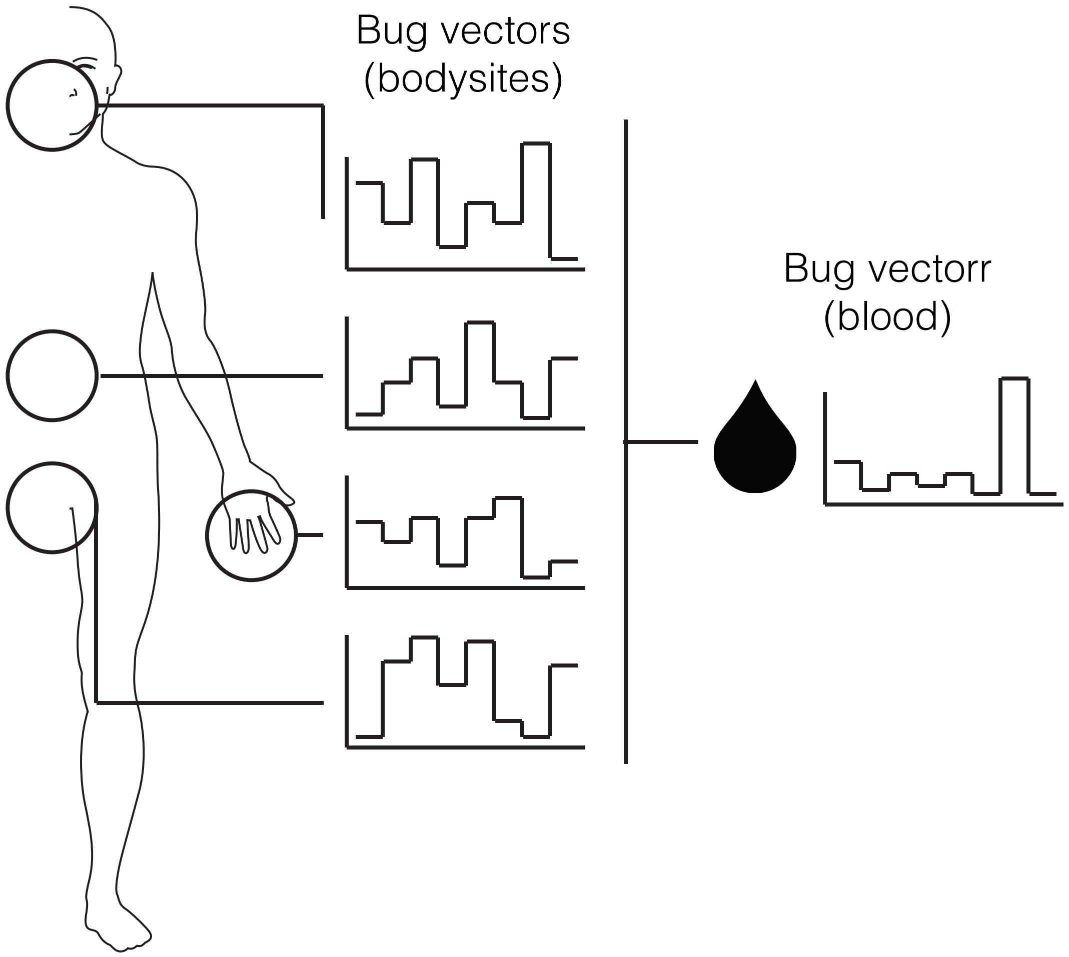
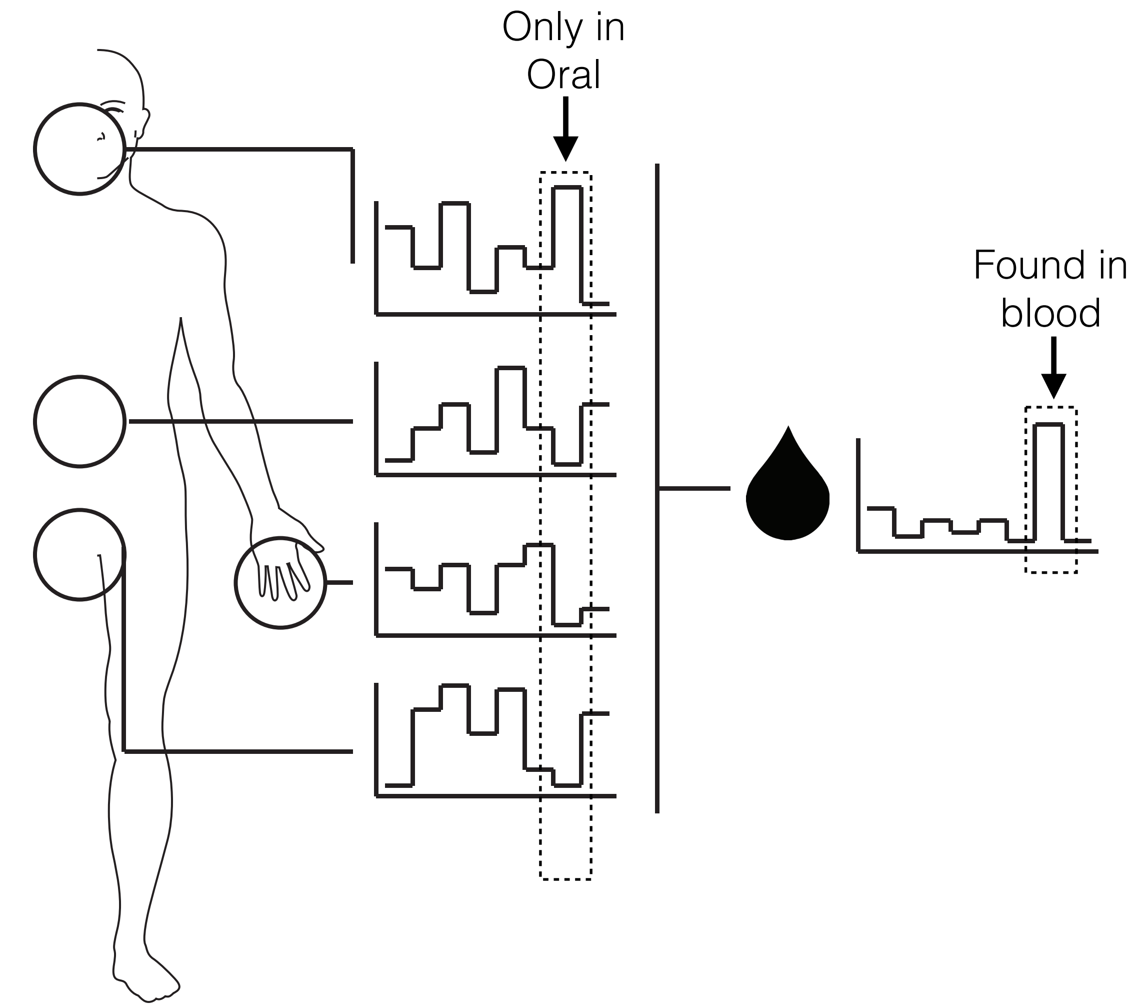
What are the tissue sources of the human blood microbiome?

