Towards unambiguous transcript mapping in the allotetraploid Brassica napus

Parkin, Isobel A. P.; Clarke, Wayne E.; Sidebottom, Christine; Zhang, Wentao; Robinson, Stephen J.; Links, Matthew G.; Karcz, Steve; Higgins, Erin E.; Fobert, Pierre; Sharpe, Andrew (Andy)

Publisher's version / Version de l'éditeur:
https://doi.org/10.1139/G10-053
Genome, 53, 11, pp. 929-938, 2010-11-05

Towards unambiguous transcript mapping in the allotetraploid *Brassica napus*

Isobel A. P. Parkin[1], Wayne E. Clarke[1,2], Christine Sidebottom[3], Wentao Zhang[1], Steve J. Robinson[1], Matthew G. Links[1,4], Steve Karcz[1], Erin E. Higgins[1], Pierre Fobert[3], Andrew G. Sharpe[3]

1. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N 0X2, Canada.

2. Department of Computing Science, 176 Thorvaldson Building, University of Saskatchewan, 110 Science Place, Canada, Saskatoon, SK. S7N 5C9, Canada.

3. NRC-Plant Biotechnology Institute, 110 Gymnasium Place, Saskatoon, SK, S7N 0W9, Canada.

4. Department of Veterinary Microbiology, WCVM, University of Saskatchewan, 52 Campus Drive, Saskatoon, SK. S7N 5B4, Canada.

Corresponding Author:

Isobel Parkin

Agriculture and Agri-Food Canada

107 Science Place

Saskatoon, SK, S7N 0X2. CANADA

Email: isobel.parkin@agr.gc.ca

Abstract

The architecture of the *Brassica napus* genome is marked by its evolutionary origins. The genome of *B. napus* was formed from the fusion of two closely related diploid Brassica species, both of which evolved from an hexaploid ancestor. The extensive whole genome duplication events in its near and distant past result in the allotetraploid genome of *B. napus* maintaining multiple copies of most genes which predicts a highly complex redundant transcriptome that can confound any expression analyses. A stringent assembly of 142,399 *B. napus* expressed sequence tags allowed the development of a well differentiated set of reference transcripts which were used as a foundation to assess the efficacy of available tools for identifying and distinguishing transcripts in *B. napus*; including microarray hybridization and 3'anchored sequence tag capture. Microarray platforms cannot distinguish transcripts derived from the two progenitors or close homologues, although observed differential expression appeared to be biased towards unique transcripts. The use of 3' capture enhanced the ability to unambiguously identify homologues within the Brassica transcriptome but was limited by tag length. The ability to comprehensively catalogue gene expression in polyploid species could be transformed by the application of next generation sequencing technologies that will capture millions of long sequence tags.

Introduction

The study of global gene expression analysis is a powerful tool for understanding complex biological phenomenon, deciphering interacting pathways and uncovering novel genes controlling traits of interest. The application of gene expression tools in the study of segregating genetic populations has given rise to the term 'genetical genomics' which describes the use of expression data to uncover genetic mechanisms governing byzantine phenotypic variation (Jansen and Nap 2001). There is now an opportunity to apply these types of analyses to non-structured collections of diverse germplasm to combine the strength of association mapping with the acuity of transcript analysis to elucidate intractable phenotypes, which is proving to be highly informative in studying the genetic architecture of transcript variation in humans (Zhang et al. 2008).

Gene expression analysis is limited by the available data for a species of interest which dictates the form of the offered tools for performing such experiments. For many species there is now sufficient published sequence data to allow the development of gene specific microarray platforms that permit the presence and intensity of thousands of genes to be queried at once (Galbraith 2006). The scope of such platforms is determined by the available resources. For example a 60-mer oligonucleotide array has recently been developed to allow approximately 90,000 Brassica transcripts to be assessed (Trick et al. 2009a) which is expected to represent roughly 71% of the gene content based on comparison with the annotated *Arabidopsis thaliana* gene sequences (TAIRv8: www.arabidopsis.org). Serial analysis of gene expression (SAGE) is an accurate qualitative transcript profiling method that is not reliant upon existing sequence data, allowing all expressed transcripts to be captured and counted but not necessarily annotated

(Robinson et al. 2004). Short sequence tags (14-26 bp) are acquired from the 3' end of transcripts providing an accurate representation of transcript abundance and type. However, to match the tags to their genomic region of origin and extract the maximum value from the SAGE data requires a fully sequenced genome. The use of SAGE analysis to determine the *B. napus* seed transcriptome was recently assessed; almost 16,000 transcripts could be identified although it was apparent that the length of the tag was insufficient to unambiguously identify all homologous copies of each gene (Obermeier et al. 2009). Interestingly this study also highlighted significant anti-sense activity with almost 30% of the genes matched by tags originating from both genomic strands. The prevalence of such alternative transcript processing could be a function of the polyploid genome and will further complicate any tag based profiling.

The difficulty in matching sequence tags to Brassica transcripts although not unique to this species is exacerbated by the evolutionary history of the Brassicaceae. *Brassica napus* is a relatively young allotetraploid formed from the fusion of the closely related A and C Brassica genomes (U 1935; Parkin et al. 1995). More recent molecular genetic analyses of *B. napus* and the modern relatives of the progenitor genomes, *B. rapa* and *B. oleracea* respectively, indicates that the Brassica genus has evolved from a hexaploid ancestor (O'Neill and Bancroft 2000; Rana et al. 2004; Parkin et al. 2005; Town et al. 2006; Lysak et al. 2007). Comparative mapping of Brassica species with the related crucifer model *Arabidopsis thaliana* confirms these studies, in most instances for each region of *A. thaliana* three genomic segments are identified in the Brassica genome (Parkin et al. 2005; Town et al. 2006; Lysak et al. 2007). Since the triplication event dated 14-24 million years ago, random gene loss and genetic drift has differentiated both the composition of the segments and the orthologous gene copies within the A and C genome

(Town et al. 2006; Yang et al. 2006; Cheung et al. 2009). The resultant genome structure dictates that for any gene within *B. napus* there could be from two to six homologues present (Figure 1).

Notwithstanding possible alternative transcript processing events the multiple related gene copies confuse accurate transcript differentiation that can confound any gene expression study and limit the ability to elucidate the genetic mechanisms controlling a trait. The available Brassica microarray platforms cannot resolve transcripts derived from the homoeologous gene copies within the A and C genomes of *B. napus* (Trick et al 2009) and the limits of tag based expression profiling in Brassica species has yet to determined. In the present study, initial data sets generated for the purposes of expression quantitative trait loci (eQTL) and single nucleotide polymorphism (SNP) analysis respectively were studied to determine if the added expense of sequence based expression analysis is warranted by the gain in resolution of transcript determination offered compared to microarray studies. Insights into the efficacy of different gene expression techniques in Brassica species are discussed at the theoretical and practical level comparing three protocols, microarray analysis and long versus short 3' anchored tag sequencing.

