NC STATE UNIVERSITY

NC STATE College of Veterinary Medicine

BIOINFORMATICS RESEARCH CENTER

Identification of Contaminant Using Hypothesis Testing in Marker Gene and Metagenomics Data

Caizhi Huang[1], Craig Gin[2], Jung-Ying Tzeng[1,3], Benjamin Callahan[1,2]

[1]NC State University, Bioinformatics Research Center, [2]NC State University, Department of Population Health and Pathobiology, [3]NC State University, Department of Statistics

Abstract

Background: The measurement of microbial community suffers from contaminant DNA sequences that are not truly present in the sample (Figure 1). *Decontam* has been introduced to identify contaminant sequences using a classification procedure based on a pattern that contaminant appears high frequencies in low-concentration samples (Figure 2). However, it has no false discovery rate control, and clear guidance is missing to help users choose an interpretable threshold.

Results: We propose a hypothesis testing procedure, *Tcontam*, to detect contaminants using statistical p-value and control the false discovery rate using multiple testing correction procedure. We confirmed validity of *Tcontam* using simulation. In a human oral dataset, *Tcontam* reports the contaminants with false discovery rate under control and has low chance to classify the sequences with small sample size as contaminants.

Contaminants

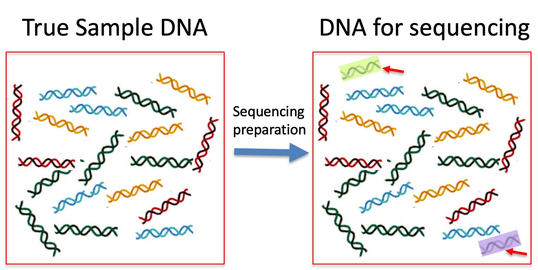


Figure 1: Schematic of contaminant DNA sequences that are not truly present been introduced in marker-gene and metagenomic sequencing (MGS) procedure.

Modeling of Contaminants

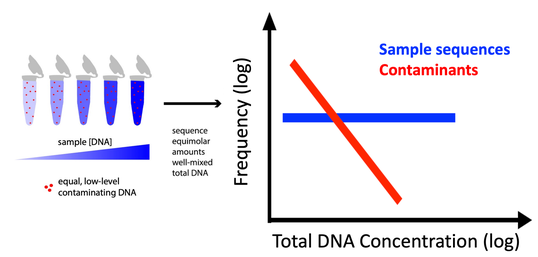


Figure 2: Mixture model of contaminants and true sample sequence in MGS experiments.

(Davis, et al. Microbiome, 2018)

Methodology of Hypothesis Testing

Step 1: Fit two models

Contaminant model (C)
$\log(freq) \sim (-1) * \log(conc)$

True Sample model (S)
$\log(freq) \sim (0) * \log(conc)$

Step 2: Compute R

$$R = \frac{RSS_c}{RSS_s}$$

***Decontam* Step3: classification procedure**

- Transform R to value P between 0 and 1 with adjusting sample size.
- Set a threshold P^* and ASVs with $P < P^*$ are classified as contaminants

***Tcontam* Step3: hypothesis testing procedure**

1. Define the null and alternative hypothesis:
   H_0: *True sample model fits better or equally better*
   H_a: *Contaminant model fits better*
2. Obtain the null distribution based on ratio of dependent Chi-square distribution
3. Compute p-values for each ASV based on null distribution giving R values
4. Perform FDR correction using q-value procedure. Reject null, i.e., ASVs classified as contaminants for q-values less than a threshold, e.g., 0.05.

