Genome-Wide Survey for Biologically Functional Pseudogenes

Örjan Svensson, Lars Arvestad, Jens Lagergren

PLoS Computational Biology 2006 May;2(5):e46.

presented by Age Tats Journal Club 10.11.2006

Pseudogenes – sequences of genomic DNA lacking the protein coding capability of their paralogous counterpart.

*similar to the original gene

*accumulate disablements (in-frame stop codons, sequence frameshifts)

*without function

*released from selective pressure (neutral evolution)

*not overlapping any known gene

~20 000 in human

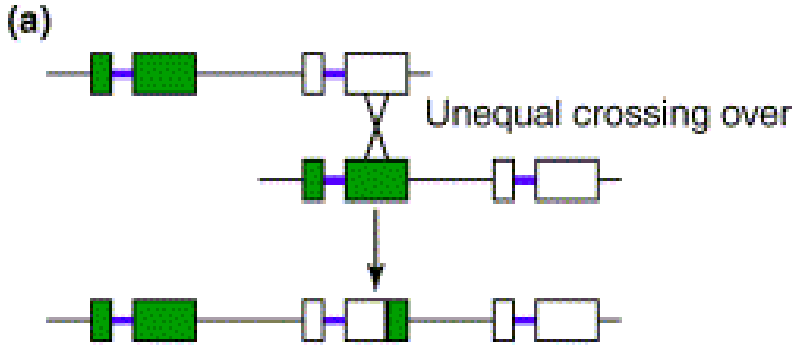
Duplicated /nonprocessed

Origin: Unequal crossing over

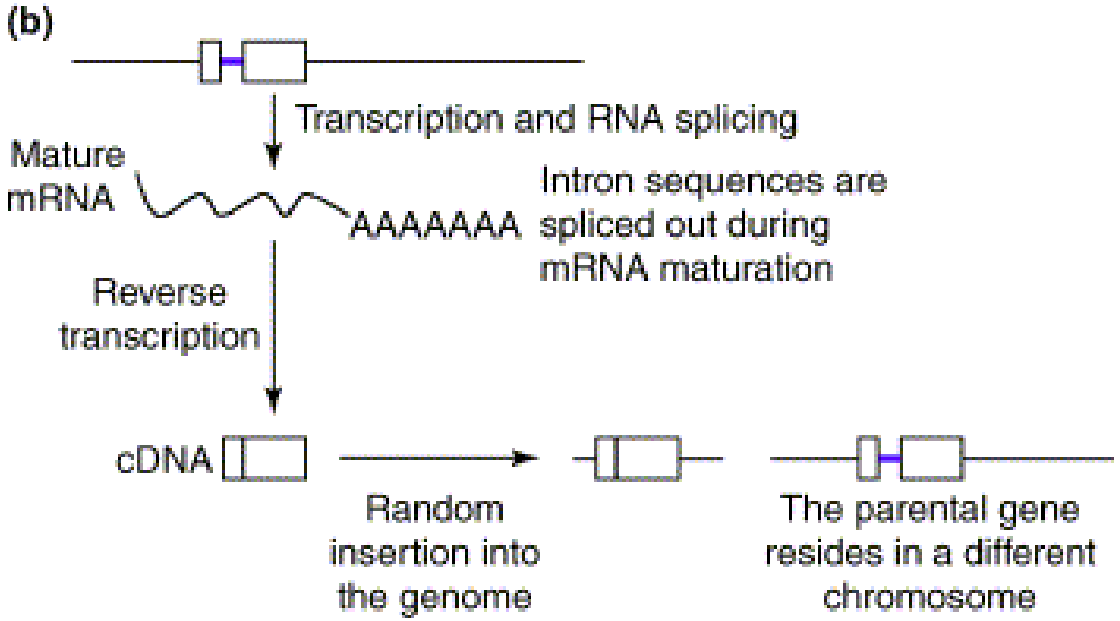
Processed (~8000 (~3600) in human)

