Algorithms in non-coding RNA structure evolution

Maria Beatriz Walter Costa

Bioinformatik
Universität Leipzig

33^{rd} TBI Winterseminar

overview

central dogma of molecular biology

NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

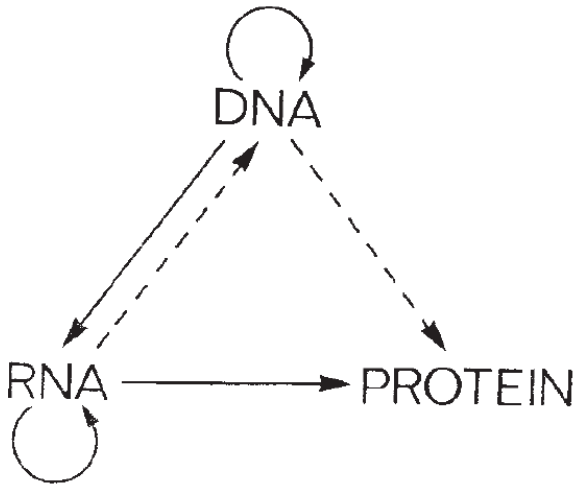


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

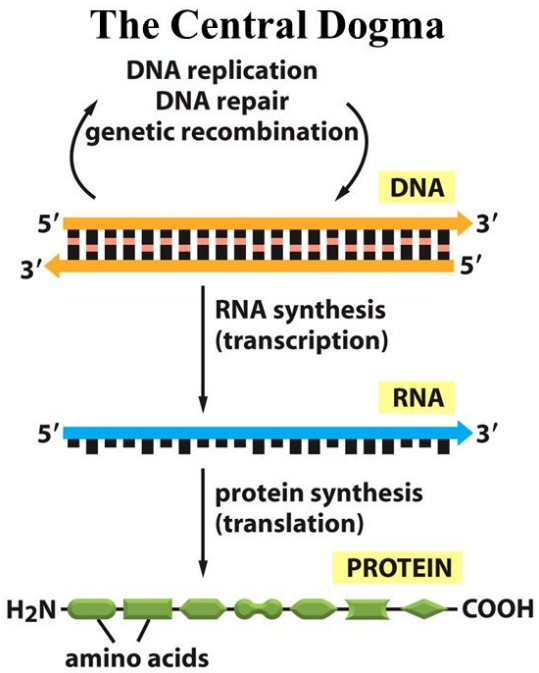


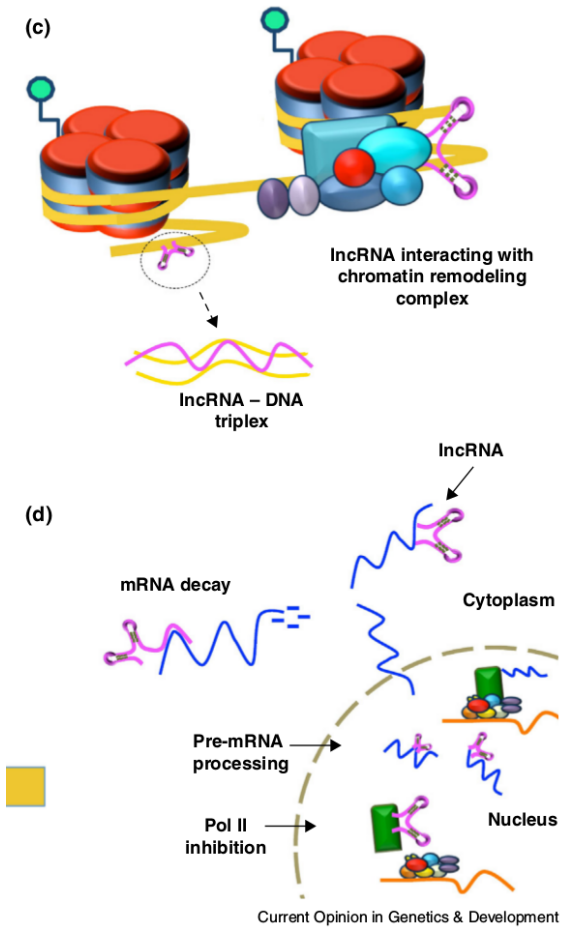
Figure 6-2

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

non-coding RNAs

- do not code for proteins
- are functional

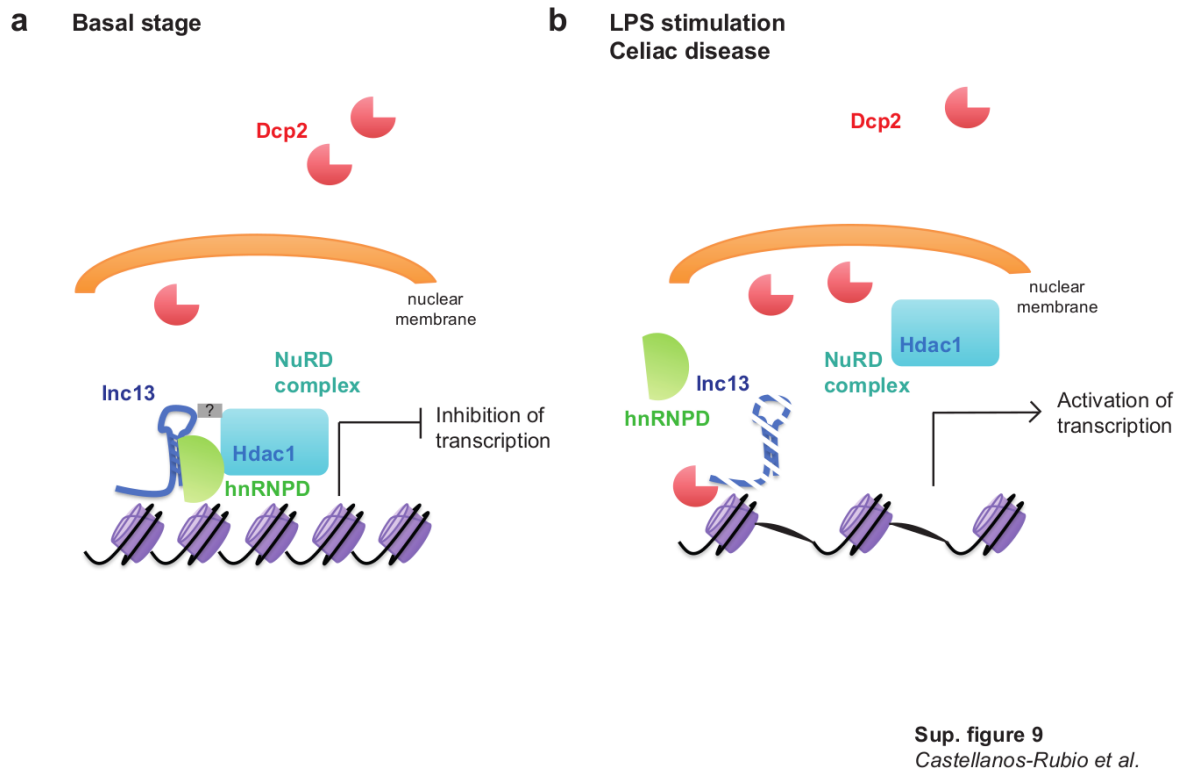
- small ncRNAs
  - $l < 200\ nt$
  - are well characterized (tRNAs, microRNAs, snoRNAs)
- long ncRNAs
  - $l > 200\ nt$
  - Play important roles in brain and all other tissues
  - Have various functions
    - guides for complexes
    - gene regulators



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

long non-coding RNAs

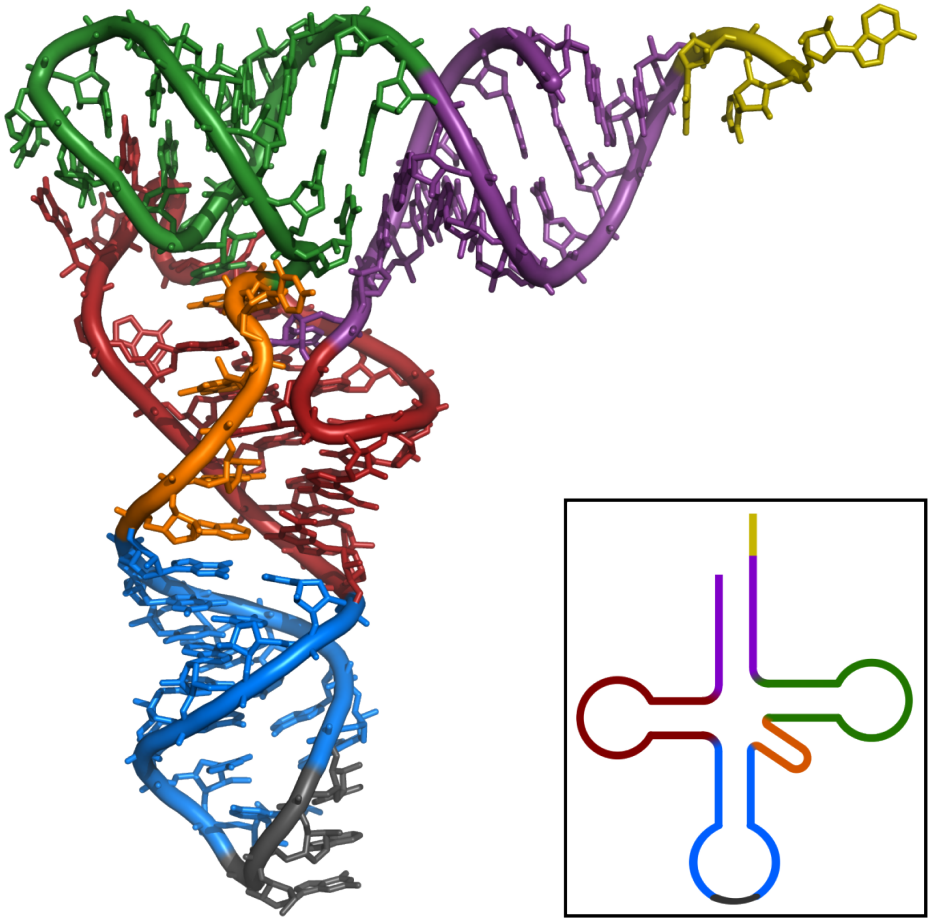
- primary sequence poorly conserved
- act through other mechanisms
- are under strong purifying selection (sequence/structures stretches, splice sites)
- Disease-associated SNPs can disrupt local structures (lnc13)

