SSS-test: a novel test for detecting positive selection on RNA secondary structure

Maria Beatriz Walter Costa, Christian Höner zu Siederdissen, Marko Dujić, Peter F. Stadler, and Katja Nowick

Long non-coding RNA (lncRNA)

- Majority of human transcriptome is ncRNA
  - ~40k – 50k human lncRNA
- No functional annotation for majority of lncRNAs
  - What function of detectable lncRNAs are biologically functional versus “junk RNA”?

Evidence for selection in lncRNAs

- Most lncRNA have low levels of sequence conservation
  - Population genetics would interpret this as low functional constraints
- As a group, lncRNA have cumulative substitution and transversion rates lower than neutrally evolving DNA
  - Suggests some level of negative selection
- Overall sequence conservation is low
- Gene structure and splice sites are usually well conserved
- Many lncRNAs located in same chromosomal positions and have similar expression patterns across species

SSS-test: Selection on the Secondary Structure Test

- **Goal: Identify and quantify the selective pressures on RNA secondary structures**

SSS-test

- **Goal: Identify and quantify the selective pressures on RNA secondary structures**
  - Focus on smaller blocks of lncRNA
  - Principally work on identifying lineage-specific positive selection
- Previous work in this area done on compensatory mutations to identify negative selection

Structural Conservation

- Only tolerates small deviation around well-defined consensus structure
  - Mutated sequences must have enough compensatory mutations to preserve structure
  - Mainly occurs in small ncRNAs and structured regulatory elements
- LncRNAs almost never structurally conserved

Negative Selection

- Less stringent than structural conservation
- Structural variation is *more constrained than it would be* given no selective pressures
  - Observed in ncRNAs: DNA sequence usually evolves rapidly but signs of selection on local secondary structures
- At least 10% of non-repetitive sequences in human genome under negative selection on RNA secondary structures

Negative Selection in Human lncRNA

- LncRNA evolve on average like unconstrained background
- Evidence of conserved gene structure
  - Splice sites
- Selective pressures don’t enforce large conserved consensus structures

Positive Selection on Secondary Structure

- Very little known
- Control for ncRNA structures: Human Accelerated Region 1 (HAR1)
  - 118-nucleotide region
  - Very conserved in non-human mammals
  - 18 human-specific single nucleotide substitutions
  - Fastest evolving region in human genome
  - Forms a stable structure in humans
  - Might be part of cortex development
  - Unknown if function depends on secondary structure

Detecting Positive Selection

- No available method to systematically detect positive selection on RNA secondary structure
- Simple approaches:
  - K_a/K_s test (and variants)
  - Divergence and diversity modeling

K_a/K_s test for coding sequences

- $K_a = \frac{\#\ non-synonymous\ substitutions}{\#\ non-synonymous\ sites}$
- $K_s = \frac{\#\ synonymous\ subsitutions}{\#\ synonymous\ sites}$

- $\frac{K_a}{K_s} > 1$ suggests positive selection

Divergence and Diversity Modeling

- $\rho \leftarrow$ fraction of sites under selection
- $\lambda \leftarrow$ polymorphism rate
- $\eta \leftarrow$ divergence rate
- Normalize parameters by a neutral control group
- Analyze for signs of selection
- Mainly used for groups of loci
  - Has shown strong evidence of selective pressures on regulatory elements

Measuring Phenotype

- Effect of indels and structural variation not well understood
- If ncRNA function depends on secondary structure, can be a proxy for phenotype
- Accumulation of mutations that change structure as evidence for positive selection

Intuition for Selection Identification

- Some previous work considered SNPs impact on secondary structure
  - Excess of structure-changing SNPs implies positive selection
  - Excess of structure-conserving SNPs implies negative selection
- Develop a statistical test
- Identify candidate lncRNAs for human-specific positive selection

SSS-test Theory

- $\boldsymbol{\mathcal{A}} \leftarrow$ multiple sequence alignment of orthologous RNA sequences from a set of species of interest
  - Use a primary structure alignment
- $\boldsymbol{x} \in \boldsymbol{\mathcal{A}}$ is the focal sequence
- $\overline{\boldsymbol{\mathcal{A}}} = \boldsymbol{\mathcal{A}} \setminus \{\boldsymbol{x}\}$ is the background distribution
- $\overline{\boldsymbol{z}}$ is the consensus sequence of $\overline{\mathcal{A}}$

SSS-test Theory

- $\mathcal{A} \leftarrow$ multiple sequence alignment of orthologous RNA sequences from a set of species of interest
  - Use a primary sequence alignment
- $x \in \mathcal{A}$ is the focal sequence
- $\overline{\mathcal{A}} = \mathcal{A} \setminus \{x\}$ is the background distribution
- $\overline{z}$ is the consensus sequence of $\overline{\mathcal{A}}$
- **Do mutations to produce $\overline{z} \rightarrow x$ change secondary structure more than expected?**

Candidate families

- To identify lineage-specific positive selection on secondary structure, only consider well-conserved families
  - Suggests structure is biologically relevant
- Quantify family's structural uniformity
  - $\boldsymbol{d_s} \leftarrow$ species distance scores
  - $\boldsymbol{d} \leftarrow$ family divergence score
    - $d = median\ (\{d_s : s \in family\})$
- Only consider families with $d \leq t$
  - Empirically determine t

Family divergence d

- Quantify structural divergence in family of orthologs
- $\boldsymbol{A_s} \leftarrow$ base pair probability matrix for aligned sequence $s \in \mathcal{A}$
- $\boldsymbol{B} \leftarrow$ base pair probability matrix of alignment $\bar{\mathcal{A}}$
- $\boldsymbol{P_s} \leftarrow$ set of base pairs in s
- $\boldsymbol{Q} \leftarrow$ set of base pairs in $\bar{z}$
- $\boldsymbol{W_s} = P_s \cap Q$, shared base pairs
- $\boldsymbol{X_s} = P_s \setminus Q$, unique base pairs
- $\boldsymbol{Y_s} = Q \setminus P_s$, absent base pairs

Family divergence d

- Divergence of sequence s from alignment $\mathcal{A}$ is

$$d_s = \frac{100}{length(\mathcal{A})} \times \left(\sum_{ij \in W_s} |A_{s,ij} - B_{s,ij}| + \sum_{ij \in X_s} A_{s,ij} + \sum_{ij \in Y_s} B_{s,ij} \right)$$