Origin: Retrotransposition

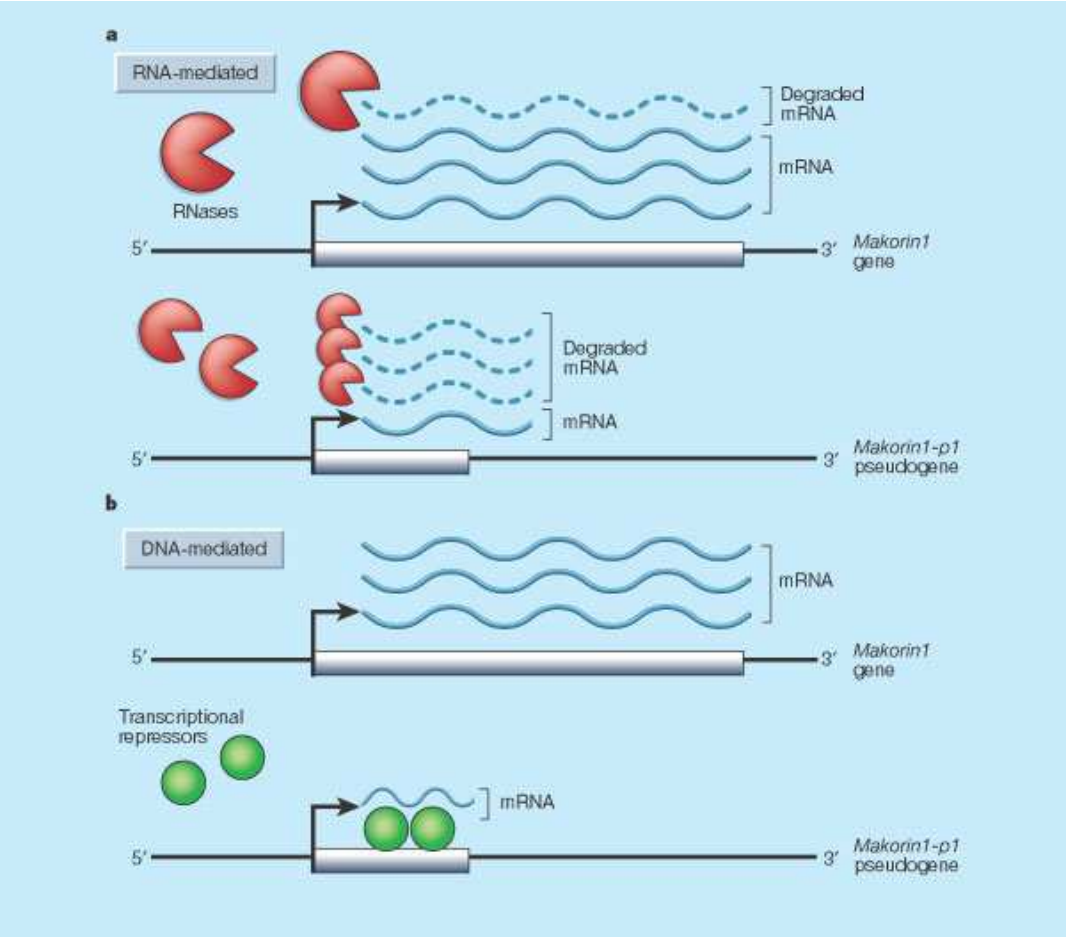
Unequal crossing over



Retrotransposition



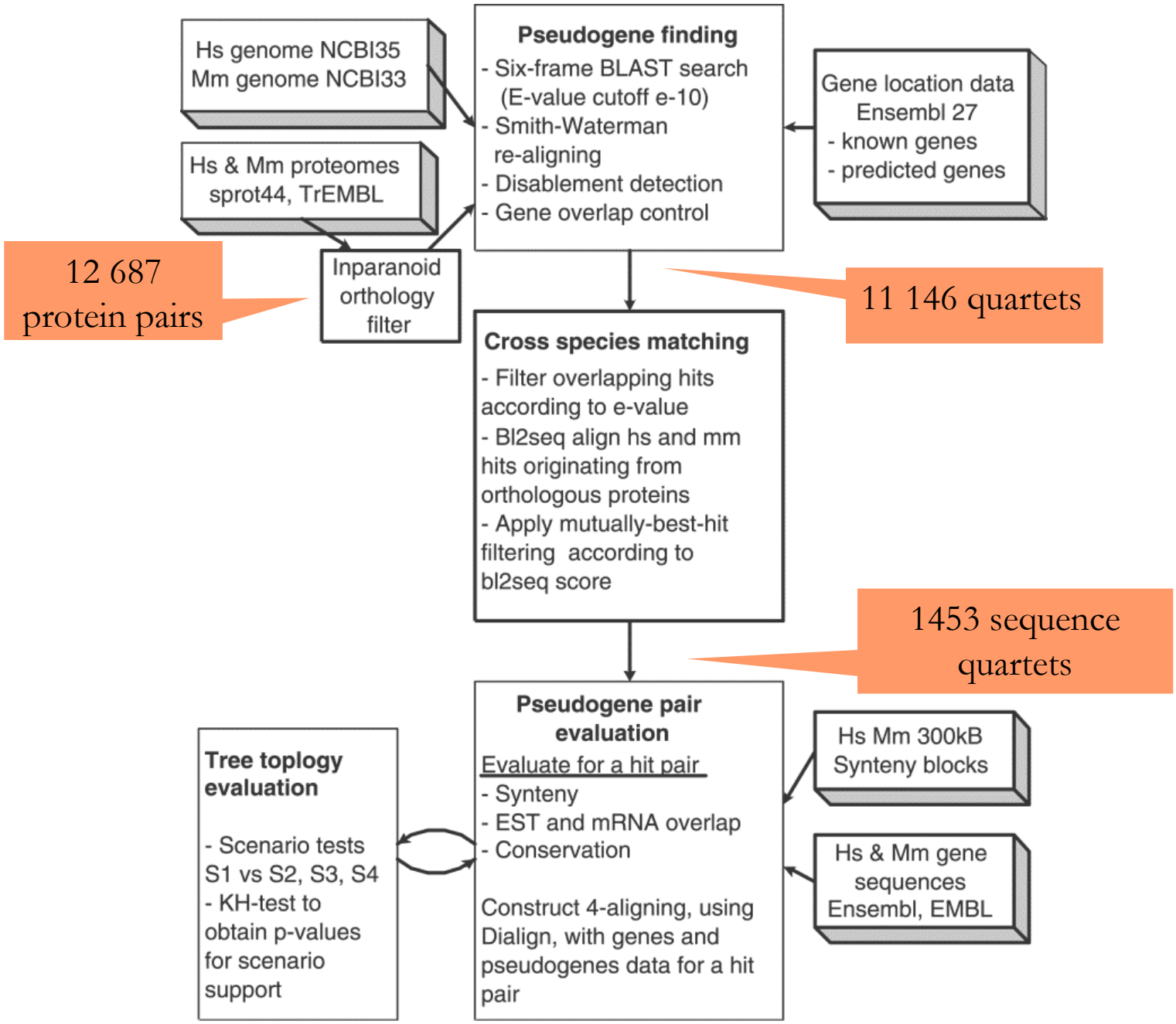
The expression of the *Makorin1* gene is controlled by one of its pseudogene copies, *Makorin1-p1*. The figure shows two ways in which this might happen. **a,** An RNA-mediated mechanism. Here, messenger RNA copies of the pseudogene and gene compete for a destabilizing protein that binds a crucial 700-nucleotide region near the beginning of the mRNAs. This destabilizing protein might be an RNA-digesting enzyme (RNase). **b,** A DNA-mediated mechanism. Here, regulatory elements in the 700-nucleotide region of the pseudogene and gene compete for transcriptional repressors.



