Biometrics 64, 1–7 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Supporting Information for A Semiparametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data by Liu, J. et al.

Liu, J.[1,9] , Zhang, X.[1,*], Chen, T.[2], Wu, T.[1,9], Lin, T.[1], Jiang, L.[1,12],

and

Lang, S.[3], Liu, L.[1], Natarajan, L.[1], Tu, J.X.[4], Kosciolek, T.[5,6], Morton, J.[7]

and

Nguyen, T.T[8,9], Schnabl, B.[3], Knight, R.[5,10,11,12], Feng, C.[13], Zhong, Y.[14], and Tu, X.M.[1,9]

[1]Department of Family Medicine and Public Health, UC San Diego, San Diego, California, U.S.A.

[2]Department of Mathematics, University of Toledo, Toledo, Ohio, U.S.A.

[3]Department of Medicine, UC San Diego, San Diego, California, U.S.A.

[4]Physical Medicine and Rehabilitation, University of Virginia Health System, Charlottesville, Virginia, U.S.A.

[5]Department of Pediatrics, UC San Diego, San Diego, California, U.S.A.

[6]Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

[7]Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, New York, U.S.A.

[8]Department of Psychiatry, UC San Diego, San Diego, California, U.S.A.

[9]Stein Institute for Research on Aging, UC San Diego, San Diego, California, U.S.A.

[10]Department of Computer Science and Engineering, UC San Diego, San Diego, California, U.S.A.

[11]Department of Bioengineering, UC San Diego, San Diego, California, U.S.A.

[12]Center for Microbiome Innovation, UC San Diego, San Diego, California, U.S.A.

[13]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.

[14]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

email:* **xizhang@health.ucsd.edu

This paper has been submitted for consideration for publication in *Biometrics*

Supplementary Materials

0.1 *Proof of Theorem 1.*

Without loss of generality, consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n$. A Taylor's series expansion gives

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\left(-\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_n(\boldsymbol{\theta})\right)^{-\top}\sqrt{n}\mathbf{U}_n(\boldsymbol{\theta})+\mathbf{o}_p\left(1\right). \quad (1)$$

From the theory of multivariate U-statistics that (Kowalski and Tu, 2008),

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_n(\boldsymbol{\theta})=\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\frac{\partial}{\partial\boldsymbol{\theta}}(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}h_{\mathbf{i}}(\boldsymbol{\theta}))\to^p E\left(\frac{\partial}{\partial\boldsymbol{\theta}}h_{\mathbf{i}}(\boldsymbol{\theta})\left(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}\right)^{\top}\right)$$
$$=-E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^{\top}\right)=-B,$$

where $\mathbf{o}_p\left(1\right)$ denotes the stochastic version of $\mathbf{o}\left(1\right)$. Since $\mathbf{U}_{n,\mathbf{i}}$ is a U-statistic-like quantity, it again follows from the theory of multivariate U-statistics that:

$$\sqrt{n}\mathbf{U}_n=\sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\mathbf{U}_{n,\mathbf{i}}=\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}E\left(\mathbf{U}_{n,\mathbf{i}}\mid\mathbf{y}_{i_1},\mathbf{x}_{i_1},\mathbf{z}_{i_1}\right)+\mathbf{o}_p\left(1\right) \quad (2)$$
$$=\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p\left(1\right)\to_d N\left(\mathbf{0},\Sigma_U\right),$$

By combining (1) and (2), we have:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\left(-\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_n\right)^{-\top}\sqrt{n}\mathbf{U}_n+\mathbf{o}_p\left(1\right)=B^{-1}\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p\left(1\right)\to_d N\left(\mathbf{0},\Sigma_{\boldsymbol{\theta}}\right).$$

0.2 *Proof of Theorem 2.*

Again consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n$. By the theory of multivariate U-statistics that (Kowalski and Tu, 2008):

