Perspective

Sequencing Disparity in the Genomic Era

Kyle T. David[1*], Alan E. Wilson[2], and Kenneth M. Halanych[1]

[1]*Molette Biology Laboratory for Environmental and Climate Change Studies, Department of Biological Sciences, Auburn University, Auburn, AL 36849*

[2]*School of Fisheries, Aquaculture, and Aquatic Sciences, Auburn University, Auburn AL 36849*

*Corresponding Author: kzd0038@auburn.edu

Abstract:

Advances in sequencing technology have resulted in the expectation that genomic studies will become more representative of organismal diversity. To test this expectation, we explored species representation of nonhuman eukaryotes in the Sequence Read Archive. Though species richness has been increasing steadily, species evenness is decreasing over time. Moreover, the top 1% most-studied organisms increasingly represent a larger proportion of total experiments, demonstrating increasing bias in favor of a small minority of species. To better understand molecular processes and patterns, genomic studies should reverse current trends and adopt more comparative approaches.

Who to Sequence?

The use of model organisms, such as maize (*Zea mays*), the mouse (*Mus musculus*), and the fruit fly (*Drosophila melanogaster*), have contributed greatly to our understanding of biology via their tractability and large research communities (Müller and Grossniklaus 2010). Thus, when whole genome sequencing came of age, focusing on organisms that have been widely used as models was sensible. With cost-effective, high-throughput sequencing, however, many barriers that limited the use of non-model organisms have been removed. Advances in non-model and reduced representation genome sampling approaches have enabled researchers to sequence virtually any organism more cheaply and easily than ever before (Goldstein and King 2016). These advances enable comparative studies with broad sampling from across the tree of life that can elucidate the origins and variation in cellular mechanisms thereby ushering in a new era of discovery (Dunn and Ryan 2015). In light of this new approach, many researchers have predicted that the lines between model and non-model organisms will blur or disappear entirely (Davis 2004; Müller and Grossniklaus 2010; Goldstein and King 2016; Bolker 2017).

Trends in High-Throughput Sequencing Biodiversity

To explore how high-throughput sequencing efforts are distributed across the diversity of eukaryotic life, we accessed all nonhuman eukaryotic sequencing experiments in the Sequence Read Archive (SRA) using the Entrez Direct suite of UNIX commands. The SRA is a high-throughput sequence database administered by the DNA

Data Bank of Japan, European Nucleotide Archive, and the National Center for Biotechnology Information (NCBI). Note that experiments are defined in the SRA as "a unique sequencing result for a specific sample" and can be from experimental or descriptive research. Experiments may use one of many different sequencing strategies, though RNA-seq (37.6% of experiments) and whole genome sequencing (22.6% of experiments) are the most common. The search was executed on January 20th 2019 and returned 1,874,638 experiments. Of those, 29,578 (1.6%) experiments were removed either because they had pooled samples from multiple species or missing data. Experiments were restricted from those published between 2010, the first year with a sufficient (>100) number of species for our analysis, and December 2018, the most recent month with complete data at the time of the search. The final dataset includes 1,808,136 high-throughput sequencing experiments from 24,288 unique species.

The top 1% of species with the most experiments represent 85.3% of all experiments accessed. The top 1% includes 120 animals, 99 plants, 14 fungi, and 9 protists. Fifteen phyla are represented, although 84.1% of species are either streptophytes (green plants; n=98), chordates (n=73), or arthropods (n=30). At the species level, the mouse *Mus musculus* is the most represented by a wide margin with 523,192 experiments, 4.4x more than any other species, and representing 28.9% of all experiments. All 13 model organisms officially recognized by the National Institute of Health (NIH) (https://web.archive.org/web/20161123070020/http://modelorganisms.nih.gov/) are featured in the top 1%, which together represent 44.7% of accessed experiments.

As expected, species richness has increased over time from 453 unique species sequenced in 2010 to 9,696 in 2018. However, despite an increase in the number of unique species sequenced over time, species evenness has decreased significantly (Fig. 1A). This trend appears to be mediated at least in part by an increasing preference toward relatively fewer study species. The top 1% represented 49.7% of experiments in 2010; however, in 2017 the top 1% represented 80.4%. The top 1% of species for each month has been increasing at a rate of about 5.0% year^{-1} (Fig. 1B).

