1   **Title** Using RNA-seq for genomic scaffold placement, correcting assemblies, and genetic map

2   creation in a common *Brassica rapa* mapping population.

3   **Author Information**

4   RJ Cody Markelz*[,2], Michael F Covington*[,1,2], Marcus T Brock [‡], Upendra K Devisetty*, Daniel

5   J Kliebenstein[†], Cynthia Weinig[‡], Julin N Maloof*

6   *University of California at Davis, Department of Plant Biology, Davis, CA 95616

7   [†]University of California at Davis, Department of Plant Sciences, Davis, CA 95616

8   ‡University of Wyoming, Department of Botany, Laramie, WY, 82072

9   [1]Amaryllis Nucleics, Inc. Berkeley, CA, 94710

11   [2]These authors contributed equally to this work.

12   **Data Availability**

13   Sequence Read Archive- SRP022220

14

1   **Running Title**

2   *Brassica rapa* genome improvements using RNA-seq

3   **Key Words**

4   RNA-seq, genetic map, *Brassica rapa*, genome assembly correction

5   **Corresponding Author**

6   Julin N. Maloof

7   jnmaloof@ucdavis.edu

8   Department of Plant Biology, LS 1002

9   University of California, Davis

10   1 Shields Ave

11   Davis, CA 95616

12

1 **Abstract**

2 *Brassica rapa* is a model species for agronomic, ecological, evolutionary and translational

3 studies. Here we describe high-density SNP discovery and genetic map construction for a

4 *Brassica rapa* recombinant inbred line (RIL) population derived from field collected RNA-seq

5 data. This high-density genotype data enables the detection and correction of putative genome

6 mis-assemblies and accurate assignment of scaffold sequences to their likely genomic locations.

7 These assembly improvements represent 7.1-8.0% of the annotated *Brassica rapa* genome. We

8 demonstrate how using this new resource leads to a significant improvement for QTL analysis

9 over the current low-density genetic map. Improvements are achieved by the increased mapping

10 resolution and by having known genomic coordinates to anchor the markers for candidate gene

11 discovery. These new molecular resources and improvements in the genome annotation will

12 benefit the *Brassicaceae* genomics community and may help guide other communities in fine-

13 tuning genome annotations.

14

1 **INTRODUCTION**

2 The *Brassica* genus is important for human diets throughout Asia, providing micronutrients and

3 up to 12% of oil calories and a wide diversity of agricultural products (Dixon, 2007; Wang *et al.*,

4 2011b). Within this genus genome sequences have recently been published for *Brassica napus*,

5 *Brassica rapa*, and *Brassica oleracea* (Chalhoub *et al.*, 2014; Liu *et al.*, 2014; Wang *et al.*,

6 2011b). *Brassica rapa* is a physiologically and morphologically diverse diploid species that has

7 87% gene exon similarity to the model plant *Arabidopsis thaliana* (Cheng *et al*. 2013). This

8 makes *Brassica rapa* an excellent species for comparing and translating knowledge of biological

9 processes from Arabidopsis to a crop species. For example, homologous Arabidopsis gene

10 information has been used to infer the action of *B. rapa* genes in glucosinolate metabolism (Li

11 and Quiros, 2001; Wang *et al.*, 2011a), flowering time, leaf development (Baker *et al.*, 2015),

12 and seed yield (Brock *et al.*, 2010; Dechaine *et al.*, 2014). All of these important traits contribute

13 to our understanding of plant growth in agricultural settings and the underlying genetic

14 understanding of these traits is made possible by a reference genome sequence (Wang *et al.*,

15 2011b), gene annotation information (Cheng *et al.*, 2013; Devisetty *et al.*, 2014), and genetic

16 mapping populations (e.g. Iniguez-Luy *et al.* 2009).

17 The annotated *Brassica rapa* genome assembly is 283.8 Mb spread over 10 chromosomes A01-

18 10 (Wang *et al*. 2011b). Although the current genome is diploid, there are three ancient

19 subgenomes derived from genome duplication events. These subgenomes are designated as least

20 fractionated (LF), most fractionated one (MF1), and most fractionated two (MF2) corresponding

21 to the fraction of gene loss in each subgenome (Cheng et al., 2012; Wang et al., 2011b). These

22 three subgenomes share many paralogous genes and contiguous regions complicating genome

23 assembly. This has prevented about 10.8% of the gene-containing genomic scaffolds in version

4

1   1.5 of the genome (http://brassicadb.org/brad/index.php) from being assigned to chromosomes.

2   The lack of chromosomal assignment is largely because these scaffolds have no molecular

3   markers that would have enabled their placement on the genetic map. This suggests that

4   identifying more markers can help to make the *B. rapa* genome assembly more comprehensive

5   (Wang *et al.*, 2011b).

6   For this study we utilized an existing RIL population of *Brassica rapa* that has been used

7   extensively for QTL mapping of physiological, developmental, and evolutionarily important

8   traits (Bra-IRRI; Baker et al., 2015; Brock et al., 2010; Dechaine et al., 2007, 2014; Edwards et

9   al., 2011; Iniguez-Luy et al., 2009; Lou et al., 2011, 2012). Recently, we completed deep RNA-

10 sequencing of the parents of the Bra-IRRI population providing a large SNP set and improved

11 gene annotation information (Devisetty *et al.*, 2014). Using a new set of RNA-seq data collected

12 on the entire population, we extend these SNP discovery methods to 124 genotypes in the

13 population for placing scaffolds, correcting assemblies, and the creation of a saturated genetic

14 map.

15 **MATERIALS AND METHODS**

16 **Plant Growth and Tissue Collection**

17 The field site for plant growth was located at the University of Wyoming Agricultural

18 Experimental Station in Laramie, Wyoming, USA. This study focused on 124 RILs and the two

19 parental genotypes (R500 and IMB211) of the *Brassica rapa* IRRI population. Individual

20 replicates of each RIL were sown into peat pots filled with field soil and topped with 1 cm LP5

21 potting soil (Sun Gro Horticulture, Agawam, MA, USA). Seeds were planted in the first week of

22 June 2011, and pots were transplanted to the field 2.5 weeks later following established protocols

23 (Dechaine *et al.*, 2014). One biological replicate of each genotype was planted into each of 5

1   fully randomized blocks. After plants were established in the field for three weeks, apical

2   meristem tissue was collected from individual replicate plants into 1.5 mL Eppendorf tubes,

3   immediately flash frozen in liquid nitrogen, and stored at -80 ºC until RNA-Seq library

4   preparation.

5   **RNA-Seq library preparation and sequencing**

6   RNA-Seq libraries were prepared using a high-throughput Illumina RNA-Seq library extraction

7   protocol (Kumar *et al.*, 2012). The enriched libraries were then quantified on an Analyst Plate

8   Reader (LJL Biosystems) using SYBR Green I reagent (Invitrogen). Once the concentration of

9   libraries was determined, a single pool of all the RNA-Seq libraries within each block was made.

10   The pooled libraries were run on a Bioanalyzer (Agilent, SantaClara) to determine the average

11   product size for each pool. Each pool was adjusted to a final concentration of 20 nM and

12   sequenced on 7 lanes on Illumina Hi-Seq 2000 flow cell as 50-bp single end reads. Any failed

13   samples from the 5 blocks were run on 2 additional lanes.

14   **RNA-Seq Read Processing**

15   Pre-processing and mapping of Illumina RNA-Seq raw reads was done as described in detail in

16   Devisetty *et al.* 2014 with one exception. The raw reads were quality filtered with FASTX tool

17   kit's (http://hannonlab.cshl.edu/fastx_toolkit/) fastq_quality_filter with parameters [-q 20, -p

18   95]. The qualified de-multiplexed reads were then mapped to *B. rapa* reference genome (Chiifu

19   version 1.5) using BWA v0.6.1-r104 (Li and Durbin, 2009) with parameters [bwa_n 0.04] and

20   the unmapped reads were in turn mapped with TopHat with parameters [splice-mismatches 1,

21   max-multihits 1, segment-length 22, butterfly-search, max-intron-length 5000, library-type fr-

22   unstranded]. Finally, mapped reads from both BWA and TopHat were combined for genotyping

23   purposes.

