CHAPTER 4 RESULTS AND DISCUSSIONS

4.1 Consideration of the maximum distance parameter

The maximum distance which is one of REannotate parameters indicates the greatest distance that is allowed for joining two fragments of the same element. This parameter was required in the step of HERV defragmentation, a process in the data preparation. There is still no standard value for this parameter in the defragmentation. However, from the previous works, the values of the parameter were varied between 500 bp to 30kb [54, 55].

To making a decision how much the value should be, REannotate was run several times with the varied distances to observe the sensitivity of this value to the defragmentation results. The distances used in this consideration are 10, 20, 50, 100, 200, 300, 400, 500, 1k, 2k, 5k, 10k, 20k, 40k, and 50k in base pairs. The number of defragmented elements, non-defragmented elements, and total elements were observed while changing the values of the distance parameters. The defragmented elements are the HERV elements originated from combining more than one fragments together. Unlike the defragmented, the non-defragmented elements are composed of only one fragment, no joining to any other fragments. Total number of all elements resulting from the defragmentations is the summation of both the number of the defragmented elements and the non-defragmented elements. The numbers of all elements resulting from the defragmentations are shown in Figure 4.1.

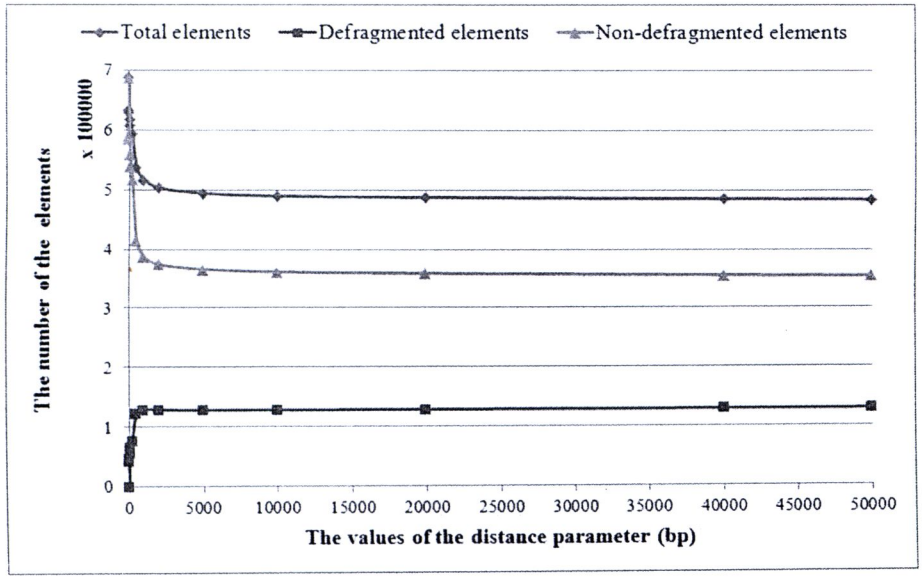


Figure 4.1 The numbers of the elements resulting from the defragmentations using different values of the distance parameter

According to Figure 4.1, the number of all elements and non-defragmented elements tend to decrease when the parameter values are increased, because there are more single fragments that must be used in the joining events to originate more defragmented

elements. This also leads to increasing the number of the defragmented elements, when the values of the distance tend to increase likewise.

In addition, the number of all elements tends to change rapidly in the beginning of varying the parameter values. This could be suggested that the optimal distance should not be too less, because many defragmentable elements would be ignored in that case. Moreover, from Figure 4.1, the number of all elements seems to be not changed anymore, when the distance values are increased more than around 10kb or 20kb. Thus, there is no need to use a large number for the distance parameter, because the smaller value of the distance can almost represent the overall results of the defragmentation as well. Why the small value of the distance is preferred is that we would like to avoid collecting too long non-retroviral sequences in the defragmented elements. In conclusion, the suitable value of the distance should be the minimum value which leads to missing the least number of the defragmented elements. Anyway, it is not necessary to obtain the exact value of the distance in this work, but there is only a need for an approximate value with an acceptable result.

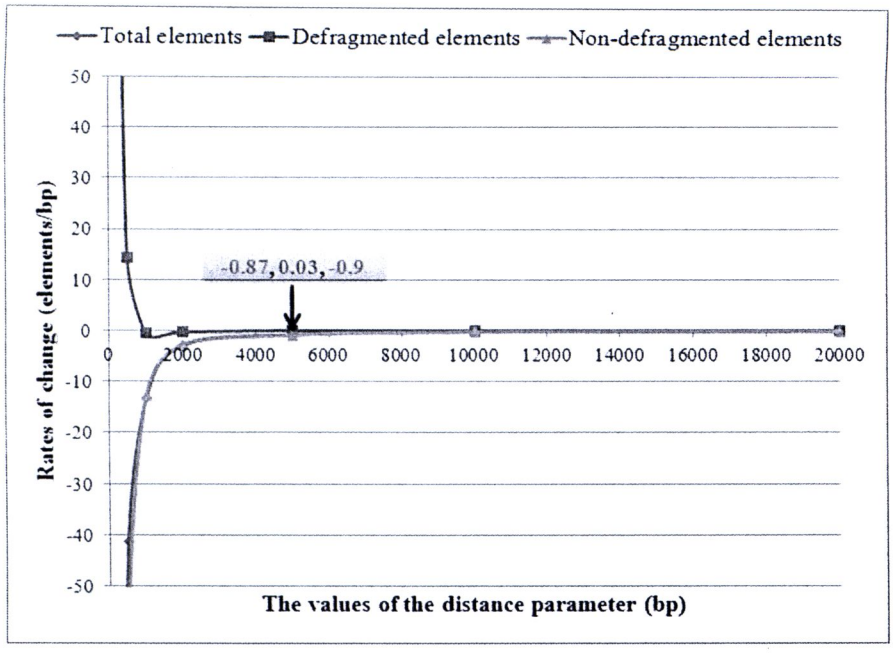


Figure 4.2 The rates of change of all elements resulting from the defragmentation using different values of the distance parameter

In Figure 4.2, the rates of change of the number of all elements were additionally observed to clarify the distance value which could be considered no changes in terms of the number of all elements. It is somewhat clearly that the rates of change begin converging to zero at the distance value equal to 5,000 bp, pointed by an arrow in Figure 4.2. In other words, it would be considered that there is no change occurred even though the distance values will be increasing from 5,000 bp to 10kb. Therefore, in this case, the chosen value should not be more than 5,000 bp.

In addition, the number of the defragmented elements, resulting from the defragmentation, and the coverage percentages was finally observed to determine the value of the distance parameter (Table 4.1). Only the maximum distances lower than 5,000 bp were considered here. To calculate the coverage percentages, the number of the defragmented elements of each distance value was compared to the number of the defragmented elements equal to 130,060. This number is the maximum number of the defragmented elements among of the distances considered already stable in terms of the number of the defragmented elements obtained.

Table 4.1 The number of the defragmented elements, resulting from the HERV defragmentation using REannotate with different values of the distance parameters, and their coverage percentages

The maximum distances (bp)	The number of defragmented elements	Coverage percentages
10	42,489	32.67%
20	45,678	35.12%
50	57,284	44.04%
100	66,500	51.13%
200	77,070	59.26%
300	91,343	70.23%
400	111,250	85.53%
500	122,437	94.14%
1,000	129,607	99.65%
2,000	129,007	99.19%

As mentioned before, the appropriate value of the distance parameter should lead to the least missing of the defragmented elements, so the coverage percentages of defragmented elements should not be lower than 90%. Furthermore, covering too many defragmented elements could not be a good way for two reasons. The first one is that the higher value of the distance parameter is required to obtain the higher number of the defragmented elements, which could lead to collect long non-retroviral sequences in the defragmented elements. In addition, the longer distance in the higher coverage percentages would lead to collect more wrong fragments in the defragmented elements, because they are just accidentally able to be joined.

Therefore, according to Table 4.1, the appropriate value of the distance parameter would be equal to 500 bp, leading to cover the defragmented elements 94% of the maximum number of the defragmented elements. Furthermore, this value of the distance parameter is also consistent with the value used in Transposon Cluster Finder (TCF) [54], a software package for identifying and defragmenting transposon clusters in the human genome.

4.2 Web interface

HERV Profiler has provided two main tracks for investigating HERVs nearby users' genes interested. The first one is to visualize neighboring HERVs of the genes. The second track is the HERV profiling that would purpose a set of characteristics which are over-represented among a gene list by performing a statistical approach. An overview of the tool is summarized and written on the homepage, as shown in Figure 4.3.

