Gene Conversion and the Evolution of Three Leucine-Rich Repeat Gene Families in *Arabidopsis thaliana*

Mariana Mondragon-Palomino\(^1\) and Brandon S. Gaut

Department of Ecology and Evolutionary Biology, University of California in Irvine

The high number of duplicated genes in plant genomes provides a potential template for gene conversion and unequal crossing-over. Within a gene family these two processes can render all members homogeneous or generate diversity by reassorting variants among paralogs. The latter is especially feasible in families where gene diversity confers a selective advantage and thus conversion events are likely to be retained. Consequently, the most complete record of gene conversion is expected to be most evident in gene families commonly subjected to positive selection. Here, we describe the extent and characteristics of gene conversion and unequal crossing-over in the coding and noncoding regions of nucleotide-binding site leucine-rich repeat (NBS-LRR), receptor-like kinases (RLK), and receptor-like proteins (RLP) in the plant *Arabidopsis thaliana*. Members of these three gene families are associated with disease resistance and their pathogen-recognition domain is a documented target of positive selection. Our bioinformatic approach to study the major family features that may influence gene conversion revealed that in these families there is a significant association between the occurrence of gene conversion and high levels of sequence similarity, close physical clustering, gene orientation, and recombination rate.

We discuss these results in the context of the overlap between gene conversion and positive selection during the evolutionary expansion of the NBS-LRR, RLK, and RLP gene families.

Introduction

Eukaryotic genomes are replete with genetic duplication, particularly plant genomes that owe their duplication, in part, to an extensive history of polyploid events (Blanc et al. 2000; Blanc, Hokamp, and Wolfe 2003; Blanc and Wolfe 2004; Adams and Wendel 2005; Lockton and Gaut 2005). For example, ~28% of existing *Arabidopsis thaliana* genes were duplicated during ancient whole-genome duplication events (The Arabidopsis Genome Initiative 2000). In addition, eukaryotic genomes contain a large fraction of genes that appear to have duplicated via single-gene events. Such events also appear to be common in plants, where, for example, up to 18% of *Arabidopsis* genes belong to tandemly arrayed gene clusters (Zhang and Gaut 2003).

Duplicated genes provide a potential template for intergenic gene conversion and unequal crossing-over, but the contribution of gene conversion to gene subfamily evolution is not clear. Traditionally, gene families have either been defined as following the “birth-and-death” or “concerted-evolution” models of evolution. In the former, gene conversion is rare and does not strongly affect phylogenetic relationships among gene family members (Zimmer et al. 1980). In the latter, gene conversion homogenizes genetic variation among family members and thus acts to retard divergence among paralogs (Smith 1973). The homogenizing effects of gene conversion appear to be relatively common, as has been demonstrated recently for tandemly duplicated genes in yeast (Gao and Innan 2004).

Nonetheless, gene conversion can also generate diversity by reassorting variants among paralogs (Slightom, Blechl, and Smithies 1980; Zimmer et al. 1980). For example, gene conversion appears to generate diversity in the major histocompatibility complex (MHC) (Weiss et al. 1983; Parham and Ohta 1996) and immunoglobulin (Ig) gene families (Ohta 1991; Huber et al. 1993) by reassorting point mutations (Flavell et al. 1986; Gyllensten et al. 1991; Parham and Ohta 1996; Ohta 1997). Furthermore, gene conversion may also “create” point mutations. In *Cf* genes from tomato, for example, Parniske et al. (1997) noted the occurrence of nonsynonymous changes at the break points of gene conversion events.

The reasortment and introduction of point mutation via gene conversion may be particularly important in gene families that have been subjected to positive (or diversifying) selection, as opposed to conservative evolution. Under positive selection, a relatively high proportion of rare genetics variants is expected to be evolutionarily successful. Indeed, the MHC, Ig, and *Cf* gene families mentioned above have been subjected to positive selection (Weiss et al. 1983; Ohta 1991; Huber et al. 1993; Parham and Ohta 1996; Parniske et al. 1997). A corollary to this point is that the record of gene conversion may be most complete (or evident) in gene families commonly subjected to positive selection because variants produced by conversion are more likely to be retained. Thus, positively selected gene families may be the best source for characterizing genome-wide patterns of gene conversion.

Here we compare the organization and history of gene conversion among three *A. thaliana* gene families that have been subjected to positive selection: the nucleotide-binding site leucine-rich repeat (NBS-LRR), receptor-like kinase (RLK), and receptor-like protein (RLP) gene families. All three gene families contain members that act as disease resistance (R) genes and mediate plant-pathogen interactions (Jones et al. 1994; Shiu and Bleecker 2001; Meyers et al. 2003). The three families also contain LRR domains, which typically form the pathogen-recognition region and appears to be the primary target of positive selection in NBS-LRR and RLK genes (Parniske et al. 1997; Meyers et al. 1998; Wang et al. 1998; Noel et al. 1999; Ellis, Dodds, and Pryor 2000; Mondragon-Palomino et al. 2002; Strain and Muse 2005).
Of these three gene families, the evolution of NBS-LRR genes has been studied most thoroughly to date. The *A. thaliana* NBS-LRR subfamily contains ~200 members, many of which are found in tightly linked clusters (Meyers et al. 1999, 2003). Initially, Michelmore and Meyers (1998) argued that gene conversion must occur infrequently in the NBS-LRR gene subfamily. They based their argument on phylogenetic evidence in which orthologs were more similar than paralogs (Michelmore and Meyers 1998). More recently, however, both Baumgarten et al. (2003) and Kuang et al. (2004) have established that gene conversion among NBS-LRR genes may be common, and a number of other studies have noted apparent intergenic or intragenic gene conversion events as well as unequal crossing-over between NBS-LRR members (Hulbert and Bennetzen 1991; McDowell et al. 1998; Ellis et al. 1999; Noel et al. 1999; Chin et al. 2001; Sun et al. 2001; Baumgarten et al. 2003).

