PLOS ONE

RESEARCH ARTICLE

Human microbiome privacy risks associated with summary statistics

Jae-Chang Cho *

Institute of Environmental Science and Department of Environmental Science, Hankuk University of Foreign Studies, Yong-In, Korea

* chojc@hufs.ac.kr

Citation: Cho J-C (2021) Human microbiome privacy risks associated with summary statistics. PLoS ONE 16(4): e0249528. https://doi.org/10.1371/journal.pone.0249528

Editor: Pedro H. Oliveira, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: September 10, 2020

Accepted: March 21, 2021

Published: April 2, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. Python code for the simulation is available at https://colab.research.google.com/drive/1dOZi8OSo5qHmF7JPyGAP1I_BQdBVSSiC?usp=sharing.

Funding: This work was supported in part by a 2021 HUFS internal research grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

Abstract

Recognizing that microbial community composition within the human microbiome is associated with the physiological state of the host has sparked a large number of human microbiome association studies (HMAS). With the increasing size of publicly available HMAS data, the privacy risk is also increasing because HMAS metadata could contain sensitive private information. I demonstrate that a simple test statistic based on the taxonomic profiles of an individual's microbiome along with summary statistics of HMAS data can reveal the membership of the individual's microbiome in an HMAS sample. In particular, species-level taxonomic data obtained from small-scale HMAS can be highly vulnerable to privacy risk. Minimal guidelines for HMAS data privacy are suggested, and an assessment of HMAS privacy risk using the simulation method proposed is recommended at the time of study design.

Introduction

Humans have coevolved with an immense number of diverse microorganisms that inhabit our bodies, collectively referred to as the human microbiome [1]. Together with the development of metagenomics, recognizing that microbial community composition within the microbiome is associated with the physiological state of the host has sparked a large number of human microbiome association studies (HMAS), which are also referred to as human metagenome-wide association studies (MWAS) [2] in analogy to genome-wide association studies (GWAS) [3]. As in the field of GWAS [4–7], the privacy risk is increasing with the increasing size of publicly available HMAS data.

The privacy threats of HMAS data are based on the fact that individual microbiomes harbor personally identifiable information in the form of microbial community composition. Several prominent studies demonstrated that individual identity can be revealed using the human microbiome. Fierer et al. [8] showed that an individual who touched an object (e.g., computer keyboard) could be identified by matching the compositional profile of the microbiome on the surface of the object to that of the individual's skin microbiome. While the authors' approach can be properly applied to forensic analyses, similar microbiome-based approaches can also be used to reveal an individual's location or intimate partner as shown by Lax et al. [9] and Kort

et al. [10]. Besides, Franzosa et al. [11] presented the possibility of a different type of privacy threat that uses information stored in databases; the authors showed that metagenomic codes, as sets of differentiable features of any given microbiome, can be used to identify individuals in the Human Microbiome Project dataset.

Although the development of biological as well as computational/statistical tools for analyzing individual microbiomes helps us to better understand human microbiomes, such tools can be viewed as double-edged swords. The more a tool has resolving power, the more privacy risks confront people as described above. Unfortunately, we might not be able to prevent privacy threats by attackers who utilize microbiome-based forensic techniques because the attackers only need to seize microbiomes from victims. On the other hand, privacy threats by data breaches can be prevented when the HMAS community develops privacy-preserving methods for HMAS data analysis as shown by Wagner et al. [12] and by storing HMAS data in access-controlled databases such as the dbGaP [13], which is currently used as a secure database of human-related genotypic and phenotypic data. However, there is another type of privacy threat that can be caused by the publication of HMAS data. Although raw or detailed data are not presented, summary statistics (e.g., mean frequencies of prokaryotic taxa in study microbiomes) are frequently provided in tables and figures contained in HMAS-based papers. Concerns over privacy breaches due to publishing summary statistics was first raised by Homer et al. [14] with respect to GWAS privacy, and this type of privacy attack was later termed 'attribute disclosure attacks under the summary statistic scenario' by Erlich and Narayanan [15]. To my knowledge, there have been no reports evaluating the privacy risk of HMAS summary statistics, which led me to perform a simple, foundational study in order to urge the HMAS community to be aware of privacy risks associated with HMAS dataset summary statistics.

In this paper, I demonstrate that the membership of an individual in the samples of an HMAS (e.g., case group or control group) can be revealed easily in the summary statistics of taxonomic compositions calculated from the microbiomes of the samples. Using a simple test statistic that was calculated from binary (presence/absence) taxonomic profiles, my simulation studies showed that publication of species-level taxonomic data obtained from small-scale HMAS can be highly vulnerable to privacy risk. This study asserts that the taxonomic profiles of the human microbiome should be treated as sensitive biometric information in that the HMAS metadata could contain the behavioral history of individuals in addition to medical conditions. I propose minimal guidelines for HMAS privacy and suggest that researchers use the simple simulation presented here to assess the privacy risk at the time of study design with the acknowledgment that a more advanced method could have greater resolving power for privacy breach.

