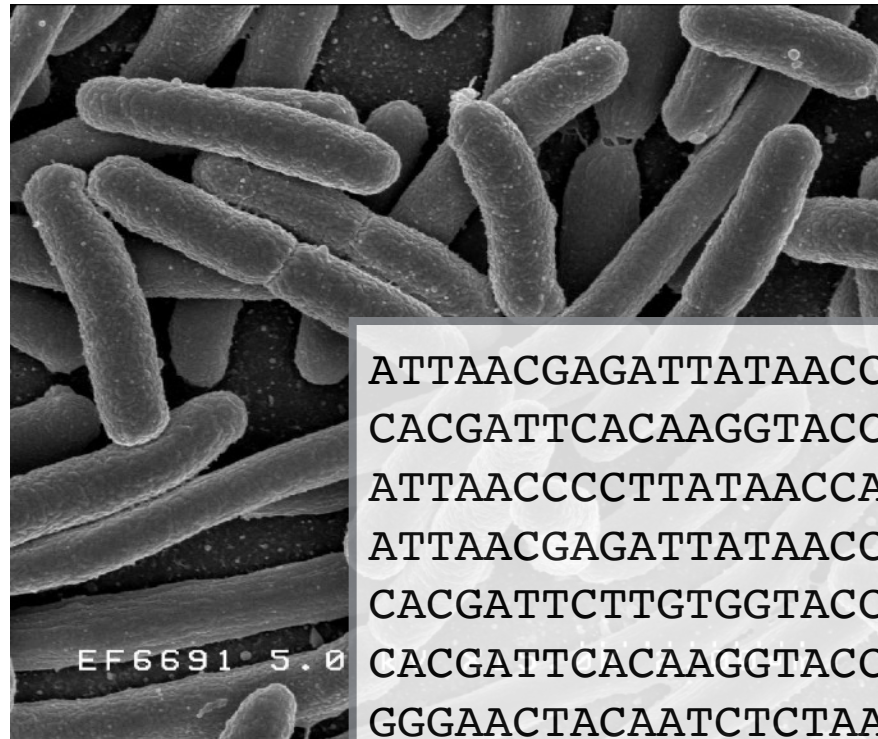


Contamination, controls and
accurate sequencing-based **measurement**
of microbial communities

A Microbial Census



ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAAC
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGA ACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC

<i>Lactobacillus crispatus</i>	1300	5	0	882	596
<i>Ureaplasma urealytica</i>	15	0	220	0	0
<i>Gardnerella vaginalis</i>	22	0	1	0	412
<i>Prevotella intermedia</i>	0	0	8	12	0
...

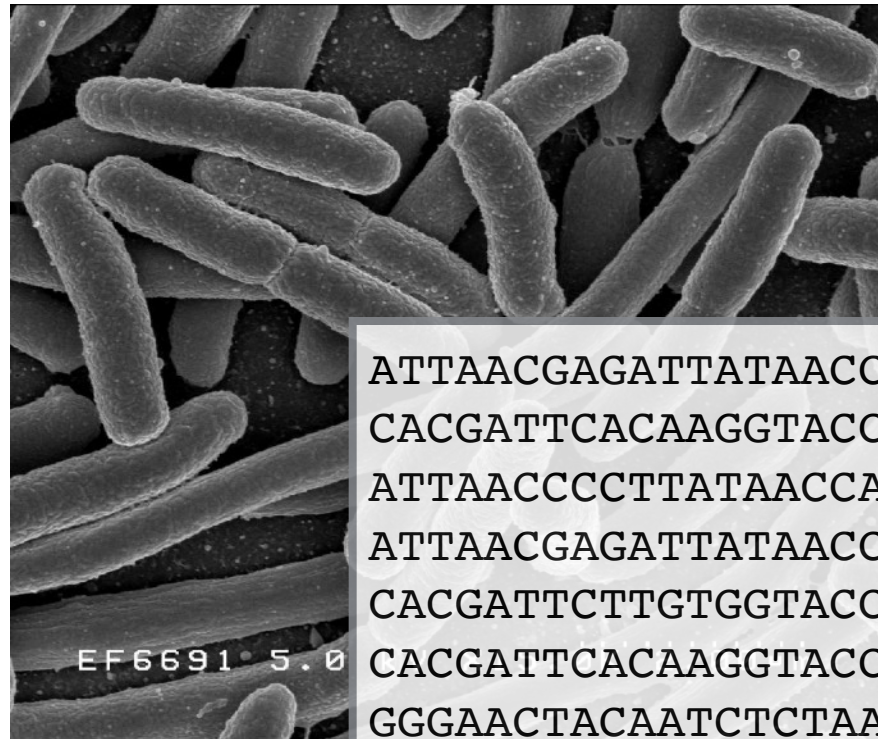
Visualization

Exploration

Inference

A Microbial Census

Marker-gene or Metagenomics Sequencing (MGS)



ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGA ACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC

<i>Lactobacillus crispatus</i>	1300	5	0	882	596
<i>Ureaplasma urealytica</i>	15	0	220	0	0
<i>Gardnerella vaginalis</i>	22	0	1	0	412
<i>Prevotella intermedia</i>	0	0	8	12	0
...

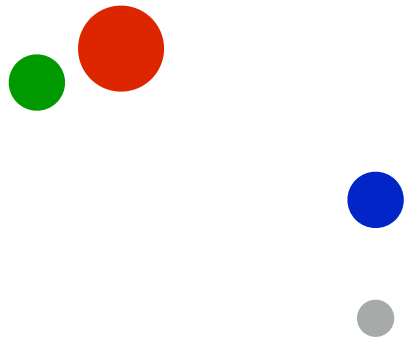
Visualization

Exploration

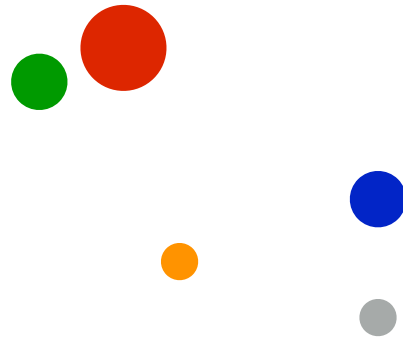
Inference

MGS: What is really there?

**Sample
Sequences**

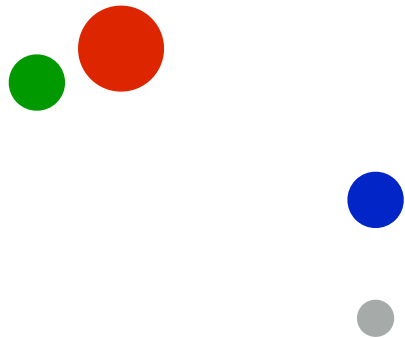


Extraction

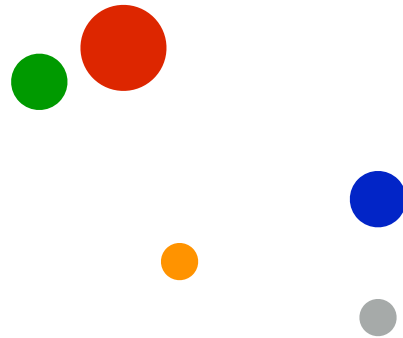


MGS: What is really there?

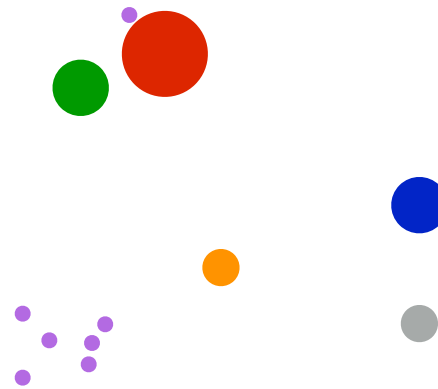
**Sample
Sequences**



Extraction

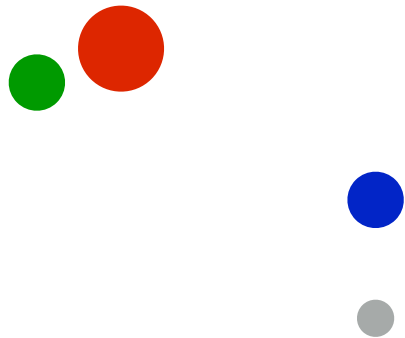


Library Prep

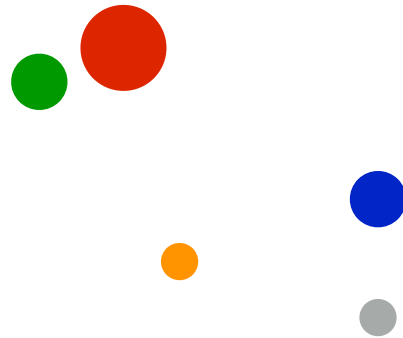


MGS: What is really there?

