INTRODUCTION

Transposable elements, or “jumping genes,” are the most abundant entity of many eukaryotic genomes, including that of humans, and have played a major role in their evolution and expression. Helitrons are a recently discovered superfamily of eukaryotic transposable elements. Despite a high variability in both sequence and length, Helitrons share conserved sequences of fifteen to twenty base pairs at the 5’ end and of about thirty-five base pairs at the 3’ end. Helitrons are highly abundant; the consensus hypothesis is that they transpose via a rolling circle mechanism. They frequently capture gene sequences and multiply them in different regions of the genome. The mechanism and biological relevance of gene capture remains undetermined. Captured genes may be transcribed, giving rise to chimerical transcripts containing coding regions of several different genes. This transcription may form a driving mechanism for the evolution of new genes within a genome. Many of the interpretations of the role of Helitrons are still hypothetical in part because no active organism has been observed in any organism thus far.\(^1\)

The objective of this project is to implement computational approaches to discover Helitrons in the maize genome and determine the structure of these Helitrons, including the genes captured by these elements. These methods provide clues to the mechanism of gene capture and attempt to identify captured genes that are potentially transcribed. In identifying these elements, we may shed light on the possible biological function of Helitrons in an organism’s genome. We propose to provide experimental evidence to validate the efficacy of computational searches, and to compare the role of these Helitrons across several strains of maize.
One putative *Helitron* identified by computational means as potentially containing a full-length transcribed gene was confirmed as a *Helitron* using plus/minus polymorphism through PCR. The *Helitron* was also evaluated for a paralogous sequence, and the presence of the gene of interest was confirmed in several maize inbred lines, though not necessarily within the *Helitron*. RT-PCR was used to test the expression of the *Helitron*-captured gene and its wildtype gene analog in the maize inbred lines. Using this molecular analysis of the potential *Helitron*, we demonstrated the validity of our computational methods as well as contributed an interesting case of gene capture to existing *Helitron* literature.

**BACKGROUND**

*Helitrons* are a unique, recently discovered class of transposable elements. They are known to capture gene fragments and move them within the genome. To date, *Helitrons* have been found in several eukaryotic species including: *Arabidopsis*, rice, maize, morning glory, *C. elegans*, and the brown bat.ii

*Helitrons* are postulated to replicate via a rolling-circle mechanism, and have characteristic, conserved 3’ and 5’ ends as well as a palindrome sequence near the 3’ end which forms a hairpin structure. *Helitrons* vary in length from 100bp-25kb and insert between adenine and thymine bases.

*Helitrons* are of interest due to their potential contribution to a mechanism of evolution—the potential to capture and move gene fragments could lead to the creation of novel genes. Additionally, through a “copy and paste” mechanism, a *Helitron* could potentially increase gene dosage and insert full length genes by copying it elsewhere in the genome, thus contributing to heterosis, or hybrid vigor. Although it is clear that *Helitrons* capture gene fragments, there has not yet been a documented case of a full-gene capture.

Although *Helitrons* apparently have an important role in the genome of many eukaryotes, little is known about the mechanism of gene capture. In order to characterize the biological mechanisms of *Helitrons*, our objective focused on creating a better way of locating *Helitrons* within the genome. Although *Helitron*Finder (Du & Dooner)iii is a program designed to
find Helitrons, we found that some of its specifications were not stringent enough, while others did not allow for enough flexibility. We developed the program HelRaizer to discover high-quality Helitrons. High-quality Helitrons would be those which have captured many gene fragments, or even full genes.

In order to verify the Helitrons discovered by HelRaizer, we performed molecular analysis. To do this, we investigated one specific case of potential full-gene capture in maize. We used maize as our model organism. The maize genome is completely sequenced and available in an online database through NCBI. This, combined with the availability of ten inbred lines makes it an ideal organism to study Helitrons.

