Diverse origins of high copy tandem repeats in grass genomes

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

[1] Department of Plant Sciences, University of California, Davis, Davis, California, United States
[2] School of Life Sciences, Jiangsu Normal University, Xuzhou, China
[3] Department of Horticulture, University of Wisconsin-Madison, Madicon, Wisconsin, United States
[4] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, United States
[5] College of Bioscience and Biotechnology, Beijing Forestry University, Beijing, China
[6] Department of Plant Sciences, Center for Population Biology, and Genome Center, University of California, Davis, Davis, CA, United States

Corresponding Authors: Paul Bilinski, Jeffrey Ross-Ibarra
Email address: pbilinsk@gmail.com, rossibarra@ucdavis.edu

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 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2314v1 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

Diverse origins of high copy tandem repeats in grass genomes

Paul Bilinski[1], Yonghua Han[2], Matthew B. Hufford[3], Anne Lorant[1], Pingdong Zhang[4], Jiming Jiang[2], and Jeffrey Ross-Ibarra[1,5,6]

[1]Dept. of Plant Sciences, University of California, Davis, CA, USA
[2]School of Life Sciences, Jiangsu Normal University, Xuzhou 221116, China
[3]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
[4]College of Bioscience and Biotechnology, Beijing Forestry University, China
[5]Dept. of Horticulture, University of Wisconsin-Madison, Madison, WI USA
[6,7]Genome Center and Center for Population Biology, University of California, Davis, CA, USA

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 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.

Keywords: Centromere, tandem repeats, heterochromatin

INTRODUCTION

Sequencing technologies have facilitated genome assembly for many non-model organisms, bringing a tremendous amount of data to the field of comparative genomics. Assembly of repetitive regions is limited by shotgun sequencing, leading to overrepresentation of genic regions in assembled genomes. Though repetitive DNA was once 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 (Consortium et al., 2012). Plant genomes can be highly repetitive, and individual repeat classes are often present at extremely high copy numbers in the genome (Pearce et al., 1996). Plant repeats can be classified in two broad categories: dispersed repeats 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 (Waterland and Jirtle, 2003; Makarevitch et al., 2015) and chromatin status (Miura et al., 2001), functions which can have strong impacts on overall phenotype.

In comparison to the wealth of TE data across organisms, little is known about the function and evolutionary history of tandem repeats. Tandem repeats contribute fewer base pairs to the genome than TEs, but the total number of nucleotides derived from tandem repeats varies substantially across phylogenetic groups (Melters et al., 2013). Tandem repeats are commonly found in the gene poor but

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2314v1 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

structurally important telomeres and centromeres. Tandem repeats do not appear necessary for the formation of centromeres (Jiang et al., 2003), however, and may instead serve as placeholders for an epigenetic signal that governs heterochromatin formation (Kagansky et al., 2009) or function in repair of double strand breaks (Wolfgruber et al., 2016). Tandem repeats are also found in other types of heterochromatin such as the large chromosomal features known as knobs in the genus *Zea* and closely related taxa (Albert et al., 2010). Knobs suppress local recombination (Chang and Kikudome, 1974)) and in some backgrounds are involved in meiotic drive (Buckler et al., 1999), but little is known of their origin.

In an effort to better understand tandemly repeated sequence, researchers have applied a combination of sequencing technologies and molecular biology. For example, studies that have paired chromatin immunoprecipitation (ChIP) against centromere proteins with bioinformatic identification of repetitive sequence have successfully identified centromere repeats (Gong et al., 2012; Neumann et al., 2012; Zhang et al., 2014). However, high-throughput ChIP across a broad sample of taxa is difficult to perform, costly, and labor intensive, leading some researchers to instead use bioinformatic approaches to explore whole genome short read data. RepeatExplorer (Novák et al., 2013), for example, clusters reads to identify repeat groups and their genomic abundance, and has been used in several studies to identify the repetitive landscape of plant genomes (Weiss-Schneeweiss et al., 2015), and paired with ChIP data to identify centromere clusters (Zhang et al., 2014). Taking a different approach, Melters *et al.* (Melters et al., 2013) conducted *de novo* repeat assembly of published short read sequence data, using consensus sequences to identify tandem repeats across 280 plant and animal species. One critical assumption of this latter approach, however, was that the most abundant tandem repeat in all taxa was the centromere repeat. While comparison to known repeats in several model organisms suggestions this assumption works well for animals, earlier work suggests that it may not apply broadly to plants. Using a similar pipeline and 454 shotgun reads from *Solanum* species, Torres *et al.* Torres et al. (2011) identified two subtelomeric repeats as the most abundant tandem repeats genome wide based on the highest frequency kmer.

