Received: 19 September 2018 | Accepted: 17 October 2018

DOI: 10.1111/1755-0998.12964

NEWS AND VIEWS

WILEY MOLECULAR ECOLOGY RESOURCES

Perspective

# The rising tide of high-quality genomic resources

Scott Hotaling | Joanna L. Kelley

School of Biological Sciences, Washington State University, Pullman, Washington

**Correspondence**
Joanna L. Kelley, School of Biological Sciences, Washington State University, Pullman, WA.
Email: joanna.l.kelley@wsu.edu

KEYWORDS
bioinfomatics/phyloinfomatics, fish, genome assembly, genomics, long-read sequencing

Few images are more iconic of coral reef ecosystems than an orange clownfish (*Amphiprion percula*) nestled among the tentacles of its mutualistic partner, the sea anemone (Figure 1a). Popularized as the Disney character, "Nemo," clownfish are more than a charismatic on-screen presence. Among biologists, they are an ecological and evolutionary research model, shedding light on everything from social organization (Wong, Uppaluri, Medina, Seymour, & Buston, 2016) to mutualisms (Schmiege, D'Aloia, & Buston, 2017). Now, clownfish have yet another reason to be in the spotlight. In this issue, Lehmann et al. (2018) present a chromosome-level genome assembly for *A. percula* with 908.8 Megabases (Mb) of the assembled sequence placed into 24 chromosomes. The *A. percula* genome is the third, and most contiguous, in a flurry of clownfish genomes published in 2018 (see also *Amphiprion frenatus*, Marcionetti, Rossier, Bertrand, Litsios, & Salamin, 2018; *Amphiprion ocellaris*, Tan et al., 2018). Beyond strengthening future research efforts and allowing for intragenus comparisons, what is most striking about the *A. percula* genome is its completeness. With a scaffold N50 of 38.4 Mb and 98% of the assembly in chromosomal scaffolds, the *A. percula* genome is one of the most contiguous fish genomes published thus far, surpassing the high water marks of many model fishes that preceded it. Notably absent from the efforts of Lehmann et al. (2018), however, was the need for a large-scale consortium effort or an assembly approach that required unusually outsized financial resources. Instead, Lehmann et al. (2018) took advantage of a confluence of maturing technologies: long-read sequencing, dedicated assembly algorithms, and proximity-based techniques for orienting genomic regions into chromosome-scale groupings. The absence of a sequencing consortium highlights a sea change in genome biology. Gone are the days when chromosome-level genome assemblies required herculean sequencing efforts. Rather, we have entered an age where genome sequencing and assembly tools have largely bridged the gap between raw sequence data and meaningful genomic order.

Earlier this decade, reference genomes for the threespine stickleback (Jones et al., 2012) and zebrafish (Howe et al., 2013) were published in *Nature*, heralding in a new era of genome biology in fishes. As models of evolutionary, ecological and medical research, the assemblies were highly contiguous (scaffold N50, stickleback = ~10.8 Mb, Jones et al., 2012; scaffold N50, zebrafish = ~1.6 Mb, Howe et al., 2013) and empowered an array of studies. (Both genomes have been cited more than 900 times according to Google Scholar.) Of the 27 published chromosome-level fish genomes, the *A. percula* genome stands alone with ~98% of the assembled genome ordered into chromosomes (Figure 1b, Lehmann et al., 2018). This impressive feat highlights the power of modern genome sequencing and assembly tools when paired with appropriate resources for the task, a thoughtful approach, and a bit of genomic luck (e.g., few difficult to assemble regions).

This raises an important question: How exactly did Lehmann et al. (2018) construct such a high-quality, contiguous genome? First, they took full advantage of sequencing technology that was in its infancy when the zebrafish and stickleback genome projects were underway, namely long-read sequencing [i.e., read lengths in excess of ~10 kilobases; also referred to as "third-generation sequencing," Hayden (2009)]. Most modern genome sequencing efforts take advantage of this technology, typically by generating large amounts (e.g., >100x coverage) of short-read (~250 base pairs or less) data produced on an Illumina platform and overlaying lower coverage (e.g., ~10–20x) long-read data to form "hybrid" assemblies. Lehmann et al. (2018) flipped this common script by completely focusing their initial efforts on long reads produced on the Pacific Biosciences (PacBio) RS II platform. In total, they generated ~121x coverage of the *A. percula* genome with PacBio reads which corresponded to ~114 Gigabases (Gb) of sequence data. At an approximate current rate of ~1 Gb of output per run for ~$400 USD, the financial resources needed for this portion of the effort were not trivial, likely

