Research in Bioinformatics: Evolution and Adaptive Selection of ncRNAs

Pesquisa em Bioinformática: Evolução e Seleção Adaptativa de ncRNAs

Maria Beatriz Walter Costa

Institute of Laboratory Medicine,
University Hospital Leipzig,
Leipzig, Germany

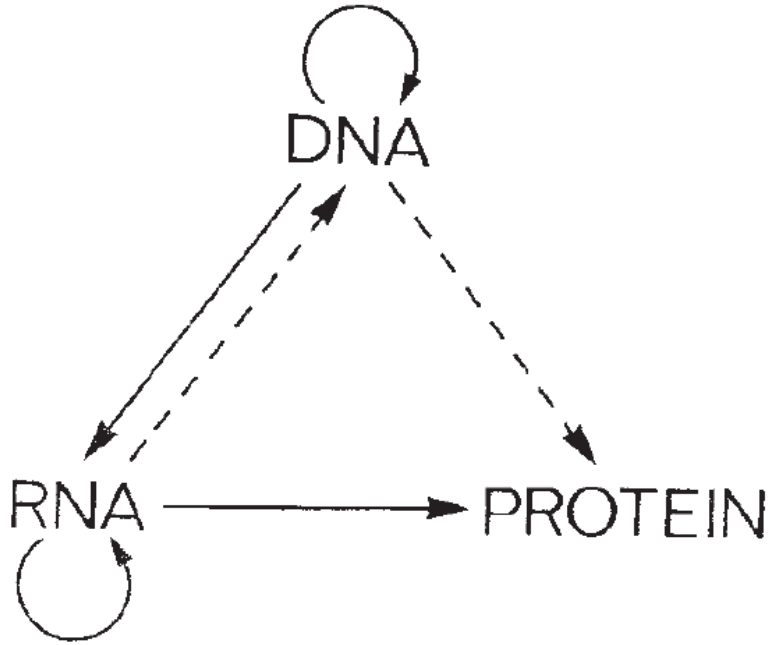
19th June 2020

Overview

Motivation

- What makes us human?
- Molecular basis of phenotypes
- Comparative genomics
- Biological molecules

⟨https://peppermintmag.com/from-the-archives-jane-goodall/⟩



Central Dogma of Molecular Biology, Francis Crick 1970

Motivation

- What makes us human?
- Molecular basis of phenotypes
- Comparative genomics
- Biological molecules

⟨https://peppermintmag.com/from-the-archives-jane-goodall/⟩

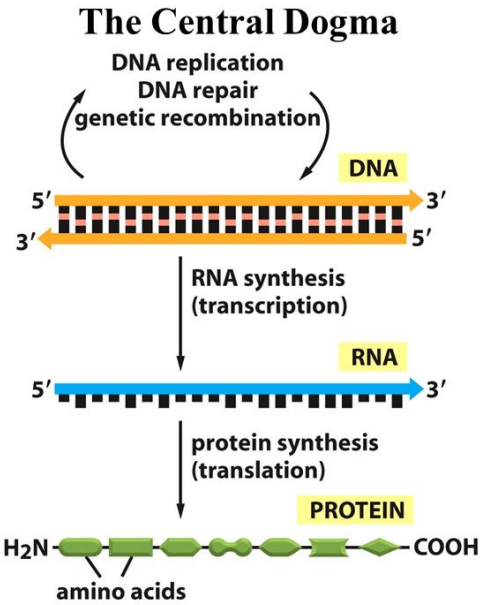


Figure 6-2

Molecular Biology of The Cell $4th$ ed., Alberts *et al.* 2002

Motivation

- What makes us human?
- Molecular basis of phenotypes
- Comparative genomics
- Biological molecules

⟨https://peppermintmag.com/from-the-archives-jane-goodall/⟩

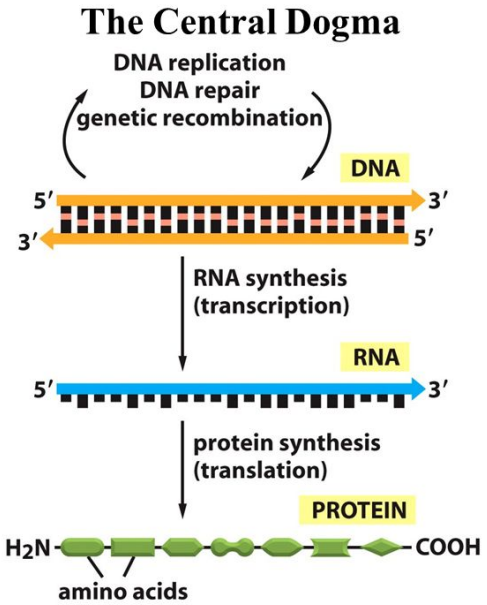
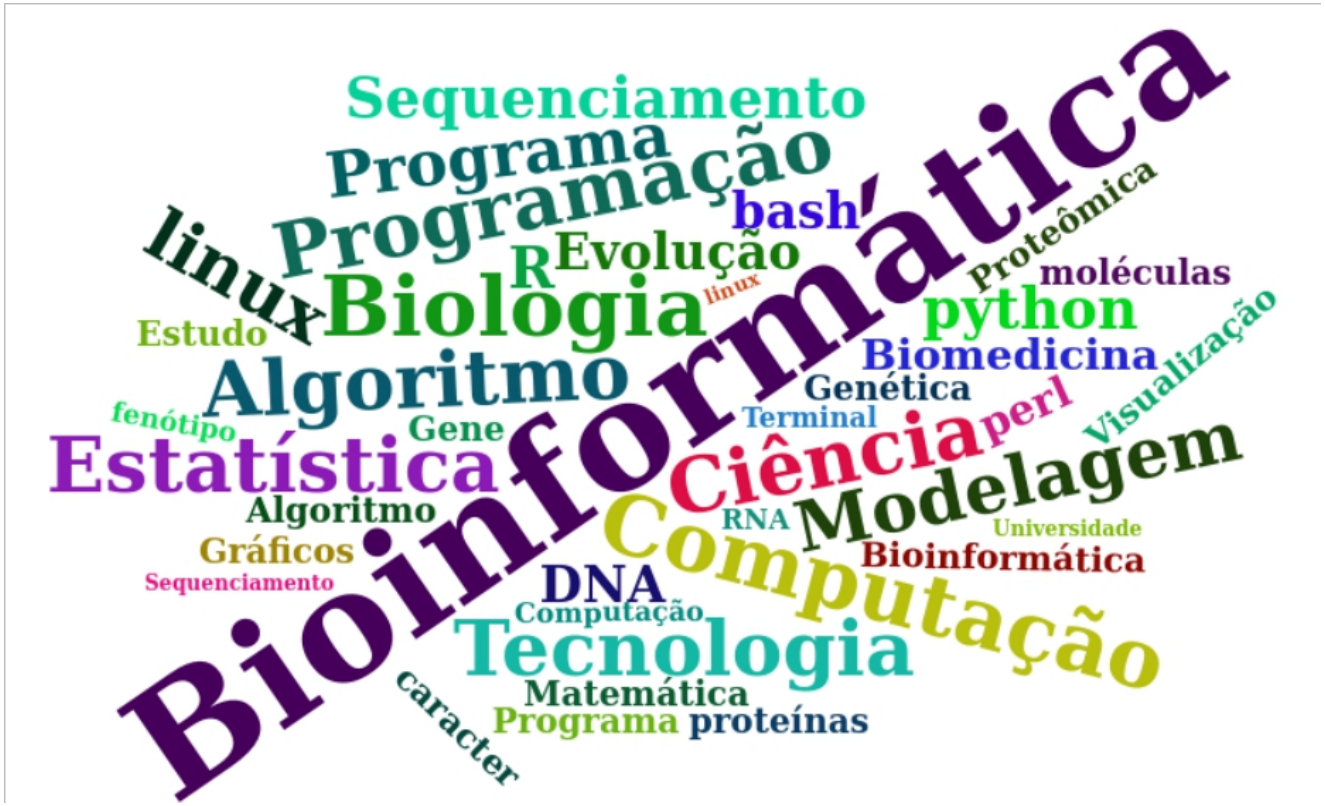


Figure 6-2

Alberts *et al* Molecular Biology of The Cell $4th$ ed. 2002

- Chimpanzee and human: different phenotypes/similar genomes
- Evolutionary approach: look for rare and interesting differences

Bioinformatics



Types of selective pressures

- **Negative selection** (purifying, conserved)
  - Eliminates variants
  - Conservation
- Neutral selection
  - Accepts some level of variants
- **Positive selection** (adaptive)
  - Selects new variants
  - Goes towards some new function
  - Indicates adaptive evolution

