Bioinformatics Protocol for Assessing Contamination Level and Quality on Genomics Data of *Ensifer meliloti*

Soohyun Lee,[1] Mario Cerón Romero,[2] and Katy Heath[2]

Parkland College & University of Illinois at Urbana-Champaign, Champaign, IL[1]

Department of Plant Biology, College of Integrative Biology, University of Illinois at Urbana-Champaign[2]

PRECS Phenotypic Plasticity Research Experience for Community College Students

Abstract

Nitrogen (N) fixing bacteria have a symbiotic relationship with host plants. The legume plants provide sugar, a product of its photosynthesis, to rhizobacteria. Rhizobacteria, one of the various N fixing bacteria, utilize the sugar for its energy source needed for conversion of N_2 into NH_4^+ [1][2]. The rhizobacteria would provide fixed nitrogen to legume plant for its growth in exchange of energy source[1][2]. There are various symbiotic relationships between microbes and plants, and the Heath Lab is especially interested in relationship between rhizobacteria and legume plants.

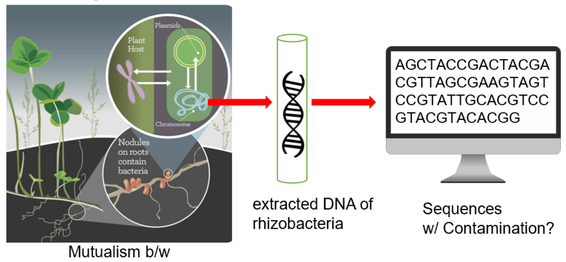
Various genomic methods to study rhizobacteria require sequence data in Heath Lab. However, there is high possibility of contamination in the sequence data, which may lead to false result in research [4]. Possible non-rhizobacteria that reside within the legume nodules, which don't participate in N fixation but in survival of the legume plant, could affect the research as well [3]. It was recently found that rhizobacteria other than *Ensifer meliloti* reside within the legume nodules [5]. Due to the existence of other rhizobacteria, we need a protocol to differentiate between these bacteria.

Here, we designed a protocol based on comparing the sequences of *E. meliloti* from the Heath Lab against public database to determine the level of contamination.

Project Description

- Creating a **protocol** to explore the **quality of genetic (sequence) data of 200 strains of *E. meliloti***, which can be applied to future work on other bacteria strains and for other type of projects in the Heath Lab
- Comparison of sequence similarity against public databases of Bacteria using BLAST, custom python, and R scripts → estimate the **level of contamination** in the Heath Lab databases
- Produce a 'database of contamination' in the Heath Lab as resource for future work on the same bacteria strains

Background



Mutualism b/w legume and rhizobacteria

extracted DNA of rhizobacteria

Sequences w/ Contamination?

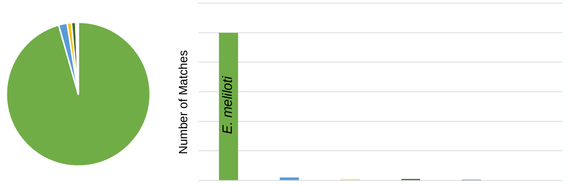
assess contamination of sequence by sequence comparison against public bacteria database through BLAST: Basic Local Alignment Search Tool

BLAST ®

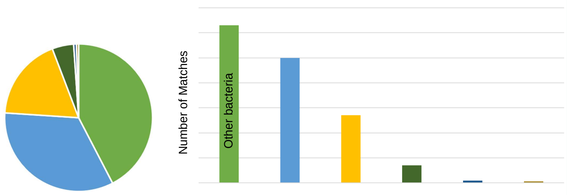
Sequences of *E. meliloti*

BLAST w/ human gene → Bad statistical values, No similarities found

BLAST w/ rhizobacteria → **Good statistical values**



Sample is **NOT contaminated**. Good to use.



