VCU Bioinformatics and Bioengineering Summer Institute

Closing the gaps in the *Streptococcus sanguis* genome: Problem Set

1. Ribosomal RNA genes often give sequencers headaches, since they frequently appear in multiple copies in bacterial genomes. Long repeated sequences are the bane of sequencing projects, since it is difficult to read past them. Given the published sequence of S. sanguis 16S rRNA (Genbank accession #AF003928), determine whether our file of contig ends contains any partial 16S rRNA sequences. If so, is the match/are the matches perfect?

2. Determine which ends of *Streptococcus sanguis* are good candidates for joining by PCR. Use the same *Streptococcus* genome and Blast program as before, i.e.:

	BlastN	**TBlastX**
S. pneumoniae	Peter, Mark, Josh	Tom, Danielle, David
S. mutans	Emily, Greg, Rebecca, Lyndsey	Gaurav, Robert, Luke, Chris

3.* What are the relative merits of using *S. pneumoniae* vs *S. mutans.* Forget theory. This time around, look at the actual results. Did the two genomes predict the same contig pairings? If not, did their predictions overlap or instead did one set encompass the other?

4.* What are the relative merits of using BlastN vs TBlastX. Again, look at the actual results. What were the benefits and disadvantages of each approach?

5. (*All those not comfortable writing programs*) Write a program that takes as input a Genbank file and outputs the embedded sequence in FastA format. This is a program you might use time and again. Of course you should not begin by staring at a blank page. Instead, write by subtraction: take a parsing program that works and whittle away at it.

**** To answer Questions 3 and 4, you will need to consult with colleagues that used the strain or program you didn't use.***