Detecting positive selection

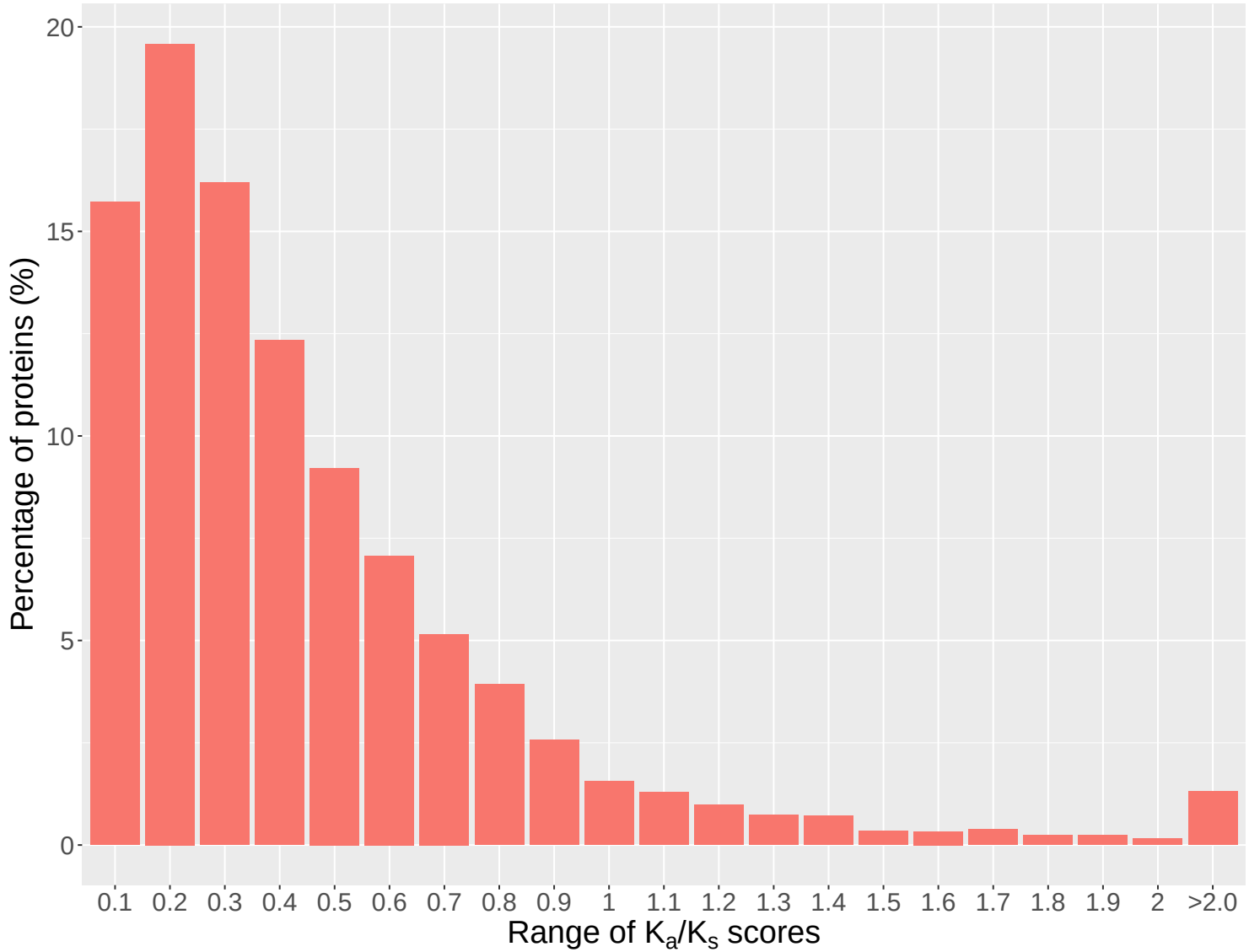
- Tests are available for proteins: K_a/K_s (d_N/d_S)
- Search for species specific functions
- Change rates are measured

$$K_a/K_s = \frac{\text{non synonymous changes/non synonymous sites}}{\text{synonymous changes/synonymous sites}}$$

- Synonymous changes do **not** change the protein
- Non-synonymous changes **do** change the protein
- Score $K_a/K_s > 1$ indication of positive selection
- Example: FOXP2 (Transcription Factor): language and speech[1]

1 Konopka & Roberts Cell 2016

Selection overview of human proteins

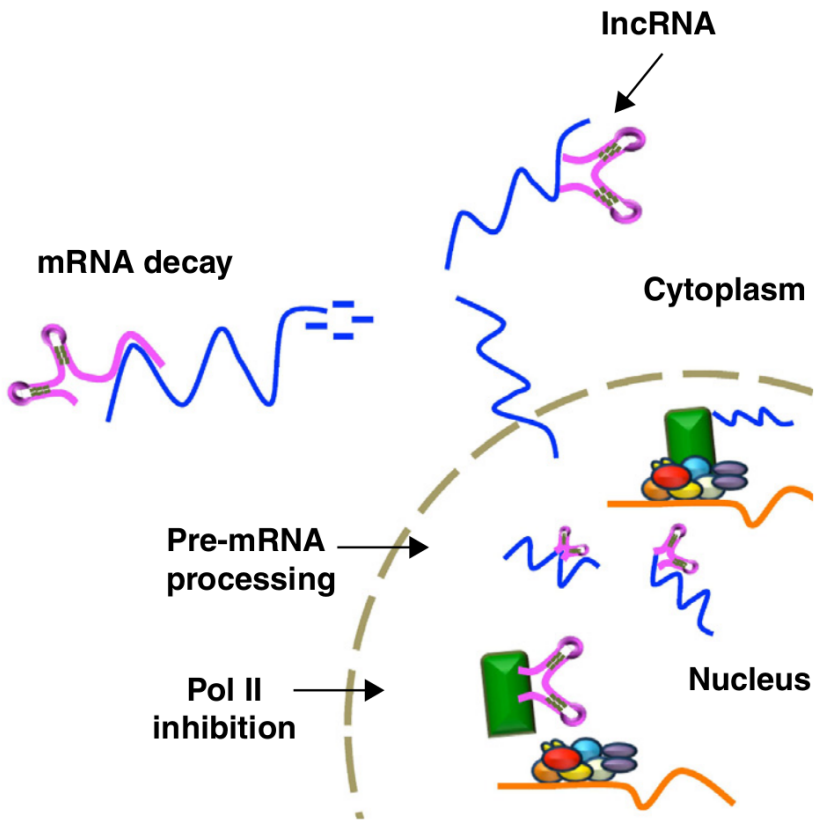


K_a and K_s data retrieved from ENSEMBL (BioMart), analysis by MBWC

Non-Coding RNAs

- Do not code for proteins
- Are functional

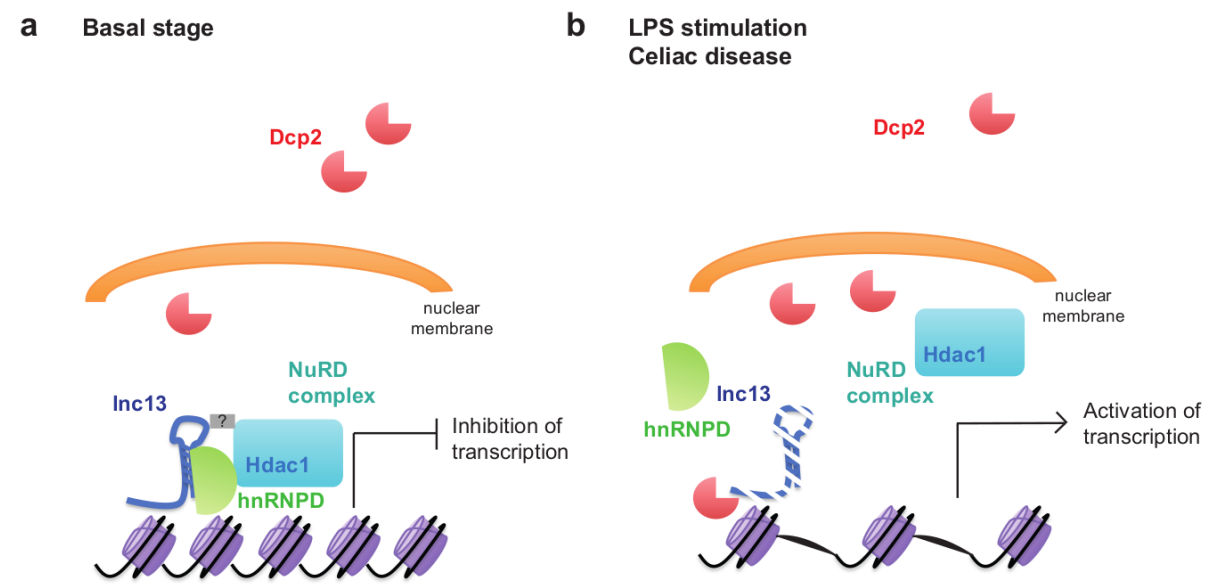
- Small ncRNAs
  - $l < 200\ nt$
  - Well characterized (tRNAs, microRNAs, snoRNAs)
- Long ncRNAs
  - $l > 200\ nt$
  - Play important roles in brain and all other tissues
  - Have various functions
    - Guides for complexes
    - Gene regulators



