Distribution of processed pseudogenes on each chromosome for different sequence identity and E-value cutoffs

The following table shows the number of processed pseudogenes on each chromosome for different cutoffs. PSSD1 indicates the number of processed pseudogenes that have frame disruptions; PSSD2 indicates the number of "putative" processed pseudogenes that do not have frame disruptions. The last row lists the correlation between the number of pseudogenes and the individual chromosome length. As can be seen, for different cutoffs, the numbers of processed pseudogenes on each chromosome are still proportional to the length of the chromosomes. This analysis corresponds to Figure 2 in the manuscript.

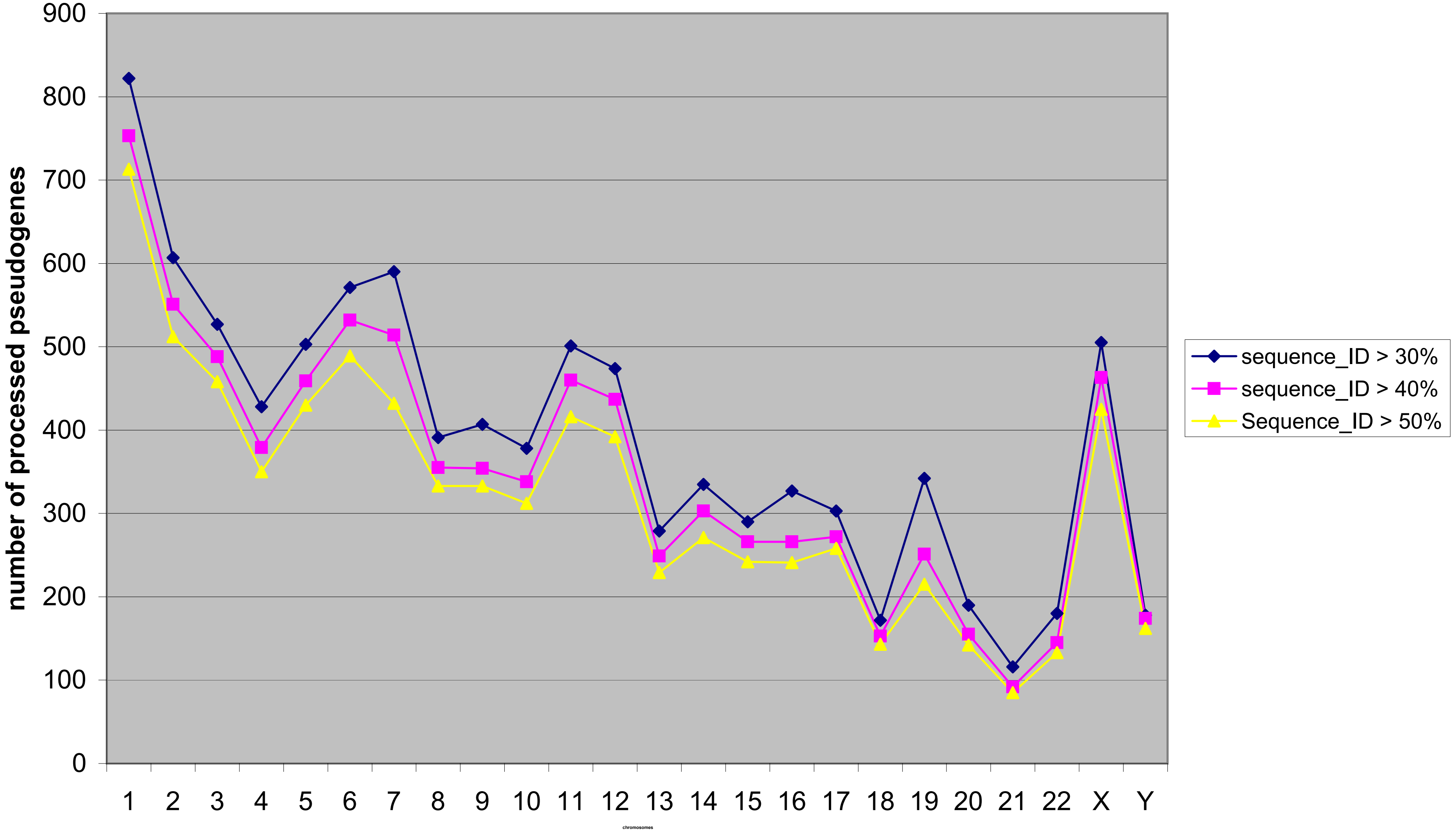
	Cutoffs	**1e-6, 0.3**		**1e-6, 0.4**		**1e-6, 0.5**		**1e-8, 0.3**		**1e-8, 0.4**		**1e-8, 0.5**	
Chr.	Chr. length (Mb)	PSSD1	PSSD2	PSSD1	PSSD2	PSSD1	PSSD2	PSSD1	PSSD2	PSSD1	PSSD2	PSSD1	PSSD2
1	247	822	78	753	73	713	63	790	74	736	71	703	61
2	241	607	64	551	60	512	59	588	63	542	59	505	58
3	195	527	55	488	46	458	41	516	45	485	42	456	37
4	192	428	29	379	25	350	23	406	28	363	24	340	22
5	181	503	53	459	52	430	49	478	50	445	49	420	46
6	170	571	35	532	29	489	28	550	35	523	29	484	28
7	157	590	74	514	57	432	48	557	68	491	53	419	44
8	144	391	24	355	23	333	22	377	24	348	23	329	22
9	132	407	29	354	26	333	23	391	28	342	25	324	22
10	134	378	22	338	22	312	17	344	21	320	21	301	17
11	137	501	72	460	65	416	57	484	69	451	63	407	55
12	131	474	53	437	50	392	37	461	50	428	47	387	34
13	113	279	21	249	19	229	19	267	20	242	18	226	18
14	104	335	46	303	43	271	40	321	43	294	42	267	39
15	99	290	18	266	18	242	18	272	15	253	15	234	15
16	82	327	34	266	33	241	31	308	32	256	32	235	30
17	80	303	43	272	42	258	39	291	40	261	39	250	36
18	78	172	16	153	16	143	15	163	16	148	16	139	15
19	60	342	39	251	36	215	27	321	38	240	35	208	26
20	63	190	12	155	11	142	9	183	10	153	9	141	7
21	45	116	6	92	2	85	2	111	6	91	2	84	2
22	48	180	22	145	21	133	21	171	22	140	21	128	21
X	149	505	39	463	37	425	32	487	39	454	37	419	32
Y	58	178	14	174	14	162	14	135	8	132	8	128	8
	Correlation with chr. length	91%	68%	92%	68%	93%	71%	91%	69%	92%	69%	93%	72%

(Continued)

