Towards unambiguous transcript mapping in the allotetraploid *Brassica napus*[1]

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

Abstract: The architecture of the *Brassica napus* genome is marked by its evolutionary origins. The genome of *B. napus* was formed from the hybridization 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 and 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 *B. napus* transcriptome but was limited by tag length. The ability to comprehensively catalogue gene expression in polyploid species could be transformed by the application of cost-efficient next generation sequencing technologies that will capture millions of long sequence tags.

Key words: polyploid, homologues, next generation sequencing, expressed sequence tags, single nucleotide polymorphism, *Brassica*.

Résumé : La structure du génome chez le *Brassica napus* est marquée par ses origines évolutives. Le génome du *B. napus* a été formé suite à l'hybridation de deux espèces diploïdes de *Brassica* très apparentées, toutes deux dérivées d'ancêtres hexaploïdes. Ces importants événements de duplication survenus à l'échelle du génome entier, tant dans un passé récent que plus ancien, font en sorte que le génome allotétraploïde du *B. napus* a conservé de multiples copies de la plupart des gènes. Cette situation permet de penser que le transcriptome de cette espèce est à la fois complexe et redondant, ce qui peut entraîner de la confusion dans les analyses d'expression. Un assemblage rigoureux de 142 399 étiquettes de séquences exprimées chez le *B. napus* a permis de développer un jeu de transcrits de référence bien différenciés, lesquels ont permis de mesurer l'efficacité des outils disponibles pour l'identification et la discrimination entre transcrits chez le *B. napus*. Les outils examinés étaient l'hybridation sur puces à ADN et la capture d'étiquettes de séquences ancrées à l'extrémité 3′. Les puces à ADN n'arrivent pas à distinguer les transcrits provenant des deux espèces progénitrices ou d'homologues proches, bien que les cas d'expression différentielle observés semblaient biaisés en faveur de transcrits uniques. L'emploi d'étiquettes capturées en 3′ a permis d'améliorer la capacité à identifier sans ambiguïté les homologues au sein du transcriptome du B. napus, mais était limité par la taille des étiquettes. Les nouvelles technologies de séquençage à faible coût permettant de générer des millions d'étiquettes plus longues pourraient révolutionner notre capacité à cataloguer de manière exhaustive l'expression génique chez des espèces polyploïdes.

Mots-clés : polyploïdie, homologues, séquençage de seconde génération, étiquettes de séquences exprimées, variation SNP, *Brassica*.

[Traduit par la Rédaction]

Received 6 January 2010. Accepted 25 May 2010. Published on the NRC Research Press Web site at genome.nrc.ca on 5 November 2010.

Corresponding Editor: M. Francki.

I.A.P. Parkin,[2] W. Zhang, S.J. Robinson, S. Karcz, and E.E. Higgins. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada.
W.E. Clarke. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada; Department of Computing Science, 176 Thorvaldson Building, University of Saskatchewan, 110 Science Place, Saskatoon, SK S7N 5C9, Canada.
C. Sidebottom, P. Fobert, and A.G. Sharpe. National Research Council Plant Biotechnology Institute, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada.
M.G. Links. Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada; Department of Veterinary Microbiology, WCVM, University of Saskatchewan, 52 Campus Drive, Saskatoon, SK S7N 5B4, Canada.

[1]This article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming".
[2]Corresponding author (e-mail: isobel.parkin@agr.gc.ca).

Introduction

The study of global gene expression analysis is a powerful tool for understanding complex biological phenomena, 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 complex phenotypic variation (Jansen and Nap 2001). There is now an opportunity to apply these types of analyses to nonstructured collections of diverse germplasm, combining the strength of association mapping with the sensitivity of transcript analysis to elucidate complex quantitative phenotypes. Such strategies are 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 are now sufficient published sequence data to allow the development of gene-specific microarray platforms that permit the presence and abundance of thousands of genes to be queried at once (Galbraith 2006). The range of suitable applications for 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. 2009*a*), which is expected to represent roughly 71% of the gene content based on comparison with the annotated *Arabidopsis thaliana* gene sequences (TAIR version 8: 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 Brassiceae. *Brassica napus* is a relatively young allotetraploid formed from the union 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, indicate that the genus *Brassica* 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 *A. thaliana* confirmed 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 (Fig. 1).