Perdomo-Sabogal *et al* Curr Opin Genet Dev 2014

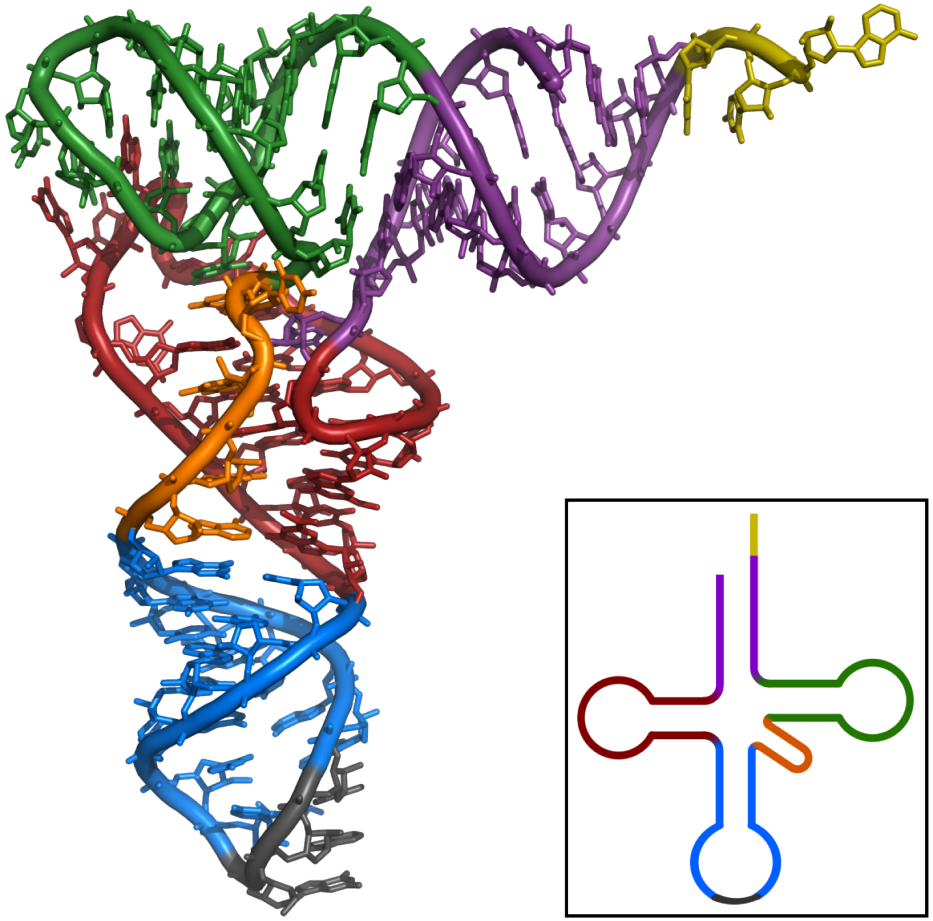
Long non-coding RNAs

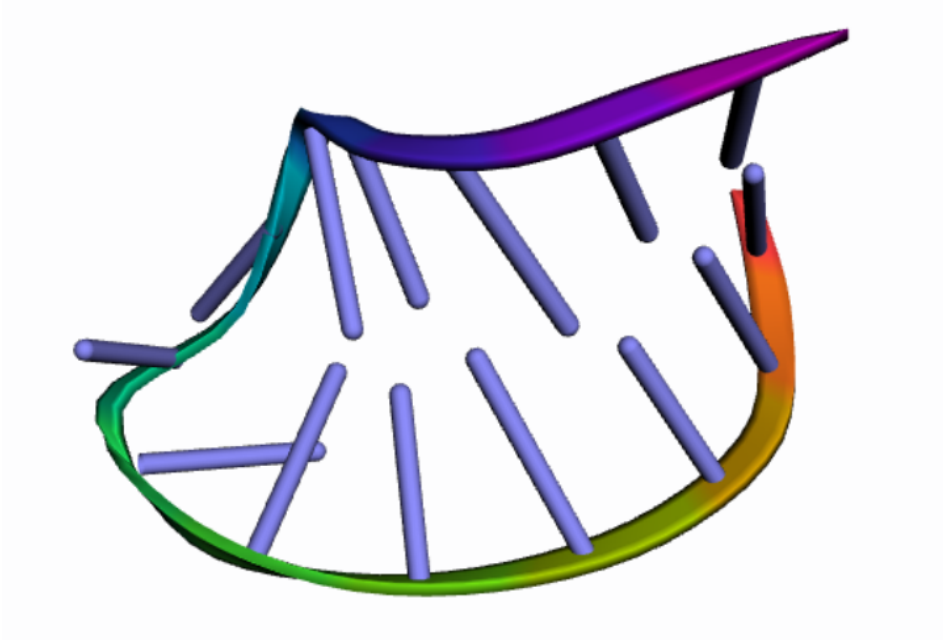
- Primary sequence poorly conserved
- Act through other mechanisms
- Are under negative selection (stretches of sequence/ structures, splice sites)
- Disease-associated SNPs can disrupt local structures (lnc13)



Castellanos-Rubio *et al* Science 2016

Structure defines function

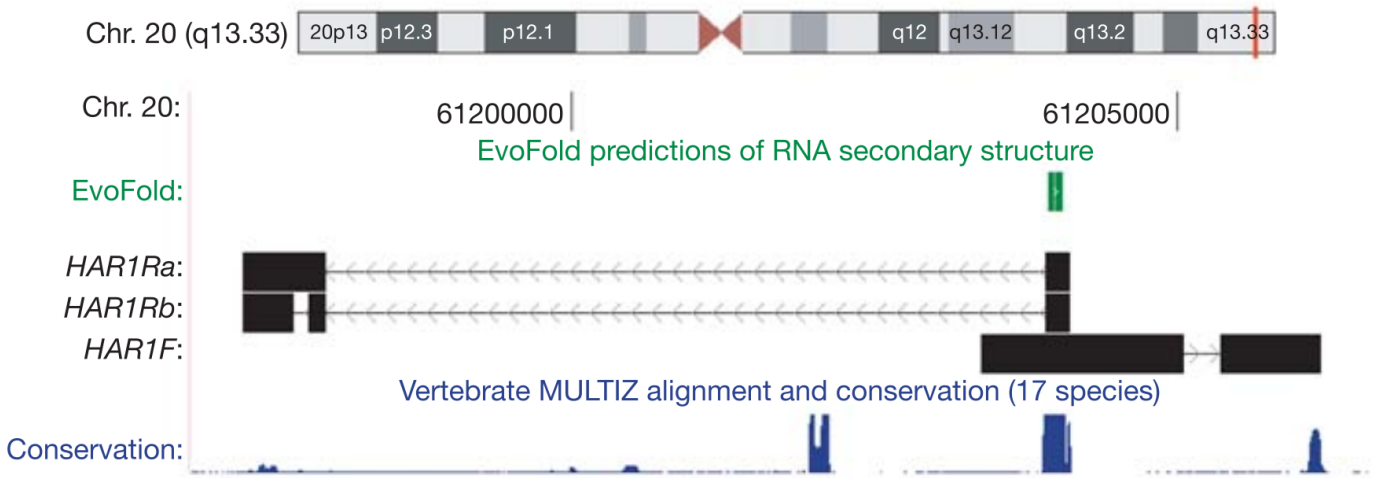




Rfam tRNA;
Duszczyk *et al* Biomol NMR Assignments 2012 XIST local structure

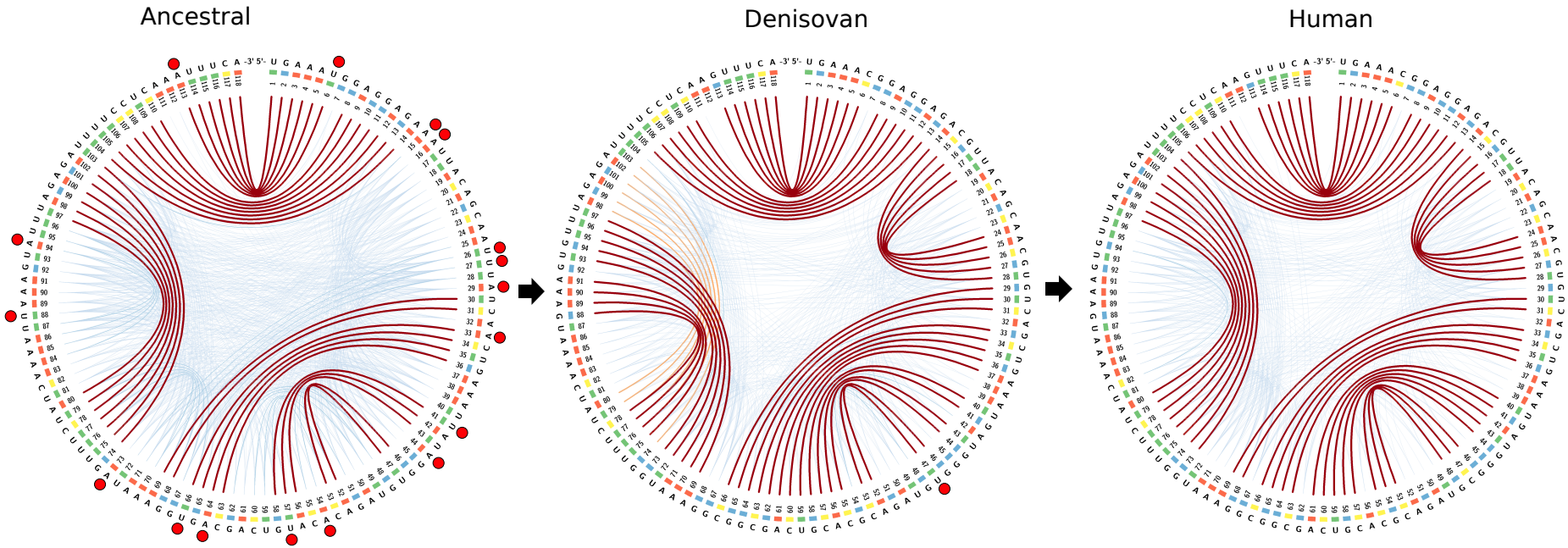
What was known about adaptive selection in ncRNAs?

