Title

Genomic abundance is not predictive of tandem repeat localization in grass genomes

Journal

PLOS ONE, 12(6)

ISSN

1932-6203

Authors

Bilinski, Paul

Han, Yonghua

Hufford, Matthew B

et al.

Publication Date

2017

DOI

10.1371/journal.pone.0177896

PLOS ONE

RESEARCH ARTICLE

Genomic abundance is not predictive of tandem repeat localization in grass genomes

Paul Bilinski[1¤]*, Yonghua Han[2,3], Matthew B. Hufford[4], Anne Lorant[1], Pingdong Zhang[3,5], Matt C. Estep[6], Jiming Jiang[3], Jeffrey Ross-Ibarra[1,7]*

1 Dept. of Plant Sciences, University of California, Davis, Davis, CA, United States of America, **2** School of Life Sciences, Jiangsu Normal University, Xuzhou, China, **3** Dept. of Horticulture, University of Wisconsin-Madison, Madison, WI, United States of America, **4** Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, United States of America, **5** College of Bioscience and Biotechnology, Beijing Forestry University, Beijing, China, **6** Dept. of Biology, Appalachian State University, Boone, NC, United States of America, **7** Genome Center and Center for Population Biology, University of California, Davis, Davis, CA, United States of America

¤ Current address: Research Group for Ancient Genomics and Evolution, Dept. of Molecular Biology, Max Planck Institute for Developmental Biology, Tuebingen, Germany
* paul.bilinski@tuebingen.mpg.de (PB); rossibarra@ucdavis.edu (JRI)

Citation: Bilinski P, Han Y, Hufford MB, Lorant A, Zhang P, Estep MC, et al. (2017) Genomic abundance is not predictive of tandem repeat localization in grass genomes. PLoS ONE 12(6): e0177896. https://doi.org/10.1371/journal.pone.0177896

Editor: Xiu-Qing Li, Agriculture and Agri-Food Canada, CANADA

Received: October 31, 2016

Accepted: May 4, 2017

Published: June 1, 2017

Data Availability Statement: Raw data are available from the online repository at figshare (https://figshare.com/articles/grasscentrepeat_tar_bz2/3494378), while scripts and analyses are available on github (https://github.com/paulbilinski/Github_centrepeat).

Funding: JRI would like to acknowledge support from USDA Hatch project CA-D-PLS-2066-H and NSF Plant Genome award IOS-0922703. PB would like to acknowledge support from the UC Davis Department of Plant Sciences and the DuPont

Abstract

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. Previous research has investigated variation in centromere repeats across eukaryotes, positing that the highest abundance tandem repeat in a genome is often the centromeric repeat. To test this assumption, we used shotgun sequencing and a bioinformatic pipeline to identify common tandem repeats across a number of grass species. We find that *de novo* assembly and subsequent abundance ranking of repeats can successfully identify tandem repeats with homology to known tandem repeats. Fluorescent *in-situ* hybridization shows that de novo assembly and ranking of repeats from non-model taxa identifies chromosome domains rich in tandem repeats both near pericentromeres and elsewhere in the genome.

Introduction

Advances in sequencing technology have facilitated development of reference genomes for many non-model organisms, providing a tremendous resource for the field of comparative genomics. Our understanding of the repetitive regions of genomes, however, has lagged behind that of gene-rich regions, mostly because the high identity shared between repeat sequences causes problems with assembly and mapping [1]. Though repetitive DNA is often disregarded as"junk DNA", research continues to unravel its many functions, spurring a growing interest in a better understanding of the evolutionary history and genomic composition of repeats [2]. Repeat sequence can be broadly classified into two categories: dispersed repeats derived from transposable elements (TEs) and tandemly repeated sequences. TE-derived

Pioneer Howie Smith Honorary Plant Breeding Fellowship.

Competing interests: The authors have declared that no competing interests exist.

repeats comprise the majority of many eukaryotic genomes and have been recognized for their potential impacts on phenotype, for example via gene expression [3, 4] or affecting chromatin status [5].

In comparison to the wealth of TE data across organisms, much less is known about the function and evolutionary history of tandem repeats. Tandem repeats are commonly found in the gene-poor regions of the genome such as telomeres and centromeres as well as heterochromatic knobs [6], B chromosomes [7], and sex chromosomes [8]. While tandem repeats generally make up less of the genome than TEs, their abundance varies substantially across phylogenetic groups [9]. In an effort to better understand tandem repeats, researchers have applied both sequencing technologies and molecular biology. Several studies, for example, have paired chromatin immunoprecipitation (ChIP) of centromere proteins with clustering algorithms [10] to identify centromeric repeats [11, 12, 13].

In a recent paper, Melters *et al* [9] conducted *de novo* repeat assembly of published short read sequence data to study the evolution of centromeric tandem repeats across 280 plant and animal species. While tandem repeats do not appear necessary for the formation of centromeres [14], they may serve as placeholders for an epigenetic signal that governs heterochromatin formation [15] or function in repair of double strand breaks [16]. Transcripts from centromere-associated tandem repeats have also been found in the nucleolus of both plant and animal taxa and are thought to be important in protein assembly [17, 18], further suggesting a potential functional role for tandem repeats. Given their likely importance, there is great potential for a bioinformatic approach that takes advantage of published sequence data. One critical assumption of the Melters *et al.* [9] approach, however, is that the most abundant tandem repeat in each genome taxa is the centromere repeat. While comparison to known repeats in several model organisms suggests this assumption works well for animals [9], earlier work suggests that it may not apply broadly to plants. Using a similar pipeline and 454 shotgun reads from *Solanum* taxa, for example, Torres *et al.* [19] identified the most abundant tandem repeats as subtelomeric rather than centromeric.