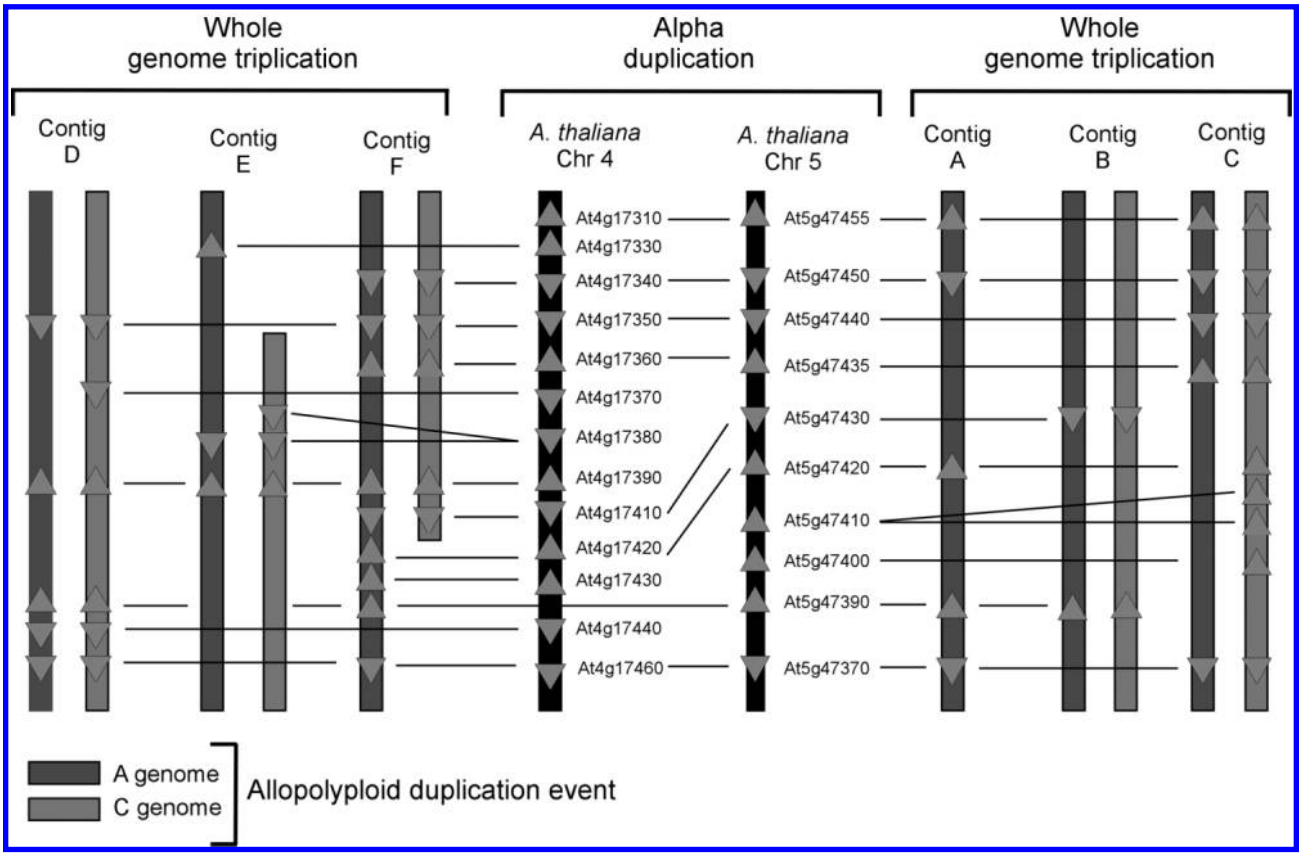
Notwithstanding possible alternative transcript processing events, the multiple related gene copies confuse accurate transcript differentiation, which can confound any gene expression study and limit the ability to elucidate the genetic mechanisms controlling a trait. The available *Brassica* microarray platforms cannot resolve transcripts derived from the homoeologous gene copies within the A and C genomes of *B. napus* (Trick et al. 2009*a*), and the limits of tag-based expression profiling in *Brassica* species have 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 G. Rakow and G. Séguin-Swartz, Agriculture and Agri-Food Canada), and PSA12, a resynthesized *B. napus* line (generated by M. Beschorner and D. Lydiate, Agriculture and Agri-Food Canada). Tissue for RNA extraction was collected from: developing seeds 21 d postanthesis grown under field conditions; etiolated seedlings (5–7 d postgermination 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 Oñate-Sánchez and Vicente-Carbajosa (2008). In summary, seed embryo tissue (fresh mass approximately 20 mg) was ground with liquid nitrogen and mixed with 700 μL extraction buffer (0.4 mmol/L LiCl, 0.2 mol/L Tris–Cl (pH 8.0), 25 mmol/L 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 mol/L LiCl for 1 h at –20 °C. After

Fig. 1. The duplication events that have shaped the *Brassica napus* genome and their impact on gene copy number. The alignment of 11 isolated and sequenced regions of *B. napus* to their collinear segments in the genome of *Arabidopsis thaliana* is shown; the data presented are taken from Cheung et al. (2009). Vertical boxes represent the *Brassica* or the *Arabidopsis* genomic sequence, filled triangles indicate the presence of a gene copy, horizontal lines indicate the homologous relationships between gene copies.



centrifuging at 13 000 r/min for 30 min, the pellet was washed with 1 mL of 75% ethanol and dissolved in 40 μL of diethylpyrocarbanate–water. DNA was removed from total RNA using DNase I (Invitrogen Canada Inc., Burlington, Ontario) according to the manufacturer's instructions. RNA was extracted from etiolated seedlings and juvenile leaves using the Qiagen RNeasy mini kit according to the manufacturer's protocol (QIAGEN Inc., Mississauga, Ontario).

Microarray hybridization and data analysis

Gene expression analysis for DH12075 and PSA12 was carried out using Agilent *Brassica* gene expression microarrays (Agilent Technologies, Mississauga, Ontario). 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 Technologies) according to the manufacturer's instructions. Amplified and labeled cRNA was purified with Qiagen RNeasy mini kit (QIAGEN) and quantified with the NanoDrop ND-1000. Labeled cRNA (2 μg) was fragmented and subsequently hybridized in dual labeled reactions to the Agilent 4 × 44 K *Brassica* array using the gene expression hybridization kit (Agilent Technologies) according to the manufacturer's protocol. Hybridization was performed in an Agilent microarray hybridization chamber (Agilent Technologies) for 17 h at 65 °C with a rotation of 10 r/min. Slides were washed with the Agilent gene expression washing buffers I and II (Agilent Technologies) according to the manufacturer's protocol.