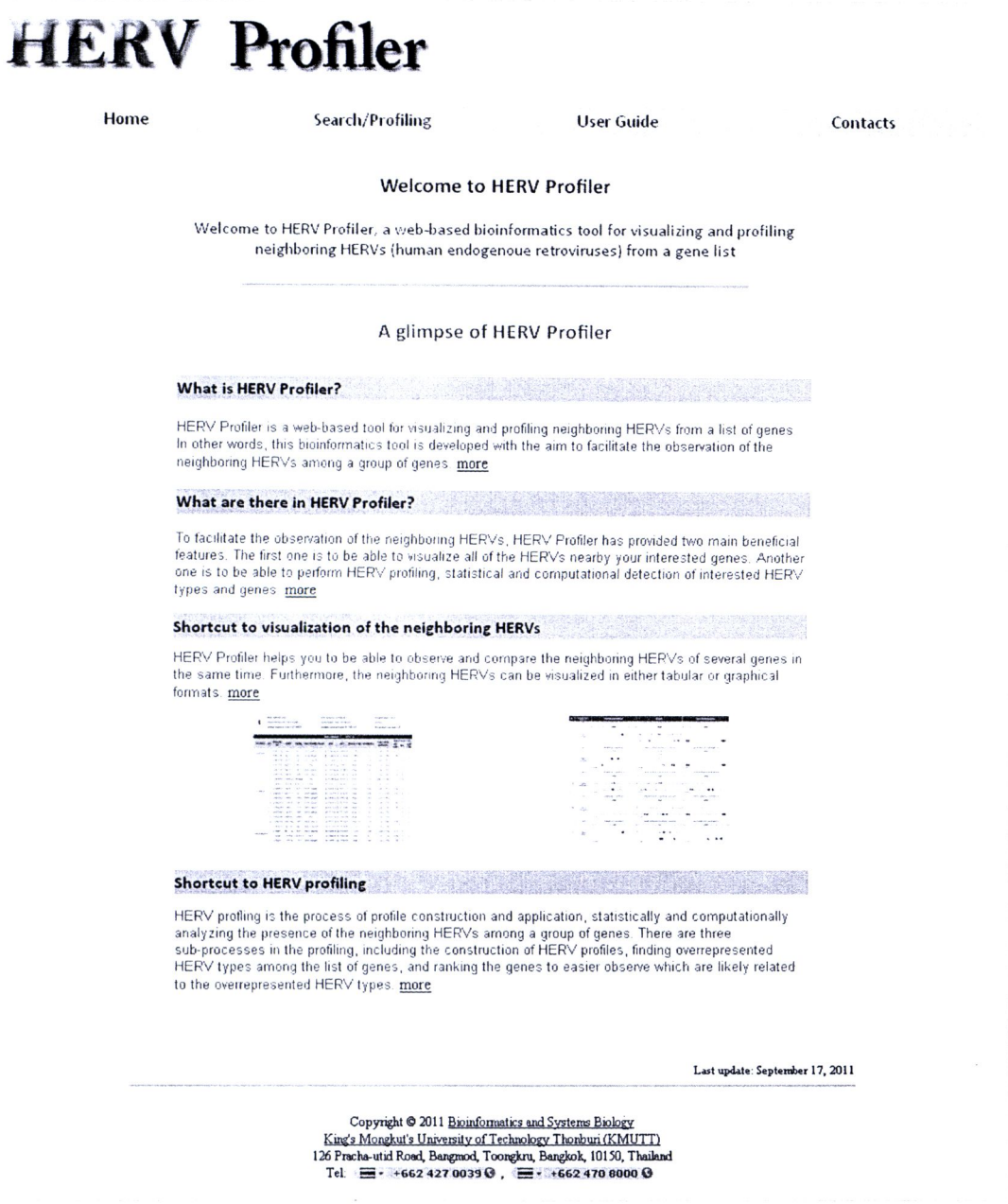


Figure 4.3 Homepage of HERV Profiler

The first welcome page contains briefly general information describing what the tool is used for and what there are provided in the tool. For more information about the tool,

users can go directly to the page of user guide or click on "more" to look forward the details of the desired section.

Users can start to begin using HERV Profiler by selecting on Search/Profiling. The page of searching and profiling is shown in Figure 4.4. There are two kinds of inputs that are required. The first one is a list of gene symbols which each should be separated by commas. The second type of inputs is a specification of HERVs that would like to include in both visualizing and profiling. There are five characteristics provided in the specification, including covering distance, HERV orientation, type of truncation patterns, superfamily, and minimum intactness ratios. The covering distance is the value determining the neighborhood of the genes where is far away from the transcription start site and termination site of the genes.

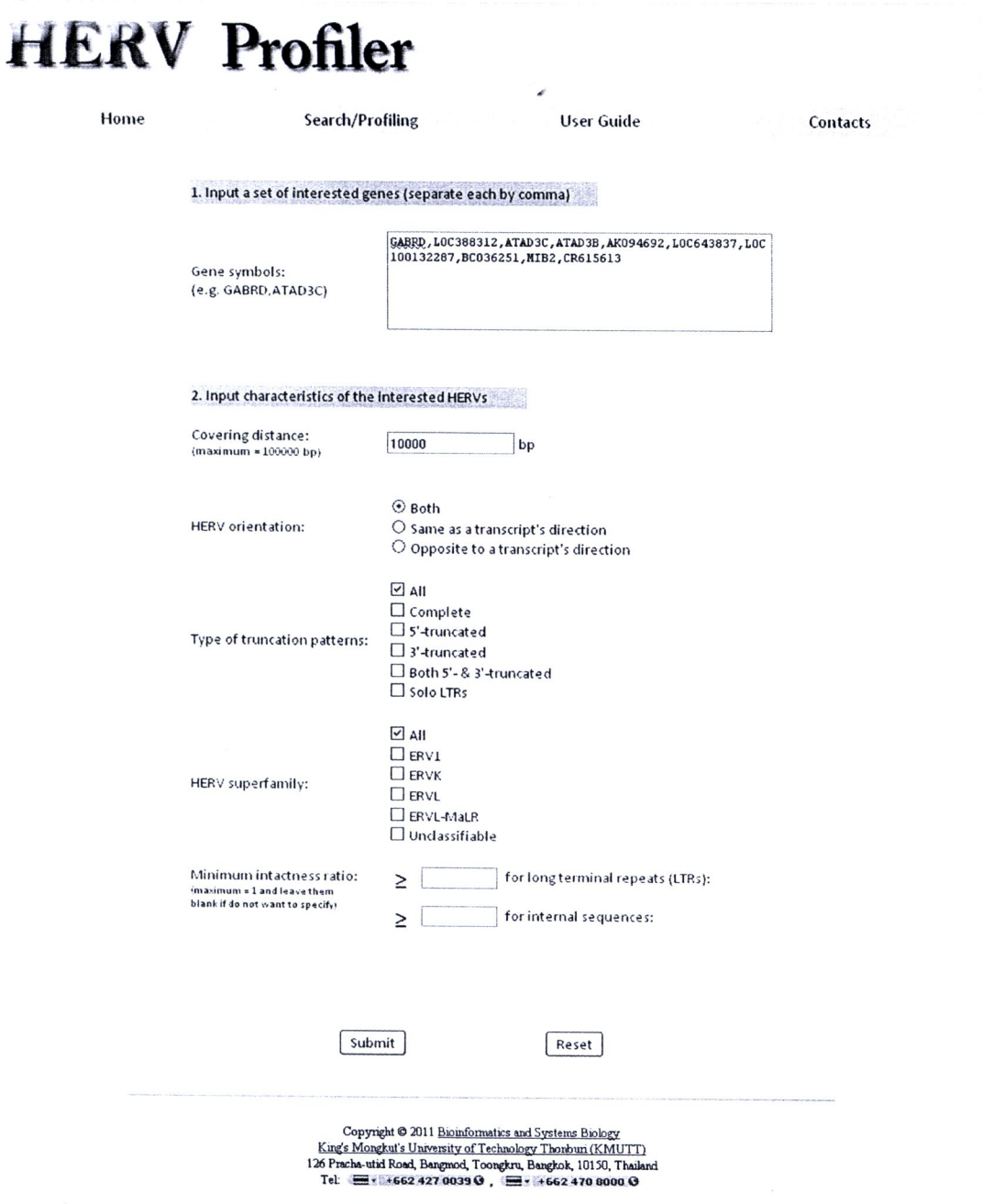


Figure 4.4 Serach/Profiling page

After specifying a set of interested genes and desired HERVs, the summary results of your genes input would be displayed. Also, the task button menu would be available now for the users to display neighboring HERVs in tabular format, to display the HERVs in graphical format, or to do the HERV profiling (Figure 4.5).

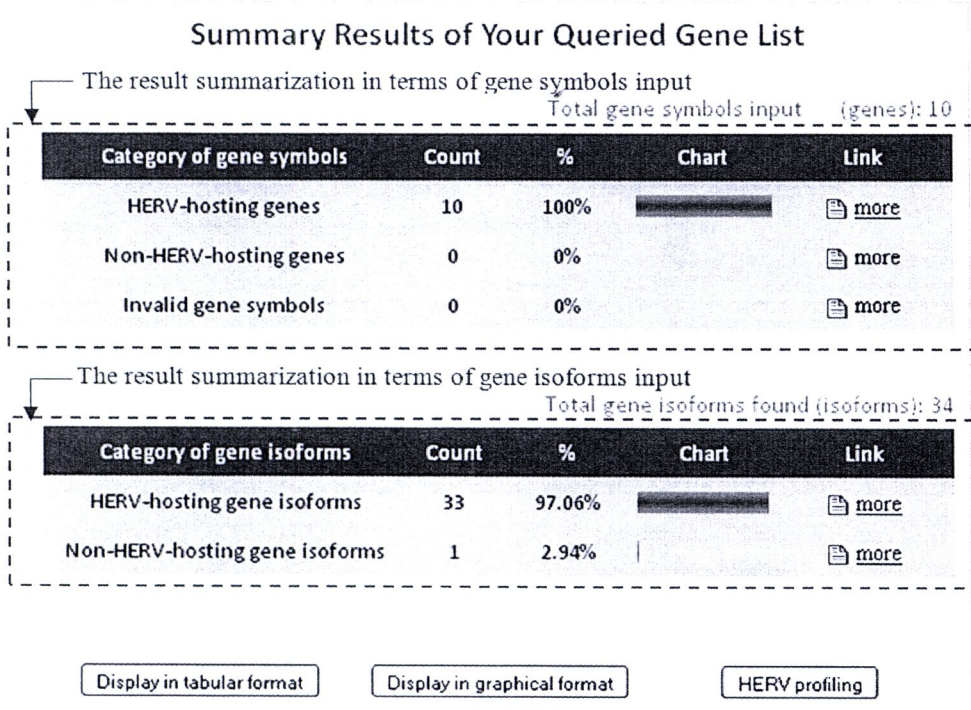


Figure 4.5 Summary results after input a gene list

The gene symbols input can be categorized into three classes, including HERV-hosting, non-HERV-hosting and invalid gene symbols. For each gene symbol, it possibly has more than one patterns of gene-HERV mapping found in the database, because a gene symbol may have more than one corresponding gene isoforms. According to Figure 4.5, the result summarization in terms of gene symbols is shown in the above table, while the summarization of the gene isoforms is shown in the below one.

