Assembly and annotation tools for analysis of large contiguous regions of the maize genome

Brent Anson Kronmiller

Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtd

Part of the Bioinformatics Commons

Recommended Citation

Assembly and annotation tools for analysis of large contiguous regions of the maize genome

by

Brent Anson Kronmiller

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Roger Wise, Major Professor
Xun Gu, Co-major Professor
Volker Brendel
Thomas Peterson
Kan Wang

Iowa State University
Ames, Iowa
2008

Copyright © Brent Anson Kronmiller, 2008. All rights reserved.
UMI Number: 3330843

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 3330843
Copyright 2008 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346
TABLE OF CONTENTS

LIST OF FIGURES iv
LIST OF TABLES v
ABSTRACT vi

CHAPTER 1. GENERAL INTRODUCTION 1

Thesis Organization 2

Literature Cited 11

CHAPTER 2. TENEST: AUTOMATED CHRONOLOGICAL ANNOTATION AND VISUALIZATION OF NESTED PLANT TRANSPOSABLE ELEMENTS 17

ABSTRACT 17

INTRODUCTION 17

RESULTS 20

DISCUSSION 39

MATERIAL AND METHODS 42

ACKNOWLEDGEMENTS 44

LITERATURE CITED 44

CHAPTER 3. COMPUTATIONAL FINISHING OF HIGHLY REPETITIVE MEGABASE SEQUENCE CONTIGS OF THE rf1-ASSOCIATED REGION OF MAIZE 49

ABSTRACT 49

INTRODUCTION 49

RESULTS AND DISCUSSION 52

CONCLUSION 75

METHODS 78
LITERATURE CITED

CHAPTER 4. COMPARATIVE SEQUENCE ANALYSIS OF THE MAIZE
rf1 LOCUS

ABSTRACT

INTRODUCTION

RESULTS AND DISCUSSION

CONCLUSION

METHODS

LITERATURE CITED

CHAPTER 5. GENERAL CONCLUSIONS

ACKNOWLEDGEMENTS
LIST OF FIGURES

CHAPTER 2

Figure 1. Recombination of Separated TE Sections. 24

Figure 2. TEnest Graphical Display; Barley, Maize, Rice, Wheat. 26

Figure 3. Chronology of Maize Whole LTR Retrotranposon Insertions. 28

Figure 4. Maize LTR Retrotransposon Quantification. 36

Figure 5. Phylogenetic Analysis of the Maize Ji and Grande Retrotransposons. 37

Figure 6. Maize LTR Retrotransposon Class by BAC. 38

CHAPTER 3

Figure 1. Combined Genetic and Physical Map of Maize Sequence Contigs. 56

Figure 2. Nested LTR Retrotransposons Cause Sequence Assembly Gaps. 59

Figure 3. TEnest Graphical Display of Maize Sequence Contigs. 62

Figure 4. Comparative Analysis of Maize Sequence Contigs to Rice and Sorghum. 74

CHAPTER 4

Figure 1. Comparative TEnest View of B73 and Wf9-BG BAC Contigs. 93

Figure 2. LTR Retrotransposon Recombination Methods to form an expanded Cassandra Element. 94

Figure 3. Sequence Comparison of B73 rfl-associated Contig 1 and Wf9-BG. 96

Figure 4. Increasing Polymorphism Frequencies nearing Recombination Location. 99
LIST OF TABLES

CHAPTER 2

Table I. Verification: TEnest results compared to GenBank annotated maize BACs 30
Table II. TEnest annotations identified across 29.3 Mb of maize sequence 34
Table III. Rates of solo LTR formation 35

CHAPTER 3

Table I. Sequenced BACs across the rf1 locus 54
Table II. Transposable elements identified with TEnest by repeat cluster 65
Table III. Predicted genes identified across the two maize sequence contigs of chromosome 3 68

CHAPTER 4

Table I. Predicted genes identified in maize Wf9-BG 92
Table II. Polymorphism Frequencies between maize B73 and Wf9 98
ABSTRACT

Sequencing projects continue to tackle larger and more challenging genomes. Many of the grass genomes have important agriculture and economic impacts. The maize genome is now underway, and others important as foods and biofuels will be sequenced in the near future. The grass genomes are very large and therefore computationally complex to assemble and annotate. Long terminal repeat retrotransposons make up significant portions of many of the longer grass genomes. Their repeat sequences across the genome, their terminal repeats, and their nested cluster configuration make assembly of sequence clones challenging and identification of gene regions difficult. Tools are needed to assist with the more difficult types of genomes that are sequenced today and will be sequenced in the future. Sequencing of the maize genome is underway, but still much is not known about the landscape of the genome. While many smaller regions of maize have been sequenced, they cannot give a full picture of the structure and layout of gene islands and of repeat clusters. In addition, because of the available small sequenced contigs of maize, a true view of the relationships between maize and other grass genomes remains elusive. In this thesis I provide tools necessary for both assembly and annotation of highly repetitive genomes, and I use these tools to construct the currently two longest maize sequence contigs. These contigs provide a resource for many of the unanswered questions of the maize genome.

In the first part of the thesis I present TEnest, annotation and visualization software for transposable elements in grass genomes. TEnest identifies all fragmented transposable elements within the input sequence and reconstructs each to the original insertion state. This provides a chronological display of the nesting pattern of clustered transposable elements. For long terminal repeat (LTR) retrotransposons, TEnest calculates an estimated age since insertion is based on the divergence of its paired LTRs. TEnest is available for use as a downloadable program or on PlantGDB as a web utility; it will also be integrated into the MaizeSequence.org maize genome annotation pipeline and the TriAnnot wheat genome annotation pipeline. In this chapter I also provide a case study of TEnest on the available maize genome sequence. TEnest shows the distribution of transposon families, ages of insertion, and frequencies of solo LTRs. In addition I provide a phylogenetic analysis of retrotransposon families showing the estimated ages since insertion of LTR retrotransposons
cluster with their sequence identity, showing that LTR retrotransposons experience specific intervals of extreme proliferation to expand across the genome.

In the second part of this thesis I introduce our two contiguous maize sequences, rf1-associated contigs rf1-C1 and rf1-C2 sequenced from maize B73. These are the two longest contiguous maize sequences and provide previously unmatched sequence quality for answering many questions surrounding the makeup of the maize genome. Here, using TEnest, we propose two maize assembly techniques for highly repetitive regions. The use of these processes has allowed us to provide the high quality contiguous sequences of the rf1-associated region and will assist researchers with assembly of difficult sequence clones. We show definite separation between gene and repeat regions. LTR retrotransposons are nested into clusters, some as long as 150 kb. With smaller sequences and gapped assemblies, the boundaries of these clusters would not be seen. With the rf1-associated contigs and with repeat annotation by TEnest, we are able to show that genes are indeed found in grouped islands between the repeat clusters. The rf1-associated contigs, when compared to the rice and sorghum genomes, show conserved macro-colinearity between genes across the long sequences. But a closer look at individual gene islands shows there is micro-noncolinearity across the analyzed grass species.

The third section of this thesis compares the B73 rf1-associated sequence contigs with two bacterial artificial chromosomes (BACs) sequenced from Wf9-BG, an Rf1 containing maize line. Here we identify four genes in an island corresponding to a similar gene island in B73, a fifth gene, however, is missing from Wf9-BG. Two repeat clusters surround the gene island; one matches its counterpart in B73. Several differences in transposable element insertions are seen in the matching sequence regions; in one case we observe a Cassandra LTR retrotransposon that is either the consequence of unequal sister chromatid recombination or has experienced unequal illegitimate recombination expanding the element yet retaining mobility. The second repeat cluster does not align between B73 and Wf9-BG. We identify a region of recombination causing the observed sequence differences and, leading up to this area, the frequency of polymorphisms between B73 and Wf9-BG drastically increase. We observe a more conserved SNP frequency between B73 and
Wf9-BG than previously reported in maize, thus highlighting the relationship between the two lines.
CHAPTER 1. GENERAL INTRODUCTION

Found in the nucleus of each eukaryotic cell, genomic DNA contains the hereditary information and instructions required for organism functionality. The genome is made up of genes, inheritable traits corresponding to coding regions and often a discernable phenotype, and non-coding DNA, containing areas responsible for gene regulation and regions of repetitive sequence. The primary objective of DNA sequencing is to locate, identify, and characterize genes, focusing on the areas of the genome that affect heritable traits. A further step is to sequence the entire genome, giving all the genes that make up an organism’s DNA. Determining the nucleotide code of an entire genome gives the ability to identify genes and gene regulatory regions, investigate genome and organism evolution, and evaluate genome structure and makeup.

Two methods are commonly used for genome sequencing. Whole genome shotgun (WGS) sequencing cuts the entire genome into small subsections, each piece is cloned into vector, sequenced, and the whole genome is re-assembled via assembly software. BAC by BAC sequencing first chops the genome into large subsections, clones these parts into bacterial artificial chromosomes (BACs), determines a minimal tiling path from each BAC sequence with restriction digest mapping, fully shotgun sequences each BAC, and finally assembles each BAC to make a spanning sequence across the genome. Each procedure has its advantages and disadvantages, while WGS is a faster and less expensive process, BAC by BAC sequencing generally provides a more complete and better assembled genome. Each of these genome sequencing procedures produces a draft assembly, contigs of sequence separated by gaps and containing regions of low quality. To close sequence gaps and enhance sequence quality specialty laboratory sequencing and computational processes are used. While an extremely time consuming process, finishing sequencing is necessary to provide complete genome sequence.

Depending on the size and complexity, sequencing an entire genome can be a costly and laboratory intensive process. The content of a genome can increase its sequence complexity. Many genomes contain extraneous sequence, regions that are not genes and do not necessarily influence the functioning of genes, this is often termed ‘junk’ DNA. Much of junk DNA is made up of repetitive elements, both autonomous and non-autonomous DNA...
that replicate throughout the genome. Repetitive elements are found in many duplicated locations across the genome and can greatly hinder sequence similarity alignments of assembly software. However, correctly sequencing and assembling the repetitive areas can provide insight to the functions of these regions and their influence on genome structure and evolution.

Grasses are the next challenge for whole genome sequencing. Covering many staple foods across the globe, the sequencing of grass genomes will have a major agricultural and economic impact. However, obtaining grass genomes will be a challenging endeavor. The genomes of many grasses are both extremely large and highly repetitive. The length and repetitive complexity makes genome sequencing, assembly, and finishing even more difficult than previously sequenced organisms, new technology is necessary to provide a fully sequenced genome. Currently several grass genomes have been sequenced. The *Oryza sativa* (rice) genome is 420-466 Mb, two subspecies, *indica* (Yu et al., 2002) and *japonica* (Goff et al., 2002) have been sequenced. Rice is 35% repetitive (IRGSP, 2005) and most of this repeat content is small MITE DNA transposons. The 770 Mb *Sorghum bicolor* (sorghum) genome has also been sequenced, a full analysis of the genome has not yet been completed (http://www.phytozome.net/sorghum). Sequencing of the *Zea mays* (maize) genome is currently underway. Maize is approximately 2.4 to 2.8 Gb (Arumuganathan and Earle, 1991) and at least 67% repetitive (Kronmiller and Wise, 2008) and may be as much as 80%. Initial analysis suggests genes are found in small groups surrounded by large regions of clustered repeats. The maize sequencing project will therefore focus on sequence and assembly of the small gene clusters and, at least at first, not attempt to correctly assemble the large repeat clusters (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0527192). In the near future the larger genomes of *Triticum aestivum* (wheat) and *Hordeum vulgare* (barley) will need to be sequenced, as well as grasses with economic interests for biofuels and biofeedstocks.

**Thesis Organization**

This thesis is organized into three sections corresponding to three manuscripts investigating highly repetitive grass genome sequence, annotation, and analysis:
1. The maize genome is thought to consist of large nested repeat clusters. Here I present the TE identification software TEnest, designed for reconstruction and annotation of densely nested LTR retrotransposon clusters. Also included in this manuscript is an analysis of TEnest annotation on the currently sequenced maize genome.

2. We have sequenced and assembled to completion two large contiguous contigs of the rfl-associated region of maize. The rfl-associated contigs are currently the two longest finished maize sequence contigs and, given the goals and procedure of the maize genome project, may be the longest contiguous sequences for some time. These contigs provide a unique resource for evaluating the genome structure and landscape of maize. Here I present the sequence and annotation of the two rfl-associated contigs and a view of long contiguous regions of maize.

3. We have also sequenced two additional BACs from Rfl Wf9-BG. Here I present a comparative sequence analysis of Wf9-BG to B73 showing differential TE insertions, a recombination breakpoint, and a SNP analysis. In the following introduction I present each of the three sections and provide an overview of their impact.

**TEnest: Nested Transposable Element Annotation**

Mobile DNA is found across the genomes of eukaryotic organisms. Transposable elements can make up significant portions of genome sequence but their introduction, modes of replication, distribution, evolutionary development, and effects on host genes and evolution are relatively poorly understood. With their mobility, repeated sequence, and abundance, TEs can induce recombination (Xiao et al., 2008), enhance gene regulation, create new genes via sequence alterations (Kidwell and Lisch, 2000) or host recruitment of TE genes, and move genes across the genome (Lal et al., 2003).

Percentage of TE content in the genome varies across organisms, even closely related species can have vastly different TE amounts. The grass genomes are highly repetitive, rice is 35% (IRGSP, 2005), maize is at least 67% (Haberer et al., 2005; Kronmiller and Wise, 2008), wheat and barley are between 60% to 80% (Flavell, 1986; Sabot et al., 2005). There are several classes of TEs characterized by their modes of replication (Wicker et al., 2007).
The two main classes; Class I retrotransposons replicate via transcription to an RNA intermediate and are reverse transcribed into DNA and inserted back into the genome, Class II DNA transposons replicate in a cut and paste method. These classes can be further subdivided into subclass, order, superfamily, family, subfamily, characterized by sequence identity. Even within the grasses the TE classes vary; the rice genome is made up of 13.0% DNA transposons and 19.3% retrotransposons (IRGSP, 2005), while maize is 1.3% DNA transposons and 63.3% retrotransposons (Kronmiller and Wise, 2008).

With such a high percentage of repetitive elements the maize genome is a difficult organism to sequence, assemble, and annotate. LTR retrotransposons are extremely abundant throughout the genome, many of the prolific families are found in multiple copies in the sequence of a single BAC clone. While they range in size many are very long, Grande is 14kb in length (Martinez-Izquierdo et al., 1997), Danelle is 15.5 kb (Kronmiller and Wise, 2008). Their high density and length provide ample opportunity for nesting, insertion of one TE within the sequence of another. (SanMiguel et al., 1996; SanMiguel et al., 1998). Nesting breaks up the sequence of the original TE insertion making annotation difficult. Clusters of LTR retrotransposons are subsequent TE insertions within one another, forming levels upon levels of TEs. In nested repeats the most recent insertions are found at the top most level of the cluster. A LTR retrotransposon is composed of a beginning LTR, an internal region containing gag and pol genes, and a final LTR, a duplicated sequence of the beginning LTR. The LTRs can be up to 4.5 kb in some families, their repeated sequence along with duplicated TE copies within a BAC can greatly hinder assembly software. Correct assembly, reconstruction, and annotation of TE clusters are essential for identification of gene locations and evaluation of the evolution of the genome. Complete annotation of TEs coupled with calculation of insertion age of LTR retrotransposons can give the evolutionary chronology of the genome, provide phylogeny of TEs and their replication path across the genome, and assist with identifying the changes that have occurred throughout the genome.

Previous repeat identification software was not designed for the highly repetitive grass genomes. Sequence similarity repeat software such as RepeatMasker (Smit et al., 1996-2004) and De-Novo TE identification software such as PLIER (Edgar and Myers,
RECON (Bao and Eddy, 2002), and RepeatScout (Price et al., 2005) were sufficient for the low repetitive genomes of human (45% repetitive) (Lander et al., 2001), Arabidopsis (10% repetitive) (The Arabidopsis Initiative, 2000), and Drosophila (3.9% repetitive) (Kaminker et al., 2002). For the longer LTR retrotransposons specific software is needed. De-novo LTR identification software such as LTR_struct (McCarthey and McDonald, 2003; Kalyanaraman and Aluru, 2006) is able to find and characterize LTR retrotransposons but cannot correctly identify the more than 70% of TEs affected by nesting. Clearly for annotation of the maize and further sequenced grass genomes a tool specifically designed for highly repetitive genomes is necessary. This software needs to identify fragmented TE pieces and reconstruct each piece to its original whole TE state. It needs to display TE clusters in such a way as to make identification of gene locations simple. It needs to be a tool for assembly of repeat collapsed BAC clones. As maize is 63.3% LTR retrotransposons, it needs to calculate age since insertion (Mya) to show the chronology or TE insertions throughout the genome.

In order to correctly identify and reconstruct nested TEs in repeat rich grass genomes and to display their chronological insertion pattern we have developed TEnest. TEnest utilizes a database of consensus TE sequences to identify all TE portions throughout the input DNA. Starting with the LTRs of LTR retrotransposons it reconstructs fragmented sections using a power set process (Suppes, 1960). LTRs are paired according to their similarity to give the boundaries of LTR retrotransposons. From here, an iterative process of identification of internal retrotransposon regions is conducted starting with smallest spanning LTR pairs first. This gives whole length reconstructed LTR retrotransposons. Based on LTR sequence divergence the time since insertion (Mya) (Kimura, 1980; Ma and Bennettzen, 2004) can be calculated. With LTR retrotransposon identification complete, TEnest locates and reconstructs DNA transposons, and finally annotates any partial fragments of TE insertions.

We present TEnest annotations on the currently finished maize BACs totaling 29.3 Mb. This analysis shows maize is at least 67% TEs, 95% of the TEs are LTR retrotransposons. LTR retrotransposon families are unequally represented; 3 families, \textit{Ji}, \textit{Opie}, and \textit{Huck} make up more than 50% of all retrotransposons. Solo LTR abundance is not
Sequence and Analysis of Large Contiguous Maize Contigs

The maize genome sequencing project is well underway, the goal of which is to obtain a full set of all genes that make up the organism (Rabinowicz and Bennetzen, 2006). With a genome size of 2.4 Gb and a repeat content of at least 67%, maize is one of the more challenging genomes sequenced to date. Several methods were considered for the maize genome sequencing project. A whole genome shotgun, while relatively quick and inexpensive, was determined to be too difficult for assembly give the high percentage of TEs in the genome. Gene enrichment sequencing methods, methylation filtration (Rabinowicz et al., 1999; Palmer et al., 2003; Whitelaw et al., 2003) and high-Cot (Yuan et al., 2003) were
investigated in-depth for selection of gene regions. While capturing a majority of maize genes (Fu et al., 2005) many genes could still be omitted. Finally a BAC by BAC sequencing approach was selected, but because of the high abundance of TEs in the maize genome and the difficulty for BAC sequence assembly repeats create, assembly and finishing will focus on the gene regions and initially ignore the clustered repeats (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0527192).

The finishing process of genome sequencing is the most time consuming stage. Sequence quality enhancement and gap closure of BACs is obtained through resequencing of plasmid subclones, sequencing off custom primers, PCR amplification off custom primers, and full length sequencing of mini-libraries or transposon-bombing of PCR products and subclones. However, the most difficult and lengthy process is not the ‘specialty’ sequencing work but in obtaining the correct configuration of misassembled sequences. Misassemblies can cause gaps, collapses, or simple wrong assemblies of BAC sequence, the most often cause are repeated sequences within the BAC. With a high repeat content largely of LTR retrotransposons, duplicated TEs and multiple copies of similar LTR sequences make maize BAC assembly challenging.

The maize genome is thought to consist of small groupings of several genes surrounded by large expanses of LTR retrotransposon repeat clusters. Several repeat clusters of maize have been sequenced and analyzed (Song et al., 2001; Brunner et al., 2005; Haberer et al., 2005). Ninety-five percent of maize TEs are LTR retrotransposons, these are long, and by their replication process, genome expanding TEs (SanMiguel et al., 1996; SanMiguel and Bennetzen, 1998; Bennetzen et al., 2005; Hawkins et al., 2006; Piegu et al., 2006). Proliferation of LTR retrotransposons across the maize genome is one of the major causes for its expansion in genome size. Expansions of the intergenic regions of maize are caused almost exclusively by numerous retrotransposon insertions. Repeat clusters are nested piles of LTR retrotransposons, a dumping ground of new insertions upon older TEs. However, little is known of repeat cluster make up. Do TE families integrate into certain repeat clusters while ignoring others? Are some repeat clusters ‘older’ based on their member’s LTR divergence than other clusters? Do repeat clusters experience more or less recombination identified by partial truncated TEs, or unequal illegitimate recombination
evidenced by solo LTRs and recombined LTR retrotransposon families? Full sequence and correct assembly of multiple repeat clusters can answer these questions. Gene islands are also not well understood or characterized. Previous studies have given a varied view, several analyses of finished maize contigs ranging from 100 to 400 kb show gene densities of 1 gene in 10 to 140 kb and evidence for and against tight clusters of genes (Song et al., 2001; Brunner et al., 2005; Haberer et al., 2005; Bruggmann et al., 2006). Large contiguous maize sequence contigs are needed for more in-depth analysis to resolve these discrepancies.

In order to provide an outlook of large contiguous sequence regions we have sequenced and finished 16 BACs from maize B73 associated with the rf1 locus. Rf1 is one of the two complementary fertility restoration genes for T (Texas) cytoplasmic male sterile (cms-T) maize (Duvick et al., 1961; Wise et al., 1996). cms-T plants do not produce viable pollen. When pollen from a plant harboring Rf1 is used to pollinate female CMS plants, fertility is restored in progeny. This genetic system was used in the 1960’s for hybrid seed production until susceptibility to the fungus Cochliobolus heterostrophus wiped out 85% of the U.S. corn production in 1970 (Wise et al., 1999). Rf1 post-processes the mRNA transcript of the novel mitochondrion gene T-urf13, truncating and reducing URF13 protein levels associated with sterility (Dewey et al., 1987; Wise et al., 1996). The molecular mechanism used by the T-urf13 protein to cause sterility and how Rf1 restores fertility by transcript modification is yet unknown.

Our 16 finished BACs from 2 contigs separated by 30 Mb on maize chromosome 3. rf1-associated contig 1 (rf1-C1) contains 11 BACs and is 961 kb, rf1-associated contig 2 (rf1-C2) contains 5 BACs and is 594 kb long. Of 12 gaps requiring extensive post-finishing assembly to close only 1 remains, caused by simple repeats and hairpin structure, we continue to attempt closure. The 11 closed gaps represented misassemblies caused by LTR retrotransposons. A common type of LTR retrotransposon gap occurred when a TE is found nested within the LTR of a previous LTR retrotransposon insertion. This causes sequence gaps in one of 3 locations in either of the 2 LTRs. Two methods of computational gap closure were designed to combat LTR retrotransposon caused gaps. The genome based approach utilizes TEnest to help identify the nesting structure of the TEs surrounding the gap and builds a sequence to span the gap region driving the BAC assembly. The sequence based
approach utilizes paired end information of plasmid subclone traces to help walk into gapped regions. The use of these two novel gap closing approaches will greatly enhance assembly of repetitive regions for the maize genome and further sequencing projects.

To provide an in-depth view of large contiguous regions of maize we conducted a comprehensive analysis of our rf1-associated sequence contigs. Based on the observed genome structure a definition of repeat clusters was developed. This allows TEs to be found inserted within gene islands, but it groups the majority of TEs within repeat clusters. Insertion ages vary within each cluster with older elements found at the top levels of TE nests. TE families are for the most part well scattered across repeat clusters, however, certain families are seen grouped together. Gene islands are observed. They contain from 1 to 9 predicted genes giving a gene density of 1 gene per 11 kb in gene islands, and 1 per 37 kb over the 2 rf1-associated contigs. Comparative sequence analysis to the rice and sorghum genomes shows small degrees of gene island colinearity, accentuating both expansion of repeat clusters and inter-gene regions. A large section of the rf1-associated region is found duplicated in the sorghum genome. This analysis of large maize contigs provides an in-depth initial view of the landscape of the maize genome.

