**Motivation**

The fossil record is heavily debated in the field of anthropology and is crucial for the understanding of a variety of topics related to human nature, evolution, and aggression [1]. Databases exist for early human and non-human fossils [2][3]; however, they lack complete references, complete information, and are not conducive for direct data analysis. This project aims to solve this issue by creating an open-source database on human fossil remains that contains core fossil information and references. The nature of this project was chosen to enhance and work upon my data science skills, particularly in relation to my project involving the creation of a global freshwater fish diet database and my upcoming summer work involving the analysis of marine protected areas.

**Methodology**

Fossil searchers were based on the Smithsonian Institute’s Human Origins website [1] and lists from Wikipedia [4]. This was only for reference and all core fossil data (site data, taxonomic information, discovery data, chronological data, anatomy data, citation data) were derived from primary sources found through targeted literature searches. The only exceptions to this were some discovery data (discoverer, date of discovery) derived from either the Smithsonian website [2] or efossils.com (a digital catalog on a limited number of human evolution fossils ran by The University of Texas at Austin) [3]. This was only done if discovery data were hard to find, gated behind a paywall, or not accessible through USD’s library. Fossil data derived through these sites are noted in the database itself.

The relevant core fossil data were extracted manually from as many papers as necessary – usually 1, but occasionally 2+ - and filled into a tidy .csv file. Example:

Citation data were stored in a separate .csv due to the potential for a given fossil requiring multiple references. Instead, a citation group number was given to each fossil. This can be entered into simple, provided R code to output the relevant citations. Latitude and longitude data were added post hoc through geocoding in R through google API. There is a chance for error in this respect as I did not have time to fully check the API's accuracy. A variety of fixes were done manually and through code in R, mostly related to minor naming issues and post hoc standardization.

**Results & Discussion**

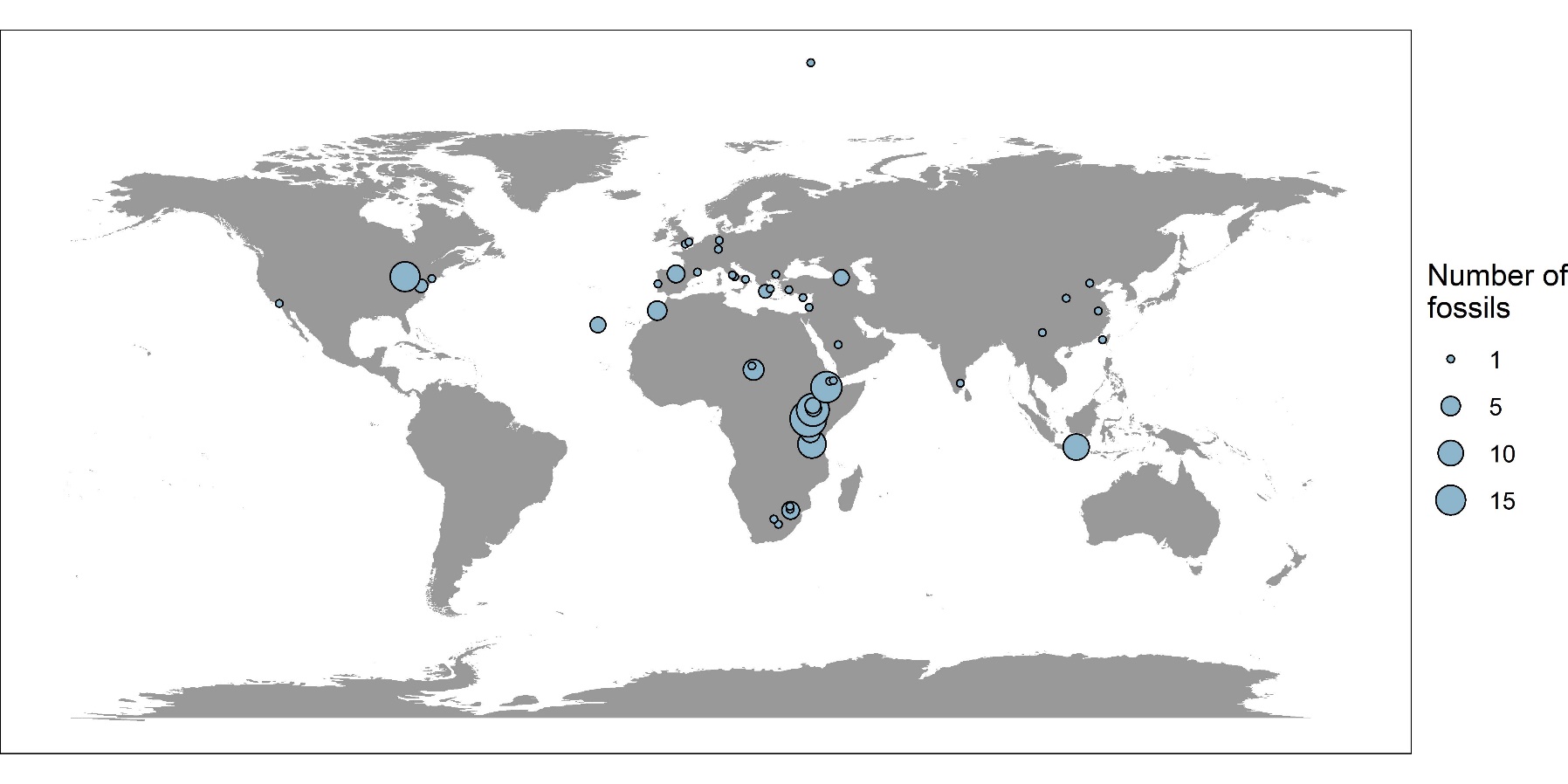
 My database contains information from 102 papers on 176 distinct fossils across 102 unique sites, 29 different species, and a time range of 0.10 – 7.24 mya. Visualizations for the distribution of fossil species taxonomically (Fig. 1) and geographically (Fig 2) can be seen below:

