THE EVOLUTION OF CENTROMERIC SATELLITE REPEATS IN MAIZE

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOSCIENCES AND BIOENGINEERING

AUGUST 2012

By

Megan M. Nakashima

Thesis Committee:

Gernot Presting, Chairperson
Paul Moore
Robert Paull

Keywords: satellite repeat, CentC, maize
# Table of Contents

Abstract ................................................................................................................................. 3
List of Tables .......................................................................................................................... 4
List of Figures ......................................................................................................................... 5
Abbreviations ........................................................................................................................ 10

CHAPTER 1. BACKGROUND ................................................................................................. 11
  1.1 Introduction ..................................................................................................................... 11
  1.2 Centromeric satellite repeats ...................................................................................... 11
    1.2.1 Alpha satellite repeats in humans .......................................................................... 11
    1.2.2 AtCon satellite repeats in Arabidopsis thaliana ...................................................... 12
    1.2.3 CentO satellite repeats in rice .............................................................................. 12
    1.2.4 CentC satellite repeats in maize ........................................................................... 13
  1.3 Centromeric Retrotransposons (CR) ........................................................................ 14
    1.3.1 Centromeric retrotransposon of maize (CRM) .................................................... 14
  1.4 CENH3 centromeric proteins ...................................................................................... 14
  1.5 Maize evolution and genome size ............................................................................ 15
  1.6 Artificial chromosomes ............................................................................................ 15
  1.7 Conclusion .................................................................................................................... 16
  1.8 Hypothesis .................................................................................................................. 16
    1.8.1 Objectives .................................................................................................................

CHAPTER 2. METHODS ........................................................................................................ 18
  2.1 Isolating genomic Mo17 DNA ..................................................................................... 18
  2.2 Haell digest, end repair, and gel extraction ............................................................... 18
  2.3 Concentration, cloning, and CentC screening ............................................................ 18
  2.4 Insert size determination of CentC positive clones and end sequencing .......... 19
  2.5 Subcloning .................................................................................................................. 19
    2.5.1 Large fragments ..................................................................................................... 19
    2.5.2 Small fragments ..................................................................................................... 20
  2.6 Isolating the B73 counterpart ..................................................................................... 21

CHAPTER 3. RESULTS AND DISCUSSION .......................................................................... 22
  3.1 Isolating genomic Mo17 DNA ..................................................................................... 22
  3.2 Haell digest, end repair, and gel extraction ............................................................... 23
  3.3 Ligation and transformation ....................................................................................... 24
3.4 CentC PCR screening ..........................................................25
3.5 Insert size determination and end sequencing ................................26
3.6 Subcloning ........................................................................28
  3.6.1 Large fragments ..............................................................28
  3.6.2 Small fragments ..............................................................30
3.7 Isolated and subcloned Mo17 CentC islands ....................................32
  3.7.1 Mo17 clones that have a B73 counterpart ...............................33
    3.7.1.1 Analysis of Clone 1 (M.L3.T1.1.G3) ..............................33
    3.7.1.2 Analysis of Clone 2 (M.L3.T1.2.E1) ..............................36
  3.7.2 Comparison of centromeric and non-centromeric B73 CentC islands .................................................39
  3.7.3 Clones that did not have an available B73 counterpart .................40
    3.7.3.1 Clone 3 (M.L3.T1.2.G11): possible CentA insertion ..........40
    3.7.3.2 Analysis of Clone 4 (M.L3.T4.2.D3): Multiple HORs ..........41
    3.7.3.3 Analysis of Clone 5 (M.L3.T4.3.H1): Three monomer HOR ........43
  3.7.4 Summary of Mo17 CentC island clones ................................44
  3.8.1 Subcloned B73 BAC C0030B18 CentC islands ............................46
    3.8.1.1 Analysis of Clone 6 (B.L2.T1.M13) ..............................46
    3.8.1.2 Analysis of Clone 7 (B.L2.T1.O22) ..............................48
    3.8.1.3 Analysis of Clone 8 and 9 (B.L2.T1.M16.1 and M16.2) ........50
      Analysis of Clone 8 (B.L2.T1.M16.1) ..................................51
      Analysis of Clone 9 (B.L2.T1.M16.2) ..................................52
    3.8.1.4 Analysis of Clone 10 (B.L2.T1.M20) ..............................55
    3.8.1.5 Analysis of Clone 11 (B.L2.T1.N14) ..............................59
    3.8.1.6 Analysis of Clone 12 (B.L2.T1.K21) ..............................61
  3.8.2 Junctions with retrotransposons ...........................................62
    3.8.2.1 CRM1 junctions .......................................................62
    3.8.2.2 cinful junctions .......................................................62
  3.8.3 Summary of B73 BAC C0030B18 subcloning .............................63
  3.9 CentC monomer comparison ....................................................63

CHAPTER 4. CONCLUSION ................................................................64
CHAPTER 5. SUMMARY ..................................................................65
REFERENCES ..............................................................................66
Abstract

Corn is an important agricultural crop and an excellent model organism for genetic studies. A better understanding of corn centromeres will help in the design of artificial chromosomes that may be used to further improve the corn germplasm in the future. The tandem repeat CentC (156 nucleotides long) has been shown by fluorescent in situ hybridization to be limited to the centromeric regions. Little is known about its function there or how CentC evolves in different maize lines. The maize inbred B73 genome has been sequenced and to date over 10,000 copies of CentC have been described in this line. I have begun to characterize the CentC repeats of the inbred Mo17 and will use the B73 reference genome to extrapolate the evolutionary history of CentC in maize.

A total of 92,112 base pairs of CentC and their junctions with non-CentC sequence were cloned and sequenced. Similar to the alpha satellite repeat in primates, there were many higher order repeats found within the CentC islands. However, unlike the alpha satellite repeat, not many were homogenous within a certain island. Unlike the AtCon sequence, there were no distinct conserved regions within the repeat. And unlike CentO found in rice, there were no other CentC-like types found other than CentC.

In this work, I have shown the presence of higher order repeats in CentC which provides an idea of the size by which CentC islands grow. By looking at the nucleotide differences between the CentC islands of two inbreds I was able to determine the relative age of the higher order repeats. I have also shown that a single CentC monomer can be lost in a CentC island and that retrotransposon insertions into the CentC island are rampant. I was also able to compare centromeric and non-centromeric CentC islands from the same inbred. While there were differences between the two, more study is required to come to definite conclusions about the sequence differences. While resequencing a CentC-containing BAC, confusing results revealed the possible instability of CentC-rich BAC clones.
List of Tables

Table 1: Count of small and large colonies from 20090107 transformation ........................................... 24
Table 2: Count of small and large colonies from 20090120 transformation .......................................... 25
Table 3: Overview of Mo17 CentC islands .................................................................................................. 32
Table 4: Centromeric and non-centromeric differences in B73 CentC islands ........................................ 39
Table 5: Overview of B73 BAC C0030B18 CentC islands ........................................................................ 46
List of Figures

Figure 1: Agarose gel analysis and gel extraction of 20090227 sheared and end repaired DNA of M.L3.T1.1.G3. (A) Before gel extraction and (B) after gel extraction of the 1 to 8 kb sheared and end repaired M.L3.T1.1G3 insert. A piece of gel corresponding to 1 kb to 8 kb in size was excised from the gel. M.L3.T1.1.G3 uncut plasmid “1”, M.L3.T1.1.G3 sheared end repaired DNA “2”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

Figure 2: Agarose gel analysis of genomic DNA extraction procedures using B73 and Mo17 inbreds. (A) Genomic DNA extraction procedure 1 with genomic B73 DNA “1 and 2”, genomic Mo17 DNA (using newer kit) “3”, genomic Mo17 DNA (using older kit) “4”, and Tri-Dye 2-log ladder “L”. (B) Genomic DNA extraction procedure 2 with genomic Mo17 genomic DNA (20081206) “1”, Mo17 (1) genomic DNA (20081212) “2”, Mo17 (2) genomic DNA (20081212) “3”, Mo17 (1) genomic DNA (20081212) “4”, Mo17 (2) genomic DNA (20081216) “5”, and TriDye 2-log ladder “L”.

Figure 3: Agarose gel analysis and extraction of HaeIII restriction digest of 20081216 Mo17 (2) using a dialysis/filter method. (A) Before insertion of the silica filter/membrane, (B) insertion of the silica filter/membrane with white boxes indicating the silica filter/membrane at the top and a strip of dialysis membrane on the bottom, and (C) after the gel was allowed to run and the DNA ran into the silica filter/membrane, with the white arrow indicating which direction the gel was run. Genomic DNA “1”, HaeIII digest “2”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

Figure 4: Agarose gel analysis and extraction of HaeIII restriction digest of 20081216 Mo17 (2) using a razor extraction method. (A) Before gel extraction and (B) after extraction of gel fragment from 6kb to well using a razor blade. HaeIII digest “1”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

Figure 5: Agarose gel analysis of 20090123 CentC PCR using primers CentC_F and CentC_R. 100 clones (1 per well) “Rows A to H”, negative control (no DNA) “-”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”. Clones of interest are labeled with their corresponding well number on a 96-well PCR plate.

Figure 6: Agarose gel analysis of 20090129 NotI digest of M.L3.T1.1 clones. CentC positive clone “A8 – H9”, left lane for each clone is uncut plasmid, right lane for each clone is plasmid cut with NotI, and NEB Tri-dye 2-log ladder “L”. Vector arms bands at 10 kb and 2.2 kb.

Figure 7: Agarose gel analysis of M.L3.T1.1 clones. CentC positive clone “A8 – H9”, left lane for each clone is uncut plasmid, right lane for each clone is plasmid cut with HaeIII, and NEB Tri-dye 2-log ladder “L”. The low molecular weight ladder represents the vector arms being digested by HaeIII. Any remaining higher molecular weight band is the insert DNA fragment.

Figure 8: Agarose gel analysis of 20090310 NotI digest and 20090722 HaeIII digest of M.L3.T1.1 clones. CentC positive clone “C4 – H6” plasmid cut with (A) NotI and (B) HaeIII, and NEB Tri-dye 2-log ladder “L”. Vector arms are represented by 10 kb and 2.2 kb bands in gel A and by a low molecular weight ladder in gel B.

Figure 9: Agarose gel analysis of 20090804 NotI and HaeIII digest of M.L3.T4 clones. CentC positive clone “1.C2 – 4.H1”, left lane for each clone is uncut plasmid, middle lane for each clone is plasmid cut with HaeIII, right lane is plasmid cut with NotI, and NEB Tri-dye 2-log ladder “L”.

Figure 10: Agarose gel analysis and gel extraction of 20100318 NotI digest of M.L3.T4.2.D3. (A) Before gel extraction and (B) after gel extraction of the 5kb 2.D3 insert. M.L3.T4.2.D3 NotI digest “2.D3”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

5
Figure 11: Agarose gel analysis of 20100408 Insert/CentC PCR using primers SL1, NZRevC, and CentC_F. Sheared subclones of M.L3.T1.2.E1 “Rows A to H, Plates 1 and 2”, PCR negative control (no DNA) “-”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”. Clones of interest are circled red. 

Figure 12: Overview of the M.L3.T1.2.G11 subcloning method. (1) HaeIII digest of genomic Mo17 DNA, (2) gel extraction of HaeIII digest, (3) BsaWI cut map of end sequences, (4) gel extraction of BsaWI digest, (5) forward and reverse sequencing of subclone fragments, and (6) the final assembly of the subclone fragments, including size of each fragment.

Figure 13: Overview of the M.L3.T4.3.H1 subcloning method. (1) HaeIII digest of genomic Mo17 DNA, (2) gel extraction of HaeIII digest, (3) BsiWI cut map of end sequences, (4) gel extraction of BsiWI digest, (5) forward and reverse sequencing of subclone fragments, and (6) the final assembly of the subclone fragments, including size of each fragment.

Figure 14: Junction viewer image of Clone 1. CentC monomers are represented by green arrows and a CRM1 element is represented by a blue block.

Figure 15: Junction viewer image of B73 BAC b0410L22 counterpart H8. CentC monomers are represented by green arrows.

Figure 16: Phylogenetic tree of individual CentC monomers from both Clone 1 consensus sequence and its B73 counterpart consensus sequence. Clone1 CentC monomers “G3-#” and B73 counterpart CentCs “H8-#”.

Figure 17: Graphic depiction of individual CentC monomers in Clone 1 and its B73 counterpart. CentC monomers “colored arrows pointing to the right”, CRM element “colored arrow pointed to the left”, nucleotide differences “colored vertical lines”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 18: Junction viewer image of Clone 2. CentC monomers are represented by green arrows.

Figure 19: Junction viewer image of the B73 counterpart of Clone 2. CentC monomers are represented by green arrows.

Figure 20: Phylogenetic tree of individual CentC monomers from both Clone 2 consensus sequence and its B73 counterpart consensus sequences. Clone2 CentC monomers “E1-#” and B73 counterpart CentCs “D8-#” and “B10-#”.

Figure 21: Graphic depiction of individual CentC monomers in Clone 2 and its B73 counterpart. CentC monomers “colored arrows pointing to the right”, CRM element “colored block”, nucleotide differences “vertical colored lines”, CentC Mo17 sequence trends “thin colored blue arrows above the CentC monomers”, and CentC B73 sequence trends “thin colored green arrows below the CentC monomers”.

Figure 22: Agarose gel analysis of HaeIII digest of BAC CH201-433P11. Red box indicates two clones that were excised and cloned into pJAZZ together, due to their close proximity to each other. HaeIII digested BAC CH201-433P11 “1”, NEB TriDye 2-log ladder “L”, and empty lane “X”.

Figure 23: Junction viewer image of Clone 3 (M.L3.T1.2.G11). CentC monomers are represented by green arrows and sequence corresponding to CentA is in orange.

Figure 24: Phylogenetic tree of individual CentC monomers from Clone 3 consensus sequence. Clone 3 CentC monomers are labeled “G11-#”. 
Figure 25: Graphic depiction of individual CentC monomers in Clone 3. CentC monomers “colored arrows pointing to the right” and CentA-like sequence “colored arrow pointed to the left”. ........................................ 41

Figure 26: Junction viewer image of Clone 4 (M.L3.T4.2.D3). CentC monomers are represented by green arrows and sequence corresponding to CRM1 is in blue. ................................................................. 41

Figure 27: Phylogenetic tree of individual CentC monomers from Clone 4 consensus sequence. Clone 4 CentC monomers are labeled “2d3-#”. ................................................................. 42

Figure 28: Graphic depiction of individual CentC monomers in Clone 4. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”. ......................................................................................................................... 42

Figure 29: Junction viewer image of Clone 5 (M.L3.T4.3.H1). CentC monomers are represented by green arrows. ......................................................................................................................... 43

Figure 30: Phylogenetic tree of individual CentC monomers from Clone 5 consensus sequence. Clone 5 CentC monomers are labeled “3H1-#”. ................................................................. 43

Figure 31: Graphic depiction of individual CentC monomers in Clone 5. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”. ......................................................................................................................... 43

Figure 32: Possible cause for Clone 5’s HOR organization. 1. Initially the HOR array was homogenous. 2. A monomer experiences nucleotide changes, thus distinguishing this HOR from the rest. 3. The HOR that contained the nucleotide change locally duplicated. ................................................................. 44

Figure 33: Graphic representation of the comparison between the subcloned B73 counterpart for Clone 1 and contigs 1 and 4 of BAC b0410L22. CentC islands are represented by color coded arrows and corresponding patterns between the BAC data and experimental data are represented by thin colored arrows above the monomers. ......................................................................................................................... 45

Figure 34: Junction viewer image of Clone 6 (B.L2.T1.M13). CentC monomers are represented by green arrows. ......................................................................................................................... 46

Figure 35: Phylogenetic tree of individual CentC monomers from Clone 6 consensus sequence. Clone 6 CentC monomers are labeled “M13-#”. ................................................................. 47

Figure 36: Graphic depiction of individual CentC monomers in Clone 6. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”. ......................................................................................................................... 48

Figure 37: Junction viewer image of Clone 7 (B.L2.T1.O22). CentC monomers are represented by green arrows and CRM1 sequence is represented by a blue block. ................................................................. 48

Figure 38: Phylogenetic tree of individual CentC monomers from Clone 7 consensus sequence. Clone 7 CentC monomers are labeled “o22-#”. ................................................................. 49

Figure 39: Graphic depiction of individual CentC monomers in Clone 7. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, and CentC sequence trends “thin colored arrows above the CentC monomers”. ......................................................................................................................... 50
Figure 40: Junction viewer image of Clone 8 (B.L2.T1.M16.1). CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.

