



Gramene Subsites: Pangenome Browsers for Crops



Marcela K. Tello-Ruiz¹, Sharon Wei¹, Joshua Stein¹, Kapeel Chougule¹, Andrew Olson¹, Yinping Jiao¹, Bo Wang¹, Ivar Meijs¹, Doreen Ware^{1,2}

¹Cold Spring Harbor Laboratory, Ware Laboratory, Cold Spring Harbor, NY, ²USDA-ARS-NEA, Robert W. Holley Center for Agriculture and Health, Ithaca, NY

Abstract: Continued advances in sequencing/assembly technologies are generating an abundance of high quality reference assemblies within crop species, ushering a transition from single-genome to pan-genome research approaches. With this transition, communities will need ready access to pre-computed comparisons of genome assemblies to identify and characterize common and variable regions. To accommodate this need, the Gramene comparative genomics project is developing Gramene subsites, each dedicated to the study of individual crop groups. We will describe current status on four pangenome subsites that support **rice**, **maize**, **sorghum** and **grapevine**. A key feature of pan-genome subsites is the application of uniform annotation protocols to minimize methodological artifacts, and the application of Ensembl and Gramene infrastructures for comparative analysis and visualization. Extending Compara gene tree output, we define conserved syntelog sets and assign conservation scores based on the proportion of genomes with membership in each set. We then score individual genomes for presence/absence and copy-number variation, additionally supported by whole genome alignments. Results in the *Oryza* genus showed that, compared to ancient families, recently emerged genes have higher rates of evolution, higher lability, more limited expression, prevalence in pericentromeric regions, reduced coding-length, and enrichment for stress-response functions.

Summary

Subsites hold collections of closely related reference genomes

- Within species, genus, or crop group
- Sourced by collaborators and funded projects
- 4 subsites in progress for rice, maize, sorghum, & grapevine



Uniform gene annotation protocol

- Species-customized repeat library & evidence sets
- RNA-seq assemblies, PacBio Iso-seq, EST, prior annotation
- Evidence-based + ab initio prediction



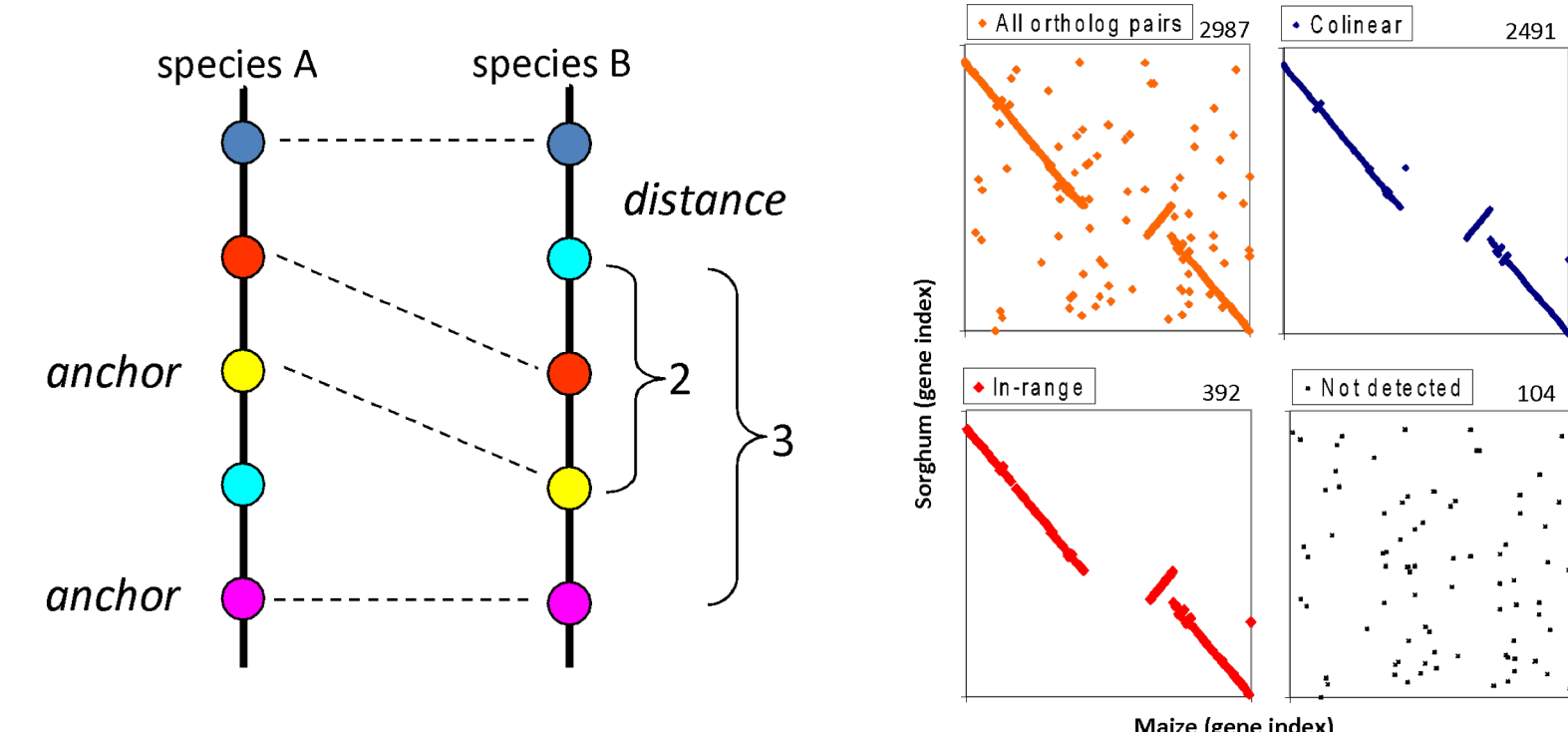
Ensembl databases, web-browsers, & pipelines
Gramene search

