Scomber scombrus genome assembly and annotation report

Jèssica Gómez-Garrido[1], Tyler S. Alioto[1,2]
[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

Methods

Genome assembly

The *Scomber scombrus* genome assembly was produced using long reads from Oxford Nanopore Technologies (ONT) and Illumina paired-ends reads for improving the base accuracy of the ONT-based genome assembly. A flowchart with the genome assembly process is shown in **Figure ASS1**.

Prior to assembly, adaptors present in the Illumina data were trimmed with TrimGalore (https://github.com/FelixKrueger/TrimGalore). A k-mer database (k=20) was subsequently built with Meryl (https://github.com/marbl/meryl) using the trimmed short-read data. The k-mer histogram generated by Meryl was used as input to Genomescope2[1] to visualize the k-mer distribution and estimate haploid genome size, heterozygosity and repeat content. The ONT data were filtered with Filtlong[2] (--minlen 700 --min_mean_q 80) prior to assembly to remove short and low-quality reads.

The filtered ONT data were assembled with Flye v2.9[3] using the 'nano-raw' mode and a minimum overlap of 1000. To improve the base accuracy of the assembly, the assembly was polished with HyPo[4] using both Illumina and ONT data. Finally, the polished assembly was purged with purge_dups[5] to remove alternate haplotypes and other artificially duplicated repetitive regions. This final assembly was named 'fScoSco3.1_cnag1.'

Genome annotation

Repeats present in the genome assembly were annotated with RepeatMasker v4-1-2 (http://www.repeatmasker.org) using the custom repeat library available for *Danio rerio*. Moreover, a new repeat library specific for our assembly was made with RepeatModeler v1.0.11. After excluding those repeats that were part of repetitive protein families (performing a BLAST[6] search against Uniprot) from the resulting library, RepeatMasker was run again with this new library in order to annotate the specific repeats.

The gene annotation of the mackerel genome assembly was obtained by combining transcript alignments, protein alignments and *ab initio* gene predictions. A flowchart of the annotation process is shown in **Figure ANN1**.

Firstly, RNA from four different tissues was obtained and sequenced with both Illumina RNAseq and ONT direct cDNAseq. After sequencing, the long and short reads were aligned to the genome using, respectively, STAR[7] v-2.7.10a and MINIMAP2[8] v2.24 with the splice option. Transcript models were subsequently generated using Stringtie[9] v2.2.1 on each BAM file and then all the models produced were combined using TACO[10] v0.7.3. High-quality junctions to be

used during the annotation process were obtained by running Portcullis[11] v1.2.4 after mapping with STAR and MINIMAP2. Finally, PASA assemblies were produced with PASA[12] v2.5.2. The *TransDecoder* program, which is part of the PASA package, was run on the PASA assemblies to detect coding regions in the transcripts. Secondly, the complete proteomes of *Carassius auratus, Cynoglossus semilaevis, Danio rerio, Oryzias latipes, Parambassis ranga, Sparus aurata* and *Scopthalmus maximus* were downloaded from Uniprot in March 2022 and aligned to the genome using Miniprot[13] 0.6. *Ab initio* gene predictions were performed on the repeat-masked assembly with three different programs: GeneID[14] v1.4, Augustus[15] v3.5.0 and Genemark-ET[16] v4.71 with and without incorporating evidence from the RNAseq data. The gene predictors were run with trained parameters for human, except Genemark, which runs in a self-trained mode. Finally, all the data were combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM)[12]. Additionally, UTRs and alternative splicing forms were annotated via two rounds of PASA annotation updates. Functional annotation was performed on the annotated proteins with Blast2go[17]. First, a Blastp[6] search was made against the nr database (last accessed March 2023). Furthermore, Interproscan[18] v5.55_88.0 was run to detect protein domains on the annotated proteins. All these data were combined by Blast2go, which produced the final functional annotation results.

