

**AgBioData : Pan-Genome Working Group**

**Running the spigot of genomes into the pan era;  
how should we pool data?**

Gerard R. Lazo  
USDA ARS Western Regional Research Center  
Albany, CA 94710-1105

gerard.lazo@usda.gov

May 1, 2023

## **AgBioData : Pan-Genome Working Group**

Established: January 2021

Last meeting: September 2022

Edoardo Bertolini (Danforth Plant Science Center),  
Clay Birkett (USDA, Triticeaetoolbox),  
Ethy Cannon (USDA, MaizeGDB) [chair],  
Steven Cannon (USDA, SoyBase/LIS),  
Alan Cleary (NCGR, LIS),  
Dave Edwards (Univ. Western Australia),  
Justin Elser (Oregon State Univ.),  
Jodi Humann (Washington State Univ., GeneSAS),  
Beant Kapoor (Univ. Tennessee),  
Marvellous O. Oyebanjo (Nigeria),  
Rashmi Jain (UC Davis. KitBase),  
Pankaj Jaiswal (Oregon State Univ., Gramene/Reactome),  
Ryan Layer (U. Colorado),  
Gerard Lazo (USDA, GrainGenes) [co-chair],  
Andrew Olson (CSHL, Gramene/SorghumBase),  
Doreen Ware (USDA, Gramene/SorghumBase),  
Kazim Wazir



## REFERENCE GENOMES

### virus:

SARS-CoV-2, V1 NC\_045512 (2020)  
TMV-U1 (1982)

### bacteria:

*Escherichia coli* K-12 (1997)  
*Salmonella typhimurium* 4/74 (2000)  
*Bradyrhizobium elkanii* USDA 110 (2006)

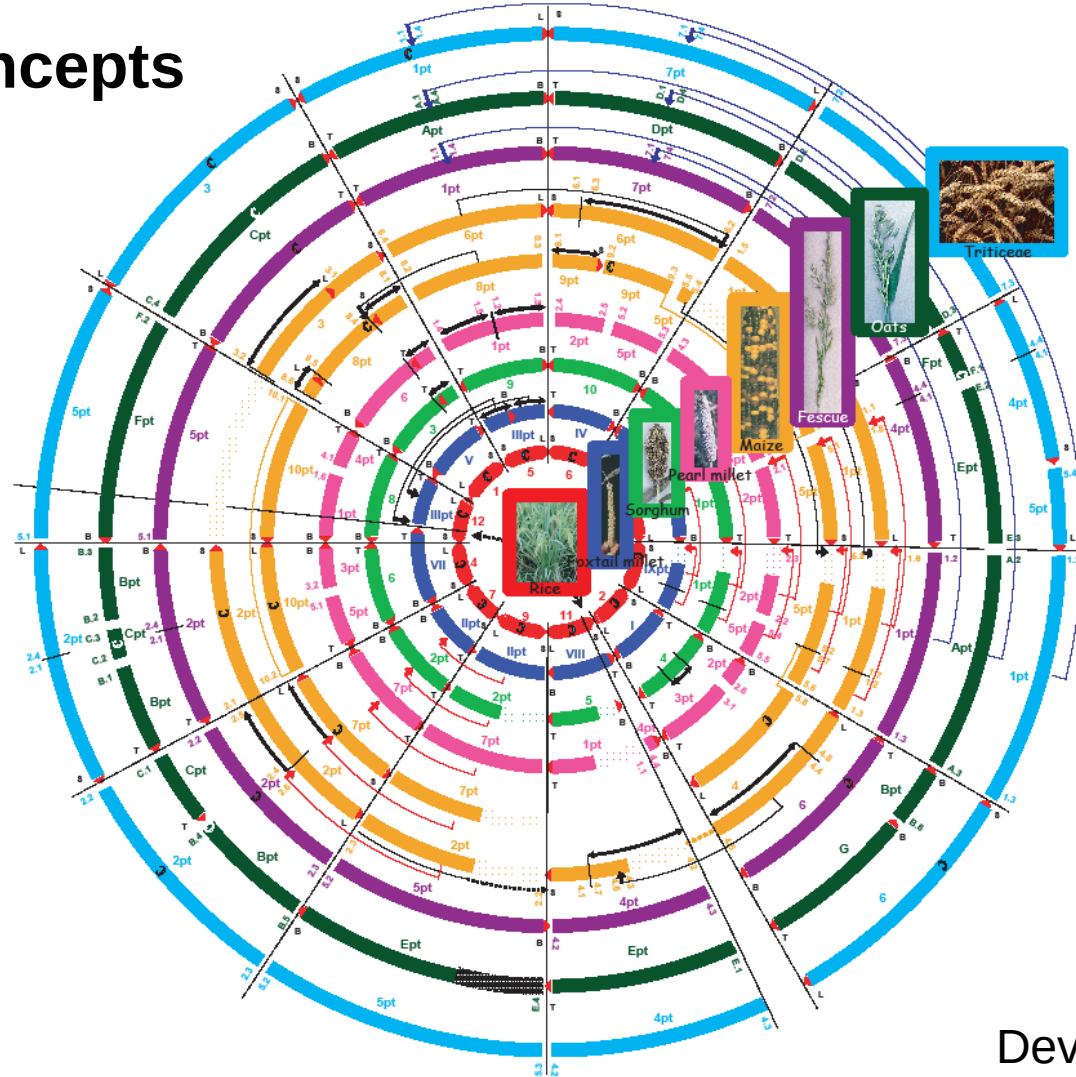
### plant:

