**Soil Microbiome Analysis Toolkit**

We want to generate a set of resources to improve the research of soil microbial communities. Ideally the resource will a curated set of genomes of soil organisms, tools for benchmarking, and a list of important yet unsequenced groups and poorly sampled environments.

Genome database

-Initially only closed genomes will be included (from Genbank's Refseq). In a second phase, draft genomes from Genbank and other repositories, as well as un published genomes will be included.

An effort to include "binned" genomes will be made, this will be a novel yet important approach in order to include the whole microbial diversity. (Second phase)

-Draft genomes will be evaluated to look for contaminants, and critical errors such as the absence of essential genes in them (Second phase).

-Genomes will be reannotated to have an uniform criteria using RAST

-Fungal genomes will be retrieved from genbank (First phase) and the 1000 fungal genomes project (second phase)

-Archaeal genomes be also included (First phase)

-Genomes from small eukaryotes will be included in the second phase

Ecology of soil organisms

Once the database is stablished the distribution of functional genes, pathways, metabolites will be studied (First phase)

Metadata and curation criteria

-Metadata for all candidate genomes will be reviewed to consider them into the database,

-Source of the organism will be a key criteria for inclusion.

-The location of the initial sample as well as publised reports will be used to associate the organism with specific soil ecosystems (e.g. praire, forest, bog, etc)

-Genomes will be assigned into different tiers of confidence according to the quality of the genome, the research published for the organism, and the genome metadata.

-Annotation can be made distributed using github.

-The extra annotations will be pushed into Genbank as third-party annotation (Second phase)

- Some people have shown interest in including data such as nutrient utilization and growth rates when available for the organisms. Some of the data is already available in publication when the organisms are officially named (International Journal of Systematic Bacteriology). (Second phase)

Benchmarkig tools

The distribution of organism in soil will be used to generate "virtual soil communities" that can be used for testing bioinformatic approaches. Ideally several communities representing different soil ecosystems will be created.

In a second phase, DNA for these mock communities will be created using DNA from pure cultures. These DNA mixtures will be made available for researchers.

Undersampled ecosystems, unrepresented organisms

-Based on the distribution of organism in different soil ecosystems a list of dominant organisms per ecosystem will be created and the presence of a reference genome will be evaluated.

With this information a list of "most wanted" organisms per ecosystem will be created to encourage sequencing of them.

-In a similar manner, the ecosystems will be ranked by representation in the database to establised underrepresented ecosystems in the database

-The EMP database will be used to stablished the distribution of organisms (OTUs)

Plan of action

Stablish an initial database and state where are the holes and the needs of the community

- Soil genomes (Bacteria, Fungi, Archaea)

- Core genes, functions, pathways, metabolites (PRMT), and transporters (PRMT) of soil bacteria only

- Most wanted organisms, undersampled ecosystems (based on EMP)

- Mock communities?

- A call for the community to provide these genomes for sequencing will be made

White paper?

-Make parallels to HMP to show how databases have improve the quality of research and made the analysis easier, more consistent, and showed the holes in the knowledge base

- Patrick mentioned the difference in mapping at the beginning and nowadays as an indicator or progress

- Use of mock databases to test assemblers

- Need for a large-scale sequencing effort to improve understanding of the soil ecosystem

Parallels with NIH for the HMP, and the support of the Moore foundation for the study of marine ecosystems.