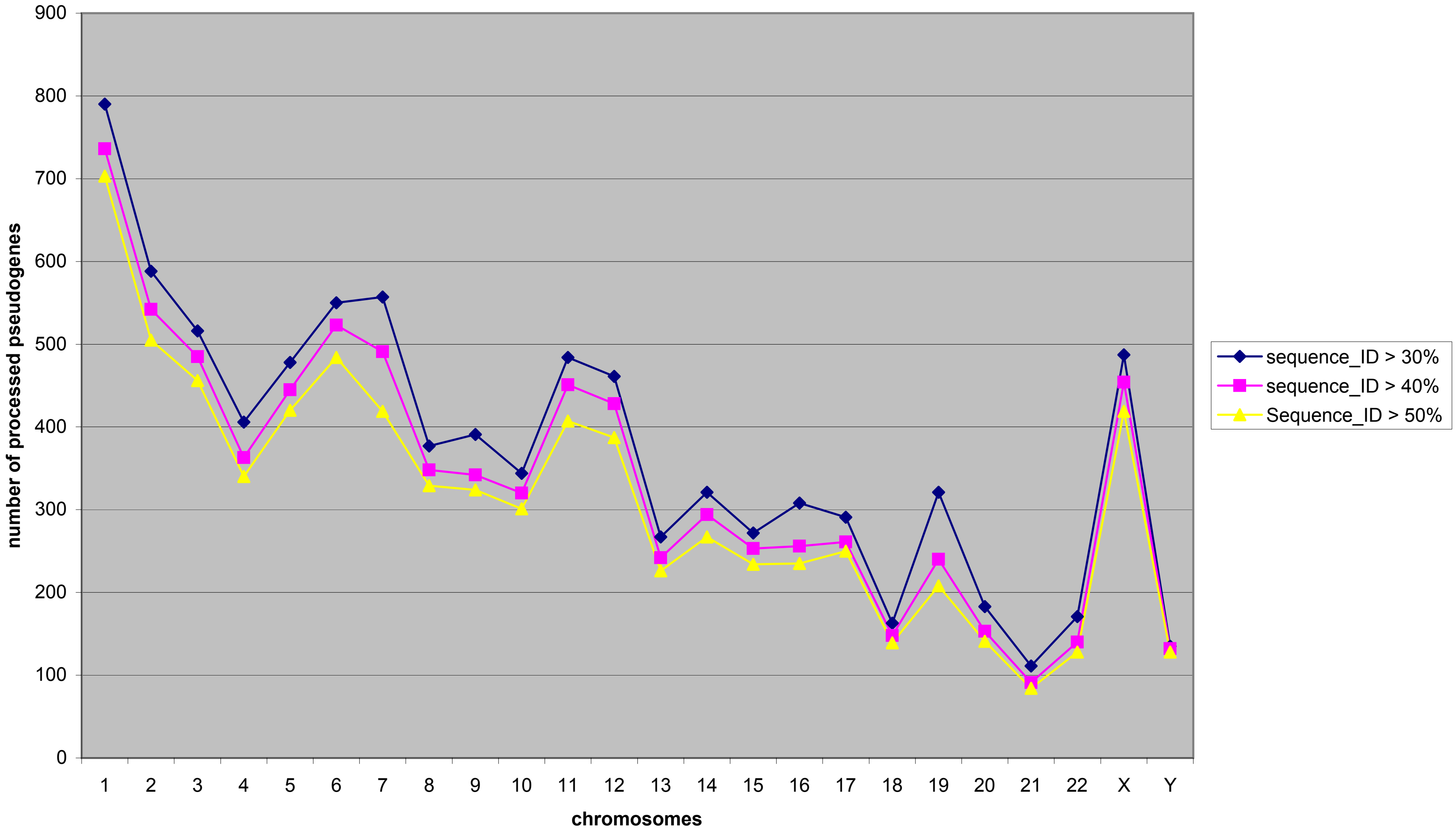
	Cutoffs	1e-10, 0.3		1e-10, 0.4		1e-10, 0.5		1e-12, 0.3		1e-12, 0.4		1e-12, 0.5	
Chr.	**Chr. length (Mb)**	**PSSD1**	**PSSD2**	**PSSD1**	**PSSD2**	**PSSD1**	**PSSD2**	**PSSD1**	**PSSD2**	**PSSD1**	**PSSD2**	**PSSD1**	**PSSD2**
1	247	703	61	764	68	719	65	694	55	734	66	698	63
2	241	505	58	565	61	526	57	495	56	532	60	506	57
3	195	456	37	502	43	474	40	450	35	482	42	460	39
4	192	340	22	390	28	351	24	331	22	374	28	338	24
5	181	420	46	457	47	431	47	408	44	438	45	418	45
6	170	484	28	530	32	508	29	474	28	511	31	492	29
7	157	419	44	512	65	454	52	400	43	478	60	432	51
8	144	329	22	361	23	339	22	324	21	347	22	328	21
9	132	324	22	374	24	332	22	318	19	366	24	329	22
10	134	301	17	323	18	309	18	297	15	303	16	294	16
11	137	407	55	476	64	446	59	405	52	456	64	431	59
12	131	387	34	443	48	416	45	380	32	429	47	407	45
13	113	226	18	255	19	236	18	222	18	246	19	230	18
14	104	267	39	307	40	286	40	265	38	292	40	274	40
15	99	234	15	262	14	247	14	231	14	245	14	235	14
16	82	235	30	296	29	250	29	233	27	281	29	242	29
17	80	250	36	271	39	250	38	241	35	260	37	244	36
18	78	139	15	153	16	144	16	138	15	146	15	139	15
19	60	208	26	287	33	224	31	198	23	268	31	215	30
20	63	141	7	178	7	151	7	140	6	165	7	145	7
21	45	84	2	101	5	87	2	81	2	88	4	80	2
22	48	128	21	160	22	135	21	125	21	149	21	129	21
X	149	419	32	461	38	436	36	411	31	444	37	422	35
Y	58	128	8	120	5	117	5	115	5	110	2	108	2
	Correlation with chr. length	93%	72%	91%	70%	92%	70%	93%	72%	91%	70%	92%	69%

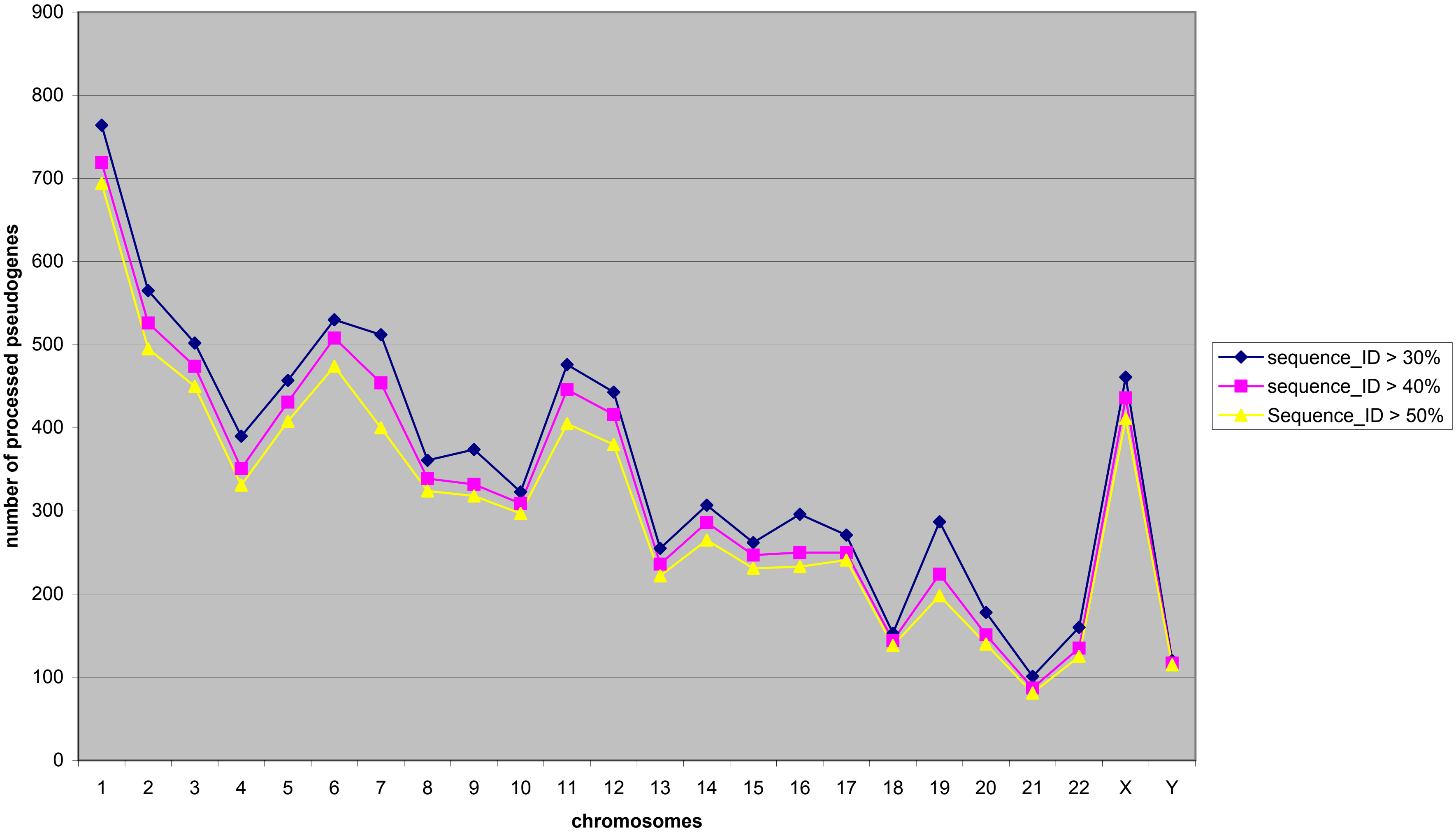
The following four graphs show the number of processed pseudogenes for the same E-value cutoff and different sequence identity cutoff.

