Lecture: Decontamination

Nikolay Oskolkov

Before we start!

```
$ cd /vol/volume/contamination
$ ls
```

You should see at least a file called <session_name>.yaml

```
$ conda activate contamination
$ which decOM
```

Should see a path with miniconda/ in it

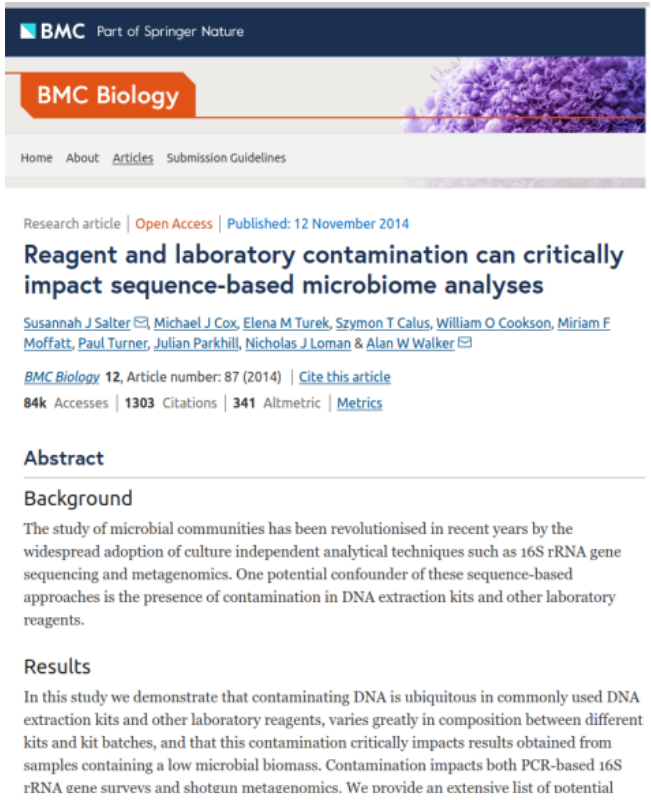
Outline

- Contamination problem
  - DNA of modern (microbial) organisms in ancient / historical samples
  - Animal / plant / bird / fish etc. references include microbial contamination

- Microbiome contamination correction
  - Decontamination via negative controls (blanks)
  - Similarity to expected microbiome source (microbial source tracking)

Contamination problem

Modern microbial and animal DNA contamination



BMC Part of Springer Nature

BMC Biology

Home About Articles Submission Guidelines

Research article | Open Access | Published: 12 November 2014

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman & Alan W Walker

BMC Biology **12**, Article number: 87 (2014) | Cite this article

84k Accesses | **1303** Citations | **341** Altmetric | Metrics

Abstract

Background

The study of microbial communities has been revolutionised in recent years by the widespread adoption of culture independent analytical techniques such as 16S rRNA gene sequencing and metagenomics. One potential confounder of these sequence-based approaches is the presence of contamination in DNA extraction kits and other laboratory reagents.

Results

In this study we demonstrate that contaminating DNA is ubiquitous in commonly used DNA extraction kits and other laboratory reagents, varies greatly in composition between different kits and kit batches, and that this contamination critically impacts results obtained from samples containing a low microbial biomass. Contamination impacts both PCR-based 16S rRNA gene surveys and shotgun metagenomics. We provide an extensive list of potential

ELSEVIER

Journal of Archaeological Science

Volume 34, Issue 9, September 2007, Pages 1361-1366

Animal DNA in PCR reagents plagues ancient DNA research

Jennifer A. Leonard [a, b, c], Orin Shanks [d, 1], Michael Hofreiter [e], Eva Kreuz [e], Larry Hodges [d], Walt Ream [d], Robert K. Wayne [b], Robert C. Fleischer [a]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.jas.2006.10.023 Get rights and content

