Sequencing and distribution analysis of Dissociation transposon events throughout the Zea mays genome

Justin Andrew Schares
Iowa State University

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

Part of the Genetics and Genomics Commons

Recommended Citation
Schares, Justin Andrew, "Sequencing and distribution analysis of Dissociation transposon events throughout the Zea mays genome" (2007). Retrospective Theses and Dissertations. 14662.
https://lib.dr.iastate.edu/rtd/14662

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Sequencing and distribution analysis of *Dissociation* transposon events throughout the *Zea mays* genome

by

Justin Andrew Schares

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

MASTER OF SCIENCE

Major: Genetics

Program of Study Committee:
Erik Vollbrecht, Major Professor
Thomas Peterson
Volker Brendel

Iowa State University
Ames, Iowa
2007

Copyright © Justin Andrew Schares, 2007. All rights reserved.
# TABLE OF CONTENTS

## ABSTRACT

### CHAPTER 1: GENERAL INTRODUCTION
- Transposon History and Mechanisms 1
- *Ac/Ds* Transposon System 6
- Anthocyanin Production and the Regulator Gene *r1-sc:m3* 8
- Thesis Organization 10
- References 10
- Figures and Tables 18

### CHAPTER 2: *Ac/Ds* PROJECT SEQUENCING PIPELINE
- Abstract 23
- Introduction 23
- Methods and Materials 26
- Results 32
- Discussion 39
- References 42
- Figures and Tables 50

### CHAPTER 3: DISTRIBUTION OF INTRACHROMOSOMAL *DISSOCIATION* TRANSPOSITION EVENTS ACROSS MAIZE CHROMOSOME 10
- Abstract 63
- Introduction 63
- Methods and Materials 65
- Results 68
- Discussion 72
- References 75
- Figures and Tables 82

### CHAPTER 4: COMPARISON OF W22 INBRED LINE SEQUENCE AGAINST REFERENCE SEQUENCE FROM B73
- Abstract 92
- Introduction 92
- Methods and Materials 95
- Results 96
- Discussion 98
- References 99
- Figures and Tables 103

### CHAPTER 5: GENERAL CONCLUSIONS

### ACKNOWLEDGEMENTS
The DNA transposon *Dissociation* is a nonautonomous element derived from *Activator*. The *Ac/Ds* project, a collaborative effort to cast stable *Ds* insertions throughout the maize genome, is a potentially valuable tool for generating a reverse genetics resource in maize. By using *Ac::immobilized*, an *Ac* element that has lost the ability to mobilize but can still produce transposase, a *Ds* element in the anthocyanin regulating allele *r1-sc:m3* can be mobilized. Through genetic crosses, *Ac::im* can be segregated away, creating maize lines with stable *Ds* insertions that can be maintained through multiple generations. These new events are sequenced and located on the maize genome using BLAST. Through selection of genetic markers based on kernel color, transposition events into either linked or unlinked locations can be preferentially enriched for, allowing a broad distribution of events beyond the donor site. As of October 5\(^{th}\), 2007, 816 unlinked families and 126 linked families have been produced, processed, and sequenced, with an approximate 77% success rate. Repeat masking shows a low content of repetitive sequence, implying that these events are moving into nonrepetitive, genic regions. By examining transposition events into the donor chromosome, transposition patterns can also be observed. 95% of hits land within a 40 cM range of the donor site. The relative concentration of events increases closer to the donor site. Also, families that are not selected for unlinked events show a bias toward transposing to locations distal to the donor locus. Because the *Ac/Ds* project uses the W22 inbred line, sequence from large fDs fragments was compared against GSS assemblies from B73. The identity
between B73 and W22 is 96.5%, comparable to previous studies comparing B73 and Mo17.
CHAPTER 1: GENERAL INTRODUCTION

Transposon History and Mechanisms

Transposons were first identified by Barbara McClintock in the 1940s. In an experiment initially designed to determine the genetic composition of the short arm of chromosome 9 in Zea mays, she analyzed the effect of a gene referred to as C (McClintock, 1950). Plants homozygous for the recessive allele of C produced nearly colorless kernels, while the wild type produced dark purple kernels. She found that the colorless mutants would spontaneously revert to wild type. This phenomenon was attributed to a locus called Dissociation (Ds). When in the presence of another locus called Activator (Ac), not only was this locus capable of excision from and breaking of the chromosome, but could also causing a variety of other effects in nearby genes. Since the chromosome containing the C allele containing Ds was known to be fragile, while the chromosome in the revertant was not, it was clear that, while Ds was present in the original, it was physically gone in the revertant. Eventually, it was understood that these loci were physically changing position within the chromosome, hence the name transposons.

Since then, transposons have been found not only in Zea mays, but across all clades of life. At first glance, transposons appear to be simple DNA parasites, relying on the host genome as a means of preserving their own existence. To a certain extent, this is true. Transposons are parasitic elements, and as such are well suited to their own existence. They are capable of replication throughout the host genome, and can increase to significant number, but uncontrolled replication would likely disrupt the host DNA to
the point that their host is no longer capable of procreation, thus preventing them from being inherited across generations. To prevent this, transposons come under a variety of control mechanisms to prevent their excessive proliferations, such as inducible promoters that limit transcription of the transposon region, epigenetic mechanisms associated with DNA modifications, or modification of proteins involved in transposition (Kunze, Starlinger, 1989; Wang et al., 1996; Wang, Kunze, 1998).

Even with these regulatory mechanisms to keep them in check, transposons may still accumulate to fairly high copy numbers, and have had a huge effect on the evolutionary process. The presence of “jumping genes” has a significant mutagenic effect. A transposon insertion in the exons or regulatory regions of a particular gene can disable its function, producing new phenotypic effects. Upon insertion, transposase generates a staggered cut in the host DNA, into which the transposon inserts. DNA polymerase fills in the gaps made by transposase, generating a repeat of the sequence at the insertion site, called a target site duplication (TSD) (Fig. 1.1). Depending on how and where it is generated, this footprint can have mutagenic effects even if the transposon itself were to be excised. Transposons can carry new genes integrated into their structure, allowing for movement of or introduction of genes around the host genome. Transposons that accumulate mutations or InDels that destroy their ability to continue transposition become permanent parts of the host genome, in some cases becoming a highly significant contribution to host genome size.

In some cases, the mechanisms of the transposon have been requisitioned by the host to develop some novel capability. There is the case of V(D)J recombination in the immune system of vertebrate animals, in which the transposase of Transib transposon
superfamily may have been modified to allow splicing of various clusters of genes to form a wide variety of antibodies (Kapitonov, Jurka, 2005). In Drosophila, the telomeres of the chromosomes are maintained and elongated by an array of tandem elements that have been formed by targeted transposition events (Biessmann et al., 1994).

Transposons can be divided into two general classes: Class I and Class II. There are cases of other transposons that don’t neatly fall into either of these classifications (i.e., Mavericks, a class of very large DNA transposons), but the majority of known transposons are classified into these groups.

Class I transposons, or retrotransposons, use an mRNA intermediate copy to allow replication of the initial insertion (retrotransposons). Retrotransposons contain two open reading frames (ORFs), *gag* and *pol*. The *gag* ORF codes for structural proteins, while the *pol* ORF codes for reverse transcriptase. The progenitor transposon is transcribed into mRNA, after which the mRNA transcript is used as a template for synthesis of a double stranded DNA copy by reverse transcriptase, which is encoded by the retrotransposon. This new DNA copy can insert itself back into the host genome via integrase, thus allowing the transposon to spread across the genome (Fig. 1.2). However, the original element itself does not transpose beyond the initial integration site. Once integrated, these elements do not excise from where they’ve been inserted. Their stationary nature and efficient replication allows for greater overall increases in copy number, and, through inactivation of the coding regions, can greatly contribute to the host’s overall genome size. In the case of maize, it is estimated that retrotransposons make up about 66% of the over 1000 megabases of the maize genome (Wei, *et al*., 2007; Haberer, *et al*., 2005; SanMiguel *et al*., 1996)
Class II transposons are DNA-based transposons that typically transfer by conservative transposition, or “cut-and-paste”, where the transposon is excised and moves to a new location. Transposase recognizes the terminal inverted repeats (TIRs) of the transposon, excises it, and integrates it into a new location. Typically, a simple conservative transposition wouldn’t allow the transposon to replicate, as it’s merely a single copy moving from place to place, but events within the host DNA yield scenarios by which a transposon may be duplicated. The most likely scenario is transposition during genome replication, where the transposon can increase in number by moving from an already replicated region to either an unreplicated region or a replicated region on the other DNA strand (Fig. 1.3). However, copy number can also be increased by other genomic events, such as unequal crossing over or recombinant events. Though conservative transposition is the most common method of transposition in Class II elements, it’s by no means the only one. Helitrons, a special subset of Class II transposons, can replicate via a rolling circle replication cycle similar to that employed by some bacteria. Helicase separates the DNA strands of the transposon, allowing replication by DNA polymerase, resulting in a new double strand DNA transposon that can be integrated elsewhere in the genome.

Class II transposons can also be divided into another two general classes: autonomous and nonautonomous elements. An autonomous element is the complete, functional transposon, carrying everything necessary for its transposition. A nonautonomous element is a transposon that has lost functionality in one or more genes necessary for its transposition, but still retains intact TIRs, which can be recognized and transposition activated in trans by the transposase of a corresponding autonomous
element. Using the enzymes produced by the functional genes of the autonomous element, the nonautonomous element can transpose. For example, consider the \textit{Ac/Ds} transposon system. The autonomous element, \textit{Ac}, is capable of transposition entirely on its own. The nonautonomous element, \textit{Ds}, is incapable of independent transposition, usually due to an interruption in the gene coding for transposase. However, the TIRs of \textit{Ds} can still be recognized by the transposase produced by \textit{Ac}, allowing the element to move if the \textit{Ac} element is present in the same nucleus.

Transposons are present in all clades of life, but their evolutionary history may be significantly different from the host in which they reside. In cases where the transposon phylogeny doesn’t coincide with the species phylogeny, horizontal transfer is commonly cited as a reason why the trees may not match up. In horizontal transfer, some gene or segment of DNA jumps species boundaries. For transposons, this is an easy scenario to envision, as their mobile nature makes excision from the original organism and insertion into a new organism relatively easy to envision. In the case of Class I retroelements, some transposons show strong similarity to some corresponding retroviruses. The mechanisms involved in replication and integration are very similar between the two, the principle difference being the protein capsids that allow viruses mobility outside of the organism. It’s likely that transposons and viruses arose from some common ancestor, and the presence or absence of capsid protein genes can turn one into the other (Xiong, Eichbush, 1990). Transposons and viruses add a very dynamic dimension to an organism’s genome, and are a significant force in the evolution of life.
**Ac/Ds Transposon System**

As stated above, the *Ac/Ds* transposon system was the first identified case of transposition. *Ac/Ds* transposons are part of a larger group of Class II DNA transposons: the *hAT* superfamily. This superfamily is named for the transposons *hobo* from Drosophila, *Ac* from maize, and Tam3 from snapdragon, though many others have been added to this family, forming a very diverse group of elements. Hallmarks of the *hAT* superfamily include a similar transposase-coding exon structure and, on integration, produce an 8 basepair target site duplication. These transposons tend to have short TIRs with an ill-defined sequence, so some variety can be found from one transposon to another. The *hAT* superfamily is one of the more ancient transposon lines, with an ancestry that seems to predate the divide between plants, animals, and fungi, as the phylogenetic trees of various *hAT* elements closely match those of their hosts. This implies that these elements inserted in some common ancestor and remained as their hosts multiplied, differentiated, and evolved into their current forms. There doesn’t appear to be any significant evidence that horizontal transfer is involved in the evolution of this superfamily. However, the phylogenetic data doesn’t rule out the possibility of trans-kingdom horizontal transfer, the ancient ancestry hypothesis is more probable (Rubin et al., 2001).

As a member of the *hAT* superfamily, *Ac/Ds* transposons are ancient. As the first transposon to be discovered, it is also one of the best characterized. *Ac/Ds* has a long history in the grass families, particularly in *Zea mays*, the organism of discovery. *Activator*, or *Ac*, is a relatively simple 4.6 kb transposon, with 11 bp imperfect TIRs (CAGGGATGAAA on the 5’ end; TTTCATCCCTA on the 3’ end) (Fig. 1.3). The
transposon produces a 3.5 kb mRNA, divided between 5 exons, with a 2.4 kb open reading frame (ORF). This ORF codes for the Ac element’s transposase. Dissociation, or Ds elements, however, can vary significantly. In Ds elements, the exons encoding transposase have been interrupted, for example by insertion or deletion. In the examples shown in Fig. 1.4, we see various Ds transposons that have been produced due to deletion. For instance, the transposon Ds6 has had a significant deletion event, after which only about 1 kb of the originally autonomous Ac element remains on each end, resulting in a 2 kb nonautonomous element. Ds9 contains a much smaller deletion in the fourth exon. Ds1 has only around 35-41 bp at the ends of the transposon show homology to Ac, the remaining 409 bp being of unknown origin (Wessler, 1991). Ac/Ds transposons move by conservative transposition, and thus can only replicate via transpositions during genome replication or events in the host genome such as strand-slipping during replication or unequal crossing over. Ac/Ds transposons tend to be biased toward closely linked transpositions, forming clusters of new insertion sites within a few centimorgans of the original donor locus (Van Shaik, Brink, 1959; Greenblatt, Brink, 1962; Greenblatt, 1984; Dooner, Belachew, 1989; Jones et al., 1990; Dooner et al., 1991; Keller et al., 1993; Healy et al., 1993; Dooner, Belachew, 1994). However, Ac elements do not appear to be restricted to linked transposition, as about one half of transposition events are to unlinked sites (Dooner, Belachew, 1989). The number of active Ac elements also has a negative dosage effect on transposition, as two active Ac elements will result in proportionally less excision events than a single active Ac. Integration and excision of Ac/Ds can have varying effects, depending on the integration site. In the case of McClintock’s study of C1, Ds was located in an intron, so the footprint left by the excised
transposon did not affect anthocyanin production. Another anthocyanin-producing gene, $A1$, has an allele called $a1-m3::Ds$. In this case, $Ds$ is inserted in an exon, yet doesn’t fully disrupt anthocyanin production, yielding a pale purple phenotype (unpublished data). If the $Ds$ element is excised, the footprint left behind results in a frame shift in the ORF of the gene, rendering it inoperable (colorless).

**Anthocyanin Production and the Regulator Gene $r1-sc:m3$**

In Barbara McClintock’s experiments involving transposons, the effects of transposon movement were visible due to the movement of an insertion in a gene affecting the coloration of the maize kernels. Anthocyanins are plant pigments resulting in a deep purple coloration. Anthocyanin production is a process mediated by a series of proteins, both structural and regulatory (Fig 1.5). A molecule of p-Courmaryl-CoA and three molecules of Malonyl-CoA are condensed by Chalcone synthase (CHS) into 4,2’,4’,6’-tetra hydrocychalcone, a yellow pigment. This is modified into naringenin and then dihydrokaempferol (DHK) by chalone isomerase (CHI) and flavanone 3-hydroxylase (F3H), respectively. DHK can be further modified into dihydroquercetin (DHQ) by flavanoid 3’-hydroxylase, and both DHK and DHQ can be converted to dihydromyricetin (DHM) by flavanoid 3’5’-hydroxylase. These three colorless dihydroflavonols are converted into anthocyanins by three more enzymes: dihydroflavanol 4-reductase (DFR) reduces the dihydroflavonols to flavan-3,4-cis-diols, or leucoanthocyanidins. Anthocyanidin synthase (ANS) converts the leucoanthocyanidins to colored anthocyanidins by oxidation and dehydration. Finally, anthocyanin glycosyltransferase (3GT), glycosylates the anthocyanin, a necessary step in cases where the anthocyanin-glucosides are further modified into anthocyanin-
rutinosides. (Holton and Cornish, 1995). These pigments accumulate in the plant tissues, turning the maize plant and its kernels a strong purple color.