number of processed pseudogenes on each chromosome
(E_value < 1e-6)



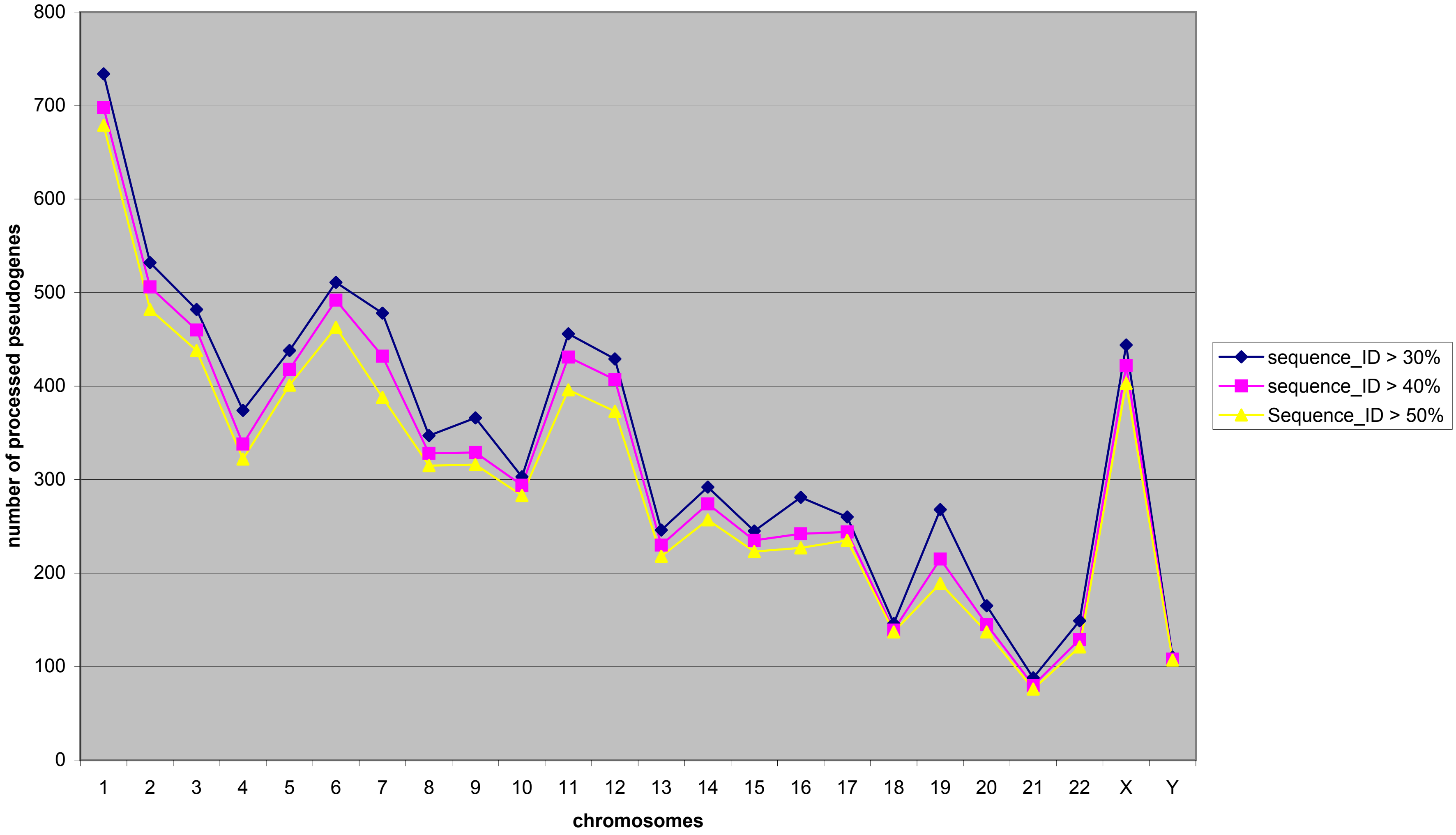
number of processed pseudogenes on each chromosome
(E_value < 1e-8)





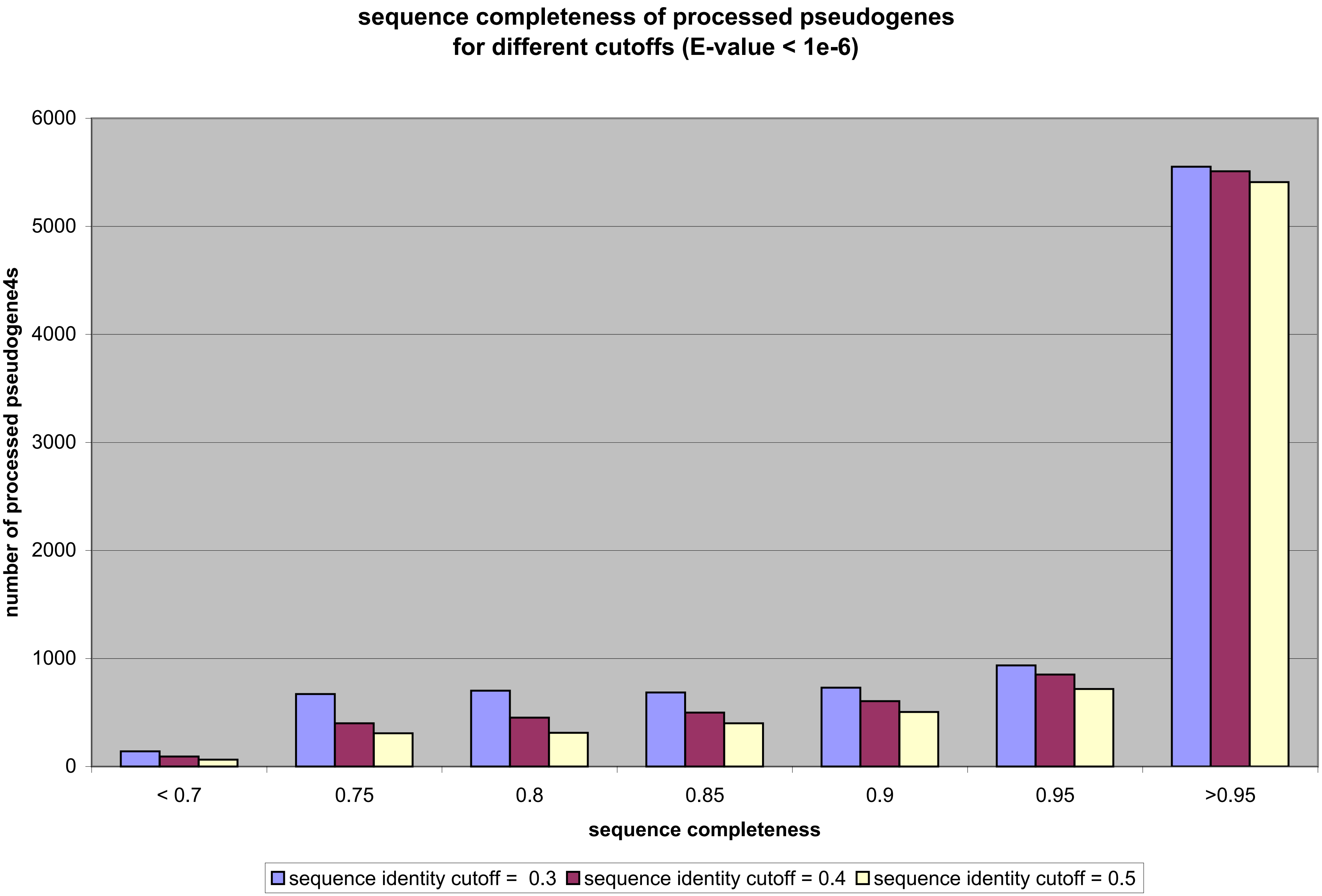
number of processed pseudogenes on each chromosome
(E_value < 1e-10)