Methods

Development of the test statistic

Suppose two samples R and C, each with size n_R and n_C, were drawn independently from population $\mathfrak{P}$ of human microbiomes. In HMAS, these samples may correspond to the sets of microbiomes of volunteers in the reference group ($\{\rightarrow y_i^R\}_{i=1}^{n_R}$) and case group ($\{\rightarrow y_i^C\}_{i=1}^{n_C}$), respectively; each microbiome is a vector of which the elements represent relative frequencies y_j^S of $j = 1,2,\cdots,t$, where t is the number of operational taxonomic units (OTUs) at different taxonomic levels from phylum to strain and $S \in \{R, C\}$. The summary statistics are vectors $\rightarrow r$ and $\rightarrow c$, of which the elements are the mean frequencies r_j and c_j of the OTUs in the pooled data obtained from R and C, respectively. Now consider the microbiome of an individual q, $\rightarrow y^q$, of which the elements are simple binary measures of presence/absence of OUT j (i.e., $y_j^q \in \{0, 1\}$), and suppose we want to determine whether $\rightarrow y^q$ is a member of R or C using

$\rightarrow r$ and $\rightarrow c$. First, calculate a distance d for OUT j using the absolute difference between y_j^q and c_j and the absolute difference between y_j^q and r_j as follows:

$$d_j = |y_j^q - r_j| - |y_j^q - c_j|$$

Assuming that OTUs are independent and invoking the central limit theorem for the large number of OTUs (t>50) examined in the HMAS, z-score of d_j across all OTUs will follow the standard normal distribution, $N(0, 1)$.

$$Z = \frac{\bar{d} - \mu_0}{\sqrt{V(\bar{d})}} = \frac{\bar{d} - \mu_0}{\sqrt{V(d)/t}} \approx \frac{\bar{d} - \mu_0}{s/\sqrt{t}} \sim N(0, 1)$$

Since the $t = \max(t_R, t_C, t_y)$ is large, the variance of d can be estimated reliably by the sample variance s^2. The test statistic Z was inspired by Homer et al. [14] and Braun et al. [16]. The authors used a similar test statistic calculated from the single nucleotide polymorphism (SNP) genotyping data to identify an individual's genotype in GWAS samples. I modified the original test statistic in order to use the binary taxonomic data in HMAS. Because an individual microbiome randomly drawn from population $\mathfrak{P}$ should be equally distant from R and C, μ_0 was presumably expected to be zero, i.e., $N(0, 1)$ was a putative null distribution. Thus, under the null hypothesis $H_0 : Z = 0 (\rightarrow y^q$ is a random draw from $\mathfrak{P})$, the alternative hypothesis $H_R : Z < 0 (\rightarrow y^q$ is a member of $R)$ or $H_C : Z > 0$ $(\rightarrow y^q$ is a member of $C)$ can be tested with an appropriate significance level α. For example, Z>1.65 rejects H_0 in favor of H_C at $\alpha = 0.05$ (one-tailed test).

Distribution simulations

I examined the feasibility of the test statistic Z in identifying the presence of an individual's microbiome in an HMAS sample with simulated datasets. Suppose that the OTUs correspond to species-level affiliations of microorganisms. Then, the presence of species j in $\mathfrak{P}$ is a Bernoulli random variable with parameter p_j. To model the probability distribution of p_j $(\rightarrow p)$, I used four different $Beta(\pi_1, \pi_2)$ distributions. For $\mathfrak{P}$ with a small number of high-frequency species, $\pi_1 = 0.1$ and $\pi_2 = 1.0$ were assumed, and for $\mathfrak{P}$ with a large number of high-frequency species, $\pi_1 = 1.0$ and $\pi_2 = 0.1$ were assumed. For $\mathfrak{P}$ with a high number of high-frequency species as well as a high number of low-frequency species, $\pi_1 = 0.1$ and $\pi_2 = 0.1$ were assumed, which might be more realistic than the above two distributions in that the high-frequency species may correspond to constitutional or autochthonous prokaryotic populations and the low-frequency species may correspond to opportunistic or heterochthonous prokaryotic populations across individual microbiomes. In addition, a $Beta(1, 1)$ (uniform) distribution was also used, which represents our ignorance of the distribution of p_j according to the principle of indifference.

