META-RESEARCH ARTICLE

Public human microbiome data are dominated by highly developed countries

Richard J. Abdill[1], **Elizabeth M. Adamowicz**[1], **Ran Blekhman**[1,2]*

[1] Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, Minnesota, United States of America, [2] Department of Ecology, Evolution and Behavior, University of Minnesota, St. Paul, Minnesota, United States of America

* blekhman@umn.edu

OPEN ACCESS

Citation: Abdill RJ, Adamowicz EM, Blekhman R (2022) Public human microbiome data are dominated by highly developed countries. PLoS Biol 20(2): e3001536. https://doi.org/10.1371/journal.pbio.3001536

Academic Editor: Ken Cadwell, New York University School of Medicine, UNITED STATES

Received: September 17, 2021

Accepted: January 11, 2022

Published: February 15, 2022

Data Availability Statement: All data has been deposited at Zenodo.org and is available at https://doi.org/10.5281/zenodo.5351179. This repository also includes the code used for data collection, along with the code used to generate each plot.

Funding: This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM128716 to RB; https://www.nigms.nih.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The importance of sampling from globally representative populations has been well established in human genomics. In human microbiome research, however, we lack a full understanding of the global distribution of sampling in research studies. This information is crucial to better understand global patterns of microbiome-associated diseases and to extend the health benefits of this research to all populations. Here, we analyze the country of origin of all 444,829 human microbiome samples that are available from the world's 3 largest genomic data repositories, including the Sequence Read Archive (SRA). The samples are from 2,592 studies of 19 body sites, including 220,017 samples of the gut microbiome. We show that more than 71% of samples with a known origin come from Europe, the United States, and Canada, including 46.8% from the US alone, despite the country representing only 4.3% of the global population. We also find that central and southern Asia is the most underrepresented region: Countries such as India, Pakistan, and Bangladesh account for more than a quarter of the world population but make up only 1.8% of human microbiome samples. These results demonstrate a critical need to ensure more global representation of participants in microbiome studies.

Background

A growing body of research shows that the human microbiome has broad relevance to human health and disease. However, identifying the specific connections between the microbiome and human health requires a broad survey of both human populations and their most common health conditions. Even among healthy individuals, human microbiome composition varies between populations in ways that are still being uncovered: Geography and geographic relocation has been found to have an influence on microbiome composition [1–3], as have host genetic variation and ethnicity [4–6]. Diet [7], lifestyle [8], and patterns in antibiotic use [9] have all been linked to microbiome composition, with other studies considering the influence of locational factors such as pollution [10]. Even within countries, interacting factors such as income, race, and education have critical impacts on health outcomes that could be mediated by the human microbiome [11]. Some microbiome studies have specifically collected

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ACCME, African Collaborative Center for Microbiome and Genomics Research; GWAS, genome-wide association study; NCBI, National Center for Biotechnology Information; PRS, polygenic risk score; SRA, Sequence Read Archive.

and compared data from global sites [12,13], but large gaps and disparities still exist in which microbiomes are being studied on a global scale. The human microbiome has been linked to a growing number of social, medical, and economic factors not directly related to host genetics, which reinforces the urgent need to evaluate the microbiomes of many populations [11,14].

Other genomics fields have developed similar gaps, in which disproportionate attention is paid to the majority populations of wealthy countries: Genome-wide association studies (GWASs), for example, have been primarily conducted in populations with European ancestry [15,16]. As a result, polygenic risk scores (PRSs) from these studies have poorer accuracy when applied to non-European groups, limiting the possible benefits of this research—including personalized medicine, early disease screening, and risk prediction—to European-descended populations [17–19]. There has been a concerted effort in genomics to include non-European individuals in GWAS studies, concurrent with calls to build research infrastructure and capacity globally [16]. It is likewise critical to identify underrepresented populations and locations in both genomics and microbiome research; otherwise, the benefits of host–microbiome research may only extend to a subset of the global population.

To investigate the geographic distribution of microbiome studies, we used metadata on all human microbiome datasets in the BioSample database, which includes metadata describing samples in the Sequence Read Archive (SRA), DNA Data Bank of Japan, and European Nucleotide Archive [20]. Our data include the country of origin and time of release for more than 444,000 samples, including both 16S amplicon sequencing and shotgun metagenomic sequencing, released over the last 11 years. These samples from the 3 largest genomic databases represent a large majority of all human microbiome samples that have been published.

Results

We downloaded metadata for 444,829 human microbiome samples across 19 body sites and 2,592 studies. These data are available from the BioSample database maintained by the National Center for Biotechnology Information (NCBI), which includes metadata describing raw sequencing data deposited in multiple international repositories, including SRA [21]. While sample-level genomic sequencing data are uploaded to SRA, information such as geographic origin is saved separately to an entry in the BioSample database. BioSamples can be tagged with any number of "attributes," including 485 standardized fields documented by NCBI [22]; we downloaded all attributes for all these samples. We used a Python script to load this metadata into a PostgreSQL database, where the information was aggregated using sample metadata such as country of origin and time of publication (see **Materials and methods**).

As expected, we found that the number of human microbiome samples with publicly available data has been increasing over time, from 3 microbiome samples in 2010 to 123,302 in 2020, the first year in which more than 100,000 human microbiome samples were released (**S1 Fig**). Although there were microbiome studies conducted prior to 2010, that was the first year of the BioSample database, which all depositors must now use if they submit sequencing data to the SRA. The most commonly used attribute in this subset of samples is the geographic origin of the sample [22], which is available for 99.5% of samples (**S1 Table**). Using this attribute, we were able to determine the country of origin for 382,711 (86%) human microbiome samples (**Fig 1A**), which originated in 115 different countries. We found that 178,960 samples (40.2%) were from the US, almost 5 times more than any other country (**Table 1**). China has the next most samples, with 36,162 (8.1%), followed by the United Kingdom, Denmark, Australia, and the Netherlands. China is the only Asian country in the top 14; the first South American country is Chile, in 16th place with 3,616 samples (0.8%). Malawi is the first African country, in 19th place with 3,052 samples (0.7%).

