amd MIDWEST

November 2019

Upcoming Events

- **APHL ID Con Submissions** - Call for Abstracts Available Late November https://www.aphl.org/conferences/ID-Lab-Con-2020/Pages/default.aspx
- **Inform 2020** - March 2020
- **APHL Annual Meeting** - June 8 - 11, 2020

Resources

- **The AMD-Midwest Regional Website:** https://staph-b.github.io/midwest-region/
- **The StaPH-B Website:** https://staph-b.github.io/

Bioinformatics Demystified

This month and next we're covering different methods of genome comparison. First we'll introduce the infamous single nucleotide polymorphism analysis. We will also talk about the many variants of multi-locus sequence typing!! Buckle up for a wild ride!

All the single... nucleotides?

Single nucleotide polymorphisms (SNPs) are differences in a single nucleotide at a specific position in a DNA sequence. SNPs are identified as differences from a reference sequence or as differences from other isolates.

Reference	**A T T G T C T A T G C G T A G T C T A G T C**
Isolate 1	A T T G T C T A A G C G T A G T C T A G T C
Isolate 2	A T T G T C T A A G C G T A C T C T A G T C
Isolate 3	A T T G T C T A A G C G T A G T C T A G T C
Isolate 4	A T T G T C T A G G C G T A T T C T A G T C
Isolate 5	A T T G T C T A T G C G T A C T C T A G T C

	SNP I	SNP II
SNP Combination I	A	C
SNP Combination II	G	T

In the figure a SNP occurs at position nine and fifteen in the DNA sequence. Isolate 1 has a 'A' nucleotide, while the reference has a 'T' nucleotide. SNP combinations and total number of SNPs can be used to identify relatedness between isolates.

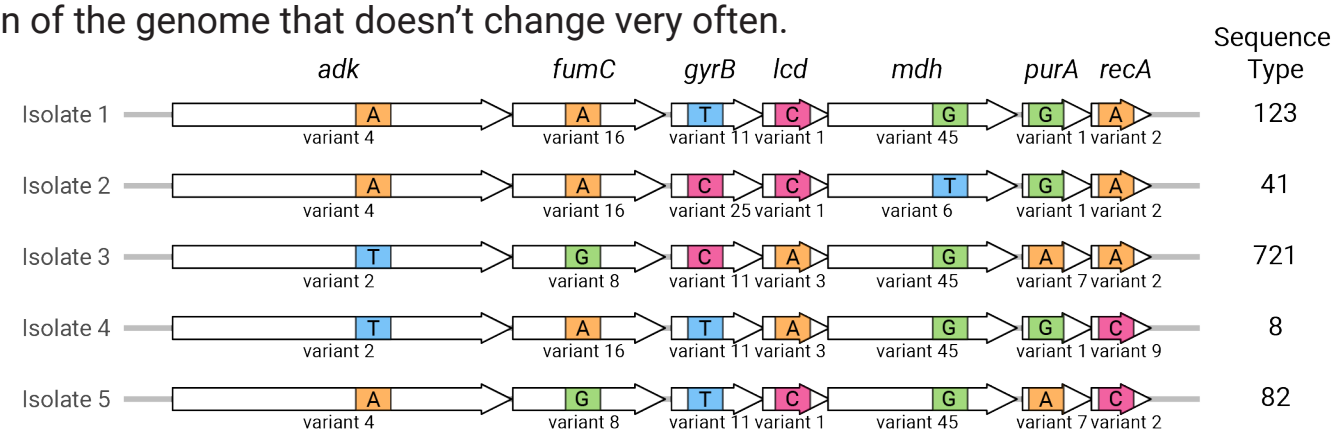
SNPs accumulate over time, which allows us to assess relatedness between individuals. In general, the more SNP differences two members of a species have, the less likely it is that they are related. Conversely, if two members of a species have very few SNP differences, the more likely it is that they are related.

Not all SNPs are equal, some may not even be included!

Our ability to accurately identify SNPs is limited by the sequencing technology we use and how we analyze our sequencing data. Sequencing technologies have associated error rates, which can introduce false SNPs into our sequence data. Additionally, bioinformatic workflows differ in the methods or quality metrics they use to confidently identify SNPs. While many organizations have defined best practices for identifying SNPs, there is no current consensus on how to confidently identify SNPs in sequencing data. The SNPs identified between isolates can vary greatly from workflow to workflow, so SNP results should always be evaluated with additional analyses.

It's all about the loci, but how many is multi?

Originally multi-locus sequence typing (MLST) was devised as a way to characterize DNA sequences. Before next-generation sequencing (NGS) became available, a method was developed to characterize and identify isolates by PCR amplifying and sequencing a small portion of 7 - 8 genes found within every organism of that genus/species. These genes are well conserved (i.e., they accumulate very few mutations over time), so it is possible to identify and type sequences using combinations of mutations. MLST works well for tracking organisms on a global level, but it is not very useful for outbreak investigations, because it uses a very small portion of the genome that doesn't change very often.



In the above example we have 5 isolates depicting SNPs occurring in a 7 gene MLST. Each SNP variant corresponds to a different gene variant. The combination of gene variants is used to determine the MLST sequence type.

Thanks to NGS, we are no longer limited to sequencing one gene at a time. We can sequence multiple genes, and even multiple genomes, at once! So what happens if we expand MLST to include more than just our original 7 - 8 well conserved genes?

Find out next month!!