Whole Genome Assembly Assessment and Pre-Finishing

Brown, Adam, Labutti, K., Young, S., Zimmer, A. and FitzGerald, M.
Broad Institute of MIT and Harvard, Cambridge, MA

Abstract:

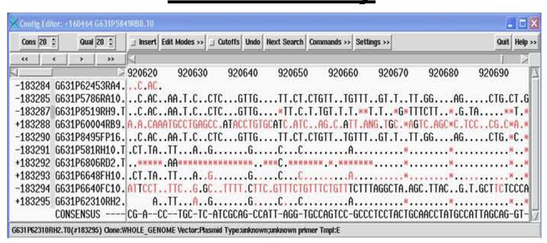
While informatics groups have improved the overall quality of genome assemblies as well as provided several useful tools for judging the integrity of such assemblies, visual inspection of these large assemblies is still necessary. During the inspection process several aspects of the assembly and genome itself can be noted, aiding the finishing process. Additionally, software tools may also be suggested as a result of this review process. An ARACHNE assembled database of the pathogenic fungus *Coccidioides immitis* was recently analyzed before the finishing process was started in order to determine misassembled and low quality regions that were missed by software tools, in addition to the difficulty of remaining gaps. By determining the signatures of regions that are missed or not counted by current software tools, new tools or revisions of old tools can be made to more fully automate the assessment process, and therefore further simplifying the finishing process as a whole.

Do we need manual assembly review?

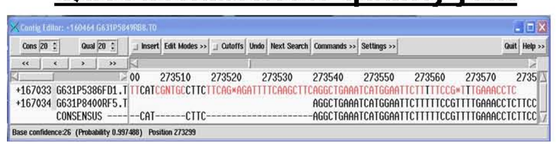
- Assemblies still contain issues
  - Misassemblies
  - Bad joins
  - Low quality regions
- These issues have proven difficult to assess in an automated fashion
- Even a short visual assessment can be useful to judge the assembly integrity and estimate the amount of finishing required

Examples of Assembly Issues:

Misassembly



Questionable/low quality join



Misassemblies and low quality joins should be counted as gaps as the resulting work required is the same

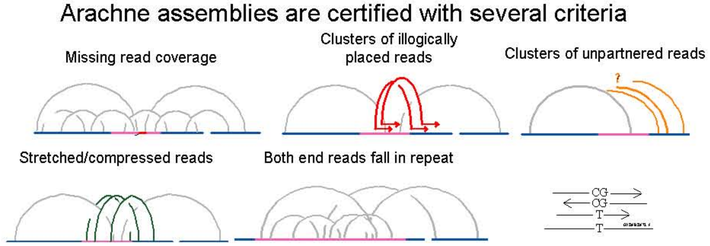
Examples of Assembly Issues (continued):

Low quality region

- Low quality regions may require less work unless several are clustered together, in which case they are treated similar to a gap
- Currently available assembly analysis tools may not correctly distinguish or recognize these issues

Diagnostic tools do not tell the complete story

Arachne assemblies are certified with several criteria

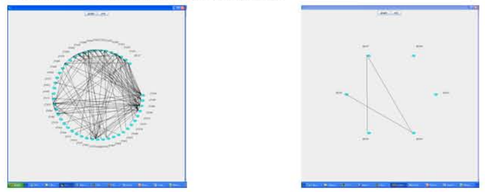


High Quality Disagreements result when reads are misplaced, but are easily scannable by a finisher

Scale of whole genome assemblies is too large for commonly used tools, such as this reverse plot diagram, which were designed for BAC sized assemblies and may miss some of these issues

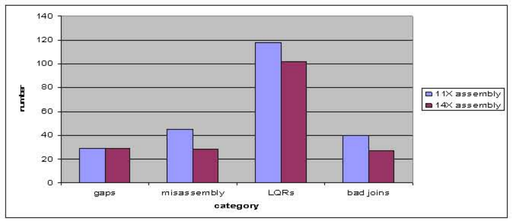
New tools give a better picture of assembly

- Newly designed tools, such as mate pair analyzer can give an accurate snapshot of an assembly
- Revplot tool visualized mates across an assembly; mate pair analyser shows mate placement across entire genome and supercontigs



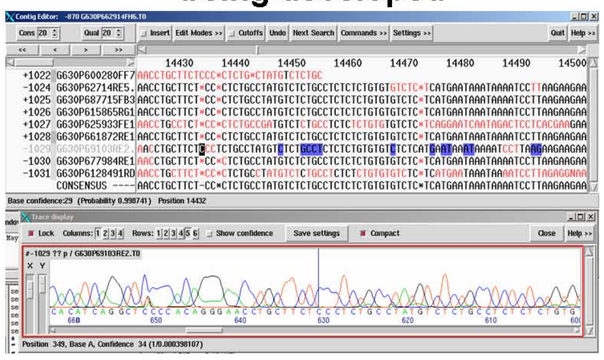
Link output of *Magnaporthe* (left) and *Coccidiodes* (right) genomes. *Magnaporthe* genome shows several misplaced links across supercontigs, whereas *Coccidiodes* assembly has far fewer misplaced reads.

Analysis of *Coccidiodes* Genome at Different Coverage Levels



- Manual assessment shows other issues besides gap count
- May not be consistent between analyses or individual analysts

Automated HQD/LQR finder being developed



- Shows some promise but currently results in a large number of false positive High Quality Disagreements (HQD) and inflates Low Quality Region (LQR) counts
- Some level of manual finishing will likely be required even with these tools to confirm tagged issues

Genome Assemblies benefit from Assessment and pre-finishing

- *Magnaporthe* and *Coccidioides* appeared to be straightforward finishes but both assemblies had issues
- Both genomes were of similar sizes (40 Mbp and 30 Mbp respectively) but the *Magnaporthe* assembly had severe assembly issues
- *Magnaporthe* assembled poorly and assessment would have resulted in a different approach
- *Coccidioides* was manually assessed and gave a much clearer picture of the amount of work required to finish

Future Directions

- Integrating draft assembly data with optical map data
- Refining the Automated HQD and LQR finders will result in quicker assembly assessment and improve prospects for pre-fin
- Developing other informatics tools, such as a gap edge analyzer
- Ultimately, the goal is to have a lot of the prefinishing work be pre-ordered, but at least a quick manual step will be required

Conclusions

- By manually assessing assemblies, we can get a better idea of how to approach finishing them
- Informatics will help this process by improving assemblies and generating auto assembly assessment
- A combination of manual and automated methods will likely be required for the most accurate appraisal of genomic assemblies