1   **Population-based Polymorphism Identification**

2   Variant Call Format (VCF) files were generated for each of five replicate blocks of samples

3   using samtools and bcftools. These tools were run as 'samtools mpileup -E -u -f

4   Brapa_sequence_v1.5.fa [all alignment files for the current block] | bcftools view -bvcg - |

5   vcfutils.pl varFilter'.

6   The VCF files were summarized using 'summarize-vcf.pl' Perl script

7   (https://github.com/mfcovington/snps-from-rils). For each block of replicates, this script (run

8   using the parameters: '--observed_cutoff 0.3 --af1_min 0.3') ignores INDELs and variant

9   positions with more than two alleles, ignores variants with site allele frequency (AF1) values too

10   far from 0.5 ($>= 0.7$ or $<= 0.3$), and ignores variants with missing information in 30% or more

11   of the population. For variants that passed these filters, the numbers of reads matching the

12   reference and the number of alternate allele reads were recorded in a VCF summary file.

13   These VCF summary files from the different replicate blocks were merged using the 'merge-vcf-

14   summaries.pl' (https://github.com/mfcovington/snps-from-rils) Perl script. Using the default

15   parameters ('--replicate_count_min 2 --ratio_min 0.9'), this script merges the information in the

16   VCF summaries and records a putative SNP as an actual SNP if the variant is identified as a SNP

17   in at least 2 replicate blocks and if the proportion of reads matching the major allele is at least

18   0.9. This was done on a RIL by RIL basis.

19   **Genotyping, Plotting, and Identification of Genotype Bins**

20   The Perl script 'extract+genotype_pileups.pl' (https://github.com/mfcovington/detect-boundaries)

21   was used with the '--no_nr' parameter to extract genotype information from the RNA-seq

7

1   alignments at each SNP location for each member of the RIL population. The resulting genotype

2   files were used to detect and remove SNPs with excessive noise.

3   Due to the crossing scheme used to create the RIL population, each individual is expected to be

4   nearly homozygous for one parent or the other. The 'filter-noisy-SNPs.pl'

5   (https://github.com/mfcovington/noise-reduction-for-snps-from-pop) Perl script performs noise-

6   reduction for SNPs derived from such a population. It does this by identifying and ignoring

7   positions that have an over-representation of heterozygosity in individual lines across the entire

8   population. Using the default parameters ('--cov_min 3 --homo_ratio_min 0.9 --

9   sample_ratio_min 0.9'), SNPs were discarded as noisy if more than 10% of the lines in the

10   population showed evidence of heterozygosity as defined by a line having at least 3 reads per

11   SNP position with a major allele with a ratio less than 0.9.

12   After noise-reduction, the 'extract+genotype_pileups.pl'

13   (https://github.com/mfcovington/SNPTools) Perl script was re-run without the '--no_nr'

14   parameter for each RIL. The resulting genotype files were used to create genotype plots using the

15   'genoplot_by_id.pl' Perl script (https://github.com/mfcovington/SNPTools) and to define

16   genotype bins for the individual RILs.

17   The 'filter-snps.pl' Perl script (https://github.com/mfcovington/detect-boundaries) was used to

18   identify regions of adjacent SNPs with alleles from the same genotype. Using the default

19   parameters ('--min_cov 10 --min_momentum 10 --min_ratio 0.9 --offset_het 0.2'), it detects

20   boundaries between genotype bins when a sliding window of at least 10 SNPs. Within each

21   sliding window a depth of at least 10 reads each exhibit major allelic ratios of at least 0.9. The

22   major allele represents the opposite genotype from the previous bin (or exhibit major allelic

8

1   ratios less than 0.7 for transitions from regions of homozygosity to those of heterozygosity). For

2   each member of the RIL population, this script generates one file with boundary between

3   genotype bins.

4   The 'fine-tune-boundaries.pl' Perl script (https://github.com/mfcovington/detect-boundaries) is an

5   automated tool for rapid, fine-scale human curation of boundaries between genotype bins that we

6   used for the RIL population. As described in Devisetty et al. 2014, "This command-line tool

7   displays color-coded genotype data together with the currently-defined bin boundaries. Using

8   shortcut keys, the operator can quickly and easily approve or fine-tune a boundary (at which

9   point, the next boundary is instantly displayed for approval)."

10   The 'merge-boundaries.pl' Perl script (https://github.com/mfcovington/detect-boundaries) was

11   used to merge all of the boundaries in the collection of the boundaries files that were generated

12   by 'filter-snps.pl' and 'fine-tune-boundaries.pl'. A comprehensive list of bins and their locations

13   resulting from the merge are written to a file: bins.tsv. The script also prints the boundary and

14   bin stats (count, min size, max size, and mean size) to the screen to allow visual analysis of the

15   resulting file. This information was used for human curation of the boundaries.

16   The 'get-genotypes-for-bins.pl' Perl script (https://github.com/mfcovington/detect-boundaries)

17   was used to convert the comprehensive bins file and all the individual boundaries files into a

18   summary of bins and their locations across the genome and their genotypes across the entire RIL

19   population (Supplemental Table 1).

20   Composite genotype plots (Figure 3) were created using the 'plot-composite-map.R' R script

21   (https://github.com/mfcovington/detect-boundaries).

9

1   **Validating and Reassigning Genomic Scaffolds**

2   Using the genotypic value for each genotype bin across the RILs, we calculated the asymmetric

3   binary distance between all central SNP pairs using the dist(method = "binary") function in R.

4   The pairwise correlation matrix was then ordered by maximal correlations to place the map in a

5   linear order and compared to the predicted bin order based on version 1.5 of the *Brassica rapa*

6   genome. Comparisons between v1.5 of the genome and binary distance plots were manually

7   inspected to ensure proper placement or reassignment.

8   **Genetic Map Construction**

9   Because each genotype bin across the RILs represents each observed recombination breakpoint

10   in the population, we used one SNP per genotype bin to create a saturated genetic map. Aside

11   from the possibility of rare, unobserved double cross over events, the mapping resolution in this

12   population is no longer limited by the number of SNPs but instead by recombination events. The

13   genetic map was constructed using the chromosomal position of each of the SNPs as a starting

14   point for marker ordering along the chromosomes. Each chromosome was treated as a large

15   linkage group and each SNP was tested for linkage disequilibrium with all other SNPs using the

16   R/QTL package (Broman *et al*., 2003) in the R statistical environment (R Core Team, 2015).

17   Larger gaps in RNA-seq information corresponding to low gene density centromeric regions

18   were problematic when ordering markers using the ripple() R/QTL function (Broman *et al*.,

19   2003). In chromosomes A08 and A09, after local marker order was established we used the

20   physical position of the SNPs to connect the two arms in the correct orientation.

21   **QTL Comparisons**

22   To test how increased marker coverage affected QTL mapping and identification for

23   physiological traits, we remapped two traits from (Brock *et al*., 2010) that had been mapped

10

1   using the previous genetic map (Iniguez-Luy *et al*., 2009). We used R/QTL (v1.39-5) to compare

2   mapping results derived from the previous and updated genetic maps using the Brock *et al*.

3   (2010) flowering time phenotype data. Specifically we used the cim() function with three marker

4   covariates and determined LOD significance cutoffs after 1,000 permutations.

5   **RESULTS AND DISCUSSION**

6   **R500 vs. IMB211 polymorphism identification**

7   We performed deep RNA sequencing of 124 individuals of a RIL population derived from a

8   cross between the *Brassica rapa* accessions R500 and IMB211 (Iniguez-Luy *et al*., 2009). We

9   sequenced five replicates of each RIL and mapped 5.26 million reads mapped per RIL. We had

10   previously identified SNPs and INDELs between R500 and IMB211, the parents of the

11   population (Devisetty *et al*. 2014) using v1.2 of the *Brassica rapa* genome. This set of R500 vs.

12   IMB211 polymorphisms was used to genotype each member of the RIL population individually.

13   The crossing scheme used to create the RIL population should create homozygous regions of

14   contiguous R500 alleles alternating with homozygous regions of contiguous IMB211 alleles in

15   the different RILs. However, when using the R500 vs. IMB211 polymorphism set to genotype

16   the RILs there were multiple regions where R500 and IMB211 alleles were randomly

17   interspersed. This suggested that the RIL population might be derived from a different parent or

18   parents than those that we had sequenced.