*Arabidopsis thaliana* ec. Landsberg erecta (2000)  
*Glycine max* cv. Williams 82 (2005)  
*Hordeum vulgare* subsp. *vulgare* cv. Morex (2007)  
*Triticum aestivum* cv. Chinese Spring (2018)  
*Oryza sativa japonica* cv. Nipponbare (2005)  
*Zea mays* B73 (2005)

### animal:

*Homo sapiens* NC\_000001.1 (2001)  
*Bos taurus* UMD\_3.1 (2009)  
*Drosophila melanogaster* (2000)  
*Sus scrofa* NC\_003187 (2005)

# PanGenome Concepts



- Triticeae
- Oats
- Fescue
- Maize
- Pearl Millet
- Sorghum
- Foxtail Millet
- Rice

Devos and Gale (1997)

# REFERENCE GENOMES (e.g. *Triticum aestivum*)

- Triticum aestivum* cv. Chinese Spring (2018 -2021)
- Triticum aestivum* cv. Arina (2020)
- Triticum aestivum* cv. Julius (2020)
- Triticum aestivum* cv. Lancer (2020)
- Triticum aestivum* cv. Landmark (2020)
- Triticum aestivum* cv. Jagger (2020)
- Triticum aestivum* cv. Mace (2020)
- Triticum aestivum* cv. Mattis (2020)
- Triticum aestivum* cv. Norin (2020)
- Triticum aestivum* cv. Stanley (2020)
- Triticum aestivum* cv. Attraktion (2022)
- Triticum aestivum* cv. Fielder (2021)
- Triticum aestivum* cv. Kariega (2022)
- Triticum aestivum* cv. Renan (2022)





## Search

- Search & Browse GrainGenes
- Genetic Maps at GrainGenes

## Submit Your Data to GrainGenes

- Submit Your Data to GrainGenes
- GrainGenes Data Formats

## Community Services

## Quick Links



Search & Browse GrainGenes



Genome Browsers



BLAST



CMap



## GrainGenes Updates

- March 2023: Genome Browser External Links Tutorial (Video)
- March 2023: Barley Gene links to NordGen updated
- March 2023: SNP World was revamped and is available under the GrainGenes Tools menu
- February 2023: Stripe Rust QTL curated from the Vavilov wheat diversity panel
- February 2023: 2022 and 2013 Uniform Regional Scab Nursery for Spring

Legacy BLAST service

Tutorials

Feedback

Examples:

- Wheat
- Barley
- Oat
- Rye
- Multi Sequence
- Protein Sequence
- TREP Wheat

**Note: default BLAST parameters: -max\_target\_seqs 6**

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...