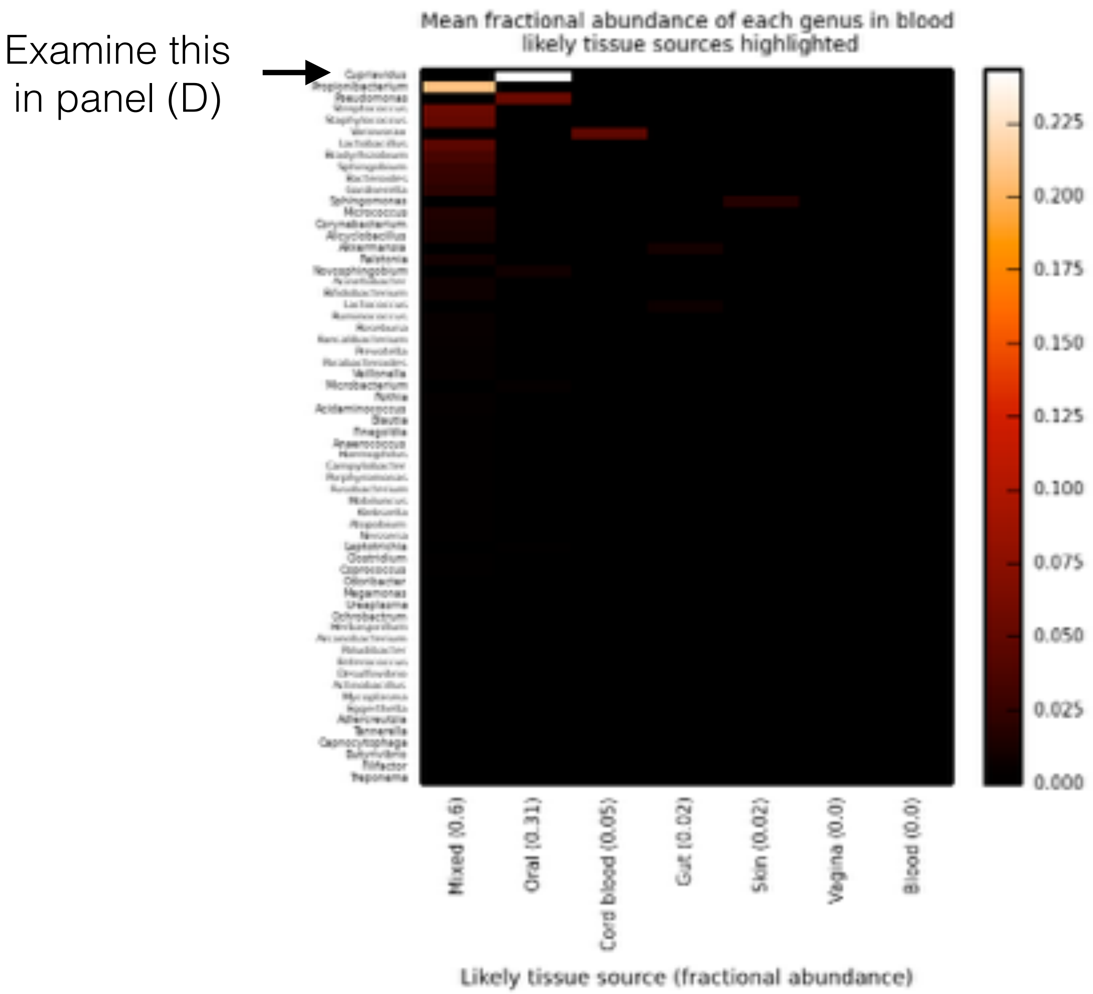


The tissue sources bugs found in human blood is a clinically important question. We have ~ 10-fold more microbial cells than human cells within our bodies. The human microbiome is now recognized as an important contributor to human health. The advent of next-generation sequencers are making it increasingly easy to measure the diversity of microbes on our bodies. Human blood is an appealing medium for making these measurements, as it is rich with information in the form of cell-free DNA (~ 100 billion molecules per mL) that is largely derived from dead human and microbial cells. We have shown that deep shotgun sequencing of human blood can be used to diagnose infection, which is useful for monitoring in certain illnesses (e.g., inflammatory bowel disease) as well as cases of high infection risk (e.g., pregnancy, cancer, immunosuppression). Here, we explore the tissue sources that contribute to human blood using ~ 200 microbiome measurements from four sampled body site (Oral, Vagina, Gut, Skin) with temporally matched (~60) microbiome measurement of blood measurements within a cohort of pregnant women at Stanford hospital.

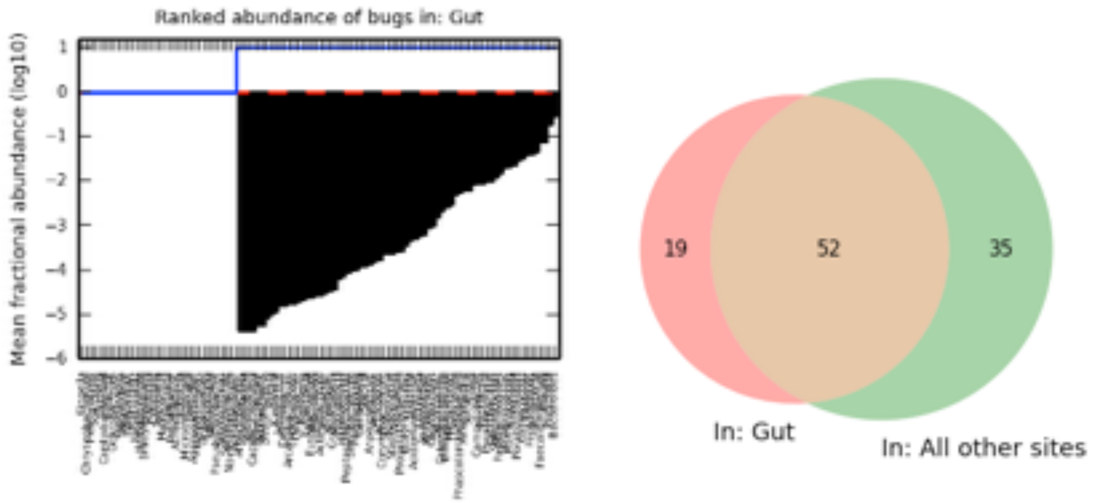
The majority of bugs found in blood are not specific to any body site.