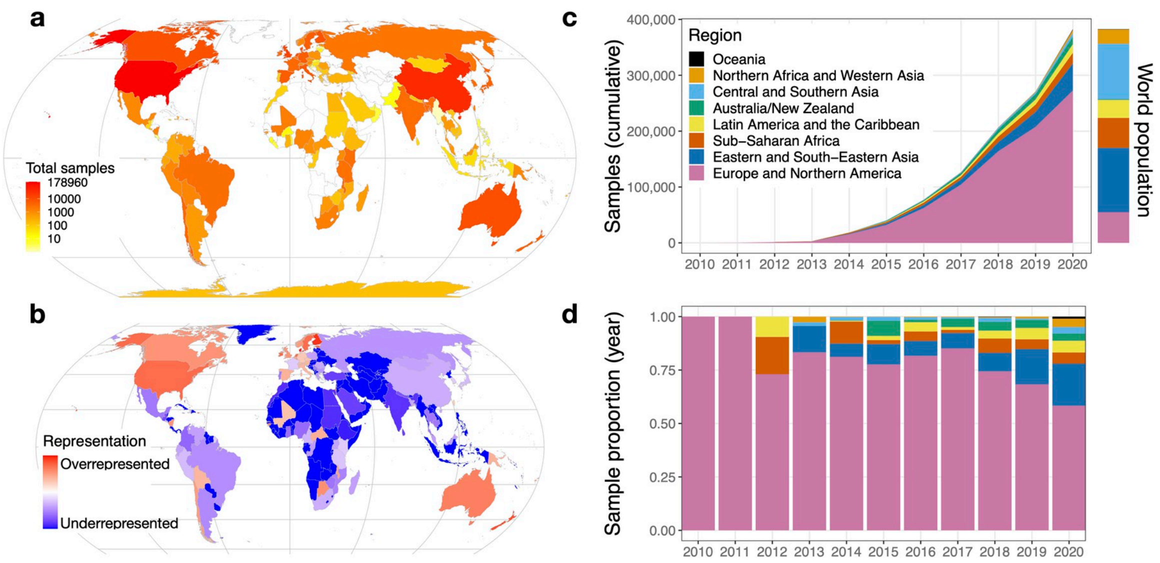

Fig 1. Global microbiome representation. **(a)** Total samples by country. The color of each country indicates the total number of samples originating in that country using a log10 scale. **(b)** Relative representation by country. The color of each country indicates its representation in human microbiome datasets, relative to its share of world population. Red colors mark countries that are overrepresented relative to their population, and blue colors mark countries that are underrepresented. Countries with zero samples in the dataset are marked with dark blue. **(c)** Cumulative microbiome samples by world region. The x-axis indicates the year, and the y-axis indicates the cumulative microbiome samples available at the end of that year. Colors indicate the cumulative microbiome samples from each of the world regions specified in the legend. The colored bar to the right of the plot indicates the share of the world population living in each of the regions using the same colors. **(d)** Proportion of annual samples. The x-axis indicates the year, and the y-axis indicates the proportion of samples from each world region published in that year. Colors correspond to the world regions shown in panel C. The data and code needed to generate this figure can be found at https://doi.org/10.5281/zenodo.5351179. All maps are based on public domain Natural Earth data; the base layer is available for download at https://www.naturalearthdata.com/http//www.naturalearthdata.com/download/50m/cultural/ne_50m_admin_0_countries.zip.

https://doi.org/10.1371/journal.pbio.3001536.g001

We also evaluated patterns specific to body sites. The number of countries represented in each body site is roughly proportional to the number of overall samples, with the most frequently sampled body site, the human gut, also holding data from the most countries, 96 (**Table 2**). This number drops quickly, however: For example, there are 44 countries represented in the skin microbiome category, and only 22 in the nasopharyngeal microbiome. Even if we consider only the 115 countries that appear in this dataset, it appears most body sites exclude most countries. When we consider body sites per country, rather than countries per body site, we can also evaluate the best characterized country-level microbiomes: China has samples in 17 of the 19 body sites, the most of any country (**S3 Table**), followed by the US with 16. The first South American country on the list, Brazil, has only 9, and South Africa, the first African country, appears in 8 body sites. Next, we used these data to assess country-level patterns at the 5 most prevalent body sites: the gut, mouth, skin, vagina, and lung (**S2 Table**). The US has the most samples in all 5. The skin microbiome category differs notably from the overall top 10: Although the US and China again appear at the top, the remainder of the top 10 includes Chile, Bangladesh, Papua New Guinea, Hong Kong, India, Puerto Rico, Australia, and Peru. However, this is also the body site with the most lopsided difference between the US and the rest of the world: The US total (19,706 samples) is 12.6 times that of the number 2 country, China (1,562 samples), and more than 50 times that of the 10th country, Peru (391 samples).

To examine patterns of under- and overrepresentation of countries, we compared human microbiome sample counts to each country's population, according to United Nations

Table 1. Samples per country.

Position	Country	Samples	Share
1	US	178,960	40.2%
	Unknown	62,118	14.0%
2	China	36,162	8.1%
3	UK	16,076	3.6%
4	Denmark	11,497	2.6%
5	Australia	9,266	2.1%
6	the Netherlands	9,173	2.1%
7	Canada	8,829	2.0%
8	Finland	7,855	1.8%
9	Italy	6,265	1.4%
10	Germany	5,531	1.2%
11	Spain	5,517	1.2%
12	Sweden	5,248	1.2%
13	Israel	4,831	1.1%
14	New Zealand	4,354	1.0%
15	Japan	4,298	1.0%
16	Chile	3,616	0.8%
17	Bangladesh	3,502	0.8%
18	France	3,402	0.8%
19	Malawi	3,052	0.7%
20	India	2,997	0.7%
	Rest of world	52,280	11.8%

https://doi.org/10.1371/journal.pbio.3001536.t001

estimates for 2020 [23]. The US is dramatically overrepresented relative to its population: Although the country has about 4.3% of the global population, 40.2% of human microbiome samples originate there. Proportionally, Denmark is the most overrepresented country, with 11,497 samples from a country of about 5.8 million people (**Fig 1B**). Of the 235 countries and territories included in the United Nations population estimates, 120 have zero human microbiome samples available in these public databases.

