## Mapping total microbial communities and waterborne pathogens in

### household drinking water in China by citizen science and metabarcoding

1

2	Xinyi Wen <sup>1,2,3</sup> , Chutong Fang <sup>1,4</sup> , Lihan Huang <sup>1,5</sup> , Jiazheng Miao <sup>1,6</sup> , Yajuan Lin <sup>1,7</sup> *
3	
4	<sup>1</sup> Division of Natural and Applied Science, Duke Kunshan University, Kunshan, Jiangsu, China
5	<sup>2</sup> Nicholas School of the Environment, Duke University, Durham, NC, United States
6	<sup>3</sup> Department of Civil and Environmental Engineering, Pratt School of Engineering, Duke University,
7	Durham, NC, United States
8	<sup>4</sup> Blum Center for Developing Economies, University of California - Berkeley, Berkeley, CA, United
9	States

- <sup>5</sup>Department of Civil and Environmental Engineering, School of Engineering, Stanford University,
- 11 Stanford, CA, United States
- <sup>6</sup>Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, MA,
- 13 United States
- <sup>7</sup>Department of Life Sciences, College of Science, Texas A&M University Corpus Christi, Corpus
- 15 Christi, TX, United States
- 16
- 17 \* Corresponding author
- 18 E-mail: yajuan.lin@tamucc.edu
- 19
- 20 Keywords: Citizen Science, Tool kit, Household drinking water, Microbial communities,
- 21 Waterborne pathogens, Metabarcoding.

# 22 Abstract

23 Ensuring safe drinking water is one of the top priorities in public health as waterborne diseases 24 remain a global challenge. In China, microbial contamination in drinking water is of particular 25 concern and comprehensive survey/monitoring of the drinking water microbiome is necessary. 26 However, traditional culture-based microbial monitoring methods have significant limitations, and 27 nationwide tap water survey/monitoring in China would require significant resources. Here, a cost-28 effective and culture-independent citizen science approach was developed to sample the microbiome 29 in household drinking water (n = 50) from 19 provinces in China from December 2020 to August 30 2021, including a few opportunistic samples collected in situ right after extreme weather events such 31 as the 2021 Henan Floods and Typhoon In-Fa Landfall. Using a protocol optimized for low-biomass 32 samples, 22 out of 50 tap water samples were tested positive for microbial DNA. 16S rRNA gene 33 metabarcoding were conducted on pooled samples, yielding 7,635 Amplicon Sequence Variants 34 (ASVs), which revealed a diverse microbiome in household tap water. Alarmingly, pathogenic 35 bacteria including *Mycobacterium* spp., *Acinetobacter* spp., and *Legionella* spp. were detected in all 36 PCR positive samples. Despite the limited number of samples, a significant number of pathogenic 37 species (e.g., Salmonella enterica) and/or toxin-producing cyanobacteria (e.g., Microcystis spp.) were 38 detected in local tap water samples from Zhengzhou and Changzhou following the 2021 Henan 39 Floods and Typhoon In-Fa Landfall, respectively. Overall, this study underscores the utility of citizen 40 science in enhancing microbial monitoring and informs future public health strategies for water 41 safety.

42

# 43 **1** Introduction

With the growing demand for safe drinking water, microbial contamination in water resourcesand its related diseases is a focus of the world's water quality control today. Numerous studies show

that drinking water with microbial contamination can cause both acute and chronic damage to human health, with waterborne diseases posing a significant global health burden [1, 2]. There has been a rising incidence of waterborne human diseases such as gastrointestinal illness (e.g., diarrhea) and liver cancers which could be fatal [3-5]. The prevalence of human pathogens and toxin-producing bacteria in ambient and drinking water is a severe problem recognized globally [6]. Therefore, the scarcity of water resources that many countries including China have been facing is further exacerbated by water contamination [4, 5, 7].

53 A systematic review of China's drinking water sanitation from 2007 to 2018 shows that 54 microbial contamination in drinking water is a particular concern in China [7]. To assess the potential 55 health risks and obtain necessary data for microbiological water safety management in China, water 56 monitoring, which supplies a comprehensive understanding of spatiotemporal patterns in drinking 57 water microbiome, is urgently needed, especially at the point of use [8]. China CDC (Centers for 58 Disease Control and Prevention) at all levels sample drinking water twice a year to obtain copious 59 water quality data [7]. Due to the vastness of China, this nationwide water monitoring requires 60 considerable investments of capital, time, personnel, and technology [7]. Fortunately, previous 61 research has shown that citizen science can be an effective tool to increase spatial and temporal 62 coverage of data [9]. In the context of China, citizen science could be a cost-effective approach to 63 supplement China's national professional water monitoring systems [10, 11].

64 Citizen science can be broadly defined as a scientific approach in which the public (i.e., 65 people who have limited knowledge and skill in the targeted field) participate in the generation of 66 scientific knowledge [11-14], commonly in data collection [11-14]. Citizen science has a history of 67 several centuries in western societies, particularly in the environmental domain of which the breadth 68 is immense [10-12]. Among citizen-based environmental monitoring programs, water resources 69 monitoring is one of the major emerging fields [10]. It is especially active in Western countries [13, 67 15-17] since the provision of safe drinking water is a defining aspect of a developed country [18].

3

The National Water Quality Monitoring Council (NWQMC) website, for example, has over 350 volunteer monitoring groups registered across the U.S. in 2018 [15]. A more recent research pointed out that citizen science will play an increasing important role in promoting freshwater research, improving public understanding of the necessity to protect aquatic ecosystems, and engaging local communities and stakeholders in freshwater resource management [19].

76 However, several research gaps persist. First, compared with the long history and prosperity 77 of citizen science development in Europe and the US, little research has been done in developing 78 countries including China [15]. This can be attributed to multiple barriers, such as the late 79 commencement of citizen science initiatives in China, low participation levels, and issues concerning 80 data quality control, etc. Consequently, the cooperation between Chinese scientists and the public are 81 limited to a few citizen-based environmental projects mainly focusing on birds and plants 82 monitoring. However, with escalating concerns over environmental issue and the growing prevalence 83 of big data and social media in China, a new era of citizen science in China is emerging [20]. For 84 example, a recent study conducted by Wu et al. [21] revealed that most of the citizen science projects 85 in China aiming to improving water quality are still ongoing, indicating great potential of the citizen 86 science approach for water monitoring in the country.

87 Second, a global review of citizen science projects related to water quality measurement in 88 the past 20 years [15] shows a significant focus on chemical-physical parameters such as nutrients, 89 water transparency, and temperature, with very few addressing waterborne pathogens. Among these 90 waterborne pathogen assessments, only E. coli, an indicator organism of microbial contamination 91 risk in water used worldwide including China [6], was targeted. However, the efficacy of current 92 indicator organisms in representing the potential presence of pathogens in water resources is still a 93 subject of ongoing debate [2, 6]. Furthermore, a study suggests that China should incorporate 94 additional microorganisms as alternative contamination indicators to improve its water quality

95 management [5]. Therefore, it is necessary to holistically assess microbial community compositions96 and pathogens in drinking water.

97 Third, traditional microbiological monitoring of drinking water generally relies on culture-98 based methods, such as the heterotrophic plate counts (HPC) of certain microbes [22, 23]. However, 99 these methods can only account for a very small fraction (< 1%) of the drinking water microbiome 100 [23-25]. To overcome this limitation, this study employed a culture-independent approach: microbial 101 16S rRNA gene metabarcoding. This method enabled a holistic analysis of the microbial 102 communities in the water samples [23, 26, 27].

103 To the best of our knowledge, this study is one of the first to utilize a citizen science approach 104 to sample microbiome samples from household drinking water (i.e., tap water) across different 105 regions in China, through a simple yet standardized methodology. Following the protocol developed 106 by this study, citizen scientists (college student volunteers) collected 50 tap water samples from 107 households across 32 administrative regions within 18 provinces/regions of China from December 108 2020 to August 2021. This included several opportunistic samples collected after extreme 109 precipitation events such as 2021 Henan Floods and Typhoon In-Fa Landfall. We successfully 110 extracted microbial DNA from these low-biomass tap water samples, pooled all samples with 111 positive 16S rRNA genes PCR amplification signals, and conducted high-throughput sequencing of 112 the purified amplicons. This allowed us to characterize the total drinking water microbiome and 113 identify potential waterborne pathogens.