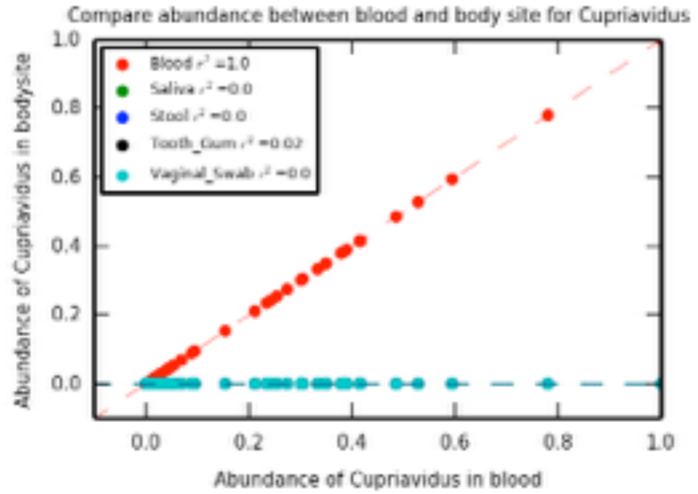
(A) **Define tissue specific bugs.** We start by determining the bugs that are specific to certain sampled body sites and evaluating their abundance in blood in order to evaluate which sites blood may be sampling.



(C) **Assign likely source to each bug found in blood.** Ranked abundance of bugs in blood (y) with the tissue source of each (x).

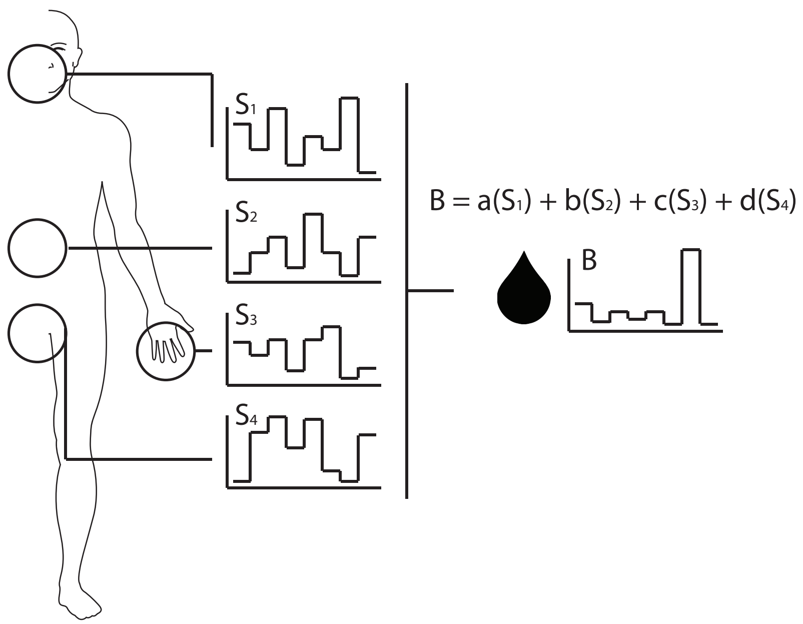


(B) **Bugs at each site are re-cast as bit vectors.** Discretize abundance of bugs measured at each body site and evaluate relative all other sites.



(D) **The most abundant bug in blood (Cuprividius) has trace abundance in Oral and is found is not other sites.**

Supervised learning with linear models to determine tissue composition of blood.



(A) **Assume blood is a linear combination of bugs found at the sampled sites**. If so, we can simply use use source and blood data to learn the parameters that weight each source.

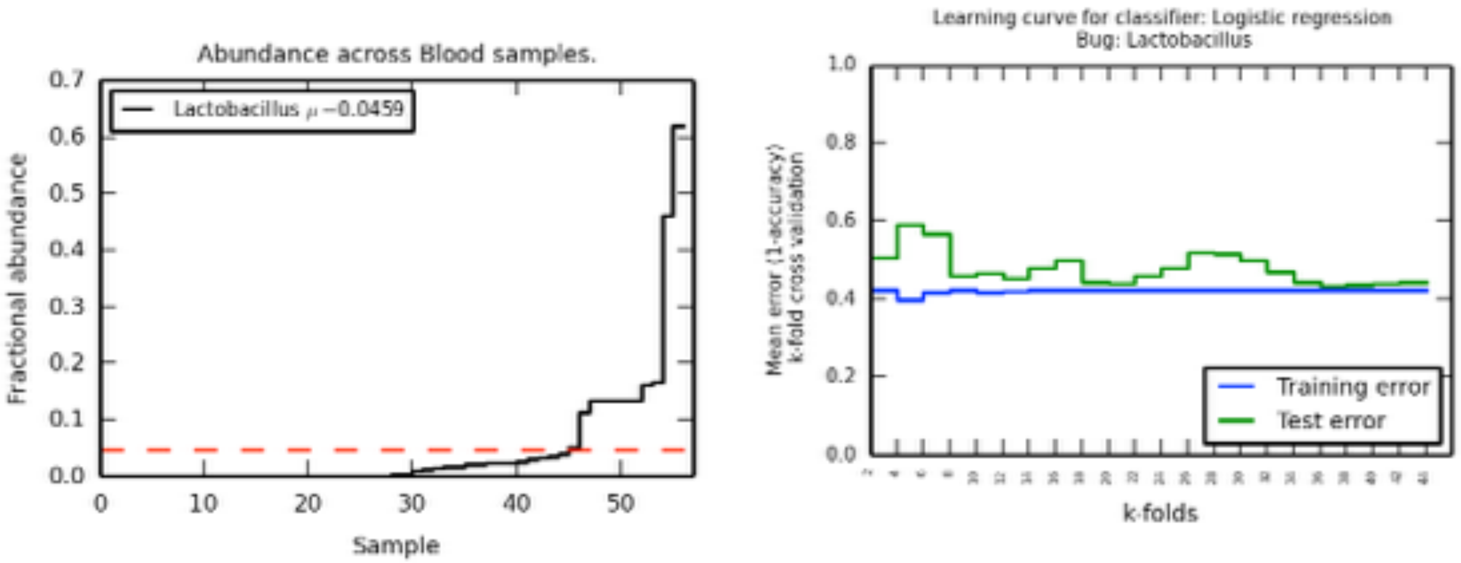
$$Y = \frac{1}{1 + e^{-\theta^T x}}$$

- θ: set of tissue weights that are learned for a specific bug.
- x: vector of bug fractional abundances per tissue in the given sample.
- Y: label indicating whether a bug is found in blood in the given sample.

