|  |
| --- |
| Genome Analysis  **Characterization of a *Brassica napus* QTL affecting flowering time and root development using homology-guided assembly**  Zhisheng Xu 1,\* and Rich Fletcher2  1Department of Computer Science, Colorado State University, Fort Collins, CO 80523  2Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Motivation:** The QTL mapping approach has provided great insight into the genetic architecture of complex phenotypes, yet the elucidation of the specific genes underlying such phenotypes is still described as a ‘grand challenge’ in modern biology. Recent advancements in DNA sequencing technology in combination with improved sequence analysis methods provide a means by which a small list of candidate genes can be rapidly identified.

# introduction

Most of the traits we observe in nature are under the control of many genes and, due to genetic variation within those genes, have a large distribution of trait values within a population. These traits have been termed quantitative traits and the loci controlling them are known as quantitative trait loci (QTL). QTLs, at the most basic definition, are regions of the genome which contribute to continuous phenotypic variation of a trait (Mackay, 2001). These regions may contain a gene or many genes which act in an additive and independent fashion, interact with one another (epistasis) or may have ever more complex effects such as pleiotropy (Falconer and Mackay, 1996). Due to the very small effects some QTLs may have, dissecting the genomic landscape of an organism to understand these genetic controls, additive or otherwise, can be a complex process. Further, the large regions that normally make up QTL, and the many genes contained therein, often leave the biologist with a long list of candidate genes.

Mapping QTL depends upon the construction of a genetic map by analyzing the segregation of alleles in a recombinant population (Haldane, 1919). In plants, a recombinant population is often generated from a directed cross between two inbred lines, hereafter know as the parents. Recombination during meiosis results in a population of individuals whose genotypes are a mosaic of the parental alleles. Parental alleles can be tracked via a number of marker methods including morphology (Sturtevant, 1913; Sax, 1923), isozyme (Hunter and Markert, 1957) and DNA (Botstein, 1980; Nakamura, 1987). A DNA marker is simply a DNA sequence that shows sequence polymorphism among individuals within a species (Andersen and Lubberstedt, 2003). Analysis of the co-segregation of marker polymorphisms (alleles) within the recombinant population allows a researcher to construct a map of the estimated locations of each allele relative to others. Marker alleles that are found together more often are assumed to be located closer together because of a limitation in crossover events due to physical the physical constraints of crossing-over (Muller, 1916). Pairwise comparisons of the frequency of each allelic combination are used to create a genetic map which provides a visual illustration of their location on an organism’s chromosomes. QTL mapping is the simple extension of associating regions of the genetic map with variation in the quantitative phenotype of interest (Mauricio, 2001).

Identification of candidate genes can be performed after QTL mapping by capitalizing upon gene annotation data available for genes located within the defined QTL interval. However, gene annotation data sets are rare for non-model organisms. The solution to this problem is to utilize knowledge about the synteny (conserved gene order among species) between model organisms, where gene annotations exist and related species. Fourmann *et al*. (1998) successfully employed just such an approach to choose candidate genes in syntenic regions of the *Arabidopsis thaliana* and *Brassica napus* genomes with annotations consistent with a functional relationship to erucic acid accumulation.

The next step in the progression towards identifying the functional gene underlying a trait is to locate a DNA polymorphism that distinguishes one parent line from the other. Traditionally, the approach was to amplify candidate genes via PCR and sequence the product via Sanger sequencing methods. This proved to be a slow and laborious process that required either an enormous amount of serendipity or a comprehensive understanding of the biological pathway to enable effective selection of the candidate gene(s). The number of functional genes identified via this method is small, suggesting an alternative approach would be highly beneficial to quantitative geneticists.

It is with this motivation that we performed the research described herein. Specifically, we set out to characterize the polymorphisms existing in a QTL region on chromosome A10 previously identified in a canola (*Brassica napus* L.) mapping population (manuscript in prep). The QTL explains variation in two ecologically and agriculturally important quantitative phenotypes. The first phenotype, flowering time, is a key life history trait and, accordingly, has implications on local adaptation, as shown in results demonstrating flowering time adaptation along altitudinal and longitudinal clines (Stinchcombe *et al*., 2004, Montesinos-Navarro *et al*., 2011). Additionally, it is a major focus for plant breeding programs of all major crop species due to its strong influence on grain yield (Buckler *et al*., 2009). The second trait, root pulling force (the force required to vertically remove a plant from the soil), has been shown to associate with increased root volume (Hayes and Johnson, 1939), root strength (Kevern and Hallauer, 1982; Arihara and Crosbie, 1982), nitrogen uptake and yield (Kamara, 2003) and drought tolerance (Ekanayake, 1985; Tuberosa, *et al*., 2002), all traits with major implications on plant productivity and sustainable yield gains during an age of increasing global temperatures and ever-more unpredictable weather patterns.