Several genes are responsible for regulation of the anthocyanin pathway in maize, including where in the plant the pigments are produced. The $R$ locus ($red-1$), a gene on chromosome 10 coding for a basic helix-loop-helix protein, acts as a tissue specific regulator of anthocyanin production in the kernel and tassel. The $B$ locus, a homologous gene on chromosome 2 that is similar in function to $R$, controls anthocyanin production in the leaf and stalk. Both work in conjunction with the $C1$ and $Pl1$ loci, which code for DNA binding proteins. Together, these genes control the expression of the CHS, DFR, and 3GT enzymes in the anthocyanin production pathway (Holton and Cornish, 1995). If these gene products are not expressed, anthocyanins are not produced, resulting in a yellow or white kernel. In some alleles, the $R$ locus is inactive due to the insertion of a transposon, which interrupts the coding region and prevents the regulator protein from being properly synthesized.

Several transposon alleles exist of the $R$ locus, with varying degrees of reversion. For these experiments, the allele of principle interest is $rI-sc:m3$. $R-sc$ alleles are responsible for a strong coloration effect in the kernels. $m3$ refers to specific mutant produced by integration of a $Ds6$ transposon. $rI-sc:m3$ is colorless but has a strong tendency to revert, and upon reversion may result in a very strong colored phenotype (Alleman and Kermicle, 1993). This strong phenotype allows for relatively easy visual assessment of the functionality of the $R$ gene, and along with it, the presence or absence of the $Ds6$ transposon.
Thesis Organization

My thesis as follows will focus on the protocols and results of the Ac/Ds Project, an effort to cast Ds elements and track them throughout the genome of Zea mays (www.nsf.gov/awardsearch/showAward.do?AwardNumber=0501713; www.plantgdb.org/prj/AcDsTagging). This is a collaborative project between the labs of Tom Brutnell (Boyce Thompson Institute), Erik Vollbrecht (Iowa State University), and Volker Brendel (Iowa State University). First, I will cover the basics of the sequencing pipeline developed for the identification and localization of the newly cast Ds elements. Over the past year and a half, we’ve developed and refined a protocol to streamline isolating, sequencing, and processing our newly created stable Ds insertion families, with the intention of providing lines with insertions that may be useful for further study to maize researchers.

Second, using data produced from the pipeline, we will look at the transposition patterns of the linked transposition events on maize chromosome 10 to determine general traits of Ac/Ds transposition.

Third, we will compare the sequence produced by the sequencing pipeline (W22 inbred line background) against archived sequence of the maize genome (B73 inbred line background) as a way of assessing differences that may have been selected for in the development of each line.

References


Alleman, M., Kermicle, J.L. 1993 Somatic variegation and germinal mutability reflect...
the position of transposable element *Dissociation* within the maize R gene. Genetics 135: 189-203


Dooner, H.K., 1979. Identification of an R-locus region that controls the tissue specificity of anthocyanin formation in maize. Genetics 93. 703-710.


Genetics 78. 691-701.


Kapitonov, V., Jurka, J. 2005 RAG1 core and V(D)J Recombination signal sequences were derived from Transib transposons. PLoS Biology. Vol. 3, Iss. 6. 0998-1011.


Kunze, R., 1996. The maize transposable element Activator (Ac). Current Topics in
Microbiology and Immunology 204. 161-194.


Figure 1.1: Generation of a target site duplication by a transposon. When the enzyme transposase generates a site for transposon integration, an uneven cut is generated in the host DNA. When the transposon inserts, these gaps are filled in by DNA Polymerase, creating a duplicated region on either side of the transposon. If the transposon were to excise, this duplication would remain.
Retrotransposons rely on an mRNA intermediate for replication. An mRNA copy is generated from the original transposon. A reverse transcriptase encoded for by the transposon generates a single stranded DNA copy of the transposon, and DNA polymerase synthesizes the complementary strand. Integrase, another enzyme encoded for by the retrotransposon, nicks the host DNA and inserts the newly synthesized transposon.

Figure 1.2: General model of replication of a Class I Retrotransposon
Retrotransposons rely on an mRNA intermediate for replication. An mRNA copy is generated from the original transposon. A reverse transcriptase encoded for by the transposon generates a single stranded DNA copy of the transposon, and DNA polymerase synthesizes the complementary strand. Integrase, another enzyme encoded for by the retrotransposon, nicks the host DNA and inserts the newly synthesized transposon.
Figure 1.3: Replication of a Class II DNA transposon during host DNA replication. Because most DNA transposons do not generate copies of themselves, they replicate through other means. If a transposition event were to occur during replication of the host DNA, the extra copies of the transposon could be generated if it is replicated before the surrounding regions. If a replicated transposon mobilizes to a region not yet replicated, that transposon would be replicated again as the host site is (below). If the replicated transposon mobilizes to the other replicated strand, one of the host genome copies would carry two transposons and the other none (above).
Activator (Ac) Element

Dissociation (Ds) Elements

Ds9

Ds 2dl

Ds 2d2

Ds6

Ds 1

Figure 1.4: General structure of Activator element and various types of Dissociation elements.

Activator elements are a fairly simple 4.6 kb DNA transposon. The transposon is flanked by 11 bp imperfect terminal inverted repeats, and carries five exons, which code for the enzyme transposase. Several types of Dissociation elements are generated by various insertions, deletions, or mutations that render the transposase-coding regions inactive, but remain mobile as transposons because their TIRs remain intact. Element Ds1 only carries the very ends of the Ac element, but the remaining 409 bp is of unknown origin.
Figure 1.5: Anthocyanin synthesis pathway. Anthocyanins are produced from p-Coumaryl-CoA and Malonyl-CoA through a series of enzymatic reactions. Enzymes circled in purple are regulated by the $R$ and $B$ loci.
CHAPTER 2: Ac/Ds PROJECT SEQUENCING PIPELINE

Abstract

As part of the Ac/Ds project, which is dedicated to producing maize lines with multiple, stable insertions of the corresponding nonautonomous Dissociation (Ds), we have been developing a pipeline that will allow for efficient production, isolation, and analysis of these new maize lines. This system makes use of Ac-immobilized, an Ac element on chromosome 7 with a deletion in its 5’ terminal inverted repeat (TIR), as well as r1-sc:m3, an anthocyanin regulating gene on chromosome 10 that is interrupted by a Ds6 element that is used as the Ds donor for producing new insertion sites. Through selection of specific kernel phenotypes, we can select for donor-unlinked and -linked transposition events. As of October 5th, 2007, the pipeline has produced 941 unlinked transposition families and 151 linked transposition families. Of these, 816 unlinked transpositions and 126 linked transpositions have been successfully sequenced and compared against the current maize sequence. At current specifications, we can produce around ~200-300 stable tr-Ds maize lines every month, and, of these, ~149-224 can be sequenced in a single pass through the pipeline.

Introduction

Due to their transportable nature, transposons can be a useful genetic tool. Their ability to carry other segments of DNA, including functional coding DNA, allows integration of new genes or regulatory features into a host genome. If the sequence of said transposon is known, it’s possible to move it around a host genome, and then, through the use of transposon specific primers or probes, analyze the region the
transposon has integrated into. Transposons can be used for mutagenic tests, and can be especially useful to plant geneticists, as there is yet to be an efficient method of gene replacement.

However, transposons do have their drawbacks. There is no guarantee that when a transposon removes itself from one site that it will successfully integrate into another. The transposon’s ability to move is itself a double-edged sword. Though the movement of a transposon around the host genome can be used for exploration, this same mobility also makes any new insertions unstable, as the transposon may just as easily leave the new site as it could the donor site.

The DNA transposon *Activator*, or *Ac*, can potentially be useful for forward genetic screening. *Ac* transposition is biased toward closely linked sites, allowing for one insertion near a gene of interest to be used as a donor for regional mutagenesis (Greenblatt, 1984; Dooner, Belachew, 1989). This localized transposition can also yield multiple alleles of a specific gene due to multiple locations of transposon insertion. Despite these abilities, *Ac* still has its drawbacks. There are relatively few *Ac* elements inserted into convenient reporter genes. More are placed near translocation breakpoints or molecular markers on recombinant inbred lines, requiring mapping experiments or molecular assays to determine the integrity of the line (Dellaporta and Moreno 1994; Auger and Sheridan 1999; Dooner *et al*. 1994; Kolkman *et al*. 2005)

The goal of our *Ac/Ds* project is to cast *Ds* elements throughout the maize genome for development of stable insertions than can be propagated through multiple generations. This is accomplished through the use of a two-component transposition system involving the nonautonomous *Dissociation* element *Ds6* interrupting the regulatory gene *r1-sc:m3*
(Fig. 2.2) and the modified Ac element Ac-immobilized (Ac::im). Ac::im is an immobilized Ac element on chromosome 7 in which the 5’ terminal inverted repeat (TIR) has been lost, effectively removing the ability of Ac::im to transpose. However, Ac::im can still produce transposase. A Ds element, with its TIR regions intact, can be transposed by this transposase supplied in trans, thus allowing Ds to move around the maize genome (Conrad, Brutnell, 2005).

An advantage here is that, while transposition is possible when both elements are present in the host genome, the two elements are unlinked (Ac::im on chromosome 7, and r1-sc:m3-Ds on chromosome 10). This allows for segregation of the elements through genetic crosses. The presence of an active transposase source can result in some regions of the endosperm having an activated R gene, causing a spotted phenotype. This allows us to select against Ac::im in further crosses, resulting in new Ds insertion lines without transposase moving them further in subsequent generations. A similar two component system has been used successfully in tomato, casting a genetically marked Dsneo element throughout the genome using an immobile Ac3 (Healy et al., 1993). Because the donor Ds is present in a gene regulating anthocyanin production, excision results in distinct phenotypic changes that can be readily observed. This colored kernel phenotype can be used to select events that may contain new Ds insertions. We can also test linkage of new insertions to the donor locus. The Ac/Ds transposon system is biased toward transposing to closely linked sites, limiting its typical range. However, through selection based on the colored phenotype, we can select for those events that are unlinked to the donor site, allowing us to focus our efforts on new transposition insertions that may move further throughout the maize genome.
Materials and Methods

Genetic crosses for casting of *Ds* Elements

The casting itself is accomplished through a series of crosses designed to allow the *Ds* element to move from its donor site, while preventing the transposed *Ds* from making subsequent transpositions in future generations (Fig 2.1). In the first cross (Stage 0), plants homozygous for *Ac::im* and the *Ds* insertion in *r1-sc:m3* (female) are crossed by *r1-sc:m3* tester plants (no *Ac::im*, *Ds* inserted in *r1-sc:m3*) (male). This produces Stage 1 ears, which carry two types of seed: purple seeds (Colored), where the *Ds* was transposed before differentiation of the gametophyte, and yellow seeds with purple spots (spotted), where the *Ds* was transposed after differentiation. The spotted kernels are discarded. The purple seeds are then grown up and the plants (female) crossed with testers (male) once again (Stage 2). Three types of seed are produced: purple (Colored), yellow (colorless), and yellow with purple spots (spotted). As before, spotted kernels are not useful. The colorless kernels are collected for further analysis and are referred to as the “R-unlinked” families.

To determine if any of the purple kernels still carry *Ac::im*, several purple kernels from each ear are grown up (female) and crossed with testers (male). If the ears produced have purple, yellow, and spotted kernels, *Ac::im* is still present, and the kernels are not used. If the ears produced purple and yellow kernels only, *Ac::im* has been segregated out. The purple kernels from these ears are assembled into the “R-linked” families.
The kernels are organized into families named by when and where the plants were grown: (BTI or ISU field).(season and year).(family number of that crop) (i.e., B.S05.0006: BTI field. Summer of 2005. family number 0006).

**DNA Extraction and Southern Blotting**

These steps are performed at the Boyce Thompson Institute. Ten kernels from each family are germinated in a sand bench and grown until approximately four inches tall, at which point tissue samples are taken for DNA extraction using phenol/chloroform. The genomic DNA is digested using either PvuII or SacII enzymes (depending on situation) and Southern Blots are performed using a $Ds$-specific probe. Bands on the Southern blot not corresponding to the band produced by the donor $Ds$ from $r1-sc:m3$ (3.0 kb band in PvuII digests, 2.0 kb band in SacII digests) indicate a new tr-$Ds$ present in the family.

**Inverse PCR**

These steps are performed at the Boyce Thompson Institute. Genomic DNA digested by either PvuII or SacII is once again used. The digest is then diluted by and used in a ligation reaction. The ligated DNA is then used as a template for IPCR, using platinum Taq enzyme (Invitrogen) and primers Lc24 5’-TTGTTGCAGCAGCAATAACACAGCAT-3’ and Lc18 5’-CCTTGGTTTGATTGGCTGCTA-3’, which are located between the PvuII and SacII sites in the donor $Ds$, and thus present regardless of which enzyme is used (Fig. 2.2). DMSO and 0.5M Betaine are also added to the reaction. However, in cases of larger product, these additives are removed, as their presence appears to benefit amplification of the parental fragments more than tr-$Ds$ fragments. IPCR products are separated on a 1.2%
agarose gel containing ethidium bromide. The bands produced from the IPCR reaction are compared against the bands present on the Southern blot to insure the fragment’s identity (Fig. 2.3).

**Isolation of IPCR Bands and Reamplification**

These steps are performed at the Boyce Thompson Institute using protocols that were designed in collaboration between BTI and myself at ISU. IPCR bands are excised from the agarose gel and extracted using QIAGEN’s gel extraction kit and protocol, and resuspended in 50µL HPCL grade water. The purified IPCR product is used as template in a standard PCR reaction using primers specific to the digestion reaction used for IPCR (PvuII digestion: JSR01 (5’-GTTCGAAATCGATCGGGATA-3’ and Lc18; SacII digestion: JGp2 5’-CCGGTTCCCGTCCGATTTCG-3’ and Lc45 5’-GTGCTGTACTGCTGTGACTTGTG-3’ (Fig. 2.4 and Fig. 2.5). The reactions are assessed by agarose gel for product quantity and size (note: due to placement of primers, reamplified product is approximately 1.0 kb smaller in PvuII-digested samples and 0.8 kb smaller in SacII-digested samples, relative to the IPCR product size). The product is purified either by ethanol precipitation (single band) or by QIAGEN gel extraction (multiple bands) and freeze-dried for concentration and shipping.

**Sequencing**

Freeze-dried IPCR product is resuspended in 10µL of HPCL grade water. Concentrations are assessed using an ND-1000 spectrophotometer (Nanodrop). Ideal concentrations for sequencing are 2.5 ng/100 bp/µL. Lower concentrations can yield sequence, but at a lower success rate. A minimum of 1.5 ng/100 bp/µL appears to be necessary for reliable sequencing reactions. Higher concentrations are diluted to 2.5
ng/100 bp/µL. Samples are submitted in 96-well plate format using primers specific to the digestion used in the IPCR reaction (PvuII digestions use primers JSR03 (5’ CGATCGGGATAAATACAAAATC 3’) and Lc18; SacII digestions use JSR05 (‘5 CGTCCCGCAAGTTAAATATGA 3’) and JSR04 (5’ ACACAACAGCTTGGTGCAAT 3’)) (Fig. 2.4 and Fig. 2.5). Sequencing reactions were performed by the DNA sequencing facility at Iowa State University. An average sequencing reaction typically yields ~600 bp of viable sequence.

