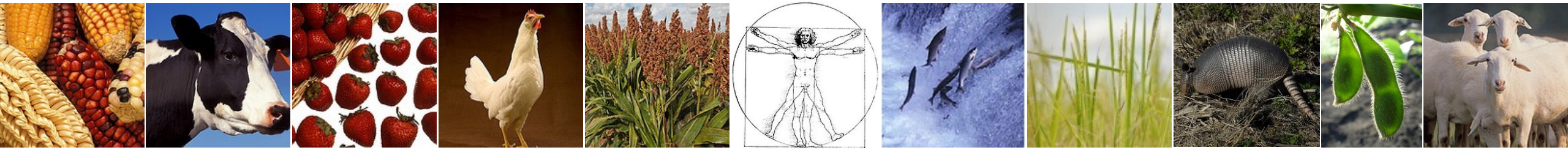


Genome Annotation of Arabidopsis thaliana T2T Col-CC By NCBI RefSeq

Terence D. Murphy – RefSeq Eukaryotes and CGR Project Team Lead

NIH National Library of Medicine
National Center for Biotechnology Information

Outline

- The CGR project and NCBI tools for Arabidopsis
- NCBI annotation of Col-CC
- What's Next

NCBI Genome Resources Workshop

Monday January 15, 2024, 4:00 – 6:10 pm, Palm 8

Time	Topic
4:00 – 4:20	SRA: Submission Improvements and Accelerating Discoveries through Lighter Data *Yuriy Skripchenko*
4:20 – 4:40	Lean, Mean, Genome Cleaning Machine: Removing Contamination with FCS-GX *Eric Tvedte*
4:40 – 5:00	GenBank Submission: How Do I Submit Diploid Genome Assemblies to NCBI? *Shelby Bidwell*
5:00 – 5:20	Advances in Eukaryotic Annotation at NCBI *Vamsi Kodali*
5:20 – 5:40	Using NCBI Datasets to Easily Gather Data from across NCBI Databases *Sally Chang*
5:40 – 6:00	The NIH Comparative Genomics Resource (CGR) *Terence Murphy*

Visit NCBI Booth **615**

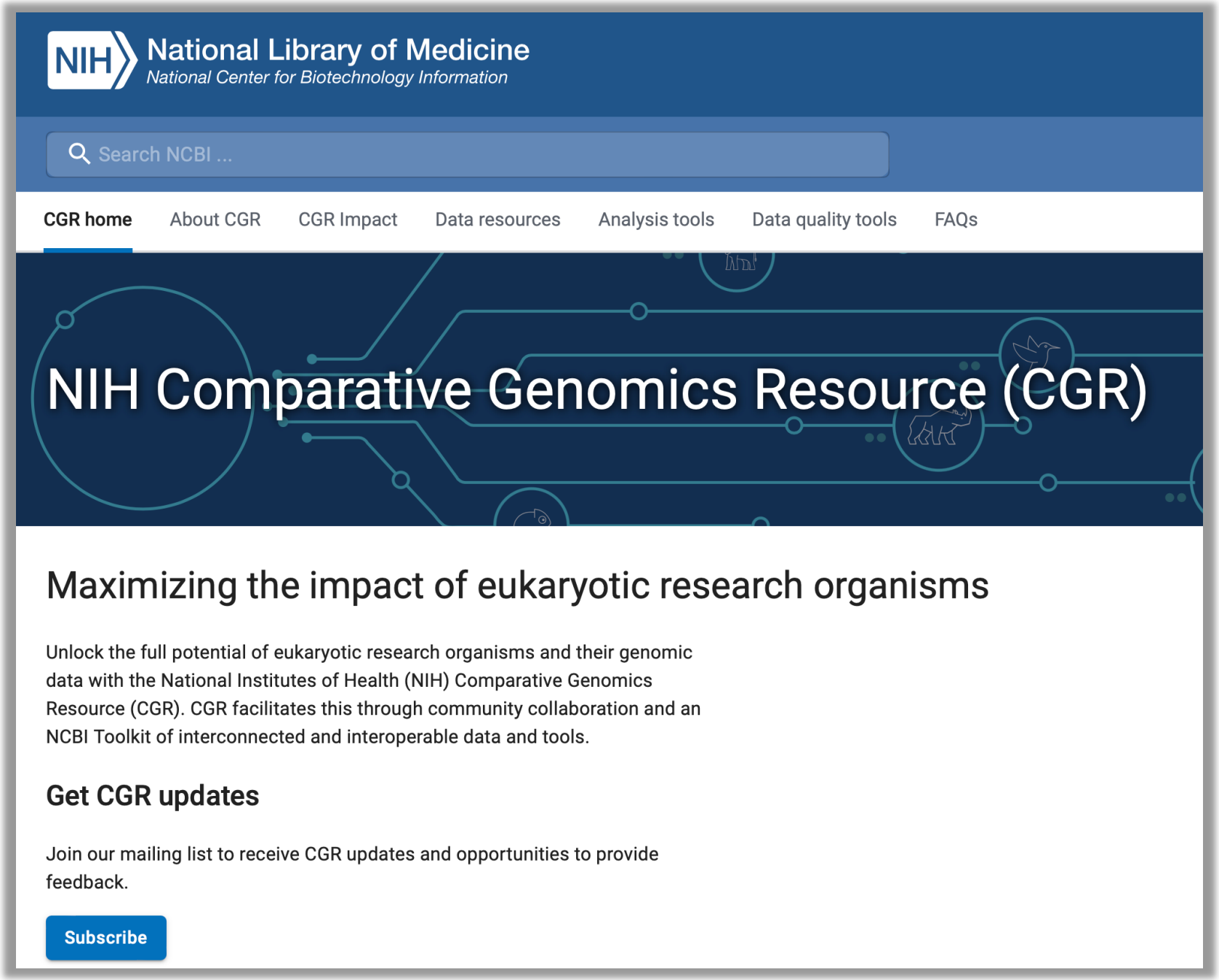
Contact us **info@ncbi.nlm.nih.gov**

Watch NCBI News for updates!

