INVESTIGATION OF GENOMIC STRUCTURAL VARIATION AND THEIR EVOLUTIONARY IMPLICATIONS

by

JIANING LIU

(Under the Direction of R. Kelly Dawe)

ABSTRACT

Genomic structural variants (SV) have a significant impact on phenotypic differences. Characterization of such variants can provide insights into genomic evolution and reveal population dynamics. However, our understanding of structural variants is limited by technical and methodological challenges in genome-wide SV cataloging. In this study, we characterized genomic differences with a variety of next-generation sequencing (NGS) resources, and evaluated their performances on SV detection. As resequencing has its limitations in SV calling, we constructed high-quality genome assemblies and provided insights for structure comparison in repeat domains. To the end, we derived a full spectrum of structural variants across 26 maize lines from whole genome assemblies, and investigated their implication on genome evolution.

Through SV identification with short- and long-reads, we identified differing levels of genomic damages in biolistic transformants of rice and maize. Our results indicated a high likelihood of unintended genomic damages by the transformation method and pointed out the importance of whole-genome variant detection prior to detailed analysis. To capture the true genomic diversity between the B73 reference and a maize line (B73-Ab10) with an abnormal

chromosome 10, we created a *de novo* assembly of B73-Ab10. Through the integration of PacBio, Oxford Nanopore and Bionano technologies, we achieved a highly contiguous assembly which contains two gapless chromosomes and spans multiple tandem repeat arrays. Two adjacent inversions were characterized in Ab10 with genome alignment and the internal structure of tandem repeat arrays was revealed. With the availability of whole-genome assemblies of 26 maize lines, we inferred structural differences and studied the genomic history of *Zea* genus. We generated a full SV catalog across 26 maize inbreds, identified ancient haplotypes and evolutionary strata in pericentromeric area and repeat arrays, and finally inferred the divergence dynamics of maize. We found that recurrent segregation and introgression events took place over the past million years among maize ancestors, and the ancient genomic remnants contribute to the great diversity of maize.

INDEX WORDS: structural variants, DNA damage of repair, next-generation sequencing, Illumina, PacBio, Oxford Nanopore, Bionano, centromere, knobs, whole-genome alignment, genome evolution, haplotype structure, evolutionary strata, maize

INVESTIGATION OF GENOMIC STRUCTURAL VARIATION AND THEIR EVOLUTIONARY IMPLICATIONS

by

JIANING LIU

B.S., Shandong Agricultural University, People's Republic of China, 2012

M.S., University of Nebraska-Lincoln, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

© 2021

Jianing Liu

All Rights Reserved

INVESTIGATION OF GENOMIC STRUCTURAL VARIATION AND THEIR EVOLUTIONARY IMPLICATIONS

by

JIANING LIU

Major Professor:	R. Kelly Dawe
Committee:	James H Leebens-Mack
	Jason Wallace
	Magdy S Alabady
	Paul Schliekelman

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2021

ACKNOWLEDGEMENTS

I would like to thank my advisor Kelly Dawe for encouraging me to think deeply about science and bringing my work to a higher level. I would like to thank the members of my committee, Jim Leebens-Mack, Jason Wallace, Magdy Alabady and Paul Schliekelman, to have provided insightful feedback for my research projects.

I would like to thank all members in the Dawe lab for creating a friendly and pleasant working environment. I would like to give my special thanks to Jonathan Gent and Natalie Nannas, who have contributed to my projects and helped me learn Bioinformatics.

I would also like to thank my collaborators, Todd Michael and Kevin Fengler, for providing wonderful resources for my genomics studies and guiding me through the construction of genome assemblies.

Finally, I would like to thank my family and friends for being supportive and caring.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Genomic structure variation and their phenotypic effects

Structural variations are genomic alterations (>1Kb) in copy number, orientation or chromosome location. These genomic variations are typically categorized into two forms, a balanced type, which includes inversions and translocations, and an unbalanced type that involves loss and gain of sequence, such as duplications, deletions and insertions (Escaramís, Docampo, and Rabionet 2015). Though most SVs are described as single events, multiple types of complex genomic rearrangement patterns have been reported, from overlapped or nested SVs such as an inverted tandem duplication (Sedlazeck et al. 2018), to massive chained events including chromothripsis, breakage-fusion-bridge (BFB) and chromoplexy (Yi and Ju 2018).

Structural variants are major contributors to genomic divergence and have significant impacts on phenotypic variation (Weischenfeldt et al. 2013). A wide range of SV types have been implicated as the cause of human diseases, where most SVs disrupt functional domains like genes and regulatory elements (Carvalho and Lupski 2016). For instance, ALK gene fusions are associated with tumor formation (Stransky et al. 2014), and an architecture change in cis-ergulatory elements could lead to Williams-Bueren syndrome (Carvalho and Lupski 2016). Copy number changes of tandem repeats have also been suggested to cause a range of disorders. For example, expansion of repeat ATTCC is found to be associated with Parkinson's disease (Schüle et al. 2017). Complex SV events such as chromothripsis and BFB cause regional shattering and

copy number switches, and potentially have a catastrophic effect on gene expression (Maher and Wilson 2012). These genomic phenomena are associated with cancer and congenital disorders (Maher and Wilson 2012; Kloosterman and Cuppen 2013).

In plants, structural variants play an important role in environment adaptation, domestication, flowering time and stress response (Saxena, Edwards, and Varshney 2014). In maize, the transposon insertion upstream of genes *ZmCCT9* and *ZmCCT10* affects flowering time and leads to domestication in higher latitude environments. In barley, the gene duplication of *Bot1* increases the tolerance of boron-toxicity in a landrace population (Sutton et al. 2007). In peach, a 1.67Mb inversion accounts for the alteration of fruit shape (J. Guan et al. 2021). Chromothripsis and BFB have also been identified in plants and affect reproductivity and yield (B. McClintock 1941; Mandáková et al. 2019; Carbonell-Bejerano et al. 2017). For instance, a chromothripsis event caused hemizygous deletions of 313 genes, and resulted in low gamete viability and decreased fruit production in grapevine (Carbonell-Bejerano et al. 2017).

Mechanisms of SV formation

There are four major mechanisms that contribute to the origination of structural variants, including errors in recombination, DNA break repair, and replication as well as insertion of mobile elements. Non-allelic homologous recombination (NAHR) is a type of erroneous homologous recombination event that involves misalignment of highly identical sequences (Carvalho and Lupski 2016). Occurrences of NAHR during mitosis and meiosis mediate the formation of large SVs, including segmental duplications, deletions, and translocations (Carvalho and Lupski 2016). Error-prone repair pathways such as Non-homologous end joining (NHEJ) and Microhomology-mediated end joining (MMEJ) fuse the ends of double-stranded

breaks (DSB) with little or short homology. Therefore, DNA repair with these two mechanisms could lead to insertions, deletions and translocations, and even complex events such as chromothripsis (Ottaviani, LeCain, and Sheer 2014). Replication-based processes that result in SV formation include break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR)/falk-stalling and template switching (FoSTeS), and serial replication slippage (SRS) (Carvalho and Lupski 2016; P. Guan and Sung 2016). BIR and MMBIR/FoSTeS can be triggered by a nick on the template strand, which causes fork collapsing or stalling. The main outcome of BIR is copy number changes, while MMBIR/FoSTeS results in complex SV patterns represented by a deletion accompanied by deletion or inversion at breakpoints (Yang et al. 2013; Carvalho and Lupski 2016). SRS normally occurs in the tandem repetitive areas where polymerases are paused and disassociated, causing deletions or expansions within repeat arrays (Usdin, House, and Freudenreich 2015). Finally, mobile element insertion plays an important role reshaping the genomic architecture through creating insertion events and mediating formation of indels (Xing et al. 2009). As the above mechanisms result in unique variant features, the causal events could be inferred from unique characters of SV types and breakpoint junctions (Carvalho and Lupski 2016).

Identification of genomic variation over the past century

Cytogenetic studies

Genomic variant studies were initiated in the early twentieth century. The first chromosomal rearrangement was revealed in *Drosophila* via gene linkage analysis in the 1910s (Strutevant 1913). Translocations and inversions were later identified in *Zea mays* via cytological observation (Barbara McClintock 1931), and tandem duplications were found in

Drosophila through chromosomal banding (Bridges 1936). The power of genomic variations detection was limited by the low resolution of cytogenetic methods at that time, when only large visible features could be characterized, such as chromosomal number changes and highly abnormal variants. The assessment of common variants was not enabled until the emergence of fluorescence in situ hybridization (FISH) technology in the 1970s. Standard FISH could accurately identify large variants above 100Kb, and Fiber-FISH further improved the theoretical power to ~1Kb (Florijn et al. 1995). Up to now, FISH remains to be one of the most reliable methods to identify large translocations and is widely applied for variant characterization (Abel et al. 2014). While FISH is effective in comparative genomic study, conventional implementation of FISH needs to be performed by highly trained professionals and thus has limited its application in a quantitative way (Abel et al. 2014).

Microarrays

Systematic assessment of copy number variation has become possible since the availability of microarray technologies in 1997 (Solinas-Toldo et al. 1997). Two major platforms were developed for genome-wide SV screening, including array Comparative genomic hybridization (aCGH) and SNP array. The aCGH method was developed earlier than SNP array, and enabled detection of copy number changes from 5 to 10Kb at the whole-genome scale in a high-throughput manner (Ren et al. 2005). Compared with aCGH, SNP arrays are more widely applied due to its high specificity and ability to detect heterozygosity (Peiffer et al. 2006; Levy and Burnside 2019). Microarray technologies are advantageous in cost and high throughput, but not adequate to detect copy number neutral rearrangements or identify genomic changes in repetitive regions (Yi and Ju 2018).

Short-read resequencing

Genomic variant studies were revolutionized by the development of next generation sequencing (NGS) technologies in the 2010s. A broad range of variant types, including SNPs, rearrangements and copy number variations, could be characterized in a single experiment at single-nucleotide resolution. Along with the prevalent use of short-read sequencing in genomic studies, different algorithms and tools were developed to characterize variants over the past decade (Kosugi et al. 2019). Four major approaches were used for variant type identification, including read pair, read depth, split read, and assembly, where the read pair and split read methods utilize discordant reads and split reads to identify rearrangements, and the assembly method produces contigs and compares it to reference (P. Guan and Sung 2016; Huddleston et al. 2017; Kosugi et al. 2019; Balachandran and Beck 2020). Aside from simple variants, complex genomic rearrangements were characterized with short-read technologies, such as chromothripsis (Tan et al. 2015; Maher and Wilson 2012), breakage-fusion-bridge (BFB) (Nones et al. 2014), and chromoplexy (Shen 2013).

While the short-read resequencing method is well established, the efficiency of SV detection is limited by the nature of the data (Huddleston et al. 2017). One major concern is its ability to identify insertions. While deletions and duplications could be inferred via read depth, alignment fails for reads derived from non-reference regions. As a result, the ability of insertion identification is dependent on the read length, and short read data could only capture insertions smaller than 1Kb (Sedlazeck et al. 2018). Another matter of short-read data is alignment issues, including a mis-alignment problem and alignment failures in highly repetitive areas. Alignment errors lead to high error rate in variant detection, while alignment failures result in low sensitivity in difficult regions (Sedlazeck et al. 2018).

Long-read technologies and optical mapping

Long-read technologies have provided the opportunity to overcome challenges of short-reads. An average length of 13.5Kb and 25Kb was achieved for PacBio and Oxford Nanopore PromethION platforms respectively (Dohm et al. 2020; Wenger et al. 2019). The length of long reads enabled confident mapping in difficult regions and increased the possibility of spanning the variant breakpoints. Compared with Illumina short-reads, both sensitivity and specificity of SV detection have been greatly improved for the long-read technologies (Sedlazeck et al. 2018; Mahmoud et al. 2019). In addition, nested SVs and complex genome patterns like chromothripsis could be better characterized with long-reads (Lei et al. 2020). The improvement of sequencing accuracy to 99.8% in PacBio HiFi reads further increased the precision for small indel detection, matching that of Illumina reads (Wenger et al. 2019). While long-read technologies have improved whole-genome mapping and the power to accurately detect SVs, large insertions remain to be a challenge when the variant size exceeds read length. In addition, mapping over high repetitive domains still lacks confidence.

Optical mapping is another long-range platform that was developed to improve genome assembly quality and SV calling. This approach creates optical images through marking a specific sequence motif with fluorescent labels. The average molecule length of optical maps is 225Kb, which is almost 10 times longer than both long-read technologies (Shelton et al. 2015). As a result, it is superior in large variant characterization, and could capture mega-base level insertions (Yuan, Chung, and Chan 2020). However, it does not resolve breakpoints at single-nucleotide level and could not provide genomic sequences of the alternative form (Shelton et al. 2015). In addition, as the resolution for alignment depends on label density, sensitivity and specificity for variant characterization varies across the genome.

Whole-genome comparison

The dramatic growth of sequencing technologies over the past decade has allowed us to construct high-quality assemblies at a reduced cost in both resources and computation (Cheng et al. 2021). We have observed remarkable improvements for contig assembly and scaffolding, where assembly contiguity and accuracy have been increased by up to ~100 folds (Jayakumar et al. 2020). With a combination of long-read sequencing and optical maps, we can potentially overcome challenges in tandem repeats and heterozygosity and achieve a telomere-to-telomere construction with fully resolved haplotypes (Miga et al. 2020; Cheng et al. 2021). Along with the capability to rapidly generate genome assemblies, we have now entered a pan-genome era, where multiple reference genomes are generated for a single species and integrated for comprehensive studies (Bayer et al. 2020).

Theoretically, whole-genome assemblies could resolve the two main issues with the read-mapping methods. First, as whole-genome assemblies have stored the information for genetic variants among available genomes, and all forms of SVs including large insertions could be captured. Second, whole-genome alignment greatly improves mapping in repetitive regions and consequently increases the SV detection accuracy (Chakraborty et al. 2019). In addition, comparative studies can be carried out among tandem repeat arrays when assemblies traverse through these areas, which were previously regarded as the ‘dark matter’ among genomes (Liu et al. 2020). Furthermore, a pangenome graph approach to integrate all the SVs across multiple assemblies would lead to benefits in population-scale SV genotyping (Eggertsson et al. 2019). However, to take advantage of the information embedded within whole-genome assemblies, more tools need to be developed for comparative studies. In addition, the power of SV detection depends on the assembly quality of the given input genomes, which is still challenging for complicated genomes, especially regions that involve heterozygosity or tandem repetitiveness.

Inference of evolutionary trajectory from SVs

Despite the long history of structural variant characterization, single nucleotide polymorphisms (SNPs) have been the major focus for population genetic studies. Evolutionary inference with structural variants has been impeded by technical difficulties of full-spectrum SV detection (Conrad and Hurles 2007). In addition, methods to integrate such datasets in population genetics studies are still under exploration (Redon et al. 2006). Now, availability of high-quality multi-genome assemblies has given us an opportunity to obtain a full set of SVs at population level. Through analyzing the frequency of SVs and their distribution in population, we could understand evolutionary processes related to the formation and persistence of these variants.

Like single-nucleotide mutations, SVs follow evolutionary trajectories (Mérot et al. 2020). Genomic structural variants can spread from a single genome to a population or even across species through recombination, introgression, and gene drift (Conrad and Hurles 2007). Upon their formation, variants experience selection and adaptation, which contributes to the development of population stratification (Wellenreuther et al. 2019). In addition, clustered SVs are associated with the establishment of haplotype blocks (Todesco et al. 2020), and could result in linkage disequilibrium through recombination reduction (Mérot et al. 2020). These haplotype regions can reflect demographic isolation and introgression or gene flow. Thus, investigation of genomic variant distribution reveals population structure and provides insights into the traces of introgression events. Comparison of SVs across populations could further enable a survey of evolutionary history across the whole-genome and allow inference of archaic events (Almarri et al. 2020). As this is still a relatively new area, we need to modify existing genomic methods and implement new tools to enable such studies.

The subsequent content of this dissertation details our efforts to study genomic structural variants along with the improvement of technologies and resources. We started the genomic variant study primarily with short-reads in 2016, when Illumina sequencing was widely applied for variant characterization due to its low cost and well-established pipelines. With the significant advancement of long-read technologies in 2018, we started to recognize the necessity of creating high-quality assemblies and employ them for structural variant detection and genome evolution inference. Later in 2020, the release of genome assemblies for 26 maize lines provided a remarkable resource for us to study genomic differences with whole-genome assemblies at population level, and investigate the evolutionary of maize with inferring comparative genomic architecture.

Chapter II is a published manuscript reporting the novel finding that biolistic transformation could lead to a wide range of genomic outcomes, from simple insertions to massive rearrangements. We employed short-reads, PacBio sequencing and optical mapping to characterize the intactness of the transgene and the genomic changes of target plants. Chapter III is another published manuscript describing a method for the gapless assembly of maize chromosomes. We developed an automatic approach to integrate the relative sequencing advantages of PacBio and Oxford Nanopore, and created a highly contiguous assembly of maize. Chapter IV details research aimed at inferring evolutionary history through comparing the genomic structure across 26 maize inbred lines. We characterized the genomic variants across 26 lines, identified haplotypes and evolutionary strata, and interpreted the evolution history of *Zea* through estimation of divergence time.

References

Abel, Haley J., Hussam Al-Kateb, Catherine E. Cottrell, Andrew J. Bredemeyer, Colin C. Pritchard, Allie H. Grossmann, Michelle L. Wallander, John D. Pfeifer, Christina M. Lockwood, and Eric J. Duncavage. 2014. "Detection of Gene Rearrangements in Targeted Clinical next-Generation Sequencing." *The Journal of Molecular Diagnostics: JMD* 16 (4): 405–17.

Almarri, Mohamed A., Anders Bergström, Javier Prado-Martinez, Fengtang Yang, Beiyuan Fu, Alistair S. Dunham, Yuan Chen, Matthew E. Hurles, Chris Tyler-Smith, and Yali Xue. 2020. "Population Structure, Stratification, and Introgression of Human Structural Variation." *Cell* 182 (1): 189–99.e15.

Balachandran, Parithi, and Christine R. Beck. 2020. "Structural Variant Identification and Characterization." *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 28 (1): 31–47.

Bayer, Philipp E., Agnieszka A. Golicz, Armin Scheben, Jacqueline Batley, and David Edwards. 2020. "Plant Pan-Genomes Are the New Reference." *Nature Plants* 6 (8): 914–20.

Bridges, C. B. 1936. "THE BAR 'GENE' A DUPLICATION." *Science*. https://doi.org/10.1126/science.83.2148.210.

Carbonell-Bejerano, Pablo, Carolina Royo, Rafael Torres-Pérez, Jérôme Grimplet, Lucie Fernandez, José Manuel Franco-Zorrilla, Diego Lijavetzky, et al. 2017. "Catastrophic Unbalanced Genome Rearrangements Cause Somatic Loss of Berry Color in Grapevine." *Plant Physiology* 175 (2): 786–801.

Carvalho, Claudia M. B., and James R. Lupski. 2016. "Mechanisms Underlying Structural Variant Formation in Genomic Disorders." *Nature Reviews. Genetics* 17 (4): 224–38.

Chakraborty, Mahul, J. J. Emerson, Stuart J. Macdonald, and Anthony D. Long. 2019. "Structural Variants Exhibit Widespread Allelic Heterogeneity and Shape Variation in Complex Traits." *Nature Communications* 10 (1): 4872.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm." *Nature Methods* 18 (2): 170–75.

Conrad, Donald F., and Matthew E. Hurles. 2007. “The Population Genetics of Structural Variation.” *Nature Genetics* 39 (7): S30–36.

Dohm, Juliane C., Philipp Peters, Nancy Stralis-Pavese, and Heinz Himmelbauer. 2020. “Benchmarking of Long-Read Correction Methods.” *NAR Genomics and Bioinformatics* 2 (2): lqaa037.

Eggertsson, Hannes P., Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T. Hardarson, Daniel F. Gudbjartsson, Kari Stefansson, Bjarni V. Halldorsson, and Pall Melsted. 2019. “GraphTyper2 Enables Population-Scale Genotyping of Structural Variation Using Pangenome Graphs.” *Nature Communications* 10 (1): 5402.

Escaramís, Geòrgia, Elisa Docampo, and Raquel Rabionet. 2015. “A Decade of Structural Variants: Description, History and Methods to Detect Structural Variation.” *Briefings in Functional Genomics* 14 (5): 305–14.

Florijn, R. J., L. A. Bonden, H. Vrolijk, J. Wiegant, J. W. Vaandrager, F. Baas, J. T. den Dunnen, H. J. Tanke, G. J. van Ommen, and A. K. Raap. 1995. “High-Resolution DNA Fiber-FISH for Genomic DNA Mapping and Colour Bar-Coding of Large Genes.” *Human Molecular Genetics* 4 (5): 831–36.

Guan, Jiantao, Yaoguang Xu, Yang Yu, Jun Fu, Fei Ren, Jiying Guo, Jianbo Zhao, Quan Jiang, Jianhua Wei, and Hua Xie. 2021. “Genome Structure Variation Analyses of Peach Reveal Population Dynamics and a 1.67 Mb Causal Inversion for Fruit Shape.” *Genome Biology* 22 (1): 13.

Guan, Peiyong, and Wing-Kin Sung. 2016. “Structural Variation Detection Using next-Generation Sequencing Data: A Comparative Technical Review.” *Methods* 102 (June): 36–49.

Huddleston, John, Mark J. P. Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A. Graves-Lindsay, et al. 2017. “Discovery and Genotyping of Structural Variation from Long-Read Haploid Genome Sequence Data.” *Genome Research* 27 (5): 677–85.

Jayakumar, Vasanthan, Hiromi Ishii, Misato Seki, Wakako Kumita, Takashi Inoue, Sumitaka Hase, Kengo Sato, Hideyuki Okano, Erika Sasaki, and Yasubumi Sakakibara. 2020. “An Improved de Novo Genome Assembly of the Common Marmoset Genome Yields

Improved Contiguity and Increased Mapping Rates of Sequence Data." *BMC Genomics* 21 (Suppl 3): 243.

Kloosterman, Wigard P., and Edwin Cuppen. 2013. "Chromothripsis in Congenital Disorders and Cancer: Similarities and Differences." *Current Opinion in Cell Biology* 25 (3): 341–48.

Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. "Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing." *Genome Biology* 20 (1): 117.

Lei, Ming, Desheng Liang, Yifeng Yang, Satomi Mitsuhashi, Kazutaka Katoh, Noriko Miyake, Martin C. Frith, Lingqian Wu, and Naomichi Matsumoto. 2020. "Long-Read DNA Sequencing Fully Characterized Chromothripsis in a Patient with Langer–Giedion Syndrome and Cornelia de Lange Syndrome-4." *Journal of Human Genetics* 65 (8): 667–74.

Levy, Brynn, and Rachel D. Burnside. 2019. "Are All Chromosome Microarrays the Same? What Clinicians Need to Know." *Prenatal Diagnosis*.

Liu, Jianing, Arun S. Seetharam, Kapeel Chougule, Shujun Ou, Kyle W. Swentowsky, Jonathan I. Gent, Victor Llaca, et al. 2020. "Gapless Assembly of Maize Chromosomes Using Long-Read Technologies." *Genome Biology* 21 (1): 121.

Maher, Christopher A., and Richard K. Wilson. 2012. "Chromothripsis and Human Disease: Piecing Together the Shattering Process." *Cell* 148 (1-2): 29–32.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1): 246.

Mandáková, Terezie, Milan Pouch, Jordan R. Brock, Ihsan A. Al-Shehbaz, and Martin A. Lysak. 2019. "Origin and Evolution of Diploid and Allopolyploid Camelina Genomes Were Accompanied by Chromosome Shattering." *The Plant Cell* 31 (11): 2596–2612.

McClintock, B. 1941. "The Stability of Broken Ends of Chromosomes in Zea Mays." *Genetics* 26 (2): 234–82.

McClintock, Barbara. 1931. *Cytological Observations of Deficiencies Involving Known Genes, Translocations and an Inversion in Zea Mays*. University of Missouri, College of Agriculture, Agricultural Experiment Station.

Mérot, Claire, Rebekah A. Oomen, Anna Tigano, and Maren Wellenreuther. 2020. “A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation.” *Trends in Ecology & Evolution* 35 (7): 561–72.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84.

Nones, Katia, Nicola Waddell, Nicci Wayte, Ann-Marie Patch, Peter Bailey, Felicity Newell, Oliver Holmes, et al. 2014. “Genomic Catastrophes Frequently Arise in Esophageal Adenocarcinoma and Drive Tumorigenesis.” *Nature Communications* 5 (October): 5224.

Ottaviani, Diego, Magdalena LeCain, and Denise Sheer. 2014. “The Role of Microhomology in Genomic Structural Variation.” *Trends in Genetics: TIG* 30 (3): 85–94.

Peiffer, Daniel A., Jennie M. Le, Frank J. Steemers, Weihua Chang, Tony Jenniges, Francisco Garcia, Kirt Haden, et al. 2006. “High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping.” *Genome Research* 16 (9): 1136–48.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444 (7118): 444–54.

Ren, Hua, Wendy Francis, Amber Boys, Anderly C. Chueh, Nick Wong, Phung La, Lee H. Wong, Jacinta Ryan, Howard R. Slater, and K. H. Andy Choo. 2005. “BAC-Based PCR Fragment Microarray: High-Resolution Detection of Chromosomal Deletion and Duplication Breakpoints.” *Human Mutation* 25 (5): 476–82.

Saxena, Rachit K., David Edwards, and Rajeev K. Varshney. 2014. “Structural Variations in Plant Genomes.” *Briefings in Functional Genomics* 13 (4): 296–307.

Schüle, Birgitt, Karen N. McFarland, Kelsey Lee, Yu-Chih Tsai, Khanh-Dung Nguyen, Chao Sun, Mei Liu, et al. 2017. “Parkinson’s Disease Associated with Pure ATXN10 Repeat Expansion.” *NPJ Parkinson’s Disease* 3 (September): 27.

Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. 2018. “Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing.” *Nature Methods* 15 (6): 461–68.

Shelton, Jennifer M., Michelle C. Coleman, Nic Herndon, Nanyan Lu, Ernest T. Lam, Thomas Anantharaman, Palak Sheth, and Susan J. Brown. 2015. “Tools and Pipelines for BioNano Data: Molecule Assembly Pipeline and FASTA Super Scaffolding Tool.” *BMC Genomics* 16 (September): 734.

Shen, Michael M. 2013. “Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome.” *Cancer Cell*.

Solinas-Toldo, Sabina, Stefan Lampel, Stephan Stilgenbauer, Jeremy Nickolenko, Axel Benner, Hartmut Döhner, Thomas Cremer, and Peter Lichter. 1997. “Matrix-Based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances.” *Genes, Chromosomes & Cancer* 20 (4): 399–407.

Stransky, Nicolas, Ethan Cerami, Stefanie Schalm, Joseph L. Kim, and Christoph Lengauer. 2014. “The Landscape of Kinase Fusions in Cancer.” *Nature Communications* 5 (September): 4846.

Strutevant, A. 1913. “The Linear Arrangement of Six Sex-Linked Factors inDrosophila as Shown by Their Mode of Association.” *Zeitschrift Fur Induktive Abstammungs- Und Vererbungslehre* 10 (1): 293–94.

Sutton, Tim, Ute Baumann, Julie Hayes, Nicholas C. Collins, Bu-Jun Shi, Thorsten Schnurbusch, Alison Hay, et al. 2007. “Boron-Toxicity Tolerance in Barley Arising from Efflux Transporter Amplification.” *Science* 318 (5855): 1446–49.

Tan, Ek Han, Isabelle M. Henry, Maruthachalam Ravi, Keith R. Bradnam, Terezie Mandakova, Mohan Pa Marimuthu, Ian Korf, Martin A. Lysak, Luca Comai, and Simon Wl Chan. 2015. “Catastrophic Chromosomal Restructuring during Genome Elimination in Plants.” *eLife* 4 (May). https://doi.org/10.7554/eLife.06516.

Todesco, Marco, Gregory L. Owens, Natalia Bercovich, Jean-Sébastien Légaré, Shaghayegh Soudi, Dylan O. Burge, Kaichi Huang, et al. 2020. “Massive Haplotypes Underlie Ecotypic Differentiation in Sunflowers.” *Nature* 584 (7822): 602–7.

Usdin, Karen, Nealia C. M. House, and Catherine H. Freudenreich. 2015. “Repeat Instability during DNA Repair: Insights from Model Systems.” *Critical Reviews in Biochemistry and Molecular Biology* 50 (2): 142–67.

Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O. Korbel. 2013. "Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease." *Nature Reviews. Genetics* 14 (2): 125–38.

Wellenreuther, Maren, Claire Mérot, Emma Berdan, and Louis Bernatchez. 2019. "Going beyond SNPs: The Role of Structural Genomic Variants in Adaptive Evolution and Species Diversification." *Molecular Ecology* 28 (6): 1203–9.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0217-9.

Xing, Jinchuan, Yuhua Zhang, Kyudong Han, Abdel Halim Salem, Shurjo K. Sen, Chad D. Huff, Qiong Zhou, et al. 2009. "Mobile Elements Create Structural Variation: Analysis of a Complete Human Genome." *Genome Research* 19 (9): 1516–26.

Yang, Lixing, Lovelace J. Luquette, Nils Gehlenborg, Ruibin Xi, Psalm S. Haseley, Chih-Heng Hsieh, Chengsheng Zhang, et al. 2013. "Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes." *Cell* 153 (4): 919–29.

Yi, Kijong, and Young Seok Ju. 2018. "Patterns and Mechanisms of Structural Variations in Human Cancer." *Experimental & Molecular Medicine* 50 (8): 1–11.

Yuan, Yuxuan, Claire Yik-Lok Chung, and Ting-Fung Chan. 2020. "Advances in Optical Mapping for Genomic Research." *Computational and Structural Biotechnology Journal* 18 (August): 2051–62.

CHAPTER 2

GENOME-SCALE SEQUENCE DISRUPTION FOLLOWING BIOLISTIC TRANSFORMATION IN RICE AND MAIZE[1]

[1] Liu, J., Nannas, N. J., Fu, F. F., Shi, J., Aspinwall, B., Parrott, W. A., & Dawe, R. K. (2019). Genome-scale sequence disruption following biolistic transformation in rice and maize. *The Plant Cell*, *31*(2), 368-383. Reprinted here with permission of the publisher.

Abstract

Biolistic transformation delivers nucleic acids into plant cells by bombarding the cells with microprojectiles, which are micron-scale, typically gold particles. Despite the wide use of this technique, little is known about its effect on the cell's genome. We biolistically transformed linear 48 kb phage lambda and two different circular plasmids into rice (*Oryza sativa*) and maize (*Zea mays*) and analyzed the results by whole genome sequencing and optical mapping. While some transgenic events showed simple insertions, others showed extreme genome damage in the form of chromosome truncations, large deletions, partial trisomy, and evidence of chromothripsis and breakage-fusion bridge cycling. Several transgenic events contained megabase-scale arrays of introduced DNA mixed with genomic fragments assembled by non-homologous or microhomology-mediated joining. Damaged regions of the genome, assayed by the presence of small fragments displaced elsewhere, were often repaired without a trace, presumably by homology-dependent repair (HDR). The results suggest a model whereby successful biolistic transformation relies on a combination of end joining to insert foreign DNA and HDR to repair collateral damage caused by the microprojectiles. The differing levels of genome damage observed among transgenic events may reflect the stage of the cell cycle and the availability of templates for HDR.

Introduction

The creation of genetically modified crop lines through transformation is typically performed using *Agrobacterium*-mediated gene transfer (Gelvin, 2017) or particle bombardment (Klein et al., 1989). Both modes of transformation insert recombinant DNA in a random and uncontrolled manner. *Agrobacterium* is viewed as superior because it often delivers complete

gene constructs bounded by known left and right borders (Gelvin, 2017). The integration of *Agrobacterium* transfer DNA (T-DNA) occurs at existing double strand breaks through the activity of native polymerase theta and microhomology-mediated repair (van Kregten et al., 2016). Despite its relative precision, most T-DNA insertions are at least dimers (van Kregten et al., 2016) and many are composed of long arrayed multimers (Jupe et al., 2018; Krizkova and Hrouda, 1998; Cluster et al., 1996). In addition, *Agrobacterium* transformation may result in multiple T-DNA insertions at different locations, large deletions (Kaya et al., 2000; Takano et al., 1997), chromosomal inversions, translocations, and duplications (Jupe et al., 2018; Zhu et al., 2010; Clark and Krysan, 2010; Nacry et al., 1998; Takano et al., 1997; Anderson et al., 2016).

Biolistic transformation offers the advantage that it can deliver any form of DNA, RNA, or protein (Altpeter et al., 2005; Shi et al., 2017; Svitashev et al., 2015; Gil-Humanes et al., 2017), a property that has been exploited to facilitate gene editing technologies (Belhaj et al., 2015; Liang et al., 2017; Begemann et al., 2017; Altpeter et al., 2016). When conditions for biolistic transformation are carefully calibrated, the results can be comparable to Agrobacterium-mediated transformation in terms of transformation efficiency and transgene copy number (Jackson et al., 2013; Lowe et al., 2009). Biolistic transformation is also free of the constraints associated with *Agrobacterium*-host plant interactions. Unaltered BAC sequences larger than 100 kb (Ercolano et al., 2004; Phan et al., 2007) and an intact linear 53 kb molecule (Partier et al., 2017) have been integrated into plants by biolistic methods. Similarly, very long PCR products containing >100 kb of a simple repeating structure (ABS arrays) were co-bombarded with a selectable marker plasmid to create maize transgenics with inserts ranging from ~200 to 1000 kb in size (Zhang et al., 2012). However, transgene copy number following biolistic transformation can be very high (depending on the amount of DNA delivered into cells (Altpeter et al., 2005))

and very little is known about the process or mechanism of insertion following biolistic transformation. Prior literature based primarily on DNA gel blots indicates that sequence breakage and reassembly is common (Shou et al., 2004; Makarevitch et al., 2003; Svitashev et al., 2002; Pawlowski and Somers, 1998, 1996). The only detailed sequence-level analysis of transgenes following biolistic transformation revealed a few large fragments and many small shattered pieces, with 50 of 82 insertions being less than 200 bp in length (Svitashev et al., 2002). These limited sequence data suggest there may be unexpected and severe genomic consequences associated with biolistic transformation.

As a means to better understand the mechanistic underpinnings of biolistic transformation, we transformed linear and circular DNA molecules into rice (*Oryza sativa*) and maize (*Zea mays sp mays*) and subjected the lines to whole genome sequencing and analysis. The data revealed a wide spectrum of insertions and outcomes, from simple insertions to extraordinarily long shattered arrays. Multiple forms of genome damage were observed, including chromosome breakage and shattering and extreme copy number variation. We also found evidence of homology-directed repair (HDR) at sites that had been damaged during transformation. The data indicate that transformation involves both damage to the genome and fragmentation of the input DNA, creating tens to thousands of double stranded breaks that are repaired by end-joining and HDR in ways that can either create simple insertions or cause large structural changes in the genome.

Results

General assessments of the genomes after co-bombardment with lambda and plasmid

We biolistically transformed 48 kb linear lambda phage DNA (Casjens and Hendrix, 2015) and appropriate selectable marker plasmids into rice and maize using a 2-fold (rice) or 4-fold (maize) molar excess of lambda. All sequence analyses were carried out on genomic DNA extracted from cultured callus tissue to obtain an unvarnished view of the transformation process; however three of the rice lines and all of the maize lines were also regenerated to plants (Table S2.1). After screening the transformed callus by PCR to confirm the presence of lambda, we sequenced 14 rice lines and 10 maize lines at low coverage. The data revealed that over a third of the rice events contained less than one copy of lambda while the remaining two-thirds contained approximately 1 to 43 copies (Table S2.1, where copy number is a sequence coverage value, and does not imply that any single lambda is intact). The maize transgenic events showed a similar wide range from approximately 1 to 51 copies (Table S2.1). The selectable marker plasmids were observed at lower abundances reflecting their lower representations during transformation.

To interpret the distribution and structure of the insertions, eight rice lines and four maize lines were sequenced at 20X coverage by 75 bp paired-end Illumina sequencing (Table 2.1). The data were then interpreted using SVDetect, which employs discordant read pairs to predict breakpoint signatures through clustering (Zeitouni et al., 2010), and Lumpy, which uses discordant read pairs and split reads to determine SV types by integrating the probabilities of breakpoint positions (Layer et al., 2014). The paired end Illumina reads were aligned to the rice or maize reference genomes with the complete lambda and plasmid sequences concatenated as separate chromosomes. Insertions were detected as inter-chromosomal translocations between lambda, plasmid and genome, whereas rearrangements were identified as intra-chromosomal translocations. Based on simulations using in silico modified forms of rice chromosome 1 with

randomly inserted lambda/chromosomal fragments, we estimate that our approach identifies about 84% of the breakpoints involving lambda and about 66.5% of the junctions involving two chromosomes but no lambda (Table S2.2).

The sequence data also allowed us to identify deletions and duplications of genomic DNA by changes in read depth as assayed by CNVnator (Abyzov et al., 2011). Unique breakpoints and regions showing copy number variation were plotted using the Circos chromosome visualization software (Figure 2.1 and Figure S2.1 and 2.2). We found a wide range of sequence complexity, ranging from simple insertions to long complex arrays and massive genome-scale disruptions.

Simple low-copy insertions

Four rice lines and two maize lines had one or few insertions and otherwise did not show evidence of genome damage (Figure 2.1, Figure S2.1 and 2.2). In these events, there were fewer than 40 detected breakpoints between lambda, plasmid and chromosomes (Table 2.1), and there were small chromosomal deletions of less than 20 kb around insertion sites (Figure 2.1, Figure S2.1 and 2.2). For example, in rice λ-1 there is a 27 kb insertion composed of rearranged lambda (5.8 kb) and plasmid (21.2 kb) fragments in a region of chromosome 8 that has sustained an 18 kb deletion. Similarly, maize λ-1 contains 86.3 kb of combined lambda and plasmid DNA in chromosome 9 with no deletion at the point of insertion, and rice λ-4 (discussed in detail below) contains a long array of lambda and plasmid fragments in chromosome 2 and a small 9 bp deletion at the site of insertion. In these and other cases of simple insertions, there was no other evidence of chromosome truncation or duplication as judged by read depth.

Creation of long arrays

Several transformants had large amounts of lambda DNA. Rice λ-4 is the simplest of these, with lambda junctions involving three genomic locations (chromosome 2, 9 and 12), and no other evidence of genome damage (Figure 2.1A, 2.1E). As assayed by sequence coverage and SV estimates, this event contains the equivalent of 37 copies of lambda broken into a minimum of 552 pieces. Local sequence assembly indicated that the apparent insertions in chromosome 9 and 12 are small sections of chromosomal DNA flanked on both sides by lambda fragments. Two fragments of chromosome 9 are 102 and 464 bp in length, and one fragment of chromosome 12 is 108 bp in length (see examples in Figure 2.2, Figure S2.3). In contrast, on chromosome 2, the assemblies revealed two simple lambda-genome junctions. These data suggest that rice λ-4 has a large insertion on chromosome 2 and that small sections of chromosome 9 and 12 are intermingled within it. Analysis of 23 self-cross progeny from rice λ-4 supported this view showing that the fragments of chromosome 9 and 12 and the junctions on chromosome 2 are genetically linked (Figure S2.3B).

To confirm our interpretation of the rice λ-4 event, we analyzed the original T0 plant by Bionano optical mapping, where long DNA molecules were fluorescently labeled at the restriction site BspQI, imaged, and assembled into megabase-scale restriction map contigs (Udall and Dawe, 2017). The data revealed no insertions on chromosome 9 or 12, but an unequivocal large insertion on chromosome 2 at the location predicted. There are two assemblies over this region, one for the wild type chromosome 2 and one showing an insertion of at least 1.6 Mb containing novel sequence. The 48-kb lambda molecule contains six BspQI sites in a distinctive pattern. However, Bionano alignment software failed to detect any similarity between lambda and the BspQI recognition pattern within the array on chromosome 2, as expected if lambda

molecules were broken and rearranged. To more accurately assess the internal structure of the array, we sequenced a T1 plant that was homozygous for the insertion on chromosome 2 at 25X coverage using PacBio technology. A total of 1810 (45280/25) lambda fragments ranging in size from 31 to 11387 bp were identified. Over 96% of the lambda fragments were less than 2kb with a mean fragment size of 410 bp (Figure 2.2C).

Evaluation of breakpoint junctions provided information on the mechanisms of repair that operate to create long arrays (Figure S2.4A). The two major forms of non-homologous repair are nonhomologous end joining (NHEJ), which is typified by blunt end junctions and short insertions (Pannunzio et al., 2017), and microhomology-mediated end joining (MMEJ), which is characterized by junctional microhomology of at least 5 bp (McVey and Lee, 2008). Computational analyses of the junctions in rice and maize transgenic events revealed blunt-end connections (25%), short insertions varying in size from 1 to 80 bp (21%) and junctions displaying microhomology in the range of 1-4 nucleotides (50%) and 5-25 nucleotides (4%), suggestive of both NHEJ and MMEJ (Figure S2.4B). It is also possible that some of the longer insertions were an outcome of synthesis-dependent strand annealing (SDSA), an alternative form of homology directed repair pathway (HDR, see below). The four relative orientations of lambda fragments (tail-head, tail-tail, head-tail, and head-head), were nearly uniformly distributed (Figure S2.4C) as expected for a random rejoining process. We also investigated whether the natural overlapping single stranded ends of lambda (the 12-bp cos sites (Casjens and Hendrix, 2015)) may have played a role in multimerization. The data showed that five rice lines and three maize lines contained a single annealed cos site; a low frequency that supports the view that homology-based annealing and ligation have minor roles in the assembly of broken fragments.

Evidence of HDR

A second major form of repair is homology-directed repair (HDR) where double stranded breaks are seamlessly corrected using undamaged homologous molecules such as sister chromatids as templates. If a segment of the genome is broken away and not repaired, we expect to find a deletion at the original coordinates, whereas if the damaged region is repaired by HDR, we expect to find no evidence of damage at the original coordinates. Incorporation of a displaced fragment at a new location followed by repair of the original site will result in a total of three copies of the region affected.

The analysis of rice λ-4 revealed that small sections of chromosome 9 and 12 were included in a long array of lambda fragments but that there were no changes from wild type where the original damage occurred (as assayed by optical mapping, Figure 2.2A). We also analyzed the coordinates surrounding the affected sites on chromosomes 9 and 12 (plus or minus 1 Mb) for a clustering of discordant reads or significant changes in read depth and found no evidence of sequence disruption. Further, PCR analysis of the T0 line revealed no evidence of small deletions at these coordinates. These data are consistent with a model where chromosome 2 and 9 were damaged, broken fragments were included in the assembly of the long chromosome 2 array, and the damaged chromatids were repaired by HDR.

To determine if HDR had occurred in any of the other lines assayed, we identified 78 additional displaced genomic fragments in 4 rice events and 4 maize events. We then systematically checked for increases in read depth and clusters of discordant reads that map to the native locations of these displaced fragments. The data provide evidence of HDR in 3 rice events and 3 maize events (Table S2.3, 2.4). For example, a 110 bp displaced fragment from chromosome 1 and a 69-bp fragment from chromosome 9 in rice λ-5, both flanked by lambda

pieces, exhibited increased sequence coverage by 50% and no apparent deletions at the original coordinates (Figure 2.3). While most of the displaced genomic fragments in lambda arrays were on the order of a few hundred bases, we also found evidence of breakage and repair among the chromosomes on a larger scale (Table S2.3, 2.4). For example in rice λ-8, a 21 kb and a 34 kb region from chromosome 2 were broken away, connected by a small fragment from lambda and reinserted in the genome, followed by repair at the original locations. This resulted in duplication regions clearly visible by read depth (Figure 2.3B). The limits on the size of a deletion that can be repaired by HDR are not known, but in animals HDR can be used to incorporate new (knock-in) constructs as large as 34 kb (He et al., 2016).

It is formally possible that some of the displaced fragments are an outcome of SDSA (Gorbunova and Levy, 1997). SDSA occurs when one strand from a double stranded break invades an intact DNA molecule and begins to initiate DNA synthesis, but is then released and processed by end joining (Verma and Greenberg, 2016). Under this model the DNA scored as displaced would actually have been copied from an undamaged location. However, SDSA events tend to be short (<50 bases; (Kleinboelting et al., 2015)) and this mechanism probably cannot explain the longer displaced regions we have observed (13 are larger than 1 kb, Table S2.3, 2.4). The fact that the majority of displaced fragments are associated with deletions at the original location is also tends to favor the HDR model over the SDSA model.

Deletions and evidence of breakage-fusion-bridge cycling

Copy number profiling provided evidence for many deletions ranging in size from 3.5 kb to 11.9 Mb in rice and 115 kb to 62 Mb in maize. Deletions and duplications/triplications greater than 1 Mb were found in four rice events and three maize events (Figure 2.1E, Figure S2.2B).

Deletions were particularly common around transgene insertions and at the ends of chromosomes, and the majority were associated with the presence of lambda or plasmid DNA, indicating that the breaks occurred as a consequence of the transformation process. Deletions that appeared to have no connection with lambda or plasmid may either reflect our imperfect (84%) ability to detect such junctions, or identify regions that were damaged and repaired without the involvement of introduced DNA. No deletions were observed in the single non-transformed rice callus line used as a control.

Chromosome breakage is expected to yield a double stranded break that is repaired by ligation to an introduced DNA molecule or to another broken chromosome. The fusion of two different chromosomes can cause the formation of a dicentric chromosome that is unstable during mitosis. When the centromeres on a dicentric chromosome move in opposite directions during anaphase, the pulling forces cause a re-breaking of the chromosome that initiates a breakage-fusion-bridge (BFB) cycle that may repeat for many cell divisions (Storchová and Kloosterman, 2016; McClintock, 1942; Zakov et al., 2013). The BFB cycle can lead to local duplications and higher order expansions (Mardin et al., 2015; Campbell et al., 2010). Chromosomes 4 and 7 in rice λ-8 show complex rearrangements and evidence of trisomy (Figure 2.4) that is consistent with errors at the level of chromosome segregation. Copy number gains were observed on chromosome 6 in maize λ-4, where the amplified regions are adjacent to a terminal deletion (Figure 2.4D). At least two inversions of 3.9 Mb and 2.8 Mb were found in the amplified area. Read depth increases adjacent to a terminal deletion were also found on chromosome 9 in maize λ-3, where the amplified region displayed switches from 2 to 6 copies (Figure 2.4C).

Shattering and chromothripsis-like outcomes

Animal cells sustaining chromosome loss or breakage undergo a process known as chromothripsis that results in complex genomic rearrangements in localized areas, generally consisting of tens to hundreds of small pieces (Korbel and Campbell, 2013; Stephens et al., 2011). The reassembly process involves a reshuffling and loss of sections of the genome. Instead of uniform coverage, a region that has undergone chromothripsis shows oscillations from the normal copy number state of two to a copy number state of one (haploid) and occasionally three (triploid) in the context of numerous rearrangements. Analysis of the rice and maize transgenic events revealed similar oscillating copy number states in regions surrounding what appear as "impact sites" on Circos displays: large areas of genome damage with multiple lambda and plasmid fragments.

We found particularly complex rearrangements with copy number oscillations and interspersed lambda and plasmid fragments in three rice events. In rice λ-7, broken fragments (44bp to 7858bp) from localized regions of chromosome 3, 5, 6, 7, 9 and 11 were interlinked along with lambda and plasmid fragments in inverted and non-inverted orientations (Figure 2.1B, 2.1C, 2.1D, Figure S2.5). These patchwork assemblages are presumably integrated into one or a few arrays. The damage imparted during transformation caused large swathes of the same regions on chromosomes 3, 5, 6, 7, 9 and 11 to be deleted. The combination of retained displaced fragments and deletions results in oscillating patterns between 1 and 2 copy number states (Figure 2.1C, 2.1D). Higher order oscillation patterns were identified in rice λ-8, where numerous fragments from chromosome 1 were linked with segments of chromosome 2, 4, 7, 9 and 11 in what is likely another complex array (Figure 2.4A). However in this case the read depth data indicate that the damaged regions of chromosome 1 were repaired by HDR. The

combination of retained displaced fragments and repaired regions result in oscillating patterns between 2 and 3 copy number states (Figure 2.4B).

Similar results were found in three maize events where large deletions and duplications occurred. The sensitivity of our assay is significantly lower in maize because of the high repeat content and necessity of using only perfectly mapped reads. Although we can only detect a fraction of the rearrangements present, the linking patterns between displaced genomic segments and lambda and plasmid is obvious (Figure 2.4C, 2.4D and Figure S2.2A). For example, maize λ-4 shows lambda and plasmid within an inter-chromosomal network including sections of chromosomes 1, 5, 6, 7, and 9, as well as evidence of copy number switching (Figure 2.4D).

Similar genome scale disturbances in single plasmid transformations

We were concerned that the linearity of lambda or the high concentration of DNA used when transforming lambda may have led to new or extreme forms of genome damage. To test whether this was the case, we transformed rice with circular plasmids designed to knockdown (pANIC10A-OsFPGS1) or overexpress (pANIC12A-OsFPGS1) folylpolyglutamate synthetase 1 (chosen for its presumptive role in regulating lignin content). Approximately 125 ng of DNA was delivered to 100 mg of callus tissue per shot, which is considerably lower than the 585 ng of DNA delivered for lambda. In addition, we only sequenced the genomes of fully regenerated plants in these experiments.

The rice lines transformed with single plasmids showed a narrower span of transgene copy numbers (ranging from 0.5 to 12.3X, Table S2.5), consistent with the lower amount of DNA used in transformation (Lowe et al., 2009). However, the genome-level damage (average inter- and intra-chromosome breakages, 17.9) was nearly identical to what we observed for the

lambda transformation experiments (19.5 for rice and 17.8 for maize). In general, while there is a natural relationship between transgene copy number and the number of junctions between the plasmid or lambda and the genome (the transgenes must insert somewhere) the copy number of the transgene was not correlated with the level of collateral damage at other genomic sites. Lines with one copy of the transgene are just as likely to have sustained damage elsewhere in the genome than lines with multiple copies (Figure 2.5E).

As in the lambda experiments, single plasmid transformations caused large-scale deletions, inversions, duplications consistent with BFB, and rearrangement patterns indicative of chromothripsis-like processes (Figure 2.5, Figure S2.6 and 2.7). For example, in event 12A-6, chromosome 4 sustained a large deletion and the remainder of the chromosome was duplicated to create a region of partial trisomy (Figure 2.5D). Evidence of alternating copy number states was found on chromosome 1 in event 10A-6 (Figure S2.6B), and chromosome 8 in event 12A-3 (Figure S2.7B).

Discussion

Here we provide data showing that biolistically transformed rice and maize plants contain a wide diversity of transgene copy numbers ranging from a fraction of a single copy to as many as 51. While it is known that lowering the amounts of input DNA (<1 ng/kb of input DNA per shot) can result in more single copy insertions (as high as 54% in maize (Lowe et al., 2009)), single copy insertions are also commonly observed when higher amounts of input DNA are used to improve transformation efficiency (~10 ng/kb of input DNA per shot; e.g. (Raji et al., 2018; Li et al., 2016)). Seven of the 24 events we analyzed had less than 1.5 copies of the plasmid by read depth (Table 2.1, Table S2.5). Our expectation based on prior work was that lines with multiple

transgenes would contain complex arrays of broken and rearranged plasmids (Register et al., 1994; Gorbunova and Levy, 1997; Kohli et al., 1999; Svitashev et al., 2000; Makarevitch et al., 2003; Jackson et al., 2001; Shou et al., 2004). Key among the early studies was work from the Somers lab (Makarevitch et al., 2003; Svitashev et al., 2002) showing that plasmids transformed biolistically are frequently broken into small (<100 bp) pieces and scrambled with genomic segments. Our results strongly support these interpretations, illustrated most vividly by our analysis of the long lambda array in rice λ-4, which contained total of 1810 lambda fragments ranging in size from 31 to 11387 bp (Figure 2.2C). The Somers group further speculated that DNA was broken randomly and rejoined at blunt ends often containing microhomology (Svitashev et al., 2002). Our more extensive analysis implicates NHEJ as the primary mechanism for rejoining broken fragments and that MMEJ and perhaps SDSA is involved as a secondary pathway.

In addition to confirming the broken and rearranged fate of transgenes following biolistic transformation, we found massive genome rearrangements on a scale that would have been difficult to anticipate. Our focus on callus tissue gave us a perspective on the outcome of transformation than might not have been visible had we worked entirely with regenerated plants. Callus is known to tolerate chromosome instability (Lee and Phillips, 1988) and is presumably more tolerant of mutations than differentiated tissue. Likewise, our use of long linear molecules allowed us to visualize DNA rearrangements with greater ease than would have been possible with plasmids alone. Nevertheless the same types of breakages and copy number variation were observed with single plasmid transformants assayed in regenerated plants. Most of the major events were associated with fragments of introduced DNA, implicating the microprojectiles themselves as the primary mutagens. Such damage is to be expected, as the 0.45 μm gold beads

used for rice transformation are about a quarter of the diameter of a rice nucleus (about 2 µm (Jones and Rost, 1989)) and 225 times larger than the diameter of DNA. When the genome is damaged in this manner, it can be repaired in one of three ways (Figure 2.6):

1) Repair can occur by homology-directed repair such that the damaged region is completely restored to its original state (Figure 2.6D).
2) Repair can occur by NHEJ or MMEJ, where the end of any other broken DNA molecule is used as a substrate. Broken fragments of introduced DNA are a likely substrate particularly when they contain markers that are under selection. The other end of the newly joined fragment may then be ligated to a second fragment of introduced DNA or to another segment of the genome. If this process culminates by reconnecting the two pieces of the original chromosome, the result will be a "simple insertion" containing a variable number of conjoined foreign DNA fragments (Figure 2.6A).
3) Repair can be initiated by the process above but not culminate in the reconstitution of the original chromosome. The break may not be repaired at all or it may culminate in connecting of two different chromosomes. In this case there can be severe genomic consequences including large terminal deficiencies, chromosome fusion and BFB cycling, and more complex events resembling chromothripsis (Figure 2.6B, 6C). These dramatic chromosomal rearrangements are a natural outcome of the same processes that are used to create a simple insertion.

The stage of the cell cycle may have a significant impact on the outcome of biolistic transformation. Data from non-plant systems indicate that while NHEJ is active throughout interphase, it is particularly important in G1. In contrast, HDR is more likely in S and G2 phases after DNA replication has provided additional templates for repair (Heyer et al., 2010; Ceccaldi

et al., 2016; Karanam et al., 2012). Simple insertions may be more probable when the cell is transformed in S or G2 so that NHEJ can insert the foreign DNA while HDR serves to repair extraneous damage. Simple insertions may also be an outcome of transformation during mitosis when chromosomes are distributed in the cytoplasm. DNA introduced during metaphase or anaphase might find its way into newly-forming telophase nuclei, and subsequently be inserted into the genome as consequence of routine DNA repair (similar to T-DNA (van Kregten et al., 2016)).

When chromosomes are broken in G1, deletions and translocations are to be expected. We observed many examples of chromosomes that were missing large terminal segments of chromosome arms (Figure 2.1B, 2.3A, 2.4C, 2.4D, and 2.5C, 2.5D). The formation of a stable truncated chromosome requires that the end be healed by formation of a telomere which is a process that occurs over a period of cell divisions (McClintock, 1941; Chabchoub et al., 2007). In the period when there is an unattended double strand break without a stable telomere, the break is likely be repaired by NHEJ using any other broken chromosome. As famously described by Barbara McClintock (McClintock, 1941), the fusion of broken chromosomes can initiate a breakage-fusion-breakage (BFB) cycle and amplification of genome segments on the affected chromosomes. In several cases we observed copy number states of 4, 5 and 6 that are difficult to explain by any other mechanism. We also observed partial and fully trisomic chromosomes (Figure 2.4A, 2.5D). Such large-scale chromosome abnormalities may also be the result of the tissue culture process itself (Lee and Phillips, 1988), and we cannot rule out the possibility that some of the chromosomal changes were either present before transformation or occurred after transformation. However, for most of the large duplications and deletions we observed, there was

either evidence of inserted foreign DNA or evidence that the lost DNA had been fragmentation and rejoined with foreign DNA.

In addition, our analyses revealed extreme shattering and chromothripsis-like outcomes. Chromothripsis was originally described as a process whereby "tens to hundreds of genomic rearrangements occur in a one-off cellular crisis (Stephens et al., 2011)." Our data meet this definition in a descriptive sense but the biological underpinnings are presumably different. For cancer lines, the simplest model (as it relates to our study, for other models, see (Rode et al., 2016)) requires that a chromosome be partitioned from the primary nucleus, generally as a result of an error in chromosome segregation that leaves it stranded in the cytoplasm (Zhang et al., 2015). The resulting micronuclei show aberrant DNA replication (Leibowitz et al., 2015; Crasta et al., 2012; Zhang et al., 2015) and appear to have fragmented chromatin (Crasta et al., 2012). The partially degraded chromatin can then be reincorporated into the primary nucleus where it is evident as broken and reassociated fragments (Rode et al., 2016). Recent data indicate that when plants sustain errors in chromosome segregation, they too show evidence of chromosome fragmentation with oscillating copy number states confined to single chromosomes (Tan et al., 2015). In contrast, our biolistically transformed lines are not expected to undergo regular loss of chromosomes during cell division. It is possible that microprojectiles severely damage nuclei such that portions of the genome are released into the cytoplasm. Another plausible explanation is that acentric fragments formed during the repair process (Figure 2.6B, 2.6C) are lost during anaphase, become partially degraded, and are reincorporated into a nucleus during a subsequent cell division.

Taken together our data help to explain the long nearly continuous arrays of 156 bp repeats we observed following biolistic transformation of PCR products in maize (Zhang et al.,

2012). At the time we were unable to determine whether the long PCR products had been transferred intact or were broken and reassembled in planta. Based on the data here it seems more likely that the PCR products were fragmented and reassembled by NHEJ to create the observed long arrays. While we did recover simple low copy insertions, the conditions used were not ideal for recovering this type of event at high frequencies. Researchers wishing to do so would be well served to lower the amounts of DNA and consider employing linearized plasmids or amplified fragments that are more likely to be inserted at low copy numbers (Fu et al., 2000; Tassy et al., 2014). Constructs as long as 53 kb have been recovered with careful selection for low copy inserts (Partier et al., 2017) although this kind of success is rare.

From a product development perspective, genomic rearrangements were initially considered to be a food/feed safety hazard (Kessler et al., 1992). To put this hazard in perspective, Anderson et al (Anderson et al., 2016) noted that the genomic rearrangements from *Agrobacterium*-mediated transformation were an order of magnitude lower than those created by fast-neutron mutagenesis. In turn, rearrangements from fast-neutron mutagenesis were an order of magnitude lower than the standing genomic structural variations in the cultivated soybean germplasm pool, all of which has a history of safe use. The frequency of rearrangements from biolistic transformation may be more comparable to that induced by fast neutrons. Regardless, as of yet, there is no evidence that a genomic rearrangement has compromised the safety of a plant used as food (Weber et al., 2012), though its agronomic performance can be compromised. Since poor agronomic performance is not tolerated in modern cultivars and hybrids, there is a rigorous selection process that eliminates deleterious mutations during the breeding process (Glenn et al., 2017).

From a research perspective, such rearrangements may be acceptable in some cases, while in others it may be necessary to consider that undetected rearrangements could be influencing the phenotype. Gene editing applications are a special case where the intent is usually to make a single precise change. Although there is great appeal in directly introducing Cas9 ribonucleoproteins (Liang et al., 2017) and repair templates (Altpeter et al., 2016) for this purpose, our data suggest that there is strong likelihood that the delivery method itself will cause unintended genome damage. Until new transformation methods become available, the *Agrobacterium*-based methods that have been in regular use for decades (Gelvin, 2017) remain the superior alternative in terms of minimizing genome rearrangements.

Methods

Rice transformation

Rice (*Oryza sativa*) variety Taipei 309 was transformed as described previously using 0.45 μm gold beads (Phan et al., 2007). For the lambda experiments, we mixed 33 ng of the 5839 bp plasmid pPvUbi2H (Mann et al., 2012) which confers hygromycin resistance and a two fold molar excess (552 ng) of purified lambda DNA cI857 (New England Biolabs #N3011S). This equates to 5.6 ng/kb of plasmid DNA per shot and 11.0 ng/kb of lambda per shot. After screening for lambda by PCR (forward primer 5'-GACTCTGCCGCCGTCATAAAATGG and reverse primer 5'-TCGGGAGATAGTAATTAGCATCCGCC), 14 callus lines were chosen for sequence analysis. Three of these callus lines were regenerated to mature rice plants (Table S2.1).

