CHAPTER 2 BACKGROUNDS AND LITURATURE REVIEWS

2.1 Human endogenous retroviruses (HERVs)

2.1.1 What is HERV

Typically, endogenous retroviruses (ERVs) are termed for DNA sequences within the genome that are similar to sequences of infectious retroviruses. They likely represent the remnants of ancient infections that became incorporated in the germ lines [2]. This resulted that the retroviral sequences integrated into the genome, so-called proviruses, could be inherited from generation to generation without the infections. In other words, they are permanently fixed and present in the host genome. The endogenous retroviruses can be found in humans, mammals, and other vertebrates [12]. Thus, human ERVs (HERVs) are generally referred to the endogenous retroviruses found in the human.

HERVs constitute approximately 8% of the human genome, which is significantly substantial when compared to protein-coding genes constituting around only 3% of the human genome [1, 13]. This resulted from retrotransposition, amplifying themselves in a genome via RNA intermediates, induced when they were highly active. According to the transposition ability, HERVs are thus included as a member of transposable elements.

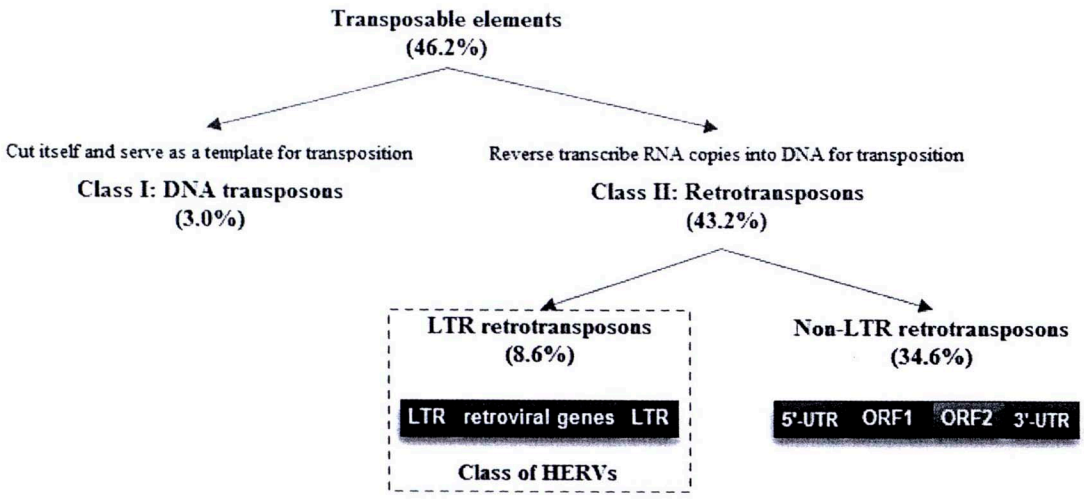


Figure 2.1 Diagram showing classification of the transposable elements and their fractions of the human genome (adopted from [3])

Generally, the transposable elements are DNA sequences that are able to move around and integrate into new sites within the genome [14]. As shown in Figure 2.1, the transposons can be separated into two main classes, including DNA transposons and retrotransposons. Retrotransposons require to be transcribed before and use those RNAs as the intermediates in the transposition, while DNA transposons employ themselves as the intermediates. In case of retrotransposons, they can be further classified into two

different classes, including LTR and non-LTR retrotransposons, according to possessing of the LTRs in their sequences (Figure 2.1). Instead of LTRs, non-LTR retrotransposons contain 5'UTR and 3'-UTR to flank the internal sequences. HERVs are a member of LTR retrotransposons. Moreover, most of the LTR retrotransposons are HERVs, because there is only 0.6% remaining which is other LTR retrotransposons, not a member of HERVs.

2.1.2 Genomic structure of HERVs

Normally, the proviruses, the retroviral sequences initially integrated into the host genome, are composed of two flanking LTRs (5'-LTRs and 3'-LTRs) and a set of viral genes (Figure 2.2a). There are at least three genes in the proviruses: *gag* encoding the structural proteins of the viral core; *pol* encoding the reverse transcriptases; and *env* encoding the surface glycoproteins of the viral envelope. The expression of the retroviral proteins is controlled by several regulatory elements in the long terminal repeats (LTRs), such as promoters, enhancers, and polyadenylation signals. Other regulatory sequences are also present in the viral genome, including the site of splice donor (SD) and splice acceptor (SA) for the *env* expression and a primer-binding site (PBS) for a complementary to a host transfer RNA (tRNA) to initiate the reverse transcription. In general, the provirus is about 7-11 kb in length [1, 5].