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}_n\left(\boldsymbol{\theta}\right)=\begin{pmatrix}\frac{\partial}{\partial\boldsymbol{\theta}_{(1)}}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) & \frac{\partial}{\partial\boldsymbol{\theta}_{(1)}}\mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right)\\ \frac{\partial}{\partial\boldsymbol{\theta}_{(2)}}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) & \frac{\partial}{\partial\boldsymbol{\theta}_{(2)}}\mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right)\end{pmatrix}\to_p B=E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^{\top}\right)=\begin{pmatrix}B_{11} & B_{12}\\ B_{12}^{\top} & B_{22}\end{pmatrix}. \quad (3)$$

It follows from a Taylor's series expansion and (3) that

$$\begin{aligned}\mathbf{0} = \mathbf{U}_{n(1)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) &= \mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) + \left(\frac{\partial^\top}{\partial \boldsymbol{\theta}_{(1)}}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right)\right)\left(\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\ &= \mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) + B_{11}\left(\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

Thus,

$$\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)} = -B_{11}^{-1}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \tag{4}$$

Similarly, since $B_{12}^\top = B_{21}$, we have:

$$\begin{aligned}\mathbf{U}_{n(2)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) &= \mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right) + \left(\frac{\partial^\top}{\partial \boldsymbol{\theta}_{(1)}}\mathbf{U}_{n(2)}\right)\left(\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \qquad (5)\\ &= \mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right) + B_{12}^\top\left(\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\ &= \mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right) + B_{21}\left(\widetilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

It follows from (4) and (5) that

$$\begin{aligned}\mathbf{U}_{n(2)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) &= \mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right) + B_{21}\left[-B_{11}^{-\top}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right] + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\ &= \mathbf{U}_{n(2)}\left(\boldsymbol{\theta}\right) - \left[B_{21}B_{11}^{-\top}\mathbf{U}_{n(1)}\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right] + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\ &= \begin{pmatrix} -B_{21}B_{11}^{-1} & \mathbf{I}_q \end{pmatrix}\mathbf{U}_n\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\ &= G\mathbf{U}_n\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

By the central limit theorem,

$$\sqrt{n}\mathbf{U}_{n(2)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) = \sqrt{n}G\mathbf{U}_n\left(\boldsymbol{\theta}\right) + \mathbf{o}_p\left(1\right) \to_d N\left(\mathbf{0}, \Sigma_{(2)} = G\Sigma_U G^\top\right). \tag{6}$$

The asymptotic normality of $\mathbf{U}_{n(2)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right)$ implies that the score statistic $S_n\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right)$ has the asymptotic χ_q^2 distribution.

0.3 *PERMANOVA*

If $\mathbf{x}_i$ consists of only one categorical variable for groups, PERMANOVA can be used to compare Beta-diversity across different groups. Consider a total of K groups for this categorical variable, PERMANOVA uses the pseudo-F statistic for inference about overall group differences in Beta-diversity:

$$\text{pseudo-}F = \frac{tr\left(HGH\right)/\left(p-1\right)}{tr[(\mathbf{I}_n - H)G\left(\mathbf{I}_n - H\right)]/(n-p)}, \tag{7}$$

$$G = \left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right)A\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right), \quad A = \left(-\frac{1}{2}d_{\mathbf{i}}^2\right),$$

where $tr\left(\cdot\right)$ denotes the trace of a matrix, X is the design matrix that contains the group information, p is the length of $\mathbf{x}_i$, $H = X(X^\top X)^{-1}X^\top$ is the projection of the design matrix X, G is the Gower's centered matrix obtained from the distance matrix $D = (d_{\mathbf{i}})$, $\mathbf{1}_n$ denotes a $n \times 1$ column vector of 1's, and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. For example, if $K = 2$, and $x_i = 1$ if the ith subject is from diseased group and $x_i = 0$ otherwise, then $X = \left(\mathbf{1}_n, \mathbf{x}^\top\right)$, where $\mathbf{x}^\top = \left(x_1, x_2, \ldots, x_n\right)^\top$.