Compara Gene Trees & whole genome alignment

- Gene family assignment
- Phylogenetic tree build
- Ortholog & paralog calling
- Taxonomic dating
- Pairwise WGA (BLASTZ-CHAIN-NET)

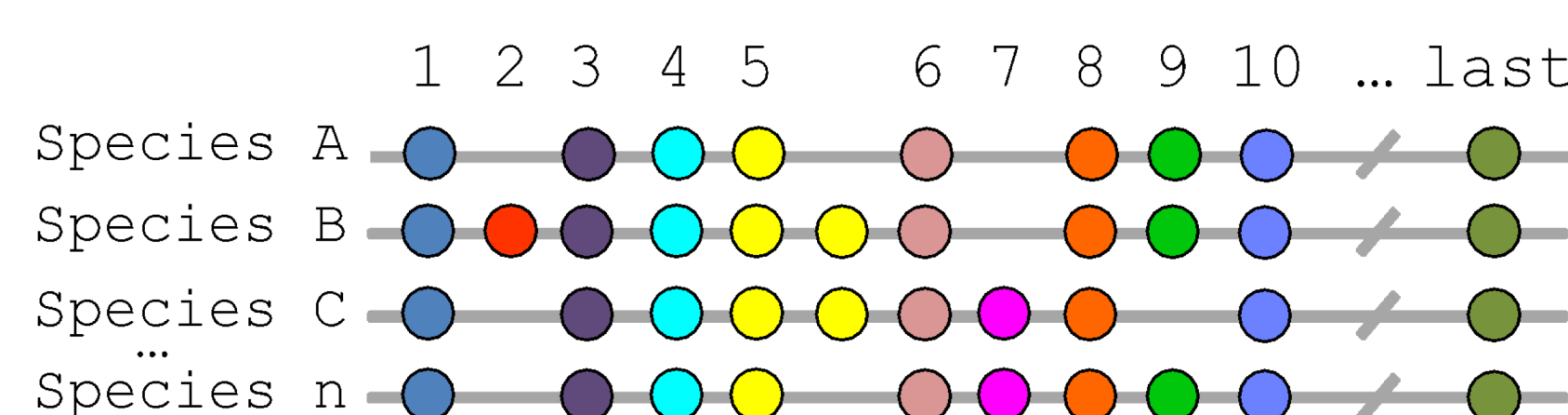
Gene-centered pairwise synteny maps