The two other gene families in this study—RLKs and RLPs—encode transmembrane receptors with a series of solvent-exposed LRRs and a transmembrane region. RLP genes encode transmembrane receptors that consist of 25–38 LRRs. Here we focus on members of the RLP family genes encode transmembrane receptors that consist of solvent-exposed LRRs and a transmembrane region. RLP—encode transmembrane receptors with a series of 2001; Baumgarten et al. 2003). The LRR genes has been studied most thoroughly to date.

The purpose of this article is to examine patterns of gene conversion (or unequal crossing-over) in the context of the genomic organization of the NBS-LRR, RLK, and RLP gene families. We use *A. thaliana* sequence data to characterize patterns of gene conversion and ask: Can gene conversion events be identified? If so, does gene conversion primarily take place within gene clusters, or does gene conversion extend to physically separated genes? Further, do these events extend to noncoding genic regions? Are there any genome-scale patterns, such as recombination rate, that correlate with patterns of gene conversion? And finally, is there any evidence that gene conversion is detected more often in genes where positive selection has occurred?

**Materials and Methods**

**Sequence Data**

The members of NBS-LRR and RLK families were obtained in May 2002 from the database of *A. thaliana* NBS-LRR—encoding disease resistance gene homologs (http://niblrs.ucdavis.edu/At_RGenes/) and the Plant Receptor Kinase Resource (http://www.botany.wisc.edu/prkr/), respectively. To identify members of the RLP subfamily from *A. thaliana*, we queried the *Arabidopsis* proteome, based on a May 11, 2002 download, with BlastP from Blast version 2.2.3 (Altschul et al. 1990). The queries were the amino acid sequences of experimentally characterized RLPs Cf-2 (U42445); the Cf-9 resistance gene cluster (AJ002236) that includes Cf-9, Hcr9-9A, Hcr9-9B, Hcr9-9D, and Hcr9-9E from *Lycopersicon pimpinellifolium*; the Cf-4 resistance gene cluster (AJ002235) that contains Cf-4, Hcr9-4C, Hcr9-4A, Hcr9-4B, and Hcr9-4E from *Lycopersicon hirsutum*; and Cf-5 (AF053993) from *Lycopersicon esculentum*. We first retrieved all hits with a cutoff e value of >10−5. These hits were considered as part of the RLP subfamily if they came from complete genes, had been annotated as similar to members of the *Lycopersicon Cf* subfamily, or if LRRs had been detected. For all three gene families, we retrieved amino acid and nucleotide sequences as well as annotation from the MIPS *Arabidopsis thaliana* Database (http://mips.gsf.de/proj/thal/db/index.html).

**Division of Groups in Subfamilies and Clusters**

We divided each gene family into clusters and subfamilies. The clusters represent physical proximity on the chromosome; within each family, a cluster was defined as two or more homologous genes that occur within a maximum of eight open reading frames (ORFs) between any two genes (Richly, Kurth, and Leister 2002; Meyers et al. 2003). Meyers et al. (2003) also employed this definition for identifying clusters in the NBS-LRR family and demonstrated that the numbers and sizes of the groups did not change substantially when the maximum number of intervening ORFs was increased up to 50 (Meyers et al. 2003).

Following previous definitions, we defined a subfamily of genes as a group of two or more homologous proteins that are at least 50% identical at both the amino acid and nucleotide levels (Dayhoff 1978). In each family, we detected subfamilies by using BlastP (Altschul et al. 1990) to make pairwise comparisons of each member at the amino acid level. Each sequence was aligned with ClustalW (Thompson, Higgins, and Gibson 1994) to all its Blast hits with an e value of zero or the group of Blast hits with the smallest nonzero value. Next, these preliminary alignments were used to calculate the matrix of mean number of pairwise differences (adjusted for missing data) with PAUP* version 4.0b10 (Swofford 2000). Sequences that were >50% different from the other members of each alignment were removed, and the remaining sequences were realigned. This grouping process finished when all members from a family were assigned to a subfamily or classified as singletons. The amino acid alignments of clusters and subfamilies were visually inspected and employed to generate the nucleotide alignments required to detect gene conversion. Alignments are available at http://gautlab.bio.uci.edu/data.htm.