114

# 115 2 Materials and Methods

116 **2.1 Sample Collection** 

### 117 **2.1.1 Sites and Participants**

The sampling sites (Fig 1) were determined based on the coverage area (i.e., to cover as large an area as possible) and the possibility to recruit undergraduate student volunteers from the Duke Kunshan University (DKU) community. Due to the distribution of volunteers, samples mostly came from central and eastern China (latitude: ~22°N - 40°N; longitude: ~100°E - 122°E), including Beijing, Shandong Province, Jiangsu Province, Guangdong Province, etc.

123

#### 124 **Fig 1. Geographical locations of sampling sites.**

Map showing the geographic locations of the citizen science samples and the PCR amplification results (Software: Datawrapper). Red dots denote tap water samples showing positive PCR signals, blue dots indicate negative PCR signals, and yellow diamonds represent samples that were collected but did not pass quality control. Some locations were sampled multiples times at different time points.

130

131 All student volunteers were recruited after a simple screening process. First, their relevant 132 experiences (e.g., majors) were considered and those with natural science majors and lab experience 133 were preferred. Besides, ideal volunteers would go home or go on a trip to any place(s) in China 134 during the sampling period. Volunteers who met those two criteria or at least one of them were 135 recruited through one-to-one conversation on WeChat (a popular Chinese communication app similar 136 to WhatsApp). Communication between volunteers and investigators is a key component of this 137 study to ensure the quality of the samples as much as possible. This is the main reason why 138 volunteers were recruited from the DKU community – they can easily reach out to us either in person 139 or via WeChat once any problem emerges.

140

#### 141 **2.1.2 Preparation and Tool Kit**

142

A sampling protocol was developed based on Buxton et al.'s [16] research on citizen science

143 methods because volunteers in this study performed highly similar tasks (i.e., water sample collection 144 and filtration) as their research. A detailed version of the protocol is provided in the "Citizen science 145 sampling protocol and materials" (CS 1). Two innovative aspects of this protocol are (1) the easy-to-146 use and low-cost Corning syringe filters (instead of pumps and Sterivex filter cartridge) were adopted 147 for water sampling, and (2) the disinfection procedure was emphasized by listing all the possible 148 exposed objects and surface during the entire sampling process. Briefly, it is recommended to 149 conduct the experiment on the day or at most one day before shipping the sample. Volunteers first 150 put on the gloves and disinfect their hands as well as everything that they may touch during the 151 operation using the disinfectant wipes (Brand: Mian Zhi Run, 75% Ethanol and RO purified water). 152 Then, a sterilized 1 L stand-up bag with sodium thiosulfate to remove residual chlorine was unsealed 153 and filled with 1 L tap water. A disposable sterile 50 mL syringe was used to pass the sample water 154 across a sterile Corning syringe filter unit (CLS431229, 0.20 µm pore size, 28 mm diameter) and 155 refilled until 1 L of water has been filtered or the filter unit has become blocked. Afterwards, a 156 syringe of air was pushed through the filter unit to reduce the amount of residual water in the sealed 157 unit. The filter unit was then sealed in a Ziplock bag and kept frozen in a household freezer prior to 158 transportation to the laboratory (DKU Environmental Research Lab in Kunshan). For transportation, 159 the protocol requires burying the filter unit sample among four to five ice packs in a Styrofoam box. 160 Depending on the distance between the sampling site and Kunshan, along with the student's mode of 161 travel (i.e., same-day flights or high-speed train), samples were either delivered via express delivery 162 service or personally carried to the laboratory by volunteers.

To further clarify the procedures and reduce variability in sample collection, an 11-minute video tutorial (720p resolution) was created, providing volunteers with a visual guide on the tools and procedures for sample collection, filtration, storage, and shipping (Fig 2 & 3 and CS 3). Furthermore, in-person demonstrations were provided at DKU Environmental Research Lab for available volunteers following the demonstration model of Willis et al. [28]. A compact sampling kit was

168 distributed to each volunteer before their trip (Fig 2). Each kit (in a 78.7×51.2×49.2 inch Styrofoam 169 box) contained one set of all tools mentioned in the protocol and a sampling information form (CS 1 170 & 2) adapted from Buxton et al.'s [16]. This form collected information such as volunteer names, 171 sampling time, geographic coordinate of the sampling site obtained from cell phone, etc. Alongside 172 the form, volunteers were requested to label the Ziploc bag containing the sample with their names 173 for convenient sample tracking in the stage of data collection. Upon recipient in lab, each sample was 174 associated with an anonymous ID number and any data with identifiers was securely eliminated, 175 ensuring no specific identity-related information was present during data processing. The ID number 176 of each sample consists of two parts separated by an underscore -a two-letter abbreviation of the 177 sampling site (city) followed by a four-digit sampling date (S1 Table). For example, the sample 178 collected in Beijing on June 20 was named as "BJ\_0620" and the full names of other mentioned 179 samples are provided in the "Non-standard abbreviations" section.

180

#### 181 Fig 2. The sampling tool kit.

182

# Fig 3. Screenshots from the demonstration video showing the sampling process and the syringefilter.

185

An online sharing document (Tencent Sharing Document, with English translation provided in CS 3) was sent to each volunteer via WeChat, including the brief introduction to the research project, sampling kit, and sampling protocol as well as the links to the video tutorial and the electronic sampling information form (a Qualtrics survey was created to supplement the printed form). Notably, key elements in the document were highlighted in red and a larger font size for clarity. All the materials were provided in Chinese to facilitate understanding among the recruited student volunteers who were all Chinese. The online sharing document enabled them to easily see

any modifications that were made later.

Before the recruitment of volunteers, several trial runs of sampling were conducted by the authors, including sampling sites in Kunshan, Hangzhou, Changzhou, and Beijing. Along with the sampling protocol and kit, methods and conditions for sample shipping (i.e., express delivery vs. same-day bullet trains or planes) were tested and compared to identify the most effective procedures for DNA preservation.

- 199
- 200 2.1.3 Sample Collection and Filtration

The formal sample collection by student volunteers was conducted from December 2020 to August 2021. At each sampling site, one to four samples were collected. For sites with more than one sample, all valid samples were included to represent the site's microbiome. With the help of 25 volunteers, 50 household drinking water samples were collected from 32 administrative regions spanning 19 provinces/regions in China.

During this study, two extreme weather events were captured by opportunistic sampling. Tap water samples were collected by volunteers from Changzhou, Jiangsu Province and Hangzhou, Zhejiang Province before and after the landfall of Typhoon In-Fa (July 22 – 31, 2021 in China), and from Zhengzhou, Henan Province following an extremely destructive flood event (2021 Henan Floods).

Typhoon In-Fa (number 2106) was a Category 2 typhoon (SSHWS) which has been the secondwettest tropical cyclone ever recorded in China. As a tropical storm, it consecutively hit Putuo District of Zhoushan and Pinghu in Zhejiang Province on July 25 and 26 respectively. Typhoon In-Fa passed nearby Hangzhou from July 25 to 26 as a typhoon and Changzhou from July 26 to 27 as a tropical storm [29]. On the other hand, 2021 Henan Floods in Zhengzhou (July 17 – 23, 2021) was indirectly influenced by Typhoon In-Fa. From July 19 to 21, Zhengzhou suddenly encountered the

217 most severe rainstorm of the last 50 years which has led to extremely severe urban inland inundation, 218 floods, and landslide. In particular, the 24-hour precipitation in Zhengzhou from 8:00 on July 20 to 219 8:00 on July 21 was 624.1 millimeters (2.05 ft), which is only about 14.4 mm (0.57 in) lower than the 220 average annual precipitation in Zhengzhou in the past 22 years (638.5 mm) [30]. According to the 221 Investigation Report on 2021 Henan Floods issued by Ministry of Emergency Management of the 222 People's Republic of China [31], the flood impacted over 14 million individuals and killed 398 223 people in Henan, with the number of deaths in Zhengzhou accounting for 95.5% of the total death. 224 Moreover, it resulted in 16 million hectares of submerged agricultural land and direct economic 225 damages amounting to \$20.69 billion, accompanied by significantly higher indirect expenses.

