IJDC | *Conference Paper*

Artificial Intelligence Assisted Curation of Population Groups in Biomedical Literature

Latrice Landry
University of Pennsylvania

Mary Lucas
Drexel University

Anietie Andy
Howard University

Ebelechukwu Nwafor
Villanova University

Abstract

Curation of the growing body of published biomedical research is of great importance to both the synthesis of contemporary science and the archiving of historical biomedical literature. Each of these tasks has become increasingly challenging given the expansion of journal titles, preprint repositories and electronic databases. Added to this challenge is the need for curation of biomedical literature across population groups to better capture study populations for improved understanding of the generalizability of findings. To address this, our study aims to explore the use of generative artificial intelligence (AI) in the form of large language models (LLMs) such as GPT-4 as an AI curation assistant for the task of curating biomedical literature for population groups. We conducted a series of experiments which qualitatively and quantitatively evaluate the performance of OpenAI's GPT-4 in curating population information from biomedical literature. Using OpenAI's GPT-4 and curation instructions, executed through prompts, we evaluate the ability of GPT-4 to classify study 'populations', 'continents' and 'countries' from a previously curated dataset of public health COVID-19 studies.

Using three different experimental approaches, we examined performance by: A) evaluation of accuracy (concordance with human curation) using both exact and approximate string matches within a single experimental approach; B) evaluation of accuracy across experimental approaches; and C) conducting a qualitative phenomenology analysis to describe and classify the nature of difference between human curation and GPT curation. Our study shows that GPT-4 has the potential to provide assistance in the curation of population groups in biomedical literature. Additionally, phenomenology provided key information for prompt design that further improved the LLM's performance in these tasks. Future research should aim to improve prompt design, as well as explore other generative AI models to improve curation performance. An increased understanding of the populations included in research studies is critical for the interpretation of findings, and we believe this study provides keen insight on the potential to increase the scalability of population curation in biomedical studies.

Submitted date 20 February 2024 ~ *Accepted* 22 February 2024

Correspondence should be addressed to Latrice Landry,. Email: latrice.landry@pennmedicine.upenn.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction: Bias Considerations Resulting from the Lack of Diversity in Biomedical Research

Biomedical research plays a central role in the evidence base for developing 'precision medicine' therapeutics and technologies. However, recent studies have reported a lack of diversity in sample populations included in research, with a large portion of the global population being mostly excluded (Bustamante, Burchard, & De la Vega, 2011; Landry et al., 2018). This lack of diversity has important implications for the translation of biomedical discoveries to global populations. The myopic nature of this problem has contributed to its continuance. Visibility, discussion and evaluation of the problem was enabled through documentation of population groups in individual studies via manual curation of digitized records of biomedical literature (Buniello et al., 2019; Wojcik et al., 2019; Sollis et al., 2023). Curation plays a key role in the field of genetics which collectively curates biomedical literature for genetic information, which is organized and accessed in digital repositories (Manotas, Rivera, & Sanabria–Salas, 2023). However, the inclusion of population groups has not been a consistent component of the field's curation practices. The recent highlighting of the disparities in research and lack of documentation of population groups has demonstrated a need to curate 'populations' in biomedical literature.

The call for consistent documentation of population groups in biomedical literature presents an additional need for a developed ontology around populations. "Race", "Cultural Ancestry", "DNA Ancestry" and "Ethnicity" are some of the more common terms used by Western populations (Byeon et al., 2021). However, many of these terms remain problematic in their controversy and the lack of consistency in use across communities and geographic locations (Mauro et al., 2022). In addition to population ontologies, the field also requires technology to alleviate the burden of manual curation, which is labor- and cost-intensive (Ravichandran et al., 2019). Many articles do not include a description of participants by demographic category in the abstract, requiring full curation of manuscripts and supplements to identify information. In some cases, the information does not exist in either the abstract or the manuscripts, resulting in a need for curators to contact authors to obtain the necessary information. The amount of time required for curation could potentially be reduced with support from technology and developed automated resources.