Arrays were scanned at 5 μm resolution with the GenePix 4000B scanner, the fluorescence data were 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 were exported from BASE and imported to GeneSpring GX 10.0 (Agilent Technologies) for further analysis. The parental lines were analyzed 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 twofold cutoff.

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 and (or) adapters modified for the Titanium chemistry were

implemented in the protocol: A-adapter-top 5′-CCATCTCATCCCTGCGTGTCTCCGACTCAGCAT-3′; A-adapter-bottom 5′-CGATGCTGAGTCGGAGACACGCAGGGATGA-3′; B-adaptor 5′biotin-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGTTTTTTTTTTTTVN-3′. Roche 454 Titanium sequencing was carried out at the National Research Council Plant Biotechnology Institute, Saskatoon, Saskatchewan 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 were assembled against a *Brassica* reference of 46 648 transcripts (described below in Results) using NGen (DNAStar Inc, Madison, Wisconsin) 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 (*i*) to parse the individual ACE files from the resultant assembly; (*ii*) to replace the consensus sequence with the reference to allow comparisons between analyses; and (*iii*) 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 (available at sourceforge.net/projects/aped). SNPs were identified in each assembly using the PolyBayes 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 PolyBayes output to identify variation resulting from homoeologous and paralogous loci.

Serial analysis of gene expression

In silico SAGE analysis was performed on the *B. 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 13 anchoring enzymes (*Acc*II, *Aci*I, *Alu*I, *Civ*RI, *Dpn*I, *Hae*III, *Hha*I, *Hpa*II, *Mae*I, *Nla*III, *Rsa*I, *Taq*I, and *Tsp*EI) and three SAGE methodologies, SAGE, LongSAGE, and SuperSAGE, that result in tags of the following respective lengths: 14, 21, and 26 bp (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 data set

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 software (Pertea et al. 2003) with optimized parameters to generate a *B. napus* distinct set of reference transcripts. Parameters were established for assembling the ESTs 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 estimated that the maximum level of sequence similarity between orthologues was 98% (Fig. 2). Assembly of the sequence data from *B. napus* and its progenitors was completed in a series of progressively more stringent assemblies using default Cap3 parameters except for varying the overlap parameter, within the range 82%–99%. Comparing the resulting assemblies to a set of single copy *Arabidopsis* gene sequences (excluded all *Arabidopsis* genes with a BLAST match to another gene at $E \leq 1 \times 10^{-5}$) confirmed that the maximal resolution of paralogues occurred with a Cap3 overlap parameter of 98%, since at this cutoff the assemblies for the progenitor genomes were being resolved for the single copy genes. 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 aafc-aac.usask.ca/cgi-bin/gbrowse/gbrowse/BAGI2/.

The *Brassica* oligosequences 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. As the *Brassica* oligonucleotides 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 lower signal intensity depending on the position of the nucleotide variation within the oligonucleotide (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 because of the length of the probes and the complexity of the genome, 3530 oligonucleotides (22%) matched multiple reference transcripts, with 23 probes matching 10 or more unique transcripts.

Array-based expression analysis in *Brassica napus*

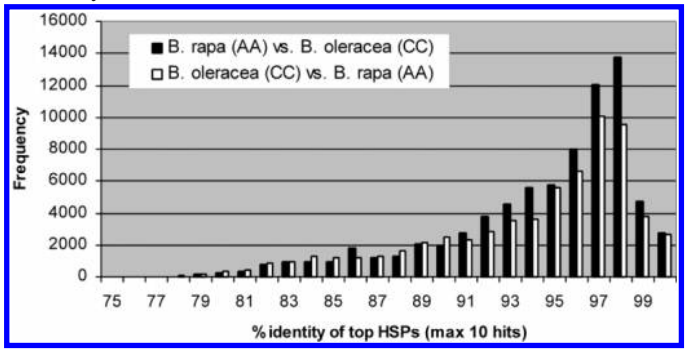
The Agilent *Brassica* 4 × 44 K array was selected for microarray experiments in *B. napus* because of the increased efficiency offered by the array design. These arrays contain 43 809 gene-specific 60-mer oligonucleotides arranged in four replicate arrays. The oligonucleotides were designed from 40 206 *Brassica* unigene sequences that were derived from a mixture of *B. napus* genotypes (available at compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=oilseed_rape). BLASTN analysis (Altschul et al. 1990) with an expect value threshold of 1×10^{-6} identified the closest homologous gene identifier in *A. thaliana* for