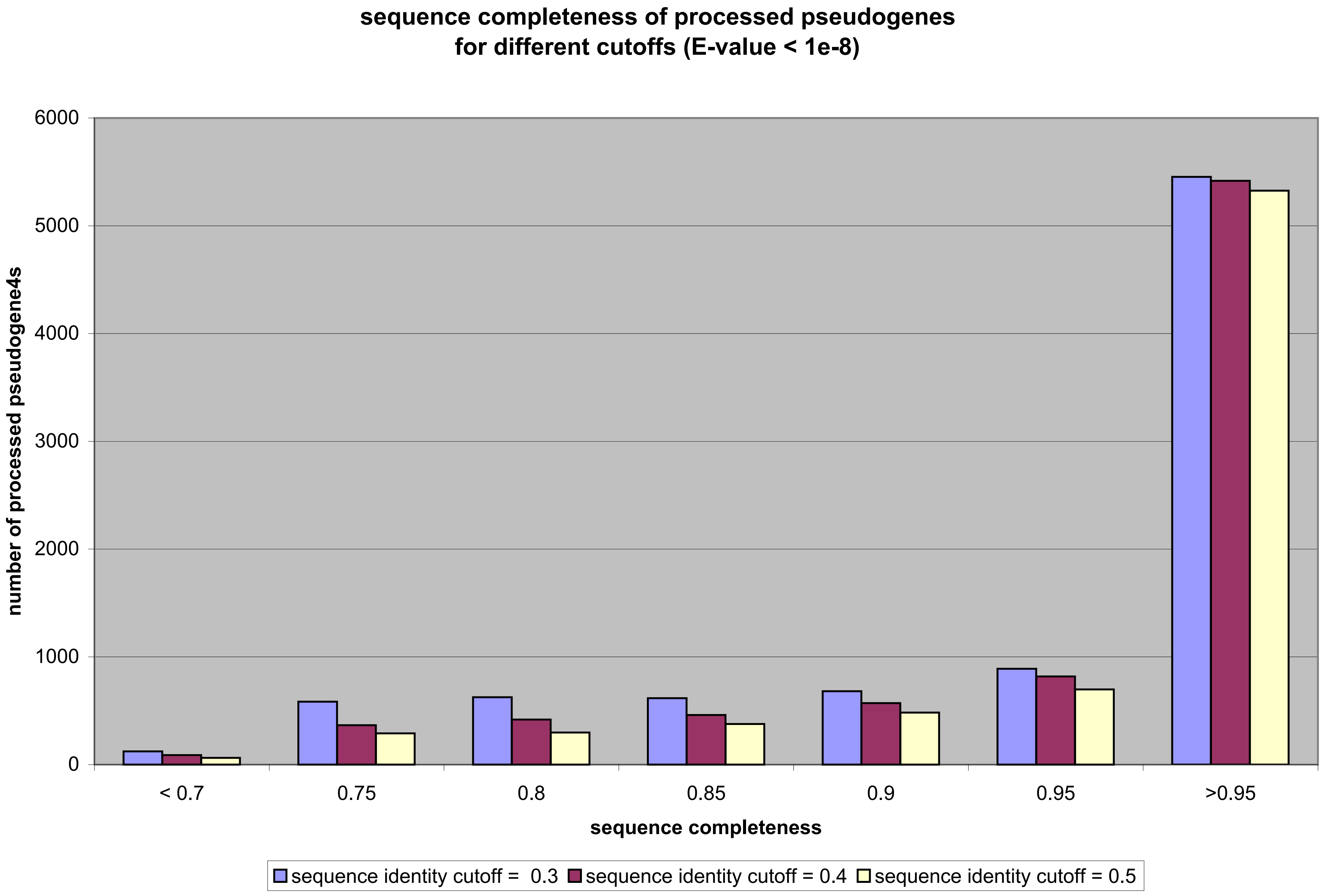
number of processed pseudogenes on each chromosome
(E_value < 1e-12)



Sequence Statistics of Processed Pseudogenes After Different Cutoffs

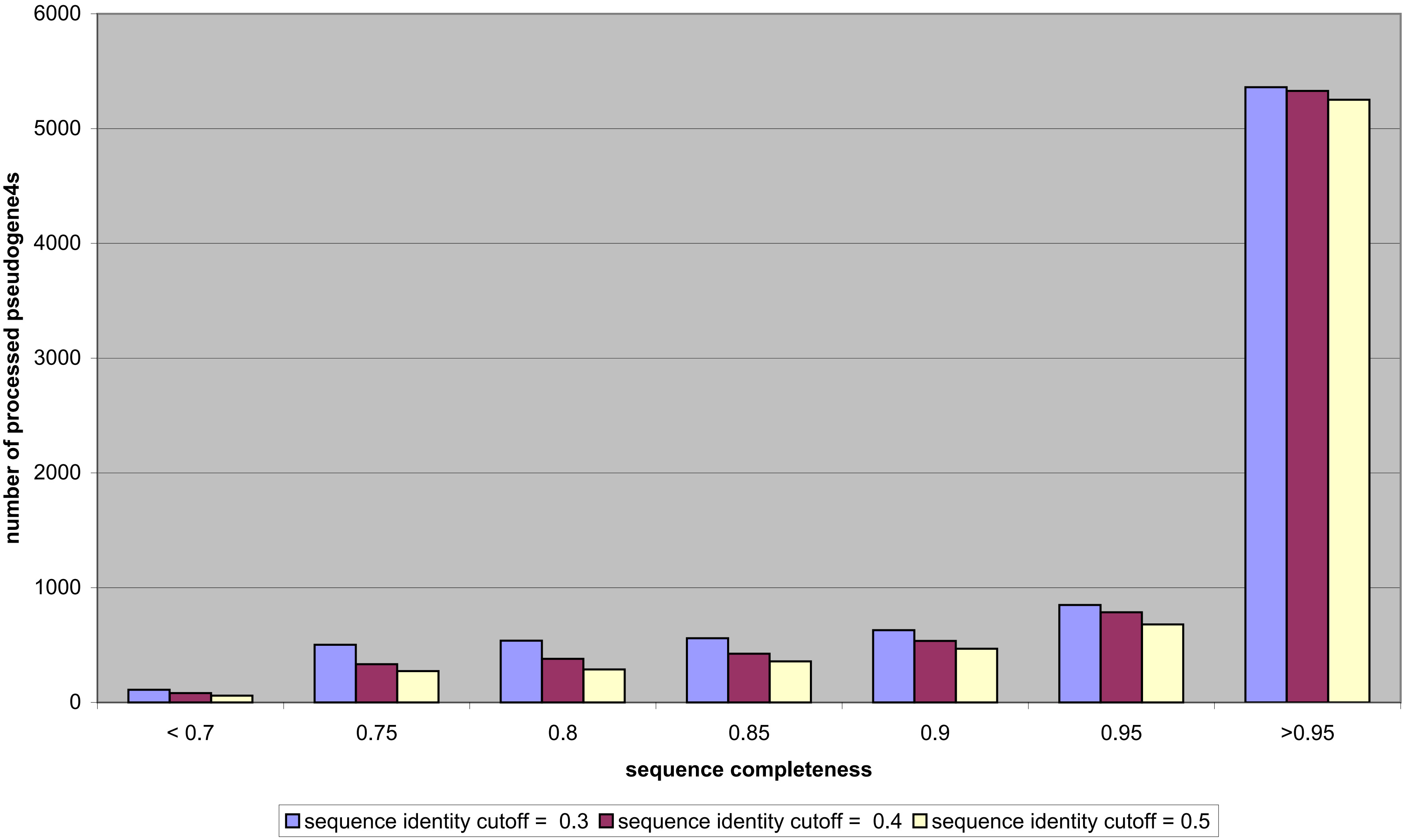
The following graphs show that, after different combinations of sequence identity cutoffs and E-value cutoffs, the observations shown in Figure 3A, 3B and 3C are still true. Namely, the majority of the processed pseudogenes still have very high sequence completeness (even after 0.5 sequence completeness cutoff). Also the numbers of frame disruptions in the pseudogenes still have a power-law behavior (See Figure 3C in the paper), regardless of the cutoffs used in the selection procedure.

