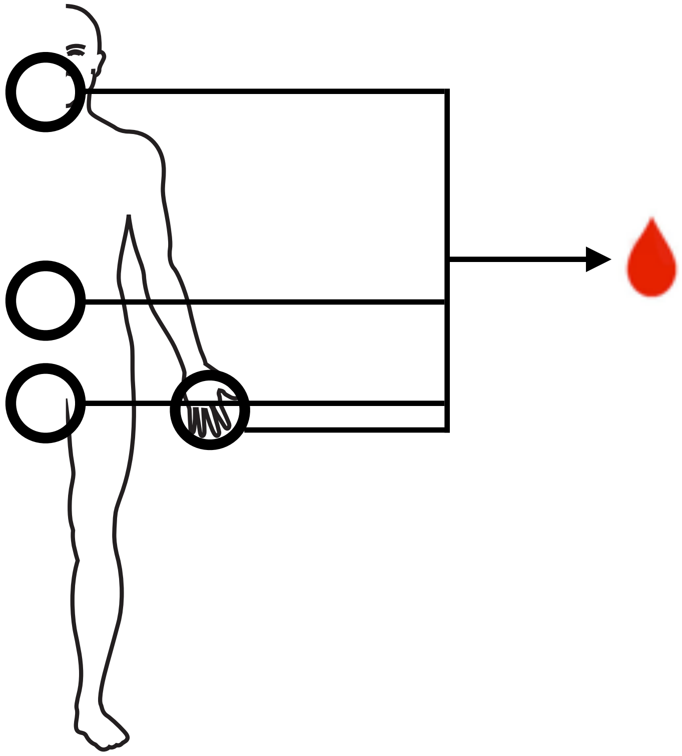
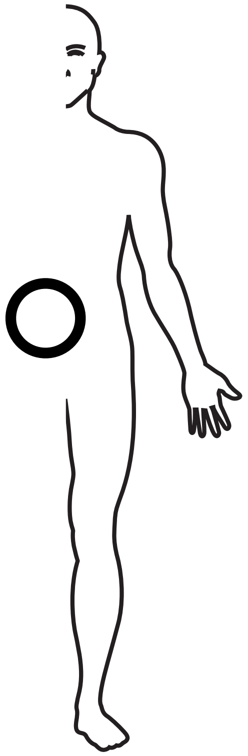
Unsupervised microbial source detection from human blood

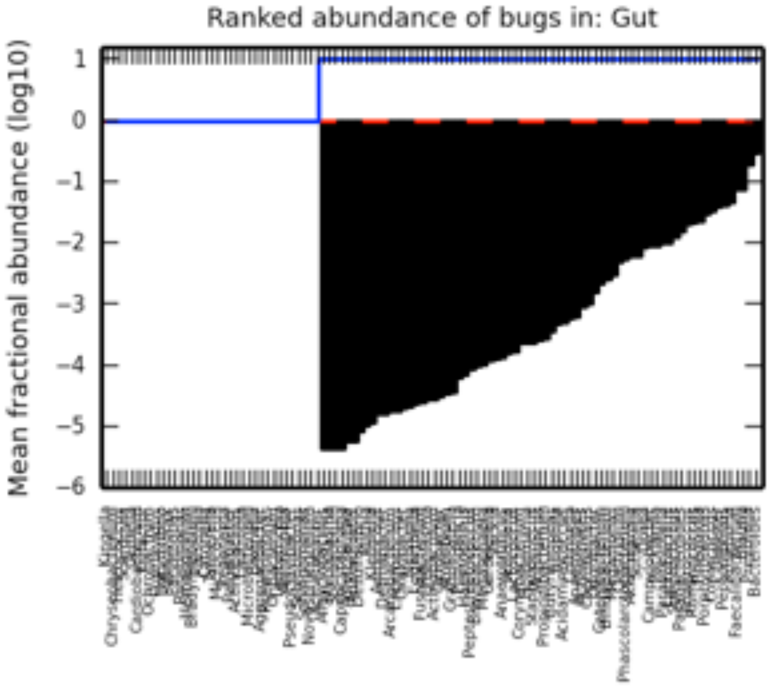
Examine microbiome composition of blood relative to body sites.

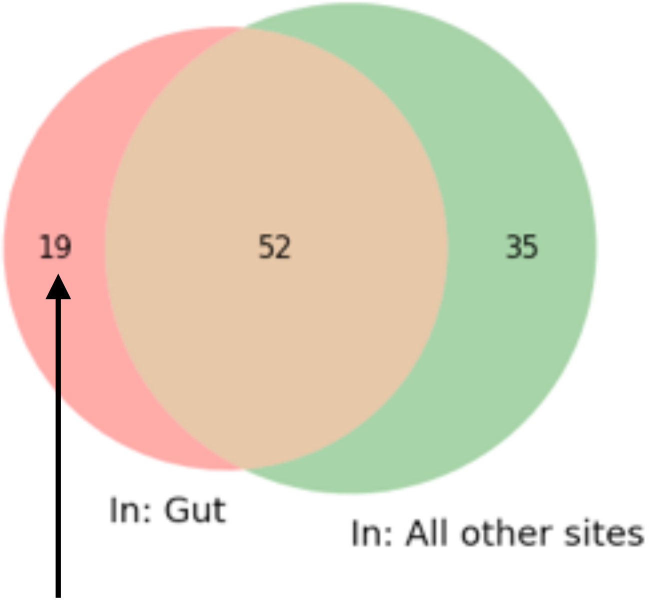


58 blood samples with matched body sites (oral, vagina, skin, gut) from healthy pregnancy cohort.

