Supporting Information: 3 figures and 3 tables across 6 pages

Influence of Library Composition on SourceTracker Predictions for Community-Based Microbial Source Tracking

Clairessa M. Brown[1], Prince P. Mathai[1], Tina Loesekann[1,2], Christopher Staley[1,3], and Michael J. Sadowsky[1,4*]

[1]BioTechnology Institute, University of Minnesota, St. Paul, MN

[2]Department of Microbiology & Immunology, University of Minnesota, Minneapolis, MN

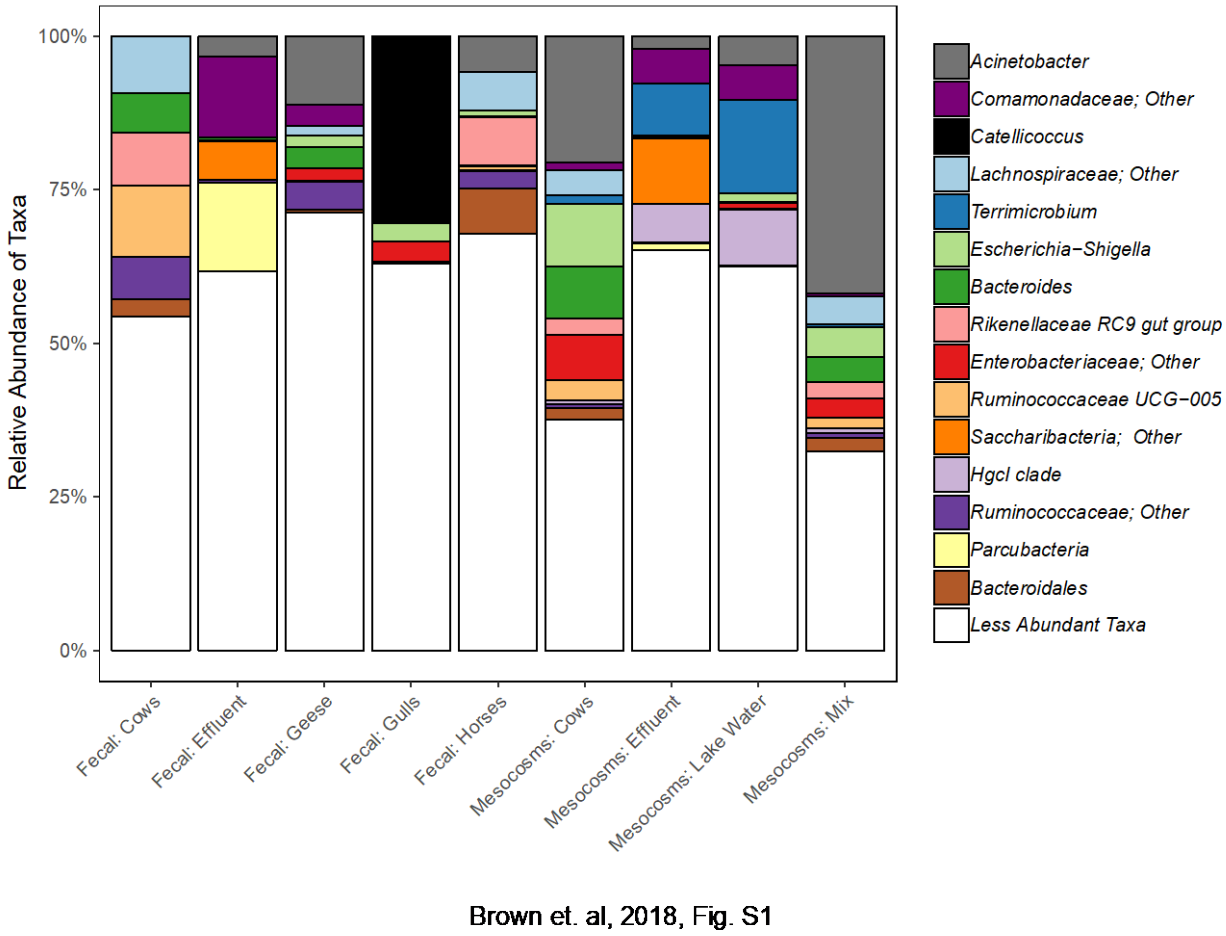
[3]Department of Surgery, University of Minnesota, Minneapolis, MN