Hirotsune et al (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**, 91-96.

1. Look for conserved pseudogenes common to human and mouse, originating from one duplication predating the human mouse species split and having evolved as pseudogenes since species split.

2. Test the potential functionality of the found pseudogenes using enrichment of transcription and synteny.


Hs genome NCBI35
Mm genome NCBI33
Hs & Mm proteomes
sprot44, TrEMBL
Inparanoid orthology filter
12 687 protein pairs
Pseudogene finding
- Six-frame BLAST search (E-value cutoff e-10)
- Smith-Waterman re-aligning
- Disablement detection
- Gene overlap control
Gene location data Ensembl 27
- known genes
- predicted genes
11 146 quartets
Cross species matching
- Filter overlapping hits according to e-value
- Bl2seq align hs and mm hits originating from orthologous proteins
- Apply mutually-best-hit filtering according to bl2seq score
1453 sequence quartets
Pseudogene pair evaluation
Evaluate for a hit pair
- Synteny
- EST and mRNA overlap
- Conservation
Construct 4-aligning, using Dialign, with genes and pseudogenes data for a hit pair
Hs Mm 300kB Synteny blocks
Hs & Mm gene sequences Ensembl, EMBL
Tree toplogy evaluation
- Scenario tests S1 vs S2, S3, S4
- KH-test to obtain p-values for scenario support

The substitution rates are specified by the rate matrix Q = $\{q_{ij}\}$ defined by:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position in a codon triplet} \\ \mu\pi_j, & \text{differ by asynonymous transversion} \\ \mu\kappa\pi_j, & \text{differ by asynonymous transition} \\ \mu\omega\pi_j, & \text{differ by anonsynonymous transversion} \\ \mu\omega\kappa\pi_j, & \text{differ by anonsynonymous transition} \end{cases}$$