**Second Reamplification**

In cases where the purified and freeze-dried samples have failed to sequence, low DNA concentrations are a major factor. In such cases, another round of PCR may be necessary to increase sample concentration. PCR is performed using the failed sample as a template. The primers are those used for the sequencing reaction (see below), as these are nested primers, increasing overall success and quality of the PCR reaction. The polymerase used is EconoTaq (Lucigen). Samples that reamplify properly are purified by ethanol precipitation and resuspended in HPCL grade water. Concentration optimization and submission are the same as above.

**TA Subcloning**

Due to the cost and time intensive nature of this method, it is reserved for samples that appear to contain multiple overlapping bands that are difficult or impossible to separate properly. If more PCR product is needed, ExTaq (Takara) is used for any PCR reactions to increase sequence fidelity. 2µL of sample is fed into a TOPO TA vector cloning reaction (3µL total reaction), to be used to transform 10µL of TOP10 chemically competent cells (Invitrogen). The cells are plated on luria broth/carboxycillin/X-gal agar
plates overnight. Depending on the number of colonies present, 4-12 white colonies are picked and transferred into a 96 well culture block containing LB/carboxycillin media for overnight culture. Plasmid extraction was performed by the DNA sequencing facility at Iowa State University. After extraction, plasmids are diluted 1:100 and their inserts assessed by PCR reaction (using sequencing primers). Those containing the properly sized band may be submitted for sequencing using the undiluted stock (Since each colony is the result of a single plasmid insert propagated over generations, multiple samples must be sequenced to insure accuracy between samples). The optimal concentration for the plasmid sequencing reaction is 0.25 µg/µL. Sequencing reactions were performed by the DNA sequencing facility at Iowa State University.

**Sequence Processing**

Sequencing reads are first passed through a perl script designed to convert the file names to our standard name format: (Family ID)_ (Sequencing Primer).ab1. (In the case of reads from the restriction enzyme end, naming format is (Family ID)_ (Reverse Sequencing Primer)_ (Forward Sequencing Primer)_dir.ab1). All sequences are opened in Mek and Tosj’s 4peaks trace file viewer (mekentosj.com/4peaks/) to confirm the status of the read. If a reaction has failed or the read is less than ideal, annotations are added to the file name. Traces are then fed into Pregap4 assembler (Staden), which aligns the reads, cleaves off $Ds$ sequence, and prepares the files for entry into Gap4 (Staden), an assembly and editing program. Editing functions include checking alignments made by Gap4, correcting errors between reads, and confirming the TSD and restriction enzyme junctions. Sequences are arranged so that the TSD junction point is at the 5’ end of the sequence and the restriction enzyme site at the 3’ end.
Pipeline Repetition

In cases where a sample has failed to sequence and cannot be reamplified to a satisfactory degree, that family may need to be repeated from the beginning of the pipeline. Seed from said family is again germinated in sand bench, and passes through the pipeline as before.

Sequence Organization

Five cumulative, multi-sequence fasta sequence files are built using information produced from each round of sequences. File names are based on the date of the sequencing reaction. 1.) cons_new_MMDDYY.fa contains all the R-unlinked sequence obtained from the most recent sequencing reaction. 2.) cons_R_new_MMDDYY.fa contains all the R-linked sequence obtained from the most recent sequencing reaction. 3.) cons_all_MMDDYY.fa contains all the R-unlinked sequence produced up to the current date. 4.) cons_R_all_MMDDYY.fa contains all the R-linked sequence. 5.) cons_net_MMDDYY.fa contains all sequence produced up to the current date, except all samples containing the exact same sequence are collapsed into a single entry.

BLAT Analysis

The newest fasta files are queried against the previous cons_net using the BLAT (BLAST Like Alignment Tool) alignment program developed by Jim Kent at University of California Santa Cruz (http://genome.ucsc.edu/). Samples in the newest files that are similar to previously produced sequences are examined. Samples with the same sequence, including the same TSD and restriction enzyme junction points, are collapsed into a single entry in the newest cons_net file. The cons_net file also contains the Ds
sequence, as well as \textit{r1-sc:m3} with the donor \textit{Ds} insertion, allowing for filtering of parental and transposition events within \textit{r1-sc:m3}.

**BLAST Analysis**

All the sequence produced is queried against the maize genome using megaBLAST (highly similar sequences) on the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/). Comparison is made against the high throughput genomic sequences (HTGS) of the maize genome. Hits were classified as having at least 95% coverage of the query sequence and at least 95% identity against the reference sequence. The latest BLAST analysis was performed October 5th, 2007.

**Repeatmasking**

Sequences are run through Smit, Hubley, and Green’s RepeatMasker program, version open-3.1.8 (www.repeatmasker.org), using the TIGR maize repeat database version 4.0 (http://maize.tigr.org).

**Results**

**Current Pipeline**

The \textit{Ac/Ds} project sequencing pipeline as it stands now consists of several steps (Fig. 2.6). First, maize homozygous for \textit{Ac::im} and \textit{r1-sc:m3-Ds} is crossed by maize homozygous for \textit{r1-sc:m3-Ds} (tester). This can produce two types of kernel, depending on when \textit{Ac::im} mobilizes the \textit{Ds} from \textit{r1-sc:m3}. If \textit{Ds} is removed before or during division of the gametophyte, anthocyanin is produced in all cells of the kernel, resulting in purple kernels. If \textit{Ds} is mobilized after the gametophyte cells begin to divide, some cells will still carry the \textit{Ds} insertion in \textit{r1-sc:m3}, while other cells will not, resulting in a yellow kernel with purple spots. Because of the chimeric nature of the spotted seeds,
they are not useful for developing stable $Ds$ lines, and thus discarded. The purple kernels are collected. These plants are heterozygous for $Ac::im$, so to produce a stable tr-$Ds$ line, another generation must be produced. The purple seeds are grown up the next season and crossed again with the tester line. This can produce three types of kernel. If $Ac::im$ is present in the resulting seed, two types of seed can be produced: purple and yellow spotted. If $Ac::im$ has not been inherited, the two types of kernel produced are purple and yellow. Since these yellow kernels don’t carry the recently reverted $R$ allele that the tr-$Ds$ originated from, any tr-$Ds$ insertions found in these seeds are more likely to be unlinked to the donor. These yellow kernels are collected and fed into the pipeline, where they are referred to as $R$-unlinked families. Since the goal of the $Ac/Ds$ project is broad distribution of $Ds$ throughout the maize genome, these are the lines that are used for distribution to the maize community. The purple kernels, which may still carry tr-$Ds$ insertions linked to the donor site, are still useful for comparisons and further analysis, but it is impossible to determine which among these kernels still carries $Ac::im$ and which does not. Another generation of crosses could determine the genotype by crossing the plants germinated from these purple kernels by the tester line. If $Ac::im$ was passed down to the purple kernel, this cross will generate purple, yellow, and spotted seeds. Spotted seeds, since they represent transposition activity during endosperm development, allow us to determine which plants still carry the transposase-producing $Ac::im$. If $Ac::im$ was not inherited by the seed, the cross will yield only purple and yellow seeds. The purple kernels from the ears lacking $Ac::im$ are collected and passed through the pipeline. These kernels are assembled into the $R$-linked families.
Once the maize lines have been produced, it is necessary to determine which of these lines contain new transposition events. These early steps are performed by the Brutnell lab at the Boyce Thompson Institute at Cornell University. Seed from these lines is germinated in a greenhouse sand bench. Genomic DNA is extracted from the plants. Restriction enzyme digestions are performed using PvuII and SacII. Southern blotting is performed to determine which families carry new tr-Ds insertions, depending on the size of the fragments produced from the digest. Parental fragments, which represent the Ds present in r1-sc:m3, are present in all families. IPCR is then performed using primers located between the PvuII and SacII digestion sites in Ds, allowing these primers to be used regardless of the digestion reaction. These samples are compared against the Southern Blot results to insure that the IPCR product is what was predicted. These samples are run out on a gel, and the desired bands excised and purified for another round of PCR using primers specific to the digestion reaction. These samples are purified either by ethanol precipitation or gel extraction, depending on how many bands are produced by the reamplification. These samples are freeze-dried and shipped to the Vollbrecht lab at Iowa State University.

Once we receive the freeze-dried samples at ISU, the DNA is resuspended and the concentrations determined and optimized. Then the samples are sequenced using a specific pair of sequencing primers. For samples that don’t produce sequence, PCR is performed using the corresponding sequencing primers. If no product can be produced from the received sample, they are reattempted from the seed stocks. If product is produced, it is purified and sequenced again. If sequencing still fails, further and larger PCR reactions may be used to increase recovery. If a sample still cannot be sequenced,
or if there appear to be multiple unique sequences present (perfect $Ds$ sequence, but multiple peaks after the junction), the fragment is cloned for isolated, high-quality sequence. The sequences are assembled and processed, then compared against previous sequences produced as well as the parental sequence. These processed, cleaned sequences can then be analyzed to determine where these insertions may be in the host genome.

**Pipeline Development**

Though we’ve developed an efficient and reliable method for finding and sequencing new $Ds$ flanking locations (fDs), the earliest stages of the pipeline were significantly different. Originally, bands were isolated and sequenced entirely by TA subcloning (Fig 2.7). While this worked well in the beginning, the high cost and lack of rapid throughput relegated this method to a last resort of isolating sequences. Under this method, multiple colonies must be selected, sequenced, and compared to insure reliable sequence is obtained, limiting how many samples could be processed at once. Also, due to the time necessary to transform, grow, and isolate cells carrying the proper insert, weeks may be required to process even a relatively small batch of samples. While this method may be reliable for producing high quality sequence, such a method is too long and intensive to consider for any streamlined pipeline. In these stages, we were also sequencing using either the original IPCR primers or primers some distance away from the junction points. Because of the extra $Ds$ sequence produced by these primers, less flanking sequence is obtained from each direction, limiting the size of the final sequences and complicating the informatics.

**Pipeline Results**
As of October 5th, 2007, 1222 fDs regions have been identified and isolated by
IPCR (1081 from unlinked families, 141 from linked families). Of those, 941 have been
successfully sequenced (77.0% success rate), the majority obtained on the first
sequencing attempt (See Table 2.1). Of these, 816 are from $R$-unlinked families, while
the remaining 126 are from $R$-linked families. The average PCR fragment after
reamplification is \(~0.6\) kb from an IPCR reaction using PvuII, while the average SacII
product is \(~1.6\) kb. Removing residual $Ds$, the finished sequences are around 0.5 kb and
1.3 kb, respectively. The current projected production rate of fDs fragments by IPCR is
\(~200-300\) per month. If success rates hold, \(~149-224\) should be sequenced every month.

**tr-$Ds$ Distribution**

The sequences were compared against bacteria artificial chromosomes (BACs) of
the ongoing maize genome sequencing effort via megaBLAST (see Table 2.2). The most
recent BLAST was performed October 5th, 2007. Of the 816 $R$-unlinked sequences, 371
fail to align to any known maize BACs. 351 align to a BAC or BACs at a single location,
while the remaining 94 align with multiple BAC locations. Of the 126 $R$-linked
sequences, 56 fail to align to any known BACs, while 54 align to a single BAC location.
The remaining 16 also hit multiple BAC locations. The samples that hit a single BAC
with chromosomal annotation were also compared. In the $R$-unlinked families, a broad
distribution can be seen across all chromosomes. In the $R$-linked families, 61% are
present in the chromosome 10, the $Ds$ donor chromosome, while the remaining events are
lightly distributed over the other chromosomes. Distribution tends to follow with
chromosome size, with the larger chromosomes typically receiving more hits than the
smaller ones. Chromosome 10, though, as the $Ds$ donor chromosome, would logically
have a disproportionally higher number of insertion events, and has not been included in any of the following statistical analyses. Based on the genetic map, the distribution across the remaining chromosomes is not what is expected by random distribution ($\chi^2=15.87$, p-value=0.0443). Based on statistical analyses of genetic distances of individual chromosomes, there appears to be a bias against events in chromosome 7 ($\chi^2=3.41$, p-value=0.050). This may be due to the fact that Ac::im is located on chromosome 7, and by removing it through successive crosses, we may also be selecting against tr-Ds events that hit that particular chromosome. The only other chromosome to show bias was chromosome 8 ($\chi^2=3.83$, p-value=0.0383). The reason for this is not known, but it may represent a region that is less susceptible to Ds insertion. If chromosomes 7 and 8 are removed from the statistical analysis of distribution, the distribution appears random ($\chi^2=8.63$, p-value=0.1955). Based on the physical map, the distribution across the chromosomes does coincide with random distribution ($\chi^2=8.72$, p-value=0.3665). The distribution across individual chromosomes also appears random.

The maize genome can be broken into two general classes: repetitive and non-repetitive regions. Repetitive regions contain various retroelements, duplications, etc. Non-repetitive regions carry genes. The fDs sequences were fed into RepeatMasker, using the TIGR repeat database for identification, to determine transposition biases relative to gene regions of the maize genome. These repeat results were compared against those for random segments of the whole genome (Messing et al. 2004) as well as for TIGR GSS contigs and TIGR methyl-filtered contigs and high c0t contigs (Table 2.3). In Messing’s sequence assemblies, repeat content was 58% across 307 Mb of sequence. The TIGR GSS, which is considered to represent the gene space, shows 17.3%. Two
subsets of the GSS sequences, methyl-filtered and high c_{ot}, have 18.4% and 8.6%, respectively. When the fDs sequences were processed through RepeatMasker, 8.0% of the sequence was repetitive elements. There did not appear to be any significant difference in repeat content between the R-unlinked and R-linked samples. Because of the low repeat content found in the fDs sequences, it appears that tr-Ds are biased toward transposition into non-repetitive, and thus genic, regions of the maize genome.

The number of genes that contain tr-Ds events was estimated using the Ac/Ds database at PlantGDB (http://www.plantgdb.org/ZmGDB/DisplayGeneAnn.php?ds=1). To estimate gene locations, polypeptide sequences from rice proteins (TIGR 5.0) are aligned to sequenced maize BACs by use of genomethreader (Gremme, et. al., 2005). The fDs sequences are also placed on these BACs. Of the 405 fDs fragments sequenced and aligned to current maize sequence, 112 align to regions containing putative maize genes (~27.7%). Of these, 93 are in R-unlinked families, and 19 are in R-linked families. We can expect these numbers to rise as more genes are identified and annotated to the maize sequence.

Redundant Transposition Events

Of the 941 fDs fragments we have sequenced, not all these events are unique. Of these, 114 sequences collapse into only 45 unique contigs (Table 2.4). These sequences have identical TSD junctions, implying that these are not merely very close events but likely the exact same one. Seven pairings appear to be cases of incomplete digestion in the Southern and IPCR steps, giving the illusion of two transposition events in a family when there is in fact only one (i.e., B.W06.0916A and B.W06.0916B). The remaining 38 contigs usually contain between two and four sequences, though there are cases of five
and six-sequence contigs. 23 of these contigs are composed of families from the same field, implying that they may be a repeating event within that generation. However, seven of these contigs appear to jump across generations. These events may signify nucleotide hot spots, or could be cryptic $Ds$ insertions ubiquitous across these families.

