Gametophytic cross-incompatibility in maize: Resequencing the *Ga1* locus

by

Marianne Lynn Emery

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Plant Breeding

Program of Study Committee:
M. Paul Scott, Co-Major Professor
Thomas Lubberstedt, Co-Major Professor
Michael Muszynski

Iowa State University

Ames, Iowa

2015

Copyright © Marianne Lynn Emery, 2015. All rights reserved.

For my parents, who instilled a work ethic and tenacious nature in their daughters that bound them for success

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Special thanks are extended to Nathan Weeks, Ethy Cannon, and Michelle Graham for their bioinformatics expertise in various aspects of the project.

ABSTRACT

Maize is an important staple crop for many countries. Culture dictates maize use, processing, and incorporation into foods. The crop has a rich history of domestication and improvement. With its relative ease of genetic manipulation, maize is considered a model crop for plant genetic experimentation. Recent biotechnological advances, as well as the completed B73 reference genome sequence, have expedited maize improvement. One such profound advance that has greatly increased profitability of maize is the use of transgenes. Despite the many benefits, transgenic plants are problematic when they contaminate transgene-free maize. Maintaining the purity of transgene-free maize is crucial, but often difficult when in close proximity to transgenic fields. Past literature suggests the use of the *Ga1* gametophytic cross-incompatibility system to control pollen flow and minimize contamination of transgene-free maize. Yet, information about how the gametophytic cross-incompatibility system functions at the molecular level is still lacking. Our research sought to assemble BACs containing the *Ga1-m* locus to better understand sequence variation with the B73 reference genome that may be causative of the male function in the *Ga1* gametophytic cross-incompatibility system.

CHAPTER ONE: INTRODUCTION

Fertilization in Maize

Maize is a monoecious plant, possessing separate male and female reproductive organs on the same plant. Both organ types produce gametes, or haploid sex cells. Maize plants possess one male reproductive organ, referred to as the tassel. It is situated at the very tip of the main stem. One tassel can produce up to one billion pollen grains. These pollen grains are the male gametes. In contrast, several female reproductive organs can be present on one maize plant; the female reproductive organs are commonly known as ears. The female gametophyte, the embryo sac, is located within the ears. The ears are positioned at one or more nodes down the length of the main stem and are connected via a sheath. Fertilization occurs when gametes fuse to create a zygote; the maize fertilization process can be outlined in five main steps: (1) pollen hydration, (2) pollen tube germination and penetration of the stigma, (3) pollen tube growth in the transmitting tract and entering the embryo sac, (4) pollen tube exiting the transmitting tract, and (5) bursting of pollen tube which releases sperm nuclei and results in fertilization of the egg cell and two polar nuclei (Heslop-Harrison, 1982; Dresseslhause & Franklin-Tong, 2013; Johnson and Preuss, 2002). Mature pollen becomes dehydrated on the tassel and is dehisced where upon it travels on the wind until it lands on silks of the same (self-pollination) or different (cross-pollination) maize plants. Via osmosis, the silks quickly hydrate the pollen grain. Once hydrated, the pollen grain produces a pollen tube that enters the transmitting tract of the silk. In the event of a successful maize fertilization, the pollen tube continues to grow down the length of the silk in oscillating bursts and pulses until it reaches the ovary (Heslop-Harrison, 1987). Here, the tip of the pollen tube bursts releasing two sperm nuclei. One sperm cell fertilizes the egg to produce the diploid zygote while the other sperm cell fuses with two polar nuclei to develop the triploid endosperm. This process is referred to as double fertilization.

Pollen-stigma interactions in pollen tube growth are not clearly understood. Pollen tube germination is recorded at growth rates close to one centimeter (cm) per hour, commencing five minutes after the pollen grain is deposited on the silk (Barnabas & Fridvalszky, 1984; Mascarenhas, 1993). Maize pollen tubes have been reported to grow to over 30 cm in length (Lu et al, 2014). Pollen tube growth is from the tip of the tube; the tip region is known to have intense secretory activity that is highly sensitive to Ca^{2+} gradients (Derksen et al, 1995; Feijo et al, 1995; Giampiero et al, 199; Steer and Steer, 1989). Cytoplasmic streaming and rearrangement of vesicles, membranes, and other organelles at the tip of the pollen tube are essential for germination and growth (Franklin-Tong, 1999; Heslop-Harrison & Heslop-Harrison, 1990; Heslop-Harrison & Heslop-Harrison, 1991; Mascarenhas, 1993; Pierson et al, 1990).

Despite highly optimized germination media, pollen tubes in vitro reach only 30-40% of their comparative in vivo lengths (Read et al, 1993). It appears that the silk plays a crucial role in pollen tube germination and growth. Research suggests that proteins encapsulating the pollen, waxes, and certain lipids may assist in initiating signaling required for both adhesion and germination of the pollen tube (Franklin-Tong, 1999). Despite experimental observations of pollen-stigma interactions, a complete understanding of the requirements and mechanisms of a germinating pollen tube have yet to be clearly defined.

Gametophytic Incompatibility Systems in Maize

Gametophytic incompatibility was first observed by Correns in 1902. In a breeding experiment with White Rice Popcorn and a *sugary1* (*su1*) mutant, Correns observed distorted F_2 ratios for the sugary-starchy phenotype. Later, while researching the white maize phenotype, Demerec (1929) noted that he could only set seed with popcorn lines when they were used as a female in the cross. Demerec demonstrated that while the popcorn genotype was self-fertile, it was not fertile to non-popcorn pollen even in the absence of any competitive pollen. Demerec concluded the selective fertilizations were a result of a dominant factor linked to the *sugary1* (*su1)* locus. Emerson (1934) later noted that the crosses in White Rice Popcorn were not controlled by the *su1* locus, but by an allele linked to *su1*.

The inability of certain genotypes to successfully pollinate other genotypes is attributed to unique components referred to as gametophyte factors. Both male and female gametophyte functions have been described (Nelson, 1994). Gametophyte factors regulate the success of pollen-stigma interactions and are credited for the aberrant Mendellian genetic ratios in certain crosses which in turn can influence gene flow. More specifically, the female function is a unique component found in silks of select genotypes that allows for discrimination against certain pollen types. The male function, on the other hand, refers to a unique component found in pollen of select genotypes that allows the pollen to overcome the silk barrier. Though the exact interaction is still unclear, results by Kermicle and Evans (2005, 2010) suggest that incompatibility is due to the lack of matching alleles and not active rejection. Eventual cloning of the gametophytic cross-incompatibility genes will hopefully provide insight into the molecular and biochemical mechanisms responsible for these interactions.

Gametophytic cross-incompatibility systems have been shown to play a role in isolating sympatric Mexican maize landrances with teosinte populations (Kermicle $ Evans, 2010). Three gametophytic incompatibility systems in maize have been described: *Gametophtye factor-1* (*Ga1*), *Gametophyte factor-2* (*Ga2*), and *Teosinte crossing barrier* (*Tcb1*).

Ga1

Ga1 has been the most well studied gametophyte factor. It was mapped to the short arm of chromosome 4 in maize (Bloom and Holland, 2011; Liu et al., 2014; Mangelsdorf & Jones, 1926; Zhang et al., 2012). Three variants at the *Ga1* locus have been identified: *ga1*, *Ga1-s*, and *Ga1-m*. The *ga1* locus is found in most conventional grain production fields (i.e. #2 yellow dent). Plants with the *ga1* locus do not contain the male or female function. The *ga1* haplotype can be pollinated by *ga1*, *Ga1-s,* and *Ga1-m; ga1* pollen is discriminated against, however, by the *Ga1-s* silks (Kermicle, 2006; Kermicle & Evans, 2005; Nelson, 1952).

Ga1-s is considered the "strong" variant of *Ga1*. Plants with the *Ga1-s* haplotype possess both the male and female function. *Ga1-s* plants can be pollinated by *Ga1-s* and *Ga1-m* pollen. However, *ga1* pollen fails to successfully pollinate *Ga1-s*, even in the absence of competing pollen (Kermicle & Evans, 2005). When *Ga1-s/ga1* heterozygous plants are self-fertilized, *ga1* pollen is discriminated against and virtually all seed set is by *Ga1-s* pollen (Emerson, 1934). When only *ga1* pollen is present, fertilization of *ga1/Ga1-s* silks will occur to varying degrees (Schwartz, 1950; Nelson, 1952). *Ga1-m* genotypes contain the male function only. Plants with the *Ga1-m* haplotype can self-pollinate and can be used to a pollen parent to cross-pollinate *ga1* and *Ga1-s* plants. *Ga1-m* silks can be successfully pollinated by *ga1*, *Ga1-m* and *Ga1-s* pollen (Jimenez & Nelson, 1964; Kermicle & Evans, 2010; Kermicle et al, 2006). In the *Ga1* system, Kermicle and Evans (2005) demonstrated that the presence of the dominant allele (*Ga1-s* or *Ga1-m*) led to successful fertilization of dominant silks; the presence of the *ga1* allele was not causative of pollen tube growth arrest. A translocation B-4Sa was introgressed into the *W22* inbred line, resulting in the creation of disomic pollen grains. The disomic pollen grains verified what is now referred to as the congruity model.

Ga2

Ga2 was mapped to the long arm of chromosome 5 in maize and teosinte populations (Longley, 1960; Kermicle & Evans, 2010). Four alleles of the *Ga2* locus have been identified: *Ga2-s* (strong), *Ga2-w* (weak), *Ga2-m* (male), and *ga2* (null) (Longley, 1930, Kermicle & Evans, 2010). Past experiments suggests that *Ga2-s* is found only in teosinte lines, *Ga2-w* is found only in Mexican landraces, and *Ga2-m* is found in both teosinte and Mexican landraces. Nonetheless, *Ga2* was proven to be a parallel, but separate, system to that of *Ga1* and *Tcb1* (Kermicle & Evans, 2010). All three systems contain a null allele (with no female or male function), a *–m* allele (male function only), and *–s* allele (female and male function). Similar to the experiments done with *Ga1*, Kermicle and Evans (2010) created disomic pollen grain (*Ga2/ga2*). The disomic pollen was able to successfully pollinate dominant *Ga2* silks, suggesting, as in the *Ga1* system, a congruity model rather than an active rejection (Kermicle & Evans, 2010).

<u>*Tcb1*</u>

Tcb1 was mapped to chromosome 4, a distance of 44 centimorgans (cM) from *Ga1* and 6 cM from *su1* (Evans & Kermicle, 2001). *Tcb1* is found only in teosinte populations, unlike *Ga1* and *Ga2* which are found in both maize and teosinte populations (Kermicle and Evans, 2010). Male and female factors have been described for the *Tcb1* locus. Lu et al (2014) created attenuated lineages of *Tcb1-s*, thus demonstrating that pistil function can be gradually lost via recurrent backcrossing to maize without losing pollen function.

In all three gametophytic incompatibility systems, the barrier is stronger in homozygous compared to heterozygous plants, suggesting a co-dominant effect (Kermicle & Evans, 2005). The barriers do not always exclude 100% of the incompatible pollen, however, which leads to greater difficulty in distinguishing between active pollen rejection and gametophytic incompatibility.

The gametophyte factor has been shown to interact weakly. Attenuated *Tcb1* lines were shown to be more compatible with *Ga1-s* than with *ga1* (Evans & Kermicle, 2001); *Ga1* has been shown to weakly interact with *Ga2* as well resulting in successful fertilizations (Kermicle & Evans, 2010). All systems, however, are associated with premature pollen tube termination (Lu et al, 2014; Zhao et al, 2014). Interestingly, pollen tube growth patterns vary among incompatibility systems with incompatible pollinations. In the *Ga1-s* barrier, pollen tubes do not grow straight and demonstrate heavy accumulation of clustered callose plug deposits; the *Ga2* barrier also leads to clustered callose plug deposits, with lateral kinks in the pollen tube at each callose plug site; in the *Tcb1-s* barrier, pollen tubes grow straight with spaced callose plugs (Lu et al, 2014).

Zhang et al (2012) performed a genetic analysis and *Ga1-s* fine mapping study using the popcorn line SDGa25 (Zhang et al., 2012). Four BC_1F_1 mapping populations were created with Jing24, W22, HN287, and JKN2000F lines. SDGa25 was used a a tester to phenotype the BC_1F_1 populations. SSR markers were used to fine map the *Ga1-s* locus to a 2.2 Mbp region on the short arm of chr 4. Pollen tube growth studies were also performed. The following pollen-pistil combinations were used: W22 pollen presented on SDGa25 pistils (incompatible reaction), SDGa25 pollen presented on SDGa25 pistils (compatible reaction), and SDGa25 pollen presented on W22 pistils (compatible reactions). Pollen tube growth was fixed and stained with aniline blue at 0.15, 0.5, 1, 2, 5, 10, and 20 hours. The experiment provided additional insight into the mechanism underlying an incompatible reaction. In both compatible and incompatible reactions, pollen tubes germinated and entered the transmitting tract in all cases, but once in the silk, significant differences in tube growth were observed. Pollen tubes in compatible reactions grew at a rate of 10 mm h^{-1} versus the incompatible reactions that grew only 2.8 mm h^{-1}. Obvious significant differences in growth were seen after two hours. After 20 hours of growth, pollen tubes in compatible reactions grew the full length of the pistil and reached the ovary; in incompatible reactions pollen tube growth arrested 5.5 cm distal to the ovule and fertilization never occurred.

Despite the amount of research that has been done on the topic of pollen tube growth, a complete picture of pollen tube growth has yet to be fully realized. The mechanisms surrounding pollen-stigma interactions also remains a question not entirely answered. Both pollen tube growth and pollen-stigma interactions do, however, remain a topic of avid interest.

Gametophytic Self-Incompatibility

Similar to gametophytic cross-incompatibility, gametophytic self-incompatibility is the inability of a plant, producing both fertile male and female gametes to create zygotes after self-pollination (Nettancourt, 1977). Darwin (1877) first described self-incompatibility. He proposed that systems in which plants were unable to successfully self-pollinate were integral to the evolution of flowering plants and ultimately encouraged allogamy, also known as cross-pollination. Since the time of Darwin, self-incompatibility has been extensively researched. The genetic control of self-incompatibility varies among species. In the Solanaceae family, a single locus governs the system; in most grasses, two loci (S and Z) are responsible for the barrier (Takayama, et al., 2012); four loci control self-incompatibility in sugarbeet (Lundqvist et al, 1973).

Protein-protein interactions determine fertilization outcomes in the gametophytic self-incompatibility systems. Both the pollen and pistil produce proteins that interact during pollination. If the proteins match, as is the case in self-fertilization, pollen tube growth never occurs (active rejection); if the pollen pistil proteins do not match, the pollen tube elongates (Takayama & Isogai, 2005). The S-locus controls specific protein expression in the pistil and pollen. The locus is made up of several tightly linked genes. There exist many alleles of the S-locus.

Gametophytic vs. Sporophytic Incompatibility

A main difference between gametophytic and sporophytic incompatibility reactions is the tissue type that exerts control over the system. Gametophytic-level control is contingent solely on the haplotype of the pollen or the egg (haploid tissue);

whereas, sporophytic-level control pertains to the pistil or stamen (diploid tissue) (Kermicle & Evans, 2005; Franklin-Tong & Franklin, 2003; Takayama & Isogai, 2005).

Sporophytic incompatibility, similar to gametophytic self-incompatibility is controlled by the S-locus; the proteins involved are, however, created before meiosis is complete, whereas in gametophytic incompatibility proteins are synthesized upon pollen-stigma interaction after meiosis (Franklin-Tong & Franklin, 2003; Takayama & Isogai, 2005). Another point of dissimilarity is in pollen tube arrest. In gametophytic incompatibility, the pollen tube arrests within the stigma, while in sporophytic incompatibility, the pollen tube arrests at the surface of the stigma and penetration of the style never occurs (Pandey, 1958). Roberts et al, (1980) demonstrated sporophytic control in a self-incompatibility system in *Brassica oleracea*. The research demonstrated that the pollen coat carries information for plant recognition and alterations in the pollen coast can lead to incompatibility.

As in the case of gametophytic cross-incompatibility, the pistil barrier and the genotype of the pollen grain work together to determine if pollination is compatible or incompatible (Kermicle & Evans, 2010).

Rationale

The cultivation and harvest of genetically modified (GM) crops have continued to increase since the release of the first GM crop, the FlvrSvr tomato, in 1994 (USDA-ERS, 2014). Since that time, additional GM maize varieties have been created and gained popularity among US farmers. USDA-ERS (2014) reported that in 2014, 76% of all planted maize acres in the United States contained stacked traits for both herbicide tolerance (Ht) and insect resistance (Bt). A parallel increase in organic maize production has been observed. Often fueled by consumer concerns regarding GM crops safety, increasingly large populations of consumers demand organic maize for consumption in both fresh and processed foods, as well as livestock rations. From 1995 to 2008, acreage of organic maize harvested in the United States had increased by 161,987 acres (Brester, 2012). With increasing acreage of organic maize grown alongside GM maize, the potential for cross pollination has increased as well. The USDA (2015) requires that products qualified for the USDA organic seal are void of genetically modified organisms. Being a value added, specialty product, maintaining purity of organic maize fields is an economic necessity.

The implication of controlling adventitious presence, the unwanted presence of transgenes, extends beyond assisting the organic sector of maize production. Maize biotechnology companies own patents on GM varieties and monitor the production of maize under such patents. Therefore, maintaining purity of the remaining maize market classes is of utmost concern. Successful field isolation of market classes, such as white maize used in the food industry and other specialty maize crops, such as high amylose maize and sweet corn, is difficult and cross pollination with neighboring GM fields and other non-specialty maize fields often occurs.

Steps to control pollen flow between neighboring fields can be taken. Physical borders and buffer zones are planted between GM and organic fields. Additionally, delayed plantings help ensure that neighboring fields are at differing reproductive stages. A delay of three to four days between field plantings has been recorded to reduce adventitious presence by 75% (Della Porta et al, 2008). This technique is often used in the cultivation of sweet corn. Unfortunately, pollen can travel great distances on the wind. Maize transgenes were found as high as 47% in non-GM fields residing adjacent to GM fields (Goggi et al, 2006). Della Porta et al (2008) demonstrated that a distance greater than 100 meters must be maintained between fields to maintain a cross-contamination threshold of 0.1%. Insects can also be a source of contamination. A more accurate means to protect market classes and maintain purity of value added maize is required. A naturally occurring biological reproductive barrier, such as gametophytic incompatibility, that prevents selective cross pollination may be a better solution.

The objective of this study is to assemble re-sequencing data of the *Ga1* region in maize, seeking to further examine the region of interest and potentially expose components that would lead to a greater understanding of how the system functions.

Challenges

The availability of only one published reference maize sequence, B73 v3, was a major disadvantage. The ability to identify possible sequence variation between maize lines, in particular, that from which our BAC libraries was derived, would have been especially useful in this project. The lack of such reference sequences led to difficulties clearly identifying sequence gaps due to sequence variation from that of causative polymorphism. Additionally, not having a mate paired library severely hindered our ability to span gaps in repetitive regions.

The lack of effective alignment tools also posed a major challenge. Scaffold sequence that had a small overlap region with a contiguous scaffold sequence, but could be overlapped manually, prevented a more continuous coverage of our region of interest. Additionally, it was difficult to determine if indeed the sequences should be combined or were a result of a smaller region that was repeated in the region of interest. Inherent challenges included processing large data files, visualizing genomic sequences of great length, and implicit error in gene prediction software's ability to correctly predict genes.

The DNA sequence of the intergenic space of the region of interest is extremely repetitive. Due to the abundance of transposons throughout the region of interest, many reads mapped to multiple positions not only in the region of interest, but also the genome as a whole. This genetic architecture led to difficulties in distinguishing overlapping regions of each BAC in comparison to the other three BACs.

Role of Student Researcher

As student researcher, my role was to use bioinformatics tools to assemble bacterial artificial chromosome (BAC) next generation sequence data. Determination of the region of interest via a mapping study was carried out in the lab of Dr. Michael Muszynski's. Construction of the BAC library was carried out in the lab of Dr. Matt Evans. Selection of the BACs for sequencing was carried out jointly in the labs of Dr. Muszynski and Evans. I processed, aligned, and assembled reads from all BAC files. Computation was performed in a Linux environment. Sequence variation between the BAC sequences and the reference genome was identified. Additionally, in the BAC assembly process, a macro in Microsoft Office Excel was created to analyze overlapping reads mapped to the reference genome. Subsequently, this allowed for the compilation of mapped reads and the determination of an overall start and stop position of contigs, in relation to the B73 v3 reference genome, derived from overlapping reads. The project contributed sequence data, a component of published literature that until recently was absent. This absence impeded understanding of gametophytic incompatibility.

Furthermore, as part of my Master's experience, I served as a coauthor for the maize introduction chapter in the Encyclopedia of Food Grains. This contribution serves not only as writing experience, but also as a source of references for those individuals seeking additional information regarding maize. Together, this research and writing experiences serve as partial requirement for the Masters in Plant Breeding degree.

Thesis Organization

This thesis is organized into four chapters. The first chapter includes an introduction to gametophytic cross-incompatibility and literature review. Chapter two describes work aimed at re-sequencing the *Ga1* region of maize to be published in a peer reviewed journal. A chapter that has been accepted for publication in the Encyclopedia of Grains Science is presented in Chapter Three.

References

Barnabas B., Fridvalszky L. (1984) Adhesion and germination of differently treated maize pollen grains on the stigma, Acta Botanica Hungarica. pp. 329-332.

Bloom J.C., Holland J.B. (2011) Genomic localization of the maize cross-incompatibility gene, Gametophyte factor 1 (ga1), Maydica. pp. 379-387.

Correns C. (1902) *Scheinbare Ausnahmen von der Mendel'schen Spaltungsregel für Bastarde*, Gebrüder Borntraeger.

Darwin C. (1877) The different forms of flowers on plants of the same species, London.

Della Porta G., Ederle D., Bucchini L., Prandi M., Verderio A., Pozzi C. (2008) Maize pollen mediated gene flow in the Po valley (Italy): Source-recipient distance and effect of flowering time, European Journal of Agronomy. pp. 255-265.

Demerec M. (1929) Cross sterility in maize, Molecular and General Genetics. pp. 281-291.

Derksen J., Rutten T., Van Amstel T., A d., Doris F., Steer M. (1995) Regulation of pollen tube growth. *Acta Bot. Neerl* 44:93-119.

Dresseslhause T., Franklin-Tong N. (2013) Male-female crosstalk during pollen germination, tube growth and guidance, and double fertilization, Molecular plant. pp. 1018-1036.

Emerson R.A. (1934) Relation of the differential fertilization genes, *Ga ga*, to certain other genes of the *su-tu* linkage group of maize, Genetics. pp. 137-156.

Evans M.M.S., Kermicle J.L. (2001) Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize. Theoretical and Applied Genetics 103:259-265.

Feijó J., Malhó R., Obermeyer G. (1995) Ion dynamics and its possible role during in vitro pollen germination and tube growth. Protoplasma 187:155-167.

Franklin-Tong N., Franklin C.H. (2003) Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms, TRENDS in Plant Science. pp. 598-605.

Franklin-Tong V.E. (1999) Signaling and the modulation of pollen tube growth. The Plant Cell 11:727-738.

Giampiero C., Moscatelli A., Cresti M. (1997) Cytoskeletal organization and pollen tube growth. Trends in Plant Science 2:86-91.

Goggi S., Caragea P., Lopez-Sanchez H., Westgate M., Arritt R., Clark C. (2006) Statistical analysis of outcrossing between adjacent maize grain production fields, Field Crop Research. pp. 147-157.

Herrero M., Hormaza J.I. (1996) Pistil strategies controlling pollen tube growth. Sexual Plant Reproduction 9:343-347.

Heslop-Harrison J. (1982) Pollen-stigma interaction and cross-incompatibility in the grasses. Science 215:1358-1364.

Heslop-Harrison J. (1987) Pollen germination and pollen-tube growth.

Heslop-Harrison J., Heslop-Harrison Y. (1990) Dynamic aspects of apical zonation in the angiosperm pollen tube. Sexual Plant Reproduction 3:187-194.

Heslop-Harrison J., Heslop-Harrison Y. (1991) Restoration of movement and apical growth in the angiosperm pollen tube following cytochalasin-induced paralysis. *Philos. Trans. R. Soc. London Ser. B*331:225-235.

Jimenez J.R., Nelson O.E. (1964) A fourth chromosome gametophyte locus in maize, Journal of Heredity. pp. 259-263.

Johnson M., Preuss D. (2002) Plotting a course: Multiple signals guide pollen tubes to their targets. Developmental Cell 2:273-281.

Kermicle J. (2006) The *gametophyte*-1 locus and reproductive isolation among *Zea mays* subspecies, Maydica, Maydica. pp. 219-225.

Kermicle J.L., Evans M.M.S. (2005) Pollen-pistil barriers to crossing in maize and teosinte result from incongruity rather than active rejection. Sexual Plant Reproduction 18:187-194. DOI: 10.1007/s00497-005-0012-2.

Kermicle J.L., Evans M.M.S. (2010) The *Zea mays* sexual compatibility gene ga2: Naturally occurring alleles, their distribution, and role in reproductive isolation, Journal of Heredity

Knox R.B. (1984) Pollen-pistil interactions, Encyclopedia of Plant Physiology. pp. 508-608.

Lausser A., Kliwer I., Srilunchang K.O., Dresselhause T. (2010) Sporophytic control of pollen tube growth and guidance, Journal of Experimental Biology. pp. 673-682.

Lu Y., Kermicle J.L., Evans M.M.S. (2014) Genetic and cellular analysis of cross-incompatibility in Zea mays. Plant Reproduction 27:19-29. DOI: 0.1007/s00497-013-0236-5.

Lundqvist A. (1956) Self-incompatibility in rye, Hereditas. pp. 293-348.

Mangelsdorf P.C., Jones D.F. (1926) The expression of Mendelian factors in the gametophyte of maize, Genetics. pp. 423-455.

Mascarenhas J.P. (1993) Molecular mechanisms of pollen tube growth and differentiation. The Plant Cell 5:1303-1314.

Nelson, O.E. 1952. Non-reciprocal cross-sterility in maize. Genetics. p. 101.

Nelson OE. (1994). The gametophyte factors of maize. In: Freeling M, Walbot V, editors. *The maize handbook*. Berlin (Germany): Springer-Verlag. p. 496–503.

Nettancourt D. (1977) Incompatibility in angiosperms, Springer, Berlin Heidelberg.

Pandey K.K. (1958) Time of the S allele action, Nature. pp. 1220-1221.

Pierson E., Lichtscheidl I., Derksen J. (1990) Structure and behavior of organelles in living pollen tubes of Lilium longiflorum. Journal of Experimental Biology 41:1461-1468.

Read S., Clarke A., Bacic A. (1993) Stimulation of growth of cultured Nicotiana tabacum W38 pollen tubes by poly (ethylene glycol) and Cu(II) salts. Protoplasma 177:1914.

Steer M., Steer J. (1989) Pollen tube tip growth. New Phytol 111:323-358.

Takayama S., Isogai A. (2005) Self-incompatibility in plants, Annual Review of Plant Biology. pp. 467-489.

USDA. (2015) Organic Agriculture.