Figure 41: Phylogenetic tree of individual CentC monomers from Clone 8 consensus sequence. Clone 8 CentC monomers are labeled “M16.1-#”.

Figure 42: Graphic depiction of individual CentC monomers in Clone 8. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 43: Junction viewer image of Clone 9 (B.L2.T1.M16.2). CentC monomers are represented by green arrows and sequence corresponding to a hypothetical protein is represented by a gray block.

Figure 44: Phylogenetic tree of individual CentC monomers from Clone 9 consensus sequence. Clone 9 CentC monomers are labeled “M16.2-#”.

Figure 45: Graphic depiction of individual CentC monomers in Clone 9. CentC monomers “colored arrows pointing to the right”, hypothetical protein “block with yellow fading to red”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 46: Proposed theory on how the Clone 9 CentC island has evolved. Like HORs are represented by same colored arrows. First a red and blue arrow pair of HORs duplicates locally (1), followed by a blue arrow HOR duplicating locally (2), then an insertion of a few CentC monomers which happen to be related to part of the red arrow HOR (3), finished by a local duplication of a pair of blue and green arrow HORs (4).

Figure 47: Junction viewer image of Clone 10 (B.L2.T1.M20). CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.

Figure 48: Phylogenetic tree of individual CentC monomers from Clone 10 consensus sequence. Clone 10 CentC monomers are labeled “M20-#”.

Figure 49: Graphic depiction of individual CentC monomers in Clone 10. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 50: Clustal alignment of Clone 8 (B.L3.T1.M16.1) and Clone 10 (B.L3.T1.M20).

Figure 51: Junction viewer image of Clone 11 (B.L2.T1.N14). CentC monomers are represented by green arrows, and cinful sequence is represented by a gray block.

Figure 52: Phylogenetic tree of individual CentC monomers from Clone 11 consensus sequence. Clone 11 CentC monomers are labeled “N14-#”.

Figure 53: Graphic depiction of individual CentC monomers in Clone 11. CentC monomers “colored arrows pointing to the right”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 54: Junction viewer image of Clone 12 (B.L2.T1.K21). CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.
Figure 55: Phylogenetic tree of individual CentC monomers from Clone 12 consensus sequence. Clone 12 CentC monomers are labeled “K21-#”. ................................................................. 61

Figure 56: Graphic depiction of individual CentC monomers in Clone 12. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”. ........................................ 62

Figure 57: The evolution of CentC islands can be seen by looking at their HORs and nucleic acid differences. Higher order repeats show evolution through local duplications and subtle monomer changes. When comparing CentC sequences between inbreds, nucleic acid differences may give a hint at older and newer parts of the sequence. .................................................................................. 64

Figure 58: All the clones obtained in this study. Green lines indicate clones from B73 DNA and blue lines indicating Mo17 DNA. Black arrows indicate a match between Mo17 and B73 clones. ........................................ 65
Abbreviations

BAC – bacteria artificial chromosome
bp – base pairs
BSSS – Iowa Stiff Stalk Synthetic
CENH3 – centromere specific histone H3
ChIP – chromatin immunoprecipitation
CR – centromeric retrotransposon
CRM – centromeric retrotransposons of maize
CRR – centromeric retrotransposons of rice
DNA – deoxyribonucleic acid
FISH – fluorescent in situ hybridization
Gb – gigabase
HAC – human artificial chromosome
HOR – higher order repeat
IPTG – isopropyl β-D-1-thiogalactopyranoside
kb - kilobase
LTR – long terminal repeat
µl - microliter
MAC – maize artificial chromosome
Mb – megabase
MSC1 – maize centromere associated sequence
Mya – million years ago
ORF – open reading frame
PCR – polymerase chain reaction
rpm – revolutions per minute
TB – terrific broth
UTR – untranslated region
UV – ultra violet
X-gal – 5-bromo-4-chloro-indolyl-β-D-galactopyranoside
CHAPTER 1. BACKGROUND

1.1 Introduction

The centromere is located at the point on the chromosome where the two sister chromatids join. It is the site of spindle attachment during cell division, and because of this it is very important for chromosome segregation. The function of the centromere must be conserved for all eukaryotic species since cell division is a fundamental requirement for cells.

The deoxyribonucleic acid (DNA) sequence of the centromere, however, shows little or no homology between species. In most cases, the centromere is embedded in long arrays of highly repetitive DNA known as centromeric satellite repeats (Csink & Henikoff, 1998). They can be found in mammals such as humans as well as plants such as Arabidopsis, rice, and maize.

Also found in the centromere is the centromere-specific CENH3 histone variant, which can be used in chromatin immunoprecipitation methods to detect the functional centromere region of chromosomes. Retrotransposons are also interspersed amongst the centromeric satellite repeats in rice and maize.

Although the rice genome has been sequenced and the centromeric repeats in rice have been studied extensively (Cheng et al., 2002), maize makes for an ideal model organism to study centromeric satellite repeats for higher plants because of its more representative large genome size (Rabinowicz & Bennetzen, 2006).

Maize artificial minichromosomes were constructed by inverted centromeric satellite arrays and other centromeric elements (Ananiev et al., 2009). The possibility of these being produced de novo was suggested, but additional study is required. More knowledge of the centromeric satellite repeats in maize can contribute greatly to this biological application and to the study of centromeres as a whole.

1.2 Centromeric satellite repeats

At first, repetitive DNA was labeled as “junk DNA” (Ohno, 1972). Orgel and Crick stated that junk DNA lacked specificity and gave little or no selective advantage to the organism it was found in (Orgel & Crick, 1980).

The term “satellite DNA” was originally used on the genomic DNA that was separated from the majority of the DNA by cesium chloride buoyant density centrifugation, in which the density and the G/C content differed from the main band of DNA (Kit, 1961). This definition was later changed when it was found that ribosomal DNA, mitochondrial DNA, and chloroplast DNA could also form bands separate from the main band of DNA during centrifugation.

Restriction satellite DNA was characterized by one or more sites for specific restriction endonucleases found within tandemly arranged satellite DNA. This form of satellite DNA was found to be mainly in the subtelomeric and centromeric regions of the chromosome (Flavell, 1986). Tandem arrays with short monomers of about 15-30 base pairs (bp) each are termed minisatellite DNA and those with extremely short monomers (2 to 6 bp each) are called microsatellite DNA (Kashi, King, & Soller, 1997).

Hemleben et al. (2007) defined satellite DNA as “as species- or genus- specific component of eukaryotic genomes, and, in particular, satellite DNA of a species or one of its closest relatives consists of numerous tandemly arranged repeats that are non-coding, late-replicating in S-phase, and mostly located in the constitutive, non-transcribed heterochromatin”.

Satellite DNA is found in large arrays at the centromeres of many model organisms including humans, Arabidopsis thaliana, rice, and maize. Because of the size of these arrays and its uniformity in sequence, it has been very difficult to sequence with standard sequencing techniques and is usually left as gaps in otherwise well-sequenced model organisms (Lamb, Theuri, & Birchler, 2004).

1.2.1 Alpha satellite repeats in humans

The alpha satellite repeat is a satellite sequence that has been found in all primate chromosomes studied so far. It is a tandem repeat with a monomer that is 171 bp in length. The monomer is organized into different higher order repeat (HOR) arrays that are representative of
different subfamilies of the alpha satellite repeat. They are present multiple times in the centromeres of single or a specific subset of chromosomes (Willard & Waye, 1987).

One hypothesis on the sequence variation in these highly homogenized arrays is that these arrays occur and are maintained by the unequal crossover of sister chromatids. Evidence of this mechanism is diverged monomers at the edges of the arrays (Smith, 2010).

Rudd, Wray, and Willard (2006) demonstrated that the higher order and monomeric alpha satellite repeats in chimpanzees and humans evolve at different rates. Orthologous HORs are less conserved than orthologous monomeric repeats. Alpha satellite HORs within an array are found to be extremely homogenous, with very few other sequences inserted into the array. In contrast, monomeric alpha satellites are more heterogenous and have numerous other sequences interspersed within the array (Schueler, Higgins, Rudd, Gustashaw, & Willard, 2001).

1.2.2 AtCon satellite repeats in Arabidopsis thaliana

Previously called the 180 bp satellite repeat found in the centromeres of Arabidopsis thaliana (Simoens, Gielen, Montagu, & Inze, 1988), AtCon was found to actually have a median length of 178bp (Heslop-harrison, Murata, Ogura, Schwarzacher, & Motoyoshi, 1999). Chromatin immunoprecipitation (ChIP) data suggests that the satellite repeat is the main component in the centromere/kinetochore complex, although only a subset of it (~15%) is actually incorporated in the centromere/kinetochore complex. Athila retroelements were not detected by ChIP assays, implying that the Athila elements rarely insert into the functional centromere of Arabidopsis (Nagaki, Talbert, et al., 2003).

Each chromosome has specific AtCon sequences, suggesting that they either became homogenized on each chromosome independently, or were distributed amongst the chromosomes and then amplified with specific variants on each chromosome (Heslop-harrison, Murata, Ogura, Schwarzacher, & Motoyoshi, 1999).

Hall, Kettler, and Preuss (2003) collected satellite sequences from 41 Arabidopsis ecotypes and created consensus sequences for each, as well as a consensus sequence for the species as a whole. They found that there were prominent sequence differences between the species consensus and the ecotype consensus sequences. They also identified three 15 bp conserved domains (C1, C2, and C3), as well as a 25 bp variable domain (V1) within the satellite repeat monomer. The presence of the conserved regions may be proof of important protein-binding sites and the presence of variable region may be an area where such strict sequence conservation is not required. Another explanation for the variable domain is that it may be under selection to remain a polymorphic region in order to create diversity amongst monomers in arrays.

In a comparison with human alpha satellite DNA, Hall, Kettler, and Preuss (2003) observed that the edge of Arabidopsis arrays were not any more variable than random genomic sequences. This suggests that Arabidopsis and human alpha satellite arrays may have different maintenance mechanisms.

The satellite repeats were more variable across the Arabidopsis populations than any other single-copy sequence that was examined. As a whole, there is significantly less variation in Arabidopsis satellite repeats than in human alpha satellite repeats. The authors speculate that this may be because Arabidopsis is a self pollinating plant and does not require as much heterozygosity and genetic diversity as organisms that need to outcross (Hall, Kettler, & Preuss, 2003).

1.2.3 CentO satellite repeats in rice

In rice centromeres, the main satellite repeat, CentO, is 155 bp in length. Centromeric retrotransposons of rice (CRR) are dispersed amongst the CentO repeats within the centromere (Cheng et al., 2002). Although CentO is the dominant satellite sequence in rice, it is not the only one. Sequences that show some relationship to CentO either have replaced it in some species or coexist with CentO in the centromere.

Of the different genomes found in the genus Oryza (AA, BB, CC, BBCC, CCDD, EE, FF, GG, and HHJJ), three (genomes CC, FF, and GG) do not produce hybridization signals to CentO. Using chromatin immunoprecipitation (ChIP) cloning libraries, two classes of tandem repeats,
CentO-C1 (126 bp) and CentO-C2 (366 bp), in *O. rhizomatis* (genome CC) and a single class tandem repeat, CentO-F (154 bp), in *O. brachyantha* (genome FF) were found at the respective centromeres (Lee et al., 2005).

Fluorescent in situ hybridization (FISH) analysis showed that CentO-C1 and CentO-F were located exclusively at the centromeres of their respective species. The varying intensities of the FISH signals suggest that the amount of each repeat varied between chromosomes of their respected organism. CentO-C2 was detected in both centromeric and telomeric regions of the chromosomes inspected.

Through phylogenetic comparison, it was found that CentO-C1 is derived from a common ancestor of CentO and CentC, the maize centromeric repeat equivalent. The CentO-C1/CentO divergence predates the divergence between rice and maize. The CentO-C2 repeat is thought to have originated outside of the centromere at the telomeres and was later recruited into the centromere. CentO-F is believed to have some property that allows it to be strongly favored in meiosis so that the repeats are fixed in the population. This is due to the absence of CRR elements, which are susceptible to increased rates of satellite homogenization, in centromeres with CentO-F (Lee et al., 2005).

Another variation of the CentO repeat, named CentO-C was found in *O. officinalis* (Bao et al., 2006). This sequence is the same length as CentO-C1 (126 bp) and shares about 76% sequence similarity. Whether the sequence originated in *O. officinalis* or *O. rhizomatis* could not be determined due to the close evolutionary relationship of the two species, and was thus given a separate name. Southern blot confirmed the presence of low amounts of CentO in both *O. officinalis* and *O. rhizomatis* (Bao et al., 2006) where in previous Southern blots CentO was absent from the CC genome (Lee et al., 2005). In both species CentO is confined to one centromere.

### 1.2.4 CentC satellite repeats in maize

Similar in structure to rice centromeres, short tandem repeats and centromere-specific and noncentromere-specific retrotransposons make up the maize centromere. CentC is a 156 bp tandem repeat limited to the centromeric regions of maize chromosomes. The monomers that make up these repeats may differ by a single nucleotide difference (Ananiev, Phillips, & Rines, 1998).

Tandem arrays of the CentC satellite DNA are frequently interrupted by other DNA sequences, usually corresponding to retrotransposons. The restriction enzyme HaeIII has a four nucleotide cut site (GG|CC) that is not found within the CentC DNA sequence, and therefore does not cut within the satellite repeat. It is possible to isolate large stretches of tandem CentC repeats flanked by short regions of non-CentC DNA by digesting high molecular weight maize genomic DNA with HaeIII. These sequence fragments will be referred to as CentC islands.

The amount of CentC in each maize centromere varies significantly. Variation in CentC quantity between chromosomes of a single inbred can be visualized by FISH. In addition, variation in the amount of CentC in a specific chromosome can also be seen between inbred lines (Kato, Lamb, & Birchler, 2004).

For example the amount of CentC found in centromere 1 of B73 is qualitatively greater than centromere 2 as seen in the FISH picture. Variation in CentC amount between inbreds can be seen in centromere 1 of B73 and Mo17 (Kato, Lamb, & Birchler, 2004).

The recently sequenced genome sequence of maize inbred B73 contains about 54% of the maize genome’s CentC content as determined by comparisons to B73 whole genome shotgun data (Schnable et al., 2009). This number was then added to by Presting Lab. They draft sequenced 101 CentC containing bacteria artificial chromosomes (BACs) and anchored them to both genetic and physical maps. In total, CentC makes up about 0.17% of the 2,300,000,000 bp genome (Schnable et al., 2009).

Wolfgruber et al. (2009) anchored centromeric B73 BACs to their respective chromosomes using repeat junction and transposon display methods and were able to map the positions for all ten centromeres. They were also able to produce physical maps for entire span of centromeres 2 and 5. They observed that the CentC content of centromere 5 was higher than that of centromere 2. CentC blocks of 196 kb and 192 kb span centromeres 2 and 5 respectively.
Although the CentC content varies between the two centromeres, they both contain very little in comparison to the other eight centromeres. In centromere 2, there is a central region of CentC that is flanked by centromeric retrotransposons in maize (CRM). Centromere 5 differs in that the largest block of CentC is skewed to the right of the centromere and a small amount of CentC is present outside of the functional centromere (Wolfgruber et al., 2009).

Nagaki, Song, et al. (2003) constructed a Mo17 BAC library and screened it using plasmid clone probes for CentA and CentC. BAC 16H10 contained four CentC tracts and was sequenced to more than 10X coverage. Seven CRM elements were also found in this BAC. BAC15C5 contained three CentC tracts and a total of 15 retrotransposons including two complete cinful elements and one complete Zeon1 element. FISH results of subclones of this BAC reveal that the insert sequence is made mostly of degenerated retrotransposons.