Materials and Methods

Plant Material

The *B. napus* lines used in this study included the parental lines for a reference doubled haploid mapping population: DH12075, a Canadian annual canola line (generated by Gerhard Rakow and Ginette Séguin-Swartz, Agriculture and Agri-Food Canada), and PSA12, a resynthesised *B. napus* line (generated by Monica Beschorner and Derek Lydiate, Agriculture and Agri-Food

Canada). Tissue for RNA extraction was collected from: developing seeds 21 days post-anthesis grown under field conditions; etiolated seedlings (5-7 days post germination under sterile conditions in the dark); and juvenile leaves (1st-4th true leaves). Total RNA was extracted from developing seeds using a method modified from that of Onate-Sanchez and Vicente-Carbajosa (2008). In summary, seed embryo tissue (fresh weight approximately 20 mg) was ground with liquid nitrogen and mixed with 700 µL extraction buffer (0.4 mM LiCl, 0.2 M Tris-Cl pH.8.0, 25 mM EDTA, 1% SDS and 2% polyvinylpyrrolidone, 2% β-mercaptoethanol), followed by treatment with an equal volume of chloroform, an equal volume of phenol and finally an equal volume of phenol: chloroform: isoamyl alcohol (25:24:1). The final supernatant was precipitated with 1/3 volume of 8 M LiCl for 1 hour at -20°C. After centrifuging at 13,000 rpm for 30 min, the pellet was washed with 1 ml of 75% ethanol and dissolved in 40 µL of DEPC-water. DNA was removed from total RNA using DNase I (Invitrogen, Catalogue: 18068-065) according to the manufacturer's instructions. RNA was extracted from etiolated seedlings and juvenile leaves using the Qiagen RNeasy Mini kit according to the manufacturers protocol (Qiagen Inc, Mississauga, Ontario).

Microarray Hybridisation and Data Analysis

Gene expression analysis was carried out using Agilent Brassica Gene Expression microarray (design id: 022520) for DH12075 and PSA12. Four biological replications were performed with dye-swaps applied to the biological replicates to minimize dye incorporation bias from the two colour system (Lee et al. 2004). cRNA was amplified and labeled with either cy3 or cy5 from total RNA (2 µg) using the Quick Amp labeling kit, Two Color (Agilent, Catalogue: 5190-0444) according to the manufacturer's instructions. Amplified and labeled cRNA was purified with

Qiagen RNeasy Mini Kit (Qiagen, Catalogue 74104) and quantified with the NanoDrop ND-1000. Labeled cRNA (2 µg) was fragmented and subsequently hybridized in dual labeled reactions to the Agilent 4x44K *Brassica* array using the Gene Expression hybridization kit (Agilent, Catalogue: 5188-5242) according to the manufacturer's protocol. Hybridization was performed in an Agilent Microarray Hybridization Chamber (Agilent, Catalogue: G2534A) for 17 h at 65°C with a rotation of 10 rpm. Slides were washed with the Agilent Gene Expression Washing buffers I and II (Agilent, Catalogue: 5188-5325 and 5188-5326) according to the manufacture's protocol.

Arrays were scanned at 5 µm resolution with the GenePix 4000B scanner, the fluorescence data was extracted from the resulting image files using Gene Pix Pro 6.0 and normalized with the LOWESS method using BASE 1.2 (Dudoit et al. 2002). Normalized Data was exported from BASE and imported to GeneSpring GX 10.0 (Agilent Technologies) for further analysis. The parental lines were analysed for differential gene expression using the Student's t-test with a P value significance threshold of 0.05, a False Discovery Rate (FDR) of 0.05 and a minimum two fold cut-off.

Differentiating homologous Brassica transcripts

Single nucleotide differences were determined between Brassica homologous transcripts based on a protocol described in Eveland et al. (2008). 3'-anchored cDNA libraries were generated from the parental line DH12075 as in Eveland et al. (2008) except *Aci*I was used to generate 3' cDNA fragments of the optimal size range for amplification during the Roche 454 Titanium sequencing protocol, as determined by *in-silico* digestion of the Brassica reference transcripts.

Additionally, primer/adapters modified for the Titanium chemistry were implemented in the protocol. A full description of the modified protocol will be published separately. Roche 454 Titanium sequencing was carried out at the NRC-Plant Biotechnology Institute following the procedure described by Margulies et al (2005) with modifications for the Titanium chemistry as described in protocols supplied by the manufacturer (Roche, Laval, Quebec).

The 454 data was assembled against a Brassica reference of 46,648 transcripts (described below in results) using NGen (DNAStar Inc, Madison, WI) with the following parameters: match size: 19; match spacing: 10; minimum match percentage: 90; match score: 10; mismatch penalty: 25; gap penalty: 25; and maximum gap: 15. Custom Perl code was developed to: parse the individual ACE files from the resultant assembly; convert the ACE files to maintain the coordinates of the reference sequence so as to allow comparison across all six analyses; and to follow the origin of each transcript within the assembly. The code for conversion of the ACE files is maintained in the software package APED (http://sourceforge.net/projects/aped). SNPs were identified in each assembly using the POLYBASES polymorphism software package (Marth et al. 1999) with parameters based on the work of Barbazuk et. al. (2007). Custom Perl scripts were written to parse the POLYBASES output to identify variation resulting from homoeologous and paralogous loci.

SAGE analysis

In silico SAGE analysis was performed on the *Brassica napus* reference transcripts using the method described by Robinson et al. (2004) except that in the absence of a complete genome sequence no assumptions were made regarding the extension of UTR sequences. The orientation

of the transcript collection was determined by sequence alignment to Arabidopsis gene sequences using BLASTN (Altschul et al. 1990). Directional cloning of cDNA sequences during library construction allowed the orientation of unigenes to be assigned for many of the Brassica specific sequences. Unigenes where the orientation remained ambiguous were excluded from further analysis. Custom Perl scripts were used to perform *in silico* restriction digests and tag capture for each reference transcript using thirteen anchoring enzymes (*Acc*II, *AciI*, *Alu*I, *CivR*I, *Dpn*I, *Hae*III, *Hha*I, *Hpa*II, *Mae*I, *Nla*III, *Rsa*I, *Taq*I and *TspE*I) and three SAGE methodologies SAGE, LongSAGE and SuperSAGE that result in tags of length 14 bp, 21 bp and 26 bp respectively (Velculescu et al. 1995; Saha et al. 2002; Matsumura et al. 2005). Only the canonical tag (the tag proximal to the 3' end of each transcript) was used to discriminate among Brassica transcripts.

Results

Generating the DH12075 Reference Transcript Dataset

A collection of 142,399 *B. napus* expressed sequence tags (ESTs) generated from 18 cDNA libraries derived from different tissues of DH12075 as shown in Table 1 (CN825827-CN829362, CN829364-CN829515, CN829517-CN831073, CN831075-CN831324, and EV090678-EV227586) were assembled using TGICL (Pertea et al. 2003) to generate a *B. napus* distinct set of reference transcripts (Figure 2). Parameters were established for assembling the ESTs in order to maximize separation of homologous transcripts within the collection. Through the alignment of EST data from the two progenitor species, *B. rapa* (A genome) and *B. oleracea* (C genome), it was determined that the maximum level of sequence similarity between orthologues was 98% (Figure 3). Assembly of the *Brassica* sequence data was completed in a series of

progressively more stringent assemblies using default Cap3 parameters except for varying the overlap parameter, within the range 82% to 99%. Comparing the resulting assemblies to Arabidopsis gene sequences confirmed that the maximal resolution of paralogues occurred with a Cap3 overlap parameter of 98% and the resultant assembly of the *B. napus* EST collection using these parameters identified 46,648 unique transcripts. Comparison of the reference sequences with *A. thaliana* estimated the gene coverage to be 46% and the number of full length gene models in excess of 20% (Table 1). The Brassica reference transcript set, including the sequence data of individual members of the assembled contigs, and the alignment of the reference set against the *A. thaliana* genome can be accessed at http://aafc-aac.usask.ca/cgi-bin/gbrowse/gbrowse/BAGI2/.