**Comparative Sequence Analysis of Maize Lines**

Through evolution areas of genome sequence are prevented from changing over time, these conserved sequences are generally important for organism functionality. Comparative sequence analysis identifies conserved areas, assisting with locating genes, their functions, and regulatory regions. Comparative sequence analysis is a vital tool for determining evolutionary associations between both closely and distantly related species.

In maize, comparative sequencing has shown a varying degree of conservation. Between grass species macro-colinearity but micro-non-colinearity of genes is the norm. Rice as an ancestral representation of grass genomes characterizes large chromosome building blocks for construction of subsequent grass genomes (Moore et al., 1995). Within these blocks gene content and order is retained (Ahn and Tanksley, 1993; Devos et al., 1994; Gale and Devos, 1998). However, a closer look at small sections of conserved sequences show few genes truly retain this colinearity, many genes having been newly introduced, deleted, or translocated (Tarchini et al., 2000; Lai et al., 2004). Between lines of maize
micro-colinearity of genes is more well conserved but gene movement still exists (Fu and Dooner, 2002; Song and Messing, 2003).

The causes of differing genome lengths have been a long standing question of grass genomes. LTR retrotransposons have been shown to be a major cause of this sequence expansion, mainly inserting into and enlarging intergenic regions (SanMiguel et al., 1996). Between species genome regions surrounding gene islands are populated with differing clusters of TEs. Between maize lines repeat clusters are both conserved and replaced for differing clusters (Fu and Dooner, 2002; Brunner et al., 2005). High amounts of TE insertions within grass genomes also introduce high potential for recombination between genome regions inducing even more variation (Vicient et al., 1999; Devos et al., 2002).

With comparative sequence SNPs between maize can give a great resource for gene mapping and mapping genes across maize lines. SNP studies in maize have shown very consistent frequencies; 1 SNP per 28 bp over genomic sequence (Tenaillon et al., 2001), 1 SNP per 129 bp in coding regions (Ching et al., 2002; Rafalski, 2002; Batley et al., 2003), shown in both genomic and EST sequencing. Using 454 pyro-sequencing SNP frequencies have been more conserved with 1 SNP per 300 bp (Barbazuk et al., 2007).

To provide a comparative sequence analysis across a large sequenced region of maize we have sequenced 2 BACs of Wf9-BG complementing our rf1-associated contig 1 of B73 (EF517601, 961 kb) (Kronmiller et al., 2008). Two overlapping BACs total 201 kb in length corresponding to 412 kb to 533 kb of rf1-C1. The Wf9-BG contig contains a gene island encompassed by repeat clusters. The gene island contains four genes corresponding to genes 15, 16, 17, 18 of rf1-C1, the gene island in B73 contains a fifth gene (gene 19) that is not present in Wf9-BG. The predicted protein of Wf9-BG gene 1 (B73 rf1-C1 gene 15) has one amino acid altering mutation changing a glutamine to a histidine.

The Wf9-BG left side repeat cluster is well conserved to B73, with one difference of LTR retrotransposons, a Prem insertion in Wf9-BG. The gene island has 2 TE insertion differences, inserted in B73 an expanded Cassandra LTR retrotransposon caused by unequal sister chromatid recombination or inter-element illegitimate recombination, and an Opie solo LTR in Wf9-BG. The right retrotransposon clusters are not shared between B73 and Wf9-BG. Extremely close to the Opie solo LTR a recombination has altered the alignment off the
two genome sequences. SNP frequencies seen here are much more conserved than previously reported in maize, 1 SNP per 244 bp over the aligning sequences, 1 SNP per 833 bp in predicted gene exons, showing frequencies of SNPs vary according to relationships of maize lines. However, leading up to the observed point of recombination SNP frequencies drastically increase. Evidences of recombination, micro-colinearity differences, and differing frequencies of SNPs illustrate the variability seen with maize comparative sequencing.

**Literature Cited**


variation in *Gossypium*. *Genome Res*, 16(10), 1252–1261.


CHAPTER 2. TENEST: AUTOMATED CHRONOLOGICAL ANNOTATION AND VISUALIZATION OF NESTED PLANT TRANSPOSABLE ELEMENTS

A paper published in Plant Physiology

Brent A. Kronmiller and Roger P. Wise

ABSTRACT

Organisms with a high density of transposable elements (TEs) exhibit nesting with subsequent repeats found inside previously inserted elements. Nesting splits the sequence structure of TEs and makes annotation of repetitive areas challenging. We present TENest, a repeat identification and display tool made specifically for highly repetitive genomes. TENest identifies repetitive sequences and reconstructs separated sections to provide full length repeats and, for long terminal repeat (LTR) retrotransposons, calculates age since insertion based on LTR divergence. TENest provides a chronological insertion display to give an accurate visual representation of TE integration history showing timeline, location, and families of each TE identified, thus creating a framework from which evolutionary comparisons can be made among various regions of the genome. A database of repeats has been developed for maize (Zea mays), rice (Oryza sativa), wheat (Triticum aestivum), and barley (Hordeum vulgare) to illustrate the potential of the TENest software. All currently finished maize BACs totaling 29.3 Mb were analyzed with TENest to provide a characterization of the repeat insertions. Sixty-seven percent of the maize genome was found to be made up of TEs; of these, 95% are LTR retrotransposons. The rate of solo LTR formation is shown to be dissimilar across retrotransposon families. Phylogenetic analysis of TE families reveals specific events of extreme TE proliferation, which may explain high quantities of certain TE families found throughout the maize genome. The TENest software package is available for use on PlantGDB under the tools section (http://www.plantgdb.org/prj/TE_nest/TE_nest.html), and the source code is available from http://wiselab.org/.

INTRODUCTION

Transposable elements (TEs) are mobile DNA found throughout eukaryotic organisms. Although abundance is extremely high in some organisms, little is known about the processes governing the distribution of TEs across the genome. Each classification level of TE may exhibit a different genetic makeup, different modes of replication, or a preference for different genomic habitats. By the nature of their mobility, TEs have the potential to induce change throughout an
organism’s genome. As a consequence of multiple TE copies, unequal cross-over and recombination can occur between chromosome regions. TE insertions can cause gene or regulatory mutations, altering levels of transcripts or provide new genetic material for novel gene functions to evolve (Kidwell and Lisch, 2000). TE genes may be recruited by the host organism for cellular functions, and can serve as transportation systems for genes to new genomic locations (Lal et al., 2003).

Abundance of TEs vary widely across different organisms. Human (Homo sapiens) DNA is composed of 45% (Lander et al., 2001) repetitive sequences, Drosophila melanogaster is 3.9% (Kaminker et al., 2002), and maize is 67% (Haberer et al., 2005). Even closely related organisms can have vastly different amounts of repetitive elements, for example, rice (Oryza sativa) is 35% repetitive (IRGSP, 2005) compared to 67% for maize (Zea mays). Classes of TEs also vary between organisms, the ratio of long terminal repeat (LTR) retrotransposon/non-LTR retrotransposon/DNA transposon of human repetitive DNA is 8.3/74.5/2.8, while Drosophila melanogaster is 68.6/22.5/8.0, and maize is 94.4/0.1/1.9, respectively. The high amounts and different repetitive makeup of genomes coupled with their potential for inducing evolutionary changes make the annotation of TEs in DNA sequences crucial to decipher genomic processes. Annotation of TEs goes beyond identification of repetitive sequence sections. Many TE sequences are historical artifacts of past replications, and have become inert or truncated by evolution. Reconstruction of these past insertion events can not only show how genes and regulatory regions have altered, but shed light on the evolutionary dynamics of the entire genome. Based on insertion order and calculation of age of insertion, complete annotation of TEs can provide a historical chronology of a genomic region. A genome-wide annotation of TEs can also provide insight to repeat biology, including insertion site preferences, family distribution, and differences in repetitive density.

High quantities of TEs, especially the LTRs of retrotransposons, greatly impede sequence assembly as well as genome annotation (Rabinowicz and Bennetzen 2006). As subsequent TEs integrate into a clustered location there is a high likelihood the repeat will insert within the boundaries of existing elements. This incident, nesting of one element within another, seen on a small scale in some organisms (Quesneville et al., 2005) is widely observed throughout the grass genomes (SanMiguel et al., 1996; SanMiguel et al., 1998). When TEs nest within one another the existing repeat is fragmented by the sequence of the inserting element. Successive insertion events nested within a cluster will create highly fragmented sequences; correct identification of the repeats in nested groups requires reconstruction of the original sequences.

Current repeat annotation tools have not adequately addressed the issue of nested TEs and are unable to rebuild fragmented elements. Three distinct methods of TE identification have been
developed. RepeatMasker (Smit et al., 1996–2004), uses a repeat database to locate sequence matches. This provides correct identification of fragmented TEs in nested repeat clusters, but reconstruction of whole TEs and evolutionary timeline of insertions is not possible. LTR retrotransposon detection software such as LTR struct (McCarthy and McDonald, 2003; Kalyanaraman and Aluru, 2005) groups LTR pairs based on sequence alignment identity. With LTR pair locations one can infer a general retrotransposon insertion order, however nested repeats are not specifically addressed and an LTR broken from subsequent insertions will not be identified. Furthermore, LTR retrotransposon identification software is unable to identify internal regions of retrotransposons, and will not locate non-LTR retrotransposons, DNA transposons, or other TEs. De novo TE identification software, PLIER (Edgar and Myers, 2005), RECON (Bao and Eddy, 2002), and RepeatScout (Price et al., 2005), has the ability to locate and classify previously unknown repeats based on sequence identity of repeated regions. These programs do not have the ability to reconstruct fragmented TEs or provide insight to the genomic evolutionary process.

In order to fully analyze repeat dense grass genomes, we have developed TEnest. Using a community updated repeat database of LTR retrotransposons, non-LTR retrotransposons, DNA transposons, and other repetitive elements, TEnest will identify all TE insertions in the input sequence. With additional repeat database construction TEnest will annotate TEs in any organism’s genome. For LTR retrotransposons, TEnest will identify the two flanking LTR sequences and calculate the time since insertion based on the rate of mutation accumulation in repetitive sequences of grasses (1.3 x 10^-8) (Kimura, 1980; Ma and Bennetzen, 2004). TEnest identifies the internal regions of the TE insertion and for all repetitive element types will reconstruct sequence fragmentations caused by nesting of TEs. TEnest outputs the coordinate locations of TEs identified as well as a publication quality graph representing the chronology of the DNA sequence.

Recent evidence suggests the high percentage of repetitive elements, especially LTR retrotransposons in maize are due to the replication activities of just a few element families (Meyers et al., 2001). Here, analysis with TEnest shows the retrotransposons Ji, Opie, and Huck of maize each make up 11-13% of the total genome sequence. A single TE that rapidly replicates throughout the genome can have significant consequences not only on the evolution of the TE family, but on the whole genome. TEnest gives the user the ability to reconstruct the ancient TE insertions to their pre-nested sequence states, allowing a phylogenetic analysis upon all the existing members of the TE family throughout the evolutionary history of the organism. A detailed phylogenetic analysis of each high copy TE family, associated with each element’s time since insertion, suggests there have been isolated events of extreme TE proliferation across the genome, allowing Ji, Opie, and Huck to
replicate seemingly without restraint. In addition to reconstructing ancient fragmented TEs, TEnest also identifies solo LTRs of retrotransposons formed by unequal recombination. Solo LTRs found throughout the maize genome are shown to be inconsistent across both retrotransposon families, inconsistent with TE length or LTR length.

Throughout this manuscript we follow the TE nomenclature format outlined in Wicker et al. 2007. TEs are hierarchically classified into 6 divisions from class through subfamily. Classifications most used in this manuscript are class, separated into retrotransposons and DNA transposons; superfamily, including *Gypsy* and *Copia* retrotransposons; and family, individual TEs grouped together by sequence similarity. Throughout this text superfamily and family names are italicized (Wicker et al., 2007).

**RESULTS**

The nested TE identification software package TEnest has three sections for use in genome sequence analysis: the organism specific repeat databases; TEnest, a program for identification of TE coordinates; and svg_ltr, a graphical display program for visualization of TE insertions.

**TEnest uses a Repeat Database Kept up to Date with new Sequences and User Input**

The repeat databases are kept up to date by two methods. First when new maize genomic contigs are completed, they are entered into PlantGDB (http://www.plantgdb.org) (Dong et al., 2004) and a TEnest insertion graph is produced. This triangle insertion graph is examined for unidentified or fragmented nested insertions within TEs. These insertions are compared against a set of potential new TEs to determine if it has been previously characterized. A similar process will be implemented for wheat sequence contigs in the TriAnnot wheat annotation pipeline (urgi.versailles.inra.fr/projects/TriAnnot). In maize, several TEs have been identified by this process including *Danelle* (GenBank accession: EF562447), *Stella* (GenBank accession: EF621725), *Tavish*, *Tenzig*, *Klaus*, and *Hodge*.

Second, users of TEnest on PlantGDB can update the repeat databases with newly identified TEs. This submission system requires information about the TE such as the organism, TE classification, sequence locations of identification, and the proposed name. The new TE is aligned to known TE families in the organism and flagged for manual review, when review is complete, users will be notified of the status of their TE submission. TEnest users can also use this submission system to suggest revisions or repairs to TE database entries.

**Annotation of Transposable Elements Using TEnest**

With use of the plant repeat databases, TEnest identifies all TE insertions in the input sequence, reconstructs fragmented elements, and determines age of insertion for LTR
retrotransposons producing a list of coordinates for each TE sequence location. TE insertions are classified as one of four data types: SOLO, corresponding to solo LTR sequences; PAIR, right and left LTRs of a LTR retrotransposon grouped by base pair similarity and the corresponding internal sequences of the TE; NLTR, full length TEs of classes not containing LTRs (non-LTR retrotransposons and DNA transposons); and FRAG, partial sequences of the NLTR class or internal fragmented regions of LTR retrotransposons.

(1) Identification of Retrotransposon LTR Sequences

Throughout the TEneest process a two alignment approach is used to quickly identify exact coordinate locations of TE alignments. First WU-BLAST blastn (Gish, 1996-2004) is used to rapidly identify possible TE sequence regions, then FASTA LALIGN (Huang and Miller, 1991) is used to retrieve the exact coordinates of each possible TE type in these regions. The TEneest process begins by identifying LTR sequences within the input sequence. Users can select or de-select specific TEs to include in the analysis. LTR database sequences are aligned to the input sequence with WU-BLAST blastn. If an alignment is found the coordinates are retrieved, expanded on either side (by the size of the matching LTR), and excised from the input sequence. This set of excised sequences will contain short incomplete matches to LTRs of multiple TE families. Each excised LTR sequence is sent to the FASTA LALIGN process where a pairwise local alignment between it and the database LTR is performed. If this alignment provides a passing score (default of E-value $10^{-20}$), this coordinate set is entered into the alignment list, which now contains full length annotations over the extent of the identified region.

Each pairwise LTR alignment coordinate set is entered into the recombination process where a powerset algorithm is used to rejoin separated LTR sections. A powerset is the set of all the subsets of a set (Suppes, 1960). For the TEneest algorithm, the original set is the separated LTR sections, the powerset is the list of all subsets of the separated LTR set. The powerset recombination process is based on coordinates of the matching database LTR (also referred to as the ‘subject’ of the alignment, while the input sequence is the ‘query’), and does not allow overlapping sequence regions to result in joined sections containing duplicated sequence regions. For example, three separated LTR sections are entered into the powerset recombination process; section A with LTR based coordinates of 1 to 350, section B with coordinates of 50 to 400, and section C with coordinates of 351 to 500. The powerset of the separated LTR set is (\{\}, \{A\}, \{B\}, \{C\}, \{A,B\}, \{A,C\}, \{B,C\}, \{A,B,C\}). Sets with overlapping LTR based coordinates are not allowed, this removes \{A,B\}, \{B,C\}, and \{A,B,C\}. From the remaining five the set with the largest sequence length is returned, each LTR section must be
returned exactly one time. In this case two sets are returned, sets \{A,C\} and \{B\}, with lengths of 500 and 350 basepairs.

LTR retrotransposons replicate into new locations across the genome by means of reverse transcription and integration. A seven step process produces an exact DNA intermediate of the retrotransposon with one exception; the two new LTR sequences are reverse transcribed from an intermediate LTR, which itself is a unique LTR sequence formed from the combination of the two original LTRs of the parent retrotransposon. This results in a new integrated retrotransposon with identical LTRs (Boeke and Corces, 1989). Over evolution, these LTRs separately acquire mutations, but over a stretch of sequence the two LTR copies from a retrotransposon insertion will be more alike than LTRs from different insertion events. Thus, sequence similarity can be used to pair LTRs for reconstruction of whole LTR retrotransposons.

For each TE family each LTR alignment returned from the powerset recombination process is excised from the input sequence and joined into a single contiguous sequence. By TE family, the LTR sequences are locally aligned to each other and the base pair substitution rate (BSR) is determined. Any insertion, deletion, or substitution of any length is scored as a single substitution for the alignment. BSR is calculated by the total substitutions divided by the alignment length of the two LTR sequences. LTR sequences are grouped according to the smallest BSR. The two paired LTRs are classified in the PAIR data type, any LTR sequences not paired are assigned to the SOLO class. LTRs can be found in a solo configuration throughout the genome. There are three possibilities where TEnest is unable to assign a LTR to a pair: 1) The second LTR sequence is missing, it is either found off the end of the contig or in a sequence gap. 2) The LTR’s true partner was incorrectly paired with another LTR. Although unlikely, if two LTRs from different retrotransposon insertions have evolved to be more similar than those from the same insertion, incorrect LTR pairing can occur, and possibly cause solo LTR identification. Any such occurrences of incorrect LTR pairing are resolved by TEnest by the discrepancy function discussed below. 3) The solo LTR is the result of homologous unequal recombination which has caused deletion of the internal retrotransposon region and one, leaving just a single LTR.

(2) Identification of Retrotransposon Internal Regions

The sequence within the boundaries of each paired LTR set is examined for the internal regions of a LTR Retrotransposon. The LTR pairs are grouped according to nesting sets, each first level LTR pair found inserted directly into the original, pre-TE insertion DNA sequence is grouped with any subsequently inserted LTR pairs found nested within the first level LTR pairs. Each LTR pair grouped in each nesting set is examined, the smallest LTR pair coordinate distance first, for
internal sequence locations. As in the LTR identification process described above, a similar two alignment method is performed to quickly and accurately identify the exact coordinates of the LTR retrotransposon internal regions. First, WU-BLAST blastn is used to rapidly identify potential locations; second, LALIGN makes a pairwise alignment of all identified regions to determine exact coordinates; third, the powerset recombination process reconstructs separated regions into the final internal middle region.

However there is one significant difference between the alignment method here and the previously described LTR identification process. First, the paired LTRs are arranged by smallest sequence spanning length first, after each paired LTR is processed the identified regions are ignored by subsequent alignments. Second, during the alignment process the sequence database of the initial WU-BLAST alignment contains only one TE sequence, the TE type corresponding to the paired LTRs. These restrictions give TEnest the ability to correctly annotate the entire internal regions of nested LTR retrotransposons. The identified internal regions are added to the PAIR data type signifying classification as whole LTR retrotransposons.

(3) Identification of Non-LTR Retrotransposons and DNA Transposons

At this stage the TEnest annotations consist of full LTR retrotransposons and solo LTRs. All unidentified region are examined for potential non-LTR retrotransposons or DNA transposons (NLTRs) or fragmented TE insertions of any classification (FRAGs). Unidentified regions may be found anywhere in the sequence not classified as PAIR or SOLO, and may be found inserted into the original DNA sequence or found nested within PAIR or SOLO identifications. As described above in the LTR and internal region alignment processes, the two alignment method is used to identify coordinates of TE annotations. First, WU-BLAST blastn is used to rapidly identify potential locations, then LALIGN is used to determine exact coordinates. As illustrated in Figure 1A, sequence locations are grouped prior to the recombination process, so as to allow joined sections to be separated by subsequent TE insertions, but not separated by recombined sections or PAIR LTR-middle-LTR sets. Recombined sequences are classified by sequence length to the original database sequence, if greater than an 80% match the joined sections are assigned to the NLTR data type, if less than this value the recombined group is assigned to FRAG, to signify it is a fragmented TE insertion.

The powerset recombination process of TEnest is useful for joining sections separated by nesting of subsequent TE insertions, however this process can run into problems. Although uncommon, recombined TE sections are susceptible to coordinate discrepancies, defined as a re-joined sequence set whose grouping configuration disagrees with another re-joined set. This is seen
Figure 1. Recombination of Separated TE Sections. (A) Recombination rules for non-LTR retrotransposon and fragmented transposon annotations. Separated sections of a single type of TE insertion are shown in blue, other TE insertions are shown as green, red, or yellow. Recombination allows sections separated by subsequent insertions to be rejoined, but does not allow sections to join across previously recombined groups. With correct TE insertion based coordinates fragments 2 and 3 will form a group, and fragments 4, 5, and 7 will join as a recombined group. Fragments 1 and 6, and the joined sections 2-3 and 4-5-7 cannot join any fragments. (B) Coordinate discrepancies of recombined sections. Recombined sections are shown with dotted areas across separated sections. These cases can be caused by local inversions, sequence assembly errors, or incorrect joins from the recombination process. A weighted ratio of number of disagreements, sequence alignment identity, and sequence length of annotations is used to determine which TE recombined discrepancy is removed.

when sections from two or more recombined TE annotations are found in alternating orders across the input sequence (Fig. 1B) as opposed to nested within one another. While a biological process such as local small inversions can explain such occurrences a nested insertion display is unable to represent disagreeing re-joined sections. TEnest therefore assumes the discrepancies are caused by either incorrect powerset grouping or incorrect LTR pairing. To resolve each recombination discrepancy each TEnest data type (PAIR, SOLO, NLTR, FRAG) is self checked and each combination of data types are checked for possible coordinate discrepancies by TEnest. Any discrepancies found are scored based on alignment identities, sequence length percentage of the whole TE, and number of discrepancies; the joins of those with the worst discrepancy scores are broken to split combined sections into separate groups.
(4) TEnest processing time is decreased with use of multi-processors and clustered computers

The time required to complete a TEnest run is dependent on the amount of TEs contained within the sequence. As longer plant sequence contigs are produced the amount of TEs found in the sequence will also increase, and the time required for TEnest runs will grow proportionally. To address increasing sequence lengths TEnest has been developed with multi-processor ability and can be run on clustered computers with use of an included perl script. An average sized BAC of 164 Kb (GenBank accession: AC148161) containing 11 PAIRs, 1 SOLO, 5 NLTRs, and 30 FRAGs takes 5 minutes, and the currently largest maize contig of almost 1 Mb (GenBank accession: EF517601) containing 46 PAIRs, 6 SOLOs, 36 NLTRs, and 155 FRAGs takes 45 minutes for a TEnest run using a dual 3.2 GHz desktop PC with 4 Gb RAM. Time intensive sections of the TEnest algorithm are broken out to different processors for five subroutines; LTR alignment, LTR powerset, LTR BSR calculation, PAIR internal sequence detection, and FRAG/NLTR detection.

In addition, to make TEnest a viable resource for chromosome sized maize pseudo-molecules, a TEnest wrapper script, clusterTEnest.pl, has been developed. This script will take a large input sequence and split it into user defined lengths, send each section to a separate node of a clustered computer to run several instances of multi-processor TEnest simultaneously. Once each split sequence is complete the annotation results are regrouped, and the identified TEs are removed from the input sequence. A final TEnest is run on the full sequence ultimately providing the same output as an original TEnest submission. This split function decreases process time for long sequences and decreases incorrect LTR BSR pairing that may be found when analyzing a large number of a retrotransposon type. For example, when the same 1 Mb contig (GenBank accession: EF517601) was split into 100Kb segments and sent to 10 nodes of a clustered computer (each with dual 3.2 Ghz, 4 Gb RAM), it took 35 minutes to complete. The benefit of clusterTEnest.pl will increase with longer sequence contigs.

