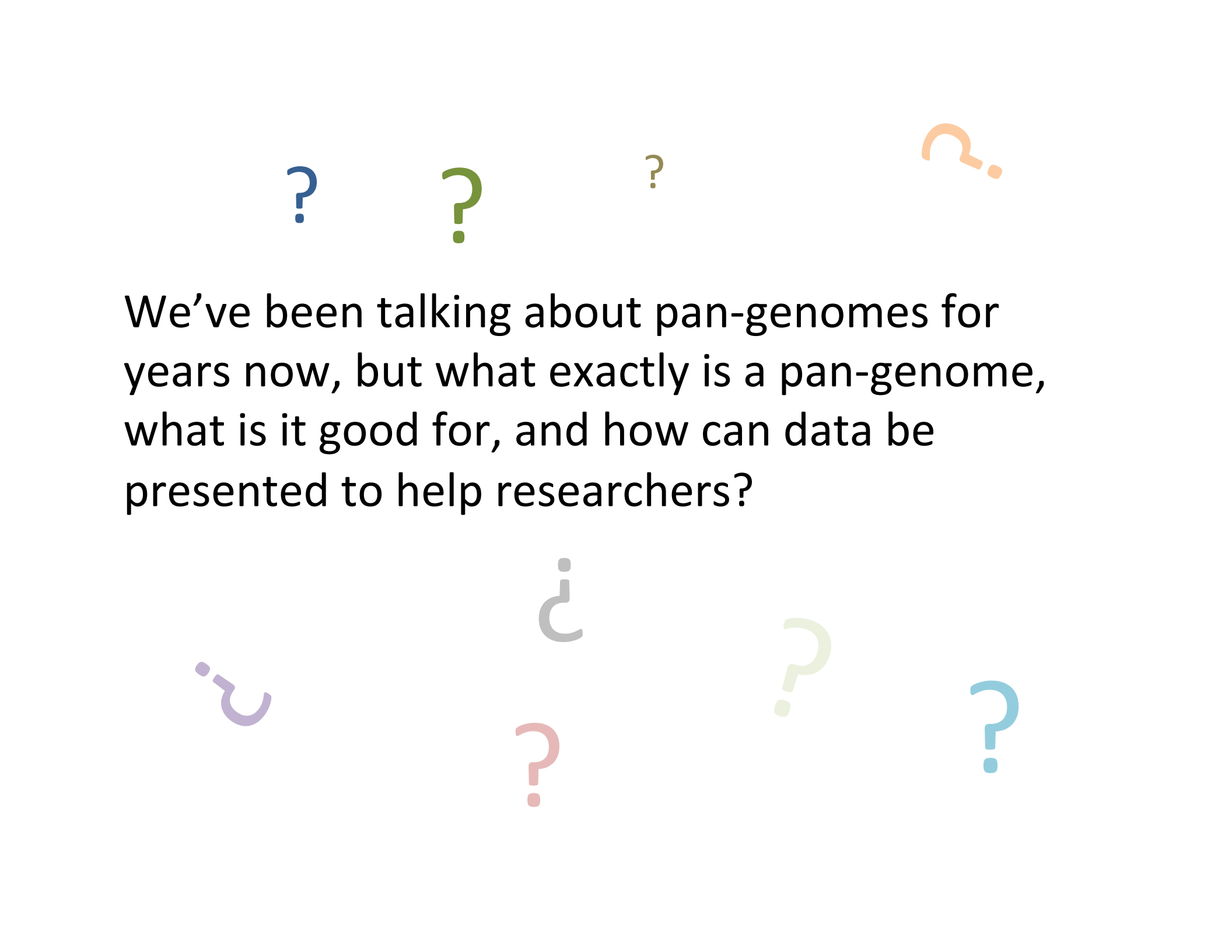
AgBioData pan-genome discussion

Wednesday, May 6th, 2020



We've been talking about pan-genomes for years now, but what exactly is a pan-genome, what is it good for, and how can data be presented to help researchers?

Topics:

1. What is a "pan-genome”
2. Examples of pan-genomes, tools, analyses.
3. Discussion: how can a pan-genome be useful, who is it useful for, how to make data accessible, what is the role of data portals?

Lightening talks:

Eloi Durant - PanacheFake pan-genome viewer
Alan Cleary − Genome Context Viewer
Marcela Tello-Ruiz - pangenome browsers for rice, maize, and grape

What is a pan-genome?

- Genome or gene focused.
- Could be reference-based or all-by-all.
- Capture large or small structural variation.
- Within a species or clade.
- Is it a graph, alignment, or set of syntentic relationships.
- When is it a pan-genome, when is it variation data? Is diversity data a form of pan-genome?

Some pan-genome portal examples

https://phytozome-next.jgi.doe.gov/cowpeapan

https://phytozome-next.jgi.doe.gov/brachypan

http://www.10wheatgenomes.com

http://animal.nwsuaf.edu.cn/code/index.php/panPig

http://animal.nwsuaf.edu.cn/code/index.php/panGoatTalks:

Some pan-genome visualizations

- Pan-tetris – standalone Java app for bacteria pan-genomes (download)
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547177/
- Rice pan-genome viewer: http://www.rmbreeding.cn/pan3k
  - https://www.ncbi.nlm.nih.gov/pubmed/27940610
- PPanGGOLiN: https://github.com/labgem/PPanGGOLiN
  - http://dx.doi.org/10.1371/journal.pcbi.1007732

Some pan-genome codefests

- https://github.com/NCBI-Hackathons/TheHumanPangenome - workshop to discuss tools in relation to pangenome analysis. Strategies and tools were presented. (Possible follow-up at Baylor College of Medicine October 11-13.)
  - https://f1000research.com/articles/8-1751
- https://graph-genome.github.io/ PantoGraph for SARS-CoV-2

How can a pan-genome be useful, who is it useful for, how to make data accessible, what is the role of data portals?

Presentations of pan-genomes, tools, analyses.

Eloi Durant - PanacheFake pan-genome viewer

https://meerketeer.ird.fr/PanacheFake

Alan Cleary − Genome Context Viewer

https://legumeinfo.org/lis_context_viewer/

Marcela Tello-Ruiz − pan-genome browsers for rice, maize, and grape

Rice: http://oge.gramene.org/ (maize, grapevine, and sorghum are under development)

How to think Pangenome Visualization?

Introducing Panache, a pangenome explorer prototype

Éloi DURANT

Introduction

eloi.durant@ird.fr
PhD student

Institut de Recherche pour le Développement
FRANCE

"Development of a tool for the visualization of plant pangenomes"

Thinking visualization

Does it to scale to big genomes?

What representations can be done?

Seeing everything at once is a fantasy.

Thinking visualization

Seeing everything at once is a fantasy.

Overly complicated data are unreadable as such, cf. the ‘Hairball effect’

Detail is reached on focusing on specific perts

Information = visualization x data processing

Panache

Summarization of information + exploration

Summarization: access to inner properties
Exploration: manipulation of multiple representations

Panache:
PANgenome **A**nalyzer with **CH**romosomal **E**xploration

Panache

https://meerketeer.ird.fr/PanacheFake

https://meerketeer.ird.fr/PanacheNapus

Panache


Minimal Presence ratio
to be part of Core
81%
Chromosome on display: 1
0
50k
100k
150k
200k
250k
300k
350k
400k
450k
500k
550k
600k
650k
700k
750k
800k
850k
900k
950k
1M
1.05M
1.1M
1.15M
1.2M
Geno10
Geno11
Geno12
Geno13
Geno14
Geno15
Geno16
Geno17
Geno18
Geno19
Geno20
Geno21
Geno22
One column represents one panBlock
A
B
C
Filled : Presence
Empty : Absence
Colour : Function
Zoom level
* At these zoom levels
lag might occur
1
2
3
4
5
6
7
8
9
10
11
12

What is missing?

Exploration through different

zoom scales

More representations

"Pangenomes, why not, but I don't want to loose all my previous analyses."

