SUPPLEMENTARY INFORMATION

Addressing Missing Data in GC×GC Metabolomics: Identifying Missingness Type and Evaluating the Impact of Imputation Methods on Experimental Replication

Trenton J. Davis[a,b†], Tarek R. Firzli[c†], Emily A. Higgins Keppler[a,b†], Matthew Richardson[d,e*], and Heather D. Bean[a,b*]

[a] *School of Life Sciences, Arizona State University, Tempe, AZ, 85287 United States*

[b] *Center for Fundamental and Applied Microbiomics, The Biodesign Institute, Tempe, AZ, 85287 United States*

[c] *School of Medicine, University of Nevada, Reno, NV, 89557 United States*

[d] *Department of Respiratory Sciences, College of Life Sciences, University of Leicester, Leicester, LE1 7RH, United Kingdom*

[e] *NIHR Biomedical Research Centre (Respiratory Theme), Institute for Lung Health, Leicester, LE1 7RH, United Kingdom*

*Email: mr251@leicester.ac.uk, Heather.D.Bean@asu.edu

[†] T.J.D., T.R.F., and E.A.H.K. contributed equally to this paper.

Table of Contents

Definitions of imputation methods

Zero: Impute all missing values to 0.

Minimum: Impute all missing values to the minimum value. This can be a global minimum (of all present data) or a local minimum of the feature in question. The local minimum was used in this study.

Half-minimum (HM): Impute all missing values to half of the global or local minimum value. The local half-minimum was used in this study.

Mean: Impute all missing values to the global or local mean value. The local mean was used in this study.

Median: Impute all missing values to the global or local median value. The local median was used in this study.

Bayesian principal component analysis (BPCA): This method utilizes Bayesian and expectation maximization methods to determine principal components. These are then used to impute missing values.

Random Forest (RF): RF imputation iteratively fits ensembles of uncorrelated decision trees—random forests—to observed data and uses these models to predict missing values.

Quantile regression imputation of left-censored data (QRILC): This imputation method was specifically designed to deal with missingness of the MNAR type where data is missing due to the limits-of-detection, also called left-censored missingness. This method draws data from the lower quantile of a distribution that is estimated based on quantile regression.

Gibbs Sampler Imputation (GSImp): GSImp is a method developed for left-censored data which begins by initially imputing data using QRILC. Elastic net regression models are then utilized to build a distribution from which updated imputation values will be drawn. This process is iteratively repeated.

Replicate Zero (RepZero): First, permute to zero any replicate group that has > 50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values using zero imputation.

Replicate Minimum (RepMin): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values using the minimum value within the replicate samples.

Replicate Half-Minimum (RepHM): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values using the half-minimum value within the replicate samples.

Replicate Mean (RepMean): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values using the mean value within the replicate samples.