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



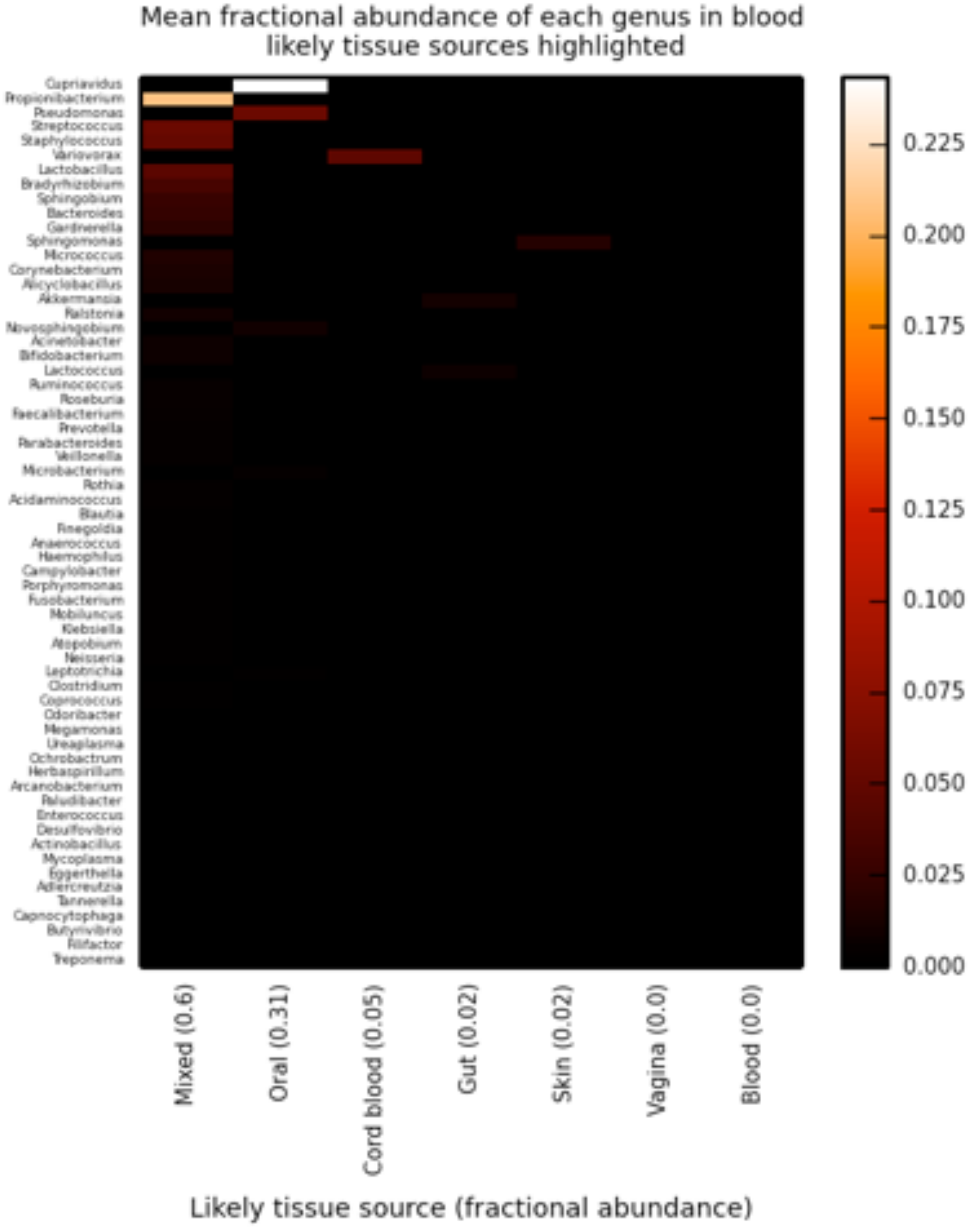

Ranked abundance of bugs in: Gut
Mean fractional abundance (log10)
1
0
−1
−2
−3
−4
−5
−6

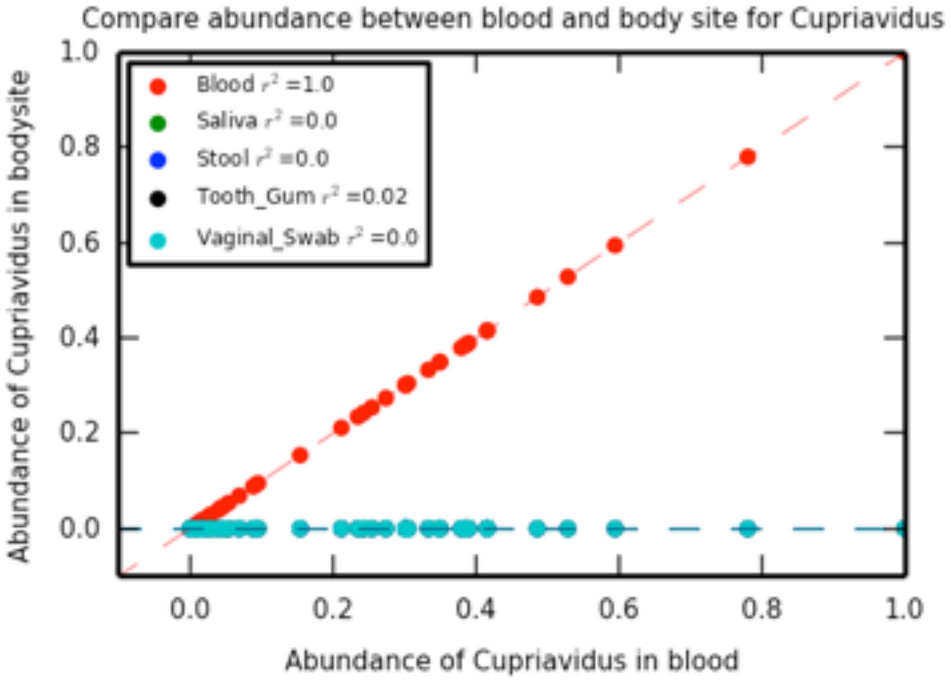


Specific to organ

By identifying bugs specific to each tissue, we can then ask whether any tissue specific bugs are found in blood to infer tissue exchange.

Bugs in blood have mixed tissue sources or trace abundance in specific (Oral) sources.

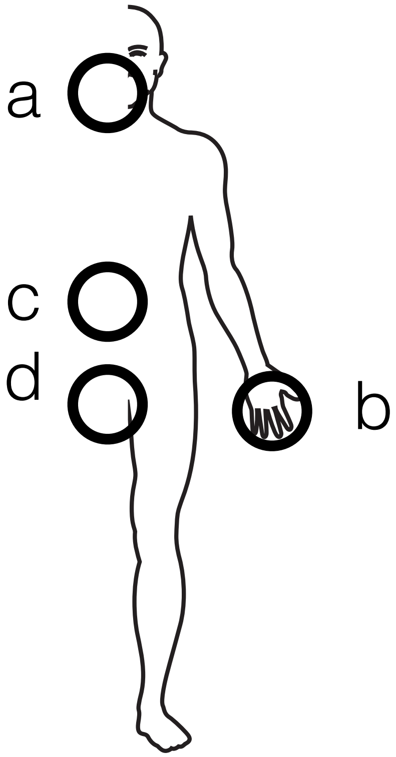




The most abundance bugs in blood are either from mixed sources or have very trace abundance in the specific source tissue (Oral). This suggests that either we have (1) under sampled the possible set of sources or (2) some bugs grow in blood (e.g., Cupridivius [1])

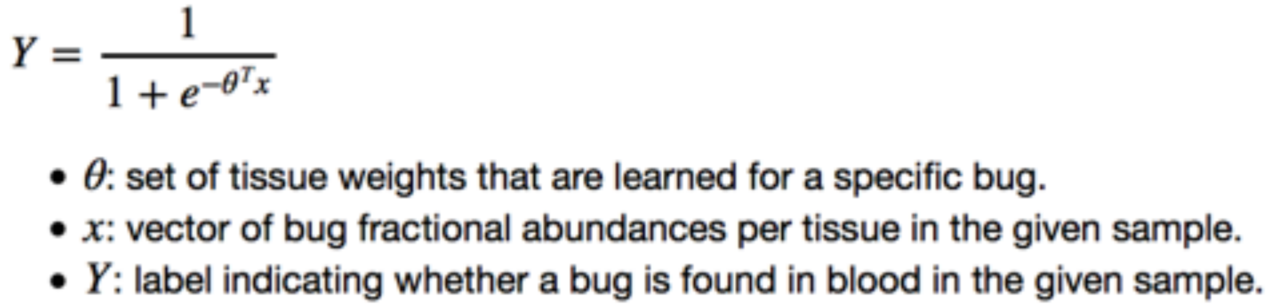
[1] Clinical Micro & Inf (2006)

Do linear models work?



💧 = a + b+ c+ d

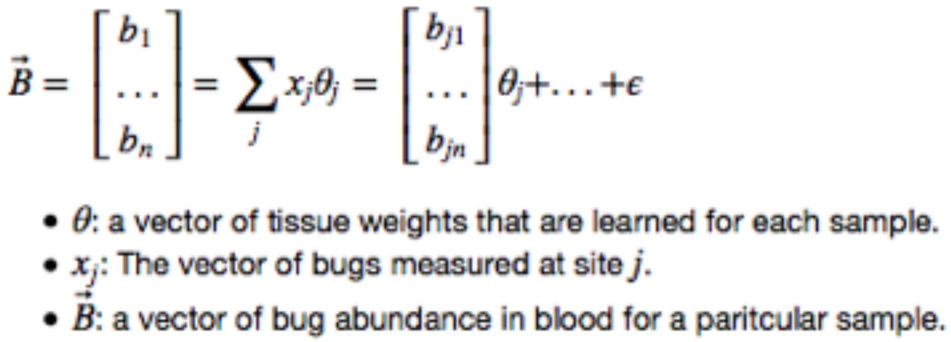
Classification -



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

- 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 -



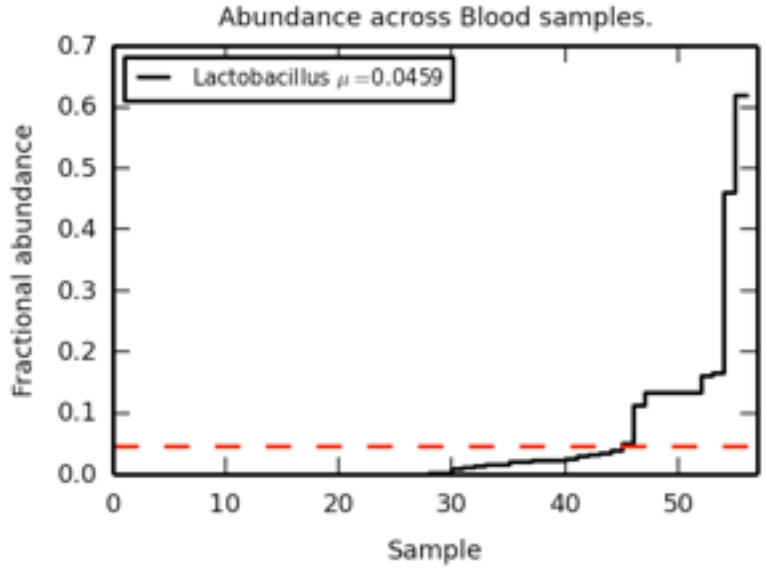
$$\vec{B} = \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix} = \sum_j x_j \theta_j = \begin{bmatrix} b_{j1} \\ \dots \\ b_{jn} \end{bmatrix} \theta_j + \dots + \epsilon$$

- 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.

