# Factors influencing taxonomic unevenness in scientific research: A mixed-methods case study of non-human primate genomic sequence data generation

Margarita Hernandez[1], Mary K. Shenk[1], and George H. Perry[1,2,3]

[1]Department of Anthropology, [2]Department of Biology, [3]Huck Institutes of the Life Sciences
Pennsylvania State University
University Park, PA 16802

Margarita Hernandez, ORCID iD: 0000-0001-6522-6455
Mary K. Shenk, ORCID iD: 0000-0003-2002-1469
George H. Perry, ORCID iD: 0000-0003-4527-3806

Corresponding Authors: M.H. (mzh235@psu.edu) and G.H.P. (ghp3@psu.edu)

Keywords: Taxonomic bias; model organisms; massively-parallel sequencing; ethnography of scientists; species conservation status

## ABSTRACT

Scholars have often noted major disparities in the extent of scientific research conducted among taxonomic groups. Such trends may cascade if future scientists gravitate towards study species with more data and resources already available. As new technologies emerge, do research studies employing these technologies continue these disparities? Here, using non-human primates as a case study, we first identified disparities in recently-generated massively-parallel genomic sequencing data and we then conducted interviews with the scientists who produced these data to learn their motivations when selecting species for study. Specifically, we tested whether variables including publication history and conservation status were significantly correlated with publicly-available sequence data in the NCBI Sequence Read Archive. Of the 179.6 terabases (Tb) of sequence data in this database for 519 non-human primate species, 135 Tb (~75%) were from only five species: rhesus macaques, olive baboons, green monkeys, chimpanzees, and crab-eating macaques. The strongest individual predictors of the amount of genomic data were the total number of non-medical scholarly publications (linear regression; $r^2$=0.37; $P$=6.15x$10^{-12}$) and number of medical publications ($r^2$=0.27; $P$=9.27x$10^{-9}$). In a generalized linear model, the number of non-medical publications ($P$=0.00064) and closer phylogenetic distance to humans ($P$=0.024) were the most predictive of the amount of genomic sequence data. We interviewed 33 authors of genomic data-producing publications and analyzed their responses using a grounded theory approach. Consistent with our quantitative results, authors mentioned that their choices of species were motivated by sample accessibility, prior published work, and perceived relevance (especially health-related) to humans. Our mixed-methods approach helped us to identify and contextualize some of the driving factors behind species-uneven patterns of scientific research, which can now be considered by funding agencies, scientific societies, and research teams aiming to align their broader goals with future data generation efforts.

## SIGNIFICANCE STATEMENT

Our study sheds lights on the species-uneven distribution of genomic sequence data generation across the order Primates. We used a combination of quantitative data analyses and qualitative interviews with authors of data-producing studies to identify factors that have driven the observed pattern of unevenness; these included the extent of prior research conducted on each species, the relevance to human medicine, phylogenetic distance to humans, and sample accessibility. While our study focused on factors influencing non-human primate genomic sequence data, similar questions can be asked about how the scientific community engages with research projects more broadly. Our goal is to bring attention to the diversity of factors that influence scientists as they plan their projects, so that this process can be considered in the future by research groups and funding agencies aiming to align their broader goals with future data generation efforts.

## INTRODUCTION AND BACKGROUND

Scholars have long observed taxonomic unevenness in terms of focal species included in published research studies. On a broader taxonomic level, birds and mammals are overrepresented in the scientific literature, while fish, amphibians, and invertebrates are included at a relative deficit to their actual abundance in nature (1–3). Conservationists specifically have observed the tendency for species to be selected based on their "charisma" or appeal (for reasons society may ascribe to certain species based on their "beauty, valor, or singularity") to scientists and/or the general public (4–6). Additionally, species that are characterized as "models" for various processes or fields – for example *Arabidopsis thaliana* in the botanical sciences or rhesus macaques (*Macaca mulatta*) in biomedicine – may continue to be disproportionately studied due to the benefit from the continuous accumulation of knowledge and research tools specific to that organism (7).

These patterns of taxonomic unevenness in scientific research matter. Specifically, future scientists will be primed to more readily and powerfully answer novel questions with species having extensive histories of prior study relative to more understudied taxa. This cascade is especially strong when the data produced in earlier studies have been made freely available to other researchers; in addition to reproducibility-related benefits, public data sharing allows for important, downstream research questions to be developed and answered using data originally generated for other research purposes.

For our study, we sought to assess whether the longstanding taxonomic unevenness in scientific research publication is similarly observed in patterns of emerging technology use. If so, then what factors are influencing or even driving this phenomenon? Conveniently, given the still-growing use of the technology that is the focus of our study, we can investigate these potential patterns of unevenness in real-time and incorporate insights from interviews with the very scientists generating these data.

Specifically, we focused on the use of massively-parallel genomic sequencing methods. The development and continued technological innovation of these tools have helped scientists answer expanding sets of questions in species biology, evolutionary history, behavioral ecology, and population dynamics (8–11). The genetics and genomics community as a whole has been a leader in the data sharing movement, with standardly-used online repositories including the National Center for Biotechnology Information's Sequence Read Archive (SRA), the Gene Expression Omnibus, and GenBank (12).

Our study aims to investigate patterns of taxonomic unevenness within publicly available genomic sequence data archives, using non-human primates as a case study. Non-human primates are among the world's most endangered taxonomic groups, with 60% of all non-human primates at risk of extinction (13, 14). Non-human primates serve important ecological, cultural, and medical purposes (13). Their extinction would threaten the ecosystems they inhabit and our opportunities to understand human biology. Given their close phylogenetic relationship with humans, non-human primate taxa have been regularly studied to help understand the progression of many human diseases (15, 16), including HIV (17) and Alzheimer's (18).

Our goal was to identify variables associated with patterns of species-unevenness in genomic sequence data across all 519 non-human primate species. We considered factors including prior publication history, geographic range, International Union for the Conservation of Nature (IUCN) Red List conservation status, and phylogenetic distance to humans. Additionally, we incorporated a qualitative component in which we interviewed first and/or corresponding authors on papers that generated non-human primate genomic sequence data to record their motivations and the factors that they explicitly considered when selecting species to study. This mixed-methods approach let us identify quantitative patterns in the existing distribution of published genomic sequence data while simultaneously investigating the contexts in which these data were generated.