The annotation of ncRNAs was obtained by running the following steps. First, the program cmsearch[19] v1.1.4 that is part of the Infernal[20] package was run against the RFAM database of RNA families[20] v12.0. Additionally, tRNAscan-SE[21] v2.0.11 was run in order to detect the transfer RNA genes present in the masked genome assembly. Identification of lncRNAs was done by first filtering the set of PASA-assemblies that had not been included in the annotation of protein-coding genes to retain those longer than 200bp and not covered more than 80% by a small ncRNA. The resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene.

Results

Genome assembly

Results obtained with Genomescope2 (**Fig ASS2**) suggest a genome-size of 769Mb and 0.68% heterozygosity rate. The base assembly obtained with Flye v2.9 comprised a total assembly span of 765Mb (9496 contigs) and the final assembly (after polishing and purging) comprised 741Mb (4,483 contigs) (see **Table Ass1**). The contig N50 of the final assembly is 1.7 Mb, and fifty percent of the sequence (L50) is placed in 110 contigs. To estimate the accuracy and completeness of the genome assembly, BUSCO[22] v5.4.0 and Merqury[23] v1.3 were run. The consensus quality (QV) of the final assembly was estimated by Merqury as 41.4 and the gene completeness reported by BUSCO v5 was 98,6% using the odb10_actinopterygii database (see **Table Ass1**).

Table ASS1: Genome assembly statistics

Assembly	**Flye**	**Flye + hypo**	**Flye + hypo + purged**
Contig N50	1,607,853 bp	1,609,045 bp	1,748,220 bp
Contig L50	117	116	110
Total sequences	9,496	9,496	4,483
Total length	765,880,130 bp	764,611,274 bp	741,290,963 bp
BUSCO* complete	98.0%	98.7%	98.6%
BUSCO* duplicated	1.1%	1.2%	0.9%
QV	32.613	40.1577	41.4058
Kmer completeness	88.3215	89.4028	88.873

*BUSCO v5 odb10_actinopterygii database

Genome annotation

In total, we annotated 26,428 protein-coding genes that produce 38,000 transcripts (1.44 transcripts per gene) and encode for 35,860 unique protein products. We were able to assign functional labels to 94.2% of the annotated proteins. The annotated transcripts contain 10.88 exons on average, with 93% of them being multi-exonic (**Table ANN1**). In addition, 7,871 non-coding transcripts were annotated, of which 6,063 and 1,808 are long and short non-coding RNA genes, respectively.

Table ANN1: Genome annotation statistics

	SCO1A annotation
Number of protein-coding genes	26,428
Median gene length (bp)	7,720
Number of transcripts	38,000
Number of exons	278,035
Number of coding exons	264,221
Median UTR length (bp)	1,142
Median intron length (bp)	431
Exons/transcript	10.88
Transcripts/gene	1.44
Multi-exonic transcripts	93%
Gene density (gene/Mb)	35.65

Figure ASS1. Workflow of the genome assembly process.

