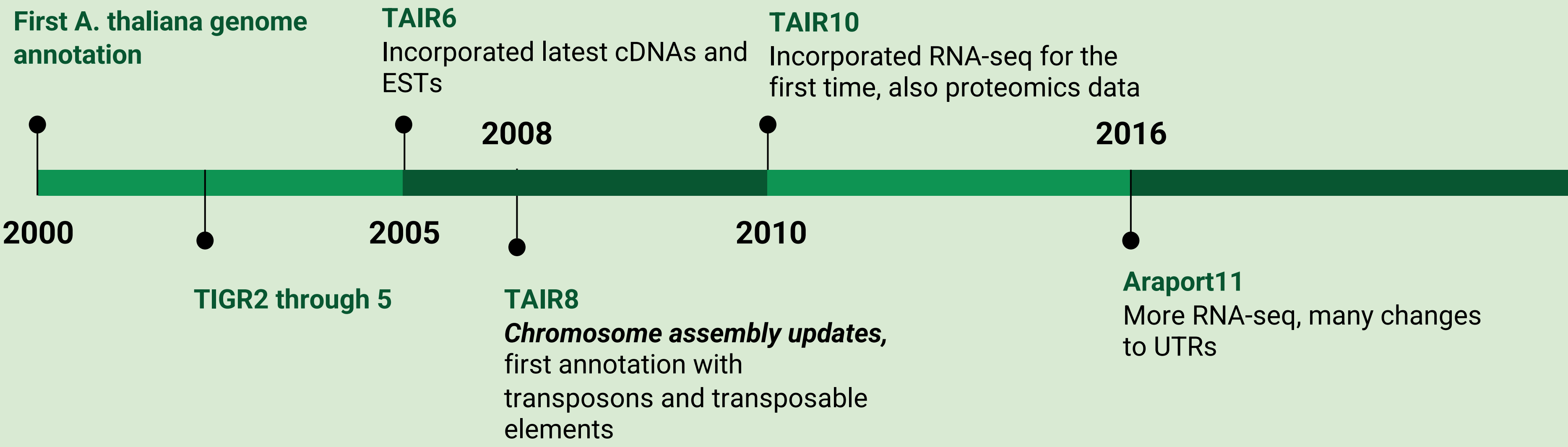
ICAR 2023: A. thaliana genome annotation v.12 update

A community effort coordinated by TAIR

tinyurl.com/Athalianav12

Timeline



2000

First A. thaliana genome annotation

TIGR2 through 5

2005

TAIR6
Incorporated latest cDNAs and ESTs

2008

TAIR8
Chromosome assembly updates, first annotation with transposons and transposable elements

2010

TAIR10
Incorporated RNA-seq for the first time, also proteomics data

2016

Araport11
More RNA-seq, many changes to UTRs

tinyurl.com/Athalianav12

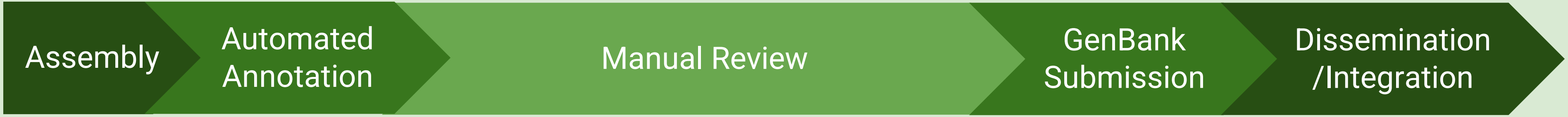
What has changed over the past 20 years?

- Greater amounts of supporting data
- Increased kinds of supporting data
- Improved sequencing technology, longer reads
- Improved genome assembly software
- Improved automated annotation pipelines

