African BioGenome Project
Genomics for the future of biological diversity across Africa

African BioGenome Project Roadmap

For Genome Annotation

Version 0.9 November 2022

Table of Contents

1. Preamble and aim of the document

The African BioGenome Project (AfricaBP) is an ambitious project aimed at generating high-quality genomes of 2500 African species in three years to improve biodiversity protection, food security and environmental health in Africa. The project will be led and implemented by African researchers in collaboration with international partners. This document compiles the state-of-the-art genome annotation practices and resources that will be used to guide genome annotation workflow and setting data standards for the AfricaBP. In addition, the document is part of a suite of documents from various AfricaBP sub-committees to serve as guidelines and training resources for researchers who may wish to gain knowledge about the topic.

The AfricaBP Genome Annotation working group will address:

- Input genome assembly quality description
- Pre-annotation data processing
- Repeat identification and annotation
- Tools and optimal pipeline for structural and functional annotation
- Annotated data standards
- Highlight resources of tools and documentations
- Highlight compute/hardware that may be accessible for potential users
- Highlight training opportunities for those interested in learning and running their own analyses

2. Genome assembly and assembly quality

The assembled genomes will be assessed for the following criteria: assembly contiguity, structural accuracy, base accuracy, haplotype phasing and gene space completeness. The AfricaBP has adopted standards and metrics for assembled quality genomes proposed by the Vertebrate Genomes Project (VGP) (Rhie et al., 2021). As genome assembly is a fast evolving field, we will also rely on updates to the Report on Assembly Standards (https://www.earthbiogenome.org/assembly-standards), a living document maintained by the Earth BioGenome Project (Lewin et al., 2018). Furthermore, as the AfricaBP will also be handling plants and microorganism genomes, there will be deviations from VGP standards. For example, under the VGP-2020, assembly threshold NG50 of 10 Mbp will be adjusted to 2-8 Mbp for plants and microorganisms due to their complex repeat structures and genome size, respectively.

Table 1: Genome assembly standards and metrics adopted by AfricaBP*. Table taken and modified from Rhie et al. (2021).

Quality category	Metric	Finished	VGP-2020	VGP-2016	B10k-2014	VGP-2021
Notation	***x.y.P.Q.C***	c.c.Pc.Q60.C100	7.c.P6.Q50.C95	6.7.P5.Q40.C90	4.5.Q30	
Continuity	**Contig NG50 (*x*)**	= Chr. NG50	>10 Mb	>1 Mb	>10 kb	1–25 Mb
	Scaffolds NG50 (*y*)	= Chr. NG50	= Chr. NG50	>10 Mb	>100 kb	23–480 Mb
	Gaps per Gb	No gaps	<200	<1,000	<10,000	75–1,500
Structural accuracy	**Reliable blocks**	= Chr. NG50	>10 Mb	>1 Mb	Not required	2.3–40.2 Mb
	False duplications	0%	<1%	<5%	<10%	0.2–5.0%
	Curation	Conflicts resolved	Manual	Manual	Not required	Manual
Base accuracy	**Base pair QV (*Q*)**	>60	>50	>40	>30	39–43
	***k*-mer completeness**	100% complete	>95%	>90%	>80%	87–98%
Haplotype phasing	**Phase block NG50 (*P*)**	= Chr. NG50	>1 Mb	>100 kb	Not required	1.6 Mb[a]
Functional completeness	**Genes**	>98% complete	>95% complete	>90%	>80%	82–98%
	Transcript mappability	>98%	>90%	>80%	>70%	96%
Chromosome status	**Assigned (C)**	>100%	>95%	>90%	Not required	94.4–99.9%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least one shared (for example, X or Z)	Fragmented	At least one shared
	Organelles (for example, MT)	One complete allele	One complete allele	Fragmented	Not required	One complete allele

*The six broad quality categories in the first column are split into sub-metrics in the second column. In the x.y.z. P.A.C notation, x = log$_{10}$[contig NG50];y = log$_{10}$[scaffold NG50]; P = log$_{10}$[haplotype phased NG50 block]; Q = Phred base accuracy QV; and C = percentage of the assembly assigned to chromosomes. c denotes 'complete' telomere-to-telomere continuity. The VGP 2021 (last column) satisfies the 6.7.6.Q40.C90 standard, but some come close to achieving a higher 7.c.7.Q50.C95 standard. These metrics apply to genomes about 1 Gb or bigger. Table and caption taken and modified from Rhie et al. (2021).