Data

Contact

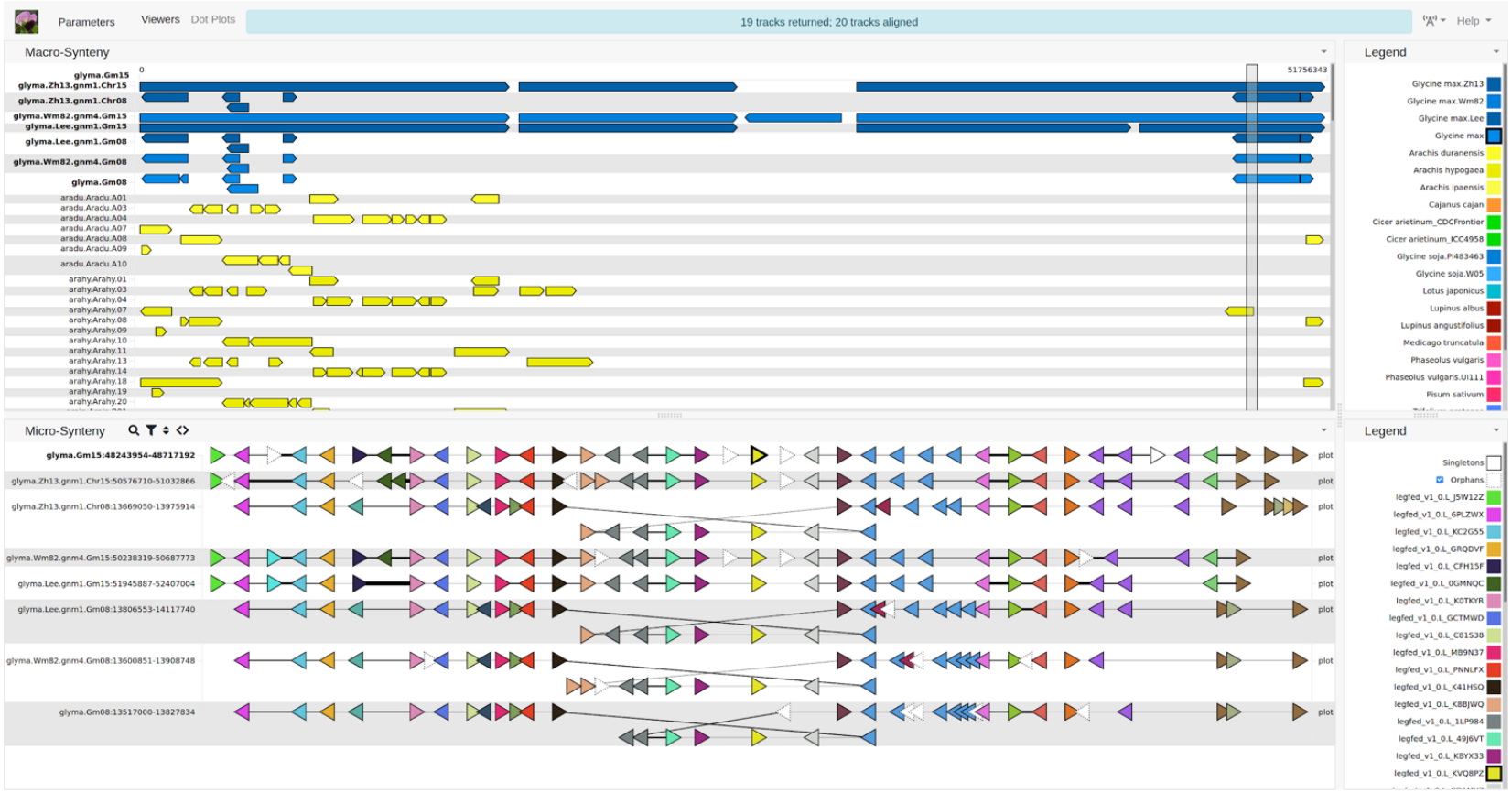
GitHub

SingingMeerkat/
Panache

eloi.durant@ird.fr

@MeerkatSinging

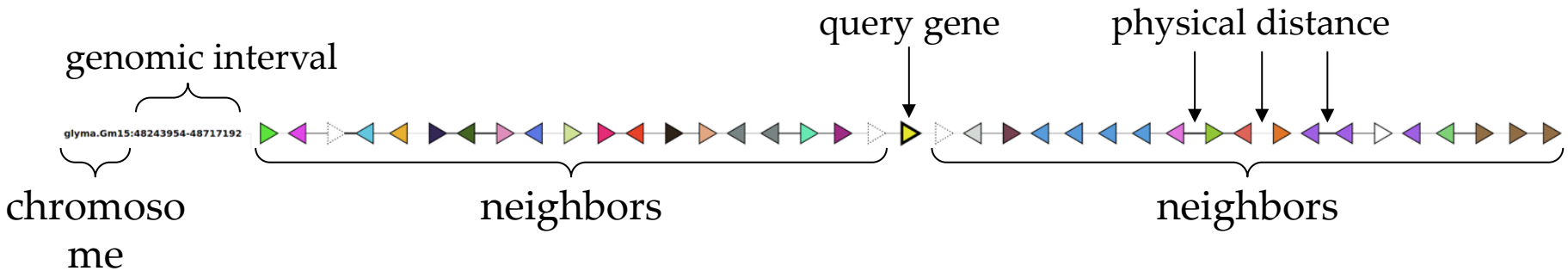
Genome Context Viewer



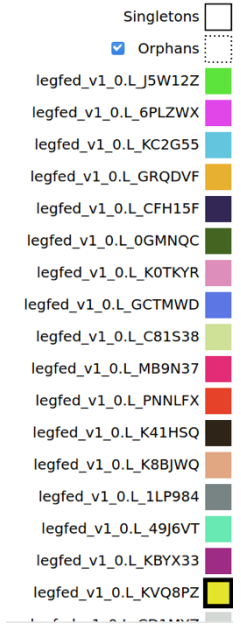
Alan Cleary

Micro-Synteny

Query Track



Functional Annotations

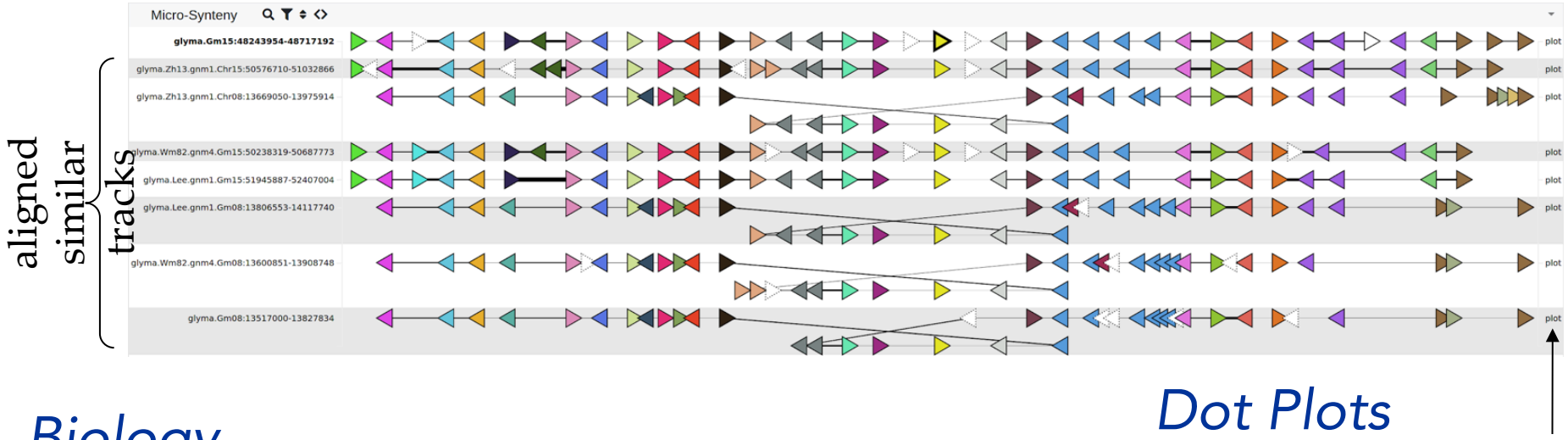


Biology

- Intra-track homology
  - Copy number

Micro-Synteny

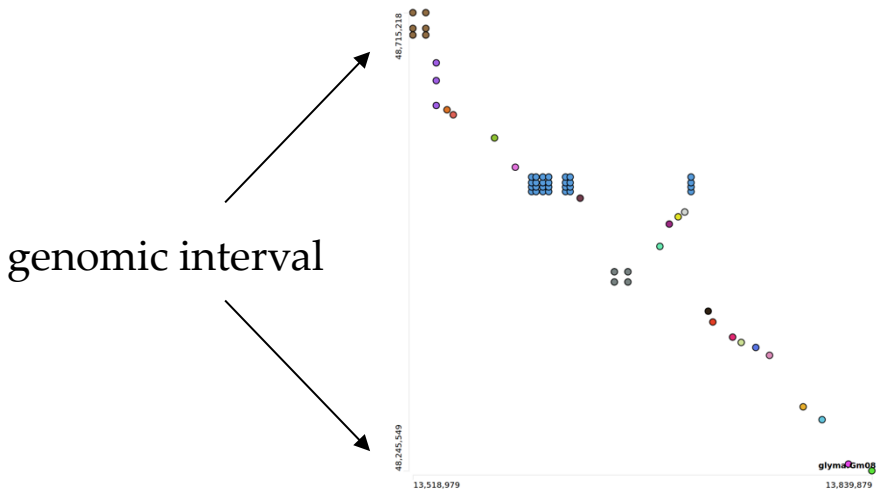
Track Search



Biology

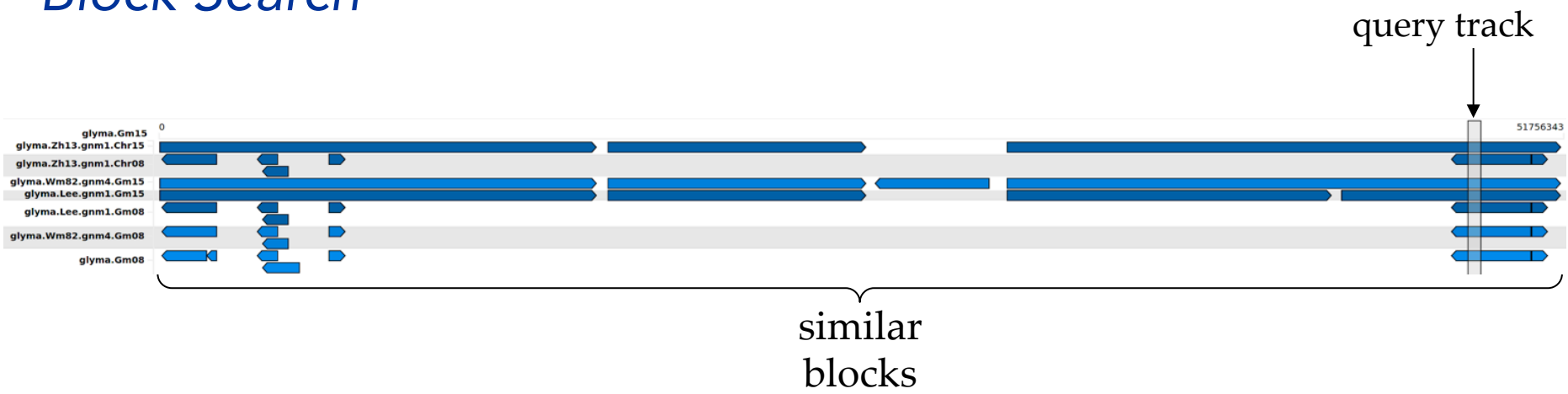
- Inter-track homology
  - Copy number variation
  - Gene presence/absence variation
  - Inversions