To verify observations based on coding regions, we also obtained alignments of the corresponding noncoding regions of the genes in each cluster or subfamily. To this end, we identified and individually aligned introns and 500-bp regions 5′ and 3′ from the coding region using the default settings of ClustalW.
RLP Family Characterization

For both the NBS-LRR and RLK gene families, there are detailed descriptions of their members in *A. thaliana* and several other species (Shiu and Bleecker 2001a, 2001b; Cannon et al. 2002; Meyers et al. 2003). Similar reports are not as detailed for the RLP family. In order to better understand the organization and characteristics of the RLP family, we analyzed the evolutionary relationships, genomic arrangement, and occurrence of positive selection. The amino acid sequences of the 48 members of this family were aligned employing standard settings in ClustalW (Thompson, Higgins, and Gibson 1994). The resulting alignment was manually edited and employed to generate the alignment of the corresponding nucleotide sequences. With this alignment, we obtained the neighboring-joining phylogeny based on GTR + I distances, as implemented in PAUP* version 4.0b10 (Swofford 2000). The reliability of this phylogeny was evaluated with 10,000 bootstrap replicates.

We employed the computer program codeml from PAML version 3.14 and analyzed subfamily groups with more than three sequences to calculate \( P \), which is the ratio of the nonsynonymous to synonymous distances (Yang 1997; Yang et al. 2000). We detected variation in \( P \) among sites by employing a likelihood ratio test between M7 and M8, M0 and M3, and M1 and M2 (Yang 1997; Yang et al. 2000). The phylogeny for each group was obtained from PAUP* neighbor-joining searches. With codeml, we employed complete deletion when analyzing alignments with gaps. In order to preserve the largest amount of data, we eliminated from every alignment sequences that contained 40% of their length or more as indels. We also analyzed \( P \) in the RLP subfamily with the online implementation of program HYPHY version 0.99 beta (Kosakovsky Pond and Frost 2005; Kosakovsky Pond, Frost, and Muse 2005). We applied the SLAC, FEL, and REL analyses; all analyses employed the HKY85 model of codon substitution. Results were considered significant at a \( P \) value \( \leq 0.05 \).

Detection of Gene Conversion

We detected gene conversion with the methods implemented in GENECONV version 1.18 (Sawyer 1999). The program was applied to cluster-based alignments and subfamily-based alignments separately. The program tests for gene conversion by finding identical fragments between pairs of sequences in a nucleotide alignment. Global and pairwise \( P \) values are calculated in order to assess the statistical significance of the observed fragment lengths. Pairwise \( P \) values are assigned based on the comparison of each fragment with the maximum fragment length that is expected from the sequence pair by chance (Sawyer 1999). Global \( P \) values are more conservative because they are based on the comparison of each fragment with all possible fragments for the entire alignment. The \( P \) values for global fragments are based on 10,000 permutations of the sequence alignment, while pairwise fragment \( P \) values are obtained by the method of Karlin and Altschul (1990, 1993). For the analysis of alignments of three or more sequences, we ran GENECONV with default settings. In order to analyze families and clusters consisting of two sequences, we adjusted the program setting to consider homomorphic sites (Sawyer 1999; Drouin 2002). Following previous publications (e.g., Drouin 2002), we considered for further characterization the conversion tracts detected by GENECONV with a global \( P \) value \( \leq 0.05 \). Although the experiment-wide type I error was therefore large, this approach was appropriate because we were interested in the outcome of the test for each alignment (Rothman 1990). GENECONV cannot distinguish between a gene conversion event and an unequal crossing-over event. Hereafter, we will refer to significant GENECONV results as conversion events or tracts.

Examining Factors That Influence the Occurrence of Gene Conversion

We calculated the average identity and physical distance (in base pairs) at both the coding and noncoding levels between pairs of genes of all clusters. To determine whether distance and average identity varied between converted and nonconverted genes, we calculated the distance or identity between converted gene pairs and compared the observed average to the distribution of averages based on 10,000 rounds of resampling from all converted or nonconverted pairwise distance or identity values of each family.

Similarly, we assessed the correlation of recombination rate with the occurrence of gene conversion in the clusters of each family. The genes we considered in this analysis belonged to cluster groups where conversion was detected at both coding and noncoding levels. For each gene, we calculated the rate of recombination by using the physical and genetic positions for each chromosome of *A. thaliana* from The Arabidopsis Information Resource and the Nottingham Arabidopsis Stock Centre Recombinant Inbred recombination map (http://www.arabidopsis.org; http://arabidopsis.info/new_ri_map.html). The rates of crossing-over were estimated by using the derivative of the third-order polynomial that describes the relationship between physical and genetic distance (Wright, Agrawal, and Bureau 2003).

Results

RLP Organization and a History of Positive Selection

The evolution of the disease-related NBS-LRR and RLK genes has been studied in depth, particularly with respect to positive selection (Mondragon-Palomino 2002; Shiu et al. 2004; Strain and Muse 2005). Although positive selection has been detected in RLP *Cf* genes from tomato (Parniske et al. 1997; Thomas et al. 1997), no study has characterized positive selection in RLP genes in other species. To better describe characteristics of the *A. thaliana* RLP family and to bring our level of knowledge on par with the other two gene families, we estimated the phylogeny of RLP genes (fig. 1). The clades on the phylogeny are defined to contain members with 50% or greater nucleotide sequence identity. Most clades in the RLP phylogeny contain tandemly duplicated genes as well as genes from two or more chromosomes, suggesting that the gene subfamily recently expanded from a few ancestral lineages by frequent duplication events and perhaps genomic rearrangement (fig.
1). However, caution must be exercised when interpreting the phylogeny of a gene subfamily as gene conversion potentially alters relationships among paralogs.