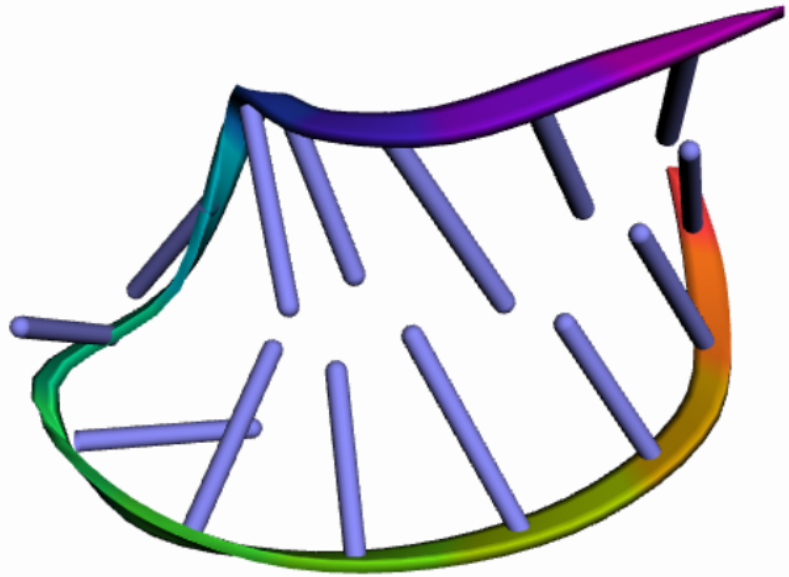


Sup. figure 9
Castellanos-Rubio et al.

Castellanos-Rubio *et al* Science 2016

structure defines function





Rfam tRNA; Duszczyk *et al.* "H, 13C, 15N and P chemical shift assignments of a human Xist RNA A-repeat tetraloop hairpin essential for X-chromosome inactivation." Biomol. NMR Assignments, 2012

selective pressures

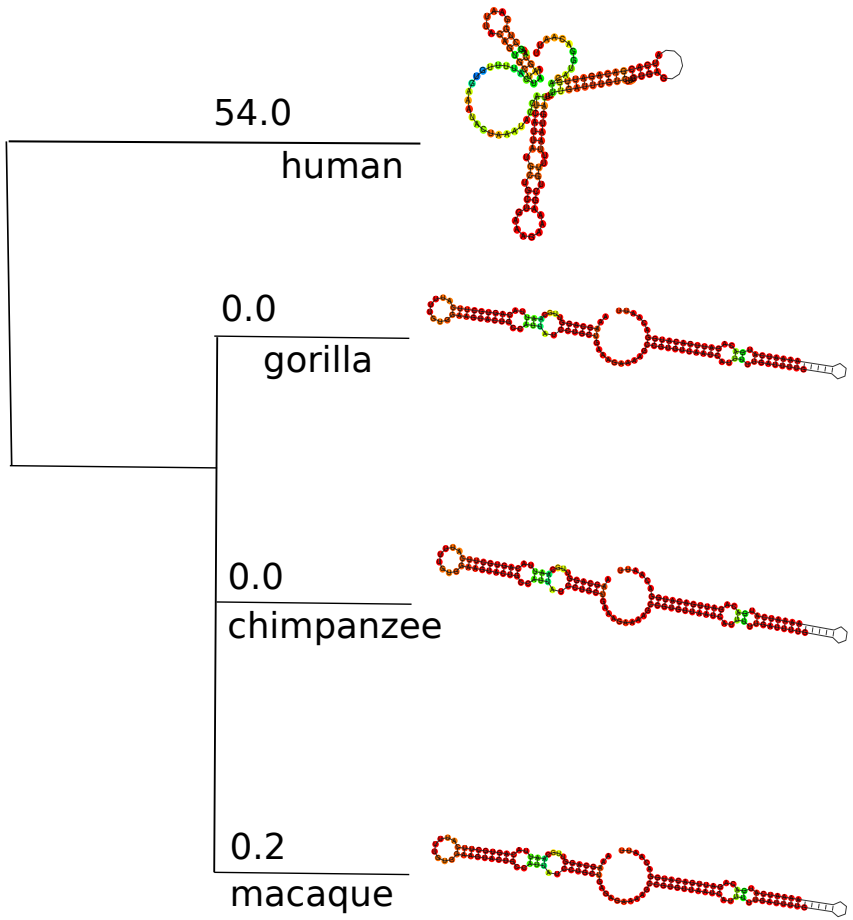
- Objective: understand the role of lncRNAs in human brain
- Selective pressures are the main cause for molecule evolution
- By understanding the selective process
  - start to characterise a molecule
  - understand species relations
  - find genes that have specialized in one species, when compared to closely related species

types of selective pressures

- Negative selection
  - goes towards eliminating variants
- Neutral selection
  - accepts some level of variants
- Positive selection
  - goes towards some new function, selecting new variants
  - detected by looking for rare events within a group

studying ncRNA selection

- Focus on conservation: RNAz, CMfinder, SISSIz, RNAalifold, LocaRNA
- Finding conserved structures can be more easily done
  - Find conserved set → functional group
  - Spot different structure → specialized functionality
- How to find structures that changed in only one lineage in a conserved set?



- Understand the role of long ncRNAs in human brain evolution
  - Develop a method for detecting positive selection in ncRNA structures (Chapter 1)
  - Search for lncRNA candidates that were positively selected in humans and are expressed in brain (Chapter 1)
  - Study the structural selection of a lncRNA that is know to be under positive selection in humans (Chapter 2)
  - Develop an algorithm for retrieving lncRNA orthologs based on splice site orthology (Chapter 3)
  - Produce a catalog of primate lncRNA orthologs based on splice site orthology (Chapter 3)

method for detecting positive selection

- Simple counting
  - analagous to the Ka/Ks
  - based on the distinction between synonymous and non-synonymous sites
- Statistical modelling
  - calculates the probability of an event using the Poisson model
  - also depends on the distinction between synonymous and non-synonymous sites
- *Combining p-values*
  - distinguishes rare and common events
  - considers every mutation uniquely

combining p-values

- **Main idea**: a selection score combining all events
  - mutations
  - indels
- Events can be rare or ordinary
- Rare events point towards interesting events $\rightarrow$ *positive selection*
- Get individual mutations p-values with RNAsnp tool

combining p-values - scoring mutations

- Get list of p-values
- Bonferroni multiple testing correction

$$p_c = p * k$$

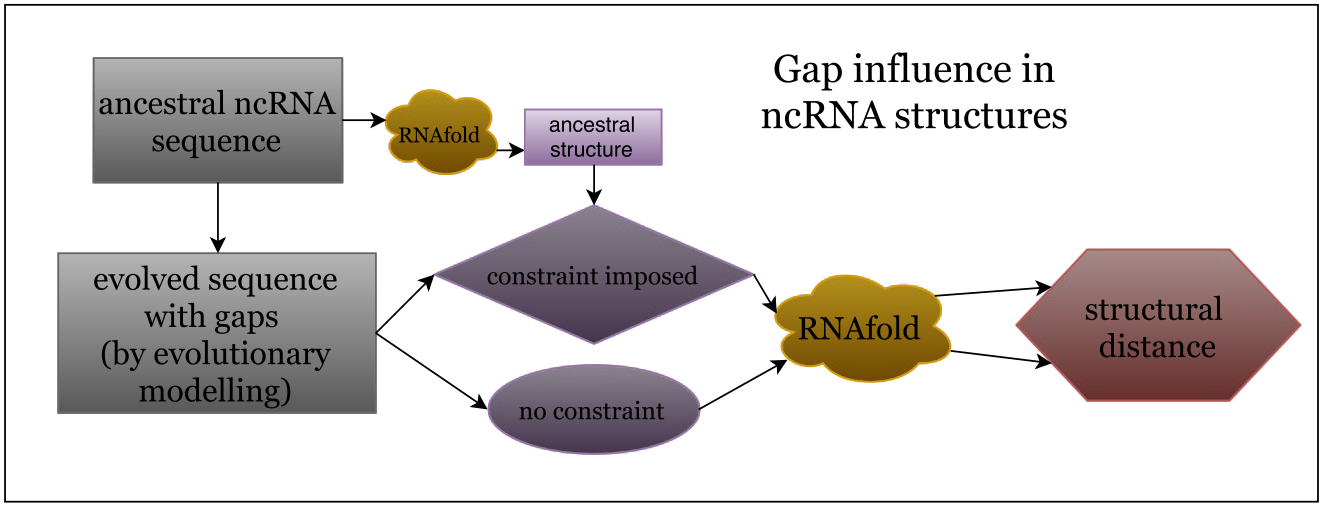
- combine them using Fisher's method

$$s = -2 \sum_{i=1}^{k} \log(p_i)$$

s: mutation selection score
p: p-value of mutation *i*
k: number of mutations

rank statistics - scoring indels

- How common or rare are the observed indels?



