Comprehensive analysis of human endogenous retrovirus group HERV-W locus transcription in multiple sclerosis brain lesions by high throughput amplicon sequencing

Katja Schmitt[a], Christin Richter[a], Christina Backes[a], Eckart Meese[a], Klemens Ruprecht[b], Jens Mayer[a]

[a]Institute of Human Genetics, Center of Human and Molecular Biology, Medical Faculty, University of Saarland, 66421 Homburg/Saar, Germany
[b]Department of Neurology, Charité - Universitätsmedizin Berlin, Germany

Supplemental file 1

Supplemental Table S1
Summary of filtering and assignment of 454/FLX sequence reads

Supplemental Table S2
Summary of filtering and assignment of Illumina/MiSeq sequence reads

Supplemental Figure S1
Location of 5'*env* and 3'*env* amplicons within the HERV-W *env* gene region

Supplemental Figure S2
Structure of HERV-W loci identified as transcribed in this study

Supplemental Table S1 – Summary of filtering and assignment of 454/FLX sequence reads*

	5'*env*																							
	MS1		**MS2**		**MS3**		**MS4**		**MS5**		**MS6**		**MS7**		**H1**		**H2**		**H3**		**H4**		**H5**		**H6**		**H7**	
total # of reads	4297		2208		4304		2203		2116		1011		4010		5726		1622		1979		736		897		494		626	
average read length [bp]	238.5		245.4		239.0		215.6		225.3		216.1		215.1		242.0		242.4		241.2		236.3		236.5		236.4		236.8	
# of sequencing artefacts filtered	155		50		111		69		45		27		144		114		42		74		13		13		8		6	
number of reads with x mismatches	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
0	3104	74.94	1606	74.42	2755	65.70	1596	78.58	1555	77.21	681	72.30	2665	71.99	4097	73.00	1250	79.11	1409	73.96	516	72.88	663	76.74	353	73.54	403	67.10
1	799	19.29	432	20.02	998	23.80	293	14.43	257	12.76	120	12.74	597	16.13	1195	21.29	272	17.22	417	21.89	120	16.95	127	14.70	65	13.54	129	21.46
2	181	4,37	97	4.50	346	8.25	84	4.12	151	7.50	104	11.04	305	8.24	232	4.13	46	2.91	63	3.31	53	7.49	49	5.67	47	9.79	46	7.65
3	39	0.94	21	0.97	75	1.79	32	1.58	25	1.24	23	2.44	70	1.89	48	0.86	8	0.51	13	0.68	6	0.85	9	1.04	11	2.29	13	2.16
4	16	0.39	1	0.05	16	0.38	14	0.69	22	1.10	4	0.42	42	1.13	19	0.34	3	0.19	3	0.16	9	1.27	11	1.27	2	0.41	6	0.10
5	3	0.07	0	0.00	3	0.07	7	0.34	4	0.20	5	0.53	14	0.38	11	0.20	1	0.06	0	0.00	1	0.14	2	0.23	1	0.21	4	0.67
>5	0	0.00	1	0.05	0	0.00	5	0.25	0	0.00	5	0.53	9	0.24	10	0.18	0	0.00	0	0.00	3	0.42	3	0.35	1	0.21	0	0.00
sum	4142		2158		4193		2031		2014		942		3702		5612		1580		1905		708		864		480		601	
reads >3 mismatches filtered	19		2		19		26		26		14		65		40		4		3		13		16		4		10	
reads contributing to further analysis	**4123**		**2156**		**4174**		**2108**		**2045**		**970**		**3801**		**5572**		**1576**		**1902**		**710**		**868**		**482**		**610**	

