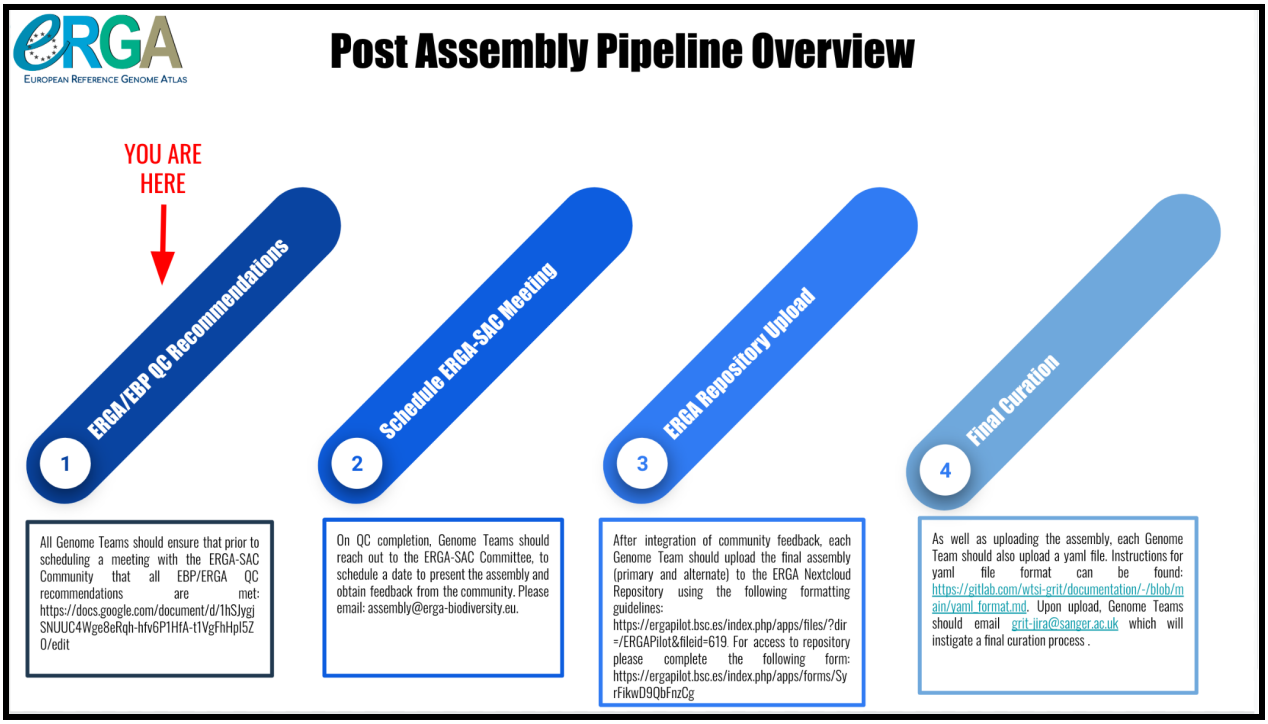
ERGA Pilot Project Post Assembly Quality Control Standards

Authors: Ann M. McCartney, Jonathan Wood, Kerstin Howe, Giulio Formenti
Version 1
Date: 06/30/2022

When the Genome Team has generated the initial assembly for their pilot species, the Team then enters into the “Post Assembly Pipeline”. It is recommended by the Pilot Project Committee that a standard QC assessment is carried out prior to scheduling a presentation at an ERGA Sequencing and Assembly Committee.



Standards

1. Run a decontamination software on your draft assembly using BlobToolKit, and NCBI’s Foreign Contaminant Screening FCS suite of tools - GX and adaptor.
2. Confirm unique single copy material representation using GENESCOPE.FK (or Genomescope2).
3. Re-map the Hi-C data, and review/edit scaffolds in PretextView (or JuiceBox, please note that Pretext maps are a requirement for final curation) to generate a set of large scaffolds that you trust to represent chromosomes. Some small contigs/scaffolds will

remain in the primary assembly, and can be submitted as unlocalised, but the goal is for these to represent less than 10% of the sequence.

4. Take a look at the assembly graph with BandageNG
5. Assign chromosome numbers to chromosomal scaffolds, using genetic data from the species if available to connect to established chromosome/linkage group nomenclature, or numerically in size order, taking into account any karyotype information that is available. If there is complete one-to-one orthology to the chromosome set of a closely related species with an established chromosome nomenclature, then it is acceptable to adopt that nomenclature. Identify and name the sex chromosomes if possible.
6. Estimate base pair accuracy using MERQURY.FK (or yak or Merqury).
7. Estimate single copy gene representation completeness with BUSCO.
8. Generate summary statistics with gfastats (or QUAST).
9. After presentation at SAC and a final assembly is production, Genome Teams should prepare the following files for final curation from the GRIT Curation Team:
    a. Pretext Hi-C image of final assembly;
    b. Fasta file of corresponding assembly; and
    c. AGP from the assembly showing underlying components.