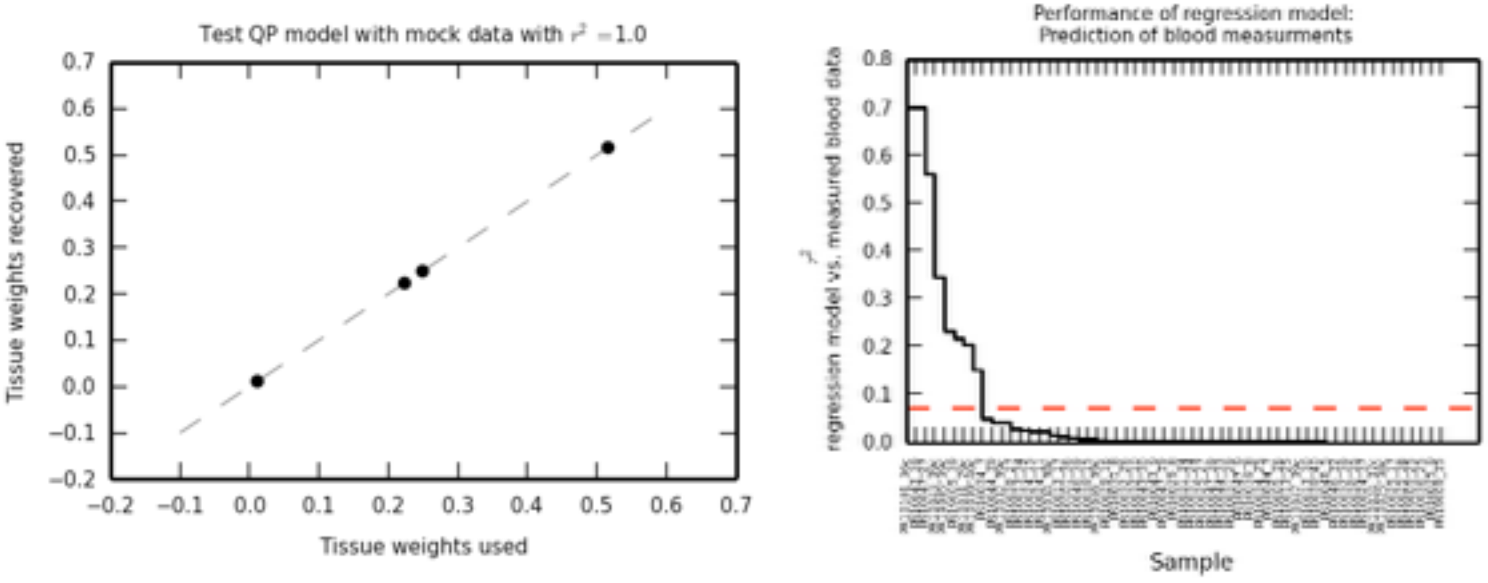
$$\vec{B} = \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix} = \sum_j x_j \theta_j = \begin{bmatrix} b_{j1} \\ \dots \\ b_{jn} \end{bmatrix} \theta_j + \dots + \epsilon$$

- θ: a vector of tissue weights that are learned for each sample.
- x_j: The vector of bugs measured at site j.
- $\vec{B}$: a vector of bug abundance in blood for a paritcular sample.

(B) **Consider two linear models**. (i) Classification: Pick one bug, discretize its detection in blood, learn weights for tissue sources that predict whether it is detected in blood. (ii) Regression: Pick one sample. Learn tissue weights that best describe the observed vector of bugs in blood using the vectors of bugs measured for each tissue source.

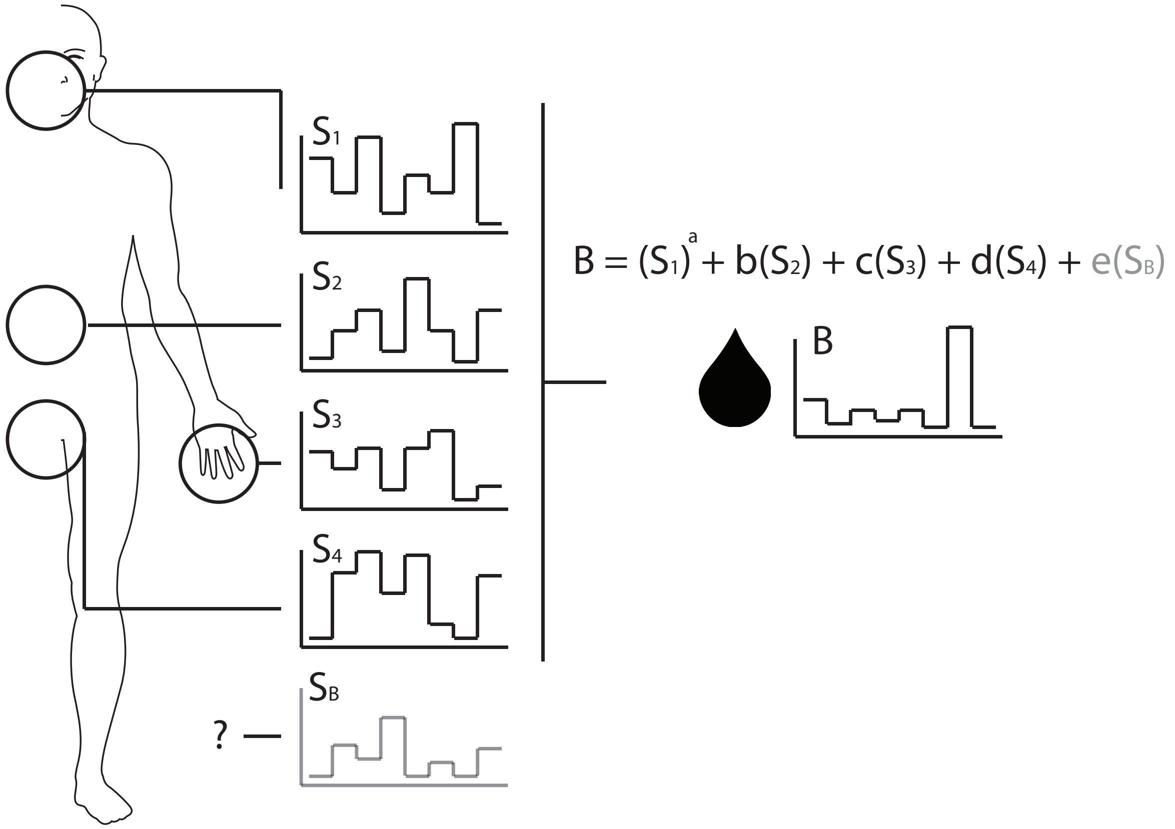


(C) **Classification with logistic regression performs poorly.** We pick a bug (Lactobacillus) with even labeling (detection) in blood (to avoid trivial model results) and compute a learning curve to evaluate whether the model.