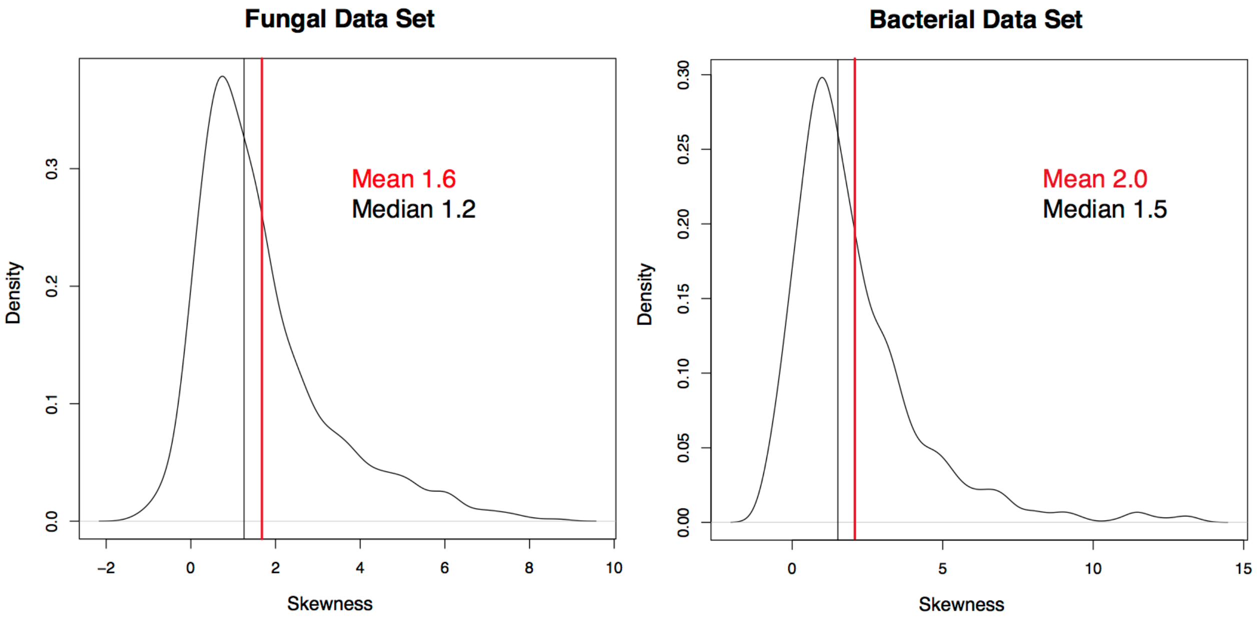
Replicate Median (RepMedian): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values using the median value within the replicate samples.

Replicate BPCA (RepBPCA): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values across the entire data set using BPCA and the present values across the entire data set.

Replicate RF (RepRF): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values across the entire data set using RF and the present values across the entire data set.

Replicate QRILC (RepQRILC): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values across the entire data set using QRILC and the present values across the entire data set.

Replicate GSImp (RepGSImp): First, permute to zero any replicate group that has >50% missing values (change the entire group of replicates to 0 for that feature). Imputation is performed on the remaining missing values across the entire data set using GSImp and the present values across the entire data set.



Supplementary Figure 1. Density plots of the full fungal and bacterial data sets.

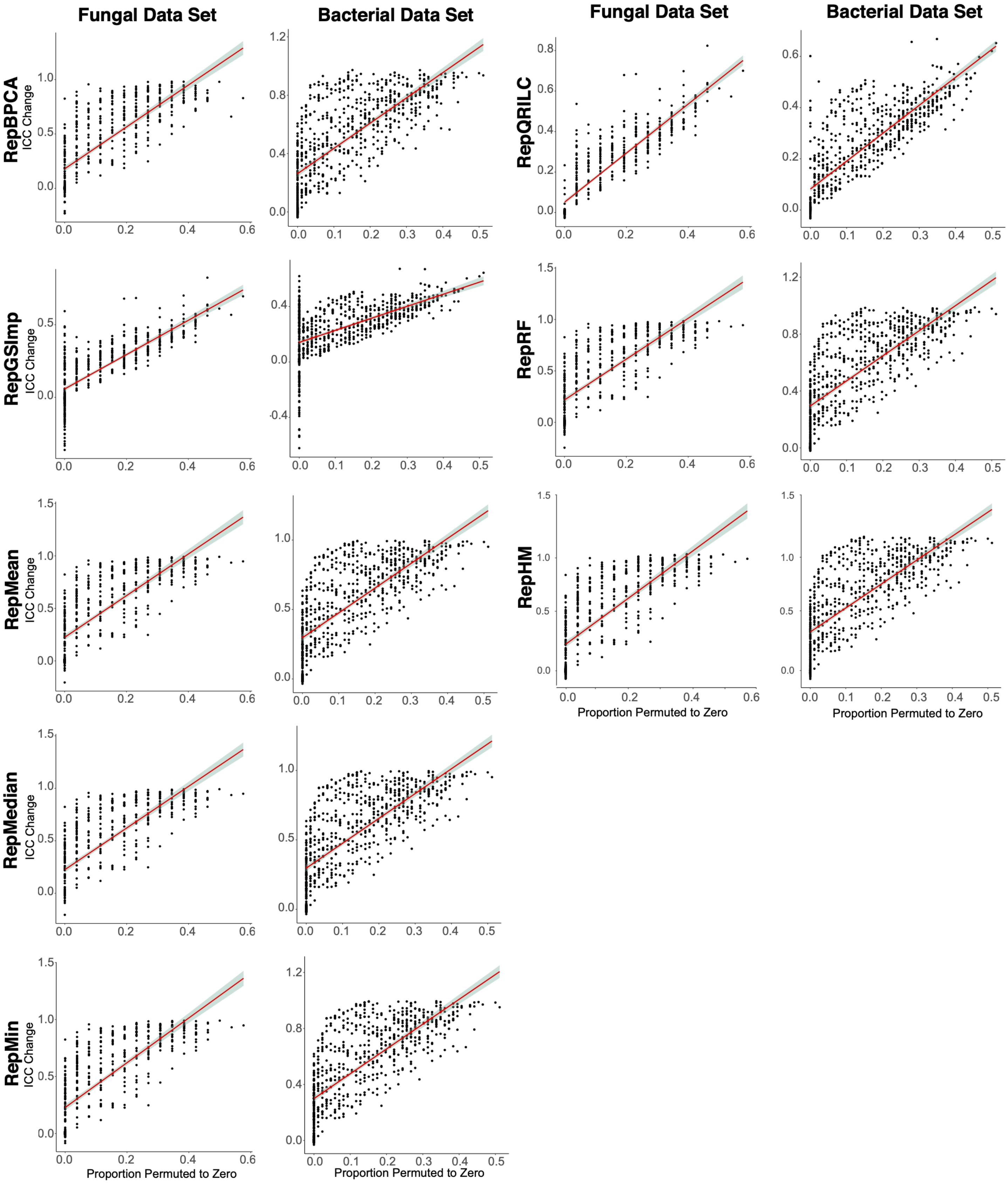
Supplementary Table 1. ICC shifts following imputation.