226 For the areas affected by the typhoon landfall, student volunteers were instructed to sample 227 tap water once pre- and post-typhoon respectively within a week. As for Zhengzhou, since the sudden 228 occurrence of the extreme rainfalls and flooding was not anticipated and predicted by the China 229 Meteorological Administration [32], pre-flooding sampling was not arranged. However, we 230 successfully coordinated with and supplied the sampling kits to two volunteers residing in 231 Zhengzhou – one living in a seriously affected region and the other in a moderately affected region – 232 during the flood. They were instructed to collect three to four tap water samples on different days 233 throughout the week following the flood.

234

#### DNA Extraction, PCR Amplification, and **Bacterial** 2.2 **16S** 235 Metabarcoding 236

237 At the DKU Environmental Research Lab, all samples were stored in a -80 freezer prior to 238 processing. For each sample, the filter membrane was removed from the sealed syringe filter unit and 239 cut into eight strips using a sterile razor blade on a disposable petri dish. DNA extraction was then 240 conducted using the Qiagen DNeasy Plant Mini Kit following the manufacturer's instructions, with

the exception of an additional bead-beating step implemented to enhance extraction efficiency. In brief, the filter strips were placed into a 2 mL microcentrifuge tube where 0.2 g of 0.1 mm Zr bead and 400 microliters of lysis buffer AP-1 were added for beat beating at 2000 rpm for 5 minutes. The final extracted DNA of each sample from the kit was dissolved in 40  $\mu$ L Qiagen elution buffer and stored in a -80 $\Box$  freezer.

246 For PCR amplification, the V4 region of 16S rRNA gene was amplified by the universal 247 primer 515F (5'pairs (5'-GTGCCAGCMGCCGCGGTAA-3') and 805R 248 GACTACNVGGGTATCTAAT-3') with dual barcode index and heterogenous spacers [33, 34]. 249 KAPA HiFi PCR Kit and the manufacturer's protocol was adopted. All PCR reactions were 250 performed in triplicates with 25 µL of each reaction mixture. Agarose gel electrophoresis was then performed to visualize amplicon fragments and PCR products were purified using QIAquick PCR 251 252 Purification Kit (Qiagen) following the manufacturer's instructions. The purified PCR amplicons of 253 each sample was dissolved in 30  $\mu$ L elution buffer and stored in a -80  $\square$  freezer.

Finally, the concentration of the purified PCR amplicons was measured by  $\text{Qubit}^{\text{TM}} 4$ Fluorometer. Equimolar amounts of purified amplicons were pooled together and sent to *Genewiz* in Suzhou for an Illumina Miseq (250 PE) sequencing run.

257

# 258 2.3 Auxiliary Data

Previous studies found that drinking water microbiome exhibited seasonality which was correlated with temperature [35, 36]. Given the difficulties in measuring water and indoor temperature with the citizen science approach, outdoor temperature was collected as the indicator of seasonal environmental temperature change. The outdoor temperatures when collecting valid samples (n = 40) were retrieved from Weather Underground [37], an online portal that provides local weather data on an hourly basis.

- In addition, to assess the potential risks of harmful bacteria detected in sampled tap water, information such as drinking water standards and microbial fact sheets was retrieved from environmental and health authorities including The United States Environmental Protection Agency (EPA) [38, 39], and World Health Organization (WHO) [40].
- 269

# 270 2.4 Data Analysis

271 **2.4.1 Sequence Processing** 

272 To profile the microbiome in sampled tap water, 16S rRNA gene sequences were processed 273 by a series of bioinformatics tools. First, paired-end sequencing reads with dual indices were 274 demultiplexed and then trimmed to remove barcodes and primers using *Cutadapt* [41]. The resulting 275 reads were then further processed following the DADA2(v1.16) pipeline [42]. Specifically, using the 276 embedded functions in DADA2, quality filtering was performed before merging paired-end 277 sequencing reads. After chimera checking, Amplicon Sequence Variant (ASV) was identified, and 278 the abundance table for each sample was constructed. Finally, the "assignTaxonomy" function in 279 DADA2 was used to assign taxonomy to each ASV based on Silva r138 reference database (DOI 280 10.5281/zenodo.4587955).

281

#### 282 **2.4.2 Microbial Community Structure**

The relative abundance (RA) of all ASV in each sample was calculated and the following analyses of microbial compositions were based on RA.

Apart for the samples collected after Typhoon In-Fa and 2021 Henan Floods ("post-weather samples"), the remaining samples ("normal samples") were divided into three nearly equal groups based on the outdoor temperatures to account for potential impacts of climate and seasonality: (a) High (H): T > 25 °C; (b) Medium (M): 15 °C < T  $\leq$  25 °C; (c) Low (L): T  $\leq$  15 °C (Table 1). The

- 289 mean RA of each ASV in the three temperature categories was calculated for the ASV-level analyses.
- 290 The most abundant ASVs in each category are presented in the supporting material S1 Text and S1
- 291 Fig.
- 292

### 293 Table 1 Sampling seasons: outdoor temperature categories.

Category	Range (°C)	Number of samples			
High	T > 25	7			
Median	$15 < T \leq 25$	7			
Low	T ≤ 15	6			

294

### 295 **2.4.3** Assessment of Potential Waterborne Bacterial Pathogens

The bacteria genera in the dataset containing pathogenic species were selected based on *Aquatic Pollution: An Introductory Text* [43] and "Guidelines for Drinking-water Quality (4<sup>th</sup> Edition)" by WHO [40]. Subsequently, the taxonomy of the resulting ASVs was further validated using the BLAST+ tool [44] against the NCBI database during February 2022. Based on our criteria, only BLAST results with percent identity (p-ident) > 97% [45], expect value (E-value) < 10e-100 [46], and query cover equals 100% were considered reliable results.

302 Subsequently, all potential pathogenic genera/species were grouped into two categories based 303 on their occurrence in the drinking water samples. Those detected in more than 30% of all samples 304 (i.e., 7 samples) were categorized as "common pathogens" while the rest were referred to as "rare 305 pathogens."

306

#### 307 **2.4.4 Statistical Analyses**

All statistical analyses were performed in R (4.1.2 and 4.2.3) and a p-value threshold of 0.05
was considered significant.

310 Fisher's exact test [47] was conducted to determine whether there is a significant relationship

311 between PCR signals (i.e., positive or negative) and average annual rainfall (AAR) (humid regions: >

312 800 mm, non-humid regions: < 800 mm).

313 Alpha-and beta-diversities were calculated at the ASV level to further describe the microbial 314 communities in the samples. Alpha diversity refers to diversity (richness or evenness) within an 315 ecosystem at the local small scale, while beta diversity measures the amount of differentiation 316 between different ecosystems or local species communities [48]. For alpha-diversity of tap water 317 microbial communities, each library was resampled in equal depth, and Chao1, Fisher, Shannon, and 318 Simpson diversity indices were then calculated from observed read counts of ASVs using the 319 "Phyloseq" package (version 1.38.0) in R [49]. The Shannon and Simpson indices including both 320 ASV richness and evenness were computed because of their reduced sensitivity to differences in 321 sample depth [50, 51]. To calculate beta-diversity, a principal coordinates analysis (PCoA) was 322 performed based on the distance matrix of Bray-Curtis dissimilarity to visualize the ordination 323 among the samples by *plot\_ordination()*. To test the significant differences among groups of tap 324 water samples, permutational multivariate analysis of variance (PERMANOVA) and a post-hoc 325 pairwise Adonis test were performed with adonis2 function [52] and *pairwise.adonis()* function [53] 326 respectively. To check if within-group variation is confused with among-group variation [54], a 327 permutation test for homogeneity of multivariate dispersions (PERMDISP) [55] was performed with 328 the betadisper function. In PERMANOVA and PERMDISP, the number of permutations was set to 329 999.

To analyze the compositions and influencing factors of waterborne pathogens in the tap water, several correlation tests based on Spearman's rank correlation coefficients were conducted. First, Mantel test [56] was conducted to examine correlations between total microbial communities (top 100 ASVs), potential pathogens, and outdoor temperatures based on Bray-Curtis dissimilarity, Euclidean Distance, and Haversine Distance matrices (permutations = 9999). Second, linear

regression fits between outdoor temperatures and RAs of total potential pathogens were examined and visualized with *ggscatter()* in the "ggpubr" package. Third, linear correlations between RAs of potential pathogens, RAs of total microbial communities (Top10 ASVs), and alpha diversity of total microbial communities were assessed with *cor()*. The corresponding p values and confidence intervals were computed with *cor.mtest()*. The matrix with significance level codes was visualized with *corrplot.mixed()*.

