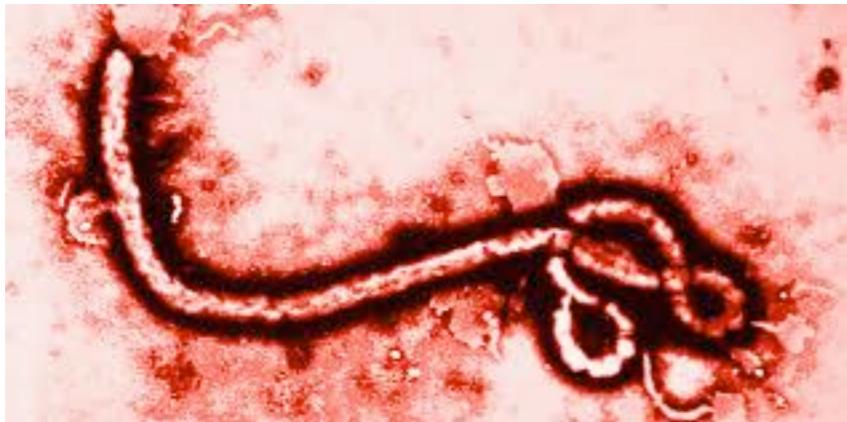
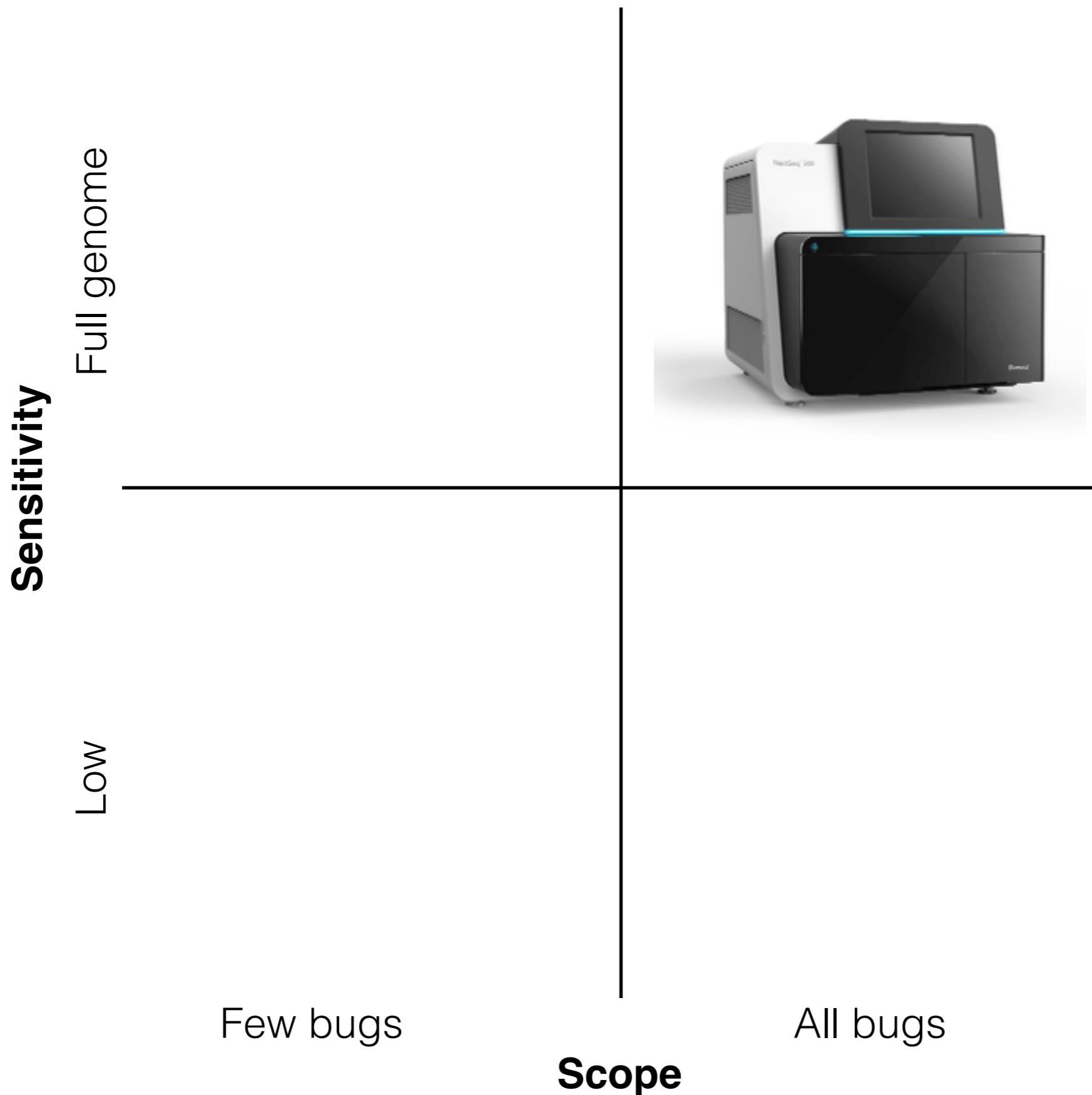


Monitoring the human infectome

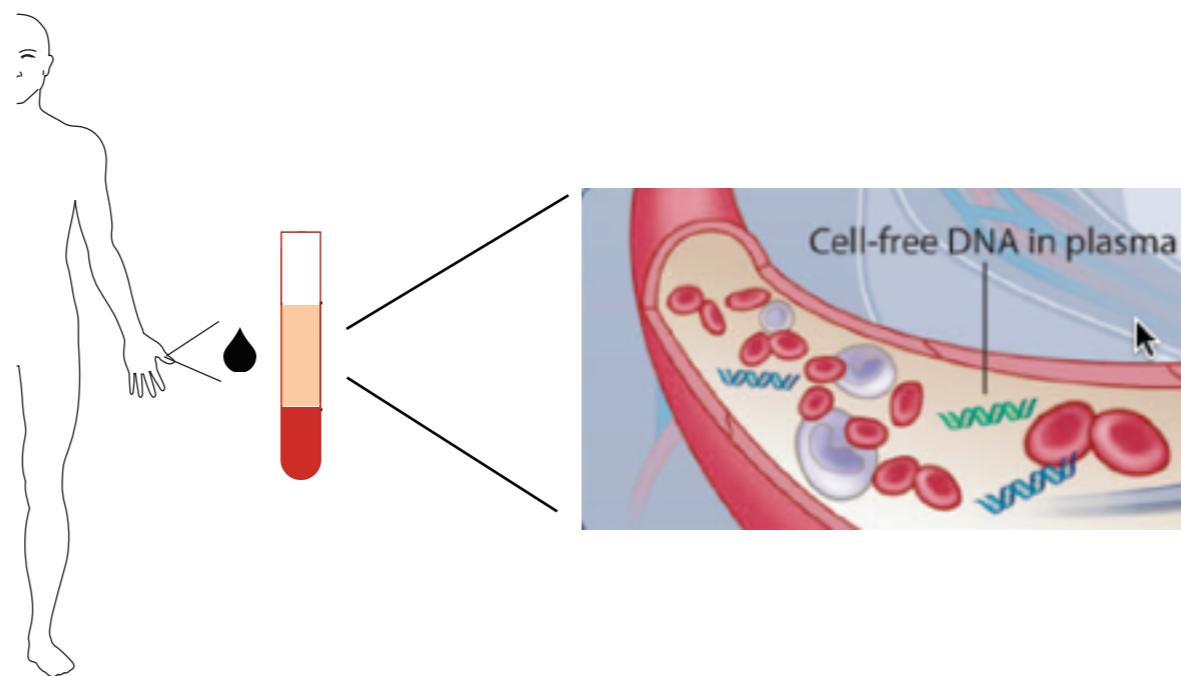
Lance Martin
Quake GM, 12/9/14



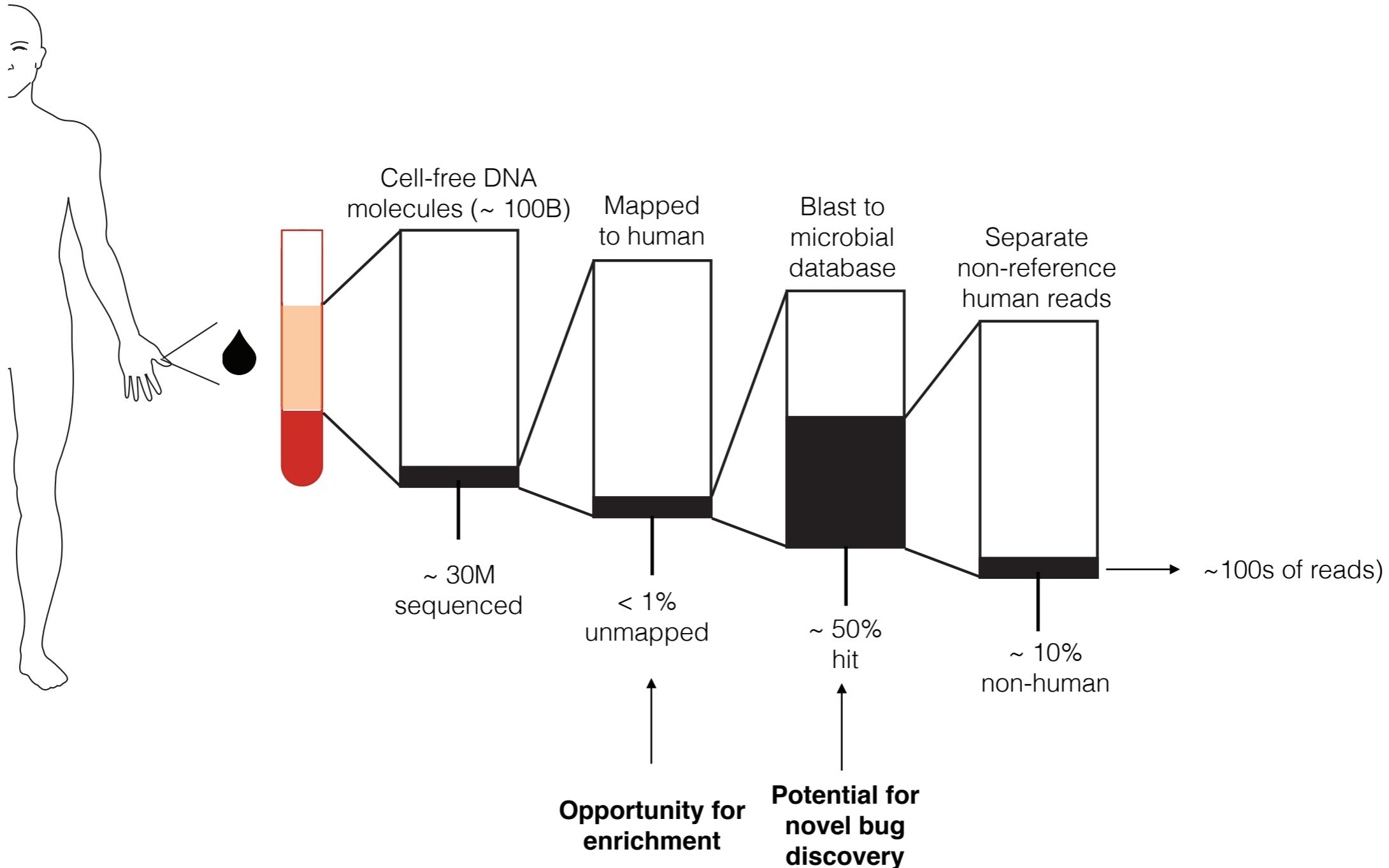
Non-invasive detection of any pathogen.



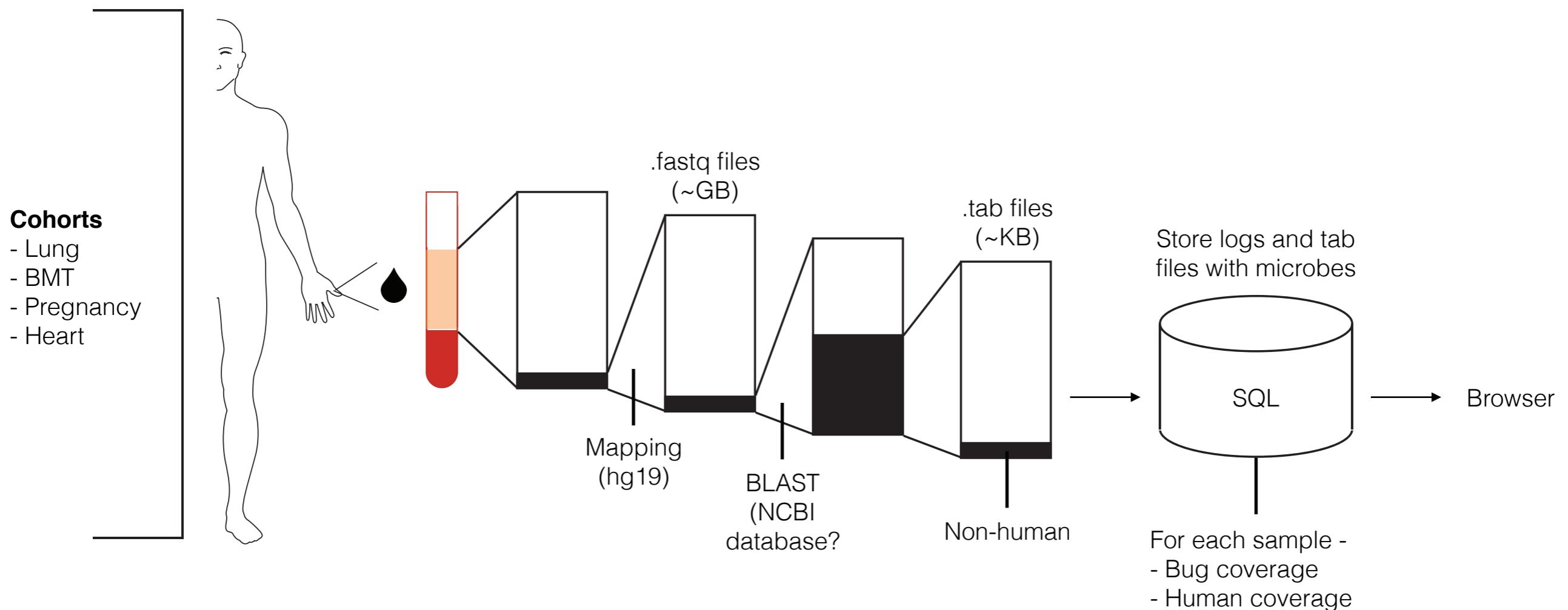
Cell-free DNA now used for non-invasive diagnostics.



Human microbiome in cell-free DNA.

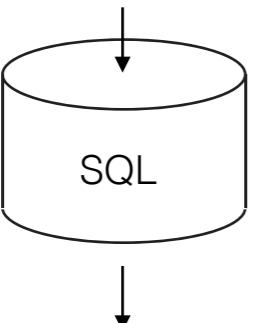


Pipeline for isolating microbial reads in cell-free data.



Browser for visualization and navigation.

1000s of samples



Infectome Explorer Cohorts - Taxonomic level -

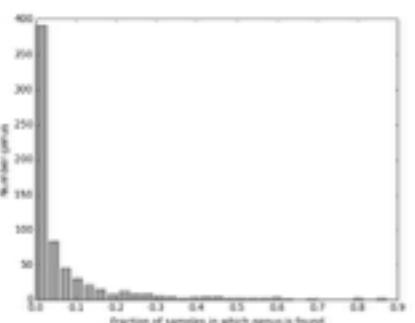
Welcome to the Infectome Explorer.

