# Abstract

Here we describe high-density SNP discovery and genetic map construction for a *Brassica rapa* recombinant inbred line population. We also use the genotype data from the population to detect and remedy putative genome mis-assemblies and to assign scaffold sequences to their likely genomic locations. These improvements to the assembly represent 7.1-8.0% of the annotated *Brassica rapa* genome. We demonstrate how this new resource is a significant improvement for QTL analysis over the current low-density genetic map.

# Introduction

## Quantitative Genetics Resources

A general problem in biology is creating causal links between changes in the DNA sequence (genotype) and measurable characteristics of an organism (phenotype). Spanning the genotype to phenotype gap has implications for understanding human and animal diseases as well as understanding drivers of plant growth and yield in agriculture. Genetic mapping populations have been created as tools to aid in linking genotype to phenotype through statistical techniques. With recent advances in DNA sequencing technologies more genotypic information can be collected on individuals in mapping populations to aid in the search of causal genes for complex phenotypic traits important for agriculture such as plant flowering time. One such population is the Bra-Irri population that was created from a cross between a rapid cycling line and a common oilseed variety of Brassica rapa (@iniguez-luy\_development\_2009).

The previous genetic resources available for this population includes 224 SSR markers spread across 10 linkage groups (chromosomes). While this has been an extremely great resource for mapping a diverse set of phenotypic traits (@lou\_genetic\_2011, @dechaine\_qtl\_2014, @brock\_floral\_2010), the existing markers were not anchored to a known location in the genome. The Brassica rapa genome was released in 2012...

## Original marker sets

### explain more about SSR and limitations

## Make case for advantages of creating high density marker set

# Methods

## Genetic Map Construction

Using one unique SNP per genotype bin, we created a saturated genetic map. The genetic map was constructed using the chromosomal position of each of the SNPs as a starting point for marker ordering along the chromosomes. Each chromosome was treated as a large linkage group and each SNP was tested for linkage disequilibrium with all other SNPs with the R/QTL package (Broman et al. XXXX, Supplemental Data X) in the R statistical environment (@R-ref). The larger gaps in the map is where there is little marker information and corresponded to centromeric regions (Figure X b). These large gaps caused a small problem when ordering the markers and connecting each of the chromosomal arms in the correct order. In these ordering chromosomes X, Y, Z we used the physical position of the SNPs to connect the two arms in the correct order. The orientation of the scaffolds cannot be determined with the current methods because of we are limited by the size of the population. Therefore, scaffolds are placed between adjacent bins determined by lowest recombination probabilities.

## QTL Comparisons

To demonstrate an improvement in coverage in mapping physiological traits, we remapped two traits from @brock\_floral\_2010 that used the existing genetic map. As the fairest comparisons between maps, marker regression was performed using the scanone() function in RQTL with 10,000 permutations to determine the significance cutoff (Figure X).

# Results and Discussion

## Deep RNA sequencing of a RIL population and its parents

We performed deep RNA sequencing of 124 members of a recombinant inbred line (RIL) population derived from the *Brassica rapa* accessions R500 and IMB211. We sequenced each member of the RIL population with ***X***-***X*** replicates at an average of ***X*** reads mapped per replicate for a combined mean of ***X*** reads mapped per RIL.(***Q: What other metrics to include? reads per expressed gene? # of genes hit per sample? Depth vs reads mapped?***)(***Q: Should we add a table with some stats/numbers?***) We have previously described the deep sequencing of R500 and IMB211 [@devisetty\_polymorphism\_2014] and use that data in this current study. These datasets will eventually be used in analyses of differential gene expression; however, we are currently using them for variant discovery and creation of a genetic map resource.

## R500 vs. IMB211 polymorphism identification

To identify polymorphisms within the RIL population, we first looked for SNPs and INDELs between R500 and IMB211, the parents of the population. We used an approach previously described for the identification of polymorphisms between R500 and IMB211 with an earlier version of the *Brassica rapa* genome annotation, v1.2 [@devisetty\_polymorphism\_2014]. In the current study, we used v1.5 of the genome for all mapping and informatics work.