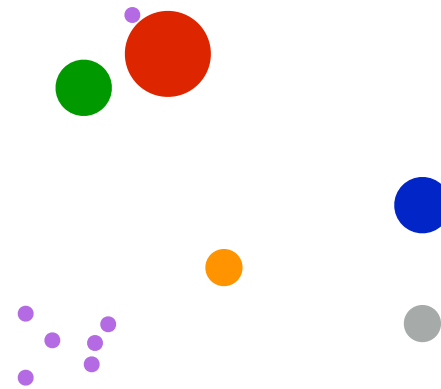
**Sample
Sequences**



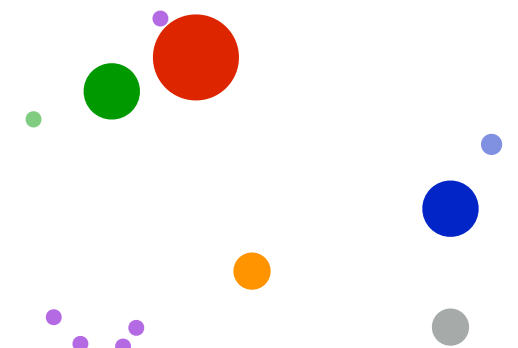
Extraction



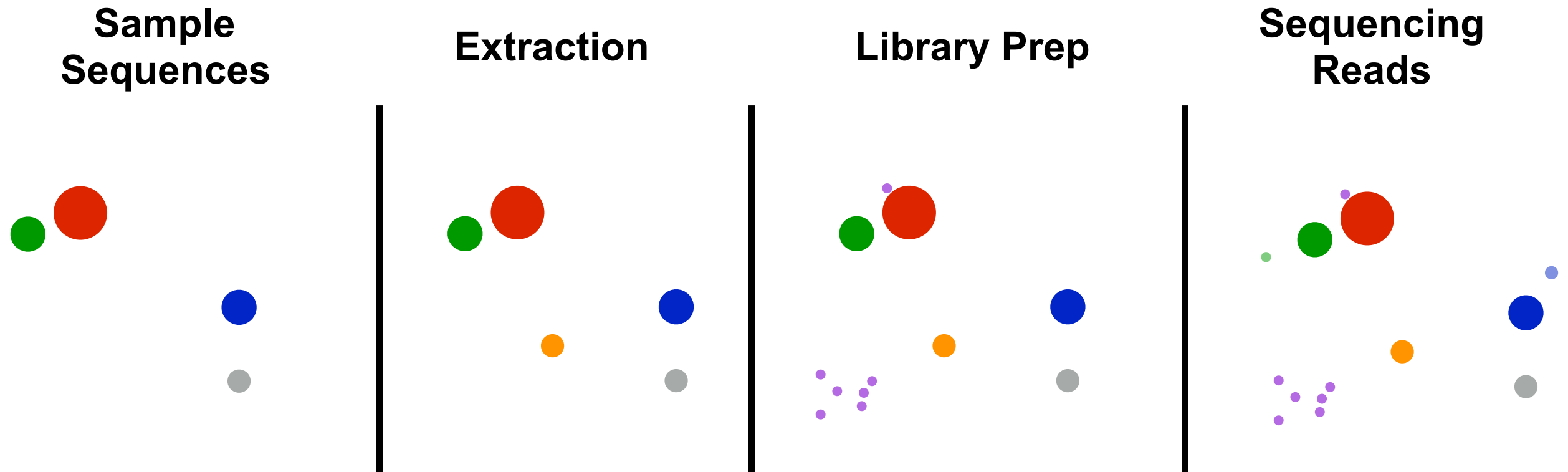
Library Prep



**Sequencing
Reads**

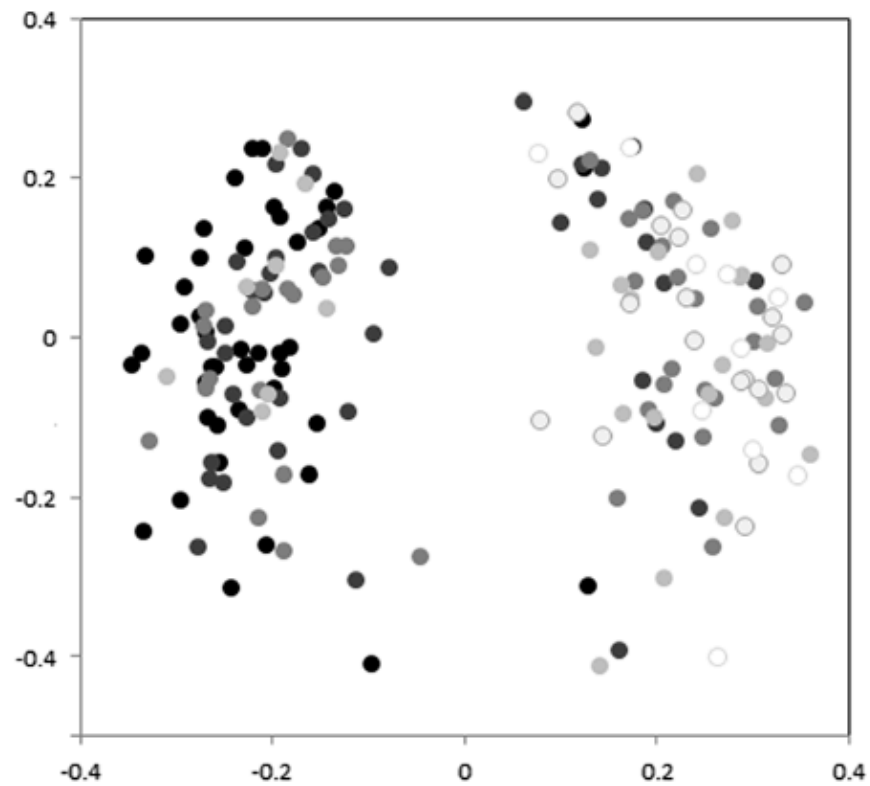


Problem: Contamination



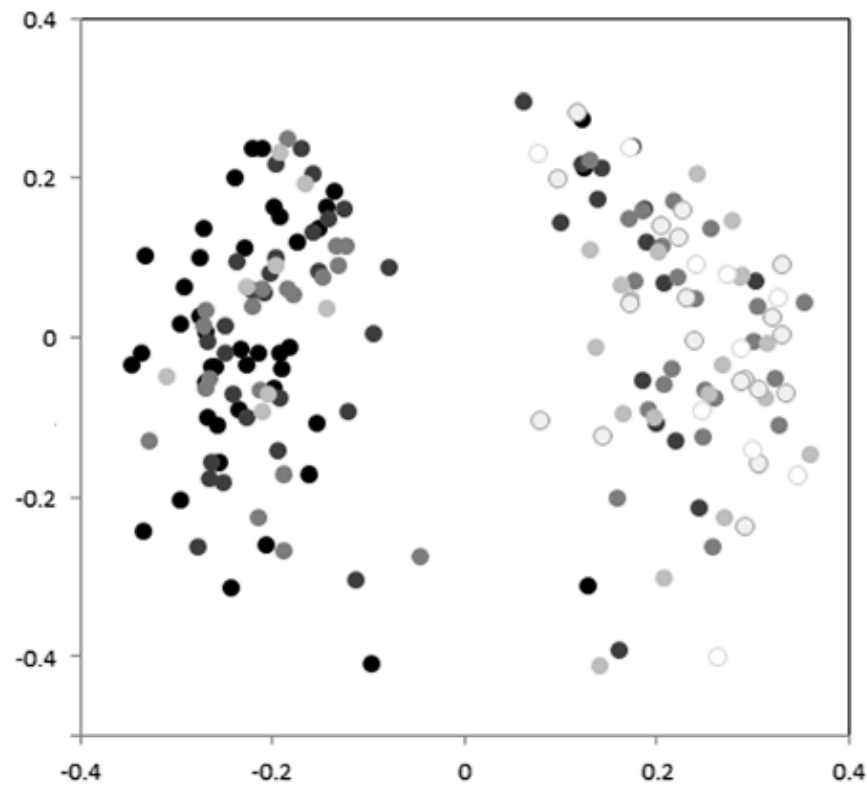
Contaminants — DNA sequences from organisms not truly present in the sample.