The plasmids pANIC10A-OsFPGS1 (17603 bp) and pANIC12A-OsFPGS1 (17501 bp) are based on the pANIC backbone (Mann et al., 2012) with inserts designed to silence or

overexpress folylpolyglutamate synthetase. In these experiments only plasmid DNA was used, delivering ~125 ng per shot. This equates to 7.1 ng/kb of plasmid per shot. All 12 lines were regenerated to plants.

Maize transformation

Biolistic transformation of the maize (*Zea mays*) inbred Hi-II was performed by the Iowa State University Transformation Facility (Ames, IA) as previously described using 0.6 µm gold beads (Frame et al., 2000). To achieve a four molar excess of lambda DNA, we mixed 20 ng of the 7121 bp plasmid pBAR184 which confers resistance to glyphosate (Frame et al., 2000) and 528 ng of purified lambda DNA cI857 (New England Biolabs #N3011S). This equates to 2.8 ng/kb of plasmid DNA per shot and 11.5 ng/kb of lambda per shot. Ten callus lines were screened for lambda by PCR (forward primer 5'-GACTCTGCCGCCGTCATAAAATGG and reverse primer 5'-TCGGGAGATAGTAATTAGCATCCGCC) and subjected to sequence analysis. All of these lines were later regenerated to mature maize plants (Table S2.1).

Library preparation and sequencing

DNA was extracted by the CTAB method (Clarke, 2009) and libraries were prepared using KAPA Hyper Prep kit and KAPA Single-Indexed Adapter kit for Illumina Platforms (KAPA Biosciences, KK8504 and KK8700). For the lambda experiments, 14 rice and 10 maize lines were skim sequenced at low coverage (~1x) using Illumina NextSeq PE35. Of those, eight rice and four maize lines were chosen for deeper sequencing using Illumina NextSeq PE75, achieving an average coverage of 20X for rice and 15X for maize. For plasmid experiments, six lines each transformed with either pANIC10A-OsFPGS1 or pANIC12A-OsFPGS1 were sequenced with Illumina NextSeq PE75 at ~20X.

Copy number calculation

The lambda and plasmid sequences were added to the rice (Kawahara et al., 2013) and maize reference genomes (Jiao et al., 2017) as separate chromosomes to construct concatenated genomes, which were then used as references for read alignment by BWA-mem (version 0.7.15) with default parameters (Li, 2013). For skim-sequenced lines, the mean coverage of lambda/plasmid and genome in each event was estimated as the division of the the total number of reads mapped to individual sequences by their respective genome sizes. For lines sequenced at high coverage, the average read depth of lambda/plasmid and genome was calculated as the mean of per-base coverage analyzed by bedtools (version 2.26). The copy number of lambda/plasmid was then derived by multiplying the mean coverage by two, considering that the insertions are heterozygous in diploid genomes.

SV calling

After adapter trimming by trimgalore (version 0.4.4) and quality checking by fastqc (version 0.11.3) at default settings, reads were aligned to the rice/maize concatenated genomes where lambda and plasmid sequences were added as separate chromosomes, using BWA-MEM (version 0.7.15) with default parameters. PCR duplicates were removed by Picard's MarkDuplicates (version 2.4.1) and MAPQ filter of 20 was applied. The output BAM files were analyzed for structural variants by SVDetect (version 0.7) (Zeitouni et al., 2010) and Lumpy (version 0.2.13) (Layer et al., 2014) to call inter-chromosomal translocations and intra-chromosomal translocations. For SVDetect, step length and window size were calculated separately for each sample and structural variants supported by fewer than two reads were filtered. For Lumpy, the mean and standard deviation of insert sizes were calculated for each sample, with two reads set as minimum weight for a call and trim threshold set as 0. For intra-

chromosomal translocations, the read cutoff for both Lumpy and SVDetect was set at 3 to increase accuracy. Structural variants in each event called (from both Lumpy and SVDetect) were filtered against those of wild type and the other events with an in-house script. Unique breakpoints were manually inspected with IGV (version 2.3.81) and plotted with Circos (version 0.69) (Krzywinski et al., 2009).

Data simulation

We performed four sets of simulations by embedding shattered and reshuffled fragments from lambda or other chromosomes into the rice reference chromosome 1 sequence at random sites (Table S2.2). Subsequently, a heterozygous diploid genome was constructed by concatenating the modified chromosome with reference chromosome 1. Paired-end Illumina reads were then simulated by ART (version 2.5.8) (Huang et al., 2012) at coverage 10X. For ART, the Illumina sequencing system was set as NextSeq 500 v2 (75 bp), average fragment size and standard deviation were set to 300 and 80 bp respectively. The inter- and intra- chromosomal translocations in each simulated data set were then identified with the SV calling pipeline described above. The output of Lumpy and SVDetect was compared with the simulated data to assess detection performance.

Junction assembly and validation by PCR

Reads that support lambda-genome junctions identified by both SVDetect and Lumpy were assembled by SPAdes (version 3.10.0) (Bankevich et al., 2012) with default parameters. The output sequences were aligned against to the reference genome with NCBI BLAST (version 2.2.26) at default parameters and used as templates for primer design. The BLAST output of all rice events transformed with lambda was then subjected to analyzing microhomology at junction

sites and identifying relative orientations between ligated fragments with an in-house script. Selected products of PCR were sequenced and aligned to the assemblies.

Copy number variation detection

CNVnator (version 0.3.3) (Abyzov et al., 2011) was employed to call copy number variation on BAM files where mapping quality was set to be at least 20. The bin sizes for rice and maize genomes were set at 500 and 5,000 bp respectively. The CNVnator output was filtered by removing calls with q0 > 0.5 and eval1 > 0.01 using an in-house script. We declared copy number variation as a deletion if the copy number in specific sample is between 0.5 and 1.5, and at least 0.5 lower than that in wild type and all other samples (unaltered regions are expected to have copy numbers between 1.5 to 2.5). We declared copy number variation as duplication if the copy number in specific sample is between 2.5 and 10, and at least 0.5 higher than in wild type and all other samples. Copy number variations in non-repetitive regions where breakpoints were identified were further inspected using IGV.

It is possible that some of the callus samples were chimeric and contained tissue from more than one independent transformation event. Using our filtering pipeline, if a callus sample contained two deletion events in roughly equal proportions we might have detected both, but if the proportions were not equal the less abundant one would most likely not have be detected. We would be unlikely to detect duplications if they were chimeric.

Bionano optical mapping

High molecular weight DNA was prepared from rice λ-4 young leaf tissue using the IrysPrep Plant Tissue DNA Isolation Kit (RE-014-05) and labeled with Nt.BspQI using the IrysPrep NRLS labeling kit (RE-012-10). Data were collected at the Georgia Genomics and Bioinformatics Core facility on a single BioNano IrysChip at 80X coverage with an average

molecule length of 248 kb. The raw data were assembled with IrysView software (version 2.5.1) set to "optArgument_human", resulting in 501 BioNano genome maps with an N50 of 1.050 Mbp. The genome maps were then aligned to an silico-digested BspQI cmap of the rice Nipponbare reference genome. Overall alignment was excellent, yielding a "Total Unique Aligned Len / Ref Len" value of 0.946, which exceeds the general recommendation of 0.85 (Udall and Dawe, 2017). Potential SVs were identified using Bionano Solve software (version 3.0.1) and analyzed individually by eye. In addition to the large insertion on chromosome 2, the SV calling software identified several other regions with small (<100 kb), potential insertions in the rice λ-4 sample relative to the Nipponbare reference (Chr3: 31,097,744, Chr12: 20,548,710, Chr1: 2,287,999, Chr3: 13,461,815, Chr3: 16,548,253, Chr6: 3,287,514, Chr7: 22,897,940). As none of these correspond to the coordinates of lambda or plasmid junctions identified by sequence analysis, they may either represent differences between the Taipei 309 line (used for transformation) and the Nipponbare reference, errors in either assembly, or small insertions caused by biolistic transformation but not involving lambda or the plasmid.

PacBio sequencing and analysis

High molecular weight DNA was prepared using a modified CTAB method (Healey et al., 2014) from young leaf tissue. The single plant was one of the 23 progeny from the original λ-4 transformant and was homozygous for the large insertion on chromosome 2. The PacBio library was prepared following SMRTbell library guidelines. The library was sequenced with three SMRT cells to generate 10.32 Gb of long reads with N50s ranging from 16-18 kb. Consensus sequences were created from subread BAM files using SMRTLink (version 5.1) with parameters: min_length 50, max_length 30,000, minPredictedAccuracy 0.8, minZScore -3.4, minPasses 0, maxDropFraction 0.34, and polish. The derived consensus reads were then mapped

to lambda sequence with BLASR (Chaisson and Tesler, 2012) at default settings. The BAM output file of mapped reads was converted to fasta format by samtools (version 1.3.1) and aligned against lambda full sequence using NCBI blast (version 2.2.26) at default parameters. The blast output was filtered by removing reads with E-values higher than 0.1 and lengths shorter than 30 bp, and by retaining the longest consecutive matches for each read using an in-house script.

Code availability

The custom code required for analysis in this study is available at the GitHub repository (https://github.com/dawelab/Genome-Rearrangements).

Accession Numbers

Raw PacBio and Illumina sequence data are available from the NCBI Short Read Archive under BioProject PRJNA508943.

Table 2.1. Copy number of introduced molecules (lambda and co-bombarded plasmid) and number of breakpoints in rice/maize transgenic genome

Transgenic Events[+]	Genome coverage	Copy Number		Number of breakpoints					
		Lambda	Plasmid	Lambda-Lambda	Lambda-plasmid	Lambda-genome	Plasmid-genome	Intra-genome	Inter-genome
Os λ-1	22.45	0.12	3.63	1	11	0	2	2	1
Os λ-2	20.95	0.95	0.80	19	1	3	0	2	0
Os λ-3	20.84	0.04	1.37	0	3	1	2	1	1
Os λ-4	19.63	32.48	3.87	517	21	14	0	1	0
Os λ-5	21.58	37.06	2.22	420	18	18	0	1	0
Os λ-6	21.78	17.35	1.18	257	6	123	0	1	13
Os λ-7	19.64	1.72	1.07	51	4	63	3	12	14
Os λ-8	21.84	7.83	0.98	152	3	99	1	67	40
Zm λ-1	14.07	1.73	1.88	22	22	17	6	1	0
Zm λ-2	18.05	19.11	13.97	31	30	15	5	14	19
Zm λ-3	15.55	10.33	2.00	12	8	8	1	1	8
Zm λ-4	17.93	40.48	20.50	241	143	73	4	10	18

[+] Transgenic events of rice (*Oryza sativa*) and Maize (*Zea mays*) are labeled as "Os" and "Zm" respectively.

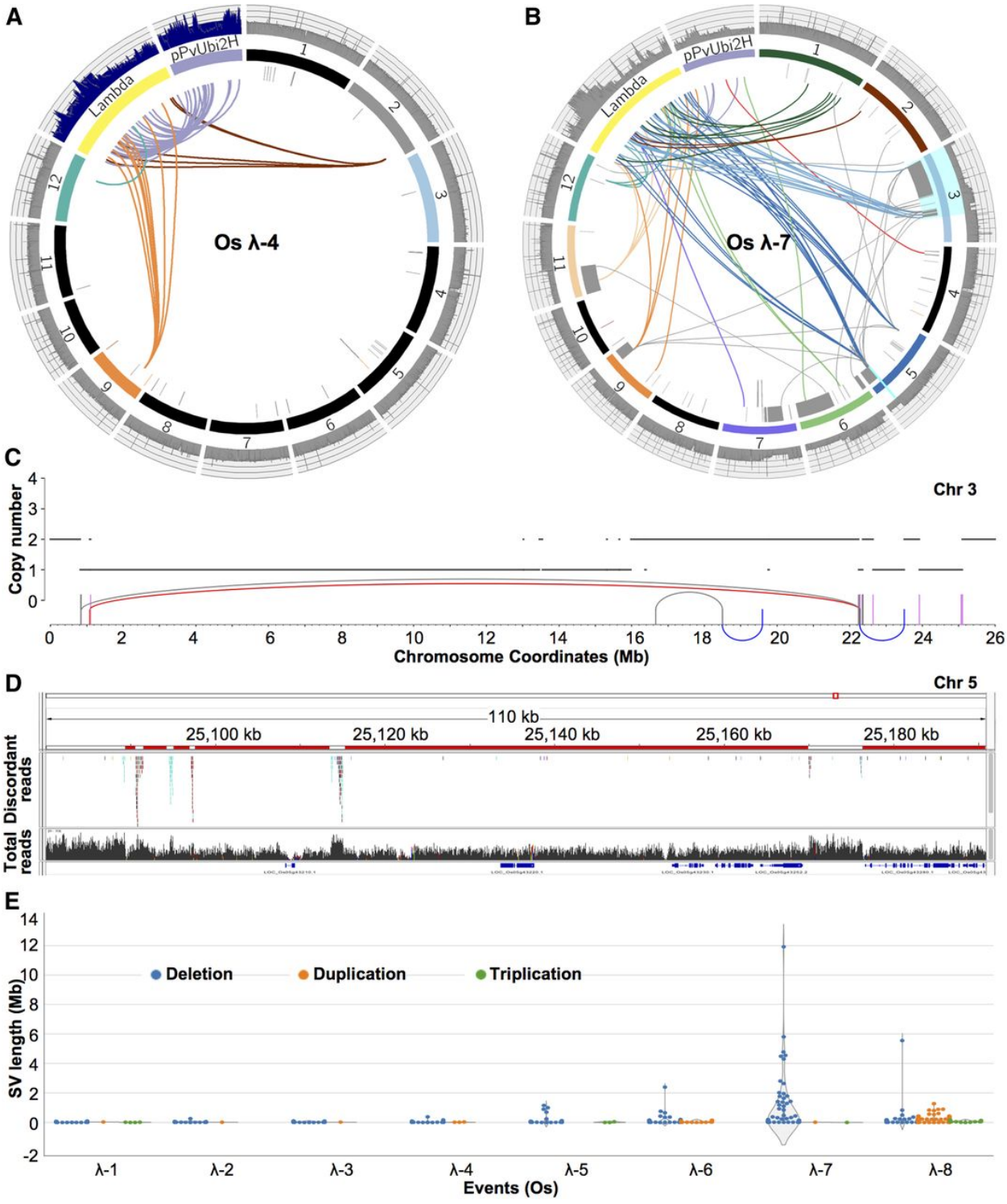


Figure 2.1. Spectrum of genomic outcomes following transformation with lambda and plasmid in rice. All Circos plots are annotated as follows. The twelve rice chromosomes are shown along with λ and plasmid pPvUbi2H magnified at 1,000X and 5,000X. The outer track shows sequence

coverage over each molecule or chromosome as histograms. The inner ring demonstrates DNA copy number profiles derived from read depth, with grey shown as 1 copy, orange as 3 copies, dark red as 4 copies and black as more than 4 copies. The inner arcs designate inter- and intra-chromosomal rearrangements. Breakpoints within the genome are colored grey while the breakpoints between λ or plasmid and the genome are colored to match the respective chromosomes.

A) Rice event λ-4, which contains a long transgene array in chromosome 2. The coverage values in histogram tracks of λ and plasmid are divided by 15 and 1.5 respectively.

B) Rice event λ-7, illustrating a complex event with severe genome damage.

C) A 26 Mb region on chromosome 3 (highlighted in cyan in Figure 2.1B) at high resolution. The horizontal lines show copy number states and vertical bars represent inter-chromosomal breakpoints (grey) and breakpoints involving λ (plum). The arcing links show local rearrangements of the deletion-type (grey), duplication-type (red), and intra-chromosomal translocation-type (blue). For a visual depiction of how local rearrangements are defined using paired end reads, see Figure S2.5.

D) A region from 25.1 Mb to 25.2 Mb on chromosome 5 (highlighted in cyan in Figure 2.1B) as visualized with IGV. Deleted regions are shown in red and retained regions in white (upper), as indicated by the alignment of discordant reads (middle) and read depth (lower).

E) Swarm and violin-plots showing the distribution of the size and number of deletions, duplications and triplications in all rice events transformed with λ. Each dot in the swarm plots represents a different SV. Violin plots represent the statistical distribution, where the width shows the probability of given SV lengths.

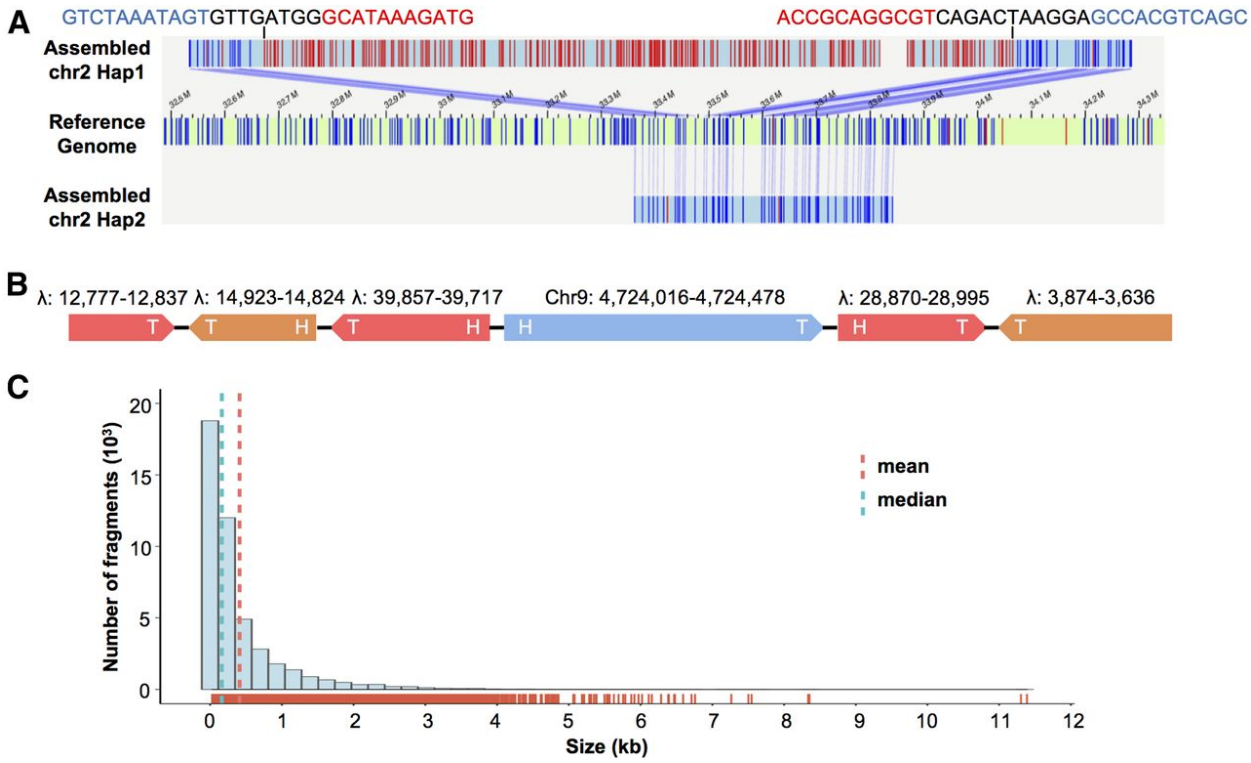


Figure 2.2. Characteristics of the long transgene array in rice event λ-4.

A) Bionano assembly depicting the 1.6 Mb insertion in chromosome 2. The middle panel represents the reference genome, and the upper and lower panels depict the assembled transgenic and wild type chromosomes in this heterozygous line. The blue bars indicate matching restrictions sites between the reference and assembled contigs, and red bars denote restriction sites within the insertion. The nucleotide sequences above the upper panel show the breakpoint sequences, with chromosome sequences highlighted in blue, λ sequences highlighted in red, and novel sequences in black.

B) A 1.1 kb region assembled from Illumina data showing five λ pieces and a single fragment of chromosome 9 in rice event λ-4. The direction of the arrows indicates the 3' ends (Tails) of λ and chromosomal genomic fragments. Four different relative orientations between intra- and inter-chromosomal pieces can be found in this sequence: Tail (3')-Head (5'), Tail-Tail, Head-Tail, Head-Head.

C) Size distribution of λ fragments in the array as determined by PacBio sequencing.

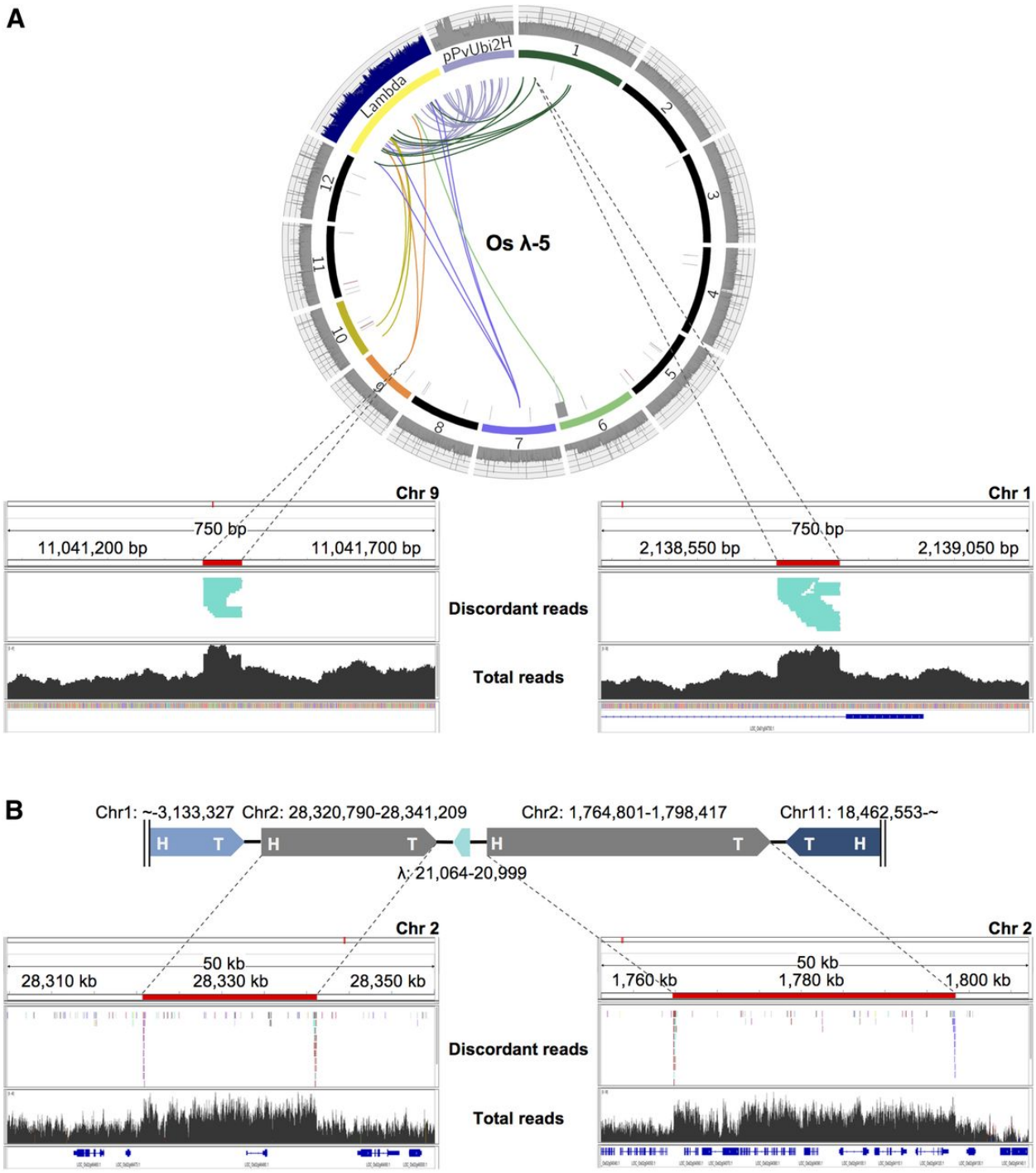


Figure 2.3. Evidence of HDR in rice transgenic events.

A) Circos plot of rice transgenic event λ-5 annotated as in Figure 2.1. The λ coverage is divided by 15. Region 2,138,442 - 2,139,257 on chromosome 1 and region 11,041,419 - 11,041,484 on chromosome 9 are displayed in IGV windows, where displaced fragments (110 bp and 66 bp) are highlighted in red. The upper panels show only discordant reads (where one end maps to the

fragment and the other maps to another chromosome). The lower panels show all reads, illustrating the ~50% increase in read depth indicative of an HDR event.

B) Complex rearrangements observed in rice event λ-8. Regions from chromosome 2 were assembled into an array with other broken fragments at an unknown location in the genome. The damaged regions of chromosome 2 were subsequently repaired as demonstrated by the ~50% increase in read depth.

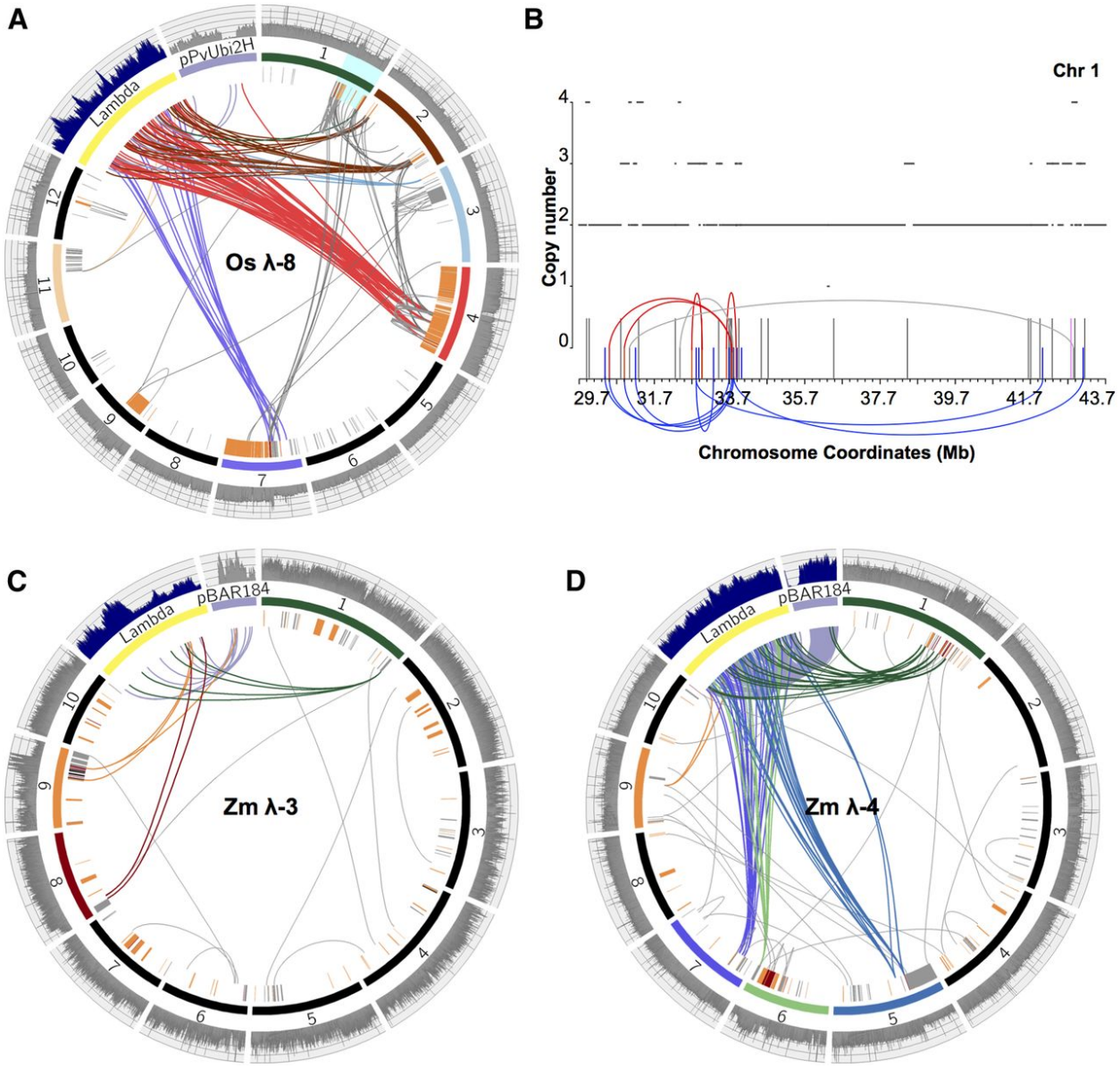


Figure 2.4. Chromothripsis-like outcomes and BFB (breakage-fusion-bridge)-like genomic rearrangements in rice and maize transgenic events.

A) Circos plot of rice transgenic event λ-8 annotated as in Figure 2.1. The coverage of λ in the histogram track is divided by 4.

B) Copy number states of region 29.7 - 43.7 Mb on chromosome 1 (highlighted in cyan in Figure 4A) annotated as in Figure 2.1C.

C) Circos plot of maize transgenic event λ-3, with coverage of λ in the histogram track divided by 5. Note the region of increased copy number states on chromosome 9 indicative of BFB.

D) Circos plot of maize transgenic event λ-4, with the coverage of λ and plasmid in the histogram track divided by 15 and 10 respectively. Note the regions of increased copy number states on chromosome 1 and 6 indicative of BFB.

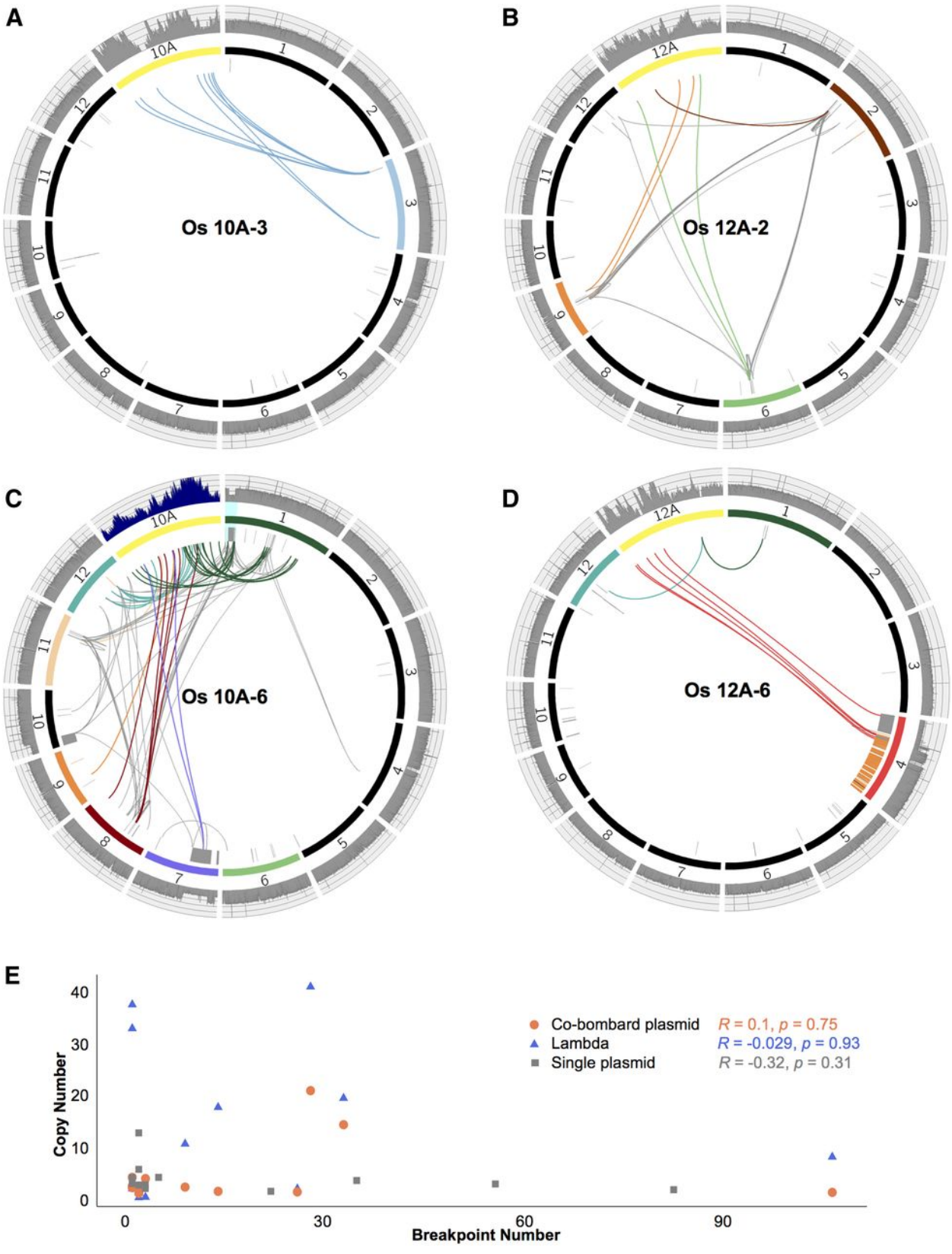

A
Os 10A-3
B
Os 12A-2
C
Os 10A-6
D
Os 12A-6
E
Co-bombard plasmid R = 0.1, p = 0.75
Lambda R = -0.029, p = 0.93
Single plasmid R = -0.32, p = 0.31
Copy Number
Breakpoint Number
0
10
20
30
40
0
30
60
90

Figure 2.5. Similar genomic disturbances following single plasmid transformations. Circos plots of rice lines transformed with plasmid pANIC10A-OsFPGS1 **(A,C)** and pANIC12A-OsFPGS1 **(B,D).**

A) Simple insertion.

B) Complex insertion showing a network of interlinked genomic regions.

C) Extensive damage with a deletion on chromosome 7 and apparent chromothripsis on chromosome 1 (coverage of the 10A plasmid is divided by 2). See Figure S2.6B for a detailed view of the chromothripsis region on chromosome 1 (highlighted in cyan).

D) Chromosome-scale disruption with a partially trisomic chromosome 4.

E) Relationship between transgene copy number and genome breakage at sites not involving the transgene (intra- and inter-chromosomal translocations). Blue triangles and orange circles show lambda and co-bombarded plasmid from the lambda transformation events. Grey squares show data from single plasmid transformations. There are no significant correlations. Pearson correlation coefficient (R) and p-value are indicated.

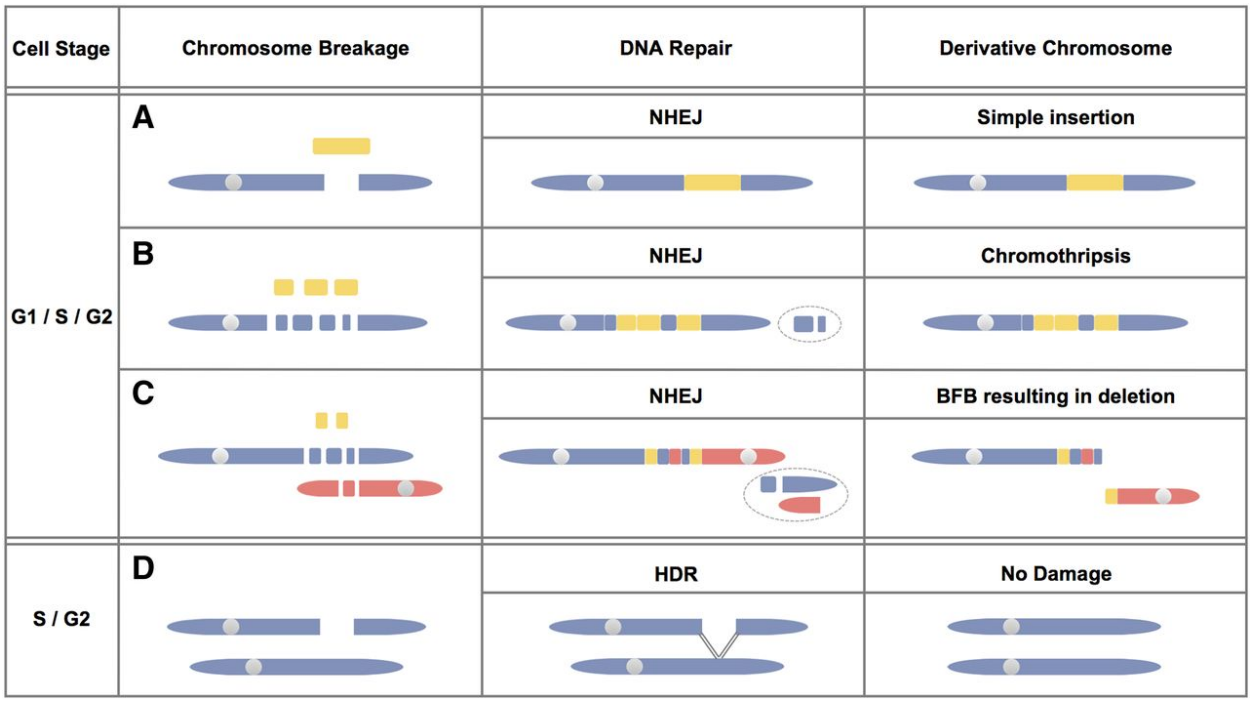


Figure 2.6. Models for genomic outcomes after biolistic transformation. The stage of cell cycle may influence the outcome of biolistic transformation. The models are based on the fact that in animals and presumably plants, non-homologous end joining (NHEJ) is the most likely repair pathway in G1 and homology directed repair (HDR) is more likely in S and G2.

A) Simple insertion. Fragments of introduced molecules (yellow) are ligated with broken ends of native chromosomes by NHEJ (non-homologous end joining).

B) Chromothripsis-like genome rearrangements. Localized regions from native genome are shattered, resulting in many double stranded breaks. Fragments of chromosomes and introduced molecules are stitched together through NHEJ, creating complex patterns that involve the loss of genomic DNA and changes in copy number state (lost regions are circled).

C) Breakage and joining of two different chromosomes and breakage-fusion-bridge (BFB)-like genome rearrangements. When two chromosomes are broken, they can be ligated together through NHEJ. The resulting dicentric chromosome is expected to undergo BFB, which can result in stable terminal deletions.

D) DNA damage repaired by HDR. Double stranded breaks in S or G2 phase may be repaired by HDR through recombination with an intact sister chromatid.

References

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21: 974–984.

Altpeter, F. et al. (2016). Advancing Crop Transformation in the Era of Genome Editing. Plant Cell 28: 1510–1520.

Altpeter, F. et al. (2005). Particle bombardment and the genetic enhancement of crops: myths and realities. Mol. Breed. 15: 305–327.

Anderson, J.E., Michno, J.-M., Kono, T.J.Y., Stec, A.O., Campbell, B.W., Curtin, S.J., and Stupar, R.M. (2016). Genomic variation and DNA repair associated with soybean transgenesis: a comparison to cultivars and mutagenized plants. BMC Biotechnol. 16: 41.

Bankevich, A. et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19: 455–477.

Begemann, M.B., Gray, B.N., January, E., Gordon, G.C., He, Y., Liu, H., Wu, X., Brutnell, T.P., Mockler, T.C., and Oufattole, M. (2017). Precise insertion and guided editing of higher plant genomes using Cpf1 CRISPR nucleases. Sci. Rep. 7: 11606.

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., Patron, N.J., and Nekrasov, V. (2015). Editing plant genomes with CRISPR/Cas9. Curr. Opin. Biotechnol. 32: 76–84.

Campbell, P.J. et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467: 1109–1113.

Casjens, S.R. and Hendrix, R.W. (2015). Bacteriophage lambda: Early pioneer and still relevant. Virology 479-480: 310–330.

Ceccaldi, R., Rondinelli, B., and D'Andrea, A.D. (2016). Repair Pathway Choices and Consequences at the Double-Strand Break. Trends Cell Biol. 26: 52–64.

Chabchoub, E., Rodríguez, L., Galán, E., Mansilla, E., Martínez-Fernandez, M.L., Martínez-Frías, M.L., Fryns, J.-P., and Vermeesch, J.R. (2007). Molecular characterisation of a mosaicism with a complex chromosome rearrangement: evidence for coincident chromosome healing by telomere capture and neo-telomere formation. J. Med. Genet. 44: 250–256.

Chaisson, M.J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC

Bioinformatics 13: 238.

Clarke, J.D. (2009). Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. Cold Spring Harb. Protoc. 2009: db.prot5177.

Clark, K.A. and Krysan, P.J. (2010). Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. Plant J. 64: 990–1001.

Cluster, P.D., O'Dell, M., Metzlaff, M., and Flavell, R.B. (1996). Details of T-DNA structural organization from a transgenic Petunia population exhibiting co-suppression. Plant Mol. Biol. 32: 1197–1203.

Crasta, K., Ganem, N.J., Dagher, R., Lantermann, A.B., Ivanova, E.V., Pan, Y., Nezi, L., Protopopov, A., Chowdhury, D., and Pellman, D. (2012). DNA breaks and chromosome pulverization from errors in mitosis. Nature 482: 53–58.

Ercolano, M.R., Ballvora, A., Paal, J., Steinbiss, H.-H., Salamini, F., and Gebhardt, C. (2004). Functional complementation analysis in potato via biolistic transformation with BAC large DNA fragments. Mol. Breed. 13: 15–22.

Frame, B.R., Zhang, H., Cocciolone, S.M., Sidorenko, L.V., Dietrich, C.R., Pegg, S.E., Zhen, S., Schnable, P.S., and Wang, K. (2000). Production of transgenic maize from bombarded type II callus: Effect of gold particle size and callus morphology on transformation efficiency. In Vitro Cell.Dev.Biol.-Plant 36: 21–29.

Fu X., Duc L.T., Fontana S., Bong B.B., Tinjuangjun P., Sudhakar D., Twyman R.M., Christou P., and Kohli A. (2000). Linear transgene constructs lacking vector backbone sequences generate low-copy-number transgenic plants with simple integration patterns. Transgenic Res. 9:11-9.

Gelvin, S.B. (2017). Integration of Agrobacterium T-DNA into the Plant Genome. Annu. Rev. Genet. 51: 195–217.

Gil-Humanes, J., Wang, Y., Liang, Z., Shan, Q., Ozuna, C.V., Sánchez-León, S., Baltes, N.J., Starker, C., Barro, F., Gao, C., and Others (2017). High-efficiency gene targeting in hexaploid wheat using DNA replicons and CRISPR/Cas9. Plant J. 89: 1251–1262.

Glenn, K.C. et al. (2017). Bringing New Plant Varieties to Market: Plant Breeding and Selection Practices Advance Beneficial Characteristics while Minimizing Unintended Changes. Crop Sci. 57: 2906–2921.

Gorbunova, V. and Levy, A.A. (1997). Non-homologous DNA end joining in plant cells is

associated with deletions and filler DNA insertions. Nucleic Acids Res. 25: 4650–4657.

Healey, A., Furtado, A., Cooper, T., and Henry, R.J. (2014). Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods 10: 21.

He, X., Tan, C., Wang, F., Wang, Y., Zhou, R., Cui, D., You, W., Zhao, H., Ren, J., and Feng, B. (2016). Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. Nucleic Acids Res. 44: e85.

Heyer, W.-D., Ehmsen, K.T., and Liu, J. (2010). Regulation of homologous recombination in eukaryotes. Annu. Rev. Genet. 44: 113–139.

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics 28: 593–594.

Jackson, M.A., Anderson, D.J., and Birch, R.G. (2013). Comparison of Agrobacterium and particle bombardment using whole plasmid or minimal cassette for production of high-expressing, low-copy transgenic plants. Transgenic Res. 22: 143–151.

Jackson, S.A., Zhang, P., Chen, W.P., Phillips, R.L., Friebe, B., Muthukrishnan, S., and Gill, B.S. (2001). High-resolution structural analysis of biolistic transgene integration into the genome of wheat. Theor. Appl. Genet. 103: 56–62.

Jiao, Y. et al. (2017). Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Jones, T.J. and Rost, T.L. (1989). The Developmental Anatomy and Ultrastructure of Somatic Embryos from Rice (Oryza sativa L.) Scutellum Epithelial Cells. Bot. Gaz. 150: 41–49.

Jupe, F., Rivkin, A.C., Michael, T.P., Zander, M., and Motley, T.S. (2018). The complex architecture of plant transgene insertions. bioRxiv.

Karanam, K., Kafri, R., Loewer, A., and Lahav, G. (2012). Quantitative live cell imaging reveals a gradual shift between DNA repair mechanisms and a maximal use of HR in mid S phase. Mol. Cell 47: 320–329.

Kawahara, Y. et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6: 4.

Kaya, H., Sato, S., Tabata, S., Kobayashi, Y., Iwabuchi, M., and Araki, T. (2000). hosoba toge toge, a syndrome caused by a large chromosomal deletion associated with a T-DNA insertion in Arabidopsis. Plant Cell Physiol. 41: 1055–1066.

Kessler, D.A., Taylor, M.R., Maryanski, J.H., Flamm, E.L., and Kahl, L.S. (1992). The safety of foods developed by biotechnology. Science 256: 1747–9, 1832.

Kleinboelting, N., Huep, G., Appelhagen, I., Viehoever, P., Li, Y., and Weisshaar, B. (2015). The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. Mol. Plant 8: 1651–1664.

Klein, T.M., Kornstein, L., Sanford, J.C., and Fromm, M.E. (1989). Genetic transformation of maize cells by particle bombardment. Plant Physiol. 91: 440–444.

Kohli, A., Griffiths, S., Palacios, N., Twyman, R.M., Vain, P., Laurie, D.A., and Christou, P. (1999). Molecular characterization of transforming plasmid rearrangements in transgenic rice reveals a recombination hotspot in the CaMV 35S promoter and confirms the predominance of microhomology mediated recombination. Plant J. 17: 591–601.

Korbel, J.O. and Campbell, P.J. (2013). Criteria for inference of chromothripsis in cancer genomes. Cell 152: 1226–1236.

van Kregten, M., de Pater, S., Romeijn, R., van Schendel, R., Hooykaas, P.J.J., and Tijsterman, M. (2016). T-DNA integration in plants results from polymerase-θ-mediated DNA repair. Nat Plants 2: 16164.

Krizkova, L. and Hrouda, M. (1998). Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions. Plant J. 16: 673–680.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15: R84.

Lee, M. and Phillips, R.L. (1988). The Chromosomal Basis of Somaclonal Variation. Annu. Rev. Plant Physiol. Plant Mol. Biol. 39: 413–437.

Leibowitz, M.L., Zhang, C.-Z., and Pellman, D. (2015). Chromothripsis: A New Mechanism for Rapid Karyotype Evolution. Annu. Rev. Genet. 49: 183–211.

Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., Liu, J., Zhang, H., Liu, C., Ran, Y., and Gao, C. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. Nat. Commun. 8: 14261.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

arXiv [q-bio.GN].

Li, J., Meng, X., Zong, Y., Chen, K., Zhang, H., Liu, J., Li, J., and Gao, C. (2016). Gene replacements and insertions in rice by intron targeting using CRISPR-Cas9. Nat Plants 2: 16139.

Lowe, B.A., Shiva Prakash, N., Way, M., Mann, M.T., Spencer, T.M., and Boddupalli, R.S. (2009). Enhanced single copy integration events in corn via particle bombardment using low quantities of DNA. Transgenic Res. 18: 831–840.

Makarevitch, I., Svitashev, S.K., and Somers, D.A. (2003). Complete sequence analysis of transgene loci from plants transformed via microprojectile bombardment. Plant Mol. Biol. 52: 421–432.

Mann, D.G.J., LaFayette, P.R., Abercrombie, L.L., King, Z.R., Mazarei, M., Halter, M.C., Poovaiah, C.R., Baxter, H., Shen, H., Dixon, R.A., and Others (2012). Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species. Plant Biotechnol. J. 10: 226–236.

Mardin, B.R. et al. (2015). A cell-based model system links chromothripsis with hyperploidy. Mol. Syst. Biol. 11: 828.

McClintock, B. (1942). The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. Proc. Natl. Acad. Sci. U. S. A. 28: 458–463.

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics 26: 234–282.

McVey, M. and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 24: 529–538.

Nacry, P., Camilleri, C., Courtial, B., Caboche, M., and Bouchez, D. (1998). Major chromosomal rearrangements induced by T-DNA transformation in Arabidopsis. Genetics 149: 641–650.

Pannunzio, N.R., Watanabe, G., and Lieber, M.R. (2017). Nonhomologous DNA End Joining for Repair of DNA Double-Strand Breaks. J. Biol. Chem.

Partier, A., Gay, G., Tassy, C., Beckert, M., Feuillet, C., and Barret, P. (2017). Molecular and FISH analyses of a 53-kbp intact DNA fragment inserted by biolistics in wheat (Triticum aestivum L.) genome. Plant Cell Rep. 36: 1547–1559.

Pawlowski, W.P. and Somers, D.A. (1996). Transgene inheritance in plants genetically engineered by microprojectile bombardment. Mol. Biotechnol. 6: 17–30.

Pawlowski, W.P. and Somers, D.A. (1998). Transgenic DNA integrated into the oat genome is frequently interspersed by host DNA. Proc. Natl. Acad. Sci. U. S. A. 95: 12106–12110.

Phan, B.H., Jin, W., Topp, C.N., Zhong, C.X., Jiang, J., Dawe, R.K., and Parrott, W.A. (2007). Transformation of rice with long DNA-segments consisting of random genomic DNA or centromere-specific DNA. Transgenic Res. 16: 341–351.

Raji, J.A., Frame, B., Little, D., Santoso, T.J., and Wang, K. (2018). Agrobacterium- and Biolistic-Mediated Transformation of Maize B104 Inbred. Methods Mol. Biol. 1676: 15–40.

Register, J.C., 3rd, Peterson, D.J., Bell, P.J., Bullock, W.P., Evans, I.J., Frame, B., Greenland, A.J., Higgs, N.S., Jepson, I., and Jiao, S. (1994). Structure and function of selectable and non-selectable transgenes in maize after introduction by particle bombardment. Plant Mol. Biol. 25: 951–961.

Rode, A., Maass, K.K., Willmund, K.V., Lichter, P., and Ernst, A. (2016). Chromothripsis in cancer cells: An update. Int. J. Cancer 138: 2322–2333.

Shi, J., Gao, H., Wang, H., Lafitte, H.R., Archibald, R.L., Yang, M., Hakimi, S.M., Mo, H., and Habben, J.E. (2017). ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. Plant Biotechnol. J. 15: 207–216.

Shou, H., Frame, B.R., Whitham, S.A., and Wang, K. (2004). Assessment of transgenic maize events produced by particle bombardment or Agrobacterium-mediated transformation. Mol. Breed. 13: 201–208.

Stephens, P.J. et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144: 27–40.

Storchová, Z. and Kloosterman, W.P. (2016). The genomic characteristics and cellular origin of chromothripsis. Curr. Opin. Cell Biol. 40: 106–113.

Svitashev, S., Ananiev, E., Pawlowski, W.P., and Somers, D.A. (2000). Association of transgene integration sites with chromosome rearrangements in hexaploid oat. Theor. Appl. Genet. 100: 872–880.

Svitashev, S.K., Pawlowski, W.P., Makarevitch, I., Plank, D.W., and Somers, D.A. (2002). Complex transgene locus structures implicate multiple mechanisms for plant transgene rearrangement. Plant J. 32: 433–445.

Svitashev, S., Young, J.K., Schwartz, C., Gao, H., Falco, S.C., and Cigan, A.M. (2015). Targeted Mutagenesis, Precise Gene Editing, and Site-Specific Gene Insertion in Maize Using Cas9

and Guide RNA. Plant Physiol. 169: 931–945.

Takano, M., Egawa, H., Ikeda, J.E., and Wakasa, K. (1997). The structures of integration sites in transgenic rice. Plant J. 11: 353–361.

Tan, E.H., Henry, I.M., Ravi, M., Bradnam, K.R., Mandakova, T., Marimuthu, M.P., Korf, I., Lysak, M.A., Comai, L., and Chan, S.W. (2015). Catastrophic chromosomal restructuring during genome elimination in plants. Elife 4.

Tassy, C., Partier, A., Beckert, M., Feuillet, C., and Barret, P. (2014). Biolistic transformation of wheat: increased production of plants with simple insertions and heritable transgene expression. Plant Cell, Tissue and Organ Culture. 119: 171–181

Udall, J. and Dawe, R.K. (2017). Is it ordered correctly? Validating genome assemblies by optical mapping. Plant Cell 30: 7-14.

Verma, P. and Greenberg, R.A. (2016). Noncanonical views of homology-directed DNA repair. Genes Dev. 30: 1138–1154.

Weber, N., Halpin, C., Curtis Hannah, L., Jez, J.M., Kough, J., and Parrott, W. (2012). Crop Genome Plasticity and Its Relevance to Food and Feed Safety of Genetically Engineered Breeding Stacks. Plant Physiol.: 112.204271.

Zakov, S., Kinsella, M., and Bafna, V. (2013). Detecting Breakage Fusion Bridge cycles in tumor genomes -- an algorithmic approach. Proc. Nat. Acad. Sci. USA 110: 5546-51

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-né, P., Nicolas, A., Delattre, O., and Barillot, E. (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26: 1895–1896.

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., and Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. Nature 522: 179–184.

Zhang, H., Phan, B.H., Wang, K., Artelt, B.J., Jiang, J., Parrott, W.A., and Dawe, R.K. (2012). Stable integration of an engineered megabase repeat array into the maize genome. Plant J. 70: 357–365.

Zhu, C., Wu, J., and He, C. (2010). Induction of chromosomal inversion by integration of T-DNA in the rice genome. J. Genet. Genomics 37: 189–196.

CHAPTER 3

GAPLESS ASSEMBLY OF MAIZE CHROMOSOMES USING LONG-READ TECHNOLOGIES[2]

[2]Liu, J., Seetharam, A. S., Chougule, K., Ou, S., Swentowsky, K. W., Gent, J. I., ... & Dawe, R. K. (2020). Gapless assembly of maize chromosomes using long-read technologies. *Genome biology*, *21*, 1-17. Reprinted here with permission of the publisher.

Abstract

Creating gapless telomere-to-telomere assemblies of complex genomes is one of the ultimate challenges in genomics. We use two independent assemblies and an optical map-based merging pipeline to produce a maize genome (B73-Ab10) composed of 63 contigs and a contig N50 of 162 Mb. This genome includes gapless assemblies of chromosome 3 (236 Mb) and chromosome 9 (162 Mb), and 53 Mb of the Ab10 meiotic drive haplotype. The data also reveal the internal structure of seven centromeres and five heterochromatic knobs, showing that the major tandem repeat arrays (CentC, knob180 and TR-1) are discontinuous and frequently interspersed with retroelements.

Introduction

Maize is a classic genetic model, known for its excellent chromosome cytology and rich history of transposon research (Nannas and Dawe 2015). Transposons make up the majority of the maize genome (Jiao et al. 2017), and their accumulation over millions of years has driven genes far apart from each other and separated genes from their regulatory sequences (Ricci et al. 2019). There are also large inversions and other structural variations that contribute to fitness (Pyhäjärvi et al. 2013; Yang et al. 2019) and significant variation in genome size caused by tandem repeat arrays (Bilinski et al. 2018). Understanding this remarkable structural diversity is important for the continued improvement of maize, but the high repeat content has impeded progress (Jiao et al. 2017; Yang et al. 2019). Here we describe an automated assembly merging approach that yields gapless maize chromosomes and dramatically improves contiguity throughout the genome, including centromere and knob regions.

The most challenging genomic regions to assemble are tandem repeat arrays that exceed the read length of the current sequencing technologies. In most eukaryotes, these arrays are enriched in centromeres and ribosomal DNA (rDNA). Maize contains a centromeric repeat of 156 bp (Wolfgruber et al. 2009), a 45S rDNA repeat of 9349 bp, and a 5S rDNA repeat of 341 bp. In addition, maize contains two abundant classes of knob repeats that are found on chromosome arms, the major knob180 repeat (180 bp) (Peacock et al. 1981) and the minor TR-1 repeat (~360 bp) (Ananiev, Phillips, and Rines 1998). Knob repeats occur in arrays that extend into the tens of megabases and present a significant barrier to full genome assembly. In most maize lines, knobs appear as inert heterochromatic bulges (Peacock et al. 1981), but in lines with a meiotic drive system on Abnormal chromosome 10 (Ab10) they have centromere-like properties and are preferentially segregated to progeny (Dawe et al. 2018). Ab10 is considerably longer than chromosome 10 and contains two inversions (Mroczek et al. 2006), three knobs, and long spans of uncharacterized DNA that include a cluster of *Kinesin driver* (*Kindr*) genes required for meiotic drive (Ananiev, Phillips, and Rines 1998). Meiotic drive systems have been documented in many organisms and often lie within large inversions that contain novel repeat arrays (Dyer, Charlesworth, and Jaenike 2007), yet no meiotic drive haplotype has been fully sequenced and assembled.

Results and Discussion

A new maize inbred, B73-Ab10, was created by backcrossing a line containing Ab10 to the B73 inbred six times and selfing it an additional five times (BC_6F_5). The B73-Ab10 inbred differs from B73 by the end of chromosome 10L which carries the Ab10 haplotype, the end of chromosome 9S which carries a kernel color gene necessary to score meiotic drive, and a 13 Mb

internal section of of chromosome 6 (coordinates between ~155 Mb-169 Mb). We used DNA from this line to prepare an optical map with the Bionano Saphyr system and sequenced it to high coverage using both PacBio and Nanopore technologies. We then implemented a genome assembly workflow based around the optical map (Figure S3.1). Briefly, the PacBio data were assembled using Canu (Koren et al. 2017), the Nanopore data assembled using miniasm (Heng Li 2016) and the two independent assemblies merged with miniasm and integrated with the optical map as hybrid scaffolds. Hybrid scaffolds were then used to guide further gap closing and create a pseudomolecule assembly (Figure 3.1A). Our approach of one-step contig merging and error correction using optical maps as a reference differs from other methods that rely on local assemblies to fill gaps and correct errors (Du and Liang 2019; Vollger et al. 2019). While PacBio provided an overall superior assembly, it tended to fail in large repetitive regions (Figure S3.2A, B) and heterozygous areas (Figure S3.2C) where the Nanopore assembly succeeded due to a longer read length distribution. This was particularly evident in TR-1, knob180, and subtelomeric arrays as well as other tandemly duplicated regions (Figure S3.2B). Alignments of the optical map to the independent assemblies (Udall and Dawe 2017) and standard genome completeness measures demonstrate that the approach is highly accurate (Table S3.1 and S3.2).

The final assembly has a contig N50 of 162 Mb (Table 3.1), which far exceeds the contiguity of any prior maize genome assembly (Jiao et al. 2017; Yang et al. 2019). Of particular note is the complete 236 Mb assembly of chromosome 3, which was assembled gaplessly without manual intervention – a first for any chromosome from a large complex genome. While the human X-chromosome was also assembled gaplessly (Miga et al. 2019), this outcome required extensive manual inspection and correction. The entire B73-Ab10 genome is represented by 63 contigs where 90% are longer than 20.4 Mb (the N90). In addition to the

expected gaps in repeat arrays, there were two gaps associated with residual heterozygosity on chromosome 9. Regions of heterozygosity reduce effective coverage and lead to assembly chimeras that are broken during hybrid scaffolding. We filled these heterozygosity-associated breaks by choosing the dominant Bionano path and performing local assemblies over the gaps. Nanopore reads were also used to span a gap within a CentC array to complete the chromosome 9 telomere-to-telomere assembly. Aside from these manual interventions, some efforts to manually improve within-knob assemblies, and a correction to the *Kindr* gene complex region of Ab10, the assembly was automated. Our success in assembling chromosomes 3 and 9 can be attributed to the fact that these chromosomes have the fewest cytologically visible repeat arrays (Albert et al. 2010). All remaining gaps in the assembly are marked at the edges by tandem repeats (Figure 3.1, Figure S3.2D).

Seven of the ten functional centromeres as defined by ChIP-seq of CENP-A/CENH3 (Wolfgruber et al. 2009) were assembled without gaps (Table S3.3). Alignment of partial BAC-based assemblies of B73 centromeres showed excellent agreement overall (Figure S3.3). Only a subset of maize centromeres are composed of long CentC arrays, and even within those arrays the majority of reads (65%) can be uniquely mapped, reflecting a high degree of sequence polymorphism (Table S3.3 and S3.4). Three centromeres have no CentC at all and are composed of transposons of different forms. These include known Centromeric Retroelements (CRM) (Wolfgruber et al. 2009) as well as other common retrotransposons. We found no tendency for CENH3 to interact with CentC and CRM over any of the other repeats present (Table S3.4). The lack of sequence specificity can be seen on centromere 3, where CENH3 localized over a 771-kb CentC array as well as a variety of other transposons in flanking sequence (Figure 3.1A, inset).