Abstract

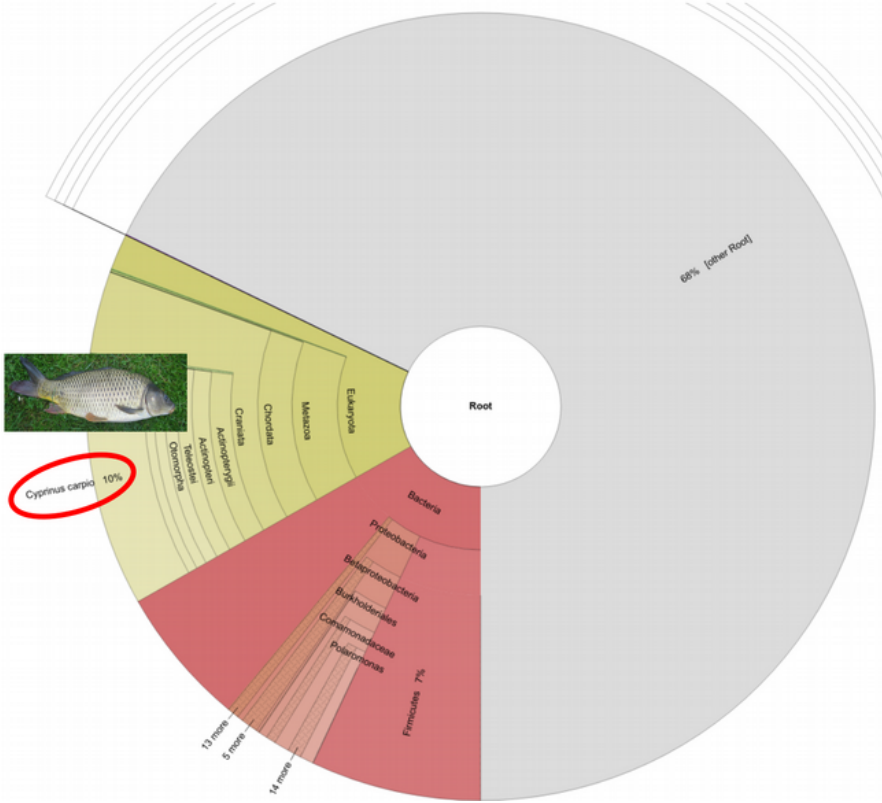
Molecular archaeology brings the tools of molecular biology to bear on fundamental questions in archaeology, anthropology, evolution, and ecology. Ancient DNA research is becoming widespread as evolutionary biologists and archaeologists discover the power of the polymerase chain reaction (PCR) to amplify DNA from ancient plant and animal remains. However, the extraordinary susceptibility of PCR to contamination by extraneous DNA is not widely appreciated. We report the independent observation of DNA from domestic animals in PCR reagents and ancient samples in four separate laboratories. Since PCR conditions used in ancient DNA analyses are extremely sensitive, very low concentrations of contaminating DNA can cause false positives. Previously unidentified animal DNA in reagents can confound ancient DNA research on certain domestic animals, especially cows, pigs, and chickens.

Previous article in issue Next article in issue

Keywords

Sus scrofa; *Bos taurus*; *Gallus gallus*; Deoxynucleoside triphosphates

Contaminated reference databases



It turns out the Carp genome is full of Illumina adapters.

One of the first things we teach people in our NGS courses is how to remove adapters. It's not hard – we use CutAdapt, but many other tools exist. It's simple, but really important – with De Bruijn graphs you will get paths through the graphs converging on kmers from adapters; and with OLC assemblers you will get spurious overlaps. With gap-fillers, it's possible to fill the gaps with sequences ending in adapters, and this may be what happened in the Carp genome.

Why then are we finding such elementary mistakes in such important papers? Why aren't reviewers picking up on this? It's frustrating.

This is a separate, but related issue, to genomic contamination – the Wheat genome has PhiX in it; tons of bacterial genomes do too; and lots of bacterial genes were problematically included in the Tardigrade genome and declared as horizontal gene transfer.

Bioinformatics Bits and Bobs

Rare and random blog posts about bioinformatics, genomics and evolution.

Monday, 29 September 2014

Why you should QC your reads AND your assembly

The genome sequence of the Common Carp *Cyprinus carpio* was published in Nature last week. By coincidence, I was doing some QC on some domesticated Ferret (*Mustela ptorius furo*) reads, which had thrown some kmer warnings in the FastQC tool. I blasted the kmers in NCBI and was quite perplexed by the number of hits that I found in the carp genome. Nearly all of the first 150 hits were all from the carp genome. Anyway, I looked a bit further into my odd kmers and it turns out that they were the ends of some Illumina adapter sequences that had presumably been incorporated into the paired-reads on the shorter ends of the insert size. This then took me back to the Carp Genome - what had creeped into that?

Microbial contamination of RefSeq genomes

Tibetan antelope (*Pantholops hodgsonii*) turns up in every metagenomic dataset 🤔

Pantholops hodgsonii as well as *Bos mutus / taurus* and a few other mammals have severe microbial contamination (Oskolkov et al. 2025)

Microbial insert

Bos mutus reference genome

Eukaryotic contamination in microbial references

RefSeq genome of the common soil bacterium *Achromobacter denitrificans* contains the entire chicken ovalbumin gene!