**COMPUTATIONAL MATERIALS AND METHODS**

The sequences of 21 experimentally determined Helitrons were examined and their 5’ and 3’ termini were isolated. These Helitron termini were aligned and a strict consensus pattern of nucleotides was composed as a template from which the maize genome could be queried. An example alignment and resulting consensus can be seen in Figures 1 and 2.

The entire database of Zea mays BACs was downloaded from the PlantGDB project’s website. A script was written in the Python programming language using modules from the BioPython project to identify putative Helitrons contained in the Zea mays BAC sequences. This program batch processed the BAC sequences and analyzed for pattern matching against the Helitron termini template using regular expression processing. When a matching putative terminus was found, the terminus type and location were recorded for further analysis. Upon completion of the putative termini identification, the results were analyzed, and areas where 5’ termini were located within 100bp – 25,000bp upstream of 3’ termini were identified as putative Helitrons. The sequences of these putative Helitrons as well as 1000bp of flanking sequence were stored for further analysis.

The computationally determined Helitrons were processed through a two-fold analysis used to identify possible Helitrons of merit. The putative Helitron sequences were evaluated by utilization of the National Center for Biotechnology Information’s Basic Local Alignment Search
Tool (BLAST). The putative Helitron sequences were batch processed through the BLAST to align the nucleotide sequences against the Expressed Sequence Tag (EST) database for *Zea mays*. The BLAST results were parsed to isolate Helitrons that have captured EST sequences of a significant length which aligned with over 99% identity. The purpose of this processing is to identify those Helitrons which have captured partial or full-length EST in hopes of finding evidence of full length gene capture. An additional analysis was performed on the flanking sequences of the computationally identified Helitrons. These flanking sequences (absent of putative Helitrons) were processed through the BLAST for alignment against the high throughput genomic sequences of *Zea mays*. These BLAST results were parsed in a manner which isolated those results which matched with high alignment around the site of the Helitron insertion. Those sequences which were isolated indicated the presence of a paralogous sequence of *Zea mays* genomic DNA which

![Figure 1: Consensus of example 5' ends of experimentally determined Helitrons.](image1)

![Figure 2: Consensus of example 3' ends of experimentally determined Helitrons.](image2)
does not contain the *Helitron* insertion. The presence of such paralogous sequences would provide confirmation that the computationally determined *Helitron* sequences correlated with genetic insertions *in vivo*.

The isolated *Helitrons* from the two-fold analysis would be taken as candidates for continued manual annotation. This annotation would potentially lead to the identification of *Helitrons* that have captured transcribed full-length protein-coding regions. The HelRaizer system has resulted in the identification of 2,376 putative *Helitron* insertions. The average length of these *Helitrons* is 7,336bp with a median length of 6,129bp. The range of *Helitron* lengths is 168bp – 25,024bp. These putative *Helitrons* compose 17.4 Mb or approximately .725% of the total maize genome. A histogram analysis of the lengths of these 2,376 possible *Helitrons* is displayed in Figure 3.

![Figure 3: Histogram of Possible Helitron Lengths](image)

The histogram shows a length distribution with an irregular abundance of *Helitrons* with lengths between 1,500-1,750bp. This could be a result of multiple insertions of the same *Helitron* across the maize genome. Further analysis of these similar length *Helitrons* should be performed in the future. Further, nucleotide pair-wise alignment of similar length *Helitrons* would lead to a better understanding of their similarities and possible significance.
An additional measure of merit was performed by attempting to quantify the number of gene fragments captured by the computationally determined Helitrons. The Helitron sequences were compared to existing gene sequences and aligned using the BLAST. Batch alignments were performed and alignments matching gene fragment of more than 50bp with at least 85% similarity were recorded as an instance of gene capture. Across the 2,376 possible Helitrons, 4,310 unique gene fragments were captured. This result of 1.81 gene fragments per Helitron indicates a higher quality Helitron prediction by HelRaizer than by previous computational methods.