An interesting feature of this particular QTL is the fact that it explains variation in both phenotypes, suggesting that the functional genes underlying the separate phenotypes are either the same (pleiotropic), or are very closely linked to one another such that they don’t segregate independently in the mapping population. This was perhaps not unexpected since going into the analysis it was known that the two traits had a high genetic correlation (rg = 0.66; p < 0.0001) which is generally considered to be the result of either a close physical proximity (genetic linkage) or pleiotropy. In the broad sense, pleiotropy refers to the effect a mutation has on two or more phenotypes. It is an area of renewed research interest (Wagner and Zhang, 2011) due to the profound effects it can have on the efficacy of selection, both natural and artificial, as a mutation advantageous to one trait can be equally detrimental to another (Falconer and Mackay, 1996). In this regard, a deeper understanding of the nature of the QTL we are studying has important implications to both plant breeders and evolutionary biologists for, if the genetic architecture of the QTL is deemed to be pleiotropy, it can be concluded that the traits cannot be decoupled by selection and their covariance is definite. However, an alternative conclusion suggests that plant breeders, and natural selection alike, can select for individuals with differential flowering times and RPF values so that a variety could be tailored with ideal values for both traits and, thus better adapted to the target environment.

To identify a list of the polymorphisms within our QTL region we started by obtaining a second generation Illumina Hi-Seq sequence data set for both parent lines at a depth of ~30X. These data were aligned to the existing reference genome of the related species, *Brassica rapa*. Next, a *de novo* assembly was performed on all reads that did not map to the reference inan effort to identify large scale structural variation existing between the parents since it is known that such variation is widespread in plant species (Springer *et al*., 2009) which has been confirmed in studies comparing the *B. napus* and *B. rapa* reference genomes (Cheung *et al*., 2009). By employing what is termed a homology-guided assembly approach, we enable our ability to identify such variation that would have otherwise been missed using traditional alignment-consensus approaches.

It is noteworthy to mention that *B*. *napus* is an allopolyploid species formed from the hybridization of two distinct but very closely relate diploid *Brassica* species, *Brassica rapa* (A genome, n=10) and *Brassica oleracea* (C genome, n=9). *B*. *rapa* and *B*. *oleracea* are estimated to have diverged between 3 and 4 million years ago as a result of the high degree of sequence homology observed between them (Cheung *et al*., 2009).

# methods

## Plant material and DNA extraction

High quality DNA was extracted from the fifth leaf of the *B. napus* mapping population parent lines, Wichita and IMC106, using the standard methods of the Qiagen (Valencia, CA) column extraction kit. The extracted DNA was then run on a 1% agarose gel to confirm DNA quality and concentrated to ~50 ng/μl. Each parent was sequenced on a single lane of an Illumina HiSeq 2000 (San Diego, CA) sequencer at the University of Missouri DNA Core Facility (<http://biotech.rnet.missouri.edu/dnacore/>) to generate 2 x 100 paired-end reads.

## Candidate gene identification

To limit the search space for identifying molecular polymorphisms, we started by identifying all of the genes present in the 1Mb QTL interval using the existing *Brassica rapa* genome (Wang *et al*., 2011). The output of this database provides the best-hit ortholog from the Arabidopsis thaliana genome which were used in finding gene annotations in The Arabidopsis Information Resource (TAIR; <http://www.arabidopsis.org/>). Annotations were then used as a filter in narrowing the list of candidate genes down to those annotated as being either:

(1) Expressed in floral and/or root tissues.

(2) Affecting flowering and/or root phenotypes.

## Shore pipeline and SNP/InDel calling

SHORE (Ossowski *et al*., 2008) is a pipeline for aligning reads and predicting SNPs and InDels. We downloaded the most recent version (0.7.1beta) from SouceForge (http://sourceforge.net/). The pipeline we used for SNPs and InDel prediction is outlined as follows:

1. Preprocessing the reference: This creates a new copy of the Fasta reference file and outputs adjusted chromosome/contig IDs into an index folder. This preprocessing step only needs to be executed once for each genome and will create the mapping indices as well as providing estimates of the local GC content and repetitiveness.

Command: shore preprocess -f Brapa\_sequence\_v1.2.fa -i ./canole

Shore will preprocess Brapa\_sequence\_v1.2.fa and outputs all results to folder ‘canole’ under the current directory.

1. SHORE import: SHORE has its own read file format called reads\_0.fl, or flat-file, necessitating that we convert the Illumina HiSeq Fastq files into this format. This process requires that we specify an output file and will create a new lane, which is a folder for storing all of the mapping results. The paired-end reads are stored separately under different folders in one lane, each is a reads\_0.fl file.

Command: shore import -v Fastq -e Shore -a genomic -x ../parent1a.fastq -y ../parent1b.fastq -o run\_01

Here we use -v to specify our reads have fastq file format, -e Shore specifies the creation of folders and files in the correct SHORE’s formate. Additional options like -e Console will make it output to console. -a genomic means we are dealing with genomic data, not mRNA, sRNA or ChIPseq data. File specified by –x and –y are the two paired-end file of Wichita. It will write the result to folder run\_01. A similar command is also run for IMC106. It will create a new lane for import, which is ./run\_01/1. Under this lane, there are two folders (./run\_01/1/1, ./run\_01/1/2) each storing a reads\_0.fl.gz file. The compressed file is about 11 Gb for Wichita and 10 Gb for IMC106. There are also some log file recording command information under each folder as well as an import\_statistics.txt for each lane.

