The Impact of Polyploidy on Genome Evolution in Poales and Other Monocots

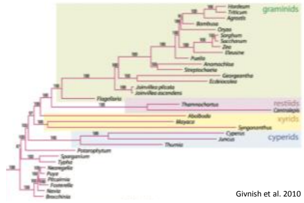
Michael R. McKain
The University of Alabama
@mrmckain

> “I don’t have to emphasize that gene duplications are the fabric of evolution in plants.”
>
> -Jan Dvorak (as quickly written down by a person with poor hearing…me)

@mrmckain

Poales Diversity

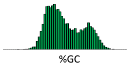
- ~22,800 species
- ~11,088 species in Poaceae

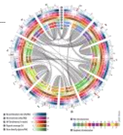


Givnish et al. 2010

Grass genomes: the choose your own adventure of genome evolution

- Transposons (McClintock, Wessler)
- GC content bias (Carels and Bernardi 2000)
- Three WGD events
  - *rho* (Peterson et al. 2004)
  - *sigma* (Tang et al. 2010)
  - *tau* (Tang et al. 2010, Jiao et al. 2014)





Schnable et al., 2009

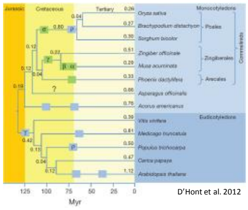
@mrmckain

How has ancient polyploidy altered the genomic landscape in grasses and other Poales?

@mrmckain

Zeroing in on WGD placement

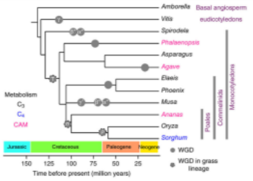
Banana genome



D’Hont et al. 2012

Recovered *sigma* after grass divergence from commelinids

Pineapple genome



Ming, VanBuren et al. 2015

Recovered *sigma* after grass+pineapple divergence from commelinids

@mrmckain

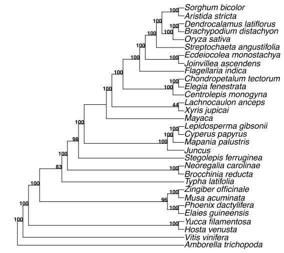
Phylotranscriptomic approach

- Sampling 27 transcriptomes and 7 genomes
- Representation for all families (except Thurniaceae) in Poales
- RNA from young leaf or apical meristem, a combination of Moncot Tree of Life and 1KP
- General steps:
  - Trinity assembly
  - Orthogroup circumscription using a curated 22-genome dataset
  - Aligned amino acids with MUSCLE and created codon alignments with PAL2NAL
  - Gene tree reconstruction using RAxML

Streptochaeta angustifolia
Picture by: Jerry Davis

@mrmckain

Coalescence-based Phylogeny of 234 Single-copy genes



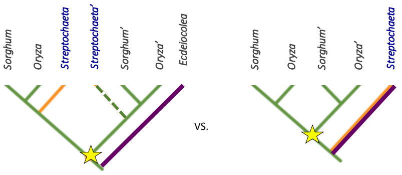
Astral v.4.7.6

- Phylogeny consistent with previous nuclear gene results
- Conflicting topology with chloroplast genome tree:
  - *Ecdeiocolea/Joinvillea* sister instead of a grade
  - *Typha* sister to Poales, not Bromeliaceae

@mrmckain

Placement of WGD Using PUG

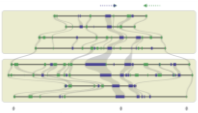
- PUG (Phylogenetic Placement of Polyploidy Using Genomes)
  - https://github.com/mrmckain/PUG
  - Queries a gene tree against a species tree using a focal putative paralog pair