This set of R500 vs. IMB211 polymorphisms was used to genotype each member of the RIL population. Based on the crossing scheme used to create the RIL population, we would expect homozygous regions of contiguous R500 alleles alternating with homozygous regions of contiguous IMB211 alleles. When using this polymorphism set to genotype the RILs; however, there were several regions where R500 and IMB211 alleles were interspersed with no apparent pattern. (***Q: Show representative RIL or region of a RIL in main or supplemental?***). This suggests that a different parental source may have been used to construct this RIL population.

To test this hypothesis, we merged all of the sequence data from the individual RILs and then genotyped the merged dataset (**Figure** [**[rils-merged.parent-based-snps.png]**](#rils-merged.parent-based-snps.png)). Given the relatively large size of the population and the expected introgression frequency and distribution, polymorphisms generated by comparing the actual parents of the RIL population should be segregating with approximately equal allelic frequency in this merged data set (black dots in figure). Most polymorphisms did display this expected distribution; however, there were several large regions that were not segregating, instead they were monomorphic for one of the putative parents of the population (indicated as orange or blue dots in the figure). Nearly all of these regions matched the putative R500 alleles. The primary exception is the bottom of chromosome A03, which displays a gradual transition from equal R500:IMB211 allelic frequency to nearly all IMB211. This pattern is consistent with segregation distortion and {==is what we would expect given that A03's centromere is at the end of the chromosome (Cheng 2013).==}{>>I don't understand this part. Is the idea Meiotic drive for chrom A03 centromere? We should say that.<<}

One explanation for the non-segregating regions is that the seed stock of one or both of our parental lines was not actually the true parent of the population(***Q: contaminated?***). To evaluate this possibility, we sequenced several additional seed stocks of R500 and IMB211 (***Q: How distinct are the sources?***). The genotypes of all ***X*** R500 seed stocks were consistent with one another across all (***Q: 198,141?***) polymorphisms. There were, however, at least ***X*** distinct genotypic variants within the ***X*** IMB211 seed stocks. Therefore, we proceeded to work under the assumption that the sequences from the R500 seed stocks that we have provide a reasonably accurate genotypic representation of the R500 parent used to create the R500-IMB211 RIL population.

## Population-based SNP discovery

Due to the uncertainty surrounding the IMB211 parent of the RILs, we switched to a population-based approach for discovering SNPs. This new strategy involves identifying variants within the RIL population and using the R500 data to determine which is the R500 allele and which is the IMB211 allele for each SNP. Using this approach, we identified 146,027 SNPs across *B. rapa*'s ten chromosomes (**Table** [**[SNP-counts.png]**](#SNP-counts.png)).

Using the population-based SNPs to genotype the merged RIL sequence data produces the expected allele frequencies throughout the entire genome (**Figure** [**[rils-merged.population-based-snps.png]**](#rils-merged.population-based-snps.png)). Over ***X***% of the genome is within ({==***Q: 2 Kbp?***==}{>>I would just figure out what the distance is for 90 and 95% and report that<<}) of a SNP; however, there are several regions with few or no SNPs. There are two primary reasons for these SNP-free regions. Most are likely gene-poor regions or regions of genes with insufficient expression under our experimental conditions (e.g., growth conditions, age, tissue, genotypes, etc.).{>>Can/should we actually show this?<<} We also found a few regions where there are significant numbers of expressed genes, but no SNPs between members of the RIL population. (***Q: Indicate these regions in the figures?? This may be time consuming since the X-axis has changed between versions.***){>>What about having a bar underneath the chromosomes that indicates gene density? This would also address my question above<<} These regions primarily correspond to the homozygous regions of **Figure** [**[rils-merged.parent-based-snps.png]**](#rils-merged.parent-based-snps.png) and, therefore, likely represent regions that are very similar between the seed stocks used to generate this RIL population.

## Genotyping the RIL population

Using the information on the entire population instead of just the parents provides many more observations of the potential SNPs in the population.

