Monitoring the human infectome

Lance Martin
Quake GM, 12/9/14

Non-invasive detection of any pathogen.

Sensitivity

Full genome

Low

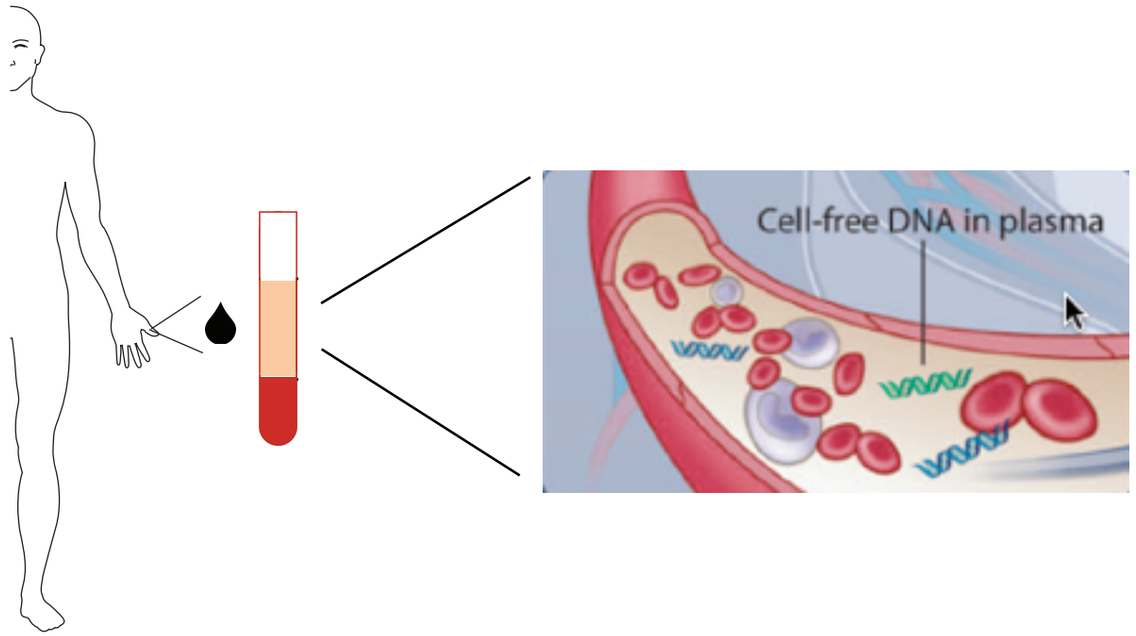
Few bugs

All bugs

Scope

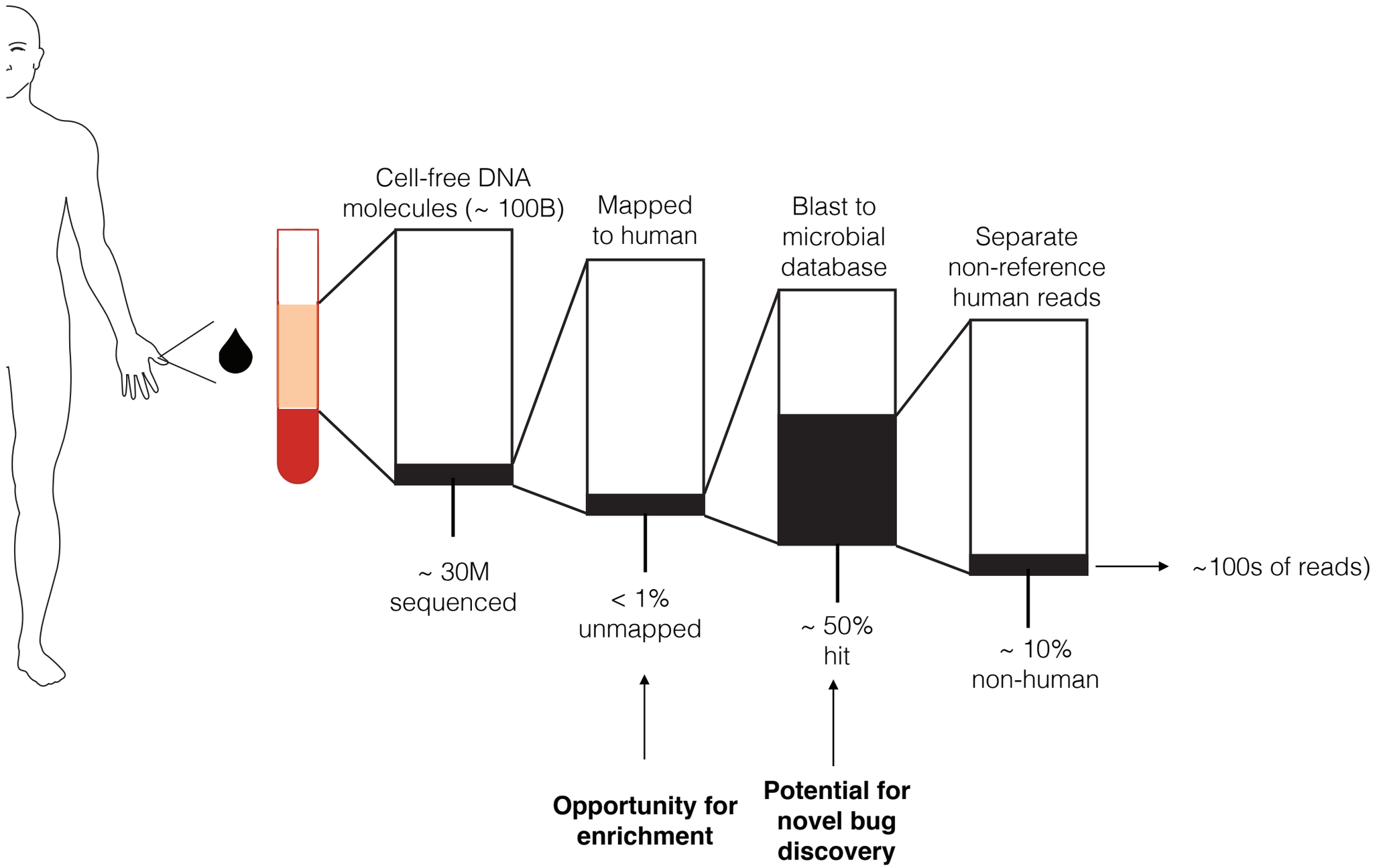
Technology Comes to Typing. Nature Biotech (2014)

Cell-free DNA now used for non-invasive diagnostics.

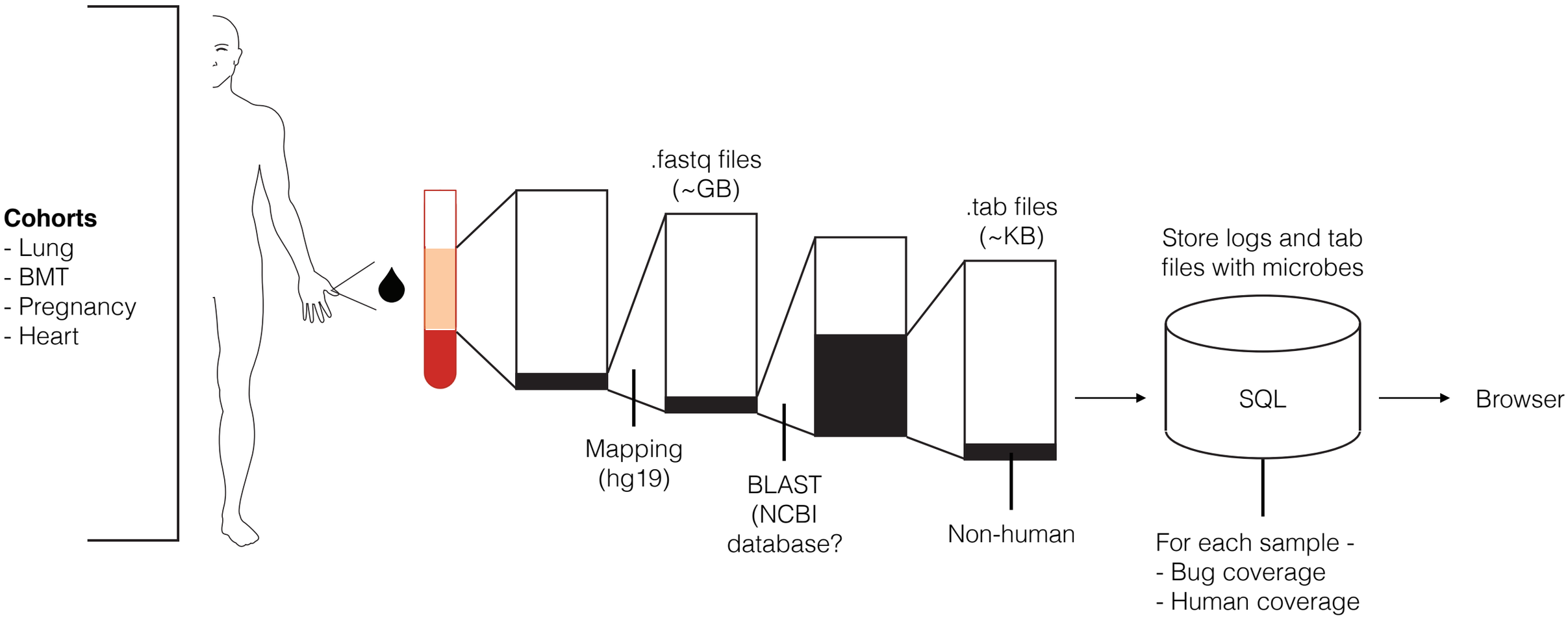


De Vlaminck et al Sci Trans Med (2014)

Human microbiome in cell-free DNA.

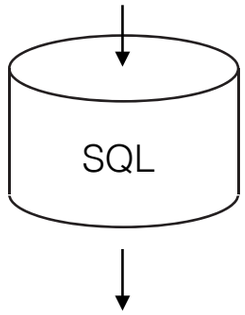


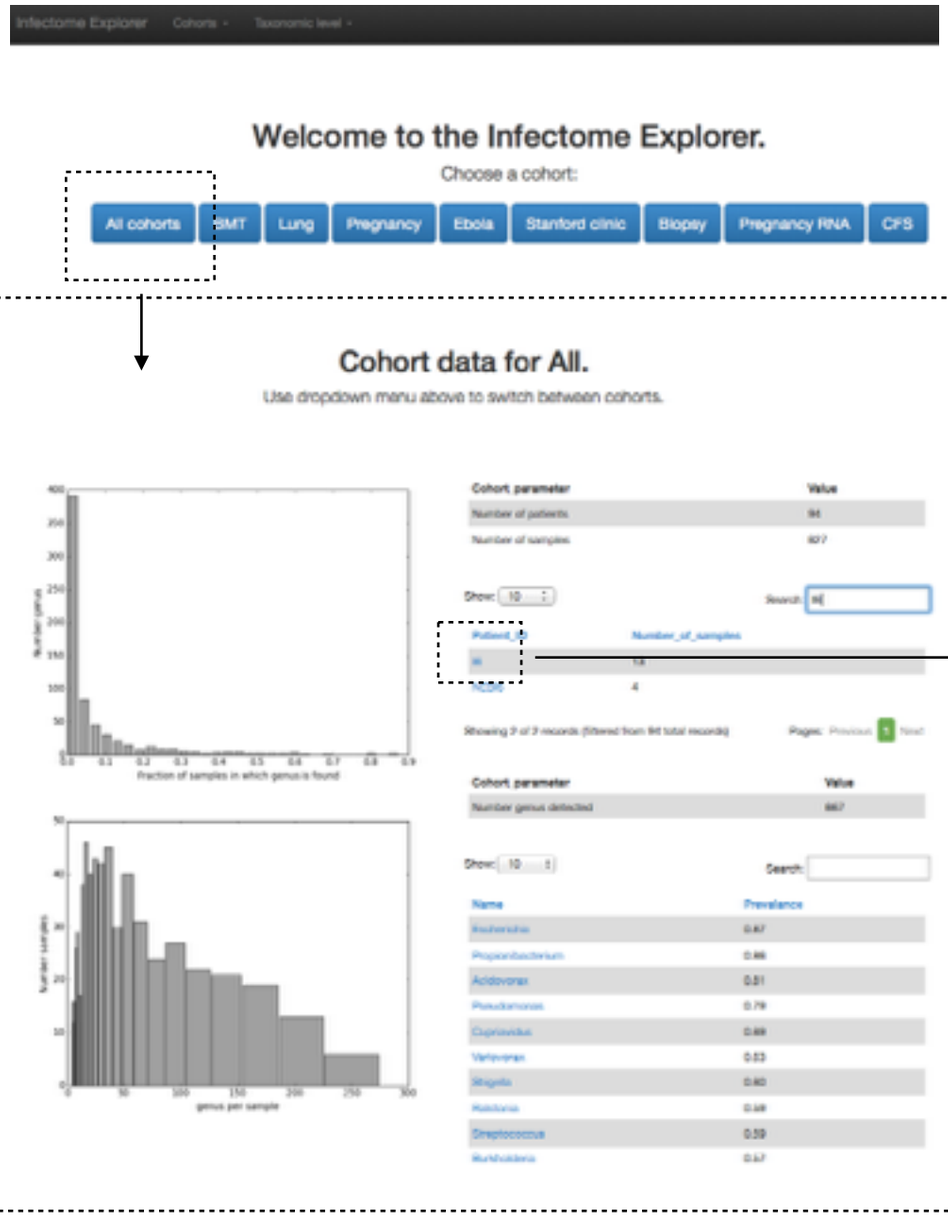
Pipeline for isolating microbial reads in cell-free data.