**Discussion**

Using a two-component transposition system of the unlinked loci $Ac::im$ and donor $Ds$, the $Ac/Ds$ project has generated over a thousand maize lines with new, stable $Ds$ insertions across the genome. Over the past year and a half, the methods by which we isolate, analyze, and locate these new insertions has changed significantly, greatly increasing both the success and the speed by which we can determine what new insertions have been generated in a particular maize family line. In its earliest stages, the time and effort necessary to adequately isolate and sequence a mere dozen new insertions could take a month or more. Now, with our refined pipeline, a few hundred can be done in a matter of weeks. Overall sequence quality and success rates have risen significantly. Our BLAST results also show how effective selection against linked transpositions works in the $R$-unlinked families. Compared against the $R$-linked sequences, of which over half hit the donor chromosome, the $R$-unlinked families show a broad distribution of tr-$Ds$ insertions across the maize genome. In previous studies, half the transposed $Ac$ elements land in sites unlinked to the donor location (Dooner *et. al.*, 1994). Our pattern of $R$-linked events is in line statistically with these studies ($\chi^2=2.66$, p-value=0.1029). Based on the genetic map, distribution does not appear to be random across the other chromosomes. Chromosomes 7 and 8 show significant biases against new tr-$Ds$ events, while the other chromosomes show a random distribution. The bias against chromosome 7 is likely due
to selection against \textit{Ac::im}, but reason for the bias against chromosome 8 is unknown. In a previous study of transpostions of \textit{Ac} from the \textit{bz1-m2} locus on chromosome 9S to unlinked sites (Dooner, \textit{et. al.}, 1994), significant biases were found in \textit{Ac} transposition patterns across the maize genome. Of the 24 events mapped to non-donor chromosomes, several biases could be found. Chromosomes 5, 6, and 7 show higher concentrations of events, while chromosomes 1, 2, 3, and 4 show lower concentrations of events. However, in our mapping efforts of 308 events, only chromosomes 7 and 8 show biases against tr-\textit{Ds} events. There are no biases for or against tr-\textit{Ds} events in the other chromosomes. It’s worth noting that while the distribution in the previous study does not appear to be random, the overall sample size was fairly small, so the distribution may not have been ideal. Another potential explanation of the observed distribution is that transposition frequency may be affected by the physical size of the chromosomes. One interpretation of these data is that the distribution of transposition events is primarily based on physical size, in terms of which chromosomes are hit, and the bias of events into gene space may be secondary. This would be supported by the results of chromosome 8, where the distribution based on physical size appears random ($\chi^2=0.90$, p-value=0.3428). Given the much broader distribution of tr-\textit{Ds} events across the maize genome in the \textit{R}-unlinked families, it appears that efforts to select against closely linked transposition events have been effective. There also appears to be a strong bias toward transposition events into genic regions of the maize genome, as the repeat content of the \textit{fDs} sequences is far lower than random sequence and comparable to GSS sequences. This pattern of transposition is very useful for reverse genetic studies, making the lines produced by the \textit{Ac/Ds} project a valuable tool for the maize community.
There are still some factors that keep the pipeline from reaching optimal conditions. Because of the crossing methodology used to create the tr-\textit{Ds} lines, there are some events that will never be isolated, such as transposition events to linked regions of the \textit{rl-sc:m3} locus on the homologous chromosome. Sample concentrations can vary significantly between samples, ranging from as high as 300 ng/100 bp/\(\mu\)L to as low as 0.16 ng/100 bp/\(\mu\)L. Sub-optimal concentrations are especially prevalent in larger fragments (>2.0 kb). At minimum, it appears that a concentration of at least 1.5 ng/100 bp/\(\mu\)L is necessary for reliable sequencing, though 2.5 ng/100 bp/\(\mu\)L is ideal. Low concentrations also present problems for further sequencing attempts, due to the low volumes of sample available. In these cases, reamplification is a necessity for any subsequent sequencing attempts. However, many of these samples cannot be reamplified properly. The reason for this is not yet known, but it may be that some contaminant is present in the purified samples that may interfere with proper PCR in samples with lower concentrations. This is especially prevalent in samples that have been purified by gel extraction, as the concentrations produced by this method are significantly lower than those obtained using ethanol precipitation. Efforts are currently underway in potentially removing the necessity of gel extraction for some samples, alleviating this issue. However, at this time it has not yet been resolved. Another related problem is contamination by other samples. This problem seems to arise due to either multiple amplified IPCR products of the same or similar size, or because of minute amounts of a smaller fragment in the sample contaminating the gel slice of a larger fragment. Reamplification of larger fragments, particularly those over 3.0 kb in size, is especially problematic due to this, in which case the parental band, though maybe insignificant in
the original sample, amplifies much easier than the band of interest. Another issue is the presence of redundant transposon hits. Though several of those found appear to be cases of incomplete digestion of the genomic DNA, the presence of identical transposition events across multiple families is a limiting factor. However, in comparison to the 941 samples sequenced, the effect of these repeating insertions appears to be fairly small.

There are also some limitations with aligning our sequences to known BACs. Of the 941 fragments sequenced, 427 cannot be located. This could be due to several factors. Fragments yielding very short finished sequences may not align to any BACs with enough identity. Some sequences do align to BACs, but the coverage or identity rates fall below our requirements.

Even with the current limitations, however, with an overall success rate of around 77.0% for an initial pass through the pipeline, many new tr-\textit{Ds} loci can be quickly and easily sequenced and analyzed. The high-throughput capacity of this pipeline also provides some flexibility in the number of sequences that can be processed at a time, and as coming seasons increase the number of new insertions to analyze, we can expect the current pipeline to easily accommodate the influx of larger data sets in the future.

References

Alleman, M., Kermicle, J.L. 1993 Somatic variegation and germinal mutability reflect the position of transposable element \textit{Dissociation} within the maize \textit{R} gene. Genetics 135: 189-203


Cooper, L.D., Marquez-Cedillo, L., Singh, J., Sturbaum, A.K., Zhang, S., Carollo, V.,


Dooner, H.K., 1979. Identification of an R-locus region that controls the tissue specificity of anthocyanin formation in maize. Genetics 93. 703-710.


Distribution of unlinked receptor sites for transposed \textit{Ac} elements from the \textit{bz-m2(Ac)} allele in maize. Genetics 136: 261-279.


Figure 2.1. Genetic methodology for generation of stable Ds insertions. Z. mays plants homozygous for Ac::im and r1-sc:m3-Ds (Stage 0 ears) are crossed by tester plants, which are homozygous for r1-sc:m3. The result is Stage 1 ears. If mobilization of the Ds in r1-sc:m3 by Ac::im occurs before or shortly after fertilization, then r1-sc:m3 is activated (reverts) and all cells will produce anthocyanin, and the result is a purple kernel. If Ds is mobilized after the fertilized endosperm begins to divide, some cells will carry active r1-sc:m3, while others will not. This results in spotted kernels. Both these types of kernels may carry new tr-Ds insertions, but it can only be passed on reliably by the purple kernels, where all the cells of the embryo would carry the same insertion event. The purple Stage 1 kernels are germinated and crossed by tester lines again, creating Stage 2 ears. In yellow Stage 2 kernels (II-3), the Ac::im has not been inherited, thus r1-sc:m3 cannot be activated, and no anthocyanin is produced. However, these kernels may still carry tr-Ds insertions from their Stage 1 parent. Because there is no anthocyanin production in these kernels, any tr-Ds events into chromosome 10 will most likely be genetically unlinked to the donor site. Spotted kernels (II-4) still carry the Ac::im gene, and are thus any insertions would be unstable. Purple kernels may or may not carry Ac::im. To determine which still carry Ac::im, purple kernels are germinated and crossed again by tester lines (not shown). Plants producing ears containing purple, spotted, and yellow kernels still carry Ac::im (II-1). Plants producing ears containing only purple and yellow kernels no longer carry Ac::im (II-2). These crosses are performed both at the Boyce Thompson Institute and Iowa State University.
Figure 2.2. Primer and restriction enzyme site map of Ds.
Several features are utilized in the identification and analysis of Ds insertions in maize. The blue box represents the Ds6 element from r1-sc:m3. The blue and yellow flanking regions represent the fDs regions isolated by IPCR using SacII and PvuII, respectively. Red lines are the SacII and PvuII digestions sites. Blue lines represent primer binding sites, and the arrows their orientation.
Figure 2.3. Reamplification of IPCR product using fragment-specific primers. The left gel image shows the reamplified product of a PvuII IPCR reaction. The right gel shows the reamplified product of a SacII IPCR reaction. The bands outlined in yellow represent parental bands created by the r1-sc:m3-Ds. Bands annotated with a star correspond to Southern Blot analysis using a Ds specific probe.
Figure 2.4. Structure of a PvuII IPCR fragment through various stages of processing. The earliest IPCR fragment generated by Lc24 and Lc18 carries a significant length of Ds sequence on the TSD end and very little on the reverse. Reamplification using JSR01 and Lc18 removes much of the Ds sequence. Sequencing reactions are performed using JSR03 and Lc18. Finally, the remaining Ds sequence is removed during processing using pregap4 and gap4 software, leaving only the fDs sequence.
Figure 2.5. Structure of a SacII IPCR product through various stages of processing. The initial IPCR fragment produced using Lc18 and Lc24 carries significant portions of \( Ds \) sequence on either side. Reamplification using JGp2 and Lc45 significantly reduces both regions of \( Ds \) sequence. Sequencing reactions are performed using JSR05 and JSR04. The remaining \( Ds \) sequence is removed through processing using pregap4 and gap4 software.
Figure 2.6. Current version of the Ac/Ds project sequencing pipeline. This diagram represents the various stages of the Ac/Ds project sequencing pipeline at present. Steps highlighted in brown are the generation and isolation of IPCR product from genomic DNA and were performed at Tom Brutnell’s lab at Cornell University. Steps highlighted in dark blue are the various stages of sequencing. Steps highlighted in light blue are the processing of sequence data. Both blue regions were performed at Erik Vollbrecht’s lab at Iowa State University.
Figure 2.7. Early version of the Ac/Ds project sequencing pipeline. This diagram represents the various stages of the Ac/Ds project sequencing pipeline in early 2006. Steps highlighted in brown are the generation and isolation of IPCR product from genomic DNA and were performed at Tom Brutnell’s lab at Cornell University. Steps highlighted in dark blue are the various stages of sequencing. Steps highlighted in light blue are the processing of sequence data. Both blue regions were performed at Erik Vollbrecht’s lab at Iowa State University.
Table 2.1. Current Pipeline Results

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Received</th>
<th>Attempted</th>
<th>Sequenced</th>
<th>% Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Unlinked</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.S05.</td>
<td>239</td>
<td>239</td>
<td>182</td>
<td>76.2%</td>
</tr>
<tr>
<td>B.W06.</td>
<td>358</td>
<td>358</td>
<td>281</td>
<td>78.5%</td>
</tr>
<tr>
<td>I.W06.</td>
<td>192</td>
<td>192</td>
<td>136</td>
<td>70.8%</td>
</tr>
<tr>
<td>B.S06.</td>
<td>88</td>
<td>88</td>
<td>67</td>
<td>76.1%</td>
</tr>
<tr>
<td>I.S06.</td>
<td>204</td>
<td>204</td>
<td>150</td>
<td>73.5%</td>
</tr>
<tr>
<td>R-Linked</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.S05.</td>
<td>79</td>
<td>79</td>
<td>72</td>
<td>91.1%</td>
</tr>
<tr>
<td>I.W06.</td>
<td>62</td>
<td>62</td>
<td>53</td>
<td>85.5%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1222</td>
<td>1222</td>
<td>941</td>
<td>77.0%</td>
</tr>
</tbody>
</table>
Table 2.2. BLAST Analysis and Distribution of fDs sequences

<table>
<thead>
<tr>
<th></th>
<th>R-Unlinked</th>
<th>R-Linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sequences</td>
<td>816</td>
<td>126</td>
</tr>
<tr>
<td>Misses the genome</td>
<td>371</td>
<td>56</td>
</tr>
<tr>
<td>Hits a single BAC</td>
<td>351</td>
<td>54</td>
</tr>
<tr>
<td>Hits multiple BACs</td>
<td>94</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>%Genome(Mb)*</th>
<th>Size (cM)</th>
<th>%Genome(cM)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch. 1</td>
<td>290</td>
<td>14.7 (15.8)</td>
<td>262.0</td>
<td>14.8 (16.0)</td>
</tr>
<tr>
<td>Ch. 2</td>
<td>234</td>
<td>11.9 (12.8)</td>
<td>207.6</td>
<td>11.7 (12.7)</td>
</tr>
<tr>
<td>Ch. 3</td>
<td>223</td>
<td>11.3 (12.2)</td>
<td>168.7</td>
<td>9.5 (10.3)</td>
</tr>
<tr>
<td>Ch. 4</td>
<td>254</td>
<td>12.9 (12.9)</td>
<td>175.8</td>
<td>9.9 (10.7)</td>
</tr>
<tr>
<td>Ch. 5</td>
<td>232</td>
<td>11.7 (11.7)</td>
<td>174.8</td>
<td>9.9 (10.7)</td>
</tr>
<tr>
<td>Ch. 6</td>
<td>162</td>
<td>8.2 (8.8)</td>
<td>168.6</td>
<td>9.5 (10.3)</td>
</tr>
<tr>
<td>Ch. 7</td>
<td>152</td>
<td>7.7 (8.3)</td>
<td>147.5</td>
<td>8.3 (9.0)</td>
</tr>
<tr>
<td>Ch. 8</td>
<td>166</td>
<td>8.4 (9.1)</td>
<td>183.3</td>
<td>10.3 (11.2)</td>
</tr>
<tr>
<td>Ch. 9</td>
<td>123</td>
<td>6.2 (6.7)</td>
<td>150.4</td>
<td>8.5 (9.2)</td>
</tr>
<tr>
<td>Ch. 10 (D&lt;sub&gt;s&lt;/sub&gt;)</td>
<td>139</td>
<td>7.0 (NA)</td>
<td>135.0</td>
<td>7.6 (NA)</td>
</tr>
<tr>
<td>Total</td>
<td>1979</td>
<td>100.0</td>
<td>1773.7</td>
<td>100.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Unlinked Hits</th>
<th>Linked Hits</th>
<th>Total Hits</th>
<th>%/Chrom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch. 1</td>
<td>55</td>
<td>5</td>
<td>60</td>
<td>14.8 (19.5)</td>
</tr>
<tr>
<td>Ch. 2</td>
<td>39</td>
<td>2</td>
<td>41</td>
<td>10.1 (13.3)</td>
</tr>
<tr>
<td>Ch. 3</td>
<td>37</td>
<td>4</td>
<td>41</td>
<td>10.1 (13.3)</td>
</tr>
<tr>
<td>Ch. 4</td>
<td>28</td>
<td>4</td>
<td>32</td>
<td>7.9 (10.4)</td>
</tr>
<tr>
<td>Ch. 5</td>
<td>40</td>
<td>1</td>
<td>41</td>
<td>10.1 (13.3)</td>
</tr>
<tr>
<td>Ch. 6</td>
<td>28</td>
<td>2</td>
<td>30</td>
<td>7.4 (9.7)</td>
</tr>
<tr>
<td>Ch. 7</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>4.4 (5.8)</td>
</tr>
<tr>
<td>Ch. 8</td>
<td>21</td>
<td>2</td>
<td>23</td>
<td>5.7 (7.5)</td>
</tr>
<tr>
<td>Ch. 9</td>
<td>21</td>
<td>1</td>
<td>22</td>
<td>5.4 (7.1)</td>
</tr>
<tr>
<td>Ch. 10 (D&lt;sub&gt;s&lt;/sub&gt;)</td>
<td>64</td>
<td>33</td>
<td>97</td>
<td>23.9 (NA)</td>
</tr>
<tr>
<td>Total</td>
<td>351</td>
<td>54</td>
<td>405</td>
<td>100.0</td>
</tr>
</tbody>
</table>