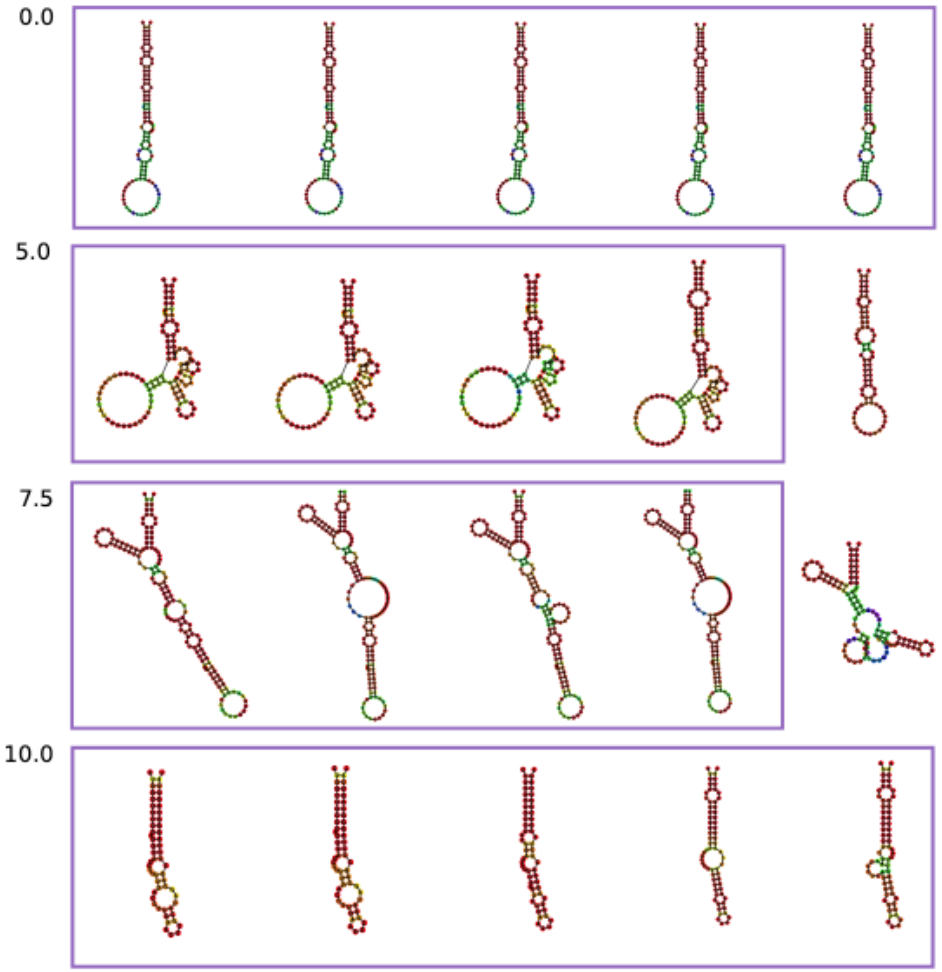
Family divergence d

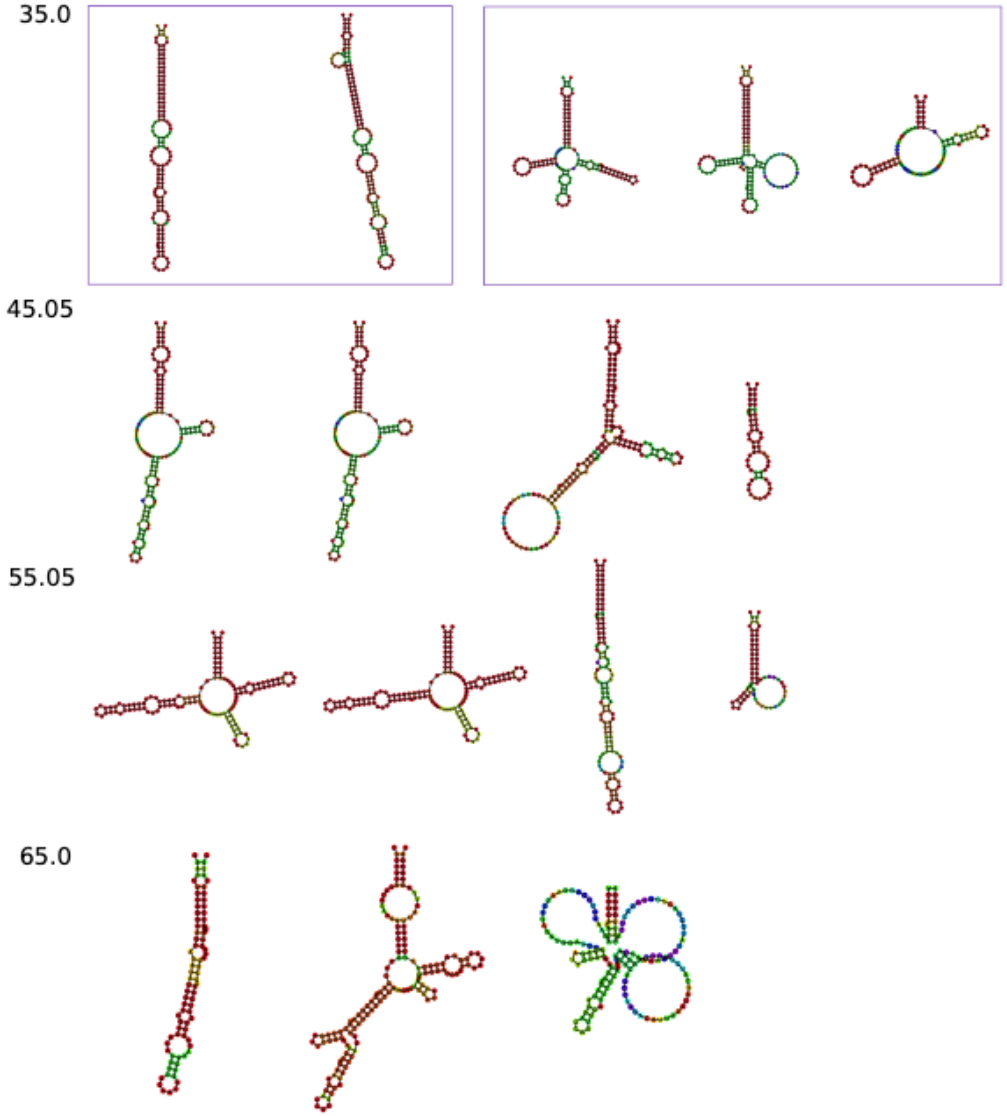
- Distance of sequence s from alignment $\mathcal{A}$ is

$$d_s = \frac{100}{length(\mathcal{A})} \times \left(\sum_{ij \in W_s} |A_{s,ij} - B_{s,ij}| + \sum_{ij \in X_s} A_{s,ij} + \sum_{ij \in Y_s} B_{s,ij} \right)$$

- Family divergence $d = median(\{d_s : s \in \mathcal{A}\})$
  - Found $d \in [0.0, 65.0]$ for 12 families of ncRNAs
  - Empirical threshold $d \leq 10.0$

Family divergence d





Candidate Selection Sites

- Interested in lineage-specific changes so only consider well conserved sites
  - Majority of $y \in \overline{\mathcal{A}}$ conform to $\overline{z}$

Candidate Selection Sites

- Interested in lineage-specific changes so only consider well conserved sites
  - Majority of $y \in \overline{\mathcal{A}}$ conform to $\overline{z}$
- $\boldsymbol{S_{\overline{z} \to x}}$ is the set of well-conserved sites that differ between $\overline{z}$ and x
  - Includes indels
- $\boldsymbol{\overline{z_i}} \leftarrow$ sequence where $\overline{z_i} = \overline{z}$ everywhere except i, and $\overline{z_i} = x$ at i
  - Score substitutions and indels separately