- Human Accelerated Regions with many human-specific mutations
- HAR1: 118 nt and 18 human specific changes is a candidate for positive selection



Pollard *et al* Nature 2006

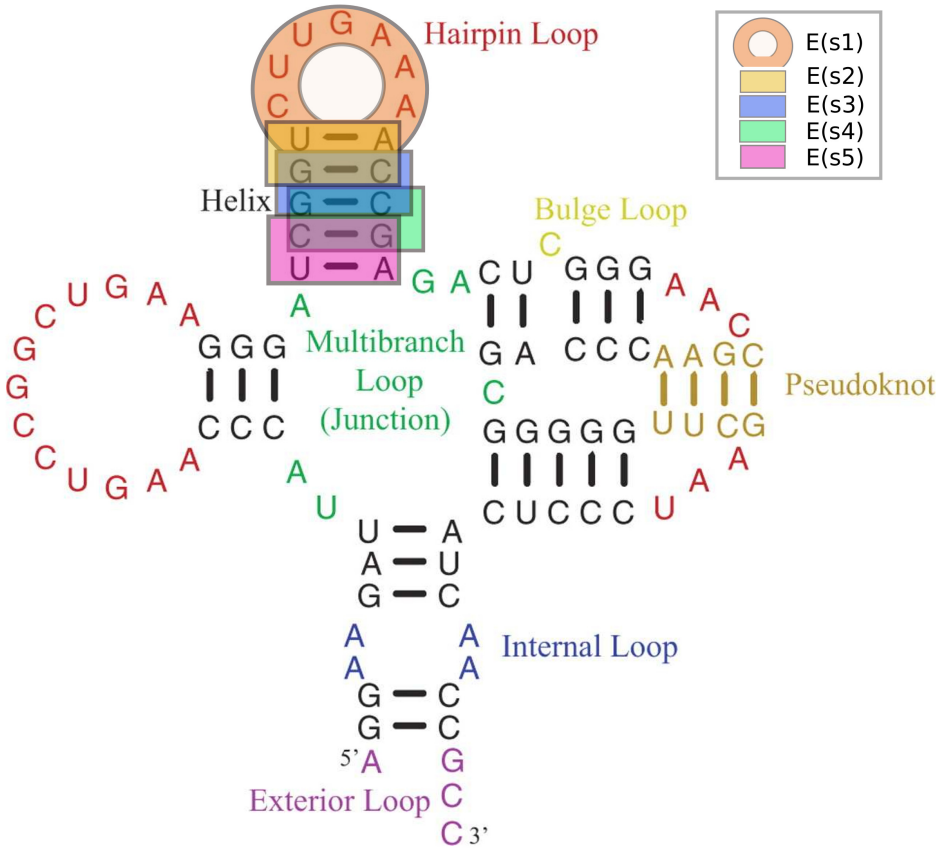
HAR1 is positively selected in humans



Walter Costa *et al.* Journal of Theoretical Biology 2018 - Dynamic Programming: `mutationOrder`
Léger *et al.* BMC Bioinformatics 2019 - Visualization: `CS2BP2-Plot`

Prediction of secondary structures of RNAs

- Free energy minimization (optimal combination of bps)
  - Thermodynamics: free energy parameters - RNA melting (**RNAfold**)
  - Statistics: train param. by well-curated DBs (Infernal)
- MFE structure: most energetic
- Centroid structure
  - Contains base pairs that are more likely to be present in the structural ensemble

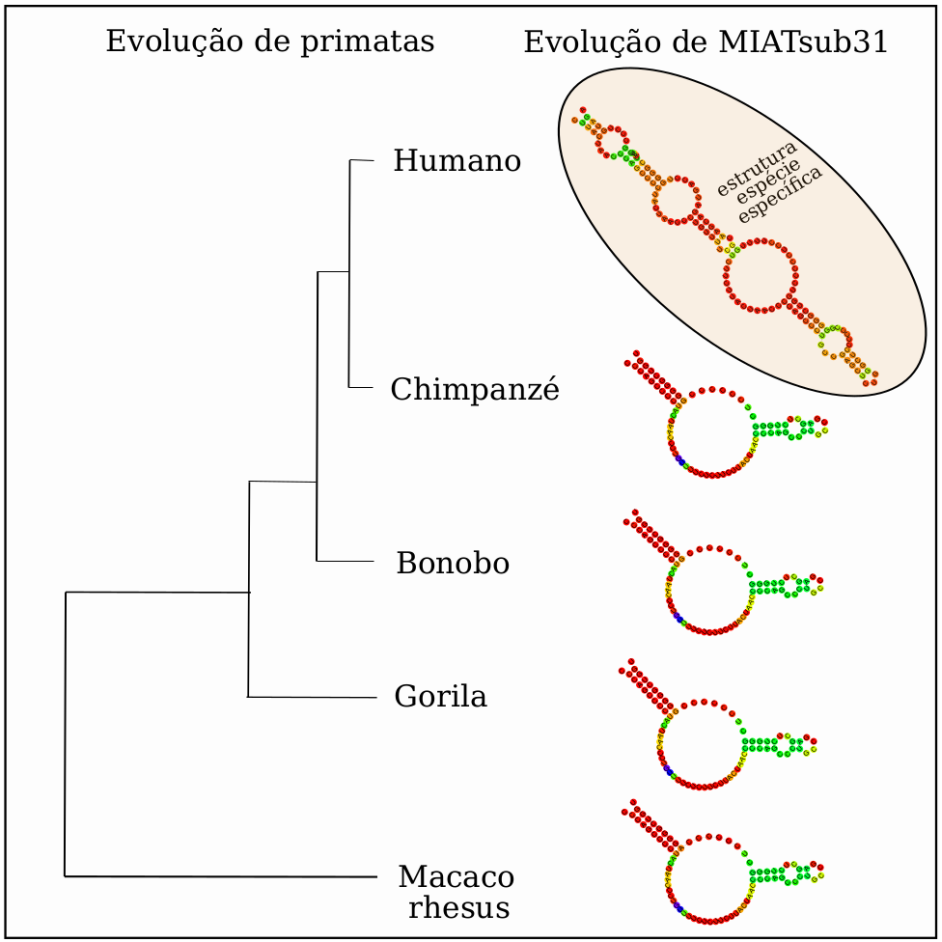


Adapted from Turner & Matthews Nuc Acids Res 2009

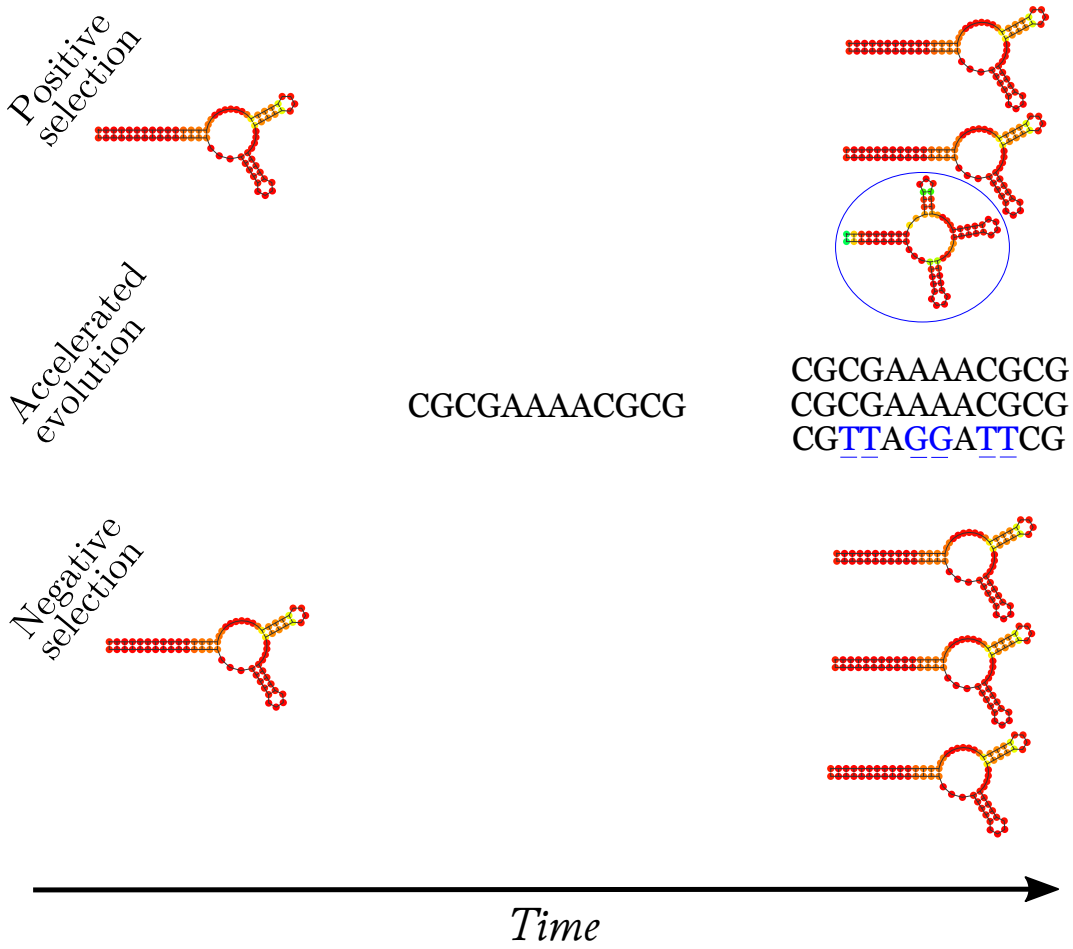
Studying selection of ncRNAs

- Focus on structural conservation
- Conservation indicates functional RNAs
- Find conserved set $\rightarrow$ functional group
- Software can detect conserved RNA structures: RNAz, CMfinder, RNAalifold...
- Spot different structure $\rightarrow$ specialized functionality
- **How to find structures that changed in only one lineage in a conserved set?**
- Understand the role of lncRNAs in the evolution of humans

Phylogeny versus ncRNA evolution



ncRNA evolution



Nowick *et al.* Evolutionary Bioinformatics 2019

ncRNA evolution

Selective pressure	Method	Level of analysis
Positive selection	SSS-test[1]	secondary structure
Accelerated evolution	Pollard *et al*[2]	primary sequence
Negative selection	R-scape[3], RNAz[4] cmfinder[5],qrna[6] AlifoldZ[7], EvoFold[8] SISSIz[9], SSS-test[1]	secondary structure

