Data and Scripts for Furman and Evans 2016

The associated publication for this data is Furman, B.L.S and Evans, B.J. Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination. G3. Accepted August 2016.

Below is a description of each directory.

nexus_sex_linked_genes/

In our paper we assess sex linkage of several genes in a lab raised family of *X. borealis*. This directory contains the alignments of those genes. Each alignment also contains the primers used to amplify the locus.

For all of these files, the sequence names contain species, individual ID, sex (F or M; female/male), and primers used. They may also have the locus name and may indicate homology to *X. laeivs* sub-genome (indicated by L or S).

The SOX3, FMR1, and AR files contain primers, sequences from parents and offspring of our *X. borealis* family. Additionally, these files contain sequences for these loci from wild caught *X. borealis* samples (indicated by the Xen and AMNH sequence names). If opened in Mesquite, coding frame should be highlighted, when known. The wild samples, mother, father and a single daughter and son were submitted to Genbank for each locus (sequence names here will match those on Genbank).

RAB6A.nex

This file contains sequences for homeologs of the gene *RAB6A*, one homeolog of which is located physically close to the *DM-W* gene in *X. laevis.* We amplified both homeologs of this gene in our *X. borealis* family and the homeolog located close to *DM-W* in a *X. laevis* lab raised family. In the *X. laevis* family, there was an indel mutation in the mother which prevented contig formation and was inherited faithfully by all tested daughters and no sons (supporting sex linkage). The ortholog of this gene in the *X. borealis* family did not show sex linked inheritance, with both sons and daughters inheriting both maternal alleles (as seen in position 131 of the alignment). The same was true for the homeolog. The genome sequences for *X. laevis*, downloaded from XenBase, are included to demonstrate homology relationships of the sequences.

The large number of heterozygous positions present for the laevis female offspring sequences are due to the indel mutations creating uncallable sections in the chromatograms. They are likely not all real.

SOX3.nex

This file has two sets of primers, one for a longer amplicon that includes a third internal primer added at the incorporation of fluorescent dNTPs for Sanger sequences, and another set for a shorter amplicon. Some individuals were amplified with the longer primer set and others with the shorter.

Sex linkage can been seen by multiple heterozygous positions in the mother of the cross, with one allele faithfully inherited by all daughters (W-polymorphisms; examples at 68,72) or all sons (Z-polymorphisms; example at 306).

FMR1.nex

This file contains the primers used and sequences from the *X. borealis* family. This amplification spans and intron with coding region for exons on either side. Sex linkage can be seen from multiple maternal heterozygous with one allele faithfully inherited by all daughters (W-polymorphisms; examples at position 333,421) or all sons (Z-polymorphisms; examples at 265,457).

AR.nex

This file contains the primers used and sequences from the *X. borealis* family. Sex linkage can be seen from a maternal heterozygous at position 162 in the alignment, with one allele faithfully inherited by all sons (Z-polymorphisms). There may be a paternal null allele causing odd heterozygous positions at other positions in the alignment.

Scripts/

The two scripts included here performed critical parts of the data generation pipeline. They are included here so interested parties can see how a these steps were done (though the code is, admittedly, pretty opaque). Do not expect these scripts to work for an problem you may be working on; they were not intended for public consumption. If you would like to use them and they are not working feel free to contact furmanbl@mcmaster.ca (or see here) and I may be able to tweak it to help.

Parse_Trees_ID_Paralogs.R

This script took in a list of trees in a directory and assessed if homeologous sequences are present in the tree. It would assess if the tree matched our knowledge of this history of the group, whereby duplication followed speciation from the outgroup (creating 2 deeply diverged lineages splitting from the outgroup). It would assess if the most recent common ancestor (MRCA) of species with multiple sequences present was deeper than MRCAs between each of the sequences and other species. Put another way, were there closer interspecific relationships than intraspecific relationships.

thresholdUngapped_Characters.pl

This script would find the most species rich alignment that matched our constraints for phylogenetic analysis. These constraints included at least two sequences present for at least one species (potential homeologs), at least 300 bp of ungapped sequence data, at least 3 ingroup species and presence of an outgroup. This script takes in a nexus file and if the constraints are met, ignores it. If the constraints are not met, then the script will test all combinations of the total number of sequences - 1 (i.e. if 10 sequences are present in the alignment, then it will start by testing all combinations of 9), and measure how long (in number of ungapped bp) each alignment is and whether the combination meets the constraints. If all combinations fail, it tests all combinations of the total number of sequences - 2 (i.e. all combinations of 8, following the previous example). The script continues testing smaller and smaller combinations until either the constraints are met, and writes a new nexus file with just the combination of sequences that meet the meet the constraints and produce the longest ungapped alignment, or indicates if the constraints are never met and does not write a new file.

nexus_phylogeny_alignments/

These are the nexus files of all 1585 loci that were used in the phylogentic analyses. These files are the result of our pipeline that identified homeologous sequences (see publication for details). Each file contains species and a classification of “alpha” or “beta” corresponding to the two homeologous lineages.

BEAST_xml/

The xml files used in the individual gene tree construction for each of the 1585 loci.

BEAST_gene_trees/

The consensus gene trees produced by BEAST.