In the gene symbol table, the class of HERV-hosting genes would contain gene symbols which at least one of their isoforms have neighboring HERVs. In terms of invalid gene symbols, they include both the gene symbols not found in the HERV Profiler database and the gene symbols which are probably typos. Users can find out a list of genes corresponding to a desired category by clicking on "more" links at the right-handed side of the tables.

4.2.1 Visualizing neighboring HERVs of genes in a gene list

There are two different views provided in HERV Profiler to display the neighboring HERVs, including a tabular view and a graphical view. User interface of these two would be illustrated in details further.

4.2.1.1 Tabular view

Users can easily go to the page of displaying neighboring HERVs by clicking on the menu button at the summary page (Figure 4.5). The information of the neighboring HERVs would then be shown in several tables as shown in Figure 4.6. The descriptions of each field in the tables are discussed in Appendix C, the user's manual guide. In

brief, a table here represents all neighboring HERVs along with their characteristics of one gene isoform, while the details of a gene isoform are written above the table.

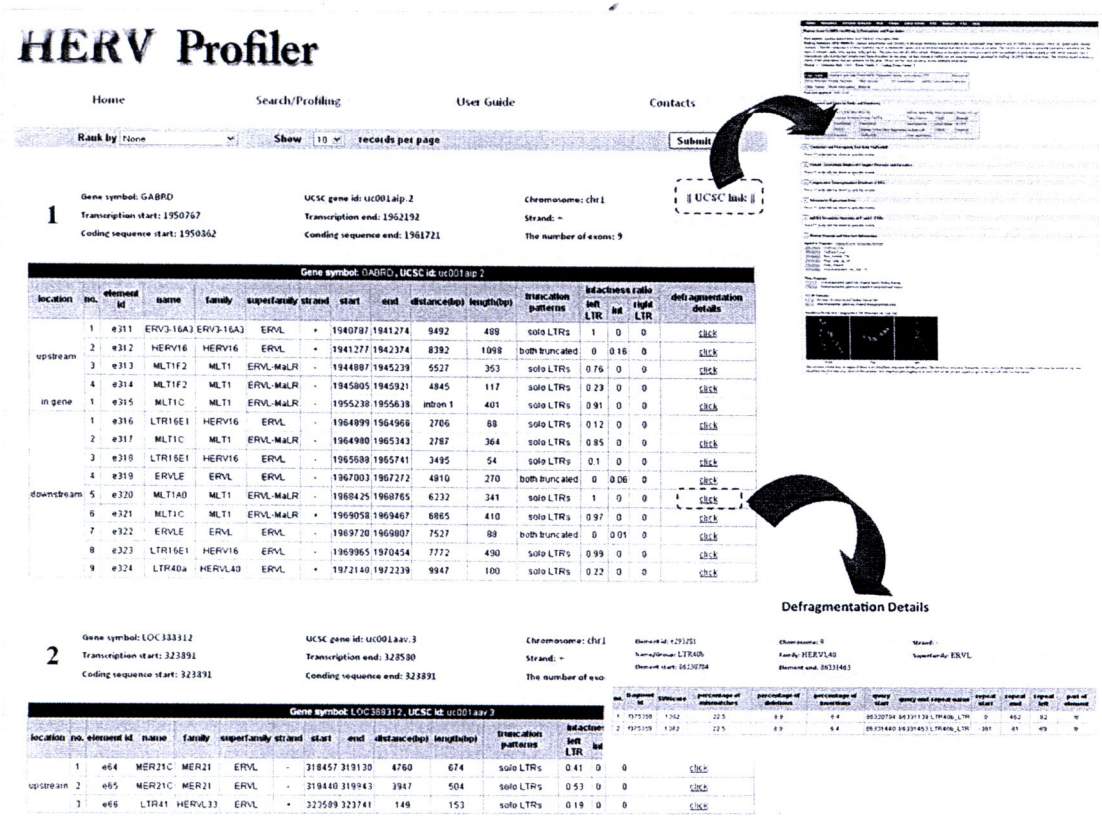


Figure 4.6 Displaying neighboring HERVs in a tabular format

Users can rank all of the tables by the number of neighboring HERVs at specific locations relative to genes or even throughout the genes. The options for ranking include ranking by the number of the HERVs throughout the genes, the number of upstream, in-gene, downstream, intron, and exon HERVs. In addition, users can look forward for more gene details at the UCSC database by selecting on the UCSC links, at the upmost right corners of the tables. The defragmentation details of each element are also provided when selecting on the links in the last columns of the tables.

4.2.1.2 Graphical view

Users can choose the menu button at the summary page to display the neighboring HERVs in a graphical view (Figure 4.5). The result interface of the graphical displays is shown in Figure 4.7.

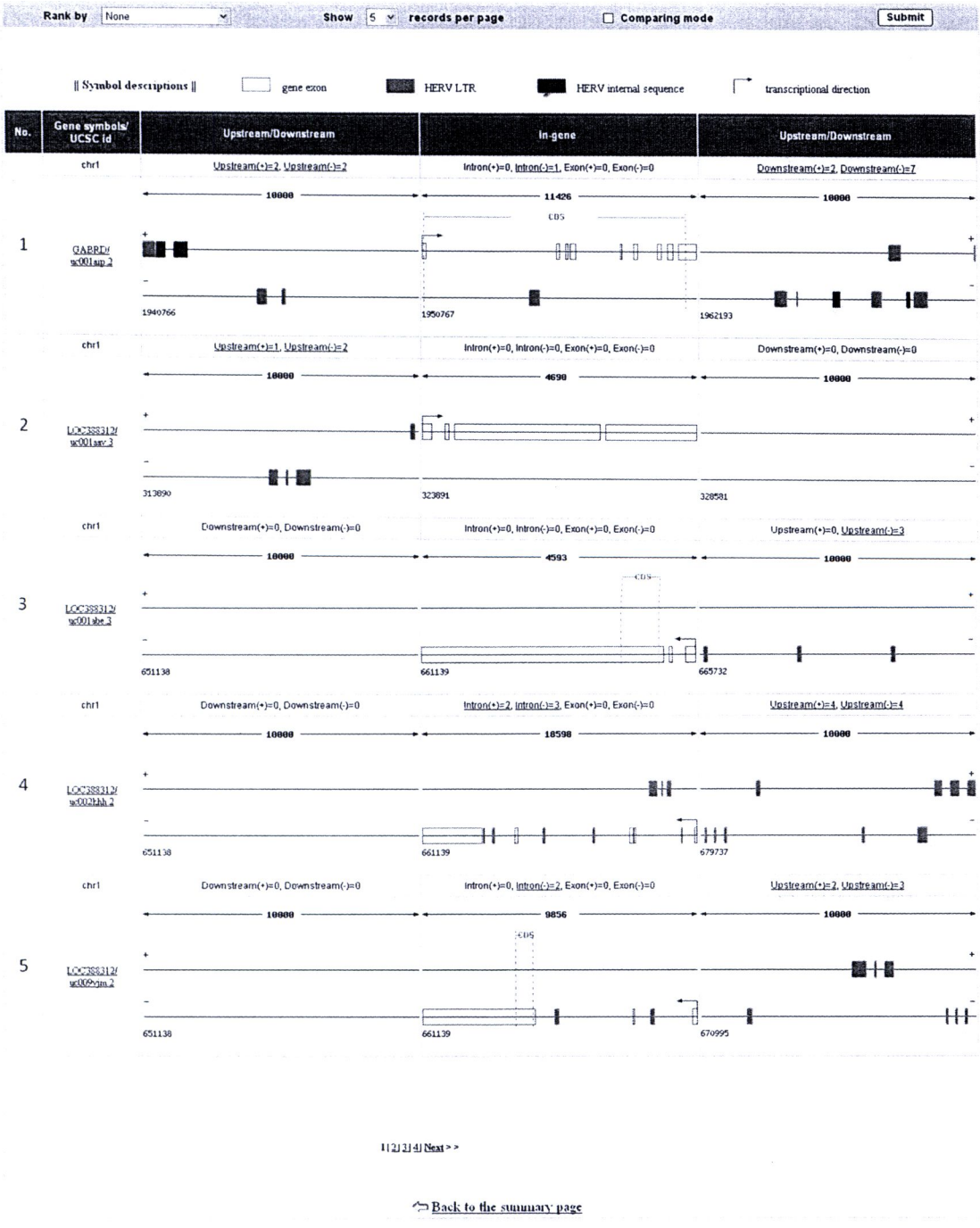


Figure 4.7 Displaying neighboring HERVs in a graphical format

According to Figure 4.7, each row, indicated with the increasing integers, represents a gene isoform. Like a tabular format, the users can also rank the genes in a gene list by the number of the HERVs throughout the genes or at specific locations relative to genes. The directions of each gene isoform drawn in the pictures are different, based on the normal transcription directions of each gene. To be easier in comparing between the pictures of each gene, the users can choose "comparing mode" to reflect the pictures of the gene isoforms laid on the reverse direction. Moreover, the users can go to the tables of a particular gene by selecting on its gene symbol and UCSC ID at the first column of

the tables (Figure 4.7). The detailed information of their HERVs would be then shown in the table as shown in Figure 4.8.