* Sample RIL genotype plot
* Boundary detection and bin construction
* Composite genotype table/plot (starting with detecting and merging boundaries of individual RILs)

## Finding and fixing genome misassemblies

* Composite plot suggests misassemblies
  + Add other evidence that supports this
* Look at pairwise distances between bins
  + individual plots and overview plot (overview first? then individual plots help determine which bins should be where.)
  + Use this info to rearrange bins (have table with rearrangement stats.)
  + Plot overview and composite map following rearrangement

## Incorporating scaffold sequences into the genome

In the current version of the *B. rapa* genome annotation (v1.5) there are 40,357 scaffolds that have not been incorporated into any of the ten chromosomes. These scaffolds range in size from 100 bp to 938 Kbp and represent 1,411 genes spanning 27.5 Mbp. For comparison, there are 39,609 genes within the 256 Mbp of annotated chromosomal sequence. Given that the scaffolds contain about as many genes as would be expected on one third of an average chromosome, we decided to extend our strategy for fixing genome misassemblies to estimate the approximate chromosomal locations of the scaffolds.

We identified 3,070 SNPs across 339 of the 40,357 scaffolds. This corresponded to a sufficient number of SNPs (10+) for only 47 scaffolds. For each of these 47 scaffolds, we found a chromosomal bin with identical or near identical genotypes, indicating very close genetic linkage. This degree of similarity made it trivial to incorporate these 47 scaffolds into the chromosomal bin structure. The incorporated scaffolds range in size from 429 to 884,746 bp and are enriched for larger scaffolds (**Figure** [**[scaffold-size-distributions.png]**](#scaffold-size-distributions.png)). The N50 values for incorporated and unincorporated scaffolds are 436.0 Kbp and 11.8 Kbp, respectively. Although we have incorporated only 0.1% of the scaffolds into the genome, the incorporated scaffolds represent nearly 7 Mbp of sequence, a disproportionately high 25% of all scaffold sequence (**Table** [**[incorporated-scaffolds-stats.png]**](#incorporated-scaffolds-stats.png)).

While most of the incorporated scaffolds represent a single genotype bin, seven are comprised of multiple bins. Scaffold000164, for example, includes 65 annotated genes across six distinct genotype bins within its 313.7 Kbp sequence. One scaffold, Scaffold000191, had bins mapping to two different chromosomes, indicating that it was misassembled. Therefore, we split its two bins and assigned them to the appropriate chromosome locations (5 genes/28.2 Kbp to A01 and 24 genes/104.1 Kbp to A05).

Possible reasons for the enrichment of larger scaffolds within the set of incorporated scaffolds include: (1) larger scaffolds are more likely to include expressed genes and, therefore, SNPs that we can detect and (2) larger scaffolds may be more likely to be accurate representations of a contiguous region within the genome. This second point is based on the assumption that large scaffolds were able to be assembled perhaps due to more abundant, more consistent, and/or more convincing experimental support than small scaffolds. The large scaffolds that we were unable to incorporate because they ***...***

## High-density genetic map

From the available SNP data we were able to create a genetic map with ten linkage groups corresponding to the 10 chromosomes of *Brassica rapa* ((**Table** [**[genetic-map-summary.png]**](#genetic-map-summary.png))). The map contains 1482 genotyped markers for 124 RILs and is completely saturated based on recombination events existing in the population. The maximum spacing between markers is 36.6 cM with an average spacing between markers of 0.7 cM (Table X), The overall map distance of 1045.6 cM. The new map is compared to the existing map containing 224 markers with an average spacing of 1.5 cM (@iniguez-luy\_development\_2009). Furthermore, the new marker set has physical anchors to the genome for each marker and includes newly placed scaffolds on chromosomes X, Y, Z.

