2022-10-27: Meeting agenda and summary

Date

27 Oct 2022

Attendees

Name	Relevant Expertise (for this effort)	Institution	Country
Nicholas Provart	community resource, sequence analysis and visualization tools	BAR/University of Toronto	Canada
Yuling Jiao	genome sequencing and assembly	Peking University	China
Bo Wang, Xiaofei Yang, Kai Ye	genome sequencing and assembly, centromere genetics	Xi'an Jiaotong University	China
Korbinian Schneeberger, Raúl Wijfjes, Xiao Dong	genome sequencing and assembly	Ludwig Maximilian University of Munich, Max Planck Institute for Plant Breeding Research	Germany
Fernando Rabanal	genome sequencing and assembly, pancentromere characterization	Max Planck Institute for Biology	Germany
Alexandros Bousios	transposable element annotation	University of Sussex	UK
Klaas Van Wijk	peptide atlas	Cornell University	USA
Craig Pikaard	rRNAs, NOR sequencing and assembly	Indiana University	USA
Michael Schatz	genome sequencing, assembly, and annotation	Johns Hopkins University	USA
Terence Murphy, Francoise Thibaud-Nissen, Anjana Raina	genome annotation pipelines	NCBI	USA
Andrew Farmer	comparative genomics and visualization	NCGR	USA
Shujun Ou	transposable element annotation	Ohio State University	USA
Todd Michael	genome sequencing, assembling plant genomes	Salk Institute	USA
Tanya Berardini, Leonore Reiser	community resource, genome annotation	TAIR/Phoenix Bioinformatics	USA

Goals

- Get all participants on the same page, provide background and impetus for this project

Agenda

a. Introductions: name, institution, interest in this effort, relevant expertise (15 mins)
b. Tanya - very brief history, overview of current motivation, TAIR's efforts since Araport11 release (10 mins)
c. Françoise/NCBI team member - overview of NCBI Eukaryotic Genome Annotation pipeline using the initial run with Naish T2T genome as example (15 mins)
d. Korbinian - overview of Col-CC (community consensus) assembly progress so far (15 mins)

- General discussion, aim to answer the following questions: (rest of time)

- Should we use the Col-CC assembly as the basis for the v12 annotation?
- If yes, is there anyone else, not currently included, who should be aware of or included in this process?
    - When is a reasonable date of completion?

- Can NCBI perform the automated annotation with their eukaryotic pipeline with that consensus assembly?

- Who can commit to participating in the manual review and update of the automated pass?
- Tool/s to use? Deployed where?
- Create list of participants, who else could we reach out to and involve in this part
- Dataset specific expertise? lncRNAs, TEs, protein-coding genes, etc
- TAIR can help in coordinating work to minimize overlap

- Who would handle submission to Genbank and how can we best prepare for a smooth submission?
- Schedule follow up meetings for subgroups (assembly, manual review, other)

Summary

General enthusiasm for the need and utility of a reannotation.

Proposed timeline: 12 calendar months to set up the framework, process, teams to get V12 released.

Funding: No dedicated, separately-sourced funding for any particular group at this time. Interested groups will contribute expertise and/or infrastructure.

1. Assembly
    a. need to work out details of tracking the metadata on BioSample provenance for the individual pieces
    a. K. Schneebeger's group's work on assembling a Col-Community Consensus (CC) assembly is likely to finish by the end of 2022, and will incorporate C. Pikaard's group's data on NOR2 and NOR4, 4 Col-0 MA lines from F. Rabanal/D. Weigel
    b. Col-CC should be submitted to NCBI as an independent assembly
    c. Idea to visualize the multiple individual assemblies that were combined to make Col-CC as a patchwork (GCV? other visualization tool?)
2. Automated Annotation
    a. NCBI will take the Col-CC assembly when accepted by NCBI and available and will run it through their eukaryotic annotation pipeline
    b. need to resolve details on whether or not to include the Araport11 proteins as evidence
    c. add isoSeq from PRJNA755474 from this paper to next run
    d. please send more recent isoSeq/RNA-seq/CAGE experimental data in GenBank to include in the next run
3. Manual Review
    a. TAIR to investigate hosting requirements/existing training tools, ease of output of information needed for NCBI submission even before manual review begins
    b. used by many MODs to maintain their genomes, concurrent editing possible, community maintained code
    a. TAIR as coordinator
    b. Klass van Wijk: anything to do with proteins (including small peptides - sORFs, etc) and protein isoforms (AS, etc)
    c. Kai Ye : We (XJTU team) would work on centromeres and microsatellite sites.
    d. Shujun Ou, Alex Bousios: TEs, ATHILAs
    e. Craig Pikaard: NOR2 and NOR4, rDNAs
    a. WebApollo as tool
    b. Community experts
4. Submission to NCBI/GenBank
    a. begin working on release early, no need to wait till manual review is done, can be done with dummy data to work out format issues
5. Dissemination
    a. broad support for authorship on V12 paper for ALL who were involved in effort, in any stage of the process
    b. V12 release to be incorporated into TAIR, BAR, etc as soon as possible after NCBI RefSeq is updated to this version

Action items

- [x] **We'll check in by email in mid-December to get an update from Korbinian and from TAIR on the assembly progress and WebApollo.**