Second-generation sequencing for gene discovery in the Brassicaceae

Alice Hayward1, Guru Vighnesh1, Christina Delay1, Mohd Rafizan Samian1, Sahana Manoli2, Jiri Stiller2, Megan McKenzie1, David Edwards2 and Jacqueline Batley1.*

1ARC Centre of Excellence for Integrative Legume Research, School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia
2Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia

Keywords: second-generation sequencing, TAGdb, Brassicaceae, symbiosis, homolog, evolution, NSP.

Introduction

The ability of crop species to differentiate friend from foe and mount the appropriate response is central to disease resistances as well as symbiotic associations beneficial to productivity. Beneficial interactions include the ancient symbiosis between almost all higher plant species (>80%) and soil fungi known as mycorrhizas (Wang and Qiu, 2006). This association evolved over 460 million years ago (Redecker et al., 2000) and improves root nutrient acquisition while supplying photosynthate to the mycorrhiza in return. Many of the genes involved in this interaction were more recently recruited (60 MYA) by legumes (Fabaceae family) to enable symbiosis with nitrogen-fixing soil bacteria known as rhizobia (Sprent, 2008). The molecular adaptations underlying the evolution of mycorrhizal and rhizobial endosymbioses remain elusive, despite substantial biotechnological interest in this field. Zhu et al. (2006) suggest that a conserved microsyntenic homolog for many of the genes involved in nodulation can be found in the non-legumes rice (Oryza sativa) and Arabidopsis (Arabidopsis thaliana). Arabidopsis is a member of the Brassicaceae, one of the most economically important plant families (Hayward, 2011). Within this family, Brassica napus (canola/rapeseed) is one of the top three most valuable oilseed crops (Qiu et al., 2006; Snowdon et al., 2006). The Brassicaceae cannot form root nodules and is also one of only a few plant families unable to associate with mycorrhiza, although this is likely the result of incompatible secondary metabolite production by Brassicaceae roots (Gerdemann, 1968; Tester et al., 1987). Nevertheless, Brassicaceae members can form beneficial interactions with other species, including the endophytic, basidiomyce fungus Piriformospora indica (Oelmüller et al., 2009; Sherameti et al., 2005; Sun et al., 2010; Verma and Varma, 1998).

Summary

The Brassicaceae contains the most diverse collection of agriculturally important crop species of all plant families. Yet, this is one of the few families that do not form functional symbiotic associations with mycorrhizal fungi in the soil for improved nutrient acquisition. The genes involved in this symbiosis were more recently recruited by legumes for symbiotic association with nitrogen-fixing rhizobia bacteria. This study applied second-generation sequencing (SGS) and analysis tools to discover that two such genes, NSP1 (Nodulation Signalling Pathway 1) and NSP2, remain conserved in diverse members of the Brassicaceae despite the absence of these symbioses. We demonstrate the utility of SGS data for the discovery of putative gene homologs and their analysis in complex polyploid crop genomes with little prior sequence information. Furthermore, we show how this data can be applied to enhance downstream reverse genetics analyses. We hypothesize that Brassica NSP genes may function in the root in other plant–microbe interaction pathways that were recruited for mycorrhizal and rhizobial symbioses during evolution.

Introduction

The ability of crop species to differentiate friend from foe and mount the appropriate response is central to disease resistances as well as symbiotic associations beneficial to productivity. Beneficial interactions include the ancient symbiosis between almost all higher plant species (>80%) and soil fungi known as mycorrhizas (Wang and Qiu, 2006). This association evolved over 460 million years ago (Redecker et al., 2000) and improves root nutrient acquisition while supplying photosynthate to the mycorrhiza in return. Many of the genes involved in this interaction were more recently recruited (60 MYA) by legumes (Fabaceae family) to enable symbiosis with nitrogen-fixing soil bacteria known as rhizobia (Sprent, 2008).

The molecular adaptations underlying the evolution of mycorrhizal and rhizobial endosymbioses remain elusive, despite substantial biotechnological interest in this field. Zhu et al. (2006) suggest that a conserved microsyntenic homolog for many of the genes involved in nodulation can be found in the non-legumes rice (Oryza sativa) and Arabidopsis (Arabidopsis thaliana). Arabidopsis is a member of the Brassicaceae, one of the most economically important plant families (Hayward, 2011). Within this family, Brassica napus (canola/rapeseed) is one of the top three most valuable oilseed crops (Qiu et al., 2006; Snowdon et al., 2006). The Brassicaceae cannot form root nodules and is also one of only a few plant families unable to associate with mycorrhiza, although this is likely the result of incompatible secondary metabolite production by Brassicaceae roots (Gerdemann, 1968; Tester et al., 1987). Nevertheless, Brassicaceae members can form beneficial interactions with other species, including the endophytic, basidiomyce fungus Piriformospora indica (Oelmüller et al., 2009; Sherameti et al., 2005; Sun et al., 2010; Verma and Varma, 1998).

