Uncovering the Structural Evolution of the Human Accelerated Region 1

Maria Beatriz Walter Costa[a,b], Christian Höner zu Siederdissen[c], Dan Tulpan[d], Peter Stadler[c,e,f,g,h], Katja Nowick[a,i,*]

[a] *TFome Research Group, Bioinformatics Group, Interdisciplinary Center of Bioinformatics, Department of Computer Science, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany*
[b] *Paul-Flechsig-Institute for Brain Research, University of Leipzig, Jahnallee 59, D-04109 Leipzig, Germany*
[c] *Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04107 Leipzig, Germany*
[d] *National Research Council Canada, Information and Communication Technologies, 100 des Aboiteaux Street, Suite 1100, NB E1A7R1, Moncton, Canada*
[e] *University of Vienna, Institute for Theoretical Chemistry, A-1090 Vienna, Austria*
[f] *Max Planck Institute for Mathematics in the Science, 04103 Leipzig, Germany*
[g] *Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany*
[h] *Santa Fe Institute, Santa Fe NM 87501, USA*
[i] *Bioinformatics, Faculty of Agricultural Sciences, Institute of Animal Science, University of Hohenheim, Garbenstrasse 13, 70593 Stuttgart, Germany*

Abstract

The Human Accelerated Region 1, HAR1, is the most rapidly evolving region in the human genome. It is part of two overlapping long non-coding RNAs, has a length of only 118 nucleotides and features 18 human specific changes compared to an ancestral sequence that is extremely well conserved across non-human primates. The human HAR1 forms a stable secondary structure that is strikingly different from the one in chimpanzee as well as other closely related species, again emphasizing its human-specific evolutionary history. This suggests that positive selection has acted to stabilize human-specific features in the ensemble of HAR1 secondary structures. To reveal the order in which the 18 human specific mutations occured, we developed a computational model that evaluates the relative likelihood of evolutionary trajectories as a probabilistic version of a Hamiltonian path problem. The model predicts that the most likely last step in turning the ancestral primate HAR1 into the human HAR1 was exactly the substitution that distinguishes the modern human HAR1 sequence from that of the archaic human Denisovan, providing independent support for our model.

Keywords: Human evolution, computational modeling, dynamic programming, non-coding RNA, secondary structure, data visualisation
2010 MSC: 92-04, 92-08

1. Introduction

Functional innovations at the phenotypic level are eventually the result of genetic changes. While most mutations are (nearly) neutral or even detrimental, occasionally they lead to innovations by affecting the expression pattern of genes or the sequence of the gene product itself. In the latter case, novel molecular and biological functions are thought to be the result of changes in the molecule's structure that in turn changes its interactions and thus its position within cellular networks. As a consequence, the mutant becomes subject to new selection pressures that may lead to rapid adaptive evolution [1]. Such scenarios are extremely difficult to model computationally, because it requires explicit models of structure formation, all relevant interactions in the network, and the function of the network. In the special case of functional RNAs it is at least possible, however, to model the adaptation towards a target structure [2, 3]. In this contribution we ask to what extent the detailed history of recent adaptive evolution can be reconstructed from the knowledge of the current and ancestral structures of a rapidly evolving RNA element.

There are some regions on the genome that have accumulated many human specific changes while remaining constant in other closely related species. These are called human accelerated regions and are candidates for generating human specific traits [4]. The Human Accelerated Region 1 (HAR1) is the region with the most human specific changes in the primary sequence. It will serve here as our paradigmatic example. HAR1 is only 118 nucleotides long and contains 18 human specific substitutions, a lot more than expected from the substitution rate of 0.27 for the other species [4]. HAR1 is located in a pair of overlapping long non-coding RNAs, HAR1F and HAR1R, both