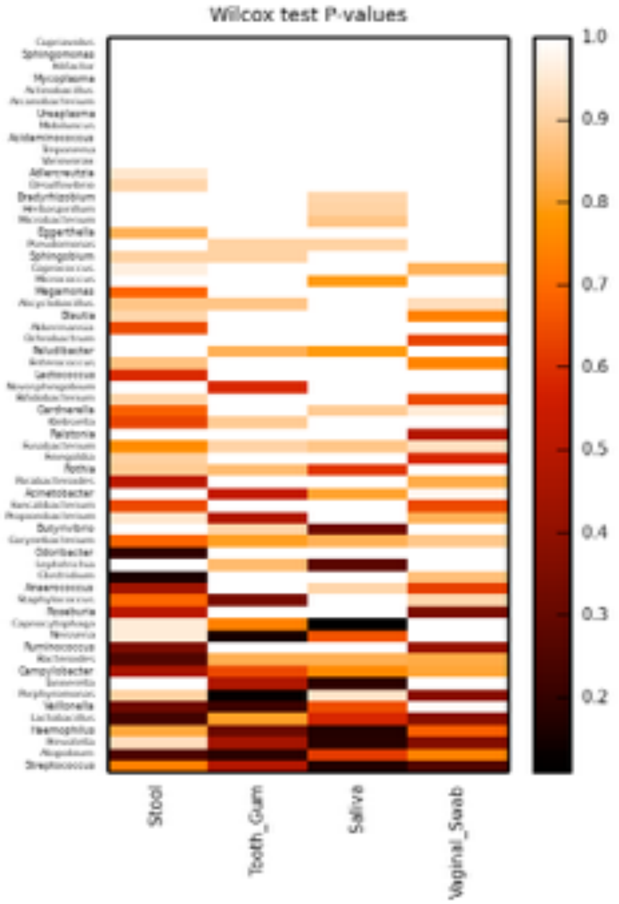


(D) **Linear regression performs poorly.** We apply linear regression with sensible constraints on the learned tissue weights (e.g., > 0). We use a quadratic and show that a test case with simulated data using tissue weights (x) shows that these weights are recovered (y) correctly. We then applied the model each sample in our cohort. We observe poor performance (as measured by correlation between measured and model-derived blood data).

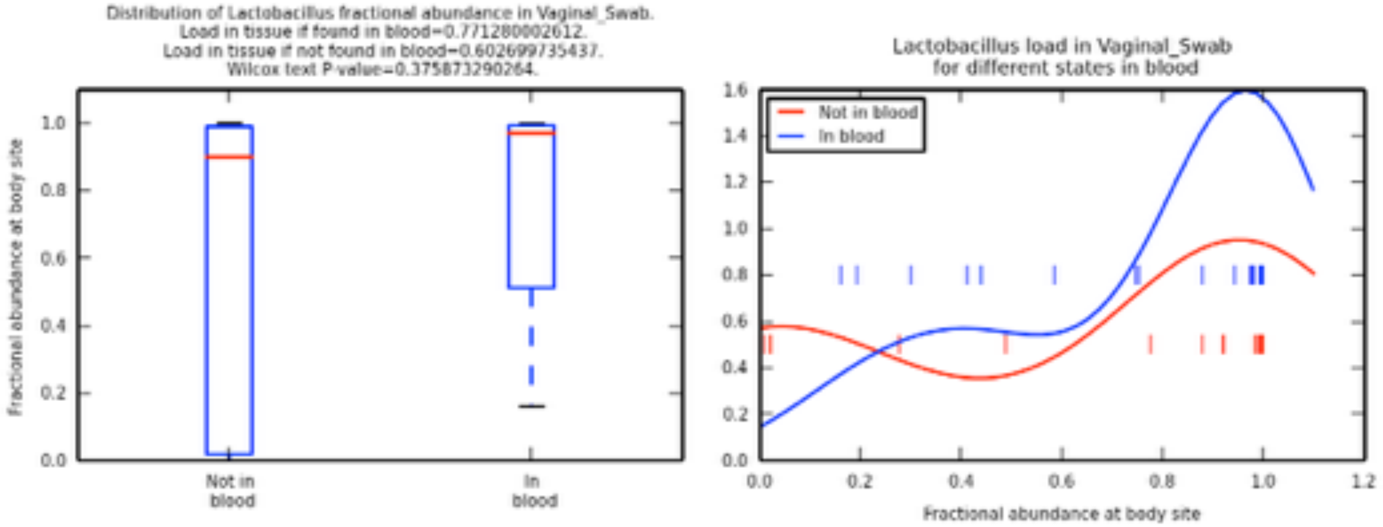
Examine why linear models doesn't work.



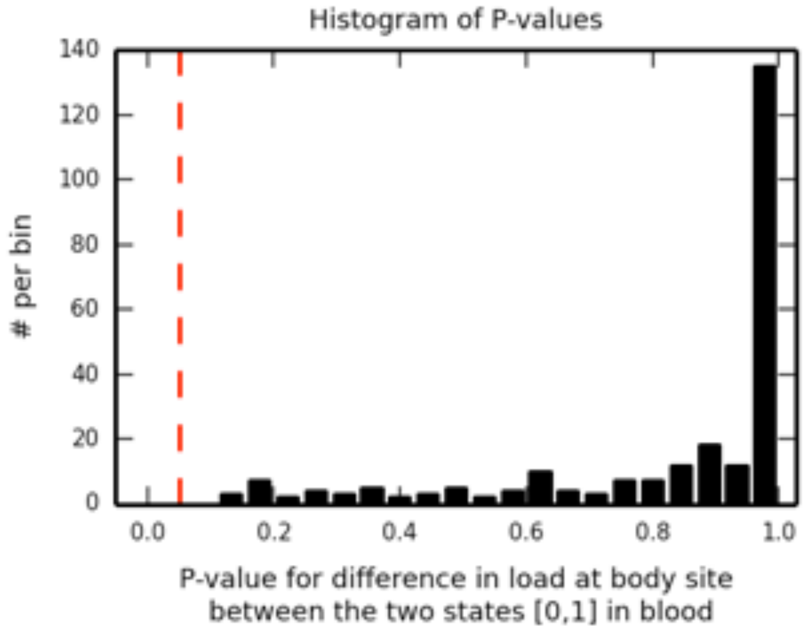
(A) **Blind sources and non-linearity frustrate performance of linear models.** Learning curve from logistic regression indicated high bias (underfit model), suggesting non-linear relationships. Residuals from linear regression show that blood samples contain some bugs that are not found in any sampled sites.