## RESULTS

We downloaded metadata for a total of 179.6 terabases (Tb) of non-human primate genomic sequence data available in the NCBI Sequence Read Archive (SRA) database as of August 16, 2018. The order Primates is comprised of 520 total species (including humans). We found that 416 of the 519 (80.2%) non-human species did not have *any* genomic sequence data deposited in SRA at the time of our analysis. Of the 103 (19.8%) species that are represented, the majority of the sequence data (133.2 Tb; 74.2%) come from only five different species (Figure 1): rhesus macaques (*Macaca mulatta*), olive baboons (*Papio anubis*), green monkeys (*Chlorocebus sabaeus*), chimpanzees (*Pan troglodytes*), and crab-eating macaques (*Macaca fascicularis*).

For each non-human primate species, we also recorded the following information based a combination of our own hypotheses and variables considered in previous species disparity studies in other taxonomic groups (19–23): current conservation status (Least Concern, Near Threatened, Vulnerable, Endangered, and Critically Endangered) and geographic species range (km$^2$) from IUCN, the number of both medical and non-medical scholarly publications featuring each species from the Web of Science database, the estimated evolutionary distance to humans (millions of years ago for most recent common ancestor) from a recent phylogenetic analysis (24), the number of individuals currently housed in >1,000 worldwide zoos and other conservation facilities who are Species 360 members, and activity pattern (nocturnal, diurnal, cathemeral) (25) (Supplementary Dataset 1). These variables were compared to the amount of genomic data (Mb) available in the SRA database for each species, both on an individual variable basis (e.g. linear regressions) and collectively (e.g. logistic regression).

### *Variables associated with the presence or absence of genomic data*

First, given the large proportion of non-human primate species without any available genomic data (n=416), we tested which variables were significantly associated with presence versus absence of genomic data (Supplementary Figure 1). We found that species with genomic data have significantly more non-medical publications (763.26 ± 2,915.19) than those without genomic data (28.82 ± 74.42; t-test; *P*=0.0126). We observed a moderately significant difference in the number of medical publications between species with genomic data (16.17 ± 84.56) compared to those without genomic data (0.11 ± 1.08; *P*=0.058). Species with genomic data available also had larger geographic ranges (905,615 ± 1,664,226 km$^2$) and more individuals in captivity (242 ± 502 individuals) than species without genomic data (385,345 ± 815,673 km$^2$; P=0.00342; 28 ± 98 individuals; $P$=3.815 x 10$^{-5}$). There was no significant difference in millions of years since last shared common ancestor with humans between species with genomic data (45.03 ± 20.97) and those without genomic data (48.80 ± 17.17; *P*=0.094). Presence/absence of genomic data were also significantly associated with red list status (Chi-square test; *P* = 0.0024) but not with activity pattern (*P*=0.1506). We also a performed logistic regression and determined that a greater number of non-medical publications ($P$=2.57x10$^{-7}$), a greater number of individuals in captivity (*P*=0.0411), and species categorized as endangered relative to critically endangered within IUCN Red List status (*P*=0.0229) were significantly predictive of the presence of genomic data in the context of all other variables.

### *Variables associated with the amount of genomic data per species*

In addition to presence/absence of genomic data, we tested whether the amount of genomic data (megabases in the NCBI SRA database) per species is significantly associated with our variables of interest (Figure 2). For the entire dataset (including species with no genomic data), the total number of non-medical publications explained 33% of the variation ($r^2$=0.33) within the genomic sequence data ($P$<2 x 10$^{-16}$). Number of medically-focused publications and frequency in captivity explained 27% and 22% of the

