REFERENCES

1. Griffiths, D., 2001, "Endogenous Retroviruses in the Human Genome Sequence", **Genome Biology**, Vol. 2, No. 6, pp. reviews1017.1011 - reviews1017.1015.

2. Blikstad, V., Benachenhou, F., Sperber, G., and Blomberg, J., 2008, "Endogenous Retroviruses", **Cellular and Molecular Life Sciences**, Vol. 65, No. 21, pp. 3348-3365.

3. Nelson, P.N., Hooley, P., Roden, D., Davari Ejtehadi, H., Rylance, P., Warren, P., Martin, J., Murray, P.G., and Molecular Immunology Research, G., 2004, "Human Endogenous Retroviruses: Transposable Elements with Potential?", **Clinical & Experimental Immunology**, Vol. 138, No. 1, pp. 1-9.

4. Cohen, C.J., Lock, W.M., and Mager, D.L., 2009, "Endogenous Retroviral Ltrs as Promoters for Human Genes: A Critical Assessment", **Gene**, Vol. 448, No. 2, pp. 105-114.

5. Jern, P. and Coffin, J.M., 2008, "Effects of Retroviruses on Host Genome Function", **Annual Review of Genetics**, Vol. 42, No. 1, pp. 709-732.

6. Medstrand, P., Landry, J.-R., and Mager, D.L., 2001, "Long Terminal Repeats Are Used as Alternative Promoters for the Endothelin B Receptor and Apolipoprotein C-I Genes in Humans", **Journal of Biological Chemistry**, Vol. 276, No. 3, pp. 1896-1903.

7. Landry, J.-R., Rouhi, A., Medstrand, P., and Mager, D.L., 2002, "The Opitz Syndrome Gene Mid1 Is Transcribed from a Human Endogenous Retroviral Promoter", **Molecular Biology and Evolution**, Vol. 19, No. 11, pp. 1934-1942.

8. Ruda, V.M., Akopov, S.B., Trubetskoy, D.O., Manuylov, N.L., Vetchinova, A.S., Zavalova, L.L., Nikolaev, L.G., and Sverdlov, E.D., 2004, "Tissue Specificity of Enhancer and Promoter Activities of a Herv-K(Hml-2) Ltr", **Virus Research**, Vol. 104, No. 1, pp. 11-16.

9. Paces, J., Pavlicek, A., and Paces, V., 2002, "Hervd: Database of Human Endogenous Retroviruses", **Nucleic Acids Research**, Vol. 30, No. 1, pp. 205-206.

10. Villesen, P., Aagaard, L., Wiuf, C., and Pedersen, F., 2004, "Identification of Endogenous Retroviral Reading Frames in the Human Genome", **Retrovirology**, Vol. 1, No. 1, p. 32.

11. Levy, A., Sela, N., and Ast, G., 2008, "Transpogene and Microtranspogene: Transposed Elements Influence on the Transcriptome of Seven Vertebrates and Invertebrates", **Nucleic Acids Research**, Vol. 36, No. suppl 1, pp. D47-D52.

12. Gifford, R. and Tristem, M., 2003, "The Evolution, Distribution and Diversity of Endogenous Retroviruses", **Virus Genes**, Vol. 26, No. 3, pp. 291-315.

13. Nelson, P.N., Carnegie, P.R., Martin, J., Ejtehadi, H.D., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J., and Murray, P.G., 2003, "Dymystified...Human Endogenous Retroviruses", **Journal of Clinical Pathology**, Vol. 56, No. 1, pp. 11-18.

14. Kazazian, H.H., Jr., 2004, "Mobile Elements: Drivers of Genome Evolution", **Science**, Vol. 303, No. 5664, pp. 1626-1632.

15. Donovan, T., 2010, "Endogenous Retroviruses and the Human Genome: Implications for Human Disease", **SPCV**, Vol. 2, No. 1, pp. 1-33.

16. Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G., and Mayer, J., 2009, "Classification and Nomenclature of Endogenous Retroviral Sequences (Ervs): Problems and Recommendations", **Gene**, Vol. 448, No. 2, pp. 115-123.

17. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J., 2005, "Repbase Update, a Database of Eukaryotic Repetitive Elements", **Cytogenetic and Genome Research**, Vol. 110, No. 1-4, pp. 462-467.

18. Mayer, J. and Meese, E., 2005, "Human Endogenous Retroviruses in the Primate Lineage and Their Influence on Host Genomes", **Cytogenetic and Genome Research**, Vol. 110, No. 1-4, pp. 448-456.

19. Smit, A.F.A., 1993, "Identification of a New, Abundant Superfamily of Mammalian Ltr-Transposons", **Nucleic Acids Research**, Vol. 21, No. 8, pp. 1863-1872.

20. Oja, M., Peltonen, J., Blomberg, J., and Kaski, S., 2007, "Methods for Estimating Human Endogenous Retrovirus Activities from Est Databases", **BMC Bioinformatics**, Vol. 8, No. Suppl 2, p. S11.

21. Blomberg, J., Ushameckis, D., and Jern, P., 2005, **Evolutionary Aspects of Human Endogenous Retroviral Sequences (Hervs) and Disease**, Eurekah.com, pp. 204-238.

22. Kurth, R. and Bannert, N., 2010, "Beneficial and Detrimental Effects of Human Endogenous Retroviruses", **International Journal of Cancer**, Vol. 126, No. 2, pp. 306-314.

23. Buzdin, A., 2007, "Human-Specific Endogenous Retroviruses", **TheScientificWorldJOURNAL**, Vol. 7, No., pp. 1848-1868.

24. Long, Q., Bengra, C., Li, C., Kutlar, F., and Tuan, D., 1998, "A Long Terminal Repeat of the Human Endogenous Retrovirus Erv-9 Is Located in the 5' Boundary Area of the Human [Beta]-Globin Locus Control Region", **Genomics**, Vol. 54, No. 3, pp. 542-555.

25. Dunn, C.A., Medstrand, P., and Mager, D.L., 2003, "An Endogenous Retroviral Long Terminal Repeat Is the Dominant Promoter for Human Beta1,3-Galactosyltransferase 5 in the Colon", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 100, No. 22, pp. 12841-12846.

26. Dunn, C.A., Romanish, M.T., Gutierrez, L.E., van de Lagemaat, L.N., and Mager, D.L., 2006, "Transcription of Two Human Genes from a Bidirectional Endogenous Retrovirus Promoter", **Gene**, Vol. 366, No. 2, pp. 335-342.

27. Di Cristofano, A., Strazzullo, M., Longo, L., and La Mantia, G., 1995, "Characterization and Genomic Mapping of the Znf80 Locus: Expression of This Zinc-Finger Gene Is Driven by a Solitary Ltr of Erv9 Endogenous Retrovrial Family", **Nucleic Acids Research**, Vol. 23, No. 15, pp. 2823-2830.

28. Feuchter-Murthy, A.E., Freeman, J.D., and Mager, D.L., 1993, "Splicing of a Human Endogenous Retrovirus to a Novel Phospholipase A2 Related Gene", **Nucleic Acids Research**, Vol. 21, No. 1, pp. 135-143.

29. Eimohng, S. and Deesomchok, U., 2001, **Systemic Lupus Erythematosus**, Text and Journal Publication Co., Ltd., pp.

30. Rahman, A. and Isenberg, D.A., 2008, "Systemic Lupus Erythematosus", **New England Journal of Medicine**, Vol. 358, No. 9, pp. 929-939.

31. Shirai, T. and Hirose, S., 2006, "Molecular Pathogenesis of Sle", **Springer Seminars in Immunopathology**, Vol. 28, No. 2, pp. 79-82.

32. Ines, C. and Robert, F.G., 2006, "Role of Endogenous Retroviruses in Autoimmune Diseases", **Infectious disease clinics of North America**, Vol. 20, No. 4, pp. 913-929.

33. Adelman, M.K. and Marchalonis, J.J., 2002, "Endogenous Retroviruses in Systemic Lupus Erythematosus: Candidate Lupus Viruses", **Clinical Immunology**, Vol. 102, No. 2, pp. 107-116.

34. Sekigawa, I., Ogasawara, H., Naito, T., Kaneko, H., Hishikawa, T., and Hashimoto, H., 2003, "Systemic Lupus Erythematosus and Human Endogenous Retroviruses", **Modern Rheumatology**, Vol. 13, No. 2, pp. 107-113.

35. Naito, T., Ogasawara, H., Kaneko, H., Hishikawa, T., Sekigawa, I., Hashimoto, H., and Maruyama, N., 2003, "Immune Abnormalities Induced by Human Endogenous Retroviral Peptides: With Reference to the Pathogenesis of Systemic Lupus Erythematosus", **Journal of Clinical Immunology**, Vol. 23, No. 5, pp. 371-376.

36. Chu, J.L., Drappa, J., Parnassa, A., and Elkon, K.B., 1993, "The Defect in Fas Mrna Expression in Mrl/Lpr Mice Is Associated with Insertion of the

Retrotransposon, Etn", **The Journal of Experimental Medicine**, Vol. 178, No. 2, pp. 723-730.

37. Balada, E., Ordi-Ros, J., and Vilardell-Tarrés, M., 2009, "Molecular Mechanisms Mediated by Human Endogenous Retroviruses (Hervs) in Autoimmunity", **Reviews in Medical Virology**, Vol. 19, No. 5, pp. 273-286.

38. Balada, E., Vilardell-Tarres, M., and Ordi-Ros, J., 2009, "Implication of Human Endogenous Retroviruses in the Development of Autoimmune Diseases", **International Reviews of Immunology**, Vol. 29, No. 4, pp. 351-370.

39. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A.H., 2008, "A Universal Classification of Eukaryotic Transposable Elements Implemented in Repbase", **Nat Rev Genet**, Vol. 9, No. 5, pp. 414-414.

40. Bergman, C.M. and Quesneville, H., 2007, "Discovering and Detecting Transposable Elements in Genome Sequences", **Briefings in Bioinformatics**, Vol. 8, No. 6, pp. 382-392.

41. Pereira, V., 2008, "Automated Paleontology of Repetitive DNA with Reannotate", **BMC Genomics**, Vol. 9, No. 1, p. 614.

42. McDonald, J.H., 2009, **Handbook of Biological Statistics**, 2nd ed., Sparky House Publishing, Maryland, pp. 24-32.

43. Ott, R.L. and Longnecker, M., 2010, **An Introduction to Statistical Methods and Data Analysis**, 6th ed., Brooks/Cole Cengage Learning, pp. 511-513.

44. Smit, A., Hubley, R., and Glusman, G., **Repeatmasker**, [online], Available: http://www.repeatmasker.org/ [2011, 8 June].

45. **Python Programming Language - Official Website**, [online], Available: http://www.python.org/ [2011, 8 June].

46. Schumacher, R. and Lentz, A., **Dispelling the Myths**, [online], Available: http://dev.mysql.com/tech-resources/articles/dispelling-the-myths.html [2011, 8 June].

47. Lerdorf, R., 1994, **Php: Hypertext Preprocessor**, [online], Available: http://www.php.net/ [2011, 8 June].

48. McCool, R., 1995, **The Apache Software Foundation**, [online], Available: http://www.apache.org/foundation/ [2011, 8 June].

49. Ihaka, R. and Gentleman, R., 1996, "R: A Language for Data Analysis and Graphics", **Journal of Computational and Graphical Statistics**, Vol. 5, No. 3, pp. 299-314.

50. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J., 2004, "The Ucsc Table Browser Data Retrieval Tool", **Nucleic Acids Research**, Vol. 32, No. suppl 1, pp. D493-D496.

51. Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D., 2006, "The Ucsc Known Genes", **Bioinformatics**, Vol. 22, No. 9, pp. 1036-1046.

52. Lee, H.-M., Sugino, H., Aoki, C., and Nishimoto, N., 2011, "Underexpression of Mitochondrial-DNA Encoded Atp Synthesis-Related Genes and DNA Repair Genes in Systemic Lupus Erythematosus", **Arthritis Research & Therapy**, Vol. 13, No. 2, p. R63.

53. Bansal, V., Halpern, A.L., Axelrod, N., and Bafna, V., 2008, "An Mcmc Algorithm for Haplotype Assembly from Whole-Genome Sequence Data", **Genome Research**, Vol. 18, No. 8, pp. 1336-1346.

54. Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., and Warburton, P.E., 2007, "Evolutionary History of Mammalian Transposons Determined by Genome-Wide Defragmentation", **PLoS Comput Biol**, Vol. 3, No. 7, p. e137.

55. Pereira, V., 2004, "Insertion Bias and Purifying Selection of Retrotransposons in the Arabidopsis Thaliana Genome", **Genome Biology**, Vol. 5, No. 10, p. R79.

APPENDIX A

List of HERV families and names/groups

Superfamily	Family	Name/group
1. ERV1	ERV9	HERV9, LTR12, LTR12B, LTR12C, LTR12D, LTR12E, LTR12F
	HERV1	HERV1_I, HERV1_LTRa, HERV1_LTRb, HERV1_LTRc, HERV1_LTRd, HERV1_LTRe
	HERV15	HERV15, LTR15
	HERV17	HERV17, LTR17
	HERV23	LTR23, LTR44, LTR56
	HERV3	HERV3, LTR4
	HERV30	HERV30, LTR30
	HERV35	HERV35I, LTR35, LTR35A, LTR35B
	HERV38	LTR38, LTR38B, LTR38C
	HERV39	LTR39
	HERV4	HERV4_I
	HERV43	LTR43, LTR43B
	HERV45	LTR45, LTR45B, LTR45C
	HERV46	LTR46
	HERV49	LTR49
	HERV70	LTR70
	HERVE	HERVE, HERVE_a
	HERVFc1	HERVFc1, HERVFc1_LTR1, HERVFc1_LTR2, HERVFc1_LTR3
	HERVFc2	HERVFc2
	HERVFH19	HERVFH19
	HERVFH21	HERVFH21, LTR21A, LTR21B
	HERVH	HERVH, LTR7, LTR7A, LTR7B, LTR7C, LTR7Y
	HERVH48	HERVH48, MER48
	HERVI	HERVI, LTR10A, LTR10B, LTR10B1, LTR10C, LTR10D, LTR10E, LTR10G
	HERVIP10	HERVIP10F, HERVIP10FH, LTR10F
	HERVP71A	HERVP71A, LTR71A, LTR71B
	HERVS71	HERVS71, LTR6A, LTR6B
	HERVW	Harlequin_I, LTR2, LTR2B, LTR2C
	HUERSP1	HUERSP1, LTR8, LTR8A
	HUERSP2	HUERSP2, LTR1, LTR1B, LTR1C, LTR1D
	HUERSP3	HUERSP3, HUERSP3b, LTR9, LTR9B
	LOR1	LOR1, LOR1a, LOR1b, LTR26, LTR26B, LTR26E
	LTR19	LTR19, LTR19A, LTR19B, LTR19C
	LTR24	LTR24, LTR24B, LTR24C
	LTR25	LTR25
	LTR27	LTR27, LTR27B
	LTR28	LTR28
	LTR29	LTR29
	LTR31	LTR31
	LTR34	LTR34
	LTR36	LTR36
	LTR37	LTR37A, LTR37B
	LTR48	LTR48, LTR48B
	LTR51	LTR51
	LTR54	LTR54, LTR54B

Superfamily	Family	Name/group
1. ERV1 (cont.)	LTR58	LTR58
	LTR59	LTR59
	LTR60	LTR60
	LTR61	LTR61
	LTR64	LTR64
	LTR65	LTR65
	LTR68	LTR68
	LTR72	LTR72, LTR72B
	LTR75_1	LTR75_1
	LTR76	LTR76
	LTR77	LTR77
	LTR78	LTR78, LTR78B
	MER101	MER101, MER101B
	MER110	MER110, MER110A
	MER31	MER31, MER31A, MER31B
	MER34	MER34, MER34A, MER34A1, MER34B, MER34C, MER34C2, MER34D
	MER39	MER39, MER39B
	MER4	MER4, MER4A, MER4A1, MER4B, MER4C, MER4D, MER4D0, MER4D1, MER4E, MER4E1
	MER41	MER41, MER41A, MER41B, MER41C, MER41D, MER41E, MER41G
	MER49	MER49
	MER50	MER50, MER50B, MER50C
	MER51	MER51, MER51A, MER51B, MER51C, MER51D, MER51E
	MER52	MER52, MER52A, MER52C, MER52D
	MER57	MER57, MER57A, MER57A1, MER57B1, MER57B2, MER57C1, MER57C2, MER57D, MER57E1, MER57E2, MER57E3, MER57F
	MER61	MER61, MER61A, MER61B, MER61C, MER61D, MER61E, MER61F
	MER65	MER65, MER65A, MER65B, MER65C, MER65D
	MER66	LTR73, MER66, MER66A, MER66B, MER66C, MER66D
	MER67	MER67A, MER67B, MER67C, MER67D
	MER72	MER72, MER72B
	MER83	MER83, MER83A, MER83B, MER83C
	MER84	MER84
	MER87	MER87, MER87B
	MER89	MER89
	MER90	MER90a
	MER92	MER92A, MER92B
	PAB	PABL_A, PABL_B
	PRIMA4	PRIMA4, PRIMAX
	PRIMA41	PRIMA41
	PrimLTR79	PrimLTR79
2. ERVK	HERVK14/ HERVK (HML-1)	HERVK14, HERVK14C, LTR14, LTR14A, LTR14B, LTR14C

