Name: Vince Buonaccorsi

Institution: Juniata College

Project Title: Transcriptome scan for positive selection on coding regions between *S. carnatus* and *S. chrysomelas*.

Summary: Detailed examination of adaptation and reproductive isolation at the molecular level is very challenging in most species, yet to achieve a full understanding of these processes we must determine the genomic underpinnings of population divergence. Continued development of new model systems is essential to test generality of early findings from model species. The genus *Sebastes* represent a rare example of a marine vertebrate species flock and presents a natural evolutionary laboratory within which one may pursue hypotheses with the rare advantage of phylogenetic replication. *S. carnatus* and *S. chrysomelas* are a particularly intriguing and important sympatrically distributed species pair because they appear to be divergent only at a small minority of loci, loci that are likely linked to genes that produced and/or maintain species boundaries (Buonaccorsi et al. 2011). In order to begin to characterize the genes and genomic processes that produced and maintain differences between these species, we propose to sequence the transcriptome of *S. carnatus* and *S. chrysomelas* from a single individual each, identify genes experiencing high non-synonymous to synonymous substitution rates and thereby identify candidate gene sets involved in their divergence.

The process of speciation in the presence of gene flow requires strong coordinated divergent selection on several pre-zygotic traits, where post-zygotic barriers have not yet taken hold (Coyne and Orr 2004; Via 2009). Speciation in allopatry may accumulate gradually (or quickly if a new environment is colonized) and may involve a relatively random selection of pre- or post-zygotic isolating mechanisms. Speciation with gene flow may be more common than previously appreciated since new mechanisms for the process have been proposed (Via 2009). For *S. carnatus* and *S. chrysomelas* speciation with geneflow appears to be occurring. The species have nearly completely overlapping species distributions, niche partitioning by depth preference, only slight morphological divergence, and introgression is still occurring to a limited degree (Buonaccorsi et al. 2011).

After a candidate set of genes with accelerated amino acid divergence has been identified between *S. carnatus* and *S. chrysomelas*, we hypothesize that genomic divergence follows patterns expected under speciation with gene flow, characterized by overrepresentation of genes involved in pre-zygotic isolation mechanisms such as mate recognition or habitat specialization (Via 2009). Follow-up studies can be performed on divergent genes to determine also whether divergence was simultaneous and dated to species origin (Buonaccorsi et al. 2011). A list of the divergent genes, their timing of divergence, and their levels of introgression should paint an interesting picture of the divergence genetics for this incipient species pair.

Quantity of sequence data needed: The genome size of *Sebastes* is estimated to be about 1Gbp, and transcriptomes are typically 1.5% of the genome size. A quarter run of the 454 will provide 250,000 reads of 400bp average each (100Mbp), which will provide 3X coverage of
each transcriptome and allow for partial assembly de novo from these two closely related species in the same subgenus, and allow us to distinguish between read error, heterozygotes, and homozygotes.

**Source of the sample, nucleic acid preparation, and timeframe:** We propose to sample liver, kidney, gonad, and brain tissue from one adult of each species for RNA isolation (Qiagen RNeasy). *S. carnatus* samples have already been obtained. *S. chrysomelas* samples should be available by the beginning of September with the aid of colleagues or small contracts to commercial fisherman in the Monterey Bay area. PolyA enrichment kits (LifeTech) will be purchased and samples sent to PSU to confirm enrichment success on their Agilent Bioanalyzer. Equal masses of RNA samples will be pooled across tissue types, cDNA will be made using the QuantiTect reverse transcriptase (Qiagen), and shipped to PSU nucleic acid facility for library preparation with barcoding, emPCR, and sequencing.

**Analysis plan:** Short reads will be assembled using the Newbler assembler or Geneious Pro 5.4.3. Protocols for protein ID, open reading frames, and calculation of non-synonymous (Ka) to synonymous (Ks) substitution ratios follow Heras et al. (2011). Briefly, gene pairs will be annotated using BLAST2GO (Conesa et al. 2005) to the Swissprot database using BlastX. To identify open reading frames EST sequences will be BlastX’d against Stickleback and Fugu genomes. Resulting files will be processed through ORF-PREDICTOR and TargetIdentifier (Min et al. 2005a,b) to identify ORFs and whether entire coding regions were covered. Protein alignments will be converted into corresponding codon alignments using Pal2Nal 2.2 (Suyama et al. 2006). Nonsynonymous to synonymous substitution rates will be calculated using KaKs_Calculator (Zhang et al. 2007), filtered for possible paralogs, and tested for enrichment of GO categories using a Fisher’s Exact Test using GOSSIP (Bluthgen et al. 2005). Ka/Ks ratios of > 1.0 are consistent with positive selection.

**Role for undergraduate students.** If funded I will plan to invite up to 10 sophomores this Fall into my research program with the goal of learning the appropriate analyses for this data using online tutorials and primary literature, and gaining experience with the actual analysis. This work and the followup work involved will also be used as the foundation for larger projects in my new 4 credit Genetic Research Methods class in the Spring semester whose goal is to introduce students to modern molecular research in the context of performing modern research projects. The results of this project will serve as basis for a publication that includes the research students that become involved.

**References**