## Conclusions

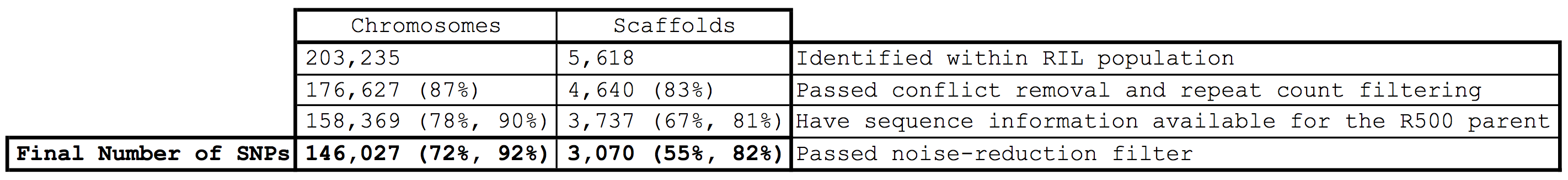
# Figures

>

**Figure [rils-merged.parent-based-snps.png]: Plot of merged data from all RILs genotyped using the parent-based SNP set. Each of the B. rapa ten chromosomes are displayed (A01-A10) with counts coverage of each SNP at each physical position on the chromosome in megabases (Mb). The color indicates the relative ratio of coverage between R500 and IMB211 for every SNP. Black is equal coverage, orange is more IMB211 and blue is more R500.**

>

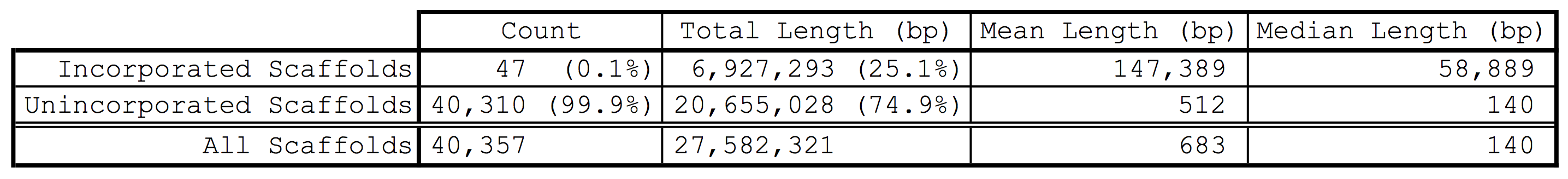
**Figure [rils-merged.population-based-snps.png]: Plot of merged data from all RILs genotyped at population-based SNP set. Each of the B. rapa ten chromosomes are displayed (A01-A10) with counts coverage of each SNP at each physical position on the chromosome in megabases (Mb). The color indicates the relative ratio of coverage between R500 and IMB211 for every SNP. Black is equal coverage, orange is more IMB211 and blue is more R500.**

> 

**Table [SNP-counts.png]: SNP counts at different steps of the SNP discovery pipeline.** The percentage of SNPs located on chromosomes or scaffolds remaining after each step are shown in parentheses. The first percentage is relative to the initial set of SNPs and the second percentage is relative to the set of SNPs from the previous step.

>

**Figure [scaffold-size-distributions.png]: Density distributions of scaffold sizes. Newly incorporated scaffolds are shown in green and unincorporated scaffolds are shown in gray.**

> 

**Table [incorporated-scaffolds-stats.png]: Incorporated scaffolds represent a disproportionately high amount of scaffold sequence.** Percentages of scaffold subset counts and total lengths relative to the set of all scaffolds are shown in parentheses.

> > **Figure [SNPs-per-scaffold.png]: Number of SNPs per scaffold.**

> > **Figure [figures/genetic-map-qtl-figure.png]: 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. Likelihood odds scores for flowering time QTL using the old (C) and new (D) genetic maps.

> > **Figure [figures/genetic-map-recombination-fraction.pdf]: Pairwise recombination fractions (upper left triangle) and Likelihood Odds Scores (LOD; lower right triangle) of SNP markers across the 10 chromosomes of *Brassica rapa*. Color scale goes from red (greater association) to blue (no association).**

> > **Figure [figures/A10-genetic-vs-physical-v2-3.pdf]: Physical position versus genetic position of each marker for genome version 1.5 (A) and our improved scaffold/marker placement (B)**

## Improved QTL results

The new genetic map improves QTL resolution using the same phenotypic data and simple QTL mapping method (Figure X). We had a hard time placing the old markers on the reference genome. We should show this...

### comparison of old and new.

### Does the new map cover the whole known genome?

### do the SSR markers cover the whole genome??

The physical locations of the markers on the genome allowed us to convert between physical and genetic distance by fitting polynomial regressions to the data (Figure X). This is a significant improvement for candidate gene finding because the LOD support intervals between known markers can be queried for genes of interest. In this case, the interval contains X number of genes, of which X have been shown to be involved with flowering time in past studies (refs).