Problem: Contamination

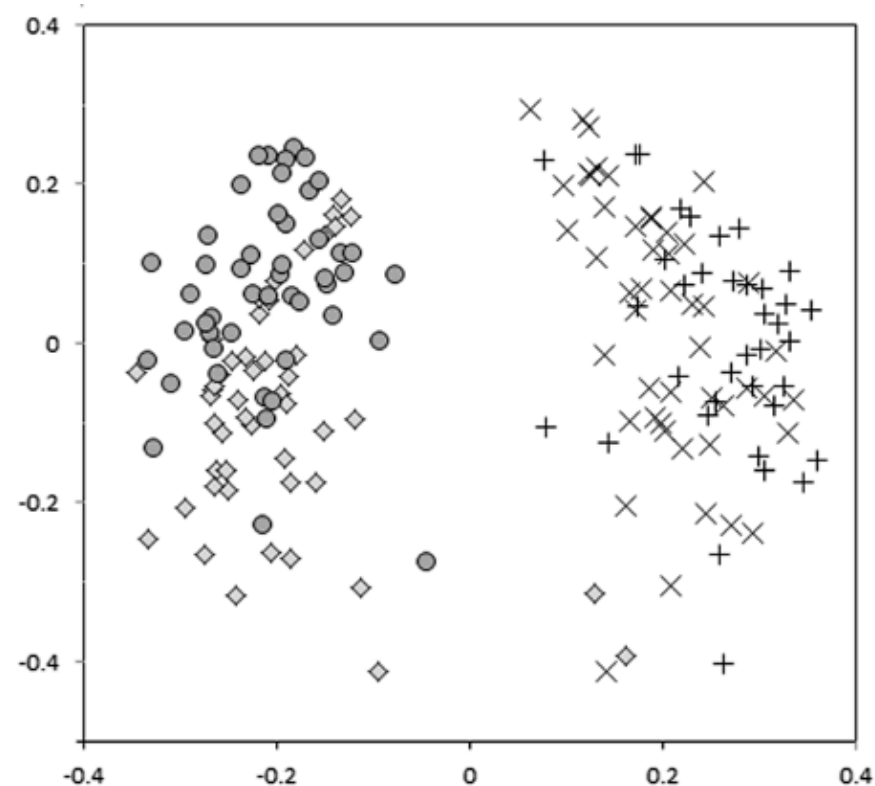


a) Full data coloured by age (months)

Problem: Contamination

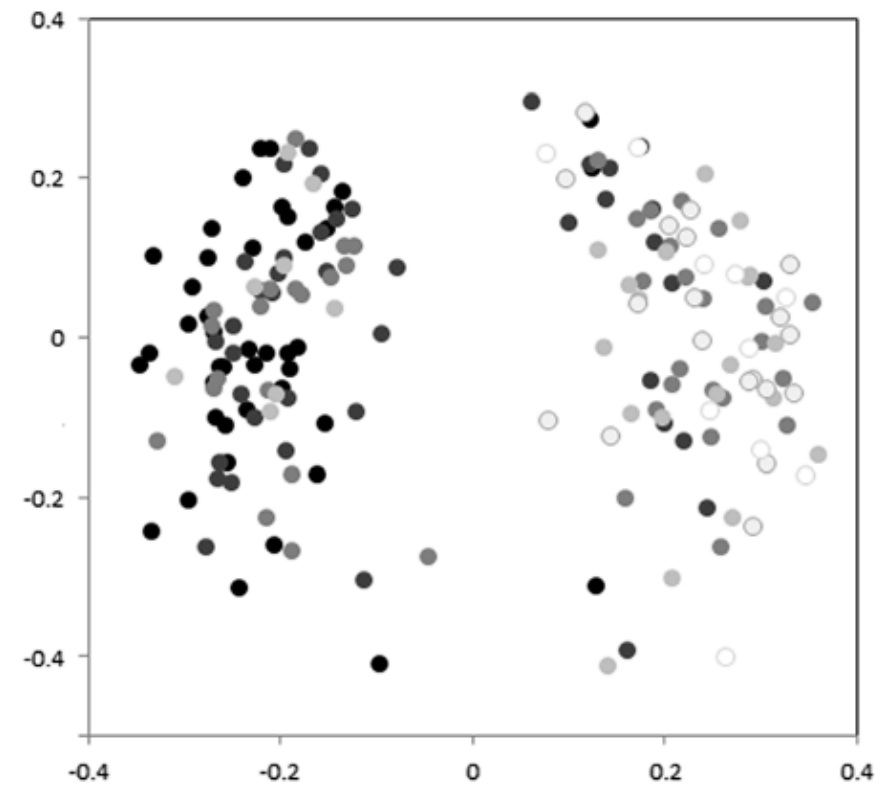


a) Full data coloured by age (months)

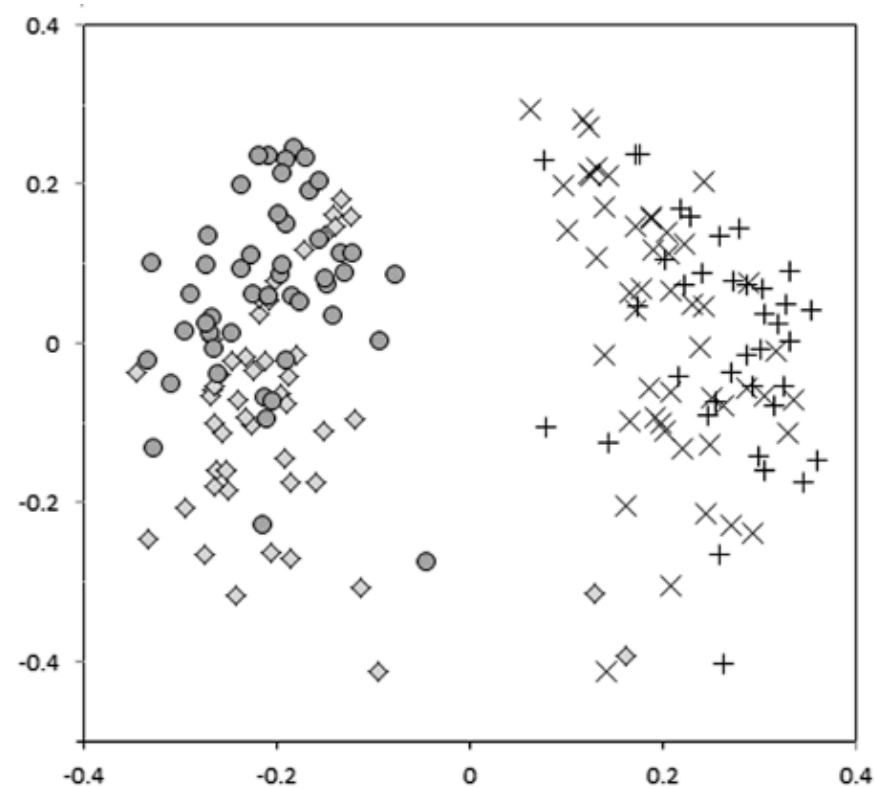


b) Full data coloured by extraction kit

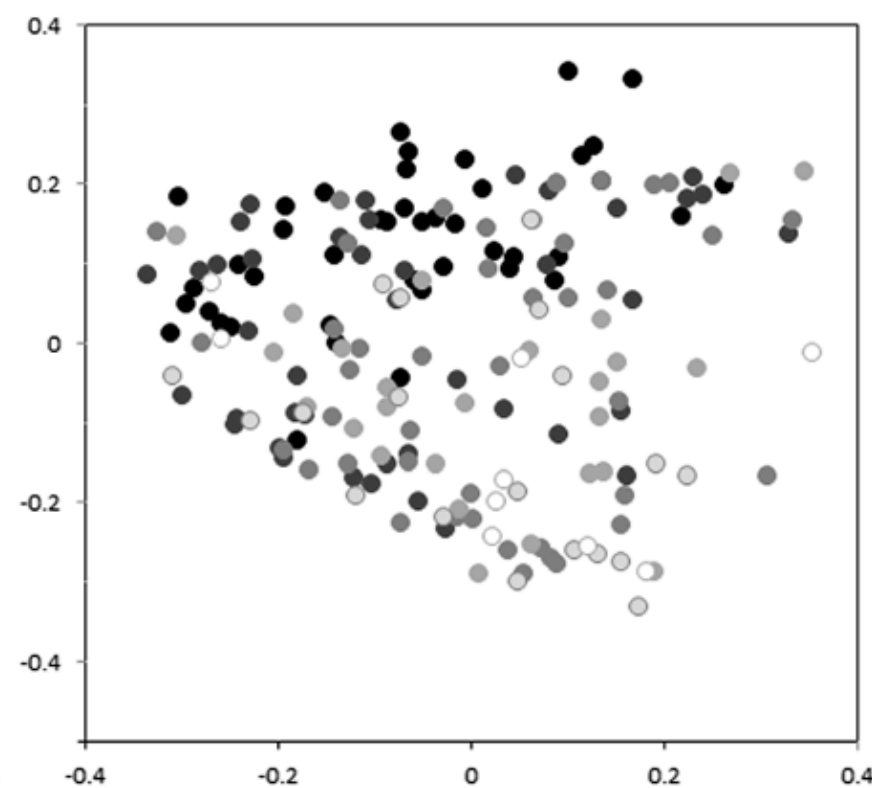
Problem: Contamination



a) Full data coloured by age (months)



b) Full data coloured by extraction kit



c) Contaminant OTUs removed

Spurious signal driven by contaminants!

