Fractionation of Synteny in a Genomic Region Containing Tandemly Duplicated Genes across *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*

JESSICA A. SCHLUETER, BRIAN E. SCHEFFLER, SCOTT JACKSON, AND RANDY C. SHOEMAKER

From the Department of Agronomy, Purdue University, West Lafayette, IN 47907 (Schlueter and Jackson); the USDA-ARS-MSA Genomics Laboratory, Stoneville, MS 38766 (Scheffler); and the USDA-ARS-CICGR, Ames, IA 50011 (Shoemaker).

Address correspondence to R. C. Shoemaker at the address above, or e-mail: randy.shoemaker@ars.usda.gov.

Abstract

Extended comparison of gene sequences found on homeologous soybean Bacterial Artificial Chromosomes to *Medicago truncatula* and *Arabidopsis thaliana* genomic sequences demonstrated a network of synteny within conserved regions interrupted by gene addition and/or deletions. Consolidation of gene order among all 3 species provides a picture of ancestral gene order. The observation supports a genome history of fractionation resulting from gene loss/addition and rearrangement. In all 3 species, clusters of *N*-hydroxycinnamoyl/benzoyltransferase genes were identified in tandemly duplicated clusters. Parsimony-based gene trees suggest that the genes within the arrays have independently undergone tandem duplication in each species.

Genetic colinearity and orthology has traditionally been identified through comparative mapping. Whereas marker colinearity has been highly conserved in the grasses (Ahn and Tanksley 1993; Devos and Gale 2000), in the legumes synteny has often been fragmented. Early marker studies identified conserved syntenic blocks among various legume genera, *Glycine max*, *Pisum sativum*, and *Vigna radiata* (Menancio-Hautea et al. 1993; Boutin et al. 1995). Choi, Kim et al. and Choi, Mun et al. (2004), using 60 mapped markers that were homologous between soybean and *Medicago truncatula*, identified 11 syntenic blocks. In these instances, synteny was limited to small intervals. Remnants of synteny were identified between soybean and *Arabidopsis* as well (Grant et al. 2000; Lee et al. 2001). A common theme of these studies is that synteny is high among the closely related species and decreases with larger phylogenetic distance. What has come to light in comparative studies in the grasses is that the conservation observed at the genetic map level is not necessarily observed at the sequence level (Bennetzen and Ramakrishna 2002; Feuillet and Keller 2002). In other words, genetic colinearity does not necessarily reflect sequence microcolinearity.

A BAC-based hybridization study between soybean and *Medicago* estimated that 27 of 50 soybean BACs possessed some level of microsynteny with *Medicago* (54%) (Yan et al. 2003). Yan et al. (2004) compared 8 groups of BACs between soybean and *Medicago* and found evidence of synteny between 6 pairs and microsynteny between 3 pairs. More recently, an analysis of 3 megabases of soybean sequence showed unusually high synteny to 2 *Medicago* chromosomes with upward of 75% gene colinearity (Mudge et al. 2005). Cannon et al. (2006) conducted a large-scale sequence-based comparison between 2 cool-season legumes that are both members of the Hologalegina clade, *M. truncatula* and *Lotus japonicus*. Those authors reported that the 2 species shared a minimum of 10 large-scale syntenic blocks with substantial colinearity.

Details of synteny between soybean (a member of the phaseoloides) and other legumes such as *Medicago* and *Lotus* as well as *Arabidopsis* are still largely undefined. In this study, genes from a genomic region defined by the presence of a tandemly duplicated *N*-hydroxycinnamoyl/benzoyltransferase gene family in soybean were compared with *Medicago*, *Lotus*, and *Arabidopsis*, 3 major dicot genomes with extensive sequence. Homologous regions in *Medicago* and *Arabidopsis* were identified revealing a network of synteny created by gene loss and rearrangement. This region was chosen because HCBT catalyzes the first committed step in phytoalexin biosynthesis (Yang et al. 1997), an important pathway involved in wound and pathogen response (Naoumkina et al. 2007). We also report that the HCBT gene clusters seem to have been created by independent tandem duplication in each species.
Materials and Methods

Previously, 2 homeologous soybean BACs, gmw1-74i13 (DQ336954) and gmw1-52d3 (DQ3336955), anchored by HCBT were identified, sequenced, and annotated (Schlueter et al. 2006). Putative homologous sequences from *M. truncatula*, *L. japonicus*, and *Arabidopsis thaliana* were identified by performing a TBLASTX search with the annotated soybean gene structures against the available genomic sequences from *Lotus*, *Medicago*, and *Arabidopsis* at National Center for Biotechnology Information. Genes were considered homologues if their BLAST-based e values were less than $1 \times 10^{-30}$. Extended syntenic relationships were then identified by searching the resulting BAC or genomic regions for genes that were syntenic to other *Lotus*, *Medicago*, or *Arabidopsis* sequences.