Because the individual microbiome in R or C is a set of t random draws of $y_j \sim Bernoulli(p_j)$, i.e., $\rightarrow y \sim Bernoulli(\rightarrow p)$, the summary statistics for samples R and C were simulated using $\rightarrow r \sim Binomial(n_R, \rightarrow p)/n_R$ and $\rightarrow c \sim Binomial(n_C, \rightarrow p)/n_C$, respectively. To construct density curves of true positives for samples R and C, random draws of $\rightarrow y^{R+}$ and y_j^{C+} from samples R and C were used to calculate the test statistics Z^{R+} and Z^{C+} $(n_Z^{R+} = n_Z^{C+} = 100)$, respectively. Density curves were estimated using the Gaussian kernel density estimator. To construct density curves of the true null distribution, random draws of $\rightarrow y^P$ from $\mathfrak{P}$ were used to calculate test statistics Z^P $(n_Z^P = 100)$. Python code for the simulation is available at https://colab.research.google.com/drive/1dOZi8OSo5qHmF7JPyGAP1I_BQdBVSSiC?usp=sharing.

Results and discussion

Overview of simulation results

A simulation study was started with the number of OTUs t =2,000, which roughly reflects the number of species (including uncultivated candidate species) in the human gut microbiome [17], and with $n_R = n_C = 10$, which could correspond to small-scale HWAS, under the assumption that p_j follows uniform distribution (Fig 1). The null distribution estimated using Z^P was very close to $N(0, 1)$ and crisply separated from the distributions of true positives (Z^{R+} and Z^{C+}), indicating the feasibility of the microbiome-based identification. The distributions of true positives moved toward the null distribution with increasing n or with decreasing t but moved away from the null distribution with decreasing n or with increasing t. Similar distribution patterns were observed for all the underlying distributions of p_j assumed (S1–S3 Figs).

For numerical interpretation of the simulation results, I focused on the probability of type II error at $\alpha = 0.05$ ($\beta^R = P(Z^{R+} > z_\alpha | H_R)$ or ($\beta^C = P(Z^{C+} < z_\alpha | H_C)$; the probability that the test statistic is not in the H_0 rejection range, given that the alternative hypothesis is true). The power of the test ($1-\beta$) is important to a privacy attacker because it would be especially difficult for the attacker with a high β to determine whether the individual under investigation belongs to a particular HMAS sample due to the high false negative rate. α level critical values were obtained from percentiles of Z^P distribution because the true null distribution might diverge from $N(0, 1)$. As expected in the density curves, β was far less than 0.01 for the experiments with a large t or small n (S1–S4 Tables).

To formulate the guidelines for human microbiome privacy, the initial simulation study was expanded for virtual HMAS samples with $n_R = n_C = 10, 20, \ldots, 100, \ldots, 1000$ and with $t = 20, 30, \ldots, 100, \ldots, 1000, \ldots, 20000$. The method was in general slightly more powerful for p_j under the uniform distribution (Fig 2) than for p_j under other beta distributions (S4 and S5 Figs). In the resulting contour plots, the yellowish area represents higher β (i.e., lower test power); thus, the HMAS samples located in the dark blue area are considered to be vulnerable to privacy risk. The power of the test decreased notably with increasing HMAS sample size (n) or with a decreasing number of OTUs (t) from which the HMAS summary statistics are calculated, and it is possible to notice a borderline where β decreases considerably, which could help in assessing the privacy risk of the HMAS data.

Effect of sample size

Samples of a large size approximate the population $\mathfrak{P}$ because $\lim_{n_{R\to\infty}} r_j = p_j$ or $\lim_{n_{C\to\infty}} c_j = p_j$ (hence, $\lim_{n_R, n_{C,\to\infty}} d_j = 0$), and the distribution of Z^{R+} or Z^{C+} would overlap with the null distribution as shown in the selected density curves, indicating that a large sample size will make the classifying method ineffective. Contrarily, for the samples of a small size, $\bar{d}$ significantly deviates from μ_0 even if the difference between r_j and p_j or between c_j and p_j is very small. Note that the sample size does not increase with the number of technical replicates used in HMAS data, since the sample points in technical replicates are not independent.

For an unequal sample size (e.g., n_R = 1,000 and $n_C = 10$), the samples with a larger size would approximate the null distribution, but the test power for identifying a microbiome in a sample with a smaller size would not be affected (S6 Fig). Considering that taxonomic profiles of individual microbiomes in the case sample could be more homogeneous than those in the control sample, the effective size of the case sample might be smaller than the census size of the case sample. This would result in much higher statistical power for testing whether an individual is a member of the case sample, which is a primary interest of the attackers.