Two microsyntenic genes with high sequence conservation between legumes, Arabidopsis and rice are NSP1 (Nodulation Signalling Pathway 1) and NSP2. These encode GRAS domain transcription factors that function downstream of rhizobia nod-factor (NF) perception to activate nodule-specific gene expression and initiate nodule organogenesis (Kaló et al., 2005; Hirsch et al., 2009; Ferguson et al., 2010; Figure 1). In rice, homologs for NSP1 and NSP2 can complement the non-nodulating phenotype of the Lotus japonicus nsp1 and nsp2 mutants (Heckmann et al., 2006; Yokota et al., 2010). This suggests a strong functional constraint on these genes during evolution, maintained after the divergence of dicots from monocots.

One hypothesis for the sequence conservation and microsynteny of symbiosis-related genes in diverse taxa is that these genes are functionally diversified from a minimal set required for plant–microbe interactions. Consistent with this, the closest Arabidopsis homolog for the cation channel gene, POLLUX, which acts upstream of NSP1 and NSP2 in nodulation (Figure 1), is required for the P. indica–Arabidopsis interaction (Shahollari et al., 2007). Furthermore, the most similar genes to the NF receptor, NFR1, in Arabidopsis (AtCERK1 (Chitin Elicitor Receptor Kinase1)) and rice (OsCERK1) are involved in responses to fungal and bacterial elicitors including chitin, flagellin and peptidoglycan. Arabidopsis mutants for AtCERK1 show increased susceptibility to various bacterial infections (Kaku et al., 2006; Miya et al., 2007; Shimizu et al., 2010; Wan et al., 2008; Willmann et al., 2011). NSP1 and NSP2 themselves were recently shown to be essential for strigolactone biosynthesis in both rice and the model legume Medicago truncatula (Liu et al., 2011). Strigolactone is required for auxin-mediated developmental responses and is involved in mycorrhizal associations and the regulation of axillary shoot and root architecture in plants (Gomez-Roldan et al., 2008; Ruyter-Spira et al., 2011;
Umehara et al., 2008). Thus, the conservation of these genes in non-legumes may relate to a core function in hormone signalling.

Given the importance of nitrogen-use efficiency in crops and our dependence on exogenous fertilizers, improving our understanding of the evolution of nutrient acquisition pathways in higher plants remains a priority. Second-generation sequencing (SGS) technologies are advancing rapidly, and it is now possible to sequence an entire genome in a matter of days. Once complete, this can be used as a reference for assembling multiple related genomes to identify polymorphisms associated with evolutionary constraints and trait variations (Duran et al., 2010; Edwards and Batley, 2010; Imelfort et al., 2009). The recent release of the Brassica ‘A’ genome from Brassica rapa (Wang et al., 2011) will undoubtedly provide an invaluable base upon which to advance our understanding of genome diversity and gene function in Brassica crop species.

In this study, we describe the use of currently available SGS databases and applications, both pre- and postrelease of the Brassica ‘A’ reference genome, to investigate whether the NSP1 and NSP2 genes from legumes are conserved across the Brassicaceae. The data will form the basis of a larger-scale evolutionary and functional analysis of symbiosis-related genes in non-mycorrhizal, non-nodulating plants.

### Results and Discussion

Prior to the availability of SGS technology, the discovery of putative gene homologs often depended on degenerate PCR, with primers designed to the species from which the gene of interest was first cloned. This process can be time-consuming and difficult if sequences have diverged significantly over evolutionary time. SGS data can be generated incredibly quickly and takes the form of millions of short sequence reads, often in pairs derived from each end of a short DNA fragment of known size. There is currently a large amount of data derived from SGS projects that is available to be freely interrogated.

TAGdb is an SGS database and alignment tool that maps available read pairs to a reference input sequence (Marshall et al., 2010). This is applicable to any species where SGS data are available and thus is particularly useful for rapid gene discovery in orphan or complex crop species based on homology to a gene from a model organism. Even in cases where data quantity is low or species divergence is high, resulting in poor read coverage, the paired nature of most SGS reads means that primers can be designed to large-insert read pairs spanning the gene of interest for traditional amplification and sequencing (e.g. Marshall et al., 2010).