EXPERIMENTAL MATERIALS AND METHODS

In order to pursue the experimental objectives in determining the efficacy of the computational methods and exploring the putative novel case of a Helitron-captured full length gene, genomic DNA (gDNA) was first extracted from ten maize inbred lines: B73, CML277, CML322, HP301, KI11, Mo17, Ms71, OH43, OH7B, and TZI8. The DNeasy® Plant Mini Kit and protocol provided by the vendor was used to isolate the gDNA from kernels from each of the respective lines, and the yield was confirmed by running samples of each of the extractions on agarose gels using electrophoresis (the kit and protocol were also utilized to attempt to extract gDNA from the leaves of two Teosinte inbred lines. However, this extraction was unsuccessful, as evidenced by lack of bands after gel electrophoresis).

The different lines of maize have been isolated and inbred over many generations, resulting in differing evolutionary lineage, with different mutations passed down through the various inbred lines. Helitrons have transposed in different regions of the maize genome after the differentiation between lines, so they rarely appear inserted at the same loci in every inbred line. Instead, they are found in lines that are related by hereditary linkages after the Helitron inserted into the maize genome. Though computational methods identify potential Helitrons, experimental determination using plus/minus polymorphism confirm its presence through molecular analysis. Plus/minus polymorphism demonstrates that the Helitron is
present in some inbred lines and not in others, thus revealing that the element detected by computational methods is an authentic transposable element.

Two separate experimental approaches were employed to determine the plus/minus polymorphism of the *Helitron* of interest. The first was to determine the presence/absence of the *Helitron* in ten different inbred lines. To do this, two sets of primers were designed. One set had one primer (FGHelUp1) encompassing a region just outside the 5′ *Helitron* insertion site and the other primer (FGHelDn1) over a region in the first exon of the gene captured within the *Helitron*, with an expected product length of 4,361 base pairs. The second set of primers had one primer (FGHelUp2) complementary to the third (final) exon of the *Helitron*-captured gene and the other primer (FGHelDn2) complementary to the sequence just outside the 3′ *Helitron* insertion site, with an expected product length of 2,059 base pairs. These primers were designed to yield a positive result only if the inbred line contained the *Helitron* in the same loci. The primers were intended to only amplify the region that contained both the flanking insertion sites of the *Helitron* along with the *Helitron* itself. Polymerase Chain Reaction (PCR) was performed using these sets of primers and the gDNA extracted from each of the ten inbred lines of maize. An annealing temperature of 55° Celsius and an extension time of five minutes was used for PCR amplification. Samples of the products from the PCR were analyzed using gel electrophoresis to determine their sizes.

The primers designed for the 3′ end of the *Helitron* were not specific enough; the lines showed multiple bands of varying lengths in at least seven inbred lines. Though some of the bands seemed to be of the expected product length, the lack of specificity precluded any conclusions from being drawn from these results. As seen in Figure 4, the 5′ primers showed more specificity.

**Figure 4: Presence of Helitron Analysis Using 5′ Primers**
Initial results showed isolated bands of amplified DNA at the expected product length in
the inbred lines B73 and Ms71, suggesting plus-minus polymorphism. However, more
experimentation is needed to confirm reproducibility.

The second experimental approach utilized for substantiating plus/minus polymorphism
involved looking for a paralogous locus; that is, looking for the empty site where the Helitron
would have inserted in other inbred lines. Two sets of primers (PKRP2Ex3Up1/Dn1 and
PKRP2Ex3Up2/Dn2) were designed flanking the insertion site, and would amplify 400 and 796
base pairs, respectively, if the Helitron were not present. If the Helitron was present in an
inbred line on which PCR was performed with these primers, the two primers would be located
too far apart within the DNA strand (by several thousand base pairs) for any PCR product to be
amplified. If the Helitron is not present, however, theoretically the PCR should return a product
of the length stipulated by the primer design, due to the absence of the Helitron and probable
splicing together of the sequences outside the 5’ and 3’ ends of the insertion site.