Compensatory Mutations

- SSS-test considers sites individually so can't account for compensatory mutations
- Removes all compensatory mutations from $S_{\overline{z}\to x}$
  - Computes consensus structure of $\overline{\mathcal{A}}$ and x with `RNAalifold` and `RNAfold`
  - Substitution/pair of substitutions considered compensatory if they form a base pair in the MFE structure of x and the MFE structure of $\overline{\mathcal{A}}$

Compensatory Mutations

- SSS-test considers sites individually so can't account for compensatory mutations
- Removes all compensatory mutations from $S_{\overline{z} \to x}$
  - Computes consensus structure of $\overline{\mathcal{A}}$ and x with `RNAalifold` and `RNAfold`
  - Substitution/pair of substitutions considered compensatory if they form a base pair in the MFE structure of x and the MFE structure of $\overline{\mathcal{A}}$
  - Removing these mutations could mask negative selection signals

Scoring substitutions

- Score all single nucleotide substitutions in $S_{\overline{z}\to x}$
  - Use RNAsnp to produce p-value for hypothesis that structural change caused by SNP is larger than expected
    - Expectation computed from same base exchange in random sequences with same length and GC content
  - RNAsnp benefits
    - Computational efficiency
    - Computes Boltzmann ensemble and not just MFE secondary structures
    - Evaluates structural change in region of maximal structural differences
      - Expect structural impact of SNP to be localized

Scoring substitutions

- Generated p-values for each SNP individually
- Benjamini-Hochberg procedure for p-value correction
  - Works well for large number of p-values that are individually ≥ 0.05
- Define $p = p_1 \geq p_2 \geq \cdots \geq p_n$
- $\tilde{p}_1 = \min\{1, p_1\}$
- $\tilde{p}_i = \min\{1, \tilde{p}_{i-1}, \frac{n}{n-i+1} p_i\}$

Scoring substitutions

- Generated p-values for each SNP individually
- Benjamini-Hochberg procedure for p-value correction
  - Works well for large number of p-values that are individually ≥ 0.05
- Define $p = p_1 \geq p_2 \geq \cdots \geq p_n$
- $\tilde{p}_1 = \min\{1, p_1\}$
- $\tilde{p}_i = \min\{1, \tilde{p}_{i-1}, \frac{n}{n-i+1} p_i\}$

- **Substitution score: $s(x) = -\sum_i \log \tilde{p}_i$**

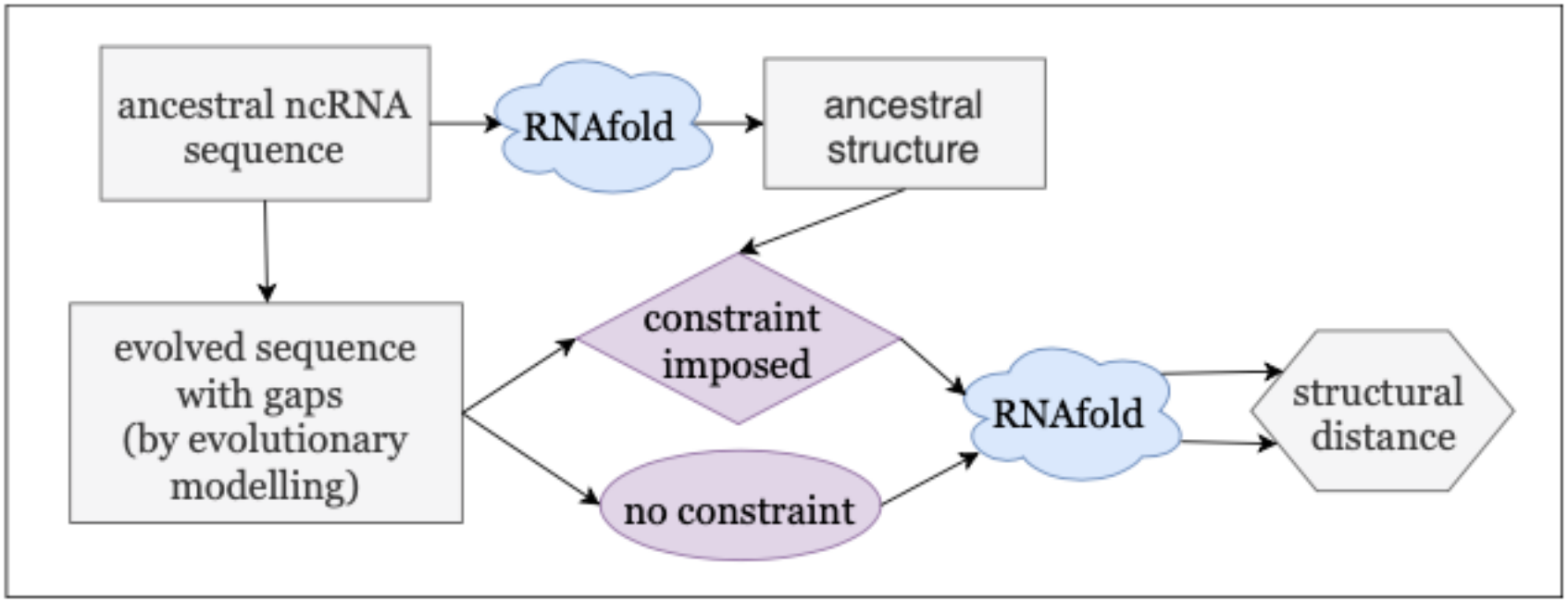
Scoring indels

- Treat indel as a single event regardless of length l
  - Most likely caused by a single evolutionary event
  - Energy penalty varies little with loop length
    - ~1-3 kcal/mol for loops from 3-30 nt
  - Experimental validation
- RNAsnp not designed to handle indels

Scoring indels

- For indel of length l:
  - construct all sequences $\boldsymbol{z_j}$ that carry indel after position j in $\overline{z}$
  - z_j and $\overline{z}$ had different lengths, so must have different structures
  - $\boldsymbol{\psi_j} \leftarrow$ modified reference structure of z_j
    - Constrained to contain all base pairs that consensus structure of $\overline{z}$ that aren't affected by the indel after position j
    - Compute with user-defined constraints using `ViennaRNA`
  - $\boldsymbol{\phi_j} \leftarrow$ unconstrained structure of z_j
  - $\boldsymbol{\delta(\phi_j, \psi_j)} \leftarrow$ quantifies structural difference with `RNAforester`

Scoring indels



Scoring indels

- Use rank statistics and relative structural impact to determine p-value for indel at location j
  - $\boldsymbol{r(j)} \leftarrow$ rank of structural impact of indel j in decreasing order
  - $\boldsymbol{p_{rank}} = \frac{r(j)}{n}$
  - $\boldsymbol{p_{struc}} = \frac{4l - \delta(\phi_j, \psi_j)}{4l}$, clamped to $\frac{1}{4l}$

Scoring indels

- Use rank statistics and relative structural impact to determine p-value for indel at location j
  - $\boldsymbol{r(j)} \leftarrow$ rank of structural impact of indel j in decreasing order
  - $\boldsymbol{p_{rank}} = \frac{r(j)}{n}$
  - $\boldsymbol{p_{struc}} = \frac{4l - \delta(\phi_j, \psi_j)}{4l}$, clamped to $\frac{1}{4l}$

