ERGA Pilot Project

Assembly Recommendations

After generating the recommended sequencing data types, each Genome Team is free to explore a variety of assembly approaches to achieve the EBP assembly standards (https://www.earthbiogenome.org/assembly-standards[1]). ERGA Pilot Project has issued the following recommended assembly approaches, for both ONT- and PacBio-based genome assemblies, to support Genome Teams in meeting these standards.

Recommended PacBio HiFi based assembly approach

For Genome Teams that have produced HiFi data, we highly recommend the use of the VGP Galaxy Pipeline for assembly. The pipeline streamlines the process of high quality reference genome production and evaluation whilst also providing the necessary computational resources - thus democratizing and decentralizing the genome assembly process. For training purposes, a step-by-step tutorial is available that demonstrates each individual step of the pipeline. For experienced users a quick-start tutorial demonstrating how to use the workflows in Galaxy is also available. If Genome Teams decide to pursue an alternative assembly approach, we still recommend that an assembly is generated using the VGP Galaxy Pipeline. This will greatly facilitate interoperability for meta-analysis for the Pilot Project Flagship Paper.

1. Genome Assessment

   Build kmer tables for HiFi and short read data (if available). Kmer tables can be used to predict genome size, heterozygosity, mean coverage and ploidy. For genome size and heterozygosity estimation using HiFi data, you can use the kmer counter **FastK** followed by **GENESCOPE.FK.** For Illumina reads (but also HiFi), you can alternatively also use the kmer counter **Meryl**[3]. Meryl kmer dbs can then be used as input to the quality evaluation tools **Merqury**[3] and **Merfin**[4]**.** If you suspect polyploidy, we highly recommend the use of **MERQURY.FK** **Smudge** or **Smudgeplot**[5]. It is good practice to keep the kmer tables (compressed) - as they can be useful later in the assembly

[1] Please note that the improvements in sequencing and in assembly algorithms now often allow to significantly exceed such standards. We encourage the generation of the best possible assemblies given the start-of-the-art approaches. For additional quality metrics please refer to Rhie et al. 2020 https://www.nature.com/articles/s41586-021-03451-0/tables/1

process. Note that all the above QC steps are natively implemented in the VGP pipeline in Galaxy.

2. Contig Assembly

To assemble the PacBio HiFi data into contigs, we recommend **hifiasm**[6], **HiCanu**[7], or **NextDenovo**. Ideally, use **hifiasm + hi-c** when HiC is available (e.g. hifiasm-hic). It is not a requirement of the Pilot Project, but if your Genome Team does have parental Illumina data, trio-binning (e.g. hifiasm-trio) should be employed to fully separate haplotypes. If employing trio-binning approaches, **LJA**[8] assembler is a possibility. Note prior to utilizing Illumina data for this purpose, it may be useful to first trim adaptor sequences using **trimmomatic**[1] or **cutadapt**[2] Organelle (mitochondrion, plastid for plants, potentially others for some organisms) contigs should be identified and removed. Animal mitochondria can be independently assembled with **MitoHifi**[9] (for HiFi data). Non-animal mitochondria and plastids are more complex, with long repeats that support recombination and variable topologies, and we are not aware yet of general automated solutions for assembling them.

3. Contaminant Removal

Most long-read assemblers create primary and alternate sets of contigs. However, it is often necessary to remove haplotypic duplication from primary contigs using **purge_dups**[10], before scaffolding. Merqury spectra plots are particularly useful to assess false and haplotypic duplications and should be generated before and after purging. At this stage, contaminants and co-bionts should also be removed, and softwares such as **BlobToolkit**[11] and **Kraken2**[12] can help identify and mark these for removal.

4. Scaffolding

Both before and after scaffolding a HiC contact map should be generated using PreText or HiGlass. Scaffold with Hi-C/OmniC data using **YaHS**[13]. **SALSA2**[14] may also be used, but YaHS is preferred.

5. Polishing

HiFi based contigs are highly accurate, however they can potentially be improved by remapping the HiFi reads with **WinnowMap**[15], calling variants with **DeepVariant**[16], followed by both **Merfin**[4] (to filter calls) and **bcftools**[17] (to apply corrections). **HyPo**[18] is an alternative tool that can be used for polishing phased assemblies. Correction with short reads may also improve accuracy, particularly for indels, but care must be taken to avoid inserting errors because of less certainty of short read mapping. Note, if both Illumina and HiFi data are available, it is recommended that polishing should implement a hybrid polishing[19] (HiFi+Illumina) strategy to help minimize effects of HiFi-specific biases.