Here, we test the assumptions of Melters *et al.* [9], applying their pipeline to species within the Andropogoneae tribe of grasses and three outgroups, *Arundinella*, rice, and bamboo, in order to better understand tandem repeat contribution to genomic composition. The Andropogoneae tribe, sometimes referred to as the sorghum tribe, includes both maize and sorghum, two model organisms with well annotated repeats [20, 21]. Many other species in this group are agriculturally and scientifically important, including sugar cane. The presence of well annotated reference genomes allows us to test the accuracy of our method and the Melters *et al.* [9] assumption regarding centromere repeat sequence and its genomic abundance. We examine the genomic composition of highly abundant tandem repeats across these species, determine their homology to known centromere repeats, and perform fluorescent *in-situ* hybridization to test whether novel high-abundance repeats show patterns consistent with known centromere repeats. We show that the common assumption that the highest abundance tandem repeat is centromeric is not supported in these taxa, but that *de novo* tandem repeat assembly can be used to identify entirely novel repeats such as a knob-like repeat in *A. nepalensis* and *U. digitatum*.

Materials and methods

Sequencing and genome size measurements

Because previous work has shown that sequencing libraries prepared through identical methods better retain relative composition of repeats [22], rather than use published data we elect to re-sequence all the species used here. Seed was requested from the GRIN database, and

Table 1. Counts of reads per sequence library for each taxa. An accession ID of NA indicates a purchase from a local nursery or sample not registered with GRIN. Taxa were selected broadly from across the Andropogoneae tribe, with higher density sampling in the *Tripsacum* genus to study tandem repeat variation within a genus. We used *A. nepalensis*, rice, and bamboo as outgroups to the Andropogoneae. Asterisks indicate genome size estimates published in this study. GS = Genome size.

Genus	Species	Reads	GS (pg/1C)	AccessionID
Apluda	*mutica*	746994	1.79*	PI 219568
Arundinella	*nepalensis*	662118	2.02[23]	PI 384059
Hyparrhenia	*hirta*	861995	1.86*	PI 206889
Ischaemum	*rugosum*	920258	0.75*	Kew 0183574
Oryza	*sativa*	599567	0.50[24]	NA
Phyllostachys	*edulis*	628030	2.1[25]	NA
Sorghum	*bicolor*	473944	0.75[24]	PI 564163
Tripsacum	*andersonii*	288175	5.8*	MIA 34430
Tripsacum	*dactyloides*	391848	3.88[24]	MIA 34597
Tripsacum	*floridanum*	743668	3.47*	MIA 34719
Tripsacum	*laxum*	723097	3.04*	MIA 34792
Tripsacum	*peruvianum*	238983	4.55*	MIA 34501
Triticum	*urartu*	435815	4.93[24]	PI 428198
Urelytrum	*digitatum*	661535	0.73*	SM3109
Zea	*mays*	4422188	2.73[24]	RIMMA0019
Zea	*perennis*	5106091	5.28[24]	NA

https://doi.org/10.1371/journal.pone.0177896.t001

accession information is available in Table 1. DNA was isolated from leaf tissue using the DNeasy plant extraction kit (Qiagen) according to the manufacturer's instructions. Samples were quantified using Qubit (Life Technologies) and 1ug of DNA was fragmented using a bioruptor (Diagenode) with cycles of 30 seconds on, 30 seconds off. DNA fragments were then prepared for Illumina sequencing. First, DNA fragments were repaired with the End-Repair enzyme mix (New England Biolabs). A deoxyadenosine triphosphate was added at each 3'end with the Klenow fragment (New England Biolabs). Illumina Truseq adapters (Affymetrix) were then added with the Quick ligase kit (New England Biolabs). Between each enzymatic step, DNA was washed with sera-mags speed beads (Fisher Scientific). Samples were multiplexed using Illumina compatible adapters with inline barcodes 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 (All scripts for this and other processes are available at https://github.com/paulbilinski/Github_centrepeat). In short, barcodes were trimmed from the sequence, paired reads were separated so that a single read could be used for assembly, allowing for much faster repetitive contig assembly. Sequence data for each species are available on FigShare (https://dx.doi.org/10.6084/m9.figshare.3494378.v2). Genome sizes were estimated using flow cytometry following [23].

Assembly and genomic composition of tandem repeats

To assemble contigs from low coverage sequence, we used MIRA [26] (version 4.0; job = genome,denovo,accurate, parameters = -highlyrepetitive -NW:cnfs = no -NW:mrnl = 200 -HS:mnr = no). We selected to use MIRA over other assemblers due to its relative speed of repetitive sequence assembly without loss of assembly quality. We ran Tandem Repeat Finder [27] (TRF) on assembled contigs, removing any unassembled reads. Previous work has shown that TRF identifies only those contigs that contain tandem repeats [9], and dot plots of the contigs confirmed the presence of tandem repeats (S1 Fig). We utilized only those contigs in all

subsequent analyses. 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. Sequence files for all contigs can be found on the project github. To discover the abundance of the tandem repeats identified in our post-TRF analysis contigs, we used Mosaik [28], which stores information about multiply mapping reads (version 1.0; parameters optimized for tandem repetitive elements as in Bilinski *et al* [22]). Low coverage libraries (<0.1X) were mapped against the contigs identified by TRF and contigs were ranked by the number of reads aligned. Previous work has shown that low coverage libraries are sufficient to recover the genomic composition of high abundance repeats [22, 29]. The top ranking contigs above 30bp were extracted, and the number of reads aligning to it was recorded from the assembly ace files. The TRF analysis that identified assembled contigs with tandem repeats also identified the consensus monomer for those tandem repeats. We used the consensus monomer from the top ranking contig to blast (-evalue 1E-1 -outfmt 7 -max_target_seqs 15000 -task blastn; these parameters were used for all BLAST analyses) against all other TRF assemblies and grouped contigs with BLAST homology. The groups of contigs identified by BLAST homology were removed from the contig library and marked as the highest abundance tandem repeat cluster. This process was repeated 4 times to identify the genomic composition of the 4 highest abundance tandem repeat groups; monomer information is available in S1 and S2 Tables. We chose to examine only the top 4 repeats as abundance was often negligible after the 4th repeat. Finally, to estimate the overall abundance of each of these four repeats, we mapped reads against a reference consisting of the most abundant monomer and all polymers with homology to the monomer as determined by BLAST. Mapping against either single monomers or groups of contigs ensured that fragment length bias did not play a large role in overall genomic composition. Contig sequence and length can be found on the project github.