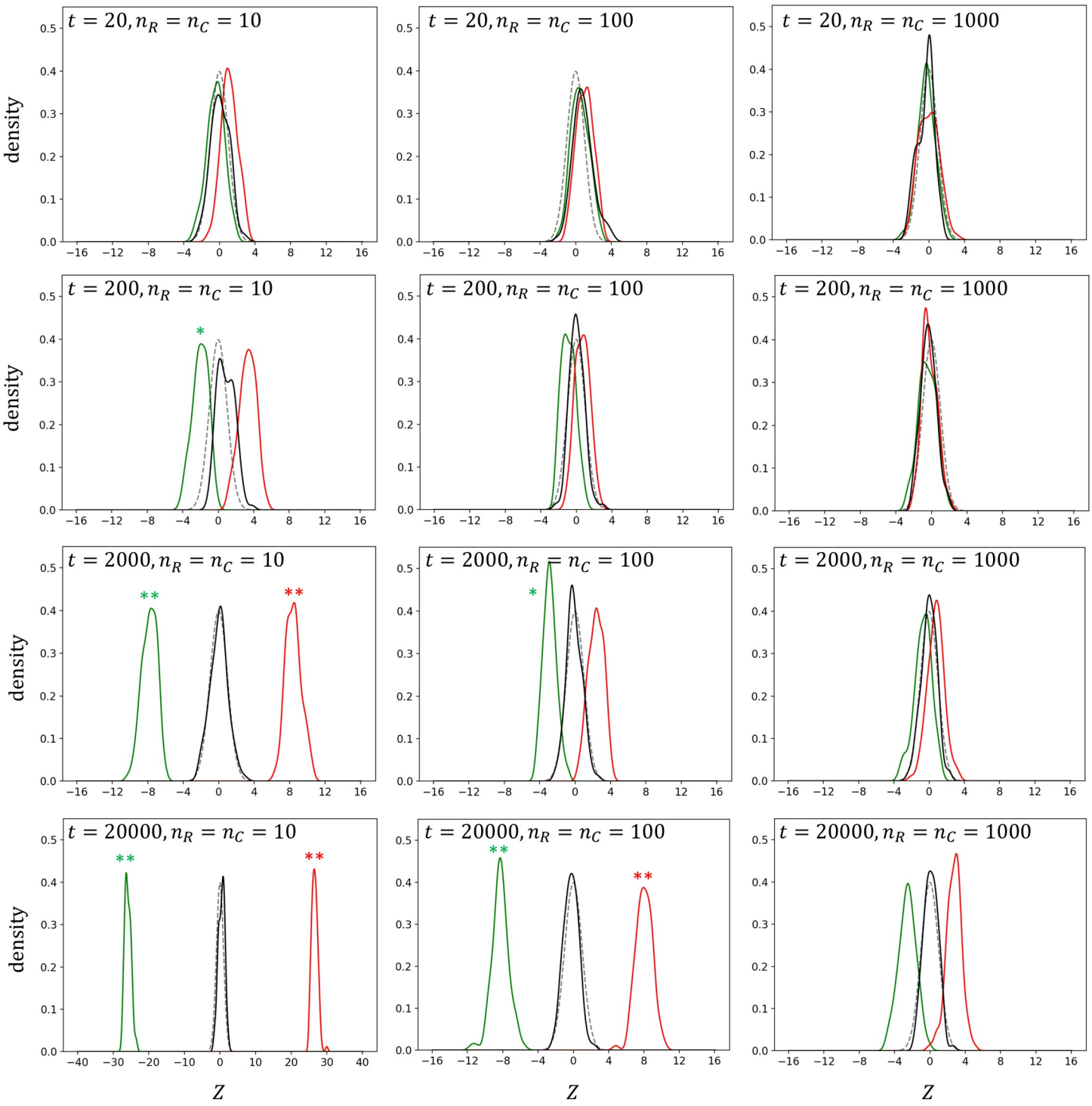

Fig 1. Distributions of the test statistic *Z* under the assumption that the population OTU frequencies follow a uniform (Beta(1, 1)) distribution. Density curves for true positives of samples *R* (Z^{R+}) and *C* (Z^{C+}) are denoted by green and red lines, respectively. Density curves of simulated null distribution and standard normal distribution are denoted by black and gray lines, respectively. Single and double asterisks represent type II error probabilities $\beta<0.05$ and $\beta<0.01$, respectively.