Spirometra erinaceieuropaei (tapeworm) seems to have human contamination (Jensen et al., Nat. Comm. 2019)

Microbiome Contamination Correction

Contaminants found on negative controls

Table 1 List of contaminant genera detected in sequenced negative `blank' controls

From: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Phylum	List of constituent contaminant genera
Proteobacteria	Alpha-proteobacteria:
	Afipia, Aquabacterium[e], *Asticcacaulis, Aurantimonas, Beijerinckia, Bosea, Bradyrhizobium*[d], *Brevundimonas*[c], *Caulobacter, Craurococcus, Devosia, Hoeflea*[e], *Mesorhizobium, Methylobacterium*[c], *Novosphingobium, Ochrobactrum, Paracoccus, Pedomicrobium, Phyllobacterium*[e], *Rhizobium*[c,d], *Roseomonas, Sphingobium, Sphingomonas*[c,d,e], *Sphingopyxis*
	Beta-proteobacteria:
	Acidovorax[c,e], *Azoarcus*[e], *Azospira, Burkholderia*[d], *Comamonas*[c], *Cupriavidus*[c], *Curvibacter, Delftia*[e], *Duganella*[a], *Herbaspirillum*[a,c], *Janthinobacterium*[e], *Kingella, Leptothrix*[a], *Limnobacter*[e], *Massilia*[c], *Methylophilus, Methyloversatilis*[e], *Oxalobacter, Pelomonas, Polaromonas*[e], *Ralstonia*[b,c,d,e], *Schlegelella, Sulfuritalea, Undibacterium*[e], *Variovorax*
	Gamma-proteobacteria:
	Acinetobacter[a,d,c], *Enhydrobacter, Enterobacter, Escherichia*[a,c,d,e], *Nevskia*[e], *Pseudomonas*[b,d,e], *Pseudoxanthomonas, Psychrobacter, Stenotrophomonas*[a,b,c,d,e], *Xanthomonas*[b]
Actinobacteria	*Aeromicrobium, Arthrobacter, Beutenbergia, Brevibacterium, Corynebacterium, Curtobacterium, Dietzia, Geodermatophilus, Janibacter, Kocuria, Microbacterium, Micrococcus, Microlunatus, Patulibacter, Propionibacterium*[e], *Rhodococcus, Tsukamurella*
Firmicutes	*Abiotrophia, Bacillus*[b], *Brevibacillus, Brochothrix, Facklamia, Paenibacillus, Streptococcus*
Bacteroidetes	*Chryseobacterium, Dyadobacter, Flavobacterium*[d], *Hydrotalea, Niastella, Olivibacter, Pedobacter, Wautersiella*
Deinococcus-Thermus	*Deinococcus*
Acidobacteria	Predominantly unclassified Acidobacteria Gp2 organisms

The listed genera were all detected in sequenced negative controls that were processed alongside human-derived samples in our laboratories (WTSI, ICL and UB) over a period of four years. A variety of DNA extraction and PCR kits were used over this period, although DNA was primarily extracted using the FastDNA SPIN Kit for Soil. Genus names followed by a superscript letter indicate those that have also been independently reported as contaminants previously. [a]also reported by Tanner *et al.* [12]; [b]also reported by Grahn *et al.* [14]; [c]also reported by Barton *et al.* [17]; [d]also reported by Laurence *et al.* [18]; [e]also detected as contaminants of multiple displacement amplification kits (information provided by Paul Scott, Wellcome Trust Sanger Institute). ICL, Imperial College London; UB, University of Birmingham; WTSI, Wellcome Trust Sanger Institute.

Compare abundances in samples and blanks

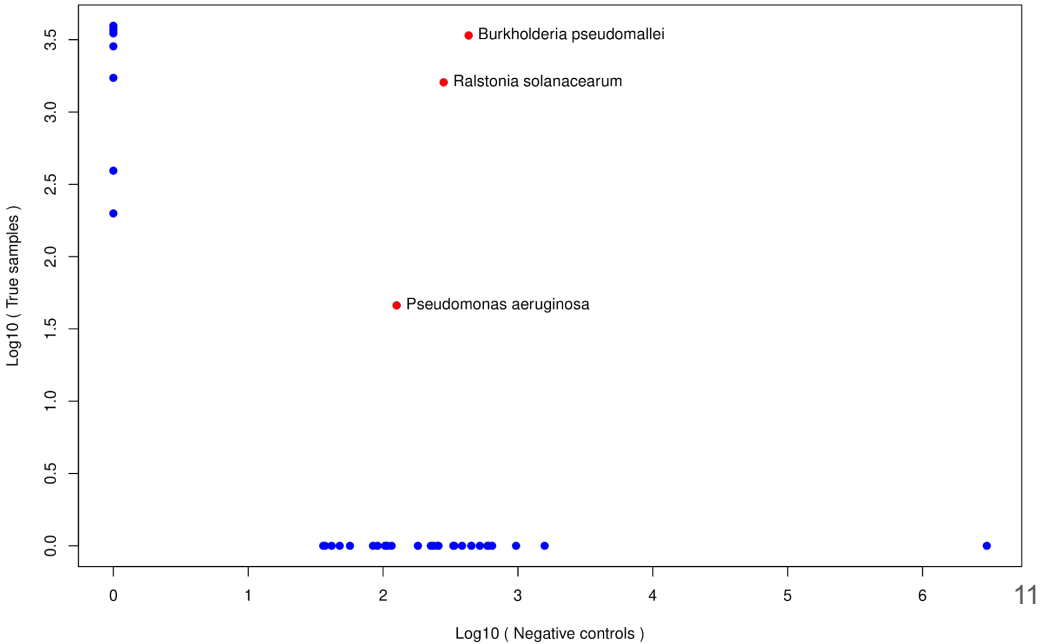
Highlight organisms abundant in both samples and blanks:

```
Rscript ${SCRIPTS_DIR}/blank_decontam.R krakenuniq_abundance_matrix.txt \
blank_krakenuniq_abundance_matrix.txt
```