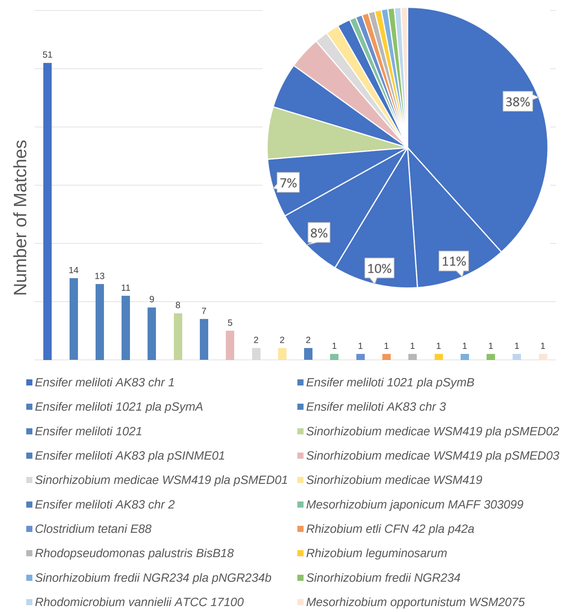
Sample is **LIKELY CONTAMINTED.**

Method

1. Collect sequence data of 200 strains of *E. meliloti* from Heath Lab, which were produced from studies of symbiosis in b/w *E. meliloti* and legume host
2. Transform FASTQ files (from sequencing facilities) into regular FASTA files (text-based sequences)
3. Compare the FASTA files of *E. meliloti* against public bacteria database (eg., refseq) through BLAST
4. Collect the results via Komodo Edit
5. Summarize the result in graphs and interpret how much contamination exist in the samples

Result

Exemplary Analysis of One of 200 Strains of *E. meliloti*



There could be possible minor contamination in this sample, but this sample looks good.

Conclusions

We have achieved a protocol to evaluate the levels of contamination in the sample of sequence data, which can be applied to any other projects.

This protocol will give us **"candidate contaminants"** for future decision of the sample.

Future Work

- Standardize the methods and create a manual for contamination detection for the Heath Lab data
- Implement the same method with other bacteria or different data

References

[1] Weese, D. J.; Heath, K. D.; Dentinger, B. T.; Lau, J. A. Long-Term Nitrogen Addition Causes the Evolution of Less-Cooperative Mutualists. *Evolution* **2015**, *69* (3), 631–642.

[2] Klinger, C. R.; Lau, J. A.; Heath, K. D. Ecological Genomics of Mutualism Decline in Nitrogen-Fixing Bacteria. *Proceedings of the Royal Society B: Biological Sciences* **2016**, *283* (1826), 20152563.

[3] Martínez-Hidalgo, P.; Hirsch, A. M. The Nodule Microbiome: N2-Fixing Rhizobia Do Not Live Alone. *Phytobiomes Journal* **2017**, *1* (2), 70–82.

[4] Goig, G. A.; Blanco, S.; Garcia-Basteiro, A. L.; Comas, I. Contaminant DNA in Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability. *BMC Biology* **2020**, *18* (1), 1–15.

[5] Brown, S. P.; Grillo, M. A.; Podowski, J. C.; Heath, K. D. Correction to: Soil Origin and Plant Genotype Structure Distinct Microbiome Compartments in the Model Legume Medicago Truncatula. *Microbiome* **2021**, *9* (1).

Acknowledgments

Financial support was provided by the National Science Foundation under grant #NSF REU 1950819/1950786, as part of the Phenotypic Plasticity Research Experience for Community College Students, through the University of Illinois at Urbana-Champaign Institute for Genomic Biology and Parkland College. http://precs.igb.illinois.edu/

Acknowledgement to Mario Cerón Romero, Katy Heath, Dr. Nathan Schroeder and Dr. C. Britt Carlson, and technical and support staff at the Institute for Genomic Biology.

PARKLAND COLLEGE

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

HEATH LAB FOR COEVOLUTIONARY GENOMICS