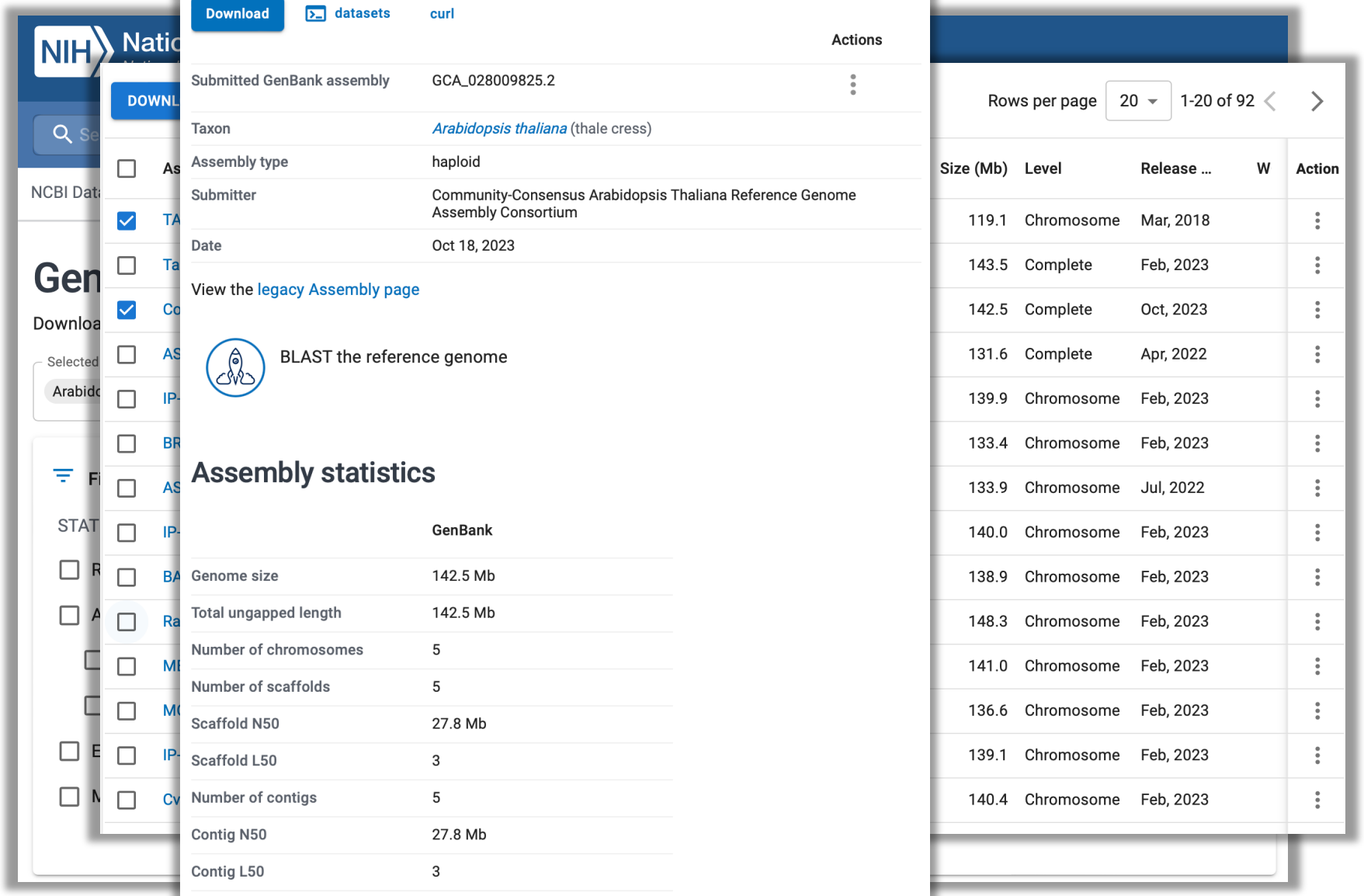
http://www.ncbi.nlm.nih.gov/news/

https://www.youtube.com/user/NCBINLM

NCBI


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
CGR home
About CGR
CGR Impact
Data resources
Analysis tools
Data quality tools
FAQs
NIH Comparative Genomics Resource (CGR)
Maximizing the impact of eukaryotic research organisms
Unlock the full potential of eukaryotic research organisms and their genomic data with the National Institutes of Health (NIH) Comparative Genomics Resource (CGR). CGR facilitates this through community collaboration and an NCBI Toolkit of interconnected and interoperable data and tools.
Get CGR updates
Join our mailing list to receive CGR updates and opportunities to provide feedback.
Subscribe

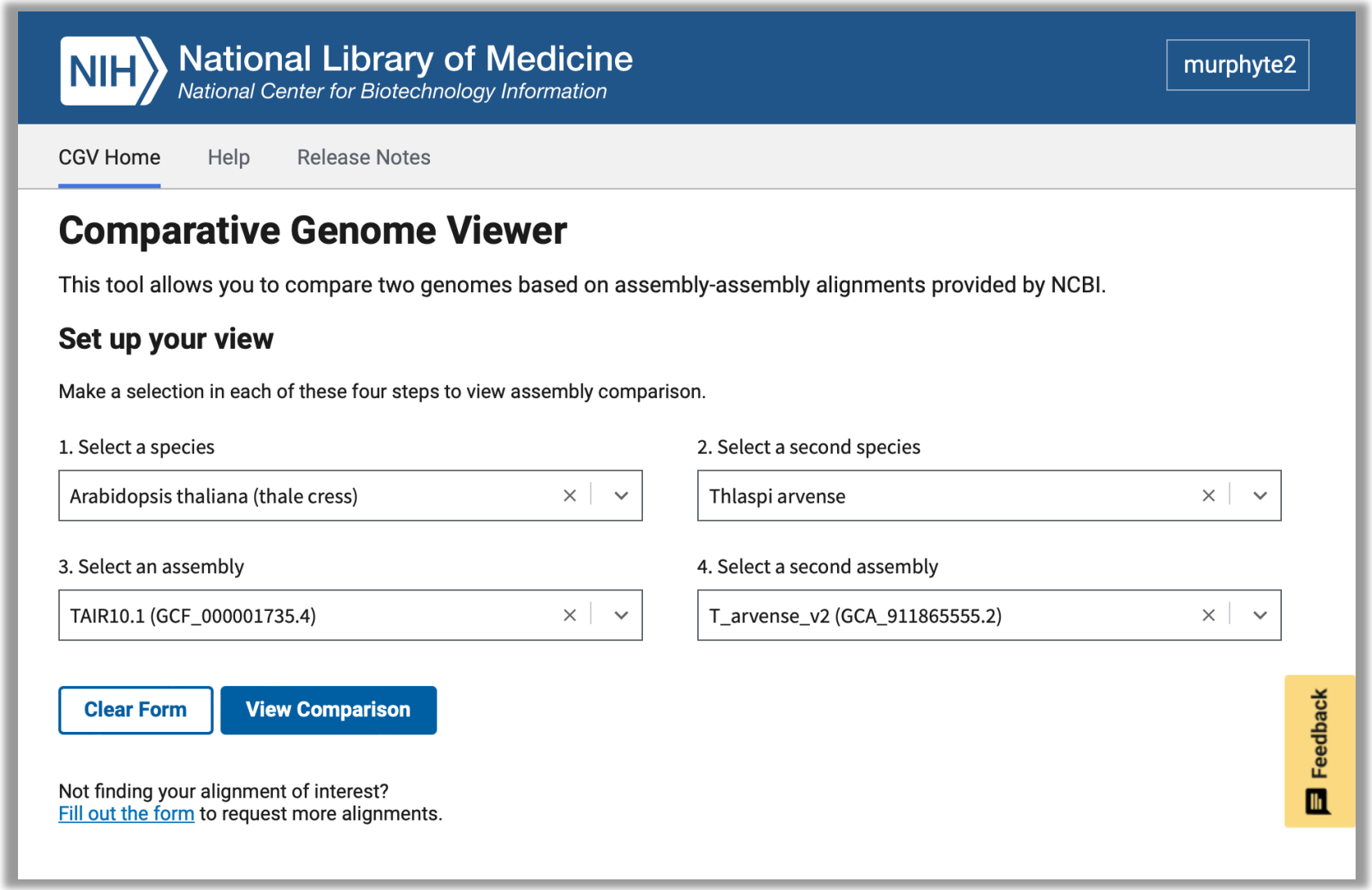
https://www.nc asets



Comparative Genome Viewer (CGV)