The Brassica oligo sequences from the Agilent array were aligned to the reference transcripts using BLASTN to allow comparisons between the microarray analyses and the transcript data generated from the 454 sequencing. Since the Brassica oligos were derived from multiple *B. napus* genotypes up to a two base pair mismatch was allowed to determine transcripts that would be expected to be identified during array hybridizations, though perhaps with a lower signal intensity depending on the position of the nucleotide variation within the oligo (Rennie et al. 2008). In addition, the designated unigenes from which each probe was designed were also matched with the *B. napus* transcripts to validate the correspondence of the individual probes. Using these criteria 15,742 probes matched 16,656 DH12075 transcripts. As might be expected due to the length of the probes and the complexity of the genome, 3,530 oligos (22%) matched multiple reference transcripts with 23 probes matching 10 or more unique transcripts.

Array Based Expression Analysis in Brassica napus

The Agilent Brassica 4 x 44 K array was selected for microarray experiments in *B. napus* due to the increased efficiency offered by the array design. These arrays contain 43,809 gene specific 60 mer oligos arranged in 4 replicate arrays. The oligos were designed from 40,206 Brassica unigene sequences that were derived from a mixture of *B. napus* genotypes (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=oilseed_rape). BLASTN analysis (Altschul et al. 1990) with an E value threshold of 1 x E-06 identified the closest homologous gene identifier in *A. thaliana* for each of the unigene sequences. Thirty-five thousand two hundred and forty-four unigenes shared sequence similarity with 90 *A. thaliana* plastid gene codes and 15,596 nuclear genes. The remaining 4,962 unigenes demonstrated no significant sequence homology with an annotated *A. thaliana* gene code.

The microarray data analysed were from experiments conducted as part of a project to carry out expression QTL analysis in *B. napus* to associate gene expression changes with genetic polymorphisms in a doubled haploid segregating population and to relate these to variation in seed quality traits. Previous studies have established that ~20 days after flowering (DAF) was the critical stage for *B. napus* seed development regarding cell proliferation and oil deposition (O'Hara et al. 2002; Dong et al. 2004). Thus developing embryos were collected from *B. napus* lines at 21 DAF in the field for RNA extraction and gene expression studies. Comparing the parental lines, 2,331 probe sequences identified genes that were significantly differentially expressed, with 1,614 up-regulated in DH12075 and 717 up-regulated in PSA12 (Figure 4). Based on the array design these 2,331 probe sequences recognised 2,235 unigene accessions. In all but one instance where different probes had been designed to a single transcript the second or

third probe displayed an identical expression pattern if not equivalent intensity. However, 238 unigene sequences that were inferred to be differentially expressed based on one or more oligos possessed additional probes which showed no change in signal intensity between the parents.

The corresponding transcripts from the reference DH12075 collection could be determined for 760 of the differentially hybridised probe sequences, which corresponded to 878 unique transcripts. Interestingly, considering the lack of resolving power for the 60 bp sequences, the majority of the probes (84%) were matched to unique transcripts (chi-square=14.31; $p<0.0002$). For the remaining 124 probes (16%) that matched multiple transcripts, 75% (93 probes) identified only two transcripts.

Differentiating Brassica Transcripts using 454 Sequence Data

An anchored 3' cDNA library was developed to provide a foundational resource for SNP discovery and mapping in *B. napus*. The method generates long sequence tags (200-400 bp) that should improve effective discrimination of transcript sequences and in multiple genotypes will maximize the likelihood of capturing equivalent transcript regions for improved SNP discovery (Eveland et al. 2008). The enzyme *Aci*I was selected through *in-silico* digestion of Brassica unigene sequences to provide an optimal profile of transcript fragments for Roche 454 Titanium sequencing. An anchored 3' cDNA library was generated from combined etiolated seedling and juvenile leaf tissue for the parental line DH12075. Sequencing resulted in 904,582 transcripts with the median read lengths of 285 and 308 bp for two different regions on a Titanium 454 sequencing plate.

The Brassica 454 EST data was assembled against the DH12075 reference transcript dataset using NGen and custom Perl scripts were developed to parse the resultant ace files. A total of 577,094 sequence reads were assembled against 23,543 reference transcripts with an average of 12 sequences per contig and 18,020 contigs with a read depth of at least 2. The individual ace files for each transcript were analysed using POLYBASES (Marth et al. 1999) to identify nucleotide variations that differentiate homoeologous and paralogous transcripts which had been co-assembled (Figure 5). The increased error rate for homopolymer calls in the 454 data can lead to spurious insertion-deletion events so such polymorphisms were ignored in the present analysis. This determined that potentially 3,259 (18%) of the contigs contained multiple homologous Brassica transcripts that could be differentiated based on simple nucleotide polymorphisms. The matching of the majority of the 454 sequences uniquely to a single transcript indicated that the generation of the reference dataset had been successful in resolving the duplicate copies. The instances where homologous 454 reads had been co-assembled likely resulted from the absence of the related duplicate copies within the reference dataset. The presence of anti-sense transcripts was also observed, with 362 of the 3,680 most highly abundant transcripts possessing at least 20 reads matched to both strands of the transcript.

Potential for Serial Analysis of Gene Expression (SAGE) in an allotetraploid genome

SAGE has the ability to provide a digital measure of an organism's transcriptome where the frequency of the tag is directly proportional to the frequency of the transcript from which it is derived. The efficacy of SAGE has been demonstrated in many model organisms including *A. thaliana* (Pleasance et al. 2003; Robinson et al. 2004; Matsumura et al. 2005). However, due to the isolation of short tag sequences and the non-random nature of genome sequence it is

impossible to unambiguously distinguish the origin of many SAGE tags. The level of ambiguity is a function of the genome size of the target species and is compounded further in polyploid species such as *B. napus*.

The theoretical discriminatory power of SAGE, LongSAGE and SuperSAGE methodologies were assessed using thirteen anchoring enzymes (AE) for *in silico* tag extraction and subsequent gene assignment in the amphidiploid species *B. napus*. The orientation of the reference transcripts was determined relative to *A. thaliana* which limited the analysis to 17,353 *B. napus* transcripts against which 454 data had been assembled.

The reference collection contained transcripts with a median length of 713 bp; however 462 transcripts were less than 256 bp in length and are unlikely to yield a SAGE tag using an AE with a four-base pair recognition site. Although the EST collection was enriched for 3' sequence spurious canonical tag assignments could occur on occasion due to incomplete gene coverage. The enzymes *Alu*I (AGCT), *TspE*I (AATT) and *Dpn*I (GATC) yielded the greatest number of potential tags from the *B. napus* transcript data and offered the greatest opportunity to unambiguously determine their origin irrespective of the tag length (Table 2). The enzymes *Alu*I and *Dpn*I have balanced restriction sites with every possible nucleotide represented with two purine residues followed by two pyrimidine residues, while *TspE*I (AATT) is targeted to A/T rich sequence regions which may bias the distribution of the available tags.

Although these data suggest that SAGE yields insufficient complexity to maximize the discriminatory power of a tag based expression analysis system in the complex *B. napus* genome,

the most effective AE, *Alu*I, would generate canonical tags that could unambiguously identify 51%, 57% or 59% of transcripts respectively with increasing tag length. Notably, the ambiguity among tags was not significantly reduced by changing the methodology from LongSAGE to SuperSAGE. The analysis was limited to discrimination of the 3' most canonical tags within the Brassica reference transcripts, with no provision made for anti-sense transcription which would further limit the ability of short tags to unravel the complexity of the transcriptome.

Alignment of the 17,353 Brassica transcripts to the *A. thaliana* gene sequences identified 8,095 unique loci. Multiple Brassica transcripts aligned to 3,964 individual *A. thaliana* gene identifiers (AGI) and of these 35 AGI loci were defined where alignments were made to greater than one gene model. These types of comparisons potentially allow the identification of Brassica transcripts that are either the result of genome duplication or are the subject of alternate transcript processing, and the resolution of these transcripts does appear to be enhanced through the capture of longer SAGE tags. Again, *Alu*I proved the most informative and could differentiate 57% of the possible 3,964 related transcripts.