1. SHORE mapflowcell: This is SHORE’s read mapping process. We used the default alignment tool, GenomeMapper (Schneeberger *et al*. 2009), as it has provided adequate results in similar studies ([Klein](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Klein%20JD%5Bauth%5D) *et al*. 2011). For each reads\_0.fl file, GenomeMapper will create a result map.list file. During the mapping, it will separate all the reads into blocks where each has about 100,000 reads. From this it will then write two files, called a map.list and a left-over.fl, for each block. After alignment of all the reads, GenomeMapper will concatenate the map.list and left-over.fl from all blocks into a single map.list file and a single left-over.fl file under each lane.

Command: shore mapflowcell -n 4 -g 3 -c 16 -p -i canole/Brapa\_sequence\_v1.2.fa.shore –f run\_01/1/\*/reads\_0.fl.gz

This command will align reads from one parent. -f is the reads file to align, -n is the edit distance, -g is the max gaps allowed indicating how many mismatches are allowed to be gaps rather than plain mistakes. We chose to use the default parameters (0 for edit distance and maxgaps) as they are fairly conservative. To accelerate the alignment, we can specify how many cores we want to use and SHORE will create that many processes. We ran the alignments on the two machines: eggs.cs.colostate.edu and bacon.cs.colostate.edu which have 16 CPUs. Finally, -p prepares the alignments for later use in correct4pe. The result is then compressed, map.list.gz for each paired-end is 10.7 Gb for Wichita and 8.7 Gb for IMC106. The left\_over.fl.gz is 8.1 Gb and 7.0 Gb for each parent, respectively.

1. SHORE correct4pe: After aligning each paired-end read separately, correct4pe finds the most likely mapping of repetitive reads by capitalizing on the information contained within the paired-end reads. In paired-end mapping, the mate-pair information can be used to increase the likelihood of an alignment by selecting the paired alignment based on the most likely distance between the pairs. The SHORE correct4pe process starts by sorting the map.list, since the original map.list is just a concatenation of map.list from within each block such that the position they mapped to is only sorted by read ID. After sorting, estimating the insert size distribution becomes a much easier task. Next, SHORE correct4pe calculates an estimate of the insert size distribution which is then translated into a probability function that is used in correcting the alignment result. A library.txt file is also output with information about the maximum, minimum and mean insert sizes being used in the algorithm.

Command: shore -l ./run\_01/1 –x 268 –e 1

After the alignment for each parent is complete, the correct4pe command will run on lane 1 of run\_01 folder, -x is the estimated insert size, and –e specifies the library number. It output a correct map file for each paired-end, which are map.list.1.gz and map.list.2.gz under each pair-end folder. SHORE also keeps a file of the distribution of the insert distance. We plotted this in Python which is shown in Supplemental Figure 1. It also outputs a library.txt file, a sample of which is illustrated by the following:

1 PE 264.583 535 101

It contains the library information necessary for SHORE wgha.

1. SHORE merge: At this point in the pipeline, the results are stored under separate paired-end folders but the following steps require that all alignments be considered at once. To deal with this, the separate map.list files will be merged into one and the two left-over.fl files will be merged into three separate files called left-over\_lowq.fl, left-over\_single.fl and left-over\_paired.fl) according to whether or not they were considered left-over based on quality or were an unmapped single-end or both ends of the paired-end reads.

Command: shore merge -l –m ./run\_01/1/1/map.list.1.gz,./run\_01/1/2/map.list.2.gz -o ./parent\_1

Shore merges the two corrected map files, -l is a command for SHORE to combine left-over reads. The folder parent\_1 now contains one map.list.gz file and three left-over files.

1. SHORE consensus: This process sequentially scans an alignment to gather all read information available at a specific locus. This information is then used to predict polymorphisms (SNPs and InDels <3 bp) existing between each *B. napus* parent line and the *B. rapa* reference.

Command: shore consensus -n canole -r -f ./canole/Brapa\_sequence\_v1.2.fa.shore -o ./parent\_1 /Analysis\_01 -i ./parent\_1/map.list.gz -b 0.51

From this process, we are able to find results in:

./parent1/Analysis\_01/ConsensusAnalysis/,

and statistics in:

./parent1/Analysis\_01/ConsensusStatistics.

-n is the sample name, which will be the first column of each Analysis file, -r plot with R (output shown in Supplemental Figure 2), -b stands for minimum agreement of base calls, Parameter 0.51 shown in this command line represents homozygous calls.

The result folder contains a list of the the insertions, deletions, homozygous and heterozygous SNPs as well as some additional supporting information for the calls. In our study, we do not expect heterozygous SNPs since we are analyzing genomic sequence from inbred breeding lines. Heterozygous calls represent “inter-homoeologue polymorphisms” (Trick *et al*., 2009) where loci from each of the progenitor genomes are being aligned.