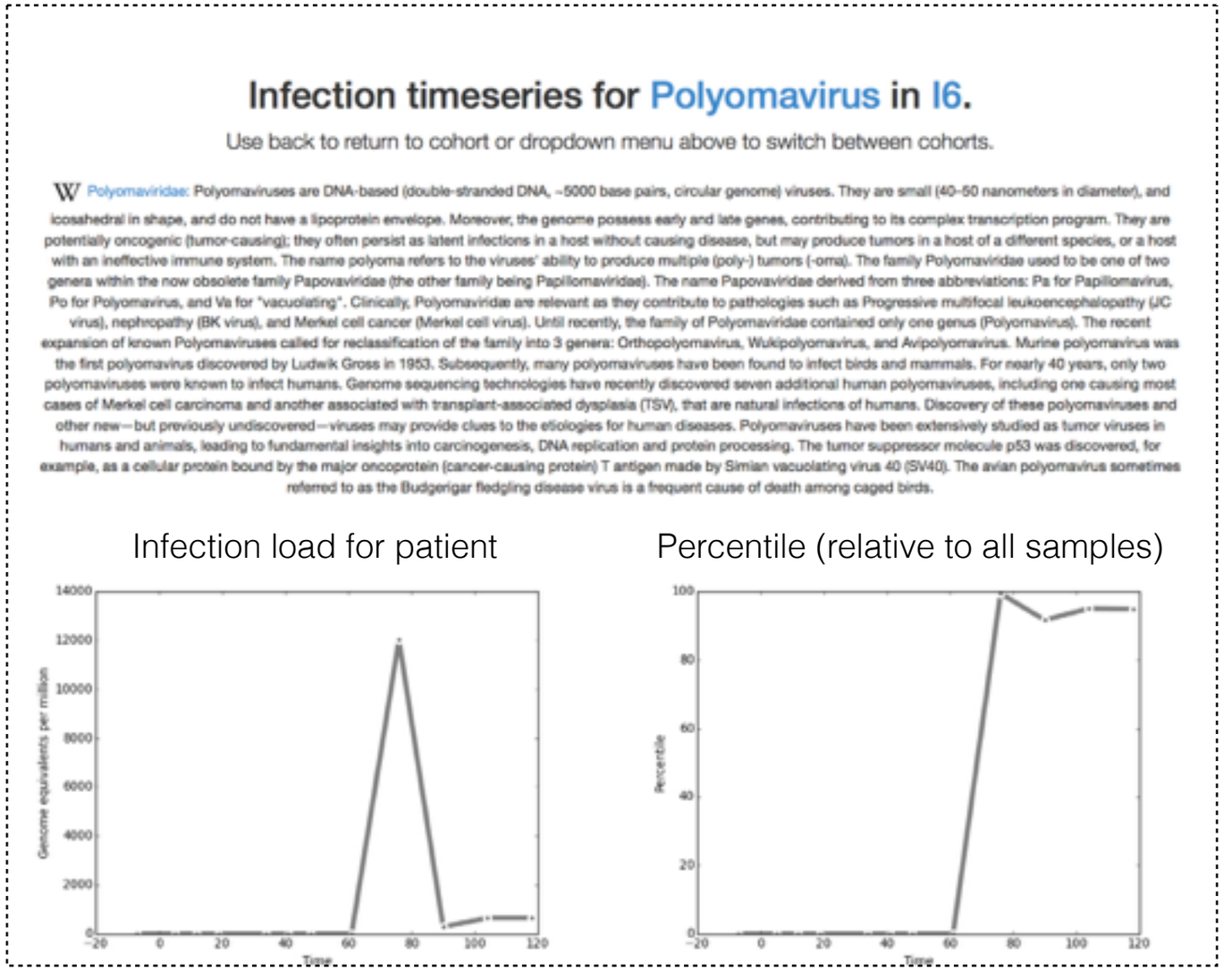


Browser for visualization and navigation.

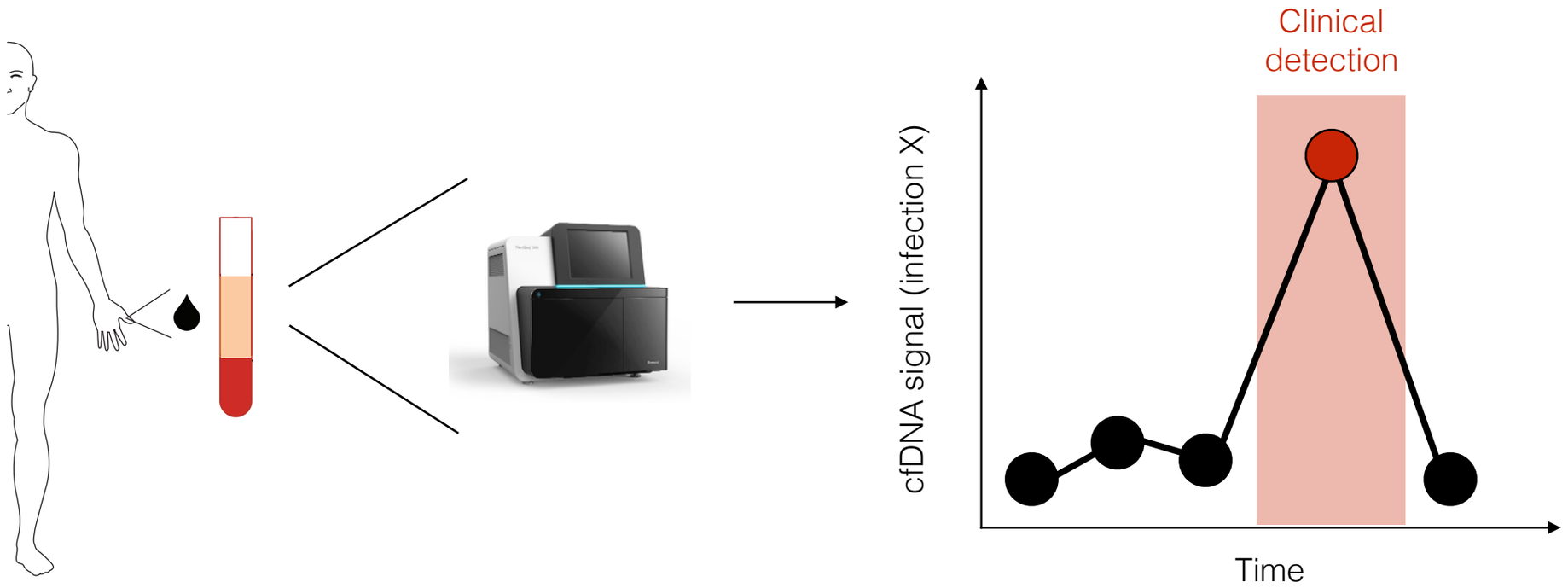
1000s of samples





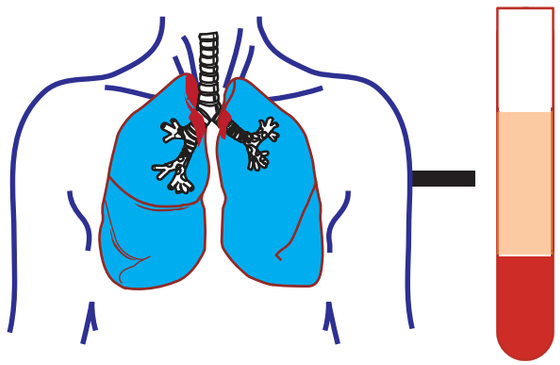


But how to know if it’s useful?

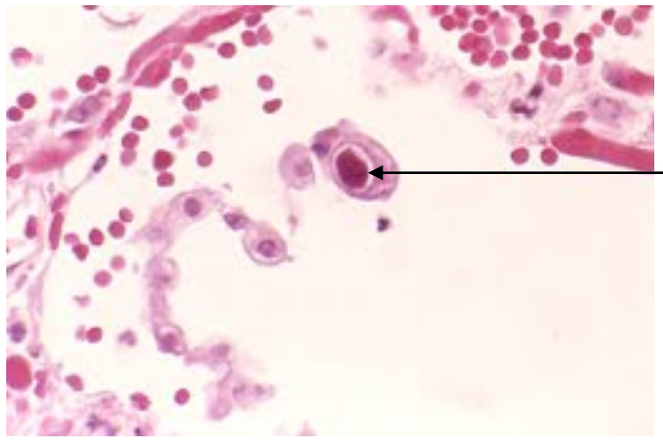


Look back at the clinical history of our cohorts.

Lung Transplant (431 samples collected and processed) -

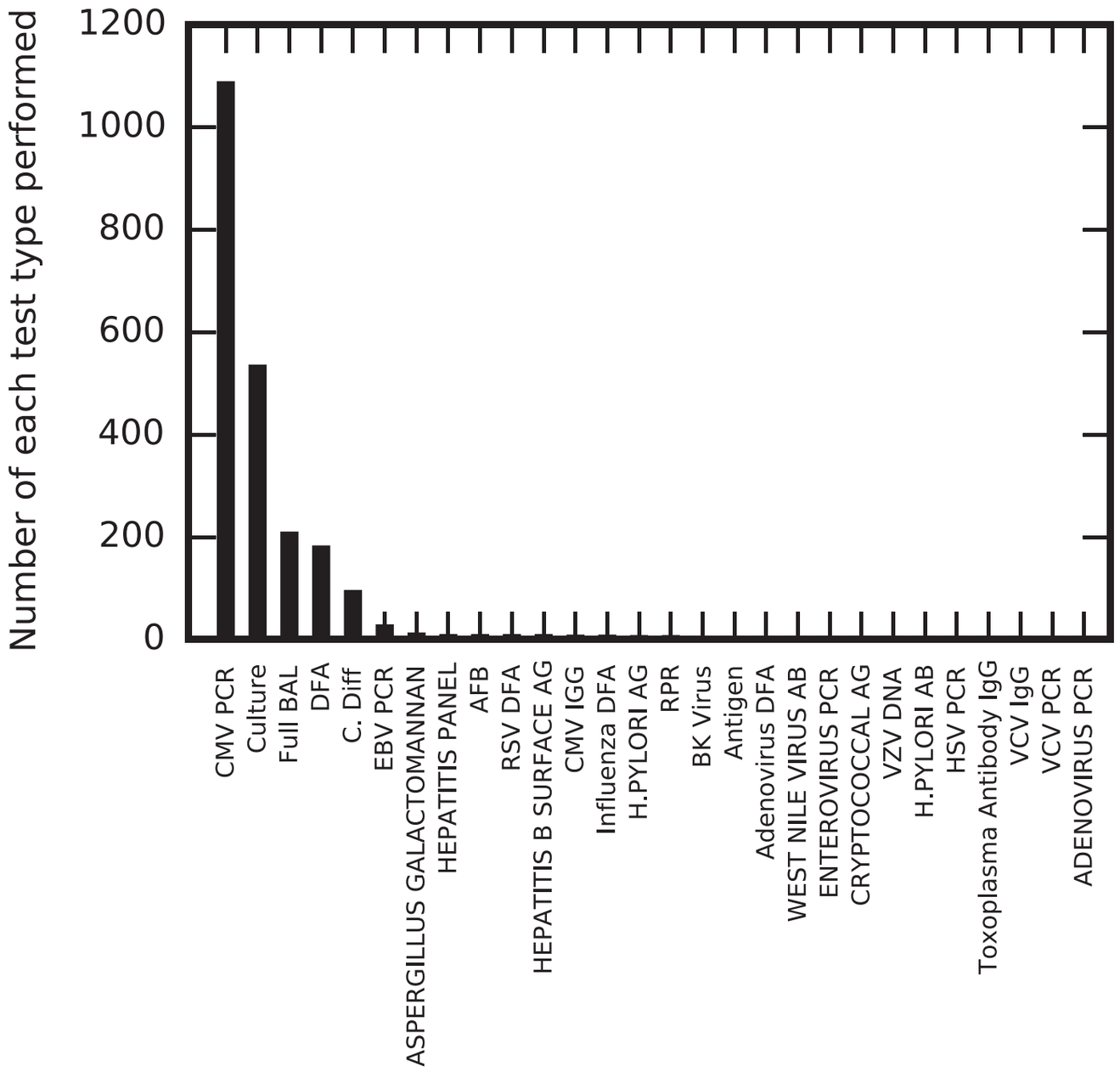


Clinical tests



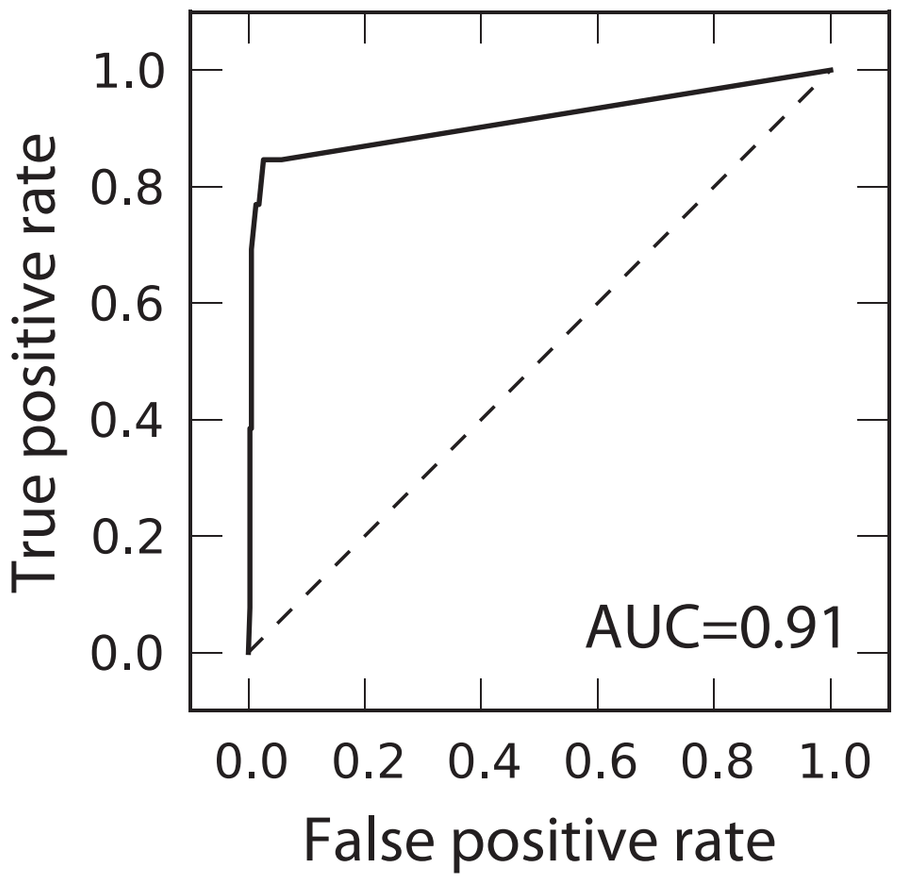
CMV infected lung cell

Thousands of tests recorded (~35k bug measurements) -



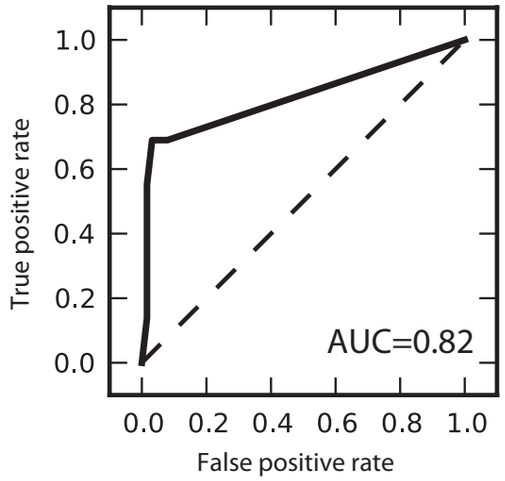
Performance on viruses.