Discussion

The ability to comprehensively analyse the expression of individual transcripts in any one species is a powerful tool for studying the phenotypic variability of gene abundance itself. The combined impact of global expression analysis with well characterised segregating populations has allowed the positioning of expression quantitative trait loci (eQTL) and promises to uncover associations between the regulation of gene expression and many elusive complex developmental traits (Kliebenstein 2009). However, such analyses are limited by the level of

discrimination offered by the available genomics tools for the species of interest. *Brassica napus* is a complex polyploid evolved from the fusion of two progenitor genomes which are themselves descended from an ancient hexaploid genome (Parkin et al. 2005). The triplication event is still prevalent in the diploid genomes although disrupted by chromosomal rearrangements and widespread gene deletion and transposition events (Town et al. 2006; Cheung et al. 2009; Trick et al. 2009b). Gene expression analysis in *B. napus* is confounded by the underlying genetic architecture and the high levels of sequence similarity observed between orthologues or homoeologues in the progenitor genomes has largely precluded their differentiation using available platforms (Trick et al. 2009a).

In the context of adopting the most cost effective and informative platform for eQTL analysis in *B. napus*, in the current study a reference transcript dataset was constructed to minimize the assembly of orthologues, and used to assess the divining properties of microarray and tag based expression analysis in the allotetraploid *B. napus*. The reference set was derived from a wide range of tissues and encompassed 46,648 *B. napus* transcripts of which 41,531 could be aligned with 12,793 *A. thaliana* genes. The majority of the *A. thaliana* genes (8,409 or 65%) shared close homology to multiple Brassica genes (on average 4.4 copies) as might be expected for a polyploid genome.

The weakest evaluation against the reference transcripts could be made for the microarray data where the nuances of probe-target interactions in hybridization experiments makes it difficult to predict the absolute limits of association between gene and 60 bp oligonucleotide sequence. However, compared to the entire probe set the differentially expressed oligo sequences aligned

with significantly more unique transcripts than expected. This raises the question of the biological relevance of this result. It could simply result from not all homologues being represented in the reference transcript dataset or it perhaps reflects the effect of transcriptional dominance in both the available EST collections and the microarray analyses. This phenomenon was observed to a limited extent in *A. thaliana* x *A. arenosa* allopolyploids (Wang et al. 2006) but was more recently suggested to be widespread in the allopolyploid species cotton (Rapp et al. 2009). Differential expression of homoeologous copies of disease responsive genes in *B. napus* in a tissue dependant and stress responsive fashion was suggested to be an adaptive response to the presence of duplicate gene copies that might confer a selective advantage (Zhao et al. 2009). The true extent and value of such expression will be underestimated until unambiguous transcript identification is routine in Brassica species.

The DH12075 reference dataset provided a platform to determine the efficacy of sequence tag based profiling in *B. napus* for both transcript determination and SNP discovery. Reference assembly of short sequence tags (~35 bp) derived from random shot-gun sequencing of cDNA has previously been employed in *B. napus* to allow extensive SNP variation to be identified (Trick et al. 2009c). However, the majority of the nucleotide variation (87.5-91.2%) was predicted to result from the co-assembly of orthologous sequences. In the present study the capture and deep sequencing of 3' fragments from a single genotype in *B. napus* generated substantial 454 sequence data. Alignment of these data against the reference transcripts from the same genotype indicated that the current assembly and SNP discovery pipeline can resolve almost 82% of the transcripts. The ability to unambiguously align the majority of the relatively long 454 reads (average 290 bp) to individual transcripts indicates that the duplicated Brassica

homologues can be differentiated; however, the expense and limited depth of 454 sequencing compared to other technologies would restrict the application of this approach. The additional benefit gained from anchoring the extracted sequence tag at the 3' end of the transcript can be observed for the *in silico* SAGE analysis, where even the shortest tags (14 bp) can allow up to 50% of the assembled transcripts to be unambiguously identified by their canonical tags. The increased complexity offered by LongSAGE can increase the resolution and robustness of the analysis even further allowing 57% of the transcripts to be differentiated by their canonical tag.

The continued reduction in costs per sequenced base as a result of next generation sequencing technologies proves to revolutionize genetic analysis in many species (Mardis 2008). In *B. napus* the use of microarrays still holds promise for the identification of unique genes or potentially genes dominantly expressed in one progenitor genome. The optimal identification of the complex mixture of transcripts generated in Brassica requires the ability to identify haplotypes across the transcript which cannot be achieved with randomly distributed short tags. However, the use of tag based technologies adopting both 3' anchoring for improved resolution and longer tag lengths which are becoming more amenable and cost effective are likely to offer the optimal balance of resolving power and transcript read depth. The release of the genome sequences for the progenitor Brassica genomes which is anticipated for 2010 will also facilitate transcript analysis by offering the ability to carry out *in silico* mapping and determining the absolute level of gene duplication present.

Acknowledgements

The research related to development and analysis of *B. napus* EST data was supported through funding from Agriculture and Agri-Food Canada Canadian Crop Genomics Initiative and Matching Investment Initiative and the National Research Council-Genomics Health Initiative IV. Analysis of gene expression analysis using microarray technology was supported by the ERA-PG project 'Associative expression and systems analysis of complex traits in oilseed rape / canola (ASSYST)'.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**(3): 403-410.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S. 2007. SNP discovery via 454 transcriptome sequencing. Plant J. **51**(5): 910-918.

Cheung, F., Trick, M., Drou, N., Lim, Y.P., Park, J.Y., Kwon, S.J., Kim, J.A., Scott, R., Pires, J.C., Paterson, A.H., Town, C., and Bancroft, I. 2009. Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. Plant Cell **21**(7): 1912-1928.

Dong, J., Keller, W.A., Yan, W., and Georges, F. 2004. Gene expression at early stages of *Brassica napus* seed development as revealed by transcript profiling of seed-abundant cDNAs. Planta **218**(3): 483-491.

Dudoit, S., Yang, Y.-H., Callow, M.J., and Speed, T.P. 2002. Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Statistica Sinica **12**: 111-139

Eveland, A.L., McCarty, D.R., and Koch, K.E. 2008. Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. Plant Physiol. **146**(1): 32-44.

Galbraith, D.W. 2006. DNA microarray analyses in higher plants. OMICS **10**(4): 455-473.

Jansen, R.C., and Nap, J.P. 2001. Genetical genomics: the added value from segregation. Trends Genet. **17**(7): 388-391.

Kliebenstein, D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. Annu. Rev. Plant Biol. **60**: 93-114.

Lee, H.S., Wang, J., Tian, L., Jiang, H., Black, M.A., Madlung, A., Watson, B., Lukens, L., Pires, J.C., Wang, J.J., Comai, L., Osborn, T.C., Doerge, R.W., and Chen, Z.J. 2004. Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species. Plant Biotechnol J **2**(1): 45-57.

Lysak, M.A., Cheung, K., Kitschke, M., and Bures, P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. **145**(2): 402-410.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. **24**(3): 133-141.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A.,

Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**(7057): 376-380.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23**(4): 452-456.

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., and Terauchi, R. 2005. SuperSAGE. Cell Microbiol. **7**(1): 11-18.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. 2009. Tablet - Next Generation Sequence Assembly Visualization. Bioinformatics: btp666.

O'Hara, P., Slabas, A.R., and Fawcett, T. 2002. Fatty acid and lipid biosynthetic genes are expressed at constant molar ratios but different absolute levels during embryogenesis. Plant Physiol **129**(1): 310-320.

O'Neill, C.M., and Bancroft, I. 2000. Comparative physical mapping of segments of the genome of Brassica oleracea var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of Arabidopsis thaliana. Plant J **23**(2): 233-243.