Choose a cohort:

All cohorts SMT Lung Pregnancy Ebola Stanford clinic Biopsy Pregnancy RNA CPS

Cohort data for All.

Use dropdown menu above to switch between cohorts.



Cohort parameter Value

Number of patients: 94

Number of samples: 807

Show: 10 Search: 86

Patient ID: 86 Number_of_samples: 16

Number of samples: 4

Showing 2 of 2 records (filtered from 807 total records)

Pages: Previous Next

Cohort parameter Value

Number genera detected: 867

Show: 10 Search:

Name Prevalence

Kinolobius 0.67

Propionibacterium 0.66

Acidovorax 0.61

Pseudomonas 0.79

Cupriavidus 0.68

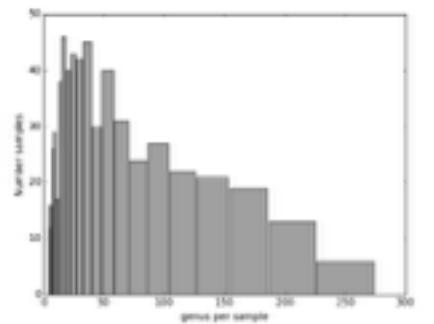
Verinimonas 0.53

Sphaerotilus 0.60

Methylophilus 0.68

Sphaerotilus 0.59

Methylophilus 0.67

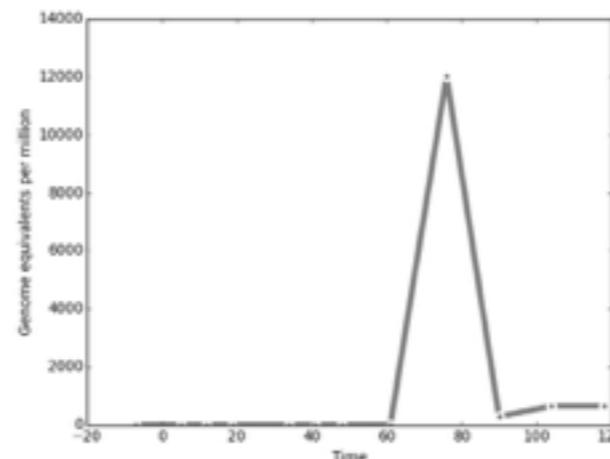


Infection timeseries for Polyomavirus in I6.

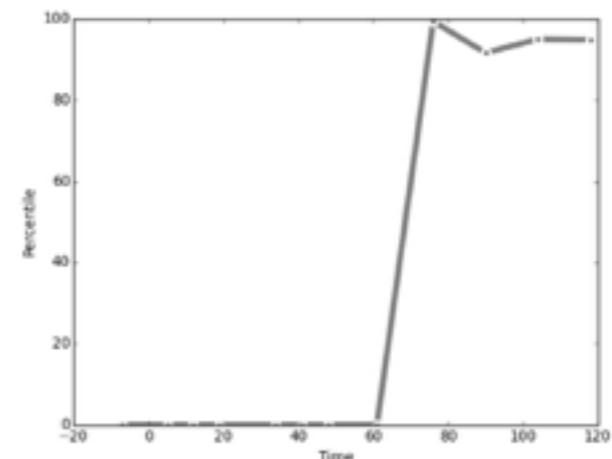
Use back to return to cohort or dropdown menu above to switch between cohorts.

W Polymaviridae: Polyomaviruses are DNA-based (double-stranded DNA, ~5000 base pairs, circular genome) viruses. They are small (40–50 nanometers in diameter), and icosahedral in shape, and do not have a lipoprotein envelope. Moreover, the genome possess early and late genes, contributing to its complex transcription program. They are potentially oncogenic (tumor-causing); they often persist as latent infections in a host without causing disease, but may produce tumors in a host of a different species, or a host with an ineffective immune system. The name polyoma refers to the viruses' ability to produce multiple (poly-) tumors (-oma). The family Polyomaviridae used to be one of two genera within the now obsolete family Papovaviridae (the other family being Papillomaviridae). The name Papovaviridae derived from three abbreviations: Pa for Papillomavirus, Po for Polyomavirus, and Va for "vacuolating". Clinically, Polyomaviridae are relevant as they contribute to pathologies such as Progressive multifocal leukoencephalopathy (PML virus), nephropathy (BK virus), and Merkel cell cancer (Merkel cell virus). Until recently, the family of Polyomaviridae contained only one genus (Polyomavirus). The recent expansion of known Polyomaviruses called for reclassification of the family into 3 genera: Orthopolyomavirus, Wukipolyomavirus, and Avipolyomavirus. Murine polyomavirus was the first polyomavirus discovered by Ludwik Gross in 1953. Subsequently, many polyomaviruses have been found to infect birds and mammals. For nearly 40 years, only two polyomaviruses were known to infect humans. Genome sequencing technologies have recently discovered seven additional human polyomaviruses, including one causing most cases of Merkel cell carcinoma and another associated with transplant-associated dysplasia (TSV), that are natural infections of humans. Discovery of these polyomaviruses in humans and animals, leading to fundamental insights into carcinogenesis, DNA replication and protein processing. The tumor suppressor molecule p53 was discovered, for example, as a cellular protein bound by the major oncoprotein (cancer-causing protein) T antigen made by Simian vacuolating virus 40 (SV40). The avian polyomavirus sometimes referred to as the Budgerigar fledgling disease virus is a frequent cause of death among caged birds.

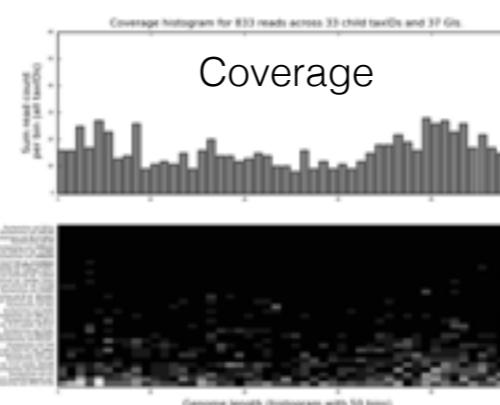
Infection load for patient



Percentile (relative to all samples)

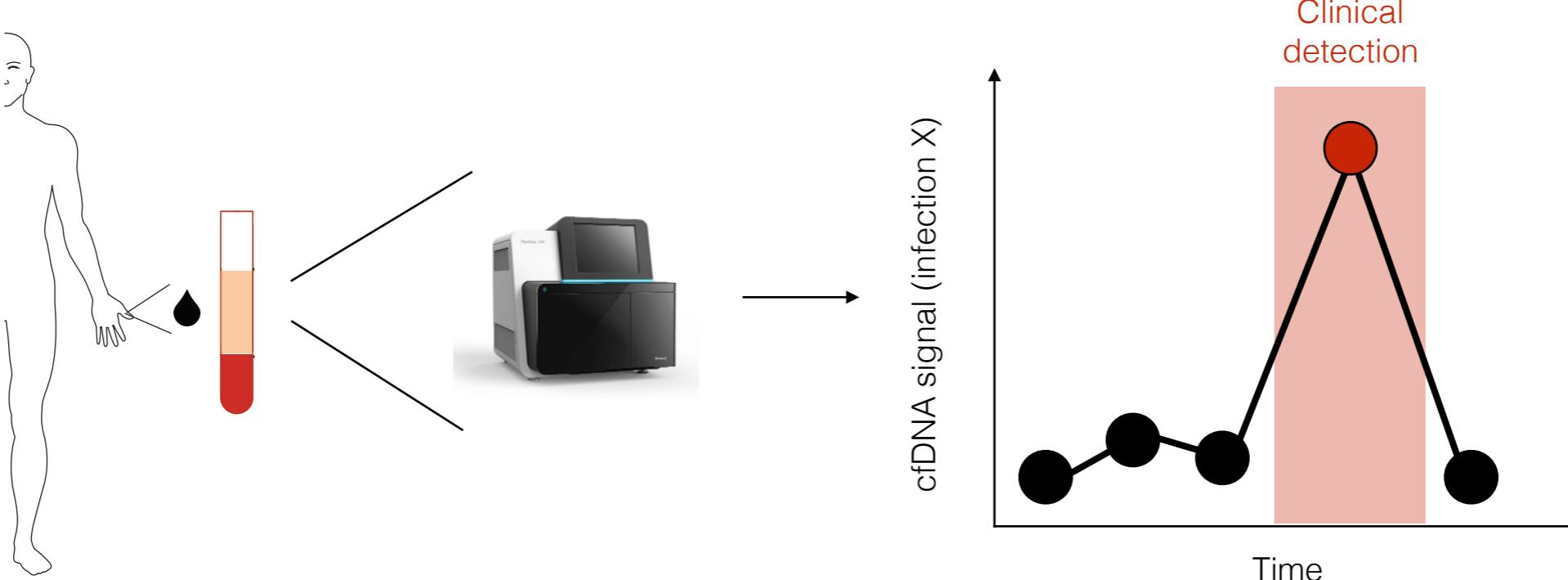


Coverage



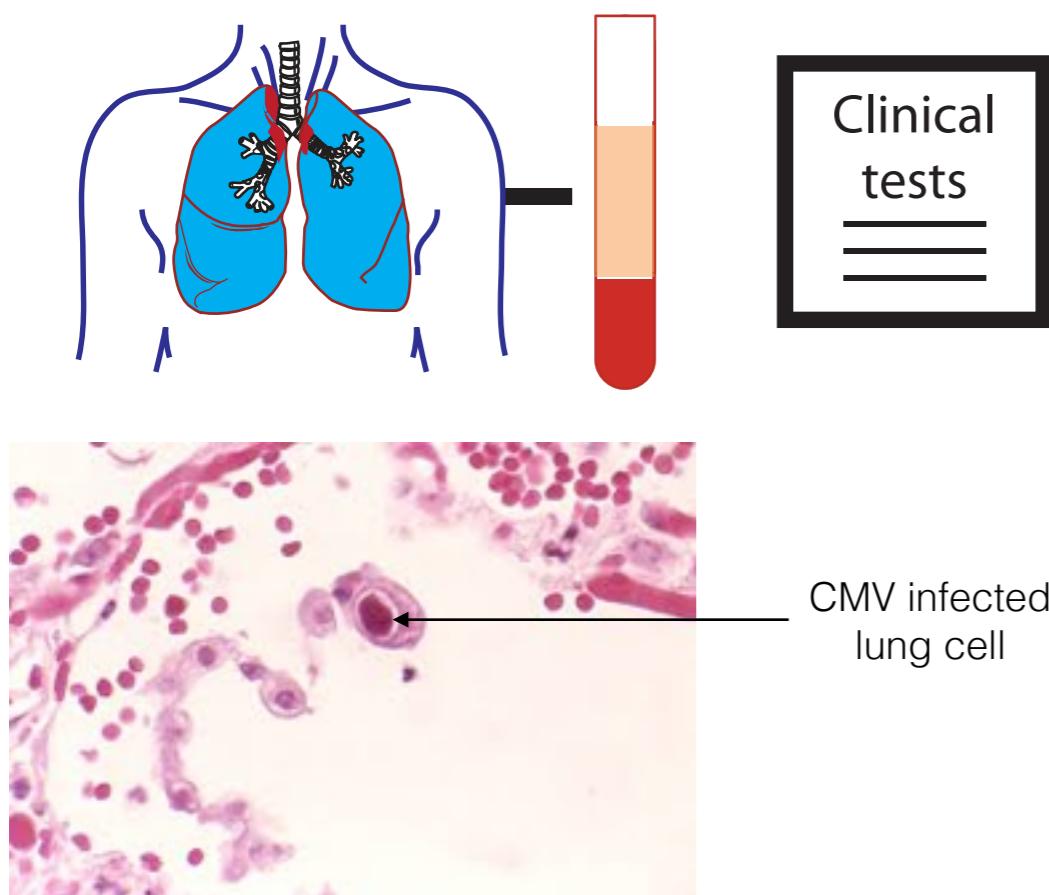
ID	Days	Date	Total Blast	C21 Coverage	Blast for inf
I6_BC	-8	2013-10-23 00:00:00	235	1.590626	27.628748
I6_D-1	-7	2013-10-24 00:00:00	379	1.001016	57.134878
I6_W1	5	2013-11-05 00:00:00	60	1.068258	22.686783
I6_W2	12	2013-11-12 00:00:00	48	0.987716	23.581704
I6_W3	19	2013-11-19 00:00:00	45	0.817058	28.274749
I6_W4	34	2013-12-04 00:00:00	65	0.893145	35.540858
I6_W5	41	2013-12-11 00:00:00	85	1.062171	45.917179
I6_W6	48	2013-12-18 00:00:00	299	1.190450	45.601895
I6_W8	61	2013-12-31 00:00:00	678	1.061075	49.335448
I6_W10	76	2014-01-15 00:00:00	807	0.878492	45.095884

But how to know if it's useful?

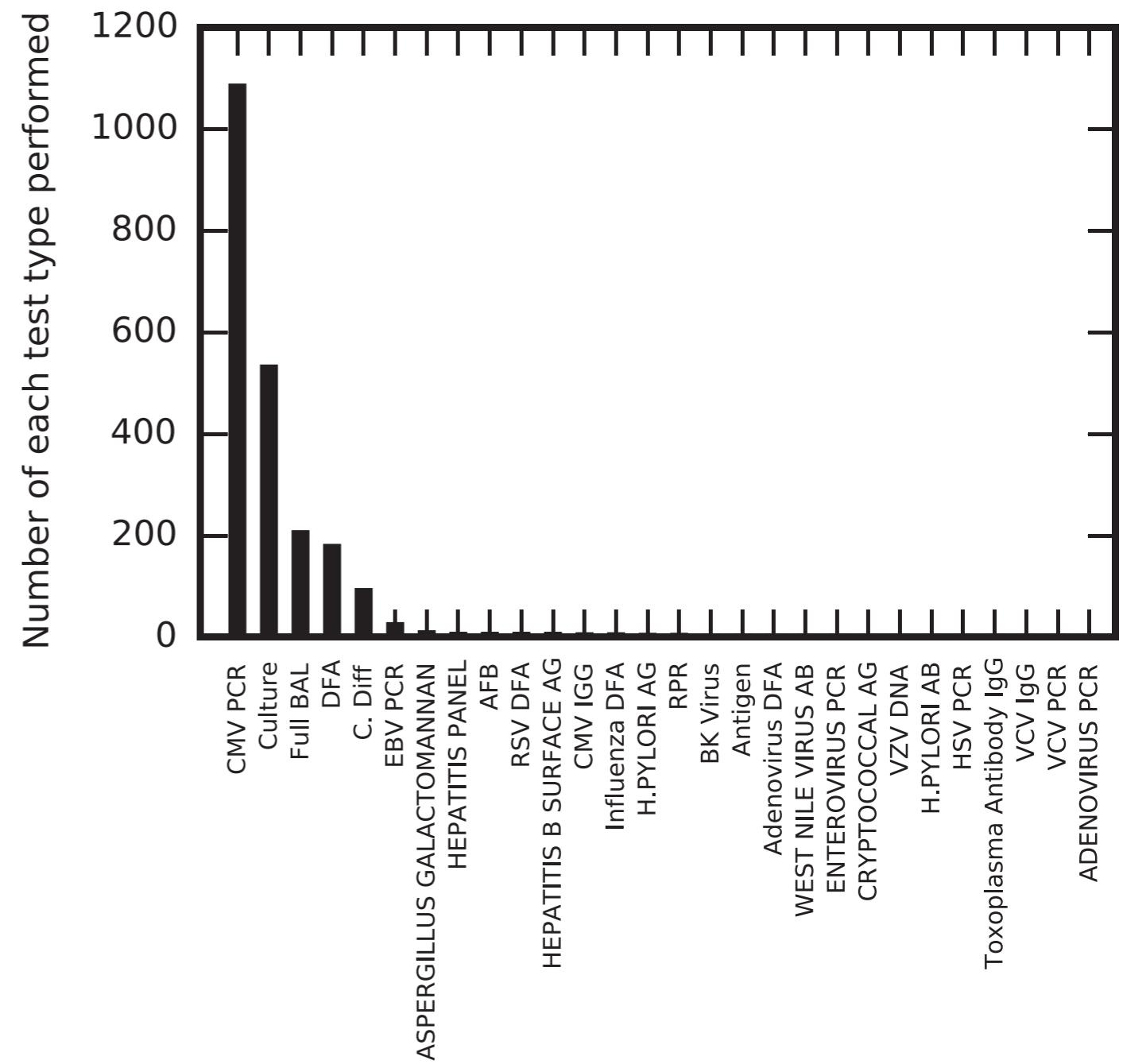


Look back at the clinical history of our cohorts.