CMV in lung cohort -

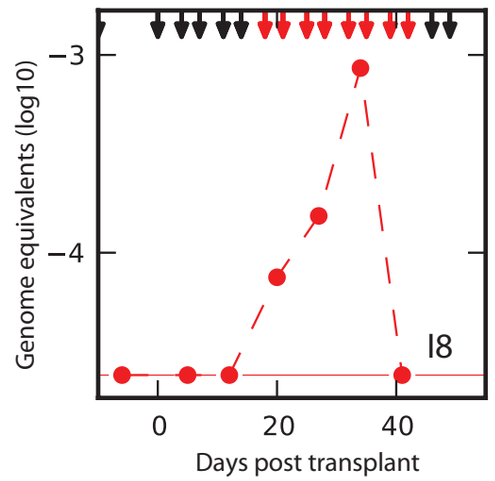


Bone marrow cohort -

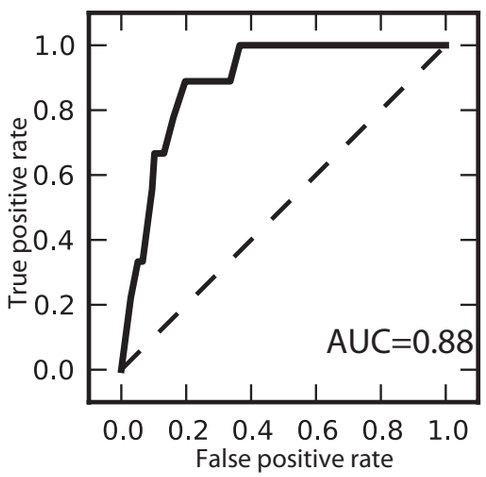
Adenovirus

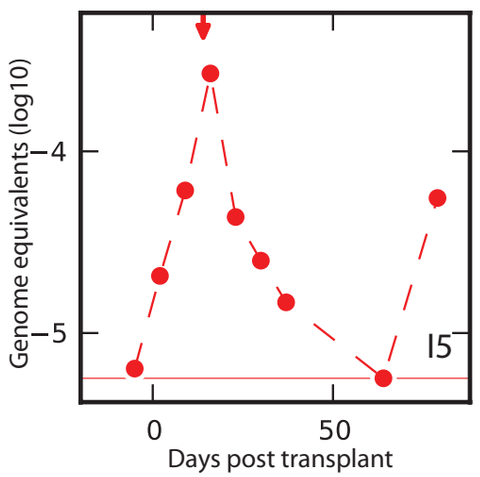


Example patient with clinical tests (pos, neg)



CMV

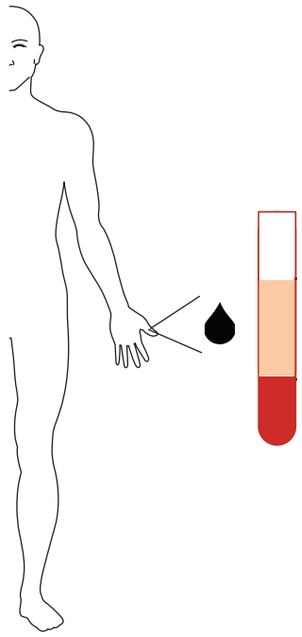


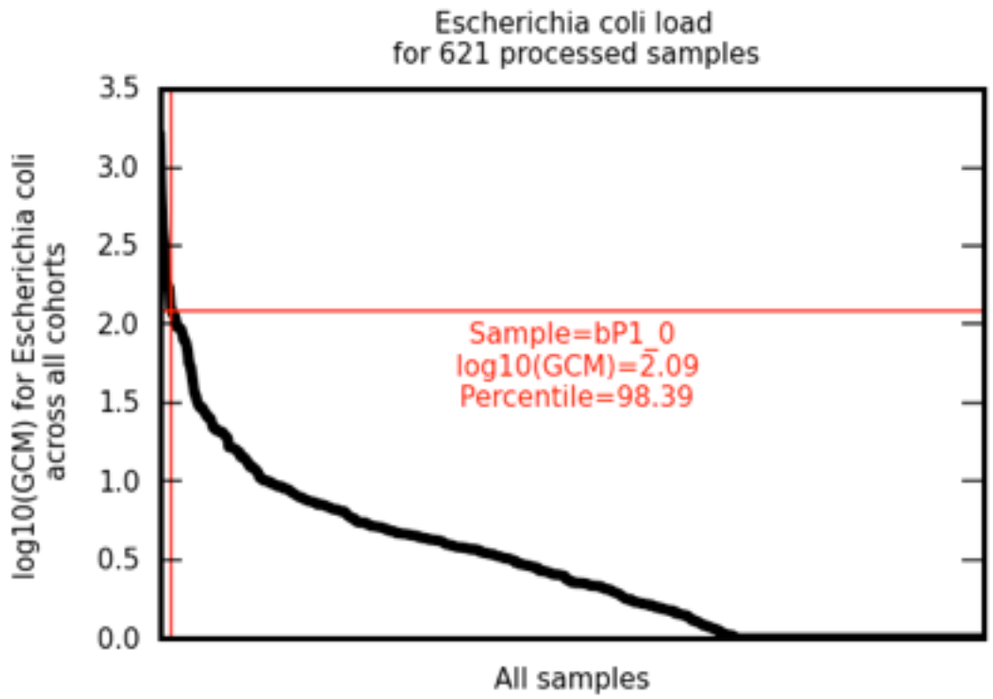


Performance on microbes (detected in deep tissues).

Stanford Pathology -
Biopsy and MALDI-tof

E. Coli





Ranking of all bugs detected in the sample -

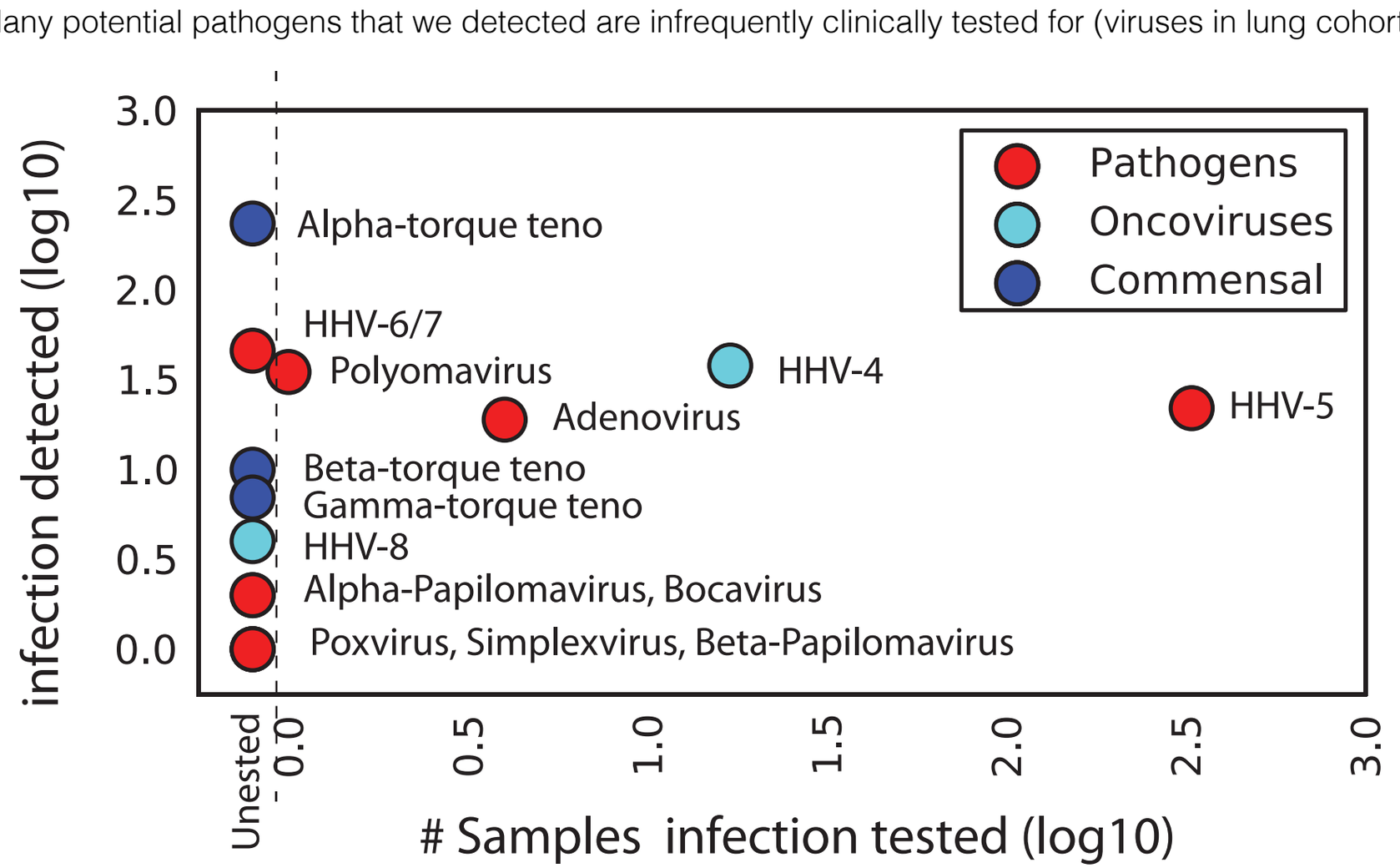
Sorted infection data for bP1_0.

Use back to return to cohort or dropdown menu above to switch between cohorts.

Show: 10 Search:

Name	Gen_Equ	Percentile
Candidatus Midichloria	0.28	99.516908
P2likeviruses	34.09	98.711755
Gammatorquevirus	169.43	98.711755
Escherichia	123.17	98.389694
Alkaliphilus	0.11	98.228663
Shigella	8.45	96.940419
Chelativorans	0.11	95.169082
Enterocytozoon	0.15	94.363929
Citrobacter	0.31	93.236715
Psychrobacter	0.39	92.914654

Common detection of un-tested / un-diagnosed infections.



Undiagnosed case of infection.

I6, Cause of death:
Respiratory failure.

Sorted infection data for I6.

Use back to return to cohort or dropdown menu above to switch between cohorts.

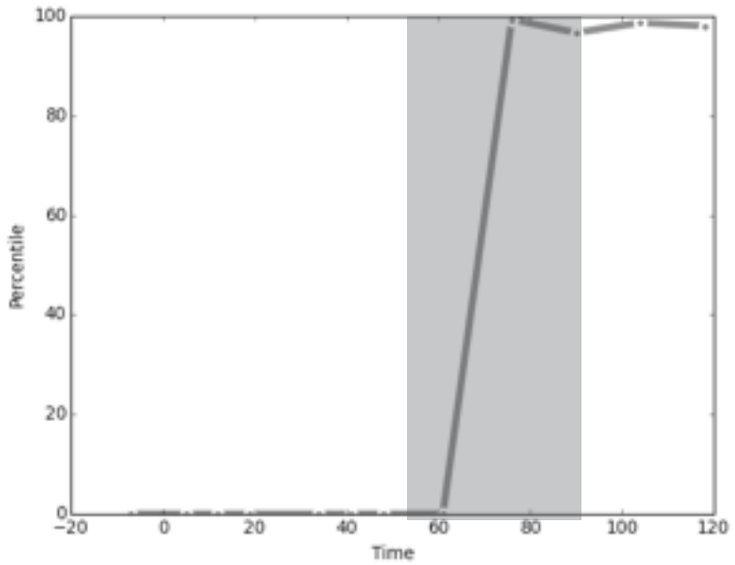
Show: 10 Search:

Name	I6_BC	I6_D-1	I6_W1	I6_W2	I6_W3	I6_W4	I6_W5	I6_W6	I6_W8	I6_W10	I6_W12	I6_W14	I6_W16
WU Polyomavirus	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	99.337748	96.688742	98.675497	98.013245
Human herpesvirus 5	0.000000	75.496689	88.079470	78.145695	0.000000	90.066225	80.132450	89.403974	91.390728	0.000000	82.781457	0.000000	0.000000
Enterocytozoon bieneusi	0.000000	0.000000	90.728477	95.364238	94.039735	0.000000	96.026490	98.013245	98.675497	99.337748	94.701987	92.715232	0.000000

.
.
.

WU Polyomavirus
(Rare virus that causes severe respiratory infection)

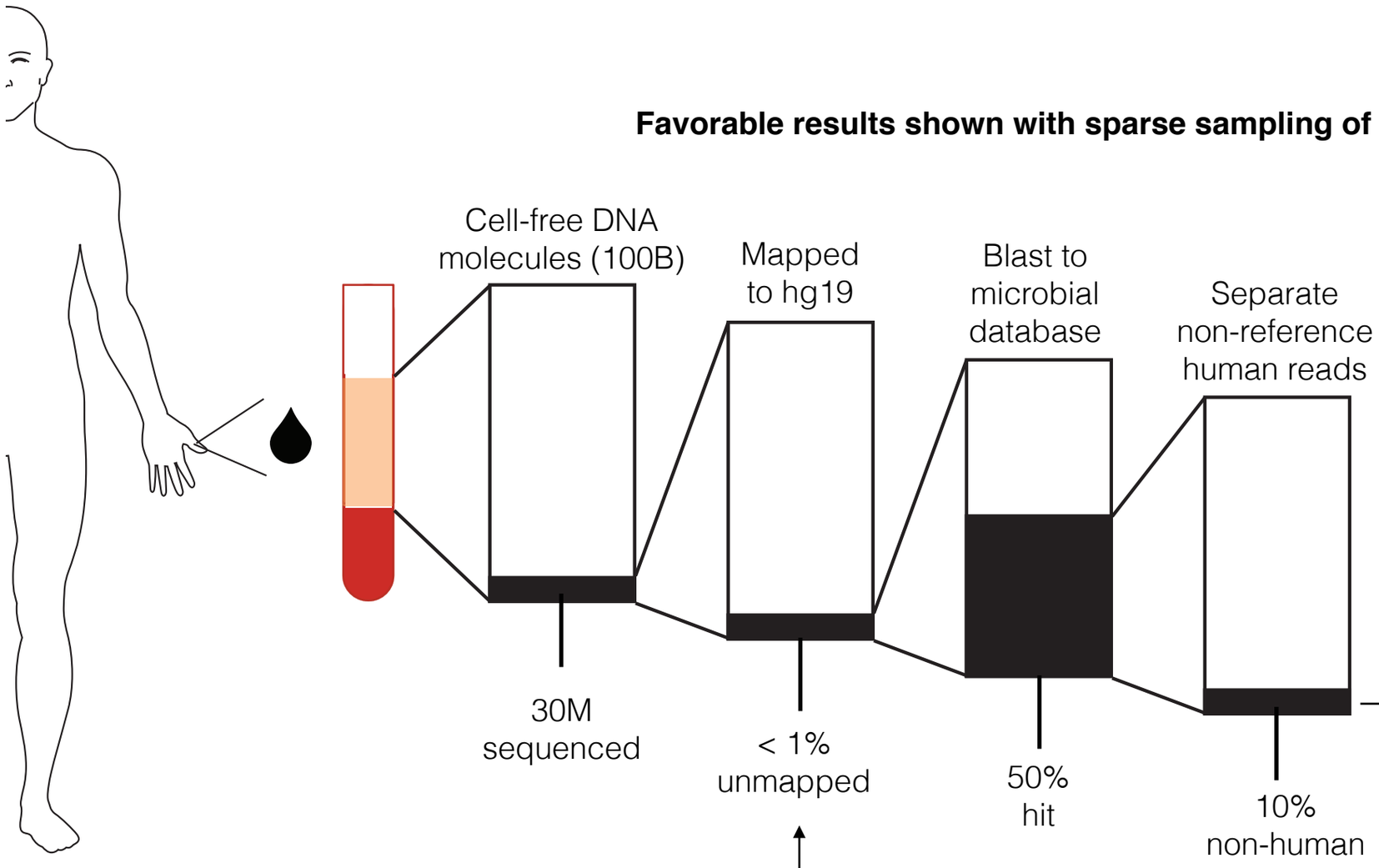
Very high load of WU Polyomavirus during time period of negative clinical test results.