Testing for Poaceae Origin Hypothesis

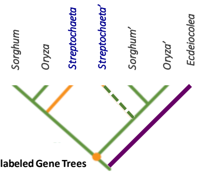
@mrmckain

How we identify putative paralogs matters

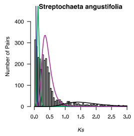


Synteny Analysis

Tang et al. 2010



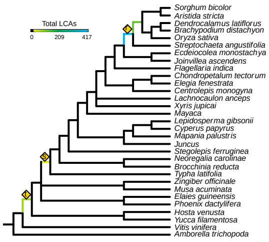
Multilabeled Gene Trees



Ks Frequency Plots

@mrmckain

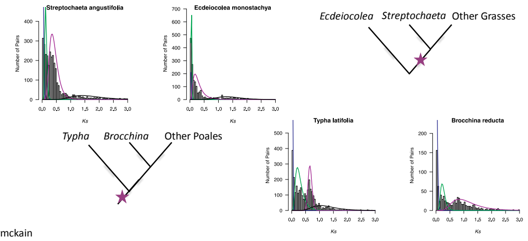
Synteny-derived Paralogs



- Synteny from rice and sorghum genomes
- Identification of
  - 8,267 putative *rho* paralog pairs
  - 3,680 putative *sigma* + *tau* paralog pairs
- Included 2,248 putative *tau* paralog pairs from Jiao et al. 2014
- After filtering for outgroups, the PUG species tree topology, and bootstrap values (80 cutoff)
- 411 *rho* LCA nodes
- 26 *sigma* LCA nodes
- 50 *tau* LCA nodes

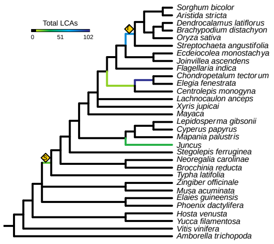
@mrmckain

Ks frequency plot-derived Paralogs



@mrmckain

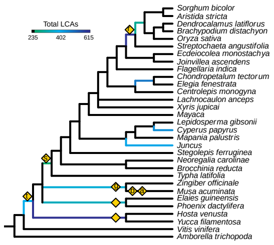
Ks frequency plot-derived Paralogs



- Ks frequency plots of all taxa
- Identification of 20,900 putative paralog pairs
- After filtering: 667 map to 343 unique LCA nodes

@mrmckain

Gene tree-derived Paralogs



- All possible pairs from multilabeled gene trees
- Identification of 1,870,214 putative paralog pairs
  - Most of these are isoforms/alleles
- After filtering: 36,567 map to 5,455 unique LCA nodes

@mrmckain

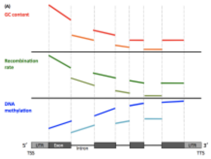
Summary of Gene Tree Polyploidy Support

Event	Synteny	Ks Plots	Gene Trees
rho (grass)	411	78	610
sigma (Poales)	26	22	235
tau (monocot)	88	0	410
Restionaceae	0	102	499
Centrolepidaceae	0	0	200
Restiid	0	15	184
Juncus	0	29	423
Cyperus	0	0	463
Zingiberales	0	0	377
Palms	0	0	345
Agavoideae	0	0	615

@mrmckain

GC Content Evolution in Grasses

- Bimodal distribution of genic GC content
- 5′ → 3′ decreasing GC content gradient
  - Positive correlation with recombination rate
  - Negative correlation with DNA methylation
- GC biased gene conversion
  - Recombination driven
- GC3 bias/codon usage bias
- Gene length (longer more GC)
- Increased expression



Glémin et al. 2014

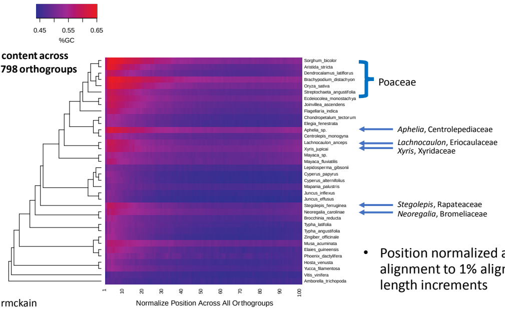
@mrmckain

How has ancient polyploidy altered the genomic landscape in grasses and other Poales?