Figure 1. A map of the distribution of the 176 fossils in my database across 102 distinct sites. Several locations (e.g. far North location not on land) seem off, probably due to google API geocoding mislocation.

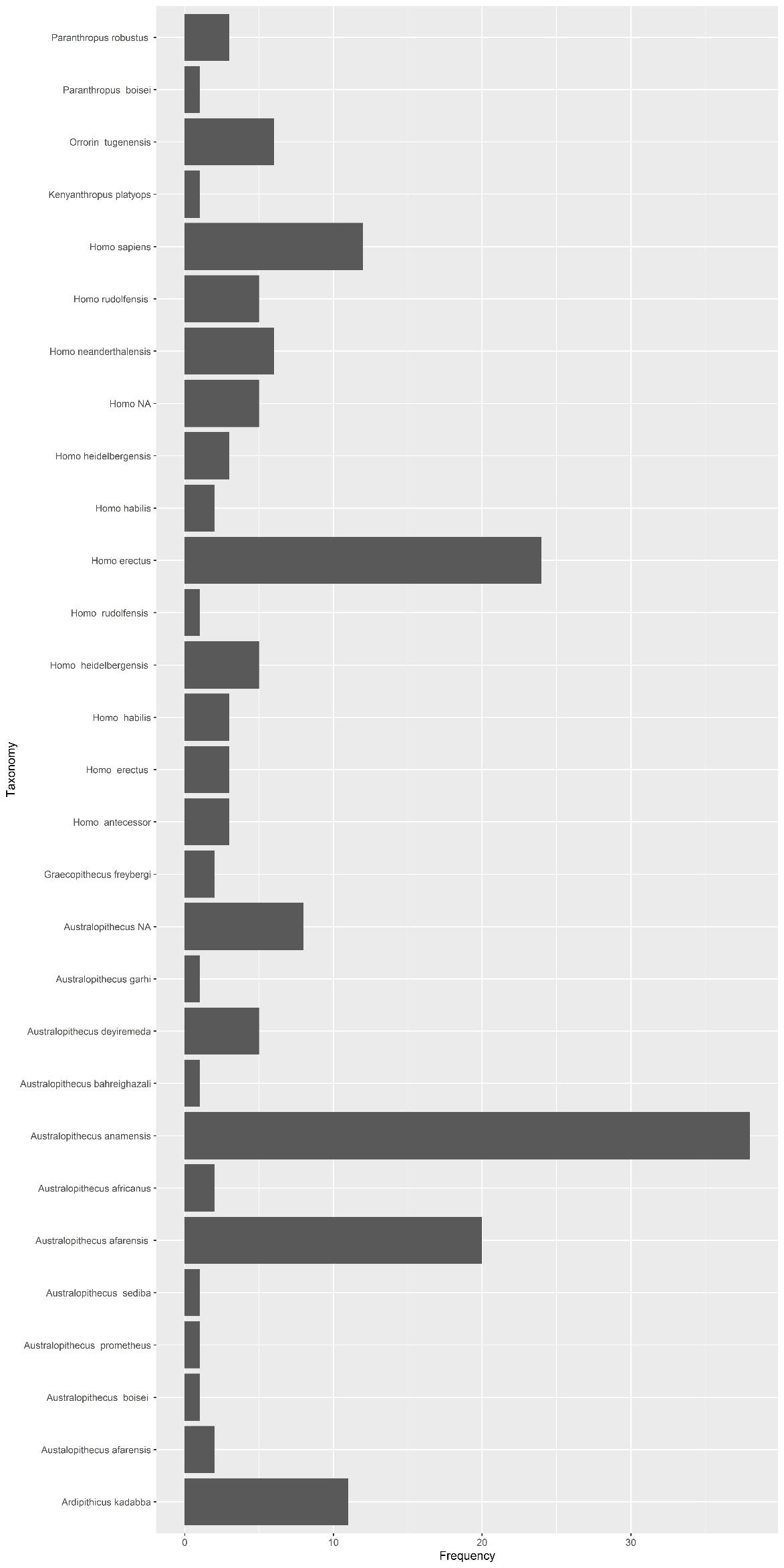


Figure 2. Bar graph of the taxonomic distribution of the 29 species of fossils in my database.

Obviously, I did not find every human evolution fossil globally. I didn’t fully make it through the lists of fossils given by Wikipedia and the Smithsonian. My ability to do so was limited by the scope of individual work, my knowledge of the field, my connections, and my ability to commit fully to the project as a busy undergraduate. As a result, the database is biased towards fossils found in the Middle Paleolithic and later. Additionally, the usage of my reference sites probably biases the database towards unique and important fossils. The idea going forward, if the project were continued, would be to establish a base with reference sites to find additional fossils through the required literature searches, additional knowledge, and citation analysis. This would greatly reduce the two biases mentioned above. However, that would still not be enough to finish this project.

Anthropological data are very decentralized compared to other fields, at least based on my experiences with other large-scale data projects. What I mean by this is that aspects of core fossil information may be contained in anything from publications, local or national news articles, universities, letters, personal notes, books that cost money, or strictly in physical locations (instead of typically just the literature for many other fields). Further complicating this is the abundance of fossils found before our current digital era. Often the only reachable or findable source of information can be citations away from the original source – with others not being accessible at all. Pairing this with the decentralized nature of the field makes accurate data extraction time-consuming, hard, and