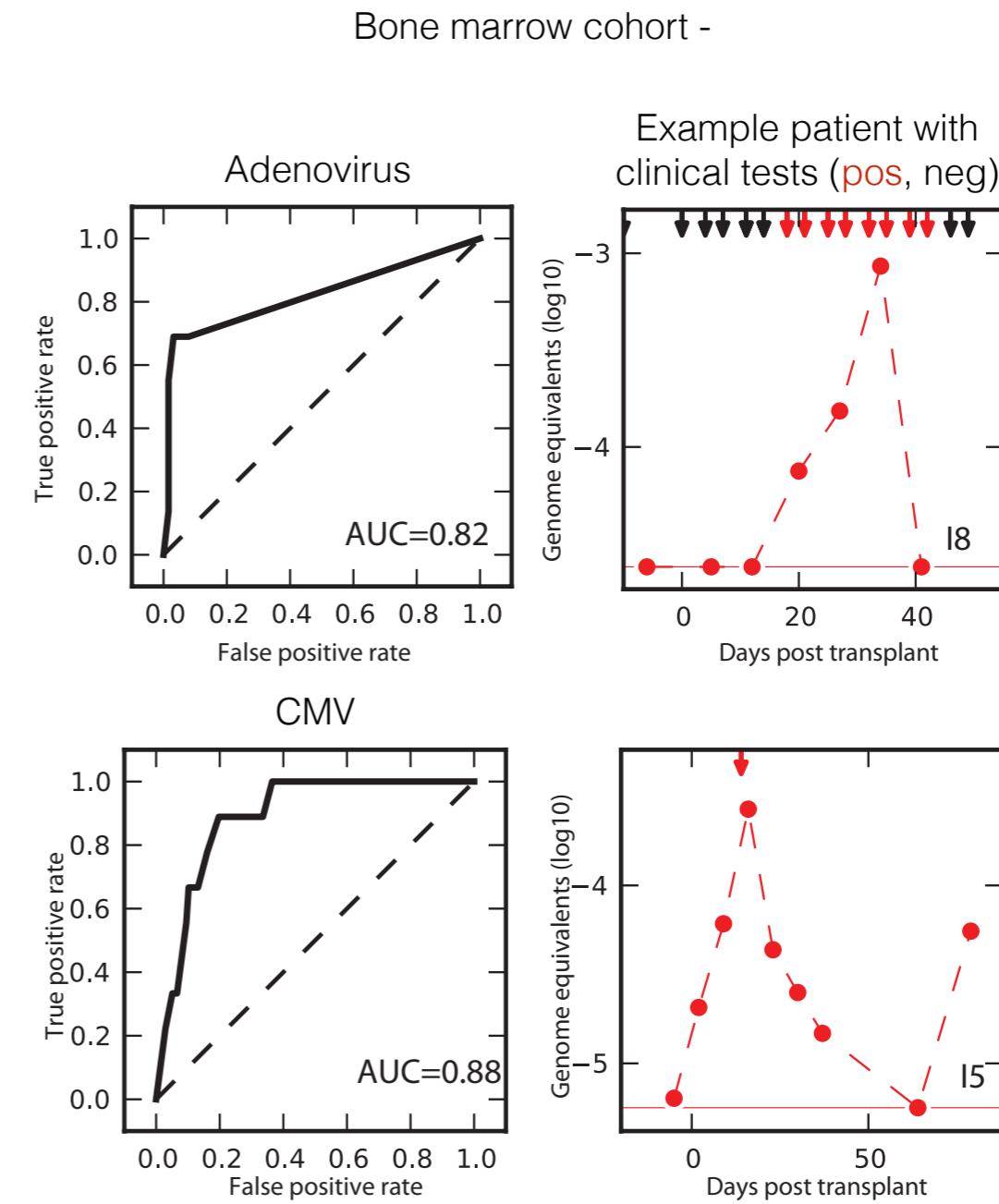
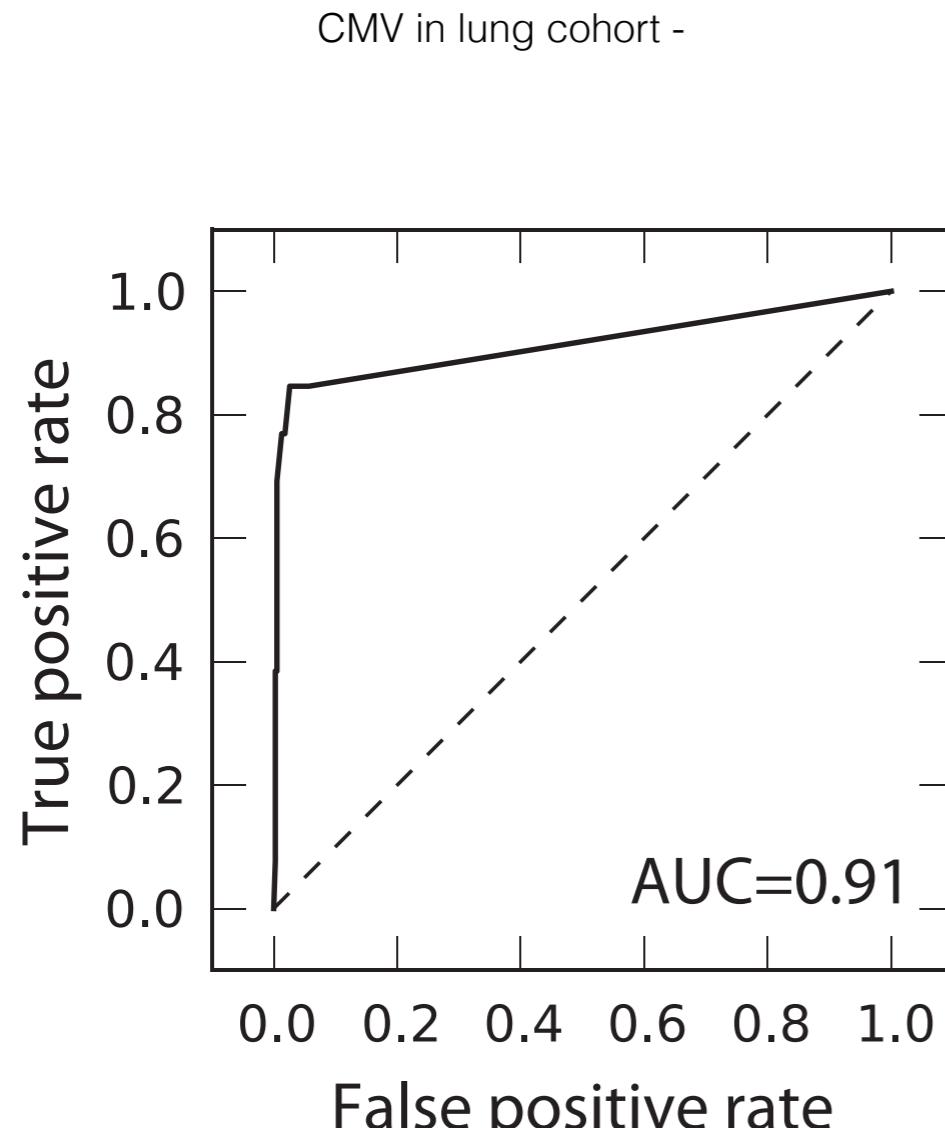
Lung Transplant (431 samples collected and processed) -



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



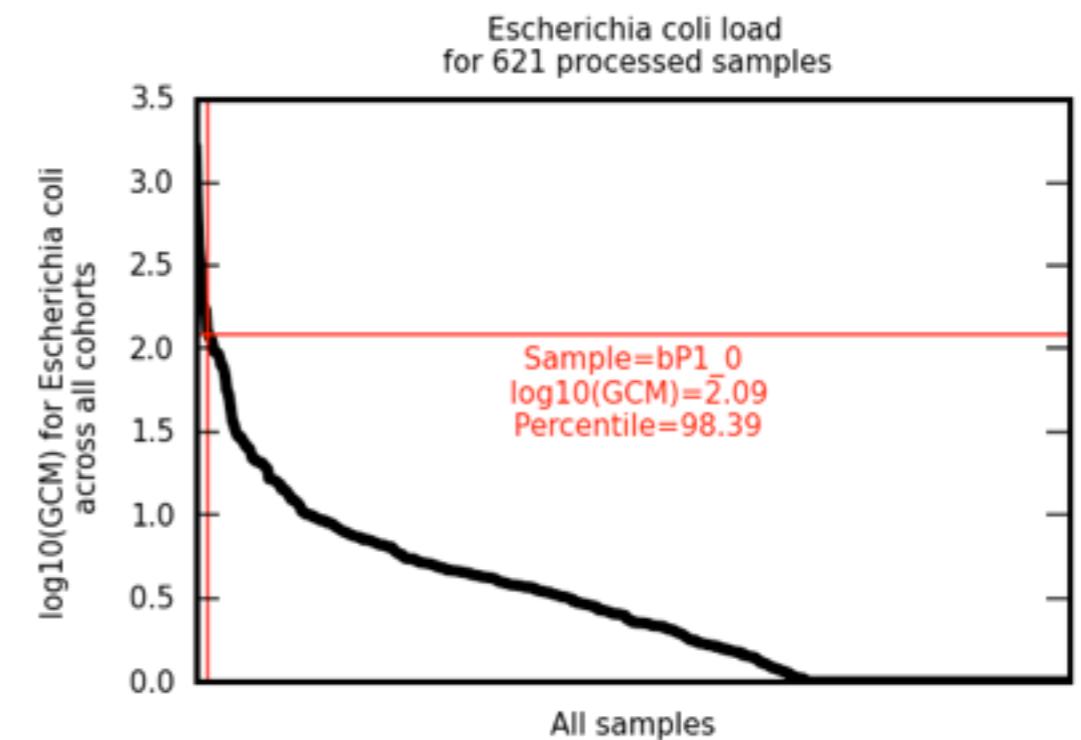
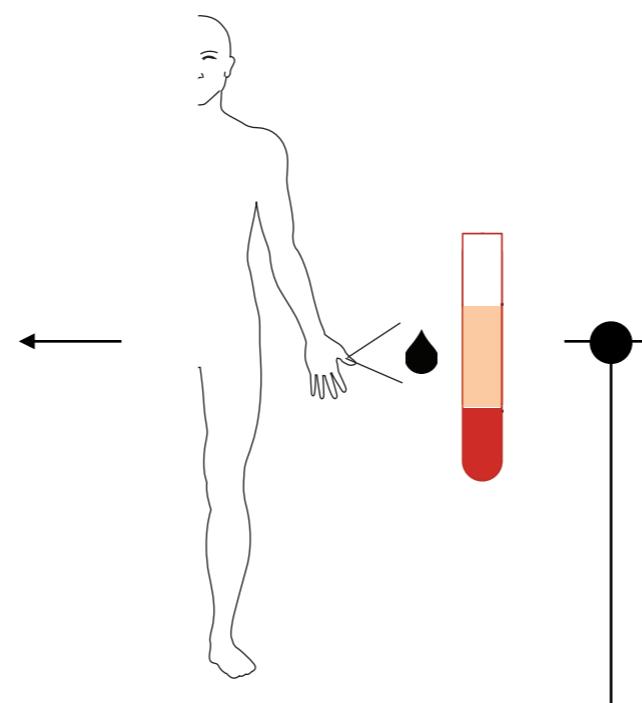
Performance on viruses.



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

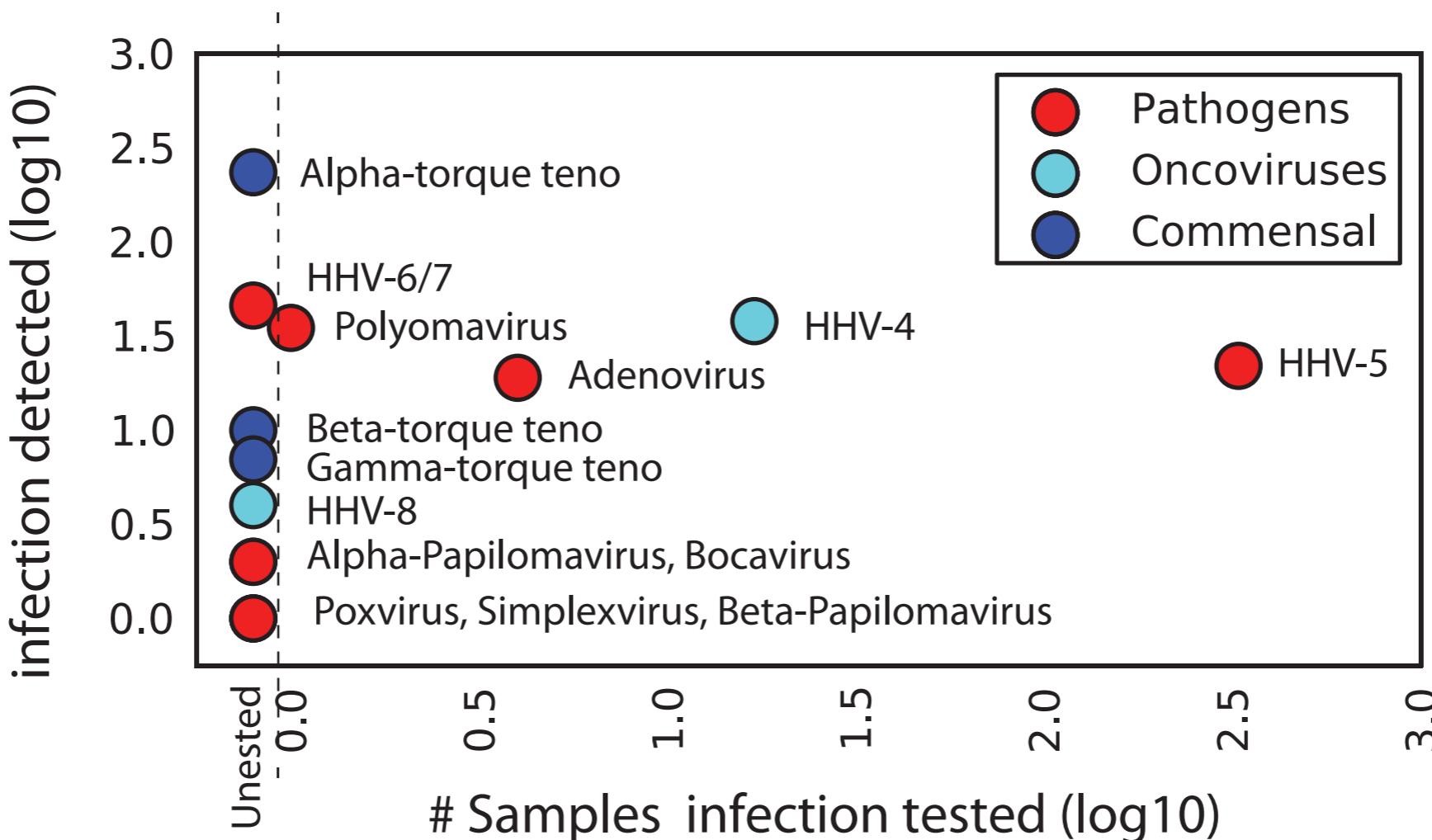
Gen_Equ

Percentile

Candidatus Midichloria	0.28	99.516908
P2likeviruses	34.09	98.711755
Gammatorquevirus	189.43	98.711755
Escherichia	123.17	98.380694
Alkaliphilus	0.11	98.228663
Shigella	8.45	96.940419
Chelatovorans	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.

