03-713: Project: The Evolution of Bacteria
Spring 2013

PROJECT SUMMARY

The bacteria that live on us and in us, and that sometimes make us sick, are constantly evolving. Strains can acquire mutations that increase or decrease their virulence, or make them resistant to antibiotics, which makes a strain very difficult to treat. These antibiotic resistant bacteria often spread through hospitals and around the world. By sequencing genomes of many strains, we can start to understand how particular mutations arise, how the strains that carry them propagate, and how genomic differences affect virulence. The goal of this project is to produce an automated pipeline to identify important differences between strains in collections of strains whose entire genomes have been sequenced.

Bacterial infections have long been studied under the paradigm that acute diseases are caused by single clones. Over the past decade, it has become clear that bacteria also play a critical role in chronic disease and human health. In this case, they are usually associated with the presence of multiple strains or species organized into complex structured communities, termed biofilms, which can persist in the body and are highly recalcitrant to antibiotic treatment.

Strains from the same species can encode very different set of genes. These differences can translate into diverse biological properties such as variation in vaccine susceptibility, antibiotic resistance, and pathogenecity.

Furthermore, bacteria can incorporate genes from neighboring cells into their genomes, through a process known as horizontal gene transfer. The widespread transfer of genes between diverse strains leads to the continuous generations of strains containing new combinations of genes.

In some species, where the strain diversity and genomic plasticity are high, novel strains are generated in the course of single infections, hospital outbreaks, or pandemics. For the human opportunistic pathogen *S. pneumoniae* there is evidence of all of the above. Strains isolated from a single infection have been shown to sequentially acquire genes leading to the change of 10% of the genome during one infection. Strains isolated from a multiple patients from a single hospital during a one-year time period have been shown to swap genes, generating variable strains with differences in genes implicated in adhesion, colonization, and resistance. Similarly, related pandemic strains, isolated in the post-antibiotic era, have been shown to acquire multiple virulence factors and antibiotic resistance genes over a 40 year period.

By performing comparative genomics across multiple strains, we can identify both the point mutations accumulated over time by vertical descent, as well as the regions acquired/lost by horizontal gene transfer. Correlations between genomic content, epidemiological data and phenotype in animal models provide functional information, especially regarding virulence determinants. Finally, analysis of the horizontally transferred regions informs on the contribution of homologous recombination, phage, or mobile elements as mechanisms for gene transfer.

The goal of this project is to produce an automated pipeline to identify genomic differences across *S. pneumoniae* strains using a collection of strains whose entire genomes have been sequenced.

Relevant Nomenclature:

1. Core Genome: region of the genome shared by all strains. For *S. pneumoniae*, the core genome corresponds to 70-80% of the genome of any single strain.

2. Distributed Genome: regions that are not shared by all strains (non-core).

3. Allelic Differences: base pair differences between orthologous.

4. Differences in Gene Content: genes that are vary in their distribution across strains.

DATA

You will be given:

* A collection of partially assembled, whole-genome sequences of 10 strains.
* The sequencing reads used to create those assemblies.

You can use:

* Any other publicly available data or databases.

GOAL

Your pipeline should perform the following analyses:

1. Identify regions that differ across the strains.
   These differences will fall into two patterns: (1) isolated base pairs which likely represent point mutations that where acquired during vertical descent, (2) differences clustered on the chromosome (may include allelic and/or genic variations) which correspond to horizontal gene transfer events.

2. Build a phylogenetic tree using the core whole genome sequence, and excluding any regions that have undergone horizontal gene transfer.
   This will provide the evolutionary relationship among strains.

3. Build phylogenetic trees for each one of the regions that have undergone horizontal gene transfer.
   This will provide the evolutionary relationship among the horizontally transferred regions, grouping together strains that may not be related.

4. Annotate the genes within the horizontally transferred regions.
   Provide a clue as to how the transfer events may influence strain phenotype.

Your software should produce a report (formatted as text, html, RTF, or as an interactive display) that describes the output of the above analyses.

Although it will be tested first on the provided data, your pipeline should work for any similar data for other species of bacteria.