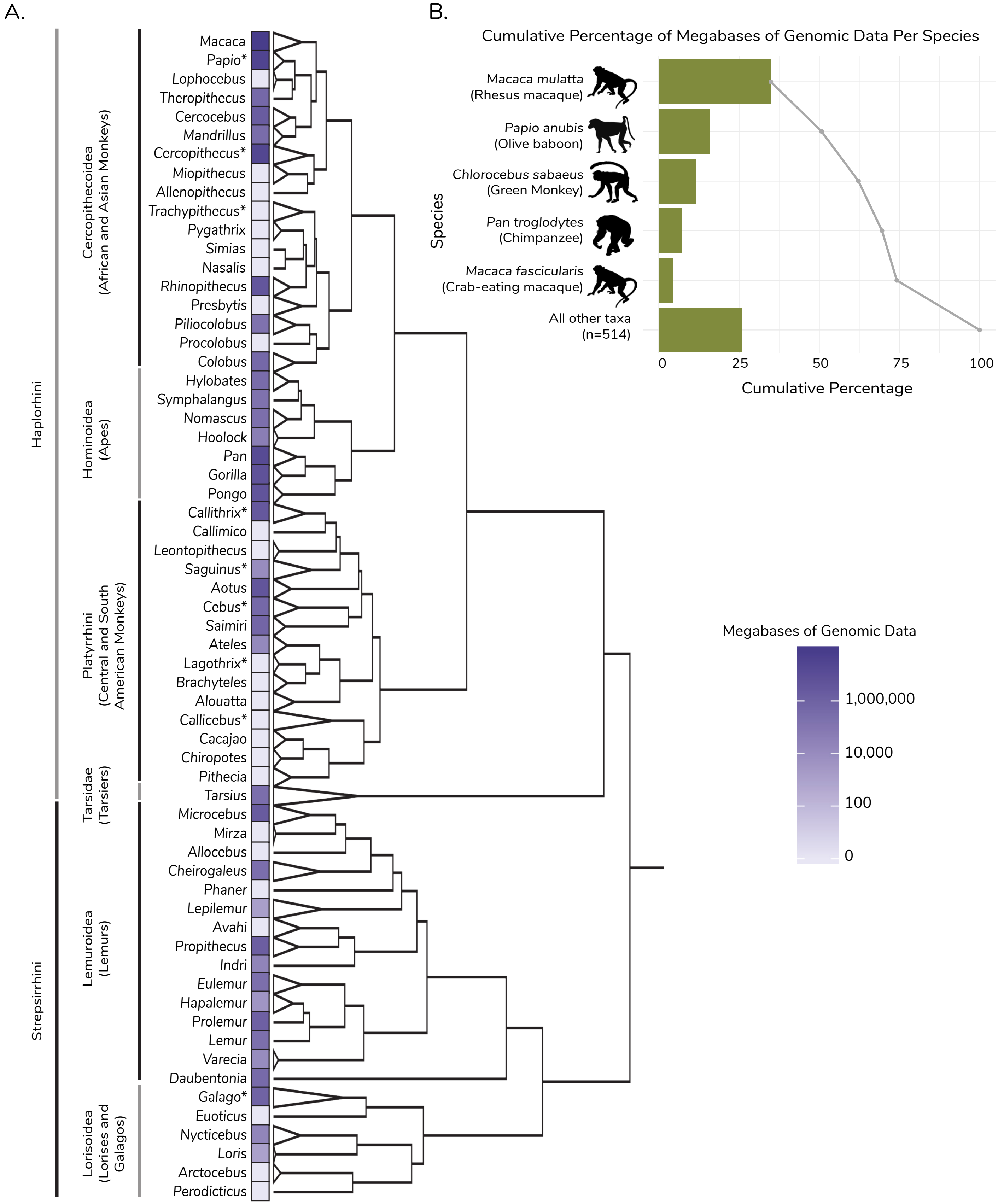


**Figure 1. Megabases of genomic data by genus across the order Primates and for the five species with the most genomic data.** A) Phylogeny of the order Primates with dark purple indicating more genomic data per genus and white indicating little to no genomic data. Paraphyletic genera are denoted with an asterisk. A complete list of genera is provided in Supplementary Table 1. Phylogeny adapted from Dos Reis et al. (2018). B) The five species with the most genomic data and the cumulative percentage of the total amount of non-human human primate sequence data represented by these taxa. Credit to T. Michael Keesey and Tony Hisgett for the chimpanzee image, under license https://creativecommons.org/licenses/by/3.0/.

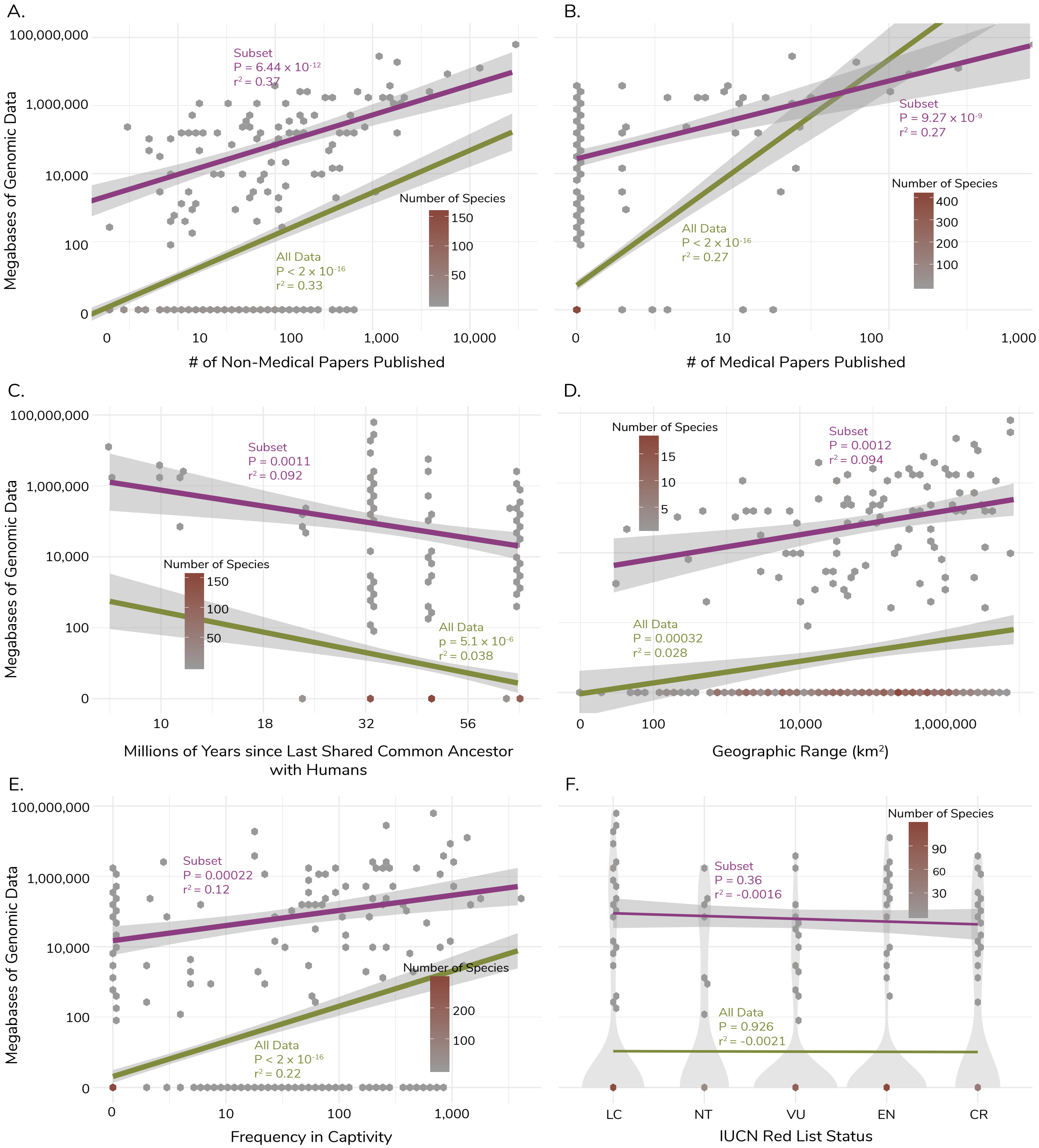


**Figure 2. Linear regressions for entire dataset and subset of species with genomic data.** Linear regressions for six variables used within the study, non-medical papers published (A), importance in medical research (B), relatedness to humans (C), geographic range (D), frequency in captivity (E), and IUCN Red List status (F). For each, the purple line represents the linear regression for the subset of species with genomic data available, while the green line represents the linear regression for all species within the dataset.

variation within genomic sequence data ($P$<2 x $10^{-16}$ and $P$<2 x $10^{-6}$, respectively). While phylogenetic relatedness to humans and geographic range were statistically significantly associated with genomic data, they had limited explanatory power ($P$=5 x $10^{-6}$ and $r^2$=0.038, and $P$=0.00032 and $r^2$=0.028, respectively). IUCN Red List status was neither significant nor explanatory (IUCN Red List status treated as an ordinal variable; $P$=0.926, $r^2$=-0.0021). Using ANOVA, there were no statistically significant differences across IUCN Red List or activity pattern categories in the amount of genomic sequence data available ($P$=0.244 and $P$=0.49, respectively).