- 341
- 342 **3 Results and Discussion**

# 343 **3.1 Sample Validity and PCR Signals**

344 Out of the 50 drinking water samples, 40 passed our quality control and they were collected 345 from 28 administrative regions across China (Table S1). 10 samples were excluded from the study 346 because of either (1) lab processing errors, or (2) improper handling during shipping and/or storage, 347 which may have compromised DNA quality. Among the 40 valid samples, 29 samples showed 348 positive PCR signals which were collected from 16 cities in 10 provinces, 4 municipalities, and 349 Macau (Fig 1). Fisher's exact test (S2 Table) reveals that no significant difference in PCR signals 350 existed between the two rainfall groups (P = 0.69). Interestingly, all the three tap water samples 351 collected in Zhejiang Province, including the post-typhoon Hangzhou sample showed no PCR signal. 352 This might be due to the high quality of source water in Zhejiang Province [57] and effective 353 drinking water treatment in those two developed cities (Hangzhou and Ningbo). Alternatively, factors 354 such as new/clean plumbing systems, suitable plumbing materials that do not support the growth of 355 microorganisms, and shorter water stagnation time in the plumbing could also account for the 356 negative PCR results [35, 58, 59].

357

# 358 **3.2 Citizen Science Sampling Approach**

This study is among the few studies on tap water microbiome via citizen science which can serve as a proof of concept for national-scale microbiological monitoring of tap water using citizen science and demonstrate its competitive advantages compared to non-citizen science sampling method.

363 Firstly, the study illustrates the cost-effectiveness and extensive coverage across China of the 364 approach. With decentralized volunteer participation and affordable sampling kits, the sampling sites 365 spanned 32 regions and various seasons, ensuring spatial and temporal diversity. It is worth 366 mentioning that due to the adoption of Corning® syringe filter (~\$2.5-4/unit) [60], the cost per 367 sampling kit was under \$7, which was around 1 - 1.5 times cheaper than using the traditional 368 MilliporeSigma® Sterivex filter (~\$8–13/unit) for aquatic microbiome sampling [61]. Our DNA 369 extraction protocol was optimized for this low-cost filter. In addition, this citizen science sampling 370 approach significantly reduces travel and personnel expenses. Although the volunteers' sample 371 collection lacked strict oversight, the high validity of 40 out of 50 drinking water samples passing 372 quality control and the above consistency in results supports the method's applicability and reliability.

373 Furthermore, due to the flexibility of this sampling approach, we conducted timely 374 opportunistic sampling of extreme weather events - specifically, Typhoon In-Fa and 2021 Henan 375 Floods - by leveraging local volunteers in Hangzhou, Changzhou, and Zhengzhou. This citizen 376 science approach proved to be particularly fast responding and effective given the limited funding 377 and the urgent nature of these events. Additionally, the sample storage tests (S3 Text) confirmed that maintaining samples for 7 days at -18°C did not generate false positives in the absence of bacteria 378 379 during collection, ensuring accurate pathogen identification and reliability of using household freezer 380 for storage, thereby enhancing the flexibility and feasibility of the citizen science sampling approach.

381 Moreover, the citizen science approach aimed to bridge the gap between the scientific 382 community and the public, raising awareness of pathogen pollution in drinking water and educating 383 the public on health protection measures.

384

# 385 3.3 Sequence Reads, ASVs, and Taxonomy Classification

386 The DNA of microorganisms in 29 household drinking water samples were successfully 387 extracted, PCR amplified with 16S primers and submitted for Illumina sequencing. However, 388 following the demultiplexing process, five samples did not yield significant reads (all < 50 per 389 library) and were excluded from the dataset, leading to a loss of two sampling sites (Shenzhen and 390 Xiamen). In addition, two samples collected after the 2021 Henan Floods were excluded after the 391 quality filtering and chimera checking because of relatively low reads (<3000 per library) compared 392 to others. As a result, the 16S rRNA gene amplicon dataset in this study contained 22 samples after 393 the preliminary processing (S1 Table). 394 In DADA2 quality filtering, proportions of output and input reads of 13 (59.1%) samples 395 were lower than 50%, indicating low quality of raw DNA reads. This may be due to 1) potential 396 degradation of DNA samples during sampling, shipping, or DNA processing; 2) residual DNA from

dead bacteria cells, which are harmless to humans. After quality filtering and chimera removing,
sequencing reads per sample ranged from 9,794 to 150,656, averaging at 60,884.

From the resulted dataset, 7635 ASVs were identified. According to the DADA2 taxonomy classification based on Silva r138 reference database, 100%, 99.5%, 96.8%, 92.3%, 81.8%, and 59.4% of the sequencing reads could be assigned to kingdom, phylum, class, order, family, and genus level respectively. At the kingdom level, 99.9% of the reads were assigned to domains of bacteria (1,294,678 reads) and archaea (43,816 reads), while the other 0.1% of the sequences belonged to eukaryotes (739 reads).

405

# 406 **3.4 Microbial Community Structure**

#### 407 **3.4.1 Compositions of Total Microbial Communities**

408 52 microbial phyla including 46 bacterial phyla and 6 archaeal phyla were detected in the 409 household drinking water samples (n = 22). In all samples, 9 bacterial phyla and 1 archaeal phylum 410 accounted for over 90.7% of total taxonomically assigned reads at phylum level (Table 2). All the ten 411 phyla are both abundant and prevalent (i.e., occur in 91-100% of all samples), indicating a relatively 412 even microbial distribution pattern in the samples at the Phylum level. The five most dominant 413 bacterial phyla were *Proteobacteria* (mean percentage  $\pm$  SD: 55.0%  $\pm$  19.8%), *Planctomycetota* 414  $(10.5\% \pm 7.9\%)$ , Acidobacteriota  $(7.0\% \pm 6.0\%)$ , Actinobacteria  $(5.9\% \pm 6.6\%)$ , and Cyanobacteria 415  $(4.7\% \pm 3.5\%)$ , comparable to a previous study in China [57]. Three of them were reported tolerant 416 to drinking water treatment and distribution processes except for *Planctomycetota* and 417 Acidobacteriota [57]. Interestingly, this bacterial composition differs from the primary bacterial 418 assemblages revealed by several highly-cited studies conducted in other countries including the U.S. 419 and Portugal, which predominantly feature Proteobacteria, Actinobacteria, and Bacteroidetes [62-420 64]. This implies a unique microbiome intrinsic to China's drinking water systems.

#### 421 Table 2 Number of sequences, ASVs, and genera for the top 10 phyla in the water samples.

Phylum	Sequence <sup>a</sup>	ASV	Genus
Proteobacteria	697642 (52.3%)	2693	281
Planctomycetota	164487 (12.3%)	998	23
Actinobacteriota	82262 (6.2%)	279	56
Acidobacteriota	77415 (5.8%)	296	17
Cyanobacteria	72219 (5.4%)	402	25
Crenarchaeota	43624 (3.3%)	31	5
Bacteroidota	38869 (2.9%)	603	61
Verrucomicrobiota	33769 (2.5%)	411	27
Bdellovibrionota	31056 (2.3%)	400	8
Desulfobacterota	21608 (1.6%)	59	11
Others	70371 (5.3%)	1209	93
N.A.	6133 (0.5%)	254	/

Those three represent the coverage, diversity, and the genus spectrum of microbial community in the water samples respectively. <sup>a</sup> The percentage of sequences was the phylum sequences in the total of assigned sequences at the phylum level which was 1,333,322.

425 The phylum-level taxonomic composition for each sample is detailed in Fig 4 and S3 Table. 426 In this study, *Proteobacteria* (class  $\alpha$ - and  $\gamma$ -*Proteobacteria*) was the most predominant phylum in 20 427 samples and the second most in others. It accounts for 29.9 - 99.0% of the reads in each sample, with 428 the highest relative abundance detected in Huizhou (HZ\_0219). Among all the samples, 429 Sphingomonas was the most abundant genus of Proteobacteria. This result was consistent with the 430 previous study [57] that found *Proteobacteria* to be dominant in tap water collected mostly from 431 central and eastern China, and the genus Sphingomonas grew during chlorination [65] or 432 monochloramine treatment [66].