$$p = \frac{n - i}{n}$$

p: probability of observing this indel event or more extreme ones
n: number of different possible indel groups

i: group in which the observed indel is

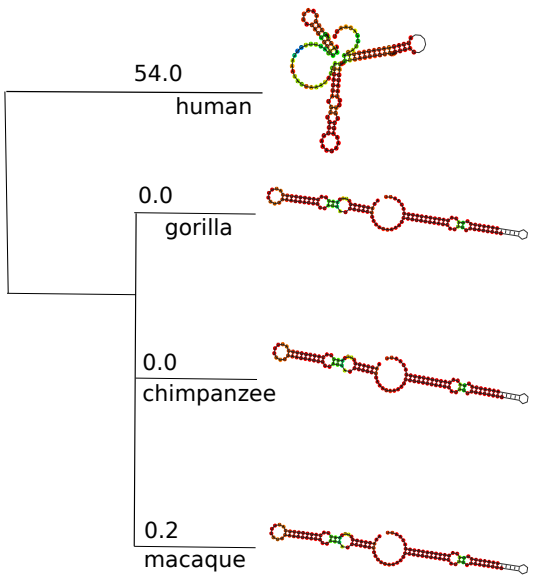
combining p-values - scoring selection

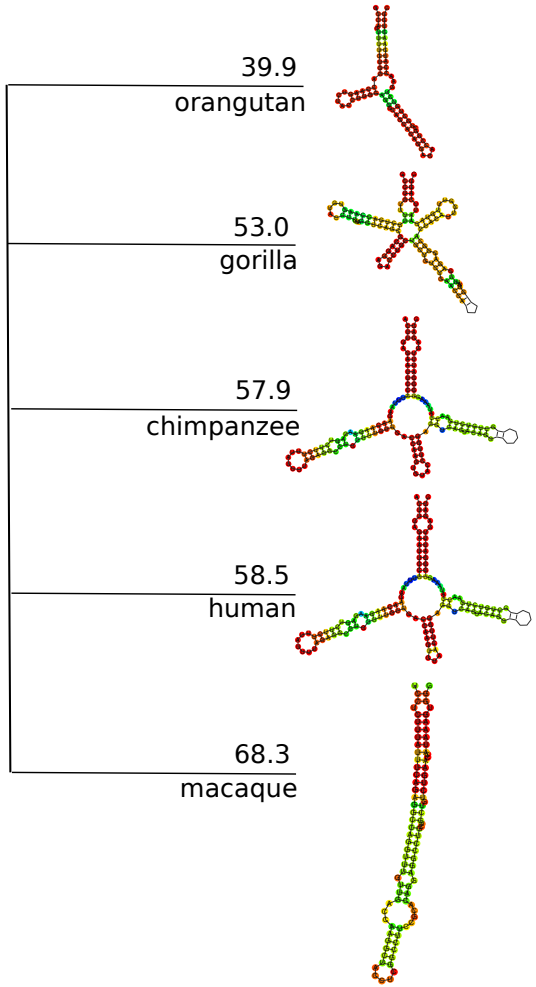
- we then have two lists:
  - (i) mutation p-values
  - (ii) indel p-value analogous (done by rank statistics)
- multiple correct them
- combine them using Fisher's method
- combine both scores for mutation and indels, getting a final selection score

evaluation and application of the SSS-test

- Species: human, chimpanzee, gorilla, orangutan, and rhesus macaque (exception of snoRNA -orang.)
- Evaluation
  - small ncRNA DB (known to be structurally conserved)
    - microRNAs (167 families)
    - small nucleolar RNAs (170 families)
    - tRNAs (611 families)
  - Synthetic DB, conserved and positive models (100 families each)
- Application: lncRNA DB Necsulea *et al* Nature 2014 (15,443 families)
  - Objective: suggestion of candidates for having evolved under positive selection in humans

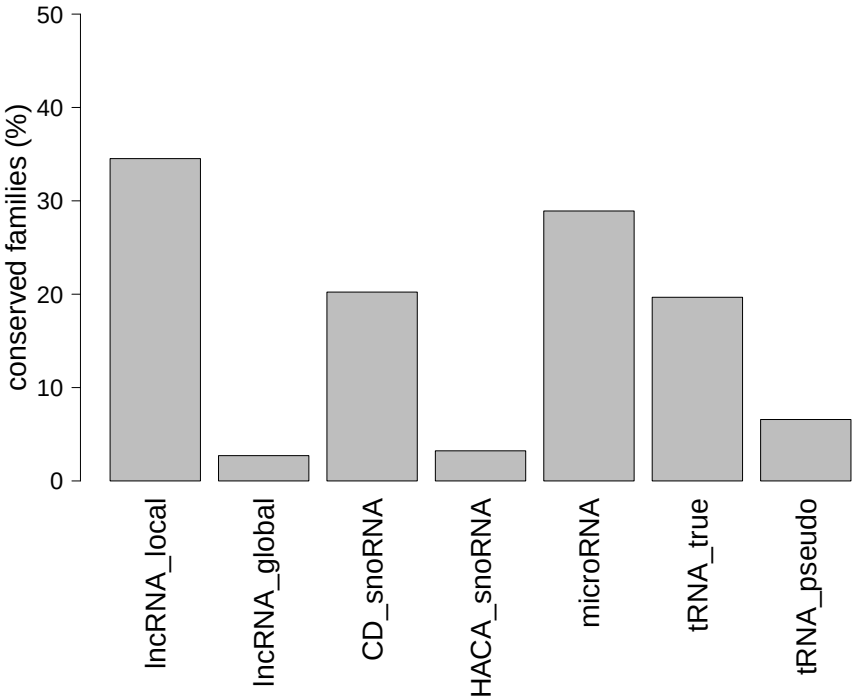
family diversity

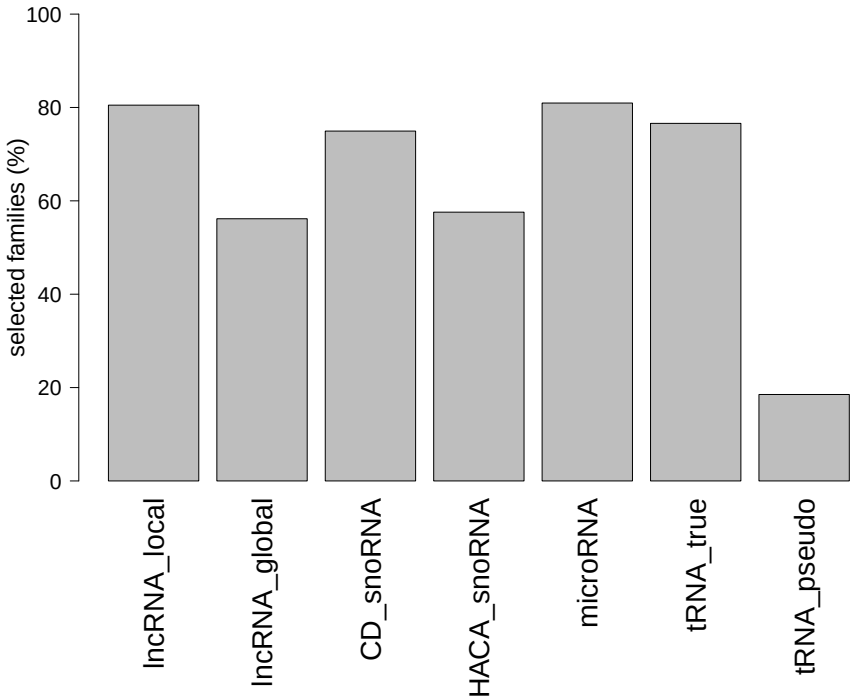




small ncRNAs have well conserved local structures

- Extremely lowly diverged families ($d = 0.0$) - full set
- Extremely conserved structures ($s = 0.0$) - conserved set





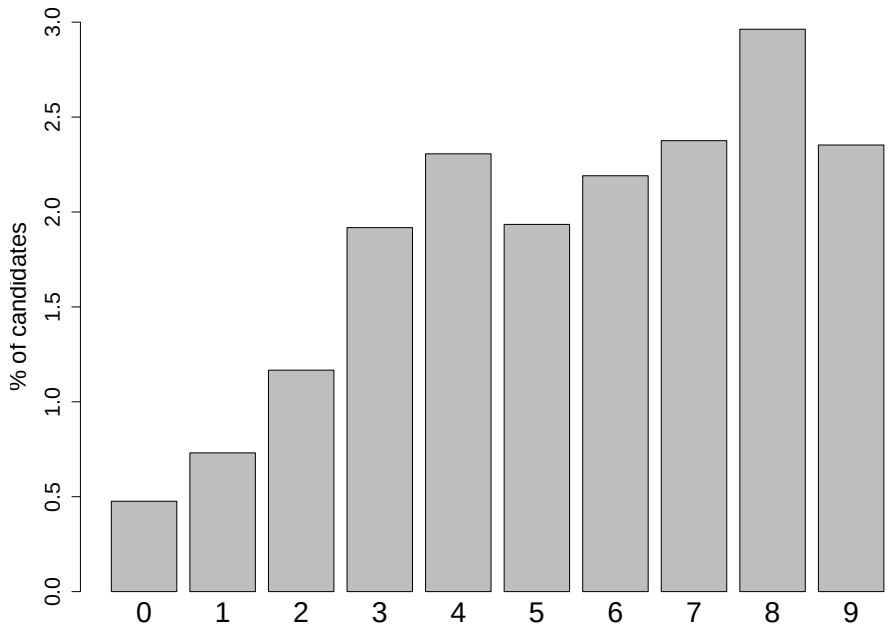
primate lncRNAs have well conserved local structures

Table: Characterization of the structural selection of local lncRNAs. Only the low diverse family set was considered in this analysis.