USDA-ERS. (2014) Adoption of genetically engineered crops in the U.S.: Recent trends in GE adoption.

Zhang H., Liu X., Zhang Y.E., Jiang C., Cui D., Liu H., Li D., Wang L., Chen T., Ning L., Ma X., Chen H. (2012) Genetic analysis and fine mapping of the Ga1-s gene region conferring cross-incompatibility in maize, Theoretical and Applied Genetics. pp. 459-465.

CHAPTER TWO: RESEQUENCING OF THE GAMETOPHYTIC INCOMPATIBILITY REGION IN MAIZE

Abstract

Gametophytic cross-incompatibility is as a biological barrier to cross pollination, preventing promiscuity between neighboring transgenic maize and organic maize fields. Interest in deploying gametophytic cross-incompatibility genes in maize to reduce unwanted pollination has fueled recent research on the topic. We identified and sequenced four BACs spanning the *Gamepthotype factor-1* (*Ga1)* locus of a line carrying the *Ga1-m* allele to better understand and characterize the male function in this gametophytic cross-incompatibility system in maize. Comparison of de novo assemblies to assemblies based on the B73 genome scaffold suggest there are extensive differences between the B73 sequence and the genome of the line the *Ga1* region was introgessed from. We therefore focused on de novo assemblies to characterize this region. A de novo assembly was performed for each of the four BACs. Repetitive sequences prevented unambiguous assembly of complete BAC sequences. The resulting contigs were compared to the region of interest in B73 to identify polymorphisms that may be responsible for *Ga1* action. Clear homology was identified between BAC contigs and six predicted genes and one transposable element in the B73 version 3 (v3) reference sequence. Polymorphisms are found in each of these genes. Six additional predicted B73 genes and two transposable elements could not be found in the *Ga1-m* region despite evidence of overlapping BAC coverage of the region in which they are found. In addition, 11 genes were predicted in our de novo assembled contigs that are not predicted

in B73. These sequence differences are candidate polymorphisms for the gametophytic cross-incompatibility function.

Introduction

There are practical applications of gametophytic cross-incompatibility as a biological barrier. It could be especially useful in specialty maize crop systems where controlling xenia effects directly influences the value of the crop. For example, if gametophytic cross-incompatibility systems are incorporated into an organic maize system, organic maize fields could be grown alongside neighboring transgenic maize fields with reduced transgene contamination.

Past studies have resulted in successful fine-mapping of the *Ga1* cross-incompatibility locus. However, to our knowledge, causative polymorphisms or causative genes have yet to be characterized. Using a mapping approach with two populations, Bloom and Holland (2012) mapped *Ga1-s* to a region on the short arm of chromosome 4. Mapping in the population B73 x Hp301 NAM RILs localized the *ga1* interval to 6,408,214 to 12,609,493 bp on the short arm of chromosome 4 in the B73 version 2 reference genome. Additionally, a diverse set of lines for which genotyping-by-sequencing (GBS) data are available were screened at SNP loci within the *Ga1* region for markers that co-segregate for the pollen exclusion phenotype. Two predicted genes homologous to sucrose-phosphate synthase genes in other plants: GRMZM2G068698 and GRMZM2G008507 were identified by this process. The W22 x *Ga1-s Su-1* mapping population delineated the *Gal-s* locus between 7,133,675 and 13,398,777 bp in the B73 AGP version 2 reference sequence. The identified region overlaps with the 2.2 Mbp (million base pairs) region previously identified by Zhang et al. (2012).

More recent studies have further delineated the *Ga1-s* locus. Liu et al (2014) defined the region to 9,491,422 to 9,591,946 bp on the short arm of chromosome 4. The study utilized a homogenous mapping population (*Ga1-s* BC_1F_1) derived from a popcorn line (SDGa25) and a Chinese line carrying the null alleles for gametophytic cross-incompatibility (Jing66), which allowed the authors to further define the region. The need for phenotyping was eliminated by taking advantage of the gametophytic cross-incompatibility system. During the creation of the population, only *Ga1-s* pollen would successfully pollinated *Ga1-s* plants; therefore, the resulting progeny were *Ga1-s/Ga1-s*. The population was screened using 14 closely-linked markers and five tightly-linked markers derived from the B73 version 2 reference genome. The work identified gene GRMZM2G039983 in the B73 reference genome as a potential candidate gene for causation of the gametophytic cross incompatibility system. The predicted gene was found to have homology to WDL1 in *Arabidopsis* which controls anisotropic cell growth and was hypothesized to have an effect on pollen tube growth. The potential role of GRMZM2G039983 has not been elucidated. After identifying a narrow region of interest, the authors demonstrated an integration proof of concept. *Ga1-s* was successfully introgressed into an elite waxy maize hybrid using marker assisted selection. These results illustrate the utility of molecular information about the locus for transfer of this trait among varieties.

Kermicle and Evans (2005) demonstrated that incongruity between pollen and silk, rather than active rejection, is responsible for the *Ga1* function. These results suggest the need for a harmonious interaction between a female factor in the silks with a male factor in the pollen. The *ga1* locus has been classified as a null allele (Kermicle, 2006). It is not understood if the null effect is due to the presence/absence of gene(s) conferring male and female functions or sequence variation(s) in genes in the region. In this project, we use of the *Ga1-m* haplotype as a means to isolate the male function. Isolation and classification of the male function may bring clarity to the role of the female function and assist in better understanding pollen-pistil interaction as a whole. The goal of this study was to understand the *Ga1* locus at the molecular level. This was accomplished by resequencing four BACs derived from a *Ga1-m* variety. Through alignments to the *ga1/ga1* inbred line B73, we seek to identify the inserted and deleted genes, as well as polymorphisms within genes in an identified region of interest. With such information, we hope to deduce how sequence variations could contribute to the male function in gametophytic cross incompatibility.

Materials and Methods

BAC library construction, BAC selection, and sequencing

Dr. Matthew Evans at Stanford University created a BAC library from a W22 inbred line containing the *Ga1-m*, *Ga2*, and *Tcb1* alleles (Kermicle & Evans, 2010). The BAC vector pIndigo-BAC5-Hind III was used in DH10B *E coli* cells. The CopyControl™ BAC Cloning Kit was used to create the BAC library. The BACs had a predicted average insert size of 120 kilobases (kb). Chloramphenicol resistance was used as a selectable marker.

Primers designed to amplify gene sequences found in the B73 region of interest, namely AC184772, GRMZM5G817995, GRMZM2G419836, GRMZM2G027021, and GRMZM2G039983, were used to identify BACs near the *Ga1* locus using PCR (Table 2.1). Primer set GRMZM2G027021 was not successful in identifying a BAC. The other four primers identified a total of one BAC per primer pair. The four BACs will be hereafter referred to as BAC1, BAC2, BAC3, and BAC4. Each BAC was sequenced at the Iowa State University DNA facility using 300 bp single end Illumina Mi-Seq technology.

Table 2.1. Markers used to identify BACs.

BAC	Gene model	Primer sequence	Amplicon size
1	AC184772.3	F: AGCTGTGTGGGGTTCTATGCGAGT R: TAGAATCCTAGCTCCTACAGCGAAGCC	350 bp
2	GRMZM5G817995	F: TCCAACTCTTTTGCTTCTTTTGATGCAC R: CGCAACCTTTGAGTAACTCTTAGC	620 bp
3	GRMZM2G419836	F: CTCCCCTCGTCTGCTTCAAATGGC R: AGAGAACAGAGCACCCAAATCGGC	640 bp
4	GRMZM2G039983	F: AAGCAGCGCTGCACAGTGGCAA R: AAGCTGGGCAGGAGGAAGACGG	578 bp

Preparing sequence reads for assembly

Unless otherwise noted, all bioinformatics work was completed on the USDA server, Lathyrus. The server is a Linux based system with 64 central processing unit (CPU) cores. It is maintained by the Corn Insects and Crops Genetics Research Unit located on the Iowa State University campus.

In the first step of processing the BAC sequence files, scythe was used to remove adapter sequences from reads (Buffalo, 2014). The sickle plugin was used to trim bases with a quality score of less than 20 and reads shorter than 50 bp in length. Using the FASTX-Tookit, fastx_trimmer was used to remove the first 15 bases of each read due to low quality base calls in that region (Pearson et al., 1997). Unique identifiers replaced original reads names. The deconseq plugin was used (Schmieder et al., 2011) to remove contaminating sequences derived from *Escherichia coli* (*E. coli*). Reads that matched the *E. coli* genome at 95% identity or better, with greater than 5% coverage, were deleted.

The remaining sequences were considered high quality reads. High quality reads averaged 280 bp in length.

Scaffold-based assembly of BAC sequences

High quality reads were aligned to chromosome 4 of the *Zea mays* v3 reference genome, obtained from Ensembl Plant (Julian et al, 2014), by BAC. Processed read files were subjected to the Burrows-Wheeler Aligner (BWA) pipeline (Li & Durbin, 2009). Post alignments, reads were divided into two groups: (1) reads that mapped to the region of interest and (2) reads that did not map to the region of interest.

Analysis of mapped reads

Positional data from mapped reads was extracted and used to identify sequences that overlap. Overlapping read sequences were formed into contigs by determination of contig start and stop positions on the reference genome; these contigs will be referred to as mapped contigs, hereafter. Mapped contigs were visualized on a custom track using the MaizeGDB Genome Browser (Figure 2.4) (Monaco et al., 2013).

Analysis of unmapped reads

The unmapped reads were subjected to de novo assembly using the MIRA 4.0.2 program (Chevreux et al., 1999) and the resulting contigs will be referred to as unmapped contigs hereafter. Parameters used in the MIRA 4 manifest file are as follows:

job = genome, denovo, accurate

parameters = -GE:not=16 (16 general number of threads)

parameters = SOLEXA_SETTINGS -CO:msr=no (tells MIRA to not merge identical reads to backbone, thus maintaining distance and orientation information)

technology = solexa

Nucleotide-Nucleotide BLAST 2.2.30+ (blastn) (Altschul et al., 1990) was used to compare unmapped contigs to the region of interest in B73. An e-value of 0.0001 was used and the default value was used for all other blastn parameters.

De novo assembly of all high quality reads by BAC

Parameters used in the MIRA manifest file are identical to those used above for assembly of the unmapped reads, except all high quality reads from each BAC were assembled separately to give four sets of contigs, one from each BAC.

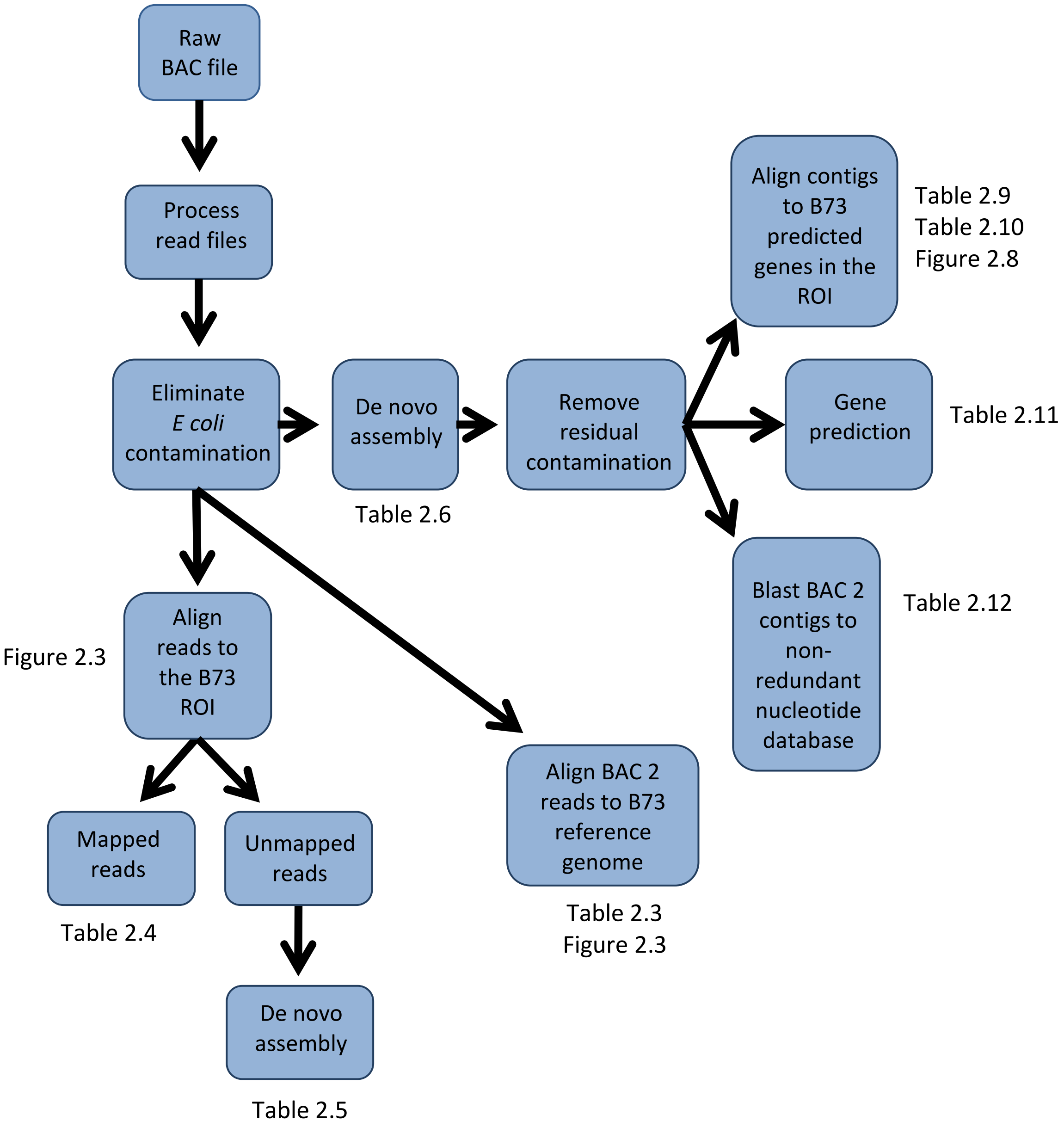


Figure 2.1. BAC assembly pipeline.

Gene prediction

Assembled contigs 5 kb (thousand basepairs) and greater in length were subjected to gene prediction using Softberry website FgenesH (Salamov & Solovyev, 2000). FgenesH ab initio gene prediction is based on monocot plant specific, trained parameters.

Gene annotation

Assembled contigs 5 kb or greater in length were blasted to the NCBI non-redundant nucleotide database. The following parameters were used: expected threshold: 10; Mismatch score: 1-2; Gap cost: linear; automatically adjust parameters for short sequences allowed. (Altschul et al, 1990). Threshold values used to declare significance were an e-value of 0.0, identity score of 15% and greater, and a query coverage of 85% and greater.

Removal of residual contamination

MIRA 4 assembly files from each BAC were blasted to the Univec database to identify residual sequence contaminates. BAC contigs that blasted to entries in the database with an e-value of .0001 or less were removed.

Results and Discussion

Identification of the region of interest

Studies completed by Bloom and Holland (2012), Zhang et al. (2012), Liu et al. (2014), and unpublished work by Dr. Michael Muszynski, identified a region likely to contain the *Ga1* locus. In this study we used a region of interest from 9.1 to 9.6 Mbp on the short arm of chromosome 4. In the B73 v3 reference genome, there are six protein coding genes and six low confidence genes characterized, ranging from 113 bp to 9,640 bp in length. Additionally there are three transposable elements situated in the latter half of the region that range from 556 to 56,722 bp in length. Figure 2.2 summarizes the region of interest. Table 2.2 presents additional details of the region of interest including the model type of each gene (low confidence, protein coding, or transposable elements), as well as, the start and stop position and orientation.

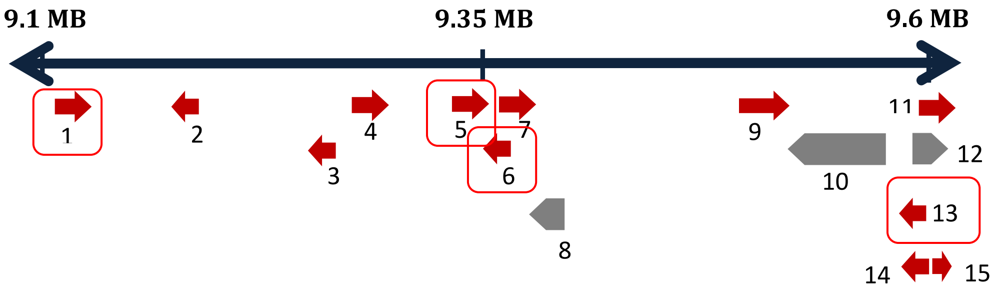


Figure 2.2. Predicted gene model in the region of interest for B73 v3 reference genome. Red arrows represent genes; grey arrows represent transposable elements. Boxed genes were used in BAC marker sequences.

Table 2.2. Position, strand, and model type of predicted genes in the B73 v3 reference genome.

Gene #	Gene	Model type	Start	Stop	Strand
1	AC184772.3_FG001	LC	9,106,014	9,106,855	Forward
2	AC201986.3_FG002	PC	9,187,173	9,187,685	Reverse
3	GRMZM2G702344	PC	9,263,791	9,264,870	Reverse
4	GRMZM2G122484	LC	9,272,045	9,272,566	Forward
5	GRMZM5G817995	PC	9,329,468	9,329,770	Forward
6	GRMZM2G419836	PC	9,355,159	9,358,375	Forward
7	AC205010.4_FG001	LC	9,358,025	9,359,830	Reverse
8	GRMZM2G535727	TE	9,375,747	9,375,860	Reverse
9	GRMZM2G027021	PC	9,490,258	9,499,402	Forward
10	GRMZM2G027368	TE	9,517,545	9,574,267	Reverse
11	AC204382.3_FG010	LC	9,588,810	9,589,611	Forward
12	GRMZM2G507805	TE	9,589,653	9,590,209	Reverse
13	GRMZM2G039983	PC	9,594,061	9,597,440	Reverse
14	GRMZM2G039971	LC	9,597,755	9,598,020	Reverse
15	GRMZM2G039928	LC	9,598,535	9,599,547	Forward

LC= low confidence; PC= protein coding; TE= transposable element

The BAC library was screened using the primers found in Table 2.1. Primer sequences originated from predicted genes in the region of interest. Marker placement is shown in Figure 2.7. We verified the presence of marker sequences in assembled contigs.

Four BACs were identified as containing molecular markers in the region of interest. Post sequencing, BAC 1 yielded 3,526,222 reads; BAC 2 yielded 4,995,350; BAC 3 yielded 1,849,985; BAC 4 yielded 2,472,846 reads. Average read length after processing is 280 bp.

We first sought to determine what proportion of reads mapped to the entire B73 reference genome, or if they mapped to the genome at all. We used the genome alignment exercise to verify that the BACs were derived from the region of interest. Since BAC 2 generated the largest number of reads and was hypothesized to reside in the middle of the region of interest, it was selected for this analysis. Visualization of BAC 2 reads mapped to the entire B73 reference genome, using BWA, revealed that the highest density of reads is indeed within the region of interest located on chromosome 4 (Figure 2.3). Homology to BAC sequences was found outside of the region of interest as well.

Read mapping outside of the region of interest could be the result of one or more of the following: 1) reads map to repetitive regions found inside and outside of the region of interest, 2) the region of interest in the *Ga1-m* haplotype is smaller than B73; BAC sequences, therefore, extend out of the region of interest defined by the B73 reference genome, and/or 3) the sequences of the BACs may differ from that of B73. These variations could be the result of genome rearrangements where sequences are not deleted from the genome, but simply moved to a new genomic location (Springer et al., 2009).

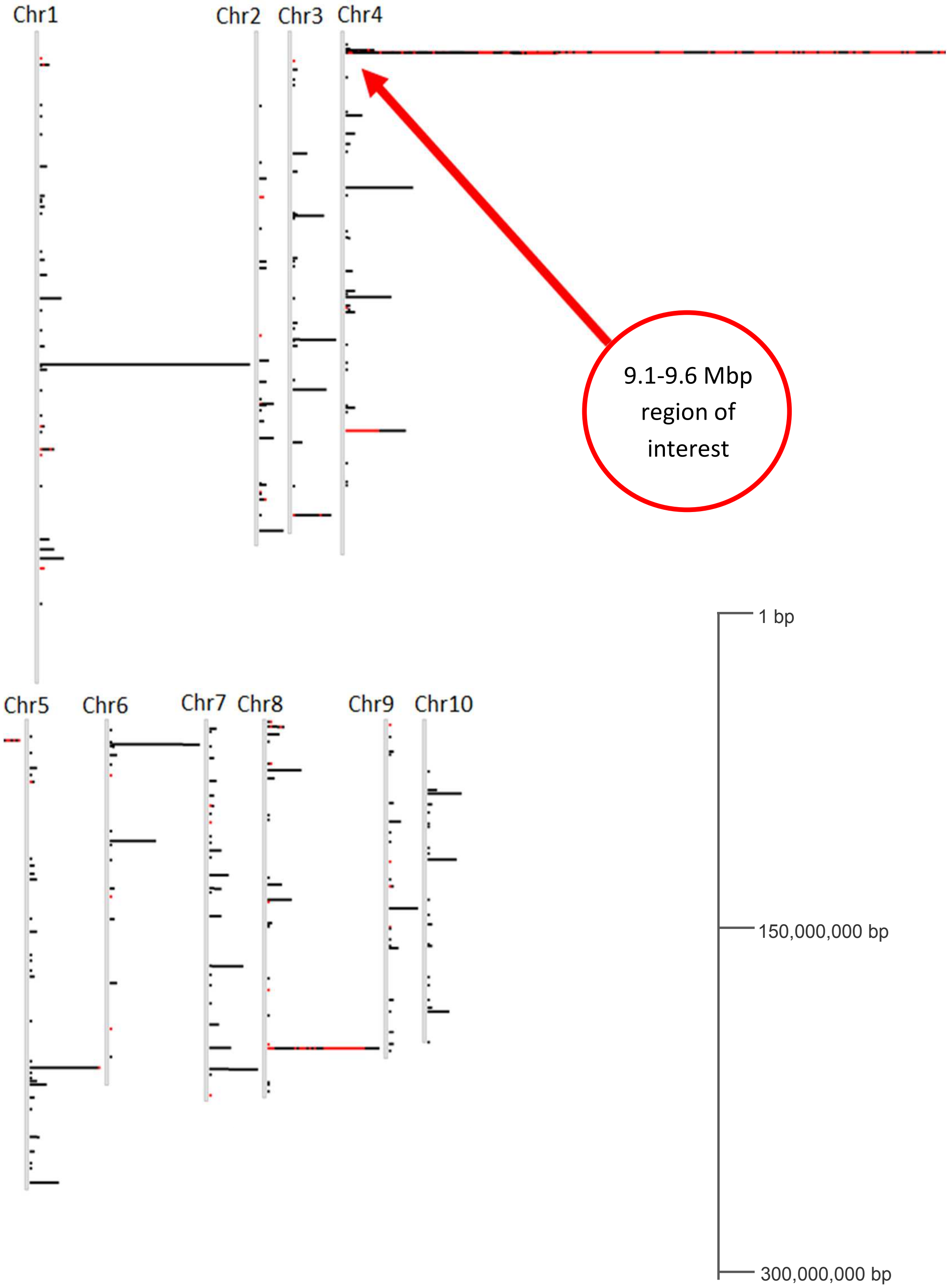


Figure 2.3. Visualization of BAC 2 reads mapped to the entire B73 genome.

Of all BAC 2 reads, less than 3% mapped to regions outside the region of interest. 20% of total BAC 2 reads mapped to the genome. These results can be seen in Table 2.3. It was therefore concluded that BAC 2 originated from the identified region of interest. Some of the reads not mapping to the genome could be the result of sequence differences in *Ga1-m* and not in B73. Residual contamination may also have resulted in unmapped reads.

Table 2.3. BAC 2 BWA alignment to the B73 genome vs region of interest.

	Number of reads	**Percent of total reads**
Total reads	4,995,350	
Reads mapped to region	889,424	17.8%
Reads mapped outside of region	122,932	2.5%
Reads mapped to genome	1,012,356	20.3%

High quality reads per BAC were then mapped to the region of interest of B73 using BWA. The distribution of the mapped reads is shown in Figure 2.4. A small proportion of reads map to locations across the region of interest. We believe this result is once again due to the mapping of repetitive reads. The majority of the aligned reads for each BAC fall within the same genomic region as the marker sequence (and corresponding gene) used to select the BAC, verifying hypothesized BAC order. The distribution of the mapping locations of reads from each BAC suggest that BACs do originate from the region of interest and do so in an overlapping BAC arrangement. Collectively, we conclude the following BAC order: BAC 1, BAC 2 and BAC 3, and BAC 4, with BAC 3 falling within BAC 2.

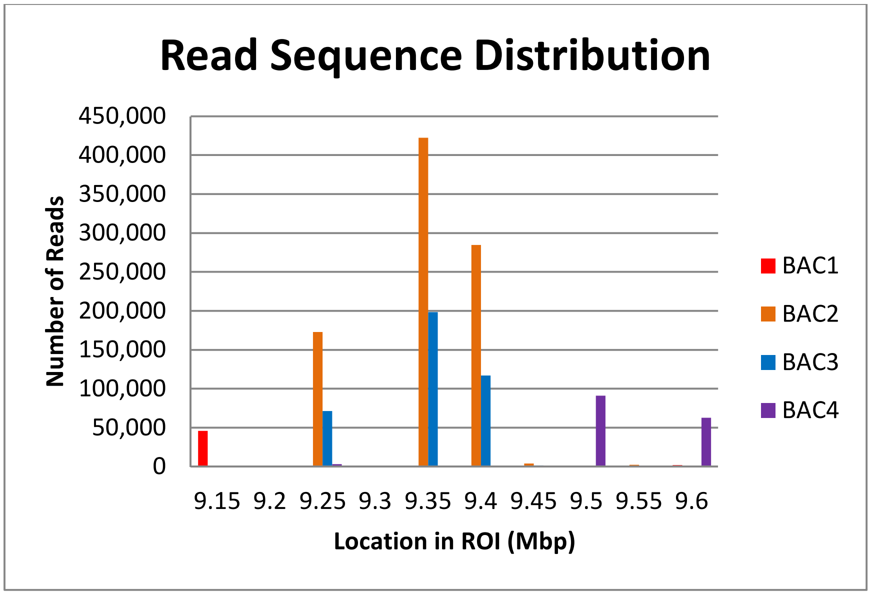


Figure 2.4. BAC read sequence distribution over the B73 region of interest.

Two different approaches (Figure 2.1) were used to assemble the sequence reads. The first approach was a comparative genome assembly. Reads are first mapped to the B73 reference genome and those that did not map to the region were subjected to de novo assembly. This was accomplished as follows.

Step One: Read files from each BAC were aligned to the region of interest using BWA. The percentages of reads that map to the region are presented in Step 1 of Table 2.4. BAC 1 (1.4%) and BAC 4 (6.4%) have a much lower percentage of mapped reads compared to BAC 2 (17.8%) and BAC 3 (21.1%). A lower quantity of mapped reads from BAC 1 and BAC 4 and alignment of BAC 1 and 4 to the boundaries of the region of interest suggest that these BACs extend out of the region of interest.