Of the 24,288 species we accessed, 1,146 have significant increases in the number of experiments over time. Given that high-throughput sequencing has increased exponentially, seeing large increases in number of experiments over time for highly studied organisms is not surprising. Indeed, with the exception of *Plasmodium falciparum,* the top five species with the greatest increases were all NIH recognized models (in descending order: *M. musculus*, *Saccharomyces cerevisiae*, *P. falciparum, Arabidopsis thalina*, and *D. melanogaster*).

We also explored change in relative frequency of experiments for each organism over time. In terms of relative frequency, the only NIH model that maintained a significant increase over time was the mouse. The two species with the largest decreases in relative frequency were *D. melanogaster* and *Caenorhabditis elegans*. While longstanding models are becoming less dominant, other organisms, such as the olive baboon (*Papio anubis*) and mummichog (*Fundulus heteroclitus*), appear to be receiving more attention (Data S1).

As anticipated, species richness in high-throughput sequencing experiments increased significantly over time (Fig. 1A). This finding demonstrates a trend toward sequencing greater taxonomic diversity, driven by recent initiatives such as those undertaken by Genome 10K and the Global Invertebrate Genomics Alliance (GIGA). Notably, in November 2018, the Earth BioGenome Project was launched, which aims to sequence all eukaryotic species genomes in 10 years.

Although more unique species are being sequenced over time, the disparity of sequencing efforts is widening, suggesting more focus is being put on relatively fewer species. In particular, the mouse has 1.8x more experiments than all other NIH models combined, and the number of experiments is growing linearly at a rate of 150.8 per month, 4.3x faster than any other organism. Mouse is also the only NIH model whose relative representation has increased, taking up a larger proportion of total sequencing experiments over time (increasing at a rate 3.5x that of any other organism).

"All Models are Wrong"

Model organisms have been a fundamental aspect of biology for at least 150 years (Müller and Grossniklaus 2010); however, they are not without problems. As statistician George Box famously articulated, "All models are wrong, but some are useful." Many researchers have previously commented on the pitfalls of model-centric research (Davis 2004; Bolker 2017), which can result in disastrous and even lethal consequences as with the infamous fialuridine and thalidomide drug trials during the 20th century (Warkany 1988; Xu, et al. 2014).

In particular, there are many questions for which mouse models are not well-suited, in spite of their popularity. In recent years, the mouse has been shown to be an imperfect representative of human disorders, particularly with regard to neurological and immune disease as well as cancer, resulting in frustratingly few real-world applications relative to the investment (Schnabel 2008; Geerts 2009; Seok, et al. 2013; Baker and Amor 2015; Perlman 2016). Additionally, early approximations of human gene count based on homology with mouse were overestimated by ~60,000, a number which was reduced to ~10,000 by a study that used the more compact genome of the pufferfish *Tetraodon nigroviridis* (Crollius, et al. 2000). Current estimates of human gene count rely on comparisons across many different species.

While the model organism philosophy is instrumental to understanding certain intraspecific mechanisms, it is singularly inappropriate to address these questions in an evolutionary context and therefore not capable of answering questions on the origins and variation of such mechanisms. Broad taxonomic sampling is a prerequisite to assess evolutionary processes and patterns. For example, a centralized nervous system was thought to be an ancestral character of bilaterians with a single evolutionary origin based on similar expression patterns in *D. melanogaster*, the annelid worm *Platynereis dumerilii*, and vertebrates. However a study in 2018, based on novel sequence data from additional groups within *Bilateria*, found different nervous-system architectures even between closely related taxa, suggesting nerve cords evolved within *Bilateria* multiple times (Martín-Durán, et al. 2018). Similar patterns remain to be explored for many other biological characters, such as gastrulation and segmentation (Dunn, et al. 2014).

The Sangerian Shortfall

Just as the Linnaean Shortfall describes how few species have been formally described, we define the Sangerian Shortfall as the lack of knowledge regarding most species' genomes. The 24,288 species represented in the SRA represent 2.0% of the

1,186,221 described eukaryotic species in NCBI's taxonomy database and only 0.0027% of the 8.7 million eukaryotic species thought to exist on earth. For species with whole genome sequences available, representation is even lower, 9,613 species as of December 2018. Of the 14,927 species currently listed as endangered or critically endangered by the IUCN, only 2.6% had high-throughput sequence data.