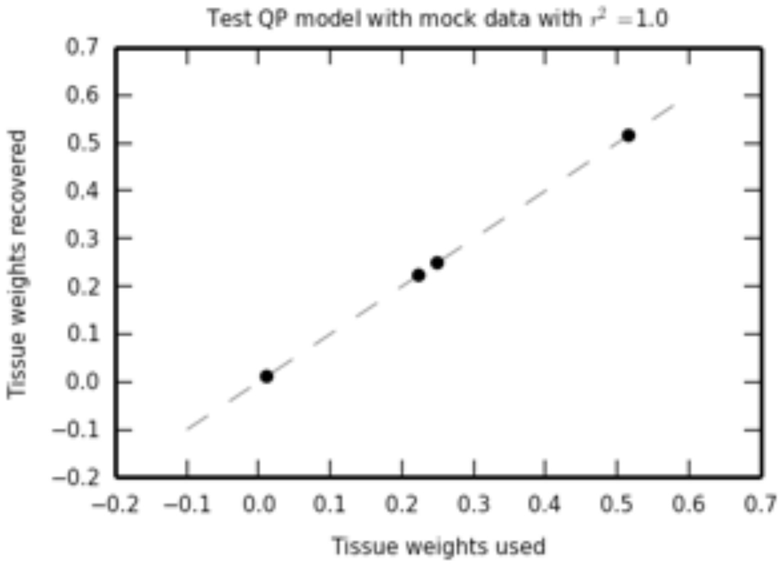
Naive approach is to ask whether linear model can be applied to the sampled body site in order to explain the composition of bug in blood. (1) for any specific bug, we discretize its presence in blood in a given sample and have an associated set of measurements for the body sites. In turn, we can apply classification to learn parameters that weight the contribution of the given bug from the sampled sources. (2) for any sample, we also have a vector of bugs measured in both blood and the tissues. We can try to learn parameters that weigh each tissue. In both cases, we obtain a set of parameters that explain the observed abundance of a specific bug (1) or all measured bugs (2) in blood with respect to sampled body sites.

Poor performance of linear models.

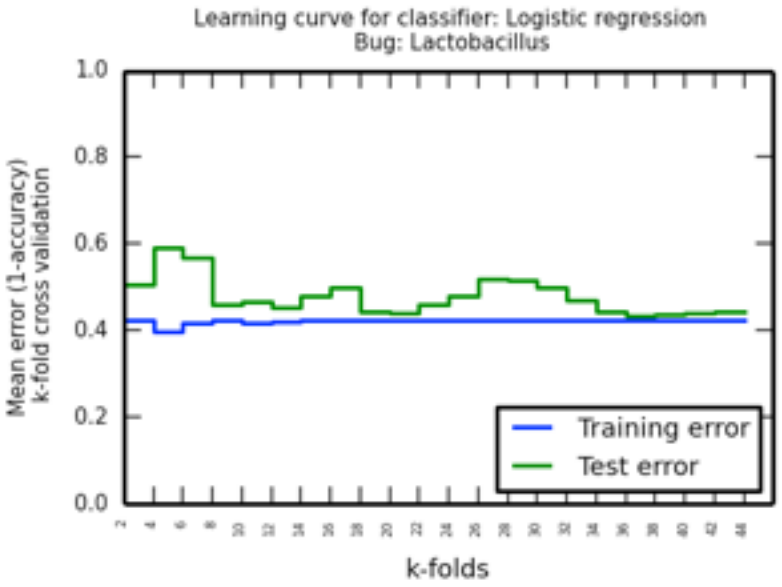
Good classification target bug (present in ~ 50% of blood samples) -



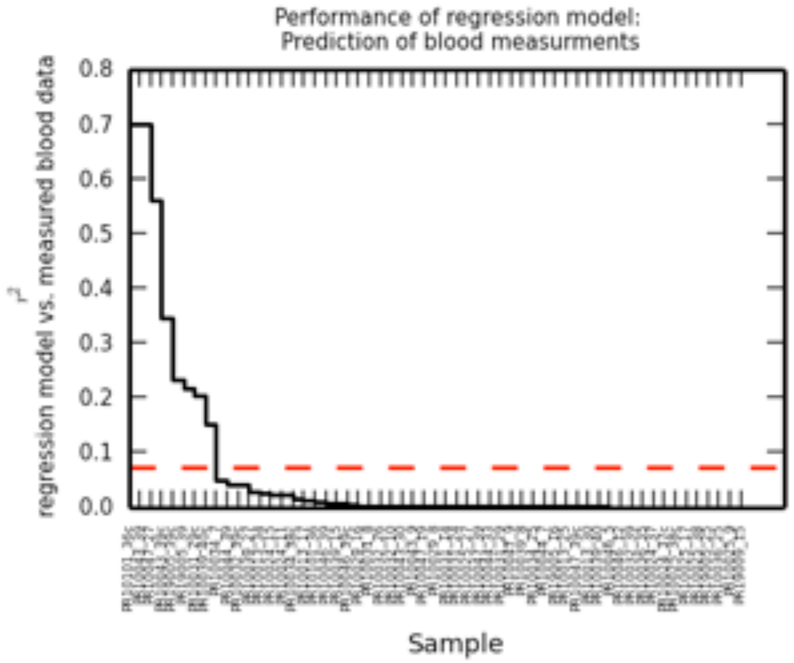
Linear regression model (see appendix) works on “test” case -



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



Poor results on cohort. Residuals indicate blind sources. -

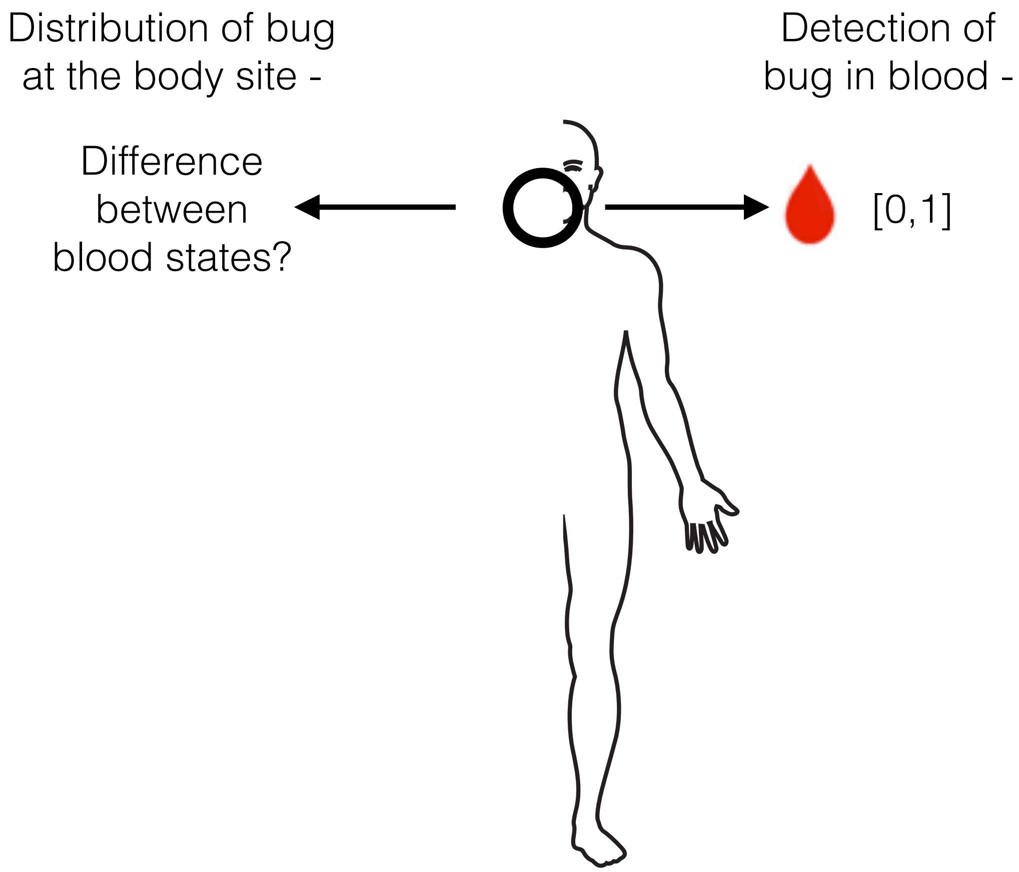


Intuition: Likely “complex” mixing in blood, such that load in blood is not coupled to load in sites (see appendix).