Location and number of reads mapped to the reference genome are not identical across BAC sequences; however, some mapped regions are shared between BACs. These similarities and differences in coverage suggest overlap, but of four distinct BACs.

Some part of the region of interest contained no mapped reads.. We believe these gaps in coverage are the result of sequence differences between the BACs and B73.

Step Two: The reads that did not map to the region were subjected to de novo assembly (Step 2 in Table 2.5). Compared to the mapped reads, the de novo assembled reads resulted in contigs with greater overall length. The percentage of unmapped reads assembled into contigs from each BAC ranged from 16%-28% (ranked from lowest to highest: BAC 3, 4, 2, 1). BAC 1 and BAC 2 have slightly higher percent read usage and a substantially larger number of total contigs (1,083 and 1,249); however, average contig length is on average 1.5 fold smaller (757 bp and 755 bp). Data suggest that assembly of BAC 1 and BAC 2 unmapped reads resulted in a myriad of short contigs that cannot be assembled into longer contigs. Fewer unmapped reads were used in BAC 3 and BAC4 compared to BAC 1 and BAC 2. The number of contigs is smaller (213 and 360); however, average contig length is much higher (1,170 bp and 1,084 bp), possibly a result of fewer mapped reads being removed.

Overall, unmapped reads yielded a greater number of contigs that are, on average, over 2.5 fold longer than mapped contigs. Additionally, more unmapped reads are assembled in comparison to mapped reads. Interestingly, the number of reads per contig base for mapped contigs is much higher than unmapped contigs. This result may be due to mapped contigs representing reads that are derived from repetitive regions. For example, if there are two similar regions in the region of interest (repetitive region 1 and repetitive region 2), the reads derived from repetitive region 1 and reads derived from repetitive region 2 will both map to region 1. This would cause the coverage of such repetitive regions to be artificially inflated. This may explain why the coverage of the mapped contigs is high. This would also suggest that the unmapped contigs are unique sequences and may be also unique to the *Ga1-m* genome.

Table 2.4. Summary of comparative genome assembly: Mapped reads.

Step 1: BWA- Identifying mapped reads

	BAC 1	**BAC 2**	**BAC 3**	**BAC 4**	**Total**
# Rds/BAC	**3,526,222**	**4,995,350**	**1,849,985**	**2,472,846**	**12,840,834**
# Rds mapped to region of interest	51,063	889,424	390,147	159,069	1,489,703
% Rds mapped	1.4%	17.8%	21.1%	6.4%	11.6% (Avg)
# Contigs	209	180	124	119	632
# Rds used in contigs	50,508	889,361	390,083	159,000	1,458,952
% Rds used	99%	>99%	>99%	>99%	>99%
Avg contig length (bp)	181	313	407	382	301
Avg #Rds/contig	242	4,941	3,146	1,336	2,308
Total length of contigs	37,752	56,396	50,421	45,474	190,043
# Rds/contig base	1.3	15.8	7.7	3.5	7.7

Table 2.5. Summary of comparative genome assembly: Unmapped reads.

Step 2: De novo assembly of unmapped reads

	BAC1	**BAC2**	**BAC3**	**BAC4**	**Total**
# Rds/BAC	3,475,159	4,105,926	1,459,838	2,313,777	11,351,131
# Rds in contigs	558,957	725,206	232,077	648,703	2,164,029
% Rds used	16.1%	17.7%	15.9%	28.0%	19.1% (Avg)
# Contigs	1,046	1,292	248	405	3,068
Avg contig length (bp)	764	748	1,083	1,026	811
Avg # rds/contig	534	561	936	1,602	705
Total length of contigs	799,118	966,157	268,669	415,684	2,487,382
# Rds/contig base	0.7	0.8	0.9	1.6	0.9

Visualization of the positions of the mapped reads in the region of interest, aligned with BWA, is shown in Figure 2.4. The gaps between clusters of mapped reads suggest there are many differences between the B73 reference genome and the BAC sequences. Sequence variation among maize lines is known to exist (Fu & Dooner, 2002). Not only organization of gene sequences, but also intergenic retrotransposon sequence can drastically differ between inbred lines (Fu & Dooner, 2002; Springer et al., 2009). Sequence differences between B73 and the *Ga1-m* haplotype could explain why a limited number of reads from the BAC files successfully aligned to the reference sequence.

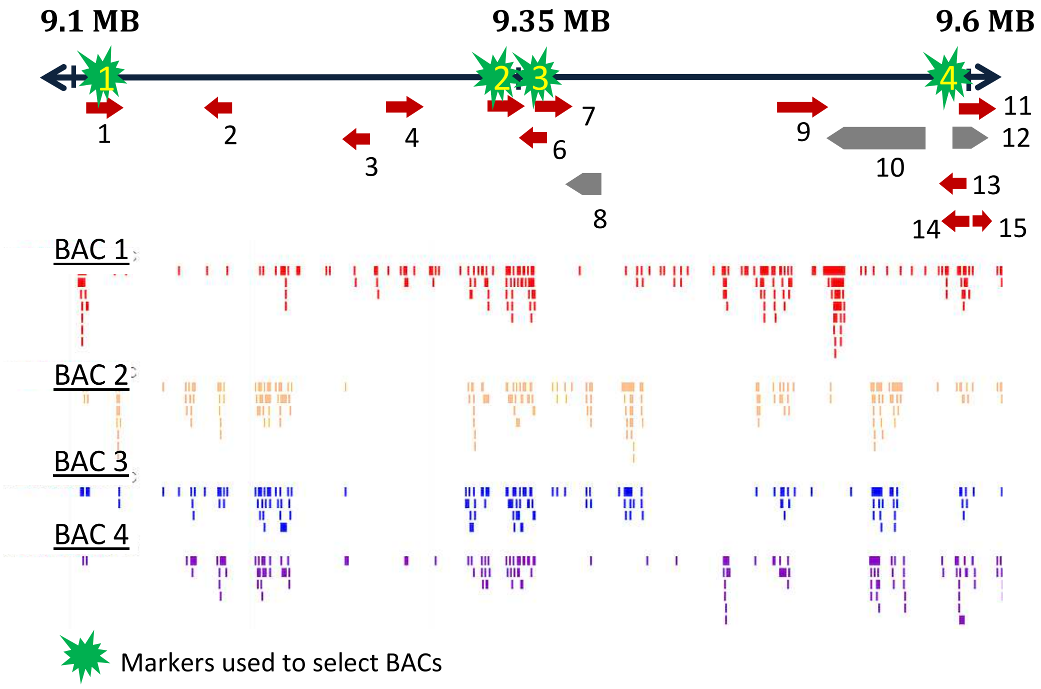


Figure 2.5 Visualization of reads mapped with BWA to the region of interest in the B73 genome. Red arrows represent genes; grey arrows represent transposable elements. *Mapped contigs are not drawn to scale.

Despite mapped reads from each BAC file resulting in coverage throughout the region of interest, this coverage was very sporadic and was separated by many areas of no coverage at all. The result was many small contigs and many unassembled reads. De novo assembly of unmapped reads allowed us to fill in gaps in the alignment. We therefore concluded that the extent of sequence differences between the BAC sequences and the B73 genome was too great to obtain an accurate assembly with the comparative genome assembly approach. One possible problem with assembling the mapped and unmapped reads separately is that that neither set contains the reads necessary to assemble large contigs. Our results suggest that the BACs consist of sequences that map to the reference genome, frequently interspersed with sequences that don't map to the reference genome. We reasoned that it may be possible to obtain longer contigs by assembling all of the reads from a BAC in one de novo assembly. The use of a reference genome has been used in previous research to address such situations (Pop et al., 2003). However, as shown in the mapping results of this project, there exists too much sequence variation and possible genome rearrangements for the use of a reference genome to be of much benefit in assembly the BAC sequences. A new assembly approach was sought.

De novo assembly by BAC

Our approach was to assemble the BAC files as completely as possible without aid of the reference genome and then align the resulting contigs to homologous portions of the reference genome. Post assembly comparison of the B73 reference genome and BAC contigs should highlight sequence differences that are candidates for causative polymorphisms responsible for gametophytic cross-incompatibility. Therefore, each BAC file was subjected to individual de novo assemblies.

Compared to the data derived from the mapped and unmapped assemblies, whole BAC de novo assembly yielded contigs that were much longer with greater total contig length. Assembled contigs were blasted against the Univec database by BAC to determine the extent of residual contamination. A total of 289 assembled contigs were identified as containing contamination. A total of 459,690 bp of contaminants were removed across all BAC files. Remaining contigs are believed to be of high quality BAC sequences.

BAC reads were screened for *E coli* sequence using deconseq and the *E coli* genome before assembly; however, residual contamination remained at the read level. It is possible that the cloning vector, DH10B *E coli*, contain a slightly different genome than that found in the *E coli* database. Genome variation could cause contamination to not be fully removed. Contig contamination was also derived from Enterobacteria. The Enterobacteria classification extends to include more genera of bacteria than *Escherichia*, such as *Salmonella* and *Shigella,* and could further explain why all contamination was not removed. Furthermore, if sequencing errors were present in the reads, accurately identifying *E coli* sequences at the 95% identity may become difficult and may lead to the reads not being removed.

Combined BAC de novo assembled contig length, after the removal of contaminated contigs, totaled 2,109,499 bp. BACs appeared to overlap substantially to cover the entire region of interest, suggesting actual length of the region is lower than combined total contig length. Individual de novo assembly results before and after sequence contamination removal can be seen in Table 2.6. The distribution of contig lengths for each BAC file demonstrates that the assembly process yielded many small contigs. Across BAC files, contigs 5 kb and greater accounted for approximately 0.4% to 2.5% of total contigs per BAC files. These results can be seen in Figure 2.6.

Table 2.6. Results from individual de novo assemblies of BAC files before and after contamination removal.

Before contaminate removal	BAC 1	BAC 2	BAC 3	BAC 4
Total number of reads	3,526,222	4,995,350	1,849,985	2,472,846
Number of reads assembled	557,271	695,794	199,820	333,466
Total length of contigs (bp)	867,582	1,026,154	261,723	419,774
Number of contigs	1,113	1,380	235	381
Largest contig (bp)	38,064	25,649	78,062	36,895
Average coverage	179	406	377	353
After contaminate removal				
Total number of reads	2,968,951	4,299,556	1,650,165	2,139,380
Number of reads assembled (bp)	354,926	348,137	12,676	262,507
Total length of contigs (bp)	750,013	871,928	141,782	345,776
Number of contigs	1,028	1,257	192	343
Largest contigs (bp)	17,149	25,649	6,150	36,895
Average coverage	591	700	3,265	370

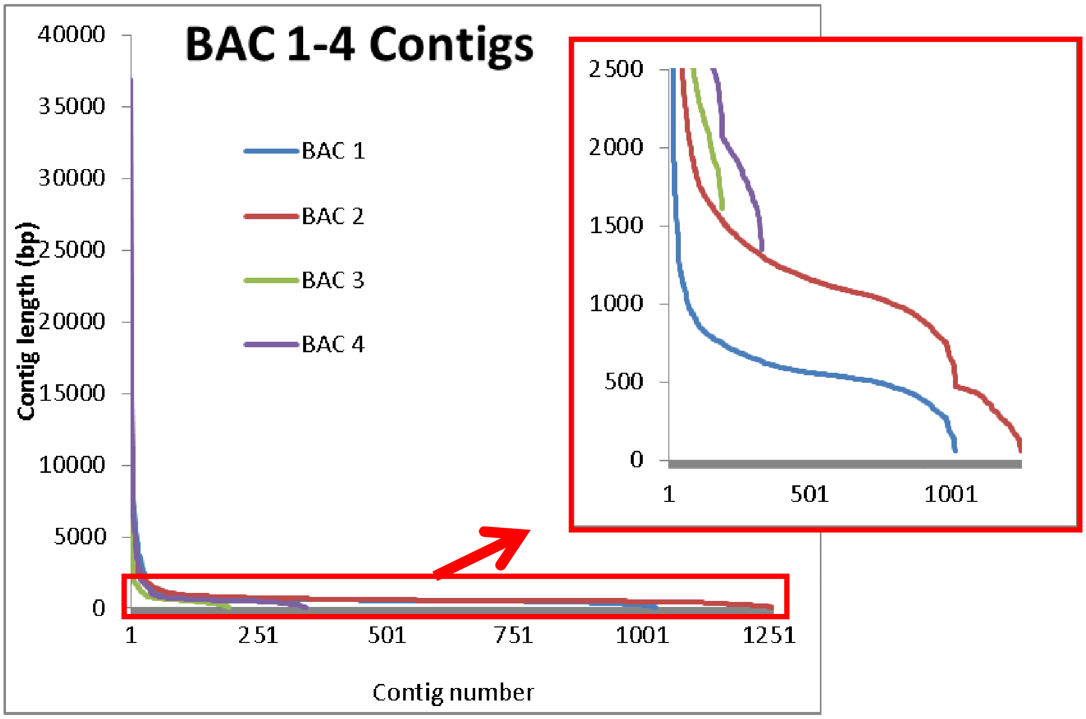


Figure 2.6. Distribution of contig length of MIRA 4 assembled contigs by BAC after contamination removal.

Despite the creation of longer contigs in individual de novo assemblies, there remained a large number of reads that were not assembled, as well as, many short contigs. Contig breaks and unassembled reads may be the result of repetitive regions (interspersed between successfully assembled regions) that were difficult to assemble. De novo assembly literature refers to repetitive reads as the biggest impediment to assembly (Phillippy & Schatz, 2008). MIRA 4 aborts contig extension, labeling the contig with "rep", in a situation in which one or more reads could be used to extend the contig and/or the contigs contains repeative sequence. Instead of inaccurately assembling sequences, contig extension is stopped. In the BAC files (1-4), the percentage of "rep" contigs in comparison to total number of contigs are as follows: 76%, 79%, 55%, and 60%. These numbers demonstrate that over half of the contigs from each BAC file was stopped due to occurrences of repetitive regions. These results likely explain why the assembly yielded many contigs.

Several MIRA4 parameters were altered in attempt to optimize the assembly and utilize more reads to create longer contigs. These results can be seen in Table 2.7. The altered parameters did change the MIRA 4 output; however, no substantial effects were observed. Because of the lack of significant improvements, we went forward with the more stringent parameters from the initial assembly.

Table 2.7. Parameter optimization for de novo assembly.

Denovo assembly	Parameters altered	Total reads	Longest contig	Number of contigs	Total length	N50 contig size	Total avg coverage	Notes
1		694,468	25,773	1,404	1,043,218	656	1651.83	
2	AS:ard=no	693,875	31,448	1,412	1,044,318	669	1690.50	automatic read detection
3	AS:urd=no	692,255	31,448	1,434	1,058,057	664	1596.03	uniform read distribution
4	AL:mo=10	746,108	30,602	2,048	1,420,361	636	1765.04	minimum overlap
5	HS:ldn=no	693,488	42,262	1,442	1,064,815	673	1599.73	mask repeats in reads; small reads will not span repeats and will be put in debris file
6	SK:percent_ required=50	691,900	26,740	1,460	1,079,847	668	1649.28	controls relative % of exact matches for overlap (typically in sync with –AL:mrs)
7	SK:percent_ required=30	694,184	28,900	1,408	1,046,836	671	1635.38	compare to above
8	HS:mnr=yes	692,672	26,741	1,447	1,071,580	670	1637.84	mask nasty repeats
9	AL:mrs=75	662,567	32,662	2,365	1,641,010	655	1601.24	minimum relative score (typically set at 95)
10	AL:mo=10 Hs:ldn=no AL:mrs=75	710,363	22,257	3,380	2,300,261	657	1593.06	

Contigs that cannot be increased in length and remaining unassembled reads could be the result of several situations. Sequencing errors in the read files may exist. Despite trimming reads to increase read quality, the files may remain error prone. Sequencing errors would prevent overlapping reads from being assembled. If overlapping reads have especially high coverage, or are repetitive in nature, MIRA4 would not assemble these regions either. Classification or repetitive reads shown in Table 2.8 suggests that many reads have been indeed tagged as "crazy" repeats and "nasty" repeats. Most reads assembled had average coverage; however, some reads did have above average coverage. Small contigs of low coverage could also be problematic and lead to a higher number of contigs. Our data, however, does not suggest that low coverage is a problem in the assembly.

Table 2.8. Read coverage and repeat classification.

BAC	Coverage classification			Repeat classification			
	HAF2	HAF3	HAF4	HAF5	HAF6	HAF7	MNRr
1	0	479,061	41,440	0	1,342	3,527,253	3,554,658
2	0	619,285	62,513	0	1,771	4,968,939	1,435,064
3	0	210,861	0	0	278	1,925,053	1,926,413
4	0	337,760	28,398	0	671	2,585,096	2,590,338

HAF5-reapeat; HAF6-heavy repeat; HAF7-crazy repeat

HAF2-low coverage; HAF3-average coverage; HAF4-above average coverage

Furthermore, contamination at a different molar concentration than the BAC DNA, as well, chemical-physical properties of the genome (such as GC rich regions) could lead to erroneous, biased coverage and lead to assembly challenges. Additionally contig alignments to the B73 genome reveal that overlapping contigs marked as "rep" are not assembled into a consensus sequence during the de novo assembly. Because of this, and the high number of "rep" contigs present, many overlapping contigs are not combined. This may explain the high number of contigs present in the assembly and why total contig length exceeds the length of the region of interest.

We next sought to compare differences in the de novo assembled contigs and the B73 reference genome in an effort to identify candidate gene polymorphisms responsible for gametophytic cross-incompatibility. We first used BWA to align genes predicted in B73 to our BAC contigs (Table 2.9).

Table 2.9. Genes from B73 reference genome present in BAC contigs determined by BWA alignment.

Gene #	**Predicted genes in reference genome**		**BAC 1**	**BAC 2**	**BAC 3**	**BAC 4**
1	AC84772.3	LC	X	X		X
2	AC201986.3	PC				
3	GRMZM2G702344	PC				
4	GRMZM2G122484	LC				
5	GRMZM5G817995	PC	X	X	X	X
6	GRMZM2G419836	PC	X	X	X	X
7	AC205010.4	LC				
8	GRMZM2G535727	TE				
9	GRMZM2G027021	PC				X
10	GRMZMG027368	TE	X	X	X	X
11	AC204382.3	LC				
12	GRMZM2G507805	TE				
13	GRMZM2G039983	PC		X		X
14	GRMZM2G039971	LC		X		X
15	GRMZM2G0339928	LC				

LC: low confidence; PC: protein coding; TE: transposable element
*Shaded cell indicates a previously identified putative gene by Liu et al. (2014).

Alignments of predicted gene sequences from the region of interest and assembled contigs from each BAC reveal sequence homology with genes 5 (GRMZM5G817995) and 6 (GRMZM2G419836) to all BACs. Predicted gene 1 (AC184772.3), 9 (GRMZM2G027021), 10 (GRMZM2G027368), 13 (GRMZM2G039983), and 14 (GRMZM2G039971) also clearly align to assembled contigs from at least one BAC. The functions of these genes have yet to be determined;

however, 3 of the genes we found in the BAC sequences do contain characterized conservative domains. Predicted gene 1 (AC184772.3) contains a thioredoxin-like fold conserved domain; predicted gene 9 (GRMZM2G027021) has a GTP-binding protein hgIX domain; and predicted gene 13 (GRMZM2G039983) has an XKlp2 targeting protein conserved domain.

Genes 2 (PC), 3 (PC), 4(LC), 8(TE), 11(LC), 12(TE), and 15(LC) have no recognizable homology to any BAC contigs. If the BACs overlap, as the data suggests, these genes may be absent from the *Ga1-m* haplotype and may be contributors to the gametophytic incompatibility phenotype. Alternatively, they may be found elsewhere in the genome.

Alignment information was used to predict a BAC order as seen in Figure 2.7. Our data suggests that BACs overlap to cover the entire region of interest. We hypothesize the following BAC arrangement: BAC 1, 2, 3, and 4 and BAC 3 falls within BAC 2.

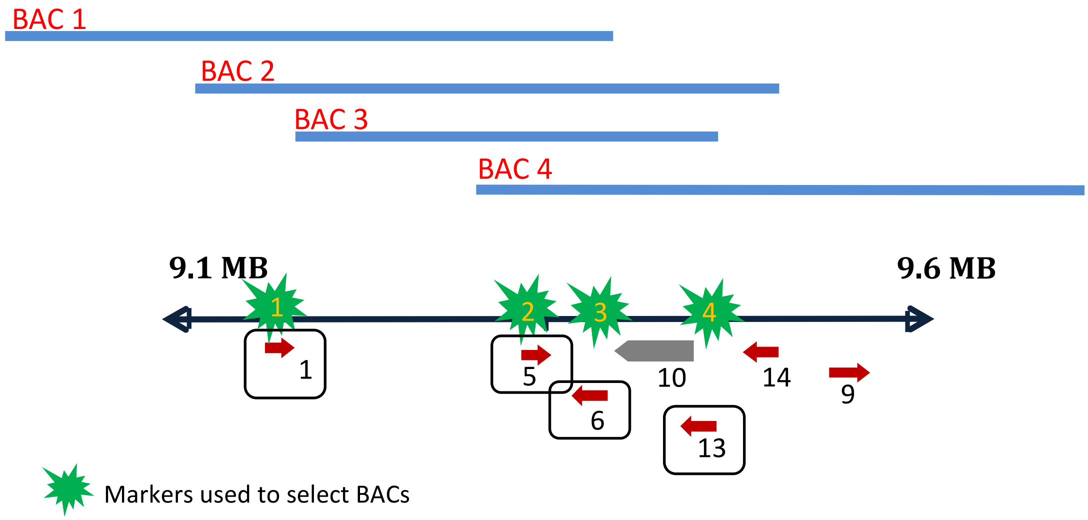


Figure 2.7. Hypothesized arrangement of BAC sequences and presence of B73 predicted genes in the *Ga1-m* haplotype. Boxed arrows indicate genes where BAC markers originated.

Figure 2.7 illustrates the contig alignment data with the predicted genes in the B73 genome. Interestingly, gene 9 (GRMZM2G027021) aligns to only BAC 4. This result suggests the gene is found in the right most boundary of the region of interest, possibly due to reorganization of the *Ga1-m* haplotype. Furthermore, marker sequence 1 and gene 1(AC184772.3) are found in BAC 1, 2, and 4. If the predicted BAC order is correct, this result suggests that gene AC184772.3 is either duplicated within the region or the gene is found downstream of its predicted location in B73. Gene 13 (GRMZM2G039983), previously annotated as a putative causative gene by Liu et al. (2014), was also identified in our BAC sequences. It is highlighted in Table 2.9.

The region of interest originally identified by Liu et al. (2014) was determined using the B73 version 2 reference genome. We sought to determine if the region remained identical in the current, B73 version 3 genome. The region identified in the version 2 genome was between markers dCS1 and ID7 from

9,491,422 to 9,591,946 bp. The version 2 region contained genes GRMZM2G027021, AC204382.3_FG010, and GRMZM2G039983. We could not find marker dCS1 (as published: TCTGTGGAGCTTTGATAAGC) in either version 2 or version 3; however, we could find the following sequence: TCTGTGGAGCTTTGATTGC. Using the identified sequence and the ID7 marker sequence, we identified the region of interest in the B73 version 3 genome to be from 9,496,453 to 9,596,169 bp on chromosome 4. The region contains genes GRMZM2G027021, GRMZM2G027368, and AC204382.3_FG010. The putative gene identified by Liu et al. (2014) is no longer present in the region of interest.

BAC 2 assembled contigs from our research appeared to span the approximately 100 kb region of interest. To identify sequence differences between the BAC sequences and the B73 reference genome, BAC 2 assembled contigs were aligned to the region. A total of 664 contigs aligned to the 100 kb region. A total of 32 contigs covered the region with a total length of 26,696 bp. The alignment suggests no coverage in some parts of the region. Lack of coverage could be due to 1) large sequence deletions in the BAC sequences resulting in a region that is smaller than that found in B73 or 2) reads that remained unassembled could fill in regions with no coverage.

We next determined the presence of polymorphisms in each gene alignment. Table 2.10 describes insertions and deletions found within the gene alignments (see appendix for additional information on alignments). Polymorphisms led to missense mutations, frameshift mutations, and premature stop codons in the protein sequences. Closer observation of GRMZM2G027021 alignment with BAC sequence reveals possible transposon activity. A deletion starting at bp 13,125 flanked by inverted terminal repeats are suggestive structures of the Ac/Ds transposon system. Such observed changes in predicted protein structure may lead to altered function which may underlie the causative polymorphisms of the gametophytic cross-incompatibility system. The gene is also protein coding found both in the version 2 and version 3 100 kb B73 region of interest. Genes within the identified region in B73 have yet to be annotated. Therefore, we can conclude that we did find sequence polymorphism in the BAC sequences; however the extent of those polymorphisms cannot yet be determined.

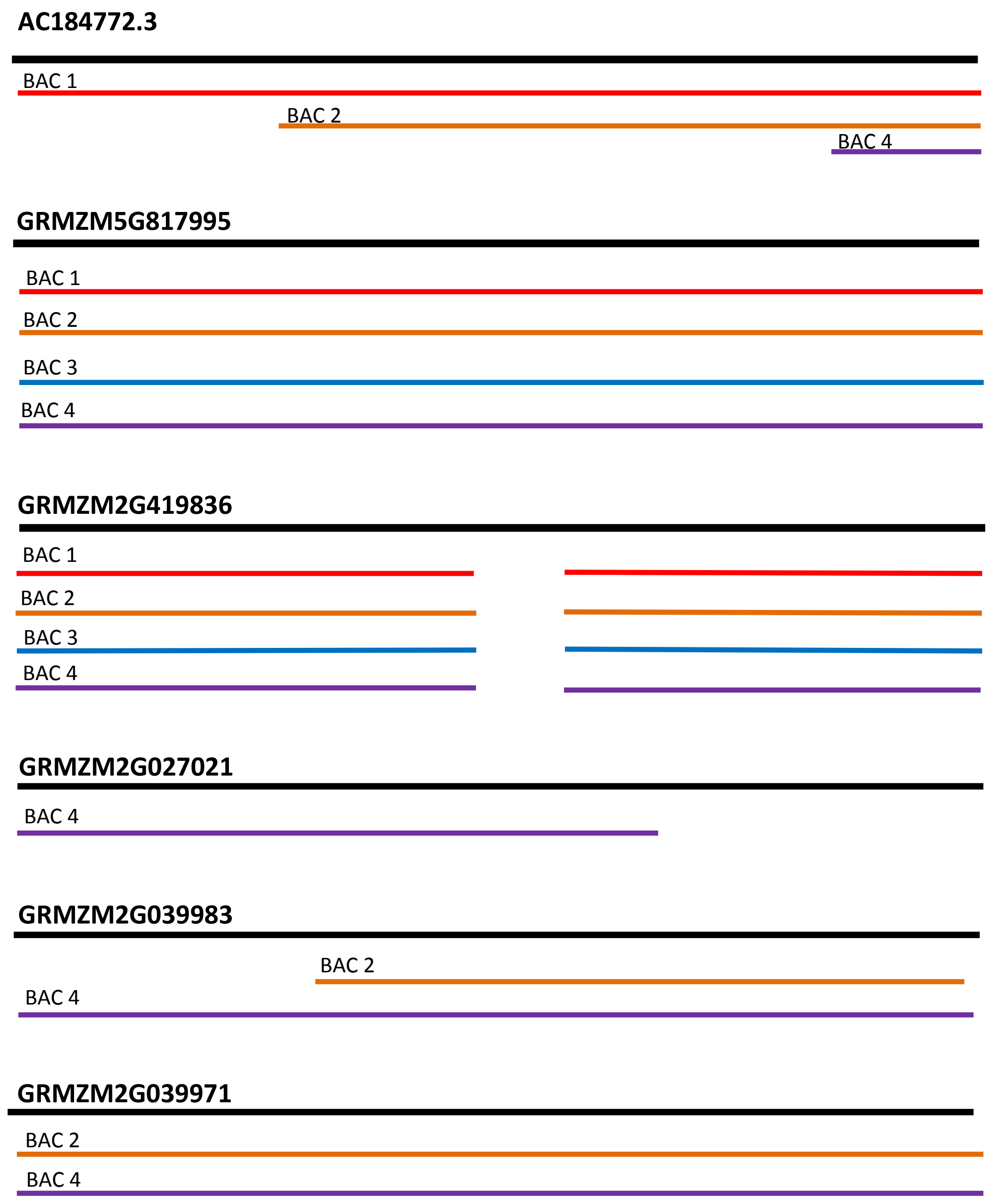


Figure 2.8. Contig alignment to predicted genes in region of interest.