Concluding Remarks

Advances in high-throughput sequencing technology have enabled researchers to sample from more species than ever before. In spite of this, genomic sampling is becoming more model-focused as relatively more attention is being paid to fewer species. Negative effects of this trend extend beyond biodiversity and evolutionary studies to many fields including pharmacology, development, genetics, and neurobiology. To improve taxonomic diversity in genomic studies we recommend 1) developing and improving funding resources to increase our understanding of biodiversity (such as the Dimensions of Biodiversity and PurSUit programs offered by the USA National Science Foundation) and 2) taking steps to ensure reviewer panels are represented by researchers working on a variety of study systems to reduce bias in funding decisions. We predict that these recommendations, if acted upon, will not only improve our understanding of variation and diversity in nature but also foster collaborations across different research fields and systems. We conclude that molecular researchers should attempt to select models based on their relevance to biological questions over ease of us and use comparative approaches to address questions in an evolutionary framework whenever possible.

All code required to update experiments and reproduce results/figures are available at https://github.com/KyleTDavid/SRA2019. Original data files are available at https://figshare.com/projects/SRA2019/39296.

Acknowledgments: We would like to thank Doug Levey, three anonymous reviewers, and all members of the Molette and Phyletica labs, in particular Damien Waits, Caitlin Redak, Michael Tassia, and Drs. Yuanning Li and Jamie Oaks for their feedback and comments. This project was supported by the Alabama Agricultural Experiment Station, the Hatch Program of the National Institute of Food and Agriculture, U.S. Department of Agriculture, Cell and Molecular Biosciences Fellowship, and the Schneller Endowed Chair Fund. This is Molette Biology Laboratory contribution XX and Auburn University Marine Biology Program contribution XX.

References

Baker D, Amor S. 2015. Mouse models of multiple sclerosis: lost in translation? Current pharmaceutical design 21:2440-2452.
Bolker JA. 2017. Animal Models in Translational Research: Rosetta Stone or Stumbling Block? BioEssays 39:1700089.
Crollius HR, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quétier F. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nature genetics 25:235.
Davis RH. 2004. The age of model organisms. Nature Reviews Genetics 5:69.
Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. Annual review of ecology, evolution, and systematics 45:371-395.
Dunn CW, Ryan JF. 2015. The evolution of animal genomes. Current opinion in genetics & development 35:25-32.
Geerts H. 2009. Of mice and men. CNS drugs 23:915-926.
Goldstein B, King N. 2016. The future of cell biology: emerging model organisms. Trends in cell biology 26:818-824.
Martín-Durán JM, Pang K, Børve A, Lê HS, Furu A, Cannon JT, Jondelius U, Hejnol A. 2018. Convergent evolution of bilaterian nerve cords. Nature 553:45.
Müller B, Grossniklaus U. 2010. Model organisms—a historical perspective. Journal of proteomics 73:2054-2063.
Perlman RL. 2016. Mouse models of human diseaseAn evolutionary perspective. Evolution, medicine, and public health 2016:170-176.
Schnabel J. 2008. Neuroscience: standard model. Nature News 454:682-685.
Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, Richards DR, McDonald-Smith GP, Gao H, Hennessy L. 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences 110:3507-3512.
Warkany J. 1988. Why I doubted that thalidomide was the cause of the epidemic of limb defects of 1959 to 1961. Teratology 38:217-219.
Xu D, Nishimura T, Nishimura S, Zhang H, Zheng M, Guo Y-Y, Masek M, Michie SA, Glenn J, Peltz G. 2014. Fialuridine induces acute liver failure in chimeric TK-NOG mice: a model for detecting hepatic drug toxicity prior to human testing. PLoS medicine 11:e1001628.

