

Computational Analysis of the AK092578 Locus

Stefan Washietl, Dominic Rose, Peter F. Stadler

February 27, 2009

1 Introduction

Genome browser annotation of the locus of interest (hg18.chr4:166,017,786-166,038,125) shows an unannotated cDNA with accession number AK092578. To analyze whether this transcript is likely to be a non-coding or coding RNA, it was analyzed with *RNAz* and *RNAcode*. In the following we summarize the results of the computational analysis.

2 Analysis of the AK092578 for structured RNA

An *RNAz*-based (v1.0, [10]) search for structured RNAs within the 32k-region hg18.chr4:166,012,000-166,043,999 yields only five ncRNA candidates. The input for *RNAz* was generated with the *mafsInRegion* program, which is part of Kent's UCSC-source-tree. For this locus, *mafsInRegion* reports 260 conserved sequence blocks within the UCSC 28-way vertebrate alignments. Further pre-processing of those 260 alignments was done using *rnazWindows.pl*, a part of the *RNAz* package, which selects up to six of the aligned sequences for screening by *RNAz* in such a way that their pairwise sequence similarity is optimized for the performance of *RNAz*. This is necessary because *RNAz* at present can evaluate alignments of no more than 6 species. The tool also prepares alignment windows of appropriate length and sequence composition matching the *RNAz* training scope. Actually only two of the five *RNAz* hits have an acceptable classification probability. The coding potential calculator (CPC, [8]) scores all *RNAz* hits as non-coding regions, but it votes for coding when it scores the complete AK092578 locus. Details of the *RNAz* hits are given in table 1 and predicted structures are available at <http://www.bioinf.uni-leipzig.de/data/REVERSADE/>.

Tab. 1. *RNAz* summary

name	start	stop	probability	annotation
locus1	166,023,781	166,023,848	0.563	intron
locus2	166,026,479	166,026,584	0.542	intron, khaito
locus3	166,027,353	166,027,407	0.975	intron
locus4	166,035,549	166,035,667	0.519	intron
locus5	166,037,681	166,037,739	0.851	exon, 3'UTR

The UCSC browser provides no *EvoFold* prediction for the whole locus.

In order to check whether there is any indication for extensive production of short RNAs from this local, we extracted the short reads that map to the region from a collection of brain libraries (produced by Phillip Khaitovich in Shanghai). 39 entries map to the AK092578, 79 to the entire 32k-region. Only one of them overlaps with *RNAz*-locus2.

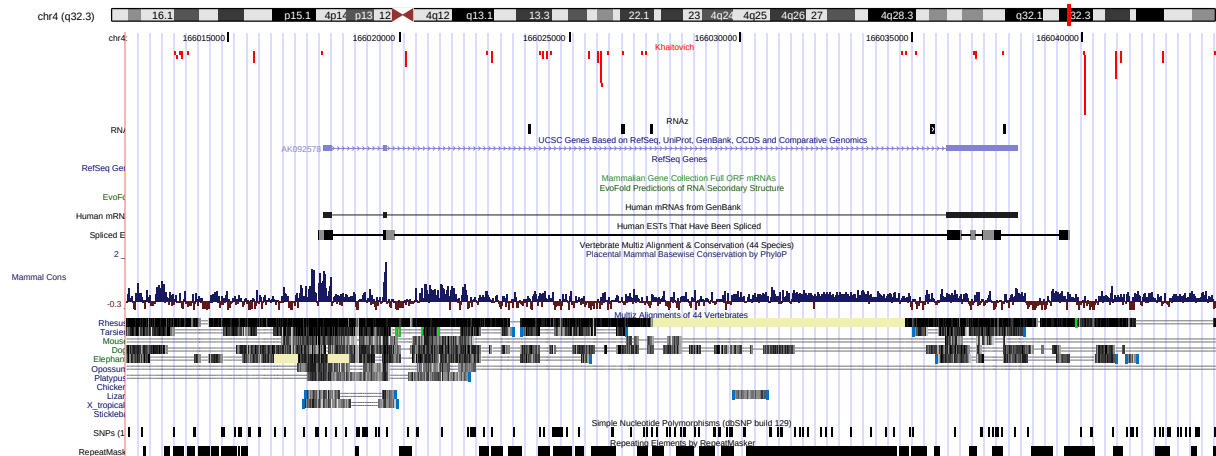


Fig. 1. AK092578

Overall, there is no compelling evidence for functional structured elements, with the possible exception of **locus5**, which is located in the 3' UTR of the human AK092578 RNA.