Table 2.10. Polymorphisms between B73 and BAC de novo assembled contigs.

Gene #	Gene ID	Total bp inserted	Total bp deleted	Change in protein length (aa)	Impact on translated product
1	AC184772.3	8	13	-1	Missense Frameshift
5	GRMZM5G817995	0	0	0	Missense
6	GRMZM2G419836	7	278	+165	Missense Nonsense Frameshift
9	GRMZM2G027021	176	7,791	+869	Missense Nonsense Frameshift
13	GRMZM2G039983	2,227	13	+737	Missense Nonsense Frameshift
14	GRMZM2G039971	1	1	0	Missense

We next determined if the assembled contigs contained predicted genes not present in the B73 genome. Based on previous observations that BAC 2 falls within the region of interest and overlaps with the other BACs, and that BAC 1 and 4 likely extend out of the region of interest, it was concluded that BAC 2 would be the best BAC to analyze in order to find predicted genes not present in B73. BAC gene prediction was performed only on contigs 5kb and larger due to the large number of small contigs. BAC 2 was assembled into 984 contigs shorter than 5 kb. Predicted genes found in BAC2 can be seen in Table 2.11.

Gene prediction on BAC 2 contigs yielded 12 predicted genes. Predicted gene 1 from contig AP2_c38 is found to overlap with the B73 predicted gene 6 (GRMZM2G419836). The predicted gene from AP2_c38 is 3,011 bp smaller than the gene model found in B73. Mutations within AP2_c38 alters the protein sequence.

Therefore, the protein structure found in BAC2 is not identical to the gene in B73. The remaining eleven out of the 12 predicted genes in BAC 2 contigs 5 kb and longer, were not present in B73, suggesting that the *Ga1-m* haplotype contains unique genes not found in the reference genome.

The six remaining genes shared between B73 and the *Ga1-m* haplotype are identified in contigs of approximately 500 to 5,000 bp in length. The predicted genes from the BAC 2 contigs 5 kb and greater were then blasted to the non-redundant nucleotide database using the NCBI web browser. Top blast outcomes can be seen in Table 2.12.

Table 2.11. Gene prediction of BAC 2 assembled contigs 5 kb and longer.

Contig	**Predicted genes**	**Predicted exons**	**Genes previously annotated in B73**
AP2 _c38	2	3	1 of 2
AP2 _rep_c126	2	3	no
AP2_rep_c137	1	1	no
AP2_rep_c138	2	5	no
AP2 _rep_c134	0	0	no
AP1_c1	1	2	no
AP2_c5	2	8	no
AP2_c23	1	2	no
AP2_c53	1	6	no
AP2_rep_c142	0	0	yes

Results suggest that the BAC 2 contigs share homology with regions of chromosome 5. Interestingly, *Ga2*, an independent gametophytic cross-incompatibility system, is found on the long arm of chromosome 5. It may be possible that the *Ga1-m* haplotype shares sequence similarities to the *Ga2* system. Similar to *Ga1*, *Ga2* possesses both a *–s* (strong) and *–m* (male) allele and has been shown to be analogous to *Ga1* (Kermicle & Evans, 2010). It is possible that during the domestication processes, an ancestral gametophytic incompatibility locus was duplicated and the duplicates diverged to become the functionally distinct *Ga1* and *Ga2* loci.

Several of the predicted genes found on BAC 2 contigs 5 kb and greater have homology to genes with functions that could play a role in pollen cross-incompatibility. Blast results for AP2_c53 and AP2_rep_c137 suggest sequence homology to transcription factors. AP2_rep_c142 shows similarity to a zinc finger. If truly present in the region of interest, transcription factors/zinc fingers could be responsible for regulating the transcription of genes necessary for pollen tube growth or hormone secretion. Altered gene expression could lead to unsuccessful pollinations. Additionally, our data suggests gene AC184772.3 is potentially duplicated in the region. The presence of a zinc finger domain could result in a dimer of the two proteins with function that contribute to the incompatibility system.

AP2_rep_c134 shows sequence similarities to an Etr2-like ethylene-receptor protein. Ethylene receptors have been shown to be responsible for plant growth and development. Disrupted hormone levels could potentially result in arrested pollen tube growth as well as other imbalances in the silks.

AP2_rep_c142 also shows sequence homology to a repressor of a protein kinase-like protein. The roles of kinases in gametophytic self-incompatibility systems in *Brassica* have been well documented. Proteins expressed by the male and female tissues interact, leading to phosphorylation of a kinase domain that ultimately inhibits pollen tube growth (Takasaki et al., 2000). It could be possible that kinases play a similar role in the inhibition of pollen tube growth in the gametophytic cross-incompatibility system.

Assembled contig AP2_rep_c126 demonstrated no homology to any known nucleotide sequences in the non-redundant database, despite gene prediction revealing two predicted genes in the contig sequence. It is possible that we discovered novel genes that have not been previously annotated in the maize genome. It could also be possible that we discovered genes unique to maize. Genes that are unique to a particular species are referred to as orphan genes. Orphan genes are thought to make up 0.5% to 8% of eukaryotic genomes (Li & Wurtele, 2015). It is hypothesized that the creation of orphan genes may be driven by genome duplication and rearrangements (Tautz & Domazet_loso, 2011), both of which our results suggest may have occurred in the *Ga1-m* haplotype. Only through further experimentation did Li and Wurtele (2015) determine the function of the orphan gene Qua-Quine-Starch (QQS) after primary sequence comparison identified no sequence homolog. It may be possible that an orphan gene is responsible for the male function in the gametophytic cross-incompatibility system. This may account for some of the difficulties in identifying the causative gene.

Table 2.12 BAC 2 contigs 5 kb and longer blasted to non-redundant nucleotide database.

Contigs	Length (bp)	e-value	Query coverage (bp)	Percent identity	Description	NCBI accession	Overlap w/ predicted gene
AP2 _c38	25,649	0.0	16014-25362	86	*Zea mays* BAC clone from chr 7	AC229875.2	
			16021-25366	86	*Zea mays* BAC clone from chr 10	AC231756.2	
			16012-25364	84	*Zea mays* BAC clone from chr 9	AC229877.2	
			16012-25363	82	*Zea mays* BAC clone ZMMBBb-37E5	AC165179.2	
			16764-25362	82	*Zea mays* BAC clone from chr 5	AC203284.4	
			18321-25362	84	*Zea mays* retrotransposon Cinful-1	AF049110.1	yes
			18321-25363	84	*Zea mays* alcohol dehydrogenase 1 genes	AF123535.1	yes
			20203-24899	88	*Zea mays* BAC clone from chr 5	AC203071.4	
			20393-25362	86	*Zea mays* BAC clone from chr 5	AC196774.5	
			20043-25362	84	*Zea mays* cultivar B73 clone genomic sequence; identified as flowering time locus on chr 10	GU142949.1	
			20393-25363	86	Genomic sequence for *Zea mays* BAC clone ZMMBBb0448F23	AC160211.1	
			20551-24898	88	*Zea mays* putative transposase	AF466646.1	
			20393-25362	85	*Zea mays* putative growth-regulating factor 1	AY530951.1	
			16012-20603	85	Contiguous genomic DNA; 19-KDA-zein family from *Zea mays*	AF546188.1	yes
AP2 _rep_c126	17,493	0.0	13830-16660	86	*Zea mays* clone FS2 19 chr B	EF190061.1	
			13830-15322	88	*Zea mays* clone from chr 6	AC226723.4	
AP2_rep_c137	14,619	0.0	10536-14616	89	*Zea mays* BAC clone from chr 5	AC196008.3	
			10536-14616	89	*Zea mays* BAC clone from chr 5	AC204225.4	
			10620-14618	85	*Zea mays* BAC clone from chr 5	AC201762.5	
			10620-14618	85	*Zea mays* BAC clone from chr 5	AC215174.5	
			10749-14563	85	*Zea mays* clone ZMMBBb-125C19	AC165173.2	
			11442-14609	85	*Zea mays* BAC clone from chr 5	AC196084.4	
			11442-14610	85	*Zea mays* BAC clone from chr 5	AC194844.5	
			11514-14610	85	*Zea mays* BAC clone from chr 5	AC210260.5	

Table 2.12 continued

			11805-14615	87	*Zea mays* m19 gene for putative MADS-domain transcription factor allele ZMM19	AJ850298.1	
AP2_rep_c138	12,648				Blast hits did not meet criteria		
AP2_rep_c134	12,298	0.0	8317-12186	83	Zea mays clone BACs ZMMBBb0345O22, ZMMBBc0294D02, ZMMBBb0103L15, ZMMBBb0622H01, and ZMMBBb0335C07	EF517600.2	
			1519-4106	88	*Zea mays* clone FS2 19 chr B	EF190061.1	
			1518-4119	88	*Zea mays* BAC clone from chr 6	AC226723.4	
			2666-4121	89	*Zea mays* B73 Etr2-like ethylene receptor (ETR61) pseudogene	AY359583.1	
			2666-4121	89	*Zea mays* full-length cDNA clone ZM BFb0095N09 mRNA	BT084267.2	
AP1_c1	8,315	0.0	1-2098	93	*Zea mays* BAC clone form chr 10	AC226721.2	
			17-2087	91	*Zea mays* chromosome 4 seq AGI.478 genomic sequence	GQ845080.1	
			2986-4472	96	PREDICTED: *Zea mays* uncharacterized protein	XM_008654301.1	yes
			2976-4472	95	*Zea mays* hypothetical protein mRNA	EU956244.1	yes
			720-1984	97	*Zea mays* full-length CDNA clone	BT069767.1	
			720-1982	96	*Zea mays* full-length cDNA clone	BT083566.2	
			1360-3070	89	*Zea mays* chloroplast phytoene synthase gene	AY455286.1	
			720-2002	94	*Zea mays* clone hypothetical protein mRNA	EU973310.1	
			1360-2096	94	*Zea mays* cultivar inbred line B73 teosinte glume architecture 1	AY883559.2	
AP2_c5	8,239	0.0	7240-7861	100	*Zea mays* uncharacterized LOC100501595	NM_001196280.1	yes
			6849-7936	99	*Zea mays* clone mRNA sequence	EU966398.1	yes

Table 2.12 continued

AP2_c23	7,451	0.0	1861-7268	98	*Zea mays* putative pol protein	AF466202.2	yes
			1861-7268	98	*Zea mays* clone ZMMBBb-136N21	AC165175.2	yes
			1861-7268	97	*Zea mays* genomic clone ZM15C05 sequence	AC116033.3	yes
			3334-7268	96	*Zea mays* clone from chr 5	AC210260.5	yes
			3538-5384	99	PREDICTED: Charadrius vociferous uncharacterized	XM_009883513.1	yes
			1859-3547	97	*Zea mays* BAC clone from chr 5	AC203430.5	
			1861-3547	95	*Zea mays* BAC clone from chr 2	AC229873.2	
			1860-3547	88	*Zea mays* BAC clone from chr 10	AC225944.3	
			20-1454	88	*Zea mays* BAC clone from chr 5	AC207417.4	
			135-1454	88	*Zea mays* BAC clone from chr 5	AC216353.5	
			50-1454	86	*Zea mays* cultivar Mo17 locus 9008	AY664418.1	
			50-1414	86	*Zea mays* cultivar B73 locus 9008	AY664414.1	
			228-1454	88	*Zea mays* BAC clone from chr 1	AC226722.2	
			297-1454	89	*Zea mays* BAC clone from chr 10	AC226721.2	
			77-1431	85	*Zea mays* clone mRNA	EU942949.1	
AP2_c53	5,767	0.0	1023-5730	89	*Zea mays* BAC clone from chr 5	AC204225.4	yes
			1023-5730	89	*Zea mays* BAC clone from chr 5	AC202177.4	yes
			1023-5761	85	*Zea mays* unknown putative heme oxygenase, anthocyanin biosynthesis regulatory protein, putative growth-regulating factor 1, and putative aminoalcoholphosphotransferase genes	AY530952.1	yes
			1023-5767	85	*Zea mays* clone ZMMBBb/125O19	AC165173.2	yes
			1023-5763	85	*Zea mays* BAC clone from chr 5	AC216070.4 AC201762.5 AC215174.5 AC202076.4 AC197049.5	yes
			1023-5762	85	*Zea mays* BAC clones from chr 6	AC231746.2	yes
			1032-5720	85	*Zea mays* BAC clone from chr 5	AC196774.5	yes

Table 2.12 continued

			1023-5645	85	*Zea mays* cultivar B73 locus 9009	AY664415.1	yes
			1023-5645	85	*Zea mays* cultivar Mo17 locus 9009	AY664419.1	yes
			1015-4909	85	*Zea mays* BAC clone from chr 5	AC204937.4	yes
AP2_rep_c142	5,430	0.0	1361-3575	88	PREDICTED: *Zea mays* zinc finger	XM_008676910.1	
			1361-3578	88	PREDICTED: *Zea mays* 52 kDa repressor for the inhibitor of the protein kinase-like	XM_008679685.1	
			1566-3575	88	PREDICTED: *Zea mays* zinc finger MYM-type protein 1-like	XM_008676911.1	
			1361-2842	88	PREDICTED: Chrysemys picta bellii zinc finger MYM-type protein 6-like	XM_008178212.1	
			2183-3413	89	PREDICTED: Caprimulqua carolinensis zinc finger MYM-type protein 1-like	XM_010163805.1	
			1361-2478	85	*Zea mays* CYP71C1 gene for cytochrome P-450	X81828.1	

Conclusions

The BAC assembly project concluded with assembled contigs from each BAC file. We were successful in our attempt to compare assembled sequence with the B73 reference genome to characterize entire gene insertions and deletions and gene polymorphisms. In this research, we present two assembly methods and resulting conclusions from each.

Maize has many repetitive regions. Our BAC assembly data are consistent with this. We believe the repetitive nature of the region of interest, as well as substantial sequence variation between our BAC sequences and the B73 reference genome, resulted in an inefficient comparative genome assembly method. Because of this genomic structure, a de novo assembly of the region of interest worked better than first assembling reads that mapped to the B73 reference genome. De novo assembly of individual BACs and removal of residual contaminants resulted in the creation of 2,820 contigs. Contig breaks are suggestive of repetitive regions that remained unassembled. Additional arrangement and connection of contigs is required.

The de novo assembly of BAC sequences in our research successfully identified six predicted genes and one transposable element from the B73 genome. Gene model alignments showed polymorphisms that could lead to altered protein structure in BAC 2 contigs. The lack of annotated genes in the region and significant sequence variation made the identification of causative polymorphisms in the region challenging. Our results do suggest noncolinearity between the BAC sequences and the B73 reference genome. Six predicted genes and two transposable elements from the region of interest in B73 were not found within the *Ga1-m* haplotype and therefore appear to be absent from the region. Gene alignments support both theories that gene insertion/deletions and/or gene polymorphisms may underlie the male function in this system. At this point, we cannot definitively rule out either hypothesis.

We demonstrate clear BAC alignment with the gene GRMZM2G039983, predicted by Liu et al. (2014) to have a possible role in gametophytic cross-incompatibility. This gene has five gene insertion sites and multiple polymorphisms that resulted in a modified protein structure. Our results of a modified GRMZM2G039983 gene sequence are consistent with past conclusions that the gene may play a role in the incompatibility system. Using the current B73 v3 genome, we determine that the region of interest identified by Liu et al. (2004) is smaller than originally documented. Furthermore, we found that the putative gene identified (GRMZM2G039983) is no longer in the region. Published markers were used to identify genes 9, 10, and 11 to now be putative genes in the region of interest. Gene GRMZM2G027021 is a protein coding gene found in both the version 2 and version 3 region of interest. We also found possible transposable element activity in the gene sequence in our BAC sequences. We identify GRMZM2G027021 as a gene of high interest for causation of the male factor.

Gene prediction on BAC 2 assembled contigs of 5 kb and longer from the *Ga1-m* haplotype yielded a total of 11 predicted genes not present in B73. BLAST results from the same BAC 2 contigs of 5 kb and longer suggest sequence homology on chromosome 5 and other conserved domains.

Significance

The mechanism underlying gametophytic cross-incompatibility in maize has remained a mystery since it was first identification in 1902 by Correns. Numerous research studies have been performed and much knowledge has been contributed to the field; however, many integral questions about the system remain unresolved. Interest in using the gametophytic cross-incompatibility system as a biological barrier to prevent unwanted pollination of maize has increased. Increased knowledge of the system has economic advantages. The utilization of the gametophytic cross-incompatibility system may have benefits in organic and specialty maize production. Effective isolation of transgenes from certain maize systems would benefit producers of both market types. The use of the gametophytic cross-incompatibility system as a means to control the flow of transgenes could possibly prevent future allegations between farmers and biotechnology companies producing transgenic maize. Increased efficiency and ease of isolation could also result in a decreased maize price for consumers.

The ability to easily sequence DNA, has allowed for characterization of the region on the basepair level. This project marks the first attempt, to our knowledge, to sequence and annotate the 9.1 to 9.6 Mbp region from a *Ga1-m* haplotype.

Recommendation for Future Research

The next step required to move this project forward is to determine overlap of contigs across BAC files. Contigs must be correctly ordered and assembled into a scaffold sequence spanning the region of interest. PCR primers can be created with the aim of linking assembled contigs. Purified PCR product can be sequenced and used to fill in sequence gaps between contigs. Upon completion of a consensus sequence, gene prediction and gene annotation can be performed on the entire consensus sequence. Gene prediction on a sequence that covers the entire region of interest will give a more accurate estimation of novel genes. A better understanding of the sequence homology between the region of interest and chromosome 5 (potentially *Ga2*) might shed light on genome arrangement and interaction.

PacBio sequencing may also greatly assist the assembly process. PacBio reads are much longer than reads from any other current sequencing technology, with a median length of 2,200 bp. PacBio reads could successfully span repetitive regions that are challenging to assembly with shorter reads. Additionally, PacBio reads and the Miseq reads used in this experiment could be used in a hybrid assembly with the PacBio reads. The presence of the shorter Miseq reads coupled with longer reads have been shown to offset the inherent sequencing error present with longer sequence reads and could potentially lead to a much improved assembly (Koren et al., 2012).

Further experiments could be done to assess involvement of predicted genes in the gametophytic cross-incompatibility region. The CRISPR-Cas 9 system could be used to knock out genes of interest and determine their role. Additionally, candidate genes could be transformed into a *ga1* haplotype and the outcome observed. Due to the smaller, simpler genome of *Arabidopsis*, incorporating the genes into *Arabidopsis* might be a valuable experiment.

RNA-seq work could also bring a greater understanding to the gametophytic cross-incompatibility system. Expression data of compatible versus incompatible reactions at different time points in pollen tube growth could be collected. The RNA-seq reads could then be aligned to the region of interest and differentially expressed genes in the region (including the predicted novel genes) could be determined. Mapping RNA-seq reads to the genome could also be beneficial in annotating genes found in the region.

References

Altschul, S., Gish, W., Miller, W., Myers, E., & Llipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*, 403-410.

Ashman RB. (1975). Modification of cross sterility in maize. *J Hered*. 66:5–9.

Birnboim, H., & Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA.*Nucleic Acids Research, 7*(9), 1513-1523.

Bloom, J. C., & Holland, J. B. (2011). Genomic localization of the maize cross-incompatibility gene, Gametophyte factor 1 (ga1) (Vol. 56, pp. 379-387): Maydica.

Chaisson, M., Brinza, D., & Pevzner, P. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Cold Spring Harbor Laboratory Press, 19*, 336-346.

Chevreux, B., Wetter, T., & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics, 99*, 45-56.

Core, N. J. I. B. SICKLE. UC Davis Genome Center Institution: GitHub.

Hall, T. (1997). BioEdit (Version 7.2.5).

Julian, P., Allen, J., Christensen, M., Davis, P., Falin, L., Grabmueller, C., et al. (2014). Ensembl genomes 2013: Scaling up access to genome-wide data. *Nucleic Acids Research, 42*(D1), D546-D552.

Koren, S., Schatz, M., Walenz, B., Martin, J., Howard, J., Ganapathy, G., E, et al. (2013). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology 30*(7), 693-700.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrow-wheeler transform. *Bioinformatics, 25*, 1754-1760.

Li, L., & Wurtele, E. (2014). The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. *Plant Biotechnology, 13*(2), 177-187.

Liu, X., Sun, H., Wu, P., Tian, Y., Cui, D., Xu, C., et al. (2014). Fine mapping of maize cross-incompatibility locus gametophytic factor 1 (ga1) using a homogeneous population. *Crop Science, 54*, 1-9.

Mi, H., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. (2010). PANTHER version 7: Improved phylogenetic trees, orthologs, and collaboration with the gene ontology consortium. *Nucleic Acids Research, 38*, 204-210.

Microsoft. (2010). Microsoft Excel. Redmond, Washington: Microsoft.

Monaco, M., Sen, T., Dharmawardhana, P., Ren, L., Schaeffer, M., Naithani, S., et al. (2012). Maize metabolic network construction and transcriptome analysis. *Plant Genome, 9*.

Monaco, M., Sen, T., Dharmawardhana, P., Ren, L., Schaeffer, M., Naithani, S., et al. (2013). Maize Metabolic Network Construction and Transcriptome Analysis. MaizeGDB.

Nelson OE. (1994). The gametophyte factors of maize. In: Freeling M, Walbot V, editors. *The maize handbook*. Berlin (Germany): Springer-Verlag. p. 496–503.

Pearson, W., Wood, T., Zhang, Z., & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics, 46*(1), 24-36.

Phillippy, A., Schatz, M., & Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly.Genomics Biology, 9(3), R55.

Pop, M., Phillippy, A., Delcher, A., & Salzberg, S. (2004). Comparative genome assembly. Briefings in Bioinformatics, 5(3), 234-248.

Robert, I., Stead, A., Ockendon, D., & Dickinson, H. (1980). Pollen stigma interactions in *Brassica oleracea*. *Theoretical and Applied Genetics, 58*, 241-246.

Salamov, A., & Solovyev, V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Research, 10*(5), 16-522.

Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE, 6*(3), e17288.

Springer, N., Yin, J., Fu, Y., Ji, T., Yeh, C., Jia, Y., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLOS Genetics, 5*(11), e1000734-e1000734.

Takasaki, T., K. Hatakeyama, G. Suzuki, M. Watanabe, A. Isogai, and K. Hinata (2000). The S receptor kinase determines self-incompatibility in *Brassica stigma*. *Nature* 403 (6772): 913–6.

Tautz, D., & Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics, 12*(10), 692-702.

Thompson, J., Higgins, D., & Gibson, T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*(22), 4673-4680.

Thomas, P., Campbell, M., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research, 13*, 2129-2141.

Zhang, H., Liu, X., Zhang, Y. e., Jiang, C., Cui, D., Liu, H., et al. (2012). Genetic analysis and fine mapping of the Ga1-s gene region conferring cross-incompatibility in maize (Vol. 124, pp. 459-465): Theoretical and Applied Genetics.

CHAPTER THREE: ENCYCLOPEDIA OF FOOD GRAINS: MAIZE CHAPTER

A chapter published in the *Encyclopedia of Food Grains*

Marianne Emery and M. Paul Scott

Publication status: Pending

Abstract

Maize grain in an important source of food around the world. Maize variety, processing, and cultural tradition dictate use of maize in food. The maize plant is regarded as a model system in the scientific world. Due to relative ease of working with maize, a large body of research has been compiled by the maize community, most notably the assembly of the maize genome. Further, maize is continually being improved for a variety of marketable traits. This chapter gives an overview of breeding techniques and concerns that arise in regards to such maize plant modifications.

Introduction

Zea mays, more commonly referred to as maize, is a member of the grass family *Poaceae*, or true grasses. Maize is thought to have originated 55-70 million years ago in what is now Central or South America and has since diversified into nearly 10,000 nondomestic relatives. Figure 3.1 shows a phylogenetic tree of grass species related to maize. There exists no direct ancestor for maize, however to date the closest relative to maize are the teosintes (Kiesselbach, 1949; Strable and Scanlon, 2013; Wilkes, 2004). Prehistoric selection has resulted in ears lacking seed cases called glumes and seeds that adhere to the cob until manual removal. These alterations limit the ability of maize to survive without human intervention. Maize is an annual plant with C4 metabolism making it very efficient at carbon fixation. It has the greatest global production of any crop species. Nearly eight million tons were produced worldwide in 2013, accounting for 32% of total cereal production (FAO, 2014). The top three producers include the United States, China, and Brazil. Maize is grown on more area of the planet than any other crop and is grown on every continent except Antarctica. Over 300 countries in the world rely on maize for their food supply on a daily basis (FAO, 2014). The grain of maize is used for food, feed, and industrial products including biodegradable foams, plastics, and adhesives. Additionally, maize stover, the leaves and stalk of the maize plant, is used for forage, biofuel production, and chemical production.