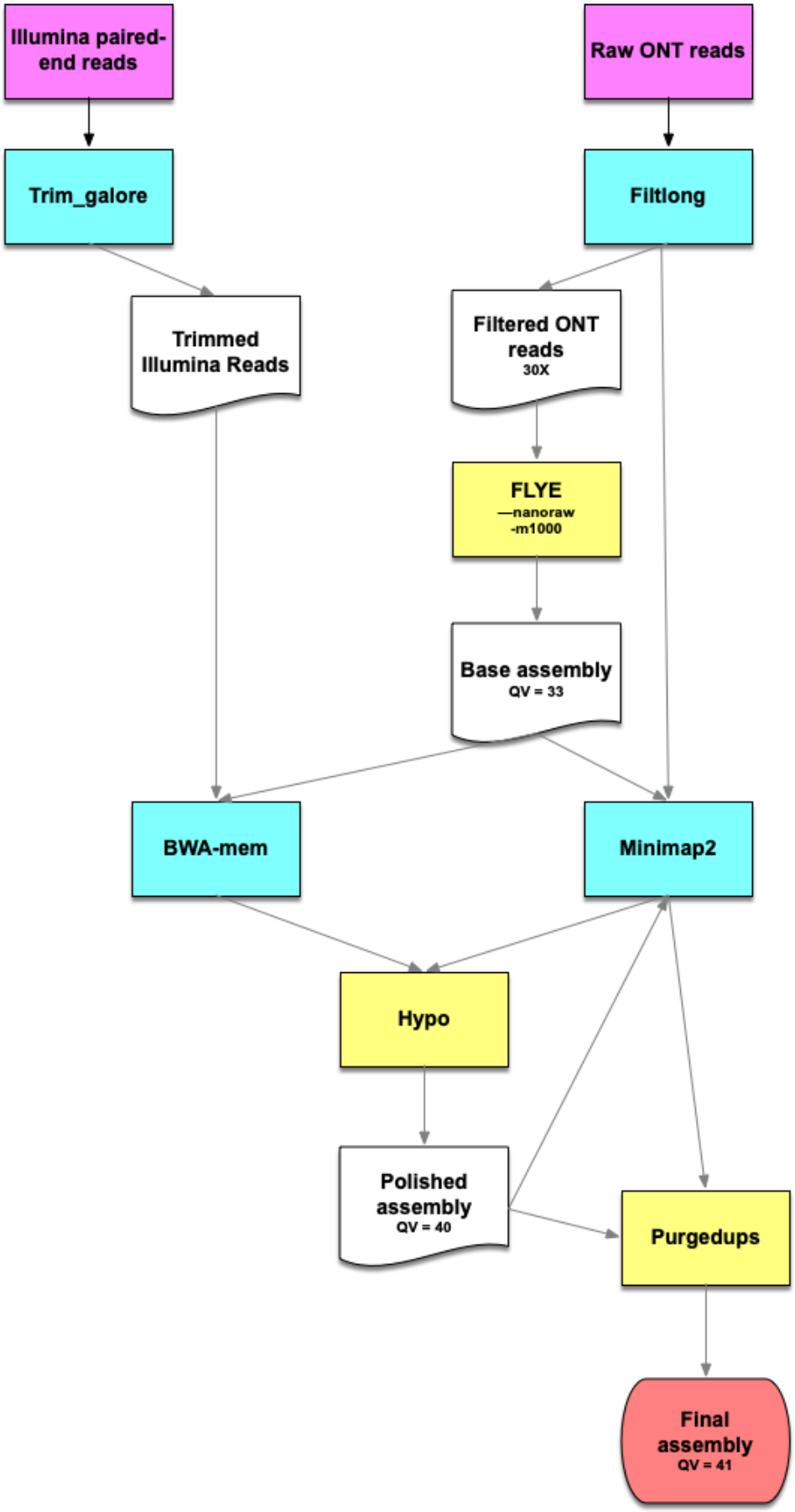


Figure ASS2. Genomescope2 transformed linear plot.