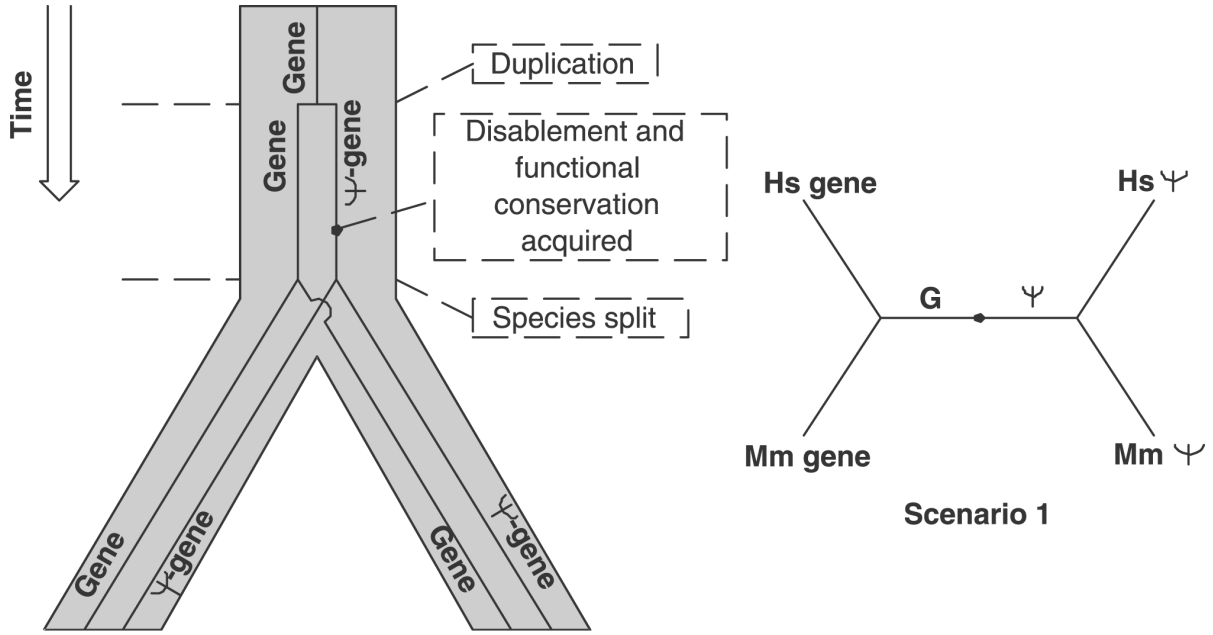
ω – nonsynonymous/synonymous rate ratio

κ – transition/transversion rate ratio

π – equilibrium frequency of codon

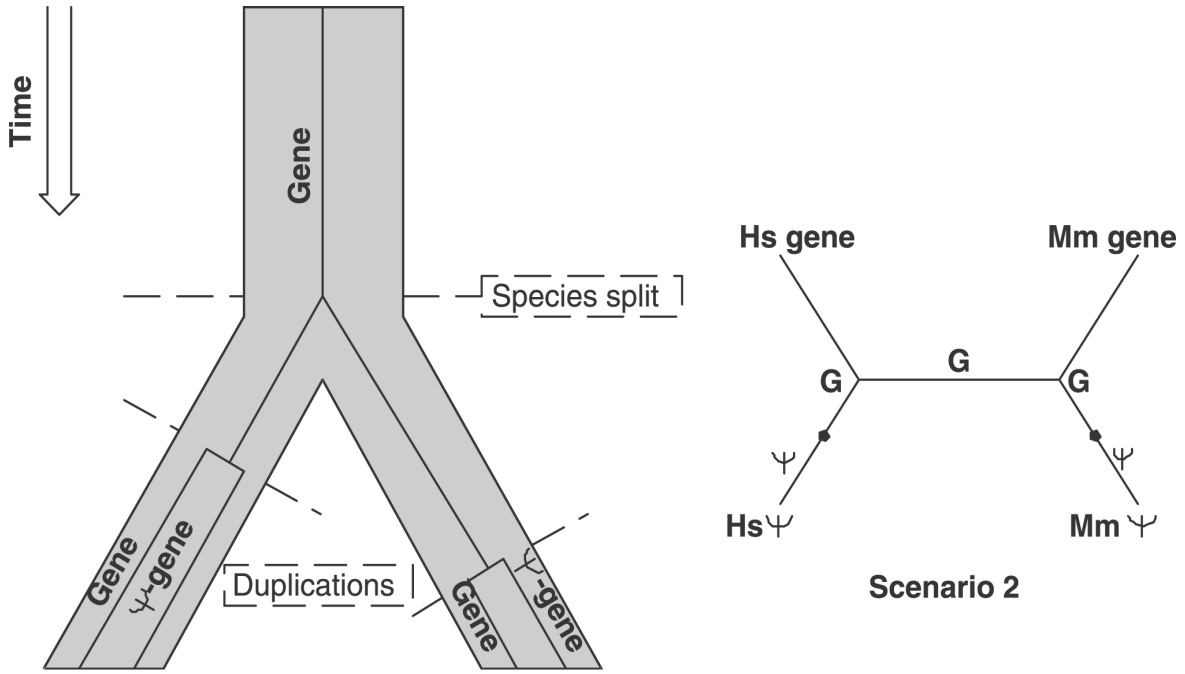
μ – normalizing rate factor

Evolutionary scenario S1



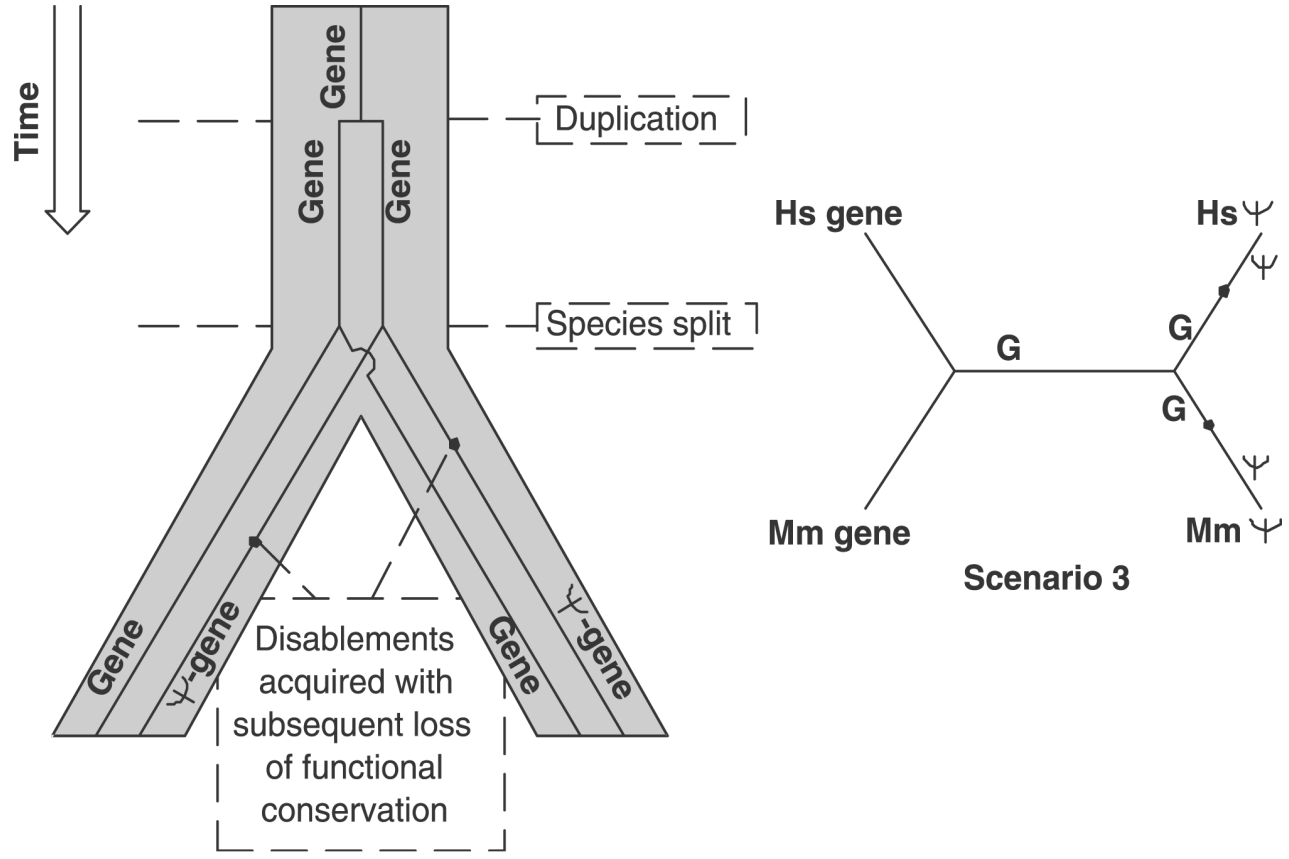
Conserved pseudogenes originating before the species split