To gain a better understanding of global representation in microbiome research, we grouped countries using the 8 United Nations Sustainable Development Goals regions [24]. We found that 71.2% of samples with a known location come from Europe and Northern America, a region that holds only 14.3% of the world's population (**Table 3**). Proportionally, Australia/New Zealand has the most lopsided presence in the database: The region's 30.3 million people is 0.4% of the population, but accounts for 3.1% of samples (**Fig 1C**). Central/Southern Asia is the most underrepresented region: It holds 25.8% of the population but makes up only 1.8% of microbiome samples. Northern Africa and Western Asia are the next most underrepresented regions, followed by sub-Saharan Africa, which is home to 14.0% of the world's population but is the source of 4.2% of human microbiome samples. These proportions indicate that a person in Europe or Northern America is roughly 14 times more likely to be studied in a microbiome project than someone from sub-Saharan Africa. The 47 countries on the United Nations list of "least developed countries" account for about 14% of the world's population [25], but 3.4% of microbiome samples; 29 of those countries have no samples at all (**S4 Table**). We also found that, although samples from Europe and Northern America are overrepresented, in recent years, there is more representation for samples from other regions, most prominently eastern and southeastern Asia (**Fig 1D**).

Table 2. Samples by body site. Each row indicates a body site related to the human microbiome. The "Samples" column indicates the total number of samples categorized under each body site, and the "Countries" column indicates the number of unique countries with at least 1 sample in that category. Samples without a known country are included in the sample count, but not the country count. Body sites map directly to categories defined in the NCBI Taxonomy Browser; see S5 Table for a list of category IDs combined for each body site.

Body site	Samples	Countries
Gut	220,017	96
Human metagenome*	69,697	58
Oral	47,798	63
Skin	36,593	44
Vaginal	17,784	31
Lung	17,307	30
Nasopharyngeal	15,646	22
Feces	6,858	13
Reproductive system	3,180	6
Blood	2,707	9
Saliva	2,503	15
Milk	2,060	9
Urinary tract	1,187	4
Tracheal	520	2
Sputum	364	3
Eye	359	8
Semen	203	3
Bile	45	2
Skeleton	1	1

*Samples under the "human metagenome" label refer to an NCBI category that does not specify a particular body site.

NCBI, National Center for Biotechnology Information.

https://doi.org/10.1371/journal.pbio.3001536.t002

Discussion

Our results show that the global distribution of human microbiome sampling is heavily skewed toward North American and European populations, both in total samples (**Fig 1A**) and in samples adjusted for population (**Fig 1B**). The US is by far the greatest contributor to the database (**Table 1**), although this is slowly beginning to change as other countries' contributions grow (**Fig 1D**). This neglect of most of the world's population represents a disparity in microbiome research that could limit the health benefits of microbiome research to those countries and populations whose microbiomes have been extensively sampled and studied. Since only a subset of the world's populations are currently being studied, the associations between the microbiome and disease may not hold in undersampled populations [26,27]. For example, Gupta and colleagues identified several differences in the microbiome of healthy individuals from various geographic locations and lifestyles across the globe; without a consistent "healthy" microbiome across global populations, identifying microbiome–disease associations is nearly impossible [26]. He and colleagues also found that microbiome-based models for predicting metabolic disease failed when applied to populations outside of the geographical location in which they were developed [27]. Additionally, by only sampling a subset of the global population, the diseases studied in the context of the microbiome are limited to diseases which impact that subset. Helminth parasite infections, for example, are common in tropical and subtropical regions of the world, but rare in North American and European populations.

Table 3. Samples and population by region.

Region	Samples	2020 population (estimated, in thousands)	% of samples	% of samples (known location)	% of population	Representation proportion*
Europe and Northern America	272,544	1,116,506	61.3%	71.2%	14.3%	4.97
Eastern and Southeastern Asia	49,007	2,346,709	11.0%	12.8%	30.1%	0.43
Sub-Saharan Africa	18,651	1,094,366	4.2%	4.9%	14.0%	0.35
Latin America and the Caribbean	15,264	653,962	3.4%	4.0%	8.4%	0.49
Australia/New Zealand	13,620	30,322	3.1%	3.6%	0.4%	9.14
Central and Southern Asia	6,685	2,014,709	1.5%	1.7%	25.8%	0.07
Northern Africa and Western Asia	5,621	525,869	1.3%	1.5%	6.7%	0.22
Oceania	1,178	12,356	0.3%	0.3%	0.2%	1.94
Unknown	62,259		14.0%			
Least developed countries	15,254	1,057,438	3.4%	4.0%	13.6%	0.29
Rest of world	367,457	6,737,361	82.6%	96.0%	86.4%	1.11
Unknown	62,118		14.0%			

*Representation proportion calculated by dividing a regions percentage of known samples by its percentage of population.

https://doi.org/10.1371/journal.pbio.3001536.t003

Undersampling of the microbiota from populations where these infections are common has led to a lack of clear understanding of the role of the microbiome in helminth colonization and resistance [28].