Superfamily	Family	Name/group
2. ERVK (cont.)	HERVK10/ HERVK (HML-2)	LTR5, LTR5_Hs, LTR5A, LTR5B, HERVK
	HERVK9/ HERVK (HML-3)	HERVK9, MER9a1, MER9a2, MER9a3, MER9B
	HERVK13/ HERVK (HML-4)	HERVK13, LTR13, LTR13A
	HERVK22/ HERVK (HML-5)	HERVK22, LTR22, LTR22A, LTR22B, LTR22C
	HERVK3/ HERVK (HML-6)	HERVK3, LTR3, LTR3A, LTR3B
	HERVK11D/ HERVK (HML-7)	HERVK11D, MER11D
	HERVK11/ HERVK (HML-8)	HERVK11, MER11A, MER11B, MER11C
	HERVKC4/ HERVK (HML-10)	HERVKC4
3. ERVL	ERV3-16A3	ERV3-16A3
	ERVL	ERVL, ERVLB4, ERVLE, HERVL, MLT2A1, MLT2A2, MLT2B1, MLT2B2, MLT2B3, MLT2B4, MLT2B5, MLT2C1, MLT2C2, MLT2D, MLT2E, MLT2F
	HERV16	HERV16, LTR16A, LTR16A1, LTR16A2, LTR16B, LTR16B1, LTR16B2, LTR16C, LTR16D, LTR16D1, LTR16D2, LTR16E1, LRE16E2
	HERV18	HERVL18, LTR18A, LTR18B
	HERV32	HERVL32, LTR32
	HERV47	LTR47A, LTR47B
	HERVL33	LTR33, LTR33A, LTR33B, LTR33C, LTR41, LTR41B
	HERVL40	HERVL40, LTR40a, LTR40A1, LTR40b, LTR40c
	HERVL42	LTR42
	HERVL50	LTR50
	HERVL52	LTR52
	HERVL53	LTR53, MER88
	HERVL54	MER54A, MER54B
	HERVL57	LTR57
	HERVL66	HERVL66, LTR66
	HERVL67	LTR67B
	HERVL68	MER68, MER68B
	HERVL69	LTR69
	HERVL70	MER70, MER70A, MER70B, MER70C
	HERVL73	MER73
	HERVL74	HERVL74, MER74A, MER74B, MER74C
	HERVL75	LTR75, LTR75B
	LTR62	LTR62
	LTR79	LTR79
	LTR80	LTR80A, LTR80B
	LTR82	LTR82A, LTR82B
	LTR83	LTR83
	LTR84	LTR84a, LTR84b
	LTR86	LTR86A1, LTR86A2, LTR86B1, LTR86B2, LTR86C

Superfamily	Family	Name/group
3. ERVL (cont.)	MER21	MER21, MER21A, MER21B, MER21C
	MER76	MER76
	MER77	MER77, MER77B
4. ERVL-MaLR	MLT1	MLT1, MLT1A, MLT1A0, MLT1A1, MLT1B, MLT1C, MLT1D, MLT1E, MLT1E1, MLT1E1A, MLT1E2, MLT1E3, MLT1F, MLT1F1, MLT1F2, MLT1G, MLT1G1, MLT1G3, MLT1H, MLT1H1, MLT1H2, MLT1I, MLT1J, MLT1J1, MLT1J2, MLT1K, MLT1L, MLT1M, MLT1N2
	MST	MST, MSTA, MSTB, MSTB1, MSTB2, MSTC, MSTD
	THE1	MLT, THE1, THE1A, THE1B, THE1C, THE1D
5. Unclassified ERVs	LTR11	LTR11
	LTR55	LTR55
	LTR87	LTR87
	LTR89	LTR89
	MER95	MER95

APPENDIX B

Data dictionary of HERV Profiler database

Table name	Attribute	Description	Type
erv_elements	e_id (PK)	HERV element ID	Text
	chr	chromosome	Text
	name	element name or group name	Text
	eStart	starting position of an element	Number
	eEnd1	ending position of a left LTR	Number
	eStart2	starting position of a right LTR	Number
	eEnd	ending position of an element	Number
	ref1	intactness ratio of the left LTR of an element	Number
	refI	intactness ratio of the internal sequence	Number
	ref2	intactness ratio of the right LTR of an element	Number
	family	name of HERV family	Text
	superfamily	name of HERV superfamily	Text
	trun_patt	type of truncation patterns of an element	Number
	e_intactness	intactness ratio of an element	Number
erv_fragments	f_id (PK)	HERV fragment ID	Text
	SWscore	Smith-Waterman score of the match	Number
	perc_div	percentage of bases substituted in bp	Number
	perc_del	percentage of bases deleted in bp	Number
	perc_ins	percentage of bases inserted in bp	Number
	qStart	starting position of match in query sequence	Number
	qEnd	ending position of match in query sequence	Number
	qLeft	no. of bases in the query prior to match beginning	Number
	rName	name of matching interspersed repeat	Text
	rStart	starting position of match in database sequence	Number
	rEnd	ending position of match in database sequence	Number
	rLeft	no. of bases in the repeat prior to match beginning	Number
	e_id (FK)	HERV element ID	Text
	part_type	part name of an element (LTR or INT)	Number
gene_erv_chr	gene_id	UCSC gene ID	Text
	e_id	HERV element ID	Text
	part_of_gene	location name relative to genes	Number
	distance	distance between an element and a gene in bp	Number
	eStart	starting position of an element	Number
	eEnd	ending position of an element	Number
	part_in_gene	in-gene location name (e.g., intron, exon)	Number
	no_in_gene	exon or intron no. an HERV inserted to	Number
kgxref	kgID (PK)	UCSC gene ID	Text
	mRNA	mRNA ID	Text
	spID	SWISS-PROT protein Accession number	Text

Table name	Attribute	Description	Type
	spDisplayID	SWISS-PROT display ID	Text
	geneSymbol	Gene Symbol	Text
	refseq	RefSeq ID	Text
	protAcc	NCBI protein Accession number	Text
	description	Description	Text
knowngene	name (PK)	UCSC gene ID	Text
	chr	Chromosome	Text
	strand	Strand of a gene (+ or -)	Text
	txStart	starting position of transcription start site	Number
	txEnd	ending position of transcription termination site	Number
	cdsStart	starting position of coding sequence	Number
	cdsEnd	ending position of coding sequence	Number
	exonCount	the number of exons	Number
	exonStarts	a list of starting position of each exon	Text
	exonEnds	a list of ending position of each exon	Text
	proteinID	UniProt display ID	Text
	alignID	Unique identifier for each pair	Text

APPENDIX C

HERV Profiler User's Guide

HERV PROFILER USER'S GUIDE

OUTLINE

- Introduction
- A glimpse of HERV Profiler
- Specifying inputs of HERV Profiler
- Visualizing neighboring HERVs of interested genes
    - Visualizing neighboring HERVs in tabular format
    - Visualizing neighboring HERVs in graphical format
- Profiling neighboring HERVs of genes in a gene list
    - Constructing HERV profiles
        - Profile display and retrieval
        - HERV type display and retrieval
    - Finding over-represented HERV types
    - Ranking genes in a gene list following to their HERV types

INTRODUCTION

Human Endogenous Retroviruses (HERVs) are retroviral-like sequences found in human genome. It has been supposed that HERVs are remnants of ancient retroviral infections that incorporated in the host genome long time ago. The ubiquitous presence of them throughout our genome leads us to question whether and how they can affect us. One that has been found and still discussing now is that HERVs could influence the functions of the neighboring genes. However, not every HERV could affect the neighboring genes, because the HERVs present in the human genome are somewhat different in several ways. In other words, it is feasible that only some HERVs with some particular characteristics could be responsible for this phenomenon. To facilitate who are interested in this point, HERV Profiler has thus been developed to provide a simpler way to investigate the HERVs adjacent the interested genes.

HERV Profiler is a web-based tool for visualizing and profiling neighboring HERVs of human genes from a gene list. In other words, this bioinformatics tool can facilitate the investigation of the neighboring HERVs of your interested genes. There are two main features provided in HERV Profiler. The first one is visualizing the neighboring HERVs of your interested genes. The information of the HERVs and their neighboring genes are shown in either tables or figures. Another feature is profiling the neighboring HERVs of genes in the gene list. This process is mainly based on the assumptions that only

HERVs with particular characteristics would be able to influence the adjacent genes, and the HERVs which are over-represented among a correlated gene list would be given the attention first in this situation. HERV profiling can be used to purpose potential candidates of the HERV types and their corresponded genes, which may be affected by the neighboring HERVs under a studied condition. HERV Profiler is freely available at http://sigma.cpe.kmutt.ac.th/herv_profiler. Finally, we hope that HERV Profiler would be a powerful tool that can help you to reveal the valuable information in the studies on the influence of the HERVs to the neighboring genes.

A GLIMPSE OF HERV PROFILER

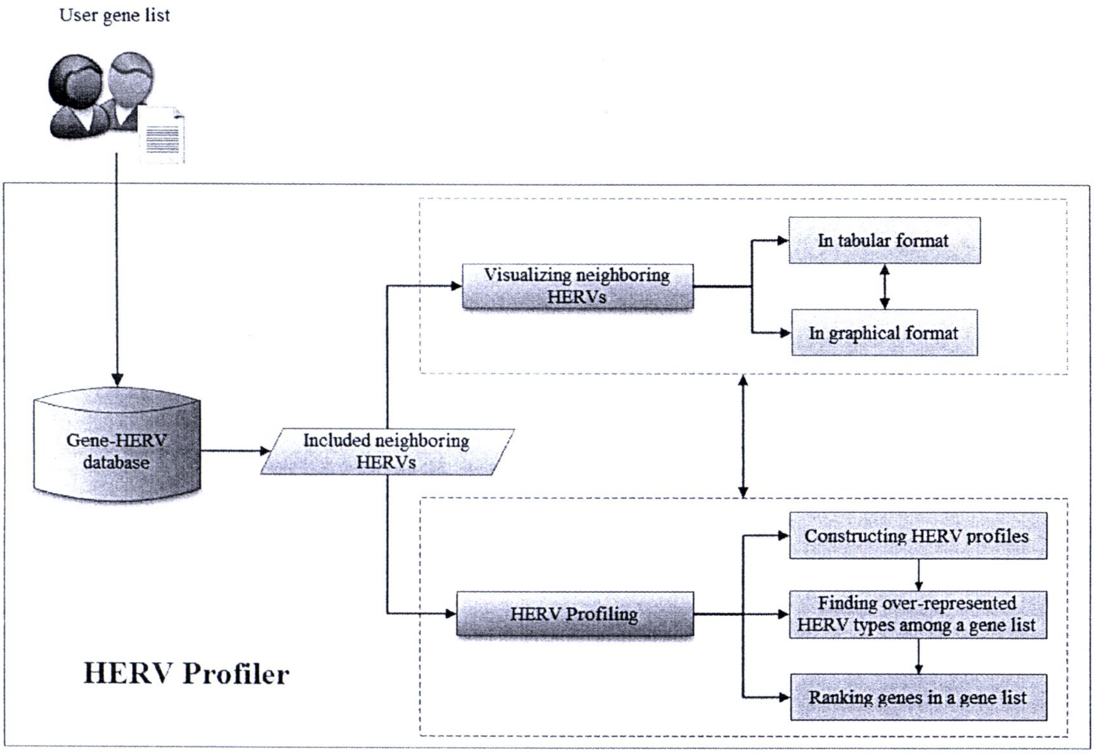


Figure C.1 A bird's eye view describing overall features of HERV Profiler

There are two main features provided in HERV Profiler to facilitate investigations of the HERVs adjacent genes, including visualizing and profiling. These two features are not completely separated from each other, but they could be co-operatively used to achieve the more efficient utilization (Figure C.1).

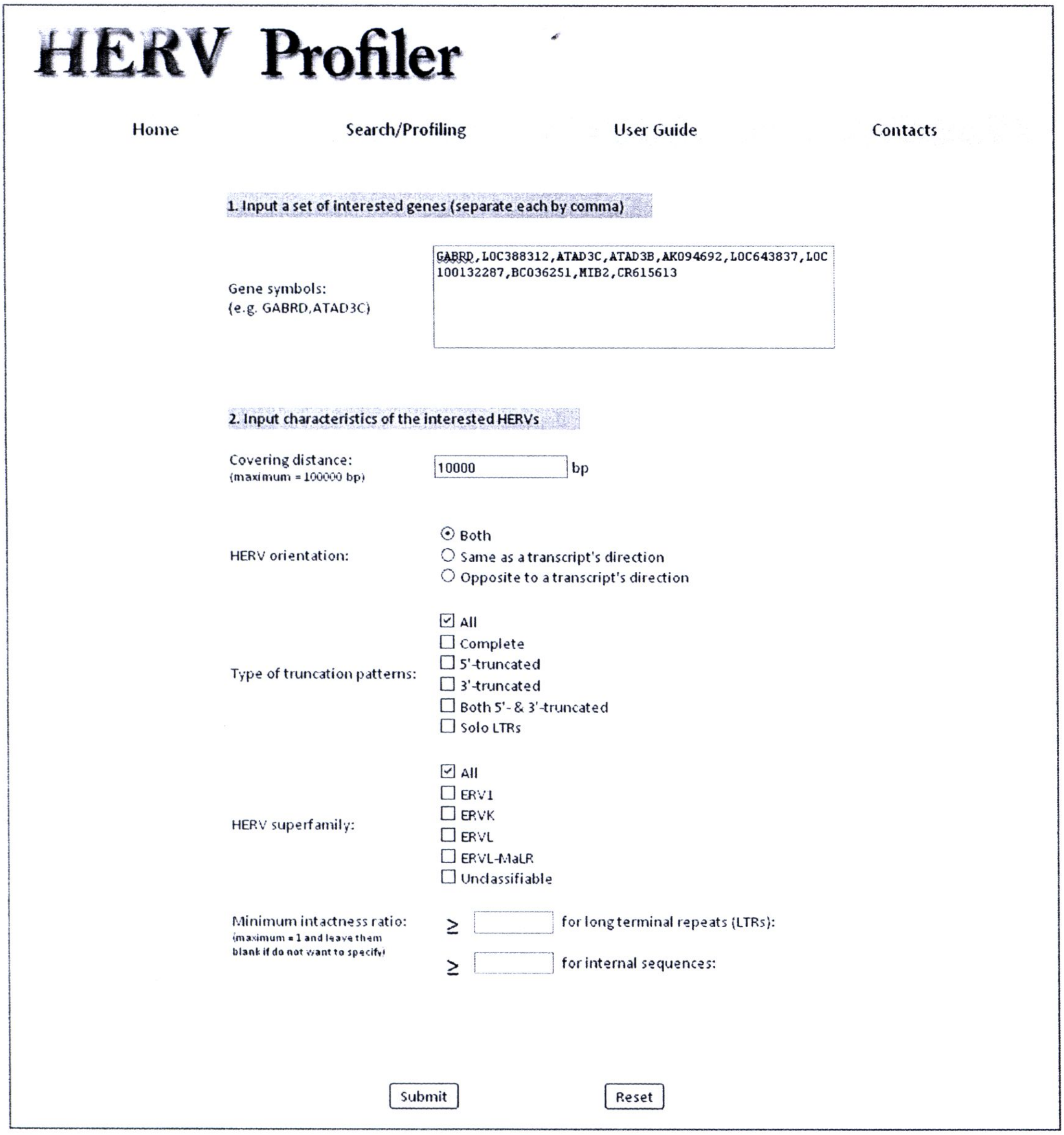
According to Figure C.1, an important input of HERV Profiler is a gene list. A list of genes which are significantly differentially expressed in a certain condition is preferred in order to look for the HERV candidates that probably influence the adjacent genes under the studied condition. Reasonable number of genes ranges from hundreds to thousands (e.g., 100-2,000 genes), not extremely low or high. After input the gene list, users need to specify the characteristics of the HERVs that would like include in both visualizing and profiling. This would be beneficial when users would like to study on a certain HERVs, such as a particular HERV superfamily. Only the included HERVs would be further used in either visualizing or profiling.