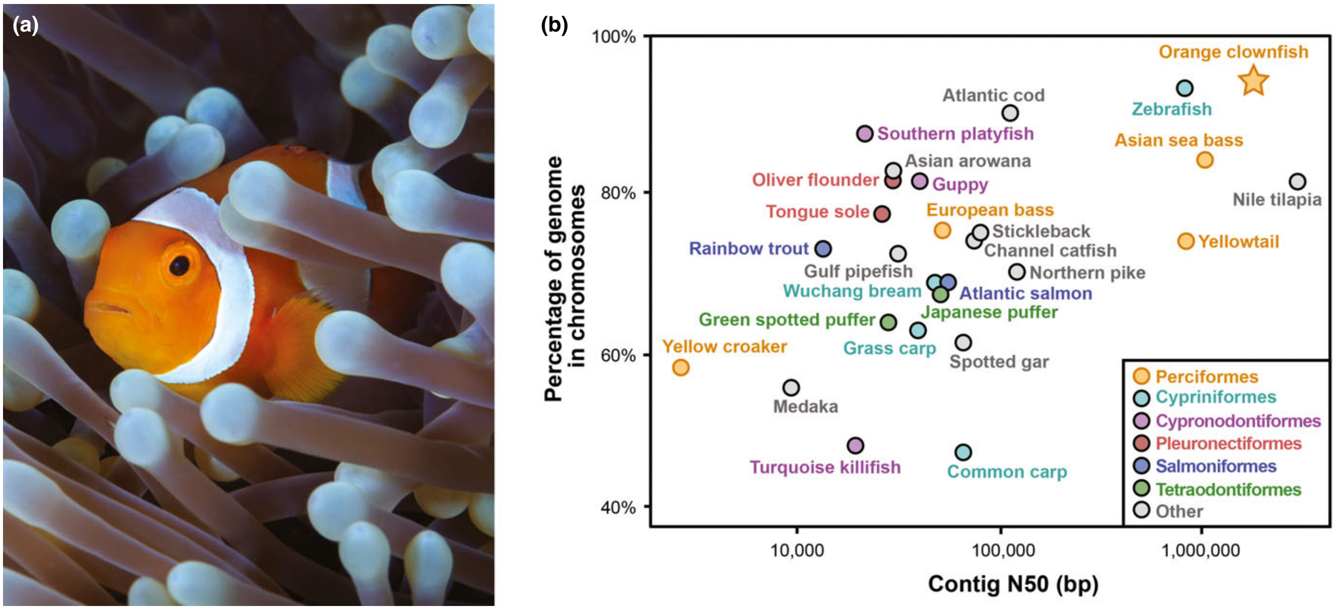

**FIGURE 1** (a) The orange clownfish, *Amphiprion percula*. Photograph by Tane Sinclair-Taylor. (b) A comparison of genome contiguity as measured by contig N50 (*x*-axis) and the percentage of the genome contained in chromosomes (*y*-axis) for *A. percula* (orange star) and the 26 previously published chromosome-scale fish genome assemblies. Points are colour-coded by order

approaching $50,000. However, with the introduction of the PacBio Sequel platform, this scale of long-read data is attainable for a fraction of the cost, perhaps as little as $10,000.

Using a combination of high-coverage PacBio data and a dedicated assembly algorithm, Lehmann et al. (2018) generated 12 “preliminary” assemblies for a range of algorithm settings. Given the contiguity of these initial assemblies—contig N50s ranging from 1.02 to 1.80 Mb with ~97% of benchmarking universal single-copy orthologs included (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015)—the authors could have stopped there and still offered a new, high-quality assembly for the community. Instead, they went further by incorporating chromatin contact maps generated by Phase Genomics with their Proximo$^{TM}$ Hi-C technology (see Lieberman-Aiden et al., 2009) to organize the *A. percula* genome into a chromosome-scale reference assembly. Alternatively, those seeking a chromosome-scale genome assembly could employ Dovetail's similar Hi-C/HiRise approach. Either way, chromatin contact mapping takes advantage of the inverse relationship between proximity of nuclear DNA and genomic distance. Physically close stretches of DNA are crosslinked and fragmented, and the ends of the resulting fragments are ligated together. The fragments are then paired-end sequenced with short-read technology. Captured within the sequence data is a frequency distribution of how often two fragments of the genome interact, and thus, how physically close they are to one another. This information allows contigs to be clustered into chromosomal groups and then oriented within chromosomes (see Burton et al., 2013). The contiguity of the *A. percula* assembly dramatically increased following this process, with scaffold N50 rising from 1.9 to 38.1 Mb, or a > 20-fold improvement. In total, >1,000 contigs comprising ~98% of the total assembly length were placed into 24 chromosomal scaffolds. To make the *A. percula* genome easier to use, Lehmann et al. (2018) mirrored larger-scale genome consortia efforts and created the aptly named genome browser, Nemo Genome DB (http://nemogenome.org/).