Species	Representatives (local structures)	Conserved ($s \leq 2$)	Positive ($s \geq 10$)
Human	8,929	8,198 (91,8%)	192 (2,2%)
Pan	8,733	8,013 (91,8%)	158 (1,8%)
Orangutan	6,433	6,001 (93,3%)	432 (6,7%)
Gorilla	8,077	7,865 (97,4%)	212 (2,6%)
Macaque	5,111	4,188 (81,9%)	923 (18,1%)

humans local structure selection

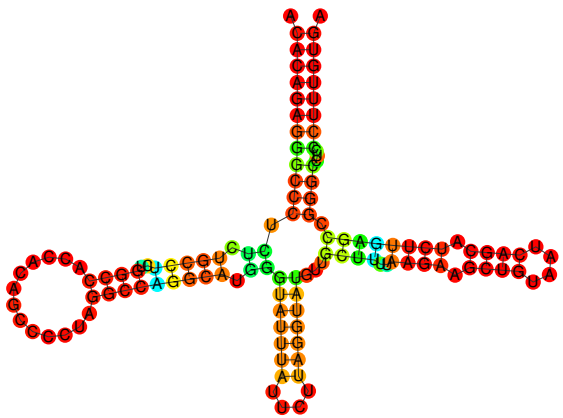
- 191 candidates for positive selection (192 local structures)
- Maybe association between candidates and evolutionary age
- Candidates seem to be expressed in multiple tissues



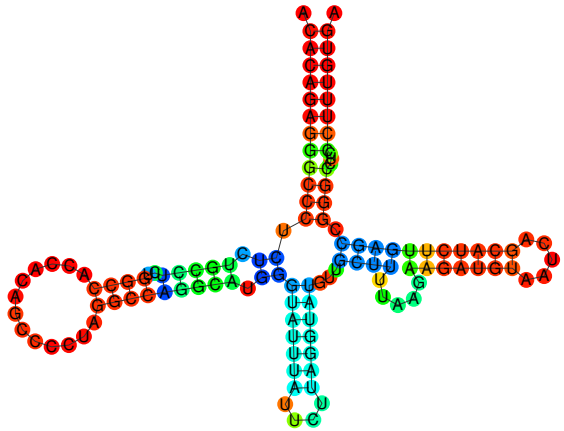
humans local structure selection

- 31 have ENSEMBLE IDs
- 2 candidates are antisense to brain proteins
  - TRPM2: ion channel, essential for cell survival, modulates mitocondrial responses associated with neuroblastoma
  - SIX3: transcription regulator, role in eye development, associated with cephalic disorder

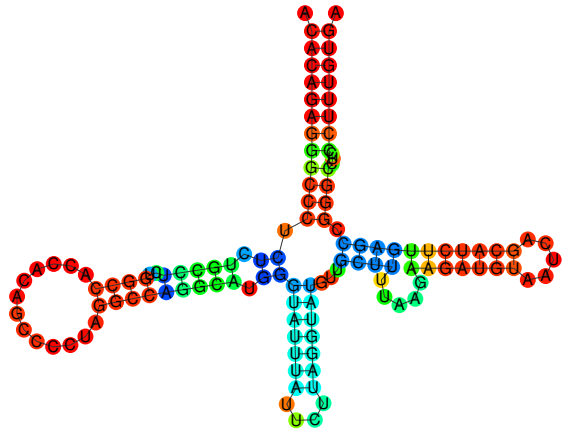
Human



Pan



Orangutan

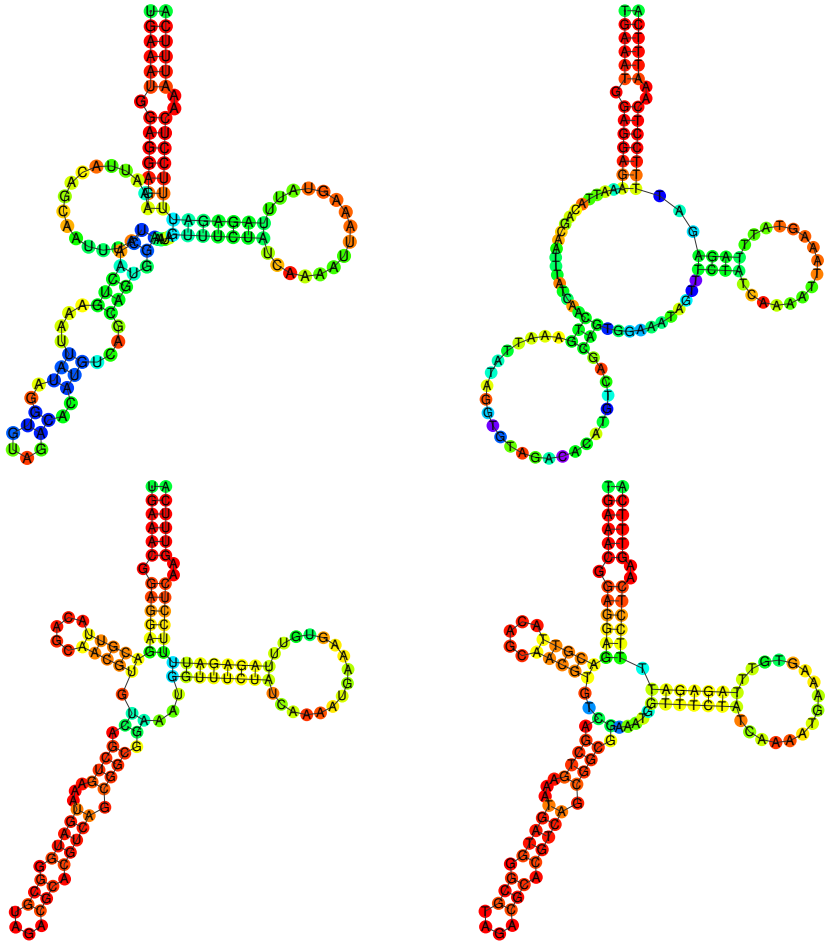


summary

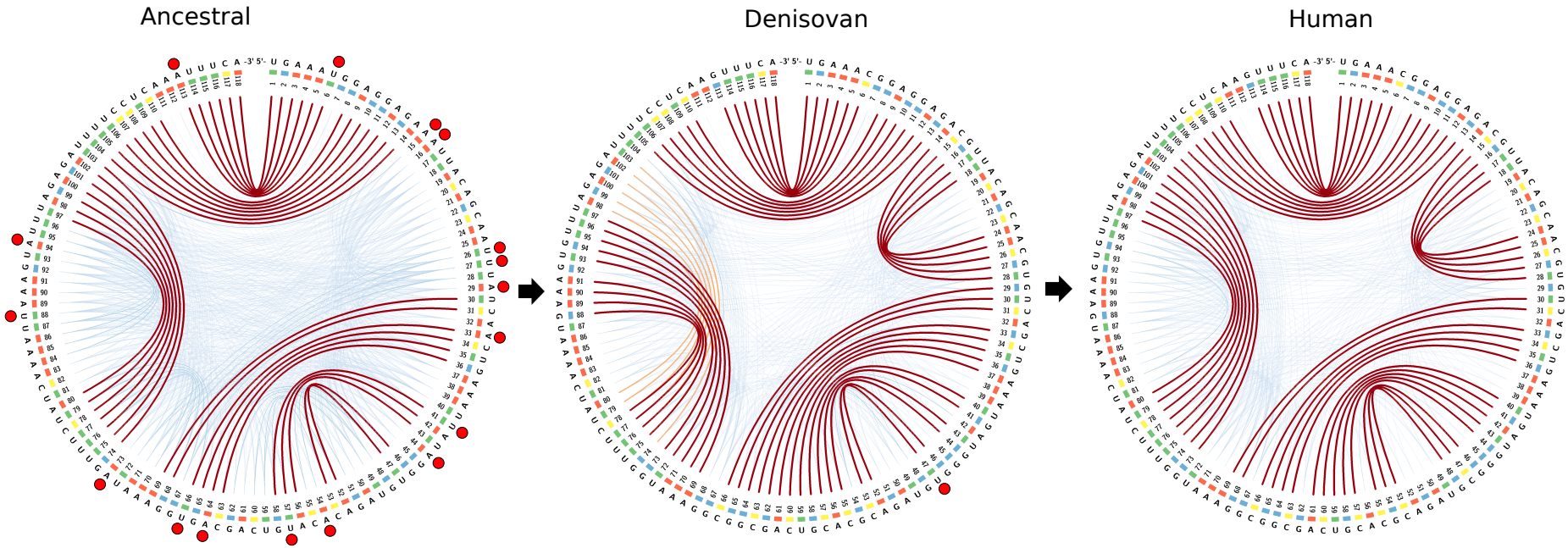
- New method for detecting positive selection in ncRNA structures: SSS-test
- 191 candidates for positive selection in humans
- Seem to be expressed in multiple tissues
- May have an association with age

structural evolution of human accelerated region 1

- 118 bp part of HAR1F/HAR1R
- 18 human specific mutations
- expressed in Cajal-Retzius cells
  - 7 - 19 gestational weeks
- co-expressed with Reelin
  - organisation of laminar cortex
- Human structure differs from Chimpanzee
- Conserved in vertebrates
  - chimpanzee's $\sim$ ancestral
- How was human HAR1 evolutionary process?
  - MutationOrder



comparing ancestral, archaic and modern human ensembles



a model to study the temporal order of mutations

- Combinatorial optimization problem (Hamiltonian path problem)
- Given ancestral sequence x, secondary structure $S(x)$
  - extant pair y and $S(y)$ (the selection target)
  - set of X of fixed mutations
- Evolutionary path π is a permutation of X
- Fitness function: $f(u) = -d(S(u), S(y))$

$$f(\pi) = \sum_{i=2}^{|X|} \left(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y)) \right)_+ \quad (1)$$

a model to study the temporal order of mutations

- Sum only includes steps in wich fitness decreases
  - distance to the target increases
- Likelihood of path π decreases exponentially with its fitness cost
  - β is a scaling parameter and Z a normalization factor

$$\text{Prob}[\pi] = e^{-\beta f(\pi)}/Z \tag{2}$$

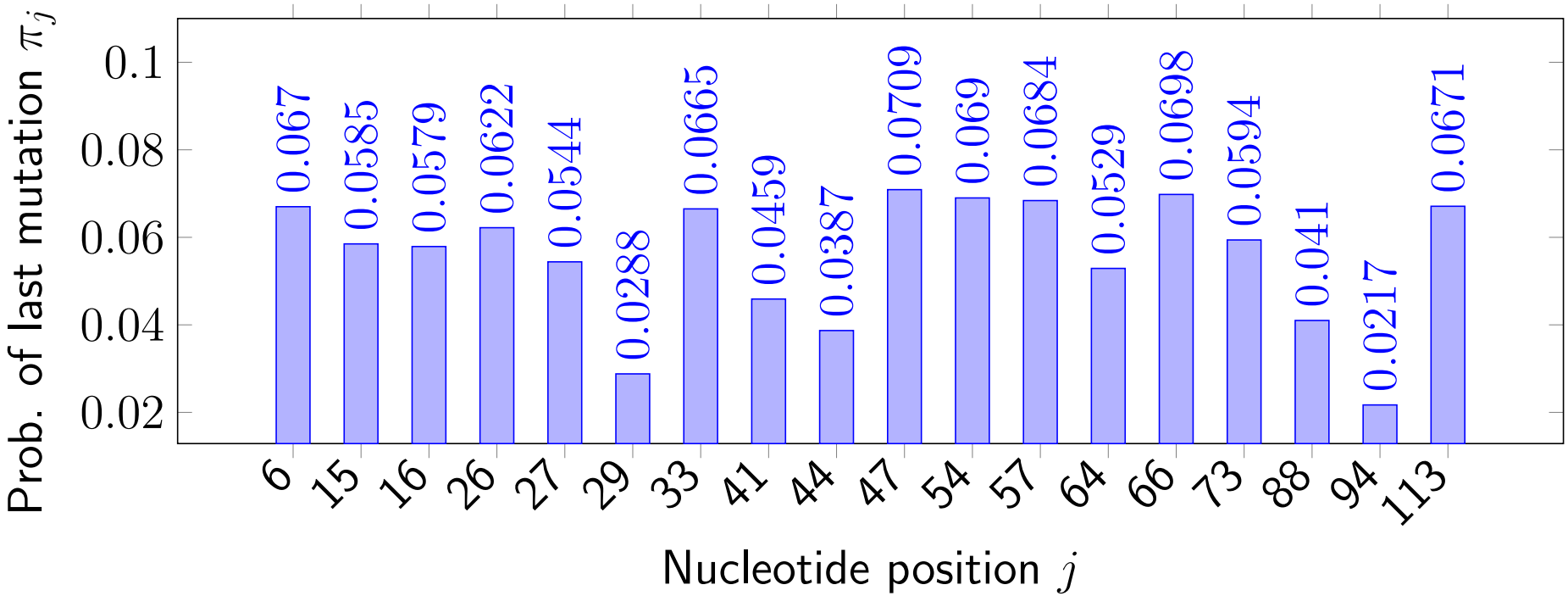
- Probability of j being the last mutation