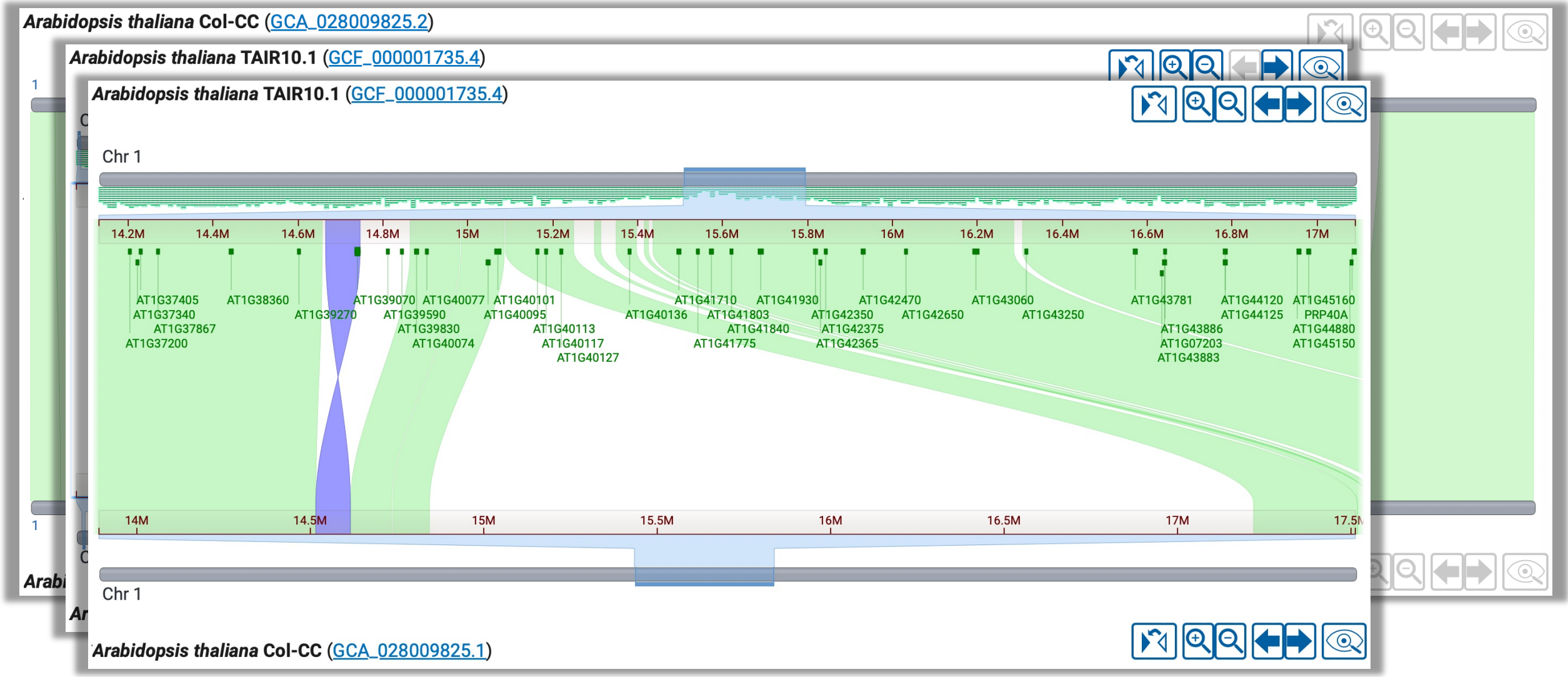
- Compare genomes through assembly alignments
- Over 700 alignments for over 300 species!
- Preprint: PMC10705539

https://www.ncbi.nlm.nih.gov/genome/cgv/



CGV: TAIR10 vs Col-CC

First Pass	Total
Col-CC (Current) Coverage: 88.74%	Col-CC (Current) Coverage: 89.58%
TAIR10.1 (Previous) Coverage: 99.03%	TAIR10.1 (Previous) Coverage: 99.32%
Percent Identity: 98.64%	Percent Identity: 98.61%



Outline

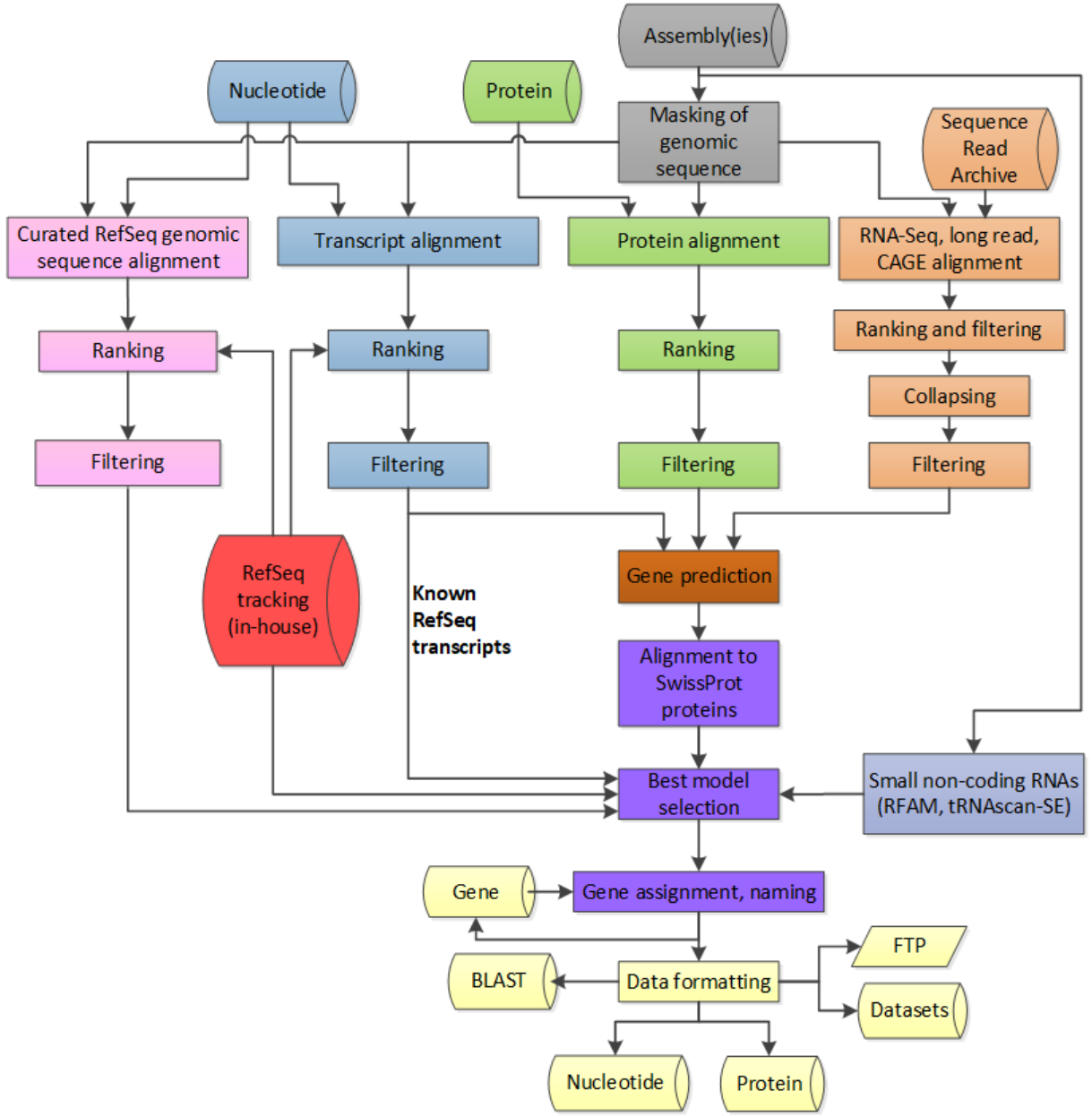
- The CGR project and NCBI tools for Arabidopsis
- NCBI annotation of Col-CC
- What's Next