Many potential pathogens that we detected are infrequently clinically tested for (viruses in lung cohort) -



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:

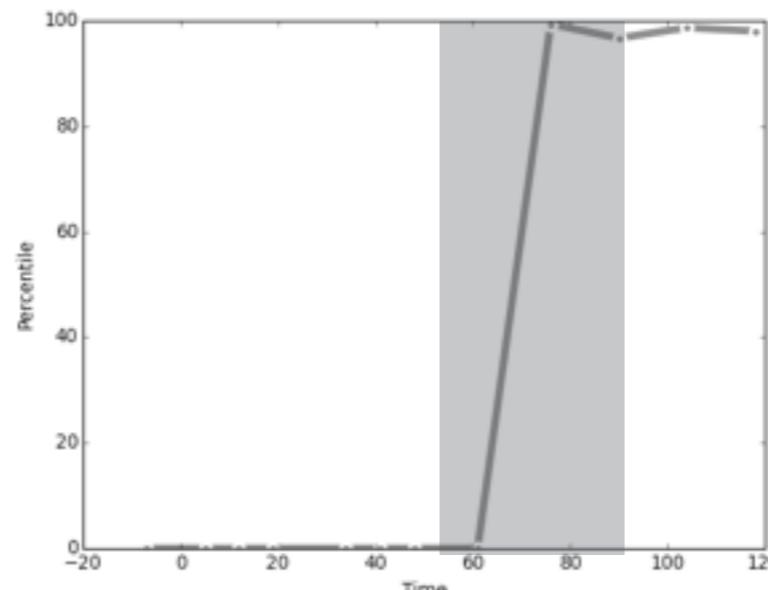
Name	I6_BC	I6_D-1	I6_W1	I6_W2	I6_W3	I6_W4	I6_W5	I6_W6	I6_W8	I6_W10	I6_W12	I6_W14	I6_W16
WU Polyomavirus	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	99.337748	96.688742	98.675497	98.013245
Human herpesvirus 5	0.000000	75.496689	88.079470	78.145699	0.000000	90.096225	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)

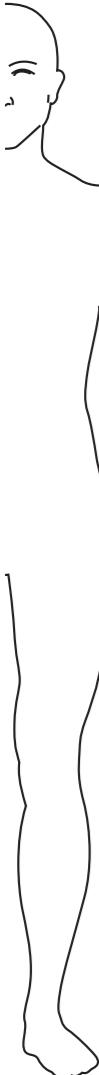


Very high load of WU Polyomavirus during time period of negative clinical test results.

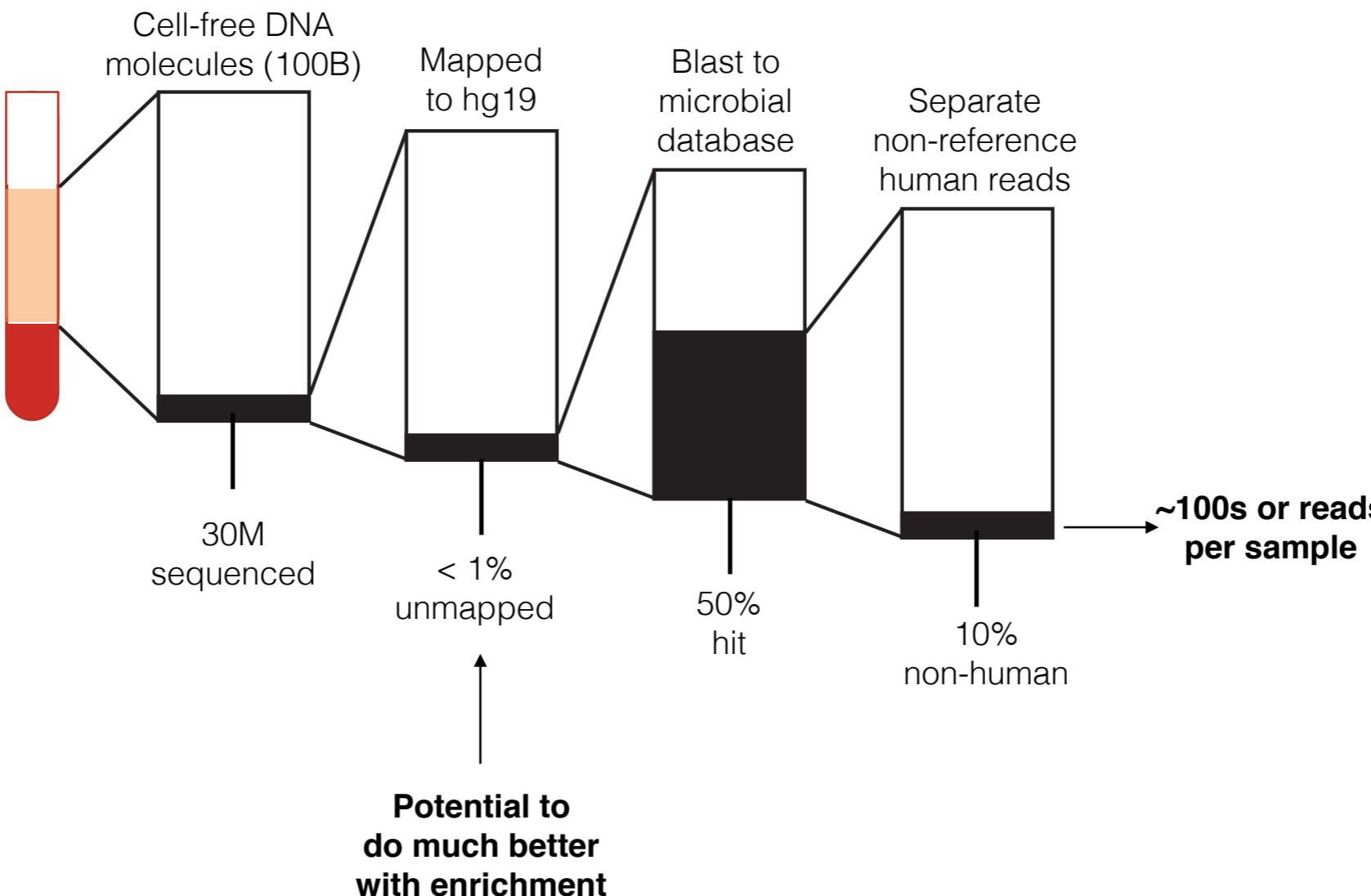


Multiple (-) tests for
BK polyomavirus

Also, headroom for improvement.



Favorable results shown with sparse sampling of microbial reads.



Clinical studies -

- Lung
- BMT

Deep tissues -

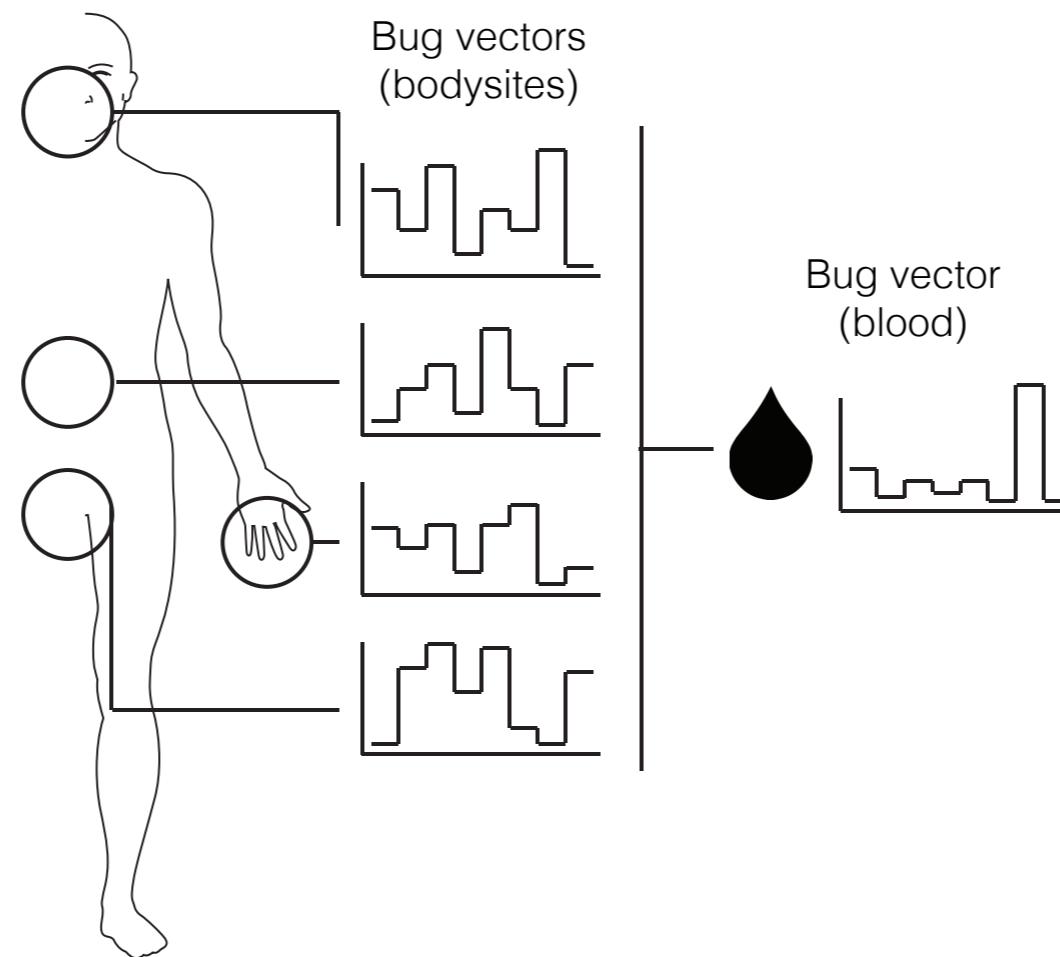
- Biopsy replacement

Undiagnosed

- Viruses (I6)
- Fungi (L78)

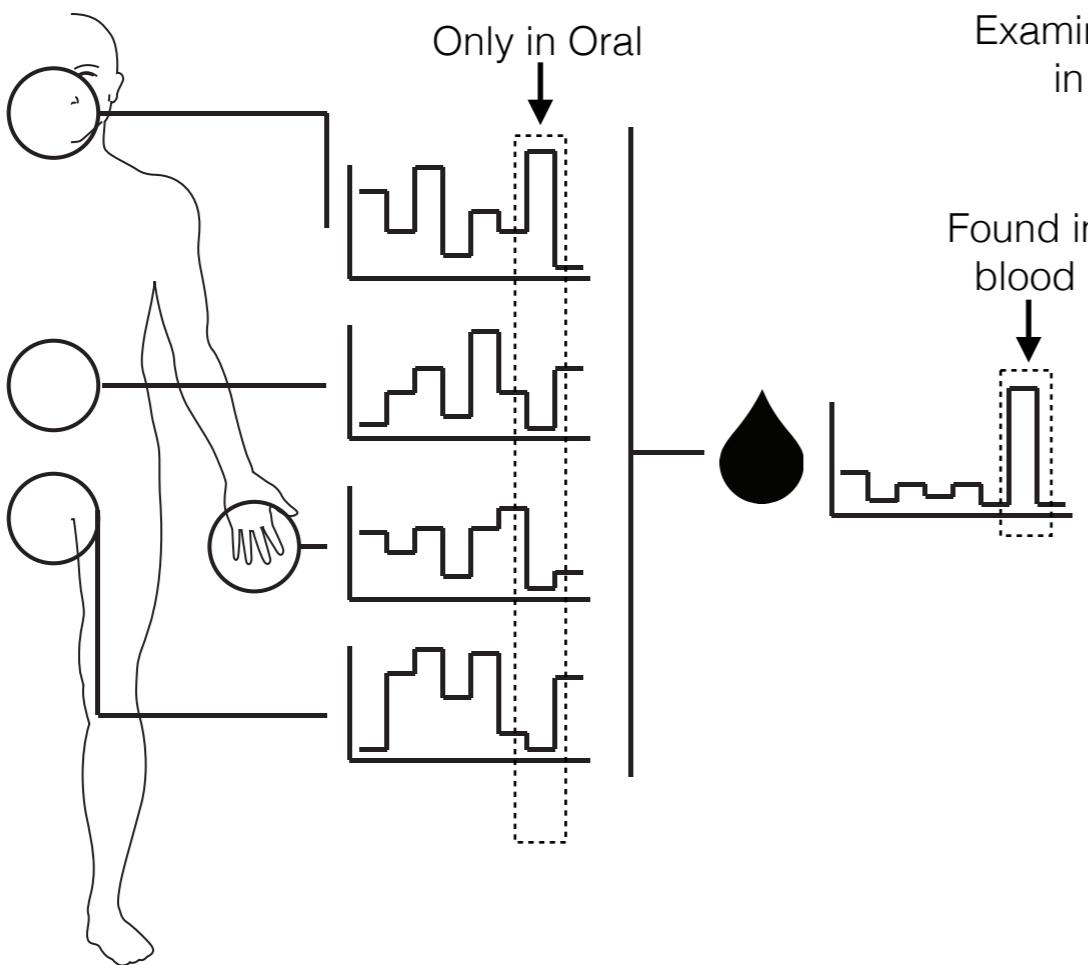
But, where does microbial cell-free DNA come from?

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

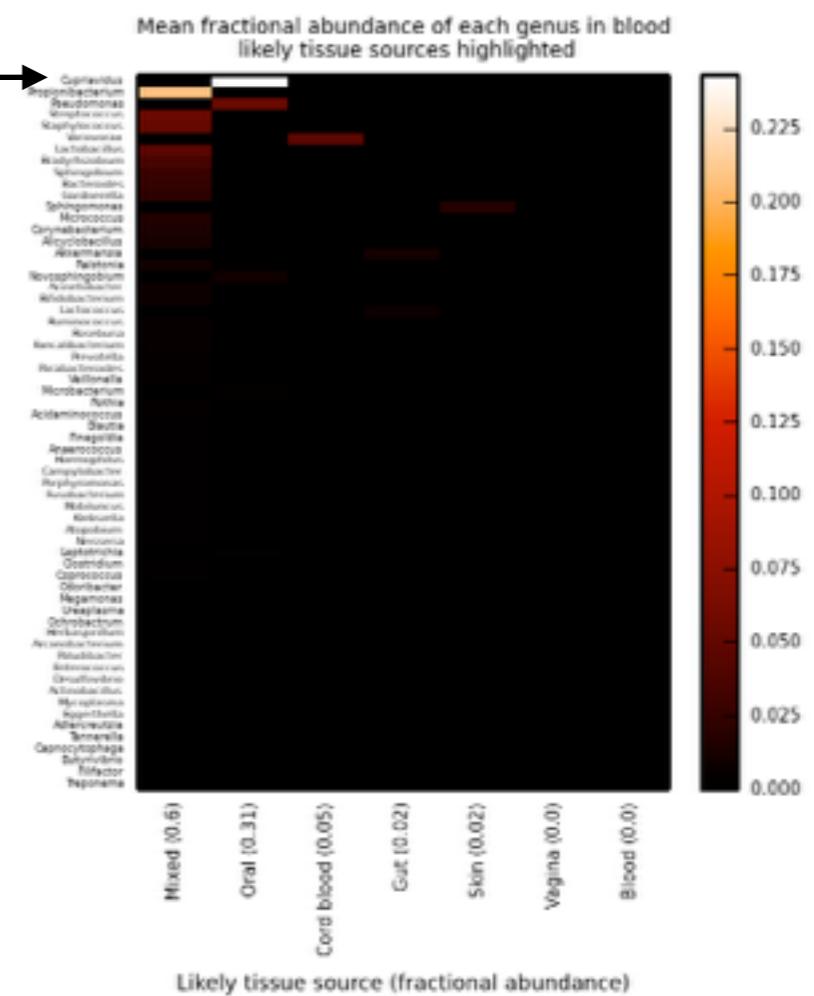


Tissue specific bugs don't explain blood composition.