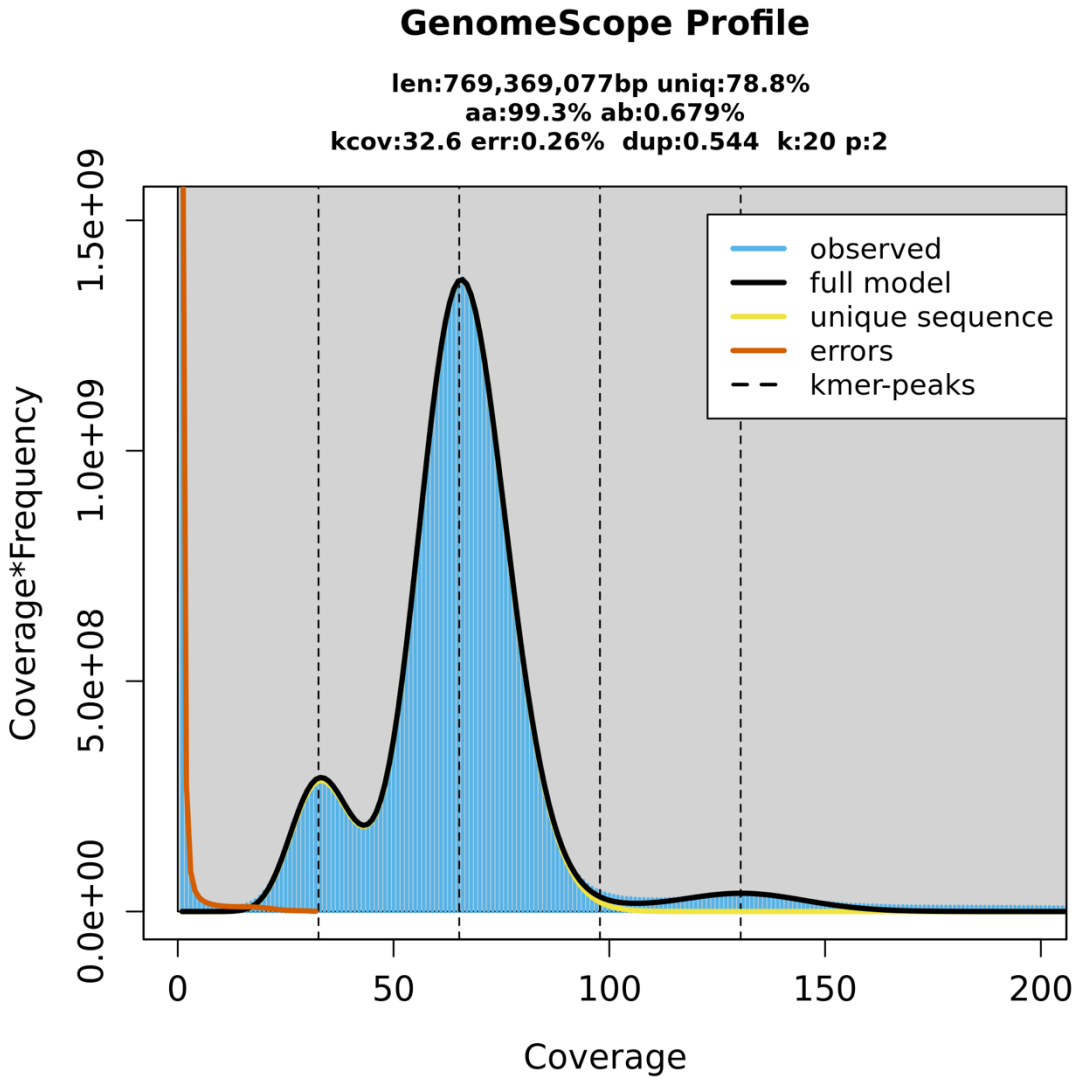
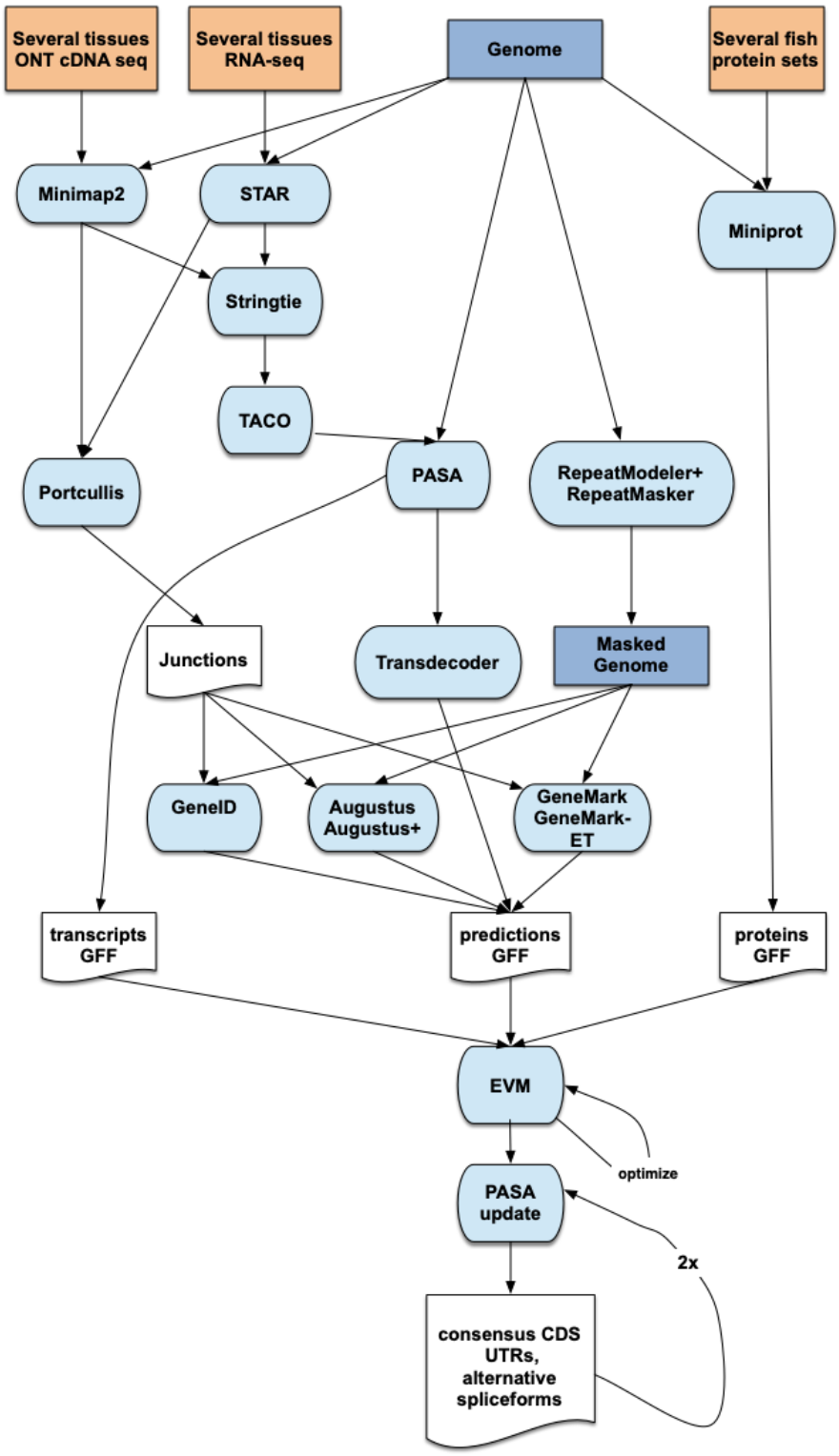