### Identification of putative NSP gene homologs in SGS data using TAGdb

In this study, the Arabidopsis genes microsyntenic to NSP1 and NSP2 (Zhu et al., 2006) were used to interrogate the available Brassicaceae SGS data on TAGdb (Marshall et al., 2010). Sequence databases available are shown in Table 1 and include paired-end (short insert) and mate-pair (large-insert) libraries for B. rapa cv. chifui and cv. kenshin (diploid A genome), mate-pair libraries for B. nigra (diploid B genome) and B.oleracea (diploid C genome) and paired-end libraries for three wild Brassicaceae species thought to be tertiary B-genome progenitors; Diplotaxis tenuifolia, Hirschfeldia incana and Sinapis alba.

In this study, read coverage across the reference genes was dependent on the level of inter-species divergence, the length and insert size of the paired reads, and the quality and quantity of the SGS library. Paired-end, high quality and quantity libraries with longer read lengths resulted in the highest coverage, particularly for more closely related species and highly conserved genes. Thus, the paired-end, 100-bp, paired-end libraries for the three tertiary Brassicaceae species more closely related to Arabidopsis mapped with highest coverage to the reference Arabidopsis sequences using TAGdb.

FASTA files of the read sequences mapping to the Arabidopsis NSP1 and NSP2 genes in TAGdb were imported into Geneious Pro (Drummond et al., 2011) and re-assembled to the reference sequence to generate consensus sequences. Complete putative coding sequences were obtained for S. alba, H. incana and D. tenuifolia (Figure S1). This demonstrates the use of TAGdb for rapidly finding putative genes of interest in SGS data for a number of diverse species in the absence of a reference genome. Following the availability of the complete Brassica ‘A’ genome (Wang et al., 2011), BLASTn analysis using Arabidopsis NSP1 and NSP2 also revealed high sequence conservation in B. rapa. The whole genomic sequences for B. rapa NSP1 and NSP2 were also used to interrogate TAGdb. This enabled improved read mapping and helped to validate the sequence consensus produced from the Arabidopsis alignments. Alignments between the complete TAGdb gene sequences and the Arabidopsis, B. rapa and L. japonicus genes are shown in Figure S1.
Sequence comparisons of putative NSP1 and NSP2 gene homologs from SGS data

Sequence topology comparisons of the *L. japonicus*, rice, Arabidopsis and predicted *B. rapa* NSP1 and NSP2 proteins are depicted in Figure 2 and summarized in Table 2. Varying levels of sequence conservation can suggest different levels of selective constraint on sequences or close physical association with additional loci under varying selection pressure. Highly divergent sequences may be indicative of functional diversification or reduced requirements for residue conservation on protein function, for example owing to redundancy.

NSP1 and NSP2 showed highest sequence conservation in the two leucine heptad repeat domains (LHRI and LHRII) and the three domains containing residues conserved across all GRAS proteins analysed to date: VHIID, PFYRE and SAW (Bolle, 2004; Kaló et al., 2005). Interestingly, an Alanine residue in the LHRI domain of NSP2 involved in hetero-dimerization with NSP1 (Hirsch et al., 2009) has been conserved in rice but not in any of the *Brassica* sequences, which all have a Glycine in that position (Figure 2 and data not shown). Alanine and Glycine have similar physiochemical properties, being small, neutral and hydrophobic, and thus it is possible that this difference has limited functional relevance. Indeed, both SNAP (Bromberg and Rost, 2007) and SIFT (Kumar et al., 2009), which estimate the likelihood that amino acid substitutions will affect protein function, predicted a neutral effect of this substitution. The N-terminal region of NSP1 and NSP2 is more variable. This domain determines specificity of protein–protein and/or DNA–protein interaction (Hirsch et al., 2009, Yokota et al., 2010), and thus divergence here may reflect differences in the functional pathways regulated by these proteins between legumes and Brassicaceae.

The NSP1 and NSP2 genes from rice rescue the non-nodulating phenotype of the respective *L. japonicus* nsp1 and nsp2 mutants (Heckmann et al., 2006; Yokota et al., 2010). This suggests a high level of functional constraint on NSP sequences during evolution. NSP2 is involved in mycorrhizal root colonization (Maillet et al., 2011) and both NSP1 and NSP2 also appear to be required for the synthesis of strigolactones (Liu et al., 2011). Such additional functions provide a likely explanation for their functional conservation between rice and *L. japonicus*. Strigolactones are intimately intertwined with auxin in the