19   To test this hypothesis, we merged the sequence data from all RILs and then genotyped the

20   merged dataset using SNPs identified by the IMB211 vs R500 comparison (Figure 1A). The

21   merged dataset gave us a much better view of segregation of putative parental SNPs in this

22   population. Given the relatively large size of the population and the expected recombination

23   frequency and distribution, polymorphisms identified in the true RIL parents should be

11

1   segregating with approximately equal allelic frequency in this merged data set (black dots in

2   Figure 1A). Most genomic regions did display this expected distribution; however, there were

3   several large regions that were not segregating, but instead were monomorphic for one of the

4   putative parents of the population (indicated as orange or blue dots in Figure 1A). In other words,

5   SNPs identified as polymorphic between the R500 and IMB211 strains are not segregating in the

6   RILs. Nearly all of these monomorphic regions matched R500 alleles, consistent with the idea

7   that the IMB211 seed strain is not the true parent of the RIL population.

8   The primary exception to the expected, Mendelian parental allele frequency in the RILs is on the

9   bottom of chromosome A03, where there is a gradual transition from equal R500:IMB211 allelic

10   frequency to nearly all IMB211. The A03 pattern is consistent with segregation distortion within

11   the population possibly caused by the centromere being located at that end of chromosome A03

12   (Cheng *et al*. 2013). With the lower recombination frequencies commonly observed near

13   centromeres in plants (Harushima *et al*., 1998; Haupt *et al*., 2001; Sherman and Stack, 1995),

14   there could be a meiotic drive allele or a local inversion in this region causing the segregation

15   distortion and this effect could be enhanced by the proximity to the centromere. There is

16   evidence for each of these mechanisms occurring across a wide range of plant species (Buckler

17   *et al*., 1999; Fang *et al*., 2012; Lowry and Willis, 2010).

18   **Population-based SNP discovery**

19   Due to the uncertainty surrounding the identity of the IMB211 parent of the RIL population, we

20   switched to a population-based approach for SNP discovery. This new strategy involved

21   identifying variants within the RIL population and using the R500 data to assign parental origin

22   for each SNP. Using this approach, we identified 146,027 SNPs across *B. rapa*'s ten

23   chromosomes (Table 1, Supplemental Table 1). These population-based SNPs segregate at the

12

1   expected allele frequencies of approximately 50/50 throughout the entire genome except at the

2   previously noted end of A03 (Figure 1B). Over 80% of the genome is within 100 kb of a SNP;

3   however, there are several regions with few or no SNPs. There are two primary reasons for these

4   SNP-free regions. Most are likely gene-poor regions or regions of genes with insufficient

5   expression under our experimental conditions (e.g., growth conditions, age, tissue, genotypes).

6   We also found a few regions where there are significant numbers of expressed genes, but no

7   SNPs between members of the RIL population. These regions primarily correspond to the non-

8   variant regions of Figure 1A and, therefore, likely represent regions that are very similar between

9   the seed stocks used to generate this RIL population.

10   **Genotyping the RIL population**

11   Using the per line transcriptomic data, each RIL was genotyped as having either the R500 or

12   IMB211 allele at each of the 149,097 SNPs identified from the population-based SNP-discovery

13   pipeline. An representative RIL genotype plot is shown in Figure 2.

14   **Collapsing Adjacent SNPs into Population-Wide Genotype Bins**

15   The next step towards creating a new genetic map was to define the largest set of non-redundant

16   SNPs. This is necessary because the 149,097 SNPs in the full dataset vastly exceed the expected

17   number of recombination breakpoints in a population of 124 individuals. We developed a

18   method to identify and summarize the "genotype bins" in the population. First, we found all

19   detectable recombination breakpoints for each RIL. Next, we consolidated these breakpoints for

20   the entire population. SNPs that were not adjacent to a recombination breakpoint in any of the

21   RILs were considered redundant and removed. This yielded bins of adjacent SNPs with

22   genotype patterns that differed from neighboring bins for at least one RIL because of a

13

1   recombination event in that specific RIL. The genotype bins for the RIL population are

2   summarized in a composite population genotype map (Figure 3).

3   **Finding and Reassigning Misassembled Genomic Regions**

4   A first version of the population genetic map revealed several markers that seemed to be

5   misplaced based on physical position, resulting in large genetic distances between them

6   (Supplemental Figure 2). Given, that we have corrected the parental genotyping issues, the

7   mostly likely explanation for this finding is that these regions represent genome assembly errors.

8   To test this hypothesis, the genotypes of representative SNPs from each bin were used to

9   calculate the asymmetric binary distances between each bin across the population. If the

10   predicted genome position of each bin is correct, the expectation is that each SNP should have

11   the lowest distance to adjacent SNPs in genome coordinates. However, consistent with genome

12   assembly problems, there were a subset of SNPs whose genotypes were more highly correlated

13   with SNPs located elsewhere in the genome rather than with SNPs near their current assigned

14   genomic position (a representative example is shown in Figure 5). To fix these issues, 13 regions

15   consisting of 19 genotypic bins were moved to different genomic locations and 4 regions

16   consisting of 66 bins were inverted in place at their original position based on assymetric binary

17   distance (Supplemental Table 2).

18   **Incorporating scaffold sequences into the genome**

19   In the current version of the *B. rapa* genome annotation (v1.5) there are 40,357 scaffolds that

20   have not been incorporated into any of the ten chromosomes. These scaffolds range in size from

21   100 bp to 938 Kbp and represent 1,411 genes spanning 27.5 Mbp. For comparison, there are

22   39,609 genes within the 283.8 Mbp of annotated chromosomal sequence. Given that the

23   scaffolds contain about as many genes as would be expected on one third of an average

14

1 chromosome, we decided to extend our strategy for fixing genome misassembles to estimate the

2 approximate chromosomal locations of the scaffolds.

3 We identified 3,070 SNPs across 339 of the 40,357 scaffolds (the remaining scaffolds had no

4 SNPs, Supplemental Table 3). To be confident in our placement we limited ourselves to the 47

5 scaffolds with 10 or more SNPs. For each of these 47 scaffolds, we were able to identify at least

6 one genomically defined chromosomal bin that had identical or near identical genotypes. This

7 indicates a very close genetic linkage between the unplaced scaffold and the placed genomic bin

8 allowing us to assign a genomic position to the unplaced scaffold, but not the orientation. The

9 incorporated scaffolds range in size from 429 to 884,746 bp and are enriched for larger scaffolds

10 (Figure 4). The incorporation of these 47 scaffolds allowed us to incorporate 25% (~ 7Mbp) of

11 the unplaced genomic sequence into the genome, representing 49% (691) of the unplaced

12 scaffold genes (Table 2; Supplemental Table 2).

13 While most of the incorporated scaffolds represent a single genotype bin, seven scaffolds are

14 comprised of multiple bins. Scaffold000164, for example, includes 65 annotated genes across six

15 distinct genotype bins within its 313.7 Kbp sequence. For six of the scaffolds with multiple bins

16 the bins were closely linked and allowed us to place the scaffold in a single location in the

17 genome. However, one scaffold, Scaffold000191, contained two bins that mapped to two

18 different chromosomes, indicating that it was misassembled. Therefore, we split its two bins and

19 assigned them to the appropriate chromosome locations (5 genes/28.2 Kbp to A01 and 24

20 genes/104.1 Kbp to A05)

21 Possible reasons for the enrichment of larger scaffolds within the set of incorporated scaffolds

22 include: (1) larger scaffolds are more likely to include expressed genes and, therefore, SNPs that

15

1   we can detect and (2) larger scaffolds may be more likely to be accurate representations of a

2   contiguous region within the genome. This second point is based on the assumption that large

3   scaffolds could be assembled perhaps due to more abundant, more consistent, and/or more

4   convincing experimental support than small scaffolds. Before (Figure 6A) and after (Figure 6B)

5   plots of genome-wide asymmetric binary distances for each marker pair show that rearranging

6   putative genomic misassemblies and incorporating scaffolds eliminates inconsistencies between

7   genome position and genotypes of adjacent markers.