Problem: Contamination

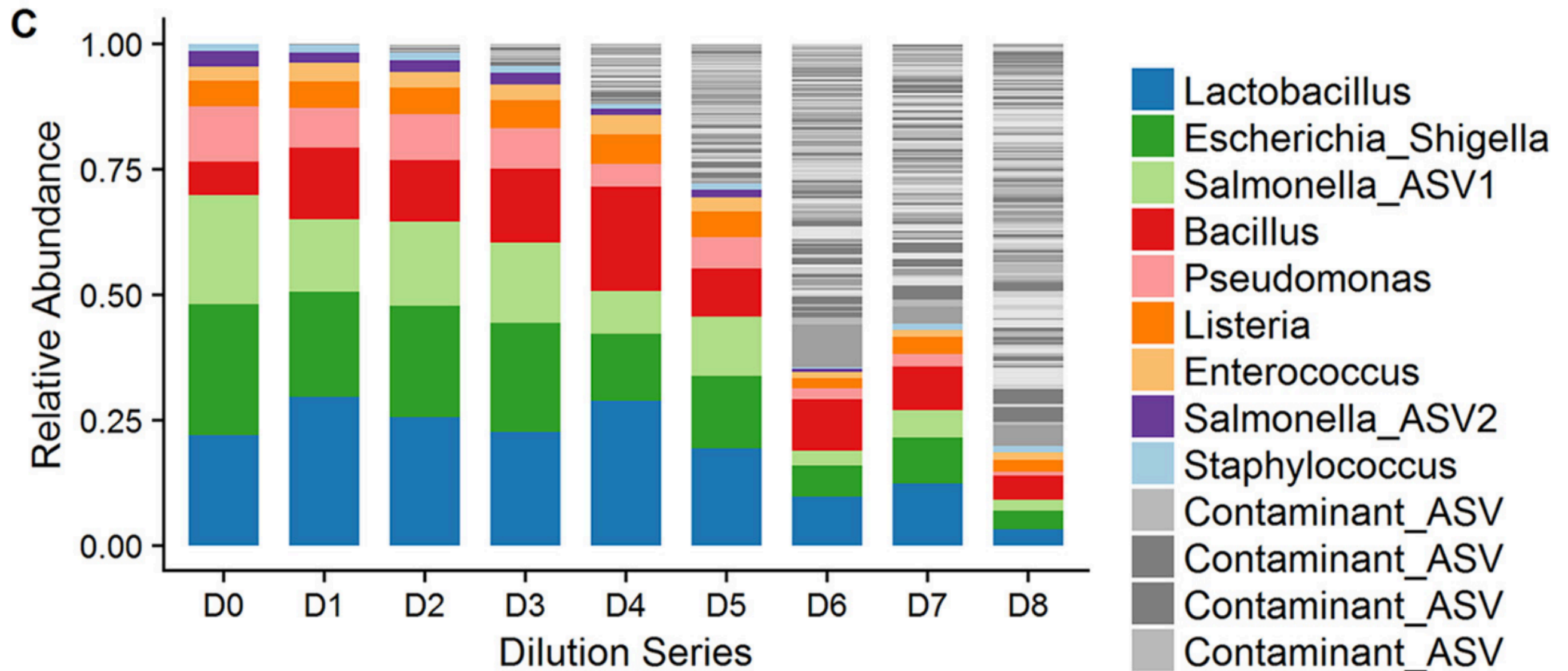


Figure: Karstens, et al. mSystems, 2018.

Modeling Contaminants

$T = S + C$, where **C** is constant

hence

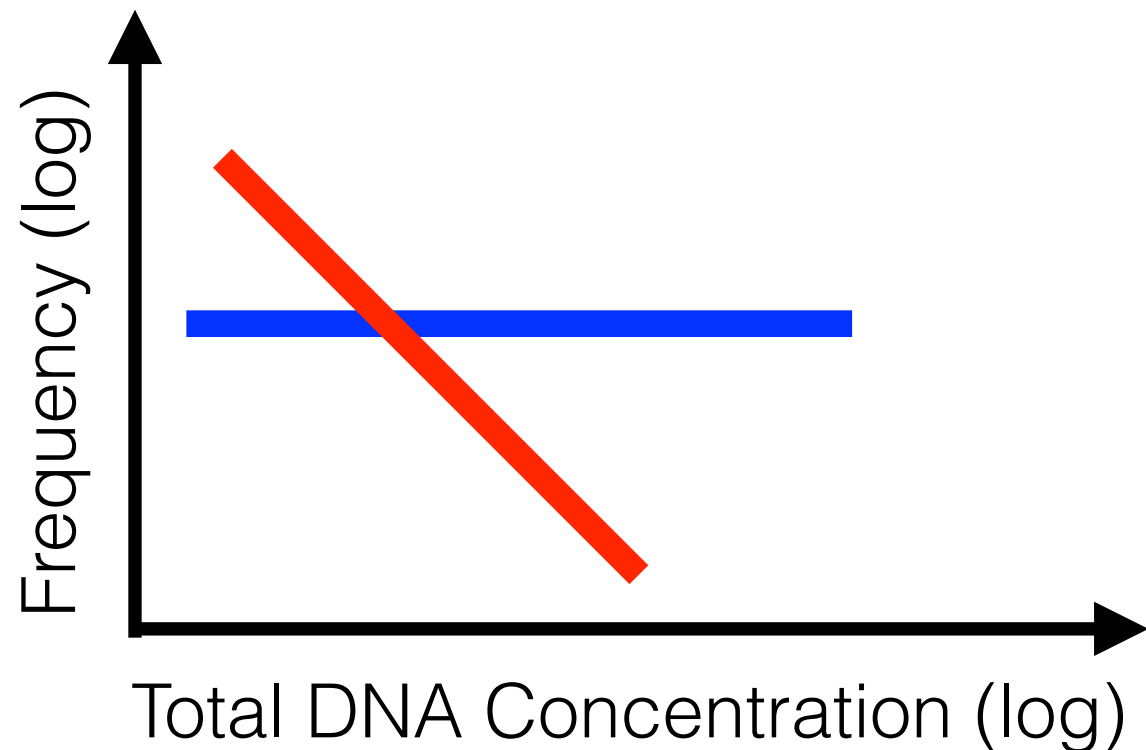
$f_c = C/(S+C) \sim 1/T$, where **$C \ll S$**