1. SHORE wgha: This is a hidden tool in SHORE that does not appear in the current manual. It seems that it is still under development and there is very little documentation for it. Based on inferences from the code, it uses map information to partition the region selected for assembly into several blocks. It then uses some, as yet, undetermined criteria for merging several blocks into superblocks so that every head block of a superblock is the tail of another. SUPERLOCAS ([Klein](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Klein%20JD%5Bauth%5D) *et al*., 2011) can then be specified to assemble each of the specified blocks. We only assembled the candidate QTL region on chromosome A10 ranging from positions 14,760,000 to 16,000,000. Assembly of the entire genome was not considered as it requires more time and memory than are presently available and also represents a goal beyond the scope of the current project.

Command: shore wgha -L ./run\_01/library.txt -i ../parent1/map.list.gz -l ../parent1/left\_over-paired.fl.gz -f ./canole/Brapa\_sequence\_v1.2.fa.shore -o ./assemble\_parent1 -s -C 10 -S 14760000 -E 16000000 -e ./parent1/left\_over-single.fl.gz

library.txt contains the insert distance information SHORE wgha requires. -i is the map file after merge. –l and –e represent the left over files. “s -C 10 -S 14760000 -E 16000000” specifies that SHORE only assembles the target region on chromosome ranging from 14,760,000 to 16,000,000. The assembly is done in folder ./assemble\_parent1. For each block, SHORE determines the reads that are aligned to a single end or both paired-ends for that region. It then arranges the files for the SUPERLOCAS input structure using: ./assemble\_parent1/BuildingSite/Input

then calls SUPERLOCAS.

The result of the assembly will be located in:

./assemble\_parent1/BuildingSite/superlocas

This directory will contain a folder for each superblock. Within each superblock folder will be a contigs.fa after running SUPERLOCAS. These contig files were then concatenated for entry into the Opera (Gao *et al*., 2011) scaffolding software.

1. The resulting list of polymorphisms output by SHORE consensus are those representing differences between the respective canola line and the *Brassica rapa* reference genome. This means that polymorphisms existing between the canola parent lines would be those identified in one parent line and not the other. An Excel macro was written to match all polymorphisms existing at the same location in both parents since these would represent monomorphic sites in the two canola lines and, thus would be considered non-informative for the present study. The Excel Macro code is as follows:

Option Explicit

Dim mRow, mDataStartRow, mDataEndRow, mDataColumn, mDataLength

Dim mIdColumn, mIdCopyColumn

Dim mItem

Dim mFound, mDataSheet, mIdSheet

'

Sub MatchMaker()

mIdSheet = "ID"

mIdColumn = 1

mIdCopyColumn = 3

mDataColumn = 2

mDataLength = 2

mDataStartRow = 2

mDataSheet = ActiveSheet.Name

ActiveCell.SpecialCells(xlLastCell).Select

mDataEndRow = ActiveCell.Row

Sheets(mDataSheet).Select

Range(Columns(mDataColumn + 1), Columns(mDataColumn + mIdCopyColumn)).Select

Selection.Insert

For mRow = mDataStartRow To mDataEndRow

Sheets(mDataSheet).Select

Cells(mRow, mDataColumn).Select

'If Len(Trim(ActiveCell.FormulaR1C1)) < mDataLength Then GoTo LoopAgain

mItem = ActiveCell.FormulaR1C1

Sheets(mIdSheet).Select

Columns(mIdColumn).Select

Set mFound = Selection.Find(What:=mItem, After:=ActiveCell, \_

LookIn:=xlFormulas, LookAt:=xlPart, SearchOrder:=xlByRows, \_

SearchDirection:=xlNext, MatchCase:=False)

If mFound Is Nothing Then

GoTo LoopAgain

Else

Selection.Find(What:=mItem, After:=ActiveCell, \_

LookIn:=xlFormulas, LookAt:=xlPart, SearchOrder:=xlByRows, \_

SearchDirection:=xlNext, MatchCase:=False).Activate

End If

Range(Cells(ActiveCell.Row, ActiveCell.Column), \_

Cells(ActiveCell.Row, ActiveCell.Column + \_

mIdCopyColumn - 1)).Select

Selection.Copy

Sheets(mDataSheet).Select

Range(Cells(mRow, mDataColumn + 1), Cells(mRow, mDataColumn + mIdCopyColumn)).Select

ActiveSheet.Paste

LoopAgain:

Next mRow

'ActiveWorkbook.Save

End Sub

## Scaffolding in Opera

Scaffolding is a problem of ordering and orienting contigs and typically incorporates paired-end read information. Opera (Gao *et al*., 2011) is a recently developed scaffolding tool. The first step is mapping paired-end reads to the contigs after which a scaffolding graph is built. Each node within the graph is a contig and an edge is directed between contigs whenever a mate-pair aligns to a separate contig. Opera attempts to find an order for the contigs where the total number of concordant edges is maximized. Ultimately, it returns improved contig results by merging some shorter contigs into longer ones. Those regions between the original contigs will be ‘N’s and the number is based on the distance information inferred from pair-end alignment.

We pulled out the paired-end reads from what SHORE prepared for SUPERLOCAS and used that as input for Opera.

Commands:

Perl preprocess.pl contigs.fa reads1.fastq reads2.fastq lib\_1.map bwa

opera contigs.fa lib\_1.map scarffolding

These are pretty easy and easy to understand commands, it first preprocess by aligning reads to the contigs using bwa (Li *et al*., 2009) followed by the scaffolding step.