Obermeier, C., Hosseini, B., Friedt, W., and Snowdon, R. 2009. Gene expression profiling via LongSAGE in a non-model plant species: a case study in seeds of *Brassica napus*. BMC Genomics **10**: 295.

Onate-Sanchez, L., and Vicente-Carbajosa, J. 2008. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. BMC Res Notes **1**: 93.

Parkin, I.A., Gulden, S.M., Sharpe, A.G., Lukens, L., Trick, M., Osborn, T.C., and Lydiate, D.J. 2005. Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana.* Genetics **171**(2): 765-781.

Parkin, I.A., Sharpe, A.G., Keith, D.J., and Lydiate, D.J. 1995. Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). Genome **38**(6): 1122-1131.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19**(5): 651-652.

Pleasance, E.D., Marra, M.A., and Jones, S.J. 2003. Assessment of SAGE in transcript identification. Genome Res. **13**(6A): 1203-1215.

Rana, D., van den Boogaart, T., O'Neill, C.M., Hynes, L., Bent, E., Macpherson, L., Park, J.Y., Lim, Y.P., and Bancroft, I. 2004. Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. Plant J **40**(5): 725-733.

Rapp, R.A., Udall, J.A., and Wendel, J.F. 2009. Genomic expression dominance in allopolyploids. BMC Biol. **7**: 18.

Rennie, C., Noyes, H.A., Kemp, S.J., Hulme, H., Brass, A., and Hoyle, D.C. 2008. Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays. BMC Genomics **9**: 317.

Robinson, S.J., Cram, D.J., Lewis, C.T., and Parkin, I.A. 2004. Maximizing the efficacy of SAGE analysis identifies novel transcripts in Arabidopsis. Plant Physiol. **136**(2): 3223-3233.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. 2002. Using the transcriptome to annotate the genome. Nat Biotechnol **20**(5): 508-512.

Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., Hine, E.E., Althoff, R., Arbogast, T.S., Tallon, L.J., Vigouroux, M., Trick, M., and Bancroft, I. 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell **18**(6): 1348-1359.

Trick, M., Cheung, F., Drou, N., Fraser, F., Lobenhofer, E.K., Hurban, P., Magusin, A., Town, C.D., and Bancroft, I. 2009a. A newly-developed community microarray resource for transcriptome profiling in Brassica species enables the confirmation of Brassica-specific expressed sequences. BMC Plant Biol. **9**: 50.

Trick, M., Kwon, S.J., Choi, S.R., Fraser, F., Soumpourou, E., Drou, N., Wang, Z., Lee, S.Y., Yang, T.J., Mun, J.H., Paterson, A.H., Town, C.D., Pires, J.C., Pyo Lim, Y., Park, B.S., and Bancroft, I. 2009b. Complexity of genome evolution by segmental rearrangement in *Brassica rapa* revealed by sequence-level analysis. BMC Genomics **10**: 539.

Trick, M., Long, Y., Meng, J., and Bancroft, I. 2009c. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J **7**(4): 334-346.

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

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. Science **270**(5235): 484-487.

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., and Chen, Z.J. 2006. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172**(1): 507-517.

Yang, T.J., Kim, J.S., Kwon, S.J., Lim, K.B., Choi, B.S., Kim, J.A., Jin, M., Park, J.Y., Lim, M.H., Kim, H.I., Lim, Y.P., Kang, J.J., Hong, J.H., Kim, C.B., Bhak, J., Bancroft, I., and Park, B.S. 2006. Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. Plant Cell **18**(6): 1339-1347.

Zhang, W., Duan, S., Kistner, E.O., Bleibel, W.K., Huang, R.S., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E., Cox, N.J., and Dolan, M.E. 2008. Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet **82**(3): 631-640.

Zhao, J., Buchwaldt, L., Rimmer, S.R., Brkic, M., Bekkaoui, D., and Hegedus, D. 2009. Differential expression of duplicated peroxidase genes in the allotetraploid *Brassica napus*. Plant Physiol. Biochem. **47**(7): 653-656.

Table 1. Summary of *B. napus* (DH12075) expressed sequence tag (EST) collections used to generate the reference transcript dataset.

cDNA Library (tissue)	Mean Read Length (bp)	Total HQ[1] EST Sequences (%)	Distinct Transcripts (98% identity)	#At[2] genes	% Full Length[3]
Apical meristem	731	7,479 (80.6)	3,315	1,730	2.5
Leaf	851	6,036 (69.7)	3,091	1,514	2.1
Senescent Leaf	577	8,464 (67.3)	3,755	1,784	1.0
Root	782	13,981 (79.8)	6,463	3,054	5.2
Stem	736	3,279 (58.9)	1,647	864	1.0
Flower	625	7,439 (69.2)	3,669	1,837	1.4
Very early anther	716	10,832 (65.6)	4,934	2,460	2.5
Early anther	720	4,414 (61.3)	1,959	1,081	1.1
Embryo	730	7,400 (82.0)	2,729	1,541	1.8
Bud	802	10,090 (75.8)	5,256	2,622	3.8
Late bud	766	3,015 (64.1)	1,794	1,108	1.3
Cotyledon	718	3,860 (71.8)	1,691	1,081	1.3
Cold stress – dark	791	6,393 (70.1)	3,867	2,000	2.9
Cold stress – light	738	8,310 (83.2)	4,416	2,189	3.5
Osmotic stress – leaf	737	13,067 (72.4)	5,420	2,676	2.1
Osmotic stress- root	615	6,163 (64.2)	3,767	2,054	2.3
Damaged cotyledon	731	4,568 (68.3)	2,457	1,378	1.7
Etiolated seedlings	724	19,514 (77.5)	9,487	4,271	3.6
Total	**726**	**144,352 (72.5)**	**46,648**	**12,793**	**20.0**

1. HQ – high quality EST sequences filtered for quality, length and vector contamination.

2. At – *Arabidopsis thaliana* gene codes identified by BLASTN analysis.

3. Estimated number of cDNA clones that were contained full length transcripts based on identification of a conserved start codon between *A. thaliana* and *B. napus* within the clone.

Table 2. Summary of *in silico* SAGE analysis in the allotetraploid *B. napus*. The most informative anchoring enzyme is indicated in bold font.

Anchoring enzyme (AE)	No. of transcripts with canonical AE site (%)[1]	No. of unique canonical tag positions (%)[2]		
		SAGE (14 bp)	LongSAGE (21 bp)	SuperSAGE (26 bp)
*Acc*II	9112 (52.5)	4913 (28.3)	5518 (31.8)	5677 (32.7)
*Aci*I	11439 (65.9)	5967 (34.4)	6609 (38.1)	6886 (39.7)
***Alu*I**	**16534 (95.3)**	**8890 (51.2)**	**9883 (57)**	**10317 (59.5)**
*Civ*RI	15174 (87.4)	8267 (47.6)	9187 (52.9)	9597 (55.3)
*Dpn*I	15889 (91.6)	8490 (48.9)	9477 (54.6)	9803 (56.5)
*Hae*III	11417 (65.8)	5748 (33.1)	6378 (36.8)	6653 (38.3)
*Hha*I	8369 (48.2)	4623 (26.6)	5117 (29.5)	5326 (30.7)
*Hpa*II	12203 (70.3)	6400 (36.9)	7060 (40.7)	7313 (42.1)
*Mae*I	12538 (72.2)	6896 (39.7)	7601 (43.8)	7787 (44.9)
*Nla*III	15286 (88.1)	8357 (48.2)	9282 (53.5)	9652 (55.6)
*Rsa*I	13181 (76)	6863 (39.5)	7613 (43.9)	7945 (45.8)
*Taq*I	14687 (84.6)	8006 (46.1)	8865 (51.1)	9143 (52.7)
*TspE*I	14609 (84.2)	887 (51.2)	9654 (55.6)	9904 (57.1)

1. Number of reference transcripts from the 17,353 orientated sequences that possess an anchoring enzyme restriction site.