To investigate positive selection in RLP genes, we calculated the \( \omega \) ratio. Using codeml from the PAML package, we detected significant evidence of positive selection in all the five RLP subfamilies that have more than two sequences (fig. 1 and Supplementary Table 1, Supplementary Material online). A total of 75% of gene family members are in these groups. Because the PAML method may overestimate the extent of positive selection, we also assessed positive selection with the SLAC, FEL, and REL routines from HYPHY (Kosakovsky Pond and Frost 2005). Although potentially subject to some of the same artifacts, the results from HYPHY analyses, FEL and REL, confirm the occurrence of positive selection in all RLP subfamily groups with more than two sequences (Supplementary Table 2, Supplementary Material online). Similar analyses identified positive selection in groups comprising up to 51% of the NBS-LRR gene family (Mondragon-Palomino et al. 2002) and ~24% of the RLK family (Strain and Muse 2005).

Organization of NBS-LRR, RLP, and RLK Gene Families

In order to distinguish the influence of identity and physical distance in gene conversion, we divided each family into clusters, which are defined as genes that are not separated by more than eight predicted genes (Richly, Kurth, and Leister 2002; Meyers et al. 2003), and subfamilies, which include genes with pairwise identity of 50% or more. Note that it is possible that gene(s) are both highly similar and physically close and thus are members of both a cluster and a subfamily.

NBS-LRR and RLP genes tend to be both highly similar and physically clustered, as reflected by the fact that 54% and 69% of the NBS-LRR and RLP genes belong
Table 1
Distribution of Family Members into Clusters and Subfamilies

<table>
<thead>
<tr>
<th></th>
<th>NBS-LRR</th>
<th></th>
<th>RLK</th>
<th></th>
<th>RLP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Genes</td>
<td>Proportion</td>
<td>Number of Genes</td>
<td>Proportion</td>
<td>Number of Genes</td>
<td>Proportion</td>
</tr>
<tr>
<td>Genes in subfamilies and clusters</td>
<td>99</td>
<td>0.54</td>
<td>56</td>
<td>0.24</td>
<td>33</td>
<td>0.69</td>
</tr>
<tr>
<td>Genes only in clusters</td>
<td>32</td>
<td>0.17</td>
<td>6</td>
<td>0.02</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Genes only in subfamilies</td>
<td>41</td>
<td>0.22</td>
<td>116</td>
<td>0.50</td>
<td>10</td>
<td>0.21</td>
</tr>
<tr>
<td>Genes neither in clusters nor subfamilies</td>
<td>13</td>
<td>0.07</td>
<td>55</td>
<td>0.24</td>
<td>5</td>
<td>0.10</td>
</tr>
<tr>
<td>Total</td>
<td>185</td>
<td></td>
<td>233</td>
<td></td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Median Size and Range (between parenthesis) of Conversion Tracts

<table>
<thead>
<tr>
<th></th>
<th>Clusters</th>
<th></th>
<th>Clusters Noncoding</th>
<th></th>
<th>Subfamilies</th>
<th></th>
<th>Subfamilies Noncoding</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NBS-LRR</td>
<td>79 (7–1491)</td>
<td>56 (7–1491)</td>
<td>82.5 (15–1873)</td>
<td>87 (18–1560)</td>
<td>32 (10–408)</td>
<td>29 (10–408)</td>
<td>24 (6–377)</td>
<td>23 (6–377)</td>
</tr>
<tr>
<td>All gene families</td>
<td>68</td>
<td>47.5</td>
<td>69</td>
<td>73</td>
<td>29</td>
<td>23.5</td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

Note.—The second column in each category corresponds to the results excluding groups with two sequences.
GENECONV, so only continuous regions of significant identity and length were considered conversion tracts. While relatively fewer events and converted genes were detected with noncoding data, the general picture was similar between noncoding and coding data—i.e., most conversion occurs between clustered genes that are \( \geq 50\% \) identical (fig. 2), but some conversion events between nonclustered genes are evident.

The second concern of GENECONV analysis is its performance on alignments consisting of only two sequences. GENECONV can detect gene conversion between two sequences by including monomorphic sites (Sawyer 1999), but the detection rate appears to be aberrantly high. To eliminate this bias, we reassessed the occurrence of gene conversion after eliminating all results coming from clusters or subfamilies with less than three members (table 2 and fig. 3). We should note that clustered \emph{Arabidopsis} genes are predominantly found in groups of two (Zhang and Gaut 2003), and thus removing alignments of only two sequences increases the proportion of genes and conversion events that can be assigned to subfamilies relative to clusters (fig. 3). Nonetheless, GENECONV results are similar in two respects (fig. 3). First, gene conversion is detected commonly. In coding regions, conversion is detected in \( \geq 20\% \) of genes whether the genes are in cluster-based or subfamily-based alignments. Second, the highest proportion of genes implicated in gene conversion events is again within the category of genes that are both clustered and \( \geq 50\% \) identical.