## Estimating coverage of candidate genes using Python

For the 19 candidate genes identified using the comparative genomics approach, we stored the start and end position in two lists: startp, endp.

The following Python code creates a mask for each gene:

for i in range( len ( startp ) ):

tmp = [0 for j in range( endp[i]-startp[i] )]

positions.append( tmp )

for each mask, the initial value is 0 and if we find some read alignment cover this position, we set it to 1.

for line in file\_handle:

strs = line.split()

ch\_id = string.atoi(strs[0])

read\_start = string.atoi(strs[1])

if ch\_id == 10 and read\_start >= 14760000-100 and read\_start <= 16000000:

process(read\_start)

For the whole map.list, every time we read in one line, if the position it aligned to overlaps with the larger region. We set the mask at this start position and base pairs after this position according to the alignment. The process function is defined as below:

def process(read\_start):

for i in range(len(startp)):

if read\_start > startp[i]-100 and read\_start <= end[i]:

for k in range(100):

tmp = read\_start - start[i] + k

if tmp >= 0 and tmp < len(positions[i]):

positions[i][tmp] = 1

break

if some position of this read is in one of the gene regions we want, we set the mask to one for that position.

# Results

## Sequencing results

The Illumina HiSeq 2000 paired-end sequencing results were reported in two separate Fastq files, one for each parent. The Wichita sequence file resulted in a final uncompressed size of 44 Gigabytes which equated to 349,355,472 reads and a total sequence size of 32.2 Gigabases. The final uncompressed file size for IMC106 was 38 Gigabytes which translated to 299,691,522 reads for a final sequence length of 27.8 Gigabases. In the end, we were left with sequence data sets equivalent to 27X and 23X coverage of the estimated *B. napus* genome size (1.2 Gb) for Wichita and IMC106, respectively.

## Candidate genes identified using comparative genomics

A total of 257 genes were predicted to be located in the QTL interval based on the *B. rapa* genome for a final estimated gene density of 20.7 genes/1000 kb. These results are similar to those obtained in the study of Cheung *et al*. where several regions were studied and the gene density ranged from 7.8 to 27.2 genes/1000 kb with a mean of 16.2. Subsequent analysis of these 257 genes using existing *Arabidopsis thaliana* gene annotations narrowed the list of candidates down significantly to a total of 19 based on what we consider to be fairly liberal selection criteria (Supplemental Table 1). The predicted locations of these 19 genes, based on the *B*. *rapa* genome, were used as a filter for prioritizing polymorphisms where those existing within the predicted gene space or 1kb upstream or downstream of that space were considered more likely to have functional consequences than those located outside of this space. The same criteria were used in a recent and very similar study for defining regions where polymorphisms would be more likely to result in functional consequences (Tollenaere *et al*., 2012).

## Polymorphisms identified via SHORE consensus

3.2.1 Total polymorphisms identified in the QTL region Summary figures of the following can be found in Supplemental Table 2. When searching the entire 1 Mb QTL interval for small InDels (<3 bp) and SNPs, a total of 6026 SNPs and 1312 InDels were discovered when comparing Wichita with the *B. rapa* reference genome. With regards to IMC106, a total of 5787 SNPs and 1248 InDels were identified and suggest that Wichita may be slightly more diverged from the reference genome than IMC106 in this region. This represents a SNP rate of 0.49% for Wichita and 0.47% for IMC106, both of which are similar, but lower than, the results reported by Cheung *et al*. (2009) when comparing the A genome of *B. rapa* to that of *B. napus*. The InDel rates of 1.05/kb for Wichita and 1.01 for IMC106 are also similar but lower than the results of the same study.

Summary figures of the following can be found in Supplemental Table 3. When comparing the output from the Excel macro identifying genomic features specifically differentiating the *B. napus* parent lines, 962 InDels and 3998 SNPs were discovered. This represents an InDel rate of 0.77 /kb and a SNP rate of 0.32%. Of the InDels, 471 were defined as insertions and 491 were defined as deletions relative to the reference genome.

3.2.2 *P*olymorphisms identified in the candidate gene search space The search focused only on those genes in the region with functional annotation to root and/or flowering discovered an expectedly much smaller number of polymorphisms was discovered. A total of 22 InDels was discovered and was evenly divided between insertions and deletions with 11 of each. When focusing on SNPs, a total of 137 were discovered. Interestingly, the Wichita parent line had nearly twice as many SNPs (88) relative to the *Brassica rapa* reference genome than did the IMC106 parent line (49).

## Running time

Running time will vary depending upon the hardware environment as well as other users’s tasks that are running in parallel. The summary below is just an approximate time for department machine like eggs, bacon, bananas, coconuts.

For shore pipeline:

1. Preprocess: 4-6 hours.
2. Import: 16-20 hours for each parent.
3. Mapflowcell: 3-4 days for each paired-end data
4. Correct4pe: 1 day for each parent
5. Merge: 6 hours for each parent
6. Consensus: 1 day for each parent
7. Wgha: 3-5 hours for the QTL region
8. For opera: 1day for each parent.