Running the PCR products using gel electrophoresis yielded no amplification for one set
of primers (PKRP2Ex2Up2/Dn2). As shown in Figure 5, the other set of primers
(PKRP2Ex3Up1/Dn1) did not show a band of DNA at the expected length, but did show a strong
band at approximately 1,500 base pairs in several inbred lines. More analysis is needed to
confirm and explain these results.

![Figure 5: Paralogous Locus Analysis of Ten Inbred Lines](image)

The Helitron of interest was chosen for intensive investigation because through
computational means it appears to have captured a full-length gene, a variation of the PKRP
gene. After confirming its existence as a Helitron, the aim was to analyze this gene’s
significance in the genome, specifically whether it is transcribed and expressed as a protein. If
it is expressed, it would be the first case of a full-length gene being captured by a Helitron and
subsequently transcribed.
In order to demonstrate a potential full-gene capture, the ten inbred lines of corn were studied for the presence of the wildtype PKRP gene, PKRP2, by designing primers to flank a 680bp sequence within the gene (FGeFLKUp/Dn), and performing PCR. As shown in Figure 6, a strong band at the expected length was amplified in all ten inbred lines, indicating the gene’s incidence in each of line and its apparent transcription.

**Figure 6: Presence of Wildtype Gene in Ten Inbred Lines**

Next, the presence and transcription of the *Helitron*-captured variation of the PKRP gene (PKRP1) versus the wildtype gene (PKRP2) was explored. Although both genes belong to the same family, they are divergent in both their 5’ and 3’ termini, which enabled the design of gene-specific primers for PCR amplification of each variant. Primers were designed to adhere only to the divergent termini, allowing amplification to occur only in each inbred line if it contained the sequences specific to the selected variation of the gene (PKR1RPUp/Dn and PKR2RPUp/Dn). Furthermore, to test the transcription of each gene rather than just its presence in the maize genome, the messenger RNA (mRNA), rather than the gDNA, of two inbred lines was evaluated. Genomic DNA contains segments of DNA known as exons, which are the portions of DNA that code for protein expression and production, and other segments called introns, also referred to as “junk DNA”. The relevance of introns is currently unknown, but it is clear that they play a lesser role in the expression of proteins. During protein production, gDNA is first transcribed into pre-mRNA, a transcript that complements each DNA strand and includes both the introns and exons. The introns from this pre-mRNA are then removed by spliceosomes, which also splice the remaining exons together to form mRNA. The mRNA is translated into proteins. By analyzing the mRNA of the ten inbred lines rather than the DNA, it is determined not only whether the specifically targeted genes are present within the line, but whether the genes are transcribed into mRNA in a specific tissue and thus are relevant in protein expression.
Extraction of mRNA from the roots of the B73 and Mo17 inbred lines was completed and the product was run on a gel to ensure the success of the extraction. Sample mRNA extracted by previous investigators was also utilized. Next, reverse transcriptase PCR (RT-PCR) was executed to obtain complementary DNA from the messenger RNA and later perform PCR using the primers described. As displayed in Figure 7, the amplification products showed the expression of the PKRP2 wildtype gene in all ten of the inbred lines, but only the expression of the PKRP1 Helitron-captured gene in the Mo17 line, and only a faint band in that line. These results, too, must be confirmed for reproducibility in the future.

**Figure 7: RT-PCR Analysis of PKRP Gene Expression in Maize Roots**

Another method used to determine whether the Helitron-captured gene was present in any of the ten inbred lines was through PCR analysis using primers designed entirely within the gene PKRP1. To do this, we designed two different forward primers (InFGHelUp1/Up2) and one reverse primer (InFGHelDn1). The two forward primers were inside the Helitron insertion site and were located in introns of the PKRP1 gene. The reverse primer was located in the first exon of the gene. We decided to design two forward primers in order to confirm the presence using amplification of two different lengths of DNA. Since this was the last experiment we did, we did not have time to optimize the PCR amplification, and the results are inconclusive.