Here, we apply the basic pipeline of tandem repeat consensus assembly to species within the Andropogoneae tribe of the grasses in order to better understand tandem repeat contribution to genomic composition. The Andropogoneae tribe includes both maize and sorghum, two model organisms with well-known repeats (Paterson et al., 2009; Schnable et al., 2009) that allow us to test the accuracy of our method and the Melters et al. (2013) assumption regarding centromere repeat sequence and its genomic abundance. Because previous work has shown that sequencing libraries prepared through identical methods better retain relative composition of repeats (Bilinski et al., 2014), rather than use published data we elect to re-sequence all the species used here. We examine 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 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 *Arundinella*.

MATERIALS AND METHODS

Sequencing

Seed was requested from the GRIN database, and accession information is available in Suppl. Table S1. 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`), and one pair of reads were used for all analyses.

Phylogenetic Tree Reconstruction

We downloaded sequence data for two ribosomal inter-genic spacers and one chloroplast gene at NCBI (sequences are available on github). Sequences were aligned using seven iterations of MUSCLE (Edgar, 2004), and concatenated in order to build a neighbor joining tree using Jukes-Cantor distance implemented in Geneious (v5.4.4) (Kearse et al., 2012). The topology of the NJ tree broadly agrees with previously published phylogenies (Wu and Ge, 2012; Skendzic et al., 2007), though variation exists where some nodes are collapsed into polytomies.

Assembly and Genomic Composition of Centromere Repeats

To assemble contigs from low coverage sequence, 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). We ran Tandem Repeat Finder (Benson, 1999) (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 percentage of genomic composition of each tandemly repeated contig, we used Mosaik (Lee et al., 2014), which stores information about multiply mapping reads (version 1.0; parameters optimized for tandem repetitive elements as in (Bilinski et al., 2014)). Low coverage libraries (<0.1X) were mapped against the contigs identified by TRF and contigs were ranked by the number of reads aligned. 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 and removed assemblies with BLAST homology. This process was repeated 4 times to identify the genomic composition of the 4 highest abundance tandem repeat groups. 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.

Fluorescent In-Situ Hybridization

Primers were designed from the computationally identified tandem repeats. Repetitive sequences were amplified using the genomic DNA isolated from the targeted species and labeled with digoxigenin-11-dUTP. FISH was performed using published procedures (Jiang et al., 1995). 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: *Arundinella* Primer F-CCATTCAAGAAATGGTGTCA; *Arundinella* Primer R-GCAAGTACGAAAGCCAAAAT; *Urelytrum* Primer F-GCACTGGCCCTGAGAGAAAT; *Urelytrum* Primer R- ACAGGCTTGGGTGGACAAAA; *Hyparrhenia* Primer F- GATCCGAAAGTCGCGAAACG; *Hyparrhenia* Primer R- TTTTTCGCAACGAACGCACA. FISH was performed using published procedures (Jiang et al., 1995).

RESULTS

Assembly of low depth Illumina data produced several thousand contigs in each species from our panel (Fig. 1, and Supp. Table S1). From these, TRF identified between 300 and 15,000 tandem repeat contigs in each taxon. The number of tandem repeat contigs varied across taxa based on coverage and overall genomic repetitive content. Illumina data were mapped against tandem repeat contigs to approximate abundance of tandem repeats in our panel (Fig. 2). 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 *Arundinella nepalensis*, though *Tripsacum* species show large variation. Based on genome size estimates from the Kew C-Value database (`http://data.kew.org/cvalues/`), the correlation between total tandem repeat content and genome size is poor (r=0.05, >0.05).