## Contig sequence size and number based on SUPERLOCAS analysis

The contig statistics are illustrated in Supplemental Table 4 of the supplemental material. For Wichita, the N50 is 395 and the total assembled length is 3.74 Mb. The length of the longest scaffold is 10.91 Kb. For IMC106, the N50 is 495 and the total assembled length is 2.99 Mb. The length of the longest IMC106 contig is 5.87 Kb. The smaller the N50 size for Wichita is surprising since the sequence data set for this parent was quite a bit larger. Perhaps the larger data set is resulting in a larger number of small fragments. In contrast, Wichita has a maximum contig size that is nearly double that of IMC106. In this case, the larger data set has improved our ability to construct longer contigs, so the longest contigs is longer than that of IMC106.

## Scaffold sequence size and number based on OPERA analysis

From figure X of the supplement material we can see that the number of contigs below 1000 bp was decreased as a result of the scaffolding procedure so that a larger number of contigs greater than 1000 bp was observed. The shorter contigs have now been merged into longer ones. Specifically, 10 non-singleton scaffolds were observed in Wichita and 12 non-singleton scaffolds were observed in IMC106. However, the final results were relatively insignificant as evidenced by little change in the final N50 values of both parents.

## Results of scaffold alignment to the reference genome

We successfully aligned the scaffolds from section 3.6 to the reference genome using bwasw found in the bwa alignment framework (Li *et al*., 2009). The results of this have yet to be analyzed in detail but the goal is to account for structural variation beyond the 3 bp limit imposed as a result of the InDel calling procedure in SHORE consensus. Structural variation of this scale is known to exist among members of the *Brassica* species.

## Results of gene coverage estimates

We estimated the coverage using the corrected paired end data for each parent, implemented in python. A detailed summary of the coverage of each gene is presented as a part of Supplementary Table 1. An average of 96% and 97% coverage for each gene was attained for Wichita and IMC106, respectively. Coverage ranged from 73% to 100% in Wichita and 62% to 100% in IMC106 with 13 of the 19 genes having 99% or greater coverage. The gene with the lowest coverage (At5g06140) was the same in each parent and was also one of the most polymorphic genes with a total of 17 SNPs and 5 InDels In fact, the 3 genes with the lowest coverage were all adjacent to one another within a 150 kb interval. These three genes accounted for 7 of the 22 InDels discovered in the list of candidate genes and 59 of 137 SNPs suggesting regions of extreme sequence divergence between the cultivars and the low level of coverage suggests this is also true for the comparison of this region with *B. rapa*.

## Problems we encountered:

Here are several problems we met in the project and the solutions:

1. The SHORE import has an error message of non-unique read id.

Solution: The original fastq file was separated into several smaller fastq files and then compressed so that data transfer would be possible without physical delivery (i.e. on a hard drive). The data was concatenate back upon recieving. The problem was that the partition lost one line at the end of each of the subdivided fastq files so that each read in the fastq file has 4 lines meaning that one missing line will cause trouble with the next read. As a solution, we just deleted some reads at the end and beginning of each sub-file.

1. The default temporary folder for SHORE is /tmp, and shore takes a lot of space while running, it may use up all the space in /tmp.

Solution: We specified the temporary directory for SHORE by using shore –T tmp\_directory each time we ran the commands.

1. In the mapflowcell step, the program stopped due to a weekly shutdown by the department while merging the map results so that running it again will take another 4 days.

Solution: We have to concatenate the map.lists and left-over.fls into one file by linux command line, and we need to sort the left-overs by read id by using another shore subprogram SHORE sort, or the merge step will tell you the left-over reads is not in sorted order.

1. We tried to assemble a larger region or the whole genome, but it takes much more space and time beyond the resource we have, so we give it up and only assembled the region of interest on chromosome A10.

# Discussion

As with most experiments ours was not without its limitations. Most notable was our use of the *B. rapa* genome as a reference since it represents only one of the two progenitor genomes present in the allopolyploid *B. napus* genome (U, 1935). The most likely consequence of this will be the mapping of reads originating from the *B. oleracea* genome onto the A genome reference which can be seen in the summary statistics output by SHORE consensus (Supplemental Figure 2). When analyzing the ‘Number of hits in genome per read’ panel, it is apparent that many reads have several hits in the genome, especially when considering the same figure from the example provided in the SHORE manual (Supplemental Figure 3) where there is a small peak at 41 hits which represents the “repetitive sequence”. However, it appears that this may not have had a major impact on our study, at least not in terms of calling SNPs (output as homozygous SNPs) and InDels. The SNP rates of 0.49% and 0.47% for Wichita and IMC106, respectively, are less than the lowest rate of 0.82% (maximum rate of 1.98%) reported by Cheung *et al*. (2009) in their comparison of the *B. rapa* and *B. napus* A genomes. However, the rate of divergence between *Brassica* genomes is also expected to be a product of the region of the genome that is being studied. This is a result of the hypothesized 2 stages of polyploidization occurring during the evolution of the *Brassica* genomes where the paralogous regions originating from the first polyploidy event (termed MF1 and MF2 by Wang *et al*. (2011)) are more divergent than those occurring as a result of the second polyploidy event (LF). In the study of Cheung *et al*. (2009), the A genome comparisons were being made between what would be considered one of the two MF regions so their results are not entirely translatable as the region we are studying exists in an LF region. With this in mind, Cheung *et al*. (2009) reported a SNP rate of 0.47% in a comparison of a less diverged region of the C genomes of *B. oleracea* and *B. napus* which is nearly exactly the same as the results we obtained in this study. Only the re-analysis of the data with the inclusion of a *B. oleracea* reference will provide a more convincing answer but our comparisons with comparable experiments provide some support for the validity of the polymorphism results we are presenting.

