Alternative Splice Evolution following WGD

manu/simen/yamile

August 2017

Analyses of gene-level evolution of alternative splicing following WGD

Here use analyse the evolution of alternative splicing folowing WGD in two vertebrate genomes, the autotetraploid *Salmo salar* (atlantic salmon) genome and the allotetraploid frog *Xenopus laevis* (African clawed frog). We use gene-level estimates of alternative splice variants and compare these to the alternative splice patterns in un-duplicated sister lineages Esox lucius and X. tropicalis.

Methods

Identification of gene triplets in *S. salar*

We used a combination of synteny information, sequence similarity and gene tree topology to identify duplicated genes originating from the salmonid WGD (**referred to as Ss4R**), genes withour a Ss4R duplicate, and their un-dupicated ortholog in pike.
In brief we. . . :

- Used longest protein translation from each gene model was used in a self blastp search (evalue<1e^10)
- Selfblast results were filtered to only contain significant hits between gene loci located in pre-defined duplicated blocks originating from Ss4R
- Blast hits were filtered using:
  - minimum pident > 80%
  - minimum hit coverage (reciprocal) > 50
- Best hit were assigned 'putative duplicates based on blast+synteny'
- Gene tre topologies were then used to:
  - remove putative duplicate pairs that did not belong to the same gene family
  - remove putative duplicate pairs that represented older non-Ss4R paralogs
- Only salmon duplicates (and singletons) from gene trees with a single E. lucius ortholog was included.

Identification of gene triplets in *S. salar*

To be added

Estimation of alternative splicing

To be added

Results

Genome wide identification of gene duplicates from WGD

We identified 10923 putative Ss4R duplicate pairs from synteny and 'best hit' filtering of the selfblastp search. Out of these 7058 pairs were found to belong to the same ortholog gene tree. After filtering on gene tree

topology and the presence of a single E. lucus ortholog we were left with 6732 gene triplets representing two salmon duplicates originating fom Ss4R WGD and their ortholog in the un-duplicated E. lucius genome. Number of 1:1 orthologs (salmon:pike, i.e. singletons) were 6145.

The frog duplicates and singletons were extracted from 'the frog paper'/Xenbase. The original duplicate table contained 6352 putative 2:1 ortholog relationships to the non-duplicated Xenopus tropicalis and 3383.

Gene-level alternative exon usage

Level of alternative exon usage per gene locus was extracted from the VAST-tools output (files named 'trulyAS2') (Table 1). We filtered the raw VAST-tools output on reliability classification (i.e. 'superAS') and merged this with the ortholog triplet dataset (see above paragraph).

% latex table generated in R 3.4.1 by xtable 1.8-2 package % Fri Aug 11 12:55:25 2017

	Type	vast.SuperAS	vast.SuperAS_dups	vast.SuperAS_sing
1	Xla_ASevents	21217.00	3206.00	1765.00
2	Ssa_ASevents	15050.00	3070.00	2758.00
3	Frog_ASevents_perGene	2.40	2.30	1.80
4	Ssa_ASevents_perGene	2.10	1.50	1.00

Table 1: Summary of AS events included in analyses

Duplicate evolution of gene-level alternative exon usage

To analyse the gene-level evolution of AS events (loss or gain) following WGD we calculated the difference between AS events between WGD duplicates and their un-duplicates orthologs. The results are plotted as barplots in Figure 1. If partitioning of AS events are symmetric among the WGD duplicates (as suggested by some) then we would expect the barplots in the first and second rows in collumn 1 and 2 to be similar to eachother. This is obviously not the case, as the duplicate copy with the 'fewest' detected AS events ('less' in the figure headings) have MUCH fewer AS events compared to the unduplicated sister group we are comparing against.

The higher AS events in pike compared to salmon (not seen for frogs) we interpret to be related to the technical ability to detect AS events when duplicates are really similar at the CDS level. This is backed up by the AS event distribution for singleton genes (Figure 2) where clearly pike has more AS compared to salmon on average, while the frog comparison seems perfectly balanced.

- NOTE: Must correct for gene expression levels as these can influence the detection of AS events.. How should we approach this?
  - Simply filter out genes that are equal in expression levels?
  - Plot fpkm's difference as boxplot for 'low AS' duplicates versus 'high AS' duplicates