Modeling Contaminants

$T = S + C$, where **C** is constant

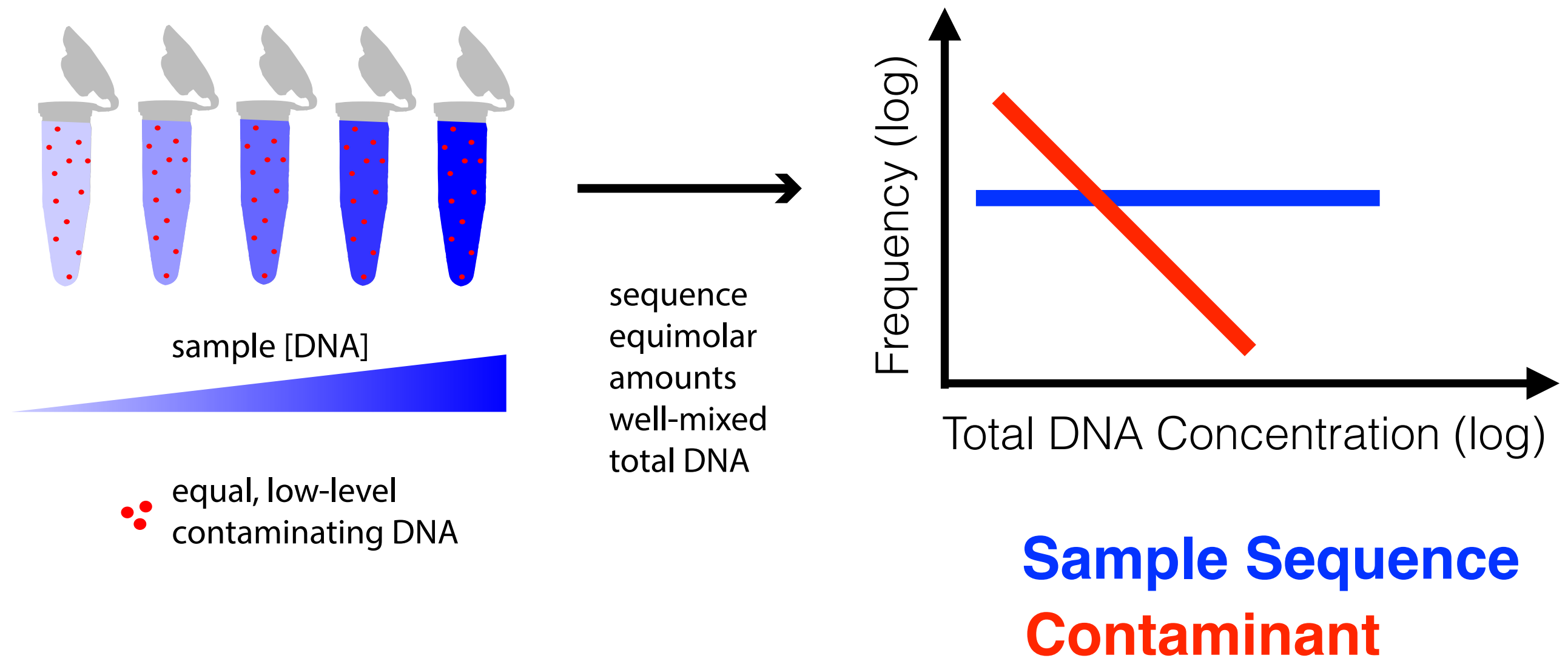
hence

$f_c = C/(S+C) \sim 1/T$, where **$C \ll S$**

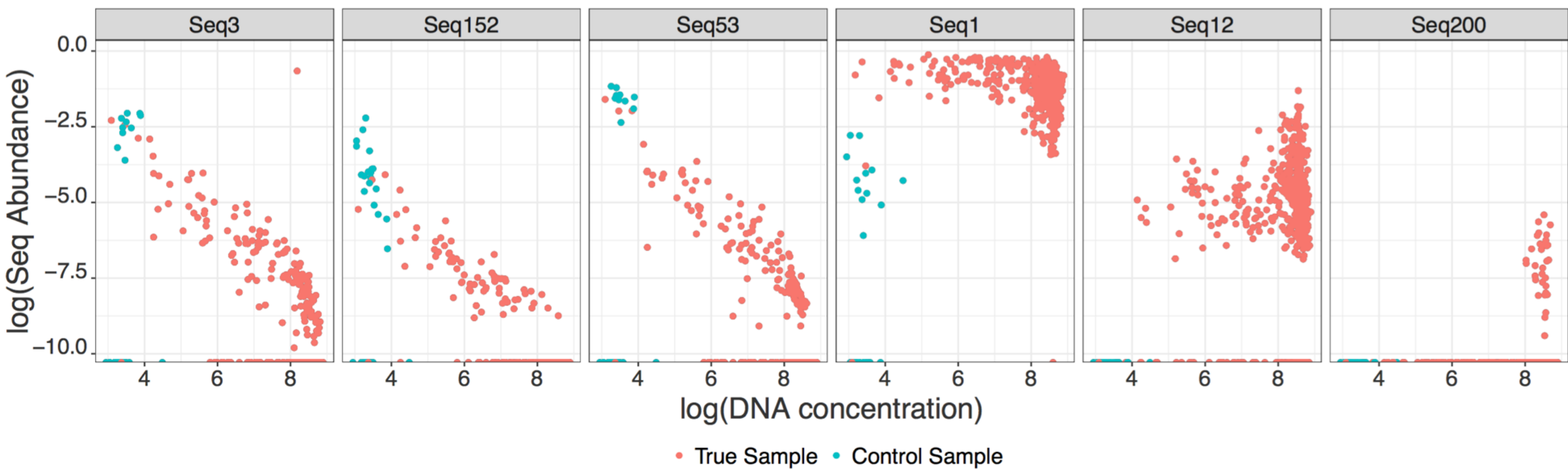


Sample Sequence
Contaminant

Modeling Contaminants

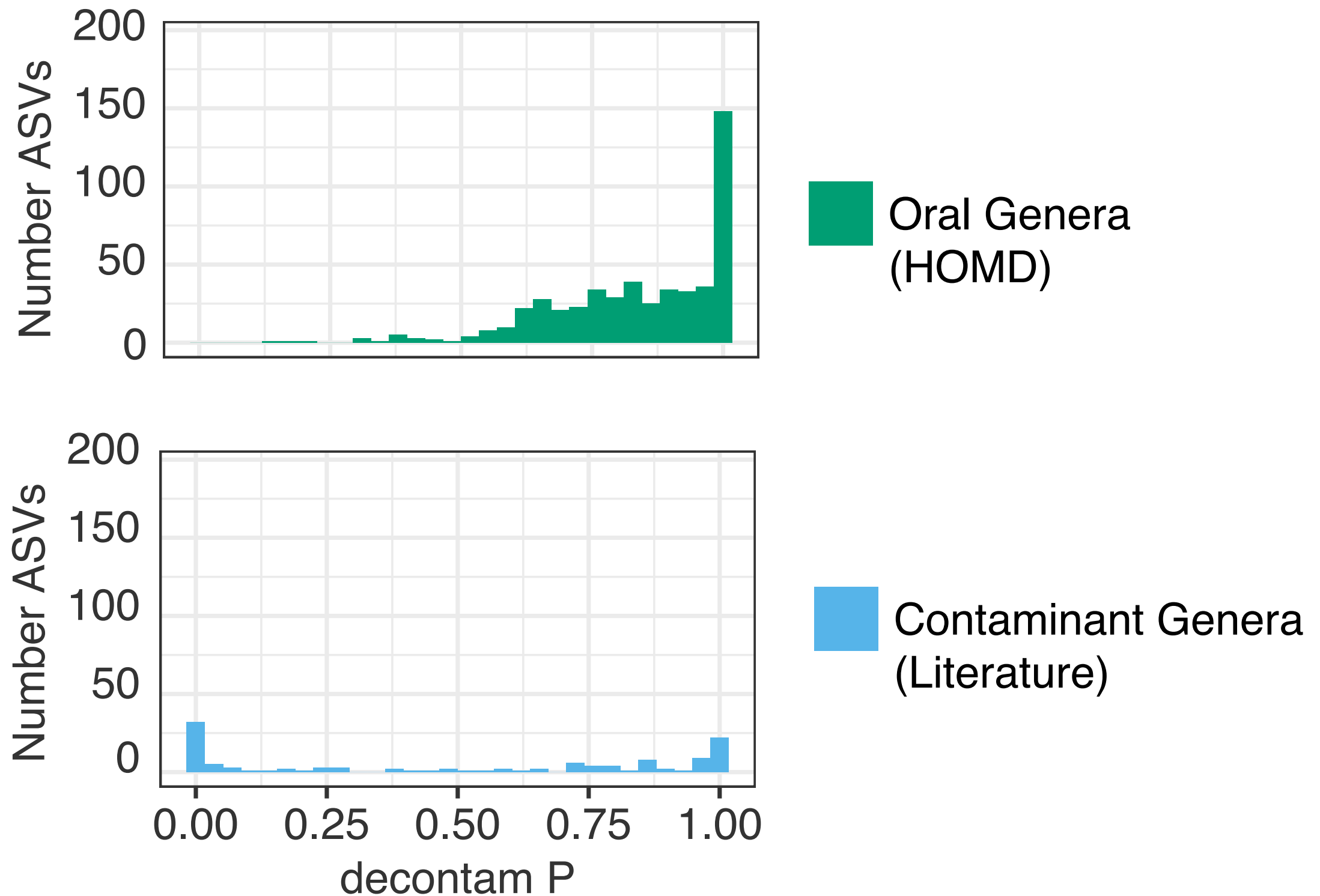


Validating the Model



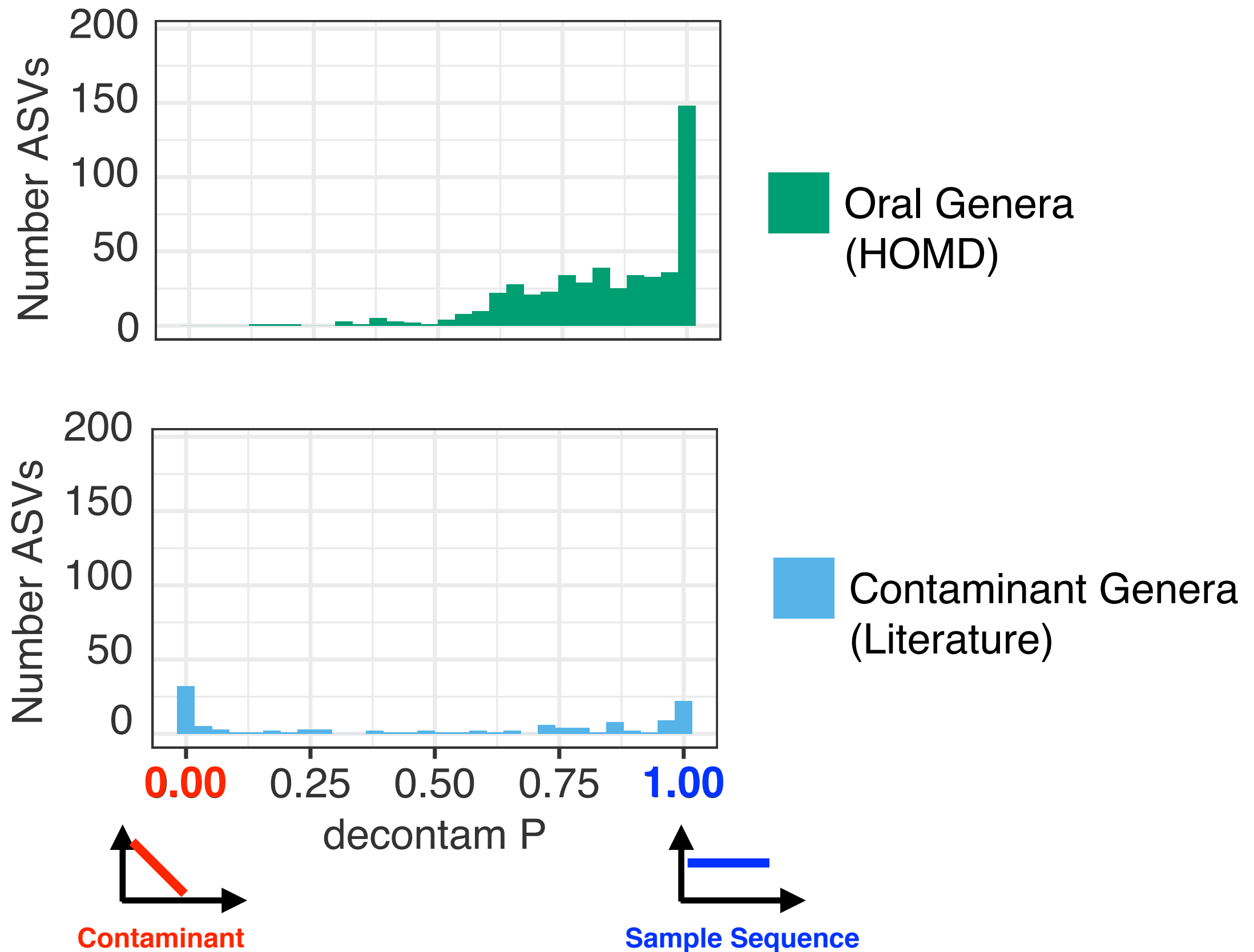
Validating the Model

Oral Mucosal Dataset



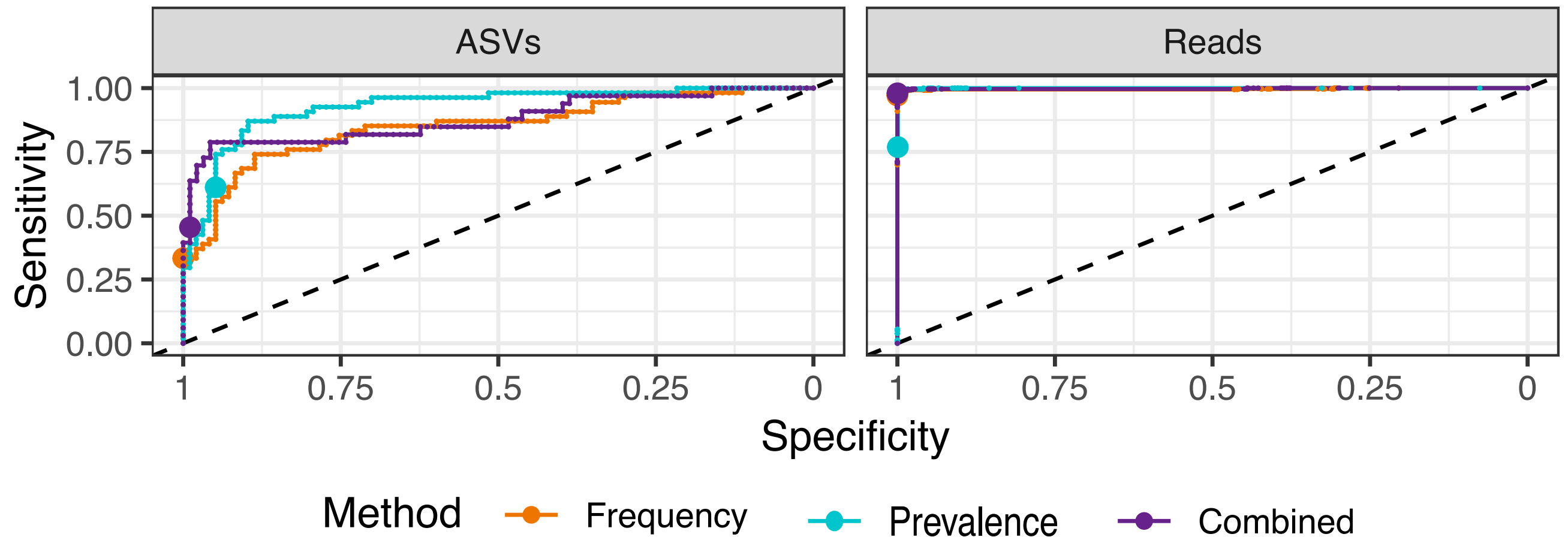
Validating the Model