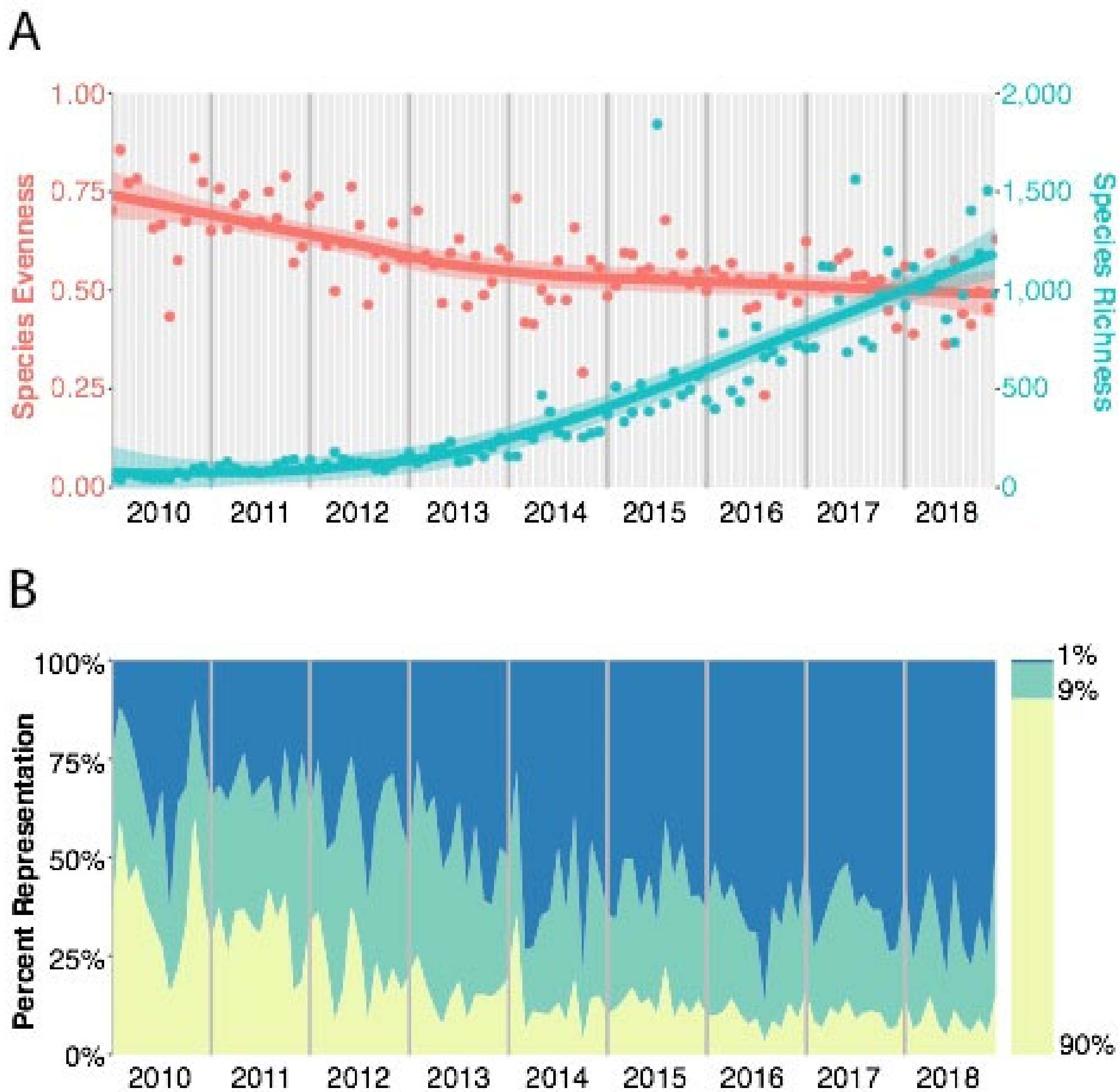

Figure 1: Biodiversity in the Sequence Read Archive. A) Pielou's species evenness (H/ln(S) where H is Shannon's diversity index and S is species richness) and species richness of all nonhuman eukaryotic sequencing experiments in the SRA calculated for each month between January 2010 and December 2018. Richness has increased over time ($p<1.1E-32$) while evenness has decreased ($p<2.5E-13$). B) Relative representation of the top 1%, top 2-10%, and bottom 90% of species in the SRA by number of experiments for each month between January 2010 and July 2018.

Data S1: Summary statistics for each species represented in the SRA between January 2010 and December 2018