From the first feature provided, the visualizing, you can search for neighboring HERVs of your interested genes. There are several characteristics collected for each HERV, such as family, superfamily, truncation patterns, intactness ratio, and inserted locations relative to genes. Furthermore, for more precisely identifying the exact locations of the HERVs relative to genes, gene isoforms, different transcription patterns of a gene, were used. In other words, there are likely more than one HERV-gene patterns for a gene symbol input in HERV Profiler. The results of your searching can be shown in either tables or figures. In addition, you can also pay your attention to the genes which host many HERVs first by ranking the genes according to the number of the HERVs.

Another feature of HERV Profiler is the HERV Profiling. This is a process for analyzing a gene list. This feature could help you to purpose the candidates of the HERVs and their corresponded genes by being based on the preference of the over-represented HERVs. The HERV profiling is composed of three ordered sub-tasks, including constructing HERV profiles, finding over-represented HERV types, and ranking genes in a gene list (Figure C.1).

Next, how to specify all of the inputs of HERV Profiler would be illustrated.

SPECIFYING INPUTS OF HERV PROFILER



1. Input a set of interested genes (separate each by comma)

Gene symbols:
(e.g. GABRD,ATAD3C)
GABRD,LOC388312,ATAD3C,ATAD3B,AK094692,LOC643837,LOC100132287,BC036251,MIB2,CR615613

2. Input characteristics of the interested HERVs

Covering distance:
(maximum = 100000 bp)
10000 bp

HERV orientation:
● Both
○ Same as a transcript's direction
○ Opposite to a transcript's direction

Type of truncation patterns:
☑ All
☐ Complete
☐ 5'-truncated
☐ 3'-truncated
☐ Both 5'- & 3'-truncated
☐ Solo LTRs

HERV superfamily:
☑ All
☐ ERV1
☐ ERVK
☐ ERVL
☐ ERVL-MaLR
☐ Unclassifiable

Minimum intactness ratio:
(maximum = 1 and leave them blank if do not want to specify)
≥ for long terminal repeats (LTRs):
≥ for internal sequences:

Figure C.2 Search/Profiling page

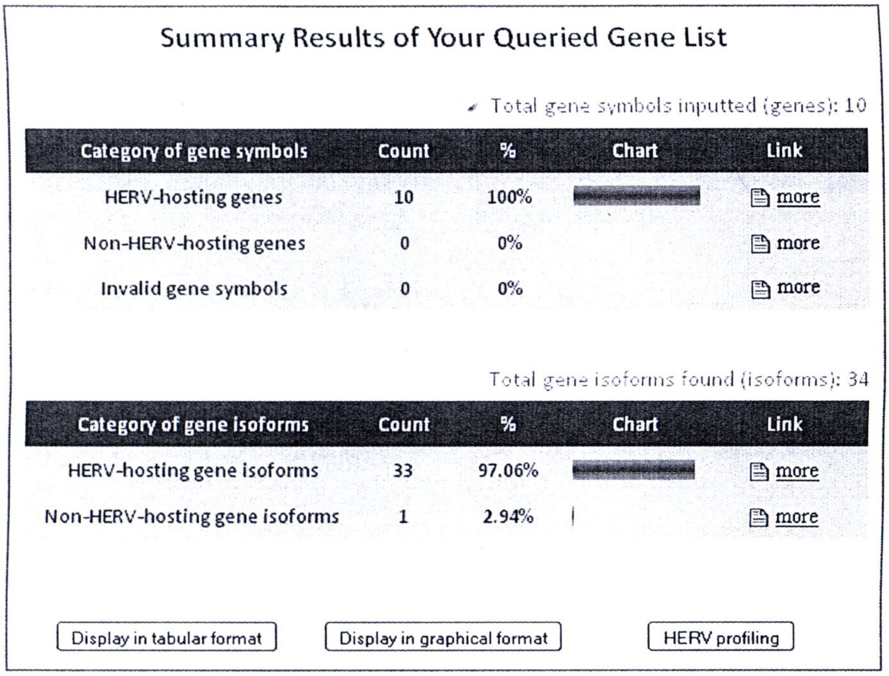
You can select on "Search/Profiling" at the top menu bar to begin using HERV Profiler. Then, the search/profiling form would be shown (Figure C.2). There are two types of inputs required on this form, including the gene list and the specified characteristics of the HERVs that would like to include into the system. The descriptions of each input fields are discussed below.

Descriptions of the input fields:

1. Specifying genes
    - **Gene symbols** = a list of gene symbols which each should be separated by commas (,).

2. Specifying HERVs
   - **Covering distance** = a distance used to determine a neighborhood of genes, where is far away upwards and downwards from a transcription start site and termination site, respectively. The maximum covering distance is up to 100000 bp. Any characters, minus values, and numbers higher than 100000 are not acceptable.
   - **HERV orientation** = a direction of an HERV relative to a transcription direction of a gene.
   - **Type of truncation patterns** = type of an HERV element classified following to the patterns of the truncated parts. For example, if an element has no a 5'-LTR, the element would be classified as a 5'-LTR truncated element. There are five types of truncation patterns here: complete, 5'-truncated, 3'-truncated, both 5'- and 3'-truncated elements, and solo LTRs.
   - **HERV superfamily** = superfamily of HERVs. This classification is based on Repbase Update database. There are five superfamilies: ERV1, ERVK, ERVL, ERVL-MaLR, and unclassified class.
   - **Minimum intactness ratio** = the least intactness ratio which is acceptable. HERV elements that have the intactness ratios higher than the minimum value would be included in both visualizing and profiling further. The minimum intactness ratios are required to specify separately between the minimum of LTRs and the minimum of internal regions. An element would be included into the system when it contains at least a part of the element that has the intactness ratios higher than the specified value. Any characters, minus values, and numbers higher than one are not acceptable. The minimum ratios can be leaved blanks if you would not like to specify them.

After submitted all of the inputs, the summary report of your input genes would be shown (Figure C.3).



Summary Results of Your Queried Gene List

Total gene symbols inputted (genes): 10

Category of gene symbols	Count	%	Chart	Link
HERV-hosting genes	10	100%		more
Non-HERV-hosting genes	0	0%		more
Invalid gene symbols	0	0%		more

Total gene isoforms found (isoforms): 34

Category of gene isoforms	Count	%	Chart	Link
HERV-hosting gene isoforms	33	97.06%		more
Non-HERV-hosting gene isoforms	1	2.94%		more

Figure C.3 An example of the summary report

According to Figure C.3, noticeably, there are two tables for showing the results that are summarized in different point of views. The above table is a summary of the results obtained in terms of gene symbols, while the below table is a summary of the results calculated in terms of gene isoforms. In the gene symbols' table, there are three categories shown there, including HERV-hosting, non-HERV-hosting, and invalid genes. The HERV-hosting-genes are gene symbols which have at least a gene isoform containing neighboring HERVs, while the non-HERV-genes are gene symbols which all corresponding isoforms do not contain HERVs. The invalid genes are gene symbols which cannot be found in the HERV Profiler database or are typed incorrectly. The count numbers are the numbers of gene symbols or gene isoforms which belong to particular categories. The percentages are calculated from the fractions to gene symbols or isoforms in total. You can find a gene list of a particular category out by selecting on "more" at the far right column of the tables. The menu buttons for performing the investigating HERVs are also shown in this summary page. There are three choices provided here, including "display in tabular format", "display in graphical format", and "HERV profiling".

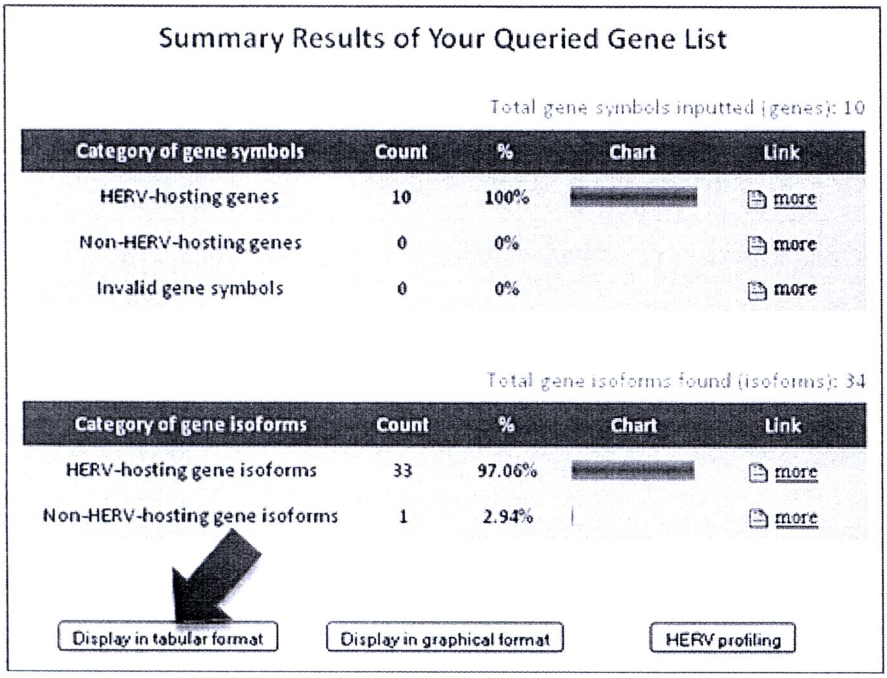
VISUALIZING NEIGHBORING HERVS
OF INTERESTED GENES

There are two different views provided in HERV Profiler to visualize neighboring HERVs, including displaying in tabular formats and in graphical formats. The illustrations how to use this feature would be described in details next.

VISUALIZING NEIGHBORING HERVS IN TABULAR FORMAT

1. Begin with specifying your gene list and HERVs that you would like to include into the system at the Search/Profiling page.

2. Choose the button "Display in tabular format" in the summary page (Figure C.4). The results would then be shown as illustrated in Figure C.5.



Summary Results of Your Queried Gene List

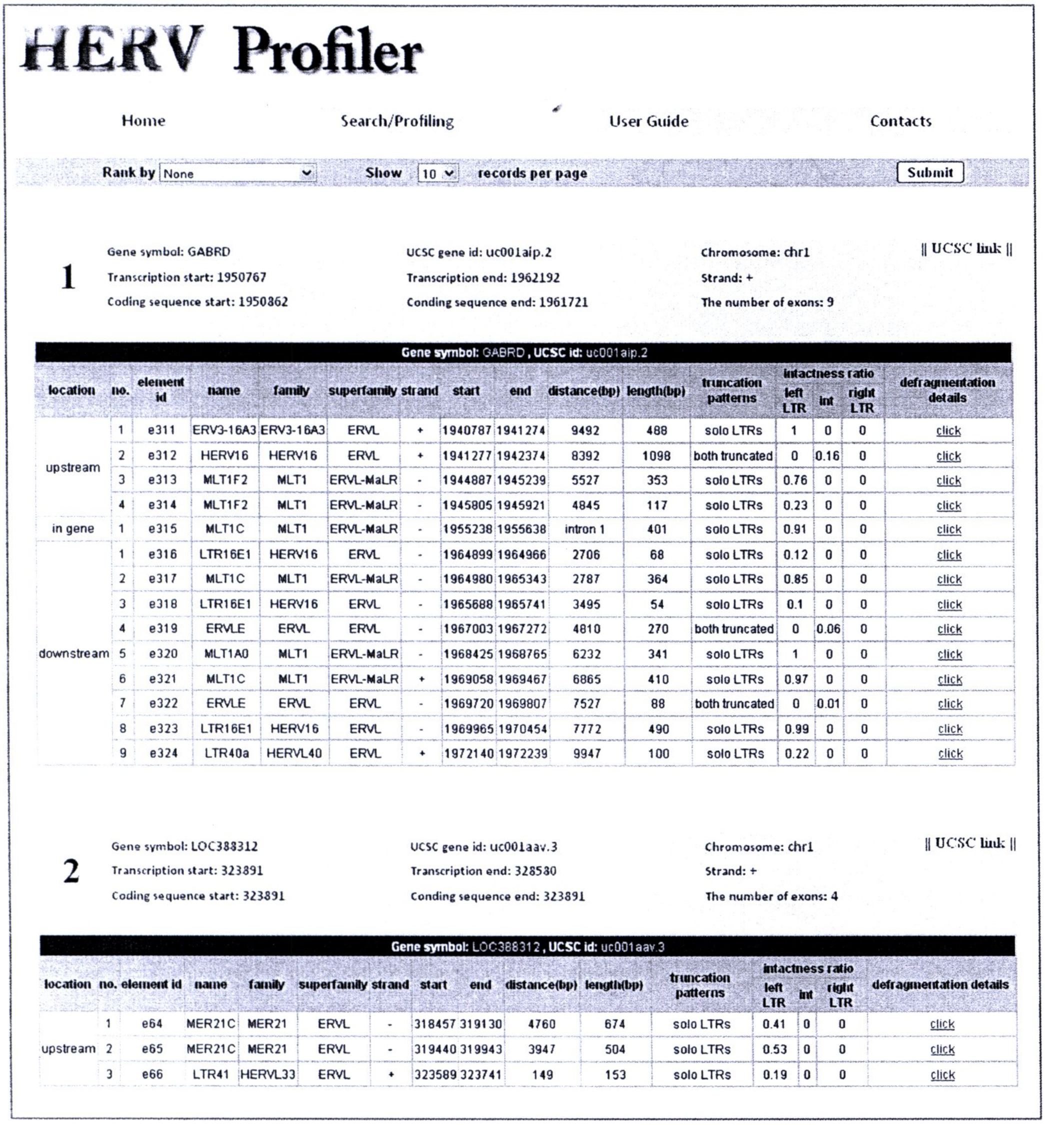
Total gene symbols inputted (genes): 10

Category of gene symbols	Count	%	Chart	Link
HERV-hosting genes	10	100%		more
Non-HERV-hosting genes	0	0%		more
Invalid gene symbols	0	0%		more

Total gene isoforms found (isoforms): 34

Category of gene isoforms	Count	%	Chart	Link
HERV-hosting gene isoforms	33	97.06%		more
Non-HERV-hosting gene isoforms	1	2.94%		more

Figure C.4 Location of the button of displaying in tabular format



Rank by None
Show 10 records per page

1

Gene symbol: GABRD
Transcription start: 1950767
Coding sequence start: 1950862
UCSC gene id: uc001aip.2
Transcription end: 1962192
Conding sequence end: 1961721
Chromosome: chr1
Strand: +
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

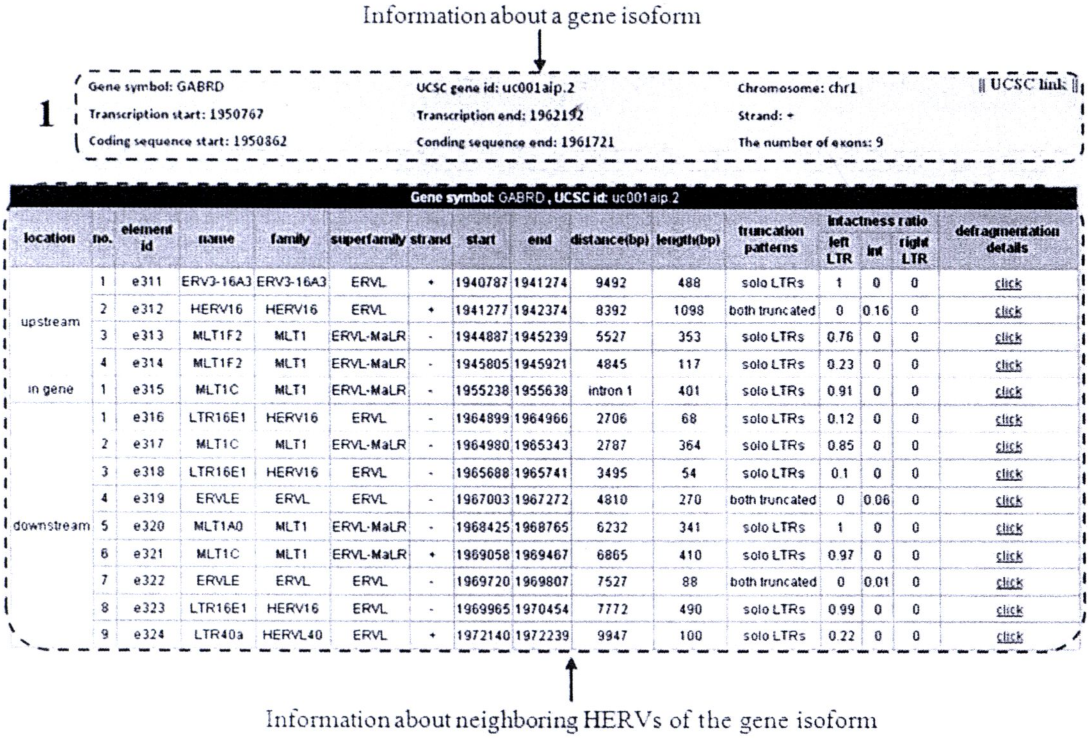
2

Gene symbol: LOC388312
Transcription start: 323891
Coding sequence start: 323891
UCSC gene id: uc001aav.3
Transcription end: 328580
Conding sequence end: 323891
Chromosome: chr1
Strand: +
The number of exons: 4

Gene symbol: LOC388312 , UCSC id: uc001aav.3

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	e64	MER21C	MER21	ERVL	-	318457	319130	4760	674	solo LTRs	0.41	0	0	click
	2	e65	MER21C	MER21	ERVL	-	319440	319943	3947	504	solo LTRs	0.53	0	0	click
	3	e66	LTR41	HERVL33	ERVL	+	323589	323741	149	153	solo LTRs	0.19	0	0	click

Figure C.5 Results of the tabular displaying of neighboring HERVs



Information about a gene isoform

1

Gene symbol: GABRD
Transcription start: 1950767
Coding sequence start: 1950862
UCSC gene id: uc001aip.2
Transcription end: 1962192
Conding sequence end: 1961721
Chromosome: chr1
Strand: +
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

Information about neighboring HERVs of the gene isoform

Figure C.6 Details of the gene isoform and HERV information shown in a rectangular area

According to Figure C.5, all of the information about a gene isoform and their neighboring HERVs would be shown in a rectangular area, and each area would be labeled with the number shown at the top left corner of the rectangular area. Each of this rectangular area is used for displaying the information of one gene isoform only. The information about the gene isoform, such as gene symbol, UCSC gene ID, transcription start and end, coding sequence start and end, strand, chromosome, and the number of exons, is shown above the table (Figure C.6). All of the HERVs in a gene's neighborhood and their characteristics are shown in the table (Figure C.6). The descriptions of the gene and HERV characteristics shown in a rectangular area are discussed below.