$$\pi_j = \sum_i \pi_{ij} \tag{3}$$

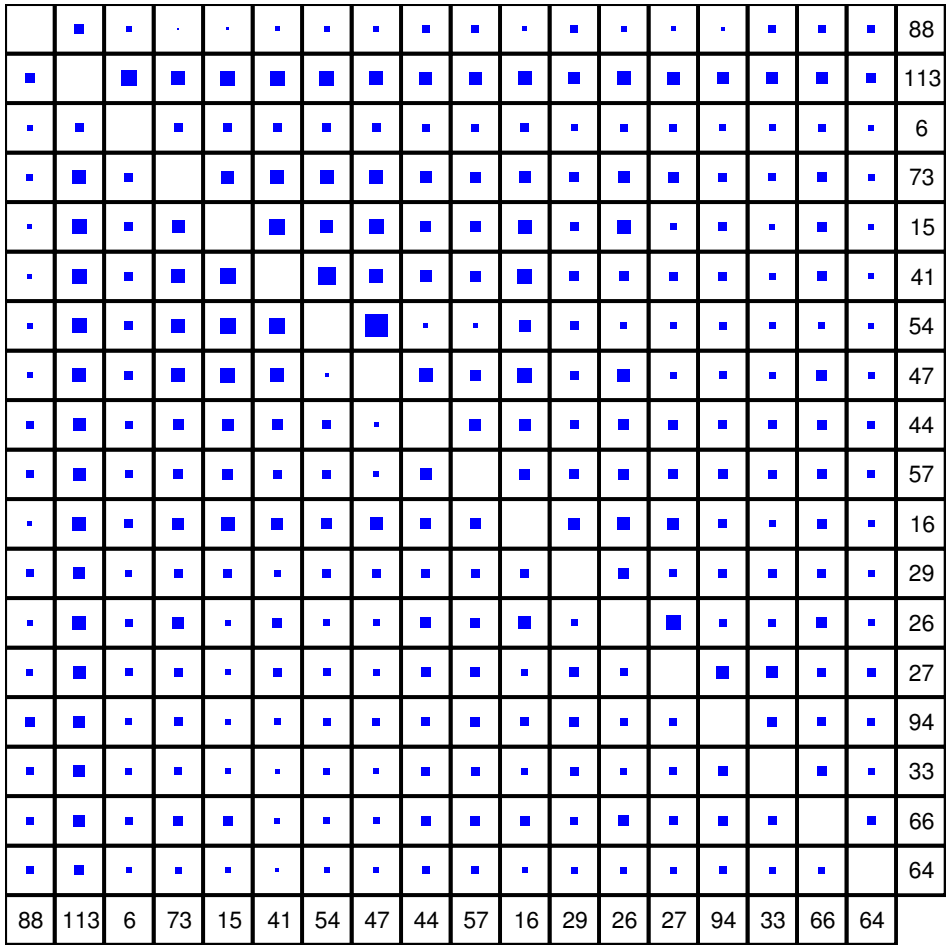
extended model including intermediate and backmutations

- Mutations occur during evolution
  - some are fixed (18 ones in HAR1)
  - some are not, and they stay in the population only temporarily
- If we want to include temporary mutations in our model
  - we will have (a lot) more evidence
  - any of the 3 other bases changing among the 100 human ones that maintained since last ancestor and changing back
  - any other 2 being intermediate to the 18 fixed ones, after the ancestral
- log-evidence $\ln n!$ in "nats" (units of information in the natural logarithm)
  - original model: ≈ 36.40 nats ($18!$ permutations)
  - backmutation model: ≈ 42.34 nats ($20!$ permutations)

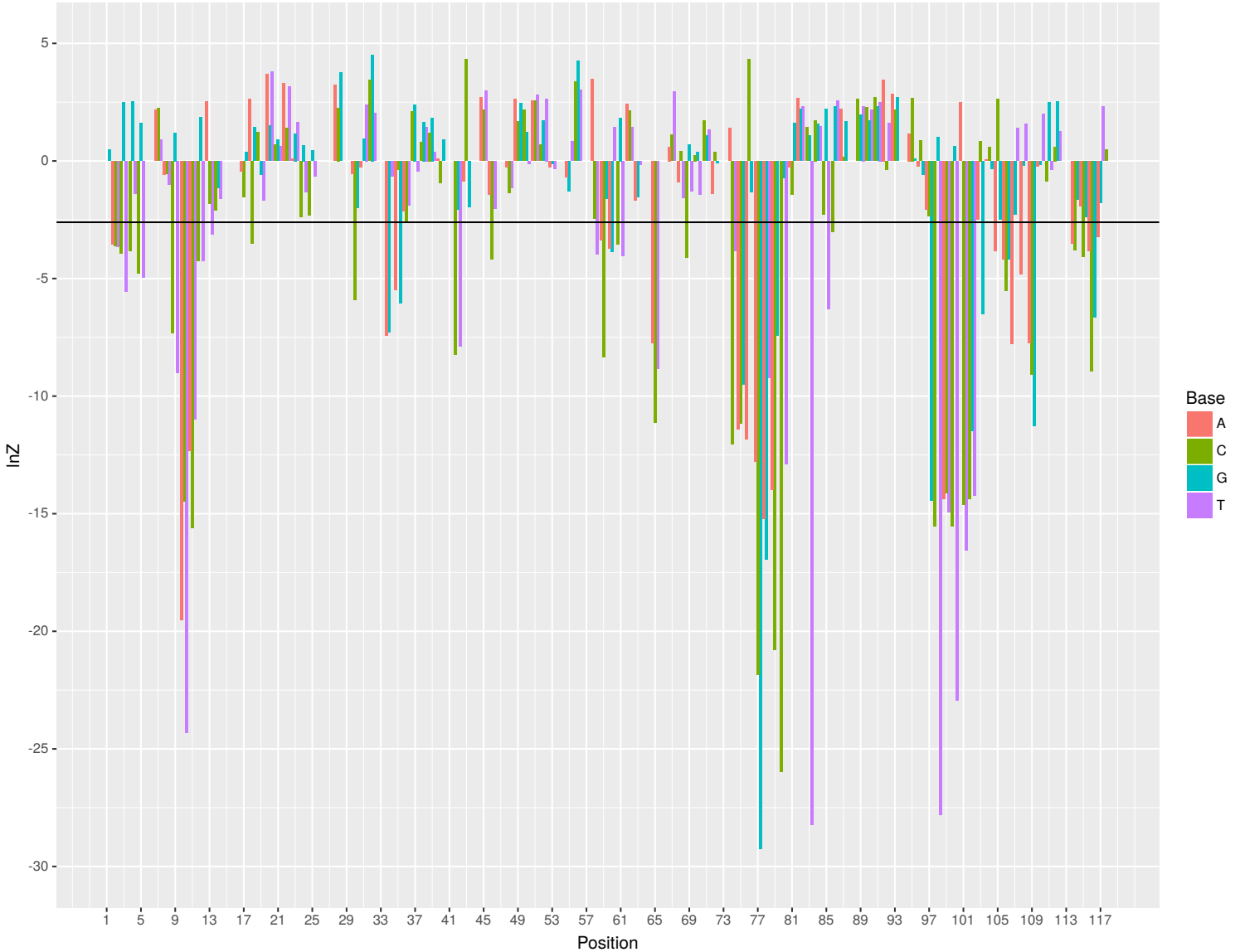
which one was the last mutation?



probability (P_{kl}) of mutation k (row) be followed by l (column)



impact of back-mutations on log-evidence of path



summary

- New method for studying structural evolution of ncRNA structures: MutationOrder
- Human HAR1 seems to have evolved under pressure to become more stable
- Last mutation seems to be position 47 (Denisovan)
  - being preceded by position 54
- Alternative paths seem equally likely
  - order does not seem to be crucial in this process