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

cDNA library (tissue)	Mean read length (bp)	Total HQ[a] EST sequences (%)	Distinct transcripts (98% identity)	No. of At[b] genes	% full length[c]
Apical meristem	731	7 479 (80.6)	3315	1730	2.5
Leaf	851	6 036 (69.7)	3091	1514	2.1
Senescent leaf	577	8 464 (67.3)	3755	1784	1.0
Root	782	13 981 (79.8)	6463	3054	5.2
Stem	736	3 279 (58.9)	1647	864	1.0
Flower	625	7 439 (69.2)	3669	1837	1.4
Very early anther	716	10 832 (65.6)	4934	2460	2.5
Early anther	720	4 414 (61.3)	1959	1081	1.1
Embryo	730	7 400 (82.0)	2729	1541	1.8
Bud	802	10 090 (75.8)	5256	2622	3.8
Late bud	766	3 015 (64.1)	1794	1108	1.3
Cotyledon	718	3 860 (71.8)	1691	1081	1.3
Cold stress, dark	791	6 393 (70.1)	3867	2000	2.9
Cold stress,– light	738	8 310 (83.2)	4416	2189	3.5
Osmotic stress, leaf	737	13 067 (72.4)	5420	2676	2.1
Osmotic stress, root	615	6 163 (64.2)	3767	2054	2.3
Damaged cotyledon	731	4 568 (68.3)	2457	1378	1.7
Etiolated seedlings	724	19 514 (77.5)	9487	4271	3.6
Total	**726**	**144 352 (72.5)**	**46 648**	**12 793**	**20.0**

[a]HQ, high quality EST sequences filtered for quality, length, and vector contamination.
[b]At, *Arabidopsis thaliana* gene codes identified by BLASTN analysis.
[c]Estimated number of cDNA clones that contained full length transcripts based on identification of a conserved start codon between *A. thaliana* and *B. napus* within the clone.

Fig. 2. Sequence alignment of expressed sequence tags from the two progenitor genomes of *Brassica napus* indicated that a significant proportion of homoeologous transcripts share 98% nucleotide similarity.



each of the unigene sequences. In total, 35 244 unigenes shared sequence similarity with 90 *A. thaliana* plastid gene codes and 15 596 nuclear genes. The remaining 4962 unigenes demonstrated no significant sequence homology with an annotated *A. thaliana* gene code.

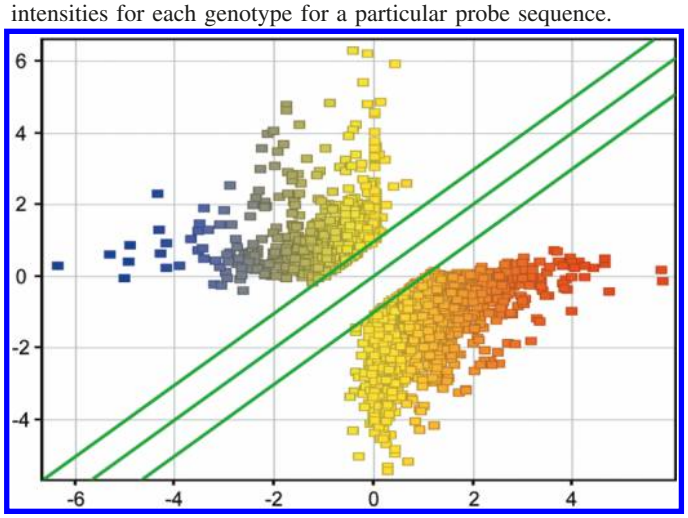
The microarray data analyzed were from experiments conducted as part of a project to carry out expression QTL analysis in *B. napus*, to associate gene expression changes with genetic polymorphisms in a doubled haploid segregating population, and to relate these to variation in seed quality traits. Previous studies have established that ~20 d 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, 2331 probe sequences identified genes that were significantly differentially expressed (see methods: $P < 0.05$, FDR < 0.05), with 1614 upregulated in DH12075 and 717 upregulated in PSA12 (Fig. 3). Based on the array design, these 2331 probe sequences recognized 2235 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 oligonucleotides possessed additional probes, which showed no change in signal intensity between the parents. Such anomalies between the observed hybridization responses of specific probes could explain discrepancies often experienced between expression platforms.

The corresponding transcripts from the reference DH12075 collection could be determined for 760 of the differentially hybridized 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 ($\chi^2 = 14.31$; $P < 0.0002$). 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 reference data set for SNP discovery and mapping in *B. napus*. The method generates long sequence tags (200–400 bp)

Fig. 3. Significantly differential expressed genes identified between developing seeds of *Brassica napus* genotypes DH12075 (*x*-axis) and PSA12 (*y*-axis). The axis represents the $\log_2$ scaled normalized intensities for each genotype for a particular probe sequence.



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 454 sequencing plate.

The *Brassica* 454 EST data were assembled against the DH12075 reference transcript data set using NGen, and custom Perl scripts were developed to parse the resultant ace files (see Materials and Methods). 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 analyzed using PolyBayes (Marth et al. 1999) to identify nucleotide variations that differentiate homoeologous and paralogous transcripts that had been co-assembled (Fig. 4). 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 3259 (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 data set 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 data set. The presence of antisense transcripts was also observed, with 362 of the 3680 most highly abundant transcripts possessing at least 20 reads matched to both strands of the transcript.

Potential for serial analysis of gene expression 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, because of the isolation of short tag sequences and the nonrandom 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*.

There are currently three developed SAGE protocols that utilize different restriction enzymes to capture sequence tags of varying lengths: 14, 21, and 26 bp for SAGE, LongSAGE, and SuperSAGE, respectively. The theoretical discriminatory power of the three SAGE methodologies was assessed using 13 anchoring enzymes 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 anchoring enzyme with a 4 bp recognition site. Although the EST collection was enriched for 3′ sequence, spurious canonical tag assignments could occur on occasion because of incomplete gene coverage. The enzymes *Alu*I (AGCT), *Tsp*EI (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 *Tsp*EI (AATT) is targeted to A- and (or) T-rich sequence regions, which may bias the distribution of the available tags.