In case of the HERVs, their structures are similar to the structure of the proviruses but typically accumulate many mutations, including point mutations (dark bands), frameshifts and deletions (particularly in *env*), as shown in Figure 2.2b. The entire central region has been frequently removed by the recombination or deletions, finally leaving the solitary LTRs behind. Although most of the HERVs are defective, the LTRs may still be active, and transcription of a few HERVs is still occurred, particularly in fetal tissue and in some certain diseases, such as autoimmune diseases and cancer [1, 5, 15].

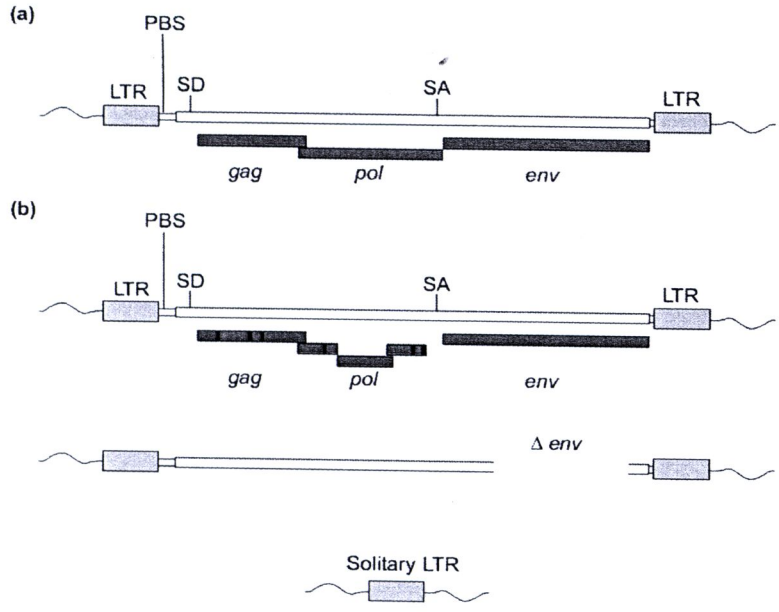


Figure 2.2 Genomic structures of retroviral proviruses (a) and HERVs (b) [1]

2.1.3 Classification of HERVs

HERVs have been usually classified into families and superfamilies, sometimes also sub-families, and those names have been referred to in the studies since the discovery of the HERVs. Nevertheless, those names previously designated could lead to considerable confusion, not just to the outsider, because the HERVs have been arbitrarily categorized and named following to manifold criteria arising from independent investigators [13]. In other words, there is inconsistency of naming and classifying for the same sequences. For example, the human DNA sequences, isolated by Callahan et al., similar to the mouse mammary tumor virus (MMTV) were named as HML-2. Subsequently, the same sequences were reported by Ono et al. and then named differently as HERV-K10 instead [16]. Some central systems would be then described in this section.

Formerly, the specific types of the tRNAs which complement to the primer binding sites (Figure 2.2) has been considered to name and classify the HERVs. For example, the members of HERV-H family contain the primer binding sites for histidine-tRNAs, and the elements in HERV-K family have the primer binding sites for Lysine-tRNAs. However, this method is still unreliable because there are some related HERVs displaying differences in terms of the primer binding sites, and otherwise some unrelated HERVs having the same type of the primer binding sites [5].

Another classification system is Repbase [17], a widely used repository of the repetitive elements. This nomenclature is based on nucleotide identity to the consensus sequences of the repeats, including HERVs, which are computationally generated. Due to a number of defective ERVs found in the human, LTRs and internal sequences of the

HERVs have been named and classified separately. Furthermore, Repbase is somewhat useful because it also contains all known alternative names of the repeats [16].

In addition, HERVs have been classified based on the phylogenetic criteria, comparing to the infectious retroviruses. The *pol* genes, the most conserved gene among the retroviruses, and *env* genes of the HERVs were used to conduct the classification of the HERVs recently [18]. This method seems to be more useful for the classification of the HERVs [5]. However, the comprehensive results are being established today.