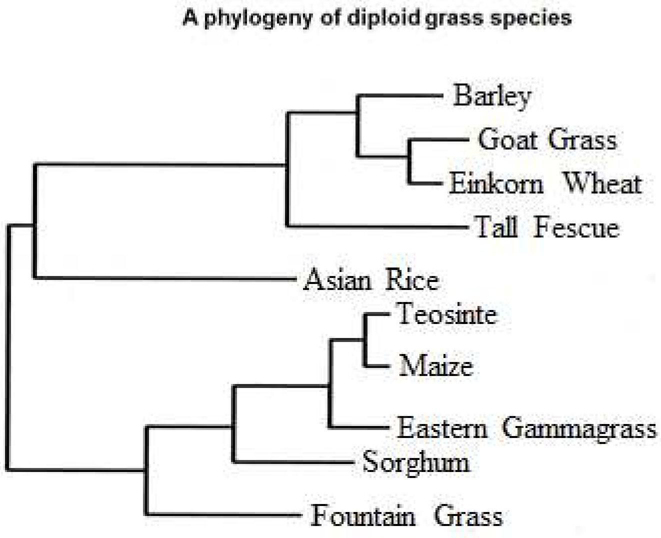


Figure 3.1. A phylogeny of diploid grass species. (Adapted from Gaut B S et al., 2000)

Maize Reproduction

Maize is a monecious plant, meaning it has both male and female reproductive organs on the same plant. Flowers mature after approximately 60-70 days of vegetative plant growth. Male staminate flowers develop into tassels and are found on the uppermost tip of the main stem. Female pistillate flowers are found in one or more ears located at nodes along of the stem. Typical maize varieties are diploid, containing two sets of 10 chromosomes. Copious amounts of pollen (up to one billion grains per plant) are shed from anthers and dispersed by air currents. While the majority of the pollen falls close to the plant, a small portion of the pollen can be carried great distances on air currents. Industry standards typically consider plants separated by a distance of 660 feet to be reproductively isolated. Fertilization occurs by the process of "double fertilization" common to angiosperm species. A pollen grain carrying two nuclei lands on a silk and germinates to produce a pollen tube. The pollen tube grows down the length of the silk until it reaches the embryo sac where it ruptures releasing the two sperm nuclei. The first sperm cell fuses with the egg cell, to produce the embryo, the organ that ultimately develops into the next generation plant. The second sperm cell fuses with the central cell of the embryo sac giving rise to the endosperm, the storage tissue that nourishes the developing seedling until it is capable of living independently. Grain fill to maturity takes about 40 days (Kiesselbach, 1949; Strable and Scanlon, 2013).

Maize Kernel Composition

The mature maize kernel is referred to as a caryopsis and is not a true seed but rather a one-seeded fruit (Keisselbach, 1949; Rooney et al, 2004). Kernels are composed of four organs: the pericarp, embryo, endosperm, and pedicel (Keisselbach, 1949). Physical properties, such as hardness, shape, size, color and composition vary among maize varieties.

The main organs of a maize kernel are shown in Figure 3.2. The outer layer of the kernel is the pericarp and encloses the kernel for protection. The endosperm comprises the majority of the kernel's inner contents. The endosperm itself is composed of four tissue types: the aleurone (outer) layer, the starchy endosperm, the basal endosperm transfer layer (BETL), and the embryo-surrounding region (ESR) (Scanlon and Takacs, 2009). The endosperm provides nutrients in the form of sugars and amino acids to the growing embryo. The embryo is composed of the following: the scutellum (the monocotyledon that absorbs nutrients during germination), the coleoptile (protective sheath of the emerging shoot), the plumule (young plant), the radicle (primary root), and the coleorhizae (protective sheath of emerging root) (Scanlon and Takacs, 2009; Rooney et al., 2004). The tip cap serves to attach the kernel to the cob and protect the kernel.

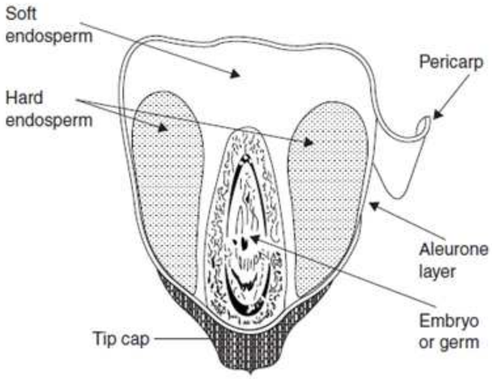


Figure 3.2. The mature maize kernel, showing component parts.
(Encyclopedia of Grain Science)

In terms of nutritive composition, the kernel can be further classified into five main components. The typical Number 2 Yellow Dent maize kernel contains approximately 72% starch, 9.5% protein, 4.3% oil, 1.4% ash and 2.6% sugar (Watson, 2003).

Starch

Starch is the most abundant component in maize kernels and serves as an efficient storage molecule for glucose. Starch accumulates in the form dense insoluble granules. It is composed of two main components: amylose and amylopectin. Amylose is predominantly a linear polymer composed of 1, 4 linked alpha D-glucan chains. In contrast, amylopectin is highly branched by alpha-1, 6 glycosidic bonds. Starch biosynthesis requires the coordinated activities of a myriad of enzymes, including starch synthases, starch branching enzymes, and starch debranching enzymes. Enzymatic activity within the kernel alters starch content and properties. Degree of branching and branch chain length are starch properties that can vary considerably among maize varieties (Campbell et al, 1994; Ji et al, 2003). Maturity also affects starch quality (Jennings et al, 2002; Pollak and Scott, 2005). Traits sought in a commercial setting include gel strength, viscosity, and thermal properties such as gelatinization. Maize starch provides four calories per gram.

Genetic mutations can confer altered starch phenotypes. Mutant alleles of *waxy1* (*wx1*) produce 100% amylopectin starch, which is useful as a thickening agent in foods. Mutation of the amylose extender gene (*ae*) leads to high amylose starch (HAS) (Vineyard and Bear, 1952) with a range of amylose values from 25-80%. HAS is known for its slow digestion in vivo. The *sugary-1* (*su1*) and *shrunken-2* (*sh2*) lead to kernel phenotypes that are sweeter than field corn, and are used to produce sweet corn varieties for canning and fresh consumption.

Oil

Oil is the second most abundant component of maize kernels. Oil from a kernel of typical Corn Belt Maize, Number 2 Yellow Dent, contains approximately 62% linoleic, 25% oleic, 10% palmitic, 2% stearic and 1% linolenic acid; saturated fatty acids equate to approximately 12% of total lipid content (Pollak and Scott, 2005; Poneleit and Davis, 1972). The oil within the maize kernel provides nine calories per gram. Linoleic, linolenic, eicosapentaenoic, and docosahexaenoic fatty acids are shown to have a positive correlation with cardiovascular health. A ratio of 6:1, linoleic to linolenic, is recommended (Wijendran and Hayes, 2004).

Similar to starch content and quality, studies demonstrate that exotic germplasm possesses extensive ranges of fatty acid composition (Jellum, 1970). Exotic lines are crossed to yield varieties with increased oil content. Oil content varies across inbred maize lines (Poneleit and Davis, 1972) and across varied environments. Total fatty acid composition varies throughout kernel development and ultimately increases as the kernel matures (Poneleit and Davis, 1972). Oil content is believed to be affected by a large number of loci (Dudley and Lambert, 1992) and is a highly heritable trait. Certain breeding schemes aim solely at increasing lipid content and/or quality (Hallauer, 2004). Duvick (2003) altered fatty acid content by introducing Tripascum genes, a wild relative of maize, into various maize lines.

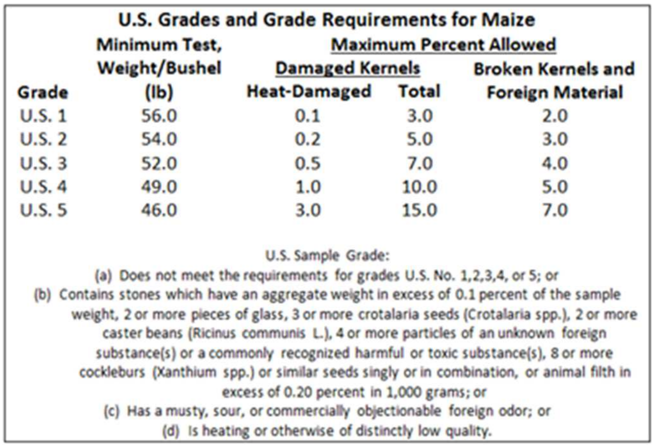
Fatty acid stability is directly correlated to saturation level. Linolenic is the least stable fatty acid, containing three points of unsaturation. Oleic fatty acids are much more stable and less prone to oxidation. Oleic fatty acids are mono-unsaturated. Once oxidation begins, it cannot be stopped or reversed and ultimately leads to rancidity.

Protein

Protein is another vital component to the maize kernel. Seed proteins are divided into four classes: albumin, globulin, prolamin, and glutelins (Rooney et al, 2004). The major storage proteins in maize are prolamins, also referred to as zeins. Eighty percent of the stored protein in maize is found in the endosperm (Flint-Garcia et al., 2009). Because of the amino acid balance of zeins and their abundance in the endosperm, lysine, tryptophan, and methionine are typically at low levels in maize (Flint-Garcia et al., 2009).

Maize is therefore not a complete protein source and must be eaten with complementary protein sources to ensure requirements for the essential amino acids are met. Many countries rely on maize as their main food source; in turn essential amino acid deficiencies such as Kwashiorkor and pellagra frequently occur (Krivanek, 1949). Maize protein provides 4 calories per gram.

Research aims to increase the quality of protein in maize. First observed in 1920, the *opaque-2* (*o2*) mutation causes a decrease in the amount of zein content and thus a higher ratio of nonzein proteins with increased levels of essential amino acids. (Krivanek, 1949; Mertz et al., 1964). Unfortunately, this mutation results in reduced kernel hardness, yield, and fungal and pest resistance (Krivnek, 1949; Vasal, 2000). To overcome this deficiency, modifier genes have been introduced into *o2* varieties that increase kernel hardness. The resulting maize is called Quality Protein Maize (QPM) and grown in many parts of the world where it has contributed to improved nutrition (Prasanna et al., 2001). In addition to *o2* mutants, *floury2* (*fl2*) mutants have shown to have improved amino acid balance (Nelson et al., 1965).



U.S. Grades and Grade Requirements for Maize

	Minimum Test,	Maximum Percent Allowed		
	Weight/Bushel	Damaged Kernels		Broken Kernels and
Grade	(lb)	Heat-Damaged	Total	Foreign Material
U.S. 1	56.0	0.1	3.0	2.0
U.S. 2	54.0	0.2	5.0	3.0
U.S. 3	52.0	0.5	7.0	4.0
U.S. 4	49.0	1.0	10.0	5.0
U.S. 5	46.0	3.0	15.0	7.0

U.S. Sample Grade:

(a) Does not meet the requirements for grades U.S. No. 1,2,3,4, or 5; or

(b) Contains stones which have an aggregate weight in excess of 0.1 percent of the sample weight, 2 or more pieces of glass, 3 or more crotalaria seeds (Crotalaria spp.), 2 or more caster beans (Ricinus communis L.), 4 or more particles of an unknown foreign substance(s) or a commonly recognized harmful or toxic substance(s), 8 or more cocklebursp (Xanthium spp.) or similar seeds singly or in combination, or animal filth in excess of 0.20 percent in 1,000 grams; or

(c) Has a musty, sour, or commercially objectionable foreign odor; or

(d) Is heating or otherwise of distinctly low quality.

Figure 3.3. U.S. maize grading scale.
(USDA, 2013)

Less abundant components of the maize kernel include: fiber, minerals, vitamins, anthocyanins, and anti-nutrients.

Maize in Food

Maize is a food ingredient that brings commonality to culinary cultures across the world. Cultural traditions and corn varieties dictate how maize is incorporated into a wide variety of foods. Main kernel components can be separated and processed into products such as corn starch for thickening and binding agents and corn oil for frying and baking; whole grain kernels are used in popped popcorn or ground into corn meal and used in breads, biscuits, and cereals. From enchiladas, tamales, totopos, tostaditas, and tortillas, virtually every Mexican dish uses maize. Maize porridges are seen across the world: referred to as puliszka and malderash in Hungary, posho in Africa, polenta in Europe, grits in the United States, and kpekple in Ghana. Maize meal can be ground and fermented into sora, a maize beer in Peru, or used to make hard alcohols such as whiskey and bourbon. Maize is truly a cross cultural food.

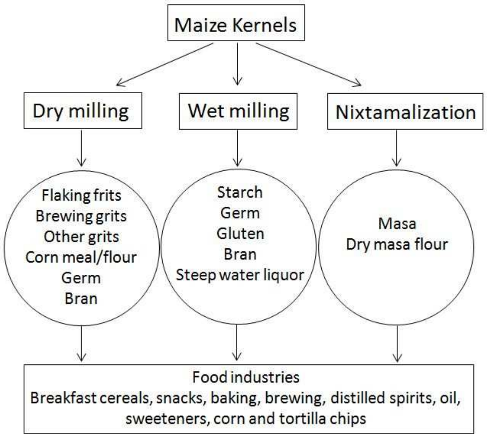


Figure 3.4. Maize food processing determines maize as food ingredient.
(Adapted from Encyclopedia of Grain Science)

Maize processing

Maize kernel quality and physical attributes determine its end use. The U.S. recognizes 5 grades of maize and three classes: yellow, white, and mixed maize (USDA, 2013). Food maize typically specifies number 1 grade yellow or white dent corn (Figure 3.3). Additionally, the manner in which maize is processed is a vital component in its incorporation into foods (Figure 3.4).

Harder kernels are desirable for storage, shipping and handling; dry-milling calls for a kernel with a harder endosperm void of cracks (Rooney et al, 2004). Dry-milling is often used to produce baked goods, breakfast cereals, and ethanol (Orthoefer and Eastman, 2004). In the dry-milling process, tempering the grain is a vital first step. A hammer mill is then used to coarsely grind the maize kernels. Several steps of size and weight separation, in addition to regrinding, yield maize grits, flour, and fiber. The quality, content, and end use of the maize must be considered before entering the wet or dry milling process.

Maize kernels with a softer endosperm perform better in the wet-milling process (Orthoefer and Eastman, 2004). The wet-milling process includes steeping maize in a dilute sulfur dioxide solution to soften the kernel and separate it into its smaller components. The germ can be first removed and later processed for oil. The remaining components are ground and separated further into grits, flour, and fiber. Further processing yields corn gluten, meal, and starch. Maize starch and high fructose corn syrup are a main end-product of the wet milling process in the United States.

Nixamalization, dating back from 1200-1500 BC, is an ancient type of maize processing that includes rendering kernels into a paste to increase the bioavailable nutrients such as calcium and digestible iron(Orthoeffer and Eastman, 2004; Rooney et al, 2004). Kernels are steeped in a water/lime solution over heat and ground into masa, also known as maize dough that is used to produce tortillas, corn chips and other food products.

Maize in Science

Maize is an important model organism in genetic research. It has several attributes that make it attractive for this purpose. It is a large plant and phenotypic analyses are easily done. Each plant produces an ear typically containing 100-400 kernels. It is broadly adapted and has tremendous genetic diversity. Maize has a moderately sized genome of approximately 2.5 gigabase pairs (Strable and Scanlon, 2013). A vast collection of mutant stocks have also been developed that assist in research; this has allowed for many genes to first be characterized molecularly in maize. Being a diploid species, genetic manipulation and analysis is less complex than in species with a higher ploidy level. Additionally, the large physical size of the maize chromosomes is a great benefit to cytogenetic researchers.

Research on maize has led to several key discoveries. Perhaps most notable is the discovery of transposons by Barbara McClintock (McClintock, 1950), for which she was awarded the Nobel Prize in Physiology or Medicine in 1983. Cytogenetic studies in maize resulted in an understanding of genetic recombination and

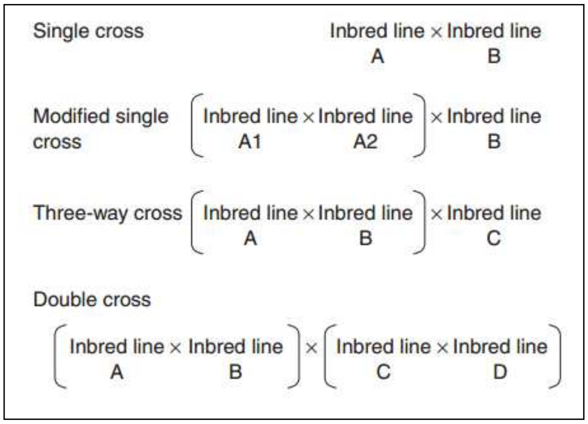

Single cross
Inbred line A × Inbred line B
Modified single cross
(Inbred line A1 × Inbred line A2) × Inbred line B
Three-way cross
(Inbred line A × Inbred line B) × Inbred line C
Double cross
(Inbred line A × Inbred line B) × (Inbred line C × Inbred line D)

Figure 3.5. Types of hybrids grown commercially in North America.
(Encyclopedia of Grain Science)

enabled genetic mapping. The role of telomeres was determined in maize. Through collaborative efforts, it was one of the first crops to have its genome completely sequenced (Schnable et al., 2009).

Maize Breeding, Genetics, and Biotechnology

Maize cultivar types

A cultivar is a plant variety has been developed for a specific use. Several types of maize cultivars are grown including inbred lines, single-cross hybrids, double cross hybrids, and open pollinated varieties (Figure 3.5). Inbred lines are created by successive generations of self-pollination. The resulting plants are genetically homozygous and phenotypically homogeneous. Due to inbreeding depression, inbred lines have low yield and are not used for grain production.   Their main purpose is in the production of hybrid seed. When two inbreds are cross pollinated, a single-cross hybrid results. Single-cross hybrids are genetically heterozygous and phenotypically uniform.   Because of the difficulties of producing seed on inbred lines, several types of hybrids have been developed. An open pollinated variety is a population of plants that is genetically heterozygous and phenotypically non-uniform. As the name implies, seed of open pollinated varieties is produced by allowing natural pollination to occur in the population. Synthetic populations are derived from inter-mating several varieties are frequently used in breeding programs to produce inbred lines.

Mechanized agriculture has led to a preferece for hybrids because of their uniformity and high yields. The process of hybrid improvement and seed production has become highly industrialized. Industrial maize breeding has led to greatly increased yields. Open pollinated varieties require much less infrastructure for seed production and genetic improvement and are often grown in developing countries.

Hybrid maize breeding

Hybrid maize breeding allows breeders to capture and fix extremely productive genotypes by taking advantage of hybrid vigor. Productivity and vigor in maize plants is generally proportional to the degree of heterozygosity. Thus, inbred lines, although uniform and reproducible are usually poor agronomic purposes. Heterozygosity and performance can be restored by crossing unrelated inbred lines to make hybrids. Inbred lines are classified into heterotic groups according to their ability to form productive hybrids in combination with other groups and their suitability as a male or female parent. For example, nearly all inbreds used as females in North American hybrids are in the Stiff Stalk heterotic group. Development and maintenance of inbred lines and testing hybrid combinations requires a great deal of infrastructure and expertise that is not available to most farmers or even small seed companies and is therefore largely done by large seed companies.

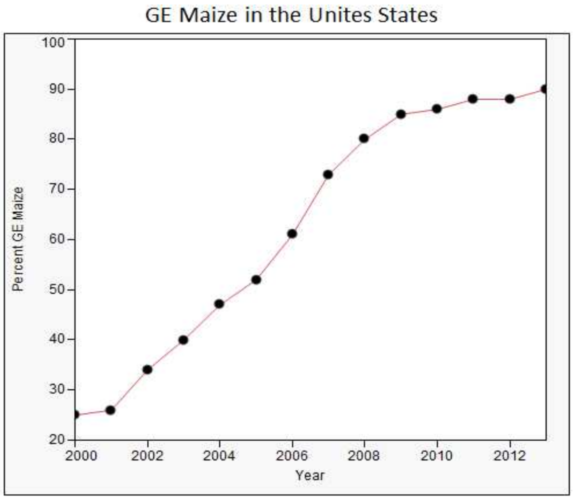


Figure 3.6. Percentage of all maize grown in the United States that is genetically engineered (GE). (USDA, 2014)

Uniformity is essential in efficient and profitable production of maize. Superior technology and machinery has assisted with such uniformity. Improved accuracy in the evaluation of cultivars has allowed for large genetic gains and the overall creation and advancement of superior maize inbreds. Superior farm equipment equipped with GPS and computer monitoring systems has led to optimal planting depth, density, and spacing and precise measurements of grain yield during harvest. In the future, precision agriculture will continue to increase productivity by optimizing inputs, such as corn variety and fertilizer amount, on a per land area basis.

Maize biotechnology

Biotechnology is the ability to introduce genes from any source into the maize genome. Two types of traits derived from biotechnology methods are currently in commercial production: insect resistance and herbicide tolerance.

Insect resistant maize decreases the need of pesticide applications directly to the plant. The use of pesticides in the United States has been reduced 6% since 1996, a total of 172 million kilograms per year (Brookes and Barfoot, 2005). Fewer pounds of chemical are applied, benefiting the health of the environment and proving economically beneficial for the farmer. From 1996-2010, the income of US farmers increased a total of $21.7 billion dollars; 23 percent of that profit was derived from 2010 alone (Brookes and Barfoot, 2005). The percentage of GE maize has increased almost 4 fold in 12 years (Figure 3.6.). The United States Department of Agriculture (USDA) Economic Research Service reports that *Bt* maize decreased the amount of insecticides per planted acres of *Bt* maize by 8% in the United States (Fernandez-Cornejo and Caswell, 2006). Herbicide tolerant maize is agreeable in environments of low to no-till agriculture. Minimal to no tillage results in decreased fuel usage and reduction of greenhouse gas emissions, as well as less soil compaction and erosion. Additionally, crop residue left on top of the soil increases levels of organic carbon sequestration. Soil and water quality are increased due to decreasing soil erosion and nutrient loss (Committee on the impact of biotechnology on farm-level economic and sustainability and national research council, 2010; National Research Council, 2010).

Of the 159 million hectares of maize grown globally in 2012, 55.1 million hectares (35%) were biotech maize (Clive, 2012). Legislation regulating such crops varies among countries. The United States regulates genetically modified organisms (GMOs) based on the end product. Three groups with differing perspectives and expertise regulate genetically modified (GM) crops in the US: the US Environmental Protection Agency (EPA), the Food and Drug Administration (FDA), and the US Department of Agriculture (USDA). GM crops must be verified free from environmental and human toxins as well as foreign proteins deemed allergenic. The FDA policy established in 1992, considers the currently approved GM crops to be “substantially equivalent” to non-GM crops and deemed “Generally Recognized as Safe” under the Federal Food, Drug, and Cosmetic Act (FFDCA); therefore, foods made with approved GM varieties do not require pre-market approval (Tucker, 2011). Acceptance of GM maize by the consumer varies by country. The European Union (EU) regulates GM crops based on the process in which they are produced. The EU tends to be cautious of GM crop consumption. The British Press often refers to such crops as “Frankenstein Foods.” Protestors of third world countries have been known to destroy entire fields of GM crop despite starvation in the country. The major concerns over production of GM maize are pollination of weedy species or non-GM maize by GM pollen resulting in undesirable transfer of the transgene (Snow, 2002) and the impact of transgenes on non-target species, particularly beneficial insects, and the development of insects resistant to the mode of action of the insecticidal transgenes in use. Researchers and regulatory agencies continue to develop new deployment strategies in an effort to minimize these risks.

References

Brookes, G., & Barfoot, P. (2005). GM crops: The global economic and environmental impact-the first nine years 1996-2004. *AgBioForum, 8,* 187-196.

Campbell, M. R., White, P. J., & Pollak, L. M. (1994). Dosage effect at the sugary-2 locus on maize starch structure and function (Vol. 71, pp. 464-468): Cereal Chemistry.

Corn Breeding: Types of Cultivars. (2014, April 9). from http://passel.unl.edu/pages/informationmodule.php?idinformationmodule=1099683867&topicorder=8&maxto=9&minto=1(2014, April 9).

Crawley, M. J., Brown, S. L., Hails, R. S., Kohn, D. D., & Rees, M. (2001). Transgenic Crops in Natural Habitats.*Nature,409*.

Dudley, J. W., & Lambert, R. J. (1992). Ninety generations of selection for oil and protein in maize (Vol. 37, pp. 1-7). Maydica.

Duvick, D. (1996). Plant Breeding, an Evolutionary Concept (Vol. 36, pp. 539-548): Crop Science.

Eckhoff, S. R. (2004). Wet Milling. In C. Wrigley, *Encyclopedia of Grain Science* (pp. 225-241): Elsevier.

FAO. (2014). from http://faostat3.fao.org/download/Q/QC/E (2014, April 9).

Flint-Garcia, S. A., Bodnar, A. L., & Scott, M. P. (2009). Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte (Vol. 119, pp. 1129-1142): Theoretical and Applied Genetics.

Gaut, B. S., Le Thierry D'Ennequin, M., Peek, A. S., & Sawkins, M. C. (2000). Maize as a model for the evolution of plant nuclear genomes. *PNAS*(97), 7008-7015.

Hallauer, A. R. (2004). Specialty Corns. In *Corn: Origin, History, Technology, and Production* (pp. 897-933). Hoboken, New Jersey: John Wiley & Sons.

Hannah, L. C. (2005). Starch synthesis in the maize endosperm. *Maydica, 50,* 497-506.

Jellum, M. D. (1970). Plant introductions of maize as a sourced of oil with unusual fatty acid composition (Vol. 18, pp. 365-370): Journal of Agricultural and Food Chemistry.

Jennings, S., Myers, D., Johnson, L., & Pollak, L. M. (2002). Effects of maturity on corn starch properties (Vol. 79, pp. 703-706): Cereal Chemistry.

Jeon, J., Ryoo, N., Hahn, T., Walia, H., & Nakamura, Y. (2010). Starch biosynthesis in cereal endosperm. *Plant Physiology and Biochemistry, 48,* 383-392.

Kiesselbach, T. A. (1949). *The structure and reproduction of corn*. Lincoln, Nebraska: University of Nebraska.

Lorenz, A., Scott, P., & Lamkey, K. (2008). Genetic Variation and Breeding Potential of Phytate and Inorganic Phosphorus in a Maize Population (Vol. 48, pp. 79-84): Crop Science Society of America.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Genetics, 36*, 344-355.

Mertz, E. T., Bates, L. S., & Nelson, O. E. (1964). Mutant gene that changes protein composition and increases lysine content of maize endosperm (Vol. 145, pp. 279-280): Science.

Nelson, O. E., Mertz, E. T., & Bates, L. S. (1965). Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science, 150*, 1469-1470.

Orthoefer, F. T., & Eastman, J. (2004). Corn Processing and Products. In *Corn: Origin, History, Technology, and Production* (pp. 867-896). Hoboken, New Jersey: John Wiley & Sons.

Pollak, L. M., & Scott, M. P. (2005). Breeding for Grain Quality Traits (Vol. 50, pp. 247-257): Maydica.

Poneleit, C. G., & Davis, D. L. (1972). Fatty acid composition of oil during maize kernel development (Vol. 11, pp. 3421-3426).

Prasanna S.K.V., B. M., Kassahun, B., & Singh, N. N. (2001). Quality protein maize. *Current Science, 81*, 1308-1319.

Rooney, L. W., McDonough, C. M., & Waniska, R. D. (2004). The Corn Kernel. In C. W. Smith, J. Betran & E. C. A. Runge (Eds.), *Corn: Origin, History, Technology, and Production* (pp. 273-303). Hoboken, New Jersey: John Wiley & Sons.

Scanlon, M. J., & Takacs, E. M. (2009). Kernel Biology. In J. L. Bennetzen & S. C. Hake (Eds.), *Handbook of Maize*. New York, NY: Springer Science+Business Media.

Serna-Saldivar, S. O. (2004). Foods from Maize. In C. Wrigley (Ed.), *Encyclopedia of Grain Science* (pp. 242-253): Elsevier.