Factors Influencing the Detection of Gene Conversion in Clusters

We compared the characteristics of converted genes against those in which we did not detect conversion. The following analyses were based on clustered genes, both because assignment to clusters is not biased by a priori limits of sequence identity and because physical distances between the genes in each cluster can be calculated.

Pairwise Identity

We calculated the level of pairwise identity and compared the averages of genes where we found conversion with the distribution of average identities calculated by
resampling. For all three gene families at the coding and noncoding levels, the degree of identity is significantly higher between the genes where there is conversion (table 3) and holds for most comparisons after eliminating results from clusters of two genes (Supplementary Table 3, Supplementary Material online). Note, however, that the statistical power to detect gene conversion with GENECONV is also a function of sequence identity, and it is thus impossible to disentangle the effect of identity on gene conversion from its effect on statistical power. Nonetheless, these results are consistent with assertions that gene conversion is favored between genes with high sequence similarity (Borts and Haber 1987; Modrich and Lahue 1996).

Physical Distance

In general, our analyses reveal that for all gene families a pairwise distance of 40 kbp is the upper limit for conversion to occur between clustered genes on the same chromosome, at either coding or noncoding levels (data not shown). For RLK coding regions, the average distance between converted gene pairs is significantly lower relative to distances between nonconverted gene pairs both when groups consisting of two genes are included in analysis (table 3) and when they are not (Supplementary Table 3, Supplementary Material online). A similar trend, but less significant, is also evident in the NBS-LRR family (Table 3).

Table 3
Degree of Pairwise Identity and Physical Distance Among Converted and Nonconverted Genes in Clusters, with Groups of Two Genes Included in the Analysis

<table>
<thead>
<tr>
<th></th>
<th>Coding</th>
<th>Noncoding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Conversion</td>
<td>Conversion</td>
</tr>
<tr>
<td>Degree of sequence identity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBS-LRR</td>
<td>0.58 ± 0.17</td>
<td>0.73 ± 0.17</td>
</tr>
<tr>
<td>RLK</td>
<td>0.66 ± 0.10</td>
<td>0.71 ± 0.20</td>
</tr>
<tr>
<td>RLP</td>
<td>0.72 ± 0.10</td>
<td>0.79 ± 0.06</td>
</tr>
<tr>
<td>Physical distance(^b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBS-LRR</td>
<td>12.15 ± 11.18</td>
<td>15.42 ± 22.18</td>
</tr>
<tr>
<td>RLK</td>
<td>21.06 ± 18.65</td>
<td>8.03 ± 8.18</td>
</tr>
<tr>
<td>RLP</td>
<td>9.14 ± 8.53</td>
<td>12.27 ± 12.23</td>
</tr>
</tbody>
</table>

\(^a\) P value obtained from 10,000 resampling replicates.

\(^b\) Distance units are kilobase pairs.
and Supplementary Table 3, Supplementary Material online).

Gene Orientation

Genes located on the same DNA strand have direct orientation, while genes on different strands have opposite orientations. Direct gene orientation between recombining genes facilitates strand mispairings that yield viable conversion events (Hulbert et al. 2001), suggesting that genes in direct orientation should undergo gene conversion more often. We counted the pairs of converted and nonconverted genes and the number of different cluster gene pairs with or without matching strand orientation for each of the three gene families. We employed these numbers to calculate the probability that a given pair of genes is in the same or different orientation and has also converted, and obtained the expected frequencies of matching orientations in converted genes. Then we compared these frequencies with the observed number of pairs that are located on the same or different strand and where conversion occurred or not. In the NBS-LRR family, the number of converted genes located on the same strand was significantly higher than expected by chance (36 converted gene pairs on the same strand; 10 pairs on opposite strands; \( P = 0.006, \chi^2_{df=1} \)). In RLK and RLP clusters, there was a tendency toward conversion on the same strand, but the tendency was not significant in either case (RLK, 12 gene pairs on the same strand, 5 on the opposite strand, \( P = 0.233; \) RLP, 11 pairs on the same strand, 6 on opposite strand, \( P = 0.442, \chi^2_{df=1} \)), perhaps reflecting low statistical power with a small number of observations. Summing over gene families, the trend is highly significant (\( P < 0.001 \)) in favor of conversion among genes in the same orientation.

Rate of Recombination

Gene Conversion in Leucine-Rich Repeat Families 2451

Table 4

<table>
<thead>
<tr>
<th>Average Rates of Recombination in Gene Clusters</th>
<th>Conversion</th>
<th>No Conversion</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBS-LRR</td>
<td>4.69 ± 1.03</td>
<td>4.36 ± 0.66</td>
<td>0.10</td>
</tr>
<tr>
<td>RLK</td>
<td>4.45 ± 0.22</td>
<td>4.16 ± 0.87</td>
<td>0.20</td>
</tr>
<tr>
<td>RLP</td>
<td>5.61 ± 1.92</td>
<td>4.73 ± 0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>All gene families</td>
<td>4.82 ± 1.22</td>
<td>4.35 ± 0.70</td>
<td>0.007</td>
</tr>
</tbody>
</table>

* Units are centimorgans per megabase pairs.
* \( P \) values were obtained with a two-tailed, heteroscedastic \( t \)-test.