1.3 Centromeric Retrotransposons (CR)

A transposable element is also found in the centromeres of grasses. The centromeric retrotransposon (CR) is part of the Ty3-gypsy retrotransposon family in grasses (Presting, Malysheva, Fuchs, & Schubert, 1998) and part of the “CRM” clade within the chromoviruses (Gorinsek, Gubensek, & Kordis, 2004). Retrotransposons transpose by the use of reverse transcription of an RNA intermediate. The structure of a CR consists of a 5’ untranslated region (UTR) and an open reading frame (ORF) flanked by long terminal repeats (LTRs).

The ORF of an autonomous CR element encodes a polyprotein that has gag, protease, reverse transcriptase, RNase H, and integrase domains, all of which are required for retrotransposition (Miller, Dong, Jackson, Song, & Jiang, 1998). Non-autonomous CR elements are deficient in one or more of the polyprotein domains and in all probability depend on the autonomous CR elements for retrotransposition (Langdon, 2000).

A single ancient family of retrotransposons is believed to be the source of all CRs in the family Poaceae. Within Poaceae, the CRs have evolved species-specifically (Langdon, 2000).

1.3.1 Centromeric retrotransposon of maize (CRM)

The CR family in maize is named CRM, centromeric retrotransposon in maize (Zhong, Marshall, Topp, Mroczek, & Kato, 2002). The elements that have been identified so far in maize are the autonomous CRM1, CRM2, CRM3, CRM4, and the non-autonomous CentA. CentA was the first CR element found in maize. It was identified in cultivar Seneca60 genomic DNA by using maize centromere-associated sequence (MCS1) as a probe (Ananiev, Phillips, & Rines, 1998). Autonomous CRM1 and CRM2 elements were then discovered (Zhong et al. 2002, Nagaki, Song, et al. 2003). A more careful search for additional elements made possible by the emerging maize genome sequence revealed CRM3 and CRM4 (Sharma & Presting, 2008).

CRM1, CRM2, and CentA were found to be all enriched in centromeres by FISH (Nagaki, Song, et al., 2003). The CRM1 and CRM2 elements were found to be inserted into blocks of the CentC tandem repeats in BAC ZM16H10 which contains centromeric sequence (Nagaki, Song, et al., 2003). The association of CRM elements with CentC was assessed by identifying the percentage of each full length CRM element that was located in the same BAC as CentC repeats. Of all the full length CRM1, CRM2, CRM3, and CentA elements, 30%, 40%, 50% and 71% were found to be in BACs containing CentC respectively (Sharma & Presting, 2008).

Maize CENH3 is a centromere-specific histone H3 variant (see below). It interacts with CentC and CRM as seen in by the enrichment observed in ChIP reads obtained from chromatin immunoprecipitated with CENH3. In situ localization of CRM shows its location at the centromere bordering or inserted into a CentC array (Zhong, Marshall, Topp, Mroczek, & Kato, 2002). Physical maps of maize centromeres show that CRM1 and CRM2 are present in very small quantities on most chromosome arms (Wolfgruber et al., 2009).

1.4 CENH3 centromeric proteins

Also found in the centromere are the protein variants centromere-specific histone H3 (CENH3). CENH3 is a variant of the histone H3 which has a divergent N-terminal tail. Jiang et al. defined the boundary of the centromere as DNA sequences that interact with CENH3 (Jiang, 2003).
The human CENH3 protein, named CENP-A, was the first CENH3 protein identified. CENP-A was found to be present only in functional centromeres and to be an essential feature of the kinetochore. The observation that CENP-A could be localized to neocentromeres lacking the human alpha satellite repeat and not be localized to inactive centromeres that contained the alpha satellite repeat, proved that DNA sequence alone was not what associated CENP-A with the centromeres (Warburton et al., 1997).

The similar size of CENH3-binding domains in different organisms suggests a minimum required size for the functional centromere: 500 kilobases (kb) in the human Y chromosome (Tyler-Smith et al., 1993); 500 kb in Arabidopsis (Murata, Yokota, Shibata, & Kashihara, 2008); about 750 kb in chromosome 8 in rice (Nagaki et al., 2004); in maize it is 0.4 to about 2 megabases (Mb) (Wolfguber et al., 2009).

CENH3 is rapidly and adaptively evolving and may serve as a type of adaptor that can relate the rapidly evolving centromeric satellite repeats to the well-conserved kinetochore apparatus.

1.5 Maize evolution and genome size

50 million years ago (Mya), the ancestor of rice and the progenitor of maize and sorghum diverged from each other. The two progenitor genomes of maize and sorghum are estimated to have diverged from each other about 11.9 Mya. The maize genome arose from the two progenitor genomes through a segmental allotetraploid event about 5 Mya (Swigonová et al. 2004).

Most single cross maize hybrids grown for commercial purposes are produced by using an inbred from each of the “Iowa Stiff Stalk Synthetic” (BSSS) and “Lancaster” breeding programs. The inbred B73 is an example of BSSS, while Mo17 is an example of “Lancaster” (Stuber, 1995). B73 and Mo17 were used as parental lines in major studies on hybrid vigor and genotype-by-environment interaction (Stuber, Lincoln, Wolff, Helenijarin, & Lander, 1992).

Maize’s 2.3 gigabase (Gb) genome makes it more representative of higher plant genome organization than rice (0.4 Gb) and Arabidopsis (0.14 Gb). It is almost as large as the human genome (2.8 Gb) and within the range of other mammals (1.4-3.7 Gb). Rabinowicz and Bennetzen (2006) suggests that the maize genome could serve as an intermediary between species with compact genomes with those of flowering plants (average size of 5.6 Gb) and important crops such as wheat (17 Gb) and barley (5 Gb).

1.6 Artificial chromosomes

Centromeric satellite DNA has the ability to form artificial chromosomes. Human artificial chromosomes (HAC) are used to investigate human chromosome function and potentially can be used to introduce large fragments of DNA into cells. They can be created by introducing telomere repeats and selectable markers into a yeast artificial chromosome containing just human centromeric satellite repeats. It was found to efficiently form human artificial chromosomes that segregated accurately and bound CENH3 variant proteins CENP-B, CENP-C, and CENP-E (Ikeno et al., 1998). In another study, a HAC was formed within a cell upon the introduction of a synthetic modified alpha satellite DNA (Ohzeki, Nakano, Okada, & Masumoto, 2002). The minimum amount of alpha satellite DNA that supported the formation of stable HACs was about 30 kb of repeat. Efficiency improved when the length of the repeat was increased to 50 to 70 kb in size (Okamoto, Nakano, Ohzeki, Larionov, & Masumoto, 2007). This signifies that the centromeric satellite DNA alone is sufficient to form a functional centromere.

Plant artificial chromosomes can be generated using techniques similar to those that were used to create the HAC. An unstable maize artificial chromosome (MAC) was found in oat-maize radiation hybrids (Riera-Lizarazu, Vales, Ananiev, Rines, & Phillips, 2000).

In another instance, circular vectors containing centromeric sequence and selectable markers were created in vitro, and then introduced into maize. These vectors were then able to be inherited from the parent to the next generation (Carlson et al., 2007). The results of these experiments were later questioned, and required confirmation through additional experiments.

A MAC was produced based on the fundamentals that artificial chromosomes relied on DNA fragments that would be capable of conveying centromeric function and the ability of the
plant cell to decorate naked DNA with the centromere-specific proteins needed to form the kinetochore. The minichromosome produced resulted from a vector designed with two inverted CentC arrays separated by CRM1 and with CentA and CRM interspersed within the arrays. The inverted array orientation was proposed based on observations of centromeres from other organisms exhibiting the same organization. The inclusion of the centromeric elements CRM, CRM1, CentA and CentC were proposed from FISH and immunostaining results that showed all four of these elements being present in maize centromeric regions (Ananiev et al., 2009).

From their data most of the minichromosomes they were able to create were produced from chromosome breakages, not a formation of a centromere de novo. A few showed promise from being formed de novo, but the authors were not able to confirm this fully.

Ananiev et al. (2009) also observed that the minichromosomes with the lowest stability were the smallest in size and the larger minichromosomes were much more stable. From this observation they concluded that a stable minichromosome is about 10% the size of the smallest native chromosome. Further characterization of the functional centromere and the mechanism behind it is needed to better develop artificial chromosomes in maize.

1.7 Conclusion

In humans, Arabidopsis, rice, and maize the centromere is located within or near long arrays of satellite DNA. Variation in the amount of centromeric satellite DNA in each chromosome and variation in the sequence of the monomers that make up the array makes determining exactly what part of this sequence lends itself to the centromere’s function very difficult. Due to its repetitive nature, it is a region that is very difficult to sequence and is usually passed over in sequencing projects.

Shedding light on the structure of centromeres in maize will contribute to the knowledge known about centromere function as a whole. Much is not known about the mechanism behind satellite evolution. This thesis will explore the evolution of CentC satellite repeats in corn by describing the arrangements of CentC between inbreds and analyzing the precise CentC sequence of a centromere-derived BAC.

1.8 Hypothesis

The FISH pictures of chromosomes from different inbred lines (Kato, Lamb, & Birchler, 2004) already illustrate that the total amount of CentC varies between centromeres from different chromosomes of a single inbred and also between inbred lines. However, we cannot see from these FISH pictures how the CentC sequences differ. The reason that the CentC sequence of the same centromere differs between inbreds may be that a single large deletion in one and not the other, with the rest of the sequence between the two inbreds being identical. The scale of the differences between sequences on the same centromere from two inbreds may vary from single nucleotides to insertions and deletions of whole monomers and retrotransposon elements. Finally, it is not known if there are HORs in the CentC sequence. The experiments detailed in this thesis will reveal how CentC islands vary between inbreds in these regards.

Based on this, my hypothesis is that CentC islands found in the centromeric regions of maize inbred Mo17 differ from their B73 counterpart.

1.8.1 Objectives

In order to test my hypothesis, a I will have to meet a few objectives.

**Objective 1** Isolate CentC islands from Mo17 genomic DNA.

Mo17 genomic DNA is needed in order to compare two inbreds. The sequence for the Mo17 genome has not been published yet, so obtaining DNA by conventional means is necessary. While DNA isolation is a standard protocol, I will have to isolate high quality high molecular weight DNA in order to obtain DNA that is least sheared in order to keep the CentC islands intact.

**Objective 2**
Find the corresponding B73 CentC island via BLAST and, if necessary, isolate and subclone it from BAC DNA.

By taking the Mo17 CentC islands that were found and finding their B73 counterpart, I will have the sequence of the same CentC island in two inbreds to compare with each other. It may be necessary to isolate and subclone the B73 CentC island from BAC DNA in order to have the same high quality sequence as the one generated from Mo17 DNA.

Objective 3

Compare the orthologous Mo17 and B73 CentC islands with respect to repeat arrangement, nucleic acid composition, the presence of monomeric or HORs, and the B73 CentC island’s location within and outside the centromere.

In depth sequence analysis will be needed in order to see the differences between the two inbreds. A CentC monomer can differ by only a nucleotide, so a comparison of nucleic acid composition is necessary. The presence monomeric or HORs and their location inside and outside the centromere will be noted in order to further characterize these islands. By comparing the two orthologous islands, I hope to find insertions or deletions that may help determine their relative ages.
CHAPTER 2. METHODS

2.1 Isolating genomic Mo17 DNA

Genomic DNA was extracted from very young leaf tissue of maize inbred Mo17 and B73 in two ways. Procedure 1 (M.L1) involved the use of the Qiagen DNeasy Maxi spin column kit. This protocol followed the instructions that came with the kit and used tissue from young corn leaves. Procedure 2 (M.L2, L3, and L4) involved grinding young corn tissue with liquid nitrogen followed by incubation with a DNA extraction buffer, a chloroform extraction, and finally an ethanol precipitation. All DNA was stored in TE buffer at 4°C for future procedures.

2.2 HaeIII digest, end repair, and gel extraction

The genomic DNA was then digested overnight at 37°C with the restriction enzyme HaeIII, which cuts GG|CC. After the HaeIII digest, the DNA was concentrated by precipitation and end-repaired using the Lucigen DNATerminator End Repair Kit. By end repairing the DNA at this step, the genomic DNA sheared during the DNA extraction process results in blunt-ended fragments which could be cloned into the pJAZZ vector.

After the end repair, the DNA was run through a 0.6% agarose gel to remove end repair buffers which could interfere with the cloning process and to isolate the band size of interest to clone. This gel used SyBr safe and blue light to visualize the DNA (M.L1, L2, L3, and L4).

In initial experiments (M.L1 and L2), silica filter and dialysis membrane were used. After visualizing the gel, strips of dialysis membrane were inserted into the gel before and after the region of interest based on the ladder. In this case that is at 6 kb and right under the well of each lane. A strip of silica filter was placed directly below the dialysis membrane strip inserted near the well. The gel was then rotated 180° and allowed to run so that the DNA present in the area could run into the silica filter. The dialysis membrane near the well would block any DNA from running through the silica filter and being lost, and the dialysis membrane located at 6kb would prevent smaller fragments from running into the silica filter.

The silica filter and membrane were centrifuged and the liquid that had spun off was then purified using the Qiagen polymerase chain reaction (PCR) purification kit (M.L1) or precipitated using sodium acetate and ethanol to concentrate the DNA (M.L2).

In subsequent attempts (M.L3 and L4), the Epoch gel extraction kit was used. This was done due to limited supply of dialysis membrane and the gel extraction kit being readily available in the lab. Following the protocol included with the kit, an area corresponding to the 6 kb region up to the well was excised from the gel with a razor with minimal ultra violet (UV) light exposure. The gel slice was then column purified and eluted with water in preparation for ligation.

2.3 Concentration, cloning, and CentC screening

The gel extracted DNA was then ethanol precipitated to concentrate the DNA and then ligated into the vector pJAZZ as per the instructions in the BigEasy v2.0 Linear Cloning Kit manual. 2 µl of the ligation reaction was used in the transformation protocol that is included in the same manual. 150 µl of the incubated transformation reaction were plated on six YT-Agar plates with kanamycin, 5-bromo-4-chloro-indolyl-β-D-galactopyranoside (X-gal), and Isopropyl β-D-1-thiogalactopyranoside (IPTG).

White colonies were picked into water on 96-well PCR plates and re-streaked onto YT-Agar plates as well. A colony PCR using vector primers SL1 (5' - CAGTCCAGTTCAGCTGGAGTC-3'), NZReC (5'-AAATGTCAGTTAATCAGTTCTC-3'), and the CentC primers (CentC_F: 5'-TCC AAAACTCATGTT TGGA-3'; CentC_R: 5'-GTGGATGCGCATGTTGCG-3') was conducted with each plate. The amplification of CentC monomers and dimers were analyzed on a 2% agarose gel (Fig. 3).

In previous experiments (M.L1, L2, L3.T1.1, L3.T1.2, L4), the colony PCR used only the CentC forward and reverse primers. Due to contamination of CentC being present in PCR buffers over time, the PCR reaction was modified to use vector primers plus one CentC primer.
2.4 Insert size determination of CentC positive clones and end sequencing

All CentC positive clones were grown in liquid terrific broth (TB) medium with kanamycin and arabinose induction solution overnight at 250 revolutions per minute (rpm) and 37°C. A portion of the culture was stored as a glycerol stock at -80°C. The remaining amount was purified using the Qiagen Miniprep kit, which uses an alkaline lysis based process to harvest the plasmid.

To isolate the CentC positive inserts from the vector arms, a NotI digest at 37°C for at least one hour was required. The NotI enzyme cuts off the two vector arms, producing three bands when visualizing on a 0.6% agarose gel.

The clones that did not produce a third band were subsequently digested using HaeIII (Fig. 6). A HaeIII digest was done in to see if any insert bands were shadowed by a vector arm. The HaeIII should not cut into CentC, so the part of the insert that is CentC would be the largest band, while the vector arms would be digested into smaller fragments. This digest was also visualized on a 0.6% agarose gel.