The annotation sub-committee will assess and verify the quality of genome assemblies provided by the assembly sub-committee based on their set metrics. Annotation of genomes will then proceed based on the assembly level i.e., contig, scaffold or chromosome, assembly metrics and the availability of supporting data e.g., RNA-seq and species-specific details e.g., whether plant or vertebrate genome. The data type of genome assembly accepted by the annotation sub-committee will depend on the level of assembly, see accepted datatypes example, https://ena-docs.readthedocs.io/en/latest/submit/fileprep/assembly.html.

For long-read data, assessing phased haplotype for polyploid species will be critical for an accurate gene prediction. Genome assembly statistics for evaluation against standards provided in Table 1 will be done using QUAST (Gurevich et al., 2013) for assembly statistics, BUSCO (Manni et al., 2021) and Merqury (Rhie et al., 2020) for assembly completeness. Additional information on these tools and additional tools that will be used for assembly and quality assessment are listed in Appendix 1.

3. Structural and functional annotation

Genomes of target species that meet AfricaBP assembly standards will be submitted to protein-coding genes prediction pipeline for annotation. To allow for rapid development and implementation, the AfricaBP will base its design of the annotation pipeline on the Ensembl gene annotation system (Aken et al., 2016). This will involve adopting the Ensembl annotation source code (https://github.com/Ensembl) for the limited compute resources of the AfricaBP annotation sub-committee. The annotation pipeline starts with identification and annotation of repetitive element contents which need to be masked before the gene prediction step.

Since novel gene models' prediction relies on an intrinsic RNA-seq dataset, *de novo* transcriptome assembly of several tissues per species will be performed based on the data from sequencing platforms such as Illumina, Nanopore, and IsoSeq. More importantly, sufficient coverage would be reached in terms of biological significance specific to each species. In addition to available protein datasets from close relatives, species will serve as bait for homology-based prediction. The key steps for the annotation include genome preparation, structural annotation, functional annotation, and data sharing and curation (Appendix Table 1)

The classical strategy for the annotation of the species will follow three major steps including repeat identification, annotation evidence, and homology-based predictions. Under a collaborative framework with Ensembl and EMBL-EBI, the annotation will follow a typical workflow (See Figure 1). After a genome assembly preparation step i.e., quality control and repeat masking, the Protein-coding Model Building stage will help to generate an initial gene prediction based on similarity evidence and curated datasets. Also, the non-coding genes will be annotated. The model filtering stage is the final step that will screen out not-well-supported genes models.