Visualization of Nested Transposable Element Structures with svg_ltr

TE insertions identified by TEnest are visualized in a triangle insertion graph with the program svg_ltr (Fig. 2). svg_ltr uses the coordinate table of identified TE locations from TEnest Supplemental Fig. 1, 2, 3, 4) to produce the main output of TEnest. The nesting display graph represents the original DNA prior to repeat insertions as a black horizontal line and the TE insertions (within it as triangles. The horizontal top of a triangle TE insertion corresponds to the length of the TE insertion, the bottom point shows the insertion location. The genome distance between any two points on a triangle graph is determined by the addition of all horizontal lines (including the black
Figure 2. TEnest Graphical Display: Barley, Maize, Rice, Wheat. TEnest graphical output examples; barley AH014393 (Caldwell et al., 2004), maize AC145481, rice AP004818, wheat DQ537335 (Gu et al., 2006). The original, pre-TE insertion DNA sequence is shown as a black horizontal line. TE insertions are shown as colored triangles, names for each type are shown below in the legend. LTRs of LTR retrotransposons are shown as black arrows at the top of the triangle. White areas within triangles are unique or unidentified insertions within the TE. These can possibly correspond to new, unidentified TEs. Insertion age of LTR retrotransposons are shown in Mya inside the retrotransposon triangle. Features including annotation type, white areas, Mya calculation, and coordinate display can be optimized and selected/deselected by the user.

DNA representation and TE triangle tops) on all levels between the two points. Spacing and alignment of triangles are adjusted to prevent overlapping triangles; however TE triangle insertion point locations are preserved to show the true location of a TE. Triangle color corresponds to TE type, shown in the legend at the bottom of the display.

Several functions are included with svg_ltr to produce graphs containing the information needed by the user. Display of data types SOLO, PAIR, NLTR, and FRAG can be toggled on or off. Arrows representing location and direction of LTRs can be shown at the top of the triangle. Either base substitution rate (BSR) or Millions of years ago (Mya) can be displayed inside a box within a LTR retrotransposon triangle, (see Figure 3 for calculation of BSR and Mya). Coordinates corresponding to either the TE or the input sequence can be displayed at the top of the TE triangle for each section of the recombined TE group. Insertions found within a TE that do not align to a TE found in the repeat databases can be omitted from the display, this is shown by a white triangle within the TE triangle.

In addition, several functions are included with svg_ltr that make this a stand alone program for display of sequence region annotations. svg_ltr can also display two user inputted data types, GENE corresponding to gene annotations, and PSDO corresponding to pseudo-gene annotations. Both are displayed along with the TE annotations in the svg_ltr display, GENE and PSDO are shown as rectangular regions with direction indicating arrows, and can show separated sections such as multiple exon genes. A script, checkTE.pl, included with TEnest, is provided to assist users in adding this additional information to the svg_ltr input file by converting to and from both the TEnest coordinate table output and a generic feature format (GFF3) table.

Computational Validation of TEnest Repeat Annotations

To assess the repetitive element identification capability of TEnest, three verification experiments were conducted. (1) The TEnest outputs of high density repetitive regions of maize and rice were compared to GenBank submitted TE annotations. (2) Permutations of TE annotations
were evaluated to exclude confounding TE annotations. (3) Simulated genomic maize sequences were made with TE insertions and examined with TEnest.

(1) Comparisons of Maize and Rice to submitted GenBank Annotations

To evaluate the accuracy of TEnest, curated maize and rice GenBank sequence contigs were compared to their TEnest outputs. Sequenced contigs were chosen that contained repeat region annotations and had 5 or less sequence gaps were chosen. Rice was selected due to its phylogenetic similarity yet differences in abundances of repetitive sequences and its complete sequence and annotation. Rice and maize have vastly different TE class proportions; rice retrotransposons make up 19.3% and DNA transposons make up 13.0% of the genome (IRGSP, 2005), while maize is composed of 63.3% retrotransposons and 1.3% DNA transposons. However, structures between element classes are similar, and some TE families found between the organisms are closely related. Eight maize (Tikhonov et al., 1999; Song et al., 2001; Fu and Dooner, 2002; Fu et al., 2002; Brunner et al., 2005) and two rice (Nagano et al., 2002) annotations were graphed with svg_ltr and visually compared to the TEnest annotations (Table I).

![Figure 3. Chronology of Maize Whole LTR Retrotranposon Insertions. Differences in paired LTR sequences are used to calculate the time since insertion in Millions of years ago (Mya). Mya is calculated by BSR divided by two times the substitution rate in repetitive regions of grasses (1.3 x 10^{-8}) (Kimura, 1980; SanMiguel et al., 1998; Ma and Bennetzen, 2004). BSR is determined by amount of mutations between the pairwise alignments of the left and right LTRs divided by the length of the LTR. Mutations are scored by counting each incidence of an insertion, deletion, or substitution between the two LTRs. Insertions or deletions of single or multiple bases are scored as one mutation event. Of 1,456 maize LTR retrotransposon insertions 50% are less than 0.875 Mya, 75% are less than 1.5 Mya.](image-url)
For maize, TEnest identified every annotation formerly found by the original curators. In addition, TEnest identified many LTR retrotransposons, solo LTRs, fragmented repeats, DNA transposons, and other repeats not found in the original annotations. For rice, TEnest found all but 7 MITE insertions originally identified by the initial annotations. These missing MITEs were truncated and below the default cutoff values in TEnest, additional runs with altered parameters to allow smaller sequence alignments identified all the missing insertions. In addition, TEnest located 11 additional DNA transposons not found in the original analysis. From the analyzed rice BACs, no TEs were seen in nested configuration, compared with 70% of maize TEs found nested. In additional rice BACs analyzed nested TEs were seen at a rate of approximately 1 per 75 to 100 Kb. This rate of nested TEs is due to the low amounts of repeats in rice as well as the small average lengths of rice TE insertions, caused by the high number of MITE insertions in the rice genome.

(2) Permutations of the Maize Repeat Database

TEnest uses a TE database of representative and consensus sequences to identify repeat insertions. While alignment to this database is much quicker than to the set of all TEs identified for each family, this method can potentially introduce a notable flaw. A TE insertion with extremely similar sequence to the representative TE in the repeat database would be identified by TEnest with little to no difficulty, however a divergent TE insertion would not match as well, therefore obtaining a complete alignment of the region would be more difficult. To insure divergent insertions are identified correctly and exact TE annotations within the repeat database do not artificially influence TEnest results, sample sequences were run with TEnest using permutated repeat databases. The permuted repeat databases were used to analyze the same eight maize GenBank submitted sequences shown in the previous TEnest verification section. For each of the eight contigs, each TE insertion found in the sequence was removed from the TE family prior to the phylogenetic consensus calculation, thus removing its influence from the database. In addition, the entire branch in the phylogenetic tree of the TE family containing this element was removed from the consensus calculation to prevent biologically similar TE insertions from influencing the TEnest results. As before, TEnest found every annotation identified by the original annotations and so did not show differences in identification between the original and permutated repeat database (Table I, with “P” notations).

(3) Construction and Analysis of Simulated Maize Genomic Sequences

Simulated maize genome sequence were constructed and analyzed with TEnest to determine correctness of repeat annotations. The percentage of repeat sequence and proportion of repeat class were randomly chosen from frequencies observed from 165 maize sequence contigs. Based on
<table>
<thead>
<tr>
<th>Maize</th>
<th>Submitted Contig</th>
<th>Total Annotated GenBank</th>
<th>Total Annotated TEnest</th>
<th>Missing GenBank</th>
<th>Missing TEnest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Retro</td>
<td>LTR</td>
<td>DNA</td>
<td>Gaps</td>
</tr>
<tr>
<td>AF391808</td>
<td></td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>AF391808 (P)</td>
<td></td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>AF123535</td>
<td></td>
<td>0</td>
<td>11</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>AF123535 (P)</td>
<td></td>
<td>0</td>
<td>11</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>AF448416</td>
<td></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AF448416 (P)</td>
<td></td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AF488416</td>
<td></td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>AF488416 (P)</td>
<td></td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>AF090447</td>
<td></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AF090447 (P)</td>
<td></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Summary of verification results of TEnest compared to curator annotated GenBank submitted sequence contigs.

* Submitted BAC sequences compared to TEnest results using the general and permutated (P) maize and rice repeat databases. In the permutated database consensus TE sequences were made from multiple alignments excluding TE sequences found on the same branch or cluster of the TE’s phylogenetic analysis. This permutation removed any influence biologically similar TE sequence could have on TEnest annotations. In each case the general and permutated database gave the same result, showing TEnest results are not artificially enhanced by related TE sequence in the repeat database.

* Summary of TE annotations found by TEnest but not by GenBank submitted annotations; (Retrotransposons, solo LTRs, DNA transposons, fragment TEs).

* Summary of TE annotations found by GenBank submitted annotations but not by TEnest; (Retrotransposons, solo LTRs, DNA transposons, fragment TEs). AY664414 contains one solo LTR identified in the GenBank submission not identified by TEnest, however this solo LTR was grouped into a whole LTR retrotransposon by TEnest. The initial run of TEnest missed several annotations in both rice contigs, these were obtained in a second run with parameter alterations to allow for smaller TE sequences.
<table>
<thead>
<tr>
<th>Submitted Contig</th>
<th>Gaps</th>
<th>Total Annotated GenBank</th>
<th>Total Annotated TEnest</th>
<th>Missing GenBank</th>
<th>Missing TEnest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Retro Tpn</td>
<td>LTR Tpn</td>
<td>DNA Tpn</td>
<td>DNA Solo</td>
</tr>
<tr>
<td>AY664414</td>
<td>5</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AY664414 (P)</td>
<td>5</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AY664416</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AY664416 (P)</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AY691949</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AY691949 (P)</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Rice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP000559</td>
<td>0</td>
<td>0</td>
<td>36</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>AP002542</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>
observed individual TE frequencies, repeat families were randomly picked for each repeat class and family and randomly assigned an insertion time estimated from observed age of insertions frequencies. A list of TE insertions and associated insertion time was produced. An original DNA sequence with equal base proportions was made and each TE sequence was inserted into the DNA sequence. The insertion location was randomly determined over the entire length of the DNA sequence, continually updated by previous TE insertions. With this process, single and nested repeat insertions were obtained. At each time point random sequence mutations were made on the whole sequence length. Mutations were made based on the rate of mutation (1.3 x 10^{-8}) (Ma and Bennetzen 2004), with a probability of either an insertion, deletion, or substitution. Insertion and deletion mutations were either single base or a sequence length based on an insertion/deletion length frequency. At each time point there was a chance of a LTR retrotransposon insertion reverting to a solo LTR, this probability depended on the amount of insertions within that retrotransposon and the divergence of the two LTR copies. For LTR retrotransposons the LTR sequence was base pair mutated before insertion and used as both the left and right LTR. Thus, the retrotransposon contained exact LTR sequence copies different from other LTRs of the same TE family, random mutations provided differences between the LTR pair to accurately model divergence for insertion age calculations. Completed simulation sequences were run with TEnest, and the output was compared with expected results. Out of 100 simulation sequences that were analyzed, six incorrect annotations were observed, 0.35% of total TE insertions. Each incorrect annotation was repaired in a second TEnest run with parameter alterations allowing for more overlap between reconstructed sections.

A Case Study: Repeat Analysis of the Maize Genome with TEnest

TEnest presents a unique ability to observe TE family distributions across plant genomes in relation to age of insertion, sequence similarity, and sequence retention. At the time of submission, the GenBank sequence database contained 165 ordered and orientated genomic maize sequence contigs greater than 100 kb. This included 56 finished contigs and 109 gapped sequence submissions with sections presented in correct order and orientation relative to the genome sequence. In total, these BACs equal 29.3 Mb, or about 1% of the maize genome. This data set contains BAC clones sequenced with intentions of gene discovery as well as BACs randomly selected to survey the entire maize genome. Therefore this set of sequence contigs may be slightly higher in gene content and lower in repetitive amounts than will be observed across the entire maize genome. These 165 contigs were evaluated with TEnest to provide a broad picture of TE clusters in the maize genome.
**Distribution of Transposable Element Insertions is Unequal Across Families**

A summary of results of TE identification by TEnest are displayed in Table II, with columns for each TE general class: LTR retrotransposons, Ty1/Copia, Ty3/Gypsy, other; solo LTRs, Ty1/Copia, Ty3/Gypsy, other; DNA transposons; and unknown. The number of copies and percent of total analyzed sequence for each class is shown. The sequence percentage for each type shows the total length of TEs divided by the total length of the contig, the last column shows the entire repetitive percentage of the contig. The average TE content for the 165 contigs analyzed is 66.95%, similar to the 65.97% reported by Haberer et al. (2005) in their analysis of 100 randomly selected BACs. The full table, showing all TE insertions annotated by TEnest in each of the 165 contigs is found in the supplementary data (Supplemental Table I).

LTR retrotransposons make up 60.59% of the total maize sequence analyzed. This is divided into 37.87% whole LTR retrotransposons, 22.23% partial or fragmented LTR retrotransposons, and 0.50% solo LTR sequences. Whole LTR retrotransposons are defined as containing both flanking LTR sequences and greater than 90% of the internal region based on the TE consensus sequence. Partial LTR retrotransposons are incomplete insertions resulting from deletions or transpositions or gaps in the sequence assembly. Partial LTR retrotransposon annotations include any amount of the internal regions of TE sequences which may be reconstructed sections from later insertions, and may also include LTR sequences. In terms of sequence length of the identified TEs, partial LTR retrotransposons cover 6.2 Mb of the BAC sequence in this analysis, whole LTR retrotransposons cover 11.1 Mb, this ratio of partial to whole is 1:1.8., showing that even in fragmented TE remnants sequence structure is moderately reconstructable.

Solo LTRs are defined as annotations greater than 50% of the LTR length and not connected to internal regions of the TE. Identified solo LTR regions that equaled less than 50% of the solo LTR length once reconstructed were not analyzed here and were classified as fragmented LTR retrotransposons. Solo LTR to whole LTR ratio varies by TE family (Table III), ranging from Gyma with 2.3 whole TEs per solo LTR, to Zeon with 121 whole TEs per solo. The majority of solo to whole LTR retrotransposon ratios lie between one solo LTR to 7 to 15 whole elements, which includes members of both Copia and Gypsy superfamilies. A much higher recombination frequency (one solo LTR to 2-3 whole elements) is seen with three retrotransposons, (Gyma, Ruda, Danelle), again in both Copia and Gypsy superfamilies. Huck and Zeon show very infrequent solo LTR formation, with one solo to 70 or 121 whole elements. In addition, many TEs have no corresponding solo, and are not shown in Table III. Sequence structure may play a role in solo LTR formation, two families of retrotransposons with similar sequence have almost exact unequal recombination rates: Ji
<table>
<thead>
<tr>
<th>TE Categorization</th>
<th>TE Amount Identified</th>
<th>TE Length Identified (Kb)</th>
<th>Percentage of Sequence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LTR Retrotransposon</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole (total [c,g,o])</td>
<td>1186 [536,424,226]</td>
<td>11095.91 [4749.53,4471.18,1872.27]</td>
<td>37.87 [16.21,15.26,6.39]</td>
</tr>
<tr>
<td>Partial (total [c,g,o])</td>
<td>2278 [804,686,788]</td>
<td>6240.90 [1760.93,2566.68,1913.29]</td>
<td>21.30 [6.01,8.76,6.53]</td>
</tr>
<tr>
<td>Total LTR Retro (total [c,g,o])</td>
<td>3464 [1341,911,1012]</td>
<td>17333.88 [6510.46,7034.93,3788.49]</td>
<td>59.16 [22.22,24.01,12.93]</td>
</tr>
<tr>
<td><strong>Solo LTR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole (total [c,g,o])</td>
<td>80 [33,16,31]</td>
<td>146.50 [49.81,14.65,82.04]</td>
<td>0.50 [0.17,0.05,0.28]</td>
</tr>
<tr>
<td>Partial (total [c,g,o])</td>
<td>706 [166,111,429]</td>
<td>272.49 [82.04,26.37,164.08]</td>
<td>0.93 [0.28,0.09,0.56]</td>
</tr>
<tr>
<td>Total LTR solo (total [c,g,o])</td>
<td>786 [249,127,450]</td>
<td>418.99 [128.92,41.02,246.12]</td>
<td>1.43 [0.44,0.14,0.84]</td>
</tr>
<tr>
<td><strong>DNA Transposon</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole</td>
<td>137</td>
<td>111.34</td>
<td>0.38</td>
</tr>
<tr>
<td>Partial</td>
<td>295</td>
<td>228.54</td>
<td>0.78</td>
</tr>
<tr>
<td>Total DNA Tpn</td>
<td></td>
<td>336.95</td>
<td>1.13</td>
</tr>
<tr>
<td><strong>Unknown</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole</td>
<td>433</td>
<td>665.11</td>
<td>2.27</td>
</tr>
<tr>
<td>Partial</td>
<td>1478</td>
<td>861.42</td>
<td>2.94</td>
</tr>
<tr>
<td>Total Unknown</td>
<td></td>
<td>1526.53</td>
<td>5.21</td>
</tr>
<tr>
<td><strong>Total TEs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.3 Mb total sequence</td>
<td>19616.35 (19.62 Mb)</td>
<td>19616.35 (19.62 Mb)</td>
<td>66.95</td>
</tr>
</tbody>
</table>

Sequence length and count of all maize TE annotations found across the 165 submitted sequence contigs by TE class (or superfamily) and TEnest classification.

* Partial TEs identified may contain more than one member from a single original TE insertion. Total TE counts, the sum of whole and partial TE amounts, are not shown due to the possible inflated value seen from un-reconstructed partial TE fragments.
Table III. Rates of solo LTR formation

<table>
<thead>
<tr>
<th>TE Type</th>
<th>TE Class</th>
<th>Whole TE Amount</th>
<th>Solo Amount</th>
<th>Ratio*</th>
<th>LTR Size</th>
<th>TE size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyma</td>
<td>Gypsy</td>
<td>42</td>
<td>18</td>
<td>2.3</td>
<td>4198</td>
<td>12067</td>
</tr>
<tr>
<td>Ruda</td>
<td>Copia</td>
<td>20</td>
<td>7</td>
<td>2.9</td>
<td>1409</td>
<td>6384</td>
</tr>
<tr>
<td>Danelle</td>
<td>Gypsy</td>
<td>53</td>
<td>16</td>
<td>3.3</td>
<td>4602</td>
<td>15397</td>
</tr>
<tr>
<td>Klaus</td>
<td>Unknown</td>
<td>22</td>
<td>6</td>
<td>3.7</td>
<td>1075</td>
<td>7040</td>
</tr>
<tr>
<td>Prem</td>
<td>Copia</td>
<td>80</td>
<td>11</td>
<td>7.3</td>
<td>3246</td>
<td>6039</td>
</tr>
<tr>
<td>Milt</td>
<td>Gypsy</td>
<td>32</td>
<td>4</td>
<td>8.0</td>
<td>565</td>
<td>9189</td>
</tr>
<tr>
<td>Xilon</td>
<td>Gypsy</td>
<td>54</td>
<td>4</td>
<td>13.5</td>
<td>2950</td>
<td>12973</td>
</tr>
<tr>
<td>Ji</td>
<td>Copia</td>
<td>268</td>
<td>18</td>
<td>14.9</td>
<td>1306</td>
<td>9030</td>
</tr>
<tr>
<td>Opie</td>
<td>Copia</td>
<td>226</td>
<td>15</td>
<td>15.1</td>
<td>1251</td>
<td>8906</td>
</tr>
<tr>
<td>Huck</td>
<td>Gypsy</td>
<td>212</td>
<td>3</td>
<td>70.7</td>
<td>1713</td>
<td>14283</td>
</tr>
<tr>
<td>Zeon</td>
<td>Gypsy</td>
<td>121</td>
<td>1</td>
<td>121.0</td>
<td>698</td>
<td>7412</td>
</tr>
</tbody>
</table>

Solo LTRs of LTR retrotransposons found in the 165 maize contigs in this analysis. Solo LTR formation rates are inconsistent across types of retrotransposons.

* Ratio of full length LTR retrotransposon to solo LTR is determined by whole TE amount divided by Solo amount.

and Opie, with 64.2% sequence identity, have 14.9 and 15.1 solo to whole ratios; Danelle and Gyma, with 55.7% sequence identity, have 3.3 and 2.3 solo to whole ratios. These are considered closely related for between family comparisons, within TE families members may have less than 60% identity to each other over their entire lengths. Solo LTRs are found on average one per 156 Kb across the analyzed sequences, however sequence AC148093, located in a near-centromeric region of chromosome 4, contains one solo LTR per 16.3 Kb. This region contains only 1 solo LTR from the high rate solo LTR families (0 Danelle, 0 Gyma, 1 Ruda), rather the high amount of solo LTRs are from a variety of lower rate solo forming LTR retrotransposon families, and suggests a high level of unequal recombination in this region.

TEnest calculates time since insertion (Mya) for each whole LTR retrotransposon using sequence identity from the paired LTRs. Fourteen hundred and fifty seven LTR pairs were identified in this set of 165 maize contigs. Fifty percent of all LTR retrotransposon insertions occurred less than 0.875 Mya, and 75% of all LTR retrotransposon insertions are less than 1.5 Mya (Fig. 3). As shown in Figure 3, the age of insertion across the 4 most abundant whole LTR retrotransposons; Ji 268 copies, Opie 226 copies, Huck 212 copies, Zeon 121 copies remains constant following this distribution of insertion times. Less represented LTR retrotransposons insertions may be younger or
older than this general distribution, however too few copies are present in the analyzed data set to accurately calculate as individual families.

Evolution of Transposable Elements Across the Maize Genome shows Clusters of Insertion Ages

Forty-seven families of LTR retrotransposons were identified, 9 Ty1/Copia, 11 Ty3/Gypsy, and 27 other. The three most abundant families, (Ji, Opie, Huck) each have 15-18% of the total amount of LTR retrotransposons identified, comprising more than half of the LTR retrotransposons found. As illustrated in Figure 4, there is a considerable drop in TE abundance in the rest of the identified LTR retrotransposon families ranging from 8% to less than 1%.

Analysis of the sequence relationship between individual elements within each TE family may give insight to the evolution and expansion of TEs across the genome and may possibly explain the unequal amounts of retrotransposon families. Using the output table of TEnest, a multiple

![Figure 4](image-url)  
**Figure 4.** Maize LTR Retrotransposon Quantification. Quantity of LTR retrotransposons identified across the 165 analyzed maize sequence contigs. Three retrotransposons types; Ji, Opie, and Huck make up almost 50% (18%, 16%, 15%) of all LTR retrotransposons identified. The three most abundant types represent both Copia and Gypsy classes, and exhibit rapid proliferation in the phylogenetic analysis. LTR retrotransposons identified in this study are noted with (*).
alignment of every element insertion for each family of LTR retrotransposon was made with ClustalW (Thompson et al., 1994). A neighbor-joining phylogenetic tree was constructed with PHYLIP (Felsenstein, 2005), each member was overlaid with the calculated age since insertion (Mya). LTR pairs incomplete due to sequence gaps or other large deletions were not included in this analysis. The retrotransposon family Ji (Fig. 5A) shows distinct clustering of insertion ages in each clade of the phylogenetic tree. In addition, branching order of the tree follows the pattern of insertion age: closely related clades have similar insertion ages. Other high copy LTR retrotransposons (Huck, Opie) also follow this phylogenetic pattern (Supplemental Figs. 5 and 6).