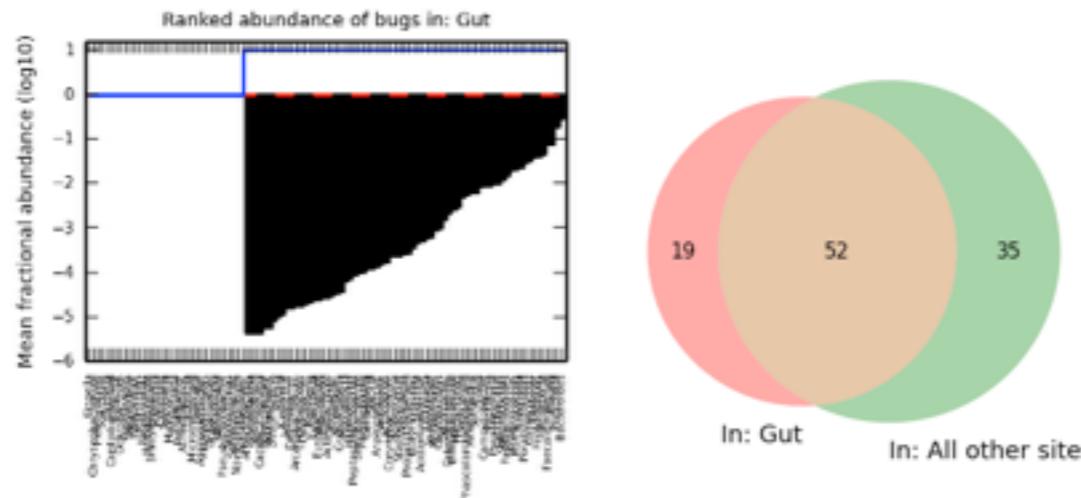
Define tissue specific bugs -



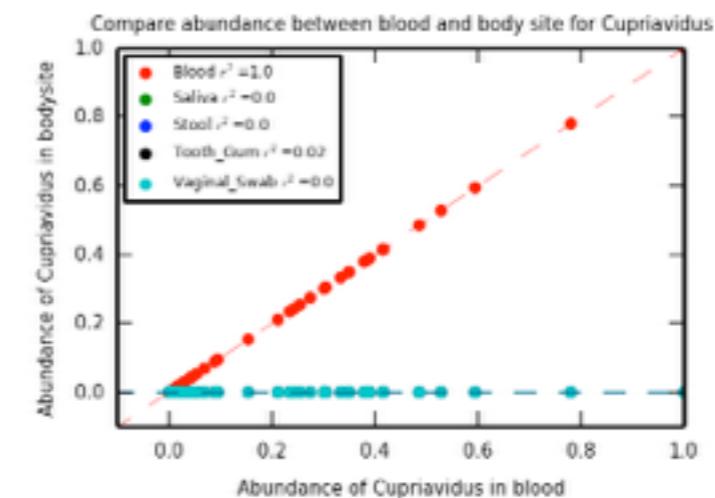
Assign likely source to each bug found in blood -



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

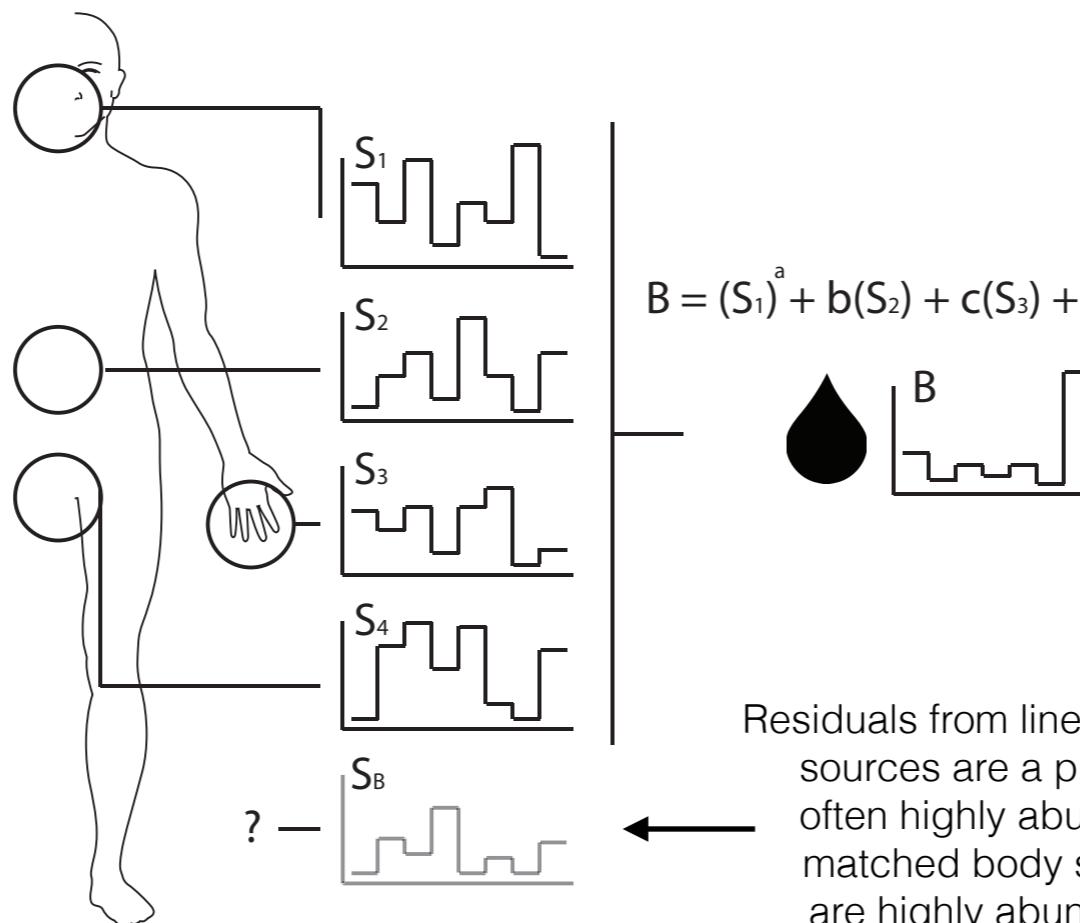


Cuprividius has trace abundance in other tissues, but high in blood -



... 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 -



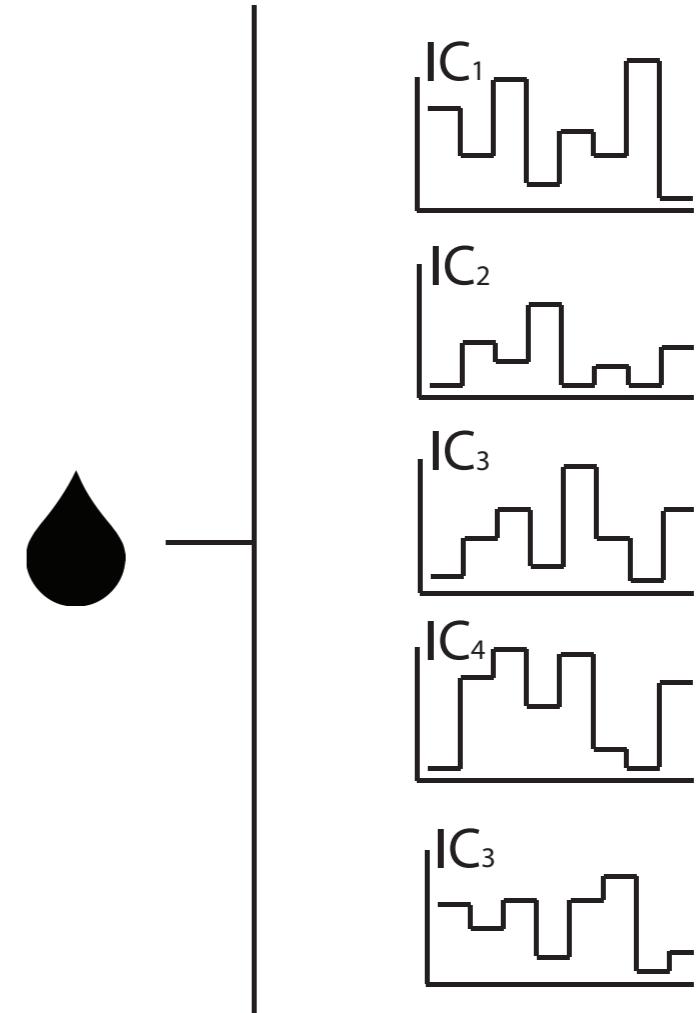
Residuals from linear regression [see appendix] show that blind sources are a problem: (1) Bugs such as *Curprividius* are often highly abundant in blood sample, but not present at matched body sites. (2) Bugs such as *Propionibacterium* are highly abundance in skin, but skin is rarely sampled.

Try approach that is not constrained to sampled sites.

The cocktail party problem -



Blood microbiome as a cocktail party problem -



Toy problem for source de-convolution from mixed signals -

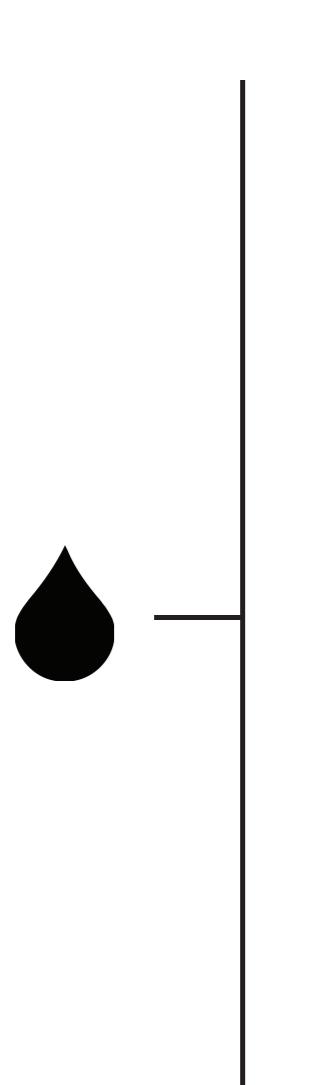
- Non-Gaussian, independent sources (voices)
- From recorded mixed conversations, it is possible to isolate the sources.
- ICA algorithm used to do this (see derivation in appendix).

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

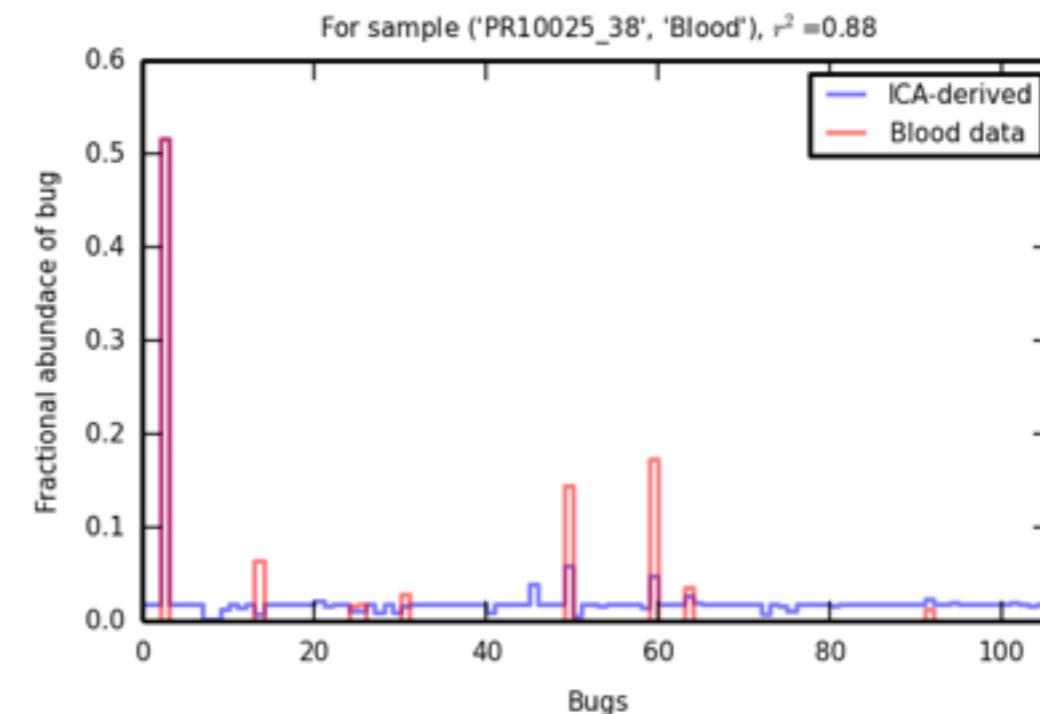
Unsupervised source detection in the pregnancy cohort.