Evolutionary scenario S2

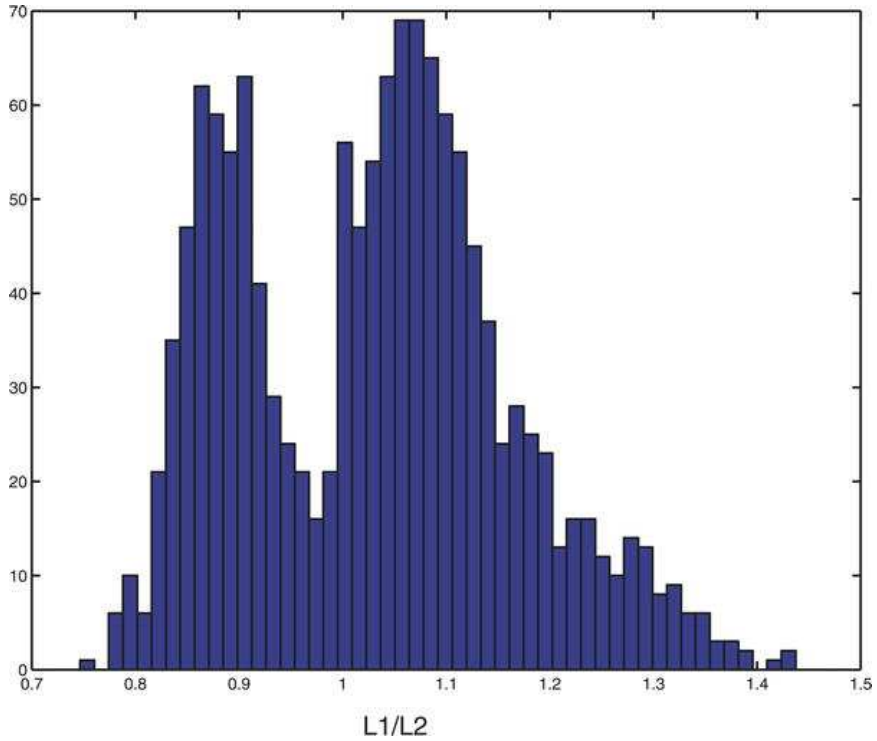


Both pseudogenes originate independently of each other, after species split

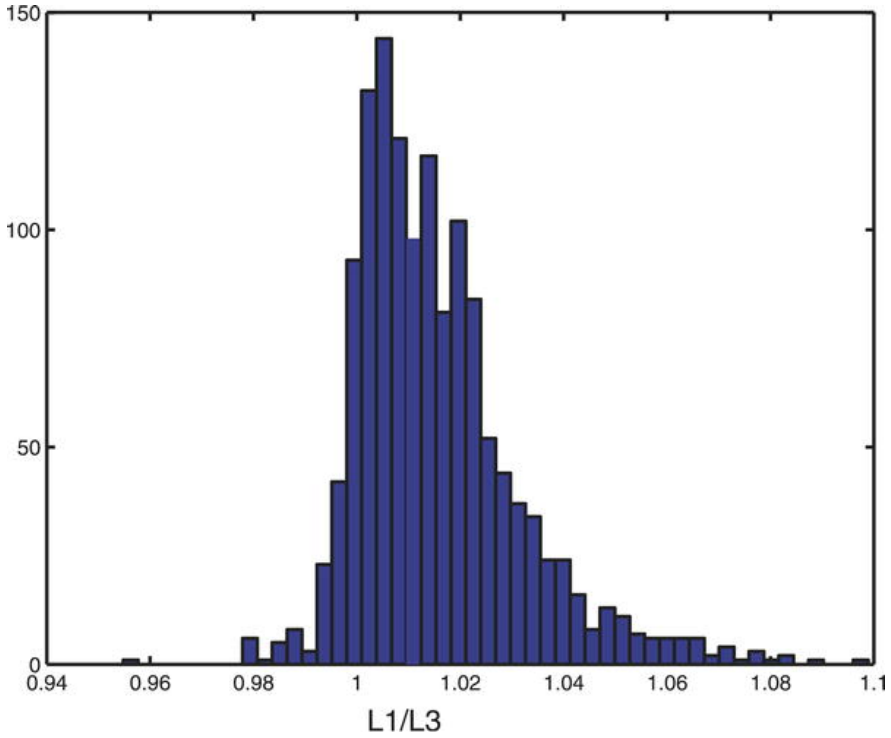
Evolutionary scenario S3



Transition from gene to pseudogene occurred subsequent to the species split



Histogram of Likelihood Quotients when Comparing Scenarios $S1$ and $S2$



Histogram of Likelihood Quotients when Comparing Scenarios $S1$ and $S3$

Number of Sequence Pairs in Each Class Favoring a Particular Scenario

Scenario	Class 1						Class 2						Class 3						Class 4						Total					
	S	R	C	U	N	Total	S	R	C	U	N	Total	S	R	C	U	N	Total	S	R	C	U	N	Total	S	R	C	U	N	Total
S1	4	0	0	0	2	**6**	2	0	0	0	0	**2**	4	0	0	0	1	**5**	10	0	5	2	0	**17**	20	0	5	2	3	**30**
S2	2	3	1	17	137	**170**	18	2	11	19	110	**160**	10	2	8	9	26	**55**	100	7	20	33	157	**317**	130	14	50	78	430	**702**
S3	2	0	4	6	14	**26**	23	1	6	2	14	**46**	18	0	3	4	3	**28**	106	4	25	6	21	**162**	149	5	38	18	52	**262**

For each scenario and class, the number of sequence pairs that are syntenic (S), reversed syntenic (R), close to synteny (C), with unknown synteny (U), nonsyntenic (N) and total (bold). *p*-Values used are 0.001 to distinguish S1 and S3 from S2 examples, and 0.1 to separate S1 from S3.