A tree with clustered insertion ages does not follow the expected phylogenetic result of continuously replicating LTR retrotransposons, with each clade representing a distinct family sub-set replicating individually and concurrently. Here one expects a tree with a similar range of insertion ages on each phylogenetic branch. Instead a LTR proliferation is observed, where at specific

**Figure 5.** Phylogenetic Analysis of the Maize Ji and Grande Retrotransposons. Full length maize LTR retrotransposon insertions for each family were identified with TEnest, excised, aligned with ClustalW, and a neighbor-joining tree was made with PHYLIP. Time since insertion in million years ago (Mya) calculated with TEnest were overlaid for each element. (A) The Ji LTR retrotransposon shows clades of the phylogenetic tree contain elements with similar times insertion ages, shown with standard deviations. We hypothesize this is caused by a number of rapid LTR retrotransposon proliferations of the Ji element. (B) Low copy number Grande LTR retrotransposon also shows the proliferation pattern, although with fewer TE insertions this analysis is less certain.
Figure 6. Maize LTR Retrotransposon Class by BAC. Sequence content of LTR retrotransposon classes of 165 maize sequence contigs. In general, contigs located in near-heterochromatic regions of the genome have a higher Gypsy to Copia ratio. In contrast, euchromatic sequences show just a slight preference for Copia retrotransposons. However high concentrations of Gypsy retrotransposons are found throughout the genome, suggesting large oceans of retrotransposons are as important as heterochromatic regions for attracting Gypsy type elements. Near-heterochromatic BACs are defined as residing at the contig ends and centromere locations of the maize WebFPC chromosomes (Coe et al., 2002).

Times throughout the genome evolution the Ji retrotransposon family has undergone cycles of rapid expansion. This suggests multiple instances of extreme proliferation events by one or a few related members propagating many similar insertions in a small time frame. The Grande retrotransposon family (Fig. 5B) also seems to follow the proposed proliferation process, although with only 39 members in this analysis the clusters of insertion ages are less obvious. This initial evidence from low copy families excludes the proliferation process as the only explanation for the extremely high copy number of Ji, Opie, and Huck retrotransposons.

Relative amounts of LTR retrotransposon superfamilies Ty1/Copia and Ty3/Gypsy are similar, 22.66% and 24.16% respectively, while the other class is much less abundant with 13.77% of the sequence content. However the similar amount of Copia and Gypsy elements does not mean they are found equally across the genome, instead they correspond to genome locations. In general, Copia and Gypsy sequence quantities per location are inversely proportional (Figure 6). With the
maize WebFPC July 19, 2005 release (Coe et al., 2002; Wei et al., 2007), general designations of near-heterochromatic or euchromatic were given to each BAC in the analysis based on proximity to centromeric and telomeric marker locations in the WebFPC BAC assembly. In this analysis, the term near-heterochromatic simply designates a BAC within the FPC contig most near the centromere or telomere, in reality few of these are truly heterochromatic BACs. No significant differences are seen between Gypsy and Copia superfamilies across near-heterochromatic or euchromatic chromosomal locations. Near-heterochromatic regions do tend to have a higher concentration of Gypsy retrotransposons, but a large percentage of both Copia and Gypsy retrotransposons are found in euchromatic and near-heterochromatic locations of the genome. In addition, chromosome location does not have an impact on total TE content found in a BAC. In general low repeat areas are found in euchromatic regions, but highly repetitive BACs can be found throughout the genome. These results show repeat oceans are as important as centromeric regions in attracting or retaining TE insertions and presumably repeat oceans may mimic structures associated with heterochromatin and attract Gypsy elements in a similar fashion. Alternatively, this can suggest local proliferation process, where retrotransposons replicate to nearby locations inflate quantities of separate types of TEs for specific regions.

DISCUSSION

TEnest: An Efficient Algorithm for Nested TE Annotation

TEnest was initially designed for annotation of maize BAC contigs (e.g., EF517601, EF517601); the repeat database has since been expanded to include rice, wheat (Triticum aestivum), and barley (Hordeum vulgare), and potentially could include other sequenced grasses such as sorghum (Sorghum bicolor) and Brachypodium (Brachypodium distachyon). TEnest can be extended further for use in a variety of organisms, however the main advantage of TEnest over other repeat identification software is the ability to annotate nested TE insertions, primarily seen in the densely repetitive grass genomes. To evaluate other organisms users can create a custom repeat database. This custom database can be used with the downloadable version of TEnest or uploaded onto the online version. In addition, users can submit and suggest edits of TE database entries, both of these systems are in place to keep TEnest up to date as more sequence is produced and more TEs are identified.

There are several important steps of the TEnest system that give it the ability to accurately annotate nested TEs and make it a viable resource for genome repeat analysis. The two alignment method, first using a quick BLAST search to locate general regions of interest, then using a pairwise local alignment to accurately identify the complete sequence alignment, greatly increases speed and
precision when using sequences of similar identity such as repeat databases. The powerset reconstruction method builds TE segments separated by nesting using coordinates based on the TE insertion, giving TEnest the ability to recognize and correctly resolve TE families nested within themselves (a whole \textit{Ji} retrotransposon nested within another \textit{Ji}) or to identify a duplicated region within a TE (an extra portion of a \textit{Ji} found within a \textit{Ji} element). TEnest pairs the left and right LTRs of retrotransposons based on their divergence, identifying the TE family and the sequence ‘ends’ of the insertion and allows TEnest to quickly build the internal region with more relaxed criteria, therefore obtaining the complete annotation.

Joining separated sections, both the powerset reconstruction method and the LTR pairing method introduce a possibility of join discrepancies, where two or more joined regions disagree (Fig. 1B). If the reconstruction or pairing processes suggest combinations that could not have occurred by TE nesting, but would require a local inversion or translocation, TEnest uses the discrepancy process to separate the most likely incorrect join.

TEnest provides the user with three output formats; an annotation table of TE insertions with insertion ages of LTR retrotransposons, a repeat masked sequence file, and a vector format graphical display of the chronology of TE insertions. Use of these output files can assist with identification of genes and other functionally important locations, and can also answer questions regarding the sequence makeup of the genome. When used to analyze a BAC or a single sequence region, TEnest gives information about sequence structure, content, rearrangement and evolutionary dynamics of the area. Expanding this analysis to multiple regions across the genome, such as the analysis of the currently finished maize BAC contigs within this manuscript, give a more in depth example of the capabilities of TEnest. With more sequence information, comparisons between TE families and classes and evolutionary analysis of single TE families can begin to answer questions about TE evolution, replication, and their effect on the genome. The included cluster submission script that splits input sequence and runs TEnest versions on each node of a cluster give users the ability to evaluate long sequences in relatively short time frames.

Three Verification Tests of TEnest Insure Accurate Results

Three analyses were used to validate the output from TEnest. The first verification examined TEnest outputs to curated submitted maize and rice contigs. The important information from this analysis is that TEnest was able to identify all known TE insertions. TEnest did identify extra insertions, however the goals of the original curators were varied and may have intentionally not included all TEs. In addition, at the time of original annotation the community repeat databases were less complete.
TEnest uses a repeat database to identify TE insertions, this repeat database is made of consensus or representative sequences. This process could allow easy identification of similar TEs within the family, while not accurately identifying distantly related TEs within the family. In the permutation verification the TE insertion and similar TEs from the sequence contig were removed from database construction in order to show TEnest correctly identified the distantly related TEs within a family. In each case TEnest was able to correctly identify the permuted TEs showing regardless of individual TEs used in the database construction TEnest gives unbiased repeat annotation. These results are possible because of the unique processes of TEnest, specifically initial identification of paired LTRs and relaxation of internal region alignment parameters.

The final validation highlights a further resource of the TEnest outputs. Construction of hypothetical ancestral sequences, prior to TE insertions, can be constructed by removing whole TE annotations, these we believe are more accurate representations than simple masking of all repeat matching sequence. These ancestral predictions can be used for comparative genome analysis to give a cleaner assessment of the shared sequence regions between genomes or sequence regions. Timepoint sequences can be made by removing TEs inserted after a certain age.

**Phylogeny of LTR Retrotransposons across the Maize Genome**

TEs cause sequence rearrangement and recombination by insertion and translocation of their own and other sequences throughout the genome. Additionally, by their seemingly unbridled expansion of genome size, LTR retrotransposons are significant drivers of sequence evolution. TEnest was designed to quickly and accurately analyze completely sequenced genomes, and to explore how TEs affect whole genome evolution. As with smaller sequence scale analyses, TEnest can provide TE insertion locations, distributions, and insertion preferences, to show the current structure of the whole genome. But on a larger scale, it can give the whole view of each TE family evolution, the sequence divergence history of each type from a common ancestor along with its age since insertion and its genome location. Combination of TE insertion age, sequence relationships, and location in the genome can be used to investigate fundamental questions about TEs such as their rates of proliferation across the genome, their paths of replications over time, and ultimately their causes on genome evolution.

Based on data presented here, different TE families experience unequal recombination that result in solo LTRs at different frequencies. The rate of solo LTR formation does not seem to be influenced by length of either the LTR or the whole retrotransposon. Only those LTR retrotransposons with at least one observed solo LTR were included in this analysis, many retrotransposons had no solo LTRs most likely due to the limited amount of maize genome sequence
in this study, as well as low rate of recombination within the retrotransposon family. *Gyma* and *Danelle*, and *Ji* and *Opie* both share similar rates of solo formation as well as relatively similar sequence identities and suggest TE structure or sequence is an important factor for unequal recombination.

Phylogenetic analysis of LTR retrotransposon families give similar age of insertion clustered in clades of the tree. We hypothesize that this is caused by proliferation of LTR retrotransposons, where at specific time points in evolution a single or related group of elements have rapidly expanded across the genome. These rapid TE expansions could correspond to times of relaxed mutation standards such as genome duplication events or environmental stress conditions where mutations caused by TE insertions are less detrimental to the organism. Alternatively these TE proliferations could be caused by advantageous mutations in the TE sequence, allowing a TE copy to replicate across the genome. Similar proliferation style phylogenetic trees are observed across many LTR retrotransposon families, and are therefore the process is not TE specific and cannot explain differences in TE amounts. The causes behind abundances of certain TE families are due to selective processes not yet understood.

Two other hypotheses for LTR retrotransposon replication do not explain the observed trees. Continual copying of TEs in a family until a mutation prevents replication of an individual will give a tree with TEs from any clades of the family with the ability to replicate. In this scenario the phylogenetic tree has clades containing a variety of insertion ages. Alternatively, genome duplication could immediately double the amount of TEs within the genome. At tree following a genome duplication event will contain two times as many TEs with every clade, each with the same insertion age, but each clade still contains a variety of insertion ages and those TEs still able to replicate will continue to increase the age ranges. However genome duplication could play a role in the proliferation hypothesis by allowing proliferation to increase with a decreased chance of harming the genome.

**MATERIAL and METHODS**

**Construction of Consensus Repeat Databases for TEnest**

Maize, rice, wheat, and barley repeat databases were been constructed from the following sources: GIRI RepBase (Jurka et al., 2005); TIGR (maize.tigr.org); the Messing lab (Messing et al., 2004); the Wessler lab (daffodil.plantbio.uga.edu/wesslerlab); ISU Maize Genome Assembly (Emrich et al., 2004), and TREP, the Triticeae Repeat Sequence Database (http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) (Wicker et al., 2002). These databases each consist of multiple fasta file formatted repeat sequences. For each organism, each multiple fasta file was combined and exact
duplicate sequence entries were removed. Using a cutoff E-value of $10^{-200}$, each entry in the combined database was aligned with WU-BLAST blastn. Sequences that passed the cutoff value were removed from the combined database and multiply aligned with ClustalW, a consensus sequence from the multiple alignment was made and added into the combined database. Consensus sequence bases are calculated as greater than 60% of each location in the multiple alignment, any base less than 60% gives a ‘N’. Sequences completely encompassed within the consensus sequence were removed from the database, those still containing unique regions were trimmed, the aligning part removed and the unique sections added back to the database. This process of clustering and making consensus sequences was repeated while raising the cutoff value until E-value reached $10^{-50}$. Most trimmed unique repeat sequences aligned to longer TEs clustered in this process were removed from the database, any remaining were classified as potential repeats with single representatives.

Each final set of clustered repeat entries was aligned with ClustalW. Neighbor-joining trees were made using the PHYLIP package. The resulting phylogenetic trees were examined for well defined separations into sub-groups such as a tree with only two distant clades. If present, these clustered tree sections were split into subgroups of the original repeat set. Consensus sequences were made from each repeat set or each subgroup within a set. Many repeat groups contained high diversity between elements, if more than 10% of the consensus sequence was ‘Ns’ or if the sequence had stretches of 90% ‘Ns’ for more than 100 bases a consensus sequence was not used. Instead a representative repeat entry was selected for use in the repeat database from a central branch off the phylogenetic analysis.

Consensus sequences were checked against the GenBank maize database (Benson et al., 2006) and the combined repeat databases, those entries previously characterized as TE families with at least partial sequences were updated with the original nomenclature. Each consensus repeat terminus was examined for LTR sequences, if found these LTRs were added to a separate LTR database.

**User Customizable Parameters of TEnest**

Customization of TEnest runs is accomplished using the many available parameter settings. All of these parameter settings are explained in further detail in the TEnest README file found with the TEnest web service or bundled with a downloaded version. Similar parameter settings are available for each TEnest identification process; LTRs, internal retrotransposon regions, fragmented and non-LTR retrotransposon regions. Users can alter the number of pairwise alignments reported (default 7), the gap open penalty (default 30 for LTRs, 75 for others), the gap extension penalty (default 15 for LTRs, 75 for others), and the pairwise alignment E-value cutoff score (default $10^{-20}$).
Amount of basepairs to allow as overlapping when joining sections are also customizable, for pairwise alignments (default 25), or when reconstructing separated sections in the powerset process (default 30). The smallest returned reconstructed LTR (default 25) can be raised to limit unnecessary annotations, the maximum distance between powerset reconstructed sections can also be altered (default 100 Kb). The LTR pairing process can be customized with gap open (default 12) and gap extension (default 4) penalties, and amount of LTR pairs to consider (default 0.1).

TE makeup across organisms is different, some TEnest settings have proved more useful when attempting annotation on other species. Rice has a high number of small MITE insertions, TEnest has better success identifying these elements when ignoring long spanning TEs (decreasing the powerset reconstruction maximum) and allowing for smaller TE alignments (decreasing the E-value cutoff for pairwise alignments, and decreasing the size of reported sections). Some success has been seen with TEnest on non-plant species. In *Drosophila melanogaster*, the LTRs of LTR retrotransposons are very small in relation to the grass species. We have achieved TE annotations with TEnest on *D. melanogaster* sequences when lowering the LTR overlap lengths, pairwise alignment cutoffs, and LTR size cutoff.

**Required Software for using TEnest**


**Supplemental Data**

GenBank sequence submissions submitted with this manuscript; LTR retrotransposons Danelle, EF562447 and Stella, EF621725.

**ACKNOWLEDGEMENTS**

The authors thank Dr. Thomas Peterson and Dr. Dan Voytas and for critical review of the manuscript.

**LITERATURE CITED**


CHAPTER 3. COMPUTATIONAL FINISHING OF HIGHLY REPETITIVE MEGABASE SEQUENCE CONTIGS OF THE rfI-ASSOCIATED REGION OF MAIZE

A paper submitted to Plant Physiology

Brent A. Kronmiller, Karin S. Gobelman-Werner, and Roger P. Wise

ABSTRACT

The architecture of grass genomes varies on multiple levels. Large long terminal repeat (LTR) retrotransposon clusters occupy significant portions of the intergenic regions and islands of protein-encoding genes are interspersed among the repeat clusters. Questions still remain, however, regarding the detailed organization and gene content of gene islands and the distribution of transposable element (TE) clusters across the large grass genomes. In order to characterize the relationships between repeat clusters and gene islands, we present two large contiguous sequenced contigs, 961 kb and 594 kb associated with the rfI locus in the near-centromeric region of maize chromosome 3. We present two methods for computational finishing of highly repetitive BAC clones that have proved successful to close all sequence gaps caused by TE insertions in this sequence. Sixteen repeat clusters were observed, ranging in length from 23 kb to 155 kb. These repeat clusters are almost exclusively LTR retrotransposons, of which the paleontology of insertion varies throughout the cluster. Gene islands contain from 1 to 9 predicted genes, resulting in a gene density of 1 gene per 11 kb in gene islands, and 1 gene per 37 kb over the entire sequenced region. The two sequence contigs, when compared to the rice and sorghum genomes, retain gene collinearity of only 19% and 31% respectively. Collinear genes on single gene islands show that while most expansion of the maize genome has occurred in the repeat clusters, gene islands are not immune and have experienced growth in both intra- and inter-gene locations.

INTRODUCTION

With the nearing completion of the Zea mays (maize) genome (Bennetzen et al., 2001; Chandler and Brendel, 2002; Wessler, 2006) the largest and most difficult to sequence and assemble plant genome will join the ranks of organisms with fully sequenced genomes.
Maize is an important economic, agricultural, industrial, and research crop, however, with a genome close to the size of human (2.8 Gb) and it’s high percentage of repetitive elements, acquiring the maize genome seemed a daunting task. Approximately 67% of the genome is made up of transposable elements (TEs) (Haberer et al., 2005; Kronmiller and Wise, 2008) increasing the difficulty of its genome assembly (Rabinowicz and Bennetzen, 2006). Much exploratory work has gone into isolating and sequencing just the gene areas and ignoring the repetitive regions, both by methylation filtration (Rabinowicz et al., 1999; Palmer et al., 2003; Whitelaw et al., 2003) and high-C\textsubscript{0}t (Yuan et al., 2003) systems, which have assisted researchers with selecting only genic regions to sequence. These methods have proved to capture a majority of the maize genes (Fu et al., 2005), but they still have the potential to miss important sequence regions. The current genome sequencing project aims to capture the entire gene set of maize including regulatory regions. However, the current strategy will not provide a fully assembled genome, but rather assembled bacterial artificial chromosome (BAC) contigs ordered and orientated to provide complete gene rejoins that are adjacent to potentially incomplete TE clusters.

The landscape of the maize genome provides an interesting challenge for both sequencing and subsequent annotation. The increased density of long terminal repeat (LTR) retrotransposons has had a direct effect on the genome size of many plant genomes including maize (SanMiguel et al., 1996; Bennetzen et al., 2005; Hawkins et al., 2006; Piegu et al., 2006). Besides expanding genome size, LTR retrotransposons can have an impact on evolution of the species (Kidwell and Lisch, 2000). LTR retrotransposon insertions tend to form nested clusters (SanMiguel and Bennetzen, 1998) which are separated by small regions of several genes. Large nested repeat clusters consist of TE insertions inside TE sequences, expanding the repeat cluster and breaking up the sequence of the TEs found within, inhibiting repeat and gene annotation and increasing the difficulty of assembly. Full sequence completion of the repetitive regions, however, can be of great benefit to understanding the evolutionary history of the maize genome. LTR retrotransposons can provide an estimated time since insertion by calculating the divergence of their LTRs (Kimura, 1980; Ma and Bennetzen, 2004), careful sequenced assemblies of nested repeat clusters can help to
illustrate their expansion, proliferation, and evolution across the genome (Kronmiller and Wise, 2008).

Previous studies of large contiguous regions of maize have given a general view of the landscape of the genome. Unfinished sequence totaling 7.8 megabases (Mb) from chromosome 1 and 6.6 Mb from chromosome 9 show a gene density of 33 and 27 respectively (Bruggmann et al., 2006). BAC contigs ranging in size from 126 kilobases (kb) to 405 kb show a gene density of 1 gene per 19 kb, and genes found in small groups between large repeat clusters (Brunner et al., 2005). Genome-wide analysis of maize BACs has painted a different picture, while gene density of 100 random BACs at 1 gene per 44 kb was similar to the above results, genes were not observed in tight clusters (Haberer et al., 2005). When investigating gene specific areas of maize this dichotomy of gene density is also seen. Analysis of gene rich regions such as the 22-kd α zein gene family on maize chromosome 4 reveals a high density of genes, with one gene observed per 10 kb over 346 kb (Song et al., 2001). The Adh1 locus on maize chromosome 1 contains two genes across 280 kb, or 1 gene per 140 kb. Perhaps the only message learned here is the gene density across the maize genome varies to a great degree, large contiguous sequenced regions can begin to capture the true diversity of maize chromosome architecture.

In order to characterize large contiguous regions of maize sequence, we have identified and sequenced two BAC contigs from chromosome 3 of maize B73. These contigs of 961 kb and 594 kb correspond to contigs 117 and 119 respectively on maize WebFPC (Wei et al., 2007) and span regions identified with the rf1 (restorer of fertility) locus. Maize Rf1 is the nuclear fertility restorer for Texas (T) cytoplasmic male sterility (cmsT) (Duvick et al., 1961; Wise et al., 1996). When pollen from a plant harboring Rf1 is used to pollinate to female cmsT plants, fertility is restored to the progeny. Rf1 post-processes mRNA transcripts of the novel mitochondrion gene, T-urf13, directly or indirectly reducing URF13 protein levels which otherwise are associated with sterility (Dewey et al., 1987; Wise et al., 1996). The cmsT system was used in the 1960’s for hybrid seed production, but in 1970 85% of U.S. maize crops were wiped out by the Southern corn leaf blight fungus, Cochliobolus heterostrophus (Wise et al., 1999). The molecular mechanism used by the T-urf13 protein to cause sterility and how Rf1 restores fertility by transcript modification is yet unknown.
As a foundation for the isolation of the \( Rf1 \) locus, a detailed genetic map of the region was made by combining visual and RFLP markers (Wise and Schnable, 1994). This placed the location of \( rf1 \) between \( umc92 \) and \( umc97 \), near the \textit{liguleless 3} (\( lg3 \)), \textit{glossy 6} (\( gl6 \)), and \textit{ragged 1} (\( rg1 \)) loci. \textit{Mutator} lines homozygous for \( Rf1 \) were created and crossed to maize B37 to screen for transposon-tagged mutants in the \( rf1 \) locus (Wise et al., 1996). Four \( rf1 \) male-sterile mutants were recovered from a screen of 123,500 flowering plants; \( rf1-m3207 \) (\( Rf1 \) allele originally from Wf9-BG), \( rf1-m3310 \) (Wf9-BG), \( rf1-m7323 \) (Wf9-BG), \( rf1-m7212 \) (\( Rf1 \) allele originally from Ky21) (Wise et al., 1996). A 5.5 kb Mu1 hybridizing \( EcoRI \) restriction fragment was identified that cosegregated with the \( rf1-m3207 \) allele. Sequences from this fragment were hybridized to an \( Rf1 \) cDNA library and probes designed from the identified cDNA, p6140-1 (Wise et al., 1999), were found to cosegregate with the \( rf1 \) locus in over 200 progeny.

Using probes designed off the 5.5 kb cosegregating restriction fragment and the p6140-1 cDNA, we have identified two BAC contigs spanning the \( rf1 \) locus. Sixteen BACs were sequenced to completion to provide high-quality finished sequence. Here we present two methods for computational finishing of highly repetitive grass genomes, which were successfully utilized to close 11 TE induced gaps. Sixteen nested repeat clusters were found each spanning as much as 155 kb and containing a variety of LTR retrotransposons types and ages of insertion. Genes are found tightly clustered showing a dense rate of 1 gene per 11 kb within gene islands. Finally, comparative analysis to \textit{Oryza sativa} (rice) and \textit{Sorghum bicolor} (sorghum), show that while many genes are retained across all three species, genes have both been lost and translocated across the genomes.