Dot Plots



Macro-Synteny

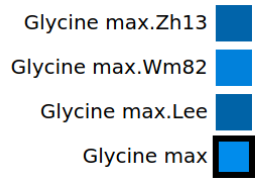
Block Search



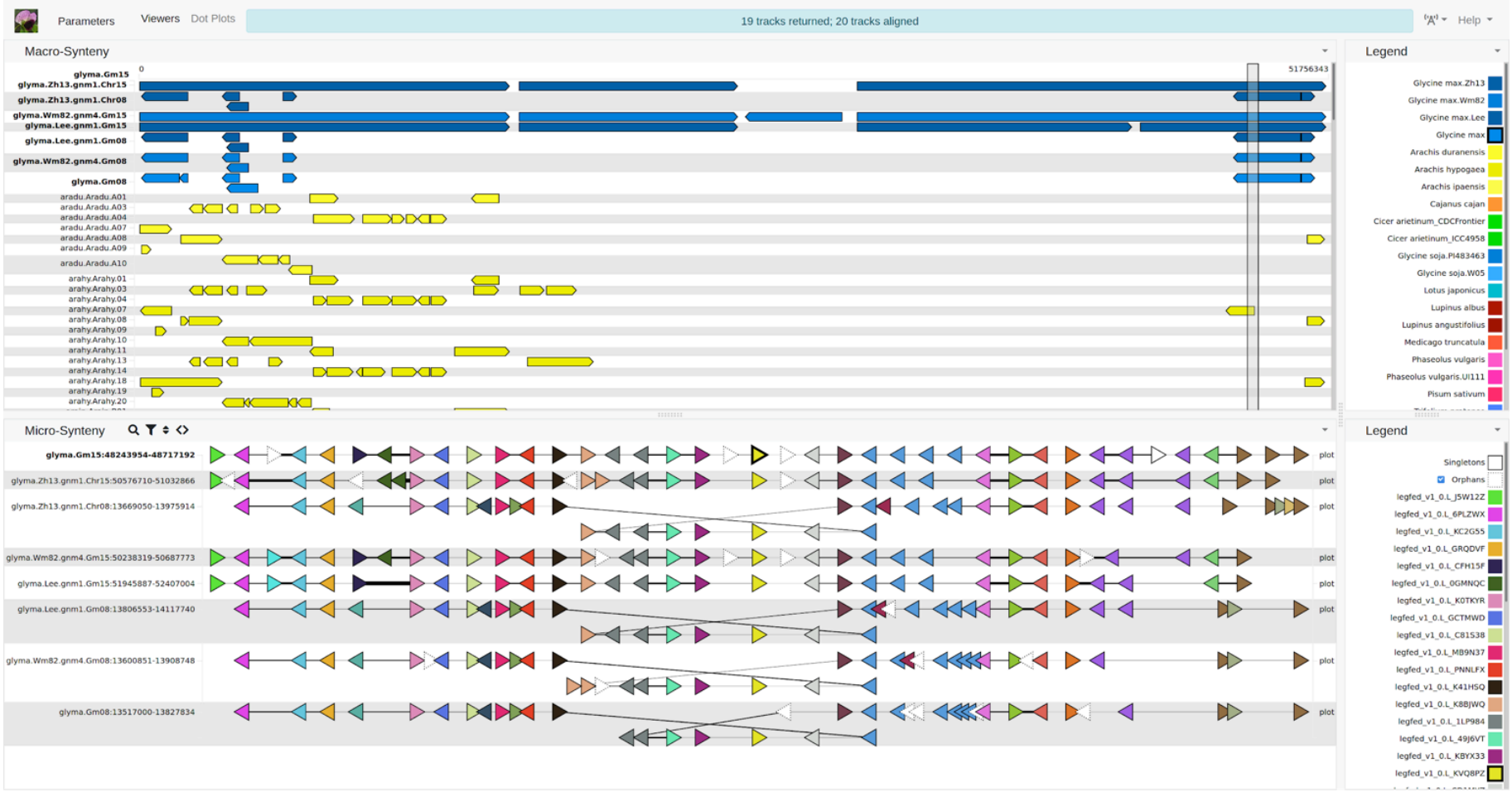
Biology

- Preservation of large structures

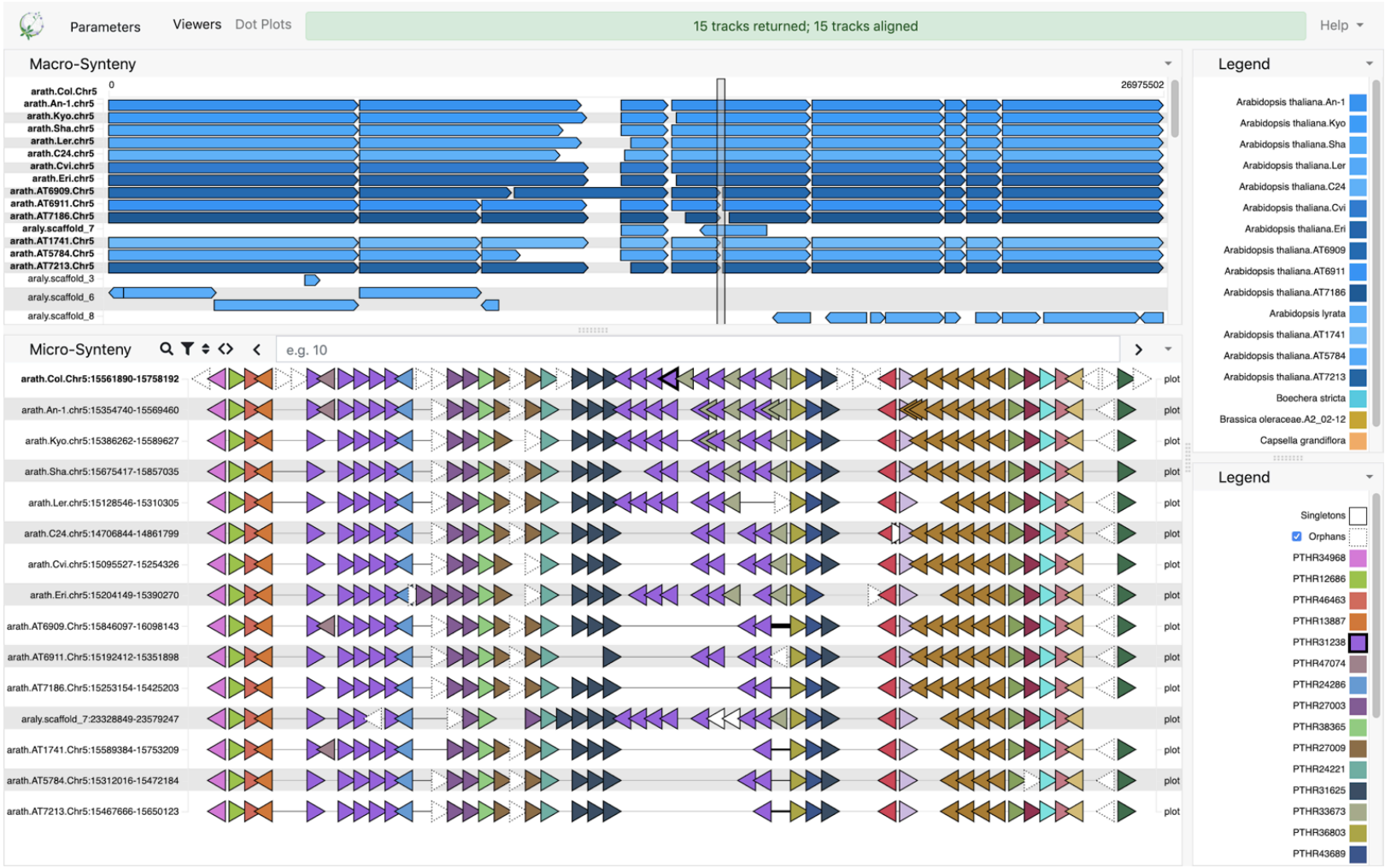
Species



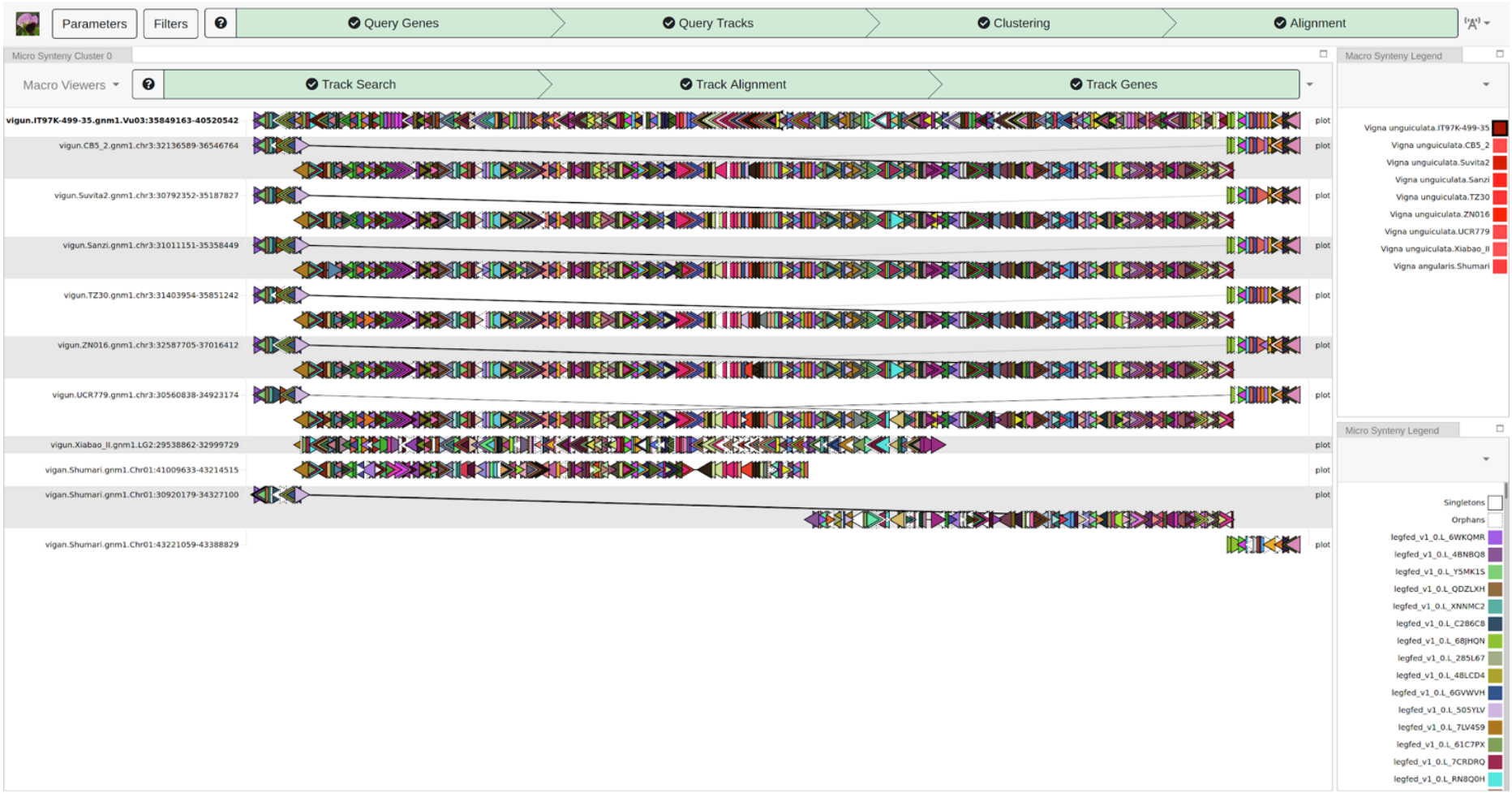
All Together



Pangenomics - *Arabidopsis thaliana*

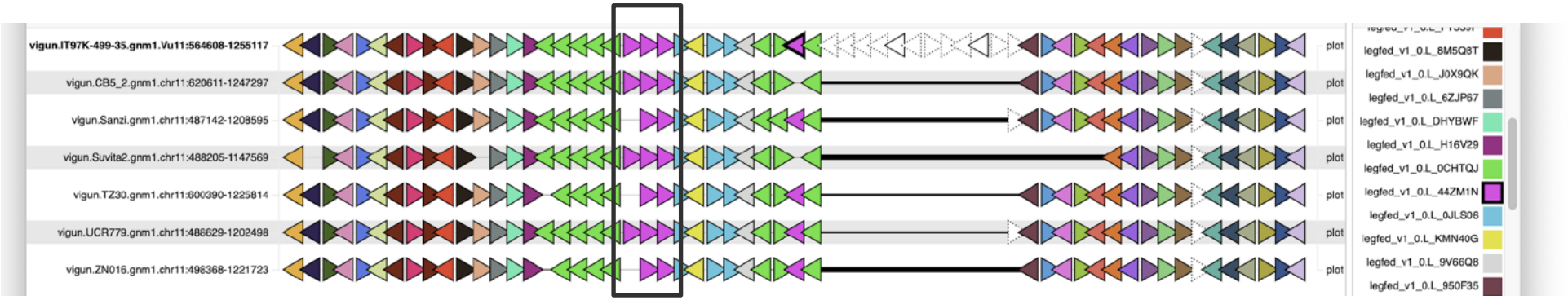