- Maps collinear & near-collinear orthologs



Pangenome index

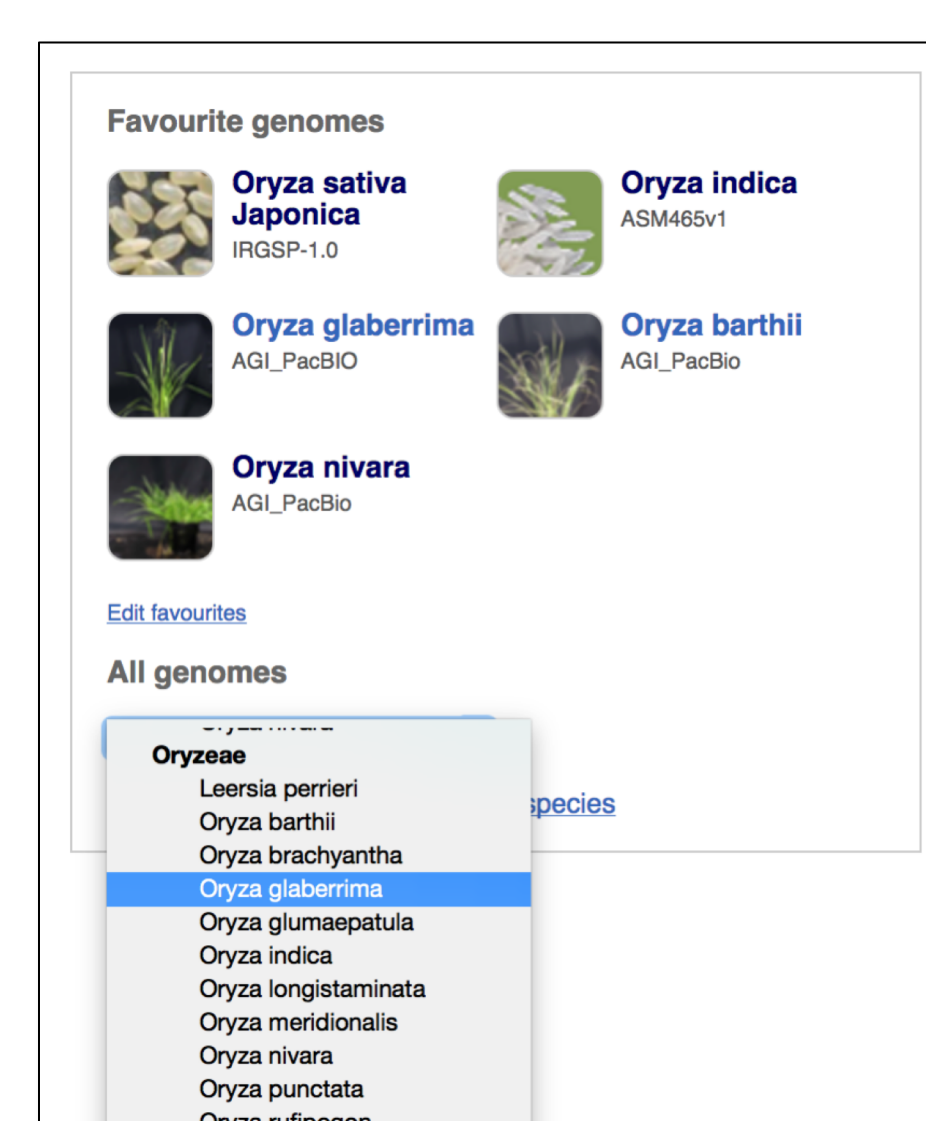
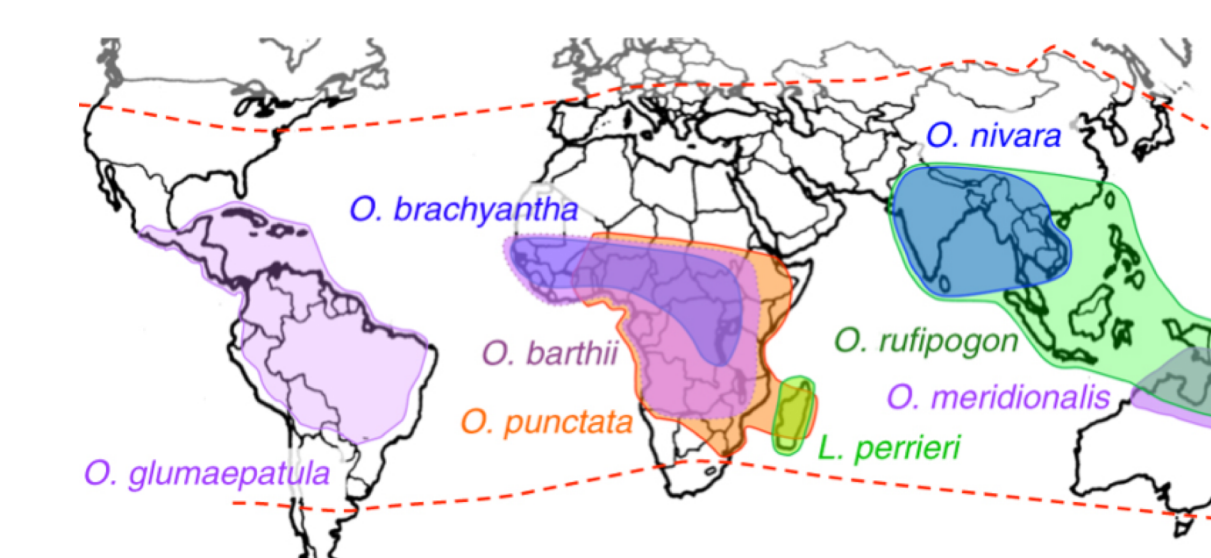
- Cluster syntelogs by transitive closure
- Presence absence variation (PAV)
- Copy number variation (CNV)
- Core & dispensable genome



