Statistical Alignment of Retropseudogenes: Supplementary Material

Miklós Csűrös* István Miklós†

March 16, 2005

This supplementary material to the paper contains the following information. Section 1 describes an initial set of recurrences referred to in the main text. Section 2 plots constraints on a molecular clock for pseudogenes. Section 3 shows conservation plots illustrating our claims on conserved syntenies around cytochrome *c* pseudogenes. We also give here the sequence preHCS that was used to compute divergence times for Class II pseudogenes.

```
>preHCS human cytochrome c hypothetical precursor CDS (somewhere before OWM)
ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTGTTCAGAAGTGTGCCCAGTGCCACACCGTTGAAAAGG
GAGGCAAGCACAAGACTGGGCCTAATCTCCATGGTCTCTTCGGGCGGAAGACAGGTCAGGCCGTTGGATT
CTCTTACACAGATGCCAATAAGAACAAAGGCATCACCTGGGGAGAGGATACACTGATGGAGTATTTGGAG
AATCCCAAGAAGTACATCCCTGGAACAAAAATGATCTTTGCCGGCATTAAGAAGAAGGCAGAAAGGGCAG
ACTTGATAGCTTATCTCAAAAAAGCTACTAATGAGTAA
```

1 Initial recurrences

For the sake of completeness, we list here the initial set of recurrences. For simplicity, these recurrences do not employ the codon-substitution model, but consider the gene sequence as a sequence of i. i. d. nucleotides. The probability $p(g, h)$ is the probability of observing homologous characters g in G and h in ΨG; $p(g)$ is the probability of observing g in G as the result of substitutions of an ancestral character or insertion; $p_\Psi(h)$ is the probability of observing h in ΨG as the result of substitutions of an ancestral character or insertion.

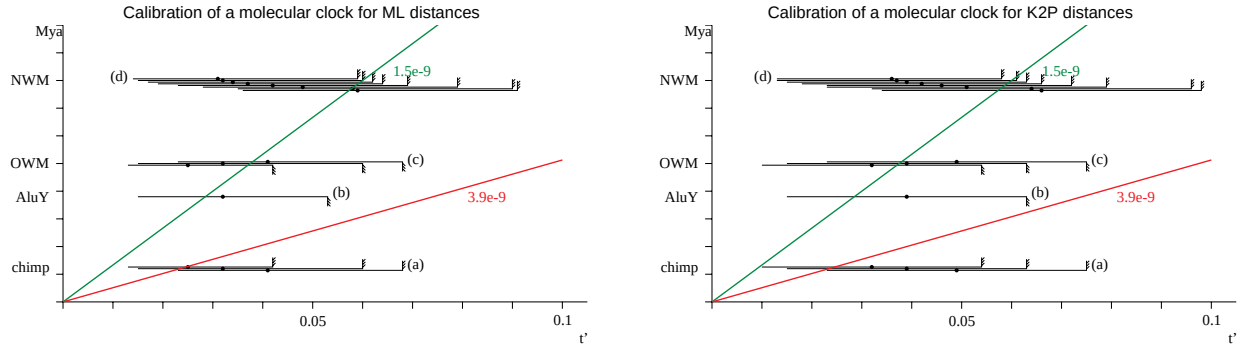
*Department of Computer Science and Operations Research, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada. E-mail: csuros@iro.umontreal.ca

†Department of Plant Taxonomy and Ecology, Eötvös Lóránd University, 1117 Budapest, Pázmány Péter Sétány 1/c, Hungary. E-mail: miklosi@ramet.elte.hu