Pangenomics - *Vigna unguiculata*

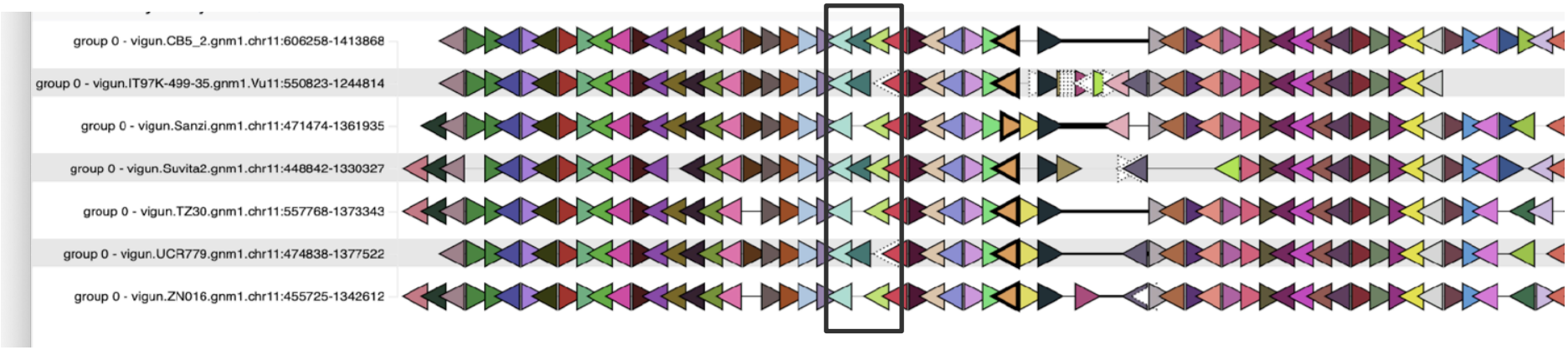


Pangenomics - *Vigna unguiculata* pan gene sets

Traditional gene families



Pan gene sets



Resources

Cleary, Alan, and Andrew Farmer. "**Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny**." *Bioinformatics* 34.9 (2017): 1562-1564.