It is important to simultaneously map the polishing reads to all the assembly material (primary, alternate, contamination, organelles) to avoid mismapping and false calls that occur if part of the genome is missing from the reference.

At this point you should have an initial draft reference genome assembly. You should now proceed to the **ERGA Pilot Project Post-Assembly Pipeline**.

Recommended ONT based assembly approach

We highly recommend that all Genome Teams generating ONT data as their primary/only long-read data type to also produce paired end Illumina short read data [>30X per haplotype]. This is important for both read assessment, and post assembly polishing and evaluation.

1. Genome Assessment

Although it is not recommended to build kmer tables for ONT data (could be possible if using amplified R10 reads), it can be useful to create these tables utilizing Illumina short read data to evaluate genome size, heterozygosity, mean coverage and ploidy. For Illumina reads, you can use the kmer counter **Meryl** [3]. Meryl kmer dbs can be used as input to the **Merqury**[3] and **Merfin**[4] quality evaluation tools, which are recommended**.** We highly recommend the use of **MERQURY.FK** **Smudge** or **Smudgeplot** [5] if you suspect polyploidy. It is good practice to keep the kmer tables (compressed) - as they can be useful later in the assembly process. Note, if after Contig assembly QC stats are low, we suggest read filtering using **Filtlong**.

2. Contig Assembly

To assemble the ONT data into contigs, if the genome size estimation for your species is large (>3Gb) we recommend an initial assembly using **wtdbg2**[20] or **Raven**[21]. For smaller genomes, or if you have the required computational resources, assemblies should be generated using either **CANU** [22], **Flye**[23] **or NextDenovo.** Noting Flye performs optimally on species with low heterozygosity. ONT read quality is continuously increasing e.g., duplex reads, and so it could be useful to test and compare an assembly using the recommended HiFi approaches. If you have parental Illumina data then you can run read binning with CANU followed by employing the trio-binning mode if available in your assembler of choice[24]. Note prior to utilizing Illumina data for this purpose, it may be useful to first trim adaptor sequences using **trimmomatic**[1] or **cutadapt**[2].

Organelles (mitochondrion, plastid for plants, potentially others for some organisms) contigs should be identified and separated out. For animal mitochondria, reassemble with **MitoVGP**[25]. Non-animal mitochondria and plastids are more complex, with long repeats that support recombination and variable topologies, and we are not aware yet of general automated solutions for assembling them.

3. Contaminant Removal

Most long-read assemblers create primary and alternate sets of contigs. However, it is often necessary to remove haplotypic duplication from primary contigs using **purge_dups**[10], before scaffolding. Merqury spectra plots are particularly useful to assess false and haplotypic duplications and should be generated before and after purging. At this stage, contaminants and co-bionts should also be removed, and softwares such as **BlobToolkit** [11] and **Kraken2**[12] can help identify and mark these for removal.

4. Scaffolding

Both before and after scaffolding a HiC contact map should be generated using **PreText** or **HiGlass**[26]. Scaffold with Hi-C/OmniC data using **YaHS**[13]. **SALSA2**[14] may also be used, but YaHS is preferred.

