Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage

Chodon Sass, William JD Iles, Craig Barrett, Selena Y Smith, Chelsea D Specht

The Zingiberales are an iconic order of monocotyledonous plants comprising eight families with distinctive and diverse floral morphologies and representing an important ecological element of tropical and subtropical forests. While the eight families are demonstrated to be monophyletic, phylogenetic relationships among these families remain unresolved. Neither combined morphological and molecular studies nor recent attempts to resolve family relationships using sequence data from whole plastomes has resulted in a well-supported, ordinal-level phylogenetic hypothesis of relationships. Here we approach this challenge by leveraging the complete genome of one member of the order, *Musa acuminata*, together with transcriptome information from each of the other seven families to design a set of nuclear loci that can be enriched from highly divergent taxa with a single array-based capture of indexed genomic DNA. A total of 494 exons from 418 nuclear genes were captured for 53 ingroup taxa. The entire plastid genome was also captured for the same 53 taxa. Of the total genes captured, 308 nuclear and 68 plastid genes were used for phylogenetic estimation. The concatenated plastid and nuclear dataset supports the position of Musaceae as sister to the remaining seven families. Moreover, the combined dataset recovers known intra- and inter-family phylogenetic relationships with generally high bootstrap support. This is a flexible and cost effective method that gives the broader plant biology community a tool for generating phylogenomic scale sequence data in non-model systems at varying evolutionary depths.

Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage

Chodon Sass[1], William J.D. Iles[1], Craig Barrett[2, 3], Selena Y. Smith[4], Chelsea D. Specht*,[1]

[1]Department of Plant and Microbial Biology, Department of Integrative Biology and the University and Jepson Herbaria, University of California at Berkeley, Berkeley, CA, USA.

[2]Department of Biology, California State University, Los Angeles, CA USA.

[3]Current address: Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV USA

[4]Department of Earth & Environmental Sciences and the Museum of Paleontology, University of Michigan, Ann Arbor, MI USA.

*** Correspondence:**
Dr. Chelsea D. Specht
111 Koshland Hall MC3102
University of California
Berkeley, CA 94720, USA
Email: cdspecht@berkeley.edu

Keywords: array-based capture, ancient radiation, exon capture, high-throughput sequencing, HTS, Heliconiaceae, Musaceae, Gingers, Banana

Abstract

The Zingiberales are an iconic order of monocotyledonous plants comprising eight families with distinctive and diverse floral morphologies and representing an important ecological element of tropical and subtropical forests. While the eight families are demonstrated to be monophyletic, phylogenetic relationships among these families remain unresolved. Neither combined morphological and molecular studies nor recent attempts to resolve family relationships using sequence data from whole plastomes has resulted in a well-supported, ordinal-level phylogenetic hypothesis of relationships. Here we approach this challenge by leveraging the complete genome of one member of the order, *Musa acuminata*, together with transcriptome information from each of the other seven families to design a set of nuclear loci that can be enriched from highly divergent taxa with a single array-based capture of indexed genomic DNA. A total of 494 exons from 418 nuclear genes were captured for 53 ingroup taxa. The entire plastid genome was also captured for the same 53 taxa. Of the total genes captured, 308 nuclear and 68 plastid genes were used for phylogenetic estimation. The concatenated plastid and nuclear dataset supports the position of Musaceae as sister to the remaining seven families. Moreover, the combined dataset recovers known intra- and inter-family phylogenetic relationships with generally high bootstrap support. This is a flexible and cost effective method that gives the broader plant biology community a tool for generating phylogenomic scale sequence data in non-model systems at varying evolutionary depths.

Introduction

Zingiberales are a diverse group of tropical monocots, including important tropical crop plants (e.g., ginger, turmeric, cardamom, bananas) and ornamentals (e.g., cannas, bird-of-paradise, prayer plants). Eight families are recognized with a total of ca. 2500 species. Fossil zingibers are known since the Cretaceous, and show a mix of characters from Musaceae and Zingiberaceae (Friis, 1988; Rodriguez-de la Rosa & Cevallos-Ferriz, 1994; Iles et al., 2015) on the basis of fruits, seeds, leaves, rhizomes, and phytoliths (Friis, Crane & Pedersen, 2011; Chen & Smith, 2013). Zingiberales are thought to have diverged from the sister order Commelinales (sensu Angiosperm Phylogeny Group, 2003) ca. 124 Ma, with diversification into the major lineages occurring from ca. 110–100 Ma (Kress & Specht, 2006). However, relationships among the families are not well resolved using multi-gene phylogenies (Kress et al., 2001; Barrett et al., 2014), likely due to this early rapid radiation. Specifically, the relationship between Musaceae, Strelitziaceae + Lowiaceae, Heliconiaceae, and the remaining four families, which form a well-supported monophyletic group (i.e., the 'ginger clade'), have conflicting support among studies. Whole plastid data for 14 taxa spanning the eight families still failed to resolve the early diverging branches of the phylogeny, perhaps owing to limited sampling and a lack of phylogenetic signal in the plastome data (Barrett et al., 2014). However challenging to resolve, rapid evolutionary radiations are thought to be a common theme across the tree of life and are thought to explain poorly-resolved phylogenies in many groups including insects, birds, bees, turtles, mammals, and angiosperms (Whitfield & Lockhart, 2007; Whitfield & Kjer, 2008).

The advent of high throughput sequencing and methods that extend the utility of new sequencing technology to non-model organisms has enabled sequence-based understanding of evolutionary relationships in previously intractable groups (Crawford et al., 2012; Faircloth et al., 2012; Lemmon, Emme & Lemmon, 2012; Bi et al., 2013). Specifically, for phylogenetic studies, multiple genes containing appropriate levels of sequence divergence can now be obtained for many phylogenetically distant individuals. Various genome enrichment methods, using hybridization to capture a targeted set of genes based on appropriately designed nucleotide probes, have enabled targeted sets of hundreds or thousands of loci to be sequenced in parallel for multiple individuals. However, the ability to capture loci across relatively deep phylogenetic scales has remained challenging because of the inverse relationship between capture efficiency and the evolutionary distance from the individual(s) used to design the probes (Bi et al., 2012; Lemmon, Emme & Lemmon, 2012; Peñalba et al., 2014; Weitemier et al., 2014). For very deep divergences in animals, to understand amniote evolution or deep divergences in vertebrate evolution for example, ultra-conserved elements (Faircloth et al., 2012) and anchored hybrid enrichment (Lemmon, Emme & Lemmon, 2012) have been used to target conserved loci that are flanked by less conserved regions. However, these regions were developed using animal genomes and are unsuitable for use in plants (Reneker et al., 2012).