Is there a connection between polyploidy and the grass bimodal GC distribution?

@mrmckain

GC content across 13,798 orthogroups

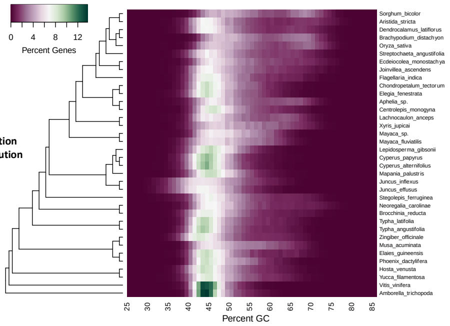


- Position normalized across alignment to 1% alignment length increments

@mrmckain

GC content across 13,798 orthogroups

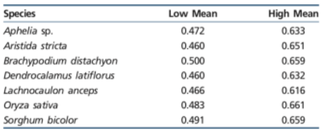
Patterns of GC:
- **Unimodal-tight distribution**
- **Unimodal-broad distribution**
- **Bimodal**



@mrmckain

Identification of GC bimodality

- Reject null hypothesis of Hartigan's dip test for unimodality for seven taxa
- Kmeans clustering used to identify the high and low means for these taxa



Species	Low Mean	High Mean
Aphelia sp.	0.472	0.633
Aristida stricta	0.460	0.651
Brachypodium distachyon	0.500	0.659
Dendrocalamus latiflorus	0.460	0.632
Lachnocaulon anceps	0.466	0.616
Oryza sativa	0.483	0.661
Sorghum bicolor	0.491	0.659

Combined GC Kmeans Results:
Low: 46.7%
High: 63.7%

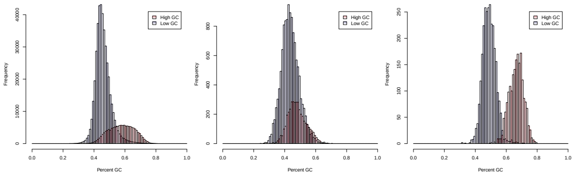
@mrmckain

Classification of Orthogroups by GC Content

- High GC—75% or more of transcripts for a given taxon in an orthogroup are clustered as high GC (63.7%)
  - 3,662 orthogroups
- Low GC—75% or more of transcripts for a given taxon in an orthogroups are clustered as low GC (46.7%)
  - 6,770 orthogroups
- Mixed GC—remaining orthogroups that do not fall into other classes
  - 2,595 orthogroups

@mrmckain

High/Low GC Orthogroup Distributions



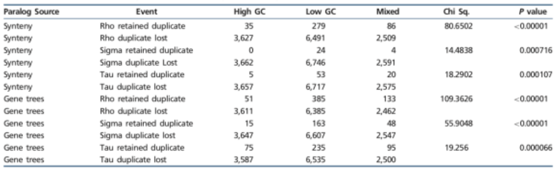
All Taxa

Juncus effusus
44.8% Mean GC

Brachypodium distachyon
56.3% Mean GC

@mrmckain

Comparison of paralog retention across WGD and high/low GC orthogroups



Paralog Source	Event	High GC	Low GC	Mixed	Chi Sq.	*P* value
Synteny	Rho retained duplicate	35	279	86	80.6502	<0.00001
Synteny	Rho duplicate lost	3,627	6,491	2,509		
Synteny	Sigma retained duplicate	0	24	4	14.4838	0.000716
Synteny	Sigma duplicate Lost	3,662	6,746	2,591		
Synteny	Tau retained duplicate	5	53	20	18.2902	0.000107
Synteny	Tau duplicate lost	3,657	6,717	2,575		
Gene trees	Rho retained duplicate	51	385	133	109.3626	<0.00001
Gene trees	Rho duplicate lost	3,611	6,385	2,462		
Gene trees	Sigma retained duplicate	15	163	48	55.9048	<0.00001
Gene trees	Sigma duplicate lost	3,647	6,607	2,547		
Gene trees	Tau retained duplicate	75	235	95	19.256	0.000066
Gene trees	Tau duplicate lost	3,587	6,535	2,500		