**RESULTS AND DISCUSSION**

**Mapping, Sequence, and Assembly of Maize \( rf1 \) Contigs**

\textit{Analysis of Multiple B73 Maize BAC libraries Leads to Two Separated 1-associated Contigs}

Three segregating \( rf1-m \) families (\( rf1-m3207 \), \( rf1-m7323 \), \( rf1-m7212 \)) (Wise et al., 1996) were analyzed by the improved AIMS (amplification of insertion mutagenised sites) method (Frey et al., 1998) (see Methods). Six tightly linked and cosegregating low-copy AIMS fragments were identified from the three \( rf1-m \) allele families and used as probes.
against B73 library filters. These probes, along with probes designed off the identified 6140-1 cDNA (Wise et al., 1999) were screened against the first two maize B73 BAC library filters, Genome Systems (St. Louis) and Clemson University Genomics Institute (CUGI) ZMMBBa. This resulted in short non-overlapping BAC contigs.

BAC ends were sequenced off each identified clone and overgo probes were designed off the resulting sequences. Nine out of 20 designed overgos hybridized to low-copy sequences and were subsequently used for the next round of hybridization to the B73 BAC library filters. Identified BACs were restriction digest fingerprinted and added to the contig assembly with the previously identified BAC contigs.

After the maize physical mapping project was underway, additional BAC clones were identified by hybridization to the ZMMBBb and ZMMBBc libraries and subsequently via in silico overlaps from the maize WebFPC database (Coe et al., 2002; Wei et al., 2007) and aligned to the BAC contig assembly. Reciprocal hybridization of digested BAC clones produced increasingly contiguous BAC contig assemblies. A minimal tiling path was constructed from a total of 796 BACs from the four B73 BAC libraries. The minimal tiling path formed two contigs both located on chromosome 3. rf1-associated contig 1 (rf1-C1), and rf1-associated contig 2 (rf1-C2), correspond to contigs 117 and 119 respectively in maize WebFPC and are located in maize bin 3.04.

**Sequencing and Initial Assembly of Maize BACs**

Sixteen BAC clones were fully shotgun sequenced to provide the most accurate representation of this region. Plasmid subclone libraries were made with inserts ranging from 2.5 to 3.5 kb; subclones were end sequenced to produce an average of 8 to 9 fold coverage. The BAC sequences were initially assembled with the phred/phrap package (Ewing and Green, 1998; Ewing et al., 1998) (http://www.phrap.org) to determine the coverage condition. If the BAC assembly was highly repetitive or assembled into many separate contigs, additional plates of sequence were produced to increase the sequence depth. Once draft sequence was completed, BACs averaged 10X coverage, however this varied across BACs depending on their repeat content (Table I).

Once Phase 1 draft sequencing was complete the BAC was assembled and evaluated in order to close gaps and resolve low quality regions. To increase quality of poor regions,
Table I. Sequenced BACs across the *rf1* locus

<table>
<thead>
<tr>
<th>BAC (ZMMB)</th>
<th>Length</th>
<th>Coverage</th>
<th>BAC % Repetitive</th>
<th>BAC % Genic</th>
<th>Post-Finish Gaps</th>
<th>Computationally Closed Gaps</th>
<th>Gaps Remaining</th>
<th>Reasons for Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>rf1</em>-C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b0322P15</td>
<td>158174</td>
<td>9.17</td>
<td>80.23</td>
<td>8.02</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0452F11</td>
<td>127944</td>
<td>12.90</td>
<td>84.47</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0153H24</td>
<td>132164</td>
<td>9.38</td>
<td>74.09</td>
<td>9.74</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c0080G23</td>
<td>167218</td>
<td>8.90</td>
<td>66.66</td>
<td>10.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0211C05</td>
<td>142755</td>
<td>9.54</td>
<td>74.97</td>
<td>10.37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0331I02</td>
<td>149683</td>
<td>11.48</td>
<td>85.60</td>
<td>2.56</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>di-nucleotide gap&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>c0360P05</td>
<td>178651</td>
<td>8.16</td>
<td>82.62</td>
<td>5.77</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0348F20</td>
<td>129989</td>
<td>11.86</td>
<td>75.49</td>
<td>5.46</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>rf1</em>-C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b0345O22</td>
<td>112741</td>
<td>10.52</td>
<td>93.01</td>
<td>0.00</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c0294D02</td>
<td>171602</td>
<td>7.38</td>
<td>78.56</td>
<td>7.07</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0103L15</td>
<td>154263</td>
<td>8.41</td>
<td>85.58</td>
<td>6.03</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0622H01</td>
<td>156570</td>
<td>10.39</td>
<td>80.46</td>
<td>8.72</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>b0335C07</td>
<td>132421</td>
<td>12.36</td>
<td>84.13</td>
<td>4.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>To not count sequence from overlapping BACs twice, information presented here is calculated from the start of a BAC sequence to the start of the next overlapping BAC sequence.

<sup>b</sup>Four areas of small nucleotide repeats causing the gap: GA x 300bp, GAP, GA x 400bp, TTAGGG x 620bp, AT x 50bp
low quality and failed subclone sequences were identified for re-sequencing. If low quality was due to the DNA structure (hairpin folding) or difficult sequence (mono/dinucleotide strings) subclone sequences were identified and re-sequenced with alternate sequencing chemistries. To close gaps, sequencing primers were designed and sequenced off the subclone and BAC template in order to walk in the direction of the gap. For larger gaps, PCR primers were designed surrounding the area and amplified to make template for sequencing into the gap. Entire plasmid subclones and PCR products were identified that spanned the gap regions and other unsequenceable areas and fully sequenced with transposon bombing insertion methods (Kimmel et al., 1997).

At this stage BAC assemblies were as close to best possible condition that finish sequencing could bring (Table I, column ‘Post-Finish Gaps’). Remaining gaps were closed with computational methods. BAC sequences were assembled with two programs, CAP3 (Huang and Madan, 1999) and phrap. Incomplete regions were examined in order to present and submit completely finished BACs. Twelve gaps made it to this stage of computational finishing. Eleven of these gaps were caused by LTR retrotransposon misassemblies, one was caused by long strings of dinucleotide/gap/dinucleotide/hexanucleotide/dinucleotide simple polymer repeats. By careful analysis of the repeats, identification the retrotransposons, their associated LTRs and their nested structure, and mapping of paired end sequences of plasmid subclones, we were able to determine the correct sequence of all of the retrotransposon caused sequence gaps, however the correct sequence for the lone simple repeat gap remains elusive (Table I, column ‘Gaps Remaining’).

Finished BAC clones were verified with restriction digest analysis. HindIII restriction digests were compared to in-silico digestion of finished sequence files (Marra et al., 1997). Any discrepancies found between the two digestions were re-examined for sequence misassemblies. For non-gap regions, basepair quality is well within sequencing standards with less than 1 error in $1 \times 10^5$ per BAC assembly. In the minimal tiling path BACs average 32 kb of overlap, although the area of the overlap between ZMMBBb0211C05 and ZMMBBb0331I02 multiple BACs were sequenced to resolve mapping discrepancies. Fully assembled, the rf1-C1 is 961 kb, the rf1-C2 is 594 kb, and they have been submitted to GenBank (Benson et al., 2006) as EF517601 and EF517600, respectively (Figure 1).
Figure 1. Combined Genetic and Physical Map of Maize Sequence Contigs. Generic Genome Browser (GBrowse) display of the rf1 BAC contigs of maize chromosome 3 showing BAC path, predicted genes, and annotated TEs. rf1-C1 is 961 kb and contains 11 repeat clusters and 32 predicted genes. rf1-C2 is 594 kb and contains 5 repeat clusters and 10 predicted genes. The two BAC contigs are separated by approximately 30 Mb. One gap remains, caused by dinucleotide and hexanucleotide polymer repeats, this is shown as a red oval on BAC ZMMBBb0331102, found at approximately 607 kb on rf1-C1.
Characterization of Repetitive Gaps in Maize Sequence Assembly

In particular, two methods proved very useful to resolve maize sequence gaps unenclosable with traditional laboratory based finishing methods. Eleven gaps in the BAC assemblies were closed with purely computational methods. Two cases of gap causing misassembly were found to be common in maize BACs, both involving the duplicated regions of LTRs of retrotransposons. The first misassembly type is much like any misassembly caused by a duplicated area within a BAC assembly; the traces for one LTR all assemble into the second copy, breaking the sequence of the first LTR and causing a gap. This was seen most often in TEs with long LTRs where the whole sequence trace or even both end sequences from an entire subclone were within the LTR boundaries. This was also commonly seen on LTR retrotransposons with a recent age of insertion, fewer polymorphisms introduced over the time since insertion between the two LTRs caused more assembly confusion.

The second common case of misassemblies was also caused by the LTRs of retrotransposons, seen when a LTR retrotransposon nested into one of the LTRs of an existing LTR retrotransposon. In this type the gap can be found in either of the two LTRs of the first retrotransposon (Fig. 2A). Once this insertion occurs the sequence of one LTR is interrupted with the sequence of the nested transposon. To cause a gap, during assembly the sequence from the complete LTR incorrectly aligns to both LTR locations removing the join between the interrupted LTR and the nested TE. This recruitment causes a gap, now one or both of the contig ends that point into the gap have assembled traces belonging to the other LTR (Fig. 2B) and can cause one of two gaps in the nested LTR, or one gap in the un-nested LTR of the original LTR retrotransposon.

The closure of the final unfinished gap, found in ZMMBBb0331I02 (Table I) has been hindered by long strings of simple repeat sequences. Simple repeats, such as homo-nucleotide polymers (AAAA), dinucleotide polymers (GAGAGA), or even larger repeated segments inhibit through sequencing by allowing the DNA polymerase to slip on the DNA template or sequencing product, resulting in either a loss of polymerase or unreadable sequence beyond the difficult region. On one contig end this gap has a 305 basepairs (bp) long string of GAs. The other side starting from the gap and traveling into the contig, has
approximately 700 bp of unique sequence, followed by 396 bp of GA repeated, followed by 620 bp of TTAGGG repeated, followed by 50 bp of ATs. Plasmid subclones surrounding the gap have not been able to close the gap when sequenced with the transposon bombing method and primers designed from surrounding the area have been unable to amplify PCR products. Sequencing off of primers designed in the most internal unique regions only provide <100 bp of sequence. All of these results suggest a strongly bound hairpin across this area preventing complete sequence, with a possible fifth simple repeat section still within the gap. Further finishing work is being conducted to close the sequence.

**Computational Methods for Closing Difficult Gaps: Genome Based Approach**

Two computational methods were designed to combat the misassemblies caused by repetitive sequences. The first method is termed the genome based approach because it uses the biological or genomic information present in the BAC sequence to determine the correct assembly configuration. As explained above, many assembly gaps occurred when similar sequences are found in multiple locations in the BAC. In maize, this occurs frequently with the long LTRs of retrotransposons, the traces for one or more location collapse their assembly into a single copy. Our genome based approach uses the structure of the nested TEs to suggest the gap filling sequence.

The first step in the genomic based approach was to run both contigs surrounding the gap with TEnest (Kronmiller and Wise, 2008). This gave two nested structure pictures of the contigs, the TE insertions leading into the gap were examined for any gap-split TE insertions. For example, for a gap presented in Figure 2, one contig end would contain a partial LTR (and possibly some internal TE sequence) of one nested retrotransposon near the end of the gap, the other contig would contain the other sections of this partial retrotransposon along with the complete sequence of the nested TE. Other TE insertions, more than presented in the simple example of Figure 2 could confound the identification of the split TE, but they could also be of assistance. If the two TEs shown in the example were both nested in an older TE insertion the older TE would also be split around the gap even further distance from the problem region and providing more evidence for the nesting pattern.

Once the nesting structure of the TEs was identified using the above process a string of DNA sequence could be filled in to span the gap. Sequence surrounding the gap was built
Figure 2. Nested LTR Retrotransposons Cause Sequence Assembly Gaps. Diagram of the commonly seen type gap caused by nested LTR retrotransposons. (A) Nested TE insertion view of the gap region. The blue TE (labeled ‘2’) is found nested within the LTR of the green LTR retrotransposon (labeled ‘1’). This can cause an assembly gap in one of three locations, at the insertion point of the blue TE on either the left or the right of the insertion, or on the other LTR of the green TE at the corresponding location of the insertion point. (B) A sequence view of the three gap locations caused by insertion of the blue TE into the LTR of the green LTR retrotransposon. The blue TE has inserted into the left LTR of the green TE, an assembly gap can be found on the left LTR to either the left or the right of the blue TE insertion. In either case, the sequence of the left LTR of the green TE has been split apart, sequences belonging to the right LTR have incorrectly assembled at this split location (shown as the arrow pointing to the red sequence) and cause the gap assembly. The assembly gap can also occur on the right LTR of the green TE. Here the join sequences between the left LTR and the blue TE, found on both sides of the blue TE insertion, can assemble incorrectly into the sequence of the right LTR and prevent the sequence from aligning. Successful closing of these types of gaps is crucial to characterization of maize nested repeat clusters.

to resemble the predicted nested TE structure. This built sequence contains 3 sections. The split LTR, formed by identifying it’s missing sequence donated by the corresponding full LTR. The join point between the split LTR and the nested TE exactly identified by the nested location on the other side of the split LTR. Finally the sequence of the nested TE is added to complete the sequence spanning the gap. A low quality backbone phd file (Ewing et al., 1998) was created from the proposed gap spanning sequence and used to drive the phrap assembly. From here the correct sequence traces were found either during the assembly or by user in Consed (Gordon et al., 1998). Several iterations were generally required to add or
remove any sequence differences between the proposed backbone and the true sequence. Ultimately sequences were found to span across the gaps, custom sequencing primers were designed to help span low quality regions if necessary.

**Computational Methods for Closing Difficult Gaps: Sequence Based Approach**

The second computational method used for difficult gap closure used the sequence information from paired end plasmids. Essentially mimicking a localized constrained assembly, the sequence based approach would back out of the gap into the contig looking for unique sequence unduplicated in the BAC assembly. This process backed up on both contigs for at least 4 kb (the largest plasmid clone length) but often much longer to find unique sequence. At the unique locations all the traces found in this area and the plasmid end pairs for these traces were built into separate assemblies. phd file backbone sequences would be made from these small localized assemblies and again overlapping sequences and their mate pairs would be added and assembled to the localized assemblies, continuing until the contigs identified and correctly assembled missing sequences or sequences incorrectly placed and walked into the gap.

This sequence based approach was most useful on the simpler gaps caused by duplicated regions in the BAC that condensed the sequence into one region. In these misassemblies the collapsed traces were identified by their plasmid mate pairs anchored in unique sequence and forced to assemble into the duplicated copy. This process also proved to be helpful to build a backbone phd sequence when closing gaps by the genomic based approach explained above. Often the sequences that were needed to span the gap were hard to identify or didn’t match the predicted backbone sequence well enough to find by assembly or by hand, this sequence based method was useful to draw them to the correct location.

**Transposable Element Annotation shows Large Repeat Clusters**

The two sequence contigs were repeat annotated with TEnest (Kronmiller and Wise, 2008) using the maize repeat database. For the *rf1*-C1 961 kb contig (EF517601), TEnest identified 60 whole LTR retrotransposons, 3 solo LTR sequences, 6 whole DNA transposons, and 42 partial TEs. For the *rf1*-C2 594 kb contig (EF517600), TEnest identified 41 whole LTR retrotransposons, 4 solo LTR sequences, 2 whole DNA transposons, and 18 partial TEs (Figs. 1, 3). The ratios of solo LTR : whole LTR retrotransposon : DNA transposon : partial
TE insertion were consistent with the genome wide analysis of maize TEs we presented in Kronmiller and Wise (2008). The families of TEs identified, the abundance of solo LTR sequences, and the estimated age of insertion for LTR retrotransposons in these two sequence contigs were also found to be consistent with the previous results. The overall TE content is greater in these contigs, 78% compared to the previously reported 67% across the sampling of the finished maize BACs, possibly showing the bias of BACs selected for sequencing with high gene and low repetitive content in previously sequenced BACs.

Definite separation between gene areas and repeat areas can be seen when large sections of the maize genome are evaluated. In maize this phenomenon is known as oceans and islands, where islands of genes are found within oceans of repetitive clusters (SanMiguel et al., 1998). For this analysis of repeat clusters, we defined a cluster or ocean as a group of nested or closely inserted TEs. TEs found inserted less than 5 kb from each other and not separated by a predicted non-transposon related gene were grouped together as a repeat cluster. Sixteen TE clusters were identified in the two rfI-associated contigs. Groups of TEs identified by this definition that contained less than three TE insertions were explained as TE insertions within a gene island so were left out of repeat clusters. Repeat clusters range in size from 23 kb to 155 kb, ranging from 3 to 18 TE insertions (Table II). While sizes of TE clusters are generally evenly distributed across the contigs, the two largest clusters are found on the smaller 594 kb rfI-C2 contig. This corresponds to its higher repeat percentage, 82% verses 75% of rfI-C1, and also its decreased density of predicted genes as explained in detail below.

TEnest displays clusters of TE insertions, with multiple layers of chronologically inserted TEs nested into one another. As repeat clusters become more dense and complex the heights or levels of these TE insertion clusters increase. The level heights of TEnest-displayed repeat clusters observed here correspond to the lengths of repeat clusters; large repeat clusters contain more TE insertions which have higher levels of nested TEs. The largest repeat group, repeat cluster 13 (RC13) at 155 kb, has 18 TE insertions, 12 of which are full LTR retrotransposons (Table II). This cluster has a height of 6 nested TEs. Estimated times since TE insertion is spread evenly throughout the repeat clusters; larger clusters do not have younger or older LTR retrotransposons insertions when compared to smaller clusters.
Figure 3. TEnest Graphical Display of Maize Sequence Contigs
Figure 3. TEnest Graphical Display of Maize Sequence Contigs. (A) TEnest insertion display output of rf1-C1 961 kb maize contig, split into two sections. (B) TEnest insertion display output of rf1-C2 594 kb maize contig. Transposable elements are shown as triangles inserted into the black DNA line, a legend of TE families is shown below. Predicted genes exons are shown as green boxes with directional arrows and have lines connecting exons. Repeat clusters identified and shown in Table 2 are displayed here as spanning lines beneath each TEnest picture, labeled RC1-RC16. Repeat clusters range from 23 kb to 155 kb. Gene islands sizes range from 4 kb to 25 kb and contain from 1 to 9 genes.
Table II. Transposable elements identified with TEnest by repeat cluster

<table>
<thead>
<tr>
<th>Repeat Cluster</th>
<th>Start</th>
<th>End</th>
<th>Size</th>
<th>TE Insertions</th>
<th>LTR retrotransposon</th>
<th>Solo DNA transposon</th>
<th>Mite</th>
<th>Partial</th>
</tr>
</thead>
<tbody>
<tr>
<td>rf1-C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC1</td>
<td>1</td>
<td>31288</td>
<td>31287</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC2</td>
<td>34881</td>
<td>82158</td>
<td>47277</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC3</td>
<td>101703</td>
<td>191706</td>
<td>90003</td>
<td>14</td>
<td>7</td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>RC4</td>
<td>289569</td>
<td>341973</td>
<td>52404</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC5</td>
<td>366102</td>
<td>389224</td>
<td>23122</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC6</td>
<td>407934</td>
<td>509832</td>
<td>101898</td>
<td>13</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC7</td>
<td>536833</td>
<td>573823</td>
<td>36990</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC8</td>
<td>615187</td>
<td>691808</td>
<td>76621</td>
<td>9</td>
<td>6</td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>RC9</td>
<td>699169</td>
<td>792216</td>
<td>93047</td>
<td>12</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC10</td>
<td>876233</td>
<td>918911</td>
<td>42678</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC11</td>
<td>924464</td>
<td>960629</td>
<td>36165</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rf1-C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC12</td>
<td>1</td>
<td>120813</td>
<td>120812</td>
<td>14</td>
<td>10</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>RC13</td>
<td>147102</td>
<td>302699</td>
<td>155597</td>
<td>18</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RC14</td>
<td>313302</td>
<td>367983</td>
<td>54681</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>RC15</td>
<td>401526</td>
<td>456413</td>
<td>54887</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>RC16</td>
<td>507665</td>
<td>593635</td>
<td>85970</td>
<td>12</td>
<td>7</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

As expected, TE nested clusters are seen with older insertions found lower in the cluster, younger insertions are found at higher levels. Partial TE insertions, resulting from whole TEs that have either undergone a deletion or rearrangement at the sequence location or that have mutated significantly so that characterization becomes increasingly difficult, are most often found at the lowest levels of nested TE clusters and so correspond to the oldest TE insertions. This is expected as after enough time for mutations to accumulate the identified TE fragments cannot be reconstructed.

LTR retrotransposons were examined for differing insertion patterns between repeat clusters. Of the three most abundant retrotransposons found in maize (Meyers et al., 2001), *Huck*, a *Gypsy* element, was found to be distributed evenly across the contigs and repeat clusters. *Opie*, a *Copia* element, was found nested almost exclusively in one location. Four
full and 2 partial Opie retrotransposons were identified in repeat clusters RC2 and RC3, while only 3 full and 1 partial Opie elements were found across the rest of the contigs. Ji, a Copia element with sequence identity similar to Opie, was found to have a scattered distribution across rf1-C1 but have 3 full and 2 partial insertions on rf1-C2 in clusters RC12 and RC13. Xilon, a Gypsy element, has 2 full length insertions found in RC11 out of 3 total Xilon elements found. Six MITE DNA transposons were identified, 3 in RC3 and 2 in RC13, and found closely inserted to each other within each cluster. This is in contrast to other results that show that MITEs preferentially insert into the 3’ upstream regulatory regions of genes (Mao et al., 2000; Zhang et al., 2000). Six solo LTRs were identified across the two rf1-associated contigs. Four Gyma, 1 Ruda, 1 Danelle solo LTRs are seen scattered evenly across the repeat clusters. It is interesting to note the high observation of full length Gyma solo LTRs, all seen in the 594 kb rf1-C2.

Distances between clusters, which can also be characterized as length of gene islands, range in size from 4 kb to 98 kb, averaging 32 kb long. These sizes heavily rely on the definition of repeat clusters and would significantly change with modifications to this rule that would separate or combine the repeat cluster sets. Gene islands are not devoid of TE insertions, as described by the definition for repeat clusters. We also attempted to characterize the differences between TEs found inserted within TE clusters verses those found in gene islands. Many of the TE insertions within gene islands are partial LTR retrotransposons, 18 TEs out of 36 total gene island TEs. This suggests ancient TE insertions have occurred in these areas and have since been mutated beyond recognition. Eleven whole LTR retrotransposons were found in gene islands. These are not young recently integrated LTR retrotransposons but rather older, yet complete insertions. Instead the recently inserted LTR retrotransposons are seen almost exclusively at the top levels of repeat clusters. There is one observed exception; a Shadowspawn LTR retrotransposon inserted into rf1-C2 at 389 kb has an estimated time since insertion (Mya) of 0.231. Also seen is nested Ji retrotransposon (0.154 Mya) inserted within an older Huck (0.654 Mya) found in a gene island between 456 kb to 507 kb on rf1-C2. The Huck TE follows the observed pattern of older LTR retrotransposons inserted into gene islands, the younger Ji does not, but because it is inserted within the Huck element, the selective pressures against its insertion may not be as strong.
than if it was to insert directly within the gene island, thus, it has less chance to disrupt nearby gene functions.