Fluorescent in-situ hybridization

Repetitive sequences were amplified using the genomic DNA isolated from the targeted species and labeled with digoxigenin-11-dUTP. Hybridization signals were detected with rhodamine-conjugated anti-digoxigenin (Roche Diagnostics USA, Indianapolis, IN). Chromosomes were counterstained with 4‘,6-diamidino-2-phenylindole (DAPI). The following primers were used on the species indicated: *A. nepalensis* (sequence ID 568) Primer F- CCATTCAAGAAATGGT GTCA; *A. nepalensis* Primer R- GCAAGTACGAAAGCCAAAAT; *U. digitatum* (sequence ID 605) Primer F- GCACTGGCCCTGAGAGAAAT; *U. digitatum* Primer R- ACAGGCTTGGGT GGACAAAA; *H. hirta* (sequence ID 520) Primer F- GATCCGAAAGTCGCGAAACG; *H. hirta* Primer R- TTTTTCGCAACGAACGCACA. We were unable to perform FISH on *I. rugosum* due to a lack of living tissue. Primers were designed based on the most abundant tandem repeat contig of the species using the program Primer3 [30]. PCR and FISH was performed using published procedures [31].

Results

Assembly of low depth resequencing data produced several thousand contigs in each species from our panel (Fig 1, and Table 1). From these, TRF identified between 300 and 15,000 tandem repeat contigs in each taxon (sequences available on the project github at https://github.com/paulbilinski/Github_centrepeat). The number of tandem repeat contigs varied across taxa based on coverage and overall genomic repetitive content. We then mapped our sequence data against tandem repeat contigs to approximate the abundance of tandem repeats in our panel (Fig 1). Our taxa vary greatly in their total tandem repeat content, ranging from over 13% to under 1%. We see high tandem repeat content across the *Tripsacum* genus and in