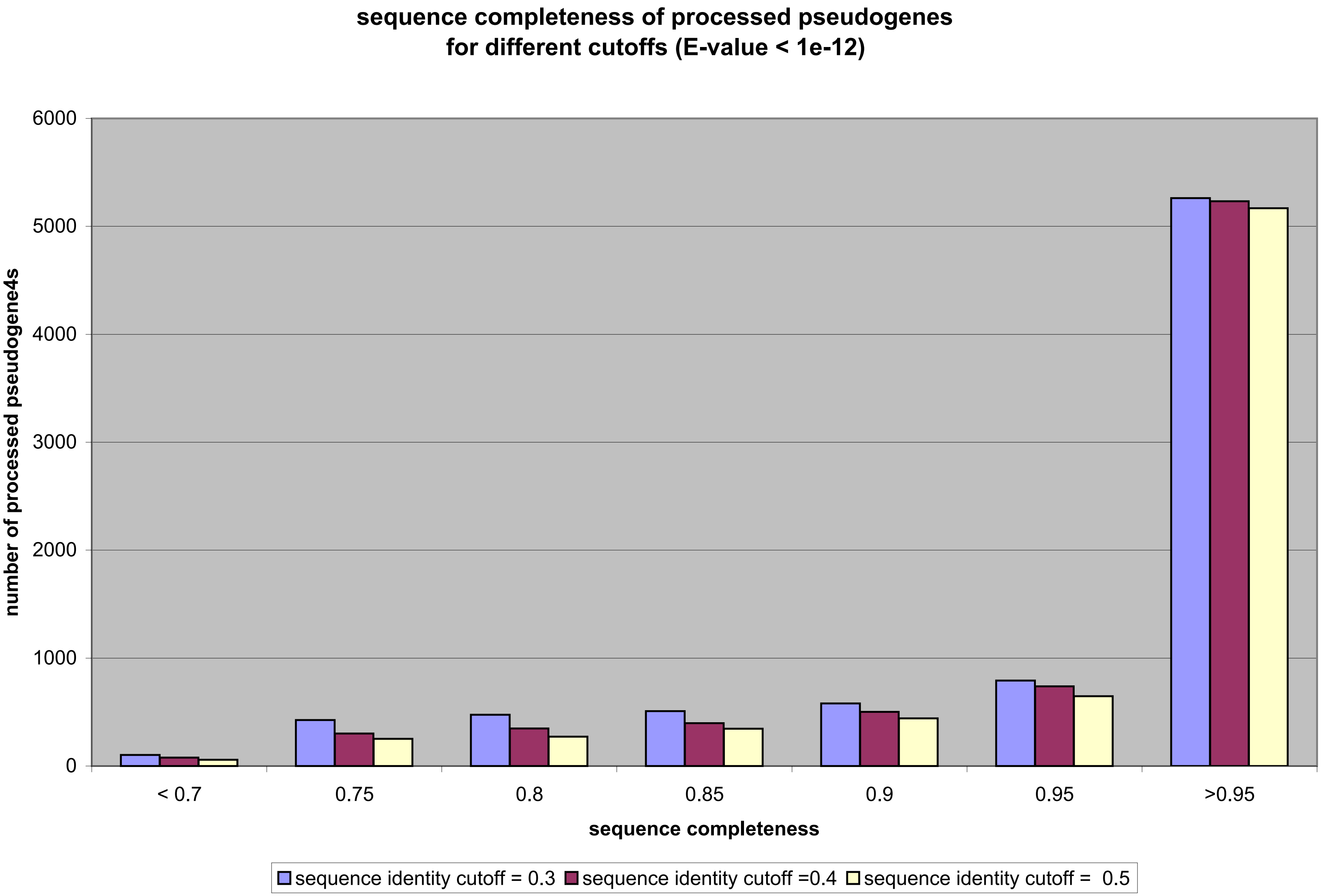

sequence completeness of processed pseudogenes
for different cutoffs (E-value < 1e-6)
number of processed pseudogene4s
6000
5000
4000
3000
2000
1000
0
< 0.7
0.75
0.8
0.85
0.9
0.95
>0.95
sequence completeness
sequence identity cutoff = 0.3
sequence identity cutoff = 0.4
sequence identity cutoff = 0.5



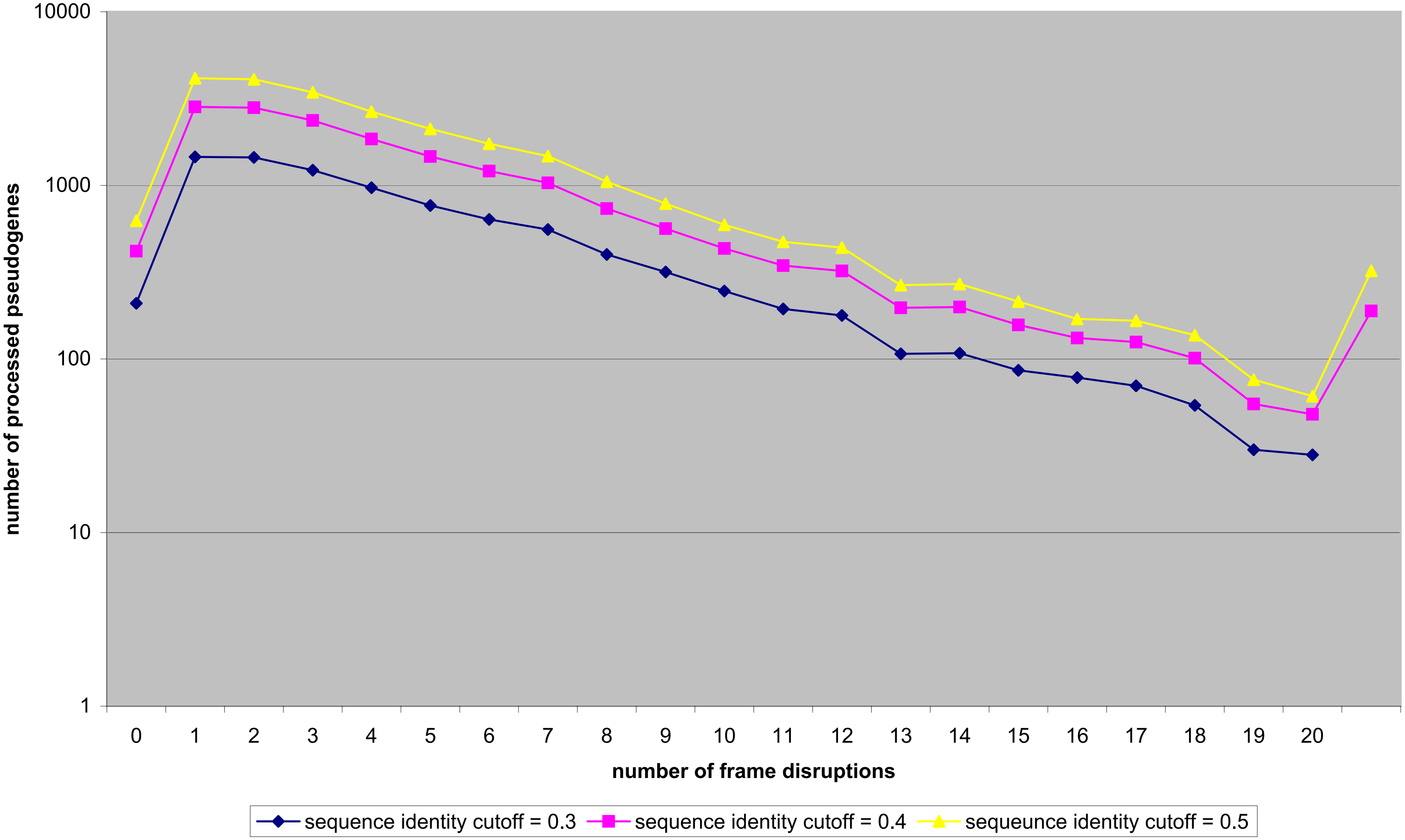
sequence completeness of processed pseudogenes
for different cutoffs (E-value < 1e-8)

sequence completeness of processed pseudogenes
for different cutoffs (E-value < 1e-10)