Interchromosomal Conversion and Chromosomal Duplication

Gene conversion and unequal crossing-over between genes in homologous chromosomes take place at a higher rate than between nonhomologous chromosomes (Petes and Hill 1988). However, we detected several potential instances of gene conversion that involved genes on different chromosomes, even when we counted only those events for which there is evidence of conversion in both coding and noncoding regions. In the RLK family, for example, three out of seven converted pairs common to coding and noncoding regions are exchanged between different chromosomes. Similarly, in the NBS-LRR and RLP families, we detected one of eight and one of two interchromosomal gene pairs, respectively. Baumgarten et al. (2003) also detected interchromosomal conversion events in the NBS-LRR gene subfamily. Their explanation for these observations is that gene conversion could conceivably take place between genes in homologous chromosomal segments that duplicated as a result of ancient polyploid events. As the genome reacquired diploid status, large-scale chromosomal rearrangements translocated genes that were once in meiotically paired, homologous chromosomes. These rearrangements can make ancient homologous recombination events appear as transchromosomal exchanges (Baumgarten et al. 2003).

We investigated whether this scenario explains our results by determining whether a pair of genes that participate in a conversion event belong to corresponding duplicated segments. We employed the definitions of duplicated chromosomal segments and gene content generated by Blanc, Hokamp, and Wolfe (2003). Using these genes in our queries, we found that none of the NBS-LRR and RLP queried pairs belongs to a duplicated block. In contrast, the six RLK family members under conversion between different chromosomes were contained in duplicated blocks. Thus, interchromosomal conversion in the RLK family can be explained by exchanges between homologous chromosomes during polyploidy or ensuing rediploidization. This does not appear to be the case for NBS-LRR or RLP interchromosomal events, but we cannot exclude the possibility that they are the result of smaller chromosomal rearrangements that are not detected as duplicated blocks.

Discussion

In this study, we divided the members of three families into clusters and subfamilies with the purpose of separating to some degree the influence of physical proximity and identity. This approach revealed that the NBS-LRR and
RLP gene families, which are organized into clusters of highly similar genes, differ significantly from that of the RLK family, which is typified by genes dispersed throughout the genome (table 1). As expected from their organization, the majority of detectable conversion events in the NBS-LRR and RLP gene families take place between highly similar, tightly clustered genes, with relatively little evidence for gene conversion among more broadly dispersed sequences (fig. 2). The RLK family is similar in that most conversion events occur between clustered genes, but there is also an appreciable amount of conversion among dispersed but similar (>50% identical) sequences. These general patterns hold for both coding and noncoding regions (fig. 2) and also when gene pairs are removed from conversion analysis (fig. 3).

For the three gene families, most detectable conversion events are between genes that are 60%–70% identical, but it is important to consider that most gene conversion events may not be detectable by GENECONV at higher levels of identity. Additionally, GENECONV is biased toward detecting recently converted regions that have not yet changed by mutation. As an alternative to GENECONV, we applied LIKEWIND (Archibald and Roger 2002), a procedure to detect conversion based on the comparison of the likelihood of phylogenetic relationships obtained from sliding windows along an alignment (data not shown). This analysis confirmed the occurrence of gene conversion in the clusters and families where we detected conversion with GENECONV. Phylogeny-based approaches like LIKEWIND are limited to analyzing sequence alignments of four to eight sequences and also do not assign conversion events to particular genes. Nonetheless, an alternative phylogenetic approach is generally in agreement with our GENECONV analysis.

The characteristics of conversion events that we detected agree with previous genome-level studies in Caenorhabditis elegans and yeast. These surveys found that gene conversion occurs more frequently in highly identical genes localized in the same chromosome (Semple and Wolfe 1999; Drouin 2002). For example, in C. elegans, Semple and Wolfe (1999) observed that 78% of the conversion events take place between tandemly duplicated members of gene families. The frequency of conversion is negatively correlated with distance between gene pairs (Drouin 2002), so an excess of intrachromosomal conversion events is observed in relation with the actual number of duplicated genes that are found in the same or different chromosomes (Semple and Wolfe 1999). This also has been observed in yeast, where the frequency of gene conversion events is 6.3% between unlinked genes but is 60% and 80% between genes that are less than four ORFs away or next to each other, respectively (Drouin 2002). Another point of agreement with previous genomic studies is the wide range in the length of conversion tracts and the bias toward short sizes in the distribution of tract lengths (Semple and Wolfe 1999). For example, the length of the conversion events we detected in coding regions of NBS-LRR clusters ranges from 7 to 1,491 bp, and 60% of the tracts are <100 bp long. In C. elegans and yeast the length range spans 12–2,958 and 8–1,181 bp, respectively (Semple and Wolfe 1999; Drouin 2002). In plants, conversion events in the RGC2 locus of Lactuca sativa appeared to span from 60 to 528 bp (Kuang et al. 2004), similar to the values we have found here (table 2).

