SUPPLEMENTARY TABLES

A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome

Jeramiah J. Smith, Kenta Sumiyama and Chris T. Amemiya

Supplementary Table S1 – Summary statistics for searches of known vertebrate repetitive elements in lamprey BACs.

Element Type	number of Elements[1]	length occupied	percentage of sequence
Retroelements	915	465414 bp	13.78%
SINEs:	437	126007 bp	3.73%
LINEs:	409	327663 bp	9.70%
L2/CR1/Rex	291	266367 bp	7.88%
R1/LOA/Jockey	10	9946 bp	0.29%
R2/R4/NeSL	2	125 bp	0%
RTE/Bov-B	32	9262 bp	0.27%
L1/CIN4	51	20465 bp	0.61%
LTR elements:	69	11744 bp	0.35%
Gypsy/DIRS1	36	24033 bp	0.71%
Retroviral	22	2396 bp	0.07%
DNA transposons (excluding *Harbingers*)	120	14001 bp	0.21%
hobo-Activator	47	4204 bp	0.12%
Tc1-IS630-Pogo	7	975 bp	0.03%
En-Spm	20	1460 bp	0.04%
MuDR-IS905	4	158 bp	0%
PiggyBac	3	543 bp	0.02%
LatiHarb1[2]	**18**	**132162 bp**	**3.91%**
Unclassified:	4	407 bp	0.01%
Total interspersed repeats:		479822 bp	14.20%

Small RNA:	258	82910 bp	2.45%
Satellites:	8	394 bp	0.01%
Simple repeats:	700	28376 bp	0.84%
Low complexity:	889	33119 bp	0.98%

The query species was assumed to be vertebrata RepeatMasker version open-3.2.5 , default mode run with cross_match version 0.990329, RepBase Update 20080801, RM database version 20080801.

1) Most repeats fragmented by insertions or deletions have been counted as one element
2) Sequence with high-identity to the *LatiHarb1* consensus, identified by alignment to a manually curated instance of the *LatiHarb1* element (i.e. external to the RepeatMasker search). RepeatMasker also identified a few fragments of Harbinger superfamily elements, which accounted for only 0.13% of *Latimeria* BAC sequence.

Supplementary Table S2 - Summary statistics for GeneScan predictions within 8 individual instances of the *LatiT1* element that were identified from the BAC sequence dataset.

Genbank Accession / Clone	Range (in clone)	Feature	Start	End	Strand	I/Ac	Do/T	CR	P	Score
VMRC4_103A21_106D14	199376-207010	Exon 1	436	916	+	93	103	514	0.999	49.22
		Exon 2	1290	1418	+	69	115	169	0.922	19.54
		Exon 3	1540	2313	+	111	81	573	0.957	51.86
		Exon 4	2762	3009	+	36	42	163	0.847	3.45
		Poly A Signal	3017	3022	+					
		Poly A Signal	3059	3054	-					
		Exon 4	4134	3193	-	-41	44	681	0.757	44.25
		Exon 3	5615	5403	-	111	72	103	0.776	11.13
		Exon 2	6038	5958	-	57	47	74	0.515	1.19
		Exon 1	6557	6435	-	94	110	198	0.995	22.39
FJ497005	82448-90459	Exon 1	379	859 (862)	+	93	103	485	0.999	46.32
		Exon 2	1233 (1161)	1361	+	88	115	205	0.976	25.04
		Exon 3	1483	2720	+	113	43	932	0.974	84.46
		Poly A Signal	2892	2897	+					
		Poly A Signal	2941	2936	-					
		Exon 6	3517 (3482)	3101	-	1	49	395	0.843	23.26
		Exon 5	3992	3686 (3621)	-	19	13	326	0.467	15.92
		Exon 4	4351	4104	-	48	89	193	0.575	12.24
		Exon 3	5971	5759	-	115	72	136	0.827	14.83
		Exon 2	6397	6317	-	94	47	86	0.976	6.09
		Exon 1	6870	6748	-	94	110	152	0.997	17.06
FJ497007	111628-120116	Exon 1	379	859	+	93	103	563	0.999	54.12
		Exon 2	1162	1346	+	94	12	158	0.388	9.51