HERV families found have been quite different in numbers, from a few to a thousand elements. For example, at least 30 HERV families were identified based on the phylogenetic approach, while more than 200 different HERV and LTR families have been mentioned in Repbase [18]. However, it is now generally accepted that HERV groups could be loosely classified into three broad classes, including class I, II, and III, based on sequence similarity to different genera of the infectious retroviruses [1, 3]. Class I, also called ERV1 superfamily, contains the HERVs related to gammaretroviruses such as murine leukemia virus (MLV) and baboon endogenous virus (BaEV). The HERVs in Class II, so-called ERVK superfamily, are related to betaretroviruses, including mouse mammry tumor virus (MMTV). Lastly, Class III HERVs, also termed ERVL superfamily, are distantly related to spumaretroviruses [1, 13]. Besides those three superfamilies, the mammalian apparent LTR-retrotransposons (MaLRs) are sometimes considered as an additional class of the HERVs, because the MaLR elements are all derived from the class III ERVs [19].

The divergence of the LTR sequences in the HERVs can be measured to estimate the age of the HERVs, given that the LTRs are identical at the time of integration [20]. Class I and III HERVs are the oldest groups and are currently present throughout the primate lineage, while class II includes the most recently integrated ERVs. A few proviruses in the HERV-K (HML-2) family are human-specific, indicating that these viruses have been active only within the last five million years [1].

2.1.4 Biological functions of HERVs

Since the HERVs were first discovered, there are a number of studies have been done with the effort to reveal the roles and effects of the HERVs. According to those studies, it could divide the functions of the HERVs into two categories: the cellular function and the genomic function. The cellular functions are related to their retained ability in expression of a few HERVs. For the genomic functions, these are related to their presence of potential regulatory sequences which may affect the functions of nearby genes.

2.1.4.1 Cellular functions

HERVs have been accumulated a number of mutations along the time resulting that most of them are incapable of being expressed. Nevertheless, there are a few HERVs still retain their ability to express the viral genes resulting in findings of retroviral

RNAs, proteins, and retroviral-like particles in several human tissues, whether normal or disease tissues [21].

The retroviral products can be beneficial for the humans. The best example is about the physiological roles of some HERVs in the host. HERV-W and HERV-FRD envelope (*env*) proteins are demonstrable, so called syncitin-1 and syncitin-2, respectively. These proteins have been highly found during the formation of the placenta, and suggested that they are responsible for mediating cell-to-cell fusion during the formation of the placental membranes [15, 22].

In contrast, the products of some ERVs can contribute the detrimental effects to the host as well. For example, the envelope proteins of some ERVs include an immunosuppressive domain. In mouse model, it has been found that the *env* proteins with this domain can promote tumor growth by allowing escape from immune surveillance [5]. In the humans, the retroviral expression has been detected in numerous patients suffering from various diseases, such as cancer, autoimmune diseases, and neurological diseases. However, it is not certainly known whether the HERVs cause the diseases or are just induced to express under the disease conditions.

2.1.4.2 Genomic functions

Besides a few of HERV genes retained, the parts of regulatory sequences of some HERVs have been reported currently active as well. The active regulatory sequences of the HERVs could affect the expression of the neighboring genes in several ways based on the active parts of the regulatory sequences as well as the placement of insertions (Figure 2.3).

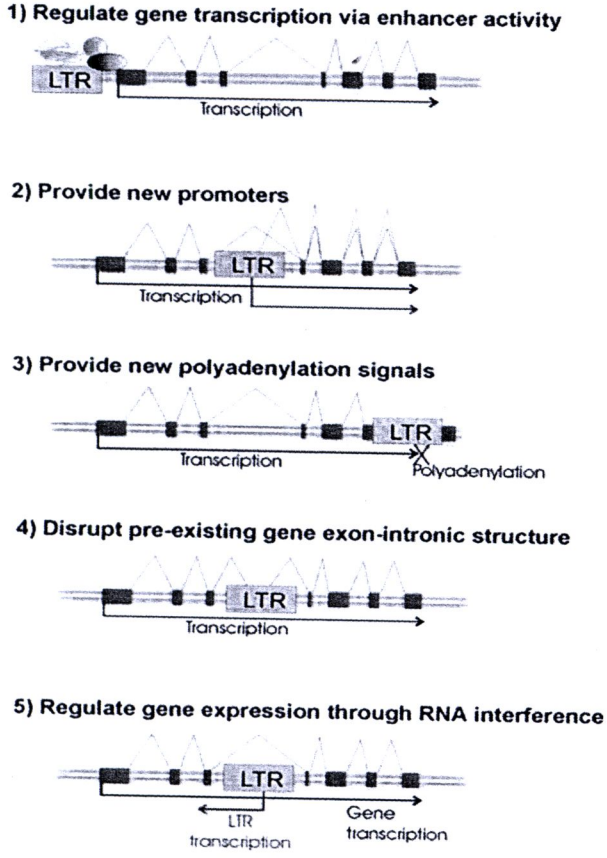


Figure 2.3 Five potential mechanisms of the HERVs for modulating the expression of the neighboring genes [23]. An HERV element is indicated an LTR box in the figure.

