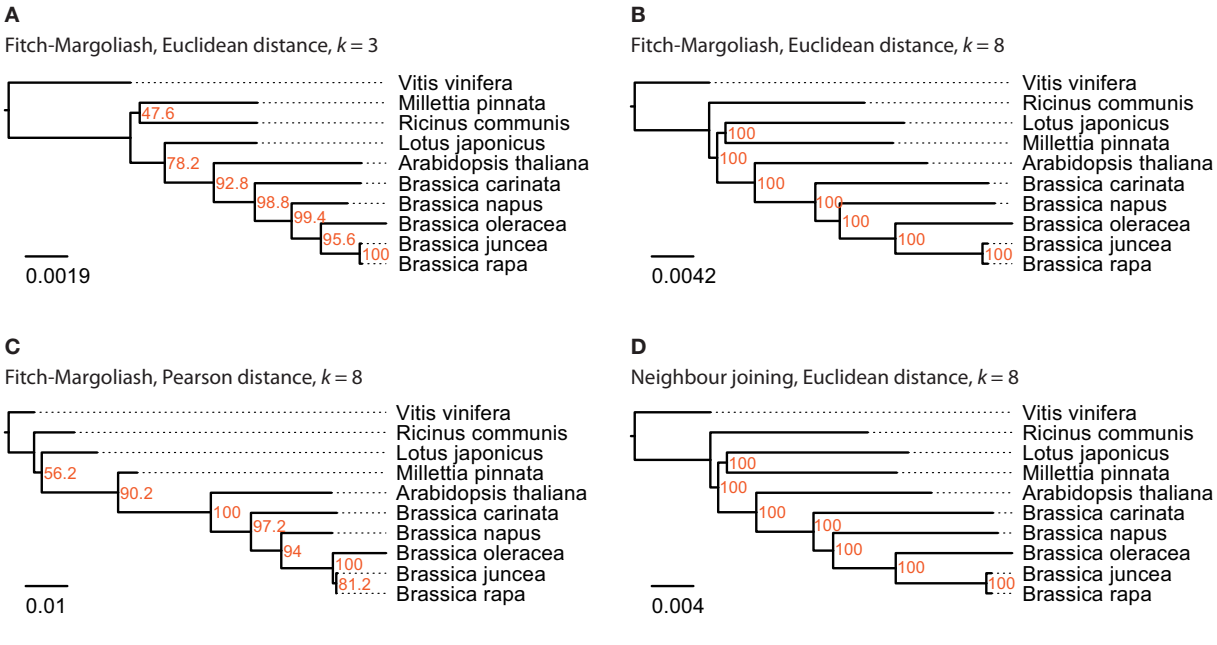


FIGURE 5 | The phylogenetic trees were generated applying different methods on FCGRs of available mitochondrial genome sequence data of the Rosids taxon. In **(A–C)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(D)** the Neighbor joining algorithm was used to calculate the tree.

DISCUSSION

In general, phylogenetic trees of species are reconstructed from amino acid or nucleotide sequence data, by comparing morphological characteristics, or by combining these data. While most of the sequence-based analyses are built on single genes, concatenated sequences are increasingly used, which could consist of even whole transcriptomes (phylogenomics). Here, we wanted to reconstruct the phylogeny of selected Brassicales species based on alignment-free sequence data. As approach we chose CGRs, which are scale-independent representations for genomic sequences (Jeffrey, 1990). Because CGRs are unique fingerprints of the corresponding sequences they cannot be compared directly. To reconstruct phylogenetic trees we therefore generated FCGRs at different resolutions. For the calculation of the distances between FCGRs we used the Euclidean (a geometric distance) and the Pearson (a statistical distance) method, and trees were reconstructed with the Fitch–Margoliash and the NJ algorithm.

Because of their different characteristics we compared three types of nucleotide sequences, nuclear genome sequences, mitochondrial genome sequences, and EST reads. Nuclear and mitochondrial genomes have been shown to have different GC contents and codon usage patterns (Zhang et al., 2007). EST data just comprise the exons and thus only part of the nuclear genome sequences. In addition EST data are potentially biased toward highly abundant genes and 5′- and 3′-terminal sequences. In order to reduce this bias we decided to include only those species for which at least 1,000 EST clones were available. Unfortunately, appropriate species from the Brassicales clade are not available for which all three types of nucleotide data have been sequenced. Therefore, we compared different sets of species for the three data types. Also, it is not known whether the mitochondrial genome data have been extracted from the whole genome datasets. As most of these are denoted as "draft assembly" we assume that the whole genome datasets still contain mitochondrial data. However, because of the very small size of the mitochondrial genomes compared to the nuclear genomes the results should be identical to those obtained from pure nuclear genome data. We would have liked to compare the results of each type of nucleotide sequence with the results of combined datasets but appropriate sequence data is not available. However, the EST and mitochondrial data do not comprise 1% of the whole genome data (**Table 1**) and a combined analysis should therefore be dominated by and be identical to the whole genome data.

The mitochondrial and whole genomes of the analyzed Brassicales species are of considerably different size, and different amounts of EST data are available. FCGRs naturally depend on the presence and frequency of the respective oligonucleotides and thus on the length of the analyzed sequence. For a reasonable result it is therefore essential to find the best balance between sequence length and FCGR resolution (oligonucleotide length), which represents the number of data available for the tree calculations and is also the main determinant for computing time. To exclude that the lengths of the concatenated sequences have an influence on the phylogenetic tree reconstructions of the Brassicales species at high FCGR resolution we calculated trees including the full-lengths sequences and specific defined subsets (**Figure 2**). At the resolution of octanucleotides, all partial sequences of whole genome assemblies containing more than 10 million nucleotides of each species group together while sets with 1 million nucleotides result in the ambiguous grouping of some species. In contrast, one million nucleotides of EST data, which correspond to the exon sequences, already result in consistent monophyly of all datasets of each species. Remarkably, this holds even true for the closely related *Brassica* species. The mitochondrial genomes of the analyzed species have sizes of 220–780 kbp. Thus, at the resolution of hexanucleotides it is not surprising that many oligonucleotides do not exist in sub-sections of 10 kbp leading to the artificial attraction of all these datasets in the reconstructed tree. Also, datasets of 100 kbp of the different *Brassica* species do not consistently group to the full-length mitochondrial genomes. Therefore, for mitochondrial data the resolution has to be reduced or full-length data to be used. As outgroup we choose *V. vinifera* in all analyses.

According to the diArk database, whole genome assemblies are available for 34 species belonging to the Malvidae/malvids (Hammesfahr et al., 2011). Twenty-two of them are *A. thaliana* strains of which we only included the reference strain into the analysis. A species tree including all these sequenced Malvidae is not available. Therefore, we assembled and annotated the CAPs α- and β-CAP, and the actin-related proteins Arp2 and Arp3 to generate a reference tree. The CAP and Arp proteins have been chosen for the reference tree because they are ubiquitous and well conserved in all eukaryotes (Goley and Welch, 2006; Cooper and Sept, 2008), and duplicates were most probably removed after the many whole genome duplication events that happened in plant evolution (Van de Peer, 2011). For example, the *A. thaliana* genome has experienced two duplications since its divergence from *Carica* (Tang et al., 2008), but has retained single copies of the CAP and Arp genes (Hammesfahr and Kollmar, 2012). Nevertheless, duplicated CAP and Arp2/3 genes have been identified in the *B. rapa* and *G. raimondii* genomes that are, however, the result of species-specific duplications. Only one of each duplicate has been used in this analysis. The phylogenetic tree of the concatenated CAP and Arp proteins is in agreement with other recent analyses containing part of the species (Bausher et al., 2006; Zhu et al., 2007; Wang et al., 2009) and can thus be regarded as reference tree. Compared to this reference tree, the FCGR tree based on the Pearson distance displays the most discrepancies followed by the tree based on low-resolution data ($k = 3$, trinucleotides). In addition, most of the branchings have low bootstrap values. The trees based on high-resolution data ($k = 8{,}65{,}536$ data points) and the Euclidean distance method show overall agreement with the reference trees independent of the method used for the tree reconstruction. Notably, *C. papaya* and *B. rapa* group wrongly, although both are only shifted by one branching event. Most surprisingly, the *Eucalyptus* species are completely wrongly grouped in all FCGR trees. Their exclusion from the tree calculation did not change the grouping of the other species (data not shown). However, the grouping of the Myrtales branch, which contains the *Eucalyptus* species, is different in all published trees (Bausher et al., 2006; Zhu et al., 2007; Wang et al., 2009) and their wrong placement in the FCGR trees might be due to some unknown characteristics of the genomes. Probably, they would group better, if species from other branches like the Crossosomatales, Geraniales, and Fabidae branches were included in the analysis. The

phylogenetic trees of the FCGRs of the mitochondrial genomes are very similar independently of the resolution, distance measure, and tree reconstruction method. Therefore either the species selection was fortunate or mitochondrial genome data is less sensitive with respect to these parameters.

When working with the EST data we observed disproportionate high frequencies for poly-A and poly-T oligonucleotides in the FCGRs. Probably, the poly-A tails were not consistently removed during the cDNA library construction. For low-resolution data (up to $k = 5$) the differences of the frequencies of these oligonucleotides to the next-highest values were not large enough to considerably bias the phylogenetic tree reconstructions. However, the topologies of trees based on high-resolution data ($k > 5$) are strongly disturbed. Therefore, we set the values for the frequencies of the poly-A and poly-T oligonucleotides to zero before we started the tree calculations. The artificial oligonucleotides generated at the boundaries of the concatenated EST reads apparently do not influence the resulting trees. The phylogeny of the *Brassica* species is slightly different compared to that obtained from the mitochondrial genome data. The genus *Brassica* includes 41 species (Velasco and Fernández-Martínez, 2010) the six with the highest economic importance being *B. rapa* (A), *Brassica nigra* (B), *Brassia oleracea* (C), *B. napus* (AC), *B. juncea* (AB), and *B. carinata* (BC). The first three comprise the three elementary species while the other three are amphidiploids that originated from natural hybridizations between two of the elementary species (Velasco and Fernández-Martínez, 2010). Thus the amphidiploid EST data contain mixtures of the hybridized species and dependent on which part is overrepresented in the data they will look closer related to one of their parent species. Although the distance in the phylogenetic tree is very small, *B. napus* seems to be closer to *B. rapa* based on the mitochondrial data. Based on the EST data, the hybrids *B. juncea* and *B. carinata* are more divergent than the parent species *B. rapa* and *B. oleracea*. Probably the part of the more divergent parent species *B. nigra* is dominating in this case.

In general we could show that FCGRs are well suited to phylogenetically group plant genomes and exonomes from even closely related species. We assume that FCGRs could also be used to group all eukaryotes provided that a balanced set of species from all lineages is taken. This has in part already been demonstrated on the phylogeny of 26 mitochondrial genomes of which only three were placed completely wrong when using the Euclidean distance method (Wang et al., 2005). However, this analysis was solely based on data from mitochondrions and biased against fish and mammalian species. Our analysis of the Brassicales clade has shown that high-resolution data (octanucleotides and longer sequences) result in better tree topologies and higher support for branchings. Trees based on the Pearson distance, which is a statistical distance measure, are less reliable than those based on Euclidean distances. The Fitch–Margoliash and NJ algorithms result in similar to identical trees. We have shown for the first time that the bootstrap concept to determine the support of the branchings in the tree, which is well established for trees based on sequence alignments since decades ("taxon-by-character" data matrix; Felsenstein, 1985), can also be applied to trees based on FCGRs. In another study it has been shown that although longer word lengths could reveal the correct clustering of the HIV-I subtypes in contrast to shorter word lengths (Pandit and Sinha, 2010) the grouping within the subtypes was always different. Also in this case a bootstrap analysis could have helped in the interpretation of the various branchings and we would recommend applying the bootstrap concept to all phylogenies based on FCGRs. FCGRs are fast to calculate and could be used in combination with alignment based data and morphological characteristics to improve the phylogenetic classification in ambiguous cases.

ACKNOWLEDGMENTS

We would like to thank Björn Hammesfahr for helpful suggestions and discussion, and Prof. Christian Griesinger for continuous support.

REFERENCES

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics* 17, 429–437.

Angiuoli, S. V., and Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–815.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Bérard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., and Lanaud, C. (2011). The genome of Theobroma cacao. *Nat. Genet.* 43, 101–108.

Basu, S., Pan, A., Chitra., and Das, J. (1997). Chaos game representation of proteins. *J. Mol. Graph. Model.* 15, 279–289.

Bausher, M. G., Singh, N. D., Lee, S.-B., Jansen, R. K., and Daniell, H. (2006). The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var "Ridge Pineapple": organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* 6, 21. doi:10.1186/1471-2229-6-21

Blair, C., and Murphy, R. W. (2011). Recent trends in molecular phylogenetic analysis: where to next? *J. Hered.* 102, 130–138.

Campbell, A., Mrázek, J., and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9184–9189.

Cantino, P. D., Doyle, J. A., Graham, S. W., Judd, W. S., Olmstead, R. G., Soltis, D. E., Soltis, P. S., and Donoghue, M. J. (2007). Towards a phylogenetic nomenclature of Tracheophyta. *Taxon* 56, 1E–44E.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31, 3497–3500.

Cooper, J. A., and Sept, D. (2008). New insights into mechanism and regulation of actin capping protein. *Int. Rev. Cell Mol. Biol.* 267, 183–206.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165.

Dassanayake, M., Oh, D.-H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R. A., Zhu, J.-K., Bohnert, H. J., and Cheeseman, J. M. (2011). The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.* 43, 913–918.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* 16, 1391–1399.

Dewey, C. N. (2012). Whole-genome alignment. *Methods Mol. Biol.* 855, 237–257.

Domazet-Loso, M., and Haubold, B. (2009). Efficient estimation of pairwise distances between genomes. *Bioinformatics* 25, 3221–3227.

Edwards, S. V., Fertil, B., Giron, A., and Deschavanne, P. J. (2002). A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.* 51, 599–613.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Goley, E. D., and Welch, M. D. (2006). The ARP2/3 complex: an actin nucleator comes of age. *Nat. Rev. Mol. Cell Biol.* 7, 713–726.

Hammesfahr, B., and Kollmar, M. (2012). Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein. *BMC Evol. Biol.* 12, 95. doi:10.1186/1471-2148-12-95

Hammesfahr, B., Odronitz, F., Hellkamp, M., and Kollmar, M. (2011). diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Res. Notes* 4, 338. doi:10.1186/1756-0500-4-338

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A. F., Weissenbach, J., Quétier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res.* 18, 2163–2170.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282.

Joseph, J., and Sasikumar, R. (2006). Chaos game representation for comparison of whole genomes. *BMC Bioinformatics* 7, 243. doi:10.1186/1471-2105-7-243

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17, 149–154.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M. L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J. K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M. C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452, 991–996.

Pandit, A., and Sinha, S. (2010). Using genomic signatures for HIV-1 sub-typing. *BMC Bioinformatics* 11(Suppl. 1), S26.

Pleissner, K. P., Wernisch, L., Oswald, H., and Fleck, E. (1997). Representation of amino acid sequences as two-dimensional point patterns. *Electrophoresis* 18, 2709–2713.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2677–2682.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* 57, 758–771.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954.

Van de Peer, Y. (2011). A mystery unveiled. *Genome Biol.* 12, 113.

Velasco, L., and Fernández-Martínez, J. M. (2010). "Other Brassicas," in *Oil Crops Handbook of Plant Breeding*, eds J. Vollmann and I. Rajcan (Springer New York), 127–153.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L. M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J. T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J. A., Sterck, L., Vandepoele, K., Grando, S. M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S. K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F., and Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326. doi:10.1371/journal.pone.0001326

Vinga, S., and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics* 19, 513–523.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D., Brockington, S. F., Alexandre, R., Davis, C. C., Latvis, M., Manchester, S. R., and Soltis, D. E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3853–3858.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., Zhang, Z., and Brassica rapa Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* 43, 1035–1039.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005). The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* 346, 173–185.

Zhang, W., Zhou, J., Li, Z., Wang, L., Gu, X., and Zhong, Y. (2007). Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in Triticum aestivum L. *J. Integr. Plant Biol.* 49, 246–254.

Zhu, X.-Y., Chase, M. W., Qiu, Y.-L., Kong, H.-Z., Dilcher, D. L., Li, J.-H., and Chen, Z.-D. (2007). Mitochondrial matR sequences help to resolve deep phylogenetic relationships in Rosids. *BMC Evol. Biol.* 7, 217.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 May 2012; accepted: 06 August 2012; published online: 29 August 2012.

Citation: Hatje K and Kollmar M (2012) A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method. Front. Plant Sci. ***3****:192. doi: 10.3389/fpls.2012.00192*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

METHODS ARTICLE
published: 08 November 2012
doi: 10.3389/fpls.2012.00243

Sequencing of chloroplast genome using whole cellular DNA and Solexa sequencing technology

*Jian Wu[1†], Bo Liu[1†], Feng Cheng[1], Nirala Ramchiary[2], Su Ryun Choi[2], Yong Pyo Lim[2] and Xiao-Wu Wang[1]**

[1] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China
[2] Department of Horticulture, Plant Genome Research Institute, Chungnam National University, Daejeon, South Korea

Edited by:
Michael Freeling, University of California Berkeley, USA

Reviewed by:
Eduard Akhunov, Kansas State University, USA
Xiyin Wang, Hebei United University, China

***Correspondence:**
Xiao-Wu Wang, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Zhongguancun Nan Da Jie 12, Beijing 100081, China.
e-mail: wangxw@mail.caas.net.cn

[†] Jian Wu and Bo Liu have contributed equally to this work.

Sequencing of the chloroplast (cp) genome using traditional sequencing methods has been difficult because of its size (>120 kb) and the complicated procedures required to prepare templates. To explore the feasibility of sequencing the cp genome using DNA extracted from whole cells and Solexa sequencing technology, we sequenced whole cellular DNA isolated from leaves of three *Brassica rapa* accessions with one lane per accession. In total, 246, 362, and 361 Mb sequence data were generated for the three accessions Chiifu-401-42, Z16, and FT, respectively. Micro-reads were assembled by reference-guided assembly using the cpDNA sequences of *B. rapa*, *Arabidopsis thaliana*, and *Nicotiana tabacum*. We achieved coverage of more than 99.96% of the cp genome in the three tested accessions using the *B. rapa* sequence as the reference. When *A. thaliana* or *N. tabacum* sequences were used as references, 99.7–99.8 or 95.5–99.7% of the *B. rapa* cp genome was covered, respectively. These results demonstrated that sequencing of whole cellular DNA isolated from young leaves using the Illumina Genome Analyzer is an efficient method for high-throughput sequencing of cp genome.

Keywords: chloroplast genome, sequencing, Solexa sequencing technology, whole cellular DNA, *Brassica rapa*

BACKGROUND

The chloroplast (cp) genome contains a wealth of information that has been shaped by speciation, rendering it a rich resource to trace population-level processes and evolutionary divergence. Therefore, the cp genome sequence is very important in several fields of plant biology, including phylogenetics, molecular biology, evolutionary biology, and cp genetic engineering. Complete sequences of cp genomes were first reported for tobacco and a liverwort in 1986 (Ohyama et al., 1986; Shinozaki et al., 1986). Since then, the sequences of cp genomes from a number of land plants and algae have been determined. However, there are still challenges in rapid and cost-effective sequencing or re-sequencing of the cp genome.

Traditional sequencing begins with the construction of plastid or other genomic libraries. Construction of these resources is a difficult part of this work, since it is complicated to isolate cps and construct libraries. For taxa that are rare and/or difficult to obtain, large-scale isolation of cps would be one difficulty in the conventional sequencing methodology. Therefore, a method to sequence cp genomes with a simple and rapid sample preparation step would greatly benefit research that requires cpDNA sequencing.

Because of the characteristics of conserved genome size, gene arrangement, and coding sequences among cp genomes, a PCR-based approach has been used for their amplification, sequencing, and assembly (Dhingra and Folta, 2005; Cronn et al., 2008). This method is useful for obtaining cpDNA sequences from divergent species. However, the complex PCR procedure and the large number of PCR reactions required to cover the cp genome of >120 kb limits the application of this approach. Therefore, there is a need for a simpler method to sequence cp genomes. Since whole cellular DNA contains chromosomal DNA, mitochondrial DNA, and cp DNA, any of the three types of DNA sequence could be derived from sequencing this DNA material. The fact that the grapevine cp genome sequence was obtained as a byproduct of whole genome sequencing indicates that it is possible to use total DNA for cp genome sequencing (Velasco et al., 2007). To ensure accurate assembly, a sufficient sequencing depth is required. However, this is difficult to achieve using traditional Sanger sequencing technology.

The development of next-generation sequencing technologies has shed new light on easy sequencing of complete cp genomes. Moore et al. (2006) were the first to attempt to use second generation sequencing technology (a 454 GS 20 system) for the cp genome. Cronn et al. (2008) developed a multiplex sequencing-by-synthesis approach to sequence cp genomes. That method combined PCR-amplified cp genomes and indexing tags for Solexa sequencing. Both of these studies demonstrated that second generation sequencing technology is a powerful tool for sequencing plastid genomes quickly and economically. However, both methods required substantial work to prepare sequencing templates either for traditional plasmid isolation or to obtain PCR-amplified cpDNA fragments.

In this study, using routinely isolated whole cellular DNA from young leaves of *Brassica rapa*, we generated highly accurate and essentially complete cp genome sequences by Illumina GA II. Our results show that this method is faster and more economical than traditional methods, and can be used for rapid sequencing of cp genomes.

MATERIALS AND METHODS

DNA SOURCES

Three *B. rapa* accessions, Chiifu-402-41, Z16, and FT were used for sequencing of the cp genome. Chiifu-402-41, the material used for whole genome sequencing (Wang et al., 2011), was donated by the Korea *Brassica* Genome Resource Bank. The accession Z16 is a Chinese cabbage line (ssp. *pekinensis*) obtained from the Chinese Academy of Agricultural Sciences (CAAS). The accession FT (CGN1010) is a fodder turnip (ssp. *rapa*) obtained from the Dutch Crop Genetic Resources Center (CGN), Wageningen, The Netherlands. DNA was isolated from young leaves using a modified CTAB method (Fulton et al., 1995).

DNA SEQUENCING

The isolated whole cellular DNA was sheared, polished, and prepared according to Illumina Sample Preparation kit (Solexa Inc, 2007). The nucleotide sequence was determined according to the Solexa sequencing method (Wang et al., 2008). Sequencing was carried out with one accession per lane to generate 35-mer micro-reads.

REFERENCE-GUIDED ASSEMBLY FOR CP GENOME SEQUENCE

After the run, fluorescent images were processed into sequences using the Illumina/Solexa Pipeline (version 0.2.2.6). Reads containing "N" were filtered out before further analyses. The strategy used for cp genome assembly is outlined in **Figure 1**. The

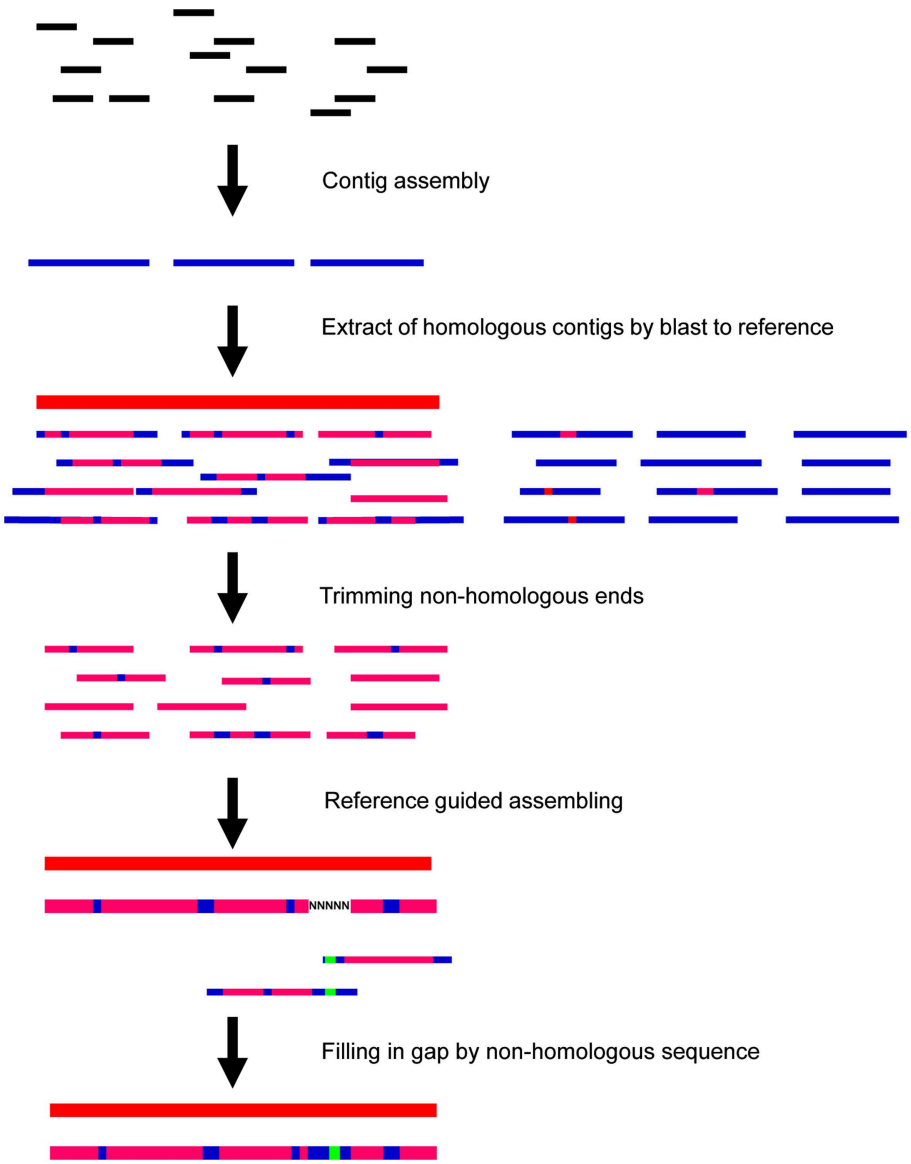


FIGURE 1 | Schematic view of micro-reads assembly method for chloroplast (cp) genome. *De novo* assembled contigs (blue bars) are aligned to reference (red bar) to extract sequences generated from cp genome (thin pink bars). Draft consensus (thick pink bar) was constructed guided by reference. Gaps were filled by extending sequence and joining two contigs that overlapped (green bar) by 10 or more nt.

micro-reads (35-mer) were first *de novo* assembled using SOAPdenovo (Li et al., 2010) with an overlap *k*-mers value of 27. Second, contigs longer than 50 nt were extracted to construct the consensus using our reference-guided assembly (RGA) program. During the RGA process, all contigs were aligned to the reference cp genome sequence by BLASTN (http://www.ncbi.nlm.nih.gov). Contigs with matches greater than 80% were selected for assembly after trimming non-homologous ends. The trimmed fragments of 50 nt or longer were assembled under the guidance of the reference. If there was no overlap between two adjoining fragments, the interval was filled with "N" to match the length in the draft consensus. To fill these gaps, the corresponding two contigs of each gap were extended to the end and merged if there was an overlap of 10 or more bp (**Figure 1**).

To explore the applicability of this cp genome sequencing strategy, we compared the assemblies guided by cp genome sequences of *B. rapa* (GenBank: DQ231548), *Arabidopsis thaliana* (GenBank: NC000932; Sato et al., 1999), or *Nicotiana tabacum* (GenBank: NC001879; Shinozaki et al., 1986).

GENE ANNOTATION

The *B. rapa* cp genes were annotated using DOGMA (Wyman et al., 2004). This program used a FASTA formatted input file of the complete cp genome sequences and identified putative protein-coding genes by performing BLASTX searches against a custom database of previously published cp genomes. Both transfer RNA and ribosomal RNA were identified by search against *A. thaliana* genes database using BLASTN.

SEQUENCE VALIDATION

To assess the accuracy of this approach for sequencing cp genome, the consensus assembled for Chiifu-402-41 cp was mapped against the sequence derived from the Sanger-based sequencing method (GenBank: DQ231548). We used BLASTN (http://www.ncbi.nlm.nih.gov) to identify mismatches between the two sequences. Once a mismatch was identified, we designed sequence-specific primers to amplify the inconsistent region and sequenced the fragment using the traditional Sanger sequencing method.

POLYMORPHISM DISCOVERY IN THE *B. RAPA* CP GENOME

To assess the applicability of this approach for re-sequencing of the cp genome, we used all the *de novo* assembled contigs to align onto the reference cp genome sequence (GenBank: DQ231548) to identify single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (InDels) among the three accessions.

To identify the SNPs based on short reads aligning, we used short reads of Z16 and FT accessions to respectively align to reference sequence of Chiifu-402-41 cp genome based on SOAP with default parameter (Li et al., 2008). And then, using these results of reads aligning, we identified SNPs between Chiifu-402-41 vs. Z16 and Chiifu-402-41 vs. FT by SOAPsnp with default parameter (Li et al., 2009).

RESULTS

ANALYSIS OF MICRO-READS GENERATED FROM WHOLE CELLULAR DNA

One lane Solexa sequencing of whole cellular DNA produced 7,015,639, 10,313,714, and 10,356,209 35-mer micro-reads for Chiifu-402-41, Z16, and FT, respectively, corresponding to the total sequenced length of 246, 362, and 361 Mb, respectively (**Table 1**). Of the total reads, 9.3% (653,057), 26.4% (2,721,148), and 10.4% (1,073,449) were mapped to the reference cp genomes for Chiifu-402-41, Z16, and FT, respectively, corresponding to an overall average sequencing depth of 103×, 550×, and 217×, of their cp genome, respectively (**Figure 2**). cp Genomes of most land plants have a quadripartite structure with large and small single copy regions (LSC and SSC) separated by two copies of large inverted repeats (IRa and IRb). The *B. rapa* cp genome is 153,482-bp long; the LSC is 83,282 bp, the SSC 17,776 bp, and each of the two IR copies is 26,212 bp (**Figure 2**). The depth of the reads showed two peaks in the corresponding inverted repeat regions after being aligned to the reference genome. This was consistent with the structure of the cp genome.

CONTIG ASSEMBLY USING A REFERENCE-GUIDED APPROACH

Small reads were first assembled into contigs, from which very short N50 lengths (length of the shortest contig among those that collectively covered 50% of the assembly) were derived for all three accessions. For the "best assembly" lane of the Chiifu-402-41 accession, there were 25,493 contigs with an N50 of 82 nt. In the second assembly step, the number of contigs that met the

Table 1 | Characteristics of reads from one lane sequencing on Illumina Solexa 1G Genome sequencer.

	Chiifu-402-41	Z16	FT
Total reads	7,015,639	10,313,714	10,356,209
Aligned reads	653,057	2,721,148	1,073,449
Aligned ratio (%)	9.3	26.4	10.4
Mean read depth (-fold)	103	550	217
N50 (bp)	13,509	3997	7461

Mean read depth was calculated by including one copy of inverted repeats.

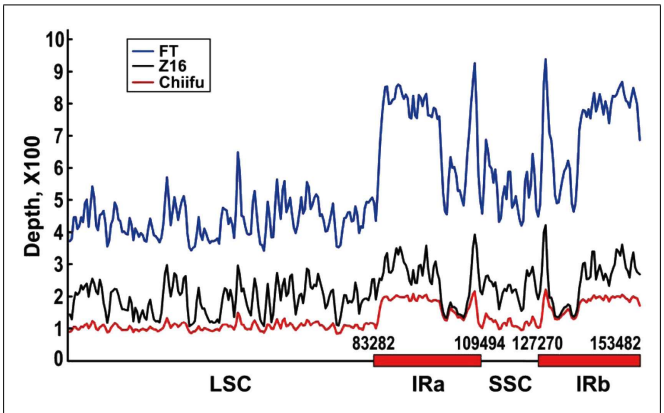


FIGURE 2 | Plot showing sequencing depth by position for chloroplast genomes of three *B. rapa* accessions sequenced by Solexa Genome Analyzer with whole cellular DNA as template. Number of micro-reads per position (*y*-axis) is plotted against position in the assembly (*x*-axis, in kb) in a window size of 100 bp. Numbers above *x*-axis indicate boundary sites of large single copy (LSC) and small single copy (SSC), and two inverted repeats (IRa and IRb).

conditions of longer than 50 bp and greater than 80% identity to the *B. rapa* reference cp genome sequence decreased to 22 with an N50 of 13,509 nt. Z16 had 86 contigs and an N50 of 3,997 nt; FT had 46 contigs and an N50 of 7,461 nt (**Table 1**). The contigs were trimmed from the positions of the first or the last hit nucleotide to build a draft consensus. Finally, to fill gaps in the consensus, the sequence of the two corresponding contigs in the direction of the gap were compared. If there was an overlap of 10 bp or more, the two contigs were joined together. Using this strategy, we achieved a minimum coverage of 99.96% of the cp genome for the *B. rapa* accessions Chiifu-402-41, Z16, and FT, with one, seven, and five gaps, respectively (see Table S1 in Supplementary Material for positions and lengths of gaps). The only gap in the consensus of Chiifu-402-41 was also present in the other two accessions. When went through the sequence of region where the gap located, it revealed that there was a satellite sequence of $(tggatatagactcatgaaag)_3$. The common 8-bp gap was located in the second repeat of the satellite.

To evaluate the applicability of this approach to sequencing those species without known cp sequences, we assembled consensuses using *A. thaliana* or *N. tabacum* as a reference and then compared the completeness of these assemblies. Guided by the cp genome sequence of *A. thaliana* and *N. tabacum*, the assembled contigs covered 99.83 and 97.77% of the cp genome from *B. rapa* accession Chiifu-402-41, respectively (**Table 2**; **Figure 3**). Among the three sequenced accessions, the lowest coverage in FT was 99.73%, which was obtained using the *A. thaliana* sequence as the reference. The lowest coverage in Z16 was 95.55%, which was obtained using the *N. tabacum* sequence as the reference. In assemblies guided by the *A. thaliana* sequence, there were four, seven, or six gaps in the consensus of Chiifu-402-41, Z16, or FT, respectively. In assemblies guided by the *N. tabacum* cpDNA sequence, there were 4, 21, or 10 gaps in the consensus of Chiifu-402-41, Z16, or FT, respectively. Table S1 in Supplementary Material shows the positions and lengths of the gaps. None of the gaps was common between the assemblies guided by *A. thaliana* or *N. tabacum*.

Using DOGMA (Wyman et al., 2004) software, a total of 89 potential protein-coding genes (including eight genes duplicated in the inverted repeat), 8 ribosomal RNA genes and 37 transfer RNA genes were assigned to the genome on the basis of similarities to the cp gene previously reported in other species (**Figure 4**; Table S2 in Supplementary Material). Fourteen genes contained one intron, while one gene had two introns. In addition, the number of genes in *B. rapa* cp genome was similar to *A. thaliana* (a total of 128 genes including 89 protein-coding genes, 4 rRNA genes, and 37 tRNA genes; Sato et al., 1999), while it was distinctly more than that in the *N. tabacum* cp genome (84 genes were identified in total; Shinozaki et al., 1986).

No base error was observed in the assemblies guided by either *A. thaliana* or *N. tabacum* for Chiifu-402-41. However, several base errors were observed in the assemblies of FT and Z16 (**Table 2**; see Table S2 in Supplementary Material for positions of errors). The rate of base error was 0.0078% for Z16 and 0.0058% for FT when guided by *A. thaliana*, and 0.022% for Z16 and 0.0019% for FT when guided by *N. tabacum*. The errors located at 37 bp and 57 bp were appeared in both assemblies of FT and Z16 guided by either *A. thaliana* or *N. tabacum*.

A fragmented loss of sequence from the consensus was observed for all three *B. rapa* accessions when using *A. thaliana* or *N. tabacum* as the reference (see Table S3 in Supplementary Material for the positions of InDel). To analyze this further, we aligned the

Table 2 | Assembly of chloroplast genomes of three *B. rapa* accessions guided by cp genome sequence of *B. rapa* (DQ231548), *A. thaliana* (NC000932), and *N. tabacum* (NC001879).

Accession	*B. rapa*			*A. thaliana*				*N. tabacum*			
	Coverage (%)	No. of gaps	Total length of gaps (bp)	Coverage (%)	No. of gaps	Total length of gaps (bp)	Error base	Coverage (%)	No. of gaps	Total gap length (bp)	Error base
Chiifu	99.99	1	8	99.83	4	156	0	97.77	4	1048	0
Z16	99.96	7	52	99.74	7	310	12	95.52	21	4161	34
FT	99.96	5	51	99.73	6	481	9	96.19	10	2339	3

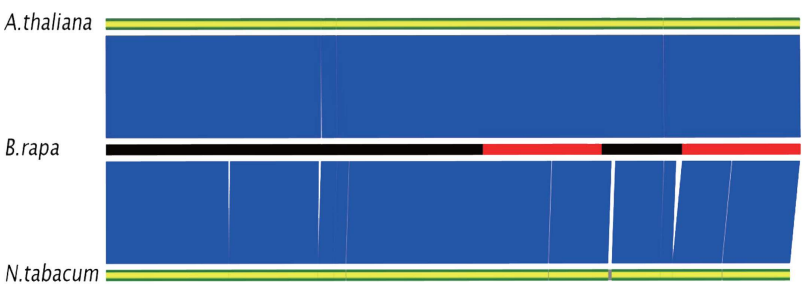


FIGURE 3 | Comparison of assemblies for Chiifu-402-41 guided by chloroplast genome sequences of *B. rapa* (DQ231548), *A. thaliana* (NC000932), and *N. tabacum* (NC001879). Assembled consensuses are represented by three bars. For *B. rapa*, black parts of bar indicate LSC or SSC; red parts of bar indicate IRa or IRb. For *A. thaliana* and *N. tabacum*, silver parts indicate gaps. Blue block between bars of *B. rapa* and *A. thaliana* or *N. tabacum* indicates identity >95%.

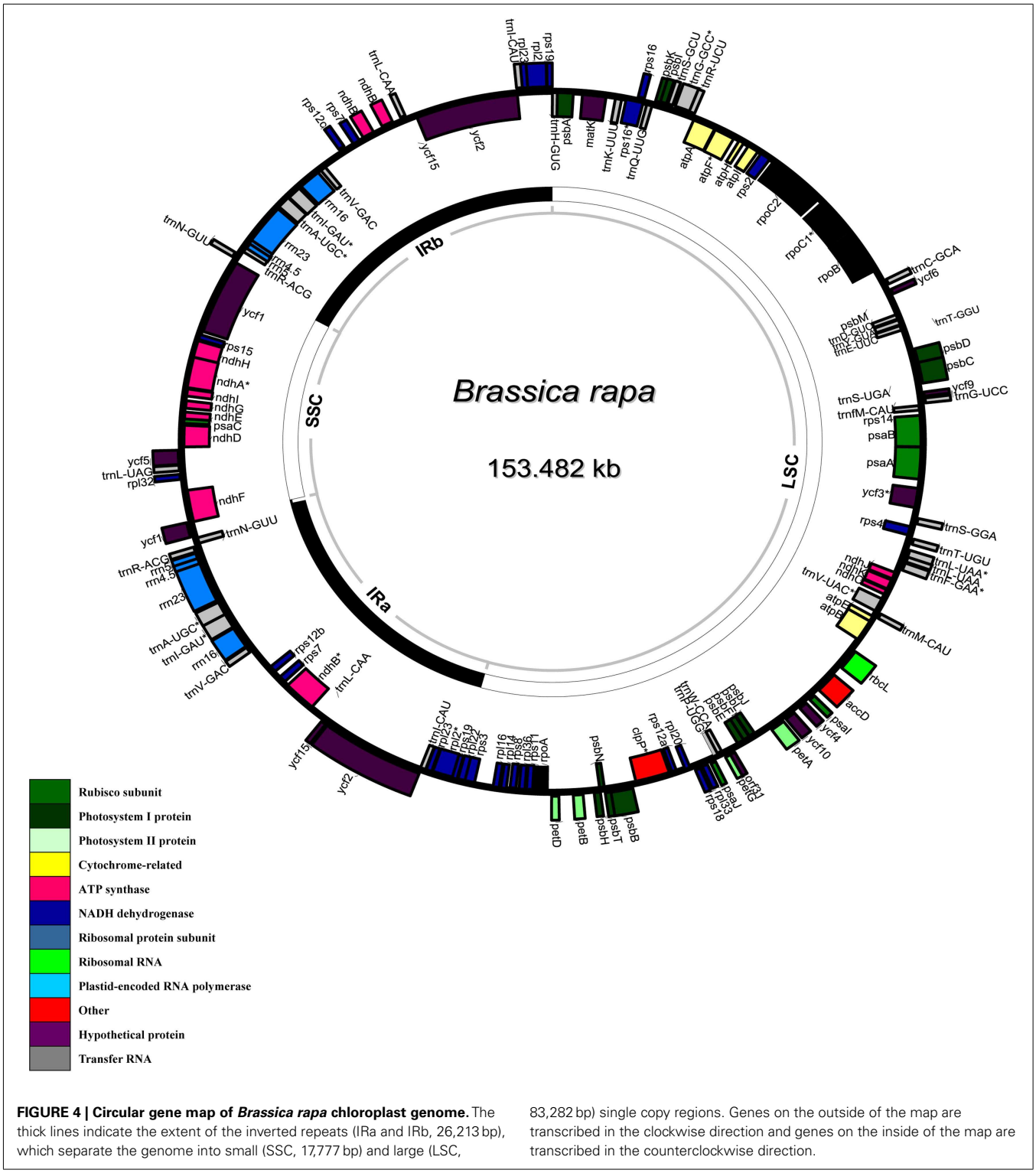


FIGURE 4 | Circular gene map of *Brassica rapa* chloroplast genome. The thick lines indicate the extent of the inverted repeats (IRa and IRb, 26,213 bp), which separate the genome into small (SSC, 17,777 bp) and large (LSC, 83,282 bp) single copy regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction.

cp sequence of *B. rapa* with those of *A. thaliana* and *N. tabacum*. This revealed that this type of error was resulted from sequence deletions in cp genomes of *A. thaliana* and *N. tabacum*. Since this type of error was reflected by calculating coverage, it was not included in the number of base errors.

SEQUENCE VALIDATION

Comparison of the consensus of Chiifu-402-41 assembled in this study with the reference sequence generated earlier by Sanger sequencing revealed one base difference (Ref: G/Solexa: A) at position of 116,457 bp at a sequencing depth of 72× by Solexa.

To confirm the nucleotide at this position, we designed primers to amplify the inconsistent region. We amplified a 508-bp cp DNA fragment using primers 5′-AAAATCATTCGTGGTAA-3′ (forward) and 5′-AAATCATTGCTTCATCTA-3′ (reverse). The produced fragment was later sequenced by conventional Sanger sequencing. The result showed that the nucleotide at the mismatched position should be A, and not G, in the reference cp sequence of Chiifu-402-41.

POLYMORPHISMS IDENTIFIED IN RE-SEQUENCING

To identify the sequence polymorphisms of cp genomes within *B. rapa* species, we mapped all *de novo* assembled contigs against the reference cp sequence (DQ231548). The hit contigs covered 100% of the cp genome for all three accessions. Comparison of Z16 and FT cp genome sequences with that of the reference cp sequence of Chiifu-402-41 revealed 31 and 8 SNPs, respectively (**Table 3**). Only 1-bp InDels were identified among the three accessions. Four InDels were observed either between Chiifu-402-41 and Z16 or between Chiifu-402-41 and FT when their consensus cpDNA sequences were compared (**Table 3**; see Table S4 in Supplementary Material for InDel positions). Amongst these SNPs and InDels, there were 14, 19, 3, and 7 mutations located in LSC, IRa, SSC, and IRb, respectively. Particularly, 27 of these 43 (63%) mutations located in nine genes of *B. rapa* cp genome (*trnH*-GUG, *rpoC2*, *rpoC1*, *rps19.1*, *rpl2.1*, *rrn23*, *psaC*, *ycf1.2*, and *rrn23*; see Table S4 in Supplementary Material).

We further compared the efficiency in sequence variants discovery by alignment of *de novo* assembled consensus with alignment of short reads. Using SOAPsnp (Li et al., 2009), a total of 10 SNPs were identified between Chiifu-402-41 and Z16, and 5 SNPs were identified between Chiifu-402-41 and FT. Comparing to SNPs identified by consensus aligning method, seven SNPs between Chiifu-402-41 and Z16, and two SNPs between Chiifu-402-41 and FT were same in genotype in both methods (see Table S4 in Supplementary Material). To confirm the nucleotides at the positions where different nucleotides were indicated by two methods, we randomly selected two positions (at the positions of 66,083–66,086 and 131,065) to design primers to amplify the inconsistent regions in Z16 accession (see Table S4 in Supplementary Material for primer sequences). The result from traditional Sanger sequencing showed that at the position 66,083–66,086 the 4 nt were "TTCT" which were consistent to those by consensus aligning method, however, the nucleotide at the position of 131,065 was "C" which was same as the result from reads aligning method.

Table 3 | Sequence polymorphisms identified by reference-guided assembly of cp genomes of *B. rapa*.

	Chiifu-402-41	Z16	FT
No. of SNP	1	31	8
No. of InDel	0	4	4

No. of SNPs (single nucleotide polymorphisms) and No. of insertions/deletions (InDels) are those compared with reference (DQ231548).

DISCUSSION

In this study, we demonstrated the feasibility of sequencing the cp genome using whole cellular DNA and reference-guided *de novo* assembly of micro-reads into complete cp genome sequences. In light of the challenges associated with established methods for cp genome sequencing, micro-read sequencing of whole cellular DNA is a rapid and cost-effective approach for sequencing/re-sequencing of cp genomes.

Large-scale isolation of chloroplasts is one of the laborious works in traditional cp genome sequencing methods. To avoid this problem, we used whole cellular DNA instead of cpDNA as the sequencing template. Unlike the alternative PCR-based methods reported previously (Dhingra and Folta, 2005; Cronn et al., 2008), this method also avoids the possibility of introducing base errors during PCR reactions.

The cp genome is well-conserved in terms of size, gene arrangement, and coding sequences, at least within major subgroups of the plant kingdom. This is the basis for this whole cellular DNA sequencing strategy, as it allows micro-reads to be assembled correctly using a reference-guided method. As expected, the closer the phylogenetic relationship between the reference and the target, the better the coverage of the assembled cp genome. However, our results of greater than 99% coverage when the assembly was guided by the *A. thaliana* cp DNA sequence and greater than 95% coverage when guided by the *N. tabacum* cp genome sequence indicated that this method could be also used for species without a closely related species in the plastid genome database. More than 150 cp genome sequences from plants belonging to 124 genera have been deposited in GenBank. This provides a base for widespread use of this approach to assemble draft cp genome sequences.

For the cp genome assembly, instead of the strategy of *de novo* assembling contigs which we used in the present study, small reads alignment to reference genome is an alternative approach. One can expect that the read mapping approach is computationally less demanding and faster than *de novo* assembling. Moreover, it has the advantage that the read coverage information can be used for reliable detection of sequence variation. However, when there are InDels or structural variations in cp genomes, which have been reported for nine grass species (Golenberg et al., 1993), Korean ginseng (Kim and Lee, 2004), *Pinus* (Parks et al., 2009), the strategy of assembling contigs is very important to identify exact mutations between cp genomes. Additionally, for the species without known cp genome sequence, aligning short reads to cp genome of other plant species was impractical based on short reads aligner, especially for the high ratio polymorphism regions, because the reads from repeated regions or homologous regions cannot be distinguished. Furthermore, conventional Sanger sequencing result on two positions (66,083 and 131,065 bp) indicated that both consensus aligning method and reads aligning method existed false positive in SNP identification. To improve the veracity in sequence variant identification, generating pair-ends reads is a potential approach.

Although we obtained reasonably high coverage of the cp genome using the present approach, there were a number of gaps in the consensus sequence. Taking the assembly of Chiifu-402-41 as an example, when the *B. rapa* cp genome sequence was used as the reference, there was an 8-bp gap in a region of three 20-bp

repeats, and this gap was also present in the other two accessions. This is caused by the limitations of the assembly software itself because of the *k*-mer value used in the assembly, rather than overlapping of micro-reads. However, when the contigs shorter than 50 bp were aligned to the reference, this gap was filled by a 42-nt contig with overlaps to the two adjoining contigs. In the assemblies guided by sequences of *A. thaliana* or *N. tabacum*, gaps could result from the limitations of the assembly software as described above, or from structure variation such as large InDels between the target and the reference.

In terms of the errors observed in assemblies guided by distant references such as *A. thaliana* and *N. tabacum*, the error rates were substantially lower than those of 0.056% reported by Cronn et al. (2008) and 0.037% reported by Moore et al. (2006). The fact that no errors were observed for Chiifu-402-41 could be because of the very evenly distributed sequencing depth for this accession.

We sequenced whole cellular DNA using one accession per Solexa lane, which resulted in sequence redundancy reaching an average read density of 103× to 550× for each base of the cp genome. These sequencing depths could be over-estimated due to reads from homologous fragments in the chromosomal genome or mitochondrial genome, and also due to reads from repeated regions of cp genome. However, the true sequencing depth is more than sufficient for cp genome assembly. A previous study showed that sequencing depths greater than 10-fold had little effect on genome coverage (Li et al., 2008). This indicates that with the present sequencing strategy, a reasonable sequencing depth and coverage can be achieved by multiplex sequencing in one lane. This will considerably reduce sequencing costs, especially considering that new developed Solexa sequencing techniques can generate approximately 35 Gb of 150-bp paired-end reads from one lane.

In addition to assembly of the cp genome, the micro-reads produced using this approach could be used to identify SNPs in species for which the whole genome sequence is available, because most of these reads are from the chromosomal DNA sequence. This could be an additional value for the data obtained using this approach.

ACKNOWLEDGMENTS

The work was supported by the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, P. R. China, and the Technology Development Program for Agriculture and Forestry, Ministry for Food, Agriculture, Forestry and Fisheries, Republic of Korea (No. 607003-05).

SUPPLEMENTARY MARTIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Plant_Genetics_and_Genomics/10.3389/fpls.2012.00243/abstract

Table S1 | Gaps in assemblies guided by cp sequences of *B. rapa*, *A. thaliana*, or *N. tabacum*.

Table S2 | Annotated genes in *B. rapa* cp genome.

Table S3 | Errors and insertion/deletions (InDels) in assemblies guided by cp sequence of *A. thaliana* or *N. tabacum*.

Table S4 | Single nucleotide polymorphisms (SNPs) and insertion/deletions (InDels) identified for *B. rapa* accessions Chiifu-402-41, Z16, and FT.

REFERENCES

Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., and Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* 36, e122.

Dhingra, A., and Folta, K. M. (2005). ASAP: amplification, sequencing and annotation of plastomes. *BMC Genomics* 6, 176. doi:10.1186/1471-2164-6-176

Fulton, T. M., Chunwongse, J., and Tanksley, S. D. (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Rep.* 13, 207–209.

Golenberg, E. M., Clegg, M. T., Durbin, M. L., Doebley, J., and Ma, D. P. (1993). Evolution of a noncoding region of the chloroplast genome. *Mol. Phylogenet. Evol.* 2, 52–64.

Kim, K. J., and Lee, H. L. (2004). Complete chloroplast genome sequences from Korean ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* 11, 247–261.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19, 1124–1132.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 713–714.

Moore, M. J., Dhingra, A., Soltis, P. S., Shaw, R., Farmerie, W. G., Folta, K. M., et al. (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.* 6, 17. doi:10.1186/1471-2229-6-17

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., et al. (1986). Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature* 322, 572–574.

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7, 84. doi:10.1186/1741-7007-7-84

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6, 283–290.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., et al. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 5, 2043–2049.

Solexa Inc. (2007). "Protocol for Whole Genome Sequencing using Solexa Technology," in *BioTechniques Protocol Guide* (New York, NY: BioTechniques), 29.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326. doi:10.1371/journal.pone.0001326

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., et al. (2008). The diploid genome sequence of an Asian individual. *Nature* 456, 60–65.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 July 2012; accepted: 12 October 2012; published online: 08 November 2012.

*Citation: Wu J, Liu B, Cheng F, Ramchiary N, Choi SR, Lim YP and Wang X-W (2012) Sequencing of chloroplast genome using whole cellular DNA and Solexa sequencing technology. Front. Plant Sci. **3**:243. doi: 10.3389/fpls.2012.00243*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 04 January 2013
doi: 10.3389/fpls.2012.00297

Inferring the *Brassica rapa* interactome using protein–protein interaction data from *Arabidopsis thaliana*

Jianhua Yang[1]*, Kim Osman[1], Mudassar Iqbal[2], Dov J. Stekel[2], Zewei Luo[1], Susan J. Armstrong[1] and F. Chris H. Franklin[1]*

[1] *University of Birmingham, Birmingham, UK*
[2] *University of Nottingham, Nottingham, UK*

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Xiangfeng Wang, University of Arizona, USA
Xiyin Wang, Hebei United University, China

****Correspondence:***
Jianhua Yang and F. Chris H. Franklin, University of Birmingham, B152TT Birmingham, UK.
e-mail: j.yang.6@bham.ac.uk, f.c.h.franklin@bham.ac.uk

Following successful completion of the *Brassica rapa* sequencing project, the next step is to investigate functions of individual genes/proteins. For *Arabidopsis thaliana*, large amounts of protein–protein interaction (PPI) data are available from the major PPI databases (DBs). It is known that Brassica crop species are closely related to *A. thaliana*. This provides an opportunity to infer the *B. rapa* interactome using PPI data available from *A. thaliana*. In this paper, we present an inferred *B. rapa* interactome that is based on the *A. thaliana* PPI data from two resources: (i) *A. thaliana* PPI data from three major DBs, BioGRID, IntAct, and TAIR. (ii) ortholog-based *A. thaliana* PPI predictions. Linking between *B. rapa* and *A. thaliana* was accomplished in three complementary ways: (i) ortholog predictions, (ii) identification of gene duplication based on synteny and collinearity, and (iii) BLAST sequence similarity search. A complementary approach was also applied, which used known/predicted domain–domain interaction data. Specifically, since the two species are closely related, we used PPI data from *A. thaliana* to predict interacting domains that might be conserved between the two species. The predicted interactome was investigated for the component that contains known *A. thaliana* meiotic proteins to demonstrate its usability.

Keywords: *Brassica rapa, Arabidopsis thaliana*, interactome, protein–protein interaction, domain–domain interaction, meiosis

INTRODUCTION

For *Arabidopsis thaliana*, large amounts of protein–protein interaction (PPI) data are available from the major PPI databases (DBs; Galperin and Fernandez-Suarez, 2012), for example BioGRID (Stark et al., 2006) and IntAct (Aranda et al., 2010). The volume of these PPI data continues to increase with information from recently published articles (Arabidopsis Interactome Mapping Consortium, 2011). Assuming the same rate of interaction as in budding yeast, researchers estimate that the protein products of the *A. thaliana* genome participate in approximately 200,000 PPIs, a large proportion of which are yet to be validated (Lin et al., 2009). Therefore, efforts have been made to predict PPIs at the level of the entire *A. thaliana* genome, i.e., to produce a predicted interactome (Geisler-Lee et al., 2007; Cui et al., 2008; Morsy et al., 2008; Lee et al., 2010; Lin et al., 2010; Gu et al., 2011). Broadly speaking, two types of strategies can be applied. One approach is based on functional conservation between orthologous proteins, so called "interologs," where *A. thaliana* protein orthologs in other species are first predicted and interacting orthologs reveal possible interactions in *A. thaliana*. An example of this type of work was reported by Geisler-Lee et al. (2007), where they surveyed PPI data in budding yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), fruitfly (*Drosophila melanogaster*), and human (*Homo sapiens*), and built an interactome based on orthologs predicted using InParanoid (Ostlund et al., 2010). This interactome is now in version 2.0 and distributed with the latest TAIR 10 release (Lamesch et al., 2012). Software tools and web servers have now been made available to enable researchers to implement the "interologs" strategy, for example see Gallone et al. (2011). The second strategy does not rely on any other species, but solely on genomic/proteomic/transcriptomic features of *A. thaliana* (Cui et al., 2008; Brandao et al., 2009; Lin et al., 2010; Gu et al., 2011). For example, in the work by Lin et al. (2009), 14 features including gene expression and domain interactions were extracted to construct positive/negative training sets, and support vector machines were built to recognize the "pattern" of interaction. Normally this type of strategy is more computationally demanding, as it needs to employ machine learning techniques in an iterative manner.

Following the production of an interactome for the model plant *A. thaliana*, the next challenge is to develop similar interactomes for crop plants. The close relationship between *Brassica* crop species and *A. thaliana* (Lagercrantz et al., 1996; Trick et al., 2009; Wang et al., 2011) provides an opportunity to infer the *Brassica rapa* interactome by utilizing the substantial amount of PPI data available for *A. thaliana*. Despite large amounts of experimental and predicted PPI data for *A. thaliana*, as of June 2012, no interactions were recorded in the NCBI Entrez gene DB for *Brassica* sub-species (Taxid 3705). Here, we have constructed the inferred *B. rapa* interactome based on *A. thaliana* PPI data from two resources: (i) *A. thaliana* PPI data from three major DBs, BioGRID, IntAct and TAIR; and (ii) ortholog-based *A. thaliana* PPI prediction data (Geisler-Lee et al., 2007). Linking between *B. rapa* and *A. thaliana* was accomplished in three ways: (i) ortholog prediction

using InParanoid, (ii) identification of gene duplications in the Plant Genome Duplication Database (PGDD; Tang et al., 2008), and (iii) BLAST sequence similarity search. In addition, we followed a complementary approach, by looking at the specificity of PPI data at the level of domains. Domains are evolutionarily conserved protein subunits and earlier studies have shown that their interactions are also conserved across species, in a manner that is more conserved than the PPIs themselves, and that these domain pairs can be used as building blocks of the PPIs (Itzhaki et al., 2006; Schuster-Bockler and Bateman, 2007). Here we used the repertoire of domain–domain interactions (DDIs) inferred from *A. thaliana* PPI data, using the message-passing (MP) algorithm (Iqbal et al., 2008) to predict novel protein interactions in *B. rapa*, as well as to validate and examine the specificity of PPIs predicted using other orthology-based methods mentioned above. We also compared and combined these DDI data with experimentally observed and computationally predicted interacting domain data from the Database of Protein Domain Interactions (DOMINE; Yellaboina et al., 2011). Briefly, Pfam domains were assigned to each *B. rapa* protein using the HMMER software (Finn et al., 2010, 2011). By combining the MP algorithm with extant information based on DOMINE, we were able to predict PPIs from protein domain information.

In constructing the interactome, special attention was paid to the fact that *B. rapa* and *A. thaliana* genes/proteins do not necessarily follow a simple one-to-one relationship. Although sequencing of the *B. rapa* genome has confirmed its almost complete triplication relative to *A. thaliana*, since formation of the postulated original hexaploid ancestor, substantial gene loss (fractionation) has occurred, and *B. rapa* contains 41,174 identified protein-coding genes compared with 33,602 in *A. thaliana* (Wang et al., 2011; Lamesch et al., 2012). In addition, it is worth noting that of a total of approximately 17,000 *B. rapa* gene families, only 5.9% appeared to be lineage-specific, with 93% shared with *A. thaliana* (Wang et al., 2011). When considering the possibility of functional divergence of genes which are duplicated/triplicated in *B. rapa* relative to *A. thaliana*, it is also worth noting that duplicated genes encoding products which interact with other proteins or are part of networks may be expected to be less likely to diverge than those which are less well connected (Zhang et al., 2005).

The inferred *B. rapa* interactome presented here, together with the *B. rapa* (Chiifu-401-42) genome sequence (Wang et al., 2011), provide a useful starting point for functional PPI studies and knowledge transfer from the model plant *A. thaliana* to *Brassica* crop species. One such example is the EU PP7 project MEIOsys (Systematic Analysis of Factors Controlling Meiotic Recombination in Higher Plants), which is aimed at identifying factors controlling crossover frequency and distribution in higher plants. This project uses affinity-based techniques to isolate meiotic protein complexes from *Brassica oleracea* for analysis by mass spectrometry (Osman et al., in press). For this, the *B. rapa* (Chiifu-401-42) genome sequence and the predicted interactome presented in this paper have already proved to be valuable resources, facilitating the screening of *B. oleracea* peptides for protein identification and the identification of possible PPIs. As such, we believe that the predicted interactome is also a useful resource for the wider *Brassica* research and crop-breeding community.

MATERIALS AND METHODS

ACCESSING PPI DBs

Usually PPI DBs provide a web-interface, where an individual or list of protein/gene IDs can be used to query the DB. Some DBs can also be downloaded in a customized format for further investigation, e.g., the Database of Interacting Proteins (Xenarios et al., 2002). An increasing number of DBs also provide a version that complies with the Proteomics Standards Initiative – Molecular Interaction (PSI-MI) standard format (Kerrien et al., 2007). However, implementations of the PSI-MI format differ slightly from each other, which limit the reusability of existing codes. As a recent effort, PSI common query interface (PSICQUIC) was introduced (Aranda et al., 2011), which aims at providing a uniform query access for different PPI DBs. Queries to supporting DBs can be performed over the web in a manner as if it was a single DB. However, querying and compiling these DBs remains a challenging task, especially for large data sets, because, for example, different DBs use different unique IDs.

Three major *A. thaliana* PPI DBs were used in the current study: BioGRID, IntAct, and TAIR. The most recent versions at the time of the analysis were BioGRID 3.1.87, IntAct 2012-03-15, and TAIR 10. The DBs were presented according to different interpretations of experimental results. The simplest case is yeast two-hybrid, where two proteins form a direct binary/pairwise interaction. Other methods of analysis, for example co-immunoprecipitation, can identify protein complexes, which result in more complicated forms of representation of the DB. A popular choice of representation is the spoke model, in which such experimental results are interpreted as a set of binary interactions between the bait protein and co-precipitating proteins. Another form of representation, so called "matrix form," assumes all co-precipitating proteins form binary interactions with each other. But this representation is considered less accurate (Bader and Hogue, 2003; Lysenko et al., 2009). Examples of both can be seen in **Figure 1**. In the current study, we downloaded all DBs in the PSI-MI TAB format, which uses the spoke model (Kerrien et al., 2007).

PPI DATA COMPILATION

An important aspect of a PPI is its detection method. Accordingly, if the same binary interaction was detected using different methods, or in different studies, all three DBs mentioned would list these binary interactions as separate entries. An example of this is seen in **Figure 2**. Although the detection method provides extra information for the DBs, in the current circumstances it leads to duplication and was thus removed during our data preparation. In fact, during the pre-processing of these DBs, we kept only the information of the two partners involved in the binary interaction, along with the original publication where the experiments appeared (i.e., PMID number); all other information provided with the PSI-MI TAB format was removed.

The compiled *A. thaliana* PPI data (denoted by D1) consists of 16,644 binary interactions from 1,398 published research articles. The total number of proteins involved in D1 is 6,451, which does not include splicing variants. The contributions of the three source DBs to D1 can be seen in **Figure 3**. BioGrid is the largest source of interactions, followed by IntAct. Although TAIR is the smallest

A

#	Aliases 'A'	Aliases 'B'	Publication Identifier	Interaction Type	Interaction AC	Expansion Method
1	Flavoprotein subunit 1 of complex II; At5g66760; MSN2.16; [+1]	At5g09600; sdh3; SDH3-1; [+5]	12970493	physical association	EBI-533336	Spoke
2	Flavoprotein subunit 1 of complex II; At5g66760; MSN2.16; [+1]	Iron-sulfur subunit of complex II; At5g40650; MNF13.21; [+2]	12970493	physical association	EBI-533336	Spoke
3	Flavoprotein subunit 1 of complex II; At5g66760; MSN2.16; [+1]	Iron-sulfur subunit of complex II; At3g27380; K1G2.9; [+1]	12970493	physical association	EBI-533336	Spoke
4	Flavoprotein subunit 1 of complex II; At5g66760; MSN2.16; [+1]	At1g47420; F16N3.26; ump6_arath	12970493	physical association	EBI-533336	Spoke
5	Flavoprotein subunit 1 of complex II; At5g66760; MSN2.16; [+1]	At1g08480; AT1G08480; q941a6_arath	12970493	physical association	EBI-533336	Spoke

B

= refTypeAc	MI:0356
= refType	identity
= id	EBI-533336
= dbAc	MI:0469
= db	intact

participantList

participant (6)

	= id
1	189398
2	189399
3	189400
4	189401
5	189402
6	189403

FIGURE 1 | Different interpretations of experimental data result in different DB formation. (A) Using the spoke model, interaction EBI-533336 shows as five binary interactions on the IntAct website; **(B)** the same interaction shows in PSI-MI XML format as an interaction with six participants.

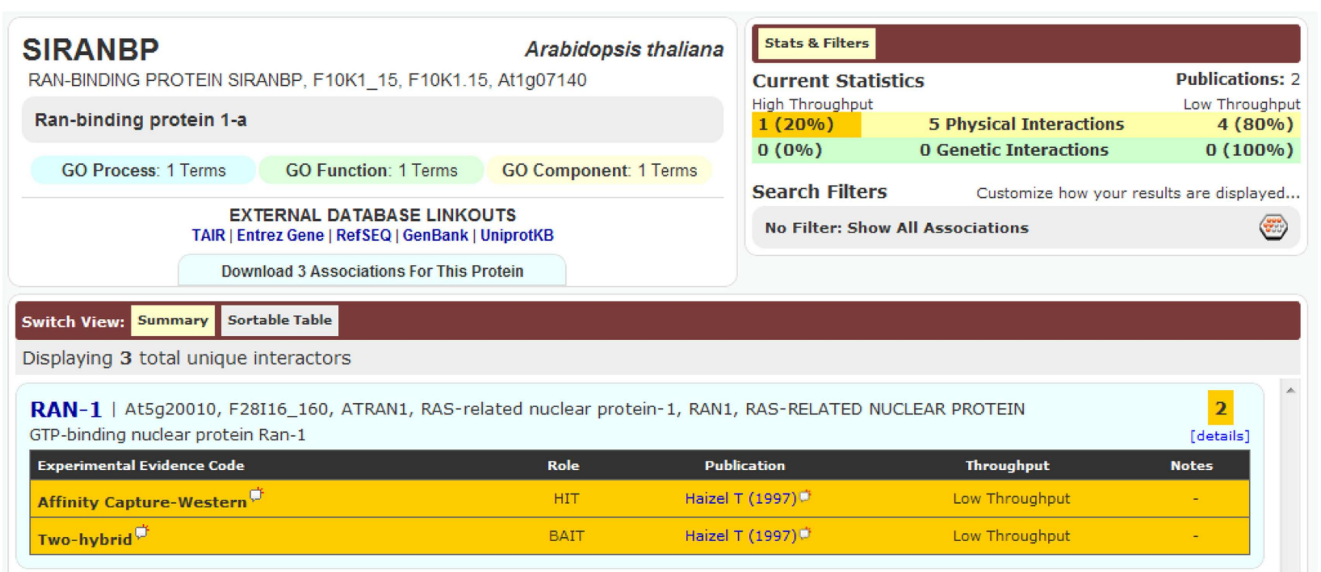


FIGURE 2 | Interaction between SIRANBP (At1g07140) and RAN-1 (At5g20010) was recorded in BioGRID as two separate entries because they were detected using two different methods, despite being from the same publication.