Gramene Rice: Oryza Genome Evolution (OGE.gramene.org)

OGE

15 domesticated & wild rice genomes



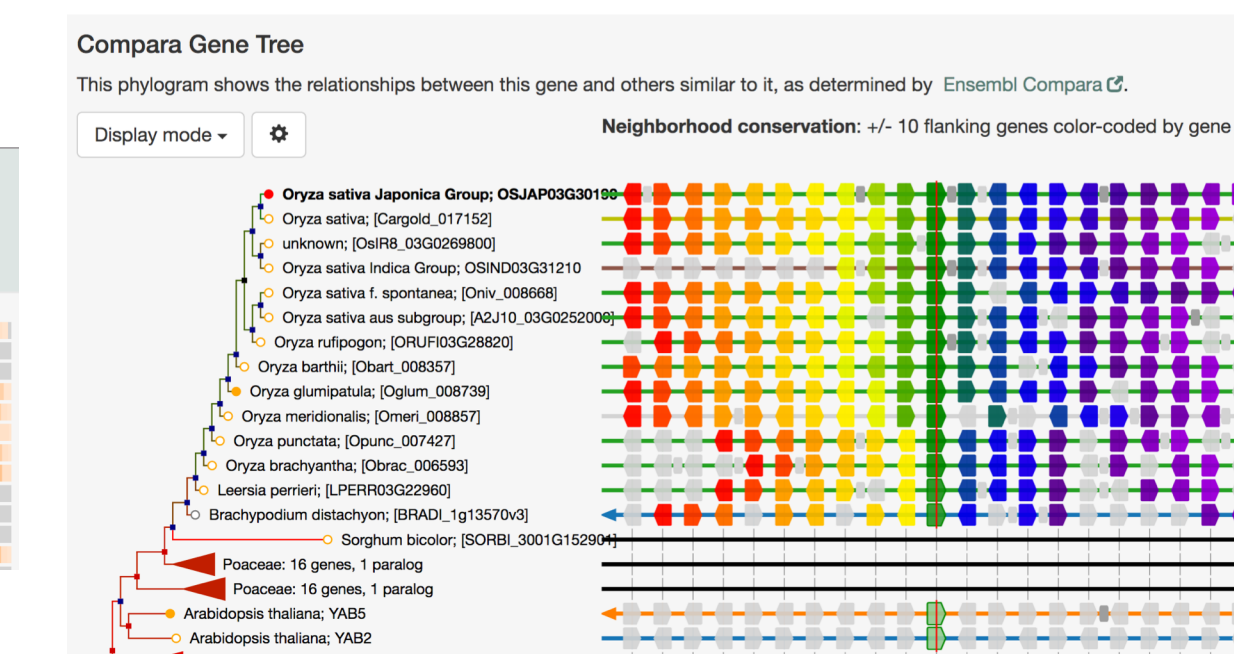
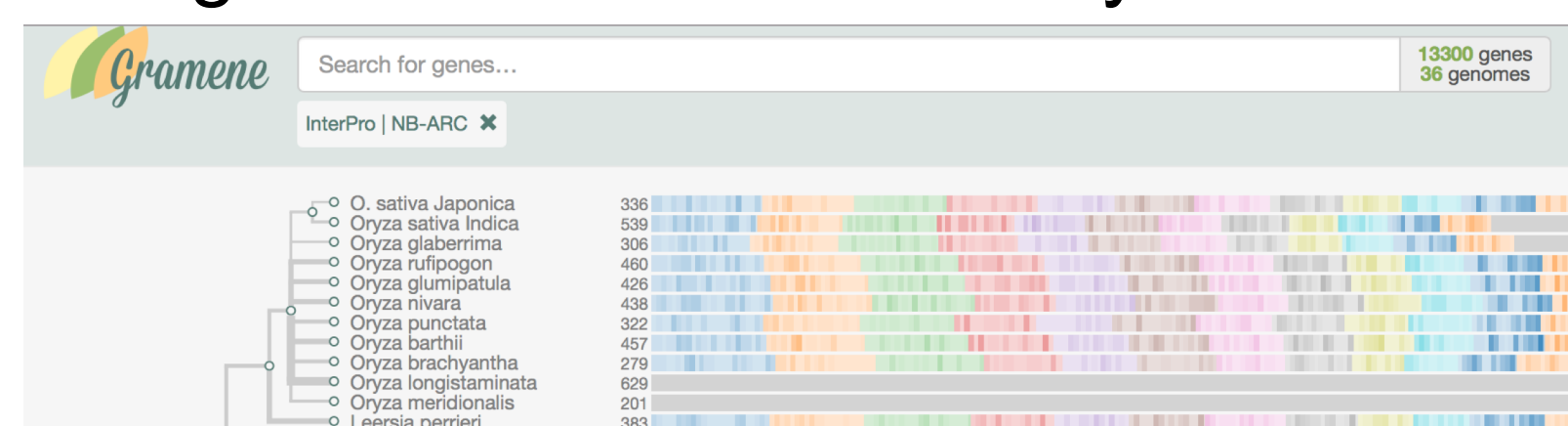
O. sativa strain or species	Release 3
IR8, indica rice	PacBio*
N22, aus rice	PacBio*
O. nivara	PacBio*
O. barthii	PacBio*
O. glaberrima (African rice)	PacBio*
O. glumaepatula	PacBio*
O. meridionalis	PacBio*
O. punctata	PacBio*
O. brachyantha	PacBio*
O. rufipogon	illumina*
L. perrieri	illumina*
O. longistaminata	illumina*
Carolina Gold rice	PacBio†
Nipponbare, japonica rice	BAC-clone Sanger‡
93-11, indica rice	Shotgun capillary‡

- 30,015 gene trees (families) comprising 577,324 genes
- 13 pairwise whole genome alignments
- 120 pairwise synteny maps

Collaborators:
*The International Oryza Map Alignment Project Consortium (I-OMAP): PI:Rod Wing (King Abdullah University of Science & Technology, Saudi Arabia)
†The Carolina Gold Rice Foundation (CGRF) sponsored project: Anna McClung (USDA-ARS, Dale Bumpers National Rice Research Center, Stuttgart, AR)
‡Brian Scheffler (USDA-ARS, Genomics and Bioinformatics Research, Stoneville, MS)