As mentioned before, most regulatory sequences of the HERVs, including promoters, enhancers, and polyadenylation signals, are located in the LTRs. Thus, the HERVs could affect the neighboring genes by providing enhancing, promoting, or terminating activities to the neighboring genes (Figure 2.3). In addition, the HERVs could change the patterns of the neighboring genes' transcripts by providing the additional splice sites. This may result in an introduction of new exons included in the transcripts, termed exonization process [15]. Furthermore, if the HERVs are located in gene introns in the antisense orientation, it could be possible that they would involve in antisense regulation of the pre-existing genes. This mechanism is based on the formation of the double-stranded RNA, followed by catalytic degradation of RNAs containing the sites homologous to the double-stranded fragments [23].

Like the case of the cellular functions, the genomic functions can be beneficial and detrimental. In some cases the HERV regulatory sequences are naturally co-opted like being a part of a host genome and in some cases the HERV regulation is abnormal and may cause diseases. The example genes which have been reported related to the HERV regulatory sequences are listed in Table 2.1.

Table 2.1 List of example human genes affected by HERV regulatory sequences

HERV regulator	HERV name	Gene name	Reference
1. Enhancer	ERV9 LTR	β-globin locus	[24]
	HERV-E	Amy1 (salivary amylase)	[5]
2. Promoter	HERV-L LTR	β1,3-galactosyltransferase 5	[25]
	HERV-E	APOCI (apolipoprotein CI)	[6]
	HERV-H LTR	DSCR4 and DSCR8 (Down syndrome critical region)	[26]
	HERV-E	EDNRB (endothelin receptor B)	[6]
	HERV-H	NAIP (neuronal apoptosis inhibitory protein)	[5]
	ERV9 LTR	ZNF80 (zinc finger protein)	[27]
3. Polyadenylation signals	HERV-K (KML2) LTR	LEPR (human leptin receptor)	[5]
4. Splice sites	HERV-H	PLA2L (phospholipase A2-like)	[28]
5. Antisense regulators	LTR91	CEBZ	[23]

2.2 Evidences linking HERVs to SLE

Systemic Lupus Erythematosus or SLE is an autoimmune disease that can affect multiple organs. It can be fatal if there are severe inflammations found in some organs, including kidneys, lungs, heart, as well as central nervous system [29]. SLE is the disease that can be found worldwide and in patients with every age. The prevalence of SLE varies from approximately 40 cases per 100,000 persons among Northern Europeans to more than 200 per 100,000 persons among blacks. 90% of the cases occur in women, especially between 15 and 50 years of age. This is an important one disease because it is a potentially fatal disease that is easily confused with many other disorders [30]. The etiopathogenesis of SLE remains partially understood. However, with the evidences accumulated over the last half century, it can be concluded that SLE is complex multifactorial disease, including genetic predisposition, environmental, as well as retroviral factors [31].

Actually, autoimmune diseases, including SLE, have been initially linked to retroviruses owing to the similarity of immune dysregulation and autoimmune manifestations between patients with SLE and known human retrovirus-related disorders, such as HIV-1 [32]. The important one evidence supporting the association between SLE and HERVs is the detection of antibodies reactive to several retroviral proteins, including *gag*, *env*, *nef*, and the p24 capsid of human immunodeficiency virus (HIV)-1 and human T cell leukemia/lymphoma virus (HTLV) in SLE patients with no history of prior infection [33]. This phenomenon was attributed to the induction by encoded retroviral proteins. Strikingly, it is found that these proteins have amino acid sequences similar to many self-nuclear antigens, such as U1 small nuclear ribonucleoprotein (70K U1 sn-RNP), topoisomerase I, and SS-B/La [32]. A given obvious example is the finding that as many as 52% of SLE patients possess circulating antibodies to HRES-1. Furthermore, from comparative sequence analysis, it was shown that there is sequence homology between HRES-1 and the 70-kDa *gag*-related region of the sn-RNP. In

respect to these findings, molecular mimicry between self-antigens and retroviral proteins is purposed as one possible mechanism in etiopathogenesis of SLE by inducing the cross-reaction between the two proteins by autoantibodies.