After insert size was determined, the sequences were sent to Pacific Biosciences Research Center Biotech Core (Honolulu, HI 96822) for sequencing using vector primers SL1 and NZRevC supplied by the Lucigen BigEasy cloning kit.

2.5 Subcloning

There are three larger CentC positive clones (M.L3.T1.1.G3, M.L3.T1.2.E1, and M.L3.T4.2.D3) and two smaller CentC positive clones (M.L3.T1.2.G11 and M.L3.T4.3.H1). Two separate protocols were used to subclone the larger and smaller islands.

2.5.1 Large fragments

The larger clones were grown in large batches of liquid TB medium and plasmid preps were done. They were digested with NotI to release the CentC island insert from the vector arms. The digest was run on a 0.6% agarose gel and the insert band extracted with a clean razor blade and purified according to the instructions included in the MN gel extraction kit.

After being extracted, the DNA obtained from the excised gel was sonicated for about 20 seconds. After shearing, the DNA was precipitated and end repaired to prepare for cloning. It was then run on a 1% agarose gel to get rid of end repair buffers and to extract a gel region of 1 kb to 5 kb of the sheared DNA.

The gel extract of sheared M.L3.T1.1.G3 selected for the region representing DNA fragments 1 to 8 kb in size (Fig. 1). The gel extract of sonicated M.L3.T1.2.E1 and M.L3.T4.2.D3 selected for the 500 bp to 3 kb region of the gel.

Figure 1: Agarose gel analysis and gel extraction of 20090227 sheared and end repaired DNA of M.L3.T1.1.G3. (A) Before gel extraction and (B) after gel extraction of the 1 to 8 kb sheared and end repaired M.L3.T1.1G3 insert. A piece of gel corresponding to 1 kb to 8 kb in size was excised from the gel. M.L3.T1.1.G3 uncut plasmid “1”, M.L3.T1.1.G3 sheared end repaired DNA “2”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

19
After gel extraction, the DNA was concentrated by ethanol precipitation, ligated into the pJAZZ vector, electroporated, and plated onto YT-KXI (kanamycin, IPTG, X-gal) plates. Colonies were picked into 100 microliter (µl) of water on 96-well PCR plates, as well as restreaked onto YT-KXI plates. A colony PCR reaction using vector primers SL1 and NZRevC and the CentC-F primer was done using the picked colony plates. The results were visualized on a 1% agarose gel.

All CentC positive clones were grown in TB medium with kanamycin and 1000X arabinose induction solution overnight at 250 rpm and 37°C. A portion of the culture was stored as a glycerol stock at -80°C. The remaining amount was purified using kits based on the alkaline lysis process to harvest the plasmid. The templates were sent to Pacific Biosciences Research Center Biotech Core (Honolulu, HI 96822) for sequencing using vector primers SL1 and NZRevC supplied by Lucigen BigEasy v2.0 Linear Cloning Kit.

Raw sequence data was formatted into .b1 and .g1 formats and input into Consed. Discrepancies were resolved by manual editing. The assembled consensus sequence was then separated into individual monomers and input into BLASTn HTGS.

If the individual monomer BLASTs had a prevalent top hit, the Mo17 CentC island would have a B73 counterpart and the BAC would be isolated and subcloned if necessary. If the individual BLASTs did not have the same top hit, the Mo17 CentC island would be considered to not have a B73 counterpart.

2.5.2 Small fragments

The smaller clones were grown up in large batches of liquid TB medium and plasmid preps were done. They were digested with HaeIII to release the insert from the vector arms. The digest was run on a 0.6% agarose gel and the insert band extracted with a clean razor blade and purified according to instructions in the MN gel extraction kit.

Clones M.L3.T1.2.G11 and M.L3.T4.3.H1 were too small to effectively subclone by sonication so an alternate method was used. A mock consensus sequence was created by taking the clone end reads and situating them on either end of a string of N ambiguities. The total size of the mock sequence was determined by the HaeIII digest gel of the clone, with the largest fragment corresponding to the insert size. Restriction enzyme cut maps were created in Sequencher using this mock consensus sequence. An enzyme that cut two or three times in the end reads, with at least one cut site being on either end of the string of N’s, was selected. After the insert was digested with the selected restriction enzyme, the fragments were end repaired and run on an agarose gel. Every band was excised individually and purified according to the instructions in the Epoch gel extraction kit. After this gel extraction, the DNA from each band was individually ethanol precipitated to concentrate, ligated into the pJAZZ vector, electroporated, and plated onto YT-KXI plates.

Colonies were picked into 100 µl of water on 96-well PCR plates, as well as restreaked onto YT-KXI plates. A colony PCR reaction using vector primers SL1 and NZRevC were used on the picked colony plates. The results were visualized on a 1% agarose gel.

Five CentC positive clones corresponding to each band size were grown up in liquid TB medium with kanamycin and 1000X arabinose induction solution overnight at 250 rpm and 37°C. A portion of the culture was stored as a glycerol stock in the -80°C. The remaining amount was purified using kits based on the alkaline lysis process to harvest the plasmid. The templates were then sent off to sequencing using vector primers SL1 and NZRevC supplied by Lucigen BigEasy v2.0 Linear Cloning Kit.

Raw sequence data for each band size was input into Sequencher. Ambiguous bases were called and vector sequence was trimmed. The group of subclones for each band size was assembled into fragments of the consensus sequence, with fragment size corresponding to the gel band sizes. They were then aligned to the end reads of the CentC islands, so that the fragments would line up along the end reads and along the specific restriction enzyme cut sites to create a consensus sequence.

The assembled consensus sequence was then separated into individual monomers and input into BLASTn HTGS. If the individual monomer BLASTs had a prevalent top hit, the Mo17 CentC island would have a B73 counterpart and the BAC would be isolated and subcloned if
necessary. If the individual BLASTs did not have the same top hit, the Mo17 CentC island would be considered to not have a B73 counterpart.

2.6 Isolating the B73 counterpart
To compare the Mo17 CentC sequence to its B73 counterpart, a decision was made to subclone the BAC that the B73 counterpart was on to obtain high quality sequence data. If there was more than one BAC that contained the counterpart, the availability of the specific BACs to the lab and the total amount of CentC within the BAC were taken into consideration. If a BAC contained a smaller amount of CentC, it was more likely that the CentC sequence isolated from that BAC DNA would match the corresponding Mo17 sequence and eliminate screening steps.

The BAC was freeze-thawed. A BAC freezing media made of LB (2.5% w/v), dibasic potassium phosphate (36 mM), monobasic potassium phosphate (13 mM), sodium citrate (1.9 mM), ammonium citrate (6.8 mM), and glycerol (4.4% w/v). Magnesium sulfate (40 mM) and 5% chloramphenicol (final concentration 12.5 µg/ml) were added after media was cooled.

5 ml of media was inoculated with 20 µl of thawed culture and 5 µl arabinose induction solution and incubated at 250 rpm, 37°C overnight. The BAC plasmid was isolated using the QIAGEN Large-Construct kit protocol.

BAC DNA was digested with HaeIII, end repaired, gel bands were extracted, ligated into the pJAZZ vector and transformed using the same protocol used to treat the Mo17 counterpart previously mentioned. CentC PCR screening of the BAC clones also followed the methods for the CentC PCR screening of Mo17 clones.

Since the approximate lengths of the B73 CentC islands are known through viewing the gel before gel extraction, clones were selected for sequencing based on it and the known size of its Mo17 counterpart. The subcloning procedure for the specific CentC island located in the BAC follows the protocol outlined for its Mo17 counterpart.

2.7 CentC island analysis and comparison

2.7.1 Mo17 CentC island and B73 counterpart analysis and comparison
If the Mo17 island had a known B73 counterpart, the Mo17 and B73 CentC islands were compared to each other as a whole using Clustal. Nucleotide differences as well as large gaps were noted. Then the islands were separated into individual CentC monomers. A phylogenetic tree was created of the monomers to better see the relationships between them. Like monomers were color coded and rearranged back into their positions on their respective island. Monomeric and HOR patterns were determined based on the patterns created by the color coding of the similar monomers.

2.7.2 Mo17 CentC island with no B73 counterpart analysis and comparison
If the Mo17 island did not have a known B73 counterpart, the CentC island was still separated into monomers and a phylogenetic tree will be created of them. They were color coded and rearranged into their positions on the island. Patterns were determined from the color coded monomers.

2.7.3 Centromeric and noncentromeric CentC island comparison
Lastly, the color patterns of centromeric and noncentromeric B73 islands were compared. This was to see if repeat patterns are present in both centromeric and noncentromeric islands. We do not know whether their location in or outside the centromere affects their monomer organization.

21
CHAPTER 3. RESULTS AND DISCUSSION

3.1 Isolating genomic Mo17 DNA

The results of multiple attempts at each procedure were visualized on a 0.6% agarose gel that contained ethidium bromide and compared to each other (Fig. 2). All lanes show the presence of DNA. All lanes in gel A and lanes 2, 3, 4, 5, and 6 in gel B have a band larger than 10 kb. The genomic DNA in gel B has much more larger fragments of DNA, with some of the very large DNA still present in the wells at the top of the gel.

![Figure 2: Agarose gel analysis of genomic DNA extraction procedures using B73 and Mo17 inbreds.](image)

(A) Genomic DNA extraction procedure 1 with genomic B73 DNA “1 and 2”, genomic Mo17 DNA (using newer kit) “3”, genomic Mo17 DNA (using older kit) “4”, and Tri-Dye 2-log ladder “L”. (B) Genomic DNA extraction procedure 2 with genomic Mo17 genomic DNA (20081206) “1”, Mo17 (1) genomic DNA (20081212) “2”, Mo17 (2) genomic DNA (20081212) “3”, Mo17 (1) genomic DNA (20081216) “4”, Mo17 (2) genomic DNA (20081216) “5”, and TriDye 2-log ladder “L”.

Two batches of genomic DNA extracted using procedure 2, Mo17 (1) 20081212 and Mo17 (2) 20081216 were chosen to move forward in the experiment. Mo17 (2) 20081212 was also of good quality, but it was decided that the two chosen to move forward should be from different DNA extraction dates to ensure that the extraction could be reproduced and useable.

By using the chloroform extraction (procedure 2) instead of the Qiagen DNeasy Maxi kit (procedure 1), shearing of DNA was avoided. The maxiprep kit requires a filter and a column in its protocol, which has a maximum DNA fragment size allowance of around 10kb that is allowed to pass through the filter in the elution step. By not using the kit, we avoided this maximum limit and were able to extract larger fragments of high quality genomic DNA. Due to the use of sodium bicarbonate in the DNA extraction buffer used in procedure 2, surplus buffer should be stored out of direct light and should be discarded after a few days, as its reactivity decreases very quickly with age.

Larger amounts of DNA were extracted using very young ear instead of leaf tissue. This batch of DNA was not discolored compared to leaf tissue DNA extraction that may be colored brown if the leaf were too old, or green if the chloroform treatment was unable to rid the extract of all its chlorophyll. In future experiments where high quality and high molecular weight DNA is required, I would strongly suggest using a very young ear of corn over leaf tissue. If leaf tissue needs to be used, a younger leaf would be better than an older one. If the leaf is too old, DNA yield and quality both decrease.
3.2 HaeIII digest, end repair, and gel extraction

The restriction enzyme HaeIII (GG|CC) was chosen to digest the genomic DNA because it lacks a cutting site in the CentC consensus sequence. We reasoned that following use of HaeIII, the high molecular weight fraction of the treated DNA would be enriched with CentC.

After the HaeIII digest, the DNA was end repaired according to instructions included in the Lucigen DNATerminator End Repair Kit. This was done in order to prepare the DNA for cloning into the pJAZZ vector. Skipping the end repair step, as was done in experiments M.L1 and L2, resulted in a very low number of colonies after transformation and plating.

After the end repair, the DNA was run through a 0.6% agarose gel in order to remove end repair buffers that could interfere with the cloning process and to isolate the band size of interest to clone. Using SyBr safe and blue light was expected to cause less point mutations than the use of UV light and ethidium bromide to visualize the DNA.

Initially a method that required dialysis membrane and silica filter was used (Fig. 3). Initially used this method because I didn’t want to use a column which is used in a lot of kits in my extraction procedure. Cutting the gel and inserting the dialysis membrane and silica filter into the 0.6% gel posed some difficulty and required practice. This process also took twice as long as the method that involved using the kit because the gel had to be run, then flipped and run for additional time in order for the DNA fragments of interest to run into the silica filter.

When the supply of dialysis membrane ran out, the Epoch gel extraction kit was used because it was readily available (Fig. 4). The protocol used by the kit involved the use of a column which would allow a DNA fragment of a maximum size of 10 kb pass through it. Larger fragments would not be able to pass through. While I had initial concerns about this limitation, the cloning kit that I chose to use had a maximum successful clone insert size of about 30 kb, although on average 15 kb fragments have successfully been cloned. With 15 kb being the average size of cloned fragments, a column allowing 10 kb fragments through it was acceptable. Using a kit also helped streamline my protocol and made it easier to replicate because it is a commercially tested kit, while the dialysis membrane and silica filter method involved more practice and technique and would possibly be harder to successfully replicate on the first attempt.
Figure 4: Agarose gel analysis and extraction of HaeIII restriction digest of 20081216 Mo17 (2) using a razor extraction method. (A) Before gel extraction and (B) after extraction of gel fragment from 6kb to well using a razor blade. HaeIII digest “1”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

3.3 Ligation and transformation

The different DNA extraction and gel extraction protocol combinations were ligated and transformed. They produced varying amounts of colonies which could be correlated to the amount of DNA present after the specific DNA extraction and gel extraction protocols used.

20081220 Ligation and transformation (M.L1.T1)

The genomic DNA extracted with the Qiagen DNeasy Maxi kit (procedure 1) was ligated to the pJAZZ vector and transformed to see if using DNA obtained from the kit and gel extraction could produce clones. The gel extraction method using dialysis membrane and filter paper was used. There were no colonies present on the YT-Agar plates. It was determined that using the other DNA extraction protocol would yield better results and all experimentation with the DNA from the Qiagen kit stopped.

20090107 Ligation and transformation (M.L2.T1)

DNA extraction procedure 2 and the gel extraction protocol using dialysis membrane and filter paper were used prior to cloning. There were no blue colonies, which indicate that the vector mixture, which contains digested vectors arms were properly digested. If there were blue colonies, that would mean that the vector was not properly digested and that the lacZ gene is still present in the vector. The white colonies were counted. The noticeably larger white colonies were counted separate from the smaller colonies.

The plates with 20081212 Mo17 (1) had fewer colonies than the 20081216 Mo17 (2) plates. Comparison of the two batches of genomic DNA used in the ligations revealed that 20081216 Mo17 (2) appears to have more DNA present in the lane (Fig. 1).

Table 1: Count of small and large colonies from 20090107 transformation

<table>
<thead>
<tr>
<th>Plate set name</th>
<th>Small colonies</th>
<th>Large colonies</th>
</tr>
</thead>
<tbody>
<tr>
<td>20081212 Mo17(1)</td>
<td>66</td>
<td>79</td>
</tr>
<tr>
<td>20081216 Mo17(2)</td>
<td>2584</td>
<td>848</td>
</tr>
<tr>
<td>pSMC121</td>
<td>1696</td>
<td>560</td>
</tr>
<tr>
<td>Control insert</td>
<td>262</td>
<td>137</td>
</tr>
<tr>
<td>Transformation control</td>
<td>92</td>
<td>66</td>
</tr>
</tbody>
</table>
20090120 Ligation and transformation (M.L3.T1)

There were no blue colonies, indicating that the vector mixture, which contains digested vectors arms were properly digested. The white colonies were counted. The noticeably larger white colonies were counted separate from the smaller colonies.

This time, two gel extractions using the Epoch gel extraction kit were performed. The second extraction was done to see if the second extraction would enrich for larger fragments. This was done by allowing smaller fragments that may have been stuck within the larger fragments of the first extraction to be separated from them during the second run on the gel. Due to less DNA collected during the second extraction, the second gel extract has markedly fewer colonies in both larger and smaller white colonies.