An improved version is PAST DUE

tinyurl.com/Athalianav12

Updating the reference genome



Assembly

Automated Annotation

Manual Review

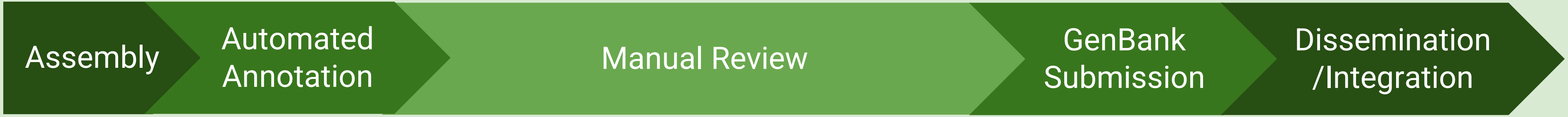
GenBank Submission

Dissemination /Integration

Past releases: specific grant funding for the entire process

tinyurl.com/Athalianav12

Updating the reference genome

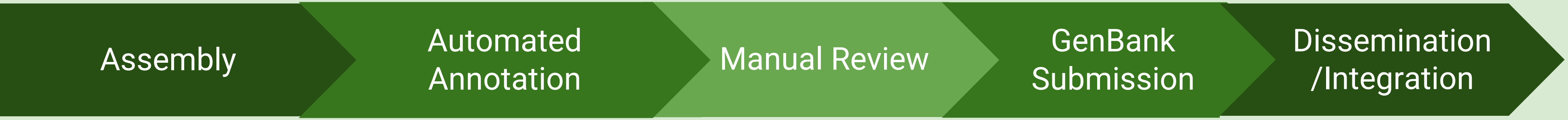


Assembly → Automated Annotation → Manual Review → GenBank Submission → Dissemination /Integration

V12: Community effort, contributions of expertise and computing resources

tinyurl.com/Athalianav12

Updating the reference genome



Who: Schneeberger lab (MPI)

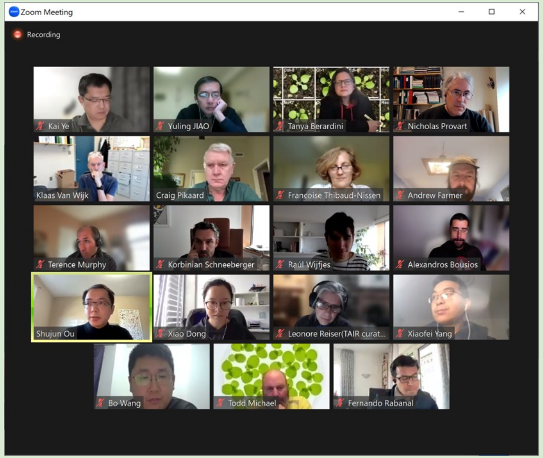
Who: NCBI (National Center for Bioinformatics)