Comparatively, curation of genetic information has become increasingly automated with the help of various laboratory tools aimed at allowing clinical geneticists quick access to knowledge synthesis for assistance in clinical decision-making (Lee et al., 2018). However, this integration of technology into the curation of genetic literature has yet to address population descriptors. Here we present the development of a framework for the use of artificial intelligence (AI) through large language models (LLMs), exampled with GPT-4, to aid in the identification and classification of population groups in biomedical literature.

Methods

The proposed framework for use of AI to assist in the curation of populations in biomedical literature includes identification of existing population curated datasets, selection of the LLM for assistance, and development of the LLM prompt. For this study, we utilize a curated public health COVID dataset which comprises publications in the CDC Public Health Genomics Knowledgebase (PHGKB) from 2020 to 2022, extracted in September of 2022. There are 1000 PubMed IDs (a unique identifier for indexed articles in the National Library of Medicine's publication database) in the manually curated dataset, out of which we selected a sample of 200 to evaluate our framework.[1]

[1] PubMed: https://pubmed.ncbi.nlm.nih.gov

The three major components of our framework are: 1) information extraction; 2) data annotation; and 3) evaluation. Figure 1 provides a high-level overview of our framework. These components are required features of the LLM workflow, which interacts with the developed approaches for prompt generation and evaluation of results. The processes are designed to be dynamic, with performance results for classification of 'population', 'country' and 'continent' impacting the automated LLM workflow and prompt generation approaches and vice versa. The accessibility and performance of the GPT-4 language model and OpenAI's GPT-4 tool were key factors in their selection for development of this framework.

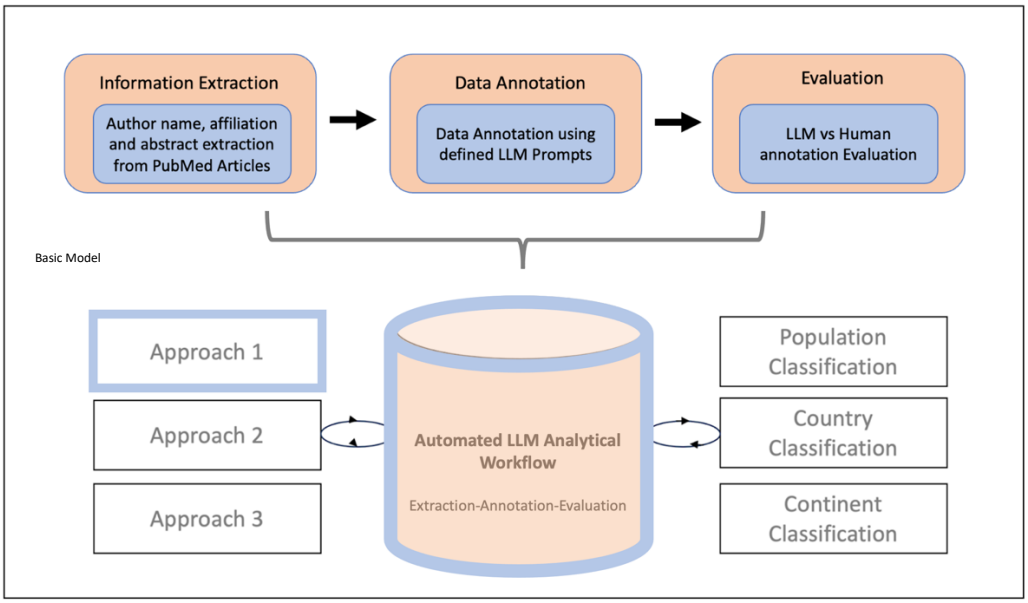

Figure 1. LLM data annotation workflow and analytical framework.

Automated LLM Analytical Workflow

The automated LLM analytical workflow includes information extraction, data annotation and evaluation (described below).

Information Extraction

Our framework provides a streamlined approach for the automatic extraction of critical data from biomedical literature using the unique PubMed ID associated with each article. Key information such as the author's name, their institutional affiliations, and the article's abstract are extracted using the BeautifulSoup Python library, which allows for seamless parsing of HTML content.

Data Annotation

We utilize OpenAI's GPT-4 as our LLM of choice as it provides state of the art performance for most natural language processing tasks. However, it is important to note that due to the modular nature of our framework, it is compatible with and can be used with any other LLMs of choice. We supply the GPT-4 model with specific prompts that instruct it to infer the population, country and continent information of the study participants from a given biomedical publication.