(C) **No signifiant differences in bug distribution.** Wilcox test performed for all bug-body site combinations.



(B) **Examine distributions.** Simply ask whether the distribution of a specific bugs differs at a body site depending upon whether it is detected in blood?

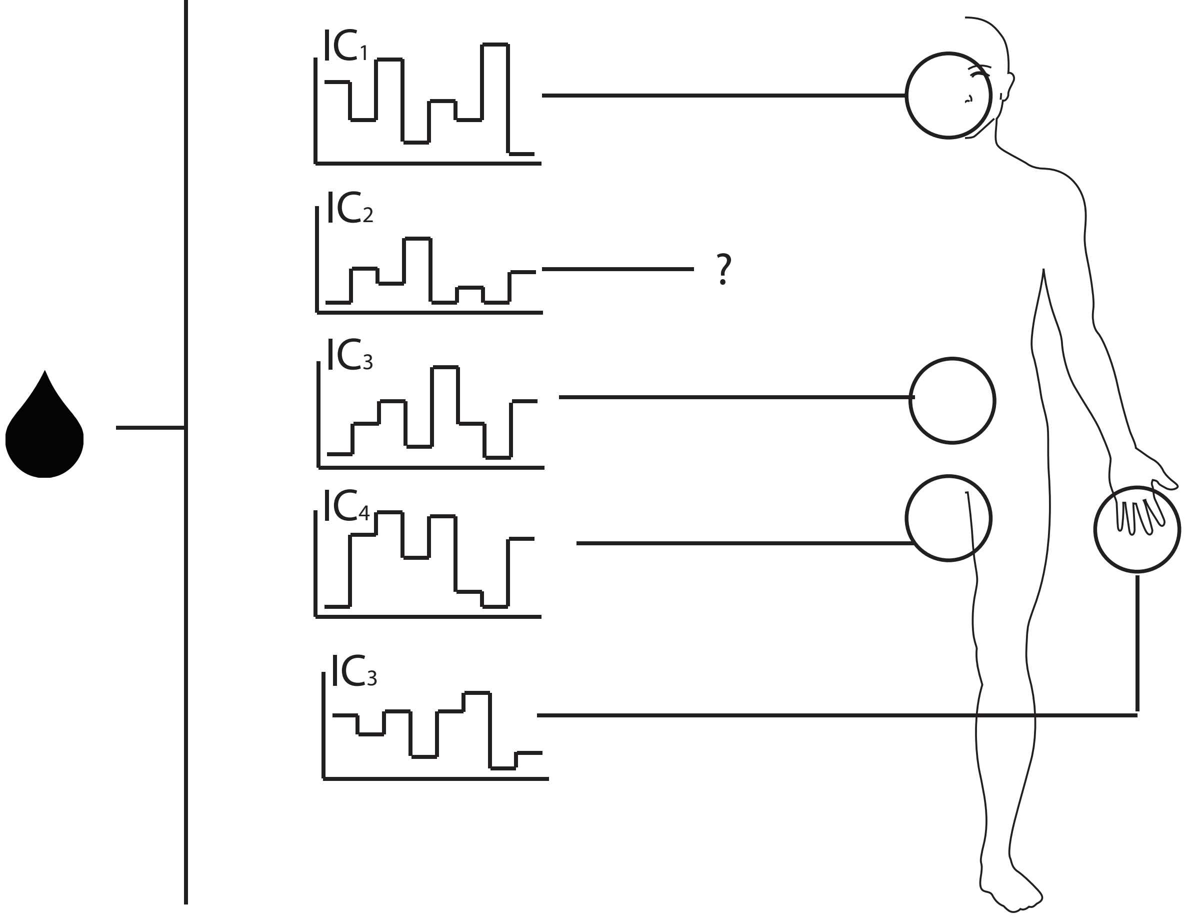


(D) **Histogram of P-values from (C) with cutoff.**

< Kosh add slides here to show how leaning can work correctly on biome data >

Unsupervised source detection from human blood.