8   **High-density genetic map**

9   From the available SNP data we were able to create a genetic map with ten linkage groups

10   corresponding to the 10 chromosomes of *Brassica rapa*. The map contains 1482 genotyped

11   markers for 124 RILs and is effectively saturated based on recombination events existing in the

12   population (Supplemental Table 4). The new map has an average marker spacing of 0.7 cM and

13   a total map distance of 1,045.6 cM. For comparison, the original map contained 225 markers

14   with an average spacing of 3.3 cM (Iniguez-Luy et al., 2009). This is also compared to a recent

15   map created on a subset of the population, 67 RILs, that had a total of 125 markers derived from

16   microarray probes (Hammond et al., 2011). Having the genetic distance of markers with known

17   genomic coordinates allowed us to fix two additional genome misassembles resulting in large

18   inversions on chromosomes A09 and A10 (Figure 7). All of these improvements combined allow

19   us to more accurately map QTL for known phenotypes such as flowering time (Figure 8). Lastly,

20   we fit spline based regressions for each chromosome to more accurately convert between genetic

21   distance and physical distance (A01 example; Figure 7B, C). These conversion equations are

22   helpful for finding candidate genes in significant QTL regions (Fulop *et al*., 2016).

23   **CONCLUSIONS**

16

1   In this study we demonstrated the flexibility and power of thoughtfully designed RNA-seq

2   experiments from tissue collected from a field experiment. RNA is a rich source of biological

3   information that can be utilized beyond expression analysis and transcriptome annotation. It is

4   our hope that these new community resources using RNA-seq are used to further genome

5   annotation, assembly, and functional analysis of the emerging model crop *Brassica rapa*.

6   **Conflicts of interest**

7   The authors declare no conflicts of interest.

8   **Acknowledgments**

12   **Supporting Information**

13   Supporting Code for genetic map construction can be found at:

14   https://github.com/rjcmarkelz/brassica_genetic_map_paper

15   **Supplemental Data**

16   **Supplemental Figure 1-** Genetic Map before misplaced markers reassigned.

17   Supplemental Table 1- RIL population SNP genotyping and genomic position.

18   Supplemental Table 2- RIL population genetic bins and scaffold original and final positions.

19   Supplemental Table 3- SNP base pair calls on scaffolds.

20   Supplemental Table 4- Final genetic map of RIL population.

1   **TABLES**

2   **Table 1.** SNP counts at different steps of the SNP discovery pipeline. The percentage of SNPs

3   located on chromosomes or scaffolds remaining after each step are shown in parentheses. The

4   first percentage is relative to the initial set of SNPs and the second percentage is relative to the

5   set of SNPs from the previous step.

	Chromosomes	Scaffolds	
	203,235	5,618	Identified within RIL population
	176,627 (87%)	4,640 (83%)	Passed conflict removal and repeat count filtering
	158,369 (78%, 90%)	3,737 (67%, 81%)	Have sequence information available for the R500 parent
Final Number of SNPs:	**146,027 (72%, 92%)**	**3,070 (55%, 82%)**	Passed noise-reduction filter

6

7

18

1 **Table 2.** Incorporated scaffolds represent a disproportionately high amount of scaffold sequence.

2 Percentages of scaffold subset counts and total lengths relative to the set of all scaffolds are

3 shown in parentheses.

	Count	Total Length (bp)	Mean Length (bp)	Median Length (bp)
Incorporated Scaffolds	47 (0.1%)	6,927,293 (25.1%)	147,389	58,889
Unincorporated Scaffolds	40,310 (99.9%)	20,655,028 (74.9%)	512	140
All Scaffolds	40,357	27,582,321	683	140

4

5

1   **FIGURES**
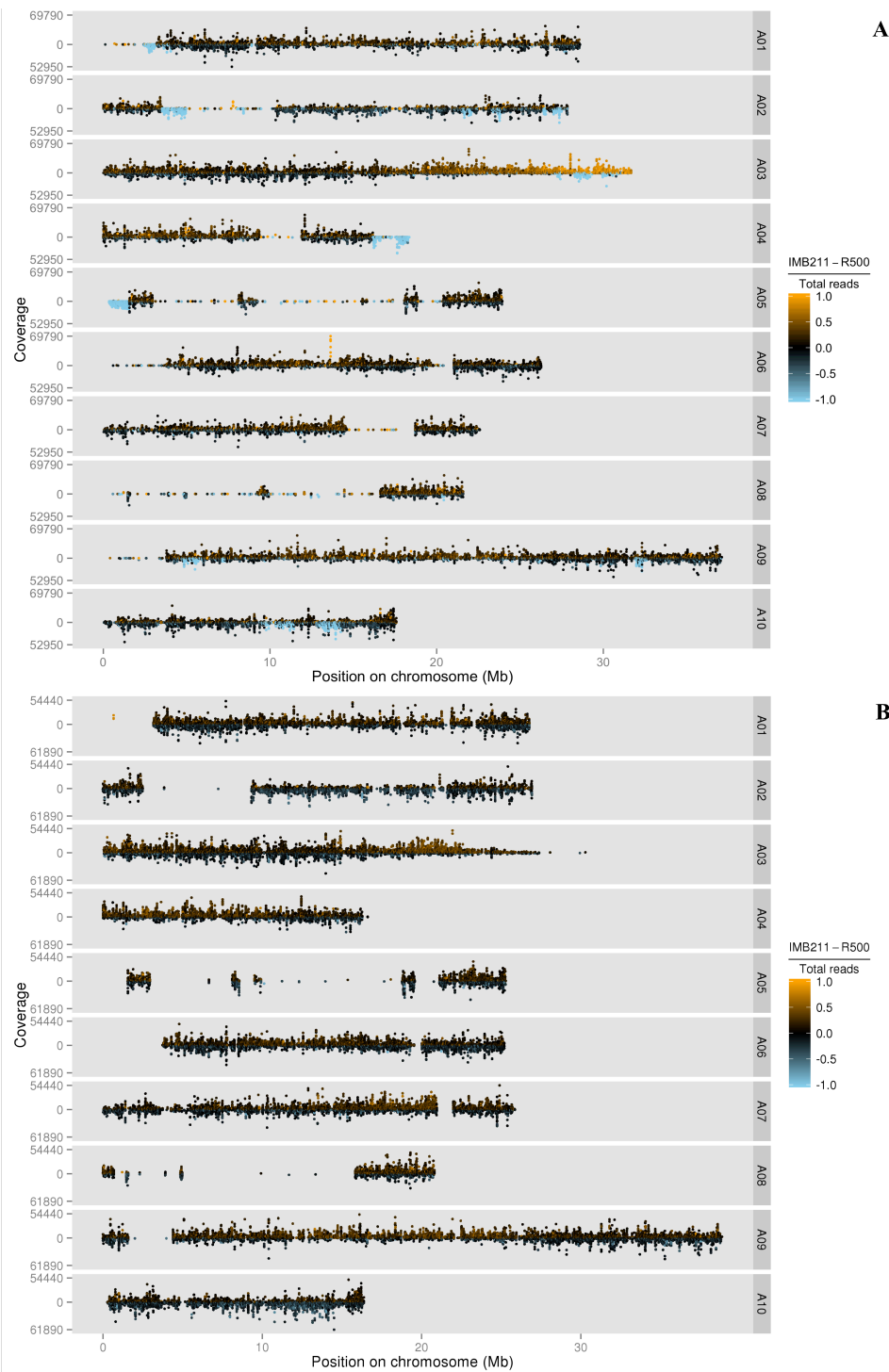
2

3   **Figure 1.** Plot of merged data from all RILs genotyped using the parent-based SNP set (A) and

4   the population based SNP set (B). Each of the *B. rapa* ten chromosomes are displayed (A01-

1   A10) with counts coverage of each SNP at each physical position on the chromosome in

2   megabases (Mb). The color indicates the relative ratio of coverage between R500 and IMB211