Prior maize assemblies have succeeded in obtaining only small fragments of knob repeat arrays. In contrast, a knob180-rich knob on chromosome 9 (850 Kb), a TR-1-rich knob on chromosome 4 (1.3 Mb) and three TR-1-rich knobs (4.2 Mb, 2.6 Mb, and 2.1 Mb) on Ab10 were fully assembled in the B73-Ab10 assembly. The data show that knobs, like centromeres (Wolfgruber et al. 2009; Jiao et al. 2017), often contain more transposons than tandem repeats (Figure 3.1C). Centromeric Retrotransposons target areas with CENP-A/CENH3 (Wolfgruber et al. 2009; Jiao et al. 2017) and occupy on average 31.9% of functional centromeres, including within CentC arrays (Figure 3.1A and Table S3.3 and S3.6). The new knob assemblies reveal that the Cinful-Zeon family of *Gypsy* elements (Sanz-Alferez et al. 2003) preferentially target knobs. Cinful-Zeon elements occupy 27.0% of the assembled TR-1-rich knobs and 8.2% of the knob180-rich knobs, but only 3.8% percent of CentC arrays (Figure 3.1A and Table S3.6 and S3.7). Cinful-Zeon elements are also abundant in other heterochromatic regions throughout the genome (Figure 3.1A).

In addition to revealing the internal structure of knobs, the data provide the first complete view of the Ab10 haplotype that provides the selective force for the accumulation and maintenance of knobs (Dawe et al. 2018). The meiotic drive haplotype on Ab10 contains three fully assembled TR-1 knobs, a much larger knob180 knob that was not assembled, and two large inversions (4.4 and 8.3 Mb) that are homologous to normal chromosome 10 (Figure 3.1C). These major structural differences help to explain why recombination between the Ab10 haplotype and normal chromosome 10 is suppressed (Rhoades 1942). Ab10 also contains 22.4 Mb of novel sequence with no synteny to other regions of the maize genome or related grass genomes. Within this domain is the complete cluster of nine *Kindr* genes that are integral components of the drive system (Dawe et al. 2018), as well as hundreds of other expressed genes, many of which have

only one exon or overlap with transposons and are likely non-functional (Table S3.8). Additional meiotic drive functions associated with the movement of knobs at meiosis and their delivery to egg cells (Hiatt and Dawe 2003) remain to be identified in this newly discovered sequence.

Conclusions

Gapless genome assemblies remove all uncertainty about the order, spacing and orientation of genes and their regulators. We have shown that this can be achieved using long reads and well-known assembly algorithms, with significant improvements in contiguity obtained by integrating independent assemblies around an optical map scaffold. Given that most contigs end in telomeres, centromeres or knobs, we presume that virtually all of the genes and associated regulatory information are represented in this genome assembly. The assembly merging pipeline also revealed the internal structure of repetitive domains that were previously known only by cytological techniques, thereby opening these regions to annotation and future epigenomic profiling. Similar results should be achievable for other complex genomes, although higher sequence coverage, longer reads, and/or additional scaffolding information may be needed for species with polyploidy or higher levels of heterozygosity.

Methods

PacBio assembly

High molecular weight DNA was extracted from young leaves using the protocol of Doyle and Doyle (Doyle and Doyle 1987) with minor modifications. Young maize leaves flash frozen at -80°C were ground to a fine powder in liquid N2 followed by very gentle extraction in CTAB buffer (that included proteinase K, PVP-40 and beta-mercaptoethanol) for 1 hr at 50C.

After centrifugation, the supernatant was gently extracted twice with 24:1 chloroform:iso-amyl alcohol. The upper phase was adjusted to 1/10th volume with 3M KAc, gently mixed, and DNA precipitated with iso-propanol. DNA was collected by centrifugation, washed with 70% EtOH, air dried for 20 min and dissolved thoroughly in 1x TE at room temperature.

Sequencing libraries were constructed following PacBio's template prep protocols (Procedure & Checklist – Preparing gDNA Libraries Using the SMRTbell Express Template Preparation Kit 2.0, PN 101-693-800 Version 01) for the Express Template Prep Kit 2.0 (Cat# 100-939-900) and sequenced using Sequel SMRTLink V5.1 and Sequel binding and sequencing chemistry v2.1. The longest 50X out of 62X PacBio raw sequences were error-corrected using falcon_kit pipeline v0.7 (Chin et al. 2016) without repeat masking by TANmask and REPmask (-e 0.75 -l 3000 --min_cov 2 --max_n_read 200). The error-corrected reads (43X, N50= 22.3 Kb) were then trimmed and assembled with Canu (Koren et al. 2017) (v1.8) with the following parameters: correctedErrorRate=0.065 corMhapSensitivity=normal ovlMerThreshold=500 utgOvlMerThreshold=150. The read error correction process that is necessary for PacBio assembly may have homogenized some repeats and limited the assembly in long repeat regions. The accuracy of the Canu-generated contigs was increased by aligning the raw PacBio reads to the assembly using pbmm2 (v1.2.0) from pb-assembly (Chin et al. 2016) and running the PacBio consensus algorithm tool Arrow (v2.3.3) (https://github.com/PacificBiosciences/GenomicConsensus) with default parameters to generate sequenced polished contigs. The contig assembly was further polished using 73X PE150 Illumina sequence by first aligning the reads to the Arrow polished assembly using minimap2 (Heng Li 2018), followed by running the assembly tool Pilon (Walker et al. 2014) (v1.22) to

correct individual base errors and small indels using the following parameters: --fix bases --minmq 30.

Nanopore assembly

Two different DNA extraction methods were used to generate high molecular weight (HMW) DNA for Oxford Nanopore (ONT) sequencing. CTAB DNA was prepared as described above for the PacBio assembly. Nuclear DNA was prepared using the protocol of Luo and Wing (Luo and Wing 2003) with minor modifications. Young leaves flash frozen at -80°C were ground with liquid nitrogen and incubated with NIB buffer (10 mM Tris-HCL, PH8.0, 10mM EDTA PH8.0, 100mM KCL, 0.5 M sucrose, 4 mM spermidine, 1 mM spermine) on ice for 15 min. After filtration through miracloth, Triton X-100 (Sigma) was added to tubes at a 1:20 ratio, placed on ice for 15 minutes, and centrifuged to collect nuclei. Nuclei were washed with NIB buffer (containing Triton X-100) and re-suspended in 40 ml of the same buffer and centrifuged again. After removal of all liquid, 10 ml of Qiagen G2 buffer was added followed by gentle resuspension of nuclei; then 30 ml G2 buffer with RNase A (to a final concentration of 50 mg/ml) was added. Tubes were incubated at 37°C for 30 min. Proteinase K (Invitrogen), 30 mg, was added and incubated at 50 C for 2 hr followed by centrifugation for 15 min at 8000 rpm, at 4°C, and the liquid gently poured into a new tube. After gentle extraction with Chloroform:isoamyl alcohol (24:1), DNA was precipitated with two thirds volume iso-propanol. The DNA pellet was washed with 70% EtOH, air dried for 20 min and dissolved in TE at room temperature.

DNA from both the CTAB and nuclear prep was used to generate either a rapid (SQK-RAD004) or one dimensional (1d; SQK-LSK109) sequencing library for ONT. The resulting

libraries were run on either a MinION or GridION sequencer running for 48 hrs. All bases were called on the GridION using Guppy (v2.1.3), and the resulting fastq files were used for genome assembly. A total of 121 Gb (~50x) of ONT sequence was generated over 27 MinION R9.4 flowcells. The data were filtered for reads >10 Kb using seqtk (https://github.com/lh3/seqtk), resulting in an estimated 30x coverage (N50=29,311 bp) of the maize genome. The resulting uncorrected reads were aligned (overlap) with minimap2 (v2.13;-x ava-ont -t 64) (Heng Li 2018) and an assembly graph (layout) was generated with miniasm (v0.3; -f <reads> <overlaps>) (Heng Li 2016). The resulting graph was inspected using Bandage (Wick et al. 2015). The fact that the Nanopore assembly was carried out with uncorrected reads may have contributed to its better performance in long repeat regions (Fig. S2). A consensus genome assembly was generated by mapping reads >10 Kb to the assembly with minimap2, and then running racon (v1.3.1) (Vaser et al. 2017); the consensus process was repeated three times. The contig assembly was further polished using 73X PE150 Illumina sequence by first aligning the reads to the consensus assembly using minimap2 (Heng Li 2016) followed by running the assembly tool pilon (v1.18) (Walker et al. 2014) two times using 73X PE150 Illumina sequence.

Optical map assembly

Ultra high molecular weight DNA was isolated from maize seedlings using a modified version of the Bionano Genomics Plant Tissue DNA Isolation Base protocol. Approximately 0.5 g of healthy aerial tissue was collected from young B73-Ab10 etiolated seedlings grown in soil-free conditions for 2 weeks. The leaves were treated with a 2% formaldehyde Bionano fixing solution, washed, chopped and homogenized using a Qiagen TissueRuptor in homogenization buffer. Free nuclei were pelleted at 2,000X g, washed, isolated by gradient centrifugation, and

embedded in a low melting point agarose plug. The nuclei were lysed by treating with proteinase K and RNase A treatments as described previously (Deschamps et al. 2018), and washed four times in Wash Buffer and five times in TE buffer. The purified high molecular weight nuclear DNA was recovered by melting the plug, digesting it with agarase and subjecting the resulting sample to drop dialysis against TE.

The Bionano Saphyr platform was used in combination with the Direct Label and Stain (DLS) process to generate chromosome-level sequence scaffolds (Belser et al. 2018). Direct labeling was performed using the Direct Labeling and Staining Kit (Bionano Genomics, San Diego CA) according to the manufacturer's protocol, except that one microgram of DNA was used and DNA Stain was added to a final concentration of 1 microliter per 0.1 microgram of final DNA. The labeled sample was loaded into a Saphyr chip and molecules separated, imaged and digitized using a Saphyr and Compute server. Data visualization, map assembly and hybrid scaffold construction were performed using Bionano Access (v1.3) and Bionano Solve (v3.4.0). A subset of 1,580,077 molecules with a minimum size of 150 Kb and combined length of 424,488 Mb were assembled without pre-assembly using the non-haplotype, no-CMPR-cut parameters without extend-split.

Assembly merging and gap closing

We developed a pipeline to integrate independent contig assemblies and curate assembly errors using Bionano maps as an anchor. The pipeline consists of five steps: 1) conflict resolution, 2) assembly error curation, 3) contig merging, 4) hybrid assembly and contig overlap removal, and 5) manual curation and gap filling (Figure S3.1). The first four steps were automated. A gapless chromosome 3 was generated upon contig merging in the third step, and

the complete assembly of chromosome 9 required manual curation. While contig merging with miniasm can be applied to any two sequence assemblies, the availability of *de novo* assembled Bionano maps is necessary to perform conflict-cutting in step 1, contig error correction in step 2, and hybrid scaffolding in step 4 of the pipeline.

Step 1: Conflicts between the optical map and DNA sequence assemblies were resolved using Bionano Solve software (https://bionanogenomics.com/support-page/data-analysis-documentation/). Sequence assembly can occasionally connect two regions that share a repetitive sequence but do not belong together (making a chimeric contig). These appear as conflicts between bionano maps and sequence assemblies when they are aligned. Optical maps were aligned to *in silico* digested representations of the DNA sequence assemblies using RefAligner (v3.4.0) and conflicts identified with the AssignAlignType.pl script. Conflicts with chimeric a quality score higher than the default threshold were split using cut_conflicts.pl (using default parameters from optArguments_nonhaplotype_noES_DLE1_saphyr.xml) and a sequence file was produced with custom script cut_conflict_NGS.py. Removing chimeric joins increases the chance of complementary contig merging in Step 3.

Step 2: Assembly errors in the conflict-resolved PacBio contigs were identified and automatically curated with ONT contigs. In this step, PacBio and ONT contigs were aligned to rescaled optical maps and structural discrepancies detected using the structural variant calling pipeline from BionanoSolve (v3.4.0). Homozygous insertions and deletions with a confidence of at least 0.1 and size larger than 1 Kb were classified as true assembly errors in the PacBio contigs. On the condition that no structural discrepancies were found in the corresponding ONT contigs, the ONT contigs were used to replace the erroneous sequences in PacBio contigs using custom script SV_fix.py.

Step 3: ONT contigs were used to close gaps and improve contiguity of the PacBio contig assembly. ONT contigs were mapped to PacBio contigs with minimap2 (Heng Li 2018) (v2.13; -k28 -w28 -A1 -B9 -O16,41 -E2,1 -z200 -g100000 -r100000 --max-chain-skip 100), and overlap regions merged using miniasm (Heng Li 2016) (v0.3; -1 -2 -r0 -e1 -n1 -h250000 -g100000 -o25000). This step creates PacBio/ONT hybrid contigs that are called unitigs. The unitigs were then combined with the remaining contigs from the PacBio backbone assembly to create a merged contig assembly. After this step a gapless chromosome 3 was generated (a region of heterozygosity from 164.5 to 166.2 Mb on chromosome 3 was automatically resolved). The merged contigs were then aligned to Bionano maps, where overlaps between adjacent contigs were detected and merged with minimap2 (v2.13) and miniasm (v0.3) using the custom script Overlap_merge.py. This step only identifies large overlaps (roughly >200 Kb) that can be detected at the level of *de novo* Bionano label alignment. Identifying all overlaps, including smaller overlaps, requires hybrid scaffolding with the optical map (Step 4). If proceeding to Step 4, overlap merging in Step 3 is optional.

Step 4: Bionano maps were integrated with the sequence contigs by hybrid scaffolding using the hybridScaffold.pl script from BionanoSolve (v3.4.0) with default parameters from optArguments_nonhaplotype_noES_DLE1_saphyr.xml. This step orders and orients sequence contigs and facilitates the resolution of remaining overlaps between contigs. As the optical maps are aligned and rescaled with the sequence maps repeatedly during hybrid scaffolding, more accurate overlaps between contigs are identified and annotated as 13N gaps. These overlaps were removed through contig merging with miniasm (v0.3), as described in Step 3. Due to the extreme repetitiveness in the 45S rDNA repeat region on chromosome 6, both the contig assemblies and

hybrid scaffolding in this area are erroneous. Therefore, we left the contigs in the NOR un-merged and marked the incorrectness with 13N gaps.

Step 5: Manual curation was performed to correct assembly errors, close gaps in repetitive and heterozygous regions, and assemble telomeres.

Repeat assembly manual curation. In highly repetitive regions, erroneous read joins at the tips of contigs were not detected as conflicts or assembly errors in Steps 1 or 2 due to the limited resolution of Bionano alignment. In these regions, we trimmed and removed the unaligned regions to reveal eligible ends for overlap merging using miniasm (v0.3). These modifications extended the contiguity of repeat arrays at the edges of longer contigs. Contigs composed exclusively of knob and CentC repeats arrays lack pan-genome anchor markers and are not present in the pseudomolecules.

Chromosome 9 manual curation. Seven gaps, ranging from 2 Kb to 236 Kb, were present in the chromosome 9 assembly after hybrid scaffolding. Two large gaps of 236 Kb and 41 Kb were caused by heterozygosity (76.29-76.80 Mb), one 21 Kb gap was due to repetitiveness in a CentC array (58.43-58.67 Mb), and the remaining four gaps were smaller than 7 Kb (two of these were in the 843 Kb knob on the tip of 9S). The four small gaps were first filled by running three iterations of LR Gapcloser (Sep 24, 2018 commit) (Xu et al. 2019) at default settings using PacBio error-corrected reads. To resolve the 236 Kb gap caused by heterozygosity, all contigs anchored to chromosome 9 were re-scaffolded using the longest chromosome 9 Bionano map as the sole anchor. This reduced the 236 Kb gap to 58 Kb. Local assemblies were run with Flye (v2.6) (Kolmogorov et al. 2019) using ONT reads surrounding gaps to fill the remaining 58 Kb and 41 Kb gaps. Flye-assembled contigs were integrated with the flanking contigs by unitigging with miniasm (v0.3), and aligned to Bionano maps for inspection. An 8 Kb gap remained, which

was filled with a single ONT read that spans it. The gap in the CentC array was filled by manually selecting two long ONT reads (>50 Kb) that spanned the gap, creating a consensus at the overlap and placing the resulting sequence in the gap.

Kindr complex manual curation. The assembly over the ~1 Mb tandem array of *Kindr* genes (each within an ~100 Kb repeat) was erroneous due to collapsing in the PacBio sequence contig and improper scaffolding. We manually selected the most contiguous ONT contig over this region, carried out hybrid scaffolding for the scaffold containing *Kindr*, placed an excluded contig in the correct area, and removed an overlap region through contig merging.

Telomere manual curation. Fifteen telomeres were assembled by extending the ends of scaffolds with the longest uniquely mapped ONT read that contained telomeric repeats TTTAGGG/CCCTAAA (>=1 Kb). The regions with newly assembled telomeres include 1L, 2L, 3S, 3L, 4S, 4L, 5L, 6L, 7S, 7L, 8S, 8L, 9S, 9L,10S.

The final scaffolds were polished with PacBio subreads using tools from pb-assembly (Chin et al. 2016). Read alignment was performed with pbmm2 (v1.2.0) and polishing was executed with GCpp (v1.0.0) at default parameters. Scaffolds were further polished with 73X PE150 Illumina reads using Pilon (v1.23) with default parameters (Walker et al. 2014). The error-corrected PacBio reads and Illumina reads often mapped incorrectly in highly repetitive regions (Figure S3.2B,C,D). Regions with excessive incorrect mapping are expected to be overpolished, whereas regions with few correctly mapped reads are expected to retain a higher frequency of sequencing errors.

AGP construction

The pseudomolecules were constructed from the hybrid scaffolds using ALLMAPS (v0.8.12) (Tang et al. 2015). Both pan-genome anchor markers (Lu et al. 2015) and the IBM (Intermated B73 x Mo17) genetic map (Lee et al. 2002) were used with equal weights for ordering and orienting the scaffolds. Pan-genome anchor markers were obtained from the CyVerse Data commons ("CyVerse Data Commons" n.d.) and processed to generate a bed file with 50bp upstream and downstream of B73 V3 coordinates. The extracted markers were mapped to a HiSat2 (v2.1.0)[29,30] indexed assembly of B73-Ab10 by disabling splicing (--no-spliced-alignment) and forcing global alignment (--end-to-end). Very high read and reference gap open and extension penalties (--rdg 10000,10000 and --rfg 10000,10000) were also used to ensure full-length mapping of marker sequence. The final alignment was then filtered for mapping quality greater than 30 and tag XM:0 (unique mapping) to retain only high-quality, uniquely mapped marker sequences. The mapped markers were merged with the predicted distance information to generate a CSV input file for ALLMAPS. Only scaffolds with more than 20 uniquely mapped markers, with a maximum of 100 markers per scaffold, were used for pseudomolecule construction. The IBM genetic markers were downloaded from MaizeGDB (https://www.maizegdb.org/complete_map?id=887740) (Portwood et al. 2019) and were processed to generate a bed file similar to pan-genome markers. For the markers with coordinates, 50 bp flanking regions were extracted from the B73 v4 genome. For markers without coordinates, marker sequences were used as-is, and those missing both coordinates and sequences were discarded. Mapping of the markers was done similar to the method described above for the pan-genome anchor markers, with all uniquely mapped markers retained. The genetic distance information for these markers was converted to a CSV file before use in

ALLMAPS. ALLMAPS was run with default options, and the pseudomolecules were finalized after inspecting the marker placement plot and the scaffold directions. Of the 50 Bionano scaffolds anchored with sequence contigs, 26 with uniquely mapped genetic markers were included in the pseudomolecules. Among the 24 unplaced scaffolds with a total size of 19.4 Mb, 22 are composed entirely of knob180 and/or TR-1 arrays (17.7 Mb).

Comparing PacBio and Nanopore assemblies in repetitive and heterozygous regions

To determine how tandem repeats and regions of heterozygosity impacted the assemblies, we identified tandemly repeated areas by chromosome self alignment with minimap2 (v2.17; -PD -k19 -w19 -m200) and heterozygous regions by manual inspection using Bionano Access software. PacBio gap coordinates were projected onto the final assembly using minimap2 (v2.17; -cx asm5 --cs), followed by coordinate liftover using paftools.js (Heng Li 2018). Gaps that were complemented by Nanopore contigs were identified as gaps present in the PacBio assembly but absent in the final assembly. The PacBio adjusted gap coordinates, complemented gaps, and final assembly gaps were mapped to tandem repeats and heterozygous regions with bedtools (Quinlan 2014) (v2.28.0; window -r 500000 -l 500000). The co-occurence of PacBio gaps with tandem repetitiveness and heterozygous regions was assessed by two-tailed Fisher's exact test using bedtools fisher (v2.28.0) at default settings.

To assess read coverage over gap areas, a total of 36.9X error-corrected PacBio reads (>=10 Kb), 20.7X error-corrected Nanopore reads (>=10 Kb), and 30X PE150 Illumina reads were mapped to the final assembly. Long-read mapping was performed using minimap2 (v2.17) with default parameters and short-read mapping was carried out with bwa (v0.7.17) at default settings. Read gap regions were defined as areas mapped with fewer than 3 reads for PacBio and

Illumina datasets, and fewer than 2 reads for the Oxford Nanopore dataset. Basepair level genome coverage was calculated with bedtools genomecov (v2.28.0; -bga) and regions with fewer reads than the cutoff were extracted. The length distributions of PacBio and Oxford Nanopore reads mapped to a tandem repeat (chr8: 31-33.5 Mb) and heterozygous area (chr3: 164-167.6 Mb) were obtained with SAMTools (v1.9).

RNA-seq

Ten tissues were sampled throughout development for evidence-based gene annotation including: primary root (1) and coleoptile (2) at six days after planting; base of the 10th leaf (3), middle of the 10th leaf (4), tip of the 10th leaf (5) at the Vegetative 11 (V11) growth stage; meiotic tassel (6) and immature ear (7) at the V18 growth stage; anthers at the Reproductive 1 (R1) growth stage; endosperm (9) and embryo (10) at 16 days after pollination. For each tissue, two biological replicates were harvested and each biological replicate was made up of tissue from three individual plants. Endosperm and embryo tissues were harvested from 50 kernels per plant (150 total per biological replicate). Tissues 1-5 above were collected from greenhouse-grown plants and tissues 6-10 were from field-grown plants. Greenhouse-grown plants were planted in Metro-Mix300 (Sun Gro Horticulture) with no additional fertilizer and grown under greenhouse conditions (27°C/24°C day/night and 16h/8h light/dark) at the University of Minnesota Plant Growth Facilities. Field grown plants were planted at the Minnesota Agricultural Experiment Station located in Saint Paul, MN with 30-inch row spacing at ~52,000 plants per hectare. RNA was extracted using the Qiagen RNeasy plant mini kit following the manufacturer's suggested protocol.

Total RNA samples were assayed by Bioanalyzer to determine RNA integrity and normalized in 25uL of nuclease-free water prior to library preparation. Sequencing libraries were prepared using KAPA's Stranded mRNA-seq kit (#KK4821) according to the manufacturer's instructions. The mRNA was enriched using oligo-dT beads, fragmented, and converted to double stranded cDNA using random hexamer priming and amplification. Libraries were pooled at equimolar ratios and sequenced on NextSeq 500 instruments using the PE75 protocol.

Gene annotation

For evidence-based predictions, genome-guided transcript assemblies were generated from five different assemblers *viz.,* Trinity (v2.6.6) (Altschul et al. 1990; Grabherr et al. 2011), StringTie (v1.3.4a) (Pertea et al. 2015), Strawberry (v1.1.1) (Liu and Dickerson 2017), Cufflinks (v2.2.1) (Pertea et al. 2015; Trapnell et al. 2012) and Class2 (Pertea et al. 2015; Trapnell et al. 2012; Song, Sabunciyan, and Florea 2016), and the best set of transcripts were identified and annotated as genes using Mikado (v1.2.4) (Venturini et al. 2018). Briefly, the RNA-seq reads from each library were mapped to a STAR (v2.5.3a) (Dobin et al. 2013) indexed B73-Ab10 genome using a 2-pass mapping approach (the initial round of alignments provides splice information for the subsequent round of mapping reads). Default options were used for mapping with few post-processing options enabled (print all SAM format attributes --outSAMattributes All; downstream compatibility --outSAMmapqUnique 10; and number of mis-matches --outFilterMismatchNmax 0). Individually mapped RNA-seq libraries were then pooled, sorted and indexed using SAMTools (v1.9) (H. Li et al. 2009), for use with the transcript assembly programs. For all genome-guided transcriptome assemblers, default options were used except, if it allowed minimum transcript length setting, it was set to 100 bp (Trinity using --

min_contig_length 100, StringTie using -m 100 and Strawberry using -t 100), and if it allowed RNAseq strandedness, it was set to stranded (Trinity using -SS_lib_type FR, Cufflinks using --library-type fr-firststrand). For Trinity, maximum intron size was also set to 10000 (--genome_guided_max_intron 10000). All assemblers generated a GFF3 as the final output except for Trinity, for which assembled transcripts in fasta format were mapped back to the gmap (v2019-05-12) indexed genome to generate a GFF3 file (by setting the output format option -f to gff3_match_cdna). Portcullis (v1.1.2) (Mapleson et al. 2018) was used to generate a high confidence set of splice junctions for the B73-Ab10 genome from the merged mapped reads. Mikado was configured to use all transcript assemblies (with strandedness marked as True for all except for Trinity, and with equal weights), portcullis generated splice sites and a plants.yaml scoring matrix. Preliminary transcripts prepared by Mikado, through merging all transcripts and removing the redundant copies, were processed using TransDecoder (v5.5.0) (Haas et al. 2013) (to identify open reading frames) and blastx (v2.9.0)(Altschul et al. 1990) against SwissProt viridiplantae proteins (for identifying full-length transcripts). Default options were used for TransDecoder, and for blastx, maximum target sequences were set to 5 (-max_target_seqs 5) and output format to xml (-outfmt 5). These were provided as input for Mikado for picking and annotating the best transcripts for each locus. The obtained GFF3 file was used to extract transcripts and proteins using the gffread utility from the Cufflinks package.

Additional structural improvements for the Mikado generated transcripts were completed using the PASA (v2.3.3) (Haas et al. 2003) genome annotation tool. The inputs for PASA included 2,019,896 maize EST derived from genbank, 83,087 Mikado transcripts, 69,163 B73 full length cDNA from genbank and 46,311 maize iso-seq transcripts from 11 developmental tissues that were filtered for intron retention (Wang et al. 2018). PASA was run with default

options, with a first step of aligning transcript evidence to the masked B73-Ab10 genome using GMAP (v.2018-07-04) (Wu and Watanabe 2005) and Blat (v.36) (Kent 2002). The full length cDNA and Iso-seq transcript ID's were passed in a text file (-f FL.acc.list) during the PASA alignment step. Valid near perfect alignments with 95% identity were clustered based on genome mapping location and assembled into gene structures that included the maximal number of compatible transcript alignments. PASA assemblies were then compared with B73-Ab10 Mikado transcript models using default parameters. PASA updated the models, providing UTR extensions, novel and additional alternative isoforms. PASA generated models were passed through the MAKER-P (v3.0) (Campbell et al. 2014) annotation pipeline as model_gff along with all the transcript and protein sequences to obtain Annotation Edit Distance (AED) (Eilbeck et al. 2009) scores to assess the quality of annotations. Transposon element (TE) related genes were filtered using the TEsorter tool (Altschul et al. 1990; Zhang et al., n.d.), which uses the REXdb (viridiplantae_v3.0 + metazoa_v3) database of TEs. Finally the gene annotations were verified for translation errors using the EnsemblCompara pipeline (Vilella et al. 2009).

BUSCO assessment

The gene space completeness of the B73-Ab10 genome assembly was assessed using the GenomeQC (Manchanda et al. 2019) tool, which provides a summary of the number of complete, fragmented and missing Benchmarking Universal Single-Copy Orthologs (BUSCO) in the assembly. The Embryophyta database (embryophyta_odb9; consisting of 1440 conserved, single-copy plant genes) and the genome assembly in the fasta file format were provided as input to the tool to calculate the BUSCO metrics.

TE annotation

The manually curated transposable element library (maizeTE11222019) derived from the Maize TE Consortium (MTEC; https://github.com/oushujun/MTEC) was used as the base TE library. Novel TEs of the maize Ab10 genome not included in the MTEC library were structurally identified using the EDTA pipeline (v1.6.5) (Ou et al. 2019) with parameters "-species maize -curatedlib maizeTE11222019". The MTEC library augmented with Ab10 specific TEs was used to annotate TE fragments using RepeatMasker. Coding sequences of the maize B73 v4 assembly were downloaded from MaizeGDB and used to remove gene sequences in the EDTA-generated TE library. Whole-genome TE annotations were generated using the EDTA augmented MTEC library (-anno 1). The LTR Assembly Index (LAI) (Ou, Chen, and Jiang 2018) scores of genome assemblies were calculated using LAI (beta3.2) within the LTR_retriever (v2.8) (Ou and Jiang 2018) package with parameters "-iden 94.8550 -totLTR 76.34".

Centromere and repeat analyses

The overall accuracy of the centromere assemblies was assessed by aligning previous BAC-based B73 centromere assemblies ("CyVerse Data Commons" n.d.) to the B73-Ab10 genome using Bionano RefAligner (v3.4.0) with default parameters. Although the BAC-based assemblies do not traverse CentC arrays, there is excellent overall agreement in sequence and contiguity (Figure S3.3).

Active centromere locations were determined by identifying the CENH3 ChIP-seq enriched regions in the final assembly using genomic reads as a control. The SE150 Illumina ChIP-seq reads were obtained from SRA (SRX2737618) (Gent, Wang, and Dawe 2017) and the

73X PE150 Illumina genomic reads were subsampled to 30X with seqtk (https://github.com/lh3/seqtk). Both the ChIP-seq reads and the genomic reads were trimmed with Trim Glore (v0.4.5; https://github.com/FelixKrueger/TrimGalore/) with default parameters and aligned to the final assembly with BWA-MEM (v0.7.17) (Heng Li 2013). Epic2 (Stovner and Sætrom 2019) was employed to call peaks with the CENH3 ChIP-seq alignment set as treatment, genomic read alignment as control, MAPQ (mapping quality) as 20, effective genome size as 0.8, bin size as 5000 and gap size as 0. The effective genome size of the final genome was calculated as the fraction of unique 150-mers over total 150-mers using Jellyfish (v2.26) (Marçais and Kingsford 2011) (-m 150 -s 2193M -out-counter-len 1 -counter-len 1). The coordinates of active centromeres were identified as islands with a score above 250 and a fold change higher than 4.

The coordinates of repeat arrays were identified by blasting the knob180 and CentC consensus sequences (Gent, Wang, and Dawe 2017), a TR-1 consensus (TTCTTTATATTCCAACTTTTTAGCAACTGTATGGTGGAAAAAGGTGTCTTACAACCT TAACCTATGTTTGGACAGTTCTCTCGTGCAATTTGGCTAAATTTCCCATGGTCTTTAT TTTATTTTGAGAAACGATGTGGTATAATGATGTGCGATGTTTTACTTGAGTGGACAT AAACACCATTTAGGTATGCCTTGAATAGAGGGGATTATTGGAAACCTGGTATCACAA AAGGTCATTAGCTAGCCCAATAACGTCTTCATCCACTAGTTATACTCTAATACCCTCT AGTGTGAATACAATGCCCACAATATCATAGAAACGTCATTTGAGGTTTAAAAGGTG ATCTATTGTTTTGAA), subtelomeric repeat (NCBI CL569186.1) and ribosomal DNA intergenic spacer sequences (NCBI AF013103.1) against the B73-Ab10 genome. Knobs were defined as repeat clusters (>=500 Kb) that are composed of at least 10% repeat consensus sequences (knob180 and TR-1) with no more than 100 Kb spacing between repeat units. This

definition of knob180 knobs excludes the subtelomeric knob180 arrays. CentC arrays are defined as repeat clusters (>=100 Kb) that are composed of at least 10% CentC consensus sequences.

Non-overlapping repeat units were quantified in each repeat array with custom script repeat_analyses.py. Five major families of the long terminal repeat (LTR)-retrotransposons in knobs, CentC arrays and active centromeres were individually quantified with bedtools (v2.28.0) (Quinlan 2014). The *Opie-Ji* family includes *Opie, Ji, Ruda* and *Giepum* and the *Prem1* family is composed of *Prem1*, *Xilon, Diguus,* and *Tekay* (SanMiguel and Vitte 2009). Centromeric retrotransposons CRM1 and CRM2 were quantified together and annotated as CRM in active centromeric regions.

To assess the enrichment of mappable repeat elements in functional centromeres, each of the elements were first classified into uniquely mappable or non-uniquely mappable groups. A cutoff of MAPQ20 was applied to the alignment file, and bedtools (v2.28.0) was used to estimate genome coverage at the base pair level (-bga). Non-uniquely mapped locations (<=2 or >=101 aligned reads) were merged into islands with a maximum interval of 1 Kb. CENH3 ChIP-seq enrichment for the unique and non-unique fractions of CentC, CRM and five major LTR retrotransposon families were then individually assessed. ChIP enrichment was calculated by normalizing ChIP-seq against the input genome-seq alignment bam files using a RPKM normalization method with deepTools (v3.2.1) (Ramírez et al. 2014). Default options were used except for the following parameters: --operation ratio --scaleFactorsMethod None --normalizeUsing RPKM.

Availability of data and materials

The B73-Ab10 inbred can be obtained as PI 690316 at the Germplasm Resources Information Network (GRIN), Ames, Iowa. All genomic sequence and Bionano data can be obtained at the NCBI SRA under Bioproject PRJEB35367 [69]. The RNA-seq data is deposited in EBI (Accession number E-MTAB-8641) [70]. The code used in this study is available at the GitHub repository https://github.com/dawelab/Ab10-Assembly [71].

Table 3.1. Assembly metrics of the B73-Ab10 genome.

	Contigs			Pseudomolecules		
	N50 (Mb)	N90 (Mb)	Max Size (Mb)	Contig Number	Total Length (Mb)	Gap[a] Length (Mb)
Nanopore	2.0	0.5	8.3	1673	2161.1	93.2
PacBio	41.2	7.1	156.3	216	2162.7	2.6
Merged	162.0	20.4	235.9	63	2162.8	1.3

[a] Gaps longer than 10 Ns.

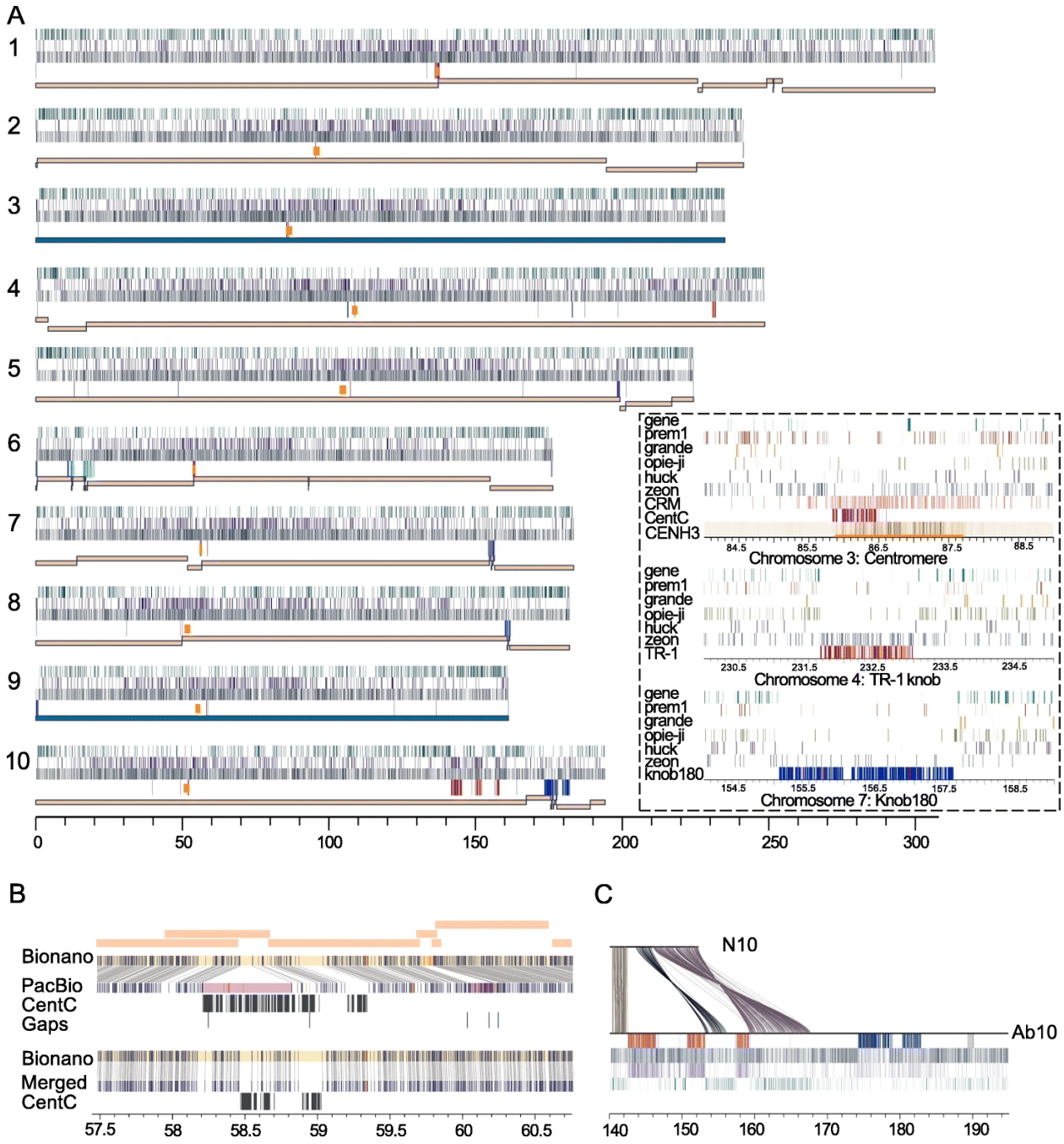


Figure 3.1. Assembly of the B73-Ab10 genome. **A**) Whole-genome view. For each chromosome, the top to bottom tracks are: gene density, cinful-zeon retrotransposon density, Gypsy superfamily retrotransposon density in 10 Kb sliding windows, repeat location (knob180 in blue, TR-1 in red, 45S rDNA in teal, CentC in magenta), and the distribution of gapless contigs. CENH3 ChIP-seq peaks identifying centromeres are marked by orange rectangles. The inset shows the centromere on chromosome 3, TR-1-rich knob on chromosome 4, and knob180-rich knob on chromosome 7. The five most common retroelement families are shown for each

panel, along with Centromeric Retrotransposons (CRM) for the centromere. CENH3 enrichment in chromosome 3 is displayed in a heatmap. **B**) The impact of assembly merging over a CentC-rich region on chromosome 9. Seven contigs (orange, above) from the PacBio assembly were originally misassembled, as can be seen in the alignment to the Bionano map (connecting lines show matching sites). CentC tracts and gaps are annotated. Assembly merging corrected the output, leaving an 11 Kb gap that was filled with nanopore reads. **C**) Sequence alignment between normal chromosome 10 from B73 (N10) (140Mb-152Mb) and Ab10 (140Mb-195Mb) from B73-Ab10. Annotation is as in A, with Kindr genes marked with black bars in the top track. Links show homologous regions larger than 500bp.

References

Albert, P. S., Z. Gao, T. V. Danilova, and J. A. Birchler. 2010. "Diversity of Chromosomal Karyotypes in Maize and Its Relatives." *Cytogenetic and Genome Research* 129 (1-3): 6–16.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Ananiev, E. V., R. L. Phillips, and H. W. Rines. 1998. "A Knob-Associated Tandem Repeat in Maize Capable of Forming Fold-Back DNA Segments: Are Chromosome Knobs Megatransposons?" *Proceedings of the National Academy of Sciences of the United States of America* 95 (18): 10785–90.

Belser, Caroline, Benjamin Istace, Erwan Denis, Marion Dubarry, Franc-Christophe Baurens, Cyril Falentin, Mathieu Genete, et al. 2018. "Chromosome-Scale Assemblies of Plant Genomes Using Nanopore Long Reads and Optical Maps." *Nature Plants* 4 (11): 879–87.

Bilinski, Paul, Patrice S. Albert, Jeremy J. Berg, James A. Birchler, Mark N. Grote, Anne Lorant, Juvenal Quezada, Kelly Swarts, Jinliang Yang, and Jeffrey Ross-Ibarra. 2018. "Parallel Altitudinal Clines Reveal Trends in Adaptive Evolution of Genome Size in Zea Mays." *PLoS Genetics* 14 (5): e1007162.

Campbell, Michael S., Meiyee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. "MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations." *Plant Physiology* 164 (2): 513–24.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (12): 1050–54.

"CyVerse Data Commons." n.d. Accessed November 12, 2019. http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Daniel_Laspisa_B73_RefGen_v4CEN_Feb_2019.

Dawe, R. Kelly, Elizabeth G. Lowry, Jonathan I. Gent, Michelle C. Stitzer, Kyle W. Swentowsky, David M. Higgins, Jeffrey Ross-Ibarra, et al. 2018. "A Kinesin-14 Motor Activates Neocentromeres to Promote Meiotic Drive in Maize." *Cell* 173 (4): 839–50.e18.

Deschamps, Stéphane, Yun Zhang, Victor Llaca, Liang Ye, Abhijit Sanyal, Matthew King, Gregory May, and Haining Lin. 2018. "A Chromosome-Scale Assembly of the Sorghum

Genome Using Nanopore Sequencing and Optical Mapping." *Nature Communications* 9 (1): 4844.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Doyle, J. J., and J. L. Doyle. 1987. "A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf Tissues." *Phytochem Bull* 19: 11–15.

Du, Huilong, and Chengzhi Liang. 2019. "Assembly of Chromosome-Scale Contigs by Efficiently Resolving Repetitive Sequences with Long Reads." *Nature Communications* 10 (1): 5360.

Dyer, Kelly A., Brian Charlesworth, and John Jaenike. 2007. "Chromosome-Wide Linkage Disequilibrium as a Consequence of Meiotic Drive." *Proceedings of the National Academy of Sciences of the United States of America* 104 (5): 1587–92.

Eilbeck, Karen, Barry Moore, Carson Holt, and Mark Yandell. 2009. "Quantitative Measures for the Management and Comparison of Annotated Genomes." *BMC Bioinformatics* 10 (February): 67.

Gent, Jonathan I., Na Wang, and R. Kelly Dawe. 2017. "Stable Centromere Positioning in Diverse Sequence Contexts of Complex and Satellite Centromeres of Maize and Wild Relatives." *Genome Biology* 18 (1): 121.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology*. https://doi.org/10.1038/nbt.1883.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith Jr, Linda I. Hannick, Rama Maiti, et al. 2003. "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies." *Nucleic Acids Research* 31 (19): 5654–66.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq: Reference Generation and Analysis with Trinity." *Nature Protocols* 8 (8). https://doi.org/10.1038/nprot.2013.084.

Hiatt, Evelyn N., and R. Kelly Dawe. 2003. “Four Loci on Abnormal Chromosome 10 Contribute to Meiotic Drive in Maize.” *Genetics* 164 (2): 699–709.

Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, et al. 2017. “Improved Maize Reference Genome with Single-Molecule Technologies.” *Nature* 546 (7659): 524–27.

Kent, W. James. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–64.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. “Assembly of Long, Error-Prone Reads Using Repeat Graphs.” *Nature Biotechnology* 37 (5): 540–46.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36.

Lee, Michael, Natalya Sharopova, William D. Beavis, David Grant, Maria Katt, Deborah Blair, and Arnel Hallauer. 2002. “Expanding the Genetic Map of Maize with the Intermated B73 X Mo17 (IBM) Population.” *Plant Molecular Biology* 48 (5-6): 453–61.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

———. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–10.

———. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352.

Liu, Ruolin, and Julie Dickerson. 2017. “Strawberry: Fast and Accurate Genome-Guided Transcript Reconstruction and Quantification from RNA-Seq.” *PLoS Computational Biology* 13 (11): e1005851.

Lu, Fei, Maria C. Romay, Jeffrey C. Glaubitz, Peter J. Bradbury, Robert J. Elshire, Tianyu Wang, Yu Li, et al. 2015. “High-Resolution Genetic Mapping of Maize Pan-Genome Sequence Anchors.” *Nature Communications* 6 (April): 6914.

Luo, Meizhong, and Rod A. Wing. 2003. “An Improved Method for Plant BAC Library Construction.” *Methods in Molecular Biology* 236: 3–20.

Manchanda, Nancy, John L. Portwood, Margaret R. Woodhouse, Arun S. Seetharam, Carolyn J. Lawrence-Dill, Carson M. Andorf, and Matthew B. Hufford. 2019. “GenomeQC: A Quality Assessment Tool for Genome Assemblies and Gene Structure Annotations.” *bioRxiv*. https://doi.org/10.1101/795237.

Mapleson, Daniel, Luca Venturini, Gemy Kaithakottil, and David Swarbreck. 2018. “Efficient and Accurate Detection of Splice Junctions from RNA-Seq with Portcullis.” *GigaScience* 7 (12). https://doi.org/10.1093/gigascience/giy131.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2019. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *bioRxiv*. https://doi.org/10.1101/735928.

Mroczek, Rebecca J., Juliana R. Melo, Amy C. Luce, Evelyn N. Hiatt, and R. Kelly Dawe. 2006. “The Maize Ab10 Meiotic Drive System Maps to Supernumerary Sequences in a Large Complex Haplotype.” *Genetics* 174 (1): 145–54.

Nannas, Natalie J., and R. Kelly Dawe. 2015. “Genetic and Genomic Toolbox of Zea Mays.” *Genetics* 199 (3): 655–69.

Ou, Shujun, Jinfeng Chen, and Ning Jiang. 2018. “Assessing Genome Assembly Quality Using the LTR Assembly Index (LAI).” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky730.

Ou, Shujun, and Ning Jiang. 2018. “LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons.” *Plant Physiology* 176 (2): 1410–22.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Doreen Ware, Thomas Peterson, Ning Jiang, Candice N. Hirsch, and Matthew B. Hufford. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *bioRxiv*. https://doi.org/10.1101/657890.

Peacock, W. J., E. S. Dennis, M. M. Rhoades, and A. J. Pryor. 1981. “Highly Repeated DNA Sequence Limited to Knob Heterochromatin in Maize.” *Proceedings of the National*

Academy of Sciences of the United States of America 78 (7): 4490–94.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33 (3): 290–95.

Portwood, John L., 2nd, Margaret R. Woodhouse, Ethalinda K. Cannon, Jack M. Gardiner, Lisa C. Harper, Mary L. Schaeffer, Jesse R. Walsh, et al. 2019. "MaizeGDB 2018: The Maize Multi-Genome Genetics and Genomics Database." *Nucleic Acids Research* 47 (D1): D1146–54.

Pyhäjärvi, Tanja, Matthew B. Hufford, Sofiane Mezmouk, and Jeffrey Ross-Ibarra. 2013. "Complex Patterns of Local Adaptation in Teosinte." *Genome Biology and Evolution* 5 (9): 1594–1609.

Quinlan, Aaron R. 2014. "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 11.12.1–34.

Ramírez, Fidel, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. 2014. "deepTools: A Flexible Platform for Exploring Deep-Sequencing Data." *Nucleic Acids Research* 42 (Web Server issue): W187–91.

Rhoades, M. M. 1942. "Preferential Segregation in Maize." *Genetics* 27 (4): 395–407.

Ricci, William A., Zefu Lu, Lexiang Ji, Alexandre P. Marand, Christina L. Ethridge, Nathalie G. Murphy, Jaclyn M. Noshay, et al. 2019. "Widespread Long-Range Cis-Regulatory Elements in the Maize Genome." *Nature Plants*, November. https://doi.org/10.1038/s41477-019-0547-0.

SanMiguel, Phillip, and Clémentine Vitte. 2009. "The LTR-Retrotransposons of Maize." In *Handbook of Maize: Genetics and Genomics*, edited by Jeffrey L. Bennetzen and Sarah Hake, 307–27. New York, NY: Springer New York.

Sanz-Alferez, Soledad, Phillip SanMiguel, Young-Kwan Jin, Patricia S. Springer, and Jeffrey L. Bennetzen. 2003. "Structure and Evolution of the Cinful Retrotransposon Family of Maize." *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 46 (5): 745–52.

Song, Li, Sarven Sabunciyan, and Liliana Florea. 2016. "CLASS2: Accurate and Efficient Splice Variant Annotation from RNA-Seq Reads." *Nucleic Acids Research* 44 (10): e98.

Stovner, Endre Bakken, and Pål Sætrom. 2019. “epic2 Efficiently Finds Diffuse Domains in ChIP-Seq Data.” *Bioinformatics* 35 (21): 4392–93.

Tang, Haibao, Xingtan Zhang, Chenyong Miao, Jisen Zhang, Ray Ming, James C. Schnable, Patrick S. Schnable, Eric Lyons, and Jianguo Lu. 2015. “ALLMAPS: Robust Scaffold Ordering Based on Multiple Maps.” *Genome Biology* 16 (January): 3.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols*. https://doi.org/10.1038/nprot.2012.016.

Udall, Joshua, and R. Kelly Dawe. 2017. “Is It Ordered Correctly? Validating Genome Assemblies by Optical Mapping.” *The Plant Cell*, December. https://doi.org/10.1105/tpc.17.00514.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46.

Venturini, Luca, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck. 2018. “Leveraging Multiple Transcriptome Assembly Methods for Improved Gene Structure Annotation.” *GigaScience* 7 (8). https://doi.org/10.1093/gigascience/giy093.

Vilella, Albert J., Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. 2009. “EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates.” *Genome Research* 19 (2): 327–35.

Vollger, Mitchell R., Philip C. Dishuck, Melanie Sorensen, Annemarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. 2019. “Long-Read Sequence and Assembly of Segmental Duplications.” *Nature Methods* 16 (1): 88–94.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PloS One* 9 (11): e112963.

Wang, Bo, Michael Regulski, Elizabeth Tseng, Andrew Olson, Sara Goodwin, W. Richard McCombie, and Doreen Ware. 2018. “A Comparative Transcriptional Landscape of Maize

and Sorghum Obtained by Single-Molecule Sequencing." *Genome Research* 28 (6): 921–32.

Wick, Ryan R., Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. 2015. "Bandage: Interactive Visualization of de Novo Genome Assemblies." *Bioinformatics* 31 (20): 3350–52.

Wolfgruber, Thomas K., Anupma Sharma, Kevin L. Schneider, Patrice S. Albert, Dal-Hoe Koo, Jinghua Shi, Zhi Gao, et al. 2009. "Maize Centromere Structure and Evolution: Sequence Analysis of Centromeres 2 and 5 Reveals Dynamic Loci Shaped Primarily by Retrotransposons." *PLoS Genetics* 5 (11): e1000743.

Wu, Thomas D., and Colin K. Watanabe. 2005. "GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences." *Bioinformatics* 21 (9): 1859–75.

Xu, Gui-Cai, Tian-Jun Xu, Rui Zhu, Yan Zhang, Shang-Qi Li, Hong-Wei Wang, and Jiong-Tang Li. 2019. "LR_Gapcloser: A Tiling Path-Based Gap Closer That Uses Long Reads to Complete Genome Assembly." *GigaScience* 8 (1). https://doi.org/10.1093/gigascience/giy157.

Yang, Ning, Jie Liu, Qiang Gao, Songtao Gui, Lu Chen, Linfeng Yang, Juan Huang, et al. 2019. "Genome Assembly of a Tropical Maize Inbred Line Provides Insights into Structural Variation and Crop Improvement." *Nature Genetics* 51 (6): 1052–59.

Zhang, Ren-Gang, Zhao-Xuan Wang, Shujun Ou, and Guang-Yuan Li. n.d. "TEsorter: Lineage-Level Classification of Transposable Elements Using Conserved Protein Domains." https://doi.org/10.1101/800177.

CHAPTER 4

ARCHAIC INTROGRESSION DIVERSIFIED AND EXPANDED THE MAIZE PANGENOME[3]

[3] Liu, J.& Dawe, R. K. (2021). To be submitted to *Nature Genetics*.

Abstract

The full extent of genomic variation in maize and its wild relatives has yet to be fully evaluated at a whole genome scale. Here we used alignments of 26 genomes to demonstrate that much of the structural diversity in maize accumulated is a result of differential transposon insertion in archaic lineages. We observed age-stratified haplotype patterns in the pericentromeric areas of all chromosomes representing multiple introgression events in the last 0.5 million years. Signatures of ancient introgression were also observed in the repeat arrays of centromeres, heterochromatic knobs and nucleolus organizer regions. These forces and large historical population sizes gave rise to a maize pangenome of roughly 7.9 Gb -- well over three times larger than any single genome. Analysis of variation in cis regulatory regions and gene expression patterns indicate that ancient genomic remnants contribute important genetic diversity.

Introduction

Maize was domesticated about 9,000 years ago from large populations of *Zea mays* subspecies *parviglumis*, a tall grass with freely shattering seeds (Matsuoka et al. 2002). The early domesticated form was grown in the lowlands and highlands of Mexico, where subsequent gene flow with *Zea mays* subspecies *mexicana* introduced additional genetic variation. Humans then transported the crop in two directions, northward through the American desert and into Canada and southward to the Andes and the islands of the Caribbean (Ross-Ibarra et al. 2009; van Heerwaarden et al. 2011; Hufford et al. 2013). Along the way, native Americans cultivated hundreds of local landraces that not only retained much of original variation, but expanded the range of diversity as a result of strong selection for local environments (Vigouroux et al. 2008).

In the early twentieth century modern breeding lines were developed from a rich mixture of US and Caribbean landraces (Unterseer et al. 2016) that are crossed in pairs to maximize yield by heterosis. This unique domestication history, involving large population sizes and selection through diverse environments, provided the underpinnings of the modern crop that is now grown on every continent.

Molecular data show that there is extraordinary variation in genome size, gene content, methylation status and repeat composition among maize lines (Chia et al. 2012; Sun et al. 2018; Haberer et al. 2020; Hufford et al. 2021). Whole-genome identity between any two maize lines can be as low as 50% (Anderson et al. 2019; Haberer et al. 2020). The genomic diversity is mainly a result of transposon insertion over the past three million years, which inflated genome size by two to five fold (Sanmiguel and Bennetzen 1998) and altered gene spacing and arrangement (Fu and Dooner 2002; Morgante et al. 2005; Brunner et al. 2005). Transposons not only restructure the genomic landscape but contribute important regulatory information that can alter genetic networks (Feschotte 2008). For example, a *Hopscotch* transposable element acts as an enhancer of the *Teosinte branched 1* gene to repress branch outgrowth (Studer et al. 2011); promoting the single stem growth habit that differentiates cultivated maize from its highly branched wild ancestors.

What the evolutionary precursors of maize looked like at the time of the major TE-mediated genome expansions can only be inferred. However the existence of ancient, deeply divergent haplotypes in modern maize suggest the ancestral populations were diverse and structured (Goloubinoff et al. 1993; White and Doebley 1999; Ching et al. 2002). In a recent study of the *sugary1* locus, the authors observed two distinct haplotype groups differentiated by 165 SNPs in a ~9 kb gene region, where each haplotype group was observed in both maize lines

and teosinte accessions, suggesting an ancient origin (Hu et al. 2021). Another dramatic example was provided by Fu and Dooner (Fu and Dooner 2002), who characterized two haplotypes of the *bz1* region. They observed almost total absence of homology in the intergenic spaces: among the 23 transposons annotated over a ~ 180kb of combined sequence, only one transposon was conserved. The authors suggested that the two regions evolved separately but survived the domestication bottleneck to segregate in modern maize. The same reasoning may apply to maize centromeres, where a small number of highly differentiated haplotypes have been described (Schneider et al. 2016).

Here we provide evidence that the deep haplotype diversity in maize can be attributed in part to introgression among highly differentiated archaic lineages. Using genome assemblies from 26 inbreds, we integrated structural variant and SNP datasets to identify haplotypes and estimate divergence times. Regions of high linkage disequilibrium, as is typical of pericentromeric areas and regions composed primarily of tandem repeats, showed clear evidence of evolutionary stratification at periodic intervals. These periods are not associated with major changes in estimated population size, suggesting that the patterns are a result of intercrossing and introgression among archaic subpopulations. The mixture of haplotypes from different lineages enlarges the functional pangenome to roughly 7.8 Gb, roughly 3.7 times larger than any single genome, and adds genetic variation that significantly alters measured gene expression patterns.

RESULTS

Extreme structural variation revealed by whole-genome alignment

We recently completed 26 high quality genomes for the founders of the Nested Association Mapping (NAM) population (Hufford et al. 2021), a rich collection chosen to

represent the diversity of maize, including temperate lines, tropical lines, sweet corn and popcorn (McMullen et al. 2009). The PacBio and optical-map based assemblies have excellent contiguity over genic and intergenic regions, as well as centromeres and heterochromatic knobs (megabase-scale heterochromatic regions; (Dawe et al. 2018)). The release included expression data from ten tissues for each inbred, and full annotation for genes, transposons, functional centromeres (by CENP-A/CENH3 ChIP-seq) and unmethylated domains representing regulatory regions. As such they present an exceptional resource to study the diversity and genomic history of maize.

Previous studies of the NAM lines have described structural variation and haplotype structure by aligning resequencing data to the primary B73 reference genome, using both short and long-reads (Hufford et al. 2021; Chia et al. 2012; Gore et al. 2009). However this approach necessarily fails for insertions that are larger than the read length and preferentially recovers deletions. Using the read alignment method, we identified ~250 Mb of deletions and fewer than ~7 Mb of insertions for each line (Hufford et al. 2021). A partial solution to this problem is the merged alignment blocks method that was used to identify structural variants with whole-genome alignments across six European flint lines (Haberer et al. 2020). While this approach identifies deletions and insertions in roughly equal proportions, it fails to capture complex events where there is variation in both query and reference (Figure S4.1). As a third approach to identify both simple and complex events, we implemented the longest increasing subsequence (LIS) algorithm (Rani and Rajpoot 2016; Abouelhoda and Ohlebusch 2005) in a two-step chaining procedure (Figure S4.2 and S4.3). The pipeline was executed in an all-by-all format such that there were 325 comparisons for each chromosome. As rearranged and small segments were incorporated in the chain, we were able to characterize ~270,000 indels embedded in rearranged domains that are not detected by the merged alignment blocks method with all-by-all comparison

(Haberer et al. 2020) (Figure S4.6). Among the ~19 million unaligned segments identified across 26 lines (Table S4.3 and S4.4), 22.5% were simple deletions or insertions and 77.5% were more complex, where the unaligned regions differed in size and SVs were called on both reference and query.

We identified over three times as many indels and ten times as many large inversions in the NAM lines than the previous estimate based on long long-read mapping (Hufford et al. 2021). We also identified 5314 large tandem duplications (>10Kb), including simple segmental duplications (Figure S4.5A) and nested duplications (Figure S4.5B). Over 50% of tandem duplications overlap with unmethylated regions (UMRs, likely regulatory domains) and expressed genes (Figure S4.5B). Many inversions and large duplications have sustained additional insertions and rearrangements, suggesting ancient origins (Figure S4.6). A complete file of all indels, inversions, and duplications in the NAM lines are available in Supplemental Dataset 1. The improved SV database will be useful for linkage and GWAS analysis in these inbreds, for which there is already extensive phenotypic information (Wallace et al. 2014).

Ancient diversity visible in retained haplotype blocks

Our whole genome alignments revealed numerous megabase-scale haplotypes around centromeres. A dramatic example of alternative pericentromeric haplotypes can be seen in a 10 Mb region of chromosome 8 (position 43-53 Mb), where CML52, HP301, IL14H, Mo18W, and P39 differ from all other inbreds (Figure 4.1A). Pairwise alignment among the five lines with the alternative haplotype showed that they share a common ancestor (Figure 4.1B). There is substantial variation in TE subfamilies for each haplotype (Figure 4.1C, 4.1D), suggesting genetic isolation over long time spans where different TE subfamilies proliferated. The

haplotypes also differed in gene distribution (Figure S4.9A, S4.9B). Among the 91 potential protein-coding genes in the chromosome 8 haplotype region, 42% were found to be non-syntenic, suggesting positive selection in the source populations. Additional haplotypes were identified in pericentromeric areas in six other chromosomes, including a 7 Mb region on chromosome 2 (95-102 Mb), a 10 Mb segment on chromosome 3 (82-92 Mb), an 8 Mb section on chromosome 5 (104-112 Mb), a 6 Mb area on chromosome 9 (55-61 Mb) and a 7 Mb region on chromosome 10 (42-49 Mb) (Figure S4.7 and S4.8). The temperate (corn belt) lines tend to have different pericentromeric haplotypes than the northern flint lines (sweet and popcorn) (Table S4.5), providing further support for the expectation that the alternative haplotypes provide beneficial alleles in different environments.