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.

Build intuition to explain why classification will not work.



Distribution of bug
at the body site -

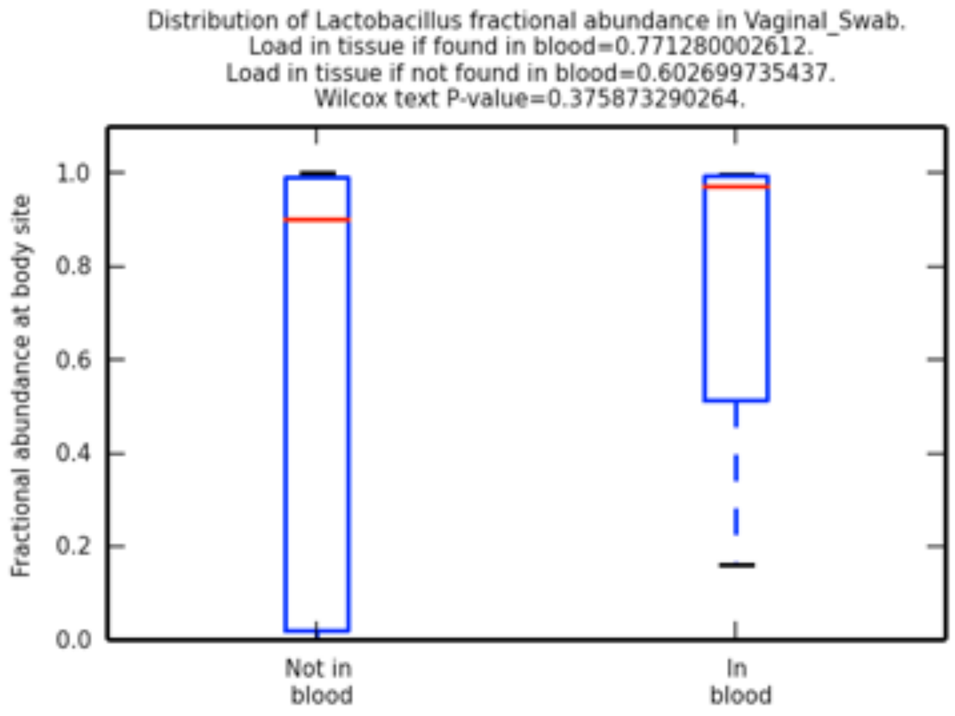
Difference
between
blood states?

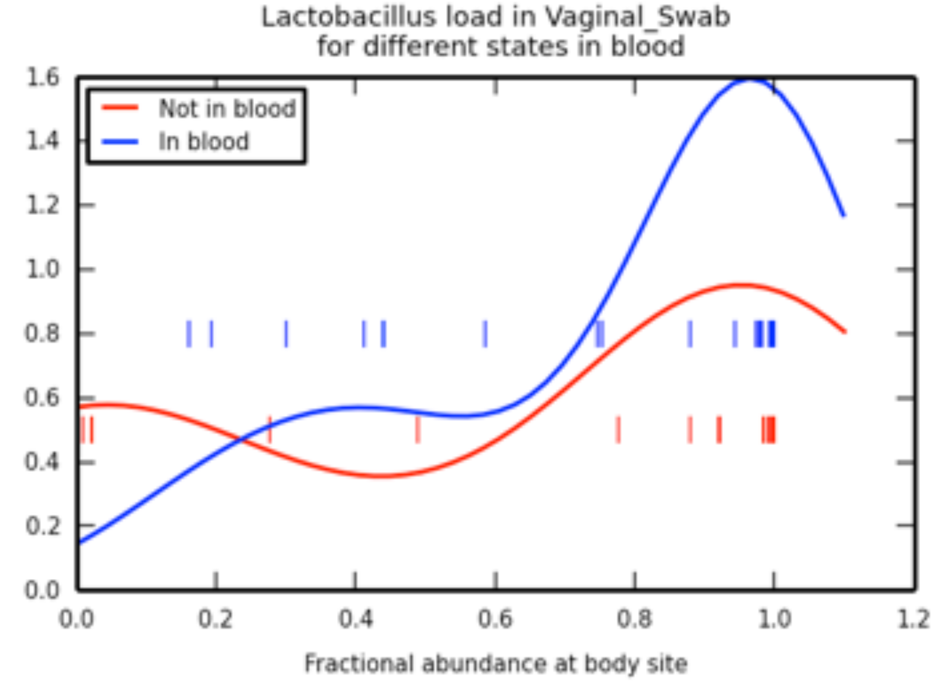
Detection of
bug in blood -

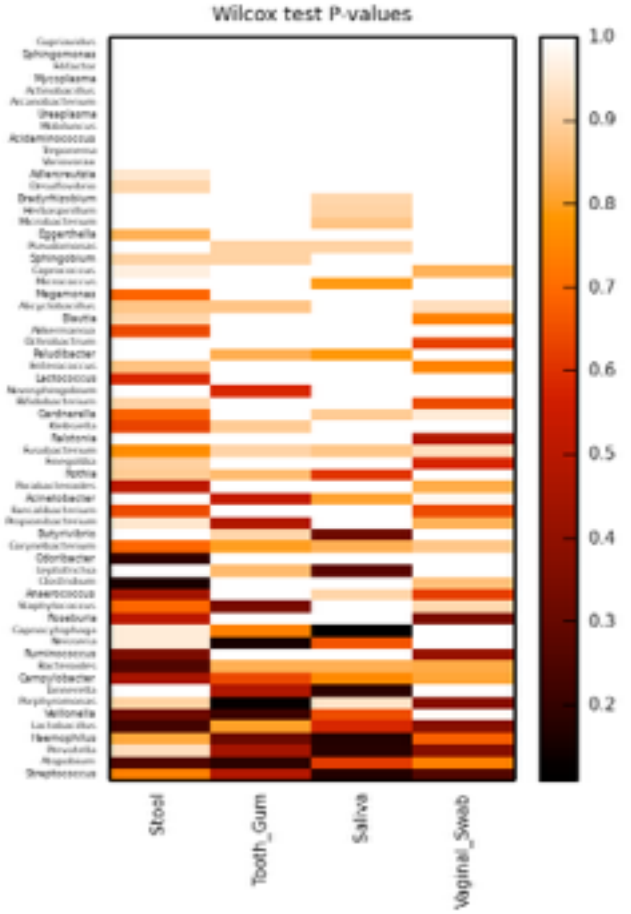
[0,1]