- $\boldsymbol{p = p_{rank} + p_{struc}}$

Scoring indels

- Use Benjamini-Hochberg procedure again
  - Produce $\tilde{p}_i$ for each indel p-value

- **Indel score: $s'(x) = -\sum_i \log \tilde{p}_i$**

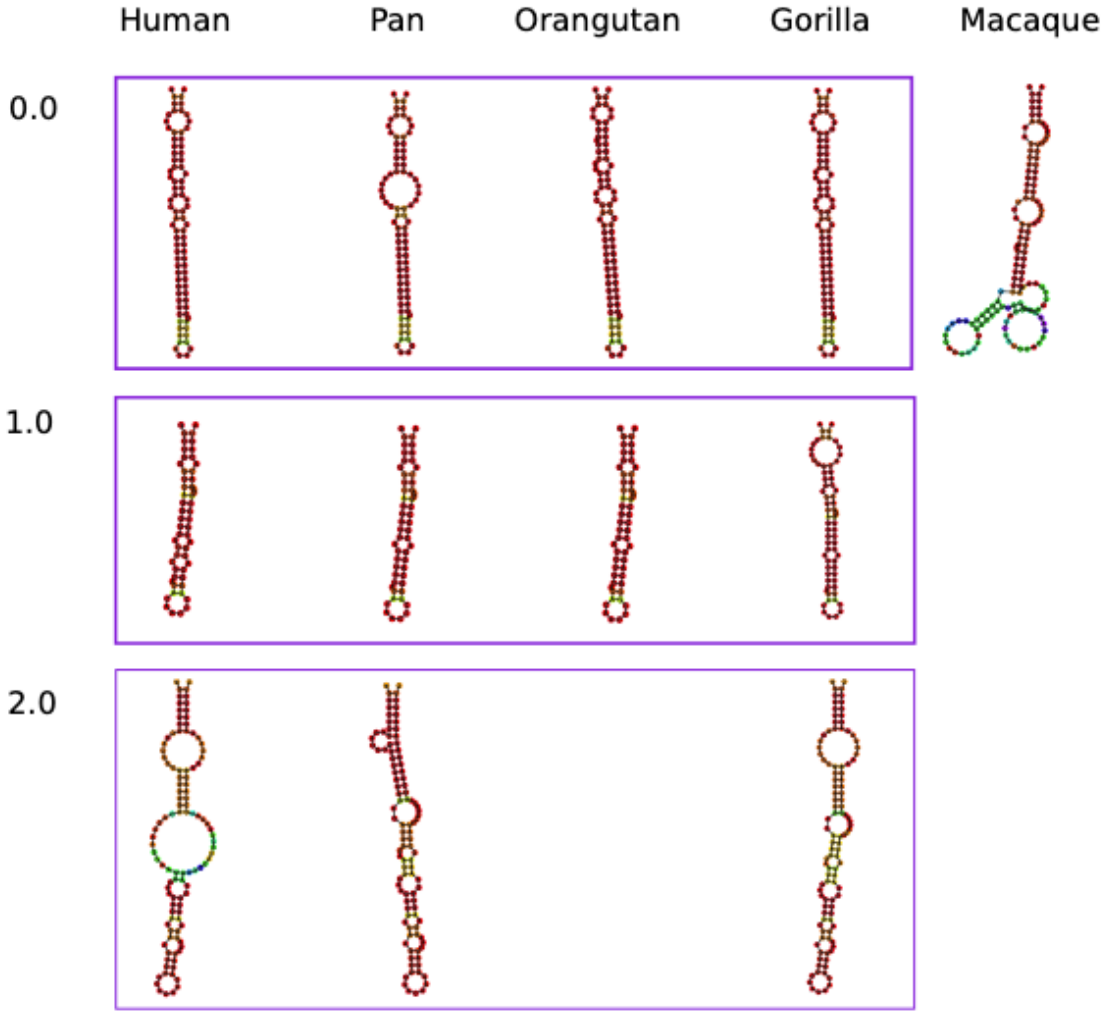
SSS-score

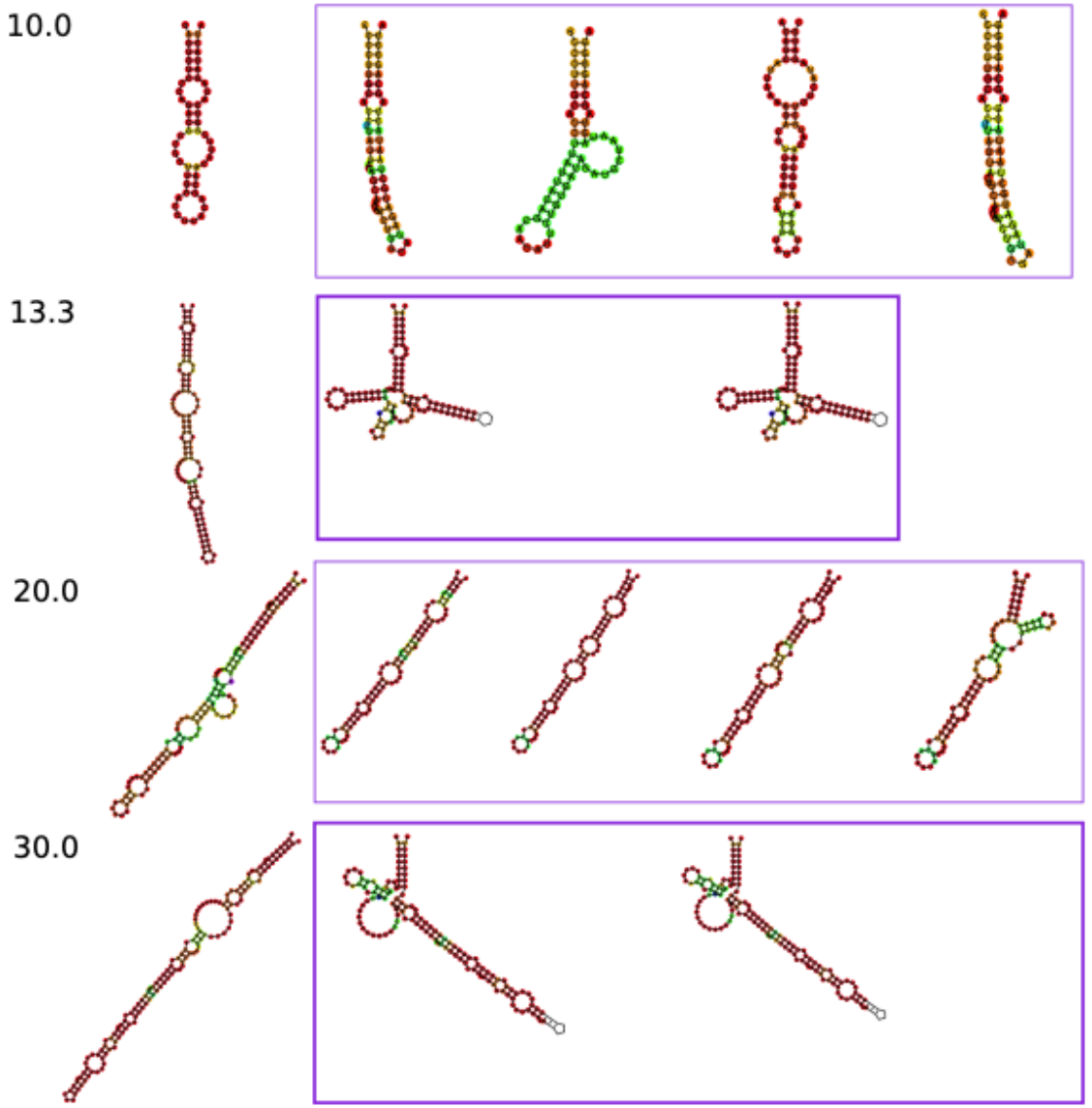
- **SSS-score** $= \boldsymbol{2s(x) + s'(x)}$
  - $s(x)$ and $s'(x)$ measure how unexpected large the impacts of observed sequence variations on secondary structure are

