Rice pan-genome resources in Gramene
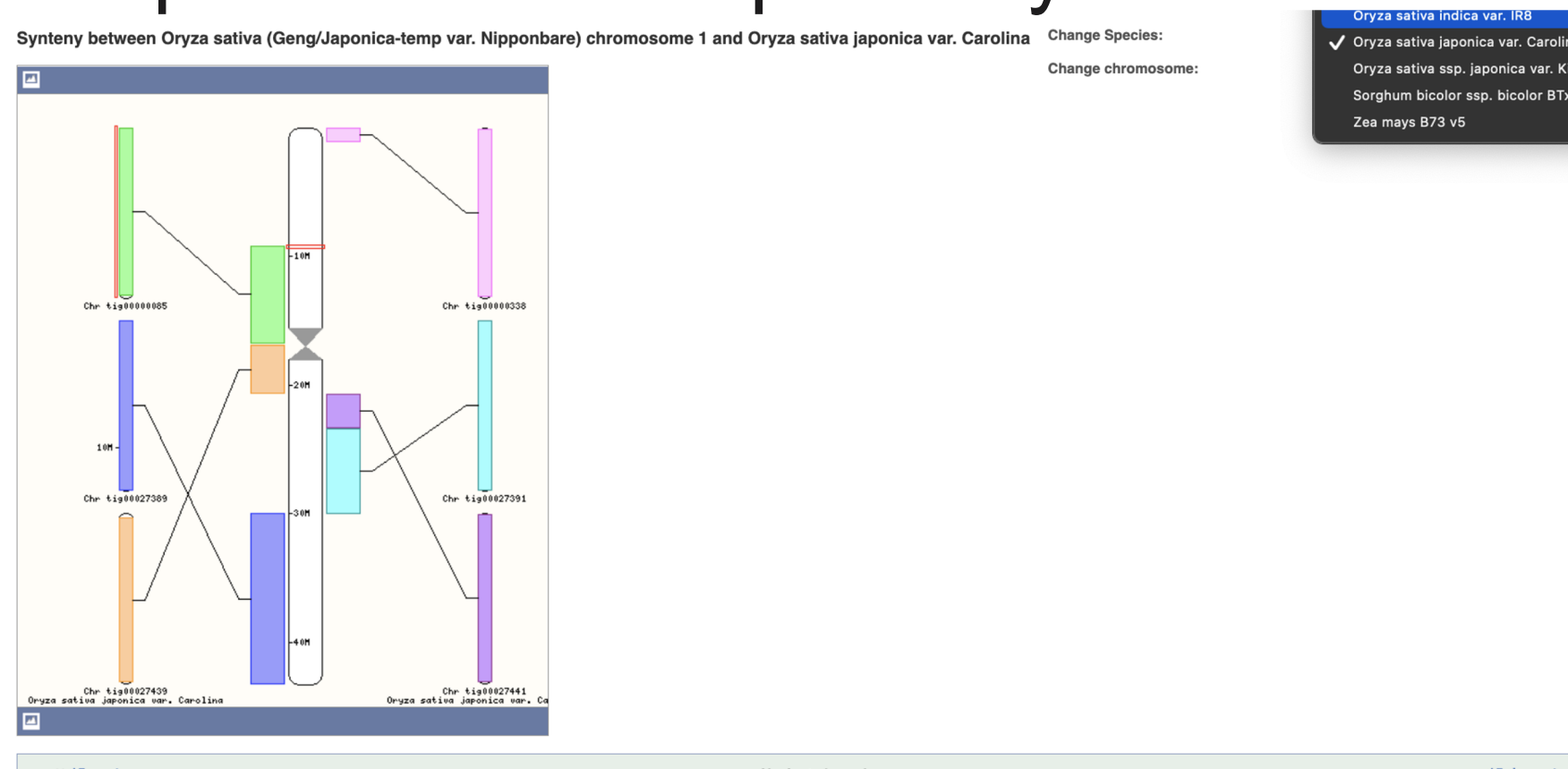
Andrew Olson[1], Kapeel Chougule[1], Vivek Kumar[1], Sunita Kumari[1], Zhenyuan Lou[1], Marcela K. Tello-Ruiz[1], Sharon Wei[1], Zhou Yong[2], Wing Rod[2,3,4], Doreen Ware[1,5]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA; [2]CEDE-BESE, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia; [3]IRRI, Los Baños, Philippines; [4]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ, USA; [5]USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Ithaca, NY, USA
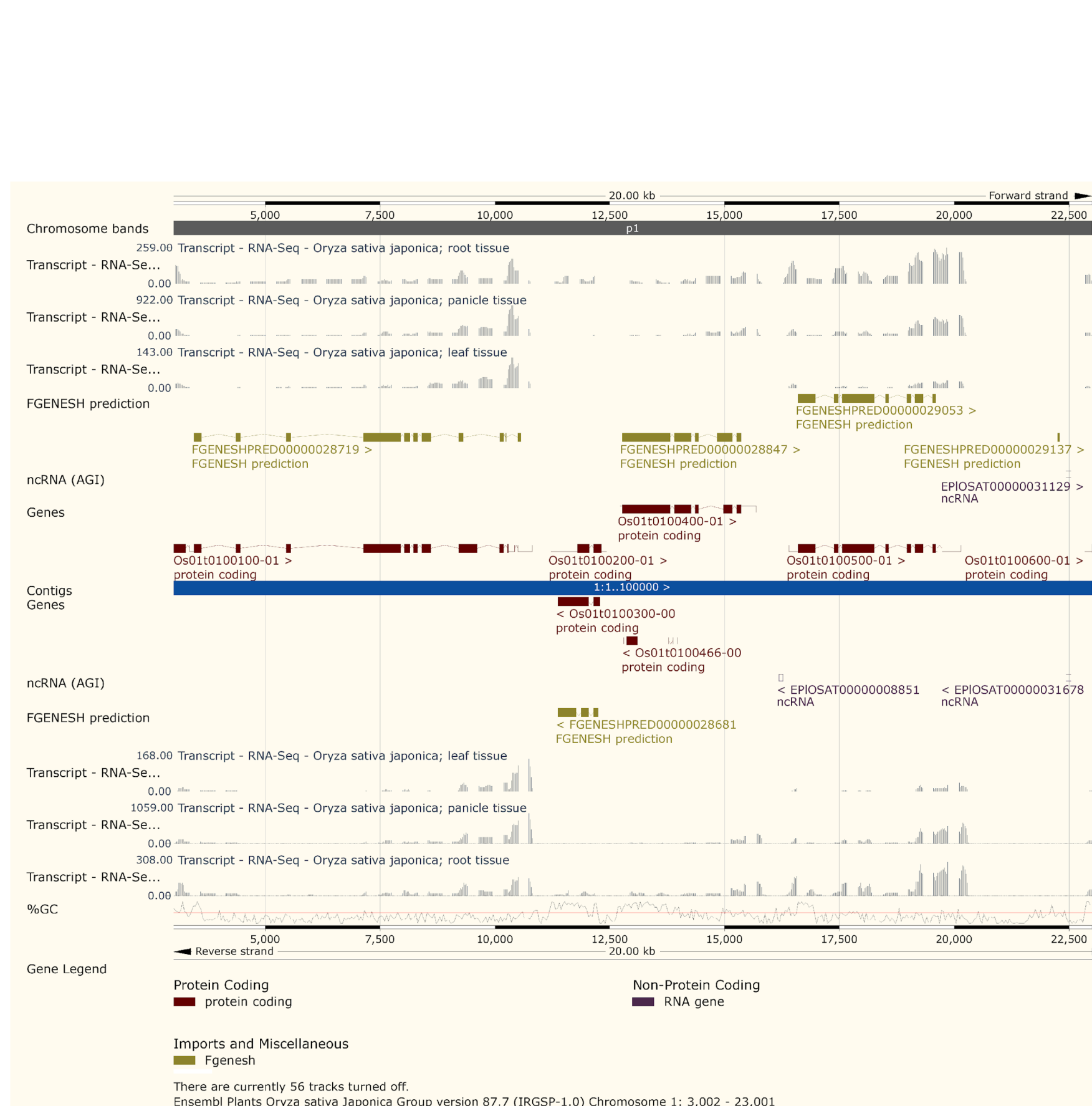
Abstract: Rice is one of the world's most important crops and, due to its small genome size (~400Mb), is a model system for plant genetics. The Gramene database was established 20 years ago to provide a comparative mapping resource for grasses, and the first rice genome was its cornerstone. Since then, Gramene has been involved in the sequencing, assembly and annotation of other rice varieties and crop genomes to expand the resource, in collaboration with community genome projects and Ensembl. Based on the Gramene/Ensembl infrastructure, we have established focused, pan-genome sites for specific crop research communities (rice, maize, sorghum, and grapevine). Gramene Oryza (https://oryza.gramene.org) release 5 (Oct. 2022) features 28 rice genomes and 8 reference genomes from other species to support comparative genomic analysis and model the evolution of gene functions. The hosted rice genomes include the MAGIC16 collection, domesticated African rice varieties, short life cycle KitaakeX, and heirloom US Carolina Gold Rice. The pan-genome site is based on 38K protein-coding gene family trees, which enable phylogenetic inference of orthologs and paralogs, provide insights into allelic genes within and across species, and catalog lineage specific expansion and contraction. In addition to individual genome browsers, we provide an integrated search interface with various ways to interrogate and visualize gene tree data. We are integrating functional annotations, phenotypes, environmental data, and genetic variation into the site and look forward to working with the community to coordinate efforts on data stewardship, biocuration and training. Gramene Oryza is funded by USDA-ARS-8062-21000-041-00D.