Examples:

[Wheat](#)
[Barley](#)
[Oat](#)
[Rye](#)
[Multi Sequence](#)
[Protein Sequence](#)
[TREP Wheat](#)

**Note: default BLAST parameters: -max\_target\_seqs 6**

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

## Wheat ABD Nucleotide Collections [\[Select all\]](#)

- [Wheat cultivar Sonmez pseudomolecules, Akpir](#)
- [Wheat Attraktion pseudomolecules, Kale et al. \(](#)
- [Wheat Attraktion contigs, Kale et al. \(2022\)](#)
- [Wheat Renan pseudomolecules, Aury et al. \(202](#)
- [Wheat Renan non-chromosomal scaffolds, Aury et](#)
- [Wheat Kariega v1 pseudomolecules \(masked\), A](#)
- [Wheat Chinese Spring IWGSC RefSeq v2.1 genom](#)
- [Wheat Fielder pseudomolecules, Sato \(2021\)](#)
- [10+ Wheat - ArinaLrFor v3.0 IPK pseudomolecu](#)
- [10+ Wheat - Jagger v1.0 IPK pseudomolecules \(:](#)
- [10+ Wheat - Julius v1.0 IPK pseudomolecules \(2](#)
- [10+ Wheat - LongReach Lancer v1.0 IPK pseudoi](#)
- [10+ Wheat - CDC Landmark v1.0 IPK pseudomol](#)
- [10+ Wheat - Mace v1.0 IPK pseudomolecules \(2C](#)
- [10+ Wheat - SY Mattis v1.0 IPK pseudomolecule:](#)
- [10+ Wheat - Norin61 v1.1 IPK pseudomolecules](#)
- [10+ Wheat - Triticum spelta PI190962 v1.0 IPK p:](#)
- [10+ Wheat - CDC Stanley v1.2 IPK pseudomolecu](#)
- [10+ Wheat - Cadenza Elv1.1 Ensembl scaffolds \(20:](#)
- [10+ Wheat - Claire Elv1.1 Ensembl scaffolds \(2020\)](#)
- [10+ Wheat - Paragon Elv1.1 Ensembl Scaffolds \(20:](#)
- [10+ Wheat - Robigus Elv1.1 Ensembl scaffolds \(202](#)

## Barley Collections [\[Select all\]](#)

- [Barley Morex v3 pseudomolecules \(2021\)](#)
- [HvBarke Barley Barke Assembly 2 Oct 2012](#)
- [Barley pan-genome - Akashinriki v1.0 pseudomolecu](#)
- [Barley pan-genome - B1K-04-12 v1.0 pseudomolecu](#)
- [Barley pan-genome - Barke v1.0 pseudomolecules \(2](#)
- [Barley pan-genome - Golden Promise v1.0 pseudom](#)
- [Barley pan-genome - HOR 10350 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 13821 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 13942 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 21599 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 3081 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 3365 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 7552 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 8148 v1.0 pseudomolecu](#)
- [Barley pan-genome - HOR 9043 v1.1 pseudomolecu](#)
- [Barley pan-genome - Hockett v1.0 pseudomolecules](#)
- [Barley pan-genome - Igri v1.0 pseudomolecules \(202](#)
- [Barley pan-genome - OUN333 v1.0 pseudomolecules:](#)
- [Barley pan-genome - RGT Planet v1.0 pseudomolecu](#)
- [Barley pan-genome - ZDM 01467 v1.0 pseudomolecu](#)
- [Barley pan-genome - ZDM 02064 v1.0 pseudomolecu](#)
- [Barley Morex V2 genome assembly by TRITEX \(2019\)](#)

