Bioinformatic analyses of genomic composition of tandem repeats in grasses

Paul Bilinski*[1]Jiming Postdoc/student[2,], Anne Lorant[1], Matthew B. Hufford[3,], Jiming Jiang[2,], Jeffrey Ross-Ibarra*[1,4]

1 Dept. of Plant Sciences, University of California, Davis, CA, USA
2 Dept. of Horticulture, University of Wisconsin-Madison, Madison, WI USA
3 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
4 Center for Population Biology and Genome Center, University of California, Davis, CA, USA

*** pbilinski@ucdavis.edu, rossibarra@ucdavis.edu**

Abstract

In studying genomic architecture, highly repetitive regions have historically posed a challenge when investigating sequence variation and content. High-throughput sequencing has enabled researchers to use whole-genome shotgun sequencing to estimate the abundance of repetitive sequence, and these methodologies have been recently applied to centromeres. Here, we utilize sequence assembly and read mapping to identify and quantify the genomic abundance of between different tandem repeat sequences. Previous research has posited that the highest abundance tandem repeat in eukaryotic genomes is often the centromeric repeat, and we pair our bioinformatic pipeline with fluorescent in-situ hybridization data to test this hypothesis. We find that de novo assembly and bioinformatic filters can successfully identify repeats with homology to known tandem repeats. Fluorescent in-situ hybridization, however, shows that de novo assembly fails to identify novel centromeric repeats, instead identifying other potentially important repetitive sequences. Together, our results test the applicability and limitations of using de novo repeat assembly of tandem repeats to identify novel centromeric repeats. Building on our findings of genomic composition, we also set forth a method for exploring the repetitive regions of non-model genomes whose diversity limits the applicability of established genetic resources.

Author Summary

Introduction

Sequencing technologies have facilitated genome assembly for many non-model organisms, bringing a tremendous amount of data to the field of comparative genomics. Unfortunately, this data has an overrepresentation of genic regions as short read data currently struggles to accurately assemble repetitive regions. Though repetitive DNA was once disregarded as "junk DNA", research continues to unravel the many functions of repetitive DNA, spurring a growing interest in a better understanding of the

evolutionary history and genomic composition of repeats. Plants are known for their high repetitive content and serve as an excellent model in which to investigate questions regarding repeat sequence evolution. Plant genomes can be up to 85% repetitive [1]. The repetitive sequence found in plants can be classified into two broad categories, either dispersed around the genome because the repeat is derived from transposable elements (TEs) or tandemly repeated sequence. TE derived repeats comprise the majority of many eukaryotic genomes and are recognized for their different modes of amplification, being divided into class I (RNA intermediate) or class II (DNA intermediate). TEs have been shown to impact gene expression [2] and chromatin status [3], functions which can have strong impacts on overall phenotype. The field of comparative genomics has shed light on the evolutionary history of TEs through the phylogeny [4], informed hypotheses about TE expansion and contraction [5], and traced TE function in related organisms [6].

In comparison to the wealth of TE data across organisms, little is known about the function and evolutionary history of tandem repeats. Tandem repeats comprise a smaller percentage of the genome, though that percentage composition can vary wildly across different phylogenetic groups [7]. The most common tandem repeats are found in the gene poor but structurally important telomeres and centromeres. Both telomeres and centromeres are highly heterochromatic, however, the mechanisms by which tandem repeat sequences contribute to the formation of centromeres or telomeres are only vaguely known. (Do I want a sentence here describing the mechanism? I dont think so, but maybe) Previous research has also shown that tandem repeats are not necessary for the formation of centromeres [8], suggesting that tandem repeats may serve as a placeholder for an epigenetic signal that governs heterochromatin formation. Tandem repeats are also found in other types of heterochromatin that suppress local recombination such as knobs in the model organism maize (*Zea mays*, [9]).