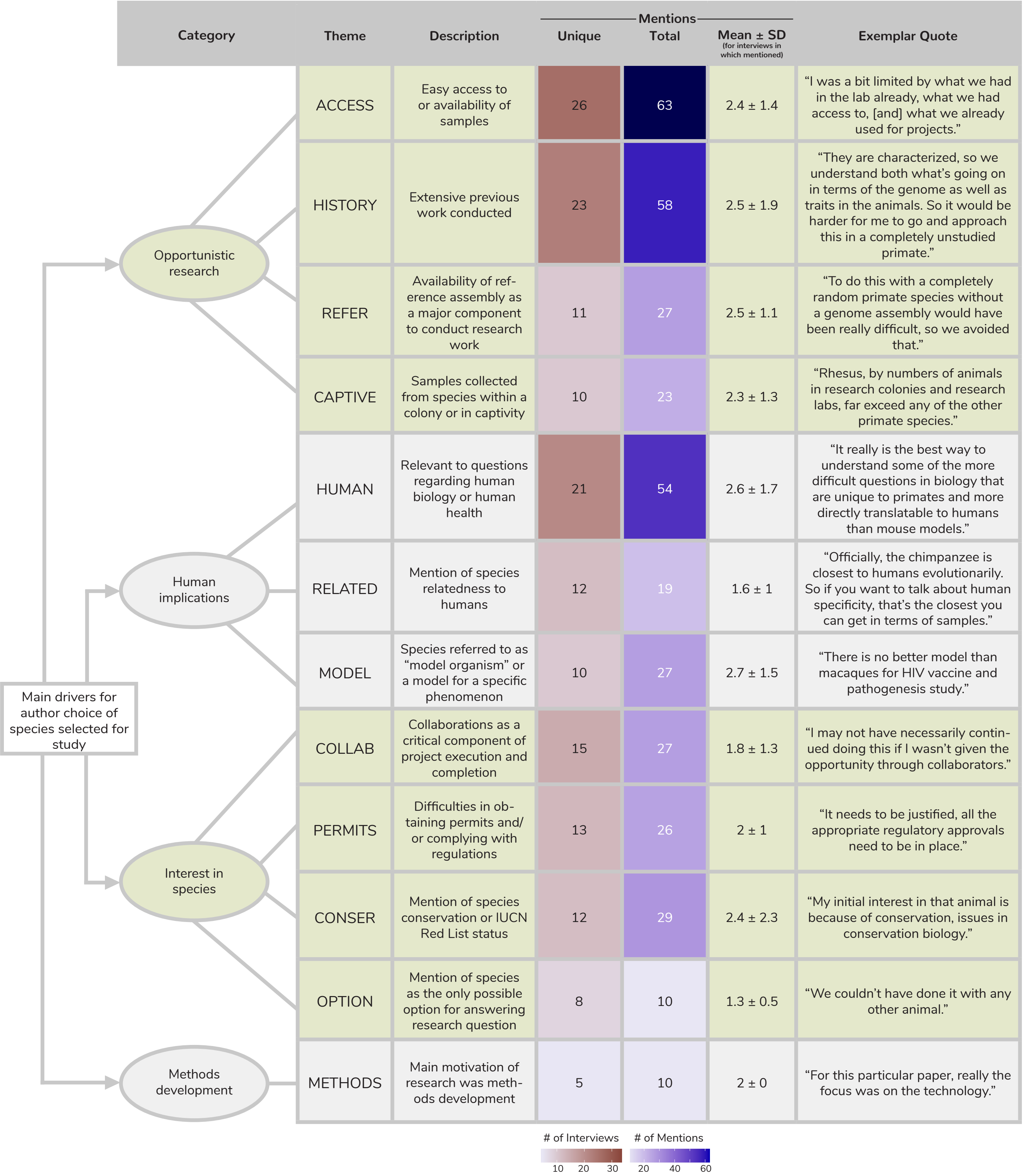
For the species with genomic data present we also performed a generalized linear model under a gaussian distribution to identify variables that best predicted the amount of genomic sequence data per species while accounting for interdependence among these factors (see Methods). Based on this model, non-medical research publications ($P$=7.62x$10^{-9}$), number of medical publications ($P$=3.74x$10^{-9}$), number of individuals in captivity ($P$=0.022), and species categorized as endangered relative to critically endangered within IUCN Red List status ($P$=0.026) were all significant predictors of the amount of genomic sequence data (P<2.2x$10^{-16}$; $R^2$=0.40).

Since the normality assumption for linear regressions was violated in our analyses of the full dataset, we also repeated these analyses on the subset of the dataset with species having at least some (i.e., non-zero) genomic data (Figure 2). In this analysis, total number of non-medical publications explained approximately 37% of the variation in genomic sequence data ($P$=6.44 x $10^{-12}$). Number of medically-focused publications was also significant and explained 27% of the variation within genomic sequence data ($P$=9.27 x $10^{-9}$). Frequency in captivity ($P$=0.00022), relatedness to humans ($P$=0.00106), and geographic range ($P$=0.0012) were all statistically significant but had limited explanatory power ($r^2$=0.12, $r^2$=0.092, and $r^2$=0.094, respectively). IUCN Red List status was neither significant nor explanatory ($P$=0.361, $r^2$=-0.0016). Our generalized linear model with the subset of species with genomic data present revealed that number of non-medical publications ($P$=0.00064) and relatedness to humans ($P$=0.024) were significant predictors of the amount of genomic sequence data ($P$=6.16x$10^{-8}$; $r^2$=0.40).

### *Author motivations in selecting species for study*

We randomly selected 300 unique SRA study numbers with the goal to contact the corresponding authors on papers for which these data were originally generated. We invited 216 authors (as some deposits did not have an associated publication and some individuals were corresponding authors on multiple papers) to participate in a semi-structured interview. In total, and after obtaining informed consent, we conducted 33 semi-structured interviews with first and/or corresponding authors on 33 publications that generated non-human primate genomic sequence data represented in our database. The 15.3% response rate is within the typical range for email/internet surveys (26). The list of interview questions is presented in Supplementary Table 2. We analyzed major themes arising from the interviews using a grounded theory approach (27). Twelve themes emerged from this analysis, grouped into four categories: opportunistic research, interest in species, human implications, and methods development (Figure 3).