https://legumeinfo.org/lis_context_viewer

https://github.com/legumeinfo/lis_context_viewer

Glycine example

Arabidopsis example

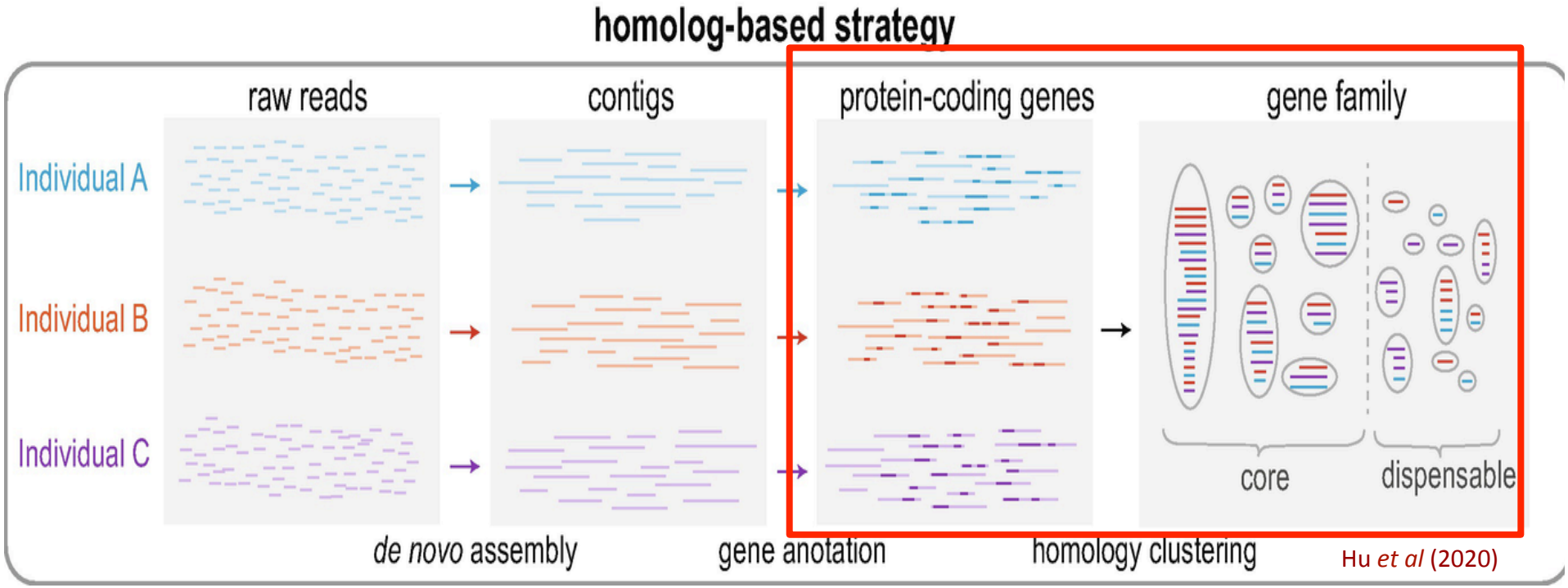
Plant PanGenome Browsers - Utilizing the Gramene & Ensembl Infrastructure

Marcela Karey Tello-Ruiz, PhD
May 6, 2020

Types of PanGenomes

Homolog-based strategy (all-by-all) - The genomes of individuals are independently assembled, and the presence/absence in a gene family is determined by clustering protein sequences into homologs.

1. **"Map-to-pan" strategy (reference-based)** - Pangenome sequences are constructed by combining a well-annotated reference genome with newly identified non-reference representative sequences, from which the presence/absence of a gene is then determined based on read coverage after individual reads are mapped to the pangenome. *Highly recommended for eukaryotic pangenome analysis.*

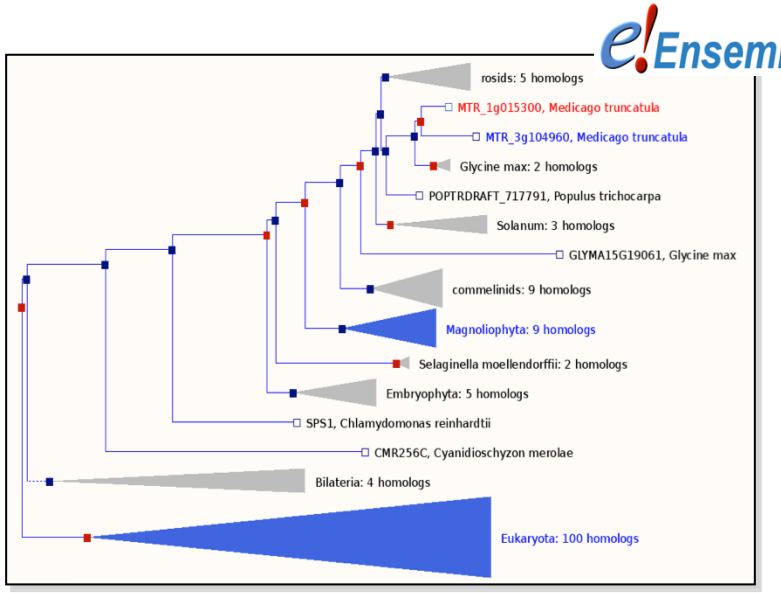


Hu *et al* (2020)

Using Comparative Phylogenomics to Support a Pan-Gene Space Browser

Phylogenetic Gene Trees

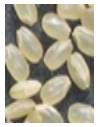
- Cluster homologous gene families
- Consensus of 5 tree-building methods
  - NJ-dN, NJ-dS, NJ-mm, Phyml-aa, Phyml-nt
- Infers orthologs and paralogs
- Taxonomic dating
- Interactive tree-browser for Cross-species



http://useast.ensembl.org/info/docs/compara/homology_method.html

Vilella *et al* (2008); Schwartz *et al* (2003); Kent *et al* (2003)

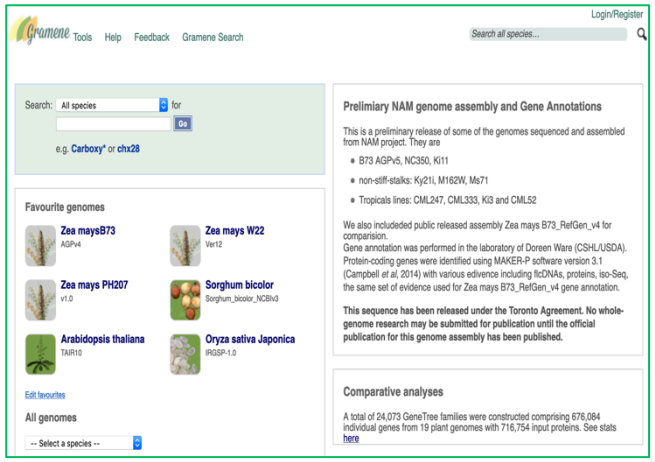
Pan-Genome (gene space) Browsers



Oryza Genome Evolution (oge.gramene.org)

Maize NAM Founders (maize-pangenome.gramene.org)

- PacBio/Bionano assembly of diverse maize inbreds
- Kelly Dawe (U Georgia), Matt Hufford (Iowa State U), Candice Hirsch, MaizeGDB: Carson Andorf, Maggie Woodhouse, Corteva: Kevin Fengler



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)

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

- Multiple PacBio &10X genomes
- Chad Hayes (USDA-ARS, Lubbock TX), Corteva, community data sets JGI, Terra Ref

Pan-Genome (gene space) Browsers

Subsites hold collections of closely related reference genomes

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

Uniform gene annotation protocol (in progress)

- Species-customized repeat library & evidence sets
- RNA-seq assemblies, PacBio Iso-seq, EST, prior annotation
- Evidence-based rediction

Gramene/Ensembl databases, Search, Views & Pipelines

Compara Gene Trees & whole genome alignment

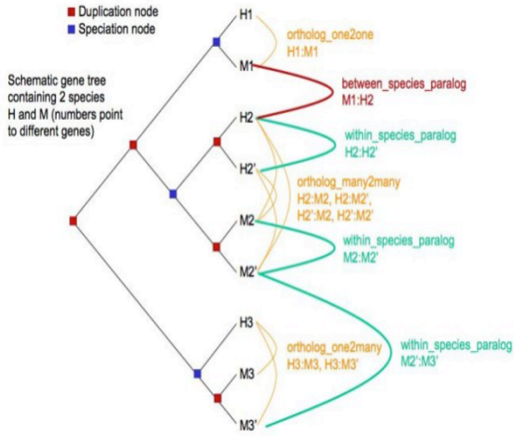
- Gene family assignment
- Phylogenetic tree build
- Ortholog & paralog calling
- Taxonomic dating
- Pairwise WGA (BLASTZ-CHAIN-NET)
- Genetic variations (SVs & SNPs)

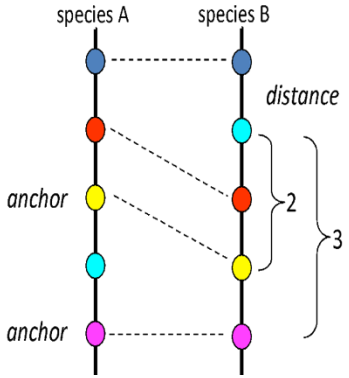
Gene-centered pairwise synteny maps

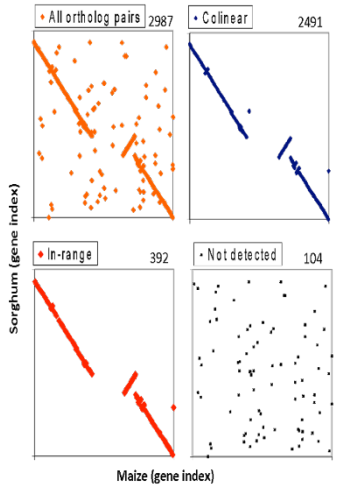
- Maps collinear & near-collinear orthologs
- Neighborhood view