In an effort to better understand tandemly repeated sequence, researchers have begun to develop new methodologies that can study sequence variation of repeats without laborious assembly. The abundance of whole genome short read data produced in a variety of organisms is an excellent and freely available resource in which to pilot novel methods. One such attempt has been to pair shotgun short read sequencing from different organisms with de novo repeat assembly [10], effectively constructing the average sequence and base pair length of tandem repeats genome wide. Melters et al [10] applied their approach broadly across more than 280 plant and animal taxa, relying on model organisms to serve as positive controls. They assume that the most abundant tandem repeat in all taxa is the centromere repeat, and use this assumption to study centromere repeat evolution broadly. While the approach appears to work well for animals, the method has known problems when working with plants.

In our study, we aim to apply the basic pipeline of tandem repeat consensus assembly to a narrower phylogenetic group within the grasses to better understand tandem repeat contribution to genomic composition. We select species from across the tribe andropogoneae, including the agriculturally important species maize and sorghum. Maize and sorghum are ideal anchor species as both have reference genomes and their centromere repeats are diverged from one another [1, 11]. Because previous work has shown that sequencing libraries prepared through a identical methods retain relative composition of repetitive content [12],we elect to re-sequence species used in the study. We examine sequence similarity and genomic composition of highly abundant tandem repeats across the phylogeny, determine their homology to known centromere repeats, and perform fluorescent in-situ hybridization to test whether novel high abundance repeats show patterns consistent with other centromere repeats. We show that the assumption of Melters et al. [10] that the highest abundance tandem repeat is a centromere is not supported in grasses, especially when the centromere repeat is novel

or unknown. However, we show that de novo tandem repeat assembly can be used to identify entirely novel repeats such as the knob-like repeat in *Arundinella*. The ability to identify novel repeats from low cost sequencing can enable a new type of comparative genomic study that tracks the evolution of genomic composition.

Materials and Methods

Sequencing

DNA was isolated from leaf tissue using the DNeasy plant extraction kit (Qiagen) according to the manufacturer's instructions. The samples were quantified using Qubit (Life Technologies) 1ug of DNA was fragmented using the bioruptor (Diagenode) with cycles of 30 seconds on, 30 seconds off. The DNA fragments were prepared for Illumina sequencing. First, DNA fragments were repaired with the End-Repair enzyme mix (New England Biolab). A deoxyadenosine triphosphate was added at each 3'ends with the Klenow fragment (New England Biolab). Illumina Trueseq adapters (Affymetrix) were added with the Quick ligase kit (New England Biolab). Between each enzymatic step, cDNA was washed with sera-mags speed beads(Fisher Scientific). Samples were multiplexed and sequenced in one lane of Miseq (UC Davis Genome Center Sequencing Facility) for 150 paired-end base reads with an insert size of approximately 350 bases. Parsing of reads was performed with in house scripts, and forward reads were used for all analyses.

Phylogenetic Tree Reconstruction

We downloaded sequence data for two inter-genic spacers and one chloroplast gene at NCBI (sequences available on `https://github.com/paulbilinski/Github_centrepeat`). Sequences were aligned using seven iterations of MUSCLE [13], and concatenated in order to build a neighbor joining tree using Jukes-Cantor distance. The topologies represent approximate relationships that have been discussed in greater detail in several studies [14, 15]. The topology varies slighty between studies, as some nodes are collapsed into polytomies.

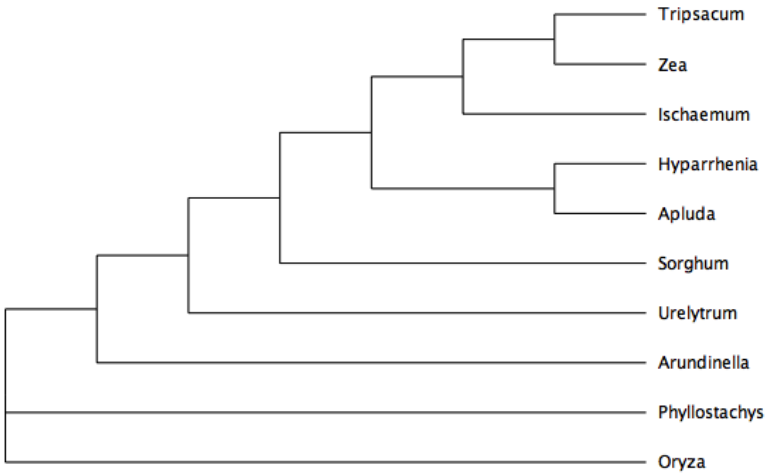

Figure 1. Neighbor joining tree of evolutionary relationships between the grasses studied. Topology is an approximate cladogram, reconstructed from 3 neutrally evolving loci. A more accurate depiction of relationships is available in [14, 15]