To ensure greater global equity in the benefits of microbiome research, many stakeholders —funders, researchers, and journals, to name a few—should consider how to ethically prioritize and incentivize improved global representation of microbiome samples, as they have begun to do in genomics with efforts such as the H3Africa initiative [29]. Others have also highlighted opportunities for growth in the microbiome field, such as developing infrastructure and processes in low-resource settings [30,31], building more comprehensive microbial reference databases, and pursuing more flexible and affordable sequencing technologies [32]. Importantly, this approach should be grounded in benefitting the populations and communities sampled, rather than simply using these microbiomes as a tool to improve health in North American and European countries [33,34]. Ongoing discussion of "helicopter research" (e.g., [35]) sheds light on ethical objections to "solving" research disparities with what essentially becomes charity, rather than collaboration: Researchers from wealthy countries obtain funding to do research in developing countries, "helicopter in" to collect data, then leave to publish their papers [36]. The result is more data from that country, but as part of a project that may not address the problems and priorities of the country under study. Local researchers, if they are consulted at all, may be excluded from authorship on the papers that are then hidden behind paywalls, written in a language they may not speak—part of much broader issues in scientific communication [37,38]. Researchers from the so-called "Global North" (as we are) would benefit from deferring to experienced scientists in these countries to find out how to avoid common extractive tropes in imbalanced collaborations (e.g., [35,39]). Research and discussion in other fields may also help scientists trying to build more inclusive research projects: Although there are no easy answers, essays in applied ecology [40–42], ocean science [43], botany [44], geography [45,46], and conservation [47], among many others, deal with the hallmarks and dangers of colonial science [48] and how researchers can change their approach to knowledge production.

The reasons for, and solutions to, global disparities in scientific research go far beyond the scope of this paper, and indeed of the microbiome field. There are broader issues of global representation in science that we and others have discussed, for example, in terms of authorship [49], language [37], and the makeup of editorial boards [40]. The complex history and current conditions driving these disparities requires a comprehensive assessment of global sociopolitical factors that we, as biologists based in North America, are not able to fully address. However, the necessity of such an assessment as a way to solve these problems illustrates an important possible reason that these problems continue to perpetuate. Most microbiome researchers are not trained in social or political science and lack the appropriate tools to assess and address these problems. The more intentional inclusion of social scientists in microbiome projects may help address not only country-level imbalances, but also remediate harmful conventions used to deal with other issues like race [50].

Despite ongoing challenges, there have been several recent success stories of microbiome initiatives set in, driven by, and focused on countries and populations who have been historically left out of microbiome research. One such example is the recently convened Microbiome Task Force from the H3Africa Consortium; their goals are to harmonize and perform meta-analyses of microbiome data from H3Africa, build capacity and knowledge sharing among members, and provide data analysis support to researchers [51]. The Pan-African Bioinformatics Network (H3ABioNet), which has worked extensively in genomics research capacity building in Africa, also recently hosted a hackathon wherein they began work on a data portal for African microbiome samples [52]. In South America, the Brazilian Microbiome project and the recently proposed Ecuadorian Microbiome project both seek to advance microbiome research capacity in their respective countries and create local infrastructure to support these goals [53,54]. Initiatives such as H3Africa's African Collaborative Center for Microbiome and Genomics Research (ACCME) [55] may be ideally positioned to make progress in these trends, although as research activity grows in these underrepresented countries, using public metadata may become a less viable measure of these disparities: ACCME's 2 existing microbiome publications, for example, do not have information about data availability [56,57], and ongoing discussions about issues such as data sovereignty [58] raise important questions about whether making data publicly available is a just and sustainable approach to biomedical research in countries or populations with comparatively little power in the global research ecosystem [59–61].

There are several limitations to our study. Metadata quality is the primary hurdle in characterizing samples [62]: For example, our results suggest that data for some microbiome samples are misclassified as "Homo sapiens" data rather than "human metagenome" data, which makes them much more difficult to locate. As a result, some of the countries listed here with zero samples do have microbiome studies that were either submitted to databases that are challenging to access in bulk (e.g., Zenodo) or mislabeled in the SRA. However, the number of these misclassified samples is likely to be minor, and given the magnitude of differences observed in our study, this is unlikely to affect our main results (see **Materials and methods**). It is also possible that not all samples identified as human in this study are indeed from humans and could, for example, include studies using human gut microbiota transferred into mice. We also did not evaluate differences in host phenotypic information: Most samples are missing even basic information such as sex (77% missing) and age (79% missing), and the most prevalent tag indicating host health status, "host_disease," is only available for 7.8% of samples (**S1 Table**). Consequently, we do not have sufficient information to draw conclusions about differences in geographic distribution between "healthy" and "disease" samples.

Although disease-specific analysis is beyond the scope of our dataset, it would be interesting to investigate differences in the types of microbiome studies, and the questions they ask, on a global scale: If the human microbiome is generally understudied in a given country, it is likely that diseases prevalent in that country may also be lacking information about microbiome associations. We have also limited our database search to 3 databases (SRA, DNA Data Bank of Japan, and European Nucleotide Archive); it is possible that different patterns of global representation are present in other databases, such as MG-RAST [63] and gcMeta [64], although they are orders of magnitude smaller than the NCBI holdings. In addition, as it has been estimated that 20% of microbiome papers do not have publicly available data [65], our study only examines the subset of microbiome studies that also shared their data in the largest international repositories.

Samples collected from the same host could occur in longitudinal studies or datasets in which biological replicates were submitted as separate BioSamples, a pattern that is difficult to evaluate across multiple studies that may identify subjects differently, if at all. If longitudinal studies happen more frequently in some regions than others, it is possible that the reported proportions of samples between countries could differ from the proportions of human subjects. However, given the differences in sample numbers between countries, this is unlikely to change the main results from our study. Moreover, since we are using sample collection as a proxy for investment in microbiome research in a given country, the identity of the subject may not be as relevant—indeed, it is likely more costly to perform a longitudinal study with subject follow-up than it is to recruit more subjects for a single sample each. Still, if longitudinal sampling is more common in studies in North America and Europe (which seems likely, given the extensive infrastructure and funding required for following patients long term), it is possible that the gap between the "Global North" and the rest of the world in terms of microbiome sampling is smaller than our results suggest, if we were to count subjects rather than samples. However, given the magnitude of the difference between countries in our study, we do not believe repeated sampling from the same individuals in the Global North alone can account for such drastic disparities in sample numbers.

