Microbial Profiling: Inferring identity from human microbiome samples

- What is your microbiome? Your microbiome is the bacteria that live on or in you.
- Recent research has shown that your microbial profile unique for different body sites, and in particular the microbial communities associated with your skin are unique and relatively stable throughout your lifetime.
- In 2011, Fierer et. al. was able to identify from bacterial residue left behind on single computer keys and mice which is easily collected that it was possible to identify who had previously handled the mice and computer keys.
- Currently Microbial Data is unprotected unlike DNA and while the metadata associated with research studies is protected the actual sequence data from samples taken from research subjects is not.
- In 2014, Lax et. al. have recently shown that the microbial profile of your home reflects your families microbial profile. After moving to a different home the new location rapidly changes to resemble you and your families microbial profiles. They were also able to determine a person’s individual contribution to the microbial profile of the surfaces of a home which degrades over time when the person leaves the home as the person is no longer there to contribute to the microbial profile of the home. This suggests that your microbial profile could be used to determine when you have left a specific location.
Motivation

• Recently Curtis Huttenhower's Lab at the Department of Biostatistics at Harvard submitted a paper for review describing an algorithm that could based on your microbial profile identify you with between 50 to 80 percent accuracy.

• Currently Microbial Data is not protected like DNA data is.

• Several Major ongoing research efforts to discover the effects of our microbial communities are currently underway including the human microbiome project which was established in 2008.

• Currently Microbial Sequencing Data is publically released and available for download with no restrictions on access. Although the subjects are deanonymized and referred to using a sample id.
Goals

- Research into the uniqueness of the human microbiome suggests that it may be possible to reidentify subjects from research studies based solely on their microbial profile i.e. collected samples generally stored as either fasta or fastq files.

- Determine given an attacker has a digital version of the microbial genetic data of an sample of an individuals microbial data or has digital versions of microbial genetic data for multiple samples from an individual whether they can determine if this individual participated in a given research study using commonly available analysis open source software packages.
Methodology

- Used publically available dataset from the study done for lax et. al. which has multiple samples for each person available stored at the European Bioinformatics Institute website in their biosamples database. For each individual samples were taken for 3 different body sites foot, hand, and nose.

- In order to get preliminary results 6 individuals were randomly selected from the 18 possible individuals. Then for each individual 10 samples for each of the body sites foot, hand, and nose were obtained.

- The Qiime Software package was used to analyze the data for the 6 individuals.

- All of the samples for all 6 individuals were subsampled so that each sample had 12000 sequences.
Methodology

• Then all the data for all the samples was pooled together for analysis.
• O T U (operational taxonomic unit) picking was used to determine the specific organisms present in the samples using usearch algorithm which searches against a database for 97% or greater matches. Clusters with 10 or less sequences were thrown out.
• Then for each cluster or O T U representative sequences were selected and assigned to taxonomies using the GreenGenes reference database and this information was then used to generate both a phylogenetic tree (a tree that shows evolutionary relationships between organisms) and a O T U table with relative abundances for each O T U per sample.
Methodology

• Finally pairwise distance matrices were created with o t u tables rarefied to 2500 sequences per a sample using weighted unifrac which calculates the distance between two different samples based on the combined phylogenetic tree. Each distance represents the amount of evolutionary history that is unique to either of the two samples this is measured as a fraction of the branch length in the phylogenetic tree.
Results

- Preliminary Results were obtained only for the body site hand for 5 people.

- The Average difference within samples from the same person was 33%.

- The Average difference between samples from different individuals was 30%.
Conclusions

• Preliminary results suggest that using the methodology above it may not be possible to reidentify individuals based solely on their microbial profile for the skin of their palm. However, it should be noted that the dataset used here was a small subset and that only o t u’s with greater than 10 sequences were considered.

• Future work:
  – Determining if having more samples increases the ability to see the differences between individuals, looking at using different algorithms such as de novo picking, considering rare o t u’s i.e. those with only 1 or 2 sequences, other body sites such as foot and nose may be more identifying, removing the core microbiome from the sample
References

- **QIIME allows analysis of high-throughput community sequencing data**


Recent research into the human microbiome or the bacteria that lives in or on us has shown that your microbial profile is unique to you. Recent research has shown that your microbial profile unique for different body sites, and in particular the microbial communities associated with your skin are unique and relatively stable throughout your lifetime. In 2011, Fierer et. al was able to identify from bacterial residue left behind on single computer keys and mice which is easily collected that it was possible to identify who had previously handled the mice and computer keys. Currently Microbial Data is unprotected unlike DNA or genomic data, and while the metadata associated with research studies is protected the actual sequence data from samples taken from research subjects is not. Moreover unlike DNA it can give information about your location or movement patterns.

In 2014, Lax et. al. have recently shown that the microbial profile of your home reflects your families microbial profile. After moving to a different home the new location rapidly changes to resemble you and your families microbial profiles. They were also able to determine a person’s individual contribution to the microbial profile of the surfaces of a home which degrades over time when the person leaves the home as the person is no longer there to contribute to the microbial profile of the home. This suggests that your microbial profile could be used to determine when you have left a specific location.

Recently Curtis Huttenhower’s Lab at the Department of Biostatistics at Harvard submitted a paper for review describing an algorithm that could based on your microbial profile identify you with between 50 to 80 percent accuracy.

Methods

Used publically available dataset from the study done for lax et. al. which has multiple samples for each person available stored at the European Bioinformatics Institute website in their biosamples database. For each individual samples were taken for 3 different body sites foot, hand, and nose. In order to get preliminary results 6 individuals were randomly selected from the 18 possible individuals. Then for each individual 10 samples for each of the body sites foot, hand, and nose were obtained. The Qiime Software package was used to analyze the data for the 6 individuals. All of the samples for all 6 individuals were subsampled so that each sample had 12000 sequences. Then all the data for all the samples was pooled together for analysis.

O T U (operational taxonomic unit) picking was used to determine the specific organisms present in the samples using usearch algorithm which searches against a database for 97% or greater matches. Clusters with 10 or less sequences were thrown out. Then for each cluster or O T U representative sequences were selected and assigned to taxonomies using the GreenGenes reference database and this information was then used to generate both a phylogenetic tree (a tree that shows evolutionary relationships between organisms) and a O T U table with relative abundances for each O T U per sample. Finally pairwise distance matrices were created with O T U tables rarefied to 2500 sequences per a sample using weighted unifrac which calculates the distance between two different samples based on the combined phylogenetic tree. Each distance represents the amount of evolutionary history that is unique to either of the two samples this is measured as a fraction of the branch length in the phylogenetic tree.

Results

Preliminary Results were obtained only for the body site hand for 5 people

The Average difference within samples from the same person was 33%
The Average difference between samples from different individuals was 30%