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
- 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*, submitted, 2017
- 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*, submitted, 2018
- 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

software and thesis

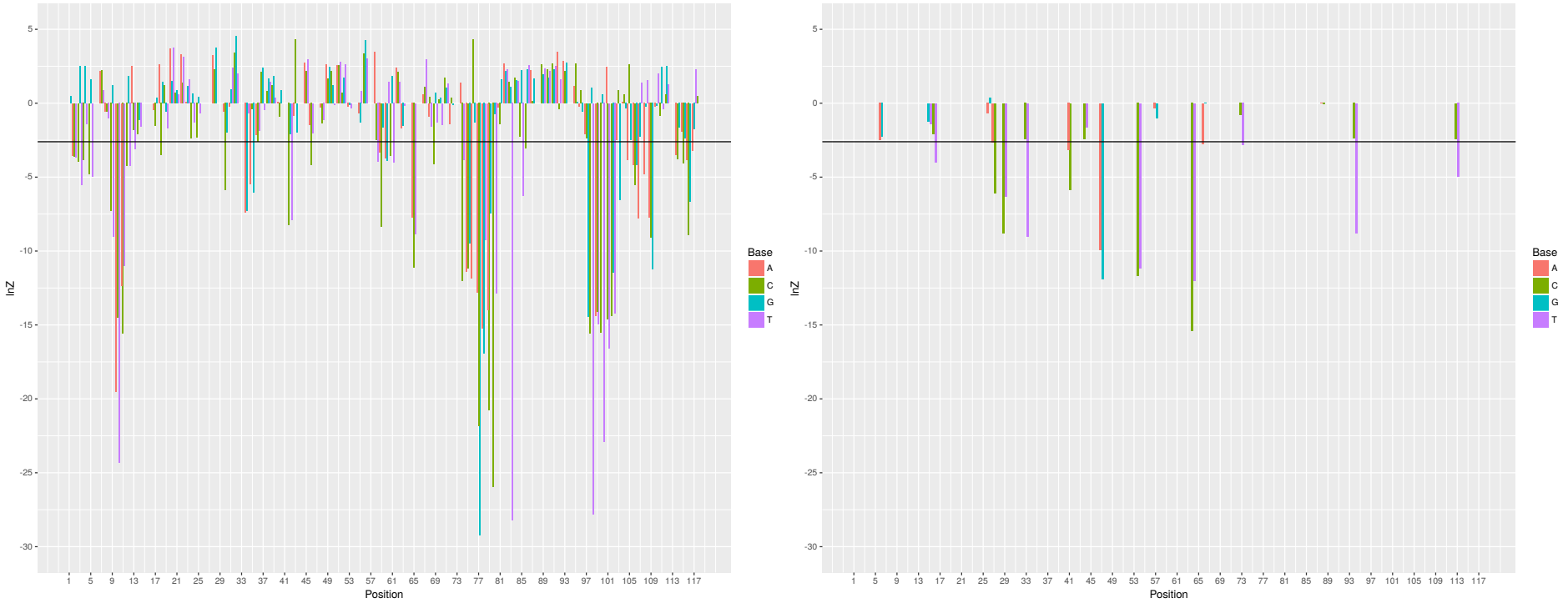
- MutationOrder
  - ⟨http://hackage.haskell.org/package/MutationOrder⟩
  - Pre-compiled binaries: ⟨https://github.com/choener/MutationOrder/releases⟩
- CS^2-UPlot in preparation as a web tool
- SSS-test: soon at ⟨https://github.com/mbwalter⟩
- buildOrthologs.pl: soon at ⟨https://github.com/mbwalter⟩

- PhD thesis (soon): *Long Non-Coding RNAs and the Evolution of the Human Brain - Algorithms in Computational Biology*

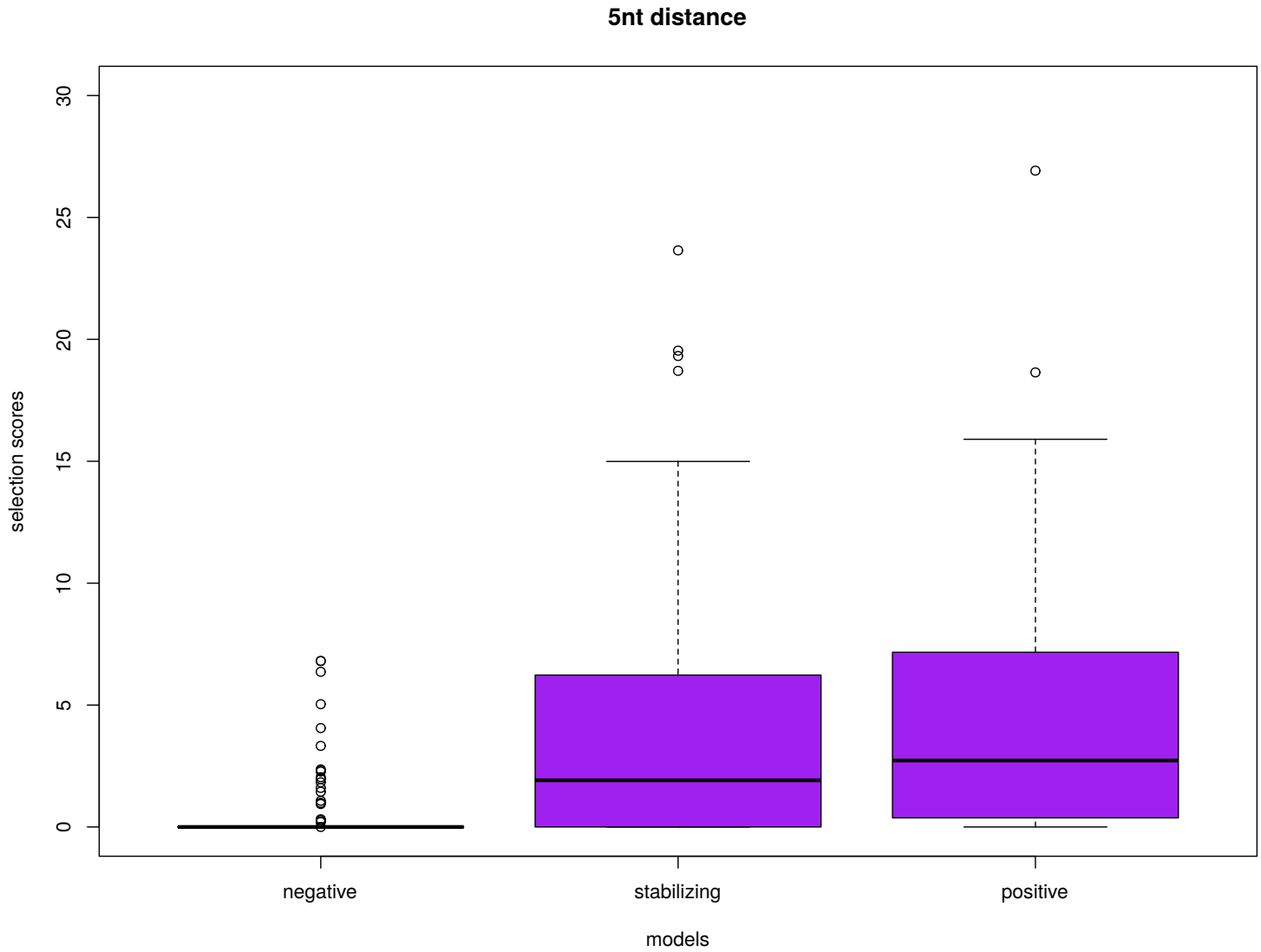
acknowledgements

Peter Stadler, Katja Nowick, Christian Höner zu Siederdissen, Nowicklab, Anne Nitsche, Irma Lozada, Dan Tulpan, Jana Hertel, Stephanie Kehr, Jens Steuck, Petra Pregel, Andrea Fallmann, Corinna Pregel, Bioinf Leipig, CNPq (Science without Borders/Brasil) and DFG 1738 for my funding, University of Leipzig

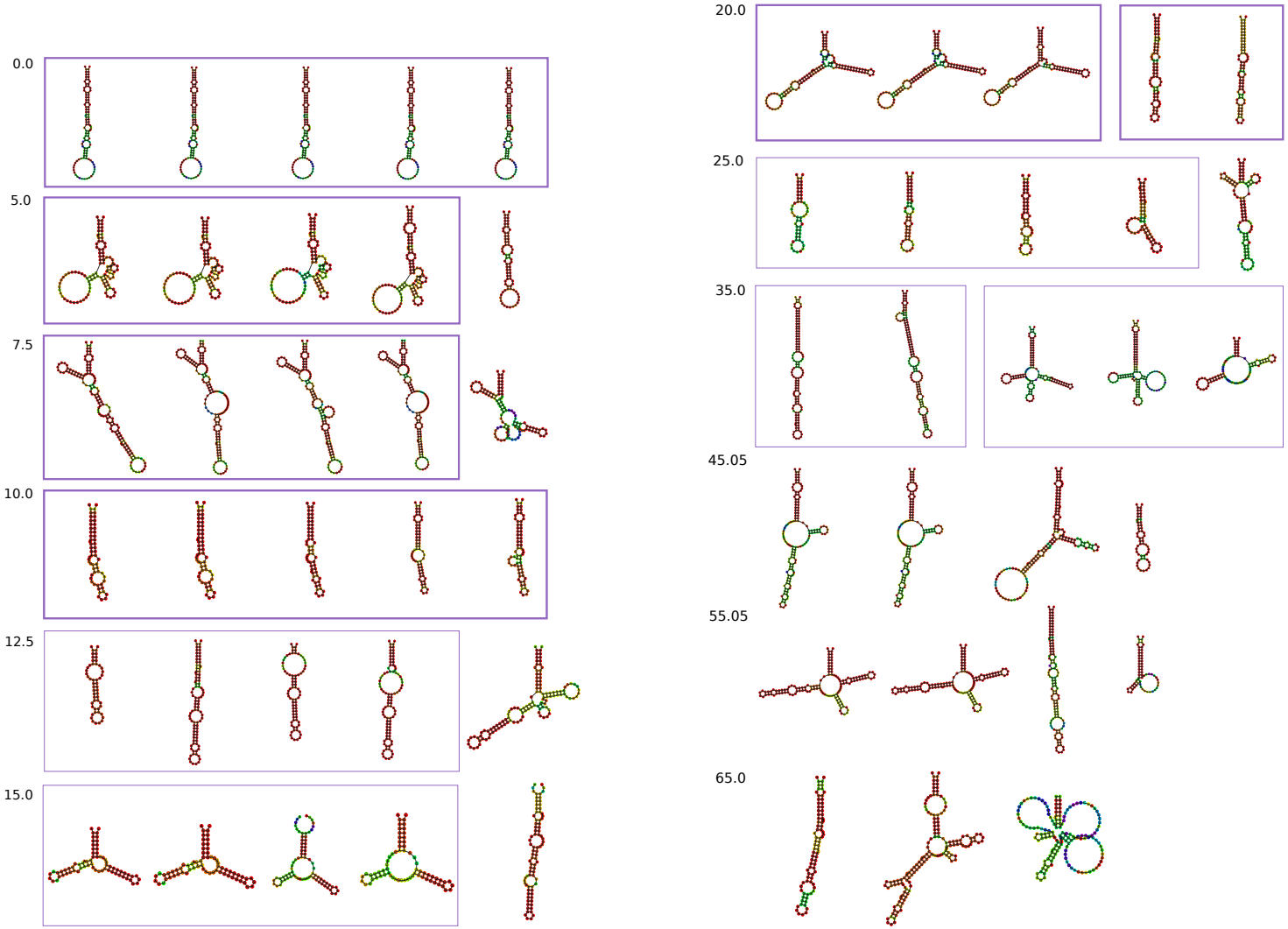
back and intermediate mutations



evaluation of the SSS-test: synthetic data (RNAdesign)