https://doi.org/10.1371/journal.pone.0249528.g001

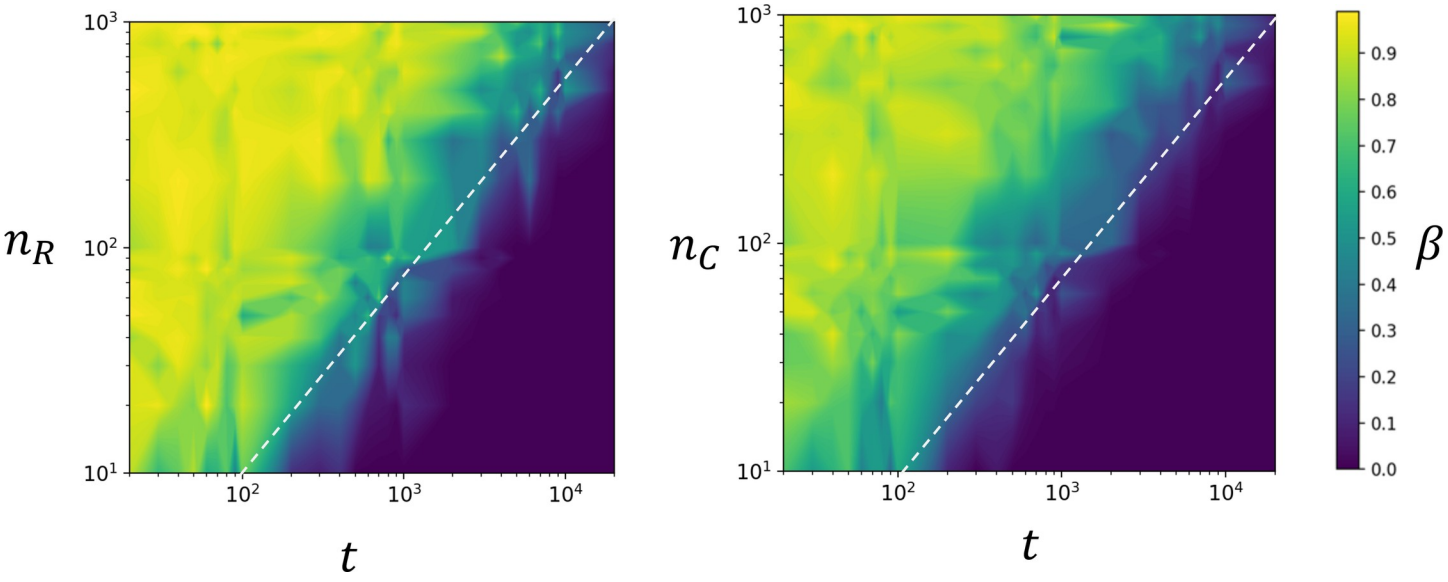

Fig 2. Contour plot representations of the type II error probabilities (β) for true positives of samples *R* and *C* under the assumption that the population OTU frequencies follow a uniform (*Beta*(1, 1)) distribution. Sample size and the number of OTUs are log-scaled. Dotted line denotes suggested minimal guidelines for HMAS privacy.

https://doi.org/10.1371/journal.pone.0249528.g002

Effect of the number of OTUs

With the sample size fixed, an increased number of OTUs would increase the test power. This situation can happen when the HMAS data contains taxonomic profiles with a resolution finer than species level. Under the arbitrary speculation that each species is comprised of ca. 10 strains (t = 20,000), the method was very powerful ($\beta \ll 0.01$) even for a moderate sample size. This is because the denominator of the test statistic shrinks by increasing t compared to the sample mean $\bar{d}$, which results in a large Z^{R+} or Z^{C+}. In the same vein, a reduced number (e.g., t = 20) of OTUs (roughly corresponding to near phylum-level) would decrease the power of the test.

Effect of correlation among OTUs

The occurrence of many OTUs in the human microbiome could be correlated, since the microbiome itself is an ecological community [18]. If the correlation among OTUs is significant, the violation of the assumption that OTUs are independent could change the test power. I analytically evaluated the effect of the OUT correlation on test power. If OTUs are not independent, the variance of $\bar{d}$ includes covariance (*Cov*) terms as follows:

$$V(\bar{d}) = \frac{1}{t}V(d) + \frac{2}{t^2}\sum_{j}^{t}\sum_{j'}^{t} Cov(d_j, d_{j'}),$$

where $j<j'$. By letting $\tilde{\rho}$ be the average correlation among d_j such that $V(\bar{d}) \geq 0$, the above equation can be written as follows:

$$V(\bar{d}) = \frac{V(d)}{t/(1+\tilde{\rho}t-\tilde{\rho})},$$

where $1 + \tilde{\rho}t - \tilde{\rho} \geq 0$. Thus, $V(\bar{d})$ increases if $\tilde{\rho}$ is positive