		Exon 3	1504	2286	+	24	12	558	0.26	34.98
		Exon 4	2338	2667	+	40	43	310	0.388	17.4
		Poly A Signal	2839	2844	+					
		Poly A Signal	2888	2883	-					
		Exon 3	3428	3049	-	114	49	480	0.998	42.72
		Exon 2	4373	3567	-	-3	65	543	0.748	35.84
		Exon 1	4783	4711	-	89	94	64	0.989	8.42
		Poly A Signal	4930	4925	-					
		Exon 5	5596	5543	-	107	49	68	0.99	3.26
		Exon 4	6429	6220	-	108	91	111	0.893	13.73
		Exon 3	7338	7172	-	-18	75	161	0.061	5.08
		Exon 2	7874	7719	-	64	68	64	0.11	3.17
		Exon 1	8192	7913	-	-38	76	240	0.066	8.5
FJ497008	469222-477589	Exon 1	336	816	+	93	103	595	0.999	57.32
		Exon 2	1119	1310	+	94	115	123	0.999	16.27
		Exon 3	1432	2579	+	137	43	943	0.998	88.25
		Poly A Signal	2751	2756	+					
		Poly A Signal	2800	2795	-					
		Exon 7	3339	2960	-	116	49	448	0.999	39.72
		Exon 6	4284	3478	-	-3	65	665	0.372	48.04
		Exon 5	5170	5140	-	121	-24	8	0.011	-8.27
		Exon 4	6344	6135	-	103	91	76	0.406	9.73
		Exon 3	6750	6564	-	102	25	116	0.288	7.6
		Exon 2	7778	7624	-	68	68	44	0.179	0.46
		Exon 1	8071	7838	-	-38	80	231	0.705	8.82
VMRC4_217L16	156711-165205	Exon 1	375	855	+	93	103	535	0.998	51.32
		Exon 2	1158	1361	+	94	115	175	0.717	21.19

		Exon 3	1483	1821	+	129	100	193	0.99	21.33
		Exon 4	1919	2664	+	38	43	675	0.591	52.87
		Poly A Signal	2836	2841	+					
		Poly A Signal	2885	2880	-					
		Exon 5	3159	3043	-	53	49	199	0.998	11.97
		Exon 4	3422	3232	-	114	75	96	0.527	11.41
		Exon 3	3655	3561	-	37	65	32	0.927	-3.43
		Exon 2	3972	3704	-	45	86	168	0.203	9.88
		Exon 1	7313	7182	-	69	75	165	0.915	13.57
AC150309	158322-166488	Exon 1	375	855	+	93	103	618	0.999	59.62
		Exon 2	1239	1361	+	69	115	155	0.696	18.03
		Exon 3	1483	2516	+	113	42	912	0.988	82.79
		Poly A Signal	2835	2840	+					
		Poly A Signal	2884	2879	-					
		Exon 5	3225	3037	-	14	49	309	0.682	18.02
		Exon 4	3966	3505	-	45	35	347	0.484	19.8
		Exon 3	4358	4008	-	-16	32	304	0.288	11.21
		Exon 2	6270	6050	-	99	81	128	0.78	12.45
		Exon 1	7071	6963	-	103	-13	129	0.084	4.8
AC150284	166488-158322	Exon 1	375	855	+	93	103	618	0.999	59.62
		Exon 2	1239	1361	+	69	115	155	0.696	18.03
		Exon 3	1483	2516	+	113	42	912	0.988	82.79
		Poly A Signal	2835	2840	+					
		Poly A Signal	2884	2879	-					
		Exon 5	3225	3037	-	14	49	309	0.682	18.02
		Exon 4	3966	3505	-	45	35	347	0.484	19.8
		Exon 3	4358	4008	-	-16	32	304	0.288	11.21

Exon 2	6270	6050	-	99	81	128	0.78	12.45
Exon 1	7071	6963	-	103	-13	129	0.084	4.8

I/Ac = initiation signal or 3' splice site score, Do/T = 5' splice site or termination signal score. CR = coding region score, P = probability of exon, Score = GENESCAN score. Values shown in parentheses are exon boundaries observed from transgenic mouse transcripts.

Supplementary Table S3 - Summary statistics for *GeneMark.hmm-E* predictions within 8 individual instances of the *LatiT1* element that were identified from the BAC sequence dataset.

Genbank Accession / Clone	Range (in clone)	Gene #	Exon #	Strand	Exon Type	Exon Start	Exon End
VMRC4_103A21_106D14	199376-207010	1	1	+	Initial	440	920
		1	2	+	Internal	1294	1422
		1	3	+	Internal	1544	1891
		1	4	+	Terminal	2049	2107
		2	2	-	Terminal	2738	3141
		2	1	-	Initial	3280	3619
		3	3	-	Terminal	5008	5061
		3	2	-	Internal	5407	5619
		3	1	-	Initial	6439	6561
		4	1	+	Initial	6744	6806
FJ497005	82448-90459	1	1	+	Initial	375	855
		1	2	+	Internal	1229	1357
		1	3	+	Internal	1479	1681
		1	4	+	Internal	2387	2513
		1	5	+	Terminal	2624	2655
		2	2	-	Terminal	2658	2710
		2	1	-	Initial	3156	3369
		3	2	-	Terminal	3673	3916
		3	1	-	Initial	4166	4347
		4	2	-	Terminal	5361	5414
		4	1	-	Initial	5755	5877