In order to investigate the proportional contribution of the most common tandem repeat classes in each of our taxa, 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 (Fig. 3. 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*,

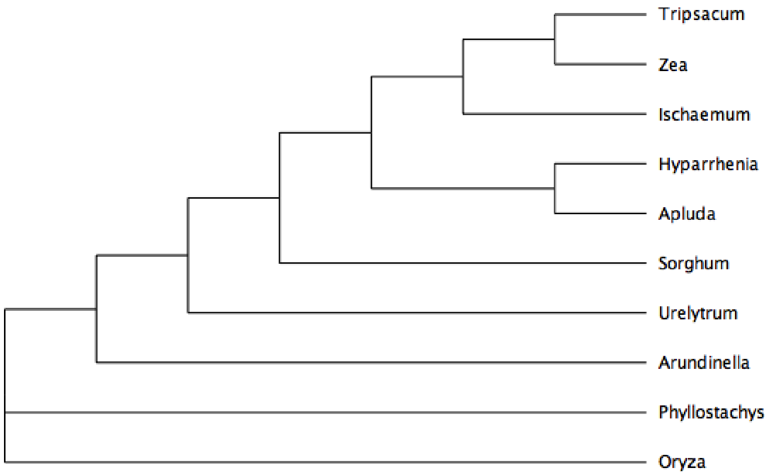

Figure 1. Neighbor joining tree of evolutionary relationships between the grasses studied. Detailed discussion of relationships among these taxa is available in (Wu and Ge, 2012; Skendzic et al., 2007). *Oryza* and *Phyllostachys* are both outside the Andropogoneae tribe.

while the top ranked contig showed immense variation, the second ranked contig had a relatively constant abundance near half a percent.

We tested the assumption that the most abundant repeat is centromeric (Melters et al., 2013) in taxa with both known and uncharacterized centromere repeats. Among taxa with known centromere repeats, the centromere repeats was found to be the most abundant tandem repeat in both *Oryza* and *Sorghum*, but in *Zea* and *Tripsacum*, while the centromere repeat was among the four most abundant, the highest abundance repeat came instead from heterochromatic knobs. While the centromere repeat was not previously known for the genus *Apluda*, its highest abundance contig shared homology and a common monomer repeat length with the *Sorghum* centromere repeat. The top-ranked contig in *Ischaemum* shared a monomer length identical to *Sorghum*, but with no sequence homology. The top ranked contigs from the remaining taxa in our panel bore no similarity to known centromere repeats. To test whether the most abundance repeat in these taxa was centromeric, we performed fluorescent in situ hybridization (FISH; Fig. 4), expecting spatial clustering of the probe in the interior (for metacentric) or end (for acrocentric) of most if not all chromosomes. FISH from the *de novo* constructed repeat of *Hyparrhenia* is widely dispersed across the genome, a pattern expected from a TE rather than a tandem repeat. In congrast, the tandem repeat from *Urelytrum* showed strong spatial clustering, but clusters were not found on all chromosomes and were associated with chromosome ends as might be expected from subtelomeric sequence. The regions probed in *Urelytrum* did not associate with visible knobs, as they were not found in regions of tightly packed heterochromatin. The probed repeat of *Arundinella* also showed subtelomeric clustering, but clusters were found in highly compacted chromatin suggesting that the probe bound to a knob-like repeat rather than a low copy subtelomeric repeat. The fact that *Arundinella* had the largest proportion of its genome comprised of tandem repeats (Fig. 2) is also consistent with a knob-like origin for this tandem repeat. While the knob repeat sequences in *Arundinella* had sequence lengths similar to those in maize (approximately 180bp and 350bp), the sequences share no identity. Our *Arundinella* FISH also showed that no single probe bound to all visible knobs. From these FISH results, we conclude that genomic abundance is not uniformly 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 sequence. Most importantly, we show that previous assumptions about repeat abundance and location within the centromere do not hold across all taxa. *De novo* assembly to identify centromere repeats only functioned in species where repeats shared homology