The first PCR reaction we ran using the two different forward primers and the reverse primer. We attempted to amplify these sequences in B73, CML322, HP301, Mo17, and Ms71. As shown in Figure 8, this reaction produced bands in four of the five lines at the expected length for InFGHelUp1/Dn1, but slightly larger bands than expected for InFGHelUp2/Dn1. Additionally, the bands for InFGHelUp2/Dn1 were significantly fainter than those of InFGHelUp1/Dn1. The fainter bands may be due to a lower level of expression, but further
In order to clarify these results, we ran another gel using only B73, CML322, and HP301. In this reaction, we used the primer combination InFGHelUp1/Dn1. As a positive control for amplification, we used a set of primers known to amplify a sequence in all inbred lines of approximately 400bp in length (PKRP2Ex3Up1/Dn1). We would have expected to see amplification in all three lines for each set of primers; however, we did not see amplification of B73 using either set.

Due to the differing results of these two experiments, we ran the PCR again using B73, CML322, and HP301, using both InFGHelUp1/Dn1 and the positive control primers. We were concerned that since no bands appeared for B73 that perhaps the gDNA had degraded. This time, in order to confirm the presence of gDNA before PCR amplification, we also ran gDNA samples on the gel. Although the gel confirms that the gDNA was still present, we did not see any amplification for InFGHelUp1/Dn1, and only obtained amplification in the positive control PCR for B73 and CML322. Since these results are not conclusive, more work needs to be done to confirm which lines contain the sequence amplified by InFGHelUp1/Dn1.

CONCLUSIONS

We feel that we have made two important contributions in this research: 1) we developed a tool that will be useful to all researchers; and 2) we have contributed evidence to support the functional hypothesis about the role of Helitrons.

We developed a program, HelRaizer, which will be an invaluable tool in advancing our understanding of the mechanisms underlying Helitrons. Our short experience with HelRaizer suggests that it is both efficient and effective in supporting this endeavor. HelRaizer will be a
crucial component in the evaluation of additional Helitrons. Additionally, further molecular analysis will be done on other Helitrons which may contain full, active genes.

We have produced supporting evidence for the hypothesis about the functionality of Helitrons. In particular, we found strong evidence for the Helitron capture of a PKRP1 gene. It remains to be investigated whether this gene capture plays an active role within the genome, although our evidence thus far supports this conclusion.

Our work will provide a basis for further discussion of Helitrons and their method of gene capture. By providing an improved way of finding high quality Helitrons as well as a novel case of potential full gene capture, we hope our work will contribute to the understanding of Helitron mediated gene capture mechanisms.