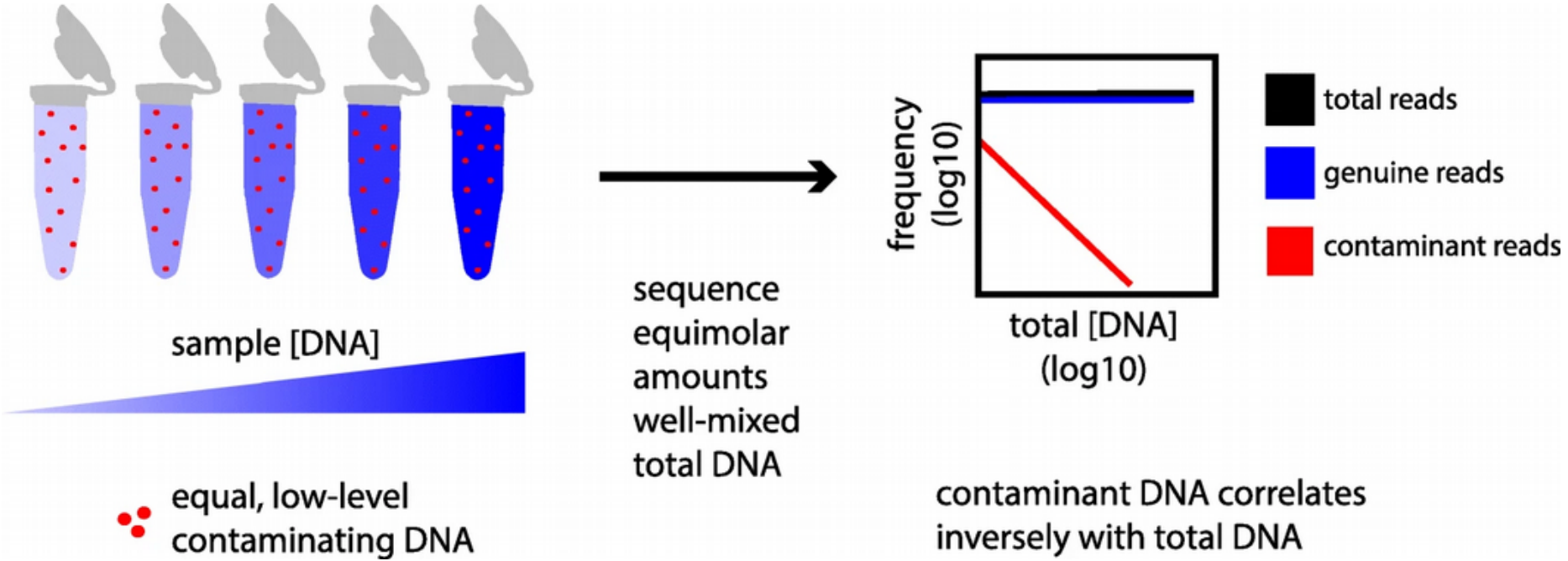
Most organisms are highly abundant either only in sample or only in blanks.

Organisms that are **highly abundant** both in samples and blank negative controls are likely of **exogenous origin**.

Recentrifuge performs similar filtering.