Gene structures for *Medicago* were taken from MtGDB and further confirmed by GENESeqER alignments of all *Medicago* expressed sequence tags (ESTs) and all plant putatively unique sequences (consensus sequences derived from EST contigs) against each identified BAC sequence (Schlueter et al. 2003; 2006; Dong et al. 2005). Gene structures from *Arabidopsis* were taken directly from AtGDB (Zhu et al. 2003; Schlueter et al. 2006).

A multiple sequence alignment for all identified HCBT genes was generated with MEGALIGN (LASERGENE; DNAStar, Madison, WI) using default CLUSTALW alignment and manually removing large gaps due to single-sequence overhang at the 5' - and 3' -end of the alignments. A parsimony-based tree was generated in PAUP* 4.0 (Swofford 1998) with manual rooting with *Arabidopsis* as the out-group (internal node with a basal polytomy), tree bisection reconnection branch swapping, and 1000 iterations to generate bootstrap values.

Results

Identification of a Network of Synteny with *A. thaliana* and *M. truncatula*

Annotated genes from homeologous soybean BACs DQ336954 and DQ336955 were used to identify potential syntetic blocks in *Lotus*, *Medicago*, and *Arabidopsis* by searching with TBLASTX against the available genomic sequences for those species. No syntetic blocks were observed between soybean and *Lotus*. For each homologous gene from *Medicago* and *Arabidopsis*, the corresponding regions were scanned for blocks of synteny. Using this method, a network of synteny between the 3 species was identified (Figure 1).

An analysis of the structure of the chromosomal regions indicates that the regions have undergone gene additions and/or losses. Three syntetic blocks were identified in *Medicago* corresponding to BACs AC148486 from chromosome 1, AC144431 from chromosome 5, and AC122728 also from chromosome 5 (http://www.Medicago.org). The first BAC, AC148486, shares 4 genes (WUSCHEL-related homeobox [gene 1], poly-A–binding protein [PABP, gene 2], zinc finger [gene 11], and basic-helix-loop-helix [gene 17]) colinear in order and orientation with DQ336954 and DQ336955, respectively. AC148486 also contains a remorin gene (gene 20) found only on DQ336955 (Figure 1 and Table 1). However, this colinearity is fragmented with 4 other genes on AC148486 that are not found in the syntetic soybean regions, unknown (gene 10), glycosyl hydrolase (gene 16), S-ribonuclease–binding protein (gene 18), and lipase 3 (gene 19). It is through the S-ribonuclease–binding protein and lipase 3 that a fragmented network of synteny to the second *Medicago* BAC, AC144431, was identified (Figure 1 and Table 1). Beyond gene 20 on
AC144431, no other genes were identified to be syntenic between these BACs. The final Medicago BAC identified, AC122728, contains a cluster of 3 HCBT genes and a gene encoding a PABP (gene 2) (Figure 1 and Table 1). No other HCBT genes with high sequence identity to those in soybean were identified from the available Medicago sequence.

Three chromosomes in Arabidopsis were found to contain sequences corresponding to sequences on the 2 soybean BACs (Figure 1 and Table 1). A region of chromosome 1 starting from At1g45180, a zinc finger protein (gene 11), to At1g45230, defective chloroplasts and leaves protein (gene 26), contains in total 11 genes, of which 5 are colinear in order and orientation with DQ336954 and 2 are colinear in order and orientation with DQ336955. Much as with the Medicago segments, the synteny is fragmented with noncolinear genes interrupting the colinearity to the soybean sequence.

The final 2 syntenic chromosomes in Arabidopsis contain only HCBT gene clusters as observed in soybean (Figure 1). The first cluster on chromosome 5 consists of At5g07850, At5g07860, and At5g07870; another region of chromosome 5 has At5g67150 and At5g67160. The chromosome 5 clusters are separated by just more than 24 Mb. On chromosome 3, another cluster of 3 HCBT genes, At3g50270, At3g50280, and At3g50300, was identified. No other genes in this region of the Arabidopsis chromosomes are syntenic with either soybean or Medicago.

Three other copies of HCBT in Arabidopsis were identified (At5g42830, At2g39980, and At5g01210) but existed as single genes, not tandemly duplicated.