$$
\begin{aligned}
\mathsf{SH1}(i,j) &= \mathsf{A}(i-1,j-1)\gamma H H_\Psi p(g^i,h^j)\\
\mathsf{SE1}(i,j) &= \mathsf{A}(i-1,j)\gamma H p(g^i)\\
\mathsf{SN1}(i,j) &= \Big(B_\Psi\big(\mathsf{SH1}(i,j-1)+\mathsf{SN1}(i,j-1)\big)+N_\Psi\mathsf{SE1}(i,j-1)\Big)p_\Psi(h^j)\\
\mathsf{SH2}(i,j) &= \big(\mathsf{SH1}(i-1,j-1)+\mathsf{SN1}(i-1,j-1)\\
&\quad +E_\Psi\mathsf{SE1}(i-1,j-1)\big)H_\Psi p(g^i,h^j)\\
\mathsf{SE2}(i,j) &= \big(\mathsf{SH1}(i-1,j)+\mathsf{SN1}(i-1,j)+E_\Psi\mathsf{SE1}(i-1,j)\big)p(g^i)\\
\mathsf{SN2}(i,j) &= \Big(B_\Psi\big(\mathsf{SH2}(i,j-1)+\mathsf{SN2}(i,j-1)\big)+N_\Psi\mathsf{SE2}(i,j-1)\Big)p_\Psi(h^j)\\
\mathsf{SH3}(i,j) &= \big(\mathsf{SH2}(i-1,j-1)+\mathsf{SN2}(i-1,j-1)\\
&\quad +E_\Psi\mathsf{SE2}(i-1,j-1)\big)H_\Psi p(g^i,h^j)\\
\mathsf{SE3}(i,j) &= \big(\mathsf{SH2}(i-1,j)+\mathsf{SN2}(i-1,j)+E_\Psi\mathsf{SE2}(i-1,j)\big)p(g^i)\\
\mathsf{SN3}(i,j) &= \Big(B_\Psi\big(\mathsf{SH3}(i,j-1)+\mathsf{SN3}(i,j-1)\big)+N_\Psi\mathsf{SE3}(i,j-1)\Big)p_\Psi(h^j)\\
\mathsf{EH1}(i,j) &= A(i,j-1)\gamma H_\Psi p_\Psi(h^j)\\
\mathsf{EE1}(i,j) &= A(i,j)\gamma\\
\mathsf{EN1}(i.j) &= \Big(B_\Psi\big(\mathsf{EH1}(i,j-1)+\mathsf{EN1}(i,j-1)\big)+N_\Psi\mathsf{EE1}(i,j-1)\Big)p_\Psi(h^j)\\
\mathsf{EH2}(i,j) &= \big(\mathsf{EH1}(i,j-1)+\mathsf{EN1}(i,j-1)+E_\Psi\mathsf{EE1}(i,j-1)\big)H_\Psi p_\Psi(h^j)\\
\mathsf{EE2}(i,j) &= \mathsf{EH1}(i,j)+\mathsf{EN1}(i,j)\\
\mathsf{EA2}(i,j) &= E_\Psi\mathsf{EE1}(i,j)\\
\mathsf{EN2}(i,j) &= \Big(B_\Psi\big(\mathsf{EH2}(i,j-1)+\mathsf{EN2}(i,j-1)\big)\\
&\quad +N_\Psi\big(\mathsf{EE2}(i,j-1)+\mathsf{EA2}(i,j-1)\big)\Big)p_\Psi(h^j)\\
\mathsf{EH3}(i,j) &= \Big(\mathsf{EH2}(i,j-1)+\mathsf{EN2}(i,j-1)\\
&\quad +E_\Psi\big(\mathsf{EE2}(i,j-1)+\mathsf{EA2}(i,j-1)\big)\Big)H_\Psi p_\Psi(h^j)\\
\mathsf{EE3}(i,j) &= \mathsf{EH2}(i,j)+\mathsf{EN2}(i,j)+E_\Psi\mathsf{EE2}(i,j)\\
\mathsf{EA3}(i,j) &= E_\Psi\mathsf{EA2}(i,j)\\
\mathsf{EN3}(i,j) &= \Big(B_\Psi\big(\mathsf{EH3}(i,j-1)+\mathsf{EN3}(i,j-1)\big)\\
&\quad +N_\Psi\big(\mathsf{EE3}(i,j-1)+\mathsf{EA3}(i,j-1)\big)\Big)p_\Psi(h^j)\\
\mathsf{NNN}(i,j) &= p(g^{i-2})p(g^{i-1})p(g^i)\Big(B\big(\mathsf{SH3}(i-3,j)+\mathsf{SN3}(i-3,j)+E_\Psi\mathsf{SE3}(i-3,j)+\\
&\quad +\mathsf{NNN}(i-3,j)\big)\\
&\quad +N\big(\mathsf{EH3}(i-3,j)+\mathsf{EN3}(i-3,j)+E_\Psi\mathsf{EE3}(i-3,j)+E_\Psi\mathsf{EA3}(i-3,j)\big)\Big)\\
\mathsf{Z}(i,j) &= \mathsf{SH3}(i,j)+\mathsf{SN3}(i,j)+E_\Psi\mathsf{SE3}(i,j)+\mathsf{NNN}(i,j)\\
&\quad +E\big(\mathsf{EH3}(i,j)+\mathsf{EN3}(i,j)+E_\Psi\mathsf{EE3}(i,j)\big)\\
\mathsf{A}(i,j) &= \mathsf{Z}(i,j)\,\delta
\end{aligned}
$$