Gramene Search & Enhanced Tree Views

Pangenomic search summary

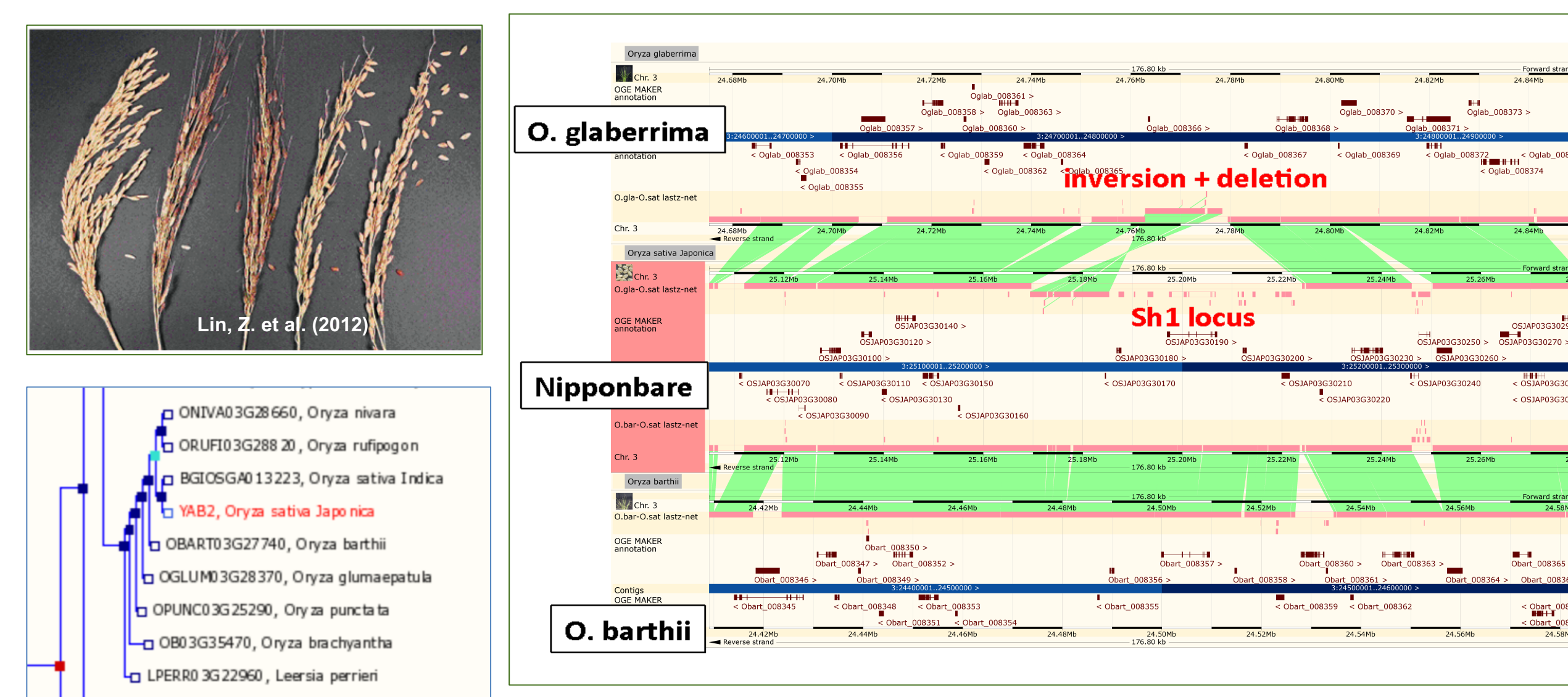


- Alternate gene-tree views
- Gene-neighborhood
- Multiple sequence alignment
- Protein-domain highlighting