*Values in parenthesis are percentages of the genome without chromosome 10.
### Table 2.3. Repeat content of fDs sequences and other databases.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>fDs</th>
<th>Messing Sequence</th>
<th>TIGR GSS contigs</th>
<th>TIGR methyl-filtered</th>
<th>TIGR high cżt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences</td>
<td>983</td>
<td>474604</td>
<td>1200</td>
<td>30000</td>
<td>30000</td>
</tr>
<tr>
<td>Base pairs, kb</td>
<td>347</td>
<td>307169</td>
<td>2989</td>
<td>21639</td>
<td>21637</td>
</tr>
<tr>
<td>Bases repeat masked, %</td>
<td>8.0</td>
<td>57.9</td>
<td>17.3</td>
<td>18.4</td>
<td>8.6</td>
</tr>
</tbody>
</table>

### Table 2.4. Redundant sequence frequency

<table>
<thead>
<tr>
<th>Redundancy per Contig</th>
<th>Sequences</th>
<th>Unique Contigs</th>
<th>Contig Frequency</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>827</td>
<td>827</td>
<td>94.95%</td>
<td>Unique</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>7</td>
<td>0.80%</td>
<td>Unique*</td>
</tr>
<tr>
<td>2</td>
<td>45**</td>
<td>23</td>
<td>2.64%</td>
<td>Hotspot/Cryptic</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>8</td>
<td>0.92%</td>
<td>Hotspot/Cryptic</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>5</td>
<td>0.57%</td>
<td>Hotspot/Cryptic</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
<td>0.11%</td>
<td>Hotspot/Cryptic</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
<td>0.11%</td>
<td>Hotspot/Cryptic</td>
</tr>
<tr>
<td>Total</td>
<td>941</td>
<td>872</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*in some cases, the restriction enzyme does not fully digest the genomic DNA. In these cases, two fDs bands appear in IPCR but represent the same tr-Ds event.

**there is a case of three redundant sequences in a contig: B.W06.0916A, B.W06.0916B, and B.W06.0908. B.W06.0916A and B.W06.0916B fall under the two-sequence unique category. B.W06.0908 qualifies under the two-sequence hotspot/cryptic category.
CHAPTER 3: DISTRIBUTION OF INTRACHROMOSOMAL DISASSOCIATION TRANSPOSITION EVENTS ACROSS MAIZE CHROMOSOME 10

Abstract

It is known that Ac/Ds transposons have a preference toward transposing to closely linked sites in the host DNA. Here, using sequence produced by the Ac/Ds pipeline, we look at patterns of intrachromosomal transpositions of Ds6, a Dissociation element located in the anthocyanin-regulating gene r1-sc:m3. Of the 119 confirmed and mapped tr-Ds on chromosome 10, 57 land proximal to the donor site and 62 land distal. The transpositions to other chromosomes show a fairly broad distribution, while the R-linked samples show a tighter distribution around the donor site. R-linked samples also show a directional bias toward transposing to one side of the donor site. Though Ds has the capability of transposing some distance away, many events are clustered immediately distal to the donor site. This implies a strong transposition bias toward these closely localized insertion events.

Introduction

The effects of transposons on genome evolution have been very significant. Transposons provide significant mutagenic potential, having the ability to insert into regions of the host genome that may alter the expression of coding regions of DNA. Also, due to their nature as mobile DNA, new genes can be carried around the host genome for potential new function. Fragmented or otherwise disabled transposons can have a significant effect on host size, in some cases becoming the single most significant factor in host genome size. However, due to their nature as parasitic DNA elements, transposons may also place a burden upon their host. When a transposon is first
introduced into a species, rapid replication is necessary for its survival, lest random mutation within the host genome render it nonfunctional. As such, it’s necessary for a transposon to replicate rapidly in order to remain functional in the host genome, but this can also cause problems for both the host and the transposon if too much transposon activity occurs. If rampant transposition were to occur in the host genome, insertions could occur in regions of DNA responsible for vital host functions. In the process of increasing their own number within the genome, the damage inflicted by their proliferation may destroy their host, and the transposon along with it (Rouzic, Capy, 2004). In order to prevent their mutual destruction, many forms of regulation have been developed to prevent excessive transposition. Inducible promoters of transcription for transposition genes, silencing due to methylation of repetitive DNA, and controls on translation and modification of proteins involved in transposition have been shown to limit the activity of transposons (Kunze, Starlinger, 1989; Wang et al., 1996; Wang, Kunze, 1998). Patterns of transposition may also be significant in the regulation of transposon activity. If transposons have a bias in their movement around the host genome, this implies a degree of specialization and control involved in the transposition mechanism. This is perhaps a function the transposon to allow it some preference to sites less likely to cause serious damage to the host genome, modulating passage of the transposon from generation to generation.

It is known that Ac/Ds elements are biased toward transposing to closely linked sites in the genome. Ac insertions show a clustering pattern closely situated around the donor site. This pattern has been shown in a variety of backgrounds, including maize, tobacco, and Arabidopsis (Van Schaik, Brink, 1959; Greenblatt, Brink, 1962; Greenblatt,
1984; Dooner, Belachew, 1989; Jones et al., 1990; Dooner, et al., 1991; Osbourn, et al., 1991; Belzille and Yoder, 1992; Keller, et al., 1993). However, many previous studies, and all of the studies in maize, involve active Ac elements, so it is difficult to determine if these new insertions are transposition events from the donor site or secondary transposition from these insertions. Also, many of these studies calculate the distances between events in centimorgans. Since this measurement is based on recombination frequency, the physical distance may vary depending on the region of the chromosome. Finally, it is not fully known how the structure of the element may affect the patterns of transposition. Ac elements are fairly uniform, but Ds elements may behave differently depending on their structure.

We have analyzed sequence data from the fDs regions isolated by the Ac/Ds pipeline, specifically fDs that are located on the same chromosome as the Ds donor locus. As of this thesis, it is estimated that around two thirds of the maize genome has been sequenced. These sequences are in the form of bacterial artificial chromosomes (BACs). By BLAST analysis, we can align our fDs sequence to these BACs, which can then be located on the assembled FPC contigs, allowing an accurate picture of where these new transposition events are located within maize chromosome 10. Using these data, we can determine where within the donor chromosome these Ds elements are inserting. We will compare this to the location of r1-sc:m3, the donor locus on chromosome 10.

Methods and Materials

Selection for Linked Transposition Events

As part of the Ac/Ds pipeline, samples are selected for having stable transposition events, or samples that have lost the transposase source Ac::im during crossing (refer to
methods and materials in chapter 2). For the R-linked families, though, it’s not possible to determine which kernels contain stable or unstable just from the phenotype of the kernels. To determine which samples contain Ac::im and which don’t, purple kernels are grown and crossed with pollen from plants homozygous for the Ds insertion in r1-sc:m3 and lacking Ac::im. If Ac::im is still present in the original kernel, the ear will have purple, spotted, and yellow kernels. If Ac::im was removed via crossing, the ear will have only purple and yellow. The purple kernels from these ears that lack Ac::im are harvested and selected for processing through the Ac/Ds pipeline as potentially containing stable, R-linked insertions.

**Ac/Ds Sequencing Pipeline**

R-linked samples are passed through the Ac/Ds sequencing pipeline in the same manner as R-unlinked samples. Seed from an R-linked family is grown up in sand bench until it reaches about four inches in height, at which point plant tissue is collected and the genomic DNA extracted. The genomic DNA is digested by either PvuII or SacII. Southern blots are performed using a Ds-specific probe to determine the presence of a new transposition event. Inverse PCR is performed on those samples using the primers LC24 and LC18. The IPCR product is run out on an agarose gel, and the bands corresponding to the Southern blots are excised and extracted. Another round of PCR is performed using primers specific to the type of fragment (JSR01 and LC18 for PvuII-digested samples, and JGp2 and LC45 for SacII-digested samples. These samples are ethanol precipitated and freeze-dried for shipping.

After being resuspended in high performance liquid chromatography (HPLC) grade water, the concentrations of the samples are assessed using an ND-1000
spectrophotometer (Nanodrop), and the samples are diluted if necessary to reach an optimum sequencing reaction concentration of around 2.5 ng/100 bp/uL. Sequencing reactions are performed by the DNA sequencing facility and Iowa State University. The sequence obtained is processed using pregap4 assembly software and gap4 editing software to assemble the sequencing reads, remove extraneous $Ds$ sequence, and allow manual editing of samples as necessary.

**BLAST analysis**

The $fDs$ fragments that are sequenced by the pipeline are run through Smit, Hubley, and Green’s RepeatMasker program, version open-3.1.8 (www.repeatmasker.org), using the TIGR maize repeat database version 4.0 (http://maize.tigr.org). These repeatmasked sequences are then queried against maize BACs via NCBI’s megaBLAST. Those sequences that align to BACs annotated to chromosome 10 ($Ds$ donor chromosome) are sorted and selected for distribution analysis. Hits were classified as having at least 95% coverage of the query sequence and at least 95% identity against the reference sequence. The latest BLAST analysis was performed October 5th, 2007.

**Distribution Analysis**

The chromosome 10 BACs that are uniquely hit by $fDs$ sequences are mapped out on the FPC map using Ensembl browser at the MaizeSequence.org website (http://www.maizesequence.org/index.html). This produces a physical map distribution of tr-$Ds$ events. To determine genetic distributions, we cross-referenced the physical map locations of markers on the Genetic 2005 genetic map. The markers were also compared against the FPC0507 genetic map. Genetic map locations of markers from the FPC0507
map were estimated with respect to the Genetic 2005 map and added to the overall map, so as to increase resolution. We also looked at a subset of samples that have been examined for both selection for and against linked transposition events, detecting all events that may have occurred. These samples are used to determine biases to unlinked regions of the donor chromosome.

Results

The overall estimated size of maize chromosome 10 is 139 Mb. Of the 941 tr-\textit{Ds} insertion site sequences currently produced by the \textit{Ac}/\textit{Ds} sequencing pipeline, a total of 108 sequences show BLAST matches to the \textit{Ds} donor chromosome 10 (Table 3.1 and 3.2). Of these, eight were confirmed to be on BACs annotated as located on chromosome 10, but the BACs themselves have not been properly located to a specific FPC contig. Of the remaining 100, 62 are from families selecting for unlinked tr-\textit{Ds} events (\textit{R}-UL) and 38 are from families without this selection (\textit{R}-L). In the tables, locations are listed by the name of the BAC, location on that BAC (based on BLAST results), locations of the BAC (base pair coordinate estimated based on maizesequence.org map data), and which FPC contig the BAC aligns to. When multiple tr-\textit{Ds} hits trace back to a single BAC they are grouped together. Because the \textit{R}-UL selections were processed through the pipeline prior to the \textit{R}-L selections, it’s possible that some of the fDs from \textit{R}-UL families may be present in the \textit{R}-L families. In other words, there may be tr-\textit{Ds} events on the map listed as from \textit{R}-UL families that may also qualify as being from \textit{R}-L families. The locus of \textit{R-sc} was also mapped to FPC contig 415 as the point of \textit{Ds} distribution. Hits to and around the donor locus are aligned to the maize 10L BAC sequences from Swigonova \textit{et al.}, 2005. RepeatMasker results showed a similar level of repeat data to that found in all fDs
sequences (discussed in Chapter 2), at around 7.7% repeat content, so the tr-Ds events on chromosome 10 show the same bias toward genic regions of the host genome.

**Mapping of Ds transpositions to the physical map**

When mapped across the entirety of the sequenced portion of chromosome 10, the mapped Ds transpositions form an interesting pattern (Fig. 3.1). R-sc, the location of the donor Ds, is located near the upstream end of FPC contig 415. Of the 100 mapped Ds hits, 48 hit FPC contigs proximal to R-sc. 33 of these are from R-UL families and 15 are from R-L families. Of the remaining 52 distal mapped Ds hits, 29 are from R-UL families and 23 are from R-L families.

Of the R-UL families, the furthest proximal event is B.S05.0027 in c0462J05 on FPC ctg392 (~118 Mb distance). The furthest distal events are I.W06.0046 and B.S06.0641 in c0031L15 on FPC ctg419 (~12 Mb distance). R-UL families show a broad distribution around the chromosome, with higher concentrations of hits around the donor locus than distant from it.

Of the R-L families, the furthest proximal event is I.W06.0993R in c0105D19 on FPC ctg401 (~65 Mb distance). The furthest distal event is B.S05.0764R in c0334G19 on FPC ctg419 (~12 Mb distance). R-L hits show a tighter distribution around the donor locus, with a higher concentration of hits immediately distal of R-sc than proximal.

In both R-L and R-UL families, there can be very tight clustering of hits in certain areas. In the sequenced contig that contains R, R-UL hits B.S06.0021 and B.W06.0222 are 7 bp away from the donor locus. In the c0105B22 BAC, R-L events I.W06.0974R and B.S05.0808R are separated by 19 bp. This close clustering within BACs is most prevalent close to R-sc.
Mapping of $Ds$ transpositions to the genetic map

To determine genetic distances, the hit locations from the physical map were compared against loci found on the Genetic 2005 genetic map of chromosome 10 (Fig. 3.2). To increase resolution of the map in early sections, the map loci on the Genetic 2005 map were overlapped with the loci on the FPC0507 recombinant inbred map. The genetic coordinates of the FPC0507 map were estimated based on surrounding loci on the Genetic 2005 map. The tr-$Ds$ locations were also inferred by comparing the physical map location with the physical map locations of loci from the genetic maps.

Of the 100 tr-$Ds$ events, 95 are located within 40 cM of the donor site. This is approximately 95% of all hits to chromosome 10. This shows a strong bias of transposition events to locations near the donor site ($\chi^2=37.87$, p-value=<0.0001). Of the subset of samples that have been analyzed for both $R$-UL and $R$-L events, all 34 of the events mapping to chromosome 10 are within 40 cM. Of the $R$-UL families, the furthest proximal event is B.S05.0027 in c0462J05 on FPC ctg392 (~90 cM distance). The furthest distal events are I.W06.0046 and B.S06.0641 in c0031L15 on FPC ctg419 (~33 cM distance). By design, tr-$Ds$ events from $R$-UL families should be placed further away from the donor site than those from $R$-L families. The tr-$Ds$ events from $R$-L families do not come under this selective pressure. Within the distance of 40 cM considered linked to the $R$-$sc$ locus, there are 57 $R$-UL hits (91.9% of mapped hits) and 38 $R$-L hits (100.0% of mapped hits). Within 30 cM, there are 47 $R$-UL (75.8%) and 32 $R$-L (84.2%) hits. Within 20 cM, there are 37 $R$-UL (59.7%) and 28 $R$-L (73.7%). Within 10 cM, there are 33 $R$-UL (53.2%) and 27 $R$-L (71.1%). Thus, fDs from $R$-L families show a higher linkage to the $R$-$sc$ locus than those from $R$-UL families. In the $R$-L families, the number of tr-$Ds$
events drops sharply outside of 10cM, whereas the $R$-UL families have a steady distribution beyond 10cM. Because the $R$-UL families show such a broad distribution pattern relative to $R$-L families, this supports the effectiveness of our selection against linked transposition events.