0.4 *Details of Data Generating Procedure with eCDF and Copula*

For notational clarity, we use upper-case to denote random variables and lower-case to denote their values. Consider a random variable X and let $F\left(x\right)$ denote the cumulative distribution function (CDF) of X. Then the probability integral transformation of X, $U = F\left(X\right)$, follows $U\left(0,1\right)$, where $U\left(0,1\right)$ is a uniform between 0 and 1 (Kowalski and Tu, 2008). Thus, if $F\left(x\right)$ is known, we can simulate X from $X = F^{-1}\left(U\right)$, where $F^{-1}\left(u\right)$ is the inverse of $F\left(x\right)$ defined by $F^{-1}\left(u\right) = \inf\left\{x \mid F\left(x\right) \geqslant u\right\}$, $0 < u < 1$. If $F\left(x\right)$ is unknown, we can instead use the empirical CDF (eCDF) of the observed X, i.e., $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leqslant x\right)$, where $I\left(A\right)$ is an indicator with value 1 if A is true and 0 otherwise.

For a $p \times 1$ random vector $\mathbf{X} = \left(X_1, X_2, ..., X_p\right)^\top$ such as OTU counts, let $F\left(\mathbf{x}\right) = F\left(x_1, x_2, ..., x_p\right)$ denote the CDF. It can be expressed in terms of uniformly distributed marginals $F_j\left(X_j\right)$ and a copula, defined as the joint CDF of a $p \times 1$ random vector $\mathbf{U} = \left(U_1, U_2, ..., U_p\right)^\top$ with uniform marginals $U_j = F_j\left(X_j\right)$ $\left(1 \leqslant j \leqslant p\right)$ (Sklar, 1959). Similar to the univariate case, we can simulate correlated multivariate random vectors $\mathbf{X} = \left(X_1, X_2, ..., X_p\right)^\top$ where $X_j = F_j^{-1}\left(U_j\right)$, with specified marginals $F_j\left(X_j\right)$ through copula.

To simulate $\mathbf{X}$ with distributions similar to those OTUs from a real study, we first use the copula to create a correlated multivariate uniform $\mathbf{U}_n$ based on the eCDF $F_n(\mathbf{x})$ of the observed OTUs, with the uniform marginals $U_{nj} = F_{nj}(X_j) = \frac{1}{n}\sum_{i=1}^{n} I(X_{ij} \leqslant x_j)$. Then by smoothing U_{nj} (Genest et al., 2017), we apply the copula again to create a multivariate normal $\mathbf{V}$ with correlations similar to those of the original OTUs. Afterward, by simulating from $\mathbf{V}$, we obtain correlated multivariate uniform $\mathbf{U}$ with correlations and marginals similar to those of $\mathbf{U}_n$. Finally, by smoothing $F_{nj}(X_j)$ and inverting the simulated U_j to X_j with $X_j = F_j^{-1}(U_j)$, where $F_j(\cdot)$ is a smoothed version of $F_{nj}(\cdot)$**,** we obtain the simulated OTUs $\mathbf{X} = (X_1, X_2, ..., X_p)^\top$ with a distribution similar to $F_n(\mathbf{x})$ of the real OTUs. Beta-diversity was then calculated from simulated OTU counts after appropriate normalization.

As this procedure does not involve analytical distributional models, population-level characteristics such as mean and standard deviation are estimated by Monte Carlo (MC) simulation with a large MC size of 5,000.

0.5 *Details of Simulation for Group Comparison Accounting for Covariates*

We simulate the two covariates from parametric distributions with $x_i^g \sim \text{Bern}(p)$ and $z_i^a \sim U(a,b)$ and then created their respective pairwise counterparts $x_{\mathbf{i}}^g$ and $z_{\mathbf{i}}^a$, where Bern(p) denotes Bernoulli with mean p and $U(a,b)$ a uniform over (a,b). We set:

$$p = 0.45, \quad a = 0, \quad b = 1$$

$$\boldsymbol{\theta}_0 = \left(\beta_0, \beta_{22}^d, \beta_{12}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top = (-0.4595, 0, 0, 0.5, 0.5, 0.5)^\top.$$