DB, it contains records complementary to the two main resources, and so is still valuable. It is interesting to note that although there were significant overlaps among the three DBs in terms of binary interactions and interacting proteins (**Figures 3A,B**), it seems that the overlap in terms of publication is not significant (**Figure 3C**). This highlights the importance of multiple data sources in the PPI prediction.

Besides experimentally verified PPIs from the three DBs, predicted PPIs were also used in our study. Geisler-Lee et al. (2007) studied PPIs in four model organisms, and predicted 72,266 PPIs based on interologs. Thus far, with information from recent publications, 3,453 of these have been confirmed. For example, the predicted interaction between AtSPO11-1 (At3g13170) and AtPRD1 (At4g14180) was later confirmed by yeast two-hybrid

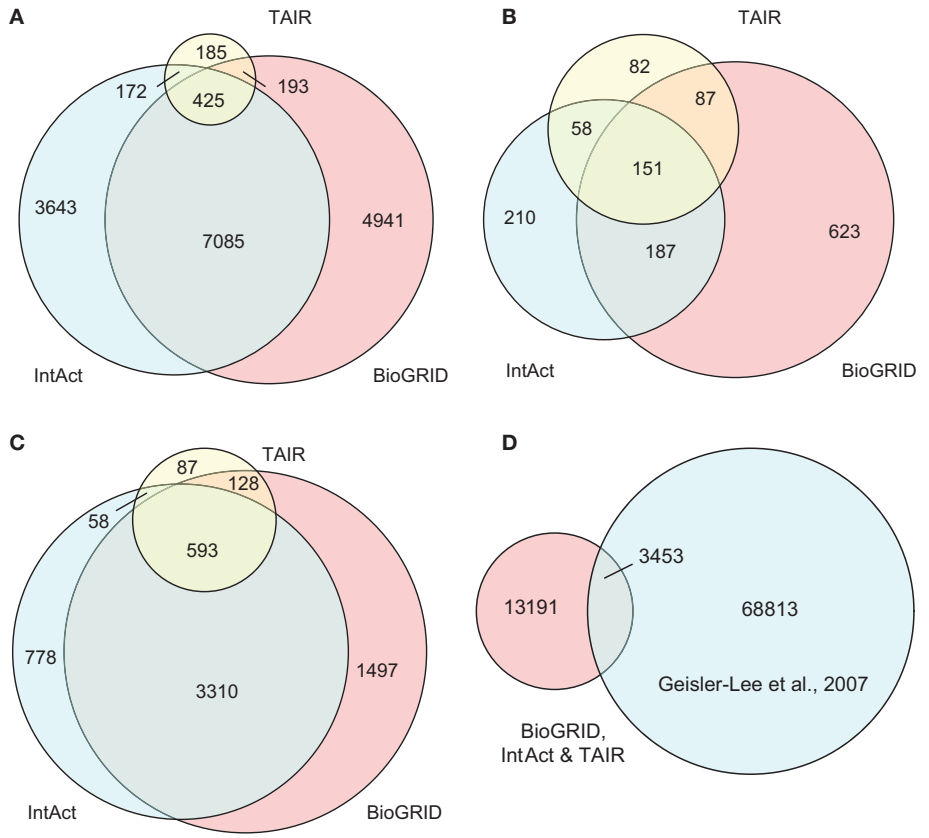


FIGURE 3 | Contributions of the three source DBs to the compiled data set (D1) in terms of: (A) binary interactions, (B) publication, and (C) interacting proteins. The largest contribution comes from BioGrid, although the overlap among the three is significant. **(D)** A small part of interaction predictions made by Geisler-Lee et al. (2007) were confirmed in PPI DBs, while the remaining form the complementary source of PPI data in our study. Venn diagrams showing correct proportions were drawn using Venn Diagram Plotter Pacific Northwest National Laboratory, http://omics.pnl.gov/software/VennDiagramPlotter.php

assay (De Muyt et al., 2007) and recorded under ID EBI-1540718 in IntAct. The remaining 68,813 PPIs that are yet to be confirmed were used in the current study as a complementary PPI source, denoted by D2. The relationship between the data compiled from the three DBs (i.e., D1) and the prediction made by Geisler-Lee et al. (2007) can be seen in **Figure 3D**.

LINKING THE TWO SPECIES

The objective of the present research is to use known *A. thaliana* PPI data in order to expand the predicted *B. rapa* interactome. It is vital that the links between the two species are established correctly. An obvious way of achieving this is to identify orthologs between them. Using InParanoid, a total of 17,859 orthologous clusters were detected, which contain 18,830 and 21,873 proteins for *A. thaliana* and *B. rapa* respectively. Note that the number of orthologous clusters is less than the number of proteins for both species. This is a desirable feature as it may be indicative of possible gene duplication events within each species. Thus, in terms of DB implementation, this creates multi-to-multi relationships within the orthologous clusters.

In general, ortholog prediction methods can be classified into two broad categories: methods based on pairwise alignments, for example InParanoid, and methods based on phylogenetic trees (Kuzniar et al., 2008). The pairwise alignment methods have been found to outperform tree-based methods (Ostlund et al., 2010), which is why they were adopted in the current study. A complementary way of identifying related proteins, however, is to look at synteny and collinearity. In fact, since the release of the *B. rapa* genome sequence, several comparative genomics DBs (Lyons and Freeling, 2008; Tang et al., 2008; Tang and Lyons, 2012) have made use of the sequence. One of these, PGDD (Tang et al., 2008), identified 682 gene/protein blocks between *A. thaliana* and *B. rapa*, each of which consists of the same number of genes/protein from both species. PGDD allows a single gene/protein to appear in several different blocks. This effectively creates a multi-to-multi relationship. The total number of proteins covered in PGDD is 18,207 and 27,536 for *A. thaliana* and *B. rapa* respectively. Combining InParanoid and PGDD, a "bridging" DB was obtained, covering 21,624 and 31,423 proteins for *A. thaliana* and *B. rapa* respectively.

The total number of protein-coding genes released in the *B. rapa* sequencing project is 41,173. This leaves 9,750 *B. rapa* proteins that are not associated with any partners in *A. thaliana*. Therefore, we performed a BLAST similarity search using these

9,750 proteins against *A. thaliana* with a cut-off *e*-value of 1.0e−6. It was found that 7,307 had a hit in *A. thaliana* and interestingly, 1,376 hits reported an *e*-value of 0 (i.e., too small to report). These one-to-one data were then added to the previously compiled set to form the final "bridging" DB, denoted as D3. *B. rapa* proteins not covered by D3 account for approximately 5.93% (2,443/41,173). This is in agreement with a previous study which found that 95.8% of gene models have a match in at least one of the public protein DBs (Wang et al., 2011).

B. RAPA PROTEIN DOMAIN ASSIGNMENTS

The total number of *B. rapa* proteins covered by D3 was 38,730, which still falls short of the *B. rapa* total of 41,173. To predict possible interactions for those *B. rapa* proteins that do not have counterparts in *A. thaliana*, as well as to complement the above mentioned methods of interactome prediction, we used other means of prediction in building the final interactome, i.e., looking at the level of DDIs. This not only increases the coverage of the interactome, but also gives a higher level of confidence. In addition, it provides more detailed information concerning which domains are potentially mediating the protein interactions. For this purpose, *B. rapa* protein domain assignments and interacting domain data (inferred using PPI data from *A. thaliana* as well as known domain interactions) can be used to predict possible protein interactions. HMMER (Finn et al., 2011) was used to search *B. rapa* protein sequences against the Pfam-A DB (Finn et al., 2010), using stringent criteria (*e*-value = 1.0e−10). As a result, 3,482 Pfam-A domains were assigned to 27,452 *B. rapa* proteins. On average, we had 1.43 domains assigned to each *B. rapa* protein. This is comparable with the TAIR Pfam annotation (1.41 domains/protein).

DOMINE: THE INTERACTING DOMAIN DATABASE

The DOMINE DB (Yellaboina et al., 2011), which contains both experimental and predicted DDIs, was used in combination with the above mentioned *B. rapa* domain assignments. Here we used only known (i.e., observed) and high confidence predictions from DOMINE, which accounts for 8,173 unique interacting domain pairs. Known interacting domain data in DOMINE come from iPfam and 3did (Stein et al., 2011). With the release of Pfam version 26.0, additional entries were added. Fusing these entries together with DOMINE, we obtained 8,366 unique interacting domain pairs (denoted D4).

THE MP ALGORITHM AND TRAINING SETS

Since *B. rapa* and *A. thaliana* are closely related, it is reasonable to assume that some interacting domains are conserved between the two species. In order to predict novel interacting domains, we employed the MP algorithm (Iqbal et al., 2008). MP is a popular method in the statistical inference community and has been applied in many hard inference problems in many fields (Berendsen et al., 1995; Richardson and Urbanke, 2001). Given the set of interacting and non-interacting protein pairs and their domain assignments, the MP method models this data as a factor graph which has two types of nodes: variable nodes which are the domain–domain pairs, and function nodes which are protein pairs (either interacting or non-interacting). The function nodes put constraints on the underlying variable nodes, as follows:

- For an interacting protein pair, at least one of the underlying domain pairs must be interacting.
- For a non-interacting protein pair, none of its underlying domain pairs should be interacting.

Given the existence of false positives in PPI data and our hypothesized negative data, the above constraints need to be "softened" to take into account the errors in the interaction map. This error is incorporated via an additional parameter ε, which ranges between 0 and 1 and quantifies our confidence in the PPI data ($\varepsilon = 0$ means the PPI network is 100% reliable). Another parameter, the *a priori* probability (β), takes into account any prior knowledge of the DDIs. Given the above constraints, the goal is to assign 1 s and 0 s to the domain pairs such that the maximum number of constraints is satisfied. For that purpose, under this factor graphical modeling framework, a powerful statistical inference method, belief propagation (BP), is employed to infer the domain–domain interaction probabilities.

Belief propagation performs exact inference if the underlying graph is a tree, which corresponds to the global minimum of a function, called Bethe free energy (Yedidia et al., 2005). Bethe free energy is a function of beliefs, which in our case are domain interaction probabilities. It has been shown that, even in the case of graphs with cycles, on convergence solutions obtained by BP correspond to the local minimum of Bethe free energy. Hence, as in Iqbal et al. (2008), an inference scheme using BP is used here by minimizing Bethe free energy which helps to estimate two known parameters in our model, i.e., ε and β. For details of the MP algorithm and BP, see Iqbal et al. (2008).

The input to the algorithm is an interaction map among a set of proteins, and a set of domain assignments for the relevant proteins. The output is a list of probabilities of interaction between each pair of domains. Domain assignments for *A. thaliana* were taken from the Pfam DB (Finn et al., 2010). The PPI data compiled previously were used as positive inputs. However, not all interaction detection methods accurately detect binary interactions, for example HTP (Lin et al., 2009). To minimize false positives and also to reduce the computational burden, only a subset of D1 (yeast two-hybrid data) was used (denoted D1-sub). The MP algorithm also requires negative samples, i.e., non-interacting protein pairs. It is difficult to build an accurate set of negative samples because it is inherently impossible to exclude non-interacting protein pairs with certainty, and hence such results do not usually appear in the literature. Researchers have used various methods for constructing "hypothetical" non-interacting protein pairs, for example those based on randomness or proteins separated in different subcellular localizations (Xu et al., 2010). In the current study, we adopt a random approach, with additional stricter rules. Two random proteins were taken to be non-interacting if: (i) they do not appear in D1, (ii) their domain pairs do not appear in D4, (iii) they must have the same GO term in terms of cellular component, and (iv) the absolute value of their co-expression is less than 0.4. The last two restrictions ensure that expression patterns

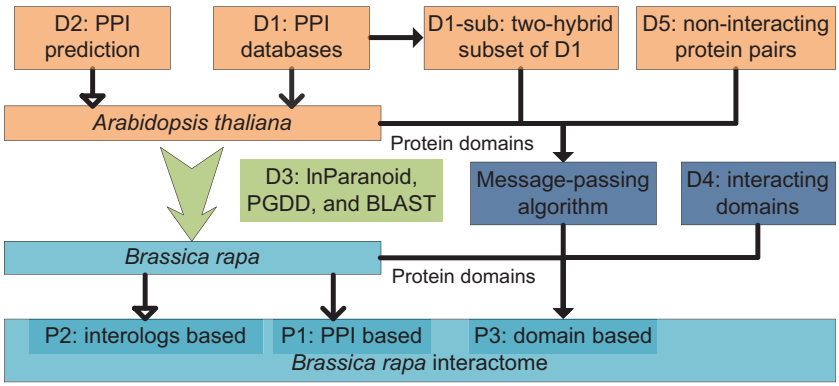


FIGURE 4 | Overview of data and methods used in the present study. Three sets of interaction predictions were obtained: P1, PPI based; P2, interologs based; and P3, domain-based.

of the two proteins/genes do not imply interaction (Allocco et al., 2004). The gene expression data were from ATTED-II (Obayashi and Kinoshita, 2010). As a result, 25,246 domain pairs and 9,076 positive/negative training samples were fed into the algorithm to make interaction domain predictions. The negative samples were denoted D5.

RESULTS AND DISCUSSION

An overview diagram illustrating data and methods used in the present study is shown in **Figure 4**. Three sets of *B. rapa* interaction predictions were obtained: PPI based interaction (denoted P1), interologs based interaction (P2), and interacting domain-based interaction (P3). P1 and P2 were obtained using physical and predicted PPI data in D1 and D2, and the "bridging" DB D3. P3 were obtained using *B. rapa* protein domain assignments and the interacting domain data, which combine both "generic" known/high confidence interacting domain data in D4, and the *A. thaliana* "specific" interacting domain predictions using the MP algorithm and D1-sub/D5.

Restriction rules were applied to P3 to reduce the number of predictions and also increase the reliability: (i) two proteins in the pair need to share the same Gene Ontology (GO) cellular component terms in order for the domain-based prediction to take effect; (ii) if not predicted to be interacting in P1 or P2, a protein pair needs to have more than one interacting domain pairs; (iii) if predicted to be interacting in P1 and P2, a protein pair can have only one interacting domain pair. GO terms were assigned to *B. rapa* sequences using Argot2 (Fontana et al., 2009) with a stringent "internal confidence" value of 0.55, based on sequence similarity (UniProtKB/Swiss-Prot) and protein domain information (Pfam-A).

NOVEL INTERACTING DOMAINS

Two parameters had to be fine-tuned for the MP algorithm to work correctly: the *a priori* probability, β, and the degree of reliability of the interaction datasets available for the inference, ε (Iqbal et al., 2008). Different values of β and ε were tested using training samples D1-sub and D5 to minimize Bethe free energy (Yedidia et al., 2005) as in **Figure 5**. For β values ranging from 0.1 to 0.8, a minimum Bethe free energy was reached for $\beta = 0.2$ (**Figure 5A**). Examining details of the minimum point, it was found that ε is equal to 0.02 (**Figure 5B**). These two values were taken forward to produce the final results.

The algorithm assigned probabilities of interactions to all 25,246 domain pairs. Special attention was paid to determine the cut-off value; on the one hand, a higher cut-off probability produces more reliable results but conversely it will produce fewer interacting domains, which does not fully represent the training sample. In the present study, a cut-off of 0.85 was used to select 2,389 high confidence interacting domain predictions. It was found that among these 2,389 domain pairs, 182 were also present in D4 (i.e., they were either physical interacting domain pairs observed in iPfam/3did, or high confidence predictions in DOMINE). A large proportion of these domain pairs (2,283) are the only domain pair in their respective protein pair in the positive training set D1-sub. They were successfully recognized; for example, domain pair PF01627 and PF03962 in protein pair AHP2 (At3g29350) and AtMND1 (At4g29170). (Interactions between AHP2 and AtMND1 were recorded under ID BIOGRID: 337481 and EBI-1555097). These predictions were considered unique contributions of the MP algorithm, and possibly conserved between *A. thaliana* and *B. rapa*. Combining results from the MP algorithm and D4, 10,573 unique interacting domain pairs were used to make prediction P3.

THE PREDICTED INTERACTOME

P1, P2, and P3 contain 77,073, 316,128, and 364,768 predicted interactions respectively; all three datasets gave a total number of 740,565 unique predicted interactions (the predicted *B. rapa* interactome, denoted by P-all). The relationship among the three sets is shown in **Figure 6A**. The histogram of the number of interacting partners for each protein in P-all is shown in **Figure 6B**. The peak in **Figure 6B** is the first bin (i.e., degree < 10), which contains nearly half of proteins present in P-all (10,254 vs. 20,677). It is also worth noting that there are a small number of protein "hubs" with interacting partners between 700 and 1,774. These hubs may be important because they link the network together. On average, each protein in P-all interacts with 71 partners, which

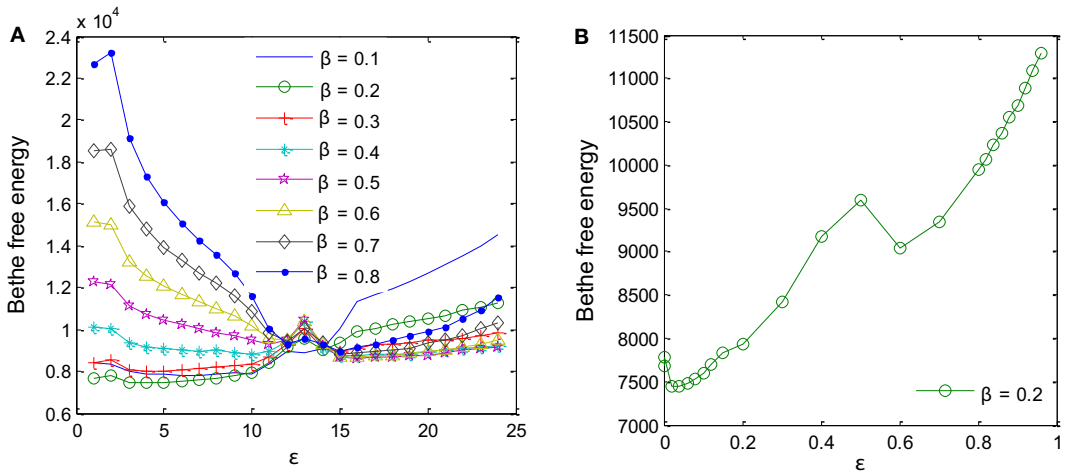


FIGURE 5 | Parameters for the MP algorithm had to be fine-tuned. (A) *A priori* probability $\beta = 0.2$ produces the minimum Bethe free energy. **(B)** For $\beta = 0.2$, minimum Bethe free energy was reached at $\varepsilon = 0.02$.

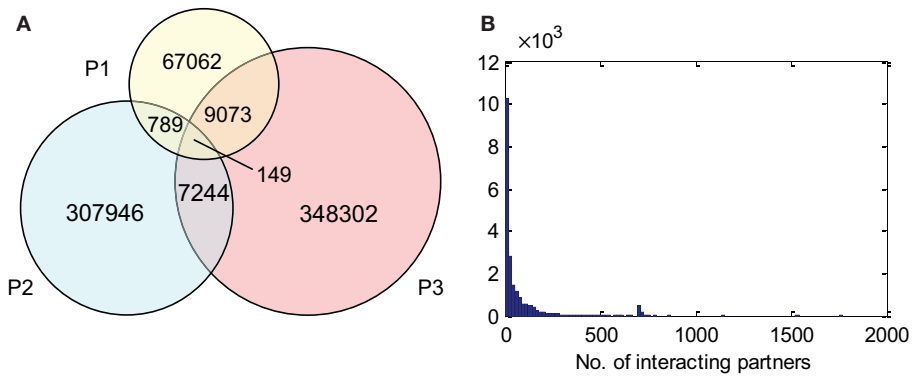


FIGURE 6 | (A) The relationship between different prediction subsets. **(B)** Histogram showing degree distribution (number of interacting partners for each protein) for P-all.

is higher than the estimation that a single protein interacts with about 5–50 proteins (Deng et al., 2002). The group of the 10 most connected hubs of P-all are shown in **Table 1**, which based on their known functions is not unexpected. Furthermore, some in this group do not have symbols, indicating that they have not been experimentally identified.

The three sets of PPI predictions constitute two levels of confidence of the predicted interactome. The high confidence prediction (Phc) has support from at least two sources of evidence, the low confidence prediction (Plc) has support from only one. Phc and Plc contain 17,255 and 723,310 interactions respectively. Some structural properties depicting P-all and the two different confidence level sub-networks were calculated using R package igraph (Csardi and Nepusz, 2006), as seen in **Table 2**. In all three cases there were large numbers of self-interactions. While these self-interactions constitute an important aspect of the interactome, they were removed from further analysis of the network structure. Interestingly, the network diameter (largest distance between two proteins) and the averaged shortest path length for Phc were significantly larger than those of Plc. This suggests that Phc contains a large sparsely connected network. It was also interesting to note that the average number of interacting partners, transitivity (i.e., clustering coefficient) and centralization of Plc are dramatically larger than those of Phc. This indicates that although Plc may contain less confident predictions, it is still useful in that it gives a densely connected network that contains all possible interactions.

INTERACTOME COVERAGE

Using Argot2 (Fontana et al., 2009), 66% of all *B. rapa* protein-coding sequences (27,179/41,173) were assigned at least one GO term. We then categorized these proteins (i.e., genome) and the proteins from P-all (i.e., interactome) in terms of GO plant slim categories using AgBase (McCarthy et al., 2006). The results are shown in **Figure 7**.

From **Figure 7** it is evident that in every category the number of proteins present in the interactome (purple line) follows

Table 1 | Top 10 interaction hubs of P-all and their *A. thaliana* counter parts.

B. rapa	Interactions	*A. thaliana*	Resources	Symbols	Description
Bra014387	1774	At2g47610	I		Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
		At3g62870	P		Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
Bra003119	1540	At5g52640	P	HSP81-1	Heat shock protein 90.1
Bra009542	1144	At1g65350	I	UBQ13	Ubiquitin 13
		At3g09790	I, P	UBQ8	Ubiquitin 8
		At4g02890	I	UBQ14	Ubiquitin family protein
		At4g05320	I	UBQ10	Polyubiquitin 10
		At5g03240	I, P	UBQ3	Polyubiquitin 3
		At5g20620	I	UBQ4	Ubiquitin 4
		At5g37640	I	UBQ9	Ubiquitin 9
Bra024839	867	At2g01950	B	VH1	BRI1-like 2
Bra024840	867	At2g01950	B	VH1	BRI1-like 2
Bra016839	794	At1g11320	P		Unknown protein
Bra032392	759	At1g30470	P		SIT4 phosphatase-associated family protein
Bra021474	755	At3g02200	P		Proteasome component (PCI) domain protein
		At5g15610	P		Proteasome component (PCI) domain protein
Bra013661	740	At4g22930	P	PYR4	Pyrimidin 4
Bra036269	738	At4g02410	B		Concanavalin A-like lectin protein kinase family protein

B, BLAST; I, InParanoid; P, PGDD.

Table 2 | Structural properties depicting the interactome P-all and two confidence levels of the sub-network: high confidence (Phc) and low confidence (Plc).

	Phc	Plc	P-all
No. of proteins	4,483	20,537	20,677
No. of interactions	17,255	723,310	740,565
No. of isolated proteins (ignore self-interaction)	155	50	54
No. of self-interaction	1,881	5,367	7248
No. of protein clusters	628	116	129
Diameter	32	10	11
Averaged neighbors	6.86	69.92	70.93
Averaged shortest path length	10.64	3.62	3.61
Transitivity	0.58	0.75	0.75
Centralization	0.01	0.08	0.08
Density	1.72E−3	3.43E−3	3.46E−3

the number of proteins in the genome (green line), and that in most categories the interactome/genome ratio is greater than 50% (bars). There are several categories with very small interactome/genome ratios, for example, cell–cell signaling and embryo development in the biological process category (highlighted by asterisk in **Figure 7A**), cell wall and nucleolus in the cellular component category (**Figure 7B**), and receptor binding in the molecular function category (**Figure 7C**). In these categories proteins do not count for a large number in either the genome or the interactome. On the other hand, most proteins from the interactome or genome fall into several specific GO slim categories, and have relatively high interactome/genome ratios. Those categories include metabolic process in biological process (highlighted by bars with solid borders in **Figure 7A**), intracellular and cytoplasm in the cellular component (**Figure 7B**), and catalytic activity in the molecular function (**Figure 7C**). From the above analysis, we concluded that the interactome is generally representative of the *B. rapa* genome. Given that a total number of 20,677 proteins are present in P-all, the protein coverage of the interactome is about 50%.

It is difficult to estimate the interaction coverage of the interactome. However, assuming the same rate of interaction as in *A. thaliana* (Lin et al., 2009), we estimated that there would be approximately 220,000 interactions for approximately 21,000 proteins in P-all. Thus the predicted interactome, with more than 740,000 interactions, is likely to have a very high false positive rate. On the other hand, the high confidence Phc contains 17,255 unique interactions, which would be coverage of approximately 78%, and thus is likely to be missing many true interactions. It is rare that, in terms of predicted interactomes, predictions match expectations exactly. For example, in PAIR (the predicted *Arabidopsis* interactome resource; Lin et al., 2009, 2010), the high confidence predictions are expected to cover 29.02% of the

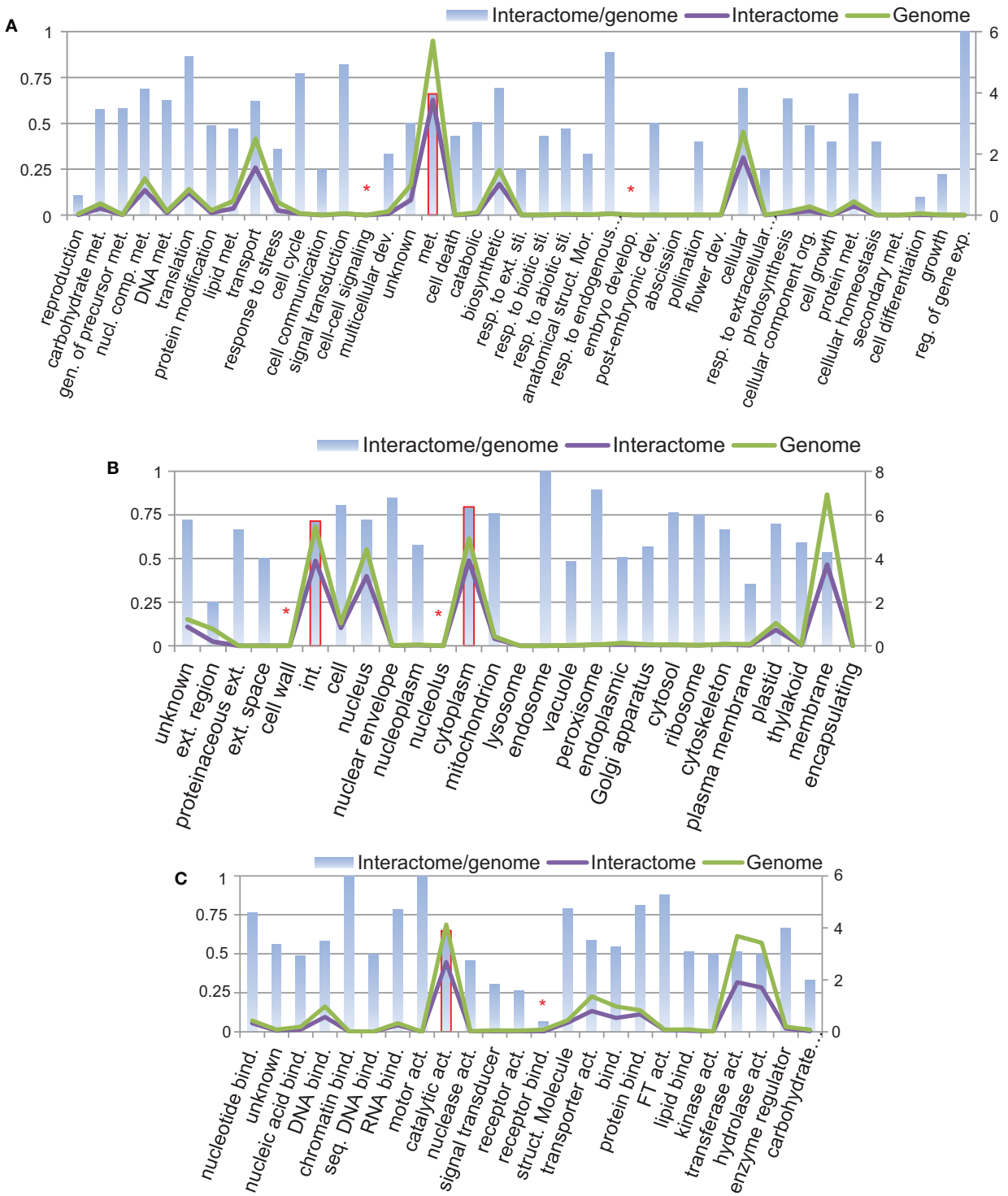


FIGURE 7 | Protein categories of the interactome in comparison with the whole *B. rapa* genome using GO plant slim categories: (A) biological process, (B) cellular component, and (C) molecular function. Left vertical axis shows the interactome/genome percentage. Right vertical axis shows protein counts in units of 1,000.

Abbreviations: act, Activity; bind, binding; comp, compound; dev, development; exp, expression; ext, extracellular; gen, generation; int, intracellular; met, metabolic; mor, morphogenesis; nucl, nucleobase; org, organization; reg, regulation; resp, response; seq, sequence-specific; sti, stimulus; struct, structure.

entire interactome. However, in the present study of *B. rapa* the problem of coverage/false positive rates seems to be exaggerated. The reasons for this are twofold: (i) Because of gene duplication/loss, genes of *A. thaliana* and *B. rapa* form a multi-to-multi relationship. However, in the interologs based prediction (P1 and P2), it is barely possible to rule out any predicted interactions (Pennisi, 2012). (ii) In the domain-based prediction P3, protein domains and GO terms were derived through computational predictions. However, parameters of the prediction algorithms, e.g., InParanoid/HMMER need to be fine-tuned to achieve higher accuracy. In addition, we used all physical interacting domain data from DOMINE, but it is possible that certain domains may only be interacting under certain cellular conditions. To address coverage/false positive rates issues, experiments need to be carried out to test predicted interactions in order that rules may be established to exclude any false positive predictions.

GENE DUPLICATION AND THE "BRIDGING" DB

The source data of the predicted *B. rapa* interactome came from *A. thaliana*. Thus it is vital that the relationships between the two genomes were correctly defined. Importantly, consideration must be given to the fact that there has been almost complete triplication of the *B. rapa* genome relative to *A. thaliana,* although since formation of the postulated original hexaploid ancestor, substantial gene loss has occurred (Wang et al., 2011). In this and the following sections we use known *A. thaliana* meiotic genes as an example to discuss gene duplication and its effect on the *B. rapa* meiosis network.

Meiosis is a key biological process that underpins sexual reproduction. During meiosis, a single round of DNA replication is followed by two rounds of nuclear division to produce four haploid gametes. Many genes/proteins participate in meiosis, for example, see reviews (Ma, 2006; Hultén, 2010; Osman et al., 2011). Here we used the list of 71 meiotic genes presented in (Yang et al., 2010),

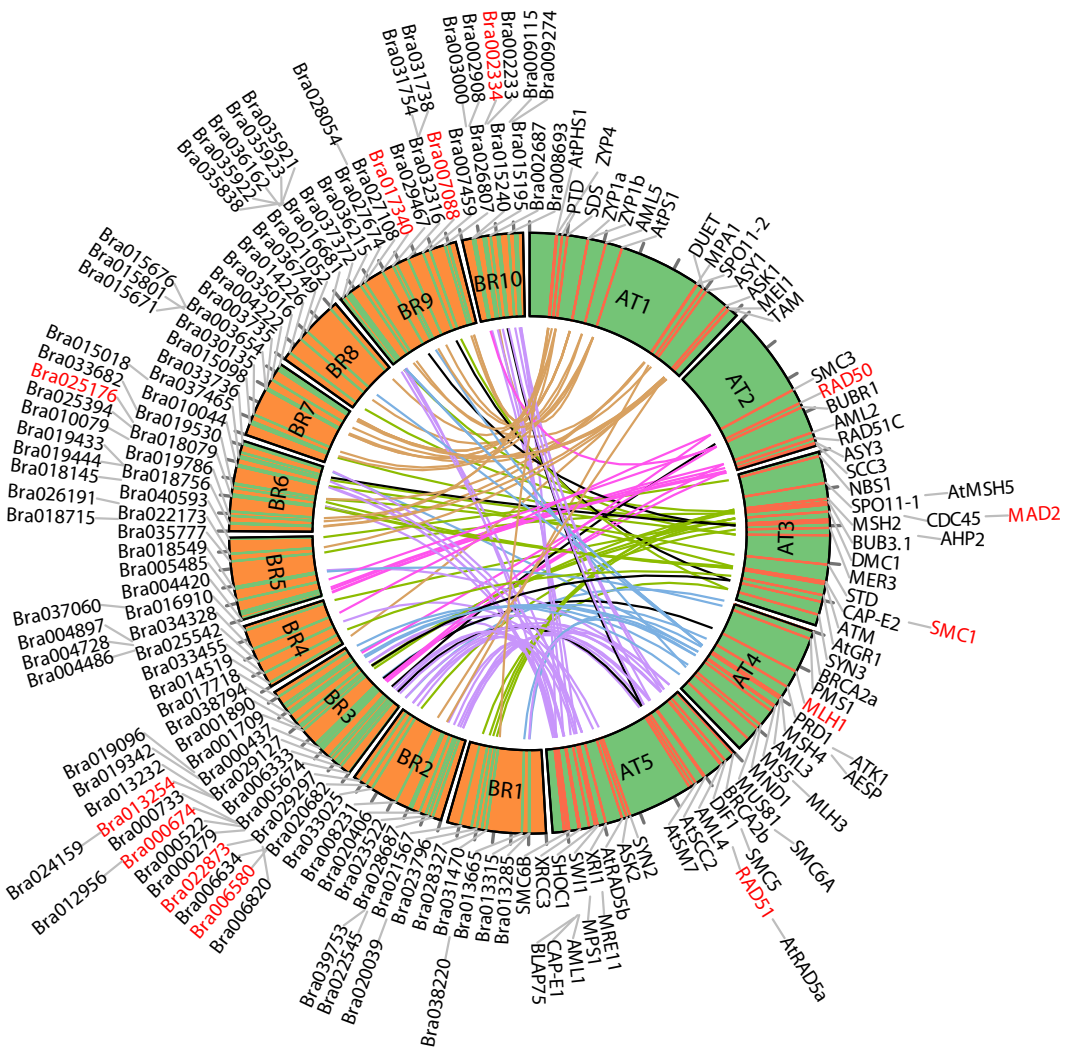


FIGURE 8 | Chromosome positions of 72 selected *A. thaliana* meiotic genes and their counterparts in *B. rapa*. The "bridging" DB represents possible collinearity and gene duplications. Genomes are arranged clockwise: *A. thaliana* chromosomes 1-5 (AT1 – AT5); *B. rapa* chromosomes 1-10 (BR1-BR10). Figure generated using Circos (Krzywinski et al., 2009).

Table 3 | Some known *A. thaliana* meiotic genes and their counterparts in *B. rapa*.

Protein locus	Gene name/ chromosome	Pfam domain	GO: biological process	GO: cellular component	GO: molecular function	Resources
At2g31970	*RAD50*	Rad50_zn_hook, AAA_23, SbcCD_C	Telomere maintenance; DNA repair; double-strand break repair; mitotic recombination; telomere capping	Nucleus; cytoplasm; Mre11 complex	Nuclease activity; ATP binding; zinc ion binding	
Bra022873	BR3			nucleus		I, P
At3g25980	*MAD2*	HORMA	Mitotic cell cycle spindle assembly checkpoint	Kinetochore; chromocenter	DNA binding; protein binding	
Bra017340	BR9	HORMA	Cell cycle		Protein binding	I, P
Bra025176	BR6		Cell cycle		Protein binding	P
At3g54670	*SMC1*	SMC_N, SMC_hinge	Chromosome segregation; sister chromatid cohesion; chromosome organization	Nucleus; chromosome; cohesin complex; chloroplast	Transporter activity; protein binding; ATP binding	
Bra007088	BR9	SMC_N, SMC_hinge, AAA_23		Chromosome		I, P
Bra013254	BR3	SMC_N, SMC_hinge, AAA_23		Chromosome		B
At4g09140	*MLH1*	DNA_mis_repair, HATPase_c	ATP catabolic process; mismatch repair; mitotic recombination; reciprocal meiotic recombination; pollen development; seed germination; fruit development; seed development	Nuclear chromatin; synaptonemal complex; nucleus; chiasma; MutLalpha complex; MutLbeta complex	ATP binding; ATPase activity; protein binding, bridging; mismatched DNA binding	
Bra000674	BR3	DNA_mis_repair, HATPase_c_3		Nucleus		I, P
At5g20850	*RAD51*	Rad51	DNA metabolic process; DNA repair; double-strand break repair; regulation of transcription, DNA-dependent; response to radiation; response to gamma radiation	Nucleus	Nucleotide binding; DNA binding; damaged DNA binding; protein binding; ATP binding; DNA-dependent ATPase activity; nucleoside-triphosphatase activity	
Bra002334	BR10	Rad51, AAA_25	DNA metabolic process	Nucleus		P
Bra006580	BR3	Rad51, AAA_25	DNA metabolic process	Nucleus		I, P

B, BLAST; I, InParanoid; P, PGDD.

Each group (separated by solid line) starts with an A. thaliana gene, followed by one or more B. rapa genes.

with the addition of *AtASY3* (At2g46980), recently described by the Birmingham meiosis group (Ferdous et al., 2012).

For ease of interpretation we have presented the relationships between the two species in a one-to-multi manner from the *A. thaliana* perspective, as shown in **Figure 8** and **Table 3**. **Figure 8** shows chromosome positions of 72 known *A. thaliana* meiotic genes and their "counterparts" in *B. rapa*. It is evident that in our "bridging" DB there are conserved collinear blocks between the two genomes, for example, between the end of *A. thaliana* chromosome 2 (AT2) and the start of *B. rapa* chromosome 5 (BR5). This is in agreement with observations by Wang et al. (2011). Furthermore, we modeled possible gene duplications of *A. thaliana* meiotic genes, for example those on AT5 migrating to BR2/BR3/BR6/BR10.

Table 3 gives some detailed information for several meiotic genes presented in **Figure 8**, where related genes from the two species are grouped together. Each group is led by an *A. thaliana* meiotic gene, followed by its *B. rapa* counterpart(s) and the inference resources. We also listed domain (Pfam) and GO term names for these genes/proteins where available. We can see that quite often the relationships were confirmed by more than one method/resource. Furthermore, most related proteins have a similar domain structure, for example AtMAD2 and its counterparts in *B. rapa* (highlighted in **Figure 8**). However, in groups containing *AtSMC1* and *AtRAD51*, it seems that *B. rapa* genes have additional functions compared to their counterparts in *A. thaliana* (i.e., additional AAA_23 and AAA_25 domains respectively). For GO terms, as we used stringent criteria, fewer GO terms were assigned to *B. rapa* proteins. However, assigned terms mostly agree with their counterparts in *A. thaliana*.

THE MEIOSIS NETWORK

The sub-network formed by putative *B. rapa* meiotic proteins was extracted from P-all (**Figure 9**) as an example to demonstrate the utility of the predicted interactome. From **Figure 9** it is obvious that there is a large number of putative *B. rapa* meiotic proteins which are sole copies of their *A. thaliana* counterparts. It is likely that these proteins are functionally identical to those in *A. thaliana*. Multi-copy proteins are also found and in some cases at least, their functions appear to have differentiated. For example, there are four

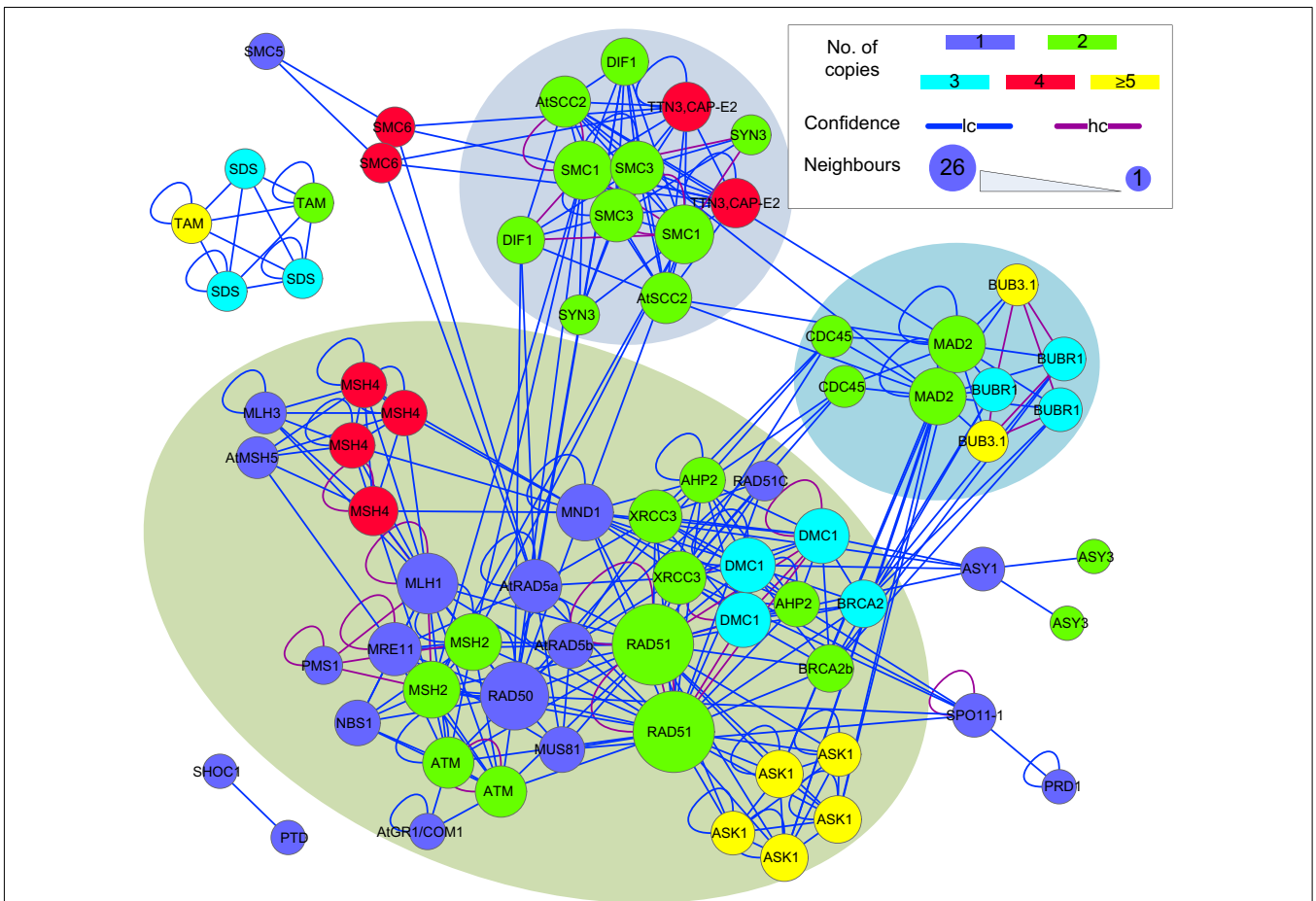


FIGURE 9 | A sub-network of the predicted interactome formed by putative *B. rapa* meiotic proteins. It is a densely connected network containing several hub proteins. *B. rapa* proteins are shown by their *A. thaliana* counterparts' names; their unique accession numbers are not shown. Node color denotes the number of copies that a single *A. thaliana* gene/protein has in *B. rapa*. The confidence of interaction prediction is presented by edge colors. Node size is set to be proportional to the number of interacting partners of each individual protein. Figure generated using Cytoscape (Smoot et al., 2011).

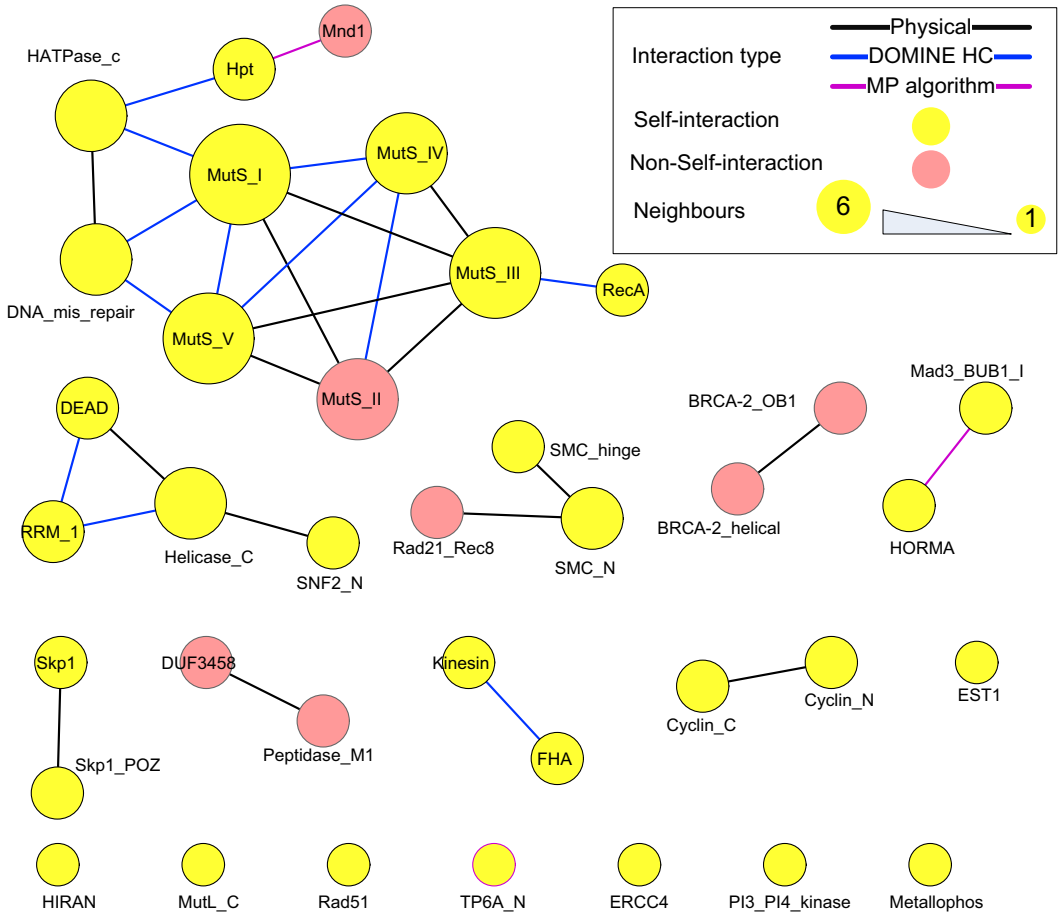


FIGURE 10 | Domain interactions contained in the putative meiotic network. Types of "evidence" were shown as edge colors. Edges of self-interacting domains were omitted, but highlighted by node (and node border) color.

B. rapa counterparts of AtSMC6, but two of them do not appear to participate in meiosis. However, for the majority of multi-copy proteins similar interacting partners are identified.

In terms of interactions, there were several hub proteins in the network, e.g., RAD51 *(26 connections)*, RAD50 *(19 connections)*, MLH1 *(15 connections)*, SMC1 *(14 connections)*, *and* MAD2 *(13 connections)*. Interestingly, these hub proteins were identified by the MCL algorithm (Enright et al., 2002) to form separate clusters with their direct neighbors (shadowed areas in **Figure 9**). Most of the interactions in the network were supported by only one piece of evidence (low confidence), and high confidence interactions were sparse and mainly self-interactions. However, it is a more dense and complex network than those predicted for *A. thaliana* (Lin et al., 2009) and rice (Aya et al., 2011) meiotic proteins.

Protein domains contained in the putative meiotic network were extracted and their interactions are shown in **Figure 10** (those of the hub proteins can be seen in **Table 3**). Overall, it is a sparsely connected network with mainly self-interactions. This suggests that although the meiotic protein interaction network has a very high density, the driving force mediating those interactions is possibly domain self-interactions. Most of the self-interactions are experimentally verified and some of them are derived from the MP algorithm, for example, self-interaction between TP6A_N. The biggest cluster was formed by the interactions among several domains, for example, MutS family domains (contained by MSH2, MSH4, MSH5), RecA (RAD51 and DMC1), and DNA mismatch repair (PMS1 and MLH1). Some of the proteins containing these domains are already thought to form protein complexes during meiosis. *In vitro* studies using purified human hMSH4 and hMSH5 have revealed that they act as complex to stabilize progenitor Holliday junctions (Holliday, 1964). Evidence suggests this is also likely the case in *A. thaliana*, for AtMSH4 and AtMSH5 (Higgins et al., 2004, 2008; Snowden et al., 2004). Other studies suggest that AtAHP2 (containing an Hpt domain) and AtMND1 (Mnd1) also form a complex (Vignard et al., 2007). During budding yeast (*Saccharomyces cerevisiae*) meiosis, interactions were found among MLH1, MLH3 (HATPase_c), and PMS1 (DNA mismatch repair and HATPase_c; Argueso et al., 2002; Nishant et al., 2008), however, these are yet to be experimentally verified in *A. thaliana*. Note that some of the self-interacting domains in **Figure 10**, for example TP6A_N (SPO11), do not show direct interactions with other domains. This does not necessarily mean that the interactome contains no predictions, but that for ease of visualization, we omitted indirect connections.

CONCLUSION

In the present study, we have inferred the *B. rapa* interactome using PPI data available from *A. thaliana*. These PPI data were either physical interactions verified through experiments, or predictions based on orthology. The relationship between the two genomes was established by studying orthologs/collinearity/sequence similarity. We also utilized domain interactions in our predictions. Both known and predicted interacting domains, as well as protein domain assignments of *B. rapa*, were used to predict possible interactions.

The inferred interactome contains 17,255 predicted interactions at high confidence level, and 723,310 predicted interactions at low confidence level. The interactome covers around 50% of the proteins in the *B. rapa* genome, and its high confidence interaction predictions give a coverage of around 78% for those proteins. As a first effort of establishing a *B. rapa* interactome, our inferred interactome could be a useful resource for experimental biologists or other researchers using *B. rapa* as a working plant. The interactome is available at http://www.meiosys.org/dissemination/ as pure text files; other formats e.g., SQL are available upon request.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement number KBBE-2009-222883.

REFERENCES

Allocco, D. J., Kohane, I. S., and Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5:18. doi:10.1186/1471-2105-5-18

Arabidopsis Interactome Mapping Consortium. (2011). Evidence for network evolution in an *Arabidopsis* interactome map. *Science* 333, 601–607.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., et al. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38, D525–D531.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8, 528–529.

Argueso, J. L., Smith, D., Yi, J., Waase, M., Sarin, S., and Alani, E. (2002). Analysis of conditional mutations in the *Saccharomyces cerevisiae* MLH1 gene in mismatch repair and in meiotic crossing over. *Genetics* 160, 909–921.

Aya, K., Suzuki, G., Suwabe, K., Hobo, T., Takahashi, H., Shiono, K., et al. (2011). Comprehensive network analysis of anther-expressed genes in rice by the combination of 33 laser microdissection and 143 spatiotemporal microarrays. *PLoS ONE* 6:e26162. doi:10.1371/journal.pone.0026162

Bader, G., and Hogue, C. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi:10.1186/1471-2105-4-2

Berendsen, H. J. C., Van Der Spoel, D., and Van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91, 43–56.

Brandao, M. M., Dantas, L. L., and Silva-Filho, M. C. (2009). AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinformatics* 10:454. doi:10.1186/1471-2105-10-454

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* 1695, 1695.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., et al. (2008). AtPID: *Arabidopsis thaliana* protein interactome database – an integrative platform for plant systems biology. *Nucleic Acids Res.* 36, D999–D1008.

De Muyt, A., Vezon, D., Gendrot, G., Gallois, J. L., Stevens, R., and Grelon, M. (2007). AtPRD1 is required for meiotic double strand break formation in *Arabidopsis thaliana*. *EMBO J.* 26, 4126–4137.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* 12, 1540–1548.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584.

Ferdous, M., Higgins, J. D., Osman, K., Lambing, C., Roitinger, E., Mechtler, K., et al. (2012). Inter-homolog crossing-over and synapsis in *Arabidopsis* meiosis are dependent on the chromosome axis protein AtASY3. *PLoS Genet.* 8:e1002507. doi:10.1371/journal.pgen.1002507

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29-37.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222.

Fontana, P., Cestaro, A., Velasco, R., Formentin, E., and Toppo, S. (2009). Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS ONE* 4:e4619. doi:10.1371/journal.pone.0004619

Gallone, G., Simpson, T. I., Armstrong, J. D., and Jarman, A. P. (2011). Bio::Homology::InterologWalk – a Perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics* 12:289. doi:10.1186/1471-2105-12-289

Galperin, M. Y., and Fernandez-Suarez, X. M. (2012). The 2012 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.* 40, D1–D8.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N. J., Millar, A. H., and Geisler, M. (2007). A predicted interactome for *Arabidopsis*. *Plant Physiol.* 145, 317–329.

Gu, H., Zhu, P., Jiao, Y., Meng, Y., and Chen, M. (2011). PRIN: a predicted rice interactome network. *BMC Bioinformatics* 12:161. doi:10.1186/1471-2105-12-161

Higgins, J. D., Armstrong, S. J., Franklin, F. C., and Jones, G. H. (2004). The *Arabidopsis* MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in *Arabidopsis*. *Genes Dev.* 18, 2557–2570.

Higgins, J. D., Vignard, J., Mercier, R., Pugh, A. G., Franklin, F. C., and Jones, G. H. (2008). AtMSH5 partners AtMSH4 in the class I meiotic crossover pathway in *Arabidopsis thaliana*, but is not required for synapsis. *Plant J.* 55, 28–39.

Holliday, R. (1964). The induction of mitotic recombination by mitomycin C in Ustilago and Saccharomyces. *Genetics* 50, 323–335.

Hultén, M. A. (2010). "Meiosis," in *Encyclopedia of Life Sciences*. Chichester: John Wiley & Sons, Ltd.

Iqbal, M., Freitas, A. A., Johnson, C. G., and Vergassola, M. (2008). Message-passing algorithms for the prediction of protein domain interactions from protein–protein interaction data. *Bioinformatics* 24, 2064–2070.

Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H. (2006). Evolutionary conservation of domain-domain interactions. *Genome Biol.* 7, R125.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., et al. (2007). Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* 5:44. doi:10.1186/1741-7007-5-44

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645.

Kuzniar, A., Van Ham, R. C., Pongor, S., and Leunissen, J. A. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24, 539–551.

Lagercrantz, U., Putterill, J., Coupland, G., and Lydiate, D. (1996). Comparative mapping in *Arabidopsis* and *Brassica*, fine scale genome collinearity and congruence of genes controlling flowering time. *Plant J.* 9, 13–20.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210.

Lee, K., Thorneycroft, D., Achuthan, P., Hermjakob, H., and Ideker, T. (2010). Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *Plant Cell* 22, 997–1005.

Lin, M., Hu, B., Chen, L., Sun, P., Fan, Y., Wu, P., et al. (2009). Computational identification of potential molecular interactions in *Arabidopsis*. *Plant Physiol.* 151, 34–46.

Lin, M., Shen, X., and Chen, X. (2010). PAIR: the predicted *Arabidopsis* interactome resource. *Nucleic Acids Res.* 39, D1134–D1140.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673.

Lysenko, A., Hindle, M. M., Taubert, J., Saqi, M., and Rawlings, C. J. (2009). Data integration for plant genomics – exemplars from the integration of *Arabidopsis thaliana* databases. *Brief. Bioinformatics* 10, 676–693.

Ma, H. (2006). A molecular portrait of *Arabidopsis* meiosis. *The Arabidopsis Book* 4, 1–39.

McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., Lawrence, M. L., Camon, E. B., et al. (2006). AgBase: a functional genomics resource for agriculture. *BMC Genomics* 7:229. doi:10.1186/1471-2164-7-229

Morsy, M., Gouthu, S., Orchard, S., Thorneycroft, D., Harper, J. F., Mittler, R., et al. (2008). Charting plant interactomes: possibilities and challenges. *Trends Plant Sci.* 13, 183–191.

Nishant, K. T., Plys, A. J., and Alani, E. (2008). A mutation in the putative MLH3 endonuclease domain confers a defect in both mismatch repair and meiosis in Saccharomyces cerevisiae. *Genetics* 179, 747–755.

Obayashi, T., and Kinoshita, K. (2010). Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways. *J. Plant Res.* 123, 311–319.

Osman, K., Higgins, J. D., Sanchez-Moran, E., Armstrong, S. J., and Franklin, F. C. (2011). Pathways to meiotic recombination in *Arabidopsis thaliana*. *New Phytol.* 190, 523–544.

Osman, K., Roitinger, E., Yang, J., Armstrong, S. J., Mechtler, K., and Franklin, F. C. H. (in press). "Analysis of meiotic protein complexes from *Arabidopsis* and *Brassica* using affinity-based proteomics," in *Plant Meiosis: Methods and Protocols*, eds W. P. Pawlowski, M. Grelon and S. J. Armstrong (New York: Springer).

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203.

Pennisi, E. (2012). Evolution. Gene duplication's role in evolution gets richer, more complex. *Science* 338, 316–317.

Richardson, T. J., and Urbanke, R. L. (2001). The capacity of low-density parity-check codes under message-passing decoding. *IEEE Trans. Inf. Theory* 47, 599–618.

Schuster-Bockler, B., and Bateman, A. (2007). Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics* 8:259. doi:10.1186/1471-2105-8-259

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432.

Snowden, T., Acharya, S., Butz, C., Berardini, M., and Fishel, R. (2004). hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Mol. Cell* 15, 437–451.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 39, D718–D723.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488.

Tang, H., and Lyons, E. (2012). Unleashing the genome of *Brassica rapa*. *Front. Plant Sci.* 3:172. doi:10.3389/fpls.2012.00172

Trick, M., Cheung, F., Drou, N., Fraser, F., Lobenhofer, E. K., Hurban, P., et al. (2009). A newly-developed community microarray resource for transcriptome profiling in Brassica species enables the confirmation of Brassica-specific expressed sequences. *BMC Plant Biol.* 9:50. doi:10.1186/1471-2229-9-50

Vignard, J., Siwiec, T., Chelysheva, L., Vrielynck, N., Gonord, F., Armstrong, S. J., et al. (2007). The interplay of RecA-related proteins and the MND1-HOP2 complex during meiosis in *Arabidopsis thaliana*. *PLoS Genet.* 3, 1894–1906. doi:10.1371/journal.pgen.0030176

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305.

Xu, F., Li, G., Zhao, C., Li, Y., Li, P., Cui, J., et al. (2010). Global protein interactome exploration through mining genome-scale data in *Arabidopsis thaliana*. *BMC Genomics* 11(Suppl 2):S2. doi:10.1186/1471-2164-11-S2-S2

Yang, H., Lu, P., Wang, Y., and Ma, H. (2010). The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *Plant J.* 65, 503–516.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* 51, 2282–2312.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.* 39, D730–D735.

Zhang, Z., Luo, Z. W., Kishino, H., and Kearsey, M. J. (2005). Divergence pattern of duplicate genes in protein-protein interactions follows the power law. *Mol. Biol. Evol.* 22, 501–505.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 August 2012; accepted: 11 December 2012; published online: 04 January 2013.

*Citation: Yang J, Osman K, Iqbal M, Stekel DJ, Luo Z, Armstrong SJ and Franklin FCH (2013) Inferring the Brassica rapa interactome using protein–protein interaction data from Arabidopsis thaliana. Front. Plant Sci. **3**:297. doi: 10.3389/fpls.2012.00297*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

METHODS ARTICLE
published: 31 July 2012
doi: 10.3389/fpls.2012.00172

Unleashing the genome of *Brassica rapa*

Haibao Tang*[1] *and Eric Lyons*[2]

[1] *J. Craig Venter Institute, Rockville, MD, USA*
[2] *iPlant Collaborative, School of Plant Sciences, University of Arizona, Tucson, AZ, USA*

Edited by:
Michael Freeling, University of California Berkeley, USA

Reviewed by:
Michael Freeling, University of California Berkeley, USA
Xiangfeng Wang, University of Arizona, USA

****Correspondence:***
Eric Lyons, iPlant Collaborative, School of Plant Sciences, University of Arizona, Keating Bioresearch Building, 1657 E. Helen St. Tucson, AZ 85745, USA.
e-mail: elyons.uoa@gmail.com

The completion and release of the *Brassica rapa* genome is of great benefit to researchers of the Brassicas, *Arabidopsis*, and genome evolution. While its lineage is closely related to the model organism *Arabidopsis thaliana*, the Brassicas experienced a whole genome triplication subsequent to their divergence. This event contemporaneously created three copies of its ancestral genome, which had diploidized through the process of homeologous gene loss known as fractionation. By the fractionation of homeologous gene content and genetic regulatory binding sites, *Brassica*'s genome is well placed to use comparative genomic techniques to identify syntenic regions, homeologous gene duplications, and putative regulatory sequences. Here, we use the comparative genomics platform CoGe to perform several different genomic analyses with which to study structural changes of its genome and dynamics of various genetic elements. Starting with whole genome comparisons, the *Brassica* paleohexaploidy is characterized, syntenic regions with *A. thaliana* are identified, and the TOC1 gene in the circadian rhythm pathway from *A. thaliana* is used to find duplicated orthologs in *B. rapa*. These TOC1 genes are further analyzed to identify conserved non-coding sequences that contain cis-acting regulatory elements and promoter sequences previously implicated in circadian rhythmicity. Each "cookbook style" analysis includes a step-by-step walk-through with links to CoGe to quickly reproduce each step of the analytical process.

Keywords: comparative genomics, synteny, CoGe, *Brassica rapa*, syntenic dotplot, *Arabidopsis*, TOC1, conserved non-coding sequences

INTRODUCTION

Cultivars of the *Brassica* genus provide humankind with a wide variety of dietary vegetables and plant oils, and are major contributors to horticultural and agricultural economies worldwide. The *Brassica* crops are frequently used as vegetable cuisine in many cultures where they are recognized as rich sources of dietary fiber, vitamins, and anti-cancer secondary metabolites including glucosinolates and sulforaphane (Hayes et al., 2008). *Brassica* oilseeds, known as "canola oil," provide about 13% of the world's supply of edible vegetable oil (Raymer, 2002).

Brassicas display the greatest diversity of leaf and floral architecture, which are manifested among many subspecies within the same species, also known as "morphotypes." For example, the species *Brassica rapa* includes familiar morphotypes known as Chinese cabbage, bok choy, turnip, canola, etc. *B. oleracea* includes morphotypes such as broccoli, cabbage, cauliflower, Brussels sprouts, and kale. *B. rapa* (A genome), along with its sister species *B. nigra* (B genome) and *B. oleracea* (C genome), make up the "Triangle of U" that describes how the pairwise combinations of these diploid species can hybridize to form allotetraploids, including *B. carinata*, *B. juncea*, and *B. napus* (Nagaharu, 1935). The extreme phenotypic "plasticity" of the diploid and tetraploid *Brassica* species are often compared to the diversity of dogs, both of which are excellent examples for the study of directed artificial selection and the domestication process.

The Brassicas are the closest crop relatives to the model plant species, *Arabidopsis thaliana*, which has a genome size of ∼120 Mb (The Arabidopsis Genome Initiative, 2000; Bennett et al., 2003). The "diploid" *Brassica* genomes are three to five times larger than that of *Arabidopsis*, ranging from 529 Mb for *B. rapa* to 696 Mb for *B. oleracea* (Johnston et al., 2005; Lysak et al., 2009). Earlier studies revealed large chromosomal blocks of conserved synteny and collinearity between *Arabidopsis* and *Brassica* by mapping genetic markers of the *Brassica* genomes onto the *Arabidopsis* reference. These well-conserved regions are often referred to as "Parkin blocks" (Parkin et al., 2003, 2005). The high-resolution whole genome all-against-all comparison between *Arabidopsis* and the recent *B. rapa* genome showed that more than 90% of the sequences from each genome are located in 24 large collinear Parkin blocks (Wang et al., 2011).

Whole genome duplications, or polyploidy events, are known to have occurred in the evolutionary history of many plant species (Tang et al., 2008b, 2010; Van de Peer et al., 2009; Jiao et al., 2011; Proost et al., 2011). The model plant *A. thaliana* was already a highly duplicated genome, with three rounds of duplication and triplications (α/β/γ), resulting in at least 12× of the ancestral angiosperm genome (Bowers et al., 2003; Tang et al., 2008a); yet the diploid *Brassica* species has experienced an additional round of genome triplication event on top of these events (**Figure 1**). The diploid *Brassica* species were first hypothesized to have been

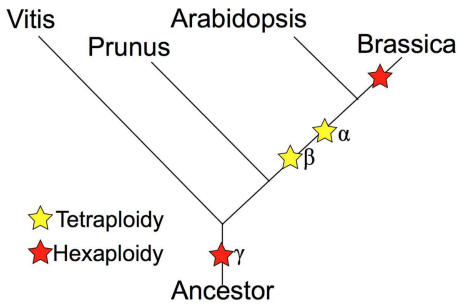


FIGURE 1 | Phylogeny of *Brassica* and relatives marking the relative placement of lineage divergence and polyploidy events. Polyploidies are named according to the *Arabidopsis* convention of α (most recent in the linage of *Arabidopsis*), β (second most recent), and γ (eudicot paleohexaploidy).

triplicated based on comparative mapping (Lagercrantz and Lydiate, 1996; Lagercrantz, 1998; Parkin et al., 2003, 2005), BAC-FISH (Lysak et al., 2005), and BAC sequencing studies (Yang et al., 2006). The genome sequence of *B. rapa* has directly confirmed the genome triplication event with almost complete coverage of the *B. rapa* genome (Wang et al., 2011). The recurring genome duplications and triplication events have created massive genetic redundancy that quickly opens the possibility of sub-functionalization and neo-functionalization for duplicated or triplicated homeologs (Force et al., 1999; Shruti and David, 2005). It is likely that the extreme morphological diversity seen within the various *Brassica* species is due, at least in part, to the genetic redundancy and functional diversification permitted by these genomic events.