Evaluation

We compared the descriptions of study 'populations', 'country' and 'continent' curated by OpenAI's GPT-4 and by human curators. For zero-shot and multi-shot processes, we utilized the following evaluation approach:

1. We examined the performance of the developed LLM approach by comparing the reported results of GPT-4-generated annotation against human annotations. Our evaluation consists of two main techniques:

    a. **Exact match:** This method involves the use of direct string comparison. We implement a binary classifier in which any difference between categories annotated by human and the LLM results in a score of 0, indicating a mismatch. Identical matches across LLM and human annotation are awarded a score of 1, indicating a perfect match.

    b. **Fuzzy match**: This approach evaluates the similarity between word sequences within sentences using Levenstein distance (Logan et al., 2023). It compares the smallest number of character changes required to change one word into another. The resulting outcome generates a similarity score ranging from 0 to 100, where a score closer to 0 signifies greater dissimilarity, and a score closer to 100 indicates higher similarity. For instance, using this approach, the terms "Asia" and "Asian" will appear similar, with a high similarity score. For our analysis, we adopted a binary evaluation criterion with a threshold score of 80 such that each pair of human–GPT-4 annotations from that exhibits a score of 80 or above are deemed to be similar for the purpose of our analysis.

Description of Experimental Approaches

To properly evaluate the effectiveness of using LLMs in curation of population groups, we employ a comprehensive evaluation methodology which utilizes four distinct approaches for data annotation.

1. A. Zero-shot [basic] approach with abstract only: This approach involves the use of LLMs for processing structured data using abstracts only without any contextual information. This approach is seen as the 'basic' or baseline approach. Results from the 'basic' approach were used to inform the rules for Approach III, multi-shot with prompt optimization.

    B. Zero-shot with author details: This approach enhances the basic approach by incorporating the author's information, such as the author's name and affiliation, in addition to the abstract.

2. Multi-shot approach with annotated examples: This approach leverages extracted biomedical publication data supplemented by human annotated examples. These examples serve to guide the LLMs with the goal of refining and improving the data curation process.

3. Multi-shot approach with prompt optimization: Building on the multi-shot approach, this approach introduces prompt optimization to further enhance model performance in the annotation task. The population curation prompt was optimized with a set of rules (see Table 1).

Table 1. Rules for population curation prompt optimization (Approach 3).

Rules
1. List any words that could be used to characterize a population from either the title of the study or the body of the abstract.
2. Using the words that describe the population, identify both the 'country' and 'continent' of the study population. If 'country' or 'continent' cannot be identified, please report 'undefined'.
3. If 'population' describes a country in the Middle East, please classify as [continent] - Middle East'. Example 1: If the study population includes Egypt list as 'Africa - Middle East'. Example 2: If the study population includes Iran, list as 'Asia - Middle East'.
4. If the study population includes United Kingdom, define country as 'England', 'Scotland', 'Wales', or 'Northern Ireland', or else as 'UK-Not Defined'.

Phenomenology

Following zero-shot 'basic' analysis, we evaluated the types of discordance between human and GPT curation of populations. Four classifications of difference were evaluated: 1) differences in grammar (including parts of speech and singular vs plural); 2) differences in spelling or spelling errors; 3) differences in interpretation (e.g., the abstract references the UK Biobank and the human curator documents country as 'England', while GPT documents the country as 'United Kingdom'); and 4) differences in curation (e.g., the human curator notes the continent is Europe when reviewing an abstract for the UK Biobank, and GPT says the information is 'not included'). Where differences between human and GPT curation were observed in the 'basic' approach, each abstract was reviewed and the nature of the difference between human and GPT-4 curation was documented. The frequency of each observed type of difference was counted and reported. Results from the difference phenomenon contributed to the prompt rules developed for Approach 3.

Reason for OpenAI's GPT-4 Classification

To provide context for OpenAI's GPT-4 classifications, we added 'provide explanation' to the prompt. Explanations are documented and tracked to provide contextualization for future prompt engineering and process improvement.