Assembly and Genomic Composition of Centromere Repeats

We used MIRA (Chevreux et al. 1999, version 4.0; job = genome,denovo,accurate, parameters = -highlyrepetitive -NW:cnfs=no -NW:mrnl=200 -HS:mnr=no)to assemble low coverage libraries. We ran Tandem Repeat Finder [16] (TRF) on all assembled contigs to select only those that contained tandem repeats. Parameters for TRF were Match = 2, Mismatch = 7, Indel = 7, Probability of match = 80, Probability of indel = 10, Min score = 50, and Max period = 2000. To discover the genomic composition of each tandemly repeated contig, we used Mosaik [17], which stores information about multiply mapping reads (version 1.0; parameters optimized for tandem repetitive elements as in [12]. Low coverage libraries were mapped against the reference and contigs were ranked by the number of reads aligning to each contig. The top ranking contig was extracted, and the number of reads aligning to it was recorded from the assembly ace files. We then blasted (-evalue 1E-1 -outfmt 7 -max_target_seqs 15000 -task blastn) the top ranking contig against all other TRF assemblies, removed assemblies with BLAST homology. This process was repeated 4 times to identify the genomic composition of the 4 highest abundance tandem repeat groups. Scripts and step by step guides are available on `https://github.com/paulbilinski/Github_centrepeat`.

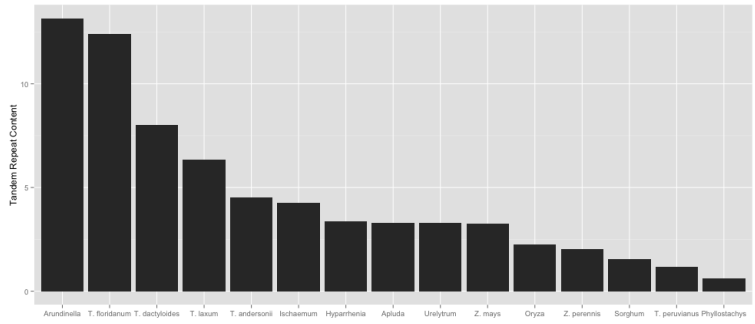

Figure 2. Percentage genomic composition of all tandem repeat contigs in monocot taxa. Values are derived from the sum of all reads mapping to any tandemly repetitive contig derived from TRF after MIRA assembly. Species are ordered from highest to lowest percentage tandem repeat content.

Fluorescent In-Situ Hybridization

Sequences for FISH probes had (Udig 96% identity; Hyp 96% identity) as they were consensus sequences from TRF assemblies of highly similar high abundance contigs. [Text from Jiming here].

Results

Assembly of low coverage whole genome shotgun Illumina data produced several thousand contigs in each species from our panel. From these, TRF was able to identify between 300 and 15,000 contigs comprised of tandem repeats in each taxa. The number of contigs made of tandem repeats varied between taxa based on coverage and overall genomic repetitive content. We mapped whole genome shotgun Illumina data against all post-TRF contigs to approximate genomic composition of all tandemly repetitive sequence in our panel 2. Mapping against all contigs enables us to capture the broad diversity of all tandem repeats. Results show that our taxa vary greatly in their total

tandem repeat content, ranging from over 13% to under 1%. We see high tandem repeat content within the *Tripsacum* and *Arundinella* genera, though *Tripsacum* taxa show large variations. We wanted to see whether total tandem repeat content correlated with genome size. Using the Kew C-Value database [18], we obtained point estimates for genome size in each of our taxa. The correlation between total tandem repeat content and genome size was poor (r=0.05).