DOI: 10.1371/journal.pcbi.0020046.t002

Class 1 – have detectable disablements and do not overlap any Ensemble gene prediction

Class 2 – have detectable disablements and overlap an Ensembl gene prediction

Class 3 – no detectable disablements and do not overlap any Ensemble gene prediction

Class 4 – no detectable disablements and overlap an Ensembl gene prediction

Result: 30 quartets for which the sequences suggest that:

1) The pseudogenes are evolutionarily conserved since before the human and mouse speciation

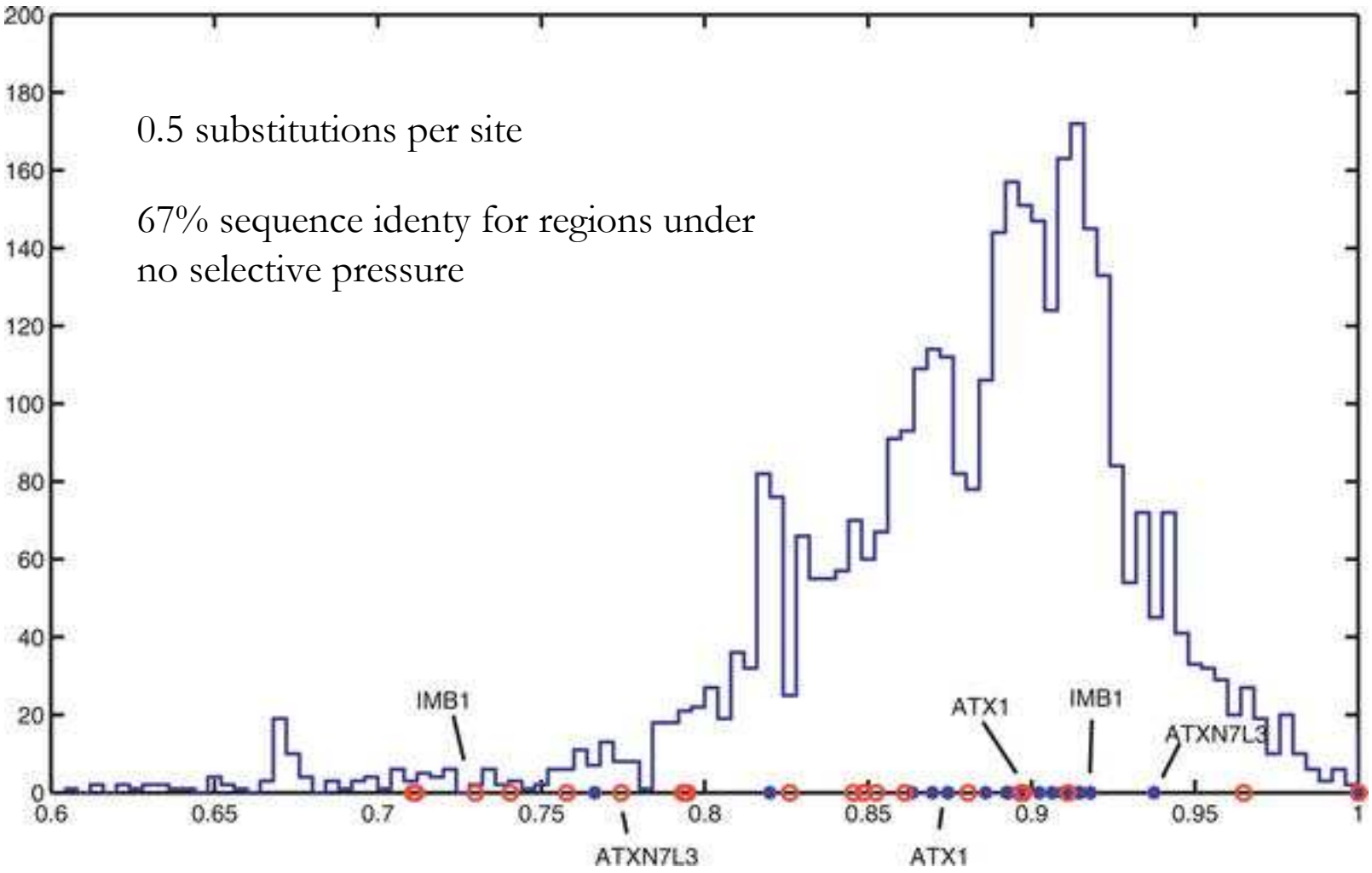
2) They have been pseudogenes since prior to the speciation

p-Values for Scenario Comparisons and Pseudogene Expression Evidence (Number of Matching EST and mRNA Sequences) for the 20 Syntenic *S*1 Quartets.