**Predicted Maize Genes are found Clustered in Islands**

Sequence files repeat masked by TEnest were used for gene prediction. These masked files were analyzed with three programs; GeneSeqer (Schlueter et al., 2003), FGENESH (Salamov and Solovyev, 2000), and GeneMark.hmm (Lukashin and Borodovsky, 1998). Each of these programs has a monocot or maize specific model. Expressed sequence tag (EST) and protein sequences from *Arabidopsis thaliana*, *Avena sativa*, *Brachypodium distachyon*, *Hordeum vulgare*, rice, *Saccharum officinalis*, sorghum, *Secale cereale*, *Triticum aestivum*, and maize were aligned to the two repeat masked contigs. Results from repeat masking, gene prediction, and sequence alignments were visually displayed with the Generic Model Organism Database (GMOD) package Generic Genome Browser (GBrowse) (Stein et al., 2002) (Fig. 1). Exon structure for each gene model identified by the three prediction programs was plotted on GBrowse and compared to the evidence based sequence alignments. A consensus approach between the results of the three gene prediction programs and the EST and protein alignments was used to pick candidate genes and build gene models. Thirty-two predicted genes were identified in rf1-C1, 10 identified in rf1-C2 (Table III). The sequence for these predicted gene exons were exported and examined for sequence similarity to the characterized genes in GenBank (Benson et al., 2006) to determine possible gene functions. PCR primers were designed in predicted gene exons for high resolution mapping.

Of the forty-two genes identified across the two contigs (Table III), complete gene models were identified for 36 predicted genes, and 6 genes contained truncated models. Gene model and exon coordinates are given in Supplemental Table S1. Non-complete gene models were kept in the gene analysis because they had high similarity to EST or protein sequences of predicted genes and may be pseudo-genes non-functional in the B73 genome or could be results of incomplete annotations. Predicted functions were assigned to 13 of the identified genes (Table III). Genes that we were unable to assign function were given one of two notations: hypothetical, if the predicted gene when compared with BLASTX (Altschul et al., 1990) to the GenBank protein database identified uncharacterized proteins; or predicted, if no proteins were identified but matched to sequenced plant ESTs when compared to the
Table III. Predicted genes identified across the two maize sequence contigs of chromosome 3

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full Length</th>
<th>Sorghum Gene a</th>
<th>Sorghum Location a</th>
<th>Rice Gene b</th>
<th>Rice Location b</th>
<th>Predicted Function c</th>
<th>E-value</th>
<th>Alignments by Length d</th>
<th>Organism</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>rf1-C1</td>
<td></td>
<td>Sb03g005150</td>
<td>3: 5376277-5376881</td>
<td>Os01g14670</td>
<td>1: 8200891-8201323</td>
<td>Germin-like protein 6a</td>
<td>1e-39</td>
<td>67% / 90%</td>
<td>Barley</td>
<td>ABG46238</td>
</tr>
<tr>
<td>1</td>
<td>34289-34705</td>
<td>Sb03g009430</td>
<td>3: 10162570-10162974</td>
<td>Os01g18170</td>
<td>1: 10164942-10165351</td>
<td>Hypothetical protein</td>
<td>1e-27</td>
<td>50% / 59%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>84721-85005</td>
<td>Sb02g026235</td>
<td>2: 61375583-61376994</td>
<td>Sb04g005300</td>
<td>4: 5243874:5245310</td>
<td>FAR1-related 5, zinc ion binding</td>
<td>4e-97</td>
<td>36% / 76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>194534-19740</td>
<td>Sb10g011816</td>
<td>10: 17470752-17473189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>218367-218814</td>
<td>Sb10g027762</td>
<td>10: 57554319-57556853</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>230228-233596</td>
<td>Sb01g014680</td>
<td>1: 13990317-14031920</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>235626-236495</td>
<td>Sb08g006252</td>
<td>8: 9338344:9340386</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>240939-244247</td>
<td>Sb03g005160</td>
<td>3: 5390359-5397235</td>
<td>Os08g30040</td>
<td>8: 18330261-18332327</td>
<td>Hypothetical protein</td>
<td>2e-54</td>
<td>21% / 70%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>265138-265823</td>
<td>Sb03g009450</td>
<td>3: 10161586-10161973</td>
<td></td>
<td></td>
<td></td>
<td>1e-22</td>
<td>80% / 49%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Predicted maize genes aligning to predicted sorghum genes of sorghum genome assembly version 1 are shown. Sorghum genome location is displayed as chromosome: start location – end location.

* Predicted maize genes aligning to predicted rice genes of sorghum genome assembly version 5 are shown. Rice genome location is displayed as chromosome: start location – end location.

* Predicted functions are proteins found to be similar to predicted rf1-associated genes by BLASTX. Hypothetical protein refers to predicted genes that align to uncharacterized proteins in GenBank, predicted protein refers to predicted genes that were identified by gene prediction software and align to sequenced ESTs found in GenBank but do not align to uncharacterized genes in GenBank.

* Percentage of matching amino acids over percentage of query sequence length as identified by BLASTX. For example for rf1-C1 predicted gene 1, 67% of amino acids match over 90% of the query sequence length.
Table III (continued)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full Length</th>
<th>Sorghum Location</th>
<th>Rice Gene</th>
<th>Rice Location</th>
<th>Predicted Function</th>
<th>E-value</th>
<th>Alignments by Length</th>
<th>Organism</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sb03g005160</td>
<td>287591-288349</td>
<td>3: 5380548-5397361</td>
<td>Sb03g009440</td>
<td>1: 10169879-10187099</td>
<td>Heavy-metal-associated domain-containing protein</td>
<td>3e-27</td>
<td>44% / 33%</td>
<td>Arabidopsis</td>
<td>NP_850876</td>
</tr>
<tr>
<td>Sb03g005180</td>
<td>344659-346112</td>
<td>3: 5406762-5408141</td>
<td>Sb03g009480</td>
<td>1: 10196462-10197835</td>
<td>Transcription regulator</td>
<td>2e-24</td>
<td>56% / 44%</td>
<td>Arabidopsis</td>
<td>NP_198156</td>
</tr>
<tr>
<td>Sb03g009445</td>
<td>349079-349445</td>
<td></td>
<td></td>
<td></td>
<td>Predicted protein</td>
<td></td>
<td>60% / 20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sb03g005190</td>
<td>350757-355509</td>
<td></td>
<td></td>
<td></td>
<td>Helix-loop-helix DNA-binding domain containing protein</td>
<td></td>
<td>99% / 75%</td>
<td>Rice</td>
<td>ABF98675</td>
</tr>
<tr>
<td>Sb03g009490</td>
<td>399956-405975</td>
<td></td>
<td></td>
<td></td>
<td>Hypothetical protein</td>
<td></td>
<td>93% / 75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sb01g050100</td>
<td>41025042-10209131</td>
<td></td>
<td>Sb02g020870</td>
<td>2: 51557727-51558799</td>
<td>Terminal acidic SANT 1</td>
<td>5e-32</td>
<td>58% / 34%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sb02g005140</td>
<td>510390-515445</td>
<td></td>
<td></td>
<td></td>
<td>Hypothetical protein</td>
<td>0.0</td>
<td>93% / 75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sb10g005480</td>
<td>516774-523756</td>
<td></td>
<td>Sb03g009540</td>
<td>3: 102009-10265042</td>
<td>Pigment defective 320</td>
<td>8e-85</td>
<td>44% / 77%</td>
<td>Arabidopsis</td>
<td>NP_566296</td>
</tr>
<tr>
<td>Sb02g040360</td>
<td>525736-526529</td>
<td></td>
<td>Sb02g009560</td>
<td>3: 10275285-10337556</td>
<td>MFS18</td>
<td>6e-22</td>
<td>98% / 99%</td>
<td>Maize</td>
<td>CAA47738</td>
</tr>
<tr>
<td>Sb08g0005140</td>
<td>533593-536953</td>
<td></td>
<td></td>
<td></td>
<td>Predicted protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sb10g005480</td>
<td>600098-601155</td>
<td></td>
<td></td>
<td></td>
<td>Cyclin laZm</td>
<td>4e-29</td>
<td>69% / 75%</td>
<td>Maize</td>
<td>AAA20239</td>
</tr>
<tr>
<td>Sb02g009560</td>
<td>613872-614931</td>
<td></td>
<td></td>
<td></td>
<td>CLAVATA1 receptor kinase-like protein</td>
<td></td>
<td></td>
<td>Rice</td>
<td>BAD33838</td>
</tr>
<tr>
<td>Sb03g009560</td>
<td>693500-696202</td>
<td></td>
<td></td>
<td></td>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Os01g19150</td>
<td>795829-799859</td>
<td></td>
<td>Os01g14860</td>
<td>1: 10821557-10823131</td>
<td>Shaggy-like kinase</td>
<td>5e-29</td>
<td>92% / 42%</td>
<td>Wheat</td>
<td>BAF36565</td>
</tr>
<tr>
<td>Gene</td>
<td>Full Length</td>
<td>Sorghum Gene</td>
<td>Sorghum Location a</td>
<td>Rice Gene b</td>
<td>Rice Location b</td>
<td>Predicted Function c</td>
<td>Alignments by Lengthd</td>
<td>Organism</td>
<td>GenBank Accession</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>----------</td>
<td>------------------</td>
</tr>
<tr>
<td>24</td>
<td>802747-803028</td>
<td>Sorghum</td>
<td>24:802747-803028</td>
<td>24:802747-803028</td>
<td>24:802747-803028</td>
<td>Hypothetical protein</td>
<td>3e-03</td>
<td>60% / 19%</td>
<td>Maize DT535841</td>
</tr>
<tr>
<td>26</td>
<td>813912-814668</td>
<td>Sorghum</td>
<td>26:813912-814668</td>
<td>26:813912-814668</td>
<td>26:813912-814668</td>
<td>Ribosomal protein L6</td>
<td>60% / 19%</td>
<td>Rice BAB89961</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>820323-820959</td>
<td>Sorghum</td>
<td>27:820323-820959</td>
<td>27:820323-820959</td>
<td>27:820323-820959</td>
<td>4-coumarate:CoA ligase</td>
<td>3e-03</td>
<td>60% / 19%</td>
<td>Rice BAB89961</td>
</tr>
<tr>
<td>28</td>
<td>823055-824609</td>
<td>Sorghum</td>
<td>28:823055-824609</td>
<td>28:823055-824609</td>
<td>28:823055-824609</td>
<td>Hypothetical protein</td>
<td>60% / 19%</td>
<td>Rice BAB89961</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>834300-836469</td>
<td>Sb01g007380</td>
<td>29:834300-836469</td>
<td>29:834300-836469</td>
<td>29:834300-836469</td>
<td>Zinc finger (zpr1-type) family protein</td>
<td>6e-30</td>
<td>80% / 78%</td>
<td>Rice ABR26166</td>
</tr>
<tr>
<td>30</td>
<td>852391-854332</td>
<td>Sorghum</td>
<td>30:852391-854332</td>
<td>30:852391-854332</td>
<td>30:852391-854332</td>
<td>Ketol-acid reductoisomerase, chloroplast precursor</td>
<td>2e-13</td>
<td>78% / 58%</td>
<td>Pea O82043</td>
</tr>
<tr>
<td>31</td>
<td>874157-875402</td>
<td>Sb03g009600</td>
<td>31:874157-875402</td>
<td>31:874157-875402</td>
<td>31:874157-875402</td>
<td>Hypothetical protein</td>
<td>6e-34</td>
<td>54% / 73%</td>
<td>Rice O82043</td>
</tr>
<tr>
<td>32</td>
<td>921361-936058</td>
<td>Sorghum</td>
<td>32:921361-936058</td>
<td>32:921361-936058</td>
<td>32:921361-936058</td>
<td>Hypothetical protein</td>
<td>1e-15</td>
<td>26% / 41%</td>
<td>Pea O82043</td>
</tr>
<tr>
<td>33</td>
<td>122827-123037</td>
<td>Sorghum</td>
<td>33:122827-123037</td>
<td>33:122827-123037</td>
<td>33:122827-123037</td>
<td>Hypothetical protein</td>
<td>3e-01</td>
<td>50% / 93%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>34</td>
<td>127688-129422</td>
<td>Sb03g013600</td>
<td>34:127688-129422</td>
<td>34:127688-129422</td>
<td>34:127688-129422</td>
<td>Hypothetical protein</td>
<td>3e-01</td>
<td>50% / 93%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>35</td>
<td>136360-142158</td>
<td>Sb03g020630</td>
<td>35:136360-142158</td>
<td>35:136360-142158</td>
<td>35:136360-142158</td>
<td>Hypothetical protein</td>
<td>5e-110</td>
<td>71% / 96%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>36</td>
<td>304382-306168</td>
<td>Sorghum</td>
<td>36:304382-306168</td>
<td>36:304382-306168</td>
<td>36:304382-306168</td>
<td>Hypothetical protein</td>
<td>3e-01</td>
<td>50% / 93%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>37</td>
<td>307915-308219</td>
<td>Sorghum</td>
<td>37:307915-308219</td>
<td>37:307915-308219</td>
<td>37:307915-308219</td>
<td>Hypothetical protein</td>
<td>3e-01</td>
<td>50% / 93%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>38</td>
<td>310843-311817</td>
<td>Sorghum</td>
<td>38:310843-311817</td>
<td>38:310843-311817</td>
<td>38:310843-311817</td>
<td>Hypothetical protein</td>
<td>3e-01</td>
<td>50% / 93%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>39</td>
<td>382024-386107</td>
<td>Sb03g013615</td>
<td>39:382024-386107</td>
<td>39:382024-386107</td>
<td>39:382024-386107</td>
<td>Hypothetical protein</td>
<td>2e-07</td>
<td>21% / 73%</td>
<td>Rice BAB94496</td>
</tr>
<tr>
<td>40</td>
<td>399585-399837</td>
<td>Sorghum</td>
<td>40:399585-399837</td>
<td>40:399585-399837</td>
<td>40:399585-399837</td>
<td>Predicted protein</td>
<td>2e-07</td>
<td>21% / 73%</td>
<td>Rice BAB94496</td>
</tr>
</tbody>
</table>
Table III (continued)

<table>
<thead>
<tr>
<th>Genea</th>
<th>Sorghum Location a</th>
<th>Rice Gene b</th>
<th>Rice Location b</th>
<th>Predicted Function c</th>
<th>E-value</th>
<th>Alignments by Lengthd</th>
<th>Organism</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>458175-460104</td>
<td>Sb03g013620</td>
<td>Os07g44140</td>
<td>Cytochrome P450</td>
<td>1e-71</td>
<td>58% / 95%</td>
<td>Wheat</td>
<td>AAR11387</td>
</tr>
<tr>
<td>42</td>
<td>503410-505209</td>
<td></td>
<td></td>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GenBank EST database with BLASTN and TBLASTX (Altschul et al., 1990). This corresponded to 7 hypothetical genes and 22 predicted genes. A gene density of 1 gene per 37 kb is similar to other observed rates of gene densities over long distances of the maize genome. For example, 1 gene per 19 kb over 2.8 Mb (Brunner et al., 2005), 1 gene per 43 kb over 14.4 Mb (Haberer et al., 2005), 1 gene per 33 kb in 7.8 Mb (Bruggmann et al., 2006), and 1 gene per 27 kb over 6.5 Mb (Bruggmann et al., 2006) are within similar densities observed here.

The length of predicted genes range from 40 bp to 2355 bp (with introns removed), having a median of 495 bp and a mean of 734 bp. Full genes (including introns) range from 40 to 14,697 bp in length, giving a median of 1496 bp and a mean of 2315 bp. Exons have a median of 119 bp and a mean of 191 bp in length. The number of exons per gene ranges from 1 to 11, and have a median of 3, and a mean of 4. Introns have a median of 244 and a mean of 558 bp in length. In two cases, TEs inserted within the introns of genes have increased the length of those introns. Gene 32 on \textit{rf1}-C1, an Ulp1 protease, has a full \textit{Xilon} and a fragment of a \textit{Xilon} retrotransposon found within intron 4. This \textit{Xilon} has an estimated age since insertion of 1.846 Mya, one of the oldest LTR retrotransposons encountered in the two sequence contigs. Gene 39 on \textit{rf1}-C2, a F-box domain containing protein, has a partial \textit{Tim} LTR retrotransposon inserted into intron 2.

We identified 14 gene islands as a result of characterization of 16 nested TE clusters. Because our repeat-cluster definition (explained above) did not allow repeat clusters to contain predicted non-transposon related genes, all of the predicted genes are found in these 14 gene islands. While within gene islands or between islands genes do not seem to form any tight clusters there is obvious clustering of genes when observed on a contig-wide scale. While some gene islands have just one or few predicted gene annotations, many gene islands contain clusters of genes. For example, two gene islands on \textit{rf1}-C1, located between 191 kb to 290 kb and between 792 kb to 876 kb contain 7 and 9 genes, giving these regions a rate of 1 gene per 14 kb and 9 kb respectively. Over the two contigs we observe a rate of 1 gene for every 37 kb, while in the gene islands the density of genes increases to 1 gene per every 11 kb. This is in direct contrast with the results of Haberer et al. (2005), where clustered genes are not observed.
Collinearity between Orthologous regions in Maize, Rice, and Sorghum

To examine sequence collinearity between grass genomes the 14 gene islands were aligned to the rice assembly (IRGSP, 2005) version 5 and the sorghum assembly version 1 (sbi1) (http://www.phytozome.net/sorghum). The maize gene islands were compared to the rice genome directly in VISTA (Dubchak et al., 2000; Mayor et al., 2000; Bray et al., 2003; Brudno et al., 2003; Couronne et al., 2003; Frazer et al., 2004) comparative genome browser. Sorghum was compared to the maize gene islands by first using WU-BLASTN (http://blast.wustl.edu) to align TEnest repeat masked gene island sequences to the sorghum genomes assembly sbi1. Each identified region of similarity was compared in the VISTA malign browser.

Eleven out of the 42 predicted maize genes align when compared to the rice genome, 9 of which are seen in a syntenic location on rice chromosome 1. As illustrated in Figure 4 and Table III, predicted maize genes 1, 10, 16, 23 on rf1-C1 correspond to gene exons of rice chromosome 1 between 8.2 and 8.4 Mb with a conserved order and orientation. rf1-C2 genes 34 and 35 also align to gene exons of rice chromosome 1 in a conserved order and orientation, approximately 5 Mb further along the rice chromosome at 13.2 Mb. Predicted maize genes 31 (rf1-C1) and 41 (rf1-C2) are also found in a conserved locations on rice chromosome 1 however these genes are found in reverse orientation. One non-predicted region on the maize contigs, a region near 85 kb on rf1-C1, aligns to rice gene Os01g14670 on rice chromosome 1 also in this conserved location, near 8.2 Mb. These conserved gene regions show expanded intra-gene distance in maize as compared to rice as expected by the increased density of repeat clusters surrounding gene islands. Two maize genes found in this observed conserved order and orientation are duplicated in the rice genome. Maize genes 1 and 23 are also found on rice chromosome 1 at 10.2 Mb and 10.8 Mb respectively. Several maize regions align to non-syntenic rice locations: predicted maize gene 3 aligns quite well to both exons and introns of the corresponding rice gene Os08g30040 on chromosome 8 and may show an alternate splicing pattern; predicted maize gene 41, in addition to its alignment to chromosome 1, also aligns to rice chromosome 6 and two locations on rice chromosome 7; a non-predicted gene region, rf1-C1 227 kb aligns to two regions on rice, a predicted gene on chromosome 8 and an intergenic region on chromosome 7. Seventeen of the 42 predicted
Figure 4. Comparative Analysis of Maize Sequence Contigs to Rice and Sorghum. Comparative analysis of maize sequence contigs to the rice and sorghum genomes. The two sequenced *rf1*-associated BAC contigs are shown in the center of the figure; predicted genes are shown as red rectangles on the black sequence contig lines with gene identification numbers found in red above. Comparative sequence analysis to rice is shown at the top half of the figure; shared sequence regions between maize and rice are shown as green connecting lines. Comparative sequence analysis to sorghum is shown at the bottom half of the figure; shared sequence regions between maize and sorghum are shown as blue connecting lines. Collinear regions are seen between maize chromosome 3, rice chromosome 1, and sorghum chromosome 3. Seven out of 42 predicted genes are found in collinear order and orientation between maize and rice; another two genes are found in collinear order and reverse orientation. Fourteen out of 42 predicted genes are found in collinear order and orientation between maize and sorghum; eight of these genes are found duplicated in a second location on sorghum chromosome 3. Two genes found in the *rf1*-associated maize contigs are found duplicated in both rice and sorghum. These are shown as red connecting lines.

maize genes align to the sorghum genome. On *rf1*-C1 predicted genes 1, 8, 9, 10, 14, 16, and 18 align with a conserved order and orientation to a 50 kb region on sorghum chromosome 3 near 5.4 Mb. This same set of predicted maize genes, along with genes 23 and 31 are found also on sorghum chromosome 3 near 10.2 Mb (Table III, Fig. 4). This shows at least 500 kb of the maize sequence is duplicated in the sorghum genome on the same chromosome, while only one copy of this region is found in rice, and only one copy of this region is found on the currently sequenced maize genome. Similar to the rice genome comparison, the nonpredicted region near 85 kb on *rf1*-C1 aligns to sorghum chromosome 3 at both 5.4 and 10.2 Mb
locations. *rf1*-C2 gene predictions show genes 34, 35, 39, and 41 are shared between maize and sorghum over the sequence of this contig in similar order and orientation. The two maize genes found duplicated on rice chromosome 1, gene predictions 1 and 23, are found in a third location on sorghum chromosome 3 near locations 13.0 and 13.9 Mb. The third duplication of gene 23 on sorghum is found in a reverse orientation in relation to the maize sequence. Predicted maize genes 3, 5, 15, 17, and 29 align to sorghum chromosomes out of order with the observed collinear genes, each of these genes align to seemingly random gene locations on sorghum with no collinearity seen between any of these gene regions. Many of these genes that align to other chromosomes match multiple locations in the sorghum genome.

The set of 9 predicted maize genes found on rice chromosome 1 in a conserved order are found in the set of 14 genes found conserved when compared to the sorghum genome. The two genes in conserved order and location but found in a reverse direction in rice are seen in the same orientation in maize and sorghum, suggesting the direction change for these genes occurred either in rice after the split to maize/sorghum, or in the maize/sorghum ancestor. There are two genes seen duplicated between the maize contigs and rice, found in three locations in sorghum. This duplicate region on rice and triplicate region on sorghum are in similar collinear locations to each other, found between the conserved locations of genes on *rf1*-C1 and *rf1*-C2. One of the copies of gene 23 found in sorghum is found in a reverse orientation and shows this direction change occurred after the maize sorghum split. Seven maize genes are found in two locations on sorghum chromosome 3, these genes are not found duplicated in the rice genome. These 7 genes are not seen duplicated in the initial maize genome sequence, either in chromosome 3 or elsewhere.

**CONCLUSION**

Based on the sequence length, 78% of the *rf1*-associated contigs consist of repetitive sequences (Table I). For an extremely repetitive organism, maize BAC clones are not overly difficult to assemble. Compared to the assembly of much less repetitive genomes: rice, 35% repetitive (IRGSP, 2005); Arabidopsis, 10% repetitive (Arabidopsis Genome Initiative, 2000), human, 44% repetitive (Lander et al., 2001); mouse, 37.5% repetitive (Waterston et al., 2002), and Drosophila, 3.9% repetitive (Kaminker et al., 2002), this near-centromeric region of maize chromosome 3 is not proportionally more difficult (Celniker et al., 2002) to
bring to sequence completion. This is due to a number of reasons. First, the maize genome has many families of transposable elements, and therefore within a given BAC there is less of a chance to contain multiple copies of a type of element. Second, the average size of transposable elements in maize is larger than those of other sequenced organisms, again decreasing the chance of obtaining multiple copies in a single BAC. Third, simple repeats are much less common in maize than in some other sequenced organisms. Simple repeats are generally small (generally less than 500 bp, similar in length to sequence traces) and tandemly duplicated, causing havoc with assembly algorithms. Fourth, the phenomenon of nesting transposable elements in maize is only seen on a small scale in previously sequenced genomes (Quesneville et al., 2005). Nesting within a transposable element will break up the repetitive sequence into smaller sections. Once broken up, these segments are flanked by unique sequences in relation to other similar elements, and so are actually easier to assemble. Unfortunately for sequence assemblies, LTRs of maize transposable elements are in general much longer than those of other sequenced genomes. LTRs are very similar to each other, and they cause much of the gaps seen in initial draft assemblies.