Descriptions of the characteristics shown in a rectangular area (Figure C.6):

1. Gene isoform information
    - Gene symbol = gene symbol of a gene isoform
    - UCSC gene id = UCSC gene ID
    - Chromosome = chromosome on where a gene isoform is located
    - Transcription start = position of the transcription start site of a gene isoform
    - Transcription end = position of the transcription termination site of a gene isoform
    - Strand = DNA strand on where a gene isoform is located. It can be either plus (+) or minus (-). In HERV Profiler, a plus (+) represents a DNA

strand in a 5'-3' direction, while a minus (-) represents a DNA strand in a 3'-5' direction.

- Coding sequence start = starting position of the coding sequence of a gene isoform.
- Coding sequence end = ending position of the coding sequence of a gene isoform.
- The number of exons = the number of exons found in a transcription unit of a gene isoform
- UCSC link = link of the UCSC database for more information about a gene isofrom

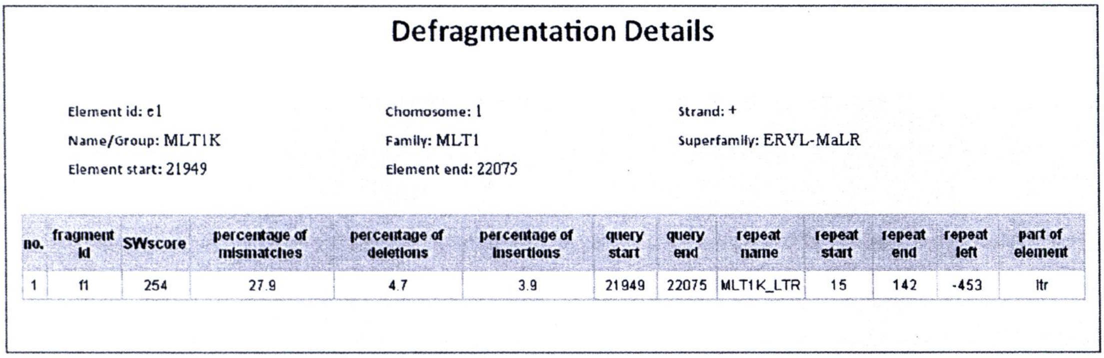
2. HERV information
    - location = location of an HERV element relative to the gene isoform (upstream, in gene, or downstream).
    - no. = the label number of an HERV element
    - element id = the ID of an HERV element, specially assigned in HERV Profiler
    - name = name or group of an HERV element
    - family = family of an HERV element
    - superfamily = superfamily of an HERV element (ERV1, ERVK, ERVL, ERVL-MaLR, and unclassified superfamily).
    - strand = DNA strand on where an HERV element is located. It can be either plus (+) or minus (-). In HERV Profiler, a plus (+) represents a DNA strand in a 5'-3' direction, while a minus (-) represents a DNA strand in a 3'-5' direction.
    - start = starting position of an HERV element
    - end = ending position of an HERV element
    - distance (bp) = distance measured in base pairs indicating how far the HERV element from the gene isoform is. When the element is inserted in the upstream region of the gene, the distance would be measured from the ending position of the element to the starting position of the gene. If the element is inserted in downstream region of the gene, the distance would be measured from the ending position of the gene to the starting position of the element. In case of in-gene insertions, the distance would be represented by a zero instead.
    - length (bp) = total length of an HERV element measured in base pairs
    - truncation patterns = type of truncation patterns of an HERV element (complete, 5'-truncated, 3'-truncated, both 5'- and 3'-truncated element, and solo LTRs).
    - intactness ratio = proportion of the total length of the element found in the human reference sequence to the length of the consensus HERV in Repbase database. This value was separately calculated in parts of the element, and should not be higher than one.

  - ▫ left LTR = intactness ratio of the left LTR of the element. If the element is located on the plus (+) strand, the left LTR refers to the 5'-LTR. In case of the rest, the left LTR refers to the 3'-LTR.
  - ▫ int = intactness ratio of the internal sequence of the element
  - ▫ right LTR = intactness ratio of the right LTR of the element. If the element is located on the plus (+) strand, the right LTR refers to the 3'-LTR. In case of the rest, the right LTR refers to the 5'-LTR.
- ▪ defragmentation details = link to the defragmentation details of the element

According to Figure C.5, there are displaying options provided above the result displaying area, including the option to rank the gene isoforms (tables) and the option to limit the genes shown in a page. The ranking option provides you to be able to rank the genes in a gene list following to the number of the neighboring HERVs. There are seven available choices for ranking the tables:

1. None = (Default) No special ranking, used an ordinary order from the input
2. Total number of HERVs = Rank the tables following to the number of the neighboring HERVs throughout the genes
3. Upstream HERVs = Rank the tables following to the number of upstream HERVs
4. In-gene HERVs = Rank the tables following to the number of in-gene HERVs, with no regarding to being located on either introns or exons
5. - Intron HERVs = Rank the tables following to the number of intron HERVs
6. - Exon HERVs = Rank the tables following to the number of exon HERVs
7. Downstream HERVs = Rank tables following to the number of downstream HERVs

According to Figures C.5-C.6, you can easily look for more details about the defragmentation of each element by selecting on "click" inside the last column of the tables. The defragmentation details of the selected element would then be shown as illustrated in Figure C.7.



Defragmentation Details

Element id: e1
Chomosome: 1
Strand: +
Name/Group: MLT1K
Family: MLT1
Superfamily: ERVL-MaLR
Element start: 21949
Element end: 22075

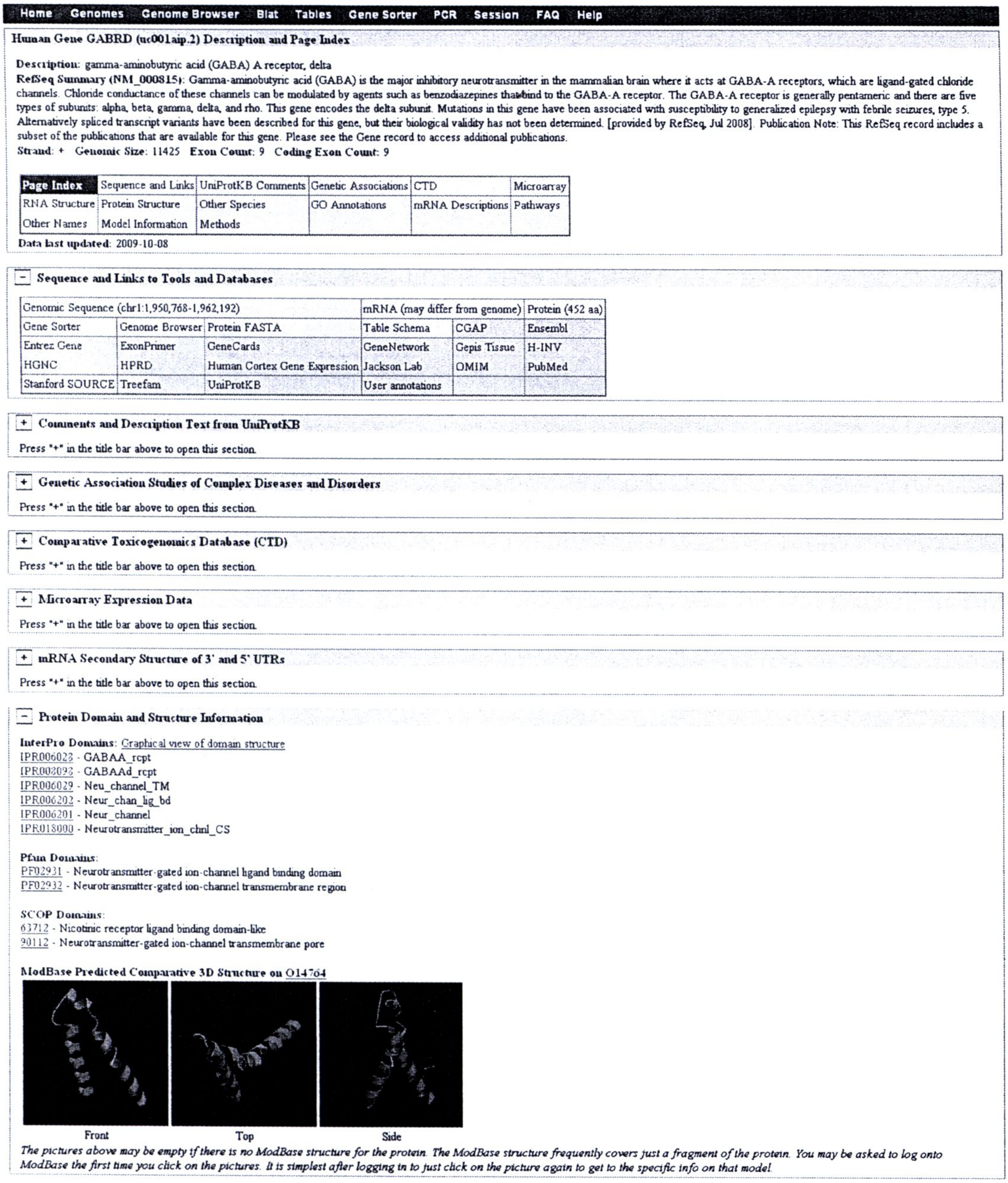
no.	fragment id	SWscore	percentage of mismatches	percentage of deletions	percentage of insertions	query start	query end	repeat name	repeat start	repeat end	repeat left	part of element
1	f1	254	27.9	4.7	3.9	21949	22075	MLT1K_LTR	15	142	-453	ltr

Figure C.7 An example of the defragmentation details

Descriptions of the defragmentation detail table:

- no. = label number of an HERV fragment
- fragment id = ID of an HERV fragment, specially assigned in HERV Profiler
- SWscore = Smith-Waterman score of the match
- percentage of mismatches = percentage of substitutions in matching region compared to the consensus
- percentage of deletions = percentage of bases opposite a gap in the query sequence (deleted bp)
- percentage of insertions = percentage of bases opposite a gap in the repeat consensus (inserted bp)
- query start = starting position of match in query sequence
- query end = ending position of match in query sequence
- repeat name = name of matching interspersed repeat
- repeat start = starting position of match in database sequence (using top-strand numbering)
- repeat end = ending position of match in database sequence
- repeat left = the number of bases in (complement of) the repeat consensus sequence prior to beginning of the match (so 0 means that the match extended all the way to the end of the repeat consensus sequence)
- part of element = type of an element part. The value of its can be either LTR (long terminal repeats) or INT (internal sequences)

Furthermore, in the result tables (Figures C.5-C.6), there are links connected to the UCSC database. These links provides you to easily look for more details about a gene isoform on the UCSC database. You can reach it by just selecting on "UCSC link" in the region of your desired gene. The details about the selected genes would be then shown as illustrated in Figure C.8.



Human Gene GABRD (uc001aip.2) Description and Page Index

Description: gamma-aminobutyric acid (GABA) A receptor, delta
RefSeq Summary (NM_000815): Gamma-aminobutyric acid (GABA) is the major inhibitory neurotransmitter in the mammalian brain where it acts at GABA-A receptors, which are ligand-gated chloride channels. Chloride conductance of these channels can be modulated by agents such as benzodiazepines that bind to the GABA-A receptor. The GABA-A receptor is generally pentameric and there are five types of subunits: alpha, beta, gamma, delta, and rho. This gene encodes the delta subunit. Mutations in this gene have been associated with susceptibility to generalized epilepsy with febrile seizures, type 5. Alternatively spliced transcript variants have been described for this gene, but their biological validity has not been determined. [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications.
Strand: + **Genomic Size:** 11425 **Exon Count:** 9 **Coding Exon Count:** 9

Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD	Microarray
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways
Other Names	Model Information	Methods			

Data last updated: 2009-10-08

− **Sequence and Links to Tools and Databases**

Genomic Sequence (chr1:1,950,768-1,962,192)			mRNA (may differ from genome)		Protein (452 aa)
Gene Sorter	Genome Browser	Protein FASTA	Table Schema	CGAP	Ensembl
Entrez Gene	ExonPrimer	GeneCards	GeneNetwork	Gepis Tissue	H-INV
HGNC	HPRD	Human Cortex Gene Expression	Jackson Lab	OMIM	PubMed
Stanford SOURCE	Treefam	UniProtKB	User annotations		

+ **Comments and Description Text from UniProtKB**

Press "+" in the title bar above to open this section.

+ **Genetic Association Studies of Complex Diseases and Disorders**

Press "+" in the title bar above to open this section.

+ **Comparative Toxicogenomics Database (CTD)**

Press "+" in the title bar above to open this section.

+ **Microarray Expression Data**

Press "+" in the title bar above to open this section.

+ **mRNA Secondary Structure of 3' and 5' UTRs**

Press "+" in the title bar above to open this section.

− **Protein Domain and Structure Information**

InterPro Domains: Graphical view of domain structure
IPR006028 - GABAA_rcpt
IPR008098 - GABAAd_rcpt
IPR006029 - Neu_channel_TM
IPR006202 - Neur_chan_lig_bd
IPR006201 - Neur_channel
IPR018000 - Neurotransmitter_ion_chnl_CS

Pfam Domains:
PF02931 - Neurotransmitter-gated ion-channel ligand binding domain
PF02932 - Neurotransmitter-gated ion-channel transmembrane region

SCOP Domains:
63712 - Nicotinic receptor ligand binding domain-like
90112 - Neurotransmitter-gated ion-channel transmembrane pore

ModBase Predicted Comparative 3D Structure on O14764

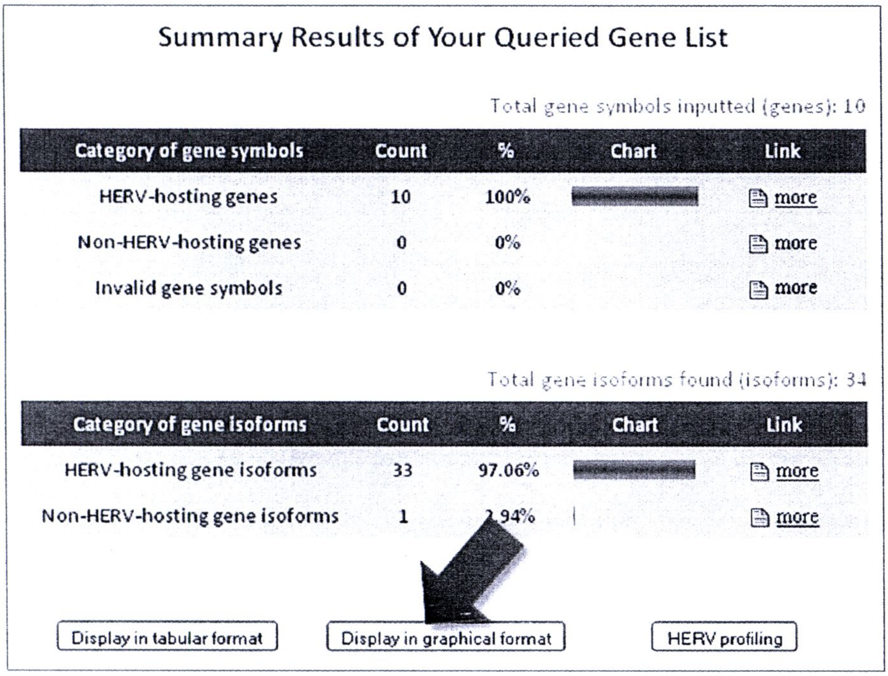
The pictures above may be empty if there is no ModBase structure for the protein. The ModBase structure frequently covers just a fragment of the protein. You may be asked to log onto ModBase the first time you click on the pictures. It is simplest after logging in to just click on the picture again to get to the specific info on that model.