Fungal Data Set

	RepHM	HM	RepMin	Min	RepMean	Mean	RepMedian	Median	RepBPCA	BPCA	RepQRILC	QRILC	RepRF	RF	RepGSImp	GSImp
Excellent to Good	1	5	0	8	0	4	0	4	0	1	0	2	1	2	3	7
Excellent to Moderate	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	1
Excellent to Poor	0	0	0	0	0	5	0	5	0	5	0	0	0	5	0	0
Good to Excellent	13	4	14	5	14	0	14	0	12	0	9	0	13	0	8	4
Good to Moderate	0	11	0	12	0	13	0	15	0	1	0	3	0	5	11	0
Good to Poor	0	1	0	1	0	12	0	9	0	12	0	1	0	9	0	1
Moderate to Excellent	43	3	43	4	43	0	43	0	32	1	7	0	42	2	8	4
Moderate to Good	5	10	2	9	2	2	2	4	10	3	21	0	4	3	20	17
Moderate to Poor	2	18	0	21	0	53	0	44	2	41	1	5	2	33	13	9
Poor to Excellent	346	0	358	1	362	0	362	2	236	0	3	0	354	0	3	1
Poor to Good	39	5	27	7	22	2	22	3	105	3	14	0	28	5	15	11
Poor to Moderate	26	27	25	28	25	13	25	15	54	29	133	1	22	33	145	45