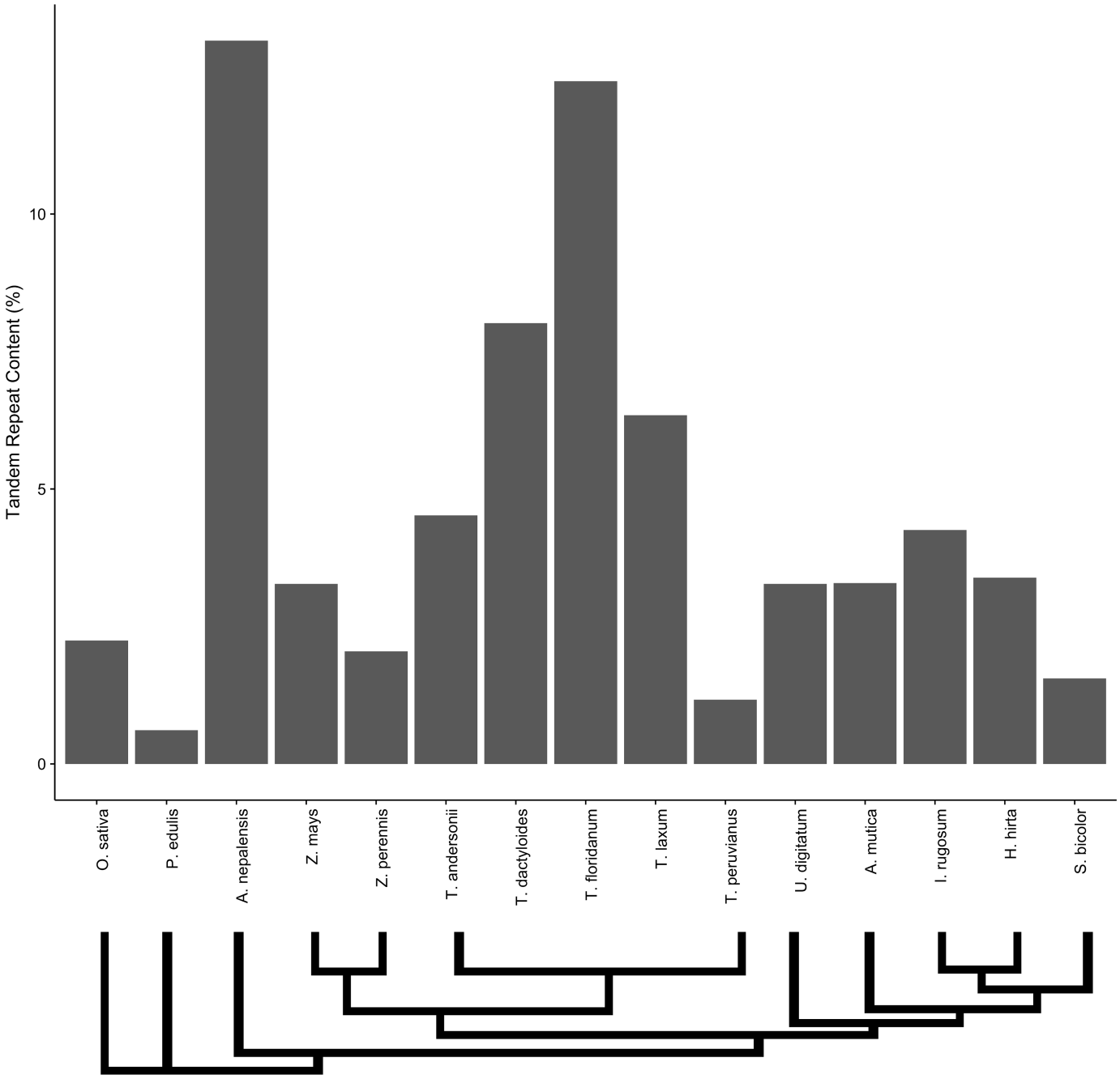

Fig 1. Percentage genomic composition of all tandem repeat contigs in monocot taxa. Values are derived from the proportion of all reads mapping to any tandemly repetitive contig derived from TRF after MIRA assembly. Species are ordered in approximate phylogenetic relationship, with a phylogenetic schematic below the graph.

https://doi.org/10.1371/journal.pone.0177896.g001

A. nepalensis, though *Tripsacum* taxa show large variation. Based on genome size estimates (see Table 1), the correlation between total tandem repeat content and genome size is poor across all taxa as well as within *Tripsacum* (Pearson correlation; $p > 0.05$).

In order to investigate the proportional contribution of the most common tandem repeat classes in each of the analyzed taxa, we ranked the mapping abundance of all contigs

containing tandem repeats as identified by TRF. 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 at much higher abundance (Fig 2). In all taxa except for *A. nepalensis*, only the top contig exceeded 1% of genomic composition. *S. bicolor*, *P. edulis*, *I. rugosum*, and *A. mutica* showed the largest difference between the top ranked contig and the second ranked contig. In the sister genera *Zea* and

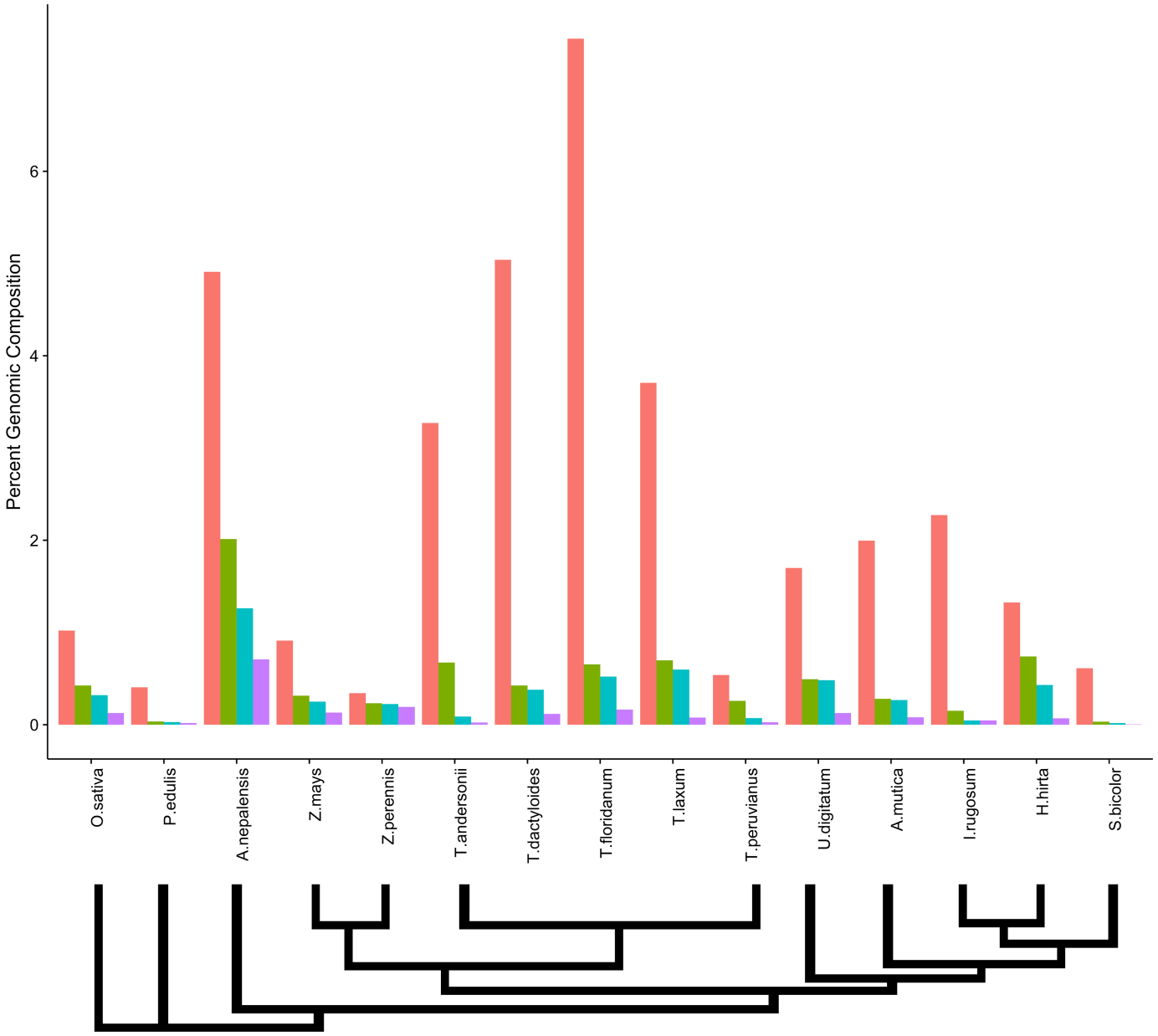

Fig 2. Genomic composition of top 4 tandemly repetitive contigs. The top 4 contigs in each species were defined as not having homology to one another, in order to identify independent repeat motifs. Species are ordered in approximate phylogenetic relationship, with a phylogenetic schematic below the graph. Values were calculated as a percentage of total genomic reads mapping to each tandem repeat family. Tandem repeat families are ordered by their genomic abundance from left to right.

https://doi.org/10.1371/journal.pone.0177896.g002

Tripsacum, while the top ranked contig showed immense variation, the second ranked contig had a relatively constant abundance near 0.5%.