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

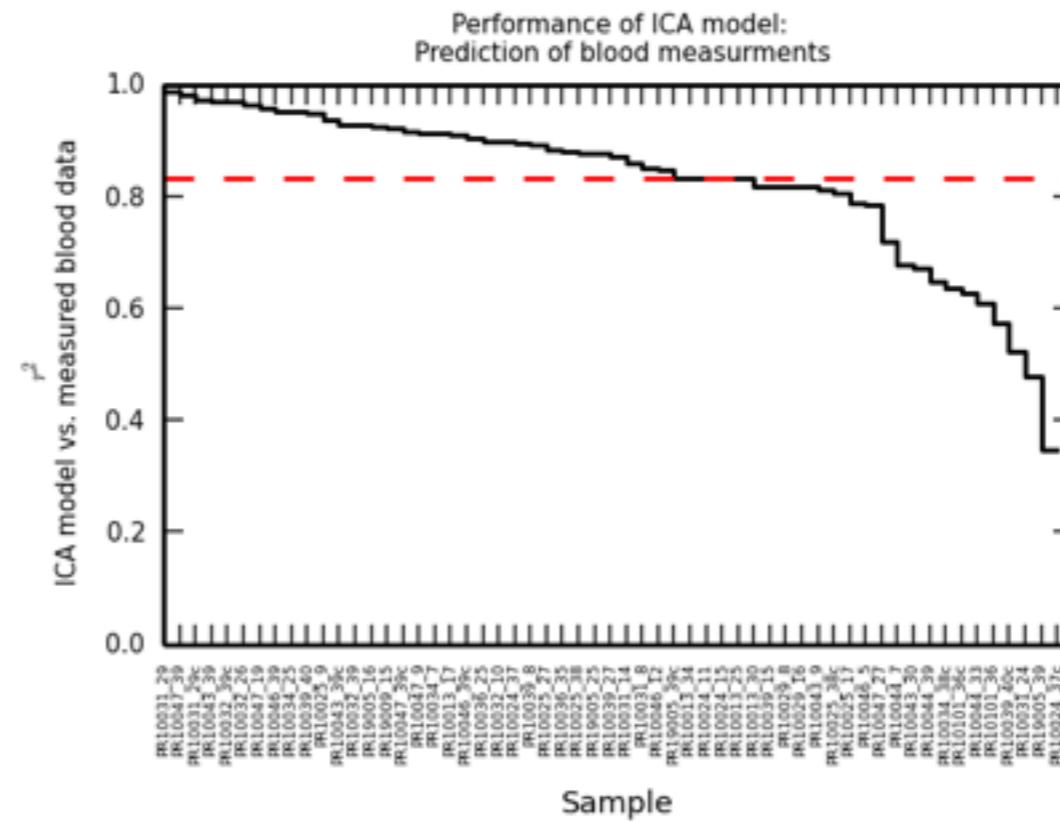
$$B_k = A_{k1}(IC_1) + A_{k2}(IC_2) + A_{k3}(IC_3) + A_{k4}(IC_4) + A_{k5}(IC_5)$$



(2) Mixing
matrix, A



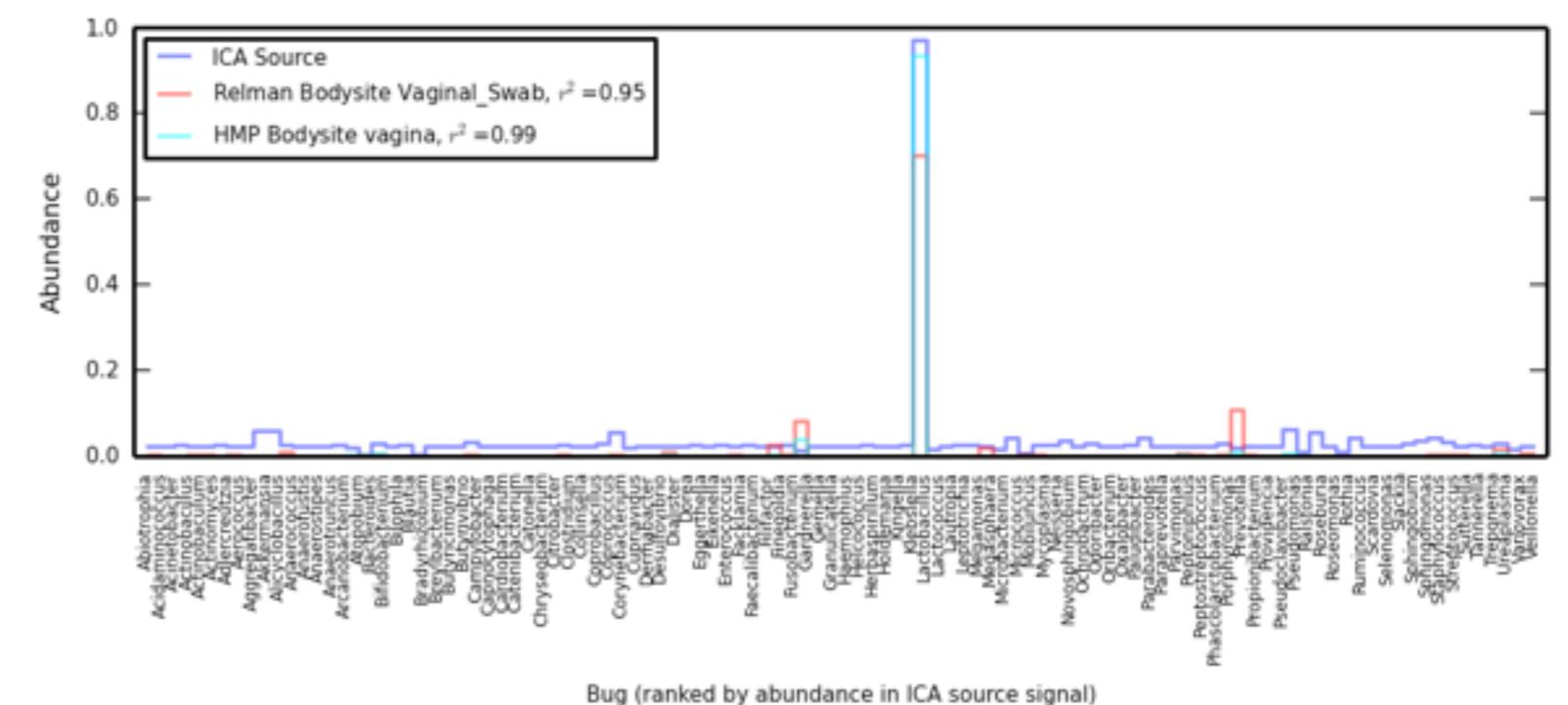
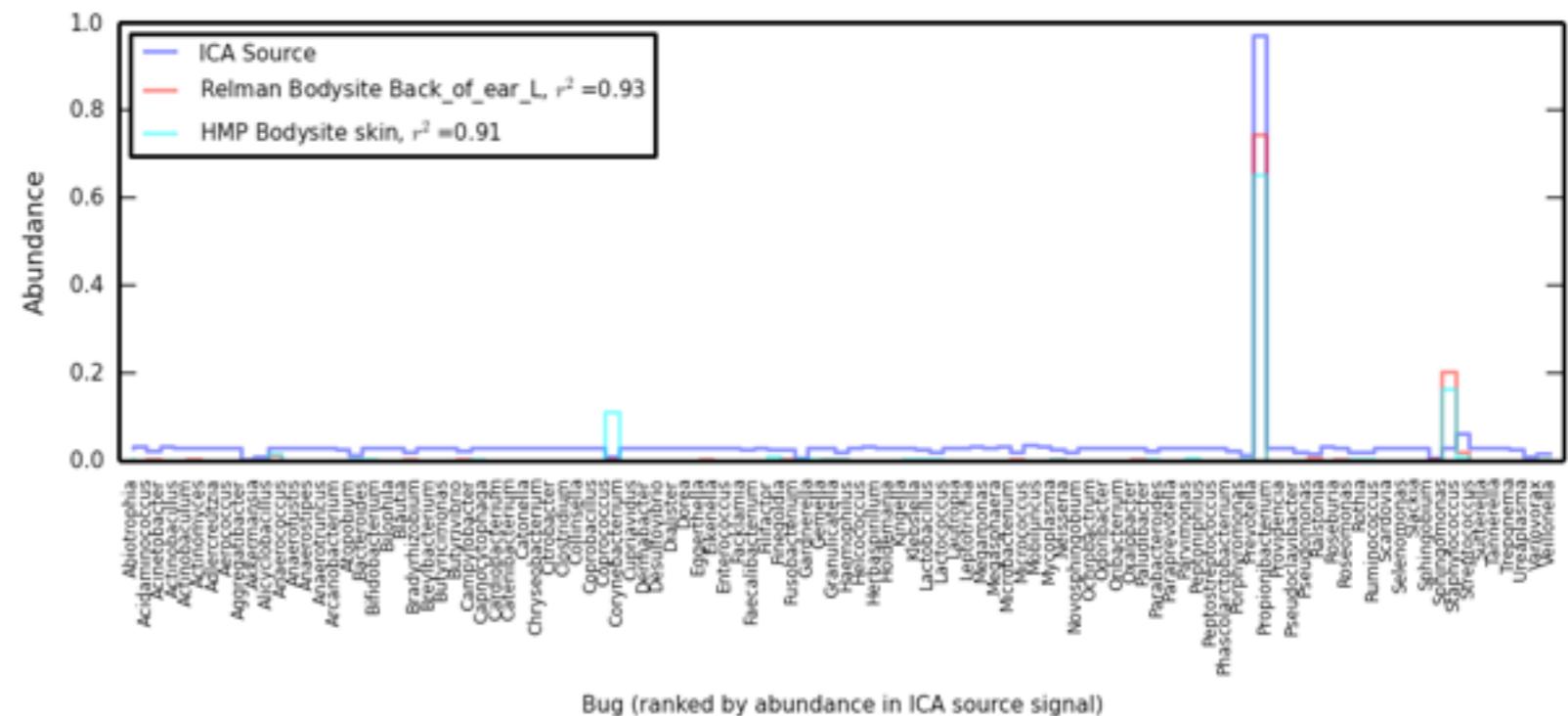
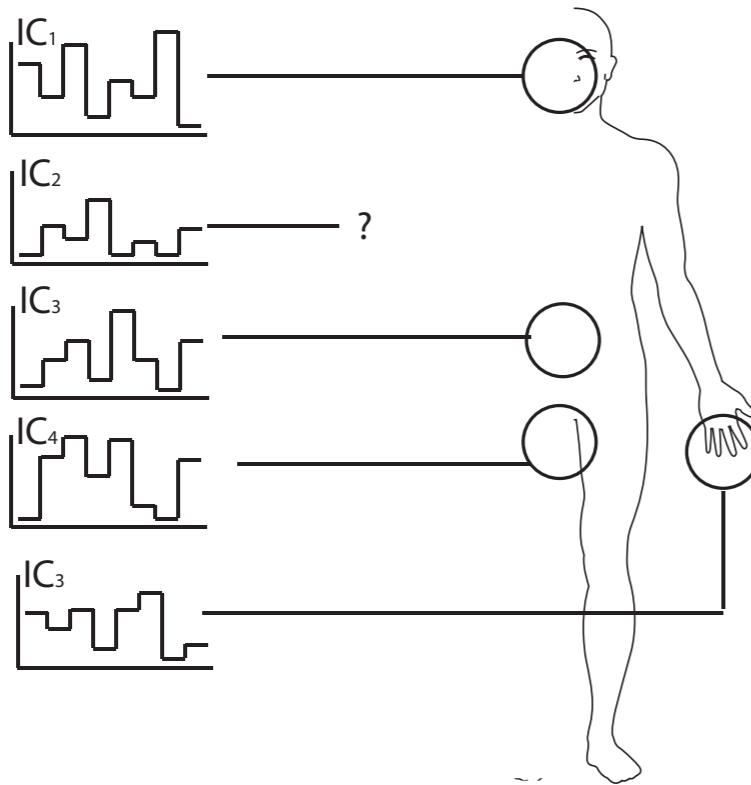
All samples (with mean highlighted) -



Assign ICA sources to sampled body sites.

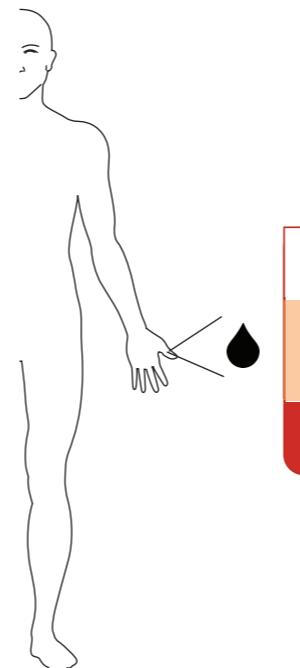
We can then try to assign “discovered” ICA sources to tissues using HMP and Relman data -

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

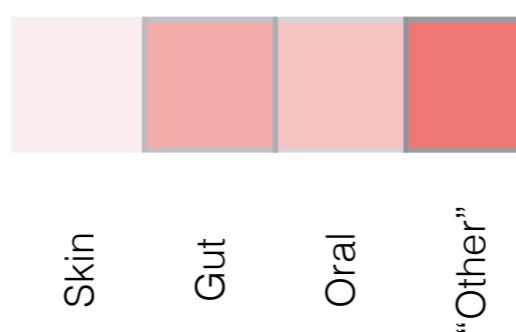


Each blood sample is a combination of ICA sources.

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



Relative weight
of ICA sources:

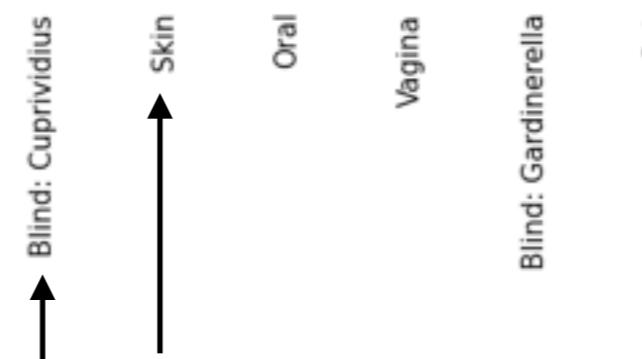
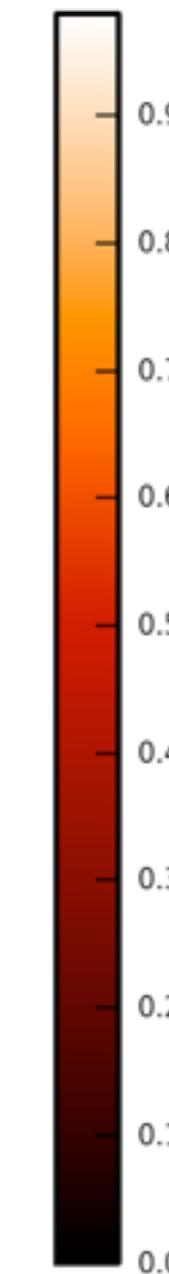
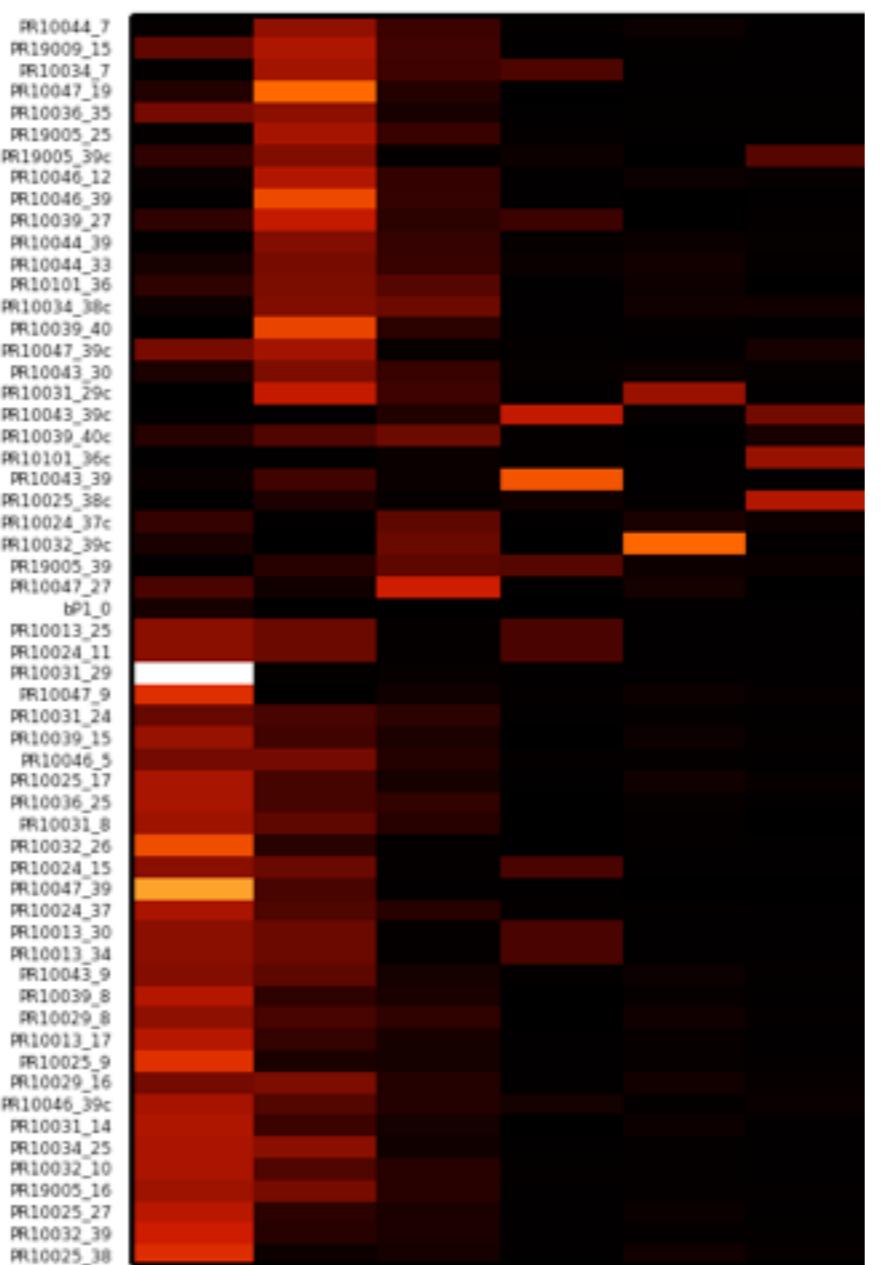


Clustered mixing matrix for pregnancy cohort.