Extensive genome-wide comparison between *Arabidopsis* and *Brassica* has revealed unusual patterns of gene loss. The three copies of the genomes within the same nucleus (subgenome) – initially similar in their sizes and gene contents – have since accumulated different amounts of gene losses (or "fractionations") following the most recent genome triplication event. Of the 24 Parkin blocks of conserved synteny, 20 showed significant deviation from the null "random gene loss" model (Wang et al., 2011; Tang et al., 2012). This striking contrast of gene loss rates among the three distinct subgenomes of *B. rapa* allows each subgenomes to be reconstructed and labeled as subgenomes I, II, III according to the number of gene losses ranging from most (I) to medium (II) to least fractionated (III; Tang et al., 2012). Note that subgenomes I/II/III (Tang et al., 2012) correspond to subgenomes MF2/MF1/LF (Wang et al., 2011), partially due to the fact that the number of genes on subgenomes I and II are very similar and substantially lower than that of subgenome III (Wang et al., 2011). Subgenomes I, II, and III have retained 5966, 7679, and 11536 genes, respectively (ignoring genes that are either unique to *B. rapa* or have transposed; Wang et al., 2011; Tang et al., 2012). Certain classes of genes, particularly subunits of large multimeric protein complexes or regulatory machineries, are retained in higher copy numbers than others (Wang et al., 2011), as predicted by the "Gene Dosage Hypothesis" (Birchler and Veitia, 2010; Schnable et al., 2012b).

By exploiting the close evolutionary relationship between *Arabidopsis* and *Brassica*, researchers have obtained a natural experimental system for understanding the evolution of genome structure following a hexaploidy event. In addition, these lineages are sufficiently diverged to permit the identification of plant conserved non-coding sequences (CNSs; Subramaniam and Freeling, 2012), which may contain cis-regulatory elements. Comparative genomics within the *Brassica* genus and Brassicaceae family will play an increasingly critical role, as many more genome sequences from this family are currently in the making. With a focus on applying these important genomic techniques, this paper uses a set of illustrative questions to walk-through various analyses using the online comparative genomics platform CoGe (Lyons and Freeling, 2008). These questions start with analyzing the entire *Brassica* genome, then dive into specific syntenic regions, and finally analyze promoter sequences of a set of genes to identify putative regulatory sequences. Since all of CoGe's tools are web-based, the techniques detailed are approachable for anyone with access to a computer and the Internet. However, because many of the analyses have interactive data visualization, using a computer with a large monitor is recommended. All datasets contain in and generated by CoGe are available for download. While *Brassica* is the focus of this paper, the techniques are applicable to any set of genomes, though the interpretations of certain analyses rely on the unique polyploid nature of plants and their relative phylogenetic positions. The examples covered herein include whole genome comparisons within and between *B. rapa* and *A. thaliana* to identify syntenic orthologous and homeologous gene pairs. We will perform in-depth analyses of these syntenic genes sets to reveal the most recent genome triplication event in *Brassica* as well as more ancient polyploidy events in the shared lineages in the crucifer family. We will also provide detailed analyses of a promoter region of a gene involved in the circadian rhythm pathway, TOC1, to identify CNSs and putative cis-regulatory elements.

RESULTS/METHODS/DISCUSSION

OVERVIEW OF CoGe

CoGe is publicly available at http://genomevolution.org. This resource contains four major systems: a data engine storing thousands of genomes, a suite of interconnected web-based tools, a wiki documentation system with hundreds of pages on comparative genomics, and a TinyURL resource for storing links to CoGe to regenerate data and analyses. The data in CoGe is constantly growing as new genomes and new versions of existing genomes become available. Currently, there are nearly 20,000 genomes from 15,000 organisms. There are over 20 tools in CoGe; each of these performs one general task, such as searching for genomes, displaying FASTA sequences, querying genomes, comparing genomic regions, etc. These tools are all interlinked with one another so that results generated in one tool may be seamlessly sent to another tool for downstream analyses (Lyons et al., 2008a). Due to the interlinking of these tools, there is no specific workflow or analytical pipeline one must follow. Instead, the questions asked and the discoveries made drive the direction of the analyses.

To learn how to use CoGe, interpret its results, and get background information on comparative genomics, there is an extensive wiki available[1]. Each tool is linked to specific

[1]http://genomevolution.org/wiki/

documentation in the wiki, along with links to over 50 written and video tutorials, as well as to FAQs and information about where to get more help.

Most analyses in CoGe return a URL along with the results that can be used to regenerate or share the analysis at any point in the future. It is important to note that there are no inherent pre-computed analyses in CoGe. New analyses are performed on-the-fly. However, large analyses may be cached for some time in case those results are revisited, which will likely incur a one-time computation cost. In order to get the computational scalability needed for its analyses, CoGe is part of the Powered by iPlant program[2], and makes extensive use of iPlant Collaborative's compute, storage, and cyber infrastructure resources (Goff et al., 2011). Anyone with an iPlant account may use those credentials to log into CoGe in order to share private data with other CoGe users.

Useful links:

- CoGe: http://genomevolution.org
- Forums: http://genomevolution.org/r/4t7m
- Tutorials: http://genomevolution.org/r/4a3
- New to CoGe: http://genomevolution.org/r/4sr7
- General news: http://genomevolution.org/r/4sr6
- How to get an account: http://genomevolution.org/r/4sr8
- How to add a private genome: http://genomevolution.org/r/4sr9
- CoGe contact list: http://genomevolution.org/r/4tal

CHARACTERIZING THE *BRASSICA* HEXAPLOID

Self–self comparisons

The phylogeny in **Figure 1** shows that the *Brassica* lineage contains a recent whole genome triplication event. This event has effectively caused the 2*n* ancestor to become a 6*n*. Over evolutionary time, such polyploidy events are followed by the diploidization process, whereby the gene content of a genome is reduced (Wolfe, 2001). The primary mechanism of post-polyploid gene loss is known as fractionation and is thought to occur through deletions by intra-strand recombination events (Woodhouse et al., 2010; Tang et al., 2012). While many duplicated genes are removed by this process, some homeologous genes are retained in multiple copies (Thomas et al., 2006). The retention of these gene pairs provides a strong evolutionary signal of polyploidy events and detection of them permits the identification of duplicated genomic regions (Tang et al., 2008a). Such genomic regions are derived from the same ancestral genomic region and are syntenic. Synteny, in a genomic context, may refer to genomic regions within the same genome or between genomes of different organisms, and are inferred through the identification of collinear sets of putatively homologous gene pairs. The parsimonious reasoning is that a collinear set of homologous genes arose through sharing a common evolutionary history.

Detecting syntenic genomic regions is the high watermark for determining whether a genome underwent a polyploidy event. If, through intra-genomic comparison, all genomic regions are syntenic to other regions, strong evidence is provided for polyploidy. By characterizing the depth of syntenic coverage across a genome, the nature of the polyploidy may be determined. For example, if a genome underwent a tetraploidy event, there would be a 2:2 intra-genomic syntenic mapping where each genomic region is syntenic to itself and one additional genomic region (Tang et al., 2011). Likewise, if a genome underwent a hexaploidy event, there would be a 3:3 intra-genomic mapping (Jaillon et al., 2007).

However, the diploidization process by fractionation can obfuscate the ability to infer synteny through collinear gene order. Over evolutionary time, the more likely any duplicated gene may be lost to fractionation. Fortunately, some gene families are resistant to fractionation, and these can continue to provide a signal to detect syntenic regions. However, concurrent with the diploidization process are additional evolutionary events that can further degrade the collinear signal (Lyons et al., 2008b). These events include gene and genomic region transpositions, chromosomal inversions, chromosomal fissions and fusions, and, most importantly, subsequent polyploidy events. While all of these increase the genomic distance between collinear genes (thus reducing the power to detect syntenic regions), the latter case most effectively reduces the collinear signal by creating an additional duplicate copy of everything in the genome followed by another round of fractionation (Schnable et al., 2012a). This results in syntenic regions of older polyploidy events becoming much more difficult to detect when overlaid by newer ones (Bowers et al., 2003).

Figure 2 shows a self–self syntenic dotplot of the *B. rapa* genome. Only syntenic gene pairs identified through collinearity are drawn on the dotplot (green). When there is a high density of syntenic gene pairs, lines are visualized with varying slopes. The variation in the slopes of syntenic regions is due to biases in fractionation along a genomic region. Syntenic regions cover the entire genome and there are at least two size-classes: large or small. By analyzing a given genomic region by traversing the dotplot vertically or horizontally, it is clear that there are three intra-genomic syntenic regions: the region to itself and to the two larger syntenic regions. The smaller syntenic regions are most likely due to an older whole genome duplication, of which the *Brassica* lineage has had at least three: two tetraploidy events shared with the *Arabidopsis* lineage, and an older hexaploidy shared among nearly all the eudicots (**Figure 1**).

Synonymous mutation values (*K*s) are often used to determine the relative ages of syntenic gene pairs and the distributions of these values for many pairs of genes may identify unique age classes (Kimura, 1977). If the larger syntenic regions in **Figure 2** are indeed from a younger contemporaneous evolutionary event than the smaller regions, they may show a bimodal makeup in their combined *K*s distribution (Blanc and Wolfe, 2004). **Figure 3A** shows this distribution for log 10 transformed *K*s values for all the syntenic gene pairs identified in **Figure 2** as calculated by CODEML (Yang, 2007). Here, there are three conspicuous peaks. The youngest peak, or the peak with the lowest *K*s distance, is on the left. The peak on the far right has a log 10 *K*s value of ∼1.9 (*K*s = 80; 80 synonymous substitutions per site), which is beyond the ability to reliably infer and indicates noise in the analysis. The two left peaks conform to the hypothesis of two recent polyploidy events. The colors from this histogram are overlaid on the dotplot in **Figure 3B**. This clearly shows that the two size-classes of syntenic regions are each derived from different peaks in the *K*s histogram:

[2]http://www.iplantcollaborative.org/ZkX

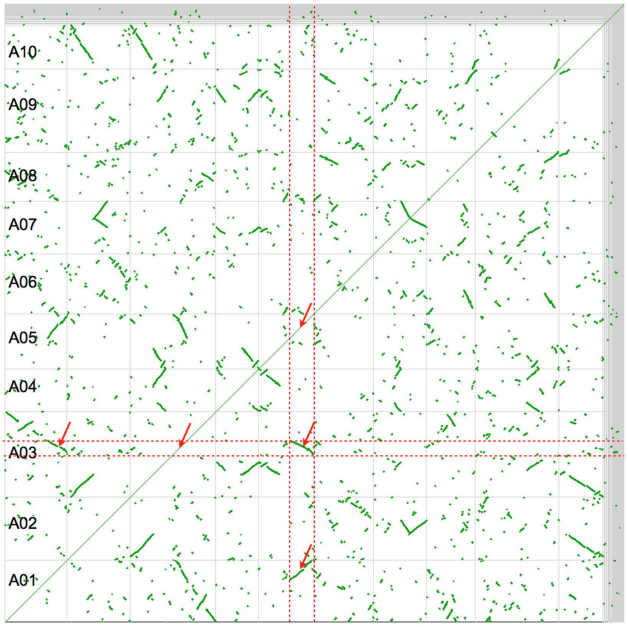


FIGURE 2 | Syntenic dotplot of self–self *Brassica rapa* comparison. Horizontal and vertical gray lines separate chromosomes. Green dots are syntenic gene pairs identified through collinearity. Red dashed lines band and red arrows point to large intra-genomic syntenic signal showing a 3:3 syntenic relationship. Results may be regenerated: http://genomevolution.org/r/4srl

larger regions are purple and younger, while smaller regions are cyan and older. Interestingly, there are several very small syntenic regions that are all colored green, which likely are derived from an even older polyploidy event. However, due to their small numbers, their peak in the distribution is not noticeable.

CoGe methods

1. Go to CoGe's homepage. Quick-link: http://genomevolution.org
2. Go to OrganismView. Quick-link: http://genomevolution.org/r/48px
3. Search for "*B. rapa*" using the search box next to "Organism Name." Quick-link: http://genomevolution.org/r/4srf
4. There may be more than one organism that matches that search term. By selecting different organisms, the page will populate with information about that organism, a list of genomes available for that organism, and information on the selected genome. Search and select for the *B. rapa* genome generated by BGI version 1.1. Quick-link: http://genomevolution.org/r/4srg
5. Under the "Genome information" panel, there is an overview of the genome including its size, number of chromosomes/contigs/scaffolds, the type of sequence (unmasked/masked), links to download the sequence and annotations, and links to various tools in CoGe. Select "SynMap" from the "Links." This loads SynMap, CoGe's tool for generating whole genome syntenic dotplots, with this genome selected for both input genomes. Quick-link: http://genomevolution.org/r/4srh
6. Once SynMapis loaded, press "Generate SynMap" to run the analysis. Quick-link: http://genomevolution.org/r/4sri
7. By default, SynMap uses LASTZ (Schwartz et al., 2003; Harris et al., 2010) for the whole genome comparison, the TangTool package (Tang, 2010) for finding tandem duplicates (Tang, 2010; Tang et al., 2011), and DAGChainer (Haas et al., 2004) to identify collinear gene pairs. These options can be adjusted under the "Analysis Options" tab. Based on empirical testing, the fastest algorithm that works well for SynMap is LAST (Kielbasa et al., 2011), which has been recently integrated into SynMap. This algorithm will become the default in the near future.
8. By default, SynMap orders the chromosomes by size along the two axes and uses the nucleotide distance for the axes.
    1. To change this order to be based on the name of the chromosome, select the "Display Options" tab and select "Name" for "Sort Chromosomes by."
    2. To change the axes distances to genes select "Genes" for "Dotplot axis metric." Quick-link: http://genomevolution.org/r/4srl
9. SynMap has the option for automatically calculating *K*s values using CODEML for all identified syntenic gene pairs.
    1. To turn this option on, select the "Analysis Options" tab and select "Synonymous (*K*s)" for "CodeML."
    2. You have the options of also changing the color scheme used, determining whether the values are log 10 transformed, and setting min/max cutoff values. Quick-link: http://genomevolution.org/r/4srm

Brassica vs. Arabidopsis syntenic dotplots

The *Brassica* hexaploidy event happened after the divergence of its lineage with *Arabidopsis*' (**Figure 1**). This means there is a 1:3 mapping of orthologous syntenic regions between *A. thaliana* and *B. rapa*, and a 2:6 syntenic mapping when including their shared most recent tetraploidy event (α; **Figure 1**). **Figure 4A** shows a syntenic dotplot between *A. thaliana* and *B. rapa*. There is a strong 1:3 syntenic mapping of large syntenic regions for a given region of *A. thaliana*. As seen in the previous example (**Figures 2** and **3**), there are many smaller syntenic regions. Applying *K*s value color markups to the dotplot (**Figure 4B**) highlights the different age classes of the syntenic regions, even though the histogram for these *K*s values does not show a strong bimodal distribution (**Figure 4D**). The peak corresponding to their shared duplication, the α event (cyan), is much smaller and reflects the degradation of the syntenic signal following the more recent *Brassica* hexaploidy event.

Often when studying genes and genomic regions between organisms, it is useful to differentiate between orthologous syntenic regions and out-paralogous syntenic regions (Koonin, 2005). **Figure 4C** shows the syntenic dotplot screened for identifying the best syntenic regions giving a 1:3 syntenic depth between *A. thaliana* and *B. rapa* using QUOTA-ALIGN (Tang et al., 2011). This figure retains the *K*s coloration of syntenic gene pairs, and through comparison to **Figure 4B**, it is clear that nearly all of the retained syntenic regions and corresponding gene pairs are orthologous. When comparing the histograms of the *K*s values from the unscreened dotplot (**Figure 4D**) to the one screened for a 1:3

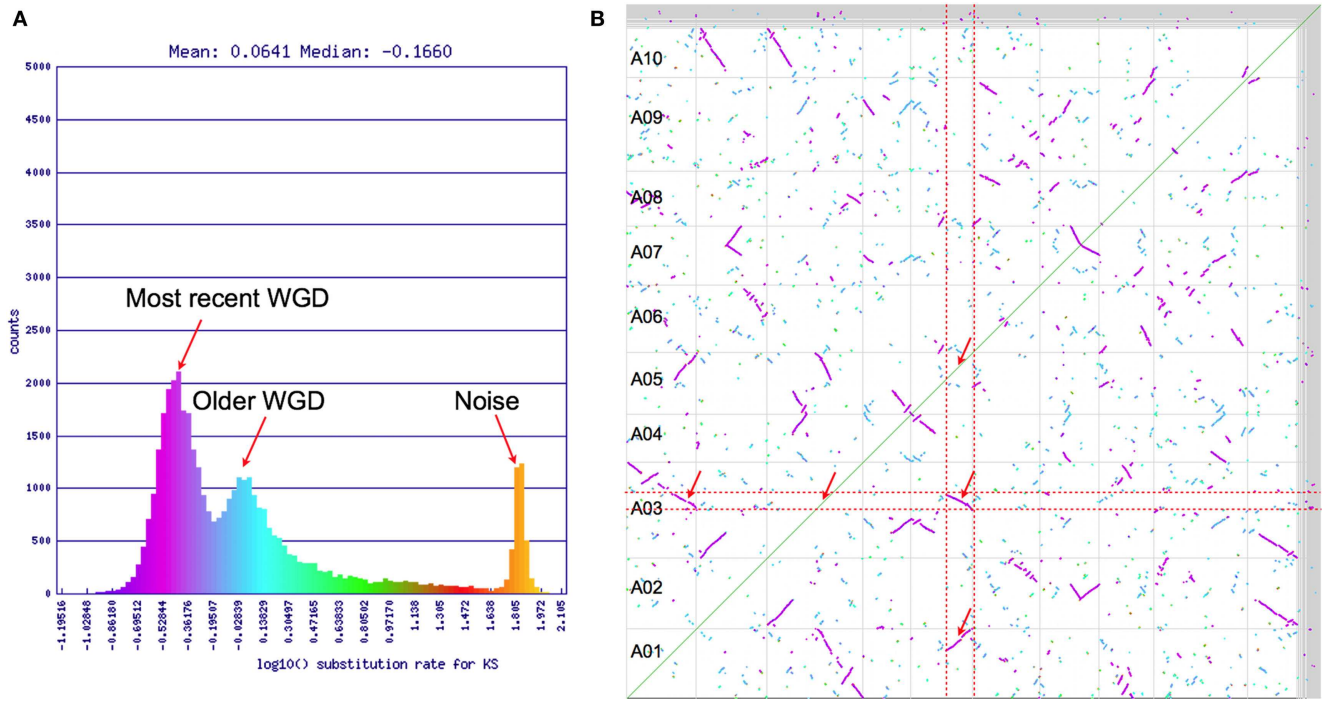


FIGURE 3 | (A) Histogram of log 10 transformed *K*s values of syntenic gene pairs identified in **Figure 2**. **(B)** Syntenic dotplot of self–self *Brassica rapa* comparison with gene pairs colored by their *K*s values shown in **(A)**. Results may be regenerated: http://genomevolution.org/r/4sr4

syntenic depth (**Figure 4E**), the tail of the distribution is greatly reduced in the screened histogram. Since all the results of these analyses are available for download (discussed below in "CoGe Methods"), anyone can quickly generate a list of all the orthologous gene sets along with the *K*s values for syntenic gene pairs.

CoGe methods

1. Go to CoGe's homepage. Quick-link: http://genomevolution.org
2. Go to SynMap. Quick-link: http://genomevolution.org/r/4ss0
3. Search for *A. thaliana* by typing "*Arabidopsis*" in the "Name" search box for Organism 1.
    1. There are many matching organisms to this name. Select "*A. thaliana* Col-0 (thale cress; id1)" from the Organism list.
    2. Of several genomes available for *A. thaliana*, select "unmasked (v10, id11022)." Quick-link: http://genomevolution.org/r/4ss1
4. Repeat the search for *B. rapa* for Organism 2 by typing "rapa" in the "Name" search box.
    1. Select organism, "*B. rapa* (id32114)," and the genome, "unmasked (v1.1, id12468)." Quick-link: http://genomevolution.org/r/4ss1
5. Sort the chromosomes by name by selecting the "Display Options" tab and selecting "Name" from "Sort Chromosomes by."
    1. Run the analysis by pressing the red "Generate SynMap" button. Quick-link: http://genomevolution.org/r/4ss2
6. Turn on the *K*s calculations by selecting the "Analysis Options" tab and selecting "Synonymous (*K*s)" for CodeML.
    1. Rerun the analysis by pressing "Generate SynMap." Quick-link: http://genomevolution.org/r/4ss3
7. You can adjust the display of the *K*s histogram and colors. To mimic **Figure 4B**, turn off the "Log 10 Transformation" of *K*s values by clearing the checkbox, selecting "2.2xRainbow" for the color scheme, and selecting a "Max Val" of 4 to exclude the noise peak in the high *K*s range. Quick-link: http://genomevolution.org/r/4sl1
8. To screen for orthologous syntenic regions, select the "Analysis Options" tab and turn on the algorithm by selecting "Quota-Align" from "Syntenic Depth."
    1. Next, select a syntenic depth of "3" *B. rapa* -to- "1" *A. thaliana*.
    2. The "Overlap Distance" specifies the number of genes by which two syntenic regions may over overlap without either being rejected (Tang et al., 2011). The default value of "40" is usually sufficient. Quick-link: http://genomevolution.org/r/4sl5
9. To download a list of the orthologous syntenic gene pairs from this last analysis, click on "Final syntenic gene set output with GEvo links" available in the "Links and Downloads" section found under the dotplot and *K*s histogram. In this section, you will also, see a link to "Regenerate this analysis" if you wish to return to an analysis in the future. These links were used in the creation of this walk-through.
10. Useful tips:
    - SynMap caches all steps of its analyses. This means that it may take awhile the first time you run a comparison, but the results are returned quickly the next time you run the

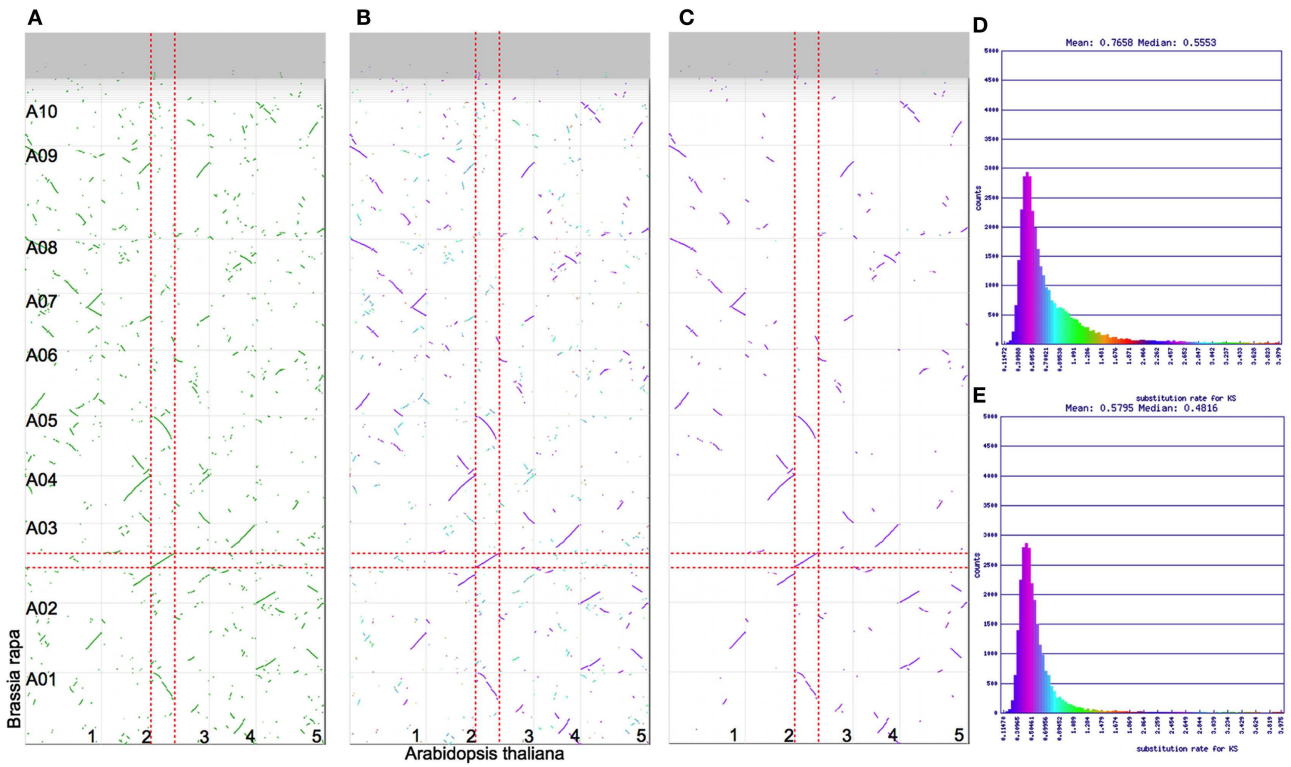


FIGURE 4 | Syntenic dotplots of *Brassica rapa* (*y* axis) vs. *Arabidopsis thaliana* (*x* axis). Vertical and horizontal gray lines separate chromosomes and contigs. Syntenic gene pairs are colored dots on the dotplot. Red dashed lines highlight sets of syntenic regions. **(A)** Syntenic gene pairs colored green. Results may be regenerated: http://genomevolution.org/r/4skv. **(B)** Syntenic gene pairs colored based on their synonymous rate values. Results may be regenerated: http://genomevolution.org/r/4sl1. **(C)** Syntenic gene pairs after screening for a 3:1 (*rapa*:*thaliana*) syntenic depth. These are orthologous gene pairs. Results may be regenerated: http://genomevolution.org/r/4sl5. **(D)** Histogram of synonymous rate values for all syntenic gene-paris. Colors in this histogram are mapped to dotplot shown in **(B)**. **(E)** Histogram of synonymous rate values for syntenic gene pairs screened for a 3:1 syntenic depth. Colors are mapped to dotplot shown in **(C)**.

analysis. If you modify one step in SynMap's analytical workflow, SynMap uses cached results of the steps leading up to the modified one.

- SynMap will run faster if CDS (protein coding sequence) is used in the comparison instead of the whole genome sequence.
- By default, SynMap auto-selects to use CDS if available.
- When using whole genome sequences, select "masked" sequence, if available.
- For large genomes (>500 Mb of sequence), there are often a number of repeat sequences caused by transposons. Comparing a large whole genome sequence to itself (especially those containing many young transposons) usually means a very long wait time for the analysis to complete (days to weeks) and uses a large amount of computer resources. Please contact the authors if there is a genome that needs to be masked.

IDENTIFYING SYNTENIC REGIONS OF INTEREST

While generating whole genome comparisons is useful for characterizing the evolution between two genomes, many researchers are interested in a particular gene or gene family. The typical method employed for identifying homologs of a particular gene uses BLAST (Altschul et al., 1990) to search genomes of interest for genes of similar sequence. However, such methods are limited in terms of characterizing the evolutionary relationship between genes, and additional analyses are often required to determine whether the genes are related through synteny or other forms of duplication (e.g., tandem duplication, transposition duplication, horizontally transferred). Syntenic dotplots help to some degree as they can be used to find a gene of interest and determine if relatives are present in syntenic regions within the same genome or in a related genome. SynMap in CoGe permits users to zoom in by clicking on a chromosome–chromosome comparison in the dotplot. When the mouse is moved over dots in these zoomed-in comparisons, the crosshairs turn red and information about the gene pair is displayed. Also, clicking on a dot opens CoGe's tool for high-resolution sequence analysis, GEvo, with genomic sequence surround the selected gene pair preloaded. GEvo will be discussed in the next section. In addition, researchers may download all identified syntenic gene pairs (described above) and can scan through those for their gene of interest using a text editor, spreadsheet, custom program, or command line tools.

CoGe has two additional tools to help identify homologous and/or syntenic genes and regions. One is CoGeBlast, which is

CoGe's interface for BLAST[3]. A detailed explanation of how to use CoGeBlast that is relevant to this discussion is available in Schnable and Lyons (2011). Briefly, CoGeBlast permits researchers to use BLAST to search their sequences of interest against any set of genomes in CoGe; the interactive display of results permits the evaluation of how well the target genome was matched and allows the user to select matched genomic feature (e.g., genes) for downstream analyses (such as GEvo for determining if genes are derived from syntenic regions).

The second tool is SynFind, which identifies all syntenic regions to a given gene in a user-selected set of genomes, regardless of whether the gene is still present in that region. SynFind is powered by an algorithm known as Synteny Score, which is available as part of the Tang Tools (Tang, 2010). The results of SynFind show a table of the matched regions with their synteny scores and whether or not a syntenic gene was identified. There is the option to download all the identified syntenic gene sets anchored on the genome from which the query gene is derived as well as a syntenic depth table. The syntenic depth table is a breakdown of the number genes in the reference genome at a particular syntenic depth. These tables are helpful in characterizing the syntenic relationship between two genomes, especially for contig-level assemblies where it is difficult to visualize large genomic structures using syntenic dotplots. Examples of various syntenic depth tables and their dotplots can be found at http://genomevolution.org/r/4suf. Importantly, at the top of the results page for SynFind is a link to GEvo to permit the analyses of these genomic regions in more detail.

In the following example, orthologs to *A. thaliana*'s TOC1 will be identified in syntenic regions in *B. rapa* using SynFind. TOC1 is part of the circadian rhythm pathway in *A. thaliana* (Strayer et al., 2000) and is one of five members in the PPR protein family (pseudo-response regulators). The PPR family is expressed in succession from morning to night (Matsushika et al., 2000). TOC1 is negatively regulated by the MYB family transcription factors CCA1 (Wang and Tobin, 1998) and LHY (Schaffer et al., 1998), through binding an Evening Element (EE) in its promoter (Alabadí et al., 2001). TOC1, in turn, negatively regulates CCA1/LHY though binding a cis-regulatory element in their promoters called T1ME (Gendron et al., 2012).

CoGe methods

1. Go to CoGe's homepage. Quick-link: http://genomevolution.org
2. Go to SynFind: Quick-link: http://genomevolution.org/r/4suh
3. Search for *Arabidopsis* TOC1 by its TAIR accession by typing "*AT5G61380*" into the "Specify Feature" Name search box.
    1. Press "Search" to run the search.
    2. Select the *A. thaliana* genome that contains "dsgid11022" from the list. Quick-link: http://genomevolution.org/r/4sui
4. Add *B. rapa* to the "genomes to search" list by typing "rapa" into the "Organism Name" search box.
    1. Select the *Brassica* genome with "dsgid12468" and press "+Add." The genome should appear in the list. Quick-link: http://genomevolution.org/r/4suj
5. Run the analysis by pressing the red "Run SynFind" button. Quick-link: http://genomevolution.org/r/4suj
6. When the results return, you may:
    1. Generate a table of all the syntenic gene sets by clicking the link "Generate master gene set table."
    2. Generate a syntenic dotplot between two genomes by clicking the "dotplot" link.
    3. Save a link to regenerate the SynFind analysis
    4. Send the identified genomic regions to GEvo for high-resolution analysis of the identified syntenic regions.

[3]http://genomevolution.org/r/4stv

HIGH-RESOLUTION ANALYSIS OF SYNTENIC REGIONS

After identifying syntenic regions of interest, it is often useful to analyze those regions in high-resolution. CoGe's GEvo tool permits the comparison of several genomic regions and provides various ways to modify the analyses and visualization of the results. **Figure 5A** shows a comparison of the *A. thaliana* genomic region containing TOC1 and the three orthologous syntenic regions in *B. rapa*. In this analysis, all three *Brassica* regions are compared to *Arabidopsis* using LASTZ. Pink-red blocks located above the gene models visualize the regions of sequence similarity. While there is extensive collinear arrangement of similar sequence between these regions, which is strong evidence for these regions being syntenic, note that there are various and different genes missing among the *Brassica* regions when compared to *Arabidopsis*. This is due to the fractionation of gene content. Of the inferred three ancestral copies of TOC1, there are two remaining ancestral copies in *Brassica*. **Figure 5B** is identical to **Figure 5A** except that gene models with overlapping regions of sequence similarity are colored purple. This shows that nearly the entire gene content of the *Arabidopsis* region is contained among the *Brassica* regions, even though no one *Brassica* region contains all the gene content of *Arabidopsis*. *Brassica* genes colored green are those that were either lost in *Arabidopsis* or transposed into the region following the divergence of these lineages. Also, note that the sizes of the *Brassica* regions are different (BR1 clearly retains more genes than Br2/3) even though they all have equivalent syntenic coverage of the *Arabidopsis* region. This is due to bias in the fractionation process (Thomas et al., 2006; Freeling, 2009; Schnable et al., 2011; Tang et al., 2012).

An important fact to keep in mind during the comparison of syntenic regions is that different algorithms are better suited for different tasks (Lyons and Freeling, 2008). This generally is due to the fine balance between sensitivity, specificity, and promiscuity of various sequence comparison algorithms. **Figure 5C** shows the results using BLASTN for comparing the same regions, and its default settings are too sensitive for this type of analysis. As a general rule, use LASTZ (or relatives) for comparing large genomic regions and BLASTN for comparing small regions. However, different algorithms may be better suited for a given problem depending on the intended resolution.

There are two syntenic orthologs of TOC1 identified in **Figure 5A** in *B. rapa*. It is important to analyze the region with the missing copy in order to determine if the missing gene happened to lie in an unsequenced gap. Such sequences are represented by a string of Ns in the genomic sequence and are colored orange in GEvo. While there are gaps in the *B. rapa* sequence, there are no gaps in the region in which the missing ortholog would be located. Therefore, we can conclude that one of the paleo-orthologs was lost to fractionation.

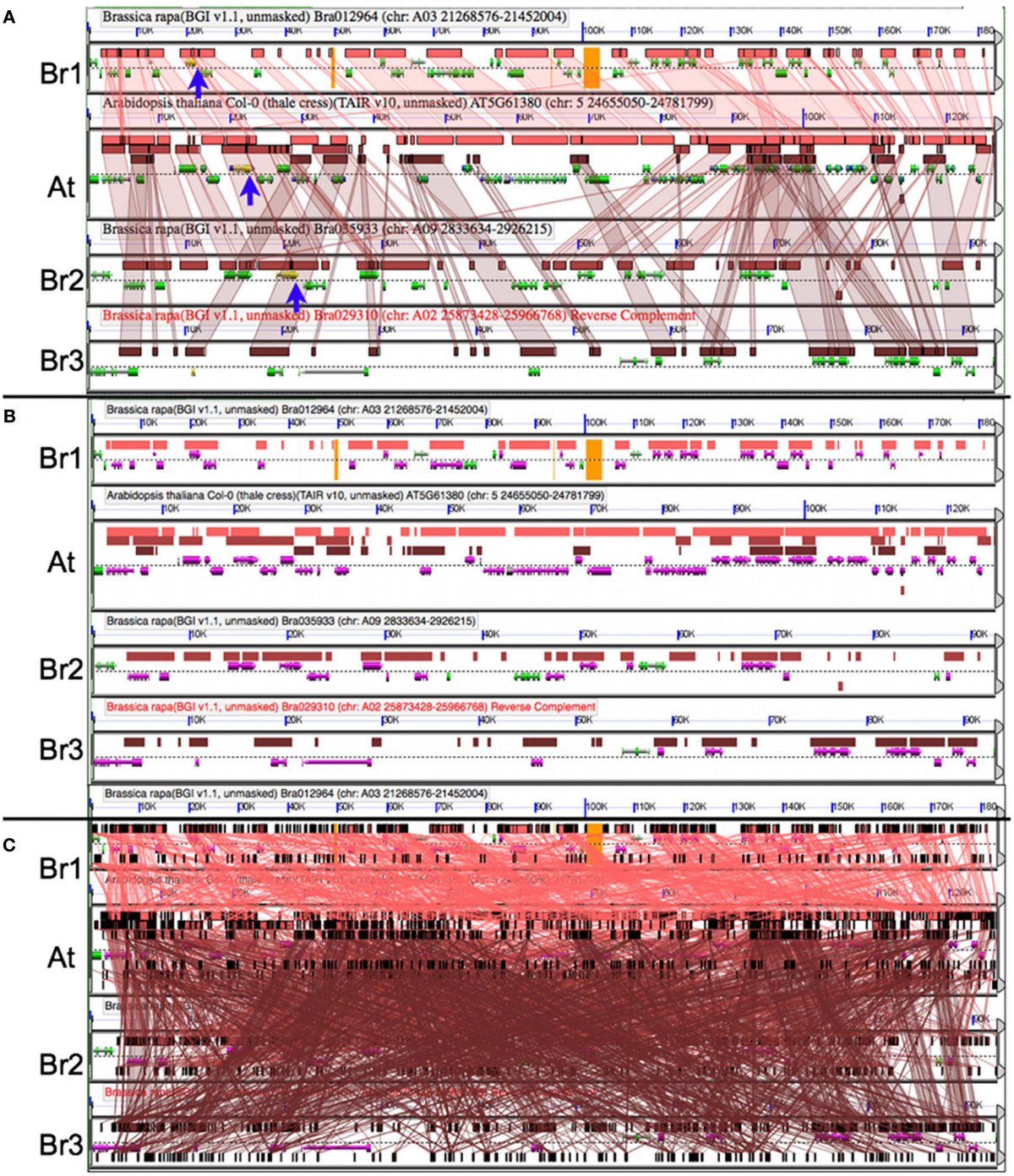


FIGURE 5 | Comparing orthologous syntenic genomic regions between *Brassica rapa* (Br) and *Arabidopsis thaliana* (At) with GEvo. Each panel represents a genomic region with the dashed line separating the top and bottom strands of DNA. Orange in the background signifies unsequenced gaps. Gene models are drawn above and below dashed line as composite colored arrows. The At TOC1 gene and its Br orthologs are colored yellow (blue arrows). Regions of sequence similarity are drawn as colored boxes, and may be connected using transparent wedges. **(A)** Syntenic pattern of collinear regions of sequence similarity as identified by LASTZ. Results may be regenerated: http://genomevolution.org/r/4sma. **(B)** Fractionation of Br's gene content; LASTZ comparison to At's gene content. Genes covered by a region of sequence similarity are colored purple. Note that At's gene content is represented among the combined Br regions. Results may be regenerated: http://genomevolution.org/r/4smb. **(C)** Picking the wrong algorithm for the comparison. BLASTN with settings for detecting CNSs used to compare sequences results in too many non-syntenic regions of sequence similarity. Results may be regenerated: http://genomevolution.org/r/4smc

CoGe methods

1. Start with the TOC1 syntenic regions identified with SynFind in the previous analysis. Quick-link: http://genomevolution.org/r/4suj
2. Follow the link to GEvo (top of the results): http://genomevolution.org/r/4sll
3. When GEvo loads, it will have those genomic regions preloaded and will automatically start running the analysis. By default, the query region from SynFind (*Arabidopsis* in this case) will be placed on the top and used as a reference sequence to which all other regions are compared.
4. When the results are returned, click on a region of sequence similarity to connect it with its partner region. For information on how to use the GEvo's interactive results viewer, see http://genomevolution.org/r/4sz5 (Pedersen et al., 2011).
5. To modify the extent of genomic region analyzed, drag the slider bars located at the end of the genomic regions to zoom in on a region, and either specify an exact amount of sequence up and downstream of the anchor gene, or modify all up and down regions by the same amount.
    1. Expand the analysis by typing "150,000" in the box labeled "Apply distance to all CoGe submissions" and rerun the analysis by pressing the red "Run GEvo Analysis" button. Quick-link: http://genomevolution.org/r/4sz6
6. Use the slider bars to adjust the regions so that only syntenic regions are compared and rerun the analysis. Quick-link:http://genomevolution.org/r/4sma
7. To change the display order of the sequences, drag the sequence submission boxes around relative to one another.
8. To color genes that are overlapped by regions of sequence similarity, select the "Results Visualization Options" tab and turn on the option "Color features overlapped by HSPs" found in the second column. Quick-link: http://genomevolution.org/r/4smb
9. To change the sequence comparison algorithm, select the "Algorithm" tab, and select an algorithm from the "Alignment Algorithm" drop-down menu. Available algorithms are BLASTN (Altschul et al., 1990), LASTZ (Schwartz et al., 2003), CHAOS (Brudno et al., 2004), GenomeThreader (Gremme et al., 2005), LAGAN (Brudno et al., 2003), TBLASTX. Quick-link: http://genomevolution.org/r/4smc

REGULATORY AND CONSERVED NON-CODING SEQUENCES

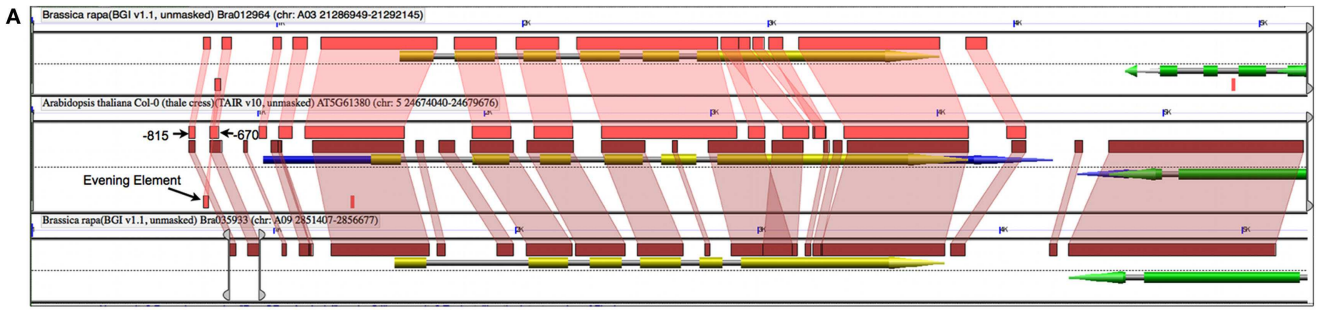
After identifying orthologs to TOC1 in *B. rapa* and confirming their evolutionary history through syntenic analysis, the next step is to identify conserved CNSs in order to identify putative regulatory elements in *Brassica* and to generate hypotheses about their regulatory evolution (Freeling et al., 2007; Subramaniam and Freeling, 2012). Plant CNSs are distinct from animals and have a specific operational definition of two or more similar sequences with an expect value less than or equal to a 15/15 BLASTN exact nucleotide match (Kaplinsky et al., 2002; Inada et al., 2003). Of particular interest, plant CNSs are often detected just above noise when comparing plant sequence (Lyons and Freeling, 2008) and GEvo's default parameters for BLASTN are set to detect plant CNSs.

Prior experimental work identified a DNA binding sequence, dubbed EE, in over 30 circadian rhythm cycling genes whose peak expression was at the end of the day (Harmer et al., 2000). An EE sequence was subsequently found in the promoter of *Arabidopsis* TOC1 and, while mutations to the EE caused a strong reduction in circadian rhythmicity, the promoter fragment (−834:−620) containing this element was essential (Alabadí et al., 2001). **Figure 6A** shows a high-resolution analysis of the *Arabidopsis*' TOC1 and its two *Brassica* orthologs which includes 1500 nt of sequence up and downstream of the genes. By changing to a higher-sensitivity algorithm such as BLASTN set to detect CNSs, smaller regions of sequence similarity are identified. While, as expected, there is extensive sequence conservation across protein coding regions, there are additional regions of sequence similarity in the CNSs. These CNSs are found 5′, 3′, and in the introns of the genes. Such conservation is assumed to be due to purifying selection providing that enough evolutionary time has passed to randomize non-functional sequences (Freeling and Subramaniam, 2009). By comparing these CNSs with the aforementioned experimental work on the regulation of TOC1, the three most 5′ CNSs identified near *Bra012964* match the promoter fragment in *Arabidopsis* determined to be essential for circadian rhythmicity; one of these CNSs contains an EE (**Figure 6B**).

While a similar set of CNSs were identified with *Bra035933*, a CNS containing the EE was not detected. However, by extracting out the entire sequence bordered by the two most 5′ CNSs in *Bra035933*, an EE is contained therein (**Figure 6C**). Interestingly, close examination of these sequences shows that all regions contain a slightly degenerate inverted repeat of EE, which may help to ensure the retention of the sequences during binding-site turnover (Dermitzakis and Clark, 2002), or to facilitate in the cooperative binding of two CCA1/LHY proteins (Eulgem et al., 1999). In any case, analysis of CNSs among homologous syntenic gene sets identifies putative regulatory sequences for further experimental functional characterization.

CoGe methods

1. While the slider bars may be adjusted from the GEvo analysis shown in the previous example to border the genes of interest, a faster method is to type "1500" in the box next to "Apply distance to all CoGe submissions."
    1. Remove the genomic region for *Bra029310* (which does not contain a syntenic ortholog of TOC1) by opening the "Sequence Options" for *Bra029310* and selecting "yes" for "Skip Sequence."
    2. Make sure that BLASTN is selected for the sequence comparison algorithm under the "Algorithm" tab for increased sensitivity, and leave it on its default settings to detect plant CNSs. Quick-link: http://genomevolution.org/r/4t5e
2. Highlight all of the connections between regions of sequence similarity by holding the Shift key and clicking on a colored box. To get information about a particular region of sequence similarity, click on that colored box without holding the Shift key.
    1. In the "GEvo Results Info" information box, you can view a summary for that particular region of sequence similarity. Click the link called "full summary" to open HSPView, which



B

	Query	Subject
Name	Bra012964	AT5G61380)
Location	21,287,698-21,287,716	24,674,802-24,674,820
Percent Identity	94.7%	94.7%
Percent Similarity	94.7%	94.7%
Length	19	19
Match	18 nt	18 nt
Similar	18 nt	18 nt
Gap	0	0
Mismatch	1	1
Orientation	+-	
E-Value	3e-04	
Sequence	AAAATATCTTATGATATTT	AAAATATCTCATGATATTT
Blast Link	CoGeBlast	CoGeBlast
SeqView Link	SeqView	SeqView
Alignment	AAAATATCTTATGATATTT |||||||| |||||||| AAAATATCTCATGATATTT	

C

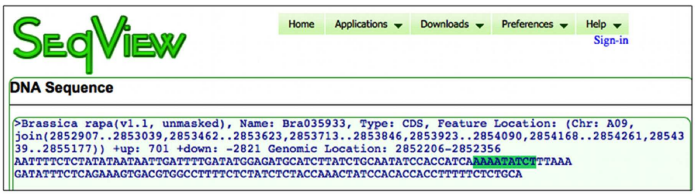


FIGURE 6 | Conserved non-coding sequence analysis of *Arabidopsis* TOC1 gene and orthologs from *Brassica rapa* using GEvo. (A) Gene models are composite arrows where green or yellow regions represent protein coding sequence, blue is mRNA, and gray is the full extent of the gene. Regions of sequence similar are as in **Figure 5** and were identified using BLASTN; such regions are in the opposite orientation if drawn below the dashed line. The bottom *Brassica* region has GEvo's slider bars adjusted to border the two most 5′ CNSs. Results may be regenerated: http://genomevolution.org/r/4t5e **(B)** HSPView's report on the BlastN HSP containing Evening Element (AAAATATCT). **(C)** Using SeqView to visualize the sequence encompassing the two most 5′ CNSs in Bra035933 with the Evening Element highlighted. Sequence can be obtained: http://genomevolution.org/r/4t6t

provides detailed information about the region of sequence similarity. Because the results from GEvo analyses are only cached on CoGe's server for 2 days, providing a quick-link to HSPView is not possible.

3. Extract the sequence upstream of *Bra035933* by dragging the slider bars to the region shown in **Figure 6A** and clicking "Get Sequence" from its sequence submission box. Quick-link: http://genomevolution.org/r/4t6t
4. Search for the EE by using the "find" option in your web-browser and typing in AAAATATCT.

CONCLUSION

While every genome is sacred, it is essential to have the appropriate computational tools to analyze a genome at various scales. Likewise, comparative analyses of a genome to itself and to related species are required in order to understand how a genome and its genetic components have evolved.

The *B. rapa* genome is of outstanding interest for a variety of reasons. Besides being from an agronomically important and morphological diverse clade of plants, its close phylogenetic relationship to the model plant system *A. thaliana* makes its genome extremely valuable. Due to the timing and phylogenetic placement of the *Brassica* hexaploidy event, and the wealth of information and genetic tools available for *A. thaliana*, the *B. rapa*'s genome provides an exceptional natural experimental system. It is sufficiently diverged from *Arabidopsis* to permit the in-depth characterization of its genome structure, gene retention patterns, and conserved CNSs. The example analyses provided above show how to extract a variety of curious patterns and scientific insights from the *Brassica* genome through comparison to *Arabidopsis*.

The next set of genomic resources of benefit to the *Brassica*, *Arabidopsis*, genome evolution, and gene regulation research communities will be extensive functional genomics data for *B. rapa* such as transcriptomes, small RNAs, and DNA methylation patterns. However, to make the most use of such data, they will need to be integrated into comparative genomics platforms such as CoGe. The vision would be to continue these analyses by overlaying and integrating functional data to investigate the regulation, usage, and timing of TOC1 in *Arabidopsis* and its syntenic orthologs in *B. rapa*. This would permit further characterization of the CNSs found between these sequences and ask questions such as: why has *B. rapa* retained two copies of TOC1 and what is their functional relevance? What is the functional consequence of retaining or losing particular CNSs? Is there something special about the truncation of intron 1 in *Bra012964*? Do these

genes have overlapping effects on the functioning of the entire circadian pathway, or have they neo/sub-functionalized their regulation? Sequencing genomes and obtaining their functional data is relatively inexpensive, analyzing these data to transform genomic information into knowledge needs to be too.

ACKNOWLEDGMENTS

We would like to thank the reviews for the their insightful and helpful comments in improving this paper. This work was supported by a grant from the US National Science Foundation (iPlant Collaborative: DBI-0735191).

REFERENCES

Alabadí, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Más, P., and Kay, S. A. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science* 293, 880–883.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Bennett, M. D., Leitch, I. J., Price, H. J., and Johnston, J. S. (2003). Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann. Bot. (Lond.)* 91, 547–557.

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62.

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of polyploidization events. *Nature* 422, 433–438.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., NISC Comparative Sequencing Program, Green, E. D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721–731.

Brudno, M., Steinkamp, R., and Morgenstern, B. (2004). The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. *Nucleic Acids Res.* 32, W41–W44.

Dermitzakis, E. T., and Clark, A. G. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19, 1114–1121.

Eulgem, T., Rushton, P. J., Schmelzer, E., Hahlbrock, K., and Somssich, I. E. (1999). Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *EMBO J.* 18, 4689–4699.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453.

Freeling, M., Rapaka, L., Lyons, E., Pedersen, B., and Thomas, B. C. (2007). G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. *Plant Cell* 19, 1441.

Freeling, M., and Subramaniam, S. (2009). Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol.* 12, 126–132.

Gendron, J. M., Pruneda-Paz, J. L., Doherty, C. J., Gross, A. M., Kang, S. E., and Kay, S. A. (2012). Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3167–3172.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W. H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S. J., Kvilekval, K., Manjunath, B. S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S. M., Cranston, K. A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D. J., White, J. W., Leebens-Mack, J., Donoghue, M. J., Spalding, E. P., Vision, T. J., Myers, C. R., Lowenthal, D., Enquist, B. J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L., and Stanzione, D. (2011). The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2:34. doi:10.3389/fpls.2011.00034

Gremme, G., Brendel, V., Sparks, M. E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *Inform. Softw. Technol.* 47, 965–978.

Haas, B. J., Delcher, A. L., Wortman, J. R., and Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643–3646.

Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J. A., and Kay, S. A. (2000). Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290, 2110–2113.

Harris, B., Riemer, C., and Miller, W. (2010). *LastZ.* Available at: http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/

Hayes, J. D., Kelleher, M. O., and Eggleston, I. M. (2008). The cancer chemopreventive actions of phytochemicals derived from glucosinolates. *Eur. J. Nutr.* 47(Suppl. 2), 73–88.

Inada, D. C., Bashir, A., Lee, C., Thomas, B. C., Ko, C., Goff, S. A., and Freeling, M. (2003). Conserved noncoding sequences in the grasses. *Genome Res.* 13, 2030–2041.

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C, Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A. F., Weissenbach, J., Quétier, F, Wincker, P., and French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., and dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100.

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., Lopez, R., and Price, H. J. (2005). Evolution of genome size in Brassicaceae. *Ann. Bot.* 95, 229–235.

Kaplinsky, N. J., Braun, D. M., Penterman, J., Goff, S. A., and Freeling, M. (2002). Utility and distribution of conserved noncoding sequences in the grasses. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6147–6151.

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267, 275–276.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics 1. *Annu. Rev. Genet.* 39, 309–338.

Lagercrantz, U. (1998). Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* 150, 1217–1228.

Lagercrantz, U., and Lydiate, D. J. (1996). Comparative genome mapping in *Brassica*. *Genetics* 144, 1903–1910.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., and Freeling, M. (2008a). Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 148, 1772–1781.

Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008b). The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* 1, 181–190.

Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A., and Leitch, I. J. (2009). The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* 26, 85–98.

Lysak, M. A., Koch, M. A., Pecinka, A., and Schubert, I. (2005). Chromosome triplication found across the tribe Brassiceae. *Genome Res.* 15, 516–525.

Matsushika, A., Makino, S., Kojima, M., and Mizuno, T. (2000). Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in *Arabidopsis thaliana*: insight into the plant circadian clock. *Plant Cell Physiol.* 41, 1002–1012.

Nagaharu, U. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot.* 7, 389.

Parkin, I. A. P., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C., and Lydiate, D. J. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* 171, 765–781.

Parkin, I. A. P., Sharpe, A. G., and Lydiate, D. J. (2003). Patterns of genome duplication within the *Brassica napus* genome. *Genome* 46, 291–303.

Pedersen, B. S., Tang, H., and Freeling, M. (2011). Gobe: an interactive, web-based tool for comparative genomic visualization. *Bioinformatics* 27, 1015–1016.

Proost, S., Pattyn, P., Gerats, T., and Van de Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. *Plant J.* 66, 58–65.

Raymer, P. L. (2002). "Canola: an emerging oilseed crop," in *Trends in New Crops and New Uses,* eds J. Janick and A. Whipkey (Alexandria: ASHS Press), 122–126.

Schaffer, R., Ramsay, N., Samach, A., Corden, S., Putterill, J., Carré, I. A., and Coupland, G. (1998). The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* 93, 1219–1229.

Schnable, J. C., Freeling, M., and Lyons, E. (2012a). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.* 4, 265–277.

Schnable, J. C., Wang, X., Pires, J. C., and Freeling, M. (2012b). Escape from preferential retention following repeated whole genome duplications in plants. *Front. Plant Sci.* 3:94. doi:10.3389/fpls.2012.00094

Schnable, J. C., and Lyons, E. (2011). Comparative genomics with maize and other grasses: from genes to genomes! *Maydica* 56, 183–200.

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103–107.

Shruti, R., and David, L. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* 5, 28. doi:10.1186/1471-2148-5-28

Strayer, C., Oyama, T., Schultz, T. F., Raman, R., Somers, D. E., Más, P., Panda, S., Kreps, J. A., and Kay, S. A. (2000). Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. *Science* 289, 768–771.

Subramaniam, S., and Freeling, M. (2012). "Conserved noncoding sequences in plant genomes, in *Plant Genome Diversity*, Vol. 1, eds J. F. Wendel, J. Greilhuber, J. Dolezel, and I. J. Leitch (Vienna: Springer), 113–122.

Tang, H. (2010). *TangTools*. Available at: https://github.com/tanghaibao/quota-alignment

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008a). Synteny and collinearity in plant genomes. *Science* 320, 486–488.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954.

Tang, H., Bowers, J. E., Wang, X., and Paterson, A. H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl. Acad. Sci. U.S.A.* 107, 472–477.

Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12, 102. doi:10.1186/1471-2105-12-102

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G. C., Wang, X., Freeling, M., and Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics*. Available at: http://www.genetics.org/content/early/2012/02/02/genetics.111.137349 [accessed April 28, 2012].

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946.

Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. *Trends Plant Sci.* 14, 680–688.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J. H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., Zhang, Z., and *Brassica rapa* Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Wang, Z.-Y., and Tobin, E. M. (1998). Constitutive expression of the circadian clock associated 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. *Cell* 93, 1207–1217.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 8, e1000409. doi:10.1371/journal.pbio.1000409

Yang, T. J., Kim, J. S., Kwon, S. J., Lim, K. B., Choi, B. S., Kim, J. A., Jin, M., Park, J. Y., Lim, M. H., Kim, H. I., Lim, Y. P., Kang, J. J., Hong, J. H., Kim, C. B., Bhak, J., Bancroft, I., and Park, B. S. (2006). Sequence-level analysis of the diploidization process in the triplicated flowering locus C region of *Brassica rapa*. *Plant Cell* 18, 1339–1347.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2012; accepted: 12 July 2012; published online: 31 July 2012.
Citation: Tang H and Lyons E (2012) Unleashing the genome of Brassica rapa. Front. Plant Sci. ***3****:172. doi: 10.3389/fpls.2012.00172*
This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 15 May 2012
doi: 10.3389/fpls.2012.00094

Escape from preferential retention following repeated whole genome duplications in plants

James C. Schnable[1], Xiaowu Wang[2], J. Chris Pires[3] and Michael Freeling[1]*

[1] *Freeling Lab, Plant and Microbial Biology, University of California – Berkeley, Berkeley, CA, USA*
[2] *Molecular Genetics Lab, Biotechnology Department, Institute of vegetables and flowers, Chinese Academy of Agricultural Sciences, Beijing, China*
[3] *Biological Sciences, Bond Life Sciences Center, University of Missouri, Colombia, MO, USA*

Edited by:
Elena R. Alvarez-Buylla, Universidad Nacional Autónoma de Mexico, Mexico

Reviewed by:
Paula Casati, Centro de Estudios Fotosinteticos-CONICET, Argentina
Amy Louise Lawton-Rauh, Clemson University, USA

***Correspondence:**
Michael Freeling, Freeling Lab, Plant and Microbial Biology, University of California – Berkeley, 111 Koshland Hall, PMB, Berkeley, CA 94720, USA.
e-mail: freeling@berkeley.edu

The well supported gene dosage hypothesis predicts that genes encoding proteins engaged in dose–sensitive interactions cannot be reduced back to single copies once all interacting partners are simultaneously duplicated in a whole genome duplication. The genomes of extant flowering plants are the result of many sequential rounds of whole genome duplication, yet the fraction of genomes devoted to encoding complex molecular machines does not increase as fast as expected through multiple rounds of whole genome duplications. Using parallel interspecies genomic comparisons in the grasses and crucifers, we demonstrate that genes retained as duplicates following a whole genome duplication have only a 50% chance of being retained as duplicates in a second whole genome duplication. Genes which fractionated to a single copy following a second whole genome duplication tend to be the member of a gene pair with less complex promoters, lower levels of expression, and to be under lower levels of purifying selection. We suggest the copy with lower levels of expression and less purifying selection contributes less to effective gene-product dosage and therefore is under less dosage constraint in future whole genome duplications, providing an explanation for why flowering plant genomes are not overrun with subunits of large dose–sensitive protein complexes.

Keywords: polyploidy, gene dosage, gene loss, genome evolution, comparative genomics, crucifers, grasses

INTRODUCTION

Plants have been colorfully labeled the "big kahuna of polyploidization" (Sémon and Wolfe, 2007). The lineages leading to the two preeminent models for plant genetics – *Arabidopsis* (a eudicot) and maize (a monocot) – each show evidence of multiple independent whole genome duplications (**Figure 1**) since monocots and eudicots diverged approximately 120 million years ago (Soltis et al., 2009). Recent evidence suggests at least two additional, shared, whole genome duplications prior to the monocot/eudicot split (Jiao et al., 2011). The cumulative ploidy numbers relative to a pre-seed plant ancestor are listed in parentheses in **Figure 1**. Whole genome duplication creates duplicate, potentially redundant, copies of all the genes within a genome. The loss of these duplicate copies from the genomes of ancient polyploid species is known as fractionation (Langham et al., 2004) and – over evolutionary time scales – the majority of genes duplicated by polyploidy will be reduced back to a single copy. If fractionation did not occur, an ancestral genome of 10,000 genes would grow to an unrealistically large 640,000 genes in maize, and 1.44 million genes in *Brassica rapa*.

Some classes of genes, particularly those encoding organelle, preferentially revert to single copy status following whole genome duplications (Duarte et al., 2010). However, other classes of genes – such as subunits of large multiprotein complexes, transcription factors, and signal transduction machinery tend to resist fractionation following whole genome duplication (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005). This observation has been explained by the Gene Dosage Hypothesis (Birchler and Veitia, 2007) which predicts that fractionation of genes encoding proteins involved in dose–sensitive interactions will be selected against, as the loss of either gene copy is expected to throw the dosage of that gene pair's product out of balance with its interaction partners, partners that also tend to remain duplicated. The topic of the influence of gene dosage-constraints on post-tetraploidy genome evolution has been well-reviewed (Sémon and Wolfe, 2007; Edger and Pires, 2009; Freeling, 2009; Birchler and Veitia, 2010). A previous study of multiple sequential tetraploidies in the *Arabidopsis* lineage found a general tendency for genes retained following one tetraploidy to also be retained following a second one (Seoighe and Gehring, 2004).

Since the divergence of the *Arabidopsis* and grape lineages, *Arabidopsis* has experienced two additional rounds of whole genome duplication. The rate of duplicate gene retention for transcription factors after single polyploidies have been observed to be approximately 25% (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004). If no mitigation of gene dosage occurred, our expectation after two rounds of whole genome duplication is that *Arabidopsis* should contain approximately 156% as many transcription factor encoding genes as grape. However, a detailed annotation of transcription factors using conserved protein domains found the number of transcription factors in the *Arabidopsis* genome is only 25.4% greater than the number found in grape

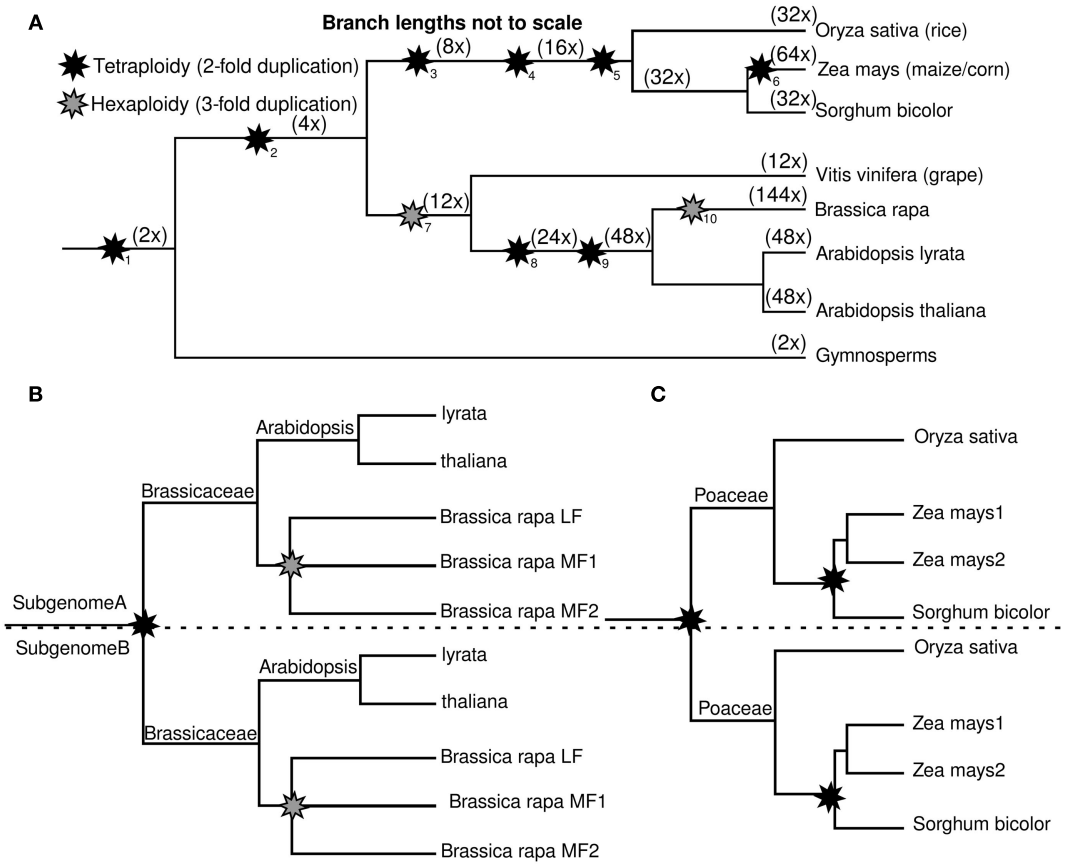


FIGURE 1 | Phylogenetic trees showing the distribution of whole genome duplications throughout the flowering plants. (A) Relationships of all species and whole genome duplications referenced in this paper. Ploidy levels relative to the common ancestor of seed plants are noted along individual branches in parentheses. Footnotes mark individual ancient whole genome duplications, described in more detail in **Table 1**. **(B)** The relationship between the multiple subgenomes created by whole genome duplications within the crucifers. The LF, MF1, MF2 terminology for Brassica subgenomes comes from The *Brassica rapa* Genome Sequencing Project Consortium (2011). **(C)** The relationship between the multiple subgenomes created by whole genome duplication within the grasses. The Maize 1, Maize 2 terminology for maize subgenomes comes from Schnable et al. (2011a).

(Lang et al., 2010). The fitness cost of changes in relative gene dosage must, to some extent, be mitigated over multiple whole genome duplications or the genomes of plants would long ago have become over-burdened with genes encoding life's most complicated machines.

This paper provides evidence that duplicate genes do not equally maintain their progenitor's preference for duplicate gene retention. Duplicate genes produced by whole genome duplication are not equivalent. Parental genomes originating from different species within a polyploid almost immediately differentiate into dominant and non-dominant subgenomes (Chang et al., 2010), and these expression differences are preserved for millions of years (Flagel and Wendel, 2010; Schnable et al., 2011a). Bias in gene loss between duplicate regions (fractionation bias) has been observed in *Arabidopsis* (Thomas et al., 2006) and maize (Woodhouse et al., 2010) and seems to be a general rule for whole genome duplications ranging from paramecium to fish (Sankoff et al., 2010). Bias in fractionation and genome dominance are linked because it is expected that genes on the underexpressed, non-dominant subgenome simply matter less to purifying selection and dosage-constraints (Schnable et al., 2011a). In maize, genes with known mutant phenotypes are indeed preferentially found on the dominant subgenome (Schnable and Freeling, 2011). As bias in expression predicts which subgenome will experience more fractionation following polyploidy, either subgenome identity or the expression patterns of individual gene pairs may also predict which copy of a duplicate gene pair will be more prone to duplicate gene retention in future polyploidies.

We addressed the issue of mitigation of gene dosage-constraints with two experimental systems, the grasses, and the crucifers. Both clades have roughly parallel histories of polyploidy among species with sequenced genomes (**Figure 1**; **Table 1**). Both grasses and crucifers contain a more ancient whole genome duplication which is shared by all sequenced species in the clade (Bowers et al., 2003; Paterson et al., 2004) and in both clades one well studied species with a sequenced genome has experienced a second subsequent whole genome duplication – maize in the grasses (Gaut and Doebley, 1997) and *B. rapa* in the crucifers (Lysak et al., 2005). In both cases any duplicate genes retained from the older clade-wide

polyploidy did not retain additional duplicate copies in the subsequent lineage-specific polyploidy. Therefore we were able to carry out parallel experiments to identify characteristics associated with preferential retention. It was possible to control, to some extent for the effect of protein function, by focusing on pairs of duplicate genes retained in the clade-wide polyploidy which had different fates in the subsequent lineage-specific polyploidy. A model is proposed to explain how the duplicate copies of dose–sensitive genes escape preferential retention in later polyploidies.