Historical whole genome duplication followed by fractionation and diploidization, genome-level processes that are common during plant evolution and occur in a lineage-specific manner, make it likely that loci with known orthology will need to be tested and developed separately for each plant lineage. Some methods have been developed for lineage specific capture, such as whole exome capture (Bi et al., 2012) that uses a transcriptome sequence and a relatively closely related sequenced genome to design lineage-specific baits. This approach was modified and recently used in plants (Weitemier et al., 2014). However, the success of these approaches to capture targeted genes is limited by the distance of the samples to the target transcriptome. A more flexible approach uses PCR products to generate a home-made, in-solution capture (Maricic, Whitten & Pääbo, 2010; Peñalba et al., 2014), but this requires some prior knowledge of locus sequence and primer optimization and likely is most useful to target 10–50 loci with known phylogenetic utility.

In the case of the Zingiberales, with approximately 110 Myr of divergence since the initial lineage diversification leading to the modern families, it is necessary to design a set of probes that can capture sequences with a relatively high percentage of polymorphisms, yet still allow the reliable assignment of orthology to captured sequences. In order to do this, we used transcriptomes that were generated as part of the Monocot Tree of Life Project (MonAToL http://www.botany.wisc.edu/monatol/) or One Thousand Plant Transcriptomes (OneKP https://sites.google.com/a/ualberta.ca/onekp/home) together with the annotated whole genome of *Musa acuminata* (D'Hont et al., 2012) to design a set of probes that were printed on an Agilent microarray chip in parallel. This parallel printing approach enables divergent taxa to be captured on a single array and alleviates binding competition between closely related and divergent individuals. Simultaneously, we captured whole plastid genomes based on published plastid genomes from one member each of the eight families (Barrett et al., 2014).

We show the utility of this cost effective method in generating phylogenetically informative sequence data by constructing a phylogenetic tree of the Zingiberales that recaptures known relationships and resolves previously recalcitrant parts of phylogeny with high support. Because of the phylogenetic breadth of transcriptomes becoming publically available across the plant kingdom, this method has the potential to aid in the design of lineage specific sequencing projects that span phylogenetic distances on the order of 100 Myr or possibly greater.

Methods

Taxon Sampling, DNA Extraction, and Library Preparation

Sampling included several members of each of the eight families: Heliconiaceae (5), Musaceae (9, including 2 previously published whole genomes, D'Hont et al., 2012; Davey et al., 2013), Strelitziaceae (3), Lowiaceae (2), Zingiberaceae (16), Costaceae (10), Marantaceae (7), and Cannaceae (3). In total, 53 individuals were sequenced *de novo* (Table S1). DNA was extracted using an SDS and salt extraction protocol (Edwards, Johnstone & Thompson, 1991; Konieczny & Ausubel, 1993) from freshly collected leaves dried in silica, eluted in TE buffer, and sonicated with a Bioruptor® (Diagenode) or qSonica Q800R machine to an average size of approximately 250bp. Sonicated DNA was cleaned and concentrated with solid phase reversible immobilization magnetic beads (Sera-Mag), and libraries were prepared according to Meyer & Kircher (2010).

Probe Design, Sequence Capture, Sequencing

To generate a nuclear probe set, the *Musa acuminata* CDS was downloaded from the banana genome hub (http://banana-genome.cirad.fr/) and split into annotated exons. Raw reads of transcriptomes for each of the remaining seven families were cleaned to remove adapters, low-complexity sequences, contamination, and PCR duplicates (Singhal & Moritz, 2012). Cleaned transcriptome reads were aligned to the *Musa acuminata* exons using NovoAlign v3.01 (http://novocraft.com) with –t 502 to allow highly divergent sequences to map. After mapping, SNPs were called using SAMtools v0.1.18 (Li et al., 2009) and VarScan v2.3.6 (Koboldt et al.,

2012) and consensus sequences for each family were made based on SNP calls. All exons were filtered for: (1) having overlapping read coverage in all 7 families (2) being longer than 150 bp (3) having between 30–70% GC content (4) being unique by reciprocal BLAST (5) not being found in the RepeatMasker database (command parameters can be found in Supplementary Methods). After filtering, a total of 494 exons from 418 genes for each of the eight families (the *Musa* reference sequence plus each sequence from the seven families) were printed with 1 bp tiling twice each on an Agilent 1M microarray chip (G3358A) (Figure 1a). A second chip was printed with one complete plastid genome from each family (Barrett et al., 2014) with slightly less than 1 bp tiling. Libraries from a total of 56 individuals were quantified by Qubit® and pooled in equimolar quantities. The total library pool was split in half and one half was hybridized to the nuclear array and the other half was hybridized to the plastid array (Hodges et al., 2009). After hybridization, pools were subject to a limited amount of PCR amplification and enrichment success was verified with qPCR using primers matching both targeted and non-targeted regions. Because of known bias toward plastid dominance in sequenced reads owing to a greater percentage of plastid DNA in the total genomic DNA extractions, the separate hybridization pools were combined in a ratio of 3 parts nuclear to 1 part plastid and sequenced (100 bp paired-end reads) in one lane of a Illumina® HiSeq® 2500 platform at the Vincent J. Coates Genomics Sequencing Facility at the University of California, Berkeley.