### Table 1. Syntenic genes between soybean, Medicago, and Arabidopsis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Putative function</th>
<th>Soybean DQ336954</th>
<th>Soybean DQ336955</th>
<th>Medicago AC148486</th>
<th>Medicago AC144431</th>
<th>Arabidopsis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WOX4 protein (homeobox-leucine zipper transcription factor protein)</td>
<td>Present</td>
<td></td>
<td>Present</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>2</td>
<td>PABP</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>3</td>
<td>Membrane-like protein</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 1</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 2</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 3</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 4</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 5</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 6</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Unknown protein</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>11</td>
<td>Zinc finger (C3HC4-type) 1</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td>At1g45180</td>
</tr>
<tr>
<td>12</td>
<td>Zinc finger (C3HC4-type) 2</td>
<td>Present</td>
<td></td>
<td>Present</td>
<td></td>
<td>At1g45190</td>
</tr>
<tr>
<td>13</td>
<td>Heat shock transcription factor</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td></td>
<td>At1g45191</td>
</tr>
<tr>
<td>14</td>
<td>Arabidopsis expressed protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Glycosyl hydrolase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Reduced form of nicotinamide adenine dinucleotide dehydrogenase subunit fragment</td>
<td></td>
<td></td>
<td>Present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Basic-helix-loop-helix transcription factor</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>S-ribonuclease–binding protein</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>19</td>
<td>Lipase 3</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
<td>At1g45201</td>
</tr>
<tr>
<td>20</td>
<td>Remorin-like protein</td>
<td></td>
<td></td>
<td>Present</td>
<td>Present</td>
<td>At1g45207</td>
</tr>
<tr>
<td>21</td>
<td>Divergent C1 domain–containing protein</td>
<td></td>
<td></td>
<td>Present</td>
<td>Present</td>
<td>At1g45243</td>
</tr>
<tr>
<td>22</td>
<td>WW domain–containing protein</td>
<td></td>
<td></td>
<td></td>
<td>Present</td>
<td>At1g45231</td>
</tr>
<tr>
<td>23</td>
<td>Arabidopsis-like expressed protein</td>
<td>Present</td>
<td>Present</td>
<td></td>
<td></td>
<td>At1g45233</td>
</tr>
<tr>
<td>24</td>
<td>Arabidopsis hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>At1g45248</td>
</tr>
<tr>
<td>25</td>
<td>Basic leucine zipper transcription factor</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td>At1g45249</td>
</tr>
<tr>
<td>26</td>
<td>DCL protein</td>
<td>Present</td>
<td></td>
<td></td>
<td></td>
<td>At1g45230</td>
</tr>
<tr>
<td>27</td>
<td>Aromatic-rich family protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Pentatricopeptide-repeat–containing protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HCBT Divergence and Possible Independent Origins

Nucleotide diversity between HCBT genes within and between soybean BACs ranged from 59% to 84% (Schlueter et al. 2006). Within the Medicago clade, nucleotide identity ranges from 81% to 88%, whereas within the Arabidopsis clade, the HCBT genes showed greater divergence with identities ranging from 40% to 82%. Amino acid similarities ranged from 58% to 90% among soybean genes, 83% to 87% among Medicago genes, and 37% to 52% among Arabidopsis genes.

The multiple sequence alignment for all identified HCBT genes, using Arabidopsis as the out-group, generated a parsimony-based phylogenetic tree, which suggests that the tandem duplication of these genes likely occurred after speciation (Figure 2). The grouping of tandemly arrayed genes suggests that each cluster was independently formed. Genes from one species always form single clades (Figure 2). This is indicative of independent duplication events or may
be the result of concerted evolution through gene conversion and gene loss through unequal crossing over.

**Discussion**

Of major interest to the legume community is the ability to utilize the developing sequences from the model legumes, *Medicago* and *Arabidopsis*, as well as from *A. thaliana* for research in *G. max* and other phaseoloids. Particularly now that soybean is being sequenced via a whole-genome shotgun approach (Jackson et al. 2006), the possibility of using the BAC-by-BAC *Medicago* and/or *Lotus* sequence to aid in assembly of the soybean genome is an attractive option. For this to be possible, however, sequence-based synteny between these legumes must be fairly strong. No syntenic blocks were identified from the available *L. japonicus* sequences. This may be due to sequences for these regions not yet available due to the ongoing process of genome sequencing or there might not be syntenic blocks in *Lotus* to these soybean sequences. The TBLASTX results (data not shown) suggest that the latter may be the case. Although 33% of the genes on DQ336954 had homologues in *Lotus* with an e value less than $1 \times 10^{-30}$, none of these genes were located on the same BACs. Marker-based studies have identified regions with upward of 54% synteny between soybean and *Medicago* (Yan et al. 2003; Choi, Kim et al. 2004; Choi, Mun et al. 2004). Previous work identified relatively
strong synteny between soybean and Medicago with up to 75% of the soybean genes colinear to Medicago (Mudge et al. 2005). This analysis shows that synteny between soybean and Medicago may not be significantly more conserved than between soybean and Arabidopsis, for some regions. This is somewhat surprising considering that the divergence of soybean and Medicago occurred approximately 50 million years ago (Lavin et al. 2005; Pfiehl et al. 2005), whereas Arabidopsis and legumes likely diverged approximately 90 million years ago (Gandolfo et al. 1998; Yang et al. 1999; Grant et al. 2000).