Figure 3: Overview of hypothesis testing procedure for contaminant identification

Comparison between *Tcontam* and *Decontam*: Human Oral Microbiome Dataset

A

Decontam \ *Tcontam*	True Sample	Contaminant
True Sample	789	0
Contaminant	43	15

B

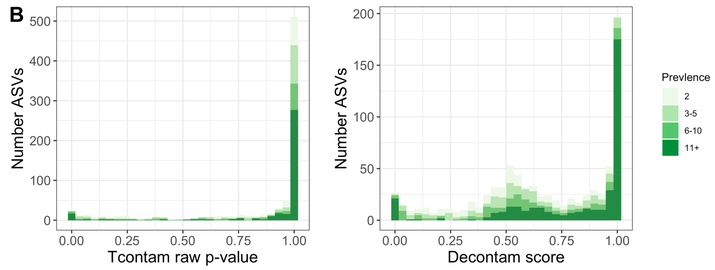
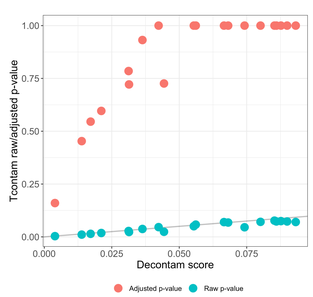


Figure 5: Difference of overall findings and scores/p-values between *Tcontam* and *Decontam* from an oral 16S rRNA gene dataset **A**: Contingency table of findings between *Decontam* and *Tcontam*; **B**: Histogram of *Tcontam* raw p-values (left panel) and *Decontam* scores (right panel) for each ASVs colored for different sample size.

A



B

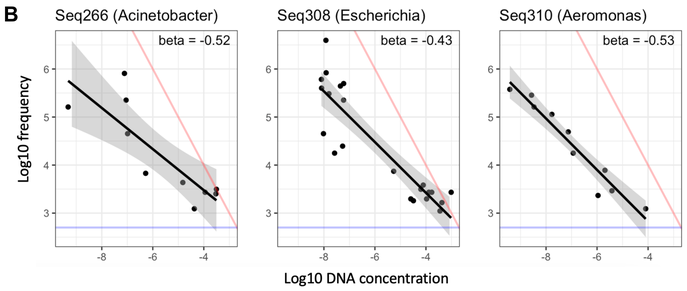


Figure 6:ASVs reported as contaminants only by *Decontam*. **A**: Scatter plot of *Tcontam* p-values vs. *Decontam* scores for ASVs with sample size < 5. **B**: Scatter plot of DNA concentration and frequency in log10 scale for 3 ASVs with sample size ≥ 10.

Validity of *Tcontam*

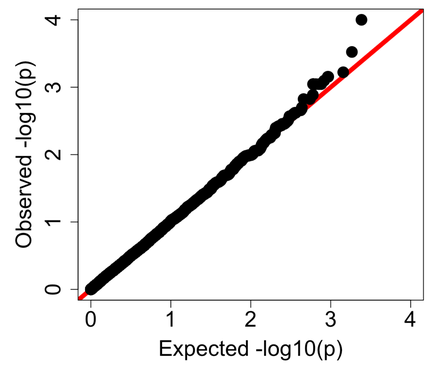


Figure 7: QQ plot of *Tcontam* p-value. It follows a uniform distribution under the null hypothesis.

Discussion and Conclusion

Tcontam is a hypothesis testing-based procedure, which assumes most of the DNA used to do marker gene or metagenomics sequencing are from true sample. *Tcontam* will call a ASV from true sample unless we have enough evidence from DNA concentration and frequency data. Compared with *Decontam*, which is based on a classification procedure, *Tcontam* reports a valid p-value, which can be (1) better interpreted with a significant level (2) connected to FDR correction procedure to control the overall FDR.

Specifically, *Tcontam* prefers not to call a ASV as a contaminant for the following two cases: (1) a strong correlation between DNA concentration and frequency with small sample size (e.g., < 5); (2) a weak to mild correlation with large sample size (e.g., > 10).

Next Steps

1. Conduct power analysis using simulation for different sample size.
2. Perform genus-level contaminant analysis using the oral dataset and validate findings use known contaminants or oral taxa reference database.
3. Conduct other real data analyses comparison between Tcontam and Decontam.

Acknowledgement

We thank Dr. Jung-Ying Tzeng and Dr. Craig Gin for their helpful discussions on this work.