2 Molecular clock for the pseudogene evolution

The following graph shows how constraints described in the main text are compatible with a constant rate of pseudogene evolution. Distances are calculated from the progenitors: HCS for Class I pseudogenes and preHCS for Class II pseudogenes. The horizontal error bars are for 95% confidence intervals: for K2P, they are calculated from the standard deviation formula, and for ML-OPT1 they are computed from simulations. Calibration points: (a) lower bound from human-chimpanzee split, (b) upper bound from human-macaque LCA, (c) lower bound from age of AluY, (d) lower bound from human-NWM LCA.

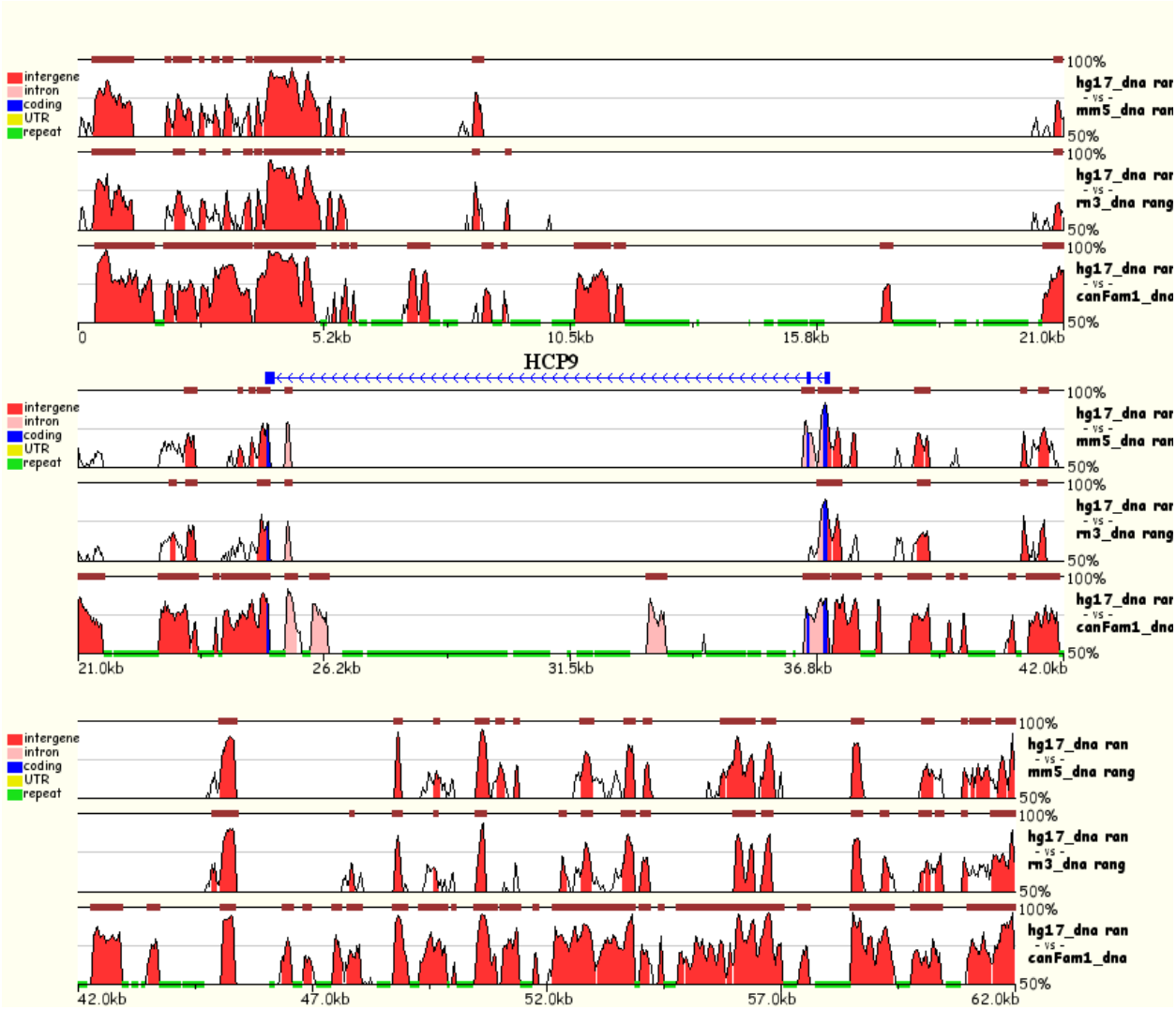


3 Conserved syntenies with cytochrome *c* pseudogenes

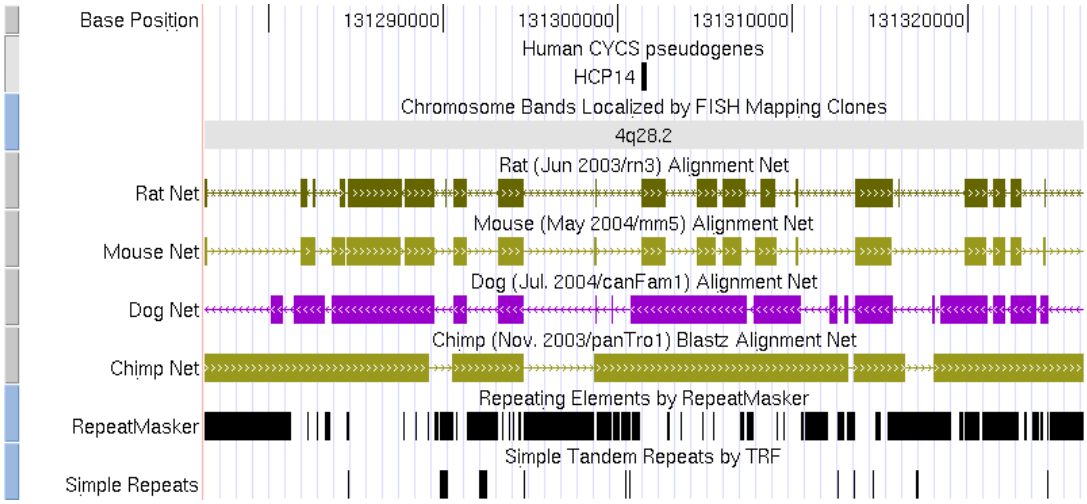
The conserved syntenies were analyzed in the UCSC Genome Browser (`http://genome.ucsc.edu/`; Karolchik et al. 2003; Blanchette et al. 2004) for human genome assembly hg17 [NCBI Build 35] (May 2004), mouse genome assembly mm5 (May 2004), rat genome assembly rn3 (June 2003), dog genome assembly canFam1 (July 2004), and chimpanzee genome assembly panTro1 (November 2003). Syntenies with the preliminary rhesus macaque genome assembly (`http://www.hgsc.bcm.tmc.edu/projects/rmacaque/`) were initially analyzed using the Genboree genome browser (`http://www.genboree.org/`), and then uploaded to MULAN (`http://mulan.dcode.org/`; Ovcharenko et al. 2005) for displaying syntenies between human, chimpanzee, and macaque genomes. A custom annotation of the hg17 human genome assembly with the locations cytochrome *c* pseudogenes can be downloaded from `http://www.iro.umontreal.ca/~csuros/pseudogenes/hg17-hcp.gff` in GFF format. The UTR and CDS annotations in the MULAN plots are based on alignment with a Refseq mRNA (accession number `NM_018947.4`) of the HCS gene.