High %GC orthogroups more likely to lose duplicated genes

@mrmckain

Summary

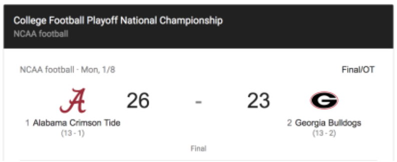
- Polyploid events likely occurred prior to the diversification of Poaceae (*rho*) and Poales (*sigma*)
- Gene tree-derived paralogs more abundant and useful in identifying WGD than synteny or Ks frequency plots
- Bimodal GC distribution result of high and low GC gene families
- Maintenance of high/low GC gene families in Poales regardless of overall GC content
- Paralogs more likely to be lost in high GC gene families

@mrmckain

Co-authors:

- **Haibao Tang**, Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, China
- **Joel McNeal**, Kennesaw State University, Kennesaw, GA
- **Saravanaraj Ayyampalayam**, University of Georgia, Athens, GA
- **Claude W. dePamphilis**, The Pennsylvania State University, University Park, PA
- **Thomas J Givnish**, Department of Botany. UW-Madison, Madison, WI
- **J. Chris Pires**, Division of Biological Sciences, University of Missouri, Columbia, MO
- **Dennis Wm. Stevenson**, New York Botanical Garden, Bronx, NY
- **Jim Leebens-Mack**, University of Georgia, Athens, GA

- Funding
  - NSF: DEB-0830009
  - NSF: DEB-1010905
- Thanks to:
  - Gane Ka-Shu Wong (University of Alberta)



College Football Playoff National Championship

NCAA football

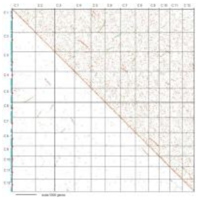
NCAA football · Mon, 1/8 Final/OT

1 Alabama Crimson Tide (13 - 1)	26	23	2 Georgia Bulldogs (13 - 2)

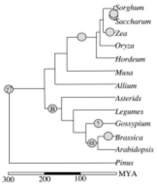
Final

@mrmckain

Identification of *rho*, the grass WGD event



Dot-plot of rice genome syntenic regions

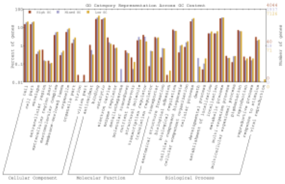


Timing of grass WGD event

rho event described as predating Poaceae divergence

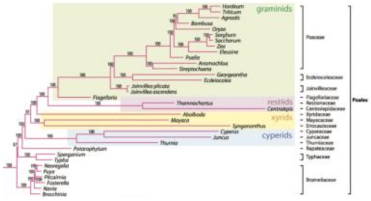
@mrmckain

Paterson et al. 2004



Fine-Scale Placement Requires Strategic Taxonomic Sampling

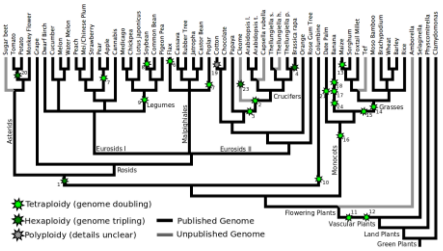
- Sampling each lineage off of the backbone
- *Streptochaeta*
  - First branch in grasses
- *Typha* (~~cattail~~ corn dog grass)
  - Contender with pineapple for first branch in Poales



Givnish et al. 2010

@mrmckain

Polyploidy is prevalent in flowering plants



CoGepedia

*<u>**Michael McKain**</u>[1,2], Haibao Tang[3], Joel McNeal[4], Saravanaraj Ayyampalayam[5], Claude W. dePamphilis[6], Thomas J Givnish[7], J. Chris Pires[8], Dennis Wm. Stevenson[9] and Jim Leebens-Mack[5]*