Gene and Its Neighboring HERV Details

Gene symbol: GABRD | UCSC gene id: uc001aip.2 | Chromosome: chr1 | || UCSC link ||
Transcription start: 1950767 | Transcription end: 1962192 | Strand: +
Coding sequence start: 1950862 | Conding sequence end: 1961721 | The number of exons: 9

Gene symbol: GABRD , **UCSC id:** uc001aip.2

location	no.	element id	name	family	superfamily	strand	start	end	distance(bp)	length(bp)	truncation patterns	intactness ratio			defragmentation details
												left LTR	int	right LTR	
upstream	1	e311	ERV3-16A3	ERV3-16A3	ERVL	+	1940787	1941274	9492	488	solo LTRs	1	0	0	click
	2	e312	HERV16	HERV16	ERVL	+	1941277	1942374	8392	1098	both truncated	0	0.16	0	click
	3	e313	MLT1F2	MLT1	ERVL-MaLR	-	1944887	1945239	5527	353	solo LTRs	0.76	0	0	click
	4	e314	MLT1F2	MLT1	ERVL-MaLR	-	1945805	1945921	4845	117	solo LTRs	0.23	0	0	click
in gene	1	e315	MLT1C	MLT1	ERVL-MaLR	-	1955238	1955638	intron 1	401	solo LTRs	0.91	0	0	click
downstream	1	e316	LTR16E1	HERV16	ERVL	-	1964899	1964966	2706	68	solo LTRs	0.12	0	0	click
	2	e317	MLT1C	MLT1	ERVL-MaLR	-	1964980	1965343	2787	364	solo LTRs	0.85	0	0	click
	3	e318	LTR16E1	HERV16	ERVL	-	1965688	1965741	3495	54	solo LTRs	0.1	0	0	click
	4	e319	ERVLE	ERVL	ERVL	-	1967003	1967272	4810	270	both truncated	0	0.06	0	click
	5	e320	MLT1A0	MLT1	ERVL-MaLR	-	1968425	1968765	6232	341	solo LTRs	1	0	0	click
	6	e321	MLT1C	MLT1	ERVL-MaLR	+	1969058	1969467	6865	410	solo LTRs	0.97	0	0	click
	7	e322	ERVLE	ERVL	ERVL	-	1969720	1969807	7527	88	both truncated	0	0.01	0	click
	8	e323	LTR16E1	HERV16	ERVL	-	1969965	1970454	7772	490	solo LTRs	0.99	0	0	click
	9	e324	LTR40a	HERVL40	ERVL	+	1972140	1972239	9947	100	solo LTRs	0.22	0	0	click

Figure 4.8 Detailed information of the neighboring HERVs of a particular gene linked from the graphical display

4.2.2 Profiling neighboring HERVs of genes in a gene list

HERV profiling is the process including both the construction and application of the HERV profiles. In HERV Profiler, there are three main steps to complete the HERV profiling: constructing HERV profiles, finding over-represented HERV types, and ranking genes in a gene list based on their neighboring HERVs. The web interfaces of the HERV profiling would be shown in order next.

4.2.2.1 Constructing HERV profiles

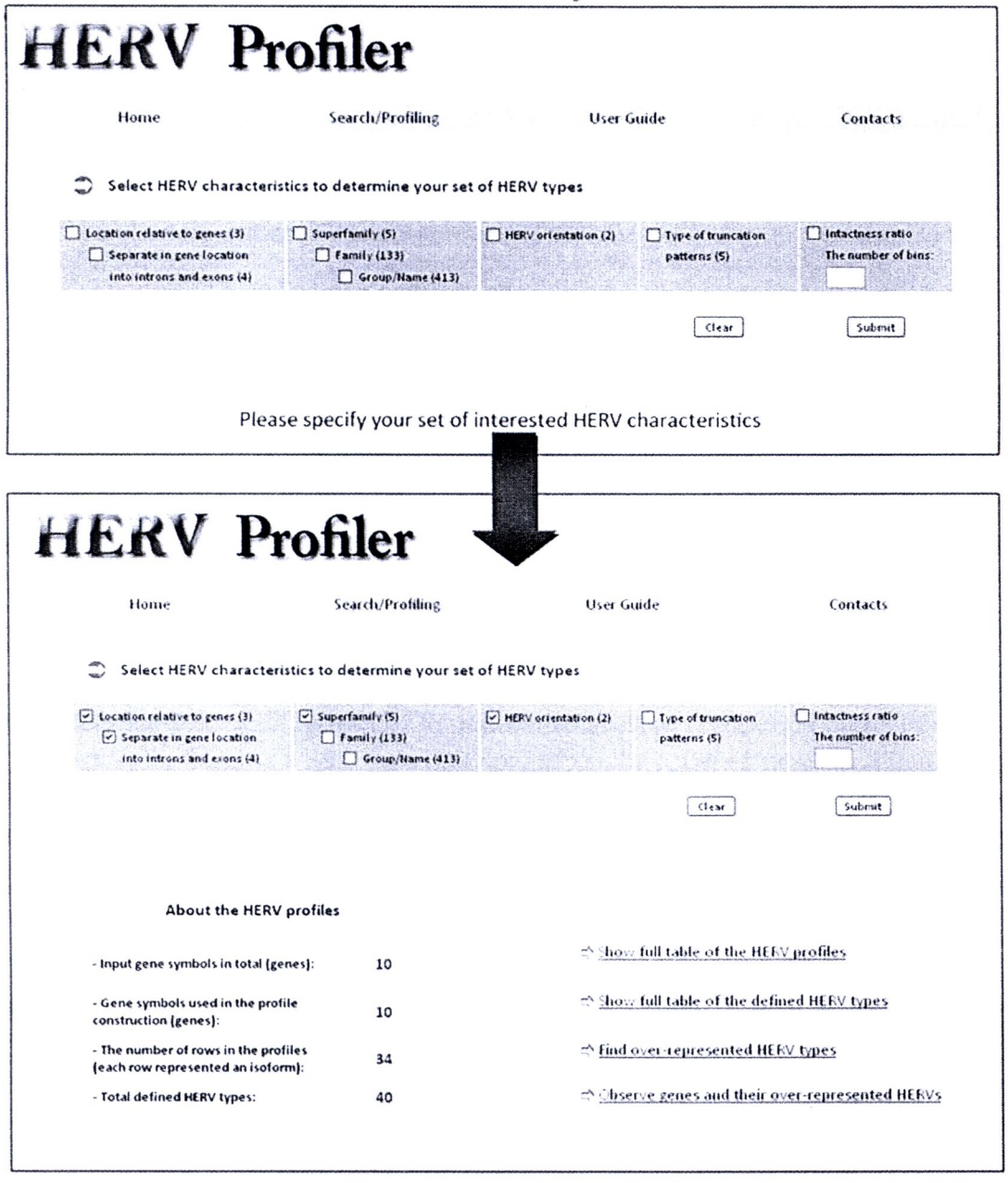
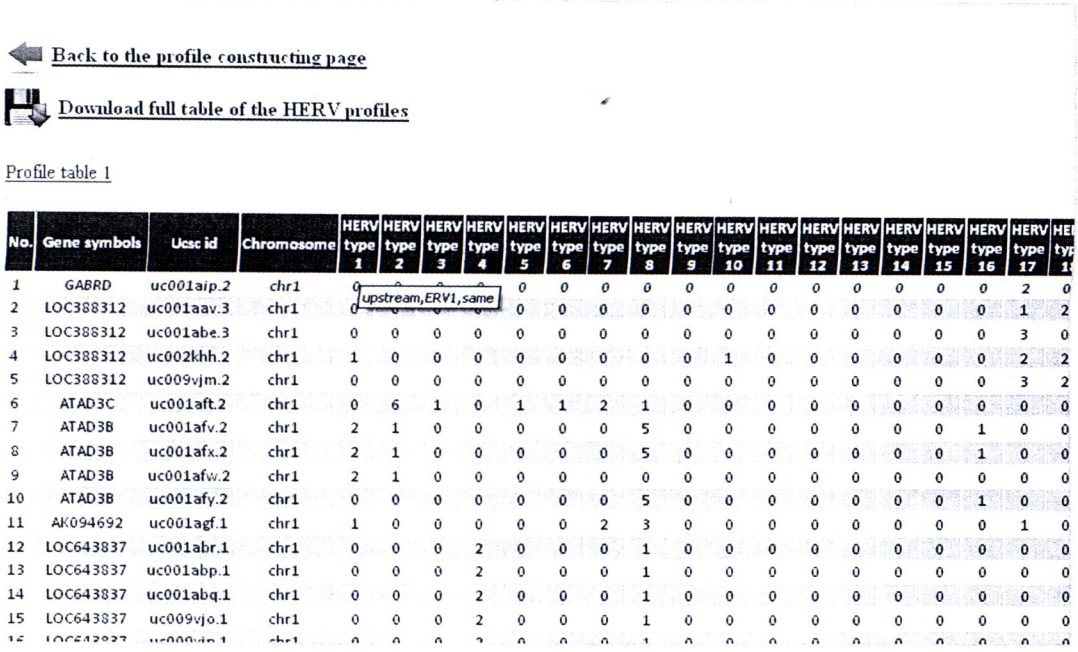