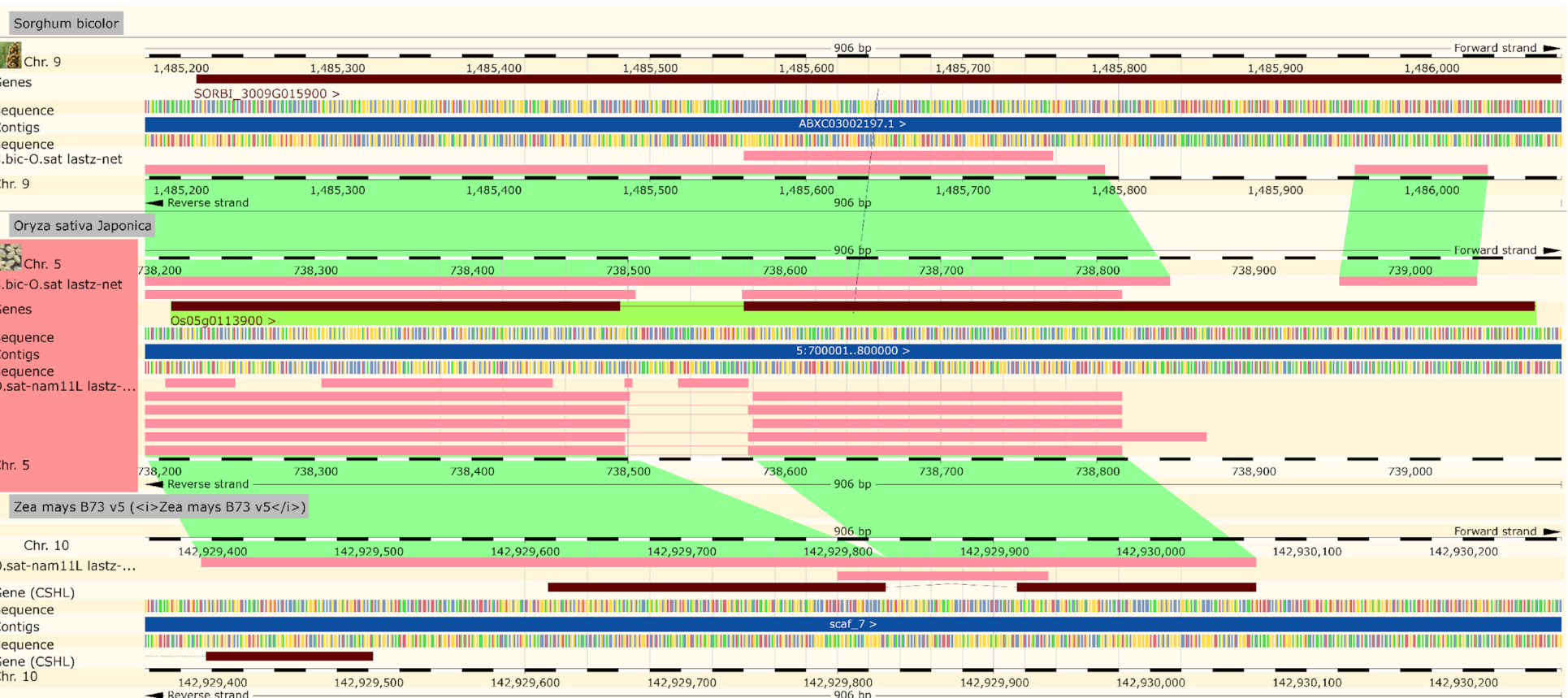## Gramene Oryza Release #6 (Jan. 2023) - Updates & highlights