We wanted to test whether the assumption that the most abundant repeat is centromeric [9] could be applied to these grass taxa with both known and uncharacterized centromere repeats. Among taxa with known centromere repeats, the centromere repeats were found to be the most abundant tandem repeat in both *O. sativa* and *S. bicolor*. The percentages of the genome comprised from each tandem repeat was similar to other studies performed in both Sorghum (1.6–1.9%) [32] and maize(<1%) [21]. In *Zea* and the closely related *Tripsacum* taxa, the centromere repeat was among the four most abundant, but the highest abundance repeat came instead from heterochromatic knobs, as has been noted previously [9, 33]. In *P. edulis*, the most abundant repeat has homology to a repeat region in *Bambusa*, but is not annotated as centromeric. While the centromere repeat was not previously known for the species *A. mutica*, its highest abundance contig shared homology and a common monomer repeat length with the *S. bicolor* centromere repeat. Previous FISH studies have shown that the *Tripsacum* centromere repeat shares homology to maize and is localized to the centromere [34]. Our data show that the ranking of the top two repeats in all *Tripsacum* species studied is the same, while the 3rd and 4th most abundant repeats vary between the species within the genus. The top-ranked contig in *I. rugosum* shared a monomer length identical to the centromere repeat of *Sorghum*, but with no sequence homology. The top ranked contigs from the remaining taxa in our panel bore no similarity to known centromere repeats. *T. urartu* did not have any tandem repeats longer than 30 bp at an appreciable frequency in the genome (see Methods). To test whether the most abundant repeat in these taxa is centromeric, we performed fluorescent *in situ* hybridization on *A. nepalensis*, *H. hirta*, and *U. digitatum* (FISH; Fig 3), expecting spatial clustering of the probe proximal (for metacentric) or distal (for acrocentric) of most if not all chromosomes. FISH from the *de novo* constructed repeat of *H. hirta* is widely dispersed across the genome, a pattern reminiscent of a TE rather than a localized tandem repeat. The tandem repeat from *U. digitatum* showed strong spatial clustering, though clusters were not found on all chromosomes and were associated with chromosome ends as might be expected from a subtelomeric sequence. The regions probed in *U. digitatum* did associate with visible knobs and the monomer repeat length is 184bp, similar to the 180bp knob repeat found in tightly packed heterochromatin in *Zea* (Fig 3). The monomer sequence of *U. digitatum* does not have homology longer than 30bp to any annotated sequence and may be a novel knob variant identifed here. The probed repeat of *A. nepalensis* also showed subtelomeric clustering, and the fact that *A. nepalensis* had the largest proportion of its genome comprised of tandem repeats (Fig 1) is consistent with a knob-like origin for this tandem repeat. In both *A. nepalensis* and *U. digitatum*, FISH signal did not occur at visible primary constriction sites (Fig 3). While the suspected knob repeat sequences in *A. nepalensis* had sequence lengths similar to those in maize (approximately 180bp and 350bp), the sequences share no identity. Our *A. nepalensis* FISH also showed that the tested 180bp probe did not bind to all visible knobs, and we speculate that FISH using the 350bp repeat, which ranks second in abundance, would likely bind to some of the other visible knobs. From these FISH results, we conclude that genomic abundance is not predictive of centromere localization in the Andropogoneae.

Discussion

Our analyses of *de novo* assembled tandem repeats in grasses provides insight into the utility of this approach for studying the evolution of repetitive sequences. Most importantly, we show that previous assumptions about repeat abundance and location within the centromere do not hold across all taxa. Identification of the most abundant tandem repeat failed to identify