Figure 4.9 The pages of profile construction before and after performing the construction

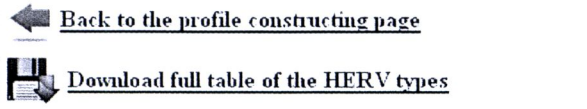
The web interface of the profile construction is shown in Figure 4.9. To construct the HERV profiles, the HERV characteristics should be selected first to define a set of HERV types. The summary information of the constructed profiles would be then shown after finishing the construction (Figure 4.9). The task menu would be available at the right-handed side of the summary information as well. The full table of the HERV profiles is shown in Figure 4.10.



No.	Gene symbols	Ucsc id	Chromosome	HERV type 1	HERV type 2	HERV type 3	HERV type 4	HERV type 5	HERV type 6	HERV type 7	HERV type 8	HERV type 9	HERV type 10	HERV type 11	HERV type 12	HERV type 13	HERV type 14	HERV type 15	HERV type 16	HERV type 17	HE ty 1
1	GABRD	uc001aip.2	chr1	0				0	0	0	0	0	0	0	0	0	0	0	0	2	0
2	LOC388312	uc001aav.3	chr1	0				0	0	0	0	0	0	0	0	0	0	0	0	1	2
3	LOC388312	uc001abe.3	chr1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
4	LOC388312	uc002khh.2	chr1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	2
5	LOC388312	uc009vjm.2	chr1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2
6	ATAD3C	uc001aft.2	chr1	0	6	1	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0
7	ATAD3B	uc001afv.2	chr1	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0
8	ATAD3B	uc001afx.2	chr1	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0
9	ATAD3B	uc001afw.2	chr1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	ATAD3B	uc001afy.2	chr1	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0
11	AK094692	uc001agf.1	chr1	1	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	1	0
12	LOC643837	uc001abr.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	LOC643837	uc001abp.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
14	LOC643837	uc001abq.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
15	LOC643837	uc009vjo.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
16	LOC643837	uc009vjn.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Figure 4.10 Full table of the HERV profiles

According to Figure 4.10, the detailed descriptions of the HERV types would be displayed when pointing a mouse over the HERV type labels. HERV Profiler also provides the users to download the constructed profiles as a tab-delimited text file. This could facilitate the users to be able to perform an analysis themselves outside the tool. In addition, the table of HERV type descriptions can be listed as shown in Figure 4.11. Also, HERV Profiler provides the users to download the descriptions as a tab-delimited text file. An example text file of HERV profiles and HERV type descriptions are both shown in Appendix C.

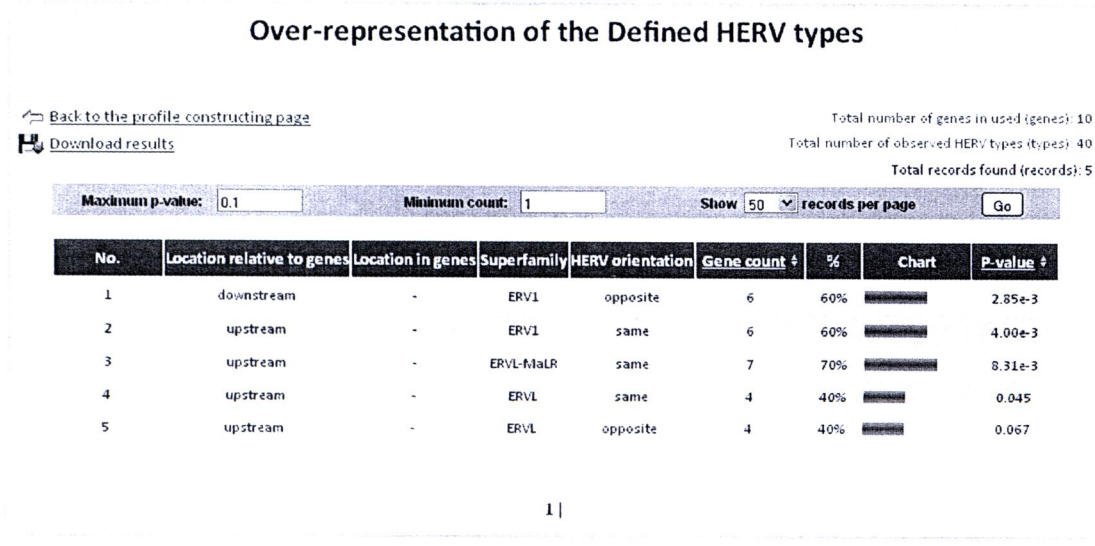


HERV type no.	Location relative to genes	Location in genes	Superfamily	HERV orientation
1	upstream	-	ERV1	same
2	upstream	-	ERV1	opposite
3	in gene	intron	ERV1	same
4	in gene	intron	ERV1	opposite
5	in gene	exon	ERV1	same
6	in gene	exon	ERV1	opposite
7	downstream	-	ERV1	same
8	downstream	-	ERV1	opposite
9	upstream	-	ERVK	same
10	upstream	-	ERVK	opposite

Figure 4.11 Full table of the HERV type descriptions

4.2.2.2 Finding over-represented HERV types among a gene list

The users can easily begin finding the HERV types that are over-represented among a gene list by selecting on the menu at the page of the profile construction (Figure 4.9). The tool would run for a while, and then the results would be shown as illustrated in Figure 4.12.



Over-representation of the Defined HERV types

Back to the profile constructing page
Download results

Total number of genes in used (genes): 10
Total number of observed HERV types (types): 40
Total records found (records): 5

Maximum p-value: 0.1 Minimum count: 1 Show 50 records per page Go

No.	Location relative to genes	Location in genes	Superfamily	HERV orientation	Gene count	%	Chart	P-value
1	downstream	-	ERV1	opposite	6	60%		2.85e-3
2	upstream	-	ERV1	same	6	60%		4.00e-3
3	upstream	-	ERVL-MaLR	same	7	70%		8.31e-3
4	upstream	-	ERVL	same	4	40%		0.045
5	upstream	-	ERVL	opposite	4	40%		0.067

1 |

Figure 4.12 Results of finding over-represented HERV types

There are two parameters used to limit HERV types shown in the result page. The first one is the maximum p-value of the HERV types. These p-values are obtained from performing the Fisher's exact tests. They indicate the significance of the testing which is concluded that an HERV type is over-represented. The lower a p-value of an HERV type, the more significant over-representation among a gene list of that HERV type. Thus, HERV types which would be shown in the result page are only the types having the p-values lower than the maximum value specified. In other words, the maximum value is the threshold used to indicate the acceptable significant level. Another parameter is the minimum count which is the lowest acceptable gene count of the HERV types among a gene list. In other words, only the HERV types which have the number of related genes more than or equal to the specified count would be shown in the result page. The default values of the maximum p-value and minimum count are 0.1 and 1, respectively. All of the accepted types can also be downloaded as a text file, as exemplified in Appendix C.

4.2.2.3 Ranking genes in a gene list based on their HERVs

Users can rank genes in a gene list base on their possessed HERV types by selecting on the menu at the page of profile construction (Figure 4.9). The result page of the ranking is shown in Figure 4.13.