Results

Our results show that GPT-4 performance varies across approaches, with Approach 2 (multi-shot with examples) resulting in the greatest similarity with human annotation of the 'population' feature and Approach 3 (multi-shot with prompt optimization) performing best for the classification of 'continent' and 'country'. Curation of 'population' had the lowest accuracy across approaches, with concordance between human and OpenAI's GPT-4 ranging between 11% for 'zero-shot with author information', 13% for 'zero-shot abstract only', 29% for 'multi-shot with prompt optimization', and 39% for 'multi-shot with examples'.

The approach to prompt design for curation plays an important role in the performance of GPT-4. Discordance between human and GPT-4 curation from Prompt 1 often resulted from differential classification of a geographic entity. For example, when human curators viewed United Kingdom (UK), they entered England for the country, whereas GPT-4 entered United

Kingdom. Additionally, when Iran was identified as the country, human curators entered Middle East/Middle Eastern, whereas GPT-4 entered Asia. Similarly, when Egypt was viewed in the abstract, human curators entered Middle East/Middle Eastern for the continent, whereas GPT-4 entered Africa (Table 2).

Table 2. Example of differences in interpretation of curated abstracts between human and OpenAI's GPT-4.

Population	Human		OpenAI's GPT-4	
	Continent	Country	Continent	Country
UK Biobank Caucasian cohort	European	England	Europe	United Kingdom
Iranian	Middle Eastern	Iran	Asia	Iran
Egypt	Middle Eastern	Egypt	Africa	Egypt

'Zero-shot with abstract only' showed consistent improvements on alignment when the model allowed for a partial or 'fuzzy' string match. Additionally, the features 'country' and 'continent' showed more than double the concordance with human annotation than the 'population' feature. Through qualitative analysis of the 'difference' phenomenon, we found 62.1% of differences between human and OpenAI's GPT-4 in the 'zero-shot abstract only' approach resulted from curation differences (curation differences are described as occurring when the human and OpenAI's GPT-4 curation showed little or no similarity). Approximately a quarter of curation differences resulted from grammatical differences, while 11.3% of curation differences resulted from differences in interpretation, and there were no discordances resulting from spelling differences. Agreement and phenomenology of Approach 1 (abstract only) are described in Figure 2. For each of the features in Approach 1, the use of a partial 'fuzzy' string to evaluate performance yielded higher concordance. However, the difference between 'full' and 'fuzzy' agreement differed by feature, with 'country' showing the smallest difference between 'full' and 'fuzzy' evaluation.

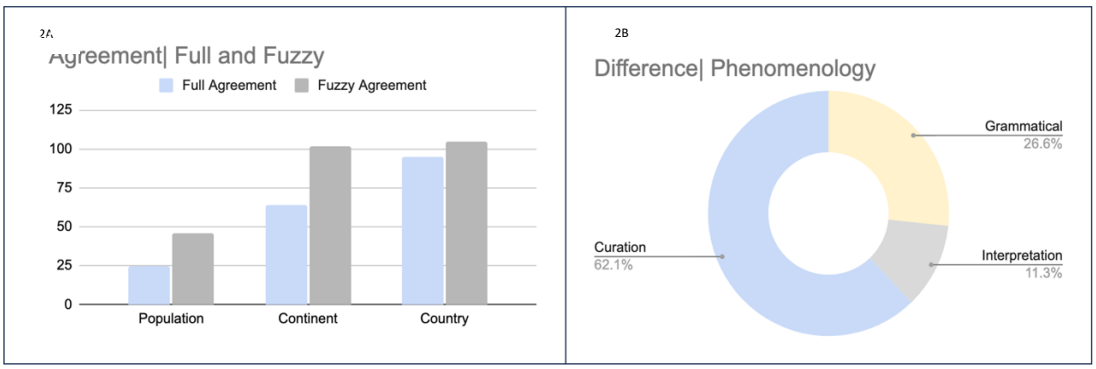

Figure 2. Agreement and phenomenology of difference in curation between human and OpenAI's GPT-4 using an abstract only approach. In 2A we present full (exact string match) and fuzzy (partial string match) comparing human and OpenAI's GPT-4 curation. The agreement between human and OpenAI's GPT-4 curators for 'population', 'continent' and 'country'. In 2B we show the qualitative analysis of differences between human and OpenAI's GPT-4 for discordant annotations. The methods for phenomenology analysis are described in the methods section.

