Alvin Tin ‘14
Computational Biology Sc. B

Validity of Single Nucleotide Polymorphism (SNP) data on *Semibalanus balanoides* genome

Principle Investigator: Professor David Rand

Finding restriction enzyme recognition sites on a given DNA contig is a common first step in the process of understanding genetic variation. Restriction enzymes are useful for surveying specific sequences in a genome because the DNA digestion patterns they generate can provide information about the underlying DNA sequence, albeit a small fraction of the total genomic sequence. Patrick Flight performed a population genomic sequencing survey of barnacle genomes from three different locations: Narragansett Bay, Rhode Island, Damariscotta, Maine, and Southwold, UK. Patrick aligned ~1kb contigs to a reference genome and found reoccurring single nucleotide differences between the two DNA fragments; this difference of a single nucleotide base is known as a Single Nucleotide Polymorphism (SNP). We took the first of many next steps to understanding the observed genetic variation. We made a python program that takes in a list of contigs, a restriction enzyme string, and the SNP data. It first finds all the restriction enzyme site locations for each contig, then find which of those sites are affected by the SNP data provided. The program seems to find an expected number of clashes between restriction enzyme and SNP. We found ~7,100 HindIII restriction enzyme (AAGCTT) sites among the contigs, and ~337 of those found sites were affected by a SNP from the data. The next step is to identify primers around those destroyed found sites for PCR and electrophoresis analysis. Observation of bands on the gel will allow us to verify the presence and ratios of the SNP. If the data is consistent, we will be able to confidently observe that there are clear genetic variations among the three populations. In which case, a good subsequent step is to identify which SNPs have noticeable difference in frequency when compared between the populations. Those are indicators for us to tackle the question of why those genetic variations occur.