MATERIALS AND METHODS

DATA SOURCES

The genome assemblies and annotation used in this study were TAIR 10 (*Arabidopsis thaliana*), *Arabidopsis lyrata* v1.0 (Hu et al., 2011), the initial release of the *B. rapa* genome (The *Brassica rapa* Genome Sequencing Project Consortium, 2011), MSU 6 (*Oryza sativa*; Goff et al., 2002), *Sorghum bicolor* 1.4 (Paterson et al., 2009), and B73_refgen1 (*Zea mays*; Schnable et al., 2009).

GENE PAIR IDENTIFICATION

Orthologous genes between *A. thaliana* and *A. lyrata* were identified using SynMap (Lyons et al., 2008) with QuotaAlign settings of 1:1 (Tang et al., 2011). *Arabidopsis–Brassica* orthologous relationships were taken from Tang et al. (2012). All orthologous and homeologous relationships between grass species are those published in Schnable et al. (2012).

EXPRESSION CALCULATIONS

Gene expression levels were calculated using previously published RNA-seq data from wild type seedlings of *A. thaliana* (SRX019140: 44.7 million reads; Deng et al., 2010) and rice (SRX020118: 8.9 million reads; Zemach et al., 2010). These datasets were selected because, at the time these analysis were originally conducted they represented the RNA-seq experiments with the most sequencing depth for these two species deposited in the sequence read archive. Reads were aligned to reference genomes using Bowtie (Langmead et al., 2009) and gene expression levels were quantified using Cufflinks (Trapnell et al., 2010). Bowtie does not perform spliced alignments, which means some reads from regions of mRNA molecules which span exon junctions were not recovered in our analysis. However, given that homeologous genes will in almost all cases posses the same intron–exon structure, any bias introduced by this approach will be equivalent between gene copies.

Table 1 | Whole genome duplications.

Footnote ID from Figure 1	One name (often of many)	One citation (often of many)
1	Pre-seed plant	Jiao et al. (2011)
2	Pre-flowering plant	Jiao et al. (2011)
3	Sigma1	Tang et al. (2010)
4	Sigma2	Tang et al. (2010)
5	Pre-grass/Rho	Paterson et al. (2004)
6	Maize Lineage WGD	Gaut and Doebley (1997)
7	Gamma/pre-eudicot hexaploidy	Jaillon et al. (2007)
8	Beta	Bowers et al. (2003)
9	Alpha	Bowers et al. (2003)
10	*Brassica* hexaploidy	Lysak et al. (2005)

MEASURING PURIFYING SELECTION

Synonymous and non-synonymous substitution rates were calculated using the synonymous_calculation package included with bio-pipeline[1] using the Nei–Gojobori method (Nei and Gojobori, 1986). All other settings remained as default.

IDENTIFICATION OF RICE CNSs

Rice CNSs were identified using version 3 of the CNS Discovery pipeline[2] (Schnable et al., 2011b).

STATISTICS

p-Values for the difference in retention frequencies between singleton genes and homeologously paired genes were calculated using Fisher's Exact Test. In the crucifers, *Arabidopsis* genes with two or three retained co-orthologs in *B. rapa* were grouped together as "retained."

[1]https://github.com/tanghaibao/bio-pipeline/
[2]https://github.com/gturco/find_cns

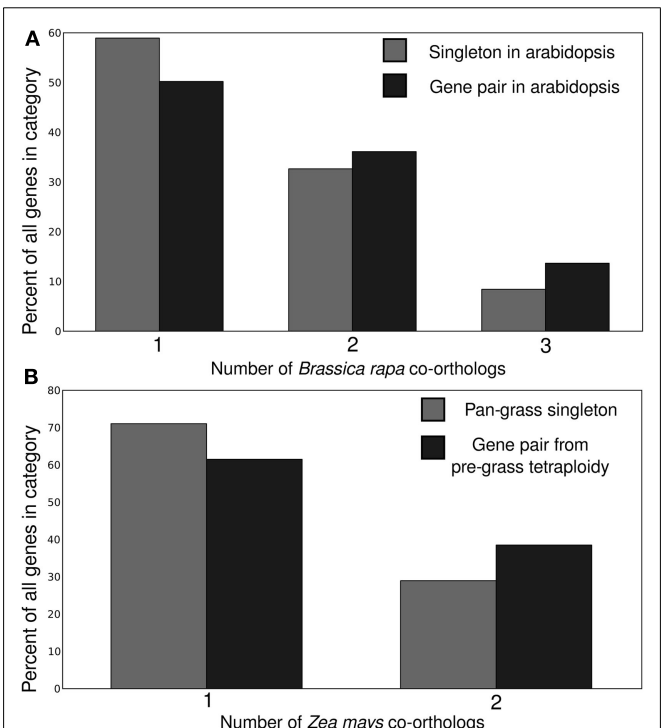


FIGURE 2 | Impact of retention in a previous whole genome duplication on retention in future whole genome duplications. (A) Proportion of *Arabidopsis* genes with one, two, or three co-orthologs in *Brassica rapa*. **(B)** Proportion of genes syntenically retained in sorghum and rice with one, or two co-orthologs in maize.

RESULTS

Genes syntenically conserved through the crucifiers or grasses were categorized as (1) those without a homeologous duplicate from the older polyploidy in each lineage (2) those with a retained homeolog from the older polyploidy in each lineage. In the crucifer lineage, the older tetraploidy is *Arabidopsis* lineage alpha (23–40 MYA); in the Poales, the earlier tetraploidy was "pre-grass" (about 70 MYA; **Figure 1**). In crucifers, these genes are classified by the number of co-orthologs conserved in *B. rapa* after the hexaploidy shared by all *Brassica* species (**Figure 2A**). In grasses, genes were classified by whether maize retained only one or both co-orthologs following the more recent tetraploidy of the *Zea*/*Tripsacum* lineage (**Figure 2B**). Retention in older polyploidies does predict retention in future polyploidies ($p < 2.2 \times 10^{-16}$ for both crucifers and grasses), as previously showing in *Arabidopsis* (Seoighe and Gehring, 2004). However in both experiments approximately half of genes previously retained as a duplicate pair in the older whole genome duplication – and therefore presumed to be sensitive to changes in gene dosage – fractionated to a single copy in the more recent whole genome duplication.

The crucifer dataset consisted of 817 *Arabidopsis* gene pairs where one copy was orthologous to only a single gene in *B. rapa* and the other possessed either two or three co-orthologs (Data Sheet S1 in Material). The grass dataset consisted of 407 gene pairs conserved in both rice and sorghum where one copy was orthologous to only a single gene in maize, its duplicate having been fractionated and the other represented by two co-orthologs in maize (Data Sheet S2 in Supplementary Material). Gene pairs result from more ancient whole genome duplications were identified and removed, as these tend to introduce confounding factors. Members of gene pairs were assigned to under and over fractionated subgenomes using differences in the number of genes syntenically retained in multiple species between homeologous regions of the rice and *Arabidopsis* genomes (Schnable et al., 2011a, 2012). In both datasets, the analysis of the relative levels of RNA encoded by duplicate genes pairs – measured by RNA-seq – was carried out in an outgroup lineage which shared only the older clade-wide polyploidy. In the grasses we used the expression of syntenic orthologs in rice and in the crucifers syntenic orthologs in *A.* thaliana (see Materials and Methods). The relative levels of purifying selection acting on each members of a gene pair were also compared using the ratio of non-synonymous substitutions to synonymous substitutions between orthologous genes in *A. thaliana* and *A. lyrata* (for the crucifers) and between rice and sorghum (for the grasses; see Materials and Methods). Promoter complexity, as measured by number of conserved non-coding

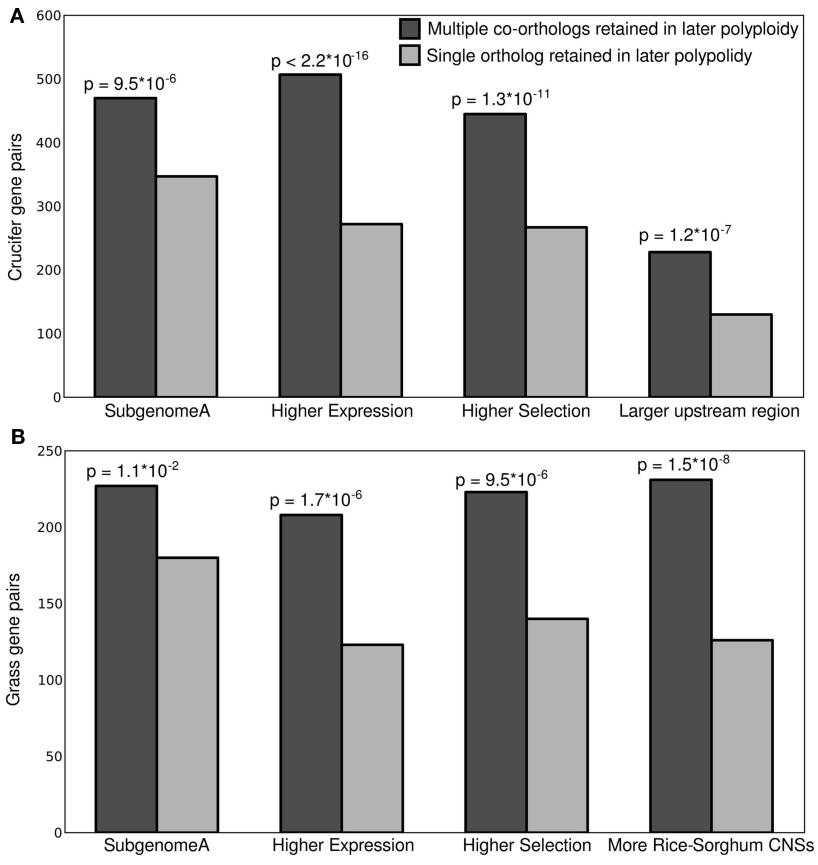


FIGURE 3 | Correlation between subsequent duplicate gene retention and a number of predicting factors including gene expression, ratio of non-synonymous to synonymous substitutions, and subgenome identity for (A) crucifer and (B) grass gene pairs. *P*-values relative to a 50/50 binomial distribution.

sequences, has previously shown to influence the odds a gene will be retained as a duplicate pair following polyploidy in the grasses (Schnable et al., 2011b) – so gene pairs were also sorted based on number of conserved non-coding sequences, in the grasses, and total quantity of upstream non-transposon sequence in *Arabidopsis*, this length being a crude proxy for promoter complexity having previously been shown to correlate with complexity of gene expression patterns (Sun et al., 2010).

All four potential markers examined showed significant power to predict which copy of a homeologous gene pair would be more resistant to fractionation in subsequent whole genome duplications (**Figure 3**). In general the gene copy retained in duplicate tended to also be the higher expressed copy, show evidence of greater purifying selection and to be associated with greater amounts of non-coding regulatory sequence. These genes also tended to be located on the dominant subgenome.

DISCUSSION

Following polyploidy, a genome possesses two or more homeologous genes, each with the same coding sequence and regulatory elements. Yet these gene copies can immediately show very different patterns of expression (Flagel et al., 2008; Buggs et al., 2011). It has been proposed that the deletion of less expressed copy of a gene following polyploidy is more likely to be selectively neutral (Schnable and Freeling, 2011; Schnable et al., 2011a). When combined with the observation that expression levels are unequal between parental subgenomes in allotetraploids (Chang et al., 2010; Flagel and Wendel, 2010; Schnable et al., 2011a), this model may explain the bias fractionation bias which has been found in ancient polyploids species (Schnable et al., 2011a).

Here we have shown that that the dominant gene copy – more expressed, under higher purifying selection, associated with more regulatory sequence – of a homeologous gene pair is more likely to retain the ancestral characteristic of preferential retention of duplicate copies in subsequent polyploidies. A number of explanations could be proposed for the link between expression and future resistance to fractionation. We propose a model based on the same link between expression and which predicts fractionation bias between parental subgenomes. If all the co-orthologs of a single ancestral gene contribute to a single pool of gene-product, the loss of less expressed gene copies would result in the smallest change in total gene-product dosage. If the total expression of a group of homeologous genes is constrained in either relative or absolute terms (Bekaert et al., 2011) smaller changes in total gene-product dosage – created by the loss of a less expressed gene copy – are predicted to be more often selectively neutral, and therefore more common (**Figure 4**). This model also predicts that, for gene pairs in *A. thaliana* where only one copy possesses any orthologous genes in *B. rapa*, it should more often be the more expressed copy; as is indeed the case (**Table A1** in Appendix).

When combined with previous results linking genome dominance with biased fractionation (Chang et al., 2010; Schnable et al., 2011a), our results suggest the Gene Dosage Hypothesis could perhaps be better thought of as the Gene-Product Dosage Hypothesis in that it can generally be considered to act on the concentration of the proteins encoded by duplicate genes, not gene copy number itself. Even when both copies of a gene are retained following whole genome duplication, the less expressed copy will often be lost in subsequent whole genome duplications. Furthermore, the greater the number of duplicate copies of a gene are found within a genome the less each individual copy contributes to total expression and the more likely it becomes that the loss of individual copies can be tolerated. In other words, the protection against fractionation provided by selection for gene dosage – either absolute or relative – becomes less powerful the less a given gene copy contributes to total expression, and the more total gene copies are present within the genome. This explains, at least in part, why despite being the "big kahuna" of whole genome duplications, plant genomes are not over-burdened with subunits of large dose-sensitive protein complexes.

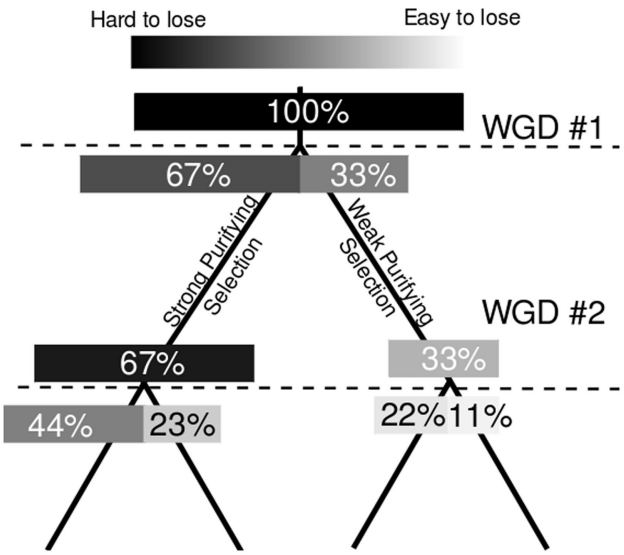


FIGURE 4 | Model for the intrinsic mitigation of gene dosage following multiple rounds of whole genome duplication. On the left, the phylogenetic tree of a perfectly retained gene after two rounds of whole genome duplication. On the right a model of how total expression is partitioned among increasing numbers of gene copies assuming genome dominance. Darkness of individual bars indicates how large an effect the loss of individual gene copies will have on total expression, and, presumably, on fitness.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Plant_Genetics_and_Genomics/10.3389/fpls.2012.00094/abstract

REFERENCES

Bekaert, M., Edger, P. P., Pires, J. C., and Conant, G. C. (2011). Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell* 23, 1719–1728.

Birchler, J. A., and Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* 19, 395–402.

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62.

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16, 1679–1691.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.

Buggs, R. J. A., Zhang, L., Miles, N., Tate, J. A., Gao, L., Wei, W., Schnable, P. S., Brad Barbazuk, W., Soltis, P. S., and Soltis, D. E. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr. Biol.* 21, 551–556.

Chang, P. L., Dilkes, B. P., McMahon, M., Comai, L., and Nuzhdin, S. V. (2010). Homeolog-specific retention and use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners. *Genome Biol.* 11, R125.

Deng, X., Lianfeng, G., Chunyan, L., Tiancong, L., Falong, L., Zhike, L., Peng, C., Yanxi, P., Baichen, W., Songnian, H., and Xiaofeng, C. (2010). Arginine methylation mediated by the *Arabidopsis* homolog of PRMT5 is essential for proper pre-mRNA splicing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19114–19119.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, J. C., Leebens-Mack, J., and dePamphilis, C. W. (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* 10, 61. doi:10.1186/1471-2148-10-61

Edger, P. P., and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17, 699–717.

Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6, 16. doi:10.1186/1741-7007-6-16

Flagel, L. E., and Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.* 186, 184–193.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453.

Gaut, B. S., and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6809–6814.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A. F., Weissenbach, J., Quétier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., and dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100.

Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D. M., Correa, L. G. G., Reski, R., Mueller-Roeber, B., and Rensing, S. A. (2010). Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* 2, 488–503.

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008). The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* 1, 181–190.

Lysak, M. A., Koch, M. A., Pecinka, A., and Schubert, I. (2005). Chromosome triplication found across the tribe Brassiceae. *Genome Res.* 15, 516–525.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van De Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. A., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. A., Wang, Y., Zhang, L., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., Rahman, M., Ware, D., Westhoff, P., Mayer, K. F., Messing, J., and Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556.

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9903–9908.

Sankoff, D., Zheng, C., and Zhu, Q. (2010). The collapse of gene complement following whole genome duplication. *BMC Genomics* 11, 313. doi:10.1186/1471-2164-11-313

Schnable, J. C., and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE* 6, e17855. doi:10.1371/journal.pone.0017855

Schnable, J. C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.* 4, 265–277.

Schnable, J. C., Springer, N. M., and Freeling, M. (2011a). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074.

Schnable, J. C., Pedersen Brent, S., Sabarinath, S., and Michael, F. (2011b). Dose-sensitivity, conserved non-coding sequences and duplicate gene retention through multiple tetraploidies in the grasses. *Front. Plant Sci.* 2:2. doi:10.3389/fpls.2011.00002

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento,

L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C. T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A. P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J. M., Deragon, J. M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115.

Sémon, M., and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512.

Seoighe, C., and Gehring, C. (2004). Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet.* 20, 461–464.

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., Sankoff, D., Depamphilis, C. W., Wall, P. K., and Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348.

Sun, X., Zou, Y., Nikiforova, V., Kurths, J., and Walther, D. (2010). The complexity of gene expression dynamics revealed by permutation entropy. *BMC Bioinformatics* 11, 607. doi:10.1186/1471-2105-11-607

Tang, H., Bowers, J. E., Wang, X., and Paterson, A. H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl. Acad. Sci. U.S.A.* 107, 472–477.

Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12, 102. doi:10.1186/1471-2105-12-102

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., Wang, X., Freeling, M., and Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model for paleohexaploidy. *Genetics* 190, 1563–1574.

The *Brassica rapa* Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 8, e1000409. doi:10.1371/journal.pbio.1000409

Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328, 916–919.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 January 2012; paper pending published: 12 March 2012; accepted: 24 April 2012; published online: 15 May 2012.

Citation: Schnable JC, Wang X, Pires JC and Freeling M (2012) Escape from preferential retention following repeated whole genome duplications in plants. Front. Plant Sci. ***3****:94. doi: 10.3389/fpls.2012.00094*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

APPENDIX

Table A1 | Expression in *Arabidopsis* and complete gene loss in *Brassica rapa*.

	Less expressed copy lost in *Brassica rapa*	**More expressed copy lost in *Brassica rapa***	***p*-Value**
All alpha pairs where one copy has been completely lost in *Brassica rapa*	428 gene pairs	217 gene pairs	$p = 3.60 \times 10^{-17}$
Alpha pairs where there are multiple co-orthologs in *Brassica rapa* of the retained copy	271 gene pairs	98 gene pairs	$p = 3.48 \times 10^{-20}$
Both copies expressed above five FPKM in *Arabidopsis thaliana*	191 gene pairs	128 gene pairs	$p = 2.49 \times 10^{-4}$

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 29 November 2012
doi: 10.3389/fpls.2012.00261

The impact of genome triplication on tandem gene evolution in *Brassica rapa*

Lu Fang†, Feng Cheng, Jian Wu and Xiaowu Wang*

Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China

Edited by:
Michael Freeling, University of California, USA

Reviewed by:
Steven B. Cannon, USDA – Agricultural Research Service, USA
Margaret Woodhouse, University of California, USA

****Correspondence:***
Xiaowu Wang, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China.
e-mail: wangxw@mail.caas.net.cn

*†**Present address:***
Lu Fang, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.

Whole genome duplication (WGD) and tandem duplication (TD) are both important modes of gene expansion. However, how WGD influences tandemly duplicated genes is not well studied. We used *Brassica rapa*, which has undergone an additional genome triplication (WGT) and shares a common ancestor with *Arabidopsis thaliana*, *Arabidopsis lyrata*, and *Thellungiella parvula,* to investigate the impact of genome triplication on tandem gene evolution. We identified 2,137, 1,569, 1,751, and 1,135 tandem gene arrays in *B. rapa*, *A. thaliana*, *A. lyrata*, and *T. parvula* respectively. Among them, 414 conserved tandem arrays are shared by the three species without WGT, which were also considered as existing in the diploid ancestor of *B. rapa*. Thus, after genome triplication, *B. rapa* should have 1,242 tandem arrays according to the 414 conserved tandems. Here, we found 400 out of the 414 tandems had at least one syntenic ortholog in the genome of *B. rapa*. Furthermore, 294 out of the 400 shared syntenic orthologs maintain tandem arrays (more than one gene for each syntenic hit) in *B. rapa*. For the 294 tandem arrays, we obtained 426 copies of syntenic paralogous tandems in the triplicated genome of *B. rapa*. In this study, we demonstrated that tandem arrays in *B. rapa* were dramatically fractionated after WGT when compared either to non-tandem genes in the *B. rapa* genome or to the tandem arrays in closely related species that have not experienced a recent whole genome polyploidization event.

Keywords: whole genome duplication, tandem duplication, tandem gene evolution, *Brassica rapa, Arabidopsis thaliana, Arabidopsis lyrata, Thellungiella parvula*

INTRODUCTION

Gene copy number can be expanded through many ways, including whole genome duplication (WGD), tandem duplication (TD), segmental duplication, and gene transposition duplication. Among these four kinds of duplications, WGD played an important role in the evolution of eukaryotes and was well documented in many sequenced genomes (Semon and Wolfe, 2007; Edger and Pires, 2009). It has been demonstrated that most eudicot plants originated from an ancient hexaploid ancestor, followed by lineage-specific tetraploidizations in many taxas: one in *Populus* (Tuskan et al., 2006), two in *Arabidopsis* (Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003), one in legumes (Cannon et al., 2006), three in *Brassica* (Wang et al., 2011), but none in *Vitis* (Jaillon et al., 2007), or papaya (Ming et al., 2008). Consequently, a single-copy gene in an ancestral angiosperm a million years ago could have expanded into a large gene family in recent species by WGD (Semon and Wolfe, 2007). TD is another important way for gene expansion. Genes expanded by TD are always distributed together as a cluster in chromosomes.

Whole genome duplication differs from TD in that WGD increases the dosage of all genes simultaneously and creates duplicate, potentially redundant, copies of all the genes within a genome. As reported previously, gene families expanded by WGD could maintain proper balance in the biological network or cascade (Freeling and Thomas, 2006). After WGD and following gene fractionation, the number of genes responding to abiotic and biotic stresses and with membrane protein functions was increased (Rizzon et al., 2006). TD is the most studied mechanism for the expansion of some gene families, such as genes that respond to environmental factors. In plants that cannot escape from stresses, who must endure turbulently changing environments and prevent themselves from being wounded (Freeling, 2009), the genes related to stress defense need to expand to resist environmental stimulation. It has been reported that TD expanded genes were more closely associated with stress-related functions than the non-TD expanded genes (Parniske et al., 1997; Michelmore and Meyers, 1998; Lucht et al., 2002; Kovalchuk et al., 2003; Leister, 2004; Shiu et al., 2004; Maere et al., 2005; Mondragon-Palomino and Gaut, 2005; Rizzon et al., 2006). Thus, the amplification of stress response genes by TD is regarded as a mechanism for protecting plants from harmful stresses. Tandem genes can be divided into classes based on gene dosage. The low-TD class appears to be represented by highly conserved, housekeeping, or key regulatory gene families, such as the transcription factor families and the proteasome 20S subunit family, while the medium- and high-TD classes are represented by gene families involved in responses to abiotic and biotic stimulus, such as pathogen defenses like NBS-LRR, or diverse enzymatic functions (Cannon et al., 2004). In comparison to WGD, TD occurred much more frequently and was responsible for the adaptive evolution of plants to rapidly changing environments (Hanada et al., 2008).

Genes are amplified through WGD and TD in a biased manner. According to gene dosage constraints (Edger and Pires, 2009), there was a functional bias in genes retained after WGD and TD.

Gene families that need to maintain proper balance in the biological network or cascade were over-retained after WGD (Freeling and Thomas, 2006). In *Brassica rapa*, the genes expanded via whole genome triplication (WGT) tend to be in functional categories such as transcriptional regulation, ribosomes, response to abiotic or biotic stimuli, response to hormonal stimuli, cell organization, and transporter functions. In contrast, TD is responsible for the expansion of genes in categories such as response to environmental stimuli, defense response, various transport functions, and metabolism. The relationship between WGT and TD shows positive or negative correlations in the expansion of different gene families. Recent studies (Hanada et al., 2008) indicated that genes involved in biotic stress responses, such as the Receptor-Like kinase family, were over-retained by both polyploidy (WGD) and local duplication (TD) in *Arabidopsis thaliana* lineages, demonstrating positive correlation. However, some gene categories, such as "transcription factors" and "ribosomal proteins," were over-retained post-WGD (Edger and Pires, 2009; Freeling, 2009), but under-retained post-TD (Freeling and Thomas, 2006), which shows the negative relationship. The Gene Balance Hypothesis can explain the reciprocal pattern (negative relationship) between WGD and TD. However, the impact of WGD on the evolution of tandem genes is still unknown.

In this work, we studied the impact of WGT on tandem gene evolution in the recently sequenced genome of *B. rapa*, which is one of the most important vegetable crops. The annotation of *B. rapa* whole genome sequences provides the opportunity for the study of gene expansion after WGT (Wang et al., 2011), and the *B. rapa* genome serves as a model system to study the impact of WGT on tandem gene evolution. *A. thaliana*, *Arabidopsis lyrata*, *Thellungiella parvula*, and *B. rapa* belong to the family Brassicaceae and have a close relationship in the phylogenetic tree (**Figure A1** in Appendix). They have undergone recurring WGD and TD in their evolutionary history. *B. rapa* underwent an additional genome triplication compared with the genomes of *A. thaliana*, *A. lyrata*, and *T. parvula*. The four species share a common ancestor from Brassicaceae, and the lineages of these three plants, *A. thaliana* (The Arabidopsis Genome Initiative, 2000), *A. lyrata* (Hu et al., 2011), and *T. parvula* (Dassanayake et al., 2011), can be regarded as controls for the pre-WGT lineages of *B. rapa*. Here, *A. thaliana*, *A. lyrata*, and *T. parvula* were used as out-groups for the investigation of the impact of WGT on the evolution of tandem genes in *B. rapa*.

RESULTS

TANDEM DUPLICATION IN *A. THALIANA*, *A. LYRATA*, *T. PARVULA*, AND *B. RAPA*, AND THE SHARED SYNTENIC TANDEM ARRAYS AMONG THEM

Tandem gene arrays contain homologous duplicates that are near to each other. Here, tandemly arrayed genes were defined as a list of paralogous genes ($\geq$2) with sequence homology satisfying the BLASTP *E*-value $< 10^{-20}$, and they should not contain more than one intervening gene among them (The Arabidopsis Genome Initiative, 2000). With the above rules, we detected 2,137, 1,569, 1,751, and 1,135 tandem gene arrays in *B. rapa*, *A. thaliana*, *A. lyrata*, and *T. parvula*, respectively. The distribution of the number of paralogous genes in each tandem array is shown in **Figure 1A**. Tandem arrays with two genes were predominant. The distribution of gene numbers in tandem arrays was not significantly different among the four species. Among these tandem gene arrays, 414 were syntenic tandem arrays shared among *A. thaliana*, *A. lyrata*, and *T. parvula* (**Figure 2**; Table 1 in Supplementary Material). The distribution of gene numbers in shared tandem arrays is shown in **Figure 1B**. It was similar to the distribution of all tandemly duplicated genes in **Figure 1A**. These shared tandem arrays should have originated from a common ancestor and been retained by all three species. This set of shared tandem gene arrays was used as the set of presumed ancestral tandem arrays, to investigate their evolution in *B. rapa* after WGT.

The number of genes that the 414 conserved tandem arrays contain in the three non-WGT species is 1,093 in *A. thaliana*, 1,090 in *A. lyrata*, and 998 in *T. parvula*. Of 414 tandem arrays, 400 had at least one syntenic ortholog in the genome of *B. rapa* (Table 1 in Supplementary Material). Among these 400 tandem arrays, 294 were syntenic to at least one tandem array in the three sub-genomes of *B. rapa*. Of 294 tandem arrays, 178 only exist in one sub-genome while 100 tandem arrays exist in two sub-genomes. However, only 16 tandem arrays exist in all three sub-genomes (LF, MF1, and MF2) simultaneously. Because *B. rapa* experienced an extra WGT, some tandem arrays had two or three copies in *B. rapa* in total, there were 426 tandem arrays in *B. rapa* corresponding to these 294 shared tandem arrays (Table 1 in Supplementary Material). The gene number in the 426 tandem arrays of *B. rapa* is 1,096. After the post-WGT fractionation, 426 tandem arrays maintained TDs, 418 tandem arrays were reduced to one copy, and 398 tandem arrays disappeared. These distributions are shown in **Table 1**.

IMPACT OF WGT ON SHARED TANDEM GENE ARRAYS IN *B. RAPA*

Compared with *A. thaliana*, *A. lyrata*, and *T. parvula*, *B. rapa* experienced an additional genome triplication, thus the ratio of syntenic genes between *B. rapa* and the three species should be 3:1. However, the ratio was much smaller for genes that were fractionized following WGT (Woodhouse et al., 2010). If WGT has no impact on TD then the ratio of syntenic tandem arrays between *B. rapa* and the other three species should be consistent with the ratio of syntenic non-tandem genes between them.

We identified 15,791 shared syntenic genes among *A. thaliana*, *A. lyrata*, and *T. parvula*, of which 13,915 genes had syntenic orthologs in *B. rapa* (Table 1 in Supplementary Material). Due to the WGT and subsequent gene fractionation, the total number of syntenic genes in *B. rapa* was 22,630 corresponding to the 13,915 syntenic orthologs (Cheng et al., 2012a; Table 1 in Supplementary Material). These 22,630 genes can be regarded as the genes retained in *B. rapa* after its genome triplication. If the WGT has no specific impact on the evolution or the loss of tandem gene arrays, the ratio of tandem arrays retained in *B. rapa* to the other three species should be in accordance with the ratio observed for all non-tandem genes. However, they were significantly different from each other (*P*-value=1.30e-06, **Table 2**). We also performed individual tests between *B. rapa* and *A. thaliana*, *B. rapa* and *A. lyrata*, as well as *B. rapa* and *T. pravula* (**Table 3**). The *P*-values were 3.47e-03, 1.35e-03, and 1.85e-17, respectively. These results indicated that tandem gene arrays in *B. rapa* significantly decreased after WGT compared with the non-WGT species.

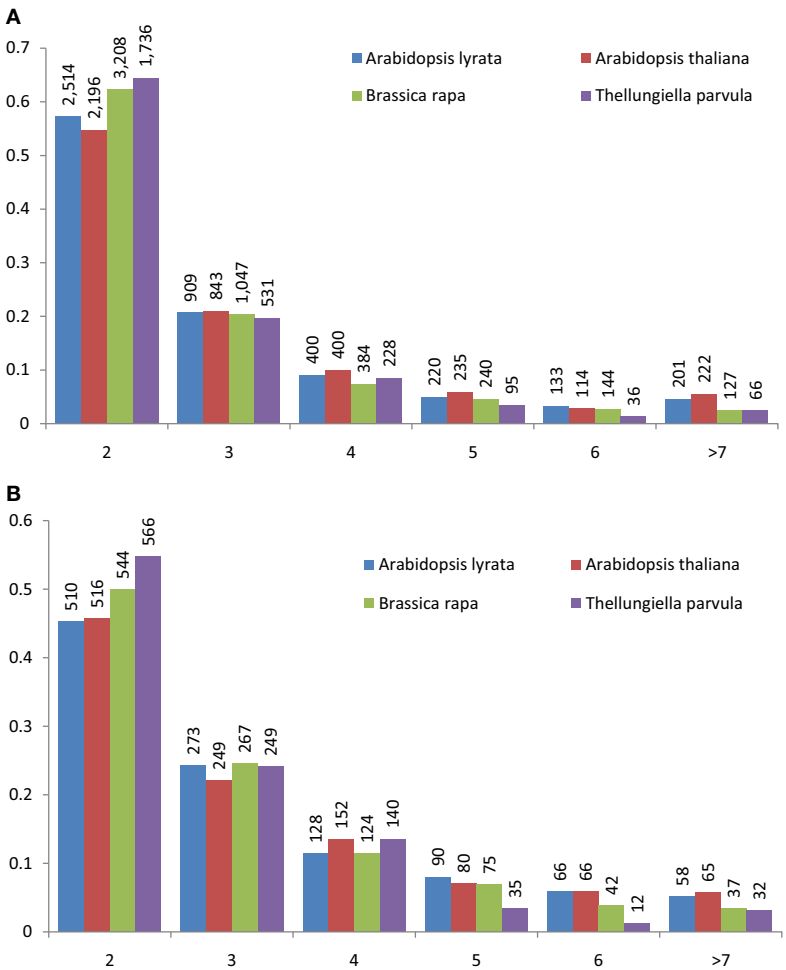


FIGURE 1 | Distribution of tandemly repeated gene arrays in the *Arabidopsis thaliana, Arabidopsis lyrata, Brassica rapa,* and *Thellungiella parvula* genomes. The number of genes in each tandem array mostly ranged from 2 to 6; data from tandem arrays with more than seven genes was combined. Tandemly repeated gene arrays were identified using the BLASTP program with a threshold of $E < 10^{-20}$. One unrelated gene among cluster members was tolerated. In both **(A)** and **(B)**, the frequency of tandem gene number is shown on the vertical axis, and the number of tandemly duplicated genes in the arrays is shown below the horizontal axis. The histogram shows the number of clusters in the genome containing two to *n* similar gene units in tandem. **(A)** The distribution of gene number in all tandem arrays of *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. parvula*. **(B)** The distribution of gene number in the shared syntenic tandem arrays among *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. parvula*.

We further looked into the loss of tandem arrays in each of the three sub-genomes of *B. rapa*: the least fractionated genome (LF), the medium fractionated genome (MF1), and the most fractionated genome (MF2; Wang et al., 2011; Cheng et al., 2012b). The *P*-values of Fisher's exact test with LF, MF1, and MF2 were 2.08e-05, 8.21e-04, and 3.29e-03, respectively (**Table 4**), indicating that tandem arrays in LF, MF1, and MF2 sub-genomes were significantly decreased in *B. rapa*. Furthermore, paired *t*-tests on gene numbers for each shared tandem array between the *B. rapa* sub-genomes and the other three species showed that the gene numbers in tandem arrays of *B. rapa* were significantly less than in *A. thaliana*, *A. lyrata*, and *T. parvula* (Table 2 in Supplementary Material). On the whole, the number of shared tandem arrays and the number of genes in tandem arrays were significantly decreased in *B. rapa* after the WGT.

THE EVOLUTION OF TANDEM GENES BETWEEN SPECIES WITHOUT THE EXTRA WGT

To verify the impact of WGT on the loss of tandem arrays in *B. rapa*, we selected *A. lyrata*, which has not experienced an extra WGT, and performed the same tests as on *B. rapa*. If the tandem arrays in *A. lyrata* decreased significantly, as was observed in *B. rapa*, then the loss of tandem genes in *B. rapa* would not be the impact of the WGT but a general process in different species. For the 391 syntenic tandem arrays shared among *A. thaliana*, *B. rapa*, and *T. parvula*, 336 can be found in *A. lyrata* (Table 3 in Supplementary Material). Meanwhile, for the 16,063 shared syntenic genes among *A. thaliana*, *B. rapa*, and *T. parvula*, 15,327 had syntenic orthologs in *A. lyrata* (Table 3 in Supplementary Material). Fisher's exact test for the loss of the tandem arrays and non-tandem genes in *A. lyrata* gave a *P*-value of 0.08733

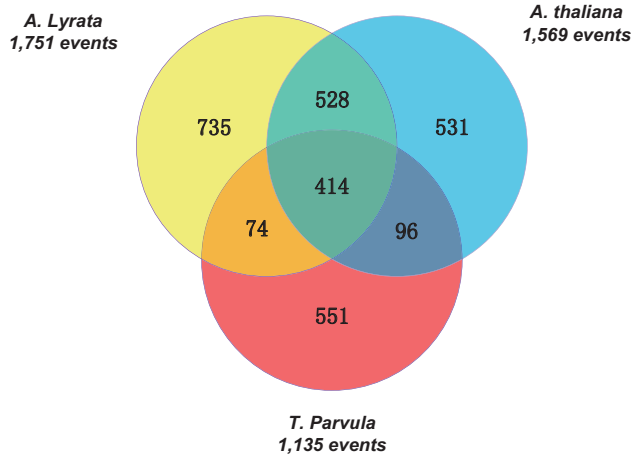


FIGURE 2 | Venn diagram showing unique and shared tandem arrays between and among the three dicotyledonous species (*Arabidopsis thaliana, Arabidopsis* lyrata, and Thellungiella parvula).

Table 1 | The distribution of gene numbers in tandem arrays from three sub-genomes of *Brassica rapa* after whole genome triplication (WGT).

	LF	MF1	MF2	Total
0 gene	88	137	173	398
1 gene	141	144	133	418
>1 genes	185	133	108	426
Total	414	414	414	1,242

"0 gene" refers to the genes in the tandem arrays of the B. rapa's ancestor that were all lost in the sub-genome of B. rapa.
"1 gene" refers to the genes in the tandem arrays of the B. rapa's ancestor that were fractionated to a single gene in the sub-genome of B. rapa.
">1 genes" refers to the tandem arrays of the B. rapa's ancestor that were maintained in the sub-genome of B. rapa.

Table 2 | Statistical test between the number of syntenic tandem arrays and all syntenic genes among *Arabidopsis thaliana, Arabidopsis lyrata*, and *Thellungiella parvula*, and in *Brassica rapa*.

	Number of shared tandem arrays	Number of syntenic genes	*P*-value[c]
[a]*At-Aly-Tp*	414	15,791	1.303e-06
[b]Syntenic in *B. rapa*	426	22,630	

[a]*The number of shared tandem arrays and syntenic genes among A. thaliana, A. lyrata, and T. parvular.*
[b]*The number of shared tandem arrays and syntenic genes in B. rapa.*
[c]*Fisher's exact test.*

(**Table 5**). The tandem arrays in *A. lyrata* did not significantly decrease. This result demonstrated that WGT accelerated the loss of tandem arrays in *B. rapa*.

Table 3 | Fisher's exact test between the number of shared syntenic tandem arrays and all syntenic non-tandem genes in *Arabidopsis thaliana, Arabidopsis lyrata, Thellungiella parvula*, and *Brassica rapa*, respectively.

	Number of shared tandem arrays	Number of syntenic gene	*P*-value[a]
A. thaliana	658	18,388	3.47e-03
B. rapa	917	29,538	
A. lyrata	603	18,125	1.35e-03
B. rapa	854	30,250	
T. parvula	524	17,303	1.85e-17
B. rapa	524	29,433	

[a]*Fisher's exact test.*

Table 4 | Statistical test between the number of tandem arrays and the number of non-tandem genes in the three species, *Arabidopsis thaliana, Arabidopsis lyrata*, and *Thellungiella parvula*, and *Brassica rapa*'s three sub-genomes.

	Number of tandem arrays	Number of gene	*P*-value[a]
LF of *B. rapa*	185	10,145	2.08e-05
At-Aly-Tp	414	15,791	
MF1 of *B. rapa*	133	6,950	8.21e-04
At-Aly-Tp	414	15,791	
MF2 of *B. rapa*	108	5,535	3.29e-03
At-Aly-Tp	414	15,791	

[a]*Fisher's exact test.*

Table 5 | Statistical test between the number of shared syntenic tandem arrays and all syntenic non-tandem genes in *Arabidopsis thaliana, Brassica rapa*, and *Thellungiella parvula*, and *Arabidopsis lyrata*.

	Number of shared tandem arrays	Number of syntenic gene	*P*-value[b]
[a]*At-Bra-Tp*	391	16,063	0.088
A. lyrata	336	15,327	

[a]*Tandem arrays with a shared syntenic relationship among A. thaliana, B. rapa, and T. parvula.*
[b]*Fisher's exact test.*

DISCUSSION

Brassica rapa experienced both a WGT and local duplications. Since its divergence from *A. thaliana*, *A. lyrata*, and *T. parvula*, *B. rapa* has undergone genome triplication, extensive gene fractionation, and genomic reshuffling that eroded its resemblance to ancestral Brassicales (Wang et al., 2011). The modes of WGD and TD in gene expansion have a reciprocal relationship. Furthermore, the WGD should have had an impact on tandem gene evolution. It is clear from our results that the shared tandem arrays among

A. thaliana, *A. lyrata*, and *T. parvula* decreased significantly in *B. rapa* compared with the syntenic non-tandem genes. However, the shared tandem arrays among *A. thaliana*, *B. rapa*, and *T. parvula* did not significantly decrease in *A. lyrata*.

Previous reports showed that tandem duplicated genes tend to be involved in responses to stress or environmental stimuli (Parniske et al., 1997; Michelmore and Meyers, 1998; Lucht et al., 2002; Kovalchuk et al., 2003; Leister, 2004; Shiu et al., 2004; Maere et al., 2005; Mondragon-Palomino and Gaut, 2005; Rizzon et al., 2006) in plants and their cells. The need for stress endurance makes plants rich in genes associated with environmental factor responses. WGD expanded all genes in *B. rapa* simultaneously, including genes that respond to environmental factors. For example, one gene that defends against abiotic or biotic stimuli is increased to three genes through WGT in *B. rapa*. Though genes were rapidly fractionized and lost following WGD, genes that respond to environmental adaptability and hormones were still over-retained after WGT (Table 4 in Supplementary Material). There were many stress resistance genes generated from WGT in *B. rapa*, so it need not amplify these genes by TD. Additionally, redundant genes in response to abiotic and biotic stimuli would be lost during the evolution of tandem genes after their WGT expansion.

It has been characterized that tandem genes experience a rapid birth-and-death evolution (Nei and Rooney, 2005). Rapid birth-and-death evolution has occurred in many gene families that have tandem duplicates, such as plant disease resistance genes (Parniske et al., 1997; Michelmore and Meyers, 1998). With this feature, tandem genes would disappear in original positions and appear or expand in other genomic regions in *B. rapa*. That would also lead to a decrease in syntenic tandem arrays among *B. rapa* and other species.

CONCLUSION

The evolution of tandemly duplicated genes in *B. rapa* has been affected by the WGT event. Following WGT, the triplicated tandem genes in *B. rapa* were largely lost. The ratio of lost tandem arrays is significantly larger than the ratio of lost non-tandem genes. All ancestral tandem arrays were triplicated by WGT in *B. rapa*. Genes in these triplicated tandem arrays then became functionally redundant and were prone to be lost in *B. rapa*, both in the number of tandem arrays and in the number of genes in each tandem.

MATERIALS AND METHODS

DATA SOURCES

Genomic data for the four plant species (*A. thaliana*, *A. lyrata*, *T. parvula*, and *B. rapa*) were obtained from the databases of The Arabidopsis Information Resource (TAIR)[1], the Joint Genome Institute[2], the *T. parvula* genome sequencing group, and the *Brassica* database (BRAD[3]; Cheng et al., 2011).

TANDEM ARRAY IDENTIFICATION

Tandem gene arrays were defined as homologous gene clusters with no more than one intervening gene located among them and sequence homology of the homologous genes in the array should satisfy BLASTP *E*-value$<10^{-20}$.

SYNTENIC GENE IDENTIFICATION

Syntenic genes between *A. thaliana* and *B. rapa*, *A. thaliana* and *A. lyrata*, and *A. thaliana* and *T. parvula* were identified using SynOrths (Cheng et al., 2012a). We took one gene from each tandem array as a representative to determine the syntenic relationships among the four species.

ACKNOWLEDGMENTS

We thank Dong-Ha Oh for his help in retrieving the *T. parvula* dataset. The work was funded by National Program on Key Basic Research Projects (The 973 Program: 2012CB113900), and the National High Technology R&D Program of China (2012AA100201) to Xiaowu Wang. The work was done in the Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture, P. R. China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Plant_Genetics_and_Genomics/10.3389/fpls.2012.00261/abstract

[1] www.arabidopsis.org
[2] http://genome.jgi-psf.org/Araly1/Araly1.home.html
[3] http://brassicadb.org/brad/

REFERENCES

Blanc, G., Hokamp, K., and Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* 13, 137–144.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol.* 4:10. doi:10.1186/1471-2229-4-10

Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., et al. (2006). Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14959–14964.

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., et al. (2011). BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biol.* 11:136. doi:10.1186/1471-2229-11-136

Cheng, F., Wu, J., Fang, L., and Wang, X. (2012a). Syntenic gene analysis between Brassica rapa and other Brassicaceae species. *Front. Plant Sci.* 3:198. doi:10.3389/fpls.2012.00198

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., et al. (2012b). Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS ONE* 7:e36442. doi:10.1371/journal.pone.0036442

Dassanayake, M., Oh, D. H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., et al. (2011). The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.* 43, 913–918.

Edger, P. P., and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17, 699–717.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453.

Freeling, M., and Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16, 805.

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., and Shiu, S. H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* 148, 993–1003.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Kovalchuk, I., Kovalchuk, O., Kalck, V., Boyko, V., Filkowski, J., Heinlein, M., et al. (2003). Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature* 423, 760–762.

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* 20, 116–122.

Lucht, J. M., Mauch-Mani, B., Steiner, H. Y., Metraux, J. P., Ryals, J., and Hohn, B. (2002). Pathogen stress increases somatic recombination frequency in Arabidopsis. *Nat. Genet.* 30, 311–314.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459.

Michelmore, R. W., and Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* 8, 1113–1130.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452, 991–996.

Mondragon-Palomino, M., and Gaut, B. S. (2005). Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana. *Mol. Biol. Evol.* 22, 2444–2456.

Nei, M., and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* 39, 121–152.

Parniske, M., Hammond-Kosack, K. E., Golstein, C., Thomas, C. M., Jones, D. A., Harrison, K., et al. (1997). Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* 91, 821–832.

Rizzon, C., Ponger, L., and Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput. Biol.* 2:e115. doi:10.1371/journal.pcbi.0020115

Semon, M., and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512.

Shiu, S. H., Karlowski, W. M., Pan, R., Tzeng, Y. H., Mayer, K. F., and Li, W. H. (2004). Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* 16, 1220–1234.

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13627–13632.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–815.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313, 1596–1604.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., et al. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 8:e1000409. doi:10.1371/journal.pbio.1000409

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 June 2012; accepted: 10 November 2012; published online: 29 November 2012.

*Citation: Fang L, Cheng F, Wu J and Wang X (2012) The impact of genome triplication on tandem gene evolution in Brassica rapa. Front. Plant Sci. **3**:261. doi: 10.3389/fpls.2012.00261*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

APPENDIX

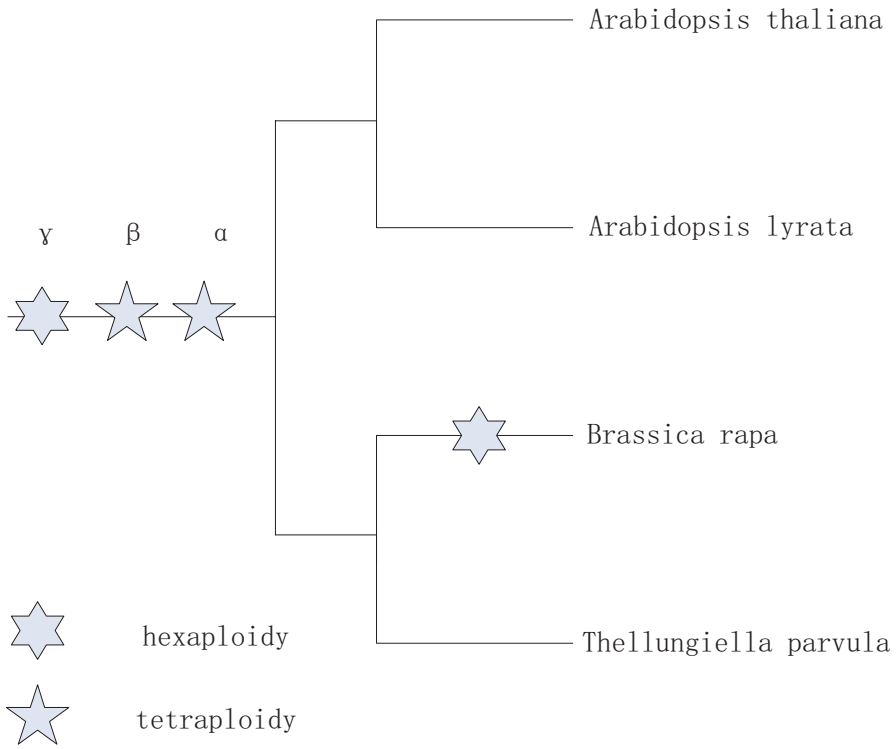


FIGURE A1 | The phylogenetic tree as above comes from website (http://www.phytozome.net/).

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 September 2012
doi: 10.3389/fpls.2012.00208

Identification and characterization of orthologs of *AtNHX5* and *AtNHX6* in *Brassica napus*

Brett A. Ford, Joanne R. Ernest and Anthony R. Gendall*

Department of Botany, La Trobe University, Melbourne, VIC, Australia

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Kees Venema, Consejo Superior de Investigaciones Científicas, Spain
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Anthony R. Gendall, Department of Botany, La Trobe University, Bundoora, Melbourne, VIC, Australia.
e-mail: t.gendall@latrobe.edu.au

Improving crop species by breeding for salt tolerance or introducing salt tolerant traits is one method of increasing crop yields in saline affected areas. Extensive studies of the model plant species *Arabidopsis thaliana* has led to the availability of substantial information regarding the function and importance of many genes involved in salt tolerance. However, the identification and characterization of *A. thaliana* orthologs in species such as *Brassica napus* (oilseed rape) can prove difficult due to the significant genomic changes that have occurred since their divergence approximately 20 million years ago (MYA). The recently released *Brassica rapa* genome provides an excellent resource for comparative studies of *A. thaliana* and the cultivated *Brassica* species, and facilitates the identification of *Brassica* species orthologs which may be of agronomic importance. Sodium hydrogen antiporter (NHX) proteins transport a sodium or potassium ion in exchange for a hydrogen ion in the other direction across a membrane. In *A. thaliana* there are eight members of the NHX family, designated *AtNHX1-8*, that can be sub-divided into three clades, based on their subcellular localization: plasma membrane (PM), intracellular class I (IC-I) and intracellular class II (IC-II). In plants, many NHX proteins are primary determinants of salt tolerance and act by transporting Na^+ out of the cytosol where it would otherwise accumulate to toxic levels. Significant work has been done to determine the role of both PM and IC-I clade members in salt tolerance in a variety of plant species, but relatively little analysis has been described for the IC-II clade. Here we describe the identification of *B. napus* orthologs of *AtNHX5* and *AtNHX6*, using the *B. rapa* genome sequence, macro- and micro-synteny analysis, comparative expression and promoter motif analysis, and highlight the value of these multiple approaches for identifying true orthologs in closely related species with multiple paralogs.

Keywords: Arabidopsis, NHX, antiporter, Brassica, sodium transport, potassium transport, pH, cation transport

INTRODUCTION

It is estimated that more than 800 million hectares of land worldwide and approximately 20% of irrigated farmland are negatively impacted by salinity (FAO, 2009). While there are many different salts that contribute to the salinization of a landscape, by far the most abundant and damaging is NaCl (Tester and Davenport, 2003). NaCl in the soil inhibits plant growth by causing an initial osmotic stress, followed by the accumulation of Na^+ ions in the plant to toxic levels (Munns and Tester, 2008). Improved crop yield under saline conditions can be achieved by identifying genes which confer salt tolerance, and introducing them to crop species by traditional breeding or transgenesis.

In plants, members of the monovalent cation/proton antiporter (CPA1) gene family exchange a sodium, potassium or lithium ion for a hydrogen ion across a cellular membrane. They can be classified into two distinct sub families, the plasma membrane (PM) localized NHAP family and the intracellular localized (IC) NHX family (Brett et al., 2005a; Chanroj et al., 2012). These proteins are important in maintaining pH and ion homeostasis, and are conserved across all phyla and kingdoms (Brett et al., 2005a). In *Arabidopsis thaliana*, there are two members of the NHAP family (SOS1/AtNHX7 and AtNHX8) and six members of the IC-NHX family, designated AtNHX1-6 (Brett et al., 2005a; Chanroj et al., 2012). The IC-NHX family can be further subdivided into two clades based on their cellular localization: the IC-I clade (*AtNHX1-4*), localized to the tonoplast; or the IC-II clade (*AtNHX5-6*), localized to endosomal compartments (Brett et al., 2005a; Bassil et al., 2011). In plants, NHX proteins are important determinants of Na^+ tolerance, detoxifying cytosolic Na^+ by transportation out of the cell to the apoplastic space, or by sequestration into subcellular compartments (Apse et al., 1999; Shi et al., 2003; Rodriguez-Rosales et al., 2008). Members of the PM (*AtNHX7/SOS1*) and IC-I (*AtNHX1*) clades have been shown to be upregulated in response to NaCl, and to confer salt tolerance upon over-expression in *A. thaliana* (Apse et al., 1999; Shi et al., 2002, 2003; Shi and Zhu, 2002). Of the IC-II clade, *AtNHX5* and the rice ortholog *OsNHX5* are both up regulated in response to NaCl (Yokoi et al., 2002; Fukuda et al., 2011) and over expressing *AtNHX5* in a variety of plant species including rice (Li et al., 2011) and tomato (Rodriguez-Rosales et al., 2008) has been shown to increase their tolerance to NaCl. Recent work has shown that *AtNHX5* and *AtNHX6* are functionally redundant,

are important for normal plant growth and development, are endosome associated and have an important role in protein trafficking to the vacuole (Bassil et al., 2011). A similar role in protein trafficking has been demonstrated for the yeast ortholog ScNHX1 (Bowers et al., 2000; Brett et al., 2005b). Interestingly, comparatively little work has been undertaken investigating the role of the Class II NHX proteins from important crop species in NaCl tolerance.

Brassica napus is a member of the Brassicaceae family and is an important oil crop species. *B. napus* cultivation has increased significantly over recent years and it is now the second largest crop for the production of oil seeds and oil meals, and the third largest crop used in the production of vegetable oil (USDA, 2009). As both *A. thaliana* and *B. napus* are members of the family Brassicaceae and therefore relatively closely related, a comparative study could be undertaken to try to identify the homologs of *AtNHX5* and *AtNHX6* to assess their importance in salt tolerance. However, there have been extensive large- and small-scale genome changes in the *A. thaliana* and *B. napus* genomes since their divergence approximately 13–17 million years ago (MYA) (**Figure 1**). *B. napus* is an allopolyploid species (AACC, $2n = 38$) arising from the hybridization of *B. rapa* (AA, $2n = 20$) and *Brassica oleracea* (CC, $2n = 18$) (**Figure 1**). Additionally, all *Brassica* species appear to have undergone a whole genome triplication (WGT, Br—α) event since they diverged from *A. thaliana* (**Figure 1**). The *A. thaliana* genome has undergone extensive duplications, deletions, re-arrangements and a reduction in chromosome number even since its divergence from its close relative *Arabidopsis lyrata* only 5 MYA (**Figure 1**) (Hu et al., 2011). These confounding factors make direct comparative studies and homolog identification in *B. napus* based solely on *A. thaliana* genomic information problematic.

To overcome this problem, Schranz et al. (2006), building on the work of Parkin et al. (2005), have proposed that the comparative studies of *A. thaliana* and the *Brassica* species should be based upon an ancient karyotype of $n = 8$, that is observed in at least 37% of the Brassicaceae species. The genome of this ancient karyotype can be further segmented into 24 conserved syntenic blocks (designated A–X) that can be identified in both the *A. thaliana* genome and the *B. rapa* (A genome) component of *B. napus* (Parkin et al., 2005; Schranz et al., 2006).

While this approach greatly aided comparative studies between *A. thaliana* and *B. napus*, the lack of high quality *Brassica* sequence still inhibited homolog identification. The recent release of the fully sequenced *B. rapa* genome (Wang et al., 2011) has enabled in-depth comparative studies between *A. thaliana* and *B. napus*.

Analysis of the *B. rapa* genome reveals that there has been significant gene loss since the most recent WGT (Wang et al., 2011). The triplicated *B. rapa* genome can be divided into three sub genomes based on their level of gene loss relative to *A. thaliana*, which have been designated the least fractionated genome (LF) with 30% gene loss, the medium fractionated genome (MF1) with 54% gene loss and the most fractionated genome (MF2) with 64% gene loss (Wang et al., 2011). However, analysis of the *B. rapa* genome shows that genes associated with response to environmental factors such as salt were over-retained (Wang et al., 2011) and as a result it is possible that there are three homologs of *AtNHX5* and three homologs of *AtNHX6* in *B. rapa*. As *B. napus* has both a *B. rapa* triplicated genome, and a triplicated *B. oleracea* C genome component there may be up to six *AtNHX5* and six *AtNHX6* homologs.

Here we describe the use of the *B. rapa* genomic sequence to aid in the clear identification of the *B. napus* homologs of *AtNHX5* and *AtNHX6* as a first step in evaluating their potential to improve salt tolerance.

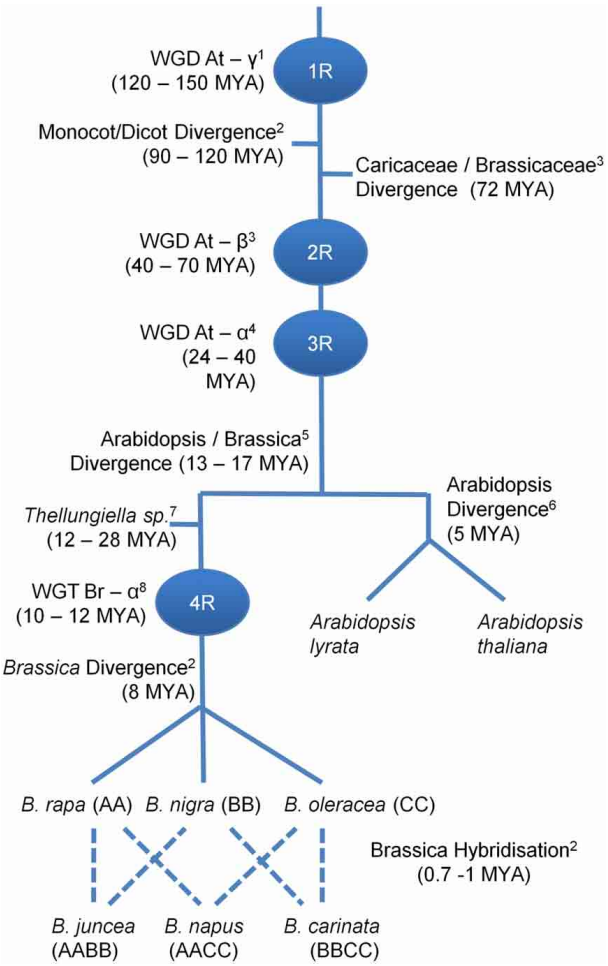


FIGURE 1 | Genome evolution of *B. napus* and *A. thaliana*. Prior to their divergence approximately 13–17 MYA *B. napus* and *A. thaliana* share a complex genome evolution including three (denoted 1R, 2R, and 3R) whole genome duplication (WGD) events. The last two WGD events post date the divergence of the Moncots/Dicots and the Caricaceae/Brassicaceae. Since their divergence the ancestor of extant Brassica species underwent a whole genome triplication event (WGT, denoted 4R) with the allopolyploid *B. napus* arising from the hybridization of *B. rapa* and *B. oleracea*. Superscript numbers relate to references for divergence times. For an in depth analysis of evolutionary history of Brassicaceae please refer to the excellent review by Franzke et al. (2011). [1](Soltis et al., 2009); [2](Mun et al., 2009); [3](Bell et al., 2010); [4](Henry et al., 2006); [5](Yang et al., 1999); [6](Koch et al., 2000); [7](Couvreur et al., 2010); [8](Lysak et al., 2005).

MATERIALS AND METHODS

IDENTIFICATION AND REANNOTATION OF *B. rapa NHX* GENES

The initial identification of the complete set of *NHX* genes in *B. rapa* used representative *A. thaliana* NHX protein sequences (See **Table 1** for accession numbers) as query sequences in TBLASTN searches of the *B. rapa* genomic database in Phytozome version 8.0 (http://www.phytozome.org; Goodstein et al., 2012). Initial searches were performed using AtNHX1, AtNHX6, and AtSOS1/AtNHX7 amino acid sequences. Hits with an *E*-value of less than 10^{-45} were classified as putative NHXs and examined further. Preliminary alignments of *A. thaliana* NHX proteins to the annotated *B. rapa* proteins revealed some minor sequence dissimilarities in annotations that often corresponded to the lack of one or more exons in the *B. rapa* gene models, or from fusion of two adjacent genes into a single mis-annotated *BrNHX* gene. To refine these predictions, two approaches were used. Expressed sequence tag (EST) or cDNA sequences were identified by using the NHX amino acid sequences as the query for a TBLASTN search of all EST databases restricting the search to the taxid *Brassica*. The EST or cDNA sequences were then assigned to the relevant *NHX* gene based on the regions of maximum match and sequence alignment to the corresponding *B. rapa* genome. These sequences were aligned with the corresponding genomic region using Spidey (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/index.html), using the "Plant" genomic sequence option and other default options. The identified exons were then incorporated into a final re-annotated open reading frame model. If no EST or cDNA support existed, amino acid sequences from the *A. thaliana* or other *B. rapa* ortholog corresponding to the mis-annotated regions were used as queries in a low stringency TBLASTN approach (expected *E*-value of 1.0), which allowed the correction of the original annotations. Gene models for *BrNHX6.1* and *BrNHX6.2* were generated using the FancyGene 1.4 gene model drawing program (http://www.bio.ieo.eu/fancygene Rambaldi and Ciccarelli, 2009). *B. rapa NHX* gene locations were determine by using the Brassica Database (http://brassicadb.org/brad).

IDENTIFICATION OF *Thellungiella parvula NHX* GENES

The identification of the intra-cellular *NHX* genes in *T. parvula* used representative Arabidopsis NHX protein sequences (See **Table 1** for accession numbers) as query sequences in BLASTP searches of the *T. parvula* predicted gene models—protein version 2.0 database in the *Thellungiella* website [http://www.phytozome.org; (Dassanayake et al., 2011; Goodstein et al., 2012)]. Initial searches were performed using AtNHX1 and AtNHX6 amino acid sequences. Hits with an E value of less than 10^{-45} were classified as putative NHXs and examined further by phylogentic analysis.

PHYLOGENETIC ANALYSIS

The predicted amino acid sequences of all predicted *B. rapa, T. parvula* and *A. thaliana* NHXs were imported into MEGA5 (Tamura et al., 2011), and aligned using ClustalW with default parameters, and the corresponding amino acid alignments used for subsequent analyses. A neighbor-joining analysis was performed, using the Poisson model of amino acid substitution, with uniform rate among sites and pairwise deletion, and the phylogeny tested with 1000 bootstrap replications.

NHX6 nucleotide sequences from *B. rapa*, *B. oleracea* and *B. napus* and *Brassica* NHX6 ESTs from Genbank were imported into MEGA5 (Tamura et al., 2011) and aligned using Clustal W with default parameters. A neighbor-joining analysis was performed using the maximum composite likelihood method with uniform rate among sites and complete deletion, and the phylogeny tested with 1000 bootstrap replications.

MICROSYNTENY OF *AtNHX5* AND *AtNHX6* AND THEIR CO-LINEARITY IN *B. rapa*

AtNHX5 and *AtNHX6* sequences were used as the query to search for homeologous regions in the three sub genomes in *B. rapa* using the syntenic paralogs search tool in the Brassica Database (http://brassicadb.org/brad). The region of *A. thaliana* chromosome 1 containing *AtNHX6* and eight flanking genes upstream and downstream was then aligned to the three co-linear genomic segments from *B. rapa*. The region of *A. thaliana* chromosome 1 surrounding *AtNHX5* was aligned in a similar manner, however due to the more complex nature of the co-linearity between this region and the corresponding *B. rapa* genomic segments the region of analysis was expanded to include 14 genes downstream and 12 genes upstream of *AtNHX5*.

BrNHX6 PROMOTER ANALYSIS

Nucleotide sequences of *AtNHX6*, *BrNHX6.1*, and *BrNHX6.2*, including 1119 bp upstream of the predicted ATG start codon, and 500 bp downstream of the predicted stop codon were obtained from Phytozome. Sequences were aligned and subjected to sliding window analysis of homology using VISualization Tool for Alignments (VISTA) web-based software (Frazer et al., 2004). Regions identified as highly conserved were analysed for conserved *cis*-regulatory elements (CRE) motifs using the Plant cis-acting regulatory DNA elements (PLACE) database (Higo et al., 1999).

PLANT GROWTH CONDITIONS

Seeds of *B. rapa* (cv Chinese cabbage), *B. oleracea* (cv White cabbage) and *B. napus* (cv Westar) were sown direct to soil in seed raising mix and vermiculite (3:1) and watered twice a week. A general purpose fertilizer (Osmocote) was applied in liquid form every two weeks and plants were grown in a 16 h light: 8 h dark photoperiod at 22°C.

For material used for qRT-PCR and isolation of cDNA, *B. napus* (cv Westar) seeds were surface sterilized by soaking in 70% ethanol for 5 min and 1% commercial bleach for 10 min. Seeds were washed three times in sterile double distilled water and plated onto half strength MS plates (Murashige and Skoog, 1962) with 0 mM or 200 mM NaCl. Three replicate plates were made for each treatment. Seeds were stratified for 72 h and grown in a 16 h light: 8 h dark photoperiod at 22°C.

CLONING OF BRASSICA NHX6 ORF SEQUENCES

Genomic DNA was extracted from *B. rapa* (cv Chinese cabbage) as described (Herrmann and Frischauf, 1987). RNA was extracted from eight week old leaf material of *B. napus* (cv Westar) and *B. oleracea* (cv white cabbage) and from two week

Table 1 | Chromosomal locations of *B. rapa NHX* genes.