To conclude, we analyzed the geographic origins of almost a half-million samples from the largest genomic repositories in the world. We find evidence that the human microbiome field may be encountering some of the same flaws that arose in human genomics [66,67], in which much of the world is excluded and progress is focused on the priorities of the wealthy. The field would benefit from a more global perspective on investigating the human microbiome's relationship to health and disease.

Materials and methods

A list of samples was exported from the NCBI BioSample database (https://www.ncbi.nlm.nih.gov/biosample) using the search string "txid408170[Organism] AND biosample sra[filter] AND "public"[filter]," which requests all samples classified under the "human gut metagenome" category in the NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). The resulting sample IDs and all associated tags were loaded into a PostgreSQL database. We repeated this for all categories described as human metagenomes (**Table 2**). We note that the term "human gut metagenome" does not describe the sequencing technique used to generate the microbiome data, including shotgun metagenomics and amplicon sequencing —specifically, 301,700 samples (72.0%) are associated with sequencing runs that list the library strategy as "AMPLICON."

We then looked in other NCBI categories nested beneath the "organismal metagenomes" category that were not explicitly labeled "human" but were likely to contain some human samples [68]. We downloaded the metadata for samples classified under any NCBI category that was the "generic" version of a human one we had already collected—the "blood metagenome" category is the generic version of the "human blood metagenome" category, for example (**S5 Table**). We downloaded all sample data for any generic categories that had at least 1,000 samples, then evaluated the metadata to find which samples indicated they were taken from a human host. To do this, we used the value of the "host_taxid" field or, if that was blank, the value of "host," to create a putative "host" value, and manually flagged any that explicitly indicated the sample was from a human—references to "human" or "Homo sapiens," for example, or if the host included words such as "patient" or "crew member" and did not indicate another species. We evaluated 4,395 unique "host" values for 173,038 samples and found 501 values assigned to 29,934 samples (17.3%) that indicated the host was a human. These were also included in the analysis. The sample data were collected between April and June 2021; to minimize the effect of collecting some body sites after others, only samples dated prior to 2021 were included here.

We then used the NCBI eUtils API to find "runs" associated with each sample, so we could ensure all the BioSamples were associated with actual sequencing data. In the NCBI system, "runs" are the entities associated with sequencing data. We also used this API to obtain information on publication date, library strategy, and the dates on which samples became publicly available. This resulted in a collection of 444,829 samples across 19 body sites (**Table 2**) after removing several hundred samples that were missing dates or sequencing data.

Representation proportions

To determine which countries were over- or underrepresented relative to their populations, we obtained the 2020 population estimates for all countries as estimated by the United Nations [23]. We used this to calculate 2 percentages for each country, one for the country's share of the global population and another for the country's share of human microbiome samples. We then calculated a representation index: For countries with a higher sample percentage than population percentage, we divided the former by the latter to obtain a number indicating how many times more samples are present than expected. For countries with a lower sample percentage than population percentage, we took the negative reciprocal of this number, indicating (in negative numbers) the number one would have to multiply the sample count by to get the number that would be proportionally representative. The interim result leaves overrepresented countries with positive scores and underrepresented countries with negative scores. After removing the scores for countries with 50 or fewer samples, we scaled the positive scores to fall between 0 and 100 and separately scaled the negative scores to fall between 0 and −100. We then plotted these on the map using the "log 10" transformation to add more variation in the color coding for the countries with middling scores. For the regional calculations (**Fig 1C and 1D**), we used top-level classifications from the same United Nations document. Antarctica is not included in a region, so those samples were added to the "Unknown" category for region-level calculations.

To better understand gaps in what data may be available outside of these large centralized repositories evaluated here, we selected several countries with zero attributed samples and did a literature search to determine whether human microbiome studies had been performed there and, if so, where the data are stored. For example, we could not confirm any samples available from Kazakhstan (population 18.7 million) in central Asia, but a human gut microbiome study from there was published in 2020 [69]; its raw sequencing data (but no phenotypic information) are available on Zenodo, a scientific data repository with many submissions but no way of searching for samples or projects. Another Kazakhstan microbiome study [70] is linked to publicly available sequencing data (BioProject PRJEB17632), but with incorrect metadata: Samples are classified as human sequencing data, rather than metagenomic, an issue addressed directly in the SRA submission instructions [71]. In addition, geolocation metadata was submitted, but listed the country of origin as Germany, the location of the senior author (and presumably the sequencing center), rather than Kazakhstan, and the geographical source of the sample, as requested by NCBI [22], although instructions can differ between repositories [62]. A study in Honduras (population 9.9 million) includes SRA data with accurate geolocation information (BioProject PRJEB31759), but the samples were again classified under "Homo sapiens" rather than "human metagenome" [72].

Visualization

All figures were made using R and the ggplot2 package [73]. Maps use the Equal Earth projection [74] and the rnaturalearth R package [75].