Oral Mucosal Dataset



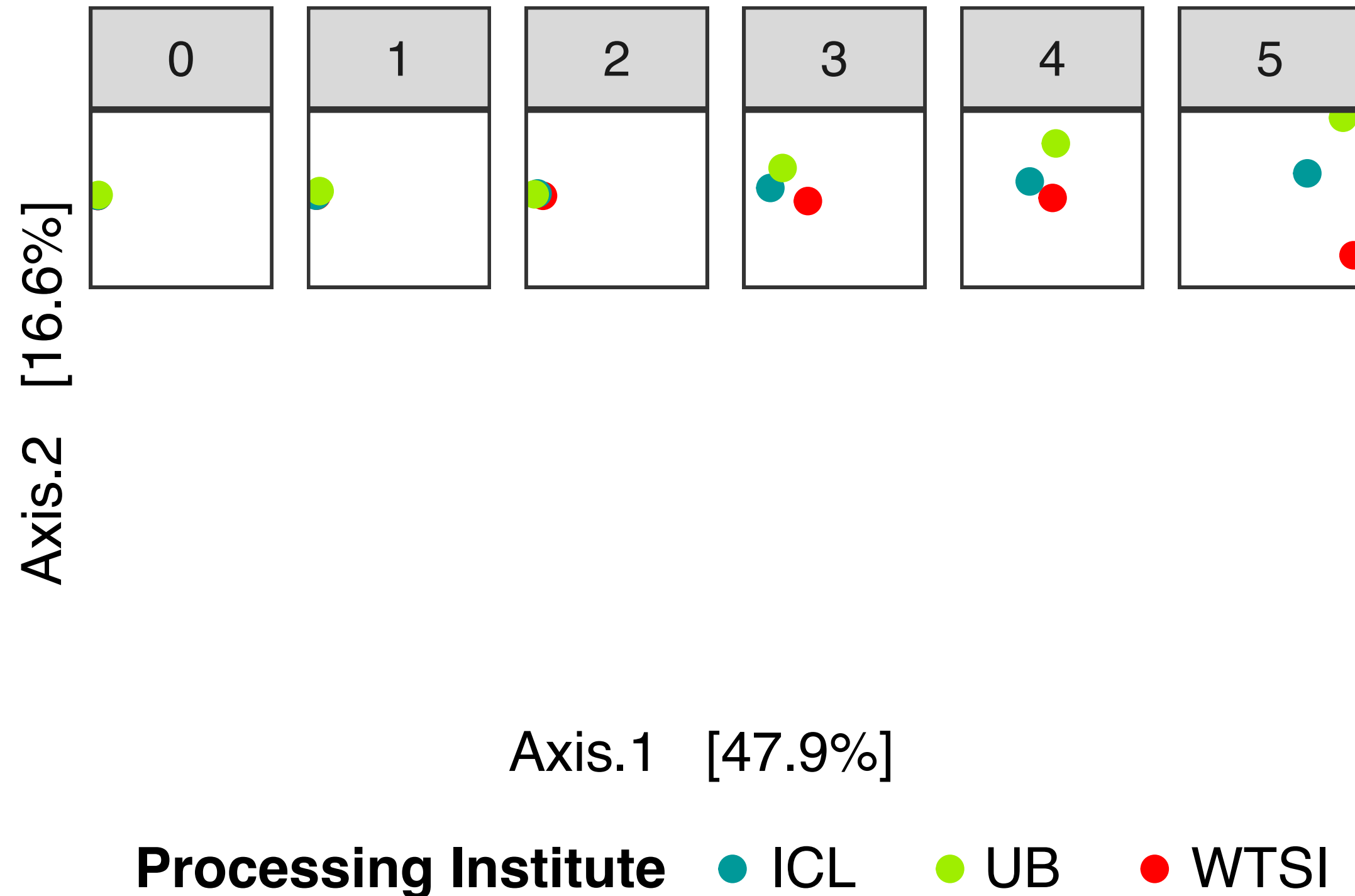
Classification Accuracy

Oral Mucosal Dataset



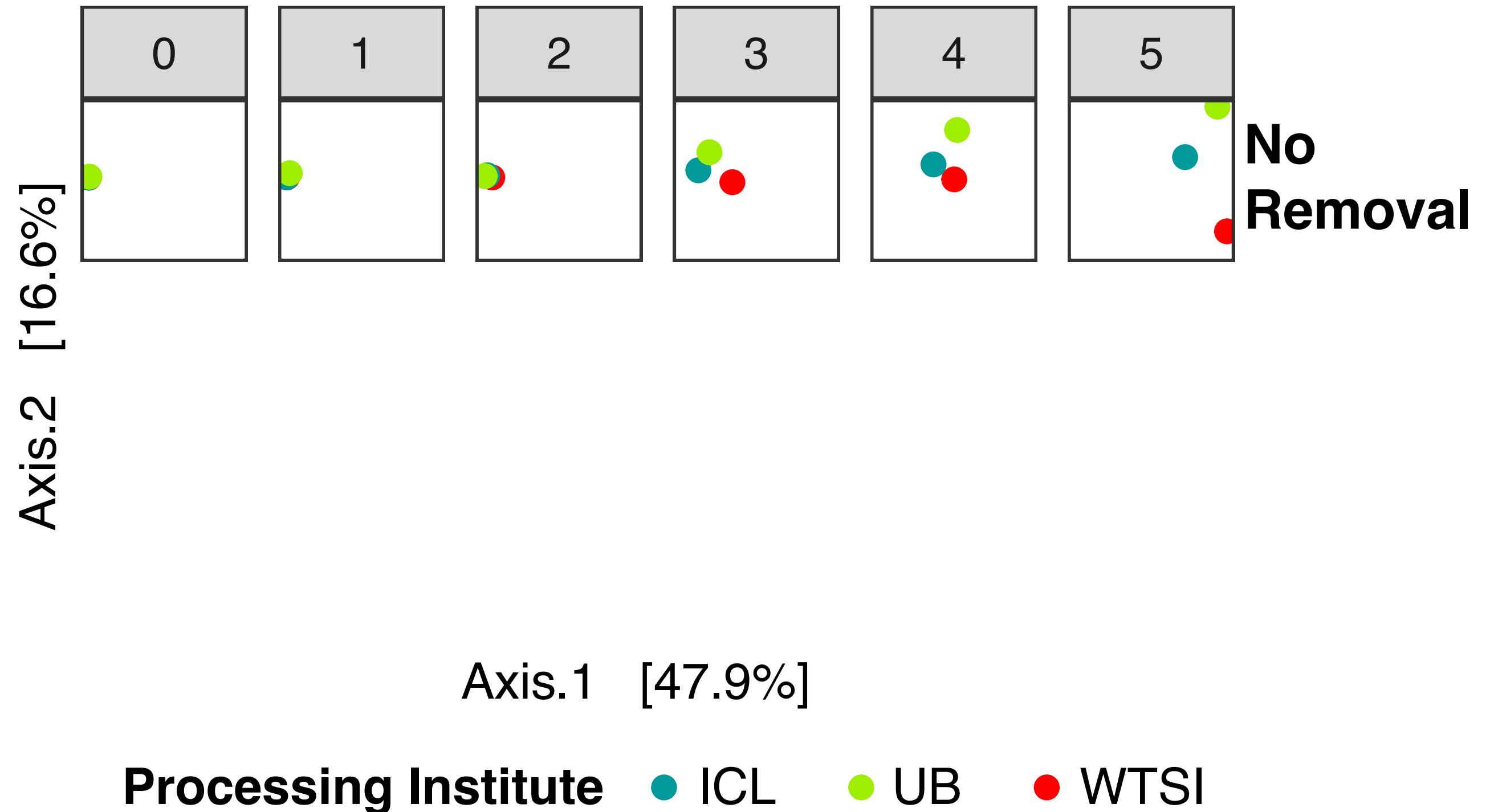
Reducing Technical Variation

Salmonella bongori: Ten-fold dilutions



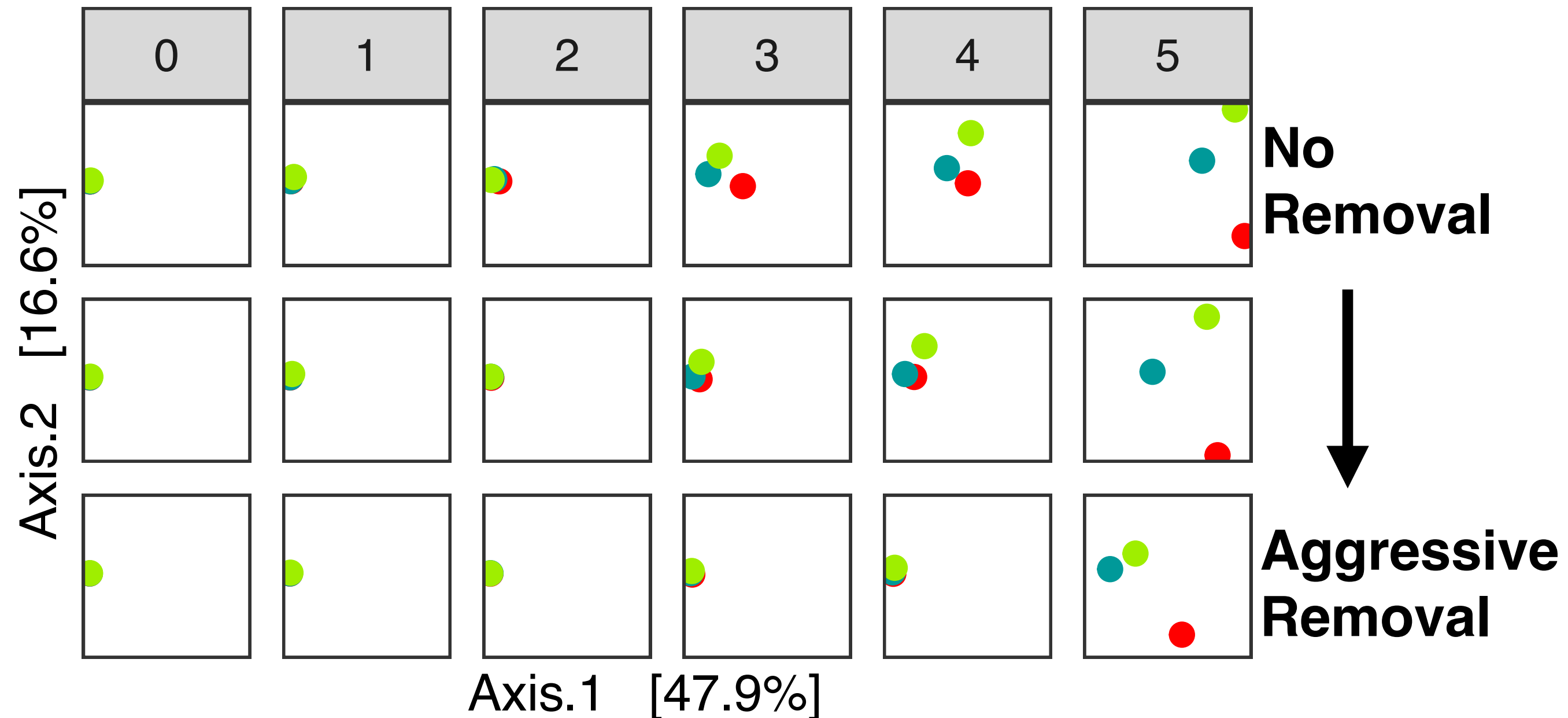
Reducing Technical Variation

Salmonella bongori: Ten-fold dilutions



Reducing Technical Variation

Salmonella bongori: Ten-fold dilutions



Processing Institute ● ICL ● UB ● WTSI

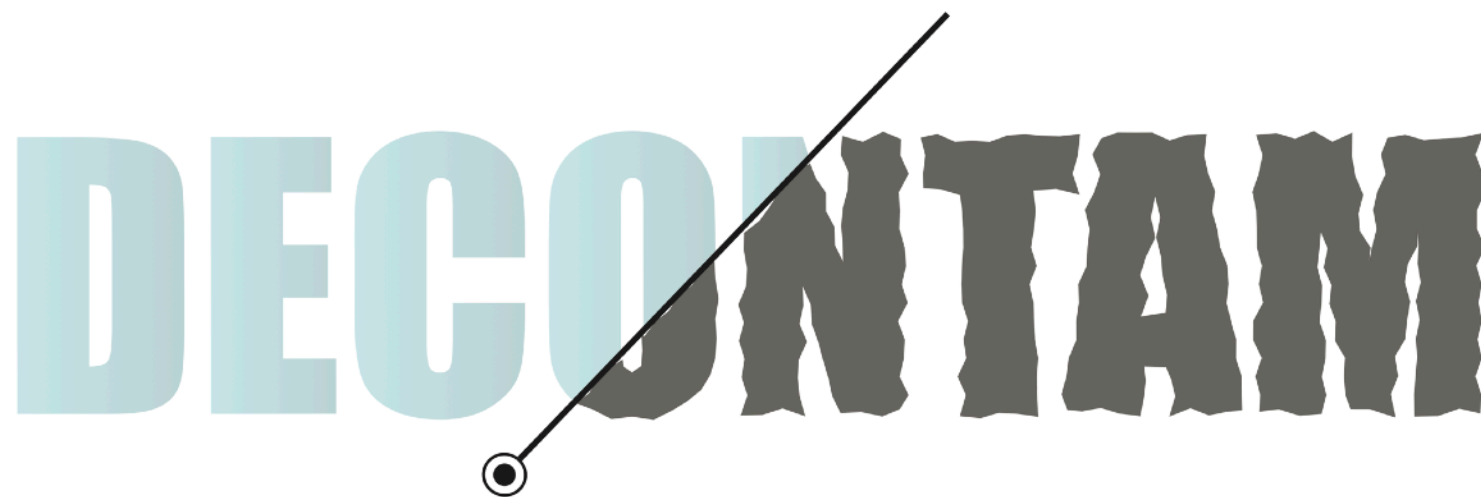