Beyond methodological curiosity, the *A. percula* assembly adds to a growing body of high-quality eukaryote genomes, unlocking the potential for new insight into general or lineage-specific patterns of genome evolution in fishes. Indeed, with the addition of *A. percula*, there are now six orders of fishes with more than one chromosome-level genome assembly (2–6 species per order; Figure 1b). Such an extensive resource sets the stage for new understanding of how genomic architecture (e.g., gene duplications, rearrangements) has evolved in a globally important and diverse group. For instance, the evolution of novel genes with important adaptive functions can greatly influence the evolutionary trajectory of species and comparisons of larger syntenic regions of interest among species will no doubt clarify the evolutionary processes underlying them.

As genome sequencing and assembly technology have steadily marched forward, it is difficult to specify exactly when generating high-quality genomes became more widely approachable. In their study, Lehmann et al. (2018) clearly showed that we have entered a new realm of eukaryotic genome biology. The authors took a species with no previous genome-scale data and constructed the most contiguous fish genome *ever* sequenced. In a world of sticklebacks and zebrafish, this was not a trivial feat. While the authors' efforts and methodological approach should be commended, a portion of their success stemmed from a nexus of technological advances, namely the maturation of long-read sequencing, the rise of methods for placing existing scaffolds into chromosomes and the decreasing costs of both. Indeed, the *A. percula* genome is indicative of a larger shift in eukaryote genomics where more and more chromosome-scale assemblies will be offered. And with this rising genomic tide will come new insight into the nature and evolution of genome structure across increasingly large swaths of the tree of life.

## ORCID

*Scott Hotaling* https://orcid.org/0000-0002-5965-0986
*Joanna L. Kelley* https://orcid.org/0000-0002-7731-605X

## REFERENCES

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*, 1119. https://doi.org/10.1038/nbt.2727

Hayden, E. C. (2009). Genome sequencing: The third generation. *Nature*, *457*, 768–769. https://doi.org/10.1038/news.2009.86

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*, 498–503. https://doi.org/10.1038/nature12111

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61. https://doi.org/10.1038/nature10944

Lehmann, R., Lightfoot, D. J., Schunter, C., Michell, C. T., Ohyangi, H., Mineta, K., … Ravasi, T. (2018). Finding Nemo's Genes: A chromosome-scale assembly of the genome of the orange clownfish. *Molecular Ecology Resources*, *19*(3), 570–585.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Sandstrom, R. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*, 289–293. https://doi.org/10.1126/science.1181369

Marcionetti, A., Rossier, V., Bertrand, J. A., Litsios, G., & Salamin, N. (2018). First draft genome of an iconic clownfish species (*Amphiprion frenatus*). *Molecular Ecology Resources*, *18*, 1092–1101. https://doi.org/10.1111/1755-0998.12772

Schmiege, P. F., D'Aloia, C. C., & Buston, P. M. (2017). Anemonefish personalities influence the strength of mutualistic interactions with host sea anemones. *Marine Biology*, *164*, 24. https://doi.org/10.1007/s00227-016-3053-1

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo: Hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *GigaScience*, *12*, gix137. https://doi.org/10.1093/gigascience/gix137

Wong, M. Y. L., Uppaluri, C., Medina, A., Seymour, J., & Buston, P. M. (2016). The four elements of within-group conflict in animal societies: An experimental test using the clown anemonefish, *Amphiprion percula*. *Behavioral Ecology and Sociobiology*, *70*, 1467–1475. https://doi.org/10.1007/s00265-016-2155-6

**How to cite this article:** Hotaling S, Kelley JL. The rising tide of high-quality genomic resources. *Mol Ecol Resour*. 2019;19:567–569. https://doi.org/10.1111/1755-0998.12964