in many cases impossible. The field is also ripe with controversy and conflicting views, adding further layers to the difficulty and accuracy of finding and extracting data. If a given fossil is not seen as important, it also may just not have a standalone publication source or complete information.

Due to the reasons above, I do not think this project is fully completable on an individual level. To finish this project on even a temporally or geographically narrow range would require one to inquire data from many scientists and institutions. There is just too much to decode and, in many cases, literally no information available. Standardization of fossil sharing practice can go a long way to address several of these issues, but collaboration on a global scale would be needed to digitize and resurface the massive amount of ‘hidden’ data. What I do know is that the authors of the papers I referenced collectively do know how to find that ‘hidden’ data – probably from a mixture of connections, better access to the papers, and decades of experience in the field. Optimistically, then, I think this collaboration would only need to be done by several key individuals who are ‘in on the knowledge.’ Due to the likely variability of personal willingness to share data, I suggest incorporating a higher level of data sharing and transparency innately into grants and research institutions in the field as a whole – for what is science if the giants are too prideful to let people stand on their shoulders?

**Github & Information for Using the Database**

All material for the project can be found at <github>. This includes the database, all code paired with explanation, and graphs. Ci

Works Cited

[1]: Wrangham, R. W. (2018). Two types of aggression in human evolution. *Proceedings of the National Academy of Sciences*, *115*(2), 245-253.

[2]: https://humanorigins.si.edu/evidence/human-fossils/fossils

[3]: http://www.efossils.org/

[4]: https://en.wikipedia.org/wiki/List\_of\_human\_evolution\_fossils