3   for every SNP. Black is equal coverage, orange is more IMB211 and blue is more R500.
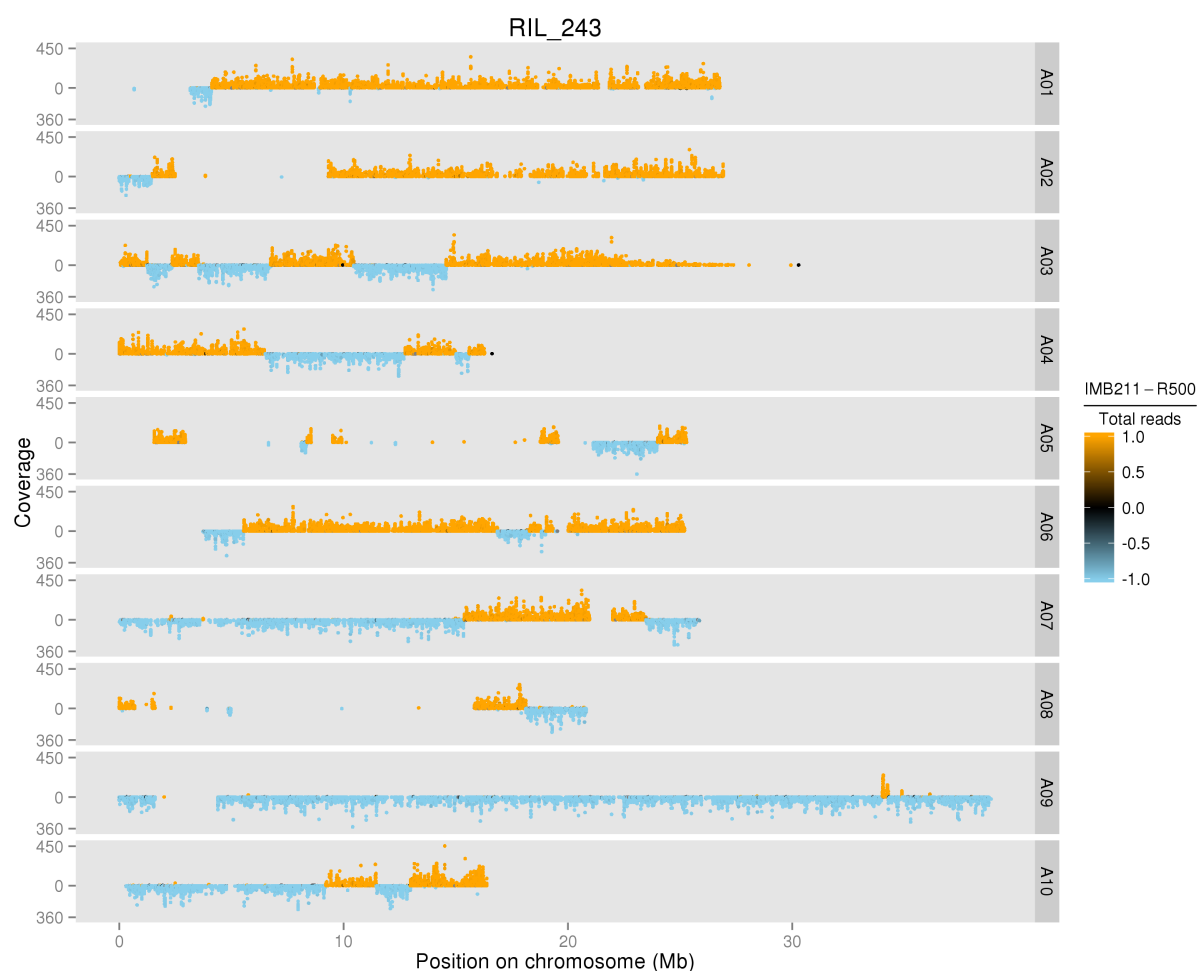
4

5   **Figure 2.** An individual plot of a RIL genotyped with the population based SNP set. Each of the

6   *B. rapa* ten chromosomes are displayed (A01-A10) with counts coverage of each SNP at each

7   physical position on the chromosome in megabases (Mb). The color indicates the relative ratio of

8   coverage between R500 and IMB211 for every SNP.
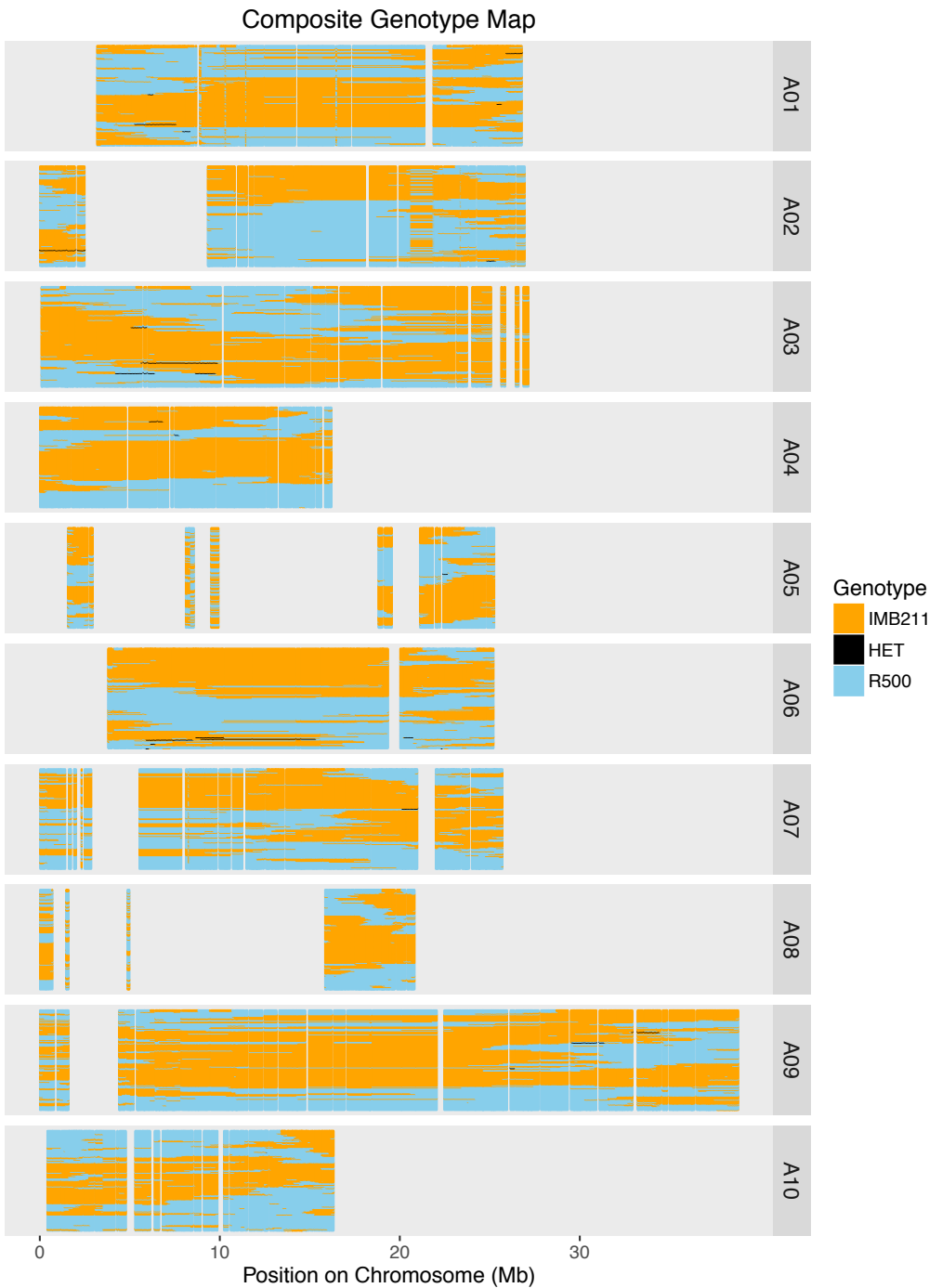
Figure 3. Composite population genotype map with the physical position for each of the ten

chromosomes. Each RIL is represented as a single row displaying the genomic region inherited