Multiple (-) tests for
BK polyomavirus

Also, headroom for improvement.

Favorable results shown with sparse sampling of microbial reads.



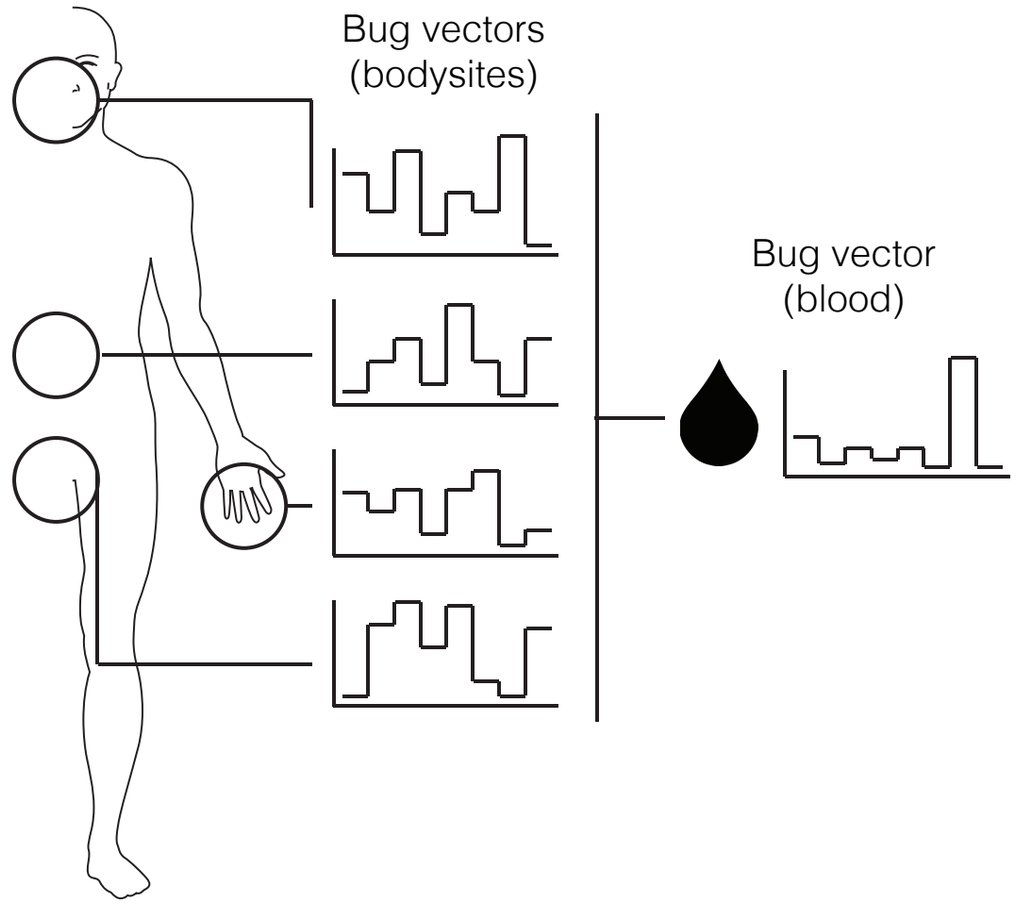
Clinical studies -
- Lung
- BMT

Deep tissues -
- Biopsy replacement

Undiagnosed
- Viruses (I6)
- Fungi (L78)

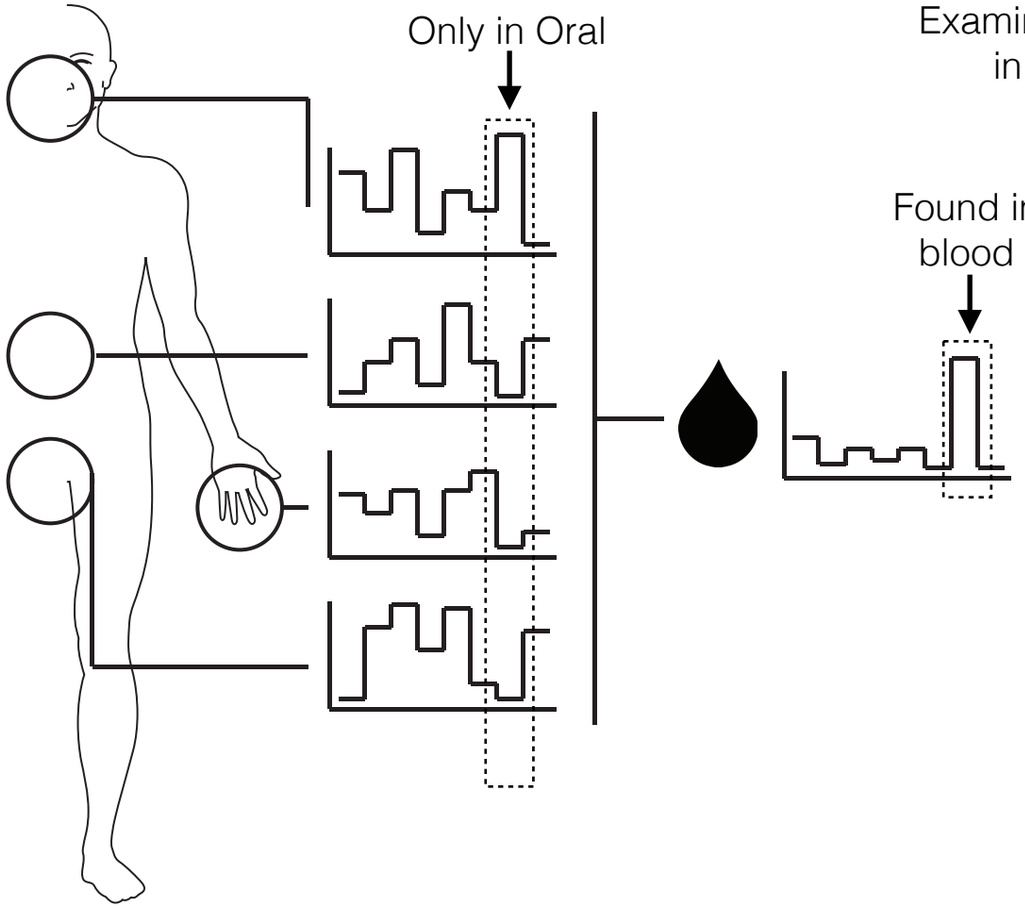
But, where does microbial cell-free DNA come from?

Examine a pregnancy cohort of with microbiome sequencing at 4 body sites with matched blood samples -



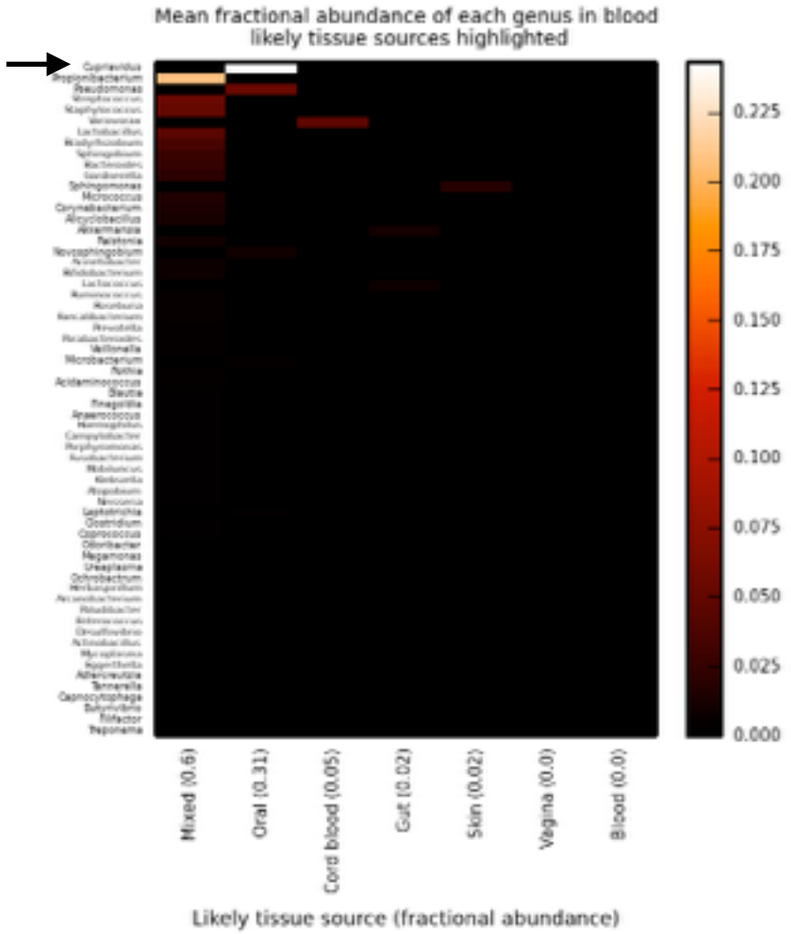
Tissue specific bugs don’t explain blood composition.

Define tissue specific bugs -

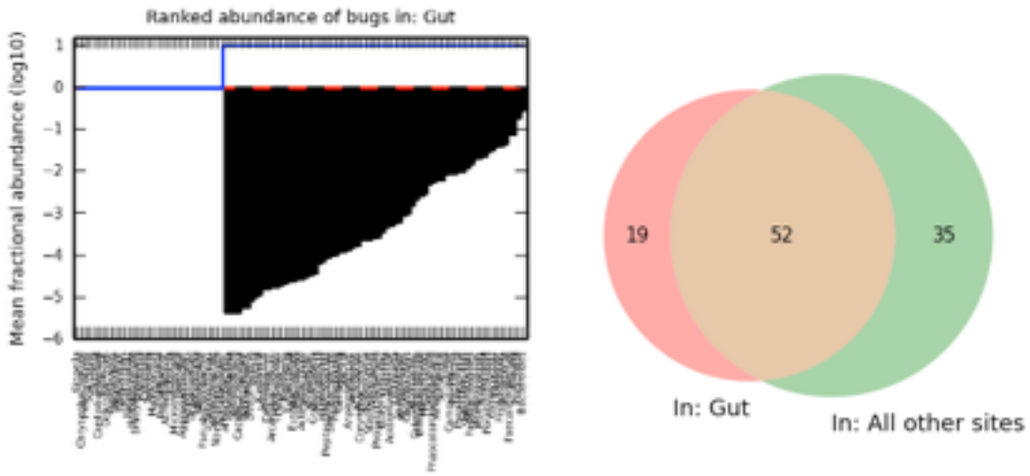


Assign likely source to each bug found in blood -

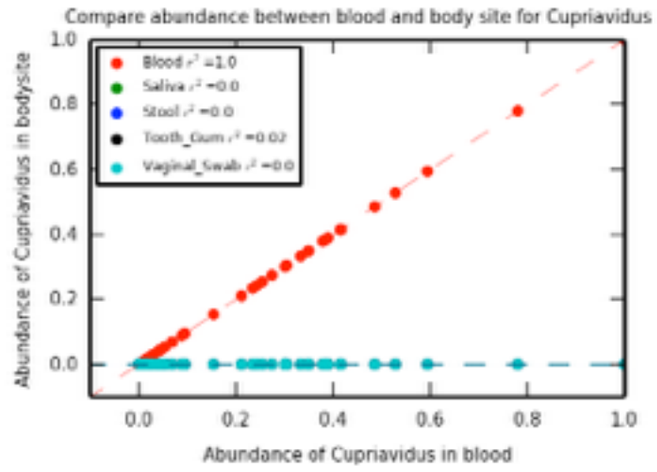
Examine Cuprividius in panel (D)



Bugs at each site are re-cast as bit vectors -

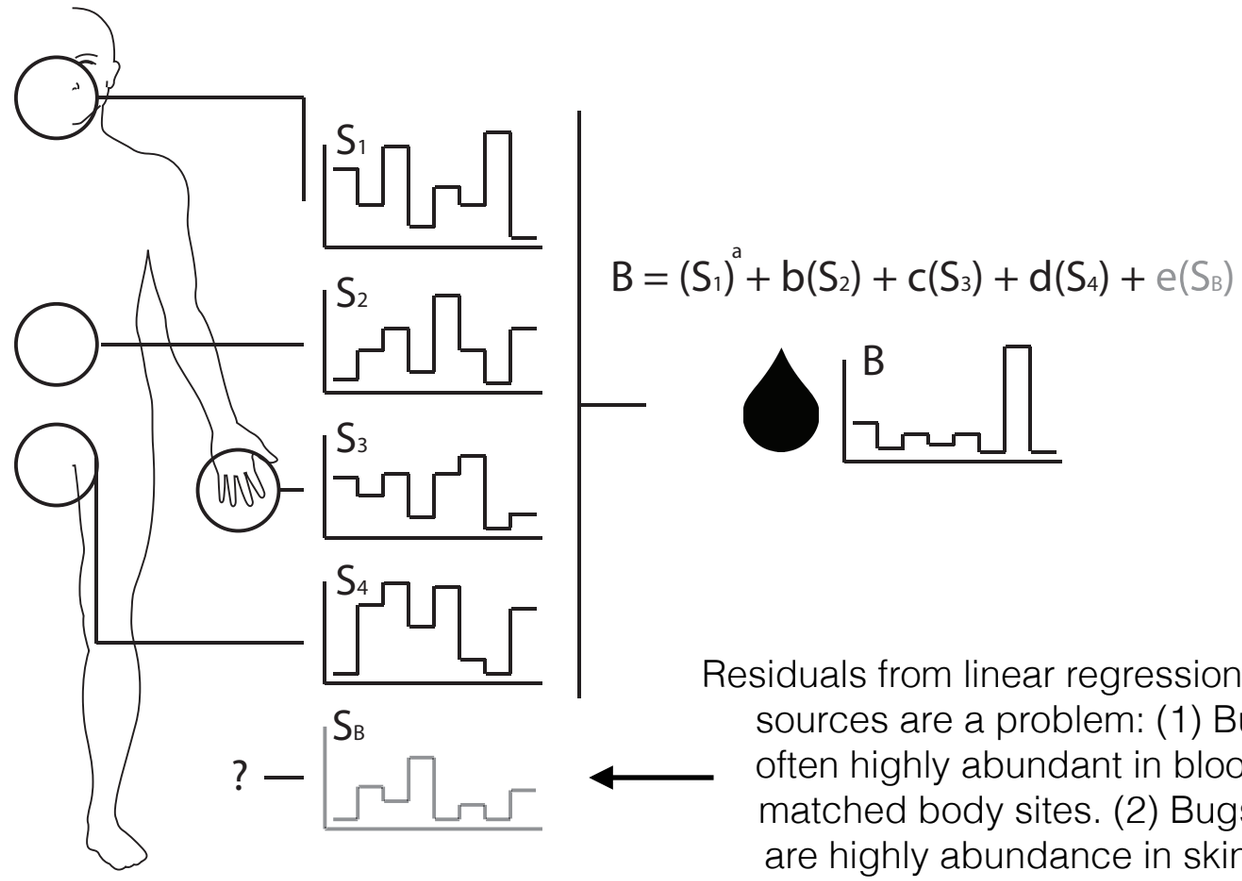


Cuprividius has trace abundance in other tissues, but high in blood -



… Neither does linear model of commonly sampled sites.