Using the information from Approach 1, we optimized the prompt (Approach 3) to include a set of curation rules (Table 1). The results from the resultant Approach 3 showed the greatest alignment with human curation for the features 'continent' and 'country'. Approach 2 showed the greatest alignment with human curators for 'population'. However, both approaches (2 and 3) had more than double the concordance between GPT-4 and human curation than that of the 'zero-shot' Approaches (A and B) for 'population' curation. A summary of results for 'exact string matches' across Approaches (1–3) for 'population', 'continent' and 'country' is presented in Figure 3.

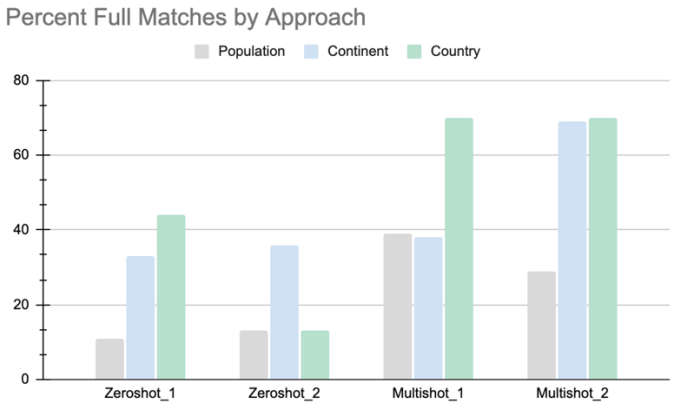

Figure 3. Percentage of full matches by approach. The approaches: Zero-shot_1, Zero-shot_2, Multi-shot_1 and Multi-shot_2 are described in the method section.

In Approach 2 (multi-shot with examples), GPT-4 was asked to explain its classification. In most cases, it provided text describing the information reviewed for the classification (Table 3). In a few cases, it returned 'not available' (Table 3, Examples 1, 2). For some of the explanations, GPT-4 asserted that the information was not specifically stated (Table 3, Examples 1, 4). In some cases, the rationale was based on structured data (e.g., author country: Table 3, Examples 3, 4). There were also some instances where population descriptions were not about geographic, ancestral or ethnic classifications (Table 3, Example 5).

Table 3. Sample of GPT-4 explanations from Approach 2 (multi-shot with examples).

Example	Population	Continent	Country	Explanation
1	Not Available	Europe	Spain	The abstract does not provide specific information about the population that the study was conducted on.
2	Egyptian	Africa	Egypt	Not available.
3	Brazilian	South America	Brazil	The population can be deduced as Brazilian due to the location of the study and the affiliations of all authors, which are from institutions located in Brazil.
4	Asian	Asia	China	The population is not explicitly stated, but it's deduced from the affiliations of the authors which are all based in China and from the context of using data from the UK Biobank and the COVID-19 host genetics initiative, which likely includes Chinese participants.

Table 3. Sample of GPT-4 explanations *(continued)*

Example	Population	Continent	Country	Explanation
5	Pregnant Women	Europe	Italy	The abstract discusses a study conducted on pregnant women. It mentions clinical features in pregnancy and third-trimester vitamin D levels, which implies that the population under study is pregnant women.

Discussion

In this study we sought to evaluate the feasibility of using an LLM as an AI curation assistant for population groups in biomedical literature. We used different prompting approaches with GPT-4 to curate study 'population', 'country' and 'continent' from a dataset of abstracts drawn from biomedical articles related to COVID. We observed that GPT-4 was able to perform some curation tasks well, and that in general the curation of our target structured information was strong. Curation of population groups yielded lower accuracy than curation of 'continent' and 'country' across all approaches (population concordance ranged from 11% to 39%). Population descriptors can vary across cultural, colloquial and geographic contexts, resulting in decreased consistency in the terms used across published abstracts. Conversely, 'countries' and 'continents' provide more fixed English language classifications, which may reflect the higher concordance between human and GPT-4 curation of 'country' and 'continent' ('country' concordance ranged from 13% in Approach 1A [zero-shot abstract only] to 70% in Approach 3 [multi-shot with prompt engineering]; 'continent' concordance ranged from 33% in Approach 1B [zero-shot with author information] to 69 % in Approach 3 [multi-shot with prompt engineering]). Accuracy in the curation of 'population' for OpenAI's GPT-4 improved when examples were provided. Given these findings, an expansion of the example set used in Approach 2 may provide further improvement for the curation of 'population'. Our findings suggest that prompt engineering and use of examples provide meaningful improvements in curation performance for OpenAI's GPT-4. Additionally, we propose use of partial string matches as opposed to exact string matches, which may be overly restrictive for evaluating curation results.