Supporting information

S1 Fig. Samples per year. The x-axis indicates the year, and the y-axis indicates the number of microbiome samples released in that year. Colors indicate the region of origin for each sample and match the colors used in Fig 1C and 1D. The data and code needed to generate this figure can be found at https://doi.org/10.5281/zenodo.5351179.
(TIF)

S1 Table. Samples per tag. Each row represents a single metadata field available for BioSample entries. The "samples" column indicates how many samples have a value for that field.
(CSV)

S2 Table. Top 10 countries by body site. Each column holds a list of the 10 countries with the most samples in a single body site. The "unknown" category is omitted here.
(CSV)

S3 Table. Samples per body site per country. This contains similar data to S2 Table, except no countries or body sites are omitted. Each column is a single body site. Each row is a country, and each cell represents the number of samples from that country that appeared in that body site.
(CSV)

S4 Table. Country-level data. Each row represents a single country or territory as defined by the United Nations. There are 10 columns; see the Supporting information documentation for a description of them.
(CSV)

S5 Table. NCBI Taxonomy IDs. Each row represents a single body site. The "human" column indicates the ID used to identify samples explicitly labeled as human (e.g., "human gut metagenome"); the "generic" column indicates the ID used to identify samples not labeled as human (e.g., "gut metagenome"). NCBI, National Center for Biotechnology Information.
(CSV)

Acknowledgments

We thank the members of the Blekhman lab for helpful discussions and the Minnesota Supercomputing Institute for providing computational resources.

Author Contributions

Conceptualization: Richard J. Abdill, Ran Blekhman.

Data curation: Elizabeth M. Adamowicz.

Formal analysis: Richard J. Abdill, Elizabeth M. Adamowicz.

Funding acquisition: Ran Blekhman.

Methodology: Richard J. Abdill, Ran Blekhman.

Software: Richard J. Abdill.

Supervision: Ran Blekhman.

Visualization: Richard J. Abdill.

Writing – original draft: Richard J. Abdill, Elizabeth M. Adamowicz.

Writing – review & editing: Ran Blekhman.

References

1. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012; 486:222–7. https://doi.org/10.1038/nature11053 PMID: 22699611
2. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, et al. US Immigration Westernizes the Human Gut Microbiome. Cell. 2018; 175:962–72.e10. https://doi.org/10.1016/j.cell.2018.10.029 PMID: 30388453
3. Kaplan RC, Wang Z, Usyk M, Sotres-Alvarez D, Daviglus ML, Schneiderman N, et al. Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. Genome Biol. 2019; 20:219. https://doi.org/10.1186/s13059-019-1831-z PMID: 31672155
4. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014; 159:789–99. https://doi.org/10.1016/j.cell.2014.09.053 PMID: 25417156
5. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 2015; 16:191. https://doi.org/10.1186/s13059-015-0759-1 PMID: 26374288
6. Brooks AW, Priya S, Blekhman R, Bordenstein SR. Gut microbiota diversity across ethnicities in the United States. PLoS Biol. 2018; 16:e2006842. https://doi.org/10.1371/journal.pbio.2006842 PMID: 30513082
7. Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, et al. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host Microbe. 2019; 25:789–802.e5. https://doi.org/10.1016/j.chom.2019.05.005 PMID: 31194939
8. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, et al. The microbiome of uncontacted Amerindians. Sci Adv. 2015; 1. https://doi.org/10.1126/sciadv.1500183 PMID: 26229982
9. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, et al. Country-specific antibiotic use practices impact the human gut resistome. Genome Res. 2013; 23:1163–9. https://doi.org/10.1101/gr.155465.113 PMID: 23568836
10. Mutlu EA, Comba IY, Cho T, Engen PA, Yazıcı C, Soberanes S, et al. Inhalational exposure to particulate matter air pollution alters the composition of the gut microbiome. Environ Pollut. 2018; 240:817–30. https://doi.org/10.1016/j.envpol.2018.04.130 PMID: 29783199
11. Amato KR, Arrieta M-C, Azad MB, Bailey MT, Broussard JL, Bruggeling CE, et al. The human gut microbiome and health inequities. Proc Natl Acad Sci U S A. 2021;118. https://doi.org/10.1073/pnas.2017947118 PMID: 34161260
12. Fragiadakis GK, Smits SA, Sonnenburg ED, Van Treuren W, Reid G, Knight R, et al. Links between environment, diet, and the hunter-gatherer microbiome. Gut Microbes. 2019; 10:216–27. https://doi.org/10.1080/19490976.2018.1494103 PMID: 30118385
13. Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell. 2021; 184:2053–67.e18. https://doi.org/10.1016/j.cell.2021.02.052 PMID: 33794144

14. Ishaq SL, Parada FJ, Wolf PG, Bonilla CY, Carney MA, Benezra A, et al. Introducing the Microbes and Social Equity Working Group: Considering the Microbial Components of Social, Environmental, and Health Justice. mSystems. 2021:e0047121. https://doi.org/10.1128/mSystems.00471-21 PMID: 34313460

15. Medina-Gomez C, Felix JF, Estrada K, Peters MJ, Herrera L, Kruithof CJ, et al. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. Eur J Epidemiol. 2015; 30:317–30. https://doi.org/10.1007/s10654-015-9998-4 PMID: 25762173

16. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally diverse populations. Nat Rev Genet. 2019; 20:520–35. https://doi.org/10.1038/s41576-019-0144-0 PMID: 31235872

17. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? Genome Med. 2018; 10:100. https://doi.org/10.1186/s13073-018-0610-x PMID: 30591078

18. Peterson RE, Kuchenbaecker K, Walters RK, Chen C-Y, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell. 2019; 179:589–603. https://doi.org/10.1016/j.cell.2019.08.051 PMID: 31607513

19. Cai M, Xiao J, Zhang S, Wan X, Zhao H, Chen G, et al. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am J Hum Genet. 2021; 108:632–55. https://doi.org/10.1016/j.ajhg.2021.03.002 PMID: 33770506

20. Nakamura Y, Cochrane G, Karsch-Mizrachi I. International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2013; 41: D21–4. https://doi.org/10.1093/nar/gks1084 PMID: 23180798

21. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012; 40:D57–63. https://doi.org/10.1093/nar/gkr1163 PMID: 22139929

22. NCBI. Biosample Attributes. In: BioSample [Internet]. [cited 2021 May 21]. Available from: https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/

23. World Population Prospects 2019, Online Edition. Rev. 1. In: United Nations, Department of Economic and Social Affairs, Population Division [Internet]. 2019 [cited 2021 May 12]. Available from: https://population.un.org/wpp/Download/Standard/Population/