3 Analysis of the locus for protein coding potential

RNAcode (Washietl *et al.*, in preparation) analyzes mutation patterns in multiple alignments for synonymous and conservative amino acid substitution, typical signatures of protein coding genes. We used the multiz 17-way alignments available at UCSC. Two coding exons with compelling P -values were identified ($P < 10^{-2}$ and $P < 10^{-5}$). These coding predictions were confirmed by a prediction of CONTRAST [6] that predicts a short protein in this region. The predicted cDNA sequence reads:

```
>cDNA
atgagatttcagcaattcctttttgcattttttatttttattatg
agtcttctccttatcagcggacagagaccagtttaatttgaccatg
agaagaaaactgcgcaaacacaattgccttcagaggagatgtatg
cctctccatttcacgagtaccctttccctga
```

It can be translated into a short protein of length 54:

```
>reversade
MRFQQFLFAFFIFIMSLLLISGQRPVNLTMRRKLRKHNC LQRRCMP LHSRVFPF
```

4 Sequence analysis of the candidate protein

4.1 Cellular localization

The predicted protein was analyzed for a signal peptide using the SignalP program [4]. Using both the Neural network model and the Hidden Markov Model, a strong signal peptide is predicted with the most likely cleavage site at between pos. 22 and 23 (...ISG|QR...).

These results clearly indicate that the predicted protein is likely to be exported. The sequence was also analyzed for a transmembrane (TM) domain. First, it is unlikely that it is a TM protein because

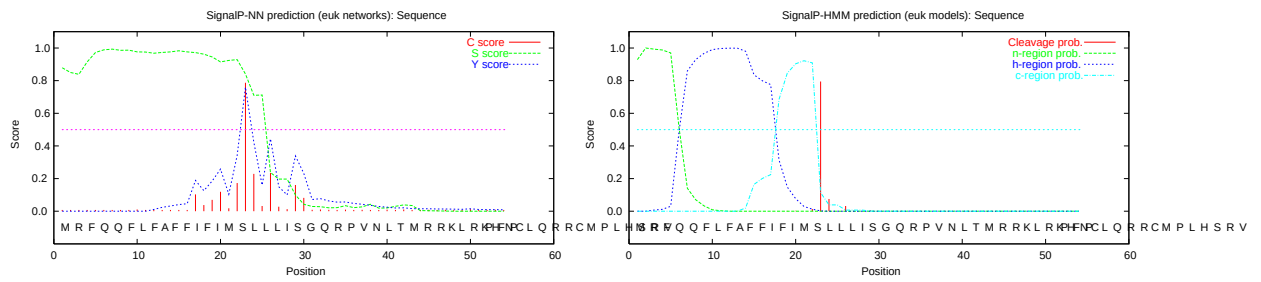


Fig. 2. SignalP output

of its length (a TM-domain would span most of the this very short protein). Second, we did not find a convincing TM domain using TMHMM [4]. Only the hydrophobic part of the Signal Peptide is predicted as potential TMHMM. However, SignalP does not predict a “signal anchor” (probability only $P=0.071$). Thus it appears likely that the signal peptide is cleaved and the protein is segregated.

4.2 Homology search

To identify a potential function of the protein we tried to find homologues in databases. A BlastP against the non redundant NCBI protein database yielded only two hits:

```
>gb|EDL28686.1| mCG22896, isoform CRA_a [Mus musculus]
Length=82
```

```
Score = 77.4 bits (189), Expect = 3e-13, Method: Compositional matrix adjust.
Identities = 37/54 (68%), Positives = 42/54 (77%), Gaps = 0/54 (0%)
```

```
Query 1 MRFQQFLFAFFIFIMSLLLISGQRPVNLTMRRLKLRKHNCQRRRCMPLHSRVFPF 54
        MRFQ + FFIF MSL L IS Q+PVN RRKL +HNC +RRC+PLHSRVFPF
Sbjct 29 MRFQPLFWVFFIFAMSLLFISEQKPVNFPRRRKLYRHNCFRRCIPLHSRVFPF 82
```

```
>ref|XP_001233480.1| UniGene infoGene info PREDICTED: hypothetical protein [Gallus gallus]
Length=54
```

```
GENE ID: 770154 LOC770154 |hypothetical protein LOC770154 [Gallus gallus]
```

```
Score = 58.2 bits (139), Expect = 2e-07, Method: Compositional matrix adjust.
Identities = 25/35 (71%), Positives = 28/35 (80%), Gaps = 0/35 (0%)
```

```
Query 20 ISGQRPVNLTMRRLKLRKHNCQRRRCMPLHSRVFPF 54
        + QRP NL +RRKL +HNC RRCMPLHSRVFPF
Sbjct 20 AAAQRPANLALRRKLHRHNCSHRRRCMPLHSRVFPF 54
```

A PSI blast search using the profile of the three sequences could not uncover new remote homologues.

We also tried to identify known protein domains in the predicted protein. Neither NCBI conserved domain search [9] nor PFAM (Pfam A+B, [5]) found any significant domains.