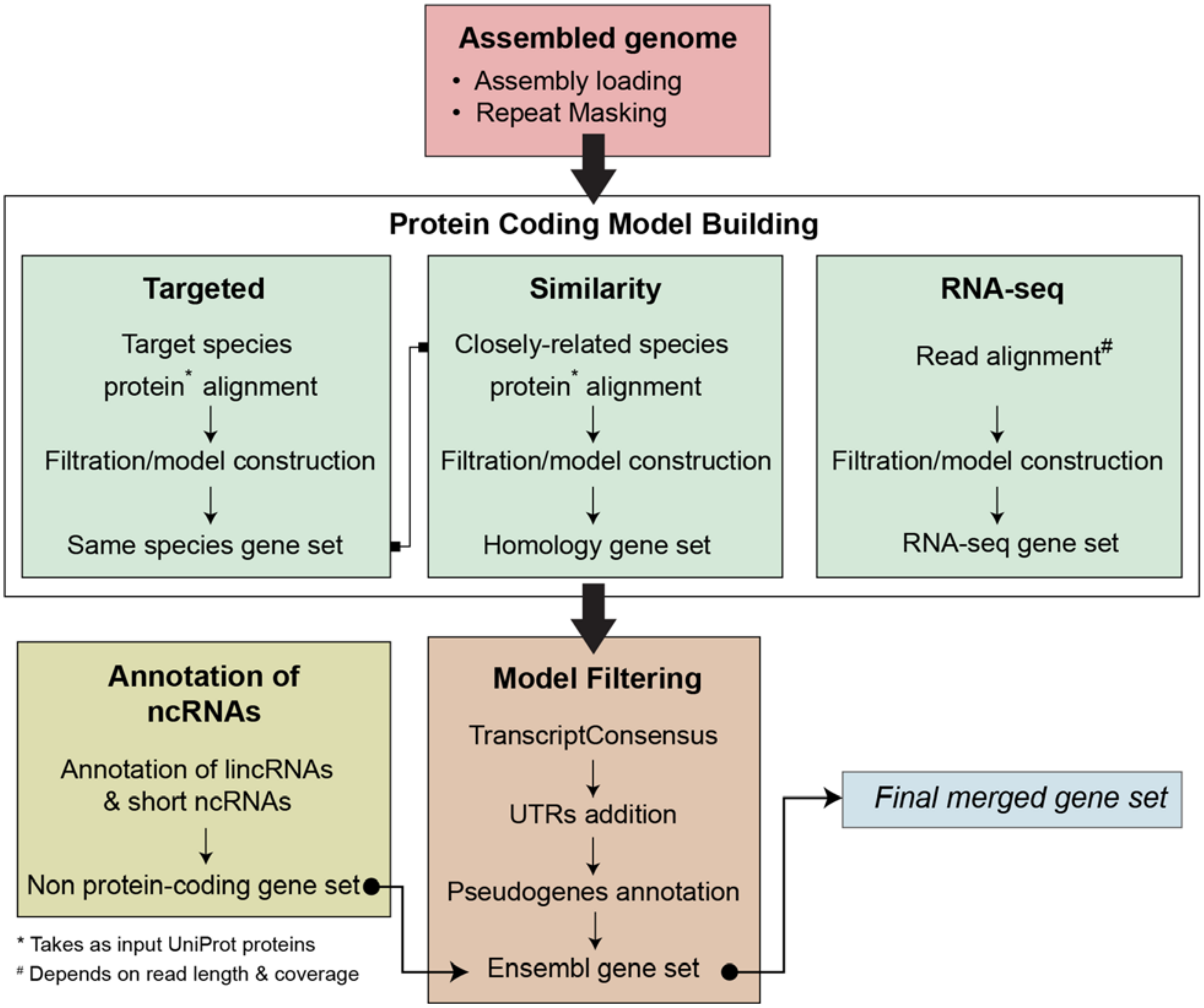
Assembled genome
• Assembly loading
• Repeat Masking
Protein Coding Model Building
Targeted
Target species protein* alignment
Filtration/model construction
Same species gene set
Similarity
Closely-related species protein* alignment
Filtration/model construction
Homology gene set
RNA-seq
Read alignment#
Filtration/model construction
RNA-seq gene set
Annotation of ncRNAs
Annotation of lincRNAs & short ncRNAs
Non protein-coding gene set
Model Filtering
TranscriptConsensus
UTRs addition
Pseudogenes annotation
Ensembl gene set
Final merged gene set
* Takes as input UniProt proteins
# Depends on read length & coverage

Figure 1: Ensembl genome annotation overview (adapted from Aken et al., 2016)

External data from high-confidence well-curated protein databases, UniRef Pfam, will be used for functional annotation of the predicted gene models. The automated annotation will further be manually curated to identify high-confidence gene models. Visualization tools, such as JBrowse, Apollo, will be employed to manually edit the automated annotations. A comprehensive list of tools required for the annotation step is presented in Appendix 1.

4. Computing Infrastructure

High-performance computing and a mix of cloud platforms will be utilized to store, download, transfer and process the generated genome sequence for endemic African species in the African BioGenome Project. Cloud infrastructures for computing and data storage constitute a key challenge that comes alongside this project. Computing infrastructures available within the African continent will be assessed to collaborate with and utilize these facilities as AfricaBP computing nodes. AfricaBP will also collaborate with various international institutions to process data including the Ensembl team at EMBL-EBI and generate high-quality reference genomes. In addition, portable, low-cost computing platforms, such as Raspberry Pi and eBioKit, and other lightweight computational pipelines that require minimal energy or the internet are anticipated to be utilized to address the challenges of energy and internet issues in Africa.

