FUSTr: a tool to find gene Families Under Selection in Transcriptomes

**Background**

Recent advances in RNA-Seq technologies have allowed for an abundance of protein coding sequence data to be generated across all levels of biodiversity[1–4]. In non-modeleukaryotic study systems, transcriptomic experiments have become the de facto approach for functional genomics in lieu of whole genome resequencing. This is due largely in part to lower costs [5], better targeting of coding sequences [6], exploration of posttranscriptional modifications and differential gene expression [7,8]. This influx of transcriptomic data has resulted in an ever-expanding need for scalable tools to automate evolutionary analysis at the genomic scale. ~~of elucidating patterns and processes involved in the adaptive evolution of genes and genomes of organisms throughout the tree of life.~~

Elucidating patterns and processes involved in the adaptive evolution of genes and genomes of *organisms is fundamental to* developing an understanding of the vast phenotypic diversity found in nature.  Speciation events ~~forming gene orthologs~~ along with frequent whole genome duplications ~~forming gene paralogs~~ has given rise toamyriadofmultigene families that span a broad range of biochemical properties. There are several families adaptive and fitness familes contain genes that contribute to organisms fitness and adaptive in *ways.* with *various* biochemical properties that contribute the vast phenotypic diversity~~which has contriuted to the vast phenotypic diversity found across all domains of life~~ (i.e  [9–11].

* phenotypic diversity  (i.e. morphological, behavioral, physiological, etc.)
* elucidating patterns and processes involved in the adaptive evolution of genes and genomes of organisms throughout the tree of life.

Grouping protein encoding genes into their respective families de novo has remained a difficult task computationally ~~that has been shown to be an NP-hard problem.~~ This typically entails homology searches in large amino acid sequence similarity networks with graph partitioning algorithms in order to cluster coding sequences into *transitive* groups[12–14]*.* This is further complicated in eukaryotic transcriptome datasets that contain several isoforms via alternative splicing, which cannot be treated as phylogenetically independent homologs, *more words* [**??]**. Further analysis of these gene families is also non-trivial, as it requires multiple sequence alignment followed by phylogenetic inference~~, both of which has been demonstrated to be NP-complete problems~~ [15–18]. Further exploration of patterns of molecular evolution in these families is also computationally intensive, requiring robust phylogenetic analysis using codon substitution models with random or mixed effects likelihood methods in addition to MCMC Bayesian statistical frameworks, in order to determine patterns of pervasive diversifying selection or episodic lineage based diversifying selection [19, 20].

        Here we present FUSTr, a tool to address the aforementioned difficulties of characterizing molecular evolution in large biodiversity transcriptomic datasets in a pipeline capable of scaling to multicore high-performance computational facilities. FUStr can be used to characterizing selective regimes on homologous groups of phylogenetically independent coding sequences in transcriptomic datasets and has been verified using Arachnoserver and simulated datasets. The presented pipeline implements simplified user experience with minimized third-party dependencies, in an environment robust to breaking changes to maximize reproducibility over a long-term time scale.

FUSTr is freely available under a GNU license and can be downloaded at <https://github.com/tijeco/Fuster>

***P4***

* Here we present things that do stuff
  + Deal with complexities in the previous stated issues
* What it is
* What it does
* Why its useful?
* How it performs/accuracy
* Where it is

**Implementation**

FUSTr is available as a Snakefile using snakemake, to ensure scalbility and reproducibility. In order to further increase the reproducibility, all necessary dependencies have been wrapped up in an Anaconda environment.

**Pre-filter data**

Input data consists of assembled transcriptome fasta files, nucleotide data. These files are first filtered to remove sequences just containing Ns as well as to removing any haphazard text found in sequences (that may be artifacts of previous assembling procedures) to ensure proper downstream analysis. Header patterns in the inputs are simulteneosly autodetected to sort out all unique and redundant identifiers for downstream isofom filtration.

|  |  |  |
| --- | --- | --- |
| gene | AOXIE | id=a |
| gene | 111 | id=b |
| gene | 2342 | id=a |
| gene | 2342 | id=b |
| gene | 2342 | id=c |

* Algorithms, bitches and money, layout of typical gene and isoform header patterns

**Predict coding sequences**

Coding sequences are determined using Transdecoder, which predicts orfs using all 6 possible reading frames, keeping the best orf per transcript. Genes containing several isoforms will only keep longest isoform for downstream stuff. protein sequences, stuff. Keep only longest isoform.

**Group by homology**

Sequence homology networks are determined using an all against all BLAST search of the amino acid sequences. Sequences are grouped into homologous groups using Silix, only adding sequences to a group that have at least ### sequence coverage, ### sequence identity, &&&& other stuff. All against all blast. Silix. Families.

**Multiple-sequence alignment and phylogenetic inference**

Homologous groups containing at least 15 amino acid sequences are aligned using Mafft (accurate and fast, auto), and then trimmed using trimal. Fastree is used to infer phylogenetic tree (accurate and fast) Mafft, trimal, codon masking. Codon sequences are masked over amino acid alignment.

**Pervasive positive selection**

Codon alignments. CODEML, pervasive positive selection.Codeml and FUBAR??

~~There are several established phylogenetic frameworks to then test the families for various forms of selection utilizing  rates of nonsysnonymous to synonymous substitution. Determining selective regimes of specific amino acid~~ *~~residue~~* ~~sites that~~ *~~may be adaptive~~* ~~involves tests of pervasive positive selection within a gene family using either fixed or random effects likelihood models to help ellucidate specific amino acid residue sites undergoing diversifying or purifying selection (FUBAR, M8). Finding specific lineages that may have undergone~~ *~~adaptive surges~~* ~~(niche diffenertiation, sexual selection, predator prey arms races, novel innovations, lots of evolution terms) requires tests for episodically diversifying lineages (MEME,BUSTED,CodemlBranchSpecific) or evolution along specific branches in gene family…~~

**Episodic positive selection**

Hyphy/MEME, FUBAR, BUSTED

**Validation**

Simulations, GLOve??? Bacteria stuff (POTION), and arachnoserver (compare to young clade paper).

Results

**Simulations**

We did stuff, yeah.

**Empirical results**

Arachnoserver, toxinbase, bacteriashit, goldenstandard.

**Conclusions**

~~An influx of~~ *~~terabytes~~* ~~of transcriptomic data has resulted in an ever expanding need for scalable~~ tools capable of elucidating broad patterns of molecular evolution within the genomic architecture of taxa spanning throughout the tree of life.