Protein Name	Hs Chr	Mm Chr	Class	S1 versus S2 *p*-Value	S1 versus S3 *p*-Value	Human EST	Mouse EST	Human mRNA	Mouse mRNA
ATX1	16	8	1	<0.001	0.030	6	4	1	1
ATXN7L3	12	10	1	<0.001	<0.001	>50	>50	2	1
IMB1	X	X	1	<0.001	<0.001	0	0	0	0
PDZRN3	12	15	1	<0.001	<0.001	4	1	0	1
DYHC	11	7	2	<0.001	<0.001	4	6	0	1
ODF3	22	15	2	<0.001	0.073	23	7	0	4
A8A1	15	7	3	<0.001	0.065	1	9	1	1
TPC3	6	10	3	<0.001	0.005	3	1	0	0
Q9P2K1	10	19	3	<0.001	<0.001	0	0	1	1
ZNF629	1	1	3	<0.001	0.002	0	0	1	1
CA1C	3	9	4	<0.001	0.058	1	0	0	1
DD17	1	1	4	<0.001	0.076	13	8	2	3
Q7Z3F3	4	5	4	<0.001	<0.001	13	4	3	3
Q8IYB1	17	11	4	<0.001	0.002	18	3	1	1
Q8N1K5	18	17	4	<0.001	<0.001	0	9	0	1
DNAH5	17	11	4	<0.001	0.007	1	1	1	2
ERBB2IP	3	3	4	<0.001	0.030	1	0	1	1
TOPORS	9	4	4	<0.001	0.047	0	4	0	2
TDR1	2	12	4	<0.001	0.035	0	0	1	2
Z142	4	8	4	<0.001	0.027	6	4	1	0

DOI: 10.1371/journal.pcbi.0020046.t003

Any reciprocal best hit longer than 100 bp and with more than 99% sequence identity to the query sequence was retrieved.



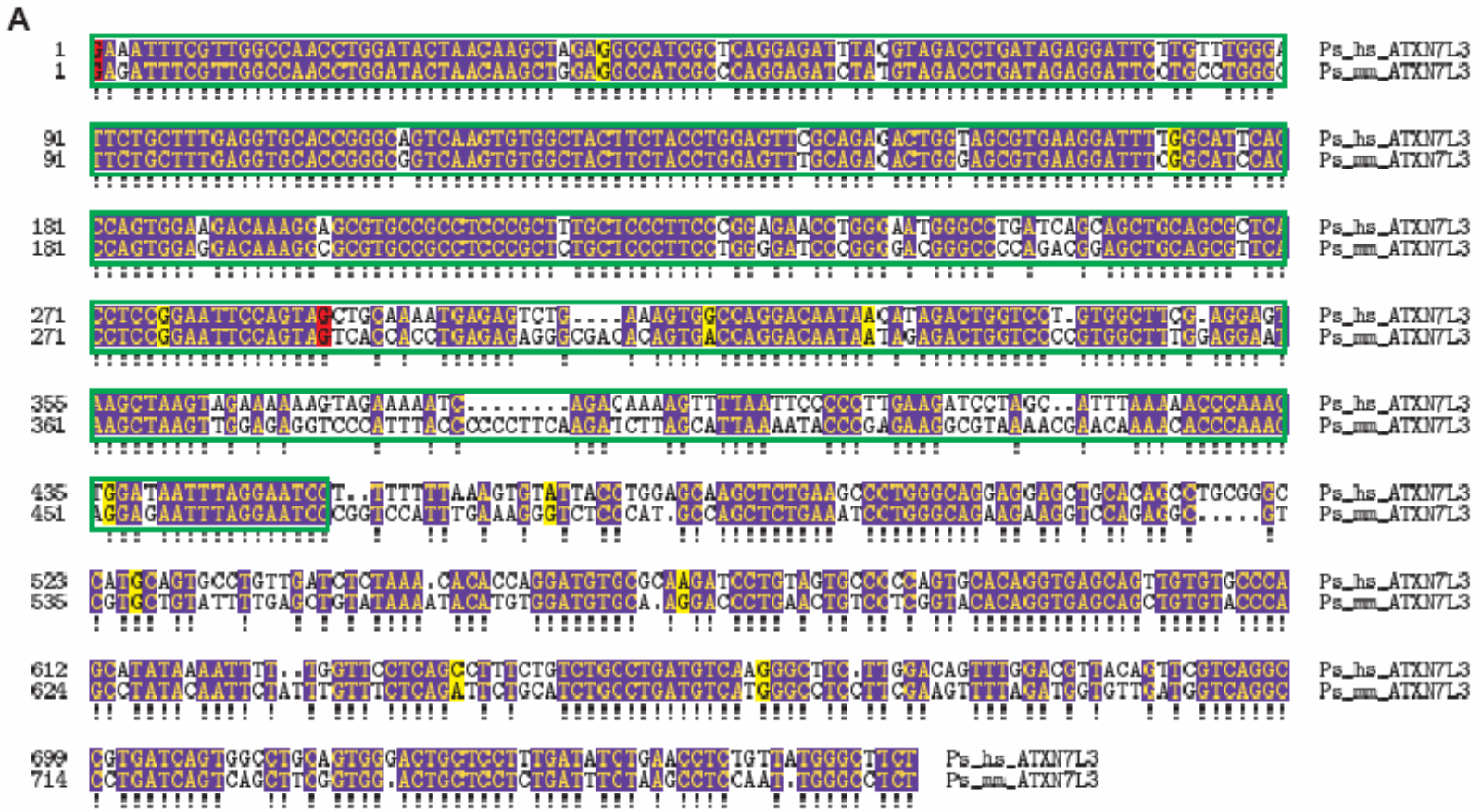
Conservation between Human and Mouse Gene and Pseudogene Sequences for the 20 Syntenic *S*1 Sequences

Blue stars indicate genes. Red circles indicate pseudogenes. The histogram shows, for reference, the conservation of all genes giving rise to pseudogenes.