Understanding the differences between GPT-4 curation and human curation through phenomenology analysis provided valuable information for prompt engineering. In a few instances, GPT-4 classification of features differed from that by human curators due to an interpretive difference in definition. We believe these discrepancies can be identified through qualitative analysis and improved with more rigorous prompt engineering and inclusion of term definitions in the prompt. Notably, these types of differences provided opportunities to review the curation process and intended definitions. Overall, we feel confident that GPT-4, and potentially other LLMs, can provide valuable assistance in the curation of biomedical literature for population descriptors. Future research will include exploration of advanced prompt optimization, an expansion of the curation example-set, and software strategies to manage tokens and computing needs in a fully automated pipeline.

Acknowledgements

The authors would like to acknowledge the AIM AHEAD Consortium as the convening body for the research team for this work.

References

Buniello, A., Amode, R., Cerezo, M., Flicek, P., Hall, P., Harris, L., … Whetzel, P. L. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research 47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bustamante, C. D., Burchard, E. G., & De la Vega, F. M. (2011). Genomics for the world. *Nature 475*(7355), 163–165. https://doi.org/10.1038/475163a

Byeon, Y. J. J., Bonham, V. L., Brody, L. C., Islamaj, R., Lu, Z., Wilbur, W. J., & Yeganova, L. (2021). Evolving use of ancestry, ethnicity, and race in genetics research: A survey spanning seven decades. *American journal of human genetics 108*(12), 2215–2223. https://doi.org/10.1016/j.ajhg.2021.10.008

Landry, L. G., Ali, N., Bonham, V. L., Rehm, H. L., & Williams, D. R. (2018). Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health affairs (Project Hope) 37*(5), 780–785. https://doi.org/10.1377/hlthaff.2017.1595

Lee, K., Anderson, M. J., Carneiro, F., Chao, E., Dixon, K., Figueiredo, J., … Zhang, L. (2018) Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Human Mutation 39*(11),1553–1568. https://doi.org/10.1002/humu.23650

Logan, R., Wehe, A. W., Woods, D. C., Tilly J., & Khrapko, K. (2023). Interpreting Sequence-Levenshtein distance for determining error type and frequency between two embedded sequences of equal length. ArXiv (Preprint). arXiv:2310.12833v1

Manotas, M. C., Rivera, A. L., & Sanabria–Salas, M. C. (2023). Variant curation and interpretation in hereditary cancer genes: An institutional experience in Latin America. *Molecular Genetics & Genomic Medicine 11*(5), e2141. https://doi.org/10.1002/mgg3.2141

Mauro, M., Allen, D. S., Dauda, B., Lewis, A. C. F., Molina, S. J., & Neale, B. M. (2022). A scoping review of guidelines for the use of race, ethnicity, and ancestry reveals widespread consensus but also points of ongoing disagreement. American journal of human genetics 109(12), 2110–2125. https://doi.org/10.1016/j.ajhg.2022.11.001

Ravichandran, V., Cadoo, K., Kemel, Y., Lipkin, S., Mandelker, D., Offit, K., … Zhang, L. (2019). Toward automation of germline variant curation in clinical cancer genetics. *Genetics in Medicine 21*(9), 2116–2125. https://doi.org/10.1038/s41436-019-0463-8

Sollis, E., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., … Ramachandran, S. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic acids research 51*(D1), D977–D985. https://doi.org/10.1093/nar/gkac1010

Wojcik, G. L., Acuña–Alonso, V., Ambite, J. L., Barnes, K. C., Bottinger, E. P., Bustamante, C., C., … Zubair, N. (2019) Genetic analyses of diverse populations improves discovery for complex traits. *Nature 570,* 514–518. https://doi.org/10.1038/s41586-019-1310-4