Figure C.8 UCSC gene link

VISUALIZING NEIGHBORING HERVS IN GRAPHICAL FORMAT

1. Begin with specifying your gene list and HERVs that you would like to include into the system at the Search/Profiling page.

2. Choose the button "Display in graphical format" in the summary page (Figure C.9). The results would then be shown as illustrated in Figure C.10.



Summary Results of Your Queried Gene List

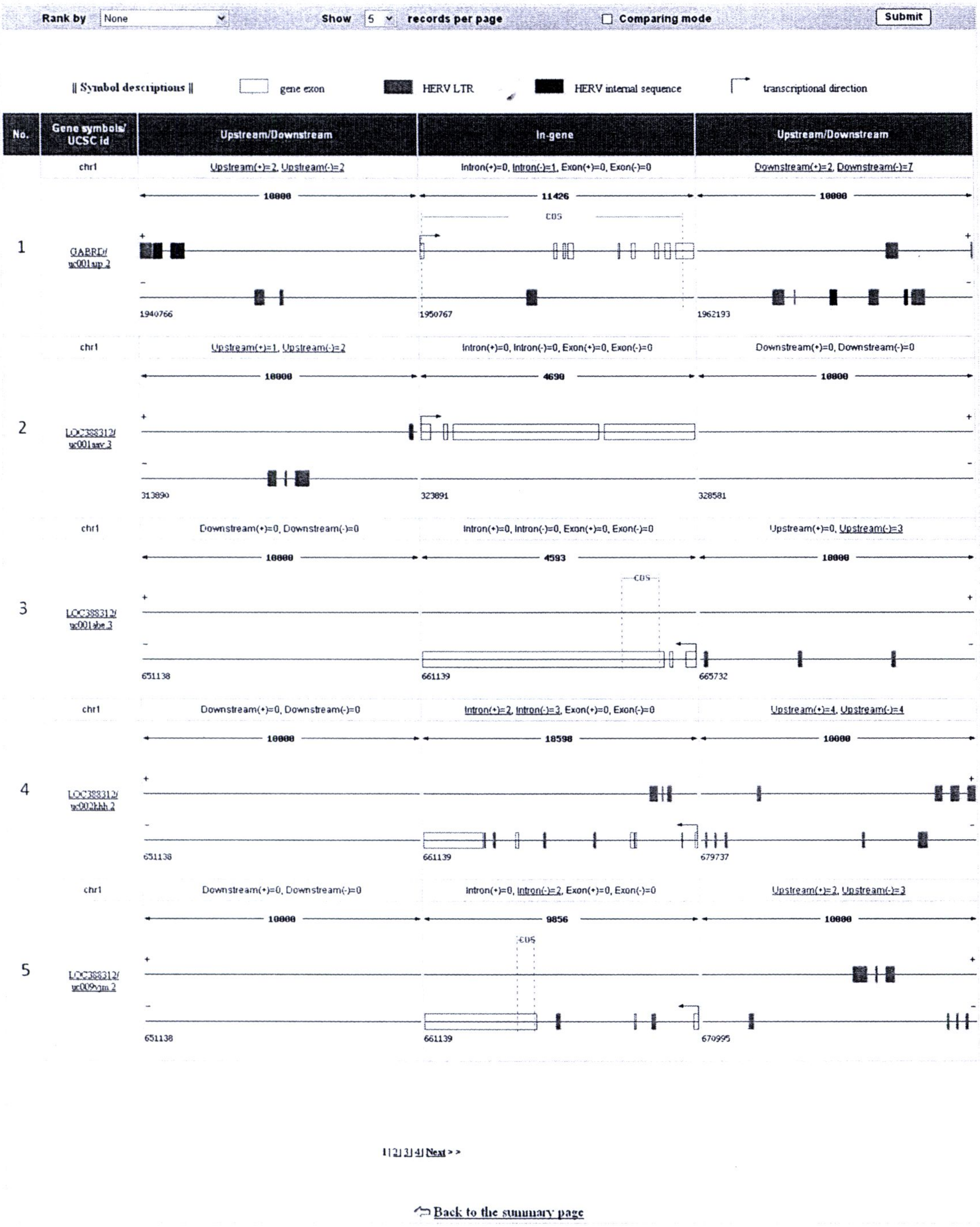
Total gene symbols inputted (genes): 10

Category of gene symbols	Count	%	Chart	Link
HERV-hosting genes	10	100%		more
Non-HERV-hosting genes	0	0%		more
Invalid gene symbols	0	0%		more

Total gene isoforms found (isoforms): 34

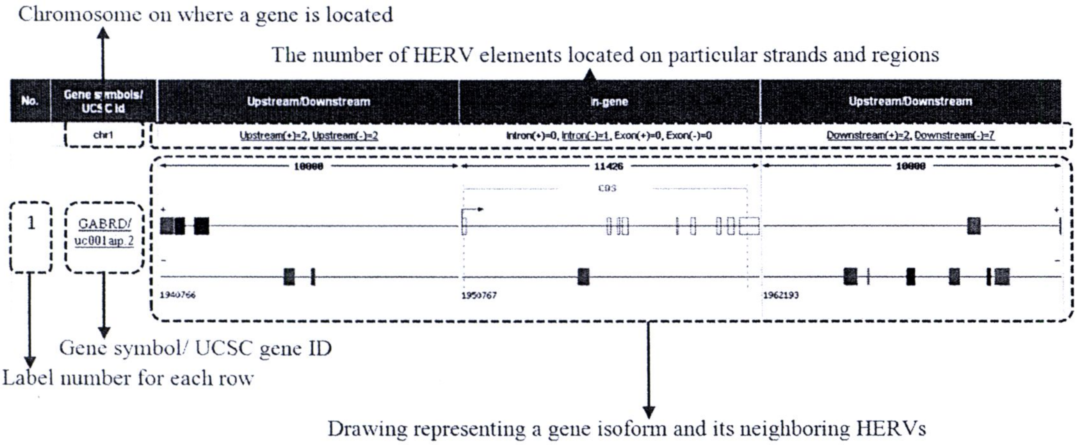
Category of gene isoforms	Count	%	Chart	Link
HERV-hosting gene isoforms	33	97.06%		more
Non-HERV-hosting gene isoforms	1	2.94%		more

Figure C.9 Location of the button of displaying in graphical format



No.	Gene symbols/ UCSC id	Upstream/Downstream	In-gene	Upstream/Downstream
	chr1	Upstream(+)=2, Upstream(-)=2	Intron(+)=0, Intron(-)=1, Exon(+)=0, Exon(-)=0	Downstream(+)=2, Downstream(-)=7
1	GABRD/ uc001up.2	10000 1940766	11426 1950767	10000 1962193
	chr1	Upstream(+)=1, Upstream(-)=2	Intron(+)=0, Intron(-)=0, Exon(+)=0, Exon(-)=0	Downstream(+)=0, Downstream(-)=0
2	LOC388312/ uc001asv.3	10000 313890	4690 323891	10000 328581
	chr1	Downstream(+)=0, Downstream(-)=0	Intron(+)=0, Intron(-)=0, Exon(+)=0, Exon(-)=0	Upstream(+)=0, Upstream(-)=3
3	LOC388312/ uc001abe.3	10000 651138	4593 661139	10000 665732
	chr1	Downstream(+)=0, Downstream(-)=0	Intron(+)=2, Intron(-)=3, Exon(+)=0, Exon(-)=0	Upstream(+)=4, Upstream(-)=4
4	LOC388312/ uc002khh.2	10000 651138	18598 661139	10000 679737
	chr1	Downstream(+)=0, Downstream(-)=0	Intron(+)=0, Intron(-)=2, Exon(+)=0, Exon(-)=0	Upstream(+)=2, Upstream(-)=3
5	LOC388312/ uc009vjm.2	10000 651138	9856 661139	10000 670995

Figure C.10 Results of the graphical displaying of neighboring HERVs



Chromosome on where a gene is located

The number of HERV elements located on particular strands and regions

Gene symbol/ UCSC gene ID

Label number for each row

Drawing representing a gene isoform and its neighboring HERVs

Figure C.11 Explanation of the components in the graphical display of a gene isoform

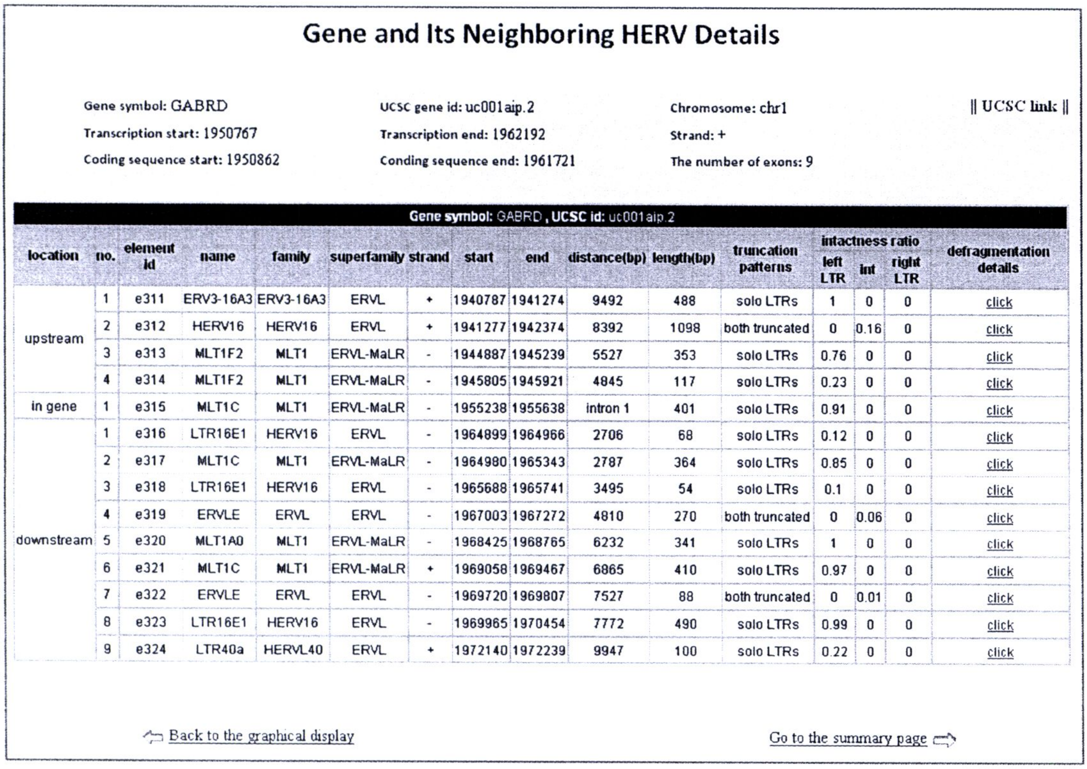
According to Figure C.10, each row, indicated by the label number, represents a drawing of one gene isoform. The description of each component in a row is shown in Figure C.11. A drawing is always divided into three regions, including upstream, in-gene, and downstream regions, relative to genes. The lengths in base pairs of each region are shown at the top of the drawing of the regions. The starting positions of each region are shown below the minus (-) strand at the bottom of the drawing. Protein-coding genes always have the labeling of CDS (coding sequences), while non-coding genes do not have it.

Like visualizing in the tabular format, you can also rank genes in a gene list following to the number of the HERVs at a particular location relative to genes or throughout the genes (see "Visualizing neighboring HERVs of interested genes" for more details about the ranking). Also, there is the option to limit the number of gene isoforms (tables) shown per page provided here.

As mentioned earlier, each drawing is divided into three regions relative to genes. Sometimes, these regions of all drawing are not in the same direction. In other words, in the same column of the result table, there can be the drawings of both the upstream regions and the downstream regions (Figure C.10). This leads to the difficulty when users need to compare across the drawings. Thus, there is an additional option provided here to draw all figures in the same direction. The option is called "comparing mode". You can create the drawings in the same direction by putting a tick in front of the terms "comparing mode" (Figure C.10). This would reflect the drawings of the genes located on the minus (-) strand resulting in the drawings in the same direction.

You can also go to the tabular result of a particular gene isoform by selecting on the gene symbol/UCSC gene ID (Figure C.10). An example of the linked table is shown in Figure C.12 (See "Visualizing neighboring HERVs of interested genes" for more descriptions of the result table). Furthermore, you can find out for more details of some HERVs in a particular regions and strands by clicking on the labels of the number of

HERV elements above the drawings (Figure C.11). This link would also lead you to the tabular result too.



Gene and Its Neighboring HERV Details

Gene symbol: GABRD
Transcription start: 1950767
Coding sequence start: 1950862

UCSC gene id: uc001aip.2
Transcription end: 1962192
Conding sequence end: 1961721

Chromosome: chr1
Strand: +
The number of exons: 9

|| UCSC link: ||

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

Back to the graphical display

Go to the summary page

Figure C.12 An example of the tabular result linked from the graphical displays

PROFILING NEIGHBORING HERVS
OF THE GENES IN A GENE LIST

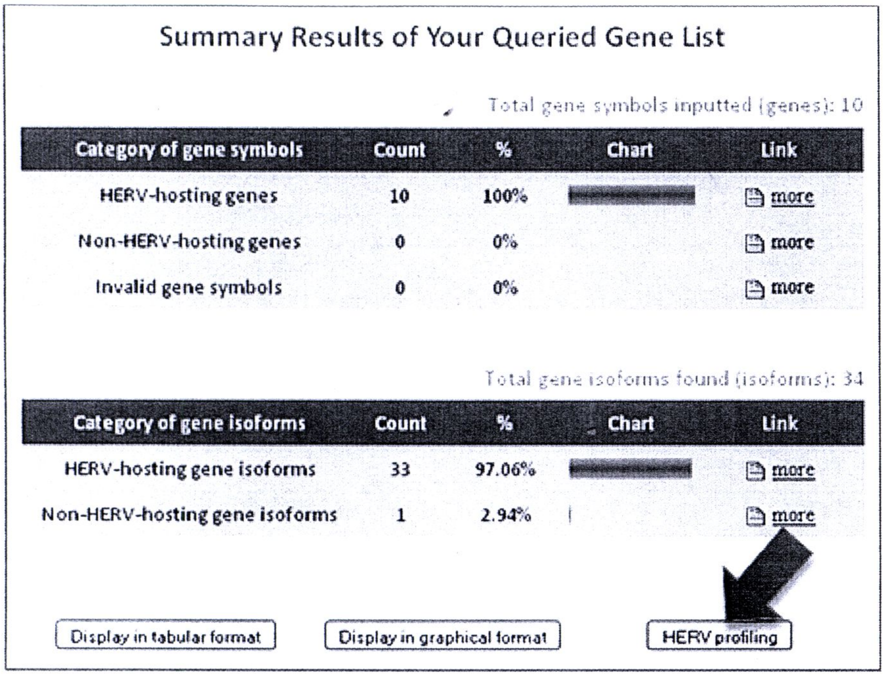
HERV profiling, or profiling for short, is another feature provided in HERV Profiler. This is a process to primarily observe shared HERV characteristics, so-called over-represented HERV types, present among a gene list. These over-represented HERV types are considered to be substantially present in this gene list more than a random chance by a statistical approach. Thus, in a group of differential expressed genes, those HERVs can serve as potential candidates which may influence their neighboring genes under a studied condition. In addition, HERV Profiler can purpose potential candidates of genes in the gene list by ranking them based on their HERV types. There are three steps to complete the profiling, including constructing HERV profiles, finding over-represented HERV types, and ranking genes in the gene list. In this section, each step of the profiling for a gene list would be illustrated.