SSS-score

- **SSS-score** $= \boldsymbol{2s(x) + s'(x)}$
  - $s(x)$ and $s'(x)$ measure how unexpected large the impacts of observed sequence variations on secondary structure are
  - Weighting determined empirically for datasets of interest
- Can't directly be interpreted as a probability
  - One area for future work
- Serves as a test statistic
  - Relevant thresholds must be determined empirically
  - For primate experiment, find SSS-score ≥ 10.0 suggests positive selection
  - For primate experiment, find SSS-score ≤ 2.0 suggests negative selection

SSS-score





Alternatives

- Extension of K_a/K_s test
  - Comparing rates of synonymous and non-synonymous substitutions in coding sequences

Extending K_a/K_s test to ncRNAs

- Don’t have analogous distinction between synonymous and non-synonymous substitutions
- Classify sites as “disruptive” and “non-disruptive”
  - Small number of sites -> lower power
  - High FPR
- ncRNA structure’s biochemical properties make this hard to binarize

Extending K_a/K_s test to ncRNAs

- Poisson distribution of “disruptive” and “non-disruptive” sites
  - Don’t directly compare substitution counts
- More robust than counts
- Still have problems due to binarization
- Suggest that K_a/K_s test does not extend well to ncRNAs

Experiments

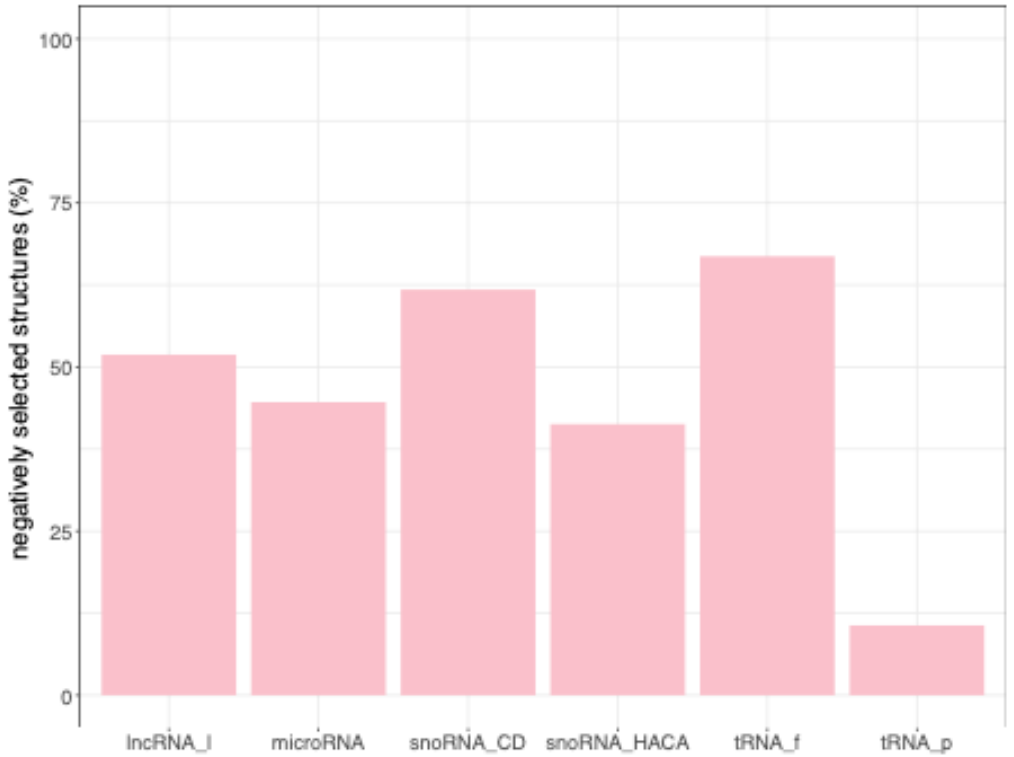
- Control experiment
- Synthetic experiments
- Primate experiments

Control Experiment

- Structurally conserved small ncRNAs
  - miRNA
  - snoRNA
  - tRNA
  - Expect low SSS-score
- Positive selection on HAR1 secondary structure

Control Experiment

- Structurally conserved small ncRNA
  - miRNA
  - snoRNA
  - tRNA
  - Expect low SSS-score



- Positive selection on HAR1 secondary structure
  - SSS-test score of 12.8 for humans
  - SSS-test score of 0.0 for other seven primates

Synthetic Experiments

- Simulate negative selection, neutral selection, and positive selection
- Two goals:
  1. Distinguish conserved families from neutrally-evolving families
  2. Distinguish lineages undergoing positive selection for otherwise conserved family

Synthetic Experiments

- Generate 150 nt origin sequence with `RNAdesign`
- Generate 100 families from origin sequence
  - Randomly mutate starting sequence
  - Accept mutation according to optimization function f
  - Continue simulation until n mutations accepted
- Lineage-specific positive selection
  - Simulate evolution from origin to one extant branch
  - Keep other four branches identical to origin sequence (extreme negative selection)

Synthetic Optimization Functions

- $f_{neg} \leftarrow$ negative selection
  - Penalize deviation from ancestral structure
- $f_{rand} \leftarrow$ no selective pressure
  - Always accept mutation
- $f_{pos} \leftarrow$ positive selection
  - Prefer mutations that move from ancestral Y-shaped structure to cloverleaf structure

Synthetic Optimization Functions

- $\boldsymbol{a} \leftarrow$ ancestral sequence
- $\boldsymbol{m} \leftarrow$ current sequence to design

Synthetic Optimization Functions

- $a \leftarrow$ ancestral sequence
- $m \leftarrow$ current sequence to design
- $\boldsymbol{\varepsilon(a, m)} = \left(\max\left(0, \mathrm{mfe}(m) - \frac{\mathrm{mfe}(a)}{2}\right)\right)$
  - Stabilizing parameter
  - Prevents degenerate structures from forming

Synthetic Optimization Functions

- $a \leftarrow$ ancestral sequence
- $m \leftarrow$ current sequence to design
- $\boldsymbol{\Delta(a, m)} = \text{base pair distance}(a, m)$
  - Constrain base pair distance