A. thaliana gene name	AGI	*A. thaliana genomic segment*[a]	Corresponding *B. rapa segment locations*[b]	Br gene name	*B. rapa* gene *(scaffold, fractionated genome*[c]*)*	Location (ATG-STOP)	*Brassica sp. ESTs*
NHX1	AT5G27150	Q	2, 6, 9	BrNHX1.1	Bra020599 (A02, MF2)	24295626–24298768	*B. juncea* HQ848296, HQ848294, HQ848295
–	–	–	–	BrNHX1.2	Bra036110 (A09, MF1)	2662958–2665986	*B. rapa* EX098258.1; *B. napus*: GU192449.1
–	–	–	–	BrNHX1.3	Bra009975 (A06, LF)	∼18416170–∼18415490	None (and likely pseudogene)
NHX2	AT3G05030	F	1, 3, 5	BrNHX2	Bra039469 (A05, LF)	23480049–23483092	*B. napus* AY189676.1
NHX3	AT5G55470	C	5, 6, 8	BrNHX3	Bra002905 (A10, LF)	6853381–6850500	None
NHX4	AT3G06370	F	1, 3, 5	BrNHX4	Bra020755 (A02, LF)	23272879–23269821	None
NHX5	AT1G54370	C	5, 6, 8	–	Not detected	–	None
NHX6	AT1G79610	E	2, 7, 7	BrNHX6.1	Bra035130 (A07, LF)	22302738–22298998	*B. oleracea* DK554651.1; DK541204.1; DK548975.1; DK489391.1
–	–	–	–	BrNHX6.2	Bra003601 (A07, MF2)	14001071–14007892	*B. napus* FG553917.1; ES983056.1: CD827425.1; *B. oleracea* EE531575.1; EE530829.1; DK495073.1; DK481606.1
NHX7/SOS1	AT2G01980	K	2, 6, 9	BrSOS1	Bra017430 (A09, MF2)	15259302–15253531	*B. rapa* HQ848290; *B. napus* EU487184.1; *B. juncea* HQ848289, HQ949287, HQ848288, EF206779
NHX8	AT1G14660	A	6, 8, 9, 10	BrNHX8	Bra026197 (A06, LF)	5641695–5637360	None

[a] *A. thaliana chromosomal blocks as described in Schranz et al. (2006).*

[b] *Numbers of B. rapa segments were as described in Wang et al. (2011).*

[c] *B. rapa chromosomal location and sub-genome were assigned using the Brassica Database* (http://brassicadb.org/brad).

old whole seedlings *B. napus* (cv Westar) using the Qiagen RNeasy Kit (Qiagen #74104) as per manufacturer's instructions. Approximately 500 ng of RNA was used to synthesize cDNA using Superscript III (Life Technologies #18080) following the manufacturer's recommendations, and the cDNA was diluted to a final volume of 100 μL with sterile double distilled water.

Full length *Brassica* NHX6 ORF sequences were amplified by PCR from cDNA derived from eight week old leaf material of *B. oleracea* (Forward 5′-CACCATGTCGGAGATTTCGCCG-3′ and Reverse 5′-TAACCGGGGGCTAAATTTCTGA-3′) and *B. napus* (Foward 5′-CACCATGTCGGAGATTTCGCCG-3′ and Reverse 5′-TAACCGGGGGCTAAATTTCTGA-3′). The PCR product was then cloned into the pENTR D/TOPO vector (Life Technologies #K2435-20) as per manufacturer's instructions. Approximately 10 independent clones from each ligation were sequenced with one unique coding sequence being identified for both *B. napus* (Genbank JX082291) and *B. oleracea* (Genbank JX082292).

IDENTIFICATION AND CLONING OF PARTIAL *Brassica* NHX6 SEQUENCES

Primer design

The *B. rapa NHX6.1* and *NHX6.2* genes, the full length *B. napus* (JX082291) and *B. oleracea* (JX082292) ORF sequences as well as other *B. napus*. *B. rapa* and *B. oleracea* EST sequences that were available were aligned in Vector NTI v11.5 (Invitrogen). From this alignment regions of 100% nucleotide conservation between all divergent *Brassica* species sequences were identified and primers (Forward 5′-GCTTGAAGCCCTAGAGGTTGT-3′ and Reverse 5′-CGTTATTACTTGTGAAGAACGTGTT-3′) designed within these regions to amplify all potential A and C genome *B. napus* NHX6 homologs.

Using these primers, partial *Brassica* NHX6 sequences were amplified by PCR from cDNA of 2-week-old whole seedlings in *B. napus* and from genomic DNA in *B. rapa*. DNA from the PCR was purified using the Promega Wizard PCR clean up kit (Promega #A9282) and then cloned into pGEM T-easy vector (Promega #A1360) as per manufacturer's instructions. Approximately 10 independent clones from each ligation were sequenced using the M13_F sequencing primer.

GENE EXPRESSION ANALYSIS

Total RNA was extracted from whole seedlings of *B. napus* (cv Westar) two weeks after germination using the Qiagen RNeasy Kit (Qiagen #74,104) as per manufacturer's instructions. Total RNA (2 μg) was DNase treated with the Promega RQ1 DNase kit (Promega #M6101) as per manufacturer's instructions to remove any genomic DNA contamination.

Primers were designed for the *BnNHX6.1* gene (Forward 5′-CATCCTTTTCTCATTCTGTTCATCG-3′ and Reverse 5′-TCGAAGTCCACTGTACCAAAG-3′) and for the *BnNHX6.2* gene (Forward 5′-GATAGCCGTGATACATCCCTTG-3′ and Reverse 5′-AGTTCTGAAAATGACTTTGCGC-3′) based on the corresponding BrNHX6.1 or BrNHX6.2 open reading frame sequences identified from the *B. rapa* genome.

Quantitative real-time PCR was performed using the Bio-Rad iCycler and the iScript One Step RT-PCR kit with SYBR green (Bio-Rad #170-8892) as per manufacturer's instructions. PCR conditions consisted of a reverse transcription step at 50°C for 10 min, a reverse transcription inactivation step at 95°C for 5 min and 40 cycles of 95°C for 10 s followed by either 61°C (*BnNHX6.1*) or 60°C (*BnNHX6.2*) for 30 s. Results were visualized using BioRad iQ5 Optical System Software. A dissociation analysis was performed and the PCR products were sequenced to confirm PCR specificity to the *BnNHX6* transcript. Expression of the *BnNHX6* genes was normalized relative to the expression of BnUBC_21 (Forward 5′-CCTCTGCAGCCTCCTCAAGT-3′ and Reverse 5′-CATATCTCCCCTGTCTTGAAATGC-3′) previously validated as a reference gene for qRT-PCR in *B. napus* (Chen et al., 2010). Transcript abundance was calculated using the Pfaffl model for relative quantification with efficiency correction (Pfaffl, 2004). All experiments were completed in triplicate on three independent biological replicates.

RESULTS

IDENTIFICATION AND PHYLOGENETIC ANALYSIS OF *B. rapa NHX* GENES

As a first step to identify the *B. napus AtNHX5* and *AtNHX6* homologs, we searched the now fully sequenced *B. rapa* genome. As *AtNHX5* and *AtNHX6* are very closely related to the other *A. thaliana* NHX members *AtNHX1-4* and *AtNHX7-8*, unmabiguous identification of the corresponding *B. rapa* homologs may be difficult. To clearly identify and unambiguously assign the *AtNHX5* and *AtNHX6 B. rapa* homologs, we identified all the *NHX* genes in the genome of *B. rapa*. To determine the complete complement of *NHX* genes, we searched the *B. rapa* genome sequence and identified ten *NHX* genes (hereafter called *BrNHX;* see **Table 1** for complete details). A phylogenetic analysis of the corresponding predicted amino acid sequences of the encoded proteins revealed that *B. rapa* is likely to express five intracellular class one (IC-I) genes, two intracellular class two (IC-II) genes and two PM clade members (**Figure 2A**). Additional evidence of one pseudogene (*BrNHX1.3*) could also be detected. In some cases, multiple *B. rapa* genes could be detected that corresponded to a single *NHX* gene in *A. thaliana,* but surprisingly, in only one case could three *B. rapa* orthologs corresponding to a *AtNHX* gene be detected (*AtNHX1* and *Br.NHX1.1, Br NHX1.2* and *BrNHX1.3*; **Table 1**) suggesting substantial gene loss of *NHX* genes in *B. rapa* following the most recent WGT. To further investigate this gene loss, we examined the chromosomal location of the *B. rapa NHX* genes, and compared these to the described fractionated genomes (Wang et al., 2011). Six of the 10 *B. rapa* NHXs were retained on the LF sub-genome, one on the MF1 and three on the MF2 (**Table 1**). The proteins encoded by the two class II-IC *NHX* genes identified in *B. rapa* are clearly differentiated from other members of the *A. thaliana* NHX family (**Figure 2A**) and are likely to be the true orthologs of *AtNHX5* and *AtNHX6*. Interestingly, both of the identified class II-IC *NHX* genes are more closely related to *AtNHX6* than to *AtNHX5* (**Figure 2A**) indicating that there may not be a *B. rapa* ortholog of *AtNHX5*.

To further investigate the apparent absence of a *AtNHX5* ortholog in *B. rapa*, we decided to try to identify the intracellular *NHX* genes in the closely related species *T. parvula.*

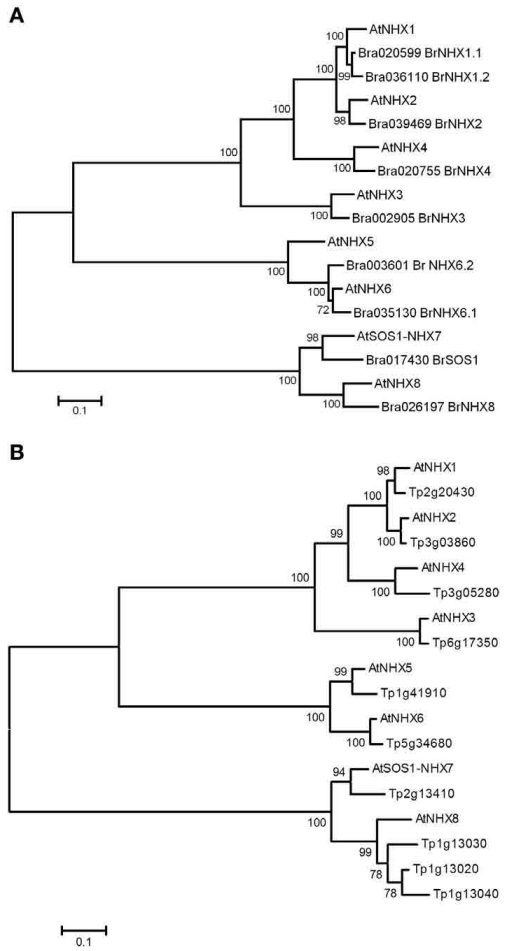


FIGURE 2 | Phylogenetic analysis of *B. rapa* and *T. parvula* NHX proteins. Amino acid sequences of NHX proteins from *B. rapa* **(A)**, *T. parvula* **(B)** and *A. thaliana* were aligned using Clustal W, and a Neighbor-joining tree generated. Numbers at nodes correspond to bootstrap support percentages following 1000 replications, and the scale is in amino acid changes per site.

T. parvula diverged from the *Brassica* species after the divergence from *A. thaliana* but before the Br—α WGT event (**Figure 1**). Six intra-cellular *NHX* genes were identified and the phylogentic analysis showed that there was one unique ortholog of each of the *A. thaliana NHX1-6* genes (**Figure 2B**), indicating that there was an *AtNHX5* ortholog present in the ancient *Brassica* ancestor species before the Br—α WGT.

MACRO- AND MICRO-SYNTENY OF *ATNHX5 AND ATNHX6*

To be confident that the *BrNHX6.1* and *BrNHX6.2* genes identified are both orthologs of *AtNHX6* and not *AtNHX5*, we analysed the macrosynentic relationship between these genes based on the 24 conserved syntenic blocks between *A. thaliana* and *B. rapa* (Schranz et al., 2006). Analysis of the conserved co-linearity in *A. thaliana* and *B. rapa* shows that *BrNHX6.1*, *BrNHX6.2*, and *AtNHX6* are all located in the retained conserved genomic block E and are likely to be true orthologs, whereas *AtNHX5* is retained within the conserved genomic Block C on the *A. thaliana* genome (**Table 1**), and appears to have no clear ortholog in *B. rapa*. We also analysed the gene models for *AtNHX5, AtNHX6, BrNHX6.1*, and *BrNHX6.2* (**Figure 3**). This showed that *AtNHX6, BrNHX6.1*, and *BrNHX6.2* all have a common 22 exon/21 intron gene structure, whereas *AtNHX5* has only 21 exons (**Figure 3**) supporting the notion that the *BrNHX6.1* and *BrNHX6.2* genes are both orthologous to *AtNHX6* and not *AtNHX5*.

Although it is clear that there was an AtNHX5 ortholog present prior to the Br—α WGT event, no *AtNHX5* homolog was identified in the *B. rapa* genome. To determine the cause of this discrepancy we decided to carefully investigate the microsynteny around *AtNHX5* (**Figure 4A**). The microsynteny in the *A. thaliana* genome, including fourteen genes downstream and twelve genes upstream of *AtNHX5*, was compared to the corresponding co-linear sub genome regions in the *B. rapa* genome (**Figure 4A**). Surprisingly, the corresponding co-linear regions of the three *B. rapa* sub genomes contain only five (LF), three (MF1) and nine (MF2) genes respectively, and only four (LF), three (MF1) and five (MF2) of these genes were homologous to *A. thaliana* genes (**Figure 4A**). The colinearity between *A. thaliana* and the *B. rapa* sub genomes is relatively well conserved from the At1G54220 gene to At1G54290 (**Figure 4A**). However, the next 14 genes upstream of At1G54290 in the *A. thaliana* genome including *AtNHX5* are completely absent from all three co-linear *B. rapa* sub genomes (**Figure 4A**). Co-linear genes are again found between the *A. thaliana* genome and the *B. rapa* LF sub genome from At1G54490 and the MF2 sub genomes from At1G5440, however, no homologous genes can be detected in the MF1 sub genome (**Figure 4A**). Interestingly, upstream of At1G54440 the co-linear region on the LF sub genome is found on chromosome BrA09, whereas the co-linear region of the LF sub genome downstream of At1G54290 was found on chromosome BrA06 (**Figure 4A**). Similarly, the region of co-linearity upstream of At1G54440 on the MF2 sub genome is found on chromosome BrA01 and the co-linear region of the MF2 sub genome downstream of At1G54290 was found on chromosome BrA08 (**Figure 4A**). This indicates that the syntenic regions around *AtNHX5* in the LF and MF2 sub genomes have probably undergone an inter-chromosomal translocation event. It also appears that there has been a large fragment deletion event in the MF1 sub genome in the syntenic region around *AtNHX5*. These chromosomal re-arrangments and deletions would account for the absence of any *AtNHX5* orthologs in *B. rapa* and may be the result of the genome reduction processes post the Br—α WGT event. This evidence, combined with the inability to identify any *AtNHX5* orthologs when searching the *B. rapa* genome, strongly suggest that there are no *AtNHX5* orthologs present in *B. rapa*.

The microsynteny in the *A. thaliana* genome, including eight genes downstream and eight genes upstream of *AtNHX6*, was compared to the corresponding co-linear sub genome regions in the *B. rapa* genome (**Figure 4B**). The corresponding three co-linear *B. rapa* sub genomes contain 13 (LF), 16 (MF1), and 10 (MF2) genes, respectively, with 11 (LF), nine (MF1), and 10 (MF2) homologous genes to the co-linear region in the *A. thaliana* genome (**Figure 4B**). The *BrNHX6.1* gene is located on the LF sub genome and the *BrNHX6.2* gene is located of

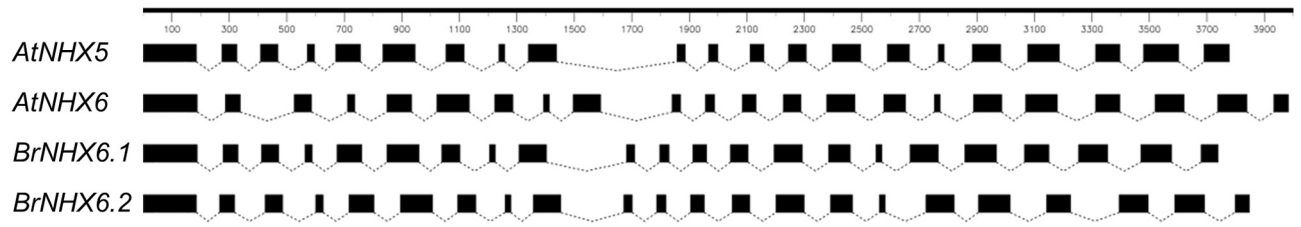


FIGURE 3 | Predicted gene models for AtNHX5, AtNHX6, BrNHX6.1 and BrNHX6.2. Predicted gene models for AtNHX5, AtNHX6, BrNHX6.1, and BrNHX6.2. Exons are shown as black boxes and introns are shown as dotted lines. Gene models were generated using FancyGene 1.4 (http://www.bio.ieo.eu/fancygene).

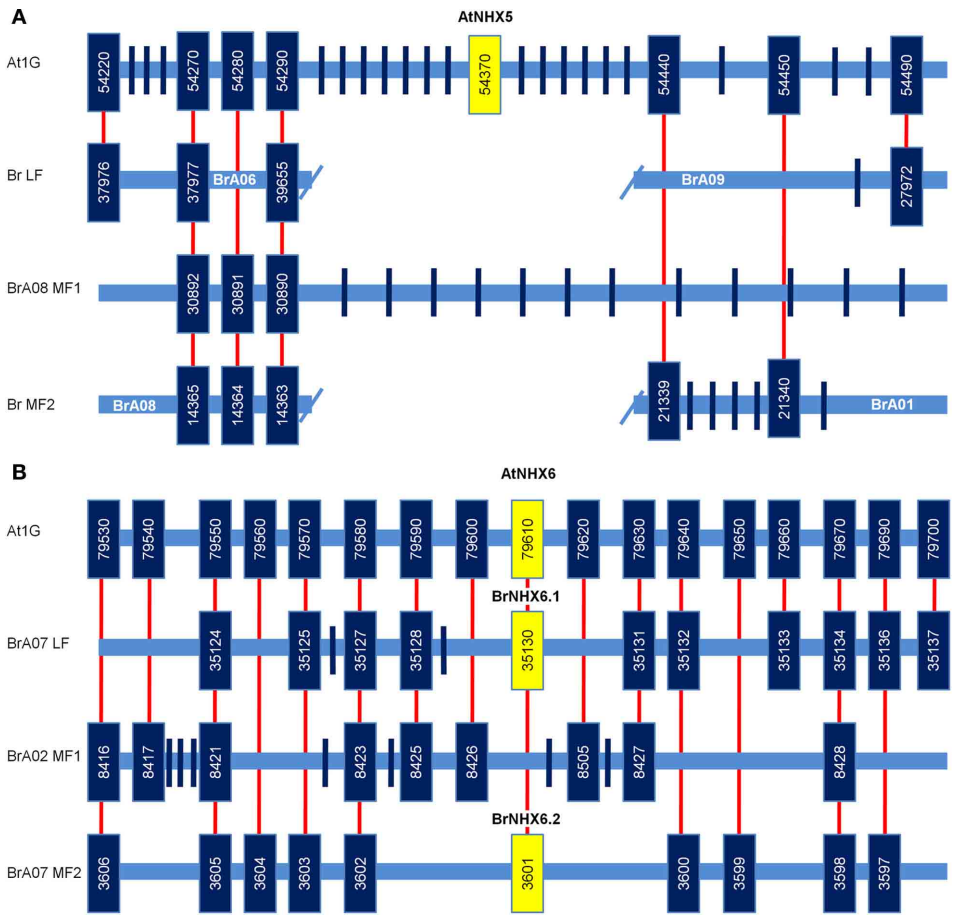


FIGURE 4 | Microsynteny of AtNHX5 and AtNHX6 and conserved co-linearity in *B. rapa*. Comparison of co-linearity between AtNHX5 **(A)** and AtNHX6 **(B)** (highlighted in yellow) in the Arabidopsis genome and the least fractionated (LF), medium fractionated (MF1) and most fractionated (MF2) sub-genomes in *B. rapa*. The boxes and bars denote annotated genes (boxes contain gene accession number) and the red lines connect syntenic homologous genes (Note corresponding genomic regions are not drawn to scale).

the MF2 sub genome, but there is no *AtNHX6* ortholog present on the MF1 sub genome (**Figure 4B**). Importantly, all of the 17 *A. thaliana* genes in the region surrounding *AtNHX6* have at least one homolog in one of the corresponding *B. rapa* sub genomes and the order of the genes in *A. thaliana* is maintained within *B. rapa* (**Figure 4B**). The high degree of co-linearity between the region in *A. thaliana* genome containing *AtNHX6* and the corresponding regions in the *B. rapa* genome strongly support the notion that there are only two *AtNHX6* orthologs in *B. rapa* with the third ortholog likely to have been lost from the MF1 sub genome since the last whole genome triplication event.

PROMOTER ANALYSIS OF BrNHX6.1, BrNHX6.2 AND AtNHX6

Although the *B. rapa* genome does not appear to contain a homolog to *AtNHX5*, there is a high level of genetic conservation between the *A. thaliana* and *B. rapa NHX6.1* and *NHX6.2*

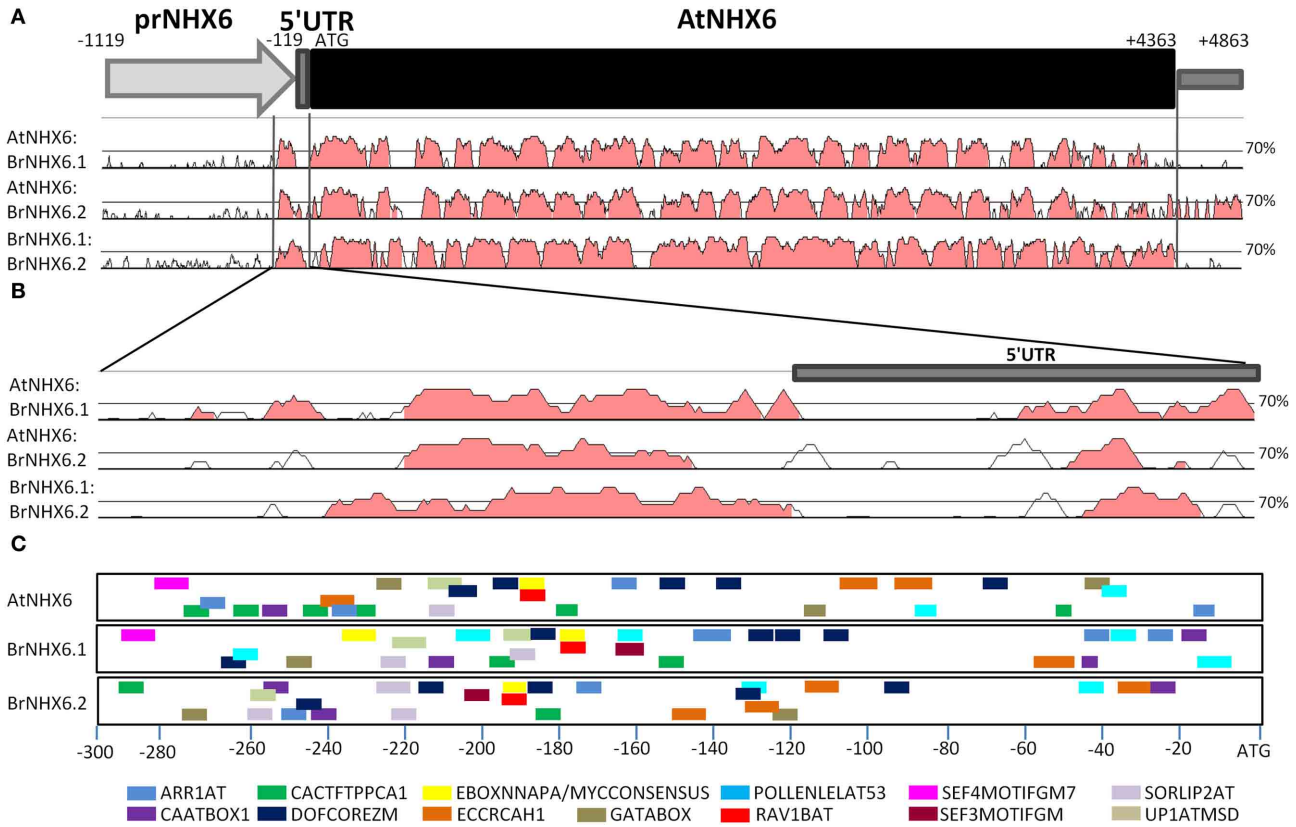


FIGURE 5 | Conservation of promoter regions of NHX6 homologues in *B. rapa*. Graphic output for comparisons of *Arabidopsis* and *B. rapa* IC II NHX genome region **(A)** and putative promoter region **(B)** using VISTA, with parameters of at least 70% identity within a window of 25 b. Regions with >70% identity shared by the three orthologs are shaded in pink. **(C)** Positions of known plant *cis*-acting regulatory elements which are common to all three genes are shown as colored bars.

orthologs. To explore the potential for differential functionality between these two genes, we investigated the level of sequence similarity between their respective promoter regions and the conservation of specific *cis*-acting regulatory motifs.

To identify conserved regions within the promoters of *B. rapa NHX6* orthologs, we compared nucleotide sequences of *AtNHX6*, *BrNHX6.1,* and *BrNHX6.2* (**Figure 5A**). The gene sequences, including 1119 bp upstream and 500 bp downstream, were subjected to sliding-widow analysis of homology. It has previously been shown that requiring 70% identity within a 25-bp window returns the greatest number of regulatory elements with acceptable specificity (Guo and Moose, 2003). Using these parameters, sliding window analysis revealed a highly conserved region of approximately 300 bp immediately upstream of the start codon, including the 5′UTR (**Figure 5B**).

Evaluation of this highly conserved promoter region (−300 bp) using the PLACE database SIGNALSCAN tool identified 10 *cis*-acting regulatory elements (CREs) which were common to all three genes: ARR1AT (5′-NGATT-3′), CAATBOX1 (5′-CAAT-3′), CACTFTPPCA1 (5′-YACT-3′), DOFCOREZM (5′-AAAG-3′), EBOXNNAPA (5′-CANNAG-3′), EECCRCAH1 (5′-GANTTNC-3′), GATABOX (5′-GATA-3′), MYCCONSENSUS (5′-CANNAG-3′), POLLENLELAT53 (5′-AGAAA-3′), RAV1BAT (5′-CACCTG-3′), SORLIP2AT (5′-GCCAC-3′), and UP1ATMSD (5′-GGCCCAWWW-3′). Additionally, the functionally similar elements SEF4MOTIFGM7S (5′-RTTTTTR-3′) and SEF3MOTIFGM (5′-AACCCA-3′) were conserved between *AtNHX6*:*BrNHX6.1* and *BrNHX6.1*:*BrNHX6.2*, respectively. The position of these motifs in the promoter sequence relative to the start codon is depicted in **Figure 5C**. Interestingly, the position of the DOFCOREZM (5′-AAAG-3′) and MYCCONSENSUS/EBOXNNAPA (5′-CANNAG-3′)/RAV1BAT (5′-CACCAG-3′) elements are conserved across the three promoter regions.

IDENTIFICATION *B. napus* NHX6 HOMOLOGS

B. napus is an agriculturally important member of the Brassicaceae family. Due to the significance of salinity in *B. napus* cultivation, we sought to identify the complement of expressed *B. napus NHX6* homologs and assess their role in salt tolerance. To identify *NHX6* homologs in *B. napus* without the advantage of a complete *B. napus* genome sequence, we designed primers in absolutely conserved regions present in all available nucleotide sequences from *B. napus*, *B. rapa*, and *B. oleracea*. These primers were first tested on *B. rapa* genomic DNA to ensure they could identify both *BrNHX6.1* and *BrNHX6.2*. Using these

"universal" *Brassica* NHX6 primers, four unique sequences were identified in *B. rapa*. A phylogenetic analysis of these sequences revealed two distinct groups including either the *BrNHX6.1* or the *BrNHX6.2* sequences from the *B. rapa* genome (**Figure 6A**). The presence of two unique sequences grouping with each *BrNHX6* sequence is most likely the result of heterozygous *BrNHX6* alleles present in the *B. rapa* genomic DNA tested. Importantly, however, the primers used were able to amplify both the *BrNHX6* homologs.

The complement of expressed *AtNHX6* homologs in *B. napus* was then investigated using the same approach, but using mRNA isolated from *B. napus* seedlings. Sequencing the cloned amplicons from *B. napus* cDNA identified only two unique sequences. A phylogenetic analysis was performed including *BrNHX6.1* and *BrNHX6.2* sequences from the *B. rapa* genome, the *B. napus* (JX082291) and *B. oleracea* (JX082292) ORF sequences and a *B. oleracea* (DK554651) EST sequence. The phylogenetic analysis showed two major groupings, each containing either the *BrNHX6.1* or the *BrNHX6.2* sequence (**Figure 6B**). Interestingly, only one sequence grouped with *BrNHX6.2* while four unique sequences grouped with *BrNHX6.1* (**Figure 6B**). The sequences that grouped with *BrNHX6.1* were divided into two sub-groups—one group has a *B. napus* sequence more closely related to the *BrNHX6.1* sequence, while the other group includes only *B. oleracea* and *B. napus* sequences, possibly indicating both an A and C genome version of *BrNHX6.1* (**Figure 6B**). From this analysis, the *B. napus* sequences that are clearly closely related to *BrNHX6.1* and *BrNHX6.2* will be hereafter designated as *BnNHX6.1* and *BnNHX6.2*, respectively.

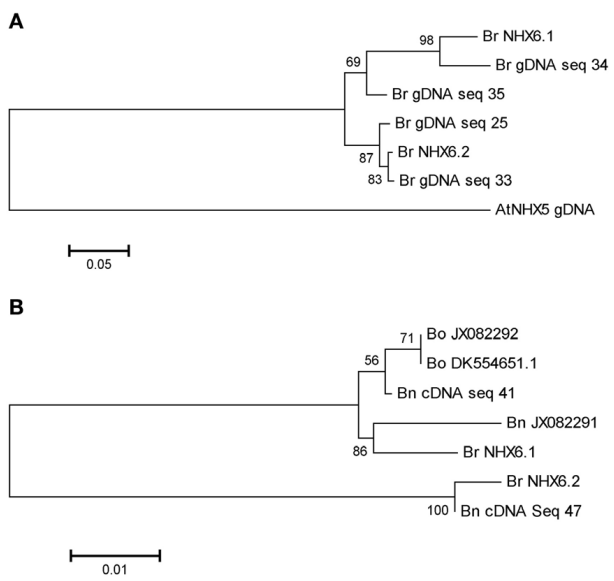


FIGURE 6 | Phylogenetic analysis of *B. rapa, B. oleracea,* and *B. napus* NHX6 partial sequences. *B. rapa* gDNA **(A)** and *B. rapa, B. oleracea* and *B. napus* mRNA **(B)** sequences were aligned using Clustal W, and a Neighbor-joining tree generated. Numbers at nodes correspond to bootstrap support percentages following 1000 replications.

EXPRESSION ANALYSIS OF *BnNHX6.1* AND *BnNHX6.2*

Having identified two *NHX6* genes in *B. napus*, we examined the relative expression of *BnNHX6.1* and *BnNHX6.2* in the presence and absence of salt stress. Firstly, to optimise the concentration of NaCl required to induce salt stress, we compared the germination percentage and fresh weight of *B. napus* cv. Westar seedlings grown in the presence and the absence of NaCl. The NaCl treatment had a marked effect on the germination and growth of *B. napus* seedlings (**Figure 7A**). Germination in the presence of NaCl was severely inhibited with only 41% of seeds germinating compared with 92% of seeds sown on plates lacking additional NaCl (**Figure 7B**). The mean fresh weight per seedling grown in the presence of NaCl was also greatly reduced compared to those grown in the absence of NaCl (**Figure 7C**). These results demonstrate that the NaCl treatment was sufficient to induce a severe stress impacting on the germination and growth of *B. napus* seedlings. We then examined the differential expression of both the *BnNHX6.1* and the *BnNHX6.2* genes in the presence and absence of the NaCl stress. The *BnNHX6.2* gene showed no significant difference in relative transcript abundance in response to the NaCl stress, whereas there was a significant increase in relative transcript abundance of the *BnNHX6.1* gene in response to the NaCl stress (**Figure 7D**). Interestingly, the relative expression of *BnNHX6.1* was extremely low compared to *BnNHX6.2* both in the presence and absence of NaCl (**Figure 7D**) suggesting that there may be some differential regulation of these two genes.

DISCUSSION

ORTHOLOG IDENTIFICATION IN *B. napus*

In this paper we have attempted to identify the *B. napus* orthologs of two *A. thaliana* proteins, AtNHX5 and AtNHX6. Although *A. thaliana* and *B. napus* are relatively closely related, precise ortholog identification is complicated by the recent WGT event in the *Brassica* species and the amphidiploid nature of the *B. napus* (AACC $2n = 38$) genome. Consequently, there are potentially six *B. napus* homologs of *AtNHX5* and *AtNHX6* comprising three triplicated paralogs from both the A and C parental genomes.

Identification of *A. thaliana* homologs in the *Brassica* species has traditionally relied upon comparison to incomplete genomic sequences, generally in the form of bacterial artificial chromosome (BAC) clones. For example identification of *FLOWERING LOCUS C* (*FLC)* homologs in *B. rapa* entailed probing *B. rapa* BAC libraries to identify BAC clones containing co-linear regions (Yang et al., 1999). While successful in this case, this methodology was reliant on having all the appropriate co-linear region of the *B. rapa* genome present in the BAC library. The recent release of the *B. rapa* genome provides a complete *Brassica* genomic resource allowing *A. thaliana* gene orthologs to be identified with a high level of confidence.

Our search of the *B. rapa* genome identified only two *AtNHX6* orthologs. Analysis of the *B. rapa* genome has shown that there has been significant loss of gene paralogs in the three sub genomes since the WGT event (Wang et al., 2011), which is likely to be the reason that there is no *AtNHX6* ortholog on the MF1 sub genome. Interestingly, we could not find any evidence of an *AtNHX5* homolog in the *B. rapa* genome. While this came as a surprise, recent evidence shows that *AtNHX5* and *AtNHX6* act redundantly

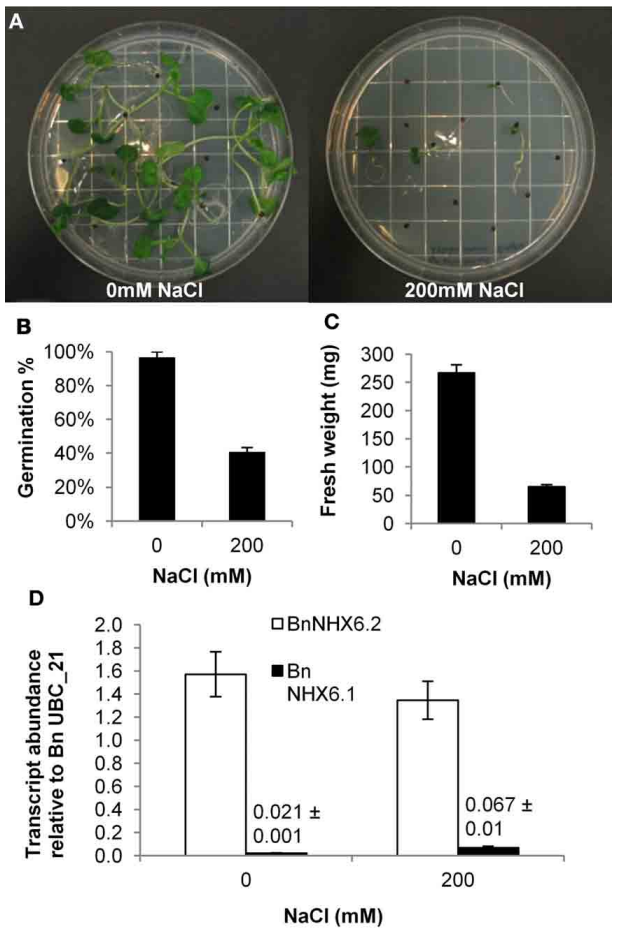


FIGURE 7 | Differential expression of BnNHX6.1 and BnNHX6.2 in *B. napus* seedlings in response to NaCl. *B. napus* (cv Westar) seeds were sown on ½ strength MS media with or without 200 mM NaCl **(A)** and the germination percentage **(B)** and fresh weight per seedling **(C)** were compared between the two treatments. RNA was extracted from seedlings and the transcript abundance of BnNHX6.1 and BnNHX6.2 was assayed relative to the BnUBC_21 reference gene **(D)**. Data are the means of three biological replicates (±s.e.).

in *A. thaliana* (Bassil et al., 2011), and therefore the the *Brassica* species may not require a functional *AtNHX5* ortholog. Our analysis of the microsynteny around *AtNHX5* and its co-linearity with the three corresponding sub genomes in *B. rapa* indicated that a large block of *A. thaliana* genes including *AtNHX5* were missing from all three co-linear *B. rapa* sub genomes as a result of inter-chromosomal translocations in the LF and MF2 sub genomes and a fragment deletion in the MF1 sub-genome. The identification of an *AtNHX5* ortholog in the closely related *T. parvula* species indicates that there was an *AtNHX5* ortholog present in the ancient *Brassica* ancestral species prior to the Br—α WGT event. The loss of *AtNHX5* in the *Brassica* species could only have occurred either due to a localized deletion event occurring in the ancestral *Brassica* genome after the *T. parvula-Brassica* divergence, but before the *Brassica* WGT event or due to the loss of all three *AtNHX5* genes following the WGT event. It seems more likely that the *AtNHX5* orthologs were lost after the Br—α WGT event due to the differential disruption in the syntentic region around *AtNHX5* in the three *B. rapa* sub-genomes. If *AtNHX5* had been lost prior to the Br—α WGT event it would be reasonable to expect that the syntenic region in all three *B. rapa* sub-genomes would display a common chromosomal disruption that was then triplicated as part of the Br—α WGT event.

Sequencing of orthologous BAC clones of the *B. rapa* A genome, the *B. oleracea* C genome and the A and C genome components of *B. napus* have shown essentially complete co-linearity between the protein coding genes in the A and C genomes and sequence identity of 94–97% (Cheung et al., 2009). It could, therefore, be anticipated that the *B. oleracea* C genome is also likely to only have two *AtNHX6* orthologs. Our sequence analysis in *B. oleracea* and *B. napus* could only identify two distinct *AtNHX6* homologs that were also homologous to the *BrNHX6.1* and *BrNHX6.2* genes. It is important to note however, that all of the *B. oleracea* and *B. napus* sequences analysed in this phylogeny were derived from mRNA; it is therefore possible that additional *B. oleracea NHX6* homologs are present in the genome but have not been detected in the limited mRNA sequencing described here. The complete complement of *B. oleracea NHX6* homologs could only be resolved with a completed *B. oleracea* genome. The phylogenetic analysis of homologous *Brassica NHX6.1* sequences shows two distinct groupings, one including *B. rapa NHX6.1* and a *B. napus* sequence, and the other containing only *B. oleracea* and *B. napus* sequences. This provides evidence of possible A and C genome versions of *Brassica NHX6.1*. Unfortunately, we do not have a corresponding *B. oleracea NHX6.2* sequence available, making it difficult to clearly identify the A and C genome versions of *NHX6.2*. It is therefore most likely that there are two *AtNHX6* homologs in both *B. rapa* and *B. oleracea* and, as a consequence, four *AtNHX6* homologs in *B. napus*.

EXPRESSION PATTERNS OF *BnNHX6.1* AND *BnNHX6.2*

The analysis of the *B. rapa* genome showed that of the eight Arabidopsis *NHX* genes only *AtNHX1* and *AtNHX6* have retained multiple paralogs since the most recent triplication event. Gene duplications arising from polyploidy events are thought to be a major source of novel variation and gene function (Moore and Purugganan, 2003) and duplicated genes may be lost, develop new functions or share the functions of the original gene through differential expression or regulation (Blanc and Wolfe, 2004). It may be the case that the two orthologs of *AtNHX6* may have different functions or expression patterns. As the coding regions of all the identified *Brassica* NHX6 genes are highly homologous with *AtNHX6* and each other, it is likely that their functions (Na^+/K^+–H^+ antiporting) remain identical. The analysis of the *BrNHX6.1* and *BrNHX6.2* promoter regions highlighted a significant region of similarity 300 bp upstream of the ATG. The conserved MYCCONSENSUS element is often associated with genes responsive to abiotic stress, such as the stress hormone abscisic acid (ABA), dehydration-responsive RD22, and cold responsive genes CBF, DREB1, and ICE1 (Chinnusamy et al., 2004). This is highly interesting given the potential role for *NHX* proteins in improving salt tolerance. Additionally, the equivalent EBOXNNAPA motif was found

to be an essential promoter element for high expression of the napA storage protein in *B. napus* seeds (Stalberg et al., 1996). The conservation of these regulatory elements suggests conservation of functionality between AtNHX6 and its *B. rapa* homologs.

The expression of the *B. napus NHX6.2* gene is significantly higher than the *B. napus NHX6.1* gene in two week old seedlings, indicating that in this tissue type at least there does appear to be some differential expression. While there is no significant difference in expression of *BnNHX6.2* in response to NaCl, the expression of *BnNHX6.1* increased approximately three-fold in response to NaCl again highlighting differences in expression. It should be noted that the relative expression of *BnNHX6.1* was 20 times lower than the expression of *BnNHX6.2* in the NaCl treatment, and is suggestive of a more prominent role for *BnNHX6.2*. This is in contrast to the situation in *A. thaliana*, where the largest difference in expression observed is a 3.5 increase of AtNHX5 transcript in leaf tissue (Bassil et al., 2011). The two Brassica NHX6 isoforms do not appear to have different functions, but the expression of BnNHX6.1 is significantly lower than that of *BnNHX6.2*. This may indicate that the *BnNHX6.1* gene is simply sitting in the genome waiting to be lost, like most of the other triplicated paralogs of the *BrNHX* gene family. It would be of great interest to examine the expression patterns of *BnNHX6.1* and *BnNHX6.2* in a variety of tissues types in the presence and absence of NaCl to better elucidate differential expression patterns.

In this study we have demonstrated the use of the *B. rapa* genome in the successful identification of two *B. rapa* NHX6 orthologs, and used this information to identify two unique *B. napus* NHX6 orthologs. Our analysis also strongly suggests that there are four *B. napus* orthologs, comprising an A and C genome version of *BrNHX6.1* and *BrNHX6.2*. This study also indicates that there may be some differential expression of the *BnNHX6.1* and *BnNHX6.2*.

ACKNOWLEDGMENTS

This work was supported by the Grains Research Development Corporation through a Undergraduate Honours Scholarship (UHS119) and a Graduate Research Scholarship (GRS161) to Brett A. Ford and GRS179 to Joanne R. Ernest.

REFERENCES

Apse, M. P., Aharon, G. S., Snedden, W. A., and Blumwald, E. (1999). Salt tolerance conferred by overexpression of a vacuolar Na+/H+ antiport in Arabidopsis. *Science* 285, 1256–1258.

Bassil, E., Ohto, M. A., Esumi, T., Tajima, H., Zhu, Z., Cagnac, O., Belmonte, M., Peleg, Z., Yamaguchi, T., and Blumwald, E. (2011). The *Arabidopsis* intracellular Na+/H+ antiporters *NHX5* and *NHX6* are endosome associated and necessary for plant growth and development. *Plant Cell* 23, 224–239.

Bell, C. D., Soltis, D. E., and Soltis, P. S. (2010). The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* 97, 1296–1303.

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16, 1679–1691.

Bowers, K., Levi, B. P., Patel, F. I., and Stevens, T. H. (2000). The sodium/proton exchanger Nhx1p is required for endosomal protein trafficking in the yeast *Saccharomyces cerevisiae*. *Mol. Biol. Cell* 11, 4277–4294.

Brett, C. L., Donowitz, M., and Rao, R. (2005a). Evolutionary origins of eukaryotic sodium/proton exchangers. *Am. J. Physiol. Cell Physiol* 288, C223–C239.

Brett, C. L., Tukaye, D. N., Mukherjee, S., and Rao, R. J. (2005b). The yeast endosomal Na+(K+)/H+ exchanger Nhx1 regulates cellular pH to control vesicle trafficking. *Mol. Biol. Cell* 16, 1396–1405.

Chanroj, S., Wang, G., Venema, K., Zhang, M. W., Delwiche, C. F., and Sze, H. (2012). Conserved and diversified gene families of monovalent cation/$H^{(+)}$ antiporters from algae to flowering plants. *Front. Plant Science* 3:25. doi: 10.3389/fpls.2012.00025

Chen, X., Truksa, M., Shah, S., and Weselake, R. J. (2010). A survey of quantitative real-time polymerase chain reaction internal reference genes for expression studies in *Brassica napus*. *Anal. Biochem.* 405, 138–140.

Cheung, F., Trick, M., Drou, N., Lim, Y. P., Park, J. Y., Kwon, S. J., Kim, J. A., Scott, R., Pires, J. C., Paterson, A. H., Town, C., and Bancroft, I. (2009). Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. *Plant Cell* 21, 1912–1928.

Chinnusamy, V., Schumaker, K., and Zhu, J. K. (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *J. Exp. Bot.* 55, 225–236.

Couvreur, T. L., Franzke, A., Al-Shehbaz, I. A., Bakker, F. T., Koch, M. A., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* 27, 55–71.

Dassanayake, M., Oh, D. H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D. J., Bressan, R. A., Zhu, J. K., Bohnert, H. J., and Cheeseman, J. M. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* 43, 913–918.

Food, and Agriculture Organization of the United Nations. (2009). "Increasing crop production sustainably. The perspective biological processes," ed *Department of Economic and Social Development* (Rome: FAO).

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., and Mummenhoff, K. (2011). Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* 16, 108–116.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279.

Fukuda, A., Nakamura, A., Hara, N., Toki, S., and Tanaka, Y. (2011). Molecular and functional analyses of rice NHX-type Na^+/H^+ antiporter genes. *Planta* 233, 175–188.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186.

Guo, H., and Moose, S. P. (2003). Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* 15, 1143–1158.

Henry, Y., Bedhomme, M., and Blanc, G. (2006). History, protohistory and prehistory of the *Arabidopsis thaliana* chromosome complement. *Trends Plant Sci.* 11, 267–273.

Herrmann, B. G., and Frischauf, A. M. (1987). Isolation of genomic DNA. *Methods Enzymol.* 152, 180–183.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., Van De Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y. L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* 17, 1483–1498.

Li, M., Li, Y., Li, H., and Wu, G. (2011). Overexpression of *AtNHX5* improves tolerance to both salt and drought stress in *Broussonetia*

papyrifera (L.) Vent. *Tree Physiol.* 31, 349–357.

Lysak, M. A., Koch, M. A., Pecinka, A., and Schubert, I. (2005). Chromosome triplication found across the tribe Brassiceae. *Genome Res.* 15, 516–525.

Moore, R. C., and Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15682–15687.

Mun, J. H., Kwon, S. J., Yang, T. J., Seol, Y. J., Jin, M., Kim, J. A., Lim, M. H., Kim, J. S., Baek, S., Choi, B. S., Yu, H. J., Kim, D. S., Kim, N., Lim, K. B., Lee, S. I., Hahn, J. H., Lim, Y. P., Bancroft, I., and Park, B. S. (2009). Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biol.* 10, R111.

Munns, R., and Tester, M. (2008). Mechanisms of salinity tolerance. *Annu. Rev. Plant Biol.* 59, 651–681.

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bio-assays with tobacco tissue cultures. *Physiol. Plant* 15, 473–497.

Parkin, I. A., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C., and Lydiate, D. J. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* 171, 765–781.

Pfaffl, M. (2004). "Quantification strategies in real-time PCR," in *A-Z of quantitative PCR*, ed S. Bustin (La Jolla, CA: International University Line), 87–112.

Rambaldi, D., and Ciccarelli, F. D. (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* 25, 2281–2282.

Rodriguez-Rosales, M. P., Jiang, X., Galvez, F. J., Aranda, M. N., Cubero, B., and Venema, K. (2008). Overexpression of the tomato K^+/H^+ antiporter LeNHX2 confers salt tolerance by improving potassium compartmentalization. *New Phytol.* 179, 366–377.

Schranz, M. E., Lysak, M. A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* 11, 535–542.

Shi, H., Lee, B. H., Wu, S. J., and Zhu, J. K. (2003). Overexpression of a plasma membrane Na^+/H^+ antiporter gene improves salt tolerance in *Arabidopsis thaliana*. *Nat. Biotechnol.* 21, 81–85.

Shi, H., Quintero, F. J., Pardo, J. M., and Zhu, J. K. (2002). The putative plasma membrane Na(+)/H(+) antiporter SOS1 controls long-distance Na(+) transport in plants. *Plant Cell* 14, 465–477.

Shi, H., and Zhu, J. K. (2002). *SOS4*, a pyridoxal kinase gene, is required for root hair development in Arabidopsis. *Plant Physiol.* 129, 585–593.

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., Sankoff, D., Depamphilis, C. W., Wall, P. K., and Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348.

Stalberg, K., Ellerstom, M., Ezcurra, I., Ablov, S., and Rask, L. (1996). Disruption of an overlapping E-box/ABRE motif abolished high transcription of the *napA* storage-protein promoter in transgenic *Brassica napus* seeds. *Planta* 199, 515–519.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5, molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739.

Tester, M., and Davenport, R. (2003). Na^+ tolerance and Na^+ transport in higher plants. *Ann. Bot. (Lond.)* 91, 503–527.

United States Department of Agriculture. (2009). "Oilseeds: world market and trade," ed *Foreign Agricultural Service Circular Series 11-09 - November*, (Washington, DC: USDA).

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J. H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Yu, J., Meng, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Z., Li, Z., Xiong, Z., and Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Yang, Y. W., Lai, K. N., Tai, P. Y., and Li, W. H. (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* 48, 597–604.

Yokoi, S., Quintero, F. J., Cubero, B., Ruiz, M. T., Bressan, R. A., Hasegawa, P. M., and Pardo, J. M. (2002). Differential expression and function of *Arabidopsis thaliana* NHX Na^+/H^+ antiporters in the salt stress response. *Plant J.* 30, 529–539.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2012; accepted: 16 August 2012; published online: 11 September 2012.

*Citation: Ford BA, Ernest JR and Gendall AR (2012) Identification and characterization of orthologs of AtNHX5 and AtNHX6 in Brassica napus. Front. Plant Sci. **3**:208. doi: 10.3389/fpls.2012.00208*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 June 2012
doi: 10.3389/fpls.2012.00129

A candidate gene-based association study of tocopherol content and composition in rapeseed (*Brassica napus*)

Steffi Fritsche[1], Xingxing Wang[2], Jinquan Li[3], Benjamin Stich[3], Friedrich J. Kopisch-Obuch[1], Jessica Endrigkeit[1], Gunhild Leckband[4], Felix Dreyer[4], Wolfgang Friedt[5], Jinling Meng[2] and Christian Jung[1]*

[1] *Faculty of Agricultural and Nutritional Sciences, Plant Breeding Institute, Christian-Albrechts-University, Kiel, Germany*
[2] *National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China*
[3] *Quantitative Crop Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany*
[4] *Norddeutsche Pflanzenzucht Hans-Georg Lembke KG, Hohenlieth, Germany*
[5] *Faculty of Agricultural Sciences, Nutritional Sciences and Environmental Management, Institute of Agronomy and Plant Breeding I, Justus-Liebig-University, Giessen, Germany*

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Antoni Rafalski, Pioneer Hi-Bred International, A DuPont Business, USA
Jianbing Yan, Huazhong Agricultural University, China

***Correspondence:**
Christian Jung, Plant Breeding Institute, Christian-Albrechts-University, Olshausenstrasse 40, 24118 Kiel, Germany.
e-mail: c.jung@plantbreeding.uni-kiel.de

Rapeseed (*Brassica napus* L.) is the most important oil crop of temperate climates. Rapeseed oil contains tocopherols, also known as vitamin E, which is an indispensable nutrient for humans and animals due to its antioxidant and radical scavenging abilities. Moreover, tocopherols are also important for the oxidative stability of vegetable oils. Therefore, seed oil with increased tocopherol content or altered tocopherol composition is a target for breeding. We investigated the role of nucleotide variations within candidate genes from the tocopherol biosynthesis pathway. Field trials were carried out with 229 accessions from a worldwide *B. napus* collection which was divided into two panels of 96 and 133 accessions. Seed tocopherol content and composition were measured by HPLC. High heritabilities were found for both traits, ranging from 0.62 to 0.94. We identified polymorphisms by sequencing selected regions of the tocopherol genes from the 96 accession panel. Subsequently, we determined the population structure (Q) and relative kinship (K) as detected by genotyping with genome-wide distributed SSR markers. Association studies were performed using two models, the structure-based GLM + Q and the PK-mixed model. Between 26 and 12 polymorphisms within two genes (*BnaX.VTE3.a*, *BnaA.PDS1.c*) were significantly associated with tocopherol traits. The SNPs explained up to 16.93% of the genetic variance for tocopherol composition and up to 10.48% for total tocopherol content. Based on the sequence information we designed CAPS markers for genotyping the 133 accessions from the second panel. Significant associations with various tocopherol traits confirmed the results from the first experiment. We demonstrate that the polymorphisms within the tocopherol genes clearly impact tocopherol content and composition in *B. napus* seeds. We suggest that these nucleotide variations may be used as selectable markers for breeding rapeseed with enhanced tocopherol quality.

Keywords: *Brassica napus*, tocopherol (vitamin E), candidate genes, association study, SNP identification

INTRODUCTION

Together with soybean and oil palm, rapeseed (*Brassica napus*, genome AACC, $2n = 38$) belongs to the most important oil crops in the world. Because of its high-quality nutritional composition it is a common source of edible oil. Recently, the focus in rapeseed breeding has turned to improving and altering the content and composition of salutary oil constituents such as carotenoids (Shewmaker et al., 1999; Yu et al., 2008a; Wei et al., 2010), sterols (Amar et al., 2008; Hamama and Bhardwaj, 2011), oleic acid and linolenic acid contents (Rücker and Röbbelen, 1996; Schierholt and Becker, 2001; Zhang et al., 2004; Wittkop et al., 2009), and tocopherols (Marwede et al., 2004; Endrigkeit et al., 2009), the latter being also known as vitamin E.

Rapeseed oil contains high amounts of vitamin E, an essential component in human nutrition and health. A sufficient uptake of vitamin E can help to prevent neurological disorders, atherosclerosis, cataracts, and cancer (Witztum, 1993; Öhrvall et al., 1996; Schuelke et al., 1999; Sheehy et al., 2000; Schneider, 2005). Vitamin E is synthesized by plants and other photosynthetic organisms. The name is a generic term, which encompasses a group of fat-soluble compounds with antioxidant activity also called tocochromanols (Grusak and DellaPenna, 1999; DellaPenna and Pogson, 2006). The basic structure of the tocochromanols is characterized by a polar chromanol ring and a hydrophobic polyprenyl side chain, products of the shikimate and 1-deoxy-D-xylulose 5-phosphate (DOXP) pathways. Tocochromanols with a fully saturated tail are termed tocopherols, whereas those with an unsaturated tail are termed tocotrienols. The number of methyl groups on the chromanol ring define the four natural occurring tocopherol and tocotrienols forms (α, β, γ, and δ; Munné-Bosch and Alegre, 2002). With regard to the vitamin E activity, α-tocopherol has the highest activity and therefore, is the most

important vitamin E form for human nutrition (DellaPenna and Last, 2006).

The tocopherol biosynthetic pathway has been elucidated several years ago (Soll et al., 1980) and genes (*VTE*, loci 1–5; *PDS1*) encoding the respective enzymes of this pathway have been isolated and characterized in *Arabidopsis thaliana* and *Synechocystis* sp. PCC3903 (Norris et al., 1998; Porfirova et al., 2002; Bergmüller et al., 2003; Collakova and DellaPenna, 2003; Van Eenennaam et al., 2003; Valentin et al., 2006). In plants, tocopherols are mainly synthesized in plastids except for the first step which is catalyzed in the cytosol. The major tocopherol form in rapeseed oil is γ-tocopherol followed by α- and δ-tocopherol (Pongracz et al., 1995). The total tocopherol content (TTC) of 87 winter rapeseed genotypes ranged from 182 to 367 mg kg^{-1} and was significantly affected by genotype and environment (Goffman and Becker, 1999, 2002). Recently, genetic dissection of tocopherol biosynthesis in crop plants has been done for maize (Wong et al., 2003; Chander et al., 2008), soybean (Li et al., 2010), tomato (Almeida et al., 2011), and sunflower (Haddadi et al., 2011). In rapeseed, between five and seven QTL with additive and/or epistatic effects were mapped for α-, γ-, and TTC and composition (α/γ ratio) on six linkage groups in a segregating DH population (Marwede et al., 2005). The first gene from *B. napus* involved in tocopherol biosynthesis was cloned by using sequence information of *VTE4* orthologs of *A. thaliana* (Endrigkeit et al., 2009). In that study, the authors verified the function of the cloned *B. napus* gene by an *A. thaliana* transgenic approach leading to a shift in the tocopherol composition in seeds of *BnaA.VTE4.a1* overexpressing plants. Finally, the gene was mapped on *B. napus* chromosome A02 to the position of two QTLs controlling α-tocopherol content (ATC; Wang et al., in preparation). Linkage mapping is a well-established approach in rapeseed and has become the main tool for identifying genomic regions which contribute to the variation of quantitative traits (Snowdon et al., 2006; Long et al., 2007; Radoev et al., 2008; Zhao et al., 2008; Mei et al., 2009; Chen et al., 2010; Yin et al., 2010; Smooker et al., 2011; Zhang et al., 2011).

In recent years, association studies have become a valuable tool in plant genetics to study the correlation between genetic variants and trait differences based on linkage disequilibrium (LD; Thornsberry et al., 2001; Gupta et al., 2005; Zhu et al., 2008; Hall et al., 2010; Rafalski, 2010). Association studies benefit from the use of genetically diverse germplasm allowing the examination of the total allelic diversity derived from historical and evolutionary recombination events, whereas linkage mapping studies simply exploit the genetic diversity present between two parental genotypes. In rapeseed, marker-trait associations have been identified in several studies using a genome-wide approach for which a large number of markers had to be screened to reach the required density (Hasan et al., 2008; Honsdorf et al., 2010; Zou et al., 2010; Jestin et al., 2011; Rezaeizad et al., 2011). So far, only one candidate gene-based study has been carried out in *B. napus*, investigating the effect of *BnaA.FRI.a* haplotypes on flowering time (Wang et al., 2011).

Up to now, the tocopherol forms α, γ, and λ have been determined by high-performance liquid chromatography (HPLC) analysis, an invasive, laborious, and expensive method, which is not considered to be suitable as routine selection procedure. Therefore, a marker-assisted strategy would be a substantial step forward toward the selection of rapeseed varieties with enhanced tocopherol content and composition and therefore, facilitate the breeding process immensely.

In the present work, we conducted a candidate gene-based association approach to identify and assess the role of polymorphisms in *B. napus* tocopherol biosynthesis genes on tocopherol content and composition. We developed gene-specific primers and sequenced fragments of the candidate genes in a diverse set of rapeseed accessions. By identifying those allelic variations associated with either tocopherol content or composition, promising candidates for the development of molecular markers were detected, verified in a second rapeseed set and can now be used for the selection of rapeseed varieties with enhanced tocopherol qualities.

MATERIALS AND METHODS

PLANT MATERIAL AND FIELD EXPERIMENTS

We investigated 229 accessions from a worldwide *B. napus* collection which were divided into two panels of 96 and 133 accessions. The 96 accessions of panel 1 are part of a core collection, established during a European project on genetic diversity in *Brassica* crop species (http://documents.plant.wur.nl/cgn/pgr/brasedb/brasresgen.htm, **Table A1** in Appendix). In 2007/2008 panel 1 was grown over winter near the city of Giessen (University of Giessen) in central Germany and near Holtsee in Northern Germany (NPZ Lembke Company, Hohenlieth, Germany). The experiments were performed as a randomized complete block design (RCBD) with two replications and 1.75 m × 2.50 m plots with 100–120 seeds per plot or in case of limited seed availability 50–60 seeds per plot. Seeds were harvested from six to eight open pollinated plants per plot and used for tocopherol and seed quality measurements. Phenotypic data of panel 1 were obtained from 91 *B. napus* accessions grown at both locations. The accessions "Wolynski," "Ridana," and "Ramon," were grown only in Giessen whereas "Tapidor" and "Ningyou 7" were planted only in Holtsee (**Table A1** in Appendix).

The second panel 2 consisted of 133 of the 140 *B. napus* accessions which were assessed by Wang et al. (in preparation) and represented a worldwide collection of rapeseed accessions including spring, semi-winter type, and winter type rapeseed cultivars. This panel was grown in 2008/2009 and 2009/2010 at Jingzhou, China (Hubei Province) as a RCBD with three replications and 3 m^2 plots with 30 plants. For panel 2, phenotypic data were obtained for 133 *B. napus* accessions, grown in 2008/2009, and for 109 *B. napus* accessions grown in season 2009/2010.

TOCOPHEROL AND SEED QUALITY TRAITS MEASUREMENTS

Contents of α-, γ-, and λ-tocopherol in seeds were determined by HPLC (Schledz et al., 2001; Dähnhardt et al., 2002; Falk et al., 2003). For the extraction, 30–80 mg seeds were disaggregated in 1500 µl *n*-heptane. The solution was incubated at −20°C for 2 h and 20 µl was used for HPLC analysis. Separation of tocopherols was performed on a silica gel column (5 µM LiChrospher® Si 60, Merck) using a mobile phase consisting of

an *n*-heptane/isopropanol-mixture (99 + 1; v + v). Quantification of tocopherols was done by fluorescence detection (excitation at $\lambda = 290$ nm, emission at $\lambda = 328$ nm). To identify specific tocopherol forms, the retention times were compared with standards of Merck's tocopherol kit (Merck, Darmstadt, Germany) and for each tocopherol form a calibration was conducted by correlating the concentration of the single forms with the signal output. The concentrations of the analyzed samples in this study were within the linear range of the calibration. Only minor traces of β-tocopherol were obtained, which were not further analyzed during this study. TTC was calculated as the sum of ATC, γ-tocopherol content (GTC), and δ-tocopherol content (DTC) and the tocopherol composition was expressed as the ratio of α- and γ-tocopherol (AGR).

Glucosinolate (GSL), seed oil (SOC), and seed protein (SPC) contents of all 96 panel 1 accessions were measured by near-infrared spectroscopy (NIRS). From each field plot, two subsamples were analyzed. For NIRS measurements 3–5 mg of intact seeds were used. Individual seed spectra from 1100 to 2500 nm were obtained with a NIRSystem 5000 Autocup sampler (Foss, Rellingen, Germany). Internal seed standards were used as control and analyses were done according to the VDLUFA (Kassel, Germany) calibration equation. The tocopherol content in the oil (OTR) was calculated as the ratio of oil and TTC for which the means of each accession was used.

DNA EXTRACTION AND GENOTYPIC ANALYSIS

DNA was extracted from panel 1 accessions grown at the location Holtsee from one single plant per plot using the NucleoSpin® 96 Plant (4 × 96) kit (Macherey and Nagel, Düren, Germany). The DNA concentration was adjusted to $5\,\mathrm{ng}\,\mu\mathrm{l}^{-1}$ using a TECAN-Freedom EVO 150® robot (Männedorf, Switzerland).

The 13 tocopherol candidate genes (*BnaX.VTE1.a*, *BnaX.VTE1.b*, *BnaA.VTE2.a*, *BnaX.VTE2.b*, *BnaX.VTE3.a*, *BnaX.VTE3.b*, *BnaA.VTE4.a*, *BnaX.VTE4.b*, *BnaX.VTE4.c*, *BnaC.VTE5*, *BnaX.PDS1.a*, *BnaX.PDS1.b*, and *BnaA.PDS1.c*) were identified by BAC library screening and characterized by functional and mapping approaches (Fritsche et al., in preparation; Wang et al., in preparation). We chose different methods for genotyping each *B. napus* panel. First, we sequenced fragments of the 13 tocopherol candidate genes in panel 1 accessions to identify polymorphisms within these genes. Therefore, extracted DNA of panel 1 accessions was used as PCR template. Gene locus specific primer pairs were developed and tested for different regions of the candidate genes. After amplification, fragments displaying the expected lengths on 1% agarose gels were sequenced by Sanger sequencing (Institute for Clinical Molecular Biology, Kiel, Germany). Only primer pairs producing a single PCR fragment were used for genotyping panel 1, which resulted also in high-quality sequence trace files for each fragment (**Table A2** in Appendix). Using DNAStar Lasergene SeqMan Pro 7.2.1 software (Madison, WI, USA), fragments were assembled and the quality of the ABI trace files was analyzed and edited manually by visual examination. We used the TASSEL software (Bradbury et al., 2007) to identify single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) within the sequences of panel 1 accessions. Alignments were constructed with CLC main workbench 5 (CLC bio, Aarhus, Denmark) or the multiple alignment tool CLUSTALW2 (Larkin et al., 2007; Goujon et al., 2010). Comparisons to publicly available sequences were done with the Basic Local Alignment Search Tool (BLAST) from the NCBI website (Altschul et al., 1997).

Second, panel 2 was genotyped with cleaved amplified polymorphic site (CAPS) markers derived from sequence information of panel 1. Restriction enzymes were selected with the restriction site analysis tool implemented in the software CLC main workbench 5 (**Table A3** in Appendix).

The population structure of panel 1 was determined by using 31 publicly available genome-wide microsatellite markers (Cheng et al., 2009). For amplification of the microsatellites M13-tailed primers were used (Schuelke, 2000). PCR reactions were performed with four primers: SSR forward primer with M13F-tail, SSR reverse primer with M13R-tail, IRD700-labeled M13F, and un-labeled M13R primers (MWG Biotech, Inc., Ebersberg, Germany). The PCR products were separated on the LI-COR 4300 DNA analyzer system (LI-COR Biosciences, Lincoln, NE, USA). Due to multiple loci amplification or multiple allelic genotypes of the SSR markers, bands were scored as 1 or 0.

STATISTICAL ANALYSIS

Mean values of each accession were calculated for each field trial and each trait. For each panel, an analysis of variance (ANOVA) was performed with SAS PRO MIXED version 9.2 (SAS Institute, 2009) to examine the effect of genotype, environment, and genotype × environment interaction on the respective traits and to estimate the variance components.

All factors were treated as random effects using the model: $Y_{ijkl} = \mu + l_i + b(l)_{ij} + g_k gl_{ik} + e_{ijkl}$, where y is the respective seed trait of the kth accession tested in the jth block of the ith environment, μ is the overall mean, l_i are the effect of the environment, $b(l)_{ij}$ the block effect, g_k the effect of the accessions, gl_{ik} the interaction effect between accession and environment and e_{ijkl} the random experimental error. Heritability was calculated as: $h^2 = V_\mathrm{g} / V_\mathrm{g} + V_{gl}/l + V_\mathrm{e}/R$, where h^2 is the broad sense heritability, V_g is the genetic variance of the test panel, V_{gl}/l is the variance of the genotype by environment interaction divided by the number of environments and V_e/R is the residual variance divided by the total number of replications.