## PanOat Collection (currently under embargo) [\[Select all\]](#)

- [PanOat Amagalon Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Aslak Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Atlantica Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Bannister Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Bilby Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Bingo Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Byzantina Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Delfin Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Eriantha Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat A. Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat FM13 Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Gehl Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat A. Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat GS7 Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Hatives Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat HiFi Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat A. Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Lion Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat A. Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Nicolas Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat A. Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat OT380 Pseudomolecules \(2023\)  \$\neq\$](#)
- [PanOat Park Pseudomolecules \(2023\)  \$\neq\$](#)

# Optimization of BLAST searches for pangenomes

- Additional genomes may have impacts on compute resources



hexaploid ABD genome  $2N = 6x = 42$ : Three cultivar example assuming 700 Mb per chromosome.



# Optimization of BLAST searches for pangenomes

## - Consolidation of Chromosomes Across Group Assignments

SequenceServer 2.0.0

[Help & Support](#)

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

### Nucleotide databases [\[Select all\]](#)

- Triticeae Group 1 chromosomes (restricted to wheat germplasm)
- Triticeae Group 2 chromosomes (restricted to wheat germplasm)
- Triticeae Group 3 chromosomes (restricted to wheat germplasm)
- Triticeae Group 4 chromosomes (restricted to wheat germplasm)
- Triticeae Group 5 chromosomes (restricted to wheat germplasm)
- Triticeae Group 6 chromosomes (restricted to wheat germplasm)
- Triticeae Group 7 chromosomes (restricted to wheat germplasm)
- Triticum aestivum L. cv. Chinese Spring IWGSC RefSeq v2.1 (2021)

Advanced parameters:

eg: -evalue 1.0e-5 -num\_alignments 100

?

Open results in new tab

BLAST

# Rice Super Pan-genome Information Resource Database

The web site and its dataset can be used freely by rice researchers.

The following data are accessible:

1. Sequences and gene annotations of each assembly
2. Reference-free whole-genome alignment of all assemblies
3. A super pan-genome graph and the associated gene annotations
4. K-mer spectrums and a variation graph used to link query data and the super pan-genome variations
5. De novo transposable element annotations of each assembly

RicesuperPIRdb.com

251 rice  
accessions



# VISUALIZATION

## Why Pan-Genomes?

Evolutionary Studies  
Identification of Introgression  
Study of Important Gene Families  
Crop Improvement

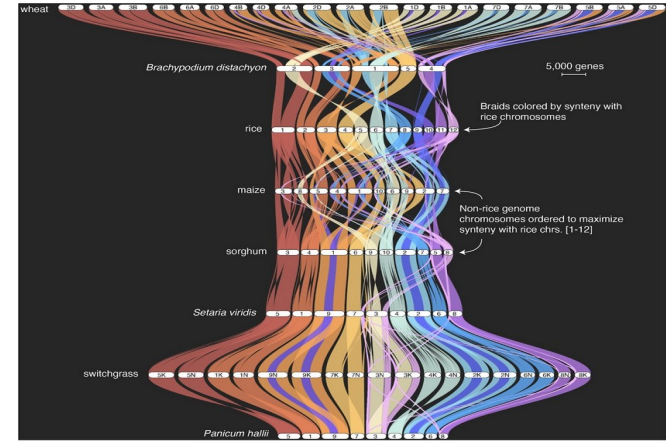
## Existing Resources

Nice Overviews – sometimes too complex  
New Tools – continued refinements  
Latest Reference Quality Release