[4]Department of Soil, Water & Climate, and Department of Microbial and Plant Biology, University of Minnesota, St. Paul, MN

*Correspondence:

Michael J. Sadowsky: 1479 Gortner Ave., 140 Gortner Labs, BioTechnology Institute, University of Minnesota. St. Paul, MN 55108, USA; Email: sadowsky@umn.edu

Keywords: DNA sequencing; operational taxonomic units; computational biology; community-based microbial source tracking; SourceTracker; predictions



Brown et. al, 2018, Fig. S1

Figure S1. **Stacked taxonomic bar charts.** The averaged relative abundances of the most abundant 15 taxonomic groups.

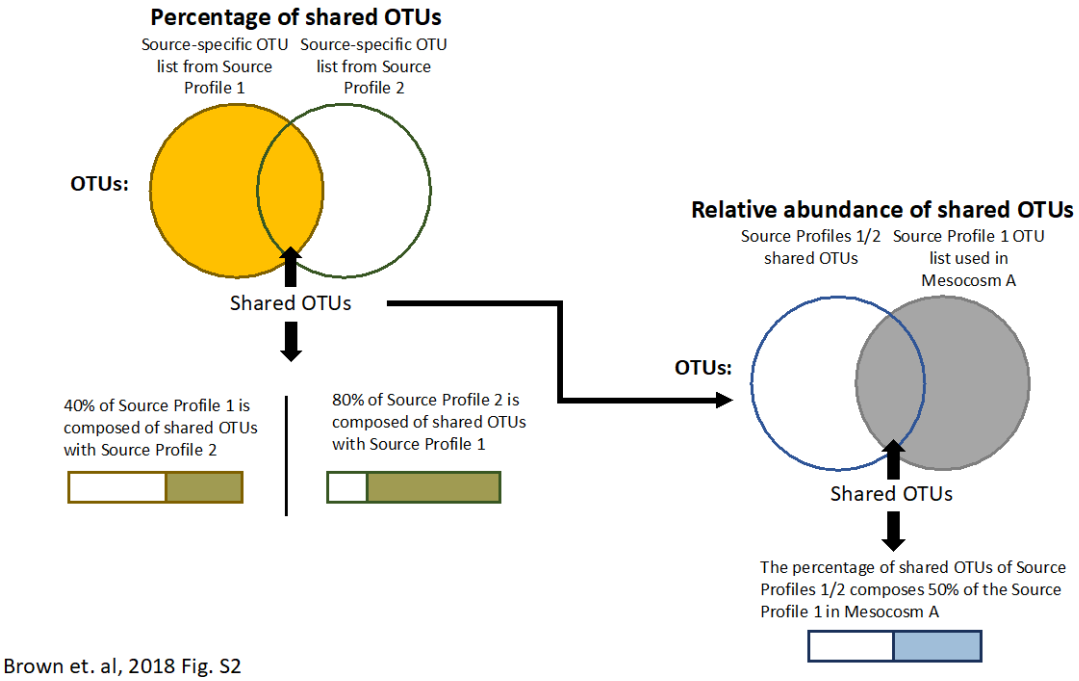


Figure S2. Depiction of the percentage of shared OTUs and relative abundance of shared OTUs. The percentage of shared OTUs are the percentages of taxa in each source profile that are shared between two different source profiles. The relative abundance of shared OTUs is the fraction of the percentage of shared OTUs that compose a source profile used to predict a source in a mesocosm.