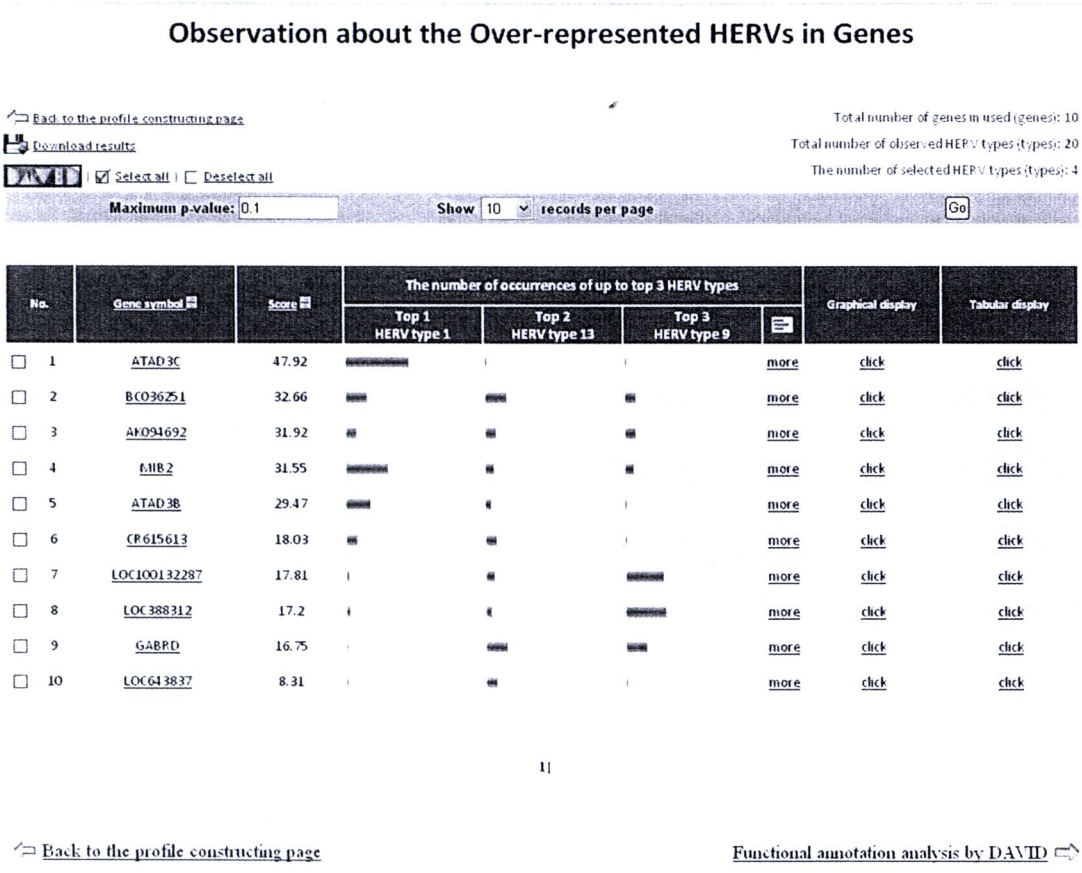


Figure 4.13 Results of ranking genes based on their possessed HERVs

The maximum p-value is required to be specified for determining the HERV types that would be included in the score calculation. Only the accepted types would be included to assign scores to every gene in a gene list. In this page, shortcuts of the profile charts of each gene are also displayed up to top three HERV types. The users can also look forward for the full profile charts of a particular gene by selecting on "more". An example of the full profile charts is shown in Figure 4.14. Also, the users can reach the gene details at the UCSC database by selecting on the gene symbols. Additionally, there are the links of DAVID or a tool for GO enrichment analysis provided in this page. The users can easily go further analyzing of DAVID by just selecting a set of desired genes, and click on the DAVID links appeared. The graphical and tabular displays of a particular gene can also be achieved from this result page, as shown in Figures 4.15-4.16. The users can retrieve all of the results, including the results of gene ranking and the profile charts, as text files also (Appendix C)

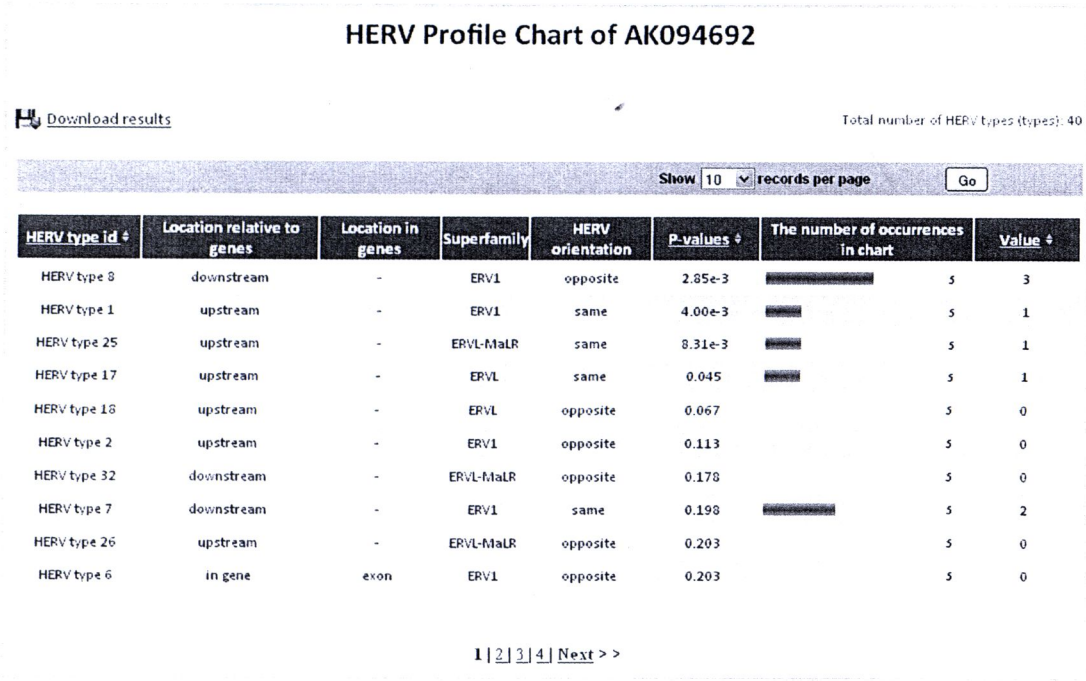


Figure 4.14 Profile charts of a particular gene

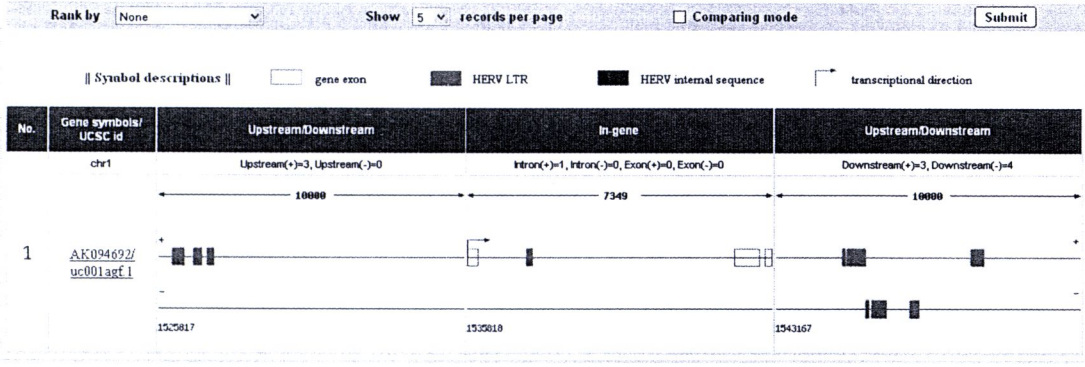


Figure 4.15 Graphical display linked from the page of gene ranking

1 Gene symbol: AK094692 | UCSC gene id: uc001agf.1 | Chromosome: chr1 | || UCSC link ||
Transcription start: 1535818 | Transcription end: 1543166 | Strand: +
Coding sequence start: 1535818 | Conding sequence end: 1535818 | The number of exons: 3

Gene symbol: AK094692, UCSC id: uc001agf.1

location	no.	element id	name	family	superfamily	strand	start	end	distance(bp)	length(bp)	truncation patterns	Intactness ratio left LTR	int	right LTR	defragmentation details
upstream	1	e271	MER83	MER83	ERV1	+	1526269	1526997	8820	729	solo LTRs	1	0	0	click
	2	e273	MER21C	MER21	ERVL	+	1527160	1527478	8339	319	solo LTRs	0.18	0	0	click
	3	e272	MLT1C	MLT1	ERVL-MaLR	+	1527022	1527607	8210	586	solo LTRs	0.52	0	0	click
in gene	1	e274	MLT1K	MLT1	ERVL-MaLR	+	1537250	1537389	intron 1	140	solo LTRs	0.24	0	0	click
downstream	1	e275	MER65D	MER65	ERV1	+	1545346	1546100	2179	755	solo LTRs	0.68	0	0	click
	2	e276	THE1C	THE1	ERVL-MaLR	+	1545562	1545913	2395	352	solo LTRs	1	0	0	click
	3	e277	LTR26B	LOR1	ERV1	-	1546101	1546209	2934	109	solo LTRs	0.2	0	0	click
	4	e278	LTR42	HERVL42	ERVL	-	1546310	1546787	3143	478	solo LTRs	1	0	0	click
	5	e279	MER4D	MER4	ERV1	-	1547553	1547794	4386	242	solo LTRs	0.26	0	0	click
	6	e280	MER49	MER49	ERV1	-	1547798	1547873	4631	76	solo LTRs	0.08	0	0	click
	7	e281	MER65D	MER65	ERV1	+	1549546	1550009	6379	464	solo LTRs	1	0	0	click

Figure 4.16 Tabular display linked from the page of gene ranking

4.3 Tool testing

4.3.1 Testing of sub-module 2.2

To check if this sub-module works accurately, the gene sets with selected over-represented HERV types were generated and applied to the tool. There are four simulated gene sets used in this step. The gene set 1, 2, and 3 have 100 genes in each and the selected HERV type as upstream-ERV1-same, in gene-intron-ERVL-opposite, and downstream-ERVK-opposite, respectively. For the gene set 4, it is intentionally generated to have three over-represented types, including the selected types of the gene set 1, 2, and 3. This results in 54 genes in the gene set 4. The over-represented types reported from the tool by using all four gene sets are shown in Figures 4.17-4.20, respectively.