from IMB211 (Orange) or R500 (Blue). Small heterozygous regions are represented in black.
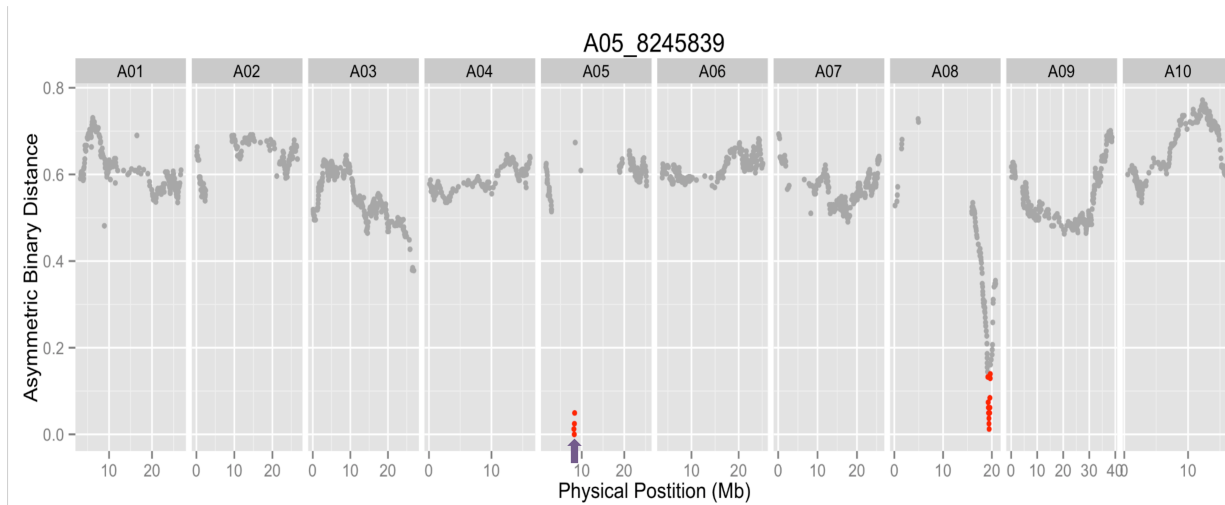
22

Figure 4. A representative asymmetric binary distance plot for a single molecular marker, A05-8245839, indicated by the purple arrow. Markers with 90% correlation to A05-8245839 are indicated in red and occur on chromosomes A05 and A08. The group of markers on A05 were moved to A08.
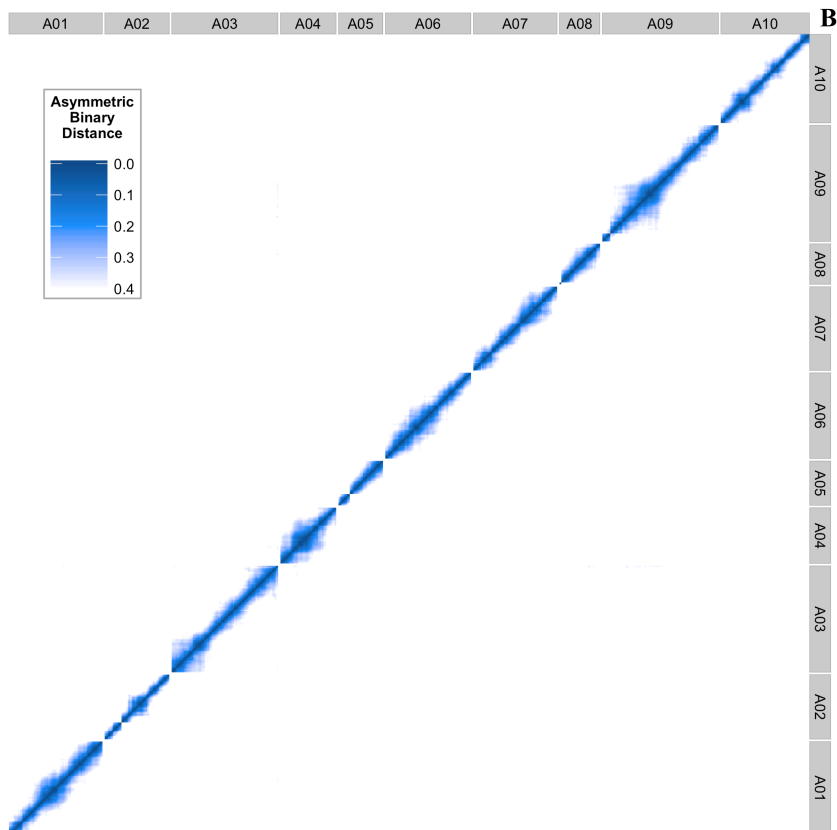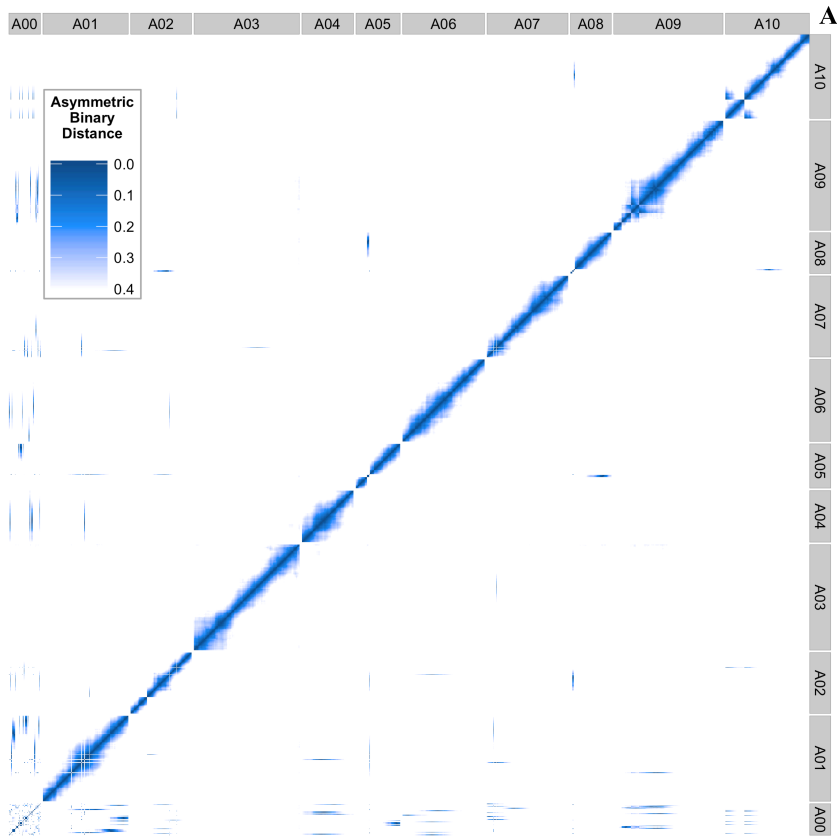
1
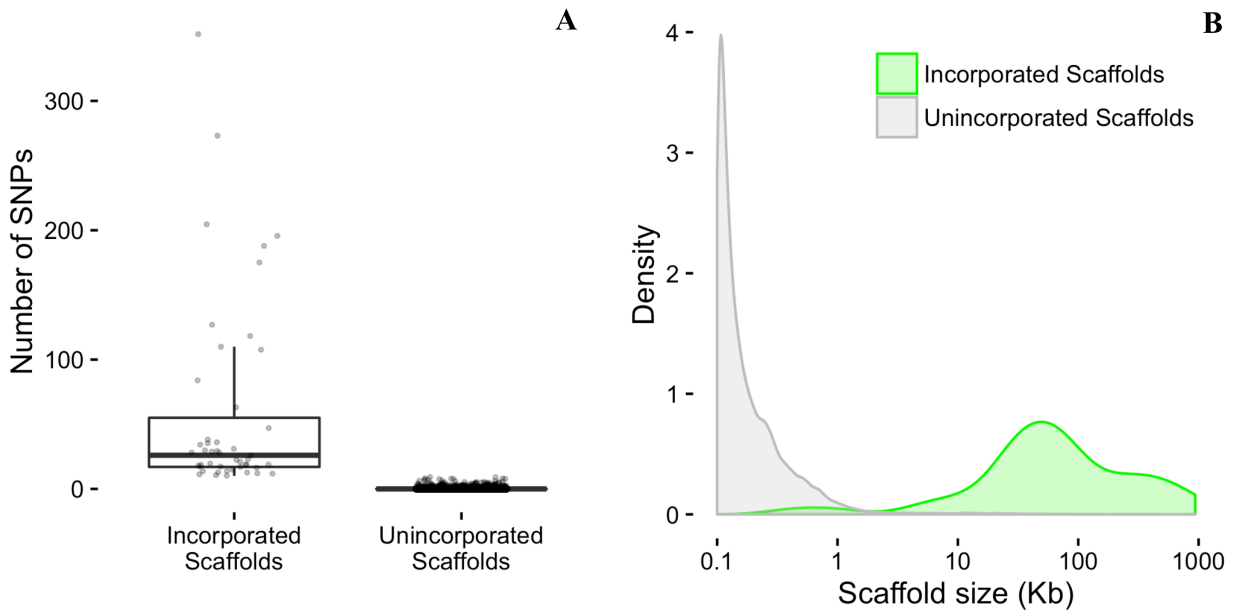
Figure 5. (A) Number of SNPs per scaffold. (B) Density distributions of scaffold sizes. Newly

incorporated scaffolds are shown in green and unincorporated scaffolds are shown in gray.