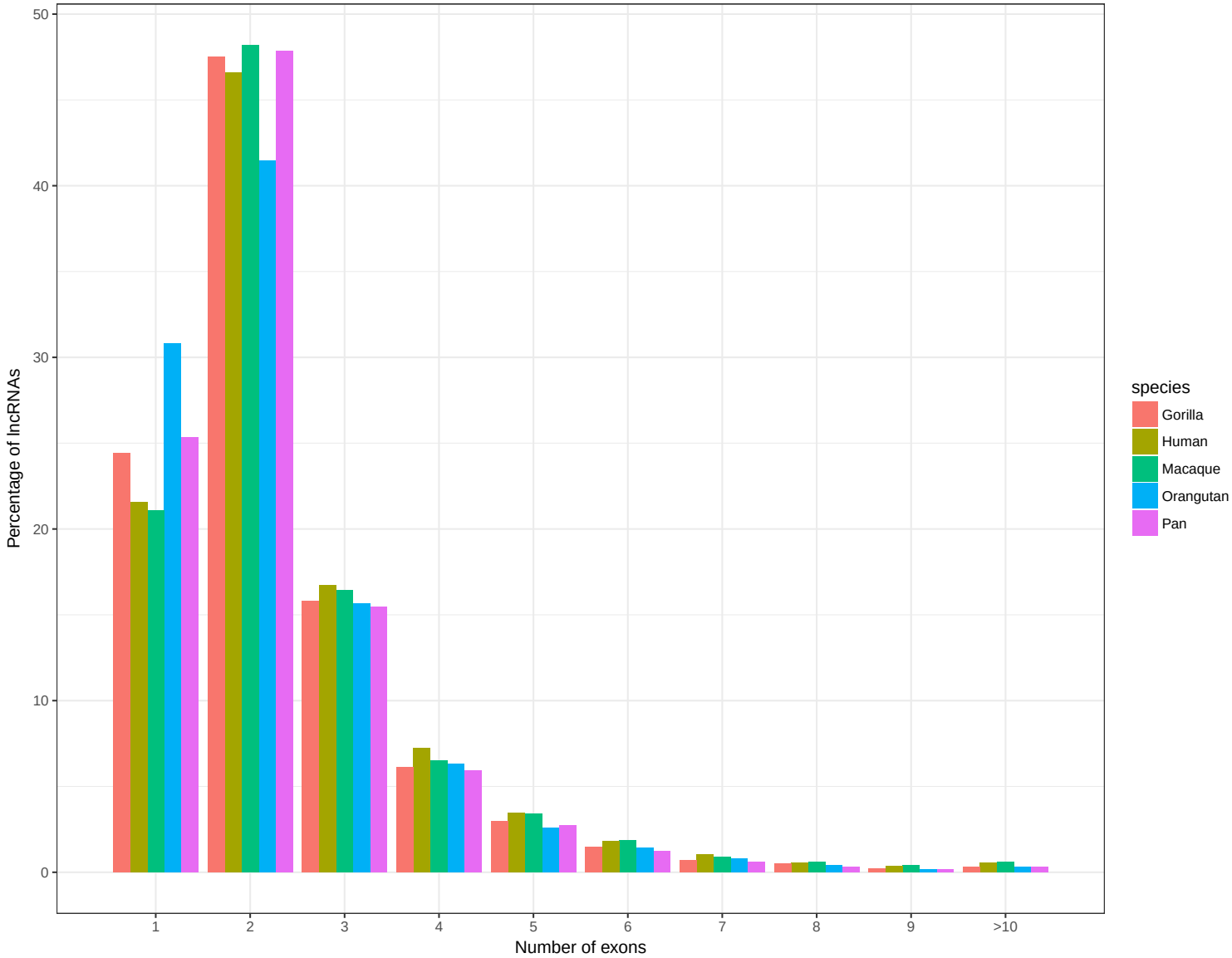
family diversity - choosing an appropriate cutoff



candidates age (31 with ENSEMBLE ID)

Gene Name	Transcrip age	Seq age	NbSp transcrip	NbSpecies seq	ENSEMBL Gene ID
DENND6A-DT	Primates	Primates	3	4	ENSG00000241933
LINC01839	Primates	Primates	4	4	ENSG00000227509
C5orf66-AS1	Primates	Primates	4	5	ENSG00000249082
LINC01861	Primates	Primates	4	5	ENSG00000251183
LINP1	Primates	Primates	5	5	ENSG00000223784
LINC01280	Primates	Primates	5	5	ENSG00000224391
LINC01802	Primates	Primates	5	5	ENSG00000225064
PLCB1-IT1	Primates	Primates	5	5	ENSG00000225479
LINC01345	Primates	Primates	5	5	ENSG00000226374
LINC01724	Primates	Primates	5	5	ENSG00000227421
LINC01693	Primates	Primates	5	5	ENSG00000227764
MACC1-AS1	Primates	Primates	5	5	ENSG00000228598
LINC00659	Primates	Primates	5	5	ENSG00000228705
TRPM2-AS	Primates	Primates	5	5	ENSG00000230061
LINC01431	Primates	Primates	5	5	ENSG00000232645
ERI3-IT1	Primates	Primates	5	5	ENSG00000233602
PLUT	Primates	Primates	5	5	ENSG00000247381
LINC01258	Primates	Primates	5	5	ENSG00000249534
LINC02501	Primates	Primates	5	5	ENSG00000249882
OPCML-IT1	Primates	Primates	5	5	ENSG00000254896
MDC1-AS1	Primates	Eutherians	2	6	ENSG00000224328
RRS1-AS1	AfricanApes	GreatApes	3	4	ENSG00000246145
LINC01939	GreatApes (max)	GreatApes	3	4	ENSG00000228799
LINC01774	Therians (max)	Therians	4	6	ENSG00000236950
LINC02042	Eutherians	Eutherians	5	5	ENSG00000240893
LINC02092	Eutherians	Eutherians	5	6	ENSG00000234721
LINC01738	Eutherians	Eutherians	6	6	ENSG00000227947
LINC02288	Eutherians	Eutherians	6	6	ENSG00000246548
LINC02217	Eutherians	Eutherians	6	6	ENSG00000248455
MIR3663HG	Therians	Therians	7	7	ENSG00000234474
SIX3-AS1	Tetrapods	Tetrapods	9	9	ENSG00000236502