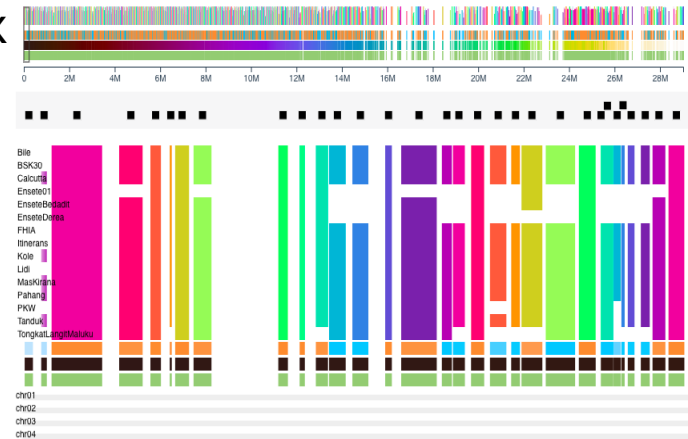
A Review of Pangenome Tools and Recent  
Studies (2020) Vernikos.

[pangenome.github.io](https://pangenome.github.io)

[cmdcolin.github.io/awesome-genome-visualization](https://cmdcolin.github.io/awesome-genome-visualization)

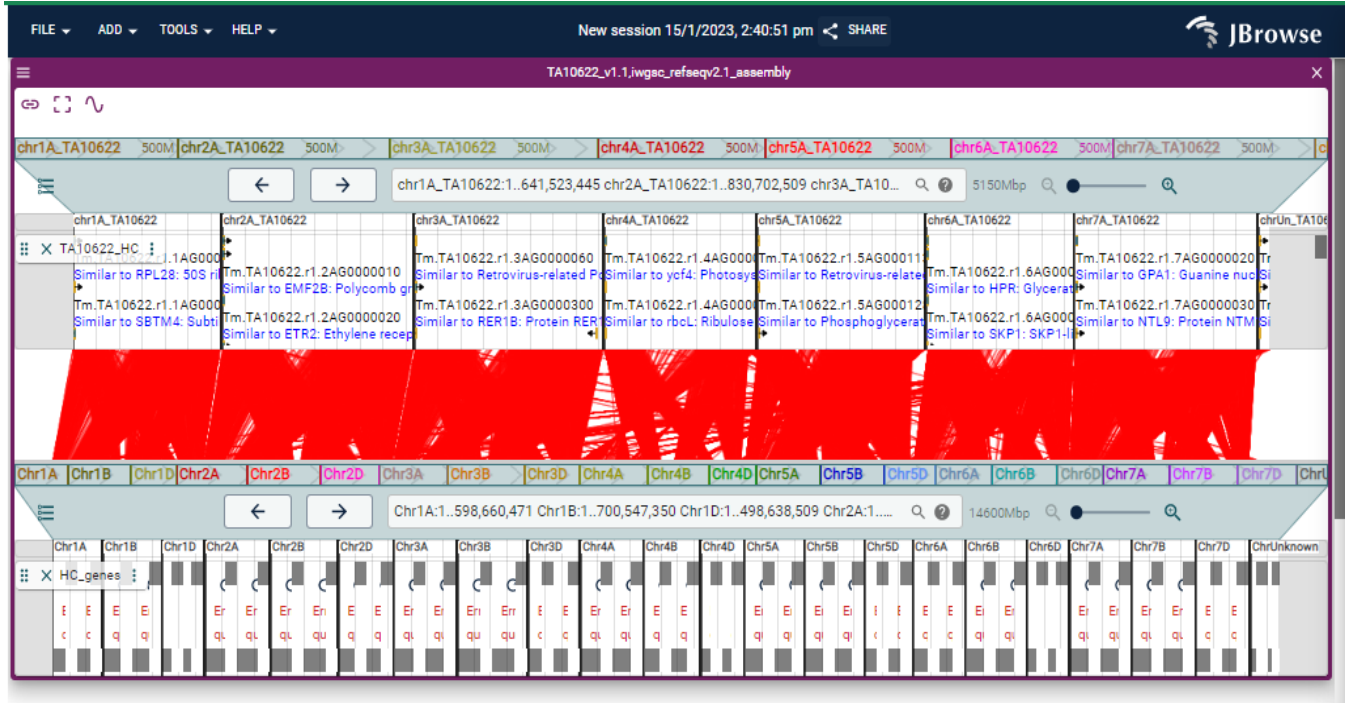


GENESPACE (bioRxiv)