One important HERV linked to SLE is HERV clone 4-1, a member of the HERV-E family, which is usually found in Japanese people. It has been reported that there is no transcription and translation of HERV clone 4-1 in peripheral blood lymphocytes (PBL) of normal individuals, whereas, in SLE patients, the *gag* region antigen and mRNA for the clone 4-1 *gag* region have been detected in PBL. The transcription of this HERV can be controlled by epigenetic mechanisms [34]. Moreover, there is one study supporting that *env*-encoded transmembrane proteins from HERV, such as p15E, could induce immune dysregulation. The study observed the mechanisms of a synthetic peptide derived from HERV clone 4-1, CKS-17, which was homologue sequence with p15E. The results from this study showed that the peptide could induce T-cell activation and anergy in normal peripheral blood mononuclear cells (PBMCs), and promote the production of interleukins IL-6 and IL-16. This phenomenon is representative immune abnormalities of SLE [35].

As a consequence from retroviral integration, the HERV LTRs can act as *cis*-regulatory sequences causing cellular activation, particularly genes involved in immune regulation. Using MRL/*lpr* mice, a murine model for SLE, the study has revealed that there is an integration of an early transposable element (ETn) in the murine *Fas* apoptosis-promoting gene. This integration results in decreased synthesis of active *Fas* proteins, and undoubtedly the failure of apoptosis in autoreactive lymphocytes, which is a primary mechanism of SLE development [36]. Furthermore, many HERVs, as well as other retrotransposons, are found within the major histocompatibility complex (MHC) genes and human complement genes. Particularly, the integration of HERVs in the MHC class I is very interesting, since there are several polymorphic genes associated with susceptibility of autoimmune diseases in that region [37]. Thus, it is suggested that the regulation mediated by HERV LTRs may also influence the expression of the MHC genes in SLE patients. In addition to the *cis*-acting roles of HERVs, they have been purposed that could *trans*-activate cellular genes, since some HERVs can encode products like *Tat* in HIV-1 or *Tax* in HTLV-1, which can act as transactivators of cellular genes. However, there is currently no definitive evidence that can proof this hypothesis [33].

Interestingly, several exogenous factors, including chemicals and UV light, are also recruited as one supporting factor that could induce immune abnormalities induced by HERVs in SLE patients (Figure 2.7). For example, a study using DNA methylation inhibitors, such as 5-aza-deoxycytidine (5-aza C), has revealed that there is significant negative correlation between the increase of HERV clone 4-1 mRNA and the decrease of DNA methyltransferase (DNMT-1) mRNA in 5-aza C-treated normal PBL. This can be implied that the level of DNA methylation may mediate the expression of HERV clone 4-1, and may also be implicated in the development of SLE [34]. In addition,

ultraviolet B (UVB) irradiation has been reported as another one factor that can activate transcription of several HERV sequences in skin biopsies of SLE patients [38].