<table>
<thead>
<tr>
<th>Species</th>
<th>Library name</th>
<th>Insert size (bp)</th>
<th>Read-length (bp)</th>
<th># Reads</th>
<th>Estimated coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brassica rapa cv. chifu</td>
<td>BrC_D3_001</td>
<td>140</td>
<td>36</td>
<td>88297796</td>
<td>6.0x</td>
</tr>
<tr>
<td>Brassica rapa cv. chifu</td>
<td>BrC_03_002</td>
<td>300</td>
<td>36</td>
<td>20466206</td>
<td>1.4x</td>
</tr>
<tr>
<td>Brassica rapa cv. chifu</td>
<td>BrC_27_001</td>
<td>2700</td>
<td>35</td>
<td>15427868</td>
<td>1.0x</td>
</tr>
<tr>
<td>Brassica rapa cv. chifu</td>
<td>BrC_37_001</td>
<td>2800</td>
<td>35</td>
<td>5578814</td>
<td>0.4x</td>
</tr>
<tr>
<td>Brassica rapa cv. chifu</td>
<td>BrC_37_002</td>
<td>2800</td>
<td>35</td>
<td>4238046</td>
<td>0.3x</td>
</tr>
<tr>
<td>Brassica rapa cv. kenshin</td>
<td>BrK_03_001</td>
<td>410</td>
<td>36</td>
<td>8466882</td>
<td>0.6x</td>
</tr>
<tr>
<td>Brassica rapa cv. kenshin</td>
<td>BrK_03_002</td>
<td>410</td>
<td>76</td>
<td>5619638</td>
<td>0.8x</td>
</tr>
<tr>
<td>Brassica nigra</td>
<td>Brn_37_001</td>
<td>2700</td>
<td>76</td>
<td>14613374</td>
<td>1.8x</td>
</tr>
<tr>
<td>Brassica oleracea</td>
<td>Bol_37_001</td>
<td>3000</td>
<td>76</td>
<td>15262520</td>
<td>1.7x</td>
</tr>
<tr>
<td>Diplotaxis tenuifolia</td>
<td>Dte_03_001</td>
<td>330</td>
<td>100</td>
<td>48459974</td>
<td>4.3x</td>
</tr>
<tr>
<td>Hirschfeldia incana</td>
<td>Hin_03_001</td>
<td>330</td>
<td>100</td>
<td>51905152</td>
<td>4.6x</td>
</tr>
<tr>
<td>Sinapis alba</td>
<td>Sal_03_001</td>
<td>330</td>
<td>100</td>
<td>36474002</td>
<td>3.2x</td>
</tr>
</tbody>
</table>

Figure 2 Visual summary of sequence similarity between lotus, rice, Arabidopsis and Brassica NSP-like proteins. Black shading represents regions of 100% homology between all sequences. Dark blue and grey= active domains and residues; Orange = Brassica rapa TILLING mutations; Light blue = NSP2 Alanine residue important for NSP1:NSP2 interaction in Lotus japonicus.
regulation of axillary meristem activity, a function that is largely conserved between the Brassicaceae member, Arabidopsis and the legume, Pisum sativum (garden pea; Gomez-Roldan et al., 2008; Umehara et al., 2008; Hayward et al., 2009). Therefore, a plausible hypothesis would be that NSP transcription factors carry out an essential function in the hormone-mediated control of meristem activity and organogenesis and that this has been recruited for nodule organogenesis.

Sequencing of putative NSP1 and NSP2 homologs in multiple Brassica accessions

To further unravel the possible functional conservation of NSP genes across the Brassicaceae as a whole, their sequences were determined for a wider range of species. In studies where SGS read coverage is insufficient to generate consensus sequences, for example for more diverse species, SGS reads that map to the 5’ and 3’UTR regions of the reference gene can be used to design species-specific or degenerate primers for PCR amplification and sequencing of the complete coding regions (Marshall et al., 2010). Where sequence divergence is too high in gene flanking regions, it is still possible to design specific primers based on reads that map to the coding sequence of the reference gene and have an un-mapped mate partner located a known insert distance up or downstream.

In this study, primers were designed in the gene flanking regions of NSP1 and NSP2 to amplify in a diverse set of Brassicaceae, from the ATFFCC (Australian Temperate Field Crops Collection), including diploid and polyploid, domesticated and wild Brassicaceae (detailed in Table S1). For optimal primer design, SGS data from TAgdb was used to determine the areas of highest interspecific sequence conservation flanking the coding sequences. The B. rapa genomic sequences proved particularly valuable as references for this purpose; yet, in the case of NSP1, this was also possible using the Arabidopsis reference (Figure 3). The resulting PCR amplicons were either directly sequenced (possible for the diploid species) or cloned and sequenced to distinguish homeologs in the polyploids. Complete sequences were obtained from B. nigra, B. oleracea, B. juncea, B. carinata, B. ruvo, B. montana, B. incana, B. oxyryhina, Crambe abyssinica and Matthiola longepetala. The ‘A’ and ‘C’ genome homeologs from B. napus were also obtained for NSP1. In most cases, comparison of cloned sequences with those from the ‘A’, ‘B’ and ‘C’ diploid genome species delimited the likely genomic source from which each homeologous copy was derived.