*Opportunistic research:* Authors frequently mentioned selecting species with sampling and analytical feasibility in mind. Four themes were categorized under opportunistic research: ACCESS (present in n=26 of 33 total interviews), HISTORY (n=23), CAPTIVE (n=10), and REFER (n=11). Having access to high-quality existing samples and/or captive individuals was repeatedly mentioned as being important. Many authors also mentioned that species with extensive histories of prior work in turn helped them to more feasibly conduct analyses (e.g. due to the presence of a high-quality, annotated reference genome), or better contextualize their own results:



Main drivers for author choice of species selected for study

| Category | Theme | Description | Mentions | | Mean ± SD | Exemplar Quote |
|---|---|---|---|---|---|---|
| | | | Unique | Total | (for interviews in which mentioned) | |
| Opportunistic research | ACCESS | Easy access to or availability of samples | 26 | 63 | 2.4 ± 1.4 | "I was a bit limited by what we had in the lab already, what we had access to, [and] what we already used for projects." |
| | HISTORY | Extensive previous work conducted | 23 | 58 | 2.5 ± 1.9 | "They are characterized, so we understand both what's going on in terms of the genome as well as traits in the animals. So it would be harder for me to go and approach this in a completely unstudied primate." |
| | REFER | Availability of reference assembly as a major component to conduct research work | 11 | 27 | 2.5 ± 1.1 | "To do this with a completely random primate species without a genome assembly would have been really difficult, so we avoided that." |
| | CAPTIVE | Samples collected from species within a colony or in captivity | 10 | 23 | 2.3 ± 1.3 | "Rhesus, by numbers of animals in research colonies and research labs, far exceed any of the other primate species." |
| Human implications | HUMAN | Relevant to questions regarding human biology or human health | 21 | 54 | 2.6 ± 1.7 | "It really is the best way to understand some of the more difficult questions in biology that are unique to primates and more directly translatable to humans than mouse models." |
| | RELATED | Mention of species relatedness to humans | 12 | 19 | 1.6 ± 1 | "Officially, the chimpanzee is closest to humans evolutionarily. So if you want to talk about human specificity, that's the closest you can get in terms of samples." |
| | MODEL | Species referred to as "model organism" or a model for a specific phenomenon | 10 | 27 | 2.7 ± 1.5 | "There is no better model than macaques for HIV vaccine and pathogenesis study." |
| Interest in species | COLLAB | Collaborations as a critical component of project execution and completion | 15 | 27 | 1.8 ± 1.3 | "I may not have necessarly continued doing this if I wasn't given the opportunity through collaborators." |
| | PERMITS | Difficulties in obtaining permits and/or complying with regulations | 13 | 26 | 2 ± 1 | "It needs to be justified, all the appropriate regulatory approvals need to be in place." |
| | CONSER | Mention of species conservation or IUCN Red List status | 12 | 29 | 2.4 ± 2.3 | "My initial interest in that animal is because of conservation, issues in conservation biology." |
| | OPTION | Mention of species as the only possible option for answering research question | 8 | 10 | 1.3 ± 0.5 | "We couldn't have done it with any other animal." |
| Methods development | METHODS | Main motivation of research was methods development | 5 | 10 | 2 ± 0 | "For this particular paper, really the focus was on the technology." |

**Figure 3. Main drivers for author choice of species selected for study.** Each theme derived from our grounded theory analysis listed with its description, the number of unique interviews the theme appeared in (max 33), the total number of times each theme appeared, the mean and standard deviation across interviews where the theme was mentioned at least once, and an exemplar quote. Themes are organized by comprehensive categories that inform author choice when selecting non-human primates for research studies. The heatmaps depict the number of interviews each theme was present in and the number of total mentions for each theme.

*"[The taxa] had data, behavioral data…hormonal data, they had super early genetic markers. It was a system that had been studied from multiple different angles, multiple Ph.D. students and the like worked on it, so it was good because we had some hypotheses for what we'd find different between the two taxa."*

*Human implications:* The second major category that emerged from our interviews was human implications. Authors frequently mentioned that their research questions were directly relevant to questions regarding human biology, and more specifically human diseases. The most repeated theme in this category was a specific mention of species being used to understand human phenomena. This category contained three themes: HUMAN (n=21), RELATED (n=12), and MODEL (n=10). One author commented:

*"One of the primary goals of biomedical research is to develop animal models which will allow us to better understand the causes and potential treatments for human diseases. And so if you understand that genetics is important for human disease, and you want to model diseases in a non-human primate, then clearly understanding genetics and genetic differences among rhesus macaques is going to be important in a couple different ways."*

*Interest in species:* Another emergent category was author interest in the species. This category contained four themes: COLLAB (n=15), CONSER (n=12), PERMITS (n=13) and OPTION (n=8). Multiple authors mentioned conservation implications as either a priority and/or a byproduct of their research questions. Some researchers specifically selected certain species because of their IUCN Red List status (e.g., Critically Endangered, Endangered, etc.) and still other authors selected species primarily for different reasons but with conservation implications also in mind. Many of the same (and other) authors also frequently mentioned difficulties in securing permits and ensuring that local governmental regulations were properly followed, especially when studying protected species. Collaborations with other research groups and international scientists was, in some cases, critical for the continuation and completion of the research work, as one author commented:

*"I may not have necessarily continued doing this if I wasn't given the opportunity through collaborators."*

*Methods development:* The final category only consists of one theme, METHODS (n=5). These authors mentioned that, in some cases, their research study was driven primarily by interest in developing a new technology, bioinformatics pipeline, and/or wet lab method. They then used non-human primate samples that were readily available in order to most efficiently develop, evaluate, and report on their method.

## DISCUSSION

Our findings contribute to the body of work supporting the idea that certain taxonomic groups, in our case individual non-human primate species, are studied more extensively than others. Specifically, certain non-human primates appear to have been selected for massively-parallel genomic sequencing studies primarily because their biological samples were available and accessible, they had extensive histories of prior published work, and they were relevant to questions pertaining to human biology, especially when investigating human diseases. As our closest living relatives, non-human primates are researched extensively to help us better understand fundamental questions regarding our evolutionary history and the diseases that plague us (17, 28).