We further wanted to investigate the proportional contribution of the most common tandem repeat classes in each of our taxa. To do so, we ranked the mapping abundance of all post-TRF contigs. We used the number of reads mapping to the top ranked contig as its abundance, and removed any similar contigs from our rankings using BLAST homology (See methods for parameters). We repeated this for the top four tandem repeats in each genome. Results showed that most taxa had one tandem repeat class at much higher abundance than other tandem repeats. In all taxa except for *Arundinella*, only the top contig exceeded 1% of genomic composition. *Sorghum*, *Phyllostchys*, *Ischaemum*, and *Apluda* showed the largest difference between the top ranked contig and the second ranked contig. In the sister genera *Zea* and *Tripsacum*, while the top ranked contig showed immense variation, the second ranked contig had a relatively constant abundance near half a percent. The *Arundinella* genome seems unique in that it has several high abundance different class tandem repeats.

In our analysis of genomic contribution from each class of tandem repeat, we found that most genomes had one class of tandem repeat far above the rest. Previous works have posited that this repeat is the centromeric repeat in many species [10]. We wanted test this hypothesis in taxa with known centromere repeats as well as in taxa with uncharacterized genomes. For *Oryza* and *Sorghum*, whose centromere repeats are known, the assumption that the highest abundance tandem repeat is the centromere repeat is true. In *Zea* and *Tripsacum*, where knob repeats are known to exist [19], the centromere repeat is not the highest abundance tandem repeat, though it is ranked within the top four. *Apluda*'s top ranked contig shared homology and a common monomer repeat length with the 137bp *Sorghum* centromere repeat, suggesting that the method is robust in *Apluda* as well. *Ischaemum*'s monomer was also 137bp long, but only shared slight homology to *Sorghum*. However, our analyses also yielded several taxa whose top ranked contig had no homology to known centromere repeats. We chose to test the robustness of the Melters et al [10] assumption in these taxa with fluorescent in-situ hybridization (FISH). FISH from the de novo constructed repeat of *Urelytrum* showed a strong spatial clustering, but clusters were not present on all chromosomes and appeared to be associated with chromosome ends as might be expected from telomeric sequence.It is possible that *Urelytrum* has multiple centromere repeats and that the tandem repeat we probed is specific to only a few chromosomes. The *Hyparrhenia* tandem repeat is widely dispersed across the genome, a pattern we would expect from a TE rather than a tandem repeat. In *Arundinella*, a taxon with a large portion of its genome comprised of tandem repeats, FISH results showed something we have yet to see though I hope it shows new knobs.

Discussion

Our analyses of tandem repeats in grasses provides insight into the evolution of genome organization. Most importantly, we show that previous assumptions about repeat abundance and function do not hold within all taxa examined in this study. Genomic abundance was not predictive of centromere localization within the three novel repeats that were examined through FISH. In *Urelytrum* and *Hyparrhenia*, FISH results did not suggest that the highest ranked contig was centromeric, as probe illumination was scattered across chromosomes, not often found near central locations, or confined to

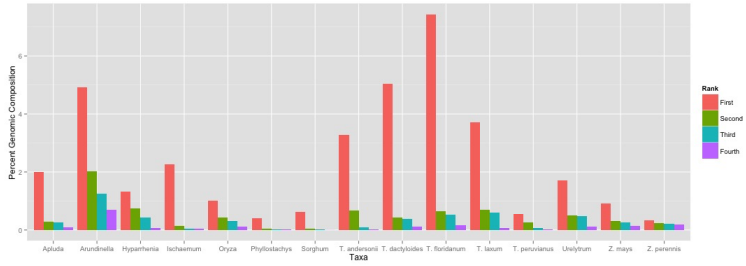

Figure 3. Genomic Composition of Top 4 Tandemly Repetitive Contig that do not share sequence homology.

only some of the chromosomes (Figure 4,5). Sentence about what we learn from the arundinella repeat goes here. The *Ischaemum* repeat, which was not probed, shared short regions of homology with *Sorghum*. The *Zea* and *Tripsacum* repeats shared homology between many of their highest abundance tandem repeats, the highest abundance tandem repeat was homologous to a known repeat, indicative of the presence of knobs and CentC within both genera. *Sorghum*'s and *Apluda*'s highest abundance repeat shared homology and sequence length. *Phyllostachys* and *Oryza*, although they were unique amongst our samples, shared homology with previously annotated bamboo and rice repeats. Altogether, we see both rapid turnover and retention of high abundance tandem repeats across our taxanomic sampling.