# Fig 4. Taxonomic composition and relative abundance of microbiota in the sampled household drinking water in China at the phylum level.

435 "\*" denotes the samples collected after extreme rainfall events. Normal samples are grouped based
436 on the outdoor temperature. Only the top 10 phyla (1-10: bottom to top in the legend) and not
437 assigned (NA) ones are shown.

438

439 However, two samples CZ 0220 (Changzhou, Feb. 20) and NC 0210 (Nanchang, Feb.10) 440 were dominated by Crenarchaeota, a common archaeal phylum. Specifically, Crenarchaeota 441 accounted for 33.3% and 35.0% with the genus Candidatus Nitrosotenuis (32.0%) and Candidatus 442 Nitrosotalea (34.9%) being most abundant in samples CZ\_0220 and NC\_0210 respectively. This 443 showed that in addition to a variety of bacteria, archaea can also grow in tap water, which could be 444 supported by the studies which detected the archaeal phylum *Crenarchaeota* in drinking water 445 distribution systems or drinking water-related environments [24, 67-70]. In particular, Inkinen et al. 446 [67] found a high abundance of archaeal reads from the genus Candidatus Nitrosotenuis and 447 Candidatus Nitrosotalea in drinking water distribution systems supplying non-disinfected waters.

448 This suggests that the disinfection processes of samples CZ\_0220 and NC\_0210 may be less 449 effective compared to others.

450 Interestingly, compared to the winter Changzhou sample CZ 0220, the pre-typhoon CZ 0712 451 and post-typhoon samples CZ\_0728 collected from the same city household in July were much more 452 similar in overall composition at the phylum level, indicating a seasonal effect. However, the post-453 typhoon sample CZ\_0728 exhibited higher Actinobacteria (increased from 2.2% to 7.1%) and 454 Cyanobacteria (increased from 5.2% to 14.9%), which are the phyla containing potential waterborne 455 pathogens and the species that produce cyanotoxins, respectively. Elevated levels of the pathogen 456 Mycobacterium spp. (more details in Section 3.5) as well as toxin-producing Cyanobacteria spp. 457 were observed in post-typhoon sample CZ\_0728. Specifically for cyanobacteria, *Microcystis* spp. 458 were higher in RA while Cylindrospermopsis sp. and Dolichospermum sp. appeared after the 459 typhoon event. Other toxic species of Cyanobacteria, including Aphanizomenon sp. and Anabaena 460 sp., were detected in normal samples collected from Shanghai, Lanzhou, Xi'an, etc. Many 461 Cyanobacteria spp. from those genera can produce a variety of cyanotoxins such as Microcystins and 462 Cylindrospermopsin, which can cause liver and kidney damage and have potential carcinogenicity 463 [39]. Similarly, a substantial rise in the RA of *Cyanobacteria* was reported in treated water samples 464 collected from a drinking water treatment plant in Jiangsu Province after Typhoon Lekima in August 465 2019 (P < 0.05) [71]. In this study, although remained detectable, the RA of *Cyanobacteria* in post-466 typhoon sample CZ\_0728 decreased to the pre-typhoon level by the third day after the typhoon event.

467

#### 468 **3.4.2 Beta-Diversities**

The beta diversity of the total microbial communities in sampled tap water is shown in Fig 5 (n = 22), S2 Text (n = 20), and S2 Fig (n = 20). According to Fig 5, Principal Axis 1, 2, and 3 for PCoA (Bray-Curtis dissimilarity) represent 20.4%, 14.5%, and 8.1% of the variation among the

472 samples respectively. The PCoA ordination illustrates that despite the different sampling sites, the 473 samples in High T category clustered closely except BJ 0620 which was greatly dominated by 474 Proteobacteria (87.5%). In contrast, samples in the other two categories, especially Medium T 475 exhibited dispersive distributions. According to PERMANOVA and PERMDISP (S4 Table), the influence of outdoor temperatures was statistically significant ( $R^2 = 0.24$ , adjusted P = 0.012 between 476 477 High T and Low T), and the differentiation was not due to differences in group dispersions (P =478 0.17). In addition, a Mantel test (S5 Table) indicates a strong correlation between microbial 479 community structures and outdoor temperatures (Spearman's rho = 0.898, P = 0.0001).

#### 480 Fig 5. Principal coordinates analysis (PCoA).

Based on the distance matrix of Bray-Curtis dissimilarity of microbial profiles (ASV-based) among all samples (n = 22). HW: the weather samples collected in summer; CZ: samples collected in the same household in Changzhou; ZZ: samples collected in two nearby households in Zhengzhou.

484

485 Research spanning China, the Netherlands, and the U.S. has consistently found that temporal 486 temperature variation, primarily driven by seasonal changes, is a controlling factor for the tap water 487 microbiome [35, 72, 73]. Complementing this, another study in China revealed that increasing air 488 temperature through indoor heating could lead to dramatic changes in the composition of the 489 bacterial community in overnight stagnant tap water [74]. These findings challenge the previous 490 discovery by Han et al. that only a weak correlation was found between air temperature and tap water 491 bacterial community in China (Spearman's rho = 0.088, P = 0.092) [57]. This might be caused by 492 differences in sampling year, sites, and geography coverage. Notable, the source water microbiome 493 could also exert a substantive influence on tap water microbiome [57]. Beyond the citizen science 494 approach, more comprehensive metadata collection associated with microbiome sampling, as well as 495 well-controlled studies systematically covering spatial and temporal variations, are needed to reveal 496 the factors shaping the drinking water microbiome.

497

# 498 **3.5** Potential Pathogenic Bacteria in Drinking Water Microbiome

In total, six bacteria genera containing pathogenic species and three pathogenic species were detected in all the PCR positive samples (n = 22) (Table 3). Five genera and one species that occurred in more than 30% of the samples (i.e., 7 samples) were categorized as common pathogens while the rest were grouped into rare pathogens (S6 Table). It is worth mentioning that the overall pattern of pathogens in the normal samples still holds when the two post-weather samples were included. This finding suggests that pathogen contamination in tap water could be a widespread phenomenon [35, 75].

506	Table 3 (	Occurrence and	mean relative	abundance	(RA)	of major po	tential pathogens
-----	-----------	----------------	---------------	-----------	------	-------------	-------------------

Pathogen Species	Occurrence	Mean RA
Mycobacterium spp.	22 (100.0%)	2.67E-02
Acinetobacter spp.	22 (100.0%)	1.43E-02
<i>Legionella</i> spp.	22 (100%)	4.59E-03
Brevundimonas spp.	20 (90.9%)	6.12E-03
Leptospira spp.	11 (50.0%)	2.04E-04
Escherichia coli	8 (36.4%)	5.80E-04
Bacillus spp.	3 (13.6%)	2.01E-05
Aeromonas hydrophila	2 (9.1%)	4.76E-05
Salmonella enterica	1 (4.5%)	1.79E-04

507

The distribution of these potential pathogens within each water sample are detailed in Fig 6 and the BLAST+ results of the potential pathogenic ASVs are provided in S7 Table. Common pathogens' mean relative abundance (RA) in all tap water samples ranged from 0.02% to 2.67% (normal samples: 0.02% - 1.95%) while that of rare pathogens was extremely low (S10 Table). *Mycobacterium* spp. (mean RA 2.67%, including rainfall events), *Acinetobacter* spp. (1.43%), and

*Legionella spp.* (0.46%) occurred in all the samples while *Leptospira* spp. (0.02%) were found in half of the samples. Notably, *Escherichia coli*. ASVs (0.06%) (E-value = 6e-132, Percent identity = 100%) were detected in 36.4% of the samples. This is alarming since *E. coli* O157:H7 is the most common pathogenic strain that causes severe gastrointestinal (GI) illness in humans [76]. In addition, two *Brevundimonas* species, *B. vesicularis* and *B. diminuta*, are particularly considered emerging global opportunistic pathogens [77] were found in most of the samples (90.9%). Compared to normal samples, three more pathogenic bacteria species were detected in the post-weather samples, including

520 Salmonella enterica, Brevundimonas diminuta, and Aeromonas hydrophila.

#### 521 Fig 6. Relative abundance of potential bacterial pathogens

Pathogens in the figure legend are in the descending order of mean RA (from top to bottom).
Samples are in the descending order of total RA of all pathogens in each sample (from left to right).
Pre- and post-weather samples are labeled in black boxes and red boxes, respectively. Changzhou
samples (typhoon) are in solid boxes, and Zhengzhou samples (floods) are in dashed boxes.

