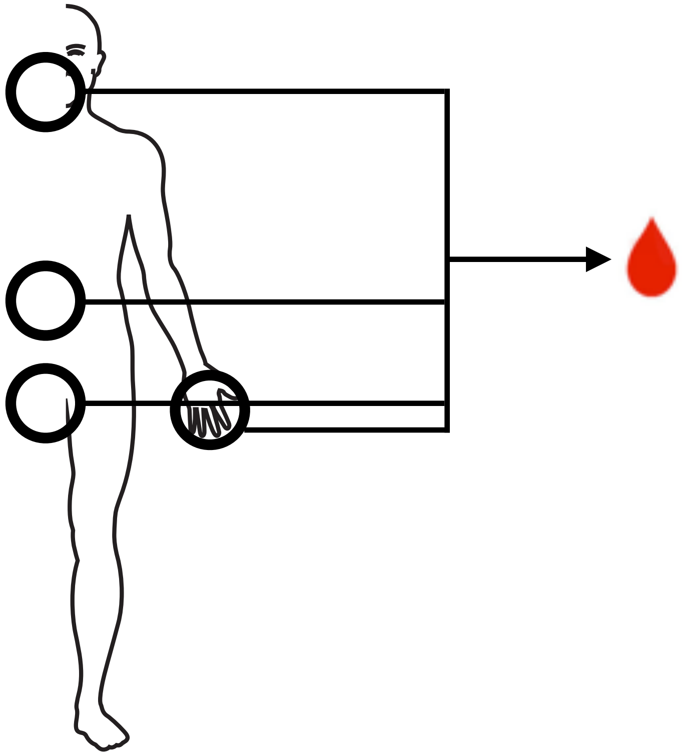
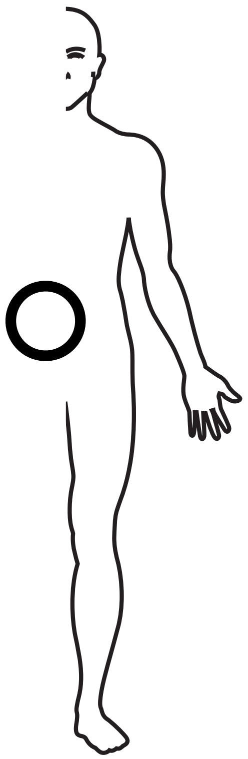
Unsupervised microbial source detection from human blood

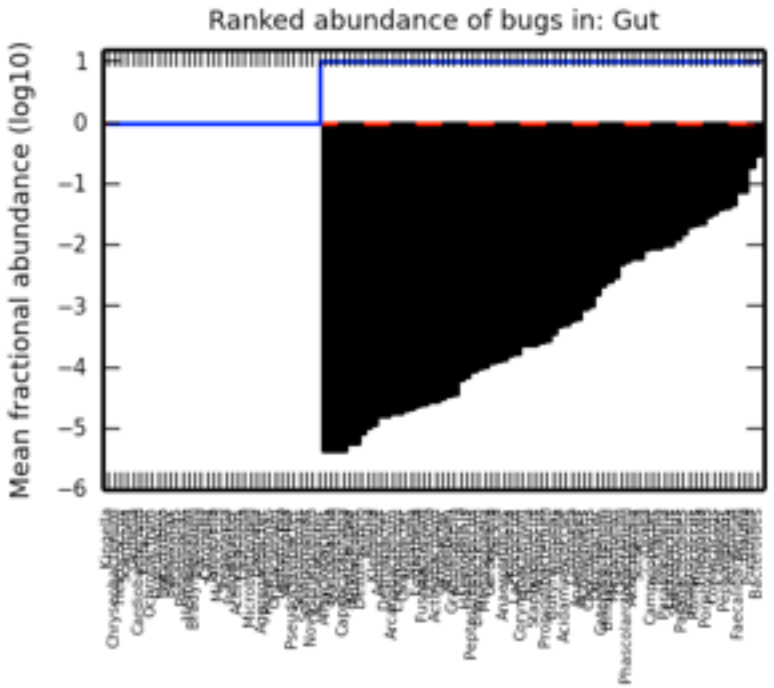
Examine microbiome composition of blood relative to body sites.

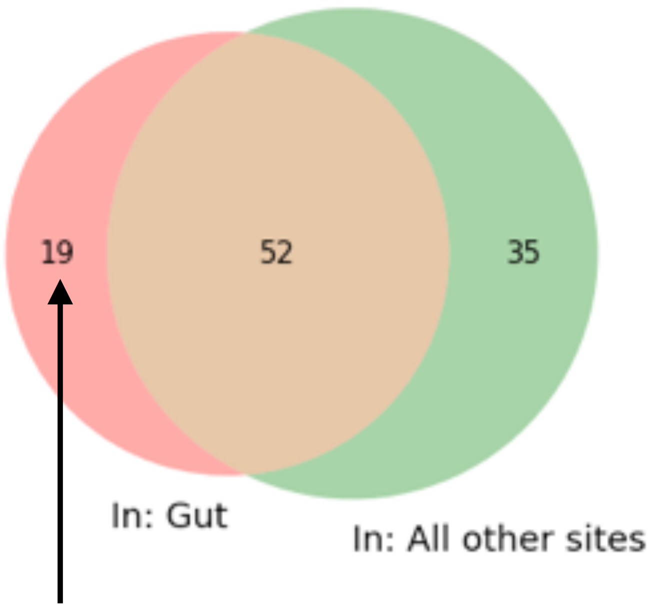


58 blood samples with matched body sites from healthy pregnancy cohort.

Discretize mean abundance data for each site and use this to evaluate blood.