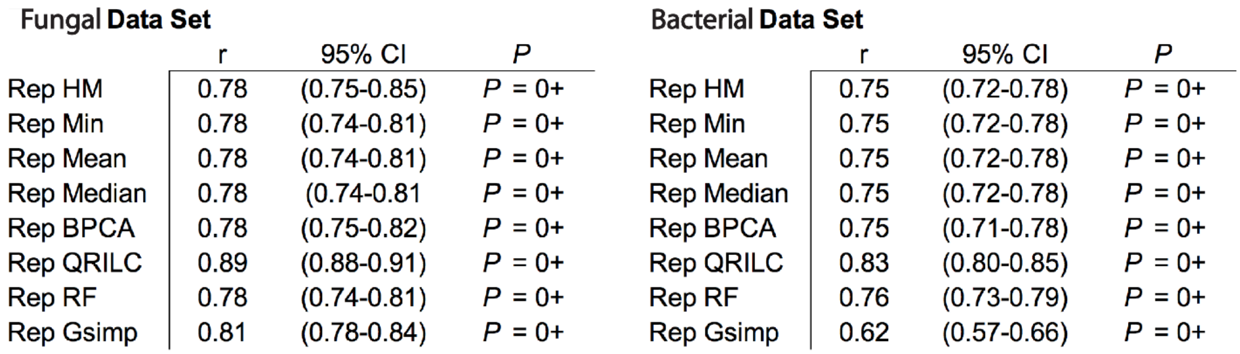
Bacterial Data Set

	RepHM	HM	RepMin	Min	RepMean	Mean	RepMedian	Median	RepBPCA	BPCA	RepQRILC	QRILC	RepRF	RF	RepGSImp	GSImp
Excellent to Good	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	6
Excellent to Moderate	0	0	0	0	0	3	0	2	0	1	0	0	0	3	0	1
Excellent to Poor	0	0	0	0	0	3	0	3	0	8	0	8	0	6	0	0
Good to Excellent	30	8	30	8	30	2	30	2	25	0	7	1	30	1	9	5
Good to Moderate	0	4	0	3	0	12	0	12	0	14	0	5	0	15	1	5
Good to Poor	0	0	0	2	0	18	0	17	0	14	0	9	0	7	0	0
Moderate to Excellent	95	0	95	1	95	0	95	0	74	0	4	0	95	0	4	1
Moderate to Good	21	63	20	63	20	19	20	25	30	19	55	4	20	37	83	57
Moderate to Poor	0	12	0	15	0	75	0	66	0	60	0	24	0	43	8	8
Poor to Excellent	464	0	471	1	473	0	473	1	375	0	2	0	466	0	2	0
Poor to Good	48	14	43	16	41	13	41	18	125	29	19	0	48	104	29	16
Poor to Moderate	39	72	37	77	36	49	36	60	42	188	254	18	34	230	278	123



Supplementary Figure 2. Relationship between the change of volatile ICCs relative to RepZero imputation and the proportion of features permuted to zero within a group of replicates. The line represents the linear regression fit and its standard error (shaded region).

Supplementary Table 2. Point and interval estimates of linear regressions of the relationship between the change of volatile ICCs and the proportion of features permuted to zero within a group of replicates using within-replicate imputation.



Fungal **Data Set**

	r	95% CI	*P*
Rep HM	0.78	(0.75-0.85)	*P* = 0+
Rep Min	0.78	(0.74-0.81)	*P* = 0+
Rep Mean	0.78	(0.74-0.81)	*P* = 0+
Rep Median	0.78	(0.74-0.81	*P* = 0+
Rep BPCA	0.78	(0.75-0.82)	*P* = 0+
Rep QRILC	0.89	(0.88-0.91)	*P* = 0+
Rep RF	0.78	(0.74-0.81)	*P* = 0+
Rep Gsimp	0.81	(0.78-0.84)	*P* = 0+

Bacterial **Data Set**

	r	95% CI	*P*
Rep HM	0.75	(0.72-0.78)	*P* = 0+
Rep Min	0.75	(0.72-0.78)	*P* = 0+
Rep Mean	0.75	(0.72-0.78)	*P* = 0+
Rep Median	0.75	(0.72-0.78)	*P* = 0+
Rep BPCA	0.75	(0.71-0.78)	*P* = 0+
Rep QRILC	0.83	(0.80-0.85)	*P* = 0+
Rep RF	0.76	(0.73-0.79)	*P* = 0+
Rep Gsimp	0.62	(0.57-0.66)	*P* = 0+

Supplementary Table 3. List of R packages loaded in MetabImpute.

abind	Matrix
doParallel	matrixStats
e1071	missForest
FNN	missRanger
foreach	nonpar
gamlss	pcaMethods
ggplot2	PEMM
glmnet	randomForest
gridExtra	reshape2
ICC	Rmisc
impute	rpart
imputeLCMD	stats
ltm	tidyr
magrittr	vegan
MASS	