1 Walter Costa *et al* 2019, 2 Pollard *et al* 2016,

3 Rivas *et al* 2017, 4 Washietl *et al* 2005,

5 Yao *et al* 2006, 6 Rivas *et al* 2001,

7 Washietl *et al* 2004, 8 Pedersen *et al* 2006,

9 Gesell *et al* 2008.

Nowick *et al.* Evolutionary Bioinformatics 2019

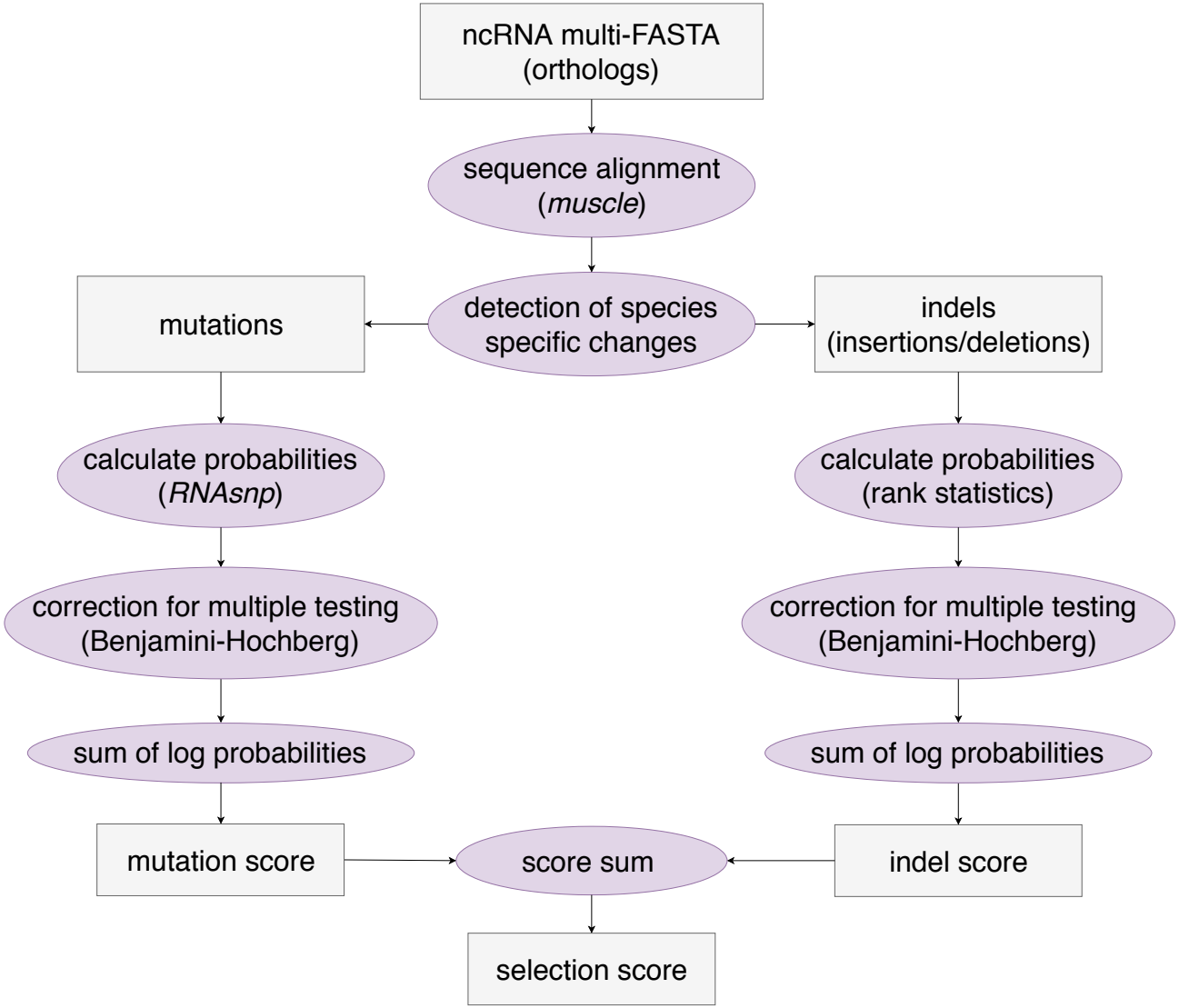
SSS-test: Structural Selection of ncRNAs

- Novel method for detecting positive selection in ncRNAs
- Discussion of the SSS-test: **S**election on the **S**econdary **S**tructure test
- Applications

Methods for detecting positive selection

- Simple counting
  - Analogous to the K_a/K_s test
  - Based on the binary distinction between synonymous and non-synonymous changes
- Statistical modelling
  - Calculates the probability of an event using the Poisson model
  - Also depends on the binary distinction between synonymous and non-synonymous changes
- *Combining probabilities of structural change: SSS-test*
  - Does not make a binary distinction of changing/not changing
  - Considers the probabilities of:
    - Mutations to impact the structure, as given by a tool (RNAsnp)
    - Indels impacting the structure, as given by rank statistics

SSS-test: Selection on the Secondary Structure test



Selection scores

- SSS-score(x) combines scores of mutations ($s(x)$) and indels ($s'(x)$):
  - SSS-score(s) $= 2s(x) + s'(x)$
- Low SSS-score indicates negative selection
- High SSS-score indicates positive selection
- Mutations: probabilities (p-values) are calculated with RNAsnp
  - Uses RNAfold to predict secondary structures
  - Uses pre-computed tables of ncRNAs with distributions of SNP effects
  - Outputs p-values related to the structural impact of the mutations

Correction for multiple testing

- Correction for multiple testing using Benjamini-Hochberg method
- Let $p = p_1 \geq p_2 \geq \cdots \geq p_n$ be the collection of p-values
- Update the corrected set of p-values $\widetilde{p}$ with:

$$\widetilde{p}_1 = \min\{1,\ p_1\}$$

$$\widetilde{p}_i = \min\left\{1,\ \widetilde{p}_{i-1},\ \frac{n}{(n-i+1)}p_i\right\}$$

Use $\widetilde{p}$ to produce the substitution score:

$$s(x) = -\sum_i \log \widetilde{p}_i\ .$$

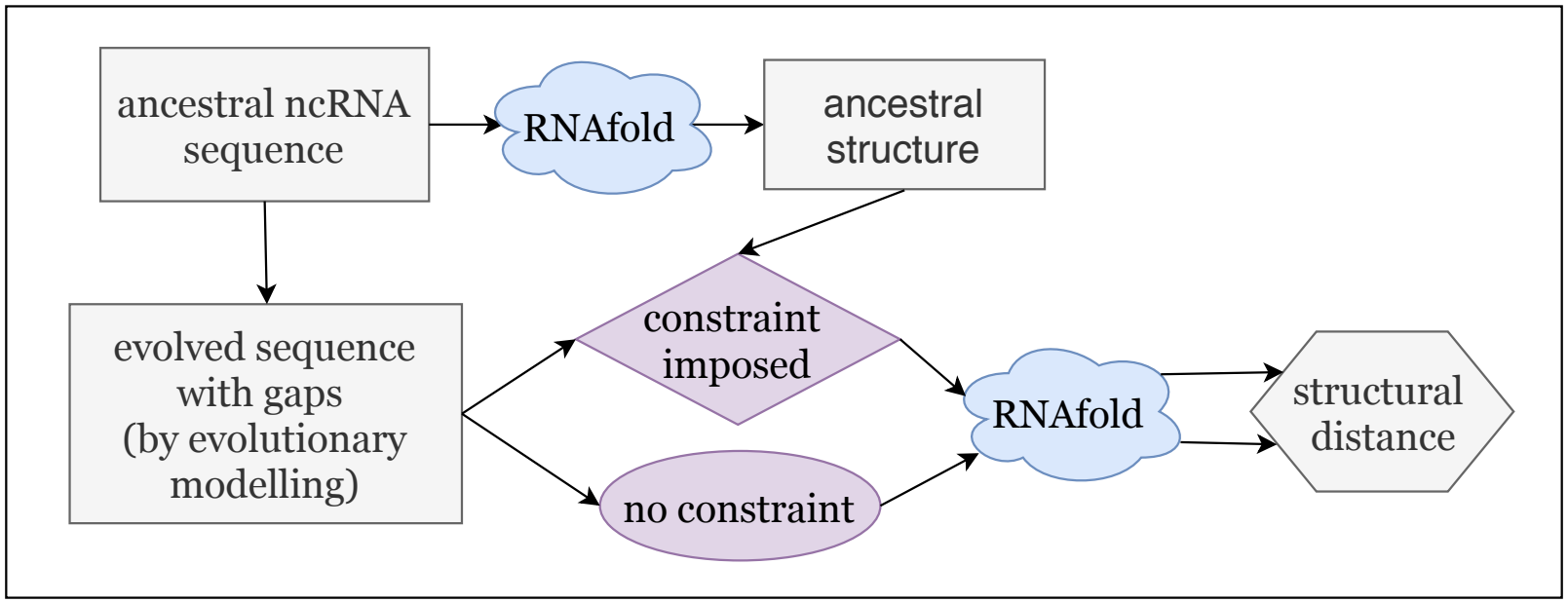
s: mutation selection score

x: sequence

Scoring structural impact of indels by rank statistics

- Assigning probabilities with rank statistics
- Correction for multiple testing using Benjamini-Hochberg method
- Use the $\widetilde{p}$ to produce the indel score