$$B_K = A_{K1}(IC_1) + A_{K2}(IC_2) + A_{K3}(IC_3) + A_{K4}(IC_4) + A_{K5}(IC_5)$$

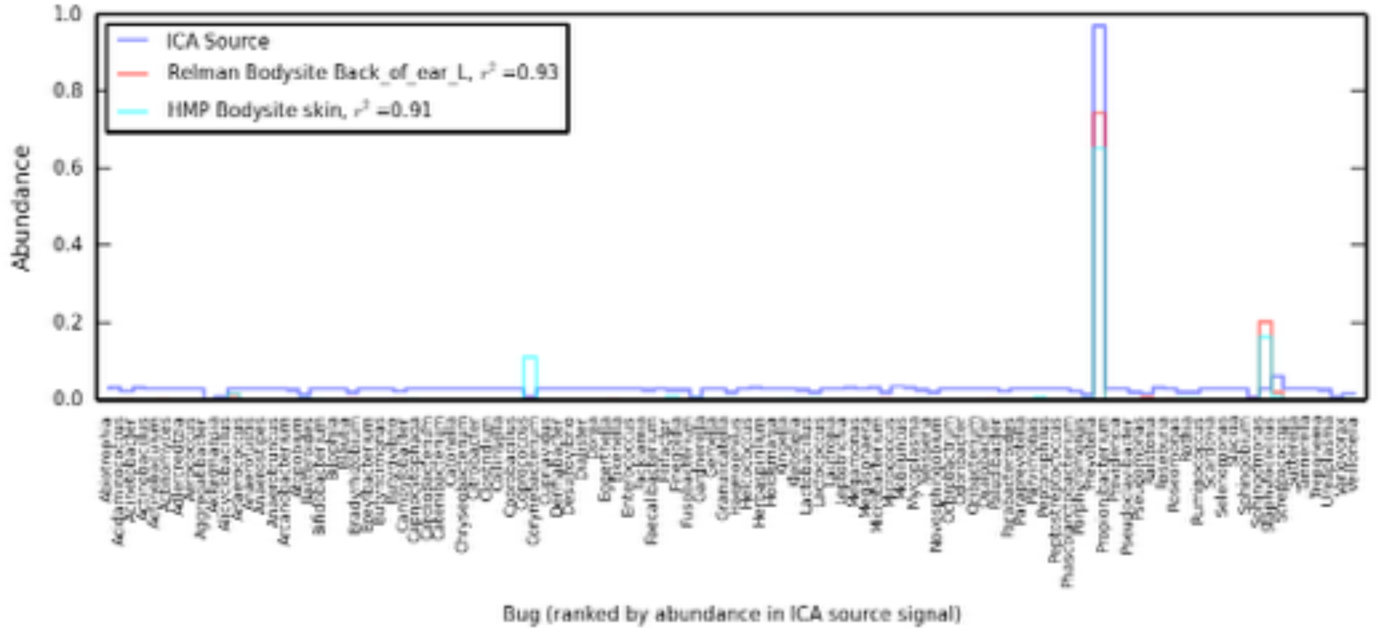


$$x_j = a_{j1}s_1 + \ldots + a_{jM}s_M$$

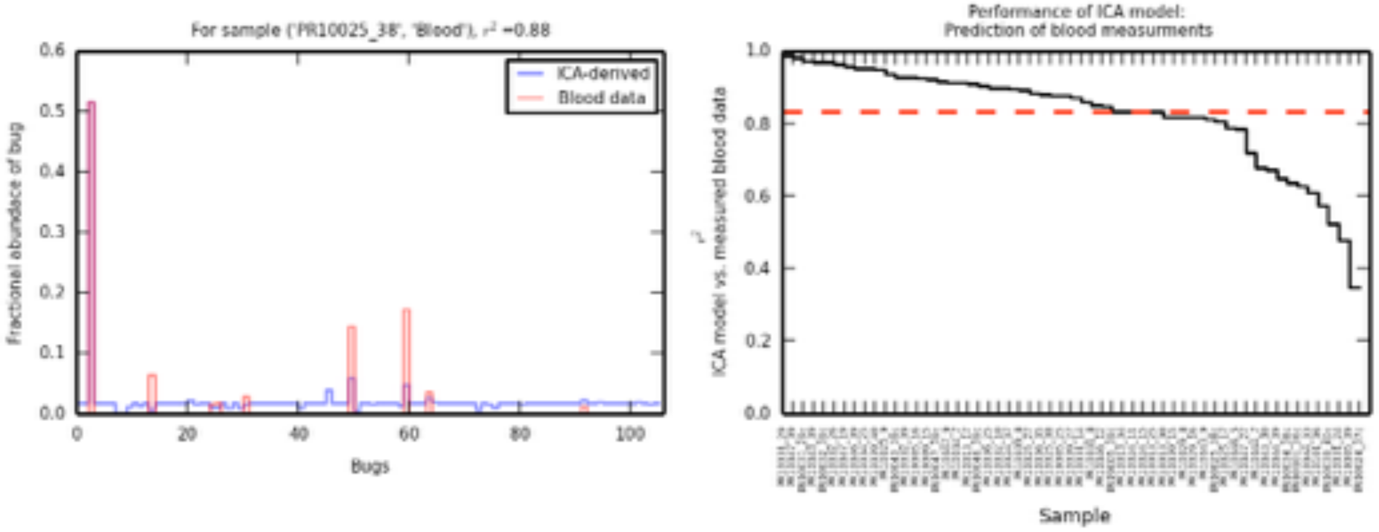
The model in matrix notation:

$$\begin{bmatrix} x_1 \\ \ldots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{11} \ldots a_{1M} \\ \ldots \\ a_{N1} \ldots a_{NM} \end{bmatrix} \begin{bmatrix} s_1 \\ \ldots \\ s_M \end{bmatrix}$$

(A) **Use ICA to compute sources and then assign them to likely tissues *ex post facto*.** ICA will compute a set of sources and a mixing matrix, A, that defines show these sources are mixed to give each measured blood sample. In this case, we can assign likely tissues to each source that ICA computes using both our cohort data as well as Human Microbiome Project data. Critically, this method is unsupervised, meaning that “blind” sources are included in the model and each sample is a unique mixture of these identified sources (using the mixing matrix defined).