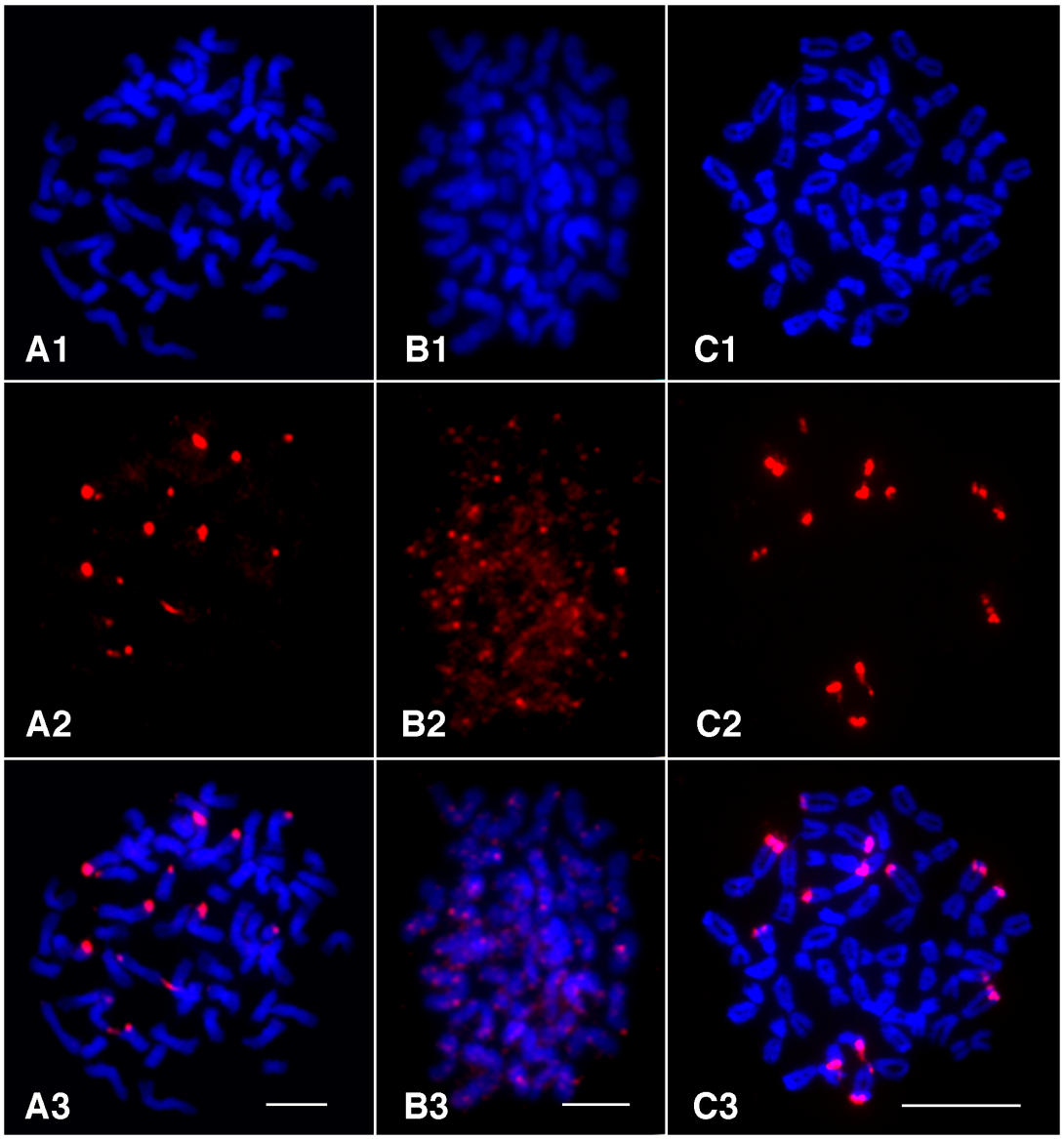

Fig 3. Fluorescent in situ hybridization of the highest abundant tandem repeats in three grasses. (A1-C1) Somatic metaphase chromosomes prepared from *A. nepalensis* (A1), *H. hirta* (B1), and *U. digitatum* (C1), respectively. (A2-C2) FISH signals derived from the three repeats identified in the three species. (A3-C3) Images merged from chromosomes and FISH signals. Scale bar = 10 microns. On all images, knobs are indicated with white arrows.

https://doi.org/10.1371/journal.pone.0177896.g003

centromeric repeats across many taxa, though in some cases it did identify sequences with homology to known centromere repeats. In *Tripsacum* taxa, previous work has shown that the maize tandemly repeated centromere element CentC cross-hybridized to the *Tripsacum* centromeres [34, 35]. As our FISH data show, *de novo* assembly and abundance ranking identified non-centromeric repeats in all taxa whose most abundant repeat did not share homology with a known repeat. Given the inconsistency of abundance as a predictor of centromere localization, we believe the alternative method of chromatin immunoprecipitation [13], despite its higher costs, is likely a more accurate and better method to reliably identify centromere repeats. Also, new sequencing technologies, such as Pacific Biosciences long reads, can also be

helpful in studying tandem repeats [22]. As costs decrease, long reads can be used alongside ChIP studies to identify higher order structure in tandem repeats and eventually assemble long tandem repeat arrays.

Though not ideal for centromere repeat identification, *de novo* assembly of tandem repeats can be an efficient, low cost method for characterizing repetitive content in non-model genomes. Our assembly of *A. nepalensis* and *U. digitatum* repeats serve as examples of novel findings that can be made regarding repeat sequences using this approach. *A. nepalensis*, sister to Andropogonae, has two highly abundant tandem repeats that do not share homology to any annotated genetic sequence, but are of similar sequence lengths of 180bp and 350bp as knob repeats in *Zea* and *Tripsacum* and found in knob-like heterochromatin. *U. digitatum* is similar, with the high abundance 184bp repeat associated with visible knobs and lacking homology >30bp to annotated sequence. Like in *Zea*, the *A. nepalensis* 180bp repeat is the highest abundance tandem repeat, and the 350bp tandem repeat is the next highest abundance tandem repeat with a different length (S1 and S2 Tables). While the sequence length of the tandem repeats are similar to those observed in many subtelomeric repeats [19], we speculate that the high genomic abundance of both the *A. nepalensis*, *U. digitatum*, and *Zea* may suggest that these new repeats are also knob-like. Knobs are associated with meiotic drive in maize [36] and suppress recombination locally but increase recombination in the intervening region between themselves and the centromere [37]. Knobs are known in a number of other plant taxa, such as maize, *Tripsacum*, rye [38], and *Arabidopsis thaliana* [39]. That we find no sequence homology between *A. nepalensis* or *U. digitatum* knobs and those in *Zea* suggests we may have identified a novel knob repeat that comprises a disproportionate fraction of the genome comparable only to certain maize and *Tripsacum* taxa, while not sharing homology to any known repeat. Further work will be necessary to identify whether the putative knobs of *A. nepalensis* or *U. digitatum* function similarly to those in maize with regard to recombination and meiotic drive, and analysis of additional taxa may reveal whether the accumulation of knobs near chromosome ends is also a common evolutionary theme [40].

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 and is associated with many important phenotypic traits such as flowering time and seed size [41, 42]. The ability to identify the percentage of the genome composed of specific types of tandem repeats can enable studies that track the components driving genome size variation. For example, identification of genomes with high abundance tandem repeats may lead to a better understanding of selfish genetic elements and how they influence long term evolution. Altogether, the results presented here show how *de novo* assembly can be used to better understand the repetitive fraction of the genome.