Oldest Class II pseudogenes

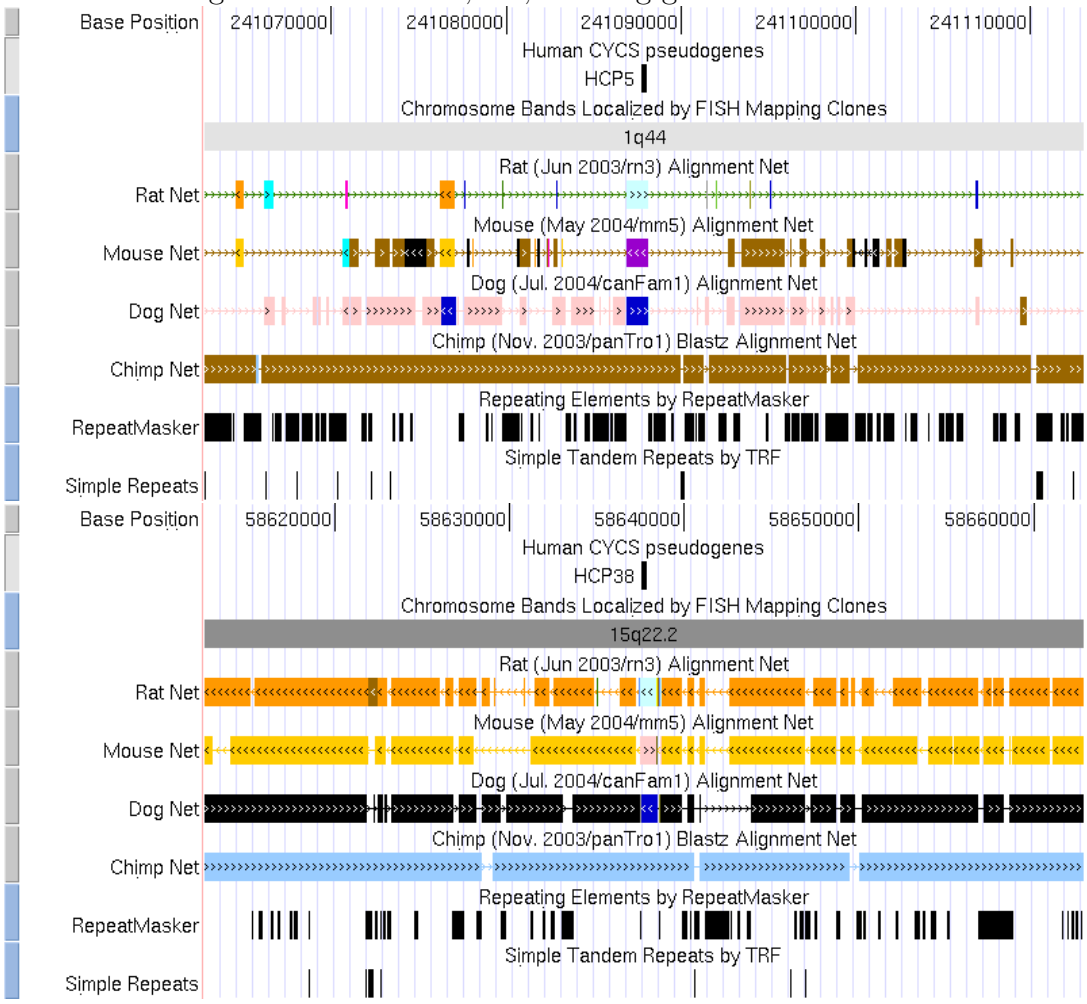
Zhang and Gerstein (2003) postulated that the hcp9 pseudogene is a disabled ortholog of rodent testis-specific cytochrome *c* genes. Indeed, hcp9 falls into conserved syntenic regions with the mouse, rat, and dog genomes.



Hcp14 is the oldest pseudogene for the somatic cytochrome *c* gene, and appears in conserved syntenies with mouse, rat, and dog.