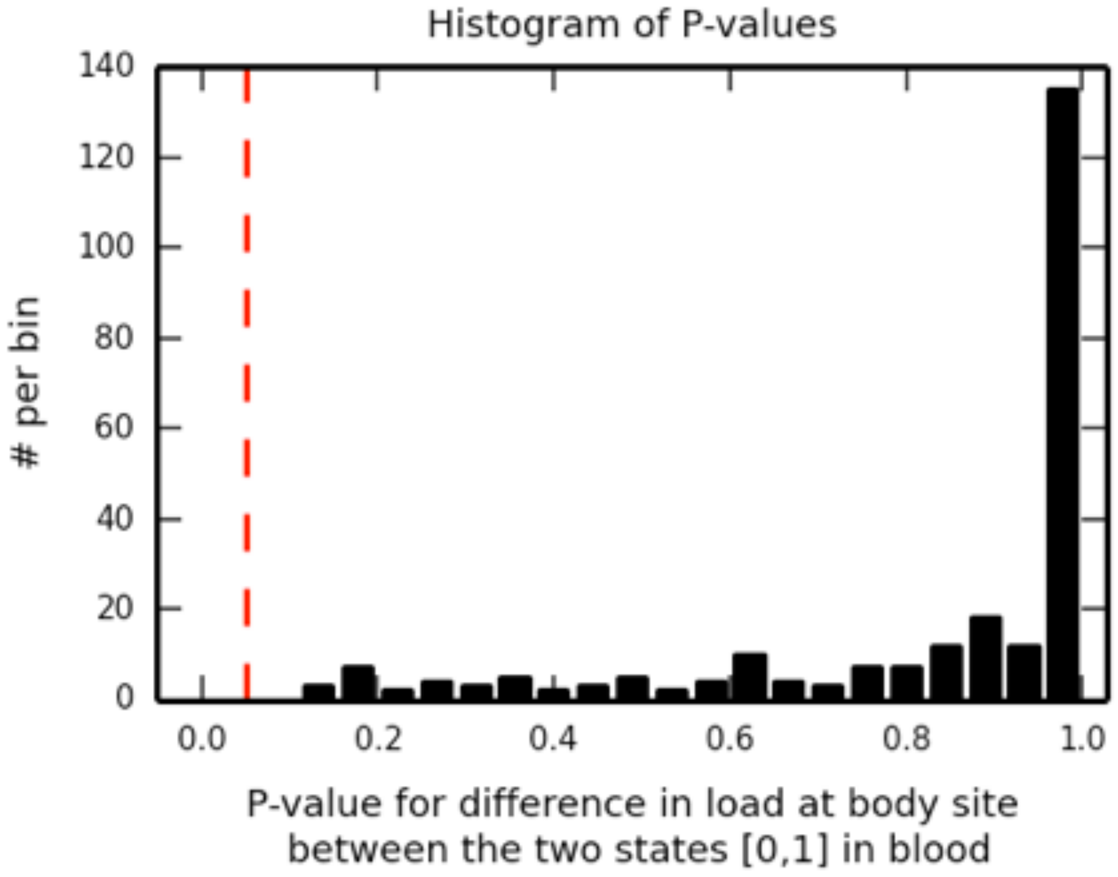
Simply evaluate the distribution of a bug at a body site with respect to its detection in blood. If there is clear coupling (that would be captured by a linear model), then we would expect to see a difference (e.g., a rise) in bug at the body site when it is detected in blood.

Evaluate abundance of bug at body sites relative to its detection in blood.



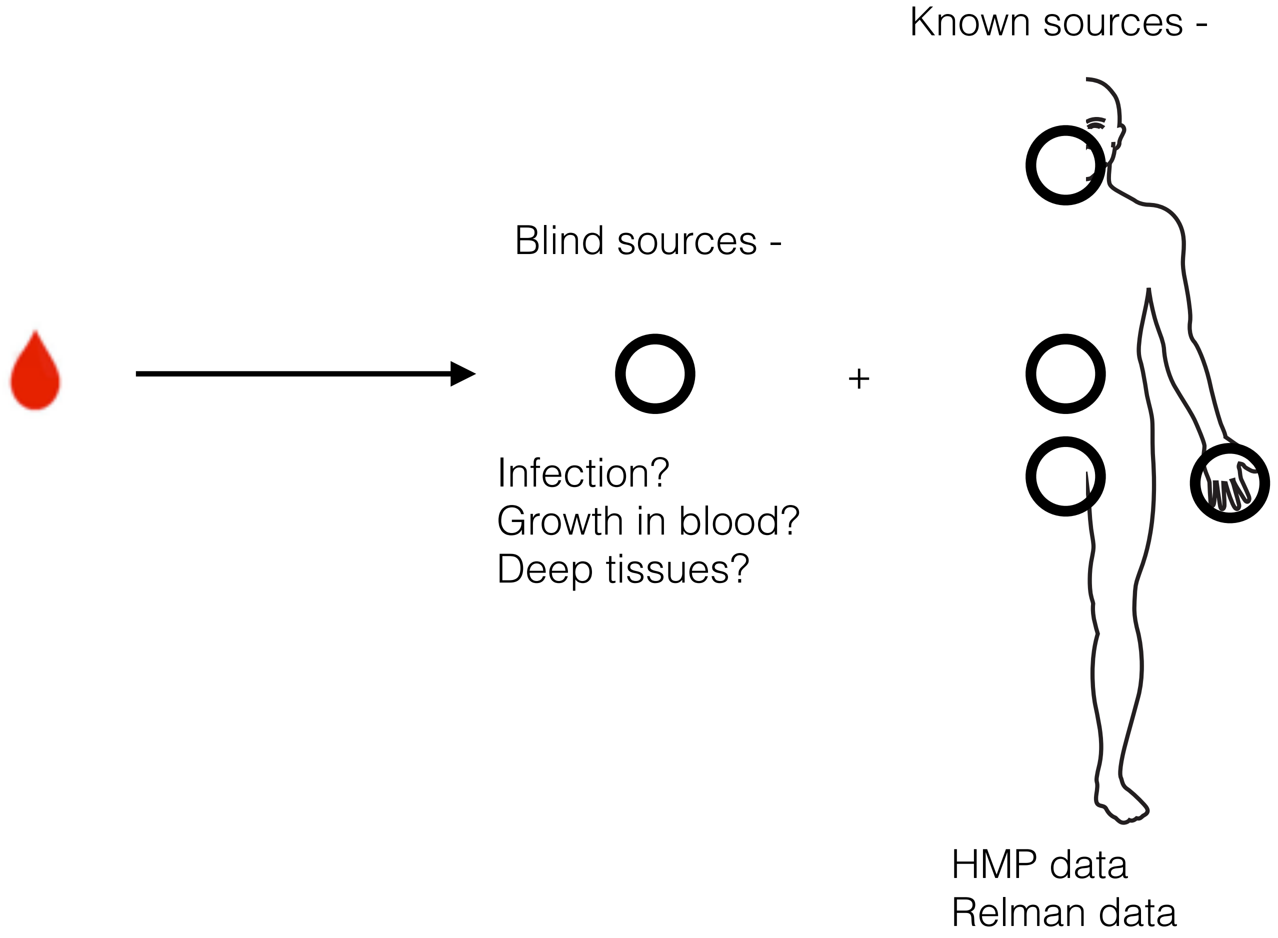






Apply Wilcox test to each bug-tissue combination relative to blood, resulting in no significant difference between body site distributions for any.