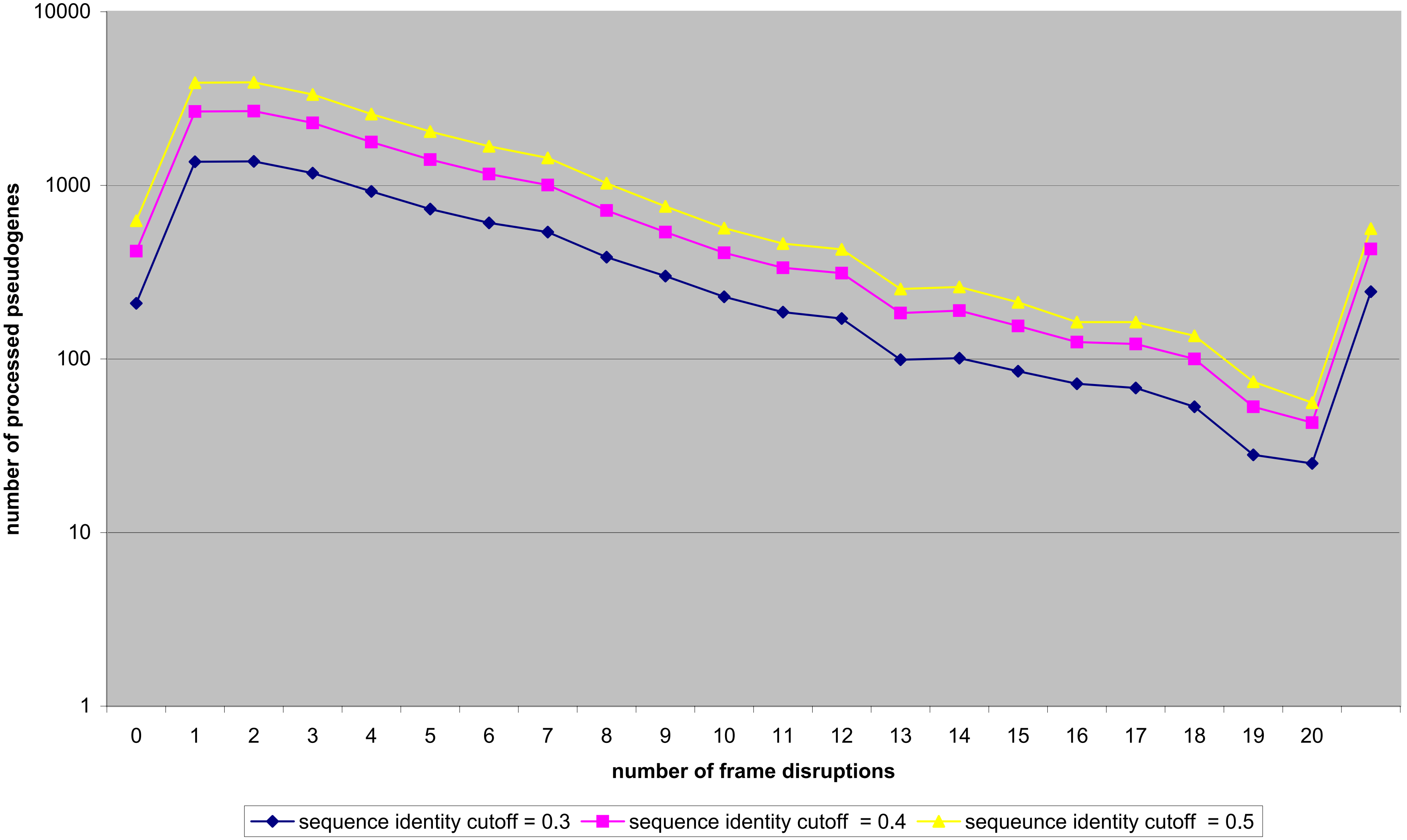


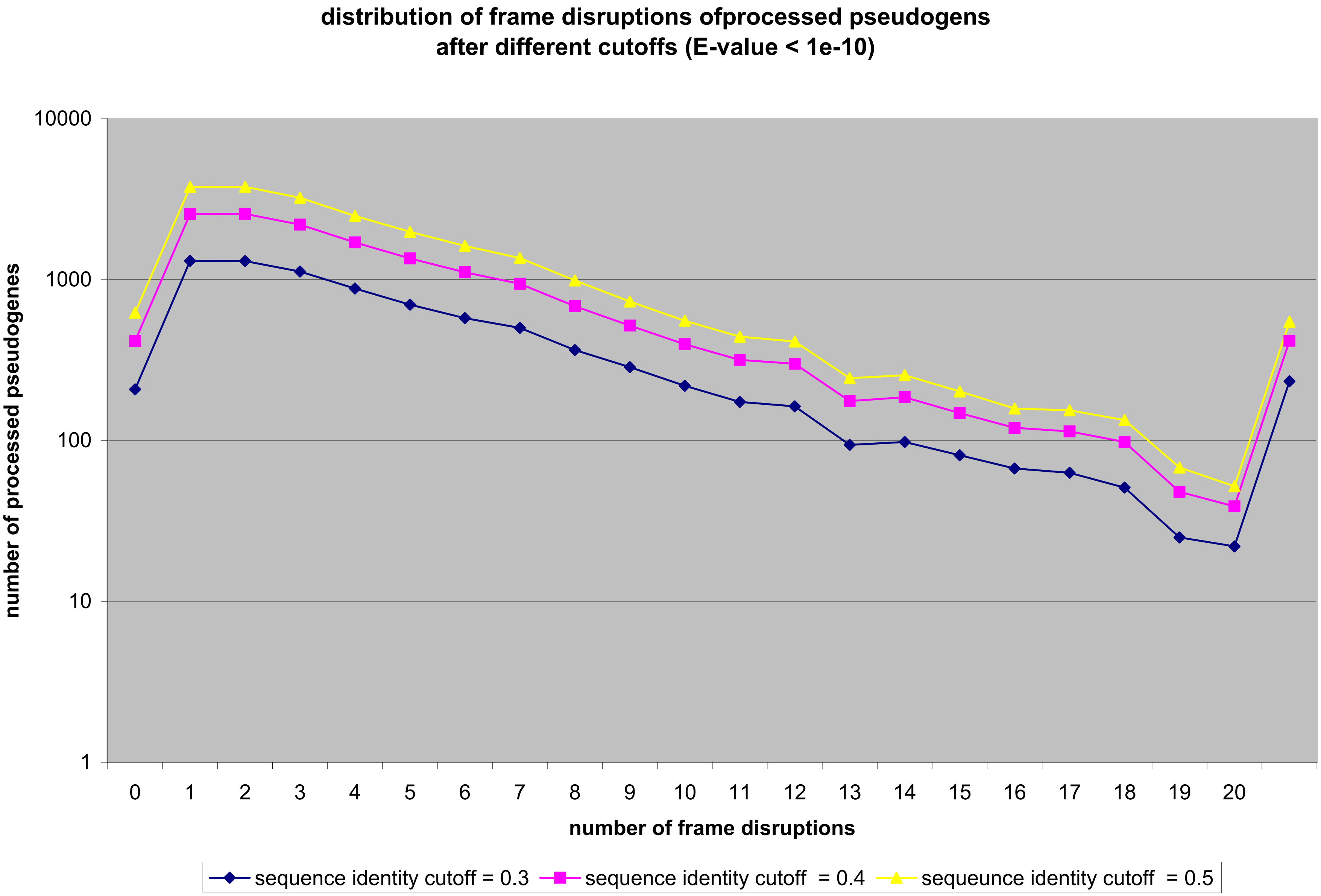

sequence completeness of processed pseudogenes
for different cutoffs (E-value < 1e-12)
number of processed pseudogene4s
6000
5000
4000
3000
2000
1000
0
< 0.7
0.75
0.8
0.85
0.9
0.95
>0.95
sequence completeness
sequence identity cutoff = 0.3
sequence identity cutoff =0.4
sequence identity cutoff = 0.5

distribution of frame disruptions ofprocessed pseudogens
after different cutoffs (E-value < 1e-6)