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. However, the most effective anchoring enzyme, *Alu*I, would generate canonical tags that could unambiguously identify 51%, 57%, or 59% of all transcripts sampled, respective to an increasing tag length (Table 2). 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 8095 unique loci. Multiple *Brassica* transcripts aligned to 3964 individual *A. thaliana* gene identifiers (AGI), and, of these, 35 AGI loci were defined where alignments were made to greater than one

Fig. 4. Sequence assembly of two homologous DH12075 *Brassica napus* transcripts aligned to a single reference transcript. The alignment has been visualized using the software tool Tablet (Milne et al. 2010) that 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.

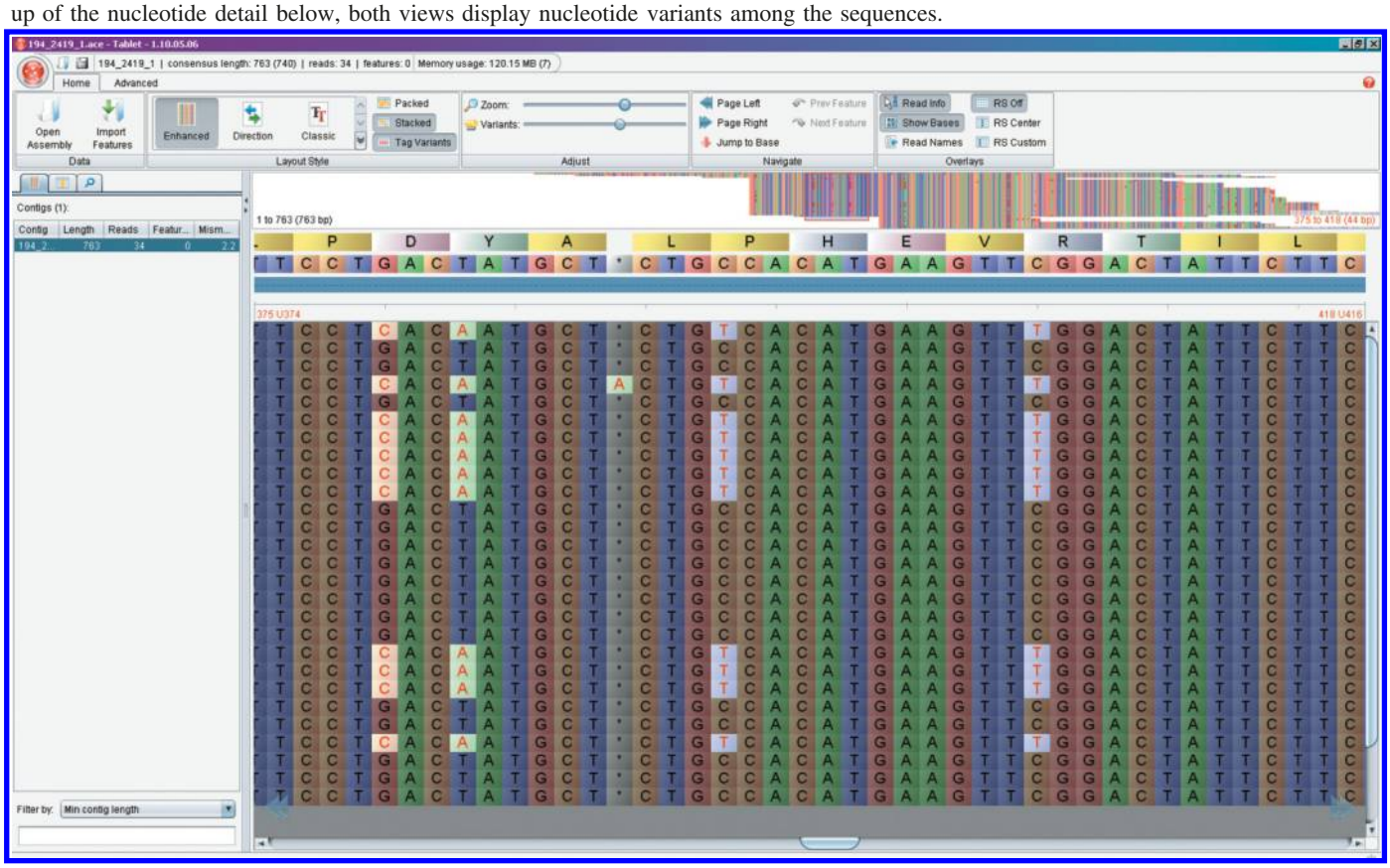


Table 2. Summary of in silico serial analysis of gene expression (SAGE) analysis in the allotetraploid *Brassica napus*.