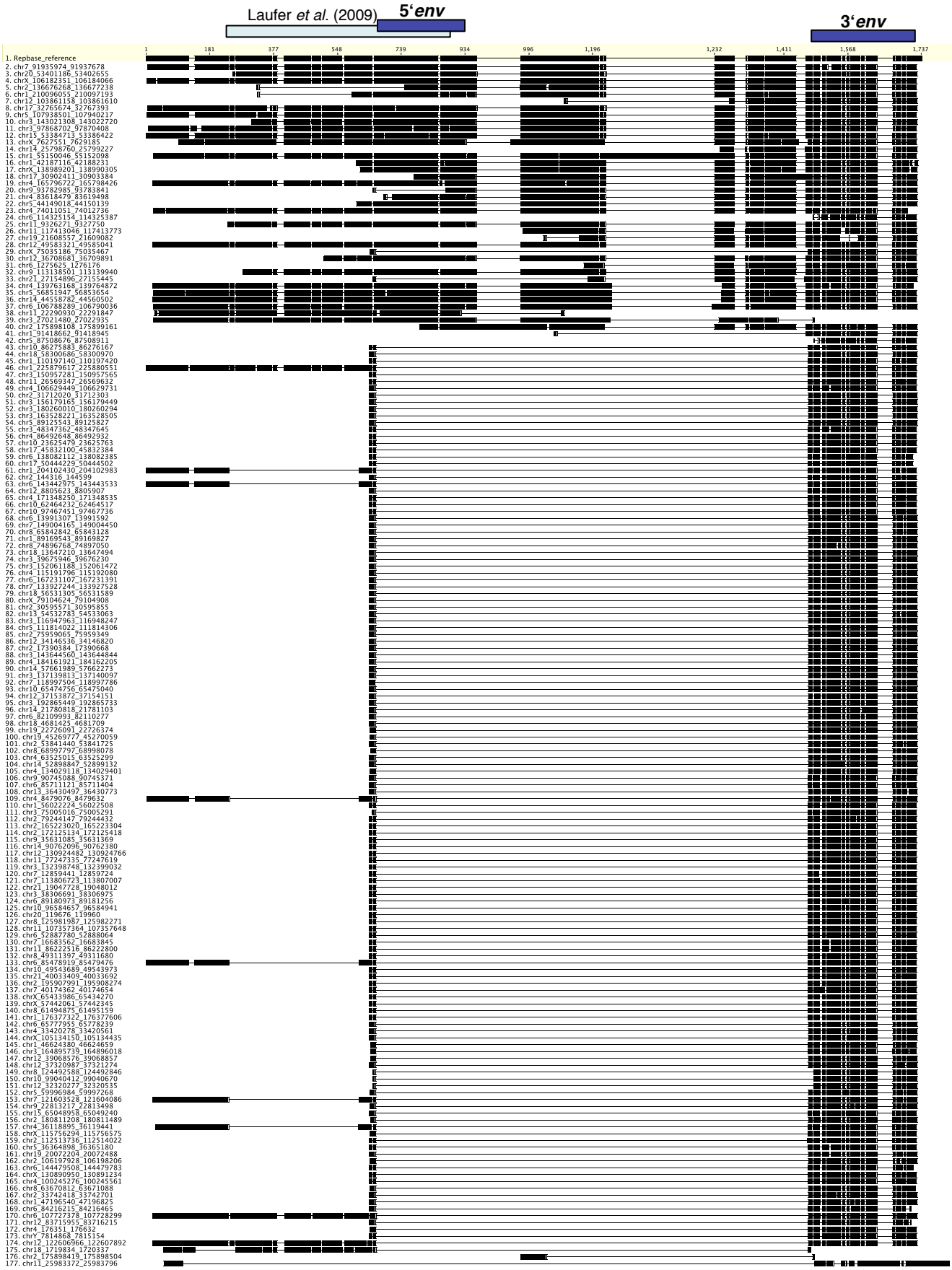
	3'*env*																							
	MS1		**MS2**		**MS3**		**MS4**		**MS5**		**MS6**		**MS7**		**H1**		**H2**		**H3**		**H4**		**H5**		**H6**		**H7**	
total # of reads	1712		3077		6547		1741		2470		2169		2201		1989		2096		2288		754		680		786		950	
average read length [bp]	233.1		237.1		232.9		150.6		162.5		149.5		154.7		238.8		233.1		236.6		155.6		165.9		167.0		148.4	
# of sequencing artefacts filtered	52		100		97		639		575		713		665		28		44		96		232		170		171		309	
number of reads with x mismatches	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
0	991	23.93	1853	85.87	3939	93.94	632	59.74	1327	73.49	1133	83.99	916	63.79	1247	22.22	1281	81.08	1472	77.27	378	75.60	380	78.35	447	76.67	455	74.84
1	495	11.95	743	34.43	1860	44.36	308	29.11	391	21.65	162	12.01	348	24.23	514	9.16	607	38.42	560	29.40	93	18.60	80	16.49	92	15.78	124	20.39
2	138	3.33	297	13.76	500	11.93	71	6.71	70	3.88	39	2.89	88	6.13	152	2.71	128	8.10	127	6.67	26	5.20	21	4.33	28	4.80	16	2.63
3	25	0.64	65	3.01	98	2.34	28	2.65	17	0.94	10	0.74	53	3.69	35	0.62	28	1.77	23	1.21	3	0.60	4	0.82	12	2.06	11	1.81
4	7	0.17	16	0.74	34	0.81	14	1.32	0	0.00	4	0.30	28	1.95	5	0.09	6	0.38	5	0.26	0	0.00	0	0.00	1	0.17	2	0.33
5	3	0.07	1	0.05	10	0.24	4	0.38	1	0.06	1	0.07	2	0.14	4	0.07	1	0.06	1	0.05	0	0.00	0	0.00	2	0.34	0	0.00
>5	1	0.02	2	0.09	9	0.21	1	0.09	0	0.00	0	0.00	1	0.07	4	0.07	1	0.06	4	0.21	0	0.00	0	0.00	1	0.17	0	0.00
sum	1660		2977		6450		1058		1806		1349		1436		1961		2052		2192		500		485		583		608	
reads >3 mismatches filtered	11		19		53		19		1		5		31		13		8		10		0		0		4		2	
reads contributing to further analysis	**1649**		**2958**		**6397**		**1083**		**1894**		**1451**		**1505**		**1948**		**2044**		**2182**		**522**		**510**		**611**		**639**	