$(0 < \tilde{\rho} \leq 1 \Longrightarrow V(d)/t < V(\bar{d}) \leq V(d))$ or decreases if $\tilde{\rho}$ is negative $(-1/(t-1) \leq \tilde{\rho} < 0 \Longrightarrow 0 \leq V(\bar{d}) \leq V(d)/t)$. One can imagine that the changes in $V(\bar{d})$ result in changes in the distribution of test statistic Z. But the effect of $\tilde{\rho}$ on test power is not easy to see with changes in $V(\bar{d})$. Rather, we can view the denominator term of $V(\bar{d})$ as an effective number of independent OUTs (t_e) as follows:

$$t_e = \frac{t}{1 + \tilde{\rho}t - \tilde{\rho}}$$

Note that $t_e = t$ if $\tilde{\rho} = 0$, $1 \leq t_e < t$ if $0 < \tilde{\rho} \leq 1$, and $t < t_e$ if $-1/(t-1) \leq \tilde{\rho} < 0$ ($t_e \rightarrow \infty$ as $\tilde{\rho} \rightarrow -1/(t-1)$). For example, if three of 10 OTUs show a strong positive correlation, two of these OTUs cannot contribute to the number of OTUs as equally as other independent OTUs. Thus, the effect of a positive $\tilde{\rho}$ could be similar to the effect of a decreased number of OTUs, resulting in decreased test power as in the case of linkage disequilibrium among SNPs [16]. However, $\tilde{\rho}$ can also be negative if negative interactions among OTUs are more prevalent than positive interactions among OTUs, which could subsequently increase test power. Even very small negative correlations (e.g., $\tilde{\rho} = -0.001$) among a moderate number of OTUs ($t = 1000$) would increase t_e to 10^6, while very small positive correlations (e.g., $\tilde{\rho} = +0.001$) decrease t_e to 500. The negative correlation overwhelmingly affects t_e because t_e is a reciprocal function of $\tilde{\rho}$ when t is fixed (S7 Fig). Nonetheless, because $\tilde{\rho}$ cannot be measured from summary statistics unless provided or assumed, the effect of $\tilde{\rho}$ should be investigated more comprehensively in the future.

Additional considerations

I used the binary vector as a query microbiome because it was considered that the binary data is less prone to experimental variations. If a frequency vector is used, the $\bar{d}$ could increase because small differences between $|y_j^q - r_j|$ and $|y_j^q - c_j|$ can accumulate across a large number of OTUs. However, it would not always result in a larger Z^{R+} or Z^{C+} because the increased sample variance could counteract the increase in $\bar{d}$. Thus, the use of a frequency vector might not guarantee more test power, rather it could make $\bar{d}$ more prone to errors in OTU frequencies.

It was assumed that R and C were samples drawn independently from the population of human microbiomes ($\mathfrak{P}$). Violation of this assumption could shift the location of the null distribution (μ_0). Suppose that populations underlying samples R and C (denoted by $\mathcal{R}$ and $\mathcal{C}$, respectively) are all different from the population underlying $\rightarrow y^q(\mathfrak{P})$. If the systemic difference between $\mathcal{R}$ and $\mathcal{C}$ is significant, the difference between summary statistics can deviate from zero, i.e., $E(\Delta) = E(r-c) \neq 0$. Let d_j^* be the distance metric that includes Δ term as follows:

$$d_j^* = |y_j^q - \Delta_j - c_j| - |y_j^q - c_j| = \begin{cases} \Delta_j & \text{for } y_j^q = 0 \\ -\Delta_j & \text{for } y_j^q = 1 \end{cases}$$

Then, according Braun et al. [16] and using $P(y_j^q = 0|p_j) = 1 - p_j$ and $P(y_j^q = 1|p_j) = p_j$, the expected value of the distance metric under the null hypothesis is as follows:

$$\mu_0^* = E(\Delta - 2\Delta p) = E(\Delta) - 2E(\Delta)E(p)$$

Because $-1 \leq E(\Delta) \leq 1$ and $0 < E(p) \leq 1$, μ_0^* ranges from -1 to 1 which is quite small compared to the distance between the distributions of Z^P and Z^{R+} or between the distributions of Z^P and Z^{C+} in the cases where the test power is very high (e.g., $t \geq 2000$ and $n = 10$ or $t \geq 20000$

and $n = 100$) (Fig 1). Moreover, as $E(p)$ deviates from its two extreme values (0 or 1) which are highly unrealistic in microbiome composition, the $2E(\Delta)E(p)$ term counteracts the deviation of μ_0^* from zero (i.e., $\mu_0^* \to 0$ as $E(p) \to 0.5$). Thus, I considered that the violation of the assumption that underlying populations are all different would not alter the main results of this study. In fact, if R (reference/control group) and C (case/treatment group) are assumed to be samples drawn from $\mathfrak{P}$ and $\mathcal{C}$, respectively, the test statistic would become very similar to that used for a two-tailed z-test (or t-test) which can be used to test the null hypothesis that y_j^q is not a member of C.