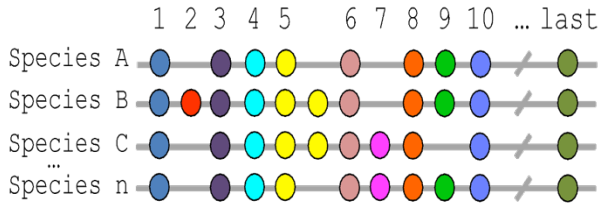
Pangenome index

- Cluster syntelogs by transitive closure
- Presence absence variation (PAV)
- Copy number variation (CNV)



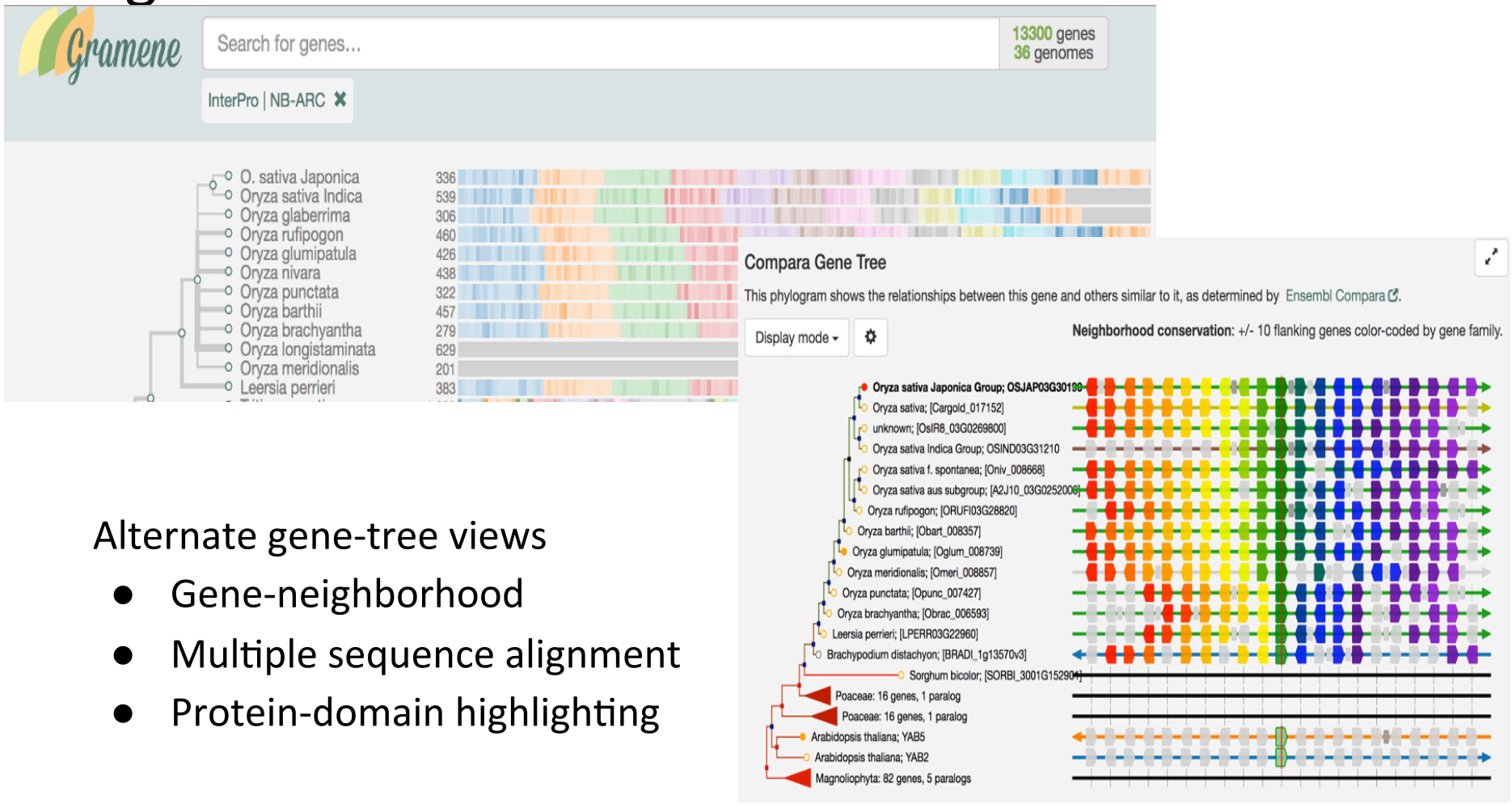






Gramene Search & Enhanced Tree Views

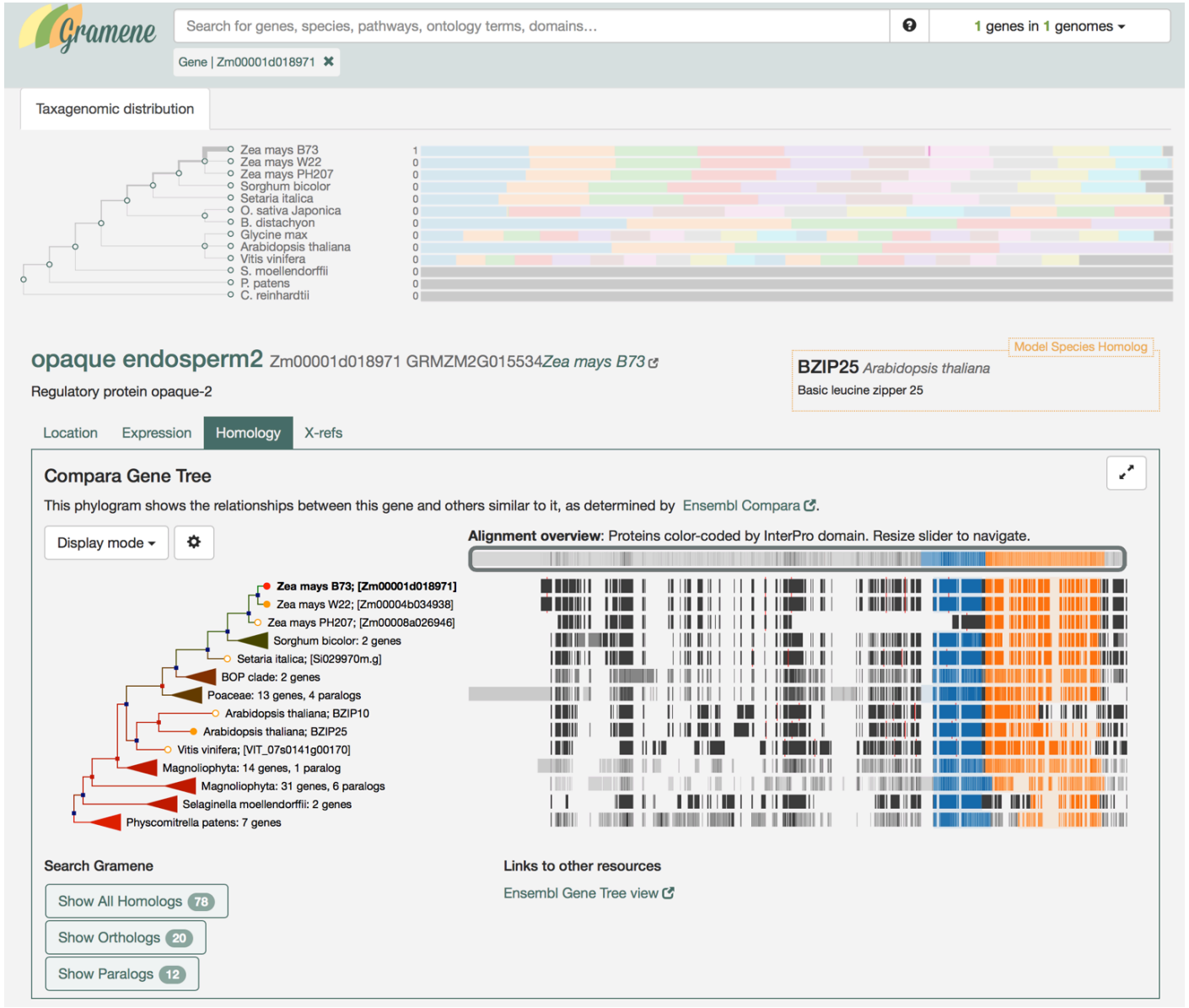
Pangenomic



Alternate gene-tree views

- Gene-neighborhood
- Multiple sequence alignment
- Protein-domain highlighting

Maize Pan-Genome: Gene tree alignment view

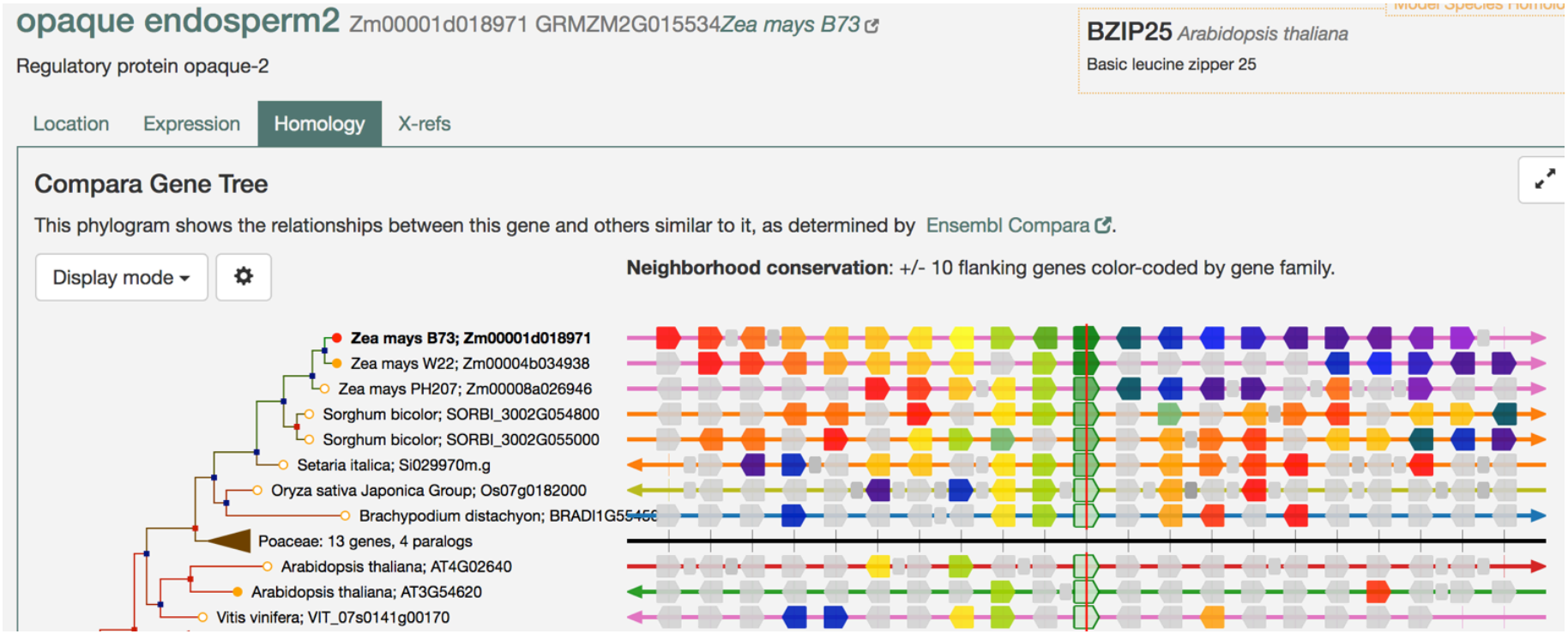


Prototype site: http://maize-pangenome-ensembl.gramene.org

Maize Pan-Genome: Multiple-sequence alignment view

Prototype site: http://maize-pangenome-ensembl.gramene.org

Maize Pan-Genome: Gene neighborhood conservation view



Prototype site: http://maize-pangenome-ensembl.gramene.org

Pan- Genome (gene space) Browsers