CONSTRUCTING HERV PROFILES

This is a primary step which should be completed before doing other further processes of the HERV profiling. How to construct HERV profiles of a gene list is illustrated step by step below.

1. Begin with specifying your gene list and HERVs that you would like to include into the system at the Search/Profiling page.

2. Choose the button "HERV profiling" in the summary page (Figure C.13). The profile constructing page would be shown like in Figure C.14.



Summary Results of Your Queried Gene List

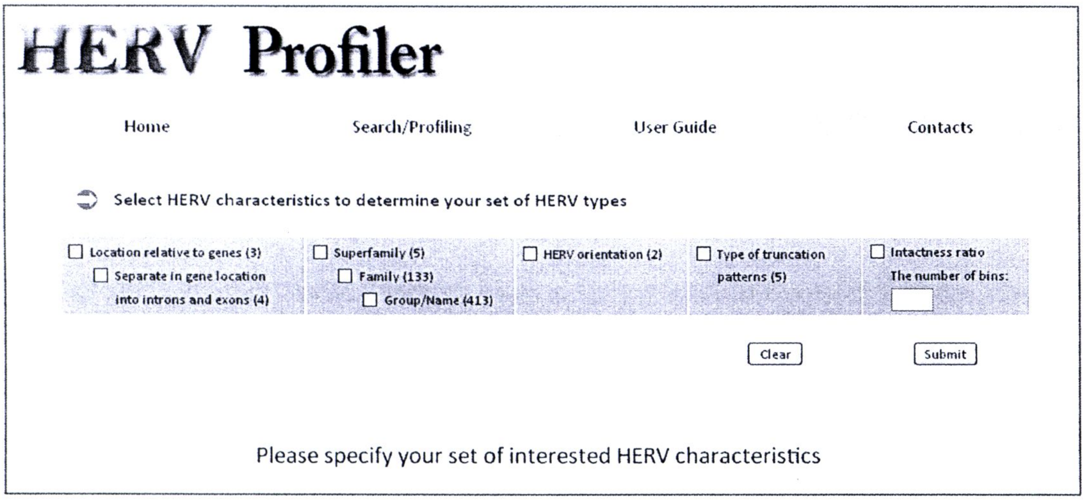
Total gene symbols inputted (genes): 10

Category of gene symbols	Count	%	Chart	Link
HERV-hosting genes	10	100%		more
Non-HERV-hosting genes	0	0%		more
Invalid gene symbols	0	0%		more

Total gene isoforms found (isoforms): 34

Category of gene isoforms	Count	%	Chart	Link
HERV-hosting gene isoforms	33	97.06%		more
Non-HERV-hosting gene isoforms	1	2.94%		more

Figure C.13 Location of the button of HERV profiling



Select HERV characteristics to determine your set of HERV types

Please specify your set of interested HERV characteristics

Figure C.14 Profile constructing page

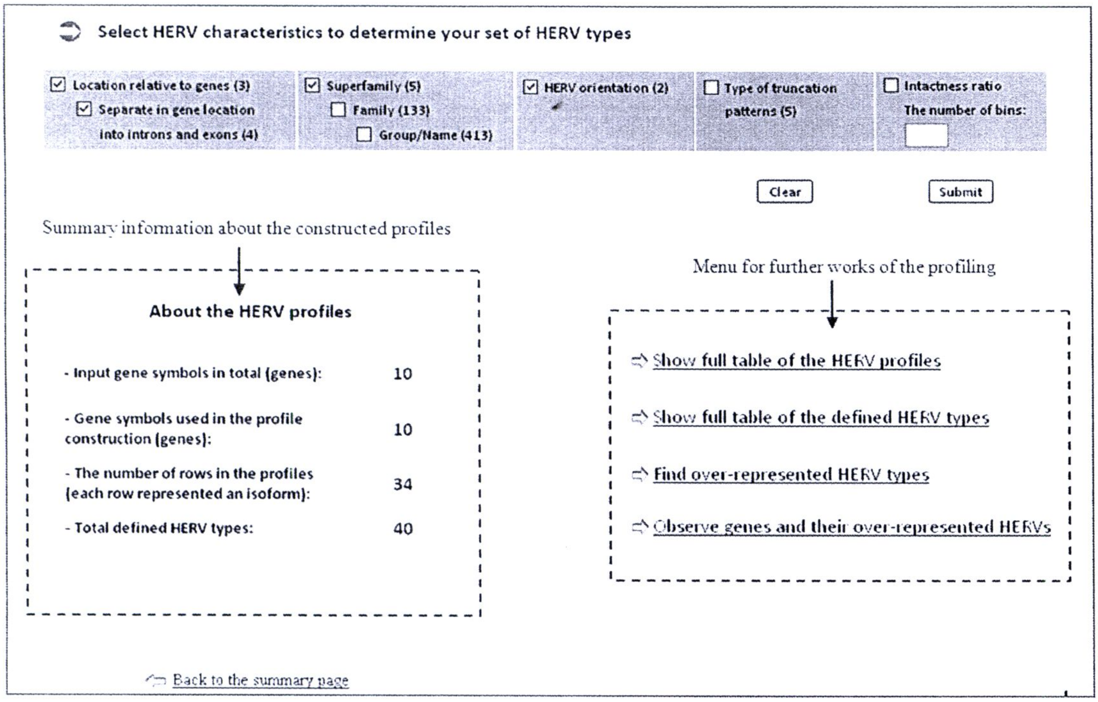
3. Select HERV characteristics to determine a set of HERV types from the profile constructing page (Figure C.14). A number in parentheses after a characteristic term indicate the number of possible categories of that characteristic. For example, "Location relative to genes" has three categories: upstream, in gene, and downstream. The all possible categories of other characteristics are listed in Table C.1.

The HERV types are the results of all possible combinations of the categories of the chosen characteristics. For example, if "Location relative to genes" and "HERV orientation" are selected, there will be six HERV types defined: sense-upstream, antisense-upstream, sense-in-gene, antisense-in-gene, sense-downstream, and antisense-downstream.

Table C.1 List of the categories of each HERV characteristics

HERV characteristics	**Description**	**The number of categories**	**Categories of the characteristics**
Location relative to genes	Location of HERVs relative to genes' regions	3	upstream, in gene, downstream
Separate in gene location into introns and exons	Location of HERVs relative to genes' regions with more specifying the regions in gene.	4	upstream, intron, exon, downstream
Superfamily	HERV superfamily	5	ERV1, ERVK, ERVL, ERVL-MaLR, unclassified HERVs
Family	HERV family	133	see Appendix A
Name/Group	HERV name or group	413	see Appendix A
HERV orientation	Direction of HERVs relative to transcription direction of a gene	2	same, opposite
Types of truncation patterns	The truncation types regarding to truncated part(s) of an HERV element	5	complete, 5'-truncated, 3'-truncated, both 5'- and 3'-truncated element, and solo LTRs
Intactness ratio	The ratio indicating how much an element is intact comparing to full length of a reference sequence	- (depend on the number of bins specified-default number of bins = 4)	List of the intervals separating the original interval lying on 0 and 1

4. After selecting the HERV characteristics to define a set of the HERV types, click the button "Submit" to begin constructing the HERV profiles. When it already finishes, the summary results of the profiles and the menu for further works would be provided as shown in Figure C.15. The descriptions of both the summary information about the profiles and the menu shown in this page are discussed in Table C.2 and Table C.3, respectively.



Select HERV characteristics to determine your set of HERV types

☑ Location relative to genes (3)
  ☑ Separate in gene location into introns and exons (4)

☑ Superfamily (5)
  ☐ Family (133)
    ☐ Group/Name (413)

☑ HERV orientation (2)

☐ Type of truncation patterns (5)

☐ Intactness ratio
The number of bins:

About the HERV profiles

- Input gene symbols in total (genes):	10
- Gene symbols used in the profile construction (genes):	10
- The number of rows in the profiles (each row represented an isoform):	34
- Total defined HERV types:	40

Figure C.15 Results shown after finishing constructing HERV profiles

Table C.2 Descriptions of terms shown in result summary of constructed profiles

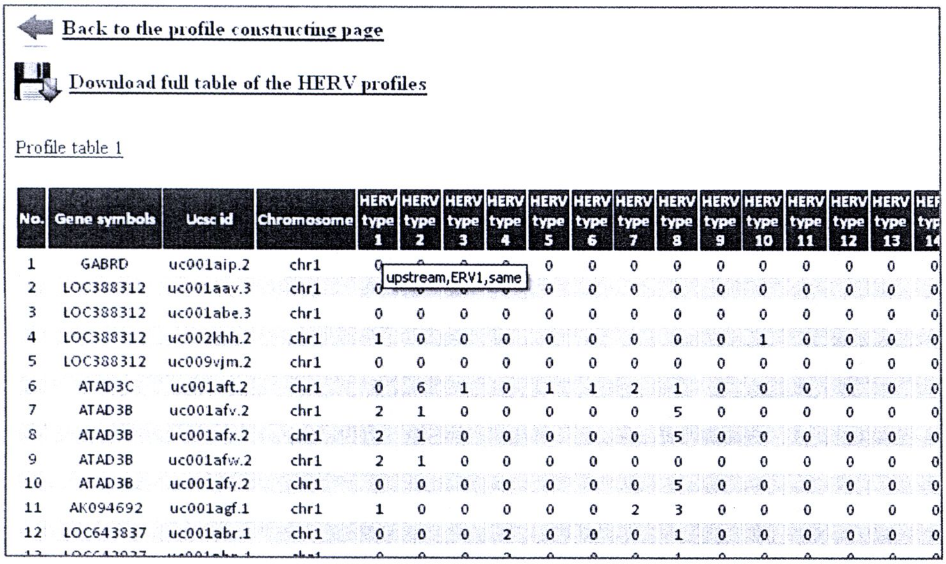
Terms	Descriptions
- Input gene symbols in total (genes)	The number of the gene symbols which are firstly input into HERV Profiler
- Gene symbols used in the profile construction (genes)	The number of the gene symbols which are included in constructing profiles (exclude only invalid gene symbols)
- The number of rows in the profiles (each row represented an isoform)	The number of rows in the profile table, which is equivalent to the number of isoforms, of the used gene symbols, found in HERV Profiler
- Total defined HERV types	The number of HERV types which are defined following to selected characteristics. This is equivalent to the multiplications of the possible categories of all selected characteristics.

Table C.3 Descriptions of terms shown in the menu after constructing profiles

Terms	Descriptions
Show full table of the HERV profiles	To show details of the HERV profiles constructed
Show full table of the defined HERV types	To show details of the HERV types defined
Find over-represented HERV types	To find over-represented HERV types among the gene list by using the previously constructed profiles
Observe genes and their over-represented HERV types	To rank genes in the gene list following to possessing of the over-represented HERVs

PROFILE DISPLAY AND RETRIEVAL

After already constructing HERV profiles, you can display the profiles in details by clicking "Show full table of the HERV profiles" in the page of profile constructing result (Figure C.15). An example of the full table of the profiles is shown in Figure C.16.



Back to the profile constructing page

Download full table of the HERV profiles

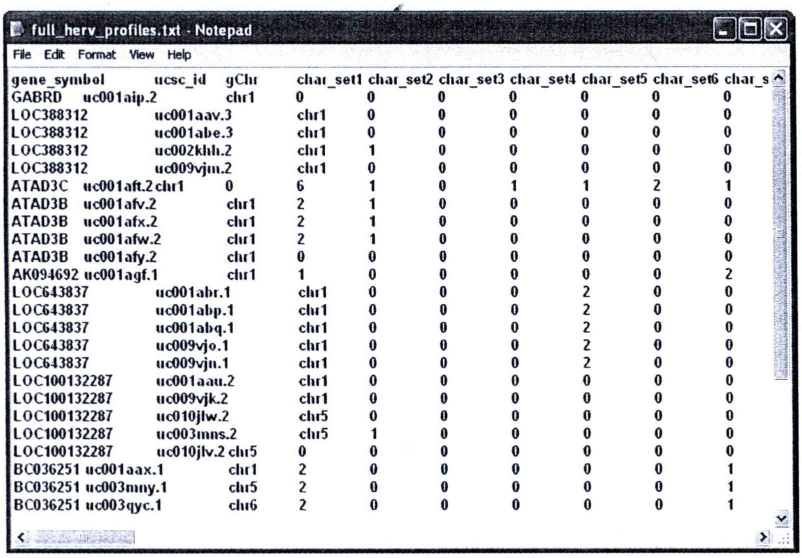
Profile table 1

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	HE typ 1
1	GABRD	uc001aip.2	chr1	0	[illegible]	[illegible]	[illegible]	0	0	0	0	0	0	0	0	0	0
2	LOC388312	uc001aav.3	chr1	0	[illegible]	[illegible]	[illegible]	0	0	0	0	0	0	0	0	0	0
3	LOC388312	uc001abe.3	chr1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	LOC388312	uc002khh.2	chr1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
5	LOC388312	uc009vjm.2	chr1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	ATAD3C	uc001aft.2	chr1	0	6	1	0	1	1	2	1	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
8	ATAD3B	uc001afx.2	chr1	2	1	0	0	0	0	0	5	0	0	0	0	0	0
9	ATAD3B	uc001afw.2	chr1	2	1	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
11	AK094692	uc001agf.1	chr1	1	0	0	0	0	0	2	3	0	0	0	0	0	0
12	LOC643837	uc001abr.1	chr1	0	0	0	2	0	0	0	1	0	0	0	0	0	0

upstream,ERV1,same

Figure C.16 Full table of the HERV profiles

According to Figure C.16, each row represents a profile of a gene isoform. The numbers in the table indicate the numbers of occurrences of an HERV type in a gene. Abbreviation information of each HERV types is shown when moving a mouse over the terms HERV type, as shown in Figure C.16. In case that the profile result contains more than 500 isoforms or rows, there will be more than one profile tables shown in this page, and each table contains no more than 500 rows only. You can freely go to any table by just selecting at the link of that profile table. Furthermore, you can export the profiles for utilizing or analyzing outside the tool by clicking on "Download full table of the HERV profiles" (Figure C.16). The profile tables would be saved as a text file where each filed is delimited by a tab character (Figure C.17).