Synthetic Optimization Functions

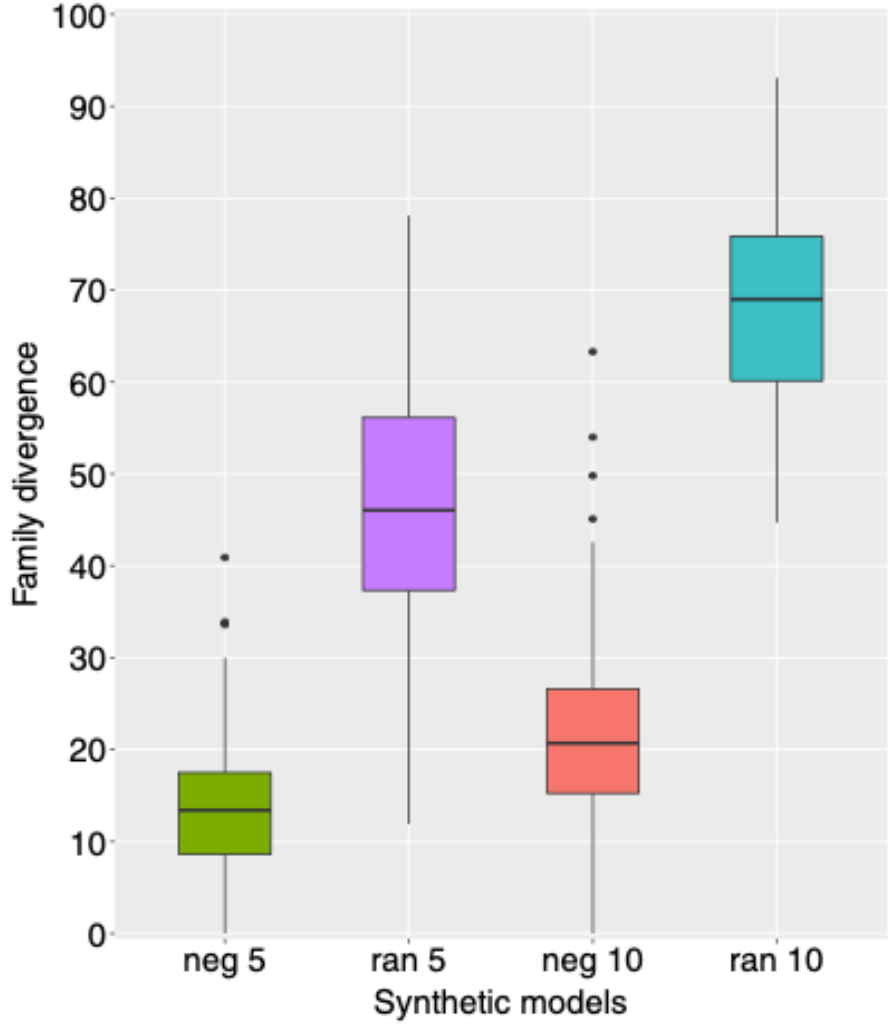
- $a \leftarrow$ ancestral sequence
- $m \leftarrow$ current sequence to design
- $\boldsymbol{\Delta_{shape:5}([\,[]\,[]\,[]\,], m)} \leftarrow$ penalize distance to cloverleaf structure
  - Shapes are coarse-grained representations of secondary structure
  - Use level 5 representation (most abstract)

Synthetic Optimization Functions

- $f_{neg}(a, m) = 1000\ (\Delta_{centroid}(a, m) + \varepsilon(a, m))$
- $f_{rand}(a, m) = 0$
- $f_{pos}(a, m) = gibbs(m) + 50\ \Delta_{shape:5}([\ []\ []\ []\ ], m) + 1000\ \varepsilon(a, m)$

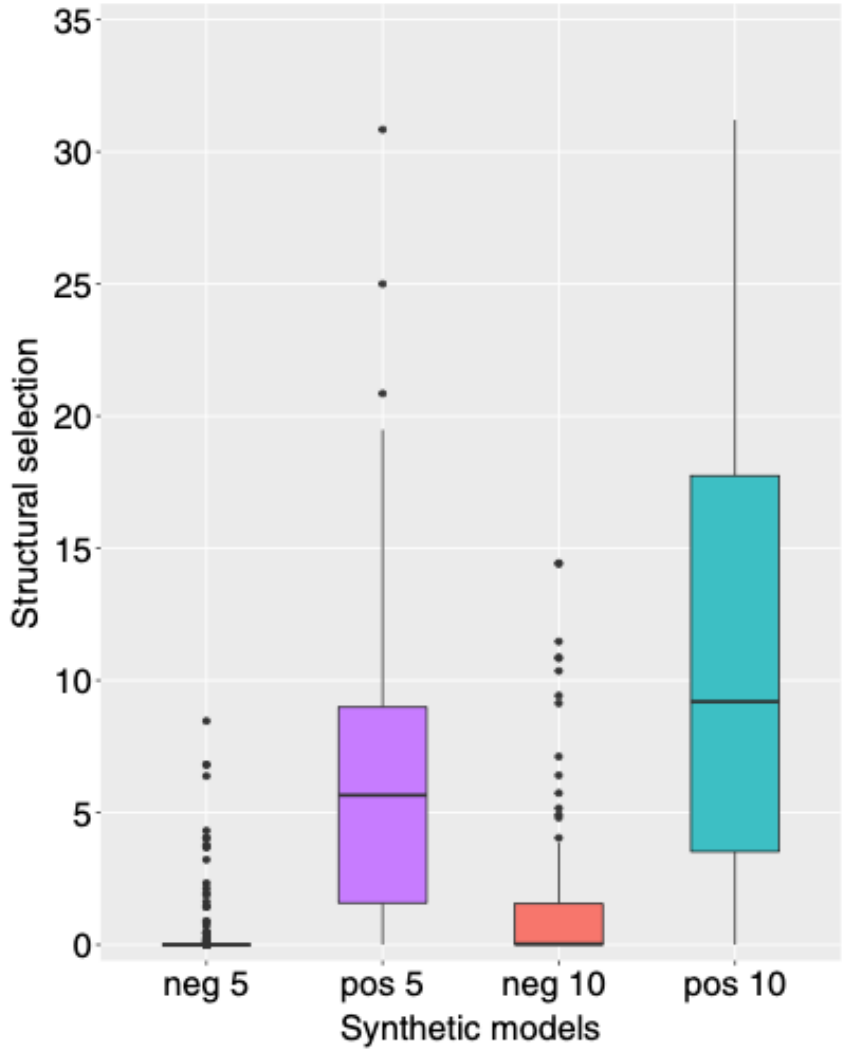
Synthetic Experiment: Conserved Families

- Goal 1: Distinguish conserved families from neutrally-evolving families
  - Found lower family divergence d for families with simulated negative selection pressure



Synthetic Experiment: Positive Selection

- Goal 2: Distinguish lineages undergoing positive selection for otherwise conserved family
  - Found higher SSS-score for lineage with positive pressure compared to negative selection



Primate Experiments

- Operate on local structural blocks, not full lncRNA
  - Most base-pairing interactions in longer RNA occur within short 150-200 bp range
  - Expect that evolution acts on local folds of lncRNA, not entire structure
  - Search for positive selection locally