526

A summary of representative potential pathogens in drinking water systems and their related diseases is provided in S8 Table [1]. The discussions of several representative pathogen species for science communication purposes as well as the potential environmental drivers of waterborne pathogens are provided in S4 Text, S5 Text, and S3 Fig.

*E. coli*, a common fecal indicator bacteria [40], was widely detected by this study. The most polluted tap water samples are CZ\_0728 (RA: 0.61%), BJ\_0620 (0.51%), ZZ\_0802 (0.05%), CZ\_0712 (0.03%), and JN\_0219 (0.02%). The elevated RA of *E. coli* in the two post-weather samples suggests that the contamination might be related to the extreme rainfall events. Notably, the proportion of *E. coli* in BJ\_0620 was dramatically higher than that of other normal samples. It is worth mentioning that the traditional *E. coli* tests can be a generally reliable indicator of enteropathogenic serotypes in drinking water, but potential viable while non-culturable *E. coli* cells

538	could result in	underestimations	of the actual	water contamination	[78]	. Therefore	, it is recommended
-----	-----------------	------------------	---------------	---------------------	------	-------------	---------------------

to use PCR or quantitative PCR (qPCR) methods for the monitoring of *E. coli* [79].

Additionally, the RA of total Legionella spp. was most abundant in ZZ\_0802 (1.32%), 540 541 CZ\_0220 (1.25%), and ZB\_0502 (1.09%). Almost all species in the genera Legionella are thought to 542 be potential human pathogens, but L. pneumophila (on Contaminant Candidate List 5 - CCL 5) is the 543 main cause of Legionnaires' disease (pneumonia) and Pontiac fever (a milder infection) [38, 40]. 544 Potential L. pneumophila ASVs was detected in 22.7% (5) of the tap water samples (3 from Low T, 1 545 from Median and High T respectively). Among those samples, the highest L. pneumophila RA was 546 observed in ZZ\_0219 (RA: 0.237%), followed by MA\_0524 (0.035%), NC\_0210 (0.019%), 547 XA\_0217 (0.011%), and SH\_0411 (0.010%). Moreover, some other pathogenic species such as L. 548 oakridgensis and L. maceachernii occurred in TJ\_0715, LZ\_0822, XA\_0217, LJ\_0722, and 549 ZZ 0802.

*Salmonella* spp., a highly pathogenic bacteria, was detected in one sample despite its known susceptibility to disinfection. Notably, *Salmonella enterica* (ASV 3082, RA: 0.39%) was observed in the post-flood tap water microbiome from Zhengzhou (ZZ\_0802). However, the potential health risks remains uncertain, as the severity of the disease depends on the serotype and host factors of *Salmonella* [40]. Nonetheless, the presence of *Salmonella enterica* after 2021 Henan Floods indicates potential fecal contamination in the household drinking water after an extreme weather event.

Spearman's correlations between potential pathogens and total microbiome alpha diversity indexes are shown in Fig 7. When the impacts of extreme rainfall events were not considered (n = 20), the RA of *Brevundimonas* spp. positively correlated with multiple pathogens and alpha diversity indexes, especially with *Mycobacterium* spp. (r = 0.84, P < 0.001) and the total RA of potential pathogens (r = 0.67, P < 0.001). This suggests *Brevundimonas* spp. to be a potential good candidate for indicator organisms. Notably, a significant correlation between the low alpha diversity of the microbial community and the samples positive for the presence of *E. coli* (Chao1: r = -0.24, P < 0.05;

563 Shannon: r = -0.21, P < 0.05) was observed. Similarly, Chopyk et al. [80] found the same pattern in 564 their pre-harvest cattle hide samples. Therefore, it is worth understanding the relationship between 565 indigenous water microbiomes and pathogenic E. coli, which might contribute to the 566 control/prevention strategies for EHEC in tap water as well as indicators for risk assessment.

567 Fig 7. Spearman's correlation of potential pathogens detected in normal samples (n = 20).

568 1-3 asterisk(s) denote the significance level of 0.05, 0.01, and 0.001, respectively.

- 569

#### **Limitations and Future Work** 3.6 570

571 The current sampling procedure exhibits several limitations and could be improved in the 572 future. First, to further minimize bias during the sample collection process, closer supervision of the 573 volunteers (e.g., through cell phone video recording) and duplicate sample collection from the same 574 location are needed. Second, systematic time-series sampling in the same location is necessary to 575 understand temporal microbial patterns. Additionally, to pinpoint contamination sources, future 576 studies could sample from water treatment system effluents to discern if contamination stems from 577 treatment or pipeline issues. Moreover, collecting a comprehensive set of metadata (e.g. water 578 temperature, nutrients, pH, etc.) associated with the microbiome sampling will help reveal the 579 environmental factors shaping the drinking water microbiome.

580 To improve the success rate of sample collection, it is important to implement stricter controls 581 on shipping temperatures and to collaborate with shipping companies. This collaboration will help 582 reduce sample transportation costs, a critical factor in expanding the citizen science outreach beyond 583 university affiliates. Meanwhile, developing an online platform could be helpful to disseminate the 584 research results to the public, thereby promoting science education and citizen engagement.

585

#### **Conclusions** 4 586

Ensuring household drinking water safety is vital for public health due to the risks associated with microbial contamination. Combining citizen science sampling and culture-independent metabarcoding, this proof-of-concept study provided profiles of tap water microbiome and waterborne pathogens from various locations in China. This method, which extends beyond basic water collection and observation (e.g., water transparency), suggests that well-structured collaborations between professional agencies and citizen science can effectively monitor water guality on a broad scale.

594 In this study, a total of 7,635 prokaryotic ASVs were detected from 40 household drinking 595 water samples from 28 cities of 18 provinces/regions in China. Despite the limited number of 596 samples, extreme weather events, such as typhoons and floods, may exacerbate the presence of 597 pathogens (e.g., Escherichia coli and Salmonella spp.) as well as toxin-producing cyanobacteria 598 *Microcystis* in local tap water. This is particularly concerning in the current context of climate 599 change, which may increase the frequency and intensity of such conditions. Additionally, this study 600 highlights the valuable role that citizen science can play in advancing our understanding of 601 environmental health risks and shaping public health policy.

602

# 603 Acknowledgements

604 First, we would like to thank Xiuwen Li for her support in sampling and data management. In

addition, we are grateful to Shuai Gu and Yuchen Meng for their help with sample processing. Last

606 but not the least, we wish to extend our special thanks to all the student volunteers who helped collect

- tap water samples across China and made this citizen science project possible.
- 608

# 609 **References**

610 Ramirez-Castillo FY, Loera-Muro A, Jacques M, Garneau P, Avelar-Gonzalez FJ, Harel J, et 1. 611 al. Waterborne pathogens: detection methods and challenges. Pathogens. 2015;4(2):307-34. Motlagh AM, Yang Z. Detection and occurrence of indicator organisms and pathogens. Water 612 2. 613 Environment Research. 2019;91(10):1402-8. 614 Zhang J, Mauzerall DL, Zhu T, Liang S, Ezzati M, Remais JV. Environmental health in 3. China: progress towards clean air and safe water. Lancet. 2010;375(9720):1110-9. 615 616 Yu X, Lin H, Zhang S. Microbial contamination in drinking water: occurrence, measurement 4. 617 and control. Bulletin of the Chinese Academy of Sciences BCAS. 2012;26(1):84-91. 618 Wen X, Chen F, Lin Y, Zhu H, Yuan F, Kuang D, et al. Microbial indicators and their use for 5. 619 monitoring drinking water quality—A review. Sustainability. 2020;12(6):2249. 620 Pandey PK, Kass PH, Soupir ML, Biswas S, Singh VP. Contamination of water resources by 6. 621 pathogenic bacteria. AMB Express. 2014;4(1):51. 622 Wang T, Sun D, Zhang Q, Zhang Z. China's drinking water sanitation from 2007 to 2018: A 7. 623 systematic review. Sci Total Environ. 2021;757:143923. 624 Altenburger R, Brack W, Burgess RM, Busch W, Escher BI, Focks A, et al. Future water 8. 625 quality monitoring: improving the balance between exposure and toxicity assessments of real-world 626 pollutant mixtures. Environmental Sciences Europe. 2019;31(1):1-17. 627 Pocock MJ, Roy HE, Fox R, Ellis WN, Botham M. Citizen science and invasive alien species: 9. 628 predicting the detection of the oak processionary moth Thaumetopoea processionea by moth 629 recorders. Biological conservation. 2017;208:146-54. 630 Albus K, Thompson R, Mitchell F. Usability of Existing Volunteer Water Monitoring Data: 10. 631 What Can the Literature Tell Us? Citizen Science: Theory and Practice. 2019;4(1). Commission E. COMMISSION STAFF WORKING DOCUMENT: Best Practices in Citizen 632 11. 633 Science for Environmental Monitoring. European Commission; 2020. 634 Brouwer S, Van der Wielen PW, Schriks M, Claassen M, Frijns J. Public participation in 12. 635 science: the future and value of citizen science in the drinking water research. Water. 2018;10(3):284. 636 Roy HE, Pocock MJ, Preston CD, Roy DB, Savage J, Tweddle J, et al. Understanding citizen 13. 637 science and environmental monitoring: final report on behalf of UK Environmental Observation 638 Framework. 2012. 639 Harper L. The Importance of Citizen Scientists American Forests2018 [Available from: 14. https://www.americanforests.org/blog/the-importance-of-citizen-scientists/. 640