Table 2: Count of small and large colonies from 20090120 transformation

<table>
<thead>
<tr>
<th>Plate set name</th>
<th>Small colonies</th>
<th>Large colonies</th>
</tr>
</thead>
<tbody>
<tr>
<td>20081216 Mo17(2) gel extract 1</td>
<td>60</td>
<td>137</td>
</tr>
<tr>
<td>20081212 Mo17(1) gel extract 2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>20081216 Mo17(2) gel extract 2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Control insert</td>
<td>31</td>
<td>105</td>
</tr>
<tr>
<td>Transformation control</td>
<td>167</td>
<td>69</td>
</tr>
</tbody>
</table>

3.4 CentC PCR screening

White colonies were picked into water on 96-well PCR plates. CentC specific primers as well as vector primers were used in to screen for CentC in PCR. Amplification of CentC monomers and dimers were analyzed on 2% agarose gels (Fig. 5).

In M.L3.T1.1, screening for CentC used only CentC specific primers in PCR. In later experiments the addition of vector primers was used to compensate for any CentC contamination of buffers. A band at 156 bp would be evidence of the CentC primers priming to a single CentC monomer. Multiple bands were expected in lanes that had a CentC positive clone. In PCR, the CentC primers would anneal to multiple sites on the DNA – producing monomers (156 bp), dimers (312 bp), and even trimers (468 bp).

There are very faint bands that correspond to the size of a CentC monomer (156 bp) in lanes A8, B7, B8, C10, D10, E7, F6, F7, G10, and H9. There are more noticeable monomer bands in lanes C7, C8, D8, G3, G7, and H7. The only visible CentC dimer can be seen in lane G3. All the noted clones were selected to move forward in the experiment. The absence of a band in the negative control indicates that there is no contamination of CentC in the PCR reagents.

Figure 5: Agarose gel analysis of 20090123 CentC PCR using primers CentC_F and CentC_R. 100 clones (1 per well) “Rows A to H”, negative control (no DNA) “-”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”. Clones of interest are labeled with their corresponding well number on a 96-well PCR plate.
3.5 Insert size determination and end sequencing

Restriction digests replaced colony PCR in isolating the fragment of interest. Normally, primers could be made to flank either end of the insert fragment and DNA could be amplified exponentially in PCR, which could then be used in subcloning steps. The repetitive nature of CentC makes it undesirable to use in PCR. As seen in the CentC screening step, the use of primers that contain CentC sequence will cause amplification of multiple PCR products, not just of the entire clone. Point mutations caused by PCR are also not desired since each CentC monomer can differ by just one nucleotide. Due to these restraints, large batches of liquid culture were grown up and digested using restriction enzymes and the insert band was isolated using a gel extraction method in lieu of the use of PCR.

To subclone the CentC positive clones as well as determine insert size, a NotI digest was required. The NotI restriction enzyme was used to cut the flanking vector arms from the insert DNA. This would produce a 10 kb and 2.2 kb bands corresponding to the vector arms and a third band corresponding to the insert DNA.

It was discovered that in some instances a vector arm and an insert were so close in size that when the NotI digest was visualized on a gel, they would be on top of each other. This would prove deceiving in that it would appear as if there was no insert fragment, therefore a negative result. In these cases, a HaeIII digest was done with the belief that HaeIII was the restriction enzyme that was used in the initial digest of the genomic DNA; therefore it would not have a cut site present in the insert DNA fragment. It successfully digested the vector arms into smaller fragments, leaving the insert containing CentC intact and would appear as the highest molecular weight band on the gel. This technique allowed us to see insert bands that were previously shadowed by vector arms in the NotI digests.

M.L3.T1.1

The linear vector arms are 10 kb and 2.2 kb in size. Any bands other than those are considered inserts. There are distinct insert bands in M.L3.T1.1.B7, G3, G7, G10, and H7 (Fig. 6). There is a possibility that there could be insert bands hidden underneath the vector bands in the other lanes.

Figure 6: Agarose gel analysis of 20090129 NotI digest of M.L3.T1.1 clones. CentC positive clone “A8 – H9”, left lane for each clone is uncut plasmid, right lane for each clone is plasmid cut with NotI, and NEB Tri-dye 2-log ladder “L”. Vector arms bands at 10 kb and 2.2 kb.
All clones except for G3 were digested with HaeIII (Fig. 7). The clone G3 was not digested with HaeIII, since it was already determined that this clone had an insert of reasonable size. The low molecular weight bands in each lane of the HaeIII digest are the vector arms that were digested with HaeIII. The highest band in A8 and G7 might be inserts that were otherwise hidden by the shorter 2.2 kb vector arm in the NotI digest.

Figure 7: Agarose gel analysis of 20090806 HaeIII digest of M.L3.T1.1 clones. CentC positive clone “A8 – H9”, left lane for each clone is uncut plasmid, right lane for each clone is plasmid cut with HaeIII, and NEB Tri-dye 2-log ladder “L”. The low molecular weight ladder represents the vector arms being digested by HaeIII. Any remaining higher molecular weight band is the insert DNA fragment.

M.L3.T1.2

Additional clones M.L3.T1.2.C4, E1, G11, and H6 were visualized after being digested with NotI (Fig. 8). E1 has a 6 kb insert and H6 has two inserts. The two inserts in H6 could be due to selecting two colonies from the plate instead of just one, or it could be that there is a NotI site present within the insert sequence, which would produce two separate bands. Clones C4 and G11 displayed no insert bands, which could indicate an empty vector, or that the insert band is the same size as a vector arm and they appear as one band instead of two different entities.

When digested with HaeIII, there is a low molecular weight ladder in each lane, which is the vector arm. In E1, G11, and H6 there are higher molecular weight bands, indicating inserts. The insert size for E1 and H6 were already known from the NotI digest. The HaeIII digest of G11 revealed an insert the size of the small vector arm, 2.2 kb. There is no higher molecular weight band in C4, indicating that maybe it is empty vector and that the positive CentC PCR screening was due to contamination.

Figure 8: Agarose gel analysis of 20090310 NotI digest and 20090722 HaeIII digest of M.L3.T1.1 clones. CentC positive clone “C4 – H6” plasmid cut with (A) NotI and (B) HaeIII, and NEB Tri-dye 2-log ladder “L”. Vector arms are represented by 10 kb and 2.2 kb bands in gel A and by a low molecular weight ladder in gel B.
M.L3.T4
There are distinct inserts in M.L3.T4.1.D3 and 2.D3 NotI digests (Fig. 9). The 1.D3 insert is very slightly larger than the 2.2 kb vector arm, while the 2.D3 insert is about 5 kb in size. A 2.2 kb insert is revealed to be present in 3.H1 in the HaeIII digest. The rest of the clones are empty vector and do not produce insert bands in either the NotI digest or the HaeIII digest.

Figure 9: Agarose gel analysis of 20090804 NotI and HaeIII digest of M.L3.T4 clones. CentC positive clone “1.C2 – 4.H1”, left lane for each clone is uncut plasmid, middle lane for each clone is plasmid cut with HaeIII, right lane is plasmid cut with NotI, and NEB Tri-dye 2-log ladder “L”.

The sizes of each potential CentC positive clone for each digest were recorded and the ends of each clone were sequenced to verify for the presence of CentC. Those that had CentC positive BLAST hits (M.L3.T1.1.G3, M.L3.T1.2.E1, M.L3.T1.2.G11, M.L3.T4.2.D3, and M.L3.T4.3.H1) were allowed to progress to the subcloning step.

3.6 Subcloning
Two separate protocols were used to subclone the CentC positive clones based on size. Shearing by sonication was a reasonable way to subclone larger fragments, but the smaller fragments risked being sonicated too much even when being careful. In some cases, the smaller fragments had an unknown sequence region of only 1 or 2 kb.

3.6.1 Large fragments
Due to the downstream treatments of gel extraction, end repair, and sonication, a large amount of DNA was needed to begin with. At each of the mentioned downstream steps a portion of the original batch was lost in the gel, during the purification steps, or sonicated into fragments too small. Normally a PCR could amplify the DNA and synthesize enough to proceed. Because of the repetitive nature of the CentC islands, another method was required. A large amount of liquid culture was grown up and the plasmid was purified by alkaline lysis on a larger scale.

Since the sizes of the larger inserts were determined by HaeIII and NotI digests, there was no concern over column use in the gel extraction kit preventing the DNA fragment of interest to pass through. The column used in the gel extraction kit has a maximum limit of 10 kb, and the largest CentC insert based on a NotI digest gel is about 9 kb. The use of the kit helped streamline the protocol and gave uniformity to this step. The vector arms are of known length, so the extraction of the remaining band was simple (Fig. 10).

Figure 10: Agarose gel analysis and gel extraction of 20100318 NotI digest of M.L3.T4.2.D3. (A) Before gel extraction and (B) after gel extraction of the 5kb 2.D3 insert. M.L3.T4.2.D3 NotI digest “2.D3”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”.

28
After gel extraction of the insert band, it was sheared by sonication then end repaired. The end repair step is crucial because the cloning kit clones blunt ends. If the sheared DNA were to be cloned directly, the outcome would be very few colonies. The buffers used in the end repair protocol interfere with the cloning buffers, so running a gel to remove the end repair buffers before cloning is crucial. By running the sonicated end repaired DNA on a gel also allowed for size selection of the sheared fragments.

Another option to clean up the end repair buffers would be to end repair the sheared DNA as usual, then use the gel extraction kit directly on the sample to purify it without running it on a gel. The issue here was that there wasn’t a step to select for fragment size, so one did not know if the sample was sonicated for too long (i.e. the sheared DNA fragments are too small) and did not allow for fragment size selection. A sample was done this way to test this procedure and it was found that after cloning the colonies all contained very small inserts (less than a centC monomer in length) that was not useful subcloning purposes. It seemed as if somehow during the cloning step the smaller fragments were selected for as opposed to the larger ones. By running the gel, we do not run into this problem.

After gel extracting the sheared end repaired DNA, it was cloned and white colonies were picked for CentC screening in PCR. When run on a 1% agarose gel, CentC positive clones exhibited a ladder (Fig. 11) as a result of the CentC primer and a vector primer amplifying different lengths of CentC monomers. This allowed for easy identification of potential CentC positive subclones. The bands do not necessarily represent CentC monomers as the sheared DNA may start in the middle of a monomer. Annealing of a vector primer to the end of the insert and the CentC primer annealing to the first CentC could produce products that are slightly larger, or shorter than those seen in previous gels where primers were just annealing to whole CentC monomers. The prior use of forward and reverse CentC primers gave less resolution as there was a lot of background.

Figure 11: Agarose gel analysis of 20100408 Insert/CentC PCR using primers SL1, NZRevC, and CentC_F. Sheared subclones of M.L3.T1.2.E1 “Rows A to H, Plates 1 and 2”, PCR negative control (no DNA) “-”, NEB Tri-dye 2-log ladder “L”, and empty lane “X”. Clones of interest are circled red.

Subclones that exhibited a ladder PCR product were grown up in liquid culture. The plasmid was isolated and sequenced using vector primers. The raw sequence data was assembled in Consed. The assembled consensus sequence was separated into monomers and input into BLASTn HTGS in order to search for a B73 counterpart. If majority of the monomers
had an identical top hit, the Mo17 CentC island had a B73 counterpart. If not, it did not have a B73 counterpart to date.

Two clones (M.L3.T1.1.T3 and M.L3.T1.2.E1) came back with B73 counterparts. The other clone (M.L3.T4.2.D3) did not have a B73 counterpart, but since only 50% of the CentC in the B73 genome has been sequenced, there may be a possibility that a B73 counterpart will be available at a later date.

3.6.2 Small fragments

By using the subcloning procedure outlined in section 2.5.2, the CentC islands were fragmented into subclones of known length; each of which could be sequenced fully without having to subclone further due to their size being maximum 1 kb. With the known restriction sites and approximated subclone sizes, it could easily be reassembled in Sequencher.

The restriction enzyme BsaWI created a cut map for Clone M.L3.T1.2.G11 that cut once on the 5’ side and twice on the 3’ side of the unknown middle region of the mock sequence. When digested with BsaWI, only three fragments were seen on the gel (Fig. 12) were named “small”, “medium”, and “large” due to their relative sizes. Since BsaWI cut three times into the mock sequence, four bands were expected on the gel. The smallest fragment recreated by the mock sequence cut map was about 103 bp. Since the lowest band on the ladder used in the gel is at 100 bp, the 103 bp fragment may have been too faint to see and was not extracted because of that. The need to extract this 103 bp fragment was not necessary as it is the section at the very end of the clone and is can be covered by end sequencing. The visualized bands were each gel extracted, end repaired, and cloned separately. The clones were screened for CentC and five of each was sequenced.

During sequence analysis, it was found that there was another cut site that wasn’t seen by creating a cut map of the end reads. This was discovered when the sequences for subclone “small” did not assemble together. Two distinct sequences of about the same length and one assembled to the 3’ end sequence of the clone. The sequences were treated separate fragments, with the “small” batch of sequences renamed into “small1” for the short sequence closer to the 5’ end and “small2” for the sequence that assembled to part of the 3’ end sequence.

After each group of sequences were assembled together, the consensus sequence of each and the clone end sequences were used to assemble the clone consensus sequence. This was done by matching up the cut sites and anchoring the fragment consensus sequences to the clone end sequences if possible.

The cuts between the medium-small1 and small1-large fragments were not ACCGG|A as expected when digesting with BsaWI, but were ACC|GGA instead. This could be due to the vector sequence at the insert site being either CCC or GGG at either end. As a precaution, the vector sequence was trimmed entirely, so any C or G that may have belonged to the insert at this junction might have been included in the vector trim if the vector sequence did not have sufficient sequence.
Figure 12: Overview of the M.L3.T1.2.G11 subcloning method. (1) Haelll digest of genomic Mo17 DNA, (2) gel extraction of Haelll digest, (3) BsaWI cut map of end sequences, (4) gel extraction of BsaWI digest, (5) forward and reverse sequencing of subclone fragments, and (6) the final assembly of the subclone fragments, including size of each fragment.

Clone M.L3.T4.3.H1 was fragmented into three sections according to the BsiWi cut map of its end sequences (Fig. 13). The final assembly was produced from the joining of a small, medium, and large fragment. The "small" fragment gained 4 bp due to the end repair required in order to use the blunt end pJAZZ vector. The "large" fragment lost about 100 bp. This was the fragment that was estimated in size initially from the Haelll digest gel. The "medium" fragment gained 9 bp, of which a 4 bp gain can be accounted for by the end repair.

Figure 13: Overview of the M.L3.T4.3.H1 subcloning method. (1) Haelll digest of genomic Mo17 DNA, (2) gel extraction of Haelll digest, (3) BsiWi cut map of end sequences, (4) gel extraction of BsiWi digest, (5) forward and reverse sequencing of subclone fragments, and (6) the final assembly of the subclone fragments, including size of each fragment.
The consensus sequences of both M.L3.T1.2.G11 and M.I3.T4.3.H1 were separated into monomers and analyzed by using BLASTn HTGS as seen the larger clone protocol to determine if the Mo17 CentC island has a B73 counterpart. Both of the smaller Mo17 clones do not have a B73 counterpart at this time.

3.7 Isolated and subcloned Mo17 CentC islands

Table 3 shows all of the Mo17 CentC islands that were isolated and subcloned using the methods outlined in Section 2.1 to 2.5 as well as any corresponding B73 counterparts which were isolated and subcloned using methods mentioned in Section 2.6. This completes objectives 1 and 2 that were stated in 1.8.1.