Subsites hold collections of closely related reference genomes

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

Uniform gene annotation protocol (in progress)

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

Gramene/Ensembl databases, Search, Views & Pipelines

Compara Gene Trees & whole genome alignment

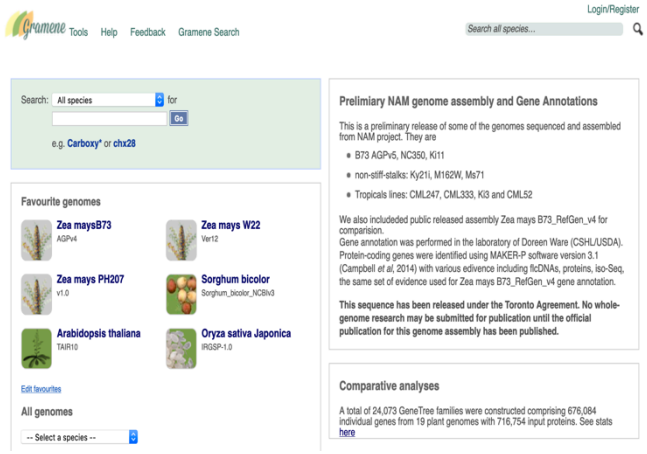
- Gene family assignment
- Phylogenetic tree build
- Ortholog & paralog calling
- Taxonomic dating
- Pairwise WGA (BLASTZ-CHAIN-NET)
- Genetic variations (SVs & SNPs)

Gene-centered pairwise synteny maps

- Maps collinear & near-collinear orthologs
- Neighborhood view

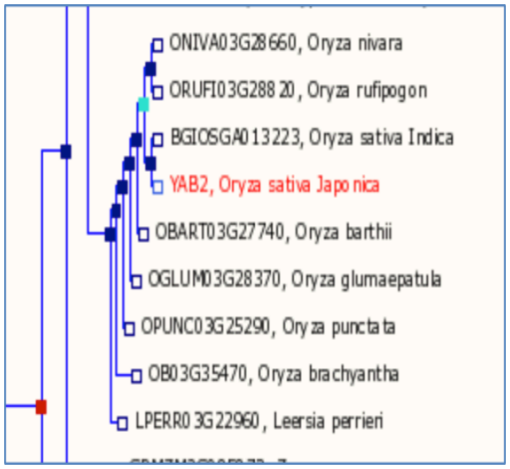
Pangenome index

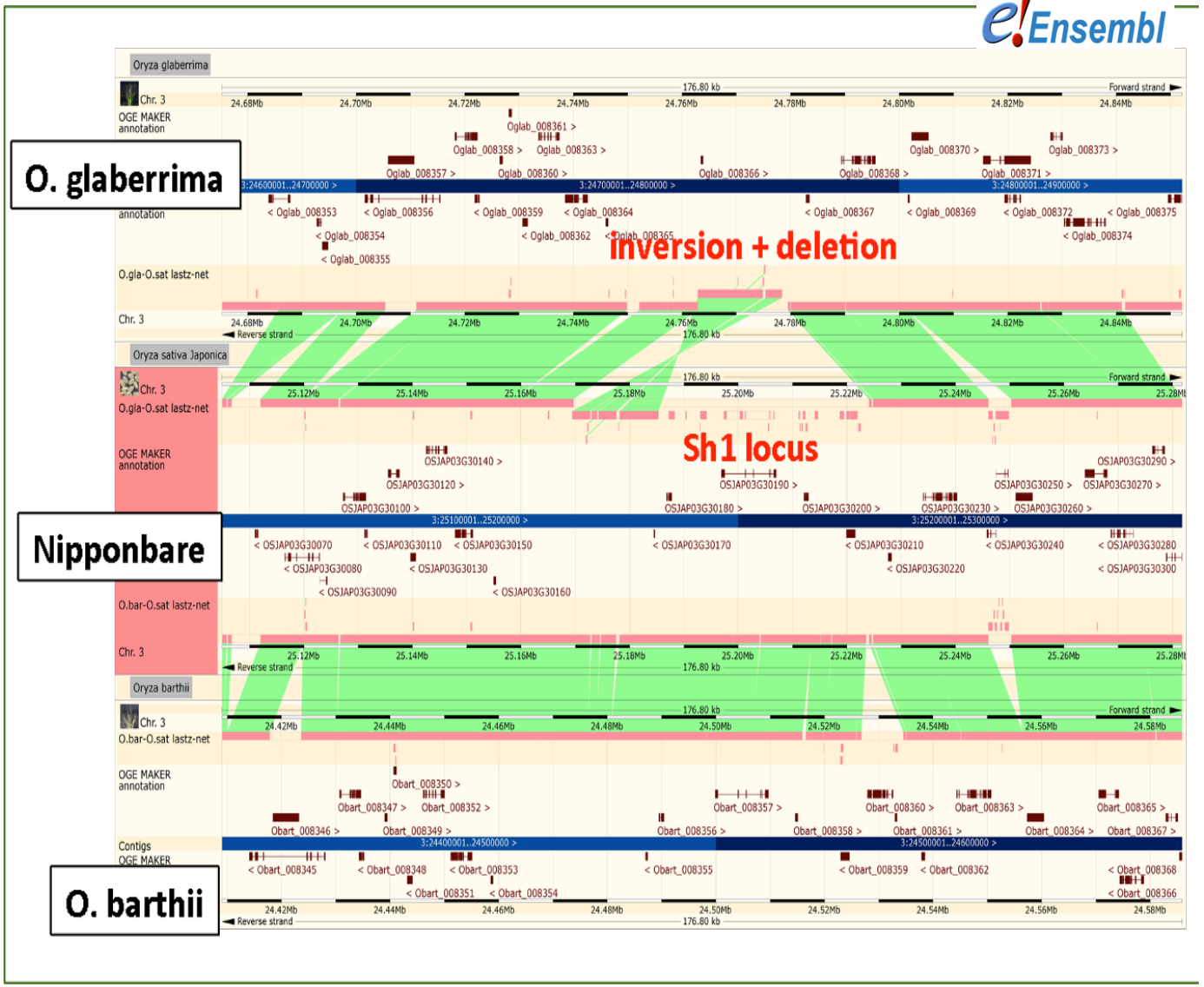
- Cluster syntelogs by transitive closure



Use case: Origin of Domestication genes for PAVs

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).