Use case: Origin of Domestication genes

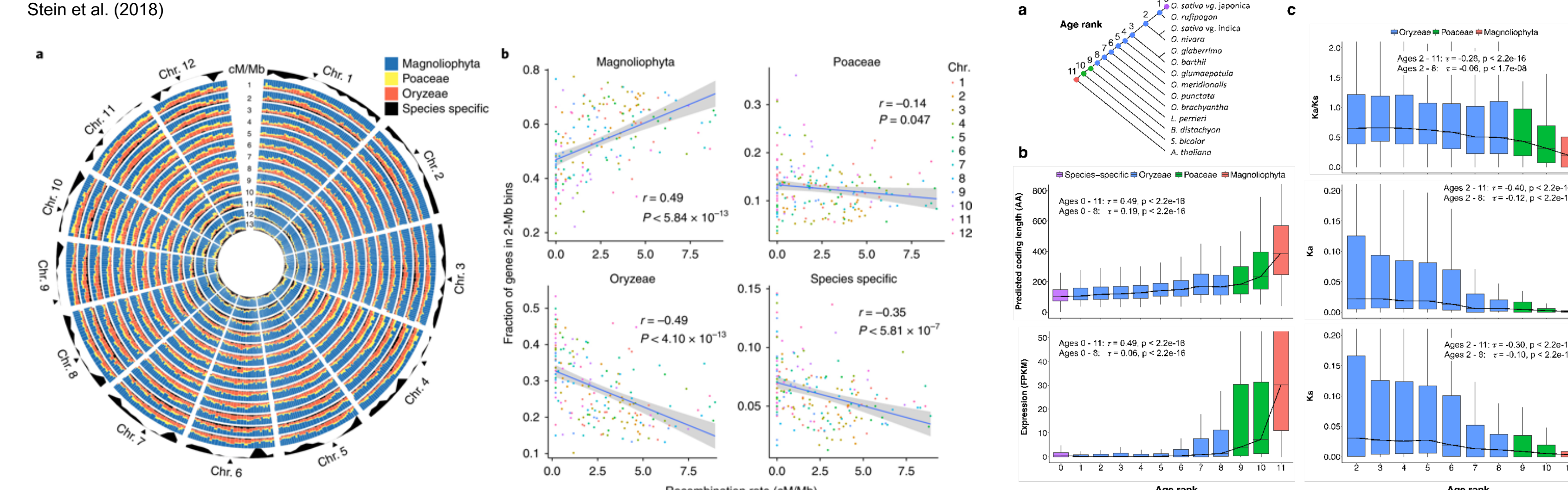
Non-shattering seed is a domestication trait enabling efficient grain harvest. Sh1 locus (a YABBY transcription factor) controls shattering (seed dispersal) and is disrupted in Asian rice & sorghum. African rice (*O. glaberrima*) was domesticated independently of Asian rice from the *O. barthii* progenitor.

Does African rice have a functional Sh1 locus?



Gene tree and whole genome alignment confirms presence of Sh1 in *O. barthii* progenitor, but absence in African rice, as previously observed (Wang et al. 2014).

Characteristics of loci derived from ancient and recent families



Subsites for 2019 Release



Home of NSF Sequencing Maize NAM Founders Project (maize-pangenome.gramene.org)

- PacBio/Bionano assembly of diverse maize inbreds
- Kelly Dawe (U. Georgia), Matt Hufford (Iowa State U.)
- Corteva, DNAnexus



USDA-ARS portal for Sorghum genomics/breeding resources (sorghumbase.org)

- Multiple PacBio & 10X genomes
- Chad Hayes (USDA-ARS, Lubbock TX)
- Corteva



Wild & cultivated Grapevine (vitis.gramene.org)

- Multiple PacBio & 10X genomes
- USDA-ARS VitisGen2 Project
- Lance Cadle-Davidson (USDA-ARS, Geneva, NY)
- Dario Cantu (UC Davis), Rachel Naegele (USDA-ARS, Parlier, CA)



We gratefully acknowledge support from grants NSF#1744001, NSF#1127112, and USDA-ARS #58-8062-7-008.

Citations:
Lin Z. et al. (2012). Nature Genetics. 44:720–724 (2012). 10.1038/ng.2281.
Stein J.C. et al. (2018). Nature Genetics. 50:285. doi.org/10.1038/s41588-018-0040-0.
Wang M. et al. (2014). Nature Genetics. 46:982. doi.org/10.1038/ng.3044.