Panache (bioRxiv)

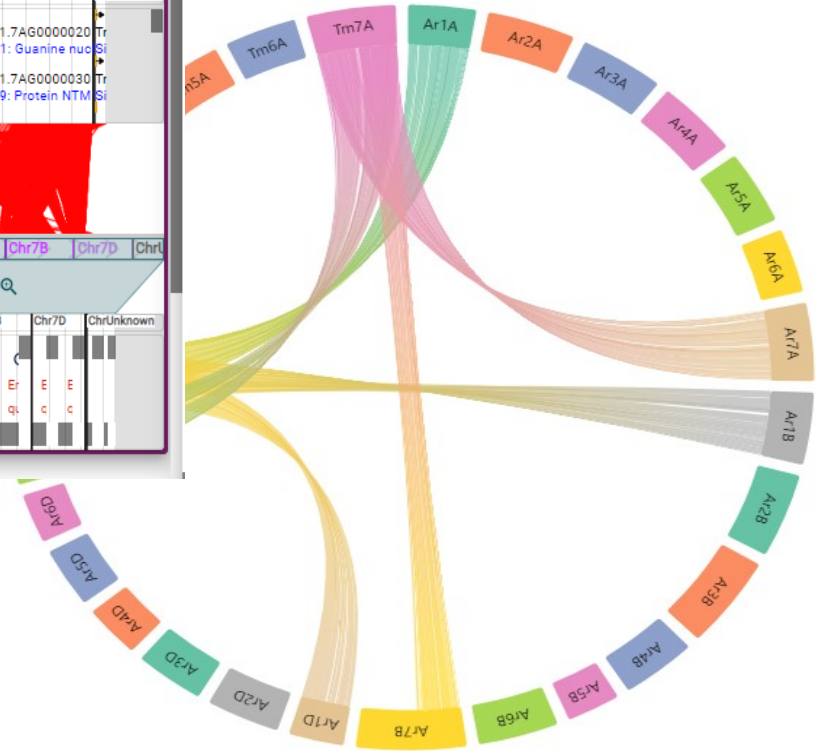
# JBrowse2 and Accusyn adapted for Comparative Displays