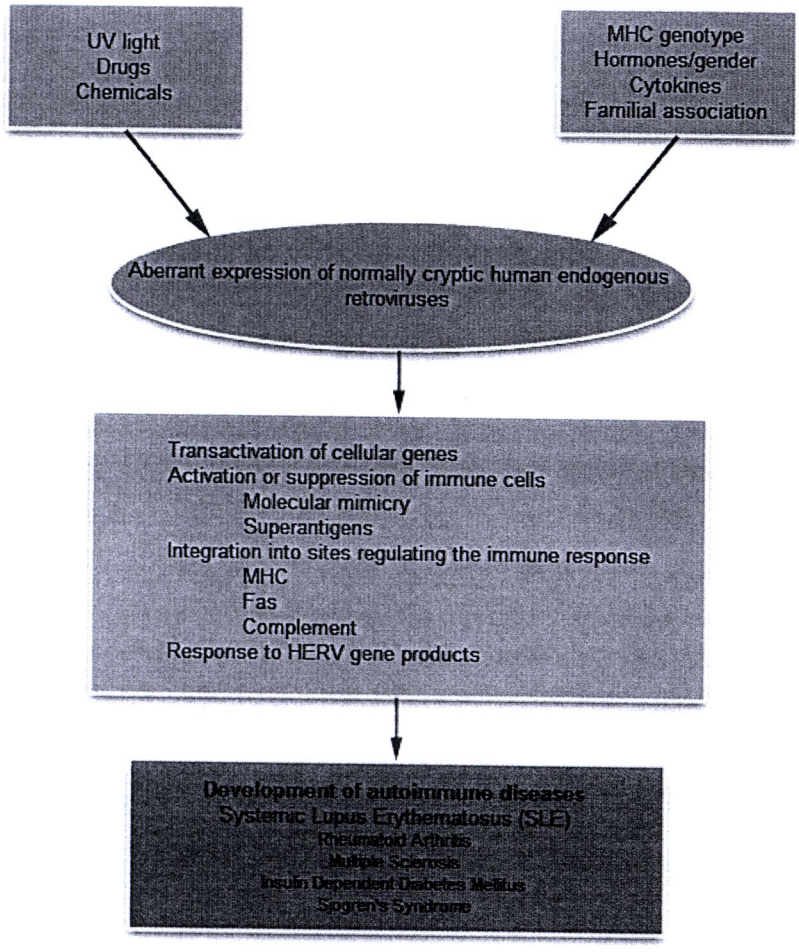


Figure 2.4 Summarization of purposed mechanisms used by human endogenous retroviruses in the etiopathogenesis of SLE and other autoimmune diseases [33]

2.3 Databases and tools related to HERVs

Although HERVs have been discovered for more than two decades, databases and tools related to the HERVs that have been developed seem to be limited in number. In this subsection, the databases and tools currently supporting the studies on HERVs are described. Because the tools currently provided are all limited for detection of HERVs and there are several of them found, the tools would be described in brief for each of them then.

2.3.1 Repbase Update (RU)

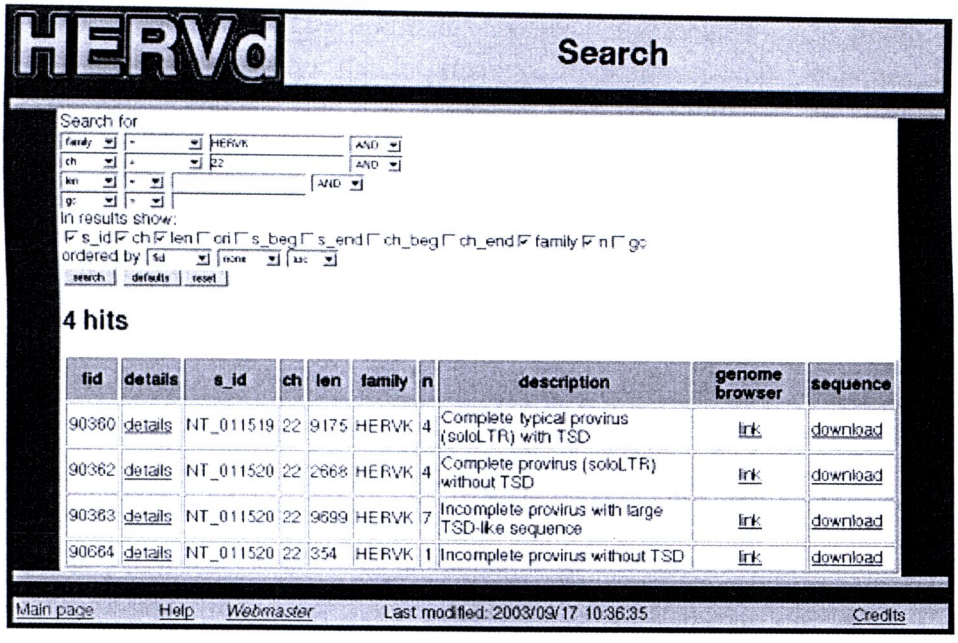
Repbase Update (RU) is a widely used database of repetitive and transposable elements, including HERVs, from human and other eukaryotic organisms. This database has been developed since 1990 to achieve a mission of Genetic Information Research Institute

(GIRI) [17]. The consensus sequences of many repetitive families and subfamilies are all collected in this database. Therefore, RU is being used as a reference collection in making and annotation of repetitive DNA by using computer programs, such as RepeatMasker and CENSOR. In addition to the collection, a systemic classification and nomenclature of the repetitive elements was also developed and implemented in RU. Currently, RU contains more than 7,600 sequences of transposable elements and other repeats, including those reported in the literature and those reported in only Repbase [39]. RU is available online for searching and downloading at http://www.girinst.org/Repbase_Update.html (last viewed on December 19, 2011).