Read Processing

Raw reads were cleaned to remove adapters, low-complexity sequences, contamination, and PCR duplicates (Singhal & Moritz, 2012). Custom Perl scripts were created to perform a series of alignment and reference adjustments using NovoAlign v3.01 (NovoCraft, http://novocraft.com), VarScan v2.3.6 (Koboldt et al., 2012) and Mapsembler2 v2.1.6 (Peterlongo & Chikhi, 2012) to generate a per individual reference for SNP calling without the need for *de novo* assembly (Figure 1b). Perl scripts are available in a github repository (https://github.com/chodon/zingiberales). The plastid sequences were processed the same way except extension with Mapsembler2 was omitted, and individual genes were extracted from the whole plastid prior to final mapping. Finally, reads were mapped with NovoAlign with –t 90 and PCR duplicates were removed with Picard v1.103 (http://picard.sourceforge.net). SNPs were called following best practices guidelines using the GATK readBackedPhasing algorithm v3.1.1 (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013), except quality scores were not recalibrated because the lack of a reference set of known variants. Consensus sequences were created based on SNP calls for regions with greater than 20× coverage (Nielsen et al., 2011). SNPs in areas with less than 20× coverage were converted to Ns and regions with less than 5× coverage were discarded. For outgroup taxa, raw reads from transcriptomes generated as part of OneKP were subject to the same pipeline as sequences generated *de novo*. The raw sequence data from the *Musa balbisiana* genome project (Davey et al., 2013) was also subject to the pipeline, but only aligned for the plastid gene set. Raw *de novo* sequence reads and the final concatenated alignment are accessible from Dryad xxx.

Alignment

After consensus sequences were made, a second pipeline was made to pass sequences through a series of alignment steps to (1) trim sequences to the *Musa* reference (MAFFT v7.164 [Katoh et al., 2002; Katoh, 2013] and mothur v1.34.4 [Schloss et al., 2009]), (2) place sequences into coding frame (MACSE v1.01b [Ranwez et al., 2011]), and (3) align by codon position (prank v140603 (Löytynoja & Goldman, 2005)). Plastid gene introns were spliced out by hand in Geneious v5.6.4 (Kearse et al., 2012) prior to step 3, above. After alignment, several additional steps were taken to eliminate genes that might contain non-orthologous sequences. Gene trees were generated with RAxML v8.1.17 (Stamatakis, 2014) and the single gene trees were assessed to identify those in which the gene of a single individual taxon accounted for greater than 15% of the total tree length (dos Reis et al., 2012). Exon sequences from one individual were BLASTed to the nucleotide collection database (BLASTN v2.2.30+, Altschul et al., 1997). Exons were removed from further analyses if significant BLAST hits were found to a whole plastid genome, or to ribosomal, transposon, or mitochondrial DNA. Exons were also removed from further analysis if they had unexpectedly high average coverage of greater than 200× or because frameshifts were introduced during codon position assignment or the alignment had too many indels to be reliable (Table S2). We also manually checked all alignments for potential problems (Rothfels et al., 2015). Command parameters for all steps can be found in Supplementary Methods.

Phylogenetic analyses

The nuclear and plastid sequence data were concatenated and analyzed using maximum parsimony (MP) and maximum likelihood (ML) approaches. For MP, PAUP* v4.0a142 (Swofford, 2002) was used to perform a heuristic search with 100 random addition sequence replicates and default parameters (TBR branch swapping with one tree held per replicate). MP support was evaluated with 1000 bootstrap replicates, each with 10 random addition sequence replicates. For ML reconstruction, gene-by-codon position partitions were created for the complete concatenated data set resulting in a total of 1128 initial partition subsets. These initial subsets were then grouped using the relaxed hierarchical clustering algorithm with a 1% search strategy (Lanfear et al., 2014) implemented in PartitionFinder v1.1.1 (Lanfear et al., 2012). The resulting partitioning scheme generated by PartitionFinder consisted of 112 subsets (see Supplementary Methods). The PartitionFinder scheme was analyzed with RAxML v8.1.24 (Stamatakis, 2014) with the GTR+Γ_4 model of sequence evolution estimated for each partition subset and the topology linked across partitions. ML support was evaluated for the same partitioning scheme with 1000 bootstrap replicates, using the rapid bootstrap algorithm (Stamatakis, Hoover & Rougemont, 2008), and using the CAT$_{25}$ approximation instead of $\Gamma_{4,}$ to model site-to-site rate heterogeneity (Stamatakis, 2006). The RAxML analysis was performed on the CIPRES web server (Miller, Pfeiffer & Schwartz, 2010). The data were not subject to coalescent methods for this initial analysis as these methods and those of statistical binning have not been shown to be more accurate than concatenation when relatively short coding sequences are being analyzed (Mirarab et al., 2014).

Results

Probe Design, Sequence Capture, and Alignment

All targeted regions for all individuals were successfully captured, although average coverage varied based on gene region (Figure 2a), individual, and phylogenetic distance to the reference sequence (Figure 2b). Members of the Musaceae, in general, captured better than any other family, likely because they are phylogenetically closest to the original genomic reference upon which the probes were designed. Within each family, close relatives of the species or taxon used to design the bait had higher success rates of capture than more distant members of the family. For example, *Siphonochilus kirkii*, had the lowest average coverage and capture efficiency for Zingiberaceae (Figure 2b, c) as predicted by its evolutionary distance from the transcriptome-sequenced taxon *Curcuma longa*. Of the total sequenced bases, the capture efficiency varied across individuals with the maximum percentage of bases mapping 3.5× higher than the minimum percentage (Figure 2c). An average of 26% of captured bases mapped to target, which is similar to capture efficiency reported in captures of human mitochondrial DNA (Maricic, Whitten & Pääbo, 2010) and transcriptome based capture of chipmunk DNA (Bi et al., 2013). Despite the attempt to capture nuclear and plastid targets evenly, sequencing was highly biased towards plastid targets (Figure 2c). There was some variability between individuals that was independent of phylogenetic distance, likely due to the standard variation in the success of DNA library preparation, which results from differences in DNA quality, genome size, and difficulties of accurately quantifying DNA for pooling in equimolar quantities. Any differences in DNA concentration were likely amplified in the post-hybridization PCR enrichment step.