<table>
<thead>
<tr>
<th>Clone</th>
<th>Official name</th>
<th>Mo17 clone size (bp)</th>
<th>Centromere/ noncentromere</th>
<th>ChIP/ non-ChIP</th>
<th>B73 BAC</th>
<th>B73 clone size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M.L3.T1.1.G3</td>
<td>7914</td>
<td>Centromere 10</td>
<td>ChIP</td>
<td>b0410L22</td>
<td>B.L1.T1.H8</td>
</tr>
<tr>
<td>2</td>
<td>M.L3.T1.2.E1</td>
<td>5775</td>
<td>Noncentromere 1</td>
<td>Non-ChIP</td>
<td>CH201-433P11</td>
<td>B.L3.T1.D8</td>
</tr>
<tr>
<td>3</td>
<td>M.L3.T1.2.G11</td>
<td>2078</td>
<td>No match</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>4</td>
<td>M.L3.T4.2.D3</td>
<td>4964</td>
<td>No match</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>5</td>
<td>M.L3.T4.3.H1</td>
<td>2092</td>
<td>No match</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

22,823 bp worth of Mo17 CentC island sequences were sequenced in total. This comes out to be about 0.6% of all the CentC in the genome. In all, five CentC islands were found. Three of them are about 5 kb or larger and two are around 2 kb in size. Clone 1 has a B73 counterpart present in Centromere 10. The B73 clone was isolated and subcloned from the corresponding BAC b0410L22 and was named B.L1.T1.H8. The B73 counterpart for Clone 1 is 652 bp longer than the Mo17 CentC island. Clone 2 has a B73 counterpart in a pericentromeric region on Chromosome 1. A B73 counterpart clone was generated from the corresponding BAC CH201-433P11 and named B.L3.T1.D8 which is 1,708 bp shorter than Clone 2. The other three Mo17 clones did not have high quality matches to the B73 genome.

A BLASTn was done on 20110323 using the HTGS database, with no organism selected. By not selecting an organism, it broadened the search for B73 BACs that were mislabeled under a different organism other than Zea mays. This search was done to update the table and see if any recently sequenced BACs matched any Mo17 CentC island.

The BLAST results for Clone 3 included a hit that covered the junction between a gene and the start of the CentC island. The number of mismatches present in the corresponding sequence though made it clear that this was not a true match to the B73 genome. The highest BLAST hit for Clone 5 covered 97% of the query and had a max identity of 94%. This isn’t enough to positively identify it as a B73 counterpart.

The decision to subclone the B73 counterpart from the BAC instead of using the BAC sequence already in GenBank was made because discrepancies in the sequence of that BAC in GenBank were discovered. This will be discussed later in section 3.8. By subcloning the BAC using the methods outlined in Chapter 2, the B73 counterpart would have the same high quality sequence that was generated for the Mo17 clones.
3.7.1Mo17 clones that have a B73 counterpart

CentC island comparison and monomer analysis between the Mo17 Clones 1 and 2 and their B73 counterparts were done as outlined in section 2.7.1.

3.7.1.1Analysis of Clone 1 (M.L3.T1.1.G3)

![Figure 14: Junction viewer image of Clone 1.](image)

CentC monomers are represented by green arrows and a CRM1 element is represented by a blue block.

Clone 1 was 7,914 bp in length. The consensus sequence for Clone 1 was input into Junction Viewer software (Wolfrubner 2010) and an image was generated (Fig. 14). The CentC island starts 357 bp into the sequence and stop at 7,719 bp. The 3’ end of the CentC sequence is followed by a CRM1 element. The Clone 1 sequence was then input into BLAST against the maize genome. Its B73 counterpart is present in BAC b0410L22 from centromere 10. It is unknown whether the Mo17 sequence was from chromosome 10 of the Mo17 genome as the final sequencing of this genome is still pending.

A junction viewer image was created for the B73 counterpart (Fig.15). The B73 counterpart, B.L1.T1.H8, is 8,566 bp in length and does not include a CRM1 element. The CentC islands start at 337 bp. There is a portion of a cinful element at the 3’ end of the B73 counterpart. Because of the 652 bp difference in size and the presence of a CRM1 element at the 3’ end of the Mo17 sequence and the presence of a cinful element at the 3’ end of the B73 sequence, a more in depth look at each individual CentC in both sequences was needed.

![Figure 15: Junction viewer image of B73 BAC b0410L22 counterpart H8.](image)

CentC monomers are represented by green arrows.
Phylogenetic trees were generated from the CentC monomers of both Clone 1 and its B73 counterpart and like monomers were color coded (Fig. 16). The monomers were then arranged back into their positions on their respective islands (Fig. 17).

**Figure 16**: Phylogenetic tree of individual CentC monomers from both Clone 1 consensus sequence and its B73 counterpart consensus sequence. Clone1 CentC monomers “G3-#” and B73 counterpart CentCs “H8-#”.
CentC monomers “colored arrows pointing to the right”, CRM element “colored arrow pointed to the left”, nucleotide differences “colored vertical lines”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

When the Mo17 and B73 CentC islands were initially aligned to each other, it was found that Mo17 was slightly shorter and had a CRM element at its 3’ end. It was proposed that the Mo17 CentC sequence was interrupted by a CRM element where the B73 counterpart sequence was not. By creating a unique primer that mirrored the CRM-CentC junction in Mo17 and a unique primer that corresponded to the 3’ CentC-cinful junction of the B73 counterpart, a PCR was done to amplify the small section of CentC located on the other side of the CRM element from Mo17 genomic DNA.

Because the PCR would produce multiple products due to the CentCs annealing to each other, potentially causing a problem during sequencing, a gel was run and the band that corresponded to the length difference between the Mo17 and B73 sequences was excised and purified using a gel extraction kit. This was then sequenced and added to the Mo17 CentC island sequence, with the CRM element’s location noted. This extended the Mo17 sequence to correspond to the entire length of the B73 island.

This is likely an insertion of a CRM element into the Mo17 CentC island as opposed to a deletion in the B73 counterpart. Since the pattern of CentC monomers matchup between the B73 and Mo17 sequence on either ends of the insertion and it is not disrupted other than the presence of the CRM, it is more likely that the CRM inserted within this array. If the CRM were deleted from the B73 clone instead of inserted into the Mo17 clone, there could be evidence of it by a disruption in the monomer pattern.

When the two consensus sequences were aligned, the B73 clone had an additional CentC where there was no corresponding Mo17 CentC. This is likely a deletion of a CentC monomer in the Mo17 sequence. Upon looking at the phylogenetic tree of the individual monomers, monomer M12 is related to monomer B12. If this is taken into account, one can
surmise that monomer B13 does not have a corresponding Mo17 monomer. The deletion of an entire monomer from Mo17 is more likely than an insertion of a monomer in B73. This is because the monomer is part of a HOR which is seen again later in the sequence at B39. The insertion of a monomer into the B73 sequence creating a HOR which is seen again later in the sequence is not likely.

The CentC monomers exhibit a few HORs within the array. There is a five monomer pattern that repeats twice (green arrow), a four monomer pattern that repeats twice (purple arrow), and another five monomer pattern that repeats four times (gold arrow). The gold arrow HOR has more nucleotide differences between Mo17 and B73, potentially making it an older HOR than that of the purple and blue arrows. Unlike other clones, where there is only one HOR found in the island – whether continuous or not, this clone has a few different HORs present in both Mo17 and its B73 counterpart. None of these HORs are seen more than twice in this specific island, in either Mo17 or B73.

There are a total of 39 nucleotide differences between Clone 1 and its B73 counterpart. This does not include the CentC monomer deletion in B73 or the CRM insertion in Mo17. Of those differences, four are nucleotide insertions or deletions, 20 are transitions and 15 are transversions, and represent one point mutation every 218 nucleotides, or about 1.4 CentC monomers.

3.7.1.2 Analysis of Clone 2 (M.L3.T1.2.E1)

Clone 2 is 5,775 bp in length. The consensus sequence for Clone 2 was input into junction viewer and an image was generated (Fig. 18). The clone is made up entirely of CentC monomers. This sequence was then input into BLAST against the maize genome. Its B73 counterpart is present in BAC CH201-433P11 from a pericentromeric region on chromosome 1. It is unknown as to whether the Mo17 sequence will be from the same place in the Mo17 genome as the finishing of the sequencing of this genome is still pending.

Clone 2 is the second Mo17 clone found to have a B73 counterpart. When the ends of the clone were used in BLASTn, it was found to match sequence found in B73 BAC CH201-433P11. Learning from the issues experienced when subcloning the B73 counterpart of Clone 1, it was assumed that part or all of the CentC sequence found in BAC CH201-433P11 was misassembled. The B73 counterpart was subcloned from the BAC using methods previously mentioned.

Figure 18: Junction viewer image of Clone 2. CentC monomers are represented by green arrows.

Figure 19: Junction viewer image of the B73 counterpart of Clone 2. CentC monomers are represented by green arrows.
The B73 counterpart to Clone 2, B.L3.T1.D8, is 4,067 bp in length and includes a part of a xilon element on the 5’ end of the sequence. The CentC island starts at 340 bp and continues for the rest of the sequence. A junction viewer image was created for the B73 counterpart clone (Fig. 19). Because of the 1,708 bp difference in sizes and the presence of a xilon element in the B73 sequence but not the Mo17 sequence, a more in depth look at each individual CentC in all sequences was needed.

The individual CentC monomers from each consensus sequence were aligned, a phylogenic tree was generated to see the relationships between CentC monomers, and like monomers were color coded together (Fig. 20). The consensus sequences were then reassembled with colored coded arrows depicting the individual monomers in the correct sequence arrangement (Fig. 21). The Mo17 and B73 consensus sequences were then compared to each other again. Patterns were observed and nucleotide differences between each corresponding CentC monomer were noted.

**Figure 20:** Phylogenetic tree of individual CentC monomers from both Clone 2 consensus sequence and its B73 counterpart consensus sequences. Clone2 CentC monomers “E1-#” and B73 counterpart CentCs “D8-#” and “B10-#”
Figure 21: Graphic depiction of individual CentC monomers in Clone 2 and its B73 counterpart. CentC monomers “colored arrows pointing to the right”, xilon element “colored block”, nucleotide differences “vertical colored lines”, CentC Mo17 sequence trends “thin colored blue arrows above the CentC monomers”, and CentC B73 sequence trends “thin colored green arrows below the CentC monomers”.

B.L3.T1.D8 has part of a xilon element on its 5’ end, suggesting a xilon insertion into the B73 counterpart compared to the Mo17 sequence. The other end of the xilon sequence was found in the GenBank sequence of the same BAC, connecting to more CentC. This piece was not subcloned along with the rest of the BAC clones due to its small size (about a monomer in length). Subcloning very small pieces is very difficult as they get lost in the small fragments of non-CentC sequence caused by HaeIII digestion.

PCR, which was used to find the 3’ end of Clone 1, was not done in this case to find the missing 5’ fragment of B73 sequence. When using PCR to find the missing 3’ fragment of Clone 1, primers were easy to create since both ends of this fragment contained junctions. Creating primers in the case of the B73 counterpart of Clone 2 is more difficult since one end is CentC without a junction to another type of sequence, which would create many different annealing sites due to the similarity between CentC monomers.

There is a 12 monomer deletion in the B73 counterpart sequence. This is more likely to be a deletion in B73 counterpart sequence than an insertion into the Mo17 sequence because they are complete monomers and the HOR is not interrupted in the Mo17 sequence. If it was an insertion into the Mo17 sequence, it would not be likely to see a continuous HOR that starts before the insertion, continues through the insertion, and ends almost one HOR after it. If it were a single HOR that was inserted, this could account for duplication of the entire HOR. This is not the case here, in which it is 1.3 worth of HOR that has been deleted.

A 12 monomer deletion is a large amount of CentC compared to the one monomer deletion in Clone 1. The possibility of a section of CentC looping itself out to artificially create this deletion became a concern. The gel of the HaeIII digested BAC was studied to confirm that B.L3.T1.D8 is a viable clone.

The red box on the gel indicates the section gel that was excised and cloned using methods mentioned in Chapter 2. Usually each band would be cloned separately, but since these two bands were so close together, they were excised together. If they were excised separately, there would have been the risk of not properly excising one or both of the bands. The bands in the green box are about 3.9 and 4.3 kb in size. Clone B.L3.T1.D8 is 4,067 bp in length, which would roughly match the size of the lower band in the green box. Clone 2 is 5,775 bp in length which does not correspond to any band sizes seen on the gel. If the 12 monomer deletion
occurred during cloning and subcloning of the B73 counterpart, the 12 monomers would be present in the CentC island in the BAC and there would be a band corresponding to 5,775bp on the gel. This is not the case. What was seen in the BAC was cloned from the BAC.

Figure 22: Agarose gel analysis of Haell digest of BAC CH201-433P11. Red box indicates two clones that were excised and cloned into pJAZZ together, due to their close proximity to each other. Haell digest BAC CH201-433P11 “1”, NEB Tridye 2-log ladder “L”, and empty lane “X”.

Instability in the BAC itself could produce an artificial 12 monomer deletion in the B73 counterpart. Further study of the BAC involving cloning CentC islands from it at different periods in time could verify this possibility. If the same CentC island experiences change in sequence composition, then the BAC could be considered instable and any further study of B73 using BAC DNA should be done with caution.

There are less nucleotide differences between B.L3.T1.D8 and Clone 2 than between B.L1.T1.H8 and Clone 1. In total, there are 14 nucleotide differences between Clone 2 and its B73 counterpart. This is excluding missing monomers and includes 5 nucleotide insertions or deletions, 6 transversions, and 3 transitions. This would come out to be a nucleotide difference about every 412 nucleotides or about every 2.6 CentC monomers.

3.7.2 Comparison of centromeric and non-centromeric B73 CentC islands

B73 counterparts to Mo17 CentC Islands were isolated from centromere 10 and from a pericentromeric region of chromosome 1. This made it possible to compare centromeric and non-centromeric CentC islands in the B73 inbred (Table 4).

Table 4: Centromeric and non-centromeric differences in B73 CentC islands

<table>
<thead>
<tr>
<th></th>
<th>Centromere</th>
<th>Non-centromere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide differences</td>
<td>Every 1.4 monomers (~218 nucleotides)</td>
<td>Every 2.6 monomers (~412 nucleotides)</td>
</tr>
<tr>
<td>Higher order repeats (HORs)</td>
<td>Multiple, some not consecutive</td>
<td>9-mer, consecutive</td>
</tr>
<tr>
<td>Monomer insertions/ deletions</td>
<td>Monomer deletion in Mo17</td>
<td>12 monomer deletion in B73</td>
</tr>
<tr>
<td>Retrotransposon insertions/ deletions</td>
<td>CRM insertion in Mo17</td>
<td>Xilon insertion in B73</td>
</tr>
</tbody>
</table>

The non-centromeric B73 CentC island has less nucleotide differences when compared to its Mo17 counterpart, but has a larger monomer deletion than the centromeric island. There are more different kinds of HORs in the centromeric CentC island, but they are not consecutive. In contrast, there is only one HOR in the non-centromeric CentC island, and it is larger than those seen in the centromeric island and consecutive. There is a CRM insertion in the Mo17 counterpart to the B73 centromeric CentC island and a xilon insertion in the non-centromeric B73 CentC island.
While there are differences between the centromeric and non-centromeric B73 CentC islands, definite conclusions cannot be made based solely on one representative of each. Future work should be done on more centromeric and non-centromeric B73 CentC islands. This would involve finding a Mo17 clone then finding its corresponding B73 counterpart. This step may be made easier when the Mo17 genome is sequenced. One could then take samples from BACs from each genome and compare. As stated previously, the stability of the BAC itself would have to be made certain.

3.7.3 Clones that did not have an available B73 counterpart

Mo17 clones 3, 4, and 5 did not produce a positive match when BLAST was used against the B73 genome. This could be due to many factors including the clone being unique to Mo17, lack of sequence data of the B73 genome, and error in assembling the B73 BACs in which the misassembly is so great that it skews the matching of B73 sequence to the available Mo17 clone. As more data becomes available, it should be more likely that a B73 counterpart can be found.

Since it is impossible to determine how different the Mo17 clone could be from the B73 counterpart, the prospect of creating a specific probe from the Mo17 clone sequence to find the B73 sequence would not be a useful approach.

3.7.3.1 Clone 3 (M.L3.T1.2.G11): possible CentA insertion

Figure 23: Junction viewer image of Clone 3 (M.L3.T1.2.G11). CentC monomers are represented by green arrows and sequence corresponding to CentA is in orange.