The main objective of the present study was to identify a large number of polymorphisms existing within the specified QTL region of the *B. napus* genome of two inbred lines. The discovery of nearly 4,000 SNPs (0.32%) within the target genomic region seems high relative to other SNP discovery studies comparing *B. napus* cultivars where rates were estimated to be between 0.047 and 0.084% (Trick *et al*. 2009). However, these results were obtained from transcriptome sequencing and ours represent polymorphisms occurring in both the intra- and intergenic regions of the genome which are expected to differ substantially in their SNP rates. When looking strictly at the intragenic regions, the SNP rate between Wichita and IMC106 drops to 0.23% but is still over twice that observed in the Trick *et al*. (2009) paper. This result does include the intronic regions within these genes so these values are expected to drop once an analysis of only the exons is conducted.

The results of these data have many applications in our understanding of the genetics underlying the quantitative phenotypes, flowering time and root pulling force. First, the thousands of polymorphisms discovered within the QTL interval can be converted into high-throughput molecular marker assays for utilization in future fine-mapping (Eshed and Zamir, 1994) efforts aimed at narrowing the QTL interval down to a much smaller genomic region. These fine-mapping studies will also provide evidence in explaining the basis of the strong genetic correlation (rg = 0.66; p < 0.0001) observed between flowering time and root pulling force in this population. More specifically, in advanced recombinant populations created in the coming months by crossing specific lines from the mapping population back to the IMC106 parent line, these markers will be utilized to track recombination events occurring within the current interval so that individuals identified as carrying recombination events between candidate gene loci can be further phenotyped for the traits of interest. Identification of individuals where trait values were inconsistent with the observed correlation would provide evidence for genetic linkage. Results to the contrary would provide evidence for pleiotropy.

One of the polymorphisms identified in our study is of particular interest and deserves further discussion. Prior to the current study, the *a priori* prediction in terms of specific gene candidates within the QTL interval having functional effects on one or both phenotypes was the major flowering time gene *FLC* (At5g10140). *FLC* is a transcription factor and one of, if not the most, well studied flowering time genes of all those discovered to date (Michaels 2009, Pose *et al*. 2012). Its effects on flowering are no longer under debate and certainly qualify it as the most likely candidate for the observed variation in flowering time. A recent study in Arabidopsis also found evidence supporting its binding to more than 500 genes where both positive and negative regulatory effects were observed (Deng *et al*. 2011). Our results have found a deletion existing in the fifth intron of the *FLC* paralog on chromosome A10 of the Wichita parent line which has been confirmed by Sanger sequencing of the cloned gene (Supplemental Figure 4). While this is by no means conclusive evidence for a functional polymorphism, it is concordant with results obtained in studies done in *B. rapa* (Yuan *et al*. 2009). In the study of Yuan *et al*. (2009), it was determined that a SNP at position 104 of the fifth intron of *FLC* was significantly associated with delayed flowering with an affect of up to 20.5 days. The deletion detected in our study is predicted to locate at position 131of the fifth intron of *FLC* so the results are not exact. However, it is in the same general vicinity of the same intron making these results highly encouraging for the efficacy of our approach. In addition, SNPs and InDels were discovered in two other strong candidate genes (AT5G10510 and AT5G10250) that are annotated to have direct impacts on root growth and development. Conversion of these polymorphisms to marker assays will be crucial in answering whether *FLC* is impacting each of these traits directly or if it is merely affecting flowering time and is linked to these root-specific genes, as outlined in the previous paragraph.

In regards to the SUPERLOCAS homology-guided assembly analysis, we feel these results are not too bad considering the extreme degree of redundancy in the *B. napus* genome resulting from the polyploidy nature of its evolutionary history (Wang *et al*., 2011). In addition, as mentioned previously, our reference only represents half of an already complicated genome. We are certain that the results of this analysis have some inherent errors as a result of these limitations so any structural variation we identify will need to be validated using other methods.

Conclusions

We have successfully used the existing second generation analysis tools SHORE (Ossowski *et al*., 2008), GenomeMapper (Schneeberger *et al*. 2009), SUPERLOCAS (Klein *et* al. 2011) and Opera (Gao *et al*., 2011) to characterize a recently identified and biologically important QTL region of the *Brassica napus* genome. The results of the SNP and InDel caller SHORE consensus have provided valuable information for the next steps involved in elucidating the genetics underlying the QTL to the extent that a candidate polymorphism has been identified. Further analysis of the homology-guided assembly results should provide further information about the extent of divergence between the *B. napus* parent lines used in this study. We suggest this approach as an effective method for improving the efficiency of moving from QTL to functional gene.

acknowledgements

The authors would like to thank C. Boucher for many helpful suggestions.