Population structure of panel 1 was examined with the software STRUCTURE version 2.2.3. (Pritchard et al., 2000) using the admixture model and correlated allele frequencies. For between 1 and 10 subpopulations (K) the burn-in length period of 100,000 iterations, followed by 100,000 Markov Chain iterations were selected (**Figure A1** in Appendix).

Principal component analysis (PCA) was performed based on the above mentioned SSR markers, which were treated as dominant markers, band by band. The first and second principal component was used (D matrix) for the association analysis.

The kinship coefficient K_{ij} between inbreds i and j were calculated based on the SSR markers according to: $K_{ij} = S_{ij-1} / 1 + T + 1$, where S_{ij} was the proportion of marker loci with shared variants between inbreds i and j and T the average probability that a variant from one parent of inbred i and a variant from one parent of inbred j are alike in state, given that they

are not identical by descent (Bernardo, 1993). For the series of T values 0, 0.025,. . ., 0.975 K matrices between all inbreds were calculated. Negative kinship values between inbreds were set to 0. The optimum T value was calculated according to Stich et al. (2008).

Population structure of panel 2 was evaluated by Wang et al. (in preparation) based on genotyping the 133 accessions with 41 SSR markers, and was provided as Q-matrix.

LINKAGE DISEQUILIBRIUM AND ASSOCIATION ANALYSIS

R^2 values of LD and corresponding p-values for all loci pairs were calculated using the software R. For LD decay analysis only SNPs with a minimum frequency of 0.05 were considered. Indels were regarded as one polymorphic site. A non-linear regression of r^2 vs. the genetic map distance (cM) was performed (Heuertz et al., 2006).

Polymorphisms were analyzed for association with the following traits: ATC, GTC, TTC, AGR, SOC, GSL, TOC, and SPC. The two models, general linear model (GLM) and PK-mixed model, were used to analyze associations between polymorphic sites and the traits in panel 1. The first model was conducted with TASSEL using the implemented GLM. Analyses were conditioned with population structure estimates, by using the Q-matrix obtained from the STRUCTURE software. Only polymorphisms with a minor allele frequency of larger than 5% were included in the association analysis. For assuming an association an adjusted p-value (Bonferroni correction) of less than 0.05 was required. The PK-mixed model was constructed as $M_{ip} = \mu + a_p + \sum_{u=1}^{z} D_{iu}\nu_u + g_i^* + e_{ip}$, where M_{ip} was the entry mean of the ith entry carrying allele p, α_p the effect of allele p, e_{ip} the residual, ν_u the effect of the uth column of the population structure matrix D, and g_i^* the residual genetic effect of the ith entry (Stich et al., 2008; Yu et al., 2008b). For panel 2, association analysis of polymorphic sites and tocopherol traits was performed using the PK-mixed model. SSR marker data were developed and provided by Wang et al. (in preparation) which were used for population structure and kinship calculations with the same method described before. Accessions with missing phenotypic or genotypic data were excluded from the analysis.

The R package EMMA (Kang et al., 2008) and the significance threshold of 0.05 was applied to perform the above outlined association analysis of all traits with the polymorphisms. Evaluation of the p-value distribution was done by generating a histogram plot (**Figure A2** in Appendix). To test the global hypothesis, the Bonferroni correction was used (Pocock et al., 1987). The percentage of phenotypic variation explained by the significant SNPs was calculated by $R^2_{\mathrm{LR}} = 1 - \exp(-\frac{2}{n}(\log L_{\mathrm{M}}) - (\log L_0))$, where log L_{M} is the maximum log-likelihood of the model of interest, log L_0 the maximum log-likelihood of the intercept-only model, and n the number of observations (Magee, 1990).

RESULTS

PHENOTYPIC VARIATION OF TOTAL TOCOPHEROL CONTENT AND COMPOSITION

In rapeseed panel 1, TTC ranged from 234.63 to 379.10 mg kg^{-1} with a mean of 304.14 mg kg^{-1} (SD ± 29.17). The mean of TTC in panel 2 was 344.80 mg kg^{-1} (SD ± 39.25) with a range of 197.54–460.07 mg kg^{-1} (**Figure 1A**). AGR varied from 0.46 to 1.51 in panel 1 and from 0.33 to 2.14 in panel 2 (**Figure 1B**). In the ANOVA highly significant ($p \leq 0.01$) effects of genotype and genotype × environment interaction were observed for all traits, except for the genotype × environment interaction effect for AGR in panel 1 (**Table 1**). High broad sense heritability values were estimated for all traits; from 0.62 to 0.78 for TTC and from 0.77 to 0.94 for AGR (**Table 1**). The heritability values for ATC and GTC ranged from 0.77 to 0.89.

Seed quality traits such as GSL, SOC, and SPC were measured with accessions from panel 1 in order to unravel any relationship with TTC or AGR. GSL contents ranged from 6.50 to 114.55 μmol g^{-1} with a mean of 76.59 (SD ± 25.37). Phenotypic variation was also found for SOC (44.64–58.88% DW) and for SPC (17.25–24.70% DW). We observed high heritability values for all three characters, ranging from 0.63 to 0.98 (**Table 2**).

ATC and GTC were significantly ($p < 0.01$) related with TTC and AGR (**Table 3**). Correlations between TTC and AGR as well as between ATC and GTC were not significant. Moreover, the correlation between tocopherol traits and SOC was not significant, whereas a negative correlation was detected between SOC and SPC ($p < 0.01$). All tocopherol traits, except AGR, were significantly ($p < 0.01$) negatively correlated with SPC. Apart from ATC and SOC, the GSL content was significantly ($p < 0.01$) correlated with all other traits.

IDENTIFICATION OF POLYMORPHISMS WITHIN TOCOPHEROL GENES

Single PCR products of the expected fragment size were detected for all 13 candidate genes which were amplified in the panel 1 accessions. However, specific primer pairs yielding high-quality sequences were developed for at least one region in nine genes (**Table A2** in Appendix). The fragments of the remaining four candidate genes had poor sequence quality and were not further investigated in the present study. The amplified regions covered between 24.0 and 72.8% of the genes and included exons as well as introns (**Table 4**).

In summary, the sequencing of fragments of nine candidate genes with a total length of 6640 bp revealed 51 SNPs and 5 indels (**Table 5**). Taking monomorphic gene fragments into account we observed a density of 1 SNP/130 bp and 1 indel/1328 bp.

The identified polymorphisms were classified according to their minor allele frequency which displayed the frequency at which the less common allele of a polymorphism occurred in the accessions of panel 1. Setting a threshold of 5%, we found polymorphic sites in two candidate genes (*BnaA.PDS1.c*, *BnaX.VTE3.a*) whereas low polymorphic sites (frequency < 5%) were detected in three genes. We found no polymorphisms in the amplified fragments of the remaining four genes (**Table A5** in Appendix).

For the gene *BnaA.PDS1.c* we identified in two amplified fragments in total 25 SNPs and three indels within 1033 bp, equivalent to an average density of 1 SNP/41 bp and an indel density of 1 indel/344 bp. Of these, 13 polymorphic sites were located in exons and 15 polymorphic sites within the only intron of this gene. LD with a mean r^2 value of 0.74, $p < 0.001$ was observed for the *BnaA.PDS1.c* polymorphisms (**Figure 2**). A LD block (mean r^2 within LD block = 0.92, $p < 0.001$) between SNP 996 and SNP 1250 was found, spanning 254 bp and including the insert region of the gene (**Figure 2**).

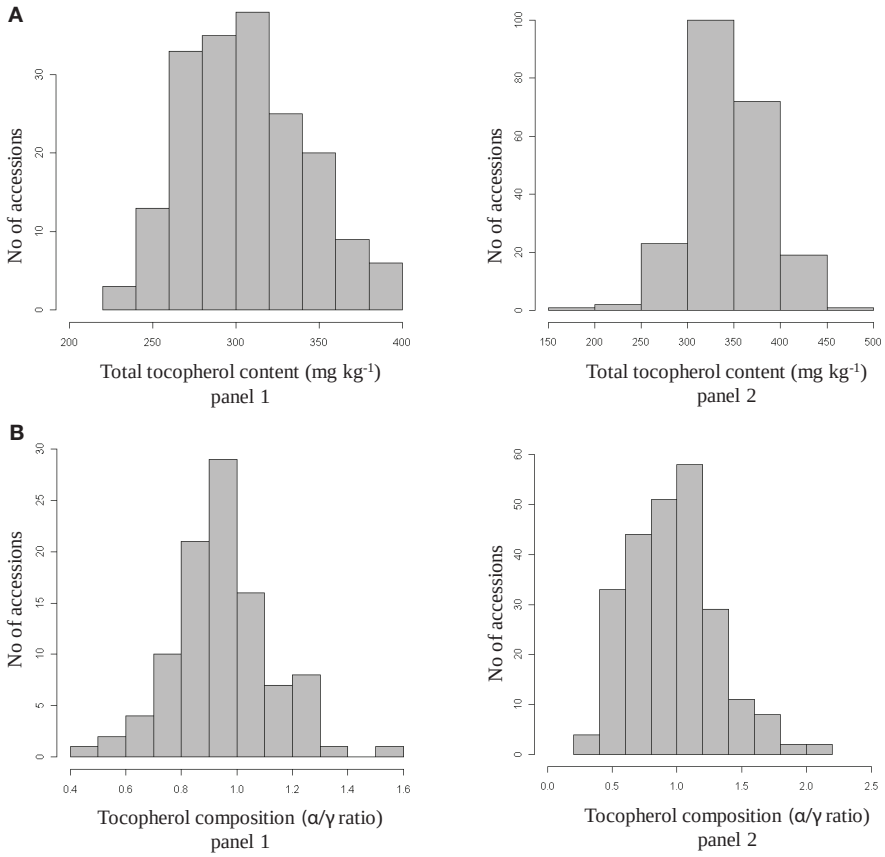


FIGURE 1 | Distribution of total tocopherol content (A) and composition (B) in two panels which consisted of 96 and 133 *B. napus* accessions, respectively. Plants were grown in the field at two different locations. Total tocopherol content is given as mg kg^{-1} and the composition as α/γ ratio.

Table 1 | Ranges, means, ANOVA statistics (components of variance of genotype, genotype × environment interaction, and residual error), and heritability estimates of two *B. napus* panels, consisting of 96 (panel 1) and 133 accessions (panel 2), evaluated in field trials for seed α- and γ-tocopherol content, total tocopherol content (mg kg^{-1}), and tocopherol composition (α/γ ratio).

Statistics	ATC		GTC		TTC		AGR	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
Min	85.48	83.39	107.76	84.99	234.63	197.54	0.46	0.33
Max	208.29	286.30	196.39	280.83	379.10	460.07	1.51	2.14
Mean	146.74	162.45	157.29	177.66	304.14	344.80	0.94	0.97
SD	21.99	37.39	18.12	36.50	29.17	39.25	0.17	0.35
V_g	435.11**	833.01**	246.93**	926.30**	621.13**	840.27**	29,150**	69,390**
V_{gl}	66.14**	471.21**	42.34**	487.11**	191.56**	950.03**	1135	40,150**
V_e	89.98	27.09	131.79	27.06	318.63	95.21	4983	180.26
h^2	0.89	0.77	0.82	0.78	0.78	0.62	0.94	0.77

*F-statistic: *significant at $p \leq 0.05$; **significant at $p \leq 0.01$.*

SD, standard deviation; ATC, α-tocopherol content; GTC, γ-tocopherol content; TTC, total tocopherol content; AGR, α/γ ratio; V_g, genetic variance; V_{gl}, variance of genotype × environment interaction; V_e, residual error; h^2, broad sense heritability.

In *BnaX.VTE3.a* six SNPs and no indels were identified within 753 bp, which corresponds to an average SNP density of 1 SNP/125 bp. The LD between pairs of SNPs ranged from 0.04 to 1 (**Figure 2**) with an average $r^2 = 0.39$ ($p < 0.001$). Two LD blocks were observed from SNP 657 to SNP 741, comprising 84 bp, and from SNP 342 to SNP 359, comprising

Table 2 | Statistics of the NIRS analysis of 96 *B. napus* accessions (panel 1) with several parameters (ranges, means, components of variance of genotype, genotype × environment interaction and residual error, and broad sense heritability) of seed glucosinolate (μmol g^{-1}), protein (% DW), oil content (% DW), and tocopherol in oil (oil/total tocopherol ratio).

	GSL	SOC	SPC	OTR
Min	6.50	44.68	17.25	1192.57
Max	114.55	58.88	24.70	2300.23
Mean	76.59	51.26	20.77	1694.30
SD	25.37	2.72	1.63	210.37
V_g	622.90**	1.21**	4.97**	n.a.
V_{gl}	13.49**	0.81**	1.41**	n.a.
V_e	22.71	1.18	1.58	n.a.
h^2	0.81	0.63	0.98	n.a.

**Significant at $p \leq 0.05$; **significant at $p \leq 0.01$; n.a., not available; GSL, glucosinolate; SPC, seed protein content; SOC, seed oil content; OTR, oil-tocopherol ratio; SD, standard deviation; V_g, genetic variance; V_{gl}, variance of genotype × environment interaction; V_e, residual error; h^2, broad sense heritability.*

Table 3 | Correlation coefficients of α- and γ-tocopherol, total tocopherol content, tocopherol composition, glucosinolate, oil, and protein content of 96 accessions in panel 1.

Correlation coefficients	GTC	TTC	AGR	SOC	SPC	GSL
ATC	0.13	0.76*	0.67*	0.04	−0.24*	−0.01
GTC		0.74*	−0.59*	0.03	−0.26*	−0.32*
TTC			−0.06	0.04	−0.32*	−0.22*
AGR				0.01	−0.02	0.22*
SOC					−0.51*	0.08
SPC						0.31*

**Significant at $p \leq 0.01$. ATC, α-tocopherol content; GTC, γ-tocopherol content; TTC, total tocopherol content; AGR, α/γ ratio; GSL, glucosinolate; SPC, seed protein content; SOC, seed oil content.*

Table 4 | Tocopherol biosynthesis genes of *B. napus*, their genomic gene length, amplified gene region, total fragment length, and number of base pairs aligned after sequencing of the gene fragment of panel 1 accessions.

Gene name	Genbank	Genomic sequence size (bp)	Amplified region	Fragment length (bp)	Sequence length (bp)			ORF coverage (%)	No. of *B. napus* accessions sequenced
					Coding region	Non-coding region	Total		
BnaX.PDS1.a	JN834026	1390	251–1332	1032	853	67	920	66.2	94
BnaX.PDS1.b	JN834015	1399	402–1199	798	598	0	598	42.7	90
BnaA.PDS1.c	JN834016	1418	1–461	461	304	0	304	72.8	96
			507–1327	821	649	80	729		
BnaX.VTE1.a	JN834017	2809	1707–2746	1040	579	281	860	30.6	92
BnaX.VTE1.b	JN834018	2866	1498–2649	1152	615	413	1028	35.9	86
BnaX.VTE2.b	JN834020	2352	1057–2086	1030	399	381	780	33.2	96
BnaX.VTE3.a	JN834021	1185	190–1045	856	612	141	753	63.5	96
BnaX.VTE3.b	JN834022	1181	48–617	570	435	43	478	40.5	96
BnaX.VTE4.b	JN834023	2062	1429–2008	580	280	214	494	24.0	95

17 bp. In total, 33 polymorphic sites formed 561 pairs of r^2 calculations, of which 59% were observed to have significant LD ($p < 0.05$). Plotting r^2 values against physical distance (bp) between linked SNP loci pairs indicated that LD decays from 0.45 to 0.25 when physical distance increased to 750 bp (**Figure 3**).

Table 5 | Polymorphic sites within tocopherol candidate genes evaluated in 96 *B. napus* accessions (panel 1), their position in the gene and exon/intron-position, predicted amino acid change, and minor allele frequency.

Position	Polymorphisms	Exon/Intron	Predicted amino acid change	Minor allele frequency (%)[1]
***BnaX.VTE2.b* (1057-2086)[2]**				
1464[3]	G/A[4]	exon 8	G → A	4.16
***BnaX.VTE3.a* (190-1045)**				
285	T/C	exon 1		16.67
342	C/T	exon 1		10.41
399	T/C	exon 1		9.60
657	G/T	exon 2		15.62
738	T/A	exon 2		15.62
741	T/C	exon 2		16.67
***BnaX.VTE3.b* (48-617)**				
326–329	4/0	exon 1	Frame shift	1.78
***BnaX.PDS1.a* (251-1332)**				
318	C/A	exon 1		1.06
336	G/T	exon 1		1.06
339	C/T	exon 1		1.06
340	0/9	exon 1	S → RTS	1.06
340	T/C	exon 1	S → A	1.06
342	T/C	exon 1		1.06
360	C/G	exon 1		1.06
456	C/A	exon 1		1.06
460	A/C	exon 1		1.06
687	C/G	exon 1		1.06
720	C/A	exon 1		1.06
800	A/G	exon 1	N → S	1.06
861	C/T	exon 1		1.06
909	C/T	exon 1		1.06
915	G/T	exon 1		1.06
927	G/A	exon 1		1.06
930	C/T	exon 1		1.06
1011	T/C	exon 1		1.06
1029	G/A	exon 1		1.06
1033	A/G	exon 1	I → V	1.06
***BnaA.PDS1.c* (1-461)**				
35	A/G	exon 1	Q → R	8.88
50–55	6/0	exon 1	DEA → A	23.40
60	T/A	exon 1		29.67
61	G/A	exon 1	A → T	5.49
174	T/C	exon 1		25.00
183	C/T	exon 1		4.44
207	G/A	exon 1		10.00
213	C/A	exon 1		41.77
297	G/C	exon 1		13.48
300	C/A	exon 1		19.05
***BnaA.PDS1.c* (507-1327)**				
543	C/T	exon 1		28.42
996	T/C	exon 1		13.04
1167	C/T	intron 1		14.29
1169	A/T	intron 1		14.29
1170	A/T	intron 1		14.29
1171	G/A	intron 1		14.29
1172	A/C	intron 1		14.29
1173	C/A	intron 1		14.29

Continued

Table 5 | Continued

Position	Polymorphisms	Exon/Intron	Predicted amino acid change	Minor allele frequency (%)[1]
1180	A/T	intron 1		14.29
1184–85	2/0	intron 1		13.19
1213	T/A	intron 1		15.11
1222	G/A	intron 1		14.61
1226	C/G	intron 1		14.44
1227	T/A	intron 1		14.44
1229	G/C	intron 1		14.44
1231	1/0	intron 1		14.44
1232	T/C	intron 1		14.44
1250	G/T	exon 2		10.00

[1] *Percentage of accessions of panel 1 with the minor allele.*

[2] *Numbers in parentheses give the gene region amplified, relative to the ATG start codon.*

[3] *Positions indicate the polymorphic site position in the gene relative to the ATG start codon. Positions of indels refer to the start of the insertion in the B. napus cv. Tapidor allele of the respective gene.*

[4] *Numbers or letters preceding the slash point the size of the insertion or nucleotide of the Tapidor allele.*

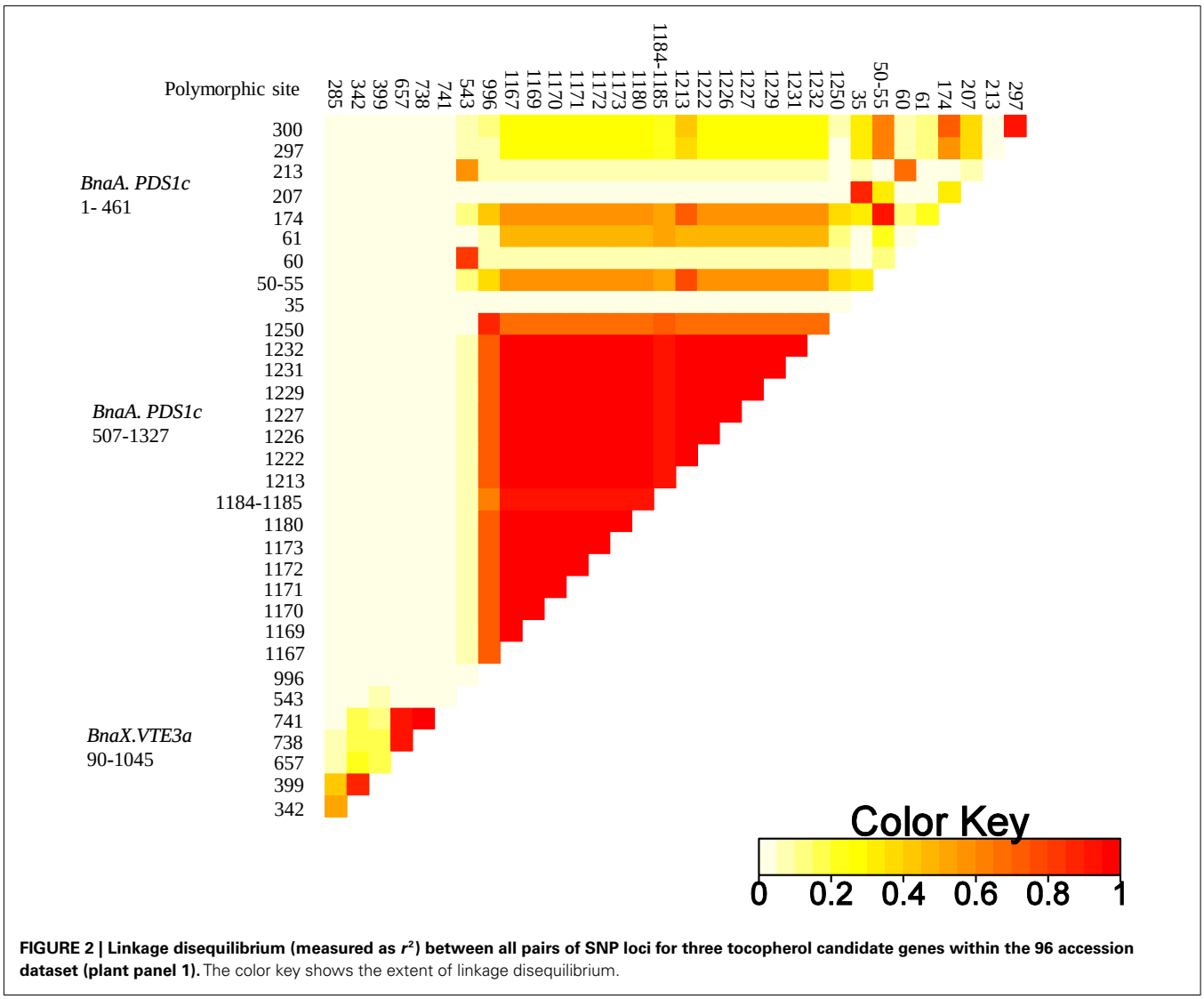


FIGURE 2 | Linkage disequilibrium (measured as r^2) between all pairs of SNP loci for three tocopherol candidate genes within the 96 accession dataset (plant panel 1). The color key shows the extent of linkage disequilibrium.

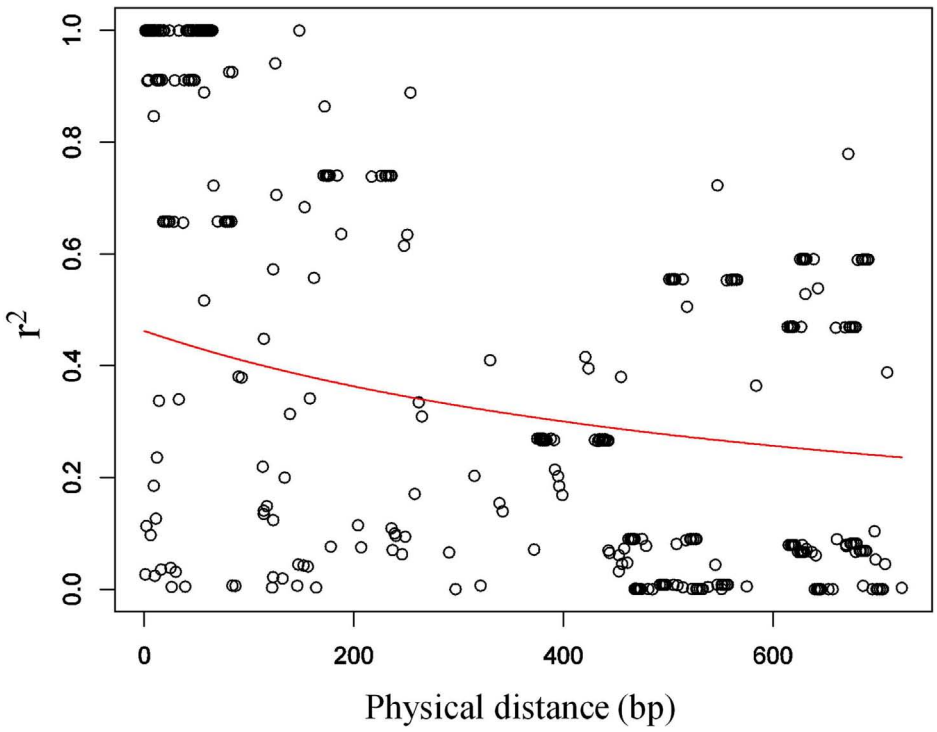


FIGURE 3 | Plot of linkage disequilibrium measured as squared correlation of allele frequencies (r^2) against physical distance (bp) between linked SNP loci pairs from tocopherol biosynthesis genes in the 96 accession dataset. The red line is the non-linear regression trend line of r^2 vs. genetic distance (bp).

Table 6 | Genotyping results of panel 2.

Gene	SNP/indel position	No of accessions per allele						Minor allele frequency (%)
		C	T	G	A	Ins	Del	
BnaA.PDS1.c	543	24	99					19.5
BnaX.VTE3.a	285	10	123					7.5
BnaX.VTE3.a	342	126	5					3.8
BnaX.VTE3.b	326–329					0	132	0
BnaX.VTE2.b	1464			2	130			1.5

Polymorphic sites within the genes BnaA.PDS1.c, BnaX.VTE3.a, BnaX.VTE3.b, and BnaX.VTE2.b and their occurrence in the 133 accessions panel determined with CAPS markers. Due to missing data accessions do not add up to the total number of 133 accessions for each gene.
ins, insertion; del, deletion.

Panel 2 was genotyped with allele-specific CAPS markers, which were based on polymorphisms of the candidate genes *BnaA.PDS1.c*, *BnaX.VTE3.a*, *BnaX.VTE3.b*, and *BnaX.VTE2.b* (**Table A3** in Appendix). CAPS marker analysis enabled the determination of the nucleotide composition at the respective position. In panel 2 the minor allele frequencies of the analyzed SNPs ranged between 1.5 and 19.5% (**Table 6**). The deletion within *BnaX.VTE3.b* which had been detected in panel 1 was not polymorphic in panel 2 and was therefore excluded from further analysis.

POPULATION STRUCTURE

We analyzed the population structure of panel 1 with 31 SSR markers. Of these, seven markers turned out to be monomorphic or gave ambiguous results and were excluded from further analysis. The remaining 24 SSR loci were polymorphic and resulted in 52 different alleles. The highest likelihood for a subpopulation was obtained with $K = 4$ and $Ln\ p(D) = -1986.6$ and a variance value of 264.4 using the software STRUCTURE (**Table 7**, **Figure 4A**). The population structure was also examined by PCA using the same data of the 24 SSR markers. The first and second principal component explained 11.5 and 7.8% of the variations, respectively (**Figure 4B**). No distinct subgroups were observed.

The kinship coefficient matrices between all accessions were calculated based on the data of the above mentioned 24 SSR markers. The highest kinship coefficient frequency (99.18%) was detected for values between 0 and 0.05, whereas 0.8% of the values

Table 7 | Population substructure estimation of 96 *B. napus* accessions (panel 1) by evaluating 24 SSR markers using STRUCTURE.

K	*Ln p(D)*	SD
1	−2226.2	22.9
2	−2048.9	83.5
3	−1997.6	152.6
4	−1986.6	264.4
5	−2082.8	493.3
6	−2317.0	1011.3
7	−2247.9	894.5
8	−2386.9	1183.7
9	−2424.4	1221.4
10	−2475.2	1246.7

The burn-in length period and the Markov Chain steps were both set to 100,000 iterations in a model allowing admixture and correlated allele frequencies. One to ten subpopulations were tested.

K, no. of populations, Ln p(D), estimated natural logarithm probability of the data; SD, standard deviation.

were above 0.05, indicating that most of the panel 1 accessions had a low level of relatedness (**Figure 5**). Population structure of panel 2 will be described elsewhere (Wang et al., in preparation).

ASSOCIATION ANALYSIS

Association analyses were performed with all polymorphic sites (minor allele frequency > 5%) of the candidate genes and the trait data for each field trial of panel 1 by using two models (GLM + Q/PK-mixed model).

With GLM + Q in total, 53 significant associations ($p < 0.05$) were found for 26 polymorphisms (**Table 8**). The second model applied during this study, the PK-mixed model, combines the kinship coefficient between individuals with population structure, estimated by PCA. Thus, some significant associations of the GLM + Q were excluded and the result was reduced to 26 significant associations of 12 polymorphisms (**Table 8**). Considering these results in total, seven polymorphisms within the candidate region of *BnaA.PDS1.c* (position 1–405) were significantly associated with ATC, one of these (SNP 61) was associated with the trait at both field trial locations. We found significant associations with AGR for four SNPs of which SNP 61 was associated with this trait at both field trial sites. Three SNPs (positions 35, 174, 207) were significantly associated with TTC as well as OTR. The indel on position 50–55 was found to be associated with TTC at the field trial site Giessen. Among the quality traits, only SOC was found to be associated with three SNPs in that gene fragment. In the second amplified region of *BnaA.PDS1.c* (positions 507–1327), SNP 543 was found to be significantly ($p < 0.05$) associated with ATC, and TTC in both locations according to both models. The phenotypic variance (R^2) explained by that single SNP was between 5.42 and 10.87% in GLM + Q and between 4.96 and 9.09% in PK-mixed model (**Table 8**). Within *BnaX.VTE3.a* (positions 190–1045), SNP 285 was found to be significantly associated with AGR (Holtsee and Giessen) as well as ATC (Giessen). This SNP explained 6.37–10.05% of the phenotypic variance of the investigated traits. Another SNP detected at position 342 was associated with the trait GTC (Holtsee). By comparing both models, 29 of the 53 significant associations of the GLM + Q were consistent with the second model. Twenty-four associations of the GLM + Q model were not significant in the PK-mixed model. None of the models found associations to GSL content.

In panel 2 we detected in total five significant associations with three polymorphic sites. SNP 543 within *BnaA.PDS1.c* was significantly associated with AGR in 2009 ($p = 0.033$). On average, all panel 2 accessions with the T-allele had −0.14 AGR (**Table 9**). Similarly, significant associations between SNP 285 within *BnaX.VTE3.a* and AGR ($p = 0.014$) as well as ATC ($p = 0.017$) in 2009 were found. The effect of the T-allele in panel 2 was on average $-25.69\ \mathrm{mg\,kg^{-1}}$ α-tocopherol. SNP 1464 within *BnaX.VTE2.b* was included in the calculations although the allele frequency was found to be 1.5% (**Table 6**). A significant association of SNP 1464 was found for GTC ($p = 0.024$) and AGR ($p = 0.035$) in 2009.

DISCUSSION

Enhancing the content and composition of tocopherol is one important step to further improve oil quality of rapeseed. In the present study we have demonstrated for the first time an association between tocopherol traits and allelic variations at various candidate gene loci. These polymorphisms represent promising candidates for the development of molecular markers for marker-assisted breeding of rapeseed varieties with enhanced tocopherol qualities. Originally, association studies were developed to dissect the genetics of human diseases but rapidly they have also become an important method in plant genetics to identify alleles and loci responsible for phenotypic trait variation. Association studies can be classified into genome-wide and candidate gene approaches (Zhu et al., 2008). In rapeseed, the genome-wide approach was applied in numerous studies but in none of them the marker density was sufficient for genome-wide association mapping (Hasan et al., 2008; Honsdorf et al., 2010; Zou et al., 2010; Jestin et al., 2011; Rezaeizad et al., 2011). Until now, only one candidate gene-based study has been carried out, investigating the association of *BnaA.FRI.a* haplotypes with flowering time (Wang et al., 2011). The genetic architecture of the tocopherol biosynthetic pathway has been almost completely unraveled in the model species *A. thaliana* (Mène-Saffrané and DellaPenna, 2010), a close relative of *B. napus*. These findings provided the incentive to study tocopherol biosynthesis genes in rapeseed as already performed for oil crops as soybean and sunflower (Li et al., 2010; Dwiyanti et al., 2011; Haddadi et al., 2011) as well as non-oil crops as tomato and maize (Wong et al., 2003; Chander et al., 2008; Almeida et al., 2011). In a separate study, we have identified all genes and several orthologs from the biosynthesis pathway of *B. napus* according to their high sequence homologies to *A. thaliana* genes (Wang et al., in preparation). Further on, we have studied their expression and function and mapped them to *B. napus* linkage groups (Endrigkeit, 2007; Wang et al., in preparation). These data together with partial coincidence between map positions and the positions of major QTL for tocopherol content and composition provided convincing evidence for their role as functional genes for tocopherol biosynthesis in oilseed rape.

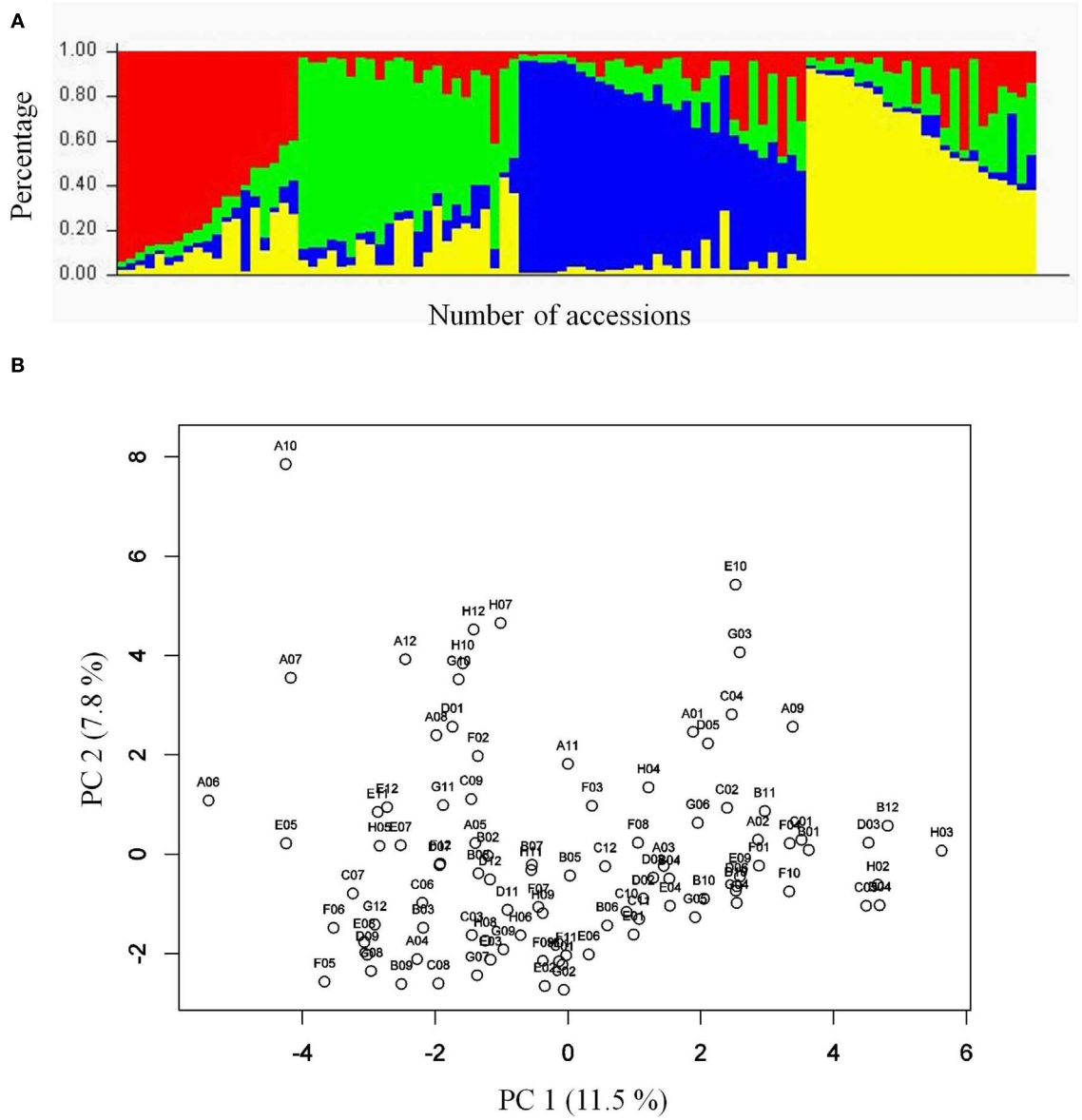


FIGURE 4 | (A) Population structure of 96 *B. napus* accessions of panel 1 based on 24 SSR markers under the assumption of subpopulation $K = 4$. *Brassica napus* accessions are represented by a bar which is divided into several parts with different colors according to the accessions estimated fractions of the four clusters. Numbers on the *x*-axis indicate the accession and numbers on the *y*-axis shows the group membership in percent. **(B)** Principal component analysis of panel 1 accessions based on 24 SSR markers. PC 1 and PC 2 refer to the first and second principal components. The numbers in parentheses refer to the proportion of variance explained by the principal components.

There is substantial phenotypic variation for tocopherol traits in *B. napus*. The two diversity sets used in our study (panels 1 and 2) displayed different ranges of variation for tocopherol content and composition. The variation in panel 1 consisting mainly of winter types was comparable to the results obtained by Goffman and Becker (2002), who found a maximum of 367 mg kg^{-1} among 87 winter type rapeseed accessions. As expected, genetic variation for tocopherol content (197.54–460.07 mg kg^{-1}) and composition (0.33–2.14 α/γ ratio) was much higher in panel 2, which is possibly due to its different composition (98 winter type and 35 spring type accessions). This higher genetic variation explains the high heritability estimates in both panels ($h^2 = 0.62$–0.94) compared to Marwede et al. (2004) ($h^2 = 0.23$–0.50), who analyzed three doubled haploid populations with a lower genetic variation for tocopherol traits.

Rapeseed is an allopolyploid species, thus it was not surprising to find several homologous sequences for each *A. thaliana* tocopherol gene (Endrigkeit, 2007; Wang et al., in preparation). In total, we analyzed 13 candidate genes for polymorphism screening in panel 1 and used between 5 and 28 primer pairs to amplify parts

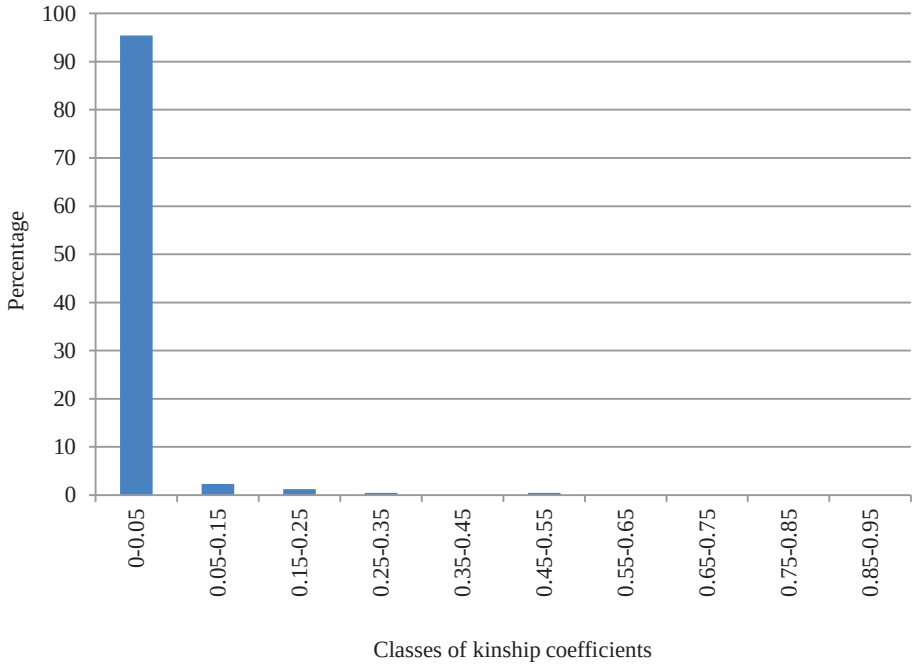


FIGURE 5 | The distribution of kinship relative coefficients between 96 *B. napus* accessions of panel 1.

of the genes (data not shown). To circumvent the known problems in direct gene sequencing of allopolyploid species, where several orthologous and paralogous gene copies often result in insufficient sequence quality for SNP detection, we used only those primer pairs producing a single PCR fragment and yielding high-quality sequence trace files. This approach was already applied successfully in several previous studies (Ganal et al., 2009; Westermeier et al., 2009; Durstewitz et al., 2010). We found large differences in the density of polymorphisms within the analyzed tocopherol genes and the allele frequencies of these polymorphisms in panel 1. In two genes (*BnaA.PDS1.c*, *BnaX.VTE3.a*) many nucleotide variations were identified, while in the amplified fragments of seven candidate genes no or only rare polymorphisms (frequency < 5%) were detected (**Table A5** in Appendix). One possible reason for these findings may be the short and intensive breeding history of rapeseed that has led to a reduced allelic diversity in conventional winter oilseed *B. napus* material (Becker et al., 1995; Hasan et al., 2006; Bus et al., 2011). The discovery of rapeseed varieties with low erucic acid and low GSL content represents major achievements in the rapeseed breeding history but also constituted genetic bottlenecks. Today's spring and winter rapeseed is derived from a limited number of genetic resources, thus most of them share the same genetic background (Friedt and Snowdon, 2009). Panel 1 almost exclusively consisted of winter rapeseed accessions, mainly from Europe; therefore, we decided to use a second panel which encompasses also spring type accessions. A further possible explanation for the low SNP frequency in panel 1 may be the short size of the amplified fragments and the high stringency conditions chosen for obtaining high-quality sequences. Future studies will have to clarify whether sequence variations detected here or any other variations beyond the amplified regions are the reasons for the observed phenotypic variations.

The SNP density of *BnaA.PDS1.c* (1 SNP/41 bp) was compared with earlier studies in rapeseed. Similar SNP densities were found in EST derived amplicons from 16 rapeseed cultivars (1 SNP/42 bp; Durstewitz et al., 2010) and in the *BnaA.FRI.a* gene after genotyping 95 rapeseed accessions (1 SNP/66 bp; Wang et al., 2011). A considerably lower rate (1 SNP/247 bp) was reported by Westermeier et al. (2009), who surveyed 18 genomic candidate sequences across six rapeseed genotypes. When considering the SNP distribution in the transcriptome of the two parents of the Tapidor and Ningyou7 DH population, which has been frequently used in genetic studies of rapeseed, an overall polymorphism rate between 1 SNP/1.2 kb and 1 SNP/2.1 kb was found (Trick et al., 2009). SNP densities were also calculated for other oil crops such as soybean (1 SNP/273 bp; Zhu et al., 2003), sunflower (1 SNP/69 bp; Fusari et al., 2008), and olive (1 SNP/156 bp; Reale et al., 2006). Taken together, SNP density varies much between species. Thus, comparisons between species should be ideally restricted to orthologous genes (Krutovsky and Neale, 2005). Consequently, we performed an *in silico* comparison of the SNP density of known *A. thaliana* tocopherol loci (**Table A4** in Appendix) by using POLYMORPH (Fitz, J. et al., personal communication). We observed SNP density values ranging from 1/30 to 1/624 bp (**Table A4** in Appendix), thus demonstrating a broad spectrum of genetic variation within tocopherol biosynthesis genes. Similar to our findings, *A. thaliana VTE1* and *VTE2* are less polymorphic than the other genes. In *B. napus*, we found a similar SNP density in *BnaX.VTE3.a* as compared to the *A. thaliana VTE3* gene (1/170 bp) but not for *PDS1* (1/415 bp). Interestingly, SNP densities for *A. thaliana* actin genes (*ACT2*, *ACT8*) were 1/194 and 1/203 bp and two randomly

Table 8 | Polymorphisms of three tocopherol candidate regions of the 96 *B. napus* accession panel significantly associated ($p < 0.05$) with tocopherol and seed quality traits and their percentage on phenotypic variation evaluated by GLM + Q and PK-mixed model.

Gene	Region	SNP/indel position	Trait	Environment	GLM + Q		PK-mixed model	
					p-Value*	R^2 (%)	*p*-Value*	R^2 (%)
BnaA.PDS1.c	1–461	35	OTR	Giessen	0.001	11.41	<0.001	14.72
			TTC	Giessen	0.011	6.92	0.003	9.75
			SOC	Giessen	0.030	5.51	0.028	5.64
			ATC	Giessen	n.s.	n.a.	0.033	5.32
			TTC	Holtsee	n.s.	n.a.	0.043	4.45
		50–55	ATC	Giessen	0.040	4.53	0.023	5.72
			TTC	Giessen	n.s.	n.a.	0.039	4.76
		60	ATC	Giessen	0.039	4.74	0.031	5.19
		61	AGR	Giessen	0.003	10.17	0.002	11.21
			AGR	Holtsee	0.005	9.12	0.008	8.23
			SOC	Giessen	0.006	8.77	0.019	5.99
			ATC	Giessen	0.010	7.42	0.005	8.79
			ATC	Holtsee	0.011	7.36	0.012	7.00
			SOC	Holtsee	0.029	5.50	n.s.	n.a.
			GTC	Giessen	0.037	4.89	n.s.	n.a.
		174	ATC	Giessen	0.009	7.59	0.006	8.46
			AGR	Giessen	0.009	7.50	0.014	6.85
			ATC	Holtsee	0.042	4.75	n.s.	n.a.
			TTC	Giessen	0.048	4.17	0.025	5.73
			OTR	Giessen	n.s.	n.a.	0.042	4.53
		207	OTR	Giessen	<0.001	14.43	<0.001	16.93
			TTC	Giessen	0.005	8.34	0.002	10.48
			SOC	Giessen	0.008	8.17	0.009	7.83
			ATC	Giessen	0.013	6.98	0.012	7.26
		213	SOC	Giessen	0.027	6.58	0.020	7.27
			SOC	Holtsee	n.s.	n.a.	0.040	5.76
		297	AGR	Giessen	0.018	6.50	0.046	4.72
			ATC	Giessen	0.026	5.77	0.045	4.74
		300	AGR	Giessen	0.007	11.98	0.006	9.35
			AGR	Holtsee	0.024	6.42	n.s.	n.a.
			ATC	Giessen	n.s.	n.a.	0.008	8.68
BnaA.PDS1.c	507–1327	543	ATC	Giessen	0.002	10.87	0.003	9.09
			ATC	Holtsee	0.004	10.01	0.011	7.26
			TTC	Giessen	0.006	8.04	0.011	7.11
			TTC	Holtsee	0.022	6.17	0.030	4.96
			OTR	Giessen	0.029	5.42	0.025	5.48
			AGR	Giessen	0.035	5.23	n.s.	n.a.
			AGR	Holtsee	0.043	4.91	n.s.	n.a.
		1167	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1169	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1170	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1171	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1172	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1173	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1184–85	SOC	Giessen	0.043	4.72	n.s.	n.a.
		1180	SOC	Giessen	0.043	4.89	n.s.	n.a.
		1213	SOC	Giessen	0.002	13.71	n.s.	n.a.
			OTR	Giessen	0.016	9.12	n.s.	n.a.

(Continued)

Table 8 | Continued

Gene	Region	SNP/indel position	Trait	Environment	GLM + Q		PK-mixed model	
					p-Value*	R^2 (%)	*p*-Value*	R^2 (%)
			SPC	Giessen	0.033	7.65	n.s.	n.a.
			ATC	Giessen	0.045	7.10	n.s.	n.a.
		1226	SOC	Giessen	0.049	4.66	n.s.	n.a.
		1227	SOC	Giessen	0.049	4.66	n.s.	n.a.
		1229	SOC	Giessen	0.049	4.66	n.s.	n.a.
		1231	SOC	Giessen	0.040	4.72	n.s.	n.a.
		1232	SOC	Giessen	0.049	4.66	n.s.	n.a.
BnaX.VTE3.a	190–1045	285	AGR	Giessen	0.003	9.50	0.002	10.05
			ATC	Giessen	0.004	8.86	0.002	9.76
			AGR	Holtsee	0.011	7.14	0.016	6.37
		342	GTC	Holtsee	n.s.	n.a.	0.028	5.35
		399	GTC	Holtsee	0.049	4.28	n.s.	n.a.

The two field trial locations of panel 1 were analyzed separately due to significant genotype × environment effects.

R^2, variance explained; ATC, tocopherol content; GTC, γ-tocopherol content; TTC, total tocopherol content; AGR, α/γ ratio; GSL, glucosinolate; SPC, seed protein content; SOC, seed oil content; n.s., not significant; n.a., not available.

**p-values after Bonferroni multiple test correction.*

Table 9 | Association between SNPs and tocopherol traits and allele mean differences of panel 2 accessions.

Gene	SNP position	Trait	Environment[1]	*p*-Value	Mean difference				R^2 (%)
					C	T	G	A	
BnaA.PDS1.c	543	AGR	2009	0.033	0	−0.14			3.49
BnaX.VTE3.a	285	ATC	2009	0.017	0	−25.69			5.70
		AGR	2009	0.014	0	−0.24			5.66
BnaX.VTE2.b	1464	GTC	2009	0.024			44.29	0	4.51
		AGR	2009	0.035			−0.44	0	4.04

R^2, variance explained; ATC, α-tocopherol content (mg kg^{-1}); GTC, γ-tocopherol content (mg kg^{-1}); TTC, total tocopherol content (mg kg^{-1}); AGR, α/γ ratio.

[1]For panel 2 we termed the two field trial seasons as environment 2008 or 2009 Figures.

chosen loci (*RPP1*, *LOV1*) involved in defense response exhibited SNP densities of 1/28 and 1/209 bp, respectively. In conclusion, the SNP densities of orthologous genes mainly depend on the genes function and on the plant material chosen for the analysis.

The decay rate of LD with distance is an important parameter determining the resolution of an association study. In our study LD decayed within a physical distance of only 750 bp. The rapid LD decline observed here confirms the power of a candidate gene-based association approach as had been demonstrated for a rapeseed *FRI* homolog tested for association with flowering time (Wang et al., 2011). Comparable to our results, LD decayed rapidly within 2 kb. Other association studies in rapeseed were based on mapped markers, where LD was found to extend over 2 cM in different cultivars (Ecke et al., 2010; Bus et al., 2011) and over 5 cM in parental lines (Würschum et al., 2012). A whole chromosome association approach limited to chromosome A09 revealed an LD extent of 1 cM, when r^2 dropped to 0.2 (Wang et al., in preparation). Altogether, these data demonstrate that in rapeseed the degree of LD strongly depends on the plant material and the genomic region analyzed as previously examined for other crops (Rafalski, 2002; Jung et al., 2004; Stich et al., 2006; Gore et al., 2009; Ecke et al., 2010).

A further key aspect in conduction of an association approach is the careful consideration of population structure to avoid confounding effects. We applied two different methods (analysis by STRUCTURE software and PCA) with differing results for panel 1. STRUCTURE suggests four subpopulations, indicating a sufficient number of markers for subpopulation calculation, whereas PCA contrasted these findings. No individual subgroups were separated by the principal components 1 and 2, which is probably due to the large proportion of winter rapeseed accessions in panel 1. In comparison, Ecke et al. (2010) could not find population substructure after genotyping 85 winter rapeseed varieties with 89 markers. This reflected the low genetic diversity of winter type rapeseed (Hasan et al., 2006; Bus et al., 2011). We got supporting evidence that the number of background markers were sufficient

for population structure and kinship estimation by calculating the *p*-values for the association of the SSRs with the phenotypic traits. The corresponding QQ-plot showed a uniform distribution of the *p*-values for most of the SSR loci (**Figure A3** in Appendix). This indicates that population structure and kinship are adequately modeled by the markers. In our second panel, three subgroups were detected with STRUCTURE which proved to be in accordance with the results from the PCA (Wang et al., in preparation). The data of panel 2 indicate that winter and spring cultivars of *B. napus* represent genetically distinct groups and corroborate results from former studies (Diers and Osborn, 1994; Hasan et al., 2006).

Because the integration of population structure is considered to be an important factor for straight analysis in association models (Flint-Garcia et al., 2005; Myles et al., 2009; Hall et al., 2010) we decided to integrate two different models in our analysis of which the GLM + Q model was first applied. In order to eliminate spurious associations as a result of relatedness between individuals we also performed the PK-mixed model, which included both population structure (by PCA) and kinship (Yu et al., 2006; Stich et al., 2008; Stich and Melchinger, 2009). The number of associations was corrected when applying the PK-mixed model: we identified an extra of seven significant associations but also enabled us to reduce associations, indicating that kinship can cause confounding effects on associations in panel 1. We used panel 2, which has a higher variation in tocopherol content and composition, to verify results obtained in the first experiment. The panel 2 accessions were genotyped with CAPS markers which were based on SNPs with significant associations in panel 1. First analyses indicate a validation of the previously detected associations of panel 1. Although association analysis of this panel was based on PK-mixed model, not all associations could be confirmed. This might be explainable by the fact that population structure and kinship relationships of this panel is known to contribute strongly to the phenotype variation for all traits (Wang et al., in preparation). Therefore, future association studies with panel 2 will provide further evidence on whether nucleotide variations in the tocopherol candidate genes detected in this study can explain phenotypic variation.

Several SNPs were found to be associated with more than one trait and of those, some were consistently associated with a trait at both field trial sites, indicating an independent environmental effect of these allelic variations. Two of them, the functional SNPs 35 and 61, were located within exon 1 of *BnaA.PDS1.c* and represent non-synonymous substitutions, whereas there was no evidence of SNP 285 and SNP 543 being functional (both are synonymous) or linked to the functional SNPs (**Table 5**, Figure2). The associations of the synonymous SNPs to the traits indicate that they are linked to other adjacent causative functional polymorphisms in LD distance. Moreover, silent SNPs can be involved in regulatory functions like alteration of mRNA splicing, stability, and structure and therefore, can affect the structure, function, and expression level of proteins (Chamary and Hurst, 2009; Hunt et al., 2009). However, due to the selection of candidate genes based on their properties like high homology to *A. thaliana* genes, mapping position, or function in tocopherol biosynthesis, it was more likely to identify associations with tocopherol traits. We identified several polymorphisms in *BnaA.PDS1.c* and *BnaA.VTE3.a* correlated with tocopherol phenotypes and other seed quality traits. Indeed, both genes encode for enzymes required for tocopherol synthesis, e.g., *PDS1* being responsible for an enzyme which catalyzes the formation of homogentisate, an essential substrate for the formation of the aromatic head group of the tocopherol forms. To obtain DMPBQ, a precursor of γ-tocopherol, the enzyme MPBQ/MSBQ methyltransferase, encoded by the gene *VTE3*, is needed. In consistence with our results, *BnaA.PDS1.c* was mapped on chromosome A10 next to a QTL for tocopherol composition in the Tapidor × Ningyou7 population (Wang et al., in preparation) whereas *BnaA.VTE3.a* was mapped on chromosome A07 in the Mansholt × Samurai population in close proximity to a QTL for α-tocopherol (Endrigkeit, 2007).

In addition to the tocopherol traits, we analyzed seed quality traits in panel 1 and could also find allelic variants that are associated with SOC or OTR. These results may reflect the fat-solubility property of tocopherols and also, the role of tocopherols in the oxidative stability of oil (Jung and Min, 1990; Isbell et al., 1999; Kamal-Eldin, 2006). Interestingly, the principal constituents and the corresponding pathways needed for the biosynthesis of tocopherol and seed oil are not inter-linked (Somerville et al., 2000; DellaPenna and Pogson, 2006). Nonetheless, the relation between fatty acids, natural components of triglycerides, and phospholipids, with tocopherols was demonstrated by some studies (Hasan and Erbas, 2004; Rani et al., 2007; Richards et al., 2008). Moreover, the mapping position of *BnaA.PDS1.c* in the TN-DH population is located within a QTL region for SOC (Qiu et al., 2006) which gives further evidence for our results.

The association between SNPs and tocopherol content/composition could be also of interest for rapeseed breeders. Tocopherols are essential components of human diet and animal feed and hold important functions in plants, such as the protection of lipids and membranes, response to abiotic stress, and oil stability. Hence, they represent an important target for rapeseed breeding. Rapeseed oil contains high amounts of tocopherol and is therefore an important dietary source. So far, the breeding of rapeseed with higher tocopherol is hampered by the ineffective phenotypic selection procedure by HPLC, a destructive, laborious, and costly method. This study provides rapeseed breeders with molecular markers as a tool for the selection of germplasm with higher tocopherol content and quality.

ACKNOWLEDGMENTS

The research was supported by the Deutsche Forschungsgemeinschaft (DFG, JU205/14-1) and the Stiftung Schleswig-Holsteinische Landschaft. We thank Jens Hermann from the Institute of Botany and Monika Zuba from NPZ Lembke Company for their excellent technical assistance in HPLC and NIRS analytics and Monika Bruisch for her support in the greenhouse. We are grateful to the Institute for Clinical Molecular Biology, University Kiel, Germany for performing the Sanger-based sequencing.

REFERENCES

Almeida, J., Quadrana, L., Asís, R., Setta, N., de Godoy, F., Bermúdez, L., Otaiza, S. N., Corrêa da Silva, J. V., Fernie, A. R., Carrari, F., and Rossi, M. (2011). Genetic dissection of vitamin E biosynthesis in tomato. *J. Exp. Bot.* 62, 3781–3798.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Amar, S., Ecke, W., Becker, H. C., and Moellers, C. (2008). QTL for phytosterol and sinapate ester content in *Brassica napus* L. collocate with the two erucic acid genes. *Theor. Appl. Genet.* 116, 1051–1061.

Becker, H. C., Engqvist, G. M., and Karlsson, B. (1995). Comparison of rapeseed cultivars and resynthesized lines based on allozyme and RFLP markers. *Theor. Appl. Genet.* 91, 62–67.

Bergmüller, E., Porfirova, S., and Dörmann, P. (2003). Characterization of an *Arabidopsis* mutant deficient in γ-tocopherolmethyltransferase. *Plant Mol. Biol.* 52, 1181–1190.

Bernardo, R. (1993). Estimation of coefficient of coancestry using molecular markers in maize. *Theor. Appl. Genet.* 85, 1055–1062.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635.

Bus, A., Körber, N., Snowdon, R., and Stich, B. (2011). Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. *Theor. Appl. Genet.* 123, 1413–1423.

Chamary, J. V., and Hurst, L. D. (2009). The price of silent mutations. *Sci. Am.* 300, 46–53.

Chander, S., Guo, Y. Q., Yang, X. H., Yan, J. B., Zhang, Y. R., Song, T. M., and Li, J. S. (2008). Genetic dissection of tocopherol content and composition in maize grain using quantitative trait loci analysis and the candidate gene approach. *Mol. Breed.* 22, 353–365.

Chen, G., Geng, J., Rahman, M., Liu, X., Tu, J., Fu, T., Li, G., McVetty, P. B. E., and Tahir, M. (2010). Identification of QTL for oil content, seed yield, and flowering time in oilseed rape (*Brassica napus*). *Euphytica* 175, 161–174.

Cheng, X. M., Xu, J. S., Xia, S., Gu, J. X., Yang, Y., Fu, J., Qian, X. J., Zhang, S. C., Wu, J. S., and Liu, K. (2009). Development and genetic mapping of microsatellite markers from genome survey sequences in *Brassica napus*. *Theor. Appl. Genet.* 118, 1121–1131.

Collakova, E., and DellaPenna, D. (2003). Homogentisate phytyltransferase activity is limiting for tocopherol biosynthesis in *Arabidopsis*. *Plant Physiol.* 131, 632–642.

Dähnhardt, D., Falk, J., Appel, J., van der Kooij, T. A., Schulz-Friedrich, R., and Krupinska, K. (2002). The hydroxyphenylpyruvate dioxygenase from *Synechocystis* sp. PCC 6803 is not required for plastoquinone biosynthesis. *FEBS Lett.* 523, 177–181.

DellaPenna, D., and Last, R. L. (2006). Progress in the dissection and manipulation of plant vitamin E biosynthesis. *Physiol. Plant* 126, 356–368.

DellaPenna, D., and Pogson, B. (2006). Vitamin synthesis in plants: tocopherols and carotenoids. *Annu. Rev. Plant Biol.* 57, 711–738.

Diers, B. W., and Osborn, T. C. (1994). Genetic diversity of oilseed *Brassica napus* germplasm based on restriction fragment length polymorphisms. *Theor. Appl. Genet.* 88, 662–668.

Durstewitz, G., Polley, A., Plieske, J., Luerssen, H., Graner, E. M., Wieseke, R., and Ganal, M. W. (2010). SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. *Genome* 53, 948–956.

Dwiyanti, M., Yamada, T., Sato, M., Abe, J., and Kitamura, K. (2011). Genetic variation of gamma-tocopherol methyltransferase gene contributes to elevated alpha-tocopherol content in soybean seeds. *BMC Plant Biol.* 11, 152. doi:10.1186/1471-2229-11-152

Ecke, W., Clemens, R., Honsdorf, N., and Becker, H. C. (2010). Extent and structure of linkage disequilibrium in canola quality winter rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* 120, 921–931.

Endrigkeit, J. (2007). *Identifikation und Charakterisierung von Genen der Tocopherol-Biosynthese aus Raps (Brassica napus L.)*. Dissertation, Christian-Albrechts-University, Kiel.

Endrigkeit, J., Wang, X. X., Cai, D. G., Zhang, C. Y., Long, Y., Meng, J. L., and Jung, C. (2009). Genetic mapping, cloning, and functional characterization of the BnaX.VTE4 gene encoding a gamma-tocopherol methyltransferase from oilseed rape. *Theor. Appl. Genet.* 119, 567–575.

Falk, J., Andersen, G., Kernebeck, B., and Krupinska, K. (2003). Constitutive overexpression of barley 4-hydroxyphenylpyruvate dioxygenase in tobacco results in elevation of the vitamin E content in seeds but not in leaves. *FEBS Lett.* 540, 35–40.

Flint-Garcia, S. A., Thuillet, A. C., Yu, J. M., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064.

Friedt, W., and Snowdon, R. J. (2009). "Oilseed rape," in *Handbook of Plant Breeding Vol. 4: Oil Crops*, eds J. Vollmann and I. Rajcan (New York: Springer), 91–126.

Fusari, C. M., Lia, V. V., Hopp, H. E., Heinz, R. A., and Paniego, N. B. (2008). Identification of single nucleotide polymorphisms and analysis of linkage disequilibrium in sunflower elite inbred lines using the candidate gene approach. *BMC Plant Biol.* 8, 7. doi:10.1186/1471-2229-8-7

Ganal, M. W., Altmann, T., and Roder, M. S. (2009). SNP identification in crop plants. *Curr. Opin. Plant Biol.* 12, 211–217.

Goffman, F. D., and Becker, H. C. (1999). "Inheritance of tocopherol contents in seeds of rapeseed (*Brassica napus* L.)," in *Proceedings of the 10th Rapeseed Congress*, Canberra.

Goffman, F. D., and Becker, H. C. (2002). Genetic variation of tocopherol content in a germplasm collection of *Brassica napus* L. *Euphytica* 125, 189–196.

Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., and Buckler, E. S. (2009). A first-generation haplotype map of maize. *Science* 326, 1115–1117.

Goujon, M., McWilliam, H., Li, W. Z., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 38, W695–W699.

Grusak, M. A., and DellaPenna, D. (1999). Improving the nutrient composition of plants to enhance human nutrition and health. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 133–161.

Gupta, P., Rustgi, S., and Kulwal, P. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* 57, 461–485.

Haddadi, P., Ebrahimi, A., Langlade, N., Yazdi-samadi, B., Berger, M., Calmon, A., Naghavi, M., Vincourt, P., and Sarrafi, A. (2011). Genetic dissection of tocopherol and phytosterol in recombinant inbred lines of sunflower through quantitative trait locus analysis and the candidate gene approach. *Mol. Breed.* 29, 1–13.

Hall, D., Tegström, C., and Ingvarsson, P. K. (2010). Using association mapping to dissect the genetic basis of complex traits in plants. *Brief. Funct. Genomics* 9, 157–165.

Hamama, A., and Bhardwaj, H. (2011). Characterization of total and individual sterols in canola sprouts. *J. Am. Oil Chem. Soc.* 88, 361–366.

Hasan, B., and Erbas, S. (2004). Influence of seed development and seed position on oil, fatty acids and total tocopherol contents in sunflower (*Helianthus annuus* L.). *Turk. J. Agr. Forest.* 29, 179–186.

Hasan, M., Friedt, W., Pons-Kühnemann, J., Freitag, N., Link, K., and Snowdon, R. J. (2008). Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (*Brassica napus* ssp. *napus*). *Theor. Appl. Genet.* 116, 1035–1049.

Hasan, M., Seyis, F., Badani, A., Pons-Kühnemann, J., Friedt, W., Lühs, W., and Snowdon, R. (2006). Analysis of genetic diversity in the *Brassica napus* L. gene pool using SSR markers. *Genet. Resour. Crop Evol.* 53, 793–802.

Heuertz, M., De Paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., and Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* 174, 2095–2105.

Honsdorf, N., Becker, H. C., and Ecke, W. (2010). Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (*Brassica napus* L.). *Genome* 53, 899–907.

Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M., and Kimchi-Sarfaty, C. (2009). Silent (synonymous) SNPs: should we care about them? *Methods Mol. Biol.* 578, 23–39.

Isbell, T. A., Abbott, T. P., and Carlson, K. D. (1999). Oxidative stability index of vegetable oils in binary mixtures with meadowfoam oil. *Ind. Crops Prod.* 9, 115–123.

Jestin, C., Lodé, M., Vallée, P., Domin, C., Falentin, C., Horvais, R., Coedel, S., Manzanares-Dauleux, M., and Delourme, R. (2011). Association mapping of quantitative resistance

for Leptosphaeria maculans in oilseed rape (*Brassica napus* L.). *Mol. Breed.* 27, 271–287.

Jung, M., Ching, A., Bhattramakki, D., Dolan, M., Tingey, S., Morgante, M., and Rafalski, A. (2004). Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. *Theor. Appl. Genet.* 109, 681–689.

Jung, M. Y., and Min, D. B. (1990). Effects of α-, γ-, and δ-tocopherols on oxidative stability of soybean oil. *J. Food Sci.* 55, 1464–1465.

Kamal-Eldin, A. (2006). Effect of fatty acids and tocopherols on the oxidative stability of vegetable oils. *Eur. J. Lipid Sci. Technol.* 108, 1051–1061.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723.