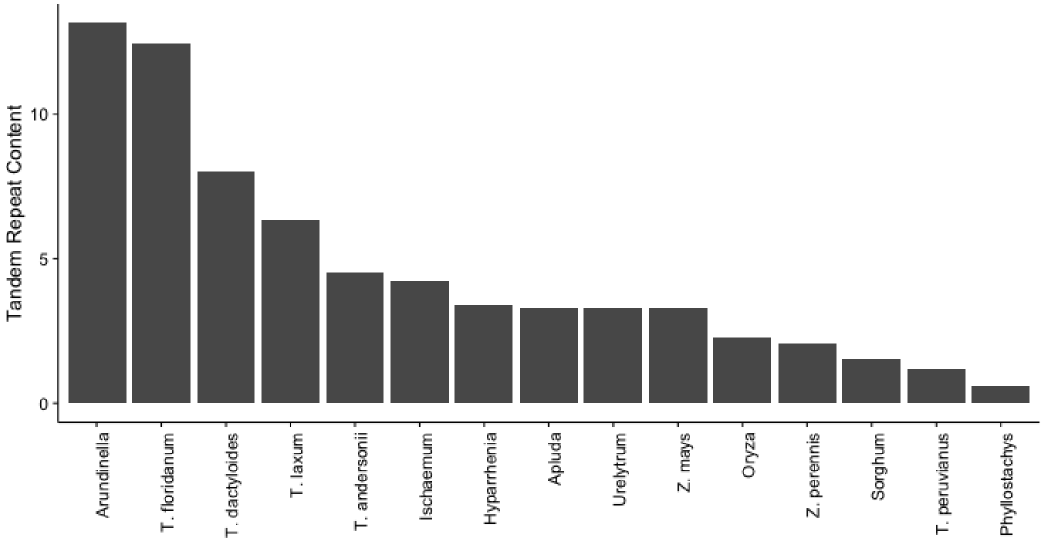

Figure 2. Percentage genomic composition of all tandem repeat contigs in monocot taxa. Values are derived from the proportion of all reads mapping to any tandemly repetitive contig derived from TRF after MIRA assembly. Species are ordered from highest to lowest percentage tandem repeat content.

to known centromere repeats. 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 with CenH3 proteins (ChiP) (Zhang et al., 2014) is likely a better method to reliably identify centromere repeats.

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, an area of study generally left untouched due to the difficulties of traditional assembly. Our assembly of *Arundinella* repeats serves as an example of evolutionary inferences that can be made regarding repeat sequences using this approach *Arundinella*, sister to all other species in this study, has two highly abundant tandem repeats that do not share homology to any annotated genetic sequence. Our cytological work suggests that these two sequences derive from knob-like heterochromatin. Knobs are associated with meiotic drive in maize (Dawe and Cande, 1996) and suppress recombination locally but increase recombination in the intervening region between themselves and the centromere Buckler et al. (1999). Knobs are known in a number of other plant taxa, such as maize, *Tripsacum*, rye (Gill and Kimber, 1974), and *Arabidopsis* (Fransz et al., 2000). That we find no sequence homology between *Arundinella* knobs and those in *Zea* suggests we have identified an entirely novel knob system. Interestingly, the lengths of the knob variants in *Arundinella* and *Zea* are similar, centered around 180bp and 360bp. These approximate lengths are observed in many subtelomeric repeats Torres et al. (2011), though the high genomic abundance of the *Arundinella* and *Zea* repeats may be unique. Further work will be necessary to identify whether the knobs of *Arundinella* function similarly to those in maize with regard to recombination and meiotic drive, but our findings suggest that knobs may be a more common genomic feature than previously believed. Future investigations in additional taxa may reveal whether the accumulation of knobs near chromosome ends is a common evolutionary theme.

The ability to look broadly across a phylogeny at consensus repeats and idnetify novel repeats in previously unstudied organisms has the potential to produce phylogentically relevant data, shedding light on the evolution of the repetitive fraction of the genome. Recently, researchers have shown that information from the repetitive fraction of genomes has phylogentically relevant signal (Dodsworth et al., 2015), showing one possible avenue of using repeat sequence to inform species relationships. Consistent with this idea, we found closely related taxa had similar rank abundance of tandem repeats. Future work with a higher density of sampling could provide insight into sequence turnover in repetitive regions