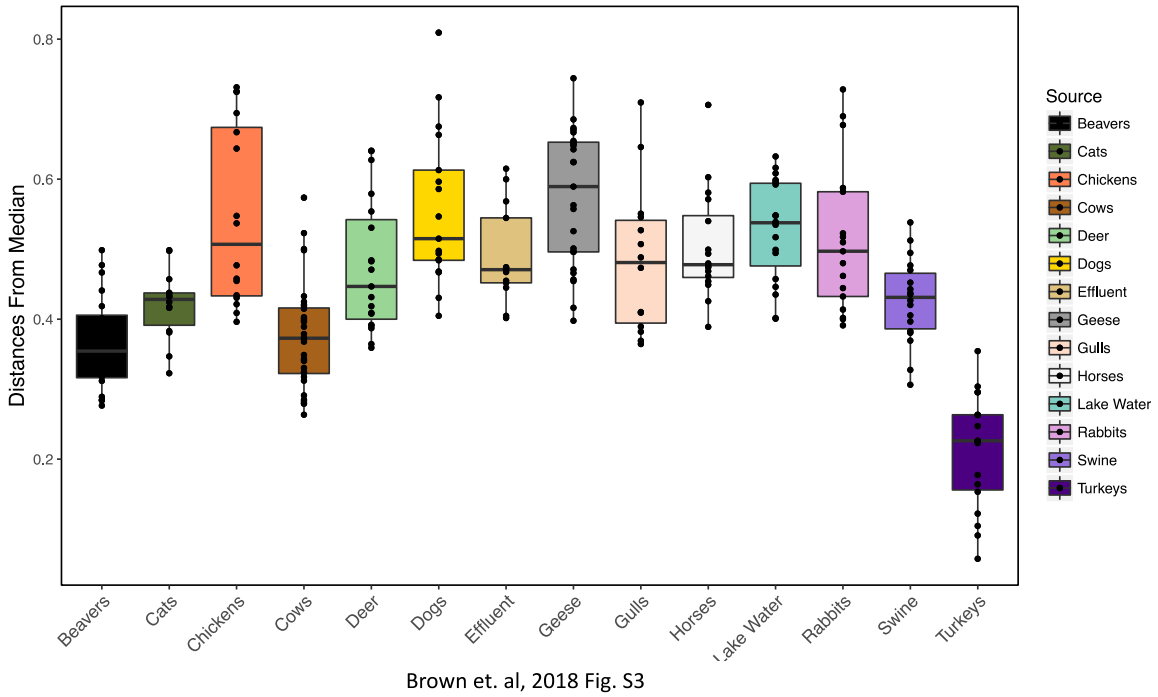


Figure S3. Boxplot depicting intra-group variances of source groups. Black dots are all samples within a group. A multivariate version of Levene's test for homogeneity of variances was performed on source group samples. Higher distances from the median indicate higher variation within the group. To have sample numbers that resembled other source groups, the cow source group was reduced to 20 samples instead of 32 for this analysis.