The four sampled body sites are insufficient to model blood, as it likely draws from additional tissue sources -



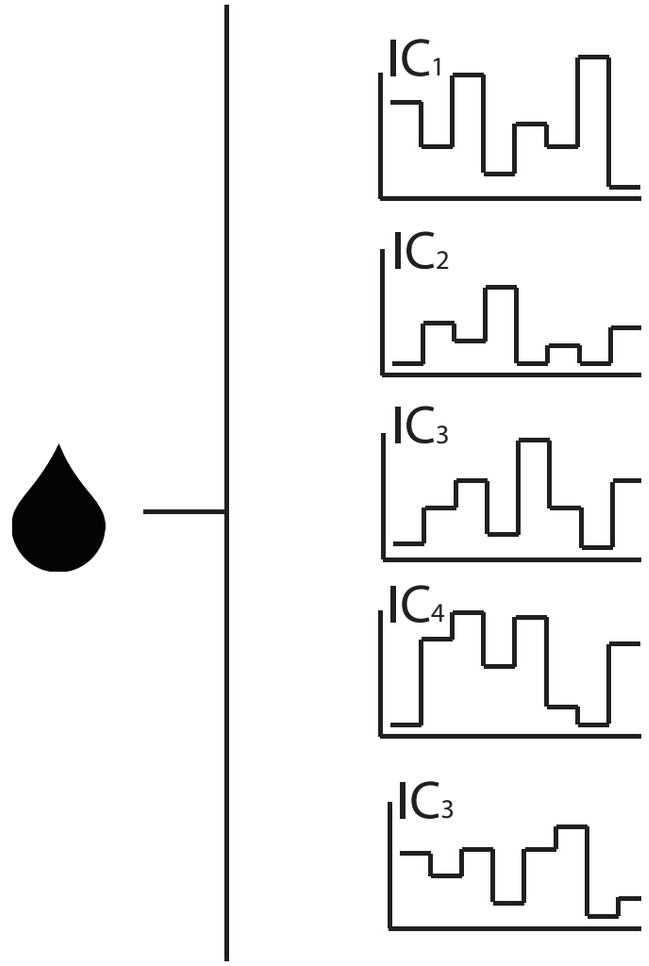
Residuals from linear regression [see appendix] show that blind sources are a problem: (1) Bugs such as *Curprividius* are often highly abundant in blood sample, but not present at matched body sites. (2) Bugs such as *Propionibacterium* are highly abundance in skin, but skin is rarely sampled.

Try approach that is not constrained to sampled sites.

The cocktail party problem -

Toy problem for source de-convolution from mixed signals -
- Non-Gaussian, independent sources (voices)
- From recorded mixed conversations, it is possible to isolate the sources.
- ICA algorithm used to do this (see derivation in appendix).

Blood microbiome as a cocktail party problem -



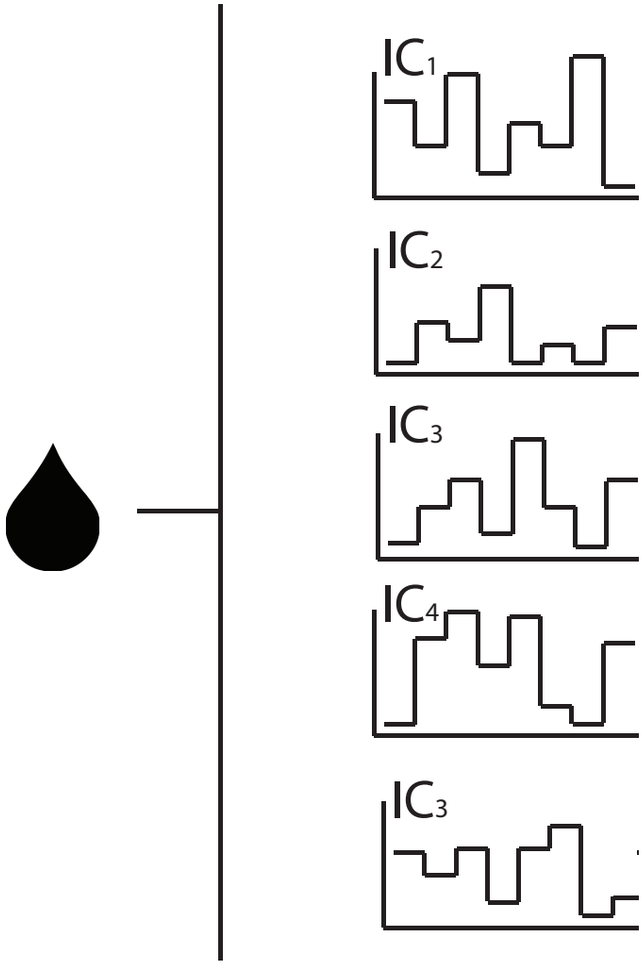
- Non-Gaussian, independent sources (body sites)
- Measured mixtures (blood samples)
- Use ICA to to isolate the body site sources

Unsupervised source detection in the pregnancy cohort.

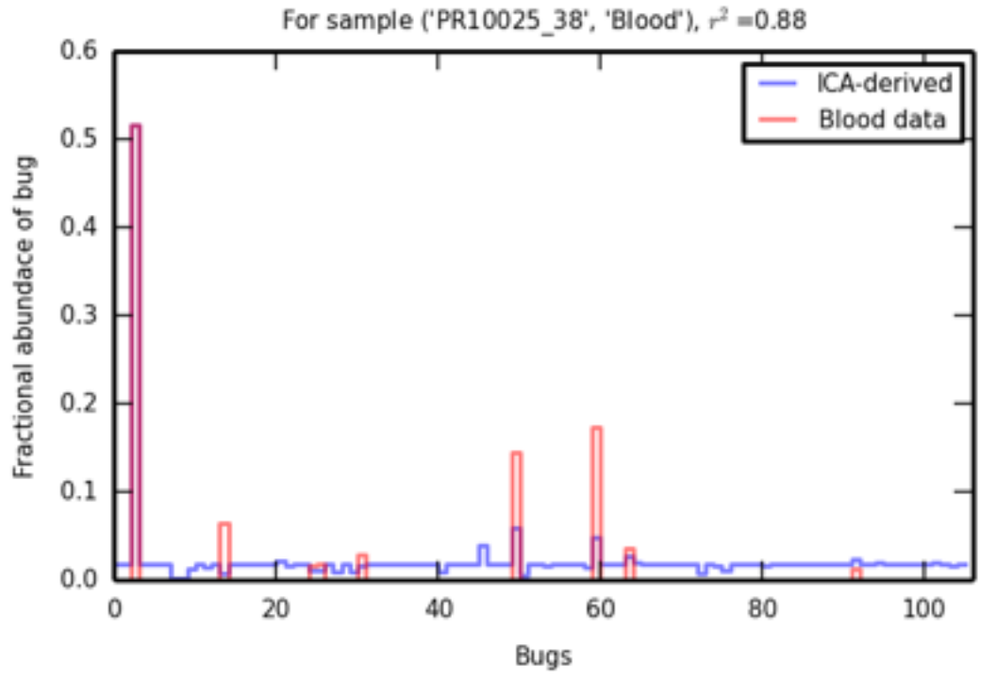
Blood data for any sample re-capitulated using ICA sources and mixing matrix -

