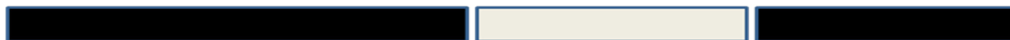
Understanding the structure of bacterial species by MLST

Introduction

Many different methods for discriminating variation among bacteria have been developed, each somewhat distinct in its sensitivity, reproducibility, cost, and broad utility. In the first lecture we learned of methods based on **R**andomly **A**mplified **P**olymorphic **D**NA, or RAPD, which target short sections of DNA that are repeated and distributed somewhat randomly throughout bacterial genomes. By designing a PCR primer that complements these stretches of repetitive DNA, we can use the products of this PCR to detect relatively subtle variation among different bacterial clones.

However, typing methods like RAPD (or pulsed-field gel electrophoresis, PFGE, which just analyzes the arbitrary fragments of genomes cut by site-specific restriction enzymes) do not reveal *where* genome differences exist or *whether they accumulated by mutation or by recombination*. These last two points are critical: if differences accumulate only by mutation, then the two samples are related by shared descent, and they can be traced backwards to the same clonal lineage. However, if differences accumulated by recombination, then samples may come to resemble one another by sharing similar DNA that originated in a different lineage (Fig **1A**), or they may even appear quite different (**1B**) because of unique recombination events, despite being mostly the same.

Fig 1A: 

Typing here ↑ makes them appear similar

Fig 1B:

Typing here ↑ makes them appear different

Why bother?

Because understanding how bacteria are related to one another can often be a really big deal, especially when tracing an ongoing outbreak, performing retrospective epidemiology (ie "Are all of these infections in our hospital related, and if so, how did the patients become infected?"), or searching for the source of a novel, useful bacterial function, like nitrogen fixation or chemical decontamination.

The purposes of this laboratory are
1) to become familiar with a method that has become the gold standard for such studies, **M**ulti-**L**ocus **S**equence **T**yping, or **MLST**
2) to use a few methods of analyzing these data to understand how different types of bacterial species are “structured,” or related to one another.

Exercises:

1) To begin to learn a bit about MLST, please visit here: http://www.mlst.net/misc/further.asp

2) Now let’s begin the analysis by going to the main page, http://www.mlst.net/, and then clicking on “Databases.” Now click on “S.aureus,” which is the database for all isolates typed and deposited to the *Staphylococcus aureus* database. You should be at: http://saureus.mlst.net/.

3) Click on “Organism specific information,” where you should read about the methods and protocols used for *S. aureus* MLST.

    a. List the seven genes that were amplified and sequenced, along with their gene products.

        i. gene: product:

        ii. g: p:

        iii. g: p:

        iv. g: p:

        v. g: p:

        vi. g: p:

        vii. g: p:

    b. What kinds of genes are they? What do they do in the cell?

    c. What are the intermediate goals of the DNA extraction protocol?

    d. How long, on average, are the MLST PCR primers? How do these compare with primers used for RAPD, and why are they different? (Google “Box PCR primer” or “RAPD primer” for comparison)

4) Now go back to the main *S. aureus* MLST page and click on “MLST maps,” then Google Maps (the link is: http://saureus.mlst.net/earth/). This links each of the isolates in the database to their global point of isolation.

   a. How many isolates are in the database from the UK? __________ From the US? From Brazil? ___________ From Russia?

   b. What does this say about the incidence of Staph infections globally?

5) Now, click on the USA link and then select “Click to view STs.” ST’s are Sequence Types. Scan down the list of ST’s, noting how frequently isolates of each sequence type have been typed.

   a. Describe the frequency distribution of these ST’s.

   b. Now click on the links to the right, roughly below “other countries (number of isolates elsewhere with the same ST). What do these findings tell you about S. aureus global epidemiology?

   c. What ST’s do you find most interesting?

6) Now do the same as above (#5) for Norway, by first clicking on the right to return to the global isolates map, then on Norway, which has a large number of isolates deposited. What do these data tell you? How do they compare with those from US?

7) Go back to the US database and click “eBURST instructions click here.” Please read the first 3 pages of the eBURST tutorial, through “Identifying the founder of a group and the patterns of evolutionary descent.”

   a. What defines a sequence type?

   b. What defines a relationship between sequence types?

   c. What’s an SLV?

8) Click: “Click to view epidemiological data from the USA,” for your own information.

9) Now select “**click to view eBURST to compare isolates from USA with global population**.” This will launch Java, click OK, and then OK if you encounter a warning. -Under the **Analysis** tab, and in the Analysis box, change the numbers so that they read “7 **0** 3 1000” instead of the default “7 **6** 3 1000.” These numbers refer to: a) the number of loci analyzed, b) minimum number of identical loci for each group, c) minimum number of SLV’s (single locus variants) to define subgroups and d) number of re-samplings for bootstrap (a statistical method). By changing to 0 you gain a complete snapshot of the population. Click Compute.

   -Now click on the **Diagram** tab, and under the Diagram menu label, remove the tick for “Show ST labels.” Now click **Draw**. Read this figure given: **Black** - STs found only in the Reference dataset, Green - STs found only in the Query (USA) dataset, Cyan - STs found in both the Reference and Query datasets.

   What does it all mean? Feel free to perform the same analysis with another country compared to the total database to aid your interpretation. Use the back if necessary.