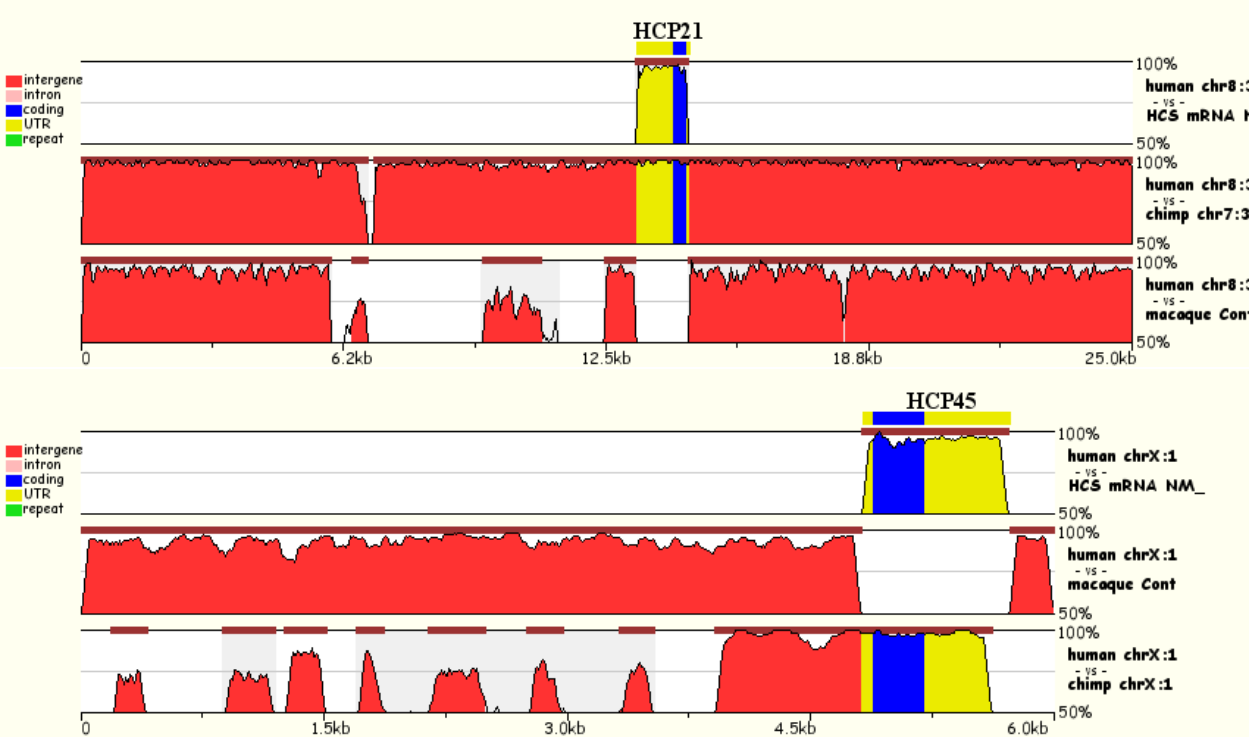


Hcp5 and hcp38 are the second and third oldest pseudogenes in our study, they fall between conserved regions in the mouse, rat, and dog genomes.

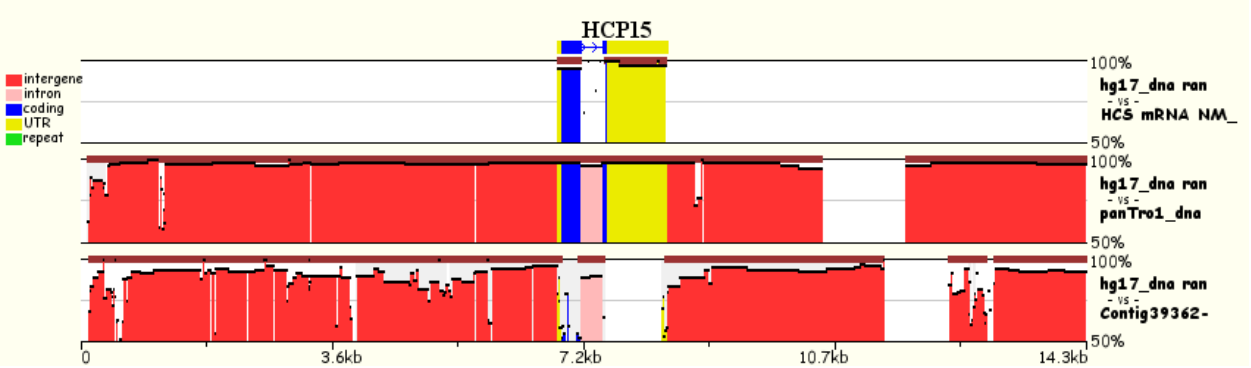


Class I pseudogenes

Class I pseudogenes (hcp21, hcp15, and hcp45) appear in syntenies with the chimpanzee genome, but not in the macaque genome.