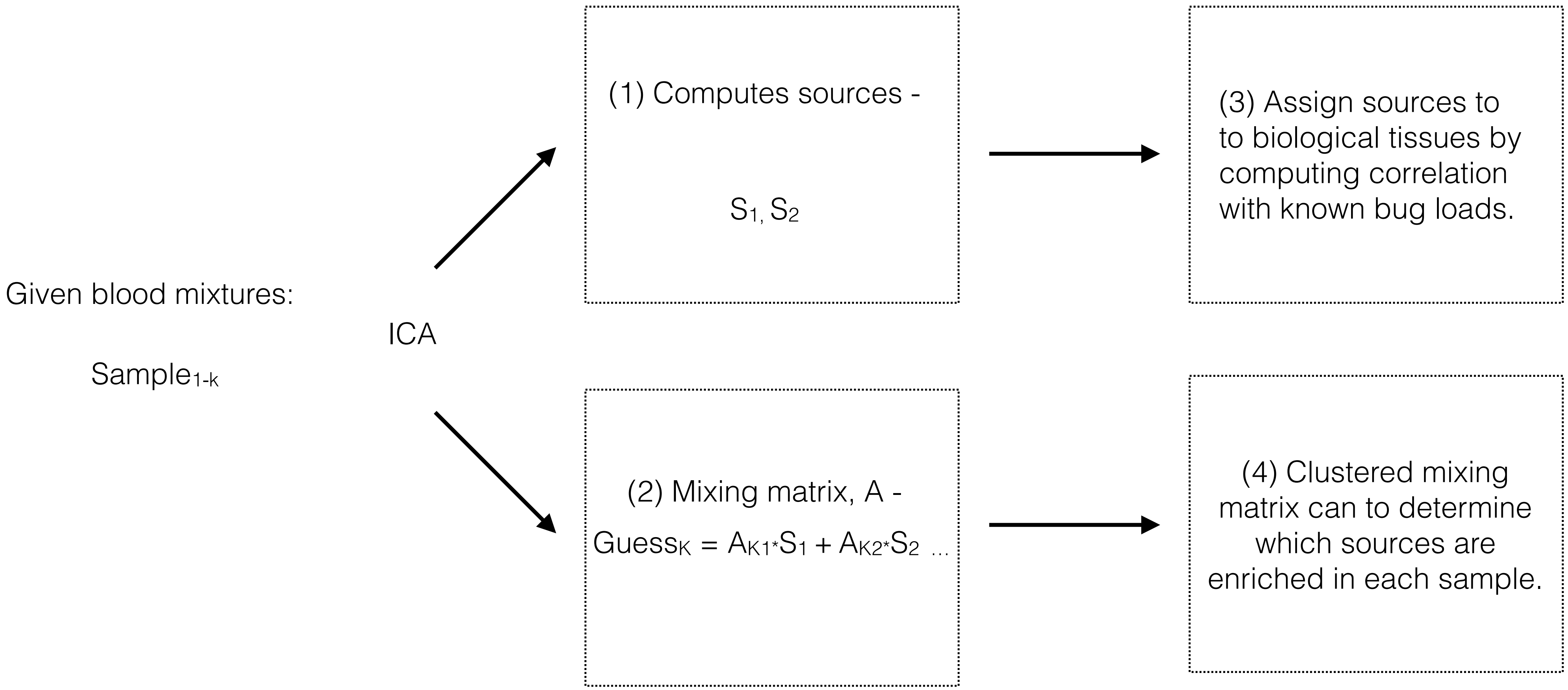
Other approach. Start from blood and try to de-compose it.



ICA 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 this blind source problem.



Given blood mixtures:

Sample$_{1-k}$

ICA

(1) Computes sources -

S_1, S_2

(2) Mixing matrix, A -

$Guess_K = A_{K1}*S_1 + A_{K2}*S_2$ ...

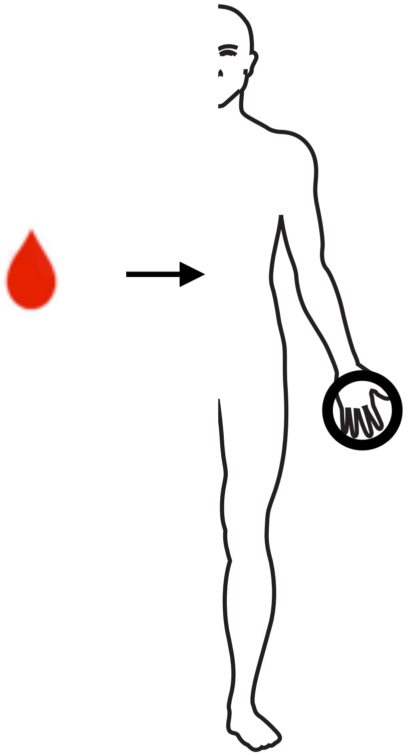
(3) Assign sources to to biological tissues by computing correlation with known bug loads.

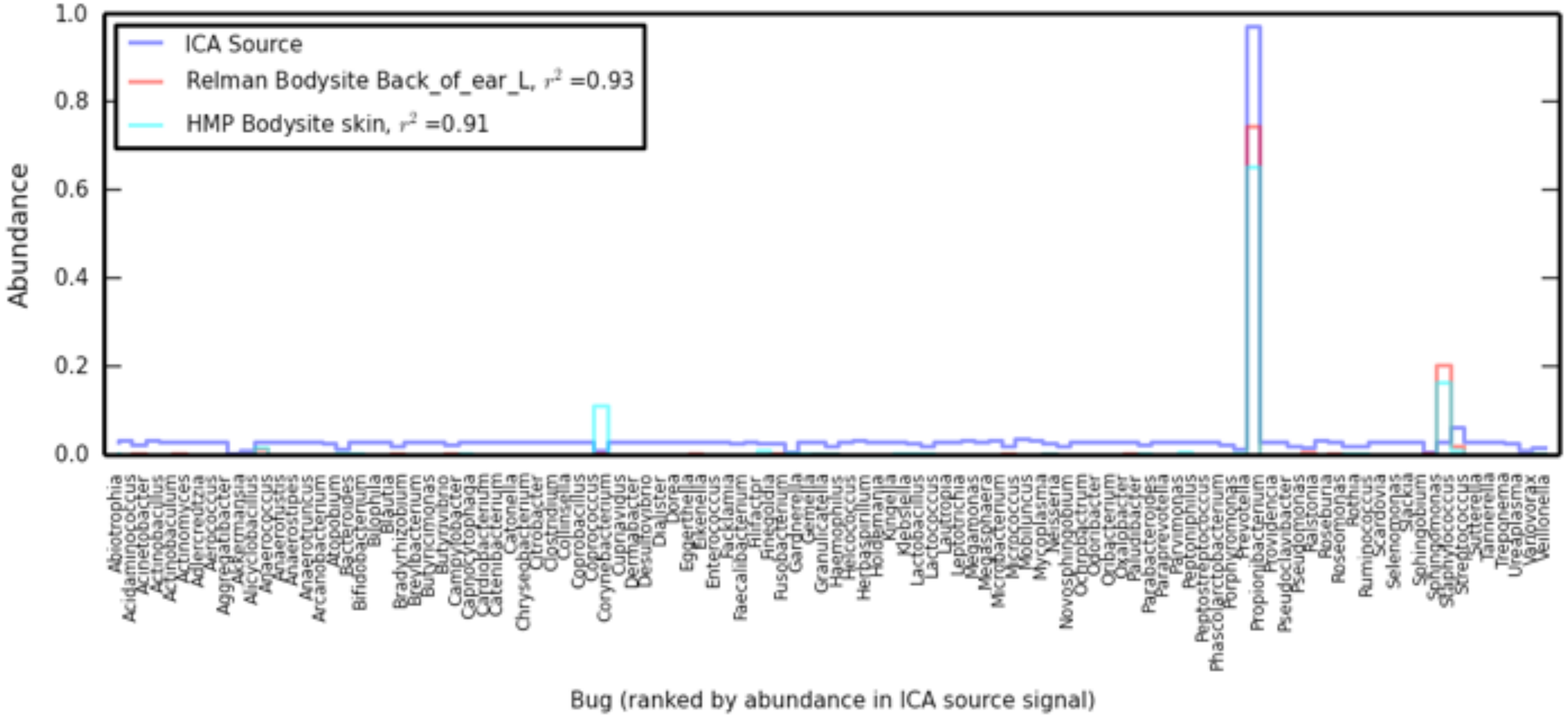
(4) Clustered mixing matrix can to determine which sources are enriched in each sample.

Given the set of mixtures (blood samples), it will simply compute a set of microbial sources as well as a mixing matrix that explains how each sample is computed from these sources. The samples can be analyzed with respect to known data to assign a likely “tissue source.” The mixing matrix can be analyzed to learn about each sample.