Location relative to genes	Location in genes	Superfamily	HERV orientation	Gene count ↕	%	Chart	P-value ↕
upstream	-	ERV1	same	100	100%		3.64e-74
upstream	-	ERV1	opposite	40	40%		2.04e-6
in gene	intron	ERV1	same	30	30%		7.07e-4
in gene	intron	ERV1	same	30	30%		1.57e-3

Figure 4.17 Reported over-represented HERV types using the gene set 1. The selected type of this gene set is upstream-ERV1-same.

Location relative to genes	Location in genes	Superfamily	HERV orientation	Gene count	%	Chart	P-value
in gene	intron	ERVL	opposite	100	100%		7.99e-57
in gene	intron	ERVL-MaLR	opposite	85	85%		2.30e-19
in gene	intron	ERV1	opposite	64	64%		6.00e-15
in gene	intron	ERVL-MaLR	same	57	57%		8.73e-11

Figure 4.18 Reported over-represented HERV types by using the gene set 2. The selected type of this gene set is in gene-intron-ERVL-same.

Location relative to genes	Location in genes	Superfamily	HERV orientation	Gene count	%	Chart	P-value
downstream	-	ERVK	opposite	100	100%		3.83e-159
in gene	intron	ERVK	opposite	22	22%		5.28e-9
in gene	exon	ERVK	opposite	7	7%		1.15e-8
downstream	-	ERV1	opposite	39	39%		1.99e-7

Figure 4.19 Reported over-represented HERV types by using the gene set 3. The selected type of this gene set is downstream-ERVK-opposite.

Location relative to genes	Location in genes	Superfamily	HERV orientation	Gene count	%	Chart	P-value
downstream	-	ERVK	opposite	54	100%		5.82e-87
upstream	-	ERV1	same	54	100%		1.83e-40
in gene	intron	ERVL	opposite	54	100%		4.55e-31
in gene	intron	ERVK	opposite	19	35.19%		8.54e-12
in gene	intron	ERV1	opposite	35	64.81%		4.74e-9
in gene	intron	ERV1	same	35	64.81%		1.05e-7

Figure 4.20 Reported over-represented HERV types by using the gene set 4. The selected types of this gene set are upstream-ERV1-same, in gene-intron-ERVL-opposite, and downstream-ERVK-opposite mutually.

According to Figure 4.17-4.20, it can be noticed that all of the selected types, independently with the number of the selected types, were all reported as the over-represented types for all gene sets. This could indicate that this sub-module can function accurately in detecting over-represented HERV types among a gene list.

4.3.2 Testing of sub-module 2.3

To check whether this sub-module works accurately, a gene set composed of ten for the genes related to the selected types and another ten for the genes not related to those types were generated and applied to the tool. The expected result is that the genes related to the over-represented types should have higher scores than the genes not related to those types. There are four gene sets, with different over-represented types pre-determined. The gene set 1, 2, and 3 are pre-determined as the upsteam-ERVK-opposite, in gene-exon-ERVL-same, and downstream-ERVK-same, respectively, to be the over-represented types. For the gene set 4, there are three pre-determined over-represented types, including upstream-MaLR-opposite, in gene-intron-ERV1-opposite,

and downstream-MaLR-same, mutually. The maximum p-value threshold was always set as 0.05 in this testing. The results performed by the sub-module 2.3 are summarized in Table 4.2.

Table 4.2 Results of gene ranking in all four gene sets. Highlighted rows indicate the genes which are related to the pre-defined over-represented types among each gene set.

Rank	Gene set 1		Gene set 2		Gene set 3		Gene set 4	
	Gene symbol	Score	Gene symbol	Score	Gene symbol	Score	Gene symbol	Score
1	C17orf51	44.38	LONRF2	127.82	DQ573033	116.181	PRUNE2	47.77
2	GLYATL2	41.71	C14orf86	48.74	BLVRA	91.94	ZNF22	36.15
3	KIAA1982	37.39	BC032899	40.76	ZNF879	52.82	RRM2B	31.91
4	ABCA11P	32.73	AK090681	37.61	SYK	52.28	SP100	26.78
5	HIGD1A	30.40	ZBTB42	37.61	UCMA	36.99	RAPGEF4	26.16
6	ZNF677	28.52	BX247990	32.79	LOC196394	35.89	RELL1	26.05
7	DO585694	24.75	BC036195	27.99	CRYGB	33.77	PRMT3	25.31
8	FHR4	24.75	SPINLW1	18.67	DQ570592	33.77	BC037215	23.25
9	SERHL2	24.75	NCRNA00173	12.80	AGAP8	26.13	OTOP1	21.46
10	D2HGDH	22.16	MEIS3	12.65	ZNF586	19.54	RPH3AL	17.86
11	UBQLN1	11.31	HBII-52-45	7.97	AX747775	7.91	BTF3L1	8.95
12	BC030116	6.99	LMBRIL	5.60	DM004401	4.76	KIAA1351	5.13
13	BCL2L11	1.88	TAS2R10	4.82	TORC2	0	USP50	4.66
14	VAMP2	0	SFRS13A	4.82	VPS16	0	B3GNT9	4.48
15	NUBP2	0	KIAA0960	3.15	SCG2	0	LYG2	3.35
16	ARL6IP1	0	FAM194B	2.41	RPP38	0	DQ585588	3.35
17	KIAA0421	0	RTEL1	0	KIAA0684	0	OR10S1	0
18	C21orf94	0	GPR173	0	BC048278	0	KIAA1772	0
19	GUCY1B2	0	DQ588584	0	AK127578	0	syndecan4	0
20	CXXC5	0	DIS3	0	AF086294	0	VPS36	0

For all gene sets, the results are obviously noticeable that the rank 1 to 10 are all genes that are related to the pre-determined over-represented types. Therefore, it could be concluded that this sub-module can be used to rank the genes in a gene list based on their relation to the over-represented types.

4.4 Case studies related to SLE

4.4.1 Case study 1: Significance of the number of HERVs under the SLE condition

This case study is aimed at observing the number of neighboring HERVs of the SLE-relevant genes, significantly differentially expressed genes under the SLE condition. To guarantee the relationships found in the SLE-relevant genes, a set of insignificant genes with the same size as the SLE-relevant genes was used to compare the number of neighboring HERVs. In this case study, the SLE-relevant genes were selected from the top 500 genes following the p-values obtained from the t-test, while a set of insignificant genes were selected from the bottom 500 genes.

The tool was used to construct the HERV profiles of both the top and bottom 500 genes to obtain the number of neighboring HERVs throughout the genes as well as at each location relative to genes. The specific locations considered in this case study are composed of upstream, intron, exon, and downstream locations. The profiles generated from the tool were then retrieved to be analyzed outside the tool. To guarantee the relationships found in a group of SLE-relevant genes independent to the number of the genes considered, the comparison between a group of significant and insignificant genes were performed on different numbers of the genes in the groups. The numbers of the top and bottom genes used are 100, 200, 300, 400, and 500. The results of the comparison by using z-test are shown in Table 4.3.

Table 4.3 P-values of the z-tests performing on the number of HERVs at each location between a group of significant and insignificant genes

The number of genes used (genes)	P-value of each location relative to genes				
	Throughout the genes	Upstream	Intron	Exon	Downstream
100	0.065	0.47	0.0097	0.53	0.00345
200	0.360	0.94	0.20	0.07	0.0149
300	0.301	0.80	0.17	0.028	0.0188
400	0.207	0.72	0.10	0.068	0.0037
500	0.138	0.93	0.054	0.10	0.0040

According to Table 4.3, the number of neighboring HERVs throughout the genes is not significantly different between the groups of significant and insignificant genes. When observing on the specific locations relative genes instead, there is no explicitly significantly difference in the number of neighboring HERVs at each location, except at the downstream location. At the 95% confidence level, the numbers of the downstream HERVs are always significantly different between the groups of the genes which are significantly and insignificantly differentially expressed under the SLE condition. The box plots comparing the number of HERVs at the downstream location between the groups of significant and insignificant genes were shown in Figure 4.21.

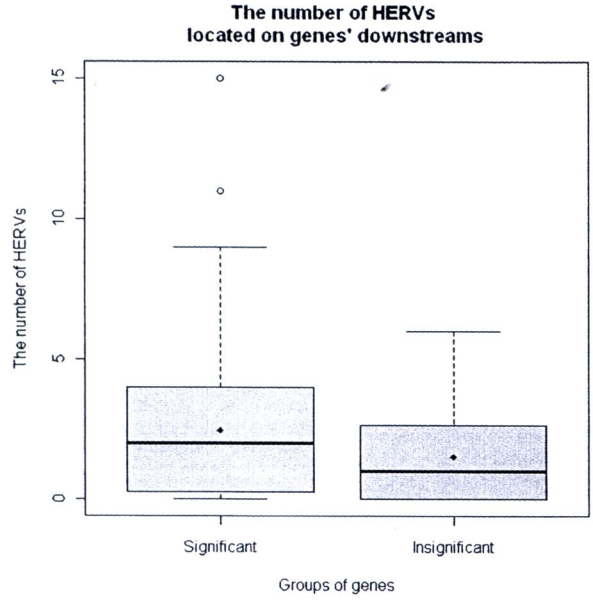


Figure 4.21 Box plots of the number of HERVs located at the downstream location comparing between top 100 significant genes and bottom 100 insignificant genes. Red points indicate average values among the groups.