Of the 494 nuclear probe exons, 124 were removed from further analyses based on coverage, BLAST results, skewed tree length, or alignment anomalies (Table S2). These 124 exons were from 110 genes. Twenty exons from 14 genes had greater than 200× average coverage suggesting that these regions are part of highly repetitive areas. It is probable that these regions were either incorrectly annotated as nuclear regions in the *Musa* draft genome, or were transferred to the nuclear genome from more high copy genomes, especially considering that 15 of these exons were annotated as having an "unknown chromosomal location" in the *Musa* draft genome (Figure 2a). A total of 37 exons from 34 genes were removed from the nuclear dataset and 13 genes from the plastid dataset due to skewed tree length. Four nuclear exons from two genes were removed because of introduced frameshifts and *ycf*1 from the plastid was eliminated because of insertions and deletions in the alignment apparent after manual inspection. Finally, 63 additional exons from 61 genes were removed because of a top BLAST hit to a whole plastid genome, mitochondrial, transposon or ribosomal DNA. Of these 63 exons, the 27 ribosomal and 21 mitochondrial exons could likely be included in further analyses or within family specific analyses in future work after analyzing secondary structure and genomic location.

The final dataset of 308 nuclear genes had a total aligned length of 81,546 bp with 24,379 (29.9%) parsimony informative sites. The 68 gene plastid dataset had a total aligned length of 56,202 bp with 8,336 (14.8%) parsimony informative sites (Table S2).

Phylogenetic Analyses

The recovered topology (Figure 3) places Musaceae as sister to all other families with 100% parsimony bootstrap support (pb) and maximum likelihood bootstrap support (mlb). The ginger families (Cannaceae, Costaceae, Zingiberaceae and Marantaceae) are well supported (100 pb/mlb) as monophyletic. The MP and ML trees are largely congruent and support values are generally high from shallow to deep phylogenetic relationships (Figure 3).

Discussion

This method functions to capture numerous loci across 100 Myr of divergence, with successful capture across individual species that are divergent from the genomic data for which the baits were generated. Using several different taxa as bait and filtering genes for those found in all families ameliorated the problem of decreased capture efficiency as phylogenetic distance from probes increases. Furthermore, this protocol can be customized to any plant group and can often be generated with publically available data generated from previous studies. Despite deep phylogenetic divergence, the array-based capture was effective, enabling the avoidance of high efficiency, but costly, in-solution capture protocols. Future work will focus on limiting mistaken high copy and excessive plastid capture as well as minimizing the introduction of PCR duplicates.

Family relationships within Zingiberales have been studied since the mid-1950s (Tomlinson, 1956, 1962). Based on morphological, anatomical, and developmental data a monophyletic ‘ginger’ clade (Zingiberaceae, Costaceae, Cannaceae and Marantaceae) has long been established (Dahlgren & Rasmussen, 1983; Kirchoff, 1988). However, there are no reliable estimates for the relationships among the other four families (i.e., the ‘banana’ lineages: Musaceae, Heliconiaceae, Lowiaceae, and Strelitziaceae) and the ginger clade despite several phylogenetic studies from combined genomic compartments and morphological data (Kress, 1990; Kress et al., 2001; Johansen, 2005). Even studies using plastome scale datasets failed to produce a well resolved phylogeny near the root of the Zingiberales (Barrett et al., 2014). Here, we show that a targeted exon capture generates phylogenomic scale data that can fruitfully address this problem and may be adapted for resolving ancient radiation in other plant groups. Our main finding suggests that Musaceae is the sister group to the remaining families of Zingiberales and that many other deep relationships within Zingiberales are well supported (Figure 3). Recent studies of gene family evolution and gene duplication (Bartlett & Specht, 2010; Yockteng et al., 2013; Almeida, Yockteng & Specht, 2015) further support this placement of Musaceae. Relationships within individual Zingiberales families are also well supported (Figure 3). Importantly, these are not in conflict with existing well supported hypotheses for generic-level relationships (Kress, Prince & Williams, 2002; Johansen, 2005; Prince & Kress, 2006; Specht, 2006; Kress et al., 2007; Prince, 2010; Li et al., 2010; Cron et al., 2012), indicating that our method is identifying orthologs and that the data produced should be useful at finer phylogenetic scales as well a deep ones.

This pilot study is a first attempt at harnessing phylogenomic data from both the nuclear and plastid genomes to address the global phylogeny of Zingiberales. We have planned substantially increased taxon sampling for both ingroups and out groups and work is ongoing to incorporate morphological data from living and fossil representatives into a phylogenetic reconstruction pipeline to co-estimate fossil placement and lineage divergence times. This will permit us to make full use of information recorded in both the fossil record and genetic data to understand morphological evolution of floral and vegetative traits across the Zingiberales, and estimate ages of diversification for the major lineages, testing the hypothesis of an ancient and rapid radiation at the base of the order.

Acknowledgments

We thank Lydia Smith at the Evolutionary Genetics Lab and Ke Bi at the Computational Genomics Resource Lab at UC Berkeley for invaluable help and discussion with lab and bioinformatics work; Igor Antoshechkin at CalTech for help with the QSonica; the Huntington and UC Botanical Gardens for collections and hosting CS; Jerry Davis for collections and DNA; OneKP and MonAToL, especially Jim Leebens-Mack and Dennis Wm. Stevenson, for early access to raw transcriptome reads used for baits.

References

Almeida AMR, Yockteng R, Specht CD. 2015. Evolution of petaloidy in the Zingiberales: An assessment of the relationship between ultrastructure and gene expression patterns. *Developmental Dynamics*.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Angiosperm Phylogeny Group. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141:399–436.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43:11.10.10–11.10.33.

Barrett CF, Specht CD, Leebens-Mack J, Stevenson DW, Zomlefer WB, Davis JI. 2014. Resolving ancient radiations: Can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Annals of Botany* 113:119–133.

Bartlett ME, Specht CD. 2010. Evidence for the involvement of *GLOBOSA*-like gene duplications and expression divergence in the evolution of floral morphology in the Zingiberales. *New Phytologist* 187:521–41.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: Next-generation museum population genomics. *Molecular Ecology* 22:6018–6032.

Chen ST, Smith SY. 2013. Phytolith variability in Zingiberales: A tool for the reconstruction of past tropical vegetation. *Palaeogeography, Palaeoclimatology, Palaeoecology* 370:1–12.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters* 8:783–786.

Cron G V, Pirone C, Bartlett M, Kress WJ, Specht C. 2012. Phylogenetic relationships and evolution in the Strelitziaceae (Zingiberales). *Systematic Botany* 37:606–619.