Who: Community experts for review, TAIR for coordination and WebApollo hosting

Who: TAIR + NCBI

Who: BAR, TAIR, Ensembl Plants, NCGR GCV, AtPeptide Atlas, more

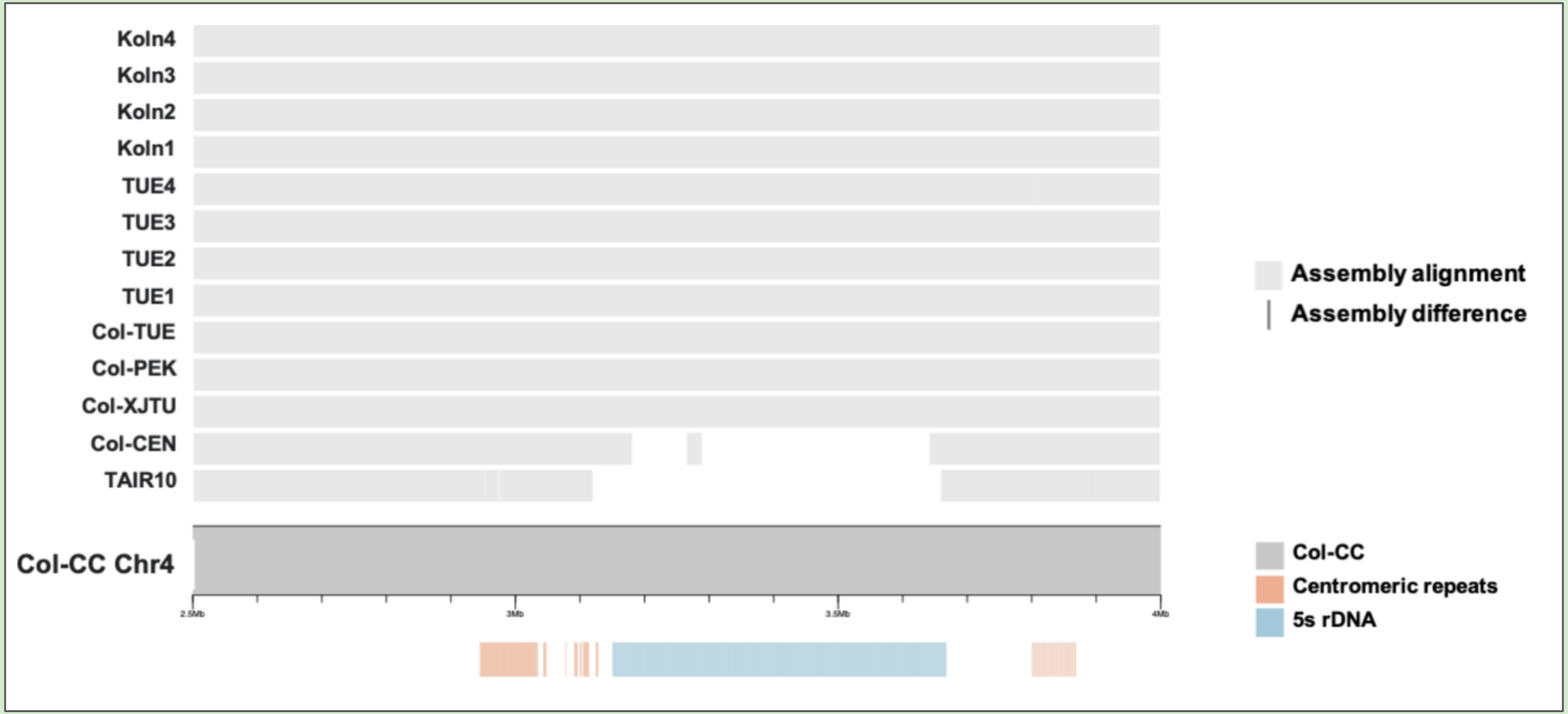
Community PUBLICATION



tinyurl.com/Athalianav12

Col-CC assembly (Xiao Dong,Raúl Y. Wijfjes, Korbinian Schneeberger)

- 13 Col-0 assemblies (Xiao Dong: P-686)



NCBI Eukaryotic Genome Annotation Pipeline

- Standard pipeline used for 1000s of genomes
- Highly dependent on experimental data
  - cDNAs
  - Proteins (100k Arabidopsis and Brassicaceae entries)
  - RNA-Seq (9.24 billion reads from ~20 tissues types)
  - IsoSeq, ONT (62 million long reads)
  - CAGE (3.94 billion reads from ~20 tissue types)