EGAP: Eukaryotic Genome Annotation Pipeline

Used by NCBI to annotate ~1100 species

Evidence used for gene prediction:

- ✓ ESTs
- ✓ cDNAs
- ✓ Same and cross-species proteins
- ✓ RNA-Seq
- ✓ PacBio IsoSeq, ONT transcriptomes
- ✓ CAGE
- ✓ PhyloCSF (for targeted use)



NCBI EGAP annotation

- Software: EGAP v10.1
- Annotation Name: GCA_028009825.1–RS_2023_03 (internal release)

Input type	amount	notes
Oxford Nanopore	41 runs, 61M reads	~90% identity/coverage
PacBio	33 runs, 17M reads	99+% identity/coverage
Illumina	332 runs, 9.3B reads	Mostly unpaired, unstranded
CAGE	52 runs, 3.4M reads	63% aligned
Transcripts	195k cDNA + 1.5M ESTs	99+% identity/coverage
Proteins	• TAIR A. thaliana annotation • Subset of Brassicaceae proteins	

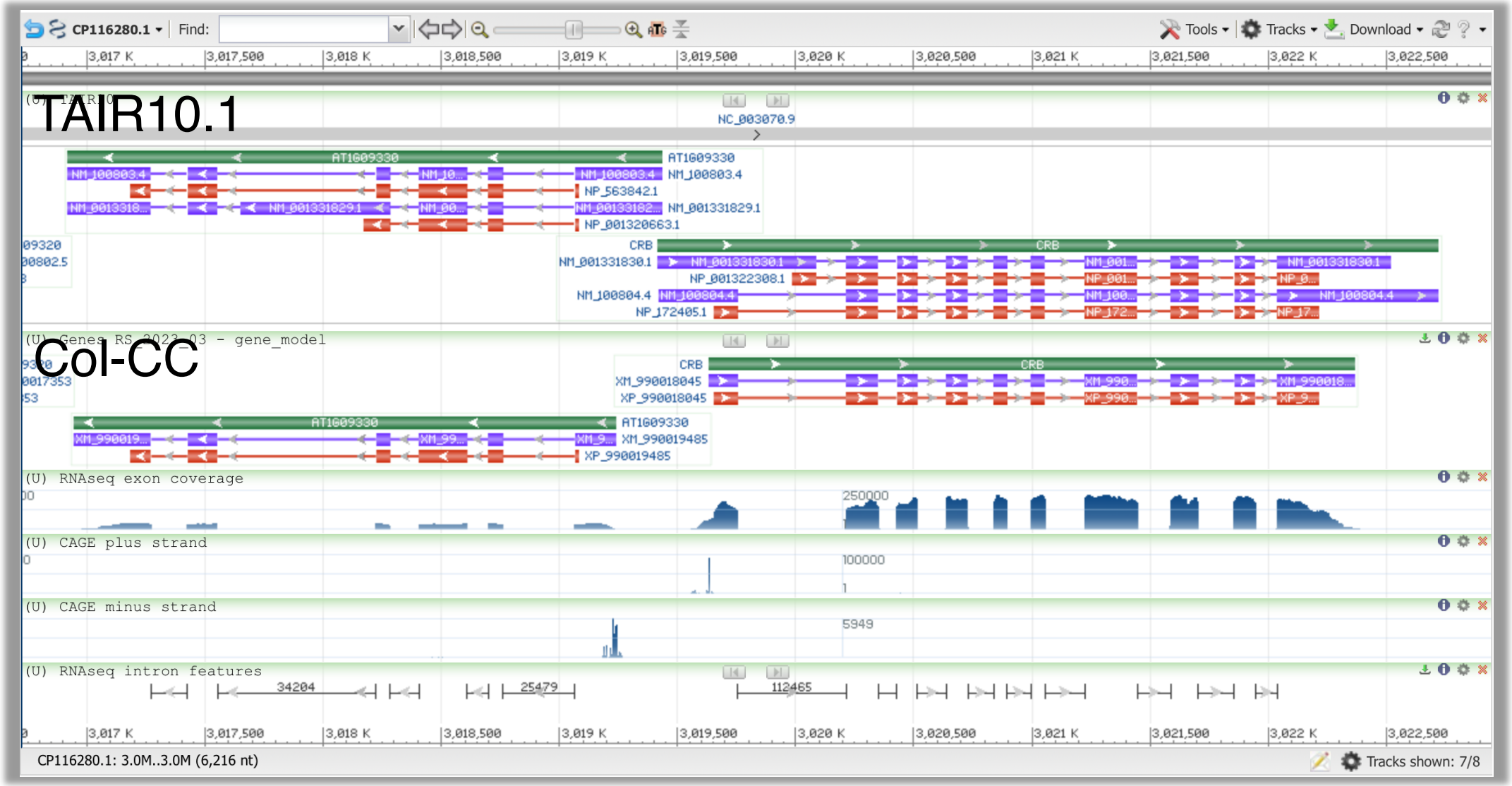
Gene counts

Gene biotype	source	Col-CC	TAIR10.1
Protein-coding	Gnomon	25,565	27,444
with alt transcripts		5,574	10,701
with major corrections		570	
Pseudogene	Gnomon	2,337	4,842
lncRNA	Gnomon	833	3,480
rRNA (v1 assembly)	Rfam	3441	4
snRNA	Rfam	62	82
snoRNA	Rfam	578	287
tRNA	tRNA-scan	545	625
miRNA	previous	325	325
Antisense	previous	79	79