We also observed a 3 Mb segment of DNA with apparent ancestry to *Tripsacum*, the sister genus to *Zea*. This introgressed region is embedded in the nucleolus organizer region of a subset 20 NAM lines (NOR) (Figure S4.13C). There is no homology between the introgressed segment and any other region in maize, making it difficult to interpret the origin of this region.

Evidence for repeated genetic isolation and introgression over a 0.5 mya timespan

To estimate the age of the visible haplotypes and the overall divergence of the NAM inbreds, we identified SNPs differentiating the NAM genomes, and inferred divergence time using an average mutation rate of 3.3×10^{-8}. We scored SNPs in all syntenic alignments using B73 as a reference. Ancestry profiling of the pericentromeric haplotypes demonstrated that they are archaic, with divergence times as old as 0.5 million years (Figure 4.2A). Ancient haplotypes often contained segments of lesser age (Figure 4.1E, Figure S4.10, Figure 4.2C), indicating periodic recombination among haplotypes. Clustering of divergence times over pericentromeric areas revealed strata across every chromosome with differing dates of divergence (Figure 4.2C).

For example, while a divergence of 0.45 million years was found between the two major haplotypes in region 43.5-46.5 Mb on chromosome 10, we observed a younger stratum of 0.13 million years in lines CML322, Ki3, M162W, Tx303 and Tzi8 (Figure 4.2C, Figure S4.9).

The estimated age of the haplotypes greatly exceeds the time frame of maize domestication (9,000 years) as well as the radiation of the *Zea* lineage (~100-300 kya; (Ross-Ibarra et al. 2009)). Therefore we expected to see similar haplotype diversity in the wild teosinte relatives of maize. Maize was originally domesticated from *Zea mays ssp parviglumis* and later introgressed with *ssp mexicana* (Hufford et al. 2013). The *Zea* clade also includes *Zea mays ssp huehuetenangensis* and the related species *Zea diploperennis* (Iltis et al. 1979; Doebley and Others 1990). Short read data are available for numerous accessions of these teosintes. Divergence analysis confirmed all the observed maize pericentromeric haplotypes and revealed additional haplotypes (Figure S4.11). In the NAM lines there is only one major pericentromeric haplotype on chromosome 1 and 6, while in teosinte there are two alternative haplotypes (Figure S4.11), consistent with prior data showing that domestication reduced genetic diversity (Wang et al. 2017). There were also differences in the distribution of haplotypes. The B73 haplotype of chromosome 2 is only present in six maize inbreds, but it is the prevalent haplotype (82%) among teosinte lines (Figure S4.10 and S4.11).

When all pericentromeric regions were analyzed together, a multi-modal distribution was observed, with local peaks found at 0.001, 0.015, 0.05, 0.08, 0.13, 0.3, and 0.45 million years (Figure 4.2D). A similar distribution was identified for divergence time across the whole genome, with local maxima at the same intervals (Figure 4.2B). The enrichment at particular divergence times could be a result of introgression between isolated populations, contractions or expansions in effective population size, or combinations of introgression and changes in

population size. We tested for historical changes in population size using the Multiple Markovian Coalescent (MSMC) model (Schiffels and Durbin 2014). Consistent with previous studies (Wang et al. 2017), we observed a continuously decreasing trend in effective population size from 0.5 million years ago to 200 years ago (Figure 4.2E). The absence of periodicity in historical population size estimates suggests that the observed evolutionary stratification most likely reflects repeated introgression.

Ancient haplotypes in repeat arrays of centromeres, knobs and NOR

The fact that evolutionary strata are most apparent in pericentromeric regions can be attributed to the fact that recombination is low in these domains. Low recombination helps to preserve ancient genetic linkages. In contrast, high recombination intersperses older and newer regions and erases visible stratification. By this reasoning, any region where recombination is low, such as centromeres, nucleus organizer regions (NOR) and knobs (Gore et al. 2009; Shi et al. 2010; Ghaffari et al. 2013), and should contain visible signatures of ancient introgression with periodicities roughly matching what we observed genome wide.

We assessed structural similarity in centromeres (CentC), knobs (knob180 and TR-1) and NOR (rDNA) repeat arrays through independent alignment of monomers and syntenic transposons. Repeat monomers do not align well using whole-genome alignment methods due to the low frequency of unique kmers in long arrays. Instead of LIS we used a dot matrix alignment method where the similarity of each monomer pair was assessed by BLAT (Mount 2007). Divergence time of knobs was estimated with SNPs in uniquely aligned TEs. The data show that the haplotypes inferred from pericentromeric regions correspond to distinctive centromere repeat array haplotypes (Figure 4.3A, Figure S4.16 and S4.17). For example, CentC arrays in lines with

the alternative haplotype on chromosome 8, including CML52, HP301, IL14H, Mo18W, and P39, align well to each other, while no homology was identified between them and the haplotype represented by B73 (Figure 4.13A). Nevertheless, multiple structural variants were detected among centromeres in the same each haplotype group, suggesting rapid evolution in these repeat-rich regions (Figure 4.3A, Figure S4.16).

A hierarchical degree of divergence was also observed in the NOR and knobs (Figure 4.3C). All by all alignment revealed three distinct clusters for the NOR array (Figure S4.15) and the upstream array of knob180 repeats on the short arm of chromosome 6, with progressive divergence times of 0.1, 0.22 and 0.3 million years (Figure 4.3D, 4.3E). Likewise, among the eight classical knobs we surveyed (Albert et al. 2010; Hufford et al. 2021), five have diverged for over 0.2 million years (Figure S4.19). The knob on the short arm of chromosome 9 is found in 22 inbreds and these separated into three clusters, where two clusters diverged from B73 over 0.3 million years ago (Figure S4.14). However, we do not expect all knobs in maize to be ancient; knobs are subject to meiotic drive and should occasionally sweep across populations (Hall and Dawe 2017). The knob180-rich knob on 5L and 7L, and the TR-1-rich knob on chromosome 4 have divergence times less than 0.1 million years (Figure S4.18, S4.19). These knobs also show low diversity in *parviglumis*, *mexicana*, *huehuetenangensis*, and *diploperennis*, consistent with a more recent emergence (Figure S4.11, S4.12C).

Total pangenome size far exceeds the size of any single genome

The observation that maize retains structural polymorphism dating to introgression events from hundreds of thousands of years ago suggests a large and fluid pangenome. To estimate the total pangenome size, we merged the overlaps from all-by-all comparisons to arrive at a total

pangome estimate of 7.9 Gb (Figure 4.4A) which is ~3.7 times larger than the average assembled size of any single genome (Hufford et al. 2021). Permutation of 26 genomes showed that over 80% of the pan-genome is identified in 10 inbreds, and beyond that additional sequences are mainly line-specific (Figure 4.4A). Only ~4.9% of the total pangenome (0.38 Gb) is conserved among all lines with the remaining 7.5 Gb segregating among lines at various frequencies (Figure 4.4B).

Ancient regions significantly expand gene regulatory diversity

The sequence diversity introduced during ancient introgression events likely affected phenotype, perhaps with adaptive value (Arnold 2004). Although we cannot assess the phenotypic impacts, we can assess how the divergent sequences impact gene expression. We were particularly interested in testing how divergent sequences in the intergenic spaces might alter gene expression. Regulatory domains can be identified in the NAM genomes by unmethylated regions (UMRs) (Hufford et al. 2021), which are known to correlate with cis regulatory elements (Ricci et al. 2019). As many as a third of these potential regulatory regions lie in sequence with homology to transposons (Oka et al. 2017).

High polymorphism was observed for UMRs between NAM lines and B73, among which more than 20% was found to be non-syntenic due to genomic dissimilarity, and 10% overlap with regions with a divergence time over 0.2 million years (Figure 4.5A, 4.5B). To test the effect of UMR divergence on gene expression, we divided them into groups according to their relative genetic distance to B73, and evaluated the corresponding gene expression change for UMR groups of various diversity. A significant differential gene expression level was identified between the groups of high diversity (>0.2 million years) and the 0-50K group (Figure 4.5C),

indicating that the high sequence polymorphism in UMRs affected transcription factor binding and in turn altered gene expression level.

Discussion

Like single-nucleotide mutations, SVs follow evolutionary trajectories (Mérot et al. 2020). SVs can spread from a single genome to a population or even across species through recombination, introgression, and drift (Conrad and Hurles 2007). Inversions or clustered SVs can reduce local recombination (Rowan et al. 2019), creating haplotype blocks identifiable by patterns of linkage disequilibrium and distinct age estimates (Thompson and Jiggins 2014). Haplotype blocks that link favorable alleles can spread through populations (Kirkpatrick and Barton 2006) and have outsized impacts on phenotype. Adaptive haplotypes formed by the accumulation of SVs are known to be associated with sex chromosomes (Ross et al. 2005), social organization in ants (Purcell et al. 2014), wing color in butterfly (Joron et al. 2011) and flowering time in sunflower (Todesco et al. 2020).

The distribution of haplotypes can also reflect demographic isolation and introgression, as illustrated by the large literature on admixture among early humans and extinct Neanderthal and Denisovan lineages (Li and Durbin 2011; Almarri et al. 2020). These archaic, reproductively isolated populations accrued different SV and SNP profiles, which are identifiable as segregating polymorphisms in modern humans. Similar arguments have been made for many plant lineages (Rieseberg et al. 2003; Arnold 2004; Harrison and Larson 2014; Suarez-Gonzalez et al. 2018) which are far more likely to undergo hybridization than animals (Mallet et al. 2016). Whether or not these events provide adaptive value is often unclear (Abbott et al. 2013), however, it is

evident that hybridization and introgression are diversifying forces that rapidly introduce complex alleles and multigene traits.

Our detailed analysis of SV-based haplotype profiles show that multiple introgression events occurred during the evolutionary history of maize over the last 1 million years (Figure 4.6). We identified ancient haplotypes among the maize NAM population in the areas of high linkage disequilibrium, including pericentromeric areas, centromeres, knobs and NOR. Introgression was particularly frequent around 0.5, 0.35, 0.25, and 0.1 and 0.015 million years ago. The most recent hybridization event that accounted for 13.2% of genomic diversity took place around 6,000 years before domestication. A 3 Mb introgressed segment of DNA from a relative of *Tripsacum* was also identified, however it is difficult to estimate the date of this ancient hybridization event.

During domestication, there were introgressions between the early domesticated maize derived from *Z. parviglumis* and the closely related teosinte *Z. mexicana* (Matsuoka et al. 2002; Hufford et al. 2013). We observed that the major archaic haplotypes are present in both Z. *parviglumis* and *Z. mexicana* with maize having higher similarity to *Z. parviglumis* (Figure S4.11). We also observed higher polymorphism in the *parviglumis* population, consistent with the known loss of genetic diversity that occurred during the domestication bottleneck (Wang et al. 2017). The *Z. mexicana* introgressions are thought to have alleviated the severity of the bottleneck by bringing in lost diversity (Hufford et al. 2013) and serve to illustrate the longer-term dynamics of the *Zea* lineage, where multiple waves of hybridization and introgression help to build the rich genetic diversity in this species.

It is noteworthy that the most conspicuous structural variation in maize is localized around centromeres. We observed that seven chromosomes segregate for at least two major,

ancient haplotypes. Regions of reduced recombination are expected to show a higher rate of deleterious variants and have particularly strong contributions to inbreeding depression (Charlesworth and Willis 2009). Genotyping data from maize demonstrate that pericentromeres are more likely to retain heterozygosity in serially inbred lines (McMullen et al. 2009). Further, a higher proportion of quantitative trait loci (QTL) for hybrid vigor (heterosis) map to pericentromeric areas than expected by chance (Thiemann et al. 2014; Martinez et al. 2016), These data support the dominance theory for hybrid vigor, where deleterious alleles in one genotype are complemented by the second genotype. In this context, the hybridization and retention of distinct centromeric haplotypes may reflect natural selection for improved vigor in the wild. Genomic comparisons across multiple inbred lines may help to explain the differential heterosis effects that are observed among maize elite lines (Springer and Stupar 2007), facilitate the fine mapping of candidate loci that help predict heterotic effects.

Hybridization and introgression over hundreds of thousands of years have created a remarkably large and diverse pangenome. We find that only 4.6% of the sequence is conserved across all 26 lines. We estimate that the pangenome totals ~7.9 Gb in 26 genomes, however, it is clear that pangenome size will continue to increase with the addition of new genomes (Figure 4.4A). Each new accession will contain additional structural variation with novel genes and neoUMRs that further expand the pool of genetic variation available to breeders.

Methods

Genome alignment and structural variant characterization

The accuracy of structural variant detection is dependent on the reliability of sequence alignment. As maize genomes are highly repetitive and divergent, non-syntenic alignments were frequently observed when comparing sequence assemblies with whole-genome aligners (Sup

Fig3). To remove the background alignment noises caused by transposable elements in genome alignment output, we aimed to identify the syntenic aligned segments (anchors) in the optimal chain through implementing the Longest Increasing Subsequence (LIS) problem (Rani and Rajpoot 2016; Abouelhoda and Ohlebusch 2005) (Figure S4.2 and S4.3).

Our workflow for SV identification is consisted of three phases: 1) perform pairwise whole-genome alignment, 2) chain aligned segments with LIS, 3) characterize structural variants through identifying alignment gaps. For the second phase, we performed independent chaining for anchors located in syntenic aligned regions and rearranged segments, as rearrangements (inversions and translocations) break colinearity and form a separate chain.

Alignment

Pairwise genome alignments were carried out with minimap2 (Li 2018) (v2.17) using parameters: -c -cx asm5 --no-kalloc --print-qname --cs=long. Prior to chaining, alignments were sorted according to the position in the reference sequence.

Chaining

A two-round chaining procedure was implemented to identify the longest set of anchors, where the first round identifies the optimal chain and the second round finds lower-scoring anchors to fill the gaps in the first chain (Figure S4.2). During each round, we calculate the chaining score for individual anchors, and identify non-overlapping anchors in the global optimal path using the backtracking approach. The computation of chaining score differs between these two rounds, as the second one was carried out aiming to incorporate anchors of low mapping quality.

The chaining score of anchor i in the first round was calculated as: $f(i) = max\{f(j) + len(i) * log10(q(i) + 0.001) - gap(i,j)/100\}, i > j > 1$, where len(i) and q(i) are

respectively the length and the mapping quality of anchor i. Gap (i,j) is the distance between anchors i and j, which was computed as $abs\ (ix - jy)$, and x and y are the start and end coordinates. Upon score calculation, the backtracking method was used by repeatedly finding the best predecessor of anchor i. After the first round, anchors identified in the optimal chain were combined with the remaining anchors with a size larger than 15Kb and were subjected to round two. To eliminate the effect of mapping quality, the score calculation of anchor i was modified from step1 as: $f(i) = max\{f(j) + \ len(i) - gap(i,j)/100\}, i > j \ > 1$.

Structural Variant characterization

Independent transposable elements insertion and indels mediated by this process result in a pairwise unaligned pattern among maize lines. This special variant structure could not be characterized by softwares developed to score small variants or variants with simple junctions. To accurately characterize structural changes in both reference and query genomes among maize lines, we defined variants as pairwise unaligned regions (Figure S4.1B). For each pair of unaligned regions, the region in reference is a deletion, and its counterpart in query is an insertion (Figure S4.1B). As to rearrangements, breakpoints for inversions, translocations and tandem duplications were inferred from alignment chains and orientation. True variants were further filtered with a 20Kb cutoff for inversions, 10Kb for tandem duplications, and 50Kb for translocations.

Variant identification among NAM lines

To identify structural changes among 26 NAM lines, we performed 325 pairwise alignments for each chromosome and carried out chaining and SV characterization with the workflow described above. Chaining was conducted with script “chaining.py” and SV calling

was accomplished with “sv_detect.py” (see WholeGenome-SV section in github). The number and size of variants, including un-alignments, tandem duplications, and inversions, were quantified for individual genomes and plotted with custom script using karyploteR.

Pan-genome analysis

Frequency of B73 genome space in NAM population

The frequency distribution of B73 genomic sequences was calculated through quantifying the presence/absence of every loci among 25 NAM lines. The start and end coordinates of each unit are intervals between adjacent alignment breakpoints of B73. For each chromosome, the SV breakpoints were extracted and sorted by position, and adjacent breakpoints smaller than 20bp were merged as their midpoint. Intervals between breakpoints were derived, and the occurrences of each interval were counted across NAM based on genome alignment. B73 segments present in 25 lines were represented by an allele frequency of 26, and an allele frequency of 1 depicts B73-specific regions. The above steps were conducted with script allele_frequency_cal.py. Genes and UMRs that overlap with each interval were identified with bedtools intersect, and subsequently quantified for every allele frequency.

Pangenome space

We employed the all-by-all syntenic alignments among NAM lines to calculate the pan-genome space. The added non-redundant genome size was calculated upon the addition of each genome, which was subsequently used as the reference to investigate the expanded genomic space. Aligned segments between the *n*th genome and all its predecessor genomes (*n-1*) were merged, and unaligned segments of the *n*th genome were derived. The unaligned parts of each additional line are the novel regions added to the pan-genome space. Order of NAM lines was

shuffled for a 1000 times, and pan-genome was calculated for every case. The pipeline for pangenome computation and permutation was implemented in script pangenome_cal.py.

Structural comparison among repeat arrays

To measure the genetic distance of syntenic repeat arrays among NAM inbreds, we employed the dot-matrix method to perform pairwise sequence alignment for repeat arrays, and calculated pairwise distance based on the number of monomer matches between reference and query. This pipeline was carried out for the structural comparison of ten CentC arrays, ten classical knobs and one nucleus organizer region (NOR) across 26 NAM lines. As large knobs are intermingled with knob180 and TR-1 monomers (Liu et al. 2020; M. B. Hufford, Seetharam, and Woodhouse 2021), we performed alignment with the dominant monomer type in the array.

Syntenic repeat arrays identification

The coordinates and of syntenic knobs, CentC, and NOR arrays were obtained from study (M. B. Hufford, Seetharam, and Woodhouse 2021). We identified CentC arrays located within 5Mb upstream and downstream of the active centromeres for each chromosome as true centromeric arrays. Classical knobs located on 2L, 3L, 4L, 5L, 6S1, 6S2, 6L, 7L, 8L, 9S were selected for structure analysis. The nucleus organizer regions (NOR) are present in syntenic areas on the short arm of chromosome 6 across all lines.

Pairwise alignment via dot-matrix

As minimap2 failed to align tandem repetitive areas, we employed the dot-matrix approach to perform pairwise alignments between repeat arrays (Gibbs and McIntyre 1970). Traditional dot-matrix method compares two sequences through identifying nucleotide or amino acid matches on the main diagonal. In our pipeline, repeat arrays from reference and query were

regarded as two sequences, where each repeat monomer was analyzed as a single residue. To identify the monomer pairs that share a common ancestor, we aligned all monomers from the reference array to those from the query array and measured their genetic distance. A match was assigned to a monomer pair when their similarity exceeds a certain threshold, and a dot was placed in the matrix. Structural similarity between the two repeat arrays was evaluated through manually inspecting the main diagonal in the dot-matrix.

To construct the dot matrix, monomer indexes from reference and query arrays were written along the two axes, where *n* represents the *n*th monomer from each array. Sequences of indexed monomers were extracted with bedtools getfasta (v2.29.2; -nameOnly -s). Genetic distance between any monomer pairs was measured through all-by-all alignments (*i* x *j*) with BLAT (vxx;-minIdentity=70 -maxGap=10 -minScore=0 -repMatch=2147483647). The similarity score for each monomer alignment was calculated with Jaccard Index: $Len(A,B) \,/\, \{Len(A) + Len(B) - Len(A,B)\}$, where $Len(A)$ and $Len(B)$represent the lengths of monomers A and B, and $Len(A,B)$is the number of matched nucleotides between them. Monomer pairs with a jaccard index above 0.98 were classified as matches and marked in the matrix. Dot matrices were plotted with R and structural similarity between any two repeat arrays from NAM lines was manually evaluated.

Similarity calculation and clustering

To assess the overall similarity between two repeat arrays in a quantitative way, we measured the total number of monomer matches for each alignment, and normalized it against the length of the smaller array to account for the difference in array length. The similarity of each pair of repeat arrays among NAM was calculated and used as input to construct a correlation matrix. This correlation matrix was visualized as a network through qgraph in R.

Divergence-time estimation with the whole-genome alignment method

Whole chromosome and intergenic spaces

As the syntenic anchors represent the true common ancestry between reference and query genomes, SNPs located in these syntenic regions could be used to accurately infer the divergence time for individual aligned segments. The syntenic aligned segments in a minimap2 paf format were identified based on coordinates derived from the synteny identification step, and were subjected to variant calling with paftools.js call (v2.17) using parameters: -L50 -q0 -l50. To obtain a more accurate estimation of divergence time of intergenic spaces, we filtered SNPs located in genes and unmethylated regions with bedtools intersect (v2.29.2), and computed the SNP ratio of each block by counting the alignment length and SNP number for intergenic regions. Divergence time for each aligned segment was estimated with a molecular clock of $3.3X10^{-8}$.

Knobs

As knobs are interspersed with a high variety of transposable elements (Liu et al. 2020; M. B. Hufford, Seetharam, and Woodhouse 2021), a great portion of knob sequence could be uniquely aligned between genomes and thus high-confidence syntenys were detected within knobs. Inspection of alignment blocks suggested that a certain amount of syntenic monomer sequences were included in the alignment chain and SNPs were called within these tandem repeats. To eliminate the effect of incorrect chaining of repeat sequences as well as the differential mutation rate between monomers and transposable elements on divergence calculation, we removed SNPs in TR-1 and knob180 repeat sequences with bedtools intersect

(v2.29.2) and computed the divergence time of each aligned segment with SNPs located in non-tandem repetitive areas.

Divergence-time estimation with short-read mapping

Source of data

Paired-end Illumina data of 49 *parviglumis* lines from Palmar Chico in Balsas river drainage of Mexico were obtained from Bioproject PRJNA616247 (SRR11448786-SRR11448838). Illumina reads of 14 *parviglumis* lines TIL01 (SRR447882), TIL02 (SRR447886), TIL03 (SRR447894-SRR447895), TIL04 (SRR447962-SRR447964), TIL05 (SRR447755-SRR447757),TIL06 (SRR447827-SRR447829), TIL07(SRR447960-SRR447961), TIL09 (SRR447954-SRR447955), TIL10 (SRR447825-SRR447826), TIL11 (SRR5976511), TIL12 (SRR447997), TIL14 (SRR447780-SRR447782), TIL15 (SRR447859-SRR447860) and TIL17 (SRR447896-SRR447898) were from HapMap II project SRP011907. Paired-end reads of two *mexicana* lines, TIL08 (SRR447933-SRR447934) and TIL25 (SRR447936-SRR5976310), were obtained from study SRP011907, and the data of other *mexicana* lines (SRR7758236 and SRR7758237) were downloaded from project PRJNA487810. Reads for *Huehuetageneis* samples Hue2 and Hue4 were downloaded from PRJNA384363, and that of diploperennis (SRR13687522) is from project PRJNA700589. Paired-end data for two *tripsacum* lines TDD39103 (SRR447804-SRR447807) and Trip_ISU_1 (SRR7758238) were respectively from SRP011907 and PRJNA487810. Short-reads of NAM lines were obtained from PRJEB31061 and PRJEB32225.

SNP calling

Illumina reads were trimmed with trimgalore (v0.6.5) and aligned to B73 reference genomes with bwa-mem (v0.7.17). Variant calling was conducted on bam files with a mapping quality above 20 using bcftools mpileup (v1.6; -Ou -f -C50). To remove the artifact of read mapping on variant calling, we filtered the sites with too low or too high read depth with bcftools filter, where the lower and upper bounds were respectively defined as ¼ and 4 times of the mean read depth of input bam files with MAPQ>20. High-confidence calls were obtained by further applying a quality cutoff of 20, and homozygous SNPs were extracted with bcftools view.

Divergence calculation and normalization

To estimate the divergence of each line from the reference with short reads, we calculated the genetic distance between each sample and reference in a fixed window. The reference genomes were divided into 20Kb non-overlapping windows with bedtools (v2.29.2). Genetic distance was measured as the proportion of intergenic SNPs over effective SNPable lengths in each window. We applied the same coverage cutoff used for SNP calling to estimate effective length, which is between ¼ and 4 times of the mean read depth. The portion of SNPable segments that overlap with genes and UMRs were removed with bedtools (v2.29.2) to account for intergenic regions. Divergence time over each window was estimated with $d/2/u$, where u = $3.3x10^{-8}$.

Upon divergence calculation in each 20Kb window, we compared the values estimated by short-reads with those by whole-genome alignment among NAM lines (Table S4.1 and S4.2. We sampled 10,000 points from the whole dataset, and investigated the relationship between divergence time and biases of the short-read method (Figure S4.13A, S4.13B). In the cases where divergence time is younger than 0.4 million years, a small variation was identified between values estimated with the two methods (Figure S4.13A, S4.13B). However, the short-

read alignment approach is not sensitive enough to capture SNPs in highly divergent regions and often underestimate the divergence time when it's ancient (Supplementary Fig 13A, B). To estimate the introgression time of the 3Mb segment on chromosome 6 (17.8-20.96Mb), we calculated the mean divergence time of the target area between two tripsacum samples and B73 using SNPs called by short-reads (Figure S4.13C). This value was later normalized against the median divergence fold difference between the short-read method and whole-genome alignment method to accommodate biases of the Illumina read-mapping approach.

MSMC (Multiple Sequentially Markovian Coalescent) analysis

To infer the dynamics of maize effective population size over the past million years, we employed the MSMC method to analyze syntenic SNPs between 25 NAM lines and B73. As low residual heterozygosity was identified among NAM inbreds, one haplotype was used for each line for MSMC analysis. To generate input files for msmc2 (Schiffels and Durbin 2014; Schiffels and Wang 2020), SNPs in VCF format across 25 lines were merged with bcftools (v1.6; http://samtools.github.io/bcftools) and phased with BEAGLE (Browning, Zhou, and Browning, n.d.). Syntenic aligned regions between NAM and B73 were used as mappability masks to define high-confidence mapping. Phased haplotypes and mask files for 25 lines were concatenated into a single file with generate_multihetsep.py, and used as input for msmc2 (Schiffels and Wang 2020). The time segment patterning parameter was set as 5*4+25*2+5*4 for msmc2 analysis.

UMR and differential gene expression

We evaluated the effect of UMR divergence on gene expression between each NAM and B73. UMR coordinates for 26 lines were obtained from MaizeGDB

(https://maizegdb.org/download). For each line, UMRs were classified into syntenic UMRs and neoUMRs based on their genetic background compared with B73. UMRs located on syntenic aligned segments were classified as syntenic UMRs, and the ones completely overlapped with query-specific sequences were categorized as neoUMRs. UMRs syntenic to B73 were further divided into groups, including 0-50K, 50-250K, 250-500K, 500K+, according to the divergence time of each syntenic segment they overlap with using bedtools intersect (v2.29.2; -f 0.5).

We identified the genes that individual UMRs regulate across NAM lines based on the condition that -10bp to +400bp from gene transcription start site overlap with UMR. Annotated genes of each NAM line were classified according to their corresponding UMR group, and selected for differential expression analysis. The pan-gene matrix was employed to correlate annotated geneIDs from 26 lines for synteny identification, where TPM values were scored with TPMCalculator (v0.0.4). Differential expression value for each syntenic gene pair between NAM and B73 was calculated as log2FC (fold change). The distribution of differential gene expression values among the 5 groups described above were compared.

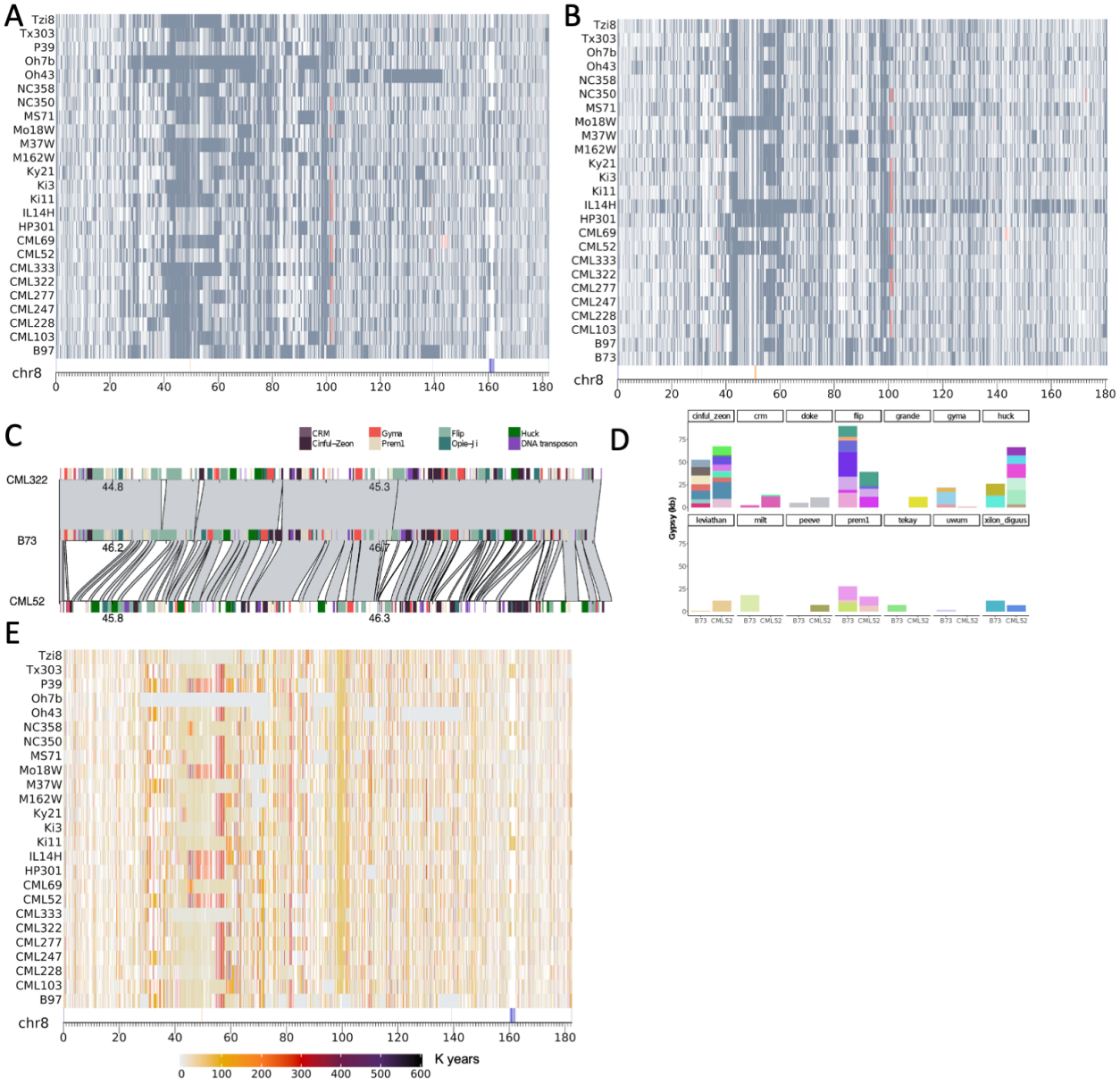


Figure 4.1. Diverse haplotype blocks in pericentromeric regions of chromosome 8. **A)** Whole-chromosome alignments between NAM lines and B73. Syntenic aligned regions and inverted segments are respectively colored as grey and red. Tandem repeats and CENH3-defined centromeres are displayed in the bottom track, where centromeres are marked by a yellow box, and CentC, Knob180, TR-1, NOR, and subtelomere repeats are represented as orange, blue, red, cyan, and black lines. **B)** Whole-chromosome alignments between 25 lines and P39. Annotation is the same as A). **C)** Pairwise alignments between CML322, CML52 and B73 from region 46.1Mb to 47.1 Mb on chromosome 8 (B73 coordinates). **D)** Comparison of transposable element content between B73 and CML52 over the 1Mb region in C). The abundance of

individual TE types (Gypsy superfamily) was shown in a stacked bar plot. Transposons classified as the same subfamilies are colored the same. **E)** Divergence time (K years) between NAM lines and B73, estimated with syntenic SNPs. Maximum divergence time was set as 0.6 million years in the heatmap, where the lowest divergence colored as light grey and highest divergence as black. White spaces depict unaligned regions between query genome and B73. Tandem repeats are shown in the bottom track and annotated as A).

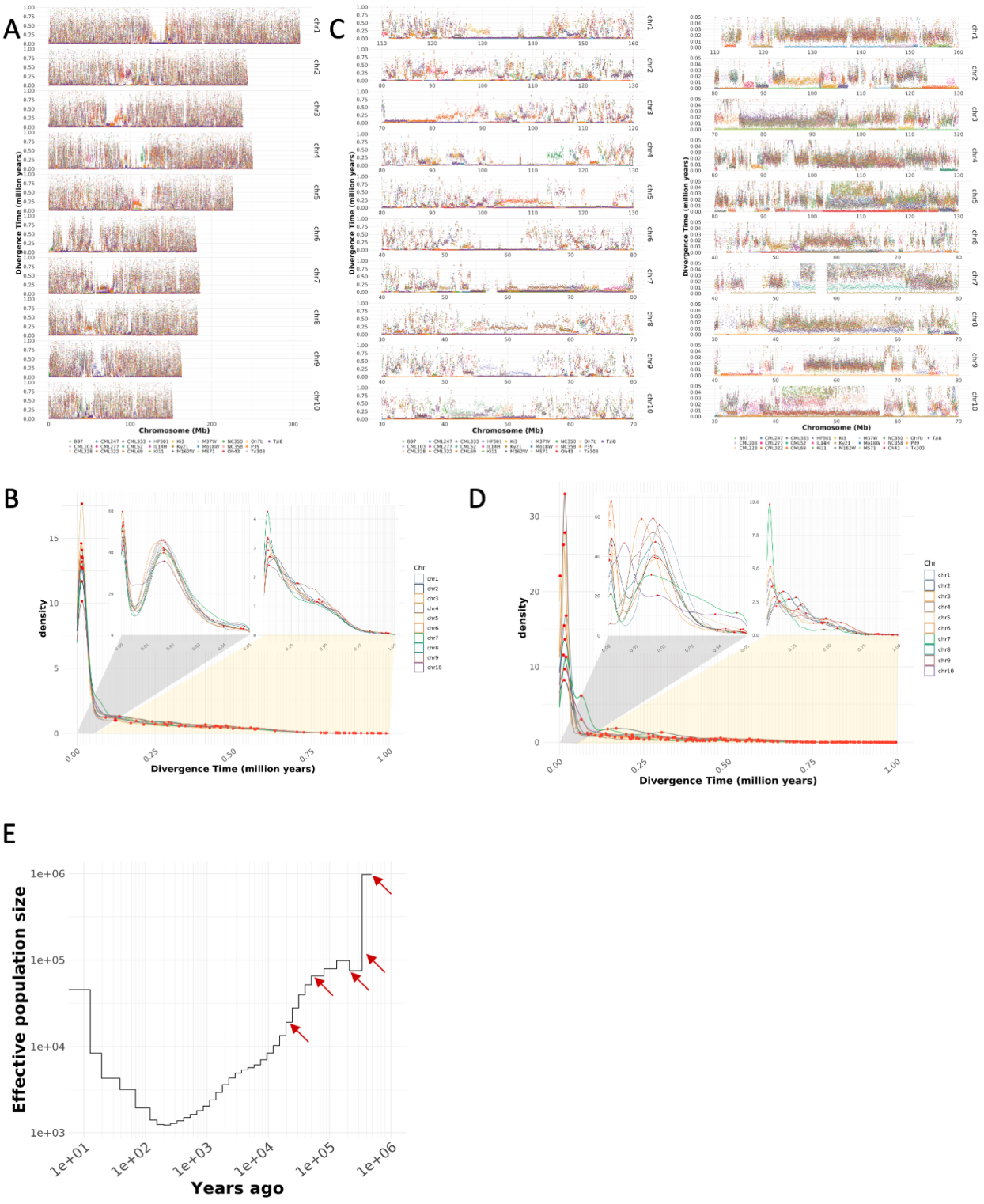


Figure 4.2. Repeated isolation and introgression over the past 0.5 million years inferred from divergence time. **A)** Divergence time between NAM and B73 over a 20Kb window across the whole genome. **B)** Density plot of divergence times displayed in A), with the left embedded

panel depicting the distribution between 0 to 0.05 million years and the right embedded panel representing that of 0.05 to 1 million years. Local maximums in density plot are highlighted as red dots, which were calculated with stat_peaks in R. **C)** Divergence time between NAM and B73 over pericentromeric areas. The left panel shows the range from 0 - 1 million years, and the right panel highlights divergence time between 0 to 0.05 million years. **D)** Density plot of divergence times displayed in the left panel of C). The insets were annotated as B). **E)** Effective population size of maize over the past 0.5 million years, estimated with syntenic SNPs via MSMC analysis. The arrows are labeling local peaks found in B) and D), respectively at 0.015, 0.05, 0.1, 0.25, and 0.5 million years.

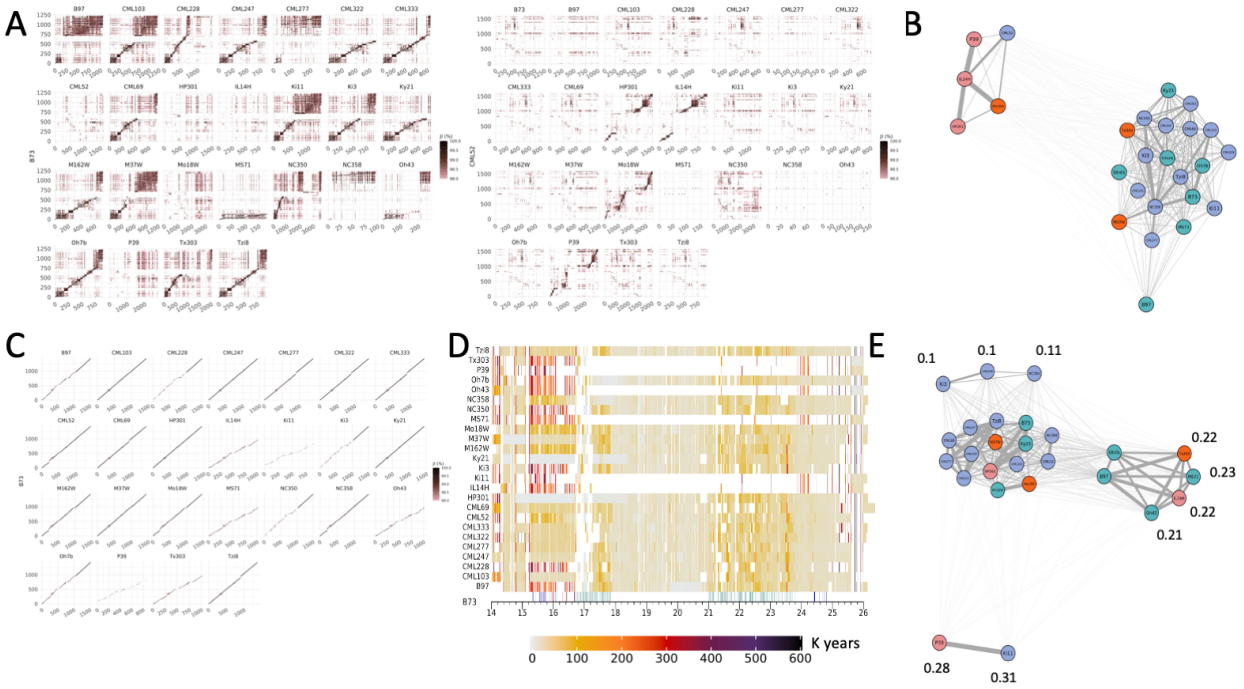


Figure 4.3. Haplotypes and evolutionary strata in repeat arrays. **A)** Pairwise alignments between 25 lines and reference genome (left: B73 and right: CML52) over the CentC repeat arrays on chromosome 8. Each alignment is represented in a dotplot, where the x and y axes depict the monomer indexes of query and reference arrays. Each dot reflects the similarity (Jaccard Index) between a pair of monomers, where jaccard similarity values higher than 0.98 are regarded as matches and shown in a gradient format. **B)** Clustering of chr8 CentC arrays across 26 lines based on all-by-all alignment in A). Flint, temperate, mixed, and tropical lines are respectively highlighted in pink, blue, red and green. **C)** Pairwise alignment between 25 NAM lines and B73 over the 6S knob180, which is adjacent to NOR (chr6:18-21 Mb). **D)** Alignment and divergence time estimation over the 6S knob and NOR region. Tandem repeats are annotated as Figure 1A. **E)** Clustering of NAM lines based on all-by-all alignment in C). Divergence time (million years) estimated is labeled for each group.

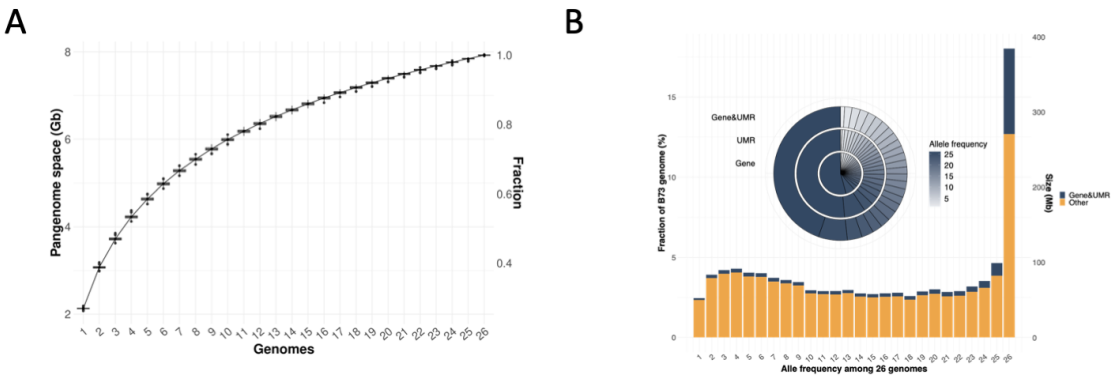


Figure 4.4. Pangenome size and allele frequency of B73 segments across 26 lines. **A)** Pangenome space with each additional line. The error bar is derived from 1000 times shuffling of the order for input genomes. **B)** Allele frequency of B73 segments among 26 lines. Genes and UMRs highlighted in the bar plot (blue) were summarized in the pie chart, where tracks from inner to outer circle represent genes, UMRs, and total space of gene and UMRs.

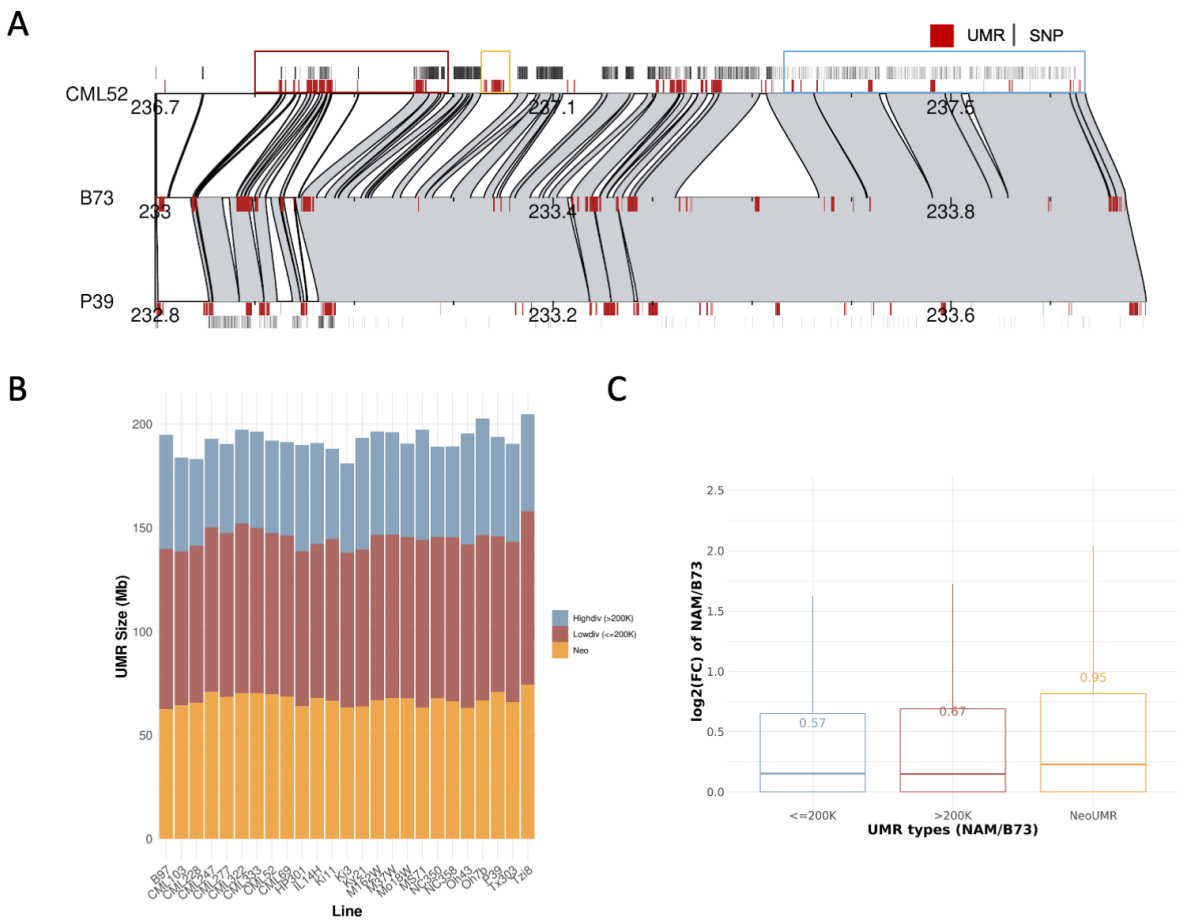


Figure 4.5. Gene regulatory diversity and impacts of divergent UMRs on gene expression. **A)** Example of UMR diversity between CML52, P39 and B73 (Chr2: 233-234 Mb). UMRs for each line are depicted in red boxes and SNPs in CML52 and P39 are shown in black lines. Examples of NeoUMRs, highly divergent UMRs and lowly divergent UMRs in CML52 are respectively highlighted with yellow, red and blue boxes. In certain cases, UMRs are splitted into small segments due to transposon insertion (e.g., second UMR cluster in B73). **B)** Distribution of UMRs divergence among NAM lines compared with B73. **C)** Gene expression affected by UMRs of different diversification level.

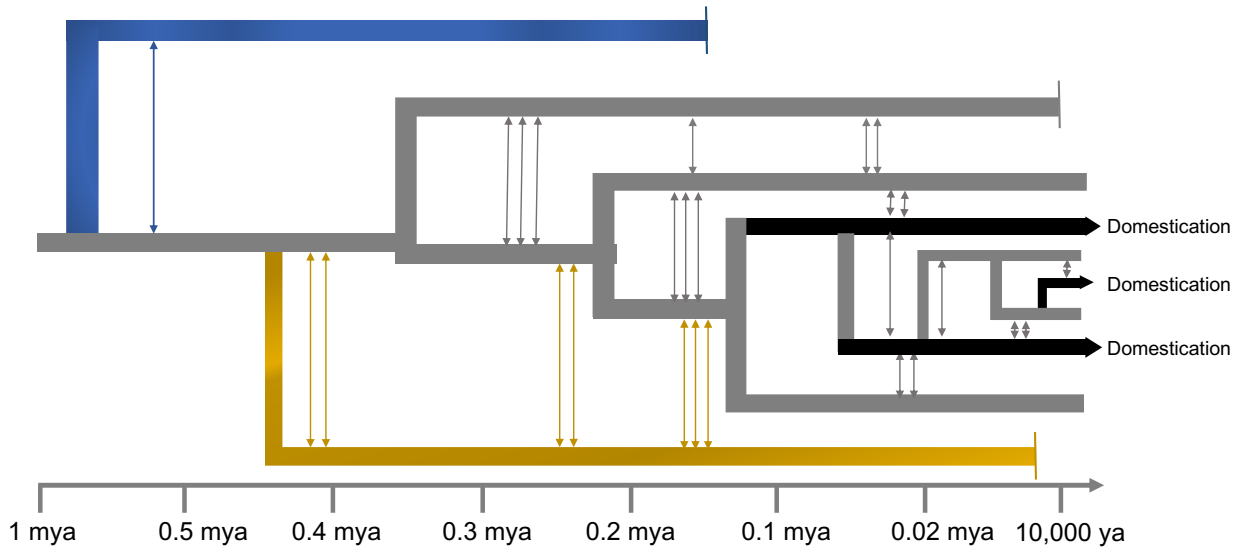


Figure 4.6. Schematic model for pangenome accumulation of maize over the past 1 million years, through recurrent isolation and introgression. Introgression among archaic populations were represented by arrows and teosinte groups that underwent domestication were highlighted in black.

References

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, et al. 2013. "Hybridization and Speciation." *Journal of Evolutionary Biology* 26 (2): 229–46.

Abouelhoda, Mohamed Ibrahim, and Enno Ohlebusch. 2005. "Chaining Algorithms for Multiple Genome Comparison." *Journal of Discrete Algorithms* 3 (2): 321–41.

Albert, P. S., Z. Gao, T. V. Danilova, and J. A. Birchler. 2010. "Diversity of Chromosomal Karyotypes in Maize and Its Relatives." *Cytogenetic and Genome Research* 129 (1-3): 6–16.

Almarri, Mohamed A., Anders Bergström, Javier Prado-Martinez, Fengtang Yang, Beiyuan Fu, Alistair S. Dunham, Yuan Chen, Matthew E. Hurles, Chris Tyler-Smith, and Yali Xue. 2020. "Population Structure, Stratification, and Introgression of Human Structural Variation." *Cell* 182 (1): 189–99.e15.

Anderson, Sarah N., Michelle C. Stitzer, Alex B. Brohammer, Peng Zhou, Jaclyn M. Noshay, Christine H. O'Connor, Cory D. Hirsch, Jeffrey Ross-Ibarra, Candice N. Hirsch, and Nathan M. Springer. 2019. "Transposable Elements Contribute to Dynamic Genome Content in Maize." *The Plant Journal: For Cell and Molecular Biology* 100 (5): 1052–65.

Arnold, Michael L. 2004. "Transfer and Origin of Adaptations through Natural Hybridization: Were Anderson and Stebbins Right?" *The Plant Cell* 16 (3): 562–70.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. n.d. "A One Penny Imputed Genome from next Generation Reference Panels." https://doi.org/10.1101/357806.

Brunner, Stephan, Kevin Fengler, Michele Morgante, Scott Tingey, and Antoni Rafalski. 2005. "Evolution of DNA Sequence Nonhomologies among Maize Inbreds." *The Plant Cell* 17 (2): 343–60.

Charlesworth, Deborah, and John H. Willis. 2009. "The Genetics of Inbreeding Depression." *Nature Reviews. Genetics* 10 (11): 783–96.

Chia, Jer-Ming, Chi Song, Peter J. Bradbury, Denise Costich, Natalia de Leon, John Doebley, Robert J. Elshire, et al. 2012. "Maize HapMap2 Identifies Extant Variation from a Genome in Flux." *Nature Genetics* 44 (7): 803–7.

Ching, Ada, Katherine S. Caldwell, Mark Jung, Maurine Dolan, Oscar S. Smith, Scott Tingey, Michele Morgante, and Antoni J. Rafalski. 2002. "SNP Frequency, Haplotype Structure and Linkage Disequilibrium in Elite Maize Inbred Lines." *BMC Genetics* 3 (October): 19.

Conrad, Donald F., and Matthew E. Hurles. 2007. “The Population Genetics of Structural Variation.” *Nature Genetics* 39 (7): S30–36.

Dawe, R. Kelly, Elizabeth G. Lowry, Jonathan I. Gent, Michelle C. Stitzer, Kyle W. Swentowsky, David M. Higgins, Jeffrey Ross-Ibarra, et al. 2018. “A Kinesin-14 Motor Activates Neocentromeres to Promote Meiotic Drive in Maize.” *Cell* 173 (4): 839–50.e18.

Doebley, J., and Others. 1990. “Molecular Systematics of Zea (Gramineae).” *Maydica* 35 (2): 143–50.

Feschotte, Cédric. 2008. “Transposable Elements and the Evolution of Regulatory Networks.” *Nature Reviews. Genetics* 9 (5): 397–405.

Fu, Huihua, and Hugo K. Dooner. 2002. “Intraspecific Violation of Genetic Colinearity and Its Implications in Maize.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (14): 9573–78.

Ghaffari, Rashin, Ethalinda K. S. Cannon, Lisa B. Kanizay, Carolyn J. Lawrence, and R. Kelly Dawe. 2013. “Maize Chromosomal Knobs Are Located in Gene-Dense Areas and Suppress Local Recombination.” *Chromosoma* 122 (1-2): 67–75.

Gibbs, A. J., and G. A. McIntyre. 1970. “The Diagram, a Method for Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences.” *European Journal of Biochemistry / FEBS* 16 (1): 1–11.

Goloubinoff, P., S. Pääbo, and A. C. Wilson. 1993. “Evolution of Maize Inferred from Sequence Diversity of an Adh2 Gene Segment from Archaeological Specimens.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (5): 1997–2001.

Gore, Michael A., Jer-Ming Chia, Robert J. Elshire, Qi Sun, Elhan S. Ersoz, Bonnie L. Hurwitz, Jason A. Peiffer, et al. 2009. “A First-Generation Haplotype Map of Maize.” *Science* 326 (5956): 1115–17.

Haberer, Georg, Nadia Kamal, Eva Bauer, Heidrun Gundlach, Iris Fischer, Michael A. Seidel, Manuel Spannagl, et al. 2020. “European Maize Genomes Highlight Intraspecies Variation in Repeat and Gene Content.” *Nature Genetics* 52 (9): 950–57.

Hall, D. W., and R. K. Dawe. 2017. “Modeling the Evolution of Female Meiotic Drive in Maize. G3: Genes, Genomes.” *Genetics DOI* 10: g3.

Harrison, Richard G., and Erica L. Larson. 2014. “Hybridization, Introgression, and the Nature of Species Boundaries.” *The Journal of Heredity* 105 Suppl 1: 795–809.

Heerwaarden, Joost van, John Doebley, William H. Briggs, Jeffrey C. Glaubitz, Major M. Goodman, Jose de Jesus Sanchez Gonzalez, and Jeffrey Ross-Ibarra. 2011. "Genetic Signals of Origin, Spread, and Introgression in a Large Sample of Maize Landraces." *Proceedings of the National Academy of Sciences of the United States of America* 108 (3): 1088–92.

Hufford, Matthew B., Pesach Lubinksy, Tanja Pyhäjärvi, Michael T. Devengenzo, Norman C. Ellstrand, and Jeffrey Ross-Ibarra. 2013. "The Genomic Signature of Crop-Wild Introgression in Maize." *PLoS Genetics* 9 (5): e1003477.

Hufford, M. B., A. S. Seetharam, and M. R. Woodhouse. 2021. "De Novo Assembly, Annotation, and Comparative Analysis of 26 Diverse Maize Genomes." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.01.14.426684v1.abstract.

Hu, Ying, Vincent Colantonio, Bárbara S. F. Müller, Kristen A. Leach, Adalena Nanni, Christina Finegan, Bo Wang, et al. 2021. "Genome Assembly and Population Genomic Analysis Provide Insights into the Evolution of Modern Sweet Corn." *Nature Communications* 12 (1): 1227.

Iltis, H. H., J. F. Doebley, R. G. M, and B. Pazy. 1979. "Zea Diploperennis (Gramineae): A New Teosinte from Mexico." *Science* 203 (4376): 186–88.

Joron, Mathieu, Lise Frezal, Robert T. Jones, Nicola L. Chamberlain, Siu F. Lee, Christoph R. Haag, Annabel Whibley, et al. 2011. "Chromosomal Rearrangements Maintain a Polymorphic Supergene Controlling Butterfly Mimicry." *Nature* 477 (7363): 203–6.

Kirkpatrick, Mark, and Nick Barton. 2006. "Chromosome Inversions, Local Adaptation and Speciation." *Genetics* 173 (1): 419–34.

Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, and Richard Durbin. 2011. "Inference of Human Population History from Individual Whole-Genome Sequences." *Nature* 475 (7357): 493–96.

Liu, Jianing, Arun S. Seetharam, Kapeel Chougule, Shujun Ou, Kyle W. Swentowsky, Jonathan I. Gent, Victor Llaca, et al. 2020. "Gapless Assembly of Maize Chromosomes Using Long-Read Technologies." *Genome Biology* 21 (1): 121.

Mallet, James, Nora Besansky, and Matthew W. Hahn. 2016. "How Reticulated Are Species?" *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 38 (2): 140–49.

Martinez, Ana Karine, Jose Miguel Soriano, Roberto Tuberosa, Rachil Koumproglou, Torben Jahrmann, and Silvio Salvi. 2016. “Yield QTLome Distribution Correlates with Gene Density in Maize.” *Plant Science: An International Journal of Experimental Plant Biology* 242 (January): 300–309.

Matsuoka, Yoshihiro, Yves Vigouroux, Major M. Goodman, Jesus Sanchez G, Edward Buckler, and John Doebley. 2002. “A Single Domestication for Maize Shown by Multilocus Microsatellite Genotyping.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (9): 6080–84.

McMullen, Michael D., Stephen Kresovich, Hector Sanchez Villeda, Peter Bradbury, Huihui Li, Qi Sun, Sherry Flint-Garcia, et al. 2009. “Genetic Properties of the Maize Nested Association Mapping Population.” *Science* 325 (5941): 737–40.

Mérot, Claire, Rebekah A. Oomen, Anna Tigano, and Maren Wellenreuther. 2020. “A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation.” *Trends in Ecology & Evolution* 35 (7): 561–72.

Morgante, Michele, Stephan Brunner, Giorgio Pea, Kevin Fengler, Andrea Zuccolo, and Antoni Rafalski. 2005. “Gene Duplication and Exon Shuffling by Helitron-like Transposons Generate Intraspecies Diversity in Maize.” *Nature Genetics* 37 (9): 997–1002.

Mount, David W. 2007. “Dot Matrix Pairwise Sequence Comparison.” *CSH Protocols* 2007 (December): db.top31.

Oka, Rurika, Johan Zicola, Blaise Weber, Sarah N. Anderson, Charlie Hodgman, Jonathan I. Gent, Jan-Jaap Wesselink, et al. 2017. “Genome-Wide Mapping of Transcriptional Enhancer Candidates Using DNA and Chromatin Features in Maize.” *Genome Biology* 18 (1): 137.

Purcell, Jessica, Alan Brelsford, Yannick Wurm, Nicolas Perrin, and Michel Chapuisat. 2014. “Convergent Genetic Architecture Underlies Social Organization in Ants.” *Current Biology: CB* 24 (22): 2728–32.

Rani, Seema, and Dharmveer Singh Rajpoot. 2016. “LIS Using Backtracking and Branch-and-Bound Approaches.” *CSI Transactions on ICT* 4 (2): 87–93.

Ricci, William A., Zefu Lu, Lexiang Ji, Alexandre P. Marand, Christina L. Ethridge, Nathalie G. Murphy, Jaclyn M. Noshay, et al. 2019. “Widespread Long-Range Cis-Regulatory Elements in the Maize Genome.” *Nature Plants*. https://doi.org/10.1038/s41477-019-0547-0.

Rieseberg, Loren H., Olivier Raymond, David M. Rosenthal, Zhao Lai, Kevin Livingstone, Takuya Nakazato, Jennifer L. Durphy, Andrea E. Schwarzbach, Lisa A. Donovan, and Christian Lexer. 2003. "Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization." *Science* 301 (5637): 1211–16.

Ross-Ibarra, Jeffrey, Maud Tenaillon, and Brandon S. Gaut. 2009. "Historical Divergence and Gene Flow in the Genus Zea." *Genetics* 181 (4): 1399–1413.

Ross, Mark T., Darren V. Grafham, Alison J. Coffey, Steven Scherer, Kirsten McLay, Donna Muzny, Matthias Platzer, et al. 2005. "The DNA Sequence of the Human X Chromosome." *Nature* 434 (7031): 325–37.