To simulate $f(\mathbf{y_i})$ for the regression with covariates, we first simulate Beta-diversity distance $d_{\mathbf{i}}(\mathbf{y_i})$ and then use the two covariates $x_{\mathbf{i}}^g$ and $z_{\mathbf{i}}^a$ to create the mean $h(\mathbf{x_i}, z_{\mathbf{i}}; \boldsymbol{\theta}_0) = \exp\left(\mathbf{u_i}^\top \boldsymbol{\theta}_0\right)$. We next center $d_{\mathbf{i}}(\mathbf{y_i})$ with the true value of β_0 $(= -0.4595)$ to create a "residual" $\varepsilon_{\mathbf{i}} = d_{\mathbf{i}}(\mathbf{y_i}) - \beta_0$, which is then added to $u_{\mathbf{i}}^\top \boldsymbol{\theta}_0$ and expenentiated to create:

$$\widetilde{d}_{\mathbf{i}}(\mathbf{y_i}) = \exp\left(\mathbf{u_i}^\top \boldsymbol{\theta}_0 \ + \ \varepsilon_{\mathbf{i}}\right) = \exp\left(\mathbf{u_i}^\top \boldsymbol{\theta}_0\right) \exp\left(\varepsilon_{\mathbf{i}}\right).$$

By setting $C_0 = E(\exp(\varepsilon_{\mathbf{i}}))$, we obtain simulated $f(\mathbf{y_i}) = C_0^{-1}\widetilde{d}_{\mathbf{i}}(\mathbf{y_i})$. This ensures that $E[f(\mathbf{y_i}) \mid \mathbf{x_i}, z_{\mathbf{i}}] = h(\mathbf{x_i}, z_{\mathbf{i}}; \theta) = \exp\left(\mathbf{u}_{\mathbf{i}}^{\top}\boldsymbol{\theta}\right)$. We estimate C_0 by the sample mean $C_0 = \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\exp(\varepsilon_{\mathbf{i}})$ using a large $n = 5,000$, where $C_0 = 1.000796$ in our setting.

0.6 *Details to Obtain Parameter Estimates from UGEE*

The method to find $\widehat{\boldsymbol{\theta}}$ is through Newton-Raphson using the pseudo-score $\mathbf{U}_n(\boldsymbol{\theta})$. For example, in a model with

$$E[f_{\mathbf{i}} \mid x_{\mathbf{i}}] = h_{\mathbf{i}}(\mathbf{x_i}; \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^{\top} g(x_{\mathbf{i}})\right\}, \ \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{8}$$

where $x_{\mathbf{i}} = \{x_{i_1}, x_{i_2}\}$, $g(\cdot)$ is some symmetric smooth function such as the Euclidean distance. Let

$$S_{\mathbf{i}} = f_{\mathbf{i}} - h_{\mathbf{i}}(\mathbf{x_i}; \boldsymbol{\theta}), \ D_{\mathbf{i}} = \frac{\partial}{\partial \boldsymbol{\theta}} h_{\mathbf{i}}(\mathbf{x_i}; \boldsymbol{\theta}), \ V_{\mathbf{i}} = Var(f_{\mathbf{i}} \mid \mathbf{x_i}, \mathbf{z_i}) = \exp\left\{\boldsymbol{\theta}^{\top} g(x_{\mathbf{i}})\right\},$$

with

$$\mathbf{U}_n(\boldsymbol{\theta}) = \sum_{\mathbf{i}\in C_2^n} \mathbf{U}_{n,\mathbf{i}} = \sum_{\mathbf{i}\in C_2^n} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \mathbf{0}, \tag{9}$$

we can obtain $\widehat{\boldsymbol{\theta}}$ by iterating through

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} &= \sum_{\mathbf{i}\in C_2^n} \left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} \\ &= \sum_{\mathbf{i}\in C_2^n} \left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} \left\{f_{\mathbf{i}} - h_{\mathbf{i}}(\mathbf{x_i}; \boldsymbol{\theta}^{(t)})\right\} \end{aligned} \tag{10}$$

until convergence, where all relevant quantities of $(D_{\mathbf{i}}, V_{\mathbf{i}})$ are evaluated at the t^{th} step with $\boldsymbol{\theta}^{(t)}$.