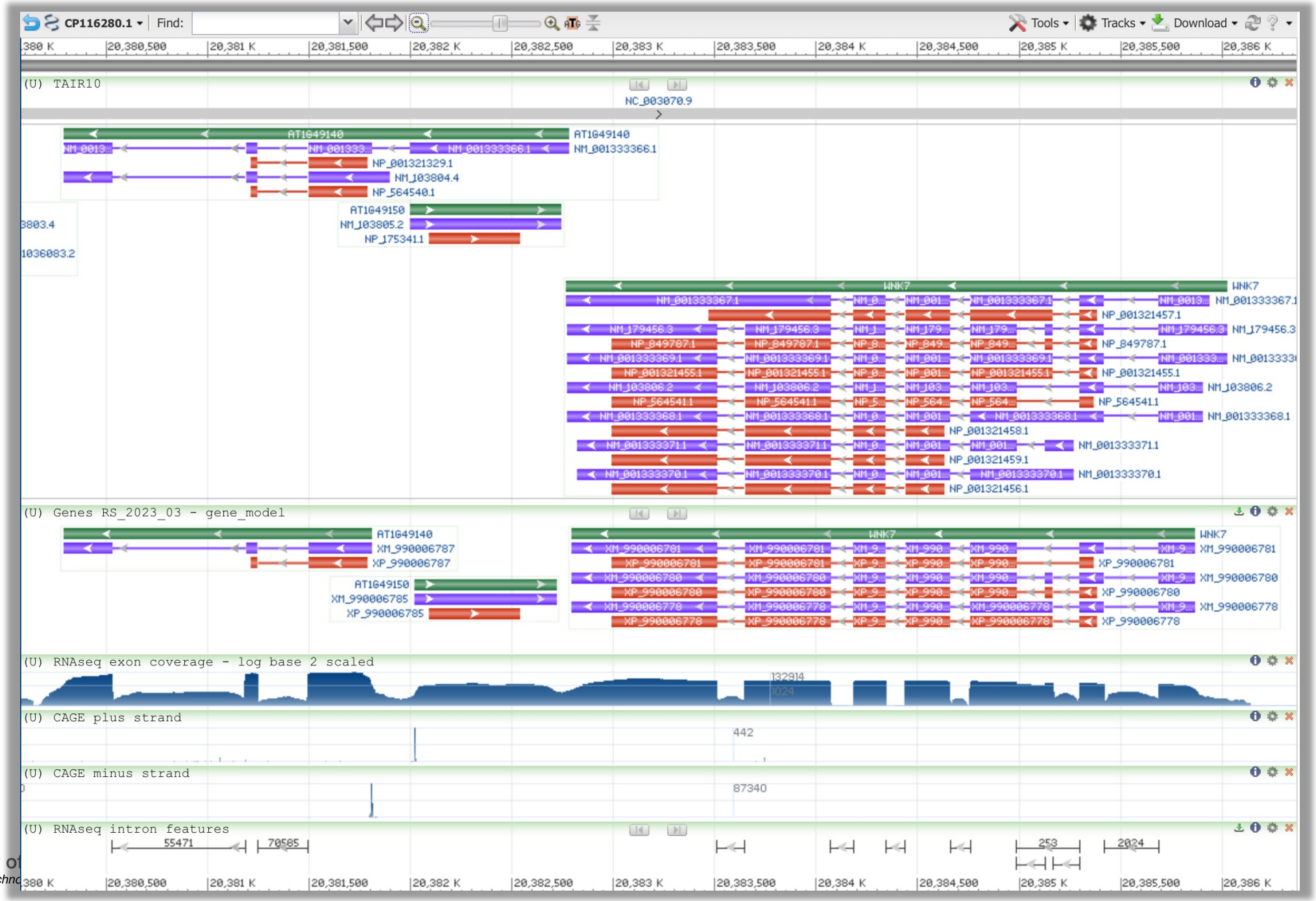
BUSCO:

- TAIR10.1: C:99.8%(S:98.7%,D:1.1%),F:0%,M:0.2%
- RS_2023_03: C:98.8%(S:97.7%,D:1.2%),F:0%,**M1.1%**

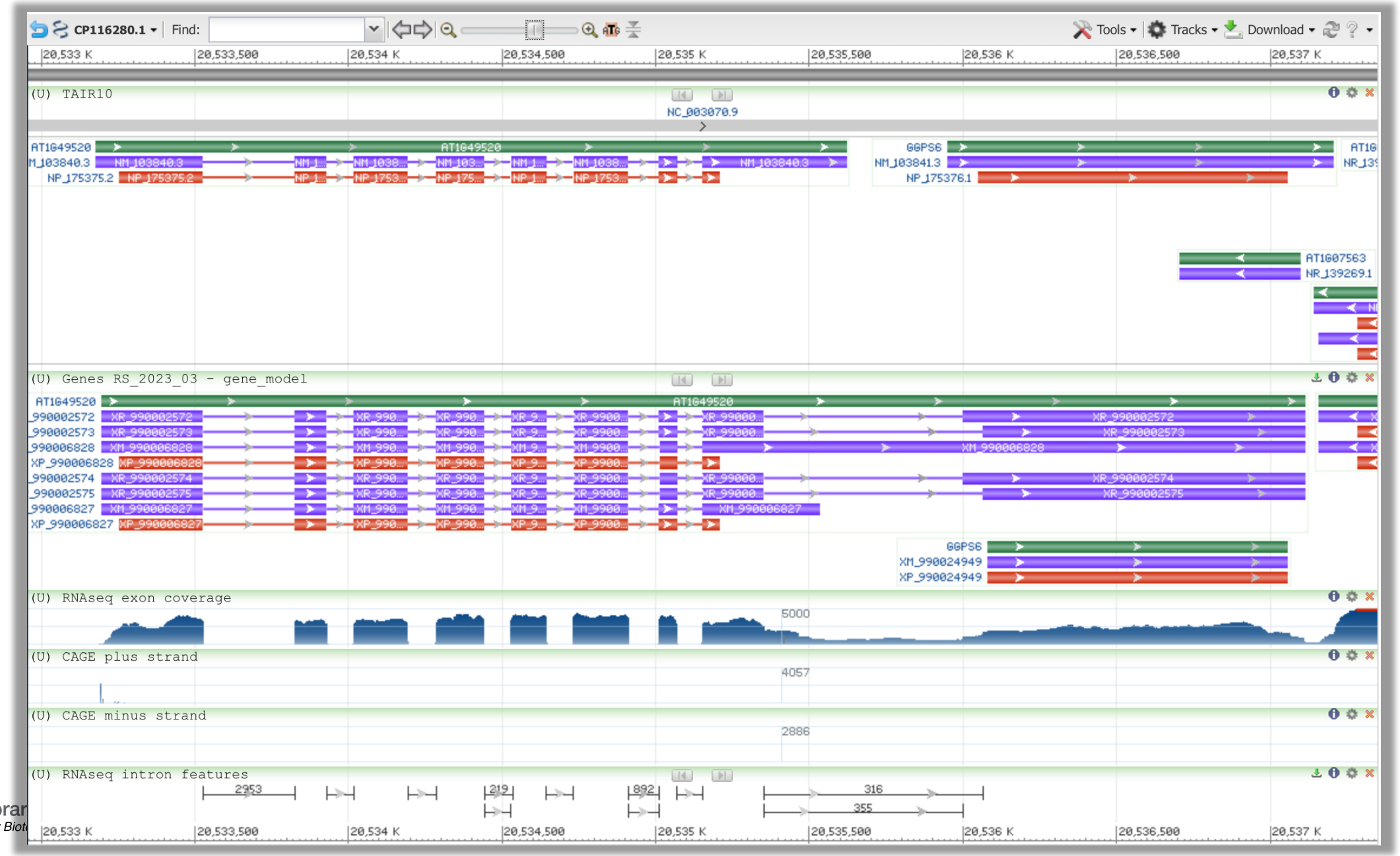
Chr1: AT1G09330 and CRB



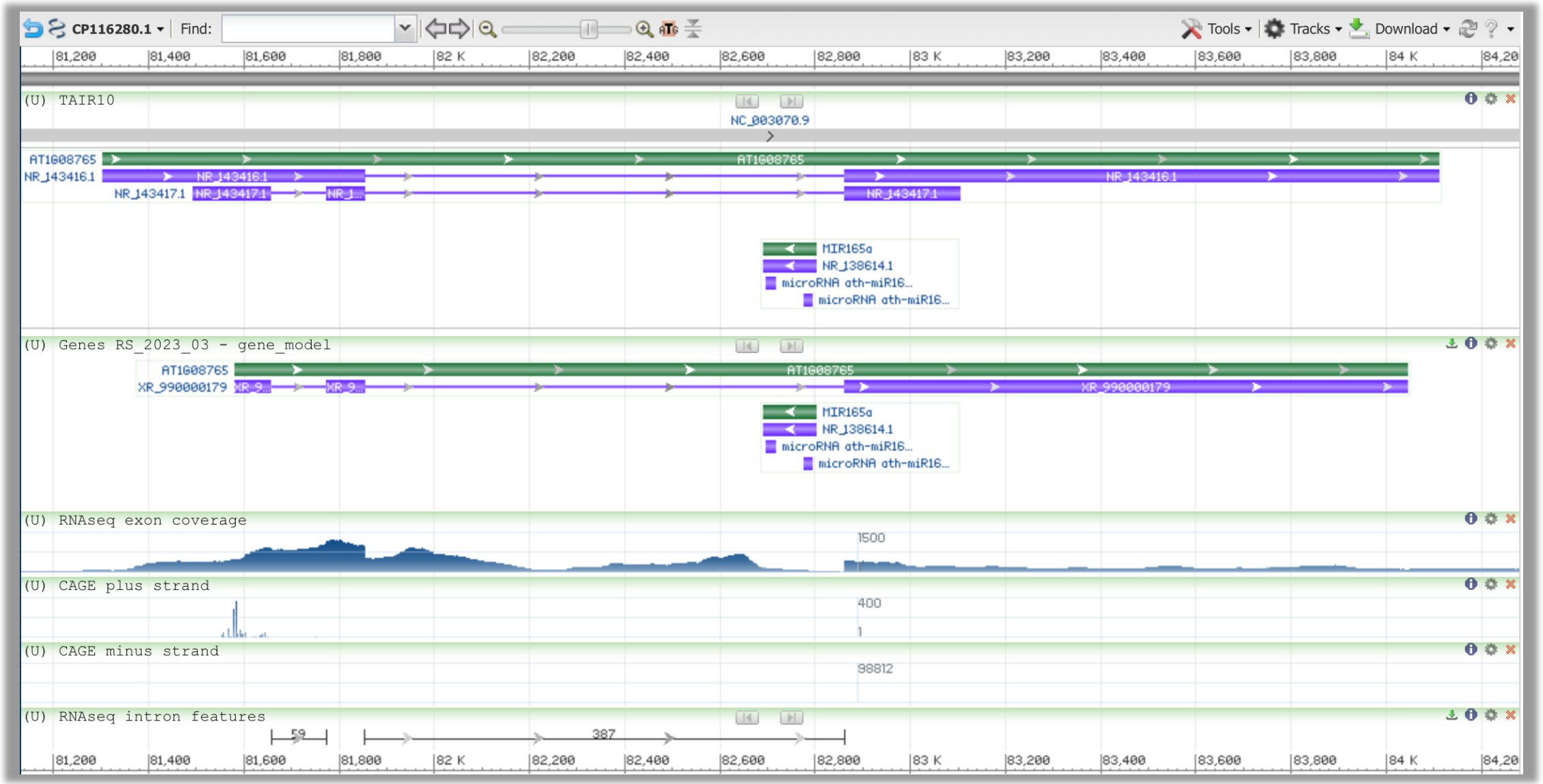
WNK7



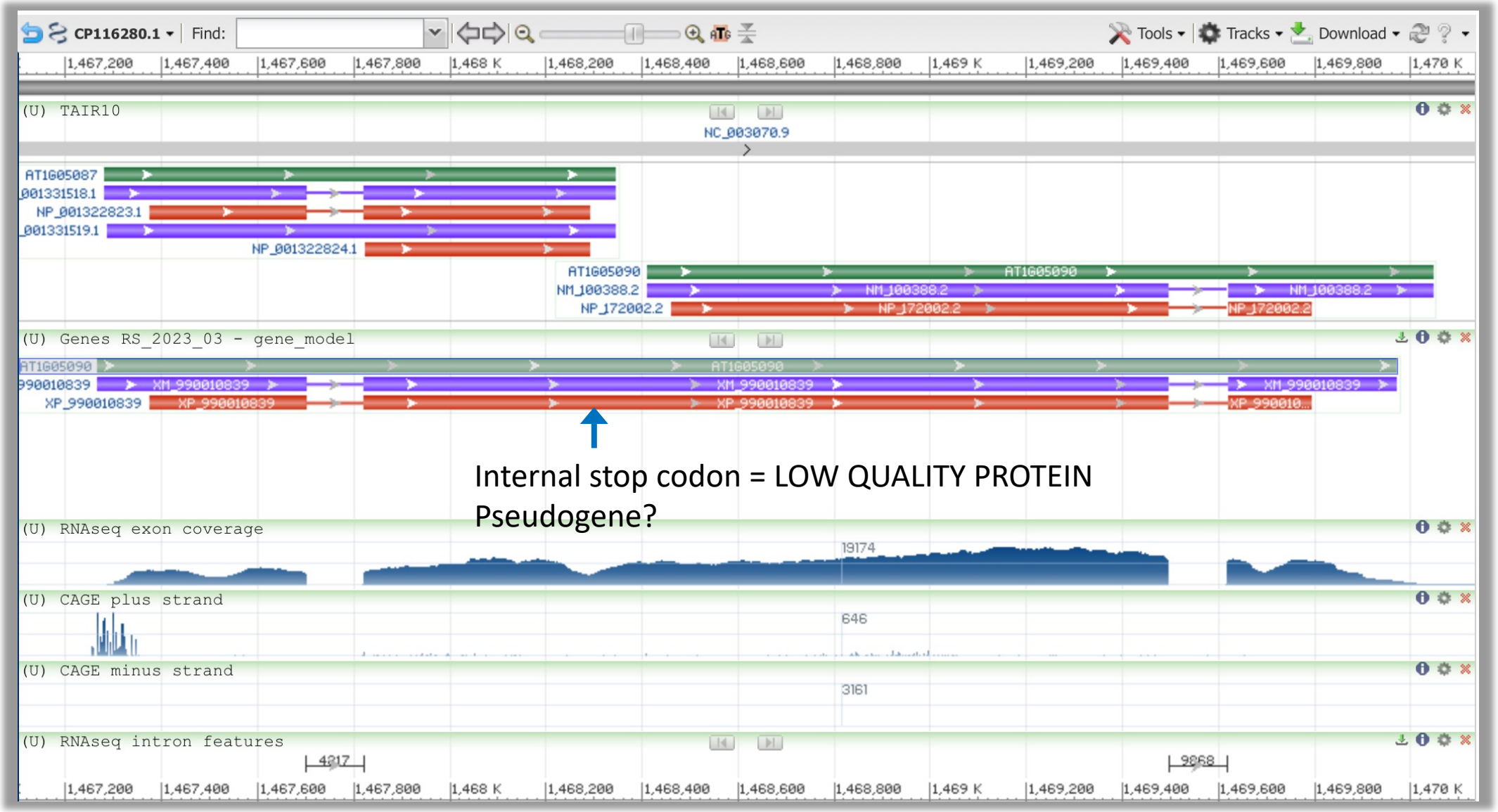
AT1G49520 and GGPS6



AT1G08765 and MIR165a



AT1G05087 and AT1G05090



What's Next

- Review and refinement
  - Major differences vs TAIR10.1
  - Low quality proteins
- Merge other annotations and shift to adjust for added NORs
- Validation and GenBank submission
- Adoption:
  - Incorporate into NCBI RefSeq
  - Switch to Col-CC in TAIR (TAIR11?)
  - Add additional data (expression, variation, etc) to the new reference
- Get back to the real science!

Thank you

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