Rowan, Beth A., Darren Heavens, Tatiana R. Feuerborn, Andrew J. Tock, Ian R. Henderson, and Detlef Weigel. 2019. "An Ultra High-Density Arabidopsis Thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features." *Genetics* 213 (3): 771–87.

Sanmiguel, Phillip, and Jeffrey L. Bennetzen. 1998. "Evidence That a Recent Increase in Maize Genome Size Was Caused by the Massive Amplification of Intergene Retrotransposons." *Annals of Botany* 82 (December): 37–44.

Schiffels, Stephan, and Richard Durbin. 2014. "Inferring Human Population Size and Separation History from Multiple Genome Sequences." *Nature Genetics* 46 (8): 919–25.

Schiffels, Stephan, and Ke Wang. 2020. "MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent." *Methods in Molecular Biology* 2090: 147–66.

Schneider, Kevin L., Zidian Xie, Thomas K. Wolfgruber, and Gernot G. Presting. 2016. "Inbreeding Drives Maize Centromere Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 113 (8): E987–96.

Shi, Jinghua, Sarah E. Wolf, John M. Burke, Gernot G. Presting, Jeffrey Ross-Ibarra, and R. Kelly Dawe. 2010. "Widespread Gene Conversion in Centromere Cores." *PLoS Biology* 8 (3): e1000327.

Springer, Nathan M., and Robert M. Stupar. 2007. "Allelic Variation and Heterosis in Maize: How Do Two Halves Make More than a Whole?" *Genome Research* 17 (3): 264–75.

Studer, Anthony, Qiong Zhao, Jeffrey Ross-Ibarra, and John Doebley. 2011. "Identification of a Functional Transposon Insertion in the Maize Domestication Gene tb1." *Nature Genetics* 43 (11): 1160–63.

Suarez-Gonzalez, Adriana, Christian Lexer, and Quentin C. B. Cronk. 2018. “Adaptive Introgression: A Plant Perspective.” *Biology Letters* 14 (3). https://doi.org/10.1098/rsbl.2017.0688.

Sun, Silong, Yingsi Zhou, Jian Chen, Junpeng Shi, Haiming Zhao, Hainan Zhao, Weibin Song, et al. 2018. “Extensive Intraspecific Gene Order and Gene Structural Variations between Mo17 and Other Maize Genomes.” *Nature Genetics* 50 (9): 1289–95.

Thiemann, Alexander, Junjie Fu, Felix Seifert, Robert T. Grant-Downton, Tobias A. Schrag, Heike Pospisil, Matthias Frisch, Albrecht E. Melchinger, and Stefan Scholten. 2014. “Genome-Wide Meta-Analysis of Maize Heterosis Reveals the Potential Role of Additive Gene Expression at Pericentromeric Loci.” *BMC Plant Biology* 14 (April): 88.

Thompson, M. J., and C. D. Jiggins. 2014. “Supergenes and Their Role in Evolution.” *Heredity* 113 (1): 1–8.

Todesco, Marco, Gregory L. Owens, Natalia Bercovich, Jean-Sébastien Légaré, Shaghayegh Soudi, Dylan O. Burge, Kaichi Huang, et al. 2020. “Massive Haplotypes Underlie Ecotypic Differentiation in Sunflowers.” *Nature* 584 (7822): 602–7.

Unterseer, Sandra, Saurabh D. Pophaly, Regina Peis, Peter Westermeier, Manfred Mayer, Michael A. Seidel, Georg Haberer, et al. 2016. “A Comprehensive Study of the Genomic Differentiation between Temperate Dent and Flint Maize.” *Genome Biology* 17 (1): 137.

Vigouroux, Yves, Jeffrey C. Glaubitz, Yoshihiro Matsuoka, Major M. Goodman, Jesús Sánchez G, and John Doebley. 2008. “Population Structure and Genetic Diversity of New World Maize Races Assessed by DNA Microsatellites.” *American Journal of Botany* 95 (10): 1240–53.

Wallace, Jason G., Peter J. Bradbury, Nengyi Zhang, Yves Gibon, Mark Stitt, and Edward S. Buckler. 2014. “Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize.” *PLoS Genetics* 10 (12): e1004845.

Wang, Li, Timothy M. Beissinger, Anne Lorant, Claudia Ross-Ibarra, Jeffrey Ross-Ibarra, and Matthew B. Hufford. 2017. “The Interplay of Demography and Selection during Maize Domestication and Expansion.” *Genome Biology* 18 (1): 215.

White, S. E., and J. F. Doebley. 1999. “The Molecular Evolution of Terminal ear1, a Regulatory Gene in the Genus Zea.” *Genetics* 153 (3): 1455–62.

CHAPTER 5

CONCLUSION AND DISCUSSION

Technology advancements over the past 10 years have greatly improved the power for structural variants characterization across genomes. The low cost and high-throughput of hybridization techniques and short-read resequencing provided us a valuable resource to study genomic variations at a large scale. Long-read technologies further improved the accuracy and sensitivity of short-read in variant identification, revealing hidden genomic variants in complicated areas (Alonge et al. 2020; Hufford, Seetharam, and Woodhouse 2021). Now with the availability of multiple high-quality assemblies for model organisms, we have transitioned from a single-reference phase to a pan-genome era, which profoundly broadened our understanding of genomic diversity and complexity (The Computational Pan-Genomics Consortium et al. 2016; Danilevicz et al. 2020).

Along with the improvement in technologies and resources during the past five years, we have performed genomic variant identification with multiple platforms, from short-reads to long-reads, from resequencing and mapping to a single reference to the multi-reference genome assemblies. Consistent with previous studies (Sudmant et al. 2015; Tattini, D'Aurizio, and Magi 2015), we found that while short reads are capable of capturing a great portion of variations, the nature of this data impeded its efficiency in complex genomes and difficult regions. Greater genome accessibility and higher accuracy have been achieved with long-reads, but variants in

complicated and highly repetitive regions can only be revealed with a combination of optical maps and long-read sequencing technologies.

With the availability of 26 maize whole-genome assemblies (Hufford, Seetharam, and Woodhouse 2021), we resolved the variant detection biases of resequencing technologies and achieved a full catalog of SVs across these lines, where we carried out all-by-all alignments and identified variants in a pairwise manner. Evolutionary dynamics of maize lines was inferred through integrating structural variants and SNPs. Functional impacts of variant regions were further explored with epigenetic profiling and gene expression analysis. The high assembly contiguity over tandem repeat arrays enabled structural comparison of centromeres, knobs and ribosomal DNA regions (NOR), of which internal structure was previously known only by cytological studies (Albert et al. 2010).

Future directions

Graph-based interpretations of pangenome data. While our approach for SV cataloging with whole-genome assemblies provides a confident set of variants, this method has the following limitations: 1) all-by-all alignment is computationally expensive, 2) SVs are not projected or concatenated to a single reference genome, 3) visualization and downstream analysis of SVs derived from all-by-all comparisons can be difficult. Theoretically, coordinate preservation and SV interpretation could be achieved by a pangenome sequence graph that encodes multiple genomes (Garrison et al. 2018; Danilevicz et al. 2020; Li, Feng, and Chu 2020). This approach represents full genomic diversity through retaining representative paths and collapsing redundant variants (Garrison et al. 2018; Danilevicz et al. 2020; Li, Feng, and Chu 2020). An alignment-based model was proposed by *Li et al* to construct pangenome graph

through iteratively adding poorly mapped regions to the reference genome (Li, Feng, and Chu 2020). This method employs progressive alignments and achieves high computation efficiency. Additionally, the final pangenome graph could be further used as a sequence reference genome for read mapping and variant calling. However, the current version does not keep sample information, and fails in complex SVs and diverse areas. The mapping algorithm needs to be reimplemented to work on highly repetitive and divergent genomes such as maize (Li, Feng, and Chu 2020). Other models such as the *De Bruijn* graph were also developed to construct pangenome graphs, yet not efficient enough to handle large numbers of genomes (Baier, Beller, and Ohlebusch 2015; Chikhi, Limasset, and Medvedev 2016; Li, Feng, and Chu 2020). Despite the above limitations, the pangenome graph approach is a promising strategy for future SV analysis, especially in cataloging, visualization, and downstream integration of various data formats.

The dark genome: repeat arrays. Comparative studies in tandem repetitive regions were recently enabled by the advancement of long read sequencing and development of highly contiguous assemblies (Miga et al. 2020; Liu et al. 2020). Our work on structural comparison over repeat arrays has revealed the substantial structural diversity in centromere, knobs, and ribosomal DNAs in maize, of which internal structure was previously known only by cytological studies (Albert et al. 2010).The above findings have helped us pave a way to investigate the evolutionary dynamics of repeat arrays, but our understanding of tandem repetitive regions is still preliminary. For example, previous studies have proposed a co-evolution model between centromeres and centromeric histones to explain the paradox where the rapid evolution of centromeric regions does not align with their stable function (Henikoff, Ahmad, and Malik 2001). Now with resources to study centromere repeat structural changes and centromeric

histone sequence evolution, we could resolve this mystery through identifying the evolutionary dynamics between these two factors. In addition, neocentromeres have been characterized in nonrepetitive areas, but its formation is poorly understood (Scott and Sullivan 2014). The *de novo* origin of neocentromeres might be attributed to the structural changes of ancestral centromeres, which can be explored via genomic comparisons. Furthermore, with the completion of multi-genome assemblies for single species, the homogenization, expansion, deletions of repeat arrays could be characterized for non-human organisms.

Interpreting the impact of SVs on phenotype. While third-generation sequencing and whole genome assemblies have facilitated the detection of structural variants, functional interpretation of SVs remains to be a challenge (Weischenfeldt et al. 2013). As SVs span genic and intergenic areas, they can cause phenotypic effects in various ways, from directly altering coding regions to indirectly influencing gene expression over a long distance (Ricci et al. 2019; Alonge et al. 2020). Impacts of SVs that disrupt coding sequences or involve in copy number changes could be estimated with the mRNA expression level of associated genes. As to SVs located in intergenic areas, they potentially interfere with regulatory elements and induce modifications in chromatin topology. These three-dimensional changes in genome structure can be characterized with Hi-C or Hi-ChIP, which associates chromatin looping with chromatin modification status (Mumbach et al. 2016; Spielmann, Lupiáñez, and Mundlos 2018). In addition to epigenetic regulation and gene expression, investigation on protein-level is crucial to verify the transcriptomic changes and characterize the linkage between genomic variants with phenotypes (Weischenfeldt et al. 2013). In summary, the integration of multi-omics provides a powerful resource for assessing the functional impacts of SVs on a large-scale (Salzberg 2019). Though a high-quality annotation in genes and cis-regulatory regions can be technically difficult,

improvements in sequencing technologies could help us overcome this challenge and lead to a better interpretation of the functional roles for SVs in the near future.

References

Albert, P. S., Z. Gao, T. V. Danilova, and J. A. Birchler. 2010. "Diversity of Chromosomal Karyotypes in Maize and Its Relatives." *Cytogenetic and Genome Research* 129 (1-3): 6–16.

Alonge, Michael, Xingang Wang, Matthias Benoit, Sebastian Soyk, Lara Pereira, Lei Zhang, Hamsini Suresh, et al. 2020. "Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato." *Cell* 182 (1): 145–61.e23.

Baier, Uwe, Timo Beller, and Enno Ohlebusch. 2015. "Graphical Pan-Genome Analysis with Compressed Suffix Trees and the Burrows–Wheeler Transform." *Bioinformatics* 32 (4): 497–504.

Chikhi, Rayan, Antoine Limasset, and Paul Medvedev. 2016. "Compacting de Bruijn Graphs from Sequencing Data Quickly and in Low Memory." *Bioinformatics* 32 (12): i201–8.

Danilevicz, Monica Furaste, Cassandria Geraldine Tay Fernandez, Jacob Ian Marsh, Philipp Emanuel Bayer, and David Edwards. 2020. "Plant Pangenomics: Approaches, Applications and Advancements." *Current Opinion in Plant Biology*. https://doi.org/10.1016/j.pbi.2019.12.005.

Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, et al. 2018. "Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference." *Nature Biotechnology* 36 (9): 875–79.

Henikoff, S., K. Ahmad, and H. S. Malik. 2001. "The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA." *Science* 293 (5532): 1098–1102.

Hufford, M. B., A. S. Seetharam, and M. R. Woodhouse. 2021. "De Novo Assembly, Annotation, and Comparative Analysis of 26 Diverse Maize Genomes." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.01.14.426684v1.abstract.

Li, Heng, Xiaowen Feng, and Chong Chu. 2020. "The Design and Construction of Reference Pangenome Graphs with Minigraph." *Genome Biology* 21 (1): 265.

Liu, Jianing, Arun S. Seetharam, Kapeel Chougule, Shujun Ou, Kyle W. Swentowsky, Jonathan I. Gent, Victor Llaca, et al. 2020. “Gapless Assembly of Maize Chromosomes Using Long-Read Technologies.” *Genome Biology* 21 (1): 121.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. “HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture.” *Nature Methods* 13 (11): 919–22.

Ricci, William A., Zefu Lu, Lexiang Ji, Alexandre P. Marand, Christina L. Ethridge, Nathalie G. Murphy, Jaclyn M. Noshay, et al. 2019. “Widespread Long-Range Cis-Regulatory Elements in the Maize Genome.” *Nature Plants*. https://doi.org/10.1038/s41477-019-0547-0.

Salzberg, Steven L. 2019. “Next-Generation Genome Annotation: We Still Struggle to Get It Right.” *Genome Biology* 20 (1): 92.

Scott, Kristin C., and Beth A. Sullivan. 2014. “Neocentromeres: A Place for Everything and Everything in Its Place.” *Trends in Genetics: TIG* 30 (2): 66–74.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. “Structural Variation in the 3D Genome.” *Nature Reviews. Genetics* 19 (7): 453–67.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.

Tattini, Lorenzo, Romina D’Aurizio, and Alberto Magi. 2015. “Detection of Genomic Structural Variants from Next-Generation Sequencing Data.” *Frontiers in Bioengineering and Biotechnology* 3 (June): 92.

The Computational Pan-Genomics Consortium, Tobias Marschall, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E. Dutilh, Ali Ghaffaari, et al. 2016. “Computational Pan-Genomics: Status, Promises and Challenges.” *Briefings in Bioinformatics* 19 (1): 118–35.

Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O. Korbel. 2013. “Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease.” *Nature Reviews. Genetics* 14 (2): 125–38.

APPENDIX A

SUPPLEMENTARY FIGURES AND TABLES – CHAPTER 2

Table S2.1. Copy number of lambda and co-bombarded plasmid in rice/maize transgenic genome

Sample name	Transgenic events	Mapped reads			Copy number†		Plant regenerated
		Genome	Lambda	Plasmid	Lambda	Plasmid	
Os λ-1	RiceE6*	7,668,253	110	242	0.23	4.20	Y
Os λ-2	RiceE10*	9,215,152	575	69	1.00	1.00	N
Os λ-3	RiceE24*	1,510,262	20	27	0.21	2.38	Y
Os λ-4	RiceE29*	200,676	527	42	42.57	27.89	Y
Os λ-5	RiceE37*	5,013,058	10,924	65	35.32	1.73	N
Os λ-6	RiceE17*	3,578,356	4,821	41	21.84	1.53	N
Os λ-7	RiceE26*	5,496,295	574	71	1.69	1.72	N
Os λ-8	RiceE16*	650,951	841	31	20.94	6.35	N
Os λ-9	RiceE20	7,174,824	90	81	0.20	1.50	N
Os λ-10	RiceE34	4,554,076	126	211	0.45	6.17	N
Os λ-11	RiceE18	5,016,781	213	578	0.69	15.35	N
Os λ-12	RiceE30	3,993,241	2,420	68	9.82	2.27	N
Os λ-13	RiceE38	3,386,957	2,709	39	12.96	1.53	N
Os λ-14	RiceE15	7,694,158	19,485	208	41.05	3.60	N
Zm λ-1	MaizeE14*	17,755,010	220	29	1.19	0.82	Y
Zm λ-2	MaizeE12*	11,246,844	2,397	68	20.42	3.03	Y
Zm λ-3	Maize E16*	17,098,938	1,703	13	9.54	0.38	Y
Zm λ-4	MaizeE4*	34,630,686	18,401	179	50.92	2.59	Y
Zm λ-5	MaizeE29	803,052	6	8	0.72	4.99	Y
Zm λ-6	MaizeE15	40,124,819	977	30	2.33	0.37	Y
Zm λ-7	MaizeE11	36,291,405	2,379	50	6.28	0.69	Y
Zm λ-8	MaizeE3	50,713,019	3,352	123	6.33	1.21	Y
Zm λ-9	MaizeE6	56,322,297	3,800	52	6.47	0.46	Y
Zm λ-10	MaizeE25	30,680,847	7,569	59	23.64	0.96	Y

† Copy number in the diploid rice/maize genome
* Lines selected for deep sequencing

Table S2.2. Sensitivity and precision evaluation of SV detection pipeline by simulation

		Simulation Set 1		Simulation Set 2		Simulation Set 3		Simulation Set 4	
Introduced fragments		Lambda:1-48502		Lambda:1-48502		Chr2:160,000-560,000		Chr3:3000,000-3500,000	
Translocation type		Inter-chromosomal	Intra-chromosomal	Inter-chromosomal	Intra-chromosomal	Inter-chromosomal	Intra-chromosomal	Inter-chromosomal	Intra-chromosomal
Number of breakpoints		20	10	100	40	484	180	100	46
Lumpy	**Total**	25	11	109	45	455	162	99	45
	TP†	18	10	73	32	306	120	71	32
	Sensitivity*	0.90	1	0.81	0.80	0.63	0.67	0.71	0.71
	Precision**	0.72	0.91	0.67	0.71	0.67	0.74	0.72	0.58
SVDetect	**Total**	22	11	127	40	699	389	129	66
	TP†	18	10	79	32	342	146	80	38
	Sensitivity*	0.90	1	0.88	0.80	0.71	0.81	0.80	0.84
	Precision**	0.82	0.91	0.62	0.80	0.49	0.38	0.62	0.58
Shared between Lumpy and SVDetect	**Total**	21	11	93	36	430	158	89	43
	TP†	18	10	70	31	301	120	71	32
	Sensitivity*	0.90	1	0.78	0.78	0.62	0.67	0.71	0.71
	Precision**	0.86	0.91	0.75	0.86	0.70	0.76	0.80	0.74

† True positives

* Sensitivity = Number of True positives/Number of total actual variants in simulation

** Precision= Number of True positives/Number of total calls by program

Table S2.3. Evidence of HDR in non-repetitive regions in rice transgenic events

Event	Fragment Size (bp)	Chromosome	Start Coordinate (bp)	End Coordinate (bp)	Copy number	Deletion (± 1 Mb)	HDR
Os λ-5	44	1	23,365,247	23,365,290	3	N	Y
	64	1	24,390,575	24,390,638	3	N	Y
	110	1	2,138,740	2,138,849	3	N	Y
	240	6	27,010,714	27,010,953	2	Y	N
	79	7	14,434,733	14,434,811	3	N	Y
	67	9	11,041,419	11,041,485	3	N	Y
	83	10	7,121,994	7,122,076	3	N	Y
	155	10	1,802,221	1,802,375	3	N	Y
Os λ-7	44	2	5,131,996	5,132,039	2	Y	N
	128	3	22,282,746	22,282,619	2	Y	N
	333	3	1,107,330	1,107,662	2	Y	N
	736	3	25,096,178	25,096,913	2	Y	N
	7,858	3	25,073,805	25,081,671	2	Y	N
	186	5	3,731,248	3,731,433	2	Y	N
	204	5	25,097,248	25,097,451	2	Y	N
	366	5	25,095,010	25,094,645	2	Y	N
	757	5	3,738,178	3,738,934	2	Y	N
	1,217	5	3,733,416	3,734,633	2	Y	N
	1,461	5	25,113,647	25,115,110	2	Y	N
	6,248	5	25,170,023	25,176,277	2	Y	N
	301	9	18,785,456	18,785,756	2	Y	N
	867	9	18,541,230	18,542,097	2	Y	N
	293	11	4,751,079	4,751,371	2	Y	N
Os λ-8	233	2	28,182,792	28,183,024	2	Y	N
	20,551	2	28,320,680	28,341,230	3	Y	N
	31,370	2	2,927,273	2,958,642	3	N	Y
	33,590	2	1,764,729	1,798,318	3	N	Y
	68,479	3	206,466	274,944	3	N	Y
	206	4	20,082,930	20,082,725	3	N	Y
	264	4	24,319,871	24,319,608	3	N	Y
	4,512	7	11,886,889	11,891,400	3	N	Y
	5,946	7	7,387,487	7,393,432	3	N	Y
	12,098	7	7,206,838	7,218,935	3	N	Y

Table S2.4. Evidence of HDR in non-repetitive regions in maize transgenic events

Event	Fragment Size (bp)	Chromosome	Start Coordinate (bp)	End Coordinate (bp)	Copy number	Deletion (± 1 Mb)	HDR
Os λ-5	44	1	23,365,247	23,365,290	3	N	Y
	64	1	24,390,575	24,390,638	3	N	Y
	110	1	2,138,740	2,138,849	3	N	Y
	240	6	27,010,714	27,010,953	2	Y	N
	79	7	14,434,733	14,434,811	3	N	Y
	67	9	11,041,419	11,041,485	3	N	Y
	83	10	7,121,994	7,122,076	3	N	Y
	155	10	1,802,221	1,802,375	3	N	Y
Os λ-7	44	2	5,131,996	5,132,039	2	Y	N
	128	3	22,282,746	22,282,619	2	Y	N
	333	3	1,107,330	1,107,662	2	Y	N
	736	3	25,096,178	25,096,913	2	Y	N
	7,858	3	25,073,805	25,081,671	2	Y	N
	186	5	3,731,248	3,731,433	2	Y	N
	204	5	25,097,248	25,097,451	2	Y	N
	366	5	25,095,010	25,094,645	2	Y	N
	757	5	3,738,178	3,738,934	2	Y	N
	1,217	5	3,733,416	3,734,633	2	Y	N
	1,461	5	25,113,647	25,115,110	2	Y	N
	6,248	5	25,170,023	25,176,277	2	Y	N
	301	9	18,785,456	18,785,756	2	Y	N
	867	9	18,541,230	18,542,097	2	Y	N
	293	11	4,751,079	4,751,371	2	Y	N
Os λ-8	233	2	28,182,792	28,183,024	2	Y	N
	20,551	2	28,320,680	28,341,230	3	Y	N
	31,370	2	2,927,273	2,958,642	3	N	Y
	33,590	2	1,764,729	1,798,318	3	N	Y
	68,479	3	206,466	274,944	3	N	Y
	206	4	20,082,930	20,082,725	3	N	Y
	264	4	24,319,871	24,319,608	3	N	Y
	4,512	7	11,886,889	11,891,400	3	N	Y
	5,946	7	7,387,487	7,393,432	3	N	Y
	12,098	7	7,206,838	7,218,935	3	N	Y

Table S2.5. Copy number of introduced molecules (single plasmid) and number of breakpoints in rice transgenic genome

Transgenic events		Genome coverage	Plasmid copy number	Number of breakpoints			
				Plasmid-plasmid	Plasmid-genome	Intra-chromosome	Inter-chromosome
pANIC10A-OsFPGS1	Os 10A-1	20.34	3.73	22	2	0	1
	Os 10A-2	21.04	12.33	50	4	1	1
	Os 10A-3	22.94	2.62	9	8	1	0
	Os 10A-4	23.18	5.34	16	8	2	0
	Os 10A-5	24.72	3.80	8	33	4	2
	Os 10A-6	22.93	1.43	17	16	36	47
pANIC12A-OsFPGS1	Os 12A-1	19.54	1.67	3	11	0	3
	Os 12A-2	18.05	2.53	5	5	27	29
	Os 12A-3	21.15	1.12	1	7	16	6
	Os 12A-4	20.78	2.36	1	4	3	0
	Os 12A-5	17.93	3.20	26	26	8	27
	Os 12A-6	20.70	2.33	3	7	2	0

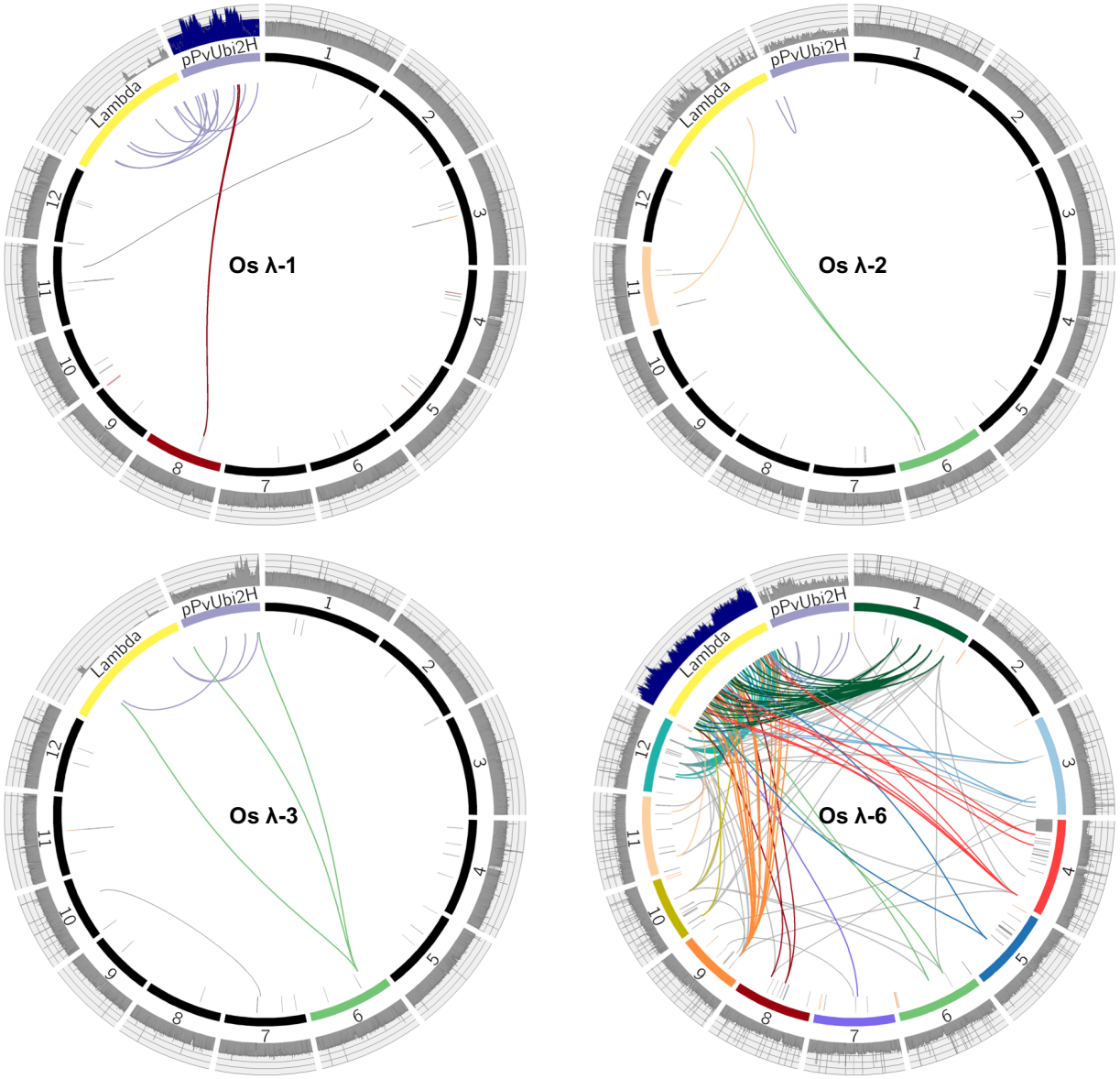


Figure S2.1. Circos plots of additional rice lines transformed with λ and plasmid pPvUbi2H. For rice event λ-6, the coverage of λ is divided by 8.

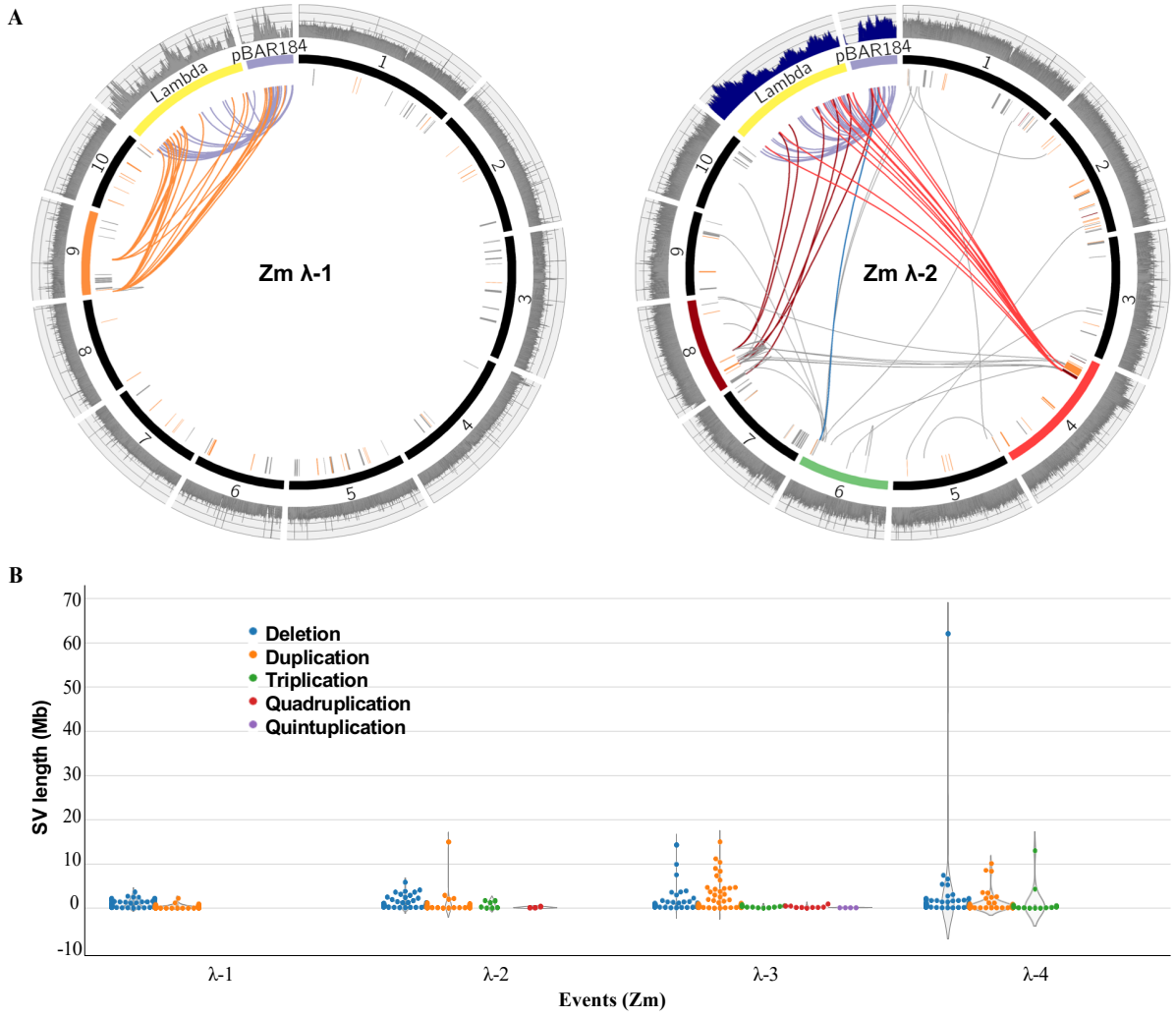
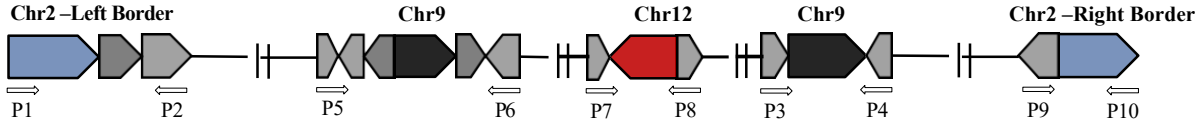


Figure S2.2. Additional data from maize lines transformed with λ and plasmid pBAR184. A. For maize event λ-2, the coverage values of λ and plasmid are divided by 10 and 8 respectively. B. Swarm and violin-plots showing deletions, duplications and triplications in all maize events transformed with λ. Each dot in the swarm plots represent a different SV. Violin plots represent the statistical distribution, where the width shows the probability of given SV lengths.

A



B

Plant	Insertion band					WT band	Genotype
	P1/P2	P3/P4	P5/P6	P7/P8	P9/P10	P1/P10	
1	✓	✓	✓	✓	✓	✓	+/–
2	✓	✓	✓	✓	✓	✓	+/–
3	✓	✓	✓	✓	✓	✓	+/–
6	✓	✓	✓	✓	✓	✓	+/–
9	✓	✓	✓	✓	✓	✓	+/–
10	✓	✓	✓	✓	✓	✓	+/–
12	✓	✓	✓	✓	✓	✓	+/–
14	✓	✓	✓	✓	✓	✓	+/–
16	✓	✓	✓	✓	✓	✓	+/–
19	✓	✓	✓	✓	✓	✓	+/–
20	✓	✓	✓	✓	✓	✓	+/–
22	✓	✓	✓	✓	✓	✓	+/–
4	✓	✓	✓	✓	✓	×	–/–
5	✓	✓	✓	✓	✓	×	–/–
8	✓	✓	✓	✓	✓	×	–/–
11	✓	✓	✓	✓	✓	×	–/–
15	✓	✓	✓	✓	✓	×	–/–
18	✓	✓	✓	✓	✓	×	–/–
7	×	×	×	×	×	✓	+/+
13	×	×	×	×	×	✓	+/+
17	×	×	×	×	×	✓	+/+
21	×	×	×	×	×	✓	+/+
23	×	×	×	×	×	✓	+/+
Os λ-4	✓	✓	✓	✓	✓	✓	+/–
WT	×	×	×	×	×	✓	+/+
λ	×	×	×	×	×	×	-

Figure S2.3. Linkage analysis of fragments from the 1.6 Mb array of rice λ-4 in self pollinated progeny. A. Schematic diagram of three genomic fragments flanked by lambda pieces embedded within the 1.6 Mb array on chromosome 2. Two of the fragments are from chromosome 9 (colored black), of size 102 bp and 464 bp respectively, while the other is a 108 bp segment from chromosome 12 (colored red). The arrows indicate the 3' ends of λ and genomic fragments. The positions of primers P1 to P10 for five amplicons are indicated. B. Three fragments from chromosome 9 and 12 are genetically linked with the insertion borders

on chromosome 2. Among the 23 T1 plants, 12 are heterozygous for the insertion on chromosome 2 (blue), 5 are wild type (grey), and 6 are homozygous for the insertion (orange).

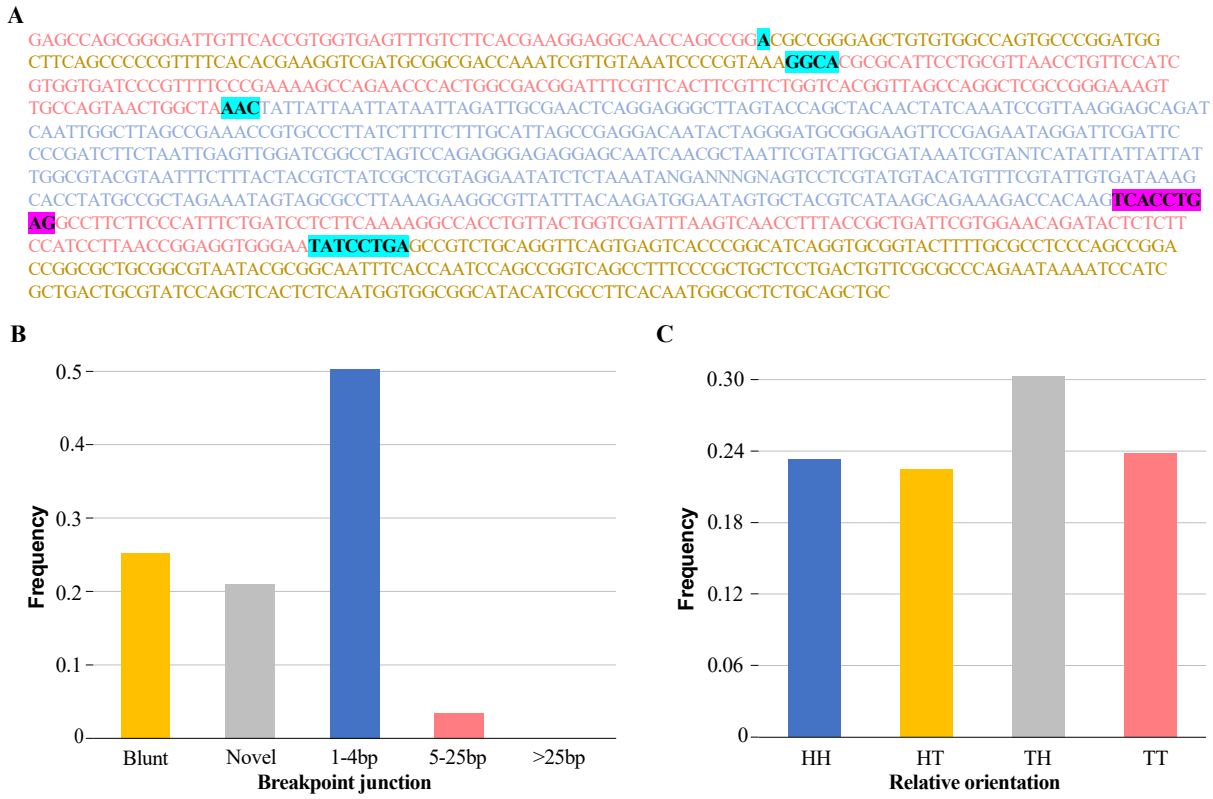
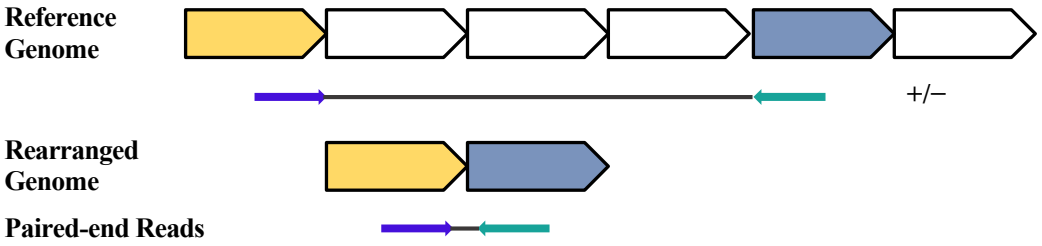
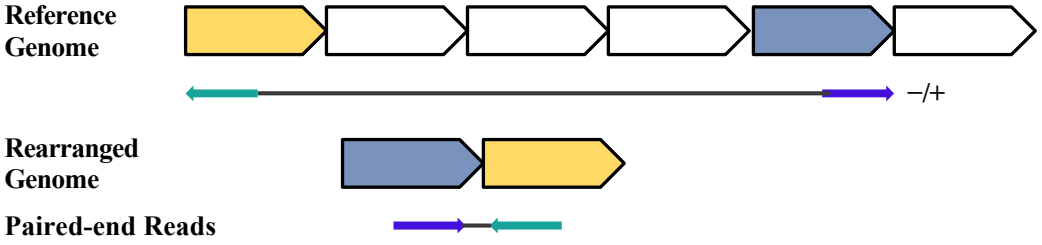


Figure S2.4. Distributions of microhomology at junction sites and relative orientations of rejoined fragments. A. Sequence of the 1.1 kb rearranged fragment from rice event λ-4, where the color code matches that in Figure 2B. Regions of microhomology at the breakpoint junctions are highlighted in cyan, and novel insertions are highlighted in magenta. B. Distribution of blunt-ends, insertion of novel sequence, and microhomology of 1-4 bp, 5-25 bp and >25 bp at breakpoint junctions. C. Distribution of four relative orientations of rejoined fragments, HH (head-head), HT (head-tail), TT (tail-tail), and TH (tail-head).

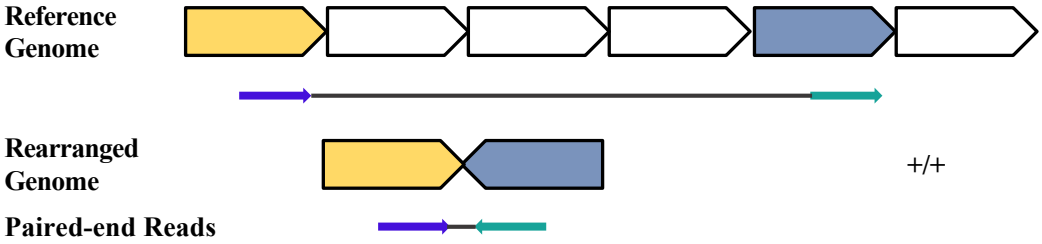
A **Deletion-type rearrangement**



B **Duplication-type rearrangement**



C **Intra-chromosomal translocation-type rearrangement 1)**



2)

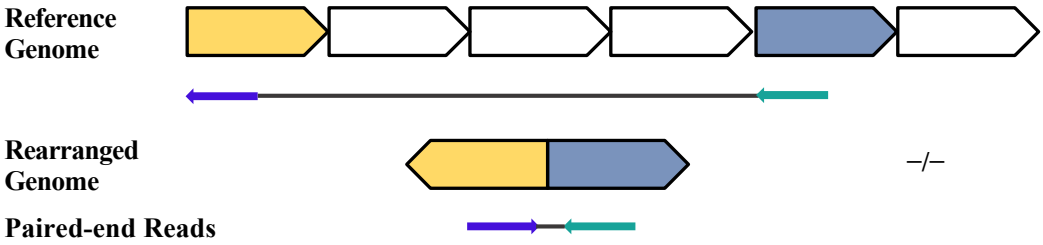


Figure S2.5. Three major intra-chromosomal SV types and the strand orientations of paired-end reads. A. Deletion-type intra-chromosomal rearrangement. B. Duplication-type intra-chromosomal rearrangement. C. Two types of intra-chromosomal translocation-type rearrangements.

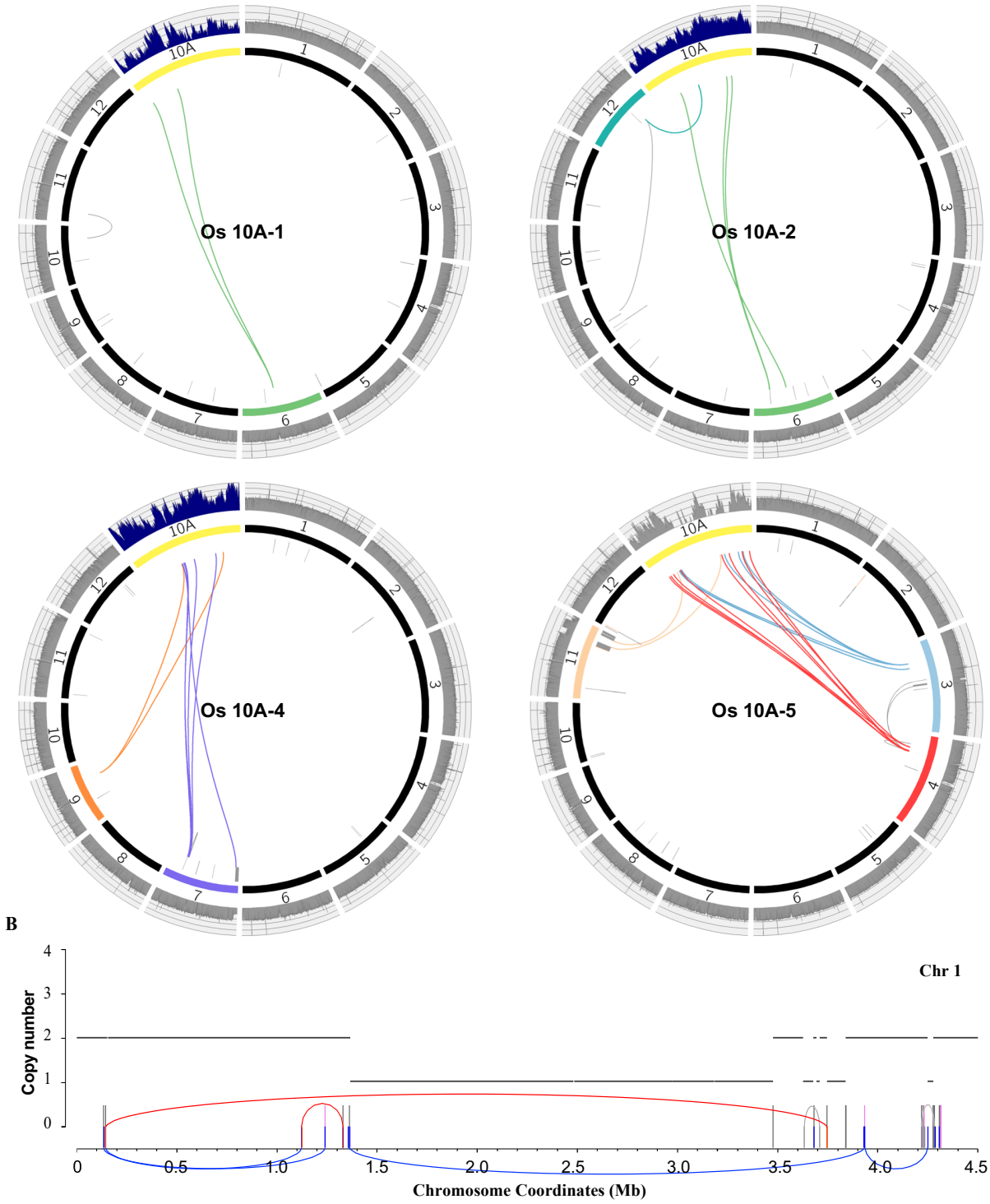


Figure S2.6. Additional data from rice lines transformed with plasmid pANIC10A-OsFPGS1. A. The coverage of plasmid 10A is divided by 2 in 10A-1 and 10A-4, and by 5 in 10A-2. **B.** Copy number states of the chromothripsis region of chromosome 1 of Figure 5C (region 0 - 4.5 Mb, highlighted in cyan) annotated as in Figure 1C.

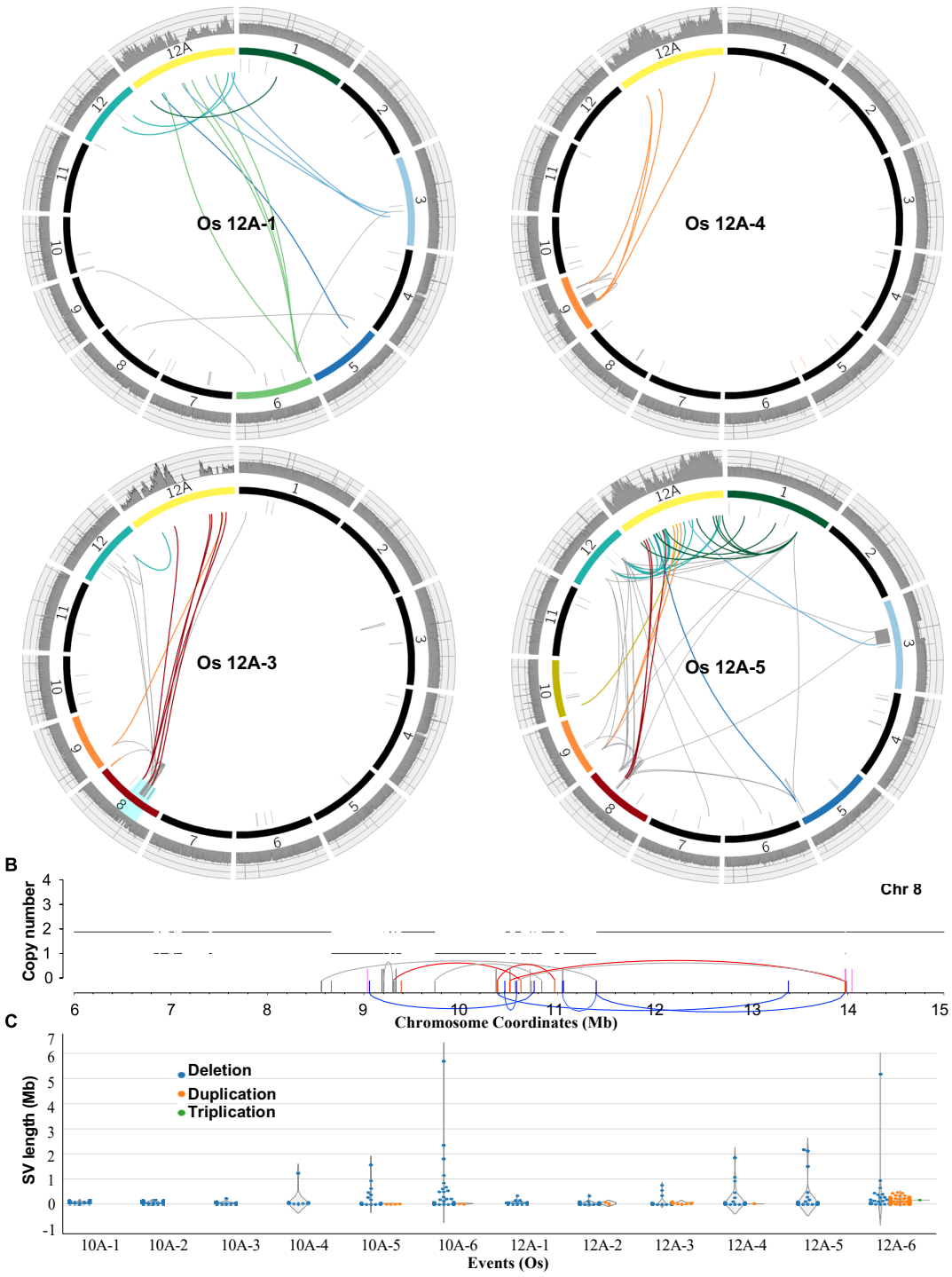


Figure S2.7. Additional data from rice lines transformed with plasmid pANIC12A- OsFPGS1. B. Copy number pattern of the highlighted region on chromosome 8 in 12A-3. C. Swarm and violin-plots showing the distribution of the size and number of deletions, duplications and triplications in all events transformed with single plasmids 10A and 12A. Each dot in the

swarm plots represent a different SV. Violin plots represent the statistical distribution, where the width shows the probability of given SV lengths.

APPENDIX B

SUPPLEMENTARY FIGURES AND TABLES – CHAPTER 3

Table S3.1. Assembly statistics and gaps in B73-Ab10 assemblies.

	Metrics	Nanopore	PacBio	Merged
Contig	N50 (Mb) / L50	2.0 / 325	41.2 / 16	162.0 / 6
	N60 (Mb) / L60	1.7 / 440	31.5 / 22	111.4 / 8
	N70 (Mb) / L70	1.3 / 584	23.8 / 30	88.8 / 10
	N80 (Mb) / L80	0.9 / 775	15.2 / 42	50.1 / 13
	N90 (Mb) / L90	0.6 / 1061	7.4 / 64	20.4 / 21
	Total contig number	1912	1103	1016
	Total sequence (Mb)	2120.5	2232.9	2241.4
	Contig number in Bionano scaffolds [a]	1699	211	132 [d]
	Contig sequence in scaffolds (Mb) [a]	2072.5	2156.5	2174.2
Scaffold	Number	43	25	50
	N50 (Mb)	125.7	162.9	161.8 [e]
	Max (Mb)	201.2	236.5	235.9
	Total length (Mb)	2170.1	2162.2	2178.1
	Contig overlaps [b]	929	114	12
	Gaps of known size [c]	728	73	51
	Total known gap size (Mb)	95.1	2.6	2.3
Pseudo-molecules	LTR Assembly Index (LAI)	8.57	27.98	27.8 [f]
	BUSCO (% Complete)	90.7	95.6	95.8
	Total length (Mb)	2161.1	2162.7	2162.8
	Contig overlaps [b]	910	114	5
	Gaps of unknown size	30	16	17
	Gaps of known size [c]	728	73	31
	Total known gap size (Mb)	93.2	2.6	1.3
	Assembled telomeres	1	9	15

[a] Contigs included in scaffolds are conflict-resolved contigs after hybrid assembly.

[b] Contig overlaps are identified when the contigs are integrated with the optical map. The Bionano hybrid scaffolding software marks them with 13 Ns.

[c] Gap sizes are estimated by Bionano maps during hybrid assembly.

[d] Only 63 are included in pseudomolecules; the remaining 59 are anchored to small scaffolds (< 3Mb) containing CentC or knob arrays that lack genetic and pan-genome markers and could not be placed on chromosomes (Suppl. Table 5).

[e] The N50 is smaller in Merged than PacBio due to the correction of the CentC array on chromosome 9 (Figure 1B).

[f] LAI is lower in the Merged than PacBio due to a reduction in the total number of LTRs, presumably because sequence overlaps were removed.

Table S3.2. Accuracy of genome assemblies as assessed by comparison to Bionano maps.

	Nanopore	PacBio	Merged
Contig misjoins / Conflict cuts	425	18	1
Number of collapsed repeats > 25 Kb	705	56	3
Sequence lost by collapse of repeats > 25 Kb (Mb)	44.80	3.65	0.13
Number of expanded repeats > 25 Kb	22	13	10
Sequence gained by expansion of repeats > 25 Kb (Mb)	1.28	1.49	0.57

Table S3.3. Coordinates and composition of centromeres defined by CENH3 ChIP-seq in the B73-Ab10 assembly.

Chr	Start (bp)	End (bp)	Size (bp)	100N[a]	CentC (%)	CRM (%)	cinful-zeon (%)	grande (%)	huck (%)	opie-ji (%)	prem1 (%)
chr1	137,090,000	138,130,000	1,040,000	Y	54.8	41.1	0	0	0	0.9	0.1
chr2	95,290,000	97,165,000	1,875,000	N	1.3	31.5	14	3.2	6.3	2.5	7.6
chr3	85,880,000	87,705,000	1,825,000	N	14.7	37.2	8.8	0	4.8	2	4.6
chr4	108,485,000	110,135,000	1,650,000	N	1	44.6	10.2	0.5	3.6	3.5	6.3
chr5	104,255,000	106,220,000	1,965,000	N	0	16.3	19.2	0.8	2.2	1.5	8.1
chr6	53,825,000	54,655,000	830,000	Y	49.1	41.4	0.8	0	1.5	0.2	0.0
chr7	56,220,000	56,860,000	640,000	Y	30.8	54.5	0.3	0	0	0	0.0
chr8	51,095,000	52,795,000	1,700,000	N	0	25.3	17.9	3.9	1.5	1.3	4.3
chr9	54,840,000	56,275,000	1,435,000	N	0	5.1	18.8	6.4	2.8	2.2	11.1
chr10	50,845,000	52,505,000	1,660,000	N	9.3	22.3	13.1	1.8	3.2	5.2	7.1

[a] Gaps marked by 100 Ns are of unknown size.

Table S3.4. CENH3 enrichment and mappability of Illumina reads in active centromeres.

Chr	Mappability[a]	CentC length (kb) / fold change [b]	CRM length (kb) / fold change	cinful-zeon length (kb) / fold change	grande length (kb) / fold change	huck length (kb) / fold change	opie-ji length (kb) / fold change	prem1 length (kb) / fold change
chr1	Unique	339.3 / 10.0	59.7 / 7.3	0 / NA	0 / NA	0 / NA	9.3 / 8.6	0 / NA
	Non-unique	226.5 / 1.7	390.6 / 4.3	0 / NA	0 / NA	0 / NA	0 / NA	0 / NA
chr2	Unique	24.9 / 32.7	167.2 / 18.2	229.1 / 16.0	24.5 / 19.2	113.0 / 11.6	27.4 / 20.6	54.3 / 21.4
	Non-unique	0 / NA	470.0 / 8.4	51.7 / 8.1	36.0 / 6.0	0 / NA	23.9 / 4.4	12.7 / 4.2
chr3	Unique	213.5 / 15.1	166.5 / 12.2	121.4 / 14.6	0 / NA	61.5 / 7.4	23.7 / 11.7	25.1 / 15.6
	Non-unique	53.8 / 1.3	598.1 / 7.8	38.6 / 9.1	0 / NA	27.0 / 1.1	18.3 / 2.2	8.5 / 7.1
chr4	Unique	17.1 / 23.6	178.3 / 14.0	138.4 / 17.7	4.7 / 42.5	59.3 / 7.3	51.1 / 16.2	52.4 / 19.1
	Non-unique	0 / NA	622.5 / 9.1	45.9 / 5.7	0 / NA	0 / NA	33.6 / 5.7	0 / NA
chr5	Unique	0 / NA	26.5 / 10.1	354.3 / 20.3	15.0 / 10.0	27.4 / 3.1	32.0 / 18.1	132.1 / 18.2
	Non-unique	0 / NA	321.7 / 6.6	33.4 / 4.0	0 / NA	15.6 / 2.3	0 / NA	0 / NA
chr6	Unique	242.4 / 11.0	57.7 / 11.4	6.3 / 5.5	0 / NA	12.4 / 2.9	1.8 / 21.8	0 / NA
	Non-unique	162.8 / 1.8	313.0 / 5.8	0 / NA	0 / NA	0 / NA	0 / NA	0 / NA
chr7	Unique	101.2 / 10.4	66.2 / 6.3	2.1 / 18.5	0 / NA	0 / NA	1.5 / 23.2	0 / NA
	Non-unique	95.0 / 2.4	317.1 / 5.4	0 / NA	0 / NA	0 / NA	0 / NA	0 / NA
chr8	Unique	0 / NA	275.8 / 17.8	54.0 / 19.8	65.4 / 18.2	25.6 / 22.7	26.9 / 22.3	53.0 / 19.7
	Non-unique	0 / NA	49.4 / 4.7	418.0 / 7.4	0 / NA	0 / NA	0 / NA	16.6 / 12.7
chr9	Unique	0 / NA	24.5 / 33.3	227.4 / 19.6	89.0 / 27.3	40.9 / 16.3	31.6 / 29.4	102.0 / 20.2
	Non-unique	0 / NA	73.6 / 15.6	46.4 / 4.2	0 / NA	0 / NA	9.0 / 9.9	29.8 / 10.3
chr10	Unique	119.6 / 13.5	140.2 / 16.5	196.1 / 22.8	19.3 / 40.0	52.5 / 7.2	73.3 / 18.0	76.3 / 22.5
	Non-unique	33.4 / 1.5	264.7 / 7.6	33.3 / 4.7	8.6 / 4.6	1.5 / 11.6	34.3 / 6.5	0 / NA
Total	Unique	1058.1 / 12.4	940.7 / 18.7	1550.9 / 14.4	218.0 / 23.9	392.7 / 10.6	257.8 / 18.8	495.1 / 19.8
	Non-unique	571.4 / 1.8	3798.1 / 5.9	298.7 / 7.4	44.6 / 5.7	44.1 / 1.9	86.2 / 5.4	67.7 / 9.3

[a] Non-unique regions are defined as sequences with too low (<3) or low high (>100) numbers of mapped Illumina reads. The total genome coverage was 30X.

[b] Fold change is expressed as the ratio of CENH3 ChIP-seq reads to genomic reads for each repeat type.

Table S3.5. Repetitive components in B73-Ab10 assemblies.

	Repeat Type	Nanopore	PacBio	Merged
Scaffold	knob180 (bp)	3,962,210	9,893,818	16,861,684
	TR-1 (bp)	2,324,807	4,984,957	7,253,994
	CentC (bp)	708,654	3,233,366	2,881,960
	rDNA intergenic spacer (bp)	185,738	789,045	755,420
	Subtelomere (bp)	99,243	504,890	563,144
Pseudo-molecules	knob180 (bp)	3,469,793	9,894,164	10,371,693
	TR-1 (bp)	1,824,356	4,984,826	4,979,336
	CentC (bp)	708,574	3,233,759	2,939,574
	rDNA intergenic spacer (bp)	185,513	789,045	622,630
	Subtelomere (bp)	99,218	554,866	598,330

Table S3.6. Composition of CentC arrays.