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema

G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci A-M, Weissenbach J, Ruiz M, Glaszmann J-C, Quétier F, Yahiaoui N, Wincker P. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488:213–217.

Dahlgren RMT, Rasmussen FN. 1983. Monocotyledon evolution: Characters and phylogenetic estimation. *Evolutionary Biology* 16:255–395.

Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J. 2013. A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. *BMC Genomics* 14:683.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43:491–498.

Edwards K, Johnstone C, Thompson C. 1991. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Research* 19:1349.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61:717–726.

Friis EM. 1988. *Spirematospermum chandlerae* sp. nov., an extinct species of Zingiberaceae from the North American Cretaceous. *Tertiary Research* 9:7–12.

Friis EM, Crane PR, Pedersen KR. 2011. *Early flowers and angiosperm evolution*. Cambridge: Cambridge University Press.

Hodges E, Rooks M, Xuan ZY, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols* 4:960–974.

Iles WJD, Smith SY, Gandolfo MA, Graham SW. 2015. A review of monocot fossils suitable for molecular dating analyses. *Botanical Journal of the Linnean Society*.

Johansen LB. 2005. Phylogeny of *Orchidantha* (Lowiaceae) and the Zingiberales based on six DNA regions. *Systematic Botany* 30:106–117.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–3066.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649.

Kirchoff BK. 1988. Floral ontogeny and evolution in the ginger group of the Zingiberales. In: Leins P, Tucker SC, Endress PK eds. *Aspects of floral development*. Berlin, 45–56.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22:568–576.

Konieczny A, Ausubel FM. 1993. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal* 4:403–410.

Kress WJ. 1990. The phylogeny and classification of the Zingiberales. *Annals of Missouri Botanical Garden* 77:698–721.

Kress WJ, Prince LM, Hahn WJ, Zimmer EA. 2001. Unraveling the evolutionary radiation of the families of the Zingiberales using morphological and molecular evidence. *Systematic Biology* 50:926–944.

Kress WJ, Newman MF, Poulsen AD, Specht C. 2007. An analysis of generic circumscriptions in tribe Alpinieae (Alpiniodeae: Zingiberaceae). *Gardens' Bulletin Singapore* 59:113–128.

Kress WJ, Prince LM, Williams KJ. 2002. The phylogeny and a new classification of the gingers (Zingiberaceae): Evidence from molecular data. *American Journal of Botany* 89:1682–1696.

Kress WJ, Specht CD. 2006. The evolutionary and biogeographic origin and diversification of the tropical monocot order Zingiberales. In: Columbus JT, Friar EA, Hamilton CW, Porter JM, Prince LM, Simpson MG eds. *Monocots: Comparative biology and evolution*. Claremont, CA: Rancho Santa Ana Botanic Garden, 619–630.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29:1695–1701.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology* 14:82.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61:727–744.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li L-F, Häkkinen M, Yuan Y-M, Hao G, Ge X-J. 2010. Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus *Musa*. *Molecular Phylogenetics and Evolution* 57:1–10.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences, USA* 102:10557–10562.

Maricic T, Whitten M, Pääbo S. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* 5:e14004.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–1303.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 2010:doi:10.1101/pdb.prot5448.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *2010 Gateway Computing Environments Workshop (GCE)*. IEEE, 1–8.

Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346:1250463.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12:443–451.

Peñalba J V., Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RCK, Moritz C. 2014. Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for non-model organisms. *Molecular Ecology Resources* 14:1000–1010.

Peterlongo P, Chikhi R. 2012. Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. *BMC Bioinformatics* 13:48.

Prince LM. 2010. Phylogenetic relationships and species delimitation in *Canna* (Cannaceae). In: Seberg O, Petersen G, Barford, Davis JI eds. *Diversity, phylogeny and evolution in the monocotyledons*. Aarhus: Aarhus University Press, Denmark, 307–331.

Prince LM, Kress WJ. 2006. Phylogenetic relationships and classification in Marantaceae: insights from plastid DNA sequence data. *Taxon* 55:281–296.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* 6:e22594.

Dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* 279:3491–3500.

Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu C-R, Korkin D. 2012. Long identical multispecies elements in plant and animal genomes. *Proceedings of the National Academy of Sciences, USA* 109:E1183–E1191.

Rodriguez-de la Rosa RA, Cevallos-Ferriz SRS. 1994. Upper Cretaceous zingiberalean fruits with *in situ* seeds from southeastern Coahuila, Mexico. *International Journal of Plant Sciences* 155:786–805.

Rothfels CJ, Li F-W, Sigel EM, Huiet L, Larsson A, Burge DO, Ruhsam M, Deyholos M, Soltis DE, Stewart CN, Shaw SW, Pokorny L, Chen T, DePamphilis C, DeGironimo L, Chen L, Wei X, Sun X, Korall P, Stevenson DW, Graham SW, Wong GK-S, Pryer KM. 2015. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *American Journal of Botany* 102:1089–1107.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75:7537–7541.

Singhal S, Moritz C. 2012. Strong selection against hybrids maintains a narrow contact zone between morphologically cryptic lineages in a rainforest lizard. *Evolution* 66:1474–1489.

Specht CD. 2006. Systematics and evolution of the tropical monocot family Costaceae (Zingiberales): A multiple dataset approach. *Systematic Botany* 31:89–106.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology* 57:758–771.

Swofford DL. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), version 4.

Tomlinson PB. 1956. Studies in the systematic anatomy of the Zingiberaceae. *Journal of the Linnean Society (Botany)* 55:547–592.

Tomlinson PB. 1962. Phylogeny of the Scitamineae—Morphological and anatomical considerations. *Evolution* 16:192–213.

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2:1400042.

Whitfield JB, Kjer KM. 2008. Ancient rapid radiations of insects: Challenges for phylogenetic analysis. *Annual Review of Entomology* 53:449–472.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends in Ecology & Evolution* 22:258–265.

Yockteng R, Almeida AMR, Morioka K, Alvarez-Buylla ER, Specht CD. 2013. Molecular evolution and patterns of duplication in the *SEP/AGL6*-like lineage of the Zingiberales: a proposed mechanism for floral diversification. *Molecular Biology and Evolution* 30:2401–2422.