RefSeq/Gene

Shashi Pujar
Tamara Goldfarb
Diana Haddad
John Jackson
Vinita Joardar
Kelly McGarvey
Michael Murphy
Barbara Robbertse
Brian Smith-White
Pooja Strope
Anjana Vatsan
David Webb

EGAP

Francoise Thibaud-Nissen
Wratko Hlavina
Avi Kimchi
Jinna Hoffman
Vamsi Kodali
Patrick Masterson
Eyal Mozes
Robert Smith
Alexandre Souvorov
Olga Ermolaeva
Craig Wallin

EGAPx / FCS

Alex Astashyn
Nathan Bouk
Victor Joukov
Deacon Sweeney
Eric Tvedte
Victor Sapojnikov
Pooja Strope
Pape Sylla
Lukas Wagner

CGR

Nuala O'Leary
Sanjida Rangwala
Tom Madden
Aron Marchler-Bauer
Anne Ketter
Katya Sukharnikov

NCBI Leadership

Valerie Schnieder
Kim Pruitt
Steve Sherry

Visit NCBI Booth **615**

Contact us **info@ncbi.nlm.nih.gov**

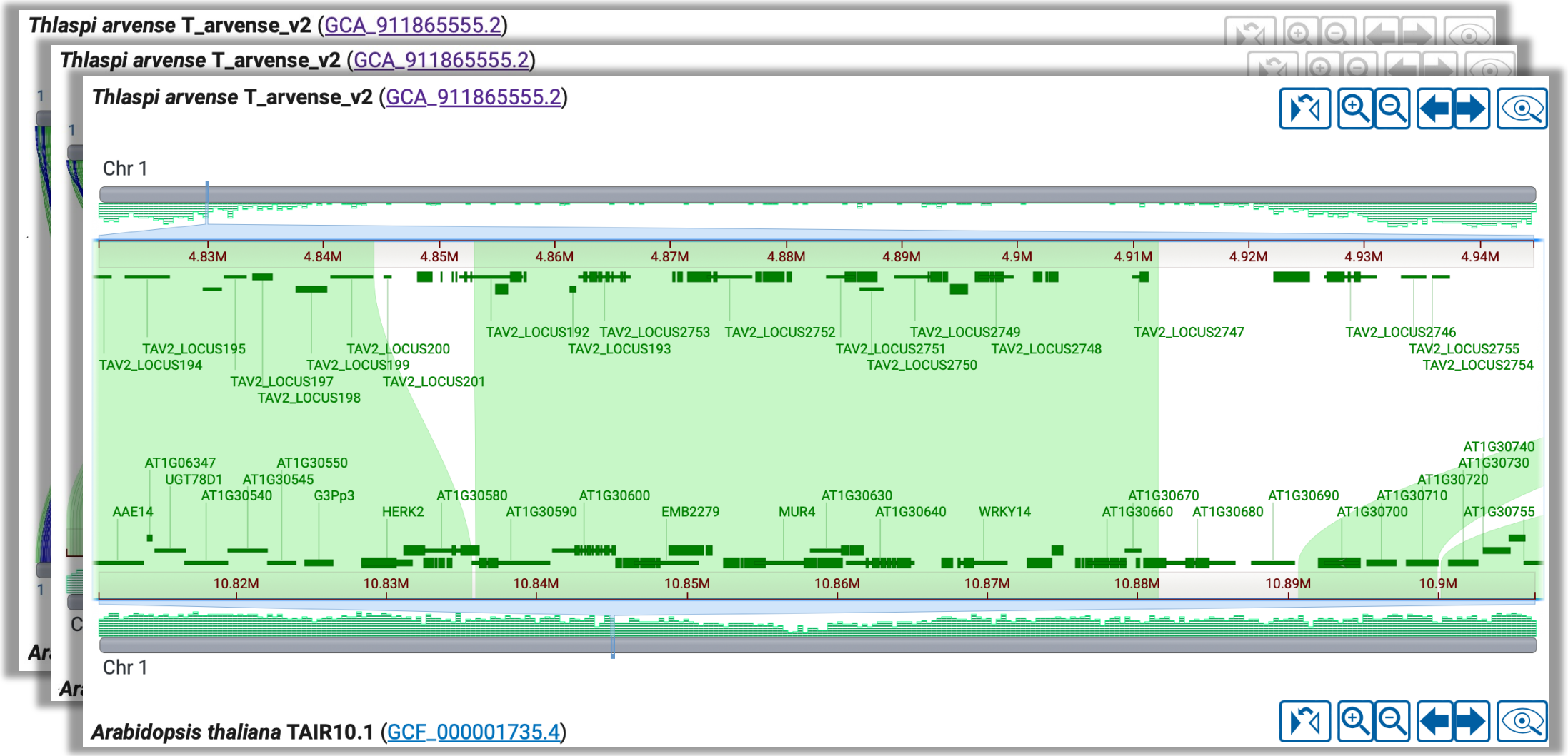
Watch NCBI News for updates!