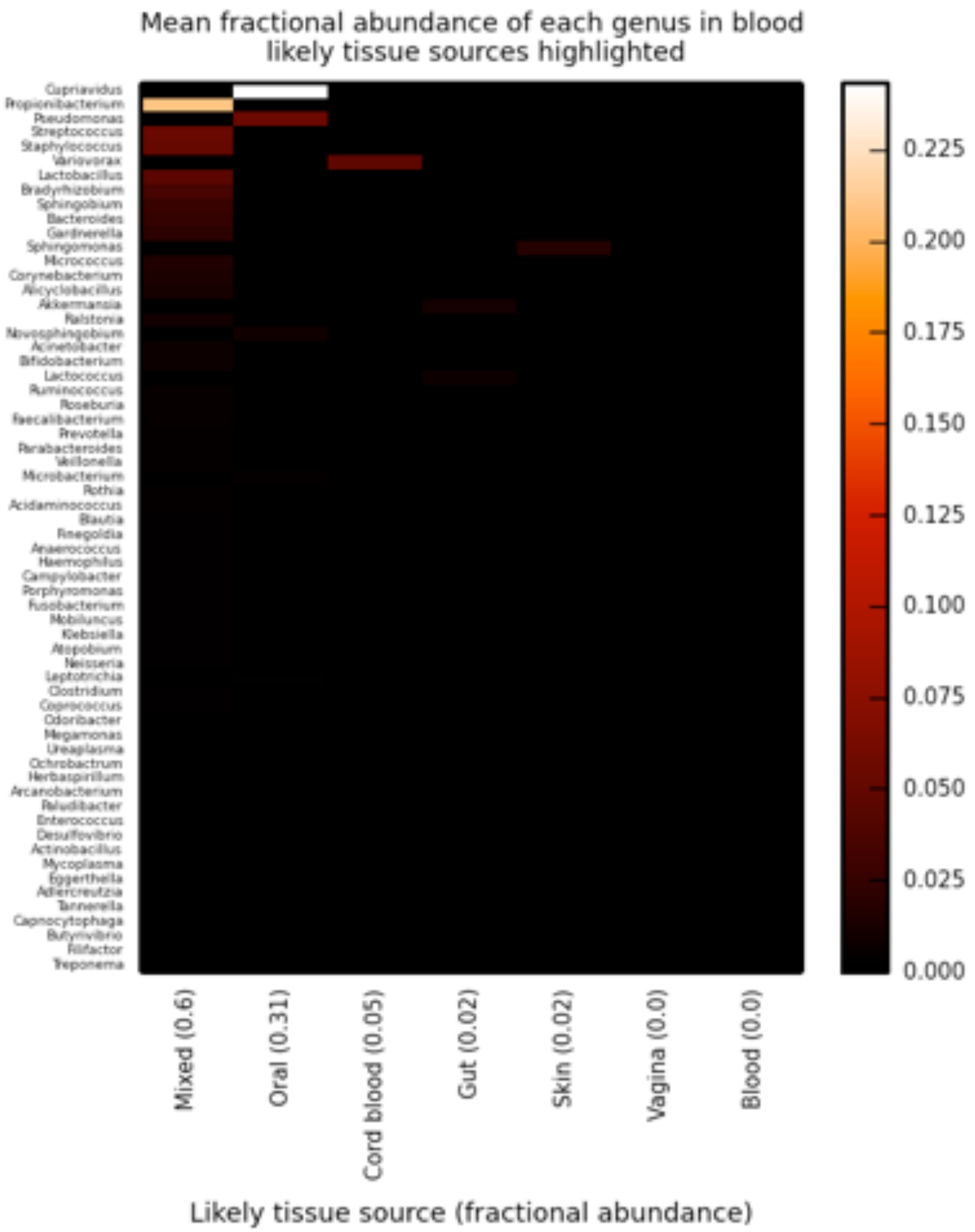


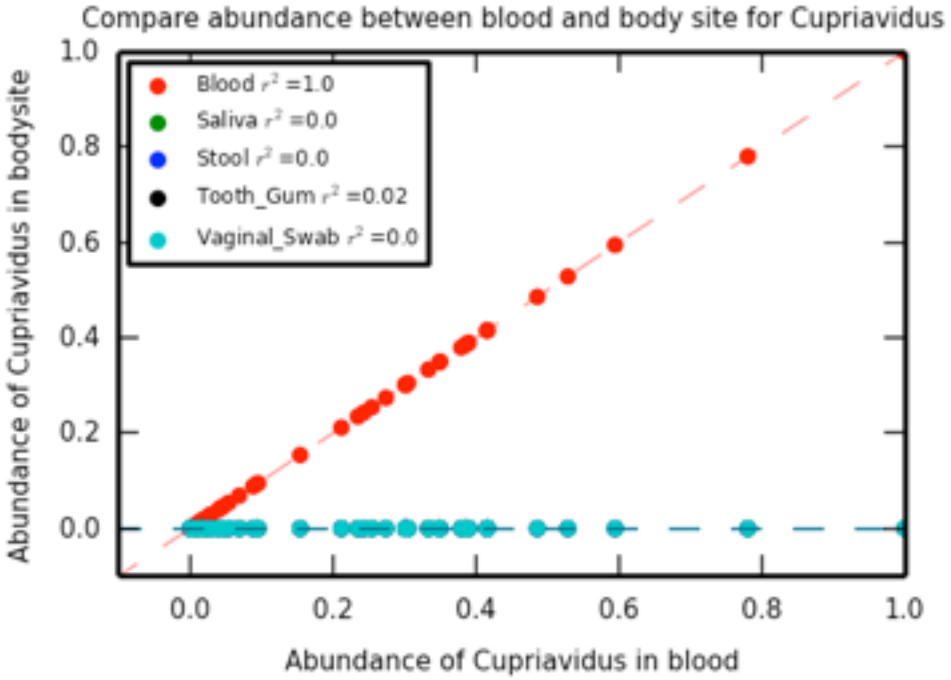

Ranked abundance of bugs in: Gut
Mean fractional abundance (log10)
1
0
−1
−2
−3
−4
−5
−6



Specific to organ

Abundant bugs in blood have mixed sources or trace abundance in specific sources.


Mean fractional abundance of each genus in blood likely tissue sources highlighted


Compare abundance between blood and body site for Cupriavidus

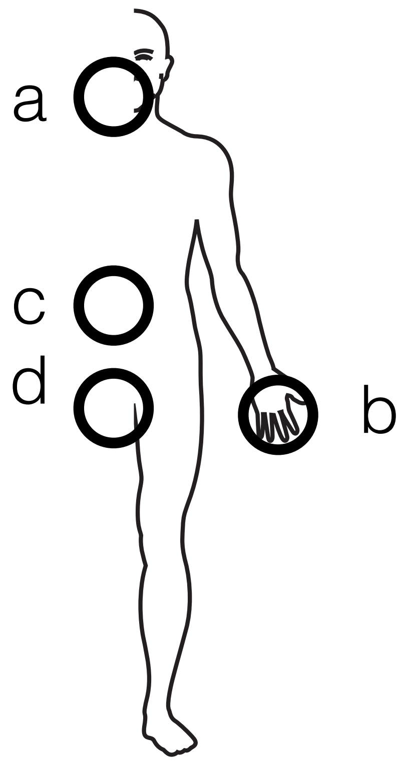
At least two possibilities:

(1) Blind sources

(2) Some bugs grow in blood (e.g., Cuprividius [1])