Future work to be undertaken includes both annotation of genes found via computational means as well as molecular analysis of cases in which full, active genes appear to have been captured.
<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Primer Sequence</th>
<th>Expected Fragment Length (bp)</th>
<th>Amplification</th>
<th>Locus (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGHelUP1</td>
<td>CCTGGATAACCGTGAGAGGA</td>
<td>4,361</td>
<td>This amplifies a sequence from outside the 5' end of the Helitron to within the first exon of the gene within the Helitron</td>
<td>83,104 to 83,123</td>
</tr>
<tr>
<td>FGHelDN1</td>
<td>AAACGGGCAAAAGAAGAAGCA</td>
<td>2,098</td>
<td>This amplifies a sequence from within the third exon of the gene within the Helitron to outside the 3' end of the Helitron</td>
<td>93,047 to 93,066</td>
</tr>
<tr>
<td>FGHelUP2</td>
<td>GTCTGTACTCAGCCGCAGAA</td>
<td>2,098</td>
<td>This amplifies a sequence within the third exon of the gene within the Helitron to outside the 3' end of the Helitron</td>
<td>93,047 to 93,066</td>
</tr>
<tr>
<td>FGHelDN2</td>
<td>AACTGAAGGAGGACGAACGA</td>
<td>2,098</td>
<td>This amplifies a sequence within the third exon of the gene within the Helitron to outside the 3' end of the Helitron</td>
<td>93,047 to 93,066</td>
</tr>
<tr>
<td>FGHelFLKUp</td>
<td>CCCCTTAATTCTCCCCTCTCAAT</td>
<td>680</td>
<td>This amplifies a region of the PKRP2 gene</td>
<td>115,512 to 115,531</td>
</tr>
<tr>
<td>FGHelFLKDn</td>
<td>GATGAGAAGCAGCTGAGGAG</td>
<td>680</td>
<td>This amplifies a region of the PKRP2 gene</td>
<td>116,171 to 116,190</td>
</tr>
<tr>
<td>PKRP2Ex3Up1</td>
<td>CCTTCACAAATGCTCAATGG</td>
<td>400</td>
<td>This amplifies the flanking sequence of the Helitron</td>
<td>83,348 to 83,368</td>
</tr>
<tr>
<td>PKRP2Ex3Dn1</td>
<td>AACTGAAGGAGGACGAACGA</td>
<td>400</td>
<td>This amplifies the flanking sequence of the Helitron</td>
<td>83,348 to 83,368</td>
</tr>
<tr>
<td>PKRP2Ex3Up2</td>
<td>CCGGATAACCGTGAGGAGA</td>
<td>796</td>
<td>This amplifies the flanking sequence of the Helitron</td>
<td>95,276 to 95,296</td>
</tr>
<tr>
<td>PKRP2Ex3Dn2</td>
<td>CGTGGAGGCGCATAGTGTGA</td>
<td>796</td>
<td>This amplifies the flanking sequence of the Helitron</td>
<td>95,276 to 95,296</td>
</tr>
<tr>
<td>PRKR1RPUp</td>
<td>GACGGCTCCCTAACGTAAAC</td>
<td>1,056</td>
<td>This amplifies a region of the Helitron-captured PKRP2 gene specific to the Helitron-captured variant of the gene</td>
<td>87,136 to 87,155</td>
</tr>
<tr>
<td>PRKR1RPDn</td>
<td>ACCTCACATCAATCCATGCT</td>
<td>1,056</td>
<td>This amplifies a region of the Helitron-captured PKRP2 gene specific to the Helitron-captured variant of the gene</td>
<td>93,244 to 93,263</td>
</tr>
<tr>
<td>PRKR2RPUp</td>
<td>GCCGAACCTCCCTAACCCCTAA</td>
<td>1,051</td>
<td>This amplifies a region of the wild-type PKRP2 gene specific to the wild-type variant of the gene</td>
<td>120,248 to 120,229</td>
</tr>
<tr>
<td>PRKR2RPDn</td>
<td>TGACGCAGTGCCACTTCTCTG</td>
<td>1,051</td>
<td>This amplifies a region of the wild-type PKRP2 gene specific to the wild-type variant of the gene</td>
<td>115,814 to 115,833</td>
</tr>
<tr>
<td>InFGHelUp1</td>
<td>GCATCGACGTAGCTGACAT</td>
<td>2,102</td>
<td>This amplifies a region within the 5' end of the Helitron to within the first exon of the gene within the Helitron</td>
<td>85,473 to 85,492</td>
</tr>
<tr>
<td>InFGHelDn1</td>
<td>CTAGCTCGAATGGATGCTGAC</td>
<td>2,102</td>
<td>This amplifies a region within the 5' end of the Helitron to within the first exon of the gene within the Helitron</td>
<td>85,614 to 85,594</td>
</tr>
<tr>
<td>InFGHelUp2</td>
<td>GTGTTCCCTCGAAGCTCCTGCTG</td>
<td>1,683</td>
<td>This amplifies a region within the 5' end of the Helitron to within the first exon of the gene within the Helitron</td>
<td>83,769 to 83,790</td>
</tr>
<tr>
<td>InFGHelDn1</td>
<td>CTAGCTCGAATGGATGCTGAC</td>
<td>1,683</td>
<td>This amplifies a region within the 5' end of the Helitron to within the first exon of the gene within the Helitron</td>
<td>87,614 to 87,594</td>
</tr>
</tbody>
</table>


See Appendix for detailed primer information