In addition to location and sequence similarity, our analyses uncovered two other correlates with gene conversion: gene orientation and recombination rate. The relative impact of gene orientation in gene conversion has been highlighted by observations on NBS-LRR genes from rice. Hulbert et al. (2001) found that linked NBS-LRR genes in opposite orientation exhibit a higher level of divergence (Hulbert et al. 2001), and they suggested that gene families arranged as direct repeats are more often free to homogenize via unequal recombination. Further, the products of crossovers between inverted repeats may lead to dicentric and acentric chromosomes that often cannot be transmitted to the next generation (Hulbert et al. 2001). Consistent with these ideas, we observe that conversion in the NBS-LRR family takes place significantly more often between genes that are in the same orientation. This same trend is evident, but not significant, in the RLK and RLP families.

The correlation between gene conversion events and recombination rate (table 4) is, as far as we are aware, yet unreported. Previously, it has been shown that the distribution of clusters or tandemly arrayed genes in Arabidopsis is biased toward high-recombination regions of the genome (Zhang and Gaut 2003). Two factors may contribute to this distribution. First, the biased distribution could be a function of the interacting forces of selection, recombination, and gene copy number (Zhang and Gaut 2003). Second, the distribution of clusters in high-recombination regions could be due to higher rates of cluster production, suggesting that rates of unequal recombination events is related to the homologous recombination rate. Because GENECONV potentially detects both gene conversion and unequal recombination events, this latter model is consistent with our analyses. However, an alternative explanation for more detectable gene conversion in regions of high recombination may simply be that there are more gene clusters in high-recombination regions. At present, we cannot determine if recombination is causative or correlative.

One issue that requires greater clarification is the mechanisms that contribute to interchromosomal conversion events. We have detected interchromosomal events in all three gene families, even when conversion between genes is subjected to the relatively strict criterion that conversion must be evident in both coding and noncoding regions of a gene. Although they used very different sequence alignments (based on different definitions of clusters and subfamilies), Baumgarten et al. (2003) also documented interchromosomal conversion events among NBS-LRR genes. A point where our results diverge with those in Baumgarten et al. (2003) is that they found that intergenic exchanges between NBS-LRR sequences on different chromosomes occurred mostly in genes that were located within duplicated genomic regions. These authors suggested that gene conversion events between genes in different chromosomes were the outcome of intergenic recombination previous to segmental duplication and rearrangement to unlinked genomic locations. Our analyses do not support this conclusion. Although we detected gene conversion between genes in different chromosomes, we
did not detect any instance of gene conversion between two NBS-LRR genes that belong to duplicated blocks. Differences between our study and Baumgarten et al. (2003) could be ascribed to different definitions of “duplicated genomic regions”—i.e., the methods that Blanc, Hokamp, and Wolfe (2003) employed to define a duplicated block differs from the approaches employed by Baumgarten et al. (2003). Further, it is not clear that one needs to invoke historical conversion prior to rearrangement. For example, it has been hypothesized that translocation of genes between clusters on the same chromosome drives the observed recombination events between genes in separate clusters of RLP Cf paralogs (Parniske and Jones 1999).

Gene Conversion and the Evolution of the NBS-LRR Family

Of the three families, the evolutionary dynamics of the NBS-LRR subfamily have received the most attention. Michelmore and Meyers (1998) initially argued that the NBS-LRR family follows a process of diversifying evolution in a birth-and-death pattern. These authors based their line of reasoning on phylogenetic analyses indicating that orthologs are more similar than paralogs in several groups of tomato and lettuce resistance genes. They did not reject the notion that gene conversion occurs in NBS-LRR genes, which has been verified by many observations (Hulbert and Bennetzen 1991; McDowell et al. 1998; Ellis et al. 1999; Noel et al. 1999; Chin et al. 2001; Sun et al. 2001), but argued that the low rate of unequal crossing-over and gene conversion between paralogs precludes either significant homogenization among paralogs or generation of new specificities (Michelmore and Meyers 1998).

More recently, Kuang and coauthors have hypothesized that there are two types of NBS-LRR genes, denoted by type I and type II (Kuang et al. 2004). Based on phylogenetic arguments, type I genes appear to be subjected to frequent conversion events whereas type II genes are not. Although all of the sequences from Kuang et al. (2004) are related to NBS-LRR genes clustered in the RGC2 locus of lettuce, most sequences were isolated by polymerase chain reaction, and it was thus unclear whether they originated from gene clusters or dispersed genes. Our results for each of the three families suggest that clustered, similar sequences are far more prone to conversion, suggesting that type I (converting) genes are likely located in clusters. In contrast, the type II (nonconverting) genes may be dispersed.

NBS-LRR genes are often characterized as two classes—the TIR and non-TIR classes—representing an ancient division that dates back to the common ancestor of angiosperms and gymnosperms (~300 MYA) (Meyers et al. 1999; Cannon et al. 2002). Based on Arabidopsis data, it was thought that phylogenies of TIR and non-TIR lineages have different properties. In the TIR subfamily branch lengths are comparatively short, but the non-TIR subfamily is divided into deep clades that contain sequences of several plant taxonomic families (Cannon et al. 2002; Meyers et al. 2003). These observations suggested that the non-TIR clades might be more ancient (Meyers et al. 2003) and perhaps less prone to gene conversion. Our results agree: we find that only 14% of NBS-LRR gene conversion events occur in the non-TIR Arabidopsis genes. In contrast, the lettuce RGC2s are non-TIR genes. It thus appears that the evolution of lettuce and Arabidopsis non-TIR genes differs, or perhaps additional data are required for further characterization of Arabidopsis non-TIR conversion events.