Chr	Start (bp)	End (bp)	Size (bp)	100N[a]	CentC (%)	CRM (%)	cinful-zeon (%)	grande (%)	huck (%)	opie-ji (%)	prem1 (%)
chr1	133,797,033	134,161,088	364,055	N	28.2	1.9	9.0	3.8	6.2	2.6	3.7
chr1	136,702,219	138,370,323	1,668,104	Y	45.7	38.8	2.4	0.0	1.5	0.9	0.6
chr2	95,886,484	96,018,046	131,562	N	19.1	45.8	10.7	0.0	0.0	6.8	6.5
chr3	83,463,110	83,579,253	116,143	N	29.9	0.0	6.4	12.3	11.2	0.0	6.8
chr3	85,837,492	86,608,170	770,678	N	37.8	41.3	3.8	0.0	3.7	1.1	3.3
chr4	109,241,724	109,363,707	121,983	N	14.2	61.3	4.0	0.0	0.0	0.0	0.0
chr5	107,727,923	107,946,850	218,927	N	41.8	34.3	4.5	0.0	0.0	0.0	3.1
chr6	53,802,144	54,696,183	894,039	Y	50.5	40.8	0.7	0.0	1.4	0.2	0.0
chr7	56,131,392	56,883,231	751,840	Y	32.6	54.5	0.3	0.0	0.0	0.0	0.0
chr7	58,725,452	59,026,217	300,765	N	73.7	2.3	0.0	0.0	0.0	6.1	5.3
chr8	49,443,032	49,660,040	217,008	N	60.9	23.8	3.4	0.0	0.0	0.0	3.7
chr8	50,101,170	50,244,111	142,941	N	47.8	28.6	9.9	0.0	0.0	0.0	0.0
chr9	58,430,781	58,672,011	241,230	N	59.7	15.8	0.0	0.0	0.0	0.0	8.5
chr9	58,867,400	59002849	135,449	N	61.2	31.0	0.4	0.0	0.0	0.0	0.0
chr10	49,448,433	49,559,148	110,715	N	67.8	0.0	0.0	0.0	11.1	0.0	8.6
chr10	52,128,660	52,443,218	314,558	N	49.2	31.1	5.2	0.0	0.0	4.2	3.2

[a] Gaps marked by 100 Ns are of unknown size.

Table S3.7. Composition of knob180 and TR-1 knobs.

	Chr	Start (bp)	End (bp)	Size (bp)	100N[a]	Ngap (%)	knob180 (%)	TR-1 (%)	cinful-zeon (%)	grande (%)	huck (%)	opie-ji (%)	prem1 (%)
TR-1	chr4	231,661,938	232,984,750	1,322,813	N	0.0	0.1	51.0	30.3	0.0	0.0	1.9	1.9
	chr10	142,321,698	146,554,927	4,233,230	N	0.0	1.1	42.7	27.3	0.7	1.2	1.2	3.1
	chr10	150,506,817	153,088,305	2,581,489	N	0.0	0.0	42.0	29.7	1.2	1.7	1.4	2.2
	chr10	157,208,255	159,276,069	2,067,815	N	0.0	0.5	36.9	20.9	0.6	2.5	4.9	5.9
knob180	chr5	198,844,996	200,124,258	1,279,263	N	0.0	59.3	5.0	12.0	1.1	1.1	2.0	5.0
	chr6	1	623,495	623,495	N	14.9	56.7	2.3	4.7	2.3	1.9	0.0	4.5
	chr6	176,451,347	177,079,220	627,874	N	0.0	28.9	18.4	11.9	4.5	1.5	4.0	0.8
	chr7	155,073,718	157,644,910	2,571,193	Y	7.0	66.0	2.7	5.9	0.0	0.6	2.8	1.6
	chr8	160,828,702	162,735,401	1,906,700	Y	15.0	58.2	3.7	6.3	0.0	0.7	1.4	1.6
	chr9	1,630	842,889	841,260	N	0.0	67.0	4.3	7.4	1.7	5.9	4.3	0.9
	chr10	174,217,005	178,146,998	3,929,994	Y	7.8	63.6	0.8	7.6	0.4	1.7	2.0	1.5
	chr10	180,357,409	182,945,132	2,587,724	N	0.0	53.9	0.0	9.8	1.1	0.5	5.6	4.0

[a] Gaps marked by 100 Ns are of unknown size.

Table S3.8. Gene and transposon distributions in the Ab10 haplotype and corresponding N10 regions.

Region	N10/Ab10 Proximal [a]	Ab10 Shared [b]	Ab10 Specific [c]
Size (bp)	20,000,000	12,716,384	22,438,721
Genes	521	580	450
Gene density (genes/Mb)	26.1	45.6	20.1
CDS content (%)	8.8	15.0	5.5
Average gene length (bp)	3,377	3,284	2,757
Average CDS length (bp)	1,465	1,425	911
Single exon gene (%)	35.5	36.0	49.1
Genes overlapped with TE by 95% (%)	4.2	4.3	7.9
TE content (%)	74.2	76.4	88.1

[a] Sequence in a 20 Mb region left of the first TR-1 knob that is not a part of the Ab10 haplotype (122.3 - 142.3 Mb).

[b] Sequence in two large inversions with shared synteny between the Ab10 haplotype and N10 (153.0 -157.4 Mb and 159.4 - 167.7 Mb).

[c] Sequence present in the Ab10 haplotype but not the B73 N10 genome, including a region between the first two TR-1 knobs (146.6 - 150.5 Mb), a region from the end of the second inversion to the large knob (167.7 - 174.3 Mb), and a region from the large knob to the end of Ab10 haplotype (183.1 - 195.0 Mb).

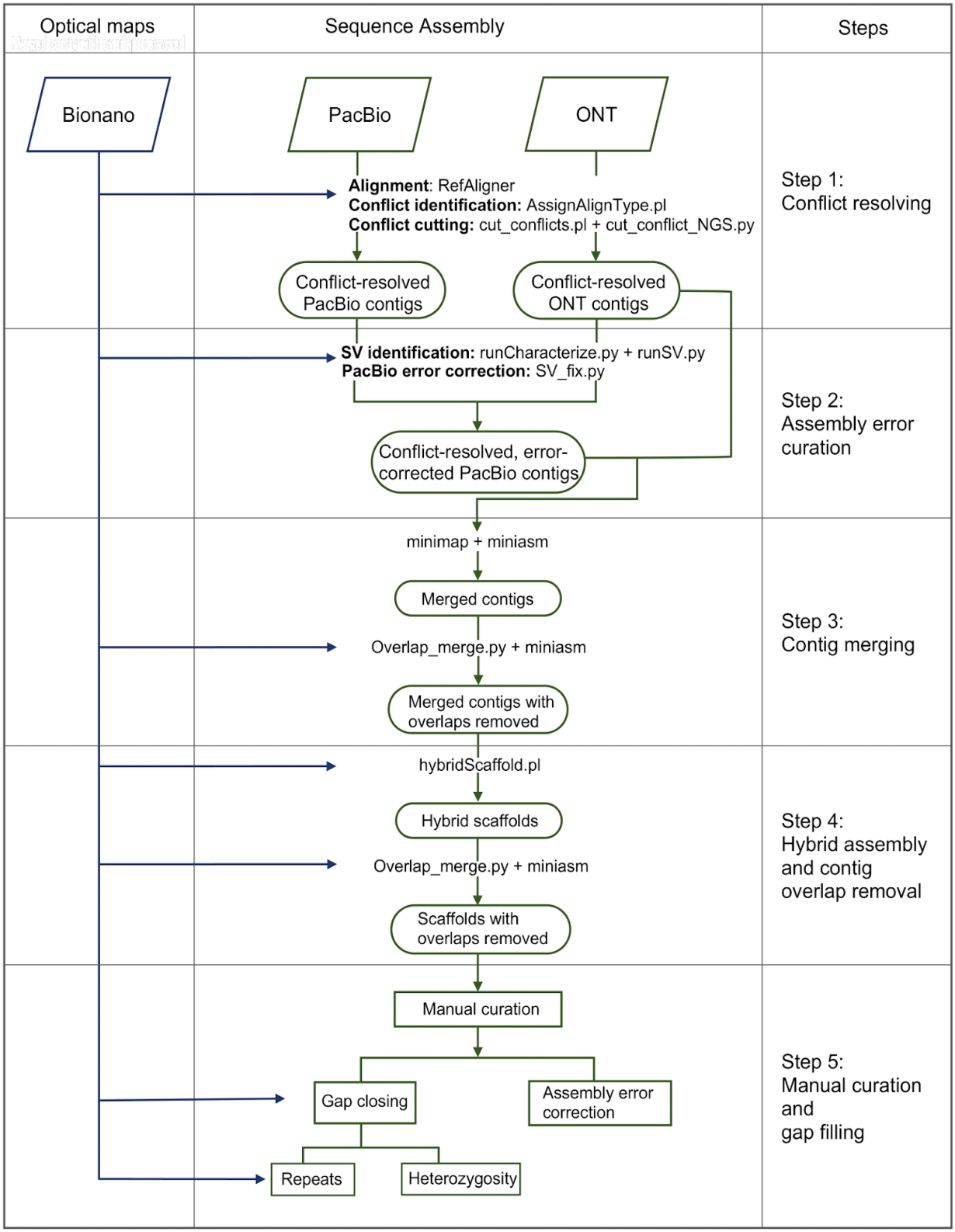
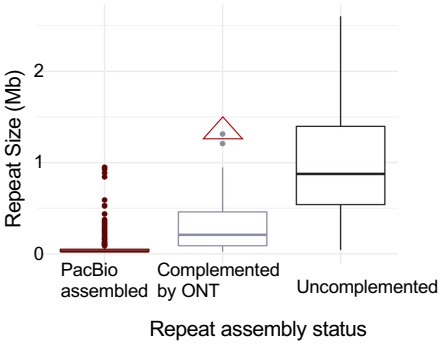


Fig S3.1. Workflow for the B73-Ab10 assembly pipeline.

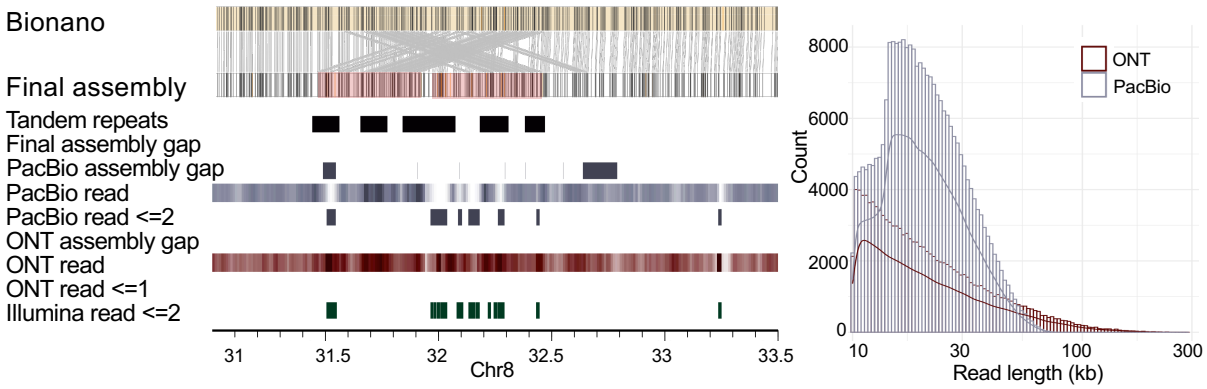
A

Gaps	PacBio pseudo-molecule	Complemented by ONT	Overlap with tandem repeat (>25kb)	Overlap with heterozygosity
13N gap	114	109	84	10
100N gap	16	0	12	0
N gap (>10kb)	42	20	37	1
Total	172	129	139	11
P-value [a]			3.9403e-156	5.3694e-09

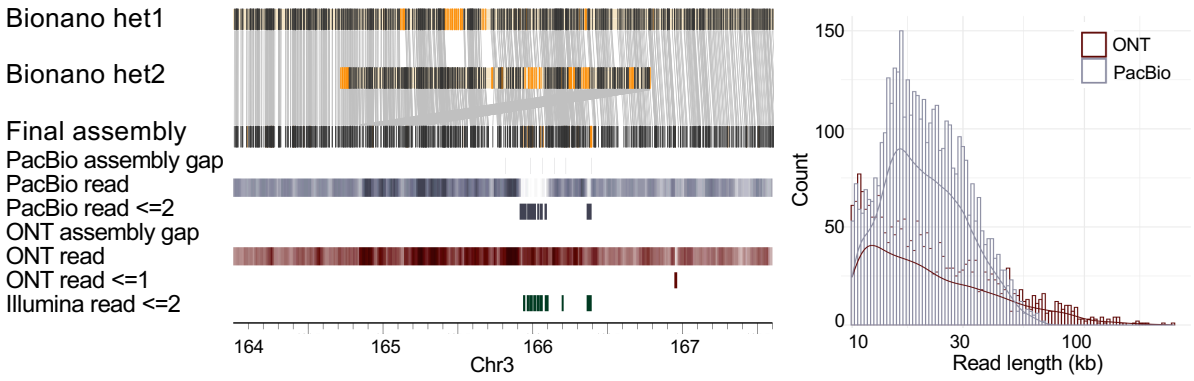
[a] Two-tailed Fisher's exact test



B



C



D

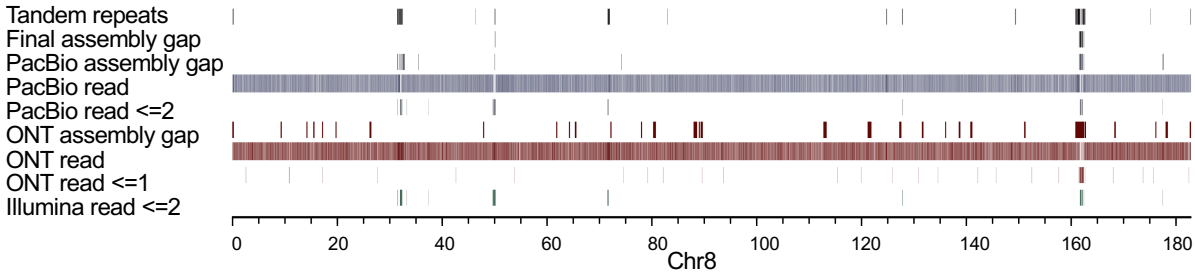


Fig S3.2. Complementation of PacBio assembly gaps by Nanopore contigs. A) Co-occurrence of PacBio assembly gaps with tandem repeats and heterozygosity. Tandem repeat arrays were defined as having at least 25 Kb of tandem repetitiveness with a maximum interval of 300 Kb between repeat units. The table illustrates that the majority of PacBio assembly gaps, either from contig overlaps (indicated by 13N) or gaps of known size, were complemented by ONT contigs. The correlations between PacBio assembly gaps and tandem repeat arrays or heterozygous regions are also shown (where P values were calculated by Fisher's exact test). The boxplot on the right shows the size distribution of repeat arrays (with at least 25 tandem repeats) that were fully assembled by PacBio data, those that were not assembled but successfully complemented by ONT data, and those associated with gaps that remain in the final assembly. The tandem repeat highlighted with a triangle is displayed in B. B) Example of a large tandem repeat region on chr8: 31-33.5 Mb. PacBio and ONT read alignment (MAPQ>=0) is shown as a heatmap with 10 Kb windows. Read gaps were defined as regions with fewer than three reads for PacBio error-corrected reads and Illumina reads, and fewer than two reads for ONT error-corrected dataset. Tracks showing no information have no gaps. Length distributions of PacBio and ONT reads mapped to the corresponding region are displayed in a histogram on the right. C) Example of a large heterozygous region on chr3: 164-167.6 Mb. Tracks are annotated as in B. Length distributions of PacBio and ONT reads mapped to the corresponding region are displayed in a histogram on the right. D) Whole chromosome view of assembly and read mapping on chr8.

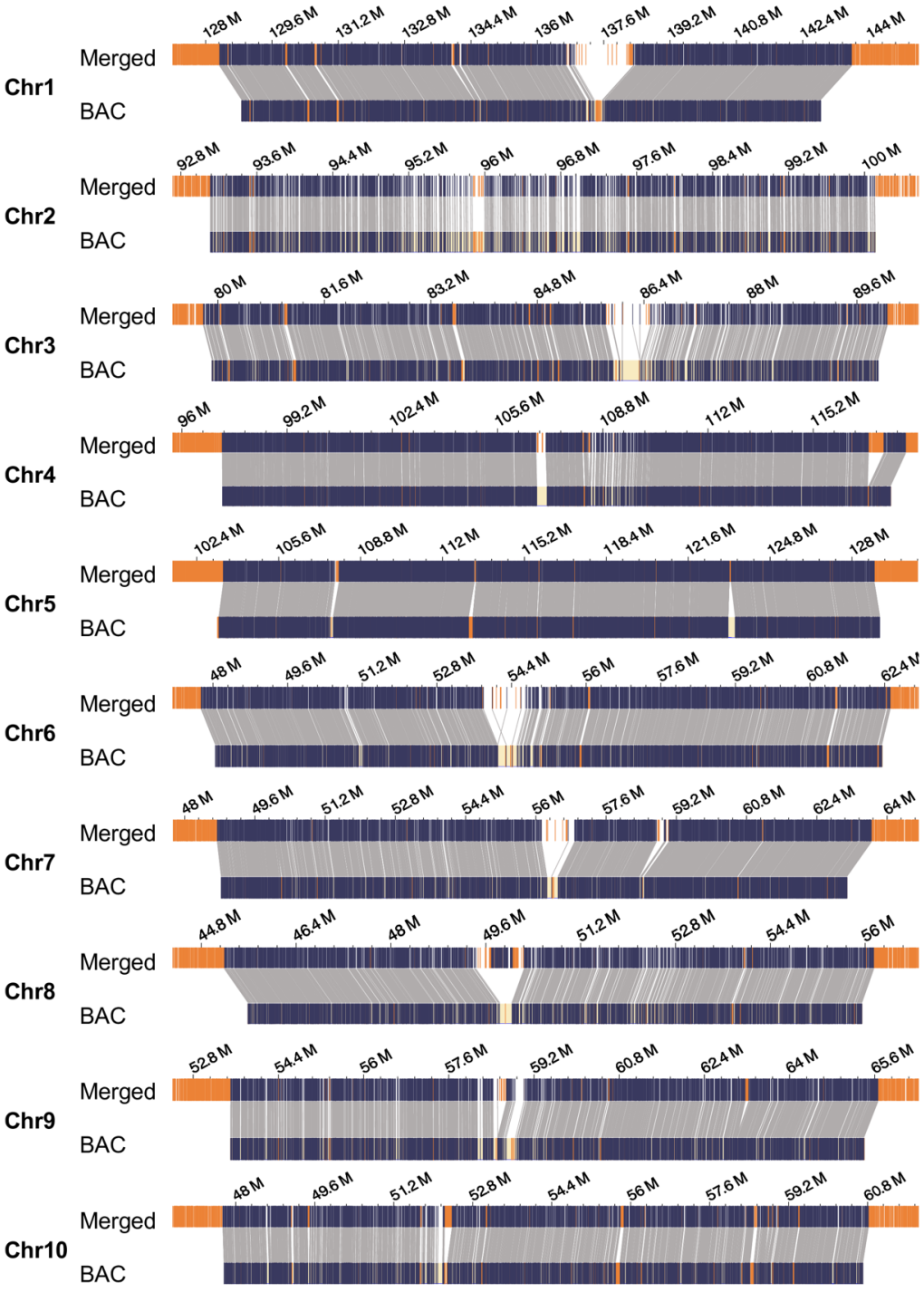


Fig S3.3. The alignment of BAC-based assemblies of B73 centromeres to the merged assembly in optical map format. The connecting lines represent matching regions between the two assemblies.

APPENDIX C

SUPPLEMENTARY FIGURES AND TABLES – CHAPTER 4

Table S4.1. Total aligned length between NAM and B73 measured by whole-genome alignment and PE 150 Illumina (~30X) mapping.

Line	Syntenic alignment (bp)	Illumina alignment (bp)	Overlap between Syntenic and Illumina (bp)	Syntenic specific (bp)	Illumina specific (bp)
B97	1,325,115,734	1,270,493,615	1,153,590,959	171,524,775	116,902,656
CML103	1,324,911,096	1,194,424,303	1,115,192,179	209,718,917	79,232,124
CML228	1,319,400,491	1,163,459,933	1,081,168,409	238,232,082	82,291,524
CML247	1,312,190,860	1,150,090,533	1,069,822,484	242,368,376	80,268,049
CML277	1,336,866,082	1,165,338,185	1,092,849,542	244,016,540	72,488,643
CML322	1,333,242,180	1,173,807,099	1,095,183,706	238,058,474	78,623,393
CML333	1,335,335,563	1,194,522,220	1,114,386,638	220,948,925	80,135,582
CML52	1,300,117,745	1,161,810,391	1,068,116,205	232,001,540	93,694,186
CML69	1,315,702,061	1,171,189,452	1,089,183,315	226,518,746	82,006,137
HP301	1,334,953,758	1,212,145,301	1,132,206,557	202,747,201	79,938,744
IL14H	1,300,631,195	1,155,244,846	1,060,541,220	240,089,975	94,703,626
Ki11	1,311,632,415	1,176,837,965	1,095,894,055	215,738,360	80,943,910
Ki3	1,337,938,727	1,169,454,355	1,095,112,089	242,826,638	74,342,266
Ky21	1,364,521,736	1,260,287,824	1,169,279,963	195,241,773	91,007,861
M162W	1,338,289,415	1,212,297,460	1,130,767,813	207,521,602	81,529,647
M37W	1,332,311,592	1,225,372,604	1,135,204,551	197,107,041	90,168,053
Mo18W	1,340,314,676	1,179,452,545	1,103,339,503	236,975,173	76,113,042
MS71	1,377,887,393	1,253,060,533	1,175,346,201	202,541,192	77,714,332
NC350	1,348,407,125	1,176,251,628	1,105,733,900	242,673,225	70,517,728
NC358	1,327,636,556	1,190,975,090	1,103,694,180	223,942,376	87,280,910
Oh43	1,345,404,924	1,256,204,921	1,158,891,636	186,513,288	97,313,285
Oh7b	1,376,148,050	1,266,086,044	1,176,882,047	199,266,003	89,203,997
P39	1,281,660,640	1,135,148,694	1,056,861,788	224,798,852	78,286,906
Tx303	1,355,443,879	1,205,445,046	1,131,753,658	223,690,221	73,691,388
Tzi8	1,354,141,118	1,186,206,493	1,110,333,969	243,807,149	75,872,524

Table S4.2. Total number of SNPs between NAM and B73 identified through whole-genome alignment and PE 150 Illumina (~30X) mapping.

Line	Syntenic SNPs	Illumina SNPs	Overlap	Syntenic specific	Illumina specific
B97	6,197,993	6,161,448	4,216,286	1,981,707	1,945,162
CML103	6,730,490	6,830,644	4,772,059	1,958,431	2,058,585
CML228	6,965,258	6,968,422	4,726,010	2,239,248	2,242,412
CML247	7,158,720	7,294,828	5,031,943	2,126,777	2,262,885
CML277	6,874,060	7,119,973	4,841,051	2,033,009	2,278,922
CML322	6,838,927	6,912,696	4,706,061	2,132,866	2,206,635
CML333	6,500,059	6,805,154	4,530,645	1,969,414	2,274,509
CML52	4,508,345	7,076,563	3,469,902	1,038,443	3,606,661
CML69	6,893,343	7,047,833	4,870,856	2,022,487	2,176,977
HP301	6,235,003	6,771,250	4,465,255	1,769,748	2,305,995
IL14H	4,609,801	7,080,577	3,518,008	1,091,793	3,562,569
Ki11	6,800,227	7,144,224	4,831,586	1,968,641	2,312,638
Ki3	6,795,702	7,015,615	4,739,719	2,055,983	2,275,896
Ky21	5,891,467	6,071,431	4,007,662	1,883,805	2,063,769
M162W	6,689,237	6,871,150	4,756,459	1,932,778	2,114,691
M37W	6,554,498	6,633,441	4,598,319	1,956,179	2,035,122
Mo18W	6,778,761	6,864,045	4,735,088	2,043,673	2,128,957
MS71	6,226,986	6,461,529	4,399,553	1,827,433	2,061,976
NC350	7,065,739	7,070,570	4,962,877	2,102,862	2,107,693
NC358	6,483,247	6,898,899	4,506,731	1,976,516	2,392,168
Oh43	6,114,159	6,310,233	4,211,563	1,902,596	2,098,670
Oh7b	5,743,370	5,764,260	3,922,016	1,821,354	1,842,244
P39	6,971,739	7,146,653	4,850,483	2,121,256	2,296,170
Tx303	6,628,981	6,792,916	4,718,395	1,910,586	2,074,521
Tzi8	6,697,618	6,905,435	4,653,253	2,044,365	2,252,182

Table S4.3. Structural variants (unalignment, inversion, tandem duplications) relative to B73 across the whole genome.

	Unalignment		Inversion (>20Kb)		Tandem Duplication (>10Kb)	
Query line	Number	Size (bp)	Number	Size (bp)	Number	Size (bp)
B97	68,177	854,555,595	16	7,911,444	42	2,965,189
CML103	75,254	912,440,700	32	9,663,765	69	4,284,020
CML228	76,732	957,952,494	33	17,588,048	98	10,227,982
CML247	79,421	977,139,534	33	9,326,837	77	6,550,994
CML277	78,097	955,286,509	35	9,045,158	130	6,740,525
CML322	76,040	930,445,780	47	50,164,694	49	5,376,904
CML333	75,460	931,264,895	33	13,235,980	59	6,488,305
CML52	77,374	970,469,033	40	36,504,989	120	7,502,433
CML69	77,768	949,798,670	34	19,113,898	55	5,052,287
HP301	73,209	893,388,514	30	11,582,644	36	3,056,711
IL14H	76,179	932,864,278	44	37,826,184	66	5,931,792
Ki11	77,548	959,218,817	37	17,999,235	58	5,937,488
Ki3	76,847	958,284,549	29	12,466,038	57	3,919,777
Ky21	68,555	849,838,736	34	19,909,991	61	3,900,881
M162W	74,937	908,212,468	32	23,492,103	31	2,235,391
M37W	73,446	905,759,884	37	19,253,776	42	3,580,080
Mo18W	76,111	937,095,984	40	23,294,264	47	3,126,914
MS71	69,614	864,510,512	29	9,868,632	33	2,559,303
NC350	77,539	975,398,973	35	15,006,973	40	3,629,420
NC358	75,690	923,553,028	24	9,498,126	46	3,856,304
Oh43	68,736	850,618,690	34	23,197,191	39	3,347,196
Oh7b	64,488	848,093,802	28	34,566,458	68	3,505,848
P39	77,436	957,486,311	49	61,198,656	32	2,366,083
Tx303	73,671	908,204,734	27	10,108,385	46	4,204,799
Tzi8	76,236	935,198,368	34	12,032,167	102	6,119,089

Table S4.4. Structural variants (unalignment, inversion, tandem duplications) across 26 lines through all-by-all alignment.

Reference line	Unalignment		Inversion (>20Kb)		Tandem Duplication (>10Kb)	
	Number	Size (bp)	Number	Size (bp)	Number	Size (bp)
B97	1,864,565	23,201,614,623	846	485,865,505	361	31,614,270
CML103	1,868,630	23,446,644,923	908	491,579,352	493	36,128,222
CML228	1,888,044	23,036,601,666	1,062	523,874,260	866	47,803,359
CML247	1,840,373	23,177,351,167	968	642,391,508	1,624	82,660,871
CML277	1,887,346	23,400,065,853	946	465,732,603	1,159	69,208,086
CML322	1,889,855	23,307,921,784	977	518,883,162	2,296	98,269,413
CML333	1,883,076	23,048,432,319	1,216	1,318,651,366	715	49,587,368
CML52	1,876,380	23,214,366,620	1,016	579,514,019	787	62,513,384
CML69	1,884,483	23,796,325,894	1,107	975,680,447	2,020	123,215,200
HP301	1,882,911	23,105,136,853	1,014	690,755,700	917	63,656,299
IL14H	1,929,453	23,839,242,115	1,145	609,019,793	666	50,888,192
Ki11	2,001,973	24,629,586,697	1,162	1,034,308,544	939	84,965,425
Ki3	1,876,900	23,344,115,131	1,009	829,422,782	892	79,134,551
Ky21	1,859,022	22,977,561,730	894	514,402,839	1,170	85,187,168
M162W	1,873,056	23,036,486,134	934	667,807,244	1,023	79,817,185
M37W	1,874,963	23,064,980,850	951	774,967,091	655	51,076,286
Mo18W	1,875,313	23,257,462,325	1,056	658,529,895	841	71,737,714
MS71	1,894,367	23,170,980,625	1,103	843,131,859	977	81,226,572
NC350	1,871,729	23,184,055,773	918	489,907,781	878	75,683,812
NC358	1,863,788	23,417,290,145	945	580,050,596	834	86,318,493
Oh43	1,859,653	22,692,949,899	793	465,656,495	1,144	94,734,533
Oh7b	1,875,159	23,255,390,826	991	774,997,779	963	77,356,690
P39	1,807,681	23,386,610,941	942	995,707,702	1,370	86,458,035
Tx303	1,892,385	23,400,128,539	1,318	1,473,044,687	759	64,723,417
Tzi8	1,861,358	22,867,471,386	873	519,414,980	1,204	105,642,153

Table S4.5. Distribution of the B73-alternative haplotypes in pericentromeric areas among NAM lines.

Chromosome	Inbreds with the alternative haplotype	Flint	Temperate	Mixed	Tropical
1	0	0	0	0	0
2	20	3	6	3	8
3	3	2	1	0	0
4	0	0	0	0	0
5	2	2	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	5	3	0	1	1
9	1	1	0	0	0
10	2	0	1	0	1

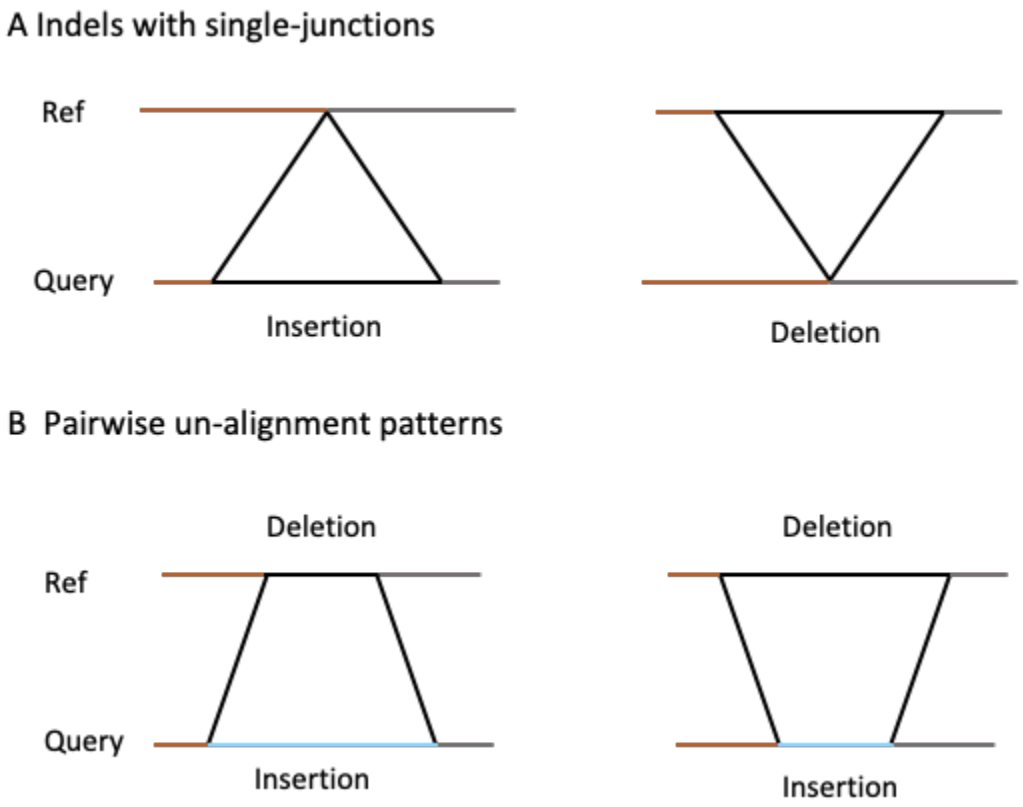


Figure S4.1. Indels with a single junction and those with pairwise-unaligned structure. A) Insertions and deletions with a single-junction feature. Variant regions are highlighted in black. Variant types and breakpoints could be inferred from read alignments. B) Insertions and deletions with an unalignment pattern. Deletions can be characterized with read mapping whereas the reciprocal insertions that exceed read length cannot be identified.

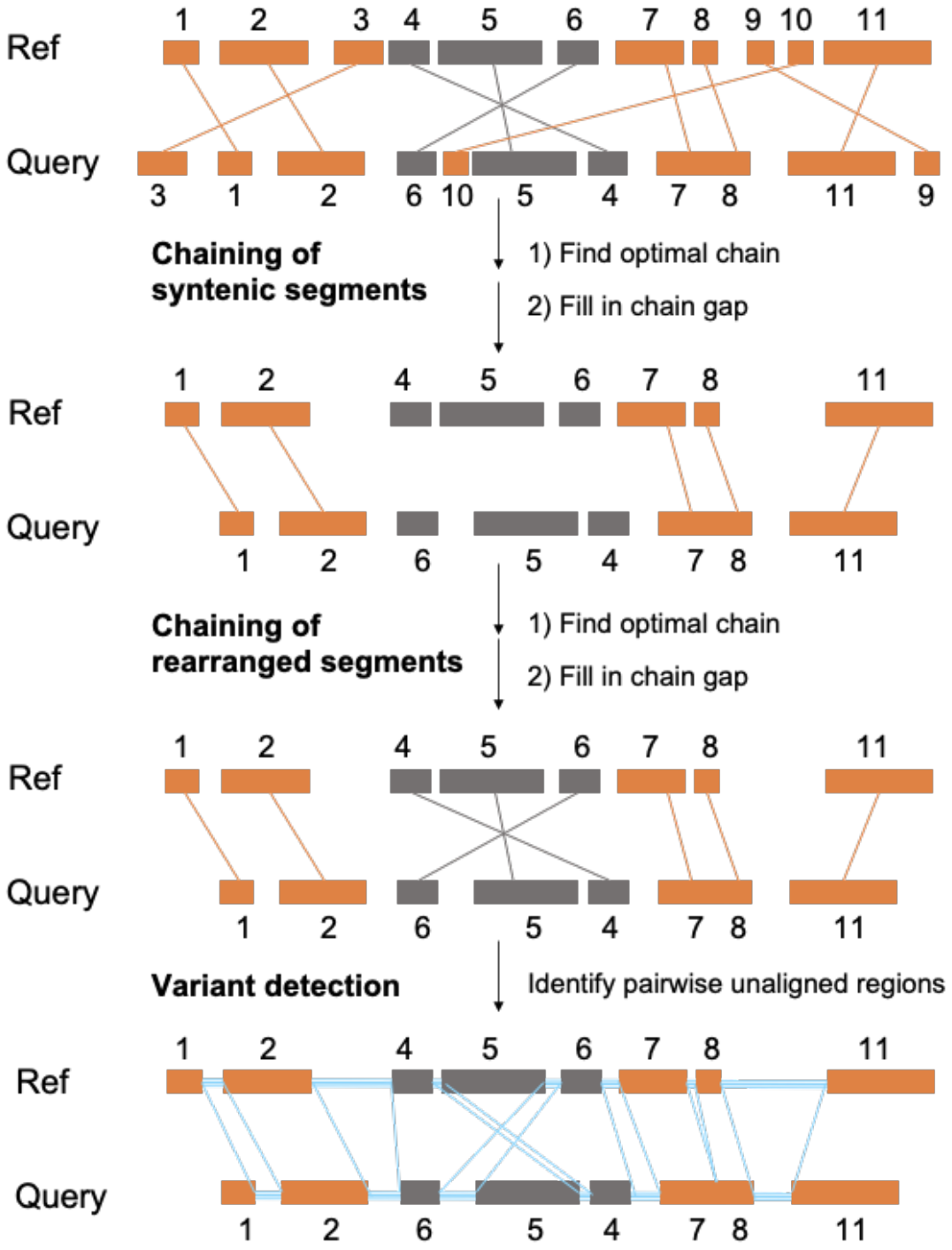


Figure S4.2. Workflow for synteny detection and structural variant characterization between reference and query genomes upon whole-chromosome alignment. Syntenic and inverted areas are respectively colored in orange and grey, and variant regions are highlighted in cyan.

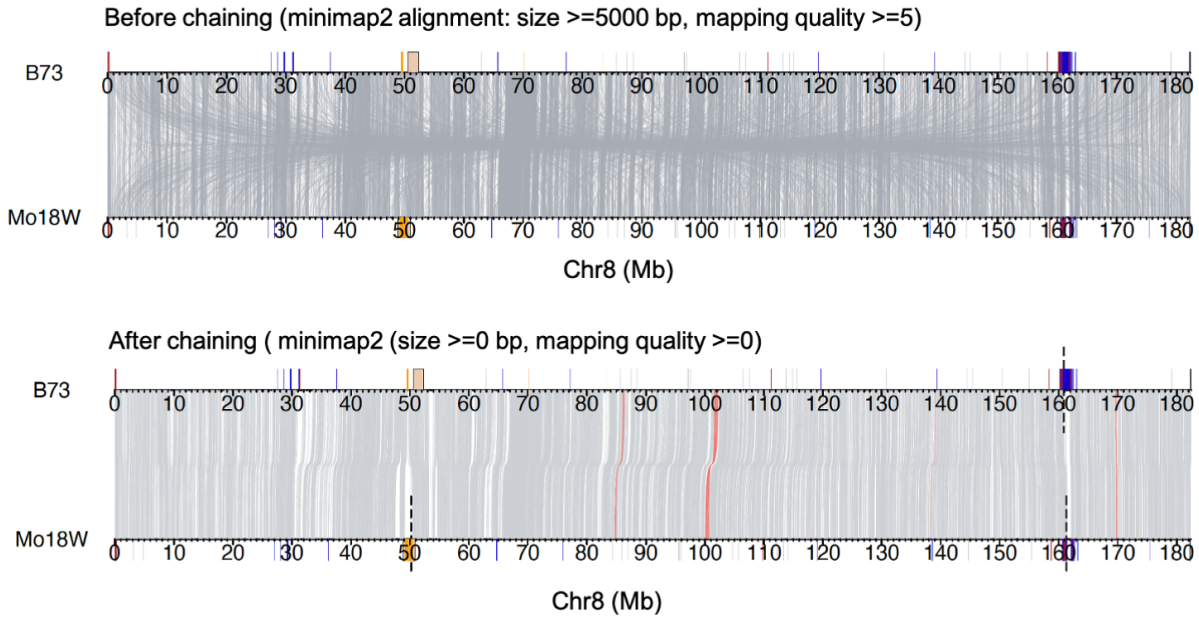


Figure S4.3. Spurious alignment removal through chaining. The upper panel exhibits the alignment between B73 and Mo18 before chaining, where blocks with a size above 5 Kb and a mapping quality higher than 5 are shown as links. The lower panel depicts alignment after chaining, where no size or mapping quality is applied. Inverted regions are highlighted in red. Centromeres and tandem repeats are annotated as Figure 1A. Dashed lines indicate 100N gaps.

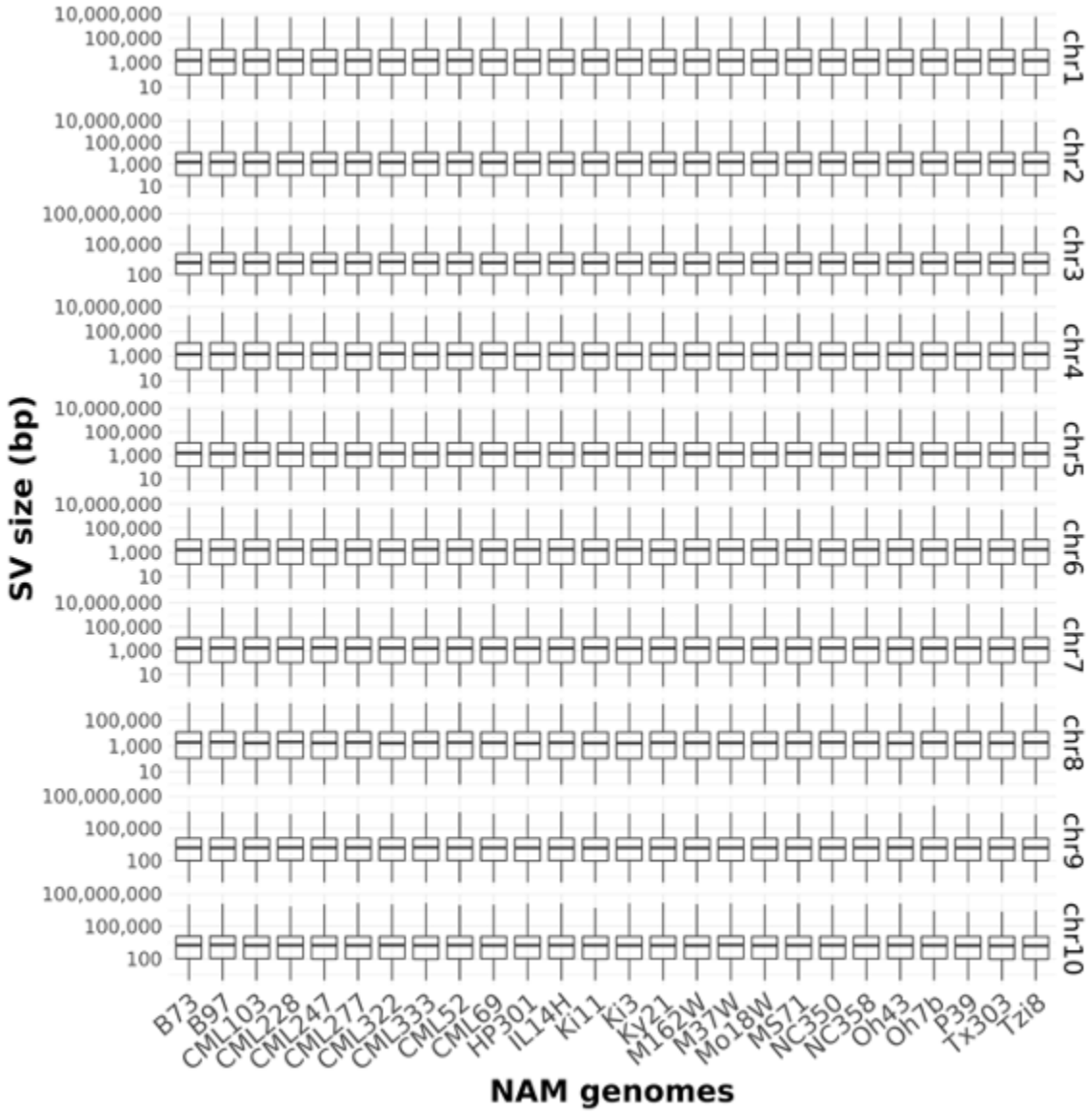


Figure S4.4. Size distribution of structural variants across NAM lines.

Figure S4.5. Examples of tandem duplications in B73. A) Simple tandem duplication on Chr3 (40Kb). B) Nested tandem duplication on Chr5 (300Kb). The Ngaps of known size indicate contig assembly failure in this area.

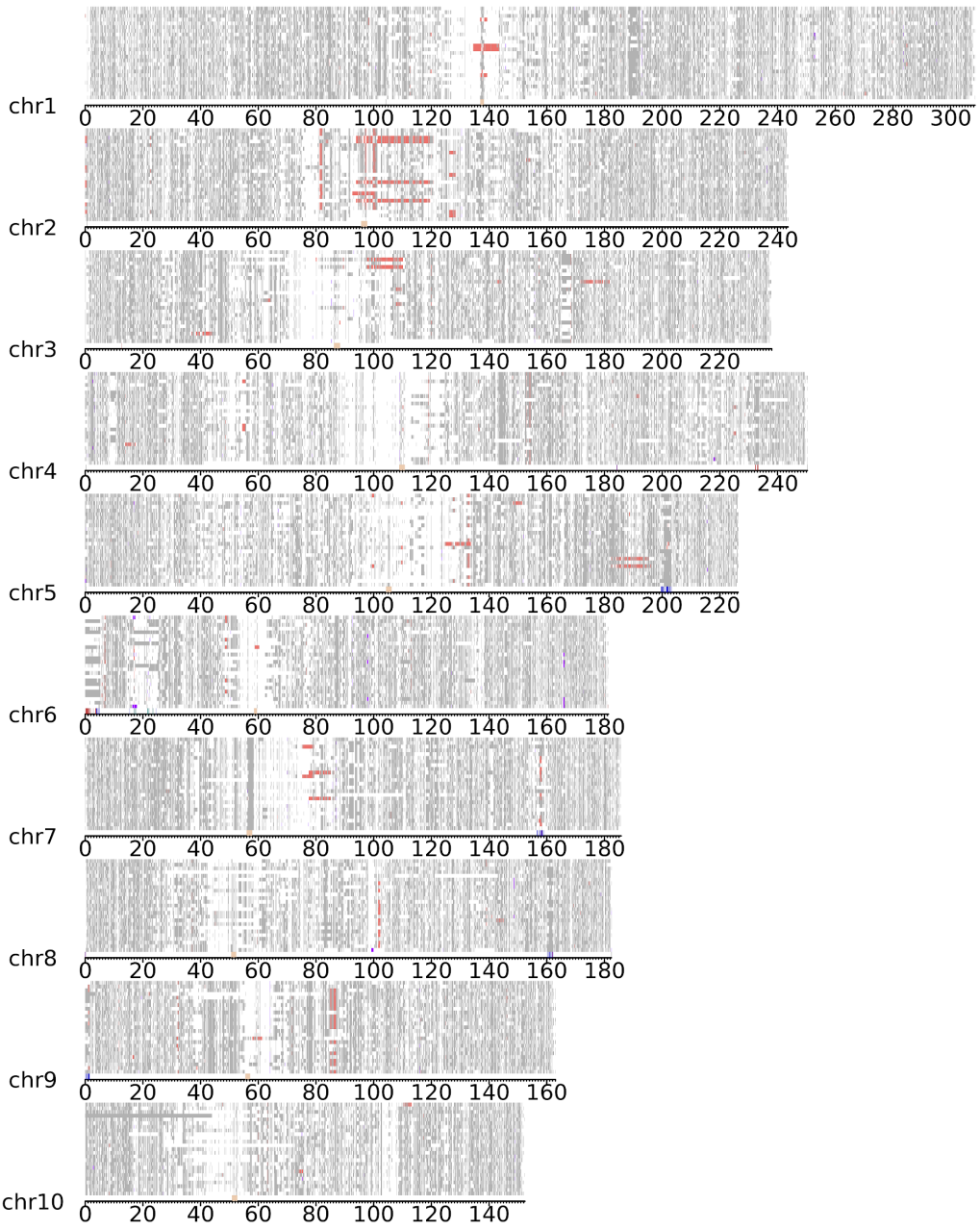
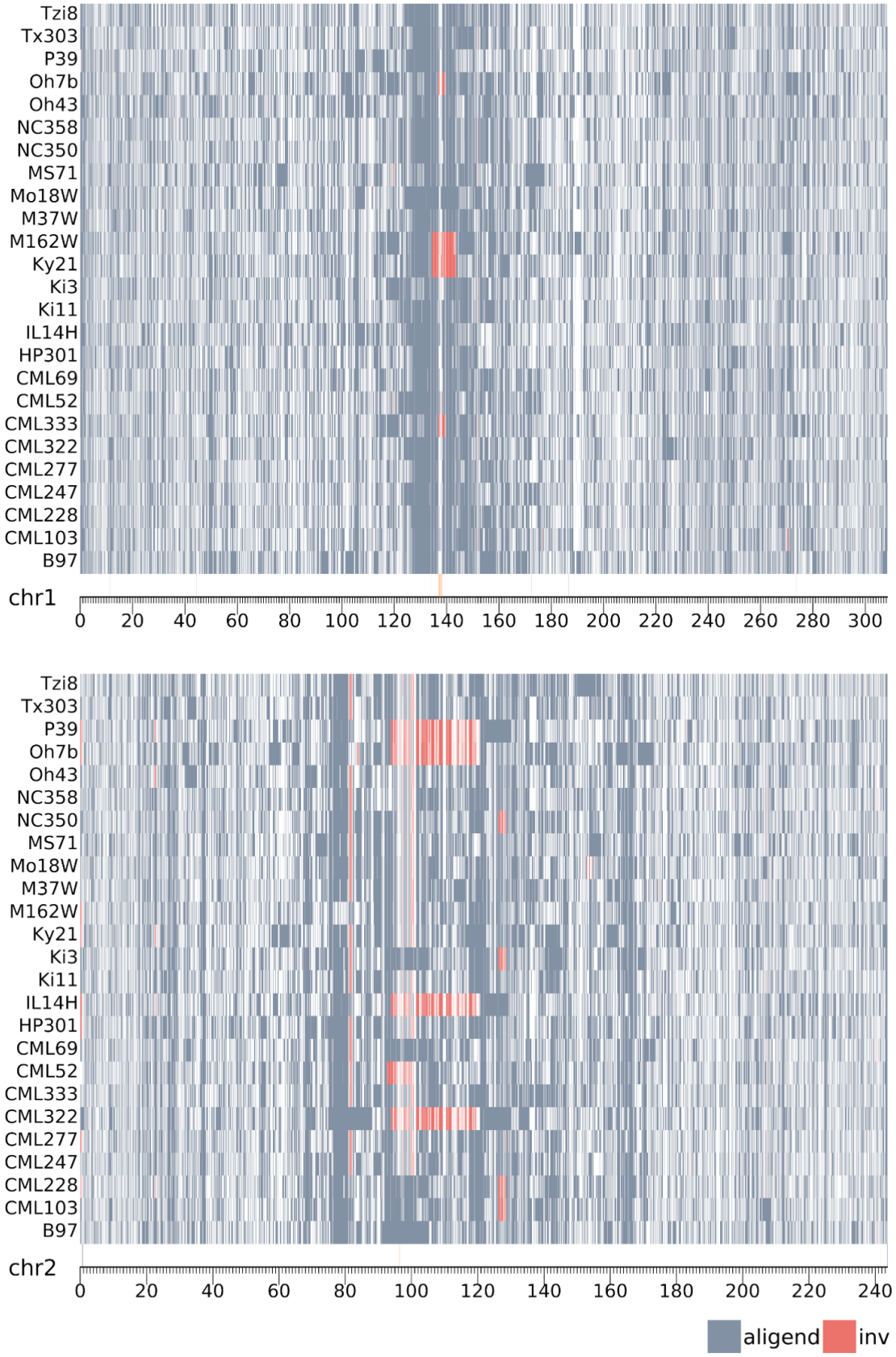
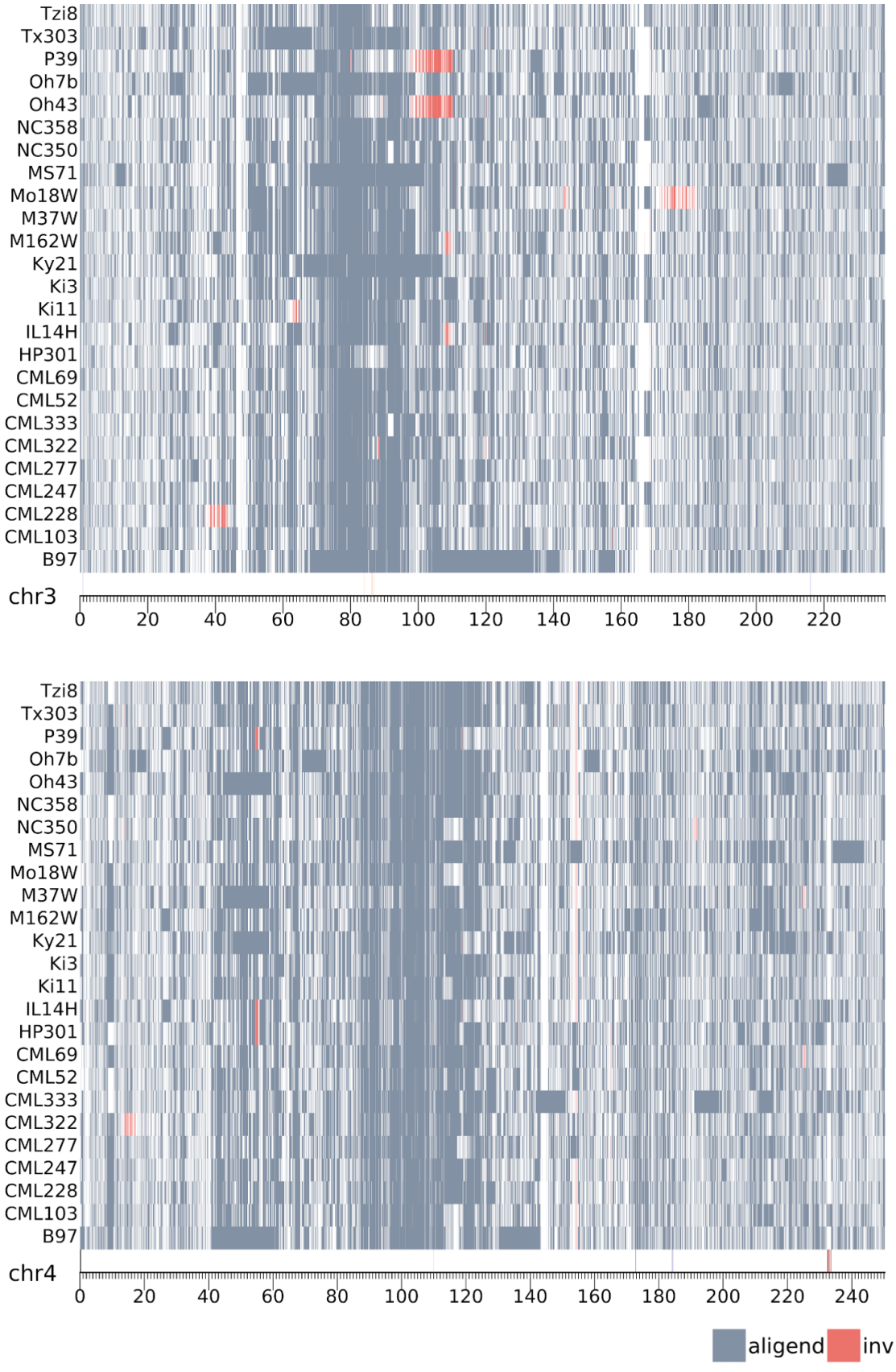
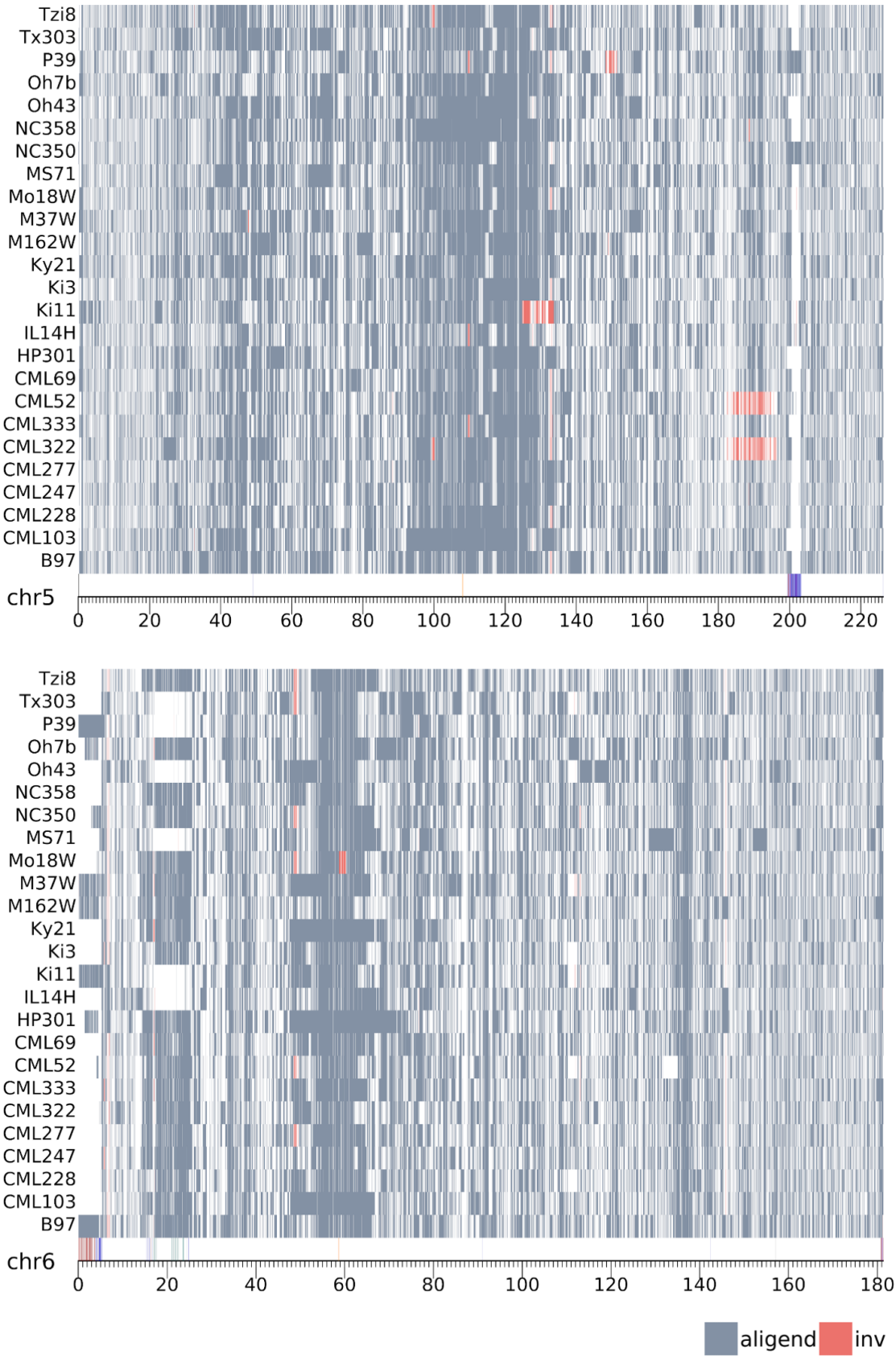
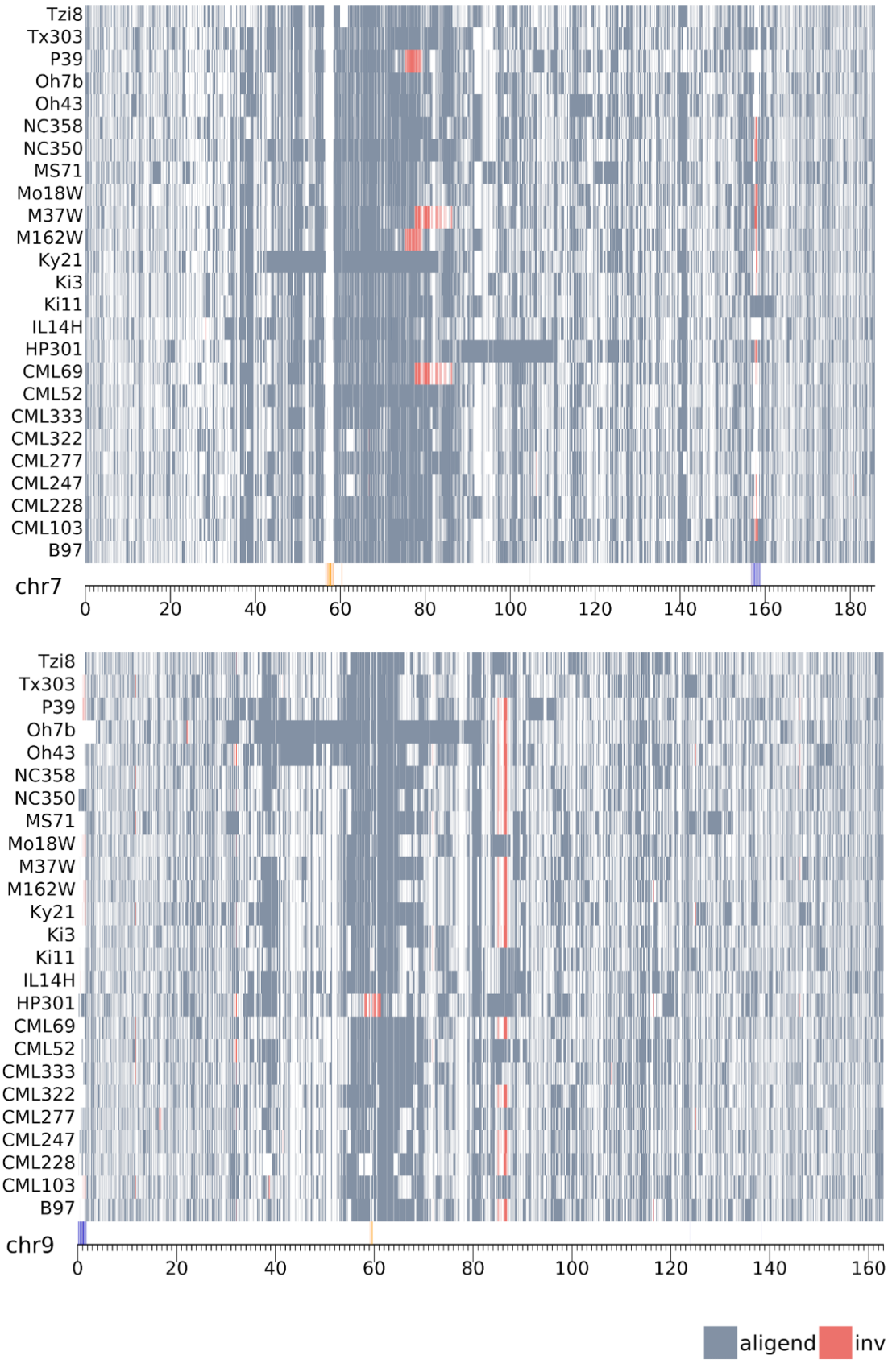


Figure S4.6. Structural variants between NAM lines and B73 reference. The bottom to top tracks represent lines: B97, CML103, CML228, CML247, CML277, CML322, CML333, CML52, CML69, HP301, IL14H, Ki11, Ki3, Ky21, M162W, M37W, Mo18W, MS71, NC350, NC358, Oh43, Oh7b, P39, Tx303, Tzi8.