exon size distribution - primate lncRNAs

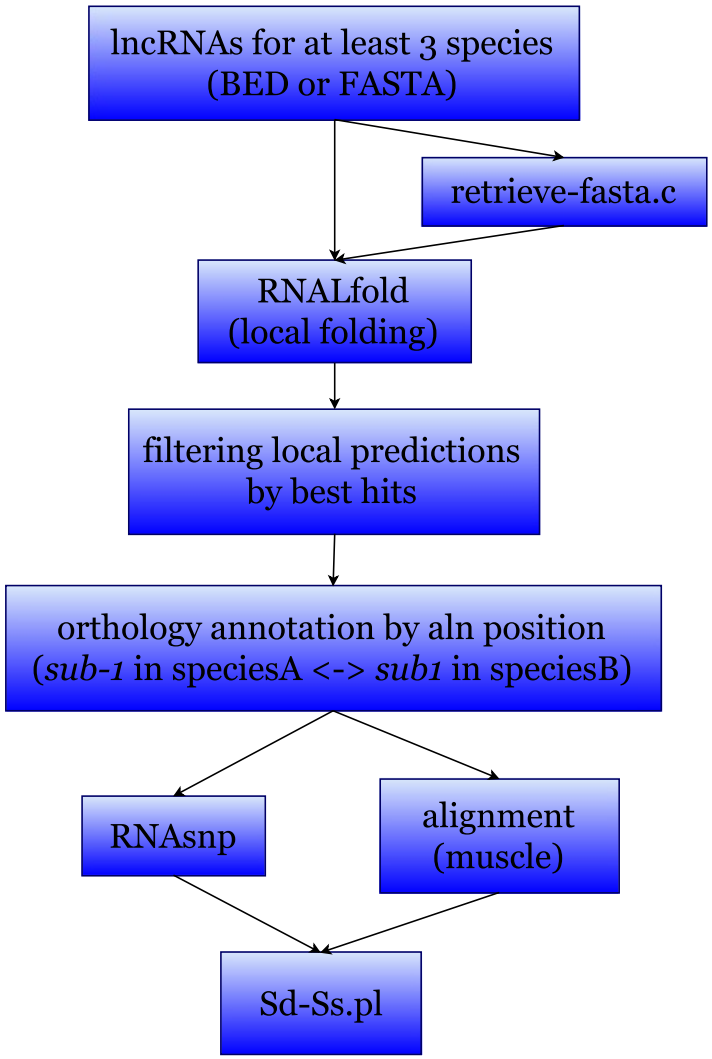


fitness cost of a path with a single back-mutation b

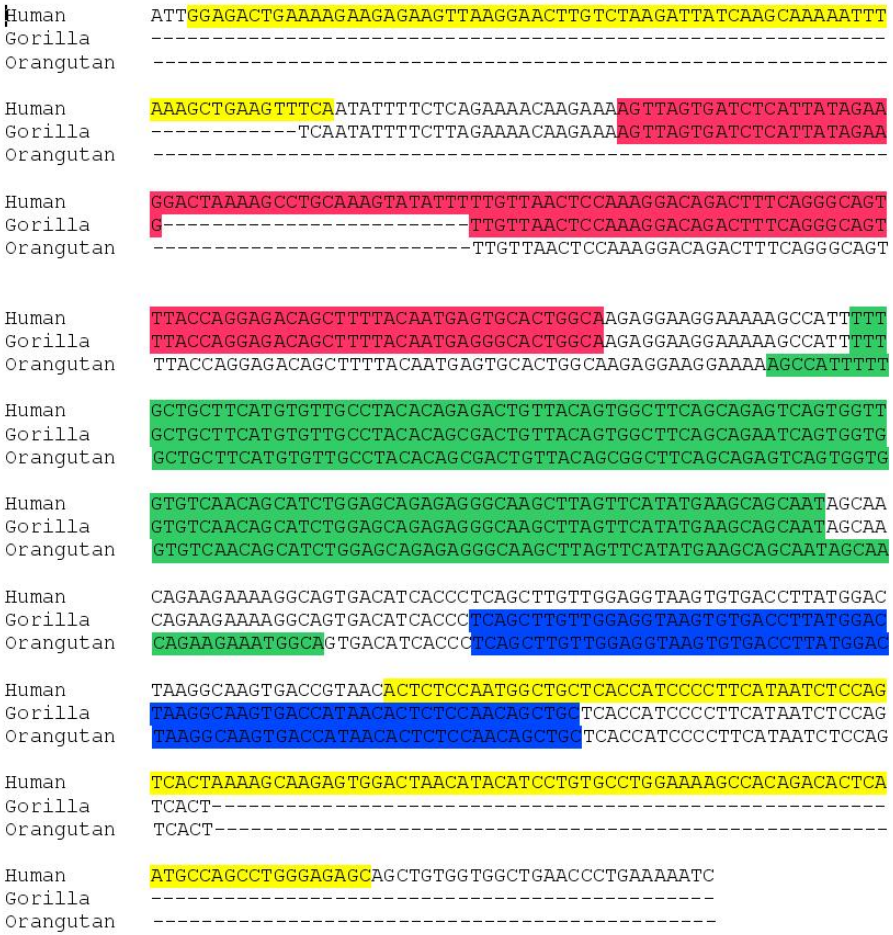
Fitness function: $f(u) = -d(S(u), S(y))$

$$
\begin{aligned}
f(\pi, b_+, b_-) = & \sum_{i=2}^{b_+-1} \left(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y))\right)_+ \\
& + d(S(\pi^B_{b_+}), S(y)) - d(S(\pi_{b_+-1}), S(y))\big)_+ \\
& + \sum_{i=b_+}^{b_--1} \left(d(S(\pi^B_i), S(y)) - d(S(\pi^B_{i-1}), S(y))\right)_+ \\
& + d(S(\pi_{b_-}), S(y)) - d(S(\pi^B_{b_--1}), S(y))\big)_+ \\
& + \sum_{i=b_-}^{|X|} \left(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y))\right)_+
\end{aligned}
\tag{4}
$$

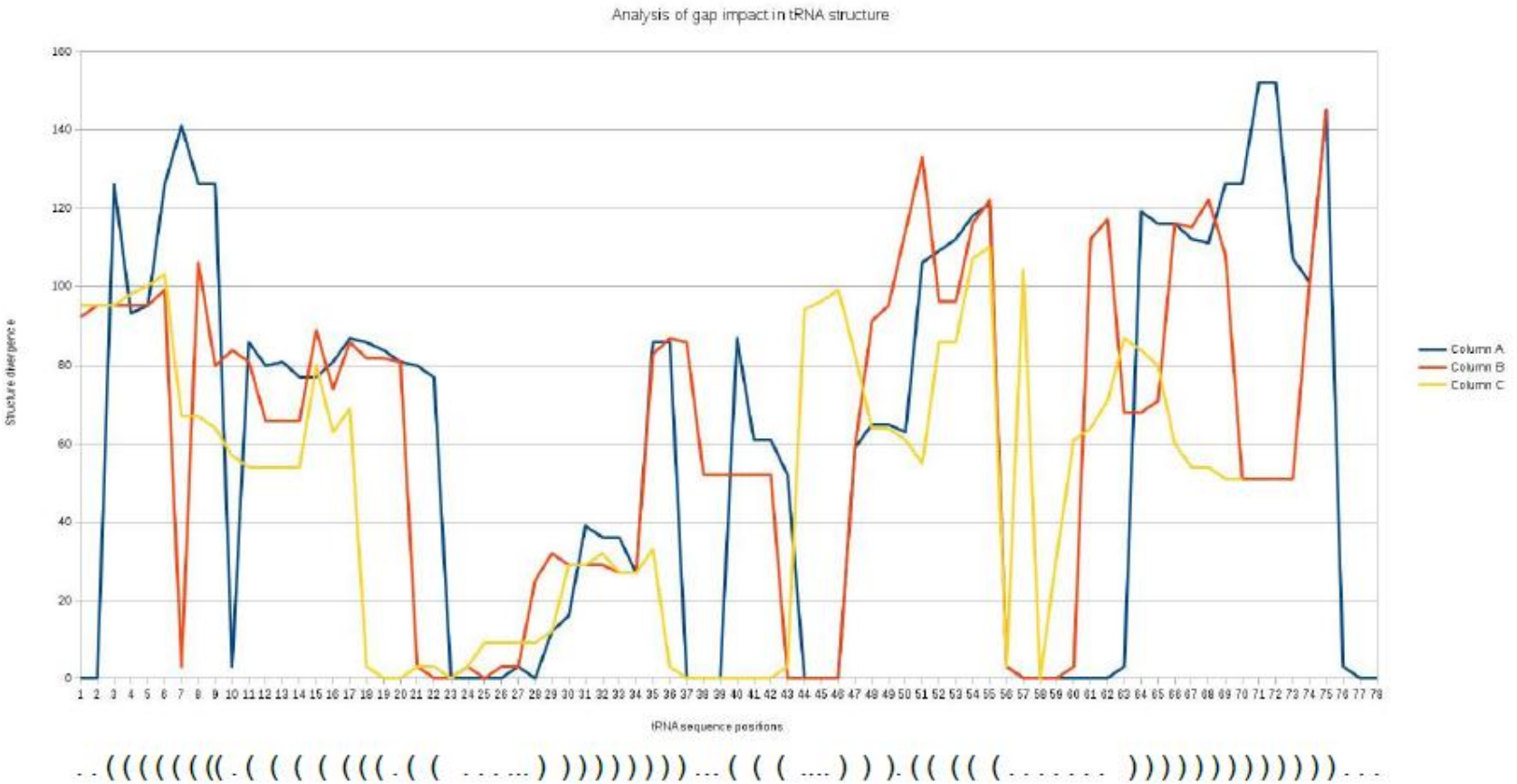
local structure pipeline for long ncRNAs



local structure orthology for lncRNAs



deletion impact in a ncRNA sequence



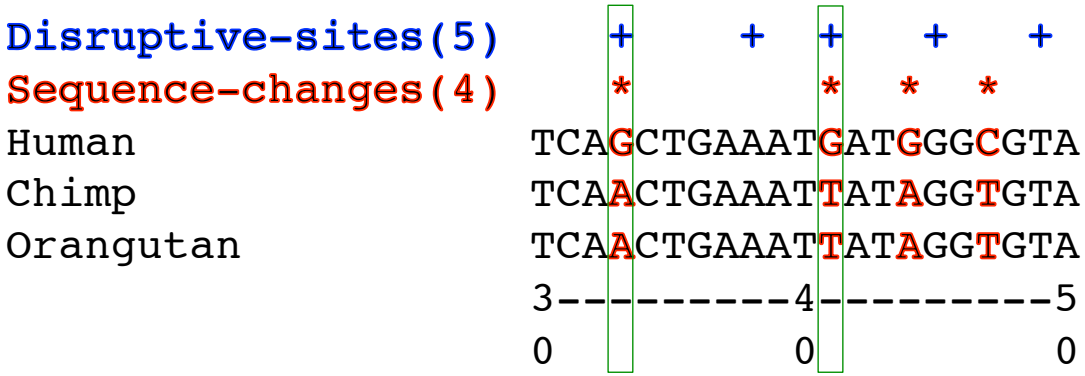
positive and negative controls

species_ID	no_ch	sp_dist	fisher_indels	fisher_mutations	sp_score
HAR1-hSapiens	13	69.3	0	3.1459	3.1459
HAR1-panTro	0	0.0	0	0	0.0000
HAR1-pAbelii	0	0.0	0	0	0.0000
HAR1-gorilla	0	0.0	0	0	0.0000
HAR1-denisovan	15	70.8	0	2.5735	2.5735

species_ID	no_ch	sp_dist	fisher_indels	fisher_mutations	sp_score
ccr_mir_430	0	16.6	0.0000	0	0.0000
dre_mir_430a_13	1	17.4	0.0000	0.0006	0.0006
dre_mir_430a_1	7	1	18.2	0.0000	0.0006
hhi_mir_430b_2	9	26.6	0.0000	0.0000	0.0000
hhi_mir_430b_1	10	43.1	0.0000	0.0000	0.0000
ola_mir_430a_2	0	21.0	0.0000	0	0.0000
ola_mir_430a_1	0	23.3	0.0000	0	0.0000
pma_mir_430a	9	32.5	0.0000	0.0000	0.0000
pma_mir_430c	8	26.9	0.0000	0.0000	0.0000
ssa_mir_430a	0	14.3	0.0000	0	0.0000
ssa_mir_430c	1	12.1	0.0000	0.6506	0.6506
ola_mir_430c_7	5	61.9	0.0000	0.0000	0.0000

simple counting

- **Main idea**: calculate the balance between synonymous and non-synonymous events



Disruptive/non-synonymous sites are characterised by RNAsnp tool

$$Sd/Ss = \frac{disruptive\ mutations/disruptive\ sites}{non-disruptive\ mutations/non-disruptive\ sites} \tag{5}$$

$$Sd/Ss = \frac{2/5}{4-2/21-5} = \frac{2/5}{2/16} = 3.2 \tag{6}$$

statistical modelling

- **Main idea**: calculate the probability of observing x mutations in the species of interest, given the information from the whole orthology group
  - Rare events, or events with low probability, point towards interesting events $\rightarrow$ *positive selection!*

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

x: observed number of mutations of the species

λ: expected mutation rate of the family, calculated by arithmetic means