Supporting information

S1 Fig. Dot plot of the *A. nepalensis*, *T. laxum*, and *H. hirta* highest abundance contigs against themselves. Lines indicate share sequence identity.
(TIFF)

S1 Table. Percentage genomic composition of the top four tandem repeat groups. Species are ordered phylogenetically.
(PDF)

S2 Table. Monomer information for taxa studied.
(PDF)

Acknowledgments

JR-I would like to acknowledge support from USDA Hatch project CA-D-PLS-2066-H and NSF Plant Genome award IOS-0922703. PB would like to acknowledge support from the UC Davis Department of Plant Sciences and the DuPont Pioneer Howie Smith Honorary Plant Breeding Fellowship. We would also like to thank R. Kelly Dawe, Luis Avila, and Michelle Stitzer for helpful commentary on the manuscript.

Author Contributions

Conceptualization: PB MBH JJ JRI.

Data curation: PB AL.

Formal analysis: PB.

Funding acquisition: PB JJ JRI.

Investigation: PB AL YH PZ MCE.

Methodology: PB MBH AL YH PZ JJ JRI.

Writing – original draft: PB JRI.

Writing – review & editing: PB JJ JRI.

References

1. Treangen T. J. and Salzberg S. L. (2012). Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46. https://doi.org/10.1038/nrg3117 PMID: 22124482
2. ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74. https://doi.org/10.1038/nature11247 PMID: 22955616
3. Waterland R. A. and Jirtle R. L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Molecular and cellular biology, 23(15):5293–5300. https://doi.org/10.1128/MCB.23.15.5293-5300.2003 PMID: 12861015
4. Makarevitch I., Waters A. J., West P. T., Stitzer M., Hirsch C. N., Ross-Ibarra J., and Springer N. M. (2015). Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet*, 11(1):e1004915. https://doi.org/10.1371/journal.pgen.1004915 PMID: 25569788
5. Miura A., Yonebayashi S., Watanabe K., Toyama T., Shimada H., and Kakutani T. (2001). Mobilization of transposons by a mutation abolishing full dna methylation in arabidopsis. *Nature*, 411(6834):212–214. https://doi.org/10.1038/35075612 PMID: 11346800
6. Albert P., Gao Z., Danilova T., and Birchler J. (2010). Diversity of chromosomal karyotypes in maize and its relatives. *Cytogenetic and genome research*, 129(1-3):6–16. https://doi.org/10.1159/000314342 PMID: 20551613
7. Klemme S., Banaei-Moghaddam A. M., Macas J., Wicker T., Novák P., and Houben A. (2013). High-copy sequences reveal distinct evolution of the rye b chromosome. *New Phytologist*, 199(2):550–558. https://doi.org/10.1111/nph.12289 PMID: 23614816
8. Hobza R., Lengerova M., Svoboda J., Kubekova H., Kejnovsky E., and Vyskot B. (2006). An accumulation of tandem dna repeats on the y chromosome in silene latifolia during early stages of sex chromosome evolution. *Chromosoma*, 115(5):376–382. https://doi.org/10.1007/s00412-006-0065-5 PMID: 16612641
9. Melters D. P., Bradnam K. R., Young H. A., Telis N., May M. R., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*, 14(1):R10. https://doi.org/10.1186/gb-2013-14-1-r10 PMID: 23363705
10. Novák P., Neumann P., Pech J., Steinhaisl J., and Macas J. (2013). Repeatexplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29(6):792–793. https://doi.org/10.1093/bioinformatics/btt054 PMID: 23376349

11. Gong Z., Wu Y., Koblížková A., Torres G. A., Wang K., Iovene M., et al. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *The Plant Cell*, 24(9):3559–3574. https://doi.org/10.1105/tpc.112.100511 PMID: 22968715

12. Neumann P., Navrátilová A., Schroeder-Reiter E., Koblížková A., Steinbauerová V., et al. (2012). Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet*, 8(6): e1002777. https://doi.org/10.1371/journal.pgen.1002777 PMID: 22737088

13. Zhang H., Koblížková A., Wang K., Gong Z., Oliveira L., et al. (2014). Boom-bust turnovers of mega-base-sized centromeric dna in solanum species: rapid evolution of dna sequences associated with centromeres. *The Plant Cell*, 26(4):1436–1447. https://doi.org/10.1105/tpc.114.123877 PMID: 24728646

14. Jiang J., Birchler J. A., Parrott W. A., and Dawe R. K. (2003). A molecular view of plant centromeres. *Trends in plant science*, 8(12):570–575. PMID: 14659705

15. Kagansky A., Folco H. D., Almeida R., Pidoux A. L., Boukaba A., et al. (2009). Synthetic heterochromatin bypasses rnai and centromeric repeats to establish functional centromeres. *Science*, 324 (5935):1716–1719. https://doi.org/10.1126/science.1172026 PMID: 19556509

16. Wolfgruber T. K., Nakashima M. M., Schneider K. L., Sharma A., Xie Z., Albert P. S., et al. (2016). High quality maize centromere 10 sequence reveals evidence of frequent recombination events. *Frontiers in plant science*, 7. https://doi.org/10.3389/fpls.2016.00308 PMID: 27047500

17. Koo D.-H., Zhao H., and Jiang J. (2016). Chromatin-associated transcripts of tandemly repetitive dna sequences revealed by rna-fish. *Chromosome Research*, 24(4):467–480. https://doi.org/10.1007/s10577-016-9537-5 PMID: 27590598

18. Wong L. H., Brettingham-Moore K. H., Chan L., Quach J. M., Anderson M. A., Northrop E. L., et al. (2007). Centromere rna is a key component for the assembly of nucleoproteins at the nucleolus and centromere. *Genome research*, 17(8):1146–1160. https://doi.org/10.1101/gr.6022807 PMID: 17623812