Snow, A. A. (2002). Transgenic crops: Why Gene Flow Matters. *Nature Biotechnology, 20,* 542.

Strable, J., & Scanion, M. (2013). Maize (Zea mays): A Model Organism for Basic and Applied Research in Plant Biology (pp. 1-9). Cold Spring Harbor Protocols: Cold Spring Harbor Laboratory Press.

Tucker, J. (2011). *U.S. Regulation of Genetically Modified Crops*, 2014, from http://www.fas.org/biosecurity/education/dualuse-agriculture/2.-agricultural-biotechnology/us-regulation-of-genetically-engineered-crops.html (2014, April 10).

USDA. (2014). *Adoption of genetically engineered crops in the U.S.*, 2014

Vasal, S. K. (2000). High quality protein corn. In A. R. Hallauer (Ed.), *Specialty corns* (2 ed.). Boca Raton: CRC Press.

Velu, V., Nagender, A., Prabhakara Rao, P. G., & Rao, D. G. (2006). Dry milling characteristics of microwave dried maize grains (Vol. 74, pp. 30-36): Journal of Food Engineering.

Vineyard, M. L., & Bear, R. P. (1952). Amylose content (Vol. 26, pp. 5): Maize Gen. Coop. Newsletter.

Vogel, K., & Burson, B. (2004). Chapter 3: Breeding and Genetics (pp. 51-94): American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Wijendran, V., & Hayes, K. C. (2004). Dietary n-6 and n-3 fatty acid balance and cardiovascular health. *Annual Review of Nutrition, 24*, 597-615.

Wilkes, G. (2004). Corn, strange and marvelous: But is a definitive origin known? In W. C. Smith, J. Betran & E. C. A. Runge (Eds.), *Corn: Origin, History,* *Technology, and Production* (pp. 3-64). Hoboken, New Jersey: John Wiley & Sons.

Wu, R., Lou, X.-Y., Ma, C.-X., Wang, X., Larkins, B. A., & Casella, G. (2002). An improved genetic model generates high-resolution mapping of QTL for protein quality in maize endosperm (Vol. 99, pp. 11281-11286): National Academy of Sciences of the United States of America

APPENDIX: ADDITIONAL ALIGNMENT INFORMATION

AC184772.3

```
             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             1                                                                         70
AC184772.3  ATGCCTCCGC CCTTGCCCTC CCCCCC–GGC AATCTCGCAT CGGCGCCCGC CCCAGCCCCG TAGAGGTCGC
            -----TCCGC CCTCGCCCTC CCCCCCCAGC AATCTCGCCT CGGCGCCCCC CCCAGCCCCG CAGAGGTCGC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             70                                                                       140
AC184772.3  ATCCGTCAGC CTCTTCCCCA CCACGGTCCC CCCTTCCCCA CCCACCGGAG ACCGCGCCCT TCCCCTCTTC
            GTCCGTCGGC CTCTTCCCCA CCGTGGTCCC GCCTTCCCCA CCCACCGGAG ACCGCGCCCT TCCCCCCTTC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             141                                                                      210
AC184772.3  CCCACCGCGG CATGGCGTCG GCTTACCCAC CAGAGACTGA TTCCTCCAAC --------CC CATCTCAACC
            CCCACCGTGG CATGGCGTCG GCTTACCCAC CAGAGACCGC TTCCTCCAAC GGTCCAACCC CATCTCGACC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             211                                                                      280
AC184772.3  ATCTGCCCTC CTCAAGTTCC TCGAGCACAG GAGCAGGGGA GGGTTCCACC AGGCCGAGGC GCCATACCAG
            ATCTGCCCTC CTCAGGTTCC TCGAGCACAG GAGCAGGGGA GGGTTCCACC AGGCCGAGGC GCCATACCAG

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             281                                                                      350
AC184772.3  TGCGCTCTCG CTGCACGTGT CGTCGCGTGG GCCTTTGACT TCAGCTTCTC CTTCCTCCCC AGCCACCACC
            TGCGCTCTCG CTGCACGTGT CGTCGCGTGG GCCGTTGACT TCAGCTTCTC CTTCCTCCCC AGCCGCCACC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             351                                                                      420
AC184772.3  GTGTTCGTGG ACCTCGCACC ACGCGATCCC TTGCATCGCC GGTATGTCCA GATCCCCACC ATCCCCAATG
            GCGTTCGTGG ACCTCGCACC ATGCGATCCC TTGCACCGCC GGTACGTCCA GATCCCCACC ATCCCCAGCG

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             421                                                                      490
AC184772.3  AGCTCTTGTC CTCCTCCGTC GTGTCGTACC AAGACGGTGT AGATCTAGAG CATTTCCTAG CACCGGATCT
            AGCTCTTGTC CTCCTCCGTC GTGTCGTACC AAGACGGTGT AGATCTGGAG CATTTCCTAG CGCCGGATCT

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             491                                                                      560
AC184772.3  CAGAGAGGCA AAGGACAAGT CGGTGTTCTA GATGATCTGT AGCCCTCCCG ATCGCCCCTC GCTCTTGCCC
            CGGAGAGGCG AAGGACGAGT CGGTGTTCTA GAGGATATGT AGCCCTCCCG ATCGCCCCTC GCTCTCGCCC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             561                                                                      630
AC184772.3  CCACGACGCG CCCGTGCCCA CACCCGGATA CAGCTGTGTG GGGTTCTATG CGAGTGCGGG GCTGGTGCGC
            CCACGACACG CCCGCGCCC- -----GGAGA TAGCTGCATG GGGTTCTACG CGGGCACGGG GCCGACGCGC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             631                                                                      700
AC184772.3  GAGGTGTGGG CGTCCGTCGA GGAGTTTGAG GCCGTGGGCG ACGGCGCCAC GCCCAACGCC GCGGTGTTCC
            GAGGTGTGGG CATCCGTCGA GGAGTTCGGG GCCGTGGGCG ACGGTGCCAC GCCCAACACT GCGGCGTTCC

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             701                                                                      770
AC184772.3  GGCACGCCGT CATAGAGCTG GGCGTGAGGT CCCCCAGTGG GGGAGGGGCC AGGCTCGACG TGCCCCCAAG
            GGCGCGCCGT CGCGGAGCTG GGCGCGAGGG CCGCCGGTGG GGGAGGGGCC AGGCTCGACG TGCCCCCG-G

             ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
             771                                                                      840
AC184772.3  GAGGTGGCTC ACGGGCAGCT TCAACCTCAC TAGCCGCTTC ACCCTCTTCC TGCACCACGG CGCGGTCATC
            GAGGTGGCTC ACGGGCAGCT TCAACCTCAC TAGCCGCTTC ACCCTCTTCC TGCATCGCGA CGCGGTCATC

             ....|....| ..
             841
AC184772.3  ATCGGCTCCT AG
            ATCGGCTCCC AG
```

GRMZM5G817995

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               1                                                              60
GRMZM5G817995  ATGGTTTATT TGCTTCTCAA ATTGGTATTG CTTTTGCCAG TAGGGGCGGA GGGCCTGTTA
               ATGGTTTATT TGCTTCTCAA ATTGGTATTG CTTTTGCCAG TAGGGGCGGA GGGCCCGTTA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                61                                                           120
GRMZM5G817995  TGGCCAAGTA TAGCCCACGC CATACCTCAA CTCAGCCCAG TCGCAGCCTC CCTGCTAGTG
               TGGCCAAGTA TGGCCCGCGC CATACCTCAA CTCGGCCCAG TCGCAGCCTC CCTGCTAGTG

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               121                                                           180
GRMZM5G817995  CTTCCCCTCT GTGGCCGCGC AACAGGCACA ACTGTAGTAT CGCAGGCGCA CAAGCGAGCC
               CTTCCCCTCT GTGGCCACGC AACAGGCACA ACCGTAGTAC AGCAGGCACA CAAGCGAGCC

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
              181                                                            240
GRMZM5G817995 ATCTGCTCAT CGTTCCGTTC GCGACCGCCT CTGCCTGGCC GGCCGCCAGC TGCCCGAGCA
              GCCTGCTCAT CGTTCCGTTC GCGACCGCCT CCGCCTGGCT GGTCGCCAGC TGCCCGAGCA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               241                                                           300
GRMZM5G817995  CGGCCGCGCG CCTGTGCCTT ACTGAGTCGC CGCCTCTCTG AAAGTTGCCT AAAGGGGGGG
               CGGCCACGCG CCCGTGCCTT ACTAAGTCGC CGCCTCTCTG AAAGTCGCCT AAAGGGGGGA

              ...
              301
GRMZM5G817995 TGA
              TGA
```

GRMZM2G419836

```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        5         15         25         35         45         55
GRMZM2G419836       GCCGTGCGCC TCACATCTTC CCTCCGCCAG TCCGTTGACA CCCCCCCCCC CCCCCCCCCC
                    GCCGTGCGCC TCCCATCTTC TC-------- ---------- ---------- ----------

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       65         75         85         95        105        115
GRMZM2G419836       CTCGGCCATC CACCGGAGAT GGGCGCCGCC GGCAAGCCTC CTCCCCTCGT CTGCTTCAAA
                    ---------- CACCGGAGAT GGGCGCCGCC GGCAAGCCTC CTCCCCTCGT CTGCTTCAAA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      125        135        145        155        165        175
GRMZM2G419836       TGGCCGTGGG GCCCTAATCC TATCCCATCG GCGAGCTCCA GCCCCAGCCC CTGCGGCGAC
                    TGGCCGTGGG GCCCTAATCC TATCCCATCG GCGAGCTCCA GC------CC CTGCGGCGAC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      185        195        205        215        225        235
GRMZM2G419836       CTCGAGCTCC CCTGGCTCTT CAAGTCCATC CGCACCCTCG CGCAGGGCCT CCTCATCGCC
                    CTCGAGCTCC CCTGGCTCTT CAAGTCCATC CGCACCCTCG CGCAGGGCCT CCTCATCGCC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      245        255        265        275        285        295
GRMZM2G419836       GGCGACATCC CCTCCCCCGC CTCTTCTCCC AGCGGAGGAG TAAGGGGCGT TCAGAGGCGC
                    GGCGACATCC CCTCCCCCGC CTCTTCTCCC AGCGGAGGAG TAAGGGGCGT TCAGAGGCGC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      305        315        325        335        345        355
GRMZM2G419836       ACGGGTGCCG CGGTGGTGGA GGTGGACCGC GGGGACGCTG AACAGCGCGC CCTGGCGGCA
                    ACGGGTGCCG CGGTGGTGGA GGTGGACCGC GGGGACGCTG AACAGCGCGC CCTGGCGGCA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      365        375        385        395        405        415
GRMZM2G419836       TCGCTCGCGA GCGGGAGGCC CGCCACGGTG CTGGAGTTCT ACTCCCCGCG CTGCCGCCTG
                    TCGCTCGCGA GCGGGAGGCC CGCCACGGTG CTGGAGTTCT ACTCCCCGCG CTGCCGCCTG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      425        435        445        455        465        475
GRMZM2G419836       TGCGCCTCTT TGCAGGGCCT CGTTCGCGAG CTCCAAGACG GTGCCAGTGG CTCCGCCAGT
                    TGCGCCTCTC TGCAGGGCCT CGTTCGCGAG CTCCAAGACG GTGCCAGTGG CTCCGCCAGT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      485        495        505        515        525        535
GRMZM2G419836       TTCGTGCTCG CTGACGCCGA GGACGACCGG TGGCTCCCCG AGGTATGTCG CCCCTTGCCA
                    TTCGTGCTCG CTGACGCCGA GGACGACCGG TGGCTCCCCG AGGTATGTCG CCCCTTGCCA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      545        555        565        575        585        595
GRMZM2G419836       TCTTCTGGGA AAATTCAGAC AATTTGTCAG ATTTGTGATG CCGATTTGGG TGCTCTGTTC
                    TCTTCTGGGA AAATTCAGAC AATTTGTCAG ATTTGTGATG CCGATTTGGG TGCTCTGTTC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      605        615        625        635        645        655
GRMZM2G419836       TCTACAGAGG AAAGATAAAC CTTTGCAATA GTGATTTAGC CACATAGGTC TTCTTCTGTT
                    TCTACAGAGG AAAGATAAAC CTTTGCAATA GTGATTTAGC CACATAGGTC TTCTTCTGTT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      665        675        685        695        705        715
GRMZM2G419836       AATTGCCTTT GCTATGGTAA TTTAGCCATA TTGGTCATGT TCTGATCAAT TTATGATGAC
                    AATTGCCTTT GCTATGGTAA TTTAGCCATA TTGGTCATGT TCTGATCAAT TTATGATGAC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                      725        735        745        755        765        775
GRMZM2G419836       TAGATGCTAT GTTGCACTTT GATGATGAGA AATTGATGAT TAGAAAATCA GTAGGTTCCA
                    TAGATGCTAT GTTGCACTTT GATGATGAGA AATTGATGAT TAGAAAATCA GTAGGTTCCA
```

```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    785        795        805        815        825        835
GRMZM2G419836    TGGTAAATGA TCCTCCCCTT TTCTTTTAAG GGGTGTTTGG ATCCCTCCAT TTTAAAGAAA
                 TGGTAAATGA TCCTCCCCCT TTCTTTTAAG GGGTGTTTGG ATCCCTCCAT TTTAAATAAA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    845        855        865        875        885        895
GRMZM2G419836    TTGGAATCTA CTTGATAAAG TATGCTATTT GTTTGGAATT TGACATTTTA CCACTTTCCA
                 TTGGAATCTA CTTGATAAAG TATGCTATTT GTTTGGAATT TGACATTCTA CCACTTTCCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    905        915        925        935        945        955
GRMZM2G419836    AAGTTTAGAT ATAAGACTCA AATTCATAGG ATGAGAGAGT TGAAATTGAT TTTATATATC
                 CAGTTTAGAT ATAAGACTCA AATTCATAGG ATGAGAGAGT TGAAATTGAT TTTATATATC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    965        975        985        995        1005       1015
GRMZM2G419836    ACTAGTCTAT GTTTCTACTC TGCAACTTAT AACACGCTCT TCAACTTAGT CCCCTATGAT
                 ACTAGTCTAT GTTTCTACTC TTCAACTTAT AACACGCTCT TCAGCTCAGT CCCCTATGAT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1025       1035       1045       1055       1065       1075
GRMZM2G419836    AGAAATGTAG CACATAAATA TCTCTCTCAT ATGGTTAGCA ATAATATACA AATACTTTCT
                 AGAAATTTAG CACATAAATA TCTCTTTCAT ATGGTTAGCA ATAATATACA AATAC-----

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1085       1095       1105       1115       1125       1135
GRMZM2G419836    ATAAAAATCA TATTAGCTTA ATTGATTTAT GTCTAAATCA CGATTATTAG AATGAAATTG
                 ---------A TATTAGCTTA ATTGATTTAT GTCTAAATTA CGATTATTAG AATGAAATTG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1145       1155       1165       1175       1185       1195
GRMZM2G419836    AATTCCAAGG ATCCAAACGA GGCGCAAGGT TATCATGTTT CATTTGTCTT ATTTACCTCG
                 AATTCCAAGG ATCCAAACTA GGCGCAAGGT TATTATGTTT CATTTGTCCT ATTTACCTCG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1205       1215       1225       1235       1245       1255
GRMZM2G419836    TACAGTGTCA GTTTGAAAAC TTAAGTTCGG TCATCACACC ATTTAGACCA AACATTGCAT
                 TACAGTGTCA GTTTGAAAAC TTAAGTTCGG TCATCACACC ATTTAGACCA AACATTGCAT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1265       1275       1285       1295       1305       1315
GRMZM2G419836    TCAGTTATGT GACTTGCGCA GCTTGAGGGA CAATGCCATG AAAATGGAAA AAAAATTTGG
                 TCAGTTATGT GACTTGCGCA GCTTGAGGGA CAATGCCATG AAAAT----- ----------

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                 Gene Deletion

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1505       1515       1525       1535       1545       1555
GRMZM2G419836    CCCATGAAAA TACCCAGGCG TTTGTTTTGA TTCTTGGACA TTGTGAAGAT TGTCACCTTA
                 ---------- -ACCCAGGCG TTTGTTTTGA TTCTTGGACA TTGTGAAGAT TGTCACCTTA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1565       1575       1585       1595       1605       1615
GRMZM2G419836    GTATTTAATT ACTTTGCACA ACAATGAAAG GGTGAGAAGG AGCTTCATAC GGTATATATG
                 GTATTTAATT ACTTTGCACA ACAATGAAAG GGTGAGAAGG AGCTTCATAC GGTATATATG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1625       1635       1645       1655       1665       1675
GRMZM2G419836    CTATCACTCT TATTACTTAG TTCCACGAGT AGATATGATT TCTAAAGGTT TGCCAGACCA
                 CTATCACTCT TATTACTTAG TTCCACGAGT AGATATGATT TCTAAAGGTT TGTCAGACCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    1685       1695       1705       1715       1725       1735
GRMZM2G419836    ACGCCAACGG TGGTGACATC AAGTGGGCAT TGGTTCCACA TTAGTATGAC TGGTTGAAGA
                 ACGCCAACGG TGGTGACATC AAGTGGGCAT TGGTTCCACA TTAGTATGAC TGGTTGAAGA
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1745       1755       1765       1775       1785       1795
GRMZM2G419836    TATGTGAATA TGTCAGATGG TTGAACATTC ATCCTTGGTG ATGGAAGCAG TGATTTGTTG
                 TATGTGAATA TGTCAGATGG TTGAACATTC ATCCTTGGTG ATGGAAGCAG TGATTTGTTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1805       1815       1825       1835       1845       1855
GRMZM2G419836    AATTGCATCA TGCCTGGCCA TCAAGGGTGT TAAGTTACAA CCAGGGACTT CGGTGCGATA
                 AATTGCATCA TGCCTGGCCA TCAAGGGTGT TAAGTTACAA CCAGGGACTT CGGTGCGATA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1865       1875       1885       1895       1905       1915
GRMZM2G419836    TCTACTTTCT TCACCAGTT- --CACTAATG GAGCATTATA TCAGTTGTTG CTGATGCATA
                 TCTACTTTTT TCACCAGTTG GTCACTAATG GAGCATTATA TCAGTTGTTG CTGATGCATA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1925       1935       1945       1955       1965       1975
GRMZM2G419836    CGGTTTACTT AACTGTTCAA GTAATTAATT GATTATGATA CAGTCGTCAA TTGGTGTCCA
                 CGGTTTACTT AACTGTTCAA GTAATTAATT GATTATGATA CAGTCGTCAA TTGGTGTCCA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1985       1995       2005       2015       2025       2035
GRMZM2G419836    TGCATAAGTA CTTCCTCTGT TCTCAAATTA TTGTTTACTT TGTCTTTGTC CTAAGTCAAA
                 TGCAGAAGTA CTTCCTCTGT TCTCAAATTA TTGTTTACTT TGTCTTTGTC CTAAGTCAAA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2045       2055       2065       2075       2085       2095
GRMZM2G419836    CTATTTTACT CTGACTAAGT TTATAGAAAA A-TGTACTAA CATCTACAAC ATCAAATTAG
                 CTATTTTACT CTGACTAAGT TTATAGAAAA AATGTACTAA CATCTACAAC ATCAAATTAG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2105       2115       2125       2135       2145       2155
GRMZM2G419836    TTTCATTAAA TTATTCATGA AATATATTTT GATATAACTC TTATTCGAAA TTGTAGGTGT
                 TTTCATTAAA TTATTCATGA AATATATTTT GATATAACTC TTATTCGAAA TTGTAGGTGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2165       2175       2185       2195       2205       2215
GRMZM2G419836    TGATACATTT TTCGAAAAAA AAAAACTGTC AAAGCTAGTG AAATTTGGCT TAATACAAAG
                 TGATACATTT TTCGAAAAAA AAAAACTGTC AAAGCTAGTG AAGTTTGGCT TAATACAAAG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2225       2235       2245       2255       2265       2275
GRMZM2G419836    CCAAAGTAAA TTATGATTCA GAGTAGAATG AGTACTATCG TTTTTAATTG GCCAATAGGT
                 CCAAAGTAAA TTATGATTCA GAGTAGAATG AGTACTATCG TTTTTAATTG GCCAATAGGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2285       2295       2305       2315       2325       2335
GRMZM2G419836    TAGTTTACAT TTTAGAAGAA GAATGTGAAT AGAGAGCTCA ACATAGGTTT ACTTGAGGGT
                 TAGTTTACTT TTTAGAAGAA GAAAGTGAAT AGAGAGCTCA ACATAGGTTT ACTTGAGGGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2345       2355       2365       2375       2385       2395
GRMZM2G419836    TGATGGAAAA CCTGCTCTGA CAATTTTGCA TGTGTACGGA TATGTGATAG TTCTGGTGGG
                 TGATGGAAAA CCTGCTCTGA CAATTTTGCA TGTGTACGGA TGTGTGATAG TTCTGGTGGG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2405       2415       2425       2435       2445       2455
GRMZM2G419836    GCTGCTAGTT TTTTAAAACA TGGACTTGTG CGACT--GTG TATTAACTGT GCACGTAAGC
                 GCTGCTAGTT TTTTAAAACA TGGACTTGTG CGACTTTGTG TATTAACTGT GCACGTAAGC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2465       2475       2485       2495       2505       2515
GRMZM2G419836    TATAGGCTGA TATCTCTTCC TTTTACAGGT ATTGGCAAAT GCCAAGTTTA AAATTACGAA
                 TATAGGCTGA TATCTCTTCC TTTTACAGGT ATTGGCAAAT GCCAAGTTTA AAATTACGAA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2525       2535       2545       2555       2565       2575
GRMZM2G419836    ATTTCCCATT CGGATGAGCT GAGGCAGTTA GAATTTATAT TATCGCGTTA AGGTGCTGAA
                 ATTTCCCATT CGGATGAGCT GAGGCAGTTA GAATTTATAT TATCGCGTTA AGGTGCTGAA
```

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2585       2595       2605       2615       2625       2635
GRMZM2G419836   GCGACCACAG GTTTCAGCAT CATAATAGTT CTTGATGATT GAAGCTAATC CACATAGAAC
                GCGACCACAG GTTTCAGCAT CATAATAGTT CTTGATGATT GAAGCTAATC CACATAGAAC

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2645       2655       2665       2675       2685       2695
GRMZM2G419836   AACCAATAAG CACTGTGGGT TGTGCTTCTG CTGCCATAAA ATGACAGTCC TTGTTTACCA
                AACCAATAAG CACTGTGGGT TGTGCTTCTG CTGCCATAAA ATGACAGTCC TTGTTTACCA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2705       2715       2725       2735       2745       2755
GRMZM2G419836   GCCTAGTTTG GATTATGACC TTATTATTTC TTGAATGTAC ATCTGCAACT CTGCACCGGA
                GCCTAGTTTG GATTATGACC TTATTATTTC TTGAATGTAC ATCTGCAACT CTGCACCGGA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2765       2775       2785       2795       2805       2815
GRMZM2G419836   GCATCATACC ACTGCTCCAA GCATATATCA TTTATGTAAA AACTGAAATG AAAATTCAAT
                GCATCATACC ACTGCTCCAA GCATCTATCA TTTATGTAAA AACTGAAATG AAAATTCAAT

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2825       2835       2845       2855       2865       2875
GRMZM2G419836   ATTCTGACAG TCAATTTGTT TTTTAACCGC TTGCAGCTTC TGCATTATGA TATCAGATAC
                ATTCTGACAG TCAATTTATT TTTTAACCGC TTGCAGCTTC TGCATTATGA TATCAGATAC

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2885       2895       2905       2915       2925       2935
GRMZM2G419836   GTCCCTTGCT TCGTGCTCCT GGACAAGCAC GGTAGAGCTC TAGCGAAGAC TGGAGTACCA
                GTCCCTTGCT TCGTGCTCCT GGACAAGCAC GGTAGAGCTC TAGCGAAGAC TGGAGTACCA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  2945       2955       2965       2975       2985       2995
GRMZM2G419836   ACCAGCCGGC AGCACGTTGT CGCCGGTCTC CATCACCTCC TGAGGATGCA GCAGCCATCC
                ACCAGCCGGC AGCACGTTGT CGCCGGTCTC CATCACCTCC TGAGGATGCA GCAGCCATCC

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3005       3015       3025       3035       3045       3055
GRMZM2G419836   GGACTGGAAG GAAACCAGAA TGCGCCTCCG TCATGAAGCC CAAATACCTG AGCAAGGCCT
                GGACTGGAAG GAAACCAGAA TGCGCCTCCG TCATGAAGCC CAAATACCTG AGCAAGGCCT

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3065       3075       3085       3095       3105       3115
GRMZM2G419836   GTATTGACAA AGAAAAATT- TTCAGAATGT GCCTTTTGTT TTTGCAAGCA TGAACAATGG
                GTATTGACAG AGAAAAAAAA TTCAGAATGT GCCTTTTGTT TTTGCAAGCA TGAACAATGG

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3125       3135       3145       3155       3165       3175
GRMZM2G419836   GCAAACATTG ATGCGTTAAT TCTTTAGCTG GTAAGTACAG ATTGAAGTTG GTGCAAAAGC
                GCAAACATTG ATGCGTTAAT TCTTTAGCTT TTTAGTACAG ATTGAAGTTG GTGCAAAAGC

               ....|....| ....|....| ....|....| ....|....| ....
                  3185       3195       3205       3215       3225
GRMZM2G419836   AAAAGGCAGG TGGTATTTTT TTTATGATAT CCGCCTTGAA ATAA
                AAAAGGCAGT TGGTATTTTT TTGGTTAGTA CAGCGTGAAG ACCA
```