There also appear to be some biases in the polarity of local $Ds$ transpositions. Within a 35 cM range, there are 11 proximal of $R$-$sc$ (31%) and 24 (69%) distal $R$-L events ($\chi^2=4.83$, p-value=0.0280). This disparity is more pronounced as the range is decreased. Within 10 cM, there are 5 proximal (20%) and 20 distal (80%) events ($\chi^2=9.00$, p-value=0.0027). This data set is still small, but it appears that in the case of linked transposition events, $Ds$ elements show a bias toward transposing distal to the original donor site.

Another note is the disparity between the genetic and physical map. On the genetic map, the most distal mapped locus on the long arm of chromosome 10 is 134.4 cM. On the physical map, this locus is located on FPC contig 419, and no mapped loci could be placed in contigs 420 or 421. Also, no tr-$Ds$ events could be located beyond the beginning of contig 420, despite a consistent series of events up to this point. Since the tr-$Ds$ events have a bias toward placement in genic regions of the genome, the area distal to 134.4 cM may be gene-sparse, making it difficult or impossible to create a genetic map covering this area.

**Mapping to a well-annotated maize 10L BAC**

To determine transposition patterns near the donor site, we’re used a BAC contig on maize chromosome 10 from Swigonova, *et. al.* 2005, which contains the $R$-$sc$ gene (Fig. 3.3). 23 events land in this BAC. Of the 16 $R$-unlinked events, 13 are back into
various points of the \( R-sc \) gene. One of the \( R-UL \) events lands in the maize \( Isr \) gene. The remaining two land in intergenic, low copy regions. Of the 7 \( R-L \) events, three land in the maize \( Isr \) gene. Another two hit the putative glycerol-3-transporter gene. The remaining two land in intergenic, low copy space.

We also compared the orientation of tr-\( Ds \) events located around the \( R-sc \) locus. Of the 22 sequences aligned to the contig, eight are oriented 5’ to 3’. The remaining 14 are oriented 3’ to 5’ (\( \chi^2 = 1.64, p\text{-value}=0.2003 \)). Within the \( R-UL \) sequences, seven are oriented 5’ to 3’, and eight are 3’ to 5’ (\( \chi^2 = 1.0, p\text{-value}=0.3173 \)). Within the \( R-L \) sequences, one is oriented 5’ to 3’ and six are 3’ to 5’ (\( \chi^2 = 3.58, p\text{-value}=0.0585 \)). The data set is small, but there does not appear to be any bias in orientation of tr-\( Ds \) events near the donor site.

**Discussion**

Overall, \( Ds \) shows a very interesting pattern of transposition along chromosome 10. Based on the 100 tr-\( Ds \) events mapped to chromosome 10, we can see that \( Ds \) transposes preferentially to locations near the donor site. In both \( R-UL \) and \( R-L \) families, over 90% of hits to chromosome 10 are within 40 cM of the donor locus. In a previous study examining unlinked transposition patterns of \( Ac \) from a donor site on the short arm of chromosome 9 (Dooner, *et. al.*, 1994) a higher proportion of events were found on long arm of the donor chromosome, implying a preference based on physical, if not genetic, linkage. Our results produced a similar result. Among a subset of 67 samples that have been thoroughly examined for any transposition events, 34 events mapped to chromosome 10 and are within 40 cM of \( R-sc \), showing a strong bias toward this area (\( \chi^2 = 720.30, p\text{-value}<0.0001 \)). The remaining 33 are distributed across the other
chromosomes, but none are located in the 60 cM of chromosome 10 outside of linkage to the donor site. The lack of events in this region could be expected based on random distribution ($\chi^2=1.25$, p-value=0.2636). The tr-$Ds$ events become more clustered closer to the donor locus. This clustering can be very tight, and in some cases multiple hits may land within a range as small as 10 bp or less. This supports previous studies showing transposition of $Ac$ to closely linked sites in maize (Greenblatt, Brink, 1962; Greenblatt, 1982; Dooner, Belachew, 1989). However, the distribution around the donor site in our study does not appear to be the same as those in the previous works. In the Dooner and Belachew study, 57% of the events mapped are within 5 cM of the donor site $bz1-m2$. 8% are located between 5 and 10 cM, and 22% are located between 10 and 20 cM. Of our subset of samples that have been analyzed for both $R$-UL and $R$-L events, 77% were mapped within 5 cM of the donor site $r1-sc:m3$. 0% are located between 5 and 10 cM, and 3% are located between 10 and 20 cM. It appears that both sample sets have a bias toward events closely linked to the donor site, but the events are more concentrated near the donor site in our tr-$Ds$ subset than the Dooner and Belachew distribution. The difference between the two sample sets would not be expected based on random variation ($\chi^2=31.42$, p-value=<0.0001). The reason for this difference in distribution is not known, but may be related to differences in the genic content and structure of the regions surrounding the donor site (in chapter 2, we discussed the preference of tr-$Ds$ events to low-repeat regions of the genome). Also, the previous studies involved an autonomous $Ac$ element, whereas our study uses a Ds6 element with a transposase source. It’s not known if the structural differences between the elements would have any effect on transposition patterns. It’s worth noting that both sample sets are fairly small, and the
distribution pattern may be further refined as more events are analyzed. The pattern of
distribution can vary depending on how the events are selected for. The $R$-UL families
tend to show a more broad distribution around the donor locus, and do not appear to show
any noticeable bias in direction. The $R$-L families show a much tighter distribution
around the donor locus, and appear to transpose preferentially to locations distal of the
donor locus. In a previous study involving an $Ac$ transposon at the $P$ locus (referred to as
$Modulator$, or $Mp$, at the time), a similar directional bias was found (Greenblatt, 1984).
To explain this bias, Greenblatt (1984) hypothesized that a replicon initiation site is
located proximal to the donor gene. It was further proposed that $Mp$ transposes when the
host genome is replicated, and transposes preferentially into unreplicated sites. This
would produce a directional bias distal to the donor site for closely linked transpositions.
However, another study involving $Ac$ elements found no bias in either direction (Dooner,
Belachew, 1989). The presence or size of a replicon at $r$ is currently unknown, but
eukaryotic replicons can vary in size from 50 to 100 kb. Within a 100 kb range around
the donor site, we mapped 19 tr-$Ds$ events. Of these, 6 are located proximal to the donor
site, and 13 are located distal to the donor site. These results are consistent with random
distribution around the donor site ($\chi^2=2.58$, p-value=0.1083). This would indicate that
either the $Ds6$ transposon in $rl-sc:m3$ does not show the directional bias found in the $Mp$
study, or that there is no replicon initiation site near the $R-sc$ locus. Both families show
some hits to regions outside of the 40 cM range, though the $R$-UL families had
proportionally more hits outside that range than $R$-L families. Because of the selection
against linked events in the $R$-UL families, this is to be expected.
It’s likely that the close-transposing nature of \( Ac/Ds \) may have had some significant influence on the evolution of the transposon and its host genome. If a transposon were to insert into a region where it has no major phenotypic detriment to its host, it’s less likely that further transposon insertions in that region would cause a significant effect. As such, the transposon could replicate and proliferate in this region without destroying the potential to be passed on from generation to generation. These areas would also become sites for increases in nonfunctional transposons, making these benign regions progressively larger, contributing to host genome size. While this may still cause a burden on the host genome, the isolation of the insertions may limit any damage caused by active transposition, representing another method of control to prevent excessive transposition. This mechanism may also explain why there is little evidence for horizontal transfer of \( hAT \) elements across species. If the transposon is strongly biased toward localized movement, events of transposition across chromosomes is less likely, let alone across species boundaries.

Because these results show similar trends to those found in previous studies involving mapping \( Ac \) transposition patterns, we can be fairly certain that we’re receiving an accurate picture of transposition. However, because of the fine degree of placement as well as the removal of secondary transposition events, the pattern of insertions produced here shows a clear view of transposition biases in \( Ds6 \) elements and in \( Ac/Ds \) elements in general.

References

Alleman, M., Kermicle, J.L. 1993 Somatic variegation and germinal mutability reflect the position of transposable element \( Dissociation \) within the maize R gene. Genetics 135: 189-203


elements Ds at the *Shrunken* locus in *Zea mays* structure of the unstable *sh-m5933* allele and several revertants. Cell 34. 383-393.


Ensembl release 46: Zea mays overview of features on chromosome 10 1-139336400.
The Maize Genome Browser. 2007.  


Kermicle, J.L., 1980. Probing the component structure of a maize gene with transposable
elements. Science 208. 1457-1459.


Van Schaik, N., Brink, R.A., 1959. Transposition of Modulator, a component of the variegated pericarp in maize. Genetics 44. 725-738

Yan, X., Martinez-Ferez, M., Kavchok, S., Donner, H.K., 1999. Origination of Ds
elements from Ac elements in maize: evidence for rare repair synthesis at the site of Ac excision. Genetics 152. 1733-1740.

Table 3.1. Distribution of $R$-$L$ $Ds$ insertion sites on $Z. mays$ chromosome 10

Total Size of Maize Chromosome 10: approximately 139 Mb

<table>
<thead>
<tr>
<th>Family ID</th>
<th>BAC hit</th>
<th>Location on BAC</th>
<th>BAC location*</th>
<th>FPC hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.W06.0993R</td>
<td>c0105D19</td>
<td>138955</td>
<td>54125400</td>
<td>ctg401</td>
</tr>
<tr>
<td>B.S05.0803R</td>
<td>c0333J13</td>
<td>10763</td>
<td>62543600</td>
<td>ctg403</td>
</tr>
<tr>
<td>I.W06.0815RA</td>
<td>c0205D13</td>
<td>100613</td>
<td>72186800</td>
<td>ctg407</td>
</tr>
<tr>
<td>I.W06.0990R</td>
<td>c0224P08</td>
<td>35225</td>
<td>74911200</td>
<td>ctg408</td>
</tr>
<tr>
<td>B.S05.0839R</td>
<td>c0472K08</td>
<td>99864</td>
<td>85054200</td>
<td>ctg411</td>
</tr>
<tr>
<td>I.W06.0839RA</td>
<td>c0472K08</td>
<td>166594</td>
<td>85054200</td>
<td>ctg411</td>
</tr>
<tr>
<td>B.S05.0997R</td>
<td>c0176D16</td>
<td>38307</td>
<td>98024500</td>
<td>ctg412</td>
</tr>
<tr>
<td>B.S05.0991R</td>
<td>c0144D12</td>
<td>151360</td>
<td>103933900</td>
<td>ctg412</td>
</tr>
<tr>
<td>I.W06.0867R</td>
<td>c0466F24</td>
<td>23116</td>
<td>109368000</td>
<td>ctg413</td>
</tr>
<tr>
<td>I.W06.0801R</td>
<td>c0248C01</td>
<td>69528</td>
<td>111768400</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S05.0827RB</td>
<td>b0427H13</td>
<td>125214</td>
<td>118682900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0841R</td>
<td>c0262H12</td>
<td>23128</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0873R**</td>
<td>c0262H12</td>
<td>23633</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0956R**</td>
<td>c0262H12</td>
<td>23633</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0818R</td>
<td>c0134O22</td>
<td>26310</td>
<td>119074900</td>
<td>ctg415</td>
</tr>
<tr>
<td>$R$-sc donor</td>
<td>maize10L***</td>
<td>75208</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0832R</td>
<td>maize10L</td>
<td>77120</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0819R</td>
<td>maize10L</td>
<td>92166</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0961R</td>
<td>maize10L</td>
<td>92678</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.1030RC</td>
<td>maize10L</td>
<td>93430</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0970R</td>
<td>maize10L</td>
<td>212804</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0818R</td>
<td>maize10L</td>
<td>267931</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0803R</td>
<td>maize10L</td>
<td>269565</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.1044R</td>
<td>c0173G07</td>
<td>128470</td>
<td>119563900</td>
<td>ctg415</td>
</tr>
</tbody>
</table>

* BAC coordinates based on www.maizesequence.org estimates.
** $tr$-$Ds$ insertions mapped to exact same site.
*** Swigonova et al. 2005 BAC.
Table 3.1. Distribution of R-L Ds insertion sites on Z. mays chromosome 10 (con)
Total Size of Maize Chromosome 10: approximately 139 Mb

<table>
<thead>
<tr>
<th>Family ID</th>
<th>BAC hit</th>
<th>Location on BAC</th>
<th>BAC location*</th>
<th>FPC hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.W06.1022R</td>
<td>c0104L01</td>
<td>71684</td>
<td>119824600</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0793R</td>
<td>c0104L01</td>
<td>83008</td>
<td>119824600</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0812R</td>
<td>b0229F11</td>
<td>112230</td>
<td>119947100</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0859R</td>
<td>c0127D18</td>
<td>46183</td>
<td>120064700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0974R</td>
<td>c0105B22</td>
<td>454</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0808R</td>
<td>c0105B22</td>
<td>473</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0766R</td>
<td>c0105B22</td>
<td>29440</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0947R</td>
<td>c0105B22</td>
<td>70910</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0849R</td>
<td>c0105B22</td>
<td>166853</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0952R</td>
<td>c0105B22</td>
<td>166883</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0943R</td>
<td>c0042L21</td>
<td>1244</td>
<td>121309300</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0792R</td>
<td>c0319O18</td>
<td>59323</td>
<td>122269700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0767R</td>
<td>b0450K23</td>
<td>116742</td>
<td>123132100</td>
<td>ctg416</td>
</tr>
<tr>
<td>B.S05.0788R</td>
<td>b0195H24</td>
<td>49141</td>
<td>126464100</td>
<td>ctg417</td>
</tr>
<tr>
<td>B.S05.0764R</td>
<td>c0334G19</td>
<td>156138</td>
<td>130309600</td>
<td>ctg419</td>
</tr>
<tr>
<td>I.W06.0974R</td>
<td>c0532P19</td>
<td>182464</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>B.S05.0808R</td>
<td>c0532P19</td>
<td>182484</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>I.W06.0850R</td>
<td>c0407E07</td>
<td>174887</td>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

* BAC coordinates based on www.maizesequence.org estimates.
** tr-Ds insertions mapped to exact same site.
*** Swigonova et al. 2005 BAC.
Table 3.2. Distribution of R-UL Ds insertion sites on Z. mays chromosome 10