Krutovsky, K. V., and Neale, D. B. (2005). Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas Fir. *Genetics* 171, 2029–2041.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948.

Li, H. Y., Liu, H. C., Han, Y. P., Wu, X. X., Teng, W. L., Liu, G. F., and Li, W. B. (2010). Identification of QTL underlying vitamin E contents in soybean seed among multiple environments. *Theor. Appl. Genet.* 120, 1405–1413.

Long, Y., Shi, J., Qiu, D., Li, R., Zhang, C., Wang, J., Hou, J., Zhao, J., Shi, L., Park, B. S., Choi, S. R., Lim, Y. P., and Meng, J. (2007). Flowering time quantitative trait loci analysis of oilseed Brassica in multiple environments and genomewide alignment with *Arabidopsis*. *Genetics* 177, 2433–2444.

Magee, L. (1990). R^2 measures based on wald and likelihood ratio joint significance tests. *Am. Stat.* 44, 250–253.

Marwede, V., Gul, M. K., Becker, H. C., and Ecke, W. (2005). Mapping of QTL controlling tocopherol content in winter oilseed rape. *Plant Breed.* 124, 20–26.

Marwede, V., Schierholt, A., Christian, M., and Becker, H. C. (2004). Genotype x environment interactions and heritability of tocopherol contents in canola. *Crop Sci.* 44, 728–731.

Mei, D. S., Li, Y. C., Wang, H. Z., Hu, Q., Li, Y. D., and Xu, Y. S. (2009). QTL analysis on plant height and flowering time in *Brassica napus*. *Plant Breed.* 128, 458–465.

Mène-Saffrané, L., and DellaPenna, D. (2010). Biosynthesis, regulation and functions of tocochromanols in plants. *Plant Physiol. Biochem.* 48, 301–309.

Munné-Bosch, S., and Alegre, L. (2002). The function of tocopherols and tocotrienols in plants. *CRC Crit. Rev. Plant Sci.* 21, 31–57.

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., and Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21, 2194–2202.

Norris, S. R., Shen, X., and DellaPenna, D. (1998). Complementation of the *Arabidopsis* pds1 mutation with the gene encoding p-hydroxyphenylpyruvate dioxygenase. *Plant Physiol.* 117, 1317–1323.

Öhrvall, M., Vessby, B., and Sundlöf, G. (1996). Gamma, but not alpha, tocopherol levels in serum are reduced in coronary heart disease patients. *J. Intern. Med.* 239, 111–117.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinic traits. *Biometrics* 43, 487–498.

Pongracz, G., Weiser, H., and Matziger, D. (1995). Tocopherole-Antioxidantien der Natur. *Eur. J. Lipid Sci. Technol.* 3, 90–104.

Porfirova, S., Bergmüller, E., Tropf, S., Lemke, R., and Dörmann, P. (2002). Isolation of an *Arabidopsis mutant* lacking vitamin E and identification of a cyclase essential for all tocopherol biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12495–12500.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Qiu, D., Morgan, C., Shi, J., Long, Y., Liu, J., Li, R., Zhuang, X., Wang, Y., Tan, X., Dietrich, E., Weihmann, T., Everett, C., Vanstraelen, S., Beckett, P., Fraser, F., Trick, M., Barnes, S., Wilmer, J., Schmidt, R., Li, J., Li, D., Meng, J., and Bancroft, I. (2006). A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. *Theor. Appl. Genet.* 114, 67–80.

Radoev, M., Becker, H. C., and Ecke, W. (2008). Genetic analysis of heterosis for yield and yield components in rapeseed (*Brassica napus* L.) by QTL mapping. *Genetics* 179, 1547–1558.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5, 94–100.

Rafalski, J. A. (2010). Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13, 174–180.

Rani, A., Kumar, V., Verma, S., Shakya, A., and Chauhan, G. (2007). Tocopherol content and profile of soybean: genotypic variability and correlation studies. *J. Am. Oil Chem. Soc.* 84, 377–383.

Reale, S., Doveri, S., Diaz, A., Angiolillo, A., Lucentini, L., Pilla, F., Martin, A., Donini, P., and Lee, D. (2006). SNP-based markers for discriminating olive (Olea europaea L.) cultivars. *Genome* 49, 1193–1205.

Rezaeizad, A., Wittkop, B., Snowdon, R., Hasan, M., Mohammadi, V., Zali, A., and Friedt, W. (2011). Identification of QTLs for phenolic compounds in oilseed rape (*Brassica napus* L.) by association mapping using SSR markers. *Euphytica* 177, 335–342.

Richards, A., Wijesundera, C., and Salisbury, P. (2008). Genotype and growing environment effects on the tocopherols and fatty acids of *Brassica napus* and *B. juncea*. *J. Am. Oil Chem. Soc.* 85, 159–168.

Rücker, B., and Röbbelen, G. (1996). Impact of low linolenic acid content on seed yield of winter oilseed rape (*Brassica napus* L.). *Plant Breed.* 115, 226–230.

Schierholt, A., and Becker, H. C. (2001). Environmental variability and heritability of high oleic acid content in winter oilseed rape. *Plant Breed.* 120, 63–66.

Schledz, M., Seidler, A., Beyer, P., and Neuhaus, G. (2001). A novel phytyltransferase from *Synechocystis* sp. PCC 6803 involved in tocopherol biosynthesis. *FEBS Lett.* 499, 15–20.

Schneider, C. (2005). Chemistry and biology of vitamin E. *Mol. Nutr. Food Res.* 49, 7–30.

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* 18, 233–234.

Schuelke, M., Mayatepek, E., Inter, M., Becker, M., Pfeiffer, E., Speer, A., Hübner, C., and Finckh, B. (1999). Treatment of ataxia in isolated vitamin E deficiency caused by [alpha]-tocopherol transfer protein deficiency. *J. Pediatr.* 134, 240–244.

Sheehy, P. J. A., Bramley, P. M., Elmadfa, I., Kafatos, A., Kelly, F. J., Manios, Y., Roxborough, H. E., Schuch, W., and Wagner, K. H. (2000). Vitamin E. *J. Sci. Food Agric.* 80, 913–938.

Shewmaker, C. K., Sheehy, J. A., Daley, M., Colburn, S., and Ke, D. Y. (1999). Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. *Plant J.* 20, 401–412.

Smooker, A. M., Wells, R., Morgan, C., Beaudoin, F., Cho, K., Fraser, F., and Bancroft, I. (2011). The identification and mapping of candidate genes and QTL involved in the fatty acid desaturation pathway in *Brassica napus*. *Theor. Appl. Genet.* 122, 1075–1090.

Snowdon, R., Lühs, W., and Friedt, W. (2006). "Oilseed rape," in *Genome Mapping and Molecular Breeding, Vol. 2, Oilseeds*, ed. C. Kole (Heidelberg: Springer).

Soll, J., Kemmerling, M., and Schultz, G. (1980). Tocopherol and plastoquinone synthesis in spinach chloroplasts subfractions. *Arch. Biochem. Biophys.* 204, 544–550.

Somerville, C., Browse, J., Jaworski, J. G., and Ohlrogge, J. B. (2000). "Lipids," in *Biochemistry and Molecular Biology of Plants*, eds B. B. Buchanan, W. Gruissem, and R. L. Jones (Rockville: American Society of Plant Physiologists), 456–527.

Stich, B., Maurer, H. P., Melchinger, A. E., Frisch, M., Heckenberger, M., van der Voort, J. R., Peleman, J., Sorensen, A. P., and Reif, J. C. (2006). Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. *Mol. Breed.* 17, 217–226.

Stich, B., and Melchinger, A. E. (2009). Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and *Arabidopsis*. *BMC Genomics* 10, 94. doi:10.1186/1471-2164-10-94

Stich, B., Mohring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S., and Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics* 178, 1745–1754.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* 28, 286–289.

Trick, M., Long, Y., Meng, J., and Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol. J.* 7, 334–346.

Valentin, H. E., Lincoln, K., Moshiri, F., Jensen, P. K., Qi, Q., Venkatesh, T. V., Karunanandaa, B., Baszis, S. R., Norris, S. R., Savidge, B.,

Gruys, K. J., and Last, R. L. (2006). The *Arabidopsis* vitamin E pathway gene5-1 mutant reveals a critical role for phytol kinase in seed tocopherol biosynthesis. *Plant Cell* 18, 212–224.

Van Eenennaam, A., Lincoln, K., Durrett, T., Valentin, H., Shewmaker, C., Thorne, G., Jiang, J., Baszis, S., Levering, C., Aasen, E., Hao, M., Stein, J., Norris, S., and Last, R. (2003). Engineering vitamin E content: from *Arabidopsis mutant* to soy oil. *Plant Cell* 15, 3007–3019.

Wang, N., Qian, W., Suppanz, I., Wei, L., Mao, B., Long, Y., Meng, J., Müller, A. E., and Jung, C. (2011). Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the FRIGIDA homologue BnaA.FRI.a. *J. Exp. Bot.* doi: 10.1093/jxb/err249

Wei, S., Yu, B., Gruber, M. Y., Khachatourians, G. G., Hegedus, D. D., and Hannoufa, A. (2010). Enhanced seed carotenoid levels and branching in transgenic *Brassica napus* expressing the *Arabidopsis* miR156b gene. *J. Agric. Food Chem.* 58, 9572–9578.

Westermeier, P., Wenzel, G., and Mohler, V. (2009). Development and evaluation of single-nucleotide polymorphism markers in allotetraploid rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* 119, 1301–1311.

Wittkop, B., Snowdon, R. J., and Friedt, W. (2009). Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe. *Euphytica* 170, 131–140.

Witztum, J. L. (1993). Role of oxidised low density lipoprotein in atherogenesis. *Br. Heart J.* 69, S12–S18.

Wong, J. C., Lambert, R. J., Tadmor, Y., and Rocheford, T. R. (2003). QTL associated with accumulation of tocopherols in maize. *Crop Sci.* 43, 2257–2266.

Würschum, T., Liu, W., Maurer, H., Abel, S., and Reif, J. (2012). Dissecting the genetic architecture of agronomic traits in multiple segregating populations in rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* 124, 153–161.

Yin, X., Yi, B., Chen, W., Zhang, W., Tu, J., Fernando, W., and Fu, T. (2010). Mapping of QTLs detected in a *Brassica napus* DH population for resistance to *Sclerotinia sclerotiorum* in multiple environments. *Euphytica* 173, 25–35.

Yu, B., Lydiate, D., Young, L., Schäfer, U., and Hannoufa, A. (2008a). Enhancing the carotenoid content of *Brassica napus* seeds by downregulating lycopene epsilon cyclase. *Transgenic Res.* 17, 573–585.

Yu, J. M., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008b). Genetic design and statistical power of nested association mapping in maize. *Genetics* 178, 539–551.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208.

Zhang, H., Shi, C., Wu, J., Ren, Y., Li, C., Zhang, D., and Zhang, Y. (2004). Analysis of genetic effects and heritabilities for linoleic and α-linolenic acid content of *Brassica napus* L. across Chinese environments. *Eur. J. Lipid Sci. Technol.* 106, 518–523.

Zhang, L., Yang, G., Liu, P., Hong, D., Li, S., and He, Q. (2011). Genetic and correlation analysis of siliquetraits in *Brassica napus* L. by quantitative trait locus mapping. *Theor. Appl. Genet.* 122, 21–31.

Zhao, J., Dimov, Z., Becker, H. C., Ecke, W., and Moellers, C. (2008). Mapping QTL controlling fatty acid composition in a doubled haploid rapeseed population segregating for oil content. *Mol. Breed.* 21, 115–125.

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1, 5–20.

Zhu, Y. L., Song, Q. J., Hyten, D. L., Van Tassell, C. P., Matukumalli, L. K., Grimm, D. R., Hyatt, S. M., Fickus, E. W., Young, N. D., and Cregan, P. B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics* 163, 1123–1134.

Zou, J., Jiang, C. C., Cao, Z. Y., Li, R. Y., Long, Y., Chen, S., and Meng, J. L. (2010). Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. *Genome* 53, 908–916.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 March 2012; accepted: 30 May 2012; published online: 26 June 2012.

Citation: Fritsche S, Wang X, Li J, Stich B, Kopisch-Obuch FJ, Endrigkeit J, Leckband G, Dreyer F, Friedt W, Meng J and Jung C (2012) A candidate gene-based association study of tocopherol content and composition in rapeseed (Brassica napus). Front. Plant Sci. ***3****:129. doi: 10.3389/fpls.2012.00129*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

APPENDIX

Table A1 | *Brassica napus* accessions which were used for field trials at Giessen and Holtsee (Germany) 2007/08.

Accession name	Geographical origin	Type
Abukuma natane	Japan	Winter
Akela	Germany	Winter
Anja	Germany	Winter
Aphid resistant rape	New Zealand	Winter
Askari	Germany	Winter
Baltia	Soviet Union	Winter
Belinda	Germany	Winter
Bienvenue	France	Winter
Binera	Germany	Winter
Bladkool	Netherlands	Winter
Brauner Schnittkohl	Germany	Winter
Brink	Sweden	Winter
Ceres	Germany	Winter
Chuosenshu	Korea	Winter
Coriander	Germany	Winter
Darmor	France	Winter
Diamant	Germany	Winter
Dippes	Germany	Winter
Doral	Germany	Winter
Dwarf Essex	United Kingdom	Winter
Edita	Germany	Winter
Elsoms giant	United Kingdom	Winter
Emerald	Germany	Winter
English giant	United Kingdom	Winter
Erra	Germany	Winter
Fertoedi	Hungary	Winter
Fora	Sweden	Winter
Gogatsuna	Japan	Spring
Goldgelber Zarter Butter	Germany	Winter
Groene Groninger Snijmoes	Netherlands	Winter
Gross-Luesewitzer Spaets	unknown	Winter
Hektor	Sweden	Winter
Hokkai 3-go	Japan	Winter
Janetzkis Schlesischer	Austria	Winter
Janpol	Poland	Winter
Jet Neuf	France	Winter
Jupiter	Sweden	Winter
Krapphauser	unknown	Winter
Kromerska	Czech Republic	Winter
Librador	Germany	Winter
Libritta	Germany	Winter
Liglory (Gelbsamige)	Germany	Winter
Limburgse Bladkool	Netherlands	Winter
Lingot, R-26	France	Winter
Liporta	Germany	Winter
Lirafit, WRF 22	Germany	Winter
Liragrün	Germany	Winter
Lirakotta	Germany	Winter
Madora	Germany	Winter
MAH 1, Jantar	Poland	Winter
Major	France	Winter
Mansholts Hamb. Raps	Germany	Winter
Marasaki natane	Japan	Winter
Markus	France	Winter
Matador	Sweden	Winter
Mestnij	Soviet Union	Winter
Moana, Moana Rape	New Zealand	Winter
Mytnickij	Ukraine	Winter
Nemertschanskij 1	Soviet Union	Winter
Norde	Sweden	Winter
Norin	unknown	Winter
Nunsdale	United Kingdom	Winter
Oleander	Germany	Winter
Ölquell, Gülzower	Germany	Winter
Palu	Italy	Winter
Panter	Sweden	Winter
Parapluie	France	Winter
POH 285, Bolko	Poland	Winter
Primor, P-R	France	Winter
Quedlinburger Platzfester	Germany	Winter
Quinta	Germany	Winter
Rafal, R 40	France	Winter
Ramses	France	Winter
Rapol	Germany	Winter
Regal	Sweden	Winter
Samo	Sweden	Winter
Sarepta	France	Winter
Siberische Boerenkool	unknown	Winter
Silona	Sweden	Winter
Skrzeszowicki	Poland	Winter
Skziverskij	Soviet Union	Winter
Slovenska Krajova	Czech Republic	Winter
Sobotkowski	Poland	Winter
Sonnengold	Germany	Winter
Start	Poland	Winter
Taisetsu	Japan	Winter
Trebicska	Czech Republic	Winter
Victor	Sweden	Winter
Vinnickij 15/59	Soviet Union	Winter
Winfred	Germany	Winter
Express	Germany	Winter
Tapidor[1]	France	Winter
Ningyou7[1]	Viet-Nam	Spring
Ramon[2]	Netherlands	Winter
Ridana[2]	Germany	Winter
Wolynski[2]	Soviet Union	Winter

They are a part of the core collection of the EU project RESGEN CT99 109-112.

[1]*Accession grown only in Holtsee.*

[2]*Accession grown only in Giessen.*

Table A2 | Primer used for the amplification of tocopherol genes.

Gene	Genbank nr	Primer sequence (5′ → 3′)
BnaX.PDS1.a	JN834026	F: ACTCGGAATGCGATTCTCCGCT R: GTAAACCTTCCCTTCCTCATCC
BnaX.PDS1.b	JN834015	F: CTCAGCATCTAATCAACGTAGCT R: CCCTCCTATCGTCCTAAACGAC
BnaA.PDS1.c	JN834016	F1: AACTCTATGGGGCACGAAAA R1: TCTGCGTCTTCTACTTCAAC F2: CCCTCCTATCGTCCTAAACGAC R2: CATCCCTCCACTCTGGTAAACTC
BnaX.VTE1.a	JN834017	F: CTGAAGAGACCGTTTGAGTAGC R: TTTTCCTGTATTTGAGCCTCAT
BnaX.VTE1.b	JN834018	F: AACCAGAAGTTAGTTCATGGC R: CCTCAGTGGACTAGCTAAGC
BnaX.VTE2.b	JN834020	F: CATCTTCACTAGTGCTTAGATC R: GCCTTAATCACTCAAAACGAGG
BnaX.VTE3.a	JN834021	F: GGTTCATCCAGCACAAGAAA R: CCTCTTCCTTTGGTCCAAGCTA
BnaX.VTE3.b	JN834022	F: GGTTTCCCCGCTTCCAATCT R: CTAAACCCAATCACACACTCTGA
BnaX.VTE4.b	JN834023	F: CATTCAGTCTTCGTTGTGCAAT R: CCCTTCAATCATCAATGG

Table A3 | Polymorphisms within tocopherol candidate genes and enzymes, which were used as allele-specific markers for genotyping the 133 *B. napus* accessions of panel 2.

Gene	Polymorphism	Position in gene	Enzyme
BnaA.PDS1.c	C/T[1]	543	*Bse*DI
BnaX.VTE3.a	T/C	285	*Hin*1II
BnaX.VTE3.a	C/T	342	*Sal*I
BnaX.VTE3.b	CCGG/–	326–329	*Kpn*2I
BnaX.VTE2.b	G/A	1464	*Bbv*I

[1] *Letter preceding the slash points the nucleotides of the Tapidor allele.*

Table A4 | SNP densities with in *A. thaliana* tocopherol genes by using POLYMORPH (Fitz, J. et al., personal communication).

Gene	Gene ID	Genomic gene length (bp)	Chromosome	No. SNPs	SNP density
VTE1	AT4G32770	3069	4	10	1/307 bp
VTE2	AT2G18950	3122	2	5	1/624 bp
VTE3	AT3G63410	1530	3	9	1/170 bp
VTE4	AT1G64970	2025	1	68	1/30 bp
VTE5	AT5G04490	1910	5	23	1/83 bp
PDS1	AT1G06570	1661	1	4	1/415 bp
ACT2	AT3G18780	2333	3	12	1/194 bp
ACT8	AT1G49240	2441	1	12	1/203 bp
LOV1	AT1G10920	3348	1	16	1/209 bp
RPP1	AT3G44480	5795	3	207	1/28 bp

The database compares the genomic sequences of 80 A. thaliana accessions. For density calculation only SNPs with a frequency of >5% were considered and then related to the genomic length. SNP densities of random chosen actin genes (ACT1, ACT8) as well as loci, involved in defense response (RPP1, LOV1), were also calculated.

Table A5 | Number of polymorphisms within the amplified gene regions of tocopherol candidate genes from *B. napus*, their properties and their frequency in panel 1 accessions.

Gene	Base pairs sequenced (bp)	Sequence variations				Frequency (in %)[1]
		Total	Intron	Silent	Non-silent	
BnaX.PDS1.a	920	20	0	16	4	1.06
BnaX.PDS1.b	598	0	0	0	0	0
BnaA.PDS1.c	304	10	0	7	3	52.13
	729	18	15	37	0	29.16
BnaX.VTE1.a	860	0	0	0	0	0
BnaX.VTE1.b	1028	0	0	0	0	0
BnaX.VTE2.b	780	1	0	0	1	4.16
BnaX.VTE3.a	753	6	0	6	0	17.77
BnaX.VTE3.b	478	1	0	0	1	1.04
BnaX.VTE4.b	494	0	0	0	0	0

[1] *Percentage of accessions different from reference genotype Tapidor.*

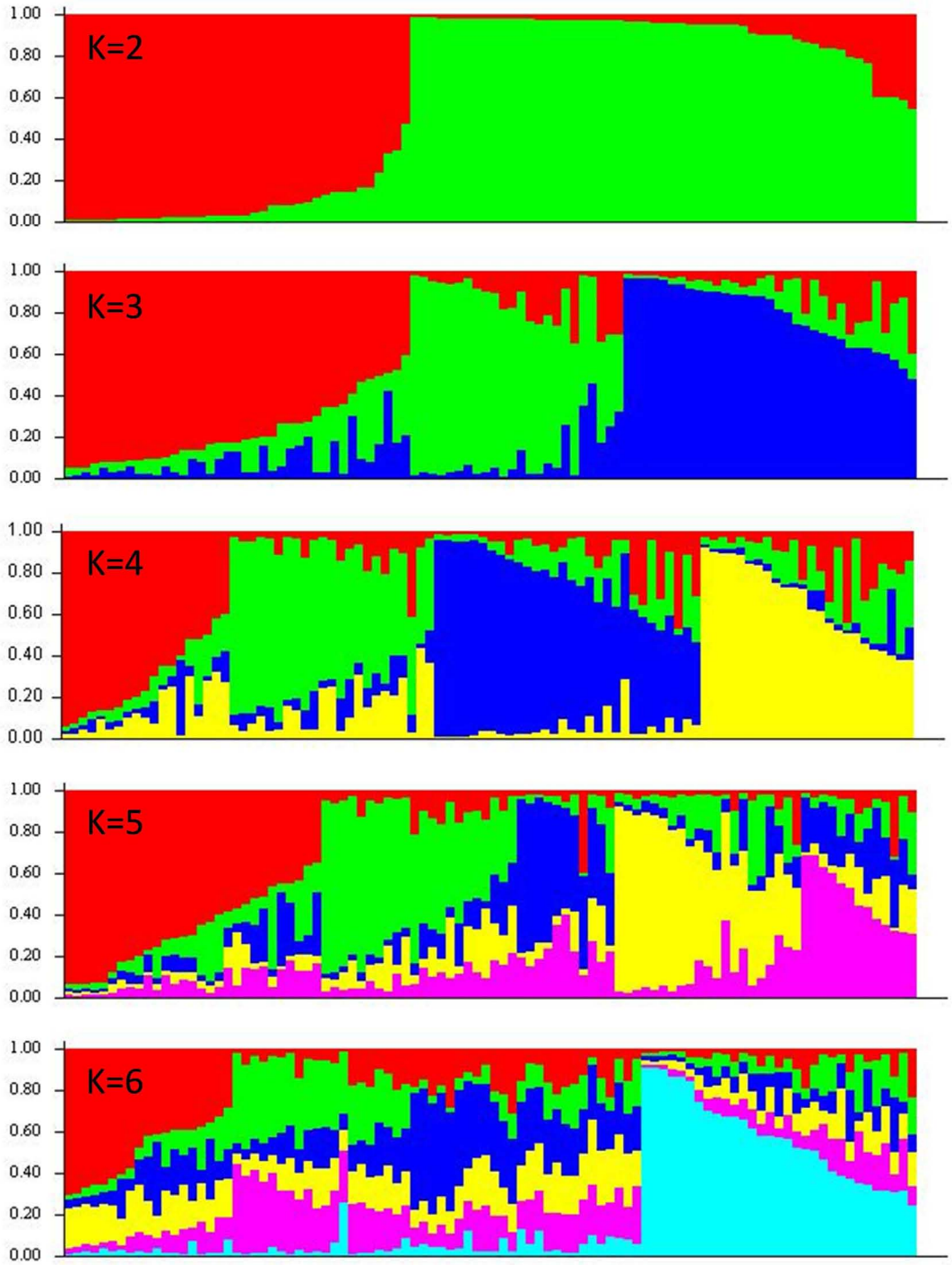


FIGURE A1 | Presentation of the population structure of 96 *B. napus* accessions of panel 1 under the assumption of subpopulation *K* = 2–6 which calculation based on 24 SSR markers.

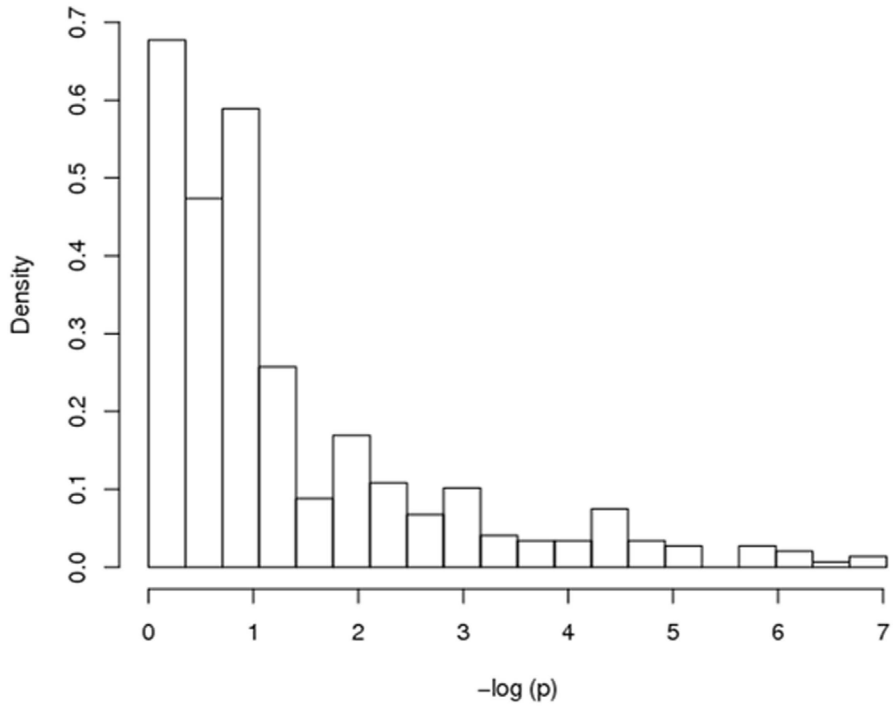


FIGURE A2 | Distribution of *p*-values of the tocopherol traits.

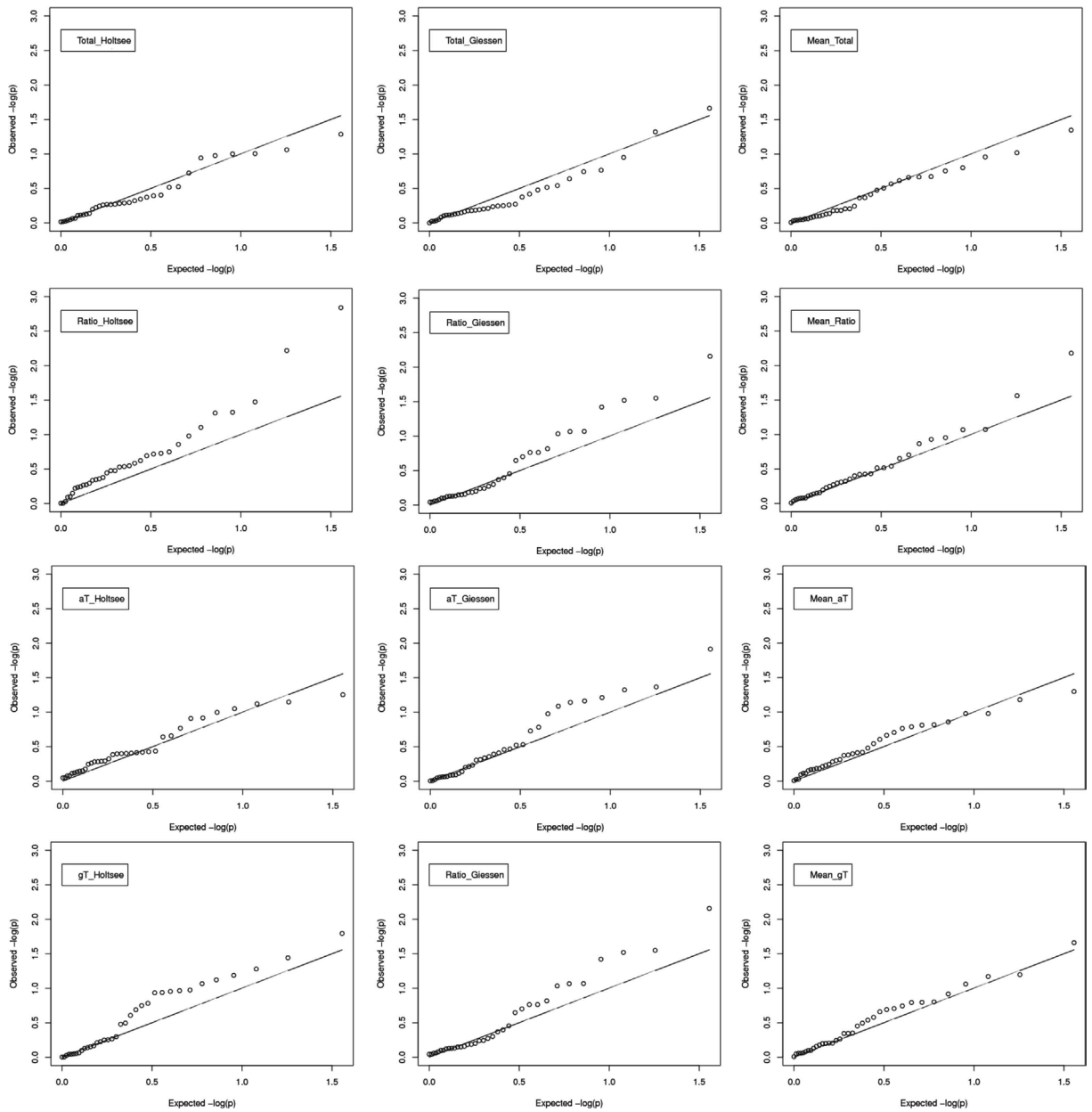


FIGURE A3 | QQ-plot shows the distribution of the *p*-values of the associations of the SSRs with the phenotypic traits.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 01 June 2012
doi: 10.3389/fpls.2012.00118

DNA-based genetic markers for rapid cycling *Brassica rapa* (Fast Plants type) designed for the teaching laboratory

Eryn E. Slankster, Jillian M. Chase, Lauren A. Jones and Douglas L. Wendell*

Department of Biological Sciences, Oakland University, Rochester, MI, USA

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Dr. Sureshkumar Balasubramanian, Monash University, Australia
Xiyin Wang, Hebei United University, China
Guusje Bonnema, Wageningen University, Netherlands

****Correspondence:***
Douglas L. Wendell, Department of Biological Sciences, Oakland University, 2200 N Squirrel Road, Rochester, MI 48309-4401, USA.
e-mail: wendell@oakland.edu

We have developed DNA-based genetic markers for rapid cycling *Brassica rapa* (RCBr), also known as Fast Plants. Although markers for *B. rapa* already exist, ours were intentionally designed for use in a teaching laboratory environment. The qualities we selected for were robust amplification in PCR, polymorphism in RCBr strains, and alleles that can be easily resolved in simple agarose slab gels. We have developed two single nucleotide polymorphism (SNP) based markers and 14 variable number tandem repeat (VNTR)-type markers spread over four chromosomes. The DNA sequences of these markers represent variation in a wide range of genomic features. Among the VNTR-type markers, there are examples of variation in a non-genic region, variation within an intron, and variation in the coding sequence of a gene. Among the SNP-based markers there are examples of polymorphism in intronic DNA and synonymous substitution in a coding sequence. Thus these markers can serve laboratory exercises in both transmission genetics and molecular biology.

Keywords: Fast Plants, rapid cycling *Brassica rapa*, marker, SNP, education, DNA fingerprinting, genetic mapping

INTRODUCTION

Rapid cycling *Brassica rapa* (RCBr), also known as Fast Plants, are a widely used model organism in biology education. They were developed by selection of *B. rapa* for the traits of short time to flowering, rapid seed maturation, lack of seed dormancy, petite growth habit, and high female fertility (Williams and Hill, 1986). The result is a plant with a 7-week generation time that can be cultivated inexpensively by novices. Their fast growth occurs at room temperature under continuous illumination by household fluorescent lights with simple and inexpensive growing materials (Williams, 1997). In addition to the plant strains, the Wisconsin Fast Plants Program[1] has developed a large assortment of educational activities and support materials. Topics covered by RCBr activities include the effect of environment on plant growth, plant-herbivore interactions, hormones and growth, and genetics (Musgrave, 2000). Seed stocks and instructional kits are available from Carolina Biological Supply (Burlington, NC, USA). Seeds are also available from the Crucifer Genetics Cooperative (Williams, 1985).

Rapid cycling *B. rapa* is an excellent organism for teaching genetics. Cross-pollination is easy for students at all levels because like other *Brassica*, and unlike *Arabidopsis*, they are self-incompatible for pollination. Lessons in Mendelian inheritance are performed using RCBr stocks that vary in easily scored phenotypes such as stem color (purple versus non-purple) and leaf color (green versus yellow-green; Williams, 1985). There are also traits with complex inheritance such as trichome density which shows additive polygenic inheritance (Lauffer and Fall, 2000) and intensity of anthocyanin pigmentation which is both polygenic and affected by environment (Goldman, 1999). Some molecular genetic markers exist, but have been slower to develop (Wendell and Pickard, 2007).

Although there is a large set of DNA markers for *B. rapa* in the form of microsatellites and single nucleotide polymorphisms (SNP), they do not lend themselves well to the teaching laboratory where simple agarose slab gels are most common, time and budgets are limited, and the students using them are novices. Microsatellites are highly desirable genetic markers because they tend to have multiple alleles and thus be highly informative (Litt and Luty, 1989; Weber and May, 1989). An extensive list of microsatellite markers for *Brassica* developed by several groups can be found at the Microsatellite Information Exchange[2], and microsatellite markers that have been developed for *Brassica* crop species are usable and polymorphic in RCBr (Burdzinski and Wendell, 2007; Iniguez-Luy et al., 2009). Instructional use is difficult because the size difference between alleles is usually in the range of 2–20 base pairs which is best resolved in polyacrylamide gels. We have previously reported a set of selected microsatellites and protocols to make them work in a teaching laboratory environment using polyacrylamide mini gels (Wendell and Pickard, 2007). However, the need for polyacrylamide gels still creates a barrier to their use by instructors of undergraduate or advanced high school laboratories who either may not have the needed equipment in a teaching laboratory, or do not wish to work with polyacrylamide. Another type of DNA marker available for *B. rapa* are SNP (Park et al., 2009). SNPs have grown in significance as genetic markers because they are present at a high density in genomes and SNP genotype data can be collected using automated high throughput methods such as microarrays. Although a single SNP is not as informative as a single polymorphic microsatellite, due to SNPs generally having

[1] www.fastplants.org

[2] http://www.brassica.info/resource/markers/ssr-exchange.php

only two alleles, a string of SNPs can be just as informative as a single microsatellite with multiple alleles (Kruglyak, 1997). However, the methods used to routinely analyze SNPs such as microarrays or automated DNA sequencing cannot be expected to be readily available in an instructional laboratory. Even simpler methods such as TaqMan assays still require more sophisticated equipment (for real time PCR) than most instructional laboratories have on hand.

In order to allow the use of DNA makers with RCBr under the conditions where these plants are most commonly used, in an undergraduate or advanced high school teaching lab, we have developed genetic markers specifically suited to use under simple conditions. The markers we report here have been selected for robust and reliable amplification by PCR, polymorphism in RCBr populations, and alleles that can be readily resolved in small conventional agarose slab gels. For repetitive DNA-based markers we have sought those with longer repeated element like variable number tandem repeats (VNTR) markers, rather than microsatellites, so that the size difference between alleles would be larger. For detection of SNPs, we have identified those that are reliably detected by the technique of PCR-RFLP (Konieczny and Ausubel, 1993).

MATERIALS AND METHODS

PLANT STRAINS

A variety of RCBr strains were used which vary in both Mendelian traits and DNA markers. One of the Mendelian loci is *anthocyaninless* which has the recessive *anl* allele for lack of anthocyanin pigment (non-purple stem) and the dominant wild type *ANL* allele that allows anthocyanin production resulting in purple stems (Williams, 2007). The other Mendelian locus used is *yellow-green* which has the recessive *ygr* allele for yellow-green color and the dominant wild type *YGR* allele for normal green color (Williams, 2007). The Wisconsin Fast Plants strains Standard *B. rapa*; Purple Stem, Hairy; Non-Purple Stem, Hairless; and Non-Purple Stem, Yellow-Green Leaf were obtained from Carolina Biological Supply Company (Burlington, NC, USA). Strain DWRCBr70 is derived from a purple-stemmed RCBr population by selection for high intensity of purple color. DWRCBr52 was derived from a non-purple stem population by selection for low trichome density. DWRCBr60 was derived from a populations of plants with purple stems and yellow-green leaves.

DNA PURIFICATION

DNA was purified from leaf tissue using a DNeasy Plant Mini Kit (Qiagen Inc., Valencia, CA, USA) following manufacturer's instructions with the exception that the tissue was disrupted using a ground glass homogenizer.

PCR

PCR was performed in a 10-μl reaction volume with 50 ng of template DNA and 10 pmol of each primer using either Accuprime Taq DNA polymerase and supplied Buffer I (Invitrogen, Carlsbad, CA, USA) or Syzygy Taq polymerase (Syzygy Biotech, Grand Rapids, MI, USA). The PCR program was an initial incubation at 94˚C for 2 min, followed by 25 cycles of 94˚C for 30 s, 61˚C for 60 s, and 72˚C for one 60 s, and a final incubation at 72˚C for 4 min.

DNA SEQUENCING

PCR amplicons were purified for DNA sequencing using a MinElute PCR Cleanup Kit (Qiagen Inc., Valencia, CA, USA) and their purity verified by analytical electrophoresis in an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Sequencing reactions were performed using ABI BigDye Terminator v3.1 Cycle Sequencing Kit and analyzed using the Applied Biosystems ABI Prism 3730 DNA Analyzer at the Wayne State University Applied Genomics Technology Center[3].

ELECTROPHORESIS

PCR products were separated by electrophoresis in 1.2% agarose (Genetic Analysis Grade from Fisher Scientific, Waltham, MA, USA) in 7 cm long slab gels at 150 V for 30–50 min. Bands were detected by ethidium bromide staining.

IDENTIFICATION OF CANDIDATE VNTR

In order to identify sequences with potential VNTR polymorphism, bacterial artificial chromosome (BAC) sequences obtained from the *B. rapa* Genome Project[4] were searched on chromosomes 1, 2, 3, and 9. (This work was performed prior to the release of the complete *B. rapa* genome sequence.) The entire DNA sequence of each BAC was analyzed using the Emboss program eTandem[5]. eTandem generates a score based on the nature of the putative repeat; for a perfect repeat, the score is equal to the length of the entire repetitive sequence minus the first repeat. From the search results, only those potential VNTR's with a score greater than 20 and sequences with three or more repeats of 6–100 base pairs were selected, with preference to the longest repeats that were available and/or the highest number of repetitions. The rationale for this choice was that a larger repeat element would produce a larger size difference between alleles and more repeats would increase the probability of polymorphism. Finally only those results with a percent consensus among repeats of 80% or greater were chosen for further analysis.

PCR primers were designed to prime from the sequences flanking the candidate VNTR using Primer-BLAST[6]. Only primer pairs that were expected to amplify a PCR product size ranging from 200 to approximately 1000 bp were accepted.

Primers designed to amplify potential VNTR markers were tested for suitability by a series of criteria. First they were tested for the ability to robustly and reproducibly amplify a product, i.e., one could always detect a "bright" band on an ethidium bromide stained agarose gel. Those that passed the first test were used to screen for potential polymorphism in a sample of 12 random plants of the strain Standard *B. rapa* as well as the strains DWRCBr52, DWRCBr60, and DWRCBr70. When more than one band size was detected, the products were tested for evidence that they segregated as alleles of the same locus by genotyping individuals from an F_2 generation previously produced by crossing DWRCBr52 and DWRCBr70 strains (Burdzinski and Wendell, 2007). Any primer pair that amplified a product from more than one locus was discarded.

[3] agtc.wayne.edu
[4] www.brassica-rapa.org
[5] http://emboss.bioinformatics.nl
[6] http://www.ncbi.nlm.nih.gov/tools/primer-blast/

IDENTIFICATION OF SNPs IN RCBr

To identify SNPs in RCBr, we resequenced sequence-tagged sites (STS) chosen from those reported by Park et al. (2009). For each STS tested, PCR primers were designed using the program Primer-BLAST (see text footnote 6). PCR was performed on three individuals of each strain tested and the amplicons were pooled to provide template for sequencing. Such pools were generated for each of the strains DWRBr52, DWRBr60, and DWRBr70. The resulting sequence data was then aligned using ClustalW2[7] to identify SNPs between strains.

DEVELOPMENT OF PCR-RFLP MARKERS FROM RCBr SNPs

Single nucleotide polymorphisms were used to develop PCR-RFLP markers using a hierarchical approach. First, the nucleotide sequence surrounding each SNP was screened using NEBcutter V2.0[8] to identify those SNPs that resided within restriction endonuclease recognition sequences. Next, PCR primers were designed so that the position of the SNP, if cut by the enzyme, would produce restriction fragment lengths on a gel that could be easily resolved from each other and from the uncut band if present.

GENOMIC SEQUENCE DATA

Information on gene sequences and *Arabidopsis* homologs connected to the markers developed was obtained through the *Brassica* database BRAD (Cheng et al., 2011).

GENETIC MAPPING

Markers expected to be on chromosome A09 were genetically mapped relative to the *anthocyaninless* (*ANL*) locus in 81 test-cross progeny generated by crossing DWRCBr70 (*ANL*/*ANL*) with DWRCBr52 (*anl*/*anl*) and backcrossing to DWRCBr52. The order of all DNA markers was determined by their position in the *B. rapa* genome sequence available from BRAD (Cheng et al., 2011) and map distances in Kosambi centimorgans were calculated using MAPMANAGER (Manly et al., 2001). The position of the *anthocyaninless* locus was determined as that which gave map distances with the highest LOD scores.

RESULTS

VNTR-TYPE GENETIC MARKERS FOR RCBr

We have developed a total of 14 genetic markers that are based on a VNTR-type repetitive DNA and meet the criteria of robust and reproducible amplification, polymorphism in RCBr strains, and alleles that can be resolved on conventional agarose slab gels (**Table 1**). Markers are available on chromosomes A01, A02, A03, and A09. With the one exception of *D9BrapaS4* which has three alleles, all markers have only two alleles in RCBr populations surveyed. From these 14 VNTR-type markers we chose three to recommend most for use in an educational setting (**Table 2**) because they are most reliable in producing "bright" bands of alleles that are most readily resolved in small agarose slab gels (**Figure 1**). We subjected these three markers to further analysis including DNA sequence of their repetitive element to determine the nature of allelic variation.

[7]http://www.ebi.ac.uk/Tools/msa/clustalw2/
[8]http://tools.neb.com/NEBcutter2/

Table 1 | Variable number tandem repeat-type markers for rapid cycling *Brassica rapa*.

Name	Genome position[1]	Primer sequence
D1BrapaS1	A01:2129419..	GGAGGAGCAAGCAGGACCAGGA
	2130033	ACGCTGTGATTTGTTGCTTCCGA
D1BrapaS2	A01:1997477..	GCGATGCGTATTGGTGGCCG
	1998260	CCGTCGCCGGTTCACAAACCA
D1BrapaS3	A01:2092628..	AAGCAAAGCCAGCGGCGGAT
	2093030	GGCTGGTCACCCACAGGCAC
D1BrapaS4	A01:2125066..	TGGGCGTTGTTCTCATGTTCGGT
	2125999	ACCCGCCATTCTCCCCACCT
D1BrapaS5	A01:1975693..	TCCAAAGTTCTGCTCGAGGGTCC
	1976021	CCAGGAATGGCAGCATTAGACCGA
D1BrapaS6	A01:2129419..	GGAGGAGCAAGCAGGACCAGGA
	2130034	ACGCTGTGATTTGTTGCTTCCGA
D1BrapaS7	A01:4323134..	TGGTCCCTGGATGCGCGGAA
	4323402	GTGGCTGGTCACCGGTGTTGT
D2BrapaS1	A02:2005018..	CCGAACCGTCTCTAACCGAATCGC
	2005276	GGAGCTAGCATCGCTCGCGG
D2BrapaS2	A02:21600350..	ACTTTGTGAGCTTTGGCTGTTGGT
	21601504	AGCCCAGAAAGCGGTTCAGGA
D3BrapaS1	A03:25334204..	ATGTGGCGCGCTGCCATTGA
	25334706	CGCTAAGCATCCTTAACATTTTCGTGC
D9BrapaS1	A09:7345386..	CCAGCCAAATCGTCACTCATGCGA
	7345818	TGCATGCCTAAGAGTTTGGAGTAACAC
D9BrapaS3	A09:1547528..	TCGTGGGACGGTCCTCTTGCT
	1548201	ACCAAACTCTCCACCTCGGACA
D9BrapaS4	A09:28258896..	AGCGATGTAGCACCCGAGTCCA
	28259404	TCGAGCTGAGAGGGAAGCTGTGA
D9BrapaS5	A09:34072636..	CCTTGGCTGCATCAGGCGCA
	34072917	TCCAAAAGTGAGGCTGCCTTAGTGA

[1]*Genome positions were determined by a BLAST search of the Brassica rapa genome sequence version 1.1 using the Brassica Database (BRAD).*

Table 2 | Best RCBr VNTR-type markers for classroom use.

Marker	Allele sizes[1]	Repeat motif[2]
D9BrapaS1	452/497	$(aataagctagtgaagaag)_{22}$
D9BrapaS4	318/462/515	$(gaaaaaacttcactttagctctaaagctaaaaaaga)_3$ $(aaagcttcaatttaagctct)_4$
D1BrapaS1	543/617	$(agttgctgtgtctcctgatgaaat)_{16}$

[1]*The allele sizes are those produced with primers given in* ***Table 1****.*
[2]*The number of repeats listed is in the largest allele.*

D9BrapaS1 resides in a segment of DNA on chromosome A09 that does not contain any annotated genes or other genomic features. Nucleotide sequencing confirms that it contains a VNTR-sized repetitive DNA element (**Table 2**). However, the two alleles present in RCBr do not differ in the repetitive DNA sequences, but instead vary in a 53-bp insertion/deletion in the single-copy DNA flanking the repetitive element (**Figure 2**). Alignment of these alleles with the *B. rapa* genome sequence indicates almost 100% identity except for the 53-bp segment.

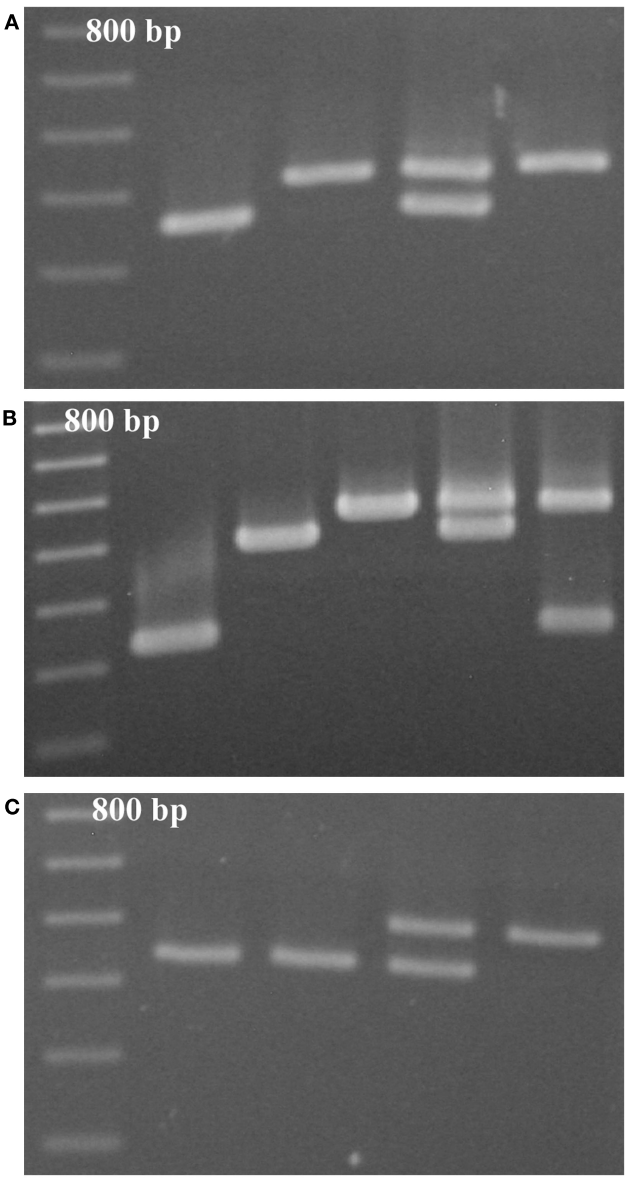


FIGURE 1 | Resolution of alleles of VNTR-type markers (A) *D9BrapaS1*, (B) *D9BrapaS4*, and (C) *D1BrapaS1* on 1.2% agarose gels. The first lane in each gel is a 100-bp DNA ladder.

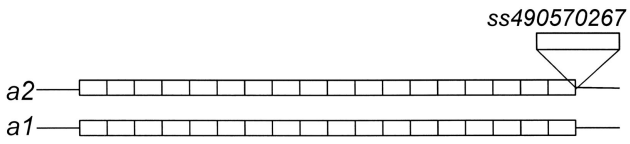


FIGURE 2 | Size variation between the alleles of *D9BrapaS1* in RCBr is due to a single insertion *ss490570267*.

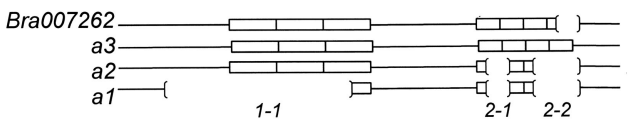


FIGURE 3 | Size variation between the alleles of *D9BrapaS4* in RCBr is due to a multiple indels designated by 1-1 (*ss490570298),* 2-1 (*ss490570286),* and 2-1 (*ss490570293).*

This indel has been deposited in to the dbSNP[9] database (ss490570267).

D9BrapaS4 contains a compound repetitive DNA element (**Table 2**) and the variation between the three alleles we have identified is in these repetitive sequences (**Figure 3**). A search of the *B. rapa* genome indicates that it resides within the first intron of the predicted gene *Bra007262*. Alignment of the comparable portion of the *Bra007262* sequence indicates that the largest allele of *D9BrapaS4* is nearly identical to the *Bra007262* sequence in the BRAD database except for a 20-bp deletion in Bra007262 in the repetitive DNA region of the marker. The indels responsible for the variation between the RCBr alleles have been deposited into the dbSNP database. ss490570286 and ss490570293 are responsible for the size difference between alleles 2 and 3, while the addition of ss490570298 produces allele 1.

D1BrapaS1 contains multiple tandem copies of a 16-bp repeat (**Table 2**), and the variation producing the fragment length difference between alleles is within the repetitive DNA, but the alleles do not vary from each other in numbers of whole repeats. Rather, each allele has several indels (relative to the other alleles) which are mostly smaller than 16 bp. Due to the repetitive DNA sequence, multiple sequence alignments are possible and we cannot presently identify the exact position of the indels.

A search of the *B. rapa* genome finds that the repetitive DNA element at the core of *D1BrapaS1* is in the second exon of the predicted gene *Bra011448* and within its predicted open reading frame. *Bra011448* is a predicted gene based on similarity to *Arabidopsis thaliana* gene *AT4G33500*. Both of these homologous genes are predicted to encode proteins with a protein phosphatase 2C (PP2C) domain near the C-terminus and a ribonuclease E (rne) domain in the second exon. Analysis of the predicted amino acid sequence of Bra011448 using NCBI Conserved Domain Search[10] finds that the *D1BrapaS1* repetitive DNA lies within the predicted ribonuclease E (rne) domain. Comparison of the nucleotide sequence of the two alleles of *D1BrapaS1* with *Bra011448,* as well as the sequence within AC189637.2 deposited in GenBank, shows that each has a different combination of indels, but all preserve the overall reading frame (**Figure 4**). They are all in-frame deletions except for one case in allele 2 of *D1BrapaS1* where two subsequent deletions preserve the reading frame. In contrast, there is very little variation in the section of the gene predicted to encode a PP2C domain. The nucleotide sequence of predicted exons 6, 7, 8, and 9 of *Bra011448* from the RCBr stocks homozygous for either allele of *D1BrapaS1* is 97% identical to the sequence in the BRAD database and there are no deletions.

[9]http://www.ncbi.nlm.nih.gov/projects/SNP/

[10]http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?

SNPs DETECTABLE BY PCR-RFLP

Among SNPs that we identified within RCBr populations, we found two that were readily assayable by PCR-RFLP under classroom conditions (**Table 3**). Both of these were detected with primers that amplified robustly in RCBr and when digested with the appropriate enzyme produced bands that were readily resolved within the same gel (**Figure 5**). The C/G polymorphism identified by *Park9-Hae*III lies within the third exon of predicted F-box protein gene *Bra026987* on chromosome A09. The substitution is a synonymous polymorphism in the third position of a serine codon. The T/C polymorphism identified by *Park14-Eco*RI lies within the seventh intron of *Bra013780*, a predicted transmembrane protein involved in defense or cell death.

LINKAGE OF A09 MARKERS TO THE *ANTHOCYANINLESS* LOCUS

Among the markers we have found to be most suitable for teaching laboratory use, two of the VNTR-type and one of the SNPs were expected to be on chromosome A09, the chromosome which we previously reported to hold the *anthocyaninless* locus (Burdzinski and Wendell, 2007). We genotyped 81 progeny of a testcross between the purple (*ANL/ANL*) DWRCBR70 and non-purple (*anl/anl*) DWRCBr53 strains and found that the *anthocyaninless* locus most likely resides within the 6.3-Mb interval between *D9BrapaS4* and *Park9-Hae*III (**Figure 6**).

MARKER POLYMORPHISM IN FAST PLANTS STRAINS

To assist instructors who obtain RCBr as Fast Plants seeds from Carolina Biological Supply, we have surveyed the allele frequencies of the markers in four popular Fast Plants strains (**Table 4**). For each strain, we determined the genotype of 20 randomly chosen plants grown from seeds obtained directly from Carolina Biological Supply. Only the strain "Standard *B. rapa*" was polymorphic for all five markers tested. It was also the only strain in which we detected all three alleles of *D9BrapaS4*. For each marker, the strains tested usually had the same major allele. The only two exceptions to this pattern were *D9BrapaS1* in the Purple Stem, Hairy strain and *Park14-Hae*III in the Non-purple Stem, Hairless strain. The distribution of the genotypes in the plants tested did not deviate from Hardy-Weinberg expectations (not shown).

NEW RCBr STRAINS WITH DEFINED MARKER GENOTYPES

We have developed three strains of RCBr with genotypes optimized for use of these markers in an instructional lab. The strains vary in both the easy to score Mendelian loci *anthocyaninless* (purple or non-purple stem color) and *yellow-green* (green or yellow-green leaf color), and the DNA markers we have developed. For most markers, a strain is fixed for a particular allele so that crosses between strains will be fully informative (**Table 5**).

DISCUSSION

The DNA-based genetic markers that we have developed were intentionally designed for science education which is the main use of RCBr (also known as Fast Plants). They can be used as both markers for transmission genetics and provide the basis for extensions into molecular biology. The DNA polymorphisms responsible for the observed alleles of *D9BrapaS1*, *D9BrapaS4*, *Park9-Hae*III, and *Park14-Eco*RI have been deposited into the dbSNP database (see text footnote 9) so that when students identify alleles of these markers using basic agarose gels as shown in **Figure 1**, they can then obtain the sequence data underlying these polymorphisms. The sequence data obtained can then be the basis of further exploration of the *B. rapa* genome through the BRAD database[11]. The markers turn out to represent a wide variety of genomic features. Of the VNTR-type markers, one is

[11] http://brassicadb.org/brad/blastPage.php

```
Bra011448    MKTFEAEENLVVEPTATVAL--SPDELVDVSPEENLVVEPTATVAVSTDELVVVSPEEDLVVEPTATVAVSTDELVVVSPEE
AC189637.2   MKTFEAEENLVVEPTATVAV--STDELVVVSPEENHVVEPTATVALSPDELVDVSPEEN----------------------
D1BrapaS1a1  MKTFEAEENLVVEPTATVELDVSPDEPVVVSPEENLVVE----LDVSPDELVVVSPEEKLVVEPTATVEL------------
D1BrapaS1a2  MKTFEAEENLVVEPTATVELDVSPDELVAVSPDENLVVEPTATVTASPDELVAMPPDELVDVASNEIVAVSPDELVAVSPDE

Bra011448    DLVVEPTATVAVSTDELVVVSPEEDLVVEPTATV--AVSTDELVVVSPEEDLVVEPTATVAVSPDELVVTSPDELISTSEAT
AC189637.2   -LVVEPTATVAVSTDELVVVSPEEDLVVEPTATV--AVTPDELVDVPPEENLVVEPTATVAVSPDELVVTSPDELISTSEAT
D1BrapaS1a1  ----------DLSPDELVVVPPEEKLVVEPTAIVELDVSPDELFVVSPEEKLVVEPTATVAVTPDELAAVSPDELVSTSEAT
D1BrapaS1a2  NLVVEST--VAASPDELVALPPDDLVDVAPNELV--AVSPDELVTVSPDENLVVEPTATVAVTPEEPVAVSPDELISTSEAT
```

FIGURE 4 | Four different alleles of *D1BrapaS1* show four different combinations of in-frame deletions in *Bra011448*. The sequence shown is the portion of *Bra011448* that is predicted to encode an *rne* domain and lies within the marker *D1BrapaS1*. *AC1896372* and *Bra011448* are Chinese cabbage sequences from public databases and *D1BrapaS1a1* and *D1BrapaS1a2* are from RCBr.

Table 3 | Single nucleotide polymorphism of rapid cycling *Brassica rapa* detected by PCR-RFLP.

Name	Genome position and primer sequences	SNP and ss#	Enzyme	PCR-RFLP, fragments
Park 9	A09:34639078 TCCTCAGCTGCTTTAGCCTC TTGCGACAAAGAAACACAGC	C/G ss490571632	*Hae*III	1022/310 + 712
Park 14	A01:7729329 TGTGCTGTAACTGCAAAGCA CGCAAATCACGAGTTCTTCA	T/C ss490571635	*Eco*RI	477 + 839/262 + 215 + 839

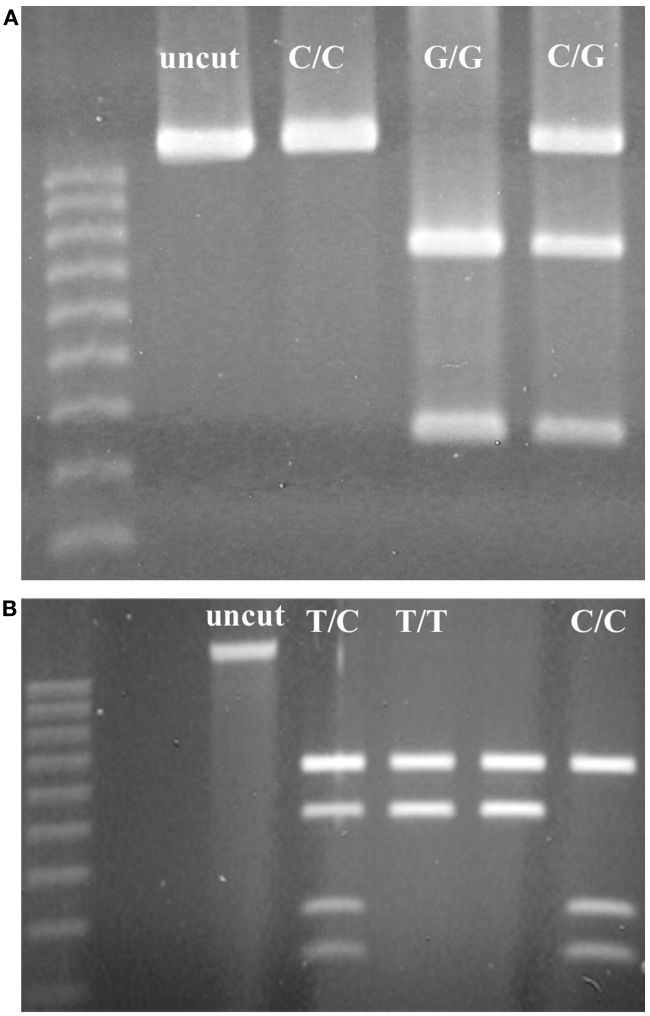


FIGURE 5 | Resolution of alleles of PCR-RFLP markers (A) *Park9-Hae*III (ss490571632) and (B) *Park14-Eco*RI (ss490571635) on 1.2% agarose gels.

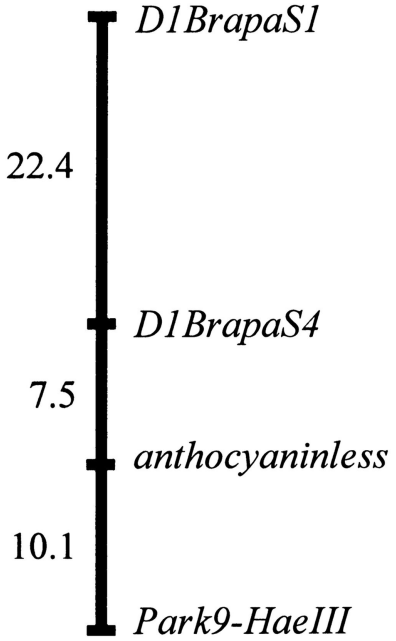


FIGURE 6 | Genetic map of the *anthocyaninless* locus relative to A09 marker reported in this paper. Map distances are Kosambi centimorgans.

in a non-genic region, one is within the intron of a gene, and one is within the open reading frame of a gene. Of the SNPs, one is in an intron and one is in the open reading from of a gene, although it is a synonymous substitution. Instructors can use the DNA sequence information we present here to develop lessons for students to study the possible impact on gene function of the sequence variation of the alleles.

Three of the markers form a linkage group with the *anthocyaninless* locus allowing them to be used in laboratory projects in genetic linkage and mapping. They are also excellent tools for projects such as paternity testing. We have previously described a laboratory project using RCBr to perform paternity testing, but the previous design used microsatellite markers (Wendell and Pickard, 2007) which can pose difficulties for lab instructors due to the need for polyacrylamide gels to resolve them. However, the markers we report here can be detected and alleles resolved in the most simple agarose slab gels.

The data we provide on population allele frequencies in RCBr strains will be valuable to instructors using these markers for educational labs. For example, we previously described a simple lab project in which the students perform paternity testing by pollinating one plant ("Mother") with a mixture of pollen from two other plants ("Alleged Fathers"), but success in this project requires that the Alleged Fathers have different alleles for the markers used (Wendell and Pickard, 2007). An instructor who wishes to perform this project using Fast Plants obtained from Carolina Biological Supply would be best served using the markers *D9BrapaS1* and *Park14-Eco*RI since these are polymorphic in all strains tested (**Table 4**). To work with a great degree of polymorphism, an instructor should use the strains described in **Table 5** because they vary greatly in their genotypes for both our markers and simple Mendelian traits. Seeds for these strains are available by request to Douglas Wendell[12].

The VNTR-type markers that we developed are not as polymorphic as expected for repetitive DNA-based markers. Except for *D9BrapaS4*, we have only found two alleles for each of the markers despite testing numerous RCBr strains, whereas VNTR markers used in mapping and DNA fingerprinting typically have multiple alleles (Nakamura et al., 1987). This could result if the repetitive DNA elements that we have selected are not prone to polymorphism, but could also result if or the RCBr populations have a low rate of polymorphism. The latter explanation is consistent with previous work in which we tested microsatellite markers

[12]wendell@oakland.edu

Table 4 | Estimated marker allele frequencies in fast plants strains from Carolina Biological Supply.

Marker	Allele	Strain and catalog number			
		Standard *Brassica rapa* 158804	Purple stem, hairy 158810	Non-purple stem, hairless 158812	Non-purple stem, yellow-green leaf 158842
D9BrapaS1	1	0.21	0.53	0.15	0.53
	2	0.79	0.47	0.85	0.47
D9BrapaS4	1	0.15	–	–	0.03
	2	0.06	–	0.13	–
	3	0.79	1.00	0.87	0.97
D1BrapaS1	1	0.97	1.00	1.00	0.84
	2	0.03	–	–	0.16
*Park9-Hae*III	C	0.09	0.03	–	0.23
	G	0.91	0.97	1.00	0.77
*Park14-Eco*RI	T	0.30	0.38	0.82	0.25
	C	0.70	0.63	0.18	0.75

Table 5 | New RCBr strains with defined marker genotypes.

Marker	Allele	DWRCBr53	DWRCBr76	DWRCBr91
D9BrapaS1	1	0.0	1.0	0.0
	2	1.0	0.0	1.0
D9BrapaS4	1	0.0	1.0	0.0
	2	1.0	0.0	0.0
	3	0.0	0.0	1.0
D1BrapaS1	1	1.0	0.0	1.0
	2	0.0	1.0	0.0
*Park9-Hae*III	C	0.0	0.8	0.0
	G	1.0	0.2	1.0
*Park14-Eco*RI	T	1.0	0.4	0.4
	C	0.0	0.6	0.6
Anthocyaninless	*ANL*	0.0	1.0	0.0
	anl	1.0	0.0	1.0
Yellow-green	*YGR*	1.0	1.0	0.0
	ygr	0.0	0.0	1.0

that had been developed for *Brassica* crop species for the usefulness in RCBr. Out of 37 primer pairs that amplified a product in RCBr DNA we only found 22 to be polymorphic and only 11 that had more than two alleles in RCBr (Burdzinski and Wendell, 2007), despite the fact that microsatellites usually have multiple alleles.

The reader may wonder why we only report two SNP markers given that SNPs are abundant in organisms, and other groups have reported huge lists of SNPs for *B. rapa* (Li et al., 2009; Park et al., 2009). We did find several other SNPs (not shown) that lie within restriction sites but we were unable to design a PCR-RFLP around them that gave legible bands. The main source of the problem was the difference in size between "cut" and "uncut" alleles when detected by PCR-RFLP in ethidium bromide stained gels. Because the intensity of staining of DNA in gels by dyes, whether fluorescent or visible, is proportional to the mass of DNA in a band, we encountered a problem of markers where the lower molecular weight bands of the cut allele were too faint for student to reliably detect. Another complication was that in some cases the restriction enzyme that recognized the SNP also had multiple recognition sites close to the SNP.