2.3.2 HERVd

HERVd is a database of human endogenous retroviruses or HERVs developed since 2001. The database is maintained at the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, and is freely accessible at http://herv.img.cas.cz (last viewed on December 19, 2011) [9]. A collection of the retroviral elements found in the human genome and their information, retrieved from the computational analysis, have been provided in HERVd.

The information in this database is based on the repetitive elements and HERV portions identified by the RepeatMasker and a defragmentation algorithm developed by them. The database can be searched by HERV families, chromosomal locations and several other features as shown in Figure 2.5.



fid	details	s_id	ch	len	family	n	description	genome browser	sequence
90360	details	NT_011519	22	9175	HERVK	4	Complete typical provirus (soloLTR) with TSD	link	download
90362	details	NT_011520	22	2668	HERVK	4	Complete provirus (soloLTR) without TSD	link	download
90363	details	NT_011520	22	9699	HERVK	7	Incomplete provirus with large TSD-like sequence	link	download
90664	details	NT_011520	22	354	HERVK	1	Incomplete provirus without TSD	link	download

Figure 2.5 An example of the search page in HERVd

2.3.3 RetroSearch

In 2004, additional information, annotated open reading frames in the HERVs, was computationally analyzed and provided for all HERVs on the database of ORF annotated HERVs named RetroSearch. In brief, the HERV-related sequences were

identified by BLAST and Gag, Pol and Env were then predicted according to homology to known retroviral proteins. All data contained in the database are available at http://www.retrosearch.dk (last viewed on December 19, 2011) [10]. The first page of RetroSearch is shown in Figure 2.6.

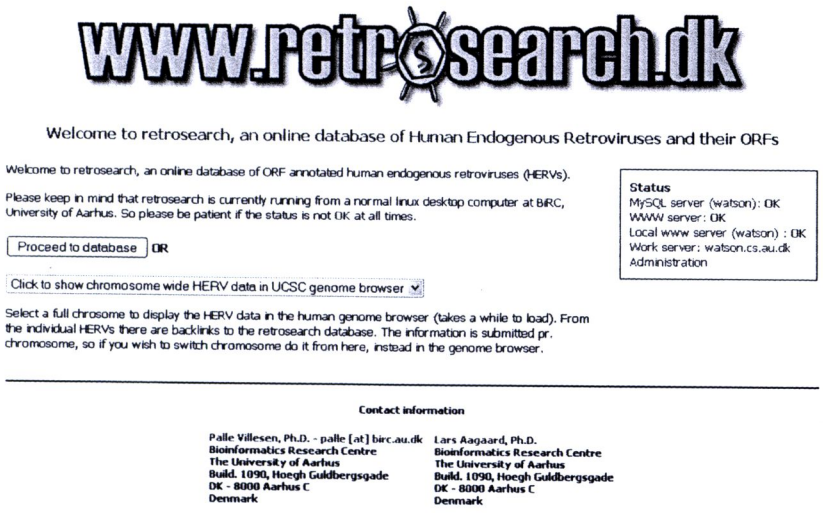


Figure 2.6 The first page of RetroSearch

2.3.4 TranspoGene

TranspoGene is a database of transposable elements (TEs) located inside protein-coding genes. This database provides the information for seven species, including human, mouse, chicken, zebrafish, fruit fly, nematode and sea squirt. In the database, the transposable elements with supporting evidences, such as RNAs, were included and classified into four categories: proximal promoter TEs, exonized TEs (insertion within an intron that led to exon creation), exonic TEs (insertion into existing exon), or intronic TEs [11]. TranspoGene is available at http://transpogene.tau.ac.il (last viewed on December 19, 2011). The first page of TranspoGene is illustrated in Figure 2.7.