4

5

6

1   **Figure 6.** Genome-wide asymmetric binary distance plots for each marker compared against

2   every other marker (A01-A10). A00 in panel A represents the unplaced genomic scaffold

3   sequences. Dark blue indicates high correlation (low asymmetric binary distance), while white

4   indicates no correlation. The right plot is the final position of each marker and scaffold after
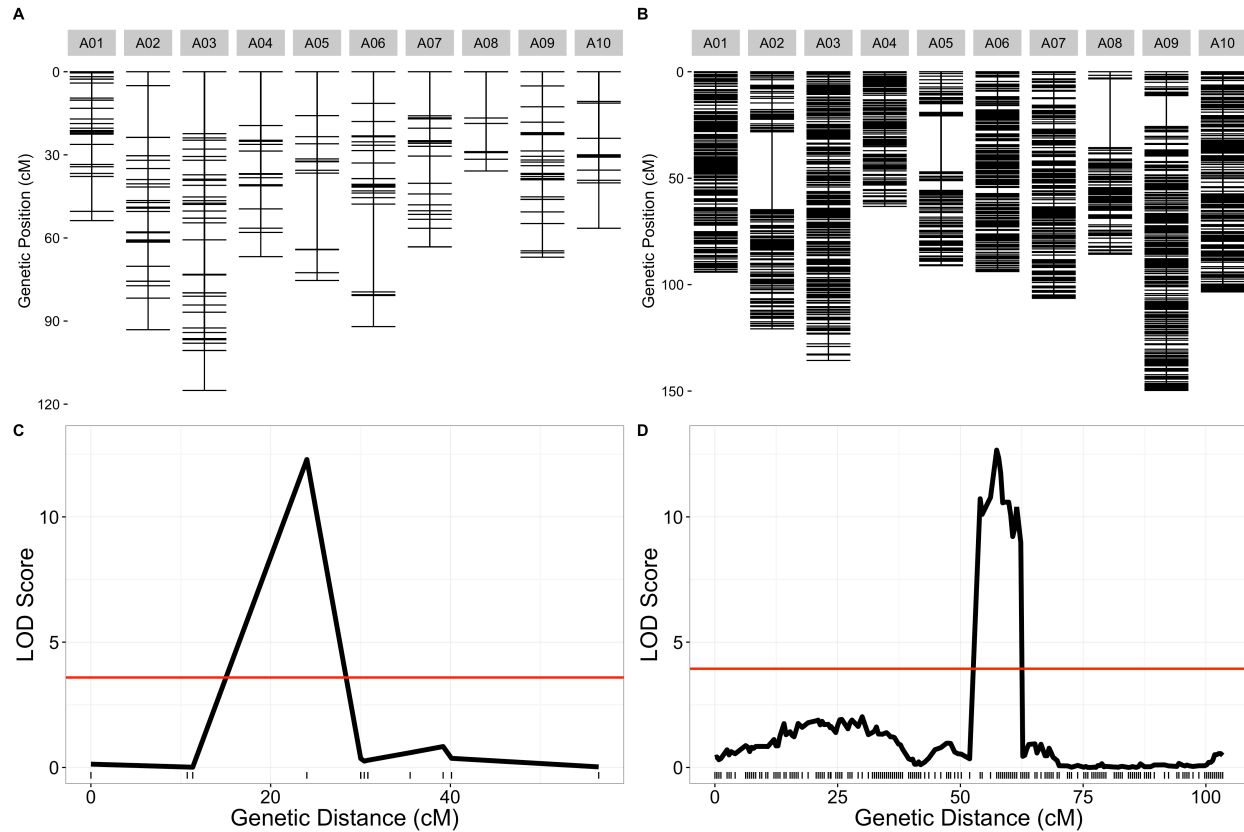
5   applying our pipeline.

6

1

2
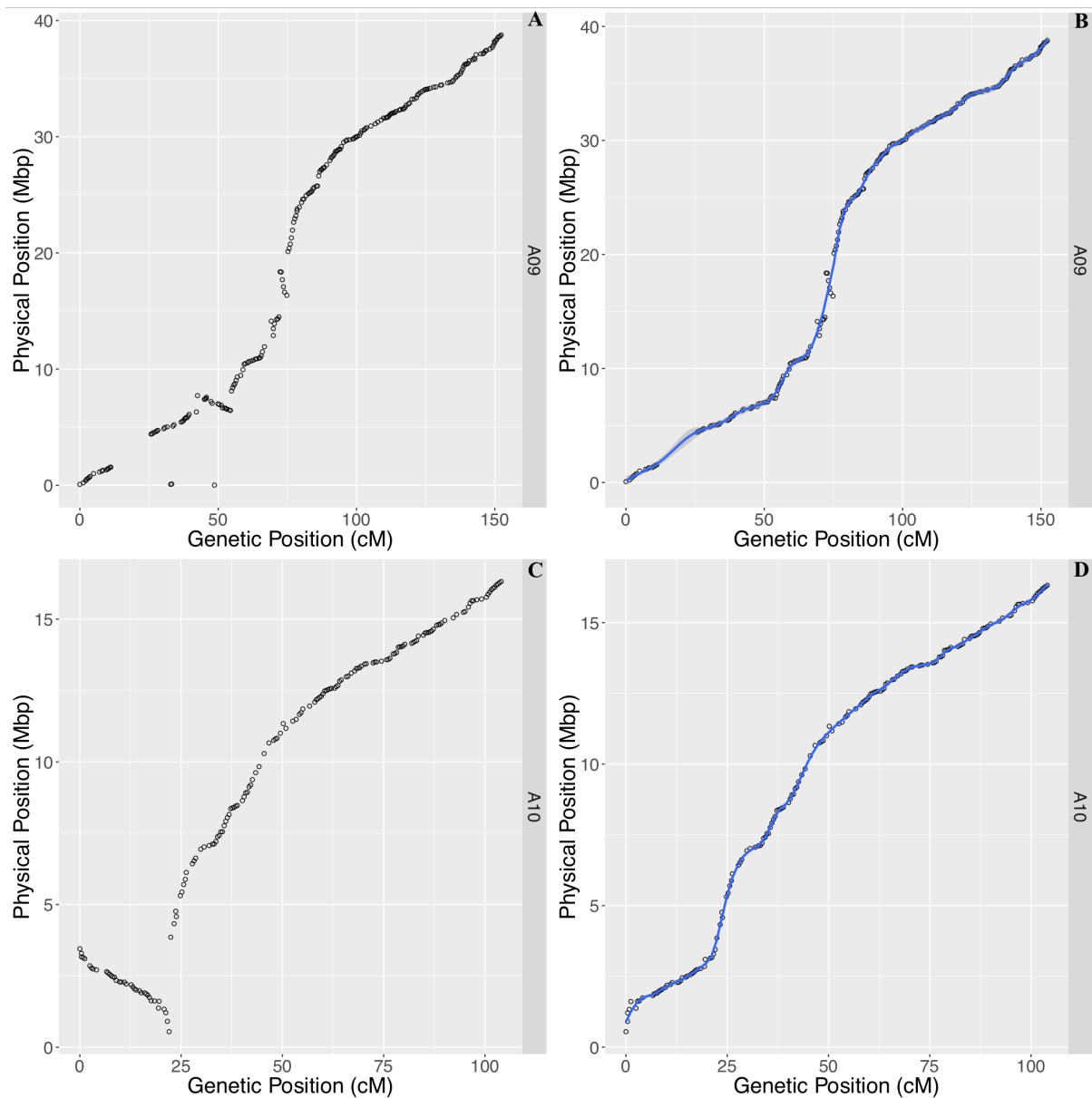
Figure 7. Physical position versus genetic position of each marker for chromosome A09 (A,B) and A10 (C,D) using genome version 1.5 (A,C) and fixed inversions using recombination information (B,D). Loess smoothing for converting between genetic and physical distance is displayed by the blue line in panels B and D.