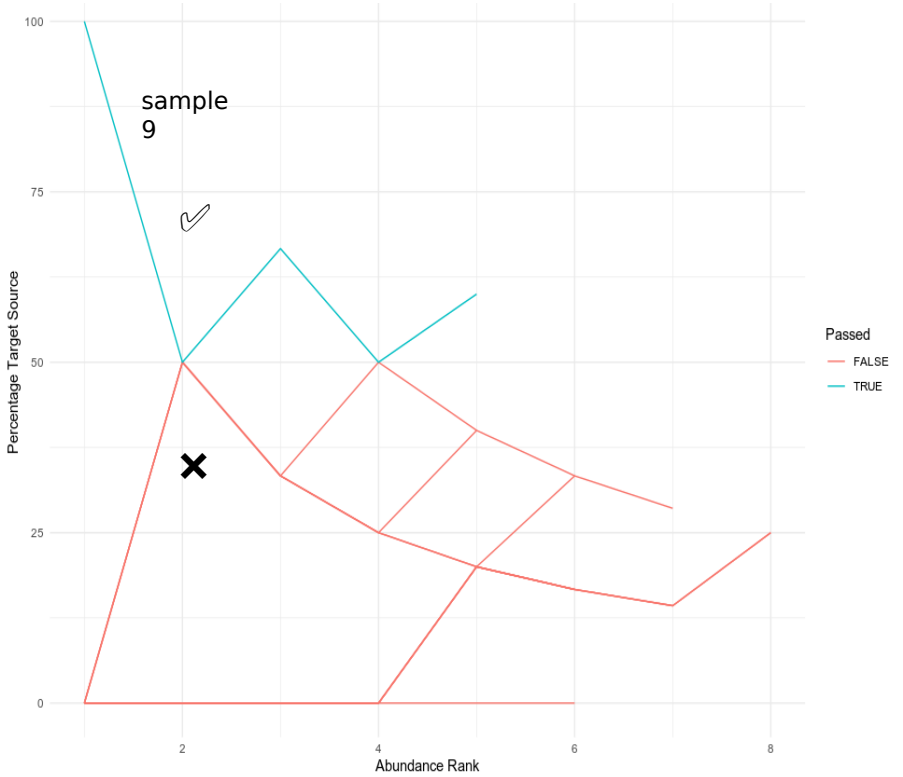


Other approaches: *decontam* R package



Other approaches: *cuperdec* R package

```
library(cuperdec);library(magrittr);library(dplyr)

# Load database (in this case oral database)
data(cuperdec_database_ex)
database <- load_database(cuperdec_database_ex,
target = "oral") %>% print()

# Load abundance matrix and metadata
abundance <- "krakenuniq_abundance_matrix.txt"
taxatable <- load_taxa_table(abundance)
metadata <- as_tibble(data.frame(Sample =
unique(taxatable$Sample), Sample_Source = "Oral"))

# Compute cumulative percent decay curves
curves <- calculate_curve(taxatable, database)
filter_result <- simple_filter(curves,
percent_threshold = 50) %>% print()
plot_cuperdec(curves, metadata = metadata,
filter_result)
```