Eleven of the predicted maize genes are found conserved in the rice genome, 17 of predicted maize genes are found in the sorghum genome. One non-predicted gene region is found conserved in both rice and sorghum, this region is close to predicted maize gene 2 and suggests an incorrect gene annotation. While we believe the current model to be correct, a mutation in this sequenced maize line may cause this conserved region to not join with the predicted gene. Nineteen percent of predicted maize genes are found in collinear locations on rice chromosome 1, 31% of predicted maize genes are found collinear to sorghum chromosome 3. Of genes found conserved across both compared organisms, 27% of shared genes are not seen collinear between maize and rice, and 23% of shared genes are not seen in collinear locations between maize and sorghum. Gene islands are not found conserved in their entirety in their orthologous locations. Rather, gene islands are made up of one to two collinear genes with additional genes found on other chromosome locations or not found in the comparison organism. In the maize to rice comparison one gene island is found containing at least two genes in the collinear region. The distance between these two genes has been retained across the organisms. In the maize to sorghum comparison three sets genes
are found with at least two genes in a gene island in the collinear region. One set of genes is seen in with a similar distance between the genes in maize and sorghum, one set has had a slight expansion in maize relative to sorghum, and the final set of genes, the same set observed in the maize to rice comparison, has experienced a large increase of inter-gene distance in the sorghum genome. While the most common increase of inter-gene distance has occurred between gene islands, increase in genome sequence is not limited to repeat clusters. In several instances genes found on the ends of collinear regions of rice and sorghum did not have a maize counterpart, however due to the increased inter-gene distances these genes could be found off the ends of our sequenced contigs. However, within the collinear gene groups several genes were found conserved between rice and sorghum but missing from the rfi1-associated maize sequence contigs.

Sixteen repeat clusters were identified across the 2 sequenced contigs. These clusters are 23 kb to 155 kb long and contain a variety of TEs and LTR retrotransposons with a range of insertion ages. In few cases several LTR retrotransposon families are seen highly clustered in tight groupings within 1 to 2 repeat clusters and may indicate preferential nesting of TEs. Recent insertions of LTR retrotransposons, those that can be considered as the currently active replicating and transposing elements, are seen almost exclusively in the top levels of nested repeat clusters. Insertions into these locations are further away from genes and therefore mutations in these regions have a less detrimental effect on the organism.

Gene islands, located between each repeat cluster, are from 4 kb to 97 kb long and contain from 1 to 9 gene predictions. The average gene density across islands is 1 gene per 11 kb, this is fairly consistent across islands; larger gene islands contain more genes. A small number of cases of large gene islands with few genes are observed, 2 genes in 41 kb and 2 genes in 51 kb. While this may be an artifact of our definition of repeat oceans / gene islands, TEs found inserted in gene islands are seen on a very small scale as opposed to the large nested repeat clusters. In all but one case LTR retrotransposon insertions in gene islands are estimated to have older ages of insertion when compared to the younger TE insertions on upper levels of repeat clusters. This suggests TEs integrated near genes are rare possibly due to their potential to cause plant altering mutations. Two LTR retrotransposons are seen within introns of predicted genes. One of these insertions is a small LTR retrotransposon and has not
significantly increased the size of the intron. The second case, a full Xilon and a partial Xilon LTR retrotransposon inserted into intron 4 (between exons 4 and 5) of predicted gene 32, Ulp1 protease, has added 9,285 bp to the size of this intron, the full intron length is now 10,136 bp. Predicted gene 32 has no orthologous counterparts in either rice or sorghum, its predicted function is based on sequence similarity to a rice protein identified on chromosome 11.

The architecture of the maize varies across its expanse. From comparative sequence analysis of related grass genomes to the clustering of genes or repeats, diversity is observed at different sequence scales and across various sequence lengths. We hope the assembly techniques presented here will assist the community, ultimately providing long contiguous maize genome assemblies that facilitate examination of the genome as a whole.

METHODS

Identification of BACs in the rf1 Region

Three rf1-m allele families (rf1-m3207, rf1-m7323, and rf1-m7212) were analyzed by a modification of the AIMS method (Frey et al., 1998). DNA was extracted from each individual plant of a segregating population. Ten µl DNA (10µg/µl) was digested with the 4-bp recognition restriction enzymes (Msei or Bfai) in a 40µl reaction volume. Adaptors with ligase and ligase buffer were added, incubated at room temperature, precipitated, and rehydrated into 30µl ddH2O. This pre-amplification product was amplified in a PCR reaction with the pre-amplification template, Mu-specific primer (AIM-Mu1: 5’-GAG AAG CCA ACG CCA ACG CCT CC) and adaptor primer (AIM-AdF: 5’-GCA CAC GCG ATT CGA TGT CGA C). Five µl of a diluted (1:500) pre-amplification product was used as a template in the exponential amplification using 5µl δ-35P ATP (Perkin Elmer, Wellesley, MA), Mu-selective primer (AIMS-Mu4: 5’-GCG CTC TTC GTC CAT AAT GGC AAT TGT CTC), and 5 µl unlabeled adaptor primer (AIMS-Ads: 5’-GAC CAC GCG TAT CGA TGT CGA G) in a 50µl reaction. Amplified product was analyzed on an electrophoresis gel and specific fragments were cloned. AIMS fragments with low copy number were used as probes on B73 BAC libraries.

Three different BAC genomic library filters were obtained from Genome Systems (St. Louis), ZMMBBa (CUGI), and ZMMBBb (CUGI). After probing, additional ZMMBBBb and
ZMMBBc (Children’s Hospital of Oakland Research Institute) BACs were computationally identified using maize WebFPC (http://www.genome.arizona.edu/fpc/maize/).

Genomic DNA, cDNA, AIMS, and RFLP probe fragments were labeled by random priming with [α-32P] dCTP (Feinberg and Vogelstein, 1983). All fragments used as probes were screened to verify the copy number of repetitive DNA by hybridizing to Southern blots of genomic DNA digested with HindIII. Hybridization reactions were performed in Church hybridization buffer (EDTA pH8.0, 7% SDS, 0.5M Sodium Phosphate buffer pH7.2, 1% Bovine Serum Albumin) at 65 °C. High stringency washes consisted of two 30-minute washes in 1X SSPE (0.2M Monobasic Sodium phosphate, 3.6M Sodium Chloride, 20mM EDTA) and 0.1% SDS (Sodium Dodecyl Lauryl Sulfate), a 60-minute wash in 1X SSPE and 0.1% SDS, and a 15-minute wash in 0.1X SSPE, 0.1% SDS at 65 °C.

BAC DNA was extracted by a modified alkaline lysis protocol obtained from Clemson University Genomics Institute (CUGI). BACs were digested with HindIII and run on a 0.9% agarose gel for fingerprint analysis. The TIFF images were edited for lane tracking, individual band calling, and size fraction with IMAGE software (Sulston et al., 1989). BAC restriction digest fingerprint data was transferred to Finger Print Contig Program (FPC) (Soderlund et al., 1997; Soderlund et al., 2000; Pampanwar et al., 2005) for contig analysis. The preliminary contigs were generated with a Tolerance of 7 and a Cutoff of 1e-12. The agarose gels were bi-directionally transferred to Hybond N (Amersham Pharmacia Biotech, Piscataway, NJ) for marker confirmation via Southern hybridization. The final contig assemblies of the Maize BAC clones were achieved by reciprocal [α-32P] dCTP random priming reactions with HindIII digested BAC DNA as the template.

Sequenced BAC ends from each of the original putative contigs were used to make low-copy overgo probes. The overgos were designed using the program Overgo Maker (http://genome.wustl.edu/tools/software/overgo.cgi). The set length of the overgos are paired 24-mer-oligonucleotides that contain an 8-bp complementary overlap with the GC range is 40-60%. The oligonucleotides were annealed to each other and a fill-in reaction was performed using [α-32P] dCTP and dATP. The BAC-end overgos were labeled by a revision of the random priming technique with α-32P dCTP and dATP. Hybridization protocol for overgos was similar to those explained above for AIMS probes, except overgos were
hybridized at 58 °C, and overgos were washed for two 15-minute wash in 1X SSPE and 0.1% SDS and a 15-minute wash in 0.5X SSPE, 0.1% SDS at 58° C.

**BAC Sequencing and Assembly**

BAC clones were sequenced by MWG Biotech (Ebersberg, Germany). BACs were sheered and cloned into 3 kb subclone libraries, subclones were end sequenced to a coverage of 8 to 10X. BACs were assembled with phred/phrap (Ewing and Green, 1998; Ewing et al., 1998) (http://www.phrap.org) and CAP3 (Huang and Madan, 1999). To improve sequence of low quality or gap regions individual subclones were resequenced, custom primers were designed for sequencing, PCR primers were designed for amplification around difficult regions and end sequenced, and entire subclones were fully sequenced using transposon bombing methods (Kimmel et al., 1997).

Assembly of repetitive gap regions was aided with use of TEnest (Kronmiller and Wise, 2008). Individual BACs and combined BAC contigs were run with TEnest using default parameters on the provided maize repeat database. Collapsed repeat spanning assemblies were manipulated with Consed (Gordon et al., 1998).

**Annotation of BAC Contigs**

Sequence files masked with TEnest were used for gene predictions. Three programs were used; GeneSeqer (Schlueter et al., 2003), FGENESH (Salamov and Solovyev, 2000), and GeneMark.hmm (Lukashin and Borodovsky, 1998). Predicted gene models were compared across the three prediction programs to determine a consensus for predicted genes. Protein and EST databases for *Arabidopsis thaliana*, *Avena sativa*, *Brachypodium distachyon*, *Hordeum vulgare*, *Oryza sativa*, *Saccharum officinalis*, *Secale cereale*, *Sorghum bicolor*, *Triticum aestivum*, and *Zea mays* were downloaded from GenBank (Benson et al., 2006) and aligned to determine gene models with BLASTX, TBLASTN, and BLASTN. Predicted gene exons were exported and examined for similarity to the plant protein, EST and predicted gene sets of GenBank to determine possible functions. Output from gene prediction programs, alignments to protein and EST sequences, and predicted genes were displayed on the Generic Model Organism Database (GMOD) package Generic Genome Browser (GBrowse) (Stein et al., 2002).
Comparative Analysis of Sequence Contigs

Orthologous regions were identified using the VISTA comparative genomics tools (Dubchak et al., 2000; Mayor et al., 2000; Bray et al., 2003; Brudno et al., 2003; Frazer et al., 2004). Identified maize gene islands were compared to the rice genome (IRGSP, 2005) using GenomeVISTA (Couronne et al., 2003). Sorghum genome assembly sbi1 (http://www.phytozome.net/sorghum) was downloaded and aligned with BLASTN and TBLASTX to maize gene islands to locate genes of similarity; these regions were compared using mVISTA. Coordinates of aligned regions were pulled out of the table of conserved regions from the VISTA output.

Supplemental Data

GenBank sequence submissions submitted with this manuscript; *rf1*-associated contig 1, EF517601 and *rf1*-associated contig 2, EF517600.

LITERATURE CITED


SanMiguel, P and Bennetzen, JL (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. Annals of Botany 82: 37-44.


CHAPTER 4. COMPARATIVE SEQUENCE ANALYSIS OF THE MAIZE RF1 LOCUS

A paper to be submitted to BMC Genomics

Brent A. Kronmiller, Karin S. Gobelman-Werner, and Roger P. Wise

ABSTRACT

Intra-species comparative sequencing provides fine-scale differences between genomes. The amounts and frequencies of polymorphisms can differ significantly between lines, which can be exploited for mapping and identification of cis–regulatory gene mutations. We have sequenced two BACs from an Rf1 containing maize line, Wf9-BG (T). These BACs were compared to our finished rf1-associated B73 contigs to characterize functional and evolutionary differences. These two maize lines are highly similar for much of the aligning sequence, indicating a shared pedigree. Twelve insertions and deletions ranging from 14 to 556 bp, and three transposable element insertions comprise the major differences. A Cassandra tiny retrotransposon in miniature (TRIM) is inserted into maize B73 and is associated with an expanded sequence structure. The duplication of the retrotransposon sequence is hypothesized to be caused by unequal sister chromatid recombination or, alternatively, a two step process of inter-element unequal illegitimate recombination. Either model illustrates that recombined retrotransposons can retain replication ability and LTR retrotransposon sequence structures are continually in flux. In addition, SNP frequencies of 1 SNP per 224 bp over the full aligned sequence, 1 SNP per 600 bp in genic regions are seen. These observed SNP frequencies are more conserved than seen in previous comparisons among maize lines. However, a major difference between Wf9-BG (T) and B73 is also found, highlighting the relationship between Wf9-BG (T) and B73 in this interval. An entire retrotransposon cluster differs between Wf9-BG (T) and B73, indicating a recombination at this location. Sequence leading up to the recombination breakpoint has a higher frequency of polymorphisms than the rest of the aligned sequence showing varying degrees of differences even within small regions of maize.

INTRODUCTION

Functional requirements of protein encoding DNA hinder the accumulation of random sequence changes. This conservation of gene sequence can be observed across related species and provides a basis for linking proteins and their function. Comparative sequence analysis of genes from
diverse species allows one to locate previously unidentified coding regions, predict the gene model and splicing patterns, and provide hypothesized functions of potential protein sequences. Comparative genome analysis across entire organisms provides a foundation for interpreting genome structure, architecture, and evolution.

In plants comparative sequence analysis offers an economical alternative to sequencing cost-inhibiting genomes. Organisms with relatively small genomes such as Arabidopsis, *Oryza sativa* (rice), and *Sorghum bicolor* (sorghum) share many genes with plants containing of much larger genomes such as *Zea mays* (maize), *Hordeum vulgare* (barley) and *Triticum aestivum* (wheat). Predicted genes and protein functions can be extrapolated from model organisms to sequences of interest. To an extent, comparative analysis of genes can also give a view of gene locations conserved throughout genomes. Comparative mapping of grass genomes reported a high extent of conservation across gene content, locality, and order (Ahn and Tanksley, 1993; Devos et al., 1994; Gale and Devos, 1998). The rice genome, when split into orthologous sections, was seen as an ancestral representation of grass genomes (Moore et al., 1995). However, the true sequence relationships between grasses are much more complex. While comparative mapping offers a large scale view of conservation, a closer analysis shows major differences at the sequence level. Comparisons of maize and rice show translocation of associated genes (Tarchini et al., 2000; Lai et al., 2004) and significant lack of micro-colinearity between the species. Intra-species sequence comparisons also show colinearity differences; between maize lines intergenic sequences are altered (Fu and Dooner, 2002; Song and Messing, 2002; Song and Messing, 2003; Brunner et al., 2005), between wheat genomes a similar genic lack of colinearity is observed (Wicker et al., 2003; Gu et al., 2004).

Beyond colinearity of genes, sequence content and lengths of the areas between genes and between gene islands shows variation again among both related organisms and the genomes of lines of the same organism. Transposable element insertions account for much of the non-genic sequence differences (Flavell et al., 1992) and are billed as major reasons for genome expansions (Wendel, 2000; Bennetzen et al., 2005; Hawkins et al., 2006; Piegu et al., 2006). In inter-species comparisons whole repeat clusters are substituted, giving distances between gene islands and differences in genome lengths (Tarchini et al., 2000). Intra-species, repeat clusters are both sustained and substituted (Fu and Dooner, 2002; Gu et al., 2003; Brunner et al., 2005) along with singular element insertions in genic regions or in retained repeat clusters.

Lack of micro-colinearity, seen not only between closely related species but even among the genomes of a single organism, results in missing, mutated, newly introduced, and translocated genes with differing regulatory regions and surrounding sequence. The consequences of these alterations
can be severe, possibly providing major plant altering effects. But what are the causes of such sequence level differences between recently evolved genomes? In the maize genome, helitron transposons have been shown to pick up and move functional genes to new locations providing not only fully functional translocated genes but also sequence fodder for gene evolution (Kapitonov and Jurka, 2001; Lal et al., 2003; Lai et al., 2005; Morgante et al., 2005). Differences in repeat clusters and transposable element insertions alter intron, intergenic, and inter-island distances, altering genome length as well as gene structure and function (Wicker et al., 2001; Gu et al., 2003; Xiao et al., 2008). LTR retrotransposons, with their abundance throughout grass genomes and their large length provide ample sequence for recombination (Vicient et al., 1999; Devos et al., 2002).

Comparisons of conserved genomic regions identify an even finer level of sequence differences; single nucleotide polymorphisms (SNPs) and small scale insertions and deletions (indels). SNPs and indels are useful as genetic markers to assist with genome mapping. As expected by functional constraints, genes show more conservation than non-genic regions; SNP frequencies across shared sequence in maize are shown to be 1 per 28 bp (Tenaillon et al., 2001), the frequency drops sharply in coding regions to 1 SNP per 130 bp (Rafalski, 2002) and 1 SNP per 128 bp (Ching et al., 2002). Expressed sequence tag (EST) sequence shows a similar rate, 1 SNP per 100 bp of EST contigs with 20 or more sequences (Batley et al., 2003). Pyro-sequencing efforts of maize transcriptomes shows a more conserved view, with SNP frequencies of 1 per 300 bp (Barbazuk et al., 2007), however necessity of stringent parameters for low coverage assemblies may artificially influence this rate.

To present a large-scale comparative sequence analysis, we identified and sequenced two overlapping BACs of maize Wf9-BG (T) corresponding to our previously sequenced rfl-associated maize B73 contigs (Kronmiller et al., 2008). Rfl is one of the two complementary fertility restoration genes for T (Texas) cytoplasmic male sterile (cms-T) maize (Duvick et al., 1961; Wise et al., 1996); Wf9-BG (T) harbors Rfl. We present two overlapping BACs totaling 201 kb in length corresponding to approximately 412 kb to 533 kb on rfl-associated contig 1 (Kronmiller et al., 2008). This sequence contains a gene island with four genes, all correspond to genes identified in B73 following a conserved colinear arrangement, however a fifth gene in B73 is omitted from the maize Wf9-BG sequence contig. Surrounding the gene island are repeat clusters, the first of which matches the corresponding B73 repeat cluster with only small indel differences and one differing LTR retrotransposon insertion. The second Wf9-BG repeat cluster does not share TE makeup or sequence identity to the B73 repeat cluster in the corresponding location. We propose that a recombination has occurred providing the differences between the two sequence contigs. Polymorphism frequencies for
shared sequence regions are calculated, giving a frequency of 1 SNP per 224 bp over the entire sequence and 1 SNP per 833 bp in predicted gene exons. Over the aligned sequence we observe significantly more conserved polymorphism rates than previously reported between maize lines. Directly before the proposed recombination point polymorphism frequencies drastically increase.

RESULTS AND DISCUSSION

Two Overlapping Wf9-BG BACs contain Four Genes and Two Repeat Clusters

PCR primers were designed in the exons of predicted genes identified in maize B73 rf1-C1 (EF517601, 961 kb). All possible PCR products were checked in-silico against maize genome and EST sequence to insure uniqueness. Primer products were amplified off associated B73 BACs, gel excised and copy number verified via strip blots. In total, 16 PCR generated probes were used, designed off of the B73 sequence; 7 were positioned uniformly across the contig and 9 clustered between 512 kb to 527 kb of rf1-C1, the region of interest for comparative analysis. Overgo probes were designed across rf1-C1 and rfI-C2 (EF517600, 594 kb) to complement the PCR probes. For both contigs, overgos were designed off repeat masked sequence files using OligoSpawn (Zheng et al., 2006). Predicted overgo pairs were computationally filtered against maize genome sequences to provide uniquely hybridizing probes. One-hundred and twenty overgo paired probes were identified on rf1-C1 and 72 on rf1-C2, these probes were distributed evenly across the sequence in both gene and non-gene unique regions to give unbiased coverage across the sequence.

A Wf9-BG (T) BAC library was constructed by Arizona Genomics Institute (AGI). Wf9-BG (T) contains both the Rf1 and Rf2 alleles necessary for restoration of fertility to cms-T plants. A 9X coverage BAC library was produced, this translated to 6 library filters. PCR and overgo probes were hybridized to the BAC library filters to identify a Wf9-BG BAC path across the B73 rf1 sequence contigs. One-thousand one-hundred and fifty-two BACs were identified and high information content fingerprinting (HICF) (Nelson et al., 2005) was performed by AGI. BAC HICF digests were entered into FPC (Soderlund et al., 1997; Soderlund et al., 2000; Pampanwar et al., 2005) to construct a BAC path in Wf9-BG. Twelve 96-well plates of hybridizing BACs provided only 3 FPC contigs with 10 or more BACs and 31 contigs containing 5 or more BACs. Reciprocal BAC Southern hybridizations were used to join the FPC contigs (see Materials and Methods). A set of 192 representative BACs from the FPC contigs were digested with HindIII and gels were prepared for hybridizations. BACs located on the ends of FPC contigs and PCR probes were hybridized to identify overlapping regions. Several FPC contigs were closed with this process but a complete Wf9-BG BAC path across the rf1-C1 sequence was not obtained. In particular, a large gap in the Wf9-BG map was located between approximately 533 kb to 700 kb of rf1-C1 B73, probes in this area do not hybridize to any Wf9-BG
BACs. Probes found after 700 kb of rfI-C1 B73 again hybridized to Wf9-BG BACs indicating a gap or difference in the Wf9-BG genomic sequence.

Two Wf9-BG BACs were chosen for sequencing, ZM_TBa0130O19 and ZM_TBa0365D18. In relation to B73, these BACs align to approximately 407 kb to 533 kb of rf1-C1. They are in the region of the rf1 mapping of cDNA 6140-1 (Wise et al., 1999) and are located immediately before the Wf9-BG BAC path gap found between 533 and 700 kb. According to HICF FPC alignments and BAC reciprocal Southern hybridizations the two BACs were predicted to have significant overlap. Draft sequence (4-6X coverage; Washington University Genome Sequencing Center) assembled with the phred/phrap package (Ewing and Green, 1998; Ewing et al., 1998) (http://www.phrap.org) provided BACs with 16 contigs (ZM_TBa0130O19) and 22 contigs (ZM_TBa0365D18), however, given their significant overlap a combination of the two BACs gave an assembly with 19 contigs, the majority of gaps found in the single BAC spanned ends.

With use of the corresponding finished B73 sequence we were able to order, orientate, and help provide backbone sequence to drive the assembly of the Wf9-BG BACs. However, due to differences of the two genomes at this location this process could not give a representation of the complete assembly. We conducted a first round of finishing work to help order the remaining sequence contigs. Primers for PCR were designed off subclones and BAC sequences surrounding gaps, nested primers were sequenced off the amplicons. This first round closed 16 gaps and provided an ordered and orientated dual-BAC assembly. Furthermore, the two remaining gaps were found in nested repeat clusters and could be ignored for the needs of this analysis.

The resulting Wf9-BG BAC assembly was repeat and gene annotated using the same process as described in Kronmiller et al. (2008). Repeat annotation with TEnest (Kronmiller and Wise, 2008) gave 13 full LTR retrotransposons, 1 solo LTR, 1 DNA transposon, and 10 partial TE insertions (Fig. 1B). This region consists of a gene island of 25570 bp surrounding by two nested repeat clusters. Gene islands were characterized as areas between repeat clusters; repeat clusters are defined as a group of nested and closely inserted TEs not separated from each other by more than 5 kb. Gene predictions identified 4 genes, their locations and hypothesized functions are provided in Table I.