		5	1	+	Initial	7459	7495
FJ497007	111628-120116	1	1	+	Initial	382	862
		1	2	+	Internal	1240	1368
		1	3	+	Internal	2097	2110
		1	4	+	Internal	2341	2467
		1	5	+	Terminal	2578	2609
		2	2	-	Terminal	3052	3431
		2	1	-	Initial	3570	3909
		3	3	-	Terminal	5546	5599
		3	2	-	Internal	6223	6432
		3	1	-	Initial	7175	7306
		4	1	-	Terminal	7578	7662
FJ497008	469222-477589	1	1	+	Initial	339	819
		1	2	+	Internal	1435	1761
		1	3	+	Internal	1879	2379
		1	4	+	Terminal	2490	2521
		2	2	-	Terminal	2963	3342
		2	1	-	Initial	3481	3820
		3	3	-	Terminal	5441	5494
		3	2	-	Internal	6138	6347
		3	1	-	Initial	7088	7219
VMRC4_217L16	156711-165205	1	1	+	Initial	375	855

		1	2	+	Internal	1233	1361
		1	3	+	Internal	1483	1821
		1	4	+	Internal	1940	2461
		1	5	+	Terminal	2572	2603
		2	2	-	Terminal	3043	3159
		2	1	-	Initial	3214	3441
		3	2	-	Terminal	5532	5585
		3	1	-	Initial	7182	7313
AC150309	158322-166488	1	1	+	Initial	378	858
		1	2	+	Internal	1242	1364
		1	3	+	Internal	1486	1824
		1	4	+	Terminal	1942	2519
		2	2	-	Terminal	3040	3156
		2	1	-	Initial	3625	3897
		3	2	-	Terminal	5510	5563
		3	1	-	Internal	6053	6172
AC150284	166488-158322	1	1	+	Initial	378	858
		1	2	+	Internal	1242	1364
		1	3	+	Internal	1486	1824
		1	4	+	Terminal	1942	2519
		2	2	-	Terminal	3040	3156
		2	1	-	Initial	3625	3897
		3	2	-	Terminal	5510	5563
		3	1	-	Internal	6053	6172

Supplementary Table S4 - Summary statistics for BLAST alignments between the prtedicted *LatiHarb1 tpase* amino acid sequence and similar proteins in Genbank and Repbase archives.

Sequence ID	Alignment Bitscore	% coverage of *tpase*	% identity over aligned region
Harbinger-5_XT*	266	73%	39%
HARBINGER3_DR*	262	59%	42%
Harbinger-1_XT*	187	53%	43%
gi|72158326|ref|XP_794452.1| PREDICTED: hypothetical protein [*Strongylocentrotus purpuratus*]	185	65%	33%
gi|72158323|ref|XP_794409.1| PREDICTED: hypothetical protein [Strongylocentrotus purpuratus]	182	65%	32%
HARBINGER2_DR*	180	62%	33%
gi|157423504|gb|AAI53379.1| Zgc:162945 [*Danio rerio*]	173	60%	30%
gi|115913060|ref|XP_001188248.1| PREDICTED: hypothetical protein, partial [*Strongylocentrotus purpuratus*]	168	40%	39%
gi|156546775|ref|XP_001605652.1| PREDICTED: putative nuclease HARBI1-like [*Nasonia vitripennis*]	163	59%	33%
gi|149022650|gb|EDL79544.1| rCG26755 [*Rattus norvegicus*]	162	45%	37%
gi|156539873|ref|XP_001599679.1| PREDICTED: similar to ENSANGP00000014470 [*Nasonia vitripennis*]	157	57%	33%
gi|72158329|ref|XP_794543.1| PREDICTED: hypothetical protein [*Strongylocentrotus purpuratus*]	148	46%	36%
Harbinger-4_XT*	146	48%	29%
gi|149537657|ref|XP_001519824.1| PREDICTED: putative nuclease HARBI1-like, partial [*Ornithorhynchus anatinus*]	145	40%	38%
gi|26334701|dbj|BAC31051.1| unnamed protein product [*Mus musculus*]	142	43%	35%
gi|26351017|dbj|BAC39145.1| unnamed protein product [*Mus musculus*]	142	43%	35%
gi|72009676|ref|XP_785880.1| PREDICTED: hypothetical protein [*Strongylocentrotus purpuratus*]	139	42%	35%
gi|115956159|ref|XP_001188637.1| PREDICTED: hypothetical protein [*Strongylocentrotus purpuratus*]	138	25%	42%
gi|115889710|ref|XP_001183641.1| PREDICTED: hypothetical protein, partial [*Strongylocentrotus purpuratus*]	122	25%	42%
gi|115938434|ref|XP_001201441.1| PREDICTED: hypothetical protein,	122	25%	41%

partial [*Strongylocentrotus purpuratus*]			
gi|156540836|ref|XP_001601490.1| PREDICTED: putative nuclease HARBI1-like [*Nasonia vitripennis*]	111	49%	29%
gi|156539627|ref|XP_001600349.1| PREDICTED: similar to ENSANGP00000014578 [*Nasonia vitripennis*]	103	40%	32%
gi|156539496|ref|XP_001600321.1| PREDICTED: similar to ENSANGP00000014578 [*Nasonia vitripennis*]	97.4	40%	31%

Sequence identifiers marked with an asterisk are from Repbase, all others are from GenBank.