Table S1. Presence and Absence of SourceTracker predictions in all mesocosms

FTL Configuration	Mesocosm	Source				
		Cow	Horse	Effluent	Lake Water	Other
Only Known Sources with lake water as source	Cow	+	+	-	+	NA
	Effluent	-	-	+	+	NA
	Mix	+	+	-	+	NA
Only Known Sources with lake water as sink	Cow	+	+	+	NA	NA
	Effluent	-	-	+	NA	NA
	Mix	+	+	+	NA	NA
	Lake Water	-	-	+	NA	NA
All Available Sources with lake water as source	Cow	+	+	-	+	+
	Effluent	-	-	+	+	-
	Mix	+	+	-	+	+
All Available Sources with lake water as sink	Cow	+	+	+	NA	+
	Effluent	-	-	+	NA	+
	Mix	+	+	+	NA	+
	Lake Water	-	-	+	NA	+
Missing Sources Sources Available: Cow & Lake Water	Cow	+	NA	NA	+	NA
	Effluent	-	NA	NA	+	NA
	Mix	+	NA	NA	+	NA
Missing Sources Sources Available: Effluent & Lake Water	Cow	NA	NA	-	+	NA
	Effluent	NA	NA	+	+	NA
	Mix	NA	NA	-	+	NA
Missing Sources Sources Available: Horse & Lake Water	Cow	NA	+	NA	+	NA
	Effluent	NA	-	NA	+	NA
	Mix	NA	+	NA	+	NA
Missing Sources Sources Available: Cow	Cow	+	NA	NA	NA	NA
	Effluent	+	NA	NA	NA	NA
	Mix	+	NA	NA	NA	NA
	Lake Water	+	NA	NA	NA	NA
Missing Sources Sources Available: Effluent	Cow	NA	NA	+	NA	NA
	Effluent	NA	NA	+	NA	NA
	Mix	NA	NA	+	NA	NA
	Lake Water	NA	NA	+	NA	NA
Missing Sources Sources Available: Horse	Cow	NA	+	NA	NA	NA
	Effluent	NA	+	NA	NA	NA
	Mix	NA	+	NA	NA	NA
	Lake Water	NA	+	NA	NA	NA

Unknown source not included in table. Results from both experiments represented in this table. Presence indicated by "+" and absence indicated by "-". NA means that the source was not available for SourceTracker to use. Sources were considered present when the SourceTracker predictions were above 1% and the RSD value was below 100%.

Table S2. RSD values associated with SourceTracker predictions

FTL Configuration	Mesocosm	Source	Average SourceTracker Prediction (%)	RSD (%)
Only Known Sources with lake water as sink	Cow	Cows	48	22
		Effluent	18	26
		Horses	20	42
		Unknown	14	22
	Effluent	Effluent	41	10
		Unknown	59	7
	Mix	Cows	22	18
		Effluent	4	18
		Horses	61	8
		Unknown	13	17
	Lake Water	Cows	4	150
		Effluent	42	16
		Horses	8	146
		Unknown	46	26
All Available Sources with lake water as sink	Cow	Cows	40	30
		Effluent	13	31
		Geese	23	63
		Gulls	9	65
		Horses	2	76
		Unknown	13	19
	Effluent	Effluent	32	15
		Geese	4	46
		Gulls	3	60
		Unknown	61	8
	Mix	Cows	19	21
		Effluent	2	17
		Geese	31	34
		Horses	37	16
		Unknown	10	20
	Lake Water	Dogs	1	149
		Effluent	32	28
		Geese	14	129
		Gulls	2	90
		Unknown	50	33

The relative standard deviation (RSD) was calculated for all sources in all mesocosms to assess confidence in the SourceTracker predictions.

Table S3. Significance values from pairwise comparisons of intra-group variances

	Beaver	Cat	Chicken	Cow	Deer	Dog	Effluent	Goose	Gull	Horse	Rabbit	Swine	Turkey	Cow Mesocosm	Effluent Mesocosm	Mix Mesocosm	Lake Water
Beaver		*	**		**	**	**	**	**	**	**	**	**	**	**	*	**
Cat			**			**	**	**		**	**		**	**	**		**
Chicken				**								**	**			**	
Cow					**	**	**	**	**	**	**	*	**	**	**	*	**
Deer						*		**					**				
Dog												**	**			**	
Effluent								*				*	**			*	
Goose									*	*		**	**			**	
Gull													**				
Horse												**	**			*	
Rabbit												**	**			**	
Swine													**	**	**		**
Turkey														**	**	**	**
Cow Mesocosm																**	
Effluent Mesocosm																**	
Mix Mesocosm																	**
Lake Water																	

Different asterisks symbolize different ranges of p-values generated from permutation test of the betadisper function. * signals a p-value below 0.05 but above 0.01. ** signals a p-value below 0.01 but above 0.001.