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr1
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr2
0 20 40 60 80 100 120 140 160 180 200 220 240
aligend
inv


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr3
0
20
40
60
80
100
120
140
160
180
200
220
Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr4
0
20
40
60
80
100
120
140
160
180
200
220
240
aligend
inv


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr5
0
20
40
60
80
100
120
140
160
180
200
220
Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr6
0
20
40
60
80
100
120
140
160
180
aligend
inv


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr7
0
20
40
60
80
100
120
140
160
180
Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr9
0
20
40
60
80
100
120
140
160
aligend
inv

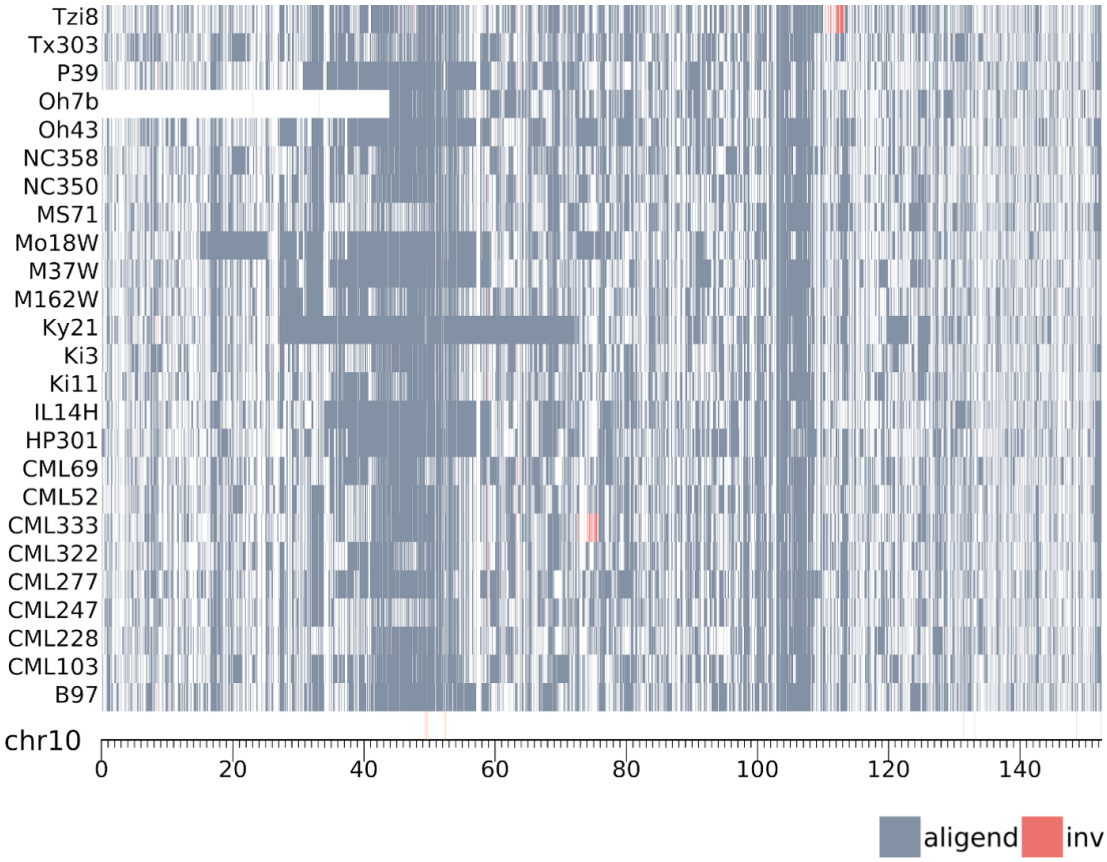
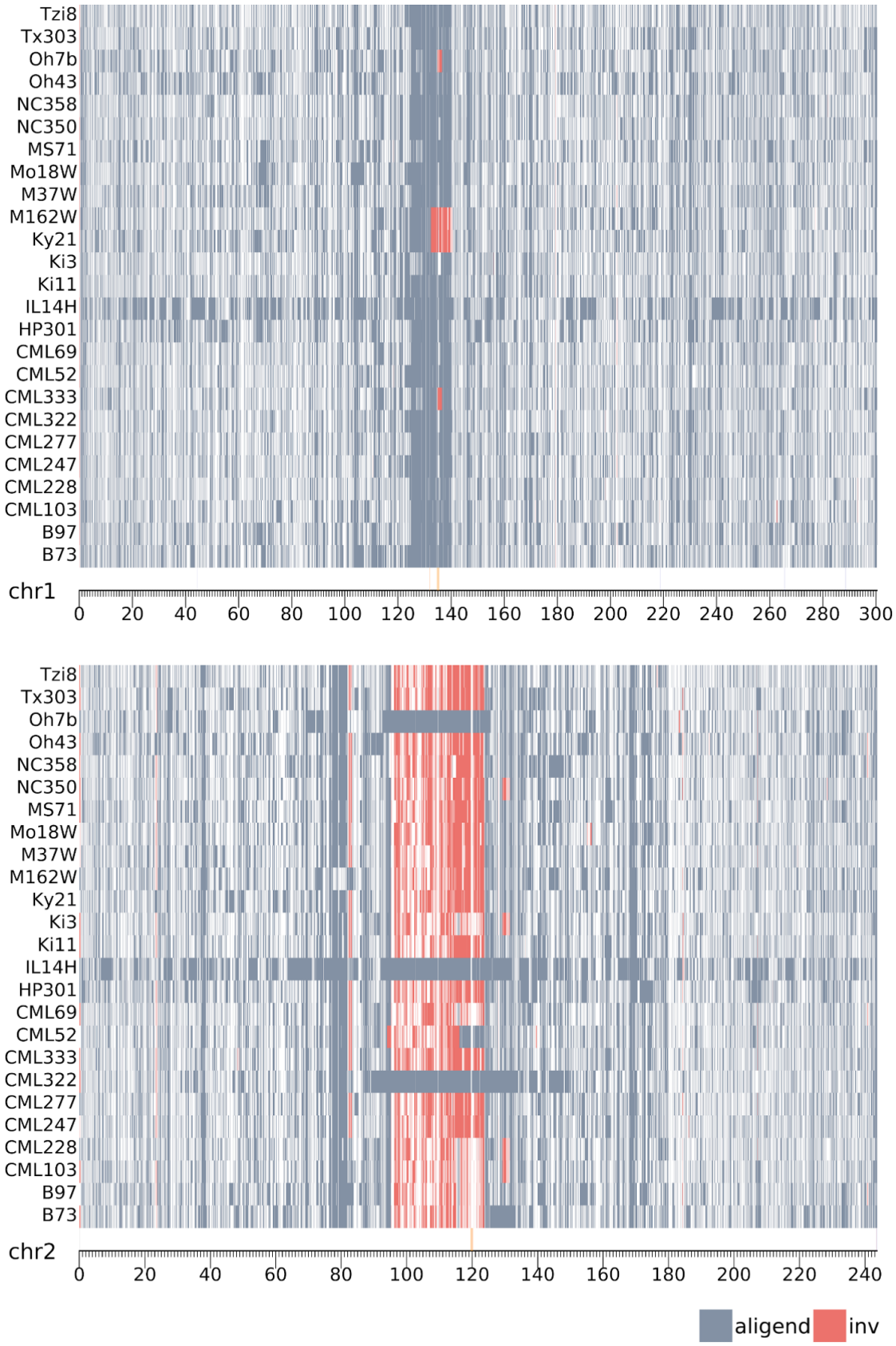
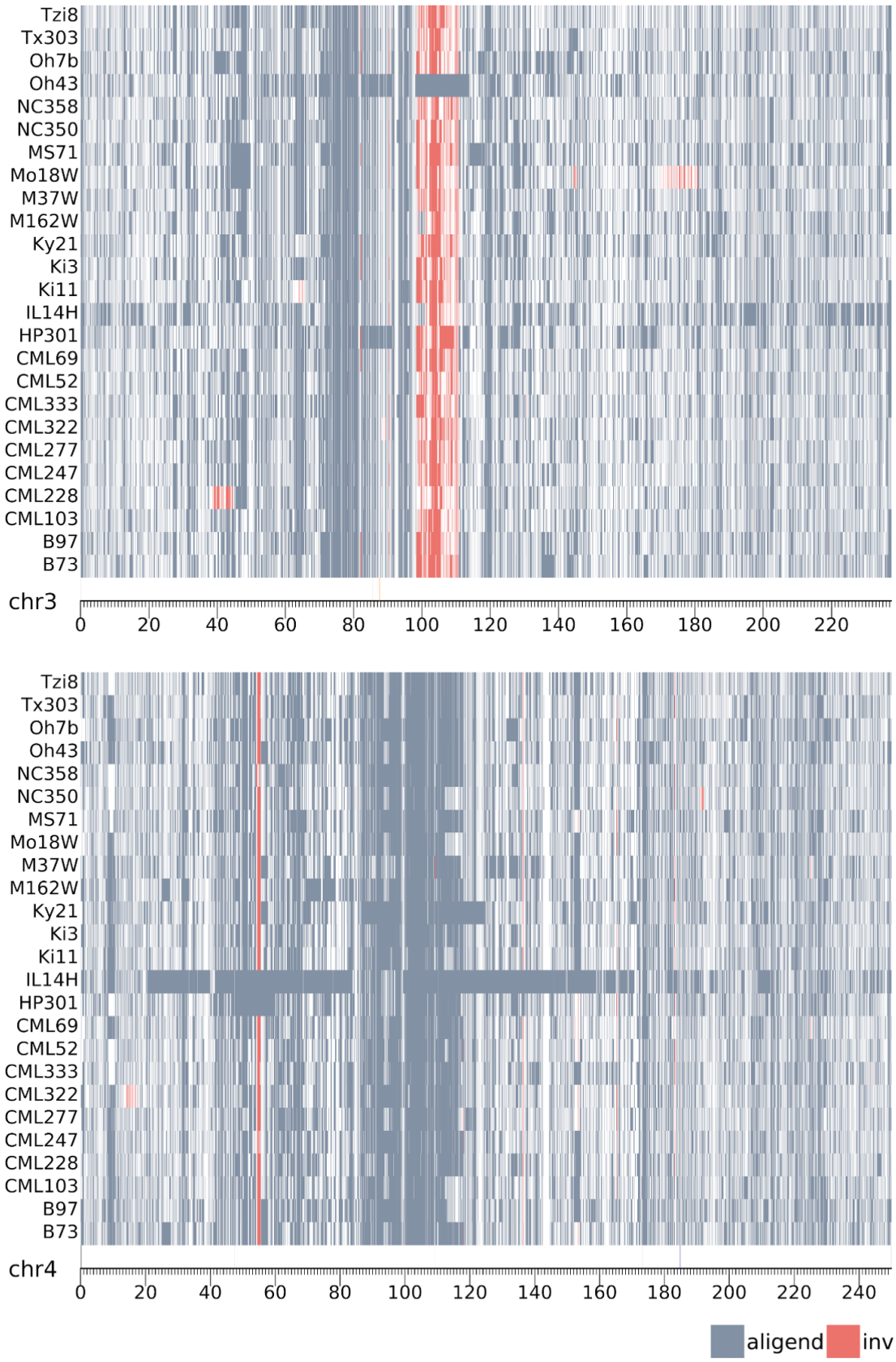
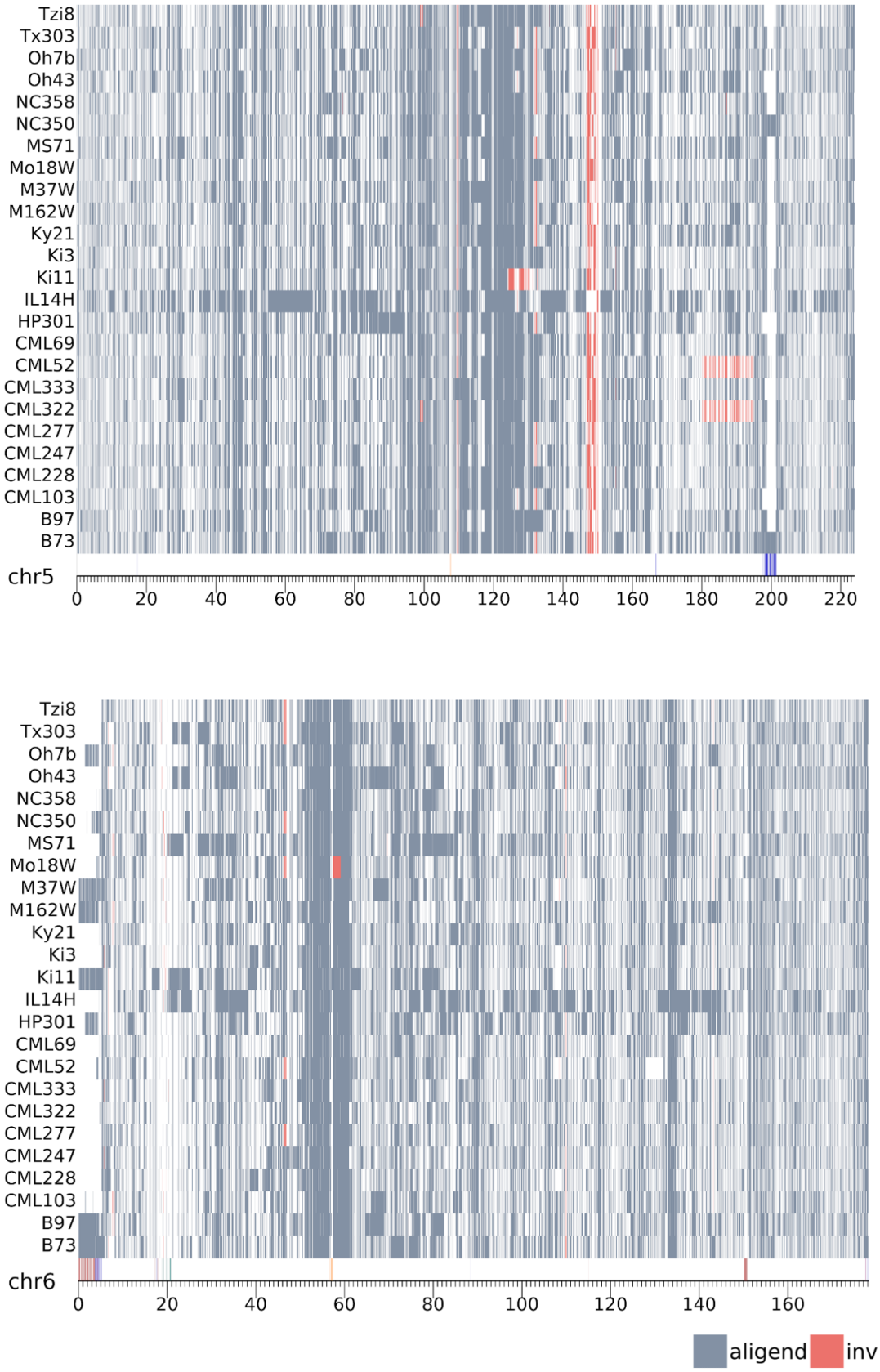
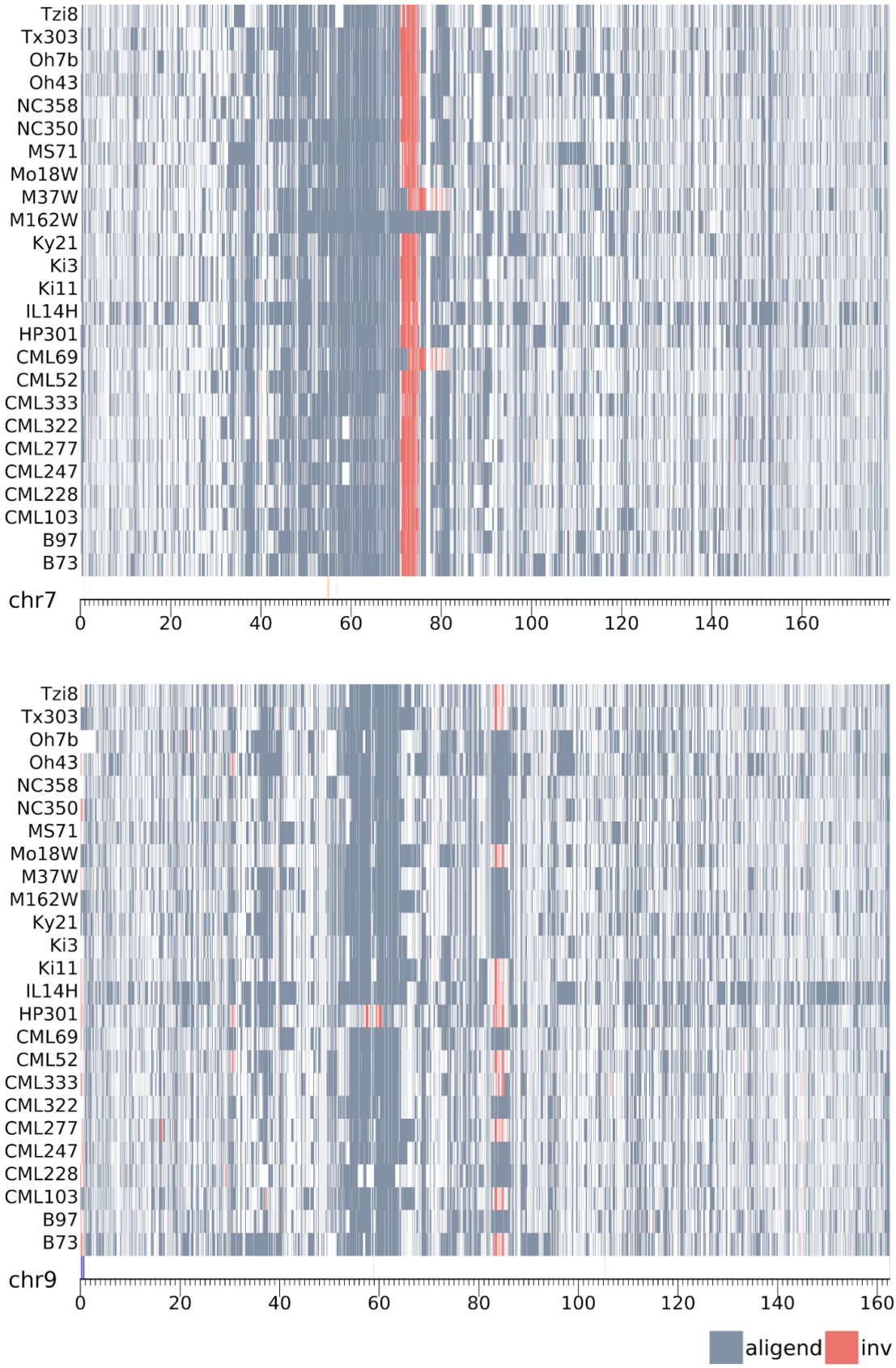


Figure S4.7. Whole-genome alignments between NAM lines and B73. Syntenic aligned regions are colored grey and Inverted segments are highlighted in red. Centromere and repeats are displayed in bottom track, and annotated as Figure 1A.


Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr1
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr2
0 20 40 60 80 100 120 140 160 180 200 220 240
aligend
inv


Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr3
0 20 40 60 80 100 120 140 160 180 200 220
Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr4
0 20 40 60 80 100 120 140 160 180 200 220 240
aligend
inv


Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr5
0
20
40
60
80
100
120
140
160
180
200
220
Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr6
0
20
40
60
80
100
120
140
160
aligend
inv


Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr7
0
20
40
60
80
100
120
140
160
Tzi8
Tx303
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
B73
chr9
0
20
40
60
80
100
120
140
160
aligend
inv

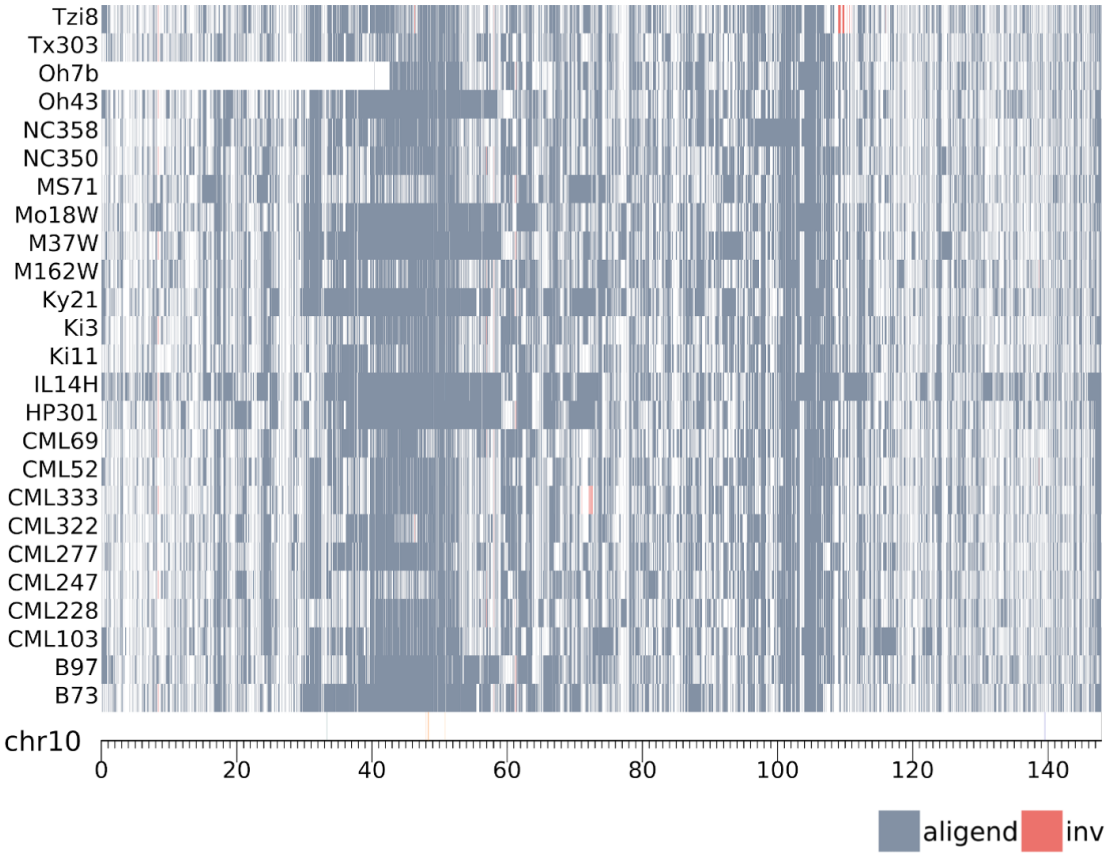


Figure S4.8. Whole-genome alignments between 25 lines and P39. Syntenic aligned regions are colored grey and Inverted segments are highlighted in red. Centromere and repeats are displayed in bottom track, and annotated as Figure 1A.

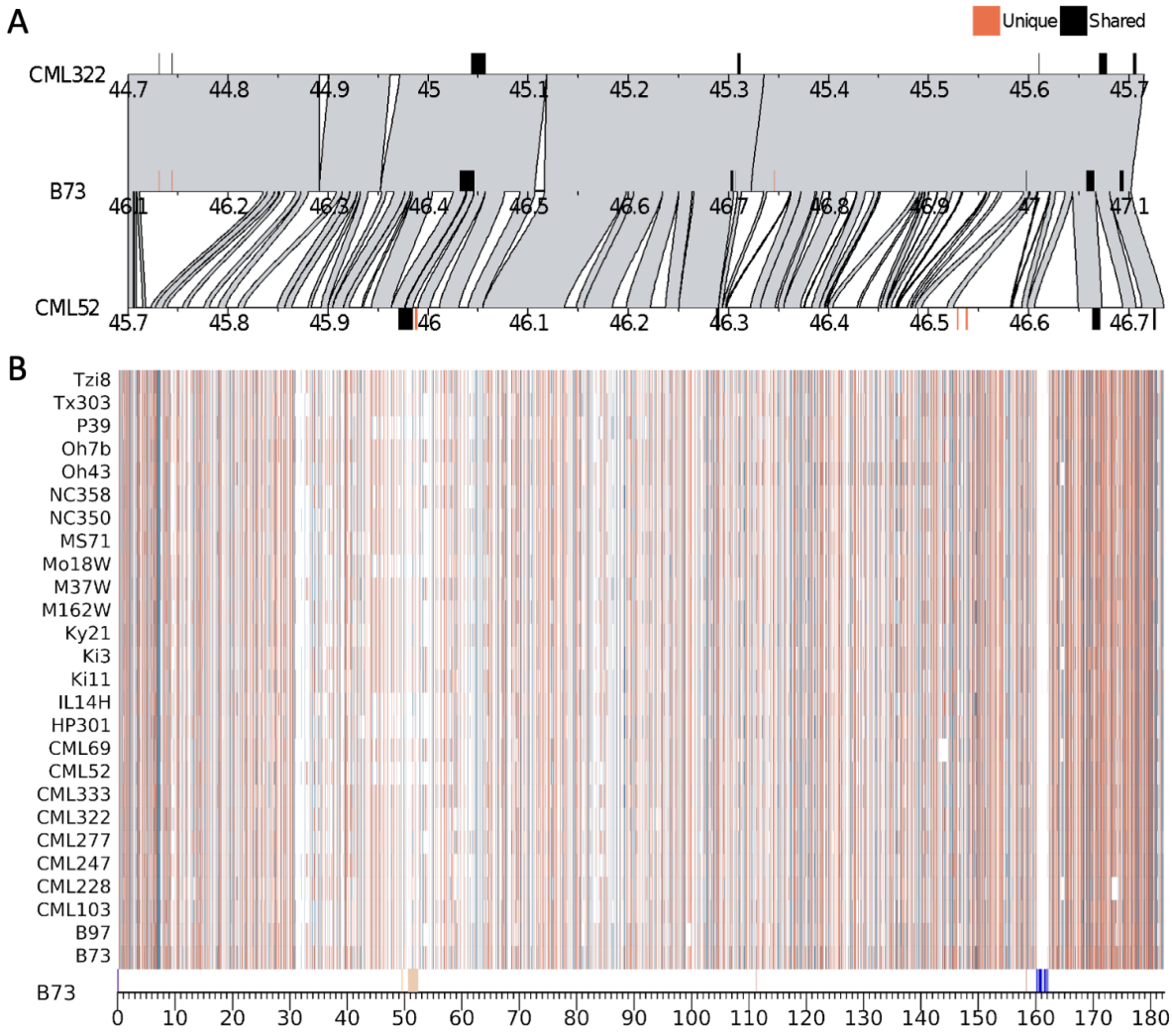
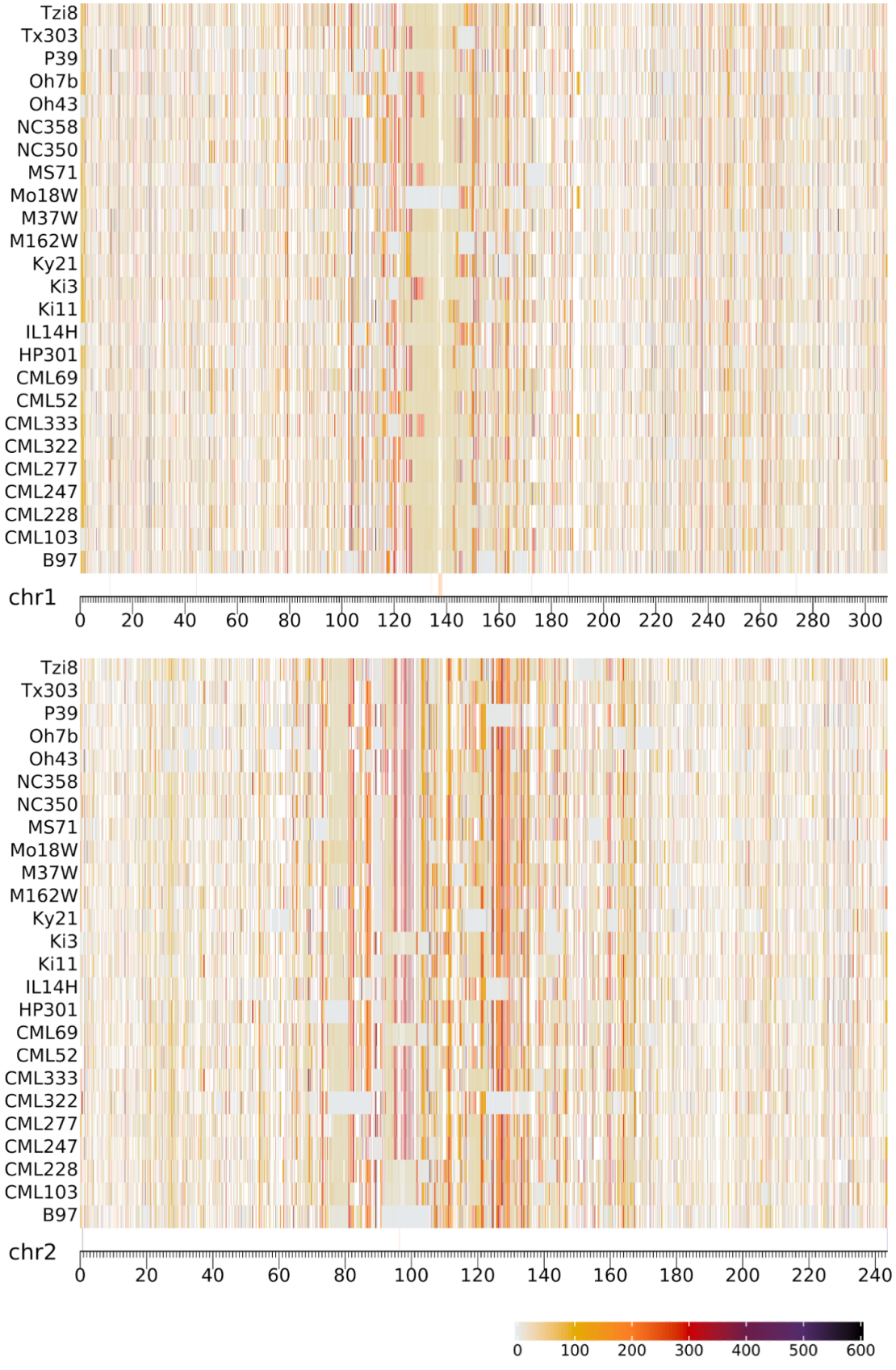
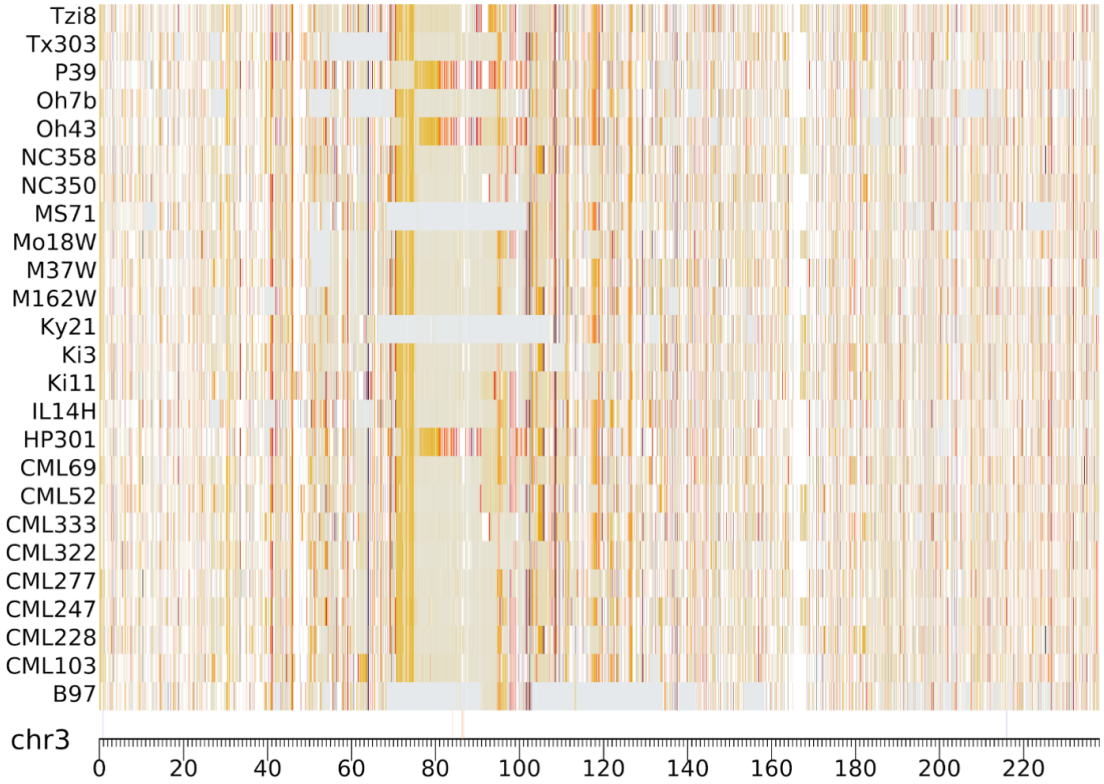
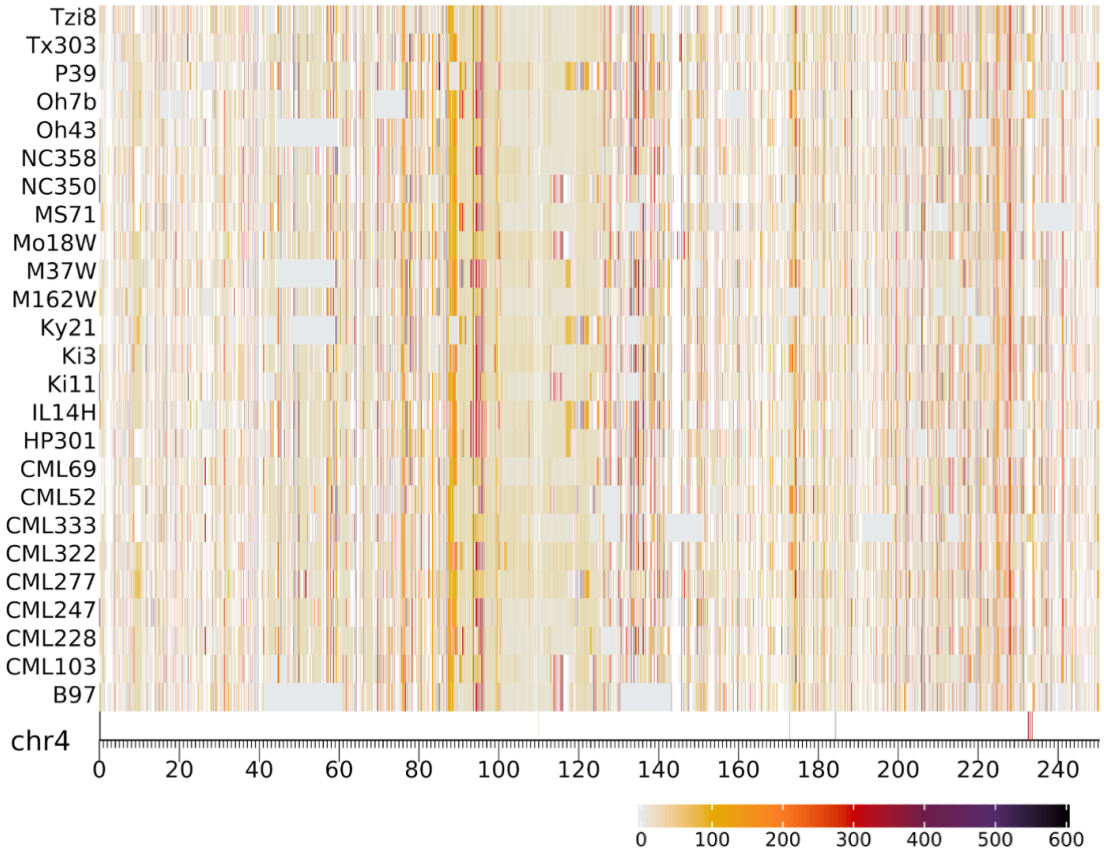
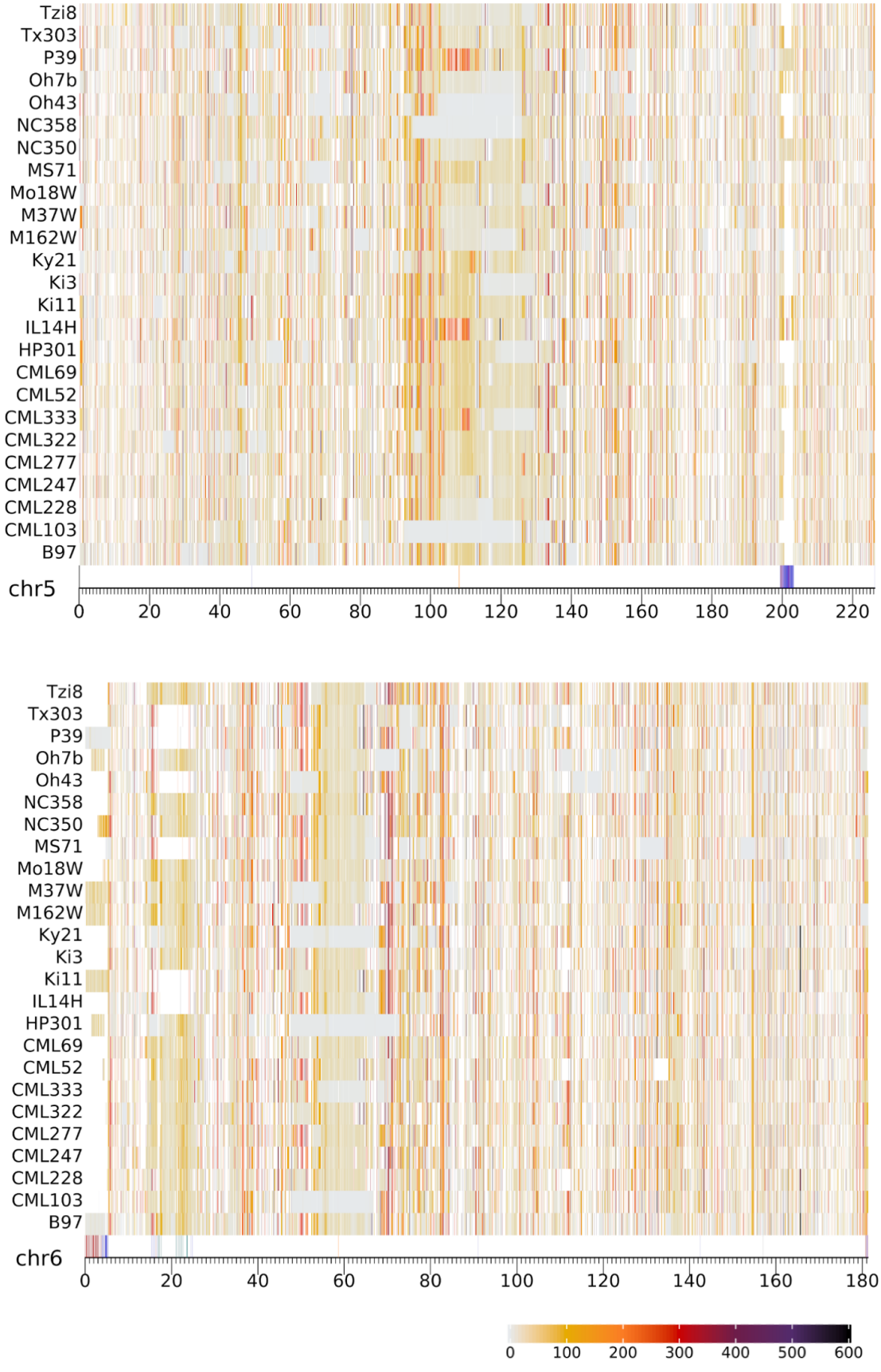


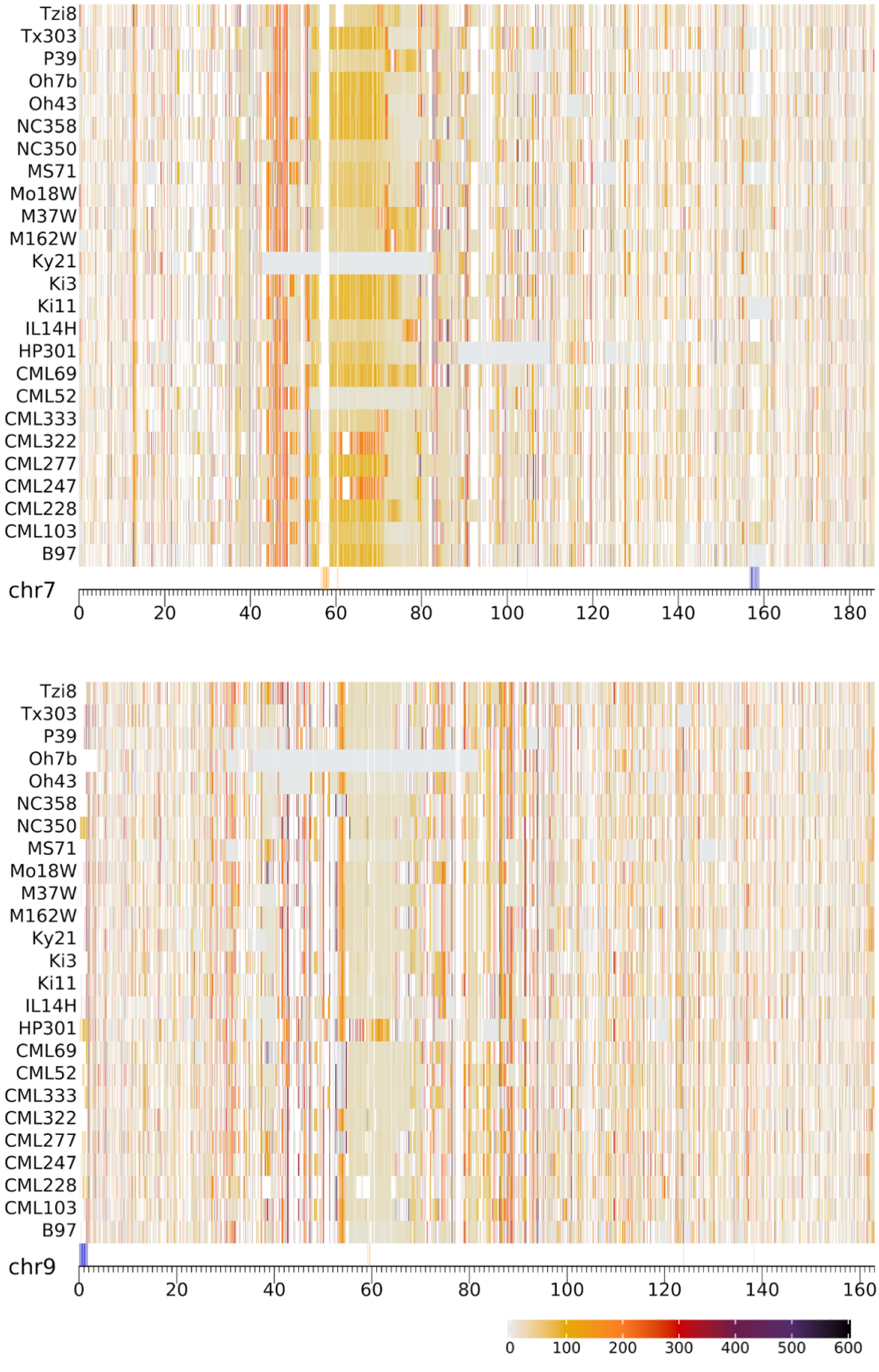
Figure S4.9. Distribution of genes in the pericentromeric haploblock on chromosome 8. A) Pairwise alignments between CML322, CML52 and B73 from region 46.1Mb to 47.1 Mb (B73 coordinates). Syntenic and non-syntenic genes are respectively colored in orange and black. D) Gene synteny between NAM lines and B73 across the whole chromosome 8. Genes in forward orientation are highlighted in red while those of reverse orientation are colored with black.


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr1
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr2
0 20 40 60 80 100 120 140 160 180 200 220 240
0 100 200 300 400 500 600


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr3
0
20
40
60
80
100
120
140
160
180
200
220


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr4
0
20
40
60
80
100
120
140
160
180
200
220
240
0
100
200
300
400
500
600


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr5
0
20
40
60
80
100
120
140
160
180
200
220
chr6
0
100
200
300
400
500
600


Tzi8
Tx303
P39
Oh7b
Oh43
NC358
NC350
MS71
Mo18W
M37W
M162W
Ky21
Ki3
Ki11
IL14H
HP301
CML69
CML52
CML333
CML322
CML277
CML247
CML228
CML103
B97
chr7
0
20
40
60
80
100
120
140
160
180
chr9
0
100
200
300
400
500
600

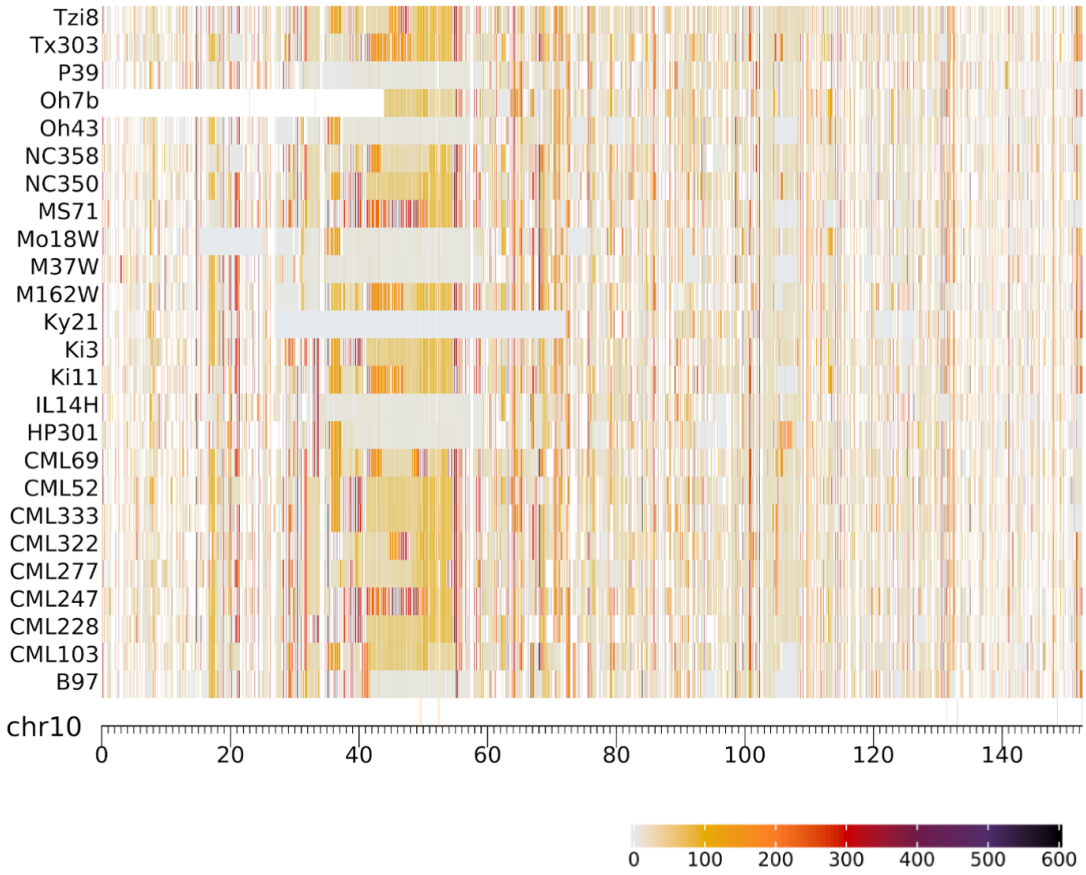
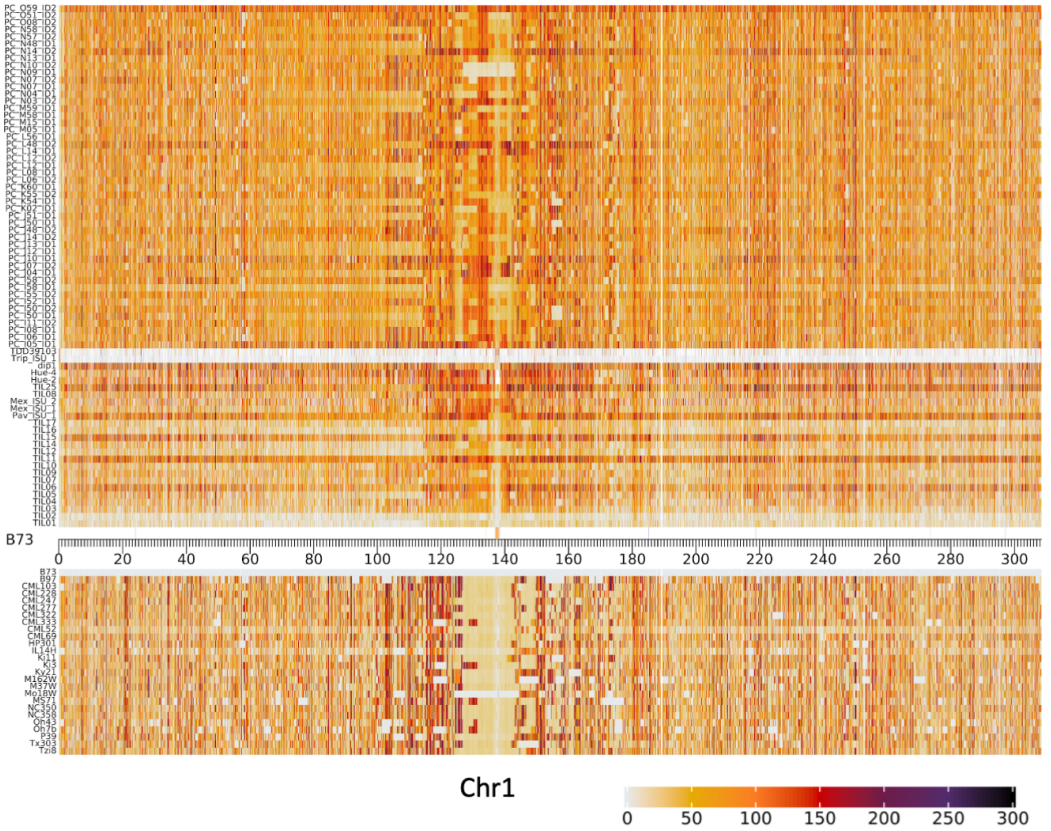
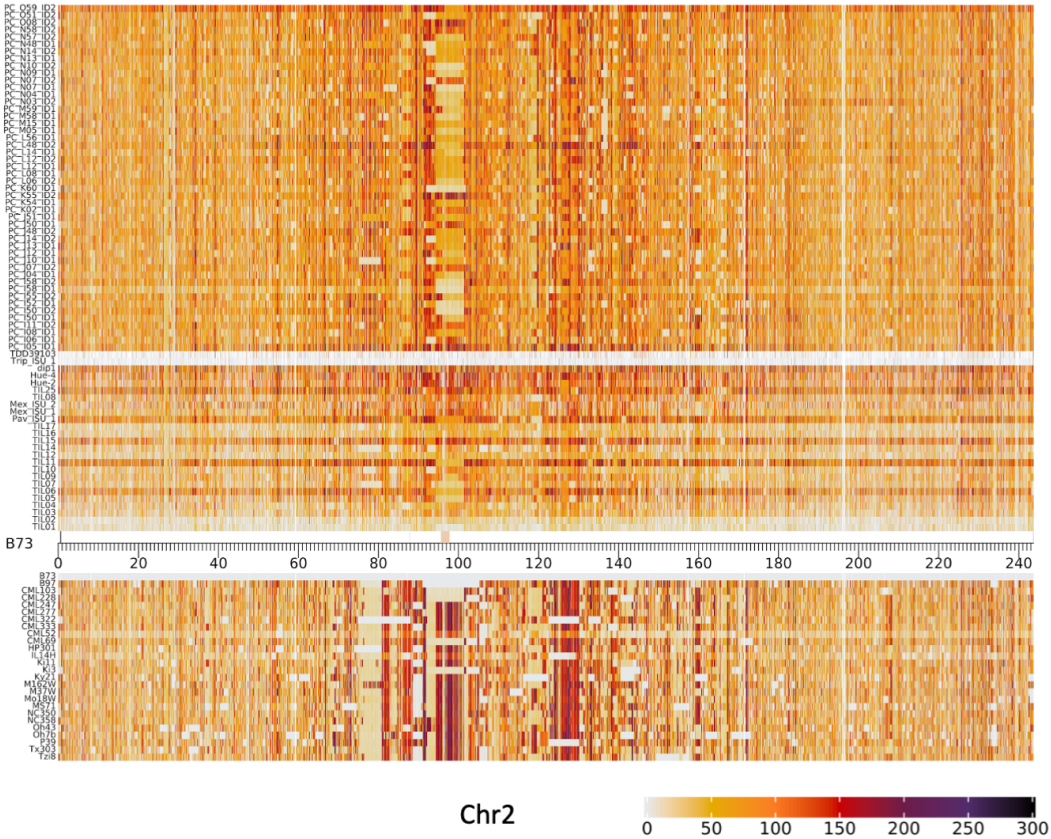
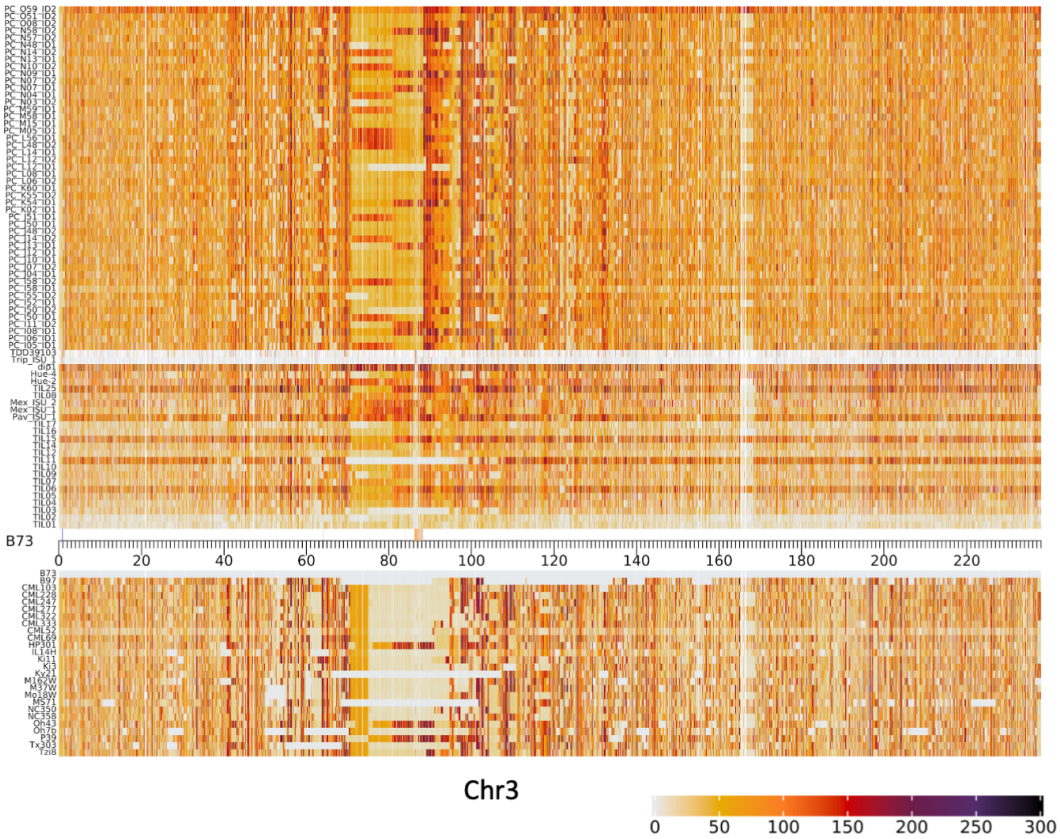
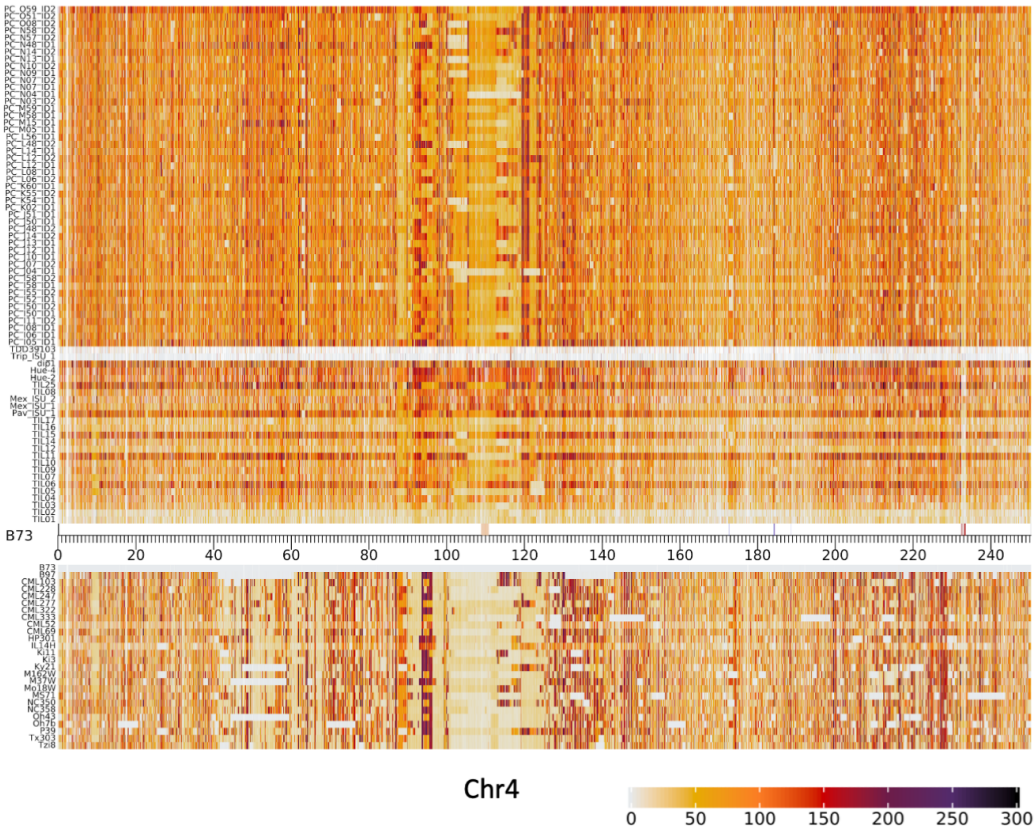


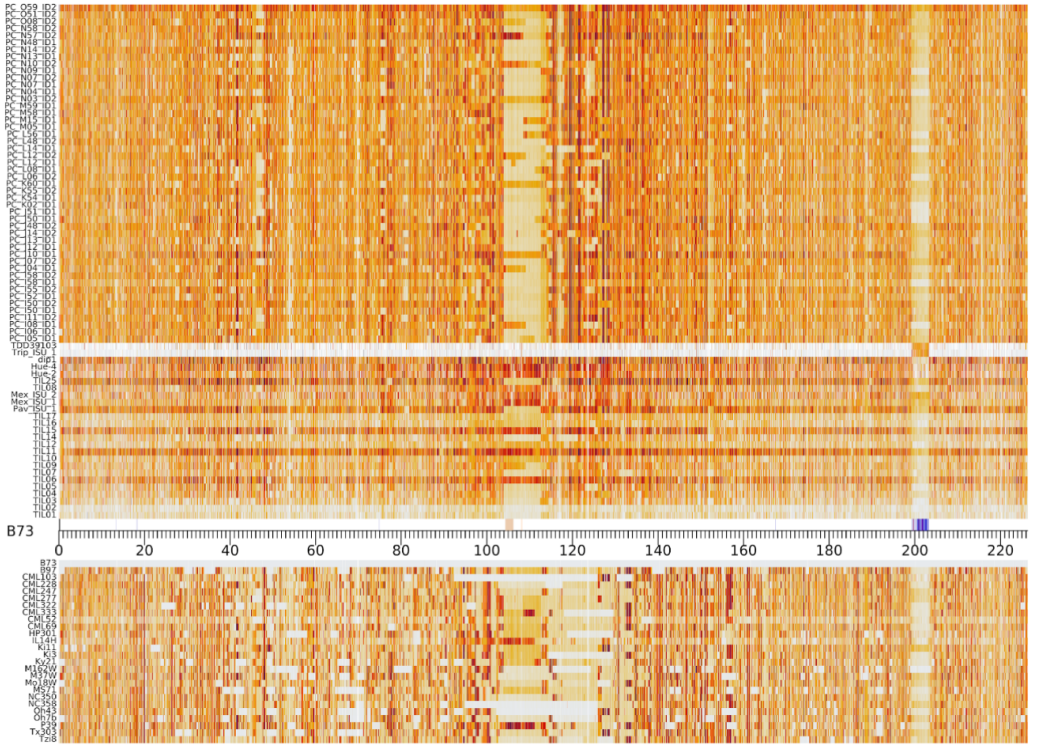
Figure S4.10. Divergence time estimated with syntenic SNPs between B73 and NAM lines across the whole genome. Divergence time (K years) between NAM lines and B73, estimated with syntenic SNPs. Maximum divergence time was set as 0.6 million years in the heatmap, where the lowest divergence colored as light grey and highest divergence as black. White spaces depict unaligned regions between query genome and B73. Tandem repeats are shown in the bottom track and annotated as A).