Gene Conversion and Positive Selection

The motivation of this work was to study gene conversion. The NBS-LRR, RLP, and RLK gene families were chosen because they have been subjected to positive selection and are thus likely to have retained evidence of gene conversion that either reassorted or introduced variation. This leads to an obvious question: Do members of the gene families that exhibit a history of gene conversion also exhibit a history of positive selection?

Unfortunately, the answer is not straightforward for two reasons. The first reason is methodological. The standard method to detect positive selection—the codeml program of the PAML package—is based on a phylogenetic hypothesis. However, a single phylogeny does not adequately describe relationship among genes that have experienced gene conversion and recombination. Further, the PAML method may overestimate the prevalence of positive selection when there is recombination (or gene conversion) among sequences (Anisimova, Nielsen, and Yang 2003; Suzuki and Nei 2004; Wong et al. 2004; Zhang 2004). When positive selection is correctly inferred, it is easier to pinpoint groups of selected sequences (not all of which may have been subjected to positive selection) than to identify a single selected sequence. As a consequence of these considerations, the correlation between positive selection and gene conversion will be tenuous, at best.

The second reason is biological. Numerous studies have documented that disease resistance genes have been subjected to positive selection (Meyers et al. 1998; Wang et al. 1998; Bishop, Dean, and Mitchell-Olde 2000; Bergelson et al. 2001; Mondragon-Palomino et al. 2002). Some of these studies are based on codeml analyses but most are not, and the overwhelming consensus is that positive selection on R genes is widespread. Here we show that gene conversion is also widespread; based on GENECONV analysis, it has occurred in >30%–70% of family members. The extent of positive selection and gene conversion in these families makes overlap inevitable. There are, however, some compelling examples. One group of seven TIR NBS-LRR genes includes 52% of conversion tracts in the family. Its members also belong to one of the few TIR groups where there is strong evidence of positive selection (Mondragon-Palomino et al. 2002), suggesting that gene conversion and positive selection may be tightly coupled. This group includes characterized disease resistance gene At4g16950, the putative ortholog of RPP5 from cultivar Landsberg, together with At1g16860, At4g16890, At4g16900, At4g16920, At4g16940, and At4g16960. Further support to the co-occurrence of gene conversion and positive selection in NBS-LRR genes comes from experimental observations in type I RCB2 genes from lettuce, which are subject to frequent gene conversion and also to positive selection (Kuang et al. 2004).
Nonetheless, the identification of gene conversion and positive selection are not completely linked. For example, ~70% of the RLP genes contain evidence of gene conversion (fig. 2), and up to 75% of the family is subject to positive selection (fig. 1). The overlap in distribution between these two processes necessarily includes most of the genes in the RLP family. Nonetheless, gene conversion operates heterogeneously throughout the phylogeny of RLP genes. While some genes from clades 2, 3 and 7 are involved in conversion, all members in clade 6 exchange regions with each other (fig. 1). On the other hand, in five out of six members of clade 5, gene conversion does not take place at any level, but this clade has evidence of positive selection (fig. 1). The loose linkage between gene conversion and positive selection also extends to the NBS-LRR: most of the NBS-LRR sequences where we found positive selection also extends to the NBS-LRR: most of the families. Gene conversion events were evident along the phylogeny of RLP genes, respectively, because each of them has the NBS-LRR and RLK genes tends to occur in the LRR regions (Parniske et al. 1997; Meyers et al. 1998; Wang et al. 1998; Noel et al. 1999; Ellis, Dodds, and Pryor 2000; Mondragon-Palomino et al. 2002; Strain and Muse 2005). Are detectable gene conversion events also biased to the LRR? To address this question, we assessed the distribution of gene conversion tracts in the coding and noncoding regions of three groups of genes with extensive evidence of gene conversion. We chose subfamilies FAT4G16860, FAT2G13790, and FAT4G13810 (the last is subgroup 6 in fig. 1) from the NBS-LRR, RLK, and RLP families, respectively, because each of them has the largest or the longest conversion events in their respective families. Gene conversion events were evident along the length of the sequences, and thus we were unable to establish a bias in the genic location of gene conversion (data not shown).

Parniske et al. (1997) proposed a mechanism that explains the co-occurrence of gene conversion and positive selection in some clusters of disease resistance genes. In Cj genes, there are several instances in which the recombination break point coincides with a hypervariable codon (Parniske et al. 1997). This suggests a link between gene conversion and point mutations in positions under positive selection in the RLP family. However, while we have shown that the RLP, RLK, and NBS-LRR families have an extensive history of gene conversion, our data cannot establish a causal link between the two processes.

Supplementary Material

Supplementary Tables 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank R. Bush and W. Fitch for comments on a previous version of the manuscript and two anonymous reviewers for constructive critique and suggestions. M.M.P. was supported by a graduate fellowship from The National Council of Science and Technology, Mexico and the University of California Institute for Mexico and the United States. This work was supported in part by National Science Foundation grants DEB-0316157 and DBI-0321467 to B.S.G.

Literature Cited


Sawyer, S. A. 1999. GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University, St. Louis.


Spencer V. Muse, Associate Editor

Accepted August 4, 2005