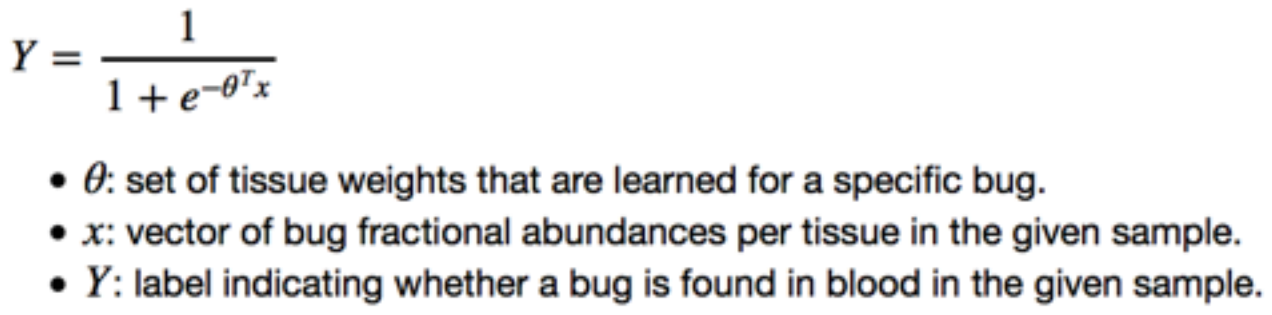
[1] Clinical Micro & Inf (2006)

Do linear models work?



🩸 = a + b+ c+ d

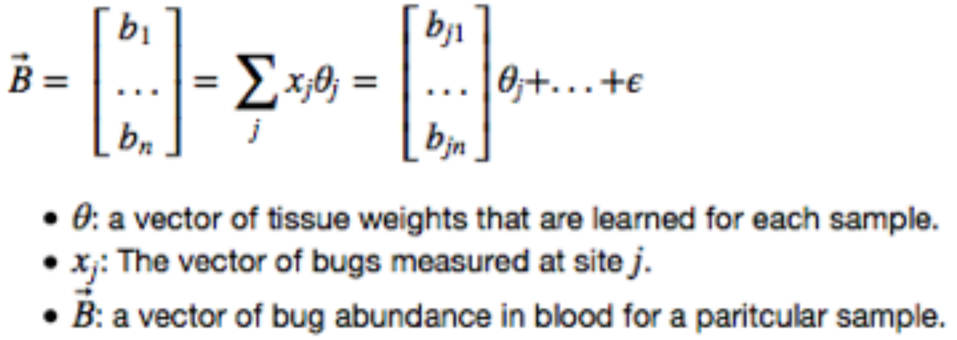
Classification -



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

- 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.

Regression -

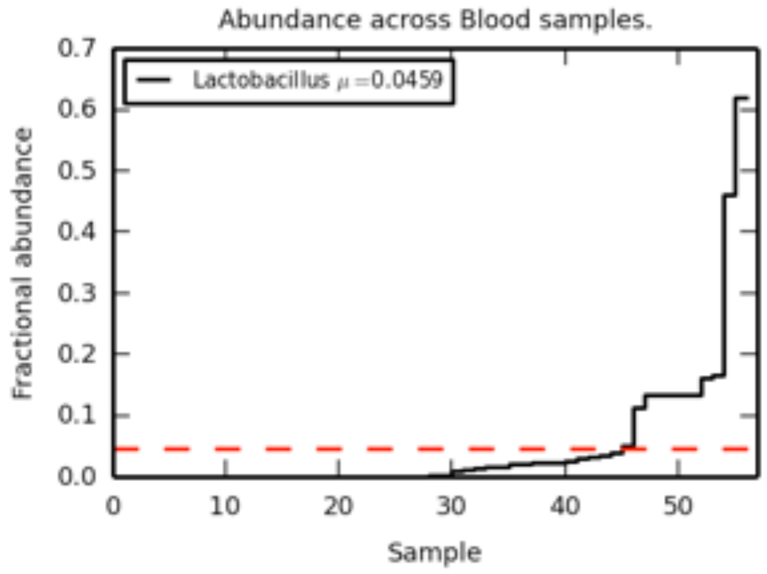


$$\vec{B} = \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix} = \sum_j x_j \theta_j = \begin{bmatrix} b_{j1} \\ \dots \\ b_{jn} \end{bmatrix} \theta_j + \dots + \epsilon$$

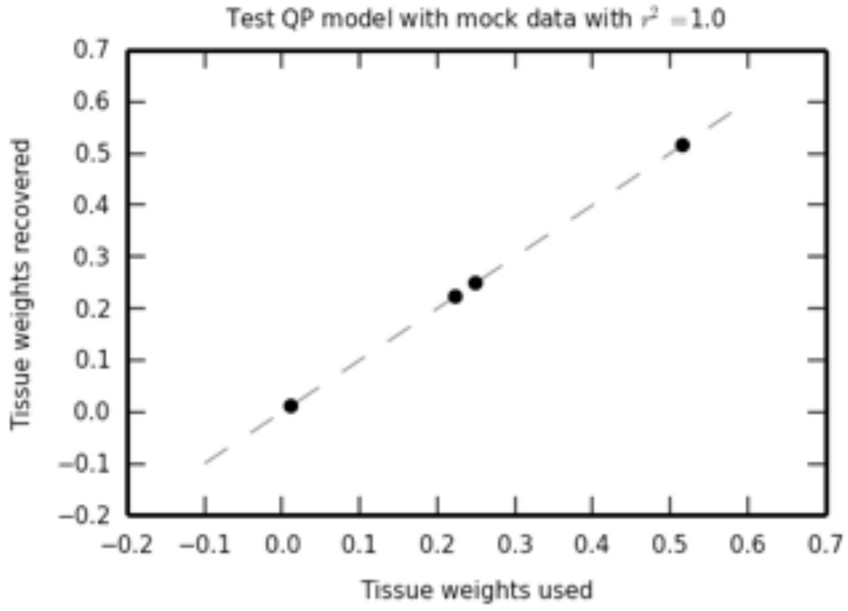
- θ: a vector of tissue weights that are learned for each sample.
- x_j: The vector of bugs measured at site j.
- $\vec{B}$: a vector of bug abundance in blood for a paritcular sample.

Poor performance of linear models.