Our qualitative results aligned closely with those from our quantitative analysis. The qualitative data in particular clearly illustrate how the relative ease of studying certain species already widely used as biomedical models for human disease may further perpetuate future data generation disparities. As one author described:

*"If we didn't have macaques, we really wouldn't have a good model for HIV/SIV, and that would be a huge problem. It's not that it's a bad thing that we focused on rhesus monkeys or cynomolgus macaques, but it has consequences just because it might be that stump-tailed macaques are a great model for Parkinson's disease, but we don't know that because we lost all the stump-tailed macaques. It might be that a particular form of spider monkey is a fantastic model for high blood pressure, but we don't know that because there aren't spider monkeys in research colonies that people can study. It's frustrating to me that there are probably outstanding models [of] human disease that we will never discover because we don't have access to those animals. Now that's unfortunate, but you can understand why the NIH can't pay for colonies of 1000 of every different species of primates just because maybe 20 years from now, somebody is going to have a need for those animals.* ***Decisions are driven by resources, by how much you can spend, and you put your resources where you think they will do the most good today. But that sometimes has long term consequences."*** [emphasis added]

We are uncertain of and unqualified to help define the best approach for developing and funding future research on non-human primate models for human disease. It is possible that the current system could be the most efficient and broadly effective. Still, it is important to recognize the manner by which this system further exacerbates patterns of taxonomic unevenness in research – including via new rounds of studies that are not themselves necessarily biomedically-motivated – due to enhanced sample accessibility and opportunities to more rapidly advance new research given existing backbones of knowledge on which to build. This phenomenon could in turn constrain opportunities for research on non-model organisms for evolutionary biology, behavioral ecology, or conservation purposes. The same consideration likely applies to other taxonomic groups other than primates. That is, even in the absence of model organisms, widely apparent biases for particular field sites (29), geographic regions (30), habitat types (31), species "charisma" (5), and societal preferences (32) likely impact taxonomic choices in successive research planning processes.

Insights from the qualitative component of our study into the processes that shape scientists' decision-making may aid funding agencies, scientific societies, and research consortia whose goals are not fully aligned with the current scientific data generation landscape. Specifically, our study demonstrates that scientific research is goal-oriented and that study organism selection is understandably based to a large degree on feasibility as well as the extent of previous published work and resources. Thus, research-oriented institutions may benefit from taking steps to increase access to biological samples and to develop and disseminate initial genomic-scale data and resources for targeted taxonomic groups. Other researchers would then be more likely to select these species for their own studies (even those funded by other agencies) and create new knowledge and further resources to the positive-reinforcement benefit of all.

## METHODS

### *Quantitative Data and Analyses*

*Non-human primate species list:* We generated a list of all non-human primate species using the IUCN Red List and supplemented using All the World's Primates by Rowe and Myers (25). 13 species were found only within the IUCN Red List, 82 species were found only within Rowe and Myers, 396 species were found in both sources, and 29 species were found in both sources but under synonymous species names. Using these two sources, we arrived at a list of 519 species. For the purposes of this study, we collapsed any subspecies under a single species name. All data recorded for each species can be found in Supplementary Dataset 1.

*Genomic and transcriptomic data:* We used the Sequence Read Archive (SRA), a public repository for biological sequence data run by the National Center for Biotechnology Information (NCBI), to record the total amount of genomic and transcriptomic sequence data for all non-human primates (hereon called genomic data) (https://www.ncbi.nlm.nih.gov/sra). SRA was searched using the broad taxonomic terms while purposefully excluding any data on humans. The search terms were as follows:

(primate OR primates) AND (genomic or genome or transcriptome or transcriptomic)
NOT (Homo sapiens)

The final list of all deposits was then reviewed to remove any species that were misclassified as primates and extinct primate taxa to arrive at a full list of genomic or transcriptomic data deposits for all non-human primates. Ambiguous deposits that were not species-specific (e.g. deposits that were listed as only "Rhinopithecus") were removed from the dataset. Genomic data from hybridized species were also removed from the dataset. All deposits that were removed from the study are listed in Supplementary Table 3. The total amount of genomic data for each species was gathered on August 16, 2018.

*Non-medical publications:* Data for this variable were collected using a similar methodology as in Wiens, 2016, described below (33). We searched for the number of publications using either the species scientific name or a common name for the non-human primate species. An example of the search criteria for each of these variables is shown below.

Research intensity: TS=(("Allocebus trichotis") OR ("Hairy-eared Dwarf Lemur"))

The total number of publications was recorded for all species within the full non-human primate species list. These data were collected in January and February of 2018. We then subtracted the number of medically-focused publications from the number of total publications to compute this variable.

*Importance in medical research:* As discussed previously, non-human primates are frequently used in the testing of vaccinations, the study of SIV progression to inform HIV studies, and for other research projects that have medical relevance. Importance in medical research was gauged using the number of papers published within Web of Science for each species under the Web of Science Category: Medicine, Research, & Experimental using either the scientific name or one of the common names for each species. An example of the search criteria is shown below:

WC=(Medicine, Research & Experimental) AND TS = (("Trachypithecus selangorensis") OR ("Selangor Silvery Langur"))

The total number of publications for each species identified using this search was recorded. These data were collected in January and February of 2018.

*Frequency in captivity:* We hypothesized that the more abundant a species was in captivity, the more opportunity there may have been for the collection of high-quality samples for genomic data analysis. To test this hypothesis, we used the number of individuals within a species found in captivity. These data were obtained from the Species 360 ZIMS database, an online repository of species currently held in captivity within institutions partnered with Species 360 around the world (https://www.species360.org/). Access to the Species 360 database was granted by the Duke Lemur Center. Each species was searched within the ZIMS database and the total number of individuals in captivity was recorded.

*Relatedness to humans:* We were interested in testing whether genomic data for species more closely related to humans were generated at disproportionately higher rates. Thus, for each species we recorded how many millions of years since they last shared a common ancestor with humans using the estimates generated in dos Reis et al (34).

*Geographic distribution:* Species with more extensive geographic distributions may be more easily accessible and therefore more frequently studied by scientists and conservationists. Geographic distribution for each available primate species was obtained from spatial data provided by the IUCN Red List. The spatial data can be accessed via the IUCN Red List online portal (https://www.iucnredlist.org/). The data were imported into ArcGIS and projected onto the Cylindrical Equal Area (sphere), where we then calculated the area of their range in square kilometers. These data were obtained on February 7, 2018.