Anchoring enzyme	No. of transcripts with canonical anchoring enzyme site (%)[b]	No. of unique canonical tag positions (%)[a]		
		SAGE (14 bp)	LongSAGE (21 bp)	SuperSAGE (26 bp)
*Acc*II	9 112 (52.5)	4 913 (28.3)	5 518 (31.8)	5 677 (32.7)
*Aci*I	11 439 (65.9)	5 967 (34.4)	6 609 (38.1)	6 886 (39.7)
*Alu*I[c]	16 534 (95.3)	8 890 (51.2)	9 883 (57.0)	10 317 (59.5)
*Civ*RI	15 174 (87.4)	8 267 (47.6)	9 187 (52.9)	9 597 (55.3)
*Dpn*I	15889 (91.6)	8 490 (48.9)	9 477 (54.6)	9 803 (56.5)
*Hae*III	11 417 (65.8)	5 748 (33.1)	6 378 (36.8)	6 653 (38.3)
*Hha*I	8 369 (48.2)	4 623 (26.6)	5 117 (29.5)	5 326 (30.7)
*Hpa*II	12 203 (70.3)	6 400 (36.9)	7 060 (40.7)	7 313 (42.1)
*Mae*I	12 538 (72.2)	6 896 (39.7)	7 601 (43.8)	7 787 (44.9)
*Nla*III	15 286 (88.1)	8 357 (48.2)	9 282 (53.5)	9 652 (55.6)
*Rsa*I	13 181 (76.0)	6 863 (39.5)	7 613 (43.9)	7 945 (45.8)
*Taq*I	14 687 (84.6)	8 006 (46.1)	8 865 (51.1)	9 143 (52.7)
*Tsp*EI	14 609 (84.2)	887 (51.2)	9 654 (55.6)	9 904 (57.1)

[a]Number of canonical short sequence tags that unambiguously identify their transcript of origin.
[b]Number of reference transcripts from the 17 353 oriented sequences that possess an anchoring enzyme restriction site.
[c]The most informative anchoring enzyme.

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 *Brassica* homologues aligned to the 3964 AGI loci.

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 characterized segregating populations has allowed the positioning of 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 hybridization 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. 2009*b*). 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 have largely precluded their differentiation using available platforms (Trick et al. 2009*a*).

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 data set 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 (8409 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 make it difficult to predict the absolute limits of association between gene and 60 bp oligonucleotide sequence. However, compared with the entire probe set, the differentially expressed oligosequences 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 data set, 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* × *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-dependent 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 data set 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. 2009*c*). 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 with 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. PMID:2231712.

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. doi:10.1111/j.1365-313X.2007.03193.x. PMID:17662031.

Cheung, F., Trick, M., Drou, N., Lim, Y.P., Park, J.Y., Kwon, S.J., et al. 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. doi:10.1105/tpc.108.060376. PMID:19602626.

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. doi:10.1007/s00425-003-1124-2. PMID:14574574.

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. Statist. 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. doi:10.1104/pp.107.108597. PMID:18024554.

Galbraith, D.W. 2006. DNA microarray analyses in higher plants. OMICS, **10**(4): 455–473. doi:10.1089/omi.2006.10.455. PMID: 17233557.

Jansen, R.C., and Nap, J.P. 2001. Genetical genomics: the added value from segregation. Trends Genet. **17**(7): 388–391. doi:10.1016/S0168-9525(01)02310-1. PMID:11418218.

Kliebenstein, D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. Annu. Rev. Plant Biol. **60**(1): 93–114. doi:10.1146/annurev.arplant.043008.092114. PMID:19012536.

Lee, H.S., Wang, J., Tian, L., Jiang, H., Black, M.A., Madlung, A., et al. 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. doi:10.1046/j.1467-7652.2003.00048.x. PMID:17166142.

Lysak, M.A., Cheung, K., Kitschke, M., and Bureš, P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. **145**(2): 402–410. doi:10.1104/pp.107.104380. PMID: 17720758.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. **24**(3): 133–141. PMID: 18262675.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature, **437**(7057): 376–380. PMID:16056220.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., et al. 1999. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23**(4): 452–456. doi:10.1038/70570. PMID:10581034.

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., et al. 2005. SuperSAGE. Cell. Microbiol. **7**(1): 11–18. doi:10.1111/j.1462-5822.2004.00478.x. PMID:15617519.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. 2010. Tablet—next generation sequence assembly visualization. Bioinformatics, **26**(3): 401–402. [btp666.] doi:10.1093/bioinformatics/btp666. PMID:19965881.

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**(1): 295. doi:10.1186/1471-2164-10-295. PMID: 19575793.

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. doi:10.1104/pp.010956. PMID:12011361.

Oñate-Sánchez, L., and Vicente-Carbajosa, J. 2008. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. BMC Res. Notes, **1**(1): 93. doi:10.1186/1756-0500-1-93. PMID:18937828.

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. doi:10.1046/j.1365-313x.2000.00781.x. PMID:10929117.

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. PMID: 18470236.

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. doi:10.1534/genetics.105.042093. PMID:16020789.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., et al. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics, **19**(5): 651–652. doi:10.1093/bioinformatics/btg034. PMID:12651724.

Pleasance, E.D., Marra, M.A., and Jones, S.J. 2003. Assessment of SAGE in transcript identification. Genome Res. **13**(6a): 1203–1215. doi:10.1101/gr.873003. PMID:12743019.

Rana, D., van den Boogaart, T., O'Neill, C.M., Hynes, L., Bent, E., Macpherson, L., et al. 2004. Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. Plant J. **40**(5): 725–733. doi:10.1111/j.1365-313X.2004.02244.x. PMID:15546355.

Rapp, R.A., Udall, J.A., and Wendel, J.F. 2009. Genomic expression dominance in allopolyploids. BMC Biol. **7**: 18. PMID: 19409075.

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**(1): 317. doi:10.1186/1471-2164-9-317. PMID:18598341.