Phylogenetic analysis of putative NSP1 and NSP2 homologs

To better understand the evolution of Brassicaceae NSP1 and NSP2 proteins, their sequences were aligned with a larger diversity set comprising complete, higher-plant NSP-like proteins identified from the NCBI protein databases using Geneious Pro (Drummond et al., 2011). Phylogenetic trees derived from these alignments were constructed in Geneious Pro using PhyML (Guindon and Gascuel, 2003; Figure 4). GRAS proteins form a large multi-gene family in higher plants. The Brassicaceae sequences most similar to NSP1 and NSP2 from legumes formed a distinct sub-group more closely related to their cognate sequences from other dicots, including the leguminosae, than to the monocot sequences. The Brassicaceae clades for these genes divided nicely into sub-clades based on putative genome derivation (‘A’, ‘B’ or ‘C’), with few exceptions. The Arabidopsis sequences tended to be more highly related to the tertiary progenitors and ‘B’ genome species, explaining the high TAgdb coverage results for these species when using the Arabidopsis reference.

The non-Brassicaceae dicot sequences for the NSP1 and NSP2 trees grouped into one sub-clade which was further divided into two distinct sub-groups. One of these groups was specific for the legumes and the other contained sequences including Ricinus communis (Rc; Castor oil plant), Populus trichocarpa (Pt; poplar), Vitis Vinifera (Vv; grape vine) and Nicotiana benthamiana (Nb; tobacco). This suggests that a common ancestral gene gave rise to the Brassicaceae NSP1 and NSP2 genes prior to the divergence of the legume-specific sequences from these other dicot species, presumably upon recruitment for nodulation.

TILLING for nsp1 mutants in Brassica rapa

Mutants for genes of interest are excellent resources for determining gene function, providing they have a quantifiable phenotype. In the past, forward genetics approaches were used to assign function to particular loci, whereby randomly generated mutants were phenotyped followed by map-based cloning of the associated genes. In reverse genetics approaches, where putative gene sequences are available in the absence of phenotype data, methods such as TILLING (Targeting Induced Local Lesions in Genomes; McCallum et al., 2000) can be used to generate targeted, allelic variants for mutant phenotyping.

In this study, the availability of Brassica SGS data and a B. rapa TILLING service provided by RevGen (John Innes Centre, Norwich, UK) presented an excellent base upon which to begin functional analysis of the NSP loci in B. rapa. EMS-generated mutant alleles within a number of individuals of the TILLING population were discovered for BrNSP1 in this way (Figure 2; Table 3). We demonstrate the use of SGS data combined with TILLING, protein functional domain prediction and mutation effect prediction, for reverse genetics studies in Brassica.

Specifically, Geneious Pro (Drummond et al., 2011) was used to assign and annotate putative functional domains and residues on the BrNSP1 protein sequence. The freely available SNAP (Bromberg and Rost, 2007) and SIFT (Kumar et al., 2009) programs were then applied to best estimate the functional tolerance of the putative BrNSP1 protein to the various point mutations in the TILLING mutants. With a few exceptions, both

<table>
<thead>
<tr>
<th>L. japonicus protein</th>
<th>Class</th>
<th>Function</th>
<th>Putative Arabidopsis homolog</th>
<th>% Identity</th>
<th>Putative B. rapa homolog</th>
<th>% Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSP1</td>
<td>GRAS-domain TF</td>
<td>Activation of nodule-expressed genes</td>
<td>AT3G13840:AtNSP1</td>
<td>50</td>
<td>BrNSP1</td>
<td>46</td>
</tr>
<tr>
<td>NSP2</td>
<td>GRAS-domain TF</td>
<td>Activation of nodule-expressed genes</td>
<td>AT4G08250:AtNSP2</td>
<td>50</td>
<td>BrNSP2</td>
<td>49</td>
</tr>
</tbody>
</table>
programs were generally consistent in these predictions. This basic analysis allowed for identifying individuals containing mutant alleles most likely to cause altered protein function and hence a quantifiable phenotype. These will be backcrossed and bulked in the homozygous state for phenotypic analysis of microbe responses and potential structural, developmental and molecular abnormalities. This pipeline of rapid gene discovery using SGS data, followed by identification of best mutant candidates in available TILLING resources, is an example of a useful reverse genetics approach for better understanding gene function in crop species, orphan species or species that are recalcitrant to genetic engineering technologies.