http://www.ncbi.nlm.nih.gov/news/
https://www.youtube.com/user/NCBINLM

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

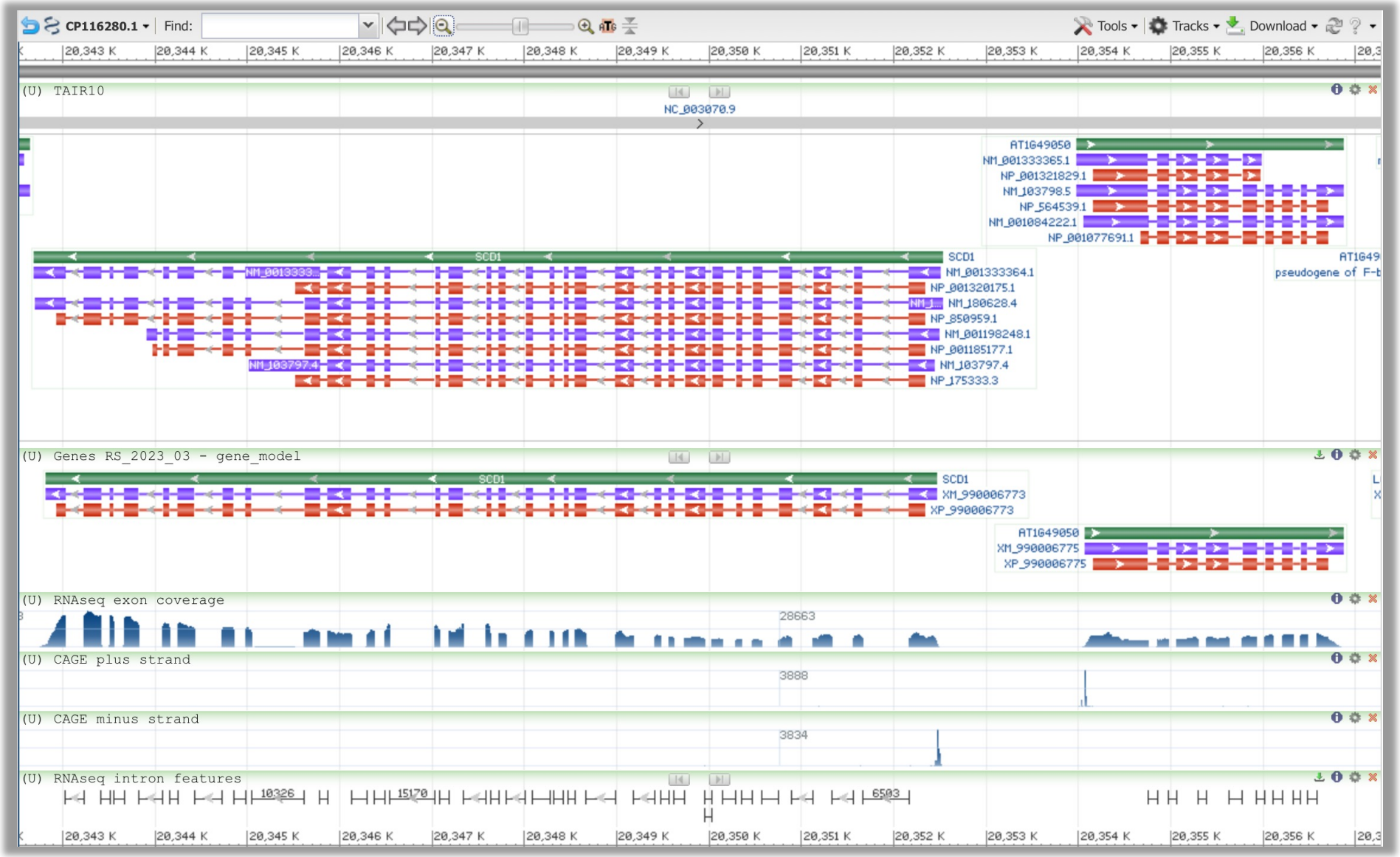
NCBI

CGV: Arabidopsis vs Thlaspi

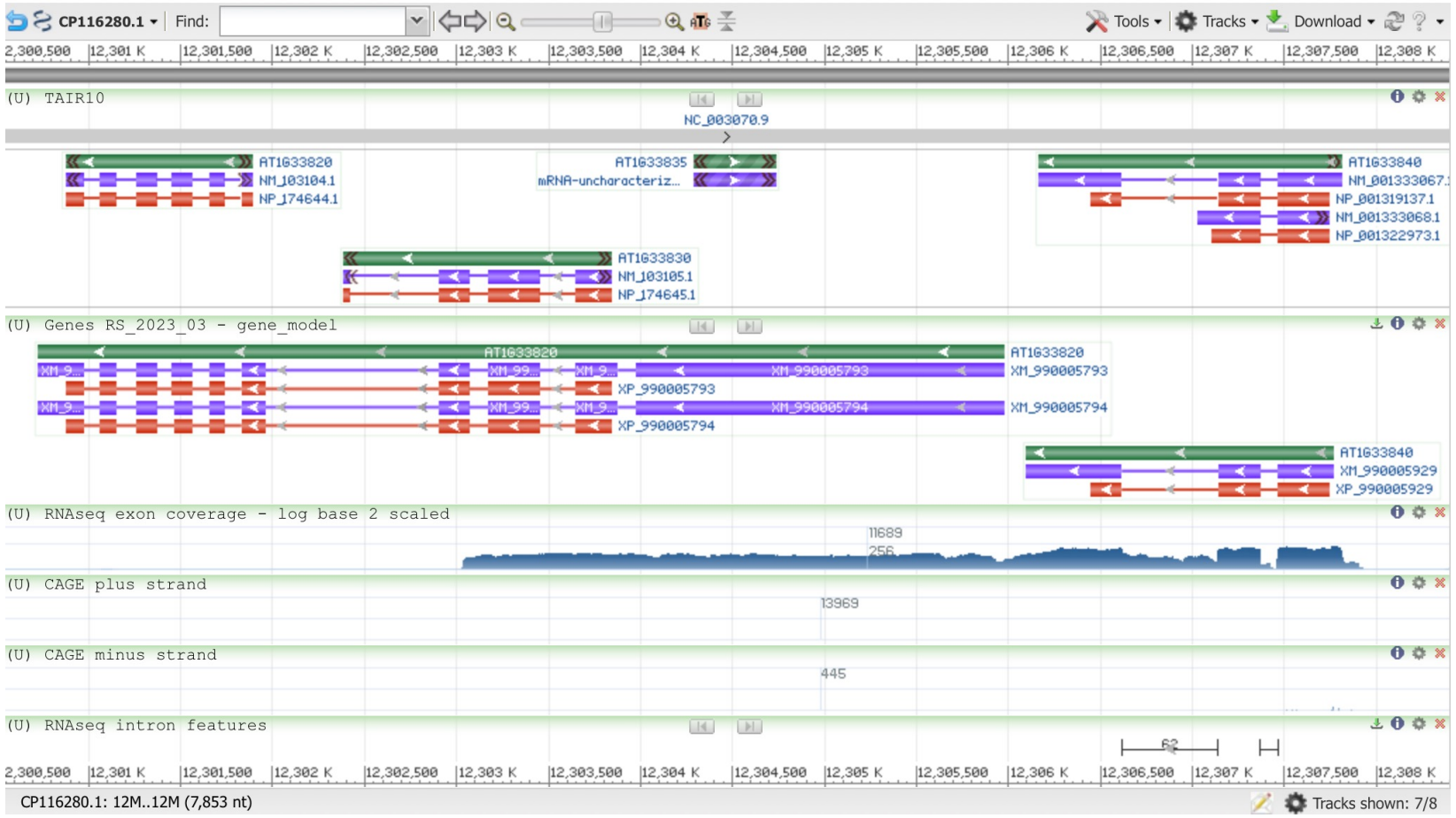


Chr1: AT1G48870

Chr1: SCD1 and AT1G49050



AT1G33820 and AT1G33840



AT1G36920 and AT1G36925