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

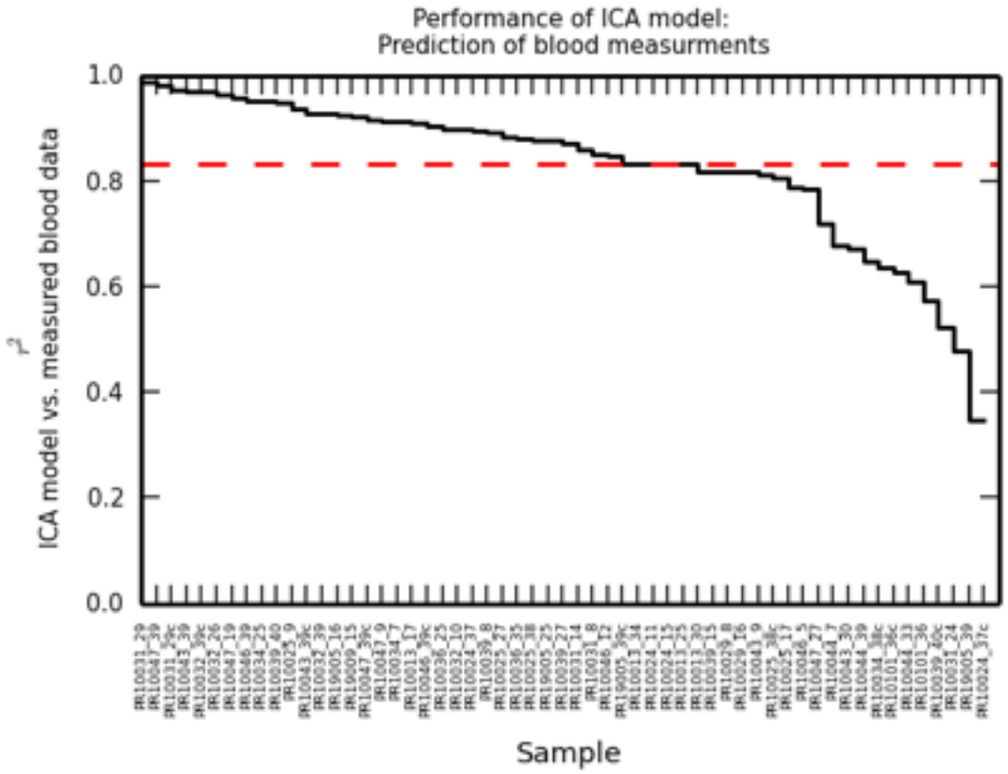
(1) Sources -



(2) Mixing matrix, A



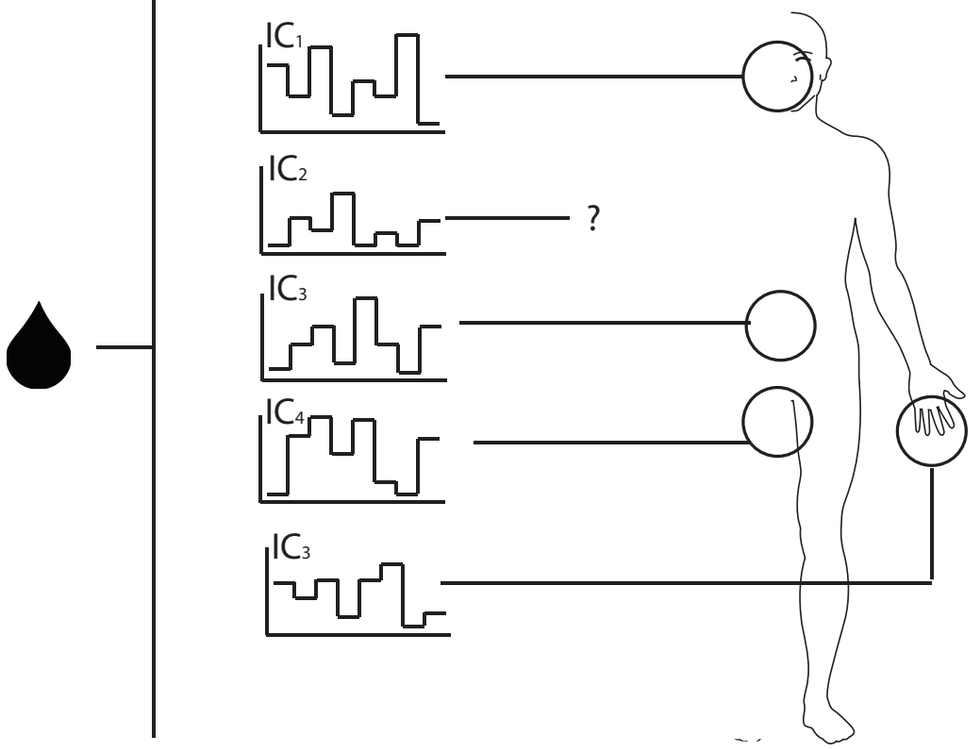
All samples (with mean highlighted) -

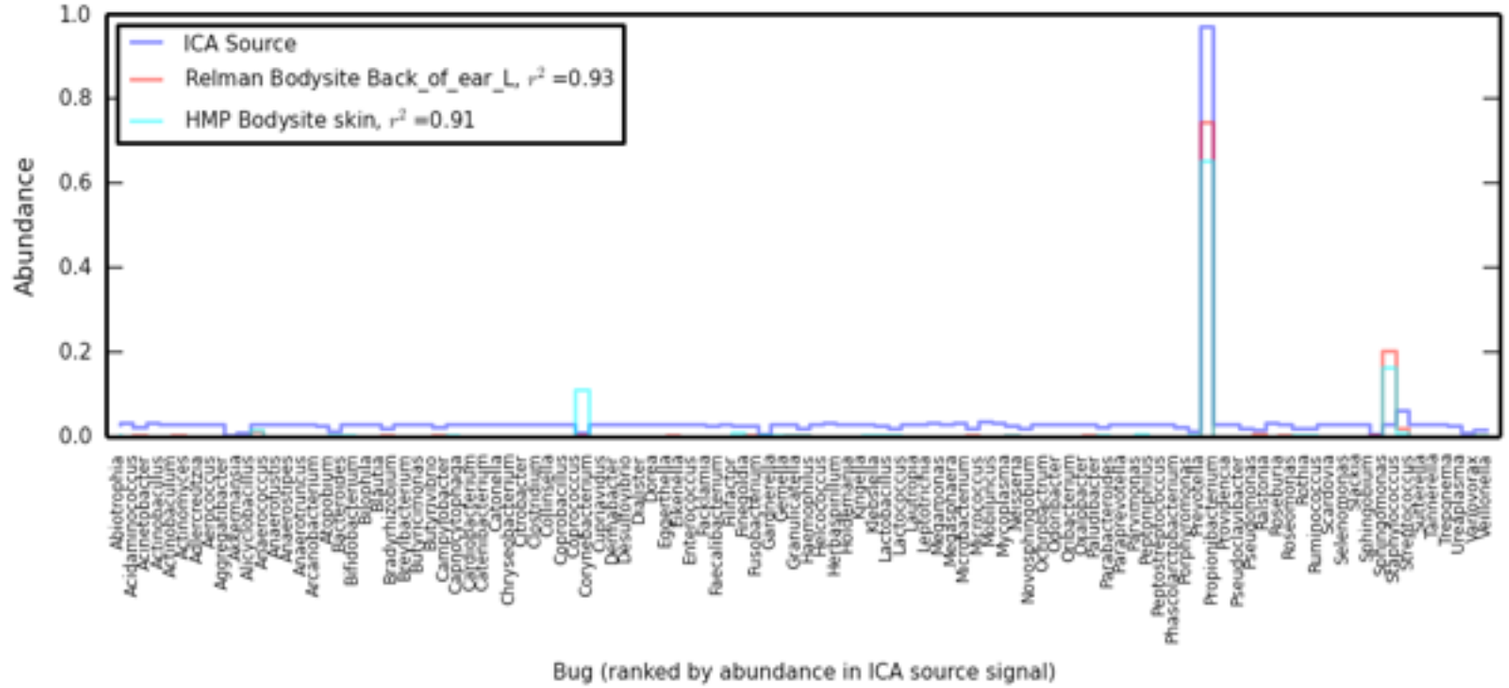


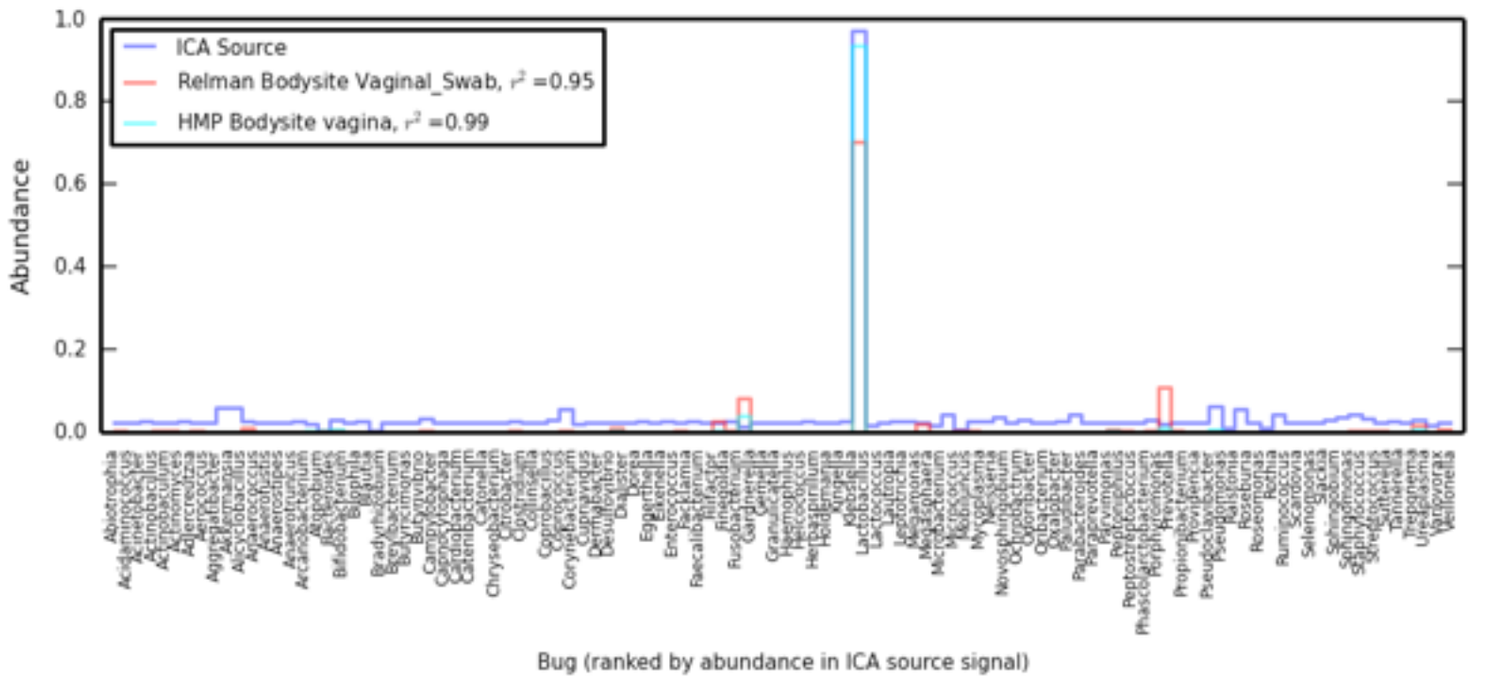
Assign ICA sources to sampled body sites.

We can then try to assign “discovered” ICA sources to tissues using HMP and Relman data -

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

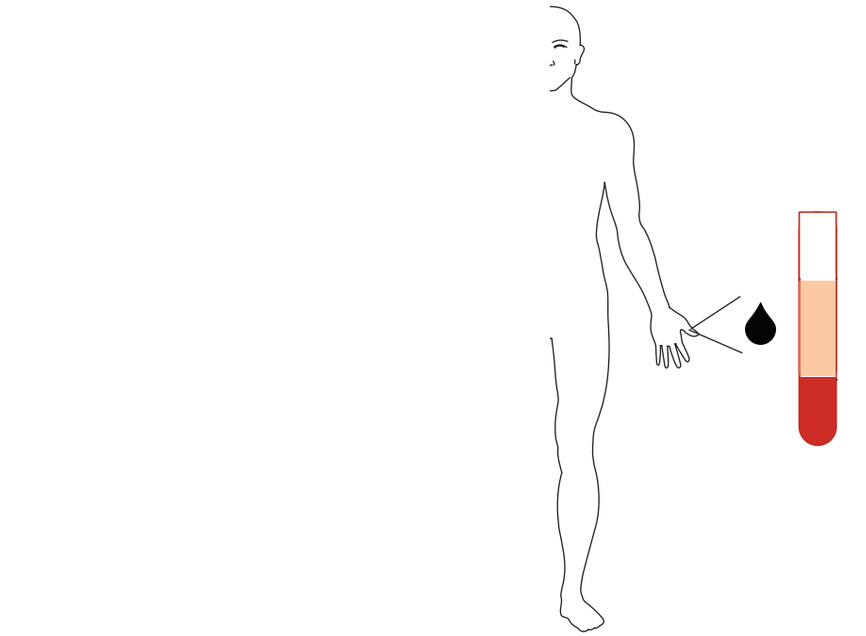




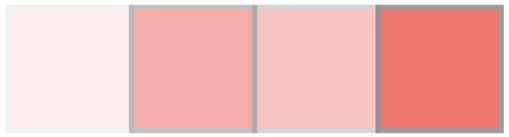


Each blood sample is a combination of ICA sources.

The mixing matrix tell us the weight of each source per sample.



Relative weight
of ICA sources:



Skin

Gut

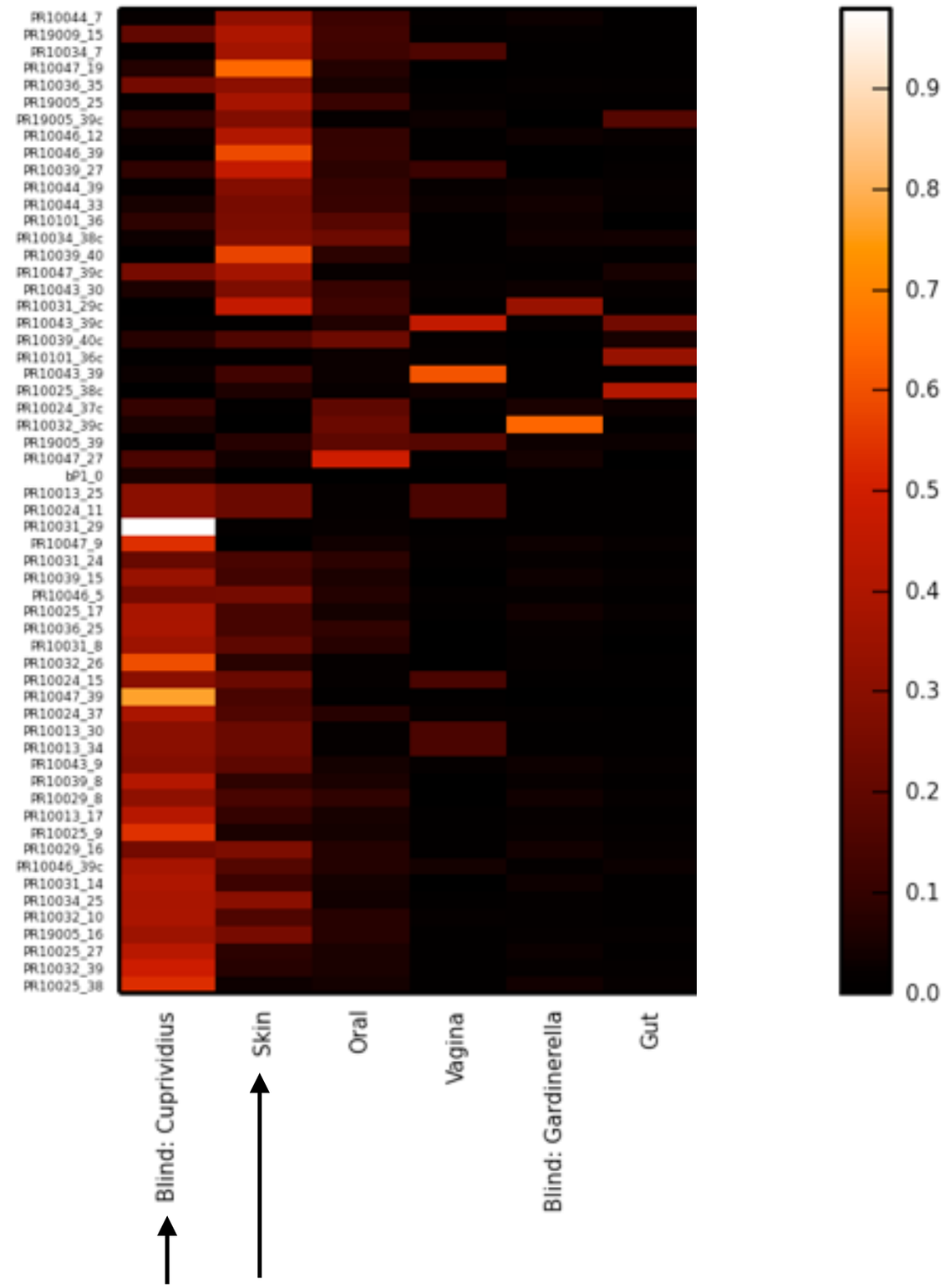
Oral

“Other”

Clustered mixing matrix for pregnancy cohort.

Cluster (1): Enriched in skin-like signal (high in *Propionibacterium*), which may come from skin or a different (unsampled) deep tissue.

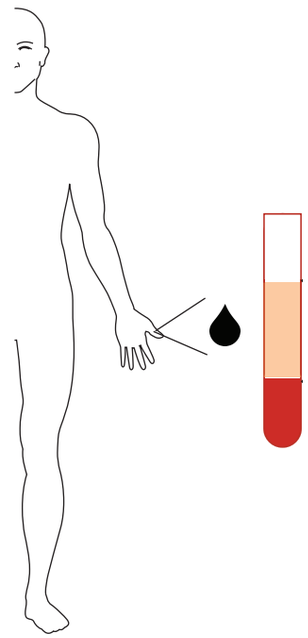
Cluster (2): Enriched in source with high load of *Curprividius,* which has no body site association (literature says it may grow in blood, so possible blood commensal).



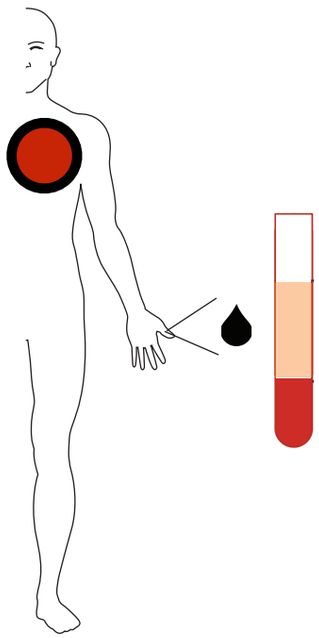
Dominant sources found in pregnancy cohort.

ICA for anomaly detection in cfDNA microbiome data.