*Funding*: Cargill Specialty Canola Oils (Fort Collins, CO) funded the creation of the doubled haploid mapping population, all resources associated with the QTL mapping experiments and the Illumina HiSeq 2000 sequencing run.

References

Andersen,J.R. and Lubberstedt,T. (2003) Functional markers in plants. *Trends Plant Sci*., **8**, 554-560.

Arihara,J. and Crosbie,T.M. (1982) Relationship among seedling and mature root system traits of maize. *Crop Sci*. **22**, 1197-1202.

Botstein D., *et al*. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*., **32**, 314-331.

Buckler,E.S. *et al*. (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714-718.

Cheung,F. *et al*. (2009) Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. *Plant Cell*. **21**, 1912-1928.

Ekanayake,I.J. *et al*. (1985) Root pulling resistance in rice: inheritance and association with drought tolerance. *Euphytica*. **34**, 905-913.

Eshed,Y. and Zamir,D. (1994) A genomic library of *Lycopersicon pennellii* in *L. esculentum*: A tool for fine mapping of genes. *Euphytica*, **79**, 175-179.

Falconer,D.S. and Mackay,T.F.C. (1996) Introduction to Quantitative Genetics. Longman Group Ltd, Harlow, Essex, England.

Fourmann,M.P., *et al*. (1998) The two genes homologous to Arabidopsis FAE1 co-segregate with the two loci governing erucic acid content in *Brassica napus*. *Theor. Appl. Genet*., **96**, 852-858.

[Gao](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Gao%20S%5Bauth%5D),S. *et al*. (2011) Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences. *J. Comput. Biol*. **18**, 1681–1691.

Haldane,J.B.S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet*., **8**, 299-309.

Hayes,H.K. and Johnson,I.J. (1939) The breeding of improved selfed lines of corn. *J. Am.Soc. Agron*., **31**, 710-724.

Hunter,R.L. and Markert,C.L. (1957) Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science*. **125**, 1294-1295.

Kamara,A.Y. (2003) Agronomic performance of maize (Zea mays L.) breeding lines derived from a low nitrogen maize population. *J. Agr. Sci*., **141**, 221-230.

Kevern,T.C. and Hallauer,A.R. (1983) Relation of vertical root-pull resistance and flowering in maize. *Crop Sci*., **23**, 357-363.

[Klein](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Klein%20JD%5Bauth%5D),J.D. et al. (2011) LOCAS – A Low Coverage Assembly Tool for Resequencing Projects. *PLoS One*. **6**, e23455.

[Li](http://bioinformatics.oxfordjournals.org/search?author1=Heng+Li&sortspec=date&submit=Submit),H. and [Durbin](http://bioinformatics.oxfordjournals.org/search?author1=Richard+Durbin&sortspec=date&submit=Submit),R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Mackay,TFC (2001) The genetic architecture of quantitative traits. *Annu. Rev. Genet*., **35**, 303-339.

Mauricio,R. (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat. Rev. Genet*., **2**, 370-381.

Michaels,S.D. (2009) Flowering time regulation produces much fruit. *Curr. Opin. Plant. Biol*., **12**, 75-80.

Montesinos-Navarro A., *et al*. (2011) *Arabidopsis thaliana* populations show clinal variation in a climatic gradient associated with altitude. *New Phytol*., **189**, 282–294.

Muller,H.J. (1916) The mechanism of crossing-over. *Am. Nat*., **50**, 193-221.

Nakamura,Y. et al. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, **235**, 1616-1622.

[Ossowski](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Ossowski%20S%5Bauth%5D),S. *et al*. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res*., **18**, 2024–2033.Posé, D. et al. (2012) The end of innocence: flowering networks explode in complexity. *Curr. Opin. Plant*. *Biol*., **15**, 45–50.

Sax,K. (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*. **8**, 552-560.

[Schneeberger](http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=PubMed&term=%20Schneeberger%20K%5Bauth%5D),K. *et al*. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol*. **10**, R98.

Springer,N.M. *et al*. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*., **11**, e1000734.

Stinchcombe,J.R. *et al*. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proc. Natl. Acad. Sci. USA*., **101**, 4712–4717.

Sturtevant,A.H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool*., **14**, 43-59.

Tollenaere,R. *et al*. (2012) Identification and characterization of candidate Rlm4 blackleg resistance genes in *Brassica napus* using next-generation sequencing. *Plant Biotechnol. J*., **10**, 709-715.

Trick,M. *et al*. (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol. J*., **7**, 334-346.

Tuberosa,R. *et al*. (2002) Identification of QTLs for root characteristics in maize grown in hydroponics and analysis of their overlap with QTLs for grain yield in the field at two water regimes. *Plant Mol. Biol*., **48**, 697–712

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

Wagner,G.P. and Zhang,J. (2011) The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet*., **12**, 204-213.

Yuan,Y. *et al*. (2009) A naturally occurring splicing site mutation in the *Brassica*

*rapa* FLC1 gene is associated with variation in flowering time. *J. Exp. Bot*., **60**, 1299-1308.

Wang,X. *et al*. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet*., **43**, 1035–1039.

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)