- Functional curation for > 4000 genes from RAP-DB and NCBI's geneRIFs
- New Papers tab in Search results
- 3K SNP data for Nipponbare, MH63, IR64 and Azucena accessions
- 30 ortholog-based synteny maps between Nipponbare and other rice accessions, as well as Leersia, sorghum and maize
- Whole-Genome Alignments between Nipponbare, sorghum and maize
- Tissue specific RNAseq tracks for Nipponbare
- Gramene Search link added to Browser Gene pages
- Genome nomenclature updated based on recommended standards from the NSF-DBI (#2029854): CIBR-BBSRC: PanOryza: Globally coordinated genomes, proteomes and pathways for rice


- The figure to the left shows an example of synteny between Nipponbare IRGSP chr1 and US Carolina Gold contigs. There are 6 syntenic regions on Nipponbare chr1 corresponding to 6 Carolina Gold contigs. The synteny was built based on colinear orthologs with less than 5 intervening genes. On the bottom is a table listing Nipponbare – Carolina orthologs in this selected region.

- The figure above shows Whole Genome Alignment region comparison view between Nipponbare, sorghum and maize. Orthologs connected with a line.

A Nipponbare region shows tissue specific expression, the leaf tissue demonstrates a different expression pattern than root and panicle.

Gramene Search link at the bottom of the side bar on gene page.
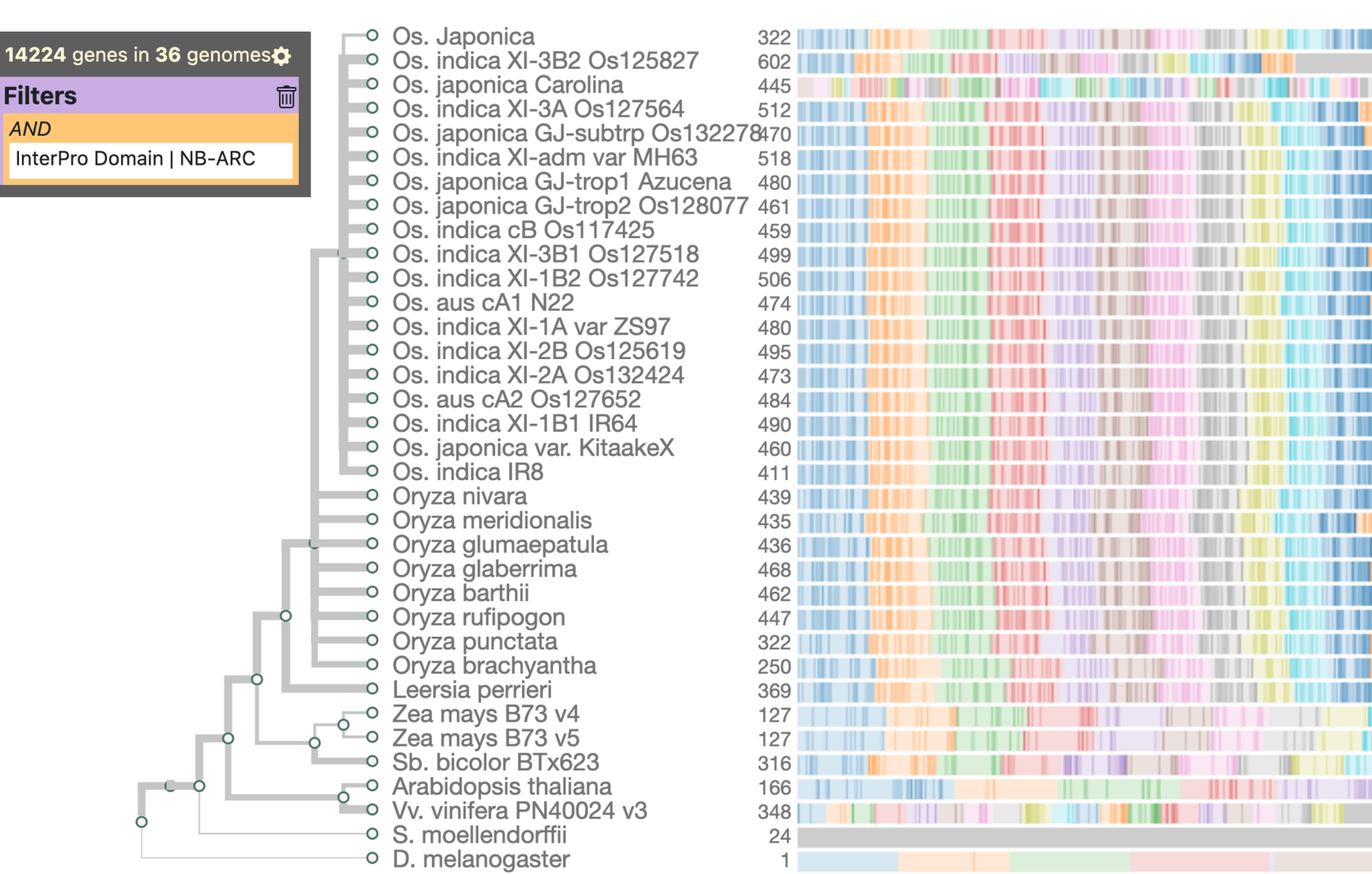
New names for each of the genomes

Gramene Oryza PanGenome (oryza.gramene.org)