641 15. Baalbaki R, Ahmad SH, Kays W, Talhouk SN, Saliba NA, Al-Hindi M. Citizen science in 642 Lebanon-a case study for groundwater quality monitoring. R Soc Open Sci. 2019;6(2):181871. 643 Buxton AS, Groombridge JJ, Griffiths RA. Comparison of two citizen scientist methods for 16. collecting pond water samples for environmental DNA studies. Ctizen Science: Theory and Practice. 644 645 2018:3(2). Ho SY, Xu SJ, Lee FW. Citizen science: An alternative way for water monitoring in Hong 646 17. 647 Kong. PLoS One. 2020;15(9):e0238349. 648 Ashbolt NJ. Microbial Contamination of Drinking Water and Human Health from 18. 649 Community Water Systems. Curr Environ Health Rep. 2015;2(1):95-106. 650 19. Metcalfe AN, Kennedy TA, Mendez GA, Muehlbauer JD. Applied citizen science in freshwater research. Wiley Interdisciplinary Reviews: Water. 2022;9(2):e1578. 651 652 Zhang J, Chen S, Chen B, Du Y, Huang X, Pan X, et al. Citizen science: integrating scientific 20. 653 research, ecological conservation and public participation. Biodiversity Science. 2013;21(6):738. 654 Wu Y, Washbourne C, Haklay M. Citizen science in China's water resources monitoring: 21. 655 current status and future prospects. International Journal of Sustainable Development & World 656 Ecology. 2022;29(3):277-90. Allen MJ, Edberg SC, Reasoner DJ. Heterotrophic plate count bacteria-what is their 657 22. significance in drinking water? International journal of food microbiology. 2004;92(3):265-74. 658 Garner E, Davis BC, Milligan E, Blair MF, Keenum I, Maile-Moskowitz A, et al. Next 659 23. generation sequencing approaches to evaluate water and wastewater quality. Water Research. 660 2021;194:116907. 661 662 24. Roeselers G, Coolen J, van der Wielen PW, Jaspers MC, Atsma A, de Graaf B, et al. 663 Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. 664 Environ Microbiol. 2015;17(7):2505-14. 665 25. van der Wielen PW, van der Kooij D. Effect of water composition, distance and season on the 666 adenosine triphosphate concentration in unchlorinated drinking water in the Netherlands. Water Res. 667 2010;44(17):4860-7. Ramirez KS, Knight CG, de Hollander M, Brearley FQ, Constantinides B, Cotton A, et al. 668 26. 669 Detecting macroecological patterns in bacterial communities across independent studies of global 670 soils. Nat Microbiol. 2018;3(2):189-96. 671 27. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal 672 catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457-63. 673 28. Willis JR, Gonzalez-Torres P, Pittis AA, Bejarano LA, Cozzuto L, Andreu-Somavilla N, et al. 674 Citizen science charts two major "stomatotypes" in the oral microbiome of adolescents and reveals 675 links with habits and drinking water composition. Microbiome. 2018;6(1):218. 676 29. NMC-Typhoon. 2106 In-fa National Meteorological Centre - Typhoon2021 [Available from: 677 http://typhoon.nmc.cn/web.html. 678 30. Chen S. Zhengzhou Rainstorm Once in a thousand years? The meteorological authority 679 responded: A year's rain fell in one day. Chinanews. 2021. China DIToTSCotPsRo. Investigation report on Zhengzhou 7.20 Rainstrom Disaster. In: 680 31. 681 China, MoEMotPsRo, editor. 2022. NMC. National Meteorological Center2021 [Available from: http://www.nmc.cn/. 682 32. 683 33. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. Towards Quantitative Microbiome 684 Community Profiling Using Internal Standards. Appl Environ Microbiol. 2019;85(5):e02634-18. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-685 34. 686 index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq 687 Illumina sequencing platform. Appl Environ Microbiol. 2013;79(17):5112-20. 688 Ley CJ, Proctor CR, Singh G, Ra K, Noh Y, Odimayomi T, et al. Drinking water 35. 689 microbiology in a water-efficient building: stagnation, seasonality, and physicochemical effects on

opportunistic pathogen and total bacteria proliferation. Environmental Science: Water Research & 690 691 Technology. 2020;6(10):2902-13. 692 36. Pinto AJ, Schroeder J, Lunn M, Sloan W, Raskin L. Spatial-temporal survey and occupancy-693 abundance modeling to predict bacterial community dynamics in the drinking water microbiome. 694 mBio. 2014;5(3):e01135-14. 695 2022 [cited January 28 2022]. Available from: https://www.wunderground.com/. 37. 696 38. EPA. Contaminant Candidate List (CCL) and Regulatory Determination 2023 [Available 697 from: https://www.epa.gov/ccl/contaminant-candidate-list-5-ccl-5. 698 EPA. Learn about Cyanobacteria and Cyanotoxins 2022 [Available from: 39. 699 https://www.epa.gov/cyanohabs/learn-about-cyanobacteria-and-cyanotoxins#what3. 700 WHO. Guidelines for Drinking-Water Quality (Fourth Edition) 2017. Available from: 40. 701 https://www.who.int/publications/i/item/9789241549950. 702 41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 703 EMBnet journal. 2011;17(1):10-2. 704 42. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-705 resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581-3. 706 Laws EA. Aquatic Pollution: An Introductory Text: Wiley; 2017. 43. 707 44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: 708 architecture and applications. BMC Bioinformatics. 2009;10(1):421. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation 709 45. 710 of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 711 2019;10(1):5029. 712 46. Vej GW. Bioinformatics explained: BLAST. CLC bio; 2007. 713 47. Upton GJ. Fisher's exact test. Journal of the Royal Statistical Society: Series A (Statistics in 714 Society). 1992;155(3):395-402. 715 48. Andermann T, Antonelli A, Barrett RL, Silvestro D. Estimating alpha, beta, and gamma 716 diversity through deep learning. Frontiers in plant science. 2022;13:839407. 717 McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and 49. 718 graphics of microbiome census data. PLoS One. 2013;8(4):e61217. 719 Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. Robust estimation of 50. 720 microbial diversity in theory and in practice. ISME J. 2013;7(6):1092-101. 721 51. Preheim SP, Perrotta AR, Friedman J, Smilie C, Brito I, Smith MB, et al. Computational 722 methods for high-throughput comparative analyses of natural microbial communities. Methods in 723 enzymology. 531: Elsevier; 2013. p. 353-70. 724 52. Oksanen J, Blanchet F, Kindt R. Vegan: commity ecology package. R package version 2.3-0. 725 2015. 726 53. Arbizu M. pairwiseAdonis: Pairwise multilevel comparison using adonis. R package version 727 0.4 github2020 [Available from: https://github.com/pmartinezarbizu/pairwiseAdonis. 728 54. Anderson MJ. A new method for non parametric multivariate analysis of variance. Austral 729 ecology. 2001;26(1):32-46. 730 Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 55. 731 2006;62(1):245-53. 732 Smouse PE, Long JC, Sokal RR. Multiple regression and correlation extensions of the Mantel 56. 733 test of matrix correspondence. Systematic zoology. 1986;35(4):627-32. 734 Han Z, An W, Yang M, Zhang Y. Assessing the impact of source water on tap water bacterial 57. 735 communities in 46 drinking water supply systems in China. Water Res. 2020;172:115469. 736 Ji P, Parks J, Edwards MA, Pruden A. Impact of Water Chemistry, Pipe Material and 58. 737 Stagnation on the Building Plumbing Microbiome. PLoS One. 2015;10(10):e0141087.