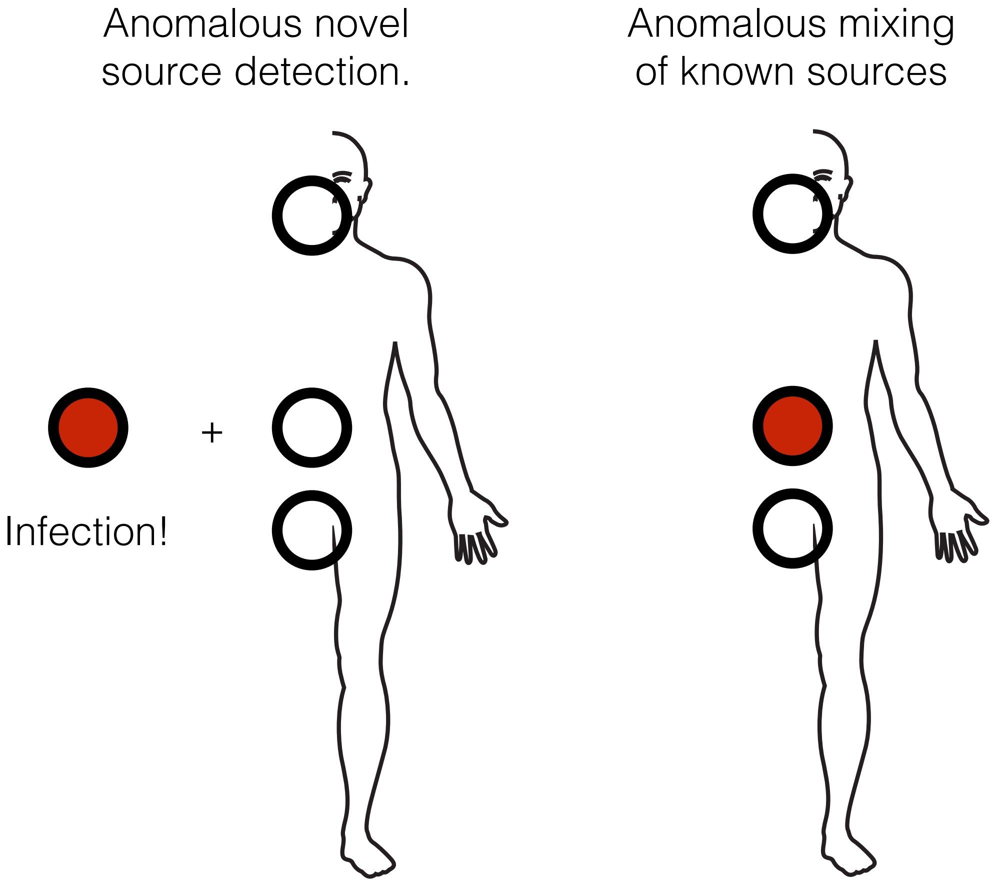


(B) **ICA sources can be assigned to tissues.** Assignment of one ICA source to a specific body site using both Human Microbiome Project (HMP) and our cohort body site data simply by correlation.



(C) **ICA performs well on the data.** For sample, the computed ICA sources and the mixing matrix can be used to compute the model guess for measured blood data (see top of panel A). We can then evaluate the model performance by computing the correlation between the computed vector and the actual blood measurement for that sample. One example sample is shown (left) and data for all samples is shown (right).

ICA as a tool for anomaly detection in human blood microbiome data.

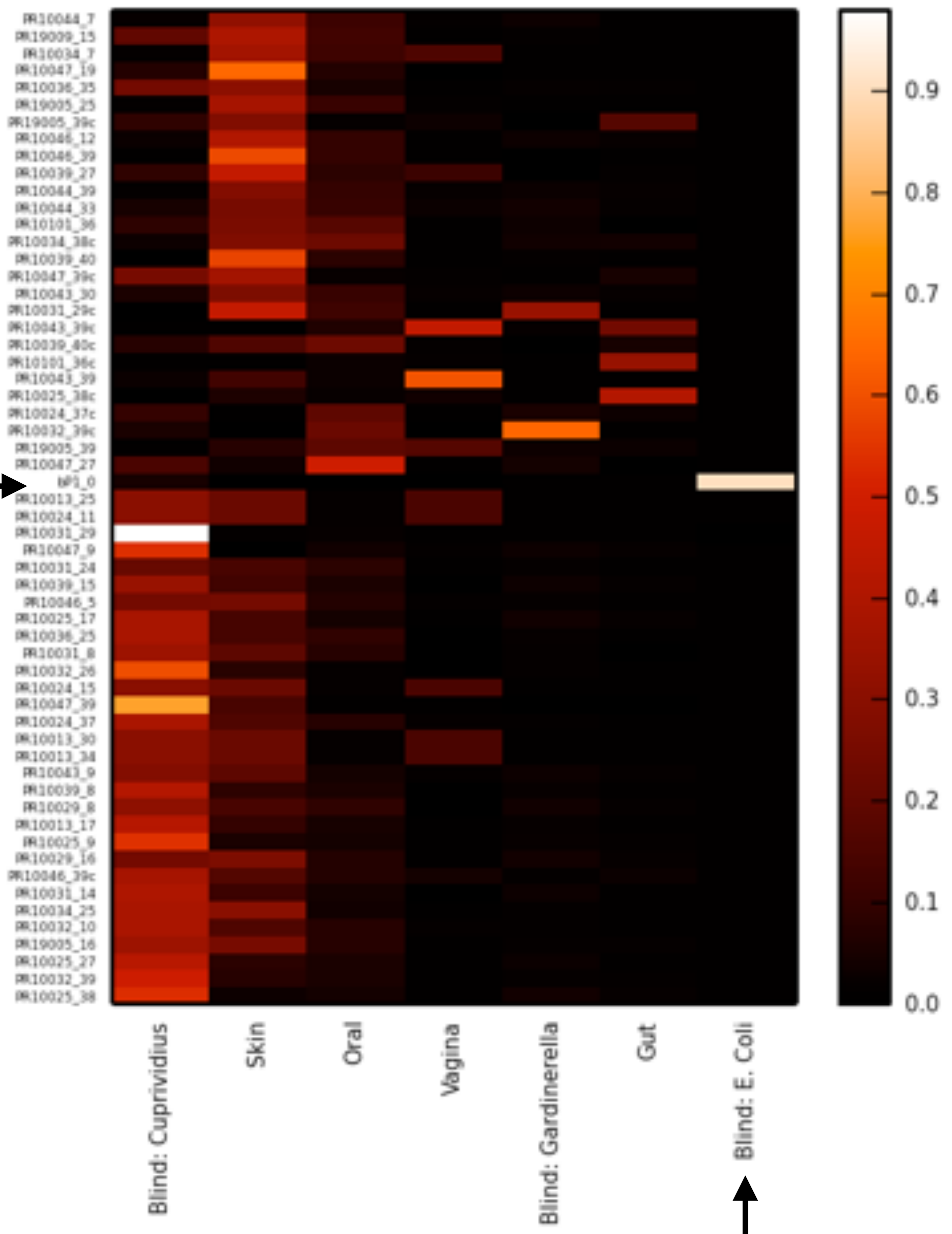


Anomalous novel source detection.

Anomalous mixing of known sources

+

Infection!



Spike in of deep tissue infection

Source detected!

(A) **Utility of ICA for blood microbiome de-convolution.** Because it is un-supervised, ICA can "discover" sources that do not correlate well to any of the sampled sites in microbiome studies. In turn, these sources may represent (i) deep tissues that contribute bugs to blood but have not been sampled, (ii) bugs that grow uniquely in blood (e.g., Cuprividius, as shown in this cohort), or (iii) pathogenic infections (see panel B). Furthermore, analysis of the mixing matrix returned from ICA can reveal (i) clustering of cohorts that show similar source mixing (e.g., which may partition healthy versus un-healthy patient groups) or (ii) outliers that show either aberrant mixing of known sources (e.g., IBD patients or chemo patient may show enrichment of gut microbes due to leakage) or strong mixing of "novel" (pathogenic) sources (see panel B).

(B) **Analysis of ICA mixing matrix can reveal insights.** After spike in of one blood sample taken from a patient with a known (diagnosed) deep tissue abdominal infection, ICA "discovers" this novel (pathogenic source) and weighs it heavily / uniquely in this single sample. The pregnancy data show clustering to the "blind" Cuprividius source and a skin-line source, though there is no significant segmentation (patient, time, pre-term birth) between these two clusters.