**Comparative Sequence Analysis: Recombination and Differences in TE Insertions**

The assembled Wf9-BG BAC sequence was compared to the corresponding B73 sequence (rf1-C1 from 407 kb to 605 kb), this was accomplished with WU-BLASTN (http://blast.wustl.edu) and a custom perl script to provide a SVG (http://www.w3.org/Graphisc/SVG) image of the alignments. Two sequence comparisons, the first using full sequence from both maize lines (Fig. 2A), the second using TEnest repeat masked sequences (Fig. 2B), show the alignment across this region.
Table I. Predicted genes identified in maize Wf9-BG

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full Model</th>
<th>Direction</th>
<th>Full Length</th>
<th>Exons</th>
<th>Predicted Function*</th>
<th>E-value</th>
<th>Alignments by Lengthb</th>
<th>Organism</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Y F</td>
<td>105134-110191</td>
<td>105134-105245, 105746-105805, 106298-106328, 106773-106924, 107168-107259, 107399-107464, 108418-108637, 108751-109100, 110063-110116, 110150-110191</td>
<td>Similar to Terminal acidic SANT 1</td>
<td>5e^-32</td>
<td>58% / 34%</td>
<td>Maize</td>
<td>AAV28560</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Y F</td>
<td>111520-111713, 112712-113088, 112232-112504, 115065-115402, 116046-116188, 117039-117201, 118290-118502</td>
<td>111520-111713, 112712-113088, 112232-112504, 115065-115402, 116046-116188, 117039-117201, 118290-118502</td>
<td>Similar to Pigment defective 320</td>
<td>8e^-85</td>
<td>44% / 77%</td>
<td>Arabidopsis</td>
<td>NP_566296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 N F</td>
<td>120479-121272</td>
<td>120479-121272</td>
<td>Hypothetical protein</td>
<td>2e^-07</td>
<td>80% / 12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Y F</td>
<td>122676-123062</td>
<td>122676-123062</td>
<td>MFS18</td>
<td>6e^-22</td>
<td>98% / 99%</td>
<td>Maize</td>
<td>CAA47738</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Predicted functions are proteins found to be similar to predicted rf1-associated genes by BLASTX.

b Percentage of matching amino acids over percentage of query sequence length as identified by BLASTX. For example for rf1-C1 predicted gene 1, 67% of amino acids match over 90% of the query sequence length.
Figure 1. Comparative TEnest View of B73 and Wf9-BG BAC Contigs. (A) TEnest transposable element annotations of B73 rf1-associated contig 1 from 412 kb to 613 kb. (B) TEnest annotations of Wf9-BG contig sequenced in this analysis. A legend is shown at the bottom of each TEnest display; colors represent different TE families. Four differences are seen between the two contigs: (1) the Wf9-BG sequence starts after the beginning of the Ji and Zeon LTR retrotransposons found at the left of the contigs, the B73 sequence is shown with the full Ji and Zeon retrotransposons; (2) a Prem LTR retrotransposon is inserted in an Opie in Wf9-BG relative to B73; (3) a Cassandra LTR retrotransposon is found inserted in a Huck LTR retrotransposon in B73 relative to Wf9-BG; (4) a shared Huck LTR retrotransposon is found in multiple locations in the non-matching sequence location; (5) a recombination breakpoint after about 2/3 of the sequences shows a different repeat cluster on the right sides of both maize contigs.

For both figures the maize B73 sequence is found on the top, the Wf9-BG sequence on the bottom. The full sequence comparison identifies almost complete alignment for the first 2/3 of the sequences, the only differences between the B73 and Wf9-BG are seen as one expansion in B73 in relation to Wf9-BG and one expansion in Wf9-BG in relation to B73, noted as 2 on Figure 2A and as 3 on Figure 2B, respectively. The two indels observed can be visualized by a comparative view of TEnest outputs (Fig. 1). The B73 TEnest output for this region (Fig. 1A) when compared to the TEnest display output for Wf9-BG (Fig. 1B) shows the expansion in Wf9-BG is caused by an insertion of a recent Prem LTR retrotransposon, nested into a Huck LTR retrotransposon. The estimated age since insertion of the Prem element calculated by TEnest is 0.038 Mya, signifying this as a very recent insertion. Indeed, the alignments of the two LTRs of the Prem LTR retrotransposon show there are 6 SNPs found between them, therefore very little time has passed since the integration of this LTR retrotransposon.
Figure 2. LTR Retrotranspon Recombination Methods to form an expanded Cassandra Element. Two methods of LTR retrotransposon recombination are proposed. Unequal sister chromatid recombination (A) mis-aligns the LTRs (represented as green arrows) of an LTR retrotransposon to produce an expanded TE. Each version of the Cassandra element is from the same insertion and so contain matching target site duplications (TSDs) represented as red circles flanking the LTR retrotransposon. After recombination the expanded TE retains matching TSDs. Unequal illegitimate inter-element recombination (B) occurs between the LTRs of two LTR retrotransposons separated by a length of DNA sequence. The two LTR retrotransposon have differing TSDs represented as red and blue circles flanking the insertions. Inter-element recombination forms an expanded Cassandra element with non-matching TSDs. This expanded element retains replication ability and copies elsewhere in the genome, upon integration it obtains matching TSDs represented as purple circles flanking the expanded Cassandra retrotransposon.

The expansion in B73 is caused by a Cassandra terminal-repeat retrotransposon in miniature LTR retrotransposon (TRIM) (Witte et al., 2001) insertion nested into a Huck LTR retrotransposon (a different Huck insertion than the one mentioned above with Prem). A Cassandra LTR retrotransposon is 809 bp long, with LTRs of 303 bp each. This Cassandra insertion is one full element and one partial element 506 bp long, both the full and partial are inserted in the same location in the Huck LTR retrotransposon on B73. The partial Cassandra starts immediately after the right LTR of the full insertion, it contains an internal region and another right LTR equaling basepairs 304-809 of Cassandra coordinates. The full Cassandra insertion here is LTR-internal-LTR-internal-LTR. TEnest pairs forward and reverse LTRs of a full length LTR retrotransposon based on their similarity (SanMiguel et al., 1996; Kronmiller and Wise, 2008). Here, listing the Cassandra LTRs from left to right as A, B, C; LTRs A and B have 1 bp difference, LTRs B and C have 4 bp differences, LTRs A and C sum the differences found in the previous two comparisons to total 5 bp differences. TEnest therefore paired LTRs A and B giving a time since insertion of 0.126 Mya, leaving the internal and furthest right LTR as an extra partial insertion. Given the small length of the LTRs the true age of the Cassandra insertion may be lower and more in line with a recent insertion.
Upon integration into the genome LTR retrotransposons duplicate 5 bp at the point of insertion. This target site duplication (TSD) is found flanking LTR retrotransposons. TSDs can be used to help identify the cause of the observed Cassandra sequence expansion. Here we observe matching TSD immediately flanking the expanded Cassandra retrotransposon. Two processes of recombination can explain this expanded sequence formation. The first, unequal sister chromatid recombination is caused by the mis-alignment of the Cassandra element (Fig. 2A). Alignment of opposite LTRs and recombination of this region will result in an expanded LTR retrotransposon. The original two Cassandra elements contain matching TSDs, these are retained in the expanded sequence version.

A more complicated process is also possible. Unequal illegitimate recombination of LTR retrotransposons is also a cause of altered retrotransposon states. There are two types: Intra-element unequal recombination produces a solo LTR with matching flanking TSDs; inter-element recombination produces a TE section with non-matching TSD in 3 possible configurations, a solo LTR, a full length LTR retrotransposon, or a full length LTR retrotransposon with an additional internal region and LTR (LTR-internal-LTR-internal-LTR). As illustrated in Fig. 2B, we propose a two step process could have formed the expanded Cassandra. At a different location in the genome sequence two Cassandra LTR retrotransposons experienced an unequal illegitimate recombination to produce a duplicated configuration with non-matching TSDs. This duplicated Cassandra did not lose replication ability, a new copy of Cassandra, using the furthest most LTRs to replicate, inserted within the Huck retrotransposon of B73, creating at its insertion matching TSDs.

Within the 26kb gene island Wf9-BG predicted genes 1-4 are equivalent to predicted genes of the corresponding B73 region genes 15-18 (Kronmiller et al., 2008). Each of these gene predictions retains its intron and exon structure despite small indel and substitutions in the UTR, exons, and introns of the predicted gene models. At approximately 130 kb, from the beginning of the sequence comparison of B73 and Wf9-BG (Fig. 3A), directly after the Wf9-BG gene island the sequence alignment ends, from this point there is no similarity between B73 and Wf9-BG save for a few LTR retrotransposons shared in both. The shared LTR retrotransposon (Huck) sequence is seen on the right of this breakpoint on Figure 3A as lines matching to multiple locations. The breakpoint is shown again in the repeat masked sequence comparison (Fig. 3B), here the 26 kb gene island is seen matching to both sequences, and the breakpoint is immediately to the right of the island, corresponding with approximately 533 kb on rf1-C1 B73. Directly after the gene island 71 kb of the Wf9 sequence does not match to B73 and continues to not align up to the end of the Wf9 sequenced BACs. The non-matching sequence consists solely of LTR retrotransposons. On maize B73, the gene
Figure 3. Sequence Comparison of B73 rf1-associated Contig 1 and Wf9-BG. WU-BLASTN alignments between each contig was parsed for E-value, and drawn according to color. Green: E-value < 1e$^{-200}$, Blue: E-value < 1e$^{-100}$, Gray: E-value < 1e$^{-60}$. (A) Full sequence alignment, (B) TEnest repeat masked sequence alignment. Five differences are noted here (number designations here are consistent with those in Fig. 1): (1) the Wf9-BG sequence intentionally starts after the beginning of the B73 contig; (2) the Prem insertion in Wf9-BG; (3) the Cassandra insertion in B73; (4) the shared Huck LTR retrotransposon found after the breakpoint in multiple locations on both maize lines; (5) the recombination breakpoint between B73 and Wf9-BG.

island continues for almost 4 kb, including gene 19, before the start of the next gene island. The truncation of the gene island could be caused by a translocated retrotransposon cluster inserted between B73 gene 18 and 19, however, a more probable cause is the difference is caused by recombination. Near the area of recombination a solo Opie LTR is found. This LTR has matching TSD suggesting it is the result of intra-element illegitimate recombination.

**Polymorphism Frequency is More Conserved than Previously Reported**

Studies of SNP frequencies in maize have focused on EST sequences and assemblies, providing polymorphism occurrences in expressed regions of the genome. In our comparison of B73 to Wf9-BG gene-rich regions, approximately 26 kb of Wf9-BG sequence matches to 23 kb of the B73 sequence. Therefore, we conducted an in-depth analysis of polymorphism frequencies across the aligned sequences. Here, using full length genomic sequences assemblies we are able to also give SNP and longer length substitutions and indels in the repeat cluster and non-expressed regions of gene islands. In addition, SNP analysis of finished genomic sequence encounters a lower rate of sequence errors than low depth EST assemblies, especially those utilizing pyro-sequencing techniques. While giving the ability to examine a large amount of sequence, low coverage EST
assemblies must correct for false positive SNP results and therefore may not arrive as close to the true count of polymorphisms as analysis with finished genomic sequence. Orthologous regions of B73 and Wf9-BG were aligned using lalign (Huang and Miller, 1991) and polymorphisms were calculated with custom Perl scripts. Single base pair insertions or deletions, each possible combination of single basepair substitution, and instances of multiple substitutions or indels were recorded. Repeat clusters and gene islands were calculated individually; for gene islands, locations of predicted genes, UTRs, introns, and exons were each separately determined.

Across the entire sequence length, 579 instances of polymorphisms were observed, a frequency of 1 polymorphism per 224 bp (Table II). This includes all observations of 1 to 10 continuous polymorphisms, each counted as one event. The shared repeat cluster has one polymorphism per 274 bp, the gene island shows one polymorphism per 129 bp. The gene island, separated into predicted genes and intergenic regions shows polymorphism frequencies of 1 SNP per 600 bp and 1 SNP per 94 bp respectively. For the 4 predicted genes in the shared sequence, 11 polymorphisms (6 SNPs) are found in UTRs, 19 polymorphisms (18 SNPs) in introns, 15 polymorphisms (14 SNPs) in exons. Twelve indels >10 bp are found between the two genomes in repeat clusters, ranging from 14, 20, 21 bp up to the 1251 bp Opie solo LTR insertion in Wf9-BG, the 1317 bp Cassandra deletion in Wf9-BG, and the 8560 bp Prem insertion in B73. Intergenic regions in the gene island show 3 insertions in Wf9-BG; 142 bp, 346 bp, 530 bp, as well as a 121 bp Wf9-BG deletion and an insertion of a 1251 Opie solo LTR. None of these large scale indels have any discernable effect on their nearby genes.

Polymorphism frequencies are seen greatly increasing up to the point of recombination. The 26 kb gene island shows 198 polymorphism events, 95% of these are seen in the final 11 kb (Fig 4). This drastic increase of SNP frequency found directly before the point of recombination may be related to the cause or effect of the recombination point. However, one expects areas that recombine to be closely related, so the cause of the divergence of this region is yet unexplained.

CONCLUSION

SNP frequencies reported here are quite different than those previously reported among maize. While demonstrating the close relationship of Wf9-BG and B73, this also illustrates the polymorphism frequency disparities between different lines of maize. The majority of differences between maize lines B73 and Wf9-BG in this sequenced area are large-scale insertions and deletions. Fourteen indels >10 bp are observed, including 3 TEs, one Prem insertion in Wf9-BG, an expanded Cassandra insertion in B73, and on Opie solo LTR in Wf9-BG.
Table II. Polymorphism Frequencies between maize B73 and Wf9

<table>
<thead>
<tr>
<th>Sequence Classification</th>
<th>Single Basepair Polymorphisms</th>
<th>Multiple Basepair Polymorphisms</th>
<th>Total Polymorphism Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transitions</td>
<td>Indels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Substitutions</td>
<td>Deletion in Wf9</td>
<td>Insertion in Wf9</td>
</tr>
<tr>
<td></td>
<td>Transversions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C/ T</td>
<td>G/ A</td>
<td>Total</td>
</tr>
<tr>
<td>Gene Island</td>
<td>47 42 89</td>
<td>9 12 21 8 50</td>
<td>139</td>
</tr>
<tr>
<td>Genes</td>
<td>5 4 9</td>
<td>9 12 21 8 50</td>
<td>12</td>
</tr>
<tr>
<td>UTRs</td>
<td>4 1 5</td>
<td>3 3 5</td>
<td>5</td>
</tr>
<tr>
<td>Introns</td>
<td>1 1 2</td>
<td>2 1 3</td>
<td>2</td>
</tr>
<tr>
<td>Exons</td>
<td>2 2 3</td>
<td>3 3 5</td>
<td>5</td>
</tr>
<tr>
<td>Intergenic</td>
<td>36 32 68</td>
<td>8 11 14 6 39</td>
<td>107</td>
</tr>
<tr>
<td>Repeat Cluster</td>
<td>90 100 190</td>
<td>25 37 29 36 127</td>
<td>317</td>
</tr>
</tbody>
</table>

*Indel <10 bp omits all insertions or deletions longer than 10 bp. This includes deletions of 14 bp, 20 bp, 95 bp, 509 bp, 1317 bp, insertions of 20 bp, 21 bp, 556 bp, 8560 bp in the repeat cluster; deletions of 40 bp and 121 bp, insertions of 142 bp, 346 bp, 530 bp, 1251 bp (an Opie solo LTR) in intergenic regions.
Figure 4. Increasing Polymorphism Frequencies nearing Recombination Location. A 26 kb gene island is shared between the B73 and Wf9-BG sequence contigs found between 105 Kb to 131 Kb. One Kb bins of SNPs found between B73 and Wf9-BG are shown plotted against the gene island. The four predicted genes identified in this region are shown as green arrows. Ninety-five percent of SNPs identified within the 26 Kb gene island are found in the last 11 Kb, directly before the point of recombination between B73 and Wf9-BG. Larger insertions and deletions are shown above the plotted bin showing and even more drastic polymorphism frequency leading up the recombination location.

The Cassandra TRIM LTR retrotransposon insertion is found within the sequence of a Huck retrotransposon. Its sequence is duplicated to contain an extra LTR and internal region. We propose this could have occurred by one of two methods. In the first possibility unequal sister chromatid recombination provided the duplicated Cassandra element, it contains matching TSDs. Alternatively a two step process of illegitimate unequal inter-element recombination could have produced this Cassandra configuration. In this process first a duplicated Cassandra is formed by inter-element recombination containing non-matching TSDs. This element retained replication functionality and inserted into the location we observe here, upon integration it obtained matching TSDs.

The Opie solo LTR is found in the vicinity of the recombination breakpoint between B73 and Wf9-BG. It has matching TSDs on either side, therefore arising from either unequal sister chromatid recombination or an intra-element illegitimate recombination between its left and right LTRs. While the Opie solo LTR is found in Wf9-BG before the recombination point neither a solo Opie LTR nor a full Opie retrotransposon are seen nearby in B73. The formation of this region, in light of the Opie
LTR, could have occurred in one of 3 ways. The full length Opie insertion could have recombined with another Opie further downstream deleting the internal sequence and giving the new sequence after the LTR seen in Wf9-BG. However, this is not a possibility because the TSDs of the solo LTR are identical, 951 bp after the solo LTR align in both genomes. Besides, there is not an Opie in the B73 sequence close to this location. Another possibility is that in the Wf9-BG line the Opie retrotransposon inserted, then was deleted to form the solo LTR. While this is entirely possible, it seems unlikely for so much to occur at one genome location. A third possibility is the differences observed at the recombination point are caused by a cross-over between Wf9-BG’s progenitors, Wf9 and IA153. At some point in the aligning sequence the chromosomes recombined. The first repeat cluster and most of the gene island are from IA153, while the area after the recombination point is from Wf9-BG. Seven kb before the breakpoint the alignment begins to worsen, 95% of polymorphisms found in the 26 kb gene island occur in the last 11 kb, suggesting this is the location of the cross-over. In order to resolve this question and identify where this new Wf9-BG sequence arose we are attempting to locate additional BACs to continue this contig into unique sequence regions.

Our comparative sequence analysis between B73 and the sorghum and rice genomes (Kronmiller et al., 2008) showed the rf1-C1 contig matched to both rice (chromosome 1 from 8.2 to 8.4 Mb) (IRGSP, 2005) and sorghum (chromosome 3 from 10.1 to 10.3 Mb) (http://www.phytozome.net/sorghum). In addition, a portion of the B73 rf1-C1 matched to another location of sorghum chromosome 3, from 5.3 to 5.4 Mb. This is an incomplete alignment to rf1-C1 and the sorghum region stops matching after gene 18, the same general location as the recombination point we observe between B73 and Wf9-BG. While not the exact location in both genomes, general proximity shows a propensity for recombination in this region.

METHODS

For PCR probes, sequences of predicted PCR products were aligned to MAGI (Fu et al., 2005) and PlantGDB GSS alignments (Dong et al., 2005). Those with less than 5 alignments with greater than 60% identity over 60% of the sequence length were kept as probes. These probes were amplified off B73 BAC clones, gel excised, and to ensure probes were unique were hybridized to genomic B73 and Wf9-BG digested with HindIII. Hybridization and wash protocols for library filters follow those explained in Kronmiller et al. 2008. For overgo probes B73 BACs were repeat masked with TEnest, overgos were designed with OligoSpawn (Zheng et al., 2006). Probes were hybridized
to Wf9-BG library filters in several pools. PCR gene probes were pooled together and several groups of oligo pools were used.

Twelve plates of BACs (1152 clones) were ordered and HICF from AGI, BAC contigs were built with FPC (Soderlund et al., 1997; Soderlund et al., 2000; Pampanwar et al., 2005). Using FPC parameters of (Tolerance: 7, Cutoff $10^{-10}$) 150 BAC contigs were constructed most were small contigs containing few BACs. Twenty-one contigs contained five or more BACs, the largest with 11 BACs. BAC clones were sequenced by Washington University Genome Sequencing Center. Draft sequence of 4-6X coverage from both BACs was sequenced together with phred/phrap (Ewing and Green, 1998; Ewing et al., 1998) (http://www.phrap.org). BAC sequence was finished with custom primer PCRs sequenced with nested primers, sequenced at Iowa State University Sequencing Facility. Two rounds of finish sequencing closed 16 gaps leaving 2 gaps found in repetitive regions.

Finished sequences were repeat annotated with TEnest (Kronmiller and Wise, 2008) using the maize repeat database and default parameters. TEnest was used to produce the comparative repeat display of B73 and Wf9-BG (Fig. 1). Genes were predicted with three programs; GeneSeqer (Schlueter et al., 2003), FGENESH (Salamov and Solovyev, 2000), GeneMark.hmm (Lukashin and Borodovsky, 1998). Predicted gene models designed from comparisons from the three prediction programs. Protein and EST databases for *Arabidopsis thaliana*, *Avena sativa*, *Brachypodium distachyon*, *Hordeum vulgare*, *Oryza sativa*, *Saccharum officinaris*, *Sorghum bicolor*, *Secale cereale*, *Triticum aestivum*, and maize were retrieved from GenBank (Benson et al., 2006) and aligned to determine gene models using BLASTX, TBLASTN, and BLASTN . The comparative sequence alignments of Figure 2 were made with custom perl scripts. Sequences were aligned with WU-BLASTN (http://blast.wustl.edu), alignments were parsed according to E-value and displayed in SVG (http://www.w3.org/Graphisc/SVG). E-value ranges were displayed in different colors: Green: E-value $< 10^{-200}$, Blue: E-value $< 10^{-100}$, Gray: E-value $< 10^{-60}$.

**LITERATURE CITED**


CHAPTER 5. GENERAL CONCLUSIONS

Genome sequencing, by identifying the genes throughout a species, can help to understand how genes maintain the organism. Grass genomes in particular have the potential to greatly impact agriculture and world food production and provide alternative fuels sources. As expenses decrease, the ability to sequence genomes beyond model organisms becomes a reality. Some grass genomes have reached this stage, and more will in the near future. As increasingly complex genomes are sequenced, the abilities of assembly and annotation tools must also increase. Ultimately, correct assembly and annotation of full genomes provide assistance with our understanding of how genes work together in regulation, development, and maintenance of the genome and the processes by which genomes have evolved to their current states.

This thesis attempts to assist with the challenges of sequencing the ever increasing complexity of future genomes. Repetitive elements in the grasses have caused major genome expansions in certain species. The program TEnest was developed to annotate nested clusters of transposable elements to assist with identification and annotation of gene regions as well as provide tools for genome evolutionary studies. In addition, this thesis provides two processes for subclone assembly of highly repetitive sequences utilizing TEnest to determine clustered repeat structures. Sequencing of large maize contigs has provided a first look at the emerging landscape of the genome showing genes are tightly grouped among large clusters of repeats. Comparative sequence analyses between maize and other grasses show macro-colinearity is retained among genes across our large sequenced contigs but micro-colinearity is not retained. Between maize lines it shows varying frequencies of polymorphisms across sequenced regions and varying rates of SNPs between different maize lines.

As more and more genomes are sequenced, the abilities of the technology and tools for analyzing these genomes must increase in unison. Large sequence contigs have shown previously unobserved characteristics of the maize genome; with the genome sequence of maize and other grasses, the tools presented here will reach their true potential.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Roger Wise, for his help and assistance throughout my graduate studies. He lured me into the laboratory with promise of sequencing hugely repetitive maize BACs and, after attempting to assemble the first of our BAC sequences, I was hooked. I would like to thank Karin Werner for letting me join her on the maize side and helping me resurrect years of background information. And yes, for all the laboratory work, too. I would like to thank the rest of the Wise laboratory for sitting through many maize and transposable element laboratory meetings and providing many helpful discussions.

Most importantly I would like to thank my wife, Danelle, without whom none of this would be possible. She has made great sacrifices to move so far from our families for such a long time. Finally I want to thank my son, Beau, whom I have taken care of at home for the last five months. He has been extremely well behaved and has made thesis writing quite an experience.