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

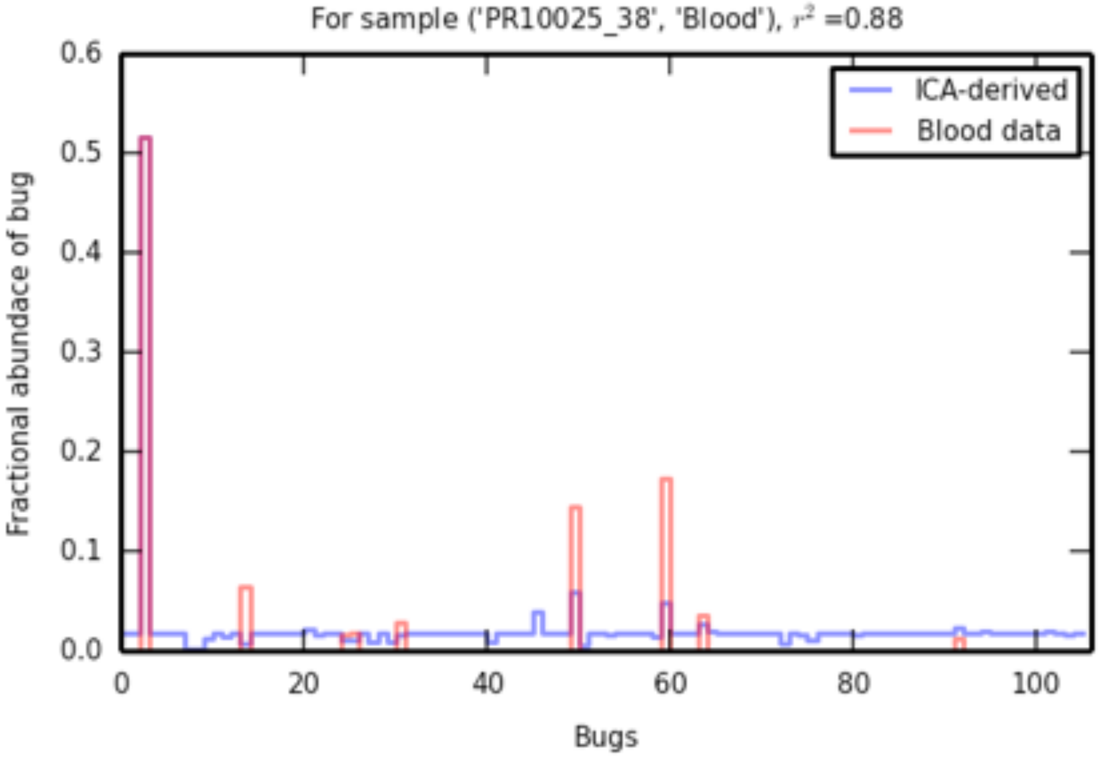
Known sources -

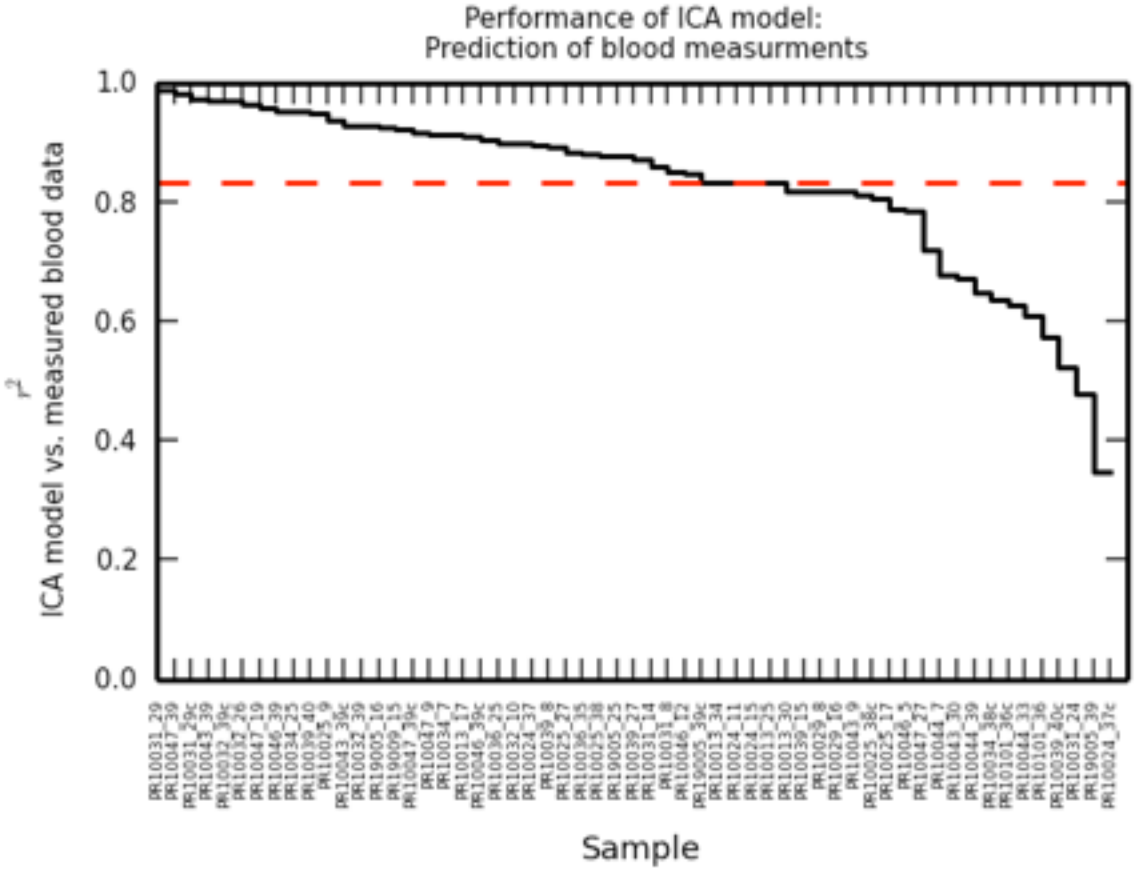




For each ICA component, we can assign it a likely tissue source using existing data (e.g., HMP).

Performance evaluation: Reasonable.

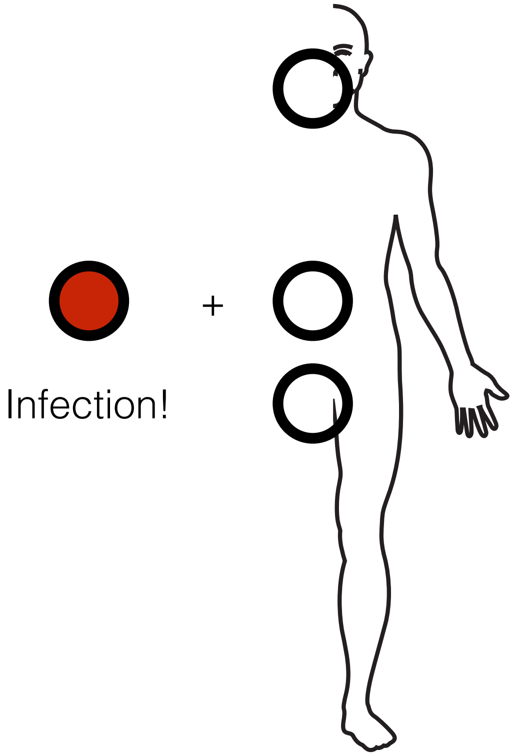




We can also evaluate its ability to re-capitulate the blood measurements, which is far better than linear models. Of course, this is because ICA is not constrained to the sampled sites; “blind sources” can be emitted by the algorithm and may not have a sensible assignment (e.g., more experiments or data may be needed to understand them).

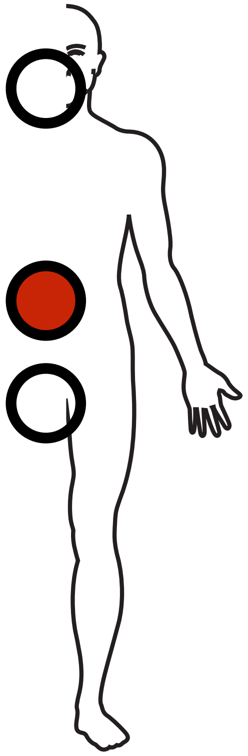
If this works, it would be useful for unsupervised detection of anomalies.

Identification of novel sources (e.g., infection!)



Normal body sites.