De novo assembly of tandem repeats is an efficient, low cost opportunity to explore repetitive content in non-model genomes, an area of study generally left untouched due to the difficulties of traditional assembly. In the assembly of the *Arundinella* repeats, we show an example of how de novo assembly of tandem repeats can add to evolutionary biology. *Arundinella*, basal to all other species in this study, has two highly abundant tandem repeats that are unique over their full length. However, the most abundant 180bp repeats shares long regions of homology to both the sorghum and maize centromere repeat. The ability to look broadly across a phylogeny at consensus repeats may provide a glimpse into the ancestral condition of tandem repeats of andropogonaeae; tandem repeats that gave rise to the structurally important centromere repeats of maize and sorghum.While far from sufficient to make evolutionary claims about the origin of functionally important tandem repeats the identification of novel repeats in previously unstudied organisms has the potential to produce phylogentically relevant data from repetitive content.Using consensus tandem repeats, evolutionary biologists can begin to piece together sequence turnover in repetitive regions [20] and inform the ways in which these heterochromatic regions of the genome evolve.

The methods presented here can also be applied to study variation in genomic composition within and between species. Genome size is highly variable across plants [18] and is implicated with many important phenotypic traits such as flowering time and seed size [21, 22]. The ability to identify what percentage of the genome is composed by tandem repeats can enable studies that track the components driving genome size variation. When applied across populations of a species, researchers can test whether repetitive components that drive genome size change are under selection. Looking across species, repetitive composition can inform our understanding of speciation, providing data toward the ideas of centromere repeat turnover and how frequently it coincides with speciation. Altogether, the results presented here show method of combining bioinformatics with molecular biology to better understand the repetitive fraction of the genome.

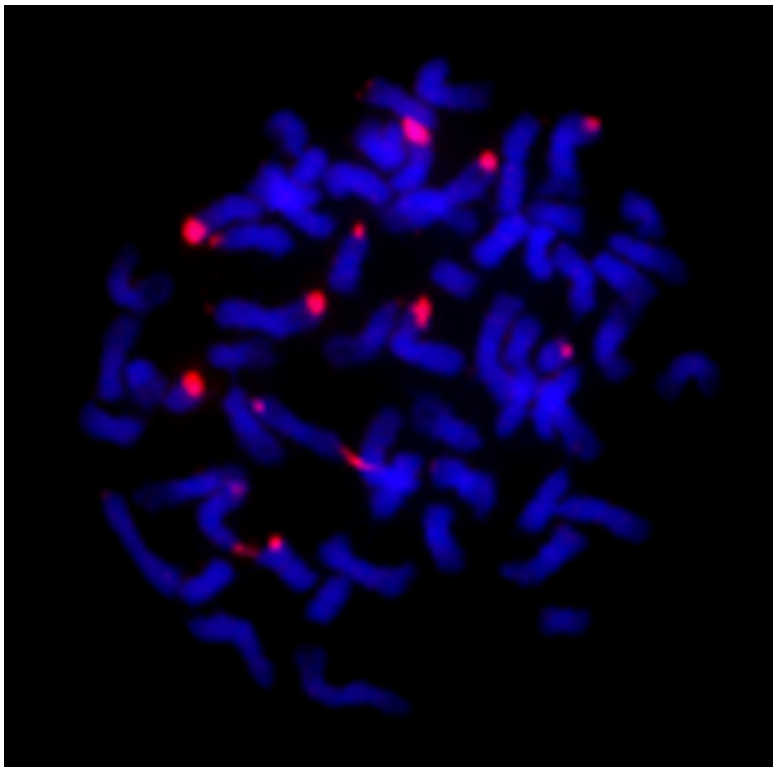

Figure 4. Fluorescent in-situ hybridization for the highest abundance tandem repeat monomer in Urelytrum.