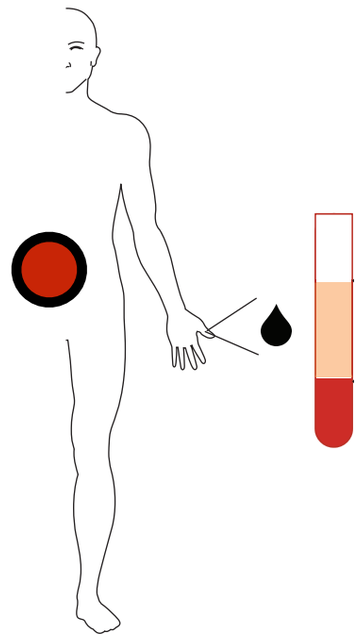
Healthy -



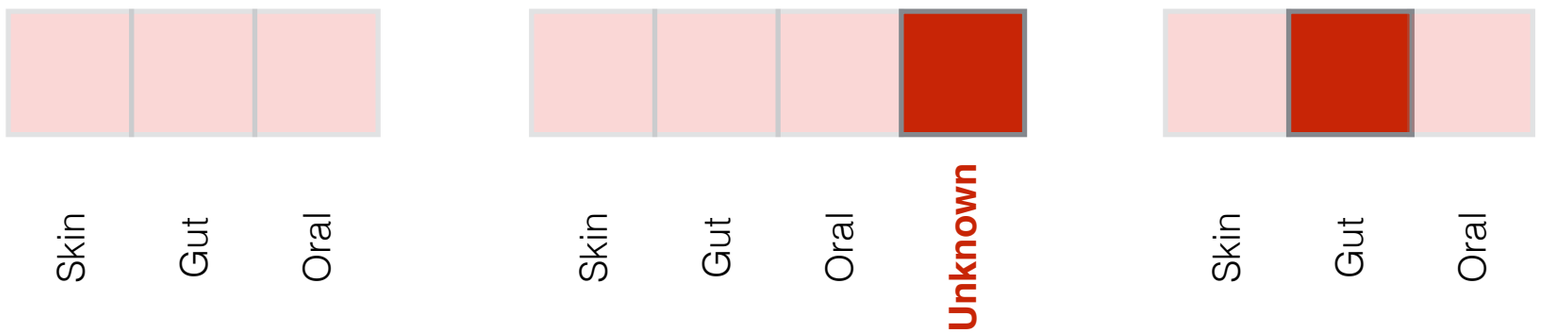
Deep tissue infection -



Inflammatory bowel disease -

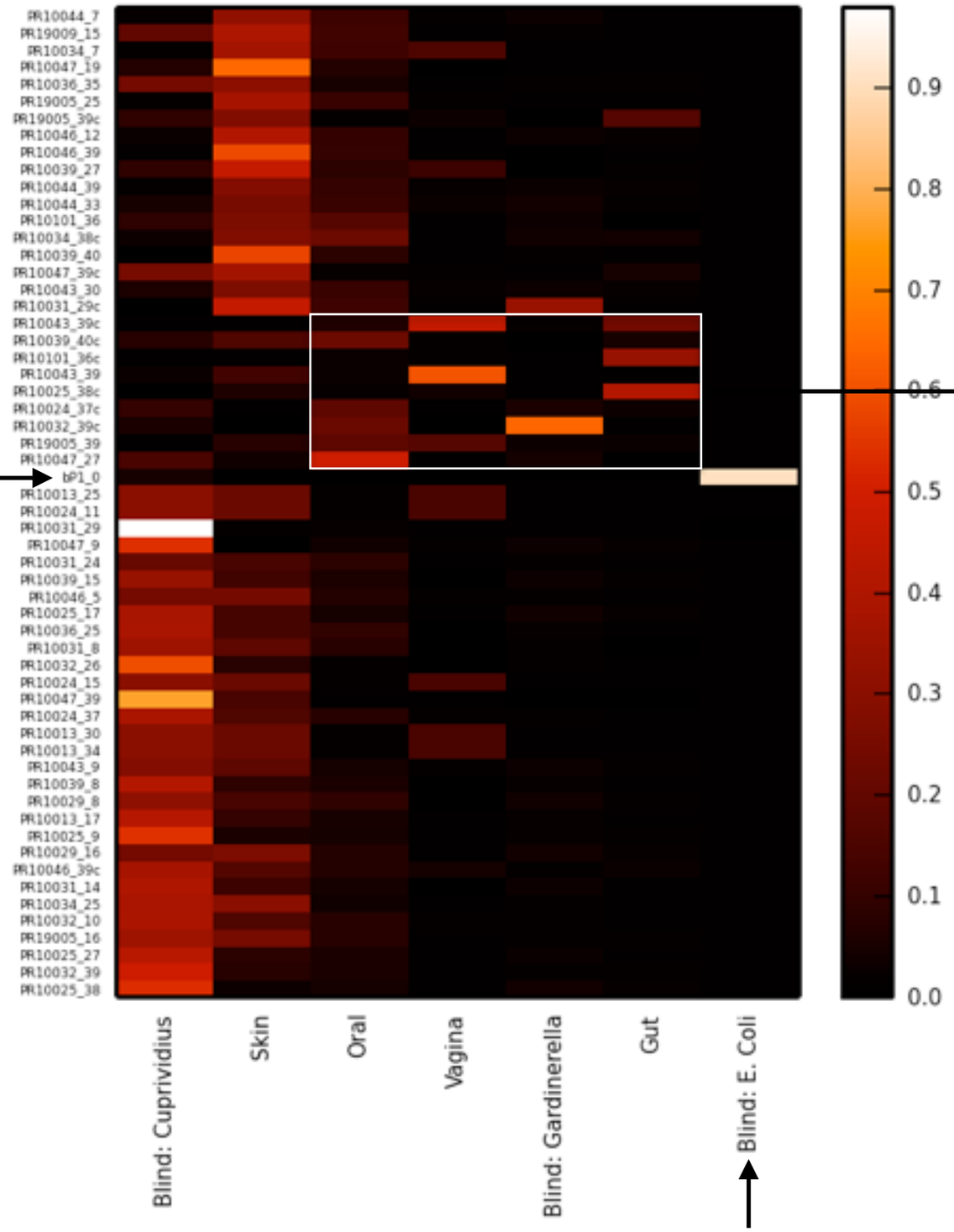


Relative weight
of ICA sources:



ICA mixing matrix can pinpoint anomalies.

(1) Spike in control of deep tissue infection with high E. Coli load.



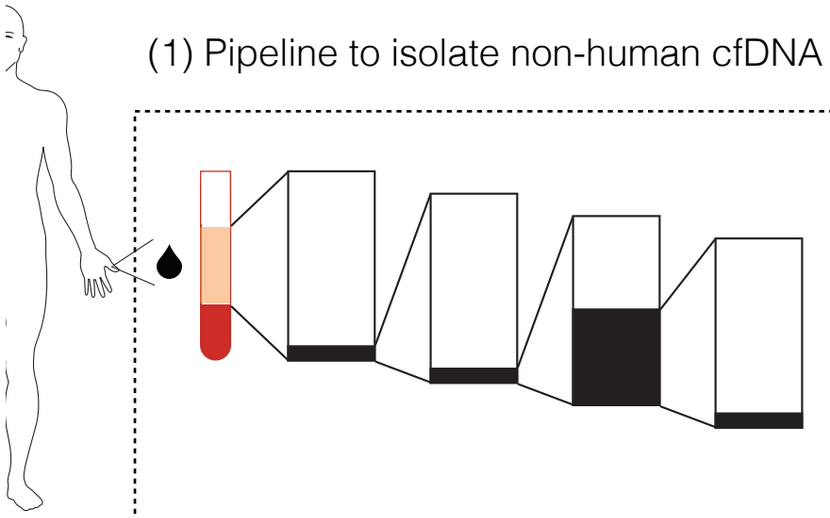
(2) Pregnancy cohort anomalies: uncommon mixing profiles.

(1) Source rich in E. Coli discovered and weighted only in this patient.

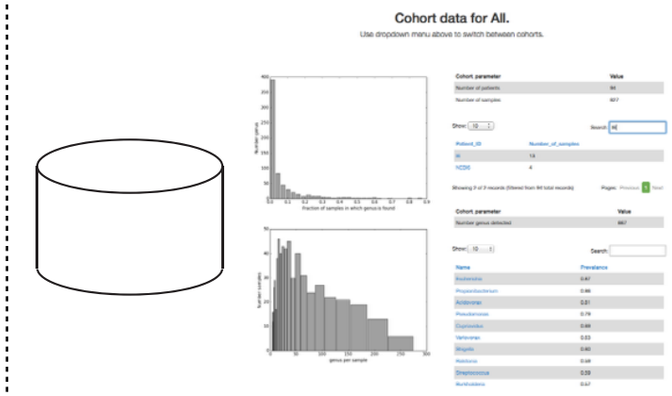
Summary

Pipeline

(1) Pipeline to isolate non-human cfDNA

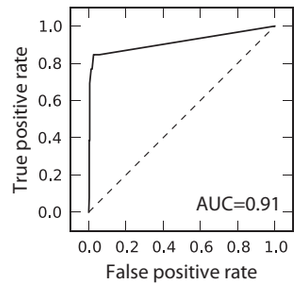


(2) Browser for visualization



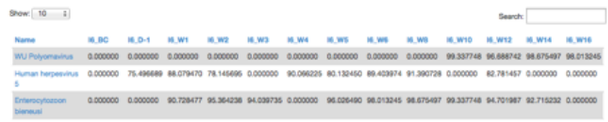
Applications

(3) Clinical studies: Lung, BMT

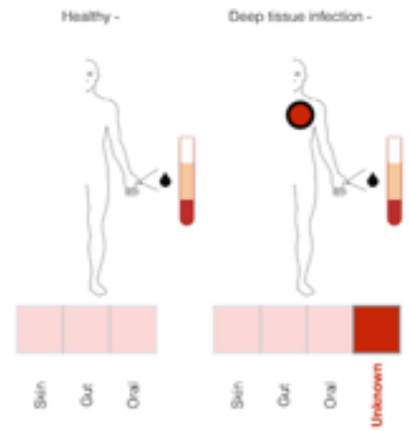


(4) Biopsy replacement
- Stanford pathology

(5) Undiagnosed infections
- I6, WU Case study



(6) Body site composition and anomalies
- Pregnancy cohort



Thanks

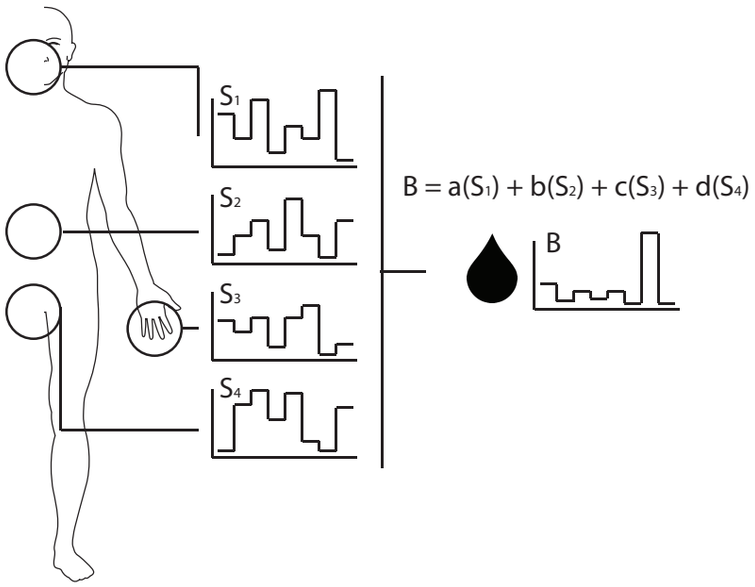
Funding

NDSEG
NATIONAL DEFENSE SCIENCE & ENGINEERING GRADUATE FELLOWSHIP

march of dimes

People

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.

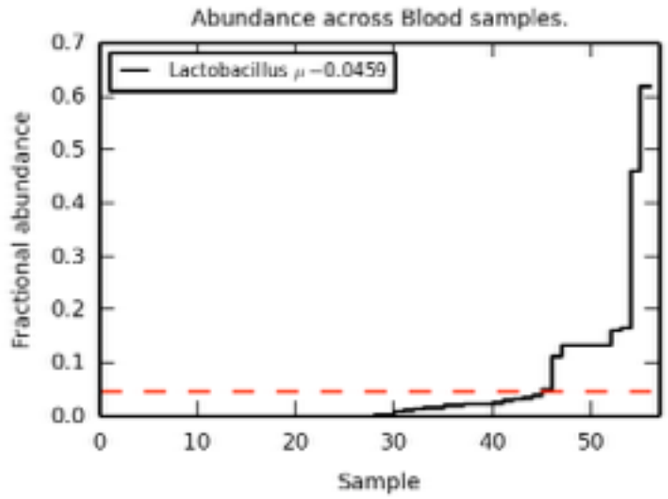
$$Y = \frac{1}{1 + e^{-\theta^T x}}$$

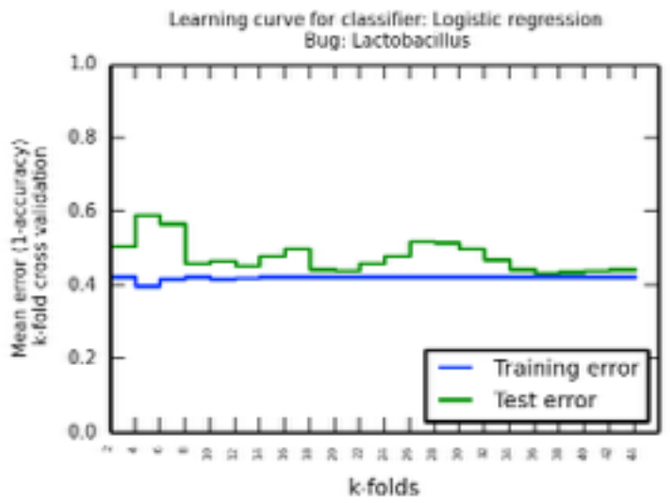
- θ: set of tissue weights that are learned for a specific bug.
- x: vector of bug fractional abundances per tissue in the given sample.
- Y: label indicating whether a bug is found in blood in the given sample.

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

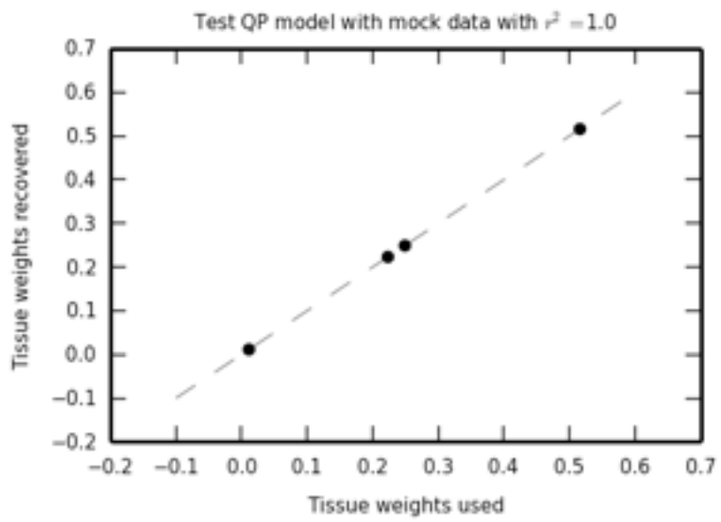
- θ: a vector of tissue weights that are learned for each sample.
- x_j: The vector of bugs measured at site j.
- $\vec{B}$: a vector of bug abundance in blood for a paritcular sample.