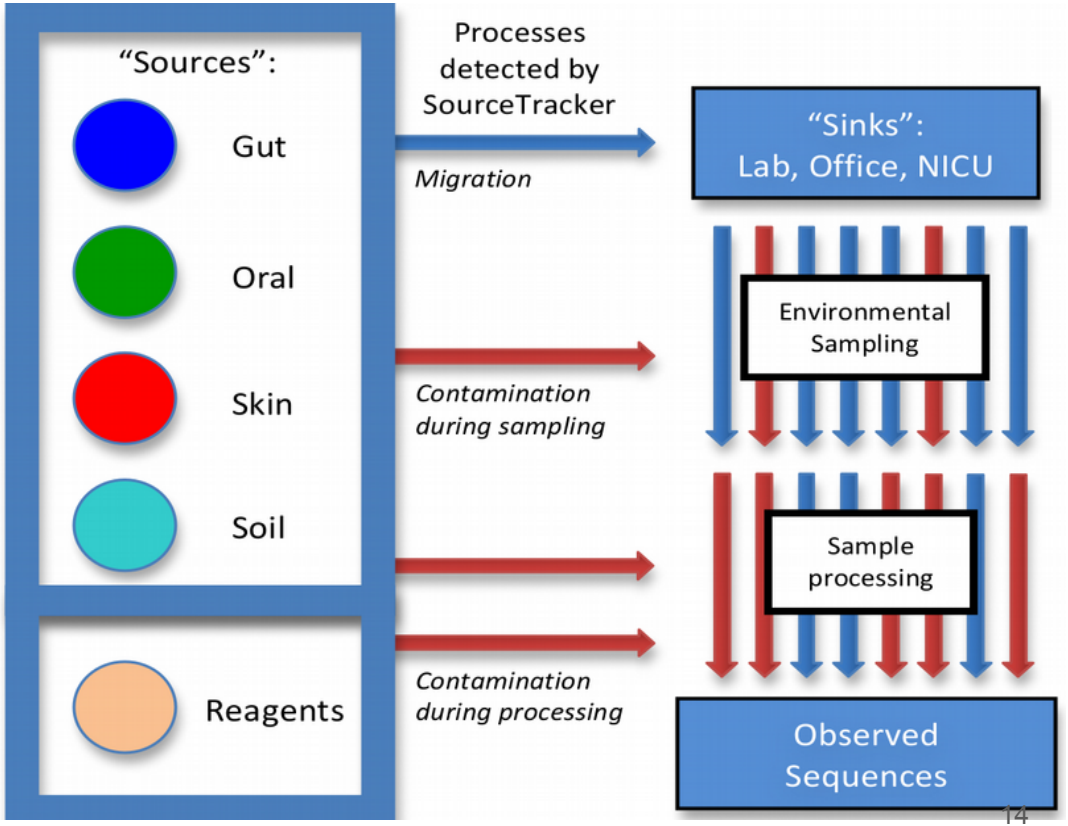


Microbial Source Tracking (MST): SourceTracker

SourceTracker is a Bayesian classifier that infers **environments** that detected microbial organisms originate from.

SourceTracker is trained on **source** samples with known environmental annotation, and makes predictions on **sink** samples.

Extensions are mSourceTracker and FEAST

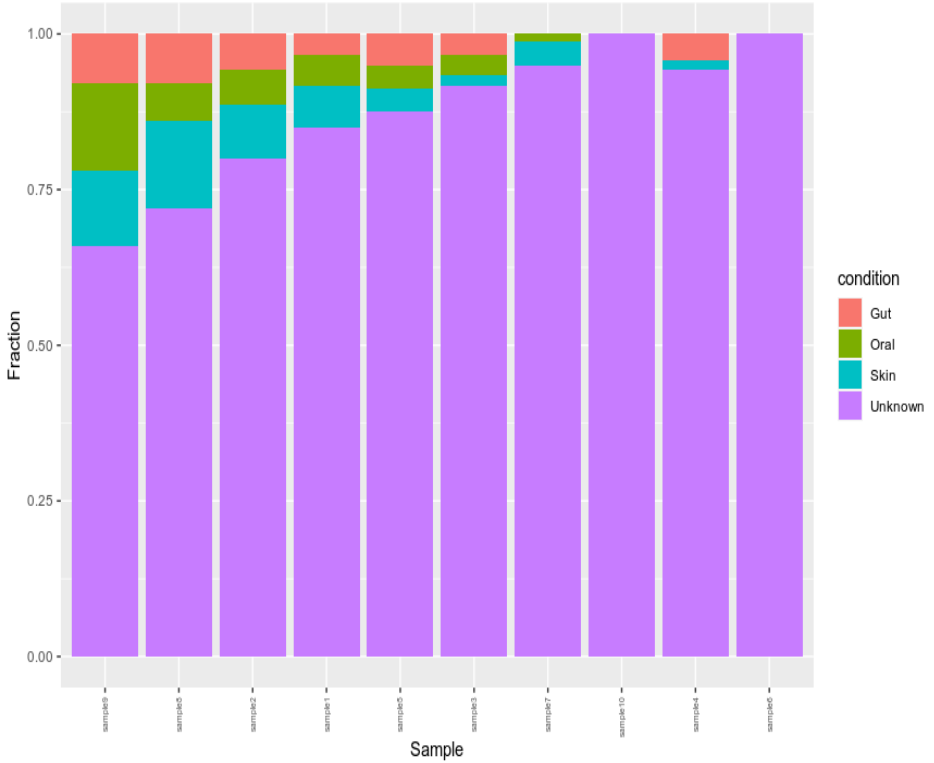


Merge sources and sinks, and train

```
otus_hmp <- read.delim("otus_hmp.txt", header = TRUE, row.names = 1, sep = "\t")
meta_hmp <- read.delim("meta_hmp.txt", header = TRUE, row.names = 1, sep = "\t")

otus_sink<-read.delim("krakenuniq_abundance_matrix.txt", header=TRUE, row.names=1, sep="\t")
otus <- merge(otus_hmp, otus_sink, all = TRUE, by = "row.names")
rownames(otus) <- otus$Row.names; otus$Row.names <- NULL; otus[is.na(otus)] <- 0
meta_sink <- data.frame(ID = colnames(otus_sink), Env = "Unknown", SourceSink = "sink")
rownames(meta_sink) <- meta_sink$ID; meta_sink$ID<-NULL; metadata <- rbind(meta_hmp, meta_sink)

otus <- as.data.frame(t(as.matrix(otus))); otus[otus > 0] <- 1; otus <- otus[rowSums(otus)!=0,]
metadata<-metadata[as.character(metadata$Env)!="Vaginal",]; envs <- metadata$Env
common.sample.ids <- intersect(rownames(metadata), rownames(otus))
otus <- otus[common.sample.ids,]; metadata <- metadata[common.sample.ids,]

# Train SourceTracker on sources (HMP) and run predictions on sinks
source('/sourcetracker/src/SourceTracker.r')
train.ix <- which(metadata$SourceSink=='source'); test.ix <- which(metadata$SourceSink=='sink')
st <- sourcetracker(otus[train.ix,], envs[train.ix])
results <- predict(st, otus[test.ix,], alpha1 = 0.001, alpha2 = 0.001)
```