References

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet J., Billis, K., García Girón, C., Hourlier, T., and Howe K. 2016. The Ensembl gene annotation system. *Database*, 206. https://dx.doi.org/10.1093%2Fdatabase%2Fbaw093

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* 115, 4325–4333. https://doi.org/10.1073/pnas.1720115115

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Nelson, T., Griffin, P., and Christiansen J. H. 2020. Genome Annotation Infrastructure Roadmap for Australia (4.0). *Zenodo*, Pp.1-22. https://doi.org/10.5281/zenodo.3942716

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology 21, 245. https://doi.org/10.1186/s13059-020-02134-9

Nelson, Tiffanie, Griffin, Philippa, & Christiansen, Jeffrey H. (2020). Genome Annotation Infrastructure Roadmap for Australia (4.0). https://doi.org/10.5281/zenodo.3942716

Appendix 1: Genome annotation tools for annotation project

A comprehensive list of genome annotation tools is provided by Tiffanie et al. (2020) are available here. Besides, additional tools were added in the appendix 2.

Appendix 2: Additional tools

1. In case of adopting an alignment driven approach to gene prediction curated protein sequences from UniProtKB can be used in the **REAT-homology** pipeline where gene models are generated using cross species protein alignment and **Mikado** selects the best scoring model.

2. One strategy to annotate a new genome is to project gene models from an annotated reference genome. This can be done by liftover tools that lift coordinates between different assemblies of the same species or closely related species (**e.g. UCSC liftOver LiftOff CrossMap**)

3. Some organisms can have non-canonical splice junctions that can complicate accurate gene prediction. **PORTCULLIS** is able to classify genuine and false positive junctions with a high-degree of accuracy.

4. In case several types of RNAseq data are available (short and long reads) **MIKADO** can integrate assemblies and score transcripts to return a set of refined gene models eliminating chimeric fragmented or short/disrupted coding sequences.

References Additional tools

Venturini L., Caim S., Kaithakottil G., Mapleson D.L., Swarbreck D. (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience, 7, giy093, doi:10.1093/gigascience/giy093

Mapleson D.L., Venturini L., Kaithakottil G., Swarbreck D. (2018) Efficient and accurate detection of splice junctions from RNAseq with Portcullis. GigaScience, 7, giy131, doi:10.1093/gigascience/giy131

Kuhn, R.M., Haussler, D., Kent, W.J. (2013) The UCSC genome browser and associated tools, *Briefings in Bioinformatics*, 14, 2, 144-161, https://doi.org/10.1093/bib/bbs038

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J-P., Wang, L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies, *Bioinformatics*, 30, 7, 1,1006-1007 https://doi.org/10.1093/bioinformatics/btt730

EI-CoreBioinformatics. Reat.(2021) Github repository, https://github.com/EI-CoreBioinformatics/reat

Appendix 3: Intellectual contributors to the AfricaBP Genome Annotation Roadmap

Taiwo Crossby Omotoriogun, PhD

Department of Biological Sciences
Elizade University
Nigeria
crossbino@gmail.com

Yedomon Ange Bovys Zoclanclounon, MSc

Department of Crop Sciences and Biotechnology
Jeonbuk National University
South Korea
yedomon@jbnu.ac.kr

Sadik Muzemil

Graduate student
School of Life Sciences, Gibbet Hill,
University of Warwick, Coventry, CV4 7AL, UK
Twitter: @sadikmz
sadik.muzeml@gmail.com

Yasmina Jaufeerally-Fakim, Professor

Department of Agriculture
University of Mauritius
Mauritius
yasmina@uom.ac.mu

Anisah Ghoorah, PhD

Faculty of Information, Communication and Digital Technologies
University of Mauritius
Mauritius
a.ghoorah@uom.ac.mu

Girish Beedessee, PhD

Department of Biochemistry
University of Cambridge

United Kingdom
gb629@cam.ac.uk

Andrew Ndhlovu, PhD
Department of Botany and Zoology
Stellenbosch University
South Africa
andhlovu@sun.ac.za

Blessing Adanta Odogwu, PhD
Department of Plant Science and Biotechnology
University of Port Harcourt
Nigeria
blessing.odogwu@uniport.edu.ng

Harish Kothandaraman
Bioinformatician,
Collaborative Core for Cancer Bioinformatics
Purdue University
hkothand@purdue.edu