7

27

1

Figure 8. Old and new genetic map comparisons. Genetic markers for each chromosome are

displayed in centimorgan distance (cM) for the old (A) and new (B) genetic maps. Comparison

of likelihood odds scores for flowering time QTL using the old (C) and new (D) genetic maps.

1   **REFERENCES**

2   Baker, R.L., Leong, W.F., Brock, M.T., Markelz, R.J.C., Covington, M.F., Devisetty, U.K.,

3   Edwards, C.E., Maloof, J., Welch, S., and Weinig, C. (2015). Modeling development and

4   quantitative trait mapping reveal independent genetic modules for leaf size and shape. New

5   Phytol. *208*, 257–268.

6   Brock, M.T., Dechaine, J.M., Iniguez-Luy, F.L., Maloof, J.N., Stinchcombe, J.R., and Weinig, C.

7   (2010). Floral Genetic Architecture: An Examination of QTL Architecture Underlying Floral

8   (Co)Variation Across Environments. Genetics *186*, 1451–1465.

9   Broman, K.W., Wu, H., Sen, Ś., and Churchill, G.A. (2003). R/qtl: QTL mapping in

10   experimental crosses. Bioinformatics *19*, 889–890.

11   Buckler, E.S., Phelps-Durr, T.L., Buckler, C.S.K., Dawe, R.K., Doebley, J.F., and Holtsford,

12   T.P. (1999). Meiotic Drive of Chromosomal Knobs Reshaped the Maize Genome. Genetics *153*,

13   415–426.

14   Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J., Belcram, H.,

15   Tong, C., Samans, B., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica

16   napus oilseed genome. Science *345*, 950–953.

17   Cheng, F., Wu, J., Fang, L., and Wang, X. (2012). Syntenic gene analysis between Brassica rapa

18   and other Brassicaceae species. Front. Plant Sci. *3*.

19   Cheng, F., Mandáková, T., Wu, J., Xie, Q., Lysak, M.A., and Wang, X. (2013). Deciphering the

20   Diploid Ancestral Genome of the Mesohexaploid Brassica rapa. Plant Cell *25*, 1541–1554.

1   Dechaine, J.M., Johnston, J.A., Brock, M.T., and Weinig, C. (2007). Constraints on the evolution

2   of adaptive plasticity: costs of plasticity to density are expressed in segregating progenies. New

3   Phytol. *176*, 874–882.

4   Dechaine, J.M., Brock, M.T., and Weinig, C. (2014). QTL architecture of reproductive fitness

5   characters in Brassica rapa. BMC Plant Biol. *14*, 1.

6   Devisetty, U.K., Covington, M.F., Tat, A.V., Lekkala, S., and Maloof, J.N. (2014).

7   Polymorphism Identification and Improved Genome Annotation of Brassica rapa Through Deep

8   RNA Sequencing. G3 Genes Genomes Genet. *4*, 2065–2078.

9   Dixon, G. (2007). Vegetable brassicas and related crucifers (CABI).

10   Edwards, C.E., Ewers, B.E., Williams, D.G., Xie, Q., Lou, P., Xu, X., McClung, C.R., and

11   Weinig, C. (2011). The Genetic Architecture of Ecophysiological and Circadian Traits in

12   Brassica rapa. Genetics *189*, 375–390.

13   Fang, Z., Pyhäjärvi, T., Weber, A.L., Dawe, R.K., Glaubitz, J.C., González, J. de J.S., Ross-

14   Ibarra, C., Doebley, J., Morrell, P.L., and Ross-Ibarra, J. (2012). Megabase-Scale Inversion

15   Polymorphism in the Wild Ancestor of Maize. Genetics *191*, 883–894.

16   Fulop, D., Ranjan, A., Ofner, I., Covington, M.F., Chitwood, D.H., West, D., Ichihashi, Y.,

17   Headland, L., Zamir, D., Maloof, J.N., et al. (2016). A new advanced backcross tomato

18   population enables high resolution leaf QTL mapping and gene identification.

19   Hammond, J.P., Mayes, S., Bowen, H.C., Graham, N.S., Hayden, R.M., Love, C.G., Spracklen,

20   W.P., Wang, J., Welham, S.J., White, P.J., et al. (2011). Regulatory Hotspots Are Associated

1   with Plant Gene Expression under Varying Soil Phosphorus Supply in Brassica rapa. PLANT

2   Physiol. *156*, 1230–1241.

3   Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin,

4   S.Y., Antonio, B.A., Parco, A., et al. (1998). A High-Density Rice Genetic Linkage Map with

5   2275 Markers Using a Single F2 Population. Genetics *148*, 479–494.

6   Haupt, W., Fischer, T.C., Winderl, S., Fransz, P., and Torres-Ruiz, R.A. (2001). The

7   CENTROMERE1 (CEN1) region of Arabidopsis thaliana: architecture and functional impact of

8   chromatin. Plant J. *27*, 285–296.

9   Iniguez-Luy, F.L., Lukens, L., Farnham, M.W., Amasino, R.M., and Osborn, T.C. (2009).

10   Development of public immortal mapping populations, molecular markers and linkage maps for

11   rapid cycling Brassica rapa and B. oleracea. Theor. Appl. Genet. *120*, 31–43.

12   Kumar, S., Banks, T.W., Cloutier, S., Kumar, S., Banks, T.W., and Cloutier, S. (2012). SNP

13   Discovery through Next-Generation Sequencing and Its Applications, SNP Discovery through

14   Next-Generation Sequencing and Its Applications. Int. J. Plant Genomics Int. J. Plant Genomics

15   *2012, 2012*, e831460.

16   Li, G., and Quiros, C.F. (2001). Sequence-related amplified polymorphism (SRAP), a new

17   marker system based on a simple PCR reaction: its application to mapping and gene tagging in

18   Brassica. Theor. Appl. Genet. *103*, 455–461.

19   Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A.P., Zhao, M., Ma, J., Yu, J.,

20   Huang, S., et al. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of

21   polyploid genomes. Nat. Commun. *5*, 3930.

1   Lou, P., Xie, Q., Xu, X., Edwards, C.E., Brock, M.T., Weinig, C., and McClung, C.R. (2011).

2   Genetic architecture of the circadian clock and flowering time in Brassica rapa. Theor. Appl.

3   Genet. *123*, 397–409.

4   Lou, P., Wu, J., Cheng, F., Cressman, L.G., Wang, X., and McClung, C.R. (2012). Preferential

5   Retention of Circadian Clock Genes during Diploidization following Whole Genome

6   Triplication in Brassica rapa. Plant Cell *24*, 2415–2426.

7   Lowry, D.B., and Willis, J.H. (2010). A Widespread Chromosomal Inversion Polymorphism

8   Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation.

9   PLOS Biol *8*, e1000500.

10 R Core Team (2015). R: A language and environment for statistical computing.

11 Sherman, J.D., and Stack, S.M. (1995). Two-dimensional spreads of synaptonemal complexes

12 from solanaceous plants. VI. High-resolution recombination nodule map for tomato

13 (Lycopersicon esculentum). Genetics *141*, 683–708.

14 Wang, H., Wu, J., Sun, S., Liu, B., Cheng, F., Sun, R., and Wang, X. (2011a). Glucosinolate

15 biosynthetic genes in Brassica rapa. Gene *487*, 135–142.

16 Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng,

17 F., et al. (2011b). The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. *43*,

18 1035–1039.

19