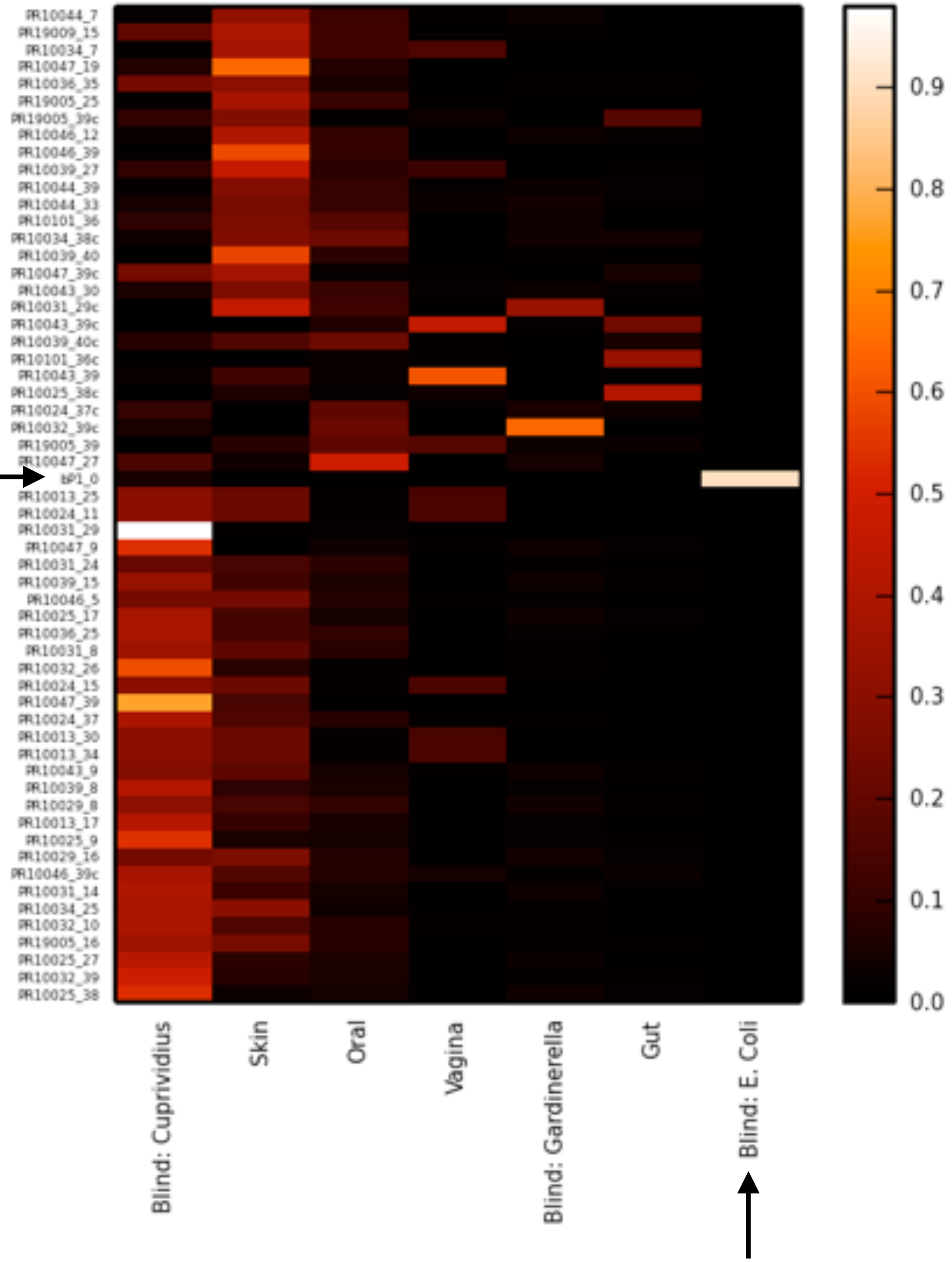
Imblance of healthy signal sources (e.g., IBD outlier)



Normal body sites.

If a patient has an un-recognized source, it will be obvious using this pipeline. For example, infection would be detected as a novel source. Furthermore, the mixing of known sources may differ between patients (or cohorts), indicating systematic imbalances.

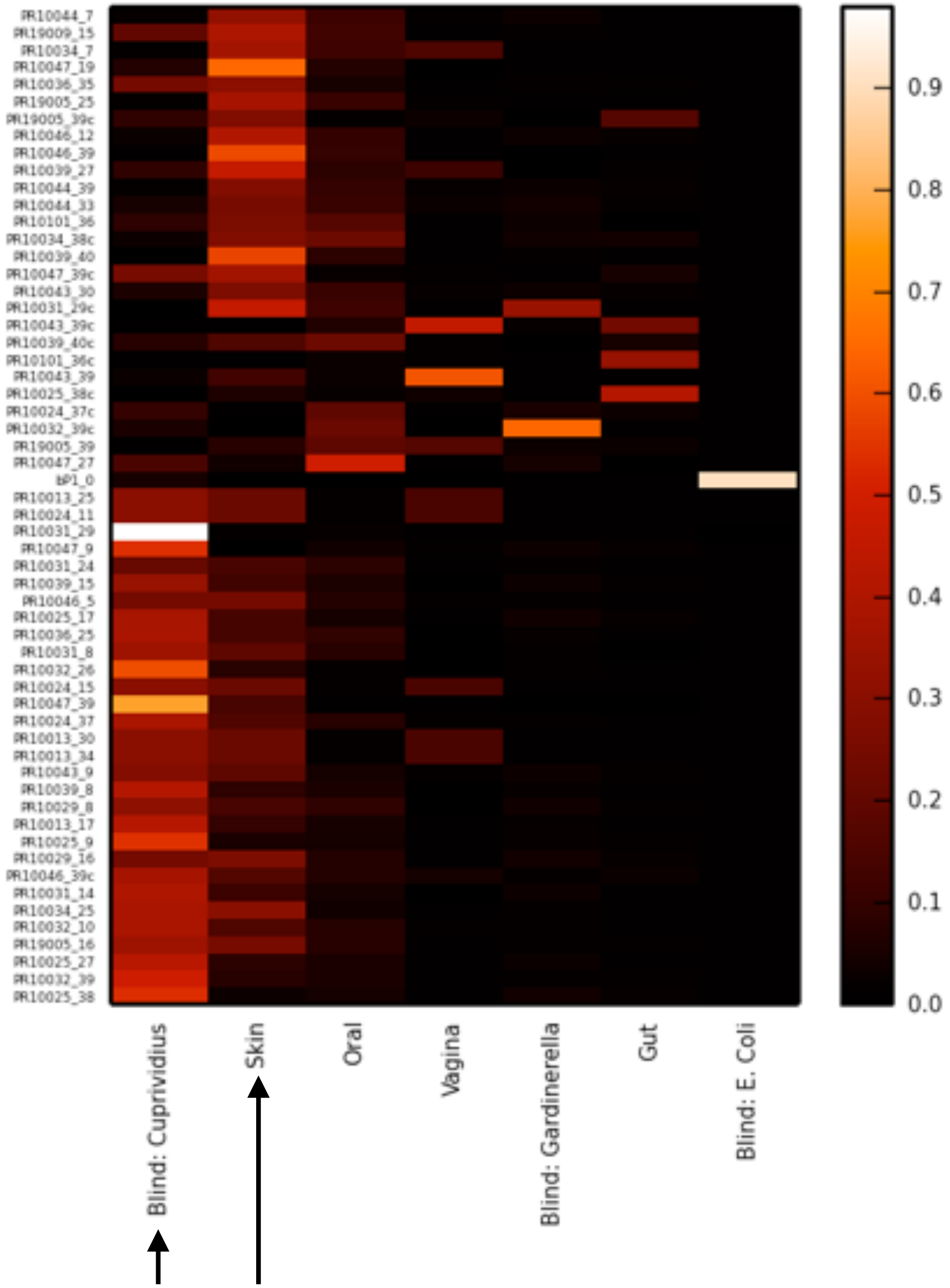
Clustered mixing matrix shows un-supervised de-convolution.



With spike-in of anomalous patient into existing cohort, ICA “discovers” a new source signal that is only weighted heavily in this case.

.. 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)

With this approach, any blood sample can be de-convoluted and signals can be analyzed with respect to a growing canon of body site signatures. This will aid in evaluation of differential mixing of body sites within cohorts. Patient segmentation can be performed based upon this (e.g., IBD sufferers likely have a higher mixing load from gut relative to normal). Furthermore, “blind” source signals will become flags for further investigation (e.g., Cuprividius in this cohort) and / or will be immediate indicators of pathology (e.g., as shown in the biopsy case here). In turn, the ICA pipeline may be useful pre-processing step applied this type of data.