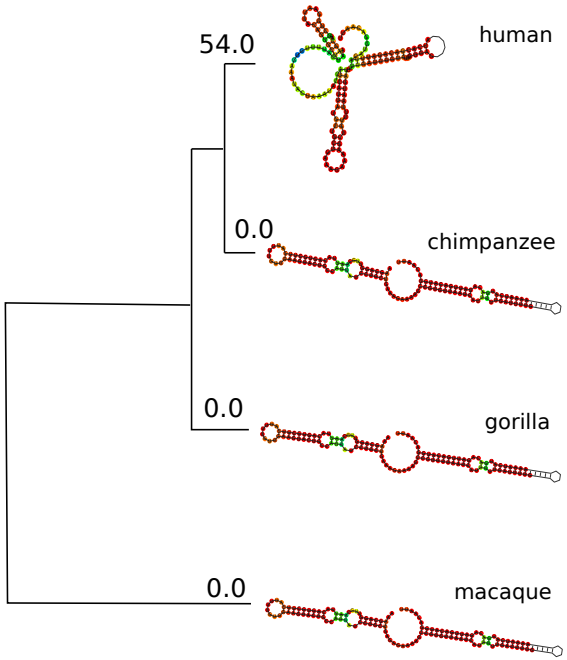
$$s'(x) = -\sum_i \log \widetilde{p}_i \ .$$



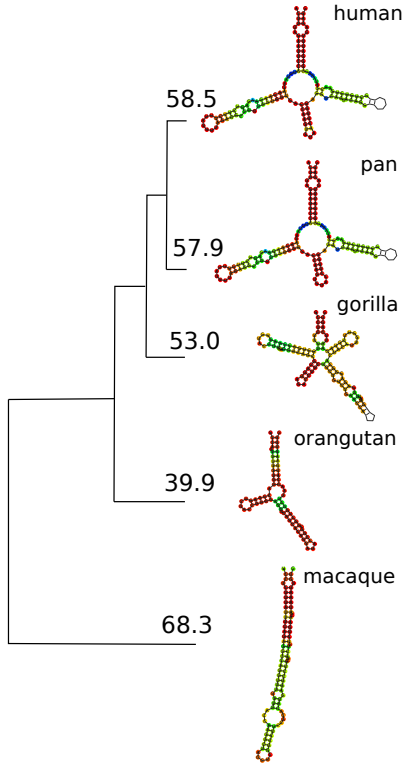
Family divergence (d score): $d = \text{median}_s\ d_s$

$$d_s = 100 \frac{\overbrace{\sum_{ij \in W_s} |A_{s,ij} - B_{s,ij}|}^{\text{shared bp difference}} + \overbrace{\sum_{ij \in X_s} A_{s,ij}}^{\text{unique bp species}} + \overbrace{\sum_{ij \in Y_s} B_{s,ij}}^{\text{unique bp consensus}}}{\text{length}(\mathcal{A})}$$

Uniform family ($d \leq 10.0$)



Diverged family ($d > 10.0$)

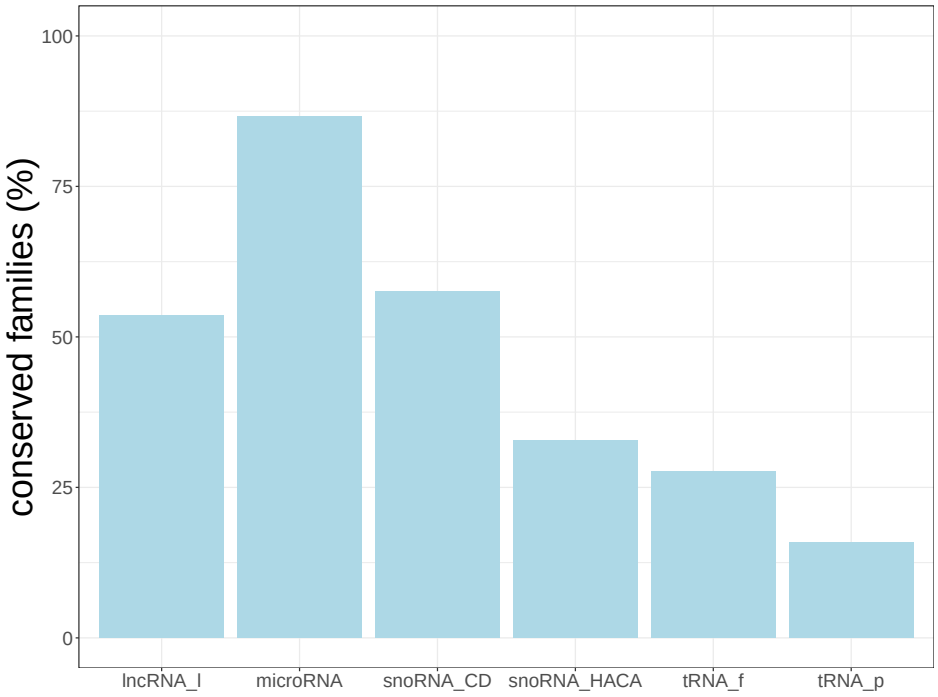


Benchmark and application of the SSS-test

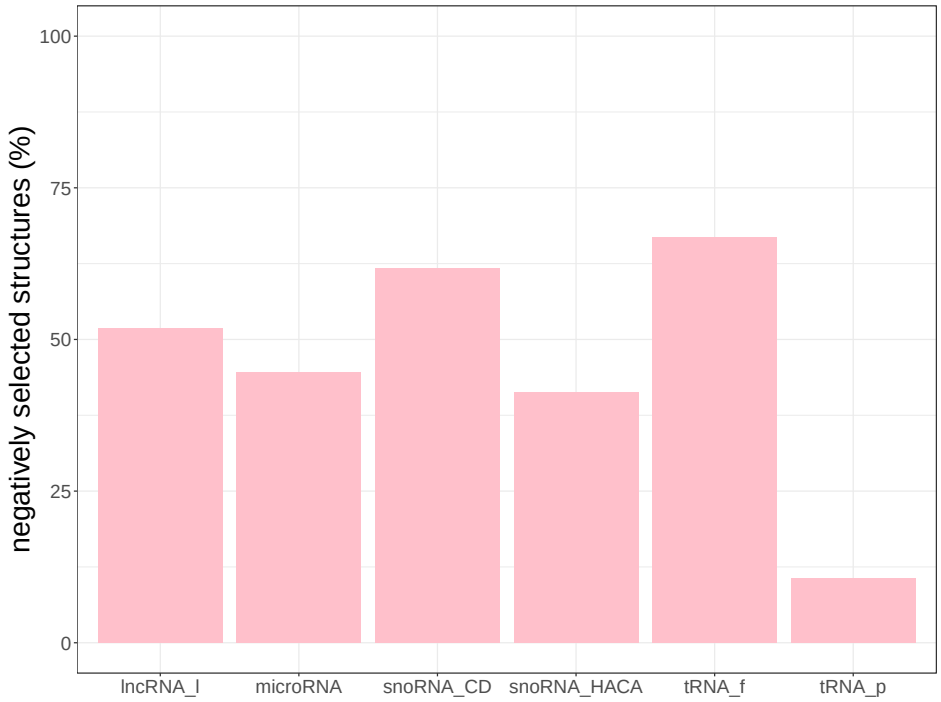
- Species: human, chimpanzee, gorilla, orangutan, rhesus macaque
- Benchmark
  - Small ncRNA DBs (negative control)
    - MicroRNAs (167 families)
    - Small nucleolar RNAs (170 families)
    - tRNAs (611 families)
  - HAR1 (positive control)
  - Synthetic DB models (*in silico* evolution)
    - Uniform family (100 families)
    - Diverged family (100 families)
    - Negative selection (100 families)
    - Positive selection(100 families)
- Application: lncRNA database of Necsulea *et al* Nature 2014
  - Objective: look for candidates of positive selection in humans
  - 15,443 families of lncRNAs

Benchmark with small ncRNA DBs

Uniform families ($d \leq 10.0$)

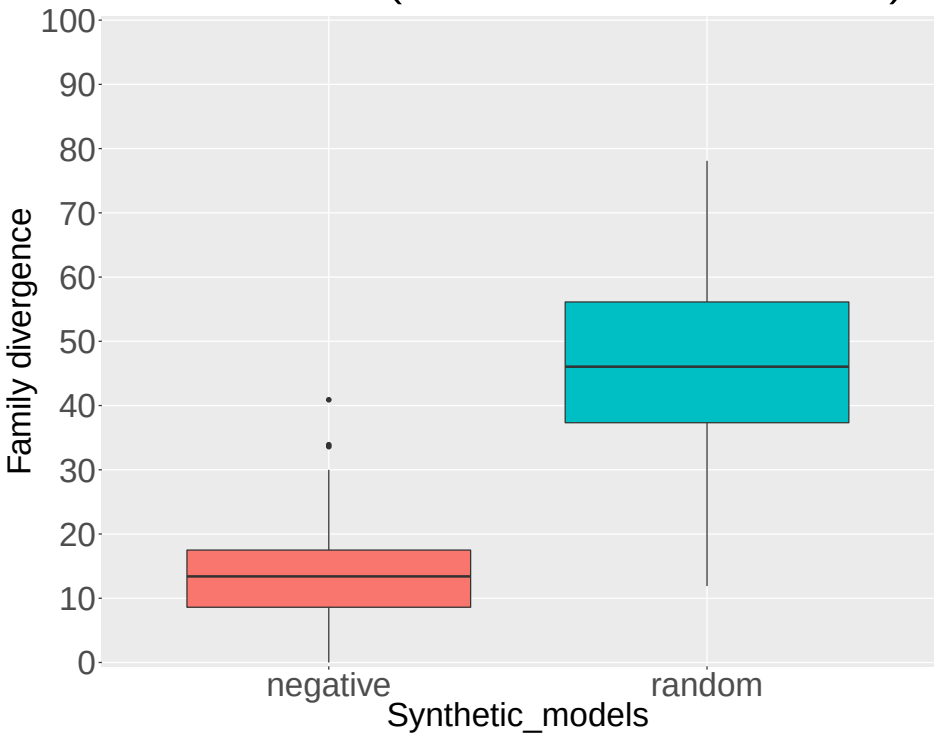


Strong negative selection ($s = 0.0$)