*Corresponding author
Email addresses: bia@bioinf.uni-leipzig.de (Maria Beatriz Walter Costa), choener@bioinf.uni-leipzig.de (Christian Höner zu Siederdissen), dan.tulpan@nrc-cnrc.gc.ca (Dan Tulpan), studla@bioinf.uni-leipzig.de (Peter Stadler), nowick@bioinf.uni-leipzig.de (Katja Nowick)

of which are very specifically expressed in Cajal-Retzius cells between 7 and 19 gestational weeks. This is a crucial period for cortical neuron specification and migration. HAR1F and HAR1R were also reported to be co-expressed with reelin (RELN), a protein involved in the organisation of the laminar cortex of the brain [4]. HAR1R and HAR1F are direct targets of the RE1-silencing transcription factor (REST) in human but not in mouse [5], indicating a change in their regulatory interactions in the human lineage. Considering the highly specific expression pattern of HAR1 in Cajal-Retzius cells, HAR1F and HAR1R may have an important role in the correct organization of the developing human brain.

The secondary structure of HAR1 is conserved among vertebrates with the exception of humans. In humans, HAR1 forms a stable cloverleaf-like structure, that differs from the other species, which was first supported by DMS structure probing [4]. The predicted divergence of the human structure was afterwards confirmed by two independent empirical methods. Chemical and enzymatic probing [6] resulted in a hairpin-like structure for the chimpanzee sequence and a cloverleaf-like structure for the human one. NMR spectroscopy confirmed the chimpanzee model but implied that the human structure contains two small hairpin domains connected by a flexible middle region [7].

Surprisingly, all 18 human specific substitutions replace an ancestral `A` or `T` with a `G` or `C`. In general `G-C` interactions are energetically more favorable than `A-T`, so that the substitutions are expected to lead to an overall stabilization of the RNA structure, which is in apparent contradiction with the empirically observed weakening of the ancestral hairpin structure in favor of a much more flexible human structure. A closer inspection, however, shows that the ancestral hairpin structure is only marginally stable and the human-specific substitutions have lead to a strategic stabilization of two of the three hairpins of the predicted cloverleaf structures, Fig. 1 (a,c). This is clear when comparing the minimum free energy (MFE) structures of the ancestral and human, and especially when comparing the centroid structures, Fig. 1 (b,d). While MFE structures are the most stable in the ensemble and require more energy to be broken, they are not the only ones occuring in the cell. The centroid structure as a representative of the ensemble has smallest average base pair distance to all alternatives. The ancestral centroid is much less stable than the human centroid, meaning that the whole ensemble for human is structurally closer to its MFE.

To understand how the human specific structure of HAR1 evolved, we developed a computational model. We revealed a strong reshaping of the HAR1 structure and the most likely last change of the 18 substitutions. Interestingly, all of the substitutions seem to drive HAR1 to a more stable ensemble. Moreover, the last predicted mutation separates the structures of the modern human from the archaic denisovan human and recreates a stem that had been weakened on the evolutionary path from chimpanzee to denisovan.

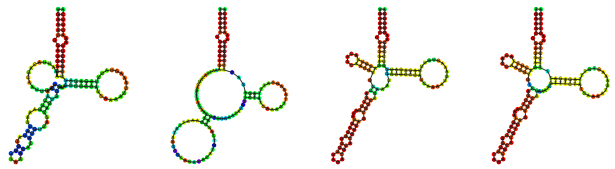

Figure 1: Chimpanzee and human HAR1 structures as start and end points of evolution simulation. Chimpanzee (a) minimum free energy (MFE) and (b) centroid structures, considered in our model as the evolutionary starting point and human (c) minimum free energy and (d) centroid as the end point in our model. Nucleotides are colored according to their pairing frequency in the ensemble. Base pairs in shades of red occur in $\geq 90\%$ of all structures in the ensemble, while green to yellow denote increasing probabilities $\geq 50\%$. For unpaired nucleotides, colors toward red denote increasing unpairedness. The centroid structures contain base pairings that occur in more than 50% of the structures of each ensemble.