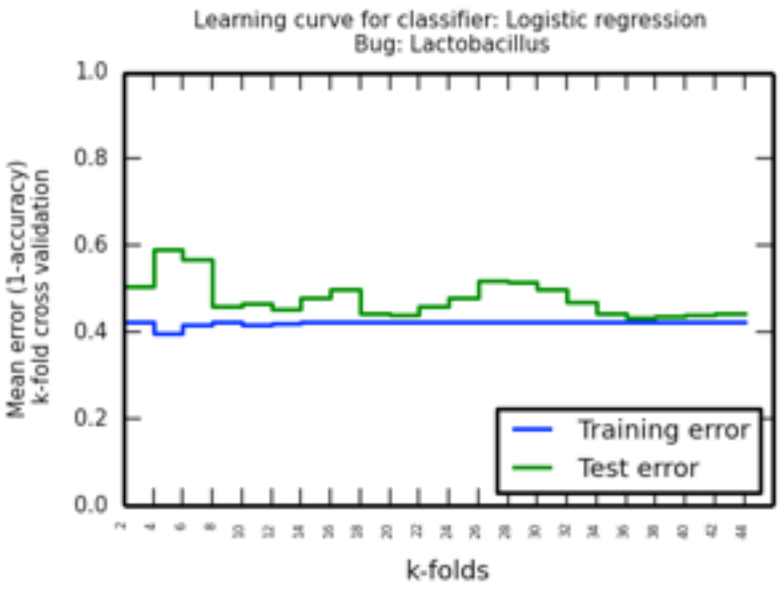
Good classification target bug (present in ~ 50% of blood samples) -



Linear regression model (see appendix) works on “test” case -

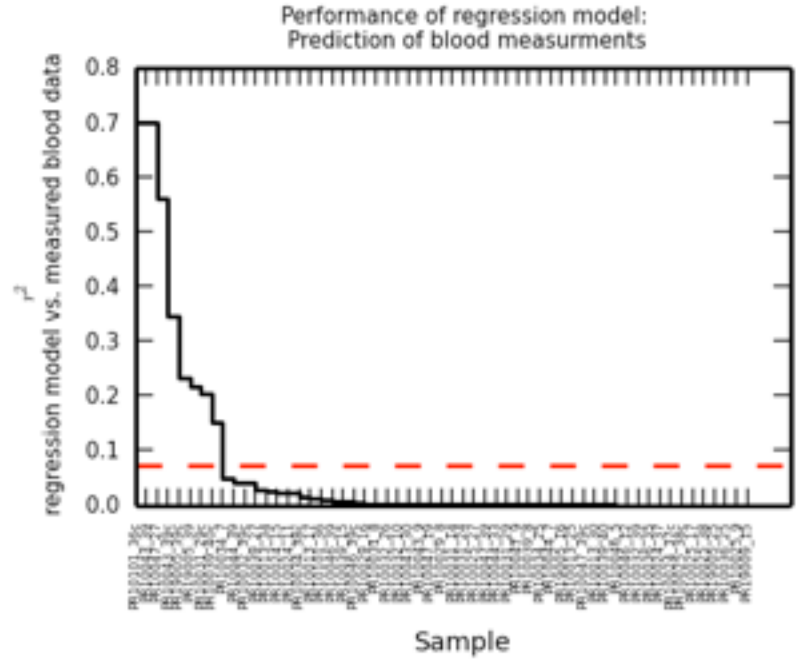


Poor results on cohort. High bias. Model underfits data. -



Intuition: Likely “complex” mixing in blood, such that load in blood is not coupled to load in sites (see appendix).

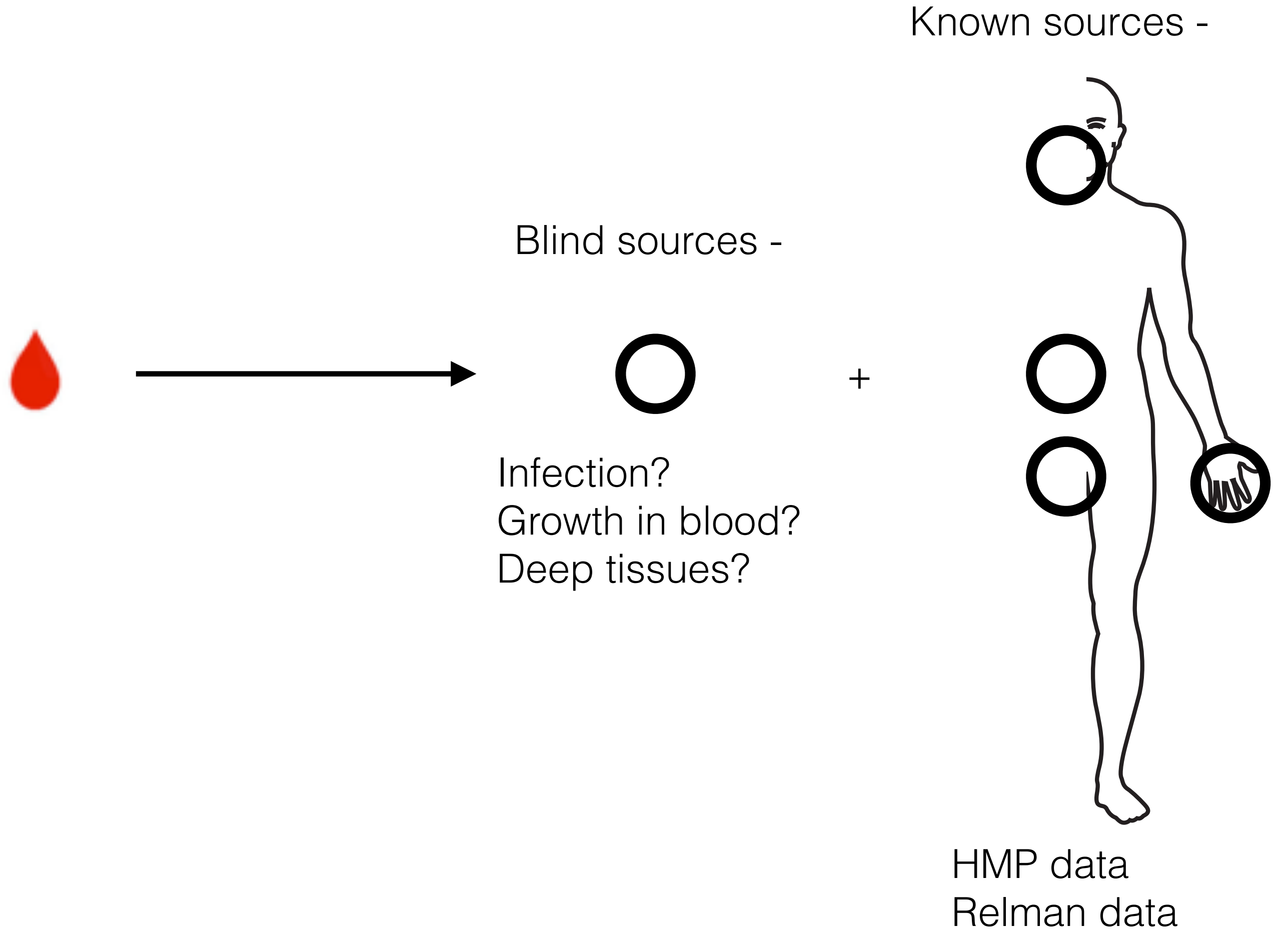
Poor results on cohort. Residuals indicate blind sources. -



Intuition: Unsampled sources frustrate model. Residuals suggest this (see appendix).