2. Number of canonical short sequence tags that unambiguously identify their transcript of origin.

Figure Captions.

Figure 1. Drawing depicting the duplication events that have shaped the *B. napus* genome and their impact on gene copy number. The alignment of eleven isolated and sequenced regions of *B. napus* to their collinear segments in the genome of *A. thaliana* is shown; the data presented is taken from Cheung et al (2009). Vertical boxes represent Brassica or Arabidopsis genomic sequence, filled triangles indicate the presence of a gene copy, horizontal lines indicate the homologous relationships between gene copies.

Figure 2. Schematic of the TGICL pipeline that was used to generate the DH12075 reference transcript dataset. The step indicated by the asterisk was optimized to facilitate maximum separation of the multiple homologous transcripts found in *B. napus*.

Figure 3. Sequence alignment of expressed sequence tags from the two progenitor genomes of *B. napus* indicated that a significant proportion of homoeologous transcripts share 98% nucleotide similarity.

Figure 4. Significantly differential expressed genes identified between developing seeds of *B. napus* genotypes DH12075 and PSA12. The axis represent the log scaled normalized intensities for each genotype for a particular probe sequence.

Figure 5. Sequence assembly of two homologous DH12075 *B. napus* transcripts aligned to a single reference transcript. The alignment has been visualized using the software tool Tablet

(Milne et al. 2009) which shows a schematic at the top of the entire assembly and a close up of the nucleotide detail below, both views display nucleotide variants among the sequences.

Figure 1.

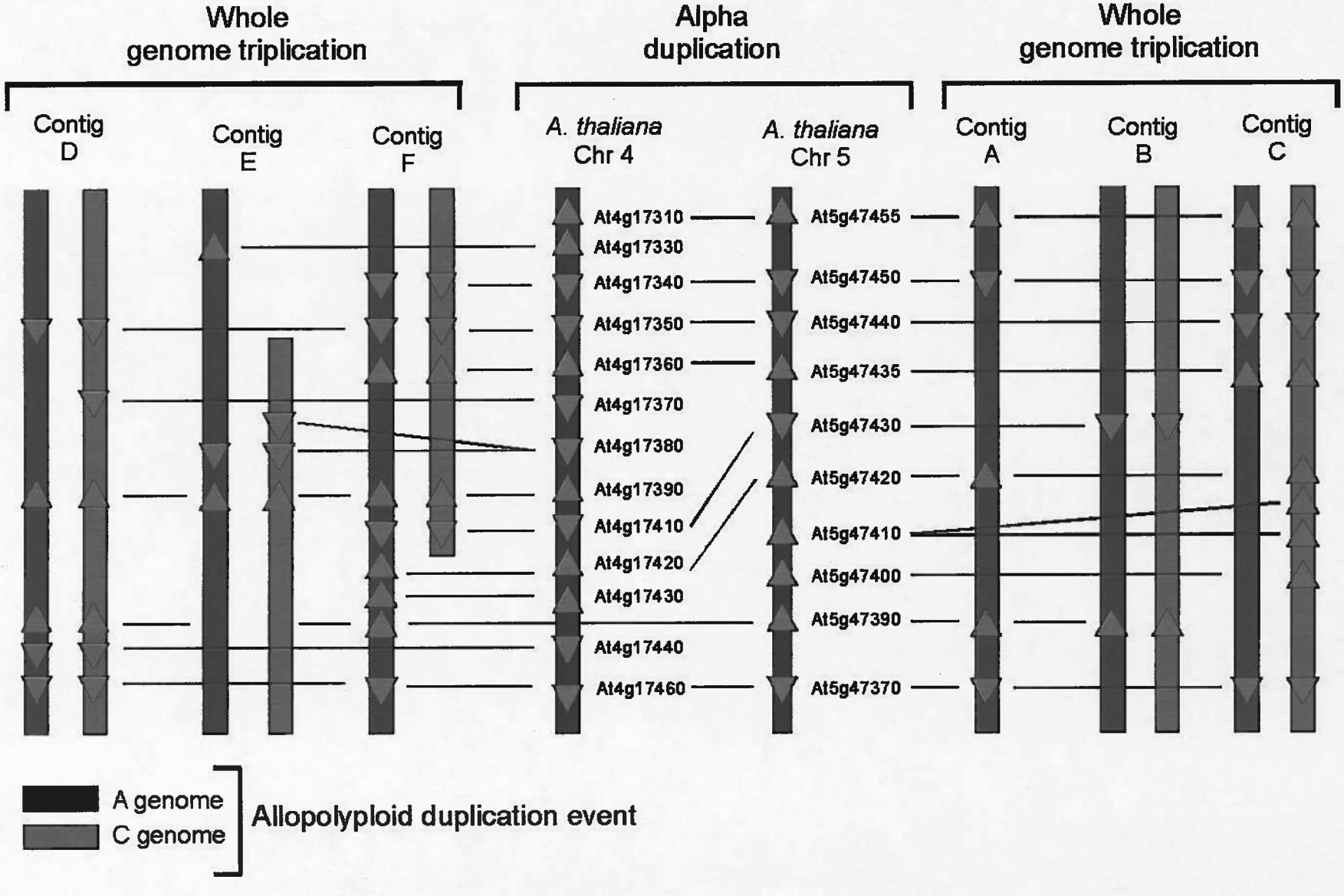

Whole genome triplication
Alpha duplication
Whole genome triplication
Contig D
Contig E
Contig F
A. thaliana Chr 4
A. thaliana Chr 5
Contig A
Contig B
Contig C
At4g17310
At4g17330
At4g17340
At4g17350
At4g17360
At4g17370
At4g17380
At4g17390
At4g17410
At4g17420
At4g17430
At4g17440
At4g17460
At5g47455
At5g47450
At5g47440
At5g47435
At5g47430
At5g47420
At5g47410
At5g47400
At5g47390
At5g47370
A genome
C genome
Allopolyploid duplication event

Figure 2.