59. 738 Ling F, Whitaker R, LeChevallier MW, Liu WT. Drinking water microbiome assembly 739 induced by water stagnation. ISME J. 2018;12(6):1520-31. 740 Corning. Syringe Filters 2023 [Available from: https://ecatalog.corning.com/life-60. 741 sciences/b2c/US/en/Cell-Culture/Filtration/Syringe-Filters/Syringe-Filters/p/syringeFilters. 742 MilliporeSigma. Sterivex 2023 [Available from: 61. 743 https://www.sigmaaldrich.com/US/en/search/sterivex?focus=products&page=1&perpage=30&sort=r 744 elevance&term=sterivex&type=product. Hull NM, Holinger EP, Ross KA, Robertson CE, Harris JK, Stevens MJ, et al. Longitudinal 745 62. 746 and source-to-tap New Orleans, LA, USA drinking water microbiology. Environmental Science & 747 Technology. 2017;51(8):4220-9. 748 Pinto AJ, Xi C, Raskin L. Bacterial community structure in the drinking water microbiome is 63. 749 governed by filtration processes. Environmental science & technology. 2012;46(16):8851-9. 750 Ivone V-M, Conceição E, Olga C N, Célia M M. Bacterial diversity from the source to the tap: 64. 751 a comparative study based on 16S rRNA gene-DGGE and culture-dependent methods. FEMS 752 microbiology ecology. 2013;83(2):361-74. 753 65. Jia S, Shi P, Hu Q, Li B, Zhang T, Zhang XX. Bacterial Community Shift Drives Antibiotic 754 Resistance Promotion during Drinking Water Chlorination. Environ Sci Technol. 755 2015;49(20):12271-9. 756 66. Chiao TH, Clancy TM, Pinto A, Xi C, Raskin L. Differential resistance of drinking water 757 bacterial populations to monochloramine disinfection. Environ Sci Technol. 2014;48(7):4038-47. 758 Inkinen J, Siponen S, Jayaprakash B, Tiwari A, Hokajarvi AM, Pursiainen A, et al. Diverse 67. 759 and active archaea communities occur in non-disinfected drinking water systems-Less activity 760 revealed in disinfected and hot water systems. Water Res X. 2021;12:100101. 761 Dai Z, Sevillano-Rivera MC, Calus ST, Bautista-de Los Santos QM, Eren AM, van der 68. 762 Wielen P, et al. Disinfection exhibits systematic impacts on the drinking water microbiome. 763 Microbiome. 2020;8(1):42. 764 69. Franca L, Lopez-Lopez A, Rossello-Mora R, da Costa MS. Microbial diversity and dynamics 765 of a groundwater and a still bottled natural mineral water. Environ Microbiol. 2015;17(3):577-93. 766 70. Bautista-de los Santos QM, Schroeder JL, Sevillano-Rivera MC, Sungthong R, Ijaz UZ, 767 Sloan WT, et al. Emerging investigators series: microbial communities in full-scale drinking water 768 distribution systems-a meta-analysis. Environmental Science: Water Research & Technology. 769 2016;2(4):631-44. 770 71. Tang W, Liu Y, Li Q, Chen L, Li Q, Li P, et al. The Impact of Extreme Weather Events on 771 Bacterial Communities and Opportunistic Pathogens in a Drinking Water Treatment Plant. Water. 772 2021;14(1):54. 773 Zhang H, Xu L, Huang T, Yan M, Liu K, Miao Y, et al. Combined effects of seasonality and 72. 774 stagnation on tap water quality: Changes in chemical parameters, metabolic activity and co-existence 775 in bacterial community. Journal of Hazardous Materials. 2021;403:124018. 776 Zlatanović L, van der Hoek JP, Vreeburg J. An experimental study on the influence of water 73. 777 stagnation and temperature change on water quality in a full-scale domestic drinking water system. 778 Water research. 2017;123:761-72. 779 Zhang H, Xu L, Huang T, Liu X, Miao Y, Liu K, et al. Indoor heating triggers bacterial 74. 780 ecological links with tap water stagnation during winter: Novel insights into bacterial abundance, 781 community metabolic activity and interactions. Environmental Pollution. 2021;269:116094. 782 Ma L, Jiang XT, Guan L, Li B, Zhang T. Nationwide biogeography and health implications of 75. 783 bacterial communities in household drinking water. Water Res. 2022;215:118238. 784 Lim JY, Yoon JW, Hovde CJ. A brief overview of Escherichia coli O157: H7 and its plasmid 76. 785 O157. Journal of microbiology and biotechnology. 2010;20(1):5.

- 786 77. Ryan MP, Pembroke JT. Brevundimonas spp: Emerging global opportunistic pathogens.
- 787 Virulence. 2018;9(1):480-93.
- 788 78. Liu Y, Gilchrist A, Zhang J, Li XF. Detection of viable but nonculturable Escherichia coli
- 789 O157:H7 bacteria in drinking water and river water. Appl Environ Microbiol. 2008;74(5):1502-7.
- 790 79. Wolf-Baca M, Siedlecka A. Detection of pathogenic bacteria in hot tap water using the qPCR
   791 method: preliminary research. SN Applied Sciences. 2019;1(8):1-9.
- 792 80. Chopyk J, Moore RM, DiSpirito Z, Stromberg ZR, Lewis GL, Renter DG, et al. Presence of
- 793 pathogenic Escherichia coli is correlated with bacterial community diversity and composition on pre-794 harvest cattle hides. Microbiome. 2016;4(1):1-11.
- harvest cattle hides. Microbiome. 201
- 795

# 796 Non-standard abbreviations

- 797 RA: Relative abundance
- 798 CS: Citizen science sampling protocol and materials
- 799 NC: Nanchang, Jiangxi
- 800 XA: Xi'an, Shanxi
- 801 NJ: Nanjing, Jiangsu
- 802 HZ: Huizhou, Guangdong
- 803 JN: Jinan, Shandong
- 804 ZZ: Zhengzhou, Henan
- 805 CZ: Changzhou, Jiangsu
- 806 SH: Shanghai
- 807 ZB: Zibo, Shandong
- 808 MA: Macau
- 809 BJ: Beijing
- 810 TJ: Tianjin
- 811 CQ: Chongqing
- 812 LJ: Lijiang, Yunnan
- 813 LZ: Lanzhou, Gansu
- 814

# 815 Supporting Information Captions

- 816 S1 Table. Sample information.
- 817 S1 Text. Top 20 ASVs in the three temperature categories.
- 818 S1 Fig. Relative abundance of top 20 ASVs in household drinking water samples.
- 819 S2 Table. Fisher's exact test: Positive PCR signal vs. Negative PCR signal in terms of annual
- 820 precipitation (non-humid and humid).
- 821 S3 Table. The phylum-level taxonomic composition for each sample.
- 822 S2 Text. Beta-diversity (Normal samples, n = 20).
- 823 S2 Fig. Principal coordinates analysis (PCoA) of the normal samples (n = 20).
- 824 S4 Table. PERMANOVA and PERMDISP of beta diversity.
- 825 S5 Table. Mantel statistic (Spearman's rank correlation, permutations = 999).
- 826 S6 Table. The relative abundance of bacterial pathogens in each sample.
- 827 S7 Table. BLAST+ results of potential pathogens.
- 828 S8 Table. Pathogens in drinking water systems and their related diseases.
- 829 S3 Text. Sample storage test.
- 830 S4 Text. Representative pathogen species.
- 831 S5 Text. Potential environmental drivers: extreme precipitation and heat.
- 832 S3 Fig. Linear correlation (normal samples, Spearman): Outdoor temperature (higher than 17°C) ~
- 833 Total relative abundance of potential pathogens.



















Sample