More complex learning models could probably boost performance, but will be very hard to understand results.

Other approach. Start from blood and try to de-compose it.



Known sources -

Blind sources -

+

Infection?
Growth in blood?
Deep tissues?

HMP data
Relman data

Allows for “blind” sources to be included in model:
- Prior approach constrained possible sources to sampled sites.
- Intuition and evidence suggests that blood can sample from more than just these few.

Better suited for “complex” mixing model:
- This allows for sample-specific mixing of the sources.
- Prior approach learned common body-site mixing coefficients for the entire cohort.

ICA applied to microbial de-composition.

Given blood mixtures:

$\text{Sample}_{1\text{-}k}$

$\xrightarrow{\text{ICA}}$

(1) Computes sources -

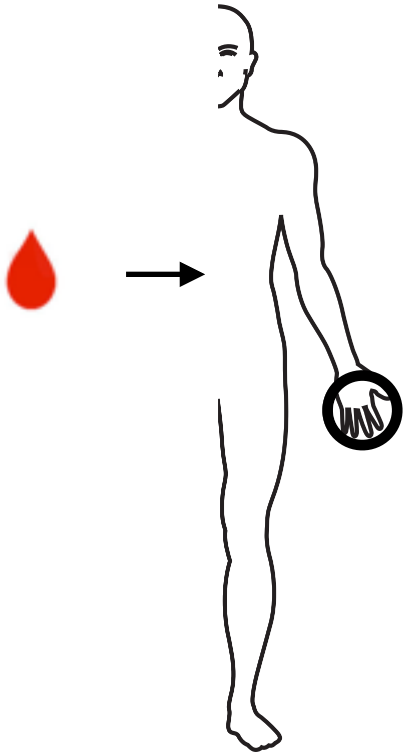
S_1, S_2

(2) Computes mixing matrix, A -

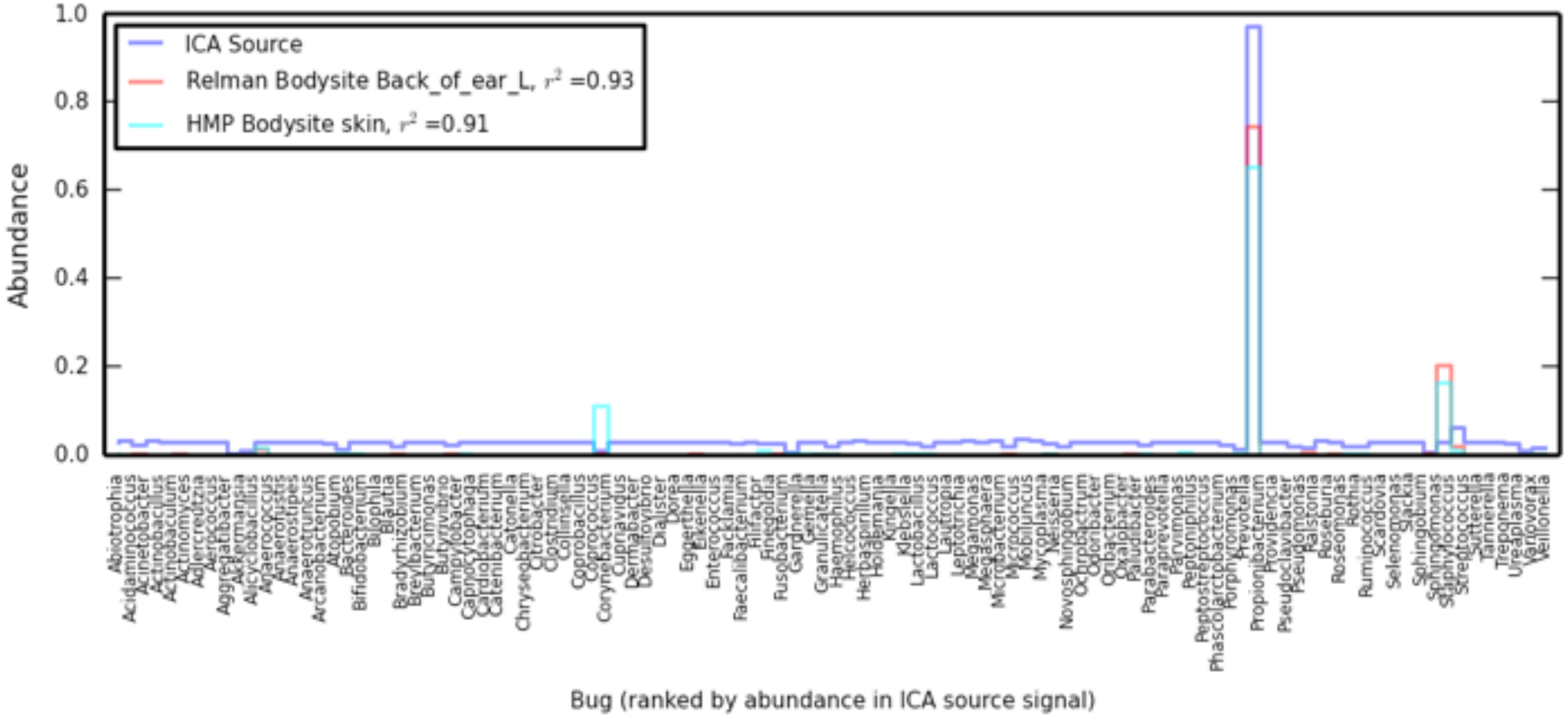
$\text{Guess}_K = A_{K1} * S_1 + A_{K2} * S_2$ ...

Source analysis: Ability to correlate ICA sources with known body site data.