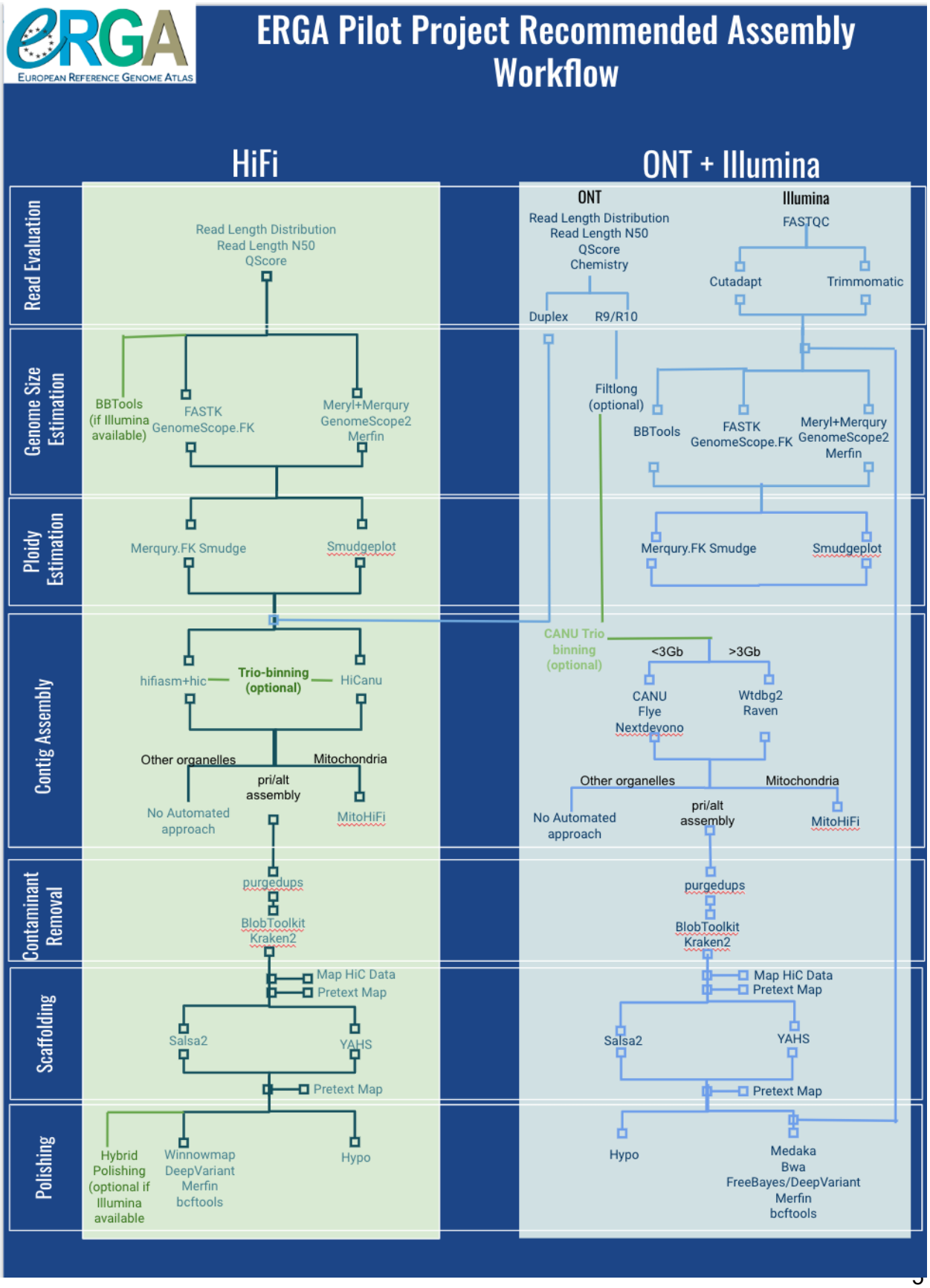
5. Polishing

To achieve higher accuracy ONT need to be polished, first using the long read data (**Medaka**) then with short read data using **bwa**[27] (to map), **FreeBayes**[28] or **DeepVariant**[16] (to variant call), **Merfin**[4] (to filter calls) and **bcftools**[17] (to apply corrections). **HyPo**[18] is an alternative tool that can be used for polishing phased assemblies. Correction with short reads may also improve accuracy, particularly for indels, but care must be taken to avoid inserting errors because of less certainty of short read mapping. It is important to map the polishing reads to all the assembly material (primary, alternate, contamination, organelles) to avoid mismapping and false calls that occur if part of the DNA is missing. Note if post assembly QC indicates a poor assembly quality, it is recommended to polish, using the above methods, after contig assembly and prior to scaffolding.

At this point you should have an initial draft reference genome assembly. You should now proceed to the **ERGA Pilot Project Post-Assembly Pipeline**.

EUROPEAN REFERENCE GENOME ATLAS

ERGA Pilot Project Recommended Assembly Workflow



References

1. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
2. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
3. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
4. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
5. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
6. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
7. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
8. Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D. & Pevzner, P. A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* **40**, 1075–1081 (2022).
9. Uliano-Silva, M., Nunes, J. G. F., Krasheninnikova, K. & McCarthy, S. A. *marcelauliano/MitoHiFi: mitohifi_v2.0*. (2021). doi:10.5281/zenodo.5205678.
10. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

11. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3* **10**, 1361–1374 (2020).
12. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
13. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *bioRxiv* 2022.06.09.495093 (2022) doi:10.1101/2022.06.09.495093.
14. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
15. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
16. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
18. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. *Cold Spring Harbor Laboratory* 2019.12.19.882506 (2019) doi:10.1101/2019.12.19.882506.
19. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
20. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
21. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nature Computational Science* **1**, 332–336 (2021).
22. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer

weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

23. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
24. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.
25. Formenti, G. *et al.* Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* **22**, 120 (2021).
26. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
28. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).