24. SDG Indicators. In: United Nations Sustainable Development Goals [Internet]. [cited 2021 May 18]. Available from: https://unstats.un.org/sdgs/indicators/regional-groups

25. About LDCs. In: United Nations Office of the High Representative for the Least Developed Countries, Landlocked Developing Countries and the Small Island Developing States (UN-OHRLLS) [Internet]. 19 Sep 2013 [cited 2021 May 12]. Available from: http://unohrlls.org/about-ldcs/

26. Gupta VK, Paul S, Dutta C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front Microbiol. 2017; 8:1162. https://doi.org/10.3389/fmicb.2017.01162 PMID: 28690602

27. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med. 2018; 24:1532–5. https://doi.org/10.1038/s41591-018-0164-x PMID: 30150716

28. Lee SC, Tang MS, Lim YAL, Choy SH, Kurtz ZD, Cox LM, et al. Helminth colonization is associated with increased diversity of the gut microbiota. PLoS Negl Trop Dis. 2014; 8:e2880. https://doi.org/10.1371/journal.pntd.0002880 PMID: 24851867

29. H3Africa Consortium, Rotimi C, Abayomi A, Abimiku A'le, Adabayeri VM, Adebamowo C, et al. Research capacity. Enabling the genomic revolution in Africa. Science. 2014; 344:1346–8. https://doi.org/10.1126/science.1251546 PMID: 24948725

30. Soo CC, Mukomana F, Hazelhurst S, Ramsay M. Establishing an academic biobank in a resource-challenged environment. S Afr Med J. 2017; 107:486–92. https://doi.org/10.7196/SAMJ.2017.v107i6.12099 PMID: 28604319

31. Mulder NJ, Adebiyi E, Adebiyi M, Adeyemi S, Ahmed A, Ahmed R, et al. Development of Bioinformatics Infrastructure for Genomics Research. Glob Heart. 2017; 12:91–8. https://doi.org/10.1016/j.gheart.2017.01.005 PMID: 28302555

32. Brewster R, Tamburini FB, Asiimwe E, Oduaran O, Hazelhurst S, Bhatt AS. Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation. Trends Microbiol. 2019; 27:824–35. https://doi.org/10.1016/j.tim.2019.05.006 PMID: 31178123

33. Benezra A. Race in the Microbiome. Sci Technol Human Values. 2020; 45:877–902. https://doi.org/10.1177/0162243920911998

34. Delgado AN, Baedke J. Does the human microbiome tell us something about race? Humanit Soc Sci Commun. 2021; 8:1–12. https://doi.org/10.1057/s41599-021-00772-3

35. Haelewaters D, Hofmann TA, Romero-Olivares AL. Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. PLoS Comput Biol. 2021; 17:e1009277. https://doi.org/10.1371/journal.pcbi.1009277 PMID: 34411090

36. Rochmyaningsih D. Did a study of Indonesian people who spend most of their days under water violate ethical rules? Science. 2018 [cited 2021 Nov 28]. https://doi.org/10.1126/science.aau8972

37. Amano T, González-Varo JP, Sutherland WJ. Languages Are Still a Major Barrier to Global Science. PLoS Biol. 2016; 14:e2000933. https://doi.org/10.1371/journal.pbio.2000933 PMID: 28033326

38. Ramírez-Castañeda V. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. PLoS ONE. 2020; 15:e0238372. https://doi.org/10.1371/journal.pone.0238372 PMID: 32936821

39. Armenteras D. Guidelines for healthy global scientific collaborations. Nat Ecol Evol. 2021; 5:1193–4. https://doi.org/10.1038/s41559-021-01496-y PMID: 34099901

40. Nuñez MA, Barlow J, Cadotte M, Lucas K, Newton E, Pettorelli N, et al. Assessing the uneven global distribution of readership, submissions and publications in applied ecology: Obvious problems without obvious solutions. J Appl Ecol. 2019; 56:4–9. https://doi.org/10.1111/1365-2664.13319

41. Baker K, Eichhorn MP, Griffiths M. Decolonizing field ecology. Biotropica. 2019; 51:288–92. https://doi.org/10.1111/btp.12663

42. Pettorelli N, Barlow J, Nuñez MA, Rader R, Stephens PA, Pinfield T, et al. How international journals can support ecology from the Global South. J Appl Ecol. 2021; 58:4–8. https://doi.org/10.1111/1365-2664.13815

43. Belhabib D. Ocean science and advocacy work better when decolonized. Nat Ecol Evol. 2021; 5:709–10. https://doi.org/10.1038/s41559-021-01477-1 PMID: 33981027

44. Antonelli A. Director of science at Kew: it's time to decolonise botanical collections. The Conversation. 19 Jun 2020 [cited 2021 Nov 28]. Available from: http://theconversation.com/director-of-science-at-kew-its-time-to-decolonise-botanical-collections-141070.

45. Noxolo P. Introduction: Decolonising geographical knowledge in a colonised and re-colonising postcolonial world. Area. 2017; 49:317–9. https://doi.org/10.1111/area.12370

46. Eichhorn MP, Baker K, Griffiths M. Steps towards decolonising biogeography. Front Biogeogr. 2020;12. https://doi.org/10.21425/F5FBG44795

47. Hazlett MA, Henderson KM, Zeitzer IF, Drew JA. The geography of publishing in the Anthropocene. Conserv Sci Pract. 2020; 2. https://doi.org/10.1111/csp2.270

48. de Vos A. The Problem of "Colonial Science." Scientific American. 1 Jul 2020 [cited 2021 Dec 12]. Available from: https://www.scientificamerican.com/article/the-problem-of-colonial-science. PMID: 34276078

49. Abdill RJ, Adamowicz EM, Blekhman R. International authorship and collaboration across bioRxiv preprints. Elife. 2020; 9:e58496. https://doi.org/10.7554/eLife.58496 PMID: 32716295