Clone 3 (M.L3.T1.2.G11) is 2,078 bp in length and includes a portion of CentA-like sequence (Fig. 23). HaeIII sites were not found on either end of the clone sequence, this could be due to the use of the BsaWI restriction enzyme. This enzyme cuts sticky ends (W|CCGGW). A small portion of initial clone which included the HaeIII site may have been lost when doing the gel extraction after the BsaWI restriction digest. When the nucleotide sequence of Clone 3 was input into BLASTn HTGS, no significant match was found to the B73 genome.

Figure 24: Phylogenetic tree of individual CentC monomers from Clone 3 consensus sequence. Clone 3 CentC monomers are labeled “G11-#”.
The individual CentC monomers of Clone 3 were aligned and a phylogenetic tree was created (Fig. 24). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig. 25).

**Figure 25: Graphic depiction of individual CentC monomers in Clone 3.** CentC monomers “colored arrows pointing to the right” and CentA-like sequence “colored arrow pointed to the left”.

A CentA-like sequence is present at the 5’ end of the clone (Fig. 22). The CentC sequence did not start with a whole CentC, but with the last half of a monomer, indicating that the CentA sequence possibly inserted within the monomer. Since there is no B73 data on this clone, it would be very hard to verify this at this time.

A two monomer HOR is present three times in the sequence. It is not a continuous array, much like the HORs seen in Clone 1. This clone could potentially be centromeric in sequence, if noncontinuous HORs are a characteristic of centromeric CentC islands. This cannot be verified until further study is done on the comparison of centromeric and non-centromeric CentC islands in an inbred.

### 3.7.3.2 Analysis of Clone 4 (M.L3.T4.2.D3): Multiple HORs

**Figure 26: Junction viewer image of Clone 4 (M.L3.T4.2.D3).** CentC monomers are represented by green arrows and sequence corresponding to CRM1 is in blue.

Clone 4 (M.L3.T4.2.D3) is 4,964 bp in length. HaeIII sites were found on both ends of the clone, thus reinforcing the use of HaeIII as a CentC island enrichment method. A junction viewer image was generated using the consensus sequence (Fig. 26). When the nucleotide sequence of Clone 4 was input into BLASTn HTGS, no significant match was made to the B73 genome.
The individual CentC monomers of Clone 4 were aligned and a phylogenetic tree was created (Fig. 27). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig. 28). Multiple HORs were found and color coded as well.

Figure 27: Phylogenetic tree of individual CentC monomers from Clone 4 consensus sequence. Clone 4 CentC monomers are labeled “2d3-#”.

Figure 28: Graphic depiction of individual CentC monomers in Clone 4. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”.

There are four distinct HORs. Three different two monomer HORs (green arrow, red arrow, and purple arrow) and a larger five monomer HOR (blue arrow) intermix with each other throughout the sequence. This could be evidence of evolution in this CentC island. In Rudd, Wray, and Willard (2006), alpha satellite HORs within an array are seen to be very homogenous. In contrast to what was found in that study, the consensus sequence for Clone 4 has many HORs that are interspersed between each other, making it a very heterogenous mixture of HORs and monomers. This could be a difference that distinguishes between the satellite repeats of Zea mays and humans, or it could demonstrate that the Mo17 inbred CentC array is a product of evolutionary monomer changes and rearrangements which would produce the variations found in the HORs.
3.7.3.3 Analysis of Clone 5 (M.L3.T4.3.H1): Three monomer HOR

Figure 29: Junction viewer image of Clone 5 (M.L3.T4.3.H1). CentC monomers are represented by green arrows.

Clone 5 (M.L3.T4.3.H1) is 2,092 bp in length (Fig. 29). This sequence was obtained by restriction digest of the clone and subcloning the pieces separately, then rejoining them in correct order based on the restriction enzyme cut map as seen in section 2.5.2. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. When the nucleotide sequence of Clone 5 was input into BLASTn HTGS, no significant match was made to the B73 genome.

The individual CentC monomers of Clone 5 were aligned and a phylogenetic tree was created (Fig. 30). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig.31).

Figure 30: Phylogenetic tree of individual CentC monomers from Clone 5 consensus sequence. Clone 5 CentC monomers are labeled “3H1-#”.

Figure 31: Graphic depiction of individual CentC monomers in Clone 5. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 29: Junction viewer image of Clone 5 (M.L3.T4.3.H1). CentC monomers are represented by green arrows.
There are two HORs present in the sequence, both being three monomers in length. Unlike the patterns seen in Clone 1, 3, and 4, these HORs are consecutive, although not identical. While this sequence is more representative of homogenous HORs within an array (Rudd, Wray, & Willard, 2006) in the sense that they are consecutive, they also differ slightly from each other.

The first HOR only differs from the second by the last monomer in the pattern. A mutation in a monomer and a local duplication could explain how this sequence came to be. Initially there was only a homogenous HOR array with three identical HORs (Fig. 32-1). At some time during their evolution, a HOR underwent a few nucleotide changes in one of its monomers (Fig. 32-2). There would have to be enough nucleotide difference to distinguish this monomer from the other monomers in the same position of the repeat on the array. The HOR with the different monomer experiences a local duplication, thus duplicating the new HOR between the two HORs that did not experience any changes (Fig. 32-3).

![Diagram of HOR organization](image)

**Figure 32: Possible cause for Clone 5’s HOR organization.** 1. Initially the HOR array was homogenous. 2. A monomer experiences nucleotide changes, thus distinguishing this HOR from the rest. 3. The HOR that contained the nucleotide change locally duplicated.

This would make the two blue arrowed HORs, older than the two red arrowed HORs. A way to possibly verify this theory would be to find the B73 counterpart of this clone and see if there are more nucleotide changes between the outer HORs than the inner ones, assuming that older HORs will have more nucleotide differences between inbred than newer HORs.

### 3.7.4 Summary of Mo17 CentC island clones

By using the methods presented in Chapter 2, I was able to successfully isolate, clone, and sequence CentC islands from both Mo17 and B73 inbreds. Only two of the five Mo17 clones have a known B73 counterpart (Clone 1 and 2). Clones 3, 4, and 5 did not have a known B73 counterpart, but this may change as more CentC sequence for the B73 genome becomes available.

The B73 counterpart for Clone 1 was found in a centromere 10 B73 BAC. In the comparison between Clone 1 and its B73 counterpart, B.L1.T1.H8, the Mo17 Clone 1 has a deletion of a monomer and an insertion of a CRM element into its sequence. There are multiple HORs present, with some displaying more nucleotide difference between the two sequences, suggesting that they are older than HORs that do not exhibit as many differences.

The B73 counterpart for Clone 2 was found in a B73 BAC that has sequence for a pericentromeric region on chromosome 1. When compared to each other, the B73 clone B.L3.T1.D8 had a 12 monomer deletion in its sequence and a xilon insertion at its 5’ end. There was a consecutive nine monomer HOR which helped verify the 12 monomer deletion.
Since there was an example of a centromeric B73 CentC island and a non-centromeric B73 CentC island, they could be compared to each other based on nucleotide differences, HORs, monomer insertions and/or deletions, and retrotransposon insertions and/or deletions. Any conclusions could only be made on these specific examples and not on the overall differences between centromeric and non-centromeric CentC islands. More examples will need to be studied before this can be done.

Clones 3, 4, and 5 did not have B73 counterparts and were analyzed solely to look for HORs and retrotransposons. Clone 3 has a CentA-like sequence at its 5’ end. Clone 4 is made entirely of CentC sequence and has multiple nonconsecutive HORs. Clone 5 has a consecutive HOR that seems to have undergone a change caused by nucleotide changes and local duplication.

### 3.8 Resequencing a B73 BAC

During the cloning and sequencing of the B73 counterpart of Clone 1, errors in the assembly of the B73 BAC began to appear. The end sequence of the corresponding B73 counterpart would align to one CentC island in the BAC, while the rest of the sequence would align to another contig containing CentC. It is possible that separating and aligning all CentC monomers within the BAC along with the sequenced B73 counterpart for Clone 1, then creating a phylogenetic tree to see like monomers, will reveal where the two CentC containing BAC contigs were assembled wrong.

The B73 counterpart to Clone 1 (B.L1.T1.H8) was present in two BACs: b0410L22 and C0030B18. B.L1.T1.H8 was initially subcloned out of BAC b0410L22 because of it being readily available in the lab, as well as having fewer CentC islands to sift through. After analysis of Clone 1, we discovered that the CentC monomers in BAC b0410L22 were misassembled. It appears that the assembly of the BAC in GenBank preserved the individual CentC monomers, but misassembled their order within the arrays (Fig. 33).

![Comparison of BAC contigs](image)

**Figure 33:** Graphic representation of the comparison between the subcloned B73 counterpart for Clone 1 and contigs 1 and 4 of BAC b0410L22. CentC islands are represented by color coded arrows and corresponding patterns between the BAC data and experimental data are represented by thin colored arrows above the monomers.
Since BAC b0410L22 contained misassembled contigs of CentC, BAC C0030B18 was subcloned. This BAC contained far more CentC monomers than b0410L22, which made it an ideal BAC to study CentC. The procedure for subcloning the BAC is the same as described in section 2.6.

3.8.1 Subcloned B73 BAC C0030B18 CentC islands

After subcloning the B73 BAC C0030B18, six unique clones with different end sequences on both 5’ and 3’ end were found. These clones were then subcloned in order to obtain the entire DNA sequence of each clone (Table 5). Each clone’s sequence was annotated by coloring like monomers as seen in previous clone analyses.

Table 5: Overview of B73 BAC C0030B18 CentC islands

<table>
<thead>
<tr>
<th>Clone</th>
<th>Official name</th>
<th>Size (bp)</th>
<th>HaeIII site?</th>
<th>Centromere/ noncentromere</th>
<th>B73 BAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>B.L2.T1.M13</td>
<td>13888</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>7</td>
<td>B.L2.T1.O22</td>
<td>11975</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>8</td>
<td>B.L2.T1.M16.1</td>
<td>6812</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>9</td>
<td>B.L2.T1.M16.2</td>
<td>6805</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>10</td>
<td>B.L2.T1.M20</td>
<td>6812</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>11</td>
<td>B.L2.T1.N14</td>
<td>4991</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
<tr>
<td>12</td>
<td>B.L2.T1.K21</td>
<td>3400</td>
<td>Yes</td>
<td>Centromere 10</td>
<td>C0030B18</td>
</tr>
</tbody>
</table>

3.8.1.1 Analysis of Clone 6 (B.L2.T1.M13)

Figure 34: Junction viewer image of Clone 6 (B.L2.T1.M13). CentC monomers are represented by green arrows.

Clone 6 (B.L2.T1.M13) is 13,888 bp in length (Fig. 34). This sequence was obtained by the methods detailed in Chapter 2 specified for large DNA fragments. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. This clone is made entirely of CentC monomers and is the largest CentC island to date. The fact that it is made up of only CentC suggests that HaeIII cut the CentC at a point mutation in the island. HaeIII was selected because it did not cut into CentC, so a point mutation must have been present.

The individual CentC monomers of Clone 6 were aligned and a phylogenetic tree was created (Fig. 35). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig.36).
Figure 35: Phylogenetic tree of individual CentC monomers from Clone 6 consensus sequence. Clone 6 CentC monomers are labeled "M13-#".
Figure 36: Graphic depiction of individual CentC monomers in Clone 6. CentC monomers “colored arrows pointing to the right” and CentC sequence trends “thin colored arrows above the CentC monomers”.

Clone 6 (B.L2.T1.M13) is an island made up entirely of CentC monomers. It was more difficult to assemble through Consed due to the similarity of sequence throughout the entire 13kb clone. A six monomer HOR is present twice, a two monomer HOR is present three times, a four monomer HOR is present twice, and a 12 monomer HOR is present three times. The 12 monomer HOR is the only HOR that is consecutive, with the second HOR and third HOR being next to each other. The rest of the HOR present in the sequence are mixed and nonconsecutive. This mixture of HORs is reminiscent of the HOR present in Clone 1. Since this is the B73 counterpart of Clone 1 comes from the BAC that Clone 6 was subcloned from, this is not surprising.

3.8.1.2 Analysis of Clone 7 (B.L2.T1.O22)

Figure 37: Junction viewer image of Clone 7 (B.L2.T1.O22). CentC monomers are represented by green arrows and CRM1 sequence is represented by a blue block.

Clone 7 (B.L2.T1.O22) is 11,975 bp in length (Fig. 37). This sequence was obtained by the methods detailed in section 2 specified for large DNA fragments. HaellII sites were found on both ends of the clone as seen in the other sequence analyses. A CRM1 element was found on the 5’ end of the clone.

The individual CentC monomers of Clone 7 were aligned and a phylogenetic tree was created (Fig. 38). Like monomers were coded the same color and the CentC monomers were
reassembled into its original order (Fig. 39). A five monomer HOR repeats itself consecutively five times. A two monomer HOR is present three times in the sequence, with the second and third being consecutive. A seven monomer HOR repeats itself three times and is interspersed with a four monomer HOR.

Figure 38: Phylogenetic tree of individual CentC monomers from Clone 7 consensus sequence. Clone 7 CentC monomers are labeled “o22-#”.
Clone 7 is the second largest clone derived from the BAC DNA. There is a junction to a CRM1 element at the 5’ end of the clone. The rest of the clone is made up entirely of CentC monomers. The first HOR (blue arrow) repeats itself consecutively five times. This arrangement is similar to that seen in Clone 5 and could suggest a more homogenous repeat pattern like that seen in human alpha satellite DNA.

After a stretch of sequence that is dotted with two monomer HORs, another pattern emerges. In the latter half of the clone, a seven monomer HOR (green arrow) and a four monomer HOR (orange arrow) are present. Both of these HORs could be combined to form a larger 11 monomer HOR since they always appear next to each other. They were separated so that the beginning of the pattern could be seen at the 3’ end, where only the green arrow is seen. Given that this end is entirely CentC and there is a HaeIII site, it could be surmised that the HaeIII site was caused by a point mutation and that it occurred in the middle of a HOR. The orange arrow would then be on the other side of the HaeIII site and the green and orange arrow would truly be one HOR.

It is very interesting to see two distinct patterns separated by such a span of monomers and two monomer HORs. This suggests that Clone 7 is evidence of an area that connects two different arrays with CentC monomers.

3.8.1.3 Analysis of Clone 8 and 9 (B.L2.T1.M16.1 and M16.2)

Initially, Clones 8 and 9 were thought to be one 12 kb clone that happened to have a HaeIII site cutting it in half into two 6 kb fragments that initially was thought to be due to a partial HaeIII digest. Because of the end sequences of the two inner ends, 3’ end of M16 contig 1 and 5’ end of M16 contig 2, not agreeing, subcloning of these two fragments was conducted to confirm
that these are two separate clones, and that the disagreeing contig ends were not a product of erroneous Consed assembly. The contigs were then renamed M16.1 and M16.2 and treated in analysis as two separate clones.

**Analysis of Clone 8 (B.L2.T1.M16.1)**

**Figure 40:** Junction viewer image of Clone 8 (B.L2.T1.M16.1). CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.

Clone 8 (B.L2.T1.M16.1) is 6,812 bp in length (Fig. 40). This sequence was obtained by the methods detailed in Section 2 specified for large DNA fragments. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. Sequence from a CRM1 element was found at the 5’ end of the clone and sequence from a cinful element was found at the 3’ end of the clone.

The individual CentC monomers of Clone 8 were aligned and a phylogenetic tree was created (Fig. 41). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig. 42).

**Figure 41:** Phylogenetic tree of individual CentC monomers from Clone 8 consensus sequence. Clone 8 CentC monomers are labeled “M16.1-#”. 
Clone 8 has a CRM1 on its 5’ end and a cinful element at the 3’ end. There are multiple HORs present in the sequence. A two monomer HOR is present twice (orange arrow), another two monomer HOR that repeats three times (green arrow), a three monomer HOR that repeats twice (red arrow), and a five monomer HOR that repeats three times (blue arrow). The only consecutive HOR is the five monomer HOR that repeats itself once. The rest of the HORs are nonconsecutive and are interspersed between each other. The different HORs interspersed between each other are much like Clone 1 and its B73 counterpart from centromere 10. Since this BAC has overlap with the BAC that the counterpart for Clone 1 came from, this may indicate a centromeric region, based on the multiple HOR that are interspersed between each other. Future study would involve studying more BACs to see whether HOR patterns can determine centromeric or non-centromeric regions of CentC.