0.7 *FDR-corrected Test Results for the Real Data Analyses*

We applied the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the family-wise FDR at 5%, and provided comparisons of p-values before and after FDR-correction for the real data analyses.

Shown in top panel of the table are estimates (Est.) of $\boldsymbol{\theta}$, Wald and score test p-values (Wald under "W.p", score under "S.p", Bootstrap Wald under "B.W.p" and Bootstrap score under "B.S.p") for testing the nulls of no difference for the diagnostic groups and no effect for the two covariates**.** The bottom panel includes Wald and score test p-values for the three major types of hypotheses and covariate effects**.**

The comparisons indicate that major conclusions in the real data application remain unchanged after FDR-corrections, except for comparing the between-group variability of AUD-HC pairs vs. the within-group variability of AH-AH pairs with $\beta_{23}^d = 0$, where the score test p-value (S.p) was .020 before and .060 after correction.

[Table 1 about here.]

0.8 *Simulation Details of Power Comparison with the Existing Approach.*

To control for the effect size that allows for appropriate power comparisons in the simulation, the data were generated from the alternative using the Dirichlet-Multinomial distribution (DM) with parameters calibrated from the real data using R package 'dirmult' (*Tvedebrink*, 2010), with effect size estimated with $\frac{\widehat{\theta}-0}{\sqrt{n}se(\widehat{\theta})}$ as a rough quantification. This allows us to vary effect sizes more easily for the power comparison and continues to generate Beta-diversity outcomes with their distributions resembling the real data as shown in the Figure S2 below.

0.9 *Supplemental Figures*

[Figure 1 about here.]

[Figure 2 about here.]

0.10 *Declaration of interests*

B.S. has been consulting for Ferring Research Institute, Intercept Pharmaceuticals, HOST Therabiomics, Mabwell Therapeutics and Patara Pharmaceuticals. B.S.'s institution UC San Diego has received grant research support from BiomX, NGM Biopharmaceuticals, CymaBay Therapeutics, Synlogic Operating Company and Axial Biotherapeutics.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300.

Genest, C., Nešlehová, J. G., and Rémillard, B. (2017). Asymptotic behavior of the empirical multilinear copula process under broad conditions. *Journal of Multivariate Analysis* **159,** 82–110.

Kowalski, J. and Tu, X. M. (2008). *Modern applied U-statistics*, volume 714. John Wiley & Sons.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris* **8,** 229–231.

Tvedebrink, T. (2010). Overdispersion in allelic counts and θ-correction in forensic genetics. *Theoretical population biology* **78,** 200–210.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

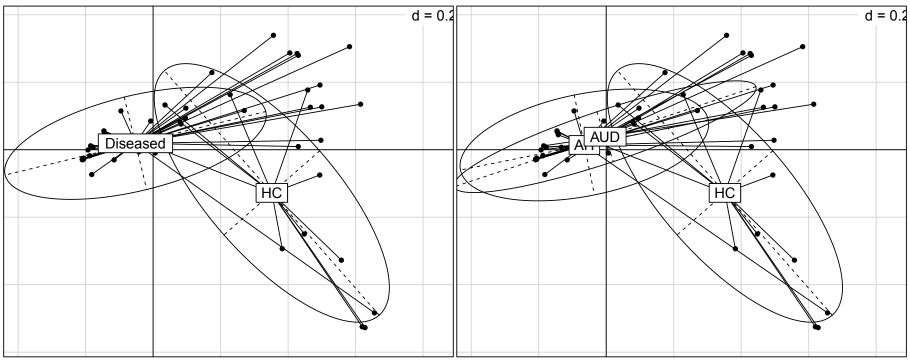

Figure S1. Principal Coordinates Analysis (PCoA) plots of Beta-diverisity distance for (1) combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (left) and (2)alcoholic hepatitis (AH) patients, alcohol user disorder (AUD) patients and non-alcoholic controls (HC) (right)