Expression analysis of *Brassica napus* NSP1 and NSP2 homeologs

The analysis of transcript abundance in different tissues at different developmental stages can reveal the likely spatiotemporal requirement for genes, delimiting putative function to a particular tissue, response or stage of development. In polyploid species, such as *B. napus*, gene homeologs derived from multiple genomes can have divergent expression profiles, in some cases with only one homeolog expressed and functional. Therefore, the ability to distinguish between homeologs for expression analyses is essential. Here we demonstrate the use of SGS data...
to distinguish the ‘A’ and ‘C’ genome homeologs of NSP1 and NSP2 in B. napus for transcript quantification using quantitative real-time PCR (qRTPCR).

Based on alignment of genomic B. napus SGS reads to the entire reference ‘A’ genome, genome-specific primers for qRTPCR were designed by targeting intergenomic SNPs within the putative ‘A’ and ‘C’ genome homeologs for each B. napus NSP homeolog (Figure 5). These specific primers were applied to analysis of the tissue-specific expression of NSP1 and NSP2 to better understand their putative role in Brassicas. The ‘A’ and ‘C’ genome homeologs of both NSP1 and NSP2 were expressed at similar levels in each tissue studied except for the roots, which expressed significantly higher levels of the ‘A’ genome copy relative to the ‘C’ homeolog (P > 0.05; Figure 6). This may suggest preferential activity of the A-genome copy in B. napus. The NSP1 and NSP2 transcripts were also co-expressed, with both genes significantly increased in roots relative to the other organs tested (P > 0.05). This is comparable with the expression profile for these genes in legumes (Heckmann et al., 2006) and consistent with a primary function for these genes in the roots of Brassica species, and in strigolactone synthesis (Hayward et al., 2009; Liu et al., 2011). Furthermore, the co-regulation of these genes supports the possibility that they may function in the same pathway and potentially...
interact in *B. napus*. Therefore, it is hypothesized that the legume NSP genes were recruited from a common root-expressed ancestor gene for nodule organogenesis. Current studies are determining if the NSP homeologs are induced by microbe and elicitor treatments, similarly to their legume counterparts, in a larger set of Brassicaceae species (Heckmann et al., 2006; Lopez-Gomez et al., 2012), with early indications that this is indeed the case (A. Hayward unpublished data).

**Experimental procedures**

**Identification of putative *Brassica* homologs using TAGdb**

The Arabidopsis and *Brassica rapa* cv. chifu genomic regions sharing highest sequence similarity with *Medicago truncatula* and *Lotus japonicus* NSP1 and NSP2 (including the gene and a variable amount of surrounding sequence up to maximum of 5 Kb) were used to query the available TAGdb (Marshall et al., 2010) Brassicaceae data sets. TAGdb is available at [http://flora.acpfg.com.au/tagdb/cgi-bin/index](http://flora.acpfg.com.au/tagdb/cgi-bin/index) (Australian Centre for Plant Functional Genomics). Arabidopsis gene accessions are shown in Table 2, TAGdb sequence libraries are listed in Table 1.

FASTA files of aligning read-pairs were downloaded and re-assembled to the corresponding genomic reference regions using Geneious Pro (Drummond et al., 2011). Consensus sequences for the CDS region were derived from these tag assemblies after careful visual filtering of reads to remove any non-conformist/divergent sequences. These sequences were then used for alignments and phylogenetic analyses.

**Plant growth and tissue harvests**

For DNA extractions, seed from a diverse set of Brassicaceae species was kindly provided by Dr Bob Redden, Curator of the Australian Temperate Field Crops Collection (ATFCC), NSW Department of Primary Industries, Australia. Plants were grown in California potting mix type C in a temperature controlled glasshouse at 18°C, 16 h-day/15°C, 8 h-night, at the University of California, Berkeley.
of Queensland, St Lucia, Australia. Young leaf tissue from single individuals per species was collected for DNA extractions.

For the tissue profile expression analysis, *Brassica napus* c.v. Ag Spectrum seedlings were potted into autoclaved vermiculite at a density of eight seeds per 15 cm diameter pot (128 plants total). Pots were covered with glad wrap and placed in the cold room for 3 days before removal of wrap and transfer to a growth chamber at 24°C and 70% humidity. Plants were grown for 3 weeks before harvesting the youngest expanding leaf (YL), oldest leaf (OL), basal internode (I), hypocotyl (H) and root (R) tissues. Tissues were harvested from 5 plants in each of two biological pools into 1.5 mL microtubes in liquid N2. All tissues for DNA and RNA extractions were stored at −80°C.