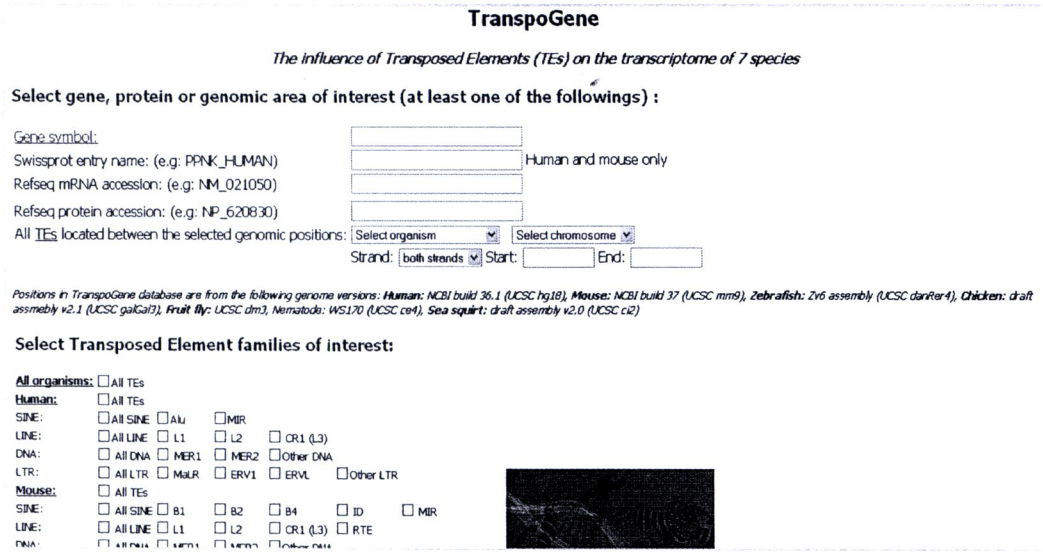


Figure 2.7 The first page of TranspoGene

2.3.5 Tools for detection of HERVs

The tools for the detection of HERVs have been developed based on several different principles. The first approach uses a set of reference sequences of the HERVs to detect the HERV-related regions in a genome. The repository is frequently employed for the purpose is Repbase. The tools based on this approach are RepeatMasker and CENSOR [2]. These tools generally used Smith-Waterman nucleotide alignment to output masked genomic DNA and a tabular summary of the detection. RepeatMasker has been reported that it efficiently detects most of the HERVs [40].

Another approach is based on the detection of retrovirus-like structure. The tools using this approach usually focus on detecting some of the internal proviral structures, such as reverse-transcriptase motifs, or other conserved motifs from *gag*, *pro*, *pol*, and *env*. Examples of tools based on this approach are RetroTector, LTR_Struc, and Genome Parsing Suite (GPS) [2].

Typically, HERVs are computationally identified as many fragmental matches instead of one with a long gap, due to large insertions and deletions accumulated during the evolutionary time. Therefore, a post-processing step, known as defragmentation, is often required to join fragments of the same element to achieve more biologically meaningful annotation [40]. There are several tools and scripts provided for this purpose, such as ProcessRepeats, LTR_MINER, Transposon Cluster Finder (TCF), MATCHER, and REannotate [41].

2.4 Fisher's exact test

Fisher's exact test is a statistical significance test for categorical data to infer about the difference between two population proportions. This is one in a class of exact tests, providing the exact probability of obtaining the observed data under the null hypothesis. Unlike the exact tests, an approximation test, such as the chi-square test, always provide

the estimated probability that would become reliable when the sample size is big enough [42]. Therefore, the Fisher's test does not depend on any large-sample distribution assumptions, and so it is appropriate even for small sample sizes.

The null hypothesis is usually based on that the relative proportions of both populations are not different, while the alternative hypothesis can support less-sided, greater-sided, or unequal comparisons between two proportions. The most common use of the Fisher's test is for 2×2 tables of the observed data, so called the 2×2 contingency tables (Table 2.2) [43].

Table 2.2 A 2×2 contingency table

Population	**Count of class I**	**Count of class II**	**Total**
1	x	$n_1 - x$	n_1
2	y	$n_2 - y$	n_2
Total	m	$n - m$	n

According to Table 2.2, the numbers of samples from the population 1 and 2 are n_1 and n_2, respectively, and n is the summation of both. Let x and y represent the numbers of the observed variable values as of class I and II, respectively, and m is the summation of both. The probability of observing a particular value of x, that is, the probability of a particular table being observed, is given by

$$P(x = k) = \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n}{m}},$$

where

$$\binom{n_1}{k} = \frac{n_1(n_1 - 1)(n_1 - 2)\cdots(n_1 - k + 1)}{k(k-1)(k-2)\cdots 1}$$

To test the difference in the two population proportions, the p-value of the test is the summation of the probabilities of all other possible tables in the way to support the alternative hypothesis. For example, if the alternative hypothesis is $H_a: \pi_1 > \pi_2$, where π represents a population proportion, we need to determine which other possible 2×2 tables would provide stronger support of H_a than the observed table. Therefore, the p-value of the case can be calculated by

$$P\text{–value} = P[x \geq k] = \sum_{j=k}^{\min(n_1,m)} \frac{\binom{n_1}{j}\binom{n_2}{m-j}}{\binom{n}{m}}$$