Analysis Exercises

MOLECULAR SIGNATURES AT A MAJOR LOCUS UNDERLYING A KNOWN ADAPTIVE TRAIT
INDELS & SWEEPS in WHOLE GENOME DATA

Background

Key Paper:
Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, et al, Myers RM, Petrov D, Jonsson B, Schluter D, Bell MA & Kingsley DM. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327 (5963): 302-305

Other Papers of Relevance:
Jones FC, Grabherr MG, Chan YF, Russell P, et al Lander ES, DiPalma F, Lindblad-Toh K, and Kingsley DM. (2012)
The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55-61.

Whole genome sequencing provides an unprecedented opportunity to study and characterise the molecular events underlying adaptation, divergence and speciation.

In today's exercise you are going to be working with an unpublished population genomic dataset *(please don't use outside this course or share with colleagues)* of benthic and limnetic stickleback species-pairs. You will study the mutations and molecular signatures of selection around a major locus previously identified to be explaining 70% of variation in the size of the pelvic (Shapiro et al 2004 Nature, Chan et al 2010 Science).

What do the patterns of molecular variation look like around this locus?
What can we learn from these patterns and how applicable are they to other regions of the genome?

BACKGROUND INFORMATION
Phenotype
Limnetic and benthic ecotypes can be found in sympatry within 4 lakes in British Columbia. They are thought to have evolved by a "double invasion" of the lakes (invasion in two waves approximately 2000-4000 years apart) and despite secondary contact are reproductively isolated. They differ considerably in numerous traits (eg habitat preference, behaviour, mate preference, morphology). In two of the four lakes benthic ecotypes have lost their pelvic spines and girdle ("pelvic reduced") while their coexisting limnetic ecotypes have complete pelves. In the other two lakes all fish have complete pelves.

	Benthic	**Limnetic**
Lake1	Pelvic reduced	Pelvic complete
Lake2	Pelvic complete	Pelvic complete
Lake3	Pelvic reduced	Pelvic complete
Lake4	Pelvic complete	Pelvic complete

Genotype
Pelvic reduction via deletion of a cis-regulatory element of Pitx1 gene (Chan et al 2010)
Forward genetic mapping previously identified the Pitx1 locus and surrounding genomic region as the major QTL explaining more than 70% variation in the pelvic

apparatus in a laboratory intercross between a Lake1 benthic fish and a marine outgroup (Shapiro et al 2004).
Subsequent allele-specific expression, fine mapping, association mapping and transgenic enhancer and rescue assays identified a 488bp deletion ~32kb upstream of the Pitx1 locus that was responsible for the loss of the pelvis in benthic fish. Transgenic integration of an undeleted marine allele upstream of Pitx1 coding sequence was capable of fully rescuing the pelvic apparatus on a benthic fish. The undeleted allele contains a regulatory element (***Pel*** cis-regulatory enhancer) that drives expression in the developing pelvic apparatus of stickleback fish.

	Benthic	**Limnetic**
Lake1	488bp deletion (1 fish)	undeleted
Lake2	undeleted	undeleted
Lake3	unknown	unknown
Lake4	unknown	unknown

Evolution & Selection
(Chan et al 2010 Science)
Pelvic reduction has evolved in numerous other isolated (non-species-pair) stickleback populations around the world via repeated and independent deletion of the same cis-regulatory element, ***Pel***. Different populations have deletions of different sizes and with different deletion boundaries suggesting parallel evolution of this phenotype involves repeated de novo mutation. Targeted genotyping of the Pitx1 locus using a high-density custom genotyping array revealed molecular signatures consistent with selection favouring the deleted allele in many of these populations: eg low levels of variation (heterozygosity, theta pi), an excess of derived alleles (Fay and Wu's H).

Open questions to think about today

- Why has pelvic reduction evolved in two of the benthic limnetic species pairs and not all four? (limited access to standing genetic variation? Limited timeline for de novo mutation?)
  Do the species pairs with pelvic reduction share the same or different deletion boundaries? What does this imply about the evolution?

- Given the low level but ongoing hybridisation between ecotypes, why do the species pair remain differentiated for this trait? (ie why haven't they merged back to being all pelvic complete or all pelvic reduced?)
  Is their evidence for selection acting on the alleles at this locus?
  What does this signature of selection look like? (size, pattern, proximity to other genomic features?).

- What are the molecular features of selection on this locus? How far do they extend? What is in linkage? Would they be detectable with other forms of reduced representation genotyping?

EXERCISE 1: Characterising the indel differences in genomic data from multiple individuals, and benthic-limnetic species pairs

EXERCISE 2: Scanning for selective sweeps across the Pel locus in benthic-limnetic species pairs.

SPECIFIC DETAILS and INFORMATION

Samples and sample bam files

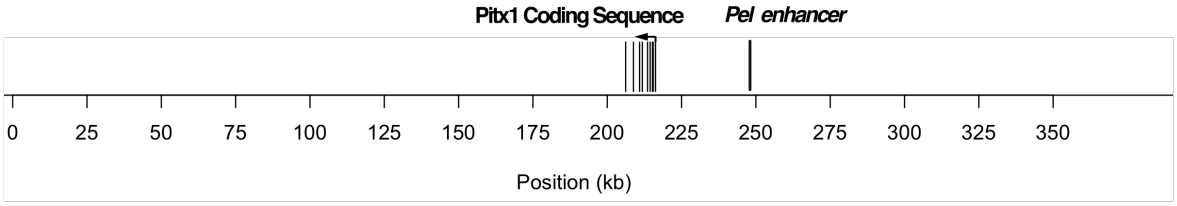
Lake	Benthic	Limnetic	Number of Individual Bam Files
Lake1	Pelvic reduced	Pelvic complete	6
Lake2	Pelvic complete	Pelvic complete	6
Lake3	Pelvic reduced	Pelvic complete	6
Lake4	Pelvic complete	Pelvic complete	6

These files already deposited on amazon server in tarball BenLim.tar. When expanded they have the file naming format **Lake1_Ben04.sorted.bam**

Reference Genome File:
(a merge of two sanger sequenced marine stickleback BAC clones covering the Pitx1 region)

/fml/chones/genome/gbdb/gasAcu1/Pitx1/SALR.Pitx1.118G22-164F21.Combined.fa [377,804bp]

This exercise introduces you to one of the dark corners of the stickleback genome. You'll be working on the far southern telomeric end of chromosome 7. So far south you'll be deep in 68bp telomeric repeats, in a region of high recombination rate that floats on a hand-assembled-scaffold tied to the end of chromosome chrVII by genetic maps. (ie its not in the assembly).



Pel enhancer coordinates:
247768-248265

Pitx1 coding sequence coordinates:

Feature	Strand	Start	Stop
PolyA	-	206193	206188
Termination	-	209315	208803
Intron	-	211079	210877
Intron	-	211952	211812
Intron	-	213745	213561
Intron	-	214559	214433
Intron	-	215397	215219
Initiation	-	215475	215459
Promoter	-	216273	216234

Extended exercise: Other genes in the region?
This scaffold likely contains more genes than Pitx1. You can run a quick analysis to identify other possible coding regions by uploading/pasting the reference file to the genscan MIT server: *http://genes.mit.edu/GENSCAN.html*
This will predict genemodels based features like start and stop codons etc etc (some of which are nonsense). It will also give you the predicted protein sequence. To find out whether there is biological support for this protein in other organisms paste the sequence into ncbi protein blast to find out what it matches.