- 28 domesticated & wild rice genomes
- Updated gene sets with Interpro protein domains annotated
- 37,423 gene trees
- 30 pairwise synteny maps
- 3 pairwise whole-genome alignments
- Gramene Search (c.v. keyword-based)
- BLAST Search (sequence-based)
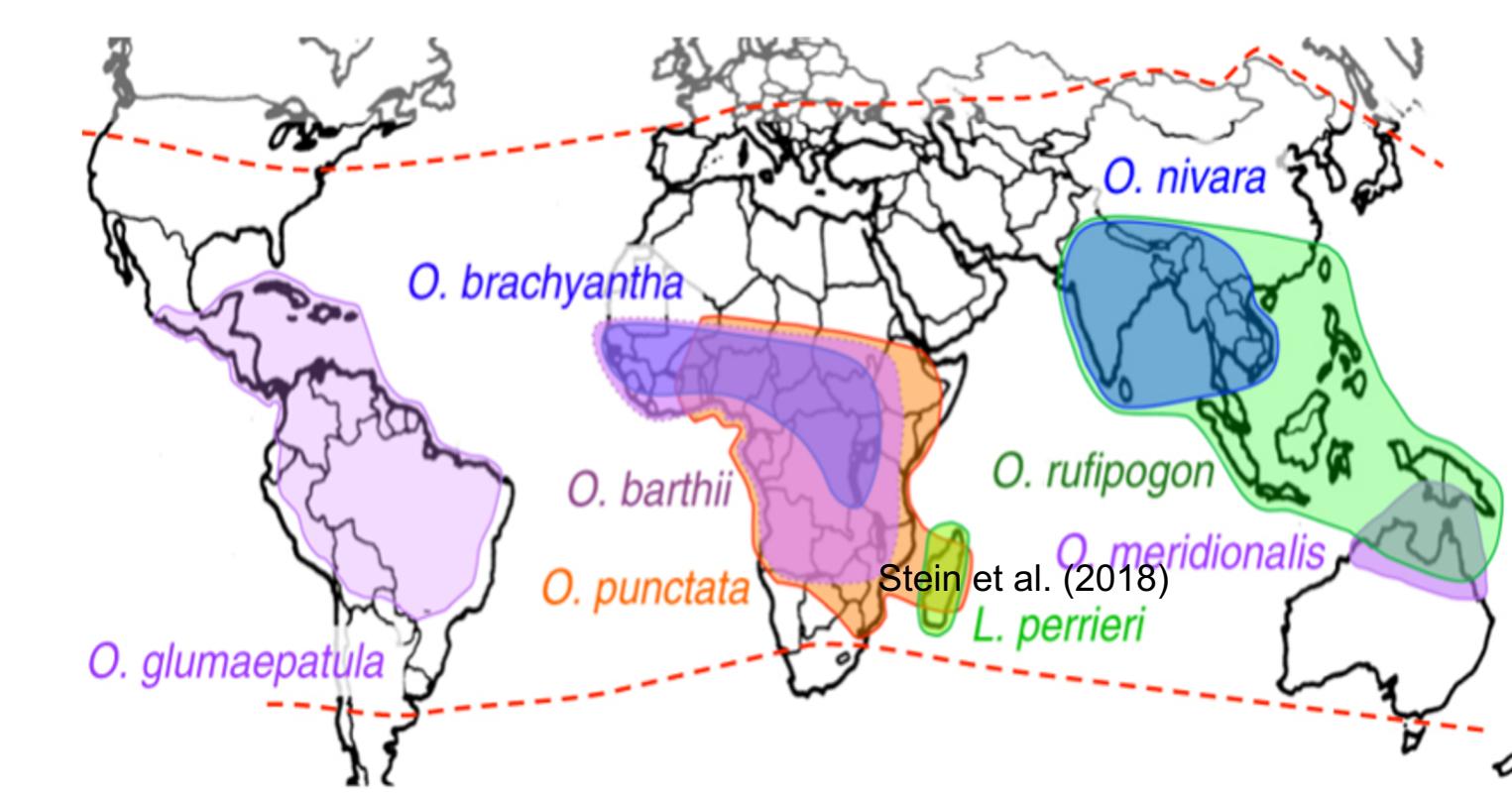
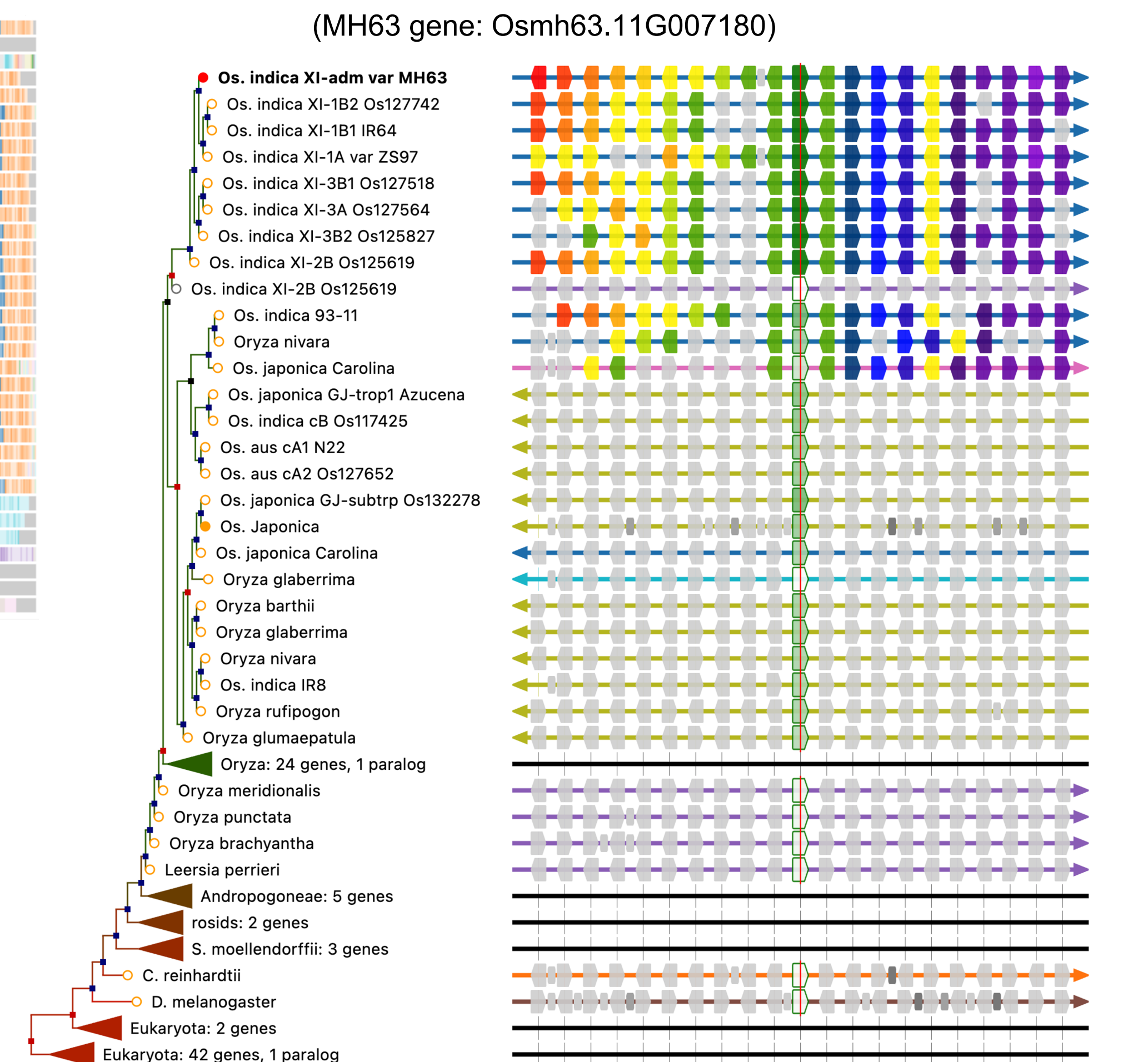### Gramene Search & Enhanced Tree Views

Pangenomic search summary

Alternate gene-tree view

(MH63 gene: Osmh63.11G007180)

The search for Interpro domain NB-ARC results in 14,224 genes from the 36 genomes hosted on Gramene Oryza panSite. The genomes are organized as a phylogenomic tree with color-coded chromosomes concatenated. These hits are marked as hashes along each genome bar (contigs are grey bars). For example, O.sativa Japonica Nipponbare has 322 hits distributed across all 12 chromosomes.

The right figure shows neighborhood gene tree view of a target gene, this gene family displays a clade specific conservation at this neighborhood region.

BLAST Search

The BLAST Search can run sequence search against up to 25 genomes at the same time, the right figure shows the results of the 25 blast jobs.

References

1. Song J.M. et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant, 14 (2021), pp. 1757-1767
   Zhou Y. et al. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci. Data, 7 (2020), p. 113
2. Stein J.C. et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nature Genetics. 50:285.
3. Vilella A.J. et al. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research 19 (2):327-35
4. Haas B.J. et al (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20 (18): 3643-46
5. Jones P. et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30 (9): 1236-40