*IUCN Red List status:* We also sought to test whether there is a relationship between an organism's perceived risk of extinction and the level of genomic sequence data generation. Red List status was obtained through the IUCN Red List of Threatened Species online portal (https://www.iucnredlist.org/). If Red List Status could not be obtained via the online portal, the status was recorded from All the World's Primates by Rowe and Myers (25). The possible categories for the Red List are as follows: Data Deficient (DD), Least Concern (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), and Critically Endangered (CR). These data were collected in February of 2018.

*Activity pattern:* Species that are diurnal or cathemeral (flexible day/night activity) may be easier to study than those that employ a nocturnal lifestyle. We hypothesized that species that were nocturnal would have less genomic data available, when controlling for the number of species within each of these categories. These data were collected using All the World's Primates by Rowe and Myers (25). The four possible categories were: diurnal, nocturnal, cathemeral, or N/A if no information was available.

*Statistical analyses:* Statistical analyses were conducted using RStudio. Code for all analyses performed is available via the GitHub Repository https://github.com/maggiehern/PrimateGenomeProject. The distribution of the amount of genomic data among non-human primates was right skewed and was normalized via log base 10 transformation prior to analyses. All continuous independent variables were also log transformed. Logistic regressions and generalized linear models were performed for the entire dataset and the subset of species with genomic data available. A list of models and their results are reported in Supplementary Table 4.

### *Qualitative Data and Analyses*

*Data collection:* We downloaded the metadata for all non-human primate deposits in the NCBI Sequence Read Archive (SRA). A list of SRA study numbers was generated by removing any SRA study numbers that had multiple entries. We randomly sampled 300 unique SRA study numbers identified in the linked publication using both the SRA study number and any other related accession number reported within SRA (this included the BioProject number, Gene Expression Omnibus deposition number, etc.) and confirmed that the paper represented the original generation of the data.

The human subjects research component of this study was approved by Penn State's Institutional Review Board (IRB) under the study number STUDY00008181. We contacted each of the corresponding authors of these studies to invite them to participate in a semi-structured interview. There were several papers that had the same corresponding author. In this case, a single email was sent listing all the papers we were requesting the authors to discuss during the interview. In total, we contacted 216 authors. There was one case where a corresponding author was not available for an interview and referred us to the first author of the paper, who consented to be interviewed. Authors were asked to participate in a 30-minute interview following a semi-structured interview format using the questions listed in Supplementary Table 2. Interviews were conducted via Skype, Zoom, phone, or in person at the participant's convenience. One survey was administered as a word document and filled out by the participant. Participants consented to being recorded prior to the start of the interview and were recorded using a Roland WAVE/MP3 Recorder R-05 and later transcribed.

*Qualitative data analysis:* We conducted a grounded theory analysis using all interviews conducted for this study (27). Grounded theory analysis is an inductive approach to understanding qualitative data where researchers read over interview transcripts (or other texts) several times while developing successively more detailed levels of coding in order to understand major themes that emerge from the texts. In this case, we transcribed and then read over transcripts of the interviews to identify and code the major themes that emerged from the responses of the participants. Interviews were then reviewed again to record the frequency of each identified theme across all 33 interviews. This process was done in the software NVIVO, used for qualitative research data collection and organization. Major themes were then grouped into larger categories and subsequently compared to the factors identified in the quantitative portion of this project.

## ACKNOWLEDGMENTS

We thank the participants of this research study that took the time out of their days to speak with us and contribute to our project. Additionally, we thank Nicholas Triozzi, Dylan Davis, Kathleen Grogan, Christina Bergey, Julie White, and Stephanie Marciniak for their discussion and analytical advice. We also thank Max Fancourt from the IUCN Red List and Sarah Zehr from the Duke Lemur Center for their assistance in data access. This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant No. DGE1255832 (to M.H.) and by NSF grant BCS-1554834 (to G.H.P.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

1. X. Bonnet, R. Shine, O. Lourdais, Taxonomic Chauvinism. *Trends Ecol. Evol.* **17**, 1–3 (2002).
2. J. A. Clark, R. M. May, Taxonomic Bias in Conservation Research. *Science (80-. ).* **297**, 191–192 (2002).
3. J. Troudet, P. Grandcolas, A. Blin, R. Vignes-Lebbe, F. Legendre, Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).
4. W. R. T. Darwall, *et al.*, Implications of bias in conservation research and investment for freshwater species. *Conserv. Lett.* **4**, 474–482 (2011).
5. F. Ducarme, G. M. Luque, F. Courchamp, What are "charismatic species" for conservation biologists? *Biosci. Master Rev.* **10**, 1–8 (2013).
6. J. Lorimer, Nonhuman charisma: which species trigger our emotions and why? *Ecos-British Assoc. Nat. Conserv.* **27**, 20 (2007).
7. S. Leonelli, R. A. Ankeny, What makes a model organism? *Endeavour* **37**, 209–212 (2013).
8. H. Ellegren, Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29**, 51–63 (2014).
9. C. E. Hinchliff, *et al.*, Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12764–12769 (2015).
10. E. K. Bors, S. Herrera, J. A. Morris, T. M. Shank, Population genomics of rapidly invading lionfish in the Caribbean reveals signals of range expansion in the absence of spatial population structure. *Ecol. Evol.* **9**, 3306–3320 (2019).
11. M. Morgado-Santos, M. F. Magalhães, L. Vicente, M. J. Collares-Pereira, Mate choice driven by genome in an allopolyploid fish complex. *Behav. Ecol.* (2018) https:/doi.org/10.1093/beheco/ary117 (February 6, 2020).
12. A. D. Baxevanis, A. Bateman, The importance of biological databases in biological discovery. *Curr. Protoc. Bioinforma.* **2015**, 1.1.1-1.1.8 (2015).
13. A. Estrada, *et al.*, Impending extinction crisis of the world's primates: Why primates matter. *Sci. Adv.* **3**, 1–16 (2017).
14. M. Gross, Primates in peril. *Curr. Biol.* **27**, R573–R576 (2017).
15. J. D. Harding, Nonhuman Primates and Translational Research: Progress, Opportunities, and Challenges. *ILAR J.* **58**, 141–150 (2017).
16. C. Abee, K. Mansfield, S. Tardif, T. Morris, *Nonhuman Primates in Biomedical Research* (Elsevier Inc., 2012) https:/doi.org/10.1016/C2009-0-01851-0.
17. K. K. A. Van Rompay, Tackling HIV and AIDS: contributions by non-human primate models. *Lab Anim. (NY).* **46**, 259 (2017).
18. D. Van Dam, P. P. De Deyn, Non human primate models for Alzheimer's disease-related research and drug discovery. *Expert Opin. Drug Discov.* **12**, 187–200 (2017).
19. I. Jarić, D. L. Roberts, J. Gessner, A. R. Solow, F. Courchamp, Science responses to IUCN Red Listing. *PeerJ* **5**, 1–11 (2017).
20. Z. M. Brooke, J. Bielby, K. Nambiar, C. Carbone, Correlates of research effort in carnivores: Body size, range size and diet matter. *PLoS One* **9**, 1–10 (2014).
21. A. J. McKenzie, P. A. Robertson, Which species are we researching and why? A case study of the ecology of British Breeding Birds. *PLoS One* **10**, 1–16 (2015).
22. P. E. Rose, J. E. Brereton, L. J. Rowden, R. L. de Figueiredo, L. M. Riley, What's new from the zoo? An analysis of ten years of zoo-themed research output. *Palgrave Commun.* **5**, 1–10 (2019).
23. C. Meyer, W. Jetz, R. P. Guralnick, S. A. Fritz, H. Kreft, Range geometry and socio-economics dominate species-level biases in occurrence information. *Glob. Ecol. Biogeogr.* **25**, 1181–1193 (2016).

