CHAPTER 1 INTRODUCTION

1.1 Background and rationale

In the human genome, surprisingly, there are DNA sequences that are highly similar to the present-day infectious retroviruses. These DNA sequences are termed Human Endogenous Retroviruses or HERVs [1]. These retroviruses represent the remnants of ancient infections that became incorporated in the germ line during the early primate evolution [1, 2]. Various HERV families constitute up to 8% of the human genome. This is substantial when compared to the proportion of the protein-coding genes which is only around 3% of the human genome [3].

A provirus, a retroviral sequence early integrated into the host genome, is generally composed of two main parts: long terminal repeats (LTRs) and internal retroviral genes. Inside the provirus, there can be many different internal regulatory regions, such as promoters, enhancers, polyadenylation signals, and splice sites, which most of them are located in the LTRs. HERVs have similar genomic structures to the provirus, but contain a lot of mutations accumulated, especially in the internal genes, during the evolution [1]. As a result, most of the HERVs have now lost of the ability to produce functional retroviral proteins. In contrast, their LTRs often still harbor functional regulatory sequences [4].

Because of their substantial quantity presenting among the human genome and their retained potentials of the regulatory sequences, HERVs have been viewed as probable influential factors of the transcription of the genes near them. Promoters inside the HERV LTRs can serve as the alternative promoters to additionally initiate the transcription of the neighboring genes with tissue- and lineage-specificity [4, 5]. Human functional genes that were reported to be related to the HERV promoters are, for instance, apolipoprotein CI (*APOCI*), human endothelin B receptor (*EDNRB*) [6], and the optiz syndrome gene *Mid1* [7]. Some HERVs can also provide a tissue-specific enhancer activity on genes nearby them, such as HERV-K (HML-2) in Tera-1 human testicular embryonal carcinoma cells [8]. The LEPRs, leptin obesity hormone receptors, have been found to exist in two variants that differ in size due to alternative splicing into a HERV-K LTR [5]. From all examples here, noticeably, the presence of nearby HERVs, whether in intra- or intergenic regions, may be significant to the gene function. Therefore, assembling genomic information and constructing a web-based resource of HERVs contribute substantially to the functional studies of HERVs.

For the past decade, there were few databases of HERVs designed for only searching of individual HERVs, such as HERVd [9] and RetroSearch [10]. The latest progress is the development of Transpogene [11] which allows users to search the transposable elements, including HERVs, located inside protein-coding genes. However, in this software, most of the HERV elements were excluded from the database, due to the restrictions of gene proximity consideration and lacking of supporting evidences. Thus,

there is still no tool that supports the investigations of neighboring HERVs of genes now.

Therefore, this thesis focuses on the development of a web-based tool, so-called HERV Profiler, which can be used to facilitate the studies on HERVs nearby functional human genes. The tool has been designed to provide two features, including visualizing and profiling neighboring HERVs of interested genes. In the visualizing HERVs, users can select to show neighboring HERVs in either tabular or graphical formats. For the HERV profiling, there are three sub-modules, including constructing HERV profiles, finding over-represented HERV types among a gene list, and ranking genes based on their over-represented HERV types. HERV Profiler has been tested with the generated gene sets to verify the implementations of each sub-module. After that, two case studies related to Systemic Lupus Erythematosus (SLE) would be demonstrated to investigate the neighboring HERVs of the SLE-relevant genes by using HERV Profiler.

1.2 Objectives

1. To develop a web-based tool for visualizing and profiling neighboring HERVs of human genes
2. To apply the developed tool to help studying the plausible relationships between SLE-relevant genes and their neighboring HERVs

1.3 Scope of work

To achieve the objectives of this study, the scope of work should be accomplished according to as these follow;

1. The data stored in the database were collected from the UCSC table browser, including the data of human repeats, human genes, and the cross-reference gene IDs annotated based on the most current version of the human genome (hg19, Feb. 2009).
2. The web-based tool was implemented with the aim to finally provide two main features, including visualizing the neighboring HERVs of the interested genes and profiling the neighboring HERVs of a gene list input.
3. After the tool implementation, the tool testing is required for checking whether the tool can accurately function or not.
4. By utilizing the web-based tool developed, the genes related to SLE disease obtained from microarray data analysis were studied on their neighboring HERVs.

1.4 Expected outputs

Upon completion of this works, the following outputs are expected:

1. The web-based tool for visualizing and profiling neighboring HERVs of human genes, which can facilitate the investigations of the HERVs near the interesting genes
2. The revealed relationships between SLE-relevant genes and their neighboring HERVs