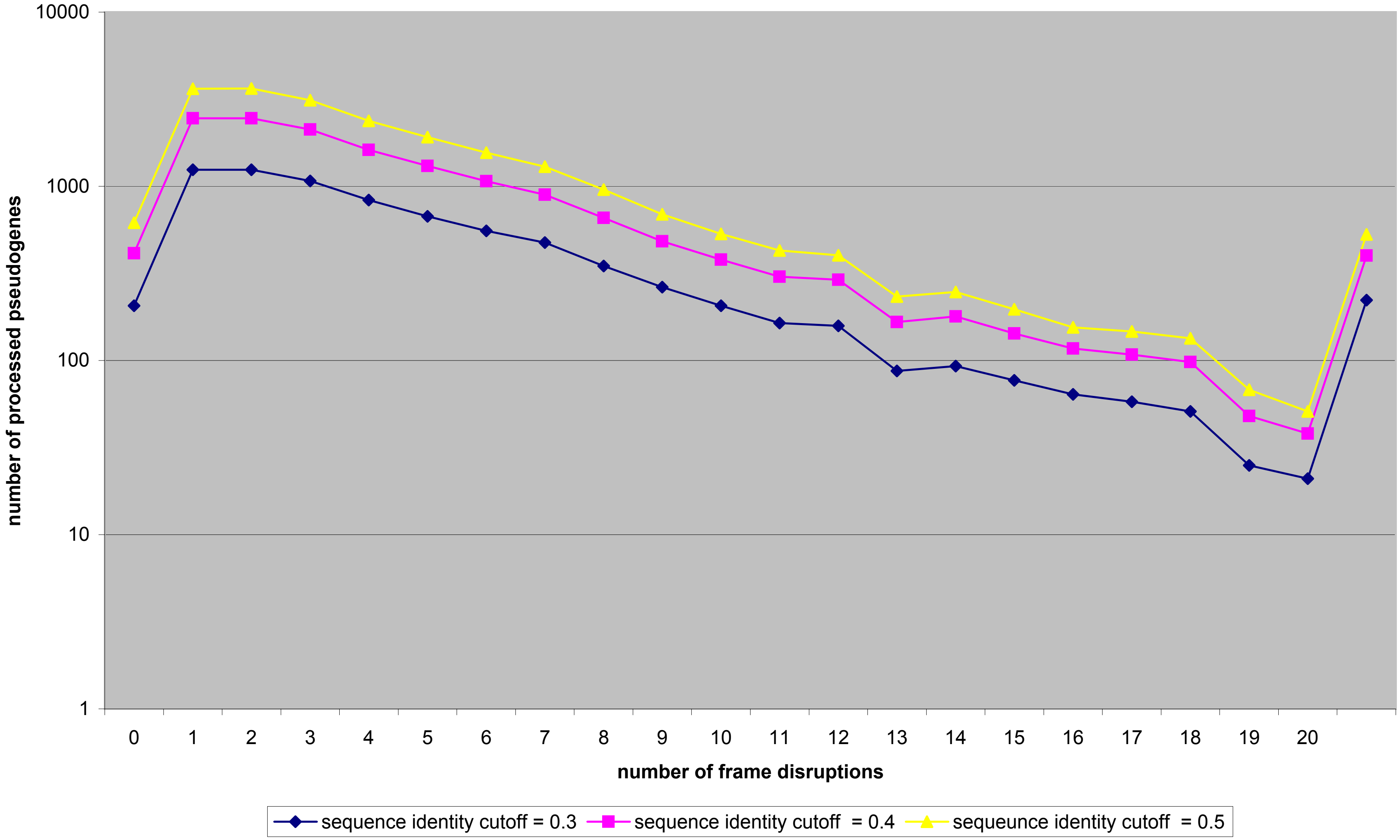


distribution of frame disruptions ofprocessed pseudogens
after different cutoffs (E-value < 1e-8)



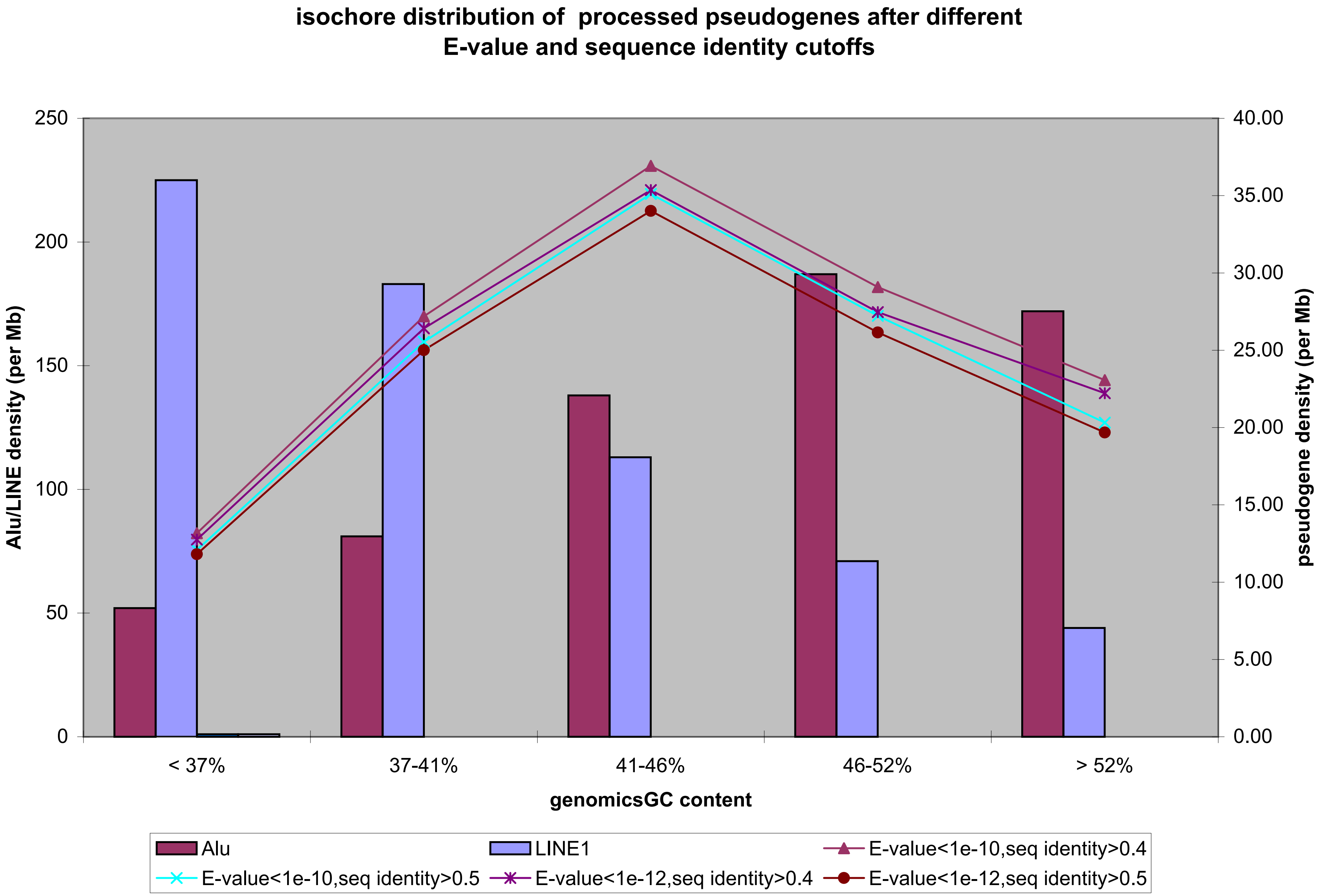

distribution of frame disruptions ofprocessed pseudogens
after different cutoffs (E-value < 1e-10)
10000
1000
100
10
1
number of processed pseudogenes
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
number of frame disruptions
sequence identity cutoff = 0.3
sequence identity cutoff = 0.4
sequeunce identity cutoff = 0.5

distribution of frame disruptions ofprocessed pseudogens
after different cutoffs (E-value < 1e-12)