Although it might be virtually impossible for the HMAS community to develop an almighty shield that protects data against all possible privacy breaching methods, we do not need to overreact to HMAS privacy concerns since the taxonomic profile of the individual's microbiome could not be a permanent identifier, while a subset of an individual's microbiome may endure for most of the lifetime. Nonetheless, the HMAS community should have guidelines for reducing the privacy risk, which could be kept minimal in order not to obstruct our understanding of the human microbiome.

Concluding remarks

This study showed the possibilities of privacy breaches using the summary statistics drawn from HMAS data. The key finding was that the publication of species composition data obtained from small-scale HMAS (e.g., $t \approx 1{,}000$ and $n_R = n_C = 10$) easily exposes the privacy of the victims. The HMAS samples with moderate size ($n_R = n_C \approx 100$) and $t \approx 1000$ were on the vicinity of the borderline. I suggest these figures as a basis for HMAS data release policy while acknowledging a sophisticated test statistic that employs better distance metric along with the violation of underlying assumptions could improve the power of the privacy breaching methods. I propose minimal guidelines for HMAS data release as follows: i) increase the sample size as much as manageable, ii) do not publish species composition data even in the form of summary statistics if the sample size is less than 10, iii) avoid publishing subspecies- or strain-level data unless the sample size is far larger than 100. I also suggest that HMAS researchers evaluate the privacy risk at the time of study design using the simulation method presented in this study. Because the test statistic used was relatively robust against the variations in the models for the distribution of p_j, HMAS researchers can simply use uniform distribution and their intended t, n_R, and n_C to estimate the 'minimal' power of the privacy-breaching method.

This study focused only on attribute disclosure attacks under the summary statistic scenario [15]. However, HMAS data privacy concerns are not limited to summary statistics; privacy breaches can occur in various manners at any stage of HMAS data management as described by Wagner et al. [12]. Thus, it is timely that the HMAS community begins comprehensive discussions on HMAS data privacy risks and developing privacy-preserving algorithms for data storage and release.

Supporting information

S1 Fig. Distributions of test statistic Z under the assumption that population OTU frequencies follow a *Beta*(0.1, 1) distribution. Density curves for true positives of samples R (Z^{R+}) and C (Z^{C+}) are denoted by green and red lines, respectively. Density curves of simulated null distribution and standard normal distribution are denoted by black and gray lines, respectively. Single and double asterisks represent type II error probabilities $\beta < 0.05$ and $\beta < 0.01$, respectively.
(PDF)

S2 Fig. Distributions of the test statistic *Z* under the assumption that population OTU frequencies follow a *Beta*(1, 0.1) distribution. Density curves for true positives of samples *R* (Z^{R+}) and *C* (Z^{C+}) are denoted by green and red lines, respectively. Density curves of simulated null distribution and standard normal distribution are denoted by black and gray lines, respectively. Single and double asterisks represent type II error probabilities $\beta<0.05$ and $\beta<0.01$, respectively.
(PDF)

S3 Fig. Distributions of the test statistic *Z* under the assumption that population OTU frequencies follow a *Beta*(0.1, 0.1) distribution. Density curves for true positives of samples *R* (Z^{R+}) and *C* (Z^{C+} are denoted by green and red lines, respectively. Density curves of simulated null distribution and standard normal distribution are denoted by black and gray lines, respectively. Single and double asterisks represent type II error probabilities $\beta<0.05$ and $\beta<0.01$, respectively.
(PDF)

S4 Fig. Contour plot representations of the type II error probabilities (*β*) for true positives of sample *R* under the assumptions that population OTU frequencies follow *Beta*(1, 1), *Beta*(0.1, 1), *Beta*(1, 0.1) and *Beta*(0.1, 0.1) distributions. Sample size and the number of OTUs are log-scaled. Dotted line denotes suggested minimal guidelines for HMAS privacy.
(PDF)

S5 Fig. Contour plot representations of the type II error probabilities (*β*) for true positives of sample *C* under the assumptions that population OTU frequencies follow *Beta*(1, 1), *Beta*(0.1, 1), *Beta*(1, 0.1) and *Beta*(0.1, 0.1) distributions. Sample size and the number of OTUs are log-scaled. Dotted line denotes suggested minimal guidelines for HMAS privacy.
(PDF)

S6 Fig. Distributions of the test statistic *Z* with unequal sample sizes under the assumption that population OTU frequencies follow a *Beta*(1, 1) distribution. Density curves for true positives of samples *R* (Z^{R+}) and *C* (Z^{C+}) are denoted by green and red lines, respectively. Density curves of simulated null distribution and standard normal distribution are denoted by black and gray lines, respectively. Single and double asterisks represent type II error probabilities $\beta<0.05$ and $\beta<0.01$, respectively.
(PDF)

S7 Fig. Relationship between the effective number of OTUs and the average correlation among OTUs.
(PDF)