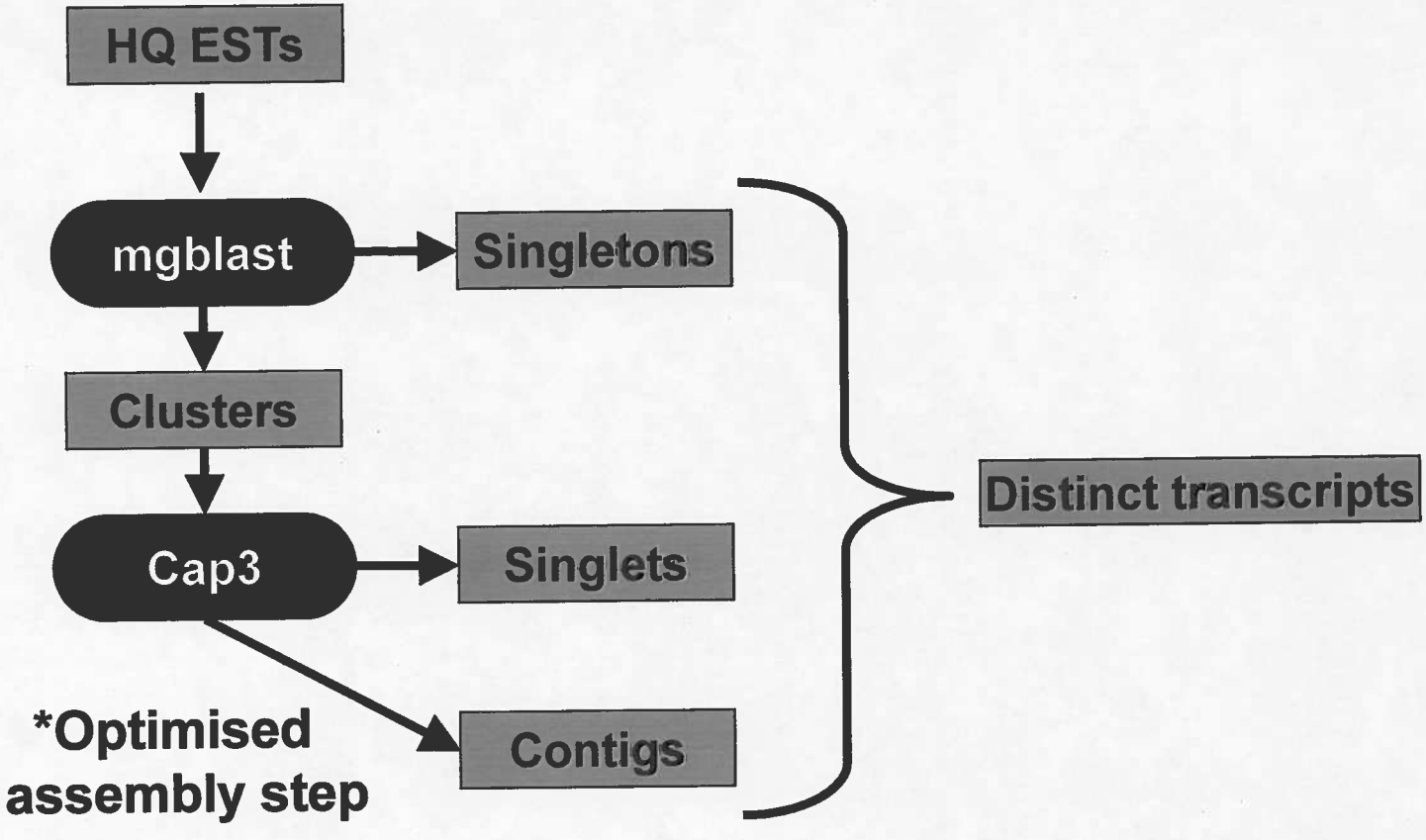

HQ ESTs
mgblast
Singletons
Clusters
Cap3
Singlets
Contigs
*Optimised assembly step
Distinct transcripts

Figure 3.

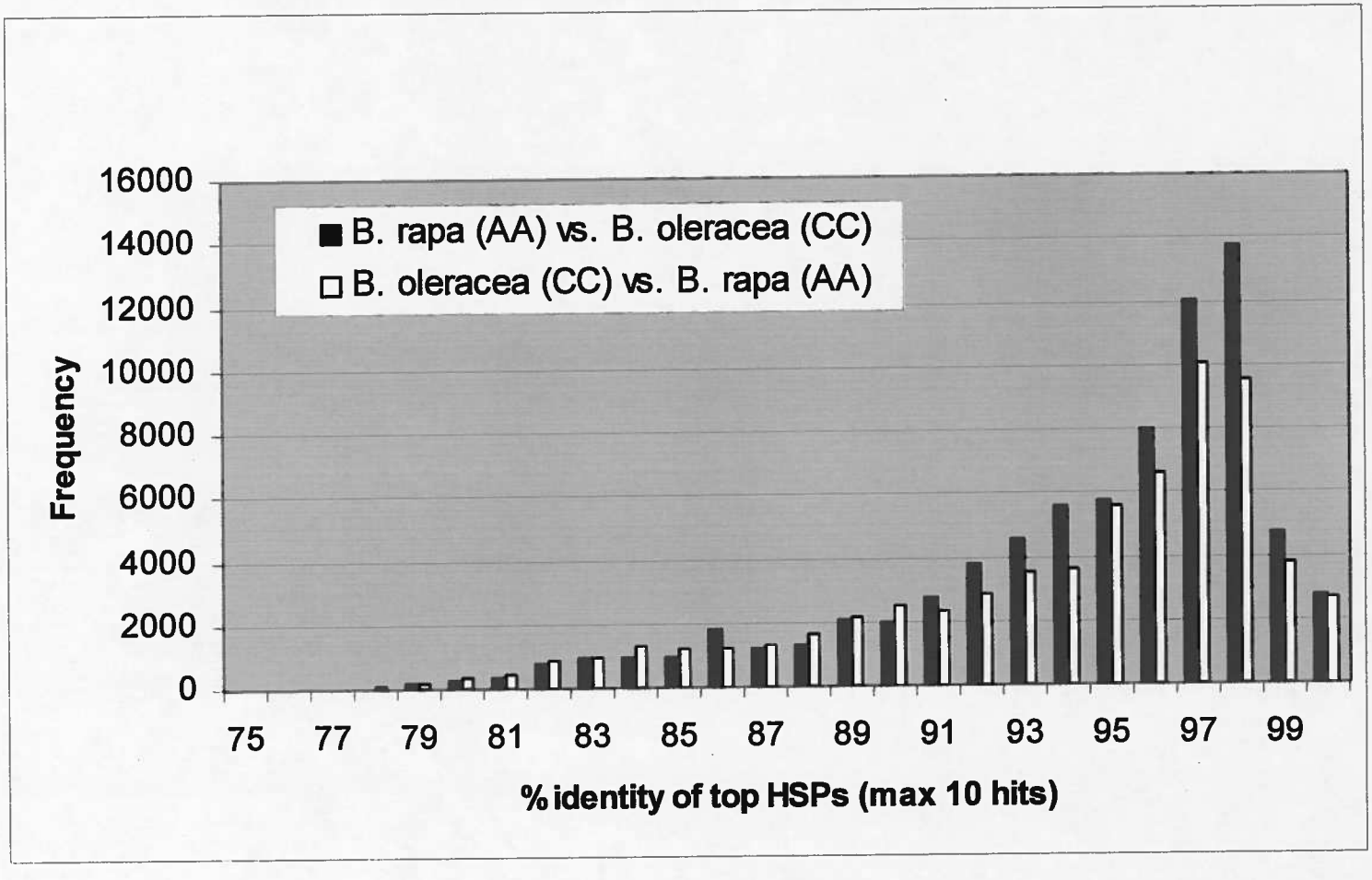

16000
14000
12000
10000
8000
6000
4000
2000
0
Frequency
B. rapa (AA) vs. B. oleracea (CC)
B. oleracea (CC) vs. B. rapa (AA)
75
77
79
81
83
85
87
89
91
93
95
97
99
% identity of top HSPs (max 10 hits)

Figure 4.