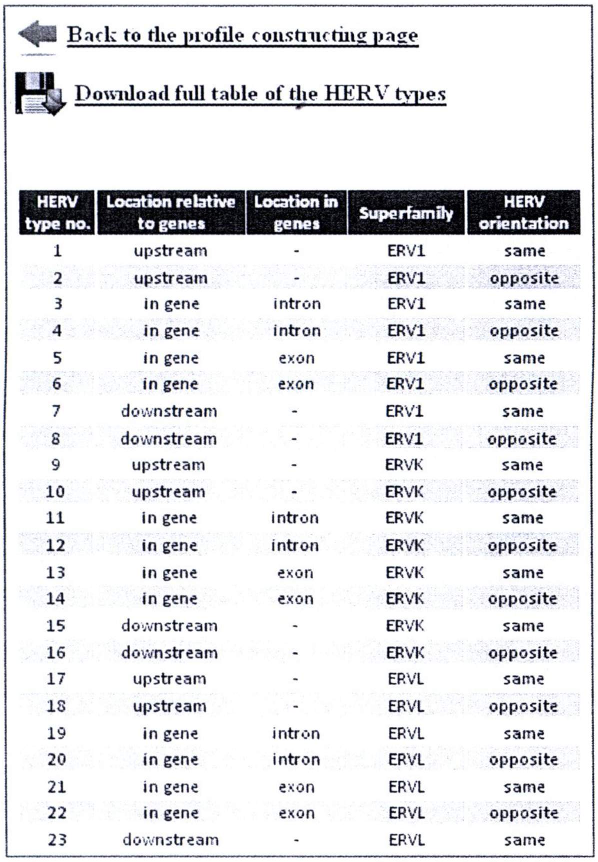


```
gene_symbol         ucsc_id   gChr      char_set1 char_set2 char_set3 char_set4 char_set5 char_set6 char_s
GABRD     uc001aip.2          chr1      0         0         0         0         0         0         0
LOC388312           uc001aav.3          chr1      0         0         0         0         0         0
LOC388312           uc001abe.3          chr1      0         0         0         0         0         0
LOC388312           uc002khh.2          chr1      1         0         0         0         0         0
LOC388312           uc009vjm.2          chr1      0         0         0         0         0         0
ATAD3C    uc001aft.2 chr1     0         6         1         0         1         1         2         1
ATAD3B    uc001afv.2          chr1      2         1         0         0         0         0         0
ATAD3B    uc001afx.2          chr1      2         1         0         0         0         0         0
ATAD3B    uc001afw.2          chr1      2         1         0         0         0         0         0
ATAD3B    uc001afy.2          chr1      0         0         0         0         0         0         0
AK094692 uc001agf.1           chr1      1         0         0         0         0         0         2
LOC643837           uc001abr.1          chr1      0         0         0         2         0         0
LOC643837           uc001abp.1          chr1      0         0         0         2         0         0
LOC643837           uc001abq.1          chr1      0         0         0         2         0         0
LOC643837           uc009vjo.1          chr1      0         0         0         2         0         0
LOC643837           uc009vjn.1          chr1      0         0         0         2         0         0
LOC100132287        uc001aau.2          chr1      0         0         0         0         0         0
LOC100132287        uc009vjk.2          chr1      0         0         0         0         0         0
LOC100132287        uc010jlw.2          chr5      0         0         0         0         0         0
LOC100132287        uc003mns.2          chr5      1         0         0         0         0         0
LOC100132287        uc010jlv.2 chr5     0         0         0         0         0         0         0
BC036251 uc001aax.1           chr1      2         0         0         0         0         0         1
BC036251 uc003mny.1           chr5      2         0         0         0         0         0         1
BC036251 uc003qyc.1           chr6      2         0         0         0         0         0         1
```

Figure C.17 An example file of the HERV profiles retrieved from the tool

HERV TYPE DISPLAY AND RETRIEVAL

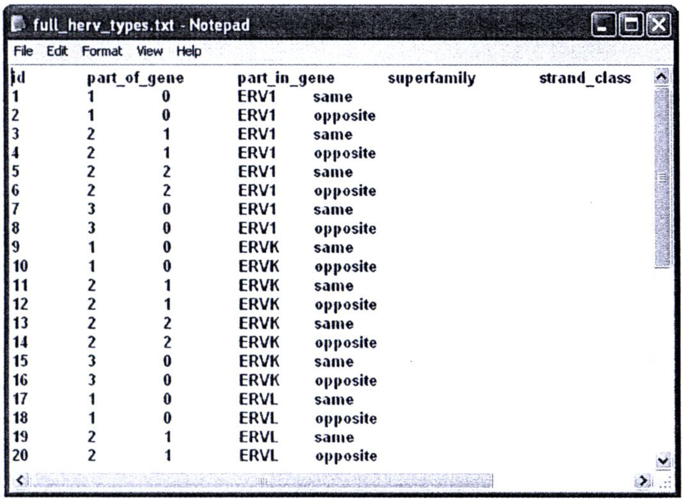
In addition to the full table of the HERV profiles, you can also look at the full table of the defined HERV types by clicking on "Show full table of the defined HERV types" in the page of profile constructing result (Figure C.15). An example of the tale of the HERV types is shown in Figure C.18.



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
11	in gene	intron	ERVK	same
12	in gene	intron	ERVK	opposite
13	in gene	exon	ERVK	same
14	in gene	exon	ERVK	opposite
15	downstream	-	ERVK	same
16	downstream	-	ERVK	opposite
17	upstream	-	ERVL	same
18	upstream	-	ERVL	opposite
19	in gene	intron	ERVL	same
20	in gene	intron	ERVL	opposite
21	in gene	exon	ERVL	same
22	in gene	exon	ERVL	opposite
23	downstream	-	ERVL	same

Figure C.18 Table of the defined HERV types

You can also download this table as a text file with tab delimits by clicking on "Download full table of the HERV types" (Figure C.18). An example of a file retrieved from the tool is illustrated in Figure C.19.

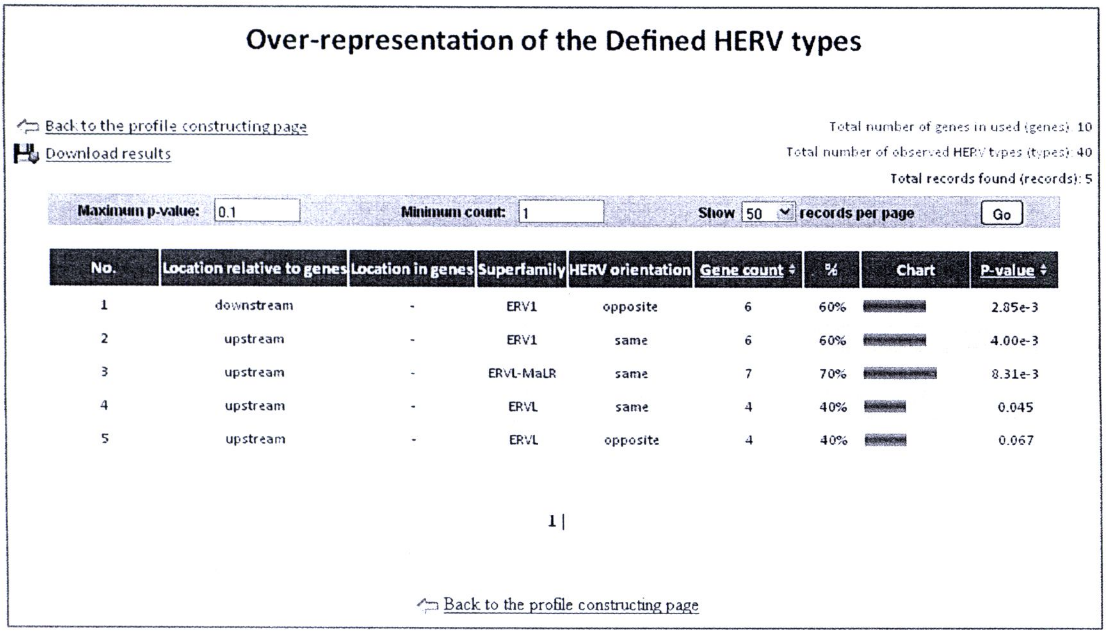


```
id      part_of_gene    part_in_gene    superfamily     strand_class
1       1       0       ERV1    same
2       1       0       ERV1    opposite
3       2       1       ERV1    same
4       2       1       ERV1    opposite
5       2       2       ERV1    same
6       2       2       ERV1    opposite
7       3       0       ERV1    same
8       3       0       ERV1    opposite
9       1       0       ERVK    same
10      1       0       ERVK    opposite
11      2       1       ERVK    same
12      2       1       ERVK    opposite
13      2       2       ERVK    same
14      2       2       ERVK    opposite
15      3       0       ERVK    same
16      3       0       ERVK    opposite
17      1       0       ERVL    same
18      1       0       ERVL    opposite
19      2       1       ERVL    same
20      2       1       ERVL    opposite
```

Figure C.19 An example file of the HERV types retrieved from the tool

FINDING OVER-REPRESENTED HERV TYPES

You can begin finding over-represented HERV types by clicking on "Find over-represented HERV types" in the page of profile constructing result (Figure C.15). HERV Profiler would perform the task for a while and then the results would be shown as illustrated in Figure C.20.



Over-representation of the Defined HERV types

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

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

Figure C.20 Results of finding over-represented HERV types among a gene list

According to Figure C.20, there are two parts of the results shown in this page. The first one written at the top right corner of this page is a result summary. Another one written in the table is a list of HERV types which are consistent with the criteria. There are two parameters and one option to limit the results shown in this page, including maximum p-value, minimum count, and the number of records shown per page. The descriptions of both the results and the limiting parameters and option are discussed in Tables C.4-C.6.

Table C.4 Descriptions of terms shown in the result summary

Terms	Descriptions
Total number of genes in used (genes)	The number of gene symbols used in constructing profiles (not include invalid gene symbols)
Total number of observed HERV types (types)	The number of HERV types previously defined in the profile construction
Total records found (records)	The number of HERV types shown in the result table

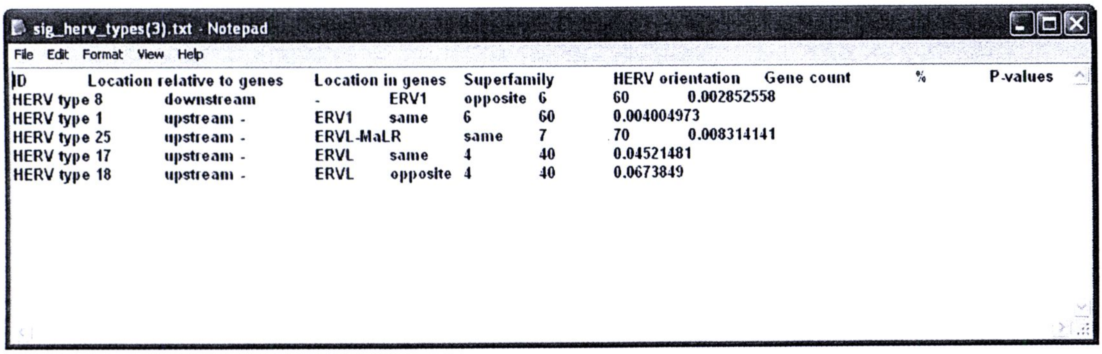
Table C.5 Descriptions of terms shown in the result table

Terms	Descriptions
No.	The integer number used to label rows or HERV types
Gene count	The number of genes which are related to a particular HERV type. The possibly lowest value is zero, while the possibly highest value should no more than the number of genes included into the system.
%	Percentages of the number of gene counts
Chart	Charts of the percentages. The width of this column represents the percentage of one hundred.
P-value	P-value of an HERV type obtained from Fisher's exact tests

Table C.6 Descriptions of limiting parameters and option

Limiting parameters and option	Descriptions
Maximum p-value	The highest p-value that can be acceptable. The input value should be between 0 and 1 (default value = 0.1).
Minimum count	The lowest gene count that can be acceptable. The input value should be a positive integer (default value = 1).
Show … records per page	The number of records shown in a page

In addition, you can rank the HERV types shown in the table following to two criteria available, including "gene count" and "p-value". To begin the ranking, you just click on either "Gene count" or "P-value" at the table headers. The result ranked by the gene count is always in a descending order, while the result ranked by the p-value is always in an ascending order. The default ranking system is following to the values of the p-values. Furthermore, you can export the results as a tab-delimited text file by clicking on "Download results". An example of a file is shown below (Figure C.21).

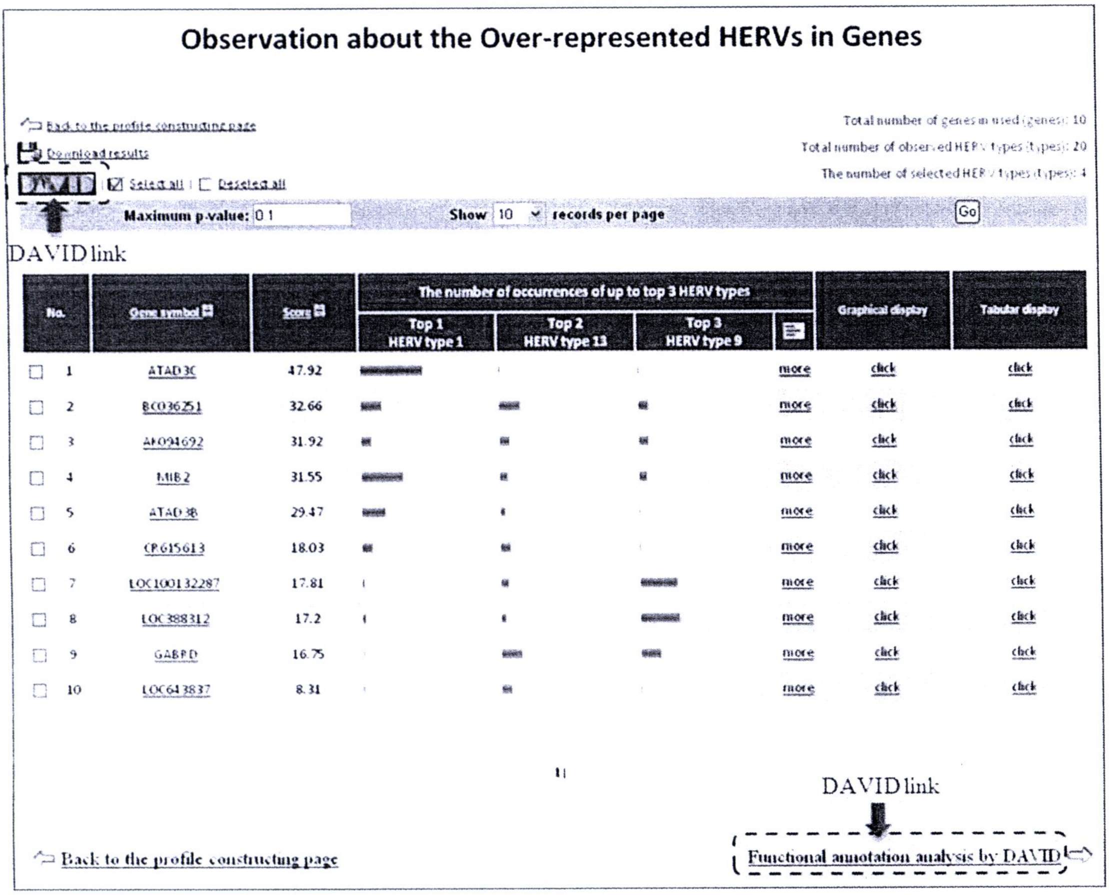


```
ID          Location relative to genes     Location in genes  Superfamily        HERV orientation   Gene count        %         P-values
HERV type 8         downstream             -          ERV1    opposite  6        60         0.002852558
HERV type 1         upstream  -            ERV1       same    6         60       0.004004973
HERV type 25        upstream  -            ERVL-MaLR          same      7        70         0.008314141
HERV type 17        upstream  -            ERVL       same    4         40       0.04521481
HERV type 18        upstream  -            ERVL       opposite 4        40       0.0673849
```

Figure C.21 An example file downloaded from the tool of the finding over-represented HERV types

RANKING GENES IN A GENE LIST FOLLOWING TO THEIR HERV TYPES

You can begin ranking the genes in a gene list by clicking on "Observe genes and their over-represented HERVs" in the page of profile constructing result (Figure C.15). After finishing the task, the results would be shown as illustrated in Figure C.22.



Total number of genes in used (genes): 10
Total number of observed HERV types (types): 20
The number of selected HERV types (types): 4

Maximum p-value: 0.1   Show 10 records per page

No.		Gene symbol	Score	The number of occurrences of up to top 3 HERV types				Graphical display	Tabular display
				Top 1 HERV type 1	Top 2 HERV type 13	Top 3 HERV type 9			
☐	1	ATAD3C	47.92				more	click	click
☐	2	BC036251	32.66				more	click	click
☐	3	AK094692	31.92				more	click	click
☐	4	MIB2	31.55				more	click	click
☐	5	ATAD3B	29.47				more	click	click
☐	6	CR615613	18.03				more	click	click
☐	7	LOC100132287	17.81				more	click	click
☐	8	LOC388312	17.2				more	click	click
☐	9	GABRD	16.75				more	click	click
☐	10	LOC643837	8.31				more	click	click

Figure C.22 Results of ranking genes following to possessing of the over-represented HERV types

According to Figure C.22, there are two parts of the results shown in this page. The first one written at the top right corner of this page is a result summary. Another one written in the table is a result of the ranking. Also, there are one parameter that is important to the ranking and one option to limit the results provided here. The descriptions of the results, the parameter, and the available option are discussed in Tables C.7-C.9.