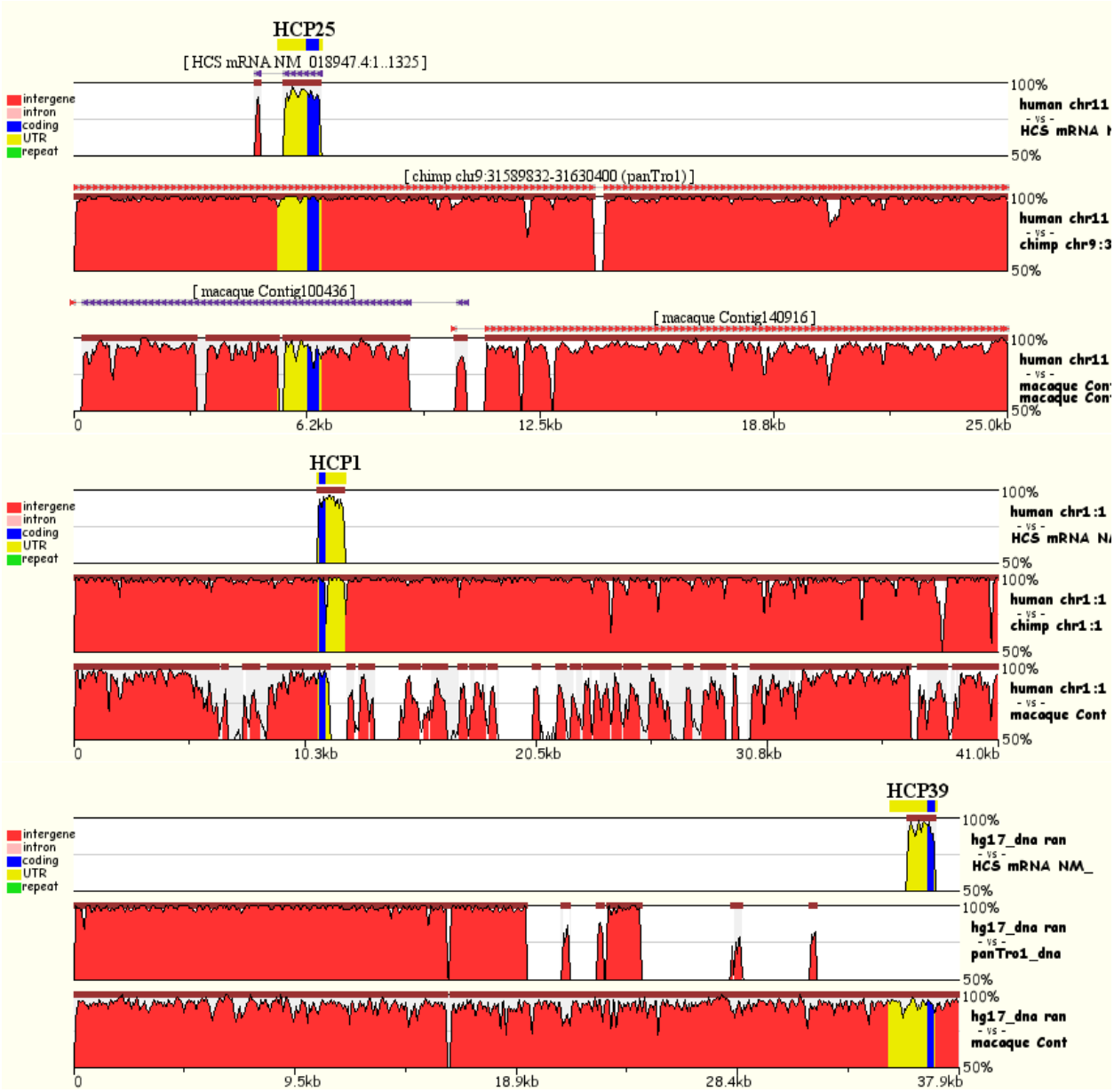


Hcp15 is disrupted by an inserted Alu element: MULAN aligns an Alu sequence in the macaque genome with it, but the coding regions do not appear in the synteny.



Youngest Class II pseudogenes

The younger Class II pseudogenes hcp25, hcp1, and hcp39 appear in conserved syntenies with the macaque genome.



References

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller (2004). Aligning multiple genomic sequences with the Threaded Blockset Aligner. *Genome Research 14*(4), 708–715.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent (2003). The UCSC genome browser database. *Nucleic Acids Research 31*(1), 51–54.

Ovcharenko, I., G. G. Loots, B. M. Giardine, M. Hou, J. Ma, R. C. Hardison, L. Stubbs, and W. Miller (2005). Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Research 15*(1), 184–194.

Zhang, Z. and M. Gerstein (2003). The human genome has 49 cytochrome *c* pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene 312*, 61–72.