**DNA and RNA extractions and cDNA synthesis**

Tissues for DNA and RNA extractions were ground to a fine powder in a mortar and pestle in liquid N2. DNA was extracted using a modified microprep protocol (Fulton et al., 1995). RNA was extracted using a NucleoSpin RNA Plant extraction kit (Machery-Nagel, Düren, Germany) with DNase-treatment as per manufacturer’s instructions and eluted in 30 μL H2O. Nucleic acids were quantified using a NanoDrop ND1000 (Thermo Fisher Scientific Australia Pty Ltd., Victoria, Australia) and visualized for integrity on 1% TAE-agarose gels containing Ethidium Bromide.

Extracted RNA was assessed for contaminating DNA by PCR with 0.5 μM Forward (5'-GGTGCCTGGATTCAATGTTAC-3') and 0.5 μM Reverse (5'-CTGATTGGAGGACACAA-3') primers, 1 x iTaq buffer with MgCl₂ (Scientifix, Victoria, Australia), 200 nM each dNTP, 1 U of iTaq (Scientifix) and 20 ng template RNA (similar to levels used for qRTPCR). Cycling conditions were: 94°C for 2 min, 45 cycles of 94°C for 15 s, 60°C for 30 s, and 72°C for 30 s, with a final extension of 72°C for 5 min. DNA-contaminated RNA samples were re-treated with rDNase (Machery-Nagel) and re-precipitated as per manufacturer’s instructions. This RNA was resuspended in 15 μL RNase-free H2O and quantified and checked for integrity as above. cDNA was synthesized from 250 ng total RNA in 20 μL using the Takara BluePrint First Strand cDNA Synthesis kit as per manufacturer’s instructions (Takara, Shiga, Japan). Reverse transcriptase-minus control reactions were performed for each RNA sample to assess any remaining DNA contamination during qRTPCR.

**PCR and sequencing of Brassica NSP1 and NSP-like genes**

PCR primers were designed in Geneious Pro (Drummond et al., 2011) to conserved regions of the predicted 5' and 3' UTRs of NSP1 and NSP2 based on alignments between available sequences and read pairs for Brassicaceae members (Figure 3). Primer properties were checked using the Sigma-Aldrich DNA Calculator available at http://www.sigmaaldrich.com/technical-service-home/web-tool-box.html. Primer details are shown in Table S2. PCR products were amplified from 50 ng of DNA from each of the species listed in Table S1 using 0.5 μM forward and reverse oligonucleotide primer, 1 U of Phusion Hot Start High-Fidelity DNA Polymerase (Finnzymes, Vantaa, Finland), 1 x Phusion GC Buffer containing 1.5 mM MgCl₂, and 200 μM each dNTP, in a gradient Thermocycler (MJ Research, Waltham, MA). Cycling conditions were 98°C for 30 s followed by 35 cycles of 98°C for 10 s, gradient Tm-based annealing for 30 s (Table S2), and extension at 72°C for 45 s. Final extension was 72°C for 10 min. Amplified products were visualized under UV light on 1% TAE-agarose gels containing Ethidium Bromide. The GeneRuler™ 1 Kb marker (Fermentas, Thermo Fisher Scientific Australia Pty Ltd) was used as a size standard. PCR products were gel-purified (Invitrogen (Life Technologies, Victoria, Australia)/Roche Applied Science (NSW, Australia)) and either directly sequenced at the Australian Genome Research Facility (AGRF), Brisbane, Australia, or cloned using the pGEM®-T-easy (Promega, NSW, Australia) and pCR®-XL-Topo® (Invitrogen) vector systems. Cloned inserts were sequenced using T7 and SP6 or M13R primers at the AGRF, Brisbane. Consensus sequences for NSP1 and NSP2 from PCR and cloned products were generated in Geneious Pro. Similarity to the ‘A’, ‘B’ and ‘C’ diploid Brassica genome sequences for each clone was calculated. All sequencing and illumina® read data generated in this study is available for download from TAGdb (Marshall et al., 2010; http://flora.acgpf.com.au/tagdb/cgi-bin/index) and/or NCBI (http://www.ncbi.nlm.nih.gov) and/or http://www.brassicagenome.net/downloads.