Available now...

Methodology | Open Access

Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman and Benjamin J. Callahan  

Microbiome 2018 6:226



- Open-source
- Well documented
- R package
- 16S or shotgun

Recommendations...

- **There is no substitute for clean lab practices**
- **Sequence** *multiple* full-process negative controls!
- Consider dilution series of a positive control
- Record DNA concentrations
- *In silico* decontamination (at high resolution)
- Be skeptical of unexpected or implausible taxa
- Report taxa in negative controls

Acknowledgements

Exact Sequence Variants



Susan Holmes



Joey McMurdie



Michael Rosen

...for ITS



Naga Betrapally

Contamination



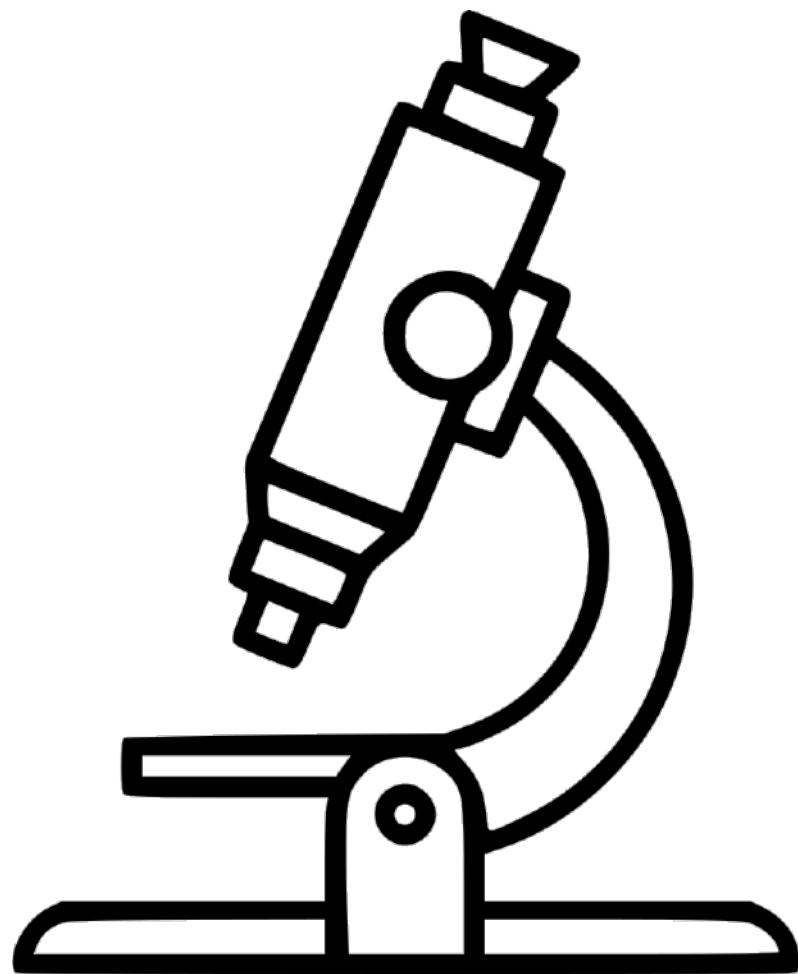
Nicole Davis

Bias

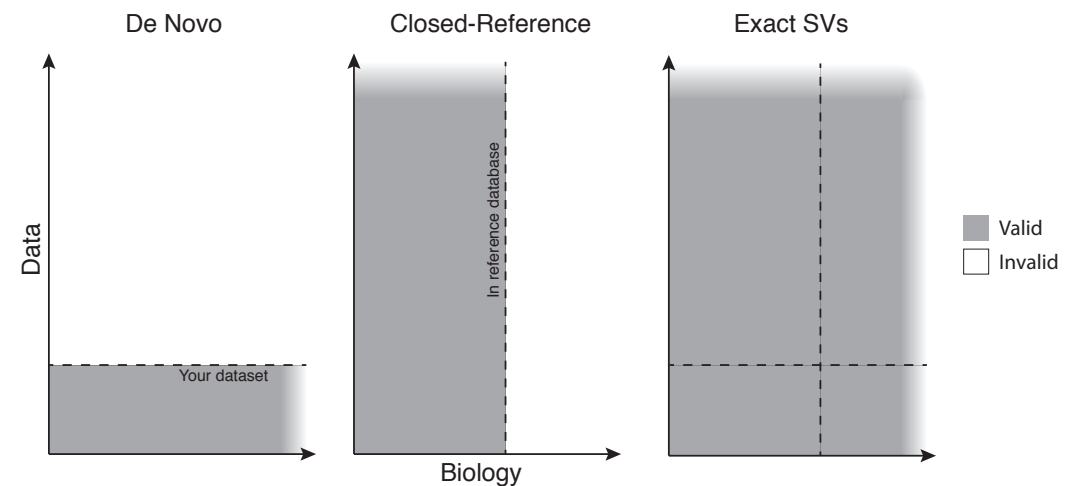


Michael McLaren

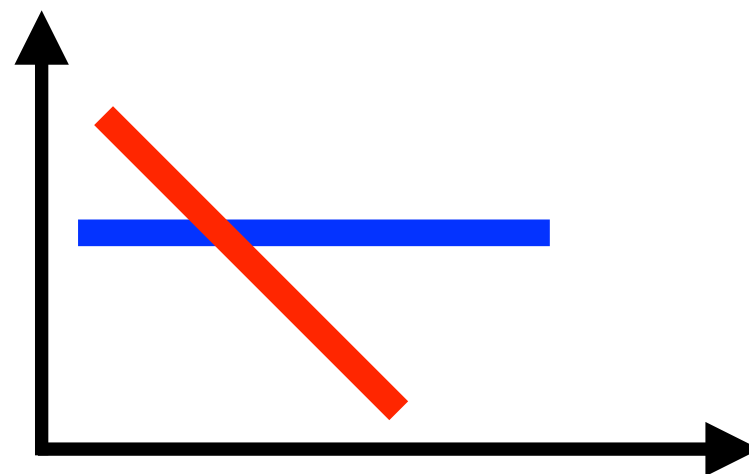




Exact Sequence Variants



(De-)Contamination



Sample Sequences
Contaminant

Bias