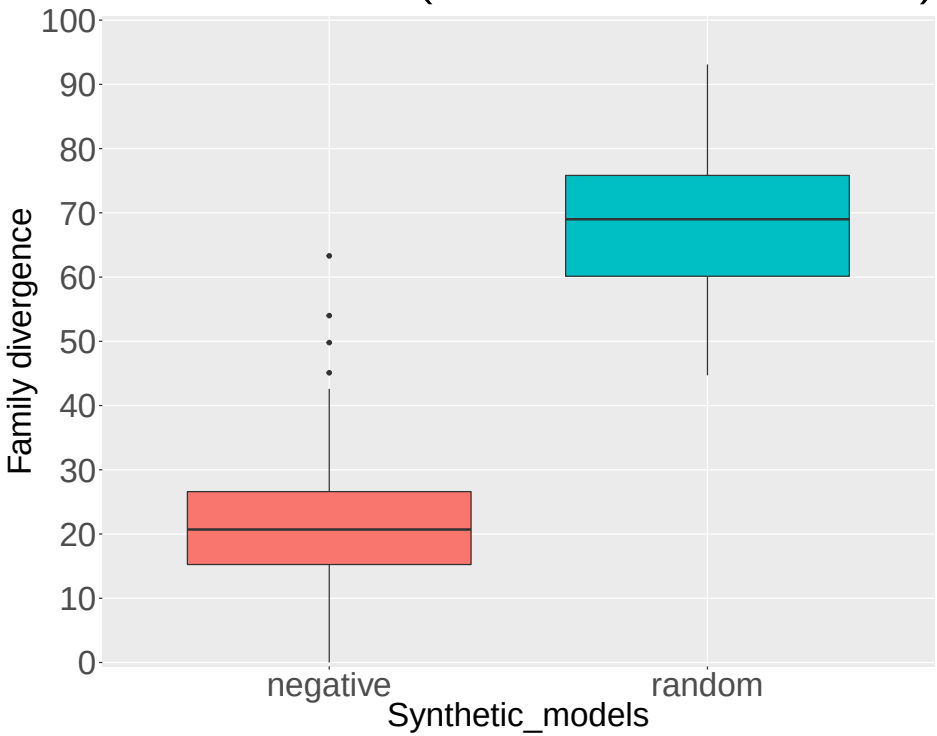
Benchmark with synthetic DB models (family divergence)

5 nt difference (ancestral → extant)



$$f_{\text{neg}}(a,m) = 1\,000\,(\Delta_{\text{centroid}}(a,m) + \varepsilon(a,m))$$

10 nt difference (ancestral → extant)

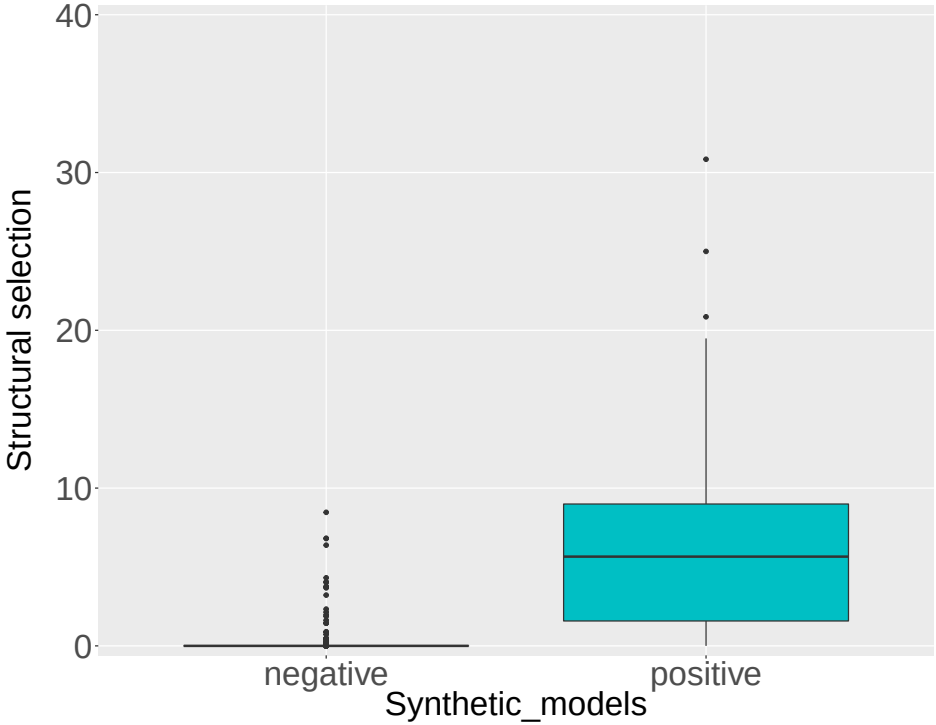


$$f_{\text{rand}} = 0$$

$$\varepsilon(a,m) = \max\left(0\ ,\ \text{mfe}(m) - \frac{\text{mfe}(a)}{2}\right)$$

Benchmark with synthetic DB models (structural selection)

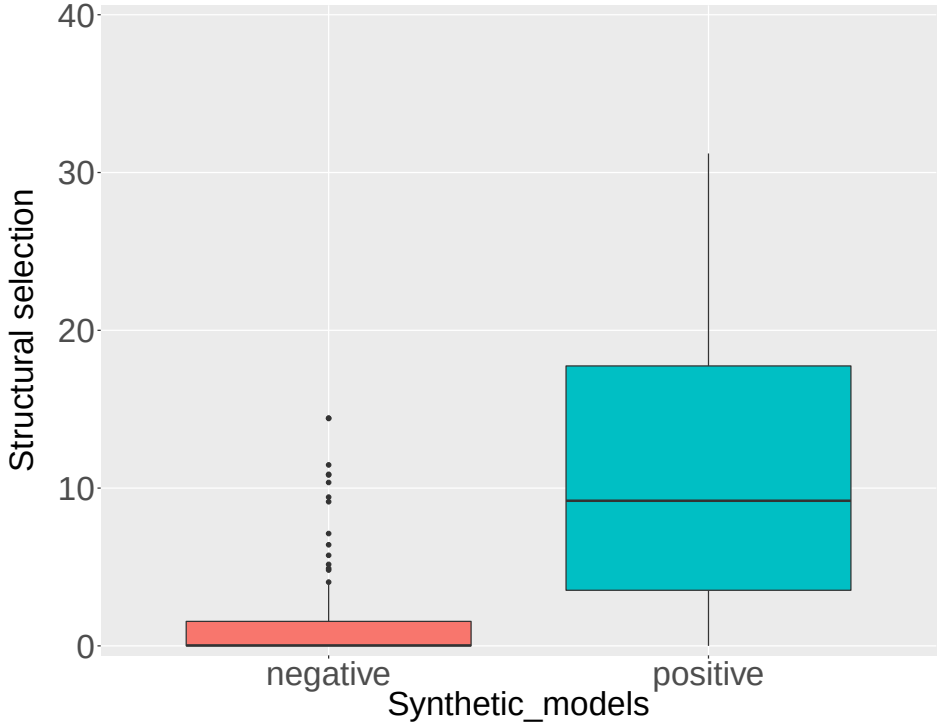
5 nt difference (ancestral $\rightarrow$ extant)



$$f_{\text{neg}}(a, m) = 1\,000\,(\Delta_{\text{centroid}}(a, m) + \varepsilon(a, m))$$

$$\varepsilon(a, m) = \max\left(0\ ,\ \text{mfe}(m) - \frac{\text{mfe}(a)}{2}\right)$$

10 nt difference (ancestral $\rightarrow$ extant)



$$f_{\text{pos}} = \text{gibbs}(m) + 50\Delta_{\text{shape:5}}(\texttt{[[][][]]}, m) + 1\,000\ \varepsilon(a, m)$$