Plot SourceTracker results

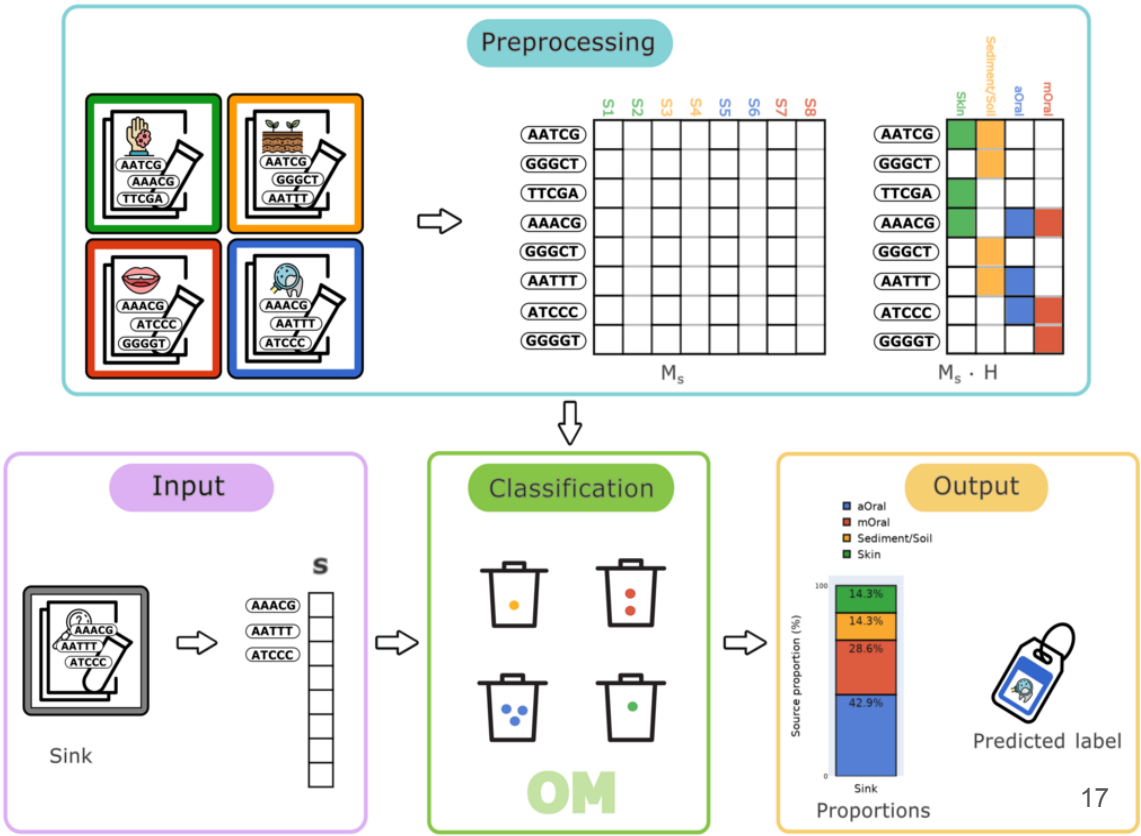
```
# Sort SourceTracker proportions for plotting
props <- results$proportions
props <- props[order(-props[,"Oral"]),]
results$proportions <- props

# Prepare SourceTracker output for plotting
name<-rep(rownames(results$proportions),each=4)
value <- as.numeric(t(results$proportions))
labels <- c("Gut","Oral","Skin","Unknown")
condition<-rep(labels, length(test.ix))
data <- data.frame(name, condition, value)

# Plot SourceTracker inference as a barplot
library("ggplot2")
ggplot(data, aes(fill=condition, y=value, x =
reorder(name, seq(1:length(name))))) +
geom_bar(position = "fill", stat = "identity") +
theme(axis.text.x = element_text(angle = 90,
size=5,hjust=1,vjust=0.5)) + xlab("Sample") +
ylab("Fraction")
```



Read-level source tracking with decOM

SourceTracker requires a microbial **abundance matrix** computed by e.g. QIIME or Kraken based on a **taxonomically annotated database**.