S1 Table. Summary statistics of simulation results obtained under the assumption that the population OTU frequencies follow a *Beta*(1, 1) distribution. Type II error probabilities less than 0.05 are in bold.
(PDF)

S2 Table. Summary statistics of simulation results obtained under the assumption that the population OTU frequencies follow a *Beta*(0.1, 1) distribution. Type II error probabilities less than 0.05 are in bold.
(PDF)

S3 Table. Summary statistics of simulation results obtained under the assumption that the population OTU frequencies follow a *Beta*(1, 0.1) distribution. Type II error probabilities

less than 0.05 are in bold.
(PDF)

S4 Table. Summary statistics of simulation results obtained under the assumption that the population OTU frequencies follow a *Beta*(1, 1) distribution. Type II error probabilities less than 0.05 are in bold.
(PDF)

Acknowledgments

I thank the anonymous reviewers for their helpful comments and suggestions.

Author Contributions

Conceptualization: Jae-Chang Cho.

Data curation: Jae-Chang Cho.

Formal analysis: Jae-Chang Cho.

Investigation: Jae-Chang Cho.

Methodology: Jae-Chang Cho.

Resources: Jae-Chang Cho.

Software: Jae-Chang Cho.

Supervision: Jae-Chang Cho.

Validation: Jae-Chang Cho.

Visualization: Jae-Chang Cho.

Writing – original draft: Jae-Chang Cho.

Writing – review & editing: Jae-Chang Cho.

References

1. Turnbaugh P, Ley R, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project. Nature. 2007; 449: 804–810. https://doi.org/10.1038/nature06244 PMID: 17943116
2. Wang J, Jian H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol. 2016; 14: 508–522. https://doi.org/10.1038/nrmicro.2016.83 PMID: 27396567
3. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009; 360: 1759–1768. https://doi.org/10.1056/NEJMra0808700 PMID: 19369657
4. Lowrance W, Collins F. Identifiability in genomic research. Science. 2007; 317: 600–602. https://doi.org/10.1126/science.1147699 PMID: 17673640
5. Lunshof J, Chadwick R, Vorhaus D, Church G. From genetic privacy to open consent. Nat Rev Genet. 2008; 9: 406–411. https://doi.org/10.1038/nrg2360 PMID: 18379574
6. Greenbaum D, Sboner A, Mu X, Gerstein M. Genomics and privacy: Implications of the new reality of closed data for the field. PLoS Comput Biol. 2011; 7: e1002278. https://doi.org/10.1371/journal.pcbi.1002278 PMID: 22144881
7. Callaway E. Microbiome privacy risk: The DNA of microorganisms living on a person's body could identify that individual. Nature. 2015; 521: 136. https://doi.org/10.1038/521136a PMID: 25971486
8. Fierer N, Lauber C, Zhou N, McDonald D, Costello E, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci USA. 2010; 107: 6477–6481. https://doi.org/10.1073/pnas.1000162107 PMID: 20231444
9. Lax S, Smith D, Hampton-Marcell J, Owens S, Handley K, Scott N, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science. 2014; 345: 1048–1052. https://doi.org/10.1126/science.1254529 PMID: 25170151

10. Kort R, Caspers M, van de Graaf A, van Egmond W, Keijser B, Roeselers G. Microbiome. 2014; 2: 41. https://doi.org/10.1186/2049-2618-2-41 PMID: 25408893

11. Franzosa E, Huang K, Meadow J, Gevers D, Lemon K, Bohannan B, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci USA. 2015; 112: E2930–E2938. https://doi.org/10.1073/pnas.1423854112 PMID: 25964341

12. Wagner J, Paulson J, Wang X, Bhattacharjee B, Bravo H. Privacy-preserving microbiome analysis using secure computation Bioinformatics. 2016; 32: 1873–1879. https://doi.org/10.1093/bioinformatics/btw073 PMID: 26873931

13. Mailman M, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39: 1181–1186. https://doi.org/10.1038/ng1007-1181 PMID: 17898773

14. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008; 4: e1000167. https://doi.org/10.1371/journal.pgen.1000167 PMID: 18769715

15. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genetics. 2014; 15: 409–421. https://doi.org/10.1038/nrg3723 PMID: 24805122

16. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet. 2009; 5: e1000668. https://doi.org/10.1371/journal.pgen.1000668 PMID: 19798441

17. Almeida A, Mitchell A, Boland M, Forster S, Gloor G, el al. A new genomic blueprint of the human gut microbiota. Nature. 2019; 568: 499–504. https://doi.org/10.1038/s41586-019-0965-1 PMID: 30745586

18. Gilbert J, Lynch S. Community ecology as a framework for human microbiome research. Nat Med. 2019; 25: 884–889. https://doi.org/10.1038/s41591-019-0464-9 PMID: 31133693