Figure 1(on next page)

Schematic diagrams for the bioinformatic work flow.

(A) Work flow to generate family specific bait sequence from transcriptomes and the annotated exons from *Musa acuminata* and (B) work flow to generate individual sequences for each gene from raw reads independent of *de novo* assembly. Base changes and SNPs are highlighted and the schematic is represented as in the SAMtools tview format (i.e., reverse reads are represented with commas and lowercase letters). The representation is condensed to show examples of how the reads are transformed but the actual coverage used to call SNPs was at least 20× (see methods).

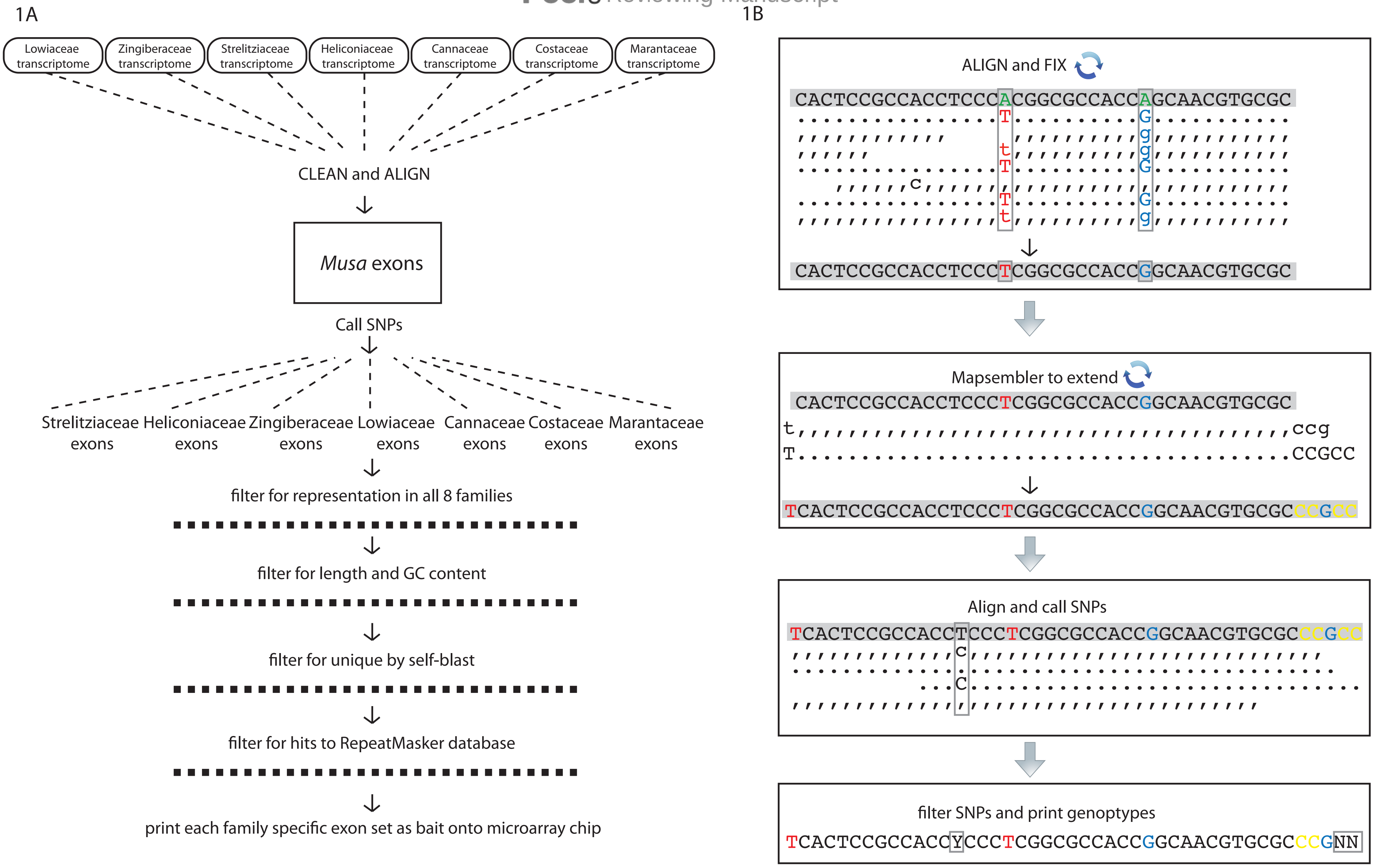

1A
Lowiaceae transcriptome
Zingiberaceae transcriptome
Strelitziaceae transcriptome
Heliconiaceae transcriptome
Cannaceae transcriptome
Costaceae transcriptome
Marantaceae transcriptome
CLEAN and ALIGN
Musa exons
Call SNPs
Strelitziaceae exons
Heliconiaceae exons
Zingiberaceae exons
Lowiaceae exons
Cannaceae exons
Costaceae exons
Marantaceae exons
filter for representation in all 8 families
filter for length and GC content
filter for unique by self-blast
filter for hits to RepeatMasker database
print each family specific exon set as bait onto microarray chip
1B
ALIGN and FIX
CACTCCGCCACCTCCCACGGCGCCACCAGCAACGTGCGC
CACTCCGCCACCTCCCTCGGCGCCACCGGCAACGTGCGC
Mapsembler to extend
CACTCCGCCACCTCCCTCGGCGCCACCGGCAACGTGCGC
t ccg
T CCGCC
TCACTCCGCCACCTCCCTCGGCGCCACCGGCAACGTGCGCCCGCC
Align and call SNPs
TCACTCCGCCACCTCCCTCGGCGCCACCGGCAACGTGCGCCCGCC
c
C
filter SNPs and print genoptypes
TCACTCCGCCACCYCCCTCGGCGCCACCGGCAACGTGCGCCCGNN

Figure 2(on next page)

Capture efficiency across individuals and exons.