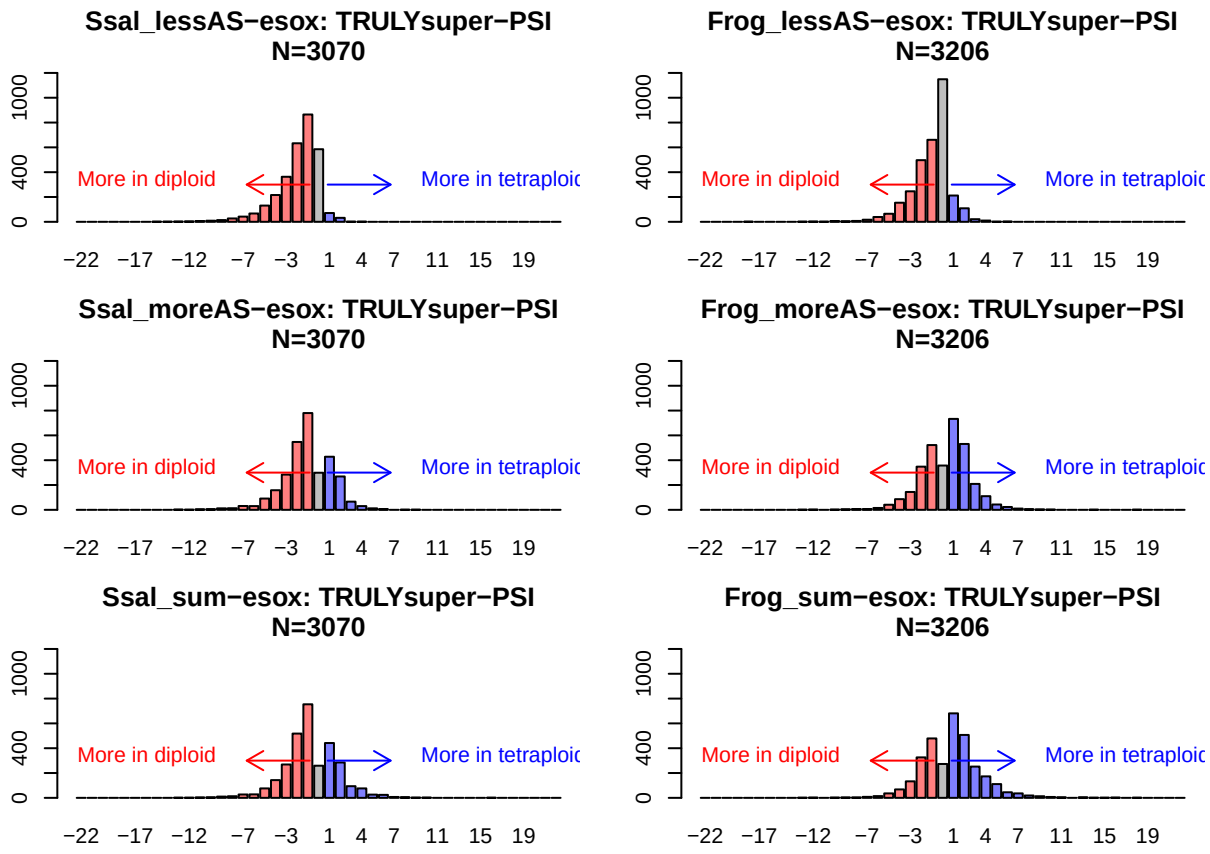

Figure 1: Figure1: Difference distribution in AS events per gene for WGD duplicates.

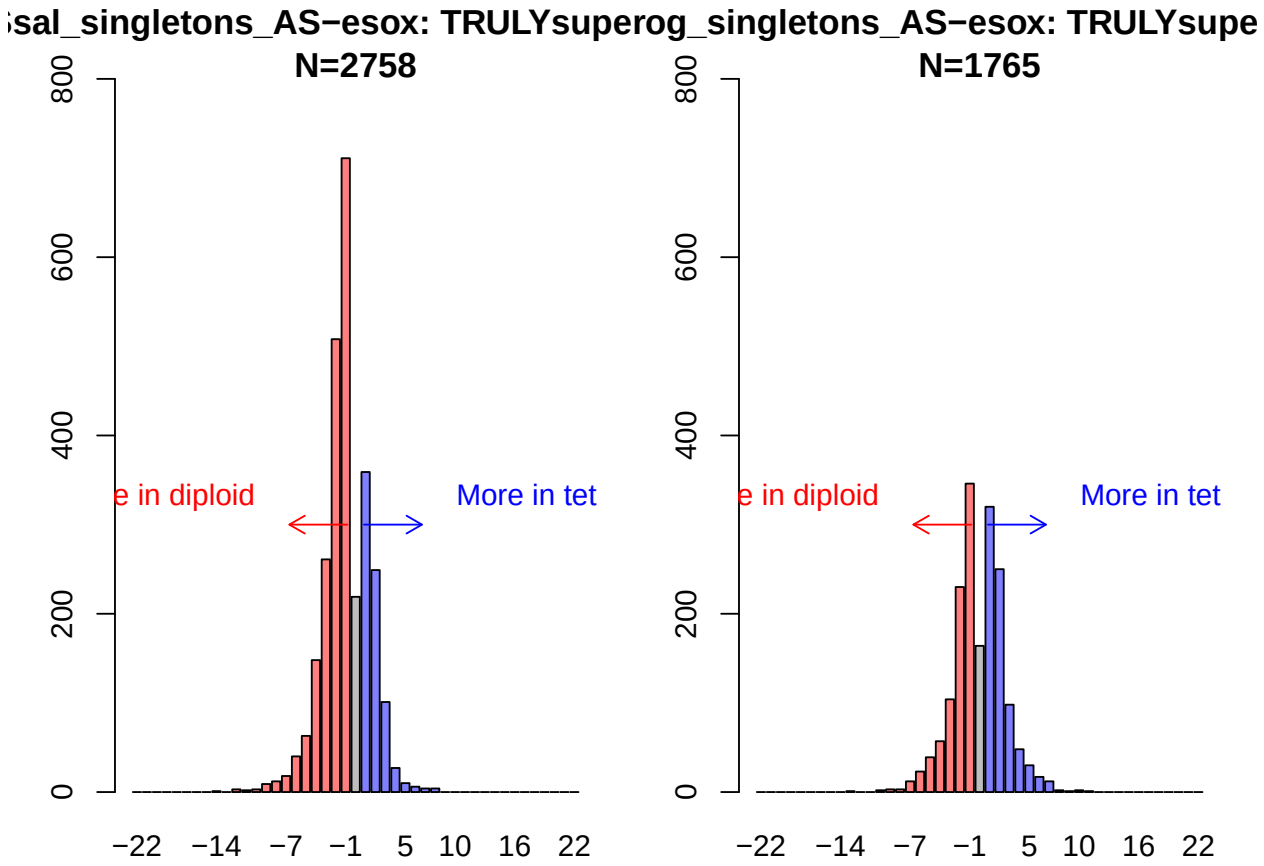

Figure 2: Figure2: Difference distribution in AS events per gene singleton in tetraploid and unduplicated ortholog. Left: salmon vs pike. Right: frog vs frog.