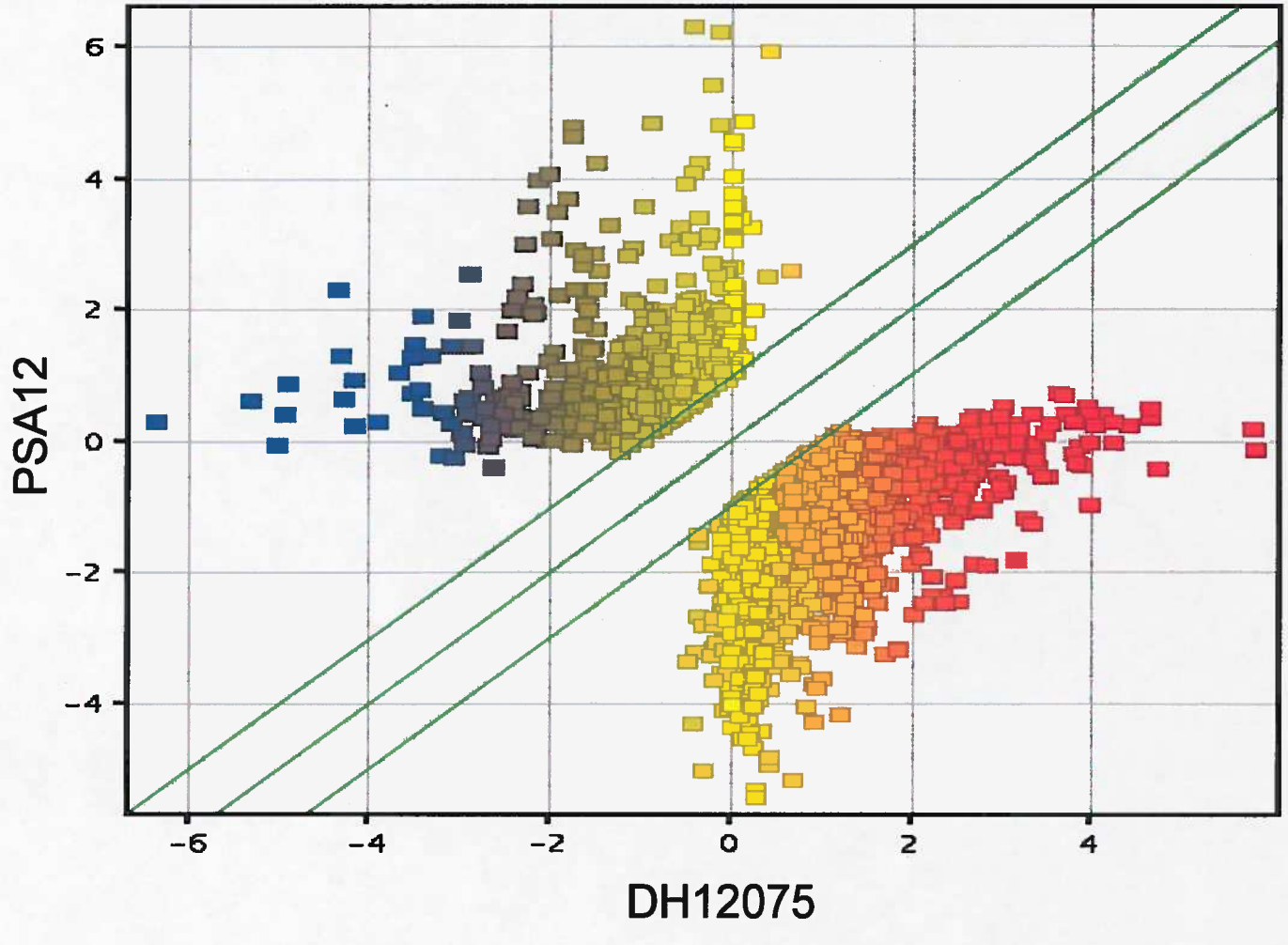
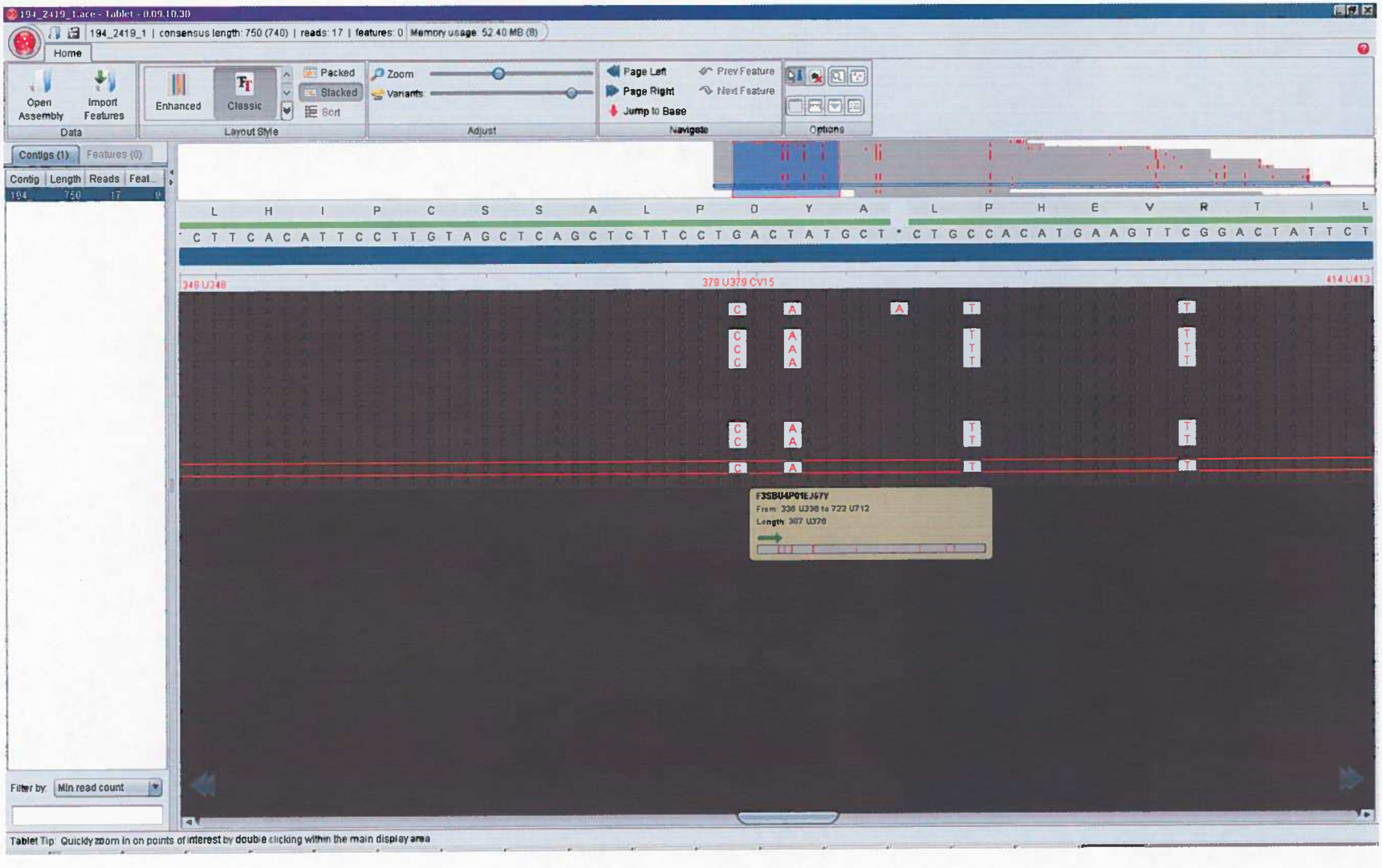


Figure 5.


194_2419_1.ace - Tablet - 0.09.10.30
194_2419_1 | consensus length: 750 (740) | reads: 17 | features: 0 | Memory usage: 52.40 MB (8)
Home
Open Assembly
Import Features
Data
Enhanced
Classic
Packed
Stacked
Sort
Layout Style
Zoom
Variants
Adjust
Page Left
Page Right
Jump to Base
Prev Feature
Next Feature
Navigate
Options
Contigs (1)
Features (0)
Contig
Length
Reads
Feat...
L H I P C S S A L P D Y A L P H E V R T I L
C T T C A C A T T C C T T G T A G C T C A G C T C T T C C T G A C T A T G C T C T G C C A C A T G A A G T T C G G A C T A T T C T
348 U348
379 U379 CV15
414 U413
F3SBU4P01EJ67Y
From: 336 U336 to 722 U712
Length: 387 U376
Filter by:
Min read count
Tablet Tip: Quickly zoom in on points of interest by double clicking within the main display area