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. doi:10.1104/pp.104.043406. PMID:15489285.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., et al. 2002. Using the transcriptome to annotate the genome. Nat. Biotechnol. **20**(5): 508–512. doi:10.1038/nbt0502-508. PMID:11981567.

Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., et al. 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell, **18**(6): 1348–1359. doi:10.1105/tpc.106.041665. PMID:16632643.

Trick, M., Cheung, F., Drou, N., Fraser, F., Lobenhofer, E.K., Hurban, P., et al. 2009*a*. A newly-developed community microarray resource for transcriptome profiling in *Brassica* species enables the confirmation of *Brassica*-specific expressed sequences. BMC Plant Biol. **9**(1): 50. doi:10.1186/1471-2229-9-50. PMID: 19426481.

Trick, M., Kwon, S.J., Choi, S.R., Fraser, F., Soumpourou, E., Drou, N., et al. 2009*b*. Complexity of genome evolution by segmental rearrangement in *Brassica rapa* revealed by sequence-le-

vel analysis. BMC Genomics, **10**(1): 539. doi:10.1186/1471-2164-10-539. PMID:19922648.

Trick, M., Long, Y., Meng, J., and Bancroft, I. 2009*c*. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol. J. **7**(4): 334–346. doi:10.1111/j.1467-7652.2008.00396.x. PMID:19207216.

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

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. Science, **270**(5235): 484–487. doi:10.1126/science.270.5235.484. PMID:7570003.

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., et al. 2006. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics, **172**(1): 507–517. doi:10.1534/genetics.105.047894. PMID:16172500.

Yang, T.J., Kim, J.S., Kwon, S.J., Lim, K.B., Choi, B.S., Kim, J.A., et al. 2006. Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. Plant Cell, **18**(6): 1339–1347. doi:10.1105/tpc.105.040535. PMID:16632644.

Zhang, W., Duan, S., Kistner, E.O., Bleibel, W.K., Huang, R.S., Clark, T.A., et al. 2008. Evaluation of genetic variation contributing to differences in gene expression between populations. Am. J. Hum. Genet. **82**(3): 631–640. doi:10.1016/j.ajhg.2007.12.015. PMID:18313023.

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. doi:10.1016/j.plaphy.2009.03.002. PMID:19345111.

This article has been cited by:

1. Baojun Li, Weiguo Zhao, Dianrong Li, Hongbo Chao, Xiaoping Zhao, Na Ta, Yonghong Li, Zhoubo Guan, Liangxing Guo, Lina Zhang, Shisheng Li, Hao Wang, Maoteng Li. 2018. Genetic dissection of the mechanism of flowering time based on an environmentally stable and specific QTL in Brassica napus. *Plant Science* **277**, 296-310. [Crossref]
2. M. Michelle Malmberg, Fan Shi, German C. Spangenberg, Hans D. Daetwyler, Noel O. I. Cogan. 2018. Diversity and Genome Analysis of Australian and Global Oilseed Brassica napus L. Germplasm Using Transcriptomics and Whole Genome Re-sequencing. *Frontiers in Plant Science* **9**. . [Crossref]
3. Muhammad Afzal, Salem Safer Alghamdi, Muhammad Habib ur Rahman, Awais Ahmad, Tahir Farooq, Mukhtar Alam, Imtiaz Ali Khan, Hidayat Ullah, Wajid Nasim, Shah Fahad. 2018. Current status and future possibilities of molecular genetics techniques in Brassica napus. *Biotechnology Letters* **40**:3, 479-492. [Crossref]
4. Abha Agnihotri, Manoj Kumar, Divya Kilam, Jasroop Aneja. Genomic and Transcriptomic Approaches for Quality Improvement in Oilseed Brassicas 31-48. [Crossref]
5. Ushan I. Alahakoon, Ali Taheri, Naghabushana K. Nayidu, Delwin Epp, Min Yu, Isobel Parkin, Dwayne Hegedus, Peta Bonham-Smith, Margaret Y. Gruber. 2016. Hairy Canola (Brasssica napus) re-visited: Down-regulating TTG1 in an AtGL3-enhanced hairy leaf background improves growth, leaf trichome coverage, and metabolite gene expression diversity. *BMC Plant Biology* **16**:1. . [Crossref]
6. A. S. Mason, R. J. Snowdon. 2016. Oilseed rape: learning about ancient and recent polyploid evolution from a recent crop species. *Plant Biology* **18**:6, 883-892. [Crossref]
7. Wayne E. Clarke, Erin E. Higgins, Joerg Plieske, Ralf Wieseke, Christine Sidebottom, Yogendra Khedikar, Jacqueline Batley, Dave Edwards, Jinling Meng, Ruiyuan Li, Cynthia Taylor Lawley, Jérôme Pauquet, Benjamin Laga, Wing Cheung, Federico Iniguez-Luy, Emmanuelle Dyrszka, Stephen Rae, Benjamin Stich, Rod J. Snowdon, Andrew G. Sharpe, Martin W. Ganal, Isobel A. P. Parkin. 2016. A high-density SNP genotyping array for Brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theoretical and Applied Genetics* **129**:10, 1887-1899. [Crossref]
8. Liyun Miao, Libin Zhang, Nadia Raboanatahiry, Guangyuan Lu, Xuekun Zhang, Jun Xiang, Jianping Gan, Chunhua Fu, Maoteng Li. 2016. Transcriptome Analysis of Stem and Globally Comparison with Other Tissues in Brassica napus. *Frontiers in Plant Science* **7**. . [Crossref]
9. Jennifer Ming-Suet Ng, Mei Han, Perrin H. Beatty, Allen Good. "Genes, Meet Gases": The Role of Plant Nutrition and Genomics in Addressing Greenhouse Gas Emissions 149-172. [Crossref]
10. Felix Hirschmann, Jutta Papenbrock. 2015. The fusion of genomes leads to more options: A comparative investigation on the desulfo-glucosinolate sulfotransferases of Brassica napus and homologous proteins of Arabidopsis thaliana. *Plant Physiology and Biochemistry* **91**, 10-19. [Crossref]
11. Ravinder Singh, Venkatesh Bollina, Erin E. Higgins, Wayne E. Clarke, Christina Eynck, Christine Sidebottom, Richard Gugel, Rod Snowdon, Isobel A. P. Parkin. 2015. Single-nucleotide polymorphism identification and genotyping in Camelina sativa. *Molecular Breeding* **35**:1. . [Crossref]
12. Christian Obermeier, Wolfgang Friedt. Applied oilseed rape marker technology and genomics 253-295. [Crossref]
13. Anoop Sindhu, Larissa Ramsay, Lacey-Anne Sanderson, Robert Stonehouse, Rong Li, Janet Condie, Arun S. K. Shunmugam, Yong Liu, Ambuj B. Jha, Marwan Diapari, Judith Burstin, Gregoire Aubert, Bunyamin Tar'an, Kirstin E. Bett, Thomas D. Warkentin, Andrew G. Sharpe. 2014. Gene-based SNP discovery and genetic mapping in pea. *Theoretical and Applied Genetics* **127**:10, 2225-2241. [Crossref]
14. Sarah Schiessl, Birgit Samans, Bruno HÃ¼ttel, Richard Reinhard, Rod J. Snowdon. 2014. Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species Brassica napus. *Frontiers in Plant Science* **5**. . [Crossref]
15. Amit A Deokar, Larissa Ramsay, Andrew G Sharpe, Marwan Diapari, Anoop Sindhu, Kirstin Bett, Thomas D Warkentin, Bunyamin Tar'an. 2014. Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics* **15**:1, 708. [Crossref]
16. France Dufresne, Marc Stift, Roland Vergilino, Barbara K. Mable. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* **23**:1, 40-69. [Crossref]
17. Lijuan Wei, Meili Xiao, Alice Hayward, Donghui Fu. 2013. Applications and challenges of next-generation sequencing in Brassica species. *Planta* . [Crossref]
18. Eric Jenczewski, A.M. Chèvre, K. Alix. Chromosomal and Gene Expression Changes in Brassica Allopolyploids 171-186. [Crossref]