Conservation Percentage in and around the Pseudogene

Protein Name	Conservation		Upstream	Downstream
	Percent	*p*-Value		
ATX1	91.1 %	$<10^{-50}$	63.3 %	75.7 %
ATXN7L3	76.7 %	$3.75*10^{-9}$	69.4 %	75.6 %
IMB1	72.9 %	$<10^{-50}$	65.3 %	63.7 %
PDZRN3	79.4 %	$1.25*10^{-8}$	61.8 %	47.2 %
DYHC	79.1 %	$<10^{-50}$	74.1 %	75.1 %
ODF3	70.7 %	0.036	58.6 %	73.4 %
A8A1	89.7 %	$1.13*10^{-16}$	71.5 %	44.2 %
TPC3	89.4 %	$1.84*10^{-13}$	47.7 %	60.3 %
Q9P2K1	85.2 %	$1.67*10^{-8}$	62.0 %	62.3 %
ZNF629	84.8 %	$1.25*10^{-35}$	69.2 %	46.5 %
CA1C	84.5 %	$2.29*10^{-20}$	70.0 %	53.3 %
DD17	86.1 %	$1.01*10^{-13}$	74.3 %	53.1 %
Q7Z3F3	100 %	$1.64*10^{-19}$	64.1 %	66.1 %
Q8IYB1	82.7 %	$1.02*10^{-20}$	76.4 %	66.4 %
Q8N1K5	73.8 %	$5.62*10^{-6}$	45.7 %	47.5 %
DNAH5	88.1 %	$1.64*10^{-8}$	43.9 %	56.2 %
ERBB2IP	75.9 %	$2.06*10^{-9}$	55.7 %	55.9 %
TOPORS	71.1 %	$7.30*10^{-4}$	70.0 %	46.0 %
TDR1	82.8 %	$2.35*10^{-7}$	45.4 %	46.5 %
Z142	96.5 %	$1.69*10^{-49}$	74.0 %	86.7 %

DOI: 10.1371/journal.pcbi.0020046.t005



An alignment of the processed copies to the ATXN7L3 human and mouse protein-coding genes. The human as well as the mouse ATXN7L3 contains 12 exons, which are all present in the respective duplicates. Approximate exon borders are shown in yellow.
The most interesting part consists of columns 1–468 (boxed green), which according to several EST and mRNA sequences is the only segment expressed. It consists of a highly conserved part, 1–288 (red), which is a potential open reading frame, followed by part 289–468 with pseudogenic disablements.

Human-chimpanzee

Conservation estimates can, even together with expression evidence, be expected to be insufficient for revealing whether an individual pseudogene is functional or not.

Results: 742 class 1 pseudogenes favoring *S*1.

Percentage of Expressed Pseudogenes in Relation to Their Conservation *p*-Values

p-Value	Total Number	EST/mRNA Expression		
		EST	mRNA	Either
0.01	27	19%	22%	33%
0.05	77	17%	19%	29%
Total	742	12%	15%	21%

DOI: 10.1371/journal.pcbi.0020046.t006

Human–Chimpanzee Conserved and Expressed Pseudogene Pairs

Hs Protein	Gene Name	Hs Chr	Hs Start	Hs End	Conservation *p*-Value	Expression	
						EST	mRNA
ENSP00000244769	ATX1	16	70441078	70443214	0.019	Yes	Yes
ENSP00000262316	RHBDF1	3	14589363	14591300	0.0022	No	Yes
ENSP00000234739	BCL9	5	66968594	66970526	$6.1*10^{-4}$	Yes	No
ENSP00000235329	MFN2	X	108617852	108619651	$1.1*10^{-8}$	No	Yes
ENSP00000327539	HNRPH1	X	142485593	142486960	0.0030	No	Yes
ENSP00000268661	RPL3L	5	60722282	60723464	0.035	No	Yes
ENSP00000313007	PABPC1	12	62502005	62503947	0.044	No	Yes
ENSP00000318000	NAB1	X	150065269	150067075	0.0021	Yes	No
ENSP00000327539	HNRPH1	6	160104224	160105428	0.014	Yes	No
ENSP00000223215	MEST	3	29103895	29104914	0.0010	No	Yes
ENSP00000349469	TPR4	X	92348904	92349903	0.0038	Yes	Yes
ENSP00000341327	SOCS4	6	113650996	113651931	0.032	No	Yes
ENSP00000313582	ZNF436	7	6465488	6467284	0.027	Yes	Yes
ENSP00000342024	ATP8A1	2	241221794	241223519	0.025	Yes	No
ENSP00000302684	DKFZp343F142	7	65814628	65815402	0.011	Yes	No
ENSP00000319053	ZNF77	19	9495628	9496170	0.0077	Yes	No
ENSP00000317614	NP444270	10	97910042	97910556	0.0090	Yes	Yes
ENSP00000307858	ZBTB4	3	142645022	142645752	0.020	Yes	Yes
ENSP00000319233	TLE3	16	70023164	70024179	0.019	No	Yes
ENSP00000256682	ARF3	17	41069429	41069752	0.048	Yes	Yes
ENSP00000257498	CTSL	10	89137032	89139195	0.037	No	Yes
ENSP00000274192	SRD5A1	X	138254579	138255358	0.050	Yes	No

DOI: 10.1371/journal.pcbi.0020046.t007