4.3 tblastn search in the DB-EST

A tblast search against RefSeq completed the above hit in the chicken to yield

```
>ref|XM_001233479.1| UniGene infoGene info PREDICTED: Gallus gallus
hypothetical protein LOC770154 (LOC770154),

Query 1      MRFQQFLFAFFIFIMSLLLISGQRPVNLTMRRKLRKHNCQRRCMPLHSRVFPF 54
            MR ++ L   F+ ++SLL + QRP NL +RRKL +HNC RRCMPLHSRVFPF
Sbjct 44     MRLRRLLCVVFLLLVSLLPAAAQRPANLALRRKLHRHNCSHRRCMPLHSRVFPF 205
```

A search of the NCBI EST database with standard parameters and the predicted human peptide as query resulted in several dozens well-conserved hits across the entire vertebrate clade. An alignment is included as Figure 3. This alignment also contains predicted homologs in the armadillo (*Dasypus novemcinctus*) and the elephant shark, which were obtained from blast searches and subsequent manual annotation of the splice site.



Fig. 3. ClustalW alignment of tblastn hits

4.4 Short sequence motifs and statistical analysis

The sequence was searched for PROSITE [7] motifs. We only found two instances of the very unspecific Asn-Glycosylation and PKC-phosphorylation motifs, which most likely do not have biological relevance.

```
PS00001 ASN_GLYCOSYLATION N-glycosylation site :
27 - 30: NLTM
```

```
PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site :
29 - 31:
```

We also analyzed statistical properties of the primary sequences using SAPS [2] but did not find any statistically unusual properties.

4.5 Structure prediction

Secondary structure prediction using JPred [3] server predicts one short helix outside of the potentially cleaved signal peptide.

```

                                1-----11-----21-----31-----41-----51--
OrigSeq      : MRFQQFLFAFFIFIMSLLLISGQRPVNLTMRRKLRKHNCLQRRCMPLHSRVFPFP

Jnet         : --HHHHHHHHHHHHHHHHHHHH-----HHHHHHHHHH-----
jhmm         : --HHHHHHHHHHHHHHHHHHHH-----HHHHHHHHHH-----
jpssm        : --HHHHHHHHHHHHHHHHHHHH-----HHHHHHHHHH-----
```

Not unexpectedly for this short protein without any homologues, Phyre [1] fold recognition did not give any useful results.

5 Conclusion

Mostly likely, the function product of the AK092578 locus is the 54nt peptide described above. Even though in particular the C-terminal side of the peptide is very well conserved across vertebrates, it shows no homology to any other protein or peptide family. The short length and the predicted extracellular localization suggest a potential function as peptide hormone.

References

- [1] R M Bennett-Lovsey, A D Herbert, M J Sternberg, and L A Kelley. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program phyre. *Proteins*, 70(3):611–25, Feb 2008.
- [2] V Brendel, P Bucher, I R Nourbakhsh, B E Blaisdell, and S Karlin. Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A*, 89(6):2002–6, Mar 1992.
- [3] C Cole, J D Barber, and G J Barton. The jpred 3 secondary structure prediction server. *Nucleic Acids Res*, 36(Web Server issue):W197–201, Jul 2008.
- [4] O Emanuelsson, S Brunak, G von Heijne, and H Nielsen. Locating proteins in the cell using targetp, signalp and related tools. *Nat Protoc*, 2(4):953–71, 2007.
- [5] R D Finn, J Tate, J Mistry, P C Coghill, S J Sammut, H R Hotz, G Ceric, K Forslund, S R Eddy, E L Sonnhammer, and A Bateman. The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–8, Jan 2008.
- [6] S S Gross, C B Do, M Sirota, and S Batzoglu. Contrast: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol*, 8(12):R269, 2007.
- [7] S Hunter, R Apweiler, T K Attwood, A Bairoch, A Bateman, D Binns, P Bork, U Das, L Daugherty, L Duquenne, R D Finn, J Gough, D Haft, N Hulo, D Kahn, E Kelly, A Laugraud, I Letunic, D Lonsdale, R Lopez, M Madera, J Maslen, C McAnulla, J McDowall, J Mistry, A Mitchell, N Mulder, D Natale, C Orengo, A F Quinn, J D Selengut, C J Sigrist, M Thimma, P D Thomas, F Valentin, D Wilson, C H Wu, and C Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–5, Jan 2009.

- [8] L. Kong, Y. Zhang, Z. Ye, X. Liu, S. Zhao, L. Wei, and G. Gao. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 35(Web Server issue):W345–W349, Jul 2007.
- [9] A Marchler-Bauer, J B Anderson, F Chitsaz, M K Derbyshire, C DeWeese-Scott, J H Fong, L Y Geer, R C Geer, N R Gonzales, M Gwadz, S He, D I Hurwitz, J D Jackson, Z Ke, C J Lanczycki, C A Liebert, C Liu, F Lu, S Lu, G H Marchler, M Mullokandov, J S Song, A Tasneem, N Thanki, R A Yamashita, D Zhang, N Zhang, and S H Bryant. Cdd: specific functional annotation with the conserved domain database. *Nucleic Acids Res*, 37(Database issue):D205–10, Jan 2009.
- [10] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA*, 102:2454–2459, 2005.