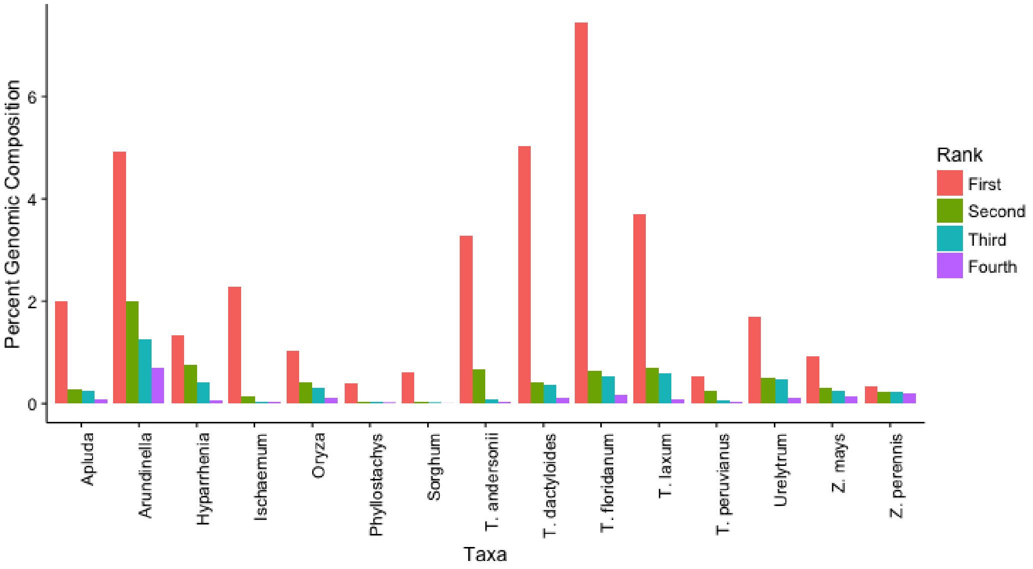

Figure 3. Genomic Composition of Top 4 Tandemly Repetitive Contigs. The top 4 contigs in each species were defined as not having homology to one another, in order to identify independent repeat motifs. Species are ordered alphabetically by genus.

(Henikoff et al., 2001) and discover 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 and is associated with many important phenotypic traits such as flowering time and seed size (Rayburn et al., 1994; Knight et al., 2005). The ability to identify the percentage of the genome composed by specific types of 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 or are under selection. Looking across species, repetitive composition can inform our understanding of speciation, showing for example how often centromere repeat divergence co-occurs with or without speciation(Pertile et al., 2009). Also, identification of genomes with high abundance of tandem repeats may lead to a better understanding of selfish genetic elements and how they may 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.

ACKNOWLEDGMENTS

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 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.

REFERENCES

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.

Benson, G. (1999). Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573.

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.

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.

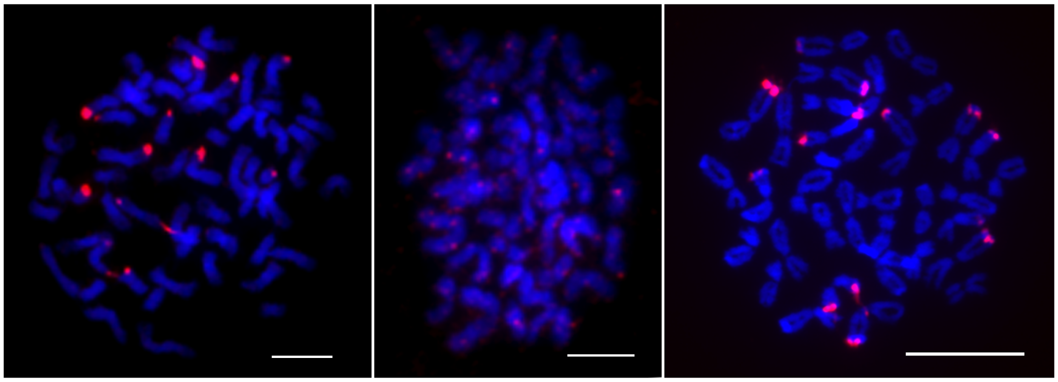

Figure 4. Fluorescent in-situ hybridization for the highest abundance tandem repeat monomer in for three grasses. (a) Arundinella, (b) Hyparhenia, (c) Urelytrum. For probe information, see methods. Scale bar = 10 microns.