Table C.7 Descriptions of terms shown in the result summary

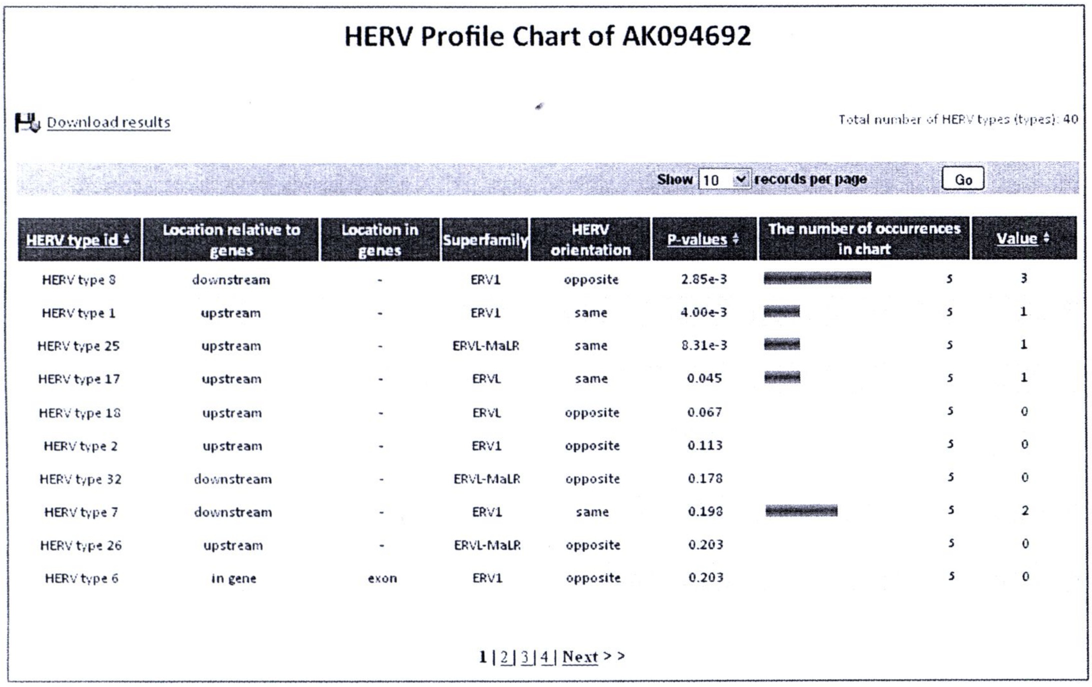
Terms	Descriptions
Total number of genes in used (genes)	The number of gene symbols used in constructing profiles (not include invalid gene symbols)
Total number of observed HERV types (types)	The number of HERV types previously defined in the profile construction
Total number of selected HERV types (types)	The number of HERV types included into the calculating scores of the genes in a gene list. This is specified by the value of the parameter maximum p-value (see Table C.9 for more details).

Table C.8 Descriptions of terms shown in the result table

Terms	Descriptions
No.	The integer number used to label rows or gene symbols.
Gene symbol	Gene symbols which are included into the system
Score	Scores of gene symbols indicating how much they are related to their over-represented HERV types. The score is the summation of the multiplications between the number of occurrences of over-represented HERV types and the scores of the over-represented HERV types.
The number of occurrences of up to top three HERV types	Charts representing the numbers of occurrences of up to top three HERV types.
Graphical display	Links for finding out a graphical display of a particular gene symbol
Tabular display	Links for finding out a tabular display of a particular gene symbol

Table C.9 Descriptions of the available parameter and option

Parameter and option	Descriptions
Maximum p-value	The highest p-value that can be acceptable for the HERV types which would be included to calculate scores for all genes in a gene list. The input value should be between 0 and 1 (default value = 0.1).
Show … records per page	The number of records shown in a page



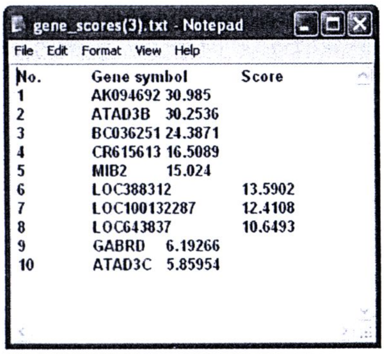
HERV Profile Chart of AK094692

Total number of HERV types (types): 40

HERV type id	Location relative to genes	Location in genes	Superfamily	HERV orientation	P-values	The number of occurrences in chart	Value
HERV type 8	downstream	-	ERV1	opposite	2.85e-3	5	3
HERV type 1	upstream	-	ERV1	same	4.00e-3	5	1
HERV type 25	upstream	-	ERVL-MaLR	same	8.31e-3	5	1
HERV type 17	upstream	-	ERVL	same	0.045	5	1
HERV type 18	upstream	-	ERVL	opposite	0.067	5	0
HERV type 2	upstream	-	ERV1	opposite	0.113	5	0
HERV type 32	downstream	-	ERVL-MaLR	opposite	0.178	5	0
HERV type 7	downstream	-	ERV1	same	0.198	5	2
HERV type 26	upstream	-	ERVL-MaLR	opposite	0.203	5	0
HERV type 6	in gene	exon	ERV1	opposite	0.203	5	0

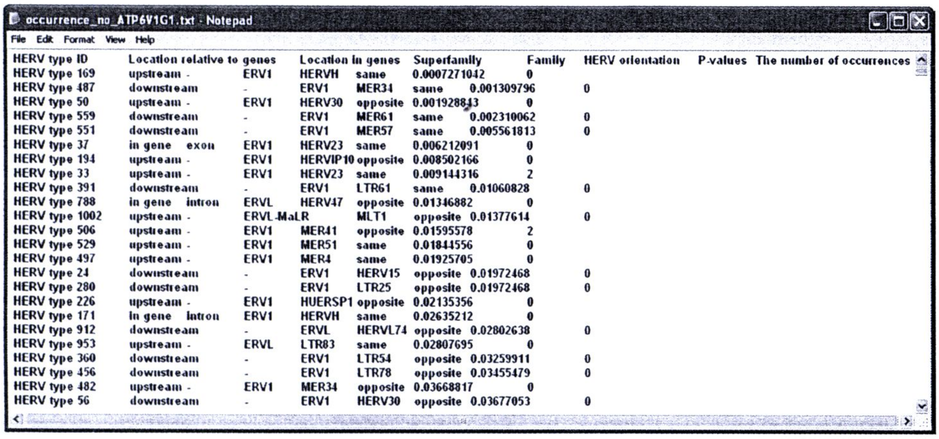
Figure C.23 Profile charts of a particular gene

According to Figure C.22, the numbers of occurrences of up to only top three HERV types are shown. However, the details of each gene can be observed more by clicking on "more" in the page of ranking results (Figure C.22). Then, the results of the number of occurrences would be shown as illustrated in Figure C.23 Furthermore, you can go further to find out the graphical or the tabular displays of the genes by selecting at "click" in the last two columns of the table. Also, the annotation details of each gene of the UCSC database can be found by clicking on the names of the desired gene symbols (See Figure C.8 for more details about the UCSC database). Also, the results of both the ranking and the profile charts of a certain gene can be downloaded as tab-delimited text files by clicking on "Download results". Examples of the files are shown in Figure C.24 and Figure C.25.



```
No.     Gene symbol     Score
1       AK094692 30.985
2       ATAD3B  30.2536
3       BC036251 24.3871
4       CR615613 16.5089
5       MIB2    15.024
6       LOC388312       13.5902
7       LOC100132287    12.4108
8       LOC643837       10.6493
9       GABRD   6.19266
10      ATAD3C  5.85954
```

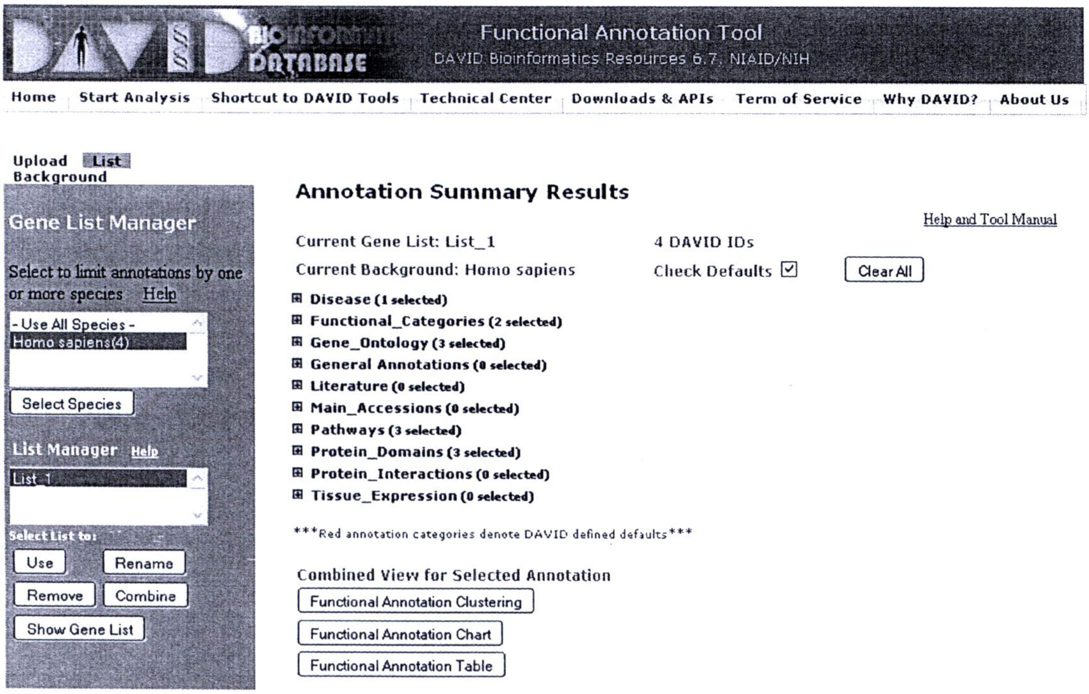
Figure C.24 An example file downloaded from the tool of ranking genes following to possessing over-represented HERV types



```
HERV type ID      Location relative to genes   Location in genes   Superfamily         Family   HERV orientation   P-values   The number of occurrences
HERV type 169     upstream  -      ERV1      HERVH      same      0.0007271042       0
HERV type 487     downstream       -         ERV1       MER34     same      0.001309796       0
HERV type 50      upstream  -      ERV1      HERV30     opposite  0.001928843        0
HERV type 559     downstream       -         ERV1       MER61     same      0.002310062       0
HERV type 551     downstream       -         ERV1       MER57     same      0.005561813       0
HERV type 37      in gene   exon   ERV1      HERV23     same      0.006212091        0
HERV type 194     upstream  -      ERV1      HERVIP10   opposite  0.008502166        0
HERV type 33      upstream  -      ERV1      HERV23     same      0.009144316        2
HERV type 391     downstream       -         ERV1       LTR61     same      0.01060828        0
HERV type 788     in gene   intron ERVL      HERV47     opposite  0.01346882         0
HERV type 1002    upstream  -      ERVL-MaLR            MLT1      opposite  0.01377614        0
HERV type 506     upstream  -      ERV1      MER41      opposite  0.01595578         2
HERV type 529     upstream  -      ERV1      MER51      same      0.01844556         0
HERV type 497     upstream  -      ERV1      MER4       same      0.01925705         0
HERV type 24      downstream       -         ERV1       HERV15    opposite  0.01972468        0
HERV type 280     downstream       -         ERV1       LTR25     opposite  0.01972468        0
HERV type 226     upstream  -      ERV1      HUERSP1    opposite  0.02135356         0
HERV type 171     in gene   intron ERV1      HERVH      same      0.02635212         0
HERV type 912     downstream       -         ERVL       HERVL74   opposite  0.02802638        0
HERV type 953     upstream  -      ERVL      LTR83      same      0.02807695         0
HERV type 360     downstream       -         ERV1       LTR54     opposite  0.03259911        0
HERV type 456     downstream       -         ERV1       LTR78     opposite  0.03455479        0
HERV type 482     upstream  -      ERV1      MER34      opposite  0.03668817         0
HERV type 56      downstream       -         ERV1       HERV30    opposite  0.03677053        0
```

Figure C.25 An example file of profile charts of a particular gene

Additionally, HERV Profiler also provides the links to a functional annotation analysis tool or DAVID (Database for Annotation, Visualization and Integrated Discovery) on this page (Figure C.22). To begin using DAVID, you have to select genes which you would like to include in the functional annotation analysis first by put ticks in front of your desired gene symbols. Then, select at the DAVID links to send a list of the selected gene symbols to the DAVID system. There are two DAVID links provided in HERV Profiler as illustrated in Figure C.22. After submitting the gene list to DAVID, DAVID would take for a while for running, and then the DAVID summary would be shown as illustrated in Figure C.26.



Select to limit annotations by one or more species

- Use All Species -
Homo sapiens(4)

List_1

Annotation Summary Results

Current Gene List: List_1 4 DAVID IDs

Current Background: Homo sapiens Check Defaults ☑

⊞ **Disease (1 selected)**
⊞ **Functional_Categories (2 selected)**
⊞ **Gene_Ontology (3 selected)**
⊞ **General Annotations (0 selected)**
⊞ **Literature (0 selected)**
⊞ **Main_Accessions (0 selected)**
⊞ **Pathways (3 selected)**
⊞ **Protein_Domains (3 selected)**
⊞ **Protein_Interactions (0 selected)**
⊞ **Tissue_Expression (0 selected)**

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

Figure C.26 Annotation summary results of DAVID

CURRICULUM VITAE

NAME Mr. Thitipong Kawichai

DATE OF BIRTH April 16, 1986

EDUCATIONAL RECORD

High school	High School Graduation Phitsanuloke Pityakom School, 2004
Bachelor's degree	Bachelor of Science (Mathematics) Chiang Mai University, 2008
Master's degree	Master of Science (Bioinformatics and Systems Biology) King Mongkut's University of Technology Thonburi, 2011

SCHOLARSHIPS

2004-Present	Full scholarship, in Bachelor's and Master's degree Development and Promotion of Science and Technology Talents Project (DPST), National Science and Technology Development Agency (NSTDA), Thailand
2009-2010	Full scholarship, in Master's degree King Mongkut's University of Technology Thonburi and National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand

PUBLICATION

1. Kawichai, T., Prom-on, S., and Meechai, A., 2011, "GENN-HERV: Gene-neighboring HERV database", **15th International Annual Symposium on Computational Science and Engineering (ANSCSE15)**, Bangkok University, Bangkok, Thailand, pp.403-408.

King Mongkut's University of Technology Thonburi
Agreement on Intellectual Property Rights Transfer for Postgraduate Students

Date......................

Name...Mr. Thitipong ...Surname...Kawichai...Student Number...52401702...who is a student of King's Mongkut's University of Technology Thonburi (KMUTT) in.... Master's Degree...Program..Bioinformatics and Systems Biology....Field of Study..... Bioinformatics and Systems Biology..Faculty/School of..Bioresources and Technology and School of Information Technology.Home Address.406/3086-10.Tambon.Arunyig. District...Mueang Phitsanulok....Province...Phitsanulok ...Postal Code....65000.... Country...Thailand...

I, as 'Transferer', hereby transfer the ownership of my thesis copyright to King's Mongkut's University of Technology Thonburi and National Center for Genetic Engineering and Biotechnology, Thailand who has appointed (Dean's name) Assoc. Prof. Narumon Jeyashoke and Assoc. Prof. Dr. Nipon Charoenkitkarn, Dean of School of Bioresources and Technology and Dean of School of Information Technology to be 'Transferees' of copyright ownership under the 'Agreement' as follows.

1. I am the author of the thesis entitled "Development of a web-based tool for visualizing and profiling human endogenous retroviruses (HERVs) neighboring genes" under the supervision of Dr. Santitham Prom-on who is my supervisor, and Asst. Prof. Dr. Asawin Meechai who is my co-supervisor, in accordance with the Thai Copyright Act B.E. 2537. The thesis is a part of the curriculum of KMUTT.
2. I hereby transfer the copyright ownership of all my works in the thesis to KMUTT and National Center for Genetic Engineering and Biotechnology, Thailand throughout the copyright protection period in accordance with the Thai Copyright Act B.E. 2537, effective on the approval date of thesis proposal consented by KMUTT.
3. To have the thesis distributed in any form of media, I shall in each and every case stipulate the thesis as the work of KMUTT and National Center for Genetic Engineering and Biotechnology, Thailand.
4. For my own distribution of thesis or the reproduction, adjustment, or distribution of thesis by the third party in accordance with the Thai Copyright Act B.E. 2537 with remuneration in return, I am subject to obtain a prior written permission from KMUTT and National Center for Genetic Engineering and Biotechnology, Thailand.
5. To use any information from my thesis to make an invention or create any intellectual property works within ten (10) years from the date of signing this Agreement, I am subject to obtain prior written permission from KMUTT and National Center for Genetic Engineering and Biotechnology, and KMUTT and National Center for Genetic Engineering and Biotechnology are entitled to have intellectual property rights on such inventions or intellectual property works, including entitling to take royalty from licensing together with the distribution of any benefit deriving partly or wholly from the works in the future, conforming with the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.
6. If the benefits arise from my thesis or my intellectual property works owned by KMUTT and National Center for Genetic Engineering and Biotechnology, I shall be entitled to gain the benefits according to the allocation rate stated in the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.

Signature..Transferor
(Mr. Thitipong Kawichai)

Signature...Narumon Jeyashoke...Transferee
(Assoc. Prof. Narumon Jeyashoke)

Signature..Transferee
(Assoc. Prof. Dr. Nipon Charoenkitkarn)

Signature..Witness
(Dr. Santitham Prom-on)

Signature..Witness
(Asst. Prof. Dr. Marasri Ruengjitchatchawalya)