Automated Annotation Results: overview

- **No numbers because the results are NOT final**

- Fewer (!) total protein coding genes
- Fewer splice variants
- More rRNA genes

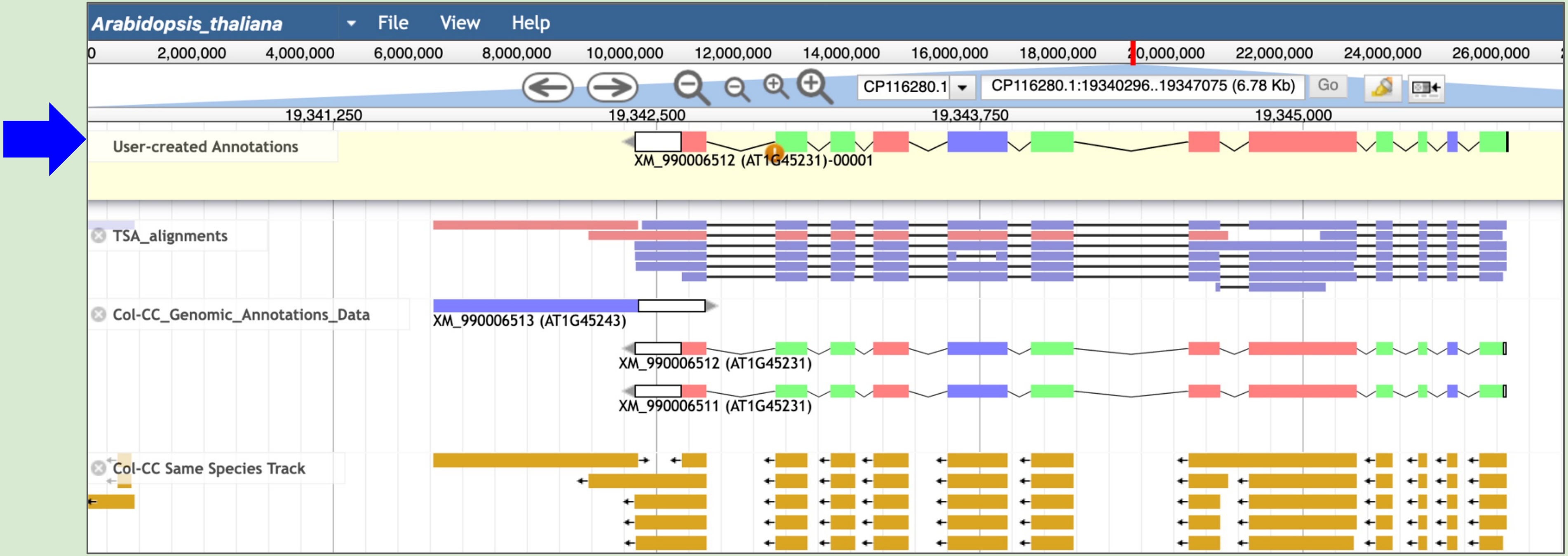
tinyurl.com/Athalianav12

To be reviewed: (numbers for protein coding genes only)

- Previous novel: 2085
- Split: 56
- Merged: 959
- Current novel: 61
- Changed locus type: 265
- BUSCO missing or fragmented: 56