Five of 9 genes contiguous on Medicago BAC AC148486 were conserved in order and orientation with at least one of the homeologous soybean BACs. However, 5 of 11 genes on a region of Arabidopsis chromosome 1 were also conserved in order and orientation with soybean. Although many of the genes that are conserved between species have the same order and orientation, synteny is fragmented between all 3 species. This suggests that there have been extensive additions and/or deletions of gene after divergence of these species.

These findings are somewhat different to those of Mudge et al. (2005) who found a region of hypersynteny between the 3 species. Although microsynteny may be used to identify orthologous regions from one legume species to another, based on our findings, sequence-based synteny in some regions of the genome will not be as strong as hoped.

Comparing both the soybean BACs with the Medicago and Arabidopsis sequences shows that synteny in this region is not straightforward, rather there is a network of synteny identified by fragmented blocks of genes. For example, gene 20 is in common between Arabidopsis and Medicago and only one homeologue of soybean. Gene 17 is in common between both homeologues of soybean and Medicago, but not Arabidopsis, whereas gene 19 is common between Arabidopsis and Medicago but is not found on either homeologues of soybean. Taken together, the information from these syntenic regions provides a consolidated view (Langham et al. 2004) of the relationships between these species. Given the fragmented sequence of the Medicago genome and the potential for more homeologues in soybean, the reconstruction of an ancestral genome sequence is not possible.

**Tandemly Duplicated HCBT Genes**

When identifying syntenic blocks in Medicago and Arabidopsis, tandemly amplified arrays of HCBT genes were identified in all 3 genomes (Figure 1). In Medicago, only 3 copies of HCBT were identified and clustered sequentially on 1 BAC, AC122728. In Arabidopsis, 3 clusters of HCBT genes could be identified; 2 clusters on chromosome 5, 1 with 3 copies and the other with 2, and 1 cluster of 3 genes on chromosome 3 (Figure 1). It should be noted that almost none of the surrounding genes in either Medicago or Arabidopsis showed any synteny to the soybean sequences. One BAC, AP007820, did identify a single HCBT gene in Lotus with an e value of 0; however, this BAC was still only Phase 1 with 25 unordered pieces.

Conserved tandem duplications of HCBT genes in all 3 species suggest that there is a biologically relevant function for these clusters. HCBT functions as the first step in phytalexin biosynthesis (Yang et al. 1997) and as such may have a role in pathogen response. Tandem arrays of resistance genes are thought to provide a broader range of response to pathogens (Graham et al. 2002). These HCBT clusters found in all these species raise the question as to when the clusters were formed, in other words, if these tandem duplications occurred prior to or postspecies divergence. Alternatively, these duplications may have occurred prior to divergence, and concerted evolution by gene conversion has homogenized each gene family such that sequence similarity suggests independent duplications.

A parsimony-based phylogenetic tree with these clustered HCBT sequences suggests that the tandem duplication of these genes likely occurred after speciation; in other words, each cluster of tandemly arrayed genes was independently formed. We found that genes from one species always form single clades, indicative of independent duplication events (Figure 2). The independent formation of these gene clusters supports the proposed role of these genes in pathogen response (Schlueter et al. 2006).

Although previous studies have suggested that hyper-synteny may exist between the legumes (Mudge et al. 2005), only a limited network of synteny could be identified between soybean, Medicago, and Arabidopsis. Further studies of homeologous regions in soybean are warranted to further understand the evolutionary history of paleopolyploid genomes. The forthcoming whole-genome sequence of soybean (Jackson et al. 2006) will permit us to answer in detail the remaining questions about macro- and microsynteny among these species.

**Funding**

United Soybean Board (4250) to R.C.S.

**Acknowledgments**

The authors would like to thank Mary Duke and Xiaofen Liu for their sequencing expertise. Names are necessary to report factually on the available data; however, the United States Department of Agriculture neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable. This article is a contribution of the Corn Insect and Crop Genetics Research Unit (USDA-ARS).

**References**


Received July 12, 2007
Accepted January 11, 2008

Corresponding Editor: Reid Palmer