Primate Experiments

- Begin with 15,443 orthologous lncRNA families
- Compute local RNA blocks with `RNALfold`
  - Computes mfe structures with restricted base pair span
  - Calculates 87,613 local blocks
- Require an orthologous block in at least 3 species
  - Defined a ‘well-conserved site’ as 60% (majority) of sequences agreeing with consensus sequence at that site
  - Filters to 19,408 conserved blocks
- Require low family divergence ($d \leq 10.0$)
  - Filters to 10,396 blocks

Primate Experiments

- Detect 1390 local structures as candidates for positive selection on secondary structure
  - Roughly proportional to evolutionary distance between species

Table 1 Characterization of local structural selection of lncRNAs

Species	Local structures	Conserved ($s \leq 2$)	Positive ($s \geq 10$)	
Human	8934	8179 (91.6%)	111	(1.2%)
Pan	8736	7997 (91.5%)	90	(1.0%)
Gorilla	8080	7199 (89.1%)	136	(1.7%)
Orangutan	6435	4802 (74.6%)	315	(4.9%)
Macaque	5113	2659 (52.0%)	738	(14.4%)

Primate Experiments: FDR

- $F \leftarrow$ number of positive test results in “foreground” dataset
- $R \leftarrow$ number positive test results in “background” dataset of same size
- $FDR = \frac{R}{F}$

Primate Experiments: FDR

- Compute background set using `SISSIz -s`
  - Simulates multiple alignments of the same dinucleotide content
  - Goal: destroy correlation of alignment columns and secondary structure
  - Consider all test results on background set to be false positives

Primate Experiments: FDR

- Randomized local blocks in humans with SISSIz
  - Produce 50 candidates for positive selection in humans
  - Estimate $FDR = \frac{50}{111} = 45\%$

Primate Experiments: FDR

- Compute background set using `SISSIz -s`
  - Simulates multiple alignments of the same dinucleotide content
  - Goal: destroy correlation of alignment columns and secondary structure
  - Consider all test results on background set to be false positives
- Empirically found that this keeps some “foreground” signal
  - Ran `SISSIz` 20 times
  - Estimated fraction of tests f where foreground signal maintained
- Updated estimate: $FDR = (1 - f)\frac{R}{F}$

Primate Experiments: FDR

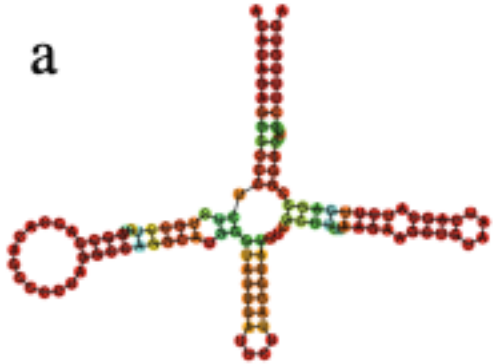
- Randomized local blocks in humans with SISSIz
  - Produce 50 candidates for positive selection in humans
  - Estimate $FDR = \frac{50}{111} \approx 45\%$
- Foreground signal maintained
  - Repeatedly running `SISSIz` on candidates from real data shows ~18.5% maintain foreground signal
  - Found that about $0.185 * 111 = 20$ of 50 predictions maintained some foreground signal in simulated alignment
  - Updated estimate $FDR = (1 - .4)\frac{50}{111} < 30\%$
- Comparable to most surveys for negative selection

Positively Selected Structures in Humans

- Detected changes in form and stability for various lncRNA
- Likely a large false negative rate due to small divergence between primates
- SIX3-AS1
  - Local structure 11 has little difference in mfe structure, but much more stable in humans
  - Increasing stability might fine-tune interactions and impact function

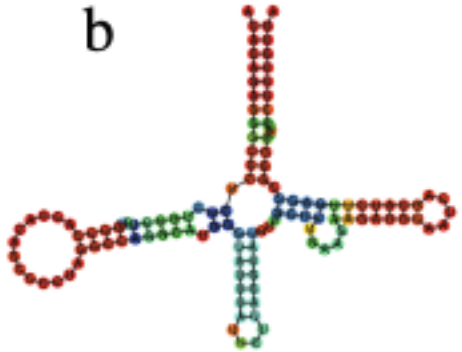
SIX-AS1 Analysis

Human



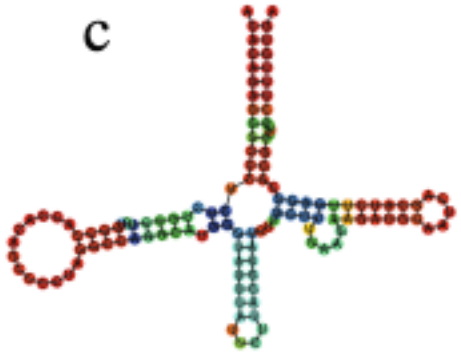
$s = 12.2$

Pan



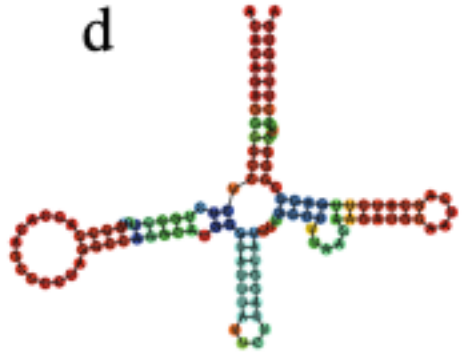
$s = 0.0$

Orangutan



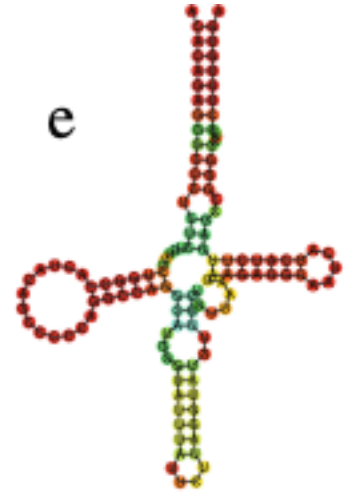
$s = 0.0$

Gorilla



$s = 0.0$

Macaque



$s = 0.0$

SIX3-AS1 Analysis

- Initially only had orthologs in human, pan, and orangutan
- Performed genome-wide scans using `Infernal v1.1.1` to find orthologs in gorilla and rhesus macaque
  - Built and calibrated a covariance model using human, pan, orangutan, and consensus structure
  - Searched for homologous structures in gorilla and macaque
    - Score of 155.1 and e-value of 1.5×10^{-31} for gorilla
    - Score of 150.7 and e-value of 1.7×10^{-30} for macaque
  - Similar structural pattern to pan and orangutan, less stable than humans