In addition to developing markers and plant strains, we have developed classroom-tested protocols for their use. We make these publically available at the web site humangeneticsmustard.blogspot.com and will be adding more instructor resources as we develop them.

ACKNOWLEDGMENTS

This work was funded by an American Recovery and Reinvestment Act grant 5 RC1 RR030293-02 from the National Institutes of Health, USA. We thank Christian Brigolin and Jay Edwards for technical assistance.

REFERENCES

Burdzinski, C., and Wendell, D. L. (2007). Mapping the anthocyaninless (anl) locus in rapid-cycling *Brassica rapa* (RBr) to linkage group R9. *BMC Genet.* 8, 64. doi:10.1186/1471-2156-8-64

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., and Wang, X. (2011). BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biol.* 11, 136. doi:10.1186/1471-2229-11-136

Goldman, I. L. (1999). Teaching recurrent selection in the classroom with Wisconsin Fast plants. *Horttechnology* 9, 579–584.

Iniguez-Luy, F. L., Lukens, L., Farnham, M. W., Amasino, R. M., and Osborn, T. C. (2009). Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*. TAG. *Theor. Appl. Genet.* 120, 31–43.

Konieczny, A., and Ausubel, F. M. (1993). A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* 4, 403–410.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* 17, 21–24.

Lauffer, D., and Fall, B. (2000). "Evolution by artificial selection and unraveling the mysteries of Hairy's inheritance," in *21st Workshop/Conference of the Association for Biology Laboratory Education (ABLE)*, ed. J. Karcher (Lincoln: Association for Biology Laboratory Education (ABLE)), 147–179.

Li, F., Kitashiba, H., Inaba, K., and Nishio, T. (2009). A *Brassica rapa* linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. *DNA Res.* 16, 311–323.

Litt, M., and Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplificaation of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* 44, 397–401.

Manly, K., Cudmore, R., and Meer, J. (2001). Mapmanager QTX, cross-platform software for genetic mapping. *Mamm. Genome* 12, 930–932.

Musgrave, M. E. (2000). Realizing the potential of rapid-cycling *Brassica* as a model system for use in plant biology research. *J. Plant Growth Regul.* 19, 314–325.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235, 1616–1622.

Park, S., Yu, H.-J., Mun, J.-H., and Lee, S.-C. (2009). Genome-wide discovery of DNA polymorphism in *Brassica rapa*. *Mol. Genet. Genomics* 283, 135–145.

Weber, J. L., and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* 44, 388–396.

Wendell, D. L., and Pickard, D. (2007). Teaching human genetics with mustard: rapid cycling *Brassica rapa* (fast plants type) as a model for human genetics in the classroom laboratory. *CBE Life Sci. Educ.* 6, 179–185.

Williams, P. (1997). *Exploring with Fast Plants*. Dubuque: Kendall Hunt.

Williams, P. H. (1985). The crucifer genetics cooperative. *Plant Mol. Biol. Rep.* 3, 129–144.

Williams, P. H. (2007). *The Rapid Cycling Brassica Collection Catalog.* Madison, WI: Wisconsin Fast Plants Program.

Williams, P. H., and Hill, C. B. (1986). Rapid-cycling populations of *Brassica*. *Science* 232, 1385–1389.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 February 2012; accepted: 17 May 2012; published online: 01 June 2012.

*Citation: Slankster EE, Chase JM, Jones LA and Wendell DL (2012) DNA-based genetic markers for rapid cycling Brassica rapa (Fast Plants type) designed for the teaching laboratory. Front. Plant Sci. **3**:118. doi: 10.3389/fpls.2012.00118*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 16 August 2012
doi: 10.3389/fpls.2012.00183

Genetic analysis of morphological traits in a new, versatile, rapid-cycling *Brassica rapa* recombinant inbred line population

Hedayat Bagheri[1,2†], Mohamed El-Soda[1,3†], Inge van Oorschot[1], Corrie Hanhart[1], Guusje Bonnema[4], Tanja Jansen-van den Bosch[5], Rolf Mank[5], Joost J. B. Keurentjes[1], Lin Meng[6], Jian Wu[6], Maarten Koornneef[1,7] and Mark G. M. Aarts[1]*

[1] Laboratory of Genetics, Wageningen University, Wageningen, Netherlands
[2] Bu-Ali Sina University, Hamedan, Iran
[3] Faculty of Agriculture, Department of Genetics, Cairo University, Giza, Egypt
[4] Laboratory of Plant Breeding, Wageningen University, Wageningen, Netherlands
[5] Keygene N. V., Wageningen, Netherlands
[6] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China
[7] Max Planck Institute for Plant Breeding Research, Cologne, Germany

Edited by:
Xiaowu Wang, CAAS, China

Reviewed by:
Joseph F. Petolino, Dow AgroSciences, USA
John Hammond, University of Western Australia, Australia

***Correspondence:**
Mark G. M. Aarts, Laboratory of Genetics, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, Netherlands.
e-mail: mark.aarts@wur.nl

† These authors equally contributed to this work.

A recombinant inbred line (RIL) population was produced based on a wide cross between the rapid-cycling and self-compatible genotypes L58, a Caixin vegetable type, and R-o-18, a yellow sarson oil type. A linkage map based on 160 F7 lines was constructed using 100 Single nucleotide polymorphisms (SNPs), 130 AFLP®, 27 InDel, and 13 publicly available SSR markers. The map covers a total length of 1150 centiMorgan (cM) with an average resolution of 4.3 cM/marker. To demonstrate the versatility of this new population, 17 traits, related to plant architecture and seed characteristics, were subjected to quantitative trait loci (QTL) analysis. A total of 47 QTLs were detected, each explaining between 6 and 54% of the total phenotypic variance for the concerned trait. The genetic analysis shows that this population is a useful new tool for analyzing genetic variation for interesting traits in *B. rapa*, and for further exploitation of the recent availability of the *B. rapa* whole genome sequence for gene cloning and gene function analysis.

Keywords: *Brassica rapa*, recombinant inbred line population, QTL analysis, plant breeding

INTRODUCTION

Brassica rapa is an important, widely cultivated crop, with various forms or "morphotypes", such as leafy vegetables, turnips, and oilseed rape (Zhao et al., 2005). While the use of *B. rapa* as an oilseed crop is relatively modest, it is important as one of the parents of *Brassica napus*, the most important oilseed crop. After the oil has been extracted from the seeds, the remaining seed components (meal) are of economic interest for feeding animals. It has been known for some time that breeding for yellow seed color is advantageous for meal quality in *B. napus*, because yellow-seeded genotypes have a thinner seed coat associated with a higher protein content and less non-energetic and anti-nutritive fiber components (Liu et al., 2012). Therefore, breeding programs aiming at combining yellow seed color with yield associated traits such as seed number, seed size, number of siliques per plant, pod shattering, carpel number, and vivipary, have been developed in *B. napus* through interspecific crosses with yellow-seeded *Brassica* species (Tang et al., 1997; Badani et al., 2006; Wittkop et al., 2009; Liu et al., 2012). A problem is that the expression of the yellow color, at least in *B. napus*, is highly dependent on environmental factors (Liu et al., 2012).

Pod shattering, caused by carpel abscission, is an undesirable characteristic in crop breeding as it decreases the yield due to seed loss during harvesting. For *B. napus* the seed yield loss can be as much as 20% of the harvest (Price et al., 1996). The absence of embryonic dormancy during seed development, which prevents seeds to germinate prematurely on the mother plant, can be expressed as vivipary. While this is more commonly observed in cereals, it can also be found in oilseed rape, leading to large economic losses due to significant reduction in seed quality. Resistance to vivipary is therefore a very favorable trait in breeding programs (Zhang et al., 2008). Next to seed related traits, plant height (Ph), branch number (Bn), and leaf number at first flower opening are factors contributing to Brassica plant architecture that differ considerably between genotypes. Plant architecture is of major agronomic importance and has a strong effect on the suitability of a plant species for cultivation, as it affects plant yield and harvest efficiency (Reinhardt and Kuhlemeier, 2002).

With the smallest genome size in the *Brassica* genus, the rapid life cycle of some of its genotypes, and the relatively close relationship to the model plant species *Arabidopsis thaliana*, *B. rapa* is considered to be one of the model dicot crops for genetic studies (Wang et al., 2011a). These studies require "immortal" mapping populations, i.e., populations that can be easily propagated through seed without altering their genotypes, as indispensable tools in identifying quantitative trait loci (QTLs) underlying traits

of interest (Koornneef et al., 2004). Doubled haploid (DH) populations are the most commonly used type of immortal mapping populations for *Brassica* species (Pink et al., 2008). However, the poor response of many *B. rapa* genotypes to DH induction (Kole et al., 1997) together with the high degree of segregation distortion often observed in DH populations (Voorrips et al., 1997), limits this use. Instead, when using self-compatible genotypes with short generation times it is feasible to develop Recombinant Inbred Line (RIL) populations through sexual propagation. In this study two *B. rapa* genotypes, corresponding to two distinct morphotypes, the leafy vegetable Cai Xin accession L58, of Chinese ancestry, and the yellow sarson oil seed DH line R-o-18, of Indian ancestry, were crossed to generate a RIL population. Both parents are early flowering and self-compatible, which facilitates rapid propagation and the ability to maintain the RILs through single seed descent.

Genetic linkage maps are required to properly query DH or RIL populations for the identification of the chromosomal regions or QTLs that harbor the genes controlling important agronomic traits. Single nucleotide polymorphisms (SNPs) represent the most abundant and common type of genetic polymorphisms that can be readily converted into genetic markers for marker assisted selection. Large-scale SNP discovery projects, using high-throughput sequencing techniques, have become a powerful complement to the standard genetic mapping procedures, and the use of resulting markers greatly improves the linkage maps of diploid crops. The Illumina GoldenGate assay is an efficient SNP genotyping tool that has been used already for soybean, tetraploid and hexaploid wheat lines, and maize (Hyten et al., 2008; Akhunov et al., 2009; Yan et al., 2010). Currently, SNP genotyping is replacing the use of the AFLP technology, which has previously been very useful for analyzing genetic diversity and relationships in many plant species, including *B. rapa*, identifying a large number of polymorphic loci (Zhao et al., 2005).

This paper describes the generation and genetic mapping of a large, versatile, rapid-cycling *B. rapa* RIL population dedicated for QTL analysis. As an illustration of the potential importance of this population, we used it to identify 47 QTLs, responsible for most of the observed morphological variation in 17 different traits.

MATERIALS AND METHODS

PLANT GROWTH AND GENERATION OF THE RIL POPULATION

The two parental genotypes L58 and R-o-18 were crossed reciprocally and from each of the two F1 offspring, one plant was randomly selected to be propagated by subsequent generations of self-fertilization using a single-seed-descent approach, aimed at minimizing any bias in selecting plants. The seeds of L58 (*B. rapa* ssp. *parachinensis*) were provided by Dr. Xiaowu Wang from the Institute for Vegetables and Flowers of the Chinese Academy of Agricultural Sciences, Beijing, China; and seeds of R-o-18 (*B. rapa* var. *trilocularis*) were obtained from Dr. Lars Østergaard, John Innes Centre, Norwich, UK. One of the two F1 combinations, L58 (♀) × R-o-18 (♂), was propagated until the F7 generation, the other remained at F5 and could be used for future fine-mapping studies. All generations were grown between April 2007 and June 2009 with four replications in a fully randomised design. Individual plants were grown in 19-cm diameter black plastic pots filled with a potting soil consisting of prefertilized peat, obtained from "Lentse potgrond" (www.lentsepotgrond.nl), in a temperature-controlled greenhouse at 21°C with artificial long day light (16 h). No cold treatment or vernalization was applied for germination or flowering respectively. For every generation, the first flower appeared about four weeks after germination in the early flowering lines. The inflorescences were covered with perforated plastic bags to prevent cross-pollination by insects. In case of poor seed set, hand pollinations were performed. The 160 F7 RILs were multiplied in the same conditions, ensuring homogeneous material for genetic studies.

DNA EXTRACTION AND GENOTYPING

DNA was extracted from frozen F7 leaves according to a modified CTAB procedure (Beek et al., 1992). The DNA was amplified with the Genomiphi-kit (*Illustra™ GenomiPhi™ V2 DNA Amplification Kit*, GE Healthcare UK) to be suitable for GoldenGate assay analysis (Akhunov et al., 2009). For SNP discovery, two *B. rapa* lines (Kenshin and Chiifu) were compared using CRoPS®-technology (van Orsouw et al., 2007) to reveal more than 1300 putative SNPs. The SNP-harboring sequences were processed with the Illumina Assay Design Tool (ADT) by Illumina (www.illumina.com). A total of 384 SNPs were selected, all having ADT scores above 0.6. 100–500 ng of genomic DNA (GenomiPhi) per plant was used for Illumina SNP genotyping at Keygene N.V. using the Illumina BeadXpress™ platform and the GoldenGate Assay. Part of the DNA was used for SSR or AFLP detection as described by Choi et al. (2007) and Vos et al. (1995), respectively. Pre-amplification and selective amplification for AFLP analysis were carried out as described by Zhao et al. (2005). For selective amplification seven combinations of EM (*Eco*RI/*Mse*I) primers (E34M15, E34M16, E37M32, E37M49, E37M56, E40M38, and E40M51) and four combinations of PM (P*st*I/*Mse*I) primers (P23M48, P23M50, P21M47, and P23M47) were used. The *Pst* I and *Eco*RI primers were labeled with IRD-700 at their 5′ ends (Zhao et al., 2005). The reaction product of selective amplification was mixed with an equal volume of formamide-loading buffer, denatured for 5 min at 94°C, cooled on ice and run on a 5.5% denaturing polyacrylamide gel using the LI-COR system 4200 DNA sequencer (Li-Cor, Lincoln, Neb.) (Myburg et al., 2001). The AFLP gel images were analyzed by the AFLP-Quantar Pro software. All distinguishable bands ranging from 50 to 500 bp were used in the data analysis. AFLP bands were scored as 1 or 0 for presence or absence of the band, respectively. All weak and ambiguous bands were scored as "unknown". In addition, 36 public SSR primer pairs (Choi et al., 2007) were used to screen for polymorphisms using the same LI-COR system to run a 5.5% denaturing polyacrylamide gel. Furthermore, 27 polymorphic InDel markers, based on DNA resequencing information of two parental lines of a DH population, which was used to construct a *B. rapa* reference map for pseudochromosome sequence assembly, were screened as described by (Wang et al., 2011b).

CONSTRUCTION OF A GENETIC LINKAGE MAP AND QTL ANALYSIS

The genetic map was constructed using JoinMap 4.0 (www.kyazma.nl). Monomorphic markers, markers with a high number

of unknown scores and markers with more than 75% allele skewedness toward either A or B were removed. Recombination frequencies were converted to centiMorgan (cM) distances using Haldane's mapping function. SNP markers positions were confirmed by comparing their primer sequences with the *B. rapa* genome using the *Brassica* database (BRAD) (brassicadb.org) of *Brassica* crops whole genome sequence and genetics data (Cheng et al., 2011). It contains the complete *Brassica* A genome sequence from the reference *B. rapa* genotype Chiifu-401-42 (Wang et al., 2011a). InDel markers were compared to the reference map (Choi et al., 2007; Wang et al., 2011b), which was previously used for chromosome alignment.

MAPQTL 6.0 (www.kyazma.nl) was used for QTL analysis. First, the interval mapping procedure was performed to detect major QTLs. For each trait a 1000 X permutation test was performed to calculate the LOD threshold corresponding to a genome-wide false discovery rate of 5% ($P < 0.05$). Markers with LOD scores equal to or exceeding the threshold were used as cofactors in multiple-QTL-model (MQM) mapping. If new QTLs were detected, the linked markers were added to the cofactor list and the MQM analysis was repeated. If the LOD value of a marker dropped below the threshold in the new model, it was removed from the cofactor list and the MQM analysis was rerun. This procedure was repeated until the cofactor list became stable. The final LOD score for each trait was determined by restricted MQM (rMQM) mapping. In some cases, rMQM mapping showed that some cofactors should be on the same linkage group, but at slightly different positions. In that case, the new marker was selected as a cofactor and the whole procedure was repeated. The linkage map was visualized using Mapchart (Voorrips, 2002).

TRAIT MEASUREMENT

The 160 RILs (four replicate plants) and both parents (five replicate plants) were phenotyped for 17 traits. These traits are categorized into two main groups. Seed related traits, including seed color, seed weight, seed oil, seed germination, and seed vivipary; and morphological traits, including flowering time (Ft), total height, Ph until the first flower, Bn, silique length (Sil), silique beak length (Bl), silique number (Sin), number of seeds per silique (Nsps), carpel number, pod shattering, total leaf number (Tln), and leaf number until the first flower (Lnf). Seed color of fully mature F8 seeds was visually scored and ranked into nine different classes ranging from yellow (1) to black (9). Seed germination data were obtained by sowing 30 seeds of each line and scoring the percentage of germination 15 h after sowing. The seeds were sterilized in 2% sodium hypochlorite for 2 min. After rinsing 2 times with sterile distilled water, they were sown in two rows of 15 seeds on square plates containing 50 ml of half MS medium +1% agar. The plates were placed vertically in a 25°C growth chamber with a 16/8 h light/dark photoperiod. Sil and Nsps were averaged from three ripe siliques. Seed vivipary was scored as either 0 (no vivipary), 0.5 (medium), or 1 (high) based on visual estimation of the number of seeds with radicles when harvested. Shattering was scored at harvesting time as either 0 (no open siliques), 0.5 (few open siliques), or 1 (many open siliques) (**Figure 1**). Seed oil was extracted by a crude method of hexane extraction, grinding 10 weighed F7 seeds of each line in 650 μl of hexane, shaking the mix for 2 min followed by 1 min of centrifugation at 14,000 rpm in an Eppendorf microfuge. 600 μl of supernatant was transferred to a new tube and left overnight in the fume hood to evaporate the hexane. The oil content was determined in mg oil per mg seed (Goossens et al., 1999). All traits were measured for each of the four replicate plants, and the average values were used for mapping, except for seed color, seed germination, and seed oil content, for which only one replication could be measured. The heritability was calculated as the ratio between the genetic variation (Vg), i.e., variance between the average values of all RILs, and the total variation (Vt), with $Vt = Vg + Ve$, where Ve is the environmental variation, i.e., variance between the replications of all lines. All statistical analysis was performed in SPSS 19.

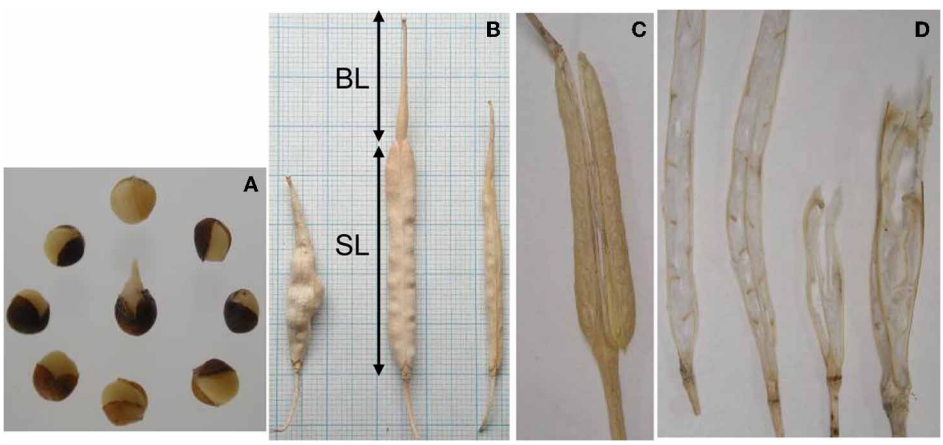


FIGURE 1 | Phenotyping of RIL population of *B. rapa* L58 × R-o-18. **(A)** Seed vivipary; i.e., premature germination of seeds still in the silique, or just after harvesting, **(B)** silique length (SL) and silique beak length (BL), **(C)** pod shattering, corresponding to the fraction of opened siliques at harvesting, **(D)** carpel number, with the left two siliques having two carpels and the two on the right having three.

RESULTS

GENOTYPING AND CONSTRUCTION OF THE LINKAGE MAP FOR THE RIL POPULATION

The availability of the complete genome sequence of *B. rapa* (Wang et al., 2011a) and the genome analysis tools provided in the BRAD database (Cheng et al., 2011), were critical for constructing a reliable genetic map of the L58 × R-o-18 RIL population suitable for QTL mapping. Out of the 384 SNPs that could be queried by the Brassica GoldenGate assay we used, 120 SNPs were polymorphic between the parents, of which 100 provided unambiguous genotype calls for mapping. Based on the sequence of the SNP primers, the position of the 100 mapped SNP markers could be linked to their sequence position on the *B. rapa* genome, thus confirming the mapping results and providing anchoring points for chromosome number assignment and proper orientation of the chromosomal linkage maps with the genome sequence. The same was done for the SSR markers previously used to create the *B. rapa* reference linkage map (Choi et al., 2007). In total 94 InDel markers were screened, from which 27 showed polymorphism between the two parental lines. Seven of these polymorphic markers have been mapped on the reference map used for *B. rapa* pseudochromosome assembly (Wang et al., 2011b), while the other markers were assigned to the chromosomes according to the position of their corresponding sequence scaffolds. The final linkage map was constructed for the L58 × R-o-18 F7 RIL population using 100 SNP, 130 AFLP, 27 InDel, and 13 SSR markers. It covers a total length of 1150 cM with an average resolution of 4.3 cM per marker (**Figure 2**).

PHENOTYPING THE RIL POPULATION

A total of 17 traits were analyzed for the F7 RIL population. **Figure 3** shows the frequency distributions of the measured traits over the whole population. Transgression beyond the parental line values was observed for most of the traits except seed color, pod shattering, seed germination, and vivipary. Broad sense heritabilities ranged from 0.35, for stem thickness, to 0.92, for Ft (**Table 1**). Heritabilities could not be determined for seed color, seed germination, and seed oil content, as for these traits only one replication could be measured. Correlation analysis of all measured traits (**Table 2**) showed that Ft was highly positively correlated with Tln and Lnf. Sin, Nsps, pod shattering, and Sil were also positively correlated. In general, plants with more siliques had longer siliques with more seeds and higher seed oil content, all contributing to traits favored for oil seed rape.

QTL ANALYSIS

In total 47 QTLs were mapped for the 17 analyzed traits (**Table 3** and **Figure 4**). Seed color was a very prominent phenotype segregating in the population. A major QTL for seed color (*Sc1*) was mapped to chromosome A9 with a LOD score of 30.8 and explaining 53.7% of the total seed color variance. This region on A9 appears to be rich in genetic variation, with several other QTLs co-located with *Sc1*, which are loci for pod shattering (*Sh*), number of seeds per silique (*Nsps1*), and seed oil (*So*). The *Sh* QTL also explains a large portion, 18%, of the phenotypic variance. Another QTL for seed color (*Sc2*), with a LOD score of 12.1, was mapped to chromosome A3, accounting for 15% explained variance. Variation in vivipary (Vi) was explained by two loci, one locus on A9 (*Vi1*), with 20% explained variance, and another on A6 (*Vi2*) that explains 13% of the variance. The carpel number QTL (*Cn1*) co-localized with the *Sil* QTL on A4, each explaining 15%, respectively 17% of the variance. This region also harbors one of the Bl QTLs (*Bl3*). As can be seen from **Figure 1**, these traits appear to be pleiotropic effects of the same locus, as the increase in carpel number often corresponds with malformed, shorter siliques with shorter beaks. Pleiotropy is also the likely cause of the co-localization of Ft QTLs *Ft3, Ft4,* and *Ft5* with QTLs for Tln (*Tln1, Tln2, Tln4*) and Lnf (*Lnf1, Lnf3, Lnf4*) on respectively A2, A7, and A8. Tln and Lnf share four of the six QTLs found for these traits, in line with the correlation found between them. The locus on A8 also seems to account for variation for Bn, harboring the major Bn QTL (*Bn1*). Ph until the first flower and total Ph (Tph) also share one common QTL, on A10 (*Ph1* and *Tph2*).

DISCUSSION

The L-58 × R-o-18 population is a new RIL population, designed for general QTL mapping studies. The parents of this population were selected for a number of reasons. Rapid-cycling and self-compatibility were two important reasons, as these would permit the rapid construction of the population and easy maintenance through single-seed-descent propagation. These are also the reasons that both parents are more and more used as reference genotypes, expanding their use for other purposes, such as the generation of a TILLING population in R-o-18 (Stephenson et al., 2010), as reference species in micro-array design (Love et al., 2010), as well as being used in setting up a diversity fixed foundation set (DFFS) and as parents in other mapping populations. For the latter purpose, currently the genome sequences and transcript profiles of both parents are being determined (Jian Wu and Xiaowu Wang e.a., unpublished results). There are not many "immortal" *B. rapa* populations available for mapping studies, with immortal meaning that the individual lines are genetically homozygous and can thus be propagated through seeds while maintaining the established genotype in their progeny. There are few other RIL populations (Kole et al., 1997; Iniguez-Luy et al., 2009), although others may still be in development (www.brassica.info). In addition, there are several DH populations available (Zhang et al., 2006; Choi et al., 2007; Lou et al., 2007; Zhao et al., 2010; Wang et al., 2011b), which are also very useful for genetic mapping studies, although they generally comprise about half the number of recombination events compared to RIL populations and often suffer more from regions with skewedness toward one of the parental alleles.

The transgression beyond the parental lines, which was observed in the F2 generation (Bagheri et al., under review) was encouraging to produce the F7 RIL family through single-seed-descent. Out of 200 F2 lines, only 160 F7 lines were available for genotyping. This 16% loss from F2 till F7 was mostly due to plant sterility, apparently from reduced pollen production. Although this may have a genetic basis, it was not obviously related to strong skewedness of the population toward one of the two parental alleles at a particular locus. To reduce the risk of

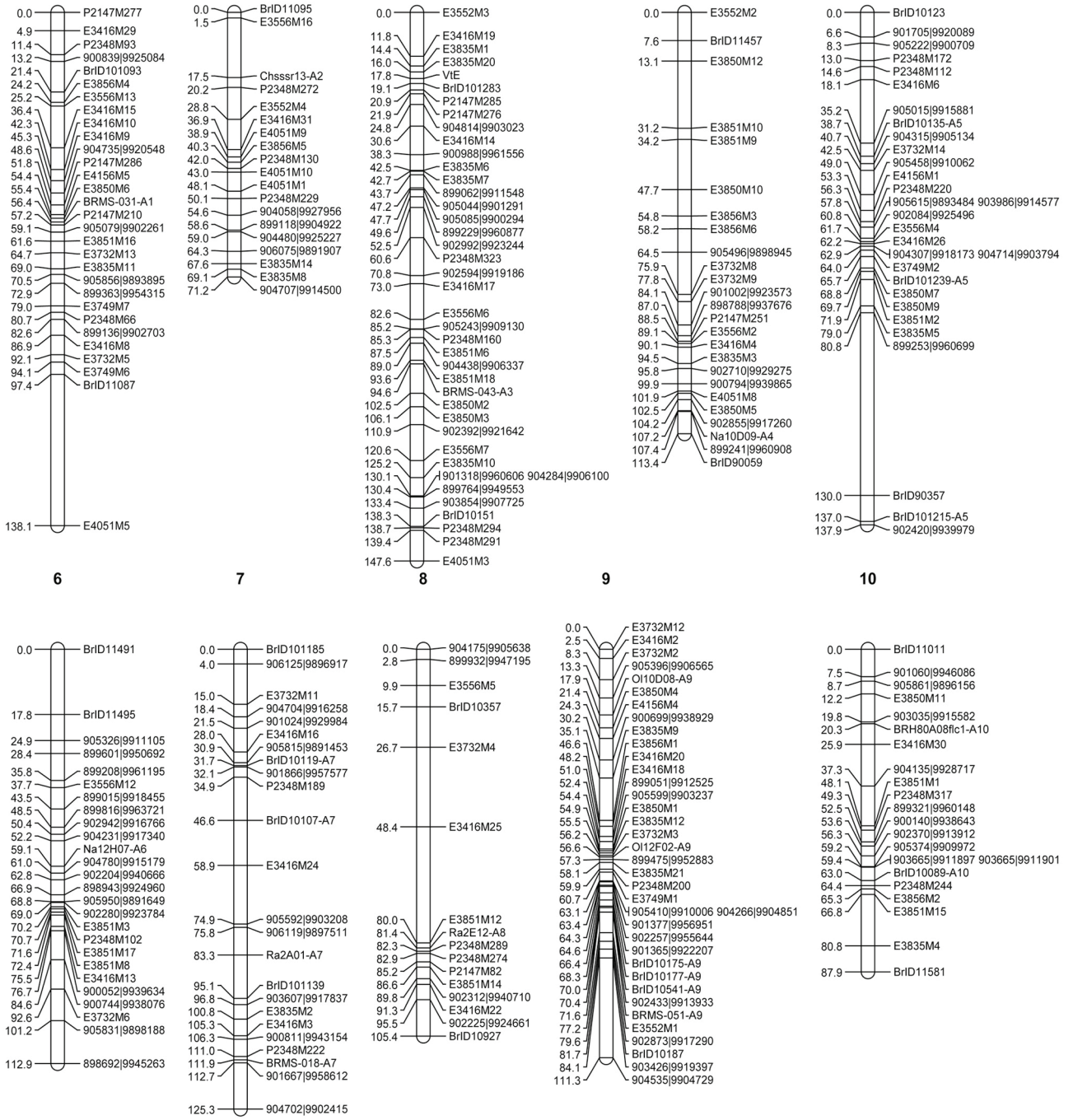


FIGURE 2 | Genetic linkage map of the *B. rapa* L58 × R-o-18 RIL population, showing the positions of 270 markers (100 SNP, 130 AFLP, 27 InDel, and 13 SSR markers) distributed over 10 linkage groups corresponding to the 10 chromosomes of the *Brassica* A genome.

Markers labeled with [6 digits|7 digits] are SNPs, markers labeled "E . . . M.." or "P . . . M.." are respectively EcoRI/MseI or PstI/MseI generated AFLPs, markers labeled "BrID . . . ." are InDels and the remaining markers are SSRs.

skewedness we started off with a relatively large population. Thus, most of the cross-overs between alleles located in skewed segregation region could be detected and correct linkage distances could be calculated along any skewed marker region. We did find the occasional marker with more than 75% allele skewedness toward either L58 or R-o-18 alleles, but since all of these markers were flanked by closely linked, non-skewed markers, the skewedness was found to be due to marker scoring problems rather than genetic skewedness, upon which the improperly scored markers were removed.

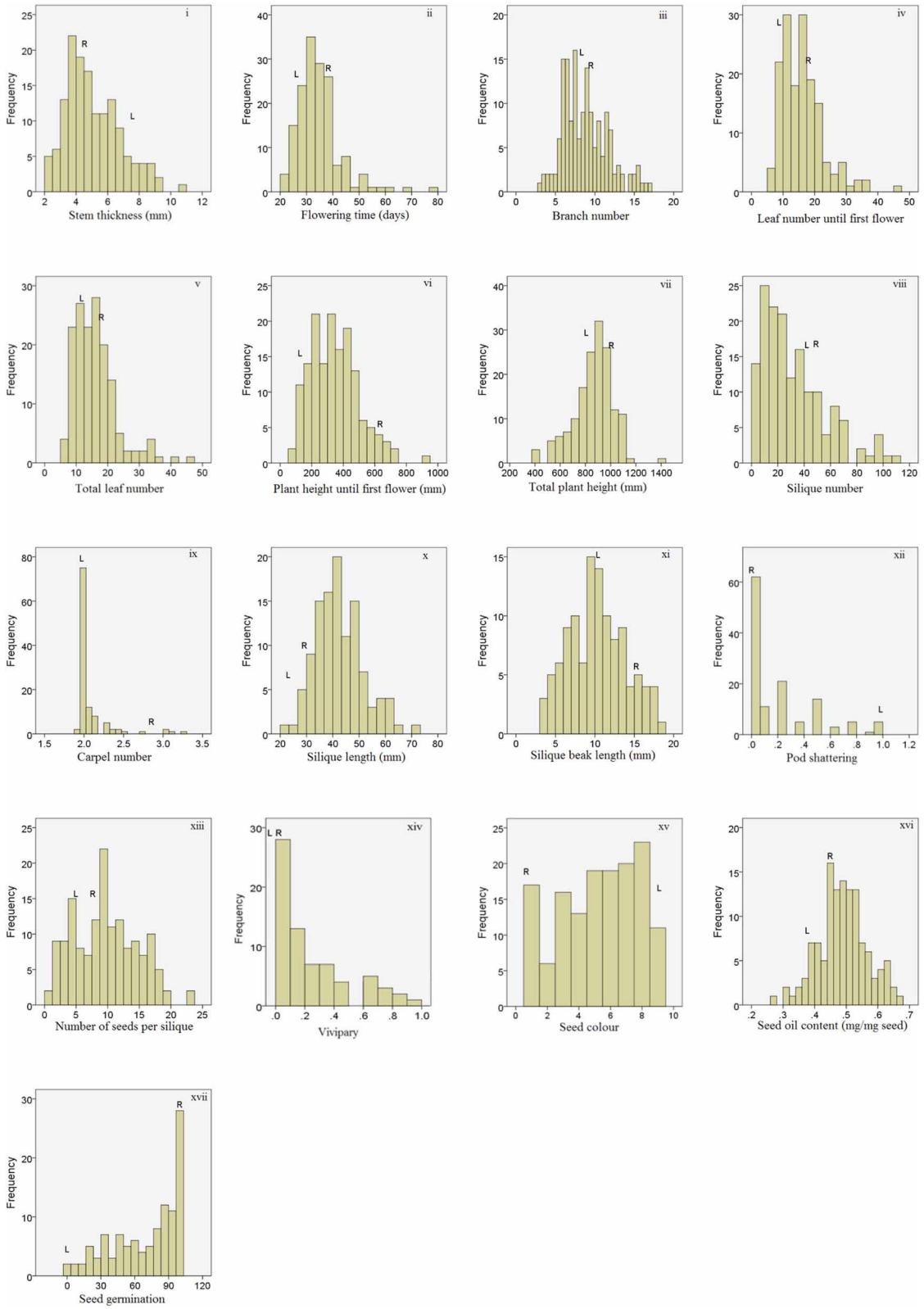


FIGURE 3 | Frequency distributions of non-normalized data of the reported traits for the L58 × R-o-18 RIL population. The vertical axes indicate the number of lines per trait value class and the horizontal axes indicate the different trait value classes. The parental values (indicated with L and R) are the mean of five replicates.

Table 1 | Phenotype data for both parental lines (L58 and R-o-18) and the RIL population, for the 17 analyzed traits.

Trait (unit)	Abbreviation	Parental lines			RIL population			Mean	SD	h^2
		L58	R-o-18	%	Min	Max	Range			
Stem thickness (mm)	*St*	7.5	4.5	60.0	2.0	10.8	8.7	5.1	1.7	0.35
Flowering time (days)	*Ft*	28.8	38.4	133.6	22.0	79.8	57.8	34.6	8.2	0.92
Branch number	*Bn*	8.4	8.9	105.9	3.0	17.0	14.0	8.7	2.9	0.69
Leaf number until first flower	*Lnf*	11.8	17.8	151.8	5.0	46.0	41.0	15.9	6.6	0.89
Total leaf number	*Tln*	13.9	17.8	128.1	5.5	46.8	41.3	15.9	6.9	0.89
Plant height until first flower (mm)	*Ph*	171	658	384	88	943	855	344	154	0.81
Total plant height (mm)	*Tph*	878	1078	123	398	1403	1005	864	156	0.76
Silique number	*Sin*	46.5	48.8	104.8	0.0	108.5	108.5	32.3	24.1	0.73
Carpel number	*Cn*	2.0	2.8	141.3	1.9	3.3	1.3	2.1	0.2	0.50
Silique length (mm)	*Sil*	22.9	34.6	151.3	23.3	70.9	47.6	42.5	9.1	0.78
Silique beak length (mm)	*Bl*	10.2	14.8	145.4	3.7	18.5	14.8	10.3	3.6	0.71
Pod shattering	*Sh*	1.0	0.0	0.0	0.0	1.0	1.0	0.2	0.3	0.50
Number of seeds per silique	*Nsps*	6.2	8.1	131.1	1.0	22.9	21.9	9.7	5.0	0.61
Vivipary	*Vi*	0.0	0.0	–	0.0	1.0	1.0	0.2	0.3	0.68
Seed color	*Sc*	9.0	1.0	11.1	1.0	9.0	8.0	5.3	2.5	–
Seed oil content (mg oil/mg seed)	*So*	0.4	0.4	112.8	0.3	0.7	0.4	0.5	0.1	–
Seed germination	*Sg*	0.0	100.0	–	0.0	100.0	100.0	68.1	29.8	–

"%" indicates the relative performance of R-o-18 compared to L58. "Min" and "Max" indicate the values of the RIL with respectively the lowest of the highest value, while "Range" indicates the difference between these values. "Mean" is the average value for all RIL lines, with standard deviation (SD), and h^2 is broad sense heritability. For all traits four replicate samples were measured, except for seed color, seed germination, and seed oil content, for which only one sample could be measured. All 160 lines have been scored.

Table 2 | Pearson correlations for the analyzed traits of the L58 × R-o-18 RIL population.

Trait	*St*	*Ft*	*Bn*	*Lnf*	*Tln*	*Ph*	*Tph*	*Sin*	*Cn*	*Sil*	*Bl*	*Sh*	*Nsps*	*Vi*	*Sc*	So	Sg
St	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ft	0.571**	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Bn	0.376**	0.382**	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Lnf	0.562**	0.847**	0.563**	1	–	–	–	–	–	–	–	–	–	–	–	–	–
Tln	0.566**	0.861**	0.557**	0.988**	1	–	–	–	–	–	–	–	–	–	–	–	–
Ph	0.183*	0.275**	0.354**	0.315**	0.289**	1	–	–	–	–	–	–	–	–	–	–	–
Tph	0.252**	−0.014	0.157*	0.016	0.012	0.629**	1	–	–	–	–	–	–	–	–	–	–
Sin	0.079	0.084	0.025	0.067	0.074	−0.109	−0.084	1	–	–	–	–	–	–	–	–	–
Cn	0.127	0.064	0.032	0.179	0.157	0.174	0.127	−0.025	1	–	–	–	–	–	–	–	–
Sil	−0.191*	−0.108	−0.078	−0.105	−0.123	0.196*	0.107	0.184	−0.181	1	–	–	–	–	–	–	–
Bl	−0.137	−0.003	−0.053	−0.028	−0.021	0.175	0.081	0.198*	−0.006	0.680**	1	–	–	–	–	–	–
Sh	0.053	0.072	0.021	0.033	0.054	−0.148	−0.028	0.419**	0.045	0.047	0.186	1	–	–	–	–	–
Nsps	0.249**	0.171*	0.120	0.139	0.143	0.085	0.097	0.367**	0.166	0.151	0.075	0.355**	1	–	–	–	–
Vi	−0.344**	−0.369**	−0.256*	−0.392**	−0.388**	−0.188	0.018	0.147	−0.150	−0.001	0.103	−0.015	−0.305*	1	–	–	–
Sc	−0.094	−0.075	−0.007	−0.121	−0.120	0.012	0.004	0.060	0.142	−0.166	0.021	0.238**	0.170*	−0.095	1	–	–
So	0.205*	0.285**	0.076	0.222*	0.221*	0.010	−0.112	0.345**	0.020	−0.040	−0.048	0.113	0.516**	−0.228	0.179*	1	–
Sg	0.017	−0.105	0.091	−0.104	−0.080	0.052	−0.020	0.032	0.024	−0.045	−0.154	−0.123	0.007	0.173	−0.127	−0.004	1

St, stem thickness; Ft, flowering time; Bn, branch number; Lnf, leaf number until first flower; Tln, total leaf number; Ph, plant height until first flower; Tph, total plant height; Sin, silique number; Cn, carpel number; Sil: silique length; Bl, silique beak length; Sh, Pod shattering; Nsps, number of seed per silique; Vi, vivipary; Sc, seed color; So, seed oil content; Sg, seed germination.

***means significant at $P \leq 0.01$; *significant at $P \leq 0.05$.*

The residual heterozygosity in the RIL population was not significantly higher than the expected value of 1.56%. Unintended selection during single-seed-descent propagation, for instance for plant size or fecundity, could lead to increased heterozygosity at some loci (Loudet et al., 2002). Since effort was made to randomly designate which plants would be selected at each propagation cycle, we were able to avoid this type of distortion in this population. The genetic map was constructed for the F7

Table 3 | QTLs detected for the analyzed traits in the L58 × R-o-18 RIL population.

Trait	QTL	Linkage group	LOD threshold	LOD	Position of LOD peak (cM)	R^2	Effect
Stem thickness	*St1*	A7	3	4.7	58.9	13.1	1.3
	St2	A4	–	2.8	7.6	8	1.1
Flowering time	*Ft1*	A7	2.8	5.4	34.9	11.2	5.5
	Ft2	A5	–	4.9	60.8	9	−5.4
	Ft3	A8	–	4.9	85.2	9	−5.1
	Ft4	A2	–	3.7	64.3	6.6	4.2
	Ft5	A7	–	3.1	106.3	6.5	−4.5
Branch number	*Bn1*	A2	3.1	5.5	64.3	13.1	2.1
	Bn2	A3	–	3.2	38.3	7.4	−1.7
	Bn3	A6	–	2.8	76.7	6.4	1.5
Leaf number until first flower	*Lnf1*	A8	2.9	8.8	91.3	15	−5.3
	Lnf2	A4	–	4.9	90.1	8	3.8
	Lnf3	A7	–	4.8	96.7	8	−3.9
	Lnf4	A2	–	3.9	64.3	6.2	3.4
	Lnf5	A5	–	3.2	35.2	5	−3.1
	Lnf6	A9	–	2.9	79.6	4.5	−2.8
Total leaf number	*Tln1*	A8	2.9	8.3	91.3	14	−5.3
	Tln2	A7	–	5.7	106.4	10.6	−4.8
	Tln3	A7	–	4.1	58.9	7.6	4.1
	Tln4	A2	–	4.6	64.3	7.3	3.8
	Tln5	A4	–	3.3	77.8	5.1	3.2
	Tln6	A3	–	3	38.3	4.6	−3.3
Plant height until first flower	*Ph1*	A10	2.8	4	59.2	9.3	95.4
	Ph2	A8	–	3.6	81.4	8.3	−90.2
	Ph3	A1	–	2.9	55.5	6.6	−85.2
Total plant height	*Tph1*	A3	3	5	38.3	11.3	−112.3
	Tph2	A10	–	4.1	64.4	9.5	96.1
	Tph3	A5	–	3.2	69.7	7.1	87.2
Silique number	*Sin*	A6	3	4	72.4	11	16.0
Carpel number	*Cn1*	A4	2.7	4.4	84.1	15.2	0.2
	Cn2	A2	–	2.5	17.5	8.5	−0.1
Silique length	*Sil*	A4	3	5	90.1	18.6	−8.1
Silique beak length	*Bl1*	A10	3	5.7	53.6	14.4	2.8
	Bl2	A8	–	4.4	0	11	−2.4
	Bl3	A4	–	4.1	90.1	10	−2.3
	Bl4	A1	–	3	72.9	7.2	−2.0
Pod shattering	*Sh*	A9	3	5.5	60	18	0.2
Number of seeds per silique	*Nsps1*	A9	3	3.5	58.1	10	3.1
	Nsps2	A3	–	2.7	82.6	7.2	−2.7
Vivipary	*Vi1*	A9	3	3.9	111.3	20.5	0.3
	Vi2	A6	–	2.6	61.1	13.3	−0.2
Seed color	*Sc1*	A9	3	30.8	56.6	53.7	3.7
	Sc2	A3	–	12.1	52.5	15	2.1
Seed oil content	*So*	A9	3	3	58.2	9.1	0.1
Seed germination	*Sg1*	A5	3.1	4.1	137	14.4	−23.2
	Sg2	A3	–	3.1	147.6	10.5	−22.1

Per trait, QTLs are numbered according to decreasing LOD score (LOD). LOD thresholds are calculated per trait based on 1000 permutation tests and an experimental error rate of $P < 0.05$. R^2 is the percentage of total phenotypic variance explained by each QTL. For each QTL, the allelic effect is calculated as μA-μB (μ = mean), where A and B are RILs carrying L58, respectively, R-o-18 alleles at the QTL.

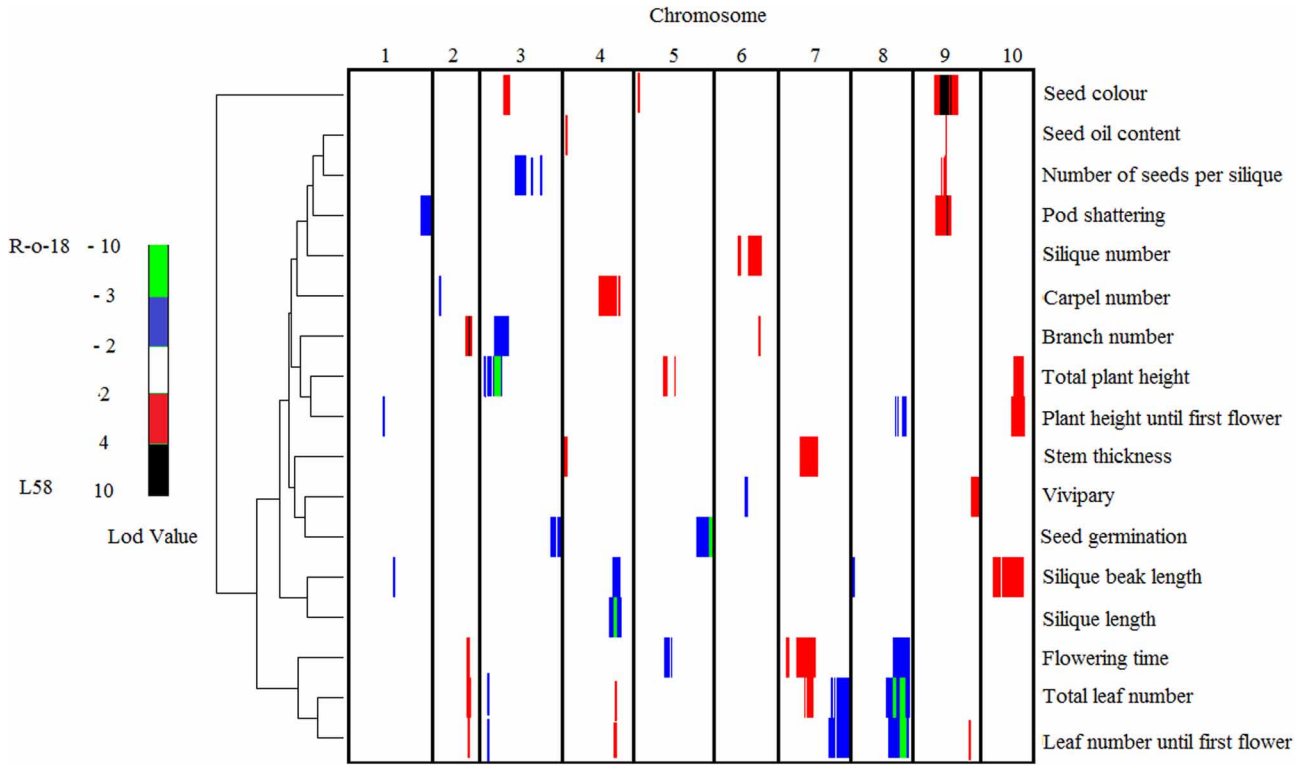


FIGURE 4 | A clustered heat map showing the LOD profiles of the measured traits. Columns indicate the 10 chromosomes in centimorgans, ascending from the left to right; rows indicate individual trait LOD profiles. A color scale is used to indicate the QTL significance corresponding to the LOD score. Positive values (red and black) represent a positive effect of the trait by the L58 allele, negative values (blue and green) represent a positive effect on the trait by the R-o-18 allele. The width of a bar indicates the significance interval of the QTL that was calculated by rMQM in MAPQTL 6. Hierarchical clustering, shown on the left, reflects the correlation between traits based on the QTL profiles.

RILs using a mix of AFLPs, SNPs, InDel, and SSR markers. The whole genome sequence information of *B. rapa* (Cheng et al., 2011; Wang et al., 2011a) ensured the correct genome location of the SNPs, SSR, and InDel markers for which primer sequences were available. This was very efficient in resolving any mapping ambiguities and in assigning chromosome numbers to linkage groups. The current map covers a total length of 1150 cM with an average resolution of 4.3 cM. This map is comparable to two *B. rapa* reference linkage maps based on DH populations, with a total length of 1182 cM (Choi et al., 2007) and 1234.2 cM (Wang et al., 2011b), and the map reported for another RIL population, of 1125 cM (Iniguez-Luy et al., 2009). The marker resolution of 4.3 cM per marker found for this population, is also in line with the reported maps. In some cases, composite interval mapping (CIM), which is one of the QTL mapping methods we used, can be affected by an uneven distribution of markers in the genome (Zeng et al., 1999), which is why non-informative markers were omitted if they did not detect additional recombination events, to keep the smallest informative marker set. Simulation studies have shown that the advantages of increasing marker density beyond one marker every 4.3 cM are less significant than those obtained when increasing the size of the population (Darvasi and Soller, 1994; Charmet, 2000). This means that with the current marker density, there is no need to screen for additional markers in order to improve mapping efficiency.

In total 47 QTLs for 17 analyzed traits were mapped. Seed coat color is a very important trait in *Brassica* oilseed crops. A yellow seed color is known to be highly correlated with meal quality, because of the thinner seed coat, corresponding to less anti-nutritive fiber components, which is also associated with higher protein content (Tang et al., 1997; Badani et al., 2006; Wittkop et al., 2009; Liu et al., 2012). Seed coat color is a maternally inherited trait, with the alleles for black seed coat acting dominantly over the alleles for yellow seed coat. In *B. napus*, seed color is inherited in different ways, probably depending on the source of the genetic variation, and is strongly affected by environmental factors (Liu et al., 2012). Earlier studies (Stringam, 1980) proposed a two locus model for seed color in *B. rapa*, involving the *Br1* and *Br2* loci, both of which were not mapped at the time. In *B. rapa* studies involving yellow sarson oilseed types, as used in this study, a major seed color locus is found on chromosome A9 (Lou et al., 2007). In this RIL population two major QTLs were detected for seed color, *Sc1* and *Sc2*, on A9 and A3 respectively, explaining about 70% of seed coat color variation. The *Sc1* locus on A9 co-located with previously reported seed color QTLs reported for both *B. napus* and *B. rapa* (Lou et al., 2007; Liu et al., 2012; Xiao et al., 2012). Near-infrared reflectance spectroscopy

measurements of acid detergent lignin (ADL) in seeds of both parental lines confirmed the expected difference in ADL corresponding to yellow and black seed (Snowdon, pers. communication) suggesting that the *B. rapa* locus we mapped to A9 affects the same gene as the A9 locus cloned from *B. napus*. This locus was found to harbor a mutation in the *CCR1* gene, encoding a cinnamoyl co-A reductase involved in lignin biosynthesis (Liu et al., 2012). In the absence of the L58 allele at *Sc1*, seeds containing the L58 allele at *Sc2* are brown, not yellow. Most of the cultivated *B. rapa* is brown-seeded, while for commercial purpose oilseed *B. rapa* with brown seeds is not preferred due to the darker coloring of the oil (Ramchiary and Lim, 2011). Introgression of the *Sc2* allele of yellow sarson types like R-o-18, could overcome this. In addition to the reported *Sc1* and *Sc2* loci, we found an additional, but very weak, *Sc3* QTL (LOD = 2.23), which mapped to A5 and accounted for 2% of the phenotypic variance. Previously, a QTL controlling yellow seed color was mapped to A5 (Teutonico and Osborn, 1994), which could concern the same locus.

Of the five QTLs detected for Ft, four Ft QTLs, *Ft1*, *Ft3*, *Ft4,* and *Ft5*, co-localized with previously mapped QTLs (Osborn et al., 1997; Lou et al., 2007; Edwards and Weinig, 2011; Lou et al., 2011). Another QTL, *Ft2* co-localized with a previously mapped, non-significant, QTL on A5 for a circadian clock parameter (Lou et al., 2011). Ft is highly co-related with plant architecture traits like Ph, Lnf, Tln, and Bn. *Ph2*, *Tln1,* and *Lnf1* co-localized with *Ft3*, while *Ft4* co-localized with *Tln4*, *Lnf4,* and *Bn1*, the only Bn QTL co-localizing with a Ft locus. Furthermore, *Ft5* co-localized with *Tln2 and Lfn3*; and finally *Ft1* co-localized with *Tln3*. *Lnf3* and *Lnf4* have been previously mapped by Lou et al. (2007), who also observed the general co-localization of *Ft* and *Lnf* loci. *Tln6* and *Lnf6* are two separate loci, mapping to A3 and A9 respectively, which did not co-localize with any *Ft* QTLs in this population, but which co-localized with *Ft* QTLs detected by Edwards and Weinig (2011).

Resistance to pod shattering is a recessive complex trait, mainly based on data from *B. napus*, which is difficult to assess because it can only be scored at maturity (Morgan et al., 2003). There are no reports related to *Brassica* loci controlling pod shattering, although work has been done on genetic engineering of pod shattering resistance, using ectopic expression of the *FRUITFULL* gene from Arabidopsis (Østergaard et al., 2006). The pod shattering QTL (*Sh*) on A9 is located in the same region as *Sc1*, but if indeed *Sc1* is caused by variation at the *CCR1* gene, as we expect, this is unlikely to be a pleiotropic effect of the same locus. Fortunately the alleles for black seeds and easy shattering are in coupling phase (**Table 2**), which means that selection for yellow-seeded lines could easily be accompanied by selection for improved shattering resistance. Since there is limited genetic variation for pod shattering resistance within the *B. napus* germplasm (Morgan et al., 2003), introducing pod shattering resistance alleles from *B. rapa* into a *B. napus* breeding program could well be an interesting approach.

Shattering has a significant positive correlation with the Nsps and the Sin. A significant *Sh* QTL co-located with *Nsps1*.The Nsps is also highly positively correlated with other silique related traits such as Sil and Bl. Therefore, Sil and Bl are likely to have an overall effect on silique-related traits. Sil and Bl shared one QTL, *Sil* and *Bl3* respectively. This co-localization is supported with high correlation between the two traits. Lou et al. (2007) reported two genomic regions on A1 and A7 and three loci on A5, A7, and A9, controlling Sil and Bl respectively. The *Sil* QTL on A4 reported here, is a new locus that explains 18.6% of the variance.

Vivipary (pre-harvest sprouting) is another important oilseed quality trait. A major QTL explaining 50.8% of the total variance for vivipary had previously been mapped to chromosome N11 of *B. napus* (Feng et al., 2009). We are not aware of previous work on seed vivipary QTLs in *B. rapa*. The two QTLs we detected on A9 and A6, explain about 30% of the vivipary variance. Vivipary is negatively correlated with seed oil content in our data. Also in *B. napus*, vivipary decreased seed viability and vigor and resulted in lower seed oil content (Ruan et al., 2008).

With the availability of the *Brassica rapa* genome sequence (Wang et al., 2011a) and the further development of molecular genetic tools based on the parental genotypes we used for the L58 × R-o-18 RIL population, we anticipate that the population can be a very useful additional tool to improve gene cloning approaches in *B. rapa* and thus contribute to more efficient *B. rapa* breeding.

ACKNOWLEDGMENTS

This work was financially supported by a personal grant to Hedayat Bagheri from the Ministry of Science, Research, and Technology of Iran, by the IOP Genomics project IGE050010 on Brassica Vegetable Nutrigenomics and by the Graduate School Experimental Plant Sciences. We acknowledge Dr. Xiaowu Wang and Dr. Lars Østergaard for their generous supply of seeds for the L58 and R-o-18 parental lines and we thank Dr. Rod Snowdon and Dr. Benny Wittkop from the Justus Liebig University in Giessen, Germany, for their help in analyzing L58 and R-o-18 seeds for lignin content and composition. The AFLP® and CRoPS® technologies are covered by patents and/or patent applications of Keygene N. V. AFLP, CRoPS, and KeyGene are registered trademarks of Keygene N. V. Other trademarks are the property of their respective owners.

REFERENCES

Akhunov, E., Nicolet, C., and Dvorak, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor. Appl. Genet.* 119, 507–517.

Badani, A. G., Snowdon, R. J., Wittkop, B., Lipsa, F. D., Baetzel, R., Horn, R., De Haro, A., Font, R., Lühs, W., and Friedt, W. (2006). Colocalization of a partially dominant gene for yellow seed colour with a major QTL influencing acid detergent fibre (ADF) content in different crosses of oilseed rape (*Brassica napus*). *Genome* 49, 1499–1509.

Beek, J. G., Verkerk, R., Zabel, P., and Lindhout, P. (1992). Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: Cf9 (resistance to *Cladosporium fulvum*) on chromosome 1. *Theor. Appl. Genet.* 84, 106–112.

Charmet, G. (2000). Power and accuracy of QTL detection: simulation studies of one-QTL models. *Agronomie* 20, 309–323.

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., and Wang, X. (2011). BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biol.* 11, 136.

Choi, S., Teakle, G., Plaha, P., Kim, J., Allender, C., Beynon, E., Piao, Z., Soengas, P., Han, T., King, G., Barker, G., Hand, P., Lydiate, D., Batley, J., Edwards, D., Koo, D., Bang, J., Park, B.-S., and Lim,

Y. (2007). The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. *Theor. Appl. Genet.* 115, 777–792.

Darvasi, A., and Soller, M. (1994). Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait loci. *Theor. Appl. Genet.* 89, 351–357.

Edwards, C. E., and Weinig, C. (2011). The quantitative-genetic and QTL architecture of trait integration and modularity in *Brassica rapa* across simulated seasonal settings. *Heredity (Edinb.)* 106, 661–677.

Feng, F., Liu, P., Hong, D., and Yang, G. (2009). A major QTL associated with preharvest sprouting in rapeseed (*Brassica napus* L.). *Euphytica* 169, 57–68.

Goossens, A., Dillen, W., De Clercq, J., Van Montagu, M., and Angenon, G. (1999). The arcelin-5 gene of *Phaseolus vulgaris* directs high seed-specific expression in transgenic *Phaseolus acutifolius* and Arabidopsis plants. *Plant Physiol.* 120, 1095–1104.

Hyten, D., Song, Q., Choi, I.-Y., Yoon, M.-S., Specht, J., Matukumalli, L., Nelson, R., Shoemaker, R., Young, N., and Cregan, P. (2008). High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theor. Appl. Genet.* 116, 945–952.

Iniguez-Luy, F., Lukens, L., Farnham, M., Amasino, R., and Osborn, T. (2009). Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa*and *B. oleracea*. *Theor. Appl. Genet.* 120, 31–43.

Kole, C., Kole, P., Vogelzang, R., and Osbom, T. C. (1997). Genetic linkage map of a *Brassica rapa* recombinant inbred population. *J. Hered.* 88, 553–557.

Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* 55, 141–172.

Liu, L., Stein, A., Wittkop, B., Sarvari, P., Li, J., Yan, X., Dreyer, F., Frauen, M., Friedt, W., and Snowdon, R. (2012). A knockout mutation in the lignin biosynthesis gene CCR1 explains a major QTL for acid detergent lignin content in *Brassica napus* seeds. *TAG Theor. Appl. Genet.* 124, 1573–1586.

Lou, P., Xie, Q., Xu, X., Edwards, C. E., Brock, M. T., Weinig, C., and McClung, C. R. (2011). Genetic architecture of the circadian clock and flowering time in *Brassica rapa*. *Theor. Appl. Genet.* 123, 397–409.

Lou, P., Zhao, J., Kim, J. S., Shen, S., Del Carpio, D. P., Song, X., Jin, M., Vreugdenhil, D., Wang, X., Koornneef, M., and Bonnema, G. (2007). Quantitative trait loci for flowering time and morphological traits in multiple populations of *Brassica rapa*. *J. Exp. Bot.* 58, 4005–4016.

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002). Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theor. Appl. Genet.* 104, 1173–1184.

Love, C. G., Graham, N. S., Ó Lochlainn, S., Bowen, H. C., May, S. T., White, P. J., Broadley, M. R., Hammond, J. P., and King, G. J. (2010). A *Brassica* exon array for whole-transcript gene expression profiling. *PLoS ONE* 5:e12812. doi: 10.1371/journal.pone.0012812

Morgan, C., Bavage, A., Bancroft, I., Bruce, D., Child, R., Chinoy, C., Summers, J., and Arthur, E. (2003). Using novel variation in Brassica species to reduce agricultural inputs and improve agronomy of oilseed rape—a case study in pod shatter resistance. *Plant Genet. Resour.* 1, 59–65.

Myburg, A. A., Remington, D. L., O'malley, D. M., Sederoff, R. R., and Whetten, R. W. (2001). High-throughput AFLP analysis using infrared dye-labeled primers and an automated DNA sequencer. *Biotechniques* 30, 348–352, 354, 356–357.

Osborn, T. C., Kole, C., Parkin, I. A., Sharpe, A. G., Kuiper, M., Lydiate, D. J., and Trick, M. (1997). Comparison of flowering time genes in *Brassica rapa, B. napus* and *Arabidopsis thaliana*. *Genetics* 146, 1123–1129.

Østergaard, L., Kempin, S. A., Bies, D., Klee, H. J., and Yanofsky, M. F. (2006). Pod shatter-resistant Brassica fruit produced by ectopic expression of the FRUITFULL gene. *Plant Biotechnol. J.* 4, 45–51.

Pink, D., Bailey, L., McClement, S., Hand, P., Mathas, E., Buchanan-Wollaston, V., Astley, D., King, G., and Teakle, G. (2008). Double haploids, markers and QTL analysis in vegetable brassicas. *Euphytica* 164, 509–514.

Price, J. S., Hobson, R. N., Neale, M. A., and Bruce, D. M. (1996). Seed losses in commercial harvesting of oilseed rape. *J. Agric. Eng. Res.* 65, 183–191.

Ramchiary, N., and Lim, Y. (2011). "Genetics of Brassica rapa L," in *Genetics and Genomics of the Brassicaceae, Plant Genetics and Genomics: Crops and Models,* Vol. 9, eds R. Schmidt and I. Bancroft (New York, NY: Springer), 215–260.

Reinhardt, D., and Kuhlemeier, C. (2002). Plant architecture. *EMBO Rep.* 3, 846–851.

Ruan, S. L., Hu, W. M., Duan, X. M., and Ma, H. S. (2008). Ultrastructural and electrophoretic analyses of viviparous and normal seeds in hybrid rape (*Brassica napus* L.). *Seed Sci. Technol.* 36, 371–378.

Stephenson, P., Baker, D., Girin, T., Perez, A., Amoah, S., King, G., and Ostergaard, L. (2010). A rich TILLING resource for studying gene function in *Brassica rapa*. *BMC Plant Biol.* 10, 62.

Stringam, G. R. (1980). Inheritance of seed color in turnip rape. *Can. J. Plant Sci.* 60, 331–335.

Tang, Z. L., Li, J. N., Zhang, X. K., Chen, L., and Wang, R. (1997). Genetic variation of yellow-seeded rapeseed lines (*Brassica napus L.*) from different genetic sources. *Plant Breed.* 116, 471–474.

Teutonico, R. A., and Osborn, T. C. (1994). Mapping of RFLP and qualitative trait loci in *Brassica rapa* and comparison to the linkage maps of *B. napus, B. oleracea* and *Arabidopsis thaliana*. *Theor. Appl. Genet.* 89, 885–894.

van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van Der Poel, H., Van Oeveren, J., Verstegen, H., and Van Eijk, M. J. T. (2007). Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2:e1172. doi: 10.1371/journal.pone.0001172

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78.

Voorrips, R. E., Jongerius, M. C., and Kanne, H. J. (1997). Mapping of two genes for resistance to clubroot (*Plasmodiophora brassicae*) in a population of doubled haploid lines of *Brassica oleracea* by means of RFLP and AFLP markers. *Theor. Appl. Genet.* 94, 75–82.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., and Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23, 4407–4414.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B.-S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A. P., Batley, J., Kim, J.-S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.-J., Choi, S.-R., Lee, T.-H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., and Zhang, Z. (2011a). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039.

Wang, Y., Sun, S., Liu, B., Wang, H., Deng, J., Liao, Y., Wang, Q., Cheng, F., Wang, X., and Wu, J. (2011b). A sequence-based genetic linkage map as a reference for *Brassica rapa* pseudochromosome assembly. *BMC Genomics* 12, 239.

Wittkop, B., Snowdon, R., and Friedt, W. (2009). Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe. *Euphytica* 170, 131–140.

Xiao, L., Zhao, Z., Du, D., Yao, Y., Xu, L., and Tang, G. (2012). Genetic characterization and fine mapping of a yellow-seeded gene in Dahuang (a *Brassica rapa* landrace). *Theor. Appl. Genet.* 124, 903–909.

Yan, J., Yang, X., Shah, T., Sánchez-Villeda, H., Li, J., Warburton, M., Zhou, Y., Crouch, J., and Xu,

Y. (2010). High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol. Breed.* 25, 441–451.

Zeng, Z. B., Kao, C. H., and Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genet. Res.* 74, 279–289.

Zhang, X.-Q., Li, C., Tay, A., Lance, R., Mares, D., Cheong, J., Cakir, M., Ma, J., and Appels, R. (2008). A new PCR-based marker on chromosome 4AL for resistance to pre-harvest sprouting in wheat (*Triticum aestivum* L.). *Mol. Breed.* 22, 227–236.

Zhang, X.-W., Wu, J., Zhao, J.-J., Song, X.-F., Li, Y., Zhang, Y.-G., Xu, D.-H., Sun, R.-F., Yuan, Y.-X., Xie, C.-H., and Wang, X.-W. (2006). Identification of QTLs related to bolting in *Brassica rapa* ssp. *pekinensis* (syn. *Brassica campestris* ssp. *pekinensis*). *Agric. Sci. China* 5, 265–271.

Zhao, J., Kulkarni, V., Liu, N., Pino Del Carpio, D., Bucher, J., and Bonnema, G. (2010). *BrFLC2* (*FLOWERING LOCUS C*) as a candidate gene for a vernalization response QTL in *Brassica rapa*. *J. Exp. Bot.* 61, 1817–1825.

Zhao, J., Wang, X., Deng, B., Lou, P., Wu, J., Sun, R., Xu, Z., Vromans, J., Koornneef, M., and Bonnema, G. (2005). Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. *Theor. Appl. Genet.* 110, 1301–1314.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2012; accepted: 26 July 2012; published online: 16 August 2012.
*Citation: Bagheri H, El-Soda M, van Oorschot I, Hanhart C, Bonnema G, Jansen-van den Bosch T, Mank R, Keurentjes JJB, Meng L, Wu J, Koornneef M and Aarts MGM (2012) Genetic analysis of morphological traits in a new, versatile, rapid-cycling Brassica rapa recombinant inbred line population. Front. Plant Sci. **3**:183. doi: 10.3389/fpls.2012.00183*
This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.