Exploring Arabidopsis thaliana Root Endophytes via Single-Cell Genomics

Scott Clingenbeel[1*], Derek Lundberg[2], Tanja Woyke[1], Susannah Tringe[1], Jeff Dangl[2]

[1] LBNL Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA - USA

[2] University of North Carolina at Chapel Hill, North Carolina - USA

**To whom correspondence should be addressed*: Email: SRClingenpeel@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was supported by NSF grant IOS-0958245.

Exploring *Arabidopsis thaliana* Root Endophytes via Single-Cell Genomics

Scott Clingenpeel[1], Derek Lundberg[2], Tanja Woyke[1], Susannah Tringe[1], Jeff Dangl[2]

[1]DOE Joint Genome Institute, [2]University of North Carolina at Chapel Hill

Introduction

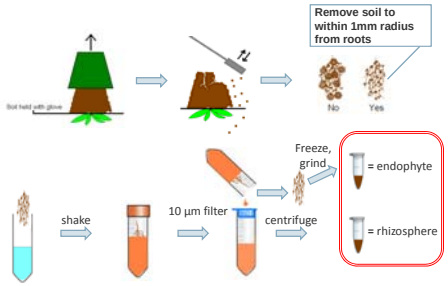
Land plants grow in association with microbial communities both on their surfaces and inside the plant (endophytes). The relationships between microbes and their host can vary from pathogenic to mutualistic. Colonization of the endophyte compartment occurs in the presence of a sophisticated plant immune system, implying finely tuned discrimination of pathogens from mutualists and commensals. Despite the importance of the microbiome to the plant, relatively little is known about the specific interactions between plants and microbes, especially in the case of endophytes.

The vast majority of microbes have not been grown in the lab, and thus one of the few ways of studying them is by examining their DNA. Although metagenomics is a powerful tool for examining microbial communities, its application to endophyte samples is technically difficult due to the presence of large amounts of host plant DNA in the sample. One method to address these difficulties is single-cell genomics where a single microbial cell is isolated from a sample, lysed, and its genome amplified by multiple displacement amplification (MDA) to produce enough DNA for genome sequencing. This produces a single-cell amplified genome (SAG). We have applied this technology to study the endophytic microbes in *Arabidopsis thaliana* roots.

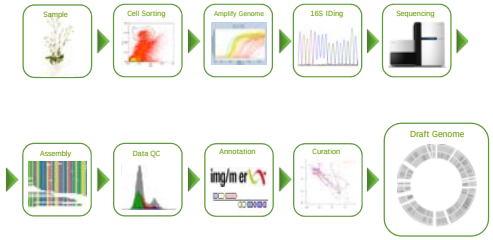
Extensive 16S gene profiling of the microbial communities in the roots of multiple inbred *A. thaliana* strains has identified 164 OTUs as being significantly enriched in all the root endophyte samples compared to their presence in bulk soil.

Methods

Obtaining Cells from Root Samples



Single-Cell Genomics Pipeline

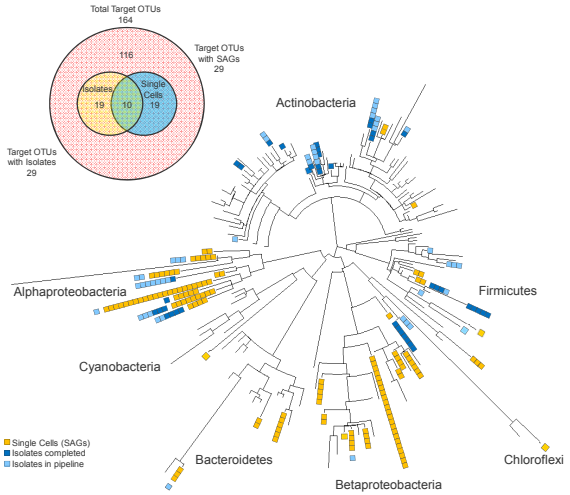


Single Cells and Isolates Obtained

~14,000 single cells were processed

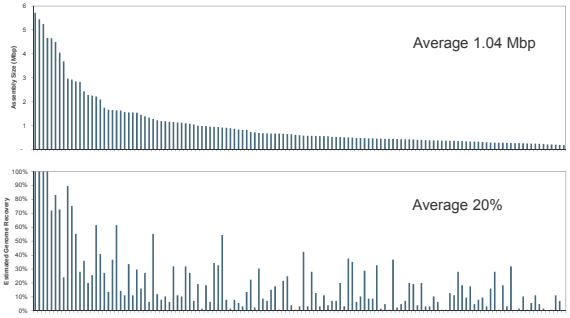
527 single cells have been obtained that match target OTUs

131 of these have been shotgun sequenced representing 29 OTUs



Maximum likelihood tree constructed from the consensus 16S pyrotag sequences for the OTUs that were significantly enriched in the endophyte community compared to bulk soil communities. Orange boxes indicate single cells that have been sequenced. Blue boxes indicate cultured isolates that have been selected for sequencing.

Single-Cell Genome Recovery



Total assembly size and estimated genome recovery from the sequenced single cells. Genome recovery was estimated based on the presence of 139 single copy genes in the final assemblies. The order of the single cells is the same in both figures.

Combined Assemblies – What to Combine

MDA amplification bias is random. Thus, combining the data from multiple single cells from the same OTU should produce a more complete genome.

What to combine

From same OTU (16S matches OTU ≥ 97%)

Average nucleotide identity (ANI) analysis

- Pairwise comparisons of genomes
- Query genome is broken into 1kbp fragments
- Fragments are blasted against comparison genome
- Hits are totaled to give the amount of overlap between genomes and the % nucleotide identity for the overlapping regions

Required ANI of ≥ 95% identity over ≥ 20kb overlap

You need to be cautious in what you decide to combine

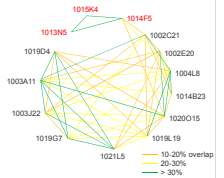
One Rhizobium combined assembly was 9.5Mbp

Estimated genome completeness for this was 86% implying the whole genome is 11.1Mbp

Largest Rhizobium genome in database is 8.7Mbp (Ave = 6.7Mbp)

Upon closer inspection, 3 of the SAGs used in this combined assembly are only weakly related to the other 11 SAGs used

When the combined assembly was rerun leaving out the 3 suspect SAGs a more reasonable assembly size of 6.1Mbp was obtained

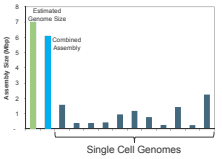


Pairwise overlap between SAGs in Rhizobium combined assembly. SAGs in red are weakly connected to the rest of the SAGs.

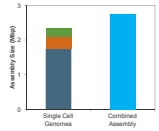
Combined Assemblies - Results

99 SAGs were combined into 22 combined assemblies (2-22 SAGs per combined assembly)

Average genome completeness in the component single-cell genomes was 18% and in the combined assemblies it is 45%



Combined assemblies produce more complete genomes



Going back to the original reads and doing a fresh assembly can improve the genome recovery

Conclusions and Future Work

- Our single-cell genomics pipeline has produced single-cell amplified genomes (SAGs) from a broad range of bacterial phyla that complements those obtained by culturing.
- The single-cell process produces incomplete genomes but combined assemblies of single-cell genomes from the same OTUs produce more complete genomes.
- Single-cell and isolate genomes will be compared to the genomes of close relatives from non-plant environments. Genes present only in the plant-associated microbes will hopefully illuminate the details of the plant-microbe interaction.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This work was supported by NSF grant IOS-0958245.