Figure ANN1: workflow of the genome annotation process



Bibliography

1. Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. 2020, GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*, **11**, 1432.
2. Wick, R. FiltLong.
3. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019, Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, **37**, 540–6.
4. Kundu, R., Casey, J., and Sung, W.-K. 2019, *HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies*. preprint, Bioinformatics.
5. Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., and Durbin, R. 2020, Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896–8.
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool. *J Mol Biol*, **215**, 403–10.
7. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
8. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–100.
9. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, **33**, 290–5.
10. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. 2017, TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods*, **14**, 68–70.
11. Mapleson, D., Venturini, L., Kaithakottil, G., and Swarbreck, D. 2018, Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience*, **7**.
12. Haas, B. J., Salzberg, S. L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.
13. Li, H. 2023, Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
14. Alioto, T., Blanco, E., Parra, G., and Guigó, R. 2018, Using geneid to Identify Genes. *Curr Protoc Bioinformatics*, **64**, e56.
15. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. 2006, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
16. Lomsadze, A., Burns, P. D., and Borodovsky, M. 2014, Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*, **42**, e119.
17. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–6.
18. Jones, P., Binns, D., Chang, H.-Y., et al. 2014, InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–40.
19. Cui, X., Lu, Z., Wang, S., Jing-Yan Wang, J., and Gao, X. 2016, CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics*, **32**, i332–40.
20. Nawrocki, E. P., and Eddy, S. R. 2013, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–5.
21. Chan, P. P., and Lowe, T. M. 2019, tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol*, **1962**, 1–14.
22. 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**, 4647–54.
23. 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.