GRMZM2G027021

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    5          15         25         35         45         55
GRMZM2G027021   ACCAATCGAA CTGAATGGAC CAGTCGACGT CATCGCCTCC CTCGCCTATC CGCTCGGCCG
                ACCAATCGAA CCGAATGGAC CAGTCGACGT CATCGCCTCC CCCGCCTATC CGCTCGGCCG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    65         75         85         95        105        115
GRMZM2G027021   TTGGCGCTCA CCAAAACCAC CCAGAAGCCT CCGCTTGACC GCTTCACTCG CTTTCCGCCC
                TTGGCGCTCA CCAAAACCAC CCAGAAGCCT CCGCTTGACC GCTTCACTCG CTTTCCGCCC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   125        135        145        155        165        175
GRMZM2G027021   GCCGCGCCAT GAGCGCCGCC GCCTGCCTGT TCGCTGCCGC CGTCTCCCTA TCATTCCCGT
                GCCGCGCCAT GAGCGCCGCC GCCTGCATGT TCGCTGCCGC CGTCTCCCTA TCATTCCCGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   185        195        205        215        225        235
GRMZM2G027021   CGACCTCCGC ACCCTCTTCC GCAAGACGCC GCCGCCTCCG GAGCCCCACC ACCCTCCTCC
                CGACCTCCGC ACCCTCTTCC GCAAGTCGCC GCCGCCTCCG GAGCCCCACC ACCCTCCTCC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   245        255        265        275        285        295
GRMZM2G027021   GCTGCTCCCC GACTCGCCGC CGTGGGCCGG TCCGGCGGAC ACTCGACGAG CGGCTGCTCG
                GCTGCTCCCC GACTCGCCGC CGTGGGCCGG TCCGGCGGAC ACTCGACGAG CAGCTGCTCG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   305        315        325        335        345        355
GRMZM2G027021   AGGCCGCGCC GGCGGAGACC GAAGACGTCC AAACCGCTGT TGATGTAGAG GATGGAGGAG
                AGGCCGCGCC GGCGGAGACC GAAGACGTAC AACCCGCTCT TGATGTAGAG GATGGAGGAG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   365        375        385        395        405        415
GRMZM2G027021   GGATCGCTGA GGGCGATGAA GTGGGAACAG AGGAGATGGA GGAGCTGGAG CAGCGACCGC
                GGATCGCTGA GGGCGATGAA GTGGGAACAG AGGAGATGGA GGAGCTGGAG CAGCGCCCGC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   425        435        445        455        465        475
GRMZM2G027021   CGACGAGGGC TTTCGTGAAG AGCAGGCGGC AGCGGCAGGA AGAGGAGGAA GCCGCGGCGG
                CGCCGAGGGC TTTCGTGAAG AGCAGGCGGC AGCGGCAGGA AGAGGAGGAA GCCGCGGCGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   485        495        505        515        525        535
GRMZM2G027021   GGCAAGACCG GTTCAAGCTC ATCAATGGCA AAGAGGTAGC GGATTGCGTA GCTTCAGCTG
                GGCAAGACCG GTTCAAGCTC ATCAATGGCA AAGAGGTAGC GGATTGCGTA GCTTCAGCTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   545        555        565        575        585        595
GRMZM2G027021   CTTGCTTTTG TTGCTCCGAC AGGCCCGCTT GGCGCCGGCC TGTTTGACAG ATTGGGCGGT
                CTTGCTTTTG TTGCTTCGAC AGGCCCGCCT GGCGCCGGCC TGTTTGACAG ATTGGGCGGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   605        615        625        635        645        655
GRMZM2G027021   TTCTACTCAG CGTGTGGAGA ATATGATAAC CTGCAGCGAT CCATCAAATT CACCGGAGAG
                TTCTACTCAG CGTGTGGAGA ATATGATAAC CTGCAGCGAT CCATCAAATT CACCGGAGAG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   665        675        685        695        705        715
GRMZM2G027021   AACTTTTGAT TGTTACATCC CCGCTAGATA TTTTGGGCCG TGACATGAAC AATAGAGCTG
                AACTTTTGAT TGTTACATCC CCGCTAGATA TTTTGGGCCG TGACATGAAC ATTAGAGCTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   725        735        745        755        765        775
GRMZM2G027021   TGAGTTGGTG TTACCTGCCA GTTTCATCAT GTCTGATTTC TG~~~~AACC TGTGGACCTG
                TGAGTTGGTG TTGCCTGCCA GTTTCATCAT GTCTGATTTC TGTCTGAACC TGTGTACCTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   785        795        805        815        825        835
GRMZM2G027021   GCTACCTGCA GATATTTCAA GAGAAGGCTT ATCTGGTTGG TGTTGAGTGC AAACGGACAG
                GCTACCTGCA GATATTTCAA GAGAAGGCTT ATCTGGTTGG TGTTGAGTGC AAACGGACAG
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   845        855        865        875        885        895
GRMZM2G027021   GAGGGAACCT GTTCGGCATA GAGGAGTCCC TTAAGGAGCT GGAGCAGTTG GCTGATACGG
                GAGGGAACCT GTTCGGCATA GAGGAGTCCC TTAAGGAGCT GGAGCAGTTG GCTGATACGG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   905        915        925        935        945        955
GRMZM2G027021   CGGGCCTTCT GGTAGTCGGC TCAACCTATC AGAAGTAAGC TTCTGTTTGA CGGGAACATC
                CGGGCCTTCT GGTAGTCGGC TCAACCTATC AGAAGTAAGC TTCAGTTTGA CTGGAACATC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   965        975        985        995       1005       1015
GRMZM2G027021   TCGACTGAGC CTGCACTGTG CTCTACTAGC AATCGTGGTT ACACGTTCTC ACCATAGATA
                TCGACTGAGC CTGCGCTGTG CTCTACTAGC AATCATGGTT ACACGTTTTC ACCATAGATA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1025       1035       1045       1055       1065       1075
GRMZM2G027021   AGATGGGACA CCACGGAAAA ACTGAGATGC CTGGTCAATC TAATTCGTGG TCCACAGAAA
                AGATGGGACA CCACGGAAAA ACTGAGATGC CTGGTCAATC TAATTCGTGG TCCACAGAAA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1085       1095       1105       1115       1125       1135
GRMZM2G027021   CTTCACGGGC AACTTGGATA GATGAAATGA TACTGTTAGT TCAGATTTTC AAAATGTACT
                CTTCACGGGC AACTTGGATA GATGAAATGC TACTGTTAGT TCAGATTTTC AAAATGTACT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1145       1155       1165       1175       1185       1195
GRMZM2G027021   CTGCAGCTGT TAGGGCCTAA GAAGGCCCAC GGAGGACTGC AGCAGCAGCA ACGATGGGCC
                CTGCAGC~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  Gene Deletion

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3245       3255       3265       3275       3285       3295
GRMZM2G027021   ATTGCCACAA TCTGAAATAA ATATAGCTCA AATTTTCCTC TTAATTTTCT GTATAAGTTG
                ATTGCCACAA TCTGAAATAA ATATAGCTCA AATTTCCCTC TTAATTTCCT GTATAAGCTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3305       3315       3325       3335       3345       3355
GRMZM2G027021   TATTGTTATG TTCTTATGTA AGATTGTAAG ACTATGG~~~ ~~~~~CATGA CATACATACC
                TATTGTTATG TTCTTATGTA AGATTGTAAG ACTATGCCAG TATGGCATGA CATACAAACC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3365       3375       3385       3395       3405       3415
GRMZM2G027021   ACATGGCTTG CCTTTTCTTA TTTCTACAGA GTACAGTGGT TCTCATGCTT TCTATTTTTC
                ACTTGGCTTG CCTTTTCTTA TTTCTACAGA GTACAGTGGT TCTCATGCTT TCTATTTTCC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3425       3435       3445       3455       3465       3475
GRMZM2G027021   AATAGGCTTT CTACCCCAAA TCCAAGGACT TACATTGGTT CAGGAAAGGT TTCTGAAATC
                AATAGGCTTT CTACCCCAAA TCCAAGGACT TACATTGGTT CAGGAAAGGT TTCTGAAATC

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3485       3495       3505       3515       3525       3535
GRMZM2G027021   AGGACTGCAA TCCAAGCACT TGATGTTGAG ACTGTAATTT TAGACGATGA GTTATCCCCT
                AGGACTGCAA TCCAAGCACT TGATGTTGAG ACTGTAATTT TGGACGATGA GTTATCCCCT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3545       3555       3565       3575       3585       3595
GRMZM2G027021   GGGTAAGATT CTCACTTATT ACTCTGCTTG TTAGAGTACC CGTTTAGGGT TTGGGGTTTA
                GGGTAAGATT CTCACTTATT ACTCTGCTTG TTAGAGTACC CATTTAGGGT TTGGGGTTTA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  3605       3615       3625       3635       3645       3655
GRMZM2G027021   CCCCGTGTAT TTACCTTCTC ACCCCTATGT AAAGGGCCAA GCCTATCTAA CTTAGTCTAT
                CCCCGTGTAT TTACCTTCTC TCCCCTGTGT AAAGGGCCAA ATCCATCTAA CTTAGTCTAT
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       3665       3675       3685       3695       3705       3715
GRMZM2G027021       TAATATATCA CCCAACCCCT TGTTAGGGTT AGGGTTTTCC ACATGGTATA GAGTTAGGTT
                    TAATATATCA CCCAACCCCT TGTTAGGG~~ ~~~~TTTTCC ACATGGTATA GAGATAGGTT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       3725       3735       3745       3755       3765       3775
GRMZM2G027021       TCTTTTTTTC CTCTTCTACT CCCACCCACC CGCCTCCACT TTCCTGCTAG CAAG~~~~~~
                    TCTTTTTT~C CTCTTCTCCT CCCACCCACC CGCCTCCACT TCCCTGCCAG CAAGCCCCAG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    Gene Insertion

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       8945       8955       8965       8975       8985       8995
GRMZM2G027021       ~~~~~~~~~~ ~~~~~~TCCA TCTAACTCAG TCTATTAATA CATAACCCAA CCCCTTGTTA
                    TATGTAATGG GCCAGACCCA TCTAACTAAG TTTATTAATA CATCACCCAA CCTCTTGTTA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9005       9015       9025       9035       9045       9055
GRMZM2G027021       GGGTTAGAGT TTCCCACACT GCTTATGTGA TTCCATTTGA TTTCCGTGTT TGTCATATCT
                    GGGTTAGGGT TTCCCACACT GCTTATGTGA TTCGATCTGA TTTCCGTGTT TGTCATATCT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9065       9075       9085       9095       9105       9115
GRMZM2G027021       GAGACCCGTC AAATGAACCC AACTGTATGA TCTTTGCCTT GTACTAATCG TTAACTATTA
                    AAGACCCGTC AAATGAACCC AACTGTATGA TCTTTGCCTT GTACTAATCG TTAACTATTA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9125       9135       9145       9155       9165       9175
GRMZM2G027021       TGCTCAAAAT ATTGGTCAGT CATCATACTT GTTATCTTCA GTTCAGAGAA TACCTGAAAG
                    TGCTCAAAAT ATTGGTCAGT CATCATACTT GTTATCTTCA GTTCAGAGAA TACCTGAAAG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9185       9195       9205       9215       9225       9235
GRMZM2G027021       AGTGTTATTT GTTAATCTCA TAAATGGATG CCGGTATGTA ATCAAATTTT TATTCTTCCT
                    AGTGTTATTT GTTAATCTCA TAAACGGATG CCAGTATGTA ATCAAATTTT TATTCTTCCT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9245       9255       9265       9275       9285       9295
GRMZM2G027021       CTATACATAA CCAGGATATA TTTGAAGAAT TTATCTTATG ATTTCGACAC CATGTATTGT
                    CTATACATAA CCAGGATATA TTTGAAGAAT TTATCTTATG ATTTCGAAAA CATGTATTGT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9305       9315       9325       9335       9345       9355
GRMZM2G027021       GTCGACCATA CATTGTTTTC AACTTCTTCC TAATCATATT TTAAACTGCT AACCACCTCA
                    GTCGACCATA CATTGTTTTC AACTTCTTCC TAATCATATT TTAAACTGCT AACCACCTCA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9365       9375       9385       9395       9405       9415
GRMZM2G027021       GATTGGCCTA ATAGTTACTC TGGTAGCTCA TATTCCCAAC AATGATTTCA GACAACTACG
                    GATTGGCCTA ATAGTTACTC TGGTTGCTCA TATTCCCAAC AATGATTTCA GACAACTACG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9425       9435       9445       9455       9465       9475
GRMZM2G027021       TAACTTGGAA AAGTCATTTG GTGGGAGTGT CCGAGTCTGT GATCGAACTG CTCTTATTCT
                    TAACTTGGAA AAGTCATTTG GTGGGAGTGT CCGAGTCTGT GATCGAACTG CTCTTATTCT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9485       9495       9505       9515       9525       9535
GRMZM2G027021       TGATATTTTT AATCAAAGGG CAGCAACACA TGAAGCTTCT TTACAGGTAA AAATCACATA
                    TGATATTTTT AATCAAAGGG CAGCAACACA TGAAGCTTCT TTACAGGTAA AAATCACATA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       9545       9555       9565       9575       9585       9595
GRMZM2G027021       CAGTAGCTTT ACCAACAGTA GTATCTTGTG GCATCATTTC TTGACATGAA GTTTGCAGCT
                    CAGTAGCTTC ACCAACAGTA GTATCTTGTG GCATCATTTC TTGACATGAA GTTTGCAGCT
```

```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9605       9615       9625       9635       9645       9655
GRMZM2G027021    TTAAGTAGAG TGAACATGTT TGTTGTCCAC GTAAGTTACT CTATATCATG TTTTCCTTTT
                 TTAAGTAGAG TGAACATGTT TGTTGTCCAC GTAAGTTACT CTATATTATG TTTTCCTTTT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9665       9675       9685       9695       9705       9715
GRMZM2G027021    TAGGTTACTT TGGCACAGAT GGAATATCAA CTTCCTAGGT TGACGAAAAT GTGGAGTCAC
                 TAGGTTACTT TGGCACAGAT GGAATATCAA CTTCCTAGGT TGACGAAAAT GTGGAGTCAC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9725       9735       9745       9755       9765       9775
GRMZM2G027021    CTGGAACGGC AGGCTGGAGG TCAAGTTAAG GGTATGGGTG AGAAGCAAAT TGAAGTTGAC
                 CTGGAACGGC AGGCTGGAGG TCAAGTTAAG GGTATGGGTG AGAAGCAAAT TGAAGTTGAC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9785       9795       9805       9815       9825       9835
GRMZM2G027021    AAGCGCATCT TGAGAACTCA AGTATTACTC TTTCTGGAAG TCATAGATTT TTTTTGCTCA
                 AAGCGCATCT TGAGAACTCA AGTATTACTC TTTCTGGAAG TCAGAGATAT TTTTTGCTCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9845       9855       9865       9875       9885       9895
GRMZM2G027021    ATAATGGACA CATGACTATG TTATTAGCTA CCACTATTGG TCAATGACAG TGCACTCCGT
                 ATAATGGACA CATGACTATG TTATTAGCTA CCACTATTGG TCAATGACAG TGCACTCCGT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9905       9915       9925       9935       9945       9955
GRMZM2G027021    CTCTAATGAC TGGAATAAAA AATAGATGGC TGGTAGACAT TTCCTAATAA AATGGCAAAC
                 CTCTAATGGC TGGAATAAAA AATAGATGGC TGGTAGACAT TTCCTAATAA AATGGCAAAC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   9965       9975       9985       9995      10005      10015
GRMZM2G027021    TTCTATTGAT AATTCATTTG TAGGACTTTA TATTTTCCAT GTCGTATTGT ACATTGCTGA
                 TTCTATTGAT AATTCATTTG TAGGACTTTA TATTTTCCAT GTCGTATTGT ACATTGCTGA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10025      10035      10045      10055      10065      10075
GRMZM2G027021    ATTTGTAGTG CTGATCTTTT TTT~GTGGAA CTTGTGGGTC TCAAACATAA GTGTCATTGA
                 ATTTGTAGTG CTGATCTTTT TTTTGTGGAA CTTGTGGGTC TCAAACATAA GTGTCATTGA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10085      10095      10105      10115      10125      10135
GRMZM2G027021    CAGATAAGTG CCTTGAGGAA AGAATTGGAA TCTGTACGGA AACACCGAAA GTTGTACCGC
                 CAGATAAGTG CCTTGAGGAA AGAATTGGAA TCTGTACGGA AACACCGAAA GTTGTACCGC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10145      10155      10165      10175      10185      10195
GRMZM2G027021    AACCGTCGCC AATCAGTTCC TATTCCTGTT GTTTCTCTGG TATAACCATG TACATTTCTT
                 AACCGTCGCC AATCAGTTCC TATTCCTGTT GTTTCTCTGG TATAACCATG TACATTTCCT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10205      10215      10225      10235      10245      10255
GRMZM2G027021    TACAATAATA AAAAACTATC ATGCTTTCTA TTCTACAAAT ATGTTCAGCT CCAAATAATT
                 TACAATAATA GAAAACTATC ATGCTTTCTA TTCTACAAAT ATGTTCAGCT CCAAATAATT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10265      10275      10285      10295      10305      10315
GRMZM2G027021    TTCAGGTAGG ATATACAAAT GCTGGAAAAA GTACACTCCT GAACCGCTTA ACTGGAGCTG
                 TTCAGGTAGG ATATACAAAT GCTGGGAAAA GTACACTCCT GAACCGCTTA ACTGGAGCTG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10325      10335      10345      10355      10365      10375
GRMZM2G027021    ATGTGCTTGC AGAGGATAAG TTATTTGCCA CATTAGATCC AACTACTAGA AGGGTTTTGG
                 ATGTGCTTGC AGAGGATAAG TTATTTGCCA CATTAGATCC AACTACTAGA AGGGTTTTGG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  10385      10395      10405      10415      10425      10435
GRMZM2G027021    TATGTTATTA GAAAACTCTC CTGGTCCATA AAAAATGGAA ACAAAAGCTT TTTTTGTTAT
                 TATGTTATTA GAAAACTCTC CTGGTCCATA AAAAATGGAA ACAAAAGCTT TTTTTCAT~~
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10445      10455      10465      10475      10485      10495
GRMZM2G027021       GTAAATTGGA TAATGGACAT GAATAAGGTC TGTATCTATT ATGATTTATA TGCCTTTGGG
                    GTAAATTGGA TAATGGACAT GAATAAGGTC TGTATTTATT ATGATTTATA TT~~~~~GGG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10505      10515      10525      10535      10545      10555
GRMZM2G027021       AAAGATTTTT TGTAAGAACT ATCCATCATT ATATCTACAT ATGACCATGA CTGAATGTAA
                    AAAGATTTTT TGTAAGAACT ATCCATCATT ATATCTACAT ATGACCATGA CTGAATGTAA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10565      10575      10585      10595      10605      10615
GRMZM2G027021       TTATGTATTA CTGTGCAGAT GAAGAATGGG ACTGAGTTCC TTCTAACTGA TACCGTCGGA
                    TTATGTATTA CTATGCAGAT GAAGAATGGG ACTGAGTTCC TTCTAACTGA TACCGTCGGA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10625      10635      10645      10655      10665      10675
GRMZM2G027021       TTCATTCAGA AATTACCCAC TATGCTGGTA CATATCCACA AAGCATATTC CTCTTGTTTA
                    TTCATTCAGA AATTACCCAC TATGCTGGTA CATATCCACA AAGCATATTC CTCTTGTTTA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10685      10695      10705      10715      10725      10735
GRMZM2G027021       CATATCCAAC TTTGCATATA TCATTTATTG ATAATACCTT TTCAGGTAGC AGCATTTAGA
                    CATATCCAAC TTTGCATATA TCATTTATTG ATAATACCTT TTCAGGTAGC AGCATTTAGA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10745      10755      10765      10775      10785      10795
GRMZM2G027021       GCAACACTAG AAGAGATATC GGAATCATCA GTTATAGTTC ATCTTGTGGA CATTAGGTAT
                    GCAACACTAG AAGAGATATC AGAATCATCA GTTATAGTTC ATCTTGTGGA CATTAGGTAT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10805      10815      10825      10835      10845      10855
GRMZM2G027021       GGAACTTATA CTAGGGGTTC TCTTCGTTGT GGATTCAATT TCATGCATCT ATATGCAGTT
                    GGAACTTATA CTAGGGGTTC TCTTCGTTGT GGATTCAATT TCATGCATCT ATATGCAGTT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10865      10875      10885      10895      10905      10915
GRMZM2G027021       ATGGACTGTC CTAATATTGT GTTATGTGTT CCAGCCATCC TTTAGCTCAA CAACAGATAG
                    ATGCACTGTC CTAATATTGT GTCATGTGTT CCAGCCATCC TTTAGCTCAA CAACAGATAG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10925      10935      10945      10955      10965      10975
GRMZM2G027021       ATGCTGTTGA AAGAGTACTG AAGGAGTTGG ATGTCGAGTC AATCCCCAAA TTAGTCGTGT
                    ATGCTGTTGA AAGAGTACTG AAGGAGTTGG ATGTCGAGTC AATTCCCAAA TTAGTTGTGT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10985      10995      11005      11015      11025      11035
GRMZM2G027021       GGAATAAGGT TTGTTTGCTC AAATATTTGA CCTGTTTGGT AAAATTTTCA ACGTTTTCAC
                    GGAATAAGGT TTGTTTGCTC AAATATTTGA TCTGTTTGGT AAAATTTTCA ACGTTTTCAC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     11045      11055      11065      11075      11085      11095
GRMZM2G027021       TTTATTTTAT ATTTTTAGAG AGTAGAGATG AGATTTTCTG ATCATAGTCT TTCTATGCTG
                    TTTATTTTAT ATTTTTAGAG AGTAGAGATG AGATTTTCTG ATCATAGTCT TTCTATGCTG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     11105      11115      11125      11135      11145      11155
GRMZM2G027021       GATTTAGTAT CAATTTCTAC TTTCCTATGC TTAATCCCCT ATTTTAAACT TCTCTCTACA
                    GATTTAGTAT CAATTTCTAC TTTCCTATGC TTAATCCCCT ATTTTAAACT TCTCTTTGCA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     11165      11175      11185      11195      11205      11215
GRMZM2G027021       AACGGTGTCA TCTACAGTTC CGCGTCGTCT ATTTTGCACG ATCCACTGAA GACAACCTTA
                    AACGGTGTCA TCTACAGTTC CGCGTCGCCT ATTTTGCATG ATTCACTGAA GACA~~~~TA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     11225      11235      11245      11255      11265      11275
GRMZM2G027021       CGGTGGACTA AAATAGTGTG AAGCTTTTTT GAGCAAAAGT TGTTGCTGAA TGTAAAAGGC
                    CGGTGGATTA AAATAGTGTG AAGCTTTTTT GATCAAAAGT TGTTGCTGAA TGTAAAAGGC
```

```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11285      11295      11305      11315      11325      11335
GRMZM2G027021    TTCTCATTTC TGATCCACCT CTGCCTATCA CTCACTCTGA ATAGATGATG TTCATATAAG
                 TTCTCATTTC TGATCCACCT CTGCCTATCA CTCACTCTGA ATAGATGATG TTCATATAAG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11345      11355      11365      11375      11385      11395
GRMZM2G027021    AAAATTAATG CAGTAGTAAA TCCCTAATAT TTATATAAAT GTTGCAGGGT TCTGTGGAGC
                 AAAATTAATG CAGTAGTAAA TCCCTAATAT TTATATAAAT GTTGCAGGGT TCTGTGGAGC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11405      11415      11425      11435      11445      11455
GRMZM2G027021    TTTGATTTGC ATTAGCTCAT TTTTTA~TCT AATCTTCAAG ATCAATCAGA ATCATAGTCA
                 TTTGATTTGC ATTAGCTCAT TTTTTAATCT AATCTTCAAA ATCAATCAGA ATCATAGTCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11465      11475      11485      11495      11505      11515
GRMZM2G027021    GGAGTTTGTA ATAATAGTGC AAATAATGAT GCAATCATGC AAACAAGACA AAATTATACA
                 GGAGTTTGTA ATAATAGTGC AAATAATGAT GCAATCATGC AAACAAGACA AAATTATACA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11525      11535      11545      11555      11565      11575
GRMZM2G027021    TTTTCAACTG GATCTGATTC TTCAAGTGCT TCCTTTTTGG AACTAAGACA TATTTGTATG
                 TTTTCAACTG GATCTGATTC TTCAAGTGCT TCCTTTTTGG AACTAAGACA TGTTTGTATG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11585      11595      11605      11615      11625      11635
GRMZM2G027021    TCATGCAGAT TGACAATACG GATGAACCAT TGAGTGTAAA AGAGGAGGCT CAGAAACAAG
                 TCATGCAGAT TGACAATACG GATGAACCAT TGAGTGTAAA AGAGGAGGCT CAGAAACAAG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11645      11655      11665      11675      11685      11695
GRMZM2G027021    GAATAATCTG CATATCAGCG ATGAATGGTG ATGGTTTGGA AGATTTATGT AATGCAGTTC
                 GAATAATCTG CATATCAGCG ATGAATGGTG ATGGTTTGGA AGATTTATGT AATGCAGTTC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11705      11715      11725      11735      11745      11755
GRMZM2G027021    AAGCAAAGTT GAAAGTATGT GTTCCCCCCT CGTAGGCAGA GGAGTTGTTT TCCCGACATG
                 AAGCAAAGTT GAAAGTATGT GTTCCCCCCT CGTAGGCAGA GGAGTTGTTT TCCCGACATG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11765      11775      11785      11795      11805      11815
GRMZM2G027021    CCTTTTTGGG TATCTACTGC ACTTATTTAT TTGGATTGGA ATGAAGGGCC TCTGTGGTCC
                 CCTTTTTGGG TATCTACTGC ACTTATTTAT TTGGATTGGA ATGAAGGGCC TCTGTGGTCC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11825      11835      11845      11855      11865      11875
GRMZM2G027021    TGATCTAAGA ATTTTAGGAG CTGGTCATAC CTAGCTCCAG AAATTATTGG AGCCAGAGCT
                 TGATCTAAGA ACTTATGGAG CTGGTCATAC CTAGCTCCAG AAATTATTGG AGCCAGAGCT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11885      11895      11905      11915      11925      11935
GRMZM2G027021    GTAGGCATAT ACGAGTACAT GTTATGCCTA TGGTGCGTCT GGGGCCTGGC CAGGACTCCT
                 GTAGGCATAT ACAAGTACAT GTTATGCCTA TGGTGCGTCA GTCAAGGGGC CTGGCCATGA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  11945      11955      11965      11975      11985      11995
GRMZM2G027021    TAGTTTTAGT TAAATAGATA GGATTAGAT~ ~~~~AAGGTT GTTAGGAGAT AGAGTTGTGG
                 CTCCTTAGTT TTAGTTAAAT AAATAGGATT AGATAAGGTT GTTAGGAGAT AAAGTTGTGG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  12005      12015      12025      12035      12045      12055
GRMZM2G027021    GATTTGTTAG GGGCTGGCTC TATGTAAAGA GAGGCACCAC AGTTAGTTGA GGCAACAATG
                 GATTTGTTAG GGGCTGGCTC TATGTAAAGA GAGGCACCAC AGTTAGTTGA GGCAACAATG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  12065      12075      12085      12095      12105      12115
GRMZM2G027021    AAGAACAGCC AGTCCAATTC CCTCAAATAC TTAGTAGTCT AATCTCCCTC AAAAACCAAC
                 AAGAACAGCC AGTCCAATTC CCTCAAATAC TTAGTAGTCT AATCTCCCTC AAAAACCAAC
```

```
                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12125      12135      12145      12155      12165      12175
GRMZM2G027021     TTTGCCCAGC TATCTTCTTG GCAATGTTAA CCCTAATGAT CTAAGGATCA TAAACACAGA
                  TTTGCCCAGC TATCTTCTTG GCAATGTTAA CCCTAATGAT CTAAGGATCA TAAACACAGA

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12185      12195      12205      12215      12225      12235
GRMZM2G027021     GGGTATTTAG CTGAGGTATT TCCTTATTTT GGATCAATGA CGGATGTCAT ACTCGGTGCT
                  GGGTATTTAG CTGAGGTATT TCCTTATTTT GGATCAATCA CGGATGTCAT ACTCGGTGCT

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12245      12255      12265      12275      12285      12295
GRMZM2G027021     GAAAGCTCCT ACACGATGTG GGGTATGGGG AATGGAATTT CTAGTTAGAG CTGCAGAAGG
                  GAAAGCTCCT ACACGATGTG GGGTATGGGG AATGGAATTT CTAGTTAGAG CTGCAGAAGG