2. The Model

Reconstructing the history of HAR1 in the human lineage is an instance of a general optimization problem. Given a (conserved) ancestral sequence x, a secondary structure $S(x)$ and a corresponding derived extant pair y and $S(y)$, we implicitly know the set X of adaptive substitutions from the alignment of sequences and x and y. What we are interested in is their temporal ordering. Each possible evolutionary path is therefore a permutation π of X. The extant structure $S(y)$ serves as proxy for the selection target. This allows us to use a measure of structural distance to $S(y)$ as a proxy for fitness, i.e., substitutions that reduce the distance to $S(y)$ can be thought as adaptive and are quickly fixated, while substitutions that increase the distance to $S(y)$ are discouraged. Thus $f(u) = -d(S(u), S(y))$ serves as a fitness function. The fitness cost of an evolutionary path π is then

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

where the sum only includes those steps in which the fitness decreases, i.e., the distance to the target increases. The likelihood of a path π decreases exponentially with its fitness cost, i.e.,

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

where the "inverse temperature" β is a scaling parameter measuring the stringency of selection, and Z is a normalization factor.

There is some freedom in modelling the distance. In this contribution, we use the energies of centroid and MFE structures as well as the base pair distances for MFE and centroid structures. Conceivable other choices include variance or Kulback-Leibler distances measured for the base pairing probabilities, as used e.g. in `RNAsnp` [8].

Finding the most likely permutation, i.e., the one that minimizes $f(\pi)$ amounts to computing the Hamiltonian

path from x to y with minimal total cost. This problem can be solved by a well-known exponential time dynamic programming algorithm, which is applicable in practice for a problem size of $n = 18$ mutations as in the case of HAR1. As shown in [9], the use of ideas from algebraic dynamic programming makes it possible to also compute the posterior probabilities P_{ij} for two mutations i and j to be consecutive along a path. Using this matrix of posterior probabilities as scoring function the same recursive algorithm can be used to compute the Maximum Expected Accuracy path.

The model also makes it simple to compute the probabilities π_{ij} that the sequence of substitutions started with mutation i and terminated with position j. These quantities give access to the probability that $\pi_j = \sum_i \pi_{ij}$ of mutation j being the last one.

3. Material and Methods

The HAR1 sequences were retrieved from the NCBI nucleotide database, including human, chimpanzee, and the archaic human denisovan. We use the chimpanzee sequence as the ancestral sequence because it has been extremely conserved among non-human primates [4]. We restrict ourselves to the 18 observed substitutions separating the human and chimpanzee sequence [4]. Since we assume that this rapid evolution was largely adaptive, back-mutations can be disregarded. We consider all subsets of the 18 observed substitutions as potential intermediates.

Both minimum energy secondary structure and base pairing probabilities were computed using the `ViennaRNA` package [10] (Version 2.3.3) with the standard Turner energy model for RNA secondary structures (dangling model `-d2`, and no lonely base pairs (`--noLP`)). These predictions were used to determine the structural and energetic differences between potentially consecutive mutations.

3.1. Visualisation of RNA secondary structures

Since the comparison of the Boltzmann ensemble of two structures is more informative and yields more detailed insights than the comparison of MFE or centroid structures alone, we used superpositions of base pairing probability dot-plots with different colors for each species[1]. While the combined dot-plots are useful to obtain a quick overview, they can be difficult to interpret.

The `CS`^{2}`-UPlot` [11] provides an alternative visualization representing the two main information components of an RNA secondary structure in two concentric graphical layers: the RNA sequence and the MFE and alternative base pairing possibilities. It uses `Circos` version 0.69-3 [12] and Perl version 5.022001 and combines base pairings with dot-plot values in a single graphical representation. It has the advantages of better highlighting similarities and differences than dot-plots and providing with the circular diagrams a graphical representation that is more intuitive to biologists.

[1] available in `http://hackage.haskell.org/package/MutationOrder`

3.2. Diversity of HAR1 in human populations

To further investigate variability of the HAR1 region in human populations we retrieved all reported SNPs in the 118 base pair HAR1 region using the ENSEMBL `Data Slicer` from the data set provided by the 1000 Genomes project [13]. We also checked the human genome for possible paralogs of the HAR1 region using Infernal tool [14], with no such paralogs being identified.