Pan- Genome (gene space) Browsers

Ensembl

Subsites hold collections of closely related reference genomes

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

Uniform gene annotation protocol (in progress)

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

Gramene/Ensembl databases, Search, Views & Pipelines

Compara Gene Trees & whole genome alignment

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

Gene-centered pairwise synteny maps

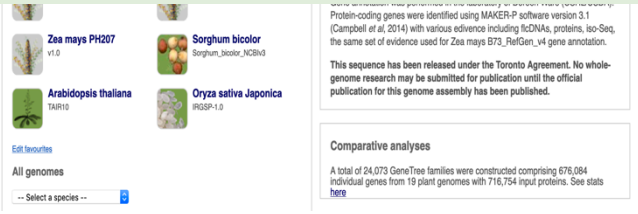
- Maps collinear & near-collinear orthologs
- Neighborhood view

Pangenome index

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

Future targets:

- Whole genome alignments compliment the protein gene trees and characterization of non-coding transcribe regions.
- Regulatory non transcribed regions



Pan- Genome (gene space) Browsers

Subsites hold collections of closely related reference genomes

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

Uniform gene annotation protocol (in progress)

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

Gramene/Ensembl databases, Search, Views & Pipelines

Compara Gene Trees & whole genome alignment

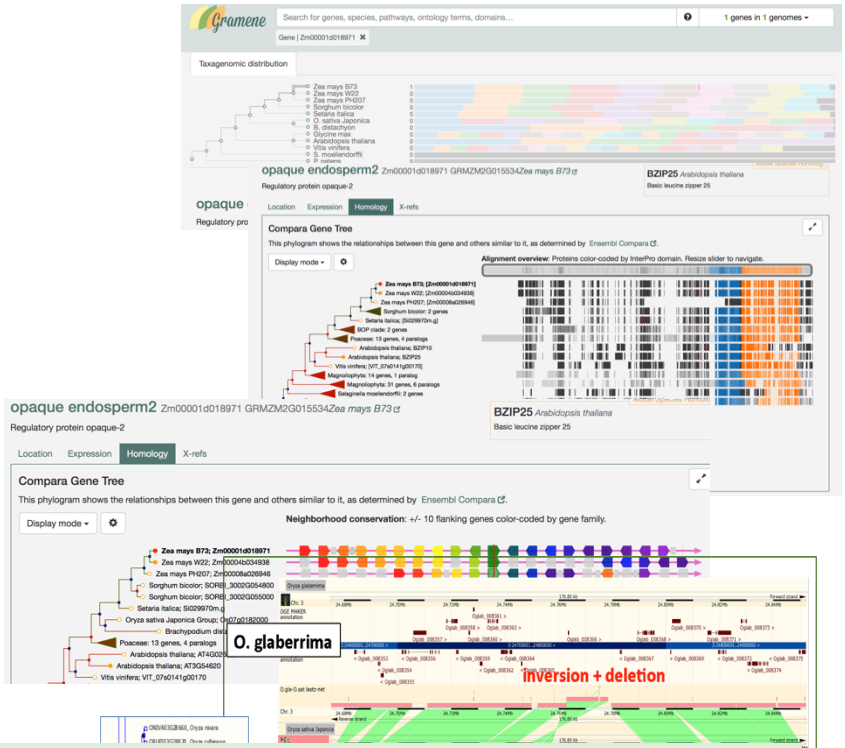
- Gene family assignment
- Phylogenetic tree build
- Ortholog & paralog calling
- Taxonomic dating
- Pairwise WGA (BLASTZ-CHAIN-NET)
- Genetic variations (SVs & SNPs)

Gene-centered pairwise synteny maps

- Maps collinear & near-collinear orthologs
- Neighborhood view

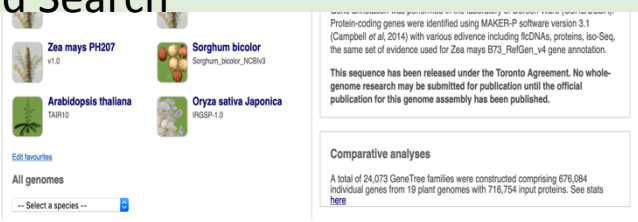
Pangenome index

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



Future targets:

- Support for analyses workflow to extract PAV & CNVs; Core & dispensable genomes
- Novel views to improve access and interpretation
- Improved Search



Thanks!

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



- **Sharon Wei** - Analyses, web & Ensembl software
- **Andrew Olson** - API development, search, views
- **Marcela K. Tello-Ruiz** - Species-specific collaborations & outreach
- Ware Lab members

Collaborators

- Ensembl - Infrastructure
- OGE project
- NAM project
- VG2 project

Some Introductory Papers

Bayer et al., 2017. Assembly and comparison of two closely related Brassica napus genomes.

Gao et al., 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor

Golicz et al., 2020. Pangenomics comes of age: From bacteria to plant and animal applications.

Montenegro et al., 2017. The pangenome of modern hexaploid bread wheat.

Sherman and Salzberg, 2020. Pan-genomics in the human genome era.

More references

Bayer PE, Hurgobin B, Golicz A, Chan K, Yuan Y, Lee HT, Renton M, Meng J, Li R, Long Y, Zou J, Bancroft I, Chalhoub B, King G, Batley J, Edwards D. (2017) Assembly and comparison of two closely related Brassica napus genomes. Plant Biotechnology Journal. 15 (12):1602-1610

Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D. (2020) Plant Pangenomics: Approaches, Applications and Advancements. Current Opinion in Plant Biology. 54: 15-25

Dolatabadian A, Bayer P, Tirnaz S, Hurgobin B, Edwards D, Batley J. (2020) Characterisation of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. Plant Biotechnology Journal. 18 (4): 969-982

Dolatabadian A, Patel DA, Edwards D and Batley J. (2017) Copy number variation and disease resistance in plants. Theoretical and Applied Genetics. 130 (12), 2479-2490

Golicz A, Bayer PE, Bhalla PL, Batley J, Edwards D. (2020) Pangenomics comes of age: From bacteria to plant and animal applications. Trends in Genetics 63(2): 132-145

Golicz AA, Bayer PE, Barker G, Edger PP, Kim HR, Martinez PA, Chan CKK, Severn-Ellis A, McCombie R, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang H, R. Teakle GR, Town CD, Batley J, Edwards D. (2016) The pangenome of an agronomically important crop Brassica oleracea. Nature Communications 7:13390

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014 Jan;26(1):121-35. doi: 10.1105/tpc.113.119982.

More references -continued

Hübner et al.. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants. 2019 Jan;5(1):54-62. doi: 10.1038/s41477-018-0329-0.he pangenome of an agronomically important crop Brassica oleracea. Nature Communications 7:13390.

Hurgobin B, Golicz A, Bayer P, Chan K, Tirnaz S, Dolatabadian A, Schiessl S, Samans B, Montenegro J, Parkin I, Pires C, Chalhoub B, King G, Snowdon R, Batley J and Edwards D. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. (2018) Plant Biotechnology Journal. 16 (7), 1265-1274

Hurgobin H and Edwards D. (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? Biology 6 (1): E21

Montenegro JDM, Golicz AA, Bayer PE, Hurgobin B, Lee HT, Chan CKK, Visendi P, Lai K, Doležel J, Batley J, Edwards D. (2017) The pangenome of modern hexaploid bread wheat. Plant Journal. 90 (5): 1007-1013

Ou et al., 2018. Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. New Phytologist Vol 220 (2): 360-363

Pinosio et al., 2016. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. Molecular Biology and Evolution, Volume 33, Issue 10, October 2016, Pages 2706–2719.

Read et al., 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature volume 499, pages 209–213.

Sun C1,2, Hu Z1,2, Zheng T3, Lu K1, Zhao Y1, Wang W3, Shi J4, Wang C3, Lu J1, Zhang D4,5, Li Z6, Wei C7,2. RPAN: rice pan-genome browser for ~3000 rice genomes. Nucleic Acids Res. 2017 Jan 25;45(2):597-605. doi: 10.1093/nar/gkw958.

More references- continued

Valliyodan et al.. 2019 Construction and comparison of three new reference-quality genome assemblies for soybean. The Plant Journal. 100 (5): 1066-1082

Varshney et al., 2019 Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nature Genetics 51, 857-864.

Wang et al., 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature volume 557, pg 43–49.

Yu J, Golicz A, Lu K, Dossa K, Zhang Y, Chen J, Wang L, You J, Fan D, Edwards D, Zhang X. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnology Journal. 17 (5): 881-892Bayer PE; Golicz A, Tirnaz S, Chan KCC, Edwards D, Batley J. (2019) Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. Plant Biotechnology Journal. 17 (4) :789-800

Zhao J, Bayer PE, Ruperao P, Saxena RK, Khan AW, Golicz AA, Nguyen HT, Batley J, Edwards D, Varshney RK. 2020 Trait associations in the pangenome of pigeon pea (Cajanus cajan) Plant Biotechnology Journal.