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

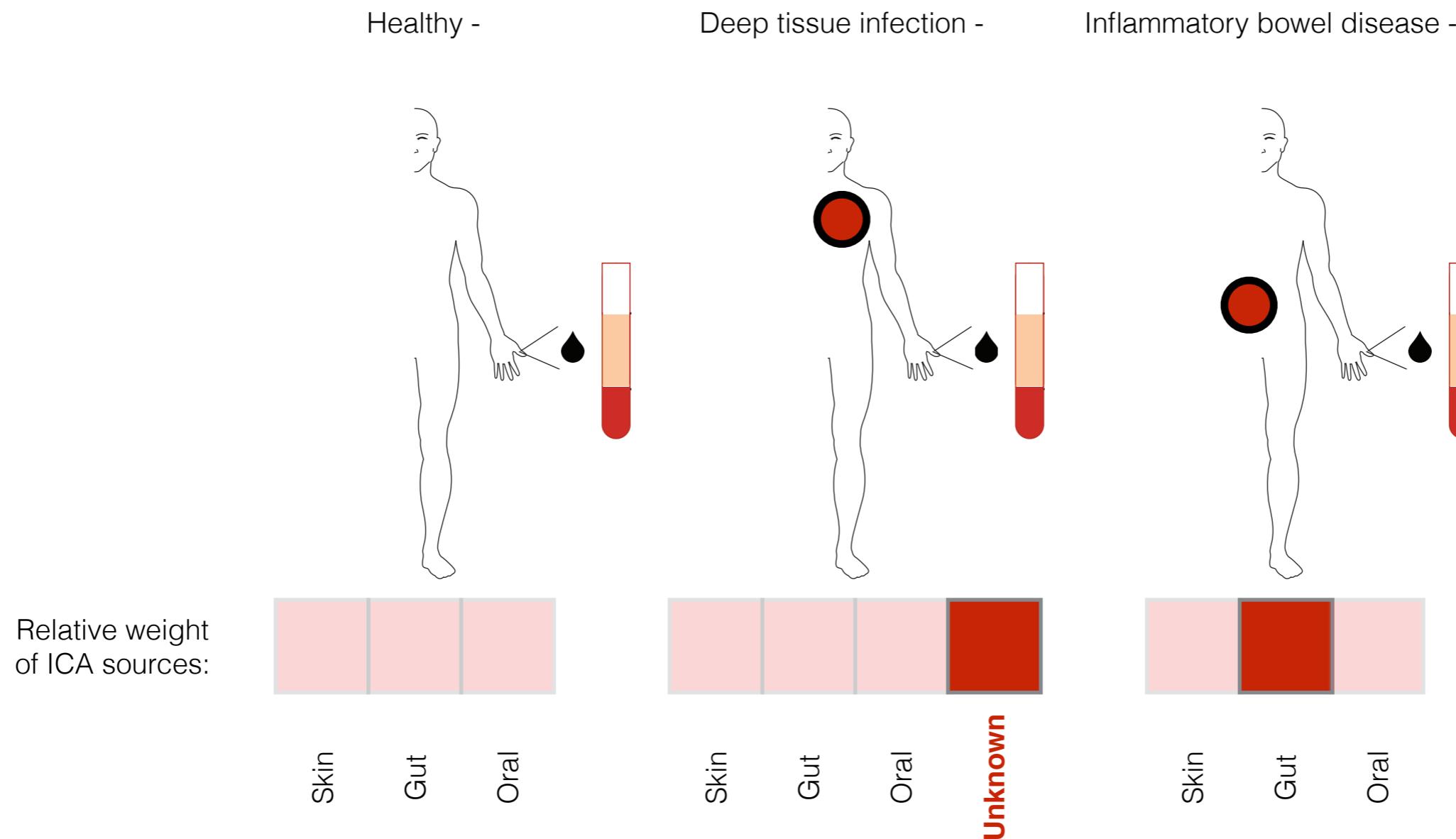


Cluster (2): Enriched in source with high load of *Cupriavidius*, which has no body site association (literature says it may grow in blood, so possible blood commensal).

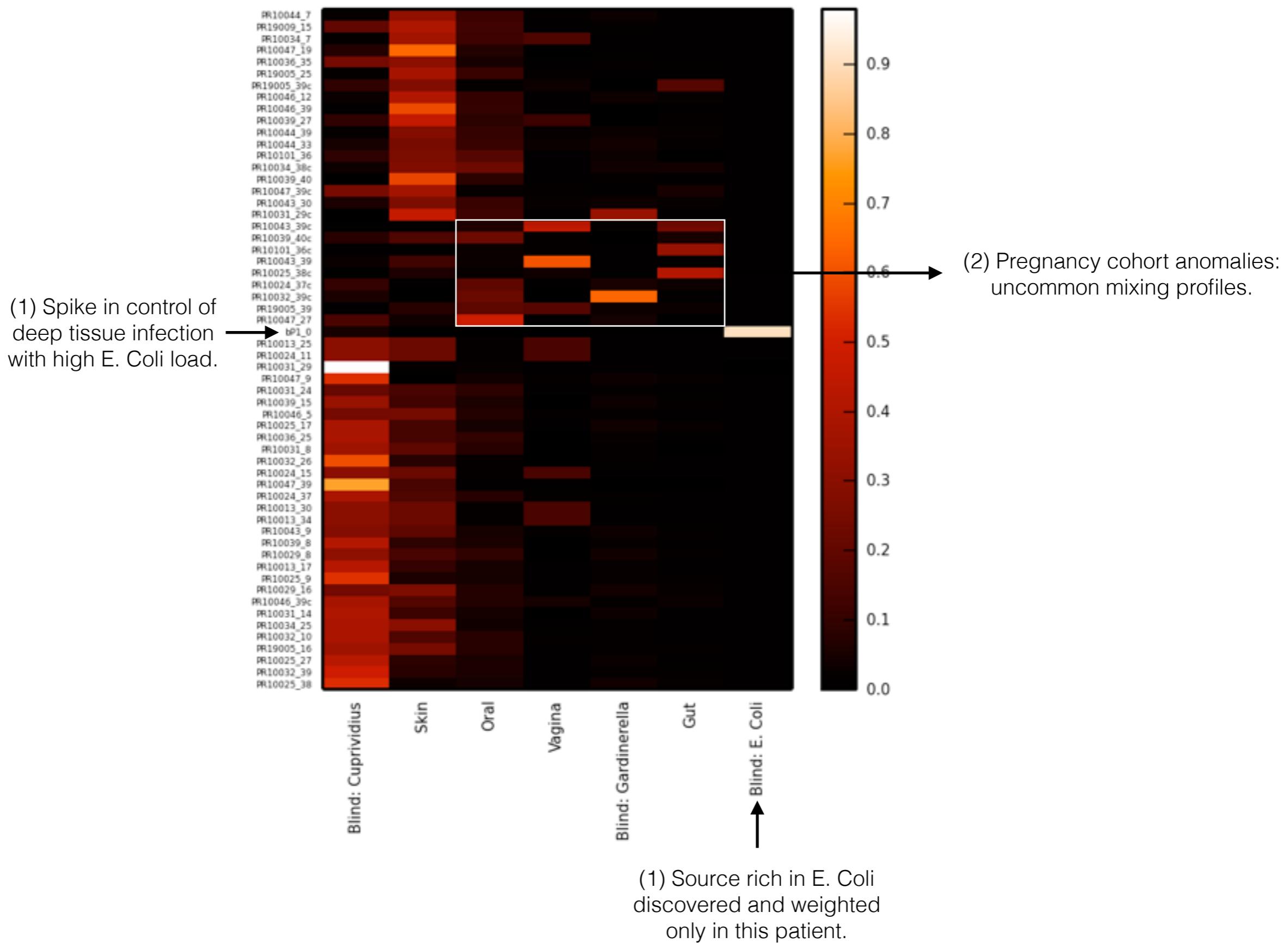


Dominant sources found in pregnancy cohort.

ICA for anomaly detection in cfDNA microbiome data.



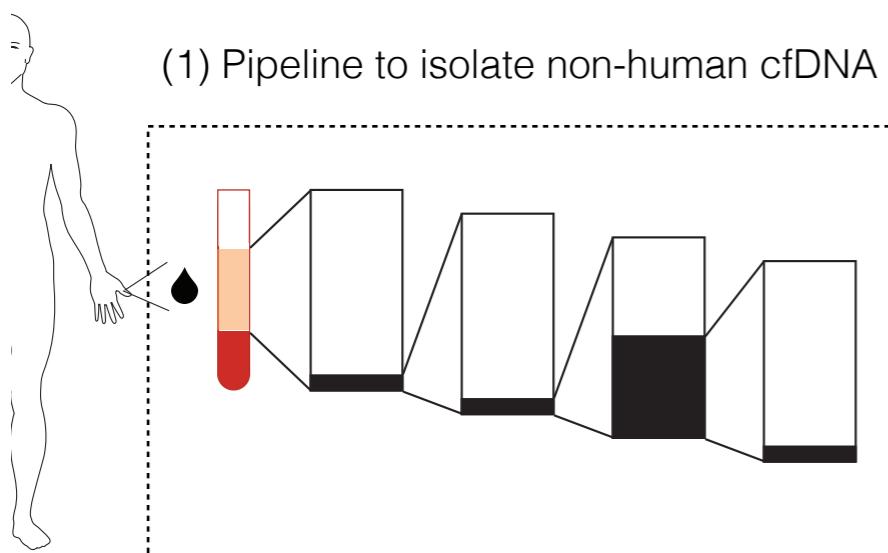
ICA mixing matrix can pinpoint anomalies.



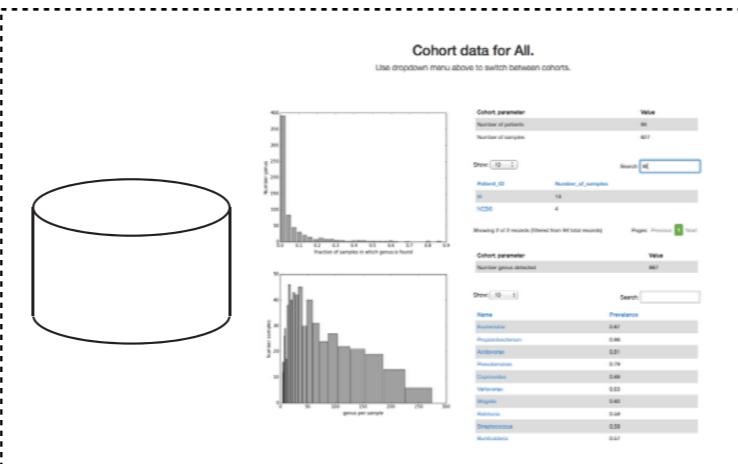
Summary

Applications

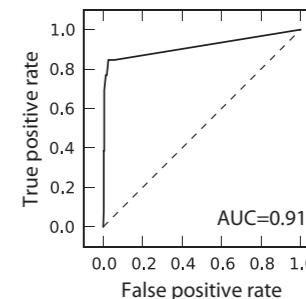
Pipeline



(2) Browser for visualization



(3) Clinical studies: Lung, BMT



(4) Biopsy replacement - Stanford pathology



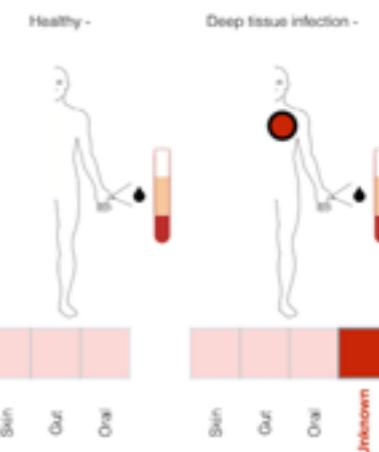
(5) Undiagnosed infections - I6, WU Case study

Sorted infection data for I6.

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

Name	I6_BC	I6_D-1	I6_W1	I6_W2	I6_W3	I6_W4	I6_W5	I6_W8	I6_W12	I6_W14	I6_W16	
WU_Polyomavirus	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	99.337748	95.685472	98.675497	98.013245
Human herpesvirus	0.00000	75.496689	88.079470	78.145495	0.00000	90.096225	80.132450	88.403974	91.390728	0.000000	82.781457	0.000000
5												
Enterocystophaga	0.00000	0.00000	90.728477	95.364238	94.039735	0.00000	95.025490	95.675497	99.337748	94.701987	92.715232	0.00000
henselii												

(6) Body site composition and anomalies - Pregnancy cohort

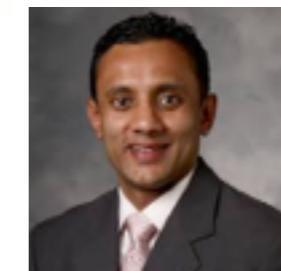


Thanks

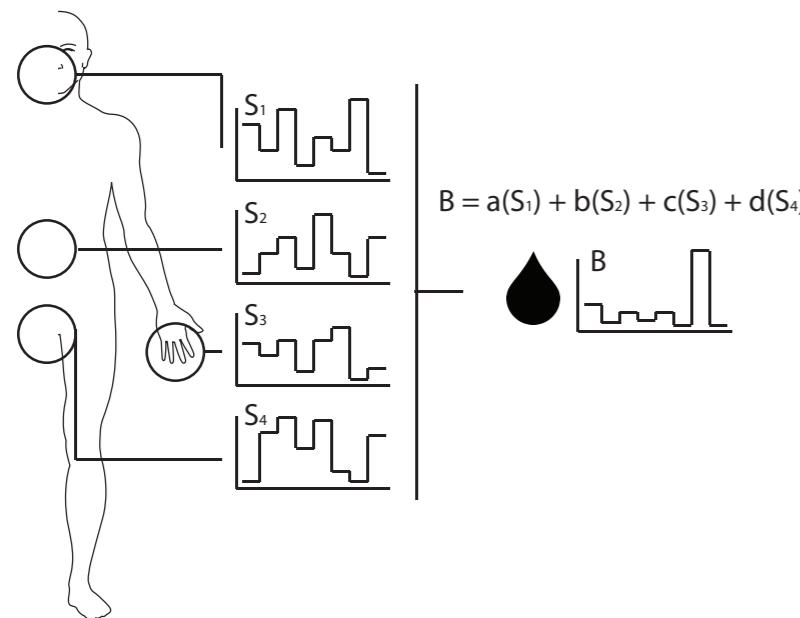
Funding



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 source and blood data to learn the parameters that weight each source.