**Sequence manipulations and phylogenetic tree construction**

All sequence manipulations and BLAST searches were carried out using Geneious Pro (Biomatters; Drummond et al., 2011). Sequence assemblies of TAGdb read pairs to the reference genes were generated using the Assembly Tool with default parameters. All sequence alignments were generated using the MUSCLE Plugin (Edgar, 2004) or the Geneious Alignment tool with default settings. Phylogenetic trees were constructed using the Geneious Pro PhyML plugin (Guindon and Gascuel, 2003) with the MREV substitution Model and 1000 x Bootstrapping.

**qRTPCR expression analyses**

Primer pairs for qRTPCR specific to the ‘A’ and ‘C’ genome homologs of the *B. napus NSP1* and *NSP2* genes were designed in Geneious Pro (Drummond et al., 2011; Table S2) based on inter-genome SNPs found in *B. napus* read pairs aligned to the *A* and *C* gene reference sequences (Figure 5). Primers were checked using the Sigma-Aldrich DNA calculator as described previously. Each qRTPCR reaction was performed in a total of 10 μL using 1 x SensiMix® PCR Master Mix (BioLine, NSW, Australia), 200 nM each primer and 7.5 ng of cDNA in an ABI Prism 7900HT Sequence Detection System (Applied Biosystems, Life Technologies, Victoria, Australia). No-template control reactions were performed for each primer and melt-curve analysis was used to check product specificity. Primer efficiencies (PE) of each gene were calculated per run by LinRegPCR (Ramakers et al., 2003). Relative quantification was determined using the comparative CT method; cycle threshold (Ct) values for technical replicates for each gene were averaged (AvCtgene) and normalized against the *Brassica napus PP2A* control gene as per Chen et al. (2010) using the formula

\[ \text{Pr}_{\text{gene}} = \frac{(\text{AvCt}_{\text{gene}})}{\text{AvCt}_{\text{reference}}} \]

Resulting values were averaged for biological replicates and plotted with standard errors relative to a chosen sample (the calibrator), as indicated on the Y-axis. In this study, the calibra-
tor is the basal internode sample for each gene. Standard errors were calculated in Microsoft Excel 2003 using the formula: $\text{STDEV} / \sqrt{n}(n)$. Statistical P-values for all phenotypic and expression data were calculated using the t-test and ‘n = 2’ referring to the number of biological pools.

**Targeting Induced Local Lesions in Genomes**

TILLING of NSP1 was carried out by the RevGen TILLING service (JIC, Norwich, UK) in *B. rapa* line R-o-18, available at: http://revgenuk.jic.ac.uk/. M2 seed lines were obtained for backcrossing and further analysis.

**Conclusions**

In reverse genetics approaches, sequence information is the first step in attempting to predict functional protein domains before in vitro or in vivo studies are undertaken. This study highlights the applicability of SGS data and associated analysis tools for rapid and efficient discovery of putative gene homologs and homeologs in complex crop genomes and orphan species with limited prior sequence information. Once sequences are known, multiple bioinformatics tools are currently available for identifying conserved protein domains, putative functional residues, secondary protein structure and even crystal structures for proteins of interest based on the plethora of information currently available in sequence databases. Once sequences have been characterized, functional analyses including gene expression quantification, targeted mutagenesis, phenotyping and proteomic assays can be used to further elucidate protein function. This has implications not only for building our wealth of biological understanding, but for maintaining allelic diversity sets and breeding improved varieties of crop species such as *B. napus* with superior allelic content.

**Acknowledgements**

The authors would like to acknowledge funding support from the Australian Research Council (LP0882095, LP0883462, LP0989200, LP110100200 and DP0985953), the University of Queensland Early Career Researcher Award Scheme and a travelling fellowship from the Australian and Chinese Academies of Sciences. Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) is gratefully acknowledged. We thank Dr Bob Redden (ATFCC, NSW DPI) for providing the diverse Brassicaceae seed, Mr Robert Simpson (University of Queensland) for qRTPCR advice and Dr Fran Robson (RevGen, UK) for identifying and providing the TILLING lines.

**References**


Supporting information

Additional Supporting information may be found in the online version of this article.

Figure S1 Multiple sequence alignments of Brassicaceae NSP genes from TAgDb. Complete gene sequences from TAgDb assemblies are aligned with the Arabidopsis and Brassica rapa reference sequences. (A) NSP1 (B) NSP2. Br = B. rapa, Dt = Diplo-taxis tenuifolia, Hi = Hirschfeldia incana, Sa = Sinapis alba.

Table S1 Brassicaceae species used for PCR amplification of NSP1 and NSP2.

Table S2 Primers designed and used in this study.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.