4. Results

4.1. Comparison between ancestral, archaic and modern human structures

Centroid structures typically yield a better impression of the consensus of the equilibrium ensemble of secondary structures than the MFE structure. The centroid of the ancestral structure has a much more flexible space for base pairing, Fig. 1, and can form both a hairpin and a cloverleaf structure, which has been reported before [6]. In contrast, the human sequence has a more constrained set of energetically low-lying structures, and hence exhibits a better defined, more stable cloverleaf-shaped structure, Fig. 1. This is consistent with the expectation of the increased GC content of the human sequence. We conclude that stabilization of the cloverleaf structure is a plausible model for how selection acted at the level of RNA structure.

The denisovan HAR1 differs from its modern human counterpart only by a `T` instead of a `C` in position 47. The archaic human structure shares small stems with modern human, which are only slightly shifted. However, the structural space of denisova is still more diverse, featuring more base pairs that are less well-defined than in modern human, thus appearing more similar to the ancestral state, see Fig. 2. A corresponding dot-plot representation is shown in the Appendix.

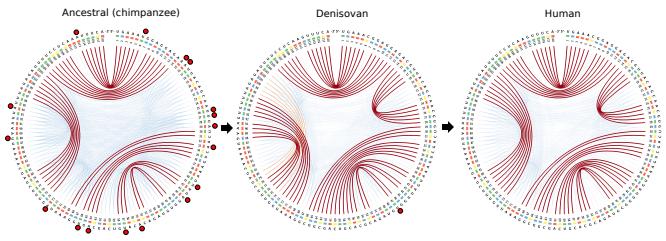

Figure 2: Comparison of the ancestral (left), Denisovan (middle) and modern human (right) ensembles of HAR1 secondary structures. The plots contain the sequence on the outer layer, the MFE base pairings in red lines and alternative base pairing possibilities in orange and blue, with orange base pairings being more likely than the blue ones. Mutations in relation to the modern human sequence are indicated by red circles.

In addition to uncovering the evolutionay path from the ancestral primate to the human version of HAR1, we also asked whether there are variants of HAR1 among modern humans. The 1000 Genomes Project [13] reports three SNPs for HAR1: C47T, C52T and G113C, each occurring in less than 1% of the surveyed populations. The variant rs374630364, corresponding to HAR1 position 47, is present in South and East Asian populations, while this variant was not detected in African, American or European populations. This is interesting, since Denisovans lived in an area ranging from Siberia to South East Asia and have inbred with modern humans who lived in the same area [15]. It thus provides independent support for our suggestion that position 47 was one of the very last steps in the evolutionary reshaping of the HAR1 structure, stabilizing a small hairpin in the human centroid structure, Fig. 2.

The variant rs183960348, which is located at position 52, is exclusive to American and African populations, while the variant rs54438677, which is located at position 113, is exclusive to Asian populations. All three variants decrease the stability of the very stable wildtype human ensemble, with the variant at position 52 having the strongest impact which can be seen especially on the centroid structure, Fig. 5. The MFE of variant 52 however still folds into a cloverleaf format, Fig. 5. Despite these impacts on the structure, no associations to diseases were reported for any af these variants in the DisGeNET database [16].

4.2. Reconstructing the Evolution of HAR1

We found qualitatively comparable features of the most likely pathways, even when using different models for the structure distances underlying the fitness model for evolutionary paths. Table 1 gives the number of co-optimal solutions when all stability-gaining mutations are assumed equally likely and mutations that decrease the stability of the structure are scored according to different criteria. We count the number of co-optimal paths for four different fitness models. The *mfe* and *centroid* models use the energy gain or loss between two structures and are essentially $\max(0, \Delta_E)$ while the *pairdist* models naturally model only losses. The basepair distance between the two structures is at least 0 and as high as the number of base pairs in the two structures. In all variants of the model there are large numbers of co-optimal permutations, suggesting that evolutionary paths along with monotonically increasing fitness were easy to find. It is particularly easy to find a large number of co-optimal paths using the MFE energy as fitness function, where all mutational steps increase the fitness. Not surprisingly we find that there are more paths that stabilize the minimum energy structure than paths that keep the centroid stable – and thereby a majority of the ensemble of structures stable.