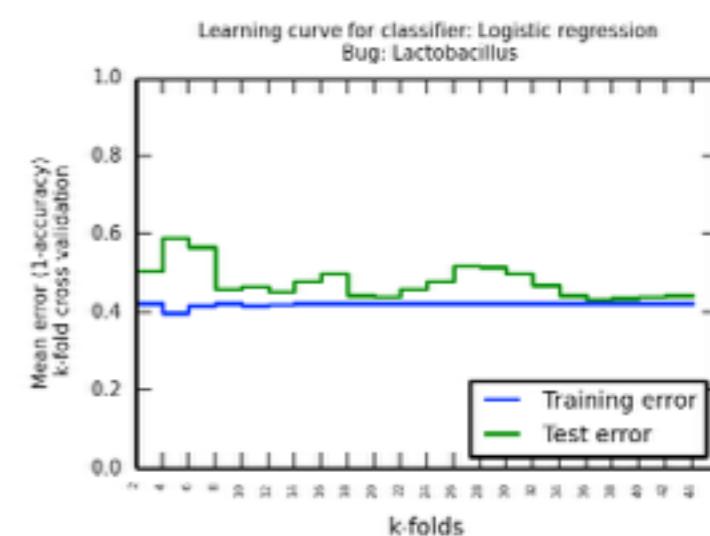
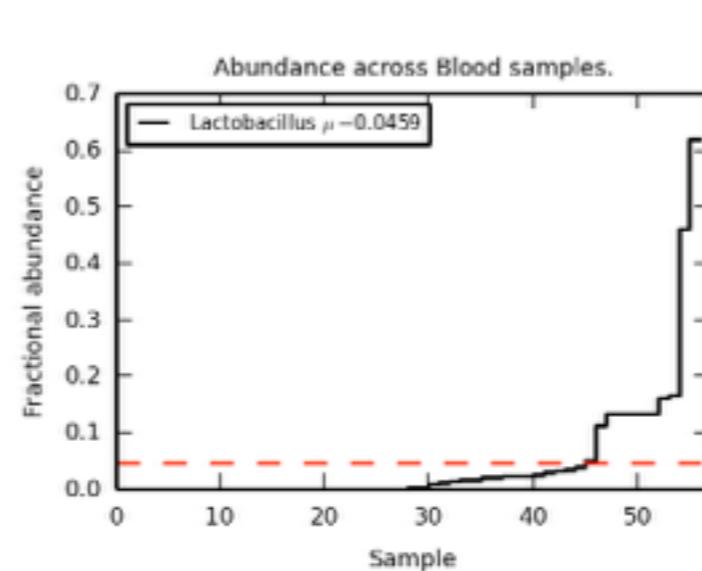
$$Y = \frac{1}{1 + e^{-\theta^T x}}$$

- 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.

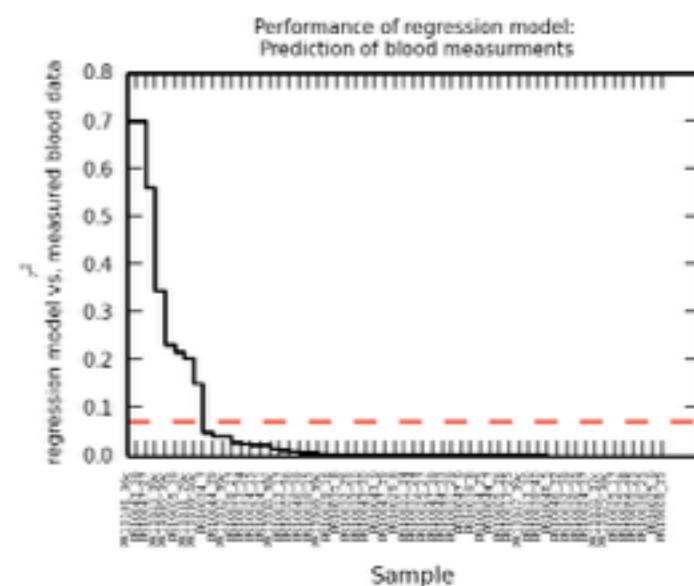
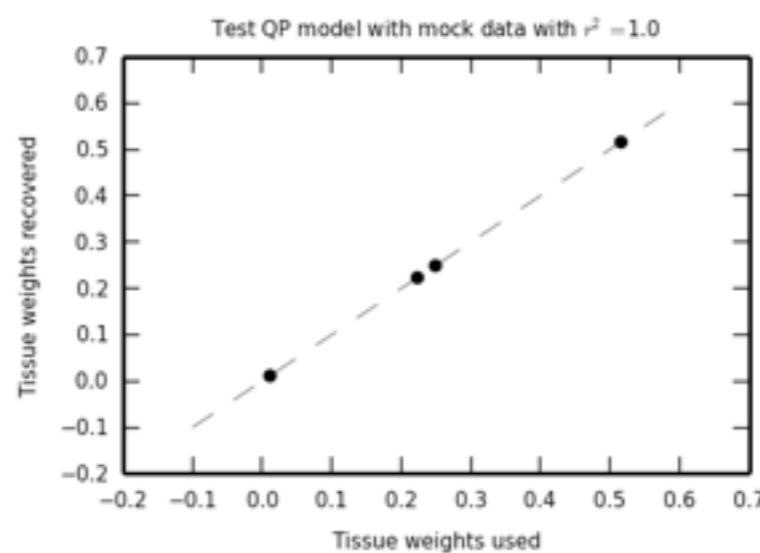
$$\vec{B} = \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix} = \sum_j x_j \theta_j = \begin{bmatrix} b_{j1} \\ \dots \\ b_{jn} \end{bmatrix} \theta_j + \dots + \epsilon$$

- 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 particular 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.

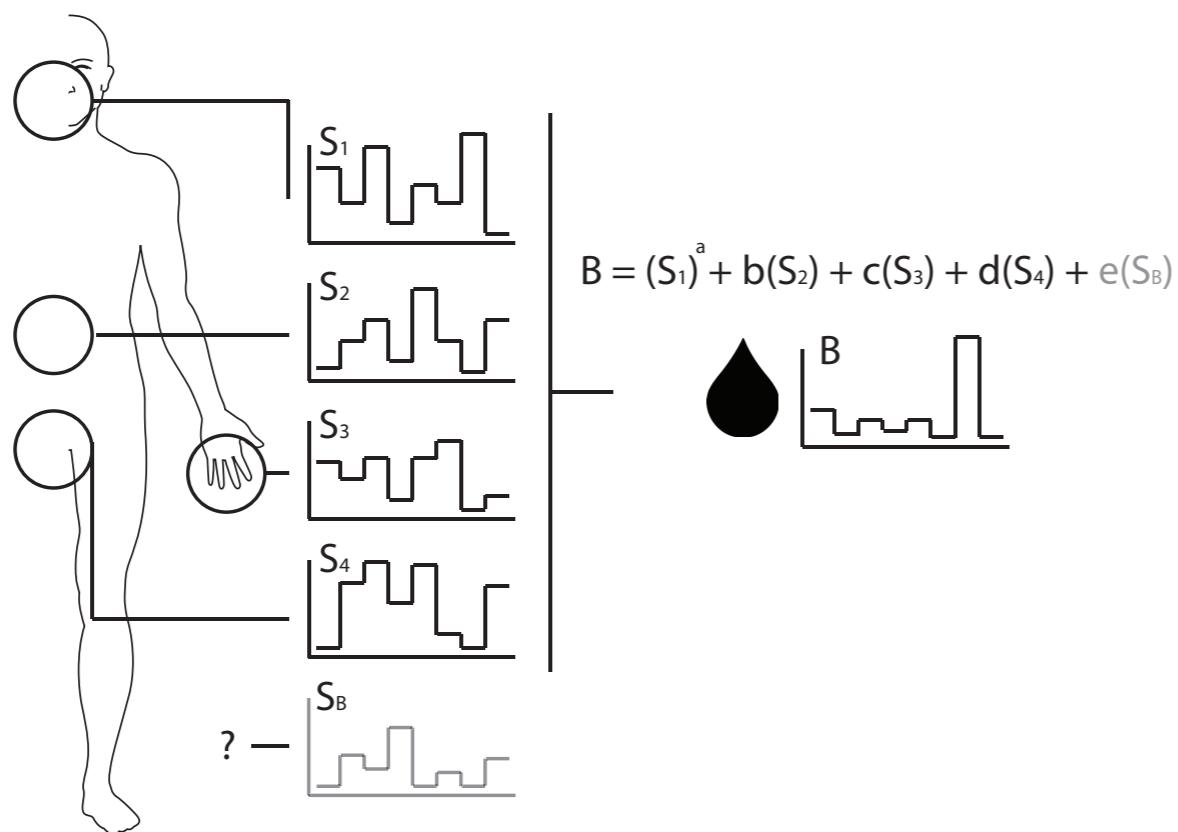


(C) **Classification with logistic regression performs poorly.** We pick a bug (Lactobacillus) with even labeling (detection) in blood (to avoid trivial model results) and compute a learning curve to evaluate whether the model.



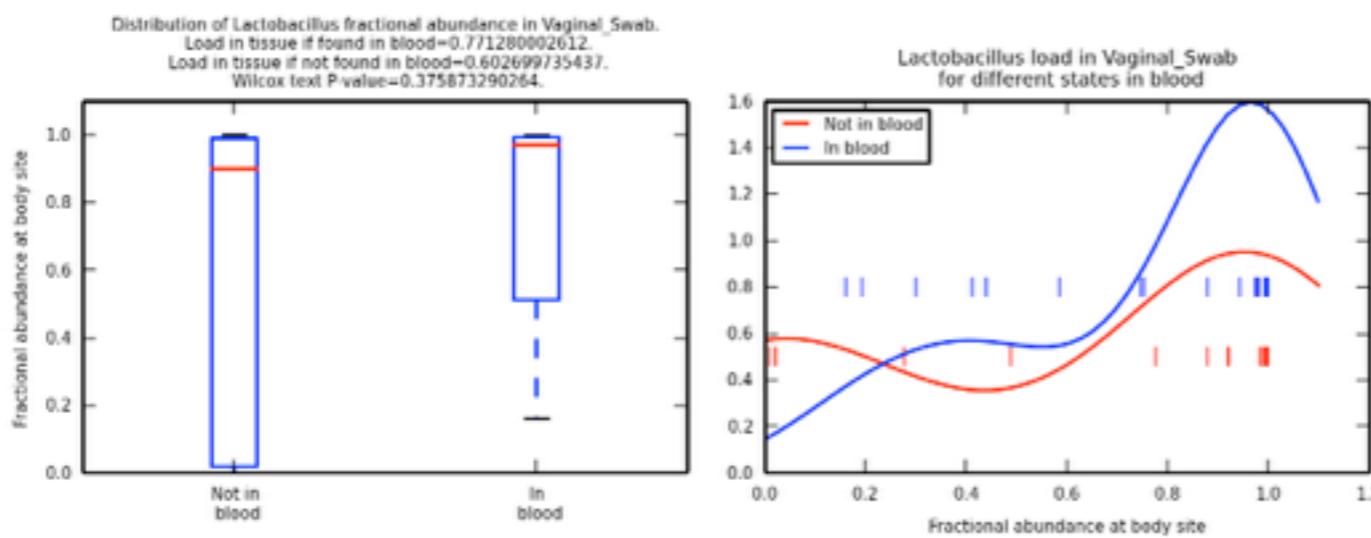
(D) **Linear regression performs poorly.** We apply linear regression with sensible constraints on the learned tissue weights (e.g., > 0). We use a quadratic and show that a test case with simulated data using tissue weights (x) shows that these weights are recovered (y) correctly. We then applied the 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.

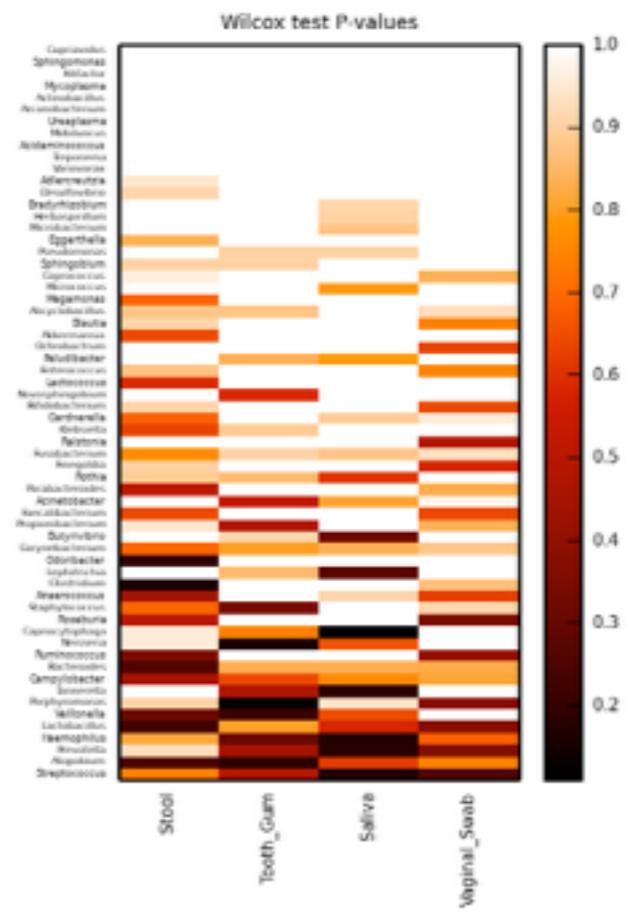


(A) Blind sources and non-linearity frustrate performance of linear models.

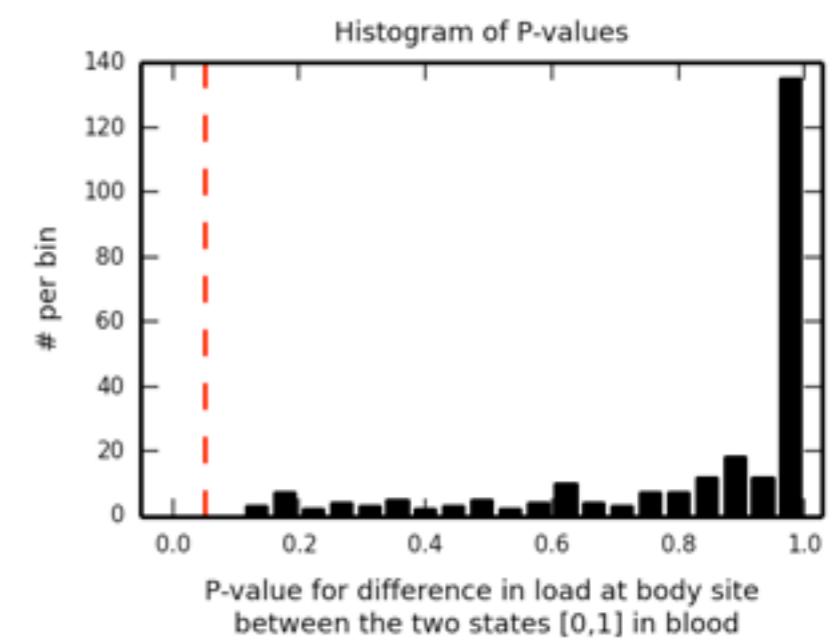
Learning curve from logistic regression indicated high bias (underfit model), suggesting non-linear relationships. Residuals from linear regression show that blood samples contain some bugs that are not found in any sampled sites.



(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 significant 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):

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

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

Mixing matrix, A (M by n):

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

- For each of M blood mixtures, it provides a set of mixing terms.
- Each mixing term is associated 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):

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

- 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 dimensionality (k).

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

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 function, then we can find the value of W that maximizes it using numerical methods (e.g., gradient ascent).

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

$$l(W) = \prod_{i=1}^m p(x_i | W)$$

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

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 function, $g(s) = \frac{1}{1+e^{-s}}$.
- 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:

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

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 function: $g'(z) = g(z)(1 - g(z))$
- $\frac{d}{dx} \log(f(x)) = \frac{f'(x)}{f(x)}$
- $\nabla_W |W| = |W|(W^{-1})^T$

The first term numerator:

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

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

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

Alot, 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 :

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