Total Size of Maize Chromosome 10: approximately 139 Mb

<table>
<thead>
<tr>
<th>Family ID</th>
<th>BAC hit</th>
<th>Location on BAC</th>
<th>BAC location*</th>
<th>FPC hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S05.0027</td>
<td>c0462J05</td>
<td>148385</td>
<td>1710100</td>
<td>ctg392</td>
</tr>
<tr>
<td>B.S05.0389</td>
<td>c0182K12</td>
<td>106365</td>
<td>6439400</td>
<td>ctg393</td>
</tr>
<tr>
<td>I.S06.1335A</td>
<td>b0308C12</td>
<td>8177</td>
<td>6904900</td>
<td>ctg393</td>
</tr>
<tr>
<td>I.S06.1234</td>
<td>c0141A12</td>
<td>82329</td>
<td>11261000</td>
<td>ctg394</td>
</tr>
<tr>
<td>I.S06.0152</td>
<td>c0135C12</td>
<td>47654</td>
<td>12784900</td>
<td>ctg395</td>
</tr>
<tr>
<td>I.W06.0993</td>
<td>c0105D19</td>
<td>117490</td>
<td>54125400</td>
<td>ctg401</td>
</tr>
<tr>
<td>B.W06.0094</td>
<td>c0255E24</td>
<td>96563</td>
<td>64279000</td>
<td>ctg403</td>
</tr>
<tr>
<td>I.S06.1736</td>
<td>c0120G06</td>
<td>22090</td>
<td>69345600</td>
<td>ctg406</td>
</tr>
<tr>
<td>B.S05.0412</td>
<td>c0020M05</td>
<td>76499</td>
<td>78778100</td>
<td>ctg409</td>
</tr>
<tr>
<td>B.W06.0767C</td>
<td>c0020M05</td>
<td>77228</td>
<td>78778100</td>
<td>ctg409</td>
</tr>
<tr>
<td>B.S05.0282A</td>
<td>c0009N15</td>
<td>199960</td>
<td>81589900</td>
<td>ctg409</td>
</tr>
<tr>
<td>B.W06.0606</td>
<td>c0022H18</td>
<td>88475</td>
<td>82569900</td>
<td>ctg410</td>
</tr>
<tr>
<td>B.W06.0382</td>
<td>c0207G07</td>
<td>12362</td>
<td>99245200</td>
<td>ctg412</td>
</tr>
<tr>
<td>I.S06.0128</td>
<td>c0124J20</td>
<td>67459</td>
<td>106678700</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S06.0609</td>
<td>b0325O03</td>
<td>39334</td>
<td>106918000</td>
<td>ctg413</td>
</tr>
<tr>
<td>I.W06.0184A</td>
<td>b0325O03</td>
<td>117094</td>
<td>106918000</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S05.0158</td>
<td>c0466F24</td>
<td>73507</td>
<td>109368000</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S06.0451A</td>
<td>c0459L10</td>
<td>160727</td>
<td>110803700</td>
<td>ctg413</td>
</tr>
<tr>
<td>I.W06.0051</td>
<td>c0291N07</td>
<td>66897</td>
<td>113871100</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.W06.0642</td>
<td>c0482A16</td>
<td>98447</td>
<td>114984200</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S05.0065B</td>
<td>c0227D13</td>
<td>5510</td>
<td>115174500</td>
<td>ctg413</td>
</tr>
</tbody>
</table>

*BAC coordinates based on www.maizesequence.org estimates.

**tr-Ds** insertions mapped to exact same site.

*** Swigonova et al. 2005 BAC.
Table 3.2. Distribution of R-UL Ds insertion sites on *Z. mays* chromosome 10 (con)
Total Size of Maize Chromosome 10: approximately 139 Mb

<table>
<thead>
<tr>
<th>Family ID</th>
<th>BAC hit</th>
<th>Location on BAC</th>
<th>BAC location*</th>
<th>FPC hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S06.0405</td>
<td>c0017M07</td>
<td>34582</td>
<td>115277400</td>
<td>ctg413</td>
</tr>
<tr>
<td>I.S06.0185</td>
<td>c0508J23</td>
<td>103621</td>
<td>116051600</td>
<td>ctg413</td>
</tr>
<tr>
<td>B.S06.0344</td>
<td>c0262H12</td>
<td>23633</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0208</td>
<td>c0262H12</td>
<td>25836</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0163B</td>
<td>c0262H12</td>
<td>108942</td>
<td>118805400</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0944</td>
<td>maize10L ***</td>
<td>68064</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0644</td>
<td>maize10L</td>
<td>68720</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S06.0268B</td>
<td>maize10L</td>
<td>68918</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0319</td>
<td>maize10L</td>
<td>69443</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0523</td>
<td>maize10L</td>
<td>70131</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0441A</td>
<td>maize10L</td>
<td>70396</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.S06.0899</td>
<td>maize10L</td>
<td>75176</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>R-sc donor</td>
<td>maize10L</td>
<td>75208</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S06.0021**</td>
<td>maize10L</td>
<td>75215</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0222**</td>
<td>maize10L</td>
<td>75215</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.S06.1009</td>
<td>maize10L</td>
<td>75222</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0520</td>
<td>maize10L</td>
<td>75329</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0026A</td>
<td>maize10L</td>
<td>75671</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0672</td>
<td>maize10L</td>
<td>76457</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.S06.1241</td>
<td>maize10L</td>
<td>92486</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S06.0439</td>
<td>maize10L</td>
<td>202624</td>
<td>119427700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0527</td>
<td>c0173G07</td>
<td>128151</td>
<td>119564900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S05.0799R</td>
<td>c0104L01</td>
<td>22714</td>
<td>119824600</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0180</td>
<td>c0105B22</td>
<td>46414</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0334A</td>
<td>c0105B22</td>
<td>58359</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.S06.0502</td>
<td>c0105B22</td>
<td>166460</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>I.W06.0176</td>
<td>c0105B22</td>
<td>166591</td>
<td>120201900</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0793</td>
<td>c0319O18</td>
<td>56576</td>
<td>122269700</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0243</td>
<td>c0114I15</td>
<td>106633</td>
<td>122852800</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0526A**</td>
<td>b0333D14</td>
<td>82478</td>
<td>122965500</td>
<td>ctg415</td>
</tr>
<tr>
<td>B.W06.0708**</td>
<td>b0333D14</td>
<td>82478</td>
<td>122965500</td>
<td>ctg415</td>
</tr>
</tbody>
</table>

*BAC coordinates based on www.maizesequence.org estimates.

**tr-Ds insertions mapped to exact same site.

*** Swigonova *et al.* 2005 BAC.
Table 3.2. Distribution of \( R \)-UL \( Ds \) insertion sites on \( Z. \) mays chromosome 10 (con)
Total Size of Maize Chromosome 10: approximately 139 Mb

<table>
<thead>
<tr>
<th>Family ID</th>
<th>BAC hit</th>
<th>Location on BAC</th>
<th>BAC location*</th>
<th>FPC hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S06.0845</td>
<td>c0056N15</td>
<td>119140</td>
<td>123980600</td>
<td>ctg416</td>
</tr>
<tr>
<td>I.S06.0038</td>
<td>b0569G13</td>
<td>31189</td>
<td>124250100</td>
<td>ctg416</td>
</tr>
<tr>
<td>B.S06.0216</td>
<td>c0321C09</td>
<td>20841</td>
<td>125116600</td>
<td>ctg417</td>
</tr>
<tr>
<td>B.W06.0391</td>
<td>c0321C09</td>
<td>50953</td>
<td>125116600</td>
<td>ctg417</td>
</tr>
<tr>
<td>I.W06.0047</td>
<td>c0009E22</td>
<td>72691</td>
<td>125983900</td>
<td>ctg417</td>
</tr>
<tr>
<td>I.S06.0386</td>
<td>b0472J14</td>
<td>19829</td>
<td>127170500</td>
<td>ctg417</td>
</tr>
<tr>
<td>I.S06.0823</td>
<td>b0612D07</td>
<td>155514</td>
<td>127748700</td>
<td>ctg418</td>
</tr>
<tr>
<td>I.S06.1672</td>
<td>b0333A22</td>
<td>48735</td>
<td>129419600</td>
<td>ctg419</td>
</tr>
<tr>
<td>B.W06.1023</td>
<td>c0214A11</td>
<td>187031</td>
<td>131217100</td>
<td>ctg419</td>
</tr>
<tr>
<td>I.W06.0046</td>
<td>c0031L15</td>
<td>70095</td>
<td>131560100</td>
<td>ctg419</td>
</tr>
<tr>
<td>B.S06.0641</td>
<td>c0031L15</td>
<td>144831</td>
<td>131560100</td>
<td>ctg419</td>
</tr>
<tr>
<td>B.W06.0060</td>
<td>c0423H22</td>
<td>51595</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>B.W06.0660</td>
<td>c0423H22</td>
<td>128846</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>B.W06.1052B</td>
<td>c0516D12</td>
<td>22466</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>I.W06.0070</td>
<td>c0445J17</td>
<td>27862</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>I.S06.1053</td>
<td>c0247L19</td>
<td>6392</td>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

*BAC coordinates based on www.maizesequence.org estimates.
**\( tr-Ds \) insertions mapped to exact same site.
***\( R-sc \) donor site location based on Swigonova et al. 2005 BAC.
Figure 3.1 Physical distance map of tr-Ds insertion BACs on Z. mays chromosome 10. The map below illustrates the distribution of tr-Ds hits across maize chromosome 10. Purple triangles represent the location of BACs containing one or more R-L tr-Ds insertions. The yellow triangles represent the location of BACs containing one or more R-UL tr-Ds insertions. The black numbers represent locations and map distance of various loci as described by the Genetics 2005 genetic map. Blue numbers are estimated map distances determined by comparing loci between the Genetic 2005 map and the IBM FPC0507 map. Because of the number of insertions in FPC ctg415, that section of the map has been enlarged to show detail.
r1-sc:m3 donor site
Fig. 3.2. Genetic distance map of tr-Ds insertion BACs on Z. mays chromosome 10
The map below illustrates the distribution of tr-Ds hits across maize chromosome 10. Purple triangles represent the location of BACs containing one or more R-L tr-Ds insertions. The yellow triangles represent the location of BACs containing one or more R-UL tr-Ds insertions. The black numbers represent locations and map distance of various loci as described by the Genetics 2005 genetic map. Blue numbers are estimated map distances determined by comparing loci between the Genetic 2005 map and the IBM FPC0507 map. Because of the number of insertions in FPC ctg415, that section of the map has been enlarged to show detail.
Fig. 3.3. Distribution of tr-Ds events around the donor locus R-sc. The BAC is the maize10L BAC used in Swigonova et. al., 2005. Figure is also adapted from said paper. Purple triangles represent hits by R-L families. Yellow triangles represent hits by R-UL families. Genes and their orientation are displayed by red arrows. The black boxes represent repetitive elements. The numbers used to represent genes are: 1) NADP-dependent malic enzyme; 2) pinhead protein; 3) anthocyanin regulatory R-S; 4) maize Isr gene; 5) putative S-domain receptor-like kinase; 6) aldose reductase-related protein; 7) conserved unknown expressed protein; 8) glutathione peroxidase-like protein; 9) putative glycerol-3-transporter; 10) response regulator; 11) 4-coumarate-CoA ligase-like protein.
CHAPTER 4: COMPARISON OF W22 INBRED LINE SEQUENCE AGAINST REFERENCE SEQUENCE FROM B73

Abstract

Since its domestication, maize has undergone selection for specific desirable traits. Various maize lines can carry significant genomic variation, as these lines may have been under selective pressure for different traits. Using sequence obtained from long fDs sequences obtained from the Ac/Ds project, we can compare W22, the inbred line used in the project, against B73, the inbred line selected for genome sequencing. Using BLAST, these sequences have been located around the host genome, as well as compared the sequence. The level of sequence identity across aligned regions is 96.5%, with an average rate of one SNP every 61 bp and one InDel every 155 bp. As efforts to sequence larger fDs fragments continue, we hope to further understand the variation between these inbred lines.

Introduction

Throughout history, humans have selected and modified their surroundings to better aid in survival. A major factor in this is the domestication of various plants to serve as food and material sources. As such, humans exert their own selective pressures upon their surroundings, forming a symbiotic relationship where a trait that may have been ineffectual or detrimental to the organism in the wild may be desirable to people who may raise and propagate them. This artificial selection force may give rise to various alterations or divisions that may never develop in the wild.

\textit{Zea mays}, as a cultivar crop plant raised by humans, comes under selective pressures not seen by wild organisms, as the wants or interests of those who raise it
fluctuate over time, typically to increase yields or make the plant more resilient to weather or disease. As such, many different traits have been selected for in maize. The ancestor of maize, teosinte, has very high genetic diversity (Buckler, et. al., 2001), which has contributed to many genetic variations to different maize lines, perhaps due to the different traits being selected for in the development of each line.

The Ac/Ds project sequencing pipeline presents the opportunity to compare genetic variation between different breeds of maize. Several different maize inbred lines exist. The current efforts for sequencing the maize genome have been on the inbred line B73. However, the Ac/Ds project uses the inbred line W22. Both inbred lines can be traced back to Reid Yellow Dent (RYD), a popular maize cultivar known for its high ear quality. In the early 20th century, maize cultivars were maintained by open pollination. Because of this, these cultivars contained greater genetic diversity than modern breeding inbreds. In the 1930s, a series of crosses using RYD landraces to select for above average stalk quality resulted in the Iowa Stiff Stalk Synthetic (BSSS) line. This line would later give rise to several useful inbred lines, among them B73 (1972). Initially, B73 showed great overall yield, but lacked a tolerance for corn borer attack. A later cross resulting in hybrid B73/Mo17 produced a very high quality and popular plant.

Reid Yellow Dent would also be used in the development of another cultivar, Funk Yellow Dent (FYD). While lacking the tight kernel placement of RYD, FYD was selected for increased disease resistance, and became a popular breed in its own right. FYD would later be used to create the Illinois B10 inbred line (Ill.B10), one of the parents of the W22 inbred line. The other parent, the W25 inbred line, can be traced back
to Golden Glow, an improved cultivar of Minnesota No. 13, a maize line known for its early maturation but large yields (See Fig. 4.1).

Though there has been little study of the differences between B73 and W22, research has been done with other inbred lines. In a previous study examining and mapping single nucleotide polymorphisms (SNPs) and InDels (insertions or deletions) between B73 and another inbred line, Mo17, an average rate of one SNP for every 73 bp and one InDel every 309 bp was obtained (Vroh Bi, et al. 2005). Another study obtained an average rate of one SNP for every 70 bp and one InDel every 160 bp (Rafalski, et al. 2001). These studies demonstrate a high degree of polymorphism between modern inbred maize lines.

To determine variations between the two inbred lines, we will be comparing sequence obtained by long fDs fragments against that of current GSS databases. Because of limitations in the sequencing reaction, the first pass of sequencing will only yield a few hundred base pairs of sequence on either end, we design new sequencing primers and work progressively across until a complete sequence is obtained. Based on the low repeat content of the fDs fragments (refer to chapter 2), we know that these events are biased toward non-repetitive regions, so this allows us to focus our comparisons on sequence of genic regions in the genome. At present, we have 17 completely sequenced fragments covering an area of ~44.6 kb, as well as another 28 at various stages of processing (~103 kb). By using sequence obtained from longer fDs fragments to compare against B73, it is possible to determine what genetic differences may have arisen since these two lines split from their common ancestor, Reid Yellow Dent, as well as pre-
Materials and Methods

Sample Selection

Large fragments are considered to be any fragment whose size estimate is equal to or greater than 2.0 kb. Samples also must have passed through the initial sequence reactions successfully.

Primer Walking

Primers are designed off the sequence obtained from the previous sequencing reactions by use of Primer3 version 0.3.0. Primers must be at least 100 bp from the end of the previous sequence. Melting temps are kept close to 60°C. The primer pair best conforming to these requirements and having the best functionality is chosen and synthesized. These primers are used in sequencing reaction as described in Chapter 2.