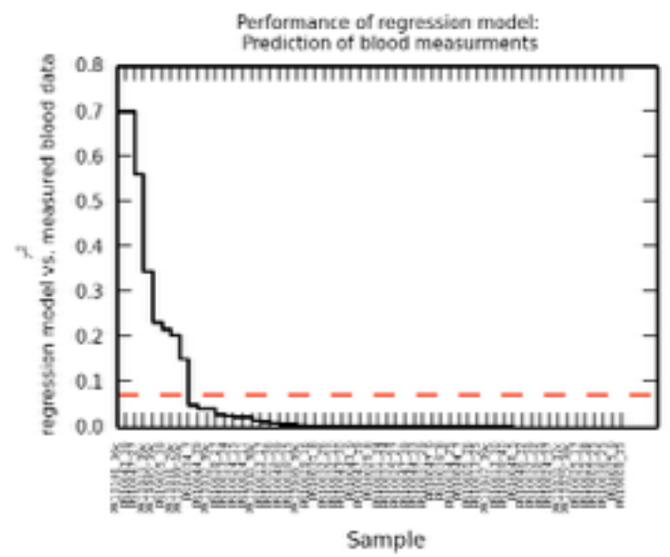
(B) **Consider two linear models**. (i) Classification: Pick one bug, discretize its detection in blood, learn weights for tissue sources that predict whether it is detected in blood. (ii) Regression: Pick one sample. Learn tissue weights that best describe the observed vector of bugs in blood using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





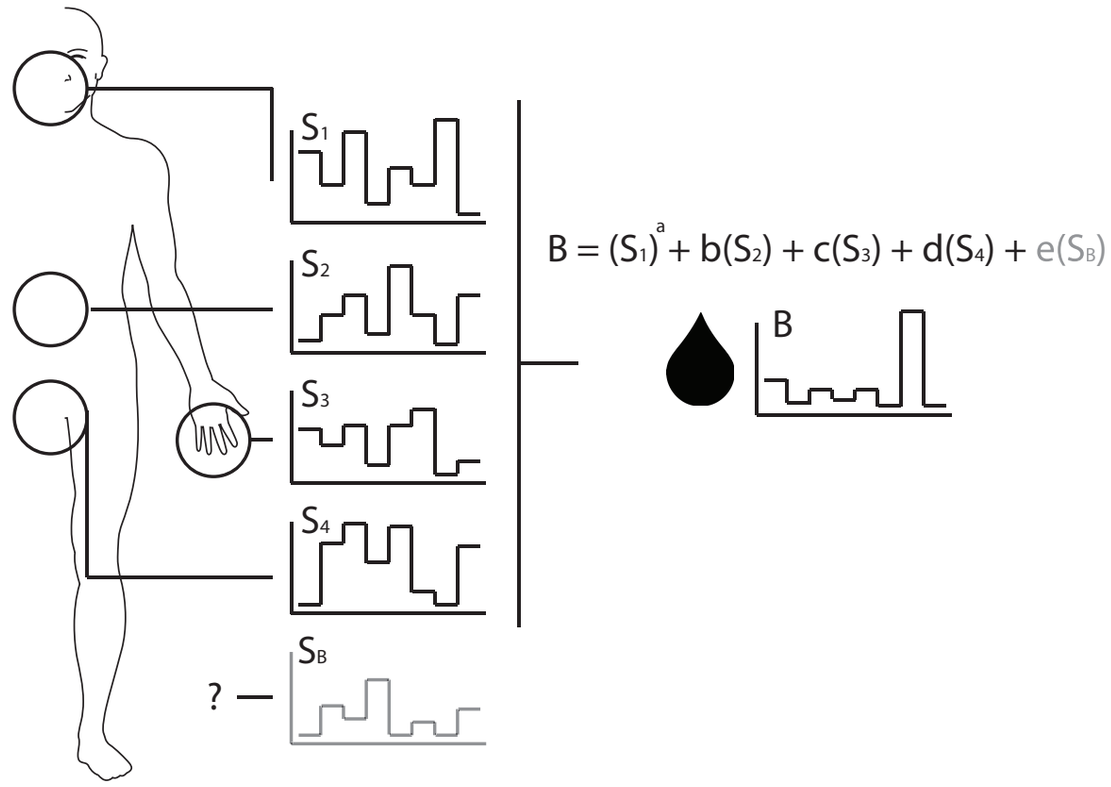
(C) **Classification with logistic regression performs poorly.** We pick a bug (Lactobacillus) with even labeling (detection) in blood (to avoid trivial model results) and compute a learning curve to evaluate whether the model.



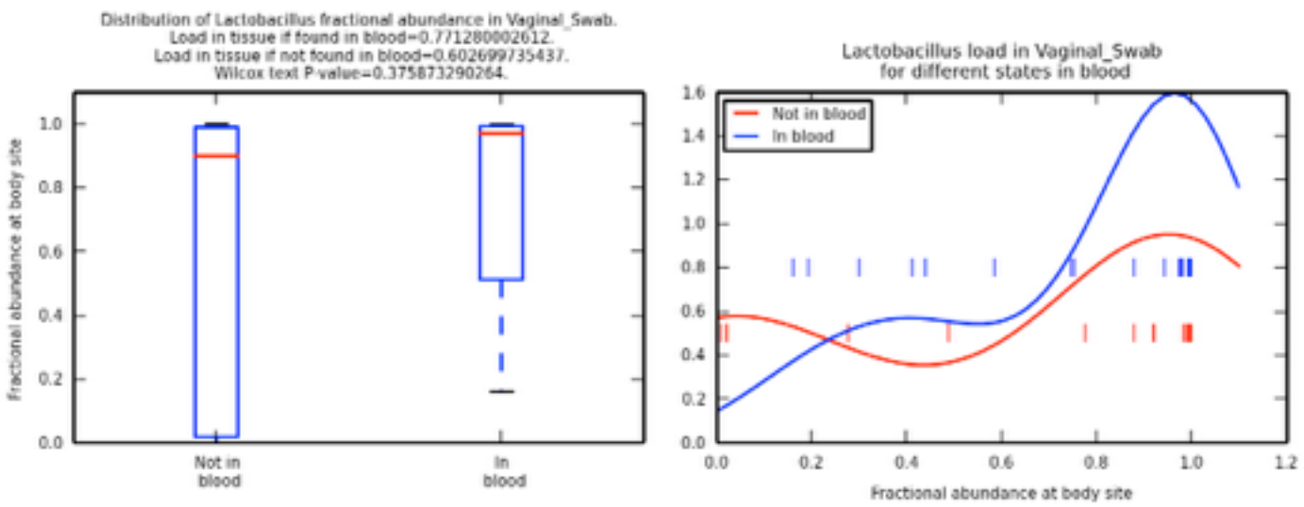


(D) **Linear regression performs poorly.** We apply linear regression with sensible constraints on the learned tissue weights (e.g., > 0). We use a quadratic and show that a test case with simulated data using tissue weights (x) shows that these weights are recovered (y) correctly. We then applied the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

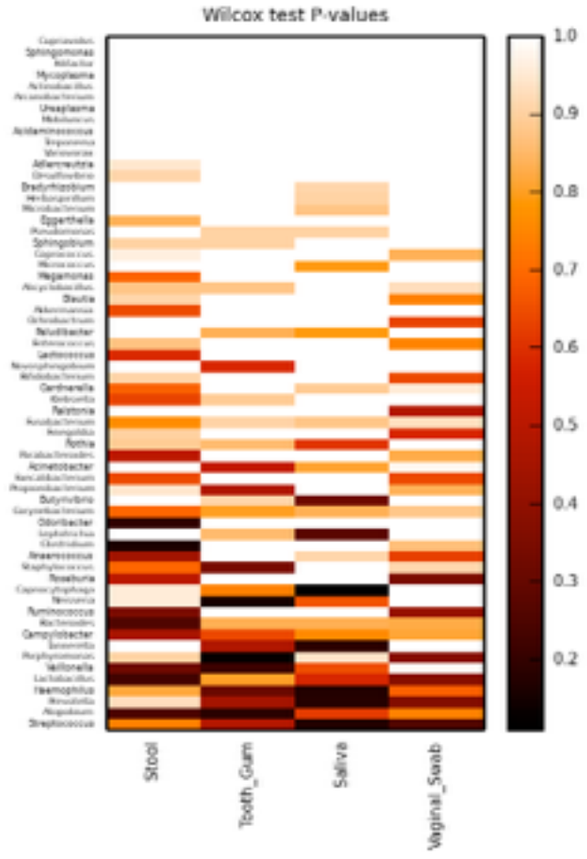
Examine why linear models doesn't work.



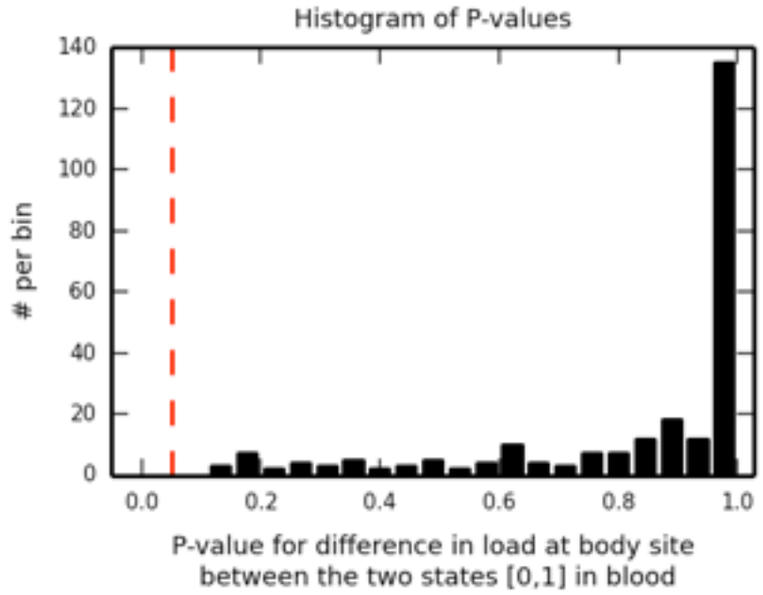
(A) **Blind sources and non-linearity frustrate performance of linear models.** Learning curve from logistic regression indicated high bias (underfit model), suggesting non-linear relationships. Residuals from linear regression show that blood samples contain some bugs that are not found in any sampled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



(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?



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



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

ICA setup

Blood mixtures, x (M by k):

$$\begin{bmatrix} x_1 \\ \dots \\ x_M \end{bmatrix}$$

- M blood mixtures.
- Each has k bugs measured.

Mixing matrix, A (M by n):

$$\begin{bmatrix} a_{11} \dots a_{1n} \\ \dots \\ a_{M1} \dots a_{Mn} \end{bmatrix}$$

- For each of M blood mixtures, it provides a set of mixing terms.
- Each mixing term is assocaited with one of the N sources.
- Each mixture in X is thus a linear combination given by: $x_j = a_{j1}s_1 + \dots + a_{jn}s_n$

Source matrix, S (n by k):

$$\begin{bmatrix} s_1 \\ \dots \\ s_n \end{bmatrix}$$

- These are the n independent components of the model.

The inner product of these terms can be written:

- The common set of sources are combined as specified by the mixing matrix to give each.
- Where each mixture, x_j, and source, s_j, have the same dimentionality (k).

$$\begin{bmatrix} x_1 \\ \dots \\ x_M \end{bmatrix} = \begin{bmatrix} a_{11} \dots a_{1n} \\ \dots \\ a_{M1} \dots a_{Mn} \end{bmatrix} \begin{bmatrix} s_1 \\ \dots \\ s_n \end{bmatrix}$$

Our goal is to find the unmixing matrix that allows us to recover the source signals:

$$W = A^{-1}$$

ICA derivation

Derivation -

Strategy:

- We want to use likelihood estimation, meaning that we will find W that maximizes the likelihood of our observed signals.
- If we can define a likelihood funtion, then we can find the value of W that maximizes it using numerical methods (e.g., gradient ascent).

The likelihood funtion takes the form of joint probability of observing our mixtures, s, given W:

$$l(W) = \prod_{i=1}^{m} p(x \mid W)$$

So, we need to find $p(x \mid W)$.

Let's first recall that:

$$x = As$$

We can start by defining the joint probability distribution of our n sources:

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

But, we want to re-write this in terms of x, our mixtures.

Recall a few rules:

- s is a vector-valued distribution with density p_s and $x = As$
- Then, $p_x(x) = p_s(Wx)|W|$

Thus, we can re-write the joint probability distribution in terms of x:

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

$$p(x) = \prod_{i=1}^{n} p_x(x)$$

$$p(x) = \prod_{i=1}^{n} p_s(W_i x)|W|$$

ICA derivation (cont.)

Now, we have written the joint probability distribution in terms of x and W.

But, we still need to know p_s, which is the probability density of our sources.

- First, let's just assume the sources have a CDF that slowly increases from 0 to 1. It cannot be Gaussian.
- Let's use the sigmoid funtion, $g(s) = \frac{1}{1+e^{-s}}$.
- Then, we know that the density is simply the derivative of the CDF.
- So, $p_s = g(s)'$.

$$p(x) = \prod_{i=1}^{n} g'(W_i x)|W|$$

Thus, the log likelihood is defined across our m training examples:

$$l(w) = \sum_{i=1}^{m} \sum_{i=1}^{n} log[g'(W_i x)] + log|W|$$

Compute partial derivative with respect to W:

$$\frac{\partial}{\partial W} l(w) = \frac{\frac{\partial}{\partial W} g'(W_i x)}{g'(W_i x)} + \frac{\nabla_W |W|}{|W|}$$

Use these rules:

- For the sigmoid funtion: $g'(z) = g(z)(1 - g(z))$
- $\frac{\partial}{\partial x} log(f(x)) = \frac{f'(x)}{f(x)}$
- $\nabla_W |W| = |W|(W^{-1})^T$

The first term numerator:

$$\frac{\partial}{\partial W} g'(W_i x) = \frac{\partial}{\partial W} g(W_i x)(1 - g(W_i x))$$
$$= \frac{\partial}{\partial W} g(W_i x) - g(W_i x)^2$$
$$= [\frac{\partial}{\partial W} (W_i x) - 2g(W_i x) \times \frac{\partial}{\partial W} (W_i x)]x$$
$$= [1 - 2g(W_i x)] \frac{\partial}{\partial W} (W_i x)x$$

Divide through by $g'(W_i x)$ in the denominator to get:

$$\frac{\frac{\partial}{\partial W} g'(W_i x)}{g'(W_i x)} = [1 - 2g(W_i x)]x$$

Alost, the second term is:

$$\frac{\nabla_W |W|}{|W|} = \frac{|W|(W^{-1})^T}{|W|} = (W^{-1})^T$$

Therefore, the derivative of the likelihood function is:

$$\frac{\partial}{\partial W} l(w) = [1 - 2g(W_i x)]x + (W^{-1})^T$$

Gradient ascent rule can be used to interavly find W:

$$W := W + \alpha \frac{\partial}{\partial W} l(w)$$
$$W := W + \alpha([1 - 2g(W_i x)]x + (W^{-1})^T)$$