19. Andrew G Sharpe, Larissa Ramsay, Lacey-Anne Sanderson, Michael J Fedoruk, Wayne E Clarke, Rong Li, Sateesh Kagale, Perumal Vijayan, Albert Vandenberg, Kirstin E Bett. 2013. Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. *BMC Genomics* **14**:1, 192. [Crossref]

20. Graham J. King. Genome Analysis 91-109. [Crossref]

21. Mahmood Gholami, Wubishet A. Bekele, Joerg Schondelmaier, Rod J. Snowdon. 2012. A tailed PCR procedure for cost-effective, two-order multiplex sequencing of candidate genes in polyploid plants. *Plant Biotechnology Journal* **10**:6, 635-645. [Crossref]

22. Rod J. Snowdon, Federico L. Iniguez Luy. 2012. Potential to improve oilseed rape and canola breeding in the genomics era. *Plant Breeding* **131**:3, 351-360. [Crossref]

23. GruberMargaret, WuLimin, LinksMatthew, GjetvajBranimir, DurkinJonathan, LewisChristopher, SharpeAndrew, LydiateDerek, HegedusDwayne. 2012. Analysis of expressed sequence tags in Brassica napus cotyledons damaged by crucifer flea beetle feeding. *Genome* **55**:2, 118-133. [Abstract] [Full Text] [PDF] [PDF Plus] [Supplemental Material]

24. H. Raman, R. Raman, M. N. Nelson, M. N. Aslam, R. Rajasekaran, N. Wratten, W. A. Cowling, A. Kilian, A. G. Sharpe, J. Schondelmaier. 2012. Diversity Array Technology Markers: Genetic Diversity Analyses and Linkage Map Construction in Rapeseed (Brassica napus L.). *DNA Research* **19**:1, 51-65. [Crossref]

25. Sheng Chen, Matthew N. Nelson, Anne-Marie Chèvre, Eric Jenczewski, Zaiyun Li, Annaliese S. Mason, Jinling Meng, Julie A. Plummer, Aneeta Pradhan, Kadambot H. M. Siddique, Rod J. Snowdon, Guijun Yan, Weijun Zhou, Wallace A. Cowling. 2011. Trigenomic Bridges for Brassica Improvement. *Critical Reviews in Plant Sciences* **30**:6, 524-547. [Crossref]

26. E.BalázsE. Balázs, W.A.CowlingW. Cowling. 2010. Exploiting genome-wide association in oilseed Brassica speciesThis article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming".. *Genome* **53**:11, 853-855. [Citation] [Full Text] [PDF] [PDF Plus]