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12305      12315      12325      12335      12345      12355
GRMZM2G027021     GATTGTTGGG GCGAAGGCGA AGACGCTACC CTTCGCTCCA AGCCTTCGTC AACCTCGTCG
                  GA~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  Gene Deletion

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12785      12795      12805      12815      12825      12835
GRMZM2G027021     AGGGATTAAA CACAGATGTT TAAAGCACCT ATTTTATCCT ACTAGTGAAA AAAATCTGCA
                  ~~~~~TTAAA CACAGATGTT TAAAGCACCT ATTTTATCCT ACTAGTGAAA AAA~TCTGCA

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12845      12855      12865      12875      12885      12895
GRMZM2G027021     TAGACTCGTT GACTCATTGT GGTTGTGAGA CCTCCCACTG CCACTAGCTT CTTTAATTCT
                  TAGACTCGAT GACTCATTGT GGTTGTGAGA CCTCCCACTG CCACTAGCTT CTTTAATTCT

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12905      12915      12925      12935      12945      12955
GRMZM2G027021     TGGTGGTGCC ATGCAGCCAG ATCTTTGCTC AAAATGGAAG AAAA~TGATT TAATTTCCTA
                  TGGTGGTGCC ATGCAGCCAG ATCTTTGCTC AAAATGGAAG AAAAATGATT TAATTTCCTA

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   12965      12975      12985      12995      13005      13015
GRMZM2G027021     GTAATCCTAT TTACTTAGGA GCTTTGGAAG TATAGGAATG TCATTATTTT TCAAGGTGTT
                  GTAATCCTAG TGATTTAGGA GCTTTGGAAG TATAGGAATG TCATTGTTTT TCAAGGTGTT

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   13025      13035      13045      13055      13065      13075
GRMZM2G027021     AGGCTAGATG TCCAAAGTGT TGTGT~~GCA GTGGTTACTG AAGGGCAGAT GTGCTGCCTG
                  AGGCTAGATG TCCAAAGTGT TGTGTCTGCA GTGGTTACTG AAGGGCAGAT GTGGTGCCTG

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   13085      13095      13105      13115      13125      13135
GRMZM2G027021     GCGTAGGGCT TTGGCCCTCT AGGACCTGGC CCTTAAGGCT GAACCACTTA GG~~~~~~~~
                  GCGTAGGGCT TTGGCCCTCT AGGACCCC~~ ~~TTAAGGCT GAACCACTTA GGCCTTGTTC

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   13145      13155      13165      13175      13185      13195
GRMZM2G027021     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                  GGTTAATCCC GTTACCTATG AATTGGACGG AATTGAAAAA AATTATGAAG AAATTTGACT

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   13205      13215      13225      13235      13245      13255
GRMZM2G027021     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                  TACTTGAGAT TTAAACCCAC ACAATCCTAA TCAATCTACA TGGATTGAGA GCTAACCGAA

                  ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   13265      13275      13285      13295      13305      13315
GRMZM2G027021     ~~~~~~~~~~ ~~GTCGAAAC GGGGCCAATA GTTTTTAAAT GTGGGTTGTA TTCCACCCTC
                  CAAGCMCTTA AGGTCGAAAC GGGGCCAATA GTTTTTAAAC GTGGGTTGTA TTTCACCCTC
```

```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13325      13335      13345      13355      13365      13375
GRMZM2G027021    TCCCCTGGAG GTGGCTAATC CATCGGGGGA TTCTTTTTTC TTTCTTAATG AAATGAAGCT
                 TCCCCTGGAG GTGGCTAATC CATCGGGGGA TTCCTTTTTC TTTCTTAATG AAACGAAGCT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13385      13395      13405      13415      13425      13435
GRMZM2G027021    CTCCTGTGTG GTTCGAGAAA AAAAATCTGC ATATGAGCTG GAGTTTTGCC AAAGATGATG
                 CTCCTGTGTG GTTCGAGAAA AAAAATCTGC ATATGAGCTG GAGTTTTGCC AAAGATGATG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13445      13455      13465      13475      13485      13495
GRMZM2G027021    TAAATCATGC ATATTTGTCT TCTACAGGAC TCGATGGTTC CTATAGAAGC TTTTGTCCCA
                 TAAATCATGC ATATGTGTCT TCTACAGGAC TCGATGGTTC CTATAGAAGC TTTTGTCCCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13505      13515      13525      13535      13545      13555
GRMZM2G027021    TATGACAAAG GAGATCTCCT GAATGACATA CATAAGGTTG GAATGGTTGA AAAAATGGTG
                 TATGACAAAG GAGATCTCCT GAATGACATA CATAAGGTTG GAATGGTTGA AAAAATGGTG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13565      13575      13585      13595      13605      13615
GRMZM2G027021    AGTGTCCTAT TTGATTTAAG ATGCAGTTTC TTTGGCAATG GTGTTTTTGA GCTTCTGGTT
                 AGTGTCCTAT TTGATTTAAG ATGCAGTTTC TTTGGCAATG GTGTTTTTGA GCTTCTGGTT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13625      13635      13645      13655      13665      13675
GRMZM2G027021    CATGTTGTCA AGTTTCTGCT TTTGTAATTT TGTTCTGGAT GAAATACACG AGTTAATTCA
                 CATGTTGTCA AGTTTCTGCG TTTGTAATTT TGTTCTGGAT GGAATACATG AGTTAATTCA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13685      13695      13705      13715      13725      13735
GRMZM2G027021    TTCAACTACC CCCAAATAGA CAACTTAGGC CTTATTTAAA TGCACTAGAG CTAATAATTA
                 TTCAACTACC TCCAAAGAGA CAACTTAGGC CTTATTTAAA TGCACTAGAG CTAATAGTTA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13745      13755      13765      13775      13785      13795
GRMZM2G027021    GCTGGCTGTT GCCCAACTAA TAGCTGATTT GGTAAAAATA GCTAATAGTT GAACTATTAA
                 GCTGGTTGTT GCCTAACTAA TAGCTGATTT GCTAGAAATA GCTAATAGCT GAACTATTAG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13805      13815      13825      13835      13845      13855
GRMZM2G027021    TTGGGCTGTT TGGATGTTTG CAGCTAATTT TAGCAACTAA CTATTATCTC CTGTGCATTC
                 TTGGGCTGTT TGGATGTTTA CAGCTAATTT TAGCAACTAA TTATTATCTC TAGTGCATTC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13865      13875      13885      13895      13905      13915
GRMZM2G027021    AAACAGGGCC TTAGTCATGG AAGCATGTGC ATGGGTTACT TGTTAAAATT TTCTTTCTGA
                 AAACATGGCC TTAGTCATGG AAGCATGTGC ATGGGTTAAT TGTTAAAATT TTCTTTCTGA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13925      13935      13945      13955      13965      13975
GRMZM2G027021    ATAATCACAC ATTTTTGCTT ATTGCAAATC TGCAAACCTA GATAATATCT AGACATTCCC
                 ATAATCACAC ATTTTTGCTT ATTGCAAATC TGCAAACCTA GATAATATCT AGACATTCCC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  13985      13995      14005      14015      14025      14035
GRMZM2G027021    AAGTACACGA TATATTGATT TCTTGAGAAG CTTTCACTTA ACAGAAAATT TGCTTTGCAT
                 AAGTACACGA TATATTGATT TCTTGAGAAG CTTTCACTTA ACAGAAAATT TGCTTTGCAT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  14045      14055      14065      14075      14085      14095
GRMZM2G027021    TATTGTTTGG ATTTAGTGAT AACTCCCCCC TCTTGCGATA TTCACTGCAG GAGTACAAGG
                 TATTGTTTGG ATTTAATGAT AACTCCCCCC TCTTGCGATA TTCACTGCAG GAGTACAAGG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  14105      14115      14125      14135      14145      14155
GRMZM2G027021    AAAGTGGGAC ATTTGTAAAA GCTCATGTGC CTCTACCTCT GGCAAGGCTT CTCACACCTC
                 AAAGTGGGAC ATTTGTAAAA GCTCATGTGC CTCTACCTCT GGCAAGGCTT CTGACACCTC
```

```
                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14165      14175      14185      14195      14205      14215
GRMZM2G027021      TACGGCAGCA GGTGGCAGCC ACTGTGTGAT GTGCATGTCC CCGATCCCTT GATGCCATTG
                   TACGGCAGCA GGTGGCAGCC ACTGTGTGAT GTGCATGTCC CCGATCCCTT GATGCCATTG

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14225      14235      14245      14255      14265      14275
GRMZM2G027021      GCACTCACAA AATTACCACA TCTTGTAGAT TCACAAAAGG AATAGCTTTG CTGTAGAAAA
                   GCACTCACAA AATTACCACA TCTTGTAGAT TCACAAAAGG AATAGCTTTG CTGTAGAAAA

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14285      14295      14305      14315      14325      14335
GRMZM2G027021      CTTA~~~~~~ GATTATCTTC ATTGTGTTTC TACGGTTCTA CCAGAGTACC GTATCAACAG
                   CTTAATCATA GATTATCTTC ATTCTGTTTC TACGGTTCTA CCAGAGTAGC GTATCAACAG

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14345      14355      14365      14375      14385      14395
GRMZM2G027021      GTGCACAGGA CTAGATAGCT GTATGTACGC ACAACAGAAA TGTAAATGTT CTCAGCAGAA
                   GTGCACAGGA CTAGATAGCT GTATGTACGC ACAACAGAAA TGTAAATGTT CTCAGCAGAA

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14405      14415      14425      14435      14445      14455
GRMZM2G027021      TTTAAG~CCC CGTTTGGTTT GGG~TAG~TC ~~~ACTTTTA GTCCCTAAAA ATATAAACAT
                   TTTAAGGTCC CGTTTGGTTT GGAGTGACTA GTTACTTTTA GTCCCTAAAG AAGCAAACAT

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14465      14475      14485      14495      14505      14515
GRMZM2G027021      GGTGACTAAA ATAGGGTAAC TAAATTTAAG TTCTTTAGTC ATCGAGGAGT GGACTAAAGT
                   GGTGACTAAA GTAGGGTGAC TAAATTTAAG TTCTTTAGCC ACCGAGGAGA C~~~TAAAGT

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14525      14535      14545      14555      14565      14575
GRMZM2G027021      AGGATTTTTA CCTCATTTGC TCTTTTCTTT TTTTTTATTG CAGCAGTCAT CCACTAATTA
                   AGGATTTTTA CCTCATTTGC CTTCTCTTTC TT~~~~AGTG CAGCAGTCAT CTACTAATTA

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14585      14595      14605      14615      14625      14635
GRMZM2G027021      ATAGGAGTAA TATAGTCATT ATTTGCATCA ATTAATGCCT TTTAGTCAGG TTTAGTCACT
                   ATTGGGGTAA TACAGTCATT ATTCGCACCA ATTAATGCCT TTTAGTTAGG TTTAGTCACT

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14645      14655      14665      14675      14685      14695
GRMZM2G027021      GGAACTAAAC CAAACGAGGT ACTTTAGTAA CTAAACTTTA GTCAGGTGAC TAAAGAAACC
                   GGAACTAAAC ~~~~~GGGGT ACTTTAGTGA CTAAAGTTTA GTCAGGTGAC TAAAGAAACC

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14705      14715      14725      14735      14745      14755
GRMZM2G027021      AAACAGGAC~ ~AACTCTCCT TTTCCCAGTT TGAGAATCAT TCTGACTACA AGCATGCGGC
                   AAACATGACC TAACTCTCAT TTTCCCCGTG TGAGAATCAT TCTATCTACA AGCATGTGTC

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14765      14775      14785      14795      14805      14815
GRMZM2G027021      TGCCAACAAG GGGGACTTGA GGGAGGGGGT GACAAGGGTT TTTTTTGGGG GGGGGGGGGG
                   GCTGTGCCAA CTCAAATAGT GAACCCTCTG GTCCCAGATT TGCAGATATA AGAGGCGTTT

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14825      14835      14845      14855      14865      14875
GRMZM2G027021      AGGAATGGGC TTCCCACCGC CGGATCGCAA TCAACGGCCC AAAACCATTC CTACGCCAGA
                   GGATCTAGAT GGCTAAATTT TAGTCTTGTC ACATCGAATT AATGTTGAAT ATTTGACTGT

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    14885      14895      14905      14915      14925      14935
GRMZM2G027021      GCTCCCGACC CCCGCACACA CATCAAACAT AAACTTTACT GTTTTATGGG TGTTTACCTC
                   TAGTTAAAAG TATTAAATAT AATATAATTA TAAAATAAAT TACCTAAATA AGGACTAAAC

                   ....|....| ....|....| ....|....| ....|....| ....|....| ....|....
                    14945      14955      14965      14975      14985      14995
GRMZM2G027021      CTAAAATTCC AAAACCCCCC CCATAGCACG GGACCCGCAA AGAGAAAACA AAGAAAAAA
                   AACAAGATGA ATTTGTTAAG TCTAATTAGT TTATGATTTT TTTTTCGAAA ACGCAGGAG
```

GRMZM2G039971

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               1                                                              60
GRMZM2G039971  CAACTCAAGT GTCCCTTCCA ATGGGTCTTT TT-CTGTGAG GTGCTTGAGA CTGTACCGGT
               CAACTCAAGT GTCCCTTCCA ATGGGTCTTT TTTCTGTGAG GTGCTTGAGA CTGTACCGGT

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               61                                                            120
GRMZM2G039971  GAAAAATAGT ATCTACATCA GCTTAATCGG GTTCTATATC GATTTGTTTG TTCCACACTA
               GAAAAATAGT ATCTACATCA GCTTAATCGG GTTCTATATC GATTTGTTTG TTCCACACTA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               121                                                           180
GRMZM2G039971  TATTTATCGG GTTCCATTGA CCAATCGTCG GATGAAAGCC TCAAGGCTCA TCCATAATCC
               TATTTATCGG GTTCCATTGA CCAATCGTCG GATGAAAGCC TCAAGGCTCA TCCATAATCC

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               181                                                           240
GRMZM2G039971  TCCTCTATCT TAAAAACCAC CAGTACCGTA CAGAGGAAAA GAAGGCGAGA AATGAGAGGA
               TCCTCTATCT TAAAAACCAC CAGTACCGTA CAGAGGAAAA GAAGGCGAGA AATGAGAGGA

               ....|....| ....| ....| ....|.
               241                       266
GRMZM2G039971  AATGGGGAAA AAAAGAAGAGA GAAAAT
               AATGGGGGAA AAAA-AAGAGA GAAAAT
```

GRMZM2G039983

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    5          15         25         35         45         55
GRMZM2G039983   GC~~~~~~GA ACGGACGAAC CCACACCATC ACCACCACCG GCCACCCTCT CCCTGCCCTG
                GCACGAGCGA ACGGACGAAC CCACACCATC ACCACCACCG GCCACCCTCT CCCTGCCCTG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    65         75         85         95        105        115
GRMZM2G039983   GCCCCCCCCG CTTCGCCTAC TCCTGCTCCT CCTCCTCCTC CGCC~~~~TCC CCCTCCCTCC
                GCCCCCCCCG CTTCGCCTAC TCCTTCCCCT CCTCCTCCTC CGCCCCCCTCC ~~~~~~~~~~

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   125        135        145        155        165        175
GRMZM2G039983   TACAAATAGC CACCACCACC ACAGTGACGC AGCCGCCGCC GCAAACGTCG CCCCCGACCG
                TACAAATAGC CACCACCACC ACAGCGACGC CGCCGCCGCC GCAAACGTCG CCCCCGACCG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   185        195        205        215        225        235
GRMZM2G039983   AAGCCTAGCC ACCACCAGCA GCACCAGCAA CCTCGCGTAG CAGCGCTCGA CACCGCTGGA
                AAGCCTAGCC ACCACCAGCA GCACCAGCAA CCTCGCGTAG CAGCGCTCGA CACCGCTGGA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   245        255        265        275        285        295
GRMZM2G039983   CGCCCCGCGC CCGCGCGAAA GCA~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                CGCCCCGCGC CCGCGCGAAA GCAGGTAATTCGC TTACTCTCCT TCGTCCTCMC CGGCCGK

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  Gene Insertion


                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1925       1935       1945       1955       1965       1975
GRMZM2G039983   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ GGACCTTGAT TCTGTCGTTG GCGATACCAT
                AAAAAAGAAA TCGTTTTGGT GTTGTACACA GGACCTTGAT TCTGTCGTTG GCGATACCAT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   1985       1995       2005       2015       2025       2035
GRMZM2G039983   GGATGTGTCC TACGAGAAGT GTGCTGATCC GTCGAACTCG GACCTGCCTA GCGCTGTTGT
                GGATGTGTCC TACGAGAAGT GTGCTGATCC GTCGAACTCG GACCTGCCTA GCGCTGTTGT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2045       2055       2065       2075       2085       2095
GRMZM2G039983   TGATGCTGAG CGATACGACG ATGGCGGCTC CGAACACCTG GGATCTGCTG TAGTAGAGGG
                TGATGCTGAG CGATACGACG ATGGCGGCTC CGAACACCTG GGATCTGCTG TAGTAGAGGG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2105       2115       2125       2135       2145       2155
GRMZM2G039983   AGCTACTGGA AACGAAGGGA ATTCGGGGAC CGAAAGTTCC GAGCAGACTG GTGATG~~~~
                AGCTACTGGA AACGAAGGGA ATTCGGGGAC CGAAAGTTCC GAGCAGACTG GTGATGGTAA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2165       2175       2185       2195       2205       2215
GRMZM2G039983   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                GTTTGATGCT TGKGCCAATT CCAGTGAAGC ATTTAGCTCT TGGATCTGGC CGTTTGCCTM

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2225       2235       2245       2255       2265       2275
GRMZM2G039983   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                GTGTTTRCTG AGTTAGATGC GCAGGRCAAT GCGTGCTGAC MCTGGATCAT GGTTGTAATT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2285       2295       2305       2315       2325       2335
GRMZM2G039983   ~~~~~~~~~~ ~~AGCGCGCT GGAGGAGGTG AAGGCTCTCC TGTTGATGTC GAAAACAGCG
                GTATGAATTC AGAGCGCGCT GGAGGAGGTG AAGGCTCTCC TGTTGATGTC GAAAACAGCG

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   2345       2355       2365       2375       2385       2395
GRMZM2G039983   CTGATAAACA AGAGAGCCAG GAGACGACGG TTCCGATGGA AGAAACAGAA ACGAGCGACG
                CTGATAAACA AGAGAGCCAG GAGACGACGG TTCCGATGGA AGAAACAGAA ACGAGCGACG
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2405       2415       2425       2435       2445       2455
GRMZM2G039983       GCACCTCGAT CACGTCGATG GAGGATGCCC TGGAACCGAA CCGTCATCAC GATCTCCCGT
                    GCACCTCGAT CACGTCGATG GAGGATGCCC TGGAACCGAA CCGTCATCAC GATCTCCCGT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2465       2475       2485       2495       2505       2515
GRMZM2G039983       CGGAGCCTGA GGATGTGGGC AACCACACTC CTGATCCTGA TCAGTCCAGC GGCAAGAACT
                    CGGAGCCTGA GGATGTGGGC AACCACACTC CTGATCCTGA TCAGTCCAGC GGCAAGAACT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2525       2535       2545       2555       2565       2575
GRMZM2G039983       CCAAAGGAAA CAGTAGCGTG TTCCAGAGCG CAAGGAGGGT GCTGGCTTCA ACCAATAAG~
                    CCAAAGGAAA CAGTAGCGTG TTCCAGAGCG CAAGGAGGGT GCTGGCTTCA ACCAATAAGG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2585       2595       2605       2615       2625       2635
GRMZM2G039983       ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                    TGGGTATATC TCCATTTCTC TGAAACCCCC TTTTTTTCCC TTCATGTATG WTCCCATCAA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2645       2655       2665       2675       2685       2695
GRMZM2G039983       ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~AAAA
                    CATTTTTTCT ATCAKAGTCA CACGGAAATA ATGCTCAACA TTTTTTTTTC TTGCAGAAAA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2705       2715       2725       2735       2745       2755
GRMZM2G039983       CTCCATCTGC AACTGCACGG AAGCCACTGC AGTTGACTAA CAGAGGTAAC CAGGATGACG
                    CTCCATCTGC AACTGCACGG AAGCCACTGC AGTTGACTAA CAGAGGTAAC CAGGATGACG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2765       2775       2785       2795       2805       2815
GRMZM2G039983       CGAAATCGTC GGCTGGAAAG GCCGCCACGG TTCCATCAGG CCCGGTTTTC CGCTGTACTG
                    CGAAATCGTC GGCTGGAAAG GCCGCCACGG TTCCATCAGG CCCGGTTTTC CGCTGTACTG

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2825       2835       2845       2855       2865       2875
GRMZM2G039983       AACGGGCCGA GAAGCGCAGA GAA~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                    AACGCGCCGA GAAGCGCAGA GAAGTATGTG ACATAACTTT CTTCTTCTTT TTTTTTTAGA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2885       2895       2905       2915       2925       2935
GRMZM2G039983       ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                    AACTATGAAT CAGAATCTTG GTAAAGGGGG GAATAATGTG GTTATGATTG TTGTTTTCAT

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       2945       2955       2965       2975       2985       2995
GRMZM2G039983       ~~~~~~~~~~ ~~~~TTTTAT ATGAAGCTGG AGGAGAAGCA TCAAGCTATG GAGGAAGAGA
                    GCTTTGCTTC GCAGTTTTAT ATGAAGCTGG AGGAGAAGCA TCAAGCTATG GAGGAAGAGA

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       3005       3015       3025       3035       3045       3055
GRMZM2G039983       AGATTCAGTT GGAGGCTAAG TTGAAG~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
                    AGATTCAGTT GGAGGCTAAG TTGAAGGTAA ATAAATTTAT CTATATGGCT GCCATTTGAC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    Gene Insertion

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       3185       3195       3205       3215       3225       3235
GRMZM2G039983       ~~~~~~~AAA GAGCAGGAGG AGGCACTGAA GCAGCTGAGG AAGAGCCTGA CCTTCAAAGC
                    ATTTCAGAAA GAGCAGGAGG AGGCACTGAA GCAGCTGAGG AAGAGCCTGA CCTTCAAAGC

                    ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       3245       3255       3265       3275       3285       3295
GRMZM2G039983       CAACCCCATG CCGAGCTTCT ACCACGAGGC GACGCCGTCC CCGAAGGCCG AGTTCAAGAA
                    CAACCCCATG CCGAGCTTCT ACCACGAGGC GACGCCGTCC CCGAAGGCCG AGTTCAAGAA
```

```
                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3305       3315       3325       3335       3345       3355
GRMZM2G039983    GCTGCCCACG ACCCGGCCCA AGTCGCCCAA GCTGGGCAGG AGGAAGACGG CCTCGACCTC
                 GCTGCCCACG ACCCGGCCCA AGTCGCCCAA GCTGGGCAGG AGGAAGACGG TCTCGACCTC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3365       3375       3385       3395       3405       3415
GRMZM2G039983    CATGGAGACG TCCAACTCGT CGTCGGAGAG CGAGGGCACG AGGCCGTGCT GCCGCGCCAG
                 CATGGAGACG TCCAACTCGT CGTCTGAGAG CGAGGGCACG AGGCCGTGCT GCCGCGCCAG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3425       3435       3445       3455       3465       3475
GRMZM2G039983    CCGCGACGGC CTCGACAGCA GCTGCAGATG CGGCGGCAGG AGCAGGCCGC AGGCCGCGAA
                 CCGCGACGGC CTCGACAGCA GCTGCAGATG CGGCGGCAGG AGCAGGCCGC AGGCCGCGAA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3485       3495       3505       3515       3525       3535
GRMZM2G039983    CGCCAAGCCG GCCGCCGGGC CCAAGAAGCC GCCGCCGCAG CAGCAGCAGC CGAAACACCG
                 CGCCAAGCCG GCCGCCGGGC CCAAGAAGCC GCCGCCGCAG CAGCAGCAGC CGAAACACCG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3545       3555       3565       3575       3585       3595
GRMZM2G039983    CGCCCACAAG ATCGCCGGCG AGGGCGCCAT CAACATCGCC GTGCACTAGC CGCCGCCGCC
                 CGCCCACAAG ATCGCCGGCG AGGGCGCCAT CAACATCGCC GTGCACTAGC CGCCGCCGC~

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3605       3615       3625       3635       3645       3655
GRMZM2G039983    GCTTCTTGAA ACTTCTTTCC GGTCGCATGC ATGCAGGACG ATGGCGATGG CGTGCGGATT
                 ~~TTCTTGAA ACTTCTTTCC GGTCGCATGC ATGCAGGACG ATGGCGATGG CGTGCGGATT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3665       3675       3685       3695       3705       3715
GRMZM2G039983    TTCCTTCTAA GTTATGAGAG TGCTTTGTCG GCTTGTGGAT TTGGTGTAGA TAATAATATA
                 TTCCTTCTAA GTTATGAGAG TGCTTTGTCG GCTTGTGGAT TTGGTGTAGA TAATAATATA

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3725       3735       3745       3755       3765       3775
GRMZM2G039983    AGTTATGGTG ACGACGAACG AACAGGGGCT GCTGCCACGA GTGAGGCCGG TCAGTCAGAC
                 AGTTATGGTG ACGACGAACG AACAGGGGCT GCTGCCACGA GTGAGGCCGG TCAGTCAGAC

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3785       3795       3805       3815       3825       3835
GRMZM2G039983    AGAGGTGGTG GTGTTTATTG CTTGCTTGCT TGTTTGTCTG TTTGTTTGTT TATTTATGCT
                 AGAGGTGGTG GTGTTTATTG CTTGCTTGCT TGTTTGTCTG TTTCTTTGTT TATTTATGCT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3845       3855       3865       3875       3885       3895
GRMZM2G039983    AATCTTATTT ATTTAATCTG CTGTCGAGGA TGGCCTGCGC ATTGCCACTG TGCAGCGCTG
                 AATCTTATTT ATTTAATCTG CTGTCGAGGA TGGCCTGCGC ATTGCCACTG TGCAGCGCTG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3905       3915       3925       3935       3945       3955
GRMZM2G039983    CTTGTTTTTT ~~~~CGTCTT CTTAATTTAT GGGGAGTGGT AAGAGAGACT TGAGCGCTGG
                 CTTGTTTTTT TTTTCTTCTT CTTAATTTAT GGGGAGTGGT AAGAGAGACT TGAGTGCTGG

                 ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    3965       3975       3985       3995       4005       4015
GRMZM2G039983    ATGTAACGTG TACAAACGAA AACGAAGGCT TGCTGGTGGT GGTGATGGAG GATTTTATCT
                 ATGTAACGTG TACAAACGAA AACGAAGGCT TGCTGGTGGT GGTGATGGAG GATTTTATCT

                 ....|....| ....|....| ....|....| ....|....| ....|....| ..
                   4025       4035       4045       4055       4065
GRMZM2G039983    GAACTATGCT CACTCGCTGC ATTTCTATTG AGTTCTTCAA GAGCTTGCTA AA
                 GAACTATGCT CATTCGCTGC ACTTCTATTG AGTTCTTCAA GAGCTTGCTA AA
```