Other Positive Selection Candidates

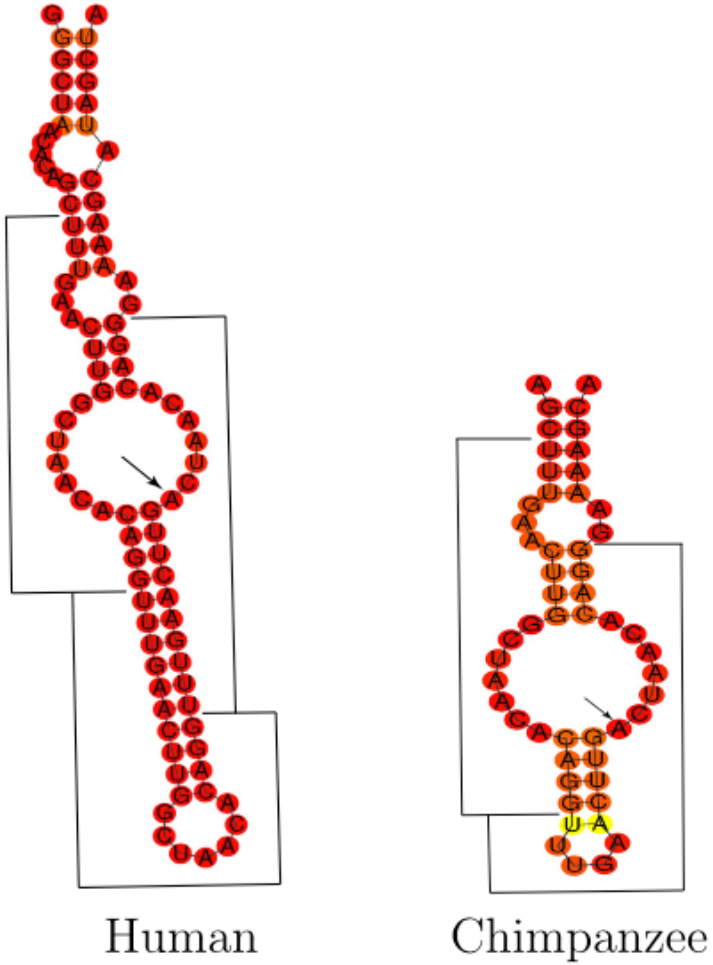
- Little/no functional annotation for most candidate lncRNAs
  - 49/110 lncRNA candidates have ENSEMBL Gene ID
  - 20/110 lncRNA candidates have HGNC gene symbol
- Tissue expression analysis for insight into function
  - 9 reported tissues: brain, cerebellum, liver, heart, kidney, placenta, ovary, testis, and stem cells
  - 6/110 lncRNAs expressed in all 9 tissues
  - 16/110 lncRNAs expressed in 1 tissue
  - 8/110 lncRNAs not detected as expressed
- Positively-selected lncRNAs tend to be expressed in more tissues than lncRNAs in general

Positively Selected lncRNAs and PDs

- Investigated link between positive-selection candidate lncRNA and psychiatric disorders (PDs)
- Used 26 lncRNAs reported to be involved in PDs
  - Filtered down to 32 local blocks as candidate lncRNAs under positive selection for secondary selection, 3 in humans
  - Manually inspected results
  - Updated thresholds to allow for candidates with SSS-score ≥ 4.5
    - Included another 11 local structures in humans

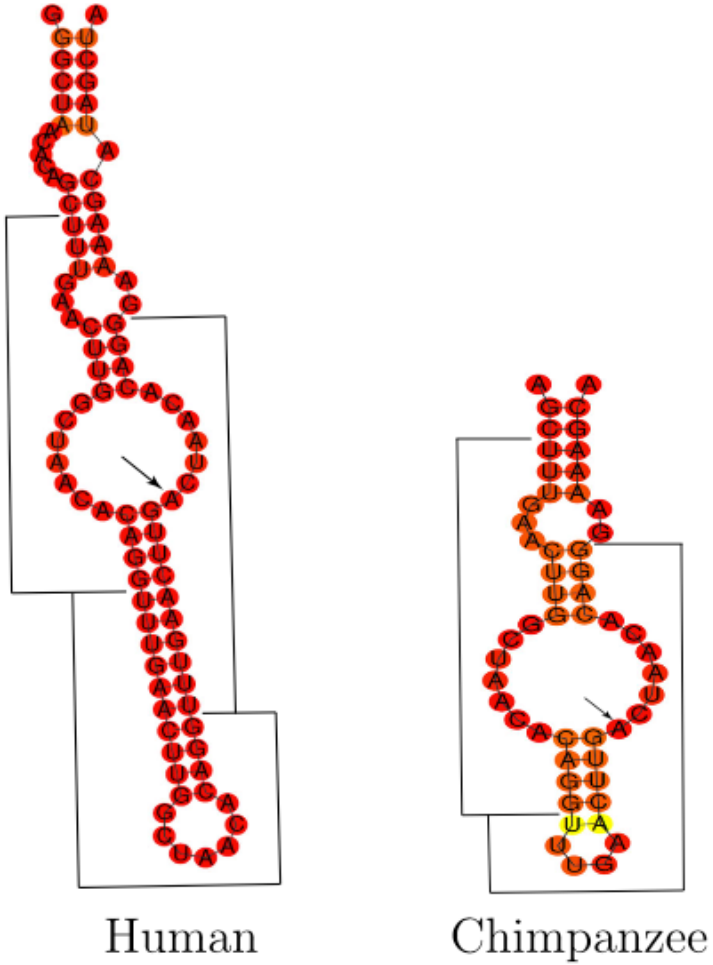
MIATsub92

- Highest selection score in humans (21.2)
- UACUAAC repeats with a substitution in one of duplications in human and chimpanzees
- Additional duplication in humans
- May have increased stability in humans relative to other primates



MIATsub92

- Repeats always in unpaired regions
- Selection seems to drive increased stability in humans while keeping UACUAAC unpaired
  - Implies importance of internal loops in recognition and binding of splicing factors
  - May cause some of the differences in splicing patterns between humans and other primates



Negative and Neutral Selection

- Have focused on detecting positive selection signals
- Extend to negative selection with SSS-score ≤ 2.0
  - Could complement other methods that assess structural conservation
- Identify relaxed selective constraints with high family divergence score

Pairwise SSS-test

- SSS-test requires 3+ orthologs
- Could extend to a pairwise version
  - Different interpretation of results
  - Unknown which sequence represents ancestral state
  - Divergent evolution vs. positive selection

Orthologs and Paralogs

- Gene duplication often but not always accompanied by positive selection
- Want to distinguish between (co)orthologs and paralogs
- Could report false positives if including paralogs by mistake
  - General concern for protein-coding genes and many ncRNAs
- Could apply pairwise SSS-test to duplicated ncRNAs to check for positive selection
- Short local duplications can also cause alignment errors
  - Should manually inspect alignments given to SSS-test

Areas for Future Work

- Find parameter with better theoretical foundation
  - SSS-score functions as a decision variable
  - Indel scoring model is very specific to SSS-score
  - Would likely take covariation of paired nucleotides into account

Discussion

- What impact does the choice to align $\mathcal{A}$ based on primary structure have on the secondary structure selection predictions?
- How could the model be changed to enable a cleaner interpretation?
- What experiments would have made the results more convincing?
- Is there a more robust or generalizable approach to the thresholds?
- How could the substitution scoring model include the relative likelihood of different SNPs?
- Could the SSS-test model be adjusted to handle compensatory mutations?