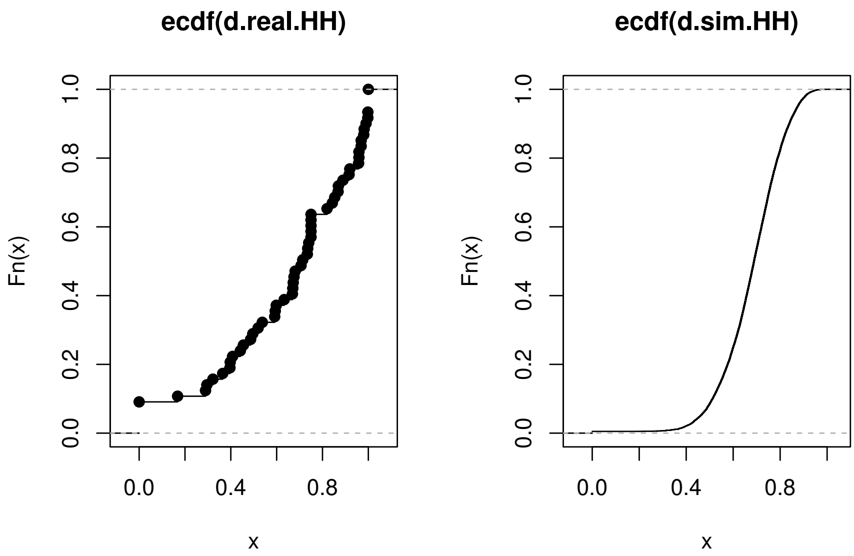

Figure S2. Empirical CDFs of Real vs. Simulated Beta-diversity.

Table S1

Table S1. Comparisons of p-values before and after FDR-correction for the Real Data Analyses

Categorical Covariate: Gender ($\boldsymbol{\beta}^g$), Continuous Covariate: Age (ξ^a)									
Parameter	Est.	p-value				FDR-corrected p-value			
		W.p	S.p	B.W.p	B.S.p	W.p	S.p	B.W.p	B.S.p
β_0	-1.042	<.0001	.0002	<.0001	<.0001	<.0001	.002	<.0001	<.0001
β_{22}^d	.226	.454	.506	.419	.662	.554	.565	.519	.674
β_{33}^d	.572	.002	.130	.007	<.0001	.005	.293	.016	<.0001
β_{12}^d	.114	.554	.565	.519	.674	.554	.565	.519	.674
β_{13}^d	.634	<.0001	.006	.002	<.0001	<.0001	.027	.006	<.0001
β_{23}^d	.672	<.0001	.020	.0004	<.0001	<.0001	.060	.002	<.0001
β_{22}^g	.125	.509	.528	.477	.613	.554	.565	.519	.674
β_{12}^g	.072	.550	.551	.511	.583	.554	.565	.519	.674
ξ^a	.006	.189	.224	.184	.348	.340	.403	.331	.626

Hypothesis	p-value				FDR-corrected p-value			
	W.p	S.p	B.W.p	B.S.p	W.p	S.p	B.W.p	B.S.p
$\beta_{22}^d = \beta_{33}^d = 0$	.007	.071	.017	<.0001	.016	.160	.038	<.0001
$\beta_{12}^d = \beta_{13}^d = \beta_{23}^d$	<.0001	<.0001	.001	<.0001	<.0001	<.0001	.005	<.0001
$\beta_{12}^d = 0$	.554	.565	.519	.674	.623	.636	.584	.758
$\beta_{13}^d = 0$	<.0001	.006	.002	<.0001	<.0001	.027	.006	<.0001
$\beta_{23}^d = 0$	<.0001	.020	.0004	<.0001	<.0001	.060	.004	<.0001
$\xi^a = 0$	.189	.224	.184	.613	.340	.403	.331	.758
$\beta_{22}^g = 0$	.509	.528	.477	.583	.623	.636	.584	.758
$\beta_{12}^g = 0$	.550	.551	.511	.348	.623	.636	.584	.626
$\beta_{22}^g = \beta_{12}^g = 0$	.733	.886	.732	1.000	.733	.886	.732	1.000