Acknowledgments

205

References

1. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science. 2009;326(5956):1112–1115.
2. Waterland RA, Jirtle RL. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Molecular and cellular biology. 2003;23(15):5293–5300.
3. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature. 2001;411(6834):212–214.
4. Rokas A, Holland PW. Rare genomic changes as a tool for phylogenetics. Trends in Ecology & Evolution. 2000;15(11):454–459.
5. Hawkins JS, Grover CE, Wendel JF. Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. Plant Science. 2008;174(6):557–562.
6. Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics. 1990;124(2):339–355.

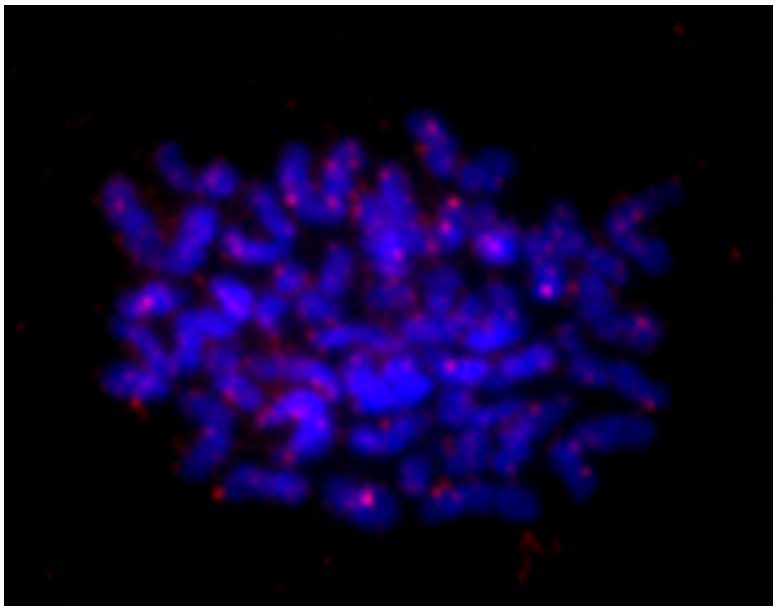

Figure 5. Fluorescent in-situ hybridization for the highest abundance tandem repeat monomer in Hypparhenia.

7. Gaut BS, Ross-Ibarra J. Selection on major components of angiosperm genomes. science. 2008;320(5875):484–486.

8. Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, et al. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. The Plant Cell. 2002;14(11):2825–2836.

9. Chang C, Kikudome GY. The interaction of knobs and B chromosomes of maize in determining the level of recombination. Genetics. 1974;77(1):45–54.

10. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14(1):R10.

11. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457(7229):551–556.

12. Bilinski P, Distor K, Gutierrez-Lopez J, Mendoza GM, Shi J, Dawe RK, et al. Diversity and evolution of centromere repeats in the maize genome. Chromosoma. 2014;p. 1–9.

13. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

14. Wu ZQ, Ge S. The phylogeny of the BEP clade in grasses revisited: Evidence from the whole-genome sequences of chloroplasts. Molecular Phylogenetics and Evolution. 2012;62(1):573–578.

15. Skendzic EM, Columbus JT, Cerros-Tlatilpa R. Phylogenetics of Andropogoneae (Poaceae: Panicoideae) Based on Nuclear Ribosomal Internal Transcribed Spacer and Chloroplast trnL–F Sequences. Aliso: A Journal of Systematic and Evolutionary Botany. 2007;23(1):530–544.

16. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999;27(2):573.

17. Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS one. 2014;9(3):e90581.

18. Bennet MD, J LI. Plant DNA C-values database;. Release 6.0, Dec. 2012. http://www.kew.org/cvalues/.

19. Dennis E, Peacock W. Knob heterochromatin homology in maize and its relatives. Journal of molecular evolution. 1984;20(3-4):341–350.

20. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293(5532):1098–1102.

21. Rayburn AL, Dudley J, Biradar D. Selection for early flowering results in simultaneous selection for reduced nuclear DNA content in maize. Plant Breeding. 1994;112(4):318–322.

22. Knight CA, Molinari NA, Petrov DA. The large genome constraint hypothesis: evolution, ecology and phenotype. Annals of Botany. 2005;95(1):177–190.