Analysis of Clone 9 (B.L2.T1.M16.2)

Clone 9 (B.L2.T1.M16.2) is 6,805 bp in length (Fig. 43). This sequence was obtained by the methods specified for large DNA fragments detailed in Chapter 2. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. Sequence corresponding to a hypothetical protein was found on the 3’ end of the otherwise completely CentC clone.
The individual CentC monomers of Clone 9 were aligned and a phylogenetic tree was created (Fig. 44). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig. 45).

Figure 44: Phylogenetic tree of individual CentC monomers from Clone 9 consensus sequence. Clone 9 CentC monomers are labeled “M16.2-#”.

Figure 45: Graphic depiction of individual CentC monomers in Clone 9. CentC monomers “colored arrows pointing to the right”, hypothetical protein “block with yellow fading to red”, and CentC sequence trends “thin colored arrows above the CentC monomers”.
There are multiple HORs present in the sequence. A five monomer HOR is present twice (red arrow), another five monomer HOR that repeats four times (blue arrow) throughout the island, and a three monomer HOR that repeats twice (green arrow).

While the HORs are not consecutive with like HORs, there is little disruption by monomers between individual HORs. Initially the first red and blue arrow HORs were one HOR, and the second blue and green arrow HOR were one HOR. I separated them into smaller HORs so that it would be easier to see how this CentC island could have evolved. Initially there were only a red and blue HOR. They duplicated locally together, creating a larger island with alternating red and blue arrow HORs (Fig. 46-1). This was followed by a local duplication of a single blue arrow HOR (Fig. 46-2). After that duplication, there is an insertion of a smaller green arrow HOR (Fig. 46-3). This HOR has monomers that are closely related to the middle portion of the red arrow HOR, possibly indicating that the green arrow HOR is a partial local duplication of the red arrow HOR. Finally, there is a duplication of a blue arrow and green arrow HOR pair (Fig. 46-4).

Figure 46: Proposed theory on how the Clone 9 CentC island has evolved. Like HORs are represented by same colored arrows. First a red and blue arrow pair of HORs duplicates locally (1), followed by a blue arrow HOR duplicating locally (2), then an insertion of a few CentC monomers which happen to be related to part of the red arrow HOR (3), finished by a local duplication of a pair of blue and green arrow HORs (4).

The appearance of two like monomers next to each other in a HOR (blue arrow) is only seen in this clone and Clone 2. It is also the only clone that shares a junction with a hypothetical protein. This could indicate that Clone 9 is a different “type” of CentC island. We do know that this island is located in the centromere of chromosome 10. This clone could be located farther away from the centromere’s center. Since the B73 counterpart for Clone 2 is from a pericentromeric region, maybe having two like monomers next to each other is characteristic of island that are farther away from the center of the centromere’s center. It could be evidence of a different CentC HOR variation jumping into this centromere from another centromere where this type is prevalent. Since Clone 2’s B73 counterpart is from a different BAC (CH201-433P11) than Clone 9 (C0030B18) and located on different centromeres, this could be the case. Further study would be required to solidify these proposed theories.
3.8.1.4 Analysis of Clone 10 (B.L2.T1.M20)

Figure 47: Junction viewer image of Clone 10 (B.L2.T1.M20). CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.

Clone 10 (B.L2.T1.M20) is 6,812 bp in length (Fig. 47). This sequence was obtained by the methods detailed in Section 2 specified for large DNA fragments. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. The sequence of a CRM1 element was found on the 5’ end of the clone. Sequence from a cinful element was found on the 3’ end.

The individual CentC monomers of Clone 10 were aligned and a phylogenetic tree was created (Fig. 48). Like monomers were coded the same color and the CentC monomers were reassembled into its original order.

Figure 48: Phylogenetic tree of individual CentC monomers from Clone 10 consensus sequence. Clone 10 CentC monomers are labeled “M20-#”.

55
Clone 10 has a CRM1 on its 5’ end and a cinful element at its 3’ end (Fig. 49). There are multiple HORs present in this CentC island. A two monomer HOR is present twice (orange arrow), another two monomer HOR is present three times (green arrow), a three monomer HOR that repeats twice (red arrow), and a five monomer HOR that repeats three times (blue arrow).

**Figure 49:** Graphic depiction of individual CentC monomers in Clone 10. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

The clone ends of Clone 10 and its HORs are identical to that of Clone 8 (Fig. 42). Upon further analysis it was found that Clone 8 and Clone 10 are not only identical in monomer pattern, but they are identical in nucleotide order as well. This was confirmed by an alignment of the two consensus sequences in Clustal (Fig. 50).
Since Clone 8 and Clone 10 are identical, this confirms that Clones 8 and 9 are indeed two separate clones that somehow were ligated together during cloning. This also verifies that the method described in Chapter 2 is a viable way to extract CentC clones. The nucleotide sequences of the two clones have no differences, indicating that there is a 0% error rate in the method from gel extraction to sequencing. This method minimized excess UV exposure that could cause point mutations, preventing accurate assembly of the sequence in Consed.

3.8.1.5 Analysis of Clone 11 (B.L2.T1.N14)

Clone 11 (B.L2.T1.N14) is 4,991 bp in length (Fig. 51). This sequence was obtained by the methods detailed in chapter 2 subcloning specified for large DNA fragments. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. Clone 11 is the only clone containing a dagaf element. This could be evidence that this clone is located on one of the ends of the BAC, and may be used to connect this BAC to another BAC.

The individual CentC monomers of Clone 11 were aligned and a phylogenetic tree was created (Fig. 52). Like monomers were coded the same color and the CentC monomers were reassembled into its original order.
A dagaf element was found on the 5’ end of the clone (Fig. 53). Sequence from a cinful element was found on the 3’ end. A five monomer pattern repeats itself twice, separated by four monomers. The presence of a five monomer pattern separated by four monomers could be evidence of a HOR that has undergone changes in its edges. This would produce a smaller HOR with monomer separations between them. Future study of this BAC would be required to verify this finding. A possible match to a Mo17 clone could also shed some light on the nucleic acid changes at the edges of these HORs.

Figure 53: Graphic depiction of individual CentC monomers in Clone 11. CentC monomers “colored arrows pointing to the right”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

Figure 52: Phylogenetic tree of individual CentC monomers from Clone 11 consensus sequence. Clone 11 CentC monomers are labeled “N14-#”.
3.8.1.6 Analysis of Clone 12 (B.L2.T1.K21)

![Figure 54: Junction viewer image of Clone 12 (B.L2.T1.K21).](image1)

CentC monomers are represented by green arrows, CRM1 sequence is represented by a blue block, and cinful sequence is represented by a gray block.

Clone 12 (B.L2.T1.K21) is 3,400 bp in length (Fig. 54). This sequence was obtained by the methods detailed in Chapter 2 specified for large DNA fragments. HaeIII sites were found on both ends of the clone as seen in the other sequence analyses. CRM1 sequence was found on the 5’ end of the clone. Sequence from a cinful element was found on the 3’ end. It is only 21 monomers long and is the shortest clone derived from the BAC.

The individual CentC monomers of Clone 12 were aligned and a phylogenetic tree was created (Fig. 55). Like monomers were coded the same color and the CentC monomers were reassembled into its original order (Fig. 56).

![Figure 55: Phylogenetic tree of individual CentC monomers from Clone 12 consensus sequence.](image2)

Clone 12 CentC monomers are labeled “K21 #”.
Figure 56: Graphic depiction of individual CentC monomers in Clone 12. CentC monomers “colored arrows pointing to the right”, CRM1 element “brown striped block”, cinful element “block with green gradient”, and CentC sequence trends “thin colored arrows above the CentC monomers”.

There is a two monomer HOR that is repeated twice (red arrow) and a four monomer HOR that repeats twice (blue arrow). The presence of multiple HORs is similar to that seen in other clones. The small size of this clone could indicate that it is actually the result of a CRM or cinful insertion into a larger array. In future work it would be interesting to find the Mo17 CentC island counterpart to this clone. It could be a larger island that is uninterrupted by retrotransposons, where the B73 island has had a CRM and/or cinful insertion.

3.8.2 Junctions with retrotransposons

In a number of B73 BAC C0030B18 subclones, the 5’ and 3’ ends feature junctions with retrotransposon elements. CRM is present at the 5’ end and cinful is present at the 3’ end of numerous clones.

3.8.2.1 CRM1 junctions

Clones 7 (Fig.37), 8 (Fig. 40), 10 (Fig. 47) and 12 (Fig. 54) all possess a CRM1 junction at their 5’ end. This end on each clone was aligned with one another to determine if the junction is identical in all. The CRM1 element takes up 327 to 330 bp of the end of each clone. Clone 8 and 10 are identical in sequence which was expected as discussed in Section 3.8.1.4. All CRM1 sequence ends with the nucleotides CATCA before the CentC sequence starts. After this site, Clone 12 begins to vary right away, while Clones 7, 8, and 10 are identical for about 1,660 bp after the CRM1 element has stopped. This alignment was done to verify that all the clones are positively different clones.

The large stretch of identical sequence after the CRM1 present in Clone 7, 8, and 10 may be evidence of evolution in B73. The homogenous sequence around the CRM1 element could suggest that the CRM1 and ten to eleven CentC monomers around it could have inserted more recently from a different location, resulting in this homogeneity. In contrast, Clone 12 may have inserted earlier, which would account for its sequence differences starting right after the CRM1 element.

3.8.2.2 Cinful junctions

Clones 8 (Fig. 40), 10 (Fig. 47), 11 (Fig. 51), 12 (Fig. 54), and the B73 counterpart to Mo17 Clone 1 (Fig 14) all have junctions containing cinful elements. These were studied the same way as the CRM1 junctions. Cinful sequences make up about 107 bp of each 3’ end of these clones and are all identical in sequence. The nucleotides TGTTG link the cinful element with the CentC sequence of each clone. Clone 11 differentiates in sequence immediately after this site, while Clones 8, 10, 12 and the B73 counterpart to Clone 1 all share 1,030 bp of CentC sequence from the cinful element. As discussed above, this could be evidence of a cinful insertion, with Clone 11’s insertion being less recent than the other clones.
3.8.2.3 Possible explanations for similar junctions

Questions arose since most of the clones from the subcloning have very similar 5’ and 3’ ends. To explain this phenomenon, multiple explanations were found.

If the BAC itself were unstable, the DNA that was used to subclone could have been the product of random rearrangement prior to DNA isolation. Future work will have to be done to ensure that the BAC is stable. A suggested way of doing this would be to subclone the same BAC using the same method and compare sequences of each clone.

Another possibility is that the CRM and cinful elements at each end are real and that there is local duplication within the centromere. This would result in many CentC islands being interrupted with the same CRM and cinful elements. The CRM and cinful elements, which are retrotransposons, were able to jump and duplicate within the CentC array.

Unstable cloning could also be a factor, even though the cloning kit used advertised a very stable clone capable of cloning repetitive DNA. An example of this would be the initial confusion of Clones 8 and 9 originally being seen as one clone instead of two clones which happened to ligate together.

3.8.3 Summary of B73 BAC C0030B18 subcloning

Due to the discovery that the sequence of the B73 counterpart to Clone 1 (B.L1.T1.H8) corresponding to two different contigs in B73 BAC b0410L22, the BAC sequence from GenBank was deemed misassembled. The other BAC that B.L1.T1.H8 belonged to was resequenced due to this discovery.

Six clones were initially identified as CentC positive using the technique outlined in Chapter 2. Of these six, one clone was determined to be a combination of two shorter fragments (Clone 8 and Clone 9) of CentC that were somehow ligated together during the cloning process. This assumption was reaffirmed when it was discovered that a completely different fragment of CentC isolated from the BAC (Clone 10) was identical to Clone 8 with no sign of any sequence from Clone 9.

CRM1 and cinful junctions present in multiple clones were verified as all unique, save the two identical clones. Most of the sequences share about 1kb of similarity before differentiating into unique sequence. An unstable BAC, local duplication, and an unstable clone are a few possibilities that could explain the similar 5’ and 3’ ends.

3.9 CentC monomer comparison

All monomers of all Mo17 and B73 CentC islands were aligned in order to see if there were any conserved regions, as seen in the AtCon satellite repeat (Hall, Kettle, & Preuss, 2003). Unlike the AtCon satellite repeat, there are no conserved regions in either the B73 CentC monomers or Mo17 monomers.
CHAPTER 4. CONCLUSION

Figure 57: The evolution of CentC islands can be seen by looking at their HORs and nucleic acid differences. Higher order repeats show evolution through local duplications and subtle monomer changes. When comparing CentC sequences between inbreds, nucleic acid differences may give a hint at older and newer parts of the sequence.

HORs shed light on the evolution of CentC satellite repeats based on nucleotide differences between the two inbreds (Fig. 57). They are a way to relatively date HORs and determine which parts of the sequence are newer or older. As seen in Clone 1, HORs that have more nucleotide differences between the Mo17 and B73 counterpart clones may be older than the HORs that have few or no nucleotide differences.

The pattern of HOR in a specific CentC island may also show how the CentC island has evolved over time. This can be seen through duplications of single or multiple HORs within the same island. In some cases, duplication may help propagate a HOR with a nucleotide change (Clone 5). And in some cases, multiple duplication events may help create a CentC island with multiple HORs (Clone 9). In this way, HORs are part of a mechanism that increases CentC island length.

The evolution of CentC is also characterized by retrotansposon insertions / deletions and monomer insertions/ deletions. In comparisons between Mo17 CentC islands and B73 counterparts, we were able to determine an insertion as opposed to a deletion of monomers and retrotansposons based on the monomeric and HOR patterns around the insertion.
CHAPTER 5. SUMMARY

Figure 58: All the clones obtained in this study. Green lines indicate clones from B73 DNA and blue lines indicating Mo17 DNA. Black arrows indicate a match between Mo17 and B73 clones.

In all, 92,112 bp of CentC and junctions were cloned and sequenced. Of this total, 63,330 bp were cloned from B73 BAC C0030B18, 22,823 bp were cloned from Mo17 genomic DNA, and 5,959 bp were cloned from B73 BAC CH201-433P11 (Fig. 58). There are many HORs found within CentC islands, but unlike the alpha satellite repeat in primates, there are not many that are homogenous within a certain island. Unlike the AtCon sequence in Arabidopsis, there are distinct conserved regions. And unlike CentO found in rice, there are no other CentC types other than CentC.

When comparing the centromeric and non-centromeric Mo17 and B73 CentC island alignments, non-centromeric B73 and its Mo17 counterpart have more nucleotide differences, but not by much. The HOR in the noncentromeric B73 and its Mo17 counterpart is a consecutive nine monomer HOR that spans a 12 monomer gap in B73. There are 3 different HORs present in the centromeric B73 and its Mo17 counterpart. Some of them are consecutive, while others are interspersed between each other.

Resequencing a B73 BAC resulted in seven unique clones. A single clone was subcloned and sequenced twice, with both times being separate from each other, producing a 0% error rate, thus validating my subcloning protocols. Multiple clones had similar CRM elements and ciful elements at their ends. These similar ends could be evidence of an unstable BAC, a problem with cloning, or valid local duplication on a centromere.

In this work I have shown the presence of HORs in CentC. This provides an idea of the size by which CentC islands grow. By looking at the nucleotide differences between the CentC islands of two inbreds, I was able to determine which HORs were older or newer. I have also illustrated that single CentC monomers can be lost in a CentC island, and also that retrotransposon insertions are rampant. I was able to begin comparing centromeric and non-centromeric CentC islands from the same inbred, and while there were differences, more study would be required to come to definite conclusions about their sequence differences. Resequencing a CentC-containing BAC revealed confusing results that may reflect the instability of CentC-rich BAC clones.
REFERENCES