Chang, C. and Kikudome, G. Y. (1974). The interaction of knobs and b chromosomes of maize in determining the level of recombination. *Genetics*, 77(1):45–54.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

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.

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novák, P., Piednoël, M., Weiss-Schneeweiss, H., and Leitch, A. R. (2015). Genomic repeat abundances contain phylogenetic signal. *Systematic biology*, 64(1):112–126.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797.

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.

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.

Gong, Z., Wu, Y., Koblížková, A., Torres, G. A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C. R., et al. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *The Plant Cell*, 24(9):3559–3574.

Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving dna. *Science*, 293(5532):1098–1102.

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.

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.

Kagansky, A., Folco, H. D., Almeida, R., Pidoux, A. L., Boukaba, A., Simmer, F., Urano, T., Hamilton, G. L., and Allshire, R. C. (2009). Synthetic heterochromatin bypasses rnai and centromeric repeats to establish functional centromeres. *Science*, 324(5935):1716–1719.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649.

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.

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.

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.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*, 14(1):R10.

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.

Neumann, P., Navrátilová, A., Schroeder-Reiter, E., Koblížková, A., Steinbauerová, V., Chocholová, E., Novák, P., Wanner, G., and Macas, J. (2012). Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet*, 8(6):e1002777.

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.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556.

Pearce, S. R., Li, D., Flavell, A., Harrison, G., Heslop-Harrison, J., and Kumar, A. (1996). Thety1-copia group retrotransposons invicia species: copy number, sequence heterogeneity and chromosomal localisation. *Molecular and General Genetics MGG*, 250(3):305–315.

Pertile, M. D., Graham, A. N., Choo, K. A., and Kalitsis, P. (2009). Rapid evolution of mouse y centromere repeat dna belies recent sequence stability. *Genome research*, 19(12):2202–2213.

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.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., and Graves, T. A. (2009). The b73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115.

Skendzic, E. M., Columbus, J. T., and Cerros-Tlatilpa, R. (2007). 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*, 23(1):530–544.

Torres, G. A., Gong, Z., Iovene, M., Hirsch, C. D., Buell, C. R., Bryan, G. J., Novák, P., Macas, J., and Jiang, J. (2011). Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3: Genes, Genomes, Genetics*, 1(2):85–92.

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.

Weiss-Schneeweiss, H., Leitch, A. R., McCann, J., Jang, T.-S., and Macas, J. (2015). Employing next generation sequencing to explore the repeat landscape of the plant genome. *Next Generation Sequencing in Plant Systematics. Regnum Vegetabile*, 157.

Wolfgruber, T. K., Nakashima, M. M., Schneider, K. L., Sharma, A., Xie, Z., Albert, P. S., Xu, R., Bilinski, P., Dawe, R. K., Ross-Ibarra, J., et al. (2016). High quality maize centromere 10 sequence reveals evidence of frequent recombination events. *Frontiers in plant science*, 7.

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

Zhang, H., Koblížková, A., Wang, K., Gong, Z., Oliveira, L., Torres, G. A., Wu, Y., Zhang, W., Novák, P., Buell, C. R., et al. (2014). Boom-bust turnovers of megabase-sized centromeric dna in solanum species: rapid evolution of dna sequences associated with centromeres. *The Plant Cell*, 26(4):1436–1447.

Genus	Species	Reads	AccessionID
Apluda	mutica	746994	PI 219568
Arundinella	nepalensis	662118	PI 384059
Hyparrhenia	hirta	861995	PI 206889
Ischaenum	rugosum	920258	Kew 0183574
Phyllostachys	edulis	628030	NA
Zea	mays	4422188	RIMMA0019
Sorghum	bicolor sp bicolor	473944	PI 564163
Tripsacum	andersonii	288175	MIA 34430
Tripsacum	dactyloides	391848	MIA 34597
Tripsacum	floridanum	743668	MIA 34719
Tripsacum	laxum	723097	MIA 34792
Tripsacum	peruvianum	238983	MIA 34501
Triticum	urartu	435815	PI 428198
Urelytrum	digitatum	661535	SM3109
Zea	perennis	5106091	NA

Table S1. 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.