Table 1: Frequency of co-optimal permutations of the 18 human-specific substitutions in HAR1 for different choices of the distance function in equ.(1). The first two distance functions penalize increases in the folding energy computed for the minimum energy and centroid structures, resp. The number of base pairs different between the structure at each step and the human target is used as an alternative model. The third colum gives the fraction of co-optimal paths among the 18 permutations.

fitness model	# co-optimal path	fraction
minimum energy	3 931 510 681 533	6.14×10^{-4}
centroid energy	1 615 195 878	2.52×10^{-7}
m.e. pairs	17 338 903 092	2.71×10^{-6}
centroid pairs	2 239218	3.50×10^{-10}

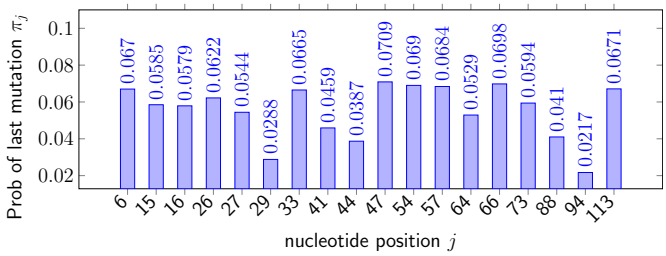

Figure 3: Probability π_j for each mutation to be the last mutational event with $\beta = 1.0$. Nucleotide position 47 is the Denisovan-human mutation and has the highest posterior probability. Position 54 has $\geq 50\%$ posterior probability to pair with Position 47. The base pairs 44-57 and 33-66 are part of the same hairpin in human with $\geq 50\%$ probability.

This large degree of redundancy, with many equivalent evolutionary trajectories, leaves no dramatic differences in the probabilities of last mutations in the sequence. Nevertheless, it is still interesting to note that the `T` to `C` transition in position 47, which separates Denisovan from modern human, is predicted as the most likely last step from our model, Fig. 3. Importantly, as the model only has information of the ancestral and the final states, but does not has information on the denisovan state as a likely intermediate, we can interpret this result as a direct support for our modelling approach.

The large number of feasible paths makes it impossible to analyze them individually. Instead, we provide further summary statistics in the form of edge probability plots. These plots identify likely neighbors in the chronological ordering of the mutational events. Fig. 4 summarizes these probabilities for the pair distance fitness model based on centroid structures. In particular, the base pair $54 \rightarrow 47$ is recognized to have high probability in this chronological order. The `A` to `G` substitution at position 54 furthermore sets the stage for the `C`/`T` polymorphism, which, despite stabilizing the structure, maintains a similar ensemble of structures.

5. Conclusion

Guided by HAR1 as the paradigmatic application, we have introduced here a suite of tools to investigate evolutionary trajectories of secondary structures in detail. We introduced a convenient visualization method for structural ensembles that enables intuitive insights into evo-