Reamplification

Typically, samples large enough to be considered large fragments are also fairly poor in concentration. In these cases, reamplification is done to increase the amount and concentration of available stock. PCR is performed using the custom primers designed above or a combination of custom and original sequencing primer. Template is usually the original stock sample, but in some cases proper reamplification can be done with genomic W22 DNA and custom primers. Successful reamplifications are ethanol precipitated and resuspended in HPCL grade water. The concentration is assessed via Nanodrop and optimized.

Sequence Assembly
Reads are processed as in Chapter 2. All the reads produced from large fragments are fed into a single Gap4 database, allowing addition on to any previous sequence as well as comparison and correction of any overlapping regions. These new larger contigs are used as the sequence for designing new primers as above. Primer walking continues until at least two sequences cover the length of the large fragment.

**BLAST**

The completed sequences are repeatmasked using RepeatMasker (Smit, Hubley, Green, 2004) and the TIGR maize repeat database 4.0 (http://maize.tigr.org/repeat_db.shtml). Completed fragment sequences are compared by BLAST against the HTGS and GSS assembly databases at NCBI and MAGI, respectively.

**Results**

**Sequencing Results**

To date, 45 tr-\textit{Ds} samples have been selected for complete sequencing. Of these, 17 have been completed with overlapping redundant sequence, covering an area of \( \sim 44.6 \) kb. The remaining 28 samples are at various degrees of processing, covering an area of \( \sim 103 \) kb.

**Sequence Placement**

Using the data from the BLAST with the HTGS database, we can determine the general distribution of the currently sequenced fragments. Of the 17 long sequences examined, 11 align to known FPC locations on the maize genome. Of these, five align to FPC locations on chromosome 10, which contains the \textit{Ds} donor site \( r1-sc:m3::Ds \). The remaining six are distributed across the genome. One lands on chromosome 1, one on
chromosome 3, two on chromosome 5, one on chromosome 6, and one on chromosome 8. Another sequence can be found on chromosome 3, but the FPC location cannot be determined (Fig. 4.2, Table 4.1).

**Sequence Analysis**

When repeatmasked long fragments are used as BLAST query against a GSS assembly database, we can compare the sequence of B73 and W22. Of the 17 completed fragments, there is a wide distribution of sequence identity. B.W06.0767C, with 1.3 kb of non-repetitive sequence, has 98.6% sequence identity across the length of the fragment. B.W06.0939, with 1.9 kb, has 38.2% identity across the length of the fragment. Overall, across the 33,910 bp of unmasked sequence, there is approximately 80.8% identity. However, this value is based on identity across the entire sequence, and gaps in the GSS alignment can reduce the percentage. These gaps account for 5,517 bp. There are two types of gaps: gaps where a region of the W22 sequence does not align to the B73 sequence (W22 gaps), and gaps where a region of the B73 sequence does not align to the W22 sequence (B73 gaps). The W22 gaps account for 1,506 bp, while the B73 gaps account for 5040 bp. There is 1029 bp overlap between the W22 and B73 gaps. If only the regions aligning to GSS sequence are used, there is approximately 96.5% identity across 28,393 bp of sequence.

SNP and InDel values were also calculated for the regions aligning to GSS assemblies. 469 SNPs and 183 InDels were located across the 28 kb of aligned sequence. This comes out to an average of one SNP for every 61 bp and one InDel for every 155 bp (Table 4.2).
Discussion

Previous studies have shown that there can be a fairly large degree of polymorphism between different maize lines, most likely due to the variable selective pressures on the already present diversity of its teosinte ancestor or genetic drift between the various inbred lines. Between the two inbred lines compared here, B73 and W22, we can see a relatively high degree of variation. Ac/Ds transposition have a bias toward gene-rich areas of the host genome, and making use of repeatmasker, we can focus on the variations present in gene regions. Based on the BLAST data using the HTGS database, there appears to be a fair distribution of sequences across the maize genome. There is a bias of hits to chromosome 10, but this is to be expected, as chromosome 10 also carries the \( r1-sc:m3::Ds \) donor site. The donor site is on FPC contig 415, and the five sequences align to contigs 409, 413, 414, and 419. These are linked transpositions, and are approximately 33 cM, 20 cM, 5 cM, and 20 cM away from the donor site, respectively. Two of these sequences are from \( R \)-linked samples, so this is to be expected. Between the GSS assembly data for B73 and the sequences obtained from W22, we find a 96.5% identity rate for aligned regions of the sequence, which is approximately 28 kb. Across these regions, we found 469 SNP and 183 InDels, coming out to an average rate of one SNP every 61 bp and one InDel every 155 bp (Table 4.2). This is comparable to the results from Rafalski’s study, which found an average rate of one SNP every 70 bp and one InDel every 160 bp (Rafalski, et al. 2001).

However, though these are average rates, these rates can vary widely. B.S05.0903R, a 3.8 kb sequence, has the highest concentration of SNPs, at a rate of about one SNP every 32 bp. I.W06.0605, a 1.3 kb sequence, has the highest concentration of
InDels, at a rate of about one InDel every 51 bp. B.W06.0767C, a 1.3 kb sequence, has the lowest concentration of each, with no SNPs or InDels present.

Considering the past data involving comparisons between B73 and Mo17, our data suggests a similar degree of variation between B73 and W22. Teosinte has a high degree of variability in its genome, and as different traits are selected for in each inbred line, it’s unsurprising to find strong differences between the genome sequences. However, it is worth noting that, as it stands now, these results are based on a fairly small data set (17 sequences covering 28,393 bp of sequences, a disproportionate number of which are located on chromosome 10). As the sequencing and analysis of these large fragments continues, we hope to create a clearer picture of the sequence variation between B73 and W22.

References


Fig. 4.1. History of B73 and W22 inbred lines. B73 is derived from Iowa Stiff Stalk Synthetic, a synthetic cultivar of Reid Yellow Dent. W22 is derived from Ill.B10 and W25. Ill.B10 was developed from Funk Yellow Dent, which was developed from Reid Yellow Dent. W25 is derived from Golden Glow, an improved cultivar developed from Minnesota No.13.
Fig. 4.2. Placement of large fragment sequences across the maize genome. Of the 11 sequences placed on the maize genome, five align to chromosome 10 (~45.5%). The remaining six are distributed among chromosomes 1, 3, 5, 6, and 8.
Table 4.1. Repeatmasking and location data for large fragment sequences.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Size(bp)</th>
<th>Unmasked(bp)</th>
<th>%Unmasked</th>
<th>Chromosome</th>
<th>FPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.W06.0605</td>
<td>1956</td>
<td>1297</td>
<td>66.3%</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>B.W06.0937</td>
<td>1486</td>
<td>1433</td>
<td>96.4%</td>
<td>3</td>
<td>131</td>
</tr>
<tr>
<td>I.W06.0603</td>
<td>1881</td>
<td>1848</td>
<td>98.2%</td>
<td>5</td>
<td>247</td>
</tr>
<tr>
<td>B.W06.0977</td>
<td>1806</td>
<td>1806</td>
<td>100.0%</td>
<td>5</td>
<td>254</td>
</tr>
<tr>
<td>I.W06.0612B</td>
<td>1914</td>
<td>1914</td>
<td>100.0%</td>
<td>6</td>
<td>264</td>
</tr>
<tr>
<td>B.W06.0939</td>
<td>1936</td>
<td>1936</td>
<td>100.0%</td>
<td>8</td>
<td>364</td>
</tr>
<tr>
<td>B.W06.0767C</td>
<td>1688</td>
<td>1331</td>
<td>78.9%</td>
<td>10</td>
<td>409</td>
</tr>
<tr>
<td>B.S05.0458</td>
<td>3157</td>
<td>3030</td>
<td>96.0%</td>
<td>10</td>
<td>413</td>
</tr>
<tr>
<td>B.S05.0903R</td>
<td>3876</td>
<td>3876</td>
<td>100.0%</td>
<td>10</td>
<td>414</td>
</tr>
<tr>
<td>B.S05.0764R</td>
<td>2076</td>
<td>1934</td>
<td>93.1%</td>
<td>10</td>
<td>419</td>
</tr>
<tr>
<td>B.W06.1023</td>
<td>1530</td>
<td>1507</td>
<td>98.5%</td>
<td>10</td>
<td>419</td>
</tr>
<tr>
<td>B.W06.0822</td>
<td>1373</td>
<td>1373</td>
<td>100.0%</td>
<td>3</td>
<td>UNK</td>
</tr>
<tr>
<td>I.W06.0333</td>
<td>2411</td>
<td>2411</td>
<td>100.0%</td>
<td>UNK</td>
<td>UNK</td>
</tr>
<tr>
<td>I.W06.0898R</td>
<td>2013</td>
<td>2013</td>
<td>100.0%</td>
<td>UNK</td>
<td>UNK</td>
</tr>
<tr>
<td>I.W06.0946</td>
<td>1748</td>
<td>1748</td>
<td>100.0%</td>
<td>UNK</td>
<td>UNK</td>
</tr>
<tr>
<td>B.S06.0568</td>
<td>2784</td>
<td>1899</td>
<td>68.2%</td>
<td>UNK</td>
<td>UNK</td>
</tr>
<tr>
<td>I.W06.0271</td>
<td>2554</td>
<td>2554</td>
<td>100.0%</td>
<td>UNK</td>
<td>UNK</td>
</tr>
</tbody>
</table>
Table 4.2. Identity and SNP/InDel content data for large fragment sequences.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Unmasked Size(bp)</th>
<th>GSS Coverage(bp)</th>
<th>Total Identity%</th>
<th>Coverage Identity%</th>
<th>SNPs</th>
<th>InDels</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.W06.0605</td>
<td>1297</td>
<td>1071</td>
<td>78.4%</td>
<td>95.0%</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>B.W06.0937</td>
<td>1433</td>
<td>1419</td>
<td>97.2%</td>
<td>98.2%</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>I.W06.0603</td>
<td>1848</td>
<td>1554</td>
<td>82.7%</td>
<td>98.3%</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>B.W06.0977</td>
<td>1806</td>
<td>909</td>
<td>48.9%</td>
<td>97.1%</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>I.W06.0612B</td>
<td>1914</td>
<td>1551</td>
<td>80.4%</td>
<td>99.2%</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>B.W06.0939</td>
<td>1936</td>
<td>741</td>
<td>38.2%</td>
<td>99.9%</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B.W06.0767C</td>
<td>1331</td>
<td>1312</td>
<td>98.6%</td>
<td>100.0%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B.S05.0458</td>
<td>3030</td>
<td>2513</td>
<td>76.8%</td>
<td>92.6%</td>
<td>68</td>
<td>39</td>
</tr>
<tr>
<td>B.S05.0903R</td>
<td>3876</td>
<td>3546</td>
<td>85.1%</td>
<td>93.0%</td>
<td>112</td>
<td>37</td>
</tr>
<tr>
<td>B.S05.0764R</td>
<td>1934</td>
<td>1689</td>
<td>86.2%</td>
<td>98.7%</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>B.W06.1023</td>
<td>1507</td>
<td>1492</td>
<td>98.4%</td>
<td>99.4%</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>B.W06.0822</td>
<td>1373</td>
<td>1361</td>
<td>97.2%</td>
<td>98.1%</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>I.W06.0333</td>
<td>2411</td>
<td>1282</td>
<td>48.9%</td>
<td>92.0%</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>I.W06.0898R</td>
<td>2013</td>
<td>1858</td>
<td>86.2%</td>
<td>93.4%</td>
<td>61</td>
<td>26</td>
</tr>
<tr>
<td>I.W06.0946</td>
<td>2004</td>
<td>1837</td>
<td>91.3%</td>
<td>99.6%</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>B.S06.0568</td>
<td>1899</td>
<td>1827</td>
<td>89.8%</td>
<td>93.3%</td>
<td>78</td>
<td>17</td>
</tr>
<tr>
<td>I.W06.0271</td>
<td>2554</td>
<td>2431</td>
<td>94.7%</td>
<td>99.5%</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>33910</td>
<td>28393</td>
<td>80.8%</td>
<td>96.5%</td>
<td>469</td>
<td>183</td>
</tr>
</tbody>
</table>
CHAPTER 5: GENERAL CONCLUSIONS

Transposons, by their mobile nature, provide a valuable resource for genetic analysis, especially mutant screens. As the first identified transposons, it’s fitting that the *Activator* and *Dissociation* elements would be one of the best characterized, and useful for such analysis.

The *Ac/Ds* project aims to cast stable *Ds* elements throughout the maize genome. By using *Ac::im*, which is incapable of transposition but still producing transposase, the *Ds* element in *r1-sc:m3* can be mobilized throughout the genome. Because *Ac::im* is unlinked to *r1-sc:m3*, it can be segregated away through subsequent crosses, preventing any new tr-*Ds* events from mobilizing again. These stable *Ds* lines can be used in further research in forward genetic screens as a valuable resource to the maize community at large. As it stands now, we have a reliable and effective system by which to cast, identify, and locate new insertions throughout the maize genome. Of the 1221 fDs fragments that have been isolated, 941 lines have been successfully sequenced and processed. The current pipeline has approximately a 77.0% success rate for sequencing new fDs fragments on a single pass. For the samples selected for unlinked transposition events, we find a fairly broad distribution across the maize genome compared to linked samples, though there are still some biases toward the donor chromosome, as well as biases against the chromosome carrying *Ac::im*.

The *Ac/Ds* project has also provided useful data for tangential studies. By mapping *Ds* insertions on chromosome 10, the *Ds* donor chromosome, we can observe patterns in transposition activity. Based on our maps, we can see that new insertion events are biased toward positions near the donor site. However, the patterns of
transposition seem to differ based on the conditions of the transposition. Families selected for unlinked transpositions tend to show a broad and even distribution around the donor site. Because these events are selected for unlinked transposition events, this is to be expected. Families not under this selection are clustered much closer and preferentially distal to the donor site.

By sequencing larger fragments produced by the Ac/Ds project, we can compare the sequence of W22 against B73. These alignments show 96.5% identity across 28 kb of sequence. The variation between W22 and B73 is comparable to that between B73 and Mo17, supporting the idea that the natural variation found in teosinte has given rise to the differences found across various maize lines. These results are still somewhat preliminary, but as more sequence data is collected, we can refine the picture of maize line variation.

In total, the Ac/Ds project can be very valuable in analysis of the maize genome, including studying the movement of transposable elements. As more data is collected and techniques further refined, the Ac/Ds project will continue to grow as a valuable resource for maize genetics studies.
ACKNOWLEDGEMENTS

I would like to thank my major professor, Erik Vollbrecht, for all his help and guidance throughout my graduate study. I would also like to thank the members of my program of study committee, Thomas Peterson and Volker Brendel. I also want to acknowledge all the members of the Vollbrecht lab, including Erica Unger Wallace, Michael Muszynski, Brandi Sigmon, Xiang Yang, Wei Li, Becky Weeks, Jacqueline Farrell, Mahima Bajpaj, Zhuying Li, as well as all the undergraduates who have worked there along side me. I want to thank our collaborators on the Ac/Ds project, Thomas Brutnell and Volker Brendel, and all the members of their labs, including Kevin Ahern, Prasit Deewatthanawong, Liza Conrad, Kazuhiro Kikuchi, Jon Duvick, and Mukul Sabharwal. Thanks also to Linda Wild for all her help and advice. I want to acknowledge my family and friends for all their support. Finally, I want to thank my fiancé Jennifer, who’s been the best support and encouragement I could ever hope for.