In contrast, **decOM** aims at **moving away from database-dependent** methods and using unsupervised approaches exploiting **read-level** sequence composition.



Perform source tracking with decOM

Prepare input *fof*-files that have a key - value format:

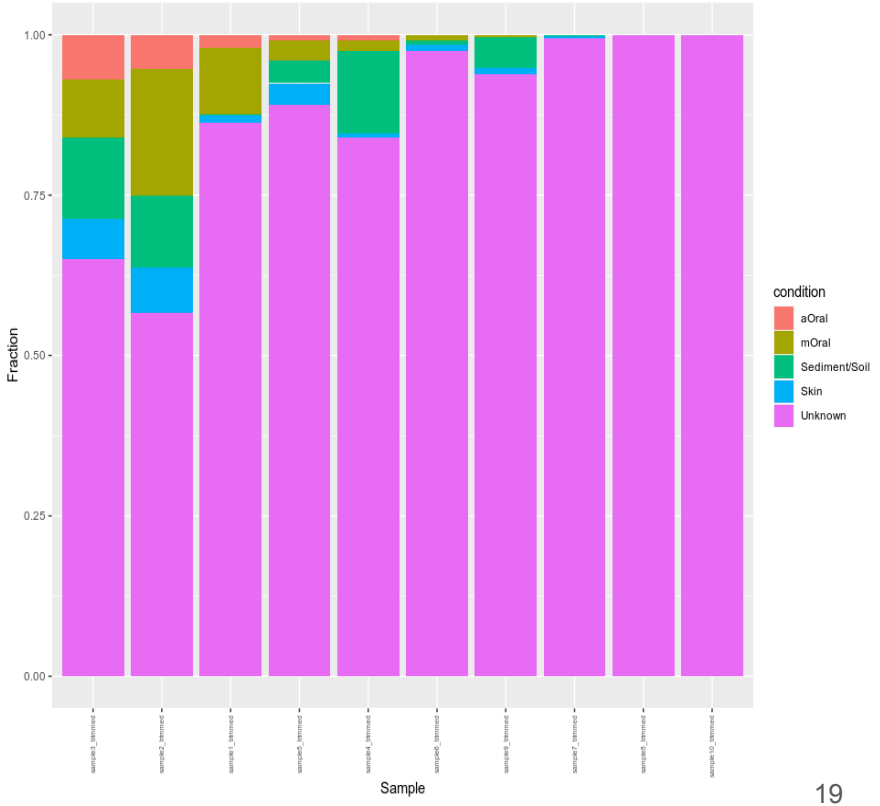
```
cd CUTADAPT # folder containing trimmed fastq-files
for i in {1..10}
do
echo "sample${i}_trimmed : sample${i}_trimmed.fastq.gz" > sample${i}_trimmed.fof
echo sample${i}_trimmed >> FASTQ_NAMES_LIST.txt
done
```

Download prebuilt kmer-matrix of sources (aOral,mOral,Sediment/Soil,Skin), run decOM:

```
wget https://zenodo.org/record/6513520/files/decOM_sources.tar.gz
tar -xf decOM_sources.tar.gz


decOM -p_sources decOM_sources/ -p_sinks FASTQ_NAMES_LIST.txt -p_keys CUTADAPT -mem 10GB
-t 5
```

Plot decOM source tracking results

```
df<-read.csv("decOM_output.csv",check.names=FALSE)

result<-subset(df,select=c("Sink","Sediment/Soil",
"Skin", "aOral", "mOral", "Unknown"))
rownames(result) <- result$Sink; result$Sink<-NULL
result <- result / rowSums(result)
result <- result[order(-result$aOral), ]

name <- rep(rownames(result), each = 5)
value <- as.numeric(t(result))
condition <- rep(c("Sediment/Soil","Skin","aOral",
"mOral","Unknown"), dim(result)[1])
data <- data.frame(name, condition, value)

library("ggplot2")
ggplot(data, aes(fill = condition, y = value,
x=reorder(name,seq(1:length(name))))) +
  geom_bar(position = "fill", stat = "identity") +
  theme(axis.text.x=element_text(angle=90, size=5,
hjust=1,vjust=0.5))+xlab("Sample")+ylab("Fraction")
```



Summary

- Negative controls are important for disentangling ancient / endogenous from modern / exogenous contamination
- Microbial source tracking is another layer of evidence that can facilitate interpretation of ancient metagenomic findings