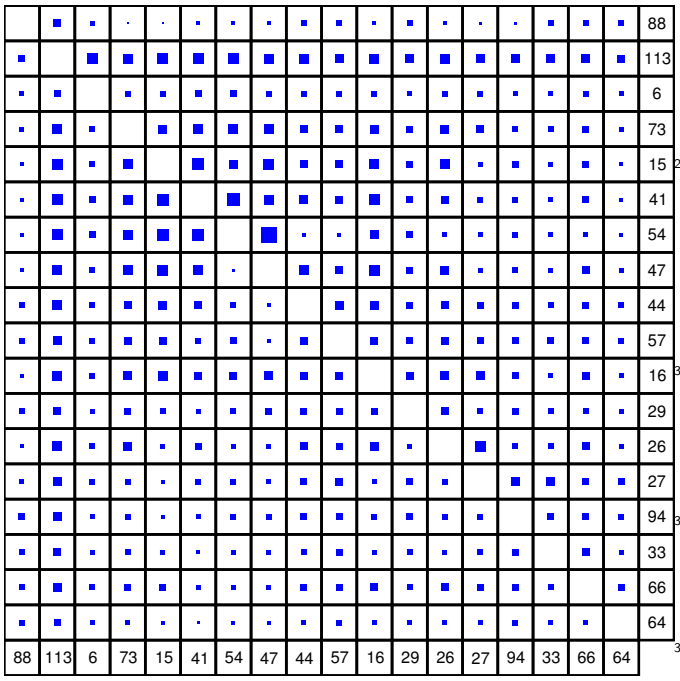

Figure 4: Probability P_{kl} of mutation k (row) to be followed by mutation l (column) following mutation k (row) using the pair distance fitness function on centroid structures. The pseudo-temperature is set to $\beta = 1.0$. Mutations are arranged in their order of appearance in the MEA path. The boxes are scaled as $1/(1 - \log P_{kl})$ to highlight the uncertainties involved in determing the most likely evolutionary path. We note the high posterior probability for the sequence $54 \rightarrow 47$. The two nucleotides form a `GC` base pair in the human centroid structure that was produced as the last step in the evolutionary trajectory. The best weight of a trajectory ending with mutation 47 is about 1/8 of the trajectory shown here.

lutionary changes of secondary structures at high resolution. A dynamic programming method makes it possible to compile exact stastistics over possible evolutionary trajectories. Despite its exponential runtime the algorithmic approach is efficient enough to handle systems with up to at least 20 substitutions, which includes at least all moderate size structured RNAs. The approach proposed here can be used to test whether rapid changes are associated with alterered selection pressures for novel RNA structures. Although beyond the scope of this contribution, the same type of model can also be used to evaluate adaptive evolution of protein structures – all that is needed is a distance measure to a target structure that correlates well with the actual fitness effects, i.e., that fitness is largely determined by structure.

A computational model assuming only selection against increasing divergence from the modern human target structure correctly identifies the single difference between human and denisovan HAR1 as the most likely last step along the evolutionary trajectories. With that, we have shown that the rapid evolution of HAR1 from the human-chimp ancestor to the modern human sequence can be explained by directional selection for the more stable, modern secondary structure.

6. Acknowledgements

This work was supported by CNPq/Science without Borders (MBWC), the Volkswagen Foundation within the initiative 'Evolutionary Biology' (KN), and the Deutsche Forschungsgemeinschaft as part of the SPP 1738 (KN & PFS).

References

[1] M. D. Laubichler, P. F. Stadler, S. J. Prohaska, K. Nowick, The relativity of biological function, Th. Biosci. 143 (2015) 143–147. doi:10.1007/s12064-015-0215-5.

[2] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, From sequences to shapes and back: A case study in RNA secondary structures, Proc. Roy. Soc. Lond. B 255 (1994) 279–284.

[3] M. A. Huynen, P. F. Stadler, W. Fontana, Smoothness within ruggedness: the role of neutrality in adaptation, Proc. Natl. Acad. Sci. (USA) 93 (1996) 397–401.

[4] K. S. Pollard, S. R. Salama, N. Lambert, M.-A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, et al., An RNA gene expressed during cortical development evolved rapidly in humans, Nature 443 (7108) (2006) 167–172.

[5] R. Johnson, N. Richter, R. Jauch, P. M. Gaughwin, C. Zuccato, E. Cattaneo, L. W. Stanton, Human accelerated region 1 noncoding RNA is repressed by REST in Huntington's disease, Physiological genomics 41 (3) (2010) 269–274.

[6] A. Beniaminov, E. Westhof, A. Krol, Distinctive structures between chimpanzee and human in a brain noncoding RNA, RNA 14 (7) (2008) 1270–1275.

[7] M. Ziegeler, M. Cevec, C. Richter, H. Schwalbe, NMR studies of HAR1 RNA secondary structures reveal conformational dynamics in the human RNA, ChemBioChem 13 (14) (2012) 2100–2112.

[8] R. Sabarinathan, H. Tafer, S. E. Seemann, I. L. Hofacker, P. F. Stadler, J. Gorodkin, `RNAsnp`: Efficient detection of local RNA secondary structure changes induced by SNPs, Hum. Mut. 34 (2013) 546–556.