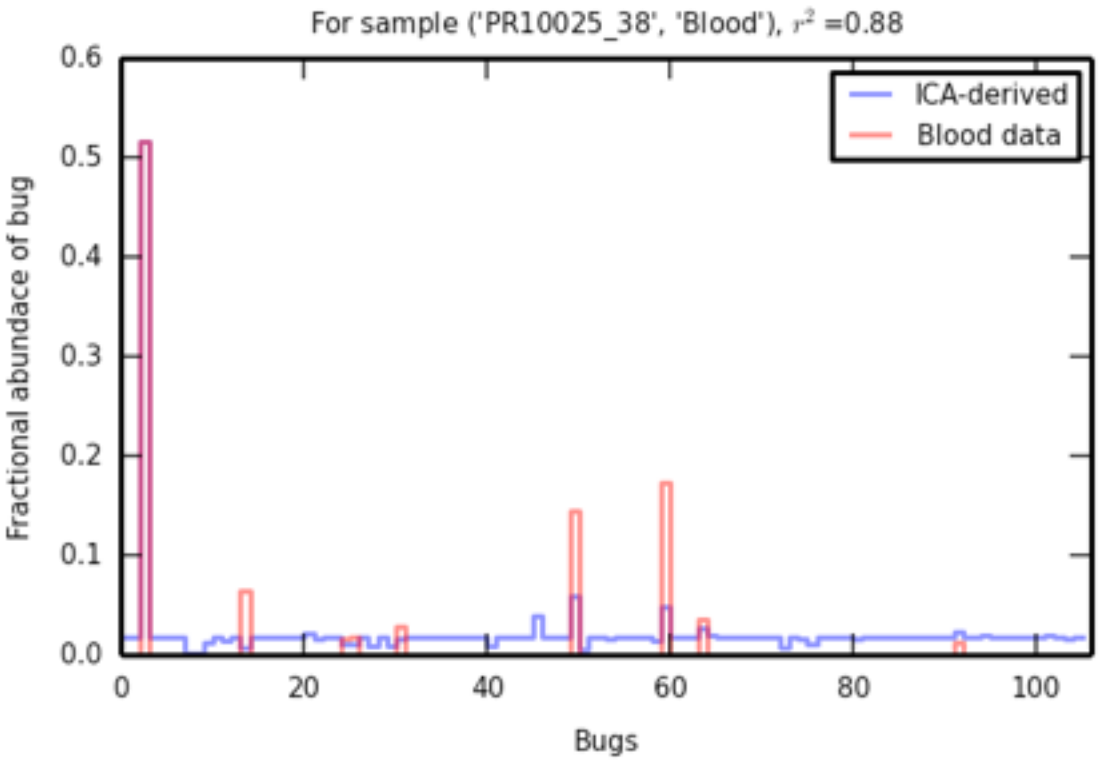
Known sources -

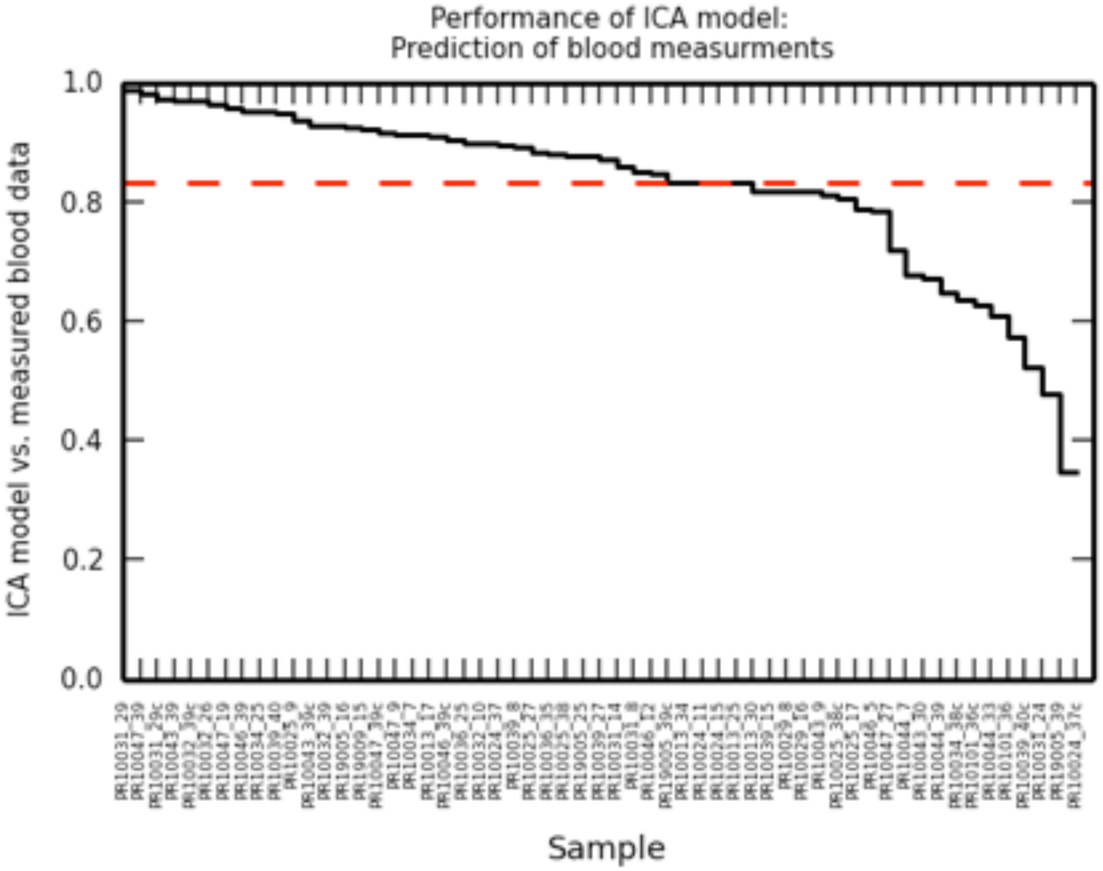


Correlate ICA source signal with known data about microbial composition of body site.
(See appendix for results on all sources and corresponding body site assignments.)