50. De Wolfe TJ, Arefin MR, Benezra A, Rebolleda GM. Chasing Ghosts: Race, Racism, and the Future of Microbiome Research. mSystems. 2021; 6:e0060421. https://doi.org/10.1128/mSystems.00604-21 PMID: 34636673

51. Microbiome Working Group. 18 Aug 2020 [cited 2021 Nov 30]. Available from: https://h3africa.org/index.php/microbiome-working-group/

52. Fadlelmola FM, Ghedira K, Hamdi Y, Hanachi M, Radouani F, Allali I, et al. H3ABioNet genomic medicine and microbiome data portals hackathon proceedings. Database. 2021; 2021. https://doi.org/10.1093/database/baab016 PMID: 33864455

53. Pylro VS, Mui TS, Rodrigues JLM, Andreote FD, Roesch LFW. Working Group Supporting the INCT Microbiome. A Step Forward to Empower Global Microbiome Research Through Local Leadership. Trends Microbiol. 2016; 24:767–71. https://doi.org/10.1016/j.tim.2016.07.007 PMID: 27498946

54. Díaz M, Jarrín-V P, Simarro R, Castillejo P, Tenea GN, Molina CA. The Ecuadorian Microbiome Project: a plea to strengthen microbial genomic research. Neotrop Biodivers. 2021; 7:223–37. https://doi.org/10.1080/23766808.2021.1938900

55. Adebamowo C, Collaborators N-A, Adebamowo S, Rotimi C. African Collaborative Center for Microbiome and Genomics Research (ACCME) Available from: https://h3africa.org/index.php/accme/

56. Dareng EO, Ma B, Famooto AO, Adebamowo SN, Offiong RA, Olaniyan O, et al. Prevalent high-risk HPV infection and vaginal microbiota in Nigerian women. Epidemiol Infect. 2016; 144:123–37. https://doi.org/10.1017/S0950268815000965 PMID: 26062721

57. Adebamowo SN, Ma B, Zella D, Famooto A, Ravel J, Adebamowo C, et al. Mycoplasma hominis and Mycoplasma genitalium in the Vaginal Microbiota and Persistent High-Risk Human Papillomavirus Infection. Front Public Health. 2017; 5:140. https://doi.org/10.3389/fpubh.2017.00140 PMID: 28695118

58. Gewin V. How to include Indigenous researchers and their knowledge. Nature. 2021; 589:315–7. https://doi.org/10.1038/d41586-021-00022-1 PMID: 33437060

59. Fox K. The Illusion of Inclusion—The "All of Us" Research Program and Indigenous Peoples' DNA. N Engl J Med. 2020; 383:411–3. https://doi.org/10.1056/NEJMp1915987 PMID: 32726527

60. Tsosie KS, Fox K, Yracheta JM. Genomics data: the broken promise is to Indigenous people. Nature. 2021:529. https://doi.org/10.1038/d41586-021-00758-w PMID: 33742179

61. CARE Principles of indigenous data governance—global indigenous data alliance. [cited 2021 Dec 9]. Available from: https://www.gida-global.org/care

62. Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data. 2019; 6:190021. https://doi.org/10.1038/sdata.2019.21 PMID: 30778255

63. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016; 44:D590–4. https://doi.org/10.1093/nar/gkv1322 PMID: 26656948

64. Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res. 2019; 47: D637–48. https://doi.org/10.1093/nar/gky1008 PMID: 30365027

65. Eckert EM, Di Cesare A, Fontaneto D, Berendonk TU, Bürgmann H, Cytryn E, et al. Every fifth published metagenome is not available to science. PLoS Biol. 2020; 18:e3000698. https://doi.org/10.1371/journal.pbio.3000698 PMID: 32243442

66. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009; 25:489–94. https://doi.org/10.1016/j.tig.2009.09.012 PMID: 19836853

67. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016:161–4. https://doi.org/10.1038/538161a PMID: 27734877

68. Organismal metagenomes. In: NCBI Taxonomy Browser [Internet]. [cited 2021 Jun 18]. Available from: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&id=410656&lvl=3&keep=1&srchmode=1&unlock

69. Yegorov S, Babenko D, Kozhakhmetov S, Akhmaltdinova L, Kadyrova I, Nurgozhina A, et al. Psoriasis Is Associated With Elevated Gut IL-1α and Intestinal Microbiome Alterations. Front Immunol. 2020; 11:571319. https://doi.org/10.3389/fimmu.2020.571319 PMID: 33117362

70. Kushugulova A, Forslund SK, Costea PI, Kozhakhmetov S, Khassenbekova Z, Urazova M, et al. Metagenomic analysis of gut microbial communities from a Central Asian population. BMJ Open. 2018; 8: e021682. https://doi.org/10.1136/bmjopen-2018-021682 PMID: 30056386

71. BioSample: Organism information. In: NCBI [Internet]. [cited 2021 Nov 27]. Available from: https://www.ncbi.nlm.nih.gov/biosample/docs/organism/

72. Walters WA, Reyes F, Soto GM, Reynolds ND, Fraser JA, Aviles R, et al. Epidemiology and associated microbiota changes in deployed military personnel at high risk of traveler's diarrhea. PLoS ONE. 2020; 15:e0236703. https://doi.org/10.1371/journal.pone.0236703 PMID: 32785284

73. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media; 2009. https://doi.org/10.1007/978-0-387-98141-3

74. Šavrič B, Patterson T, Jenny B. The Equal Earth map projection. Int J Geogr Inf Sci. 2019; 33:454–65. https://doi.org/10.1080/13658816.2018.1504949

75. South A. World Map Data from Natural Earth [R package rnaturalearth version 0.1.0]. 2017 [cited 2021 May 12]. Available from: https://cran.r-project.org/package=rnaturalearth