[9] C. Höner zu Siederdissen, S. J. Prohaska, P. F. Stadler, Algebraic dynamic programming over general data structures, BMC Bioinformatics 16. doi:10.1186/1471-2105-16-S19-S2.

[10] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA package 2.0, Algorithms for Molecular Biology 6 (1) (2011) 1.

[11] D. Tulpan, The circular secondary structure uncertainty plot (CS2-UPlot) -visualizing RNA secondary structure with base pair binding (2015).
URL `http://nrc-ca.academia.edu/DanTulpan`

[12] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: an information aesthetic for comparative genomics, Genome research 19 (9) (2009) 1639–1645.

[13] . G. P. Consortium, et al., A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.

[14] E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster rna homology searches, Bioinformatics 29 (22) (2013) 2933–2935.

[15] M. Meyer, M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. Schraiber, F. Jay, K. Prüfer, C. de, Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic Denisovan individual, Science 338 (2012) 222–226. doi:10.1126/science.1224344.

[16] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L. I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res 45 (2017) D833–D839. doi:10.1093/nar/gkw943.

Appendix

A1 Secondary Structures of Human HAR1 Variants

Variations of HAR1 within the human species are rare and are found in less than 1% of human populations. Three variants of HAR1 have been reported to date and all cause changes to the wildtype structure, Fig. 5.

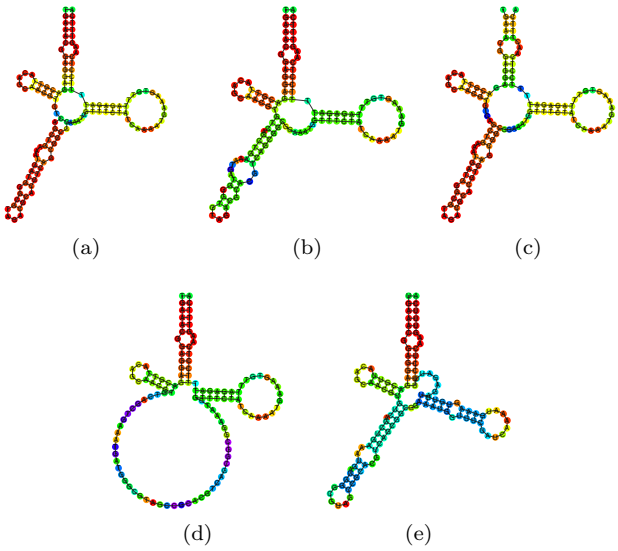

Figure 5: Wildtype and human variations of the HAR1 structure. (a) Wildtype centroid, (b) variant rs374630364 C47T centroid (the same as Denisovan), (c) variant rs544386774 G113C centroid, (d) variant rs183960348 C52T centroid and (e) variant rs183960348 C52T MFE.

A2 Comparative Dot-Plots

Comparative dot-plots provide an alternative visualization of differences between the structural ensemble of two closely related sequences. The upper right triangle shows the base pairing probabilities in two colors, one for each input sequence. The lower left triangle displays the base pairs of the minimum energy structure.

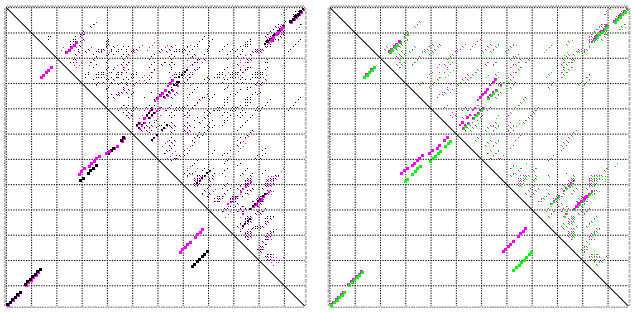

Figure 6: Base pairing patterns of the ancestral (black), denisovan (magenta) and modern human (green) HAR1 sequences. The plots show the large difference between ancestral and denisovan structures (left) and the more subtle differences between denisovan and the modern human structure. Interestingly the 3'-most stem coincides in modern human and the ancestral state, but is shifted in denisovan. On the other hand, the 5' part of the structure is already close to modern human in the denisovan structure.