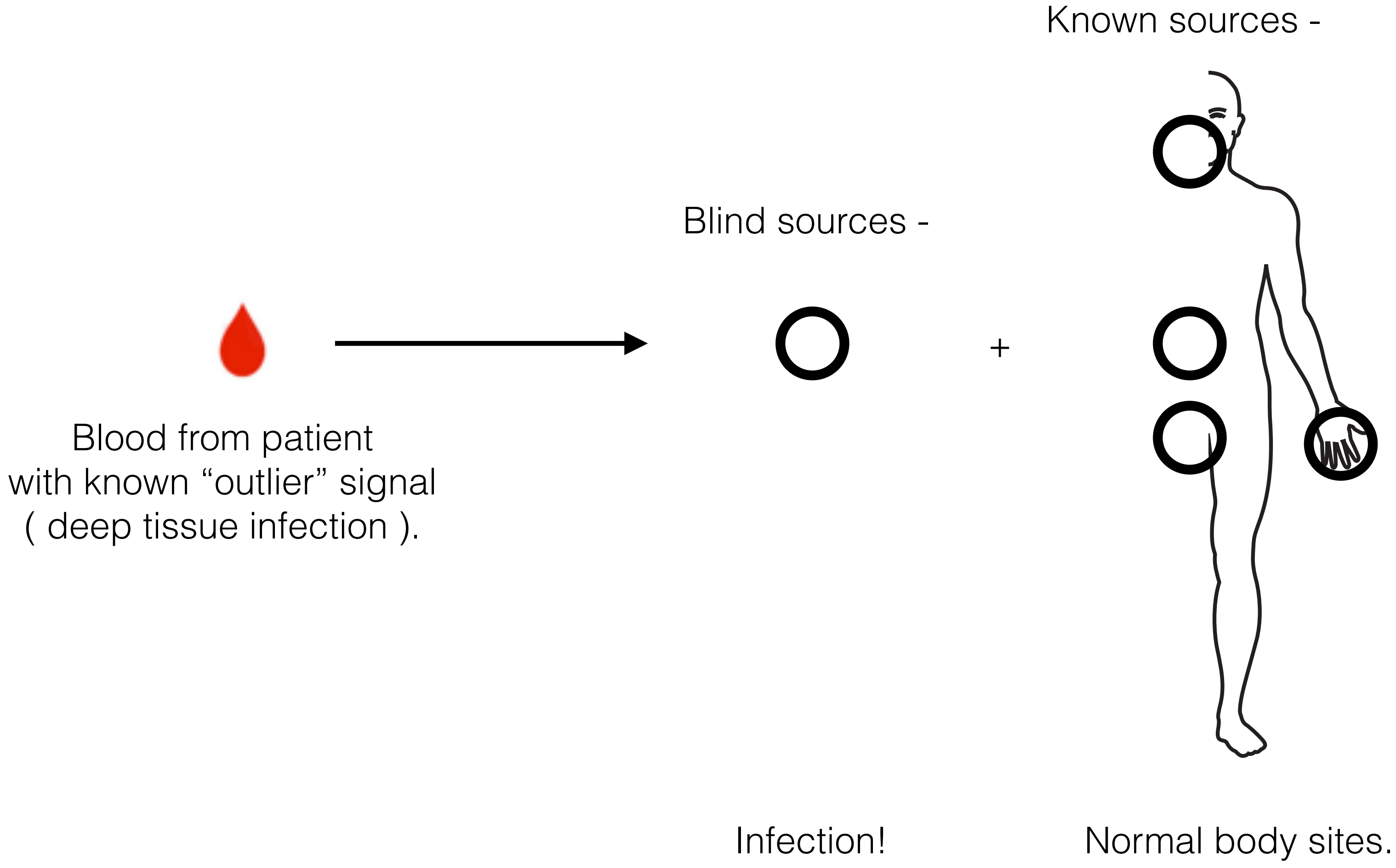


Performance evaluation: Reasonable.





If this works, it would be able to detect anomalies.



Known sources -

Blind sources -

Blood from patient
with known “outlier” signal
(deep tissue infection).

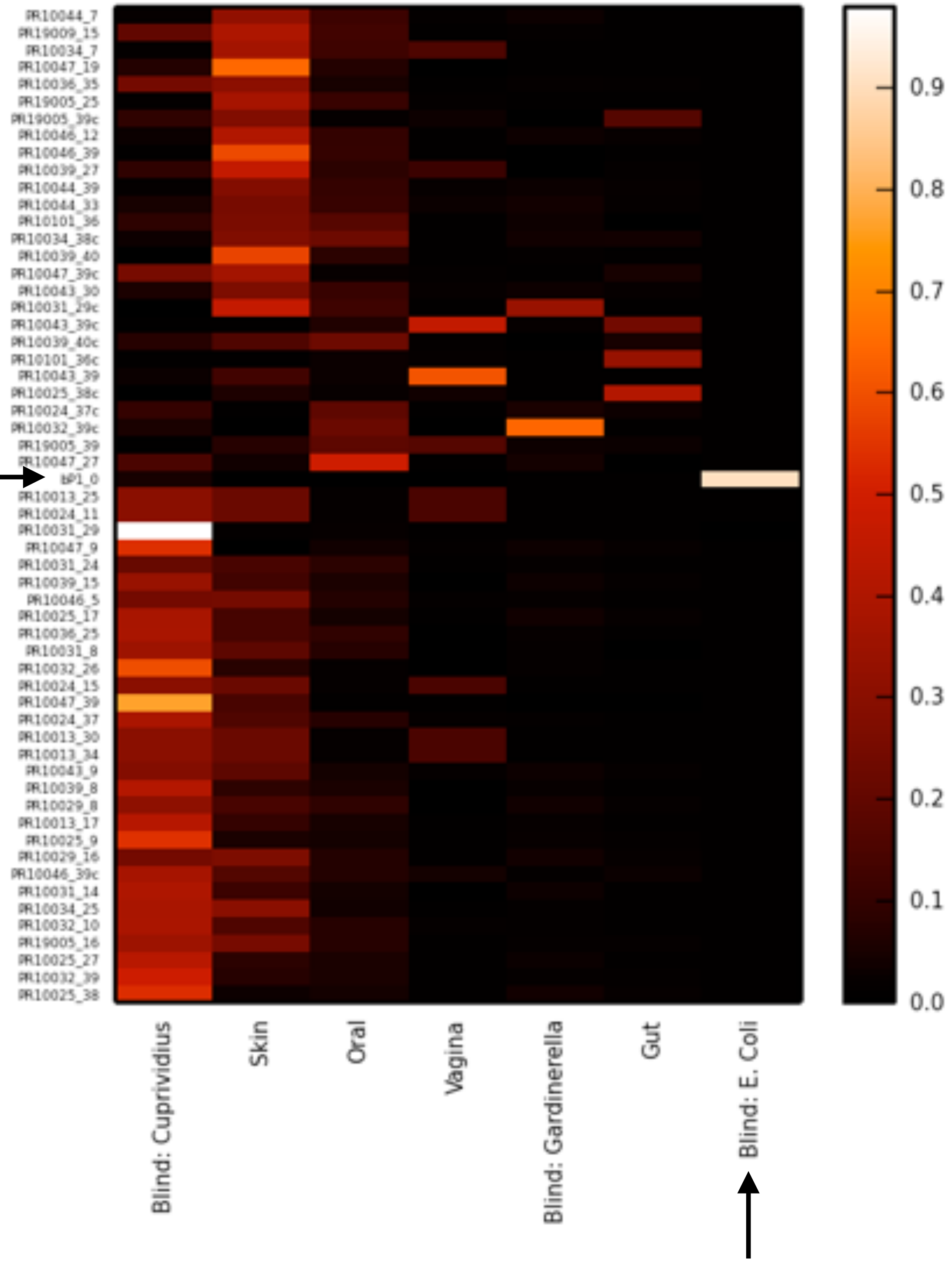
+

Infection!

