Unknown EST analysis:

1. blastall to JGI scaffolds of 20101210.fa gives many hits that look "gene-like" in that they look like they have exons/introns:

   Full blast output is here, but note that the hits mapped above use an E-value cutoff of 1e-50.

   But the fact that it has so many such hits suggests it might be hitting common motifs. Note also that although the sequence homology is great overall, there are lots of areas of disagreement, suggesting either strain differences (lots) or poor quality sequencing (SPHS's guess).

2. Compare mapping results:

   Test 1: map all the mut/cnt reads against the full laevis scaffolds, then look to see if they happen to be where the blast results (from #1 above) also hit

   Test 2: map all the mut/cnt reads against the EST plus 10% of the laevis genome (to provide some background) to see if any number of reads would map to the EST directly

   Results: Test 2 Mapping of the control/mutant reads directly to the EST (and 10% of the laevis genome, to give some background) suggests it's "real" too:
Although it has light coverage, the coverage looks reasonable and also "gene like". You can tell there are probably errors or divergence in the sequences by the small bits of color here-and-there.

I should really generate a 641 bp random sequence and map alongside this just to see...

The problem is if I look at test 1, there are essentially no reads mapped. Since the read mapping software will randomly pick one location for reads that map to multiple locations in the genome and we already know this sequence blasts to several scaffolds, it suggests that this sequence really is repeated (though not a classic repeat) more than probably 10x in the genome, though again it may be fragmented.

I didn't find any real hits to human RNA via tblastx, or to nr via blastx, or to the CDD via rpsblast, so really not clear it makes a protein. I also checked (via NCBI web site) for blastn to refseq_rna - no hits.

**Any similarity to known miRNAs?**

In short, no. There are about six ~10 bp regions that hit mirBase, but they are lower complexity sequences and hit the loop or early in the stem; no hits to mature miRNA.

Started a RapidShapes analysis...

**Any genes anywhere near it?**

The largest scaffold that the sequence hits well (eval 3e-103) in the JGI laevis 2010 Dec. assembly is Scaffold 23920 (15kb); it hits 260265 over a longer stretch but this scaffold is only 1.9kb total, so I did rpsblast to cdd for all of Scaffold 23920. trappc5 is located about 7kb upstream of this region (evalue 2e-43). I also checked 111952, 197014, 231992 which are all also short (6.4kb, 2.6kb, and 2.1kb) but rpsblast to cdd shows these might all be the same sequence - all show some subunit of NADH as either the first or 2nd hit, and the best e-val is 3e-5 for a chemosensory receptor in 231992.

This might be a lot easier with a better assembly - any idea how that's coming along?