B73
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
Chr1
0 50 100 150 200 250 300

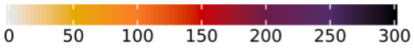

B73
0 20 40 60 80 100 120 140 160 180 200 220 240
Chr2
0 50 100 150 200 250 300

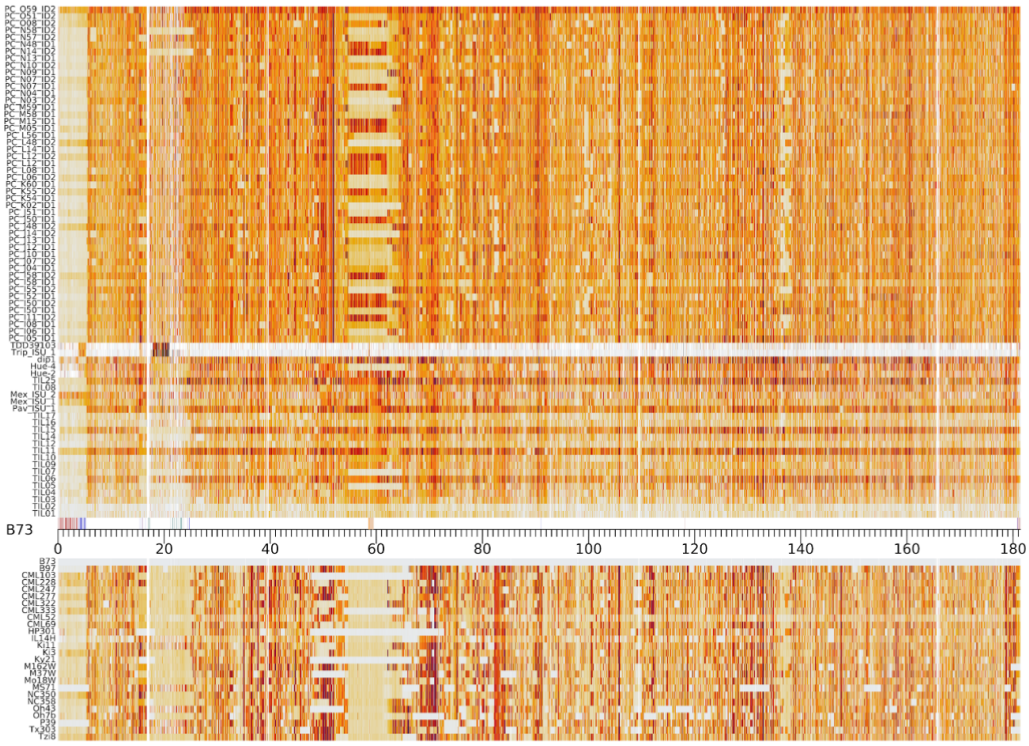

B73
0 20 40 60 80 100 120 140 160 180 200 220
Chr3
0 50 100 150 200 250 300


B73
0 20 40 60 80 100 120 140 160 180 200 220 240
Chr4
0 50 100 150 200 250 300

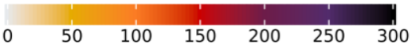

B73
0
20
40
60
80
100
120
140
160
180
200
220

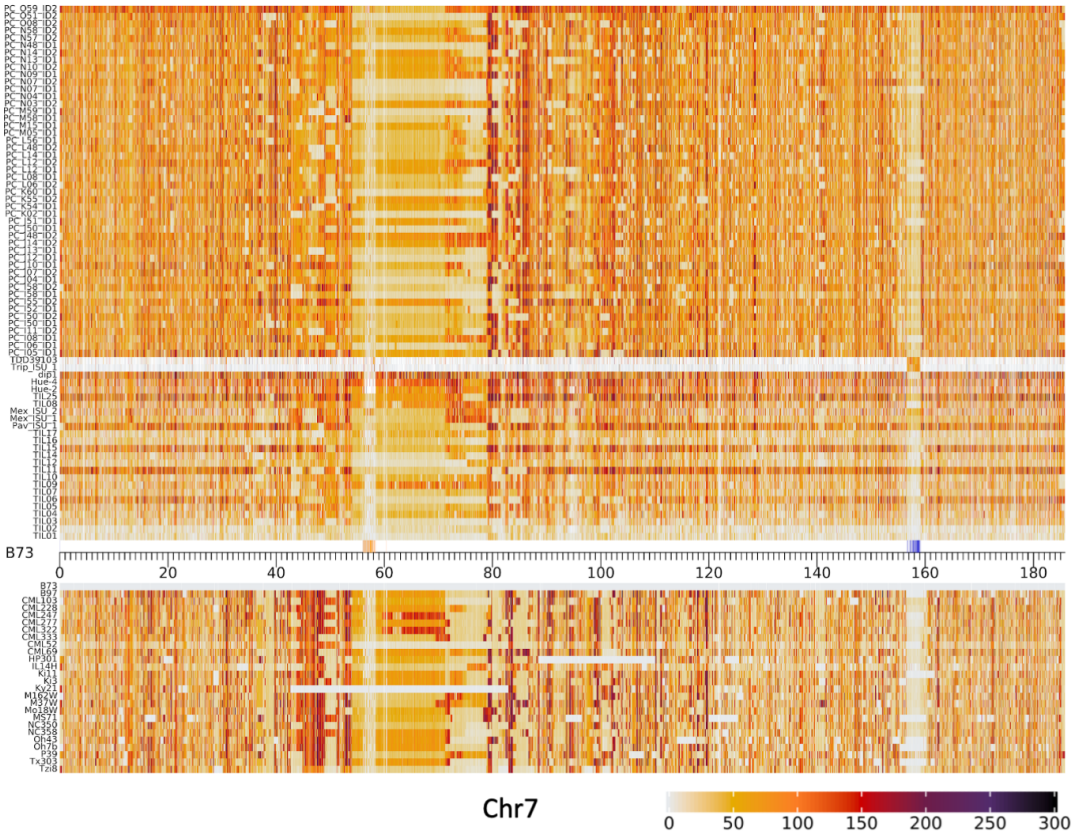
Chr5

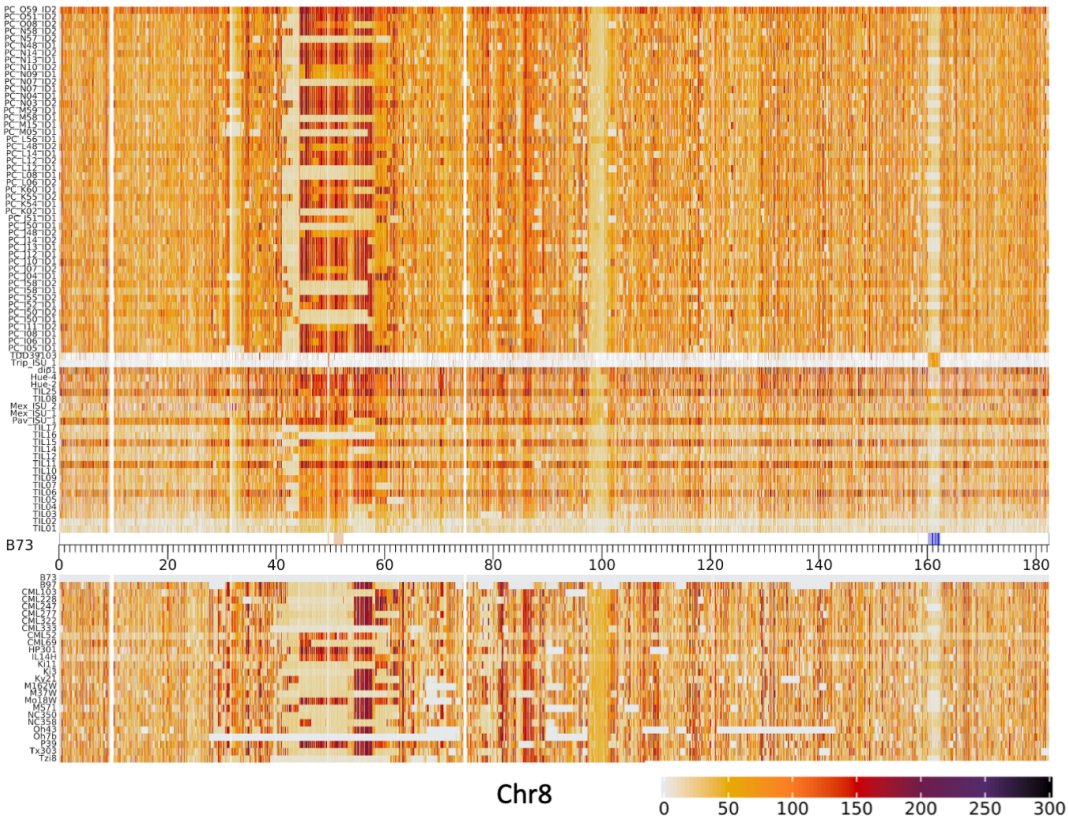

0
50
100
150
200
250
300

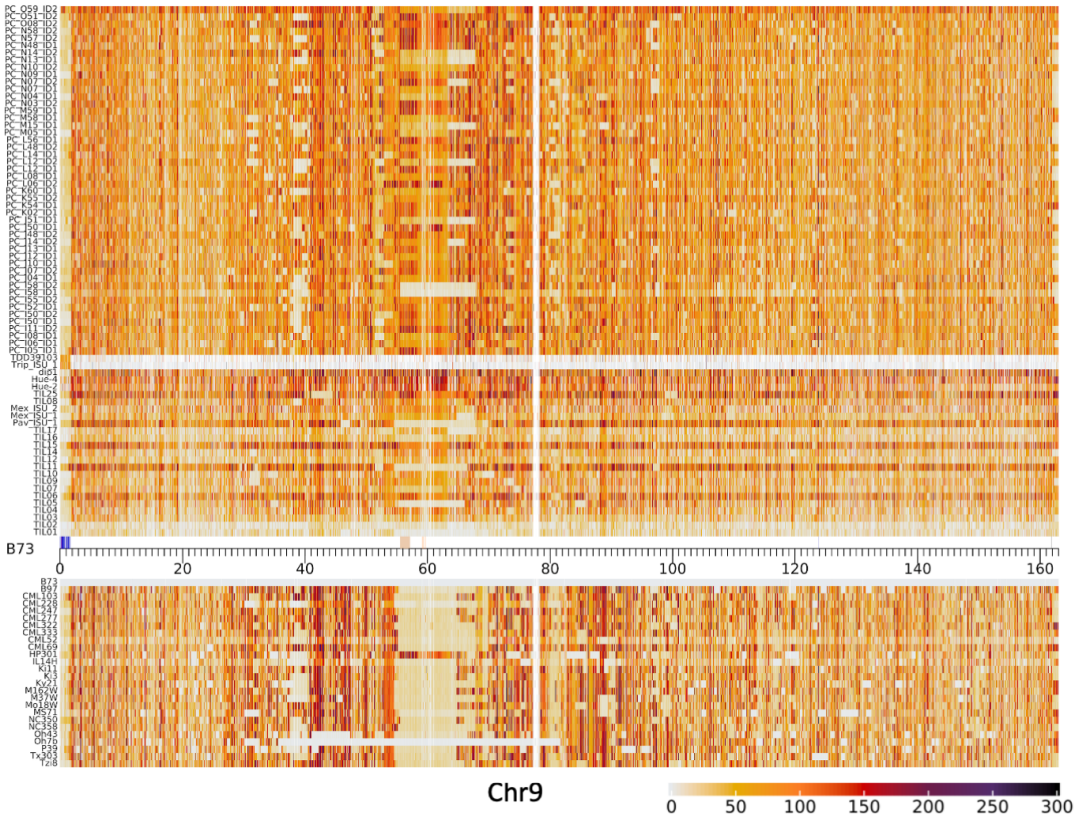

B73
0
20
40
60
80
100
120
140
160
180

Chr6


0
50
100
150
200
250
300


B73
0
20
40
60
80
100
120
140
160
180
Chr7
0
50
100
150
200
250
300


B73
0
20
40
60
80
100
120
140
160
180
Chr8
0
50
100
150
200
250
300


B73
0
20
40
60
80
100
120
140
160
Chr9
0
50
100
150
200
250
300

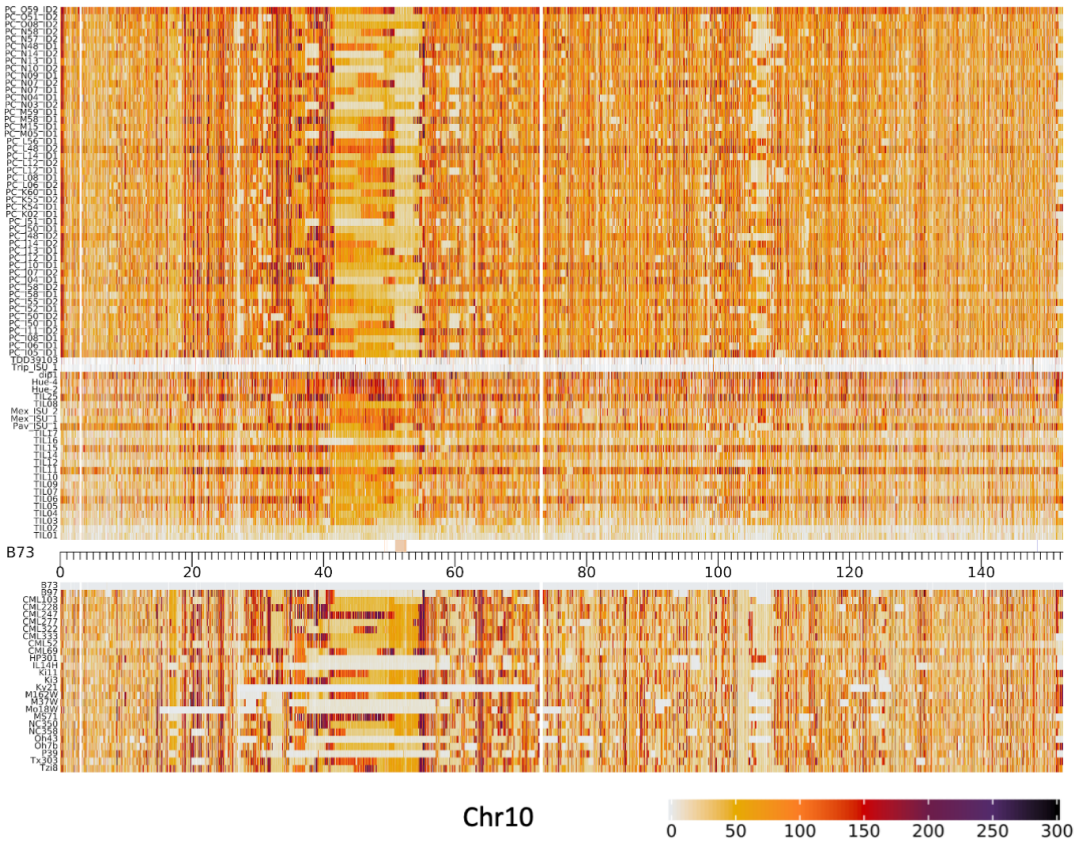

B73
0
20
40
60
80
100
120
140
Chr10
0
50
100
150
200
250
300

Figure S4.11. Divergence time (K years) estimated with short-reads mapping across 10 chromosomes. Maximum divergence time was set as 0.6 million years in the heatmap, where the lowest divergence colored as light grey and highest divergence as black. The order of samples from bottom to up is: 25 maize NAM lines, 15 *Parviglumis* lines from the HapMap II project (TIL01, TIL02, TIL03, TIL04, TIL05, TIL06, TIL07, TIL09, TIL10, TIL11, TIL14, TIL15, Pav_1), 4 *Mexicana* lines (Mex1, Mex2, ,TIL08, TIL25), 2 *Huehuetageneis* samples (Hue2, Hue4*) , 1 Diploperennis* (dip1) and 2 *Tripsacum* lines (Trip1, TDD39103), and 48 *Parviglumis* lines from Palmar Chico in Balsas river of Mexico.

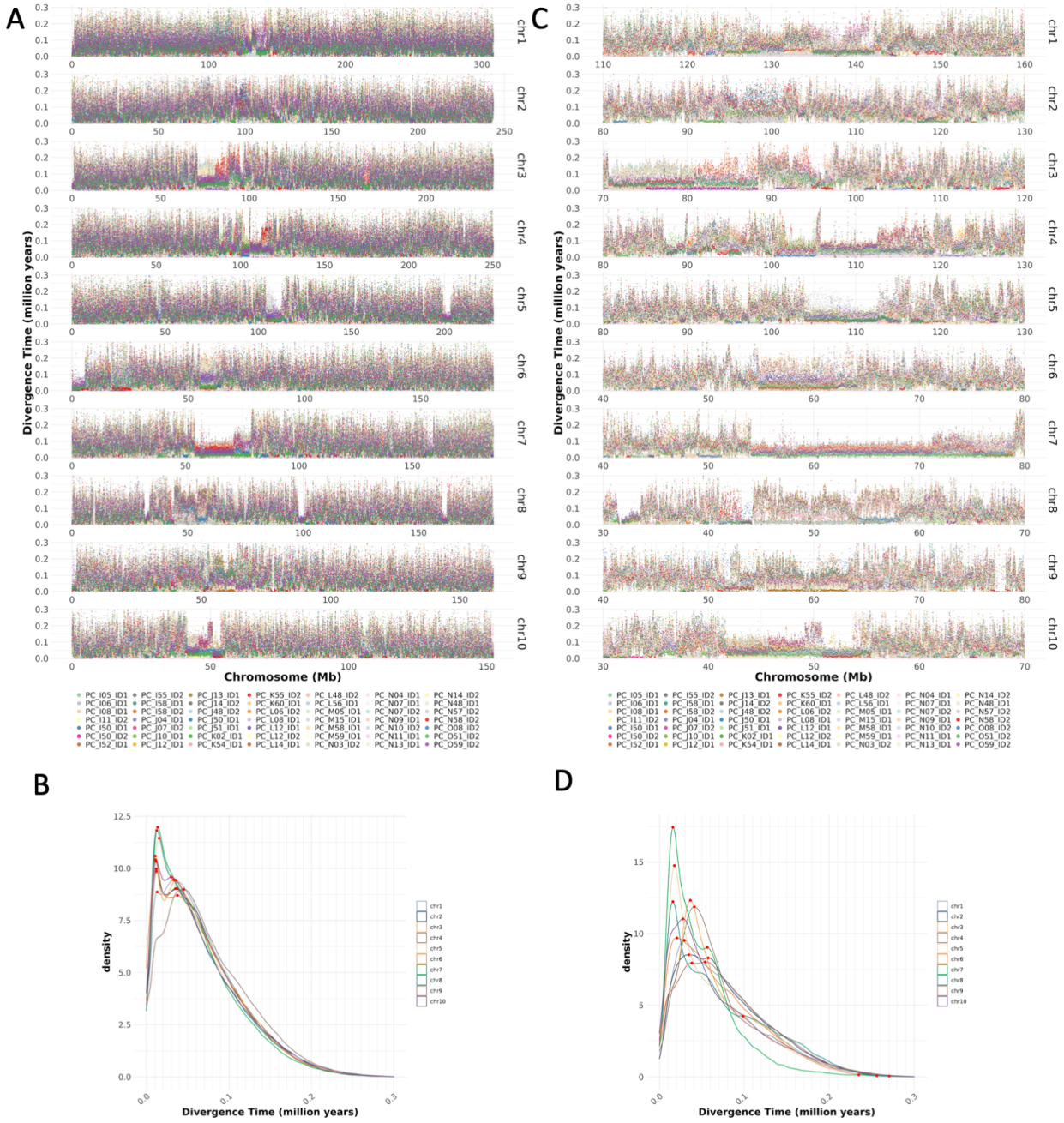


Figure S4.12. Evolutionary strata among 48 *parviglumis* lines (Palmar Chico). A) Divergence time across the whole genome. B) Density plot of divergence time in A). C) Divergence time over pericentromeric regions. D) Density plot of divergence time in C).

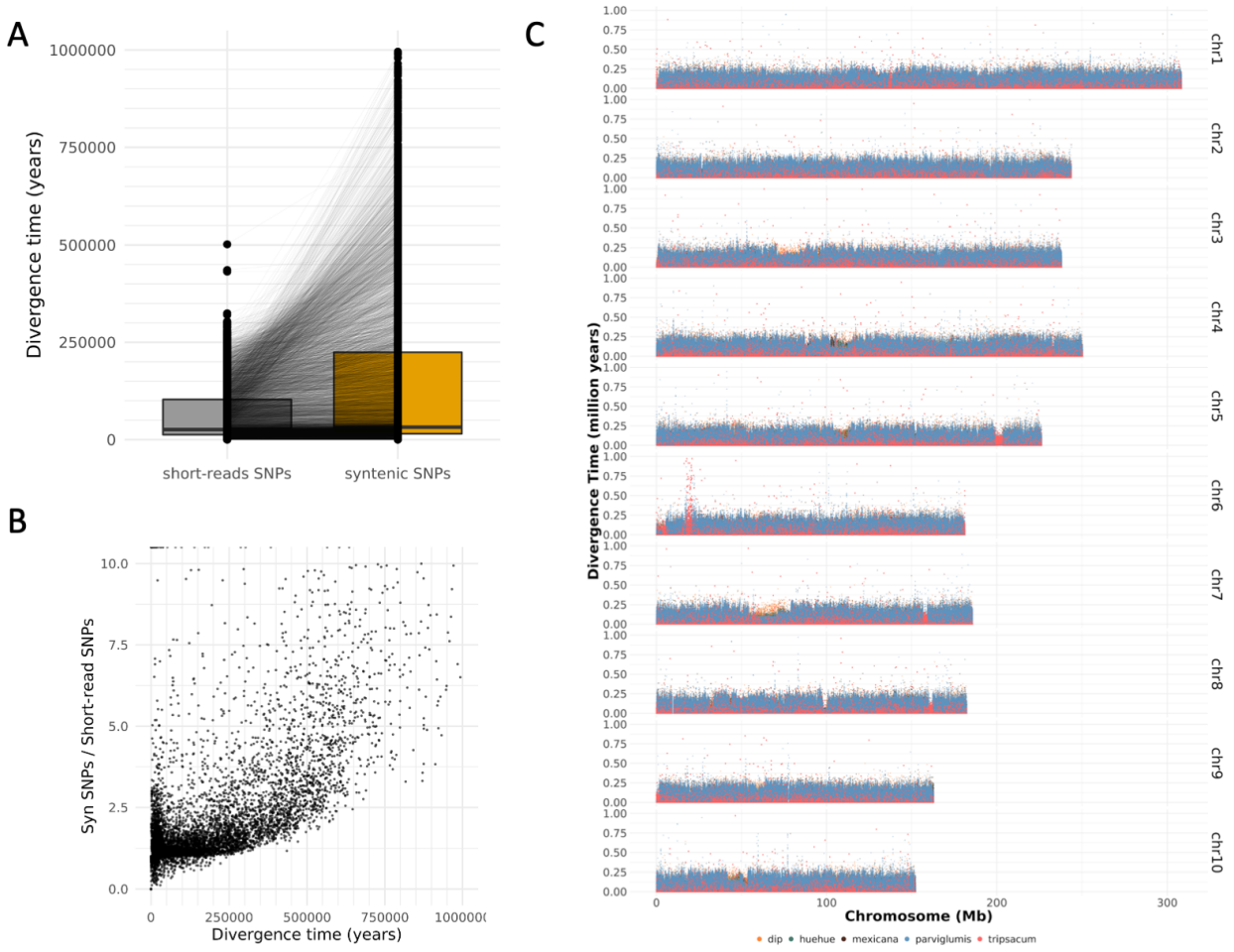


Figure S4.13. Divergence time estimation of the 3Mb segment on chromosome 6 between *Tripsacum* and maize. A) Comparison of divergence times inferred from whole-genome alignment method and short-read mapping method. B) Correlation between divergence time and the variation between the two methods. C) Divergence time estimation of the 3Mb region based on short reads.

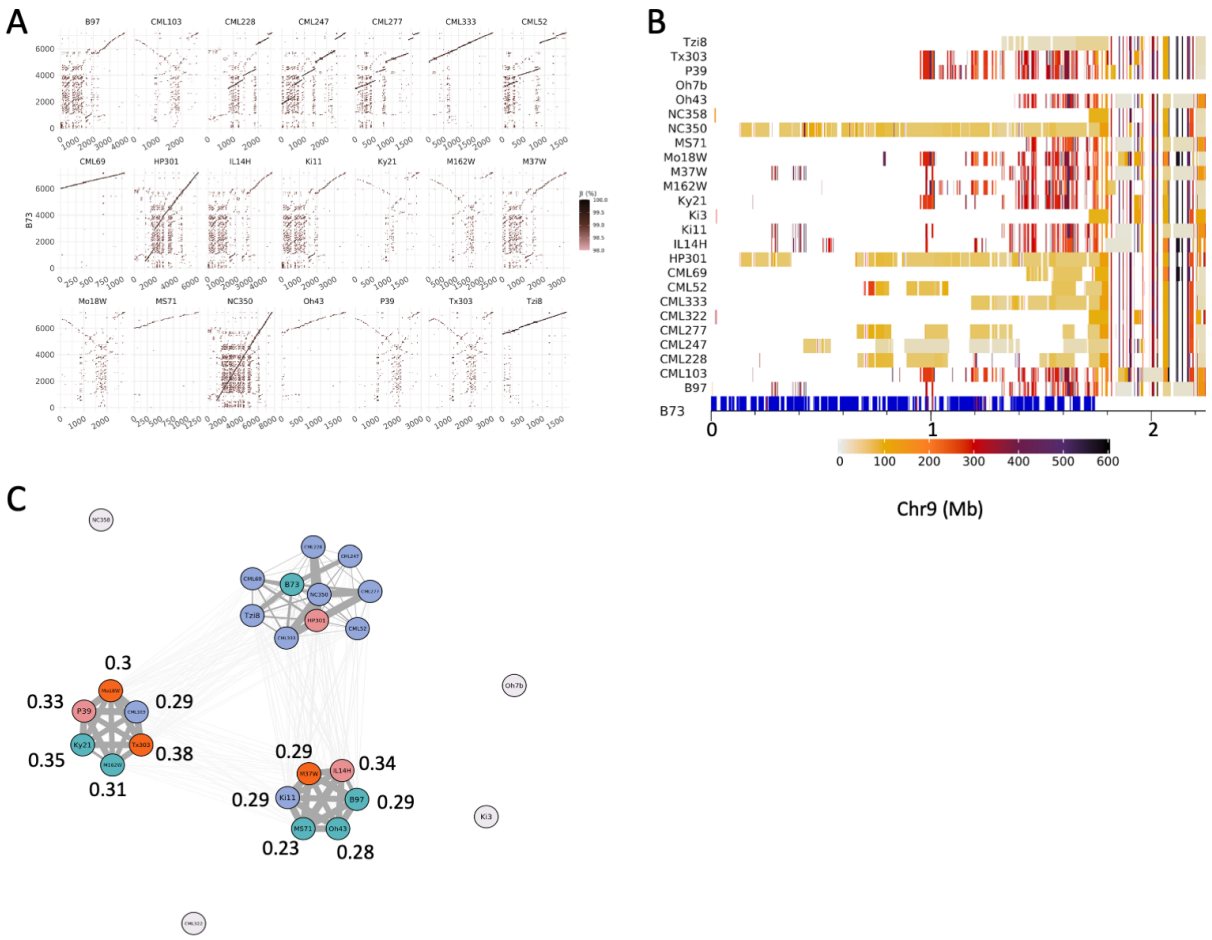


Figure S4.14. Haplotypes and divergence estimation of the 9S knob180 among NAM lines. A) Pairwise alignments between 21 NAM lines and B73 over the 9S knob180. B) Divergence time between NAM and B73. Tandem repeat Knob180 and TR-1 are colored as blue and red. C) Clustering of NAM lines based on all-by-all alignment in A). Divergence times (million years) estimated from syntenic transposable elements were labeled for each group. White circles indicate lack of syntenic knob for that line.

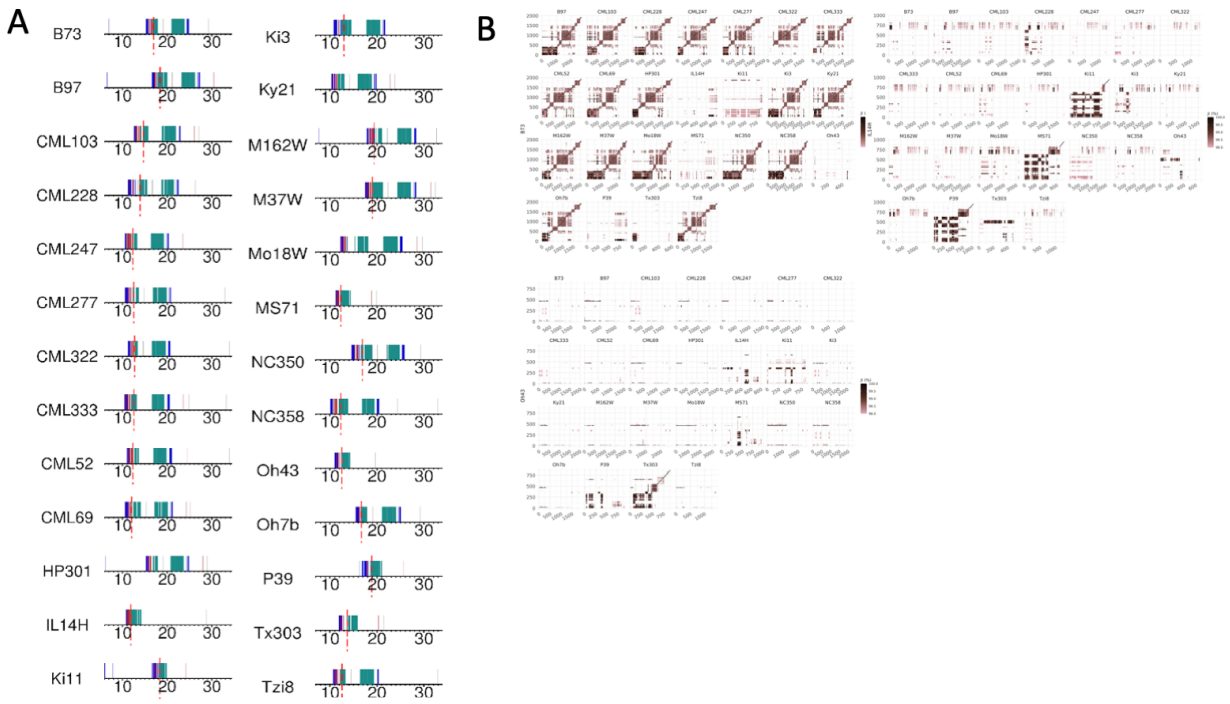
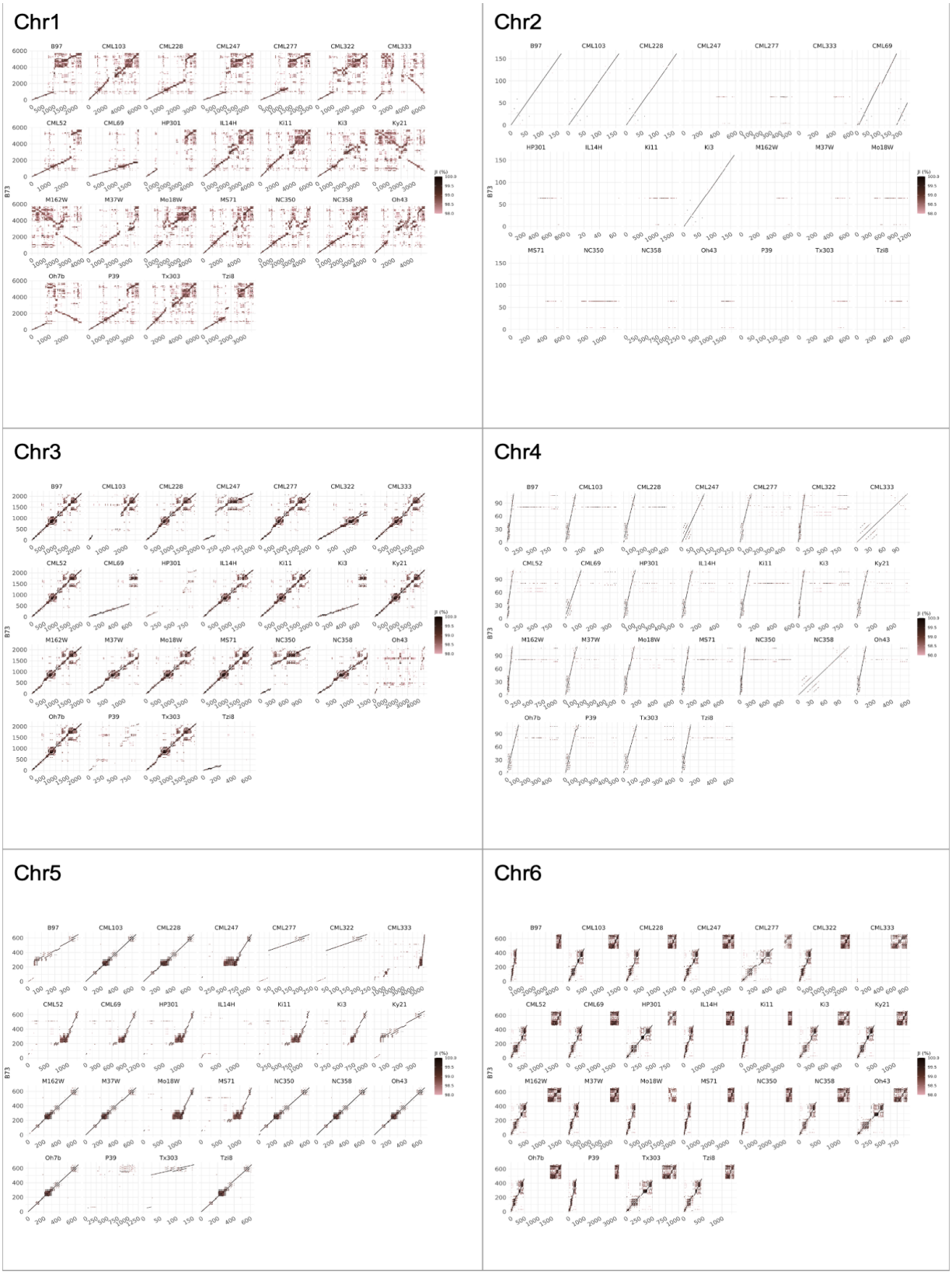


Figure S4.15. Haplotypes of NOR among NAM lines. A) Presence/Absence of the 3Mb introgressed segment in the NOR. Red dashed lines indicate 100N gaps. Tandem repeat Knob180, TR-1, and NOR are respectively colored blue, red and cyan. B) Pairwise alignment of NOR repeat arrays between 25 lines and the reference genome (B73 and IL14H).


Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
B97
CML103
CML228
CML247
CML277
CML322
CML333
CML52
CML69
HP301
IL14H
Ki11
Ki3
Ky21
M162W
M37W
Mo18W
MS71
NC350
NC358
Oh43
Oh7b
P39
Tx303
Tzi8
B73
JI (%)

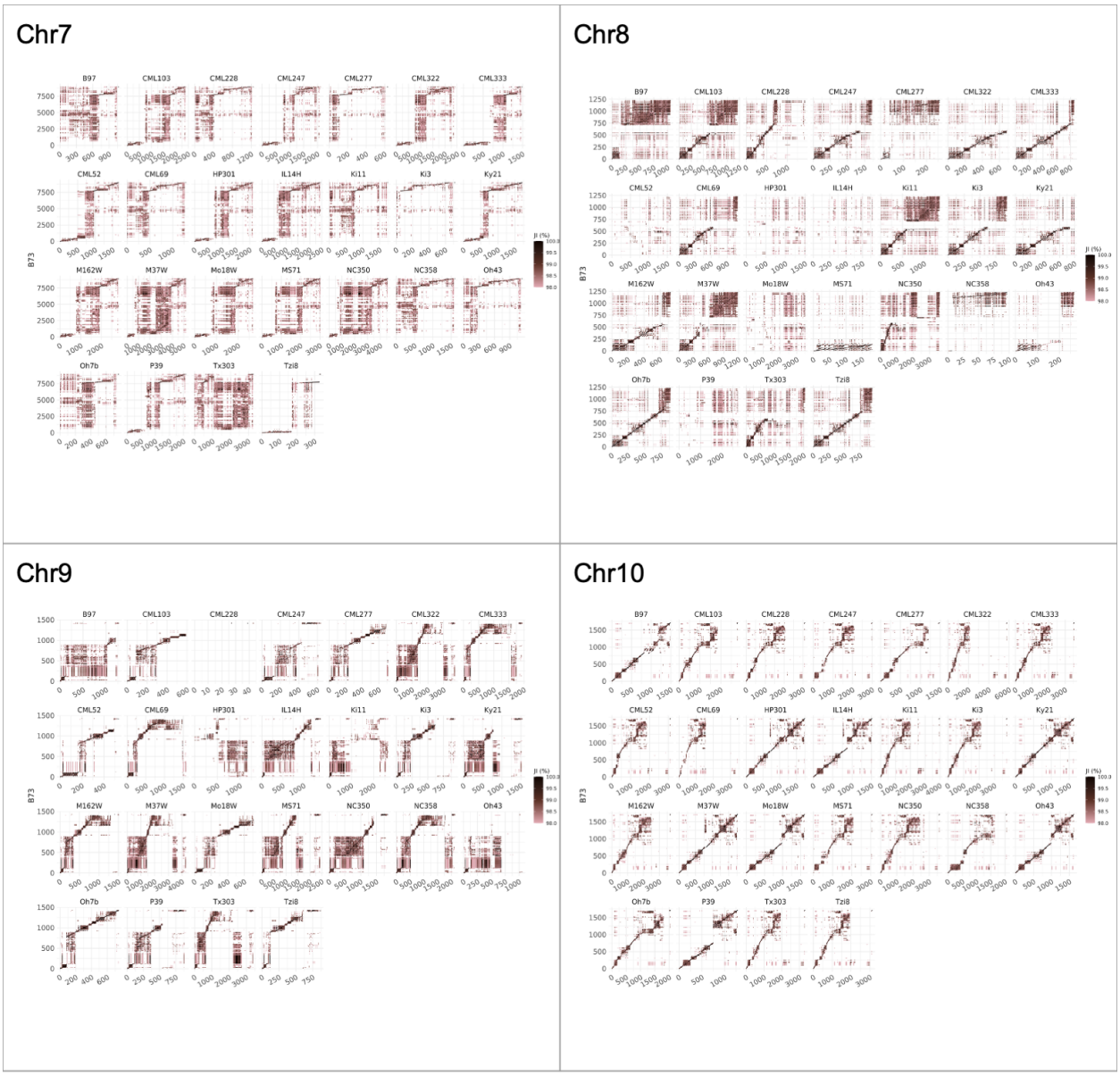
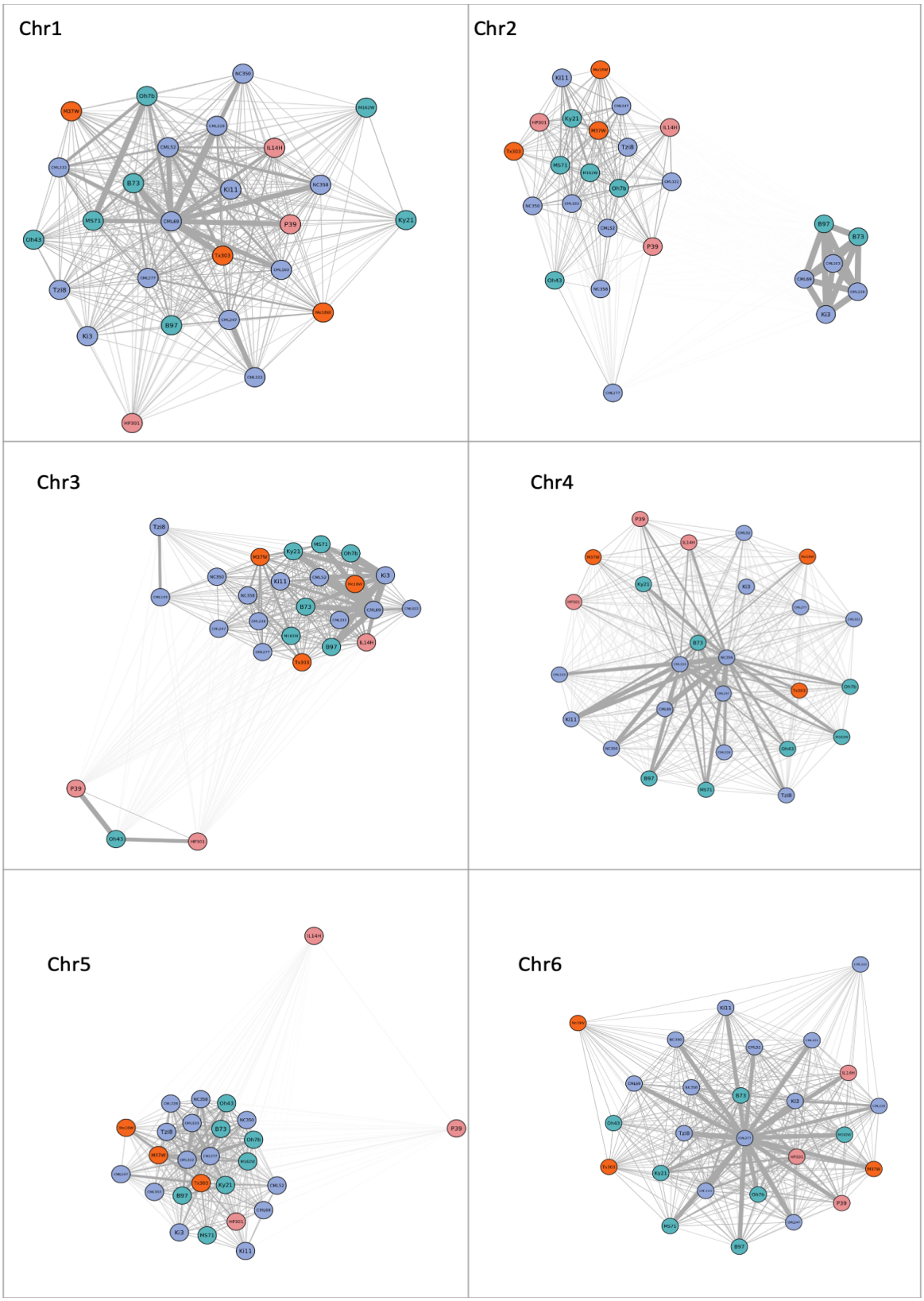


Figure S4.16. Pairwise alignments between NAM and B73 over CentC arrays across 10 chromosomes.


Chr1
Chr2
Chr3
Chr4
Chr5
Chr6

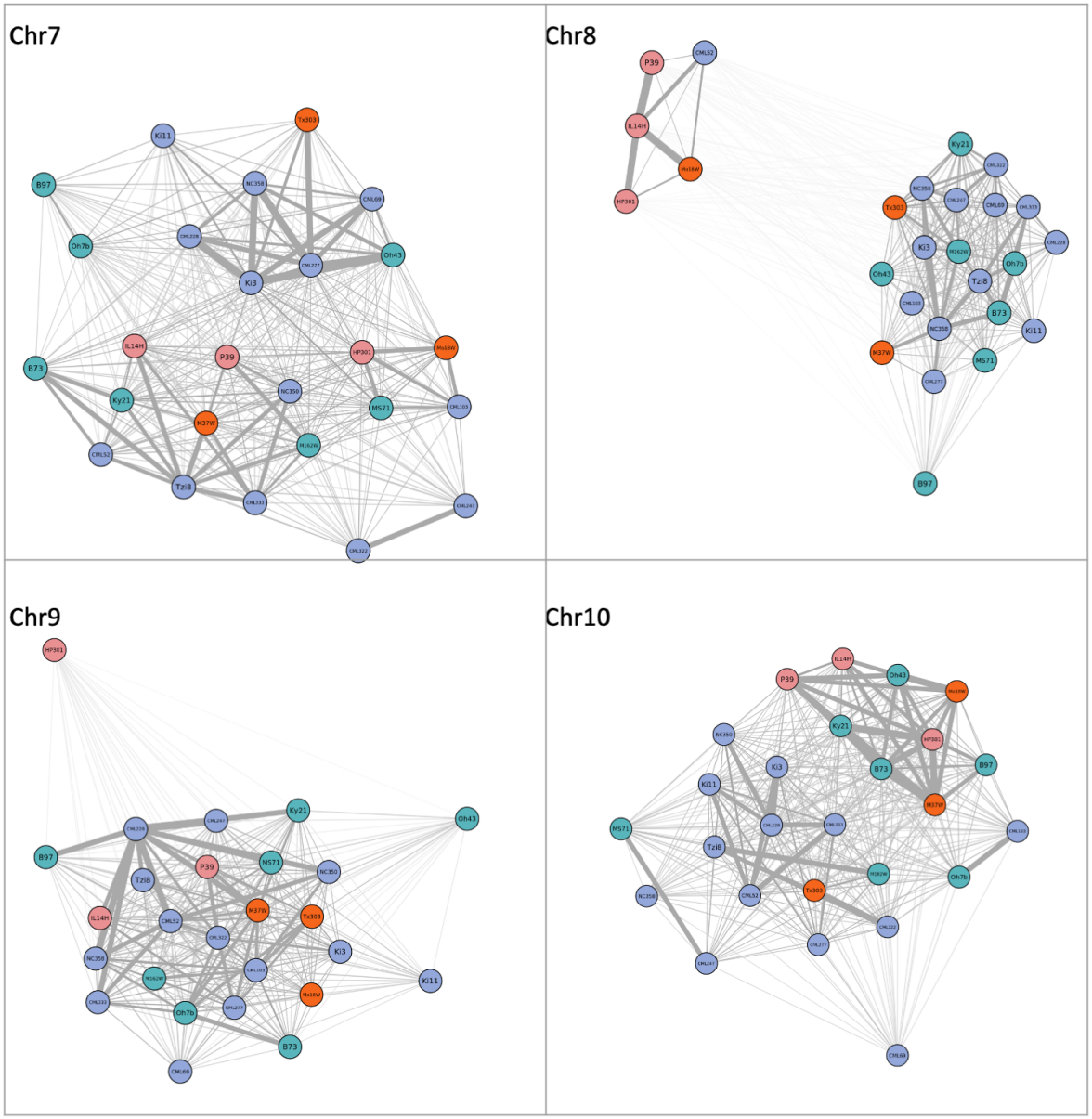
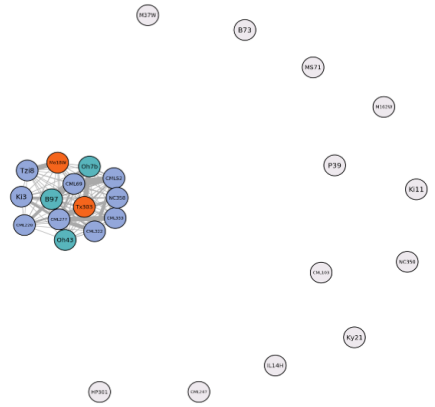
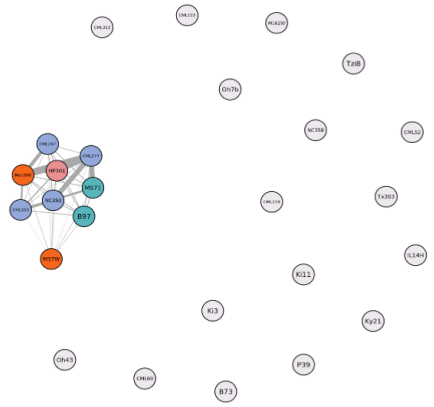


Figure S4.17. Clustering of CentC arrays based on all-by-all alignment across 26 lines. Flint, temperate, mixed, and tropical lines are respectively colored in pink, blue, red and green.

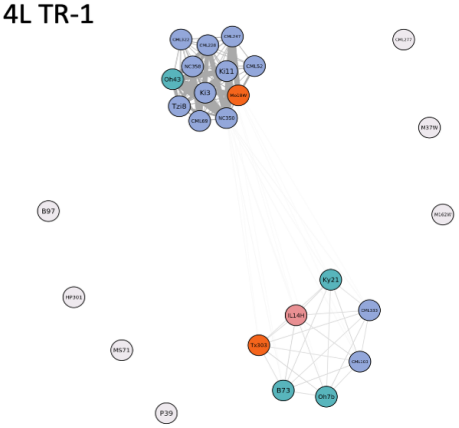
2L kno180



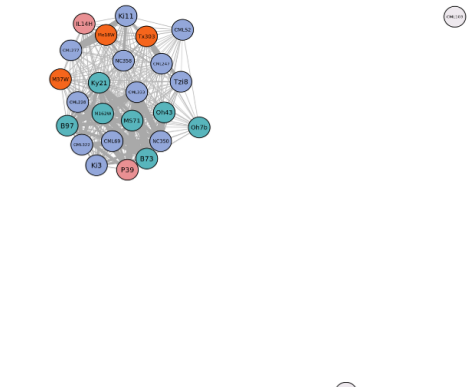
3L kno180



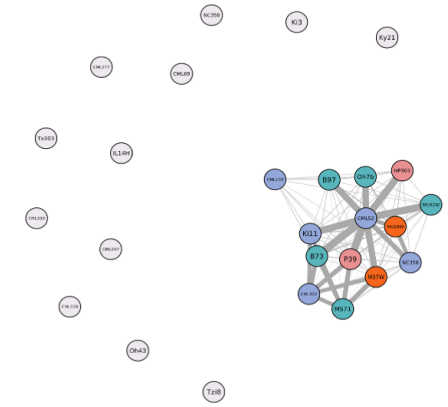
4L TR-1



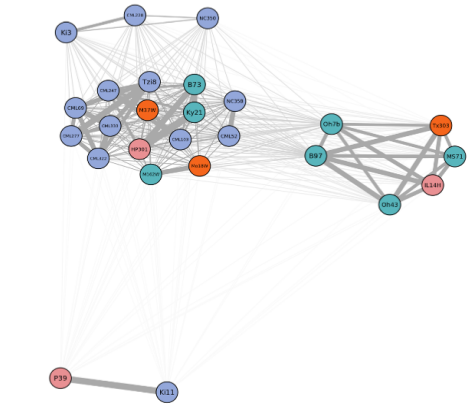
5L kno180



6S1 TR-1



6S2 kno180



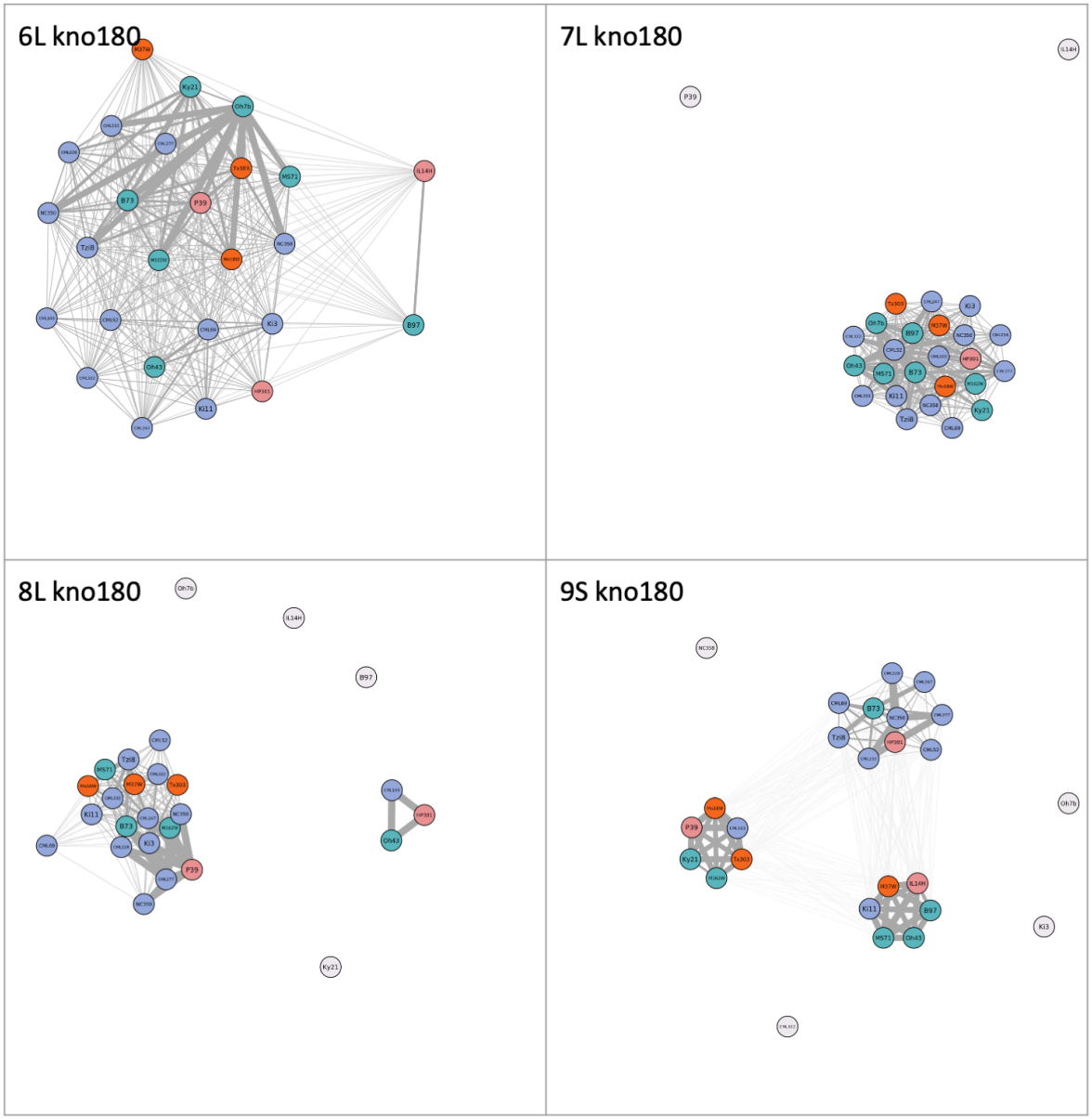


Figure S4.18. Clustering of knob arrays across NAM lines based on all-by-all alignment. Flint, temperate, mixed, and tropical lines are respectively colored in pink, blue, red and green. White circles indicate lack of syntenic knob for that line.

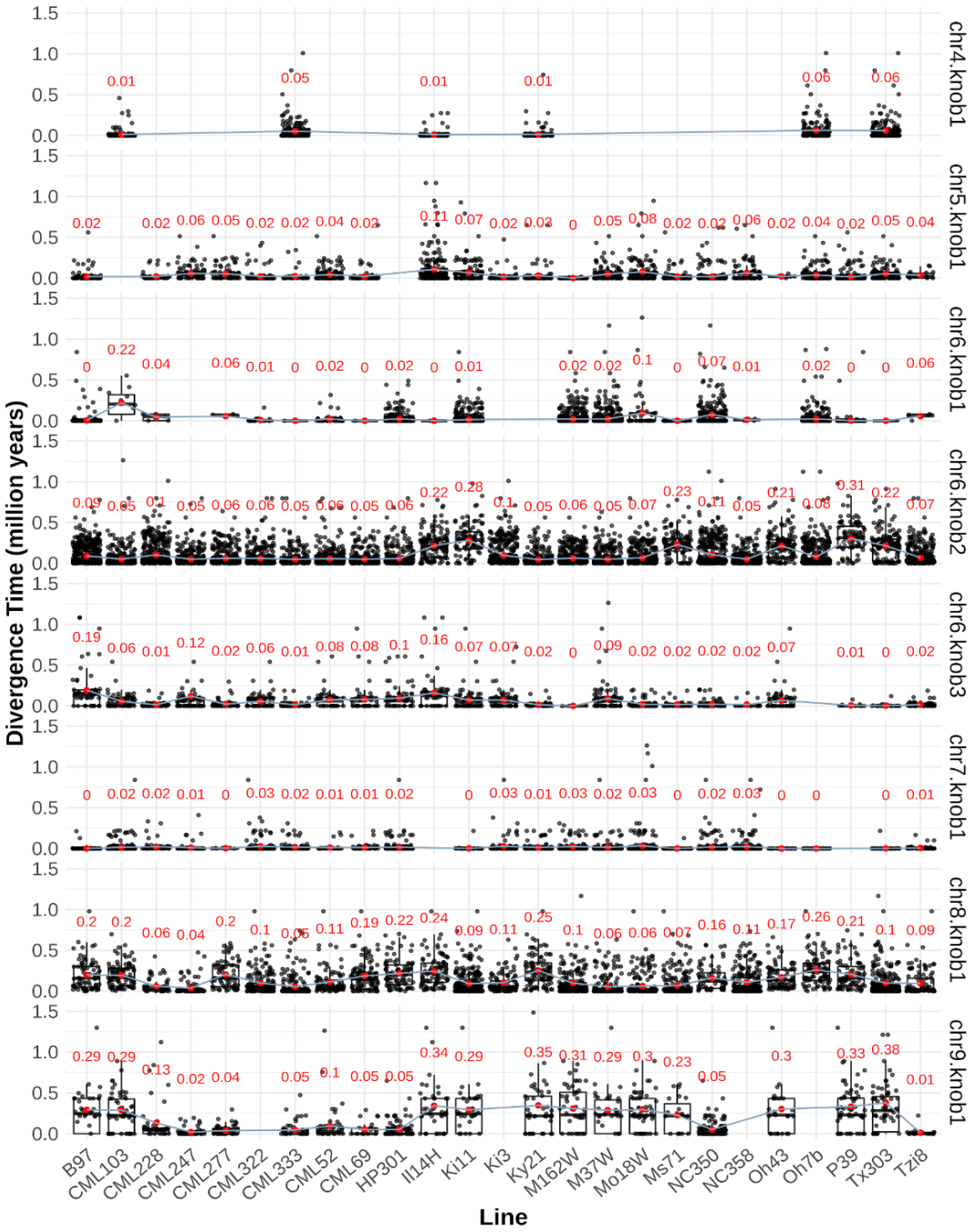


Figure S4.19. Divergence time estimation of eight classical knobs across 25 NAM lines. Each dot represents the divergence value inferred from an individual syntenic aligned transposon fragment. Mean divergence times are labelled and highlighted in red.