Application: primate lncRNAs have well conserved local structures

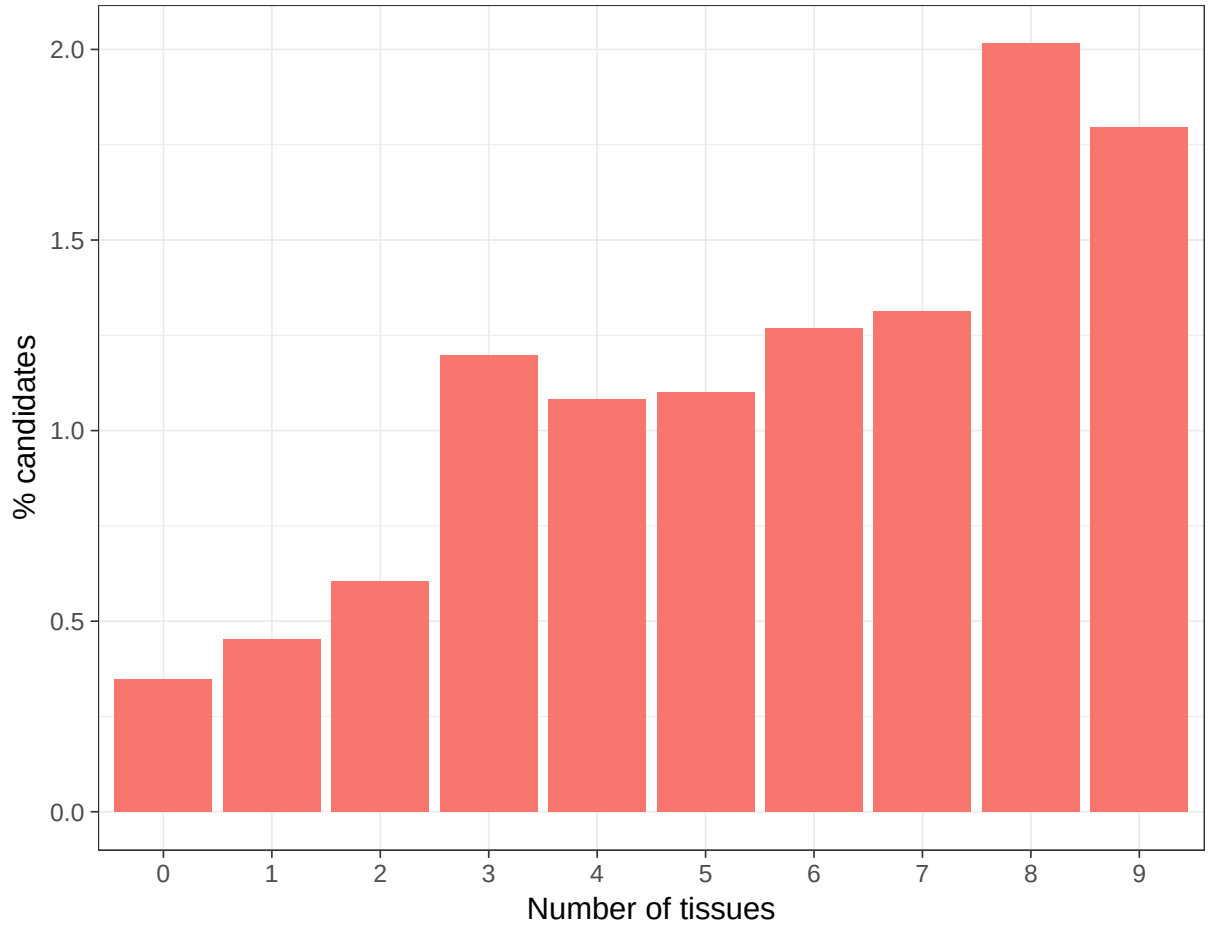
Table: Structural selection of local structures of lncRNAs. Only uniform families were considered.

Species	Representatives (local structures)	Negative ($s \leq 2$)	Positive ($s \geq 10$)
Human	8,934	8,179 (91.6%)	111 (1.2%)
Pan	8,736	7,997 (91.5%)	90 (1.0%)
Gorilla	8,080	7,199 (89.1%)	136 (1.7%)
Orangutan	6,435	4,802 (74.6%)	315 (4.9%)
Macaque	5,113	2,659 (52.0%)	738 (14.4%)

- 15,443 families of lncRNAs (Necsulea *et al* Nature 2014)

Profile of human structures with signs of positive selection

- 110 candidates for positive selection (111 local structures)
- Candidates seem to be expressed in multiple tissues



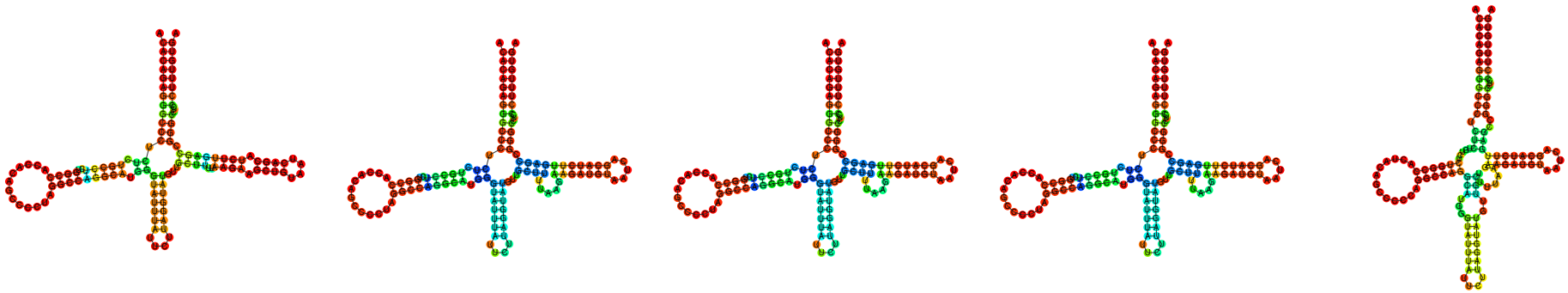
Expression data from Necsulea *et al* Nature 2014

Profile of humans structures with signs of positive selection

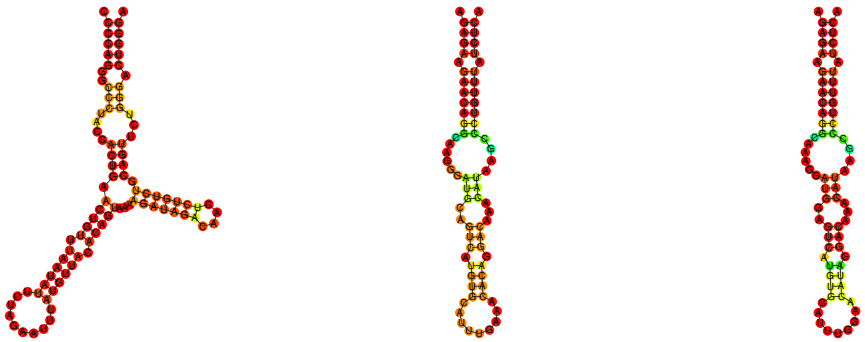
- Only 20 have HGNC names (at least some characterization)
  - SIX3: transcription regulator, associated with cephalic disorder

SIX3-AS1 →

Human Pan Orangutan Gorilla Macaque



Human Gorilla Macaque ← *LINC02217sub5*



Summary: SSS-test

- First method for detecting positive selection in ncRNA structures: SSS-test
- Benchmark with biological and synthetic data
- Application to a primate DB of lncRNAs
- We suggest 110 candidates for having evolved under positive selection in humans
- Candidates tend to be expressed in a higher number of tissues
- We suggest 20 of these candidates for wet lab validation

Summary of presentation

- Biological question: molecular basis of what makes us human
- Adaptive selection of structures of ncRNAs is largely unkown
- Bioinformatics approach based on thermodynamics and RNA folding
- `SSS-test`: first method for testing positive selection of ncRNA structures
- Benchmark and application on primates: 15k lncRNA families
- 110 candidates of human lncRNAs with local structures under positive selection
- Contribution of novel methods and algorithms and new insights into the evolution of ncRNAs structures

Articles

- Nowick K, **Walter Costa MB**, Höner zu Siederdissen C, Stadler PF. *Selection Pressures on RNA Sequences and Structures*, Evolutionary Bioinformatics, 2019
- **Walter Costa MB**, Höner zu Siederdissen C, Tulpan D, Stadler PF, Nowick K. *A novel test for detecting selection on the secondary structures of non-coding RNAs*, BMC Bioinformatics, 2019
- Kolora SRR, Weigert A, Saffari A, Kehr S, **Walter Costa MB**, Spröer C, Indrischek H, Chintalapati M, Doose G, Bunk B, Overmann J, Lohse K, Bleidorn C, Henle K, Nowick K, Faria RM, Stadler PF, Schlegel M. *Divergent evolution in the genomes of closely-related lacertids,* Lacerta viridis *and* L. bilineata *and implications for speciation*, GigaScience, 2018

Articles

- **Walter Costa MB**, Höner Zu Siederdissen C, Tulpan D, Stadler PF, Nowick K. *Temporal ordering of substitutions in RNA evolution: Uncovering the structural evolution of theHuman Accelerated Region 1*. Journal of Theoretical Biology, 2018
- Léger S, **Walter Costa MB**, McComb S, Tulpan D. *Pairwise Visual Comparison of Small RNA Secondary Structures with Base Pair Probabilities*, to be submitted
- Perdomo-Sabogal A, Kanton S, **Walter MB**, Nowick K. *The role of gene regulatory factors in the evolutionary history of humans*. Opinion in Genetics and Development, 2014

- Apresentação LGBio: ⟨https://waltercostamb.github.io/blog/2020/06/19/LiveLGBio⟩

Software and PhD thesis

- `SSS`: ⟨https://github.com/waltercostamb/SSS-test⟩
- `mutationOrder`
  - ⟨http://hackage.haskell.org/package/MutationOrder⟩
  - Pre-compiled binaries: ⟨https://github.com/choener/MutationOrder/releases⟩
- `buildOrthologs`: ⟨https://github.com/waltercostamb/lncRNA-ortholog-reconstruction⟩
- `retrieve-fasta`: ⟨https://github.com/waltercostamb/lncRNA-ortholog-reconstruction⟩
- CS^2-UPlot (web tool): ⟨https://nrcmonsrv01.nrc.ca/cs2bp2plot/⟩

- PhD thesis: *Adaptive Evolution of Long Non-Coding RNAs* ⟨https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-323898⟩

Peter Stadler, Katja Nowick, Christian Höner zu Siederdissen, Rohit Kolora, Dan Tulpan, CNPq (Science without Borders/Brazil) and DFG 1738

CIÊNCIA SEM FRONTEIRAS

UNIVERSITÄT LEIPZIG

SPP 1738