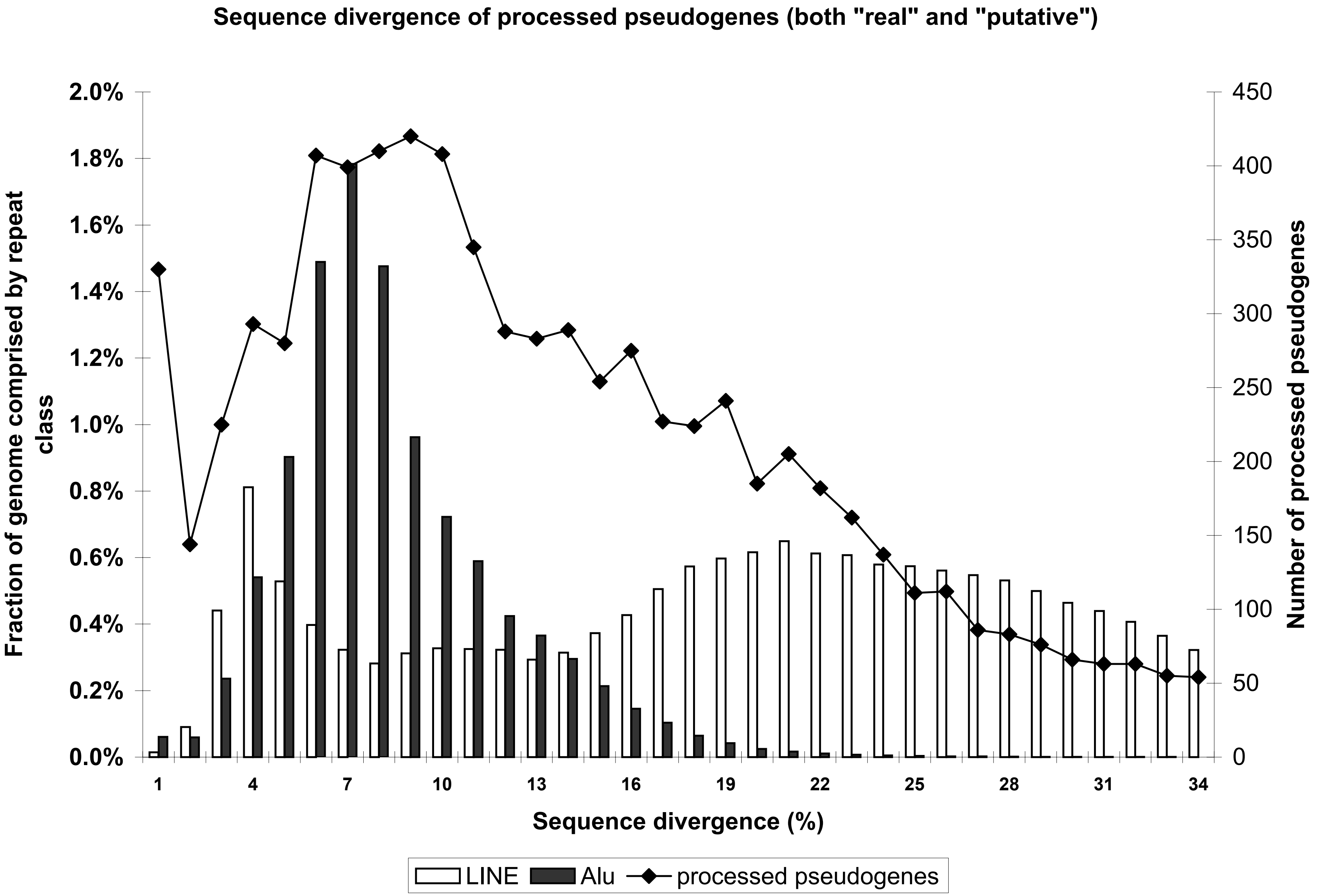
Isochore Distribution of the Pseudogenes After Different Cutoffs

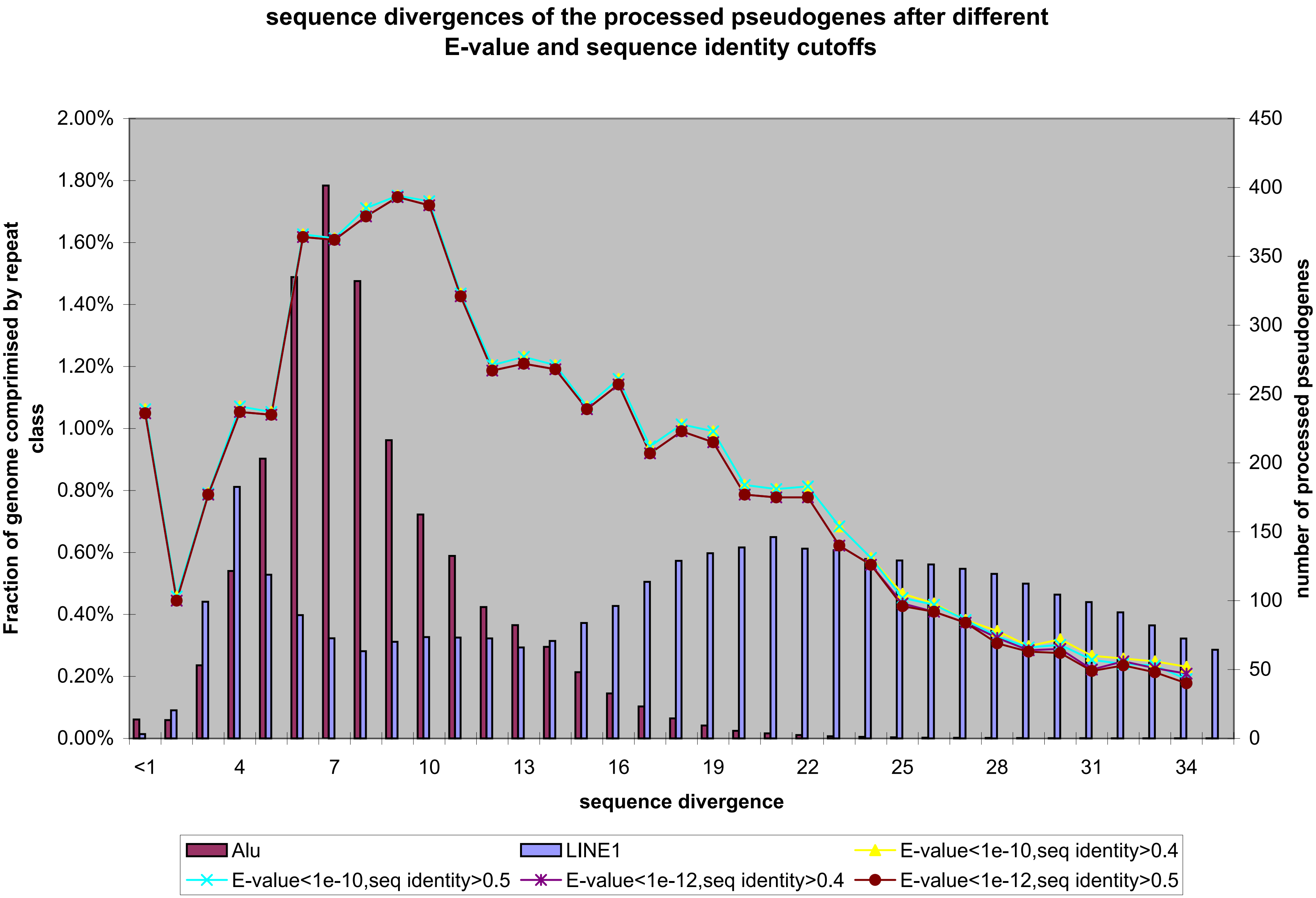
The following graphs show that, after different combinations of sequence identity cutoffs and E-value cutoffs, the conclusion drawn shown in Figure 4 is still true. Namely, after more stringent selection filtering (E-value being raised to 1e-12 from 1e-10, and sequence identity raised from 0.4 to 0.5), the isochore distribution described in the text still remains the same.


isochore distribution of processed pseudogenes after different
E-value and sequence identity cutoffs
Alu/LINE density (per Mb)
250
200
150
100
50
0
pseudogene density (per Mb)
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
< 37%
37-41%
41-46%
46-52%
> 52%
genomicsGC content
Alu
LINE1
E-value<1e-10,seq identity>0.4
E-value<1e-10,seq identity>0.5
E-value<1e-12,seq identity>0.4
E-value<1e-12,seq identity>0.5

Distribution of sequence divergence of the Pseudogenes After Different Cutoffs

The following two graphs show that, after different combinations of sequence identity cutoffs and E-value cutoffs, the conclusion drawn shown in Figure 6 is still true. Namely, after more stringent selection filtering (E-value being raised to 1e-12 from 1e-10, and sequence identity raised from 0.4 to 0.5), the age distribution of the processed pseudogenes remains almost the same. Please see the text for discussion on the comparison between processed pseudogenes and Alu and LINE1 elements.


Sequence divergence of processed pseudogenes (both "real" and "putative")
Fraction of genome comprised by repeat class
2.0%
1.8%
1.6%
1.4%
1.2%
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
Number of processed pseudogenes
450
400
350
300
250
200
150
100
50
0
1
4
7
10
13
16
19
22
25
28
31
34
Sequence divergence (%)
LINE
Alu
processed pseudogenes


sequence divergences of the processed pseudogenes after different E-value and sequence identity cutoffs
Fraction of genome comprimised by repeat class
number of processed pseudogenes
sequence divergence
2.00%
1.80%
1.60%
1.40%
1.20%
1.00%
0.80%
0.60%
0.40%
0.20%
0.00%
450
400
350
300
250
200
150
100
50
0
<1
4
7
10
13
16
19
22
25
28
31
34
Alu
LINE1
E-value<1e-10,seq identity>0.4
E-value<1e-10,seq identity>0.5
E-value<1e-12,seq identity>0.4
E-value<1e-12,seq identity>0.5