Normal body sites.

Spike this outlier into the cohort of otherwise healthy pregnant samples.

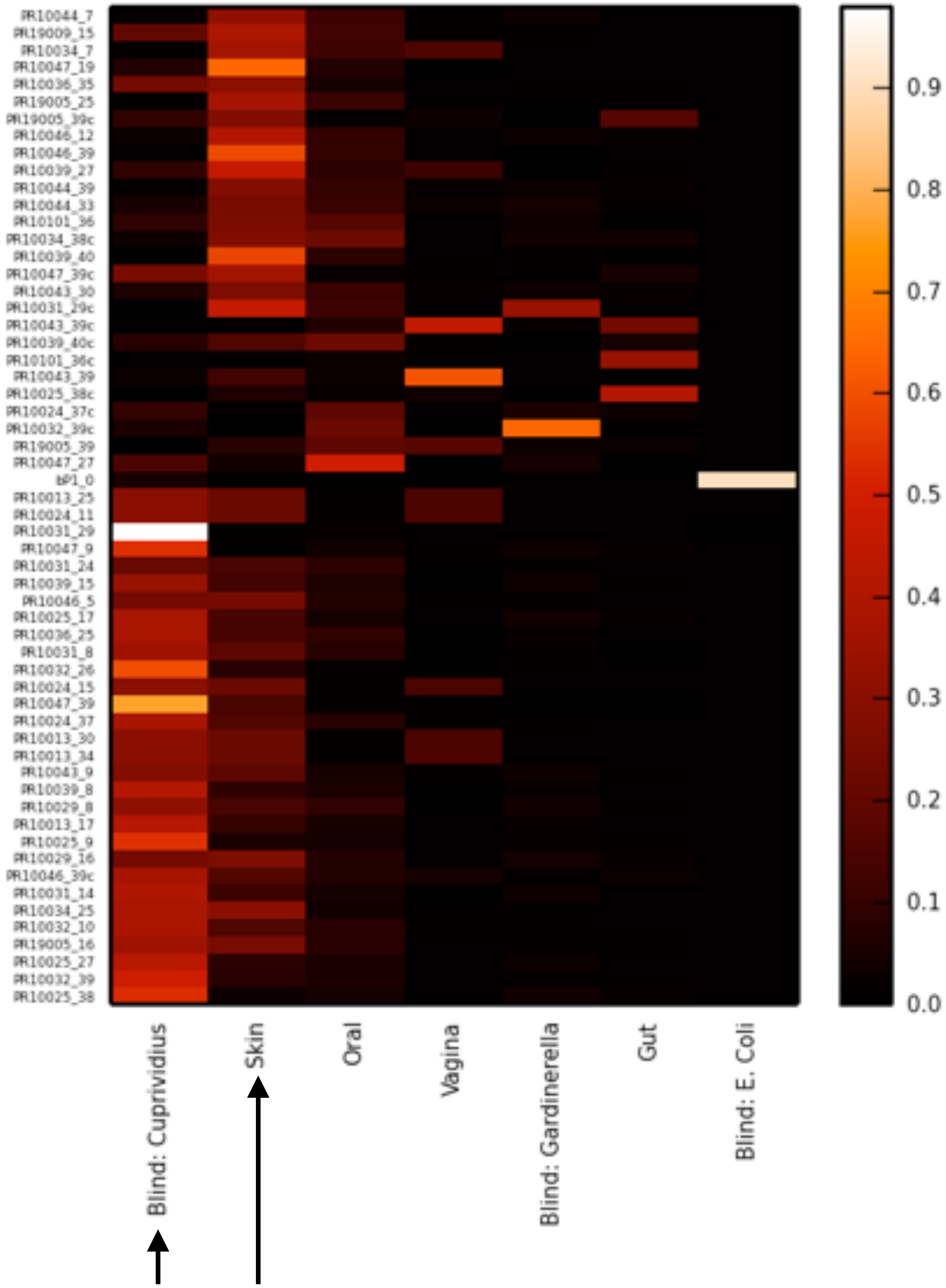
Clustered mixing matrix shows un-supervised de-convolution.



For patient with deep tissue E. Coli. infection …

.. strong mixing of a “blind” source enriched in E. Coli that ICA identified.

Analysis of pregnancy samples.



- Many 28 clustered samples enriched in blind source (Cuprividius) signal and 20 with Propionibacteria enriched (skin-like) signal.
- No significant patient segmentation between these two clusters (e,g., by pre-term status, patient, time)

Summary

Context -
- Cohort of healthy patients with 58 sequenced blood samples for microbiome.
- Matched microbiome sequencing of 4 body sites (oral, vagina, skin, gut).

Bit vector analysis -
- Some highly abundant bugs in blood (e.g., Cuprividius, Pseudomonas) have trace composition at sampled sites.
- Suggested unsampled sources or possible growth in blood.

PCA (appendix) -
- Cuprividius and Propionobacteria loads explain much of the variance.

Linear learning models -
- Classification on a single bug (e.g., Lactobacillus) with even labeling (50% of blood samples) has poor performance.
- High bias, indicating linear model is under fitting pattern in the data. Suggests complex mixing.
- Constrained regression (appendix for model setup) also has poor performance on samples.
- Residuals point to blind sources or blood-induced growth
- Bugs present at high load in blood for some samples (e.g., Cuprividius) are not found or are trace at sampled sites.

ICA and clustering -
- ICA allows for blind source inclusion in the model.
- Resulting sources can be assigned to issues based upon existing (e..g, HMP and our cohort) data.
- Good performance.
- Identifies the “blind” sources as well as tissue-specific sources (appendix).
- Correctly identifies anomalous E. Coli source in patient with known deep tissue infection.
- Clustered mixing matrix show two primary sources in pregnancy cohort, though no clear cluster segmentation of patients.

Results:
- Training on larger sets of healthy patients can further validate the expected “healthy” sources.
- Deeper sampling may explain current “blind sources.”
- Un-supervised detection of anomalies (e.g., infections, colonization, intestinal breakage) may be possible.