The box plots indicate the distributions of the number of HERVs at the downstream location in both a group of the top and bottom genes. The numbers of the downstream HERVs of the top 100 genes somewhat tend to be higher than the bottom 100 genes. This also results that the number of the downstream HERVs in average of the top 100 are higher than the average of the bottom 100 genes. This is equivalent to the results of the z-tests (Table 4.3).

HERVs located on the downstream relative to genes could provide polyadenylation signals to abnormally terminate the transcriptions of the neighboring genes. For example, the leptin obesity hormone receptor (LEPR) exists in two variants that differ in size due to the availability of a HERV-K polyadenylation signals [5]. Therefore, it is very interesting that there is more density of the HERVs present at the downstream locations in the group of SLE-relevant genes. Furthermore, this could be suggested that the downstream HERVs may be related to the SLE-relevant genes by providing the interfering regulatory to the neighboring genes. However, this issue should be subject to further study, especially on the wet lab experiments.

4.4.2 Case study 2: Significance of HERV characteristics under the SLE condition

This case study is aimed at observing the characteristics of the HERVs nearby the SLE-relevant genes, the significantly differentially expressed genes. From performing t-tests and calculating the fold-changes, there are 230 genes being significantly differentially expressed under the SLE condition. These genes were applied to the HERV Profiler to investigate the neighboring HERVs. The HERV profiling was performed to

preliminarily discover the over-represented HERV types among the gene list. The reported types, HERVs with particular characteristics, could be suggested as a factor possibly disturbing the expression of SLE-relevant genes, because of their suspicious over-representations among the group of SLE-relevant genes. The preliminarily summary result obtained after inputting the gene list is shown in Figure 4.22.

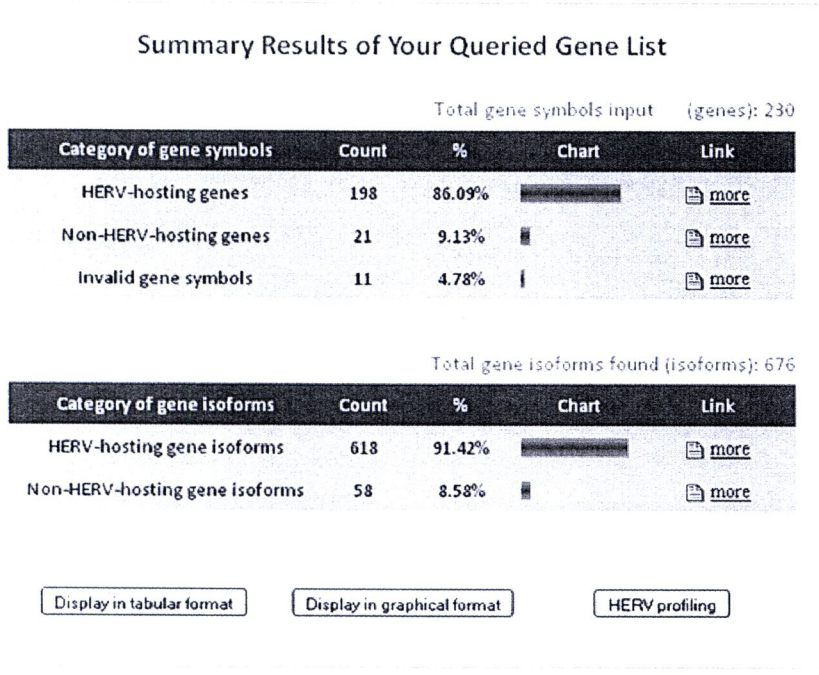


Figure 4.22 Summary results of the 230 significant genes under the SLE condition

From those 230 genes, there are 11 genes not found in HERV Profiler, and approximately 9% of the input genes that do not possess any neighboring HERVs. In other words, more than 85% of the genes input are classified as the HERV-hosting genes (Figure 4.22). To preliminary exploration the neighboring HERVs in all of the 230 genes, the visualizing module (HERV Profiler module 1) was performed and the results are shown below.

Table 4.4 Top ten genes ranked by the number of all neighboring HERVs throughout the genes

Rank	Gene symbol/UCSC ID	Total number of all HERVs (elements)	In-gene HERVs (elements)	Intron HERVs (elements)	Exon HERVs (elements)
1	PDE4D/uc003jsb.2	190	186	186	0
2	PDE4D/uc003jsa.2	99	95	95	0
3	STK3/uc003yio.2	91	89	89	0
4	PDE4D/uc003jry.2	89	85	85	0
5	PDE4D/uc003jrz.2	89	85	85	0
6	HECW1/uc003tid.1	86	81	81	0
7	HECW1/uc011kbi.1	86	81	81	0
8	PDE4D/uc010iwj.1	78	77	76	1
9	ATP8A2/uc001uqk.2	76	73	73	0
10	SEMA5A/uc003jek.2	76	74	74	0

Table 4.5 Top ten genes ranked by the number of in-gene HERVs

Rank	Gene symbol/UCSC ID	In-gene HERVs (elements)	Sense in-gene HERVs (elements)	Antisense in-gene HERVs (elements)
1	PDE4D/uc003jsb.2	186	46	140
2	PDE4D/uc003jsa.2	95	17	78
3	STK3/uc003yio.2	89	27	62
4	PDE4D/uc003jry.2	85	17	68
5	PDE4D/uc003jrz.2	85	17	68
6	HECW1/uc011kbi.1	81	17	64
7	HECW1/uc003tid.1	81	17	64
8	PDE4D/uc010iwj.1	76	26	50
9	ATP8A2/uc001uqk.2	74	25	49
10	SEMA5A/uc003jek.2	73	18	55

Table 4.6 Top ten genes ranked by the number of intron HERVs

Rank	Gene symbol/UCSC ID	Intron HERVs	Sense intron HERVs	Antisense intron HERVs
1	PDE4D/uc003jsb.2	186	46	140
2	PDE4D/uc003jsa.2	95	17	78
3	STK3/uc003yio.2	89	27	62
4	PDE4D/uc003jry.2	85	17	68
5	PDE4D/uc003jrz.2	85	17	68
6	HECW1/uc003tid.1	81	17	64
7	HECW1/uc011kbi.1	81	17	64
8	PDE4D/uc010iwj.1	76	26	50
9	ATP8A2/uc001uqk.2	74	25	49
10	SEMA5A/uc003jek.2	73	18	55

According to Tables 4.4-4.6, the neighboring HERVs of the top ten genes seem to be present in introns rather than other locations. Furthermore, the presence of the neighboring HERVs is usually located in the antisense direction rather than the sense direction among those top ten genes. Regarding to the issues, the characteristics of HERV location, with separating in-gene locations into introns and exons, as well as HERV orientation should be included in the determination of the HERV types to initially construct the HERV profiles.

There are 1,064 HERV types in total defined based on HERV locations, with separating in-gene location into introns and exons, superfamily, family, and orientation. 219 genes from the gene list were included to construct the profiles, because the remaining eleven genes could not be found in the HERV Profiler database.

In the finding over-represented HERV types, the maximum p-value used is 0.01, for more conservative selection. The results show that there are eight types over-represented among this gene list as shown in Figure 4.23.

Location relative to genes	Location in genes	Superfamily	Family	HERV orientation	Gene count ↕	%	Chart	P-value ↕
upstream	-	ERV1	HERVH	same	7	3.2%		7.27e-4
downstream	-	ERV1	MER34	same	7	3.2%		1.31e-3
upstream	-	ERV1	HERV30	opposite	2	0.91%		1.93e-3
downstream	-	ERV1	MER61	same	4	1.83%		2.31e-3
downstream	-	ERV1	MER57	same	6	2.74%		5.56e-3
in gene	exon	ERV1	HERV23	same	2	0.91%		6.21e-3
upstream	-	ERV1	HERVIP10	opposite	3	1.37%		8.50e-3
upstream	-	ERV1	HERV23	same	5	2.28%		9.14e-3

Figure 4.23 Over-represented HERV types among 219 genes which are significantly differentially expressed under the SLE condition

When ranking genes in the gene list, the results show that there are 30 genes that are related to the over-represented HERV types, the genes having scores higher than zero. The list of top ten genes related to the over-represented types is shown in Figure 4.24.

No.	Gene symbol	Score	The number of occurrences of up to top 3 HERV types			
			Top 1 HERV type 169	Top 2 HERV type 487	Top 3 HERV type 50	
1	IL8RA	21.68				more
2	DET1	14.45				more
3	STK17B	12.14				more
4	TRIM38	11.59				more
5	FCGR1A	10.95				more
6	FCGR2B	10.14				more
7	HMGCS1	9.64				more
8	IL18RAP	9.39				more
9	ATP6V1G1	9.39				more
10	TLR2	7.8				more

Figure 4.24 Top ten genes related to over-represented HERV types

When observing the random gene lists, the results show that there is no overlapping over-represented HERV type at all. Therefore, it can be certain that the over-represented types reported in Figure 4.23 would not be occurred randomly. From this case study, it could be suggested that those reported HERV types (Figure 4.23) may influence their neighboring genes under the SLE condition, because they are significantly present among the SLE-relevant genes when comparing to the presence in the whole genome. Also, the genes that are highly related to those over-represented HERV types should be possibly to be suspected genes which are affected by the neighboring HERVs. However, both reported over-represented types and genes related to those over-represented types should be paid more in further study, especially in wet lab experiments, to obtain the exact answers of these influencing.