*Results are given for the various brain tissue samples and the 5'*env* (top) and the 3'*env* (bottom) amplicons. Starting from a total number of reads, sequences were filtereds for sequencing artifacts and then BLAT searched against the hg18 human genome reference sequence. Numbers of sequences with 0 to >5 mismatches to the best matching HERV-W locus sequence are given. Sequences with >3 mismatches were excluded from the analysis. The resulting total number of sequences used for depicting transcription patterns of HERV-W loci is given at the bottom each. Note the somewhat lower number of sequence reads from samples H4–H7.

Supplemental Table S2 – Summary of filtering and assignment of Illumina/MiSeq sequence reads*

	H4	H5	H6	H7	MS4	MS5	MS6	MS7
total # of reads	1134146	877248	1028010	1124659	1171497	963202	1453740	760466
>1 best match or match <185 bp	107212	81021	77806	119896	115276	80011	83238	67669
unambiguous	1026934	796227	950204	1004763	1056221	883191	1370502	692797
in %	90.55	90.76	92.43	89.34	90.16	91.69	94.27	91.10
number of reads with x mismatches								
0	337277	287659	373024	376830	382577	338915	469821	230189
1	236466	185520	236888	268266	269981	225526	308496	182478
2	131043	97960	116222	129837	141457	113209	170553	102415
3	79721	56578	63798	67121	77828	61625	100496	54668
4	53489	37318	39965	41188	48244	37754	65272	33087
5	46459	76784	30464	29512	33818	28029	49367	22755
>5	142479	99408	89843	92009	102316	78133	206497	67205
discarded reads >3 mismatches in %	23.61	25.38	16.87	16.19	17.46	16.30	23.43	17.76
reads contributing to further analysis	784507	627717	789932	842054	871843	739275	1049366	569750
annotated as HERV17-int and contributing to further analysis	**778893**	**621236**	**785625**	**836399**	**865593**	**710593**	**1044013**	**565905**
in %	68.68	70.82	76.42	74.37	73.89	73.77	71.82	74.42

*Four MS lesions and healthy brain tissue samples each were also examined by Illumina/MiSeq sequencing technology. The total number of reads per sample is given. Reads were all 251bp in length due to the sequencing strategy. Sequences displaying more than one best match or a match length <185bp in BLAT analysis were removed from the analysis. Numbers of sequences with 0 to >5 mismatches to the best matching HERV-W locus are summarized. Note that a larger relative number of sequence reads was discarded compared to the 454/FLX dataset, yet, the total number of reads used for depicting HERV-W locus transcription patterns was nevertheless was higher. Also note that the much higher number of sequence reads also produced some non-HERV-W sequences that were identified and excluded from the analysis.



Supplemental Figure S1 – Location of 5'*env* and 3'*env* amplicons within the HERV-W *env* gene region. Depicted here is a multiple alignment of the *env* gene region of various HERV-W loci. Chromosomal positions of multiply aligned sequences are given on the left. Alignable regions are depicted as black boxes, gaps are indicated by horizontal lines. Location of an RT-PCR amplicon previously employed in Laufer et al. (2009) is also indicated. Note that a greater number of HERV-W loci uniformly lack an *env* gene region, yet those loci can be detected by the 3'*env* amplicon.

Supplemental Figure S2 – Structure of HERV-W loci identified as transcribed in this study.
Depicted is a multiple alignment of sequences of complete HERV-W loci identified as transcribed in our study. Location of HERV-W 5' and 3'LTR and proviral gene regions are indicated on the top. Alignable regions are depicted as black boxes, gaps are indicated by horizontal lines. Location of 5'*env* and 3'*env* amplicons employed in this study is also indicated.