24. M. Dos Reis, *et al.*, Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Syst. Biol.* **67**, 594–615 (2018).
25. N. Rowe, M. Myers, *All the world's primates* (Charlestown: Pogonias Press, 2016).
26. N. Sappleton, F. Lourenço, Email subject lines and response rates to invitations to participate in a web survey and a face-to-face interview: the sound of silence. *Int. J. Soc. Res. Methodol.* **19**, 611–622 (2016).
27. B. Glaser, A. Strauss, Grounded theory: The discovery of grounded theory. *Sociol. J. Br. Sociol. Assoc.* **12**, 27–49 (1967).
28. H. Svardal, *et al.*, Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nat. Genet.* **49**, 1705–1713 (2017).
29. M. Bezanson, A. McNamara, The what and where of primate field research may be failing primate conservation. *Evol. Anthropol.* **28**, 166–178 (2019).
30. M. R. Donaldson, *et al.*, Taxonomic bias and international biodiversity conservation research. *Facets* **1**, 105–113 (2016).
31. J. J. Lawler, *et al.*, Conservation science : a 20-year report card. *Front. Ecol. Environ.* **4**, 473–480 (2006).
32. I. Jarić, *et al.*, On the overlap between scientific and societal taxonomic attentions-Insights for conservation. *Sci. Total Environ.* **648**, 772–778 (2019).
33. J. J. Wiens, Climate-Related Local Extinctions Are Already Widespread among Plant and Animal Species. *PLoS Biol.* **14**, e2001104 (2016).
34. M. dos Reis, *et al.*, Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Syst. Biol.* (2018) https:/doi.org/10.1093/sysbio/syy001.

## FIGURE CAPTIONS

**Figure 1. Megabases of genomic data by genus across the order Primates and for the five species with the most genomic data.** A) Phylogeny of the order Primates with dark purple indicating more genomic data per genus and white indicating little to no genomic data. Paraphyletic genera are denoted with an asterisk. A complete list of genera is provided in Supplementary Table 1. Phylogeny adapted from Dos Reis et al. (2018). B) The five species with the most genomic data and the cumulative percentage of the total amount of non-human human primate sequence data represented by these taxa. Credit to T. Michael Keesey and Tony Hisgett for the chimpanzee image, under license https://creativecommons.org/licenses/by/3.0/.

**Figure 2. Linear regressions for entire dataset and subset of species with genomic data.** Linear regressions for six variables used within the study, non-medical papers published (A), importance in medical research (B), relatedness to humans (C), geographic range (D), frequency in captivity (E), and IUCN Red List status (F). For each, the purple line represents the linear regression for the subset of species with genomic data available, while the green line represents the linear regression for all species within the dataset.

**Figure 3. Main drivers for author choice of species selected for study.** Each theme derived from our grounded theory analysis listed with its description, the number of unique interviews the theme appeared in (max 33), the total number of times each theme appeared, the mean and standard deviation across interviews where the theme was mentioned at least once, and an exemplar quote. Themes are organized by comprehensive categories that inform author choice when selecting non-human primates for research studies. The heatmaps depict the number of interviews each theme was present in and the number of total mentions for each theme.

**Supplementary Dataset 1. PGP_Data_FINAL.csv** All quantitative data collected and analyzed for the study. DOI: https://doi.org/10.26207/j7nd-ka67.

**Supplementary Figure 1. Boxplots comparing species with and without any genomic data for each tested variable.** The box extends from the first to the third quantile. The horizontal line within each boxplot represents the median value. The whiskers represent the lower and upper extreme value limits. The black dots represent outliers.

**Supplementary Figure 2. Violin plots of per-species genomic data by activity pattern.** Violin plot width corresponds to the density of species, which is also depicted via heatmap.

**Supplementary Table 1. List of paraphyletic generic names found in genomic databases that were collapsed into a single genus for construction of Figure 1**. The single genus names listed in the first column were used in Figure 1 (denoted by asterisks in the Figure) to represent data combined for multiple genera now believed to be paraphyletic, listed in the second column.

**Supplementary Table 2. List of interview questions.** Interview questions were asked in the order presented in this table. At times, additional probing questions were asked regarding the answers provided for each question. However, overall, each interview followed the structure presented here.

**Supplementary Table 3. SRA deposits omitted from genomic dataset.** A complete list of the SRA deposits that were removed from the dataset (see Methods), including the total amount of genomic data under each deposit name. Deposits that were not identified at the species level or that came from known species-hybrid individuals were omitted.

**Supplementary Table 4. Analytical models performed**. A list of all models used within the study, including data used, test performed, distribution of data, variables included, AIC values, p-values, and $r^2$ values where applicable.