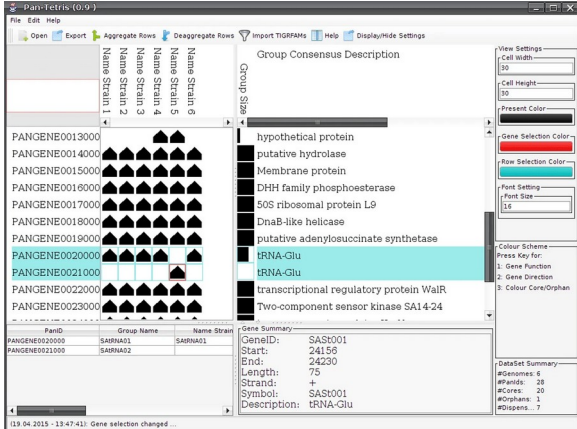
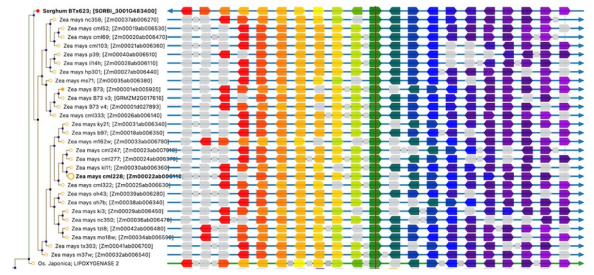
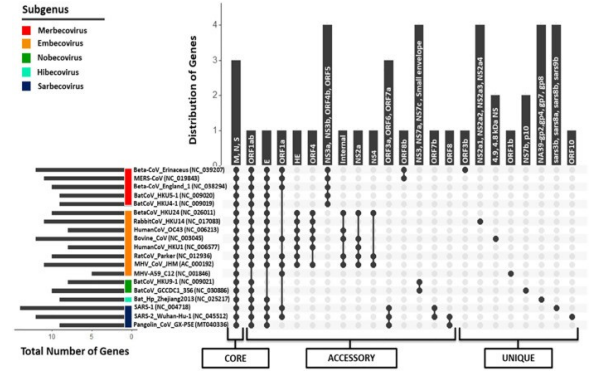
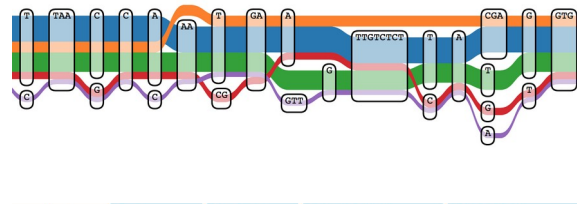
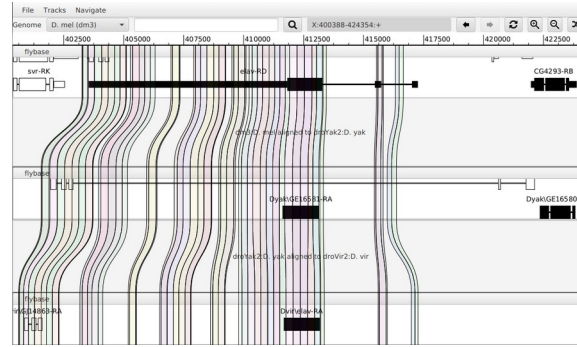
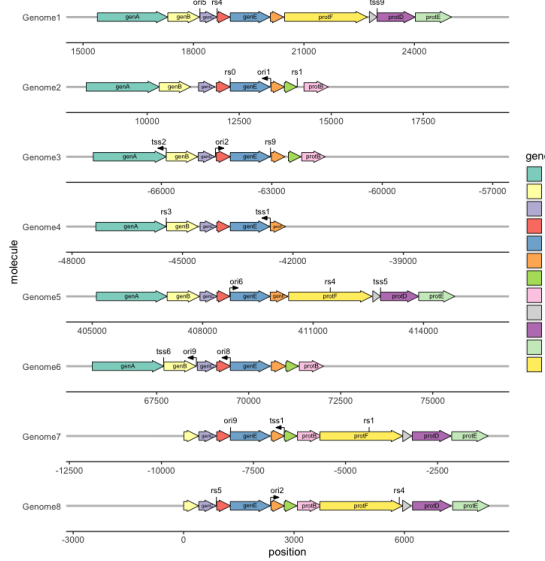
JBrowse2 comparative display of diploid to hexaploid genome

- Ar1D
  - Ar2D
  - Ar3D
  - Ar4D
  - Ar5D
  - Ar6D
  - Ar7D
  - Tm1A
  - Tm2A
  - Tm3A
  - Tm4A
  - Tm5A
  - Tm6A
  - Tm7A
- [Deselect all](#)   [Deselect all](#)

Data for Accusyn Display



# Other Viz:



# Is there a need for a Pan-Genome WG or should these be other WG sub-topics?

## Recent 6-mo assessment

### Challenges:

Mainly TIME!

Members are motivated, but other time commitments are limiting.

### Needed Interactions:

Standards for Genetic Variation.

Genome Assembly and Annotation Nomenclature.

### Goals in the queue:

Developing a web resource.

- includes catalog of visualization and analysis software and techniques.
- file formats.
- help getting started with pan-genomes.
- term definitions.