(A) Average coverage across all individuals for each of the 494 exons captured. The view is clipped at 200× coverage as this was the value above which exons were removed from further analysis. The exons are ordered by chromosome based on annotation from the *Musa acuminata* genome. (B) Average coverage over all exons for each individual after removing PCR duplicates and with strict alignment parameters that were used for SNP calling. Average coverage was calculated before and after removing the high coverage exons indicated in 2A. (C) Per individual, the percent of the total sequenced base pairs passing Illumina quality filters that mapped to target regions prior to PCR duplicate removal. Percent of base pairs mapping to chloroplast plastid and nuclear regions are indicated in orange and blue respectively. Species are grouped by family (Can=Cannaceae, Mar=Marantaceae, Cos=Costaceae, Zin=Zingiberaceae, Str=Strelitziaceae, Low=Lowiaceae, Hel=Heliconiaceae, Mus=Musaceae) and species upon which baits were generated are indicated with a filled circle (nuclear bait) or open circle (plastid bait).

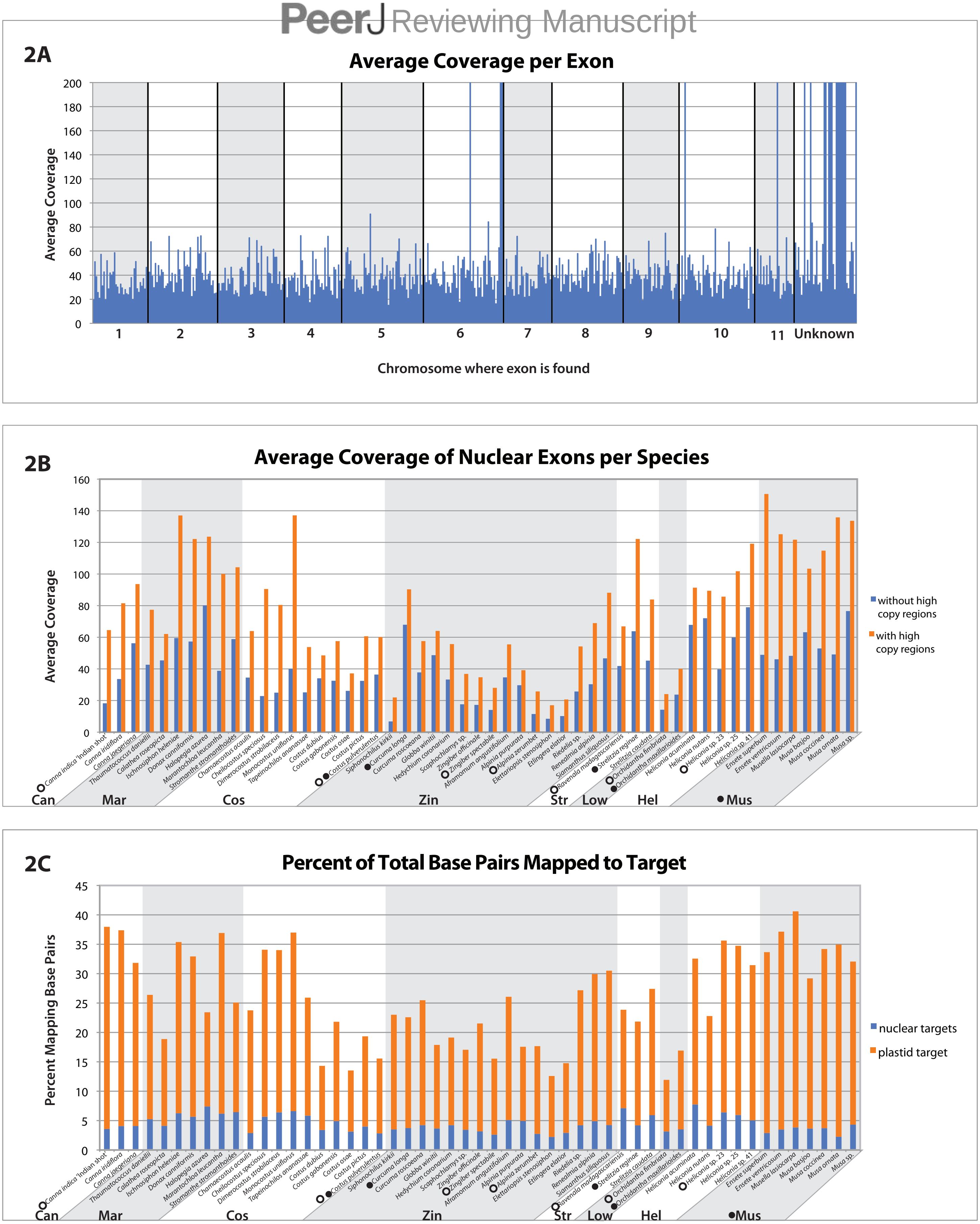
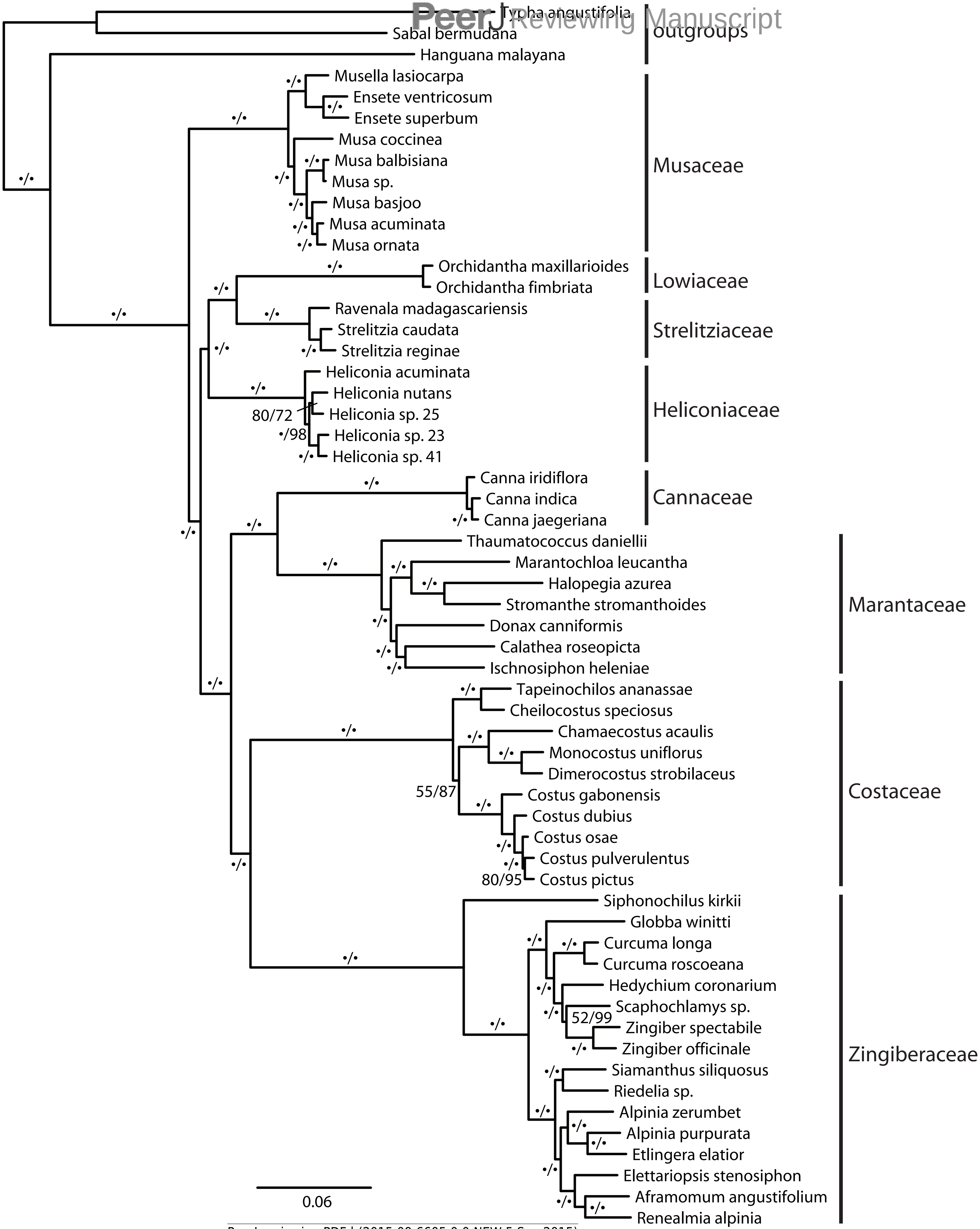