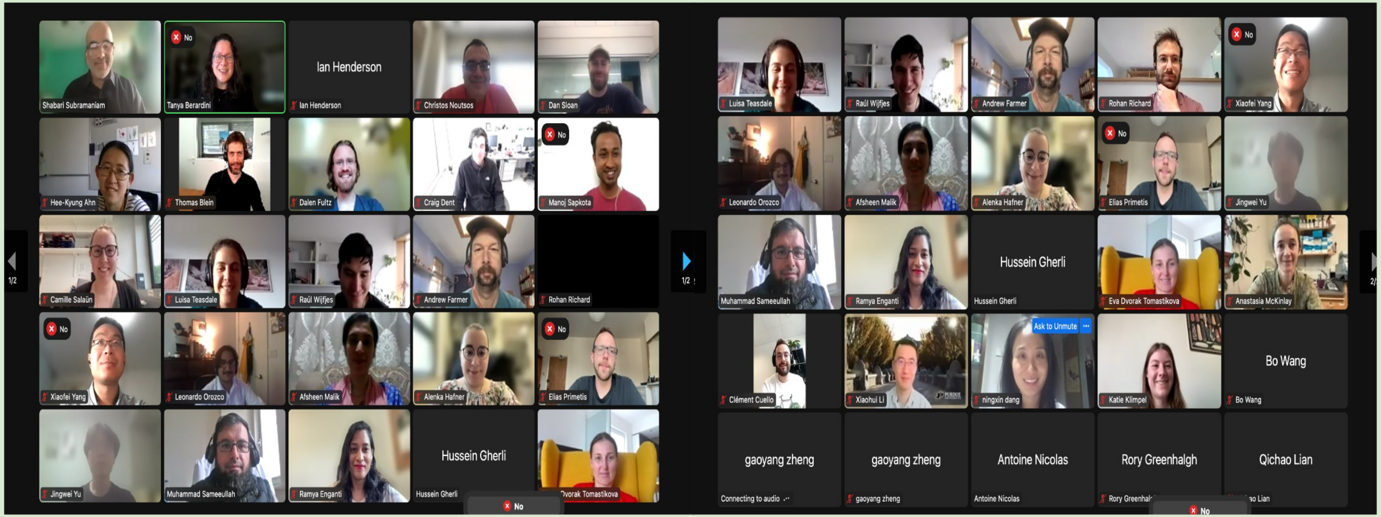
TAIR: community hub and coordination

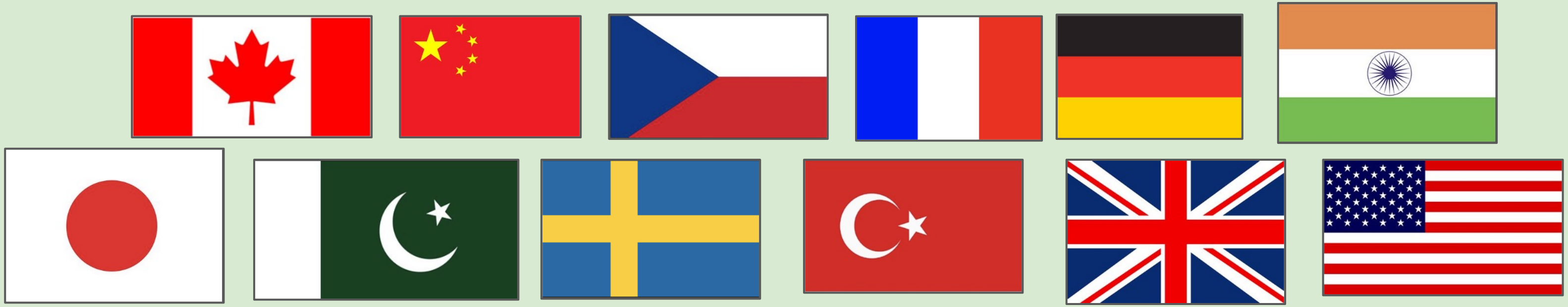
Apollo



tinyurl.com/Athalianav12

Manual review training: 4 x 90 min sessions





tinyurl.com/Athalianav12

Tandem repeat annotation: Ian Henderson, Piotr Wlodzimierz (University of Cambridge)

- analysis was done using TRASH (https://github.com/vlothec/TRASH) with sequence templates of telomeric (TTTAGGG), *AthCEN159*, *AthCEN178* and *5s* rDNA to classify major repeat classes of repeats under 1 kbp
- Followed by manual review and revision

Transposable Element reannotation: Alex Bousios (University of Sussex), Shujun Ou (Ohio State University), Zhigui Bao (Max Planck, Tübingen)

- Use a curated repeat library to capture and preserve known *Arabidopsis thaliana* TE families.
- Combine start-of-the-art tools including EDTA, RepeatMasker, ATHILAfinder, TEsorter, AnnoSINE, TRF, and SRF.
- Lift-off TE loci with functional support (e.g., *ONSEN*, *Athila*)

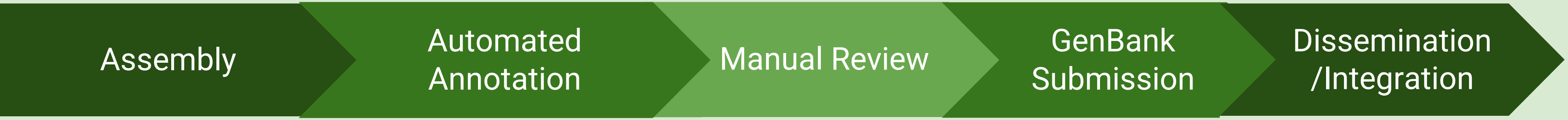
rDNA annotation - Ramya Enganti, Craig Pikaard (Indiana University); Ian Henderson, Piotr Wlodzimierz (University of Cambridge)

- Connected at this ICAR!

Bioinformatics Support: Kai Ye, Xiaofei Yang, Bo Wang (Xi'an Jiaotong University); Yuling Jiao (Peking University)

- Liftoff: Mapping of Araport11 genes onto Col-CC assembly coordinates
- Additional track for Apollo, manual review
- More to come

Updating the reference genome



Assembly

Who: Schneeberger lab (MPI)

Automated Annotation

Who: NCBI (National Center for Bioinformatics)

Manual Review

Who: Community experts for review, TAIR for coordination and WebApollo hosting

GenBank Submission

Who: TAIR + NCBI

Dissemination /Integration

Who: BAR, TAIR, Ensembl Plants, NCGR GCV, AtPeptide Atlas, more

Community PUBLICATION

tinyurl.com/Athalianav12

Stay up to date

Website: https://tinyurl.com/AthalianaV12

Social media:

- Mastodon: @TAIR@genomic.social
- Twitter: @tair_news

In-person updates at:

- Plant Biology 2023
- PAG 2024

Col-CC Assembly: Korbinian Schneeberger and lab team
NCBI Eukaryotic Genome Pipeline: Françoise Thibaud-Nissen, Terence Murphy
Apollo setup @TAIR: Shabari Subramaniam, Xingguo Chen, Trilok Prithvi, Chris Childers
Training materials: Moni Munoz Torres, Marcela Tello Ruiz, Monica Poelchau, Jason Williams
The wider Arabidopsis community
YOU

Manual Review: Join our volunteer team effort!

Need: subject matter expertise, interest, and attention to detail

What would you be signing up for?

- 2 x 90 min WebApollo training sessions
- Reviewing a set of genes (clearly defined) within a set time period
- Slack channel
- Authorship in the reannotation paper

tinyurl.com/Athalianav12