19. Torres G. A., Gong Z., Iovene M., Hirsch C. D., Buell C. R., et al. (2011). Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3: Genes, Genomes, Genetics*, 1(2):85–92. https://doi.org/10.1534/g3.111.000125 PMID: 22384321

20. Paterson A. H., Bowers J. E., Bruggmann R., Dubchak I., Grimwood J., Gundlach H., et al. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556. https://doi.org/10.1038/nature07723 PMID: 19189423

21. Schnable P. S., Ware D., Fulton R. S., Stein J. C., Wei F., Pasternak S., et al (2009). The b73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115. https://doi.org/10.1126/science.1178534 PMID: 19965430

22. Bilinski P., Distor K., Gutierrez-Lopez J., Mendoza G. M., Shi J., Dawe R. K., and Ross-Ibarra J. (2014). Diversity and evolution of centromere repeats in the maize genome. *Chromosoma*, pages 1–9. https://doi.org/10.1007/s00412-014-0483-8 PMID: 25190528

23. Estep M.C., McKain M. R., Dilys Vela D., Zhong J., Hodge J. G., Hodkinson T. R., et. Allopolyploidy, diversification, and the miocene grassland expansion. *Proceedings of the National Academy of Sciences*, 111(42):15149–15154, 2014.

24. Bennett MD and Leitch IJ. Plant dna c-values database. *Royal Botanic Gardens, Kew*, 2005.

25. Peng Z., Lu Y., Li L., Zhao Q., Feng Q., Gao Z., et al. The draft genome of the fast-growing non-timber forest species moso bamboo (phyllostachys heterocycla). *Nature Genetics*, 45(4):456–461, 2013. https://doi.org/10.1038/ng.2569 PMID: 23435089

26. Chevreux B., Wetter T., Suhai S., et al. (1999). Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics*, volume 99, pages 45–56. Heidelberg.

27. Benson G. (1999). Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573. https://doi.org/10.1093/nar/27.2.573 PMID: 9862982

28. Lee W.-P., Stromberg M. P., Ward A., Stewart C., Garrison E. P., and Marth G. T. (2014). Mosaik: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one*, 9(3): e90581. https://doi.org/10.1371/journal.pone.0090581 PMID: 24599324

29. Tenaillon M. I., Hufford M. B., Gaut B. S., and Ross-Ibarra J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and zea luxurians. *Genome biology and evolution*, 3:219–229. https://doi.org/10.1093/gbe/evr008 PMID: 21296765

30. Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M., and Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic acids research*, 40(15):e115–e115. https://doi.org/10.1093/nar/gks596 PMID: 22730293

31. Jiang J., Gill B. S., Wang G.-L., Ronald P. C., and Ward D. C. (1995). Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes.

Proceedings of the National Academy of Sciences, 92(10):4487–4491. https://doi.org/10.1073/pnas.92.10.4487 PMID: 7753830

32. Miller J., Jackson S., Nasuda S., Gill B., Wing R. A., and Jiang J. (1998). Cloning and characterization of a centromere-specific repetitive dna element from sorghum bicolor. *Theoretical and Applied Genetics*, 96(6-7):832–839. https://doi.org/10.1007/s001220050809

33. Estep MC, DeBarry JD, and Bennetzen JL. The dynamics of ltr retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity*, 110(2):194–204, 2013. https://doi.org/10.1038/hdy.2012.99 PMID: 23321774

34. Lamb, J. C. and Birchler, J. A. (2006). Retroelement genome painting: cytological visualization of retroelement expansions in the genera zea and tripsacum. *Genetics*, 173(2):1007–1021. https://doi.org/10.1534/genetics.105.053165 PMID: 16582446

35. Zhu Q., Cai Z., Tang Q., and Jin W. (2016). Repetitive sequence analysis and karyotyping reveal different genome evolution and speciation of diploid and tetraploid tripsacum dactyloides. *The Crop Journal*. https://doi.org/10.1016/j.cj.2016.04.003

36. Dawe R. K. and Cande, W. Z. (1996).Induction of centromeric activity in maize by suppressor of meiotic drive 1. *Proceedings of the National Academy of Sciences*, 93(16):8512–8517. https://doi.org/10.1073/pnas.93.16.8512 PMID: 8710901

37. Buckler E. S., Phelps-Durr T. L., Buckler C. S. K., Dawe R. K., Doebley J. F., and Holtsford T. P. (1999). Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics*, 153(1):415–426. PMID: 10471723

38. Gill B. S. and Kimber G. (1974). The giemsa c-banded karyotype of rye. *Proceedings of the National Academy of Sciences*, 71(4):1247–1249. https://doi.org/10.1073/pnas.71.4.1247 PMID: 4133848

39. Fransz P. F., Armstrong S., de Jong J. H., Parnell L. D., van Drunen C., Dean C., Zabel P., Bisseling T., and Jones, G. H. (2000). Integrated cytogenetic map of chromosome arm 4s of a. thaliana: structural organization of heterochromatic knob and centromere region. *Cell*, 100(3):367–376. https://doi.org/10.1016/S0092-8674(00)80672-8 PMID: 10676818

40. Heslop-Harrison J. and Schwarzacher T. (2011). Organisation of the plant genome in chromosomes. *The Plant Journal*, 66(1):18–33. https://doi.org/10.1111/j.1365-313X.2011.04544.x PMID: 21443620

41. Rayburn A. L., Dudley J., and Biradar D. (1994). Selection for early flowering results in simultaneous selection for reduced nuclear dna content in maize. *Plant Breeding*, 112(4):318–322. https://doi.org/10.1111/j.1439-0523.1994.tb00690.x

42. Knight C. A., Molinari N. A., and Petrov D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*, 95(1):177–190. https://doi.org/10.1093/aob/mci011 PMID: 15596465