2A
Average Coverage per Exon
Average Coverage
200
180
160
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
Unknown
Chromosome where exon is found
2B
Average Coverage of Nuclear Exons per Species
Average Coverage
160
140
120
100
80
60
40
20
0
without high copy regions
with high copy regions
○ Canna indica 'Indian shot'
Canna iridiflora
Canna jaegeriana
Thaumatococcus daniellii
Calathea roseopicta
Ischnosiphon heleniae
Donax canniformis
Halopegia azurea
Marantochloa leucantha
Stromanthe stromanthoides
Chamaecostus acaulis
Cheilocostus speciosus
Dimerocostus strobilaceus
Monocostus uniflorus
Tapeinochilos ananassae
Costus dubius
Costus gabonensis
Costus osae
Costus pictus
○ ● Costus pulverulentus
Siphonochilus kirkii
● Curcuma longa
Curcuma roscoeana
Globba winitii
Hedychium coronarium
Scaphochlamys sp.
○ Zingiber officinale
Zingiber spectabile
Aframomum angustifolium
○ Alpinia purpurata
Alpinia zerumbet
Elettariopsis stenosiphon
Etlingera elatior
Riedelia sp.
Renealmia alpinia
Siamanthus siliquosus
○ Ravenala madagascariensis
● Strelitzia reginae
Strelitzia caudata
○ Orchidantha fimbriata
● Orchidantha maxillarioides
Heliconia acuminata
Heliconia nutans
○ Heliconia sp. 23
Heliconia sp. 25
Heliconia sp. 41
Ensete superbum
Ensete ventricosum
Musella lasiocarpa
Musa basjoo
Musa coccinea
Musa ornata
Musa sp.
Can
Mar
Cos
Zin
Str
Low
Hel
● Mus
2C
Percent of Total Base Pairs Mapped to Target
Percent Mapping Base Pairs
45
40
35
30
25
20
15
10
5
0
nuclear targets
plastid target
Can
Mar
Cos
Zin
Str
Low
Hel
● Mus

Figure 3(on next page)

Phylogeny of Zingiberales based on a partitioned ML of concatenated plastid and nuclear sequence.

Bootstrap support values adjacent to branches are for MP and ML, respectively. A dot indicates 100% bootstrap support. Scale is in expected substitutions per site.


Typha angustifolia
Sabal bermudana
Hanguana malayana
outgroups
Musella lasiocarpa
Ensete ventricosum
Ensete superbum
Musa coccinea
Musa balbisiana
Musa sp.
Musa basjoo
Musa acuminata
Musa ornata
Musaceae
Orchidantha maxillarioides
Orchidantha fimbriata
Lowiaceae
Ravenala madagascariensis
Strelitzia caudata
Strelitzia reginae
Strelitziaceae
Heliconia acuminata
Heliconia nutans
Heliconia sp. 25
Heliconia sp. 23
Heliconia sp. 41
Heliconiaceae
80/72
•/98
Canna iridiflora
Canna indica
Canna jaegeriana
Cannaceae
Thaumatococcus daniellii
Marantochloa leucantha
Halopegia azurea
Stromanthe stromanthoides
Donax canniformis
Calathea roseopicta
Ischnosiphon heleniae
Marantaceae
Tapeinochilos ananassae
Cheilocostus speciosus
Chamaecostus acaulis
Monocostus uniflorus
Dimerocostus strobilaceus
Costus gabonensis
Costus dubius
Costus osae
Costus pulverulentus
Costus pictus
Costaceae
55/87
80/95
Siphonochilus kirkii
Globba winitti
Curcuma longa
Curcuma roscoeana
Hedychium coronarium
Scaphochlamys sp.
Zingiber spectabile
Zingiber officinale
52/99
Siamanthus siliquosus
Riedelia sp.
Alpinia zerumbet
Alpinia purpurata
Etlingera elatior
Elettariopsis stenosiphon
Aframomum angustifolium
Renealmia alpinia
Zingiberaceae
•/•
0.06